Big Data and Deep Learning for Understanding DoD Data

Ying Zhao, Ph.D., Naval Postgraduate School
Douglas J. MacKinnon, Ph.D., Naval Postgraduate School
Shelley P. Gallup, Ph.D., Naval Postgraduate School

Abstract. Today, Big Data infrastructure and analytics intervene with traditional data sciences. We are compelled to ask - What is new? In this article, the authors provide a pragmatic context for how Big Data infrastructure and analytics are related to traditional data sciences including statistical analysis, numerical analysis, machine learning, data mining, pattern recognition and data fusion. The authors also discuss use cases in various categories that demonstrate empirical practicality for understanding and applying Big DoD Data.

1. Why Big Data Now?

"Project pursuit," a method that seeks useful lower dimensional data, was researched extensively in the 80s with a research funding level of $10,000. When "machine learning" emerged in the earlier 90s, the funding for "projection pursuit learning network" [1], for instance, grew to $100,000. It grew to $1,000,000 for "data mining" in the late 90s. The core method remained the same, yet the data size was bigger. Today, Big Data science intervenes with traditional data sciences. We are compelled to ask - What is new? Let us examine the current breakthroughs:

- Big rise in data: data creation is remarkable for its volume, velocity, and variety. "Volume" considers the rise of new data creation platforms of multimedia, social media, mobile devices, the Internet of Things (IOT) and new sensors. "Velocity" considers these new platforms capturing millions of events per second and in real-time. "Variety" considers captured data not only just numbers but also unstructured text, images, audios, videos, geospatial data, and 3D data. Big Data are omnipresent and ubiquitous. In 2012, the Obama administration announced 84 Big Data initiatives across six departments [2].
- Big rise in needs: It is critical for business to transform data into smart data, or actionable knowledge. For example, researchers need to use Big Data to discover new drugs. Marketers need to use social networks, mobile, geo-location, and sensor data to reach more customers. The United States National Security Agency (NSA) needs to process the exabytes (1018) of data collected over the internet in the Utah Data Center [3].
- Big rise in technologies: Traditional data sciences including statistics, numerical analysis, machine learning, data mining, business intelligence, and artificial intelligence are evolved into Big Data analytics. The US Federal Government owns six of the ten most powerful supercomputers in the world [4].

These technologies can be overwhelmingly complex, requiring diversified and extensive expertise.

2. Practicality

2.1 Tools

Big Data is near impossible to process with conventional technologies, requiring instead massively parallel software on thousands of servers. The current technologies are dominated by systems that provide 1) safe storage, 2) parallel/operational processing, and 3) deep analytics.

As part of open-sourced Apache Hadoop ecosystem, Hadoop Distributed File System (HDFS) provides distributed and fault-tolerant data storage. Beehive and Pig are "SQL-like" tools for conventional database queries on a HDFS. NoSQL systems include document and graph databases in a "cloud" such as Amazon and Cloudera. Operational systems for messaging, banking, advertising and mobile devices can utilize Apache Storm to handle day-to-day transactions in real-time, or with no- or low-latency of response.

Map/Reduce is an analytic programming paradigm for Big Data. It consists of two tasks: 1) the "Map" task, where an input dataset is converted into key/value pairs; and 2) the "Reduce" task, where outputs of the "Map" task are combined to a reduced key-value pairs. Apache Spark [5] could replace Map/Reduce for its speed and in-memory computation.

2.2 Challenges

As the data size gets bigger, the statistical significance for an analysis is often guaranteed due purely to the size. This positive impact of the data size can be a great advantage. However, other challenges rise. For example, traditional data sciences used in small- or moderate-sized, analysis typically require tight coupling of the computations of the "Map" and "Reduce" steps. Such an algorithm often executes in a single machine or job and reads all the data at once. How can these algorithms be modified so they can be executed in parallel in thousands of clusters? If the data is processed in parallel and parsed into subsets, how to leverage the art and science of fusing the results as phrased in the "Reduce" step?

