BIG DATA FOR DEFENCE AND SECURITY

Neil Couch and Bill Robins

Royal United Services Institute

OCCASIONAL PAPER

About RUSI
The Royal United Services Institute (RUSI) is an independent think tank engaged in cutting edge defence and security research. A unique institution, founded in 1831 by the Duke of Wellington, RUSI embodies nearly two centuries of forward thinking, free discussion and careful reflection on defence and security matters.

For more information, please visit: www.rusi.org

About EMC
EMC is a global leader in enabling organisations in both the private and public sector to transform their operations and deliver IT as a service. Fundamental to this transformation is cloud computing. Through innovative products and services, EMC accelerates the journey to cloud computing, helping organisations store, manage, protect and analyse one of their most valuable assets – information – in a more agile, trusted and cost-efficient way.

Over the past ten years, EMC has invested more than $33 billion in R&D and acquisitions to build industry leading capabilities in Cloud, Big Data and Trust, which help our customers transform their operations.

About Cassidian
Cassidian, the security and defence division of EADS, is a worldwide leader in defence and security solutions. The company delivers advanced defence systems along the whole action chain from sensors through command and control systems to combat aircraft and unmanned air systems. In the area of security, Cassidian provides customers worldwide with border surveillance systems, cyber-security solutions and secure communications. In 2012, Cassidian – with around 23,000 employees – achieved revenues of €5.7 billion.

www.cassidian.com

Occasional Paper, September 2013

Big Data for Defence and Security

Neil Couch and Bill Robins

The views expressed in this paper are the authors' own, and do not necessarily reflect those of RUSI or any other institutions with which the authors are associated.

Comments pertaining to this report are invited and should be forwarded to:
Elizabeth Quintana, Senior Research Fellow, Air Power and Technology, Royal United Services Institute, Whitehall, London, SW1A 2ET, United Kingdom, or via e-mail to elizabethq@rusi.org

About RUSI Publications
Director of Publications: Adrian Johnson
Publications Manager: Ashlee Godwin

Paper or electronic copies of this and other reports are available by contacting publications@rusi.org

Printed in the UK by Stephen Austin and Sons Ltd.

Contents

Introduction and Summary

> Big Data holds great potential for the Defence and Security sector but [the MoD] must not fall into the trap of procuring only bespoke software solutions if it is to exploit the technology in a timely manner.
>
> Air Chief Marshal Sir Stuart Peach, Vice Chief of Defence Staff

The modern world generates a staggering volume of data and the capacity to store, broadcast and compute this information continues to grow exponentially, with one estimate suggesting that the installed capacity to store information would reach 2.5 zettabytes (2.5×10^{21} bytes) in 2012.[1] International Data Corporation research suggests that the world's digital information is doubling every two years and will increase by fifty times between 2011 and 2020.[2] The commercial sector is increasingly exploiting these vast quantities of data in a variety of ways, from sophisticated market analysis that allows precisely targeted advertising, to the real-time analysis of financial trends for investment decisions and ultimately to completely new business models.[3]

The term 'Big Data' is used to describe these very large datasets and 'Big Data analytics' to refer to the process of seeking insights by combining and examining them.[4] As noted by the Ministry of Defence (MoD):[5]

> Big Data – large pools of data that can be captured, communicated, aggregated, stored, and analysed – is now part of every sector and function of the global economy. Like other essential factors of production such as hard assets and human capital, it is increasingly the case that much of modern economic activity, innovation, and growth simply couldn't take place without data.

Against the backdrop of this huge growth in storage capacity, data collection and computation power, Big Data is arguably one of the most important global trends of the coming decade. Champions of Big Data suggest that combined with increasingly powerful and ubiquitous mobile and cloud computing and a huge increase in the spread of social-media technology, Big Data is underpinning a move to the next-generation of computing: the 'third platform', outlined in Annex A.

Over recent years, with an ever-growing reliance on network-centric operations, governments such as those of the US and the UK have allocated significantly increased budgets to improving their ability to collect intelligence, largely as a response to the demands of both the changing nature of global terrorism and military operations in Iraq and Afghanistan.

This data comes from a number of sources and platforms. It is varied and increasingly 'unstructured', with large quantities of imagery and video generated every day. A paper by the UK's Government Business Council estimates the number of unmanned aerial systems (UAS) in use to have risen from fifty a decade ago to 'thousands' now, with spending on them increasing from less than $300 million in 2002 to more than $4 billion in 2011.[6] These UAS, it is estimated, have flown more than 1 million combat hours during this period. One MQ-9 Reaper sortie collects the equivalent of up to twenty laptops' worth of data. It is therefore not surprising that much of this information can only be retrospectively analysed, rather than being fully exploited in real time. The number of intelligence analysts in the US armed forces has soared over the last decade in an attempt to manage the information deluge, leading US Air Force commanders to comment wryly that UAS are rarely 'unmanned' – also reflected in the RAF's use of the alternative label 'Remotely Piloted Air Systems'. Moreover, a recently retired senior US Army intelligence officer claimed that 95 per cent of battlefield video data is never viewed by analysts, let alone assessed.[7] Future high-end, wide-area sensors are only likely to aggravate this situation: for example, real-time ground surveillance system ARGUS, developed by the United States' Defense Advanced Research Projects Agency (DARPA), collects over 40 gigabytes of information per second.

Big Data analytics have a potentially significant role in helping to manage the data deluge and assisting analysts to focus their efforts on analysing content rather than searching for relevant information. It has already been applied in a broader context as part of the 'human-terrain mapping' exercise in Afghanistan, importing traditional databases, biometric data and imagery to extract information that would improve ISAF's understanding of the local population and key relationships within it.[8]

However, the analogues and parallels with commercial-sector initiatives are perhaps more readily identifiable outside of application to military deployments; for example, when applied to policy formulation, financial planning and management of the MoD. Financial efficiencies are welcome at any time but never more so than during a period of prolonged austerity and stiff competition for public-sector budgets. The McKinsey Global Institute estimates that Big Data represents opportunities worth around $250 billion annually to Europe's public sector (and $300 billion annually to the US health-care sector alone).[9] Big Data also has applications in cyber-security, providing network managers with the means to process millions of daily attacks and identify the more serious attacks dubbed 'advanced persistent threats'.

Whilst there are potential benefits of Big Data to operational effectiveness and financial efficiency, there are also widely held reservations regarding the temptation for central government to centralise excessively information and

control. However, other principal concerns focus on the ability to furnish and sustain the necessary skills to exploit the information, particularly when that information could save lives. Military systems will therefore need to walk the fine line between assuring that the information is adequately protected – not least in terms of individuals' rights – whilst enabling the degree of sharing essential to release the insights it contains.

The Aim

This Occasional Paper seeks to highlight to defence and security policy-makers the possibilities offered by Big Data, to warn of some of the associated risks and implications of using it, and to recommend some of the work that should now be set in motion.

Key Recommendations

The consequences of ignoring Big Data and associated 'third-platform' technologies (cloud computing, mobile devices and social media) in the defence and security sector could be profound, including the loss of life and operational failure. In addition, the growing legal obligation regarding human rights of personnel and standards of care when on operations could apply in the future as much to the information provided to commanders and servicemen as it increasingly does to physical equipment and training. Beyond the clear moral imperative, therefore, the reputational and financial impact in an increasingly litigious society should not be ignored.

The authors therefore recommend that the MoD should:

- Define a Big Data work package (possibly to be managed by DSTL) as part of technology innovation studies sponsored by the Chief Technology Officer (CTO). This should consider a broad range of candidate technologies and techniques from the commercial sector that may have application to the areas of defence outlined above
- Consider persuading two three-star officers (one in MoD Centre and one in Joint Forces Command – JFC) to act as Big Data champions or senior responsible owners for Big Data exploitation within the department and JFC respectively. The authors' suggestions in this regard are the Chief of Defence Intelligence (CDI) and Director General, Strategy in the JFC
- Consult widely on the responses to likely legal and ethical challenges that such an approach might require, particularly from a department of state
- Select two functional areas (one from the MoD Centre and one from JFC or a Front Line Command) that might benefit from pilot programmes or concept demonstrators, acting both to support the MoD as a learning organisation and as proofs of concept for Big Data techniques

- Use the proof of concepts to create a more detailed business case for change
- In the pilot areas, the MoD should:
  - Assess training and educational needs for the functional areas expected to use Big Data, covering senior management and subject-matter experts (data analysts)
  - Initial assessment of the moral and legal issues to be addressed in any Big Data policy-development activity
  - Clarify the role of industry in support of developing the capability, including potentially providing skilled data analysts to the reserve force element.

Notes and References

1. Chris Yiu, *The Big Data Opportunity: Making Government Faster, Smarter and More Personal* (London: Policy Exchange, 2012), p. 6.

2. IDC, 'Digital Universe Study: Extracting Value from Chaos', June 2011.

3. For example, the so-called 'long tail' model that characterises Amazon, eBay and the iTunes store, amongst others.

4. Yiu, *The Big Data Opportunity*, p. 6.

5. McKinsey Global Institute, 'Big Data: The Next Frontier for Innovation, Competition, and Productivity', May 2011, Preface.

