

NASA/TM-2016-219361



Big Data Analytics and Machine Intelligence Capability Development at NASA Langley Research Center: Strategy, Roadmap, and Progress

*Manjula Y. Ambur and Jeremy J. Yagle
Langley Research Center, Hampton, Virginia*

*William Reith
Booz Allen Hamilton, Hampton, Virginia*

*Edward McLarney
Langley Research Center, Hampton, Virginia*

December 2016

NASA STI Program . . . in Profile

Since its founding, NASA has been dedicated to the advancement of aeronautics and space science. The NASA scientific and technical information (STI) program plays a key part in helping NASA maintain this important role.

The NASA STI program operates under the auspices of the Agency Chief Information Officer. It collects, organizes, provides for archiving, and disseminates NASA's STI. The NASA STI program provides access to the NTRS Registered and its public interface, the NASA Technical Reports Server, thus providing one of the largest collections of aeronautical and space science STI in the world. Results are published in both non-NASA channels and by NASA in the NASA STI Report Series, which includes the following report types:

- **TECHNICAL PUBLICATION.** Reports of completed research or a major significant phase of research that present the results of NASA Programs and include extensive data or theoretical analysis. Includes compilations of significant scientific and technical data and information deemed to be of continuing reference value. NASA counter-part of peer-reviewed formal professional papers but has less stringent limitations on manuscript length and extent of graphic presentations.
- **TECHNICAL MEMORANDUM.** Scientific and technical findings that are preliminary or of specialized interest, e.g., quick release reports, working papers, and bibliographies that contain minimal annotation. Does not contain extensive analysis.
- **CONTRACTOR REPORT.** Scientific and technical findings by NASA-sponsored contractors and grantees.

- **CONFERENCE PUBLICATION.** Collected papers from scientific and technical conferences, symposia, seminars, or other meetings sponsored or co-sponsored by NASA.
- **SPECIAL PUBLICATION.** Scientific, technical, or historical information from NASA programs, projects, and missions, often concerned with subjects having substantial public interest.
- **TECHNICAL TRANSLATION.** English-language translations of foreign scientific and technical material pertinent to NASA's mission.

Specialized services also include organizing and publishing research results, distributing specialized research announcements and feeds, providing information desk and personal search support, and enabling data exchange services.

For more information about the NASA STI program, see the following:

- Access the NASA STI program home page at <http://www.sti.nasa.gov>
- E-mail your question to help@sti.nasa.gov
- Phone the NASA STI Information Desk at 757-864-9658
- Write to:
NASA STI Information Desk
Mail Stop 148
NASA Langley Research Center
Hampton, VA 23681-2199

NASA/TM-2016-219361



Big Data Analytics and Machine Intelligence Capability Development at NASA Langley Research Center: Strategy, Roadmap, and Progress

*Manjula Y. Ambur and Jeremy J. Yagle
Langley Research Center, Hampton, Virginia*

