Artificial Intelligence: Short History, Present Developments, and Future Outlook

Final Report

January 2019

Study Committee

Dave Martinez, Co-Lead
Nick Malyska, Co-Lead
Bill Streilein, Co-Lead
Rajmonda Caceres
William Campbell
Charlie Dagli
Vijay Gadepally
Kara Greenfield
Robert Hall
Andre King
Rich Lippmann
Benjamin Miller
Doug Reynolds
Fred Richardson
Cem Sahin
An Tran
Pierre Trepagnier
Joe Zipkin

MIT LL Review Team

Christopher Roeser, Lead
Konstantinos Hennighausen
Sanjeev Mohindra
Jason Thornton

This material is based upon work supported by the United States Air Force under Air Force Contract No. FA8702-15-D-0001. Any opinions, findings, conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the United States Air Force.

Preface

The Director's Office at MIT Lincoln Laboratory (MIT LL) requested a comprehensive study on artificial intelligence (AI) focusing on present applications and future science and technology (S&T) opportunities in the Cyber Security and Information Sciences Division (Division 5). This report elaborates on the main results from the study.

Since the AI field is evolving so rapidly, the study scope was to look at the recent past and ongoing developments to lead to a set of findings and recommendations. It was important to begin with a short AI history and a lay-of-the-land on representative developments across the Department of Defense (DoD), intelligence communities (IC), and Homeland Security. These areas are addressed in more detail within the report.

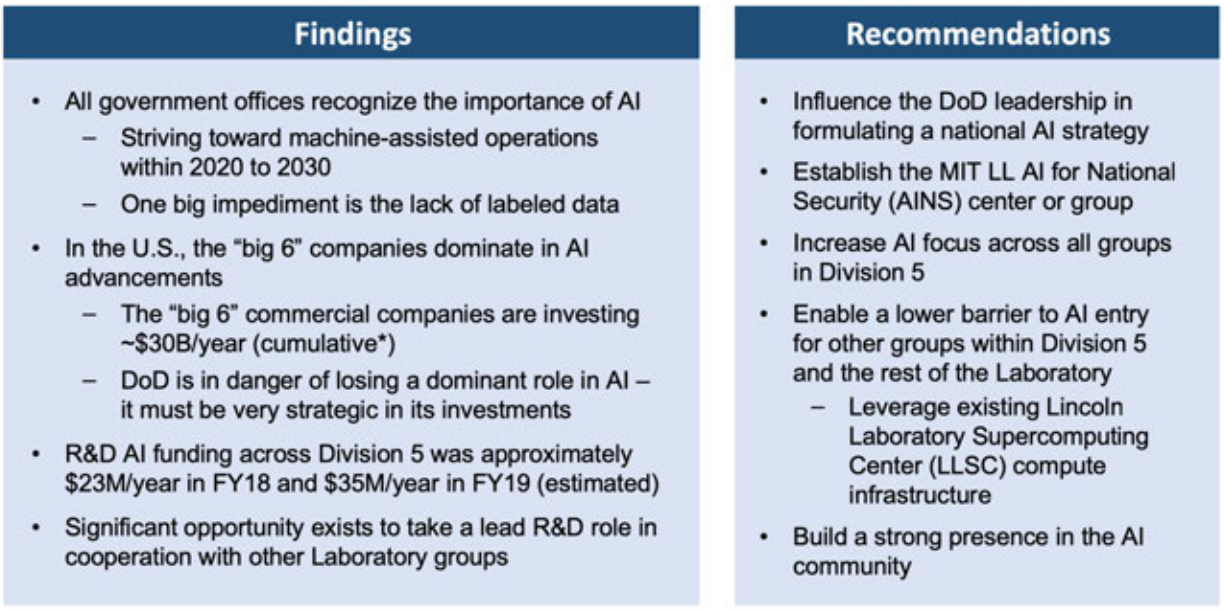
A main deliverable from the study was to formulate an end-to-end AI canonical architecture that was suitable for a range of applications. The AI canonical architecture, formulated in the study, serves as the guiding framework for all the sections in this report.

Even though the study primarily focused on cyber security and information sciences, the enabling technologies are broadly applicable to many other areas. Therefore, we dedicate a full section on enabling technologies in Section 3. The discussion on enabling technologies helps the reader clarify the distinction among AI, machine learning algorithms, and specific techniques to make an end-to-end AI system viable.

In order to understand what is the lay-of-the-land in AI, study participants performed a fairly wide reach within MIT LL and external to the Laboratory (government, commercial companies, defense industrial base, peers, academia, and AI centers).

In addition to the study participants (shown in the next section under acknowledgements), we also assembled an internal review team (IRT). The IRT was extremely helpful in providing feedback and in helping with the formulation of the study briefings, as we transitioned from data-gathering mode to the study synthesis.

The format followed throughout the study was to highlight relevant content that substantiates the study findings, and identify a set of recommendations.



An important finding is the significant AI investment by the so-called "big 6" commercial companies. These major commercial companies are Google, Amazon, Facebook, Microsoft, Apple, and IBM. They dominate in the AI ecosystem research and development (R&D) investments within the U.S. According to a recent McKinsey Global Institute report, cumulative R&D investment in AI amounts to about $30 billion per year[1]. This amount is substantially higher than the R&D investment within the DoD, IC, and Homeland Security. Therefore, the DoD will need to be very strategic about investing where needed, while at the same time leveraging the technologies already developed and available from a wide range of commercial applications.

As we will discuss in Section 1 as part of the AI history, MIT LL has been instrumental in developing advanced AI capabilities. For example, MIT LL has a long history in the development of human language technologies (HLT) by successfully applying machine learning algorithms to difficult problems in speech recognition, machine translation, and speech understanding. Section 4 elaborates on prior applications of these technologies, as well as newer applications in the context of multi-modalities (e.g., speech, text, images, and video). An end-to-end AI system is very well suited to enhancing the capabilities of human language analysis.

Section 5 discusses AI's nascent role in cyber security. There have been cases where AI has already provided important benefits. However, much more research is needed in both the application of AI to cyber security and the associated vulnerability to the so-called adversarial AI. Adversarial AI is an area very critical to the DoD, IC, and Homeland Security, where malicious adversaries can disrupt AI systems and make them untrusted in operational environments.

This report concludes with specific recommendations by formulating the way forward for Division 5 and a discussion of S&T challenges and opportunities. The S&T challenges and opportunities are centered on the key elements of the AI canonical architecture to strengthen the AI capabilities across the DoD, IC, and Homeland Security in support of national security.

[1] McKinsey Global Institute, AI The Next Digital Frontier?, June 2017

Acknowledgements

The study participants were selected from across different groups within MIT LL's Division 5. One criterion was that they needed to be practicing researchers in the AI field. This requirement was important for the study to quickly gather inputs inside Division 5 and outside MIT LL, and then formulate a set of important findings. The study participants, as researchers, had a good understanding of the key players in the AI discipline outside of MIT LL.

The AI study participants are shown here and spanned expertise in AI applied to HLT and cyber security. The MIT LL review team was chosen from researchers outside Division 5. Dr. Chris Roeser and Dr. Kosti Hennighausen came into the study with a strong background in red teaming critical technologies for national security. Dr. Sanjeev Mohindra has been working on the AI application to the intelligence, surveillance, and reconnaissance (ISR) area. Dr. Jason Thorton has shown successful use of AI in support of homeland defense.

Mr. Bob Hall, from the MIT LL Knowledge Services department, was responsible for searching AI literature, notable events, and relevant announcements. He diligently issued a weekly digest containing this information and maintained an archive of all previous literature findings. An example of this digest, which continues today, is provided in Appendix A. MIT LL's Archives team, also from the Knowledge Services Department, provided a vast number of records with information on the early history of AI at MIT LL.

The study co-leads are also very thankful to the MIT LL Director's Office for requesting this comprehensive study and their support during the course of the study. We are also thankful to all the study participants for contributing to the successful completion of the study. We are very thankful to the support personnel including Mr. Brad Dillman, Division 5 graphics artist, and Ms. Cynthia Devlin-Brooks and Ms. Kimberly Pitko for their administrative support. Finally, we are also very thankful to the Technical Communications department at MIT Lincoln Laboratory for the dedicated editorial support.

Table of Contents

Artificial Intelligence Study Motivation (D. Martinez)

> *"The greater danger for most of us lies not in setting our aim too high and falling short, but in setting our aim too low and achieving our mark."—Michelangelo*

Artificial intelligence (AI) is not a new technology. The algorithms used today have been in existence for several decades. What is new is the confluence of three key elements:

1. Advent of voluminous amounts of data. Estimates indicate that 90% of all data have been created in the past two years [1].
2. The ability to train the existing algorithms with vast amounts of data samples.
3. The use of modern computing, particularly Graph Processing Units (GPUs), that are very well matched to parallel computations. GPUs were initially developed for the gaming industry for rapid rendering of videos at low power. Researchers recognized that in many problems of interest—for example, image recognition and understanding—the same computing engine could be used for machine learning.

National security is faced with a number of challenges where AI can be instrumental, particularly in the role of augmenting human capabilities. As shown in Figure A, there is a need to employ intelligent systems and autonomy to keep U.S. armed forces out of harm's way. Similarly, the nation needs to maintain information superiority both at home and abroad. AI can accelerate the decision-making process performed by humans by leveraging machine intelligence. The human-machine teaming will result in actionable intelligence with a higher degree of confidence in environments where the consequence of an inappropriate action is high. Ultimately, the decision to take an action resides with well-trained humans in the loop.

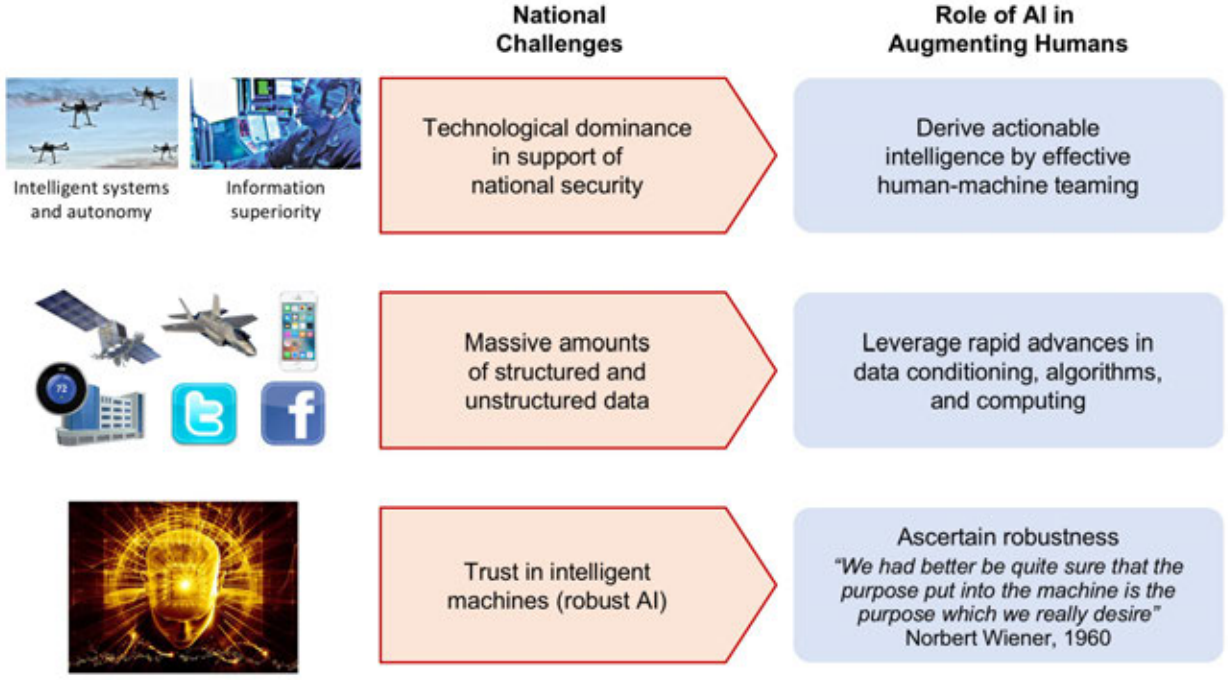


Figure A. National challenges and the role of AI.

In military and intelligence missions against radical extremists, terrorists, and peer threats, U.S. forces deal with massive amounts of both structured data (like those data provided by physical assets such as satellites, airplanes, surface platforms, and undersea platforms when supported with metadata) and unstructured data (such as data available from social media). Both of these data types must be curated through a data conditioning step before AI algorithms can ingest them. Modern computing is an enabler to rapidly process these data to reach timely decisions. Effective data conditioning of vast amounts of data, leading to the leveraging of AI algorithms, and timely processing, presents significant challenges.

Another important challenge in the proper use of AI for national security applications is trust in intelligent machines. In this report, we refer to this topic as robust AI. Any AI system employed in military and intelligence campaigns must be trusted by the operational users. Trust has many different elements, including explainability of the algorithm results, performance metrics, system validation and hardware/software verification, physical and cyber security, plus compliance with policy, ethics, and safety. All these topics are further discussed in the report.

At the start of the study, we felt it was very important to formulate an operative AI definition. Figure B depicts the AI definition used in the study. The emphasis is on *augmenting human intelligence* for important functions routinely performed by humans (such as perceiving, learning, classifying, abstracting, reasoning, and/or acting). This definition falls under the category of "narrow AI." Other researchers and practitioners refer to this definition as "specific AI" in contrast to "general AI."

There is significant angst about the advent of general AI—also referred to as artificial general intelligence (AGI). Many well-respected entrepreneurs, academics, and thought leaders have signed a manifesto on the dangers of AI on humankind [2-5]. In this report, we do not significantly address AGI. The definition we used during the course of the study is shown in Figure B. This definition focuses on *narrow AI*. The operative word in this definition is *augment* [6, 7]. We focused our study on the theory and development of computer systems that augment human intelligence. It has been very well documented that machines augmenting humans provide an immense value in accelerating the decision-making cycle.

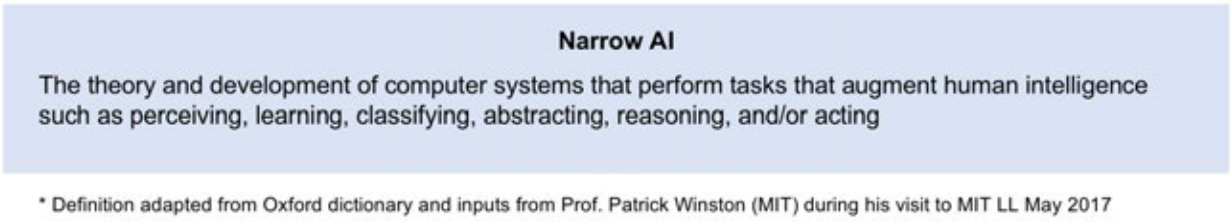


Figure B. Operative AI definition for the study.

By all accounts, AI is predicted to have a dramatic impact to all industries to the same degree that the internet revolutionizes the way we work today. As described by Thomas Malone in his recent article "How Human-Computer 'Superminds' are Redefining the Future of Work", [8] very powerful forms of collaboration will emerge by leveraging smart technologies into traditional human processes. Figure C graphs the AI domain of impact. The vertical axis is a spectrum from a limited amount of labeled data to a large amount of labeled data. The horizontal axis is a spectrum from a low consequence of action to a high consequence of action.

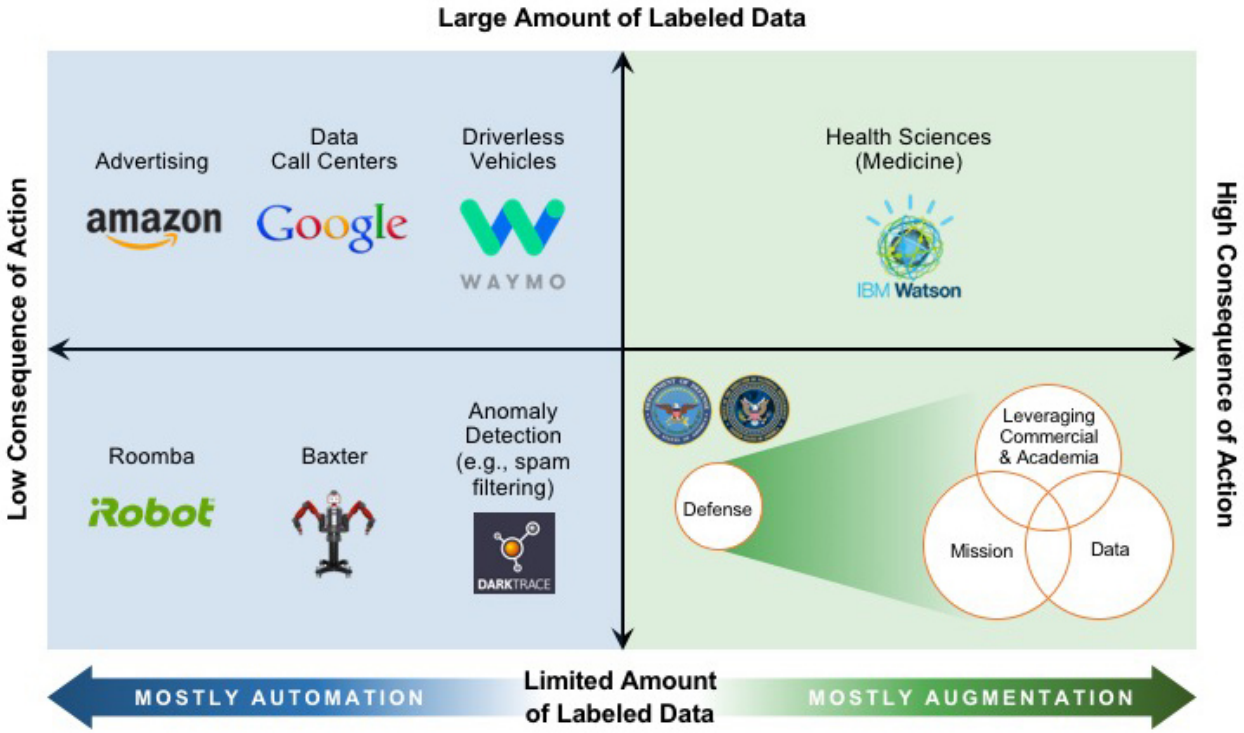


Figure C. AI domain of impact.

In Figure C, we illustrate different examples of narrow AI applications. Proceeding clockwise, the lower left-hand corner shows applications where labeled data are not significant since, for the most part, it is not needed to achieve desirable performances. In the top left-hand corner of the graph, we commonly find commercial applications, such as advertisement, data center analysis, and even driverless cars, which benefit from having abundant amounts of labeled data from social media and physical sensors. Both of these quadrants are well suited to automation—at least to date. It is true that there have been accidents with driverless cars, for example, the Uber accident in March 2018 [9]. After preliminary analysis [10], it was determined that the AI algorithms interpreted the scenario as a false positive regarding the pedestrian and the bicycle. This was caused due to a number of circumstances outside of the norm (for both the pedestrian and the vehicle). Several AI practitioners believe that self-driving cars are further away than industry admits [11]. This is one reason why, for government applications, we recommend a high level of investment in the development of *robust AI*.

On the other side of the vertical axis in Figure C, we show applications where the consequence of action is high. This means people can lose their lives. On the top right-hand corner, a very good representation is the application of AI to health sciences, as IBM Watson is doing with many hospitals [12]. In government applications—for example, in the DoD and IC—lives are at stake if the AI system recommends the wrong courses of action (CoAs) with a high degree of confidence. It is for this reason that many professionals in these communities believe that AI needs to be augmenting humans and humans are then responsible for making the final decision, as we do today. Therefore, these types of applications are better suited to augmentation of human decision making.

In national security applications, AI will have a profound effect. By AI augmenting human capabilities, routine tasks can be accelerated, leading to decisions and ultimately CoAs at speeds vastly faster than the speeds of the traditional warfare [13].

This study report was motivated by the need to identify areas where Division 5 can more significantly apply narrow AI to important national problems. The report begins with a short AI history where we address the changes that have happened in the past almost 70 years. We then present the lay-of-the-land across the commercial, academia, and the national security sectors. On the section on emerging technologies, we devote significant emphasis to the advances in key areas critical to a successful operational deployment of end-to-end AI systems. We address two application areas that are central to Division 5:

1. AI applied to human language technology
2. AI applied to cyber security

We conclude the study report with a section on future outlook. There we formulate a set of recommendations for areas the government should invest in for science and technology (S&T). The S&T recommendations are divided into three horizons:

1. Horizon 1 spans the next two years and focuses on using AI systems to deliver capabilities that augment humans by providing content-based insight.
2. Horizon 2 spans the subsequent three to four years and focuses on using AI systems to deliver capabilities that augment humans by providing more effective collaboration-based insight.
3. Horizon 3 looks at five years and beyond, and focuses on AI systems delivering capabilities that augment humans by providing context-based insight.

We identified our recommendations across the three horizons using the AI canonical architecture developed as part of the study. We also highlight the need to recognize that effective application of AI systems requires more than advances in S&T capabilities. The U.S. government must also consider improvements in existing processes to enable effective use of AI capabilities, and establish an environment of constant training of the military and civilian workforce. Toward this end, we propose an AI business model that facilitates rapid prototyping and insertion into operational systems with users intimately involved from the start.

This report provides readers with the following main contributions:

1. Increased clarity on what is the state of narrow AI relevant to national security problems
2. A description of AI enabling technologies
3. Areas for advancing AI applied to human language technology (HLT)
4. Areas for advancing AI applied to cyber security
5. Specific S&T recommendations anchored on an AI canonical architecture

References

1. Thomson Reuters, *AI Predictions.* 2018. https://www.thomsonreuters.com/en/reports/2018-ai-predictions.html
2. Bostrom, N., *Superintelligence*. 2017: Dunod.
3. Dormehl, L., *Thinking Machines: The Quest for Artificial Intelligence--and where It's Taking Us Next*. 2017: Penguin.
4. Tegmark, M., *Life 3.0: Being human in the age of artificial intelligence*. 2017: Knopf.
5. Yonck, R., *Heart of the machine: Our future in a world of artificial emotional intelligence*. 2017: Skyhorse Publishing, Inc.
6. Scoble, R. and S. Israel, *The fourth transformation.* How Augmented Reality and Artificial Intelligence Change Everything, Patrick Brewster, 2017.
7. Sullivan, J. and A. Zutavern, *The Mathematical Corporation: Where Machine Intelligence and Human Ingenuity Achieve the Impossible*. 2017: Hachette UK.
8. Malone, T.W., *How Human-Computer'Superminds' Are Redefining the Future of Work.* MIT Sloan Management Review, 2018. **59**(4): p. 34-41.

9. New York Times, *Self-Driving Uber Car Kills Pedestrian in Arizona, Where Robots Roam.* 2018. https://www.nytimes.com/2018/03/19/technology/uber-driverless-fatality.html
10. NTSB. *Preliminary Report Released for Crash Involving Pedestrian, Uber Technologies, Inc., Test Vehicle*. 2018 5/24/2018; https://www.ntsb.gov/news/press-releases/Pages/NR20180524.aspx.
11. Brandom, R. *Self-driving cars are headed toward an AI roadblock*. 2018; https://www.theverge.com/2018/7/3/17530232/self-driving-ai-winter-full-autonomy-waymo-tesla-uber.
12. IBM. *IBM Watson Health*. 2018; https://www.ibm.com/watson/health/.
13. West, D.M. and J.R. Allen. *How artificial intelligence is transforming the world*. 2018; https://www.brookings.edu/research/how-artificial-intelligence-is-transforming-the-world/.

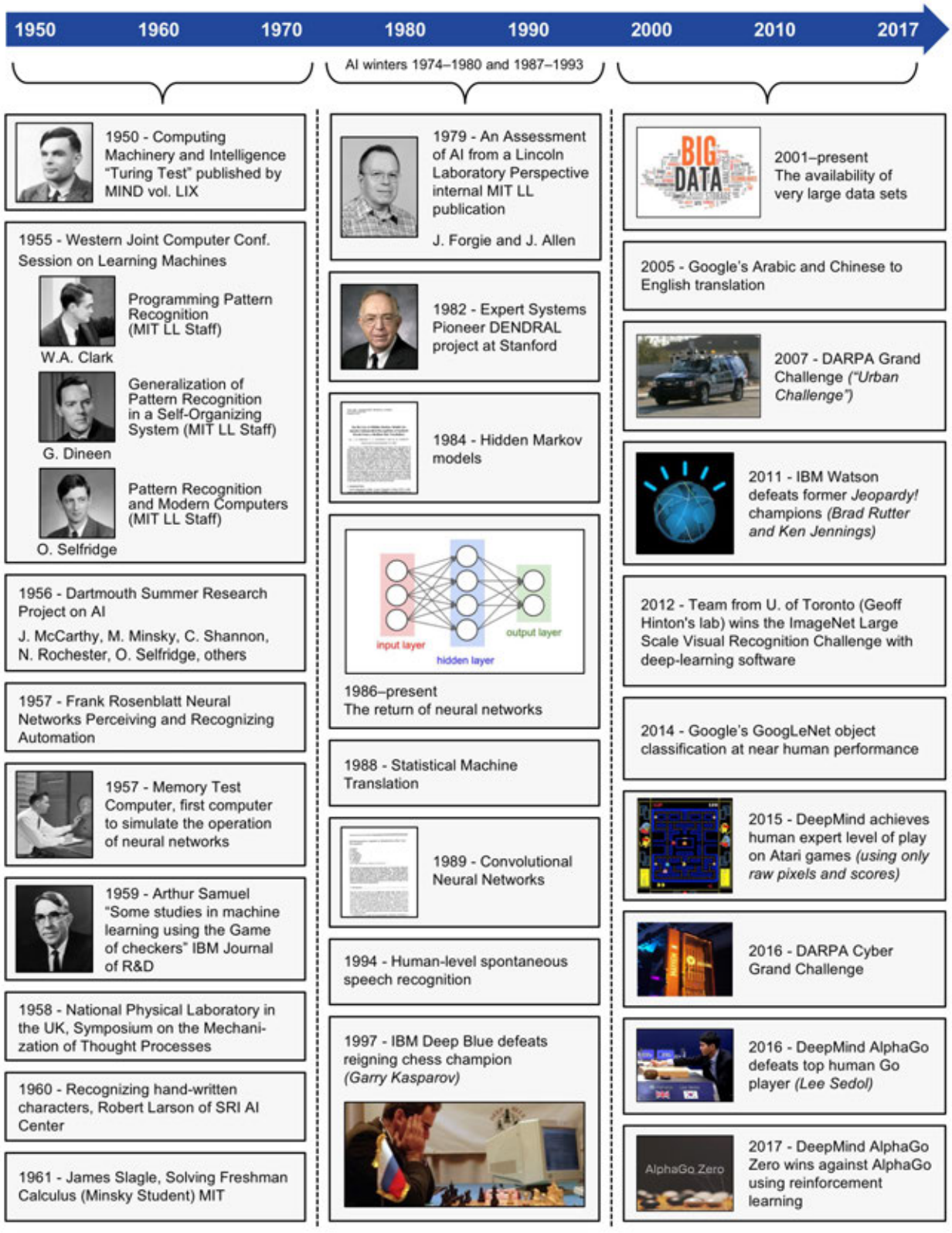
1 History of Artificial Intelligence and Trends (D. Martinez)

1.1 Notable Events in AI During the Last Seven Decades

AI developments, in the last 70 years, have shown many successes but also failures. As shown in Figure 1.1, one of the seminal papers was published by Alan Turing in 1950 and was titled: "Computing Machinery and Intelligence" [1]. Turing's main discussion centered on the key question: Can machines think? This led to the famous Turing test where a machine and a human are compared to determine—blindly—if a machine is unrecognizable from a human [2].

During this time at MIT LL, there were several researchers who were experimenting with pattern recognition techniques and their implementation in modern computers. This research work was published in the Western Joint Computer Conference under the session on Learning Machines [3]. The participants at this conference from MIT LL at the time were Wes Clark (from Group 63; Digital Development Group), Gerald Dineen (from Group 24; Data Processing Group), and Oliver Selfridge (from Group 34, Communication Techniques Group) [4].

The Dartmouth summer research project is considered the dawn of AI [5]. The actual summer gathering took place in 1956 after funding for 10 researchers was provided by the Rockefeller Foundation. John McCarthy is credited with coining the name "artificial intelligence". As shown in Figure 1.2, in addition to the AI giants who submitted the project proposal, there were others who participated in this important summer project, including Oliver Selfridge. Selfridge was Marvin Minsky's supervisor during the early years of Minsky's career. During the late 1950s and early 1960s, while at MIT LL, Minsky wrote the paper titled "Steps Towards Artificial Intelligence", which established an initial vision leading to the famous MIT AI Laboratory—where Minsky served as its first director [6].

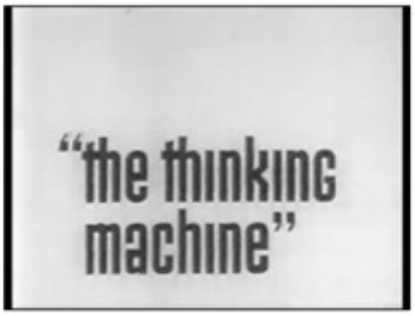


Adapted from: *The Quest for Artificial Intelligence*, Nils J. Nilsson, 2010 and MIT Lincoln Laboratory Library and Archives

Figure 1.1. Important AI milestones from 1950 to 2017.

Marvin Minsky (center) with Claude Shannon, Oliver Selfridge, and other scientists attending the Dartmouth Summer Research Project on Artificial Intelligence

Steps Toward Artificial Intelligence



Dr. Jerome Wiesner (MIT President), Prof. Claude Shannon, and Dr. Oliver Selfridge (Group Leader at MIT LL)

The Thinking Machine (Artificial Intelligence in the 1960s)

Source: YouTube, posted by Roberto Pieraccini

Marvin Minsky, "One of the fathers of artificial intelligence" worked at MIT LL 1956–1960

Figure 1.2. The Dartmouth summer research project on AI (1956) and Prof. Marvin Minsky early days at MIT LL.

In the 1950s and 1960s, there was significant activity in AI. Frank Rosenblatt, faculty at Cornell University and lead of the cognitive systems effort, published one of the first papers on the perceptron [7]. The perceptron is a simple, one-stage predecessor to what we now know as a neural network. Marvin Minsky and Seymour Papert wrote a book titled *Perceptrons: An Introduction to Computational Geometry*, where they elaborated on the mathematics of perceptrons as one of the first illustrations of a machine that could be taught to perform simple tasks by using training data as examples [8]. Perceptrons are considered an example of models within the rubric of "connectionists." Another AI model in the 1960s began to show impressive results leveraging serial reasoning of symbolic expressions—these techniques belong under the rubric of "symbolists". Minsky and Papert, in their later edition of their book [8], clarified that both connectionist learning (like perceptrons) and symbolists reasoning are important techniques within the scope of machine intelligence. In 1988, they emphasized that connectionist approaches would flourish ("…and we expect the future of network-based learning machines to be rich beyond imagining").

Another important AI milestone was the demonstration of an intelligent machine that could play checkers against a human [9]. In 1959, Arthur Samuel, at the time at IBM, showed that a machine can be programmed to play better than the human who programmed the machine. Samuel was one of the first AI researchers to introduce the term "machine learning." These were the early days of machine learning mostly built on rule-based decision trees. Although simple in comparison to today's standards, this demonstration provided an initial indication that machines could be built to exhibit a capability to learn.

Despite all the initial successes achieved during the 1950s and 1960s, in the late 1970s and again in the late 1980s, as shown in Figure 1.1, there were two so called "AI winters", when R&D funding came to a halt. These AI winters, described in Figure 1.3, were driven by the hope of achieving artificial general intelligence (AGI), mentioned earlier in this report. Although R&D funding came to almost a complete halt during those periods, there were significant accomplishments achieved, primarily based on expert systems [10]. Ed Feigenbaum, faculty at

Stanford University, demonstrated a functional expert system—DENDRAL project—applied to organic chemistry to help with the identification of organic molecules from their spectra. Although the project began in the mid-1960s, by the 1980s there were impressive results of using an expert system based on a rule-based decision process. Many more academic papers were published during the 1980s leveraging expert systems. David Martinez published a paper titled: "Systems Analysis Techniques for the Implementation of Expert Systems" [11]. The paper addressed the application of system analysis tools for designing knowledge-based expert systems; the tools were illustrated with a simplified example drawn from the oil and gas exploration application. A rigorous system analysis approach will continue to be important in many present and future applications to ascertain the robustness of AI, a concept discussed further in Section 3.

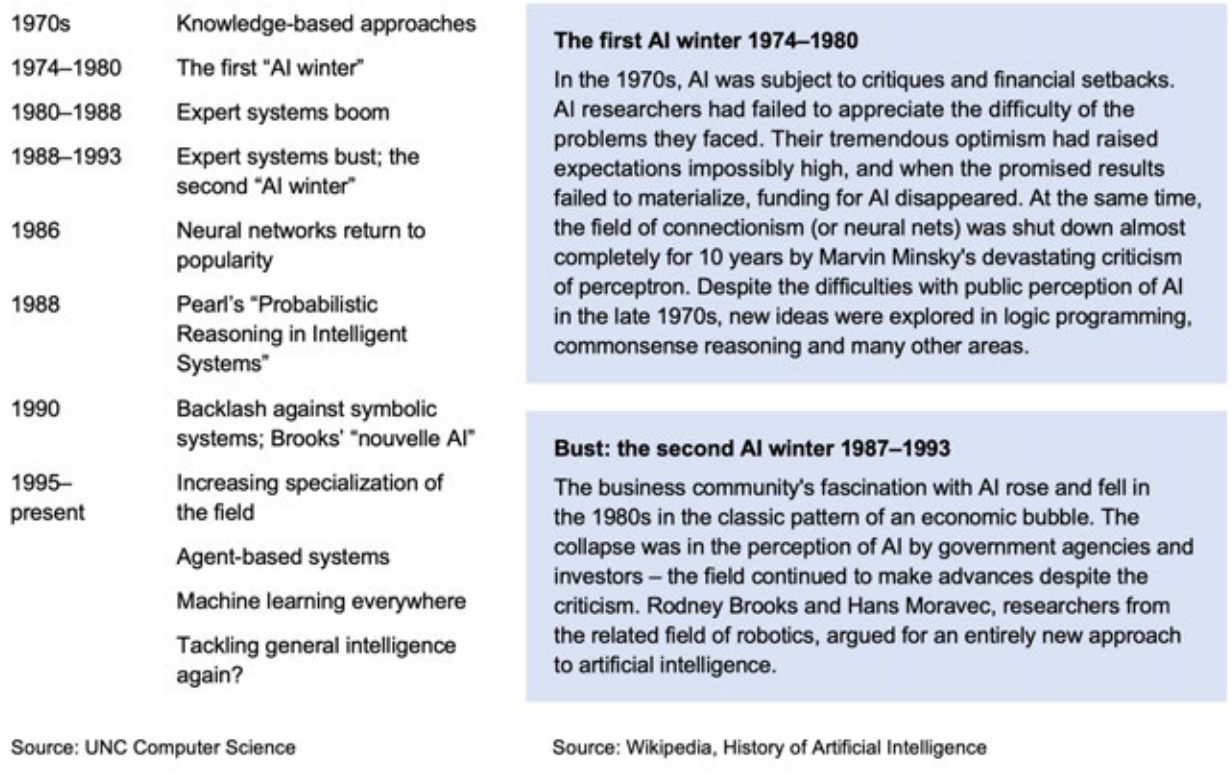


Figure 1.3. AI winters.

In 1997, there was a fundamental shift in recognizing what AI could do when IBM Deep Blue defeated reigning chess champion Gary Kasparov [12]. The chess playing program was written in the C programming language. It was capable of evaluating 200 million positions per second. In June 1997, Deep Blue was the 259th most powerful supercomputer in the world according to the well-known LINPACK benchmark used for evaluating the TOP500 list (delivering 11.38 billion floating point operations/sec) [13]. Kasparov points out that it was the ability of a computer to evaluate those 200 million positions per second that caused him to lose to IBM Deep Blue. In addition to the incredible demonstration of a machine defeating a human at as difficult a game as chess, which was revolutionary, it was also very important for the field of AI to simultaneously leverage powerful AI algorithms with a powerful computing platform, resulting in a major milestone in AI.

Since the 2000s, there has been a significant acceleration in AI milestones, as shown in Figure 1.1. In addition to advances in AI algorithms, the availability of so-called "big data" and high-performance computing have led to many important AI accomplishments in a relatively short time. The 2007 DARPA Grand Challenge demonstrated the ability of autonomous cars to navigate in an urban environment. In classical DARPA fashion, this successful grand challenge demonstration spun off a whole industry and the interest in driverless cars that we are experiencing today. We in the S&T community need more investments in AI grand challenges modeled after DARPA's successful demonstrations.

Another important AI milestone, shown in Figure 1.1, was IBM Watson defeating former Jeopardy champions Brad Rutter and Ken Jennings. This demonstration was impressive because, in contrast to the early IBM Deep Blue defeating the world chess champion through analysis of massive combinatorial chess moves, the challenge for the IBM machine was to search a massive database in real-time to find the correct question to the answer.

In 2016, the company DeepMind Technologies Limited (acquired by Google in 2014 [14]) demonstrated the ability for a machine to defeat the top Go player, Lee Sedol. This major AI accomplishment integrated advances in deep neural networks through supervised reinforcement learning trained from many examples of human plays. The AI system was named AlphaGo and consisted of more than 1200 CPUs in a distributed processing architecture and 48 Tensor Processing Units (TPUs). As described in the *Nature* paper [15], DeepMind introduced a new AI system called AlphaGo Zero with the ability to defeat the previous system, AlphaGo, by self-play reinforcement learning. AlphaGo Zero was also remarkable because instead of depending on a large number of CPUs and TPUs in a distributed computing architecture, it used a single machine with four TPUs.

Many of the breakthroughs in AI in the past several decades were based on algorithms that existed many years before the AI achievement was demonstrated. As shown in Figure 1.4, from the time an AI algorithm was first proposed to the time a breakthrough happened, on average, was 18 years. In addition to advances in computing, another important factor was the availability of relevant datasets. It took approximately three years from the time the datasets were available for building models and cross-validating the algorithms to the time the AI breakthrough happened. These gaps did not mean that algorithms were not advanced in the intervening years. In most cases, the algorithms were adapted to meet the need of the application [16].

Year	Breakthroughs in AI	Datasets (First Available)	Algorithms (First Proposed)
1994	Human-level spontaneous speech recognition	Spoken Wall Street Journal articles and other texts (1991)	Hidden Markov Model (1984)
1997	IBM Deep Blue defeated Garry Kasparov	700,000 Grandmaster chess games, aka "The Extended Book" (1991)	Negascout planning algorithm (1983)
2005	Google's Arabic- and Chinese-to-English translation	1.8 trillion tokens from Google Web and News pages (collected in 2005)	Statistical machine translation algorithm (1988)
2011	IBM Watson became the world Jeopardy! champion	8.6 million documents from Wikipedia, Wiktionary, Wikiquote, and Project Gutenberg (updated in 2010)	Mixture-of-Experts algorithm (1991)
2014	Google's GoogLeNet object classification at near-human performance	ImageNet corpus of 1.5 million labeled images and 1,000 object categories (2010)	Convolution neural network algorithm (1989)
2015	Google's DeepMind achieved human parity in playing 29 Atari games by learning general control from video	Arcade Learning Environment dataset of over 50 Atari games (2013)	Q-learning algorithm (1992)
Average No. of Years to Breakthrough:		**3 years**	**18 years**

Source: Train AI 2017, https://www.crowdflower.com/train-ai/

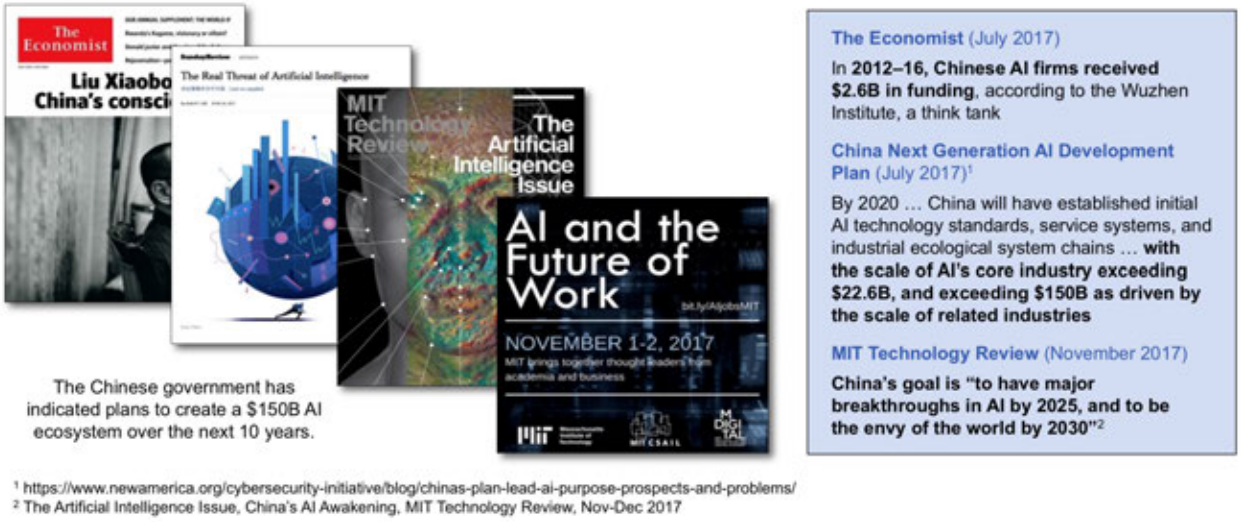
Figure 1.4. Representative breakthroughs in AI.

These notable AI accomplishments of the last few decades illustrate the potential future for AI applicable to many different classes of user needs. The major drivers are the availability of a variety of sensors and sources of data, rapid advances in algorithm models, and significant improvement in modern computing. This trend will continue as Internet of things (IoT) devices become prolific as another source of data, algorithms and simulation models get advanced, and computing continues to accelerate. We elaborate on these enabling technologies later in the report in Section 3.

1.2 AI Global Trends

Since AI can have major economic and national security implications, there is a fierce competition worldwide to dominate in AI advances [17]. As pointed out by Michael Horowitz, from the Center for a New American Security, in the national security sector, the range of applications span defending our military forces, providing timely intelligence, Homeland Security, diplomacy, and humanitarian missions. Unfortunately, AI can also be used by adversaries to inflict damage to our citizens and military forces.

As shown in Figure 1.5, China has published its short-term and longer-term plans for promoting developments in AI [18-20]. For both gaining economic advantage and for its national security strategic vision, China wants to demonstrate major breakthroughs in AI by 2025, and it wants to be the envy of the world by 2030 [21]. Toward this vision, the Chinese government has indicated its plan to create an ecosystem by 2030, together with small and large commercial companies, approximating an investment value of $150 billion. This level of ecosystem investment in AI can have a concerning implication to the United States both economically and in national security.



[1] https://www.newamerica.org/cybersecurity-initiative/blog/chinas-plan-lead-ai-purpose-prospects-and-problems/
[2] The Artificial Intelligence Issue, China's AI Awakening, MIT Technology Review, Nov-Dec 2017

Figure 1.5. China declares AI as a strategic priority.

Cumulative across 2011–2016, China reached the largest number of issued patents in AI, as shown in Figure 1.6 [22, 23][2]. Figure 1.6 also shows the top 15 patent holders in AI per patent assignee, cumulative over 2011–2016. IBM Corporation has been in the lead, with a total of 542 patents issued under their name. The State Grid of China achieved 283 patents issued to their name.

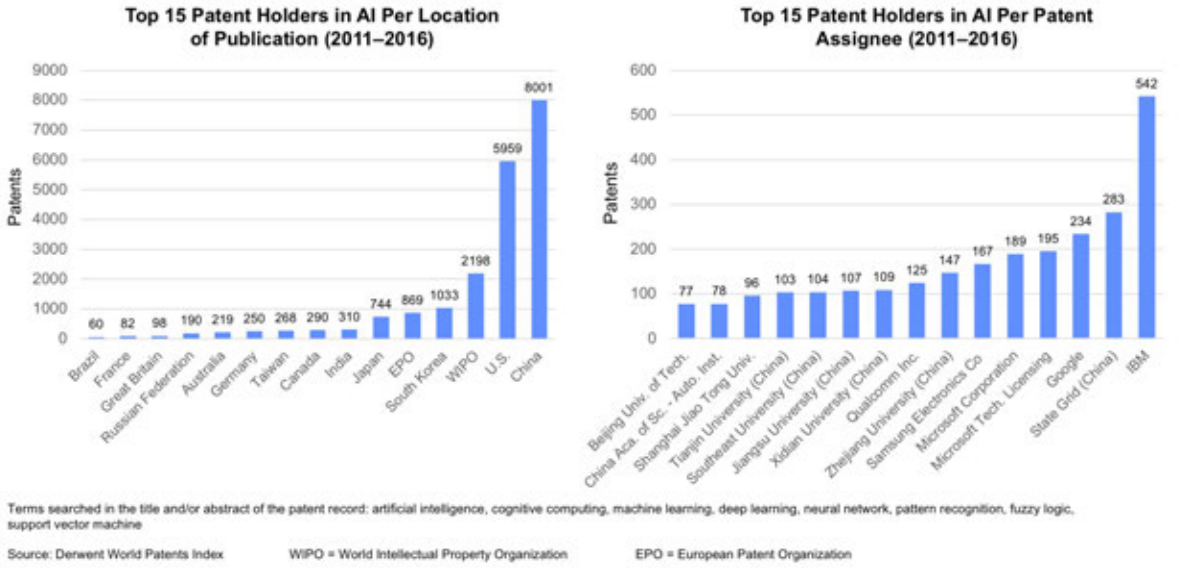


Figure 1.6. Top 15 patent holders in AI per country (2011–2016).

Many universities in the United States continue to dominate in AI research across the world. Figure 1.7 illustrates the top 15 universities/organizations from 2011–2018 [22]. CMU

[2] Source: Scopus is the largest abstract and citation database of peer-reviewed literature: scientific journals, books, and conference proceedings.

dominates in the number of published articles relating to topics within the rubric of AI[3]. MIT and Stanford are a close second and third, respectively.

Across the world, there has been significant progress made by other countries as well. As illustrated in Figure 1.8 [24], China ranked first for absolute AI citations in 2015. The United States was in the lead if self-citations were not taken into account. A better metric, as reported by G. Fabre [24], is the H-index[4], also depicted in Figure 1.8. The H-index illustrates a measure of publication influence. In 2015 the United States was in a clear lead followed by the United Kingdom. Most recently, as stated during the recent Association for the Advancement of AI conference, held in 2018, China had 290 papers accepted for publication at the conference, compared to 293 papers from the United States. There is clearly a scientific race by world powers to dominate in the field of AI.

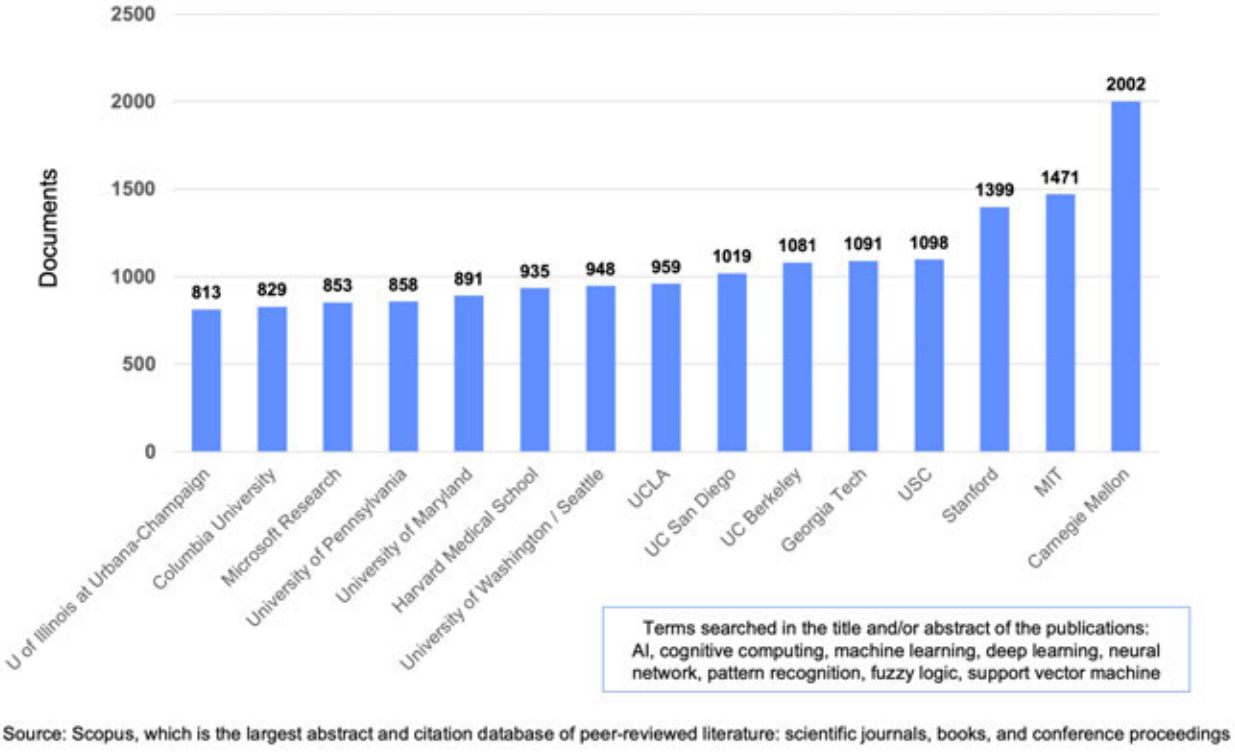
Apart from patents and publications, as discussed later in the report, there are several other factors that come into play in advancing the AI field. One important one is in the advancement of computing technologies. As of June 2018, Oak Ridge National Laboratory is officially the home of the fastest supercomputer in the world. Its computing system, named Summit, achieved 122.3 petaflops on the TOP500 benchmark [25]. The system is based on the IBM architecture using POWER9 3.07 GHz CPUs, NVIDIA Volta GV100 GPUs, and Infiniband interconnect technologies. Google also continues to advance TPUs for its datacenters.

Advanced computing, together with advances worldwide in algorithms, will enable fast acceleration of AI adaption to a large number of applications. Each application will be required to ingest the appropriate datasets that are "AI ready" for effective integration with the AI algorithms and modern computing.

Inventions in AI algorithms are originating from both commercial organizations (such as Google, Amazon, and Facebook) as well as small commercial companies. Several of these organizations are providing the algorithms to the AI community via openly accessible frameworks—for example, Google's TensorFlow and Facebook's PyTorch [26]. For the applications discussed in this report, our recommendation is to leverage all these available inventions and adapt them, via our own innovations, to our critical national security applications. As Dr. Eric Schmidt, former executive chairman of Alphabet, expressed during his testimony to the House Armed Services Committee on April 17, 2018, "success no longer goes to the country that develops a new technology first, but rather to the one that better integrates it and adapts its way of fighting."

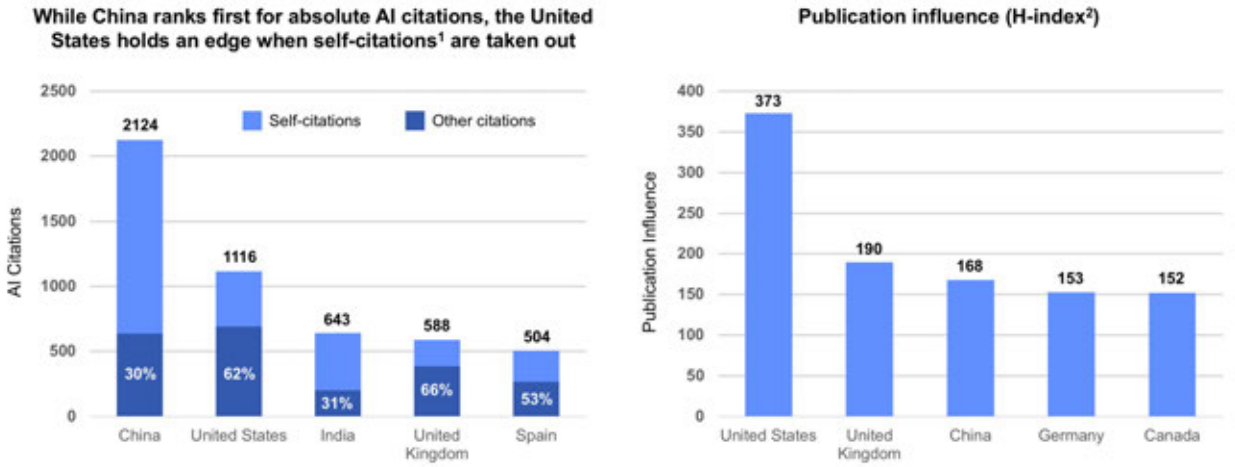
[3] Terms searched in the title and/or abstract of the publication: AI, cognitive computing, machine learning, deep learning, neural networks, pattern recognition, fuzzy logic, support vector machine.

