Big Data and Deep Analytics Applied to the Common Tactical Air Picture (CTAP) and Combat Identification (CID)

Ying Zhao, Tony Kendall and Bonnie Johnson
Naval Postgraduate School, Monterey, CA 93943, U.S.A.

Keywords: Big Data, Deep Analytics, Common Tactical Air Picture, Combat Identification, Machine Vision, Object Recognition, Pattern Recognition, Anomaly Detection, Lexical Link Analysis, Heterogeneous Data Sources, Unsupervised Learning.

Abstract: Accurate combat identification (CID) enables warfighters to locate and identify critical airborne objects as friendly, hostile or neutral with high precision. The current CID processes include processing and analysing data from a vast network of sensors, platforms, and decision makers. CID plays an important role in generating the Common Tactical Air Picture (CTAP) which provides situational awareness to air warfare decision-makers. The Big "CID" Data and complexity of the problem pose challenges as well as opportunities. In this paper, we discuss CTAP and CID challenges and some Big Data and Deep Analytics solutions to address these challenges. We present a use case using a unique deep learning method, Lexical Link Analysis (LLA), which is able to associate heterogeneous data sources for object recognition and anomaly detection, both of which are critical for CTAP and CID applications.

1 INTRODUCTION

An accurate, relevant and timely CID capability enables warfighters to locate and identify critical airborne objects as friendly, hostile or neutral with high precision. The objective of the CTAP is to provide tactical situational awareness to the decision-makers; and thereby provide critical information to support the engagement events and courses of action that protect Navy and Joint assets. An effective CID and CTAP capability supports the optimal use of long-range weapons, aids in fratricide reduction, and ultimately reduces or minimizes friendly forces' exposure to enemy fire. The CID process is an essential part of generating a CTAP.

Traditionally, CID decisions are derived from data from intelligence, surveillance, and reconnaissance (ISR) sensors. This research group has noted that the size and heterogeneity of the data from these sensors creates a Big Data environment. The current tactical information systems cannot meet the timelines required for CID in complex threat environments. Nor can they process and analyze additional types of data that may support CID, such as information from the Internet, social media, and commercial airline information. We are studying new methods such as Big Data and Deep Analytics that show promise for handling and analyzing the rising tide of sensor and non-sensor data in a timely manner.

The Aegis combat system, CEC, and Link 16 are critical systems supporting CID for sharing data among distributed platforms, correlating and fusing data, and displaying airborne object tracks. Additionally, the current CID processes include the use of Naval CTAP components and combinations of:

- Platforms: destroyers, cruisers, carriers, F/A-18s, E-2C/D, LHD/LHA's and Amphibious Assault Ships.
- Sensors: radar, Forward Looking Infrared (FLIR), Identification Friend or Foe (IFF), Precision Participation Location Identifier (PPLI), and National Technical Means (NTM)
- Networks: Cooperative Engagement Capability (CEC), Link-16 Global Command and Control System (GCCS), and Global Information Grid (GIG)
- Decision makers: Air and Missile Defense Commander (AMDC), Air Warfare (AW) Officer, Tactical Action Officer (TAO) and Air Defense Officer (ADO)

Zhao, Y., Kendall, T. and Johnson, B.
Big Data and Deep Analytics Applied to the Common Tactical Air Picture (CTAP) and Combat Identification (CID).
DOI: 10.5220/0006086904430449
In *Proceedings of the 8th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2016) - Volume 1: KDIR*, pages 443-449
ISBN: 978-989-758-203-5

The challenges for CTAP and CID include:

- An extremely short dwell time for fusion, decision making, and targeting.
- Uncertain and/or missing data outside sensor ranges (e.g., radar). For example, track pictures are uncertain with track conflicts, multiple objects per track or multiple tracks per object.
- Manual decision-making. For example, complex threat environments can create situations in which decision-makers can be overwhelmed by large amounts of data, uncertain track pictures, and complicated doctrine.
- Hard-to-detect anomalies and a lack of predictive analytic capabilities.
- Manual methods for incorporating electronic warfare (EW), electronic intelligence (ELINT), and non-cooperative sensor measurements and signature databases, into the CID process.

The contribution of this paper is to position various Big Data and Deep Analytics in the context of Big "CTAP and CID" Data. We also show a unique Deep Learning method, i.e., Lexical Link Analysis (LLA), which uses a bi-gram model to link any two entities across multiple contexts and associate heterogeneous data sources for object recognition and anomaly detection.

