CCICADA and Big Data

Delivered to the US Department of Homeland Security by CCICADA
May 2014

1. Introduction

Everyone is talking about ***Big Data***. Last year, six agencies offered $200M in ***Big Data*** grants in a wide-ranging White House initiative and now we're about to see another round of funding. But what exactly is ***Big Data***? Why is it considered so important? What about data has changed in the last five years? The last ten years? What might the situation be five or ten years from now? What role has CCICADA played in this arena and what role will it play in the future? These are the questions we address in this report.

2. What is Big Data and How Has it Changed?

There are many different ways to describe ***Big Data***. Whereas Massive Data has a precise definition (data not fitting into computer memory, thus requiring out of memory algorithms for solving complex problems), ***Big Data*** has no such definition. We, at CCICADA, look at ***Big Data*** operationally, namely data so large that what to save is at question and, in some cases, decisions on that data have to made instantaneously. The most obvious change for data science is that a few years ago there was just not enough data and now, in many cases, there is just too much. This is a relatively new problem and we need methods to identify the relevant data, or most relevant data at the given time, and even which data needs to be saved for later use as it might become relevant.

Big Data is sometimes described in terms of the three V's, volume, variety, and velocity, with the emphasis that it is not necessarily an increase in any one of them that has created a challenge but the concomitant increase in all three. Perhaps more importantly is that it is not volume, variety, or velocity even together that define ***Big Data***, but it is something more difficult to define and capture, namely ***complexity***. Simply put, **data today is very large, heterogeneous, interrelated and complex.**

Data science is an old field (Galileo Galilei was a data scientist and not the first). Data science as we have come to think of it with data mining and data analytics is enabled by

ever-increasing volumes of sensor data, the ability to transmit data over ever-higher capacity networks, storage devices that can store and retrieve massive amounts of data, growing computing power, and the demand for faster solutions to complex problems. These trends are not expected to change.

Big Data now proliferates even in the commercial world, as well as involving the academic world and large and small US Government agencies. Storing or warehousing data is a commercial success. Academic researchers continue to work on discovering new techniques for dealing with long and streaming data sets. Government agencies worry about how to regulate the things we can do with data that we couldn't just a short time ago.

This new data science is increasingly involved in simulation and modeling, especially focused on developing models that exploit all available data to predict costs, threats and threat reductions. This clearly is one of the most important strengths at CCICADA.

Big Data also involves now much more complex and heterogeneous data. Such data can be dirty (noisy), wide (more variables than cases), and fuzzy (involving uncertainty). The focus on making decisions under dirty and fuzzy data and of developing a theory of modeling heterogeneous data for applications such as anomaly detection are areas of CCICADA strength.

Big Data has led to using many different techniques for coming up with decisions. It is rare to be able to get a simple solution with such data. Most often alternative solutions need to be evaluated. CCICADA has been working on methods for choosing among packages of alternative flood mitigation strategies or packages of alternative nuclear detection algorithms in which to invest. This is similar to business and government's use of auctions involving a huge number of bidders playing at lightning speed. (This is an area where some CCICADA researchers also have considerable practical experience.) Composite auctions lead to NP-complete allocation problems. One may not even be able to feasibly express all possible preferences for all subsets of goods. Determining the winner is computationally intractable for many economically interesting combinations. And guaranteeing the privacy of auction participants requires cryptographic methods to preserve such privacy. In many auctions time for bidding is extremely limited, sometimes a matter of a few seconds. This requires software agents to act on behalf of humans and the lightning speed allows many repeated auctions where bidders can learn their opponents' preferences and utilities through repeated games. These repeated auctions represent the possibility of making sequential decisions, where one learns from earlier decisions to choose the next best one – an issue that has arisen in CCICADA work, for example on inspection procedures.

Big Data problems are cross-disciplinary, requiring diverse topical data scientists to work in teams. Examples abound: NEON (National Ecological Observatory Network), a project sponsored by NSF, involves gathering data from 20 sites across the US to get a continent-wide picture of the impacts of climate change, land use change, and invasive species on natural resources and biodiversity; with biodiversity a key indicator of environmental health, GBIF (Global Biodiversity Information Facility) is an international effort to digitize all information about all living species, and requires collating data from formalized

databases, handwritten research notes, and images and video; intelligence analysts work with data from multiple disciplines, including financial transactions, communications, and human behavior; Coast Guard fisheries law enforcement involves understanding of fish population and fish migration data, weather patterns, the economics of fish price, the sociology of interactions with commercial fishing fleets, diverse intelligence data, and changing fisheries technology; authorities dealing with disease outbreaks need to understand not just epidemiology, but sociology (will individuals follow instructions), economics (what incentives to give people to comply with those instructions), transportation (as diseases spread rapidly with modern transportation systems), ecology (as for diseases spread by migratory birds) and climate science (effect of climate change on new disease patterns). CCICADA researchers span many relevant disciplines, from the mathematical sciences (computer science, mathematics, statistics, operations research) to engineering (civil engineering, industrial engineering, transportation science) to the social and behavioral sciences to economics and business to medicine and public health, and work in cross-disciplinary teams on most of the Center's projects.

