JULY 2020

Messier than Oil: Assessing Data Advantage in Military AI

CSET Issue Brief

CENTER *for* SECURITY *and* EMERGING TECHNOLOGY

AUTHORS
Husanjot Chahal
Ryan Fedasiuk
Carrick Flynn

Executive Summary

"Data is the new oil," or so we've been told. From policy pronouncements to media reports to op-eds, many have used the attractive analogy when discussing artificial intelligence. Kai-Fu Lee, author of *AI Superpowers*, has written, "in the age of AI, where data is the new oil, China is the new Saudi Arabia."[1]

Yet reality is far messier. With a population of 1.4 billion people, robust surveillance and data collection capabilities, and access to private sector data, the Chinese government appears to have vast quantities of data.[2] But even if China has far more data than the United States, does this raw data necessarily translate into a meaningful advantage for China? And if so, is this enough to overtake the United States in AI? Both countries invest in AI for military applications; will China's potentially greater access[3] to commercial data accelerate its development of AI-enabled weapons relative to the United States?

This paper reviews the challenges in assessing whether the United States or China has a "data advantage" in the military AI realm—i.e., whether one country has access to more data in a way that confers an advantage in developing military AI systems. We provide initial insights for measuring a relative data advantage by answering three questions that are important when evaluating data competitiveness. What does it mean to have a data advantage? Does commercial data matter for military AI? Will big data stay relevant for future AI applications?

Following are the key assessments of this paper:

- Determining whether one country has a data advantage over the other is not as simple as measuring which country produces more raw

data overall. Estimates that compare raw data broadly without looking at specific application or domain areas are oversimplifications that do not accurately reflect the role of data.

- A country that first reaches the experimentation phase (i.e., where data for a specific application is digitally stored, cleaned and transformed, labeled, and optimized to train a machine learning algorithm) is at an advantage over others for that application, as it is positioned to move faster toward developing its aimed AI application.

- Commercial data, while useful, will be less relevant for military operational AI. China's access to commercial market data is unlikely to confer a military operational advantage; data needs for military AI applications are environment-specific, and little ability exists to transfer commercial data and machine learning models to military applications.

- Certain emerging approaches might make big data collected from the real world less relevant in the future, even though the applicability of these approaches to military needs remains unclear.

Data advantage is about more than the sheer quantity of raw data

Determining whether China has an AI-relevant data advantage over the United States is not as simple as measuring which country produces more raw data. Most existing comparisons of national competitiveness in AI data are based on indicators like population size and the percentage of the population engaged in digital activities. Such estimates are too broad to be useful. They attempt to measure the amount and availability of data,[4] but often neglect metrics essential for training AI algorithms, such as data quality and diversity.[5]

Furthermore, although people produce large amounts of data when they engage in various online and offline activities, not all of this data is collected and stored. Even when stored, it generally cannot be used unless cleaned and labelled.[6] Data cleaning is a crucial part of the process: data scientists spend up to 80 percent of their time collecting, cleaning, and organizing data.[7] Finally, broad country comparisons overlook AI algorithmic approaches and technological progress with a potential to make real-world data less relevant for AI development.

AI is a general-purpose technology with promising applications for both military and civilian purposes. U.S. policymakers have raised concerns that the Chinese Communist Party might use its authoritarian power to access the data of Chinese tech firms, such as Baidu, Alibaba, Tencent, and Huawei, to advance Party-state and national security interests.[8] Scholars, too, have broadly linked China's success in commercial AI to its military and espionage AI capabilities.[9] Such concerns result from generalized analogies like "data is the new oil" that don't reflect a clear understanding of AI data or the role of commercial sector's data in defense applications.

Unlike oil—which once processed can be used to fuel industry, electricity, and a range of transport vehicles—data is not an all-purpose resource. Building AI systems in different domains calls for distinct types of data, and data can consist of words, sounds, pictures, ideas, facts, statistics, or anything that constitutes digital information. The type, structure, quality, and availability of data required for military applications of AI differs significantly from the data used in commercial applications. For example, a self-driving car would train on data related to traffic, maps, and adjacent cars—vastly different from the off-road self-driving data needs of a military unmanned ground vehicle.

Broad estimates and generalized analogies, therefore, point us in the wrong direction. Answering the question of data advantage is not as straightforward as comparing countries' oil reserves. Even if we could accurately estimate which nation collects the most raw data from the real world, it would not help us determine who has the most data relevant for military AI. And even if it were possible to estimate accurately which country leads in military AI data, it wouldn't necessarily help us predict whether and how data for commercial AI applications may be used for military purposes that threaten national security interests. Nor will estimates based only on large amounts of real-world data present a complete picture that accounts for questions on its near-term relevancy for AI development.

