Data Rich – But Information Poor

Peter Bernus[(✉)] and Ovidiu Noran

IIIS Centre for Enterprise Architecture Research and Management,
Griffith University, Brisbane, Australia
{P.Bernus,O.Noran}@griffith.edu.au

Abstract. The article describes the missing link between the information type and quality required by the process of decision making and the knowledge provided using the recent developments of 'big data' technologies, with emphasis on management and control in systems of systems and collaborative networks. Using known theories of decision making, the article exposes a gap in present technology arising from the disparity between the large amount of patterns that can be identified in available data using data analytics, and the lack of technology that is able to provide a narrative that is necessary for timely and effective decision making. The conclusion is that a second level of situated logic is necessary for the efficient use of data analytics, so as to effectively support the dynamic configuration and reconfiguration of systems of systems for resilience, efficiency and other desired systemic properties.

Keywords: Situation awareness · Big data · Decision model · Situated reasoning

1 Introduction

The ability to gather very large amounts of data has preoccupied the research community and industry for quite some time, starting within Defence in the 1990s and gradually evolving in popularity so as to become a buzzword in the 2010s. 'Big data', the 'sensing enterprise', and similar concepts signal a new era in which one hopes to provide 'all necessary' decision-making information for management in the quality and 'freshness' required. Whoever succeeds in building such facility and use the data to derive decision making information efficiently and effectively, or produce new knowledge that was not attainable before, will 'win'. The hope is well founded, if one considers an analogy with the early successes of evidence-based medicine (Evidence Based Medicine Working Group 1992, Pope 2003) that transformed the way the medical profession decided what treatments to use. At the same time, it is necessary to define realistic expectations in order to avoid expensive mistakes; for example, even evidence-based medicine that relies on large scale data gathering through clinical trials and careful statistical analysis is showing signs of trouble (Greenhalgh et al. 2014) with the usefulness of the evidence gathered when it is applied in complex individual cases increasingly being questioned. Generally, it is becoming obvious that finding new ways to correctly interpret complex data in context is necessary – a claim supported by many authors (HBR 2014).

An obvious viewpoint is that when intending to use large amounts of gathered data to create useful decision making information, one must carefully consider the information needs of management and especially how the interpretation of data is influenced by context. The goal of this paper is to investigate and analyse, from a theoretical perspective, what missing links exist in order to put 'big data' to use apart from the obvious 'low hanging fruit' applications.

2   Current Issues in Big Data and Data Warehousing

The role of Management and Control (or Command and Control) is to make decisions on multiple levels, i.e.: Real time, Operational, Tactical and Strategic, often guided by various types of models. Along these lines, two notable systematic models of decision making (which is the essence of management and control) are the GRAI Grid (Doumeingts et al. 1998) and the Viable Systems Model or VSM (Beer 1972). These models are very similar and differ mainly in the details emphasised (Fig. 1). Fundamentally, these generic models identify management, command and control tasks and the information flow between them.
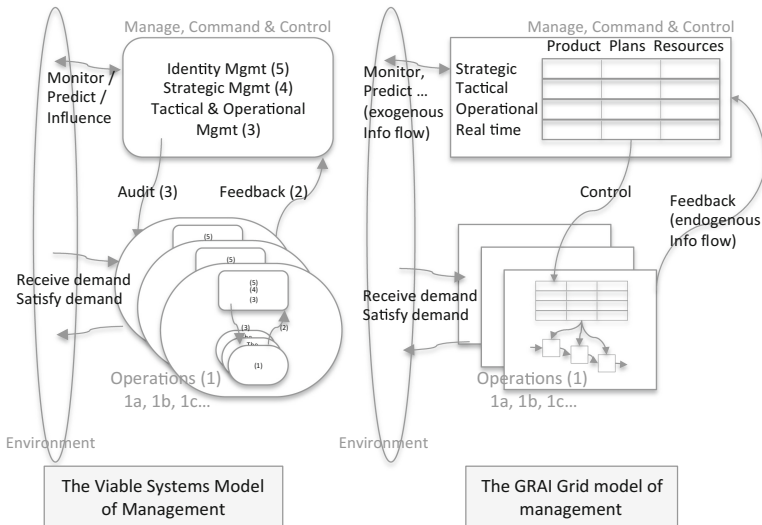

Fig. 1. A simplified view of two equivalent decision making models: Viable Systems Model (Beer 1972) and GRAI Grid (Doumeingts et al. 1998)

To make successful decisions, it is necessary to satisfy the information needs of the management functions as depicted in Fig. 1. The big data movement (discussed in Sect. 1) has a 'little brother': data warehousing (with its associated techniques), which made similar initial claims of creating meaningful insight for management. Even though there exist many success stories, there were some notable failures of data warehousing to deliver on its promises, owing to similar causes (see below).