*William Reith
Booz Allen Hamilton, Hampton, Virginia*

*Edward McLarney
Langley Research Center, Hampton, Virginia*

National Aeronautics and
Space Administration

Langley Research Center
Hampton, Virginia 23681-2199

December 2016

Acknowledgments

The authors gratefully acknowledge the following individuals for their invaluable contributions to the Comprehensive Digital Transformation (CDT) Big Data Analytics and Machine Intelligence Capability Team at NASA Langley Research Center, and for their continued support and collaboration on capability development for data analytics and machine learning. Those who contributed to the development of the initial vision, strategy, and roadmap in 2014 include Sakeba Abedin, Steve Dry, Cory Gilbert, Dana Hammond, Dave Hinton, Michael Little, Ed McLarney, Peter Mount, Brandi Quam, William Reith, Steve Scotti, Ted Sidehamer, Mia Siochi, Jim Thomas, Sharon Welch, and Bill Winfree. Big Data Analytics and Machine Intelligence (BDAMI) Team members including Lin Chen, James Ecker, Christina Heinich, Charles Liles, Raymond McCollum, Graeme Melrose, Robert Milletich, William Reith, Daniel Sammons, Travis Smith, Ted Sidehamer, and Jeremy Yagle have diligently and skillfully worked on use case development, outreach, and collaboration over the past three years. Their efforts have been supplemented by interns including Karleigh Cameron, Zachary Ernst, Moshea Fink, Evana Gizzi, Wade Hunter, Colin Lockard, Travis Moore, Macarena Ortiz, Jefferson Ridgeway, Alexander Sarris, Nicholas Sarris, Jake Spracher, and Troy Thomas. Leaders of the CDT and BDAMI activities including Damodar Ambur, Manjula Ambur, J.F. Barthelemy, Dennis Bushnell, Jill Marlowe, Ed McLarney, Joe Morrison, and Jeff Seaton have provided guidance and direction for all aspects of the initiative, collaborations, and outreach efforts. Finally, we extend our sincere thanks to the subject matter experts from our mission organizations, without whose strong collaboration and continued support this work would not otherwise be possible; Danette Allen, Dale Arney, Trey Arthur, Erik Axdahl, Randy Bailey, Kristopher Bedka, Douglas Brown, 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 Siochi, Chad Stephens, Scott Striepe, Patrick Taylor, Marty Waszak, Bill Winfree, and Kristopher Wise.

<p>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.</p>

Available from:

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

Table of Contents

Executive Summary	4
A: Big Data Analytics & Machine Intelligence Strategy and Roadmap	8
A1 Overview	8
A1.1 Background.....	8
A1.2 Current State and Approach.....	10
A1.3 Vision.....	11
A1.4 Scope	13
A2 Strategic Goals, Objectives, and Initiatives	14
A3 Use Cases	23
A4 Knowledgeability	26
A5 Research and Partnerships	27
A5.1 Federal Research Initiatives.....	28
A5.2 Partnerships.....	28
A6 Technology and Architecture.....	29
A6.1 Deep Analytics and Data Visualization	29
A6.2 Data Mining and Data Discovery	29
A6.3 Machine Learning and Machine Intelligence.....	29
A6.4 Architecture and Infrastructure	30
A7 Workforce Skills	32
A8 Findings, Key Recommendations, and Actions	33
A8.1 Findings	33
A8.2 Key Recommendations	34
A8.3 Actions.....	36
A9 The Way Ahead	37
A9.1 Organizational and Cultural Changes	38
A9.2 Funding Considerations	38
B: Progress Made to Date (2014-2017)	39
B1 Redefinition of CDT Strategy	39
B2 Vision, Roadmap, Recommendations, and Actions.....	41
B3 Progress on Data Intensive Scientific Discovery (DISD) Projects	44
B3.1 Nondestructive Evaluation: Automated Identification of Anomalies to Assess Structural Damage.....	45
B3.2 Aeroelasticity: Predicting Flutter from Aeroelasticity Data.....	47
B3.3 Crew Cognitive State Monitoring and Detection	50
B3.4 Rapid Exploration of Aerospace Designs	53
B3.5 Turbulence Modeling	55