[4] H-index, suggested by Jorge Hirsch (also known as the Hirsh number), attempts to measure productivity and citation impact of a scholar.



Source: Scopus, which is the largest abstract and citation database of peer-reviewed literature: scientific journals, books, and conference proceedings

Figure 1.7. Top 15 publishing universities in the US (2011 to 2018).



[1] Self-citation occurs when a journal cites another article published in the same journal.
[2] The H-index ranks both the productivity of scholars and the citation impact of their publications. A higher H-index number indicates more publications that are widely cited.

Source: SCImago Journal Rank 2015; McKinsey Global Institute analysis
Publication: G. Fabre, "China Digital Transformation, Why is AI a Priority for Chinese R&D?", HAL archives-ouvertes, June 2018

Figure 1.8. Top countries with widely cited AI-related papers (2015).

References

1. Turing, A.M., *Computing Machinery and Intelligence.* 1950. http://links.jstor.org/sici?sici=0026-4423%28195010%292%3A59%3A236%3C433%3ACMAI%3E2.0.CO%3B2-5
2. Toronto, U. *The Turing Test*. http://www.psych.utoronto.ca/users/reingold/courses/ai/turing.html.
3. Ware, W.H. *Introduction to session on learning machines*. in *afips*. 1899. IEEE.
4. Archives, M.L.L., *MIT Lincoln Laboratory Archives*. 2018.
5. McCarthy, J., et al., *A proposal for the dartmouth summer research project on artificial intelligence, august 31, 1955.* AI magazine, 2006. **27**(4): p. 12.
6. Minsky, M., *Steps toward artificial intelligence.* Proceedings of the IRE, 1961. **49**(1): p. 8-30.
7. Rosenblatt, F., *The perceptron, a perceiving and recognizing automaton Project Para*. 1957: Cornell Aeronautical Laboratory.
8. Minsky, M. and S.A. Papert, *Perceptrons: An introduction to computational geometry*. 2017: MIT press.
9. Samuel, A.L., *Some studies in machine learning using the game of checkers.* IBM Journal of research and development, 1959. **3**(3): p. 210-229.
10. Nilsson, N.J., *The quest for artificial intelligence*. 2009: Cambridge University Press.
11. Martinez, D.R. and M.G. Sobol, *Systems analysis techniques for the implementation of expert systems.* Information and Software Technology, 1988. **30**(2): p. 81-88.
12. Kasparov, G., *Deep Thinking: Where Machine Intelligence Ends and Human Creativity Begins*. 2017: PublicAffairs.
13. Wikipedia. *Deep Blue (chess computer)*. 2018; https://en.wikipedia.org/wiki/Deep_Blue_(chess_computer).
14. Wikipedia. *DeepMind*. 2018; https://en.wikipedia.org/wiki/DeepMind.
15. Nature, *The Go Files.* 2016(Special). https://www.nature.com/collections/hqwpvkfhrr/
16. TrainAI. *Train AI 2018*. 2018; https://www.figure-eight.com/train-ai/.
17. M.C. Horowitz, e.a., *Artificial Intelligence and Intenational Security.* 2018. https://www.cnas.org/publications/reports/artificial-intelligence-and-international-security
18. Ding, J., *Deciphering China's AI Dream.* 2018.
19. Gramham Webster, e.a., *China's Plan to 'Lead' in AI: Purpose, Prospects, and Problems.* 2017. https://www.newamerica.org/cybersecurity-initiative/blog/chinas-plan-lead-ai-purpose-prospects-and-problems/
20. Triolo, P., *Translation: Chinese Government Outlines AI Ambitions through 2020.* 2018.
21. MITTechnologyReview, *China's AI Awakening.* 2017. https://www.technologyreview.com/magazine/2017/11/
22. DerwentIndex, *Derwent Patent Index.* https://clarivate.com/products/derwent-world-patents-index/
23. Scopus. https://www.elsevier.com/solutions/scopus
24. Fabre, G., *China's Digital Transformation. Why is Artificial Intelligence a Priority for Chinese R&D?* HAL archives-ouvertes HAL ID: halshs-01818508, 2018. https://halshs.archives-ouvertes.fr/halshs-01818508v2
25. Top500, *Supercomputer Top500 List.* 2018. https://www.top500.org/list/2018/06/
26. Maladkar, K., 2018. https://analyticsindiamag.com/tensorflow-vs-pytorch-which-framework-is-better-for-implementing-deep-learning-models/

2 Lay-of-the-Land (D. Martinez)

2.1 Study Outreach

The organizations involved in the application of AI and/or developing capabilities in this field are growing very rapidly. During the course of the study, the study team interacted with a wide and diverse number of these organizations. In some instances, the interactions were in person and in other instances, it was via attendance to meetings, where several government representatives spoke about their efforts, or through emails and teleconferences.

Figure 2.1 highlights organizations from both the DoD and IC that we spoke with. These organizations are advancing capabilities ranging from R&D to operational use. Their focus in almost all cases is in augmenting human capabilities—what we defined earlier in the report as narrow AI. DARPA is looking at what it calls *Wave 3*, focusing on context and stronger human-machine symbiosis. DARPA defines *Wave 1* as knowledge-based systems typically based on a set of rules (i.e., expert systems). It defines *Wave 2* as AI systems that are based on statistical learning, as we typically find today at the S&T levels and in earlier demonstrations at the operational levels (e.g., supervised learning based on structured and unstructured data). The future is clearly in the ability to achieve context-based reasoning (*Wave 3*), and where machines are not just tools but partners with humans to achieve the desired mission.

The Under Secretary of Defense for Intelligence (USDI) has demonstrated a series of sprints through Project Maven [1] with the objective of introducing AI capabilities based on the application of deep neural networks (Wave 2). Project Maven is showing the ability to employ machine-learning techniques, in real time, to identify objects of interest that would normally take a significant amount of time by human analysts to classify. Project Maven is enabling image analysts to devote more of their time to higher cognitive tasks. It is very important to clarify that Project Maven is not trying to automate military weapons. The decision to undertake a military task is left to the military men and women responsible for deciding the courses of action.

Figure 2.1. Study outreach to government organizations.

Since the study was led out of Division 5—the Cyber Technology and Information Sciences division—the organizations we contacted were primarily working on AI for cyber security or AI for information sciences (e.g., human language technology and computing). Figure 2.2 illustrates organizations from the defense industrial base working on AI applied to national security problems, commercial sector, our peer laboratories, and AI centers.

The defense industrial base organizations were primarily advancing capabilities for their national security customers. As capability providers, their interests range from data conditioning and algorithms adaptation to decision support systems enabled by AI.

As we all know, much of the rapid acceleration in AI has been a result of significant R&D investments in the commercial sector. As shown in Figure 2.2, companies like NVIDIA are in the forefront of computing, leveraging their GPUs. IBM Watson has started, in the last few years, a whole business unit leveraging the original IBM Watson demonstrations used in the game of Jeopardy! Their applications range from addressing important societal problems to advances in computing technology [2]. Similarly, Google and Microsoft are leaders in both AI algorithms and computing technologies.

Many of our peer organizations, shown in Figure 2.2, are pursuing R&D in the application of AI. NASA Jet Propulsion Laboratory (JPL) has established a center for data science and technology to coordinate R&D for data intensive systems, methods, and technologies across the JPL organization. As we discussed earlier under the section on AI trends, Oak Ridge National Laboratory just received the TOP500 award for the fastest supercomputer in the world; in addition to nuclear weapon modeling, they are planning to leverage the immense computing for AI developments. Thus, the advances ongoing in the United States at national laboratories are accelerating rapidly, both in the application of AI to national security problems and in strengthening the talent pool.

Another important set of capabilities toward the advancement of AI is available through research centers or consortia. For example, the USC Center for AI in Society is conducting research in AI to help solve the most difficult social problems facing our world [3]. One successful application is its demonstration of game theory in critical infrastructure protection against terrorist threats. An example of a consortium is Partnership on AI, which is bringing together diverse, global voices to realize the promise of AI [4]. These are a few of the examples of advances that we can leverage in the adoption of techniques to solve challenging problems in the DoD, IC, and Homeland Security communities.

Figure 2.2. Study outreach to representative organizations from the defense industrial base, commercial, peers, and AI centers.

Shivon Zilis (advisory board member at OpenAI) published a survey of commercial organizations in the machine intelligence field [5]. Zilis points out that just in the course of one year, from 2015 to 2016, there was a growth of 30% in commercial companies that have had a significant play in AI. Figure 2.3 depicts the commercial companies broken down by different types of business sectors (the original paper appeared in the Harvard Business Review) [6]. By breaking down the commercial companies working in the machine intelligence field as shown in Figure 2.3, it gives us a very good perspective on how important it is for our applications to leverage the large investment that commercial companies have devoted to the field of AI—including open source content we can leverage.

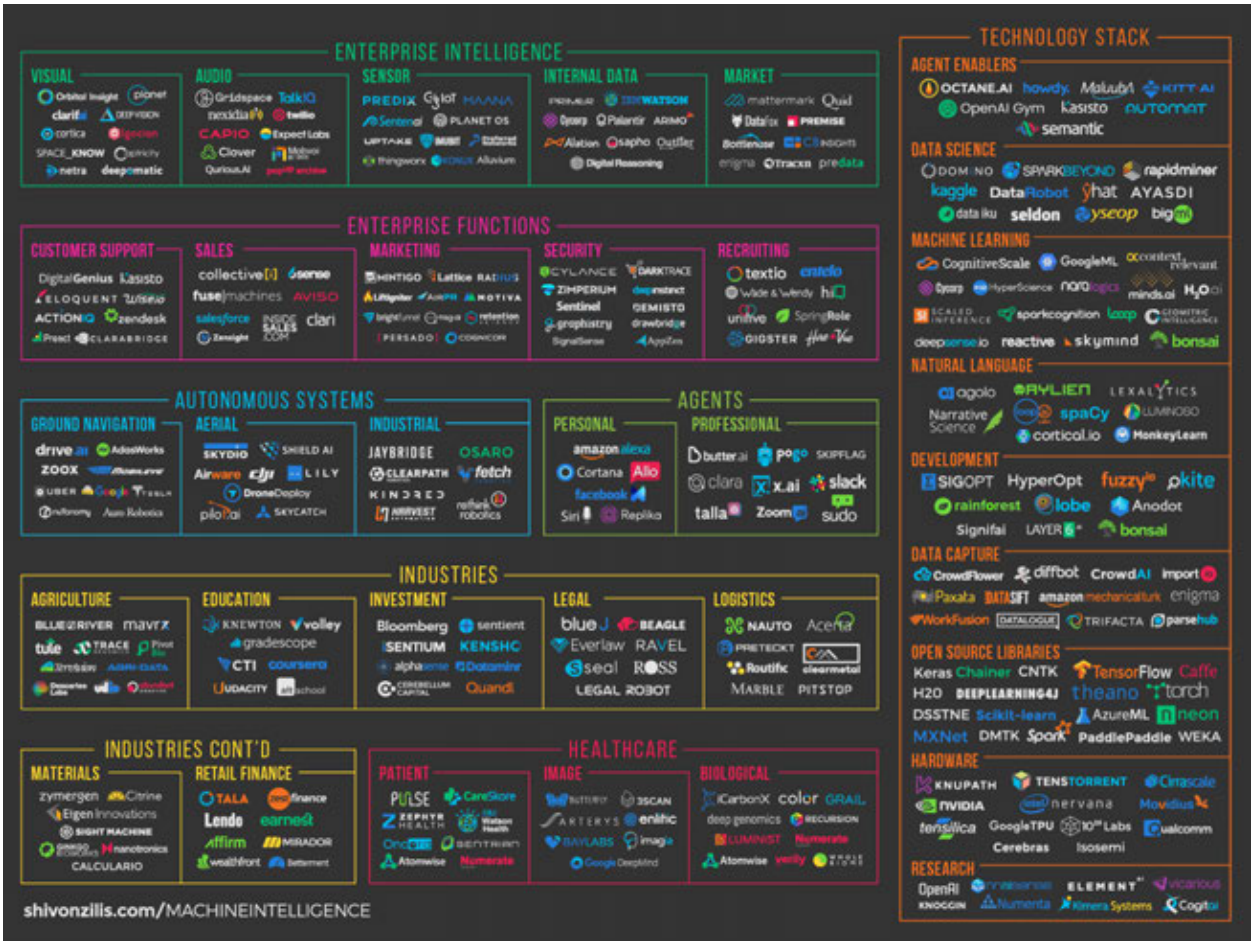


Figure 2.3. Spectrum of commercial organizations in the machine intelligence field.

The study team reached out to many scholars in academia. In Figure 2.4, we identified the principal researchers in academia that provided inputs to the study team. The breadth of research spans theoretical developments, algorithms, hardware, and applications to both cyber security and information sciences. One effort that we at MIT LL are planning to be heavily involved with is the recently announced MIT Quest for Intelligence (launched on February 1, 2018 [7]) with the objective of advancing human and machine intelligence research. MIT LL can leverage this effort by transitioning new innovations from MIT faculty and students to solve important national security problems.

Figure 2.4. Study outreach to scholars from academia.

2.2 AI Canonical Architecture

Since AI requires many different key enablers, the AI study team formulated an end-to-end AI canonical architecture. The AI canonical architecture is shown in Figure 2.5. Centering the study on an AI canonical architecture served several purposes:

1. To identify the key enablers needed to address an end-to-end AI system
2. To highlight where machine learning fits relative to the overall AI system
3. To categorize areas where the DoD should either lead, adopt/adapt, or follow relative to ongoing investments in other parts of the federal government, commercial sector, and academia
4. To formulate areas where an organization should focus its resources to rapidly advance its AI capabilities
5. To unify the enabling technologies and applications discussed in this report under one end-to-end system architecture
6. To serve as the guiding framework for organizing the study recommendations

As many AI practitioners predict, AI will continue to have a significant impact in many areas including medicine, agriculture, energy, transportation, manufacturing, financial services, human resources, logistics, national security, etc. These impacts are going to result in a change to our economic landscape, education, workforce, and global competitiveness. Therefore, there will be a need for close coupling between the AI technical community and the branches of the government establishing policies. As pointed out during a recent AI summit chaired by the Office of the President for Technology Policy [8], AI has tremendous potential to benefit the American people. Therefore, the White House has established a Select Committee on AI, under the National Science and Technology Council, to improve coordination of federal efforts and continuing U.S. leadership in AI. During this AI summit, it was pointed out that the top eight universities in AI are in the United States. It was also emphasized that the U.S. ecosystem addressing the AI challenges and developments is also very strong, with roughly three-quarters of the world's top 100 AI startups residing in the United States. The challenge to our nation will be in leveraging the commercial sector, academia, government laboratories, and the industrial base to garner all the sources of AI innovation in a coherent way.

In this AI report, we focus on the technical aspects of end-to-end AI systems in support of national security—more specifically, AI for cyber security and AI for human language

technology. However, the same AI canonical architecture and associated enabling technologies discussed in Section 3 can be used for addressing other AI application domains. We also emphasize, later in the report, the importance of sound government policies, ethics, safety, and training when we discuss robust AI as depicted in Figure 2.5.

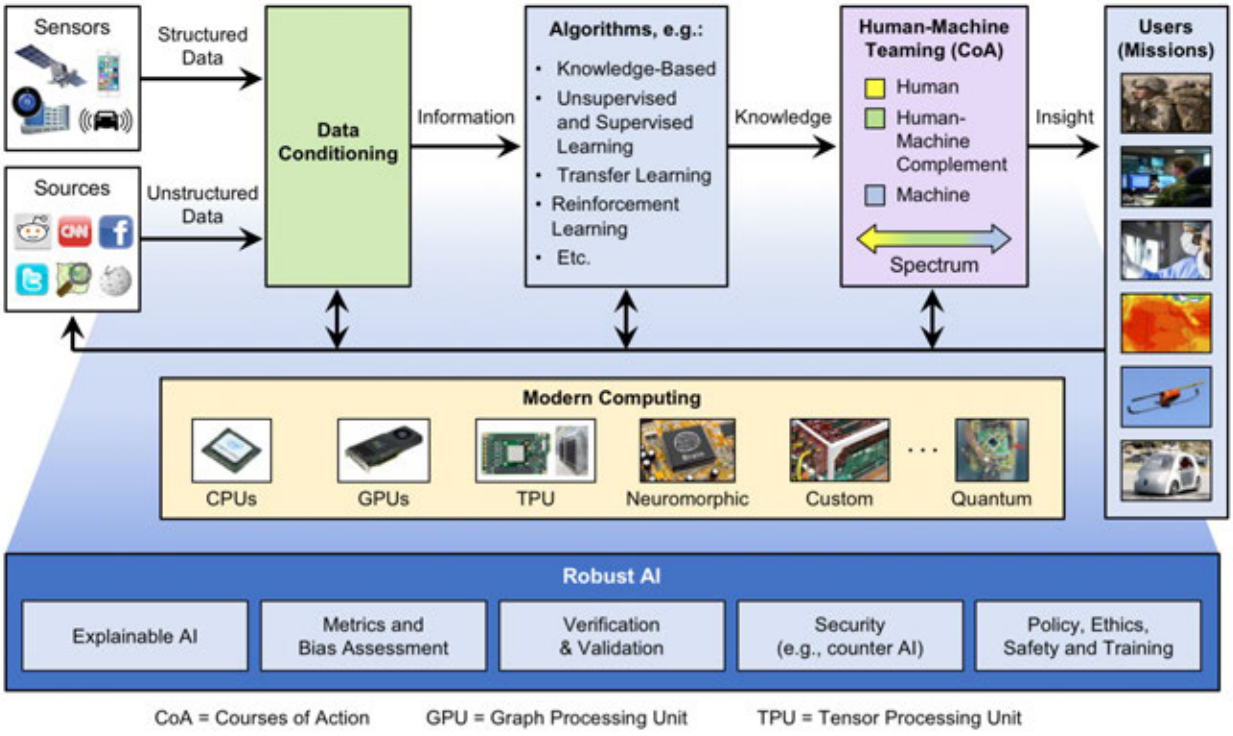


Figure 2.5. AI canonical architecture.

As discussed in the U.S. National Defense Strategy [9], the DoD will invest broadly in autonomy, AI, and machine learning, including rapid application of breakthroughs from the commercial sector, academia, and defense industrial complex, to gain a competitive military advantage. We strongly believe that AI will be a game changer, similar to other technologies, such as precision weapons, Global Positioning System (GPS), stealth, etc., developed in support of the so-called second offset strategy. However, in the case of AI, it is imperative that we think of this game-changing capability as an end-to-end system and not as a single-point solution, therefore requiring key subsystems to effectively be integrated together.

Recently, Prof. Andrew Moore, Dean of the School of Computer Science at CMU, envisioned and defined what he refers to as the AI stack [10]. As shown in Figure 2.6, Moore's AI stack shares many of the same key subsystem components of the AI canonical architecture shown in Figure 2.5. As Moore points out, "AI isn't just one thing or a single piece of software; it is a massive collection of interrelated technology blocks called the AI stack…" This perspective is very important for any organization investing in R&D and transitioning the AI capabilities into operational systems. The AI stack, similar to our AI canonical architecture, can provide a framework for identifying and organizing all of the technologies and capabilities required of an end-to-end AI system. Later in this section, we elaborate in detail on each of the subsystems shown in Figure 2.5. However, we felt it was also important to highlight key differences between the commercial sector compared to national security applications to get a better appreciation of the challenges surrounding the implementation of AI.

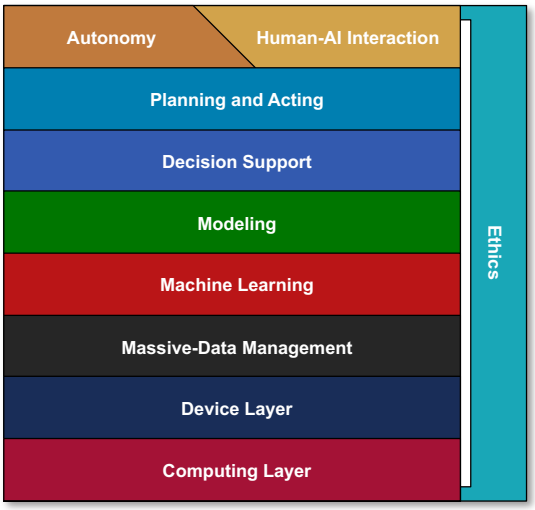


Figure 2.6. AI Stack formulated by Dean Andrew Moore of the School of Computer Science at CMU.

Many of the AI applications in the commercial sector (as it applies to Facebook, Google, Amazon, IBM Watson, Microsoft, and Apple) depend on data with high dimensionality (so called variety), as well as massive amounts of data (large volume). These are also true of applications in support of national security. As shown in Figure 2.7, there are some similarities between the commercial sector when compared to the national security domain, but there are significant differences as well. Some of the notable differences are in amounts of labeled data needed for supervised learning. Another important difference is needing to operate in a contested environment with low capacity/intermittent datalinks, reducing the access to the vast set of resources available at the enterprise level. For national security applications, we must also recognize that countermeasures will be necessary to defend against malicious uses of AI (more on this when we discuss the topic of robust AI).

However, we believe that for DoD, IC, and Homeland Security applications, the most effective way to accelerate adoption of AI capabilities is to adopt/adapt commercial techniques, since they are evolving at a very fast pace. The large number of available tools, high productivity languages, high performance databases, etc., facilitates leveraging of commercial AI advances. Some examples of available tools, discussed further in Section 3, are:

- Machine-learning frameworks: Caffe (developed at UC Berkeley), TensorFlow (from Google), PyTorch (from Facebook), Keras (open source neural network library), etc.
- High productivity languages: Matlab Simulink, Julia, Python, R, etc.
- High performance databases: SciDB, accumulo, etc.

AI prototyping and experimentation with speed and agility are also of paramount importance since AI capabilities, complementing humans in the decision cycle, can save lives when confronted with radical extremists (“finding a needle in a haystack”), terrorists networks, and defending our homeland and those abroad against peer threats. In Section 6 under Future Outlook, we recommend that the federal government implement an AI capability business model for rapid experimentation with close collaboration between researchers and users.

Commercial Sector	National Security
High dimensionality	High dimensionality
Large volume	Large volume
Mostly using enterprise cloud computing	Need for both cloud and tactical computing
Human-machine teaming is tolerant to errors	Human-machine teaming must be robust
Abundant amounts of labeled data	Limited amounts of labeled data
High capacity datalinks	Low capacity/intermittent datalinks
Competitive environment	Adversarial environment/countermeasures
Mostly consumer users	Today requires sophisticated users
Explainability is not the largest issue	Trust/explainability is core

AI will be a technological enabler (i.e., data and algorithm warfare) against:
radical extremists, terrorists, and peer nations to defend our homeland and abroad

Figure 2.7. Key similarities and differences between the commercial sector and the national security sector in the application of AI.

Another aspect of effective AI implementation to our domain is in the use of both simulated and real data. The machine-learning algorithms require a significant amount of data, at least today, to create a model through the training step. Therefore, as pointed out by Dr. Eric Schmidt during a visit in May 2017 to MIT CSAIL, we need to improve our computational resources by 100×–1000× from what we have today.

Waymo, a self-driving development company and subsidiary of Alphabet Inc., acquired 2.5 million real-world miles in several cities by operating driverless vehicles in 2016. They also modeled 1 billion virtual miles in the same year [11]. The modeling of virtual miles requires substantial amounts of compute power. This is one reason why Google developed the Tensor Processing Unit (TPU) computing hardware, which it uses in its data centers. The United States is also devoting significant federal investment to the development of exascale supercomputers (billion billion, or a quintillion, calculations per second) through the Department of Energy. These supercomputer systems are needed to enable modeling of AI algorithms for a broad range of applications.

2.3 High-Level Description of Subsystem Components in the AI Canonical Architecture

In this section, we describe, at a high-level, the key subsystem components of the AI canonical architecture shown in Figure 2.5. Section 3 elaborates in more detail on the main enabling technologies. The format we follow is to describe the flow starting from input data through the insight stage delivered to the users. An important subsystem component is what we referred to as *robust AI* shown in Figure 2.5, which is a critical area to ascertain acceptance by the DoD, IC, and Homeland Security users. If an end-to-end AI system fails to deliver trustful results with high confidence, the user will opt to revert back to the prior approaches he or she has used without the benefit of AI. This is true because we hold AI systems to a higher standard in avoiding errors compared to errors made by humans—until we get to be comfortable with the benefits provided by AI augmenting humans cognitive tasks.

In the effective application of AI, data are one of three drivers evolving this field. As discussed earlier, the other two drivers are modern computing and the algorithms. In a recent

online article, award-winning author Alexander Wissner-Gross stated very accurately: "Perhaps the most important news of our day is that datasets—not algorithms—might be the key limiting factor to development of human-level artificial intelligence" [12]. Although in this report we are not on a quest to achieve human-level artificial intelligence, as defined earlier under general AI, the statement applies equally to narrow AI, where the goal is to augment human intelligence.

We begin our high-level description of the subcomponents shown in the AI canonical architecture by providing a short description as follows.

Sensors and Sources: Data provided by physical sensors—like satellites, airplanes, submarines, and IoT devices—and cyber sources, such as different forms of social media, news articles, standard open web, plus the deep web and dark web, etc.

Often data from sensors are categorized as structured data because the raw digital data are accompanied by metadata. Metadata describes the characteristics of the incoming data; for a radar system it might be radar frequency, transmitter power, number of receiving channels, sampling frequency of an analog-to-digital converter, etc. In contrast, what we call data sources are accessible in the cyber domain and are categorized as unstructured data. There is, typically, no source of metadata contained with those data. For example, reports from a platoon in a mission are written text void of any description of what is in the report until a human reads it. AI techniques in the field of natural language processing are advancing rapidly to determine what is contained in these types of reports.

Data Conditioning: Both structured and unstructured data need to go through a data conditioning stage before algorithms can be applied to the data. Data conditioning, also referred to as data munging in the data scientist community, is quite involved. More than 90% of all data available in the world has been generated in the past two years [13]. Figure 2.8 depicts the exponential data growth in data originating from open sources. The chart illustrates the amount of open source data generated every minute of the day in 2018.

The main objective for this subcomponent is to transform data into information. An example of information is a new sensor image (after data labeling) that we need to use to classify if the object of interest is present in that image or not (like a vehicle of interest). Typical functions performed under this subcomponent are: standardization of data formats complying with a data ontology, data labeling, highlights of missing or incomplete data, errors/biases in the data, etc.

Algorithms: This subcomponent is also commonly referred to as machine learning. It is the stage where an end-to-end system is able to transform information into knowledge. Knowledge is more specific than information. For example, using the same example as used earlier in the data conditioning description, knowledge is the classification of what is in the image (a vehicle of specific make, model, and color).

There are many types of machine-learning algorithms, including unsupervised and supervised learning techniques, as shown in Figure 2.5. One of the watershed moments in AI (see Figure 1.4) happened when the AI community experienced the confluence of labeled data and the demonstration of dramatic improvement in image classification. The labeled dataset was created by Prof. Fei-Fei Li (presently at Stanford University) and students [14]. They have created to date 15 million labeled images (in 2010, they had labeled 1.5 million images) by formulating a well-defined ontology and employing Amazon's Mechanical Turk to have humans go through and label each image according to the ontology. In 2012, Geoffrey Hinton (now at University of Toronto and Google), et al., [15] used the labeled data from ImageNet and applied a deep convolutional neural networks algorithm, using multiple GPUs for training, and were able to demonstrate an error rate of 15.3% in image classification—compared to the 28% error rate achieved in 2010. The demonstration was part of the annual competition called ImageNet Large-Scale Visual Recognition Challenge (ILSVRC). In 2016, the error rate was reduced down to 3%,

compared to humans at about 5% error rate. This was an impressive demonstration of the confluence among big data, algorithms, and modern computing.

Modern Computing: This subcomponent addresses classes of modern computing suitable for the AI training and the inference stage. One of the leading processing engines is Google's TPU. It leverages variable precision to gain in performance per unit watt. Similar techniques have been used for other applications where matrix–matrix and matrix–vector multiplies are of the essence for both training and application of the weights [16]. These linear algebra operations require not just high-speed computations, but also high-performance interconnects, and access to memory.

Human-Machine Teaming: A subcomponent that allows a strong collaboration between humans and machines will be paramount to any AI system. This collaboration will achieve operational speed by providing timely insight to users, increasing scale, and reducing the level of the consequence of actions. Collaborative intelligence will permeate across many different application domains beyond national security [17-19].

Robust AI: There are many elements within the rubric of robust AI as shown in Figure 2.5. For users to trust and depend on AI, the overall system has to include the ability to explain how the machine-learning algorithms arrive at their output (information to knowledge).

Similarly, there needs to be better metrics, not just for the algorithms, but for the end-to-end AI system. Every system will need to undergo verification (verifying that the software does what it was designed to do), and validation that system performs as expected. Security, both physical and cyber, will be an important aspect for protecting the AI system. Finally, any AI capability will need to be cognizant of policy, ethics, and safety. In this context, there will need to be an active effort in training our people, both military and civilians, in the use and understanding of AI systems. As shown in Figure 2.9, AI systems will be used routinely in operation when we are able to demonstrate a high confidence level vs. an acceptable level in consequence of actions.

In Figure 2.10, we illustrate examples of AI system capabilities, either as early prototypes or use in operation. In cyber security, spam filtering routinely employs rudimentary machine-learning techniques to identify potential phishing attacks. Similarly, keyword searching, as an example of an information sciences application, is well matched to machines. These are examples of high confidence with low levels of consequence of actions. Most cases, as shown in Figure 2.10, fall in the center where machines augment humans in the decision cycle.

Users: Although this stage in the overall AI canonical architecture is not a subcomponent, per se, the users are ultimately the beneficiaries of an end-to-end AI system. Users must be able to provide feedback to the system to enable increased refinement and improvements. There are many types of users ranging from users in an enterprise environment, at the tactical edge, as well as autonomous systems. Autonomous systems will only be able to operate without the aid of a human if, and only if, the confidence in the decisions made by the machine is high and the consequence of the action is low. For most other cases, the humans are ultimately the ones making the final decisions at least as envisioned for the near future.

In the following sections, we elaborate in more detail on several important enablers in the development and advancement of AI systems. For clarity, the reader should note that in several sections we reuse or illustrate similar charts to emphasize different concepts within the AI report.

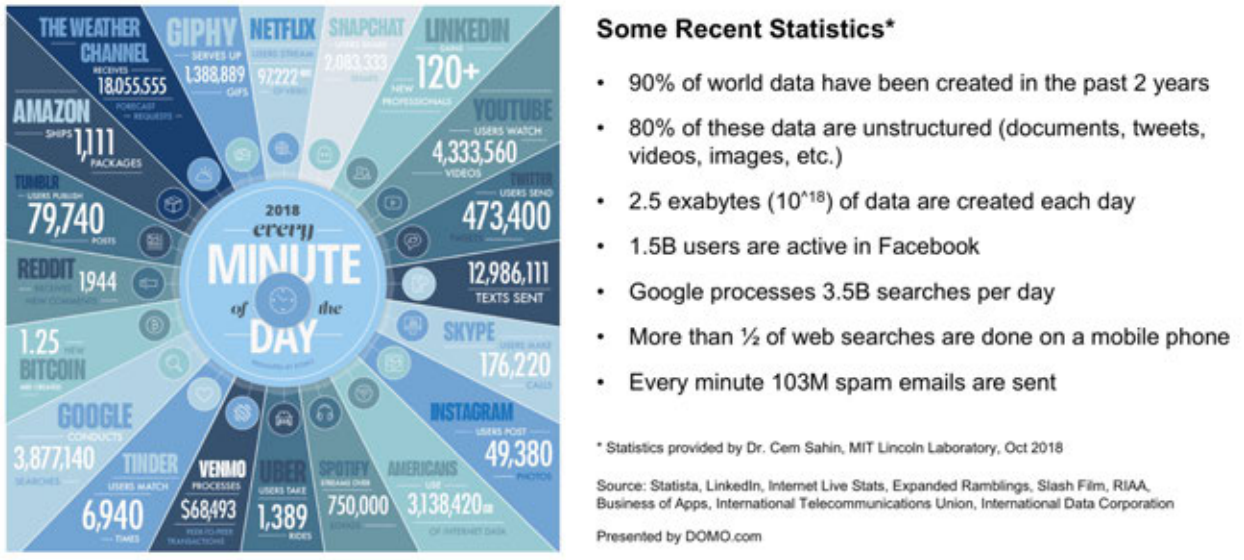


Figure 2.8. Exponential growth in open sources of data. Estimates are that the world generates 2.5 quintillion bytes of data per day.

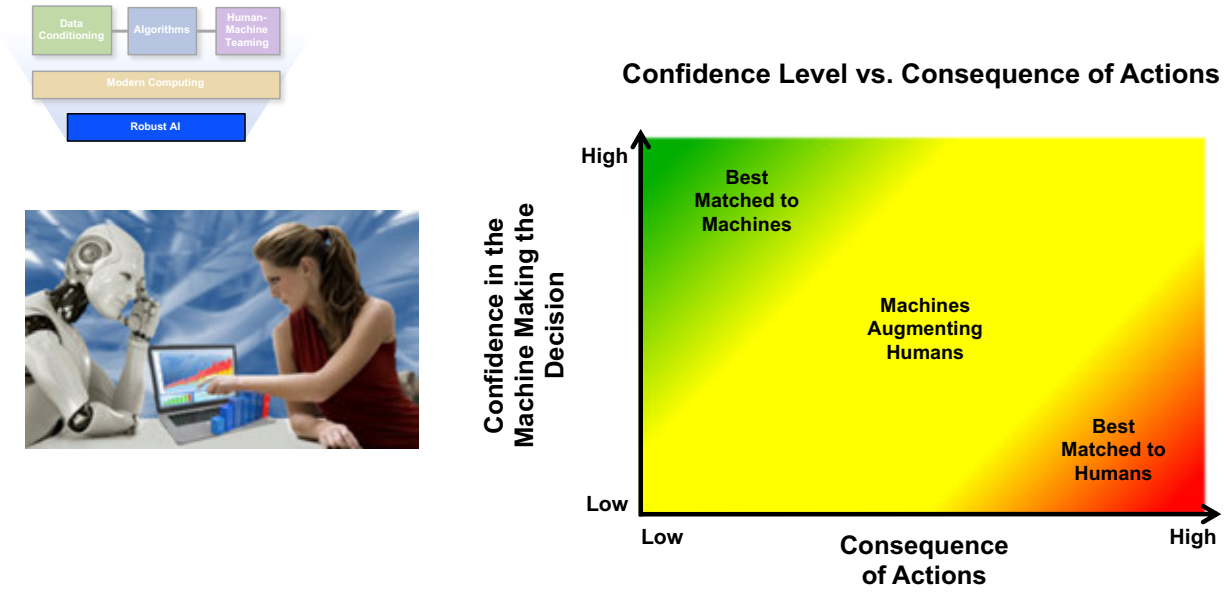


Figure 2.9. Robust AI: preserving trust in AI systems.

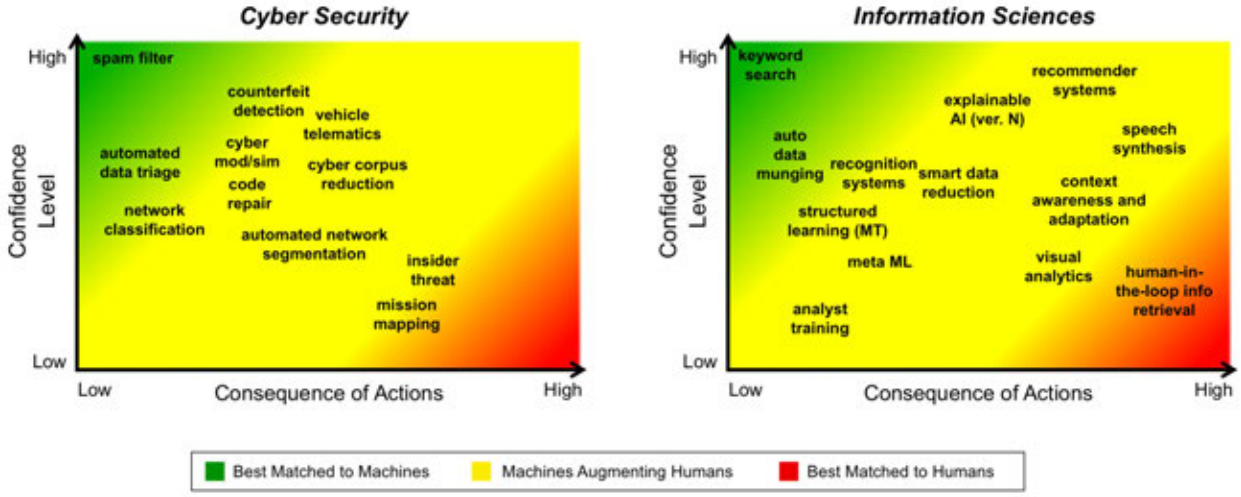


Figure 2.10. System capability space.

References

1. Cukor, C.D., *Algorithm Warfare Cross- Functional Team: Integrating Big Data and Machine Learning Across the Military.* 2017.
2. Kelly III, J.E. and S. Hamm, *Smart machines: IBM's Watson and the era of cognitive computing*. 2013: Columbia University Press.
3. Society, U.C.f.A.i.; https://www.cais.usc.edu/projects/algorithmic-experimental-game-theory/.
4. Partnership on AI. https://www.partnershiponai.org/
5. Zilis, S., *The current state of machine intelligence 2.0*. 2015, O'Reilly Media on Our Radar.
6. Shivon Zilis, *The Competitive Landscape for Machine Intelligence.* Harvard Business Review, 2016. https://hbr.org/2016/11/the-competitive-landscape-for-machine-intelligence
7. Quest, T., *The MIT Quest for Intelligence.* 2018. https://quest.mit.edu/
8. White House Summit in AI, *Summary of the 2018 White House Summit on AI for American Industry.* 2018.
9. National Defense Strategy, *Summary of the National Defense Strategy 2018.* 2018.
10. Moore, A.W., M. Hebert, and S. Shaneman. *The AI stack: a blueprint for developing and deploying artificial intelligence*. in *Ground/Air Multisensor Interoperability, Integration, and Networking for Persistent ISR IX*. 2018. International Society for Optics and Photonics.
11. Dmitri Dolgov, *Self-Driving Cars and the Future of Mobility.* 2017.
12. Alexander Wissner-Gross, *Datasets Over Algorithms.* 2016. https://www.edge.org/response-detail/26587
13. Ralph Jacobson, *IBM Consumer Products Industry Blog,.* 2013. https://www.ibm.com/blogs/insights-on-business/consumer-products/2-5-quintillion-bytes-of-data-created-every-day-how-does-cpg-retail-manage-it/
14. Deng, J., et al. *Imagenet: A large-scale hierarchical image database*. in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. 2009. Ieee.
15. Krizhevsky, A., I. Sutskever, and G.E. Hinton. *Imagenet classification with deep convolutional neural networks*. in *Advances in neural information processing systems*. 2012.
16. Martinez, D.R., M.M. Vai, and R.A. Bond, *High performance embedded computing handbook: A systems perspective*. 2008: CRC Press.
17. Wilson, J. and P.R. Daugherty, *COLLABORATIVE INTELLIGENCE Humans and Al Are Joining Forces.* Harvard Business Review, 2018. **96**(4): p. 115-123.

18. National Research Council, *Frontiers in massive data analysis*. 2013: National Academies Press.
19. Quinn, A.J. and B.B. Bederson. *Human computation: a survey and taxonomy of a growing field.* in *Proceedings of the SIGCHI conference on human factors in computing systems*. 2011. ACM.

3 Enabling Technologies (V. Gadepally)

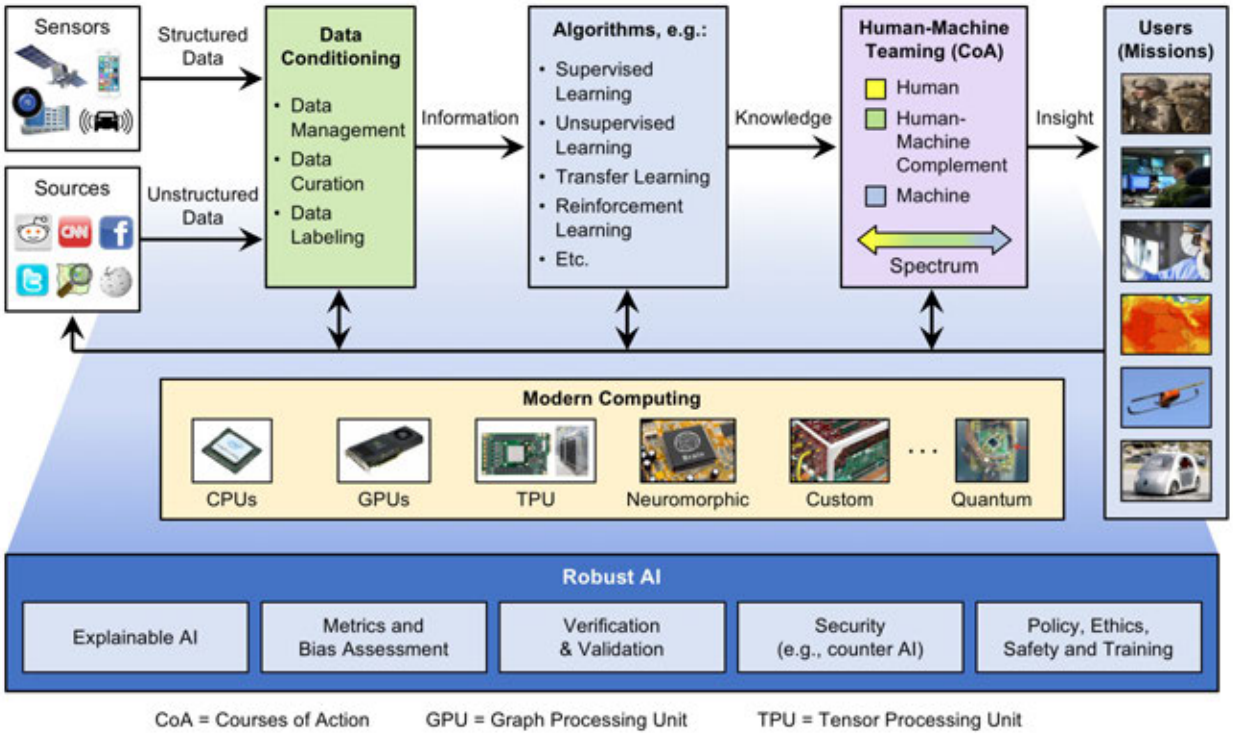


Figure 3.1. Canonical AI architecture consists of sensors, data conditioning, algorithms, modern computing, robust AI, human-machine teaming, and users (missions). Each step is critical in developing end-to-end AI applications and systems.

AI has the opportunity to revolutionize the way the DoD and IC address the challenges of evolving threats, data deluge, and rapid courses of action. AI solutions involve a number of different pieces that must work together in order to provide capabilities that can be used by decision makers, warfighters, and analysts. Consider the canonical architecture of an AI system in Figure 3.1. This figure outlines many of the important components needed when developing an end-to-end AI solution. While much of the popular press surrounds advances in algorithms and computing, most modern AI systems leverage advances across numerous different fields. Further, while certain components may not be as visible to end-users as others, our experience has shown that each of these interrelated components play a major role in the success or failure of an AI system.

On the left side of Figure 3.1, we have data coming in from a variety of structured and unstructured sources. Often, these structured and unstructured data sources together provide different views of the same entities and/or phenomenology. For example, a satellite image that indicates people in a particular region may be complemented by unstructured reports from social media or news outlets. These raw data are often fed into a data conditioning step in which they are fused, aggregated, structured, accumulated, and converted to information.

The information generated by the data conditioning step feeds into a host of supervised and unsupervised algorithms such as neural networks. These algorithms are used to extract patterns, predict new events, fill in missing data, or look for similarities across datasets. These algorithms essentially convert the input information to actionable knowledge. In our definition, we use the

term knowledge to describe information that has been converted into a higher-level representation that is ready for human consumption.

Figure 3.2. Example categories and video screen shots from the Moments in Time Dataset.

With the knowledge extracted in the algorithms phase, it is important to include the human being in the decision-making process. While there are a few applications that may be amenable to autonomous decision making (e.g., email spam filtering), recent AI advances of relevance to the DoD have largely been in fields where a human is either in- or on- the-loop. The phase of human-machine teaming is critical in connecting the data and algorithms to the end user and in providing the mission users with useful and relevant insight. Human-machine teaming is the phase in which knowledge can be turned into actionable intelligence or insight by effectively utilizing human and machine resources as appropriate.

Underpinning all of these phases is the bedrock of modern computing systems made up of a number of heterogenous computing elements. For example, sensor processing may occur on low power embedded computers whereas algorithms may be computed in very large data centers. With the end of Moore's law [1], we've seen a Cambrian explosion of computing technologies and architectures. Understanding the relative benefits of these technologies is of particular importance to applying AI to domains under significant constraints such as size, weight, and power.

Another foundational technology underpinning AI development is robust or trusted AI. In this area, researchers are looking at ways to explain AI outcomes (for example, why a system is recommending a particular course of action); metrics to measure the effectiveness of an AI algorithm (going beyond the traditional accuracy and precision metrics for complex applications or decisions); verification and validation (ensuring that results are provably correct under adversarial conditions); security (dealing with malicious or counter-AI technology); and policy decisions that govern the safe, responsible, and ethical use of AI technology. While traditional academic and commercial players are looking at these issues, some non-profit initiatives such as OpenAI or the Allen Institute are taking a leading role in this area.

In the following sections, we highlight some of the salient technical concepts, research challenges, and opportunities for each of these core components of an AI system. In order to elucidate these components, we also use a running example based on research applying high performance computing (HPC) to video classification. We would also like to note that each of the components of the AI architecture are vast academic areas with rich histories and numerous

well published results. In order to provide readers with an overall view of all the components within this section, we concentrate on high-level concepts and also include vignettes of select research highlights or application examples.

3.1.1 Video Classification Example Overview

Over the course of this section, in order to provide concrete examples of components of the AI architecture being discussed, we use a running example based on our research of using high performance computing for video classification purposes. Specifically, we concentrate on the recently developed Moments in Time Dataset [2] developed at the Massachusetts Institute of Technology (MIT) Computer Science and Artificial Intelligence Laboratory (CSAIL). This dataset consists of 1 million videos given a label corresponding to an action being performed in the video. Each video is approximately three seconds in length and is labeled according to what a human observer believes is happening in the video. For example, a video of a dog barking is classified as "barking" and a video of people clapping would be labeled as "clapping." Figure 3.2 shows a few screenshots of videos from the dataset and associated labels. Of course, there are many areas where a particular label may not be as precise. For example, videos with the action label "cramming" could imply a person studying before an exam or someone putting something into a box. As of now, each video in the Moments in Time Dataset is labeled with one of approximately 380 possible labels. Some of the video clips also contain audio, but it is not necessarily present for all videos.

The Moments in Time Dataset is an example of a well-curated dataset that can be used to conduct research on video classification. To this effect, the creators of the dataset held a competition in 2018 to encourage dataset usage and share results that highlight the state of the field. Information about this competition can be found at:
https://moments.csail.mit.edu/challenge2018/

As a metric to present the quality of a particular algorithm, the competition called for presentation of a top-*k* accuracy score. This metric is defined as follows: An algorithm will label each of the videos with one of *k* labels. The top-*k* accuracy says that a video was correctly identified if one of its top *k* labels is the correct label. For example, a video may be classified (in decreasing probability) as: *(barking, yelling, running, ...)*. If the correct label (as judged by a human observer) is "yelling", the top-5 accuracy for this would be 1. The top-1 accuracy would be 0. As of June 2018, competition winners had top-1 accuracies of approximately 0.3 and top-5 accuracies of approximately 0.6 [3-5].

3.2 Data Conditioning

Many AI application developers typically begin with a dataset of interest and a vision of the end analytic or insight they wish to gain from the data at hand. While these are two very important components of the AI pipeline, one often spends the first few weeks (sometimes months) in the phase we refer to as data conditioning. This step typically includes tasks such as figuring out how to store data, dealing with inconsistencies in the dataset, and determining which algorithm (or set of algorithms) will be best suited for the application. Larger, faster, and messier datasets such as those from IoT sensors, medical devices or autonomous vehicles only amplify these issues. These challenges, often referred to as the three V's (Volume, Velocity, Variety) of Big Data, require low-level tools for data management and data cleaning/pre-processing. In most applications, data can come from structured and/or unstructured sources and often includes inconsistencies, formatting differences, and a lack of ground-truth labels.

By some accounts, data conditioning can account for nearly 80% of the time consumed in developing a data science or AI application [6]. Within the realm of data conditioning, specific tasks include data discovery, data linkage, outlier detection, data management, and data labeling.

At a high level, the concept of data conditioning is the effort required to go from raw sensor data to information that can be used in further processing steps. Sometimes this phase is also referred to as data wrangling. Typically, each of these data conditioning tasks can be cumbersome, require significant domain knowledge, and represent a significant hurdle in developing an AI application. Many of the recent algorithmic advances have, in fact, occurred in areas where "conditioned" data can be found. For example, advances in image classification were largely driven by the availability of the ImageNet dataset [7], advances in handwriting recognition by the MNIST dataset [8], and advances in video recognition by the Moments in Time Dataset [2]. Other popular datasets such as CIFAR-10 [9], ATARI games [10], and Internet traces [11] have also played their role in advancing certain classes of algorithms and genres of applications.

There are a number of research efforts and organizations aiming to reduce the data conditioning barrier to entry. In this section, we will focus on three particular aspects of data conditioning: data management, data curation, and data labeling (for supervised learning). The input to this phase is typically raw data from heterogenous sources. This step aims to convert these data by aggregating them in a single place, designing a schema that relates all the components and often performs rudimentary anomaly detection/outlier detection. In the following subsections, we describe a few approaches to these tasks.

3.2.1 Data Management

AI and machine-learning systems are highly dependent on access to consistent and formatted data. However, it is rare that a collection of sensors such as those used in the AI pipeline directly provide this information in a consistent manner. For example, in the video classification example, certain cameras may be turned off, may have different metadata, have different compression techniques, have different frame rates, have different color normalization schemes, etc. Further, fusing different pieces of data coming from disparate sources can be a major challenge—like fusing the information contained in audio streams with video streams in the video classification example. One of the first challenges is providing a uniform platform in which data can be fused and managed.

Traditionally, database systems are seen as the natural data management approach. A database is a collection of data and supporting data structure. Traditionally, databases are exposed to users via a database management system. Users interact with these database management systems to define new data structures, schemas (data organization), to update data, and retrieve data. Beyond databases, developers may store data as files leveraging parallel file

systems such as Lustre [12]. For the remainder of this section, however, we will focus on database systems such as those shown in Figure 3.3.

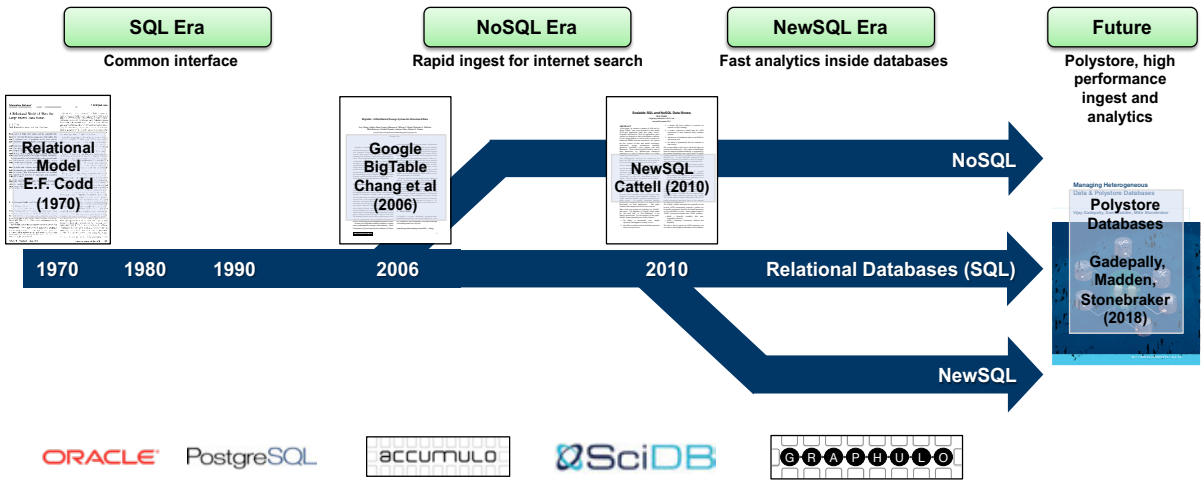


Figure 3.3. Evolution of database management systems.

