SYSTEMS ENGINEERING RESEARCH CENTER

SENSEMAKING RESEARCH ROADMAP

PREPARED FOR THE OFFICE OF THE DIRECTOR OF NATIONAL INTELLIGENCE

K.P. (SUBA) SUBBALAKSHMI, STEVENS INSTITUTE OF TECHNOLOGY
ARAM GALSTYAN, UNIVERSITY OF SOUTHERN CALIFORNIA
RAMA CHELLAPPA, UNIVERSITY OF MARYLAND, COLLEGE PARK
CHARLES CLANCY, VIRGINIA TECH

STEVENS INSTITUTE of TECHNOLOGY THE INNOVATION UNIVERSITY®

UNIVERSITY OF MARYLAND 1856

USC Viterbi School of Engineering Information Sciences Institute

VIRGINIA TECH

RESEARCH CENTER

The Systems Engineering Research Center (SERC) is a federally funded University-Affiliated Research Center managed by Stevens Institute of Technology.

This material is based upon work supported, in whole or in part, by the U.S. Department of Defense through the Office of the Assistant Secretary of Defense for Research and Engineering (ASD(R&E)) under Contract HQ0034-13-D-004.

EXECUTIVE SUMMARY

On 2-3 May 2018, the Systems Engineering Research Center (SERC) convened a workshop to examine current research trends, challenges, and open science questions in artificial intelligence (AI)-enabled sensemaking technologies. The Office of the Director for National Intelligence (ODNI), with support from the Office of the Deputy Assistant Secretary of Defense for Systems Engineering (DASD(SE)) funded this effort. Nearly 50 experts from both academia and the intelligence community (IC) gathered together to contribute. The finding of these experts formed the starting point of this research roadmap.

The consensus from the workshop was that the two most pertinent research thrusts that are needed for developing a higher form of sensemaking are (1) multi-modal analysis for sensemaking and (2) hybrid systems for sensemaking. Within these thrust areas exist several research tracks worthy of pursuit. We discuss these in Sections 3 and 4, respectively.

Currently, four elements influence multi-modal sensemaking systems: (1) the volume of information generated by both conventional and social media sensors; (2) the intentional corruption of information in these modalities; (3) the different speeds at which information propagates over the internet, (4) the often contradictory information about the same topic from different sources. Although there has been significant research done to mitigate these elements, these tend to be more on intra-modal or cross-modal modalities, using a few subsets of data. Significant research is required to progress from these analytics to true multi-modal sensemaking systems.

Therefore, we propose the following research tracks: interoperability issues, reliability, and trustworthiness, next-generation AI and sensemaking algorithms, and assessing and closing the loop for multi-modal sensemaking systems. Some of this research has already begun, some will take a decade or more to develop. In Section 5, we provide a projected timeline for the research. We believe these tracks are crucial to developing the science of holistic sensemaking.

Hybrid sensemaking faces its own set of challenges. IC-related issues often rely on human analysts, creating a bottleneck that cannot be solved with additional computing power. Hence, the goal of HS^2 is to explore the possibilities for human-machine hybrid systems. We seek to eventually develop a networked ecosystem of human analysts and machines. The HS^2 research tracks are: HS^2 taxonomy & performance measures, interactive & continuous sensemaking, HS^2 autonomy & trust, HS^2 as networks: organizational sciences perspective, and HS^2 interfaces. Although we have made some strides along the HS^2 research tracks

(e.g., developing a taxonomy, domain-specific HS^2, and HS^2 modeling) much of the hybrid sensemaking work is still to come. As you will see in Section 5, we project much of the HS^2 research to be 5-10 years out or more.

TABLE OF CONTENTS

List of Figures

1. SETTING THE STAGE

The Systems Engineering Research Center (SERC) convened a workshop funded by the Office of the Director for National Intelligence (ODNI), with support from the Office of the Deputy Assistant Secretary of Defense for Systems Engineering (DASD(SE)), on 2-3 May 2018 to examine expertise in sensemaking technologies. As of now, analysts and collectors within the U.S. Intelligence Community (IC)—whose job is to inform and warn decision-makers of potential threats or developing situations—continue to struggle to adapt to the persistent growth in available data. Moreover, practical considerations, such as personnel ceilings, govern the depth and breadth of topics that the IC can cover at any given time.

The main objective of this effort was to provide the sponsor with insights from world-class experts and technologists familiar with emerging research opportunities associated with the application of sensemaking technologies to IC tradecraft and to kick-start the process of creating a research roadmap in this area.

Participating experts (Appendix A: Workshop Participants) were collectively experienced in all aspects of the computational understanding and modeling of sensemaking technologies in complex, realistic contexts, to include business, economics, policy, or science and technology. The workshop spanned two days with approximately 50 people in attendance on the first day. Of the 50 people, 30 faculty members from 20 different universities were present as were members of the IC. The second day included only the core planning group and members of ODNI to debrief and discuss next steps.

The consensus from the ODNI sensemaking workshop held in early May 2018 was that the two major research thrusts that will help towards the quest for a higher form of sensemaking are "multi-modal analysis for sensemaking" and "hybrid systems for sensemaking". These will be discussed in Sections 3 and 4, respectively.

The rest of this research roadmap defines the research thrust areas and identifies the main research tracks within these two broad thrust areas where research is needed and can have a significant impact.

2. BACKGROUND

Intelligence generation in the context of the IC is a complex process. Data is collected by a range of different sensors, undergoes processing to extract information, and is then stored in a database. Domain-specific analysts query those databases and produce intelligence reports. All-source analysts aggregate those reports into finished intelligence that is delivered to the warfighter and policymaker. In the reverse direction, warfighters and policymakers express intelligence needs, which they then translate into requirements and tasking for the front-end sensors. Figure 1 shows this simplified view of intelligence generation (originally from [1]).

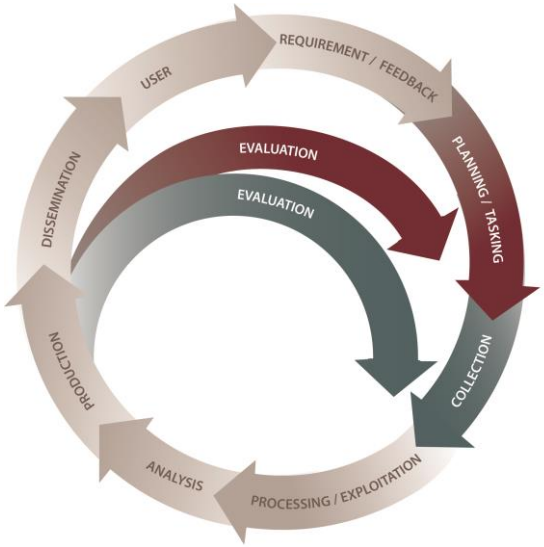

Figure 1: Intelligence Gathering

The structure of this system is somewhat rigid and lacking in agility. There is a significant feed-forward delay associated with the automated batch processes responsible for processing and databasing, and the human response times and work schedules responsible for analysis and reporting. The feedback loop is still heavily dependent on the analyst and his/her ability to push new tasking to front-end collectors.

There are numerous technological barriers for making the sensemaking cycle in figure 1 efficient and scalable to meet the growing IC needs. Those challenges affect relatively low-level tasks, such as information gathering and processing at scale, higher level problems of generating and tracking multiple hypotheses to aid an analyst and closing the feedback loop

where those hypotheses direct information gathering and sensor tasking processes. During the last couple of decades, the government, and specifically the Department of Defense (DoD) funding agencies have invested considerable funds to address those challenges. However, most of these efforts have targeted specific, isolated parts of the above sensemaking cycle, instead of addressing it holistically. As a result, there is still a sizeable technological gap between the current capabilities and the needs of the IC community.

In this roadmap, we have identified two broad, yet well-defined research challenges that we think need to be addressed for closing the technological gap. The first challenge is due to the evolving nature and the sheer volume of the information that intelligence analysts must handle. In particular, such information typically originates from highly heterogeneous sources and comes in different modalities (text, speech, imagery/video and social networks such as Facebook, Twitter, and chat sites). Although there have been significant recent efforts in addressing each of those modalities individually, there is an urgent need for technologies that will enable seamless multimodal sensemaking across multimodal information.

The second broad challenge is based on the realization that the current paradigm of an intelligence analyst working with custom-built computational tools for combing and analyzing large volumes of data is not scalable. Indeed, even for a single organization such as the National Geospatial-Intelligence Agency (NGA), the current projections indicate a need for eight million trained analysts by 2020 for handling the anticipated volume of the data generated by the agency. The above observation necessitates a qualitatively innovative approach to sensemaking, which encompasses a more integrated and synergistic system of human analysts and computational tools or machines to make the process more scalable, efficient, and accurate.

Based on these challenges, we propose a research roadmap that articulates our rationale and identifies specific research areas relevant to achieving this goal.