Data advantage derives from a robust metadata management capacity

At present, there is no consensus, conceptual or methodological, about what it means to have a data advantage. Does it imply access to large reserves of raw data? A single Air Force drone can generate 70 terabytes of data every 14 hours,[10] roughly seven times the amount of data produced by the Hubble Space Telescope in one year.[11] The next generation of wide-area motion imagery sensors will be capable of collecting 2.2 petabytes per day, which is

more than the storage needed for a 24/7 video recording at 1080p for almost seven years.[12] The wealth of data accessible today is overwhelming. In such a scenario, we need to understand when data becomes a military asset. The first step to assessing a country's data advantage is clarifying the role of data in applications of AI.

The role of data in AI

The following is a depiction of the various stages data goes through before it can deliver a successful AI application.

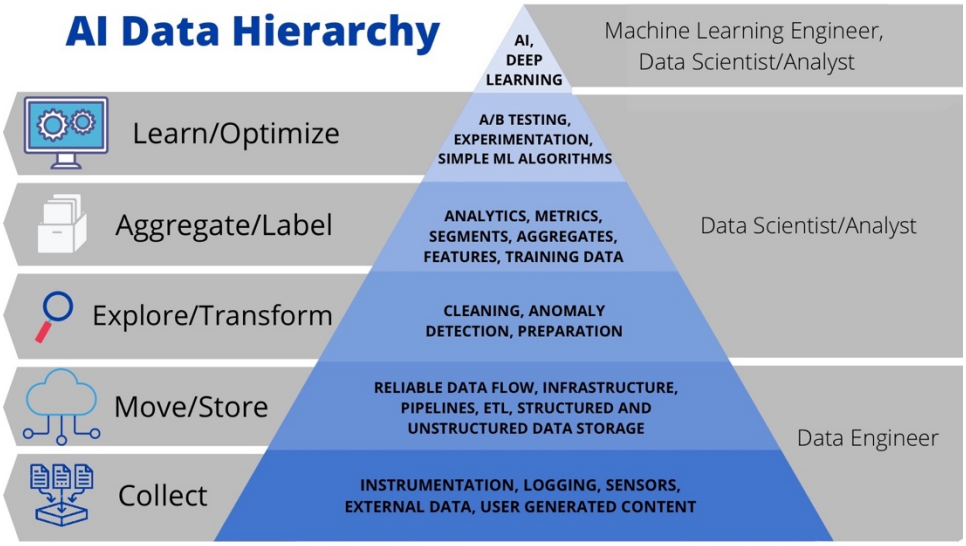

Figure 1: AI Data Hierarchy[13]

At the start of the process is **data collection**—determining what data is needed and if it is being collected. This could be data from sensors, satellite imagery, telemetry data, or even maintenance logs, depending on the system that generates it. Data collection will not mean much unless the data is accessible. Data accessibility depends on how task-relevant data is **moved and stored**—there needs to be a reliable extract, transform, and load (ETL) pipeline[14] that pulls data out of the point of generation and moves it to a useful database, making the data accessible.

Once in an accessible database, this data can be **explored**[15] and transformed further to enable data improvement and analysis. This process involves cleaning the data, finding errors[16] and missing pieces of information such that it can be reliably explored.[17] This step is far more difficult and time consuming than any that follow, especially for large datasets.[18]

When dealing with large amounts of data, sometimes improved data quality is itself a major part of a project. The Joint Artificial Intelligence Center's (JAIC) three-day "Into the Dataverse" challenge for its predictive aircraft maintenance project was tailored specifically for the purpose of getting organized data. The challenge statement read as follows:

"Develop an AI-enabling user interface that can intuitively capture both structured and non-structured maintenance data, and associated maintainer actions, in an efficient and user-friendly manner to produce more accurate maintenance logs."[19]

With clean data in hand, data scientists or relevant ML experts can build upon the information with **analytics** to explain what's happening in a system and why. This process can involve exploring patterns, assigning labels to the training data automatically or manually, and defining metrics to describe the essential information in meaningful ways; it helps researchers understand what they might learn or predict. For machine learning, this is where designing the model begins.

The final step involves **testing and experimentation** to put models in place in an incremental fashion to prevent problems later. Simple baseline ML algorithms can be deployed as part of testing and new features can be added as appropriate to refine and fine-tune the model. This is a quality control process that includes validation and verification. It is also the stage where the algorithm is fit to best match ground truth data. Once robust, the model is deployed and becomes a usable AI application.

Evaluating a lead in data

The United States and China each have military systems generating massive volumes of data, where ML applications could be incorporated to increase the speed, efficiency, and accuracy of tasks, including navigation, information processing, planning, decision support, targeting, and others. Data is most valuable for operational purposes once digitally stored, cleaned, transformed, labeled, and optimized to be deployed in ML algorithms. Therefore, the country that first reaches the experimentation phase

for a set of data can be said to have an advantage over others,[20] positioned to move faster toward developing its desired AI application.

To reach this stage, a country must have a robust metadata management infrastructure supported by reliable data flow pipelines, cloud technology, edge computing, labeling and annotation capacity, among other factors. Official reports in the United States and China indicate significant progress toward this end.

