NASA/TM-2016-219358

Machine Learning Technologies and Their Applications for Science and Engineering Domains Workshop – Summary Report

Manjula Ambur
Langley Research Center, Hampton, Virginia

Katherine G. Schwartz
Georgia Institute of Technology, Atlanta, Georgia

Dimitri N. Mavris
Georgia Institute of Technology, Atlanta, Georgia

December 2016

NASA/TM-2016-219358

Machine Learning Technologies and Their Applications for Science and Engineering Domains Workshop – Summary Report

Manjula Ambur
Langley Research Center, Hampton, Virginia

Katherine G. Schwartz
Georgia Institute of Technology, Atlanta, Georgia

Dimitri N. Mavris
Georgia Institute of Technology, Atlanta, Georgia

National Aeronautics and
Space Administration

Langley Research Center
Hampton, VA 23681

December 2016

Acknowledgments

The authors gratefully acknowledge the fifteen presenters/experts from Universities and Industry for their time and effort in enthusiastically participating in the workshop and the ensuing discussions. Our thanks to the BDAMI team members Jeremy Yagle, Robert Milletich, Ted Sidehamer, Stephen Casey, Evana Gizzi, Macarena Ortiz, Daniel Sammons, and Zachary Ernst for their help in organizing the workshop and their thoughtful inputs to the next steps section for this effort. The development and application of machine learning techniques by the BDAMI team would not have been possible without the strong collaboration and continued support from the subject matter experts from the mission organizations; we sincerely thank the following for their valued contributions - Danette Allen; Damodar Ambur; Dale Arney; Trey Arthur; Erik Axdahl; Randy Bailey; Kris Bedka; Eric Burke; Jeff Cerro; Kyle Ellis; Christie Funk; Dana Hammond; Angela Harrivel; Jeff Herath; Jon Holbrook; Patty Howell; Lisa Le Vie; Constantine Lukashin; Louis Nguyen; Jeremy Pinier; Alan Pope; Brandi Quam; Laura Rogers; Cheryl Rose; Jamshid Samareh; Mark Sanetrik; Rob Scott; Lisa Scott-Carnell; Steve Scotti; Patrick Shea; Walt Silva; Mia Sicohi; Chad Stephens; Scott Striepe; Patrick Taylor; Marty Waszak; Bill Winfree; Kristopher Wise. We have special appreciation for the following subject matter experts for their valuable time and full engagement in the workshop discussions and providing insightful inputs in their specific domains for next steps section – Chris Rumsey, Lisa Le Vie, Patrick Taylor, Bill Kimmel and Alan Pope. Bianca Clark from the National Institute for Aerospace was instrumental in communicating, setting up and managing the website, and the logistics to bring the speakers from different parts of the country and we thankfully acknowledge her efforts. The authors and the BDAMI team thank the Reid Center team, especially Andrea McGruder for their help with arrangements at the Pearl Young and Reid Center for this three-day workshop. Last but not the least, we are indebted to Marty Waszak for his encouragement and Convergent Aeronautics Solutions/CAS Project funding support to conduct this workshop.

The use of trademarks or names of manufacturers in this report is for accurate reporting and does not constitute an official endorsement, either expressed or implied, of such products or manufacturers by the National Aeronautics and Space Administration.

Available from:

NASA STI Program / Mail Stop 148
NASA Langley Research Center
Hampton, VA 23681-2199
Fax: 757-864-6500

Contents

Executive Summary

The fields of machine learning and big data analytics have made significant advances over the past several years and have been demonstrating potential to transform how the traditional disciplines of science and engineering are conducted. The new, advanced methods combined with rapidly evolving computational capabilities, has created an environment where cross-fertilization of methods and unique collaborations can achieve previously unattainable outcomes. NASA Langley Research Center has recognized these changes in the technical and scientific communities and created the Comprehensive Digital Transformation (CDT) initiative that focuses on the development and synergistic integration of four main pillars: advanced modeling and simulation, machine learning and big data analytics, high performance computing, and advanced IT infrastructure.

The CDT Machine Learning and Big Data Analytics team planned a workshop held at NASA Langley in August 2016 to bring together leading experts the field of machine learning and NASA scientists and engineers. The primary goal for this workshop was to assess the state-of-the-art in this field, introduce these leading experts to the aerospace and science subject matter experts, and develop opportunities for collaboration. The workshop was held over a three day-period with lectures from 15 leading experts followed by significant interactive discussions. The invited lecturers and their lecture topics were:

- Dr. Sebastian Pokutta, Georgia Institute of Technology
  Machine Learning in Engineering: Applications and Trends
- Dr. Ella Atkins, University of Michigan
  New Data Sources to Revolutionize UAS Situational Awareness and Minimize Risk
- Dr. Chris Codella, IBM, Watson Group
  Cognitive Computing and IBM Watson in Research, Operations, and Medicine
- Dr. Tsengdar Lee, NASA Science Mission Directorate
  NASA Earth Science Knowledge Network
- Dr. Barnabas Poczos, Carnegie Mellon University
  Applied Machine Learning for Design Optimization in Cosmology, Neuroscience and Drug Discovery
- Dr. Lyle Long, Pennsylvania State University
  Toward Human-Level (and Beyond) Artificial Intelligence
- Dr. Una-May O'Reilly, MIT
  Machine Learning: Data Driven Artificial Intelligence
- Dr. Matthias Scheutz, Tufts University
  Intelligent Agents: One-Shot Learning through Task-Based Natural Language Dialogues
- Dr. Dimitri Mavris, Georgia Institute of Technology
  Application of Machine Learning for Aircraft Design
- Dr. Karthik Duraisamy, University of Michigan
  Data-driven Turbulence Modeling: Current Advances and Future Challenges
- Dr. Heng Xiao, Virginia Polytechnic Institute and State University
  A Physics-Informed Machine Learning Framework for RANS-Based Predictive Turbulence Modeling
- Dr. Krishna Rajan, University at Buffalo SUNY

Materials Informatics: Mining and Learning from Data for Accelerated Design and Discovery

- Dr. Jaime Carbonell, Carnegie Mellon University
  Machine Learning and Data Analytics for Aircraft Design and Operation: CMU and Boeing Partnership
- Dr. Vipin Kumar, University of Minnesota
  Big Data in Climate: Opportunities and Challenges for Machine Learning and Data Mining
- Dr. Raju Vatsavai, North Carolina State University
  Global Earth Observations Based Machine Learning Framework for Monitoring Critical Natural and Man-Made Infrastructures

This report provides an overview of each of the 15 invited lectures, as well as a summary of the key discussion topics that arose during both formal and informal discussion sessions. Each participant in the workshop, whether they were an invited lecturer or an attendee in the audience, were encouraged to seek out collaboration opportunities and identify areas of synergy in the field. Four key workshop themes were identified after the closure of the workshop, and insight into these themes is provided in this report after the discussion after the workshop discussion summaries. The four identified key themes are: classification with imperfect data, usage of natural language commands, applications in science and engineering, and the concept of the 'virtual assistant'.

Upon the conclusion of the workshop, several attendees in different research fields provided their feedback on how they are already utilizing machine learning algorithms to advance their research, new methods they learned about during the workshop, and collaboration opportunities they identified during the workshop. Input was received from specialists in the field of computational fluid dynamics, aerospace system design, human machine interaction, climate science, and aircraft training and safety. Furthermore, the NASA CDT Machine Learning/ Big Data Analytics team provided their insights from the workshop with respect to key takeaways and potential follow-up opportunities.

Overview of NASA Langley's Comprehensive Digital Transformation (CDT) Initiative

NASA and the nation have unique challenges in aeronautics, space exploration, and science. Even now, it takes approximately 10 years from conceptualization to developing and deploying an evolutionary aircraft, a launch system, or an instrument for gathering earth science data. This severely impacts system affordability and our nation's global competitive position. NASA Langley Research Center (LaRC) initiated Comprehensive Digital Transformation (CDT), which is intended to serve as a catalyst to create an integrated, digital tools and technologies capability to enable transformational changes in conducting relevant and innovative research, systems analysis, and design. This is achieved by augmenting NASA's efforts by leveraging and synergistically combining non-NASA funded, state-of-the-art advancements in modeling and simulation, high performance computing (HPC), big data analytics and machine intelligence, and IT infrastructure – the four core capability areas. Applying these capabilities both individually and through convergence of these compute- and data- intensive capabilities will lead to innovative concepts, reduced design cycle time, improved affordability, and increased confidence in the designs.

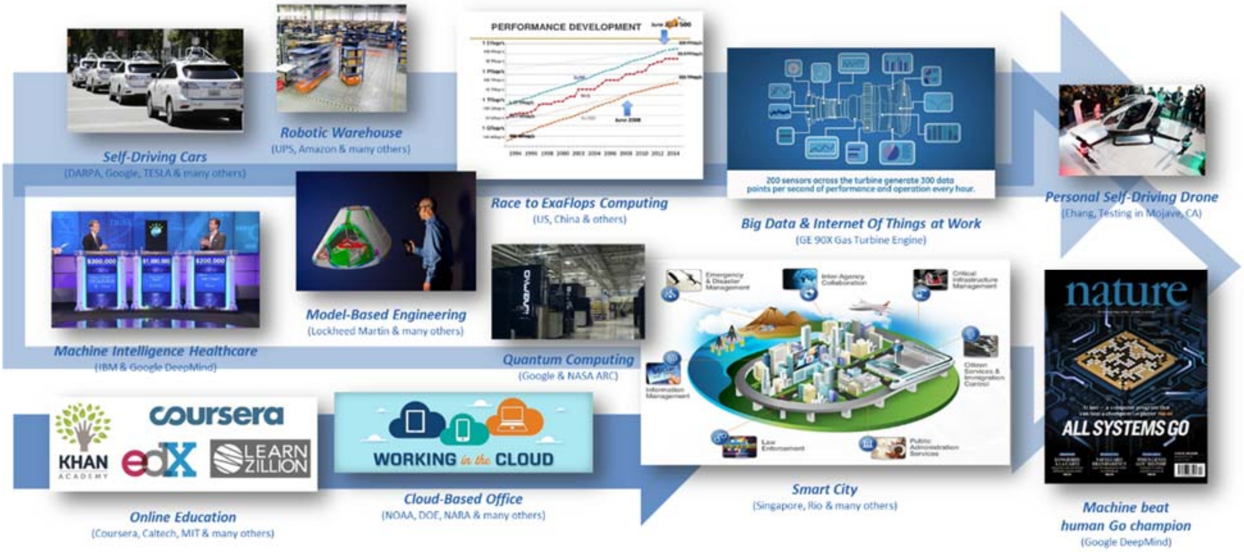
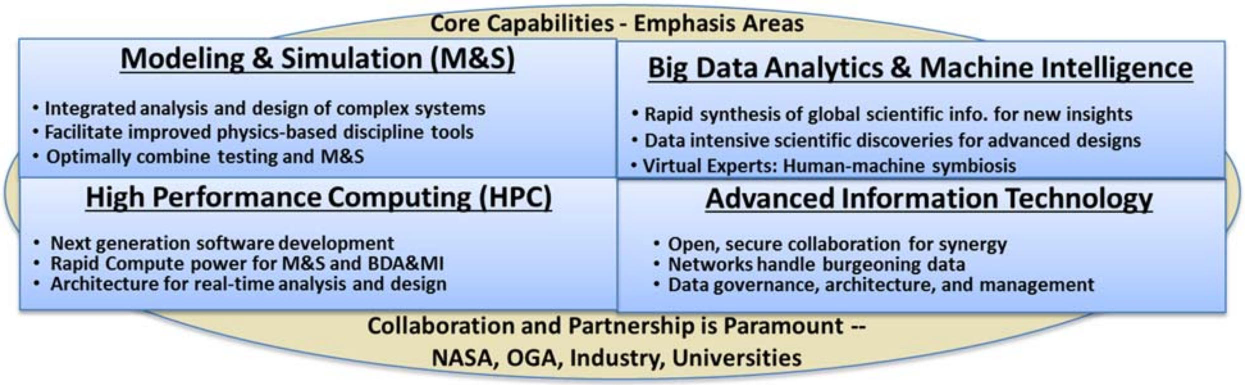

Self-Driving Cars
(DARPA, Google, TESLA & many others)
Robotic Warehouse
(UPS, Amazon & many others)
PERFORMANCE DEVELOPMENT
Race to ExaFlops Computing
(US, China & others)
200 sensors across the turbine generate 300 data points per second of performance and operation every hour.
Big Data & Internet Of Things at Work
(GE 90X Gas Turbine Engine)
Personal Self-Driving Drone
(Ehang, Testing in Mojave, CA)
Machine Intelligence Healthcare
(IBM & Google DeepMind)
Model-Based Engineering
(Lockheed Martin & many others)
Quantum Computing
(Google & NASA ARC)
Emergency & Disaster Management
Inter-Agency Collaboration
Critical Infrastructure Management
Information Management
Citizen Services & Immigration Control
Law Enforcement
Public Administration Services
nature
ALL SYSTEMS GO
coursera
KHAN ACADEMY
edX
LEARN ZILLION
WORKING in the CLOUD
Online Education
(Coursera, Caltech, MIT & many others)
Cloud-Based Office
(NOAA, DOE, NARA & many others)
Smart City
(Singapore, Rio & many others)
Machine beat human Go champion
(Google DeepMind)

Figure 1: Technology Progression across the Engineering Community.