6. William Matthews, 'Data Surge and Automated Analysis: The Latest ISR Challenge', Industry Insights, Government Business Council, January 2012.

7. Private conversation with the author.

8. NetReveal Analyser, provided by Detica, has been in service since 2010.

9. McKinsey Global Institute, 'Big Data', p. 7.

I. What is Big Data?

> Big Data is not a technology, but rather a phenomenon resulting from the vast amount of raw information generated across society, and collected by commercial and government organisations. Put simply, it is the management and exploitation of large or complex data sets.
>
> Ministry of Defence[1]

Big Data generally refers to datasets that are not susceptible to analysis by the relational database tools, statistical analysis tools and visualisation aids that have become familiar over the past twenty years since the start of the rapid increase in digitised sensor data. Instead, it requires 'massively parallel software running on tens, hundreds, or even thousands of servers in some (currently extreme) cases'. (Annex A provides an overview of these technologies and techniques.) For the purposes of this paper, Big Data is defined as 'datasets that are too awkward to work with using traditional, hands-on database management tools' and Big Data analytics as 'the process of examining and interrogating Big Data assets to derive insights of value for decision making'.[2]

That said, the label Big Data can be unhelpful for two reasons. First, it can too easily be dismissed as the latest buzzword from an information technology (IT) industry constantly seeking competitive advantage and new revenues. Some informed sources consider the currently high profile of Big Data to be fleeting. For example, the technology research and analysis firm Gartner suggests that it will have disappeared from the lexicon in three years or so. Not because it is of no consequence, but rather the opposite: because by then – they judge – it will be a prerequisite for successful business. Whilst its use may not guarantee success, its absence is likely to lead to failure.

The second reason the label is unhelpful is its misleading focus on data size, ignoring other equally important characteristics. These include:

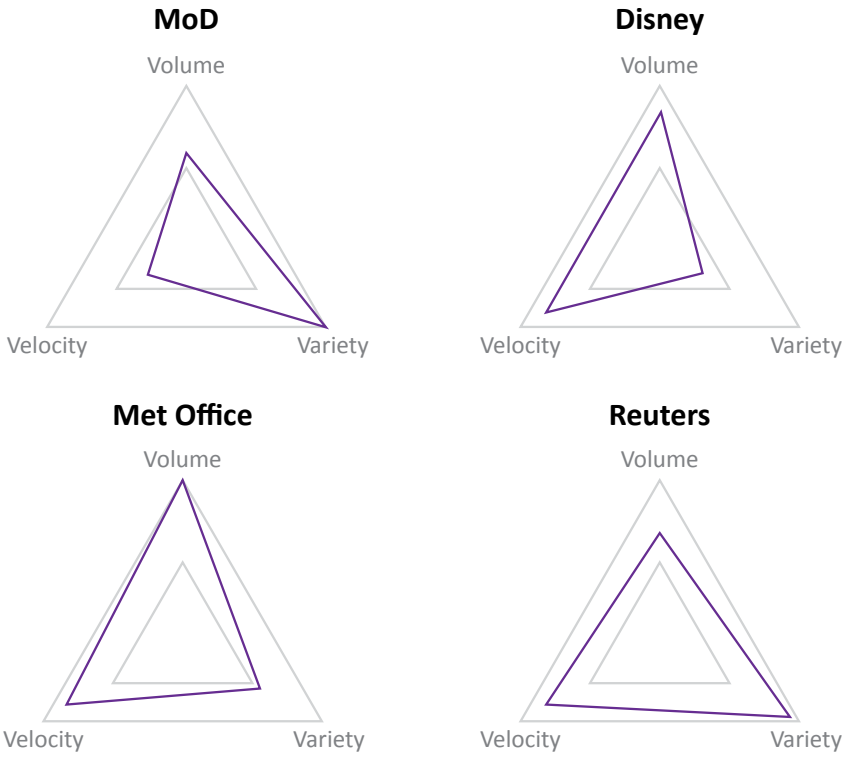
- Volume. The most obvious aspect: mass quantities of data that organisations are trying to harness to improve decision-making across the enterprise
- Variety. Variety is about managing the complexity of multiple data types, including structured and unstructured data. This aspect of Big Data is often less widely understood
- Velocity. The speed at which data is disseminated and also the speed at which it changes or is refreshed (often as a result of cyclical actions). The dynamic nature of the data demands large amounts of computing power, specific design features and fast networking. However, used optimally, it can support fast and accurate decision-making

- Variable Veracity. The level of reliability associated with certain types of data. A view held in the Big Data industry is that the precision of individual data items is becoming less relevant because of the normalising effect of analysing data at these vastly different orders of magnitude.

However, Big Data is not defined by the absolute terms of these characteristics, but by the relationship between them and the enterprise's ability to manage and exploit them. It is neither a new type of data, nor a new technology, nor a data-management methodology – although it may stimulate and promote any of these. It is a relationship between data and the organisational, procedural and cultural factors that make up the enterprise. It is primarily about the ways in which data is managed.

Figure 1 gives a relative comparison of the dominant characteristics of data stored by various entities. While the MoD's volume is regarded as relatively low compared to the other organisations listed, its variety is high.

Figure 1: The Dominant Characteristics of Data Stored by the MoD, Disney, the Met Office and Reuters.



Source: Network Technical Authority, 'Big Data: A NTA Technical Vision, Rev 0.4', DE&S Information Systems and Services, 1 August 2013, courtesy of Niteworks.

The MoD's Network Technical Authority (NTA), which is responsible for ensuring the technical coherence of the MoD's networks and for informing its future architecture and design, has proposed the following draft technical vision for Big Data in the MoD:[3]

> Data from within and outside the MoD is brought together and exploited to its optimum effect in order to establish clarity, increase efficiency and gain previously hidden intelligence. Users can confidently access and analyse timely and trusted information from the most appropriate sources, through a variety of interfaces, regardless of their geographic location, organisational unit, or underpinning technical architecture.

Notes and References

1. MoD C4ISR Joint User, 'The Challenges and Opportunities of Big Data', 4 December 12. Please note, this is an internal MoD paper that is not publicly available.

2. Chris Yiu, *The Big Data Opportunity: Making Government Faster, Smarter and More Personal* (London: Policy Exchange, 2012), p. 10.

3. Network Technical Authority, 'Big Data: A NTA Technical Vision, Rev, 0.4', DE&S Information Systems and Services, 1 August 2013, p. 5.

II. Why Should the Defence and Security Sector be Interested?

The estimated volume of data currently collected by British UAS platforms in Afghanistan equates to between 1 and 5 per cent of the Google index, causing some commentators to observe that 'the MoD doesn't have Big Data. Google and Amazon do'. This overlooks the fact that there are large variations in the structures of the MoD's data, which is stored in a multitude of different sources and repositories that do not comply with an overarching design. Nonetheless, the defence and security sector functions increasingly in a digital world for both operational activity and corporate business. This presents opportunities to accelerate and disaggregate activity, but also requires greater focus on managing and assuring large volumes of data securely across multiple, widely dispersed, fixed and deployed locations.

The impact of failing to exploit defence- and security-related data could be far higher than that faced by the private sector in terms of loss of profit or market share. From an operational perspective, the intelligence element is increasingly required to trawl through vast datasets to identify specific signatures in order to trigger specific responses. The timeframes for this are often measured in seconds, not hours or days, requiring constant and rapid analysis. For intelligence analysts, the emphasis is switching increasingly to a more immediate response to fleeting signatures and away from the time-consuming monitoring of routine 'pattern of life' in order to find those signatures. The expectations of commanders continue to grow as more digital sensors and collectors enter service, and as new signatures are identified that need to be 'washed' against multiple databases. These expectations will not be met without corresponding improvements in tools and techniques to support the search and analysis of data.

The volume and complexity of the raw data being collected threatens to overwhelm existing analytical systems and processes. Data are frequently lodged in different forms and structures, stored in isolated silos bolstered by deliberate measures to prevent cross-fertilisation in order to conform to security policies, many of which were designed for a non-digital (and slower-moving) era. This makes it difficult or even impossible to develop a coherent view of complex issues at the level that is needed.

In its capacity as a department of state (albeit not a citizen-facing one), the MoD exhibits many of the shortcomings highlighted in Sir Philip Green's 2010 review of government efficiency,[1] which exposed the lack of good-quality data on how the government spends its money as a key factor preventing effective decision-making. This has been a persistent problem within the MoD, which suffers from a lack of adequate management-information

on which to formulate decisions and with which to take evidence-based arguments into the Whitehall resource debate. The ability to find the critical piece of information among a torrent of real-time data is crucial, whilst also ensuring that collection, storage and processing conform to the laws and regulations that protect the rights of the citizen.

In the UK at least, shrinking budgets have led to similarly shrinking armed forces. Future operations are likely to be conducted with fewer forces. There is likely to be an ever-greater dependence on battle-winning intelligence in order to underpin the command and control that the Vice Chief of the Defence Staff recently reflected had refined the application of force to a fine art.[2] This will drive even greater reliance on bandwidth and the network that increasingly sits at the heart of the business of defence. Modern, capable infrastructure will therefore be key. New, more agile commercial relationships will also be needed in order to reduce the cost of operational maintenance and permit more investment in information-enablers matched to the expenditure in platforms and sensors.