An oddity to be further explored in Section 3 is that "data fusion" has been successfully performed in many DoD applications whereas commercially-tempted innovations (i.e., Thinking Machines and Cray Computers) were not successful [7].

2.3 Commercial Trends

Predictive analytics is to turn Big Data into smart data, for example, accurately forecasting high-value targets such as high-value customers, events, and social media sentiment. The topic has been thoroughly studied in supervised learning. Some algorithms are implemented using the Map/Reduce paradigm [8]. 95% of Big Data is unstructured. Text analysis methods (e.g., categorization, summarization and topic discovery) are being adapted to Big Text.

Social network analysis, product cross-selling, recommendation engines, event diffusion, and graph search require graph analyses leveraging massively parallel processors. For instance, viral path predictions are used for predicting how useful events, e.g. new ideas, videos or diseases, become proliferated to a large population or “go viral.” Graph algorithms can process petabytes of data and are considered as the core drivers of Big Data analytics. Spark, Titan and Neo4j are used for Big Graph.
One important trend is Deep Learning including unsupervised machine learning techniques (e.g., neural networks) for recognizing objects of interest from Big Data [9], for instance, sparse coding [10] and self-taught learning [11]. The self-taught learning [12] approximates the input for unlabeled objects as a succinct, higher-level feature representation of sparse linear combination of the bases. It uses the Expectation and Maximization (EM) method to iteratively learn coefficients and bases [13]. Deep Learning links machine vision and text analysis smartly. For example, text analysis Latent Dirichlet Analysis (LDA) is a sparse coding where a bag of words used as the sparsely coded features for text [10]. Our methods Lexical Link Analysis (LLA), System-Self-Awareness (SSA), and Collaborative Learning Agents (CLA) can be viewed as unsupervised learning or Deep Learning for pattern recognition, anomaly detection, and data fusion.

3. DoD Big Data Applications

Data sources for DoD applications including disparate, multi-sourced real-time sensors, and archival sources are of extremely high rates and large volumes. In DoD collaboration environments, the needs for information sharing and agility as well as strict security across all domains makes the matter more complex. While commercial applications such as massive marketing may require identifying information with popular and repeatable patterns, emerging and anomalous information are more useful for DoD applications (e.g., intelligence analysis and resource management). Deep learning regarding pattern recognition, anomaly detection, and data fusion can be even more useful. The US Navy has now begun to take initiatives to move Big Data into the battlefield [14].

In the past, at the Distributed Information Systems and Experimentation (DISE) research group at the Naval Postgraduate School (NPS), we have applied Big Data sciences to understand DoD data. In particular, Lexical Link Analysis (LLA) has been used to analyze unstructured and structured data for pattern recognition, anomaly detection, and data fusion. It uses the theory of System Self-Awareness (SSA) to identify high-value information in the data that can be used to guide future decision processes in a data-driven or unsupervised learning fashion. It is implemented via a smart infrastructure named "system and method for knowledge pattern search from networked agents (US patent 8,903,756)" also known as Collaborative Learning Agents (CLA), licensed from Quantum Intelligence, Inc. [15].

In the following sections, we first describe our approaches of LLA, SSA and CLA briefly and then categorize some DoD applications. We discuss four use cases in these categories. Some use cases were described in more detail in related publications [15-16, 25-30].

3.1 LLA, SSA and CLA

In LLA, a complex system is expressed in specific vocabularies or lexicons to characterize its features, attributes or its surrounding environment. LLA uses bi-gram word pairs as the features to form word networks. Figure 1 depicts LLA with word pairs as groups or themes. Figure 2 shows a detail of a theme in Figure 1.

Figure 1. Themes Discovered in Colored Groups

Figure 2. A Detailed View of a Theme in Figure 1
The separation of the three types is based on SSA and implemented using CLA. Figure 5(a) shows a CLA as a computer program used to separate and extract patterns and anomalies from multiple data sources. A single agent installed in a single computer node is capable of ingesting and analyzing data sources locally. Multiple agents can work collaboratively in a network and fuse multiple data sources as shown in Figure 5(b).