CDT is a capability development and demonstration initiative strongly aligned with NASA strategy and program goals. This effort requires extensive collaborations between NASA, other government agencies, academia, and the private sector to leverage knowledge, tools and methods to realize this integrated capability for addressing NASA's aeronautics, space exploration, and science mission challenges. As a catalyst, CDT is envisioned to take on five overarching functions.

1. Leverage advancements from external to NASA organizations in all digital tools and technologies.
2. Utilize seed investments internal and external to NASA to develop and demonstrate individual and integrated capabilities.
3. Leverage current NASA program work and funds to demonstrate value/benefits to the mission.
4. Advocate to NASA mission directorates and influence capability advancements in alignment with current program goals and anticipated future needs.

5. Facilitate capability demonstrations that lead to and enable transformational solutions to NASA mission challenges.



Core Capabilities - Emphasis Areas

Modeling & Simulation (M&S)

- Integrated analysis and design of complex systems
- Facilitate improved physics-based discipline tools
- Optimally combine testing and M&S

Big Data Analytics & Machine Intelligence

- Rapid synthesis of global scientific info. for new insights
- Data intensive scientific discoveries for advanced designs
- Virtual Experts: Human-machine symbiosis

High Performance Computing (HPC)

- Next generation software development
- Rapid Compute power for M&S and BDA&MI
- Architecture for real-time analysis and design

Advanced Information Technology

- Open, secure collaboration for synergy
- Networks handle burgeoning data
- Data governance, architecture, and management

Collaboration and Partnership is Paramount --
NASA, OGA, Industry, Universities

Figure 2: Comprehensive Digital Transformation Core Areas.

LaRC’s approach encompasses development and demonstration of the four core capabilities individually and together with a synergistic integration of them to conduct discipline, multidiscipline and system-level demonstrations. Individual core capabilities have identified 3 focus areas each that need to be developed and strengthened providing benefit to missions and demonstrating the potential. System level demonstrations are being worked concurrently in a spiral development model, to lead up to transformational demonstrations that are aligned with the agency-level, mission directorate goals.

In this approach, this first tier is a development of a CDT capability baseline, both at individual capability areas and system integration levels. This requires an assessment of the current state of the tools, methods and compute infrastructure to identify gaps in executing an end-to-end analysis and design of current generation aerospace systems (e.g., Blended Wing Body as a fixed wing aircraft example). Existing discipline and multidiscipline tools utilizing available code integration methods will be integrated with open-architectures to demonstrate integrated analysis and design capability on current aircraft, space systems, and science instruments. It also requires evaluation of and securing leveraging opportunities to fill these gaps from within and outside of NASA, as well as identification of necessary NASA investments in critical missing areas in order to develop this capability. Identification and implementation of tools for combining analysis codes at the discipline, multidiscipline, and systems levels is also necessary. This integrated capability, with benefits of compute- and data-intensive capabilities from the four core areas, must be demonstrated where possible on benchmark candidate aircraft, spacecraft, and science instruments to elicit improvements achievable (e.g., design reliability, reduced testing needs, reduced design cycle time,…) through integration of analysis and design tools, machine intelligence, and HPC in advanced IT architectures. This short-term demonstration of benefits through the CDT approach is expected to position us for better advocacy and moving the Center to work in a collaborative culture between research, engineering, and systems analysis and design.

The second tier of this activity is advanced capability development and demonstration. The outcomes from this effort are aimed to enable transformational changes in the state of systems level analysis and design by opening the design space for aerospace systems beyond those that are currently possible with dramatic reductions in the design cycle time. Meeting this goal requires identification of gaps in tools and methods between the above baseline and the needed state within each discipline, systems level, and end-to-end integration tools with variable fidelity to capture the physics, define critical tests needed to validate the tools, quantify the uncertainties at the discipline and multidisciplinary levels and propagate them to the systems level to improve confidence in research results and systems design. Addressing these gaps must then be prioritized to determine where the most investment and advocacy must be focused, both within NASA and externally, to advance the capability for a future state. These efforts will be undertaken in conjunction with NASA and non-NASA efforts.

Big Data and Machine Intelligence Capabilities and Workshop Goals

Big Data Analytics and Machine Intelligence Capability (BDAMI) Goals and Projects

The Big data Analytics and Machine Intelligence capability team developed a vision and roadmap in 2014. The goal is to enable LaRC to discover “unknowns” and deliver previously unimaginable capabilities by applying transformational technologies as force multipliers for scientific and engineering discoveries and systems innovation and optimization. The vision is to have a “Virtual Expert” or “Virtual Research and Design Partner” enabling NASA employees to achieve greater scientific discoveries and rapid system design optimization.

Big Data Analytics & Machine Intelligence Capability Vision: Virtual Research and Design Partner

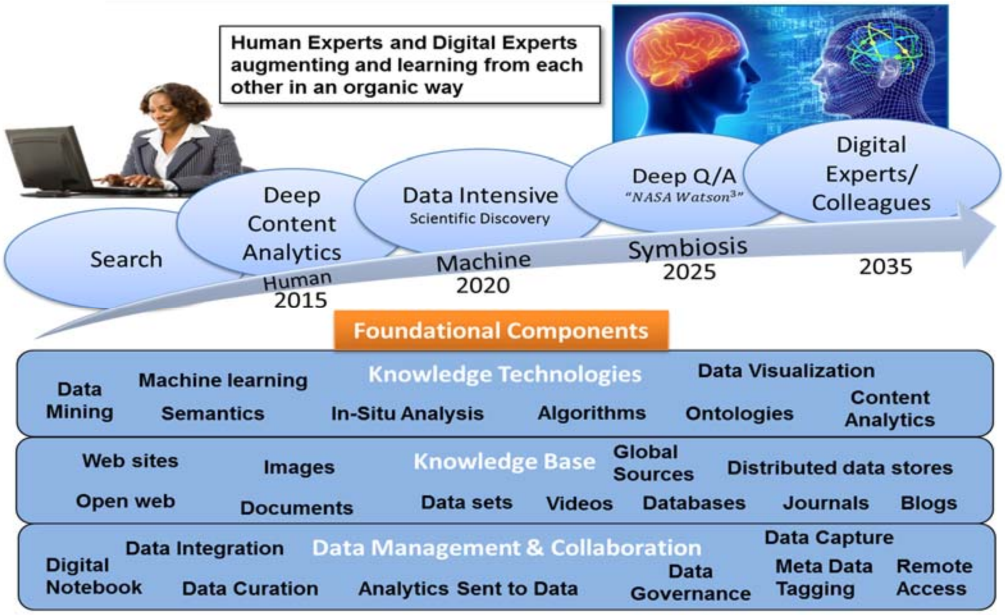


Figure 3: Big Data Analytics and Machine Intelligence Capability Vision.

The key capabilities of this Virtual Partner are:

- Able to quickly digest the latest research innovations and leverage insights.
- Deep analysis of world-wide multimedia scientific information and data enabling discovery of trends, unobvious relationships, and possible paths with evidence.
- Ability to ask engineering design-related questions and get reliable answers.
- Fusion and real-time analysis of data utilizing HPC to optimize modeling and simulation, ground, and flight testing.
- Accelerated ideation & design to increase research productivity.

The six strategic goals set are to:

1. Keep up with big data, deep analytics and machine intelligence technologies and capabilities, and advance LaRC knowledge and utilization and application of them.

2. Build a robust data intensive scientific discovery analytics capability and cognitive computing analytics to enable better science and engineering; Work collaboratively with subject matter experts.
3. Build a modular, robust and flexible big data and machine intelligence architecture and infrastructure to enable use by multiple disciplines/groups for heterogeneous data.
4. Ensure understanding, expertise and use of machine intelligence and artificial intelligence remains a long-term focus.
5. Proactively pursue, utilize and leverage partnerships and collaborations with other NASA Centers, universities, federal research organizations, and Information Technology industry.
6. Ensure buy-in at the grassroots level, resource availability with linkage to programs and investment prioritization for building and enhancing a big data, deep analytics and machine intelligence capability.

Over the last two plus years the BDAMI team has embarked on working a few selected projects/pilots to develop foundational expertise in applying big data analytics and machine learning technologies to Langley's aerospace domain data and information, working toward building a broader capability. This required investigation and experimentation, and working very collaboratively with the discipline subject matter experts (SMEs). Two of the major capability focus areas that are being worked are Data Intensive Scientific Discovery (DISD) and Deep Content Analytics (DCA). Both areas are equally important, and will eventually need to come together in the "Virtual Expert" vision, with data fusion and analyses of scientific and engineering data, scholarly literature, web, and multimedia. There are no readily available solutions for applying machine learning and big data techniques to the information and physics-based data sets in our aerospace domains. These challenges can only be researched, investigated, and developed into solutions through the ongoing collaboration of Big Data Team experts and SMEs, and by leveraging expertise and algorithms from universities, industry, and other NASA Centers. The team in collaboration with subject matter experts has been working on many pilots/projects and made progress in developing expertise and experience, and developing algorithms and solutions. A summary of each area is provided below.

Data Intensive Scientific Discovery (DISD): DISD's goal is to develop a machine learning and data mining capability for data analytics enabling our SMEs to save time, and to derive new insights and discoveries from experimental and computational data sets that would not otherwise be possible. The key pilots being worked are - 1) Rapid Anomaly Detection for the Non-Destructive Evaluation of Composite Materials: The goal is to develop algorithms that assist SMEs to rapidly detect different failure modes for better design of material compositions and structures; 2) Predict Flutter from Aeroelasticity Data: The goal is to develop algorithms that SMEs will use to accurately predict flutter precursors and flutter onset, in order to help optimize testing and to enhance vehicle design configurations; 3) Pilot Cognitive-State Assessment: The goal is to predict aircraft pilots' cognitive state using physiological data from flight simulations while performing tasks during various alertness states to help improve pilot training and safety; 4) Enhanced Launch Vehicle Designs: The goal is to develop a framework in which to apply machine learning algorithms for improved design of space launch vehicles using data from modeling and simulation codes/programs; 5) Atmospheric and Earth Science Data Analysis: The goal is to apply machine learning algorithms for rapid and enhanced data fusion and analyses, leading to better climate modeling, new insights and better science.

Deep knowledge/Content Analytics (DCA): The goal is to provide the technical community with easy-to-use analytics technologies that will allow them to quickly access and grasp internal and global knowledge without spending hours/months in reading, keep up with trends, identify experts, get answers with evidence, identify optimal paths for research, and derive actionable nuggets. Two key capabilities being in place are: Knowledge Assistants: The goal is to provide a seamless, easy-to-use capability for researchers/engineers to create 'Knowledge Assistants' for their specific domain content with deep analysis for insights, trends, and experts, using the Watson Content Analytics capability we built over the last three years. The key projects/corpus analyzed and being used by experts are – Uncertainty quantification; Aerospace vehicle design; Space radiation research/HRP; Human machine teaming; Model based Engineering. Additional activities include: Watson cognitive technologies Proof of Concepts/POC: Goal is the application of cognitive computing technologies to aerospace domains to understand and evaluate the use for our mission areas. Aerospace Innovation Advisor POC goal is to accelerate the pace of innovation and discovery, enabling researchers to identify optimal research paths. Topic being considered is Hybrid Electric propulsion; Watson Pilot Advisor POC goal is to help Pilots with decision making by providing a set of recommended actions for a given problem situation using the corpus of flight manuals, reports and data from flight deck.