Traditional database management systems such as Oracle [13] and PostGRES [14], sometimes referred to as relational databases, while compliant with ACID [15] guarantees, are unable to scale horizontally for certain applications [16]. To address these challenges, large internet companies such as Google and Facebook developed horizontally scalable database technologies such as BigTable [17] and Cassandra [18]. These NoSQL [19] (not-only SQL) technologies enabled rapid ingest and high performance even on relatively modest computing equipment. BigTable inspired databases such as Apache Accumulo [20] extended the NoSQL model for application specific requirements such as cell-level security. NoSQL databases do not provide the same level of guarantees on the data as relational databases [16]; however, they have been very popular due to their scalability, flexible data model, and tolerance to hardware failure. In the recent few years, spurred by inexpensive high performance hardware and custom hardware solutions, we have seen the evolution of a new era in database technologies, sometimes called NewSQL databases [21]. These data management systems largely support the scalability of NoSQL databases while preserving the data guarantees of SQL-era database systems. Largely, this is done by simplifying data models, such as in SciDB, or leveraging in-memory solutions such as in MemSQL and Spark. Looking towards the future, we see the development of new data management technologies that leverage the relative advantages of technologies developed within the various eras of database management technologies. A very high-level view of this evolution is presented in Figure 3.3. Looking towards the future, it is clear that no single type of database management systems is likely to support the kinds of data being collected from heterogenous sources of structured and unstructured data. In order to address this challenge, one example of an active area of research in data management is in multi-database systems [22] such as Polystore databases and a specific example is the BigDAWG system described below.

Data Management Research Example—BigDAWG

BigDAWG [16, 23, 24], short for the Big Data Working Group, is an implementation of a polystore database system designed to simplify database management for complex applications. For example, modern decision support systems are required to integrate and synthesize a rapidly expanding collection of real-time data feeds: sensor data, analyst reports, social media, chat, documents, manifests, logistical data, and system logs (to name just a few). The traditional technique for solving a complex data fusion problem is to pick a single general-purpose database engine and move everything into this system. However, custom database engines for sensors, graphs, documents, and transactions (just to name a few) provide 100× better performance than general-purpose databases. The performance benefits of custom databases have resulted in the proliferation of data-specific databases, with most modern decision support systems containing five or more distinct customized storage systems. Additionally, for organizational or policy reasons, data may be required to stay in disparate database engines. For an application developer, this situation translates to developing his or her own interfaces and connectors for every different system. In general, for N different systems, a user will have to create nearly N^2 different connectors. BigDAWG allows users to access data stored across multiple databases via a uniform common interface. Thus, for a complex application in which there is scientific data, text data, and metadata, a user can store each of these components in the storage technology best suited to each data type, but also develop analytics and applications that make use of all of this data without having to write custom connectors to each of these storage technologies. The end-to-end architecture of the BigDAWG polystore system is described in Figure 3.4. This architecture describes how applications, visualizations, and clients at the top access information stored in a variety of database engines at the bottom. At the bottom, we have a collection of disparate storage engines (we make no assumption about the data model, programming model, etc., of each of these engines). These storage engines are organized into a number of *islands*. An island is composed of a data model, a set of operations, and a set of candidate storage engines. An island provides location independence among its associated storage engines. A *shim* connects an island to one or more storage engines. The shim is basically a translator that maps queries expressed in terms of the operations defined by an island into the native query language of a particular storage engine. A key goal of a polystore system is for the processing to occur on the storage engine best suited to the features of the data. We expect in typical workloads that queries will produce results best suited to particular storage engines. Hence, BigDAWG needs a capability to move data directly between storage engines. We do this with software components we call *casts*.

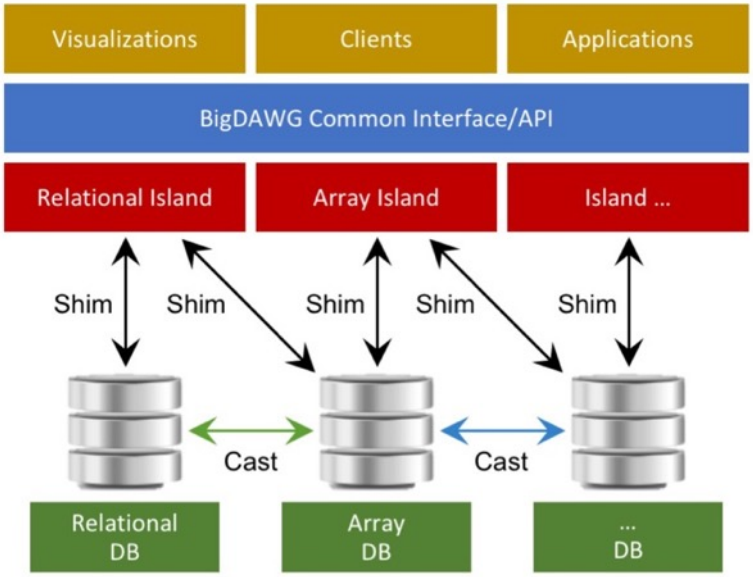


Figure 3.4. BigDAWG architecture.

Database and Storage Engines

A key design feature of BigDAWG is the support of multiple database and storage engines. With the rapid increase in heterogeneous data and the proliferation of highly specialized, tuned, and hardware-accelerated database engines, it is important that BigDAWG support as many data models as possible. Further, many organizations already rely on legacy systems as a part of their overall solution. We believe that analytics of the future will depend on many distinct data sources that can be efficiently stored and processed only in disparate systems. BigDAWG is designed to address this need by leveraging many vertically integrated data management systems. The current implementation of BigDAWG supports a number of popular database engines: PostGRES (SQL), MySQL (SQL), Vertica (SQL), Accumulo (NoSQL), SciDB (NewSQL), and S-Store (NewSQL). The modular design allows users to continue to integrate new engines as needed.

BigDAWG Islands

The next layer of the BigDAWG stack is its islands. Islands allow users to trade off between semantic completeness (using the full power of an underlying database engine) and location transparency (the ability to access data without knowledge of the underlying engine). Each island has a data model, a query language or set of operators, and one or more database engines for executing them. In the BigDAWG prototype, users determine the *scope* of their query by specifying an island within which the query will be executed. Islands are a user-facing abstraction, and they are designed to reduce the challenges associated with incorporating a new database engine. The current implementation of BigDAWG supports islands with relational, array, text, and streaming models. Our modular design supports the creation of new islands that encapsulate different programming and data models.

BigDAWG Middleware and API

The BigDAWG "secret sauce" lies in the middleware that is responsible for developing cross-engine query plans, monitoring previous queries and performance, migrating data across database engines as needed, and physically executing the requested query or analytic. The BigDAWG interface provides an API to execute polystore queries. The API layer consists of server- and client-facing components. The server components incorporate islands that connect to database engines via lightweight connectors referred to as shims. Shims essentially act as an adapter to go from the language of an island to the native language of an underlying database engine. In order to identify how a user is interacting with an island, a user specifies a scope in the query. A scope of a query allows an island to correctly interpret the syntax of the query and allows the island to select the correct shim that is needed to execute a part of the query. Thus, a cross-island query may involve multiple scope operations.

3.2.2 Data Curation

Data curation, within the context of our AI architecture, is used to refer to the process of maintaining data from creation (sensor collection) to usage in the machine-learning algorithm. In contrast to data management in the previous section, data curation typically involves some form of data normalization, data cleaning, data enrichment, and/or data discovery. Figure 3.5 describes where such curation activities may sit in relation to the data management and polystore technologies described in the previous section.

While this is often a largely manual process with a human expert looking at data and determining if there are errors, there are machine-learning techniques that can significantly reduce the amount of manual work needed. To do this, one may employ some form of unsupervised learning to look for obvious anomalies and outliers in the data. For example, in a medical dataset, one may incorrectly encode a field that is supposed to use inches as feet (thus allowing for people that are 60 feet tall!). This step may also look at simplifying data via dimensionality reduction techniques. Finally, this step often does some form of normalization and weighting in order to ensure that the further processing steps are acting on the right data.

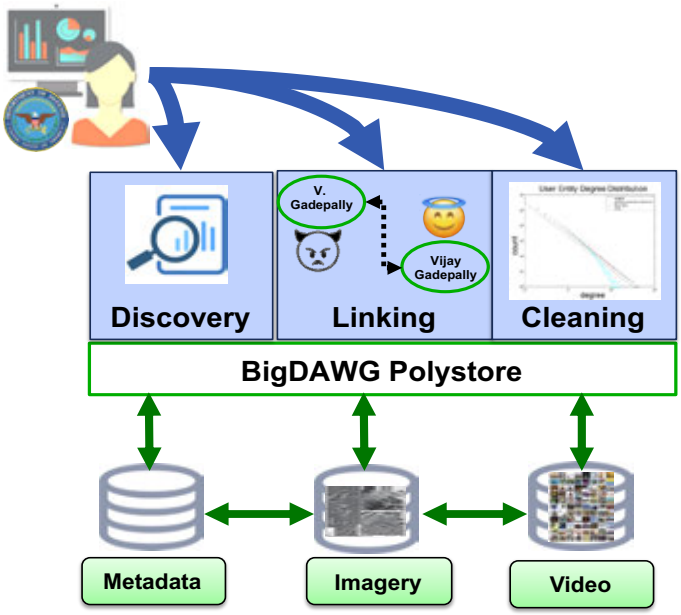


Figure 3.5. Notional overview of data curation activities on top of data management.

Anomaly and Outlier Detection

A particular task in data curation is the process of converting data, often with noisy inputs, to a version of the data that is amenable to further processing. The overall goal is to improve the signal-to-noise ratio such that further algorithms are likely to learn from the correct components of the data. The source of these errors can be due to a variety of factors such as sensor error, human error, etc. While this is a very wide field in which significant human expertise is often used, one can also make use of machine-learning techniques such as unsupervised learning to automate or simplify the process. For example, a human analyst may set rules that are used on the data to ensure sufficient quality (such as height must be less than 8 but greater than 0 feet).

As an example of a more automated technique, for example, given noisy data from a sensor, one may cluster the various data points into a set of clusters. Outliers from these clusters may be data points that are likely to be important—either because they are anomalous or otherwise. As noted in [25], there are three general approaches to outlier detection: 1) leverage unsupervised learning to look for outliers such as in the example above, 2) leverage a set of labels that correspond to normal or abnormal data in order to look for outliers, and 3) model only normal behavior with the intention of looking for samples that do not fit within the bounds of normal behavior.

The task of anomaly and outlier detection is a very important step and is often the limiting factor in quality of algorithmic performance in further processing steps.

Dimensionality Reduction

Often in a dataset, it is necessary to condense the amount of information that will be processed by the subsequent pipeline. This can be done for a variety of reasons such as improved computational performance, removal of redundant dimensions, or removal of features (dimensions) that will not play a part in further processing. In an image, for example, dimensionality reduction could be converting a color image with three channels to a single channel grayscale version that maintains important features of the original image such as edges or shapes without the additional red, green, and yellow channel information.

Techniques for dimensionality reduction look for variance in features within a dataset along with correlations across features. Through algorithms such as principal component analysis, users can quickly determine which features (or dimensions) of their dataset have the greatest variance and look for features that are closely correlated with other features. Using this information, it may be possible to keep only high-variance features and remove features that are closely correlated with other features. Thus, it may be possible to remove a large set of features within a dataset without adversely affecting future algorithmic performance.

Data Weighting and Normalization
Another technique for data conditioning is often referred to as data weighting or normalization. This process involves bringing various features, or dimensions, within a dataset to a common frame of reference or common dynamic range. If particular features have very high rates of change and/or very high dynamic range, it is possible that further processing steps will tend to overweight these changes compared to other features that may not have as high a range. Alternatively, there may be certain features in the dataset that should, in fact, have an outsized role in determining the output. In such cases, one may weigh these features higher than other features.

3.2.3 Data Labeling

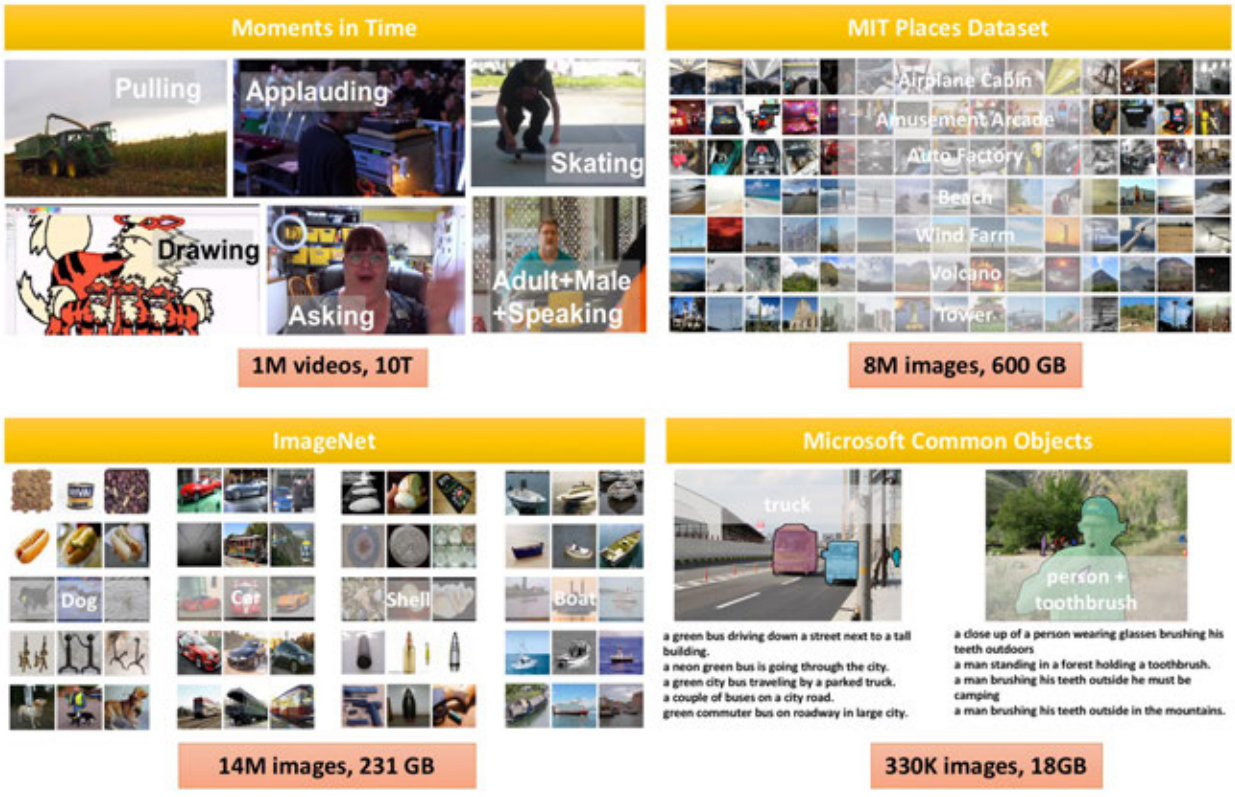


Figure 3.6. Examples of popular curated datasets ImageNet, COCO, MIT Places, Moments in Time.

One of the most time-consuming tasks in applying supervised learning techniques (such as neural networks) is in providing useful labels in the data. This task, sometimes referred to as data annotation, aims to provide the machine-learning algorithm with a set of labels that can be used to train the classification or regression model. Essentially, in supervised learning, the machine

learns the pattern that associates an input with an output label. A multitude of labeled data points can provide a robust and repeatable training process. Of course, data collected from a sensor is rarely given clear labels and there has been a great deal of effort in the wider community to come up with techniques that can be used to simplify data labeling.

One obvious method to label a large quantity of data is to use human evaluators. A human evaluator (sometimes referred to as an "oracle") is often considered as the gold standard for labeling. While a developer of an algorithm or AI application can label a few images, this may not scale to thousands or millions of data samples. Further, a single evaluator can introduce bias (perhaps the evaluator has difficulty judging particular colors that may bias their labels). Figure 3.6 describes a few well-curated labeled datasets that are used widely in the community. For example, ImageNet [26] is a dataset that has helped spur the growth of recent computer vision advances. Other datasets such as MIT Places [27], Microsoft COCO [28], and the previously mentioned Moments in Time are also widely used by AI researchers.

One common technique for manual data labeling used across the commercial world is to leverage a service such as Amazon's Mechanical Turk [29, 30]. Using such a service, a user can upload a dataset of interest and quickly leverage a pool of users who can provide labels for the dataset. While popular for tasks such as labeling of images or transcription of spoken languages, such crowdsourcing techniques often suffer quality issues [31] or do not scale to sensitive or proprietary datasets that require significant domain expertise such as those used across the DoD/IC. Within the DoD, one example of "crowdsourced" data labeling is being done via the Air Force's Project MAVEN [32]. This project aims to leverage the domain expertise of service members to label data where current techniques fail or cannot be applied due to the sensitivity of data.

Beyond manual annotation of data, the research community has been actively looking at techniques that provide varying levels of automation in the data labeling process by leveraging a human selectively or by leveraging machine intelligence on a small set of labeled data points.

In general, there are a number of algorithmic techniques that can still be used in cases where labels for data do not exist. For example, in the semi-supervised learning paradigm, a small subset of labeled data can be used in conjunction with unlabeled images. These semi-supervised techniques essentially infer the labels on data using a number of different techniques. For example, generative techniques assume that the subset of labeled data consists of labels for all classes of importance and attempts to learn statistical patterns of the labeled data. Then, assuming the unlabeled data looks similar to the labeled samples, it is possible to deconstruct the unlabeled data into the statistical components and compare with the labeled data. One particular instance is in the cluster and learn paradigm where the unlabeled data is clustered along with the labeled data. Unlabeled data is then assigned a label based on its proximity to a labeled sample. Another technique, referred to as self-training, may instead attempt to create a classifier using the labeled datasets and then iteratively apply the classifier to the unlabeled data. After application, instances where the classifier was very confident on the classification can now be added back to the pool of labeled data and the process can be repeated. Other techniques may use graph-based methods to represent the dataset where both labeled and unlabeled datasets form parts of the graph. Then, one can use graph matching techniques to apply labels to particular instances.

Another expanding area of research is in active learning. In this area, one attempts the above techniques as possible, but also makes use of a human observer to help the system. These techniques essentially allow both humans and machines to work together, with the machine applying labels to straightforward samples and focusing human attention on difficult samples.

Very generally, there are a number of ways in which one could still perform supervised learning in cases where labels are either nonexistent, limited, or difficult to collect:

1. **Semi-supervised learning**. In this paradigm, an algorithm is designed to leverage some labeled and some unlabeled data. For example, for a DoD/IC specific use-case, one may leverage a different but related dataset (using publicly available datasets on cars [34] to train models that can look for military vehicles).
2. **Active learning**. In this paradigm, a user is in the loop of the data preprocessing and labeling. In this technique, the system can apply labels to particular observations in which the confidence is high and can also leverage a human oracle for difficult or previously unseen samples.
3. **Self-supervised learning**. In this paradigm, the system does not need any explicit labels on the data samples to be provided. Rather than use labels, the algorithm can automatically extract metadata, encoded domain knowledge or other such correlations that can be used in place of explicit labels. For example, in [35], the authors learn the task of image colorization as a means to learn tracking.
4. **Automated labeling**. This technique leverages statistical techniques to apply labels to data. For example, in [36], the authors propose a technique using conditional random fields to build models that can be used to label data sequences.

3.2.4 Data Conditioning within the Context of Exemplary Application

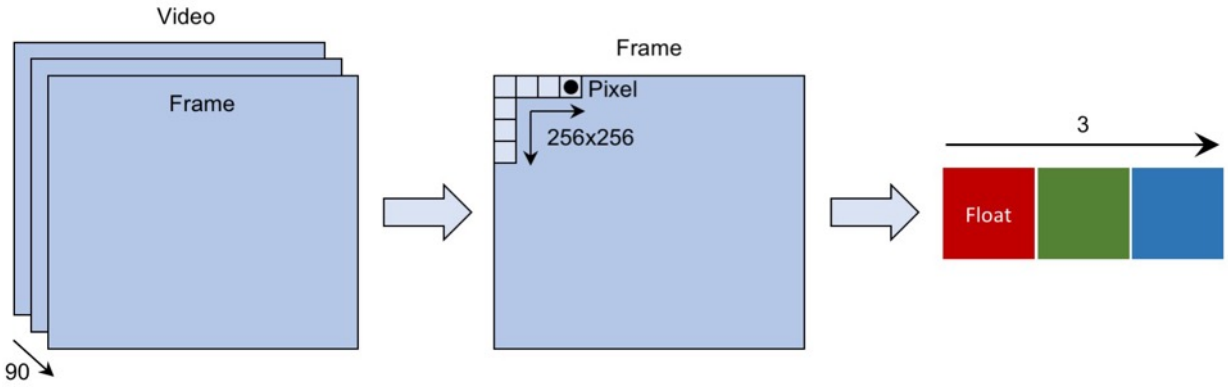


Figure 3.7. Data Conditioning within the context of Video classification example. Data starts as raw compressed .mp4 videos, are converted to individual frames and resized to consistent dimension arrays across RGB channels.

While the Moments in Time Dataset is a well curated dataset with high quality labels, there are still a number of preprocessing steps that need to be taken. In developing our pipeline, the first thing we need to do is convert the MPEG-4 encoded video files to arrays that are easy for future processing steps. In the dataset provided, each video is 3 seconds long and consists of 30 frames per second (for a total of 90 frames per video). Each frame is 256×256 pixels and has three floating point values that correspond to the red, green, and blue (RGB) channels. Thus, for each video represented in tensor format, this corresponds to a shape of 90×256×256×3 for a total of approximately 17.5 million integers per video. For 1 million videos, this corresponds to approximately 70 terabytes of raw video data. As a next step, we resize all of the frames to be a specific size—224×224 pixels in our case. Finally, we normalize the data so that all frames have a similar distribution of pixel intensities. First, we divide by 255 (the maximum range of pixel intensities) from each pixel intensity from each frame. Then we find the mean pixel intensity across all video frames and subtract this from the previous value. Next, we divide each pixel intensity by the standard deviation of pixel intensities. As a final step, in order to remove potential outlier frames, we remove all frames that contain pixel intensities below or above a

certain threshold (in order to get rid of frames that are all black or white). This process is described pictorially in Figure 3.7.

At the end of this step, the raw videos are now cleaned up in a consistent format and stored on the file system as arrays that can be read in for further processing.

3.3 Algorithms

In the past decade, much of the hype around AI has come from advances in performance of machine-learning algorithms applied to various problems such as image classification. In this section, we will describe some of the popular machine-learning developments while highlighting a few salient algorithms.

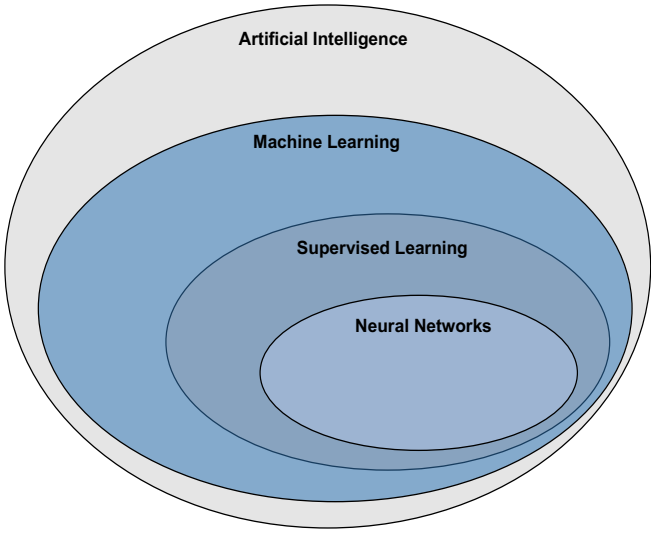


Figure 3.8. Relationship between artificial intelligence, machine learning, and neural networks. Figure adapted from *Deep Learning* by Ian Goodfellow.

From our perspective, machine-learning algorithms form the core of AI algorithms (with a few exceptions such as with expert systems). Figure 3.8, adapted from [37] describes, at a high level, the relationship between AI, machine learning, supervised learning and neural networks. While a lot of recent focus has been on neural networks, it is important to understand that there a multitude of machine-learning algorithms beyond neural networks that are used widely in AI applications. It is also interesting to note that many current and historical AI systems such as [38] actually leveraged techniques outside of traditional machine learning such as expert systems [39, 40] or the more general *knowledge-based systems* [41]. Knowledge-based systems leverage a knowledge base of information and an inference engine to apply the knowledge base to a particular application. Expert systems, a form of knowledge-based systems, utilize human experts to formulate a knowledge base that can be applied via an inference engine for decision making. Knowledge-based and expert systems continue to be used in a variety of applications such as tax software and were even included in early autonomous vehicles [42]. For domains in which collection of data is limited, rules are complex and there is significant human expertise, expert systems can still play a major role in building an end-to-end AI system. Further, knowledge-based systems are often inherently explainable and interpretable, which can make them amenable to applications in which trust is critical.

At a high level, one can think of a tradeoff that exists between the ease of codifying human knowledge, compute power, and data availability. In cases with significant codified knowledge, limited compute power, and limited data availability, knowledge-based systems can play an important role in AI systems. While the majority of our discussion is on machine-learning algorithms, knowledge- and expert-based systems are an important class of algorithms that still have wide applicability to DoD/IC applications.

With the ability to collect large quantities of data, coupled with greater computational resources, the shift in technology has been toward the world of machine learning. In the machine-learning paradigm, a user provides a system with (varying amounts of) data along with a set of rudimentary guidelines (such as generative model, number of classes, etc.). Using this

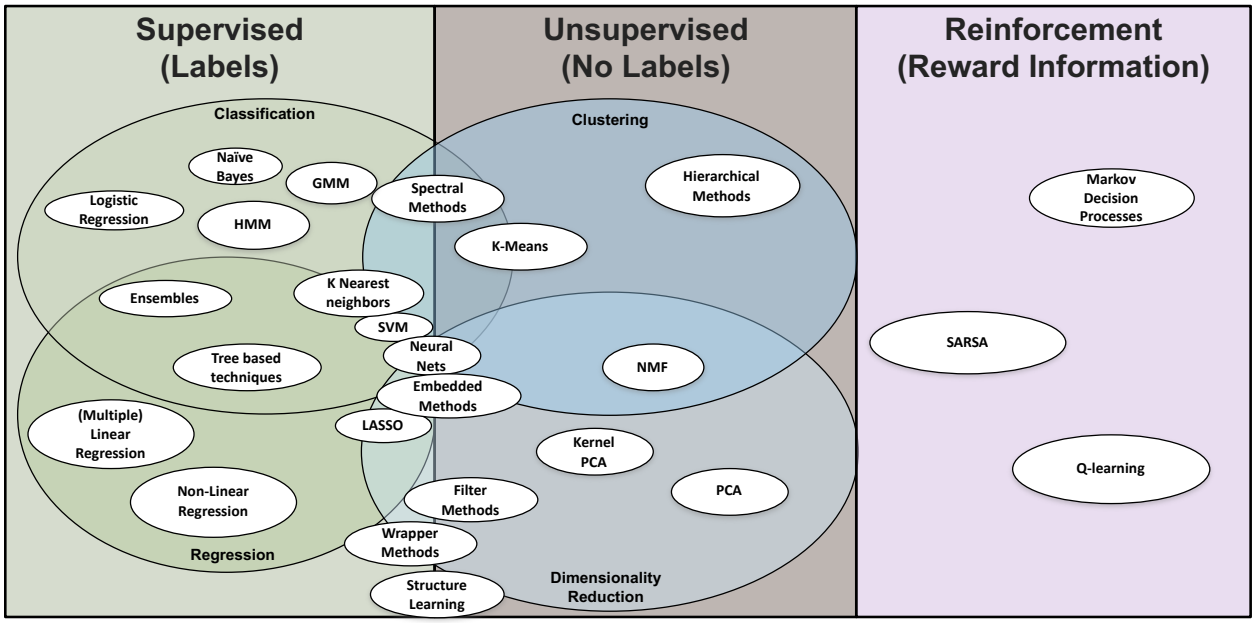


Figure 3.9. Popular machine-learning algorithms and taxonomy. We acknowledge Lauren Milechin (MIT EAPS) and Julie Mullen (MIT LLSC) for contributing to the development of this figure.

input, the machine "learns" a mapping that can be applied to the data in order to generate the required output. With this paradigm, machine-learning algorithms essentially learn the relationship between inputs and outputs.

To highlight the many techniques for machine learning, we use the taxonomy presented in Figure 3.9. Other taxonomies, such as those from [43], are alternate ways of organizing the variety of machine-learning algorithms. In our taxonomy, machine-learning algorithms are broadly broken into supervised, unsupervised, and reinforcement learning techniques. Supervised learning techniques have labels that relate inputs and outputs; unsupervised techniques typically do not; and reinforcement learning algorithms are provided with reward signals rather than explicit labels. While this is a good first-order break-up of the landscape, we would like to note that these boundaries are not meant to be rigid and in practice, there are a number of algorithms that cross these boundaries. In the remainder of this section, we provide a brief overview of each of these learning paradigms along with a discussion of where these techniques may be used in the development of AI systems.

3.3.1 Supervised Learning

Supervised algorithms start with labeled data (or ground truth) and aim to build a model that can be used to predict labels for data in which labels (or classifications) do not exist. As shown in Figure 3.10, supervised learning makes use of data and labels to train a model that can be applied on test data in order to predict future labels. Thus, the video classification task we present is a supervised learning problem because we are given labels of the video samples that are used to train a model that we use to classify new data samples. Generally, supervised learning algorithms attempt one of the following goals: regression (to predict a future continuous variable) or classification (to predict a new class or label). In most cases, the majority of computing time in supervised learning is spent training a model from training examples. The model generated by the algorithm is essentially the representation of what the system believes relates the inputs to the corresponding outputs. After training, the model can now be applied for inference tasks. During inference, the trained model is applied to previously unseen test data to predict labels in the case of classification or future values in the case of regression.

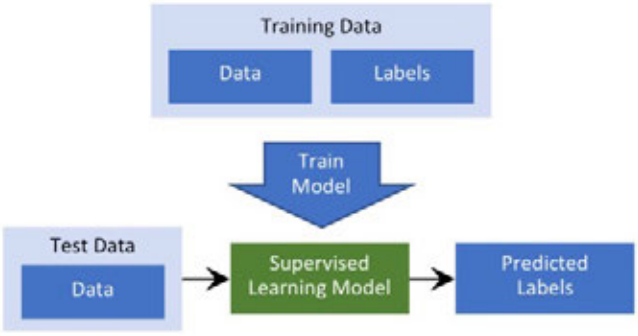


Figure 3.10. Supervised learning makes use of data and labels in order to train a model that automates this process.

To further explain supervised learning, consider the versatile *k*-nearest neighbors algorithm [44]. This algorithm can be applied to both classification and regression problems and relies on the assumption that similar points (data points that are close to each other in a feature space) have similar labels/outputs. To use this algorithm, each data point in a dataset is represented in a multi-dimensional feature space that corresponds to data features (for example, in an image, this could be the pixel intensities and/or pixel locations) that represent each individual data point. With this representation, it is possible to predict values or apply labels to new data points. For a new data point represented in the same feature space, we find the *k*-closest neighbors (i.e., we look for all data points that are close by in the multi-dimensional space). If performing regression, the predicted value is an average of the neighboring values. For classification, the *k*-nearest neighbors vote with their label and the predicted class label for the new data point is the label held by a majority of the *k*-nearest neighbors. While this algorithm is very easy to implement and has been used in a variety of applications, it can be sensitive to too many features (or dimensions) or in feature spaces where the closeness assumption or definition does not hold. Figure 3.11 describes an example of applying the *k*-nearest neighbors algorithm to a two-dimensional dataset. On the left side of the figure, we describe how one would classify a data point when we have labels corresponding to three labels. The new data point in this figure is represented by the star-shaped dot. The k=7 nearest neighbors are highlighted with circles. On the right side, we describe how one could use the same algorithm for regression. In this example, we are trying to predict the value of the second dimension. Similar to the classification example, for a new data point, we find the k=7 closest neighbors and the predicted second dimension value is given as an average of the second dimension value of the seven nearest neighbors.

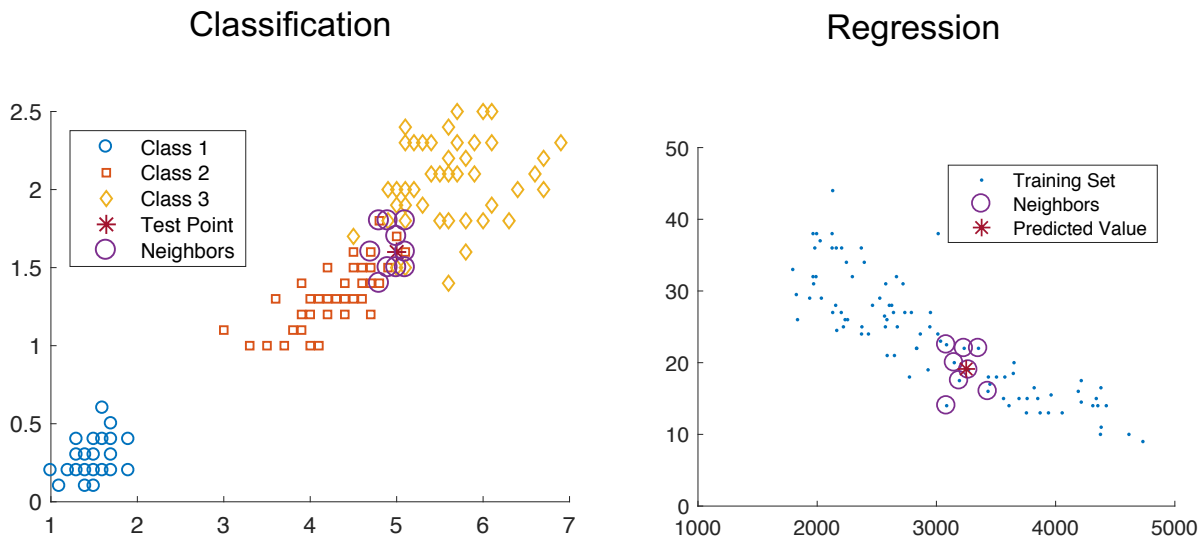


Figure 3.11. Using a *k*-means supervised learning technique for classification and regression.

Supervised learning algorithms play an outsized role in many of the recent machine-learning advances. The quality of supervised learning algorithms, however, largely relies on access to high quality labeled data. While there are a number of labeled datasets that have yielded breakthrough advances in fields such as image classification, voice recognition, game play, and regression, in DoD/IC applications, this may not be available. In such cases, unsupervised learning provides a means to analyze data.

3.3.2 Unsupervised Learning

Unsupervised learning is a technique that applies statistical analysis techniques to data in which explicit labels are not provided. Very often, upfront data conditioning is done via unsupervised learning with the aim of looking for outliers or reducing the dimensionality of data. Without data labels, it is difficult to classify data points, and unsupervised learning techniques are limited to clustering or dimensionality reduction. More formally, if we observe data points X_1, X_2, X_3,…,X_N, we are interested in looking for patterns that may occur among these data points. In this example, each data point X can be made up of a number of features or dimensions. For example, in an image, each data point X may corresponds to a single pixel and each data point can be represented by three features—the RGB value associated with a single pixel.

In unsupervised learning, there are often no clear metrics such as accuracy or recall for the algorithm. However, given that unsupervised learning algorithms can work on unlabeled data, they are often an important first step in any AI pipeline.

Typically, unsupervised learning is used for clustering and data projection. Clustering algorithms group objects or sets of features such that objects within a cluster are more "similar" than those across clusters. The definition of similarity is often highly dependent on the application. For example, in an image processing application, similarity may imply the difference in pixel intensities across different image channels; in other applications, similarity may be defined as Euclidean distance. Measuring intra-cluster similarity is often done via measures such as squared error (se):

$$se = \sum_{i=1}^{k} \sum_{p \in c_i} \|p - m_i\|^2$$

Many clustering algorithms work by iteratively trying to minimize error terms, such as the squared error term defined above, by placing different data points in different clusters and measuring the error. Within clustering, common algorithms include *k*-means, nearest neighbor search, Spectral clustering. Figure 3.12 shows a pictorial example of a notional clustering algorithm applied to data within a two-dimensional space. For the purpose of this figure, the "similarity" is defined to be Euclidean distance (the geometric distance between two points).

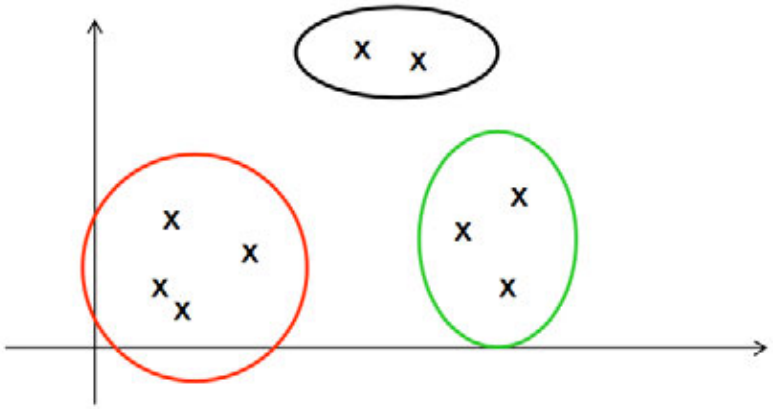


Figure 3.12. Notional clustering example for data represented by two features.

Within data projection/preprocessing, typical tasks include principal component analysis (PCA), dimensionality reduction, and scaling. Dimensionality reduction is used to reduce a large dataset into a more efficient representation comprised of high variance dimensions. These techniques can be especially useful to simplify computation or represent a dataset. Typically, one uses these techniques for selecting a subset of important features or representing data in a lower number of dimensions.

Consider the example of using PCA for dimensionality reduction. In this technique, a set of possibly correlated variables are converted to a set of uncorrelated variables using orthogonal transformations. In essence, through this technique, we are looking for a lower dimensional space representation of a dataset that still maintains some of the broad trends in the data.

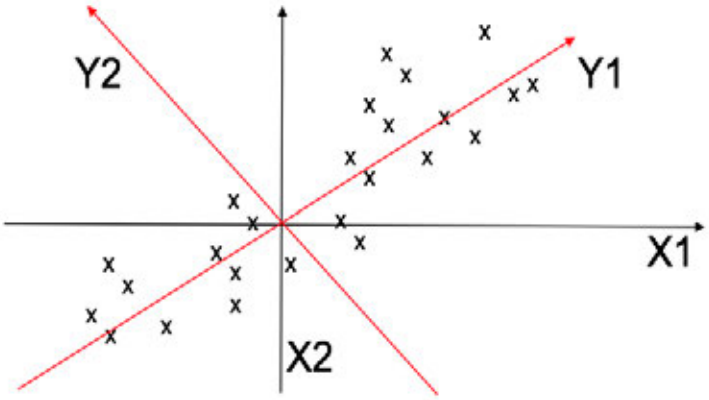


Figure 3.13. Principal component analysis. In this example, Y1 is the first principal component and Y2 is the second principal component.

For example, Figure 3.13 describes an example two-dimensional dataset that was originally represented by the axes X1 and X2. After applying PCA, the axes Y1 and Y2 are found to be the principal components. Using this new representation, the PCA technique says that if you would like a lower dimensional (in this case, one-dimension) representation of the dataset, you can convert it to the axes of Y1. As can be seen in the Figure 3.13, axis Y1 does a good job of representing the spread in the data and would be the best 1D representation of the data.

3.3.3 Supervised and Unsupervised Neural Networks

Neural networks are a popular machine-learning technique that are largely used for supervised learning but can be applied to unsupervised learning problems as well. The biologically inspired computing systems learn by repetitive training to do tasks based on examples (training data). A neural network consists of inputs, layers, outputs, weights, and biases. Deep neural networks (DNNs) differ from traditional neural networks by having a large number of hidden layers. DNNs have had much success in the past decade in a variety of applications and are supported by a number of toolboxes [45-47] and hardware platforms. Popular extensions of the neural network computation model include convolutional neural networks [7], recursive neural networks [48], and deep belief networks [49]. The network shown in Figure 3.14 in an example of a feedforward neural network that consists of one input layer, one hidden layer and one output later. The arrows in the figure indicate connections across neurons.

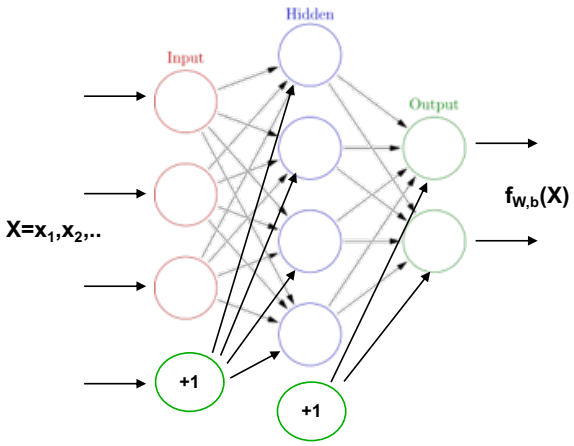


Figure 3.14. Notional neural network.

In general, for a neural network of k layers, the following expression describes the input-output behavior between any two layers (l and $l+1$):

$$y_{l+1} = f(W_l y_l + b_l)$$

Where y_l corresponds to the outputs of layer l, W_l the weights between layers l and $l+1$ and b_l the corresponding biases. In this equation, the function $f(.)$ corresponds to a non-linear activation function such as the sigmoid, rectified linear unit or tanh function [50]. During the training phase of a neural network algorithm, the weights W and biases b are iteratively adjusted in order to represent the often non-linear relationship between inputs and outputs. The training phase consists of the following steps:

1. Forward Propagation: A training example (or set of examples) is fed through the network with some initialized weights and biases. The output is computed and an error term is calculated as the difference between the computed output(s) and the intended (ground-truth) output(s).
2. Backward Propagation: Given the error from the previous step, network weights are adjusted to better predict the labels of future unlabeled examples. This is done by adjusting each weight in proportion to its individual contribution to the overall error. As the contribution to the error from weights in earlier layers depends on the contribution from weights in later layers, the error signal is described as “backward propagating” through the

network. Iterative optimization techniques built on this idea, such as the popular stochastic gradient descent (SGD), are thus known collectively as "Back Propagation."

The above procedure is repeated for as many training examples as possible and when a particular convergence criterion (such as accuracy) is met, the model is considered trained and can be used for inference. In practice, it is common to split the total training set into training and validation subsets. This can help reduce the likelihood of overfitting in which the trained model works especially well on the training data but does not generalize to new unseen samples. Typically, the training phase is where the majority of computation takes place. Once a model is trained and validated, it is ready to be used for inference.

Inference is the phase in which a trained model is applied to new data in order to predict new values (i.e., a forward pass through the network with weights and biases set by the training step). In this step, the model computes an output for a previously unseen input. If the model was trained well, the output should correspond to the correct output for a given input example.

While many uses of neural networks focus on supervised learning problems, there are many ways to use neural networks for unsupervised learning. Network architectures such as autoencoders or self-organizing maps are often used as techniques to use neural networks for unsupervised learning problems. Another technique, generative adversarial networks (GANs) [51] (not to be confused with adversarial AI discussed later in this section), uses two neural networks—a generator and discriminator—to train a model. In this architecture, the generator creates artificial samples meant to represent data from the training dataset. These artificial samples are then passed to the discriminator, which decides whether the sample is real or fake. Continuing this process iteratively allows the generator to improve the quality of its artificial samples and the discriminator to improve its discriminative capabilities. GANs can be an important tool in understanding the structure and patterns present in unlabeled datasets.

At present, the theory behind deep neural networks—why they work, how to estimate performance, performance bounds, etc.—is not well understood and is an active area of fundamental research. A user wishing to develop a neural network solution will typically be faced with a number of choices such as model architecture (what type of network), number of layers, activation functions, learning rate, batch size (for memory limited applications), etc. While there are some high-level guidelines that can be used by practitioners developing networks for their application, much of the state-of-the-art relies on exploration of parameters. There are also techniques such as in [52] that can leverage high performance computing to look at this vast parameter space more efficiently, but much of the current state-of-the-art relies heavily on domain knowledge and experience. Understanding the theory behind DNNs is particularly important for DoD and IC applications where we will need to adopt solutions from other applications. At present, without a clear understanding of the theory behind DNNs, this process is largely ad-hoc and relies on costly trial-and-error solutions.

One generally accepted finding is the notion that deeper neural networks (more layers) typically perform better. This is likely due to the fact that more layers allow for the network to provide decision boundaries that are more non-linear and much more complex decisions. Current state of the art networks such as ResNet and Inception [53] often consist of hundreds of layers. As the community looks toward even deeper networks (10^5–10^9 layers) it is likely that we will see the emergence of a new type of neural networks—sparse neural networks [54]. As opposed to the commonly used dense equivalent, sparse deep neural networks will consist of a large number of weights that are zero. This can be quite advantageous for hardware platforms where memory or computations are resource constrained.

3.3.4 Transfer Learning

As described in [31], traditional machine-learning techniques assume that the data used to train and deploy a model are of the same domain. The framework of transfer learning allows the training of models in a source domain or feature space and the deployment of this model to a target domain in which sufficient training data may not be available and domain or feature space are different. Transfer learning can happen in different settings. The first, inductive transfer learning, occurs in cases where there are labels in the source and target domain; on the other hand, transductive transfer learning is used in cases where only source labels are available but target labels are not (similar to the unsupervised learning case), and unsupervised transfer learning is a case in which neither source nor target labels are available [31].

Transfer learning may be of particular interest to DoD/IC applications where training data within domains or tasks of interest is scarce. Thus, within the framework of transfer learning, it may be possible to find domains or tasks that are related to the target domain or task with abundant training information. Then, using inductive, transductive, or unsupervised transfer learning techniques, it may be possible to transfer the knowledge gained from the source domain and task to the target domain and task with significantly less effort than trying to train a model for the target domain and task from scratch. Although a promising avenue for applications such as those in the DoD and IC where we are data rich but truth (labeled) poor, a theoretical understanding of how well transfer learning works for arbitrary applications is still limited.

3.3.5 Semi-Supervised Learning

In many cases, one has access to a small set of labeled data but wishes to make additional use of large quantity of unlabeled data to improve the training of their models. Semi-supervised learning algorithms are a class of algorithms designed to work within this paradigm. The authors of [55] provide a good overview of the variety of techniques that fall within semi-supervised learning. Semi-supervised learning differs from supervised learning, which works on all labeled data and unsupervised learning in which none of the data is labeled. Further, it should be noted that semi-supervised learning has certain assumptions about the quality and relationship between labeled and unlabeled samples. For example, it is assumed that the labeled samples provide coverage over the possible classes and are related in some feature space to unlabeled samples that should have the same label. Thus, it is an important paradigm, but simply having access to a dataset with labeled and unlabeled samples is not necessarily sufficient to implementing a semi-supervised algorithm. As noted in [55], a few of the many ways in which semi-supervised learning can be used include:

1. **Generative models**. In this case, it is assumed that your dataset can be well represented by some sort of statistical mixture model. Then, if you have at least one labeled sample for each mixture component, it is possible to leverage the unlabeled samples to improve model quality. This process can also be extended to use any clustering algorithm to apply labels to the unlabeled samples.
2. **Self-training**. In this case, we use the labeled samples to create a classifier that can be then used to classify the unlabeled samples. High confidence classifications are then added to the labeled sample pool and a new classifier is trained. This process is repeated until satisfactory results are achieved
3. **Graph-based**. In this class of semi-supervised learning algorithms, data points (both labeled and unlabeled) are represented as a graph. Using techniques such as graph similarity, it is possible to infer labels on the unlabeled samples.

While at first glance it seems that semi-supervised learning may be a natural fit to DoD and IC problems by greatly reducing the burden on labeling samples, it should be noted that the use of unlabeled samples in conjunction with labeled samples provides no guarantee of improving supervised learning techniques on the labeled samples alone. Further, using semi-supervised learning typically involves significant manual design of features and models in order to achieve meaningful results [56].

3.3.6 Reinforcement Learning

Reinforcement learning is another machine-learning paradigm that has received significant interest in the recent past [57]. Major breakthroughs of reinforcement learning can be seen in robotics [58], learning to play Atari games [59], and improving computer performance in the game of Go [60].

In contrast to other learning paradigms, reinforcement learning leverages a reward signal in order to learn a model. Thus, this reward signal can provide much higher level "labels" that the system can eventually use to learn from and can work particularly well in complex environments where it may be difficult to tease out specific rules that need to be learned. An example of reinforcement learning in action is given in [61]. In this work, researchers use reinforcement learning techniques to develop an algorithm capable of controlling a helicopter. As the authors note, the difficulty in modeling the physical properties of a helicopter make it a good candidate for reinforcement learning.

There are a number of factors that make reinforcement learning different from other learning techniques. First, within reinforcement learning, there is no supervisor but only a reward signal. Second, the reward signal or feedback is not instantaneous but often delayed. Within reinforcement learning, signals are often provided sequentially and time or order of samples and signals are important. At each step, the response to subsequent samples differ. For the example of using reinforcement learning to learn to play Atari games, at the beginning, the rules of the game are unknown and the system learns directly from interactive game play. The system picks an action on the joystick and sees a set of pixels and scores that correspond to a positive or negative signal that is used to adjust behavior.

3.3.7 Common Algorithmic Pitfalls

When designing machine learning algorithms, developers should be aware of a number of common errors that can arise. Below are a few issues that we have observed in practice:

- **Over-fitting ("variance") vs. under-fitting ("bias")**. Over-fitting is a phenomenon when a particular machine-learning model is too closely fit to the training examples. When over-fit, an algorithm may have trouble generalizing to new examples. Under-fitting, on the other hand, is when an algorithm provides an overly simplistic model that describes the training examples. There are a number of ways that one can avoid such issues—for example, selecting training data that represents the variety of examples to be seen, including a regularization term in the training objective, or choosing different algorithms.
- **Bad/noisy/missing data**. In this challenge, a model is trained on bad, noisy, or missing data. In such cases, the trained model may not work as intended and decisions may be made on the wrong set of features. In the case of missing data, the model may ignore important features or make certain assumptions of the data that are unlikely to work in practice. Overcoming this challenge often requires the use of good data conditioning techniques of human intervention to ensure the fidelity of training data.

- **Model selection**. It is imperative that the model being used to represent the desired input-output relationship be closely related to the actual input-output relationship.
- **Lack of success metrics.** Having a clear metric of algorithmic success is important. While metrics such as accuracy or precision provide some view into the performance of the algorithm, there may be other metrics that should be carefully designed prior to training a model.
- **Linear vs. non-linear models**. Picking the right type of model is also important. If there is to be a linear relationship between inputs and outputs, one should pick a linear model.
- **Training vs. testing data**. Carefully segregating data into training, validation, and testing datasets is an important best practice that can help avoid issues such as over-fitting or under-fitting.
- **Computational complexity, curse of dimensionality**. Data conditioning techniques can be used to help reduce the dimensionality of datasets which can also help machine-learning algorithms use cleaner data when determining input-output relationships.

3.3.8 Algorithms within the Context of Exemplary Application

For the exemplary video classification example, we leverage a number of open-source models. In order to greatly reduce training time on the very large array, we use pre-existing models and weights and update them in a process similar to that outlined in the transfer learning section.

To begin the process, we tested many different types of models in order to judge the model that is likely to work well for the classification task at hand. Specifically, we focused on the following considerations:

1. Model type—spatial, temporal, relational or auditory
2. Training from scratch vs using pretrained models
3. Types of input channels
4. Computational constraints such as memory and computational performance

To simplify our problem, we considered each video as a series of images. Using this simplification, we were better able to focus on spatial models such as convolutional neural networks. Given the rich variety of preexisting models for image classification, we decided to use transfer learning from existing models such as Inception and ResNet. Given the simplification of videos to series of images, we only used the red, green, and blue channels of the video and ignored the audio tracks. Each node on our system consists of NVIDIA K80 GPUs with 16GB of memory, which is not enough to load the full training dataset. Thus, we developed a batch training mechanism that could iteratively train on a smaller subset of training samples rather than the full set. Determining other parameters such as number of layers, learning rate, and batch size was done by using high performance computing techniques in order to quickly look at the thousands of possible parameter settings.