B3.6	Entry, Descent, and Landing	56
B3.7	Climate Science: Cloud Fraction Simulation	56
B3.8	Space Launch System (SLS) Booster Separation Aerodynamics	57
B3.9	Hypersonic Inlet Performance Analysis	58
B3.10	Space Launch System (SLS) Additive Manufacturing Certification	58
B3.11	Flight Deck Analytics from Trajectory-Based Optimization	58
B4	Progress on Deep Content Analytics (DCA) Projects	58
B4.1	Carbon Nanotubes	60
B4.2	Autonomous Flight	60
B4.3	Space Radiation	60
B4.4	Aerospace Vehicle Design	60
B4.5	Uncertainty Quantification	61
B4.6	Space Mission Analysis	61
B4.7	Model Based Engineering	61
B4.8	Human-Machine Teaming	61
B4.9	NASA Technical Reports	62
B4.10	NASA Lessons Learned	62
B5	Progress on Deep Q&A Projects	62
B5.1	Cognitive Computing and the “NASA Watson” Vision	62
B5.1	Current Application: Pilot Advisor Proof of Concept	64
B5.2	Current Application: Aerospace Innovation Advisor Proof of Concept	66
B5.3	Next Steps	67
B6	Collaboration	67
B6.1	Collaboration with other NASA Centers	67
B6.2	Collaboration with Universities	68
B6.3	Collaboration with Expertise in Industry and Other Federal Agencies	68
B7	Outreach and Education	69
B7.1	Seminars and Focus Groups	70
B7.2	Agency Events	70
B7.3	Conferences	71
B7.4	Center-wide Workshops	71
B8	Research Machine Intelligence	72
B9	Hire or Obtain Expertise	73
C	Moving Forward (2018 – 2020)	74
C1	Vision, Roadmap, Recommendations, and Actions	74
	Phase II Actions (2018-2020)	77
C2	Focus on NASA Langley Product Lines and Linkage to NASA Projects	77

C3 Focus on Mission Support Functions	78
C4 Enhancing Data Intensive Scientific Discovery (DISD).....	78
C5 Enhancing Deep Content Analytics (DCA)	78
C6 Enhancing Deep Q&A and “NASA Watson”	79
C7 Enhancing Collaboration.....	79
C8 Artificial Intelligence	80
C9 Enhancing Outreach and Education	80
C10 Integration with High Performance Computing and Modeling & Simulation.....	80
Concluding Remarks.....	81
Appendix A: References for Section A	83
Appendix B: References for Section B.....	84
Appendix C: Key Presentations	86
Big Data Analytics and Machine Intelligence Capability Algorithms and Software: AIAA SciTech Conference (January 2016).....	86
Comprehensive Digital Transformation: An Overview (February 2016).....	94
Knowledge Analytics and Data Analytics (August 2016).....	107
Cognitive Computing Vision and Watson Applications at NASA (April 2016).....	113

Executive Summary

The term *Big Data* refers to datasets whose volume, speed, and complexity is beyond the ability of typical tools to capture, store, manage and analyze. Big data is often described by the four V's -- volume of data, velocity of data, variety or types of data, and veracity or accuracy of the data. Analysis of the data can be enhanced by *Machine Learning* and *Machine Intelligence*. Machine Learning relies on algorithms that are capable of learning from both data and human interaction to enable insights and predictions, while machine intelligence involves an autonomous entity that can observe and act upon its environment, making decisions like a human. Each of these three terms refers to rapidly emerging capabilities that are revolutionizing how industry and government analyze data.

The fields of machine learning and big data analytics have made significant advances over the past several years and have been demonstrating the potential to transform how the traditional disciplines of science and engineering are conducted. These new, advanced methods, paired with rapidly evolving computational capabilities, have created an environment where cross-fertilization of methods and unique collaborations can achieve previously unattainable outcomes. NASA Langley Research Center (LaRC) has recognized these changes in the technical and scientific communities and created the Comprehensive Digital Transformation (CDT) initiative that focuses on four main pillars: advanced modeling and simulation, machine learning and big data analytics, high performance computing, and advanced IT infrastructure. The primary goal of CDT is to serve as a catalyst to apply these capabilities both individually and through convergence of these compute- and data- intensive capabilities to enable innovative concepts, reduced design cycle time, improved affordability, and increased confidence in the designs.

A team of researchers, engineers and information technology (IT) specialists developed the *Big Data Analytics & Machine Intelligence Strategy and Roadmap* in 2014. The vision is to have a “Virtual Expert” or “Virtual Research and Design Partner” enabling NASA employees to achieve greater scientific discoveries and system design optimization. Drawing from this document, Section A presents the BDAMI goals, objectives, initiatives, key recommendations, and phased actions designed to equip LaRC to develop near-, mid- and long-term big data capabilities.