Collaborations and Partnerships: The BDAMI team has established strong partnerships with Georgia Institute of Technology, MIT Computer Science and Artificial Intelligence Lab/CSAIL, Old Dominion University, NASA Ames Research Center Machine learning group**,** and IBM Watson Group. These partnerships are helping the team to leverage, learn and expand their expertise to develop innovative solutions for aerospace data and information challenges using data analytics and machine learning technologies.

For detailed information about BDAMI vision and progress, please refer to "Comprehensive Digital Transformation – "Big Data Analytics and Machine Intelligence Capability Strategy, Roadmap and Progress - NASA TM -2016-219361".e

Workshop Goals

The *Machine Learning Technologies and Their Applications for Scientific and Engineering Domains* workshop was hosted by the Big Data and Machine Intelligence group. It was held at NASA Langley Research Center August 16-18, 2016. The workshop aimed to bring together experts in the field of machine learning with experts from other scientific fields that could benefit from machine learning techniques. The goal of the workshop was to cultivate collaborative research opportunities to both further the development of machine learning techniques and advance other domains through leveraging opportunities.

The workshop provided presentations from 15 different experts and was attended by 300 participants over the course of three days. Each expert was invited to discuss their research topics for 45 minutes, and then were allotted dedicated break-out time for further, informal discussion. The following sections provide summaries of the workshop presentations as well as summaries of the informal discussion sessions. It is important to note that not all speakers were able to appear in person for their presentations; therefore, they were unable to conduct informal discussion sessions. The speakers that provided their talks through teleconference were Dr. Sebastian Pokutta, Dr. Barnabas Poczos, and Dr. Una-May O'Reilly.

The full agenda for the workshop as well as the full proceedings can be found at the following website: http://www.nianet.org/nasalarcmachinelearning2016/

Summaries of Workshop Presentations

Dr. Sebastian Pokutta

Presentation Summary

Dr. Pokutta, an Associate Professor in the Industrial and Systems Engineering department at the Georgia Institute of Technology, discussed the current state-of-the art in the field of Machine Learning and how he and his colleagues at Georgia Tech are working together to advance the field. Dr. Pokutta is involved in an inter-disciplinary machine learning research group at Georgia Tech (GT) that is comprised of approximately 80 faculty from across the campus. The GT machine Learning Center was founded to bring together engineers, data scientists, and statisticians for the purpose of developing new methods, encouraging the cross-fertilization of existing methods, and applying techniques to unique use cases.

Dr. Pokutta defined machine learning, in simple terms, as *gaining insight from data using computers*. It was stated that advances in computing, algorithms, and sensor technology have led to an accelerated development and progression of the machine learning field. The process of implementing machine learning ideals is iterative and can be summarized through three main steps: measure, learn, and optimize. For data analysis and learning, Dr. Pokutta stated that the availability of large amounts of data is no longer the problem, but entities need to ensure they are collecting the right data because that is what inhibits the effectiveness of data analytics techniques. For decision making and optimization concerns, black box solvers have been proven to be very efficient for real world problems, such as dispatching and scheduling in heavy industries.

The current trends and game changers in the field were stated to be the concept of cyber-physical systems, in-situ machine learning, and deep learning. Cyber-physical systems can be defined as those where the machine, sensors, and computing power are tightly integrated to provide the intended outcome, such as an autonomous vehicle. In-situ machine learning regards systems where high performance computing resources are placed near the sensors that are collecting the data, which enables the learning and processing to be performed in-situ. This enables very high performance with low energy and price points. Deep learning is the result of the convergence of larger datasets and faster computing capabilities. It is a field that was impossible to pursue five years ago, but now enables analysis like high-accuracy, real-time image recognition.

Dr. Pokutta concluded his lecture with two success stories. The first was predictive and prescriptive maintenance, where predictive implies the system is monitored and imminent failures are predicted. Prescriptive implies the inclusion of spare parts logistics and workforce scheduling for planning purposes. The use of machine learning to improve maintenance planning minimizes operational costs of assets and improves asset availability. It was mentioned that neural networks have been leveraged to categorize failure modes. The second success story described was real-time manufacturing optimization for problems where the design space is large and optimal parameter settings are desired. Dr. Pokutta described how surrogate models are being used to model the floating catalyst synthesis process for carbon nanotubes. The neural networks are then used to predict the effect of varying a parameter and determine the optimal settings for the parameters.

Dr. Ella Atkins

Presentation Summary

Dr. Ella Atkins from the University of Michigan discussed how advances in the machine learning and data analytics fields are aiding unmanned aircraft systems (UAS) situational awareness and risk mitigation. With machine learning techniques, sensor and database information can be processed quickly and used to change UAS situational awareness. The capabilities of UAS to process and store the data they collect ranges from inert packages where no data is collected or stored to storing data on-board and analyzing it in real-time to influence the mission. The current capabilities in the field are optimizing a mission when given a baseline flight plan, Guidance Navigation and Control system, vehicle dynamics, envelope constraints, energy dynamics, and constraints. Furthermore, Dr. Atkins stated that machine learning techniques can be used to capture and avoid different pilot or operator behaviors.

The first specific use case Dr. Atkins discussed was utilizing autonomy for safety through electronic geofencing. Electronic geofencing refers to invisible geographical boundaries in the airspace that UAS can be subjected during operation. There are different types of geofences, such as keep-in where UAS are restricted from leaving the boundaries and keep-out where UAS are restricted from entering an area defined by the boundary. The formulation and utilization of electronic geofences requires information on property lines, expected air and ground traffic, airspace rules, etc. However, even when they are put in place many questions still exist because standards have not been set and units have not been safety-certified. These questions include: *How do you stop the UAS if it violates the geofence? When control is returned to the operator if it is taken away?*

The second use case Dr. Atkins discussed was emergency landing planning for aircraft. Dr. Atkins explained that different data sources, such as roadways and traffic patterns, can be synthesized to prioritize potential landing sites for aircraft in distress. Cell phone data can also be used to characterize how heavily populated a roadway might be, but issues arise due to the dynamic nature of the data and inaccuracies.

Discussion Summary

Discussions with Dr. Atkins centered on policy for autonomous air vehicles and the distance they are allowed to operate within. With regards to autonomy in general, it was stated that in general people in the community are convinced that that pilots are well-trained and dependable; therefore, they are more open to autonomous cars than autonomous aircraft. Next, the topic of the operating distance was discussed. Dr. Atkins stated she is opposed to placing a 500-foot cap on autonomous aircraft. First, there are property ownership issues. Some feel that people that own a plot of land should control the airspace above it, up until a certain altitude. However, small UAS now make backyards navigable airspace and it, therefore, is the property of U.S. government. Secondly, 500 feet is too low to recover the UAS and its mission when something goes awry.

Some proposed solutions were discussed, such as increased segregation of the airspace and ownership agreements for rental or leasing of airspace over private property. The current state of UAS policy is a waiver process that can take up to six months. Congress asked the FAA to create a set of regulations for line-of-sight missions, which they did. However, there is still much more policy work to be done with regards to beyond-line-of-sight missions. Discussions concluded with the topic of the Automatic Dependent Surveillance Broadcast technology, ADS-B, which help UAS with communication and avoidance. ADS-B does not provide a form of conflict resolution, and its applicability going forward was questioned due to scalability concerns for high-density UAS scenarios.

Dr. Chris Codella

Presentation Summary

Dr. Chris Codella from IBM discussed cognitive computing and the advances made regarding IBM Watson over the past decade. Cognitive computing machines learn their behaviors through education and use forms of expression that are more natural for human interaction. In 2011, Watson was introduced to much of the public through a game of Jeopardy where it demonstrated an ability to handle more subtle context and cleverness beyond the challenges of understanding natural human language.

Dr. Codella explained that Watson parses and sorts text, generalizes and forms statistical aggregation of phrases, and then creates probabilistic hypotheses. Watson weighs evidence to synthesize answers and over time the accuracy and relevance of the answers are adjusted due to the evolution of the model from the training material Watson ingests. It was stated that the vision for Watson is to be used as an expert assistant that can bring a plethora of answers to an interpreted query, each with confidence levels.

Dr. Codella also provided insight into the different types of keyword searches. Keyword searches can be utilized to return documents containing the provided keywords where the judgment of the documents is based on importance. However, when using popular search engines like Google you can run into the issue of providing too many keywords, which limits the results you receive. Expert question and answer, on the other hand, utilizes natural language questions to produce possible answers and evidence. Expert Q&A can be rule-based or probabilistic-based, like Watson, which is better for complex combinatorial problems.

Cognitive computing can be applied in a variety of use cases, such as in aircraft and airlines operations to streamline decision making by providing several decision options. Dr. Codella acknowledged that in just a few years the basics of cognitive computing has been realized, and the next step is to work with structured or symbolic data.

Discussion Summary

Workshop participants from NASA's Armstrong Flight Research Center were interested in the subject/verb/object method Watson utilizes to interpret the meaning of a human language statement or command and were curious if there were other rules that Watson uses to extract information out of language. Dr. Codella stated this was only one way Watson extracts facts from language, and that several methods are used and the results are indexed for future searches and follow-up questions. The idea of symbolic reasoning was also mentioned, and it was stated that IBM is in the process of investigating it. Several researchers were also interested in the ability of Watson provide the virtual assistance to read and synthesize research papers. Dr. Codella stated that they are continuing along the process to develop these capabilities and are leveraging the increase in computing power. He also stated they are not currently looking into using Watson and Watson-like technology to debug computer code, but that it would be a useful application.

Dr. Codella was asked about Watson's ability to handle information communications, such as email messages or social media, and stated that there are ongoing experiments but the capability depends on the type of language. Emails are more easily handled because they tend to follow normal spoken language structures; however, social media language like Twitter requires more training because a unique vernacular is used. With regard to lack of data, Watson provides the ability to provide 'best guesses' to answer questions when insufficient data is present to definitively answer the question. Watson does this by default because it provides a ranking of answers with confidence levels for each. Dr. Codella recalled that it is important to remember there should always be a human in the loop for decision support, and that people make decisions in different manners when there is a lack of data. Watson provides what it considers to be the most relevant information to be considered, and the person in change will make a conclusion based on their own list of questions.

During the breakout session workshop attendees were interested in hearing more about how Watson handles misinformation and how it can be used for image processing. Dr. Codella stated that users must provide curated data to Watson, and possibly provide data flags regarding its usefulness. It was also stated that image processing is possible if a standard for comparison is provided. Examples that were discussed include identification of cancerous cells in medical scans, failure identification in a composite material, and the study of craters on the moon.

Dr. Tsengdar Lee

Presentation Summary

Dr. Tsengdar Lee, from NASA, discussed the advances in big data with respect to the earth science community. Dr. Lee stated that research he was presenting was based off of IBM Watson's ability to point to answers and data when provided a natural language question. In the earth science field, NASA has an unlimited amount of data from the past 40 years and is continuously collecting new data. However, the ability to utilize this data to gain knowledge and insight hinges on the ability to promoting the data, categorizing it, and useful data processing software. If these challenges can be overcome, it is desired to produce specific information about a topic or question similar to a google card, which provides information like keywords, instruments, datasets, variables, etc.

Dr. Lee discussed the desire to be able to create a database that can be easily queried, with the goal of providing a one-stop gateway that is able to proactively recommend personalized datasets, tools, algorithms, and experience. A two-step process was discussed that could lead to this capability, the creation of a science knowledge network construction and data processing workflow recommendations. The science knowledge network can employ either a fixed or open lexicon of entries: schema-based or schema-free. Schema-based inputs are fixed vocabulary whereas schema-free inputs are unstructured data, such as papers. Vocabulary and expertise is learned through ingestion of this input material. Therefore, the data must be labeled properly in order for the machine to assess and understand it on multiple levels.