Intelligence and Operations

For many intelligence experts, automated analysis technology is the top intelligence, surveillance and reconnaissance (ISR) priority. The UK Maritime Intelligence Fusion Centre recently highlighted the imbalance between investment in collectors and in the tools to support its analysis, rendering analysts incapable of taking into account all available sources when performing their assessment. According to senior MoD officials, the UK 'has reached an inflection point in data deluge. We are now in danger of data asphyxiation and decision paralysis'.[3]

Algorithms are now needed that can 'discover' useful intelligence in the surfeit of collected data to allow analysts 'to perform real analysis rather than exhausting themselves culling raw data'. The director of DARPA, a US agency responsible for developing new military technology, illustrated the case for automated analysis to a Congress hearing by comparing the intelligence collected for counter-insurgency operations with that for air defence against strategic bombing. She suggested that detecting dismounted fighters in Afghanistan requires collecting and processing about 100,000 times more data than it takes to detect a strategic bomber. 'One of two things must happen,' she said. 'Either we must give up the target set [of dismounted fighters], or we must deal with the data volume. Obviously, we do not want to give up the target set. The only choice before us, therefore, is to develop capabilities that dramatically improve our ability to manage and use the data'.[4] The US is engaged in a variety of research projects, some in conjunction with UK universities. One such classified programme proposes developing algorithms to analyse hundreds of thousands of open-source

documents each hour and compare them to billions of historical events. The aim is to produce predictive tools capable of anticipating specific incidents.

This chimes with a coherent, well-developed set of observations provided to the authors by the UK Maritime Intelligence Fusion Centre that the breadth of sources available to the UK intelligence staff far exceeds the capacity of analysts employing traditional methods of reviewing all available information before presenting a reasoned assessment. There is a need for tools to reduce the volume of material that analysts must assess, allowing them to focus on those likely to be the most fruitful. For this exact reason, the Australian Defence Imagery and Geospatial Organisation has adopted Big Data techniques to help analysts sift through satellite imagery and monitor changes in areas of interest around the world.[5]

Similarly in the field of cyber-security, where network managers can be dealing with millions of attacks every day, Big Data analytics are being applied to spot advanced persistent threats – such as socially engineered attacks designed to steal corporate intellectual property or government information – above the ever-growing background noise of everyday nuisance or opportunistic attacks. Most hackers have a modus operandi, which once identified can be used to predict the form of future attacks and put appropriate defensive measures in place.

The skill of the future analyst is likely, therefore, to concentrate more on configuring sophisticated search tools to which subject-matter experts can then apply their experience, intuition and human judgement. The Officer Commanding of the UK Maritime Intelligence Fusion Centre particularly stressed the need for tools that can perform semantic and sentiment analysis, able to indicate the most likely next step or outcome. However, funding for analytical tools appears to be lacking in the MoD core programme.

Predictive Tools

Work has already begun on examining the utility of social media and Big Data analytics as tools to assist in identifying opportunities for upstream prevention of instability, particularly where the UK might identify the early-warning signs and act before the need arises for more costly intervention and stabilisation operations. Separately, the NATO Allied Rapid Reaction Corps HQ is working with academics to find methods of applying Big Data analytics to open-source intelligence (OSINT) as a means of predicting likely targets for radicalisation (both communities and, increasingly, specific individuals) as well as to 'collapse the distance between policy, planning and operations' by capturing large datasets and seeking to identify patterns. This aims to substitute or provide support for the human analytical function in order to speed up the process by, in effect, applying a synthetic intuition based on pattern recognition.[6]

Human analysis, however, will remain essential in order to add judgement and insight to the findings of Big Data analytics. For example, automated analysis of OSINT will also require the involvement of subject-matter experts and other technical sources in order to reduce the possibility of malicious manipulation.

Corporate and Department of State

As mentioned previously, McKinsey identifies 'strong evidence' that Big Data potentially underpins substantial improvements in efficiency in the public sector, an analysis borne out by that of the Policy Exchange think tank. McKinsey identifies five categories of Big Data levers for the public sector:

- Creating transparency to make data more accessible, allow agencies to share data and minimise the repeat entry of data
- Enabling experimentation through which to discover requirements, expose variability and improve performance. Big Data analytics can reveal wide variations in performance within agencies that are not visible in the highly aggregated analyses carried out by conventional means. Big Data also offers the opportunity for predictive analysis. By examining the relationships embedded in large datasets it is possible to build a new generation of models describing likely future evolution. These can be combined with scenario planning to develop predictions for how systems will respond to policy choices or other decisions
- Replacing or supporting human decision-making with automated algorithms. Such techniques can often reveal anomalies contained in large datasets when compared to historical data and are applicable to counter-terrorism and other police, security and defence intelligence scenarios. Algorithms 'crawl' through data sources identifying inconsistencies, errors and fraud
- Segmenting population groups to allow targeted and tailored action plans
- Innovating new business models, products and services.

These categories have been developed around tax and labour agencies, but McKinsey asserts that 'their use can be just as relevant to other parts of the public sector'. McKinsey's models are founded on exploiting the mass of data existing in a citizen-facing department. Their application to the defence sector will not necessarily be straightforward in all cases. However, those regarding transparency, experimentation and decision support do potentially fit the MoD model. Some examples are described in Annex C.

Logistics

Perhaps counterintuitively, operational logistics are not seen as an immediate priority for the application of Big Data tools. However, with regards to more routine maintenance and support, commercial companies are increasingly

contracted for capability and availability. It is therefore likely that there will be a growing demand to exchange large volumes of monitoring and reporting data for predictive maintenance.

Senior logistics officers express concern about the data fidelity and utility of their many legacy information sets and applications. These support the base inventory and contribute to substantial discrepancies between inventory management (a failure to track effectively all the equipment and parts that MoD owns) and financial management in the defence sector. This view is supported by the National Audit Office in its 2012 report on the Defence Equipment Plan.[7] It is not yet clear how Big Data analytics could improve this without a basic improvement in data hygiene, but the concept is worth examining.

Complex Programme Management

Of all the defence programmes, the most complex is overall financial programme planning and management. The MoD has been criticised by various parliamentary committees on numerous occasions for shortcomings in this regard. The inability to identify how and where money is spent, or how and where it would best be spent, and to provide compelling, evidence-based cases, weakens the department's position whenever it is obliged to negotiate individual proposals or budgetary settlements with the Treasury.

Cabinet Officer Minister Francis Maude in a 2012 speech to the Policy Exchange set out the current government's determination to exploit tools such as Big Data analytics in order to make efficiencies and to reduce the huge sums currently lost or wasted as a result of inadequate data management, availability or analysis.[8] Sir Philip Green's 2010 review of government efficiency exposed the poor quality of data on 'where and how government spends its money', and the MoD is as guilty of this as other departments.[9] Green reported that data was 'inconsistent and not available', making it impossible for government to buy efficiently. Exactly the same observations have been made by interviewees in the course of this study regarding defence procurement and the Defence Equipment and Support (DE&S) organisation.

Notes and References

1. Philip Green, 'Efficiency Review by Sir Philip Green: Key Findings and Recommendations', Cabinet Office, October 2010.

2. Air Chief Marshal Sir Stuart Peach, Vice Chief of Defence Staff, Ministry of Defence, presentation to RUSI Chief of the Air Staff's Air Power Conference, London, 17 July 2013.

3. MoD C4ISR Joint User, 'The Challenges and Opportunities of Big Data'.

4. William Matthews, 'Data Surge and Automated Analysis: The Latest ISR Challenge', Industry Insights, Government Business Council, January 2012, p. 3.

5. Rohan Pearce, 'How Defence Tackles the Challenge of Big Data with Machine Learning', *Computerworld*, 13 March 2013.

6. Collapsing the policy, planning and execution timeline will also require users to look critically at their own organisations and processes in order to exploit this capability. This is consistent with remarks by the Vice Chief of the Defence Staff at a recent RUSI conference regarding the potential need to collapse future command-and-control structures.

7. The department focused on increasing the robustness of its procurement costings, and has yet to apply the same level of challenge and review to the support costs element, although it plans to do so for the Equipment Plan 2013.

8. Francis Maude, 'The Big Data Opportunity', speech to the Policy Exchange, 3 July 2012, <https://www.gov.uk/government/speeches/francis-maude-speech-to-policy-exchange-the-big-data-opportunity>, accessed 5 September 2013.

9. For example, see evidence given by Chief of Defence Materiel Bernard Gray to the House of Commons Public Accounts Committee regarding the Major Projects Report 2012 and MoD Equipment Plan, 4 February 2013, <http://www.publications.parliament.uk/pa/cm201213/cmselect/cmpubac/uc868-i/uc86801.htm>, accessed 5 September 2013.

III. Examples and Case Studies

Data Sharing for Entity Resolution

The Policy Exchange paper 'The Big Data Opportunity' includes the following example of data sharing for identity resolution, which has potential in terms of biometric and other identity-tracking applications for security purposes:[1]

> The UK does not have a single national identity database so it is not possible to retrieve reference information relating to an individual citizen from a single authoritative source. With previous generations of technology, this often meant there was no alternative to storing multiple copies of the same data (or simply coping without some of the data altogether). Modern technology, however, can enable fragments of related information to be matched and linked together quickly and non-persistently. This can be used to streamline transactions – reducing the scope for errors and avoiding asking people to provide the same information multiple times.
>
> For example, when you apply for a UK driving licence online, the Driver and Vehicle Licensing Agency (DVLA) requires a photograph and signature for your new licence. If you have a UK passport then the DVLA will try to capture these electronically from the information already held by the Identity and Passport Service (IPS). This is often held up as an example of how simple changes can deliver practical improvements for end-users.