In data warehousing, the methodology suggested by its proponents was to take copies (snapshots) of operational databases (and other data repositories) and build an interface, based on which the data could be 'mined' (analysed) to find management-relevant information. To build such a facility fast and in an affordable manner, the methodology suggested by Inmon (1992) and Kimball (1996) and others was to create the data warehouse out of existing databases and possibly transaction logs so as to gain management insight. In other words the aim was to create a *narrative* that is characterising the present or predicted future situation, which is essential for strategic decision making. Big data is no different, using traditional data analysis and machine learning techniques the protagonists derive useful interpretations, just on a larger scale than data warehouses can do, as data sources are larger and include multiple sources. However, both the big data and data warehousing movements (even though their scales are different) have in fact similar shortcomings, such as:

- The associated methodologies give not enough weight to first understanding the fundamental information needs of the decision maker;
- Very little is done to *correlate* internal- and external data sources to create useful information for decision making (i.e., relating the endogeneous and the exogeneous);
- Insufficient effort was put into realising what data *would be* needed but was unavailable (for being able to draw useful inferences); Even the recent method to limit the amount of sensor data taken into account in situation assessment – while providing the facility to switch 'on or off' additional pre-stored sensor data sources – is relying on the commander to pinpoint what data should be taken into account to possibly change the narrative;
- If the above deficiency is identified, then the need for data that is not available, but is deemed necessary, may become the source of additional data collection tasks. However, this can inadvertently result in poor data quality (HBR 2014, p. 47 and Hazen et al. 2014, p. 78). This is because the essential but problematic Information Systems point of view of the data gathering task is ignored (i.e., how to avoid data quality problems due to data entry by humans who consider it a chore), in favor of only solving the easier to manage database/computing problems (how to use various algorithms to identify patterns in data, which problem is essentially technical in nature);
- Very little has been done on transforming existing processes so that they would *produce the necessary data as a by-product of the production* (or service delivery) *process*, instead of requiring additional data entry (which was found to be the main source of data quality issues) (Hazen et al. 2014);
- There has been a propensity to disregard the *context* of the collected data, and thus creating the danger of situation mis-identification without even being aware of having committed this mistake (Santanilla et al. 2014).

The data warehouse movement largely concentrated on collecting and making available for analysis the *internal* data of the enterprise, while the big data movement, with its roots in the 1990s (e.g. the 'Dominant Battlefield Knowledge' movement (Szafranski 1995)), concentrates on the ability (initially, of the military) to access an unprecedented

amount of *external* data so as to gain advantage over adversaries - and indeed, specialised systems have been built and deployed in order to achieve this goal.

Two issues become apparent when analysing the history of creating useful decision-making information, whether through data warehousing or big data analytics and the associated business intelligence processes:

1. On each decision-making level, we must correlate internal *and* external data;
2. With the opportunity to collect and access very large amounts of data, it becomes difficult to identify patterns that are useful for decision making (there are too many patterns that algorithms can identify) – unless one uses heuristics (i.e., the result of prior learning) to discern what is relevant and what is not. Importantly, the measure of relevance changes with time and with the current interpretation of data.

3   Making Effective Decisions

Tasks that appear in each type and level of decision-making and the feedback that can be used to inform the filters used to selectively observe reality may be studied using a finer model that explains how successful decisions are made. This filter is part of the well-known Observe, Orient, Decide and Act (OODA) Loop devised by John Boyd (Osinga 2006) (see explanation below). It must be noted that on closer inspection, it turns out that OODA is *not* a strict loop because it is precisely the feedbacks inside the high level 'loop-like' structure that are responsible for learning and for decisions about the kind of filters necessary. Note that this 'loop' is often misunderstood to be a strict sequence of tasks (e.g. *cf.* Benson and Rotkoff's 'Goodbye OODA Loop' (2011)) when in fact it is actually an *activity network* featuring rich information flow among the OODA activities and the environment.