The data can be modeled as social entities and social network analysis techniques can be applied to study software recommendations. Papers can be fed into the network, parsed based upon a set structure. The information from the paper can then be used to recommend other papers, identify research trends, identify usage of datasets, etc. Dr. Lee stated that they have experienced some difficulties with the use of natural language and the ability to capture information from some aspects of research papers, like figure and table captions. The next steps of this research are to continue the definition of the algorithm and work on training the system. Realization of these capabilities can be a concern due to worker displacement from the value of information that can be inferred from seemingly innocuous data.

Discussion Summary

Dr. Lee was questioned about the ability to utilize his work to identify small, less-known datasets that researchers could utilize. He stated that these datasets can be discovered without his machine through commonly used internet search engines, and the searchers can be tuned to identify new datasets once more information is understood. Dr. Lee also discussed the transferability of tools like IBM Watson and Google knowledge graphs to other problems. He stated that all processes start with some type of metadata tied to a given domain of knowledge, and when you learn you are building in inherent knowledge. Therefore, you continue to put in domain-related knowledge. However, creating a generic system is probably beyond current capabilities.

Workshop attendees were interested in hearing how Dr. Lee and his researchers were planning on handling ITAR and proprietary datasets. Dr. Lee stated they are not currently using these types of data, but have had discussions with companies like Google and Microsoft. With current datasets, researchers start discovering information and making inferences when they are able to 'connect-the-dots', which could be a dangerous thing when dealing with a sensitive or private application. At a policy/security level, there are a lot of issues that would need to be worked out and they are aware of the challenges.

Dr. Barnabas Poczos

Presentation Summary

Dr. Barnabas Poczos, of Carnegie Melon University, discussed how the machine learning field is being leveraged for design optimization in cosmology, neuroscience, and drug discovery. Dr. Posczos stated that a scientific assistant can be created through machine learning a statistical analysis techniques to automate discoveries, conduct experiments, and perform analyses. Traditional machine learning starts with data observations, the creation of a feature vector, classification, and then training. However, distributional data exists and can be used for training. This leads to a set of feature vectors, not just a single vector, and simply taking the average results in the loss of information.

One example where machine learning can be used not just on data directly, but on distributions of data, is image recognition. Images are broken apart and each piece can be represented by a set of feature vectors. This information can then be used to detect anomalies in the images. Dr. Poczos also discussed several other applications for the techniques he uses. In the field of cosmology, the machine learning techniques can be used to estimate the dynamical mass of galaxy clusters. They can also help identify interesting or anomalous galaxy clusters by finding a set of feature functions for a galaxy. Furthermore, they can help identify the true parameters of the universe by providing the algorithm with a distribution of particles and then predicting the parameter settings of the simulated universe.

Dr. Poczos also provided an enumeration of several other use cases where the techniques can be leveraged. He explained how they can be used in drug discovery to model the drug characteristics through regressions and surrogates, which then enables a rapid search through the different drug parameters. In the field of neuroimaging, they can decode thoughts from brain scans. They are also being used in agriculture to recommend crossbreeding and experiments for corn and other crops by using sensors to scan plants and predict how future plants will grow.

Discussion Summary

Dr. Poczos was questioned about the computational expense of the classifier algorithm he developed. He stated it was not too expensive. The manner in which it operates is to calculate the distances of all points in consideration; however, if the dimension becomes too large and the computational expense comes into question, a data structure can be used to remedy the problem. Dr. Poczos also mentioned that cross-validation may be desired and that different alphas work better in different applications.

Dr. Lyle Long

Presentation Summary

Dr. Lyle Long, a professor at Penn State, provided a talk about the advances in artificial intelligence technology and how it compares to the capabilities of the human brain. Dr. Long stated that the field of artificial intelligence has been oversold for the past 60 years. Now, however, supercomputers exist that exceed human computing power with respect to memory and speed and self-aware systems will most likely be possible in the near future. There are two common approaches to intelligent systems: connectionism and symbolic. An example of connectionism are artificial neural networks and an example of symbolic are cognitive architectures. Cognitive architectures are rule-based systems with some learning. The learning curve for cognitive architectures is steep for users and it is difficult to scale them to a human-level. For example, hand writing rules are not scalable and learning is required.

The human brain has a massive amount of sensors fed into it and outputs things like attention, memory, cognition, and motor control. The brain is a complex, parallel neural network system. Artificial intelligence would need to mimic the massively parallel neural network that is the brain. Scientists could model the brain one neuron at a time, but equations are expensive. Dr. Long discussed the Hodgkin-Huxley equations for modeling the brain. These equations show voltage across one neuron and are ordinary differential equations (ODE). It takes a coupled set of four ODEs to represent a single neuron, so to model the human brain in this manner one would need to model and solve 400 billion ODEs where each is coupled to approximately 1000 other ODEs. Dr. Long provided an explanation of an efficient, flexible object oriented code he created for calculating the Hodgkin-Huxley equations.

The end goal of this research effort is to model a child system that will grow and learn over time. However, unlike a human, scientists will be able to copy the machine and create replicates. The copies will not be required to go through the learning process again. Dr. Long did acknowledge one problem that the machines will face, which is the catastrophic forgetting problem where they tend to forget old knowledge that was previously learned as new knowledge is acquired. He also acknowledged that there is still a need for efficient, large-scale learning algorithms.

Discussion Summary

During breakout discussions Dr. Long commented on the ability of a machine to mimic the child brain. It was stated that a child has to be taught constraints and then eventually laws, and it would be difficult to instill moral constraints in an emulation of a child in a provable way. The discussion topic then changed to the concern in the community that machine learning techniques could lead to the elimination of jobs. Dr. Long stated that as civilization progresses, robots will be provided basic needs like food. It is a problem that workers are also consumers in the economy, and if they are put of work they will not be buying and using as many products. Therefore, this is a problem that may be solved through economic policy surrounding the way people are taxed and supported.

Some workshop participants were particularly interested in the ability of machine learning techniques, specifically the artificial Neural Network, to mimic the behavior of a partial differential equation (PDE) that describes the physics of a fluid. Dr. Long stated that he feels a Neural Network can be used to solve a given PDE without the need for a specially designed neuron. Dr. Long also fielded some questions regarding the hardware and software he was utilizing on his research. For hardware, Dr. Long utilizes National Science Foundation supercomputers and commodity processes. For the software development, he stated that C++ was chosen based on its efficiency and other attractive properties, such as its dynamic memory and the ability to model neurons and synopses as objects. Furthermore, the code itself was made to be very general but tailorable so major changes did not have to be made for different sets of problems. He wanted something general so the code didn't have to be changed a lot. It can be used on more problems, very tailorable.

Dr. Long was questioned if supervisory mechanisms were implemented to manage the connections between neurons, and stated that his research has not gotten that far. It currently reads in topology of the network but has the capability to drop neurons and/or synopses. Lastly, Dr. Long was asked about the traction this type of research has gained in the medical community, specifically with regards to simulating the brain of an animal from data gathered from medical imaging. He stated that there has been not been that much traction in the community, but some interest has been shown from the Allen Brain Institute.

Dr. Una-May O'Reilly

Presentation Summary

Dr. Una-May O'Reilly, from the Massachusetts Institute of Technology, discussed the field of data-driven artificial intelligence. The goal of her research is to create intelligent computation to support intelligent agents and understand complex systems. Information from past examples can be used to predict future states. Currently they are studying several topics, such as forecasting physiological states, behaviors of students as they learn with open online courses, and tax monitoring. Data is now collected is a variety of ways, such as from web browsing, phones, financial institutions, and vehicles, to answer predictive problems.

The advancement of cloud computing provides a large amount of available data to scientists. The cloud concept is a cost-saving, virtualization device that provides the ability to encapsulate hardware configurations and software. Clouds support resource elasticity and efficient resource budgets. Dr. O'Reilly discussed the need for machine learning that will handle the high volume of data provided by the cloud. Machine learning techniques must be scaled to execute in cloud computing environments due to the cloud offering the computational resources and storing large amounts of data.

Dr. O'Reilly described supervised machine learning as modeling and/or regression, where a set of explanatory (aka independent) variables are used by a function that predicts an output dependent upon variable values. Data is divided into training data and testing data. The training data is used to propose candidate models and the testing data is used to validate a trained model and report its accuracy. The selected model (among training candidates) is then used to make output predictions in new but similar situations. Dr. O'Reilly discussed the differences between generalized linear models and non-linear models and stated the one of the current issues in machine learning is scalability. Most existing algorithms are mature enough to be used off the shelf, but researchers must determine which algorithm is best for the problem at hand. Dr. O'Reilly stated that non-linear models are the most accurate, but not as fast and readable as linear models because it is hard to understand how the outcome is determined.

Dr. O'Reilly discussed how genetic programming can be used to scale machine learning techniques. She introduced a project she has been working on which is an open source genetic programming tool that addresses the algorithm scaling problem for large datasets. In FLEXGP, data and learner replicates are factored across the cloud and return resulting models. These models are filtered with respect to their accuracy, and then they are fused into an ensemble. For model fusion, some data is reserved so it can be used to come up with the ensemble coefficients for each model within the fusion set. Overall, cloud computing is a large resource and FLEXGP is helping with the creation of scalable, competent machine learning algorithms that execute on the cloud.

Discussion Summary

Dr. O'Reilly was questioned about the datasets sizes she has utilized and the corresponding cloud space required for the datasets. She stated that the largest datasets she has worked is data on credit card transactions. The size of the credit card data was on the order of terabytes, but was reduced to gigabytes when the feature vectors were created for the dataset. Before solving the complete problem defined by the entirety of the credit card transaction data, Dr. O'Reilly and her team sampled the dataset and used this subset of the data to explore the problem on a smaller scale.

Dr. O'Reilly also discussed data issues they have faced. She stated that they have faced similar problems that other researchers have mentioned, which the lack of clear labeling of training data for the machine learning algorithms they are developing. Not having clear labels is a problem because supervised machine learning algorithms depend on them. The solution is to get others to help build the labels, such as through a crowd-sourcing endeavor. Otherwise, you are in a situation where unsupervised algorithms must be used, such as clustering.

Dr. Matthias Scheutz

Presentation Summary

Dr. Matthias Scheutz, of Tufts University, discussed the development of robots he is researching that utilize one-shot learning techniques with natural language dialogues. Dr. Scheutz began his talk by discussing the difference in closed world tasks and open world tasks. With closed world tasks all possible environmental, task-based, human and interaction oriented aspects are known at the beginning of the task. For open world tasks, some aspects of the task are unknown at the beginning and during the task execution. For these types of tasks, robotic architectures would be required to deal with unknowns and acquire knowledge on the fly. Dr. Scheutz provided an example, via a video, of a potential search and rescue robot. In the example the robot had to take direction via a natural language statement and use that information to determine what it was looking for and how to retrieve it.

Data driven training for robots assumes that datasets are always available and training is not time critical. However, in open-world scenarios time can be critical and new knowledge use is important. Therefore, data-driven approaches must be supplemented with knowledge-drive techniques like one-shot learning. These techniques are not data-driven; rather, they are driven by task-based natural language dialogues. It can be accomplished through object learning, action learning, or rule learning. Dr. Scheutz provided a series of video examples that demonstrated different knowledge-driven learning approaches with a variety of robots. In one example it was observed that a robot could recognize an object, a box with a red cross on it, through object learning. This required the robot to have a three dimensional vision system that can recognize color, texture, and curvature. In another example, it was observed how a robot can recognize but does not know how to complete a requested task, ask for assistance, and learn how to do the task by watching a human demonstrate.

Open-world tasks require significant changes to algorithms in all components of robotic architecture. Scaling of the techniques is not yet complete, but will get there in the future. Soon robots will be able to manipulate objects based on observations, perform complex activities from narration, and learn and perform tasks in mixed human robot teams. Dr. Scheutz stated that his future work will be to include different languages, a broader range of sentence structures, and simulating the generalizations of an object. He stated that true autonomy is the ability to be creative when no known solution exists and the field is not there yet.

Discussion Summary

Dr. Scheutz provided further discussion on how he validates that his robots are properly interpreting and learning rules. Bounds are set on each rule and its implications, with upper bound defining what is considered correct and the lower bound defining what is considered incorrect. He discussed the concept of modeling the trust you have in the person teaching the rule, and that one might not be confident that a rule is true in all potential scenarios. Meta Modeling allows for the simulation of behaviors when the robot is offline. This may be a good substitute during open world tasks when there is a lack of data, but it may not be ideal if data availability is not an issue.