3. Multi-Modal Systems for Sensemaking

Four elements influence multi-modal sensemaking systems in today's age: (1) the volume of information generated by both conventional and social media (or more generally, the internet) sources (2) the intentional corruption of information in these modalities; (3) the different speeds at which information propagates over the internet, and finally (4) the often contradictory information about the same topic from different sources.

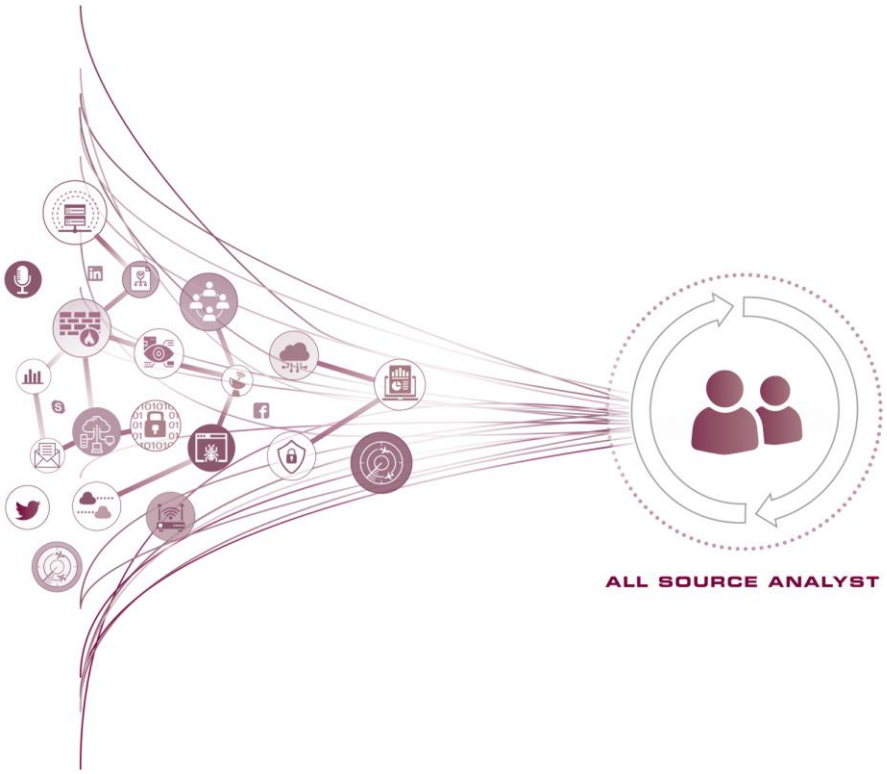

Figure 2: The all-source analyst must deal with a large volume of information from a very diverse set of sources that vary in speed of acquisition, modes of acquisition, methods of processing, and accuracy.

Over the last decade, information is being generated at a dizzying rate by conventional and social media sources (figure 2). Conventional sources of information include images, videos, language, text, while social media information sources include Facebook, YouTube, Twitter, and various online chat forums and blogs. Recent estimates indicate that 200 million tweets are being generated every day, and 250 million photographs are being uploaded to Facebook every day, and as mentioned earlier, the NGA generates such a vast amount of data that, in the estimation of the Director of the NGA, by 2020 eight million analysts will be needed to process it.

The *second* element is that some of the information being generated by these sources may be false or intentionally misleading. For example, on April 2013, a fake tweet on the Associated Press (AP)'s Twitter account about two explosions in the White House created widespread panic in the U.S. financial markets. Although the White House and AP announced that the tweet was a rumor, the Dow Jones index values dropped one percent, wiping out $136.5 billion in a matter of seconds before recovering the market's losses [2].

Multi-Modal Sensemaking

For this research roadmap, we define multi-modal sense making as follows:

Multi-Modal sensemaking is defined as a holistic analysis of multi-modal data with the aim to produce actionable intelligence. Here domains may be defined as (i) physical – air, space, surface, subsurface etc., (ii) cyber: social media, news or (iii) other types like archival data, human & machine, biometrics, human-machine etc. The information can be derived from independent information sources or from correlated applications. The data types could include: (i) text, voice, images/video; (ii) synthetic (machine generated), (iii) natural etc.

Artificial intelligence start-ups like Lyrebird[1] have made it easy to mimic the human voice. It is now possible to make a machine read the fake text in any given person's voice using technologies such as the one used by Lyrebird. Text-based deception has existed for about a decade, and significant strides have been made in detecting this kind of deception [3].

Although significant research has been done on differentiating rumors from factual information [4] and how rumors become viral on social networks [5], these studies have not been extended for multi-modal sensemaking. Even within the single domain (i.e., the internet), there are many challenges in this topic, including localizing the sources of misinformation [6]. One of the problems here is the speed with which information propagates over the internet, causing a viral spread of misinformation that can cause contamination of information at other sources.

Social media, online chats, and interactive fora can be a major supply of crowd-sourced information. For example, during the Stanley Cup riots in Vancouver, Canada, young rioters bragging on social media led to numerous viable leads. However, often, several online

[1] https://lyrebird.ai/

sources can have conflicting information about a given topic. To illustrate the confusion, we collected tweets from Twitter users in different regions on full body scans in airports. We counted the number of tweets that were (a) supportive of full-body scans, (b) opposed to full body scans, and (c) neutral to full body scans using sentiment analysis. A spatial snapshot of the data shows that if a user sends a tweet to a friend who lives in the North American region to obtain information on full body scan, s/he is likely to obtain information that is supportive of full body scan [7]. However, another user who lives in the Asian region may provide negative information on full body scan. These differing tweets can confuse the consumer who sought information. Similarly, when a sensemaking system requests information from multiple sources but receives information that can either be contradictory or partially contradictory, this can add to the confusion/noise causing a deterioration of the quality of the sensemaking system's output.

Recently, the Defense Advanced Research Projects Agency (DARPA), has instituted several programs such as Deep Exploration and Filtering of Text Low Resource Languages for Emergent Incidents (LORELEI), and Active Interpretation of Disparate Alternatives (AIDA). The broad goals of these efforts are to design media-specific transforms that look at input information (i.e., unstructured) in certain modalities (e.g., video, image, language) and produce a set of structured objects as output. The structured outputs are in the form of knowledge objects, as opposed to language objects. Specifically, the Deep Exploration and Filtering of Text program focused on creating an automated capability to transform large volumes of unstructured text into structured data for use by human analysts and downstream analytics. The Low Resource Languages for Emergent Incidents Program, which is currently underway, uses the same model of converting the unstructured input into structured, actionable output. However, this program specifically considers how to respond quickly to an emergent situation (e.g., humanitarian assistance and disaster relief missions) in a low-resource language area. The newly initiated AIDA program is looking at the text, language, images, and videos. The goal is to extract from each modality events and relations, link them to each other and other knowledge sources, and then try to put together multiple hypotheses to explain as much of the situation as possible. Errors in processing, reporting, and misinterpretation are difficult to avoid. A key distinction with AIDA is that it can promote decision making based on multiple interpretations, whereas many decisions made currently rely on a single analysis that might have ignored important evidence. More details on these programs can be found on the DARPA website[2]. Allowing an end user to consider multiple possibilities and prepare for contingencies could be particularly useful.

[2] https://www.darpa.mil/

While significant progress is being made in learning from multi-modal data in the above-referenced programs and other related programs, many challenges remain. Given increasing concerns about reliability of sources that generate information, it is important to develop measures to characterize trustworthiness of information sources. Developing these measures will be useful while integrating or fusing multi-modal data. While a well-established methodology exists for fusing information from conventional sensors [8], theoretically well-founded fusion methods for integrating data from conventional sensors and social networks are still immature. Fusion methods that can adaptively accommodate various levels of trustworthiness are needed. An obvious feature of multi-modal data is that it may be generated from different probability distributions. Domain adaptation methods [9] that can mediate over domain shifts among multi-modal data are also needed. Of particular interest is developing representations to capture the different features of multi-modal data. One may also exploit the emerging success of Generative Adversarial Networks (GANs) [10]. Classical artificial intelligence (AI) concepts such as non-monotonic reasoning, [11] common sense reasoning, [12] and explainability of decision process may also have to be explored.

The lack of data fidelity can be viewed as a multi-layer integration problem, where each layer represents a domain. AI and machine learning (ML) systems running on each layer create a layer-wise understanding of problem. These results, as well as raw information from these layers, feed into and drive overall sensemaking at all-source analyst's end and global level.

Figure 3 captures a sense of the complex process of sensemaking along with the inter-dependencies inherent in the process.