3.4 Computing

Many recent advances in AI can be at least partly credited to advances in computing hardware [26, 62]. In particular, modern computing advances have been able to realize many computationally heavy machine-learning algorithms such as neural networks. While machine-learning algorithms such as neural networks have had a rich theoretic history [63], recent advances in computing have made the application of such algorithms a reality by providing the computational power needed to train and process massive quantities of data. While the computing landscape of the past decade has been rich with numerous innovations, DoD and IC applications that require covert and low size, weight, and power (SWaP) systems will need to look beyond the traditional architectures of central processing units (CPUs) and graphics processing units (GPUs). For example, in commercial applications, it is common to offload data conditioning and algorithms to non-SWaP constrained platforms such high-performance computing clusters or processing clouds. The DoD, on the other hand, may need AI applications to be performed inside low-SWaP platforms or local networks (edge computing) and without the use of the cloud due to insufficient security or communication infrastructure. Beyond modern computing platforms, the wide application of machine-learning algorithms has also been supported by the availability of open-source tools that greatly simplify developing new algorithms. In this section, we highlight some recent computing trends along with a brief introduction to software packages that have relevance to DoD and IC applications.

3.4.1 Processing Technologies:

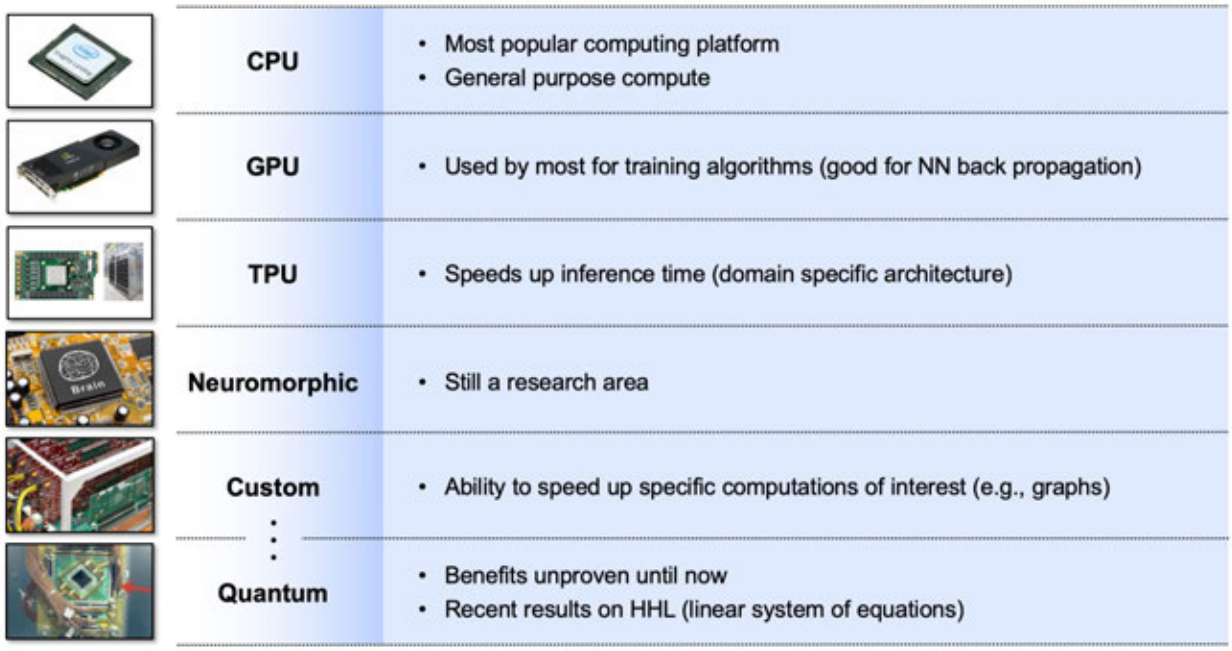


Figure 3.15. Some examples of processing technologies with salient features.

Figure 3.15 describes the primary classes of processing technologies for AI applications. While CPUs continue to dominate in terms of availability, cost, and market support, many of the recent advances in machine-learning algorithms, specifically neural networks, have been driven by GPUs. GPUs are essentially parallel vector processing engine, which have shown themselves to be adept at massively parallel problems such as training neural networks and are particularly well-suited to the back propagation algorithm described in Section 3.3.3. While GPUs will likely

dominate supervised learning model training in the near future, it is important to note that there are also a number of academic and commercial groups developing custom processors tuned for neural network inference and training. An example of such a processor is the Google TPU [62, 64], which is an application specific integrated circuit (ASIC) originally designed for machine-learning inference, while more recent versions support both inference and training [65]. There is also significant research in new hardware architectures such as neuromorphic computing [66], which may work well in low-power or resource-constrained environments.

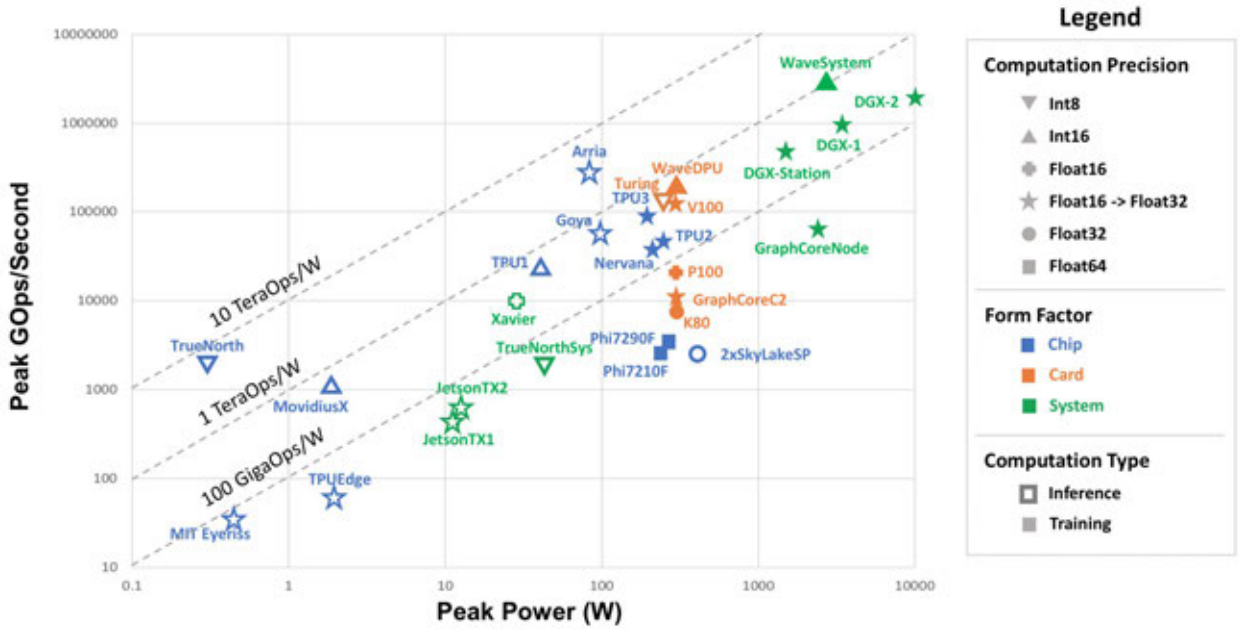


Figure 3.16. December 2018 view of AI computing systems. The x-axis indicates peak power and the y-axis indicate peak giga operations per second. (GOps/s) Note the legend on the right which indicates various parameters used to differentiate computing techniques.

Figure 3.16 graphs some of the recent processor capabilities (as of December 2018) mapping peak performance vs. power usage. As shown in the figure, much of the recent efforts have focused on processors that are in the 10–300W range in terms of power utilization, since they are being designed and deployed as processing accelerators. (300W is the upper limit for a PCI-based accelerator card.) For this power envelope, the performance can vary depending on a variety of factors such as architecture, precision, and workload (training vs. inference). At present, CPUs and GPUs continue to dominate the computing landscape for most artificial intelligence algorithms. However, the end of Moore's law [67] and Dennard scaling [68] implies that traditional commercial-off-the-shelf (COTS) technologies are unlikely to scale at the rate at which computational requirements are scaling. For example, according to [69], the amount of computation required for training popular neural network models goes up at a rate of approximately 10×/year. To address these challenges, a number of hardware developers have begun to develop customized chips based on FPGA or ASIC technologies. For example, the aforementioned Google TPU [70], Wave Computing Dataflow Processing Unit (DPU) [71], GraphCore C2 [72], and Habana Goya [73] are all ASICs customized for tensor operations such as parallel multiplications and additions. Other ASICs have been designed to push the boundaries of low-power neural network inference, including the IBM TrueNorth neuromorphic spiking neural network chip [66] and the MIT Eyeriss architecture [74]. One interesting trend of note is that many hardware manufacturers, faced with limitations in fabrication processes, have been

able to exploit the fact that machine-learning algorithms such as neural networks can perform well even when using limited or mixed precision [75, 76] representation of activation functions, weights, and biases. Such hardware platforms (often designed specifically for inference) may quantize weights and biases to half precision (16 bits) or even single bit representations in order to improve the number of operations/second without significant impact to model prediction, accuracy, or power utilization. For example, as reported by NVIDIA, the V100 GPU can perform 7.8 teraFLOPS of double-precision, 15.7 teraFLOPS of single-precision, and 125 teraFLOPS of mixed single- and half-precision [75].

While looking at published numbers from vendors provides an important view of performance, these numbers may be derived from algorithms and applications that are particularly well suited to the hardware platform. In most cases, it is also valuable to benchmark performance by testing various hardware platforms on workloads or applications of interest. For example, Figure 3.17 shows the time taken to train a dense convolutional neural network model using different hardware platforms. As expected, the NVIDIA K80 GPU performs the best when compared against the time taken for training by the Intel Xeon-E5 and Intel Knights Landing processor.

Beyond performance, power utilization, or other resource constraints, choosing the right computing technology for an application may also be driven by the software being used and what hardware platforms are supported by this software.

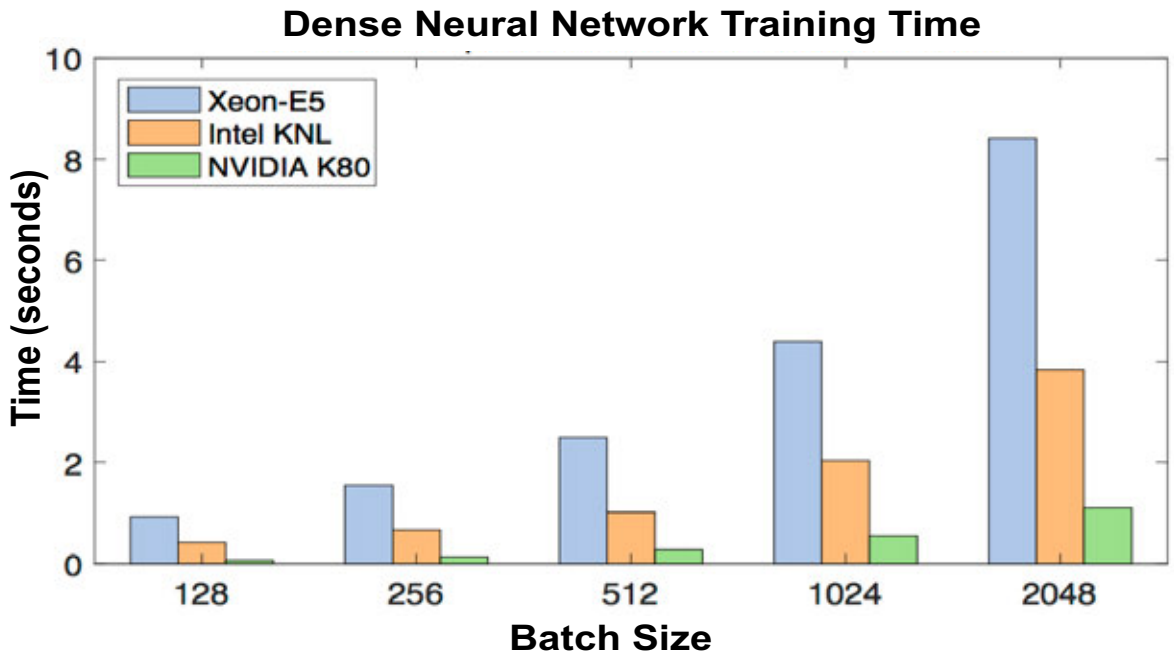


Figure 3.17. Performance results for training convolutional neural network using different hardware platforms as a function of batch size. Evaluation was performed using TensorFlow and training the AlexNet model with data from the ImageNet dataset.

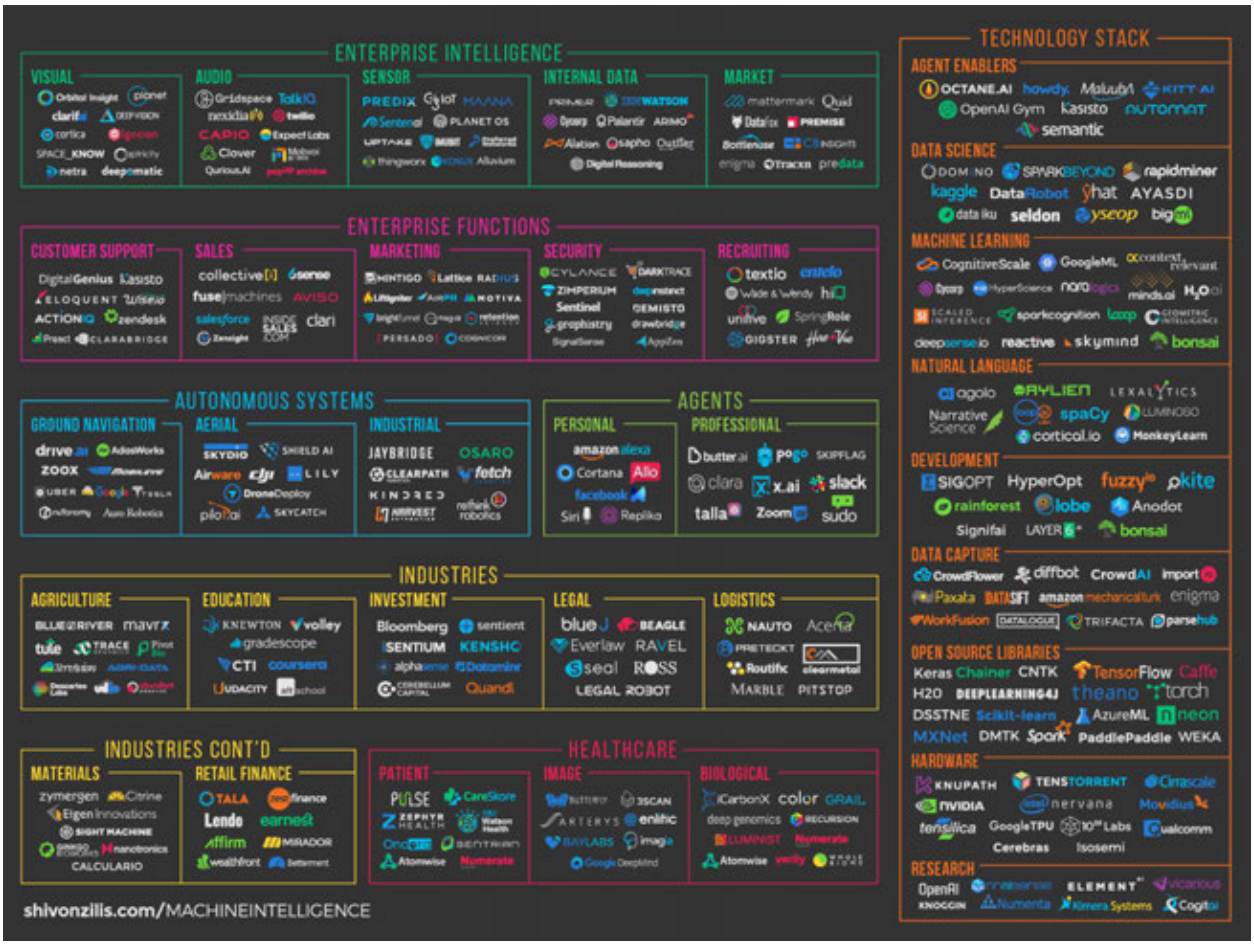


Figure 3.18. Survey of commercial organizations developing AI software tools. This infographic was developed by Shivon Zilis.

3.4.2 Machine Learning Software

The field of available software for AI and machine learning has witnessed an explosion of options in the past few years. For example, Figure 3.18 shows some of the many tools available. These tools typically provide a high-level domain-specific interface that allows users to quickly apply AI and machine learning techniques to problems and domains of interest.

As noted in the previous section, choosing a software environment for development and hardware platform for deployment are not necessarily independent decisions. Particular environments may only work with a subset of hardware platforms (or at least work well with a subset). A recent industry trend has been for well-known machine-learning software providers to develop custom hardware tuned to work well for their software environment. For example, software packages such as TensorFlow, PyTorch, MXNet, RAPIDS, CNTK are led by Google, Facebook [45], Amazon [77], NVIDIA, and Microsoft [78], respectively. Each of these vendors has also publicized the fact that they are developing hardware platforms (most often focused on inference) that will provide benefits to the users of their software packages. These benefits could include inexpensive use of proprietary cloud-based computing solutions, higher performance or additional software functionality. When deciding which of the multitude of AI and machine-learning frameworks to use for an application, it is important to understand how they fit in to the larger computing pipeline. Further, it should be noted that most of these tools output models in a

format unique to that software package. Fortunately, there are efforts such as the Open Neural Network Exchange (ONNX) to standardize model specifications and allow developers to move models from one ecosystem to another. However, as of December 2018, the portability of models across platforms is still limited. Finally, deploying individual software packages in private clouds or high performance computing (HPC) clusters still requires significant efforts to reach published performance numbers.

3.4.3 High Performance Computing

HPC systems play an important role in developing AI systems. Often, HPC systems are used for data conditioning and algorithm development. In the realm of data conditioning, HPC systems are particularly well designed for processing massive datasets in parallel by providing high-level interfaces and high-quality hardware. In the realm of algorithm development, HPC systems can be used for model design and training computationally heavy models such as neural networks. The parallel processing capabilities and abundant storage present in most HPC systems make them particularly well-tuned to such computationally heavy tasks that may require large sweeps of model parameter spaces. However, there are some potential pitfalls with using HPC systems for developing AI systems including internode communication overhead, data distribution, and parallelizing computations [79]. HPC systems can be used to train and optimize AI algorithms and models by evaluating many parameters (parallel hyperparameter training) [80, 81] or by training single algorithms and models on many compute nodes of an HPC system [82-84].

Many modern HPC systems such as the Oak Ridge National Laboratory's Summit system feature hybrid architectures consisting of heterogenous computing elements such as CPUs, GPUs, and FPGAs. While HPC systems have a long history, the iterative nature of AI and machine learning development has led to new software and tools. In our research at MIT LL, we have spent significant effort in developing new tools [85, 86] that enable interactive, on-demand rapid prototyping capabilities for AI and machine-learning practitioners interested in applying the power of HPC systems to AI and machine-learning workloads. Table 1 describes a number of salient differences between traditional HPC and AI/machine-learning application development and execution. The new MIT LL tools bridge these differences, and they enable both HPC and AI/machine-learning application development and execution on the same system and in the same user environments.

Table 1: Salient differences between designing high performance computing systems for traditional workloads vs. machine-learning workloads

	Traditional Workloads	AI/Machine-Learning Workloads
Programming Environments	OpenMP, MPI, C, C++, Fortran	Tensorflow, Caffe, Python, MATLAB, Julia
Software	Designed for clusters	Designed for laptop
Deployment	Bare Metal	VMs/Containers
Computing	Homogenous	Heterogenous
Computing Literacy	High	Low
Scheduling	Batch	Interactive
Resource Managers	Slurm, SGE	Mesos, YARN, Slurm
Data Sources	Generated by simulation	Imported from outside
Data Storage	Files	Databases (in-memory, file-based)
Interfaces	Terminal	Jupyter/Web-based

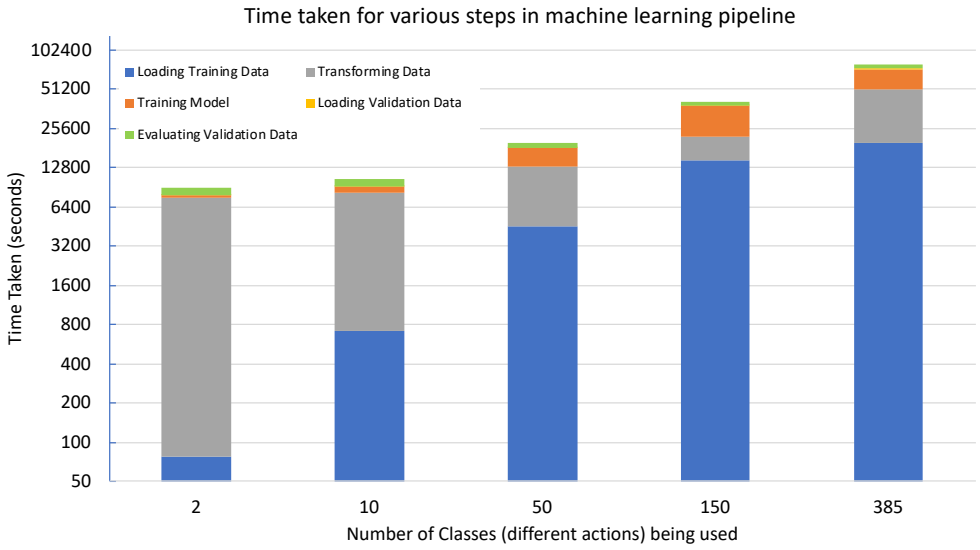


Figure 3.19. Time taken for different steps on exemplary problem's machine-learning pipeline (note log scale on y-axis).

3.4.4 Computing within the Context of Exemplary Application

For our particular application, we use Nvidia K80s as the computing platform. As described in Figure 3.17, the K80 GPU performs well when compared to other hardware platforms that

were available on our system (Intel Xeon-E5 and Intel KNL/Xeon64c). While GPUs can perform much of the model learning, we still utilize a CPU (Intel Xeon-E5) for data preprocessing and manipulation. The reason we use two different processors is three-fold: 1) our cluster has many more available CPUs that allow us to use parallel processing techniques for data conditioning, 2) each CPU has significantly more memory available (256 GB for a CPU vs. 16 GB for a GPU) that makes it amenable to manipulating large datasets, and 3) preprocessing is not amenable to vector-parallel processing. To fit datasets in the limited GPU memory for training, we used a batch training technique such as that utilized in [26] so that weight updates occur on a smaller batch of data when compared against the nearly one million videos. For the software environment, we used the widely available open-source TensorFlow and Keras [87] packages developed by Google that support the CPUs and GPUs available on our cluster. Development was performed on the MIT SuperCloud [88, 89] cluster at the Lincoln Laboratory Supercomputing Center.

Figure 3.19 describes the computational performance of various steps in the machine-learning pipeline: 1) loading training data, 2) transforming data, 3) training model, and 4) loading validation data. Training a particular model took approximately 11 hours and we were able to leverage HPC techniques in order to simultaneously test a number of competing parameters such as different learning rates and batch sizes for the neural network.

3.5 Robust Artificial Intelligence

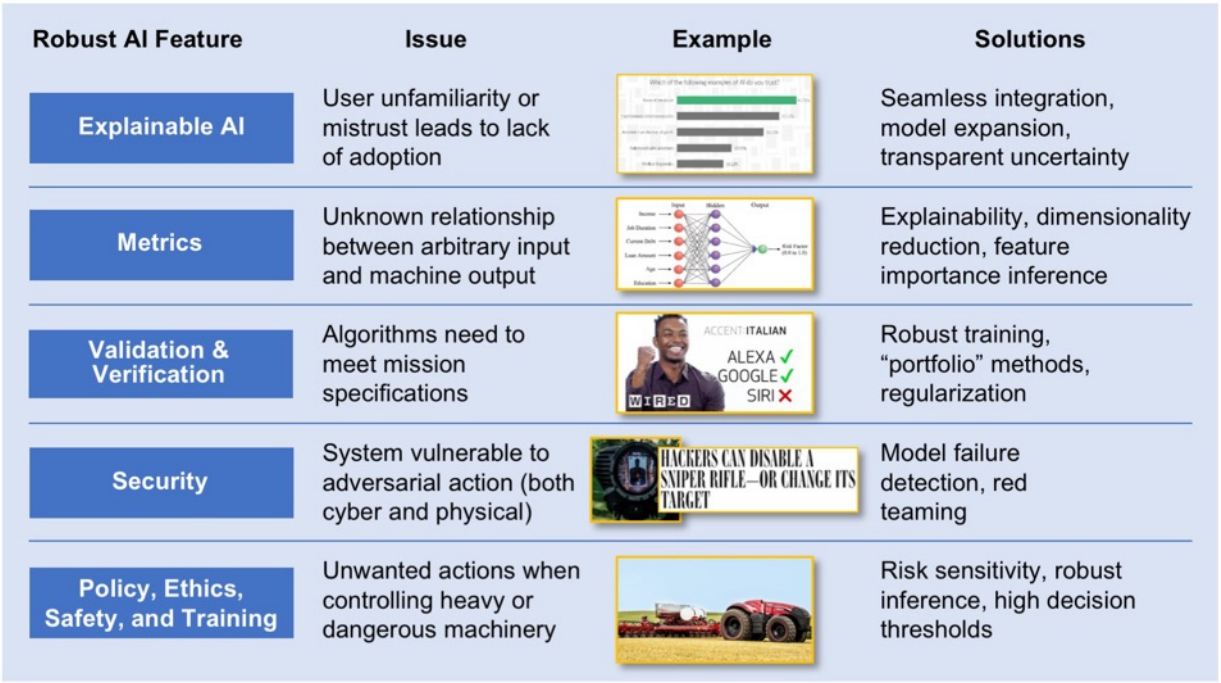


Figure 3.20. Importance of Robust AI.

A growing research area critical to the widespread deployment of AI solutions to DoD and IC problems is in the domain of robust AI. We use the term robust AI as a general term that includes explainability, verification/validation, metrics, security (cyber and physical), policy, ethics, safety, and training. Sometimes, this component is also referred to as trusted AI or adversarial AI (not to be confused with generative adversarial networks—a machine-learning technique). In this section, we describe each of the robust AI features of Figure 3.20. As this

field continues to evolve at a rapid pace, we focus on describing the challenges and highlight a few related research results as applicable.

3.5.1 Explainable AI

A key aspect of gaining trust in an AI system is in the ability of the system to explain the reasoning behind a particular decision. In particular for DoD/IC applications where AI systems may be supporting complex decisions or decisions with very high impact, it is imperative that users understand how a particular decision was reached by the system. This explanation can help users not only understand the reasoning behind an output, but also may help with the resiliency of algorithms. It is more difficult for an adversary to manipulate a model that needs to explain what it is doing.

Traditional AI systems that rely on expert or knowledge-based systems had the advantage of often being inherently explainable. Since the rules were created by experts and used tools such as decision trees, the AI system could simply output the states that were activated in reaching a decision, and the human user, ostensibly trained to use the system, could simply look at what was activated.

With more recent machine-learning algorithms, however, this is not necessarily the case. In fact, the ability of a machine-learning algorithm to combine multiple features in often non-linear ways is part of the value associated with them! Standard approaches of simply outputting the accuracy or confidence may work for domains in which the cost of an error is low but is unlikely to be sufficient for decision makers in critical applications [90]. While there are certain machine-learning algorithms such as Bayesian networks that are more amenable to explainability, this property is not necessarily present in most machine-learning techniques. For example, while there is ongoing research [74] looking at the explainability of neural networks, the explainability of most off-the-shelf models is poor. To underscore the importance of this field, the Defense Advanced Research Projects Agency (DARPA) recently began a program associated with explainable AI called XAI [91].

Closely related to explainability is the concept of interpretability. One should be careful with the terms “explainable” and “interpretable” (along with “comprehensible” and “understandable”) that are overloaded and often conflated in the AI literature. “Explainable” AI provides an explanation for the AI’s recommendation in terms a human can understand, even though the explanation might not fully describe how the AI arrived at its recommendation. The model or process that an AI uses to make its recommendation is said to be “interpretable” if a human can understand it. As examples, consider a neural network and a decision tree. The neural network is typically regarded as an opaque black box whose processing cannot be understood, so it is neither explainable nor interpretable. In contrast, the decision tree follows an explicit sequence of logical steps to make its recommendations, so it is interpretable and hence explainable. One approach to making a neural network explainable is to train another decision tree on examples of the neural network’s inputs and outputs and use the new decision tree to approximately describe what the neural network is doing.

For most applications of interest, one will need to develop algorithms that are explainable as well as interpretable. In many cases, providing the explanation of how an algorithm reached a particular decision may only be as good as how well that explanation can be interpreted by the end user. An overly complicated explanation may exceed the capacity of a human end user or domain expert to understand.

3.5.2 Metrics

In discussing metrics, we differentiate between component-level metrics and system-level metrics. Much of the presentation of AI or machine-learning results is done via component-level

metrics that provide results on how well a particular component of the AI architecture performed. For example, one may present the accuracy or precision of an algorithm but this does not indicate how well data conditioning was performed or how much of an impact this made to the overall mission.

Measuring the output of a machine-learning algorithm depends heavily on the task at hand. Within the realm of supervised learning classification, some common measurements include the true positive rate (the number of correct "positive" classifications), the true negative rate (the number of correct "negative" classifications), the false positive rate (the number of incorrect "positive" classifications), and the false negative rate (the number of incorrect "negative" classifications). These measures are often combined to report metrics such as accuracy (the ratio of correct predictions to the total number of samples), precision (the ratio of true positives to true positives and false positives), and recall (the ratio of true positives to true positives and false negatives). Higher level metrics such as the F-score can be used to represent the ratio of precision to recall for example.

The quality of regression can be measured by metrics such as residuals, which measure the algorithms output and compare them with the actual outputs. These residuals can be used to compute metrics such as mean absolute error (the average of absolute values of residuals) or mean square error (the average of the squared values of residuals). In the literature, it is also common to see metrics such as mean absolute percentage error or R^2 error.

In the case of unsupervised learning tasks such as clustering, internal structure may be measured by parameters such as the ratio of intra-cluster distances vs. inter-cluster distances, distances between cluster centroids, or mutual information. Other measurements may leverage external information such as ground-truth of cluster labels or known cluster structure [92].

While there are good component by component metrics for the AI canonical architecture, there is a major gap in end-to-end metrics. Thus, one may be able to measure the effectiveness of data conditioning or algorithms or computing but understanding how all of the components work together is a major gap in the presentation of metrics. For example, it is uncertain how one would present the overall impact to a mission by using an AI system. Of course, such metrics would likely need to tie in closely to the mission at hand.

3.5.3 Security

Security researchers look at the confidentiality, integrity, and availability of systems when evaluating threats and defenses [93]. In the same vein, AI researchers will need to evaluate the functioning of their AI system under adversarial conditions. At the core of using AI for important applications is the trust behind the system. Beyond obvious issues such as accuracy and precision of a particular AI pipeline, one must also look at issues that may arise when these systems are used in adversarial settings.

AI applications are prone to numerous attacks that can change the output often in unpredictable ways. For example, an adversary may physically manipulate an image or video that leads to incorrect classification or reduction in confidence in an algorithm. Adversaries may also be able to introduce bias into training data or manipulate sensors collecting data through cyber attacks.

The threat surface of the AI pipeline can be vast and Figure 3.21 describes some of the dimensions that may be used by security researchers in understanding where an attack on their AI system may occur. It should be noted that these dimensions are meant simply to be guidelines and not seen as rigid definitions of the dimensions associated with AI security. Further, the dimensions presented are not perfectly orthogonal to each other, and there are relationships that exist across dimensions.

The first dimension of Figure 3.21—Access—corresponds to what information or knowledge about the AI system an adversary may have access to. This dimension is meant to indicate that an adversary with only limited knowledge and access to an AI system they wish to compromise will be limited in the types of adversarial attacks they can perform. For example, given knowledge of the machine-learning model or architecture used, an adversary may deliberately poison training data such that, under certain circumstances, the model outputs an incorrect classification. There are sophisticated and relatively simple ways of doing this. A simple example of such a data poisoning attack would be to physically manipulate datasets. For example, the authors in [94] were able to fool a machine-learning algorithm into misclassifying a stop sign as a speed limit sign by simply placing a sticker on a stop sign. More sophisticated adversaries can also leverage detailed information such as model architecture, weights, and training tools that may be readily available based on knowledge of public-domain models.

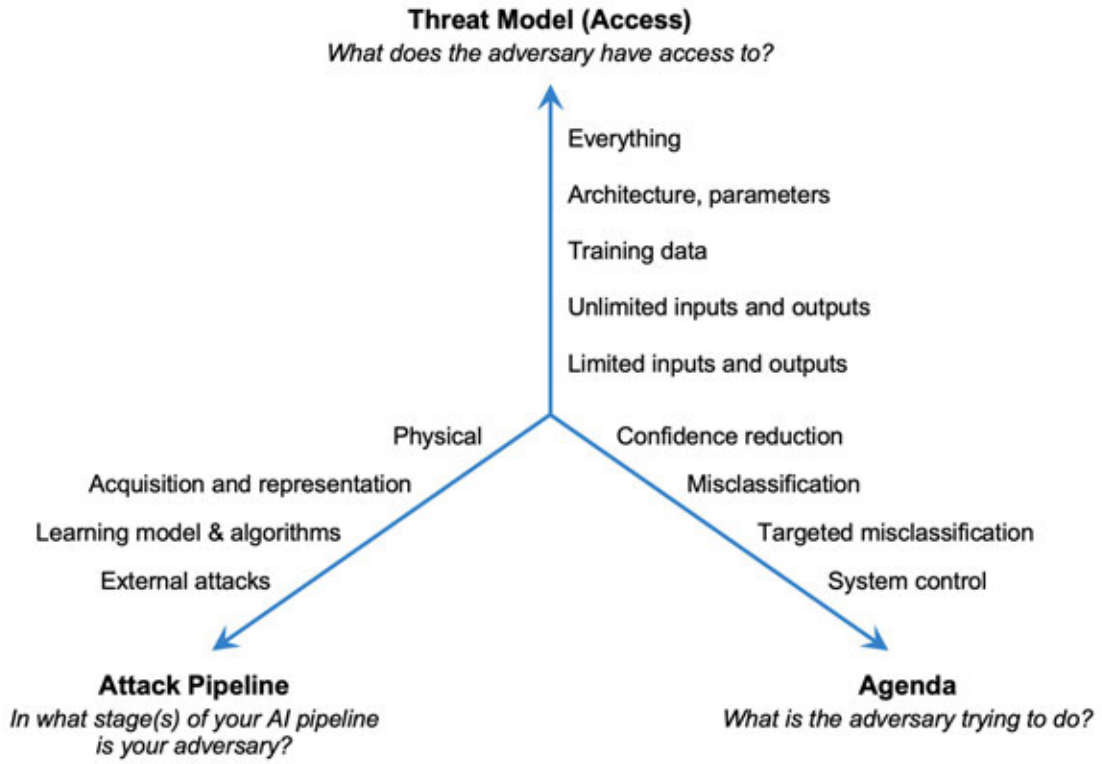


Figure 3.21. Dimensions of adversarial AI (three A's): access, agenda, attack pipeline.

In our framework, the second dimension—Agenda—corresponds to what an adversary is trying to achieve with their attack. In a simple case where an adversary is simply trying to reduce the confidence of an algorithm, they may only need access to samples or training data. For more sophisticated attacks that aim for system control, an adversary may need access to many other components or have in-depth knowledge of how a system was designed.

The final dimension in our framework—Attack Pipeline—is meant to indicate that there are different places in the AI pipeline where an attack may occur. Clearly, the types of attacks that can occur on the acquisition and representation of data are different than attacks on the learning model and algorithms. For example, an adversary may try to manipulate a sensor collecting data or attempt to manipulate the weights used in making a determination of a threat.

Again, the presentation of these dimensions is not meant to imply that each of these dimensions are mutually orthogonal or unrelated. Rather, these dimensions are meant to indicate

that there are different ways in which an adversary may attempt to infiltrate a system and that there are different capabilities needed to protect against these attacks. There are a number of well-studied and publicized examples of machine-learning algorithms being fooled by seemingly simple attacks [95-97]. Developing algorithms that are robust to such attacks is a very active area of research and there are numerous examples [98-101] of researchers developing counter-measures that can be applied depending on the threat. Beyond attacks specific to the AI pipeline, we should also note that AI system are also prone to a variety of cyber and physical vulnerabilities such as supply-chain risks.

Similar to many other areas of security research, it is likely that in the near future this field will continue to see a game of "cat-and-mouse" in which security researchers develop new examples of adversarial attacks and develop counter-measures to provide robust AI performance in the face of these attacks. For DoD and IC applications, it is imperative that developers work with security professionals to understand the types of adversarial attacks they may be prone to and develop counter-measures or techniques to minimize the effects of these attacks.

3.5.4 Other Robust AI Features

Other topics that may need to be studied or developed before widespread deployment of AI solutions to DoD/IC missions include validation and verification and policy, ethics, safety and training. Validation and verification techniques can be used to measure the compliance of various system features with specifications, rules, and conditions under which the system is intended to operate. While there is limited work in validating and verifying expert systems from a software perspective [102], there is very little current research on applying such techniques to AI systems that leverage more complex machine-learning algorithms such as neural networks.

Finally, widespread deployment of AI systems within the DoD and IC will largely be advanced or impeded by AI rules and regulations. While there are a number of technical challenges associated with developing such rules and regulations, there will also need to be a consistent effort across multiple agencies to develop best practices and share results.

3.5.5 Robust AI within the Context of Exemplary Application

Given the research nature of our exemplary application, we did not focus significant effort on security and adversaries. However, we did use simple techniques such as inspecting inputs, outputs, and selective pieces of the neural network to understand where the system was failing and for debugging errors. In measuring results, we use the top-1 and top-5 accuracies. This metric is defined as follows: An algorithm will label each of the videos with one of k labels. The top-k accuracy says that a video was correctly identified if one of its top k labels are the correct label. For example, a video may be classified (in decreasing probability) as: *(barking, yelling, running, ...)*. If the correct label (as judged by a human observer) is "yelling", this would contribute a correct classification toward the top-5 but would contribute a miss towards the top-1 accuracy. As of June 2018, leading models for the Moments in Time Dataset had top-1 accuracies of approximately 0.3 and top-5 accuracies of approximately 0.6 [3, 5].

3.6 Human-Machine Teaming

The final piece of the canonical AI architecture of Figure 3.1 is what we refer to as human-machine teaming. This piece is critical in connecting the AI system to the end user and mission. Human-machine teaming tasks are defined as tasks in which the human and machine system are interdependent in some fashion. Human-machine collaboration is a broader defined term that encompasses interdependent tasks, but also tasks that are not interdependent such as sequential or loosely coordinated tasks.

First of all, it is important to understand which tasks are mapped well to humans and which tasks are mapped well to machines. As shown in Figure 3.22, there is a spectrum that relate to how closely humans and machines work together. Borrowing the terminology from [103], we refer to broad collaboration relationships as human-in-the-loop, human-on-the-loop and human-out-of-the-loop. Human-in-the-loop collaboration is when a human is closely in the loop of the AI system. For example, a human and machine working together to jointly solve a common goal. In this relationship, the human and machine are equal contributors (or in certain cases, the human is a greater participant) to the overall system. The second type of relationship is referred to as human-on-the-loop. In this form of interaction, a human is largely participating in the system in a supervisory capacity. Human-on-the-loop systems will largely leverage automated techniques and may triage more important information to a human observer or the human may provide oversight on the functioning of the system. The final type of relationship is referred to as human-out-of-the-loop. In this relationship, the human does not participate in the AI system's operation under normal conditions.

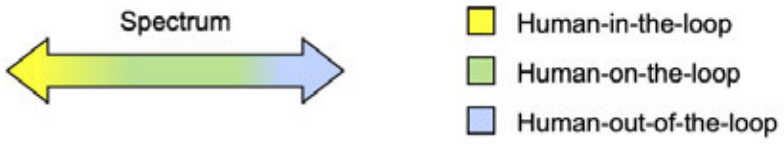


Figure 3.22. Spectrum of humans and machines interacting.

Clearly, different applications will have different requirements in terms of how closely humans and machine work together. In order to determine the appropriate level of human and machine interaction, Figure 3.23 describes a high-level framework that may assist in such a mapping. On the horizontal axis, we look at the consequence of actions—how important is it that the system provide the correct response. On the vertical axis, we look at the confidence in the machine making the decision. Clearly, for very consequential decisions that may impact lives in a significant way, such decisions are clearly mapped well to humans. On the other hand, for low consequence decisions in which we have high confidence in the machine, such decisions may be well mapped to machines. Within the context of the DoD and IC applications, in the near future, it is likely that AI systems will largely be used to augment human decision making. This is largely due to the high consequence of decisions. Certain tasks such as anomaly detection or highlighting important data may be performed by a system but high-consequence decisions such as whether to deploy troops or resources will likely still be performed by humans. Historically, systems have been designed with a static human/machine task allocation, but in current and future systems, task authority can dynamically change from human to machine depending on the context or the capabilities of the human or machine.

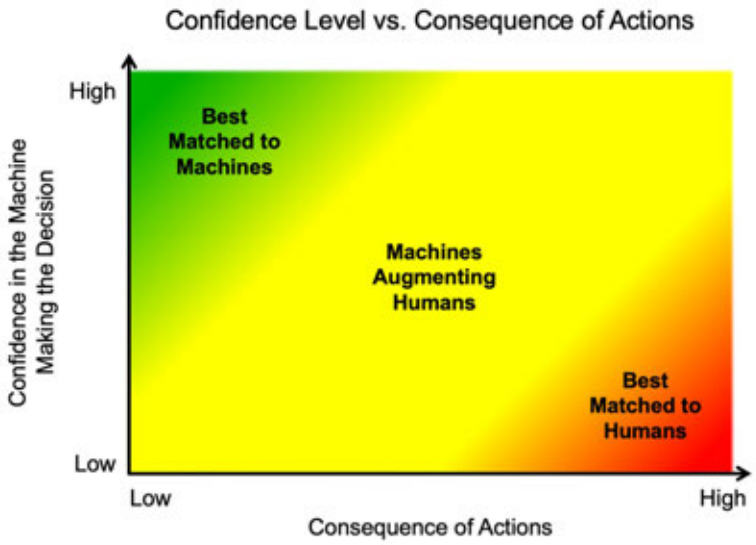


Figure 3.23. Determining which tasks map well to humans and machines.

For those systems in which the humans and machines must interact to be successful, it is important to provide the elements to enable an effective human-machine collaboration. In Figure 3.24, some of these elements are outlined. At the top are environmental elements that provide part of the context in which a collaboration occurs. There are static elements such as physics of the environment, which do not change dynamically. There are semi-static elements such as physical infrastructure including buildings, rivers, and forests. There are social constructs that provide limitations on and rules about how people interact with one another (and these could vary depending on what part of the globe one is in). Mission provides the other part of the context. Goals of the mission establish why the human-machine collaboration is taking place. Tasks outline how these goals can be accomplished. Procedures are organizationally determined structure on how the tasks are performed.

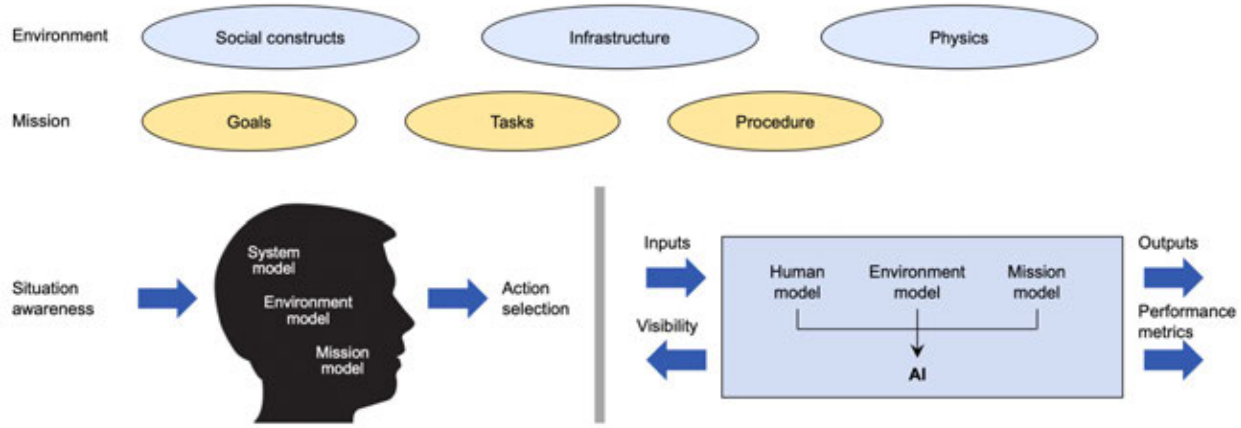


Figure 3.24. Elements of effective human-machine collaboration.

The humans then perceive information from the context and from the machine system. Using this perceived information, the human projects expected behavior for the system, the environment, and the mission into the future. With these predictions, course of action decisions can be made.

The system also receives input from the human and the environment. Ideally, the system would contain models of the human, environment, and mission in order to extrapolate expected behavior into the future. These predictions could then be used to make system behavior decisions. In reality, much of the human-machine collaboration issues stem from the fact that the human, environment, and mission models within the system are faulty or non-existent. Systems can also have more or less information visible to the human as it continues its tasks. As systems lean towards being adaptive AI, to ensure that the system model within the human is functionally accurate, more visibility into the system's processes should be provided. Likewise, the more information the system has about the human's evolving tasks and models, the better the system's decisions will be.

Johnson's research [104] has identified three aspects of the interaction between humans and machines that establish effective human-machine teaming: observability, predictability, and directability. Observability is akin to the "visibility" description above—can the human and machine see what the other is doing/planning to do? Predictability is the capability resulting from the models within the human/machine. If you have a model of the human or system or environment, you can then predict the future behavior. Directability is the means of being able to exert control on the human/system teammate. If one of the human's courses of action were to tell the system what to do (or vice versa), then the interaction would have the quality of directability. Often directability lies on a spectrum from low directability (infrequent or gross control) to high directability (continuous and fine control). Outstanding research questions remain on how much observability, predictability, and directability is required for a sufficient teaming interaction for different tasks. One interesting finding from Johnson's research is that as the system becomes more autonomous, MORE attention to human-machine collaboration, not less, is required to maintain complete system performance.

There are a number of research areas such as data visualization and algorithm interpretability that may also help move these boundaries. While there has been research on how humans and machines interact in domains such as robotics [105, 106], research on the interaction between humans and machines for more general AI applications is relatively limited. In the near future it seems that AI systems will need to be developed in close collaboration with end users in order to design how systems interact with the end users.

There are a number of techniques in the field of human-computer interaction that may be used as a starting point for researchers and developers. For example, principles of user design, usability and interface testing. New visualization research on augmented reality, 3D printing, immersive gaming environments, and brain-computer interfaces may also play a large part in developing novel human-machine teaming interfaces.

For the DoD and IC, the possibilities of humans and machine working together are tremendous. While certain tasks will likely continue to be human-only, there are a number of possibilities for applications human-in-the-loop and human-on-the-loop relationships. For example, in [103], the authors highlight AI roles such as training, interacting, and amplifying as prime candidates for either humans complementing machine intelligence or AI providing humans with superpowers (as described by P.R. Daugherty in Figure P2-1 on page 107, in *Human+ Machine: Reimagining Work in the Age of AI* [103]). Certain mundane, cumbersome, or trivial tasks may also be candidates for human-out-of-the-loop solutions.

3.7 Acknowledgements

The authors wish to thank the following individuals for their contributions, thoughts and comments towards the development of this section: Charlie Dagli, Lauren Milechin, Julia Mullen, Jonathan Herrera, David Martinez, Paul Monticciolo, Justin Goodwin, Nicolas Malyska, William Streilein, Arjun Majumdar, and Jonathan Su. Additional thanks to Laura Glazer and Brad Dillman for their editing support. Pieces of this section were coauthored by Dr. Albert Reuther and Dr. Hayley Reynolds.

3.8 References

1. Theis, T.N. and H.-S.P. Wong, *The end of moore's law: A new beginning for information technology.* Computing in Science & Engineering, 2017. **19**(2): p. 41-50.
2. Monfort, M., et al., *Moments in Time Dataset: one million videos for event understanding.* arXiv preprint arXiv:1801.03150, 2018.
3. Li, C., et al., *Team DEEP-HRI Moments in Time Challenge 2018 Technical Report.*
4. Li, Y., et al., *Submission to Moments in Time Challenge 2018.*
5. Xiaoteng, Z., et al., *Qiniu Submission to ActivityNet Challenge 2018.*
6. Press, G., *Cleaning big data: Most time-consuming, least enjoyable data science task, survey says.* Forbes, March, 2016. **23**.
7. Krizhevsky, A., I. Sutskever, and G. Hinton, *ImageNet classification with deep convolutional neural networks. Advances in Neural Information Processing Systems 25 (NIPS 2012)*. 2017.
8. LeCun, Y., C. Cortes, and C. Burges, *MNIST handwritten digit database.* AT&T Labs [Online]. Available: http://yann. lecun. com/exdb/mnist, 2010. **2**.
9. Krizhevsky, A., V. Nair, and G. Hinton, *The CIFAR-10 dataset.* online: http://www. cs. toronto. edu/kriz/cifar. html, 2014.
10. Kurin, V., et al., *The Atari Grand Challenge Dataset.* arXiv preprint arXiv:1705.10998, 2017.
11. Borgnat, P., et al. *Seven years and one day: Sketching the evolution of internet traffic*. in *INFOCOM 2009, IEEE*. 2009. IEEE.
12. Braam, P.J., *The Lustre storage architecture.* 2004.
13. Loney, K., *Oracle Database 11g The Complete Reference*. 2008: McGraw-Hill, Inc.
14. Stonebraker, M. and L.A. Rowe, *The design of Postgres*. Vol. 15. 1986: ACM.
15. Gilbert, S. and N. Lynch, *Brewer's conjecture and the feasibility of consistent, available, partition-tolerant web services.* Acm Sigact News, 2002. **33**(2): p. 51-59.
16. Gadepally, V., et al. *The bigdawg polystore system and architecture*. in *High Performance Extreme Computing Conference (HPEC), 2016 IEEE*. 2016. IEEE.
17. Chang, F., et al., *Bigtable: A distributed storage system for structured data.* ACM Transactions on Computer Systems (TOCS), 2008. **26**(2): p. 4.
18. Lakshman, A. and P. Malik, *Cassandra: a decentralized structured storage system.* ACM SIGOPS Operating Systems Review, 2010. **44**(2): p. 35-40.
19. Han, J., et al. *Survey on NoSQL database*. in *Pervasive computing and applications (ICPCA), 2011 6th international conference on*. 2011. IEEE.
20. Kepner, J., et al. *Achieving 100,000,000 database inserts per second using Accumulo and D4M*. in *High Performance Extreme Computing Conference (HPEC), 2014 IEEE*. 2014. IEEE.
21. Stonebraker, M., *Newsql: An alternative to nosql and old sql for new oltp apps.* Communications of the ACM. Retrieved, 2012: p. 07-06.
22. Tan, R., et al. *Enabling query processing across heterogeneous data models: A survey*. in *Big Data (Big Data), 2017 IEEE International Conference on*. 2017. IEEE.
23. Elmore, A., et al., *A demonstration of the bigdawg polystore system.* Proceedings of the VLDB Endowment, 2015. **8**(12): p. 1908-1911.
24. Mattson, T., et al. *Demonstrating the BigDAWG Polystore System for Ocean Metagenomics Analysis*. in *CIDR*. 2017. http://cidrdb.org/cidr2017/papers/p120-mattson-cidr17.pdf

25. Hodge, V. and J. Austin, *A survey of outlier detection methodologies.* Artificial intelligence review, 2004. **22**(2): p. 85-126.
26. Krizhevsky, A., I. Sutskever, and G.E. Hinton. *Imagenet classification with deep convolutional neural networks*. in *Advances in neural information processing systems*. 2012.
27. Zhou, B., et al. *Learning deep features for scene recognition using places database*. in *Advances in neural information processing systems*. 2014.
28. Lin, T.-Y., et al. *Microsoft coco: Common objects in context*. in *European conference on computer vision*. 2014. Springer.
29. Paolacci, G., J. Chandler, and P.G. Ipeirotis, *Running experiments on amazon mechanical turk.* 2010.
30. Sorokin, A. and D. Forsyth. *Utility data annotation with amazon mechanical turk*. in *Computer Vision and Pattern Recognition Workshops, 2008. CVPRW'08. IEEE Computer Society Conference on*. 2008. IEEE.
31. Ipeirotis, P.G., F. Provost, and J. Wang. *Quality management on amazon mechanical turk*. in *Proceedings of the ACM SIGKDD workshop on human computation*. 2010. ACM.
32. Dobkin, A. *DOD Maven AI project develops first algorithms, starts testing*. 2017; https://defensesystems.com/articles/2017/11/03/maven-dod.aspx.
33. Ralph Jacobson, *IBM Consumer Products Industry Blog,.* 2013. https://www.ibm.com/blogs/insights-on-business/consumer-products/2-5-quintillion-bytes-of-data-created-every-day-how-does-cpg-retail-manage-it/
34. Mundhenk, T.N., et al. *A large contextual dataset for classification, detection and counting of cars with deep learning*. in *European Conference on Computer Vision*. 2016. Springer.
35. Vondrick, C., et al., *Tracking emerges by colorizing videos.* arXiv preprint arXiv:1806.09594, 2018.
36. Lafferty, J., A. McCallum, and F.C. Pereira, *Conditional random fields: Probabilistic models for segmenting and labeling sequence data.* 2001.
37. Goodfellow, I., et al., *Deep learning*. Vol. 1. 2016: MIT press Cambridge.
38. Buchanan, B.G. and E.A. Feigenbaum, *DENDRAL and Meta-DENDRAL: Their applications dimension*, in *Readings in artificial intelligence*. 1981, Elsevier. p. 313-322.
39. Hayes-Roth, F., D.A. Waterman, and D.B. Lenat, *Building expert system.* 1983.
40. Waterman, D., *A guide to expert systems.* 1986.
41. Gonzalez, A.J. and D.D. Dankel, *The engineering of knowledge-based systems*. 1993: Prentice-Hall Englewood Cliffs, NJ.
42. Gadepally, V., et al. *Driver/vehicle state estimation and detection*. in *Intelligent Transportation Systems (ITSC), 2011 14th International IEEE Conference on*. 2011. IEEE.
43. Domingos, P., *The master algorithm: How the quest for the ultimate learning machine will remake our world*. 2015: Basic Books.
44. Peterson, L.E., *K-nearest neighbor.* Scholarpedia, 2009. **4**(2): p. 1883.
45. Paszke, A., et al., *Automatic differentiation in pytorch.* 2017.
46. Abadi, M., et al. *Tensorflow: a system for large-scale machine learning*. in *OSDI*. 2016. https://www.usenix.org/system/files/conference/osdi16/osdi16-abadi.pdf
47. Jia, Y., et al. *Caffe: Convolutional architecture for fast feature embedding*. in *Proceedings of the 22nd ACM international conference on Multimedia*. 2014. ACM.
48. Socher, R., et al. *Parsing natural scenes and natural language with recursive neural networks*. in *Proceedings of the 28th international conference on machine learning (ICML-11)*. 2011.
49. Hinton, G.E., *Deep belief networks.* Scholarpedia, 2009. **4**(5): p. 5947.
50. Karlik, B. and A.V. Olgac, *Performance analysis of various activation functions in generalized MLP architectures of neural networks.* International Journal of Artificial Intelligence and Expert Systems, 2011. **1**(4): p. 111-122.
51. Goodfellow, I., et al. *Generative adversarial nets*. in *Advances in neural information processing systems*. 2014.