Dr. Scheutz discussed the amount of time it takes his team to develop the robots and build new capabilities into them. Overall, it took 10-15 years to develop the robots' architecture. However, it takes less time to build in new capabilities to complete a new task now that the architecture is in place. The time to complete a task depends on what sensors or features are necessary for the task. Vision sensor and task planning are computationally expensive but speech is not. Dr. Scheutz also stated that the robots are programmed to provide a reason to the instructor when it does not complete a task. The robots are not programmed to optimize how they meet the goals or performs the task, but are only focused on simply completing the task without any drive to perfect its behavior. The robots simply do as they are told and do not learn the purpose of their action.

Dr. Scheutz also fielded several questions centered on the use of natural language commands. First, he discussed the ability to utilize a language other than English. It was stated that none of the current research is necessarily language-specific, and as long as the robot can parse the commands provided, any language can be utilized. However, it was acknowledged that the language parsing in a language other than English would be of a different difficulty. Next Dr. Scheutz addressed the rigidity required in natural language commands. All of the natural language systems are very brittle, so there are specific ways one must speak to the robot to achieve the desired performance. Certain variations can be handled, but it is important to realize the main focus of the research is not the broad purpose but trying to get the parts to work together. Lastly, Dr. Scheutz was questioned about what happens after natural language, when simulation methods are utilized to generalize actions. One big unresolved problem in generalization is how much common sense knowledge is required to recognize the dimensions. There will not be a generalized answer, but you can consider what the robot already knows about something.

Dr. Dimitri Mavris

Presentation Summary

Dr. Dimitri Mavris, from the Georgia Institute of Technology, discussed the use of machine learning techniques to aerospace and complex system design problems. Dr. Mavris explained that past analysis tools were dependent on historical data that are no longer relevant to new, advanced systems that are being designed and analyzed. Models are moving away from historical data based models and towards physics-based models of varying fidelity levels. Furthermore, systems of systems analyses are more complex and can require the integration of several different types of analysis tools. Machine learning techniques can now be utilized to enable integration of legacy tools through surrogate models and provide a means to perform multi-disciplinary analysis for system design and system of system problems. Furthermore, they also enable multi-objective optimizations and the ability to do dynamic, parametric trade-offs. Dr. Mavris explained that ongoing research at Georgia Tech is utilizing machine learning techniques for performance forecasting and uncertainty quantification and management.

Dr. Mavris discussed several example applications. One application discussed was a set of technology and vehicle performance assessments Georgia Tech (GT) performed for NASA's Environmentally Responsible Aviation (ERA) program. The ERA program focused on maturing technologies that will help the aerospace industry meet a future generation (N+2) environmental goals. Georgia Tech created a decision-support dashboard that included performance assessments at the aircraft, airport, and fleet level. The number of technologies and vehicle concepts that existed to analyze created a combinatorial problem, which was amplified by the runtime of the assessment tools. Artificial Neural Networks were utilized to speed up the assessments, which enabled a Monte Carlo analysis of all potential technology portfolios and an uncertainty quantification analysis.

Dr. Mavris also provided an example of a dynamic, data-driven application, the GT Smart Campus initiative. A large amount of raw data that described the utility consumption of each building on the Georgia Tech campus existed, and researchers were able to utilize it to model each building and the connectivity of all buildings. The first step of the project was to provide situational awareness to decision makers on campus regarding the consumption of each building and each type of utility, such as chilled water and electricity. Various data and visual analytic tools were utilized to create an interactive dashboard to display this information. The next step was to incorporate machine learning techniques to predict future performance of buildings given different weather and construction scenarios. This resulted in the creation of a unique signature for each building on the campus. The Smart Campus initiative helps illustrate how machine learning can help enable virtual experimentation by creating a virtual test bed.

Discussion Summary

Dr. Mavris provided further discussion on the GT Smart Campus initiative. He explained that there were multiple phases of the research. The first phase was data mining. This data-driven process involved gathering data from several external databases, cleaning and organizing the data to ensure there were clear labels for the types of utilities and buildings involved, and then visualizing the data. The second phase is to provide forecasts of the campus performance given different potential scenarios. Lastly, the team will focus on optimizing the campus. Currently each building is optimized on an individual basis, with no consideration of the global optimal.

Dr. Mavris next discussed how the methods and tools he presented has enabled unique collaboration opportunities with industry and government partners. The Meta modeling techniques enables his researchers to utilize industry-developed codes while still protecting the proprietary nature of them. Furthermore, the techniques have also enabled important uncertainty quantification and management analyses. It has been studied how the results of these analyses can be used for experimentation planning in a technology or system development program. Dr. Mavris explained that when technologists are planning experiments to advance the maturity of a given entity, they must first go through a 'thought experiment' and then go through the actual experimental design. These two concepts are different, but are both important for ensuring the right experiments are planned and performed without wasting resources. After experiments are performed, the data that is collected can be utilized to improve the technology performance models and quantify the amount of uncertainty reduction that was achieved. These key machine learning concepts are important as the industry moves towards cyber-physical systems.

Dr. Karthik Duraisamy

Presentation Summary

Dr. Karthik Duraisamy, of the University of Michigan, provided a talk about data-driven modeling. While the framework is general to be applicable to a wide range of complex physics problems, specific discussion was provided in the context of turbulent flows. Turbulence involves multi-scale phenomena and practical simulations requires coarse-graining the Navier-Stokes equations. Specifically, the Reynolds stress tensor has to be modeled. It was acknowledged that existing models are lacking in accuracy for many complex flows and they are calibrated on limited data. It is also stated that the balance between the terms is what matters most for model development, not the accuracy of the individual terms themselves.

The approach provided by Dr. Duraisamy is to provide large scale data driven improvements for turbulence models by using data from complex flows and embed the algorithm into the solver. This can provide non-parametric improvements and does not replace knowledge but augments it. The key innovation in this work is the use of inverse modeling to extract discrepancies between the model and the data and machine learning to transform these discrepancies into functional forms. These machine learned functional forms are embedded into a predictive solver. The concepts are tested by giving a deficient model to recover a known model with machine learning, and to observe whether the partial differential equation (PDE) solver can recover the true result. It is successful when the model, solver, and loss function are properly scaled.

The framework presented by Dr. Duraisamy have been tested on turbulence flow and separation on airfoils. Data was taken from open literature for one airfoil at two different Reynolds numbers and a range of angle of attack. It was observed that just by using lift data, this model provide improved predictions for separation on new airfoils at different Reynolds numbers. It was also observed that prediction accuracy was unaltered in low angles of attack when the baseline model was accurate enough. This machine learned model was integrated into a commercial flow solver and successfully validated, demonstrating portability of the technology. In the future, Dr. Duraisamy and his team to allow users to train their own models with their complex datasets for other use cases. Dr. Duraisamy is directing the Center for Data-driven Computational Physics at the University of Michigan, where a range of other applications including materials physics, cosmology and climate science are being explored with the same paradigm, as are novel hardware/software paradigms that enable these types of applications.

Discussion Summary

Dr. Duraisamy provided further discussion on turbulence modeling for flows over airfoils. With regards to including more features in the model development, he stated it would be good model training but not optimal when using the model for prediction. When asked how one can be sure the model is giving the right answers for the right reasons, he stated that more data can help answer these questions. For instance, he stated that pressure data is sufficient in many problems, but when shocks are present skin friction was required. They are investigating unsteady problems, but application has thus far been restricted to periodic in time problems as the Bayesian field inference becomes too expensive otherwise.

Furthermore, it was mentioned that bifurcating problems may be difficult to predict. It depends on which phenomena is important, but it indicates the physics is not completely captured or understood. Adding more dimensions to the model may remedy the bifurcation issue altogether.

Dr. Heng Xiao

Presentation Summary

Dr. Heng Xiao provided a discussion about the use of physics-informed machine learning techniques for predictive turbulence models. He explained that Reynolds-averaged Navier Stokes (RANS) models for turbulence modeling purposes are still used in the field of turbulence modeling despite low system confidence and major discrepancies. Researchers do not currently understand the physics of the problem well enough and/or cannot afford computational costs to resolve the physics; therefore, they utilize data to accompany low fidelity models.

Dr. Xiao explained that machine learning techniques can be applied to offline to reduce the discrepancy of low-fidelity turbulence models. However, machine learning techniques and algorithms developed for business applications cannot be used directly because they may violate the physics laws that define the problem. The laws of physics provide hard constraints, whereas popular machine learning applications have soft constraints. Therefore, the physics must be tied into the development of the machine learning algorithms to utilize the concepts on turbulence modeling.

There are many algorithms inspired by physical phenomena, such as simulated annealing, particle swarm, and genetic algorithms. However, Dr. Xiao emphasized that the objective for turbulence modeling is different because it is desired to use machine learning to solve physical problems and not to use physics as inspiration for algorithm development. The goal is to correct and improve existing, low fidelity turbulence models with machine learning to lead to more insight on the phenomena itself.

Dr. Xiao described a case study, a RANS-based turbulence model. RANS models have poor performance in flows with separation, mean pressure gradient, curvature, or swirling; they can be improved by quantifying and reducing the model discrepancy, which is caused by the Reynolds stress closure. Dr. Xiao explained that the training data for the machine learning algorithms is DNS data of Reynolds stress for elementary flows, and the type of algorithm selected was a Random Forest regression. The Random Forests regression was selected because it is suitable for high-dimension feature space and robust in tolerating unimportant features.

Discussion Summary

During follow-up discussion Dr. Xiao was questioned about his collaboration with experts in the machine learning experts. He stated that his work started by educating himself and colleagues without collaborating with experts. However, he acknowledged there is a lot of potential for future collaborations with machine learning experts.

Furthermore, Dr. Xiao discussed the ability of the machine learning techniques to help researchers understand the complex phenomena in addition to just modeling it. One takeaway from the current research is they have identified certain features in the phenomena that are important. Next, they will focus on what can be done to incorporate this new knowledge into the existing models and potentially form a new model to improve predictive capabilities. The objective is not just a good prediction but to close the feedback loop and improve the features and models. He stated there is a small community of researchers to investigate data-driven modeling, and that the researchers often look at small pieces of equations they are trying to remedy and not the entire thing.

A specific use case was mentioned by climate science researchers attending the workshop. Dr. Xiao was questioned about any ideas he had about cloud boundaries and the applicability of his techniques. He stated that the general principle is clear, but specific physical knowledge is required to understand the problem. He suggested that they should not start from scratch, rather build from people in the turbulence modeling community.

Dr. Krishna Rajan

Presentation Summary

Dr. Krishna Rajan, of SUNY Buffalo, provided a talk about materials informatics and the ways the field of materials science is changing. He discussed that the materials science field is entering into a new paradigm where there is computational materials science, experimental materials science, and big data. Traditionally, the field of material science does not have a lot of data, which led Dr. Rajan to the fields of machine learning and data analytics. The existing data needs to be pieced together and the gaps need to be bridged with informatics. An example from history of the successful use of data mining is the creation of the periodic table and the ability to predict new elements from it.

Classification is a big task in materials science, and classification systems do not allow for significant discoveries. It takes a long time to determine the characteristics of a material that is only slightly different from an existing material. The lack of data causes this to be an inverse design problem where machine learning can be leveraged to fill in the gaps. Furthermore, Dr. Rajan discussed the material genome concepts. There is a need to determine the important variables that define a material, classify the behavior based upon the variable settings, and then make quantitative predictions. There are many issues with classification of materials, including the ability to link together different levels and scales of information and the misidentification of outliers due to improper sorting.

Dr. Rajan stated that a lot of uncertainty exists in a material's behavior because a measured property of a material is a function of a lot of other things. Furthermore, it can take a long time to discover a cluster of materials because of sparse data. There is a need to apply theory, find important characteristics of materials, and cluster the results. The results can be clustered with either nested clustering or decision boundaries. Machine learning helps find classifiers, discover catalyst chemistries, new structure maps from bond characteristics and angles, prediction of new materials, and finding substitutions for different alloys.

