The Benefit of Thinking Small in Big Data

Kurt Englmeier[1] and Hernán Astudillo Rojas[2]
[1]*Faculty of Computer Science, Schmalkalden University of Applied Science, Blechhammer, Schmalkalden, Germany*
[2]*Departamento de Informática, Universidad Técnica Frederico Santa María, Valparaiso, Chile*

Keywords: Big Data, Data Analysis, Human-Computer Interaction, Automatic Translation, Translation Memory.

Abstract: No doubt, big data technology can be a key enabler for data-driven decision making. However, there are caveats. Processing technology for unstructured and structured data alone–with or without Artificial Intelligence–will not suffice to catch up the promises made by big data pundits. This article argues that we should be level-headed about what we can achieve with big data. We can achieve a lot of these promises if we also achieve to get our interests and requirements better reflected in design or adaptation of big data technology. Economy of scale urges provider of big data technology to address mainstream requirements, that is, analytic requirements of a broad clientele. Our analytical problems, however, are rather individual, albeit mainstream only to a certain extent. We will see many technology add-ons for specific requirements, with more emphasis on human interaction too, that will be essential for the success in big data. In this article, we take machine translation as an example and a prototypical translation memory as add-on technology that supports users to turn the faulty automatic translation into a useful one.

1 INTRODUCTION

People produce worldwide every day zillions of data when giving away transaction data to retailers and banks, opinions posted on social media, or geospatial and health data recorded by our phone, just to name the most prominent ones. We funnel these zillions of data into big data machines that, thanks to their analytic power, produce insights we even haven't dared to dream of. The journalist and futurist Patrick Tucker sees a revolutionary advance emerging from analytics: "In the next two decades, we will be able to predict huge areas of the future with far greater accuracy than ever before in human history, including events long thought to be beyond the realm of human inference." (Tucker, 2014) big data champions adhere to this view and relentlessly allege that analytics can solve a long array of problems, ranging from crime over people's nutrition and health to education and global warming.

Thanks to tools like Hadoop and Spark we can in fact combine zillions of structured and unstructured information. However, the analytical power of tools for predictive analysis, correlation analysis, time series analysis, and the like hasn't changed much in recent decades (Jagadish et al. 2014). Operating these tools with all these data does by far not automatically result in valuable information. Gary Marcus and Ernest Davis (Marcus and Davis, 2014) compiled a nice list of caveats about big data. They present prominent examples where big data projects neglected the adage "correlation is not causation". The combination of database technology with tools like Apache Spark–for integrating and standardising unstructured data–and a versatile statistical apparatus forms a powerful technology for big data that in fact can process masses of data and produce masses of results. The mere availability of computational power does not automatically open avenues to information that supports data-driven decisions. There's no such thing like one-size-fits-all technology for big data. Currently available big data technology or "Data Science as a Service (DSaaS)" provide at best off-the-shelf features for data handling and analysis.

There are always tiny, but essential details not covered by big data technology. Just think about opinion mining or sentiment analysis. There are so many methodological, language, and domain specifics to adhere to. There is no standard approach to opinion mining. Too many continually changing individualities need to be covered. (Wright, 2009; Feldman 2013; Zhang and Liu, 2011) Thinking small in big data means to pay attention to these details. This article outlines at first that data analysis may

Englmeier, K. and Rojas, H.
The Benefit of Thinking Small in Big Data.
DOI: 10.5220/0006423301860191
In *Proceedings of the 6th International Conference on Data Science, Technology and Applications (DATA 2017)*, pages 186-191
ISBN: 978-989-758-255-4

consist of standard processes, but there is always an array of specialised processes to be taken into account. It further explains that analytic proficiency emerges from combining and orchestrating of three knowledge areas, namely tool knowledge, domain knowledge, and methodological knowledge. Big data analysis requires mainstream big data technology but also add-on technologies that enable the users to ensure that the applied technology completely supports their analysis hypotheses. The mix of required technology reaches far beyond analytics architectures that cover data warehouses, statistical tools, and dashboard features. Future big data technology will include also more tools that are controlled by the information consumer. In its second part, the article presents an example of such an add-on technology that enhances mainstream technology. A prototypical translation memory helps the user to correct the results of machine translation.