Big data has been a focus of attention in China since March 2014 when the term "big data" was first written into a government work report. This was followed by the launch of the big data strategy in 2016 and subsequent efforts at its implementation and integration within the economy. Until 2019, nearly 31 provincial-level administrative units in China had released relevant documents to promote development of big data industry, including a focus on AI in the past two years. The policy design phase has gradually entered the execution phase, with several provinces and local authorities in China establishing big data bureaus for the implementation of policy objectives.[21]

On the other hand, the United States issued its first government-wide "Federal Data Strategy" in June 2019 along with a 2020 Action Plan for its implementation, offering a vision and guidance on how its various government agencies should manage and use data and what practices can help leverage data value.[22] Several agencies have released their own data strategies in recent years with an emphasis on advanced analytical technologies like AI and ML.[23] Moving from strategy to execution has been hampered as agencies such as the U.S. Department of Defense (DOD) struggle to finalize cloud computing contracts and manage data culture issues.[24]

Public statements by U.S. and Chinese officials highlight the underlying challenges facing their efforts, indicating that neither country has a data management infrastructure fully in place. For instance, during a special lecture held by the Standing Committee of the 13th National People's Congress, a People's Liberation Army expert noted the absence of real data management regulations and the existence of cumbersome bureaucratic processes as obstacles preventing China from fully exploiting its data.[25] There are also significant challenges pertaining to civil-military integration on big data, including lack of effective channels for information communication and issues related to confidential information sharing.[26]

In the United States, the former director of the JAIC, Lt. Gen. Jack Shanahan, repeatedly listed "dirty data" as the Pentagon's AI problem. He attributed it to legacy systems with challenges in integration and interoperability, and lack of manpower to label sensitive operational information. Among other issues, the cost and effort of establishing enterprise-wide data pipelines for DOD platforms will be very challenging given the siloed organizations, talent shortage, and culture of slow-moving systems and information within the Department.[27] In that regard, the JAIC's bottom-up approach to operationalize AI via small projects before scaling it across the DOD appears to support the above assessment of DOD's current lack of data management maturity at an enterprise level.

In practice, data advantage is application-specific and will rely on whether a country has access to usable data for its data scientists/analysts or ML engineers looking to deploy AI models. To assess data advantage, therefore, looking at raw data volume is less useful than evaluating who has more effective data management processes for reliable data collection, storage, and transformation as needed to make it ready for experimentation. The United States and China realize the significance of this distinction and have made policy advances to enable such an effort, yet both struggle with their own challenges and have not established such an infrastructure.

It is important to note that even with a robust data management infrastructure in place, several other factors can influence data advantage. For instance, early clarity on AI applications of strategic interest and project-specific work toward them can be a key factor in bolstering military data competitiveness for certain applications. Another crucial differentiator can be a country's military operational experience that provides unique opportunities in terms of real-world operational data collection, model testing, and discovering what kinds of AI applications can be most useful.

In theory, the United States is a clear leader on the operational experience front with its forces and systems having undergone diverse combat experiences. How much this experience matters rests on, among other things, whether data was collected and stored at every opportunity, whether the stored data is understandable, traceable, and sharable, and whether it can be made interoperable across DOD's various systems, old and new. The presence of diverse collection opportunities is an important factor to consider. However, all data—historical or current, big or small—will not mean much in the absence of a robust data management infrastructure, making it a primary element for assessing data advantages in developing AI systems of the future.

Commercial data is generally not relevant for military operational AI

Popular claims positing China as a data leader base their estimates on China's access to commercial data.[28] China's commercial market success surely has relevance to its national security given that it increases the technological capabilities available to China's military and intelligence community vis-à-vis funding, talent, and economies of scale.[29] However, the same assessment cannot be applied for data in the military domain. China's access to commercial market data is unlikely to give it a military operational advantage because data needs for military AI applications are environment-specific and, in many cases, learning derived from commercial data cannot be transferred to military applications at present.

A recent RAND report on the DOD AI posture classifies military applications of AI into three categories: 1) enterprise AI, such as AI-backed personnel management systems, 2) mission-support AI, such as Project Maven or AI-enabled logistics planning systems, and 3) operational AI, including the Skyborg wingman prototype or use-cases under the JAIC's AI for maneuver and fires.[30] AI applications in these categories vary significantly in the environment in which they will operate, the speed of information-processing, implications of failure, and the amount of available resources.

