Executive Summary:

- Create a new architecture to collect, share, and analyze data that can be mined for patterns that humans cannot perceive
- Utilize data to enable better decision-making in all facets of the Department, providing significant advantages that adversaries cannot anticipate
- Forge culture of data collection/analysis to meet the demands of a software-centric combat environment

Full Recommendation 12:

ARGUMENT

Data is the 21st century equivalent of a global natural resource, like timber, iron, or oil previously – indispensable for sustaining military innovation and advantage. The next global conflicts will be fueled by data. The rapidly expanding power of new mathematical and computing techniques to reveal insights into intentions and capabilities, and to enhance accuracy, lethality, and speed, depend on immense data sets to train algorithms and from which to extract information. The data that provide the raw materials from which to identify patterns, as well as the anomalies that defy them, constitute the fuel that powers the engine of machine learning (ML). Whoever amasses and organizes the most data first will sustain technological superiority, so it is incumbent upon the Department to collect, store, share, analyze, and protect its data faster and better than its competitors. Data must be regarded as one of the most powerful resources in the Department's arsenal.

PROBLEM STATEMENT

DoD does not view data as strategic resource. Without a data strategy to collect, protect, and make available this critical resource, DoD will not be able to create or sustain competitive advantages over our adversaries. Meanwhile, the lack of a modern approach to data is consuming vast financial and personnel resources.

BENEFITS

DoD should establish a new paradigm for the collection, exchange, availability, analysis, and protection of all DoD data. Data should be mined for patterns to train ML systems that will provide strategic and tactical insights no human could ever generate. This advancement will transform the Department in four ways:

1. Data-enabled capabilities will enhance the lethality, speed, precision, and survivability of warfighters

2. Storage of data will allow artificial intelligence (AI) and ML-based systems to identify significant cost and time efficiencies that can be deployed across the enterprise

3. Ubiquitous and deep data sources will advance our capabilities past those of our adversaries

4. Encrypting and aggregating data will allow us to protect them using new sophisticated cyber security techniques

The scope of this recommendation comprises all development, operational, tactical, and strategic data across the total DoD enterprise. Any data repositories or similar architecture will therefore require use of the encryption and access rights technologies necessary to protect secure and third party proprietary data. This approach is more secure than our current one, as DoD would be protecting the data, not merely the networks that surround them. Moreover, aggregating data allows us to use ML to pattern who accesses data and why, which enables anomaly detection that can be used to counter insider threats.

FIXABLE PROBLEMS

Every time a U.S. fighter plane takes the air or a submarine slips below the surface, it collects enormous amounts of data from its sensors, picking up vital information about targets in all domains. It generates a continuous stream of data on the performance characteristics of its systems, the mission profile it is assigned, and the behavior of the crew operating the controls. If captured and analyzed over time, these data provide unprecedented insight into enhancing performance of the operators, maximizing the performance of the platforms, detecting patterns across a fleet of planes or vessels, understanding the capabilities and patterns of potential adversaries, and reducing the cost of maintenance. However, these data are virtually never collected, and when they are, they are seldom organized in a helpful way, often discarded, and not stored in a way to discern patterns that could show us how to reduce IED-related casualties, defend against cyber attacks, identify which zip codes tend to produce the most talented recruits, or map out how our adversaries attempt to wage hidden economic warfare against us. Our inability to track these patterns disadvantages us against our competitors.

This challenge is characterized by three interdependent areas: 1) data systems; 2) data policies; and 3) data talent.

Data systems: DoD does not adopt the latest technology needed to capture and understand data, creating an infrastructure gap that becomes harder to bridge the longer it is neglected.
Data policies: Since many of the Department's challenges with data are cultural (i.e. DoD organizations are not used to collecting or sharing data), the Secretary's role in this endeavor is critical, particularly because new policy and legal frameworks will be necessary to change the status quo.
Data talent: DoD should recruit and develop data scientists that can prioritize speed and agility, and apply data science techniques common in the private sector but novel to DoD. A new breed of talent is necessary because without the requisite understanding of how to build and interpret algorithms at all levels of organizations – which is profoundly lacking among DoD personnel – advanced analytics will provide a false sense of security for prediction of catastrophic risks.