2 ANALYTIC PROFICIENCY

Economists, for instance, would welcome big data technology. They work with big data every day, since decades. The German ifo Institute for Economic Research issues every month one of the most reliable indicator for Germany's economic development, the Ifo Business Climate Index (ifo, 2017). It asks companies across all sectors to assess their current and future business situation. ifo gets thousands of replies indicating whether current situations are considered "good", "satisfactory" or "poor". Furthermore, it asks the companies whether their business situations for the next six months presumably will be "more favourable", "unchanged" or "more unfavourable". The answers are merged into sector-wide and economy-wide indicators. Summarising the respondent's feedback is certainly part of the off-the-shelf features of analysis software. However, companies are different in size and market power. This needs to be taken into account. The business situation of a company like Siemens has an impact on the overall assessment of the situation across the electrical and electronic industry that outweigh that of a small company serving a niche market with a handful of products. Merging the companies' feedbacks into one indicator is only valid if their different market positions are considered correspondingly in the statistical model. Each company feedback gets an individual weight that reflects the company's impact on the sector's economic performance. Feeding the respondents' answers and their individual weights into a statistical tool is a standard procedure. Defining these individual weights, however, reaches far beyond any standard feature. It requires solid knowledge of the market sector and continuous observation of the company's performance.

The consolidated values are then added to the time series reflecting the sector's or the economy's performance. Finally, the whole series get seasonally adjusted, because public, regional and major plant holidays distort the picture of a sector's performance. The result is a pretty realistic picture on the performance of the sector economy or the country's economy as a whole. The moving holiday components of seasonal adjustment, in turn, vary from country to country, sometimes there are even regional and sectoral differences. Hence, economists need to fine-tune them in accordance to the market in consideration. The two examples from economic analysis show: no matter how powerful analytic tools are they need to be fine-tuned, that is, adjusted to the specific individual purposes. Furthermore domain experts constantly need to check the plausibility of the results produced. We cannot simply funnel data into machines—as intelligent as they might be—and expect to get the results we require.

Analytic features of database systems, statistical software or data mining tools have clear functional profiles. Correlation analysis, trend analysis, regression analysis, and the like are instruments quite well-known to statisticians. They know how to apply them, how to fine-tune them for specific purposes, and—by involving the expertise of domain experts—how to check the plausibility of their results. Not every correlation detected automatically has a plausible meaning. In their article, Marcus and Davis give some funny examples of correlations that went awry. Quality assurance in data analysis is of outmost importance. It rests on a combination of methodological, tool, and domain knowledge and expertise or on data science and domain expertise (Liu et al, 2014). Not everything that is computationally possible is also plausible. Big data offers great opportunities, but it does not unleash the next wave of digital disruption. Big data has the potential to distract and drain budget and resources. And it is even more daunting if major investments in big data technology do not yield the informative quantum leap we expected so much. Analytic tools are powerful and can squeeze an avalanche of additional data out of the data we already have. The results we get may be interesting, exhilarating, daunting, out of focus, irritating, or overwhelming. The power of modern computing can help us solve many problems. However, we need to be level-

headed about the capabilities of big data, in particular if we are prone to believe in computing power. Just crunching numbers does not yield us all the hints we need to solve problems. It is naive to believe that big data produces all the information we require, we just have to grab what we need.

In our everyday jobs, there are many small things to discover in big data. And there are probably too many things that go unnoticed by big data technology. The problem is not so much fine-tuning of big data tools that is complicated and time-consuming and requires a lot of tool knowledge. Together with statisticians and domain experts we can design and implement reliable analytic features that produce results in accordance with our quality standards.