Today's data science is changing rapidly. Important data science subdisciplines include the following categories: Foundational (Mathematics, Statistics, Operations Research, Computer Science); data engineering (storage, data-enabled energy use, data warehousing); perception (visual analytics); applied computer science methods (machine learning, high performance computing, pattern recognition and learning, uncertainty modeling, information networks and social media).

Newer components of data science arise from the increasing multi-disciplinarity of the problems that ***Big Data*** is allowing us to address. These newer components include social psychology (to elicit information/data from groups and individuals); human computer interaction (to improve efficiency of humans interacting with tools and systems that process data); economics (cost of data analysis and storage); behavioral economics (to understand incentive structures and carry out cost-benefit analyses); perceptual and cognitive science (to place a firm scientific foundation on topics from visual analytics and related areas); public health analogies (e.g. between good cyber practices and good health practices); methods of information elicitation (often based in the social sciences).

CCICADA researchers have been at the forefront in developing some of these newer components of data science. For the next five years, it will be important to give increasing emphasis to such topics as social-science-based methods, behavioral economics, and cognitive science for foundational work on human perception of information.

Computing performance may not be growing at the rate it did a decade ago. Moshe Vardi, Editor-in-Chief, of Communications of the ACM [May 2014, Vol 57, #5] discusses how Moore's Law is dead and the resulting strongest implication, namely that performance will now require improved algorithms, something CCICADA has been a proponent of. With ever increasing data, with the ability to run numerous new algorithms, with the ability to integrate adaptive simulations with streaming data, with the harnessing of using such algorithms in dealing with erroneous, noisy and fuzzy data, the challenge is no longer just in data acquisition but in the algorithms.

3. Sources of Data

Data is coming to us from a wide variety of sources: news, text, audio, imagery, video (including from security cameras), LiDAR (including from modeling/simulation and virtual training, geophysical analyses, disaster analyses), all types of sensors, GPS systems, smartphones and tablets, and new types of media. Some of that data is of lower quality, such as web data, blogs, forums, and tweets. It is this data heterogeneity that will lead to answers to the increasingly complex questions being generated; but this requires new processing methods.

The remarkable variety of data sources present some new challenges for data science. In the rest of this section, we describe some of these challenges, and then in the next section we talk about some general CCICADA approaches to them.

***Fusion Challenge*:** Many analysis tasks require the fusion of information from various media or sources. For example: How can one combine the "hard" numerical readings of sensors monitoring emergency vehicle movements with the "soft" natural language utterances of the driver with the tweets of the public? How can one extract information from the relevant images and connect this information to that obtained from video and other sources?

It is reasonable to argue that 85% to 90% of data nowadays is unstructured (text, video data, etc.), some of which disappears quickly (cf. snapchat).

***Unstructured Data Challenge*:** The large amounts of unstructured data and its complexity prevent us from moving from the current common techniques for the discovery of simple relationships to gaining a deeper understanding of objects, context, events, in other words, what is happening at a more granular level.

Data, whether Big or Small, is not necessarily valuable in its own right, but only in how it is useful. A ***Big Data*** challenge is to utilize data to make better decisions.

Decision Support Challenge: Decision science is an old topic that is coming to have major new components. Once the domain of social scientists and economists, it is now in the domain of computer scientists and mathematicians who, working with decision scientists trained in the social science-economics tradition, need to develop tools of modeling, simulation, algorithmics, uncertainty quantification, and consensus.

Near-real-time situational awareness or what is coming to be called real-time analytics includes simulation and modeling, data and simulations from mobile applications, sensor data. Such data can be too rapid for real-time human consumption or exploration.

Real-time Analytics Challenge: Some data rates are so large that not all the data can be saved and yet real-time or almost real-time decisions must be made, e.g., sampled

astronomical data. Consider for example a smart grid where status updates which came in every two to four seconds are now approaching ten times a second using new phasor technologies. That rate may be too rapid for a human alone to absorb the presence of an anomaly in time to act upon the information, thereby requiring agent or algorithmic support.