We define System Self-Awareness (SSA) as the ability for an agent to estimate its global importance by optimizing its total value considering its relations to other agents (authorities and patterns) and its own expertise (anomalies) learned from the local data. SSA is implemented as a fusion mechanism to optimize the overall value \(R(t, j)\) using a recursion as shown in Figure 6.

![Figure 5(a). A Single Collaborative Learning Agent: Patterns are graded from medium to relevant correlations.](image1)

![Figure 5(b). Agent Collaboration](image2)

![Figure 3. An Example of Emerging Theme](image3)

![Figure 4. An Example of Anomalous Theme](image4)

![Figure 5(a). A Single Collaborative Learning Agent: Patterns are graded from medium to relevant correlations.](image5)

![Figure 6. Recursion to Compute the Overall Value of a System R(t, j)](image6)
3.2 How Can the Methodology Be Used for Future Decision Processes

We show in the following four real examples (i.e., use cases in Section 3.2.1 to 3.2.3) that the Big Data and Deep Learning methodology in Section 3.1 can be used for future decision processes by

- Processing more data in parallel
- Automating data fusion
- Learning associations and correlations from diversified data sources which may not be standard data
- Performing pattern recognition and anomaly detection

3.2.1 Data Fusion, Optimization of Distributed Resources

DoD Big Data (e.g. sensors) are collected in local compartments and specific to domains. DoD resources are distributed with the strict security requirements [19], fault-tolerance, and agility. These data need to be combined for applications. Data fusion is the process of combining information from a number of different sources to provide a robust and complete description of an environment or process of interest. Distributed and parallel processing is required but is not sufficient for data fusion where analytic algorithms that can combine the results from distributed systems are critically required. Data fusion finds application in many military systems especially when sensor data were collected and must be combined, fused, and distilled to obtain information of appropriate quality and integrity on which future decisions can be made. For many military data fusion scenarios [20], data fusion is often divided into a hierarchy of four processes. Level 1 and 2 fusion is generally concerned with processing raw data using numerical fusion methods such as probability theory or Kalman filtering. Level 3 and 4 fusion is thus concerned with the extraction of high-level knowledge from low level fusions, the incorporation of human judgment and the formulation of decisions and actions.

To understand Big Data architecture and analytics in DoD applications, we need first understand the existing decentralized and distributed data fusion architectures [20].

- A decentralized data fusion system consists of a network of sensor nodes, each with its own processing facility, which together do not require any central fusion facility. In such a system, fusion occurs locally at each node ensures that the system is scalable as there are no limits imposed by centralized computational bottlenecks. Such a system is also made survivable or fault-tolerant. The decentralized data fusion algorithms are implicitly limited in requiring full communication, e.g. a fully connected sensing network or as a broadcast system.

- In a distributed data fusion system as shown in Figure 7 requires a central processor; however, each sensor also has its own local processor which can extract useful information from the raw sensor data prior to communication. The degree to which local processing occurs at a sensor site varies substantially from simple validation and data compression up to the full construction of tracks or interpretation of information locally.

In a use case entitled "Big Data Architecture and Analytics (BDAA) for Common Tactical Air Picture (CTAP)" the NPS team showed that the data generated by intelligence, surveillance, and reconnaissance (ISR) sensors has become overwhelming and the Navy now needs to apply new architectures and analytics to improve its CTAP. More specifically, accurate, relevant, and timely

![Figure 7: From [20]](image)

**Combat Identification (CID) enables the warfighter to locate and identify critical targets.**

The NPS team applied LLA, SSA, and CLA jointly with Hadoop, Map/Reduce, and Deep Learning to 1) improve track correlation, continuity, fidelity and latency reductions, 2) discover and learn the patterns in historical data and correlate them with real-time data to detect anomaly; 3) improve real-time targeting recommendations and guide future decision making; 4) optimize the warfare resource management.

In this context, the NPS team cannot use either distributed or decentralized data fusion, instead a recursive data fusion methodology leveraging LLA, SSA, and CLA can be employed as follows:

- An agent \( j \) represents a sensor, operates on its own like a decentralized data fusion, however it does not communicate with all other sensors but only with the ones that are its peers. A peer list can be specified by the agent.