2 BIG DATA

2.1 Big Data Problem

Today, Big Data is omnipresent. Big Data science intervenes with traditional data sciences. We are compelled to ask - What is new? Here, we examine some aspects of the problem:

- 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 are also unstructured text, images, audios, videos, geospatial data, and 3D data.
- Big rise in needs: It is critical for business to transform data into *smart* data, or actionable knowledge.
- Big rise in analytics: Traditional data sciences including statistics, numerical analysis, machine learning, data mining, business intelligence, and artificial intelligence have evolved into Big Data analytics or Deep Analytics. These technologies can be overwhelmingly complex, requiring diversified and extensive expertise.

2.2 Tools and Challenges

Big Data requires massively parallel software on thousands of servers. The current technologies are dominated by systems that provide 1) data collection, ingestion, integration and safe storage; 2) parallel/distributed processing; and 3) Deep Analytics.

As part of the 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. NoSQL databases are increasingly used because of simplicity of design, horizontal scaling, and finer control over availability.

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 (Spark, 2016) is replacing Map/Reduce for its speed and in-memory computation.

As the data size gets bigger, the statistical significance for an analysis is often guaranteed due purely to the data 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?

2.3 Big CTAP and CID Data

Data sources for Department of Defense (DoD) applications including disparate, multi-sourced real-time sensors 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 make 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 Bata into the battlefield (NBD, 2014).

The data used for CID come from a combination of massive cooperative and non-cooperative sensors, organic sensors and non-sensor information. In reality, each sensor collects certain attributes. The Big CID Data need to be fused over time and space since they are collected in a distributed fashion as shown in Figure 1.

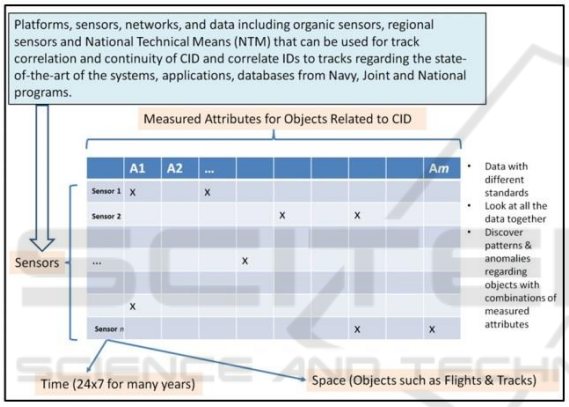

Figure 1: A holistic view of Big CTAP and CID Data.

3 DEEP ANALYTICS

3.1 Commercial Trends

It is critical to turn Big Data into *smart data*. One important trend is Deep Analytics including analytic algorithms that can be run in parallel and distributed fashion.

Predictive analytics turns Big Data into smart data, for example, accurately forecasting high-value targets. The topic has been thoroughly studied in traditional supervised learning. Some algorithms are implemented using Big Data and Deep Learning requirements such as Map/Reduce paradigm, Mahout (2016) and Spark, (2016).

Social network analysis and graph search require graph analyses leveraging massively parallel processors. Graph algorithms can process petabytes of data and are considered as the core drivers of Big Data. Spark, Titan and Neo4j are used for Big Graph.

3.2 Deep Learning

Deep Learning models, in a nutshell, are much larger machine learning models with many more parameters that are specifically designed to handle Big Data. Deep Learning models including Deep supervised machine learning models, e.g., convolutional neural networks (CNN, 2016) with much deeper hidden layers; Deep reinforcement learning models; and Deep unsupervised machine learning models for recognizing objects and patterns of interest. Sparse coding (Olshausen and Field, 1996) and self-taught learning (Le, Ranzato, Monga, Devin, Chen, Corrado, Dean, and Ng, 2012) make Deep unsupervised learning possible. The self-taught learning is also a deep unsupervised learning model that approximates the input for unlabelled 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 (LeCun, Bottou, Bengio, and Haffner, 1998). Deep Learning models links machine vision and text analysis smartly. For example, Latent *Dirichlet* Analysis (LDA, Blei, Ng and Jordan, 2003) is a sparse coding where a bag of words used as the sparsely coded features for text (Raina, Battle, Lee, Packer and Ng, 2007). Our methods Lexical Link Analysis (LLA, Zhao, Gallup and Mackinnon, 2011, 2015), System-Self-Awareness (SSA, Zhao and Zhou, 2016), and Collaborative Learning Agents (CLA, Zhou, Zhao and Kotak, 2009) can be viewed as Deep models, in the sense similar to the LDA method as a Deep Learning method (Raina, Battle, Lee, Packer and Ng, 2007).