Much data often also involves graphs and networks: IP traffic level, access logs, command logs and rapidly evolving time graphs and networks. Situational awareness requires us to translate such data into large, interpretable and manageable graphs which scale, which can be monitored to detect local changes that may not have visible effect on global metrics, and where subgraphs can be monitored in cases where we may not even know beforehand which subgraphs to monitor.

Streaming Data Challenge*:* New algorithms are needed to deal with large and possibly massive graphs over time, and in some cases streaming in real time.

Our modern society is critically dependent upon ***Big Data***: manufacturing and production systems, power and water systems, transportation systems, and financial systems, for example. Vulnerabilities are ever present: cyber attacks on our infrastructure, cascading failures, rapid spread of anomalies, biological terrorism. Indeed, it is the very ability to utilize and benefit from large amounts of data that creates vulnerabilities.

Vulnerabilities Challenge: How do we identify new vulnerabilities caused by usage of data? How do we develop tools for monitoring and minimizing such vulnerabilities?

4. Some CCICADA General Approaches to the Above

New Approaches: These include the development of new methods of abstracting information and integrating it using evolving standardized ontology-driven representation schemes; of methods to predict interpretations or textual annotations for images that do not have text; and of automated object detection such as the likely locations of known objects in an image. CCICADA advances include major new tools for image analysis and algorithms for translating images to text.

The Future: A major challenge is to be able to aggregate information from multiple sources, some credible and some of dubious quality. This is where a theory of "trust" needs to play a role (see Section 5.5).

New Approaches: At the same time, having large amounts of data provides opportunities to better understand what is happening. For example, if you want to know if a certain treatment has specific side effects, yesterday you would ask four doctors; today you can get hundreds of thousands of people to tell you what they think. This gives rise to "*citizen science*."

The Future: CCICADA researchers have begun to explore citizen science in connection with their work on urban responses to climate events. This is an intriguing area for the next five

years of CCICADA research. And it connects to a somewhat related area of using humans as sensors.

New Approaches: New tools of decision making involve ability to make decisions in sequence, learn from earlier mistakes, and take the preferences of extensive stakeholder communities into account through modern tools of algorithmic decision theory and preference analytics.

The Future: New, more complex machine learning issues arise from the need to understand the complexity of preference and utility. New ways to elicit information to build the models used in decision support are called for. New tools for solving the optimization problems arising from sequential decision making are needed.

New Approaches: "Sketches" that allow complex reports on large streaming data are a key tool that have been developed by CCICADA researchers. Developing classifiers to automatically label incoming requests (e.g., HTTP requests) as "valid" or "attack" and to identify attack types have also been explored by CCICADA researchers. Other approaches developed by CCICADA-affiliated researchers are tools allowing tracking data access anomalies and quick identification of worm signatures; use of cyclical network statistics to monitor time-varying streaming network data in an automated online fashion; metric forensics to track changes at the micro, meso, and macro levels of a volatile network; and new proof protocols for verifying computations that are streaming in nature.

The Future: A major challenge is to find ways to summarize data, without being able to store individual items, in a way that allows one to uncover patterns from the summaries, patterns that might not have been in areas of interest at the time the summaries are produced. Another challenge is to use data to go beyond understanding that an event is taking place to get at causality to aid in post-event mitigation or prevention of future events.

New Approaches: Much of the work on cyber-physical systems is aimed at understanding the vulnerabilities that result from using high-powered algorithmic tools based on extensive availability of data to run our machines. However, this subject is relatively young.

The Future: New tools are needed to identify sources of vulnerability and develop general principles for vulnerability minimization. Particular challenges arise from the anticipated many uses for unmanned and remote automated vehicle systems and the data they can gather along with the vulnerabilities created by their use.

New Approaches: You can duplicate information without much cost, but you cannot duplicate physical products so easily. The resulting redundancy of information means that original owners of information need to be protected, and spread of information without proper permission needs to be regulated. How can this be accomplished? Research in this area has emphasized compensation to original owners, fair division of costs resulting from spread of information (e.g. through multicasting), and precise definition and application of

rule-based permissions for access to copies of information. Much work has also been done on privacy, e.g., in information sharing contexts – see Section 5.4.

The Future: What is the role of government in protecting us from problems caused by ***Big Data***? What is government's role in regulating and preventing the exploitation of ***Big Data***? In terms of copyright, privacy, ownership, governments are far behind in the information world than they are in the physical world. Tools to assist governments in this realm will need to be developed through partnerships among data scientists, lawyers, and social scientists.