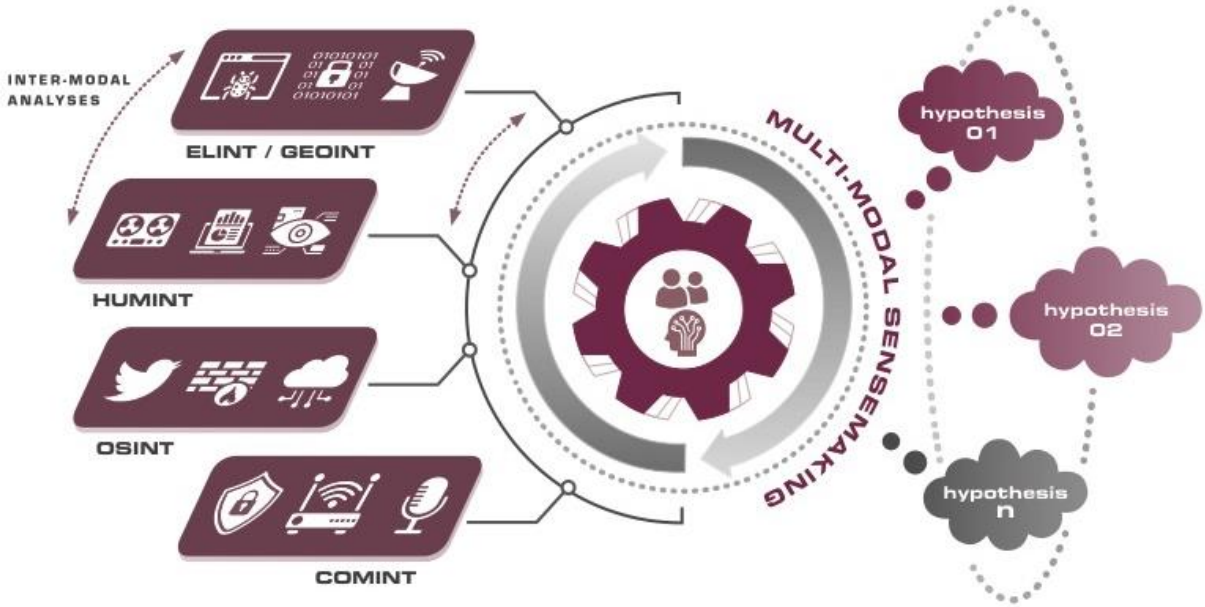

Figure 3: Multi-modal sensemaking architecture. While intra-modal/inter-modal analyses methods exist, an intelligent AI sensemaking system that uses information from all planes is the future.

3.1 KEY RESEARCH TRACKS

We have identified the following key research tracks as being crucial to driving the science of holistic sensemaking within multi-modal analysis:

- Interoperability Issues
- Reliability and Trustworthiness
- Next Generation AI and Sensemaking Algorithms
- Assessing and Closing the Loop for Multi-modal Sensemaking Systems

We provide a discussion of each in this section.

3.1.1 Interoperability Issues

One of the key factors that will determine the success of generating actionable intelligence from disparate modes is the ability of the sensemaking system to navigate the disparity between these modes successfully. The modalities of gathering information, the timescales, the reliability of the sources, and the density of available information, will all be different. ML methods for interpreting the multi-modal data must take into account many factors, such as trustworthiness, heterogeneity, both unstructured/structured data types, and varying temporal resolutions [13]. In this section, we explore some of the problems associated with interoperability issues, which often stem from the heterogeneity of the sources.

- **Fusion of conventional- and social-network information:** information-fusion and sensor-fusion methodologies have been developed for over three decades. The deluge of information in both the conventional (physical, like sensors) and cyber domains necessitates new system design to fuse data from conventional sensors such as images, video, language, and text with data collected by social networks such as Twitter, Facebook, etc. Although several inroads have been made into this burgeoning area, several challenges remain. Information from social networks is gathered from across the world, has a much different acquisition rate, and has differing levels of quality assurance than physical sensor-based information. Reliable extraction of knowledge objects from fused data is a challenging problem that may require several years of focused research. Examples of areas that researchers must address include asynchronicity of the information sources and its implication on the final hypotheses presented by the sensemaking system to the all-source analyst.

- **Fusing decisions across the multiple modes:** One key area of opportunity is to expand the use of soft decisions in processing systems. In most current systems, where sensemaking occurs with some interactions, depicted in figure 2, each processing stage advances only its most likely hypothesis, whether it be an algorithm seeking to detect a feature in a raw data stream, or an analyst concluding target activity. If an underlying philosophy of soft decisions can be implemented where multiple hypotheses are advanced at each stage, each with a likelihood and dependability scores, then context found in later stages can influence priors in earlier stages, and significantly improve the resulting decision quality. Building this into the system will also help address issues associated with noisy, incomplete, and often conflicting data.

3.1.2 Reliability and Trustworthiness

While it is well known that sensor noise and other degradations can corrupt multi-modal data collected by physical sensors due to acquisition conditions, the general view regarding data collected by conventional sensors is that the data is not intentionally corrupted. The situation is vastly different from online modes where there are several documented instances of intentional misinformation and corrupt data dissemination. Since fusion of such corrupt data with sensor data (images, videos) can be reasonably expected to affect the quality of the final intelligence generated, it is imperative that the sensemaking system is designed with these potential flaws already in mind.

Several approaches can be taken when working with the knowledge that some of the information can be intentionally corrupt. One example is to develop measures that reflect the trustworthiness of multi-modal data. Contextual reasoning must be used to evaluate the goodness of the information. One way to do this is by performing consistency checks using more reliable sensors. Preliminary efforts along this direction are being undertaken under the DARPA MediFor program, where methods exist for checking if the images and videos are doctored. We need similar methods for multi-modal data.

Below, we explore some of how reliability and trustworthiness can be increased in a multi-modal sensemaking system.

- **Detecting synthetic identities:** As mentioned earlier, deep-voice techniques [14] have made it possible to spoof anyone's voice with very high fidelity. As extensions to this problem, actors have worked to create systems of synthetic identities that involve creating fake identities with machine-generated information across multiple modes (voice, biometrics, text, etc.) Techniques are now being developed to distinguish between machine-synthesized voices and actual human voices. However, these

approaches are far from mature, and there are significant gaps in understanding these issues even within single modes. Researchers must develop systems to understand these issues over multiple modes. This research may involve several steps such as: (a) continually computing a high-dimensional data model for each mode; (b) intelligently combining these models using model stacking to improve the overall performance; and (c) updating the individual models as well as using the stacked model based on new information.

- **Analysis by synthesis using Generative Adversarial Networks (GANs):** Recently, data augmentation using GANs has proven effective for improving the performance of ML algorithms. Such data augmentation techniques can be used to mimic methods that intentionally or otherwise corrupt data so that ML algorithms can be made more robust by training them on corrupted data. It is likely that GANs can be designed to generate data that correspond to rare and unforeseen situations so that proper anomaly detection methods can be developed and evaluated. One advantage of GANs is that they can synthesize data across modalities (text to image, language to image/video, etc.); this will be useful for training multi-modal ML algorithms.

- **Information confusion**: As mentioned earlier, there are often multiple information sources with no clear a priori indication of their reliability. Often these sources have very contradicting information on any given topic [15]. In these cases, the sensemaking problem becomes even more complicated by the resulting confusion from these sources. Several fundamental questions must be answered. What is a good mathematical model for this information confusion? How does the quality of the outcome of the sensemaking system degrade due to this confusion? What are good strategies for the information providers to control the power or the intensity with which the information is transmitted? What are the fundamental bounds on the information that can be extracted when the sources are corrupt to varying degrees?

- **Characterizing the reliability of machine learning:** While performance bounds are available for linear statistical models, such bounds are not easily derived for non-linear ML algorithms based on deep networks. For humans to be able to trust that they have reliable ML algorithms to rely on, we need to develop performance bounds for test error given training error rates. Such bounds are available for many classical pattern-matching methods and support vector machines (SVMs) [16]. We need to derive similar bounds for deep networks. In [17], the authors have derived bounds on test error of a deep network as a function of network structure for cases when the training and test data distributions are the same and different. We need to develop similar bounds for a larger class of ML methods.

- **Resistance to adversarial process manipulation**: Any process can be manipulated using the knowledge of how it works. Examples range from the use of camouflage and decoys to obfuscate the locations of military equipment from reconnaissance satellites, to use of low probability of intercept/low probability of detection communications signals to prevent detection and localization. On a broader scale, information warfare seeks to achieve similar objectives. However, once AI/ML algorithms are introduced into the system, there are key questions about how they will respond to manipulated and potentially malicious inputs. Researchers recently showed that by manipulating a stop sign, an autonomous vehicle's computer vision algorithms would fail and cause the car to not stop at an intersection [18]. Prior research in cognitive wireless systems has shown that an attacker can introduce signals that slowly change decision boundaries for signal classifiers that use unsupervised learning [19], [20].