While these applications range in scope and function, the data they require is mostly not a fungible resource. This is especially true for applications in the operational AI category that will be used in in uncertain, fast-paced environments with severe consequences for failure. Data needs for operational AI are environment-specific, making data drawn from non-military environments irrelevant. To support a predictive maintenance project,[31] for example, the JAIC began gathering troves of maintenance logs for all of the H-60 series helicopters used across the Army, Navy, and Air Force. It soon realized that maintenance data for this purpose was not easily compatible across services because each operates helicopters for different tasks in different environments,[32] and AI trained on data from one service did not work for another. When the JAIC instead focused narrowly on UH-60 Black Hawk versions, it still faced significant data relevance issues due to the variety of operating environments of that particular version.[33] Results from federally-funded research and development centers indicate that the reliability of AI algorithms can vary significantly across desert, forest, ocean, and mountain environments. In the end, the JAIC decided to start by focusing on the UH-60 variant under the 160th SOAR (Special Operations Aviation Regiment) operating in desert conditions.[34]

The JAIC's experience with predictive maintenance illustrates how data cannot be treated as a substitutable resource even within the military, let alone from the commercial sector to operational defense environments. The requisite logs of the S-70 helicopter (the commercial variant of the military's H-60 platform) will differ sharply and its data may not be relevant for predicting the specific repairs needed for any of the military's helicopters. The military operating environment is characterized by frequent shocks, vibrations, dust, inclement weather, and energy constraints that commercial systems do not necessarily confront. Additionally, in evaluating the accuracy and reliability of existing training data, the parameters of the two sectors differ significantly. Whereas the amount and quality of training data providing reliable answers 80 or 90 percent of the time may be sufficient in the private sector,[35] military systems where human lives are more often at stake need better accuracy.[36]

All of this suggests that commercial data, while useful, will not be directly relevant for operational AI in military applications.[37] To create robust models in this domain, services will need to acquire task- and environment-specific data that respects DOD's performance specifications. Commercial data's usefulness will vary in the enterprise AI or mission AI categories depending on the tasks employed and the environment the applications will operate in.[38] Commercial data could indirectly benefit a country's operational AI military domain by allowing a country to mature its AI capabilities and talent in a particular application with subsequent implications for the military. For instance, researchers in the Chinese private sector working on facial recognition applications employing huge quantities of available commercial data with real-world use cases are likely to be much more experienced at developing superior military variants of the technology.

The future of AI may not be about big data

AI applications in the future may not require the same types and quantities of data used to evaluate advantages now. Two examples of research areas that could change how AI data is currently viewed include synthetic data generation and few-shot learning.[39] If these approaches prove successful and broadly applicable, then the potential strategic advantage conferred by big data for training and developing future AI applications will be significantly reduced.

Synthetic data generation

The development of machine learning AI systems typically relies on massive amounts of real-world data. Because such data is often not available or is cost/time prohibitive to assemble, researchers are testing the use of synthetic data—data manufactured artificially by a computer rather than measured or collected from real-world situations.[40] Synthetic data includes digitally created videos, images, audio, 3D environments, and more, combining techniques from gaming and movie industries (like computer-generated imagery) to create simulated environments.[41] A simple and common way of using synthetic data is to "augment" real-world data by making random changes to an existing dataset; for example, augmenting an image dataset by adding rotated and cropped copies of the images it contains, thereby multiplying the dataset's size. At the other extreme, in settings where the underlying data is to be generated from scratch, systems can be trained using purely synthetic data; for example, DeepMind's AlphaGo Zero system learned to play Go at world champion level by playing millions of games against itself, in essence training on its own data stream of Go games.

For applications where synthetic data could be generated and used for training on a broad scale, it might present a cost-effective and efficient way to train and develop AI algorithms, thus reducing the need for massive real-world data sets. However, several challenges remain in synthetic data generation. In some areas of synthetic data application, it is difficult to create high-quality synthetic data that reflects the real-world complexity contained in a large real-world dataset. Even if the generated data is excellent, it still replicates the authentic data and could miss crucial features. Nevertheless, there are instances of synthetic data propelling progress: for its autonomous vehicle testing, Waymo has used 20 million miles of data on real roads, and 10 billion miles on simulated roads as of March 2020.[42] Actual X-rays are combined with simulated X-rays to train AI algorithms to identify medical conditions.[43] Moreover, recent research demonstrated how some applications can achieve the same results using synthetic data as real data.[44] To the extent that militaries are able to obtain or create high-quality synthetic data, it would mitigate the need to invest in collecting and organizing big data from the real world for some AI applications.

Few-shot learning

Few-shot learning refers to the practice of training an AI algorithm to make predictions with very small amounts of data.[45] It was inspired by the idea of learning like humans. A four-year-old child needs to see photos of a zebra

only a few times for the animal to become instantly recognizable.[46] In few-shot learning, the model learns from pre-trained concepts (such as parts and relations) as prior knowledge and with a few examples of supervised information.[47] For instance, in the case of an image classification task of a rare bird species, the prior knowledge will be pre-trained models on images of other types of birds in general, and the supervised information will be a few (usually two to five) labeled images of that rare bird species.

Few-shot learning is useful in situations where data is hard or impossible to acquire due to safety, privacy, ethical, or other issues. A typical example is low data drug discovery campaigns, which attempt to discover properties of new compounds with scarce biological records.[48] Few-shot learning can reduce the need for collecting big data in highly specialized areas with sparse data points, and can enable unique breakthroughs in areas like object tracking, video event detection, language modeling, gesture recognition, and others.[49] However, few-shot learning still has limitations and can't yet be used reliably in a range of settings.[50] If few-shot learning were to prove reliable and widely applicable, it would significantly reduce the need (and therefore perceived value) of data for AI development.