The next two sections provide details focused on getting information from data and turning that information into knowledge for decision making.

5. Information from Data

A key challenge for data science is to be able to aggregate data from multiple sources with potentially questionable quality and credibility, and obtain useful "information" as a result. This is an area where CCICADA has put a great deal of emphasis and expects to continue to do so.

5.1. Information Access and Information Distillation

Frequently, the information of interest is buried in very large amounts of data and in varying forms that make it difficult to reveal the content of interest (e.g., text). As noted above, studies suggest that 85-95% of the data available to agencies and corporations is unstructured: text, images, video, etc., obtained from social media, email, websites and other sources. How can one deal with the huge amount of unstructured data as if it were organized in a database with a known schema? How do we locate, organize, access, analyze and synthesize unstructured data, in particular the microtext commonly used in social media? Simply put, the challenge is to transform data to Information. The difficulties have to do with both *ambiguity* of representation and the *variability* of expressing meaning. While this challenge is not new, the approaches to it have been changing.

Information Access Challenge: How can one develop high-accuracy task-driven information search and access capabilities? This challenge is not new, but what is new is to go "beyond Google" to find totally new ideas and approaches.

New Approaches: These involve deployment of advanced techniques for semantic interpretation, using pattern matching, concept mining, ontologies, "story discovery," and other methods that usually involve machine learning. CCICADA advances since 2009 include significant advances in our toolkits and our model of distributional semantics.

The Future: one new approach is to develop special extraction technology combined with classifiers to learn the current "story" being told in streaming social media and related microtext data sources across multiple dimensions of time, space, and "story" dimension.

Information Distillation Challenge**:* We would like to make inferences and derive hypotheses from large amounts of data. Data seldom exists in the form most suited for analysis. Usually, data must be extracted from its environment and distilled or reformulated into some internal format, sometimes also interpreted in terms of existing internal knowledge structures. This is a key challenge for ***Big Data science. Again, the challenge is not new, but the speed with which we must draw inferences in the age of ***Big Data*** presents a big new challenge here.

New Approaches: A long-term effort in data science has been to perform information extraction from natural language and images. Since 2009 CCICADA has developed various methods of automated pattern learning (for example, Double-Anchored Patterns) that are more expressive and much more accurate.

The Future: It is often necessary to determine the normal state of a system in order to be able to quickly detect departures from normality. Machine learning approaches can help to define normality and hence abnormality, but we need rapid, real-time methods to develop hypotheses about departure from normality that will allow us to take rapid responses (as in the example of the smart grid given above).

5.2. Information Storage and Management and New Architectures

The extracted and distilled information must usually be stored for later re-use. How does one store, query, and search data or information when there is so much of it?

***Information Storage and Management Challenge**:* We need to find ways to create very large-volume databases that support data homogenization across various sources. This will require the formation of large semantic multi-graphs and other complex representations and tools for the manipulation and management of these large-volume databases and complex representations. Most tasks with data require specialized analysis and inference in order to extract just the right combination of information units. For example, we need tools to assist an analyst in exploring various alternatives in order to arrive at valid interpretations and useful conclusions.

New Approaches: There is increased emphasis on sophisticated technology from information retrieval (vector spaces, etc.) and from artificial intelligence (knowledge representation and inference systems). New relevant tools are being developed in time series analysis, text-based inference, evidence gathering, and evidence propagation across knowledge/belief networks, evidence combination and hypothesis formation. CCICADA has contributed significant pieces of this work.

The Future: Data evolves, reflecting changing points of view, opinions, environmental conditions, etc. One challenge analysts face is follow the development dynamics of adversarial views of a topic, an interest in a technology, or an opinion. These evolving views can be modeled as a time series, which can then be studied in conjunction with corresponding text data, in the search for triggers for early warning. Research is needed to develop text mining techniques that can analyze time series variables together with

arbitrary companion text data to discover causal topics. Forecasting is another research challenge for the future. For example, in the context of social networks, can one predict evolving connections? In other contexts, can you use evolving individual characteristics to predict divergent/dangerous behavior?

Much data has grown too large to reside in one location. New architectures need to be developed. Reflecting this, there is increasing emphasis on use of the "the cloud" to do computations, store data, etc. This is a major change in the short time since CCICADA was founded.

New Architectures Challenge*:* As more computation is outsourced to a potentially untrusted third party ("the cloud"), we need assurances that computations are performed correctly as claimed. There is need for protocols for querying a database kept in the cloud that keep the query confidential and return records only when the query is authorized. And, there is need for new methods for information retrieval in new architectures other than the cloud. We also need new architectures that implement distributed algorithms to support low latency and enable real or near real-time access and analytics of unstructured, complex data from different devices.