Discussion Summary

Dr. Rajan responded to questions about how the community was responding to the use of predictive modeling for structural analyses. He stated they have been able to predict new chemistries for materials people haven't yet observed. It was also acknowledged that some materials are considered outliers, meaning they appear to not follow the normal performance trends for a given material composition. Dr. Rajan stated, however, that his research has been able to show that the outliers are not outliers after all. Instead, they need to project the information onto the right manifold. These new trends, where outliers are not observed, have not been seen because the wrong projections were being made onto the wrong manifold.

Dr. Rajan discussed how his team determines when a new material, or new properties and trends, have been discovered. He stated that literature that has been previously ignored or not utilized can yield new observations and discoveries. However, researchers must be cautious and conduct the assessments multiple times. A ceramic material example was provided, and Dr. Rajan stated that it was observed that a single defining parameter of the ceramics composition dominated the fracture mechanics.

Dr. Rajan also discussed how machine learning techniques can be used for different purposes in his field. First, it can be used as an instructive tool for exploration and can help identify regions of high change and the approaches utilized depends on the size of the data. Machine learning techniques can also be used to create virtual libraries and to find correlations between different parameters. He stated that machine learning can really be utilized in all aspects of the thorough process for designing and identifying new materials and their properties.

Dr. Jaime Carbonell

Presentation Summary

Dr. Jaime Carbonell, of Carnegie Mellon University, discussed machine learning and big data applications to aerospace. Dr. Carbonell began his talk by stating we are currently in the fourth industrial revolution with the emergence of mobile internet, smaller powerful sensors, artificial intelligence, and machine learning. Machine learning is a subset of artificial intelligence, and current trends are in deep learning, reinforcement learning, large margin methods, and graphical methods.

Supervised machine learning techniques require labeled data; however, labeled data volumes are significantly smaller than unlabeled data volumes. Some of Dr. Carbonell's work focuses on dealing with datasets with label sparsity, such as active learning, proactive learning, and multi-task learning. When there exists a limited number of labeled points, the question arises over how you select the next point to label. There are different ways to do this, such as select a point that is equal distance between the two groups along the decision line, select an outlier point to drastically alter the decision boundary, select a centroid point of a cluster, or a combination of any of these strategies. Classification error reduces as you increase the number of samples to the labeled set, but there is no right way to add samples to the labeled set.

Active learning assumes a single perfect source of information, but sometimes there are multiple sources of information that need to be combined. Furthermore, the different sources have different characteristics, such as their levels of certainty. Proactive learning is able to take into consideration these differences and attempts to jointly estimate the accuracy. The past accuracy of information sources can be taken into account, and as you probe the source more the variance will reduce. Over time, it is possible for the accuracy of information sources can change.

Dr. Carbonell has applied these methods to the F/A-18 through his work in the Boeing-Carnegie Mellon aerospace data analytics lab. Their research is helping with maintenance decision support to make it more reactive and improve overall flight readiness. They have been able to determine the minimal number of tests to perform on the aircraft and then determine the right mechanic for the job. They are now working on proactive, commercial aviation and trying to optimize the preventive maintenance. This requires getting access to and combining information from sensor data, the parts networks, and the aging information of the fleet.

Discussion Summary

Dr. Carbonell elaborated on the types of machine learning techniques he and his colleagues are utilizing for condition-based and predictive maintenance research being conducted with industry partners. While he could not divulge all details due to proprietary agreements, he did mention that they investigated transductive methods, specifically transductive support vector machines. After they were able to identify the appropriate support vectors for a given dataset, they utilized supervised machine learning methods and they are currently utilizing proactive learning techniques heavily. Dr. Carbonell also discussed the required data labeling work and how experts from his industry partner help with this. Many of the domain experts are tasked with making the aircrafts more predictable, reliable, and maintainable their knowledge is important to capture.

In the breakout discussions, Dr. Carbonell elaborated on using MRI brain scan data. He discussed two different datasets that had the same features, however he acknowledged that the sensors utilized can be different and labeling can be different due to doctor subjectivity. Dr. Carbonell also touched on the topic of autonomous flight and the type of sensors that would be required. He stated this type of problem would be a multi-modal, temporal environment with terabytes of data.

Dr. Vipin Kumar

Presentation Summary

Dr. Vipin Kumar, from the University of Minnesota, discussed opportunities and challenges with big data in climate science. There are a lot of challenges in analyzing the climate data, such as the high dimensionality, high amount of temporal variability, correlations in space and time, multi-resolution and multi-scale data, large amounts of noise and missing data, spatial heterogeneity, lack of a ground truth, and class imbalance due to changes and anomalies being rare events.

Dr. Kumar illustrated some of these challenges in the context of two specific climate science examples, with the first being global mapping of forest fires. Forest fire monitoring is important for evaluating their climate change impact. The current state of the art for forest fire monitoring is the NASA MCD646A1 product, which is available monthly for every location on the globe. But this product is known to have poor performance in the tropical forests, as it tends to miss most of the fires in these regions. It is challenging to build classification models to detect forest fires that can work well globally the absence of ground truth in many parts of the world including the tropics, and large variations in geography, seasons, land class, and time. The classification performance varies greatly depending on the location on the globe, with the best performance occurring in North America. Dr. Kumar acknowledged that for areas outside of North America, especially in tropical forest regions, scientists must assume they are working with a situation with no data labels. MODIS active fire product that captures thermal anomalies can be used to identify a forest fire because a fire will cause an increase in surface temperature. However, anomalies can exist due to small fires that can lead to false positives, and many fires are missed due to reasons such as smoke and clouds, leading to false negatives. Dr. Kumar presented a novel supervised learning methodology that can work with imperfect labels (such as those available from MODIS active fire product) and showed that it can be used to build a very high quality map of forest fires in the tropics.

The second example Dr. Kumar presented was the global mapping of inland surface water dynamics. For this example, there are a lot of samples to work with, but the problem is hard due to heterogeneity (there are many ways water and land can appear in the spectral space at varying times and locations) and due to large amount of noise and missing data. Two different set of techniques were developed to address these challenges: ensemble learning and physics guided labeling. Ensemble learning techniques are used to address the heterogeneity and physics guided labeling was used to correct any errors made by the classification model. Dr. Kumar gave a demonstration of a system to monitor the dynamics of water globally (using MODIS data) that was produced using this methodology. Results produced by this system are available in a web viewer and allow investigation of the areas where water bodies are growing or shrinking globally.

Discussion Summary

Dr. Kumar discussed the applicability of transfer learning techniques to the examples he discussed during his talk. He stated that a lot of machine learning methodologies, including multi task learning, are applicable for addressing the complex nature of climate data. The forest fire prediction model works poorly in tropical regions because it needs to be tuned for false positives, but the tuning parameters that work well for many parts of the world including North America are not appropriate for the tropics. He stated there is also a need to build model for each different land cover. Multi-task learning can aid this problem by building different models each different land classes, but they have not gotten to this point yet in the research.

Dr. Kumar also provided further discussion on the surface hydrology example. In the past NASA flew special missions to map the surface water and provide a labeled dataset for the globe. A lot information came from those missions, and it was used to build the current model, as it provided training samples for the classifier. The next missions will provide similar type of data at higher resolution and help capture the dynamics and evolution better. It was also stated physics-based labeling will still be used to address missing data and poor classification output due to noise and heterogeneity.

Dr. Raju Vatsavai

Presentation Summary

Dr. Raju Vatsavai, of North Carolina State University, provided a discussion on a machine learning framework for infrastructure monitoring. NASA and other agencies have a large number of satellites monitoring and collecting images of the earth. The images range in resolution, and the type of features that can be identified in a given areas depend on the resolution. Dr. Vatsavai acknowledged that spatial-temporal data mining needs proper use of resolution-per-purpose of monitoring.

Image mapping can be used for several unique reasons, such as agriculture monitoring for resource conservation and social monitoring of cities. Agriculture accounts for 70% of total global freshwater withdrawals, so monitoring it can be very important for conservation efforts. Dr. Vatsavai stated one potential use is prediction of soil moisture for irrigation planning purposes. Imaging also helps identify impacts of natural disasters, such as flooded regions and earthquake damage. Infrastructure monitoring for social mapping includes slum mapping. Dr. Vatsavai stated that imaging can be used to classify new constructions that are either formal or informal settlements, which can then be used to classify the socio-economic composition of the neighborhood.

Dr. Vatsavai discussed the big data problem that occurs with high resolution imaging: a very high resolution image can provide 600 trillion pixels to classify and determine if changes have occurred. Pixel change detection in high resolution images is done by dividing an area into a grid, modeling each grid as a statistical distribution, and monitoring clusters of features in the grid for changes. With a higher resolution and more pixels more information if available for classification. However, with certain things like a body of water, an increase in the resolution can cause classification mistakes due to plants growing on water surface. Therefore, increasing resolution does not necessarily work to increase classification accuracy with all applications and situations where you have prior knowledge on what you are searching for should be utilized. The object in the end should be to identify the investment that is good enough because the detailed data is expensive.

Discussion Summary

The topic of hardware was discussed in the breakout session shared by Dr. Kumar and Dr. Vatsavai. It was stated that a lot of clusters are used, with shared memory and distributed memory. The NASA server exchange has provided a quicker way for researchers to obtain data from NASA. Furthermore, researchers now have the ability to upload their algorithm to the cloud, where they will be allocated nodes to run their analyses.

The topic of classification was discussed, with questions arising about when classification became a part of machine learning. The discussion on this topic focused on how the field of statistics as a whole has undergone significant changes as computers have been developed and continuously improved over the past several decades. Computer scientists have overtaken the field of statistics, and it was stated that data science programs must be joint with statistics going forward.

Lastly, climate science researchers questioned Dr. Vatsavai and Dr. Kumar if they had considered clouds as a potential use case, such as utilizing observation data from satellites to predict and classify clouds. Dr. Vatsavai stated that he has worked with clouds, specifically helping when cloud classification provided cloud imaging information.

Key Workshop Themes

Imperfect Data

The problem of imperfect data arose during several of the presentations made by experts at the workshop. There are several different ways a dataset can be categorized as 'imperfect', including a lack of data and mislabeled data. The use of imperfect data for machine learning purposes, such as classification, can lead to several hardships. Dr. Rajan acknowledged that the issue in the field of materials science is a lack of data that can be used for classification purposes of existing and new materials, and researchers in the field must work on creating new data that can be used for material classification. In situations where data exists but is not properly labeled, it is difficult to train a classification algorithm and determine which data points should be classified next. Dr. Carbonell acknowledged that a limited number of labeled points can make it difficult to properly draw decision boundaries for classifiers. Furthermore, as new data points are classified the decision boundaries can change drastically.

When these situations arise, there are ways to improve the data to enable a better classification. One example as provided by Dr. Kumar. For his research, observable instances of forest fires is used as training data for the classification algorithms; however, areas on the earth where a fire is less likely to be observed, such as in the tropical forest regions, the data is very sparse and they must assume little or no labels exist. Dr. Kumar explained that his algorithm prediction was improved by introducing other types of data to the classifier, such as surface temperature readings. Dr. Carbonell also stated that he and his colleagues at Carnegie Mellon are improving their mislabeled datasets. Instead of supplementing their existing data with other types of data, they are utilizing subject matter experts to help them create and improve data labels. The new data labels can then be incorporated into the classifiers.

Autonomy and Robotics

One topic that was prevalent in several of the presentations during the workshop was autonomy and robotics. Machine learning techniques are huge enablers for autonomous vehicles, such as unmanned aircraft systems, and robotics in general. The topic was first discussed by Dr. Ella Atkins. She provided insight into how sensor and database information is being quickly processed by unmanned aircraft through the utilization of machine learning techniques for the purpose of providing better situational awareness to UAS operators. Dr. Atkins specifically acknowledged the field of geofencing and how it can be used to keep autonomous vehicles within accepted boundaries in the airspace. Furthermore, the idea of geofencing can be integrated into the command and control algorithms that define the vehicles to help define the concept of operations for loss of signal.

Dr. Matthias Scheutz provided a very popular presentation on the robotics work he and his colleagues at Tufts University are researching. His research focuses on utilizing one-shot learning techniques to teach different types of robots to perform a variety of tasks. Dr. Scheutz demonstrated the current results of his research through a series of videos where his different learning techniques, such as the use of natural language commands, were displayed along with the outcomes of the robots. The videos demonstrated that the machine learning techniques paired with the use of natural language, and other types of learning inputs, that the robots Dr. Scheutz and his team are developing are becoming more 'intelligent' throughout the research.