Section B provides an overview of the significant progress that has been made over the last two and a half years in developing pilots and projects in several research, engineering, and science domains, and implementing them using both institutional and aeronautics and exploration project funds. These accomplishments were possible with significant collaboration between the multi-skilled BDAMI team, mission organizations, and external partners from universities and industry. 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.

In Section C, a strategy for moving forward is presented, with a focus on achieving next steps for the near future (2018 – 2020.) Since there are few readily available solutions for applying machine learning and big data techniques to the information and physics-based data sets in our aerospace domains, our challenges must be researched and developed into solutions by enhancing current work, continuing collaboration, and providing regular outreach and education.

The six main goals of the strategy are:

- **Goal #1** – Keep up with big data, deep analytics and machine intelligence technologies and capabilities, and advance LaRC knowledge and utilization of them.
- **Goal #2** - Build a robust data intensive scientific discovery analytics capability and cognitive computing analytics to enable better science and engineering.
- **Goal #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.
- **Goal #4** – Ensure understanding and use of machine intelligence remains a long-term focus.
- **Goal #5** – Proactively pursue, utilize, and leverage partnerships and collaborations with universities, federal research organizations, Department of Energy (DoE) labs and industry.
- **Goal #6** – Ensure buy-in at the grassroots level, resource availability and investment prioritization for building and enhancing a big data, deep analytics and machine intelligence capability.

Twelve key recommendations to meet these strategic goals are outlined in Table 1, below.

12 Key Recommendations	Near-Term	Mid-Term	Long-Term
	2014-2018	2019-2024	2025+
R 1 – Educate and promote the value of big data through seminars and workshops by experts and LaRC working group to foster the understanding of its value and use by mission organizations. (Links to Goal #1)			
R 2 – Understand incubator needs and incorporate them with deep analytics and machine learning pilots and capability; Demonstrate feasibility and add value to incubator success. (Links to Goal #1, Goal #2 and Goal #6)			
R 3 –Build a big data and machine intelligence team, including data scientists, statisticians, algorithm developers, machine learning expert and comprised of civil service employees, contractors and students. (Links to Goal #2, Goal # 4 and Goal #5)			
R 4 – Develop and implement a data-driven scientific discovery capability; start with small-scale and highvalue pilots: non-destructive evaluation (NDE) images, aerelasticity data and cyber security. (Links to Goal #1, Goal #2 and Goal #3)			
R 5 – Develop an IBM Watson-like cognitive computing capability with deep analytics for research and question and answer (Q&A) for design; begin with pilots, including the Knowledge Assistant Pilot in progress. (Links to Goal #1, Goal #2 and Goal #3)			
R 6 – Identify and establish partnerships with universities, government and industry; leverage their expertise for LaRC's big data capability and participate in research when possible. (Links to Goal #1 and Goal #5)			
R 7 – Develop a data capture and management capability for automatic capture of data with context, including meta data standards and tagging, real-time uploads and ingests; start with a pilot. (Links to Goal #2 and Goal #3)			
R 8 – Develop a big data architecture capability; Research and understand technologies, tools and architectures and incorporate learnings from the pilots. Start with Hadoop and cloud pilots. (Links to Goal #1 and Goal #3)			

R 9 – Keep machine intelligence as a North Star goal by actively researching state-of-the art developments, attending seminars /conferences; developing partnerships; and pursuing pilots with the Massachusetts Institute of Technology (MIT). (Links to Goal #4 and Goal #5)			
R 10 – Develop in-situ data analysis with modeling and simulation (M&S) data and implement the capability of high performance computing (HPC), big data and M&S working together; start with pilots. (Links to Goal # 2 and Goal #3)			
R 11 – Develop operational capability for virtual colleagues, experts and intelligent agents; start with pilots. (Links to Goal #1, Goal #2 and Goal #3)			
R 12 – Define and develop metrics for big data capabilities to demonstrate and communicate value to end users and leadership. (Links to Goal #1 and Goal #6)			