New Approaches: There are new proof protocols for verifying computations that are streaming in nature. Methods have been developed that have a "guarantee" that is so strong that even if the service provider deliberately tries to cheat, there is only vanishingly small probability of doing so undetected. There has also been development of three-party architectures for safety and security with the cloud: using data manipulation, searchable encryption, conditional oblivious transfer, and secure function evaluation. Work has also been done on using a smartphone as a personalized security hub. It connects to personal computing devices (home appliances, your car, personal robots). It does security screening of these devices. If necessary it connects to secure back-end cloud for heavyweight security screening, serving as a secure conduit to the cloud. Many of these approaches are due to CCICADA-affiliated researchers.

The Future: The cloud and the smartphone both present major issues with regard to safety and security of data, as well as major new opportunities for storage and management of data. These issues should play a key role in the development of ***Big Data*** science in the future.

5.3. Information Networks and Analysis

Increasingly, data creators and sources exist in networks, with information coming from sources such as individuals of interest, sensors, databases, or any other data concentration, and information coming from relations among sources such as varying types of connections between nodes or various types of data flow between nodes. These networks are large, dynamic, and evolving rapidly in many cases.

Information Networks Challenge*:* What can one learn from network structure, data flow, individual interactions, and longer-term data trends? Typically there is too much

information for a human to handle, and the desired information is not immediately apparent.

New Approaches: CCICADA is performing network-oriented data analysis that may involve time and/or space, as in Twitter analysis. The area of information networks has changed drastically since CCICADA was formed, due to: a) scale; b) the realization that combining content and network structure is essential; and c) major advances in algorithmic frameworks for dealing with heterogeneous networks. CCICADA researchers have played a major role in this; related are major new techniques for social media analysis, privacy, and trust developed at CCICADA.

The Future: One new set of ideas is to take note of the fact that information sources are moving and developing tools to track such movement and understand, evaluate, and analyze the trajectories in order to gain insight into the underlying networks. Another key idea is to develop ways to infer leadership in dynamic networks (through tweets, snapchat, and various kinds of forums).

5.4. Information Sharing, Privacy

Secure information sharing is a key to enable agencies and individuals to work together on a wide range of homeland security issues. Privacy considerations are essential to secure information sharing. The importance of privacy is emphasized in a recent White House report on ***Big Data,*** where a majority of the recommendations were about privacy.

Secure and Trustworthy Cyberspace Challenge: With changing cyber security enhancements, what are the effects on privacy? What methods can be developed that will enable new technologies and methodologies while protecting privacy of information?

New Approaches: CCICADA's work in this area has emphasized a foundational approach to secure and trustworthy cyberspace: What is the boundary of possibility and impossibility? Are there fundamental tradeoffs (e.g., between security and efficiency? communication and computation? privacy and usability?) The insights gained from this work can lead to more informed choices and better system design. Much current work in this area takes a *cryptographic approach*. Earlier work has led to tools for: Privacy-preserving construction of Bayesian networks, privacy-preserving clustering, privacy-preserving reinforcement learning. Newer efforts emphasize privacy-preservation in preprocessing and postprocessing steps; secure and private database access; and "differentially" private information and learning theory.

The Future: We have just begun to see a new effort in *systematizing secure computation*. A systematic characterization of solutions will allow decision makers to understand essential strengths and weaknesses of different secure computation solutions and determine which best applies in a given scenario. The goal is to systematically characterize existing solutions according to their properties, including: their methods of representing functions; their adversarial models; whether or not they guarantee fairness; requirements on their execution environments; prerequisites regarding correctness, auditability, and

compliance; intractability assumptions; communication latency; and computational overhead.

Information Sharing Challenge: Information sharing requires appropriately safeguarding both systems and information; selecting the most trusted information sources; and maintaining secure systems in potentially hostile settings. How can one best accomplish these things?

New Approaches: Some key research directions are to develop an analytical basis to describe, measure and assure needed levels of privacy in multi-media data; methods to assess the trustworthiness of data and data sources; secure frameworks for information integration; and methods to ensure secure networks. Prior to CCICADA's work, existing methods for anonymization to enable information sharing were limited to structured data (e.g., database and contingency tables). CCICADA researchers found techniques for a richer set of data types: Unstructured (text, image, temporal events) and structured (semantic graphs, transactional data). CCICADA researchers also found techniques that are "tunable" to different levels of information sensitivity. *Secure multiparty computation* aims at allowing parties to jointly compute something over their inputs while at the same time keeping those inputs private. This is another area of strength for CCICADA.