  Research on combating these issues can flow along the following lines:

  - *Securing input information*: Data that is used to train any system must be secured. Its disclosure would allow an adversary to identify training gaps, and its manipulation could cause systems to have learned backdoors that an adversary could walk through.

  - *Commonsense and fallback*: Any operational AI/ML system must have a coarse model of its environment that is hard-coded, and any output that is inconsistent with this model should raise alarm and trigger fallback to a simpler, less vulnerable algorithm, or a prior set of learned weights that have been vetted.

  - *Corroboration:* A future that ubiquitously leverages data from multiple collection modalities can be used to help to build evidence from multiple sources to more effectively corroborate or discredit a hypothesis that is partially influenced by malicious data. Adoption of a soft decision philosophy will also help with these objectives.

3.1.3 Next Generation AI and Sensemaking Algorithms

To move from simple analytics to sensemaking across multiple domains, it is essential to address key scientific challenges in ML and AI. Techniques such as non-monotonic reasoning, common sense reasoning, and explainable methods have been studied in classical AI literature for over four decades. New extensions of these methodologies for multi-modal analysis will yield scalable and robust methods. For example, non-monotonic

reasoning methods that can revise the beliefs as more multi-modal data become available will yield higher quality knowledge objects. Likewise, common-sense reasoning methods can be used for addressing the challenges due to maliciously corrupted multi-modal data. One of the drawbacks of the "black-box" approach that is dominant is that the analyst does not get additional insights into why a particular decision was made. Explainable common-sense reasoning methods for multi-modal data interpretation will provide a comprehensive understanding of situational awareness to analysts.

Another key problem to study in this area is the true integration of the results from sensemaking at several layers of the multi-modal sensemaking problem into one true global set of hypotheses and conclusions. Are there fundamental theoretical limits to sensemaking? What properties of the different layers of sensemaking and their inter-relationships will influence this fundamental theoretical limit? Examples of these properties could be the orthogonality of the hypotheses generated at the intra-layer levels. How does the intentional corruption of information studied in Section 3.1.2 affect the limits on overall sensemaking?

3.1.4 Assessing and Closing the Loop for Multi-modal Sensemaking Systems

The sensemaking system engineered to solve large complex problems using information from multiple domains must be evaluated to be able to compare it with other competing systems. It is conceivable that the key performance metrics that evaluate the performance of the AI/ML system at that level may vary with each layer, with some being probabilistic and some deterministic. Metrics to evaluate the overall sensemaking system must work with these disparate metrics to synthesize one measure of performance of these systems. What are the challenges in designing this overall metric? How do metrics such as precision, recall, and accuracy at different layers translate into an overall metric that captures the trustworthiness of the conclusions of the sensemaking system? What are the appropriate metrics to capture the efficiency of an AI system to perform a specific task in the operational environment? What are the metrics that capture the efficiency with which the sensemaking system works with humans?

4. HYBRID SYSTEMS FOR SENSEMAKING (HS^2)

The primary bottleneck in analyzing and sensemaking from vast volumes of intelligence data is not necessarily computational but human power. To adapt a state-of-the-art ML method to a specific problem requires computer science knowledge, time to implement, and domain expertise to interpret the results. Furthermore, some IC problems do not easily lend themselves to a computational approach and require more human decision making/deliberation. On the other hand, there are problems for which even generic, non-customized machine models can be used with excellent accuracy and efficiency, e.g., a general time-series based forecasting method can produce more accurate results than a human analyst and with far greater efficiency. This conundrum suggests a hybrid approach, where a human analyst is assisted by machine models, the extent of which depends on the type of the problem. Such an approach is currently being tested in the context of the Intelligence Advanced Research Projects Activity (IARPA)'s Hybrid Forecasting Competition[3], which is specifically designed for forecasting problems. Another relevant example is DARPA's Communicating with Computers program (CWiC), where one of the use-cases focused on building automated reasoning agents that can help biologists to make sense of vast biomedical literature. DARPA's Neurotechnology for Intelligence Analysts (NIA) and Cognitive Technology Threat Warning System (CT2WS) programs used "brain-in-the-loop" hybrid approaches to increase the efficiency and throughput of imagery analysis, while the CT2WS used a related approach (based on real-time analysis of human brain signals) to detect potential threats during real-time surveillance operations. The core underlying idea is to maximize human sensemaking capabilities by optimizing human information processing in domains in which completely automated approaches have insufficient sensitivity due to the domain knowledge required and the vast variability of potential targets (e.g., in satellite image analysis). The NIA and CT2WS established the feasibility of combining human neural processing with computers to augment human cognitive and sensemaking abilities, leading to a ten-fold increase in image analysis throughput in the NIA program. These programs were also a proof-of-concept for operational neuroscience, referring to how our understanding of the neural bases of human cognition could be exploited for the development of hybrid brain-computer systems to tackle relevant applied problems in the IC domain. An overview of these programs can be found in [21].

[3] https://www.hybridforecasting.com/

4.1 KEY RESEARCH TRACKS

The goal of this research area is to explore the concept of hybrid human-machine systems for more general IC-relevant sensemaking problems. In this section, we explore the following key research tracks:

- HS^2 Taxonomy & Performance Measures
- Interactive & Continuous Sensemaking
- HS^2 Autonomy & Trust
- HS^2 as Networks: Organizational Sciences Perspective
- HS^2 Interfaces

Figure 4 provides a schematic illustration of some of the aspects of hybrid systems that are IC-relevant. Section (a) of the figure depicts the current paradigm of an intelligence analyst interacting with large volumes of data, building and tracking multiple hypotheses with limited and localized input from machine-based methods. Figures (b)-(d) describe progressively hybrid systems where human analysts work collaboratively with machine models. In particular, (d) shows the sensemaking process within the future HS^2 systems where the machines are used not only for simple query answering but can be tasked with generating and evaluating various competing hypotheses based on analyst input, communicating those hypotheses to the analyst, finding information in support of those hypotheses, and so on.

Finally, the symbolic timeline at the lower part of each subsection of figure 4 corresponds to where that particular subsection falls on the timeline of evolution from current hybrid human-machine systems to a networked ecosystem of human analysts and machines. The current hybrid human-machine systems include a limited role for the machines. The machines are limited to tasks such as search and classification/categorization problems. As we move along the timeline, we can see that the hybrid system progresses to more synergistic systems where both humans and machines engage in more complex tasks such as generating and evaluating various hypotheses. Finally, in (d), we arrive at a networked ecosystem where both machines, as well as analysts, can have a diverse set of qualifications and expertise.

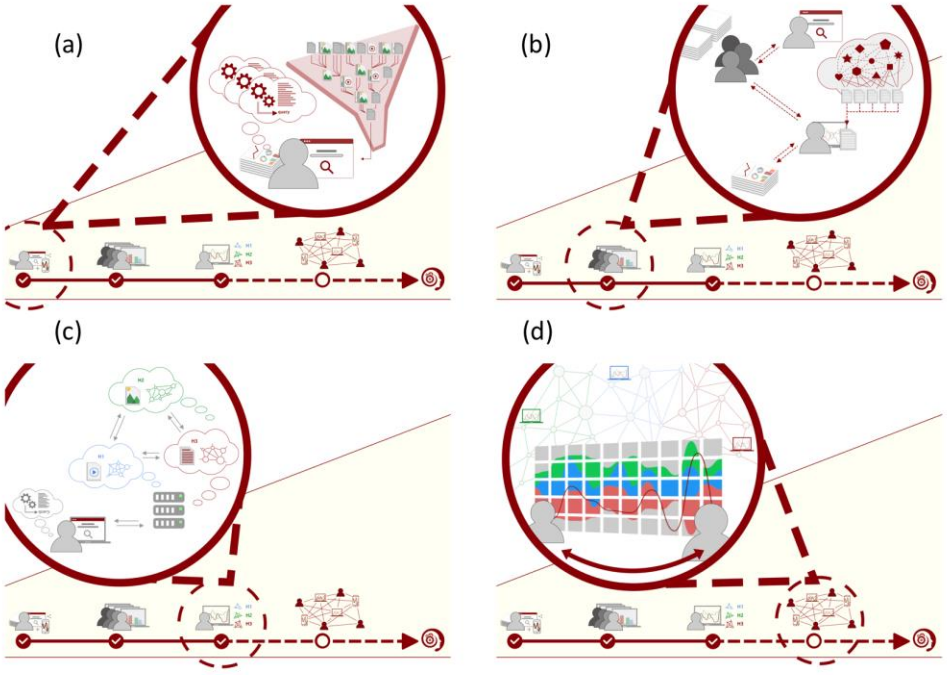

Figure 4: Schematic illustration of various aspects of hybrid systems for sensemaking. The four subfigures (a-d) depict gradually increasing levels of human-machine collaboration, with (d) representing a hybrid ecosystem of human analysts and machines with a complementary set of qualifications and expertise working collaboratively on complex sensemaking problems.