Table 1 Overall Roll-Up of 12 Key Recommendations and Projected Timeline

With constraints on NASA funding and in view of the significant ongoing external efforts to develop Big Data and Machine Intelligence technologies, LaRC must utilize a collaboration model to leverage these technologies and adapt them for our science and technology applications. The twelve recommendations reflect an emphasis on pilots, research and partnerships over capital-intensive technology acquisition. Investment in in-house personnel with specific big data and machine intelligence skills will be critical to ensure that right technologies are identified and nurtured to more efficiently and effectively address our challenges. The strategy should be to build a small team of civil servants augmented by contractors and students with the flexibility to bring in skills on an as-needed basis and manage LaRC’s analytic resources (data, talent, technologies, and time). This team will work closely with researchers and engineers from our mission organizations and external partners to implement big data and machine learning solutions for specific LaRC challenges.

A: Big Data Analytics & Machine Intelligence Strategy and Roadmap, 2014-2035

A1 Overview

Big data is a rapidly emerging, and in some cases, even mainstream capability across a range of industries that is revolutionizing how companies and government agencies analyze data. Big data enables users to perform faster analysis and achieve unprecedented and, often, counterintuitive insights from data.

The *Big Data Analytics & Machine Intelligence Strategy and Roadmap, 2014-2035*, provides goals, objectives and initiatives and lists 12 key recommendations to enable Langley Research Center (LaRC) to develop near-, mid- and long-term big data capabilities. Big data, deep analytics and machine intelligence are critical capabilities in LaRC's Comprehensive Digital Transformation (CDT), which is designed to strategically position LaRC to maximize relevant, innovative and persistent contributions to NASA and the Nation in the 21st century. In reading this document, it is important to keep in mind the recommendations outlined represent the forward thinking of the LaRC Big Data, Deep Analytics and Machine Intelligence Team and LaRC scientists and information technology (IT) specialists. They have looked beyond their own disciplines to consider best practices and lessons learned from industry, academia and other government agencies to determine where LaRC research and development efforts should be focused. Given the fact short-term recommendations are generally based on more concrete information than long-range ones, there is greater fidelity for the near- and mid-term than the long-term (e.g., 10-20 years) recommendations. This does not diminish the importance of the long-term recommendations, but merely acknowledges that they are likely to be more malleable as technologies evolve and mission focus areas change. The *Big Data Analytics and Machine Intelligence Strategy and Roadmap, 2014-2035*, will be updated periodically to reflect those changes.

A1.1 Background

The significance of big data is recognized at the highest levels of the Federal government. As described by Jeffrey Mervis in his article "Agencies Rally to Tackle Big Data," *Science*, 6 April 2012, a Federal effort is under way to improve the nation's ability to manage, understand and act upon the 1.2 zeta bytes (10^{21}) of electronic data generated annually. In March 2012, President Barack Obama's administration unveiled a "Big Data Research and Development Initiative," which allocated more than \$200 million (M) to improve the tools and techniques needed to access, organize and glean discoveries from huge volumes of data.

The inherent potential of big data, deep analytics and machine intelligence is predicated on 1) recognition of how these capabilities should be applied for best effect, and 2) recognizing that they do not exist in a vacuum. They are cross-cutting capabilities that can benefit all scientific disciplines. To ensure a clear understanding of big data, deep analytics and machine intelligence and what they entail, the terms, as used throughout this document, are defined below.

“*Big Data*” refers to datasets whose volume, speed and complexity is beyond the ability of typical tools to capture, store, manage and analyze. “Big” will be a different number for each organization that may be trying but unable to extract business advantage from its data. Big data is often described as having four challenges. Known as the four Vs, they include volume, velocity, variety and veracity.