One potential twist is that further advances in few-shot learning and related techniques, such as transfer learning[51] and fine-tuning, could mean that—counter to what is described above—commercial data might have some utility in military applications after all. For example, data from commercial self-driving vehicles could provide pre-training for autonomous vehicles in military settings. However, it remains unclear which data would be useful and how, so the overall point made above holds.

The future value of large datasets is uncertain

Further advancement of methods to reduce the need for large datasets, such as synthetic data and few-shot learning, could potentially affect national AI competitiveness in important ways. The availability of high-quality synthetic data generators could compensate for a country's lack of sufficient operational experience and associated data.[52] Few-shot learning could bolster progress in areas with access to few data points—for instance, in applications forecasting rare geopolitical events like nuclear weapon or missile tests. By reducing or eliminating the need for large amounts of real-world data, few-shot learning and synthetic data may allow militaries to avoid dealing with unstructured and incompatible data from legacy systems or the need for large scale data labeling. The use of such techniques would also minimize concerns regarding data ownership and data privacy.

Beyond foundational research, significant practical strides in few-shot learning and synthetic data have yet to be applied (and exist neither within the U.S. or Chinese government to the best of the authors' knowledge), despite a demonstrated interest.[53] There is still time before both sides start making effective use of them. If proven practical and applicable for military needs, these approaches would enable development of AI without reliance on massive datasets and might mean, in effect, that data is not the new oil; rather, it is no more than yesterday's whale oil.

Key takeaways

Measuring data advantage is difficult and offers no simple answers. Estimates that compare raw data broadly are oversimplifications that do not accurately reflect the role of data. This paper finds that a country with a robust metadata management capacity—or that first digitally stores, cleans, transforms, labels, and optimizes a set of data for specific projects—will be positioned to move faster toward its desired AI application, and hence at an advantage for that particular application. Commercial data, while useful, will be less relevant for military operational AI, rendering China's access to commercial market data not very pertinent for operational AI defense applications. Certain emerging approaches in AI might make big data collected from the real world less relevant for future AI development, even though the applicability of these approaches to military needs remains unclear.

Acknowledgments

The authors are grateful to Igor Mikolic-Torreira, Helen Toner, Margarita Konaev, and Lynne Weil for critical support throughout the research and writing process. We thank Robert Cardillo for providing crucial guidance and thoughtful feedback on this paper. Thanks also to Dewey Murdick, Elsa Kania, Andrew Imbrie, and Ben Murphy for their invaluable insights and comments.

We thank Paul Scharre of the Center for a New American Security and Matt Sheehan at the Paulson Institute for reviewing the report and providing excellent feedback and suggestions. Thanks to Autumn Toney for facilitating the graphical work, and Alexandra Vreeman and Matt Mahoney for their editorial support.

Document Identifier: doi: 10.51593/20190002

Endnotes

[1] David Fanning and Neil Docherty, "In the Age of AI", *Frontline* (film transcript), November 5, 2019, https://www.pbs.org/wgbh/frontline/film/in-the-age-of-ai/transcript/; Kai-Fu Lee, *AI Superpowers: China, Silicon Valley, and the New World Order* (Boston: Houghton Mifflin Harcourt, 2018).

[2] It is not clear that China actually does have more raw data than the United States, and if it does, it does not seem to be a lot more. The United States and China were estimated to produce 6.9 and 7.6 zettabytes of "raw data" in 2018. [David Reinsel, Lianfeng Wu, John F. Gantz, and John Rydning, "The China Datasphere: Primed to be the Largest Datasphere by 2025," Seagate (white paper), IDC #US44613919, January 2019, https://www.seagate.com/files/www-content/our-story/trends/files/data-age-china-idc.pdf; John F. Gantz, David Reinsel, and John Rydning, "The U.S. Datasphere: Consumers Flocking to Cloud," Seagate (white paper), IDC #44601919, January 2019, https://www.seagate.com/files/www-content/our-story/trends/files/data-age-us-idc.pdf] The United States remains the leader by a wide margin in storing raw data. [Synergy Research Group, "No Change at the Top as AWS Remains the Leading Public Cloud Provider in all Regions," *AP News*, November 20, 2018, https://apnews.com/eb1244f58464ba06193625946acefc6b] This is additionally complicated by the United States sitting on the "backbone of the internet" such that a large majority of all internet traffic passes through the United States [Elizabeth Chang, and Chris Alcantara, "Northern Virginia, center of the (data) world," *The Washington Post*, July 6, 2017, https://www.washingtonpost.com/apps/g/page/lifestyle/northern-virginia-center-of-the-data-world/2226/].

[3] Zhizheng Wang, "Systematic government access to private-sector data in China," *International Data Privacy Law*, 2, no. 4 (November 2012): 220-229, https://academic.oup.com/idpl/article/2/4/220/676863.