52. Chan, M., et al. *Learning Network Architectures of Deep CNNs under Resource Constraints*. in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 2018.
53. Szegedy, C., et al. *Inception-v4, inception-resnet and the impact of residual connections on learning*. in *AAAI*. 2017.
54. Gadepally, V., J. Kepner, and A. Reuther, *Storage and database management for big data.* Big Data: Storage, Sharing and Security. CRC Press. Retrieved on July, 2016. **30**: p. 2017.
55. Zhu, X., *Semi-supervised learning literature survey.* Computer Science, University of Wisconsin-Madison, 2006. **2**(3): p. 4.
56. Zhu, X., *Semi-supervised learning*, in *Encyclopedia of machine learning*. 2011, Springer. p. 892-897.
57. Sutton, R.S. and A.G. Barto, *Reinforcement learning: An introduction*. 2018: MIT press.
58. Kober, J. and J. Peters, *Reinforcement learning in robotics: A survey*, in *Reinforcement Learning*. 2012, Springer. p. 579-610.
59. Mnih, V., et al., *Playing atari with deep reinforcement learning.* arXiv preprint arXiv:1312.5602, 2013.
60. Silver, D., et al., *Mastering the game of Go without human knowledge.* Nature, 2017. **550**(7676): p. 354.
61. Abbeel, P., et al. *An application of reinforcement learning to aerobatic helicopter flight*. in *Advances in neural information processing systems*. 2007.
62. Jouppi, N., *Google supercharges machine learning tasks with TPU custom chip.* Google Blog, May, 2016. **18**.
63. Minsky, M.L., *Computation: finite and infinite machines*. 1967: Prentice-Hall, Inc.
64. Jouppi, N.P., et al., *A domain-specific architecture for deep neural networks.* Communications of the ACM, 2018. **61**(9): p. 50-59.
65. Dean, J., D. Patterson, and C. Young, *A New Golden Age in Computer Architecture: Empowering the Machine-Learning Revolution.* IEEE Micro, 2018. **38**(2): p. 21-29.
66. Merolla, P.A., et al., *A million spiking-neuron integrated circuit with a scalable communication network and interface.* Science, 2014. **345**(6197): p. 668-673.
67. Schaller, R.R., *Moore's law: past, present and future.* IEEE spectrum, 1997. **34**(6): p. 52-59.
68. Frank, D.J., et al., *Device scaling limits of Si MOSFETs and their application dependencies.* Proceedings of the IEEE, 2001. **89**(3): p. 259-288.
69. OpenAI. *AI and Compute*. https://blog.openai.com/ai-and-compute/.
70. Jouppi, N.P., et al. *In-datacenter performance analysis of a tensor processing unit*. in *Computer Architecture (ISCA), 2017 ACM/IEEE 44th Annual International Symposium on*. 2017. IEEE.
71. Hemsoth, N. *Startup's AI Chip Beats GPU*. 2017 Aug. 23, 2017; https://www.nextplatform.com/2017/08/23/first-depth-view-wave-computings-dpu-architecture-systems/.
72. Lacey, D. *Preliminary IPU Benchmarks*. 2017; https://www.graphcore.ai/posts/preliminary-ipu-benchmarks-providing-previously-unseen-performance-for-a-range-of-machine-learning-applications.
73. Merritt, R. *Startup's AI Chip Beats GPU*. 2018 Sept. 17, 2018; https://www.eetimes.com/document.asp?doc_id=1333719.
74. Chen, C., et al., *This looks like that: deep learning for interpretable image recognition.* arXiv preprint arXiv:1806.10574, 2018.
75. Micikevicius, P., et al., *Mixed precision training.* arXiv preprint arXiv:1710.03740, 2017.
76. Gupta, S., et al. *Deep learning with limited numerical precision*. in *International Conference on Machine Learning*. 2015.
77. Chen, T., et al., *Mxnet: A flexible and efficient machine learning library for heterogeneous distributed systems.* arXiv preprint arXiv:1512.01274, 2015.
78. Seide, F. and A. Agarwal. *CNTK: Microsoft's open-source deep-learning toolkit*. in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2016. ACM.

79. Keuper, J. and F.-J. Preundt, *Distributed training of deep neural networks: theoretical and practical limits of parallel scalability*, in *Proceedings of the Workshop on Machine Learning in High Performance Computing Environments*. 2016, IEEE Press: Salt Lake City, Utah. p. 19-26.
80. Domhan, T., J.T. Springenberg, and F. Hutter. *Speeding Up Automatic Hyperparameter Optimization of Deep Neural Networks by Extrapolation of Learning Curves*. in *IJCAI*. 2015.
81. Lorenzo, P.R., et al. *Particle swarm optimization for hyper-parameter selection in deep neural networks*. in *Proceedings of the Genetic and Evolutionary Computation Conference*. 2017. ACM.
82. Goyal, P., et al., *Accurate, Large Minibatch SGD: Training ImageNet in 1 Hour*. 2017.
83. Kurth, T., et al. *Deep learning at 15pf: supervised and semi-supervised classification for scientific data*. in *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*. 2017. ACM.
84. You, Y., A. Buluç, and J. Demmel. *Scaling deep learning on GPU and knights landing clusters*. in *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*. 2017. ACM.
85. Gadepally, V., et al. *D4m: Bringing associative arrays to database engines*. in *High Performance Extreme Computing Conference (HPEC), 2015 IEEE*. 2015. IEEE.
86. Prout, A., et al., *MIT SuperCloud portal workspace: Enabling HPC web application deployment.* arXiv preprint arXiv:1707.05900, 2017.
87. Chollet, F., *Keras*. 2015.
88. Prout, A., et al., *Enabling on-demand database computing with MIT SuperCloud database management system.* arXiv preprint arXiv:1506.08506, 2015.
89. Reuther, A., et al., *LLSuperCloud: Sharing HPC systems for diverse rapid prototyping*. 2013: 2013 IEEE High Performance Extreme Computing Conference (HPEC). p. 1-6.
90. Biran, O. and C. Cotton. *Explanation and justification in machine learning: A survey*. in *IJCAI-17 Workshop on Explainable AI (XAI)*. 2017.
91. Gunning, D., *Explainable artificial intelligence (xai).* Defense Advanced Research Projects Agency (DARPA), nd Web, 2017.
92. Wang, K., B. Wang, and L. Peng, *CVAP: validation for cluster analyses.* Data Science Journal, 2009. **8**: p. 88-93.
93. Fuller, B., et al. *Sok: Cryptographically protected database search*. in *Security and Privacy (SP), 2017 IEEE Symposium on*. 2017. IEEE.
94. Evtimov, I., et al., *Robust physical-world attacks on machine learning models.* arXiv preprint arXiv:1707.08945, 2017. **2**(3): p. 4.
95. Carlini, N. and D. Wagner. *Adversarial examples are not easily detected: Bypassing ten detection methods*. in *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*. 2017. ACM.
96. Kurakin, A., I. Goodfellow, and S. Bengio, *Adversarial examples in the physical world.* arXiv preprint arXiv:1607.02533, 2016.
97. Moosavi-Dezfooli, S.-M., A. Fawzi, and P. Frossard. *Deepfool: a simple and accurate method to fool deep neural networks*. in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016.
98. Lu, J., T. Issaranon, and D.A. Forsyth. *SafetyNet: Detecting and Rejecting Adversarial Examples Robustly*. in *ICCV*. 2017.
99. Meng, D. and H. Chen. *Magnet: a two-pronged defense against adversarial examples*. in *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*. 2017. ACM.
100. Papernot, N., et al. *Distillation as a defense to adversarial perturbations against deep neural networks*. in *2016 IEEE Symposium on Security and Privacy (SP)*. 2016. IEEE.
101. Shaham, U., Y. Yamada, and S. Negahban, *Understanding adversarial training: Increasing local stability of neural nets through robust optimization.* arXiv preprint arXiv:1511.05432, 2015.
102. Vermesan, A. and F. Coenen, *Validation and Verification of Knowledge Based Systems: Theory, Tools and Practice*. 2013: Springer Science & Business Media.

103. Daugherty, P.R. and H.J. Wilson, *Human+ Machine: Reimagining Work in the Age of AI*. 2018: Harvard Business Press.
104. Johnson, M., et al., *Coactive design: Designing support for interdependence in joint activity.* Journal of Human-Robot Interaction, 2014. **3**(1): p. 43-69.
105. Goodrich, M.A. and A.C. Schultz, *Human–robot interaction: a survey.* Foundations and Trends® in Human–Computer Interaction, 2008. **1**(3): p. 203-275.
106. Parasuraman, R., et al., *Adaptive automation for human-robot teaming in future command and control systems*. 2007, Army research lab aberdeen proving ground md human research and engineering directorate.

4 AI Applied to Human Language Technology (N. Malyska)

4.1 Background

Many technologies researched, developed, and tested at MIT LL are in the area of information sciences. As illustrated in Figure 4.1, MIT LL makes significant contributions to this area across ISR, human language technology (HLT), bioengineering, informatics and decision support, cyber, and supercomputing areas.

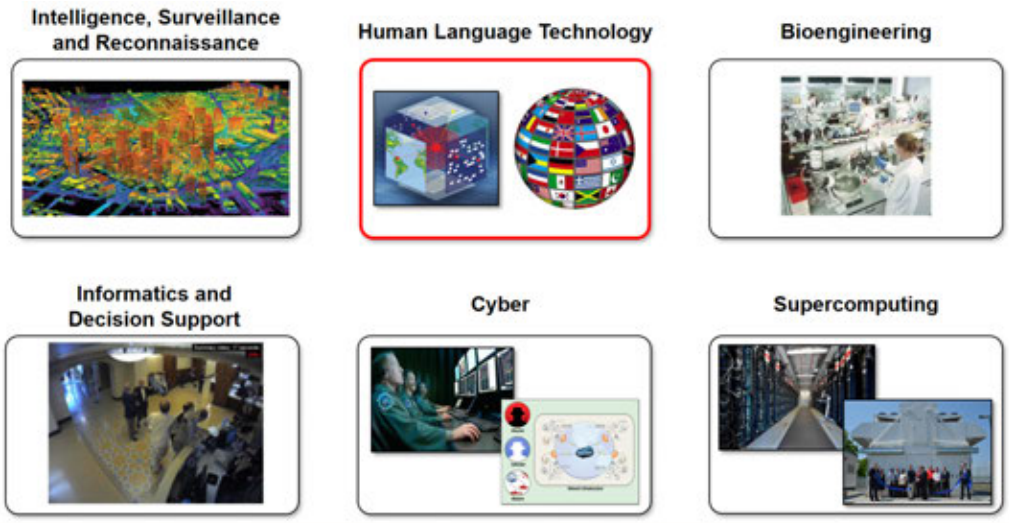


Figure 4.1. Information science technology areas at MIT LL.

While all of these areas are important, work in HLT is a long-standing area of contributions in machine learning and narrow AI.

4.2 Early Work to Recent Developments in AI

HLT work—in text and other forms of communication, such as Morse code—has been conducted at MIT LL since the 1950s. For example, in the mid-50s, Ben Gold, one of the primary founders of the later MIT LL speech group, designed a Morse code translation machine. More broadly, he was involved in pattern recognition under the leadership of Oliver Selfridge, who led "one of the first groups in Artificial Intelligence". [1]

The application of AI to speech began in the 1980s when large datasets became available. For example, TI and MIT released the TIMIT speech database of read sentences in 1986. TIMIT [2] and other datasets like it allowed researchers to conduct experiments on high-quality, carefully curated data. These datasets were disseminated widely across the community and became the basis for shared benchmarks in speech, speaker, and language recognition, as well as other emerging areas.

The MIT LL Speech Systems Technology group was a lead in the resulting machine-learning algorithms development boom, with staff creating capabilities in emerging HLT technology areas. Early milestones included the DARPA Neural Network Study [3] and the LNKnet machine-learning software toolkit [4]. In 1996, the group pioneered Gaussian mixture model (GMM)-UBM-based speaker verification [5] and in the 2000s was a leading group in the use of support vector machines (SVM) for speaker verification [6]. Figure 4.2 shows two milestone publications in AI for HLT at MIT LL.

Since that time, the group has continued to develop new AI approaches, including statistical signal processing for machine translation in the mid-2000s [7, 8] and sparse coding for human language technology in 2018 [9]. The group has more than 30 years of expertise in machine learning and AI for HLT.

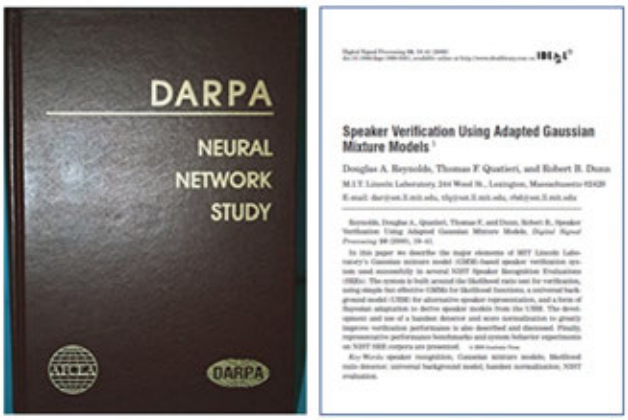


Figure 4.2. Milestone publications in AI for HLT at MIT LL.

4.3 Technology Landscape and Representative Capabilities

AI for HLT rises from a set of technology foundations, viewed through the lens of Mission Perspective, and applied to a set of capabilities and delivered to government sponsors. Figure 4.3 depicts the full landscape of AI for HLT, including technology foundations, mission perspectives, and representative applications.

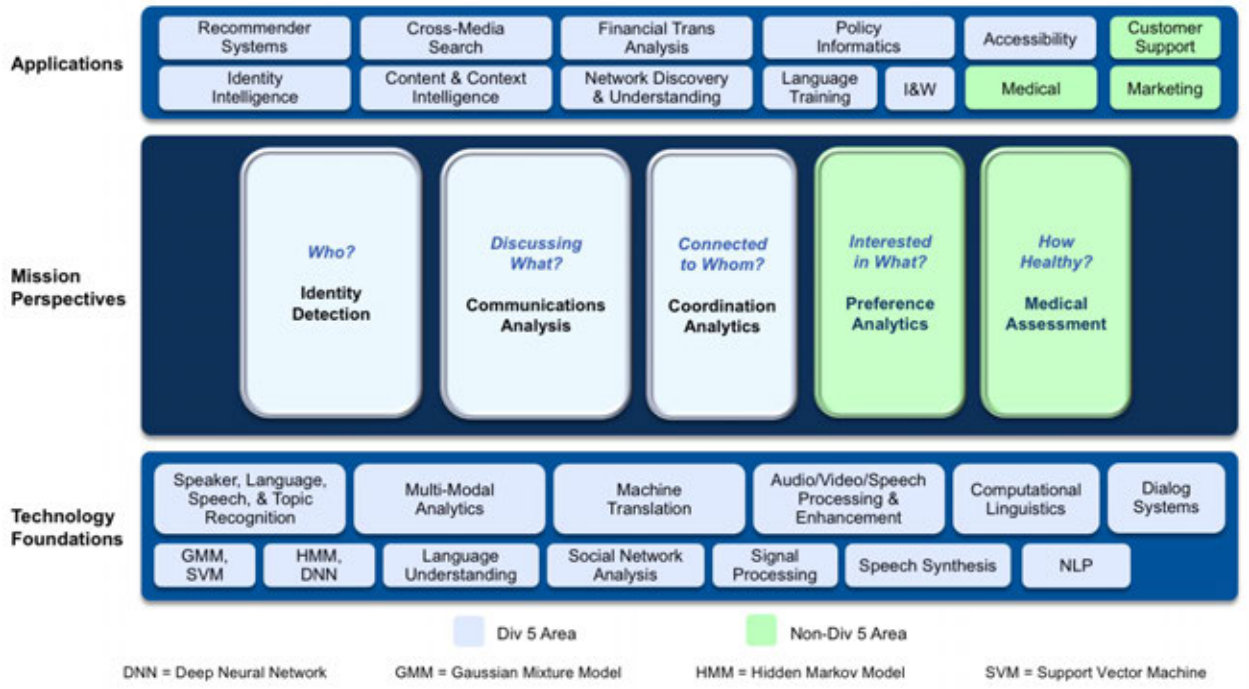


Figure 4.3. The full landscape of HLT across Division 5 and non-Division 5 areas.

Technology foundation areas represent deep technical areas where fundamental research has been conducted. This work is often the result of research in academia or with academic partners, and requires years of investment to reach a mature level. For example, fundamental speaker, language, speech, and topic recognition work has been conducted by MIT LL for more than three decades across academia, industry, and government. The mission perspective is the fundamental need that drives the creation of an HLT system. For example, the need to analyze a flood of communications data in an environment without sufficient analyst resources is a core need for the communications analysis pillar. Finally, specific applications steer how technology foundations are applied to a particular mission. For example, human network discovery and understanding applications address coordination analytics needs and integrate core computational linguistics, natural language processing, image processing, and other technology foundations.

Although there is a broad array of possible missions, at MIT LL, the mission focus is more constrained than the general landscape. As depicted in Figure 4.4, we focus on three different areas: identity detection, communications analysis, and coordination analytics. The first focus, identity detection, involves the determination of a talker's identity and related characteristics like language and gender, from the speech signal using AI. In this area, MIT LL also works to fuse the speech signal with image video modalities to robustly identify individuals from multimedia sources. The second area, communications analysis, focuses on enabling analysts to extract the content of communications with cross-language information retrieval, translation, speech recognition, and signal enhancement. As part of this work, we also build tools to train and augment foreign-language analysts and warfighters in languages needed in government-specific scenarios. The final focus area, coordination analytics, develops AI capabilities to understand networks of bad actors by analyzing the content-in-context of their interactions with others.

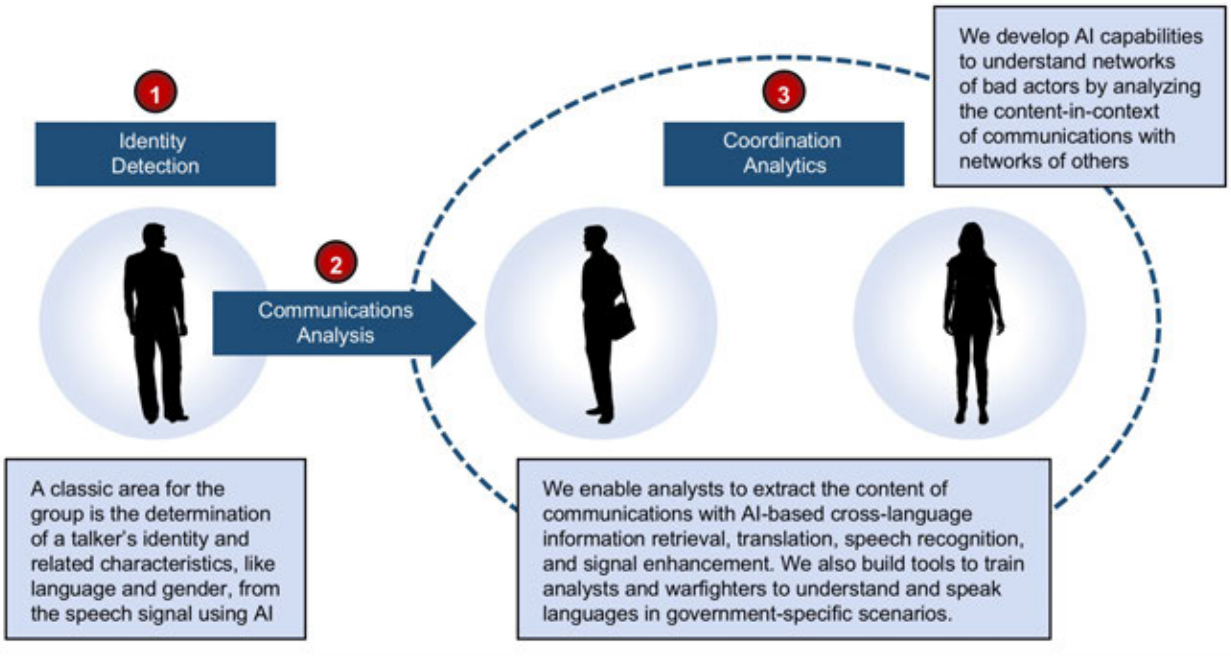


Figure 4.4. MIT LL AI for HLT mission focus.

Two representative capabilities, forensic speaker comparison and cross-language document retrieval demonstrate MIT LL's approach to developing AI capabilities in the area of HLT. In the

first example, forensic speaker comparison [10, 11], shown in Figure 4.5, law enforcement and the IC need to determine whether two speech audio recordings are from the same individual. Manual comparison is error-prone, subjective, inconsistent, and time consuming. To approach this application, data are conditioned by extracting a set of representative features. In addition, file and channel characteristics are detected. Then, machine learning algorithms, based on GMMs and DNNs, are applied to train models on large sets of telephone and microphone speech. The trained models are then applied to compare the two audio files. A human analyst is then presented with a confidence score that helps them interpret how likely the two audio files were created by the same individual.

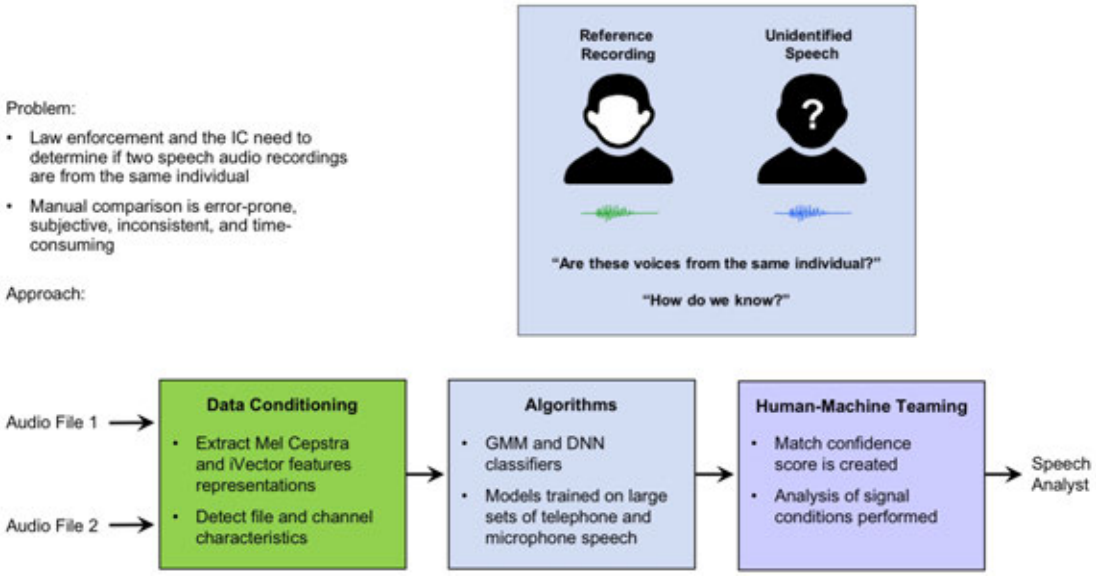


Figure 4.5. Example of an AI pipeline designed for forensic speaker comparison.

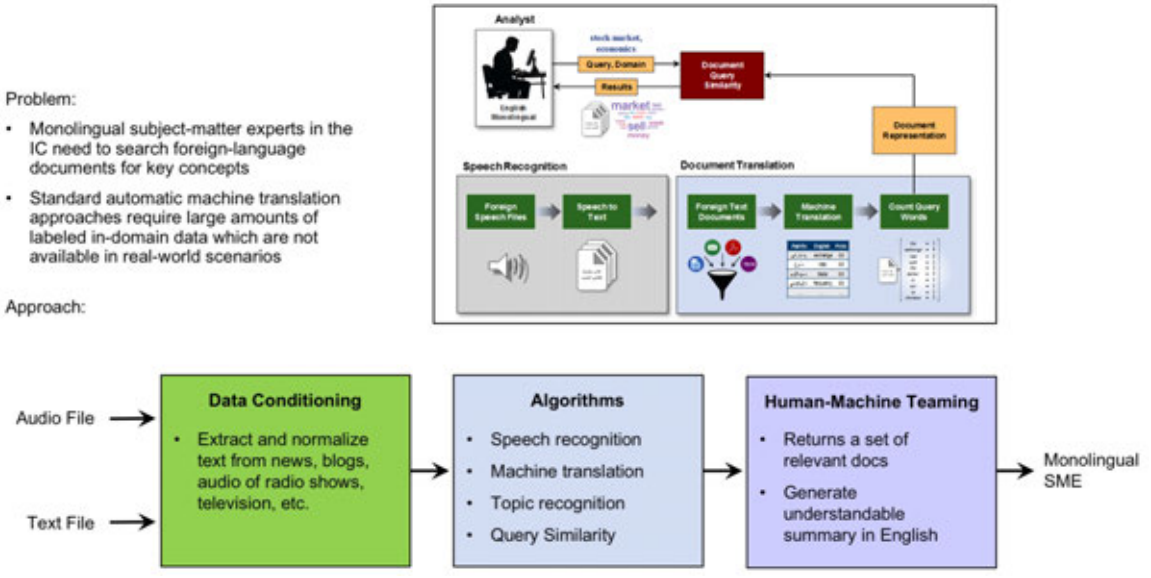


Figure 4.6. Example of an AI pipeline designed for cross-language document retrieval.

In another example—cross-language document retrieval, shown in Figure 4.6—monolingual subject matter experts in the IC need to search foreign-language documents for key concepts. Standard automatic machine translation approaches require large amounts of labeled in-domain data that are not available in real-world scenarios. As with forensic speaker comparison, data conditioning, algorithms, and human-machine teaming aspects are considered.

Our representative examples demonstrate that AI systems are not simply about developing algorithms. The entire chain, from sensor to user mission, built upon a fabric of modern computing and robust processing, forms the complete AI capability. Each step is a technical challenge, as well as an opportunity for innovation, in the process of building an AI system. We can view this chain, a canonical architecture enabling AI for HLT, depicted in Figure 4.7.

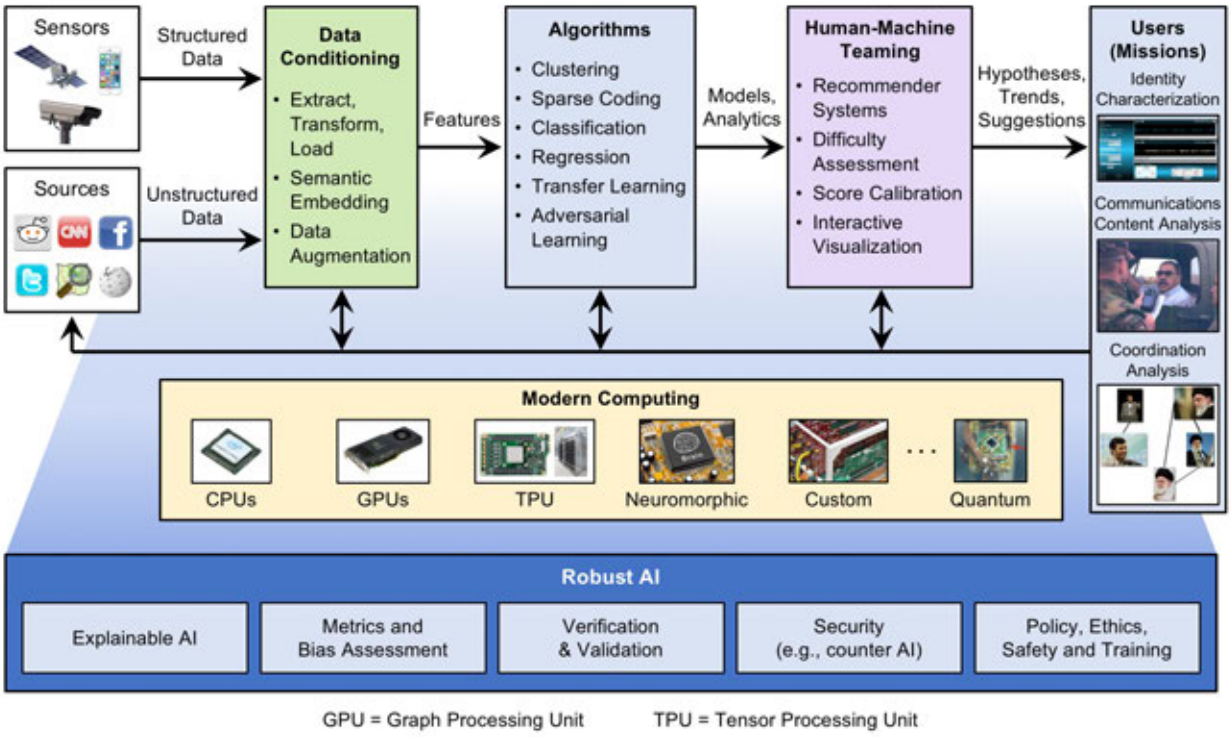


Figure 4.7. A canonical architecture enabling AI for HLT.

Of particular note in the pipeline is the importance of well curated data and early data conditioning steps. Recent discussions of the AI pipeline, including as part of the DARPA D3M program [12] on teaching machines to learn machine learning, focus on automatic data preprocessing as a crucial step [13]. It is often reported that about 80% of data analysis is spent in this critical step [14] and doing so requires highly trained data scientists. A key part of this process, data cleaning [15] has been the topic of recent research focusing on ways that data can be efficiently and automatically conditioned.

4.4 Global Trends Transforming AI for HLT

Several global trends are dominating the transformation of AI for HLT over the 10 years from 2010 to 2020. Across government, industry, and academia, two of the most important trends are the commoditization of a wide range of text and speech analytics and AI for HLT spreading across application domains and being adopted with a wide range of expertise in AI. This situation is a dramatic shift from the earlier environment, where AI solutions, for example

for automatic speech recognition, were customized solutions, often specific to a particular domain and developed and deployed by a specialized team. Now, subject matter experts and software teams with little or no machine-learning background are empowered to try a range of off-the-shelf tools for their problems.

While this trend affects many application areas for AI, it is particularly apparent in face recognition and text, speech, and video processing, where performance levels have enabled available tools to be useful in a wide range of applications. In app development for iPhone and Android, for example, application developers can make use of AI primitives already tuned to those devices [16].

4.5 Academic, Commercial, and DoD/IC/LE Roles in AI Systems

Important to the canonical architecture we have presented is the idea that no one team can create all of the necessary components. As shown in Figure 4.8, academia, commercial industry, and the DoD/IC/LE make up the community of teams creating the future of AI for HLT. We can view the critical contributions of each of these contributors in terms of three critical components for successful government AI systems:

1. Government-specific user perspective
2. Access to mission-specific data
3. Algorithms and system architects

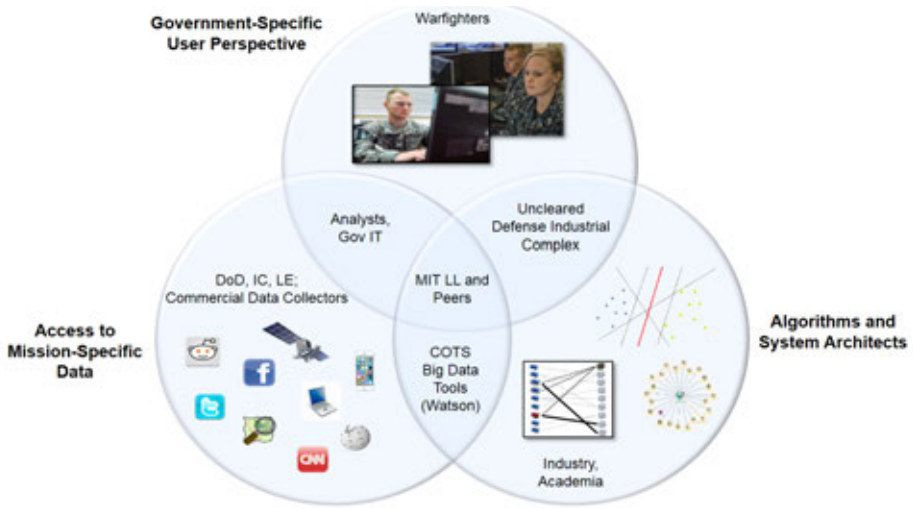


Figure 4.8 The role of academia, commercial industry, and government teams in AI systems.

Creation of emerging AI solutions has a set of associated challenges and opportunities. Different participants in the AI ecosystem have different roles in addressing these challenges. For example, industry is driving advances in processor technology to handle AI system training, and academia and industry are leading the creation of new machine-learning algorithms. Data, especially sensitive operationally relevant data, are the domain of the government IT and analysts. They are often available with only very controlled sharing, even with trusted government partners. Warfighters and analysts have unique perspective on government missions, and they help expose critical technical issues that need solution.

Using the lens of the canonical architecture allows each organization to better understand its role in the AI ecosystem. For example, as an FFRDC, MIT LL has a role as a connector in the AI ecosystem, evaluating early technologies from academia and industry for effectiveness on

government missions and datasets. As required, AI researchers formulate and implement components that are necessary to accomplish the mission. Table 4.5.1 shows an analysis of challenges and opportunities in AI from an MIT LL perspective, highlighting areas to potentially lead, adapt, and leverage.

Table 4.5.1 Challenges and opportunities in AI trends

Challenges		Opportunities
Deluge of many kinds of data, structured and unstructured and often special	Lead	Automated data conditioning
AI training currently takes weeks or months on powerful compute clusters	Leverage	Follow industry advances to create mission capability: GPUs, DNN processors
Integration of machines into human-driven HLT process	Lead	Demonstrate performance of systems in sponsor scenarios; develop proxy problems and data sets; increase interpretability
Persistent performance gap applying academic and industry AI solutions to government	Adapt	Use FFRDC mission understanding to adapt academic and commercial algorithms
Latest open algorithms not available across hardware architectures	Adapt	Adapt algorithms to MIT LL high-performance computing platforms
Peers and adversaries have access to a common set of AI algorithms	Lead	Develop robust HLT AI systems and adversarial models
Narratives in gray-zone warfare are difficult for humans to determine	Adapt	Develop AI to identify signatures, discover narratives, and understand threats

4.6 Key Findings in the Application of AI to HLT

In this section, we discuss observations and trends related to each element of the canonical architecture. These trends will be used to inform our recommendations and way forward.

Data Conditioning: AI in commercial industry is fueled by refined data pipelines, both ahead of an effort to bootstrap new capabilities and as in-the-loop subcontracted annotation to develop a steady stream of labeled data. Apple and Google, for example, both have carefully engineered data acquisition and labeling as part of their AI development process [17-20]. As part of the pipeline, large-scale data procurement and labeling is used to meet the needs of commercial systems as they evolve. Commercial data collection is a large industry, with companies such as Appen providing in-country, in-context data collection. With some exceptions, government does not yet purchase large amounts (millions of dollars) of labeled data to fuel AI programs, with more focus tending to be on algorithms. What data are purchased are often purchased at the beginning of the program, while our study revealed that many companies build continued large procurements into the ongoing lifecycle of a product. They use these later purchases to incrementally improve general performance and also to address special cases, such as automatic recognition of a new phenomenon in social media.

Data used for training and evaluating systems in industry and academia also differ significantly from data as shown in Table 4.6.1. Government data are typically noisier and come from a wider variety of sources than in industry [21]. When operational data are not available, opportunities exist to create proxy data for government problems, such as in the recent IARPA ASpIRE Challenge for speech recognition of noisy and reverberant speech [22], and in the Speakers in the Wild dataset, which has many characteristics unseen in telephony [23].

Table 4.6.1 Factors typically differentiating data in government from data in commercial and academic settings

Factor	Data in Industry and Academia	Data in Government
Sensors	Small set of known sensors	Many possible sensors; often sensor is unknown
Recording Conditions	Controlled conditions, i.e., user speaking directly into telephone	Unconstrained and dynamic conditions
Noise	Low-to-medium	Medium-to-high
Languages	Single, common language	Multiple, low-resource languages
Content Structure	Fixed content structure, i.e., user requesting schedule information	Content structure is dynamic and mixed, ranging from news to conversations
Subject Compliance	Compliant: Subject attempts to produce input that maximizes success of AI	Adversarial: Subject is not aware of AI or produces input to minimize its effectiveness

Algorithms: Open-source toolkits allow users to leverage machine learning with low barrier to entry. Commercial companies are building business on such AI toolkits that can be applied easily. Academics also iteratively develop packages of algorithms to encourage adoption of their research. In this sense, academia and the commercial sector are advancing algorithms and AI capabilities.

In the area of HLT, there are several speech and text tool kits that are widely used. Three examples are:

- **Kaldi:** Kaldi is an open-source toolkit produced by John's Hopkins University to perform automatic speech recognition, speaker recognition, and language recognition [24]. Over the past several years, it has seen widespread adoption, including in government systems. Kaldi developers integrate new algorithms as they are released, in "recipes."
- **Open NMT:** Open NMT is an open-source neural machine translation package involving multiple academic and commercial partners [25]. This toolkit allows rapid development of translation systems, which are updated regularly.
- **Stanford NLP:** Stanford has open sourced a variety of natural language processing (NLP) tools, including a parser, a part-of-speech tagger, and a named-entity recognizer [26]. For English in common domains, these tools are used often and allow rapid prototyping of text processing systems.

As with data, open and commercially available toolkits are designed to operate under different conditions than are often present in government scenarios. Those with access to and an understanding of government challenges can help guide development of off-the-shelf technologies through interactions with the research community. One approach that has seen strong positive results is involvement, including sponsoring, of public challenge problems. These challenge problems, which provide a relevant scenario and curated data, allow HLT AI performance benchmarking. An example of success with this model in government is the IARPA ASpIRE Challenge [22]. Here, procurement of a small operationally relevant dataset has led to a dramatic shift in the state of the art for room audio transcription. Additionally, participation in non-government-affiliated challenge problems and hackathons offer an opportunity to learn from and help shape the state of the art.

Human-Machine Teaming: Voice assistant and recommender systems are driving commercial HLT AI. As with data and algorithms, though, commercial recommender scenarios differ from government-relevant scenarios [27]. A selection of these differences is shown in Table 4.6.2.

A major difference between industry and government recommender systems is that government teams often need to distribute work. In other words, the same content in many cases should not be provided to multiple team members, as this would lead to redundant work. In addition, domains for recommender use by government are high-consequence—suggesting intelligence for review rather than suggesting movies or purchases.

Table 4.6.2 Differences between typical commercial and government recommender-system scenarios

Factor	Industry	Government
User Independence	Each user receives independent recommendations	Users share workload—recommendations must be in context of whole team
Cost of Error	Low	High
AI Explainability	Not required	Required for action in government
Languages	User language	Multilingual
Content Structure	Highly structured (i.e., item specifications)	Less structure (multimedia documents)

Modern Computing: In modern computing, a significant focus is swinging back from cloud processing to edge computing [28, 29]. For example, Apple is promoting on-device machine-learning implementations, supported by accelerated hardware on mobile devices. Recently, strategies for combining edge with DNNs [30], especially distributed DNNs [31], have been pursued.

The key feature of edge processing is robustness to limited or non-existent communications back to centralized resources. In commercial industry, these issues can cause delayed responsiveness to customers [28]. In government applications, lack of edge capability increases the time from field-forward data ingest to insights on those data.

An example demonstrating the importance of edge processing to the government is document and media exploitation (DOMEX) [32]. Here, large numbers of hard drive and media images are ingested in forward-operating locations. Information on these drives is of great importance and may be of high intelligence value, but it currently may take weeks or months for the full drive images to be physically shipped back to the enterprise, due to the volume of data being transferred. In the meantime, analysts at headquarters have a very limited ability to summarize the content of a drive to determine if it is of interest or to transfer large multimedia files for inspection. Figure 4.9 depicts several potential strategies for media transmission. Smart compression of the data, using select AI on the tactical edge can potentially reduce the amount of data that needs to be sent.

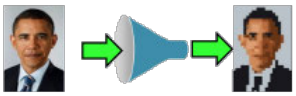
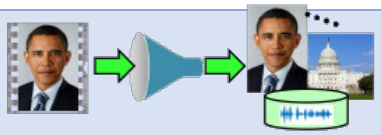
Approach	Description	System Impact
Physical Shipment 	Physically ship media	• **High latency** • **High bandwidth**
File Compression 	Transmit entropy-based coding of multimedia files	• **High latency** • **Low bandwidth**
Smart Representation 	Transmit content-based coding of multimedia files	• **Low latency** • **Increased effective bandwidth**

Figure 4.9. Comparison of potential strategies for media transmission.

Robust AI: A key component of usable AI systems is providing confidence that a system will work in a predictable way in practice. Robustness is a critical component of this confidence. As depicted in Figure 4.10, in cases where confidence in AI systems is high and consequences are low, problems are best matched to machines. In cases where confidence is low and consequences are high, problems are best matched to humans.

We have found that there are cases where a problem that initially appears best matched to humans can be decomposed into portions that are matched well to people and portions that are handled by machines. One example is in AI-augmented language learning. Here, one particular element of test creation is the creation of a large body of foreign language materials, labeled by difficulty level. Although overall test creation is arguably too critical of a process to replace with a machine, automatic leveling quickly creates sets of documents that can be used by experts to build a test.

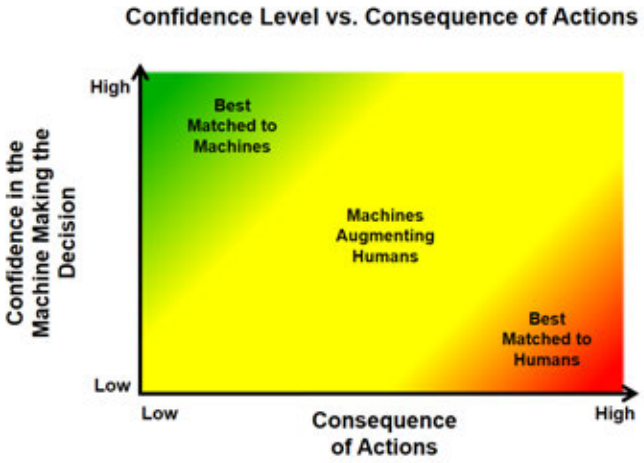


Figure 4.10. Effect of confidence level and consequence of actions on the role of AI in a process.

With worldwide access to data, algorithms, models, and computing for HLT AI, the information battlespace is changing. Here, counter AI is critical as adversaries gain high-performance AI capabilities. AI for HLT must be robust to be effective. By gaining access to an AI system, an adversary can potentially learn and then introduce imperceptible perturbations to inputs that render the system unusable. Adversarial attacks can limit the effectiveness of AI solutions, leading to incorrect behavior [33].

Real-world HLT AI systems are currently deployed worldwide, supporting the needs of different actors. These include surveillance, mission planning, biometric security, and forensic analysis. In the area of surveillance, nation states are reportedly employing AI to monitor their own population using, for example, Megvii Face++ [34].

Currently, best practices for robust systems focus on measuring reduced performance as systems move from prototype data to data encountered in operational settings. Measuring this potential real-world performance gap requires access by testers to operational data, from which an evaluation set is constructed, as illustrated in Figure 4.11.

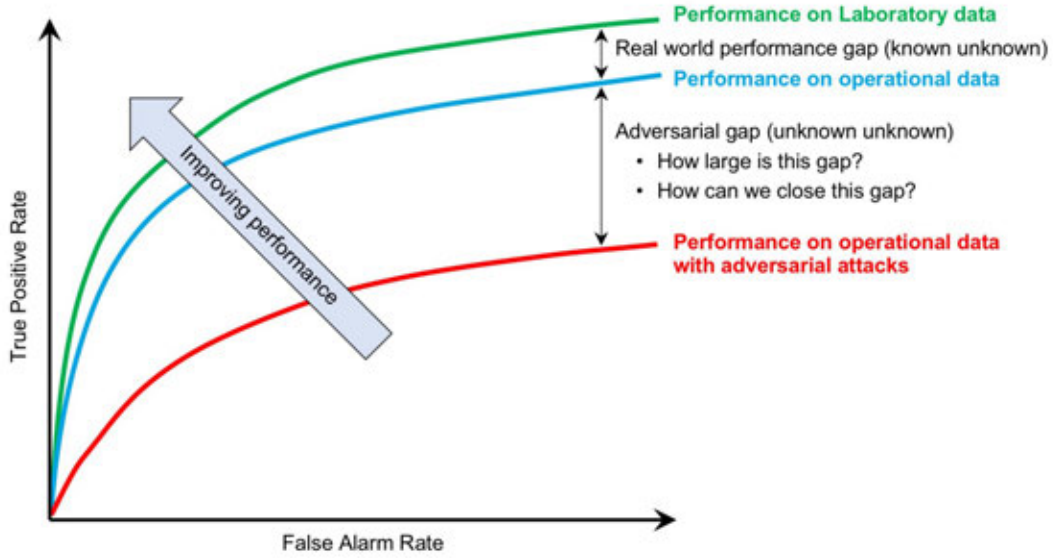


Figure 4.11. Notional illustration of potential performance gaps due to operational data and in the presence of adversarial attacks.

Despite increasing reliance by U.S. Government, however, little known capability exists to secure HLT AI systems to adversarial attack. In the face of this threat, called the adversarial gap here, system performance can potentially be degraded in a way that is not well understood and for which rigorous tests may not currently exist, even with access by T&E teams to operational systems and data. Figure 4.12 illustrates this concept. In particular, it is critical to understand mechanisms by which adversaries can affect performance, how this interference can be detected, and if/how the gap can be closed by mitigating vulnerabilities.

Challenges include: understanding the attack surface (adversarial access to machine-learning system data, models, users), quantifying the “adversarial gap” for systems under attack, developing and assessing defensive capabilities, and in-situ system evaluation and best practices (i.e., machine-learning resilience testbed). Figure 4.12 illustrates some of the complexity of the potential space of attacks on AI systems, which includes aspects of access and which stage of the machine-learning pipeline is targeted.

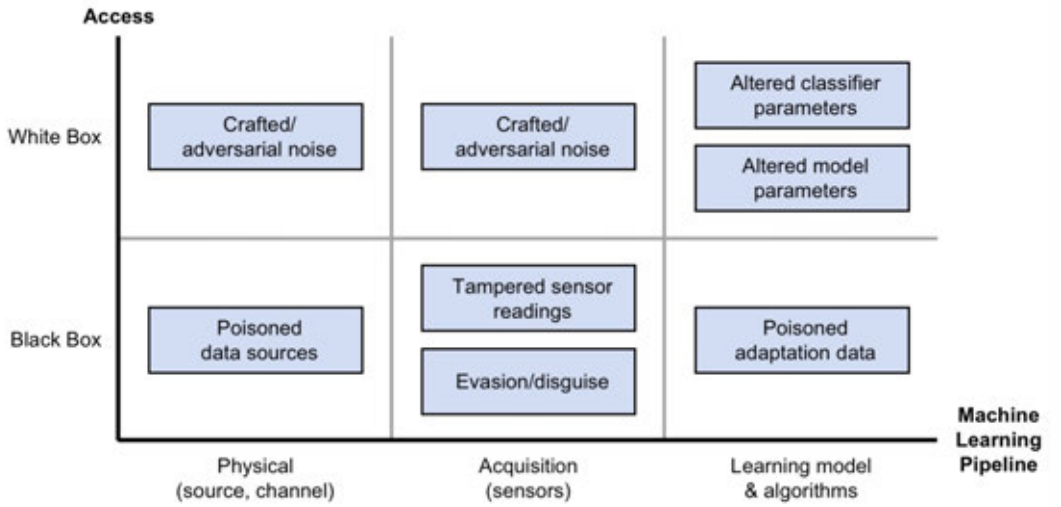


Figure 4.12. Potential space of attacks on AI systems, broken down by stage of the machine-learning pipeline and by the type of access required.

Adversarial effects on AI systems have the potential to interfere with a wide variety of functions. For surveillance tools, for example, attacks could prevent detection of an individual or connection with activities. Forensic analysis tools could potentially be manipulated to hide relevant evidence from investigation. In information retrieval, attacks could potentially make systems fail to return items relevant to query.

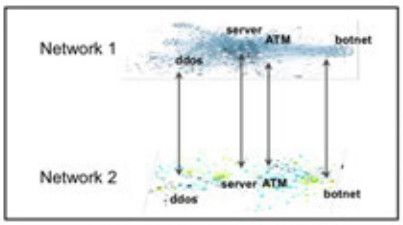
To achieve a specific goal or to reveal a potential attack, a specific system must be broken down into its component parts. Specific successful attacks may occur at points that are both 1) accessible using either white box or black box approaches and 2) important to the outcome of the AI system in terms of the specific adversary goal.

User/Mission: A key part of the AI chain is understanding emerging needs for user missions. As an example, recent DARPA hackathons, illustrated in Figure 4.13, demonstrate that there is a growing need to analyze coordination across multimodal sociocultural networks.

- Financial Crime
- Social Media Analysis
- GISR Sensor Processing
- Traceroute Analysis
- MoveINT Processing

- Multimedia Indexing/Search
- Multi-Platform Persona Linking (Drugs, Guns, Hacking)
- Syrian Refugee Screening
- Evidence-Driven Policy Making

- Patent Litigation Detection
- Yemeni Ceasefire Detection
- Underage Sex Ad Detection
- Indicators of Human Trafficking
- Chat Forum Understanding



Persona Linking across Dark Web Forums

Counter Human Trafficking

Refugee Screening

Figure 4.13. Recent DARPA hackathon topics.

There is also a trend toward complex narratives arising in gray-zone warfare. Gray-zone warfare is complex because it involves multiple simultaneous fronts, different scopes, and is adversarial in nature.

4.7 Recommendations and Way Forward

A summary of study findings is shown in Figure 4.14.

Summary of Findings: AI Applied to Human Language Technology

Data	1. Industry purchasing and collecting huge stores of data to support their AI development 2. Government data is noisier and comes from a wider variety of sources than in industry
Algorithms	3. Many available commercial and open HLT AI systems and components 4. Public challenge problems, with curated data, allow HLT AI performance benchmarking
Teaming	5. Voice-assistant and recommender systems driving commercial HLT AI success 6. Commercial recommender scenarios differ from government-relevant scenarios
Computing	7. Significant focus is swinging back from cloud processing to edge computing
Robust AI	8. Worldwide access to data, algorithms, models, and computing for HLT AI 9. Adversarial attacks can limit effectiveness of AI solutions, leading to incorrect behavior
User and Mission	10. Growing need to analyze coordination across multimodal sociocultural networks 11. Trend towards complex narratives arising in Gray Zone Warfare

Figure 4.14. Summary of study findings.