WHAT CHANGE LOOKS LIKE

Modern data set storage and analysis capabilities no longer make it necessary to devise a common labeling scheme for all data that might be addressed by an application; rather, it is more effective to connect and collect already-labeled data and then, using ML-enabled algorithms, automatically label non-labeled data based on similar available labeled data. Google Scholar, Apple photos, and Evernote work this way, for example.

To this end, it is useful to describe how DoD *currently* treats data and how it *should* treat data, per the standard of the most advanced software companies.

Access

- Current approach: DoD agrees on a set of common tags, sets up a database, and forces everyone to enter and/or convert data into a standard form and submit it to the database that responds to standard queries.
- New approach: Make existing data "discoverable" by putting it on NIPRNet or other accessible networks, create code that scans all databases, assemble a knowledge base (such as a Knowledge Graph) to store structured and unstructured information, and start to create linkages between the data (the goal is not just to view the raw data but also to create pattern-matching APIs (Application Programming Interfaces) that allow access to whatever structure exists for those data).

Analytics (data integrity protection)

- Current approach: Using the same database for transactions and analytics.
- New approach: Making an offline copy of data for analytics and decision support.

Usage

- Current approach: Each use case constructs and maintains a separate database.
- New approach: Data can be pulled from various databases for many different purposes, some of which we can't even identify right now, but will be useful in the future.

Standards

- Current approach: Joint standards committees develop mandated one-size-fits-all schemas.
- New approach: DoD promotes flexible opt-in cooperation across the Department to respond to emerging demands and opportunities.

COURSES OF ACTION

While these are not mutually exclusive, they represent tiered options, from the most comprehensive to the most specific:

1. Establish a new DoD-wide data agency

This new agency would make all of DoD's information available to employees, possess the authority to access all data on NIPRNet, and establish a distributed cloud-based set of services that enable DoD personnel to write applications that tap into the database and provide access to this information.

With a budget of $500 million to $1 billion per year, creating one virtualized and distributed logical facility to store data and provide the tools and methods to access and analyze it, will form the foundation – common among the private sector's largest innovative organizations – of strategic and tactical data analytics and machine learning.

2. Make data accessible across silos

Modern data access and translation APIs now provide the power to make existing data across disparate silos available without the need to comprehensively process and store data according to unified standards. In myriad DoD verticals (budgeting, personnel, acquisition, healthcare, logistics, etc.), there are often dozens of databases that are not interoperable, a software design flaw that severely undermines DoD's ability to understand what its vast arrays of data can tell it. Sample topics amenable to analysis from more comprehensive data extraction might include: money flows, why certain officers succeed more than others, how we can purchase superior weapons systems for a greatly reduced price, how best to keep service members and their families healthy, and where and why breakdowns in supplying warfighters with materiel occur.

Traditionally, it has been very difficult to access data across multiple databases within one vertical or Service or Combatant Command or single installation/facility (a "location"), undercutting DoD's ability to analyze its own data. By creating an application at each data storage "location" that mines the data and feeds them into an ML-enabled system, DoD personnel will be able to view patterns never seen before that help them make better decisions in each vertical, Service, COCOM, etc.

3. Apply machine learning to existing data within silos

This approach is a straightforward effort that DoD can and should do now, but it won't solve the underlying data challenges outlined in this recommendation. In this case, data in one location, agency, Service, etc. that has already been labeled can be mined by an ML-enabled application that pre-labels data to enable an exponentially larger increase in efficiency in looking at data normally viewed by humans.

For example, whether it is drone footage viewed by analysts at Creech Air Force Base or sonar scans viewed by analysts in the Fifth Fleet, there is simply too much incoming data for humans to analyze quickly and accurately. As DoD's sensor collection ability increases, this problem will only get worse, and increasing the number of humans watching monitors is neither practical nor effective. Image and video recognition by machines, which are demonstrably faster, more accurate, and more resilient than humans, will provide the needed, timely analysis mission commanders need. Project Maven is an example of how DoD is addressing this issue the right way, showing that this approach is feasible and replicable across all DoD verticals, agencies, or locations.