Quite often we even look for insights from data where the whole big data machinery is not of much help. There is thus a new concept emerged recently called Small Data or Little Data. The concept sounds a bit misleading at first, because it is sometimes referred to as data analytics considering small pieces of data processed by a spreadsheet application. The attribute "small" refers more to the complexity of the problem we address, in terms of the complexity of the analytic processes required to get the insights to solve this problems. In simple terms, for the majority of analysis problems we do not need sophisticated statistical models or machine learning algorithms. The attribute "small" refers also to the individuality of the problem, that is, there only a small community of information consumers requiring a big data solution for a particular problem.

Big data has the potential to leverage data-driven decision making to a higher level of quality and efficiency. From a first glance on the topic big data, we might get the impression that the processing power of big data technology can turn raw data into useful information that discloses actionable insights to us. The more we get into this topic, the more we realise that success in data-driven decision making is only partly technology-driven. The example from economic research demonstrates that we most likely need to combine tool knowledge, domain knowledge, and methodological knowledge if we want to benefit from big data. Each analysis project requires a different mix of knowledge from these areas, that is, each knowledge area contributes only to a certain degree to the required proficiency. Not each of these knowledge areas may even be required in some big data project. The success of and the benefit from big data thus depends on our capabilities to find the right knowledge and technology mix.

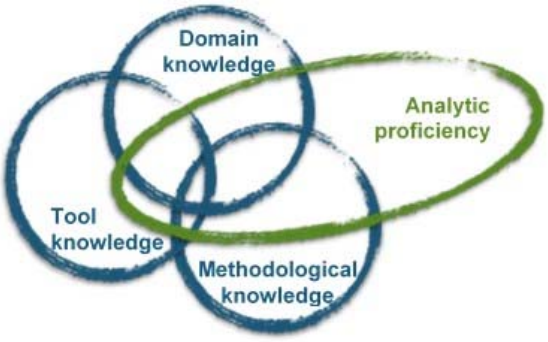

Figure 1: We get the insights we expect from big data if we efficiently combine domain knowledge, methodological knowledge and knowledge of and experience with analytical tools.

Essentially, with big data we see more and highly distributed data emerging from more places at a higher rate. This attribute reflects mainly the technological challenge of big data. Many of these data are unstructured and need to be transformed, standardized, and integrated into our information environments. Big data technology usually comes with methods to distil useful information from raw data as long as its respective functionalities cover mainstream requirements, that is, they have features requested by a broad user community. The actual discussion about big data is rather technology-driven and has a strong focus on the development of tool knowledge. Success in big data requires more than just powerful tools as the success of analysis projects depend on different knowledge areas and even on interdisciplinary teams working together.

3 THE SMALL THINGS IN BIG DATA

Distilling information from unstructured data and/or from the combination of unstructured and structured data, however, can reach quite quickly beyond mainstream requirements. For instance, gathering unstructured or text data from social media, generating insights about products and services, and developing chatbots for customer service and engagement, require quite individual text mining and communication features. They depend on the domain and the nature of information we want to mine, such as opinion, customer request, complaint, etc. This means our distilling requirements can quickly demand more domain and methodological knowledge as we might have considered in the first place. We

may even find ourselves suddenly completely outside the focus of mainstream big data technology.

The mere quantity of new emerging information is useless if it is too imprecise for the information consumers to act upon. Thinking small means to take care of the small and individual requirements that are not covered by mainstream big data technology. Even though mainstream technology takes us close to the insights we yearn to obtain from big data, in most analytical situations we do not get useful results without fine-tuning the tools we use. Discussions around thinking small in big data or small data (Hendershot, 2015, Kavis, 2014) is pointing out that the Internet of Things (IoT) produces mainly small data that do not require big analytic machinery. The remote control of devices of our home (heating, blinds, lights, etc.) does in fact not depend on many data. Sensors mostly produce only small pieces of data.