These lead to the set of five recommendations in Figure 4.15.

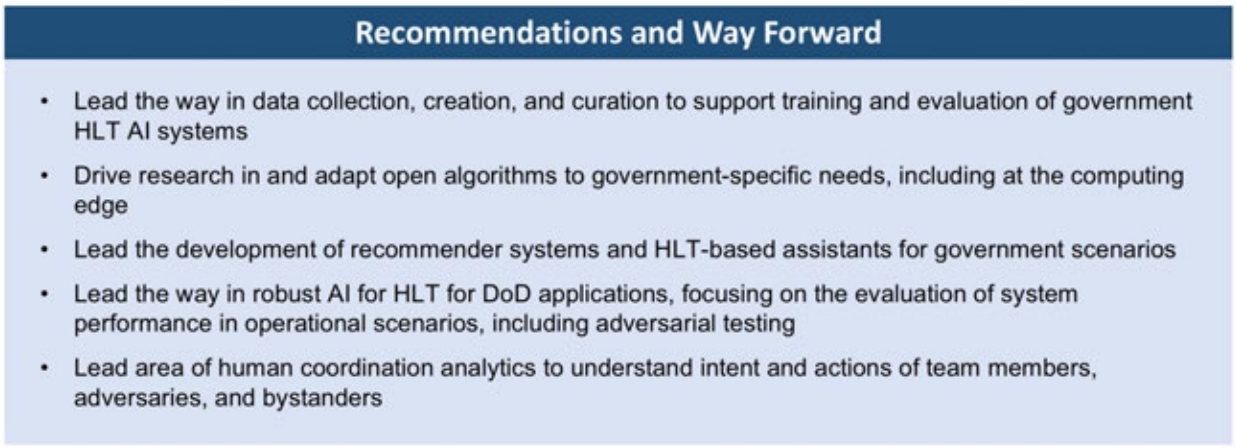


Figure 4.15. Study recommendations and way forward.

The potential benefits of leading the way in data collection, creation, and curation are to become a government touchpoint for HLT data collection and curation, to bring best data practices from industry to government, and to build trust in the ability of AI systems to work in operational settings. To accomplish this, the way forward is to design and support sponsor-funded data collections for AI training and evaluation, create and advocate for strong government evaluation pipelines using operational data, follow commercial data processes closely, and develop and adapt synthetic data augmentation.

In adapting open algorithms to government-specific needs, the benefits are to leverage best-of-breed for DoD applications, to lead in AI HLT areas specific to DoD, e.g., specific needs at the tactical edge, and to influence academia and industry toward DoD problems. To accomplish these benefits, the way forward is to carry a message of the state of the art to sponsors, transform available capabilities for DoD applications through adaptation and re-engineering, and to define proxy problems, baselines, and metrics for open evaluations to drive research to areas of sponsor needs.

When we lead the development of recommender systems, we allow the government to maximize use of available analyst time, subject matter expertise, and language skills to reduce the time from analyst knowledge to insight and to enable warfighters to access and use timely mission-relevant multimodal information. The way forward in this area is to follow commercial HLT developments closely; work with analysts to understand where AI can help, develop training and human augmentation tools to maximize warfighter effectiveness; and develop and adapt recommender systems specific to government scenarios.

For robust AI, the benefits are to assure mission success in the face of advanced adversary AI capabilities, to enable quantification of AI vulnerability, and to move the community forward in leveraging AI in support of DoD missions. The way forward here is to follow academic research closely, prototype robust AI capabilities within DoD programs, develop programs in counter-AI for HLT applications, and build a new test range for AI/information battlespace using cross-division resources.

Finally, in the area of human coordination analytics, the benefits are to leverage decades of world-class machine-learning research and development in the HLT Group for new focus areas, to lead applications of AI to DoD and IC problems, to defend against rapidly advancing threats, and to effectively address gray-zone information warfare. We will achieve these benefits by broadening sensor modalities to include social-cultural networks, continuing to develop a

portfolio of demonstrations in the area of human coordination analytics, and formulating a "moonshot" prototype for gray-zone information warfare. Figure 4.16 shows the initial concept for a "moonshot" prototype for gray-zone information warfare.

In summary, the AI landscape is changing for HLT. There are more off-the-shelf capabilities, as well as adversaries and peers who can use them. There are widely available machine-learning toolkits. In addition, government adoption of these capabilities requires support to adapt to mission scenarios and operational data relevant to important missions. AI developers should design large sponsored data collections, especially in emerging multimodal and multichannel areas. We should also adapt latest algorithms to mission needs. We must also develop robust AI solutions, especially in the face of adversarial threats.

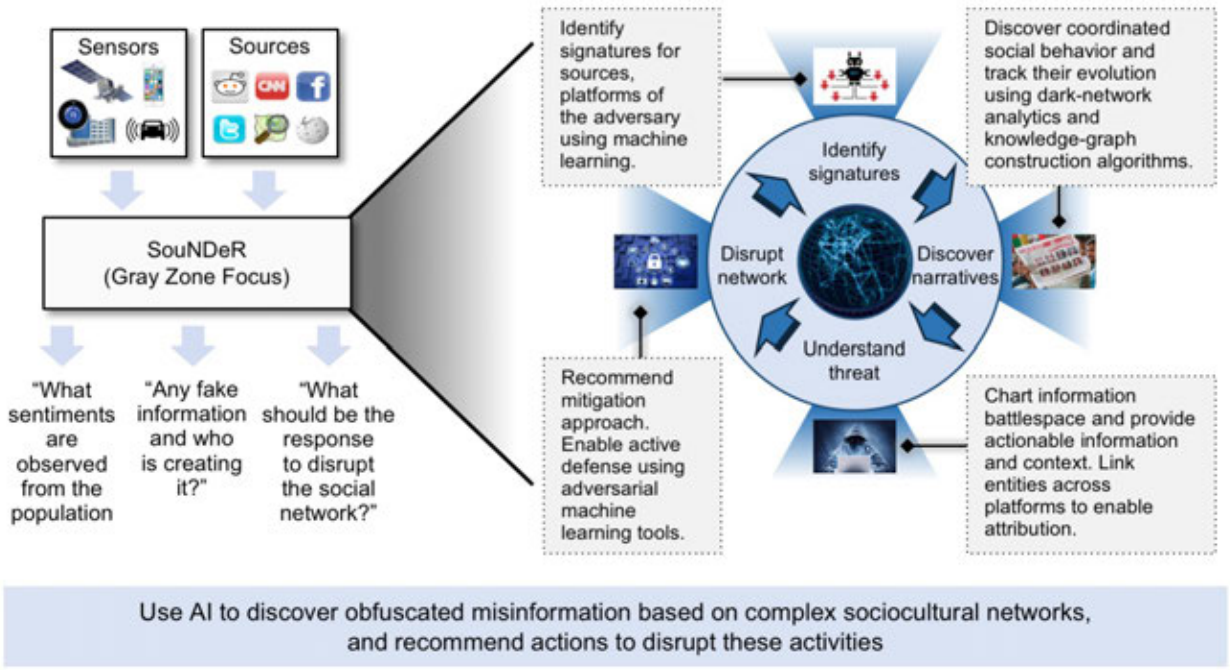


Figure 4.16. Sociocultural Network Attack Discovery and Response (SouNDeR).

References

1. Goldstein, A. *Ben Gold, an oral history*. 1997; https://ethw.org/Oral-History:Ben_Gold.
2. *The DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus: NIST Speech Disc CD1-1.1*. [CD-ROM] 1990;
3. Study, D.N.N., L. Laboratory, and U.S.A.F.S. Command, *Defense Advanced Research Projects Agency neural network study final report*. 1989: The Laboratory. https://books.google.com/books?id=QeeHuAAACAAJ
4. Lippmann, R.P., L. Kukolich, and E. Singer, *LNKnet: neural network, machine-learning, and statistical software for pattern classification*. 1993, MIT Lincoln Laboratory.
5. Reynolds, D.A., T.F. Quatieri, and R.B. Dunn, *Speaker Verification Using Adapted Gaussian Mixture Models.* Digital Signal Processing, 2000. **10**(1-3): p. 19-41.
6. Campbell, W.M., D.E. Sturim, and D.A. Reynolds, *Support vector machines using GMM supervectors for speaker verification.* IEEE Signal Processing Letters, 2006. **13**: p. 308-311.
7. Shen, W., B. Delaney, and T. Anderson. *The MIT-LL/AFRL MT System*. in *International Workshop on Spoken Language Translation (IWSLT) 2005*. 2005.
8. Shen, W., et al. *The JHU workshop 2006 IWSLT system*. in *International Workshop on Spoken Language Translation (IWSLT) 2006*. 2006.
9. Gwon, Y., et al. *Sparse-coded net model and applications*. in *2016 IEEE 26th International Workshop on Machine Learning for Signal Processing (MLSP)*. 2016.
10. Campbell, J.P., et al., *Forensic speaker recognition.* IEEE Signal Processing Magazine, 2009. **26**(2).
11. Schwartz, R., et al. *USSS-MITLL 2010 human assisted speaker recognition*. in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*. 2011. IEEE.
12. Lippmann, R., W. Campbell, and J. Campbell, *An Overview of the DARPA Data Driven Discovery of Models (D3M) Program*, in *NIPS '16*. 2016: Barcelona, Spain.
13. Samulowitz, H. *Automated Feature Engineering for Predictive Modeling* in *ICDM17-D3M*. 2017. New Orleans, LA.
14. Dasu, T. and T. Johnson, *Exploratory data mining and data cleaning*. Wiley series in probability and statistics. 2003, New York: Wiley-Interscience. xii, 203 p. http://www.loc.gov/catdir/toc/wiley031/2002191085.html
15. Wickham, H., *Tidy Data.* 2014, 2014. **59**(10): p. 23. https://www.jstatsoft.org/v059/i10
16. Wiggers, K. *Apple's Core ML 2 vs. Google's ML Kit: What's the difference?* 2018.
17. Brown, T.B. and C. Olsson, *Introducing the Unrestricted Adversarial Examples Challenge* in *Google AI Blog*. 2018.
18. Hughes, T., et al. *Building transcribed speech corpora quickly and cheaply for many languages*. in *INTERSPEECH*. 2010.
19. Nellis, S. *Apple's Siri learns Shanghainese as voice assistants race to cover languages*. 2017 March 9, 2017; https://www.reuters.com/article/us-apple-siri-idUSKBN16G0H3?feedType=RSS&feedName=technologyNews&utm_source=Twitter&utm_medium=Social&utm_campaign=Feed%3A+reuters%2FtechnologyNews+%28Reuters+Technology+News%29.
20. Polyzotis, A., et al., *Data Management Challenges in Production Machine Learning*, in *Proceedings of the 2017 ACM International Conference on Management of Data*. 2017: New York, NY. p. 1723-1726.
21. Sturim, D., P. A. Torres-Carrasquillo, and J. P. Campbell, *Corpora for the Evaluation of Robust Speaker Recognition Systems*. 2016. 2776-2780.
22. Harper, M. *The Automatic Speech recogition In Reverberant Environments (ASpIRE) challenge*. in *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*. 2015.
23. McLaren, M., et al. *The Speakers in the Wild Speaker Recognition Challenge Plan*. 2016.
24. Povey, D., et al. *The Kaldi speech recognition toolkit*. in *IEEE 2011 workshop on automatic speech recognition and understanding*. 2011. IEEE Signal Processing Society.

25. Klein, G., et al., *OpenNMT: Open-Source Toolkit for Neural Machine Translation.* ArXiv e-prints, 2017. https://arxiv.org/abs/1701.02810
26. Manning, C., et al. *The Stanford CoreNLP natural language processing toolkit*. in *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*. 2014.
27. Gadepally, V.N., et al., *Recommender systems for the department of defense and intelligence community.* Lincoln Laboratory Journal, 2016. **22**(1).
28. Satyanarayanan, M., *The Emergence of Edge Computing.* Computer, 2017. **50**(1): p. 30-39.
29. Satyanarayanan, M. *Edge Computing: Vision and Challenges*. in *HotCloud '17*. 2017. Santa Clara, CA: USENIX Association.
30. Huang, Y., et al. *When deep learning meets edge computing*. in *2017 IEEE 25th International Conference on Network Protocols (ICNP)*. 2017.
31. Teerapittayanon, S., B. McDanel, and H.T. Kung. *Distributed Deep Neural Networks Over the Cloud, the Edge and End Devices*. in *2017 IEEE 37th International Conference on Distributed Computing Systems (ICDCS)*. 2017.
32. Garfinkel, S.L., *Document & media exploitation.* Queue, 2007. **5**(7): p. 22-30.
33. Laskov, P. and R. Lippmann, *Machine learning in adversarial environments*. 2010, Springer.
34. Jacobs, H. and P. Ralph *Inside the creepy and impressive startup funded by the Chinese government that is developing AI that can recognize anyone, anywhere*. Business Insider, 2018.

5 AI Applied to Cyber Security (B. Streilein)

Recent improvements in AI have resulted in advances in many technological and scientific fields including medicine, transportation, communication, and data analysis. In these cases, AI techniques have assisted humans in several ways, including dealing with large amounts of information (i.e., big data), recognizing anomalous behaviors or trends, and making complex decisions. As will be shown in the sections below, cyber security faces many of the same challenges as these other areas, and thus, AI has the potential to have similar impact if applied in appropriate ways.

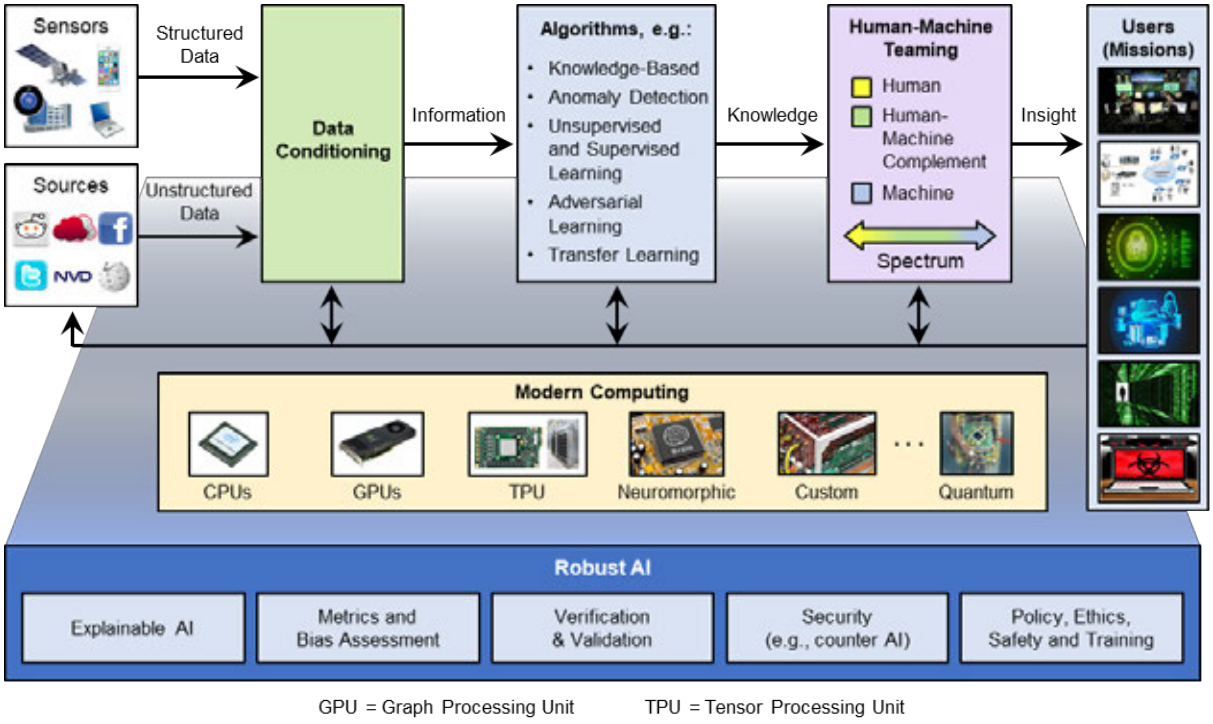


Figure 5.1. Architecture for application of AI to cyber security.

Figure 5.1 presents an architecture for applying AI to cyber security. The first step in the architecture involves conditioning (cleaning, normalizing, etc.) of both structured and unstructured data in order to prepare them for use by AI algorithms. AI algorithms, including both unsupervised and supervised learning algorithms, consume conditioned data in the next block to develop knowledge about the data. Output from trained AI algorithms can be leveraged by human-machine teams to develop insight that is relevant to particular cyber missions, such as vulnerability discovery, intrusion detection, and others, shown to the right of Figure 5.1. To effectively meet cyber domain timescales, developed capabilities will rely upon modern computing technologies, such as GPUs and TPUs, and others. Finally, the architecture highlights the need for robust AI solutions that can be verified, validated, measured, are robust to attacks (e.g., adversarial learning), and are understandable or explainable. Section 5.3 will refer to the AI architecture in Figure 5.1 to present key findings and recommendations from the study of the application of AI to cyber security.

5.1 Cyber Background

At its most basic level, a cyber system comprises a user transacting data with a computer system, as depicted in Figure 5.2. Given this simple representation, the critical threat surfaces of such a system are threefold: the compromised user, the compromised input, and the compromised system components. These surfaces represent the main entry points for an attacker and capture at a high level all the ways that a cyber system can be compromised. That is, all known attacks at some level can be binned into these categories of attacks.

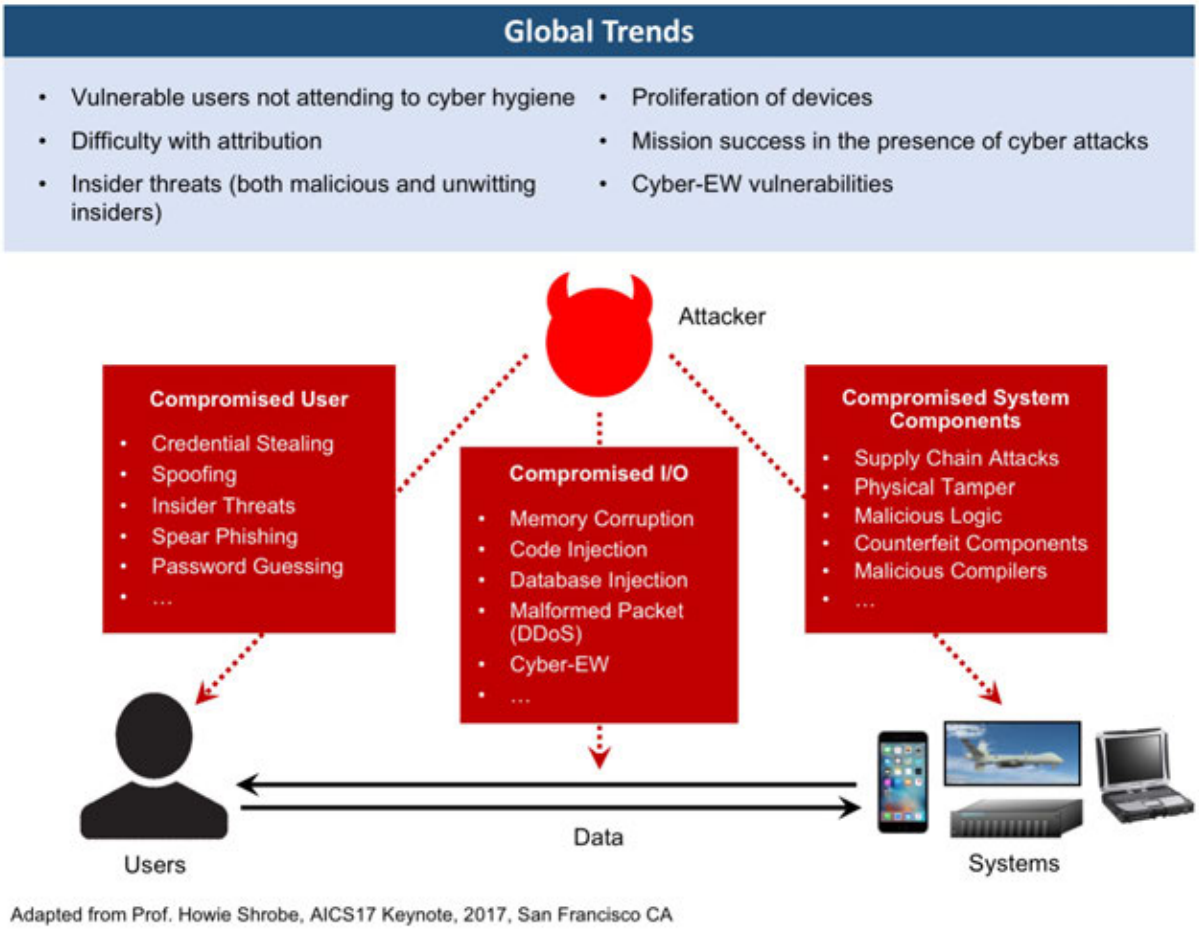


Figure 5.2. Critical threat surfaces for a cyber system as well as notable trends.

As depicted in Figure 5.2, the user is one of the main ways a cyber system can be compromised; in fact, the user continues to be one of the weakest points in any such system. Ways in which an attacker can compromise a user include credential theft through keylogging capabilities, physical theft, or brute force password theft [1, 2]. Through well-crafted spear-phishing attacks, the attacker can trick an unsuspecting user into providing credentials or banking information in order to gain access to a protected system [3]. Spear-phishing continues to be a pervasive concern for security professionals and users alike.

Compromising the input of a cyber system can take place in many ways, including corrupting internal memory [4], in which attackers alter important storage locations in running programs to achieve their effects, such as leaking data or changing program behavior. Code

injection into the control path [5] is another way to compromise the input of a cyber system. Through this method, direct control of the system is taken by the attacker. Finally, the sending of malformed messages and control packets can cause a system to behave counter to original design specifications in order to support attacker goals [6].

The final category of attacks against a cyber system involves compromise of the system components. In this class of attack, the attacker is able to insert malicious or counterfeit components into the development or operational chain of a cyber system. In [7], an example is presented of a supply chain attack against an industrial control system (e.g., SCADA) with the purpose of compromising its security. Of the three types of attacks, component compromise takes more planning on the part of the adversary to accomplish, but can lead to more widespread impact [8].

A number of trends have been noted when discussing the cyber domain, including the continuing vulnerability of users, the potential for insider threats, and the proliferation of connected devices, such as IoT devices. However, one of the major trends in cyber security is the use of automation by an attacker in carrying out their actions; this trend is behind the incredible rise in the number of unique attacks that are seen and is overwhelming defenders. Moreover, as depicted in Figure 5.3, although the sophistication of attacks continues to increase, automation of these capabilities enables less sophisticated hackers to carry out the attacks. Metasploit is an example of one such tool that is freely downloadable online and readily provides attack capabilities to users [9], regardless of their sophistication.

Noteworthy Facts

- 250K new malware programs are registered each day
- There were 357M new email malware variants in 2016—36% more new variants than in 2014
- There were 463M new variants of ransomware in 2016—36% more new variants than in 2015
- Attackers have become more aggressive and sophisticated
- 99 days to detect compromise—adversary gains access in three
- Internet of things and cloud are hot targets (e.g., Mirai botnet)—two minutes to compromise
- Projected cyber attack costs in 2019: $2.1T

Sources:
https://www.symantec.com/security-center/threat-report http://expandedramblings.com/index.php/cybersecurity-statistics/
http://www.nato.int/docu/Review/2016/Also-in-2016/cyber-defense-nato-security-role/EN/index.htm
https://www.fireeye.com/blog/threat-research/2017/03/m-trends-2017.htm https://www.ag-test.org/en/statistics/malware

Figure 5.3. Global trend: sophisticated attacks more easily accomplished with automation.

Recasting the cyber system representation in a slightly different manner reveals the tension between the attacker and defender within the cyber battleground. As depicted in Figure 5.4, we see an enhanced version of the cyber system, showing an expanded view of potential cyber systems, including processing, storage, and communication components, all of which can be attacked. While the system is being leveraged to carry out a mission, the attacker is executing his or her "kill chain", indicated by red blocks at the top of the diagram. The steps in the kill chain enable the attacker to know the target (prepare), launch an appropriate attack (engage), establish persistence, and finally, achieve and assess his or her effect. Figure 5.4 presents a simple representation of the Lockheed Martin Cyber Kill Chain [10]. While this is happening, but not at the same time or in synchrony, the defender is also executing stages in a multi-step defensive process, as indicated by the blue blocks at the bottom of the diagram. The defender process starts

by identifying critical components of the cyber system and installing protections for them. From this point, the defender must continue to be vigilant by monitoring and detecting new attacks that require responses or online defensive actions. Finally the defender must support the mission recovery so that critical functionality can be restored. The defender steps are laid out in the NIST Cybersecurity Framework [11].

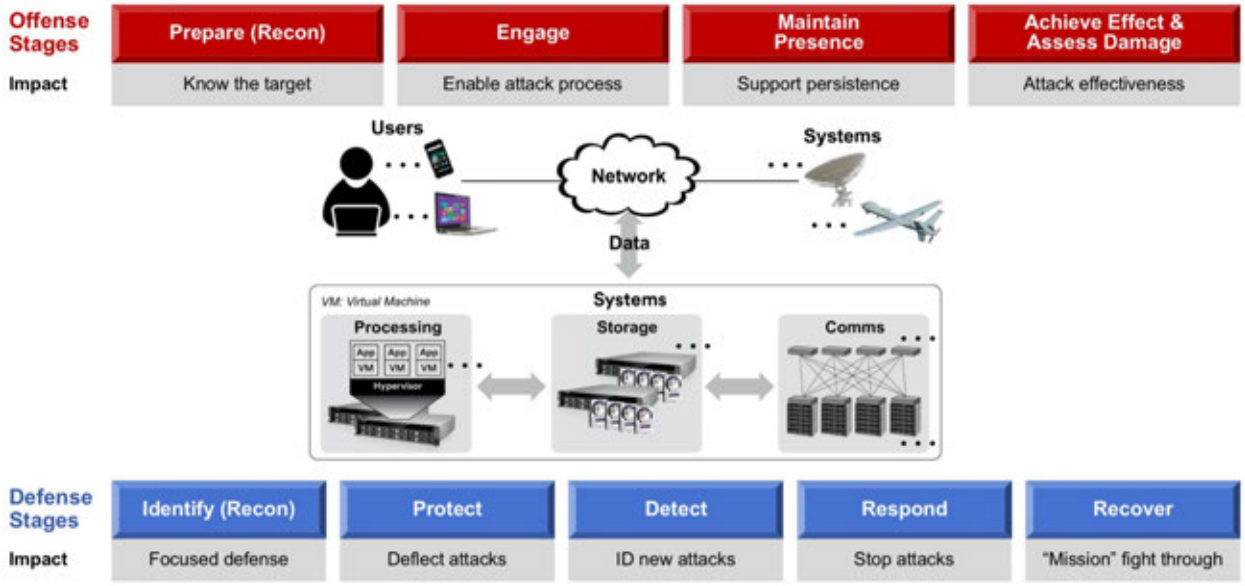


Figure 5.4. The cyber battleground showing attacker kill-chain steps (red blocks) and defender defense steps (blue blocks).

Having outlined the cyber battleground, it is important to enumerate the major challenges facing the defender while trying to stop the attacker and protect the cyber system. AI capabilities that can help in dealing with these challenges will be discussed in the following section.

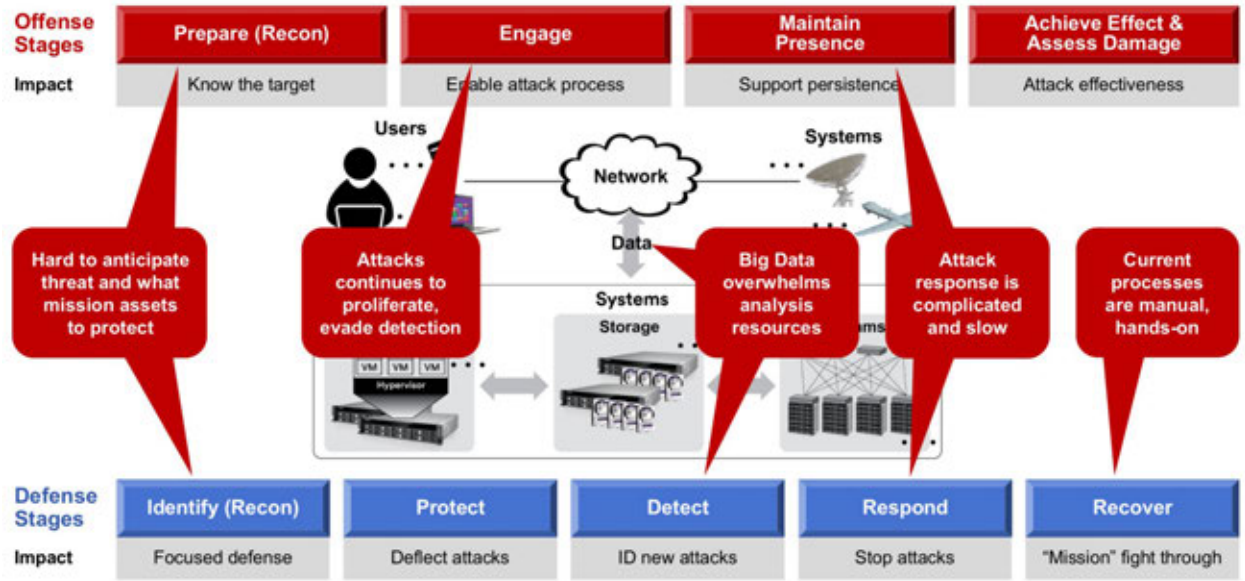


Figure 5.5. Major challenges to cyber security including anticipating the threat and identifying assets, as well as dealing with big data and responding to detected attacks.

First among the challenges faced by the defender is the need to identify the critical assets that need protection. Through the analysis of network and system data as well as business need procurement information, systems that are important for mission and user function are identified. The Camus system is an example of an automated capability that performs identification of critical assets and how they support the mission, so-called "mission mapping [12]."

Once identified, these critical systems must be protected through the use of cyber security best practices and secure technologies, such as data encryption and cyber moving-target capabilities [13]. This step continues to be a challenge as the attacks continue to evolve and proliferate, evading protection mechanisms [14].

Another significant challenge for the cyber defender is dealing with the overwhelming amount of relevant data in order to develop situational awareness. While some of these data, specifically are from network protocols, structured, much of them are not, which adds to the difficulty of understanding them. Situational awareness, which involves the analysis of cyber big data, enables a real-time understanding of the state of both red and blue activities in the context of mission goals. This understanding supports the detection of attacks that have made it past network and host protections that are in place. Although it has been investigated for many years, the development of cyber situational awareness capabilities continues to be an important research area [15].

Once an attack has been detected, it is necessary to respond by stopping the attack and removing its effects, such as re-imaging systems that are affected. This step remains a significant challenge for cyber defenders as it typically involves human intervention to decide amongst many potential responses, thus slowing the response and potentially overwhelming the human. In some cases, it may be necessary to determine who caused the attack through attribution determination [16]; however, attribution remains a significant challenge for cyber responders as the nature of the Internet supports anonymity.

After the attack is neutralized, it is necessary to restore mission functionality. This step is difficult and time-consuming as the systems that are supporting the mission are identified and restored. The processes for mission restoration for determining mission impact are typically manual and slow, though work is being done in this space [17, 18] to improve speed and efficacy.

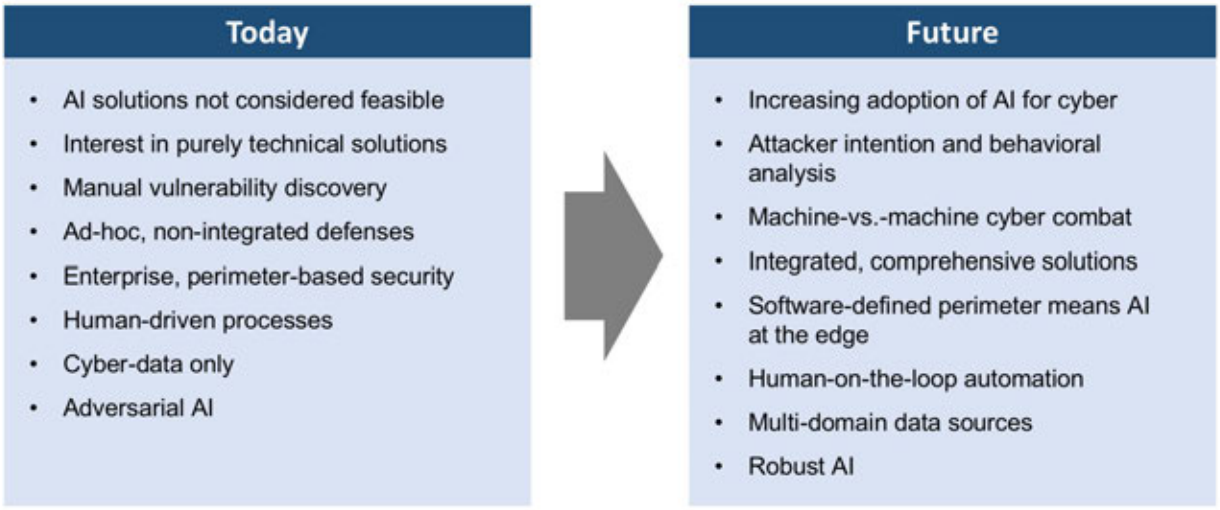


Figure 5.6. Notable trends relating to the use of AI for cyber security include increased adoption of techniques, machine-to-machine combat and robust AI solutions.

As we look to the future, several overarching trends are worth noting that can lead to an AI-enabled cyber security future (see Figure 5.6). Despite the challenges with the fast-evolving cyber domain, we expect to see increased adoption of AI capabilities for cyber. The commercial market has exploded with a large number of start-ups hoping to capitalize on applying AI to cyber security problems. Many of these solutions leverage big data architectures to triage large amounts of data, allowing the human security analyst to focus on more relevant and potentially more threatening data.

As both the defender and attacker increasingly use automation and AI, we expect to see more machine-to-machine interaction. This reflects the fact that the cyber domain requires machine-speed response, something beyond human capabilities. The Cyber Grand Challenge is an excellent example of this capability being developed with support from the DoD research community [19, 20].

Another trend of note is the move from perimeter-based security to one that is integrated with the edge of the network and is defined by software rather than hardware components such as routers and proxies. Finally, we see a future for the application of AI to cyber in which machine-learning-based solutions are robust to adversarial AI, in which an attacker subverts developed capabilities through data poisoning or model inversion.

Challenges		Opportunities
Deluge of cyber data, structured and unstructured and often special	Lead	Automated data conditioning of cyber mission data
Ineffective cyber detection algorithms, high false alarm rate	Adapt	Adapt academic advances in algorithm improvement
Insufficient computing resources to support cyber need	Leverage	Follow industry advances to create mission capability: GPUs, DNN processors
Integration of machines in human-driven cyber processes due to lack of explainability	Adapt	Augment human capabilities with automated processes; engender trust in AI solutions
Applying non-mission solutions to mission problems	Lead	Leverage FFRDC mission knowledge and access to adapt, including offense
Cyber detection, characterization algorithms vulnerable to attacks	Lead	Develop robust AI techniques based on adversarial learning

Figure 5.7. Challenges and opportunities for MIT LL in adopting AI for government mission needs.

Challenges and opportunities related to the application of AI to cyber are numerous (see Figure 5.7) and one needs to choose its approach carefully in order to achieve the maximum impact possible to meet government mission needs. In some cases, MIT LL should lead the way, such as in developing automated conditioning capabilities, leveraging our deep mission familiarity and knowledge to adapt capabilities, and finally, to help the government maintain resilience of capabilities in the presence of adversarial learning attacks. In other cases, MIT LL should adapt and leverage what exists elsewhere, such as advanced algorithms being developed in academic and commercial communities, automation technologies that augment human capabilities and engender trust in AI, and industry advances in special-purpose hardware that can be used for AI.

5.2 Representative Capabilities and Technologies

Division 5 possesses a number of ongoing cyber efforts that leverage AI (e.g., machine learning) capabilities in order to achieve impactful results for government mission needs. As depicted in Figure 5.8, example programs include those for detecting online cyber discussions of relevance, threat forecasting and risk-based decision assistance, as well as others that detect counterfeit components and support embedded device red-teaming. In this section, we briefly review a few of particular relevance and interest in order to demonstrate the current state of Division 5 use of AI for cyber.

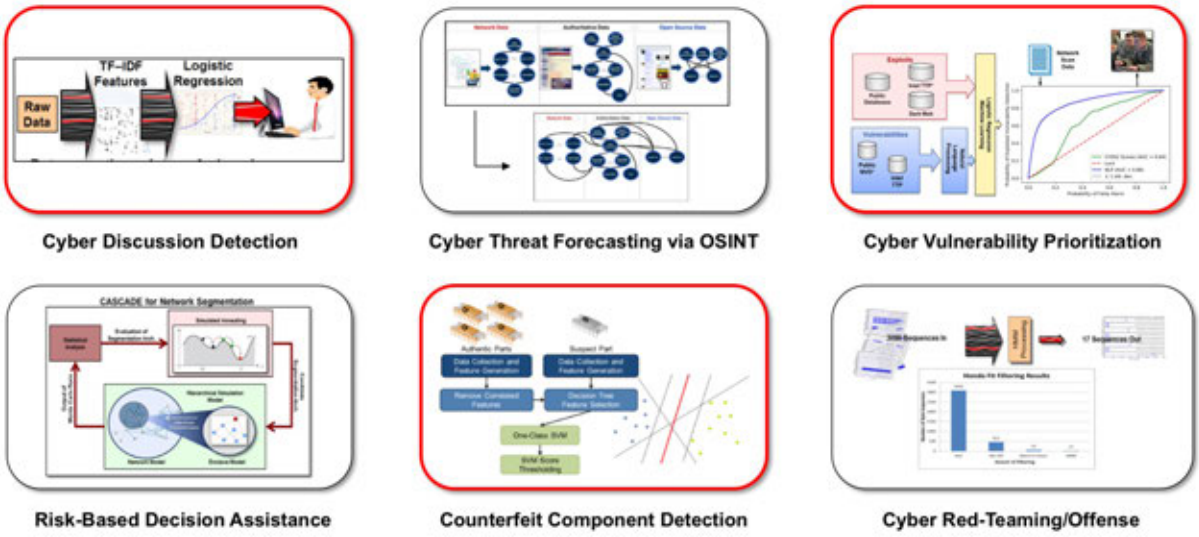


Figure 5.8. Example cyber AI programs in Division 5 include detection of online cyber discussions, risk-based cyber decision assistance, and counterfeit component detection.

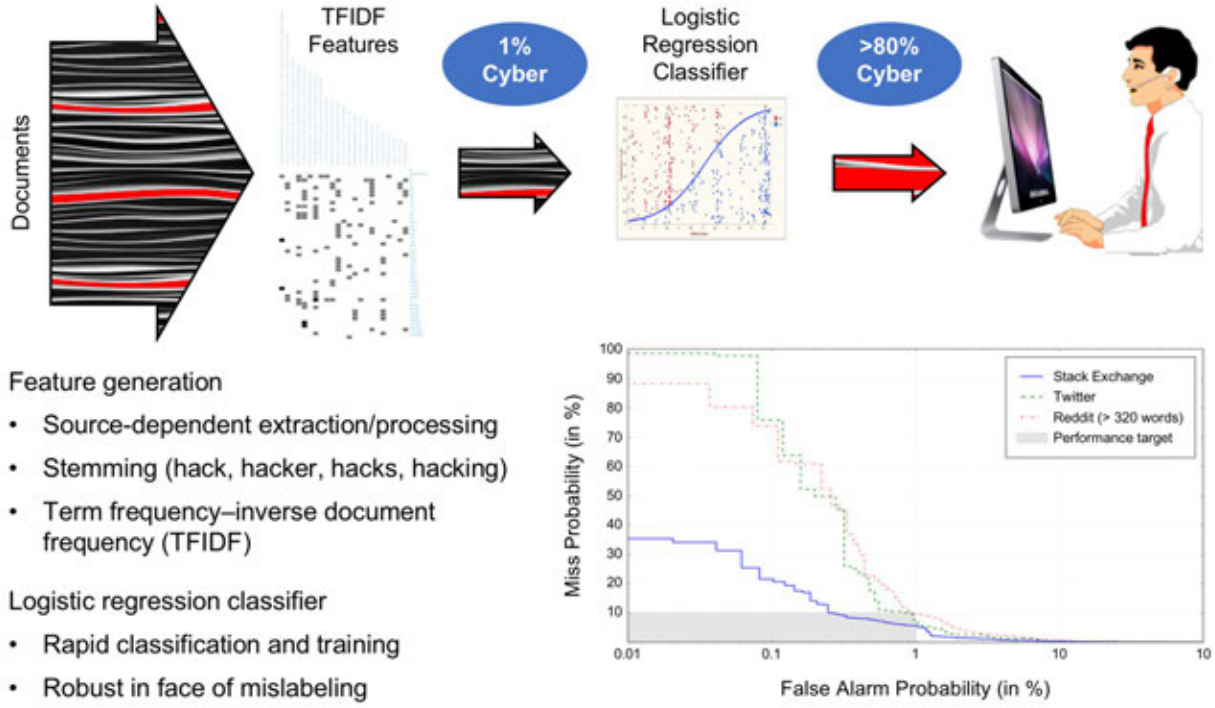


Figure 5.9. The CHARIOT capability automates the detection of online cyber discussions for the cyber analyst.

An important capability for cyber analysts within the U.S. government is the ability to stay ahead of the attacker in order to anticipate their intended targets or method of attacking. Attackers often coordinate their activities online—especially in the case of large-scale attacks,

such as denial-of-service (DoS) attacks—making use of public discussion forums and more recently, social media capabilities to incite and recruit others and to coordinate their attacks [21].

Among the challenges faced by cyber analysts is the lack of clear detection signals that enable ready detection of the relevant discussions. Unlike examples of malicious code or executables, where specific sequences of bytes, also known as "signatures," can be used to identify directly the presence of the offending attacker [22], detection of relevant information from forums relies upon inferred information gleaned from unstructured discussion text. Another challenge relates to the sheer volume of information and data that must be looked through in order to find the discussions of relevance. Current analysts' processes leverage manual capabilities and human intervention limiting their ability to cover a large amount of data, thereby leading to missing of important discussions. However, given the increasing number of attacks, it is important to be able to process online data to anticipate attacker actions and put in place relevant defenses.

To address these and other challenges of cyber analysts, the Cyber Analytics and Decision Systems Group in Division 5 has developed a capability known as CHARIOT to discover cyber-related discussions in large amounts of online media [23]. The system, depicted in Figure 5.9, leverages HLT to triage and featurize data and a logistic regression classifier to detect discussions of interest for analysts dealing with large amounts of unstructured data.

As depicted in the figure, the CHARIOT system does well to detect discussions of interest in three representative online data sources: Reddit, Stack Exchange, and Twitter. In each case, the system is able to achieve a high rate of detection while at the same time maintaining a low false alarm rate. The performance of the system meets the strict performance parameters of the analyst, indicated by the light gray rectangle in the performance curve graph. The CHARIOT system has been transitioned to the sponsor environment and is undergoing testing for operational deployment.

Another example of ongoing work in Division 5 that leverages AI capabilities in support of cyber security mission needs is the SIde Channel Authenticity Discriminant Analysis (SICADA) system. The SICADA system leverages machine learning capabilities to detect counterfeit parts that may be created by adversaries attempting to introduce malicious behavior into working U.S. government systems. As depicted in Figure 5.10, SICADA relies upon a SVM to discriminate between authentic and counterfeit components. The SVM is trained on electronic signals, such as power and voltage traces, gathered during normal operation of known good parts. Similar signals collected from suspected components are presented to a trained SVM, which in turn uses the signals to classify the parts as authentic or counterfeit.

The system achieves a high level of accuracy and is able to detect suspected counterfeit parts which can then be further investigated to determine their status.

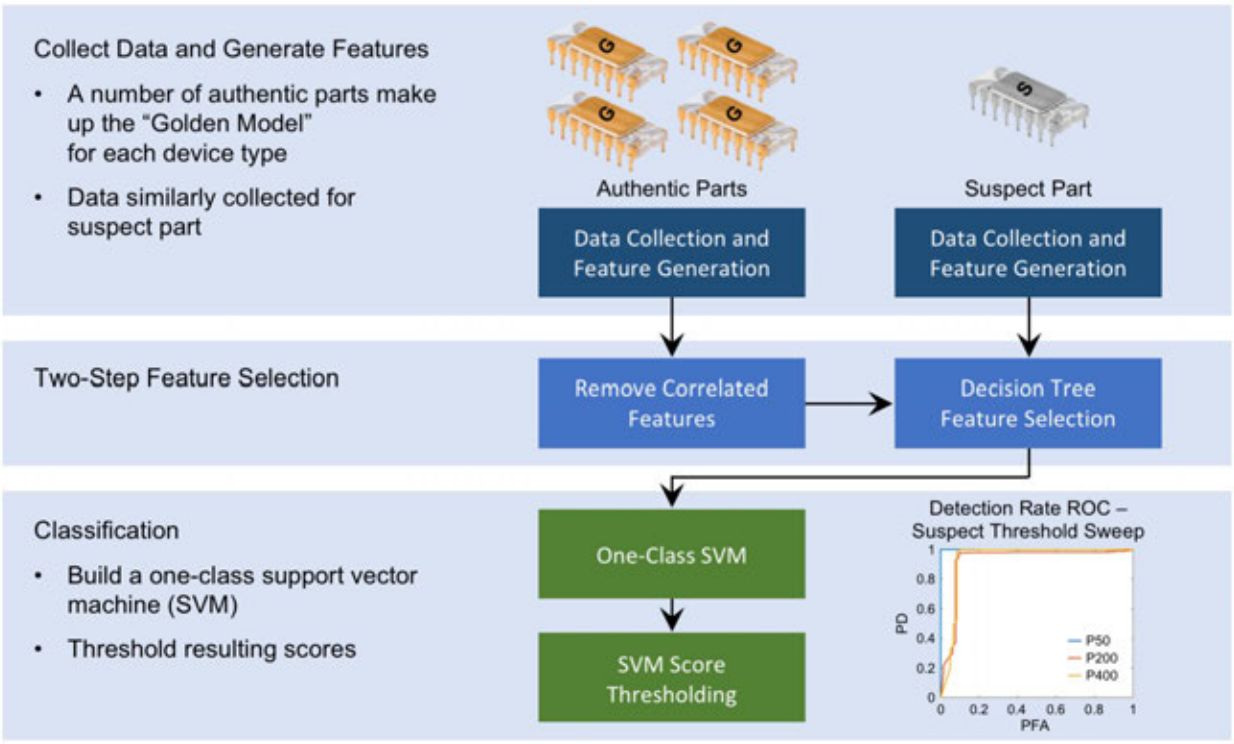


Figure 5.10. The SICADA system leverages machine learning (SVM) to support the detection of counterfeit cyber system components.

A final example of relevant work in Division 5 that leverages AI capabilities in support of cyber security is depicted in Figure 5.11. The capability leverages natural language processing to associate unstructured descriptions of vulnerabilities with critical exploits witnessed in practice to predict likely vulnerabilities that will be exploited by hackers. The system achieves an improved level of accuracy over raw CVSS (Common Vulnerability Scoring System) scores. The system output could enable a cyber analyst to prioritize which vulnerabilities should be patched first.

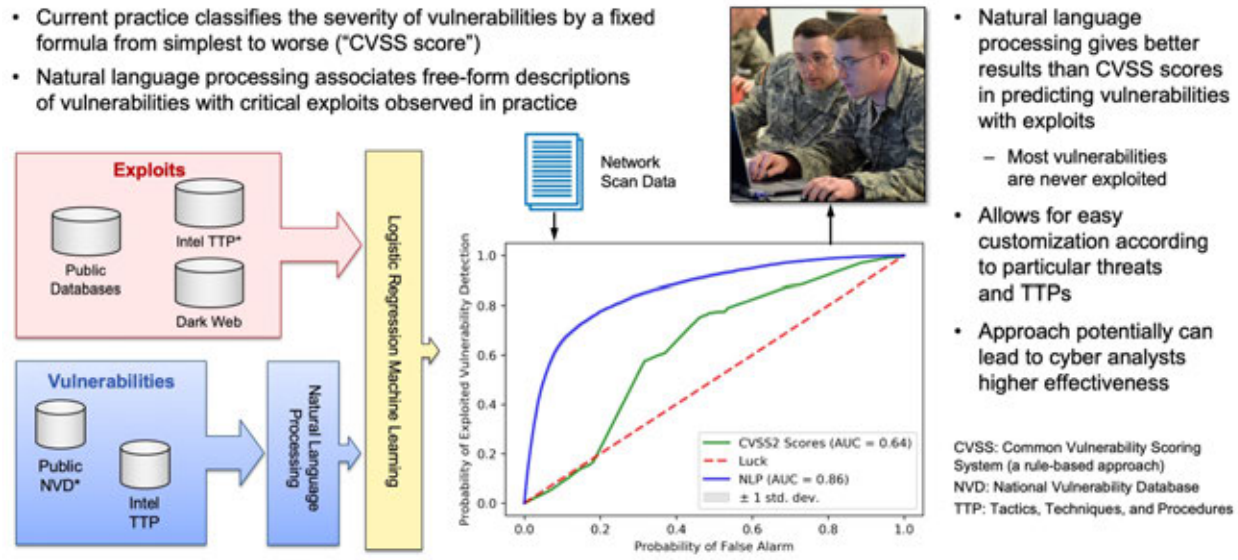


Figure 5.11. Machine-learning capabilities such as natural language processing and logistic regression can be applied to the problem of cyber vulnerability prioritization, yielding improved performance.

5.3 Key Findings in the Application of AI to Cyber Security

The successful application of AI to any domain requires acknowledging and addressing that domain's challenges; cyber security is no exception [24]. Our study has identified several key findings that have particularly important bearing on the state of cyber security and the future of research, development, and operations. For the discussion below, we have organized our findings along the lines of the AI architecture presented in Figure 5.1: data conditioning, algorithms, human-machine teaming, computation, and robust AI. For each component of the canonical architecture, we have culled the key takeaway of concern.

First and foremost is the finding that one of the major challenges facing the application of AI to cyber security relates to the huge amount of data that must be leveraged to make progress. As reported in a recent report by Internetworking giant CISCO, network-related data are getting created at a rate that continues to increase so much that by 2021, the Internet will produce 3.3 zettabytes of data. [25]. As depicted in Figure 5.12, this huge amount of open-source data corresponds to traffic between computers and users, but also the large number of smart devices, IoT devices, and cell phones, and represents huge opportunities for analysis and intelligence production.