The Future: Expanded efforts on secure information sharing for richer data sets are called for. Secure multiparty computation at scale is still very expensive but new approaches to it are needed. These approaches will involve exploring the tradeoff between efficiency and security: Is it okay to have some security loss? One should also consider the secure implementation of privacy-critical subroutines and insecure implementation of the rest.

5.5. Trustworthiness

Data comes from multiple sources. What sources are more accurate than others? Multiple information sources often provide information that is not consistent and is often conflicting — either maliciously, or due to various kinds of noise. This is especially so in emergency situations where heterogeneous information flows in multiple information streams, providing information on the situation in different locations affected. To utilize the vast amounts of data available to us in this age of ***Big Data***, we have to understand what sources we can trust.

The Trustworthiness Challenge: We can view the information trustworthiness problem as *sources* producing *claims* (or sets of claims) and we must find the appropriate degree of trust in each of these elements, and provide *evidence* for why this is so. How can we develop a computational framework that addresses the problem of trustworthiness in disasters and other situations?

New Approaches: CCICADA researchers have developed computational models that are based on a precise definition of the factors that contribute to trustworthiness, including accuracy, completeness, and bias, and finding a satisfactory way of measuring these.

Work has included development of trust metrics that reflect typical characteristics and reflect the roles of truthfulness, completeness and bias in identifying trustworthiness. There are also constrained trustworthiness models that allow us to go beyond simple majority, and incorporate prior beliefs and background knowledge. Finally, there has been work on incorporating evidence for claims sufficient to convince users of the trustworthiness of sources and claims.

The Future: Work is needed to develop claim verification systems, with automated claim verification by finding supporting and opposing evidence. We also need to understand how a trustworthiness framework can be integrated with information extraction systems.

6. Turning Information into Knowledge for Decision Making

Big Data is not necessarily a problem in and of itself. It is the need or desire to use it that presents the challenge. So, we shouldn't worry about the data alone without considering the analytics and applications. Along these lines, there are pressures for using as much data as is out there, just because it is there. We heard VADM Chuck Michel of the Coast Guard tell us that he could be rightfully criticized if there is an emergency and he hadn't made use of all the available data to address it.

There are many uses of ***Big Data***. Here, we shall emphasize use of ***Big Data*** to make better decisions.

Today's decision makers in fields ranging from engineering to medicine to homeland security have available to them: a) remarkable new technologies; b) huge amounts of information; and c) ability to share information at unprecedented speeds and quantities. These tools and resources will enable better decisions if we can surmount concomitant challenges:
a) The massive amounts of data available are often incomplete or unreliable or distributed and there is great uncertainty in them
b) Interoperating/distributed decision makers and decision-making devices need to be coordinated
c) Many sources of data need to be fused into a good decision, often in a remarkably short time
d) Decisions must be made in dynamic environments based on partial information
e) There is heightened risk due to extreme consequences of poor decisions
f) Decision makers must understand complex, multi-disciplinary problems

When faced with such issues, decision makers need the help of data-driven methods to support better decisions. Decision makers today have few highly efficient algorithms to support decisions. The new field of algorithmic decision theory, which CCICADA researchers have helped to create, aims to exploit algorithmic methods to improve the performance of decision makers (human or automated). Algorithmic approaches to decision support, which make use of ***Big Data***, are in their infancy and much work is needed to develop this field of algorithmic decision theory. This is especially true of the need to gain real-time situational awareness (discussed in Section 3) and to make decisions

in real-time.

6.1. Sequential Decision Making

Decisions are rarely made in isolation and a major area of study today is how one can learn from the effects of earlier decisions to make better future decisions. This issue arises in such diverse areas as network security (testing connectivity, sequencing tasks), manufacturing (testing machines, fault diagnosis), artificial intelligence (optimal derivation strategies in knowledge bases), and medicine (diagnosing patients, sequencing treatments).

Sequential Decision Making Challenge: How do we best learn from earlier decisions to make better subsequent decisions?

New Approaches: The CCICADA work on sequential inspection and data-based sequential testing has developed new foundational tools such as binary-decision-tree-based approaches, layered defense, and the concept of a polytope of paths. This work has had applications in stadium security, container inspection, nuclear detection, and other areas.