To recognize an event, big data technology is completely obsolete in many cases. Take an app on a smart watch, for instance, that can be used to detect if a person wearing that watch needs immediate help because she or he had collapsed and remained motionless on the floor for a critical while. There are not masses of data required to distinguish the different movements of the watch. The app must clearly distinguish the movement when a person is collapsing from similar but less critical situations like the watch falling down or being put on the table.

Many devices of an aircraft provide status information. The landing gear, for instance, has a series of sensors attached that produce a series of small data during touchdown. These data help to assert if a hard landing occurred and, therefore, landing gear and wheels need a special check before the next take-off.

The two examples show that in these situations tool knowledge is not so much required to get useful and actionable information. The focus is more on domain and methodological knowledge, plain physics in these cases. When do speed and direction of the smart watch point to a falling person or when does compression and angle of the landing gear indicate a hard landing? In data analysis, the hard work is to formulate and test a solid hypothesis about the data patterns that indicate the event we expect to observe. Knowledge in physics is essential in both cases that is unsurprisingly not covered much by mainstream analytic technology.

If we take all data together that are produced by the sensors of an aircraft during take-off, en route, and landing, we see, in fact, masses of data. Big data technology, such as database systems and integration tools like Spark (or Hadoop) can certainly support quite a lot in data handling. Even though we do not get all data required for the information we expect, they lay important groundwork work for efficient data analysis. Nevertheless, in many cases remains essential data processing and analysis work to be done. Actually, big data is mainly addressed as a technical challenge requesting new capabilities that reach beyond current database technology and embrace features and methodologies from Artificial Intelligence (AI). Both fields have an essential impact on the development of successful tools and technologies to master the big data challenge. However, economy-of-scales limit their development to generic analytic needs. Final process steps towards specific analytic solutions addressing individual needs of the information consumers are widely neglected or considered as a problem for statisticians and programmers. It is obvious that a wide variety of tools is required to cover these individual needs. Furthermore, information consumers increasingly expect to handle these tools by themselves. They want to avoid the delegation of analytical tasks as far as possible. In the future, we will see many tools that enable the information consumer to adapt and fine-tune the results of big data analysis to their individual needs.

4 DISCOVERY LIFECYCLE

Hence, we suggest to shift the focus a bit from the technical capability of big data technology to the roles and needs of the information consumers. Undoubtedly, big data tools can do the heavy lifting in data analysis and intelligent data processing. The final and indispensable fine-tuning of the results, however, reaches beyond this capability. Human-centred design of add-on tools or features for big data technology closes the gap between the needs of information consumers (including analysts) and technological capabilities.

We take machine translation as an example to present the nature of a human-centred technology component for big data. Automatic translation services can translate large quantities of text within seconds. As we all know, automatic translations contain many, probably too much errors and thus do not satisfy our expectations, in general. There are many types of errors, like wrong sentence structure, incorrect translation of pronouns, and the translation of a word inconsistent to the context of the phrase or the text. The first two classes of errors can be easily corrected by humans. The latter class of errors is more

difficult to handle. The quality of translation in general and of translation support depends on words reflected in the right context. The correct disambiguation of the meaning of a word is key to correct translations. Translation memories (TM) can help the users to find suitable translations for key expressions in a given context. TMs record the translation work and serve as a digital vocabulary for specific translations in the domain(s) of the users (Casacuberta et al., 2009). This means, they contain enough information to set up the required domain context. In the following we present a prototypical TM that supports the retrieval of context-specific translations of critical words or expressions.