Persons of interest can often be found across multiple databases. Sharing such information can be complicated, particularly when names and locations are in foreign countries that do not use Latin script, as the translation of these names into English can vary. As a very simple example, during NATO Operation *Unified Protector* in Libya in 2011, British journalists reported on action from 'Misrata' whereas US journalists referred to the same city as 'Misurata'. Big Data analytics can help to resolve such 'entities', thereby allowing analysts to make links that would otherwise have remained obscured.

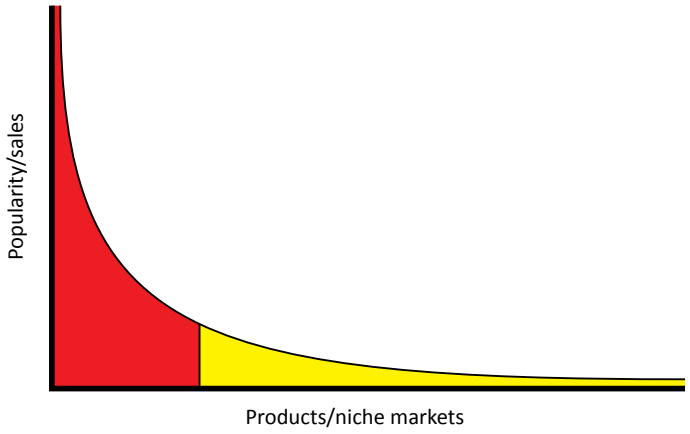
In another case, a data analytics house was invited to examine records of insurance claims from a broad set of UK insurance companies and advise on the level of fraudulent claims. By correlating a sufficiently large dataset and applying entity-resolution techniques, it revealed a pattern indicating that organised crime was a greater problem than individual cases of fraudulent claims. A similar system exists in the US to detect fraudulent health-insurance claims.

Personalising and Focusing

There are many examples in the commercial world of the power of Big Data for personalisation: for example, Amazon generates tailored recommendations based on its knowledge of the customer and of the histories and buying

patterns of similar customers. This technique has allowed the creation of 'long tail' business models, which have inverted previous mass-marketing models that sought to sell the largest possible quantity of mass-produced items to the largest possible group of customers. The long-tail model, shown in Figure 2, instead seeks to sell the broadest possible product range (assuming that the distribution channel and the market are large enough). In this way, Amazon has been able to stock and sell a far broader range of books than traditional bookshops.[2]

Figure 2: The Amazon Long-Tail Business Model.



This is a beneficial harnessing of technology and analytical techniques to reduce the burden on the customer of searching for what they want. Others see this as an intrusion into individual privacy, using sophisticated techniques to 'piece together' a broader and more detailed picture of personal behaviour than the individual would willingly grant. Google, for example, 'personalises' or tailors its search function according to geographic location, past behaviour and, in some cases, online contacts. On first appearance, it would appear to have utility in intelligence applications because it not only tracks mass behaviour (for example, movements like the Arab Spring) but can concurrently identify what the minority population is doing.

Notes and References

1. Chris Yiu, *The Big Data Opportunity: Making Government Faster, Smarter and More Personal* (London: Policy Exchange, 2012), p. 13.

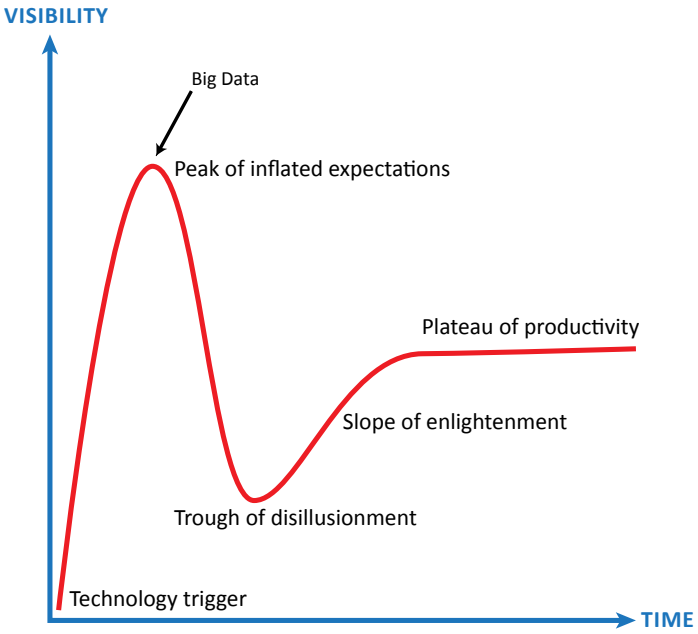
2. Based on products that are in low demand or that have a low sales volume collectively making up a market share that rivals or exceeds the relatively few current bestsellers and blockbusters, if the store or distribution channel is large enough.

IV. Issues and Risks

It is unlikely that Big Data amounts to a silver bullet and it should be approached with some caution. It offers opportunities to maximise return on shrinking investment, but natural enthusiasm must be moderated by a wariness of government tendency at times to grasp at novel IT trends as a panacea to deal with more deeply rooted dysfunction.

Gartner sees the life of Big Data on the 'Hype Cycle' (see Figure 3) as potentially short-lived, judging that within a few years it will become a 'given' – an assumption in terms of doing business in a highly competitive global market. It is possible that the same will be true in certain aspects of defence, particularly if the UK intends to remain a 'preferred partner' of the US. Big Data competence and capability could well become an entry-level requirement for future operations in much the same way as the ability to join 'the network' is a current prerequisite. There are, however, a number of associated risks and challenges that will need to be addressed.

Figure 3: The Gartner Hype Cycle.



Data Aggregation: Confidentiality and Legality

The use of Big Data analytics will call for a clear understanding of the moral and legal underpinning of the actions of subject-matter experts, commanders and policy-makers. The power of these analytical techniques offers capabilities previously undreamt of by totalitarian regimes for spying on their citizens,

but which are now readily available. The possibility of abuse or misuse by any state, including liberal democracies, should not be ignored. Indeed, the aggregation of Big Data across multiple sources to form a virtual repository has already given rise to questions of security (in terms of confidentiality), legality and ethics.

Big Data potentially allows individuals to be identified by linking previously unconnected datasets. Removing anonymity in this way can be a powerful intelligence tool, but one that raises both moral and legal issues. For example, trumpeting the decision not to build a national ID database while continuing to assemble a virtual capability by other means might be a tempting course for officials to take; but it potentially raises challenging questions of legality and of governments' right to access the data and their ability to protect it. Tensions between the privacy of citizens and the need for intelligence have been highlighted by a number of recent high-profile cases, including Julian Assange and Wikileaks; PRISM and Edward Snowden; and the reported collection of information about Verizon customers by the National Security Agency (NSA) in the US. Other less sensational incidents further highlight the challenge to government of making effective use of available data for improved outcomes and reduced cost. For example, it was reported that a recent cross-Whitehall initiative intended to 'join up' multiple data sources in order to allow government to follow citizens from childhood upon school entry, through school and university, and ultimately into the workplace, partly in order to track the payment of student loans, was apparently blocked by data-protection legislation.[1] Even in the commercial world, commentary on *Forbes* and elsewhere suggests that consumers prefer to have control over the information that companies hold and how it is used, arguing that there is a fine line between being an intelligent service provider and stalking.[2]

In the majority of cases in the UK, these issues will not be the direct concern of the MoD but of other departments. However, they could raise significant issues relating to 'operations amongst the people', depending on the law and the jurisdiction that applies to the information, when collected, analysed and stored. Cloud hosting is not an essential prerequisite of Big Data but the two are often associated. If Big Data sources are hosted in the 'cloud', there will need to be careful consideration of the legal jurisdiction that applies and the extent to which it permits assurance and confidentiality, including the protection of citizens' rights.

Understanding the Answer: Visualisation Tools

Sophisticated visualisation tools will be needed to aid understanding of the results of the analysis. Most, if not all, of these possible analytical tools will not realise their full potential to generate insights and actionable intelligence without also

applying recent advances in visualisation tools that mean it is now feasible to bring granular and up-to-date evidence to bear on leadership decision-making.

A recent article in *The Economist* highlights that 'A new generation of statisticians and designers – often the same person – [is emerging] working on computer technologies and visual techniques that will depict data at scales and in forms previously unimaginable… Three dimensional network diagrams show ratios and relationships that were impossible to depict before'.[3] 'Translating data into images', it argues, 'allows people to spot patterns, anomalies, proportions and relationships', which could potentially reduce the level of blind trust commanders and analysts need to place in algorithms that they do not understand and cannot satisfactorily interrogate.