The Future: Work is needed to extend research on sequential decision making to approaches that scale to the size and speed of modern homeland security decision problems, specifically for sequential decision making in real-time. The optimization problems arising from some of the modern paradigms of sequential decision making need to be solved for those modern paradigms to find practical use.

6.2. Preference Analytics

Learning preferences and/or opinions and, more generally, being able to infer ordering relations out of direct or indirect observation of human behavior is a key issue in order to be able to design intelligent, flexible and reliable decision support systems.

Preference Analytics Challenge: In today's large decision making contexts, the preferences of a decision maker can rarely be expressed by functions that are easily evaluated. More often, they are partially observed based on choices made in a few specific instances. How do we elicit preferences from stakeholders and infer preferences from data?

New Approaches: Automatically inferring or "learning" preference models based on empirical observation is an important task, and a modern trend is to use new variants of machine learning. In this context, preferences can be represented as utility functions or as binary relations. In the latter case, the machine learning challenge is to predict complex binary relations such as rankings or partial orders, rather than single values as in traditional machine learning. In many settings, preferences need to be learned from incomplete training data, relative preferences, or implicit data (as opposed to specific comparisons of alternatives), for example by inference from a person's actions. A classic example is estimating user preferences based on "click-through" data, where we assume

that a click on a page indicates a user's preference for that page over another. The beginning of work on preference analytics has been developed in part by CCICADA researchers.

The Future: Major new challenges for machine learning in this age of ***Big Data***-driven decision support is to learn rankings or preference orders in dynamic or changing environments, under adversarial conditions, and for composite objects consisting of portfolios of items. Also, how do we best use learned preferences to make recommendations for policy options? Note that the term "preference" is too specific. We also need to "learn" other kinds of binary relations, such as "riskier than" or "more likely than."

6.3. Information-driven Modeling and Simulation and Uncertainty Quantification

A key tool of the analyst or decision maker is modeling. Simulation and modeling are needed when the information is so complex as to resist other analytical methods, as is so often the case in this age of ***Big Data***.

Modeling and Simulation Challenge*:* In order to assist decision makers in utilizing tools of modeling and simulation, we need to apply methods that will allow us to understand the data they have, the goals and objectives of their work, their problem-specific business rules, and the parameter values needed as inputs to the models we build to assist them. How do we get reasonable estimates of parameters needed for a model? How do we work with decision makers to make their "business rules" specific?

New Approaches: A great deal of CCICADA's work with the Coast Guard, DNDO, and other agencies has involved making precise their business rules, formulating their objective functions in a precise way, and identifying values of key parameters. This involves developing new methods of elicitation but also design of experiments that can cut down significantly on the relevant range of parameter values that need to be considered to find meaningful distinctions. The tools include the development of powerful tools of combinatorial experimental design.

The Future: We need new methods of information elicitation to support model building. These methods will most likely result from combined efforts of data scientists with social scientists.

How do we make precise the uncertainty in our understanding of parameter values and of the conclusions from a model? Uncertainty comes from parameter values, model relationships, recorded observations. Uncertainty also comes from conflicting sources.

Uncertainty Quantification Challenge*:* How can we obtain results about uncertainty from limited data? How is it best to present levels of uncertainty? How can we interface complex modeling of processes and data and use this to obtain validated computational predictions from data?

New Approaches: We have begun to see validated computational predictions from data. New tools for model validation, model calibration, code verification are being developed. These include methods of data assimilation, error bars, and sensitivity analysis (an old topic with new directions).

The Future: A key challenge is to develop tools to compare levels of uncertainty across different models and to find tools for developing consensus when different models lead to at least seemingly different conclusions. This problem arises in a big way in climate science, but also in the need to compare alternative models leading to different conclusions in intelligence analysis and in other applications.

7. Education

Education is a key component in development of data science experts who will produce cutting edge tools and techniques for gaining insight from data as well as who will develop the tools to protect us against the vulnerabilities created from ***Big Data***-enabled systems. Cyber security is a case in point, where it has been estimated that we will have a shortfall of 700,000 cyber security experts within five years. We need to start early with education in data science, to attract people into the field in general and into cyber security in particular. While the emphasis in cyber security education has been on educational programs in computer science and electrical and computer engineering, we need to understand the relevance of and develop new programs that interface data science with such disciplines as political science, cognitive science, business, law, engineering, and public health. Thus, a key challenge arises from the relevance of multiple disciplines to the problems of concern in homeland security.

The Education for Multi-Disciplinarity Challenge: How can we train the next generation of homeland security workers (and more generally the workforce of the future) to work across disciplines and incorporate multiple points of view both in research and application?