4.1.1 HS^2 Taxonomy & Performance Measures

One of the key issues that arose during the planning workshop was the need to establish a taxonomy of various hybrid human-machine sensemaking systems. There are several dimensions of those systems that a proper taxonomy should acknowledge, depending on the type of sensemaking tasks the taxonomy is addressing. For example, for a forecasting task discussed during the workshop, event ontology, data selection and provision, and data integration were suggested as relevant dimensions.

In addition to the taxonomy, there is also a need for developing measures and metrics for characterizing different aspects of HS^2 performance. One issue includes:

> **Developing measures that can characterize the reliability of human-based algorithms:** This problem becomes even more daunting for human-based algorithms as humans are not able to articulate how well they make their decisions. This problem was highlighted in the case of forensic examiners in a National Research Council report, ***On Strengthening Forensic Science in the United States: A Path Forward*** [22], [23]. In the most comprehensive study to date [24], forensic facial

examiners were found to be superior to both motivated control participants and to students on six tests of face identity matching. However, the performance was not supported by any quantitative measures or bounds. Because identification decisions in a forensic laboratory typically require days or weeks to complete and are made with the assistance of image measurement and manipulation tools, [25] the performance of forensic facial examiners reported in [24] represents a lower-bound estimate of the accuracy of examiners in practice. This problem needs solutions from cross-cutting disciplines drawn from human psychology, human factors, etc.

4.1.2 Interactive & Continuous Sensemaking

Most tasks that are of interest to IC are not one-shot but rather involve a multi-stage process of data gathering and hypothesis generation and refinement. The goal of this research track is to explore various concepts for developing interactive sensemaking systems where the analyst and machine models are engaged in multiple collaborative sessions for addressing those certain problems.

4.1.3 HS^2 Autonomy & Trust

A central question for HS^2 is the problem of autonomy, e.g., who has control over different aspects of the sensemaking process, how this control evolves in time, etc. Again, there are several important dimensions to consider when thinking about the autonomy in HS^2. For instance, in one scenario the human analyst will have full control over hypothesis generation, and the machines will be mostly tasked with finding relevant information (e.g., by multi-modal, automated sensemaking) either supporting those hypotheses or contradicting them. In another scenario, a machine can generate hypotheses themselves, perhaps with some input from an analyst.

Closely related to the issue of autonomy is the issue of *trust and algorithmic aversion*. Formal statistical models consistently make more accurate predictions than expert judgments due to the inconsistency of human judges. The inability of humans to decipher between predictive and unpredictive cues, and the inability of humans to detect how misrepresentative their experiences are [26], [27]. Two recent meta-analyses show that statistical (i.e., model-based) predictions beat clinical (i.e., human judgment) predictions by a factor of about 10-13 percent on average, though the rate changes with prediction format and context. Individuals tend to choose human forecasts over algorithmic predictions even after viewing information about the superiority of the algorithm. Such "algorithm aversion" is at least partly due to a more rapid loss of trust in algorithms when seeing them err [28]. Despite this general preference, algorithms influence human judgments more than other

USE CASE 1: HYBRID FORECASTING

The ability to anticipate and forecast events has long been a key underpinning of the U.S.' ability to shape effective responses when they occur. Previous IARPA programs such as Aggregative Contingent Estimation and Open Source Indicators have driven unprecedented but independent advances in human-based and automated (machine-based) forecasting systems. Research advances on those programs have been accompanied by insights into the individual weaknesses of these systems. Machine-based methods typically perform well on problems for which there is sufficient historical data but are ill-suited to forecast *rare event*s for which such data may not exist, or when the underlying context has changed in ways not reflected by the historical data. Human analysts, on the other hand, can often accurately forecast outcomes without relying exclusively on historical data, and by leveraging their judgment, domain knowledge, and prior experience. However, even the best analysts may not match machine performance where solid historical data is available and can be cognitively overwhelmed if they must deal with many problems in short time periods, thereby significantly limiting the scalability of a forecasting system that relies solely on human judgment. Motivated by experience in domains such as weather forecasting, IARPA's Hybrid Forecasting Competition program seeks to blend the distinct strengths of human- and machine-based forecasting systems to mitigate their individual weaknesses. Although the program is still in its initial stages, there are some encouraging signs pointing toward complementarity of human- and machine-generated forecasts.

When humans and machines interact for hybrid forecasting, it may induce an implicit hierarchical structure (e.g., a human domain expert at the top of the hierarchy). Each layer in such a hierarchy may observe inputs from multi-modal sensors, operate at different time scales, or produce vastly varying forecasting error rates, etc., e.g., [54]. Research is needed to analyze the fundamental limits of such hierarchical human-machine learning systems including mathematical modeling, convergence analysis, and system stability under different types of (stochastic) perturbations from the sensed environment.

humans [29]. Evidence suggests the cognitive mechanisms leading to use an algorithm is twofold. First, an individual's default forecasting mode is usually judgment-based, and people are consistently overconfident in their judgment [30], so it requires a substantial disruption to switch to an alternate method. Second, individuals tend to have non-normative, unrealistic expectations for algorithmic performance that far exceeds their success rate [31].

A recent review of trust in automation found three overlying factors that determine trust [32]. Dispositional trust is the individual's tendency to trust automation, situational trust

describes the contextual factors that alter trust, and learned trust is how interacting with automation promotes trust. The same typology seemingly applies to trust in algorithms, although the literature is sparser. First, algorithmic attitudes vary with *dispositional* traits of the judge. Individuals who view her or himself as more expert in the task are more likely to discredit the algorithm [29], and individuals who rely on their emotions to make decisions are less trusting [33]. Second, the *situation*, including the type and context of judgment, alters preferences of algorithms. People trust algorithms more for objective decisions and less for subjective and moral decisions [29], [31], [33]. Trust means different things when ascribed to algorithms, where it means efficiency and objectivity, compared to humans, where it means social recognition and authority. Third, trust in algorithms is malleable and can be *learned*. Trust increases when humans are shown feedback about machine performance [34], or allowed to interact, even with limited degree of modification [28].

Other areas relevant to proposed topics are: analysis of performance, process, and purpose in hybrid human-machine systems [35]–[37]; limitations of machines in terms of reliability, validity, robustness, and false-alarm rate [38]–[41], impact of machine performance on trust, [42], feedback and transparency issues [32]; trust with noisy data and detectability of trend [43]; and aggregation of human and machine-generated predictions [44].

4.1.4 HS^2 as Networks: Organizational Sciences Perspective

The main objective of this research track is going beyond the paradigm of a "single-analyst interacting with single machine" to allow for networked ecosystems of analysts and machines, where both machines, as well as analysts, can have diverse sets of qualifications and expertise. There is a large body of research on human teams suggesting that optimal team composition depends on the task at hand. Research under this track will extend this type of analysis to hybrid systems.

4.1.5 HS^2 Interfaces

The importance of the role of communication interfaces for HS^2 is hard to overestimate. This track is concerned with researching and eventually developing natural interfaces to allow seamless human-machine interactions for solving IC-relevant problems. Note that recent years have witnessed tremendous progress in natural language communication between humans and computers. Also, groundbreaking work, (e.g., the aforementioned DARPA Neurotechnology for Intelligence Analysts and Cognitive Technology Threat Warning System programs) has demonstrated feasibility of direct, highly efficient communication between brains and computers to augment human sensemaking abilities through real-time analysis of neural signals.

USE CASE 2: HYBRID IMAGE PROCESSING

Since the early nineties, DARPA and the IC have explored many approaches for how image analysts can work with image analysis and computer vision algorithms. As part of the DARPA-supported efforts known as the Research and Development for Image Understanding Systems [55] done during 1992-1996, the researchers interacted with image analysts and learned how they prioritize their tasks using quick-look profiles. Image analysis algorithms were then developed to populate the quick-look profiles automatically. In a subsequent DARPA effort known as Automatic Population of Geospatial Databases, many algorithms were developed for terrain feature extraction with the goal of assisting NGA analysts. While these efforts had good goals, due to less than expected performance of machine vision algorithms, the humans preferred to work by themselves. The common complaint was that humans had to spend a lot of time to fix the errors generated by the algorithm.

Dramatic improvements in the performance of machine learning algorithms for tasks such as object/face detection and verification have regenerated interest in the hybrid approach to the image processing task. It was shown recently [7] how forensic examiners and machine learning algorithms developed as part of the IARPA Janus program can work together to realize the best forensic performance. This study asked the following question: How can we achieve the most accurate face identification: using people and/or machines working alone or in collaboration? In a comprehensive comparison of face identification by humans and computers, the investigators compared the performance of forensic facial examiners, facial reviewers, and super-recognizers on a challenging face identification test. On the same test, four deep convolutional neural networks developed [56]–[58] under IARPA support, identified faces within the range of human accuracy. The accuracy of the algorithms increased steadily over time, with the most recent deep convolutional neural network scoring above the median of the forensic facial examiners. It was found that single forensic facial examiners fused with the best algorithm were more accurate than the combination of two examiners. Therefore, collaboration among humans and between humans and machines offers tangible benefits to face identification accuracy in important applications. These results offer an evidence-based roadmap for achieving the most accurate face identification possible.