4 DEEP ANALYTICS FOR CID

4.1 The CTAP Cloud Concept

We first explored how Big Data and Deep Analytics could address the challenges of CID. We developed a CTAP Cloud Concept as shown in Figure 2. Conceptually, it can be physically associated with a Big Data cloud implementation such as the Naval Tactical Cloud (NTC). It could store traditional CTAP and CID data sources as well as the additional non-traditional data sources, such as temporal, spatial and organic sensor data that are collected but not currently used (e.g. Aegis residual data), open sources flight schedules, advanced

(EW/ELINT) signature data sources, and intelligence data. These new data sources could be fused and analyzed in parallel using Deep Analytics in a CTAP Cloud. The resulting knowledge repository, i.e., *smart data*, could be searched, matched, and cross-validated with real-time new data streams. For example, the cloud could send or push the smart data (e.g. early warnings or alerts) to various platforms within a battlespace. A platform with partial or uncertain sensor/track data could send a real-time query to the cloud to find a higher certainty match. The smart data push and pull would have a relatively small data size and therefore not strain current networks for transmission between platforms.

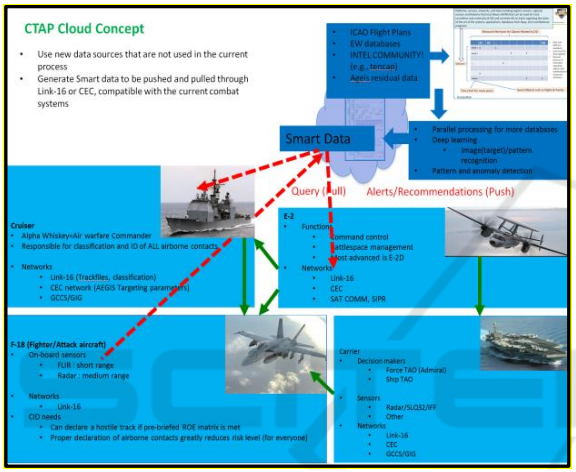

Figure 2: The CTAP Cloud Concept.

The CID/CTAP application domain is an extremely complex and amazingly interesting field in terms of the roles that many Big Data and Deep Models can play. We investigated Big Data and Deep Analytics to address CTAP and CID challenges including the following areas:

- Machine vision and Deep Learning models: These algorithms have the potential to improve object recognition, classification accuracy and probability of correctly identifying air objects by associating, correlating, and fusing heterogeneous data sources that do not share data models. This process is demonstrated with unclassified tactical data samples of infrared (IR) and Electro-optical (EO) images in this paper (Section 4.1).
- Pattern recognition, anomaly detection and unsupervised learning models: We developed and selected pattern recognition and anomaly detection algorithms that could be used for identifying intent, air picture event anomalies or launch predictions.
- Optimization, decision making and deep reinforcement learning models: We investigated Big Data optimization, decision making and reinforcement learning models such as Q-learning in Soar (2016) and DeepMind (2016) that can be used for CTAP and CID. The models could not only automate many current manual CTAP and CID processes but also have the potential to enhance future CTAP capabilities such as uncooperative game theory and total battle management.
- Fast, parallel and distributed computing models: Commercial tools for Big Data may not satisfy CTAP and CID which requires fast, parallel and distributed computing. Tools such as associative arrays (Kepner, Chaidez, Gadepally and Jansen, 2014), BigDAWG polystore (2016) and GraphBLAS (2016) may have the potential to address the requirements.

4.2 Machine Vision and LLA

LLA is an unsupervised deep learning method, implemented in parallel and distributed fashion. By using LLA, a complex system can be expressed in a list of attributes or features with specific vocabularies or lexicon terms to describe its characteristics and surrounding environment. LLA uses bi-gram word pairs, compared to LDA, are potentially more meaningful and sparse coded features. Specifically, LLA is a form of text analysis. For example, word pairs or bi-grams as lexical terms and features can be extracted and learned from a document repository. For a text document, words are represented as nodes and word pairs as the links between nodes. Figure 3 shows an example of such a word network, for example, “cash dividend”, “dividend report”, and “market influence” are examples of bi-gram word pairs from a financial news data sample. LLA is related to Latent Semantic Analysis (LSA, Dumais, Furnas, Landauer and Deerwester, 1988), Probabilistic Latent Semantic Analysis (PLSA, Hofmann, 1999), WordNet (Miller, 1995), Automap (CASOS, 2009), and LDA (Blei, Ng and Jordan, 2003). LDA uses a bag of single words (e.g., associations are computed at the word level) to extract concepts and topics. LLA uses bi-gram word pair. LLA was previously used in many examples for understanding DoD data (Zhao, McKinnon and Gallup, 2009, 2011, 2015).