The following example gives an impression of benefits and odds in machine translation. The benefit of machine translation is indisputable: You get the translation at a fingertip. However, it does not correctly grasp the context of the word to be translated. For example, much like many other words, the word “stock” can carry different meaning depending on the context. Its correct translation into Spanish or German, for instance, requires context information. Does it refer to stock market, stock-keeping or cooking? In the first case, it is translated by “acciones” or “Aktien”. In the cooking context, “caldo” or “Brühe” are suitable translations. Machine translation works well if the immediate context is doubtlessly known to the system. “100 mls stock or water” is correctly translated into Spanish: “100 ml de caldo o agua”. If we add a cooking instruction, the machine translation suddenly loses the context! “100 mls stock or water. Bring the stock to the boil in a large pot.” is translated into “100 ml de caldo o agua. Llevar la población a hervir en una olla grande.” (see figure 2). The German translation is similarly wrong (see figure 3).

The reason for the inaccurate translation of machine translation emerges from so-called concordance lists, which, in turn, result from numerous translations. In simple terms, they store the translation frequencies of words, frequencies of co-occurrences of words, and the frequencies of certain word sequences. If the word "stock" follows the term "100 mls", the probability of the word "caldo" being the correct Spanish translation is very high. If more words follow then the probability quite likely changes. Correlation between "100 mls" and possible subsequent expressions (including “stock”) may now point to the expression “la población” or “las acciones” as correct translations. The human translator immediately recognizes that both expressions lie outside the context of the phrase to be translated. Machine translation, in contrast, cannot not recognize that all sentences have to do with cooking and consequently takes the most probable translation according to the frequencies mentioned above.

The translation work of the users appears almost always in a specific thematic context: translation of an operating manual for a specific product, the pages of a web site or the user interface for a cooking recipes app, all these translations have a thematically narrow focus. Our prototypical TM serves the context-driven translation of words. As soon as contexts are available, it determines whether word use is compatible to a context and, if necessary,

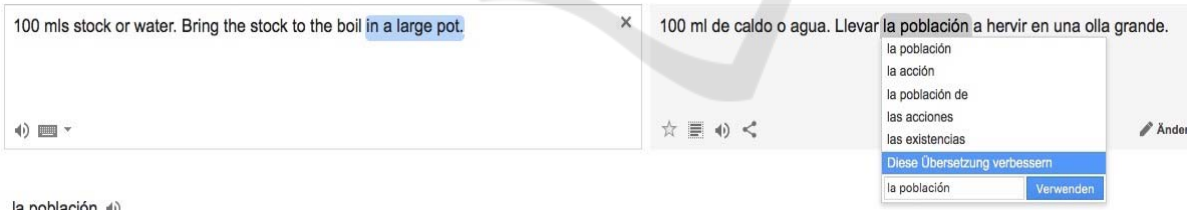

Figure 2: Example for machine translations using Google’s translator: if only the first phrase is entered the system correctly identifies the context of the word “stock” and translates it correctly (“caldo”). As soon as the second phrase is added it no longer recognizes the correct context.

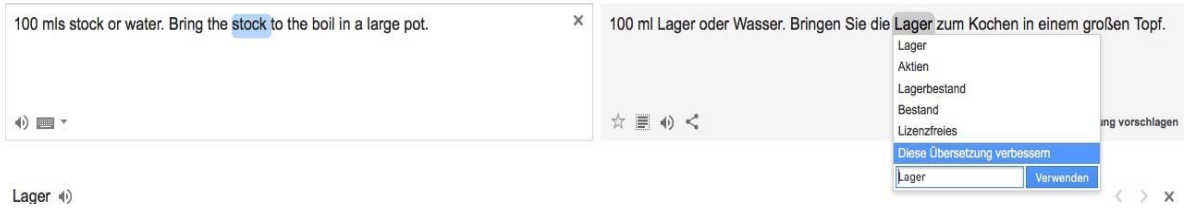

Figure 3: If translating the same phrases into German the system loses the correct context of the word “stock” as well when the second phrase is added and uses the wrong translation for both occurrences of “stock”.

derives improvement suggestions, that is, recommends words or expressions that are more appropriate in the given situation.