These successes raise the exciting prospect of further advances in hybrid sensemaking systems that exploit recent insights into the neural bases of human sensemaking to achieve optimal integration of humans and computers. Traditional human-computer interaction systems usually use hand-crafted templates to complete tasks in specific domains, which do not generalize well to other scenarios [45]. Alternatively, statistical learning and neural network models can automatically learn from corpora to generate responses, which generalize to open domains [46]–[51]. The desire to maintain engagement relevance has

motivated the development of personalized dialogue agents [52]. Dialogue agents have been studied in the commercial 'chit-chat' space where engagement is valuable for its own sake but can be leveraged to the task completion use case by responding to queries in a user-specific way. Both context sensitivity and increased engagement over long sessions to prevent fatigue are worthwhile goals. An additional problem specific to the IC is that the task completion domain can change rapidly. Recent advances in zero-shot adaptation [53] enable trained task completion agents that can effortlessly pivot to new tasks without task-specific training data. These two threads of research can be usefully combined yielding personalized, action-focused natural language interactions to accomplish sensemaking tasks. While the research in this track should focus on adapting recent results to more IC-relevant problems, we also envision a need for fundamental research that focuses on efficient non-verbal modes of communications in HS^2.

5. Projected Timelines for Research Tracks

We present the projected timeline for both research tracks in this section. We include some the main sections of the research tracks as well as some more nuanced areas for your consideration.

The progress of research in multi-modal sensemaking and hybrid sensemaking systems are given in Figure 5 and Figure 6 respectively.

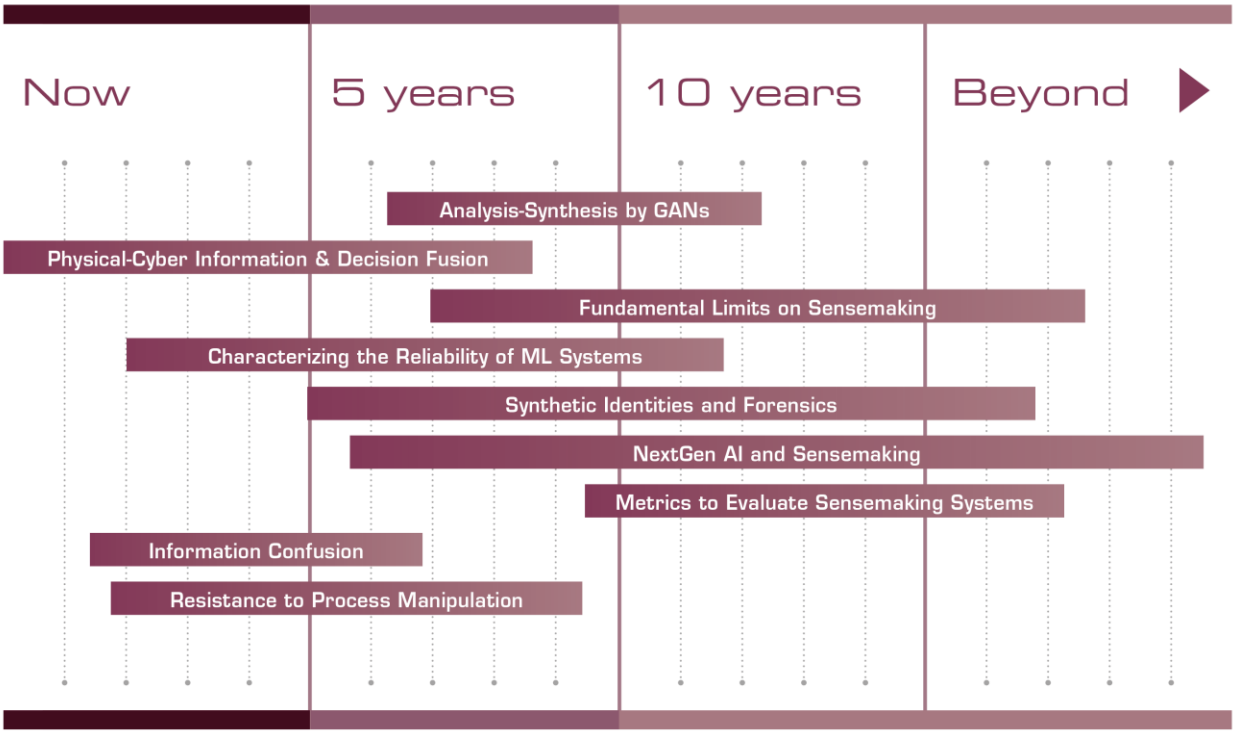

Figure 5: Multimodal Sensemaking: Projected Timeline

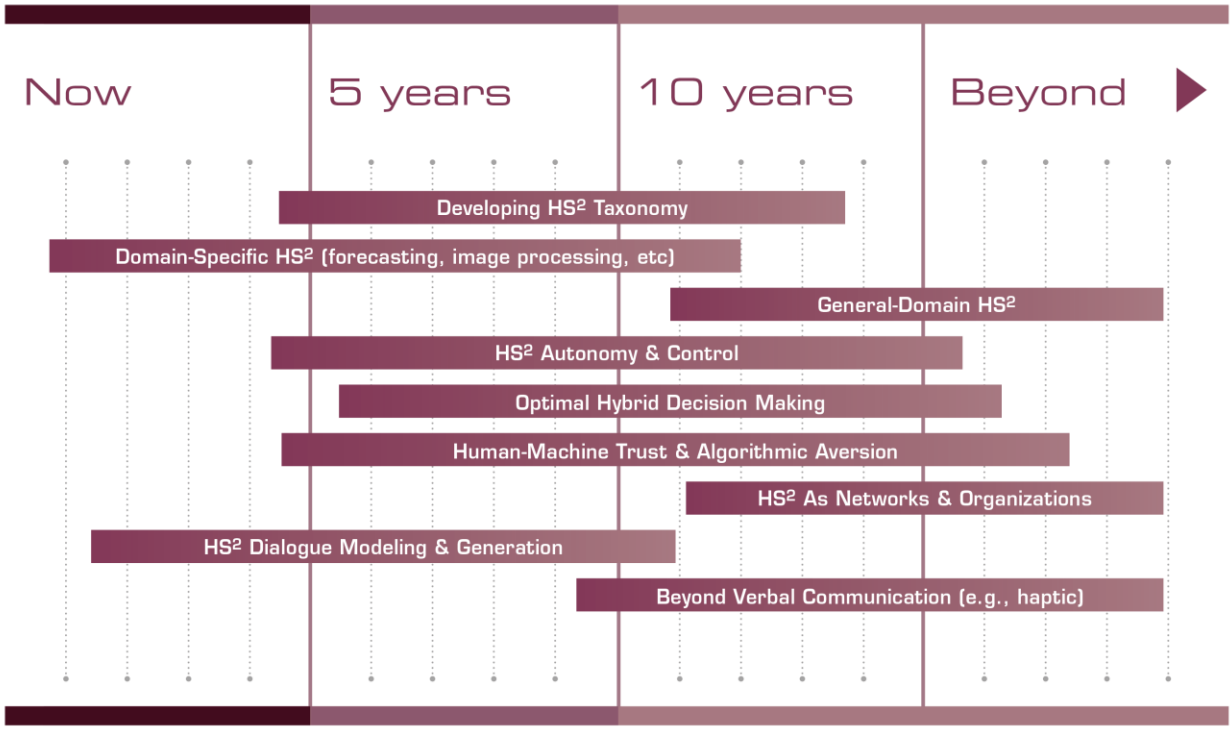

Figure 6: Hybrid Systems for Sensemaking (HS²): projected timeline

6. About the Authors

Professor K.P. (Suba) Subbalakshmi is the Founding Director of the Stevens Institute for Artificial Intelligence and a Professor in the Department of Electrical and Computer Engineering at Stevens Institute of Technology. She is also the co-founder of two technology start-up companies.

Prof. Subbalakshmi was named a Jefferson Science Fellow in 2016 by the National Academy of Sciences, Engineering and Medicine. As a Jefferson Science Fellow, she served as a senior science and technology adviser to the U.S. Department of State and worked on technology policy issues in Information Communication Technologies like the Internet of Things and 5G communications as well as AI and ML. She served as a subject matter expert for the National Spectrum Consortium in 2015. She is a Founding Associate Editor of the Institute of Electrical and Electronics Engineers (IEEE) Transactions on Cognitive Communications and Networking. She is the Founding Chair of the Special Interest Group on Security, IEEE Communications Society (ComSoc)'s Technical Committee on Cognitive Networks. She is a recipient of the New Jersey Inventors Hall of Fame, Innovator Award. She has given several tutorials, keynote addresses, and participated in several panel discussions for IEEE and other international conferences and events. Prof. Subbalakshmi has published several papers, book chapters, a book, and holds five patents. Her research is supported by the National Science Foundation, the National Institutes of Justice, the Air Force Research Laboratory, the U.S. International Student and Scholar Office, and other DoD agencies.