The unique characteristic of LLA is that the Bi-gram also allows it to be extended to data other than text (e.g., numerical or categorical data). For example, structured data from databases can be discretized or categorized to word-like terms. For

example, features, such as "age_older_than_65" and "gender female" can be generated from the "age" and "gender" attributes.

The word pair model can further be extended to a context-concept-cluster (CCC, Zhao and Zhou, 2014) model. A *context* is a word or attribute that are shared by multiple data sources. A context can be a location, a time point or an object that are shared across data sources. Using this generalization, a bi-gram or word pair model can used to link any two entities across multiple contexts. This is the key point for LLA used in the CTAP and CID analytics to associating heterogeneous data sources (see the use case in Section 4.3).

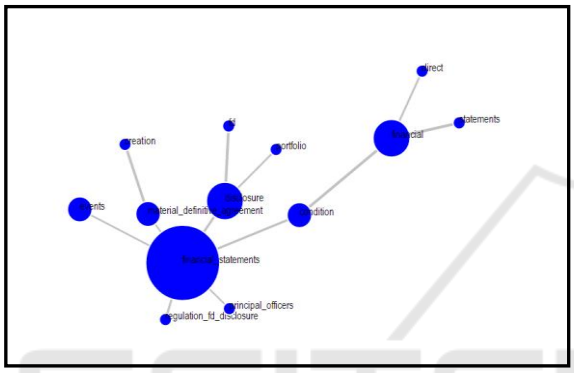

Figure 3: An example of a theme or topic discovered by LLA for a text data.

4.3 Use Case

4.3.1 Data Samples

The sample data contains a large collection of visible and IR imagery collected by the US Army Night Vision and Electronic Sensors Directorate (NVESD). It contains 207 GB of IR imagery and 106 GB of visible imagery along with an image viewer, ground truth data, meteorological data, photographs of the objects, and other documentation to assist the user in correctly interpreting the imagery. All imagery was taken using commercial cameras operating in the IR and visible bands.

The data was pre-processed using SiFT-like code (SiFT,2016) to generate 400 visual "words" (histograms to the centers of k-means) so LLA bi-gram models can be applied. Figure 4 summarizes the processed data, consisting of 4500 total training images with 400 features or visual words for nine classes of objects (target vehicles) and two different modalities (i.e., IR and EO sensors). Therefore with 4500 total images per test, there were a total of 9000 images. Each object in each mode contained 500 images. The baseline object recognition for this data was given using the method of representation learning through topic models (Flenner, 2015).

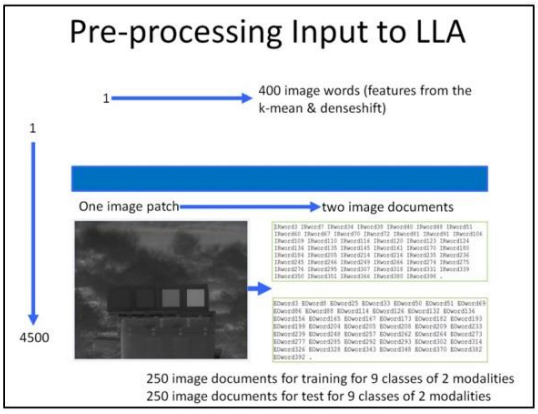

Figure 4: Images data were pre-processed to feed to LLA.

4.3.2 Associating Data Sources

Another challenge to improving CID is that traditional ISR sensor data does not have standardized or common data attributes; and often there are missing attributes. For example, IR and EO sensors use completely different features (vocabularies). We used a generalized LLA model of bi-gram co-occurrence of spatial locations (i.e., image patches) to link two modalities. For example, an IR image feature (i.e., the concept in a CCC model) describes the same image characteristics with an EO image feature because these two features are frequently used in the same image patches (i.e., contexts in the CCC model). This learning paradigm is a generic framework to fuse two data sources. The data sources do not share vocabularies and some data are even missing or uncertain. Nevertheless, they can all be fused into one picture using this method.