- An agent \( j \) includes a learning engine CLA that collects, analyzes from its domain specific data knowledge base \( b(t_{j}) \), for examples, \( b(t_{j}) \) may represent the statistics for bi-gram feature pairs (word pairs) computed from LLA.

- An agent \( j \) also includes a fusion engine SSA with two algorithms SSA1 and SSA2 that can be customized externally. SSA1 integrates the local knowledge base \( b(t_{j}) \) to the total knowledge base \( B(t_{j}) \) that can be passed along to its peers and used globally in the recursion in Figure 6. SSA2 assesses the total value of the agent \( j \) by separating the total knowledge base into the categories of patterns, emerging and anomalous.

In a use case entitled "Big Data Architecture and Analytics (BDAA) for Common Tactical Air Picture (CTAP)" the NPS team showed that the data generated by intelligence, surveillance, and reconnaissance (ISR) sensors has become overwhelming and the Navy now needs to apply new architectures and analytics to improve its CTAP. More specifically, accurate, relevant, and timely
• The total value \( V(t,j) \) is used in the global sorting and ranking of relevant information.

In this recursive data fusion, the knowledge bases and total values are completely data-driven and automatically discovered from the data. Each agent has the exact same code of LLA, SSA, and CLA, yet has its own data apart from other agents. This agent work has the advantages of both decentralized and distributed data fusion. It performs learning and fusion simultaneously and in parallel. Meanwhile, it categorizes the patterns and anomalous information. In many use cases investigated, the NPS team found the discovered patterns are often correlated with authoritative information, while anomalies are correlated with new and interesting information requiring further investigation. For example, sorted and ranked information according to authority and anomalousness can be used to improve and automate future decision processes of CID with higher precision and lower latency that optimize the use of long-range weapons, aid in fratricide reduction, enhance battlefield situational awareness, and reduce exposure of U.S. Forces to enemy fire.

3.2.2 Situation Awareness (SA), Decision Making and Command and Control

Situational Awareness (SA) in military parlance is the ability to maintain a constant, clear mental picture of relevant information and the tactical situations (e.g. friends and threats). The traditional SA exists in three levels: the perception of elements in the environment within time and space, the comprehension of their meaning, and the projection of their status in the near future.

SA models focus heavily on human factors for perceiving, comprehending and projecting including mental and team models, sensemaking [21], and communication models in computational linguistics and machine learning [22].

3.2.2.1 DoD Budget Processes

How much and when to buy: control, oversight, balance

Warfighter’s Requirements

Why and what to buy: mission success at lowest cost, service advocacy

Acquisition

Final Products for Procurements

How to buy: balancing cost, schedule, performance

Related to SA is a Decision Support System (DSS) which is a computer system that supports decision-making and command and control (C2) activities. A DSS architecture typically includes a knowledge database, a model and a user interface. DSS models often are based on machine learning and artificial intelligence, e.g. decision trees [23] and intelligent agents [24].

There are many differences between SA and DSS. DSS, for instance, may rely on traditional analytics and apply to less dynamic data. SA emphasizes the real-time information gathering, communication methods and collective knowledge that might result in better operational capabilities. Therefore, it may require not merely DSS technologies such as machine learning systems and decision making algorithms but also smart infrastructures to achieve real-time (e.g., in a crisis response situation) and collective intelligence (e.g., in a social web).

In a use case, the NPS team has been studying the DoD acquisition decision making [16, 25-30] since 2009. The US DoD acquisition process is extremely complex. There are three key processes that must work in concert to deliver the capabilities: the warfighters’ requirements/needs; the DoD budget planning and the final products for procurement as in Figure 8. Each process produces Big Data. There has been a critical need for automation, validation, and discovery to help acquisition professionals, decision makers and researchers understand the data and optimize the DoD resources.