A brief review of Boyd's OODA 'loop' can be used to point to potential development directions for a 'big data methodology' for decision support. Accordingly, decisions can be made by the management/command & control system of an entity, in any domain of action and on any level or horizon of management (i.e., strategic, tactical, operational and real-time, performing four interrelated tasks:

- Observe (*selectively perceive data [i.e., filter]* – measurement, sensors, repositories, real-time data streams, using *existing* sensors);
- Orient (*recognise* and become *aware of the situation*);
- Decide (*retrieve existing-*, or *design/plan new* patterns of behaviour);
- Act (execute behaviour, of which the outcome can then be observed, etc.).

According to Boyd and subsequent literature analysing Boyd's results (Osinga 2006), to make the actions effective, decision makers must repeat their decision making loops faster than the opponents, so as to disrupt the opponent's loop. Note that the first caveat: one can only 'observe' using *existing* sensors. Since there is no chance to observe absolutely everything, how does one know that what is observed is relevant and contains (after analysis) all the necessary data, which then can be turned into useful situation awareness (Lenders et al. 2015)? The likely answer is that one does *not* know *a priori*; it is through learning using past positive and negative experiences that a decision system

approaches a capability level that is timely and effective in establishing situation awareness. This learning will (or has the potential to) result in decisions, which are able to identify capability gaps (and initiate capability improvement efforts). This is the task of *self-reflective management*, comparing the behaviour of the external world and its demands on the system (the future predicted action space) with the action space of the current system (including the current system's ability to sense, orient, decide and act). In this context, the 'action space' is the set of possible outcomes reachable using the system's current technical-, human-, information- and financial resources.

This learning loop depends on self-reflection and *it is in itself an OODA loop* analogous to the one discussed above, although the ingredients are different and closely associated with strategic management. The questions are: (a) what to observe, (b) how to orient to become situation aware and (c) what is guiding the decision about what to do (within constraints and decision variables and the affordances of action) so as to finally be able to perform some move. The action space of this strategic loop consists of transformational actions (company re-missioning, change of identity, business model change, capability development, complete metamorphosis, etc.).

Essentially, such strategic self-reflection compares the current capabilities of the system to desired future capabilities, allowing management to decide whether to change the system's capabilities (including decision making capabilities), or to change the system's identity (re-missioning), or both. Note that management may also decide that part of the system will need to be decommissioned due to its inability to fully perform the system's mission. Such transformation is usually implemented using a separate programme or project within a so-called Plan-Do-Check-Act (PDCA) loop (Lawson 2006, p. 102) (not discussed further as it is outside the scope of this paper).

The above analysis can be put to use in the following way: to achieve situation awareness, which is a condition of successful action, 'big data' (the collective technologies and methods of data analysis and predictive analytics) has the *potential* to deliver a wealth of domain-level facts and patterns that are relevant for decision-making, which were not available before. However, this data needs to be interpreted, which calls for a *theory of situations* ultimately resulting in a *narrative* of what is being identified or predicted; without such narrative, there is no true situation awareness, which can significantly limit the chances of successful action.

It is therefore argued that having the ability to gather, store and analyse large amounts of data using only algorithms is not a guarantee that the patterns thus found in data can be turned into *useful* information that forms the basis of effective decision-making, followed by appropriate action leading to measurable success.

The process works the other way around as well: when interpreting available data (however large the current data set may be), there can be multiple fitting *narratives* and unfortunately it is impossible to decide which one is correct. Appropriate means of reasoning with incomplete information could in this case identify a need for new data (or new *types* of data) that can resolve the ambiguity. Thus, supporting decision-making using 'big data' requires the collection of a *second level* of data, which is not about particular facts, but about the creation of a 'repertoire' of situation types, including facts that *must* be true, facts that *must be not* true, as well as constraints and rules of *causes and effects* matching these situation types. Such situation types can also be considered

as models (or model 'archetypes') of the domain, which then can be matched against findings on the observed data level. Given the inherently perpetual changing nature of the world, these situation types are expected to evolve themselves; therefore, one should *not* imagine or aim to construct a facility that relies on a completely predefined ontology of situation types. Rather, there is a need for a facility that can continuously improve and extend this type of knowledge, including the development and learning of new types that are not a specialisation of some previously known type (to ensure that the 'world of situations' remains open, as described by Goranson and Cardier (2013)).