Furthermore, advances made in the field of autonomous vehicles was mentioned in the context of advancements made in computing capabilities over the past couple of decades. This was acknowledged by almost all workshop participants, and several important impacts were highlighted. A current observed trend mentioned by several researchers was the field of cloud computing. Cloud computing enables greater access to large datasets, increased collaboration capabilities, and more accessible computing resources. Another trend acknowledged in the workshop was the pairing of increased computing power, increased memory capabilities, and the decreased size of sensors for the purpose of autonomous operations. Dr. Atkins acknowledged that these hardware and software combinations have greatly increased the capabilities of small assets like UAS. UAS are now able to collect data, store it on the platform, analyze the data in real-time, and utilize the results for the purpose of mission or course adjustments, data-collecting changes, etc. This type of in-situ computing is applicable to any system that desires autonomous functionality.

Applications in Science and Engineering

The purpose of the workshop was to bring together experts in the fields of machine learning and big data with scientists and engineers that could leverage the advances to tackle their complex problems. Several of the experts that presented their research identified unique use cases where they are either currently exploring the use of machine learning techniques or where there is potential for the methods to have a positive impact in their research. A few use cases that were prominently mentioned throughout the workshop were climate science studies, aerospace vehicle design studies, and turbulence modeling.

Dr. Xiao and Dr. Duraisamy both discussed how they are leveraging machine learning techniques for modeling and analysis of turbulent flows. It was explained that there are instances where researchers do not have a high level of confidence in the current models utilized for turbulent flows because some of the defining physics is not well-understood. Dr. Duraisamy discussed the use of machine learning techniques to remedy large scale differences within turbulent flows that create discrepancies between reality and the model. Dr. Xiao acknowledged that machine learning can be utilized to solve physical problems with the goal of correcting and improving existing, low fidelity turbulence models.

Dr. Kumar discussed his use of machine learning algorithms to analyze different climate science problems. It was highlighted that existing climate science data has some challenging characteristics, such as multi-resolution and scale, large amounts of noise and missing data, and missing labels. Dr. Kumar provided discussions on two specific climate science examples, the global mapping of forest fires and global mapping of inland surface water dynamics. The approach explored for both of these use cases involves classification through the use of imaging data. It was identified that utilization of classification techniques in isolation may not do a faultless job, but the performance can be greatly improved through the inclusion of physics guided labeling. Dr. Vatsavai also utilized methods involving image mapping and clustering on climate science related use cases, such as agriculture.

Dr. Dimitri Mavris provided insight into how he and his team of researchers are utilizing machine learning techniques to advance the field of complex system design. It was highlighted by Dr. Mavris, as well as other researchers, that legacy analysis tools can be largely reliant on historical datasets that are no longer applicable to new, advanced systems. Machine learning techniques have provided the capability of integrating toolsets into multidisciplinary assessment environments that can be optimized and provide efficient design space exploration. This results in the identification of more potential designs that meet a given set of constraints and the ability to optimize the design settings for a given set of goals. Furthermore, it was acknowledged that aerospace system architects are also leveraging machine learning techniques to conduct uncertainty assessments developing technologies.

Virtual Assistant using Natural Language Processing

Several different use cases were discussed throughout the course of the workshop, but one that was mentioned several times was the concept of a ‘virtual assistant’. A ‘virtual assistant’ concept is a man-made, computer-based machine that is capable of performing basic research tasks that would otherwise be done with a human in the loop. The concept of the virtual assistant is coming into play now largely due to advances in natural language algorithm development and computing power. Several researchers presented work where there was extensive usage of natural language, with the first being the description of IBM Watson from Dr. Chris Codella. Watson utilizes natural language in several ways. First, it is able to take commands, or questions, in the form of spoken natural language. It then breaks down the meaning of the question and accesses its knowledge-base to determine an appropriate answer or set of answers. Watson is then able to reply in the form of a natural language expression.

In addition to Dr. Codella, Dr. Scheutz provided similar natural language research through the robots he and his collaborators are developing. Techniques utilized for systems like Watson and the robots described by Dr. Scheutz are directly applicable to the concept of the virtual assistant. The virtual assistant would take in direction through natural language commands, decipher the topics it is searching for, determine if information it is sorting through is relevant, synthesize the information, and report it back in an understandable manner.

Other researchers, such as Dr. Lee, discussed the importance of machines to be able to search through natural language data sources (i.e., published articles and papers) and find relevant information. In these scenarios the machines are not taking direction or providing answers through spoken natural language, but they still require the ability to understand sentence structure. This leads to a key concept that was discussed, language parsing. Language parsing is a key concept when dealing with natural language information. When a machine is provided a command or question in natural language form, it must dissect it into terms it can understand and determine the appropriate knowledge base for synthesizing an answer. The parsing logic built into the machine will determine the rigidity requirements for the user’s commands, and it was acknowledged that some machines are currently able to only accept a certain type of structured rhetoric, and changes would need to be made for it to accept dialect from sources like social media.

The final key topic discussed with regards to natural language was the ability to use languages other than English. All researchers that presented natural language work have only currently experimented with the English language. Most researchers said they were not planning on incorporating other languages in the near future, but that it could be done. It was acknowledged that there may be an increased complexity in the way a natural language command is parsed if a language other than English is considered.

Next Steps

The *Machine Learning Technologies and Their Applications for Scientific and Engineering Domains* workshop provided valuable insight into the fields of machine learning and data analytics by domain experts and users alike. The engineering and scientific communities, including NASA and the aerospace industry, can benefit from the advances made in these domains in a multitude of ways. Furthermore, workshops like this provide the opportunity for harvesting relationships among the domain experts and the scientists that can benefit from the methods in their respective fields. It is the hope that these opportunities and the relationships they build will provide mutually beneficial, collaborative research projects that will enable a further expansion in many scientific domains.

Reflecting on these thoughts, five key areas discussed during the workshop have been targeted for discussion about their future roles. These areas are machine learning techniques and methods, computational fluid dynamics, aerospace system design, human machine interaction, and climate science. Several key workshop participants and attendees have provided their insights with respect to these areas and the summaries are provided in the following sub-sections.

Machine Learning Techniques and Methods

The Big Data Analytics and Machine Intelligence/BDAMI team of the CDT initiative at NASA Langley provided their key takeaways from the workshop and the follow-up opportunities they identified for further collaboration.

Key Takeaways

Several key takeaways from the workshop were provided from the team. They are as follows:

- Dr. Poczos had an excellent presentation on how to best deal with complex problem spaces such as pictures (essentially break up into smaller spaces and represent the input to the learner as a set of the smaller spaces) and how to deal with sparse data through anomaly detection; these two ideas are very relevant to CSM. Dr. Xiao also had a good presentation that included a discussion on what to do when you aren't sure about ground truth/ you can't 100% trust your models.
- Active learning is a semi-supervised approach in which a learning algorithm is able to interactively query the user to obtain desired output at new data points. This concept has many applications to NASA-related projects and overall missions.
- Machine learning on distributions is a powerful nonparametric approach to anomaly detection and clustering. Again, many applications to NASA-related projects and overall missions.
- The 'physics' of a problem is never completely eliminated from machine learning models. This means that the physics are either directly or indirectly a part of the machine learning pipeline, which may explain why some use cases fail in machine learning because physics of the underlying problem are ignored.
- Deep learning models may be able to achieve human-level performance across many other areas. However, the big bottleneck is the hardware needed to train the models. Specialized hardware being developed such as tensor processing units (TPUs) to help keep up with user needs.
- One-shot learning is an object categorization problem in computer vision that has important applications to first-responders.

- Approaches to using Machine Learning on Distributions in order to deal with complex objects was found very interesting. Specifically, Dr. Poczos' research on techniques to break objects into sets, and treat the elements of sets as points from an unknown distribution.
- Research presented by Dr. Carbonell on strategies for effectively managing issues with reliability, reluctance, and cost per query.
- Effective methods for incorporating physics into machine learning methodologies and understanding that even basic applications of physical properties (Kumar) can produce results.
- Sparse data is a challenge that most experience and several Machine learning techniques could be used such as transfer learning and active learning and adaptive learning.
- Domain experts and computer scientist/machine learning experts working together can develop best possible solutions for scientific and engineering challenges.
- Machine learning techniques and tools are maturing and focused workshops in their optimal use are needed; potential for insights and intelligence from our aerospace data and information is significant and collaboration with Universities is critical to leverage the emerging techniques and technologies.

Topics with Follow-up Potential

A potential avenue for expansion and improvement is the inclusion of both introductory level and more in-depth, follow-up workshops --- separate from the high-level seminars/ workshops which are currently featured --- in order to bring in an a fully encompassing experience for interested attendees. Such topics of value could include a base-level introduction to Machine Learning and related methodologies, and, possibly, a short workshop on the applications of machine learning in regards to the use cases presented by the attendees (such as: how one reframes the way they think about their data to fit a ML model).

Furthermore, Dr. Carbonell of CMU was identified as a potential guest for this workshop due to his research into active learning methodology for the validity to the prospective use cases site wide. Also, the possibility of pulling in high profile individuals to the future seminar, such as: Andrew Ng (Stanford Professor, Chief Scientist at Baidu Research), Sebastian Thurn (AI Professor at Stanford and Ga Tech) , Pete Warden (Tensor Flow Mobile team lead, google), Demis Hassabis (Co-founder, Deepmind [now Google Deepmind), or one of the many talented individuals working at OpenAi were suggested.

Members of the team also identified the researchers/Professors from the workshop they are most interested in having follow-up collaborations with. These included: Dr. Barnabas Poczos. Dr. Jaime Carbonell. Dr. Una-May O'Reilly, and Dr. Ella Atkins.

Computational Fluid Dynamics

Dr. Christopher Rumsey, a Senior Researcher at NASA Langley Research Center, provided the following feedback on the workshop and how the field of computational fluid dynamics can utilize the techniques presented.

Machine learning is starting to make inroads into the field of turbulence modeling for computational fluid dynamics (CFD). Recently, several researchers (such as Karthik Duraisamy of University of Michigan and Heng Xiao of Virginia Tech) have made substantial progress. While they are apparently not yet at the point of devising new globally-useful turbulence models for CFD, they have already started to use the power of big data analytics to help guide possible improvements. In particular, we hope that the power of big data can eventually help us to improve turbulence model performance for problematic cases for which current models are typically poor, such as separated aerodynamic flows.

Over the next decade, close collaboration between turbulence modeling experts and machine learning experts could help to drive breakthroughs. In the past, turbulence modelers have often looked for insights from experiments or mathematical analyses, but these techniques have not born much fruit in a long time. It is hoped that big data analytics can be used in an intelligent way to tease out insights from hundreds or thousands of direct numerical simulations (which of course have to be found and/or created, and collated).

Aerospace System Design

William Kimmel, a vehicle design and system analyst at NASA Langley Research Center, attended the workshop and provided his feedback on how machine learning techniques are, and will remain, relevant to the domain of aerospace system design. Mr. Kimmel provided three main areas of interest he took away from the workshop. The first area of interest is the use of machine learning enablers for next generation MDAO frameworks. Currently, he is collaborating with Dr. Dimitri Mavris at Georgia Tech on MDAO frameworks to extend and increase the level, and discipline scope, of physics-based variable fidelity MDAO frameworks with the purpose of increasing the utility and efficiency of their application to advanced system concepts involving disruptive technologies. Together they are also focusing on methods that enable meaningful insight into how "good" an analysis result is in both the systems and architecture (system of systems) context. Such a need exists to dramatically increase the credibility and utility of systems analysis products from the Decision Maker/Stakeholder viewpoint. These results cannot be presented to a stakeholder simply as a "magic black box" output and are challenged by the scale and complexity in doing so already with current state of approaches/methods. Key subtopics of interest here are reasoning under uncertainty, uncertainty quantification and management, and deep learning.