Since 2009, the NPS team has been working on the research questions, for example, can the Big Data be used to produce the awareness of the fit between DoD programs and warfighters’ needs? Can gaps be revealed? The NPS team performed studies in the following areas:

• Compare Urgent Need Statements with Trident Warrior technologies
• Compare congressional budget documents with the warfighters’ needs
• Compare categories of data in the Acquisition Visibility Portal

The NPS team took a detailed look at the Research, Development, Test and Evaluation (RDT&E) budget modification practice from one year to the next over the course of ten years and about 450 DoD Program Elements. The NPS team found a pattern that the programs with fewer links (measured by LLA) to warfighters’ requirements, received more budget reduction in total but less on average, indicating the budget reduction may have focused only on large and expensive programs rather than perhaps cutting all the programs that do not match warfighters’ requirements. Furthermore, the programs with more links to each other received more budget reduction in total, as well as on average, indicating a pattern of good practice of allocating DoD acquisition resources to avoid overlapping efforts and to fund new and unique projects. These findings were useful as validation and guidance for future decision processes for automatically identifying programs to match warfighter’s requirements, limit overall spending, minimize efficiencies, eliminate unnecessary cost and maximize the return of investment.

Figure 8. DoD Acquisition Decision Making
3.2.3 Prediction, Deep Learning, Pattern Recognition and Anomaly Detection

Predictive models are paramount to machine learning their predictive power comes from empirically reviewed data. Machine learning algorithms are divided into supervised learning and unsupervised learning. Supervised learning is accurate but more expensive due to intervention costs. Unsupervised learning focuses on data-driven discovery and often is linear scale-up on the number of machines and the size of the data. Thus unsupervised learning is the proposed core analytic strategy for both commercial and DoD Big Data. The NPS team show LLA, SSA and CLA are important unsupervised learning methods for pattern recognition, anomaly detection and data fusion.

In many cases of the DoD applications, Big Data is buried in the complex business processes and the data fusion has to be performed on a vast amount of data sources from the complex business processes. In a use case entitled “Comprehensive Approach to Identifying and Sourcing NATO’s Future Capability Requirements,” the NPS team applied LLA, SSA and CLA to identify and predict NATO capabilities and force requirements to improve the US EUROPE COMMAND (EUCOM)’s visibility and recommend new collaborations toward “Smart Defense” projects. The NPS team first interviewed USEUCOM’s desk officers and planning specialists and gained an understanding of their business processes and the Big Data involved in these processes shown in Figure 9.

The NPS team then conducted the following studies using LLA, SSA and CLA:

1. Compare Chicago Summit Open Sources and Smart Defence (SD) Database, the SD database contains structured and unstructured data about all the SD projects
2. Compare Minimum Capability Requirements (MCR) and SD Database
3. Compare 28 Bluebooks of NATO countries

Figure 10 shows an example of visualization from (1). Themes were discovered automatically and can be drilled down to the original data or the features (word pairs) that describe the consensus and gaps between the Chicago Summit Open Sources and the SD database. Themes were further categorized into popular, emerging, and anomalous concepts. The NPS team showed that popular concepts are highly correlated with the discovered consensus among compared data sources. In contrast, the emerging and anomalous themes are highly correlated with the gaps in the business processes which may need further investigation and could provide guide for future decision processes, for example, discovery of interesting resource relocation opportunities that can be used by the USEUCOM and Smart Defense programs to advance US interests.

3.2.4 Knowledge Management, Collaboration and Network Analysis

Graph and network analysis are important for knowledge management and collaboration. The current research focuses on direct social links among social entities of people or organizations regardless of the contents [30]. The study of centrality has been a focal point for the social network structure studies to discover mavens, leaders, bridges, isolated nodes and peripheries.

So called “metadata analysis,” applied to social entities whose profiles collected from structured data (e.g., Palantir [31]) has also drawn attention. It can infer two people are linked because they share the same metadata attributes [32], for example, two people may share a same metadata attribute such as belonging to a same social club. LLA-generated semantic networks can infer that two people are linked because they share the same content, for example, two people are both interested in “information assurance.” LLA can be used to discover such keywords for interesting connections.