Dr. Aram Galstyan is the Director of the Machine Intelligence and Data Science (MINDS) group at Information Sciences Institute, University of Southern California, and a Principal Scientist at USC-ISI. He is also research associate professor at the USC Computer Science department. His work focuses on various problems at the intersection of ML, information theory, and statistical physics. His research includes both theoretical effort and more application-oriented work geared toward describing various real-world phenomena. Various U.S. funding agencies, including the National Science Foundation, National Institutes of Health, DARPA, IARPA, and Army Research Office have supported his research.

Prof. Rama Chellappa is a Distinguished University Professor and a Minta Martin Professor of Engineering and in the Department of Electrical and Computer engineering at the University of Maryland. He is a recipient of the K.S. Fu Prize from the International Association of Pattern Recognition IAPR, the Society, Technical Achievement and Meritorious Service Awards from the IEEE Signal Processing Society, and the Technical Achievement and Meritorious Service Awards from the IEEE Computer Society. He also received the Inaugural Leadership Award from the IEEE Biometrics Council. At UMD, he has received college and

university level recognition for research, teaching, innovation, and mentoring of undergraduate students. He has been recognized with an Outstanding Electrical and Computer Engineering Award and a Distinguished Alumni Award from Purdue University and the Indian Institute of Science, respectively. He is a Fellow of IEEE, International Association on Pattern Recognition, Optical Society of America, American Association for the Advancement of Science, Association of Computing Machinery, and American Association for Artificial Intelligence and holds five patents. His research program has been funded by the National Science Foundation, DoD, DARPA, IARPA, and several companies. Recently, he graduated his 100th doctoral student.

Dr. Charles Clancy is the executive director of Virginia Tech's Hume Center for National Security and Technology and is a professor of electrical and computer engineering. With 80 professors, researchers, and staff, the Hume Center engages over 400 students annually in research and experiential learning focused in national security and technology through $10M to $15M per year in grants and contracts from defense and intelligence agencies.

Dr. Clancy is an internationally-recognized expert in wireless security, having chaired standards groups and regularly testifying to the U.S. Congress as an expert on issues of telecommunications and cybersecurity. Before joining Virginia Tech in 2010, he led a $75M portfolio of wireless research and development programs at the Laboratory for Telecommunications Sciences, a National Security Agency research laboratory located near the University of Maryland campus.

Dr. Clancy received his BS in Computer Engineering from the Rose-Hulman Institute of Technology, MS in Electrical Engineering from the University of Illinois, and Ph.D. in Computer Science from the University of Maryland. He has over 230 peer-reviewed technical publications and patents, is co-author to five books, and co-founder of four venture-backed startup companies that collectively employ over 150 people in the Boston and D.C. metro areas.

7. ACKNOWLEDGMENTS

We want to thank R. Chandramouli, Max Reisenhuber, Fred Morstatter, Seth Goldstein, Jon Wade, Tom McDermott, and the attendees of the ODNI Sensemaking Workshop for their input, discussions and useful comments.

APPENDIX A: WORKSHOP PARTICIPANTS

Guest Name	Affiliation
Andy Monje	Department of Defense
Aram Galstyan	University of Southern California
Barry Boehm	University of Southern California
Barry Horowitz	University of Virginia
Bill Rouse	Stevens Institute of Technology
Bill Webster	Office of the Director of National Intelligence
Boyan Onyshkevych	Defense Advanced Research Projects Agency
Charles Clancy	Virginia Tech University
Chris North	Virginia Tech
Dan Delaurentis	Purdue University
Daniel McAdams	Texas A&M
David Honey	Defense Advanced Research Projects Agency
David Isaacson	Office of the Director of National Intelligence
David Jenkins	Pennsylvania State University
David Nahamoo	Pryon Inc
Dean Souleles	Office of the Director of National Intelligence
Debra Stanislawski	Office of the Director of National Intelligence

Guest Name	Affiliation
Dinesh Verma	Stevens Institute of Technology
Erica Briscoe	Georgia Tech
Fred Morstatter	Information Sciences Institute
Gary Witus	Wayne State University
Gerry Dozier	Auburn University
Heather Lench	Texas A&M
Hortense Gerardo	Associate Professor of Anthropology at Lasell College
Jeff Moulton	Louisiana State University
Jeff Solka	Naval Surface Warfare Center–Dahlgren
Joe Olive	Defense Advanced Research Projects Agency
John Sustersic	Penn State University
Jon Wade	Stevens Institute of Technology
K.P. Subbalakshmi	Stevens Institute of Technology
Kevin Sullivan	University of Virginia
Maximilian Riesenhuber	Georgetown University
Megan Clifford	Stevens Institute of Technology
Melvin L. Eulau	Office of the Director of National Intelligence
Ming Dong	Wayne State University

Guest Name	Affiliation
Naren Ramakrishnan	Virginia Tech
Nicole Hutchinson	Stevens Institute of Technology
Patrick Wolfe	Purdue University
Paul Brigner	Georgetown University
Peter Perolli	Institute for Human & Machine Cognition
Philip Odom	GA Georgia Tech
Rajarathnam Chandramouli	Stevens Institute of Technology
Rama Chellappa	University of Maryland
Richard Malak	National Science Foundation
Scott Lucero	Department of Defense
Seth M. Goldstein	Intelligence Advanced Research Projects Activity
Steve Thompson	Office of the Director of National Intelligence
Tom McDermott	Georgia Tech
Tomas Diaz de La Rubia	Purdue University
Vanessa Lewis	National Reconnaissance Office
William Millonig	Office of the Director of National Intelligence

APPENDIX B: ACRONYM LIST

AI	Artificial Intelligence
AIDA	Active Interpretation of Disparate Alternatives
CT2WS	Cognitive Technology Threat Warning System
CWiC	Communicating with Computers
DARPA	Defense Advanced Research Projects Agency
DASD(SE)	Office of the Deputy Assistant Secretary of Defense for Systems Engineering
DoD	Department of Defense
GANs	Generative Adversarial Networks
IARPA	Intelligence Advanced Research Projects Activity
IC	Intelligence Community
IEEE	Institute of Electrical and Electronics Engineers
LORELEI	Deep Exploration and Filtering of Text Low Resource Languages for Emergent Incidents
ML	Machine Learning
NGA	National Geospatial-Intelligence Agency
NIA	Neurotechnology for Intelligence Analysts
ODNI	Office of the Director for National Intelligence
SERC	Systems Engineering Research Center

SVM	Support Vector Machines

WORKS CITED

[1] "Intelligence for everyone."

[2] G. Strauss, A. Shell, R. Yu, and B. Acohido, "SEC, FBI probe fake tweet that rocked stocks," *USA Today Online*, p. 1, 2013.

[3] X. Chen, R. Chandramouli, and K. P. Subbalakshmi, "Scam Detection in Twitter," *Data Min. Serv.*, 2014.

[4] S. Kwon, M. Cha, K. Jung, W. Chen, and Y. Wang, "Prominent features of rumor propagation in online social media," in *Proceedings - IEEE International Conference on Data Mining, ICDM*, 2013.

[5] E. Bakshy, I. Rosenn, C. Marlow, and L. Adamic, "The Role of Social Networks in Information Diffusion," *Proc. 21st Int. Conf. World Wide Web SE - WWW '12*, 2012.

[6] A. Louni and K. P. Subbalakshmi, "Who Spread That Rumor: Finding the Source of Information in Large Online Social Networks with Probabilistically Varying Internode Relationship Strengths," *IEEE Trans. Comput. Soc. Syst.*, vol. 5, no. 2, pp. 335–343, 2018.

[7] P. Jonathon Phillips, Amy N. Yates, Ying Hu, Carina A. Hahn, Eilidh Noyes, Kelsey Jackson, Jacqueline G. Cavazos, Géraldine Jeckeln, Rajeev Ranjan, Swami Sankaranarayanan, Jun-Cheng Chen, Carlos D. Castillo, Rama Chellappa, David White, Alice J. O'Toole, "Face recognition accuracy of forensic examiners, superrecognizers, and face recognition algorithms" Proceedings of the National Academy of Sciences May 2018,

[8] M. Liggins II, D. Hall, and J. Llinas, *Handbook of Multisensor Data Fusion: Theory and Practice, Second Edition*, 2nd ed. CRC Press, 2009.