4.3.3 Applying LLA

We applied LLA to the data set as follows:

Step 1: Divide data into a training data set and a test data set: each object has 500 images which are divided into 250 images for training and 250 images for test. Bi-gram and association learning are performed on 250 training images. There are 36 data sets of nine training sets and nine test sets for the two modalities for the nine objects.

Step 2: Extract bi-gram features for each data set in a distributed fashion. Uni-gram or bi-gram features for each of 36 data sets are then processed separately.

Target Type	
Pickup	0
Sport Utility Vehicle	1
BTR70 – Armored Personnel Carrier	2
BRDM2 – Infantry Scout Vehicle	3
BMP2 – Armored Personnel Carrier	4
T62 – Main Battle Tank (broke down early, rarely used)	
T72 – Main Battle Tank	5
ZSU23-4 - Anti-Aircraft Weapon	6
2S3 – Self-Propelled Howitzer	7
MTLB – Armored Reconnaissance Vehicle Towing a D20 Artillery Piece	8

Figure 5: Target types.

An unsupervised learning system ideally should discover nine clusters. According to the data description in Figure 5, some of the nine (0-8) objects are very similar in nature. An automatic unsupervised method is expected to see fewer clusters of objects. Figure 5 shows that 36 data sets are grouped into five clusters.

4.3.4 Results

We first applied the "uni-gram" setting: this is related to the "bag-of words" approach where only the 400 features are used to distinguish the objects. The correlation for any of the two data sets is flat and similar, indicating a uni-gram or a bag-of-words. This indicates that the 400 features are not good for separating, recognizing and distinguishing these objects.

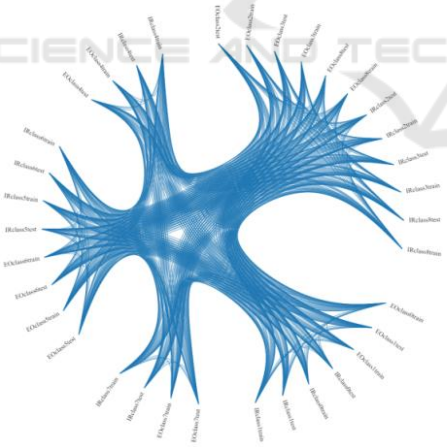

Figure 6: LLA discovered five clusters of objects.

The second setting of LLA we used generated both full bi-gram and association learning between IR and EO. This is shown in Figure 6. There are five clusters for nine classes of the objects as follows:

- Cluster 1: class 0 (pick up) and class 1 (sport utility vehicle)
- Cluster 2: class 2 (infantry scout vehicle), class 3 (armored personal carrier) and class 8 (armored reconnaissance vehicle towing a D20 artillery piece)
- Cluster 3: class 4 (armored personal carrier)
- Cluster 4: class 5 (main battle tank) and class 6 (anti-aircraft weapon)
- Cluster 5: class 7 (self-propelled howitzer)

Five clusters are consistent with the ones marked in Figure 5. Initial results in the use case show Deep Analytics such as LLA can automatically discover categories of objects in a Big Image Data.

5 FUTURE WORK

Our team plans to combine and test sample CID track data with FAA and twitter data. We will test several behavior-based Deep Learning algorithms to see if there are normal patterns and anomalies for the military aircraft and commercial ones. The goal is to see if added databases and Deep Analytics will improve CID and the CTAP.

6 CONCLUSIONS

We identified and assessed the current CTAP and CID Big Data problems and challenges; and identified key Deep Analytics required to address the challenges. Big Data and Deep Analytics were found to have potential in improving object recognition and classification through the utilization of more databases, distributed computation, and data fusion. These applications could be realized by the adoption of our cloud architecture concept which includes continuous monitoring in time and space; and collecting and processing data in a cloud. Finally, the team found that the unique LLA method is able to associate heterogeneous data sources and perform Deep unsupervised Learning; which implies a future application to the CID and CTAP.

ACKNOWLEDGEMENTS

Thanks to our research sponsors, Mr. William A. Treadway and Mr. Richard Heathcote from the OPNAV Combat Identification Capability Organization. Thanks to Dr. Arjuna Flenner in the U.S. Naval Air Warfare Center, who provided insightful domain expertise and discussion for the research. Thanks to the Naval Postgraduate School Research Program for funding this project.