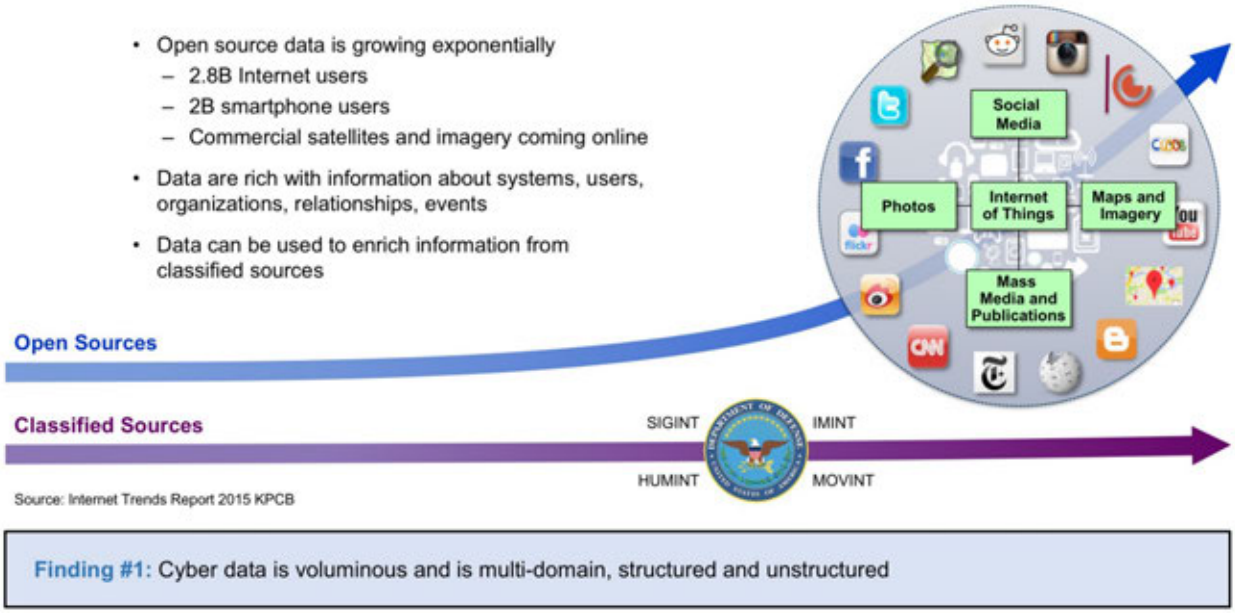
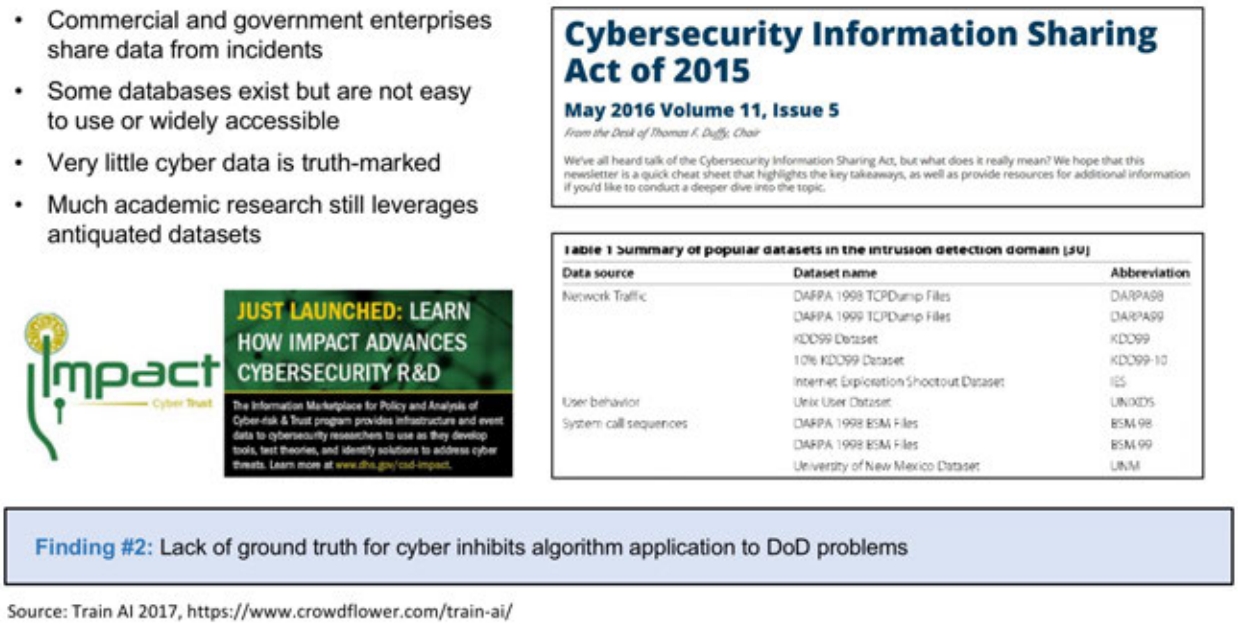


Figure 5.12. Data conditioning: the open-source intelligence big data opportunity.



Source: Train AI 2017, https://www.crowdflower.com/train-ai/

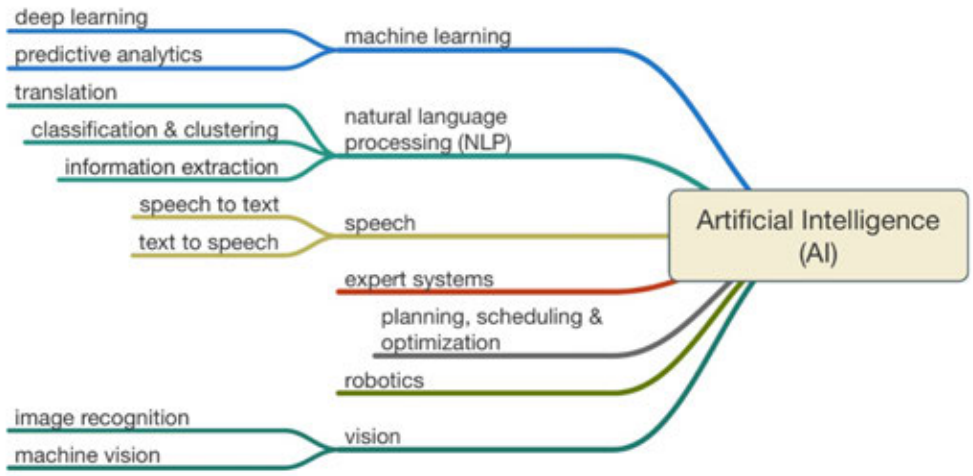
Figure 5.13. Large truth-marked cyber datasets are difficult to find.

In other application areas of AI, such as image recognition and health care (e.g., medicine) large truth-marked datasets have enabled major advances in capability. These large datasets support training of the algorithms over a wide variety of domain conditions that imbue the models with the ability to generalize to unseen cases. In addition, these common well-known datasets enable sharing of techniques and results across the research communities, which leads to more rapid evolution of capability as a whole.

However, as outlined in Figure 5.13, the cyber domain continues to lack agreed-upon truth-marked datasets that can support research and capability advances. Despite the enormous amount

of data being generated on a daily basis [25], these data are largely not truth-marked and thus, not suitable for training today's AI algorithms; the data's designation as malicious or non-malicious is not known. In response to this, research in cyber security has had to rely upon older datasets, such as the DARPA Intrusion Detection Evaluations from 1999 [26], which are not representative of today's network environments.

There is promise in recent efforts to create common repositories of data for use in cyber security research. These repositories, such as the IMPACT database (formerly known as PREDICT [27]), contain large collections of network and host traffic, some of which contain interactions between attackers and defenders, such as during government exercises and capture the flag events. However, because of the sensitivity of the data (it may contain proprietary or personal information), access is restricted to those who can justify a strong research need and who promise to delete the data once the research is complete. This tight control of the community data contrasts with the open datasets in other communities, such as MNIST [28] for handwritten character recognition and ImageNet [29] for image classification, where any researcher can download and work with the data.



Finding #3: Many algorithms exist which can be applied to cyber security

Figure 5.14. A variety of AI algorithms exist, many of which can be applied to cyber security.

As we consider the algorithms themselves (see Figure 5.14), there are several intelligent capabilities that hold promise for improving current cyber security processes. As depicted in Figure 5.15, the defender stages (in blue) can all benefit from the application of AI. Predictive analytics can support the defenders' task of identifying network components that need protection. Among these types of capabilities are those that support mission mapping, which refers to the capability to identify how computer components and systems support mission capability and thus need particular protection. Algorithms that leverage open and special data sources to monitor and anticipate the threat can help to point out weaknesses that are at particular risk for exploitation. In addition, clustering and classification algorithms can be leveraged to infer and prioritize vulnerabilities that must be addressed on mission assets [30, 31].

After protecting those devices and systems that have been identified as vulnerable, the defender must still work to detect attacks as they arrive in order to repel and respond, and finally restore mission functionality. This stage will require triage of large amounts of real-time system

data in order to detect attacks that succeed in making it past protection measures. AI capabilities that recognize anomalous traffic patterns by comparing with baseline activity can help here, although false alarms can plague this stage [32].

The decision about how to respond to detected cyber attacks must take into consideration many variables, including effectiveness at thwarting the activity, impact on the network itself and its users, and the role of the mission being undertaken. Planning and optimization algorithms can help significantly at this stage by considering a large of number of potential scenarios and outcomes through simulation until an optimal solution is discovered [33] .

Finally, the defender must work to restore mission functionality. Recent advances in recommender systems and human-computer interface technology based on HLT can help the human enact solutions that work best [34].

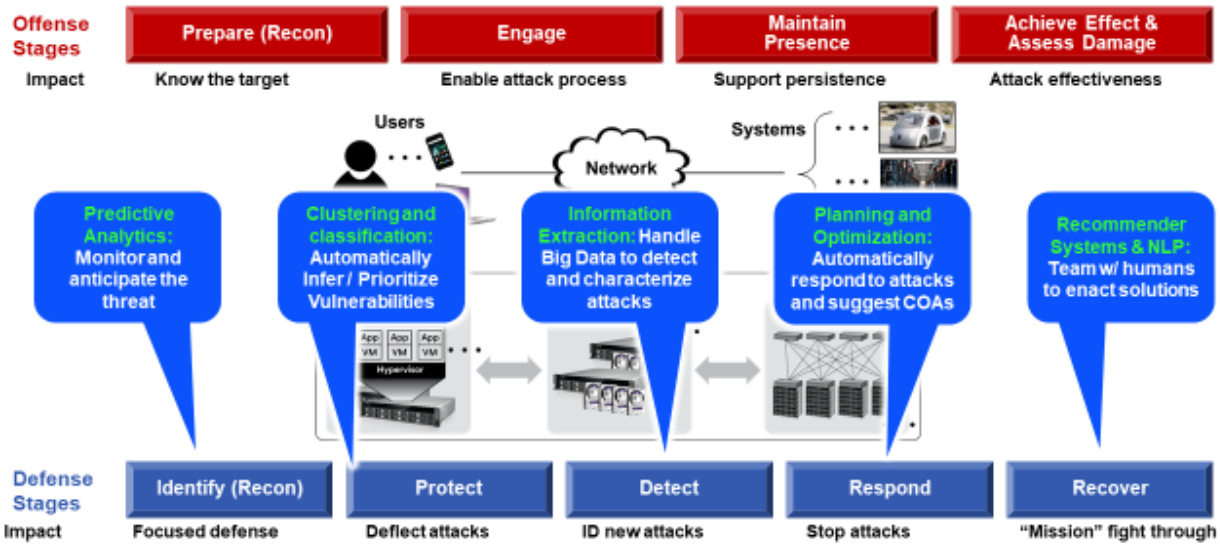


Figure 5.15. AI can help defender processes in a number of ways, including predictive analytics and planning and optimization.

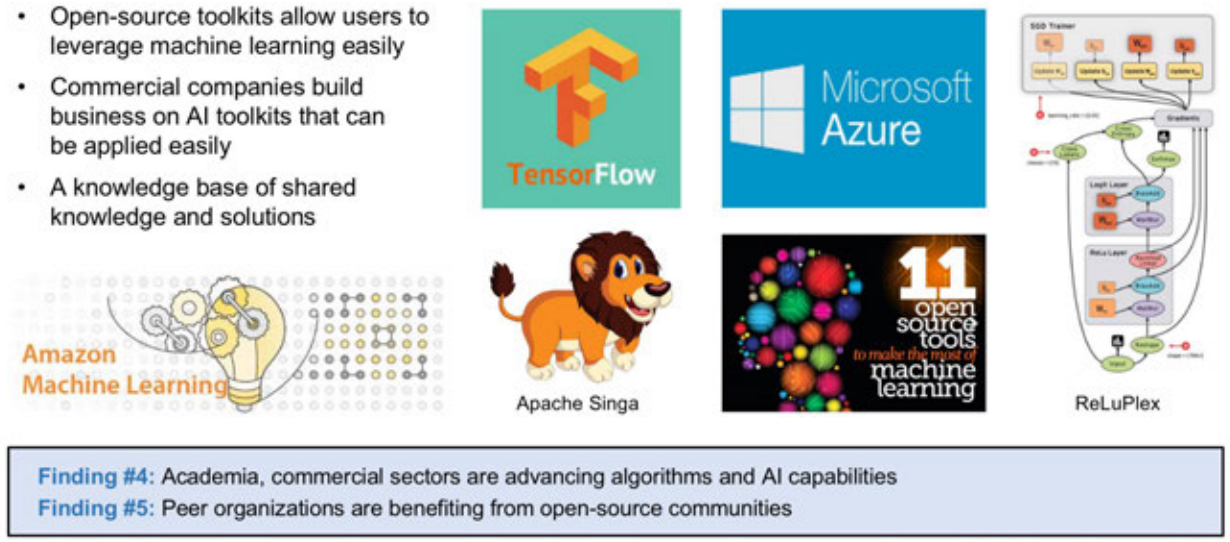


Figure 5.16. Many AI resources are available online that should be leveraged by MIT LL and the government to bring to bear impactful mission-relevant capabilities.

One of the major factors contributing to the widespread interest in and use of AI technologies is the ready availability of open-source toolkits and collaboration communities (see Figure 5.16). Researchers can quickly download software packages that contain working implementations of the latest algorithms ready to be applied to new problems. While private companies will protect special updates and additions to the basic algorithms, as it represents a business advantage, these open-source resources have helped to move things along at a very rapid pace [35].

This trend in availability of open-source solutions can be applied to cyber security as well. Techniques and algorithms that show promise can be bundled into libraries that others can use to move the field forward. However, detailed technical knowledge of the cyber domain is required to ensure that these techniques and algorithms are applied correctly to achieve maximal impact.

Figure 5.17. Human-machine teaming can help with resource challenges faced by the government.

Once an algorithm has been trained and is ready for use, it must be incorporated into an existing human-centric workflow. Through this human-machine teaming, the algorithm can enhance and augment the human's capabilities to improve overall effectiveness. This is especially important given the government's stagnant or declining resources available to apply to the problem, as highlighted in Figure 5.17.

Recommender systems, in particular, have been leveraged with great success, helping humans find relevant information for entertainment and shopping purposes [34, 36]. These same capabilities could be helpful in making the cyber analyst and security specialist more effective at their job functions, as the capabilities enable all individuals to benefit from the successes of a few.

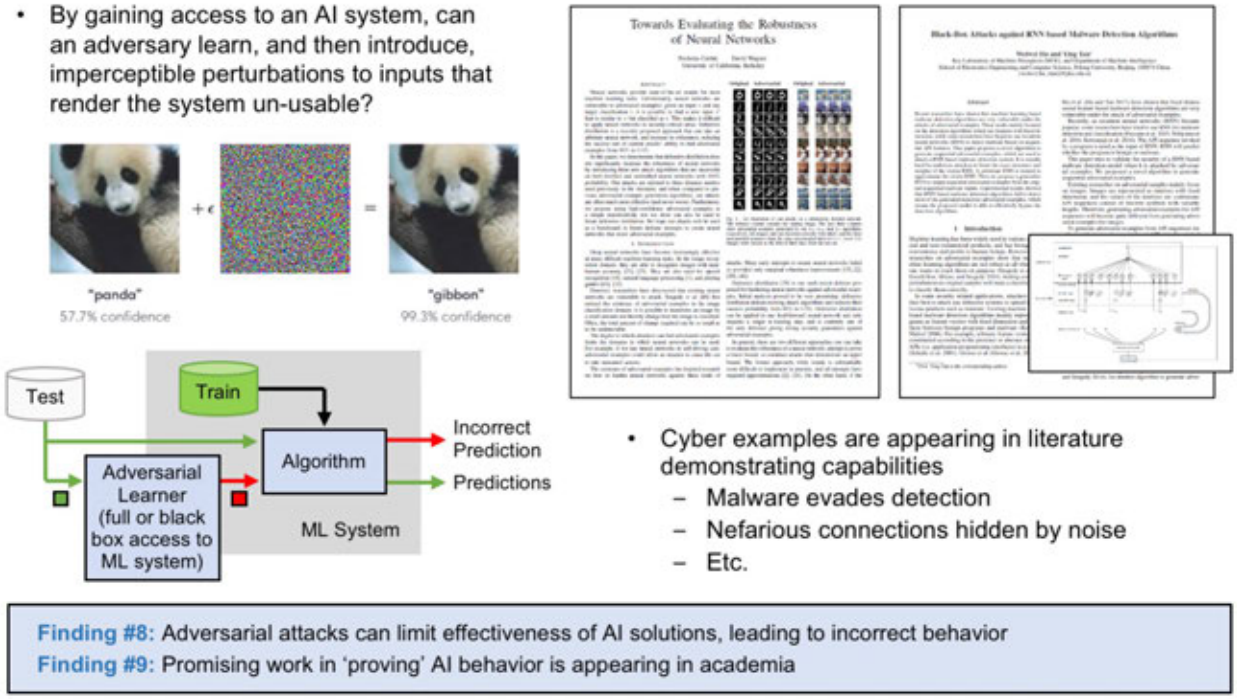


Figure 5.18. AI for cyber must be robust, especially against a motivated and capable adversary who can launch adversarial learning attacks.

Machine-learning capabilities have recently been shown to offer astounding ability to automatically analyze and classify large amounts of data in complex scenarios, in many cases matching or surpassing human capabilities. However, it has also been widely shown that these same algorithms are vulnerable to attacks, as depicted in Figure 5.18, known as adversarial learning attacks, which can cause the algorithms to misbehave or reveal information about their inner workings [37]. In general, attacks take three forms: 1) data poisoning attacks inject incorrectly or maliciously labels data points into the training set so that the algorithm learns the wrong mapping; 2) evasion attacks perturb correctly classified input samples just enough to cause errors in classification; and 3) inference attacks repeatedly test the trained algorithm with edge-case inputs in order to reveal the previously hidden decision boundaries. Protection against adversarial learning attacks include techniques that cleanse training sets of outliers in order to thwart data poisoning attempts, and methods that sacrifice up-front algorithm performance in order to be robust to evasion attacks [38]. As machine-learning-based AI capabilities become incorporated into facets of everyday life, including protecting cyber assets, the need to understand adversarial learning and address it becomes clear.

Adversarial learning is a particular type of cyber attack in which the implementation of an algorithm (i.e., machine learning) is attacked through adversary actions. In fact, recent protections against adversary learning are leveraging other typical cyber security defenses, such as differential privacy [39]. AI techniques such as co-evolutionary computation may help make AI systems robust to adversarial interference [40].

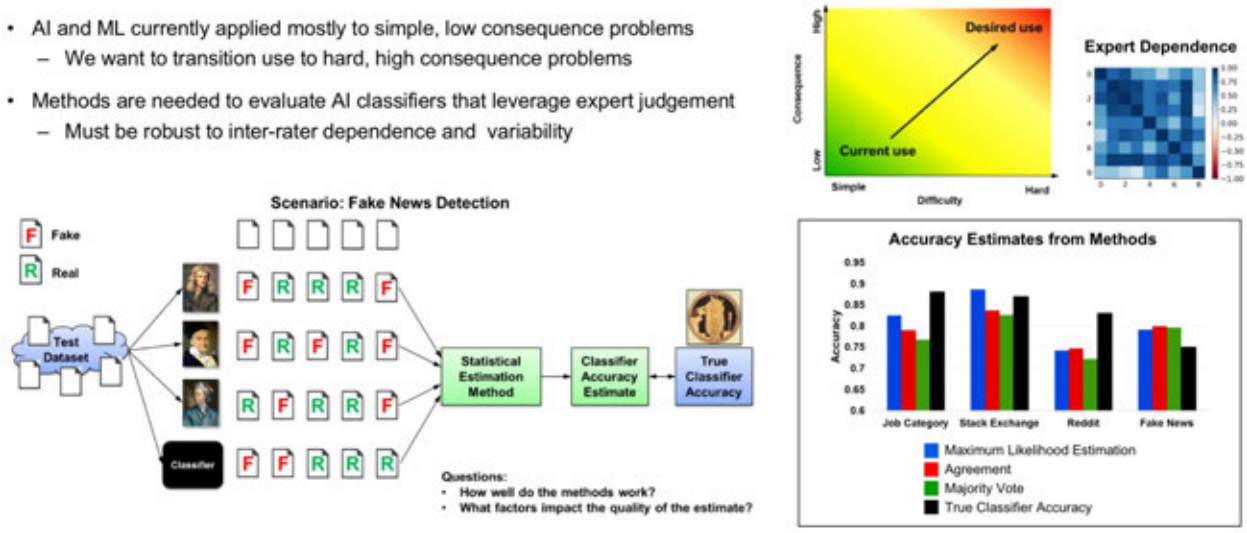


Figure 5.19. As AI systems become more capable, the evaluation classifier performance using human expert judgment shows promise.

Evaluating AI performance remains a challenge, especially when that performance exceeds human capability. One approach to measuring AI performance involves leveraging expert human judgment to assess black-box classifiers. As depicted in Figure 5.19, current approaches include statistical methods that combine individual expert judgment through maximum likelihood estimation (MLE), agreement [41], or majority vote to arrive at an estimate of true classifier accuracy [42]. In these cases, it is seen that performance of the methods greatly depends upon the independence of the experts as well as the difficulty of the problem at hand.

5.4 Recommendations and Way Forward

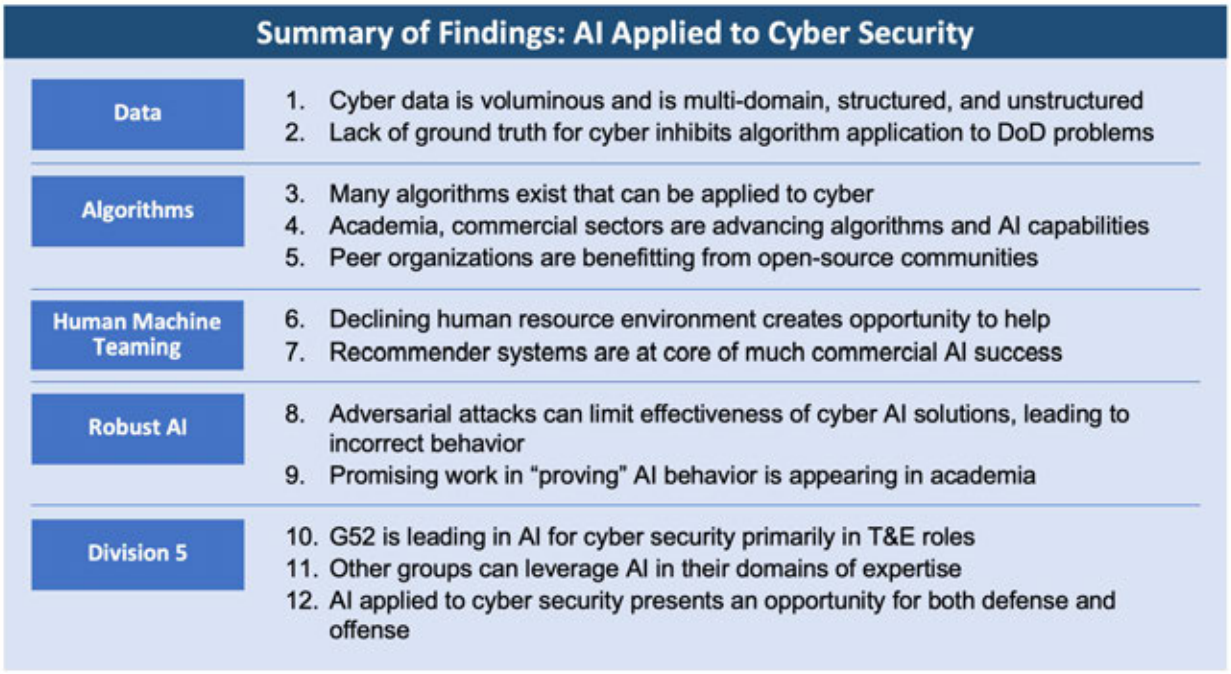


Figure 5.20. Findings from the AI for cyber study include the need for truth-marked cyber data, advanced algorithm research in academia, and robust AI.

In summary, several key findings were discovered from our survey of AI capabilities for cyber. As listed in Figure 5.20, these findings indicate that although AI is having a large impact on many technology areas (e.g., image recognition, health care) much remains to be done for similar impact to be seen in the cyber domain.

The needs as specified by the findings provide an opportunity for the development of AI systems to shape how the DoD adopts AI in order to be more operationally effective and resilient to adversary attacks.

First among the recommendations from our study is that the AI developers should lead the way in addressing the truth-marked dataset gap that currently exists in cyber security research. Based on years of experience in cyber as well as other fields, AI research laboratories, such as MIT LL, can collect, create, and curate multiple large cyber datasets along with relevant truth data to help develop and enhance AI algorithms for DoD use.

The second recommendation will leverage MIT LL's strong academic connection to remain aware of the latest advances in AI so that they can be brought to DoD problem spaces. In addition, we should immediately begin incorporating open-source toolkits and libraries for AI so as to jumpstart relevant capabilities. MIT LL will rely upon its extensive mission knowledge and data access to develop solutions that the academic and commercial communities are not focused on, at this point.

Building upon recent efforts in workflow understanding and task automation, MIT LL can play a key role in automating mundane cyber tasks, such as initial data triage and alert processing in order to allow over-worked analysts and security personnel to focus on the most relevant concerns of the time.

Our fourth recommendation is to continue current research into automated decision making so that the current prototype capabilities (e.g., CASCADE) can be leveraged in a real-time

operational setting [43]. Further refinement of mission and threat models will enable the considered and suggested solutions to be more relevant and appropriate for real-world scenarios. In addition, enhanced optimizations algorithms as well as instantiation within a big data computational environment will enhance its potential impact on mission decision needs.

Developing AI capabilities that can be trusted and understood and that are resilient to adversarial attacks is a major challenge facing the adoption of AI solutions across the DoD today. This lack of assurance that these data-driven algorithms will perform as expected during operational scenarios keeps them from being used on a regular basis. AI researchers can play a major role in helping to uncover ways to make AI algorithms, such as DNNs and other machine learning capabilities, robust to adversary attacks. Recent work into methods to measure high performing classification systems will lead to improved methods for training and evaluating AI solutions.

As shown in Figure 5.21, the DoD can benefit greatly from recent advances in AI. Capabilities for data understanding, counterfeit detection, intelligence gathering, and even offensive planning can be augmented by intelligent algorithms that can learn from gathered data. AI researchers with access to DoD sponsor problems, as well as their data, can enable research and development along these lines and will help to develop end-to-end AI systems enabling the DoD to adopt AI for cyber security and operations.

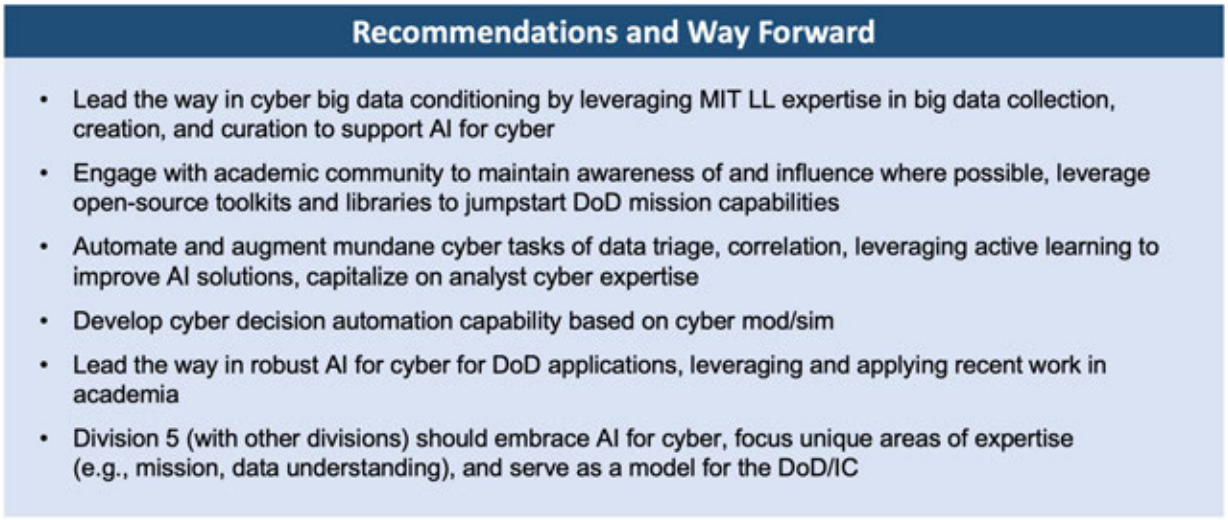


Figure 5.21. Recommendations for MIT LL to support AI for the DoD cyber mission include leading the way in data curation and collection, automating mundane cyber tasks, and in developing and transitioning robust AI solutions.

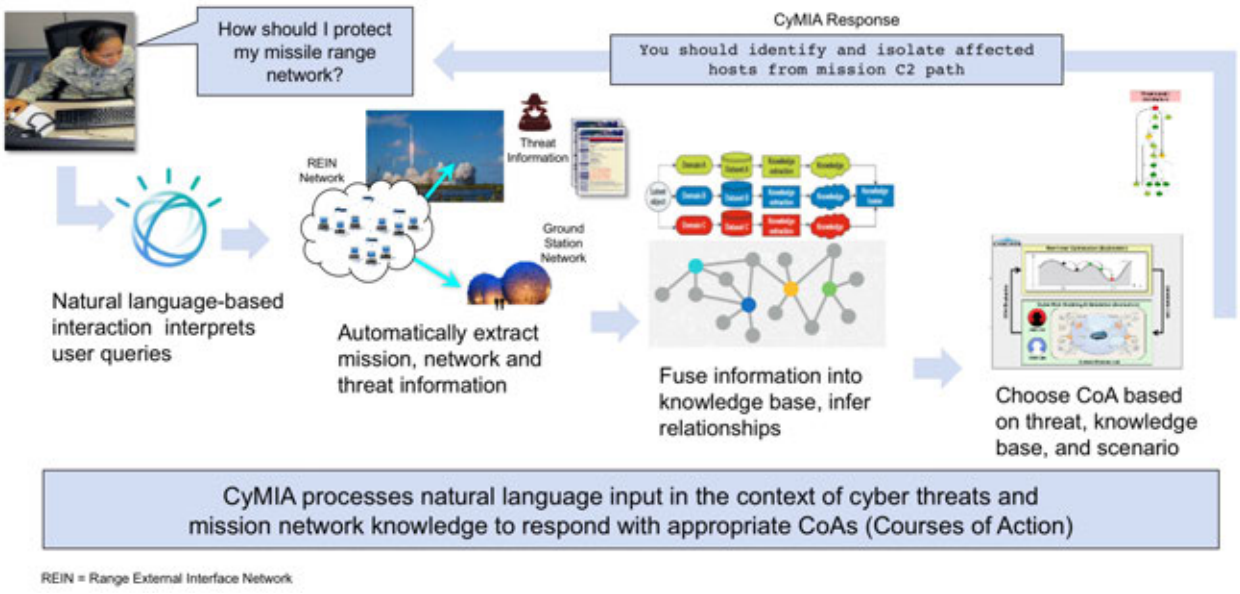


Figure 5.22. Cyber Machine Intelligent Assistant (CyMIA) leverages natural language processing, automated mission and threat mapping, and automated decision assistance to enable a security analyst protect their mission critical system against evolving cyber threats.

As cyber attacks continue to grow in sophistication and impact, the mission cyber defender grows evermore overwhelmed in defending critical cyber mission assets. Challenges faced by the defender include: inability to track potentially relevant cyber threats and to understand which cyber assets are critical to mission function, difficulty in fusing threat and mission information in order to explore potential impacts, and determining optimal responses to the threat while still protecting mission functionality. Current approaches for dealing with this involve manual defensive processes relying upon technical-only solutions and using cyber-only data. AI holds promise for shifting the balance of the cyber struggle in favor of the cyber defender by enabling early attack warning and preemptive mitigation deployment. Here, deeper analyses of attacker intention and behavior against a specific mission system target support machine-vs.-machine cyber combat by leveraging a range of multi-domain data sources. As depicted in Figure 5.22, CyMIA leverages existing AI capabilities in online social media mining for early indications of cyber attack, as well as those for automated cyber decision to realize leap-ahead AI-enabled cyber warriors whose effectiveness at defending mission cyber systems is greatly enhanced.

CyMIA consists of four distinct technology components. At the front of the system, a human-machine interface based upon a current off-the-shelf natural language detection capability (e.g., IBM Watson) permits the expression of a cyber-concern of the mission security analyst. In preparation for consideration of the concern in a mission context, the next stage leverages automatic critical asset identification capabilities (e.g., mission mapping) and autonomous capabilities for identifying and collecting threat information to build a knowledge graph of relevant cyber, mission, and threat information. The third stage of CyMIA mines the constructed knowledge graph to build probabilistic scenario models of the attacker, defender, mission, and network for use by the automated decision engine, CASCADE [44]. CASCADE explores the cyber mission scenario by combining a hill-climbing optimization algorithm with a mod/sim-based evaluation engine in an iterative fashion in order to arrive at an optimal COA.

The following tasks must be undertaken in order to develop the end-to-end CyMIA system and demonstrate its capability:

1. **Human-Machine Interface:** Leverage and augment off-the-shelf capabilities in human language detection to develop an analyst interface that accepts and interprets a query about the state of the mission network.
2. **Automated Mission and Threat Mapping**: Develop a mapping of the mission network and the relevant cyber threats by leveraging and extending automated capabilities. To develop mission mapping capabilities, bottom-up statistical inference methods (e.g., co-occurrence) as well as top-down, specification methods must be explored.
3. **Knowledge Fusion and Inference:** Develop a graph-based knowledge base of information combining mission system and cyber threat information for the purpose of discovering relevant relationships. Analysis and inference upon the knowledge base will support the creation of probabilistic models necessary for specification of the relevant scenario to be used as input to the next stage of COA selection.
4. **Scenario Evaluation and COA selection:** Leverage and extend the CASCADE automated decision tool for evaluating and selecting COAs suitable for mitigating relevant threats to the mission system, as represented by the four scenario models created in the previous stage: attacker, defender, mission, and network.

As cyber mission stakeholders become increasingly overwhelmed with the challenge of collecting and assimilating threat information and determining its relevance to their mission system, CyMIA will provide an effective, near-real-time automated capability for responding to concerns with an automated decision analysis capability and optimal COA suggestion.

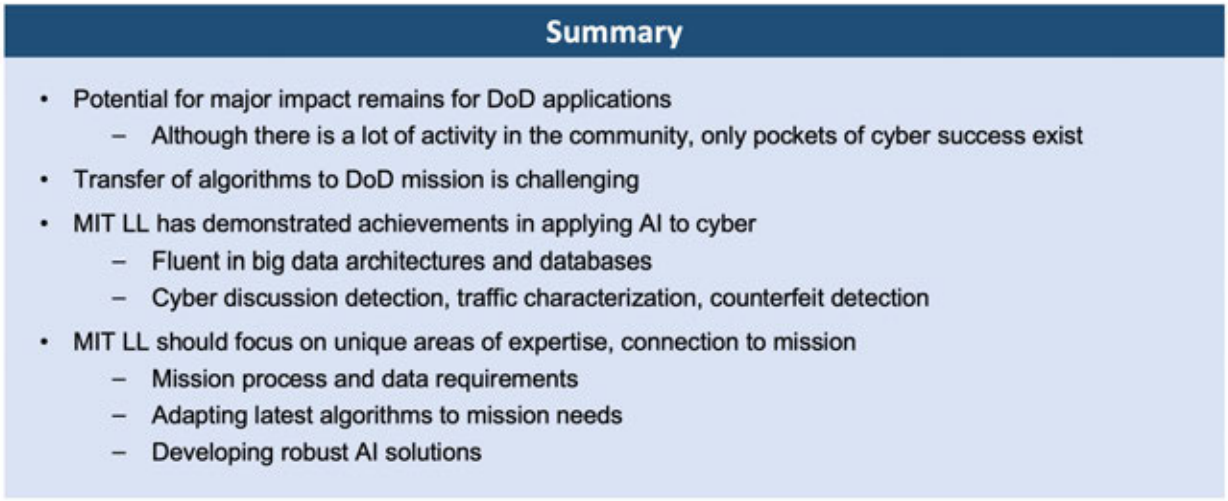


Figure 5.23. AI for cyber study summary.

In summary (see Figure 5.23), there remains significant potential for major impact of AI to DoD applications. At the present time, while there are pockets of activity, there should be more pervasive and comprehensive activity to leverage AI in cyber operations, both offensively and defensively.

Challenges in transitioning algorithms and AI-enabled capabilities to the DoD continue to remain. Among them are the need for sufficient truth-marked data to train the algorithms, the need to map the algorithms to the DoD problem set, and the need to establish trust for these difficult-to-explain capabilities.

MIT LL—and Division 5, in particular—has demonstrated achievements in applying AI to cyber problems. These include design and development of big data architectures and data platforms and the development of algorithms to address pressing mission challenges in cyber anomaly detection, network traffic characterization, and in the detection of counterfeit computer components, such as chips.

To enable the government to leverage the current AI boom to the fullest extent possible, AI researchers should focus on unique areas of expertise and its strong connection to the mission. In particular, the AI prototype developers should leverage knowledge of mission processes and purpose as well as access to relevant data to help map non-DoD algorithmic solutions to DoD problems. Finally, AI researchers should focus on developing robust AI solutions, by focusing on explainability, verification, and validation of developed capabilities, and techniques for ensuring AI-enabled systems are secure against adversarial learning attacks, including poisoning, evasion, and model inversion.

References

1. Marechal, S., *Advances in password cracking.* Journal in computer virology, 2008. **4**(1): p. 73-81.
2. Halevi, T., N. Memon, and O. Nov, *Spear-phishing in the wild: A real-world study of personality, phishing self-efficacy and vulnerability to spear-phishing attacks.* 2015.
3. Hiltgen, A., T. Kramp, and T. Weigold, *Secure internet banking authentication.* IEEE Security & Privacy, 2006. **4**(2): p. 21-29.
4. Meer, H., *Memory corruption attacks the (almost) complete history.* Blackhat USA.(Jul. 2010), 2010.
5. Younan, Y., W. Joosen, and F. Piessens, *Runtime countermeasures for code injection attacks against C and C++ programs.* ACM Computing Surveys (CSUR), 2012. **44**(3): p. 17.
6. Bykova, M. and S. Ostermann. *Statistical analysis of malformed packets and their origins in the modern Internet*. in *Proceedings of the 2nd ACM SIGCOMM Workshop on Internet measurment*. 2002. ACM.
7. Abrams, M. and J. Weiss, *Malicious control system cyber security attack case study–Maroochy Water Services, Australia.* McLean, VA: The MITRE Corporation, 2008.
8. Smith, G.E., et al., *A critical balance: collaboration and security in the IT-enabled supply chain.* International journal of production research, 2007. **45**(11): p. 2595-2613.
9. Maynor, D., *Metasploit toolkit for penetration testing, exploit development, and vulnerability research*. 2011: Elsevier.
10. Martin, L., *Cyber Kill Chain®.* https://cyber.lockheedmartin.com/hubfs/Gaining_the_Advantage_Cyber_Kill_Chain.pdf, 2014.
11. Barrett, M.P., *Framework for Improving Critical Infrastructure Cybersecurity Version 1.1*. 2018.
12. Goodall, J.R., A. D'Amico, and J.K. Kopylec. *Camus: Automatically mapping cyber assets to missions and users*. in *Military Communications Conference, 2009. MILCOM 2009. IEEE*. 2009. IEEE.
13. Okhravi, H., et al., *Finding focus in the blur of moving-target techniques.* IEEE Security & Privacy, 2014. **12**(2): p. 16-26.
14. Marpaung, J.A., M. Sain, and H.-J. Lee. *Survey on malware evasion techniques: State of the art and challenges*. in *Advanced Communication Technology (ICACT), 2012 14th International Conference on*. 2012. IEEE.
15. Franke, U. and J. Brynielsson, *Cyber situational awareness–a systematic review of the literature.* Computers & Security, 2014. **46**: p. 18-31.
16. Wheeler, D.A. and G.N. Larsen, *Techniques for cyber attack attribution*. 2003, Institute for Defense Analysis, Alexandria VA.
17. Musman, S., et al., *Evaluating the impact of cyber attacks on missions.* MITRE Corporation, 2010.

18. Musman, S., et al. *Computing the impact of cyber attacks on complex missions*. in *Systems Conference (SysCon), 2011 IEEE International*. 2011. IEEE.
19. Walker, M. *Machine vs. Machine: Lessons from the First Year of Cyber Grand Challenge*. in *Proceedings of the 24th USENIX Security Symposium*. 2015.
20. Price, B., et al., *House Rules: Designing the Scoring Algorithm for Cyber Grand Challenge.* IEEE Security & Privacy, 2018. **16**(2): p. 23-31.
21. Sapienza, A., et al. *Early warnings of cyber threats in online discussions*. in *Data Mining Workshops (ICDMW), 2017 IEEE International Conference on*. 2017. IEEE.
22. Griffin, K., et al. *Automatic generation of string signatures for malware detection*. in *International workshop on recent advances in intrusion detection*. 2009. Springer.
23. Lippmann, R.P., et al., *Finding malicious cyber discussions in social media*. 2016, MIT Lincoln Laboratory Lexington United States.
24. Sommer, R. and V. Paxson. *Outside the closed world: On using machine learning for network intrusion detection*. in *Security and Privacy (SP), 2010 IEEE Symposium on*. 2010. IEEE.
25. Cisco, V., *Cisco Visual Networking Index: Forecast and Methodology 2016-2021.* 2017.
26. Lippmann, R., et al. *Analysis and results of the 1999 DARPA off-line intrusion detection evaluation*. in *International Workshop on Recent Advances in Intrusion Detection*. 2000. Springer.
27. Scheper, C., S. Cantor, and D. Maughan. *PREDICT: a trusted framework for sharing data for cyber security research*. in *Proceedings of the First Workshop on Building Analysis Datasets and Gathering Experience Returns for Security*. 2011. ACM.
28. Deng, L., *The MNIST database of handwritten digit images for machine learning research [best of the web].* IEEE Signal Processing Magazine, 2012. **29**(6): p. 141-142.
29. Deng, J., et al. *Imagenet: A large-scale hierarchical image database*. in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. 2009. Ieee.
30. Benjamin, P., *System for intrusion detection and vulnerability assessment in a computer network using simulation and machine learning*. 2010, Google Patents.
31. Ingols, K., R. Lippmann, and K. Piwowarski. *Practical attack graph generation for network defense*. in *Computer Security Applications Conference, 2006. ACSAC'06. 22nd Annual*. 2006. IEEE.
32. Axelsson, S., *The base-rate fallacy and the difficulty of intrusion detection.* ACM Transactions on Information and System Security (TISSEC), 2000. **3**(3): p. 186-205.
33. Carlini, N. and D. Wagner. *Adversarial examples are not easily detected: Bypassing ten detection methods*. in *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*. 2017. ACM.
34. Adomavicius, G. and A. Tuzhilin, *Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions.* IEEE Transactions on Knowledge & Data Engineering, 2005(6): p. 734-749.
35. Marvin, R.H., Brian, *10 Steps to Adopting Artificial Intelligence in Your Business.* PC Magazine, 2018. https://www.pcmag.com/article/350965/10-steps-to-adopting-artificial-intelligence-in-your-business
36. Bobadilla, J., et al., *Recommender systems survey.* Knowledge-based systems, 2013. **46**: p. 109-132.
37. Huang, L., et al. *Adversarial machine learning*. in *Proceedings of the 4th ACM workshop on Security and artificial intelligence*. 2011. ACM.
38. Papernot, N., et al. *The limitations of deep learning in adversarial settings*. in *Security and Privacy (EuroS&P), 2016 IEEE European Symposium on*. 2016. IEEE.
39. Lecuyer, M., et al., *On the Connection between Differential Privacy and Adversarial Robustness in Machine Learning.* arXiv preprint arXiv:1802.03471, 2018.
40. Hemberg, E., et al. *Adversarial co-evolution of attack and defense in a segmented computer network environment*. in *Proceedings of the Genetic and Evolutionary Computation Conference Companion*. 2018. ACM.

41. Lehner, P.E. *Estimating the Accuracy of Automated Classification Systems Using Only Expert Ratings that are Less Accurate than the System*. 2014.
42. Platanios, E.A., A. Dubey, and Mitchell, *Estimating accuracy from unlabeled data: A bayesian approach*. 2016.
43. Wagner, N., et al. *Towards automated cyber decision support: A case study on network segmentation for security*. in *Computational Intelligence (SSCI), 2016 IEEE Symposium Series on*. 2016. IEEE.
44. Wagner, N., et al., *Quantifying the mission impact of network-level cyber defensive mitigations.* The Journal of Defense Modeling and Simulation, 2017. **14**(3): p. 201-216.

6 Future Outlook (D. Martinez)

We tend to overestimate the effect of a technology in the short run and underestimate the effect in the long run.—Amara's Law

In this section, we present a roadmap for AI investments. In previous sections, the recommendations were focused on specific applications to for example, AI for Human Language Technology in Section 4, and AI for Cyber Security in Section 5. In both of these specific recommendations, the emphases were on areas of relevance to Division 5. In this last section of the report, we make more broadly applicable investment recommendations relevant to the DoD, IC communities, and Homeland Security.

6.1 Three Horizons for AI S&T Investments

To effectively characterize S&T investments, we formulated the conceptual framework shown in Figure 6.1. The vertical axis is notional representing AI capability impact. The horizontal axis represents development time.

For example, for Horizon 1, we recommend starting AI investments now so that the capabilities are operational in one to two years. Similarly, for Horizon 2, we also recommend starting investing now, so AI capabilities are available in operations in three to four years. For Horizon 3, if we invest now, at the S&T level, we would look at those capabilities to be available operationally in five or more years.

Horizon 1 targets achieving robust content-based insight. Recall that the AI canonical architecture represents a framework for transforming data into insight that users can then use to augment their capabilities to make timely decisions. More specifically, this near-term investment would be focused on applying AI to gain insight on interesting content present in disparate types of data—both structured and unstructured. The main benefits of these investments are:

- Reduced user workload
- Improved confidence in AI (see Figure 2.9)
- Lower consequence of actions (see Figure 2.9)
- Robust AI

Horizon 2 focuses on collaboration-based insight. This requires that we extend AI roles to include multiple human-machine teams working together. It has been demonstrated that a group of humans aided by a machine (or multiple machines) can outperform an expert in difficult cognitive tasks. Humans are much better than a machine at making subjective judgements, disambiguating options, and understanding context [1, 2]. The main benefits of these investments are:

- Enriched insight
- Collaboration across intelligent machines
- The ability to migrate more routine tasks to AI enabled systems

Horizon 3 addresses context-based insight. In contrast to content-based insight, performing intelligent tasks by machines using context is an area AI systems do not do well today. An AI system, incorporating context, should be able to refine its recommendations with high degree of confidence by incorporating relevant knowledge from other related inputs. For example, in the

application of terrorist countermeasures, one can envision an AI system that can classify a suspected vehicle but also refine the answer by knowing that such vehicle is parked in a compound identified as housing a terrorist cell. The main benefits of these investments are:

- Reduced time from data to reaching insight with high confidence
- Increased symbiosis between humans and machines

DARPA is emphasizing further research consistent with Horizon 3. It coined the term *Wave 3* (contextual reasoning) in reference to increasing human-machine symbiosis and making machines more of a partner to the user. The new thrust at DARPA is referred to as the AI Next Campaign.

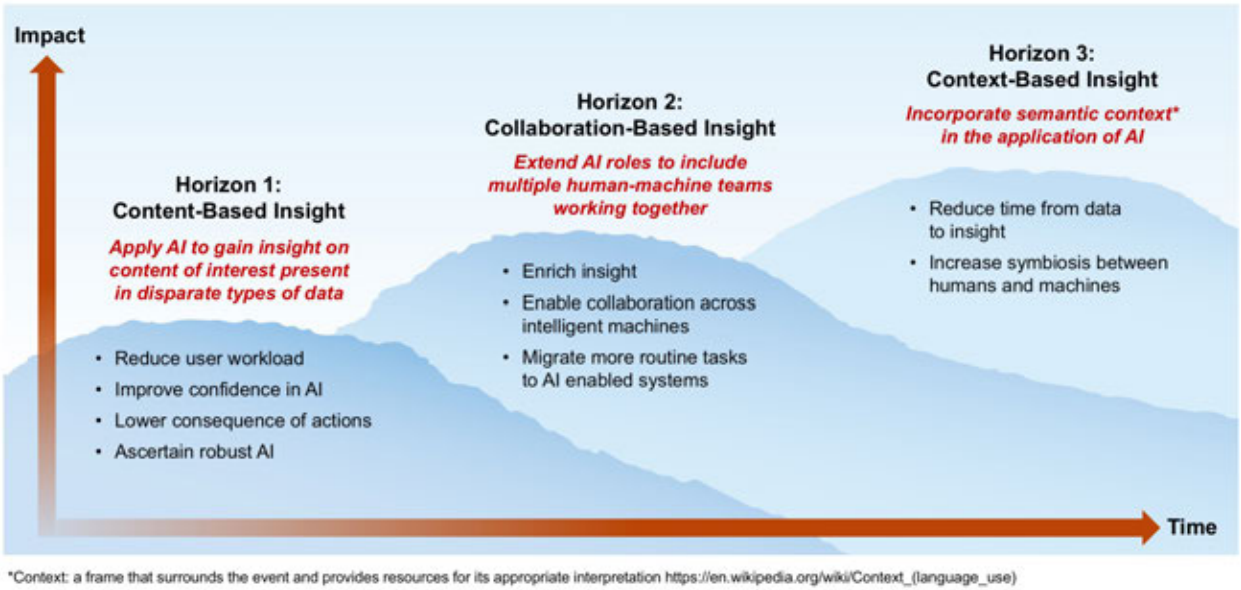


Figure 6.1. Definition of AI investment horizons.

In Figure 6.2, we map investment horizons across key subcomponents described in the AI canonical architecture. The high-level recommendations, which we elaborate in more detail in Section 6.2, are:

- For data conditioning: Develop common database formats across a wide range of multi-modal sensors and sources for both structured and unstructured data.
- For algorithms: Demonstrate machine-learning techniques across multi-domains to allow us to transform information into knowledge.
- For modern computing: Achieve real-time performance on end-to-end AI systems. This is most critical for tactical edge applications where timely courses of action are required.
- For human-machine teaming: Augment the human capabilities by leveraging intelligent machines. This is at the core of any narrow AI system.
- For robust AI: Build a culture of rapid development cycles with user in the loop. This synergistic interplay between developers and users will likely increase trust in the acceptance of AI systems.

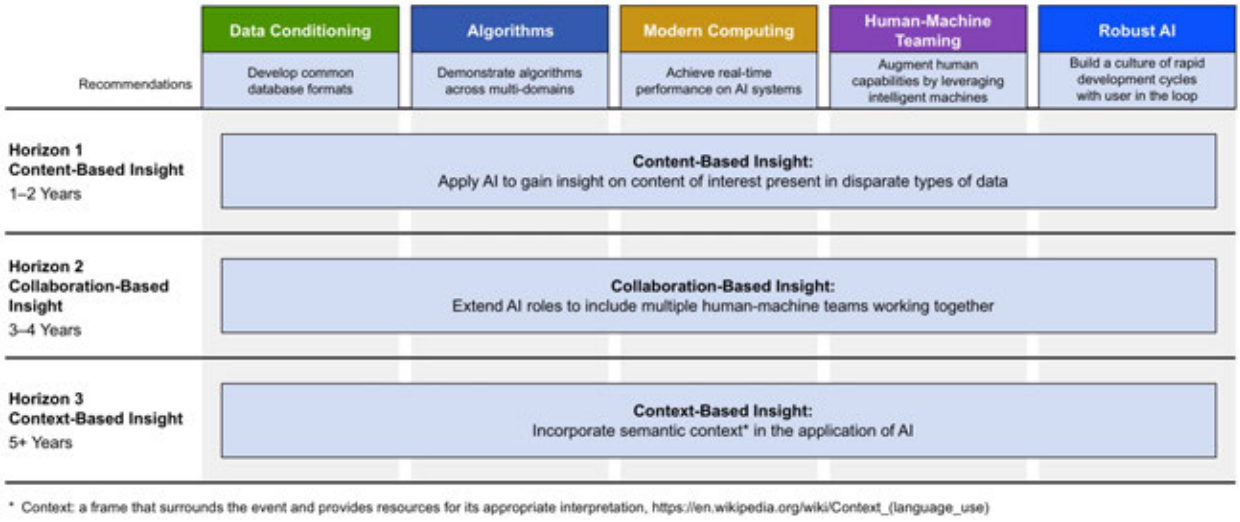


Figure 6.2. Investment horizons across key subcomponents of the AI canonical architecture.

The number of specific S&T recommendations can be substantial for any given application. Thus, in Figure 6.3, we opted to identify specific recommendations that are needed across a broad range of applications. These recommendations are important to achieve substantial AI capabilities across each of the respective horizons without being limited to any one application, and relative to each respective AI canonical architecture subcomponent. In Sections 4 and 5, we chose two areas of interest, and made recommendations relevant to those specific applications. In Section 6.2, we elaborate in more detail on the entries shown in Figure 6.3 and associated benefits.