[9] V. M. Patel, R. Gopalan, R. Li, and R. Chellappa, "Visual Domain Adaptation: A survey of recent advances," *IEEE Signal Process. Mag.*, 2015.

[10] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, "Generative Adversarial Attacks," *arXiv [stat ML]*, p. 1206.2661, 2014.

[11] C. Strasser and G. A. Antonelli, "Non-Monotonic Logic," *Stanford Encyclopedia of Philosophy*. [Online]. Available: http://plato.stanfor.edu/index.html.

[12] E. Davis, *Representations of Commonsense Knowledge*. San Francisco, CA: Morgan Kaufmann series in representation and reasoning, 2014.

[13] L. Casola, Ed., *Challenges in Machine Generation of Analytic Products from Multi-*

Source Data: Proceedings of a Workshop. Washington, DC: The National Academies Press. National Academy of Sciences., 2017.

[14] S. O. Arik *et al.*, "Deep Voice: Real-time Neural Text-to-Speech," Feb. 2017.

[15] S. Anand, K. P. Subbalakshmi, and R. Chandramouli, "A quantitative model and analysis of information confusion in social networks," *IEEE Trans. Multimed.*, vol. 15, no. 1, pp. 207–223, 2013.

[16] J. Shawe-Taylor and N. Cristianini, *Kernel Methods for Pattern Analysis*. 2004.

[17] M. Kabkab, E. Hand, and R. Chellappa, "On the size of Convolutional Neural Networks and generalization performance," in *Proceedings - International Conference on Pattern Recognition*, 2017.

[18] K. Eykholt *et al.*, "Robust Physical-World Attacks on Deep Learning Models," Jul. 2017.

[19] T. Newman and T. Clancy, "Security threats to cognitive radio signal classifiers," in *Virginia Tech Wireless Personal Communications Symposium*, 2009.

[20] T. C. Clancy and A. Khawar and T. R. Newman, "Robust Signal Classification Using Unsupervised Learning," *IEEE Trans. Wirel. Commun.*, vol. 10, no. 4, pp. 1289–1299, 2011.

[21] T. A. Miranda *et al.*, "DARPA-funded efforts in the development of novel brain-computer interface technologies," *Elsevier J. Neurosci. Methods*, vol. 244, pp. 52–67.

[22] "Strengthening Forensic Science in the United States: A Path Forward," Washington, DC, 2009.

[23] D. White, K. Norell, P. Jonathon Phillips, and A. J. O'Toole, "Human factors in forensic face identification," in *Advances in Computer Vision and Pattern Recognition*, 2017.

[24] D. White, P. Jonathon Phillips, C. A. Hahn, M. Hill, and A. J. O'Toole, "Perceptual expertise in forensic facial image comparison," *Proc. R. Soc. B Biol. Sci.*, 2015.

[25] "Guidelines for Facial Comparison Methods, Version 1.0," 2012.

[26] R. M. Dawes, D. Faust, and P. E. Meehl, "Clinical versus actuarial judgment," *Science (80-.).*, 1989.

[27] P. E. Meehl, *Clinical versus Statistical Prediction*. Minneapolis, MN: University of Minnesota Press, 1954.

[28] B. J. Dietvorst, J. P. Simmons, and C. Massey, "Algorithm Aversion: People Erroneously Avoid Algorithms after Seeing Them Err," *SSRN Electron. J.*, vol. 143, no. 6, pp. 1–13,

2014.

[29] J. M. Logg, J. A. Minson, and D. A. Moore, “Algorithm Appreciation: People Prefer Algorithmic to Human Judgment,” *Harvard Bus. Sch. Work. Pap. No. 17-086*.

[30] W. R. Sieck and H. R. Arkes, “The recalcitrance of overconfidence and its contribution to decision aid neglect,” *J. Behav. Decis. Mak.*, 2005.

[31] B. J. Dietvorst, “Consumers and Managers Reject (Superior) Algorithms Because They Fail to Compare Them to the (Inferior) Alternative,” *Adv. Consum. Res.*, vol. 45, pp. 302–306, 2017.

[32] K. A. Hoff and M. Bashir, “Trust in automation: Integrating empirical evidence on factors that influence trust,” *Hum. Factors*, 2015.

[33] N. Castelo, M. Bos, and D. Lehmann, “Consumer adoption of algorithms that blur the line between human and machine. (Under review),” *J. Mark. Res.*, 2018.

[34] M. T. Dzindolet, L. G. Pierce, H. P. Beck, and L. A. Dawe, “The Perceived Utility of Human and Automated Aids in a Visual Detection Task,” *Hum. Factors J. Hum. Factors Ergon. Soc.*, 2002.

[35] J. Lee and N. Moray, “Trust, control strategies and allocation of function in human-machine systems,” *Ergonomics*, 1992.

[36] B. M. Muir, “Operators’ trust in and use of automatic controllers in a supervisory process control task. Doctoral Dissertation.,” University of Toronto, Canada, 1989.

[37] S. Zuboff, *In the Age of the Smart Machine: The Future of Work and Power.* New York: Basic Books, 1988.

[38] R. R. Hoffman, M. Johnson, J. M. Bradshaw, and A. Underbrink, “Trust in automation,” *IEEE Intell. Syst.*, 2013.

[39] J. D. Lee and K. A. See, “Trust in Automation: Designing for Appropriate Reliance,” *Hum. Factors J. Hum. Factors Ergon. Soc.*, 2004.

[40] B. M. Muir and N. Moray, “Trust in automation. part ii. experimental studies of trust and human intervention in a process control simulation,” *Ergonomics*, 1996.

[41] R. Parasuraman and V. Riley, “Humans and Automation: Use, Misuse, Disuse, Abuse,” *Hum. Factors J. Hum. Factors Ergon. Soc.*, 1997.

[42] M. T. Dzindolet, S. A. Peterson, R. A. Pomranky, L. G. Pierce, and H. P. Beck, “The role of trust in automation reliance,” *Int. J. Hum. Comput. Stud.*, 2003.

[43] M. S. Gönül and P. Goodwin, “Why Should I Trust Your Forecasts?,” *Foresight Int. J. Appl. Forecast.*, 2012.

[44] D. Önkal, P. Goodwin, M. Thomson, S. Gönül, and A. Pollock, “The relative influence of advice from human experts and statistical methods on forecast adjustments,” *J. Behav. Decis. Mak.*, 2009.

[45] T. Misu and T. Kawahara, “Speech-based interactive information guidance system using question-answering technique,” in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 2007.

[46] N. Peng, M. Ghazvininejad, J. May, and K. Knight, “Towards Controllable Story Generation,” pp. 43–49, 2018.

[47] D. Bahdanau, K. Cho, and Y. Bengio, “Neural Machine Translation by Jointly Learning to Align and Translate,” Sep. 2014.

[48] A. Sordoni *et al.*, “A Neural Network Approach to Context-Sensitive Generation of Conversational Responses,” Jun. 2015.

[49] N. Peng, H. Poon, C. Quirk, and K. Toutanova, “Cross-Sentence N -ary Relation Extraction with Graph LSTMs,” *TACL*, 2017.

[50] M. Shang, Z. Fu, N. Peng, Y. Feng, D. Zhao, and R. Yan, “Learning to Converse with Noisy Data : Generation with Calibration,” pp. 4338–4344, 2016.

[51] J. Li, M. Galley, C. Brockett, G. P. Spithourakis, J. Gao, and B. Dolan, “A Persona-Based Neural Conversation Model,” Mar. 2016.

[52] S. Zhang and J. Dinan, Emily Urbanek, Jack Szlam, Arthur Kiela, Douwe Weston, “Personalizing Dialogue Agents: I have a dog, do you have pets too?,” in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2018, pp. 2204–2213.

[53] T. Zhao and M. Eskenazi, “Zero-Shot Dialog Generation with Cross-Domain Latent Actions,” no. May, p. 4803, 2018.

[54] M. A. Haleem and R. Chandramouli, “Adaptive downlink scheduling and rate selection: a cross-layer design,” *IEEE J. Sel. Areas Commun.*, vol. 23, no. 6, pp. 1287–1297, 2005.

[55] O. Firschein and T. Strat, Eds., *RADIUS, Image Understanding for Imagery Intelligence*. San Francisco, CA: Morgan Kaufman Publishers, 1998.

[56] J. C. Chen *et al.*, “Unconstrained Still/Video-Based Face Verification with Deep Convolutional Neural Networks,” *Int. J. Comput. Vis.*, 2017.

[57] R. Ranjan *et al.*, “Deep Learning for Understanding Faces: Machines May Be Just as Good, or Better, than Humans,” *IEEE Signal Process. Mag.*, 2018.

[58] R. Ranjan *et al.*, “Crystal Loss and Quality Pooling for Unconstrained Face Verification and Recognition,” vol. 14, no. 8, pp. 1–15, 2018.