REFERENCES

Blei, D., Ng, A. and Jordan, M. 2003. Latent Dirichlet allocation. Journal of Machine Learning Research, 3:993-1022. Retrieved from http://jmlr.csail.mit.edu/papers/volume3/blei03a/blei03a.pdf.

BigDAWG, 2016. A Demonstration of the BigDAWG polystore system. http://livinglab.mit.edu/wp-content/uploads/2016/01/bigdawg-polystore-system.pdf.

Center for Computational Analysis of Social and Organizational Systems (CASOS) 2009. AutoMap: extract, analyze and represent relational data from texts. Retrieved from http://www.casos.cs.cmu.edu.

DeepMind, 2016. https://deepmind.com/

Dumais, S. T., Furnas, G. W., Landauer, T. K. and Deerwester, S. 1988. Using latent semantic analysis to improve information retrieval. In Proceedings of CHI'88: Conference on Human Factors in Computing, 281-285.

Freeman, L.C. 1979. Centrality in social networks I: conceptual clarification. Social Networks, 1: 215-239.

Flenner, A. 2015. Representation learning through topic models. NFCS NAWCWD, China Lake, in the 2015 National Fire Control Symposium.

GraphBLAS, 2016. Graph algorithms for basic linear algebra subprograms (BLAS). http://istc-bigdata.org/GraphBlas/

Hofmann, T. 1999. Probabilistic latent semantic analysis. Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence, Stockholm, Sweden.

Kepner J., Chaidez, J., Gadepally, V., Jansen, H. 2014. Associative Arrays: Unified Mathematics for Spreadsheets, Databases, Matrices, and Graphs. http://db.csail.mit.edu/nedbday15/pdf/Paper7.pdf.

LeCun, Y., Bottou, L., Bengio, Y. and Haffner, P. 1998. Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11):2278-2324. Retrieved from http://yann.lecun.com/exdb/publis/pdf/lecun- 01a.pdf.

Miller, G. A. 1995. WordNet: a lexical database for English. Communications of the ACM, 38(11).

Mahout, 2016. http://mahout.apache.org/

NBD, 2014. Navy Big Data. http://defensesystems.com/articles/2014/06/24/navy-onr-big-data-ecosystem.aspx.

Olshausen, B. and Field, D. 1996. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. Nature.

Raina, R., Battle, A., Lee, H., Packer, B., and Ng, A.Y. 2007. Self-taught learning: transfer learning from unlabeled data. In ICML.

Ranzato, M., Monga, R., Devin, M., Chen, K., Corrado, G., Dean, J., Le, Q. V. and Ng, A.Y. 2012. Building high-level features using large scale unsupervised learning. Proceedings of the 29th International Conference on Machine Learning (ICML-12). Retrieved from http://arxiv.org/pdf/1112.6209v5.pdf.

Soar, 2016. http://soar.eecs.umich.edu/

Spark, 2016. http://spark.apache.org/

SiFT, 2016. http://www.vlfeat.org/

Zhou, C., Zhao, Y., and Kotak, C. 2009. The Collaborative Learning agent (CLA) in Trident Warrior 08 exercise. In *Proceedings of the International Conference on Knowledge Discovery and Information Retrieval (KDIR)*, Madeira, Portugal.

Zhao, Y., Gallup, S.P. and MacKinnon, D.J. 2011. System self-awareness and related methods for improving the use and understanding of data within DoD. Software Quality Professional, 13(4): 19-31. http://asq.org/pub/sqp/

Zhao, Y., Mackinnon, D. J., Gallup, S. P. 2015. Big data and deep learning for understanding DoD data. Journal of Defense Software Engineering, Special Issue: Data Mining and Metrics.

Zhao, Y., Mackinnon, D. J., Gallup, S. P. 2015. Big data and deep learning for understanding DoD data. Journal of Defense Software Engineering, Special Issue: Data Mining and Metrics.

Zhao, Y., Zhou, C. 2016. System Self-Awareness Towards Deep Learning and Discovering High-Value Information. The 7th IEEE Annual Ubiquitous Computing, Electronics and Mobile Communication Conference, New York City, USA 20 - 22 October 2016.

Zhao, Y. and Zhou, C. 2014. System and method for knowledge pattern search from networked agents. US patent 8,903,756. https://www.google.com/patents/US8903756.