4   Using Big Data in Decision-Making for System of Systems

The Dominant Battlefield Knowledge movement (Szafranski 1995) pointed out quite early that in spite of the ability to deploy a large number of sensors, in order to achieve situation awareness data needs to get filtered based on 'relevance'. The word 'relevant' is in inverted commas, because it is precisely the possible situations of interest that dictate what data are or would be relevant – however, one does not know *a priory* (and/ or unambiguously) what situation one is actually in! Therefore, the *continual narrative* of the situation changes the data needs (Madden 2012), as well as what needs to be filtered out and should be kept.

In the real world of cooperative action (whether business, government or military), such as in collaborative networks (Camarinha-Matos et al. 2011) or virtual organisations created by them (both being socio-technical kinds of systems of systems (SoS)), decisions are not taken in a completely centralised way. The participants of a system are systems themselves, in control of their own resources; moreover, usually each participating system has its own *system identity* as well as multiple commitments at any given time (with one of these commitments being to belong to the SoS in question). The types of strategies that need to be used in such scenarios have recently been reviewed in the extensive state of the art report by the Committee on Integrating Humans, Machines and Networks (CIHMN 2014), which calls for an interdisciplinary approach, similar e.g. to the collaborative networks research area (instead of relying on a purely computational viewpoint as a background discipline).

A SoS must be robust to cope with the situation when a participating system is not performing (e.g. becomes faulty, is destroyed, or otherwise unavailable, or due to communication channels being compromised, etc.). Successful SoS level decision-making must be framed as a *cooperative conversation* of information exchange and commitments, however with the added complexity that important systemic properties (e.g., availability) of the SoS need to be maintained, without being able to completely rely on the same property (i.e., availability) of individual participating systems.

To overcome this difficulty, the architecture of a successful SoS must be dynamically reconfigurable, so that the functional integrity of the SoS is preserved, including its mission fulfilment and its management and control. The robustness of the decision system is only achievable, if (i) the decision function is built to cope with incomplete information (at least for a limited time), (ii) the decision function can pro-actively provide guidance regarding its information needs to the contributing systems that

'observe', so as to resolve ambiguity or to replace information sources that became unavailable, (iii) the allocation of the OODA loop functions to resources is dynamic – similar to how cloud computing can achieve required capacity, availability, scalability, and other desirable systemic properties (the 'ilities') (Lehrig et al. 2015).

This self-awareness requirement for a SoS is in addition to the self-reflection requirement discussed (Sect. 3), as it requires operational (and real-time) reconfiguration, based on the need for timely and always available reliable narrative.

Although it is not possible to go into the technical details within the limits of this article, the theory that allows the two levels – situation theory and domain level theory(ies) to coexist is *channel logic* (Barwise and Seligman 2008). Mathematically, given the category of situations (that represent situation types) there exist a mapping between situation types that regulates the way complete lines of reasoning can be 'transplanted' from one situation type to another. This transplanting works as follows: when there exist a logic in a known situation type A, and the facts suggest that the situation is of a related type B, many but not all facts and inferences should also be valid in type B. As a result, if we have a known situation (of type A) with facts supporting this claim, and we only have scarce data about another situation of interest (of type B), channel logic allows us to deduce the need for data that can be used to 'fill in the details' about this second situation.

The mapping from one category to another is a morphism between categories, and can be implemented using functional programming techniques. The practical consequence is that the decision maker can use this (strictly formal) analogical reasoning to come to valid conclusions in an otherwise inaccessible domain (or if this is not possible, narrow down the need for specific data that can support a valid conclusion). This is a rudimentary explanation of the ability of the situation theoretic logic to infer that for decision making there is a need for specific unavailable data that can disambiguate the interpretation of data available at the time.

5   Conclusions and Further Work

The conceptual analytical work presented in this paper can be extended and used as the concept of a solution creating an ongoing *situation awareness* capability. All application domains (e.g. business, government, military, etc.) have typical situations and thus maintain a specific 'repertoires' of actions known to work. The knowledge of these situations can be acquired partially through education and partially through practical experience. The efficient use of these patterns depends on them being utilised fast, without too much explicit thought; in other words, efficient behaviour is typically based on the use of *tacit* skills and knowledge (irrespective of the fact that some of this knowledge would also be available in an explicit, formal form).