After the Haiti earthquake in 2010, US military and civil organizations provided rapid and extensive relief operations. In a use case entitled “Open Source, APAN Network and Haiti Operation Data Analysis [29][30],” the NPS team applied LLA to show an overall picture of how military and civil organizations actually collaborated. The NPS team first examined ~2600 open source data from the social media platforms Twitter, Facebook and news-feed web sites [29]. The NPS team discovered the synergy patterns and the organizations involved in disseminating information orderly and efficiently in the operation. The NPS team also analyzed the All Partners Access
Network (APAN) data [30] with official briefings of 317 PDF files of situation reports, 1400 forum posts, and 3900 blog messages. By using social, metadata and LLA-generated semantic networks, the NPS team found that these were the organizations which had no social connections with others, however, they shared similar metadata attributes, discussion content, and therefore, may be predicted as potential high-value targets in the future decision and collaboration processes.

4. Acknowledgements
   The authors thank Major Henry R. Salmans III, USMC (Retired) of CSC, Technology Services Organization, Programs & Resources, HQMC at the Marine Corps Information Technology Center, who provided many relevant insights and in-depth discussions.

4.1 Funding grew exponentially as the commercial practicality of using “data mining” to gain competitive edge was identified and articulated to industrial executive leadership. This article explores the landscape.

4.2 NoSQL databases are increasingly used in Big Data and real-time applications because of simplicity of design, horizontal scaling, and finer control over availability. The data structures used by NoSQL databases make some operations faster than those used in relational databases.

ABOUT THE AUTHORS

Dr. Ying Zhao is a research associate professor at the Naval Postgraduate School and frequent contributor to DoD forums on knowledge management and data sciences. Her research and numerous professional papers are focused on knowledge management approaches such as data/text mining, Lexical Link Analysis, system self-awareness, Collaborative Learning Agents, search and visualization for decision-making, and collaboration. Dr. Zhao was principal investigator (PI) for six contracts awarded by the DoD Small Business Innovation Research (SBIR) Program. Dr. Zhao is a co-author of four U.S. patents in knowledge pattern search from networked agents and data fusion and visualization for multiple anomaly detection systems. She received her Ph.D. in mathematics from MIT and is the Co-Founder of Quantum Intelligence, Inc.

E-mail: yzhao@nps.edu

Dr. Doug MacKinnon is a research associate professor at the Naval Postgraduate School (NPS). Dr. MacKinnon is the deputy director of the Distributed Information and Systems Experimentation (DISE) research group where he leads multi-disciplinary studies ranging from leading the Analyst Capability Working Group (ACWG) for the U.S. Air Force, studying Maritime Domain Awareness (MDA), as well as Knowledge Management (KM) and Lexical Link Analysis (LLA) projects. He also led the assessment for the Tasking, Planning, Exploitation, and Dissemination (TPED) process during the Empire Challenge 2008 and 2009 (EC08/09) field experiments and for numerous other field experiments of new technologies during Trident Warrior 2012 (TW12). He teaches courses in operations research (OR) and holds a PhD from Stanford University, conducting successful theoretic and field research in Knowledge Management (KM). He has served as the program manager for two major government projects of over $50 million each, implementing new technologies while reducing manpower requirements. He has served over 20 years as a naval surface warfare officer, amassing over eight years at sea and serving in four U.S. Navy warships with five major, underway deployments.

Dr. Shelley Gallup is a research associate professor at the Naval Postgraduate School’s Department of Information Sciences, and the director of Distributed Information and Systems Experimentation (DISE). Dr. Gallup has a multidisciplinary science, engineering, and analysis background, including microbiology, biochemistry, space systems, international relations, strategy and policy, and systems analysis. He returned to academia after retiring from naval service in 1994 and received his PhD in engineering management from Old Dominion University in 1998. Dr. Gallup joined NPS in 1999, bringing his background in systems analysis, naval operations, military systems, and experimental methods first to the Fleet Battle Experiment series (1999–2002) and then to Fleet experimentation in the Trident Warrior series (2003–2013). Dr. Gallup’s interests are in knowledge Management and complex systems field experimentation.
REFERENCES