Another article suggests the *Guardian* newspaper's use of data visualisation as the 'high-water mark' of data journalism.[4] Using the classified American information released by Wikileaks in 2010, it took thousands of American field reports and extracted information on IEDs, then mapped the location and timing of each incident to produce an interactive infographic that described a key aspect of the conflict by data alone.[5] There are a number of impressive examples of the power of open data sources presented using effective visualisation tools on the *Guardian* website, in its Datablog and World Data Store.[6]

Assurance and Confidentiality

A number of concerns have been raised concerning assurance of the outcomes of Big Data. (Here, assurance includes the availability, integrity, protection of confidentiality (protective security), authentication and non-repudiation of data.) There is some anxiety over the ability to understand what is 'going on under the bonnet'. This is most acute when the analysis is being used to support safety-critical decisions: for example, targeting. It will therefore be important that the MoD develops sufficient understanding of its overall processes and data flows to ensure that a human being is kept in the loop where it is critical to the application of judgement and to meet legislative or regulatory requirements. This suggests that the development of Big Data as part of information exploitation must include intimate involvement of the war-fighting practitioner and cannot be left to the information specialist alone.

The information specialist will, however, need to take the lead in constructing a robust but agile Information Assurance regimen to protect the data. This will need to address the dilemma created by the fact that the benefits of Big Data are derived in part from the automated cross-referencing and correlation of multiple datasets, which will require permissive cross-boundary policies, unlike those that exist currently. The aggregation of data has long been treated as a reason to raise related security classifications and protection requirements, but the scale of aggregation in Big Data solutions is

likely to make this an acute problem deserving of careful attention. A form of federated identity control will be essential to confine access to appropriate users, which will require automated, real-time 'understanding' of the increase in sensitivity of the data as it is merged, cross-referenced or in some other way combined with other information.

Joint, Multinational and Multi-Agency Operations

Recent coalition campaigns in Iraq and Afghanistan have accelerated enormously the move to operations that exploit and rely on the network. The pace of progress that has been achieved would have seemed unlikely when the concepts of Network Centric Warfare (NCW) and Network Enabled Capability (NEC) were first gaining traction at the start of the last decade. Led by the US, nations including the UK have realised the power of the network and, at the same time, their reliance on it to provide timely, mission-critical information.

However, national and departmental policies (not only security policies, but also data-management, data-definition and data-storage policies) can still present obstacles to the smooth passage of information across organisational interfaces and boundaries. The role of the information specialist in providing adequate information assurance will be even more challenging in multinational and multi-agency operations and more challenging still to plan for information sharing on contingency operations alongside unspecified coalition partners. Identifying appropriate planning assumptions for the implementation of Big Data analytics on such contingency operations will require leadership and fine judgements surrounding the balance of risk in order to achieve optimised operational effectiveness. However, it must remain clear that the challenge rests ultimately with the war-fighter or end-user of the information, and they should also therefore be responsible for identifying the appropriate assurance and sharing policy.

Duty of Care and Legislation

There are risks in failing to adopt Big Data (or equivalent methods) for handling the increasing quantity of information now available to government departments. First and foremost must be the danger of operational failure through the inability to turn existing information (acquired at great cost and sometimes risk) into actionable intelligence. In legal terms, and in terms of external confidence in the performance of the department, the handling and exploitation of information and intelligence in this context must be viewed in the same light as the equivalent for physical weapon systems and equipment. If any reinforcement of this were needed, it can be found in recent verdicts concerning the application of civil law and regulation in the battlespace, such as the application of UK Health and Safety on deployed operations and the judgement of the UK Supreme Court that the provision contained in the European Convention on Human Rights regarding the right to life also applies in operational zones.

These outcomes increasingly provide a legislative framework for compelling a high standard with regard to the duty of care. Although legal experts have suggested that such concerns should not apply to judgements made by commanders on the battlefield, they do place a proper responsibility on the state to ensure that proper training and equipment is made available. It may, however, be no more than a matter of time before the same case is made in regard to ensuring the provision of potentially life-saving intelligence, particularly if that intelligence can be shown to have existed within the 'system' or to have been available through other accessible sources at the time.

At least one case is already being brought against the MoD in which it is claimed that the death of a soldier in Iraq in 2003 was avoidable, had the intelligence known at the time to the relevant commanders been used effectively. In this context, it is of note that a recently retired senior US military intelligence officer assesses that 95 per cent of battlefield video data is never viewed by analysts, let alone assessed.[7] Lack of adequate tools to 'mine' data may soon no longer be an acceptable explanation for failure. Similar tools are also required to allow the large volumes of diverse archived data recovered from operational theatres to be managed and searched in order to meet legal and other obligations, such as responding to coroners' inquests, parliamentary enquiries and Freedom of Information requests, as well as supporting historical research and the lessons-learned processes.[8]

Skills

As with many other information and process change initiatives, skills will be an essential prerequisite to success. In this way, recruiting and retaining sufficient analytical talent will be a critical factor and not one that the public sector will necessarily find easy, given the likely competition from the private sector, which will normally be better placed to match scarcity of such qualified people with attractive remuneration.

The same report by the McKinsey Global Institute referred to earlier highlights the potential shortfall of skilled personnel in the US, while it also concludes that the shortage of deep talent in this field will be a global phenomenon.[9] Its analysis suggests that by 2018 the US will require 4 million 'data-savvy managers and analysts' suited to a Big Data world. These would not be the deep experts but those with enough 'conceptual knowledge and quantitative skills to be able to frame and interpret analyses in an effective way'; in other words, functional subject-matter experts. McKinsey estimates that on current trends the US will have a shortfall of 1.5 million people of this type. Its corresponding analysis of the shortfall of deep analytical talent – the data scientists of the future – is even more stark, suggesting a need to increase supply by nearly two-thirds from 300,000 to about 490,000 in 2018. Furthermore, it is arguable that many of these subject-matter experts will operate in functional areas benefiting from Big Data and will need to

be able to apply their judgement and knowledge of, for example, ethnicity, intelligence and policy formulation to make assessments of the information and analysis presented to them.

If McKinsey is correct in its analysis that this trend will be reflected worldwide, and if Big Data analytics does indeed become 'the entry standard' for military operations as well as for commercial enterprises, then militaries should begin now to address the implications for recruiting, developing and retaining the necessary skills. Given the inevitable competition with the commercial sector for a limited pool of talent, the MoD should consider potential opportunities for collaboration with industry, including rooting a substantial part of the capability in the reserve element where professional practitioners would be best-placed to keep their skills in line with the latest emerging techniques.

Notes and References

1. It is perhaps worth noting how difficult it would be to express this sentence without the words 'track' or 'follow' that alarm liberal commentators suspicious of government's alleged Big Brother ambitions.

2. Jerry Michalski, 'Big Data and the Stalker Economy', *Forbes*, 10 March 2012.

3. *The Economist*, 'Inforgraphics: Winds of Change', 6 July 2013.

4. *Ibid*.

5. *Guardian*, 'Wikileaks Data Journalism: How We Handled the Data', Datablog.

6. Simon Rogers, 'All of Our Data Journalism in One Spreadsheet', *Guardian* Datablog, 31 January 2011.

7. Private conversation with the author.

8. Staff members at PJHQ explained in conversation with the author the challenges they were facing in storing and searching archived data from operations in Iraq, which they expected to increase when data was retrieved from Afghanistan and archived. They speculated that Big Data techniques could assist in addressing these.

9. McKinsey Global Institute, 'Big Data: The Next Frontier for Innovation, Competition, and Productivity', May 2011, p. 7.

V. Creating a Big Data Capability in Defence

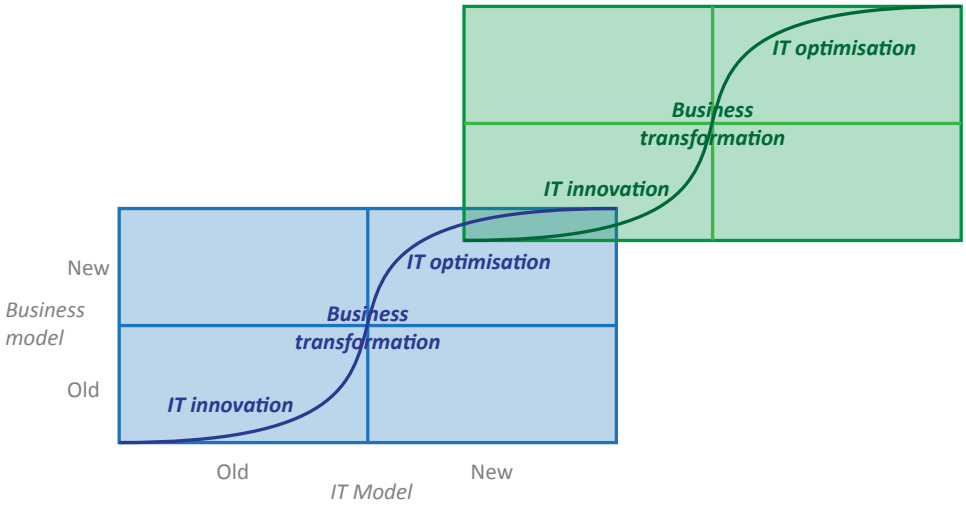
Implementation Strategy

The approach to the promotion of organisational competence in Big Data outlined below has been developed by EMC, a commercial player in this capability.[1]

EMC believes that the typical company which uses existing second-platform technology (fixed infrastructure and relational database management systems) – represented by the lower-left box in Figure 4 – could only improve by a limited amount even with business transformation processes in place. Adopting third-platform technologies (mobile devices, cloud computing, Big Data and social media) would move the company forwards, but such a large transformation process would inevitably incur an exorbitant cost and could fall foul of an organisation's conservative instincts.