As pointed out, the effectiveness of the OODA loop is also dependent on its efficiency – hence, *whoever does it better and faster will win*. If, for example, one is not able to process the gathered information in a timely manner, the resulting action/s may become irrelevant, because one's adversary may have already acted and thus, changed the situation. Given the complexity of situations and the amount of data

available, using data analytics *in conjunction with situation recognition* could dramatically speed up the loop, hence increasing the chance of success.

The technology for data analytics and predictive analytics is currently subject to substantial ongoing effort in research and in industry. The authors are observing the technical side of this movement, while concentrating their effort on the promising, however difficult to implement and mathematically challenging technology that is developed on the basis of situation theory (Goranson and Cardier 2013). The main aim is to demonstrate the use of such technology to construct resilient Systems of Systems through dynamic management, command and control of a large number of cooperating participating agents.

References

Barwise, J., Seligman, J.: Information Flow: The Logic of Distributed Systems. Cambridge Tracts in Theoretical Computer Science. Cambridge University Press, Cambridge (2008)

Beer, S.: Brain of the Firm. Allan Lane Penguin Press, London (1972)

Benson, K., Rotkoff, S.: Goodbye, OODA loop: a complex world demands a different kind of decision-making. Armed Forces J. **149**(3), 26–28 (2011)

Camarinha-Matos, L.M., Pereira-Klen, A., Afsarmanesh, H. (eds.): PRO-VE 2011. IAICT, vol. 362. Springer, Heidelberg (2011). doi:10.1007/978-3-642-23330-2

CIHMN (Committee on Integrating Humans, Machines and Networks): A global review of data-to-decision technologies. Complex operational decision making in networked systems of humans and machines a multidisciplinary approach. National Academy of Sciences. National Academy Press, Washington (2014)

Doumeingts, G., Vallespir, B., Chen, D.: GRAI grid decisional modelling. In: Bernus, P., Nemes, L., Schmidt, G. (eds.) Handbook on Architectures of Information Systems, pp. 313–337. Springer, Heidelberg (1998). doi:10.1007/978-3-662-03526-9_14

Evidence Based Medicine Working Group: Evidence based medicine. A new approach to teaching the practice of medicine. JAMA **268**, 2420–2425 (1992)

Goranson, H.T., Cardier, B.: A two-sorted logic for structurally modeling systems. Progress Biophys. Mol. Biol. **113**(2013), 141–178 (2013)

Greenhalgh, T., Howick, J., Maskrey, N.: Evidence based medicine: a movement in crisis. BMJ 2014 **348**, g3725 (2014)

Hazen, B.T., Boone, C.A., Ezell, J.D., Jones-Farme, A.: Data quality for datascience, predictive analytics, and big data in supply chain management: an introduction to the problem and suggestions for research and applications. Int. J. Prod. Econ. **154**, 72–80 (2014)

HBR: From data to action (a Harvard Business Review Insight Center report). Harvard Business Review (2014)

Inmon, B.: Building the Data Warehouse. Wiley, New York (1992)

Kimball, R.: The Data Warehouse Toolkit. Wiley (1996). ISBN 978-0-471-15337-5

Lawson, H.: A Journey Through the Systems Landscape. College Publications, London (2006)

Lehrig, S., Eikerling, H., Becker, S.: Scalability, elasticity, and efficiency in cloud computing: a systematic literature review of definitions and metrics. In: Proceedings of the 11th International ACM SIGSOFT QoSA, pp. 83–92. ACM, New York (2015)

Lenders, V., Tanner, A., Blarer, A.: Gaining an edge in cyberspace with advanced situational awareness. IEEE Secur. Priv. **13**(2), 65–74 (2015)

Madden, S.: From databases to big data. IEEE Internet Comput. **16**(3), 4–6 (2012)

Osinga, F.P.B.: Science, Strategy and War: The Strategic Theory of John Boyd. Routledge, London (UK) (2006)

Pope, C.: Resisting evidence: the study of evidence-based medicine as a contemporary social movement. Health **7**, 267–282 (2003)

Santanilla, M., Zhang, D.W., Althouse, B.M., Ayers, J.W.: What can digital disease detection learn from (an external revision to) Google Flu trends? Am. J. Prev. Med. **47**(3), 341–347 (2014)

Szafranski, R.: A Theory of Information Warfare. Airpower J., **9**(1), 1–11 (1995)