New Approaches: CCICADA researchers have pioneered in some areas of multi-disciplinary education (such as developing materials that can be simultaneously used by faculty in different disciplines).

The Future: Development of materials for and training courses for multi-disciplinary collaborations should be an important area of emphasis for the CCICADA educational program going forward in the next phase of the Center.

CCICADA partner institutions have developed broad-based data science educational programs that are addressing not only the need for new data science and data scientists, but the multi-disciplinary training required in today's data science. In the Appendix we present a sample of relevant degree/certificate programs (emphasizing the Masters level) at the CCICADA institutions. We mostly give those that have data science or data specifically in the name, though many other programs have data science components.

At this writing, there are also major new ***Big Data*** initiatives, including plans for new degrees and certificate programs or tracks, at our university partners.

Appendix: A Sampling of Data Science Degree/Certificate Programs at CCICADA Partner Institutions, Mostly with Data or Data Science in the Name

Carnegie Mellon University

CYLAB has various MS programs relating to Cyber Security (see https://www.cylab.cmu.edu/education/index.html):

Master of Science in Information Security Technology and Management (MSISTM)
Master of Science in Information Technology - Privacy Engineering (MSIT-PE)
Master of Science in Information Networking (MSIN)
Master of Science in Information Security Policy and Management (MSISPM)
Kobe Master of Science in Information Technology, Information Security (MSIT-IS)

CS Department:

Master of Information Technology Strategy (MITS)
http://www.cmu.edu/mits/index.html

MS of Computational Data Science (used to be MLSI = MS in Large-Scale Informatics)
http://mcds.cs.cmu.edu/
2-year MS program focusing on large-scale software

MS in Intelligent Information Systems (MIIS)
http://www.lti.cs.cmu.edu/education/miis/

Language Technologies Institute:

MS in Intelligent Information Systems (MIIS)
http://www.lti.cs.cmu.edu/education/miis/
16-month MS program focusing on IR

CS Department, Language Technologies Institute, Machine Learning Department (jointly):

Data Sciences MS program (formerly known as Very Large Integrated Systems Program)
http://www.cmu.edu/graduate/data-science/data-science-masters-table-landscape.pdf

City College of New York

CS Department

Masters in Information Systems
http://css8a0.engr.ccny.cuny.edu/csblog/

Rensselaer Polytechnic Institute

Lally School of Management
MS in Business Analytics
http://www.lallyschool.rpi.edu/academics/ms_ba.html

Multidisciplinary Degree Program in Information Technology and Web Science
MS in Information Technology (concentration in Data Science and Analytics)
http://www.rpi.edu/dept/IT/graduate/index.html

Rutgers University

Graduate School – New Brunswick
Professional Science Masters Program – Concentration in Analytics
http://psm.rutgers.edu/programs/analytics

Rutgers Discovery Informatics Institute
Graduate programs in discovery informatics and data science
http://rdi2.rutgers.edu/education-training/opportunities-professional-graduate-students
Master of Business and Science in Discovery Informatics and Data Sciences
Graduate Certificate Program in Computational and Data-Enabled Science and Engineering (CDS&E)
Graduate Certificate Program in Discovery Informatics

Edward J. Bloustein School of Planning and Public Policy
Geospatial Information Science Certificate
http://policy.rutgers.edu/academics/uppd/certificates.php

Business School
MBA in Analytics and Information Management
http://www.business.rutgers.edu/mba/concentrations/aim

School of Communication and Information
MS in Communication and Information Studies
http://comminfo.rutgers.edu/master-of-communication-and-information-studies/mcis-home.html

Statistics Department
MS in Statistics with Option in Data Mining
http://stat.rutgers.edu/home/kolassa/Graduate/msoptions.html

University of Illinois-Urbana Champaign

Dept. of Statistics
MS Concentration in Analytics

http://www.stat.illinois.edu/students/msanalytics.shtml

Dept. of Computer Science
Certificate Programs
http://cs.illinois.edu/prospective-students/graduate-students/non-degree-certificate-program
Security
Networks and Distributed Systems
Information Systems
Software Engineering
Systems Software

University of Massachusetts Lowell

Department of Computer Science
MS Computer Science (Data Science) forthcoming 2014
http://www.uml.edu/Sciences/computer-science/default.aspx

University of Southern California

Department of Computer Science
MS in Computer Science (Data Science)
http://www.cs.usc.edu/academics/masters/msdata.htm

Marshall School of Business, Department of Data Science and Operations
PhD in Data Science and Operations
http://www.marshall.usc.edu/faculty/iom/curriculum