[4] Daniel Castro, Michael McLaughlin, and Eline Chivot, "Who is Winning the AI Race: China, the EU or the United States?" (Center for Data Innovation, August 19, 2019), https://www.datainnovation.org/2019/08/who-is-winning-the-ai-race-china-the-eu-or-the-united-states/.

[5] Matt Sheehan, "Much Ado About Data: How America and China Stack Up" (Macro Polo, July 16, 2019), https://macropolo.org/ai-data-us-china/.

[6] This step may not be necessary for a few big tech firms with established advanced data pipelines for their specific business purposes that transform the relevant data before it gets stored in a database or a data warehouse.

[7] Tamraparni Dasu, and Theodore Johnson, *Exploratory Data Mining and Data Cleaning* (New York: Wiley Interscience, 2003) pp. ix; "2016 Data Science Report" (CrowdFlower, 2016), http://www2.cs.uh.edu/~ceick/UDM/CFDS16.pdf; Steve Lohr, "For Big-Data Scientists, 'Janitor Work' is Key Hurdle to Insights," *The New York Times*, August 17, 2014, https://www.nytimes.com/2014/08/18/technology/for-big-data-scientists-hurdle-to-insights-is-janitor-work.html; Armand Ruiz, "The 80/20 data science dilemma," *The*

Cognitive Coder, InforWorld, September 26, 2017, https://www.infoworld.com/article/3228245/the-80-20-data-science-dilemma.html.

[8] "Senator Hawley Introduces Bill to Address National Security Concerns Raised by Big Tech's Partnerships with Beijing," Josh Hawley, U.S. Senator for Missouri, November 18, 2019, https://www.hawley.senate.gov/senator-hawley-introduces-bill-address-national-security-concerns-raised-big-techs-partnerships.

[9] Gregory C. Allen, "Understanding China's AI Strategy," (CNAS, February 6, 2019), https://www.cnas.org/publications/reports/understanding-chinas-ai-strategy.

[10] E. Dill, M. Uijt de Haag, "3D Multi-copter navigation and mapping using GPS, inertial, and LiDAR. Navigation" 63, no. 2 (2016): 205–220 https://doi.org/10.1002/navi.134, cited in Himansu Das, Rabindra K. Barik, Harishchandra Dubey, and Diptendu Sinha Roy, *Cloud Computing for Geospatial Big Data Analytics: Intelligent Edge, Fog and Mist Computing* (Cham: Springer International Publishing, 2018), 193.

[11] National Aeronautics and Space Administration, *About the Hubble Space Telescope* (Washington, DC: National Aeronautics and Space Administration, December 18, 2018), https://www.nasa.gov/mission_pages/hubble/story/index.html.

[12] William M. Arkin, *Unmanned: Drones, Data and the Illusion of Perfect Warfare* (Boston: Little, Brown, 2015); https://www.makeuseof.com/tag/memory-sizes-gigabytes-terabytes-petabytes/.

[13] This chart is inspired by the Data Science Hierarchy of Needs. Monica Rogati, "The AI Hierarchy of Needs," *Hackernoon*, June 12, 2017, https://hackernoon.com/the-ai-hierarchy-of-needs-18f111fcc007; Christopher Bolard, "Data Engineer VS Data Scientist," Towards Data Science, December 5, 2018, https://towardsdatascience.com/data-engineer-vs-data-scientist-bc8dab5ac124.

[14] Alex Woodie, "The Real-Time Future of Data According to Jay Kreps," *Datanami*, July 8, 2015, https://www.datanami.com/2015/07/08/the-real-time-future-of-data-according-to-jay-kreps/.

[15] Among other things, data exploration involves understanding what is in the dataset, finding characteristics of the data involving size, completeness, correctness, relationship between elements, etc.

[16] Joseph I. Naus, Data Quality Control and Editing (New York: M. Dekker, 1975).

[17] In more mature stages, an organization can build a data ETL pipeline which automatically presents data in reliable format with real-time integration.

[18] Tamraparni Dasu, and Theodore Johnson, *Exploratory Data Mining and Data Cleaning* (New York: Wiley Interscience, 2003) pp. ix

[19] NSIN & JAIC, "Into the Dataverse: A hackathon to turn maintenance actions into structured data," NSIN & The JAIC Present Into The Dataverse Hackathon, Turning Maintenance Actions into Structured Data, https://nsin-into-the-dataverse.devpost.com.

[20] This holds true assuming that the data they have is representative of the actual environment they plan to operate in.

[21] China Academy of Information and Communications Technology, *Big Data White Paper (2019)* (CAICT, China's Ministry of Industry and Information Technology, December 2019), http://www.caict.ac.cn/english/research/whitepapers/202003/P020200327550643303469.pdf.

[22] United States Government, *Federal Data Strategy: Leveraging Data as a Strategic Asset* (U.S. Federal Government, June 2019), https://strategy.data.gov.

[23] United States Government, *2020 Action Plan Progress* (U.S. Federal Government, December 2019), https://strategy.data.gov/progress/.