5 CONCLUSIONS

Big data pundits, mainly the providers of database technologies, allege that big data technology creates a new generation of analytical tools that help us to solve many problems in crime prevention, health care, and global warming, just to name a few. There is no doubt, big data technology is very good at detecting correlations. However, it does not necessarily mean that there is a plausible causation behind every correlation detected. Even if we focus just on correlation analysis, we have to admit that it can get quite complex and versatile. Usually, it starts with forming a hypothesis, that is further transformed to an analysis model and finally to a set of instructions for the analytic tool. The results are then checked against the hypothesis, the model is adapted and tested again. After a couple of iterations the causation behind the correlation is ideally validated. We can easily imagine that there any many and subtle parameters to fine-tune in this analytic cycle. Hence, there is no such thing like general purpose big data analysis.

Big data technology can take us quite close to the information we require. However, in many situations we need to adapt the technology or we need additional tools for the essential final processing step. The actual discussion in big data bypasses this missing link in data analysis. For a successful big data architecture, we need a broader technology mix that reaches beyond database, statistical, and dashboard features. Without closing the technology gap in analysis we cannot fully leverage big data. There is a series of tools required that ranges from domain to domain and from methodology to methodology. Even if we just look at text mining, there are standard mining tools for texts, but things like opinion mining and analysis of customer requests or complaints, for instance, can get very specific and reach definitely beyond mining features. The automatic detection of positive, neutral or negative opinions in social media about products or service of a specific company needs to be gauged continuously. It can be quite beneficial if the information consumers, in this case the decision-makers, can immediately adapt the mining features.

This article outlines the problem of the technology gap in the context of automatic translation. With an additional tool that enables the users to retrieve essential expression in the right context, they can correct diction errors that occur frequently in machine translation. Even though machine translation does the heavy lifting, the users need to correct its results. Only by this final step they get the expected quality in their translations. We consider the context-sensitive translation memory as an example among many, many add-on tools that will be required to produce the insights we can expect from big data.

REFERENCES

Casacuberta, F., Civera, J., Cubel, E., Lagarda, A.L., Lapalme, G., Macklovitch, E., Vidal, E., 2009. *Human interaction for high-quality machine translation*, Communications of the ACM, vol. 52, no. 10, 135-138.

Feldman, R., 2013. *Techniques and applications for sentiment analysis*, Communications of the ACM, vol. 56, no. 4, 82-89.

Hendershot, S., 2016. *Data done right*, Project Management (PM) Network, March 1, 2016, published on http://www.pmi.org/learning/library/data-done-right-project-success-9989, retrieved on March 13, 2017.

ifo Institute for Economic Research, *Ifo Business Climate Index*, http://www.cesifo-group.de/ifoHome/facts/Survey-Results/Business-Climate.html, retrieved on February 28, 2017.

Jagadish, H.V., Gehrke, J., Labrinidis, A., Papakonstantinou, Y., Patel, J.M., Ramakrishnan, R., Shahabi, C., 2014. *Big data and its technical challenges*, Communications of the ACM, vol. 57, no. 7, 86-94.

Kavis, M., 2015. *Forget big data - Small Data Is Driving The Internet Of Things*, Forbes, February 25, 2015, published on https://www.forbes.com/sites/mikekavis/2015/02/25/forget-big-data-small-data-is-driving-the-internet-of-things/#6f1ab9cb5d7e, retrieved on March 13, 2017.

Liu, J., Wilson, Gunning, D., 2014. *Workflow-based Human-in-the-Loop Data Analytics*, Proceedings of the 2014 Workshop on Human Centered big data Research, 49-52.

Marcus, G., Davis, E., 2014. *Eight (No, Nine!) Problems With big data*, New York Times, April 6, 2014.

Tucker, P., 2014. *The Naked Future: What Happens in A World That Anticipates Your Every Move?*, Current.

Wright, A., 2009. *Our sentiments, exactly*, Communications of the ACM, vol. 52, no. 4, 14-15.

Zhang, L., Liu, B., 2011. *Identifying noun product features that imply opinions*, HLT '11, Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers – vol. 2, 575-580.