Secondly, Mr. Kimmel is interested in the initial and beginning exploration of potential collaboration with Chris Codella on his cognitive computing/IBM Watson enabling research. Currently, there is an active study now with the CDT team, NASA Langley's Systems Analysis and Concepts Directorate (SACD), and IBM. Mr. Kimmel is very interested in specifically how this technology of radical scale integration and data aggregation could aid NASA as various subject matter experts (SME) retire. Many of these SMEs physically possess the expertise of accessing and working with a tremendous volume of existing scientific and technical information (STI) databases. The desire is to determine if NASA can retain corporate knowledge for future aerospace researchers despite the exodus of those with broad wisdom and extensive longitudinal history of NASA research. Furthermore, it is questioned if, and how, a more efficient and deeper understanding of the STI database can inform future study of systems/architectures where revolutionary/disruptive/highly coupled interdisciplinary technologies are involved. SACD is beginning to explore this discovery-to-options-to-decision thought process through the current IBM collaboration. Moving forward, SACD would like to know more about the "Watson Discovery Advisor" utility to understand how this could further their future MDAO direction.

The third topic of interest was the materials informatics discussion provided by Dr. Krishna Rajan of the University of Buffalo. Dr. Rajan's ideas and application examples stirred up a number of questions that Mr. Kimmel would like to explore with CDT and others moving forward. The questions include:

- How can the Big/Deep Data functions presented be adapted to the mesa-mega scales SACD is targeting in their systems analysis and advanced concepts discipline area?
- Are these atomistic/molecular level approaches limited to the nano-micro scale or if they are readily extensible to multi-disciplinary, multi-objective studies?

- What are the barriers to incorporating this and what is the tractability especially in highly coupled, complex interdependent systems and architectures?

Through his presentation, Dr. Rajan offered a tantalizing potential in the application example scheme starting with x-ray crystallography and resulting decision system for materials research. A key intriguing concept is the "Forming the Inverse Design Problem" as a data mining problem that incorporates multi-scale scope and is reliable despite limited and uncertain information. Furthermore, another interesting term was introduced, which was the concept of identifying so called "solution manifolds" in a trade space.

Human Machine Interaction

Lisa Le Vie, a Human/Machine researcher at NASA Langley Research Center, provided her perspective on the potential advancements of the Human Machine Interaction field in the future. Ms. Le Vie attended the workshop due to the breadth and depth of the provided material, and was specifically interested in the discussions about human intelligence modeling learning. The creation of autonomous systems requires machine learning techniques, and as the field of autonomous systems progresses researchers in her domain need to survey all method options and develop a 'jack of all trades' mentality. Two different learning approaches for autonomous systems were discussed at the workshop, data-driven and one-shot learning. Ms. Le Vie believes in order to achieve adult human level intelligence, pieces of both types of machine learning will be required.

The domains of big data and machine learning are somewhat new, but use of open-source and cloud-based collaboration provide the opportunity for collaborators to make their own journey in a replicable manner. This type of sharing is better than separate entities protecting their solutions because collaborators work together, learn from each other, and provide a mutually beneficial working relationship. Ms. Le Vie acknowledged that open environments allow herself and other researchers to take datasets and learn from them, which enables looking at the data in different ways. There are many available techniques for discovering information from a new dataset, which can result in a seemingly endless number of ways to look at it. The results of these collaborative environments are better science and engineering, the development of better algorithms, lessons learned, etc. These results will be beneficial to the human-machine interaction domain, as well as other scientific domains.

Climate Science

Dr. Patrick Taylor, a climate scientist at NASA Langley Research Center, provided his insight into where the field of machine learning can provide the most benefits in the climate science domain. One opportunity for leveraging machine learning techniques is the creation of a better model for predicting cloudiness within climate models. Many of the current methods for calculating cloudiness are over 20 years old and climate scientists do not utilize machine learning techniques to do cloud retrievals due to a lack of good training datasets. However, NASA has been collecting information from high quality sensors over the last decade that could be used to inform the basic assumptions/models for cloud formation that are used within bigger climate models. The data includes high fidelity measurements of where clouds are located in the vertical direction and lower-fidelity measurements in the horizontal direction. Machine learning techniques, such as artificial neural networks, could be used with these datasets to create a better model for predicting cloudiness within climate models. These techniques could provide a means for accelerating progress by facilitating the infusion of new information into the models and model development.

Another issue is that this new, high fidelity data is sparse, which could make utilizing neural networks challenging. However, several methods for dealing with sparse data were discussed at the workshop and Dr. Taylor feels collaborating with these researchers could be beneficial to this problem. Furthermore, the high fidelity vertical data could be fused with the lower-fidelity, broad swath data through machine learning techniques to aid the learning about the cloud that can be achieved, such as predicting the vertical profile of the cloud in locations where vertical measurements do not exist.

Another uncertainty Dr. Taylor acknowledged was the hesitation of climate scientists to use the results of something like neural networks, because they are sometimes viewed as 'black box' models with no physical intuition built into it. However, Dr. Xiao acknowledged this topic during the workshop through his presentation on turbulence modeling. It was highlighted that it is possible to build physical intuition into neural networks by using data to build an addendum to an existing model, which will increase the overall model fidelity. For climate science, this could mean that physical intuition, such as the relationship between temperature and water vapor, is utilized to aid the selection of a neuron function when building the neural network.

Aircraft Training and Safety

Dr. Alan Pope, a Senior Research Scientist at NASA Langley Research Center, provided insight into how machine learning techniques, such as those discussed at the workshop, are beneficial to his work with flight crew training. Dr. Pope works with multiple time series collected at high data rates, and the machine learning techniques provide the opportunity to understand the data better and improve the flight crew training processes. The Commercial Aviation Safety Team (CAST), an international government-aviation industry partnership, has established a Safety Enhancement (SE211) entitled "Training for Attention Management," which calls for research and development on the detection of attention-related human performance limiting states (AHPLS). Such cognitive states can cause pilots to lose airplane state awareness, which CAST has specified as a casual factor in commercial aviation accidents and incidents. The Crew State Monitoring (CSM) group at the NASA Langley Research Center (LaRC) has implemented an end-to-end system to test the capability of predicting the occurrence of cognitive states using psychophysiological data recorded via multiple sensing modalities. Their collaboration with LaRC' s Big Data/Machine Learning experts has enabled the development and testing of classification algorithms able to predict multiple cognitive states simultaneously in real time within that system.

During training, the CSM team puts the pilots into a flight simulator and subjects them to scenarios where they are likely to experience identified cognitive states. However, they don't know when during the task the pilot will necessarily experience the cognitive state and how much they will experience the state. If models were available, the data collected from the training exercises could be put into the models to answer these questions and provide insight into the likely experience a person has at a given moment. Before they investigated the use of advanced machine learning techniques, this research was being done with more basic, less sophisticated statistical models. The process for training is to first run some baseline or benchmarking task with commercial airline pilots they are working with. Each task corresponds to something that relates to a given cognitive state, such as surprise. This data forms the training data for the classification models being developed using the machine learning techniques. Displaying this information during the training scenarios is valuable to instructor pilots at airline training facilities because it provides more intelligence than simply eyeballing situations. It enables instructors to train the pilots to improve the management of their attention and inform them when they are entering into a challenging state.

Presentations at the Machine Learning Workshop confirms that the collaboration between the Big Data/Machine Learning team and the CSM team is the effective path to achieve the CAST objectives of developing methods for identifying and controlling AHPLS that lead to crew or operator error, the largest source of accident causes, in order to improve the safety of transportation systems. In particular, Dr. Carbonell's presentation of the partnership between Carnegie Mellon and Boeing applying machine learning to preventative maintenance provided a useful analogy with the ongoing cognitive state detection research. Both research endeavors are aimed at predicting the future based upon past data augmented by current data. Whereas the preventative maintenance research works to predict the need for repair or replacement based on a part's historical data, CSM research strives to predict an aircrew's cognitive state based on their past baseline physiological data. Dr. Carbonell's presentation also provided a useful machine learning tutorial as well as other enlightening use cases. Dr. Poczos of Carnegie Mellon, in his presentation on Neuroscience, posed the intriguing possibility of decoding thoughts from brain scans, analogous to the ongoing LaRC CSM research to decode cognitive state from brainwave physiological signals through use of machine learning methods.

Summary Remarks

The importance of machine learning technologies, and big data analytics to bring about transformational changes in conducting science and engineering research, engineering, and systems design is being understood. The willingness, enthusiasm, and participation by 15 external expert presenters from universities and industry and 300 of our scientists and engineers in this three-day workshop is a resounding acknowledgement of that. Examples of discipline areas where machine learning technologies are developed and utilized by the workshop presenters is broad ranging and includes materials, aero sciences, aircraft design, climate science, global earth observations, autonomy, and knowledge assistants.

While there are success stories based on more than a decade of work in several cases in medicine and some science and engineering areas, there are several technical challenges that need to be overcome – e.g., data accuracy and classification, fusing data of varied fidelity, techniques to deal with missing data strings in a database, natural language processing, barriers to applications to complex systems, and teaching machines science and engineering domains. The exponential progress in machine learning, natural language processing , cognitive technologies, artificial intelligence technologies coupled with advances in computing research and development including quantum computing being made in universities and industry will offer solutions to these and others to expand applications to science and engineering domains that benefit us. It is imperative for computer scientists and domain experts working together both at Universities and NASA Centers to develop and advance solutions for scientific and engineering domains. In fact, at NASA Langley BDAMI team of CDT has been increasingly modeling this collaboration model during 2015 and 2016. The next steps captured in this report will be pursued in the coming months and years to leverage current knowledge and ongoing work. As a user community of the machine learning and big data analytics technologies, we must continuously stay connected with all key organizations that advance the state-of-the-art and leverage their efforts to advance solutions to our discipline which benefit NASA mission challenges. As CDT increasingly focuses on advancing simulation based science and engineering for innovation solutions for NASA mission challenges, applications of machine learning and artificial intelligence technologies will become more critical, and continue to deepen the expertise of applying these technology to aerospace domains is necessary and essential.

REPORT DOCUMENTATION PAGE	*Form Approved* *OMB No. 0704-0188*

1. REPORT DATE *(DD-MM-YYYY)* 01- 12 - 2016	**2. REPORT TYPE** Technical Memorandum	**3. DATES COVERED** *(From - To)*

4. TITLE AND SUBTITLE Machine Learning Technologies and Their Applications for Science and Engineering Domains Workshop - Summary Report	**5a. CONTRACT NUMBER**
	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S) Ambur, Manjula Y.; Schwartz, Katherine G.; Mavris, Dimitri N.	**5d. PROJECT NUMBER**
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER 736466.07.08.07.02
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) NASA Langley Research Center Hampton, VA 23681-2199	**8. PERFORMING ORGANIZATION REPORT NUMBER** L-20772
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES) National Aeronautics and Space Administration Washington, DC 20546-0001	**10. SPONSOR/MONITOR'S ACRONYM(S)** NASA
	11. SPONSOR/MONITOR'S REPORT NUMBER(S) NASA-TM-2016-219358

12. DISTRIBUTION/AVAILABILITY STATEMENT

Unclassified - Unlimited
Subject Category 82
Availability: NASA STI Program (757) 864-9658

13. SUPPLEMENTARY NOTES

14. ABSTRACT

The fields of machine learning and big data analytics have made significant advances in recent years, which has created an environment where cross-fertilization of methods and collaborations can achieve previously unattainable outcomes. The Comprehensive Digital Transformation (CDT) Machine Learning and Big Data Analytics team planned a workshop at NASA Langley in August 2016 to unite leading experts the field of machine learning and NASA scientists and engineers. The primary goal for this workshop was to assess the state-of-the-art in this field, introduce these leading experts to the aerospace and science subject matter experts, and develop opportunities for collaboration. The workshop was held over a three day-period with lectures from 15 leading experts followed by significant interactive discussions. This report provides an overview of the 15 invited lectures and a summary of the key discussion topics that arose during both formal and informal discussion sections. Four key workshop themes were identified after the closure of the workshop and are also highlighted in the report.

15. SUBJECT TERMS

Artificial intelligence; Big data analytics; Comprehensive digital transformation; Machine intelligence; Machine learning; Virtual assistants

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON**
a. REPORT	**b. ABSTRACT**	**c. THIS PAGE**			STI Help Desk (email: help@sti.nasa.gov)
U	U	U	UU	42	**19b. TELEPHONE NUMBER** *(Include area code)* (757) 864-9658