[24] JAIC, "A Roadmap to Getting "AI-Ready"," CHIPS: The Department of the Navy's Information Technology Magazine, June 19, 2020, https://www.doncio.navy.mil/chips/ArticleDetails.aspx?ID=13598; Jory Heckman, "DoD overcoming culture challenges to turn data 'snapshot' into predictive analytics," Federal News Network, November 21, 2019, https://federalnewsnetwork.com/big-data/2019/11/dod-overcoming-culture-challenge-to-turn-data-snapshot-into-predictive-analytics/.

[25] Chen Yu, "There are three difficulties in big data sharing: "unwilling," "dare not," "will not"" *Science and Technology Daily*, October 29, 2019, http://digitalpaper.stdaily.com/http_www.kjrb.com/kjrb/html/2019-10/29/content_433625.htm.

[26] Cao Lantian, Zhao Yinan, and Wang Xin, "Challenges facing the application of big data in army-civilian integration and their countermeasures," *Chinese Journal of Medical Library and Information Science*, April 2018, 27(4):40-43, http://cjml.ijournals.cn/ch/reader/create_pdf.aspx?file_no=20180409&flag=1&journal_id=zhyxtsqbzz&year_id=2018.

[27] Josh Mayo, "DoD CDO Describes Challenges in Pushing Data Culture," *MeriTalk*, February 27, 2019, https://www.meritalk.com/articles/dod-cdo-describes-challenges-in-pushing-data-culture/.

[28] Jeffrey Ding, "Deciphering China's AI Dream" (Future of Humanity Institute, March 2018) https://www.fhi.ox.ac.uk/wp-content/uploads/Deciphering_Chinas_AI-Dream.pdf ; [civil-military fusion papers]

[29] Gregory C. Allen, "Understanding China's AI Strategy," (CNAS, February 6, 2019) https://www.cnas.org/publications/reports/understanding-chinas-ai-strategy.

[30] Tarraf, Danielle C., William Shelton, Edward Parker, Brien Alkire, Diana Gehlhaus Carew, Justin Grana, Alexis Levedahl, Jasmin Leveille, Jared Mondschein, James Ryseff, Ali Wyne, Dan Elinoff, Edward Geist, Benjamin N. Harris, Eric Hui, Cedric Kenney, Sydne Newberry, Chandler Sachs, Peter Schirmer, Danielle Schlang, Victoria M. Smith, Abbie Tingstad, Padmaja Vedula, and Kristin Warren, "The Department of Defense Posture for Artificial Intelligence" (RAND Corporation, 2019), https://www.rand.org/pubs/research_reports/RR4229.html; U.S. Department of Defense, "Transcript: Lt. Gen. Jack Shanahan Media Briefing on A.I.-Related Initiatives within the Department of Defense," Office of the Department of Defense Chief Information Officer, August 30, 2019, https://www.defense.gov/Newsroom/Transcripts/Transcript/Article/1949362/lt-gen-jack-shanahan-media-briefing-on-ai-related-initiatives-within-the-depart/.

[31] Predictive maintenance of military platforms (like the H-60 helicopter in the JAIC's case) is likely to fall under both operational AI and mission-support AI depending upon its usage.

[32] For example, the U.S. Navy uses the SH-60 Seahawk as a multi-mission maritime helicopter – the SH-60F Oceanhawk for anti-submarine warfare, and the MH-60S Knighthawk for troop transport and vertical replenishment. The U.S. Air Force uses the HH-60G Pave Hawk for combat search and rescue. The U.S. Marine Corp has used the VH-60N White Hawk as Presidential and VIP transport helicopter.

[33] Elias Groll, "The Pentagon's AI Chief Prepares for Battle," *Wired*, December 18, 2019, https://www.wired.com/story/pentagon-ai-chief-prepares-for-battle/.

[34] Adam Stone, "The Pentagon's top AI official explains 'computer vision'," *C4ISRNET*, September 13, 2019, https://www.c4isrnet.com/thought-leadership/2019/09/13/the-pentagons-top-ai-official-explains-computer-vision/; Sydney J. Freedberg Jr, "Fix it before it breaks: SOCOM, JAIC Pioneer Predictive Maintenance AI," *Breaking Defense*, February 19, 2019, https://breakingdefense.com/2019/02/fix-it-before-it-breaks-socom-jaic-pioneer-predictive-maintenance-ai/.

[35] Leading-edge machine learning research in the private sector tends to focus on improving mean average precision (mAP) by increasing recall as opposed to precision; for examples, see Naigang Wang, Jungwook Choi, and Kailash Gopalakrishnan, "8-Bit Precision for Training Deep Learning Systems," *IBM Research Blog*, December 8, 2018, https://www.ibm.com/blogs/research/2018/12/8-bit-precision-training/; and Ankar Agrawal et. al, "Approximate Computing: Challenges and Opportunities," *IEEE Conference on Rebooting Computing*, October 2016, https://ieeexplore.ieee.org/document/7738674.