As such, EMC's approach recommends building a new organisation from scratch, based on the principles of third platforms. The new organisation is shown in the upper-right box in Figure 4, and is kept separate from the day-to-day business of the operation. The process of transforming innovation into optimised capability takes place in each box but the risks involved in the upper-right box are higher, hence the need for protection of the fledgling capability. As confidence and competence grow, operations should increasingly move to this upper box. The authors of this paper believe that an evolutionary transformation of the type envisioned by EMC may be worth consideration by the MoD.

Figure 4: EMC's Evolutionary Implementation Model.



Source: EMC.

A Possible Approach
An organisation in the initial stages of developing competence in Big Data might consider:

- Appointing a 'champion' (maybe a senior responsible owner) to take the capability forward and protect the developing capability from departmental conservatism
- Developing an outline business case and allocating initial resources accordingly
- Testing the concept with a pilot programme or concept demonstrator, possibly based in a functional area that already has some awareness of, and commitment to, Big Data
- Assessing the skill base required and moving to develop a core of the necessary skills
- Developing relationships with relevant industry players and with relevant universities.

The Current Situation in Defence
As regards the UK MoD, it appears to the authors that:

- With the exception of a few pockets of awareness, there is as yet little understanding at any level of the benefits and risks involved in exploiting Big Data
- There is no champion for Big Data capability and no case for Big Data exploitation has yet been made
- There is some potential for pilot programmes. A few functional areas may welcome a pilot programme to exploit Big Data if the resources were available
- The skill sets for Big Data exploitation (analytics, presentational tools, advanced data management) are largely lacking.

However, some work is already in progress. For example, a well-structured approach to developing the technical aspects of Big Data is proposed by the NTA in DE&S Information Systems and Services (ISS).[2] It recommends that the move towards technical competence in Big Data should take place within the mainstream Information Management/Information Exploitation programme, under the aegis of the Defence Information Strategy and the leadership of the Chief Information Officer (CIO).

Approach Proposed for Defence
The authors propose that at the technical level, the MoD Director Capability, JFC should consider working with the MoD CIO and Chief Technology Officer (CTO) on the approach suggested by the NTA paper. As part of this technical-level implementation, it is also recommended that a work package be undertaken as a set of technology innovation studies under the CTO to

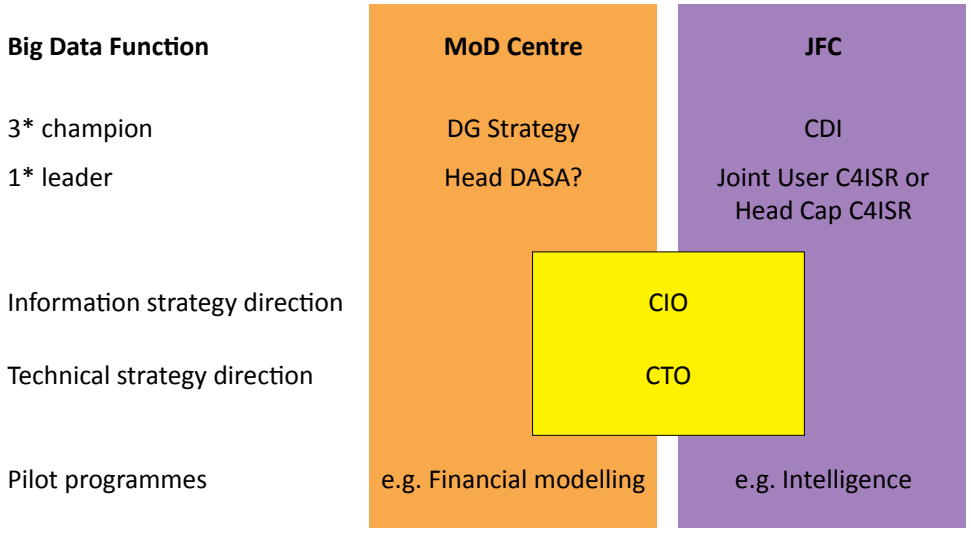
look at a broad range of candidate technologies and techniques from the commercial sector with possible application to the areas of defence that this paper highlights. DSTL may be the best focus for managing this work.

But more is needed: the authors of this paper believe that although the NTA work proposed would create an effective 'engine room' for Big Data, the management and operational levels must create their own drives for exploitation if the department as a whole and its management, operational, intelligence and logistics capability is to benefit fully. It is therefore proposed that the Director Capability should:

- Use material from this initial paper to create wider Big Data awareness within the relevant areas of the department
- Create an outline business case to identify start-up resources and engage the relevant people
- Consider persuading two three-star officers (one in MoD Centre and one in JFC) to act as Big Data champions or senior responsible owners for Big Data exploitation within the department and Joint Forces Command, respectively. This role may be best performed by the Director General Strategy and the Chief of Defence Intelligence
- Select two functional areas (one from the MoD Centre and one from JFC or a Front Line Command) which might benefit from pilot programmes, acting both to support the MoD as a learning organisation and as proofs of concept for Big Data techniques.

The proposed approach is shown diagrammatically in Figure 5.

Figure 5: Possible MoD Implementation Model.



Outline Business Case

The outline business case might be assisted using material from this paper. This could most effectively be done by Director Capability Joint Forces Command staff assisted by CIO/CTO staff and by the staff of the departmental champion proposed by this paper, the Director Strategy. It might usefully cover:

- The rationale for adoption of Big Data as a capability
- An outline of anticipated benefits and costs
- The penalties of not exploiting Big Data
- The proposal for leads in the MoD Centre and JFC
- The proposal for pilot schemes, based on these criteria. Each pilot should be:
  - Driven by a functional area that sees benefits in doing it
  - Supported by an informal grouping of MoD technology staff selected on their ability to 'get the point of Big Data', as well as by competent industry players, DSTL and selected academic foci[3]
  - Underpinned by clear project management for pilot programmes: objectives, resources, timescales and governance
  - Supported by an assessment of training and educational needs, covering senior management, subject-matter experts for the functional areas expected to use Big Data[4] and data analysts
- An initial assessment of the moral and legal issues to be addressed in any Big Data policy-development activity
- The role of industry in support of developing the capability, including potentially providing skilled data analysts to the reserve force element.

Such an approach might help to prove the capability and avoid institutional resistance to 'yet another' change programme.

Notes and References

1. The authors are grateful to Chris Roche, CTO of EMEA, EMC for the use of this diagram.

2. Network Technical Authority, 'Big Data: A NTA Technical Vision, Rev 0.4', DE&S Information Systems and Services, 1 August 2013.

3. The Institute of Security Studies at Exeter University has volunteered to support this work.

4. As an example of the sort of mid-career education that may be needed here, it should be noted that an Exeter University Islamic Studies Specialist has had to teach himself Bayesian Analysis to become competent in using Big Data. Similar personal development may be needed in MoD functional areas.

Conclusions and Recommendations

This paper has described the recent large and continuing expansion in the volumes of data collected and stored by both the commercial and public sectors. It has highlighted the shift occurring in the commercial sector towards a new type of relationship between organisations, their data and information management, and their processes. The move to a Big Data approach allows them to deal not only with the high volume of data but also with its volatility, velocity and variability to extract ever-greater insights, business intelligence and, ultimately, more reliable predictions of the likely evolution of events.

These developments in the commercial sector have parallels in governmental and military arenas. The intelligence world already collects more raw data than it can analyse, with perhaps as much as 95 per cent of imagery never being viewed by analysts. Similarly, government policy-makers collect more data than they can handle yet simultaneously often lack robust, data-based evidence on which to plan and manage the business of the Ministry of Defence.

The consequences of not dealing effectively with these challenges in the defence and security sector are potentially profound – extending beyond those associated with market competition that drives the commercial sector – and include loss of life and operational failure. In addition, the growing legal obligation regarding human rights and standards of care when on operations could apply in the future as much to the information provided to commanders and servicemen as it increasingly does to physical equipment and training. Beyond the clear moral imperative in this regard, the reputational and financial impact in an increasingly litigious society should not be ignored.

The MoD does not hold petabytes and zettabytes of data like the Internet giants, but its data is largely either unstructured or in diverse structures; it is often dynamic; in many cases its provenance and reliability cannot be confirmed; and operational demands often dictate a need for real-time analysis to create actionable intelligence. Existing approaches to data and information management are unable to cope. Austerity measures exacerbate the challenge by reducing the available resources whilst there are continued demands to do more with less.