Recommendations	Data Conditioning Develop common database formats	Algorithms Demonstrate algorithms across multiple domains	Modern Computing Achieve real-time performance on AI systems	Human-Machine Teaming Augment human capabilities by leveraging intelligent machines	Robust AI Build a culture of rapid development cycles with user in the loop
Horizon 1 **Content-Based Insight** 1–2 Years	• Automate data labeling • Create dataset benchmarks (gold standard)	• Blend unsupervised and supervised learning • Demonstrate reinforcement learning	• Accelerate model generation • Deploy computing to the edge	• Relationship graphs updated in real time • Advance natural language processing (NLP)	• Develop robust AI metrics • Demo adv. learning • Design adversarial AI countermeasures
Horizon 2 **Collaboration-Based Insight** 3–4 Years	• Aggregate real data, simulated data, and prior knowledge • Access intermediate results	• Advance algorithm accuracy through collaboration • Exploit physics and causal relationships	• Compute across distributed platforms • Reduce SWaP for embedded and IoT devices	• Transparent human-machine teams • Collab. based on achieving mission goals	• Strengthen full end-to-end system security • Enable explainable AI
Horizon 3 **Context-Based Insight** 5+ Years	• Generate models from limited infor. • Exploit the what, how, who, and why to build context • Operate in denied and degraded environments	• Train with limited data • Low-shot or one-shot learning • Context-aware learning • Advance research on goal reasoning	• Advance cognitive computing • Deterministic to probabilistic computing • Scalability based on mission objectives	• Understand & shape human-machine networks • Sentiment analysis • Scale to very large human-machine teams	• Gain user confidence through probabilistic courses of action • Leverage context to defend against adv. learning attacks • Employ formal math to verify performance
DoD Investment Priority	High	Medium	Medium	High	High

Note: Investment priorities are relative to DoD uniqueness vs. commercial applications

Figure 6.3. Specific investment recommendations.

6.2 Investment Recommendations

Based on the AI canonical architecture and the investment horizons illustrated earlier, in this section, we elaborate on the recommendations for S&T investments. Figure 6.4 describes a set of recommendations that are relevant to all investment horizons. One of the important issues facing the application of AI to national security problems is the need to rapidly insert AI capabilities into operational use. The typical insertion of capabilities based on prototyping initial operating capability (IOC) and production taking decades is not viable for AI. The AI advances progress much faster than what most people are used to in national security applications. Therefore, there is a need to establish a culture of innovation, rapid experimentation, and deployment. Innovation hubs in Silicon Valley use the driving principle of rapid, disciplined risk reduction to "fail fast, fail often, fail forward."

Another important broad recommendation, shown in Figure 6.4, is to focus on end-to-end capabilities across all elements of the AI canonical architecture discussed in earlier sections. The users care about the end-to-end capability provided by AI. In some cases, a suboptimal algorithm might be sufficient to still achieve substantial improvements. For example, a machine-learning technique based on time frequency–inverse document frequency (TFIDF) followed by a logistic regression classifier is very effective in reducing a cyber analyst's workload. The main takeaway from this recommendation is that unless we look at capabilities through the lens of an end-to-end AI system, we can end up over-designing the subcomponents needed for an effective operational system.

As shown in Figure 6.4 and in all subsequent figures that elaborate on investment recommendations, the format we use is to outline the specific recommendations and the benefits accrued from those recommendations. Thus, to address the challenge of advancing AI capabilities while rapidly inserting them into operational use, we need to act on the specific recommendations highlighted in Figure 6.4. Respective recommendations would result in multiple benefits, also highlighted in Figure 6.4. Early user buy-in is paramount for AI systems to be accepted operationally because ultimately, users need to judge how well an AI capability is improving their present workflow approach. Users will have an easier time doing this assessment if we have quantifiable metrics and benchmarks, including the ability to explain why the AI system is reaching its output. These metrics should also incorporate an assessment on AI biases.

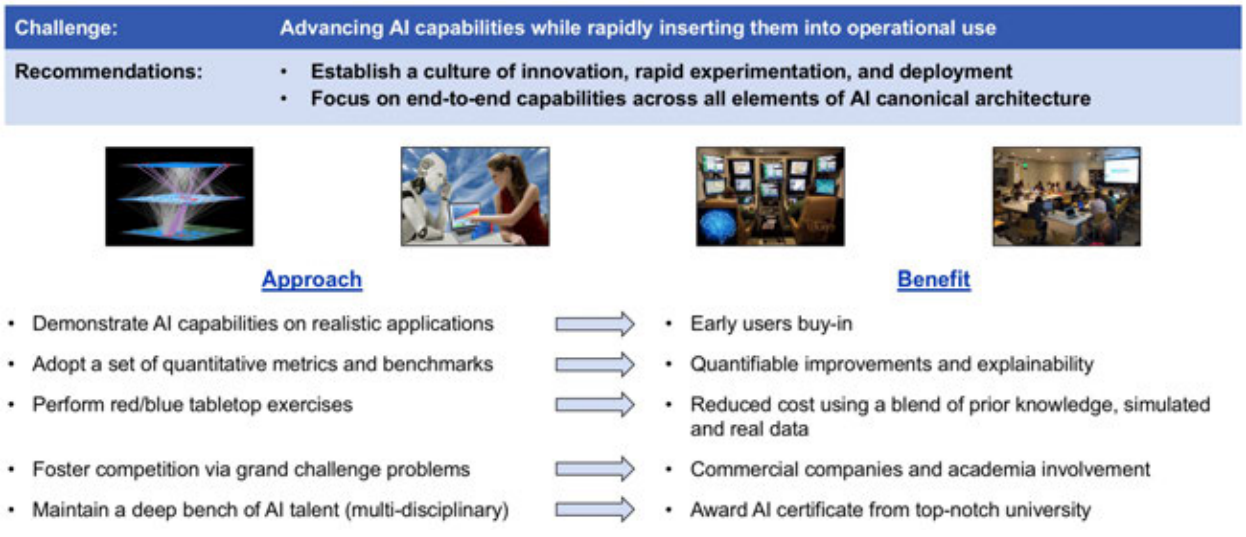


Figure 6.4. Broad recommendations on way forward—relevant across all horizons.

A very effective way to demonstrate the value of an AI system is to perform red/blue tabletop exercises. This approach facilitates relatively quick assessments under somewhat realistic scenarios. These tabletop exercises have the additional benefit of being able to incorporate prior knowledge, simulated data, emulated data, and real data. One variant of this approach is known as serious games, described in a number of books and demonstrations performed, for example, by MIT LL and others, including commercial agile environment communities [3-5].

Fostering competitions via grand challenges will result in motivating commercial companies and academia to be involved. For example, from 2004 through 2007, DARPA carried out a number of driverless cars grand challenges [6, 7] that transformed commercial industries, as we know it today. There were several competing vehicles developed at a number of universities, including CMU, Stanford, MIT, and others. This type of investment is also needed in AI to advance the AI ecosystem in significant ways.

In addition to advancing AI capabilities, there is a need to maintain a deep bench of AI talent. This talent pool must be cross-disciplinary including data scientists, experts in machine-learning algorithms, social scientists, and personnel that know how to rapidly test and evaluate AI systems. One approach, among many, is to develop curriculums that lead to either micro-degree programs or university certificates.

Section 3 presented details on enabling technologies critical to any AI system. In Figure 6.5, we highlight the recommendation objective for the data conditioning subcomponent of the AI canonical architecture. The objective for this subcomponent is to enable common database formats (including storing of all intermediate results to facilitate refinement of the AI algorithms). This step will enable transforming structured and unstructured data into information suitable for insertion into the algorithm subcomponent of the AI canonical architecture.

The format we adopted for each of the architecture subcomponents is to outline the specific recommendations shown in Figure 6.3, across all horizons followed by benefits that these recommendations would provide to the AI system user. For example, by investing in techniques to enable automated data labeling, then one would end up with automated tools to discover, link, and store heterogenous data. Also, by automating the process of labeling data, the user would benefit by achieving significant reduction in the time needed to develop an end-to-end AI system.

Likewise, many of the specific recommendations outlined in Figure 6.5 lead to important benefits such as automated cleaning and preprocessing of data. Also, by having the ability to access intermediate results, it will help the user increased understanding of AI intermediate outputs to better explain the output of the AI algorithms. Data conditioning exploiting the what, how, who, and why to build context will facilitate low-shot or one-shot learning [8, 9]. These classes of learning will be necessary for cases where having large amounts of labeled data will not be feasible. Finally, it is very important to invest in technologies where the AI system needs to operate in denied and degraded environments. This type of investment, in many cases, is unique to the DoD, IC, and Homeland Security applications.

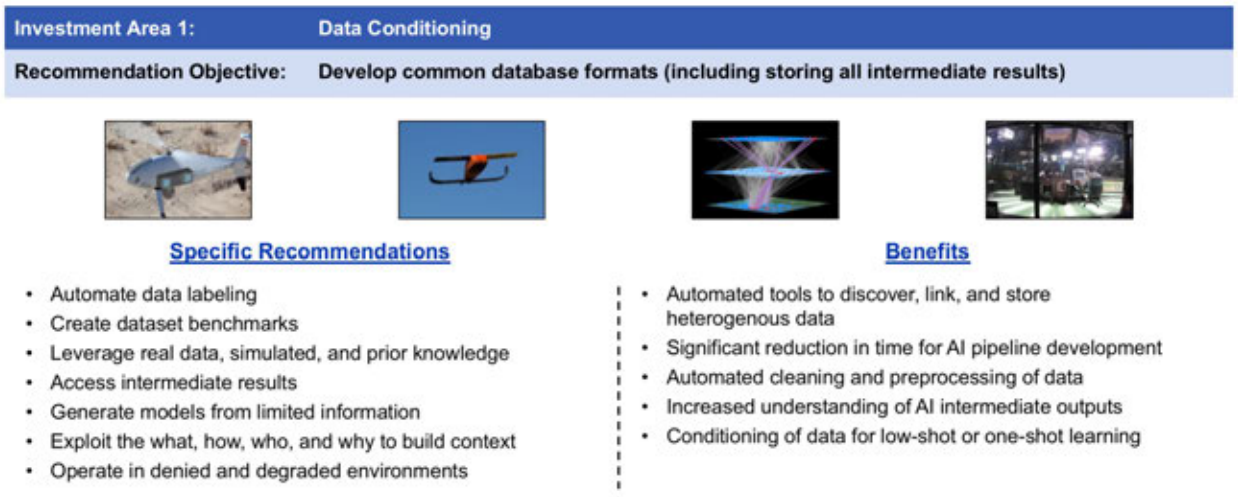


Figure 6.5. Recommendations on data conditioning.

Broadly speaking, the recommendation objective for the algorithm subcomponent is to demonstrate its application across multi-domains; multi-domains refer to inputs in the form of videos, images, text, speech, cyber, etc. The algorithm subcomponent shown in the AI architecture enables transforming information into knowledge. One example of information, at the input of the algorithm subcomponent from the output of the data conditioning subcomponent, is a list of viruses plus a database containing those viruses that have been converted into exploits. It has been noted that today only 10% of the identified viruses are converted into exploits. A type of algorithm, to transform information into knowledge, is to use a supervised machine-learning logistic regression algorithm. The output will be a list of exploited viruses (knowledge) that a group of cyber protection teams can then attend to as most critical in their protection of their systems (as described earlier in Section 5).

Most algorithm approaches today use supervised learning techniques. In the future, we will need a blend of both supervised and unsupervised learning techniques. Unsupervised learning techniques have the benefit of not needing labeled data. Another approach receiving significant attention in the literature is reinforcement learning. Reinforcement learning is based on a reward-based approach where the neural networks learn as they converge closer to the right answer and are therefore rewarded for it.

We also envision that algorithm accuracy will be improved by getting feedback from the user through collaboration. Another way to improve algorithm accuracy is to leverage physics of the environment (e.g., standard vehicles are not going to appear to be driving at speeds exceeding, say, 300 MPH)—this type of algorithm output would be in error. Similarly, leveraging causal relationship would help in introducing context into the algorithms [10]. Another algorithm technique still at the very early stage of research is based on goal reasoning. This approach focuses the algorithm on the expected goals as information contained in the input data is transformed into desired (goal) knowledge [11]. AI agents would need to deliberate and self-adjust their ultimate goals, particularly in complex environments.

This specific set of recommendations offers many benefits as outlined in Figure 6.6. For example, by leveraging both unsupervised and supervised learning, we expect to lower the algorithm dependency on large volumes of labeled data. Algorithm collaborations can be beneficial by incorporating social-cultural networks relevant to transforming information into knowledge. Although most of the discussion in this section is focused on leveraging AI for

defense applications, it is natural to expect that we might also have to employ AI for offense applications.

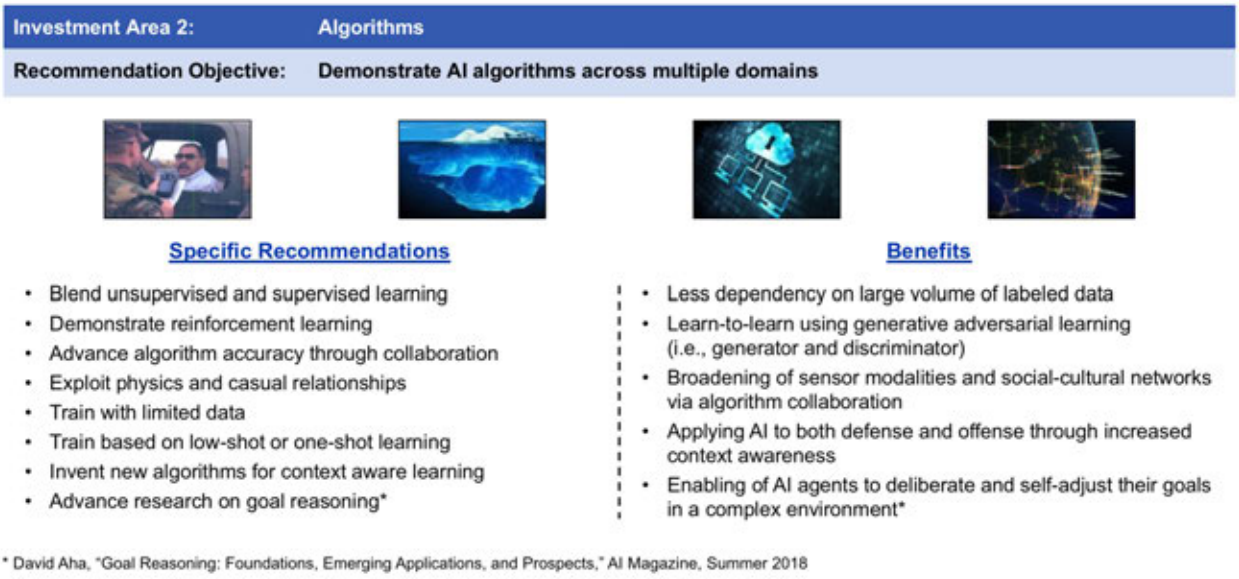


Figure 6.6. Recommendations on AI algorithms.

As discussed in Section 3, modern computing permeates across all elements of the AI canonical architecture, from data conditioning through human-machine teaming. Ultimately the objective is to achieve near-real-time or real-time performance. There are several important investment recommendations outlined in Figure 6.7 that are still not developed enough to make AI systems useful in DoD, IC, and Homeland Security applications. Some of the needed investments are in the deployment of computing to the edge (for example, available to users forwardly deployed requiring AI systems to operate in near-real time or real time). Similarly, these users must operate across distributed platforms operating under stringent low SWaP. The advent of IoT will be more ubiquitous than what we see today.

As we introduce more context into AI systems, modern computing will need to evolve into cognitive computing vs. content extraction computing as done today. This advancement also requires that the computing be able to adapt from deterministic operations to more probabilistic operations, where computation precision is not as critical compared to timely outputs. As described by David Patterson and John Hennessy during their Turing award lecture [12], modern computing will need to be based on domain-specific architectures with the ability to include variable precision. Variable precision results in lower power consumption when high precision is not needed. For example, the classifier stage (inference) in a machine-learning algorithm does not need high levels of precision—integer arithmetic suffices. At the recent Supercomputing conference [13], Jesen Huang (Nvidia's CEO) described advancements in GPUs with rapid adoption into datacenters. Nvidia's latest T4 GPU is based on a variable precision capable of 260 TOps (trillion operations per second) at integer precision of 4 bits. It is also capable of delivering 8.1 TFlops at a floating-point precision of 32 bits. Server companies, such as Dell EMC, IBM, Lenovo, and others, are all featuring the Nvidia T4 GPU.

Ultimately, computing systems must scale to meet the objectives of the mission. For example, in adversarial and tactical environments, embedded computing must be able to be

resilient to unforeseen cyber attacks. AI systems must be trusted under these very complex environments.

As most AI practitioners know, most of the time is spent on the creation of the models—and it is at this stage where floating-point precision is needed. Therefore, it is important to invest in computational advances to reduce the time to train AI models. Also, the development of context-based computing will provide improvements in the ability to sense, reason, and respond to stimulus closer to models of the human brain. The development of DSA (design-specific architectures) will require closer coupling between hardware and software designers—an important byproduct of the DSA approach is in providing computing improvements in spite of Moore's law continuing to slow down. Finally, since most of the AI developments today exists in cloud-based environments (for example TensorFlow, PyTorch, etc.), these tools must be adapted to also be relevant to tactical and embedded systems.

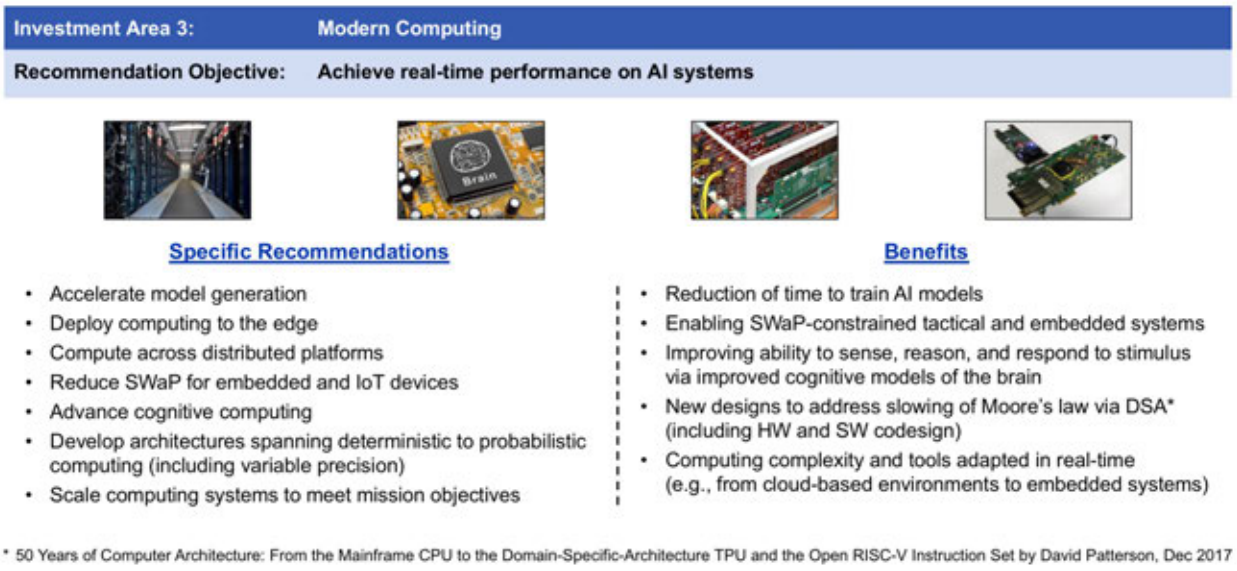


Figure 6.7. Recommendations on modern computing.

The human-machine teaming (HMT) subcomponent is one of the most important elements of the end-to-end AI system. Unfortunately, this capability is one of the least attended in today's research and development investments. HMT will enable augmentation of human capabilities by leveraging intelligent machines. Of course, as discussed in earlier sections and illustrated in the AI canonical architecture, the HMT fits across a broad spectrum. In some instances, the human would have a stronger role. In other instances, the intelligent machines will be more prominent in the overall architecture.

As discussed by A. Ilachinski [14], regarding AI, robots, and swarms, there are several levels of autonomy in the context of HMT frameworks. A good way to categorize this HMT is based on the following three categories:

- Human-in-the-loop: this is also categorized as semi-autonomous operations where the intelligent machine acts upon direction from the user.
- Human-on-the-loop: this is also categorized as human-supervised actions, where the intelligent machines are designed to allow for the human operator to intervene and terminate any undesirable actions.

- Human-out-of-the-loop: this is also categorized as instances where the intelligent machine is allowed to operate independently of the user. As we discussed in earlier sections under robust AI, these instances will be very unique where the human must have complete trust in the decisions made by the intelligent machine. Most cases today, and in the foreseeable future, will fall in the category of human-in-the-loop or human-on-the-loop for most national security applications. Only in those cases where the confidence is high that the machine is making the correct decisions and the consequence of actions is low will the intelligent machine will be allowed to operate completely autonomously (human-out-of-the-loop).

Since HMT still needs significant and consistent investment, relative to its importance in a robust and trustful end-to-end AI system, there are several important recommendations outlined in Figure 6.8. For example, HMT involves inter-relationships and connectivity among multiple humans and machines in a teaming environment. These relationship graphs must be updated in real time. Also, there has to be a significant level of human and machine transparency. An important aspect of this interaction is the ability to communicate in a very natural way, as we commonly communicate among humans. Thus, there is a need to make advances in natural language processing to improve HMT. Ultimately the human-machine teams must collaborate to achieve mission goals—recall that intelligent machines must augment the capabilities of the humans for them to be useful in critical missions. This investment recommendation is closely coupled with the need to have the ability to understand and shape the human-machine networks—and sentiment analysis is an important element in augmenting human capabilities. Long-term, the goal is to scale to very large human–machine teams, as we commonly find in human–human teams.

The benefits from these investments are numerous. Figure 6.8 outlines a few. For example, the recommended investments will lead to strengthening coupling between HMTs. Since HMT will lead to effectively transforming knowledge resulting from the algorithms to insight, these recommended investments will increase the value of the resulting insight. Insight from an end-to-end AI system is what users can use to make informed CoA decisions.

Most of the important DoD, IC, and Homeland Security applications will require adapting to environments that evolve rapidly. Thus, the recommended investments will lead to the ability for HMT to adapt to rapidly changing complex environments. A close corollary to these changing environments is the ability to leverage social networks via HMT collaborations. A longer-term benefit is to have the ability for an AI system to reason based on context. Again, most AI systems are able to operate on content present in the input data. AI systems are not able to easily integrate context, as humans commonly do daily. Another long-term goal is to be able to reconfigure HMTs to meet scale of missions.

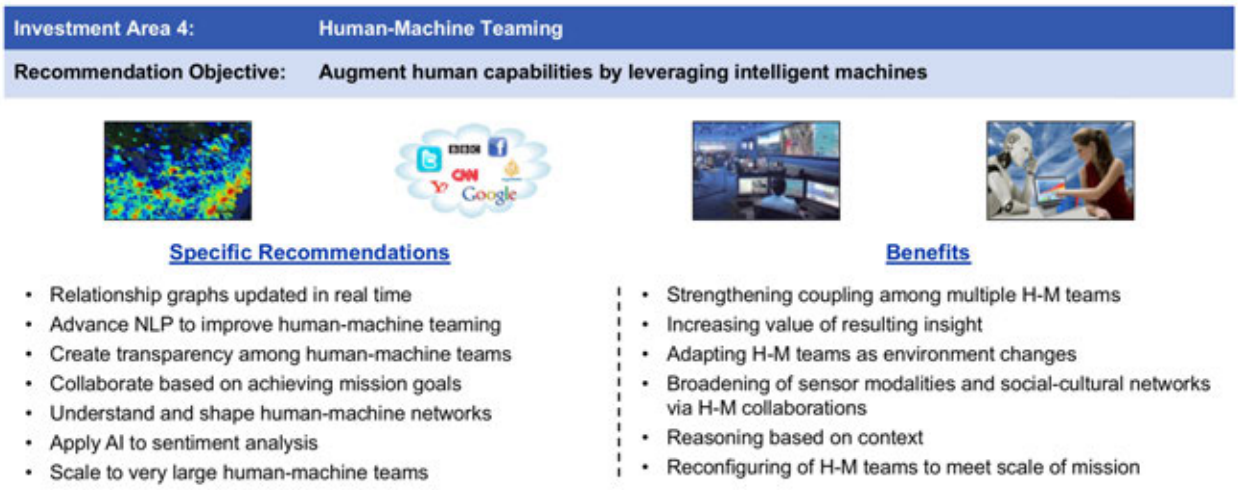


Figure 6.8. Recommendations on human-machine teaming.

The next investment area is robust AI. This is very critical because there are many techniques that can be used in a malicious ways by an attacker [15]. As discussed earlier in Section 3, machine learning techniques are very easily fooled to result in the incorrect classification. There are several techniques that an adversary can employ. These are known in the literature as white-box attacks and black-box attacks. White-box attacks assume the adversary has full knowledge of the machine learning algorithm subcomponent. A black-box attack represents the case where an attacker can infer the internal working of the algorithm by tapping into the data input and data outputs. This latter form of attack is much more likely in real-world scenarios. Therefore, it is very important to invest in robust AI metrics, which will lead to rigorous AI assessments. A field well established in computer science known as formal methods could be employed in verifying and validating an AI system. However, much research is needed to implement formal methods techniques.

There is a need to understand the vulnerabilities of an end-to-end AI system and adversarial learning approaches to avoid any element of surprise—red-teaming is highly recommended to be a key part of any AI development. A close corollary is adversarial AI counter-measure. This investment plays an important role in strengthening both the physical and cyber security of an AI system.

Ultimately, the goal of robust AI investments is for the users to gain confidence in the results produce by the AI system. As discussed earlier, gaining user confidence can be attained through grand challenges and/or AI hackathons (leveraging, for example, Kaggle competitions and serious games approaches). Users are responsible for making decisions based on a set of CoAs. Therefore, insights derived from an AI system must be robust.

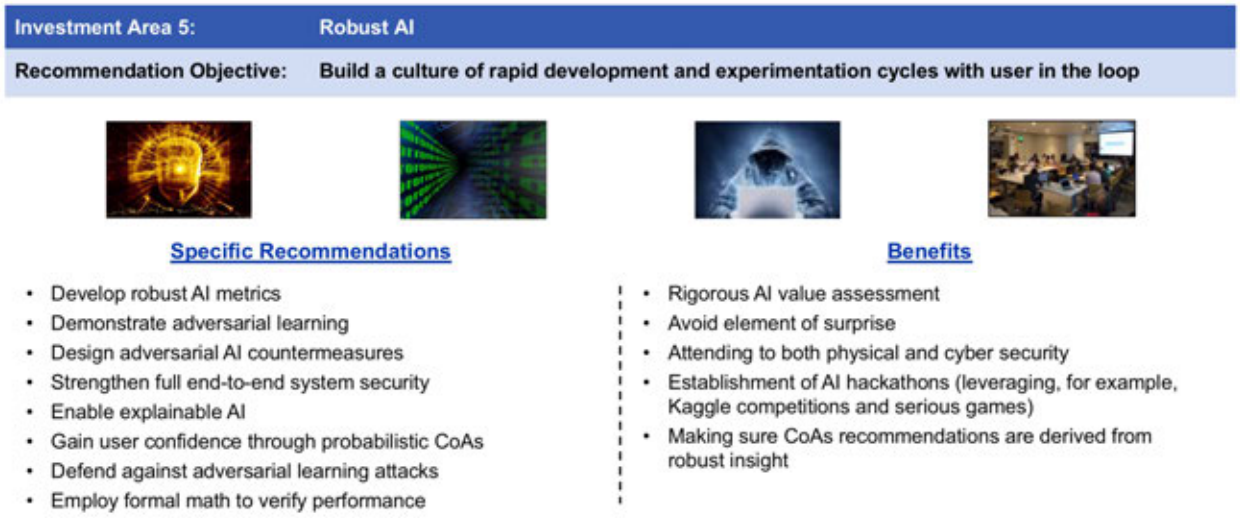


Figure 6.9. Recommendations on robust AI.

6.3 Transitioning AI Capabilities to Users

We should continue to build an ecosystem where researchers from academia and laboratories work elbow-to-elbow with government users to make rapid advances in AI for national security applications. The field evolves so rapidly that previous approaches based on sequential developments are too long to stay relevant. We should also learn from great successes made at the national levels. As stated by giants such as MIT professor Vannevar Bush:

> *If America wants to put a man on the moon, which is really a tough engineering job, they just gather enough thousands of scientists, pour in the money, and the man will get there. He may even get back.—Vannervar Bush [16]*

Future AI developments, in our nation, will need a very strategic vision and approach. The United States decisively won the Cold War against the Soviet Union by strategically investing in systems and technologies that were game changing, and, therefore, allowed deterrence from war conflicts. AI falls in the category of game changing, but the challenge is how to be strategic about its implementation at a fast and a diligent pace. A congressional commission, established by the House Armed Services Committee, has been created and named the National Security Commission on Artificial Intelligence [17]. This commission will evaluate the usefulness of AI and related technologies in national security efforts, potential future applications, global use trends, data standards, ethical questions, and workplace and education incentives.

Although this comprehensive report addresses capabilities needed to advance systems leveraging AI technologies, technology developments are necessary but not sufficient to successfully transition them to operational users. Figure 6.10 shows two other critical elements for a successful transition: people and process. Organizations investing in AI must build a culture of rapid development cycles with user in the loop. There is also a need to maintain a talent pool, as discussed earlier and in Figure 6.4, with a multi-disciplinary background.

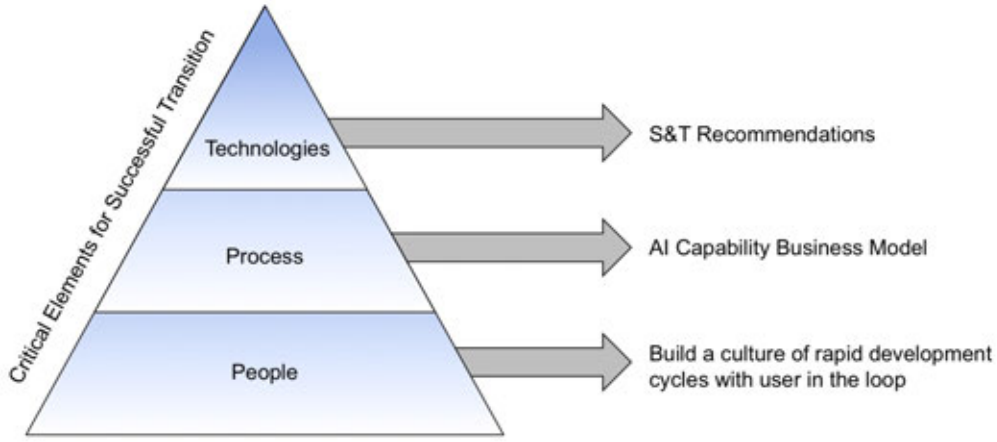


Figure 6.10. Critical elements for successful transition of AI capabilities to users.

In the context of process, as part of this AI study, we formulated an AI capability business model. This capability model, shown in Figure 6.11, starts by requiring a clear understanding of the AI capabilities desired. There are modern techniques to perform this phase of the model. One approach is referred as *design thinking* [18, 19]. Design thinking integrates an end-user focus with multi-disciplinary collaborations, including iterative improvements. It is a very powerful tool for achieving desirable, user-friendly, and economically viable design solutions and innovative capabilities.

Since many of the AI advances are being developed across several scientific and research organizations, another important step is to perform a survey on the lay-of-the-land relative to the AI capabilities desired. The identified capabilities are then rapidly prototyped within an end-to-end AI system. As it is commonly done with complex systems, that are depending on a high degree of robustness, a rigorous system analysis is also required together with a lens towards rapid experimentation and a verification and validation phase.

Most successful prototype developments rely on an environment consisting of a gold standard. A gold standard would consist, as a minimum, of a test harness, datasets, performance metrics, and benchmarks. The output of the AI prototype will then be transition into an operational environment. In some cases, an 80% solution that still meets most of the AI capabilities desired is better than a postulated 100% solution that it is too late to be useful to the ultimate users.

Users would need to be an integral part of this AI capability business model. They would provide feedback in the form of operational gaps, together with an assessment on how well the prototype performed in an operational setting via performance metrics. This recommended approach is very consistent with processes well established in, for example, Silicon Valley. Most products developed in fast moving industries, like advanced electronics, require that systems be put in the hands of users to get feedback and for the developers to iterate on the product. AI systems should not be any different except for the need to make them robust, as discussed in the previous section.

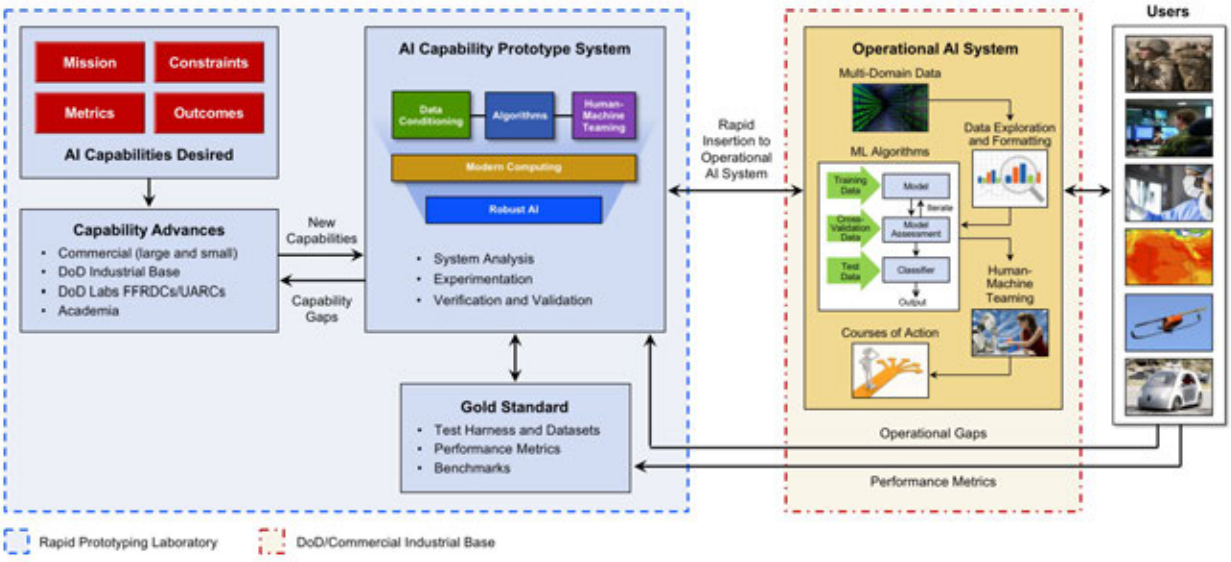


Figure 6.11. AI capability business model. Rapid feedback between research and users.

Consistent with the need to simultaneously address people, process, and technologies, and the AI capability business model discussed earlier, organizations must also formulate a strategic blueprint or roadmap. In Figure 6.12, we depict a strategic planning model adapted from D. Nadler and Mike Tushman [20]. The main components of this strategic model is first to formulate candidate strategic directions based on inputs from an envisioned future, long-term needs, and organization core values. The key operating "engine" of this model is shown in the center of Figure 6.12, consisting of identified AI implementation opportunities (based on customers' inputs), people available to execute the implementation, the AI infrastructure in the organization, and finally a culture of rapid innovation and prototyping. This strategic model results in a set of deliverables that form the strategic roadmap, including goals, actions, strength, weaknesses, opportunities, and threats (SWOT) analysis, regrets, ascertaining that capabilities meet AI needs, and end-to-end AI system assessment via performance metrics. Of course, no strategic roadmap is static. Users must weigh in by identifying critical gaps missing from the strategic roadmap. This roadmap is always a journey, not a destination, so it needs to be updated as one learns from implementing AI prototypes shown in Figure 6.11.

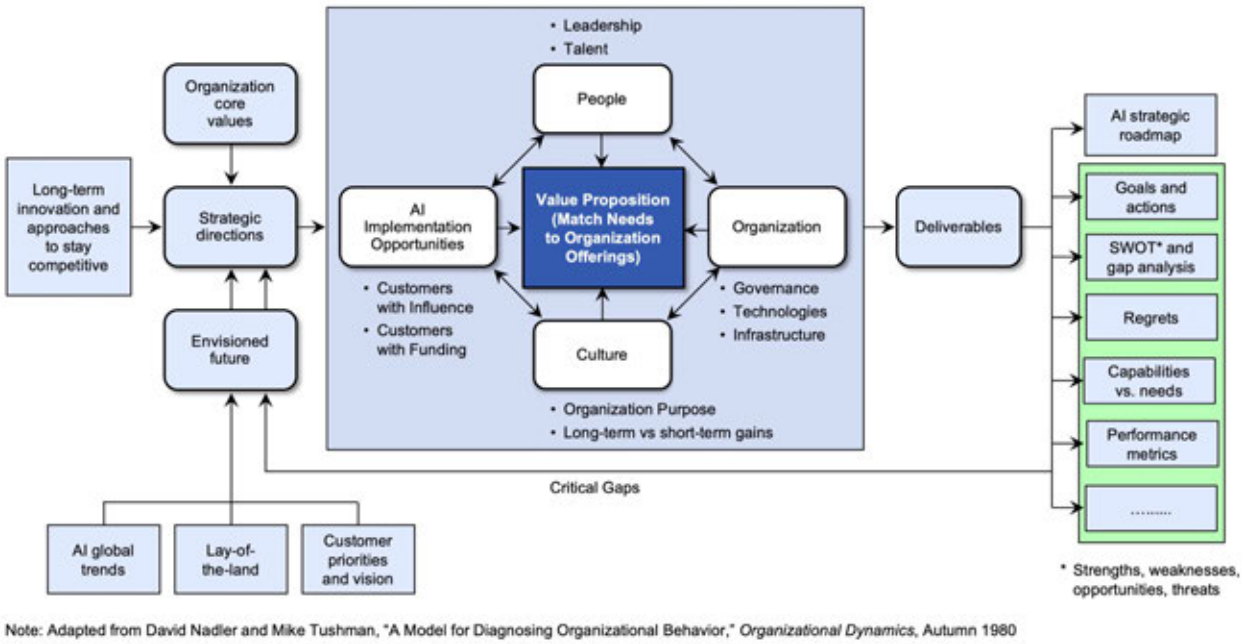


Figure 6.12. Strategic planning model for effective AI implementation and transition.

In this report we documented the AI study that we undertook:

- To present an AI short history
- To elaborate on present developments
- To discuss enabling technologies, and applications of AI to HLT and cyber security
- To address a future outlook consisting of S&T investment recommendations

We also discussed, briefly, the need to have developers work closely with the users, build an environment of innovation and rapid prototyping, with timely user feedback.

We hope this report serves as a blueprint for AI developments across three horizons:

- Horizon 1: capabilities available to the user in 1–2 years
- Horizon 2: capabilities available to the user in 3–4 years
- Horizon 3: capabilities available to the user in 5+ years

Each horizon above must start its respective investment now to have the desired capabilities consistent with the rapid evolution of AI technologies avoiding obsolescence.

This study and written report should be revisited at least every two years or less to document and assess AI advancements in the intervening years.

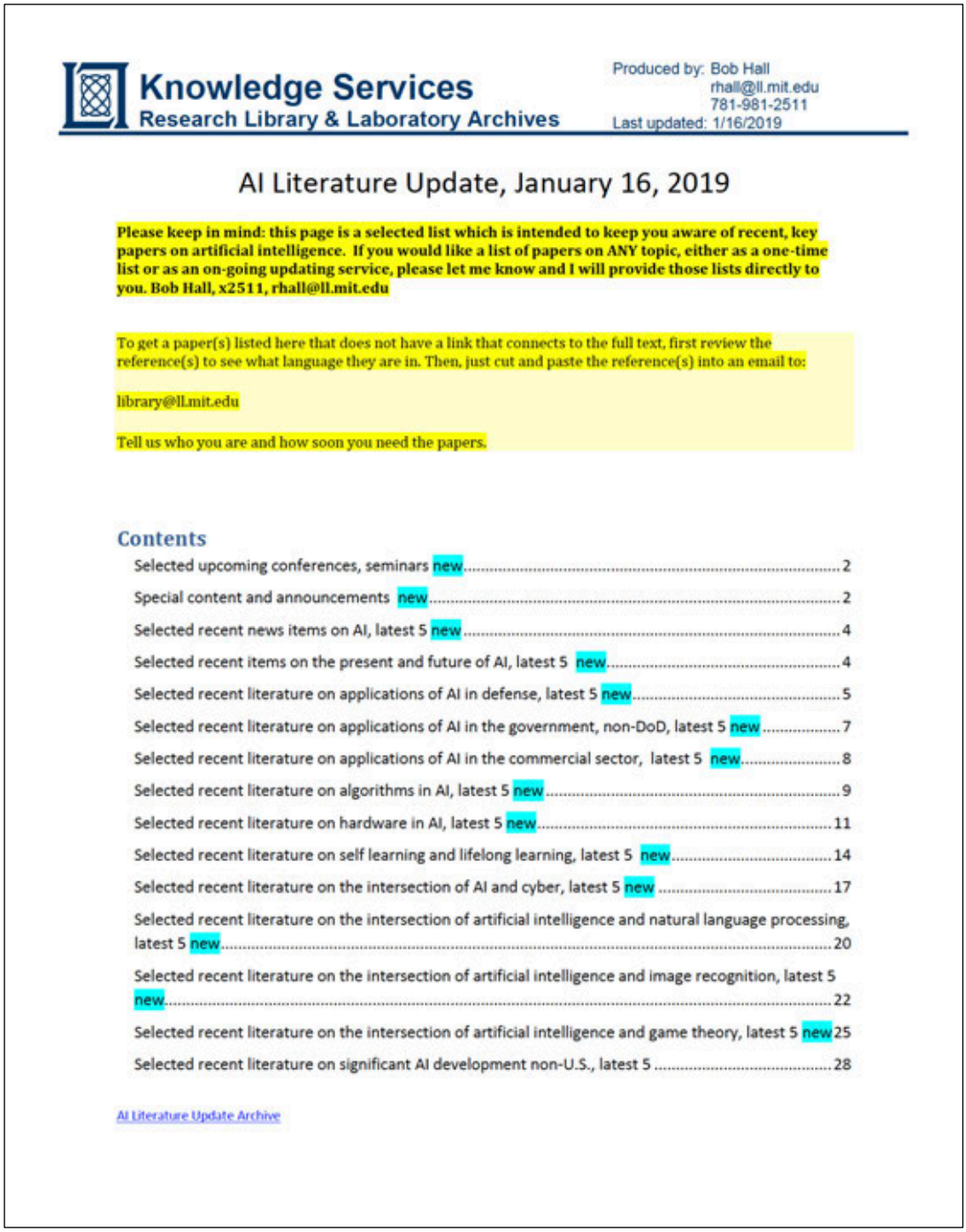
"Things don't have to change the world to be important."— Steve Jobs

References

1. National Research Council, *Frontiers in massive data analysis*. 2013: National Academies Press.
2. Quinn, A.J. and B.B. Bederson. *Human computation: a survey and taxonomy of a growing field.* in *Proceedings of the SIGCHI conference on human factors in computing systems*. 2011. ACM.
3. MIT Lincoln Laboratory, *Strike Group Defender.* 2017. https://www.ll.mit.edu/r-d/projects/strike-group-defender
4. Laamarti, F., M. Eid, and A.E. Saddik, *An overview of serious games.* International Journal of Computer Games Technology, 2014. **2014**: p. 11.

5. Michael, D.R. and S.L. Chen, *Serious games: Games that educate, train, and inform*. 2005: Muska & Lipman/Premier-Trade.
6. Seetharaman, G., A. Lakhotia, and E.P. Blasch, *Unmanned vehicles come of age: The DARPA grand challenge.* Computer, 2006. **39**(12).
7. Buehler, M., K. Iagnemma, and S. Singh, *The DARPA urban challenge: autonomous vehicles in city traffic*. Vol. 56. 2009: springer.
8. Hariharan, B. and R.B. Girshick. *Low-Shot Visual Recognition by Shrinking and Hallucinating Features*. in *ICCV*. 2017.
9. Santoro, A., et al., *One-shot learning with memory-augmented neural networks.* arXiv preprint arXiv:1605.06065, 2016.
10. Pearl, J. and D. Mackenzie, *The Book of Why: The New Science of Cause and Effect*. 2018: Basic Books.
11. Aha, D.W., *Goal Reasoning: Foundations, Emerging Applications, and Prospects.* AI Magazine, 2018. **39**(2).
12. John Hannessy and David Patterson, *Turing Lecture at International Symposium on Computer Architecture (ISCA) 2018.* 2018. https://www.acm.org/hennessy-patterson-turing-lecture
13. Jensen Huang, *Superconputing Conference 2018 Speech on New High Performance Computing.* 2018. https://www.youtube.com/watch?v=c29B-sfzJvY
14. Andrew Ilachinski, *AI, Robots, and Swarms: Issues, Questions, and Recommended Studies.* 2017.
15. Brundage, M., et al., *The malicious use of artificial intelligence: Forecasting, prevention, and mitigation.* arXiv preprint arXiv:1802.07228, 2018. https://arxiv.org/pdf/1802.07228
16. Zachary, G.P., *Endless frontier: Vannevar Bush, engineer of the American century*. 2018: Simon and Schuster.
17. House Armed Services Committee, *National Security Commission on Artificial Intelligence.* 2018. https://fcw.com/blogs/fcw-insider/2018/11/nov15quickhits.aspx?s=fcwdaily_151118
18. Plattner, H., C. Meinel, and U. Weinberg, *Design thinking*. 2009: Springer.
19. Razzouk, R. and V. Shute, *What is design thinking and why is it important?* Review of Educational Research, 2012. **82**(3): p. 330-348.
20. Nadler, D.A. and M.L. Tushman, *A model for diagnosing organizational behavior.* Organizational Dynamics, 1980. **9**(2): p. 35-51.

7 Appendix A: AI Literature Update



Knowledge Services
Research Library & Laboratory Archives

Produced by: Bob Hall
rhall@ll.mit.edu
781-981-2511
Last updated: 1/16/2019

AI Literature Update, January 16, 2019

Please keep in mind: this page is a selected list which is intended to keep you aware of recent, key papers on artificial intelligence. If you would like a list of papers on ANY topic, either as a one-time list or as an on-going updating service, please let me know and I will provide those lists directly to you. Bob Hall, x2511, rhall@ll.mit.edu

To get a paper(s) listed here that does not have a link that connects to the full text, first review the reference(s) to see what language they are in. Then, just cut and paste the reference(s) into an email to:

library@ll.mit.edu

Tell us who you are and how soon you need the papers.

Contents

AI Literature Update Archive

8 Appendix B: Additional Readings, Conferences, and Other Venues

8.1 Additional Readings

1. Alden, Edward et al., *Technology and National Security: Maintaining America's Edge*, The Aspen Institute, January 2019.
2. Council, N.R., *Complex operational decision making in networked systems of humans and machines: A multidisciplinary approach*. 2014: National Academies Press. https://www.nap.edu/catalog/18844/complex-operational-decision-making-in-networked-systems-of-humans-and-machines
3. DarkTrace, *The Next Paradigm Shift: AI-Driven Cyber-Attacks* 2018. https://www.darktrace.com/en/resources/wp-ai-driven-cyber-attacks.pdf
4. Deep Knowledge Analytics. *Artificial Intelligence Industry in the UK 2018*. 2018 [cited 2018; Report posted on website]. Available from: https://www.dka.global/ai-in-uk-report.
5. Evtimov, I., et al., *Robust physical-world attacks on deep learning models.* arXiv preprint arXiv:1707.08945, 2017. **1.**
6. Fuller, S.H. and L.I. Millett, *Computing performance: Game over or next level?* Computer, 2011. **44**(1): p. 31-38.
7. Gautam, S., *The Intelligent Web. Search, smart algorithms and big data*. 2013, Oxford University Press.
8. Gonzalez, R., Deep Convolutional Neural Networks [Lecture Notes]; MSP Nov. 2018 79-87.
9. Goodfellow, I., P. McDaniel, and N. Papernot, *Making machine learning robust against adversarial inputs.* Communications of the ACM, 2018. **61**(7): p. 56-66.
10. Michael Horowitz, *The promise and peril of military applications of artificial intelligence.* 2018. https://thebulletin.org/landing_article/the-promise-and-peril-of-military-applications-of-artificial-intelligence/
11. Jouppi, N.P., et al., *A domain-specific architecture for deep neural networks.* Communications of the ACM, 2018. **61**(9): p. 50-59.
12. Jeff Loucks Tom Devenport David Schatsky, *Deloitte State of AI in the Enterprise, 2nd Edition.* 2018. https://www2.deloitte.com/insights/us/en/focus/cognitive-technologies/state-of-ai-and-intelligent-automation-in-business-survey.html
13. Tom Markiewicz and Josh Zheng, *Getting Started with Artificial Intelligence.* O'Reilly Media Inc. and IBM Corporation, 2018.
14. National Research Council, *Frontiers in massive data analysis*. 2013: National Academies Press.
15. National Science & Technology Council, *Frontiers of Visualization II: Data Wrangling Workshop Summary.* 2018. https://catalog.data.gov/dataset/frontiers-of-data-visualization-workshop-ii-data-wrangling-workshop-summary
16. Potember, R., *Perspectives on research in artificial intelligence and artificial general intelligence relevant to DoD*. 2017, The MITRE Corporation McLean United States.
17. Simon, P., *Too big to ignore: The business case for big data*. Vol. 72. 2013: John Wiley & Sons.
18. Stoney Trent and Scott Lathrop, *A Primer on Artificial Intelligence for Military Leaders.* Small Wars, 2018. http://smallwarsjournal.com/jrnl/art/primer-artificial-intelligence-military-leaders
19. Sydney Freedberg, Joint Artificial Intelligence Center Created Under DoD CIO. 2018. https://breakingdefense.com/2018/06/joint-artificial-intelligence-center-created-under-dod-cio/

8.2 Conferences and Other Venues

- NeurIPS: Neural Information Processing Systems (formerly abbreviated NIPS). Held in early December https://nips.cc/
- ICML: International Conference on Machine Learning. Held in July https://icml.cc/
- ICLR: International Conference on Learning Representations. Held in May https://iclr.cc/
- AAAI: Association for the Advancement of Artificial Intelligence. Held in February http://www.aaai.org/
- CVPR: Computer Vision and Pattern Recognition. Held in June https://www.thecvf.com/
- ICCV: International Conference on Computer Vision. Held in the Fall of odd years https://www.thecvf.com/
- KDD: Knowledge Discovery and Data Mining https://www.kdd.org/kdd2019
- IJCAI: International Joint Conference on Artificial Intelligence http://ijcai19.org/
- MIT Artificial Intelligence Course https://ocw.mit.edu/courses/electrical-engineering-and-computer-science/6-034-artificial-intelligence-fall-2010/lecture-videos/
- Google Advancing AI for Everyone https://ai.google/
- Facebook's F8 and Apple's WWDC
- Microsoft https://academy.microsoft.com/en-us/professional-program/tracks/artificial-intelligence/
- Coursera Machine Learning https://www.coursera.org/learn/machine-learning
- NVIDIA's GPU Technology Conference https://www.nvidia.com/en-us/gtc/
- AI Conference (Presented by O'Reilly & INTEL AI) https://www.quora.com/What-are-the-top-AI-conferences
- TTI Vanguard (By Membership Only) https://www.ttivanguard.com/

https://www.quora.com/What-are-the-top-AI-conferences
http://www.guide2research.com/topconf/machine-learning

9 Appendix C: Glossary

ASIC	application specific integrated circuit
AGI	artificial general intelligence
CPU	central processing unit
COTS	commercial-off-the-shelf
Cyber-EW	cyber electronic warfare
CyMIA	Cyber Machine Intelligent Assistant
DNNs	deep neural networks
DARPA	Defense Advanced Research Projects Agency
DENDRAL	dendric algorithm
DoS	denial-of-service
DoD	Department of Defense
DSA	design specific architecture
FFRDC	federally funded research and development center
GANs	generative adversarial networks
GISR	global intelligence, surveillance, and reconnaissance
GMM	gaussian mixture model
GPS	global positioning system
GPU	graphics processing unit
HPC	high performance computing
HLT	human language technology
HMT	human-machine teaming
IC	intelligence community
IoT	Internet of things
MLE	maximum likelihood estimation
NLP	natural language processing
NMT	neural machine translation
ONNX	Open Neural Network Exchange
S&T	science and technology
SE	squared error
SICADA	SIde Channel Authenticity Discriminant Analysis
SNR	signal-to-noise
SouNDeR	Sociocultural Network Attack Discovery and Response
SVM	support vector machine
SWaP	size, weight, and power
SWOT	strength, weaknesses, opportunities, and threats
TFIDF	time frequency–inverse document frequency
Tops	trillion operations per second
TPU	tensor processing unit
UBM	universal background model