[36] For more information about system requirements and ruggedization, see "Test and Evaluation Management Guide," *Defense Acquisition University*, January 2005, https://www.dau.edu/guidebooks/Shared%20Documents/Test_and_Evaluation_Mgmt_Guidebook.pdf.

[37] It is important to note that as research progresses, approaches that allow us to transfer knowledge from commercial data to military applications, like transfer learning, might make it possible to tailor more private sector data for military needs. However, the presence of so

many failure modes and no reliability guarantee in transfer learning at present prevents employment of commercially trained models to military AI application, particularly those needed to pursue successful operations on the battlefield against an adversary. This is also discussed in the section on few-shot learning.

[38] Boundaries between operational, enterprise, and mission-support AI are fuzzy and depend upon how we choose to define them. For instance, depending upon whether we consider AI-enabled cyberattacks on foreign countries to be mission-support or operational AI, commercial sector data in this domain (and the military's access to it) is likely to confer an advantage in maximizing the spread or damage of a given attack.

[39] Other approaches that do not require significant data include one-shot learning, zero-shot learning, and low-shot learning.

[40] There are several variants of synthetic data tools available in the commercial market, and governments have divisions dedicated to modeling and simulation for training, testing, and processing information. The ones referred to in this section utilize cutting-edge AI methods like neural networks or deep learning (specifically Generative Adversarial Networks).

[41] Synthetic data has been used for creating interactive simulation environments for training autonomous platforms like robots, drones, and self-driving cars. Sergey I. Nikolenko, "Synthetic Data for Deep Learning," *Arxiv*, September 26, 2019, https://arxiv.org/pdf/1909.11512.pdf.

[42] Waymo, "Waymo raises first external investment round," Waymo Via Company News, March 2, 2020, https://blog.waymo.com/2020/03/waymo-raises-first-external-investment.html.

[43] Bernard Marr, "Does Synthetic Data Hold the Secret to Artificial Intelligence?," *Forbes*, November 5, 2018, https://www.forbes.com/sites/bernardmarr/2018/11/05/does-synthetic-data-hold-the-secret-to-artificial-intelligence/#18a5b7b842f8.

[44] Stefanie Koperniak, "Artificial data give the same results as real data - without compromising privacy," *MIT News*, March 3, 2017, http://news.mit.edu/2017/artificial-data-give-same-results-as-real-data-0303

[45] Few-shot learning is a type of Transfer Learning (TL). Like TL, it is most useful to have few-shot learning when you have a lot of data on the problem you're transferring from and very less data on the problem you're transferring to. Other related approaches include meta-learning, generative modeling, and embedding learning, among others.

[46] A four-year-old likely has 20,000+ hours of experiential data. The child could identify a zebra with only a few pictures also because it is relying on a lot of prior data on other related topics and transferring that knowledge.

[47] Yaqing Wang, Quanming Yao, James T. Kwok, and Lionel M. Ni, "Generalizing from a Few Examples: A Survey on Few-Shot Learning," *Arxiv*, March 29, 2020, https://arxiv.org/pdf/1904.05046.pdf.

[48] Han Altae-Tran, Bharath Ramsundar, Aneesh S. Pappu, and Vijay Pande, "Low Data Drug Discovery with One-Shot Learning," *ACS Cent. Sci.*, 3, 4, (April 3, 2017): 283-293, https://pubs.acs.org/doi/10.1021/acscentsci.6b00367.

[49] Yaqing Wang, Quanming Yao, James T. Kwok, and Lionel M. Ni, "Generalizing from a Few Examples: A Survey on Few-Shot Learning," *Arxiv*, March 29, 2020, https://arxiv.org/pdf/1904.05046.pdf.

[50] A major issue with this learning approach is that the approximated values based on so few data points are unreliable. The performance potential of few-shot learning under varying observation conditions (like scale, size, illumination, occlusion, rotation, etc.) is yet to be demonstrated.

[51] Researchers are working on a technique called Transfer Learning (TL) that may one day allow learning to be transferred between related applications. For instance, knowledge gained while learning to recognize dogs could be transferred to instead recognize cats. This is an area of ongoing research and it is uncertain if large scale transfer learning on a level useful to many military applications will be possible.

[52] For instance, the operating environment faced by U.S. forces in Iraq and Afghanistan necessitated greater experimentation with counter-IED capability in comparison to the PLA that likely has had fewer opportunities deploying it and developing systems that produce IED-relevant data. In this case, high quality synthetic data generators creating necessary IED data could prove useful for the PLA struggling to develop AI capabilities in a sector where it has had lesser operational experience.

[53] Y. Fang et al., "Few-Shot Learning for Chinese Legal Controversial Issues Classification," in IEEE Access, vol. 8, pp. 75022-75034, 2020, https://ieeexplore.ieee.org/abstract/document/9069958; National Geospatial-Intelligence Agency, "Synthetic Data for Computer Vision in Remote Sensing," U.S. Department of Defense, SBIR 2020, NGA201-001, https://www.sbir.gov/node/1654761.