Government and industry face similar pressures to increase performance and speed of reaction, driving a need to collapse the distance between policy, planning and action. Emulating the intuitive understanding of human analysts with automated processes designed to search for patterns in huge, complex datasets could help. The most immediate benefits are likely to be in areas such as departmental and financial planning, programme management,

and in key JFC enablers such as operational logistics and intelligence, as well as operational planning as a whole. The experience of the commercial sector suggests that a Big Data approach may offer:

- Faster transition from collection to analysis, decision and action
- Greater confidence in the analysis and the conclusion
- More reliable insight and foresight, resulting in more reliable predicted outcomes potentially both at the level of large groups and that of the individual
- Better use (and re-use) of the data that has been collected, sometimes at considerable cost in blood and treasure.

The potential benefits of Big Data analytics are significant but need to be approached with an awareness of the associated risks and challenges. Security management across boundaries must balance the need to protect information with the potential benefits of sharing. The majority view amongst those spoken to during the course of this short study was that currently this balance does not exist and the barriers to sharing are too high. Big Data will not solve this. Indeed, the aggregation of datasets is likely to increase the challenge. There is a need to develop a culture, supported by corresponding policies, that incentivises data sharing for the greater good as well as more sophisticated risk-assessment models and techniques. These may need to be reflected in agreements with principal partners and may be particularly challenging in an era of contingency operations when prospective partners may not be known prior to a specific operation. Access control and authentication algorithms that dynamically limit user permissions to correspond with changes to the sensitivity of merged datasets will be necessary, as will an awareness of where information is held, the jurisdiction that applies, and the legal obligations and freedoms that implies. Individuals must be protected against any temptation on the part of government and associated agencies to misuse or abuse the personal information gained through these techniques, particularly perhaps those which identify individuals in what otherwise would be an anonymous dataset.

Competence in these areas will quite possibly become a fundamental requirement for military forces in the future, as is also expected of successful enterprises in the commercial sector. They may also become entry-level expectations for participation alongside the UK's high-end partners in a similar way as is now the ability to participate fully in network-enabled operations. In the US, the Pentagon and security agencies are already beginning to move forward quickly with developing Big Data capabilities. The Australian government is also committed to developing its data-analytics capability.[1] The UK defence and security sector should guard against slipping behind key partners in its ability to exploit a new opportunity that will potentially offset some of the impact of reduced force levels and resources elsewhere.

A collaborative approach with the commercial sector that allows defence to benefit from cutting-edge skills, possibly maintained in the reserve forces, merits investigation.

Whilst this does not require a major investment programme, it will not succeed without improved information mining and harvesting tools and a better balance between investment in collectors, sensors and platforms as well as visualisation tools to aid understanding of the results produced. There is a view that this currently remains out of kilter; on the evidence gathered for this paper, this appears to be justified.

Progress will demand senior leadership in both the operational and corporate areas of the defence business, by both practitioners and end-users supported by the CIO and CTO. It appears to the authors that this leadership needs to be at the three-star level and, as such, the Director Capability, JFC might approach the Director General, Strategy and the Chief of Defence Intelligence to take on these roles, though the authors stress that consultation at this level has not been undertaken during this short study.

Nevertheless, under the combined aegis of these two senior leaders, the Director Capability, JFC and the CIO, should begin a programme to generate wider awareness of Big Data within the relevant areas of the department and create a business case covering the themes set out in Chapter IV, beginning by identifying the start-up resources to take pursue the capability offered by Big Data analytics.

Implementation should follow an evolutionary path within the wider Information Management/Information Exploitation Strategy and be coherent with the Defence ICT Strategy. However, the authors are inclined to accept the conclusions and advice drawn from commercial experience that this can most effectively be stimulated by sponsoring experiments and capability concept demonstrators (pilots) initially isolated from the mainstream of current business in order to prevent them from being stifled early on by the prioritisation of 'business as usual'. It is recommended that a work package (possibly managed by DSTL) be investigated as part of technology innovation studies sponsored by the CTO. This should consider a broad range of candidate technologies and techniques from the commercial sector with possible application to the areas of defence that highlighted in this paper. The authors recommend selecting two functional areas (one from the MoD Centre and one from the JFC or a Front Line Command) that might benefit from pilot programmes, acting as learning environments and proofs of concept. Academia stands ready to support the work.

As the Vice Chief of the Defence Staff has outlined in his quotation given at the start of this paper, it would be a mistake to rely solely on bespoke Big

Data capabilities for defence and security. Rather, the MoD should build on the huge investment in this area being made by the commercial sector and in doing so ensure that the MoD is well-positioned to track and exploit further commercial technological developments as and when they occur.

Notes and References

1. Australian Government Department of Finance and Deregulation, 'Big Data Strategy: Issues Paper', March 2013.

Annex A: The Scale of the Challenge and Examples of the Toolset

Depending on the sensor suite used, a medium-altitude, long-endurance remotely piloted aerial system such as the UK's MQ-9 Reaper or a Hermes 450 can collect the equivalent of up to twenty laptops' worth of raw data in a single sortie.[1] There are at least five of the former and nine of the latter in active operation. As such, the weekly data output of these UAS can be as much as 1,950 laptops' worth of data. The UK currently has a further five Reaper and fifty-five Watchkeeper UAS on order,[2] which would generate up to 4.5 petabytes – equivalent to over 9,000 laptops' worth – of data, which is roughly the size of all the web pages viewed globally over a five-day period. This is Big Data by any definition and does not include intelligence collected through surveillance sorties by aircraft, such as the Tornado RAPTOR and Sentinel ISTAR, sensors in the field, signals intercepts, or e-mail and Internet traffic.

Storage Lags behind Processing

Whilst global capacity for automated data processing doubles roughly every fourteen months, capacity for storage lags far behind, only doubling every thirty-four months. The ability to store data is lagging way behind the ability to process it.[3] The key challenges within ISR are:

- Storing the initial data streaming in from UAS and sorties
- Accommodating 'unstructured' data
- Scaling-up human analytics in proportion to these increased data volumes.

As with other areas of IT, the first challenge in ISR is simply to ingest and store this data, particularly when the majority of it is 'unstructured' (in the form of images, videos, documents, e-mails, and so on) and not sitting in a conventional relational database. Despite processing power becoming cheaper and increasingly powerful, a key challenge is automating the analysis and moving analytics from being retrospective to predictive, thereby automating the identification of risks and anomalies and visualising the analysis in new ways.[4] Human analysts simply do not have the time, nor are there enough of them to wade through the ISR equivalent of five days of global Internet web page traffic each week.

As things stand, then, the norm is post hoc, retrospective analysis of 'stale' data, as opposed to predictive warnings based on active, near real-time data. Emerging technologies and techniques aim to offer this analysis, promising predictive tools based on anomalies and patterns, as opposed to reports dissecting events that have already taken place.

Examples of Technology and Techniques

When defining a Big Data technology platform that underpins the deployment of analytical and predictive information services to military operational personnel, it is worth considering this platform as the convergence of three technology layers or 'fabrics': the application fabric; the data fabric; and the cloud fabric. IDC (an American market research firm) refer to this future platform as the third-generation platform upon which millions of data-intensive applications for billions of users will be built and delivered through mobile devices and apps.

The first-generation platform was the mainframe. The second-generation platform, which has been in existence since the early eighties, was built on client-server technology and an underlying relational database management system, which stores data in rows and columns – intuitively familiar to users of Excel spreadsheet software. The focus of this new, third platform is the productivity of the data scientists and application developers who have to build and deploy the time-critical analytical solutions.

In the new world of real-time analytics, the major design principles that pervade all three layers are speed, collaboration, choice and value. The instantiation of these four principles within each layer of a third-generation technology platform is paramount if organisations are to move from proprietary, expensive, inflexible data models of the second-generation business-intelligence platforms era, which have struggled to cope with the new data workloads and use-cases required by service personnel today. Second-generation platforms tended towards vendor lock-in, which ultimately leads to less choice and greater cost through expensive licensing fees.

For example, in the context of the military, a new future application requirement might be the rapid deployment of an intelligence application that allows a front-line soldier to identify, in real time, who he is talking to, what their links are to other organisations, who they have been in touch with recently and more. The big issue for him will be absorbing the information in time to use it.

Looking first at the data fabric, from a technology perspective, it will need to deploy technologies that underpin the high-capacity, real-time ingestion and querying of data. Today, that is a technology combination of three things.

First, massive parallel-processing (MPP) databases, which allow data scientists and developers to take advantage of large clusters of increasingly powerful, increasingly inexpensive commodity servers, storage and Ethernet switches is required. These architectures allow the analytic models that data scientists build to be processed inside the database over vast petabyte

datasets, thereby reducing the need to move large datasets around and increase the performance a hundred times over non-MPP databases. This technology also facilitates the rapid ingestion of data. This speed affords the data scientist the ability to iterate many models and hypotheses – what is termed 'fail fast' capability.

Second, in-memory database capability is also required. Sometimes the speed of the data is such that in-memory database (IMDB) technology is required to facilitate the capture and immediate processing of data. IMDBs primarily rely on main memory for computer data storage and are faster than disk-optimised databases since the internal optimisation algorithms are simpler and execute fewer CPU instructions. When MPP databases and in-memory databases are coupled together, the concept in which fast data meets Big Data becomes a reality. This is particularly important when the real-time location of an entity is desired or, for instance, data is flooding in from millions of sensors in the battlefield and needs to be correlated with vast quantities of historic data.

Third, Hadoop processing and storage technology is required.[5] As the variety of data types increases – especially machine-generated data – traditionally structured relational-database management systems become a limiting factor. These also require data to adhere to a strict pre-defined schema prior to loading which means they must be intensely scrubbed and formatted before loading into the database (known as 'Schema on Write'). In many instances, much of the data is discarded at the ingestion point to adhere to the schema. Hadoop allows the user to 'dump' all data types, videos, web-logs, machine-generated and flat files, for example, into the database and at a later point build the relevant schema ('Schema on Read') through which to process the data. Hadoop has two components: a MPP engine called MapReduce and a file system known as HDFS. Hadoop is an emerging discipline and not all use-cases are suitable to its construct today. However, as this element of the data fabric is improved, observers of the data fabric predict a convergence to an in-memory Hadoop-based platform.

In reality, the ability to build a data fabric that combines and integrates all of the above technologies seamlessly and not just to focus on one area is a requirement of a new third-generation platform. Importantly, for choice, the data fabric must also be capable of processing and interacting with many different analytical models and statistical software packages. Data scientists tend to use many different tools when analysing and visualising data. In addition to being productive in dealing with data, the data fabric must also offer a productive environment for the data scientist. As data-science skills are in demand, ensuring any platform makes the data-science team as productive as possible is important. Productivity in data science comes from collaboration and the ability to create data workspaces within the data

fabric that facilitate the import, exploration, and visualisation of data and communication between scientists.

Once an insight has been uncovered, the military may wish to take action on such insight. In today's world, the visualisation and presentation of data and insight is more and more via a multitude of devices including smartphones, tablets and direct to the sensor. If the data fabric is optimised from the data and data-science perspective, it would be counterintuitive then to have to wait months for an application to be built (coded) or re-architected to deliver so that it could then be used, or to build an application that is not fit for purpose because it has taken so long to commission. In the third generation, an application fabric would have to support the rapid build and deployment of analytic applications.

Rapid application programming techniques such as pair-driven programming and test-driven programming are best deployed within open-source application frameworks and frameworks that are integrated with the data fabric. These ensure that developer time is totally focused on the bespoke code that is needed and the programming techniques ensure that their coding time is optimised.

Open-source software is particularly important in the third-generation platform. Open-source software can be defined as any computer software, generally developed as a public collaboration, whose source code is made freely available. There are many reasons why an organisation may wish to develop on top of open-source software – cost being one of them. As important, if not more so, is that when businesses turn to open-source software, they free themselves from the severe vendor lock-in that can afflict users of proprietary software packages.

All applications run on physical hardware such as servers and storage. In the past, the development of applications has had to take into account the hardware it will be run on. This clearly poses challenges in environments with multiple hardware versions installed (not unusual in today's data centres). This can result in multiple versions of the same code being produced. The principle behind the cloud fabric is to ensure that code is portable across all infrastructure and not dependent on any single infrastructure. As the majority of infrastructure today is deployed in cloud environments, the cloud fabric needs to be cloud independent. This independence offers greater choice and lower cost when deploying applications.

Further Information
A further overview of analytical techniques and technologies for Big Data can be found in the McKinsey Global Institute report 'Big Data: The Next Frontier for Innovation, Competition and Productivity', <http://www.

mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation>.

Notes and References

1. Figures provided by EMC. These are based on 10TB of data being collected per sortie and assume that one laptop can store 500GB in data.

2. The Watchkeeper UAS will replace the Hermes 450 Tactical UAS.

3. Martin Hilbert and Priscila Lopez, 'The World's Technological Capacity to Store, Communicate, and Compute Information', *Science* (Vol. 332, No. 6025, 1 April 2011).

4. See an analysis of global population data for 200 countries over 200 years depicted in just four minutes: <http://www.youtube.com/watch?v=jbkSRLYSojo>, accessed 5 September 2013.

5. Apache Hadoop is an open-source software framework that supports data-intensive distributed applications.

Annex B: An Example of Implementing Big Data

In 2012, one of the UK's central government ministries (not the MoD) reviewed its general IT strategy for 2013–14, including examining Big Data and analytics. The department had a substantial Analytics Directorate team of experienced statisticians, and a variety of business management information systems (MIS) and business intelligence (BI) teams. However, it was not clear whether Big Data was even relevant. There was a view that it appeared to be typical 'technology hype' from vendors, possibly of importance but probably overstated.

The department CTO and strategy office worked with a commercial partner over four months in early 2013 to explore the potential benefits. Questions included:

- Is Big Data real?
- Does it offer the department any potential for saving cost or improving operational excellence?
- How do Big Data technologies differ from what they currently use for MIS and BI?
- If it is relevant, what will a Maturity Assessment show?
- Is cloud computing also valuable in the Big Data space?
- What specific use cases exist within the department?
- If Big Data has potential, what specific use cases should they focus on as an exemplar or flagship?

The first phase of the work has recently concluded, revealing that:

- Use-cases presented a circular problem. Internal users were needed to help find the best potential use-cases, but found it hard to identify these without some examples as a framework. By definition, there were none
- The key priorities became:
    - Identification of genuine use-cases
    - The feasibility of using a private cloud off-premises
    - Skills gap among the analysts and end-users
- Existing MIS and BI teams were helpful but had a small-scale view of data. Typically, they work with highly processed, summarised and abstracted datasets. For them, Big Data was a structured database of 30–60 terabytes
- Conversely, the analytics teams understood Big Data's potential for first, handling much larger datasets, second, handling unstructured data and, third, consolidating data into a structured database later – when appropriate, rather than when they simply run out of space
- The narrowest skills gap existed within the analyst community

- In the end, the team identified eleven potential use-cases, of which the top three were in the same operational area. These were recommended as the focus for future work
- The use-cases listed by management initially turned out to be red herrings. The best use-cases only emerged as a result of this discovery exercise
- A service provided via a private cloud off-premises appeared to have potential, provided the requirement to gain accreditation for the information protection (Impact Level 3 as a minimum).

The next phase of work is under consideration. It will potentially focus on the three top use-cases and define work packages that will yield the most immediate returns in terms of both cost saving and operational excellence.

Open-Source Intelligence for Prediction

Suggested examples of the use of open-source intelligence on the Internet include:

- An 80 per cent correlation between search intensity for crisis-related terms and yields on government bonds in distressed economies
- The number of searches made by a population for 'flu' as a reliable indicator of the probability of members becoming infected
- Fast-moving consumer-goods companies use search intensity to predict levels of sales and plan supply chains
- Google search trends for the second week of the month as the best predictor of car sales for the month.

About the Authors

Neil Couch

Neil Couch served for over thirty years in the British Army with the Royal Corps of Signals, commanding operational signals units at every level, including squadron command with the UN in Namibia and regimental command in Bosnia and Kosovo. After command of 1st (United Kingdom) Signal Brigade in support of the Headquarters Allied Rapid Reaction Corps in Germany, his last two tours were spent in the Ministry of Defence, London where he was responsible at the strategic level for communications and information systems support to current operations; for conceptual development of Network Enabled Capability; and for planning and prioritising the equipment capability programme for the full range of future Defence ICS. He is now an independent consultant to the defence and telecommunications sectors.

A graduate of the UK Higher Command and Staff Course and of the Royal College of Defence Studies, he holds a Master's degree in International Security and Strategy from King's College London where he specialised in the threat to the Mexican state from corruption and organised crime. His study was recently published in the Defence Studies journal. He is a Fellow of the British Computer Society and a Chartered IT Professional.

Bill Robins

Major General (Rtd) Bill Robins CB OBE led tactical communications units in parachute, mechanised, armoured and infantry formations of the British Army.

During a spell in Whitehall, he led the requirements team for a strategic protected Whitehall bunker, directed Command and Information Systems for the Army and as Director General of Information and Communications Services, attempted to unify UK Defence information services across MoD, Cabinet Office and other Government systems and into the theatres of operations.

On leaving the army in 1998, he was appointed Chairman of the Royal Signals Institution and worked as a consultant for the Treasury before joining Marconi and then BAE Systems where his last appointment was as Director of Advanced Concepts for the newly formed C4ISR Group. He left BAE Systems in 2003 and now runs his own consultancy specialising in Defence and Security Information Management. He chairs the Board for the Defence Fixed Telecommunications Service PFI programme for both partners, MoD and BT. He is an Associate Fellow at RUSI, a Freeman of the Worshipful Company of Information Technologists and a Visiting Professor at Cranfield University, assigned to the UK Defence Academy.

RUSI Membership

RUSI membership packages provide privileged networking opportunities and benefits tailored to meet the needs of both individuals and large organisations.

Individual Memberships
Individual memberships are suitable for those individuals who wish to join RUSI's growing network of policy-makers and practitioners. Benefits include regular updates from RUSI, including invitations to members' lectures and seminars, subscription to the *RUSI Journal* and *RUSI Defence Systems*. This package also offers members access to our renowned Library of Military History.

Corporate membership
RUSI's Corporate Level Membership packages, offering discounts to all RUSI conferences, are open to all organisations concerned with defence and security matters, and can be tailored to meet the business interests of both public and private sectors.

Concessions
Discounted student and young persons rates are available for those under the age of 35. Concessions are also available for those over the age of 65. We also offer Online Membership to those wishing access to RUSI's content of analysis and commentary.