- **Volume** refers to the scale of big data. About 2.5 quintillion bytes (2.3 trillion gigabytes) of data are created daily. By 2020, there are expected to be 40 zeta bytes (43 trillion gigabytes) of information, or 300 times the amount of data in existence in 2005.
- **Velocity** involves the analysis of streaming data. The sheer velocity at which data is being created today is exceptional. The New York Stock Exchange (NYSE) alone captures one terabyte of trade information each session. At another level, a modern car has nearly 100 sensors to monitor data such as fuel level and tire pressure.
- **Variety** involves the numerous forms of data. In 2011, the global size of healthcare data was about 150 exabytes (161 billion gigabytes). As hospitals increasingly adopt systems for electronic medical records, this number will rise. There are an estimated 420 million personal wearable wireless health monitors in use, storing constant data never monitored so extensively before. At the same time, people watch 4 billion hours of video on YouTube and share 30 billion pieces of content on Facebook each month.
- **Veracity** involves the effort to mitigate the uncertainty of data. The fourth big data challenge is keeping it organized in order to distinguish between accurate and inaccurate information. Today, one in three business leaders do not trust the information they use to make decisions. Poor data quality costs the U.S. economy around \$3.1 trillion each year, according to estimates.

“*Deep Analytics*” involves getting value from big data using various analysis techniques, often involving natural language processing. The major outcomes from deep analytics are knowledge, discovery and prediction. Deep analytics depends on:

- Technologies to collect, store, analyze and share huge quantities of data
- Harnessing and using massive data for sensing, perception and decision support to make autonomous systems possible
- Creating human-computer interaction tools for customizable visual reasoning
- Scalable data management analysis and visualization for simulation data

“*Machine Learning*,” and the related field of artificial intelligence, is about the construction and study of systems that can learn from data. Machine learning focuses on prediction, based on known properties learned from training data. It differs from data mining, which focuses on the discovery of (previously) unknown properties of the data.

“*Machine Intelligence*” involves the study and design of intelligent agents, where an intelligent agent is a system that perceives its environment and takes actions that maximize its chances of success. The goal of machine

intelligence is to sufficiently replicate human intelligence to create virtual experts to support researchers and scientists in their work.

Each of these areas has been used to determine everything from potential target markets for products based on social media data to identifying the correlation between retailer food/beverage sales and pre-hurricane weather conditions. The biomedical field, most notably through the use of IBM Watson, and the financial industry have also benefitted in recent years from big data and deep analytics. Nevertheless, it is important to keep in mind applying big data and deep analytics is not a one-size-fits-all panacea for every problem.

LaRC has been involved with big data for several years, even before the term “big data” became popular. Research projects at LaRC generate large amounts of raw data, of which only a fraction is analyzed. While the Office of the Chief Information Officer (OCIO) has analytics tools to process large amounts of data, it needs to expand this capability to optimize its ability to manage knowledge, provide greater opportunities for discovery and, ultimately, use data as a predictive mechanism.

A basic application of big data, deep analytics and machine intelligence involves automating human decision making with automated algorithmic functions, freeing staff from performing repetitive tasks to address more creative and productive tasks. A larger opportunity is to create discoveries and draw conclusions that might not be reached with existing processes and technologies. Sophisticated analytics can substantially improve decision making, minimize risks and unearth valuable insights that would otherwise remain hidden.

A1.2 Current State and Approach

Overall, LaRC is not currently positioned to reap the benefits of digital age advances and risks being left behind. LaRC does not have data scientists or cloud experts on staff and little expertise in over-the-horizon technologies like machine intelligence.

Externally, big data has proven to be a significant force multiplier in the retail, financial, social media and medical fields. However, its application in the scientific and engineering fields is relatively nascent, outside of large entities such as Conseil Européenne pour la Recherche Nucléaire (CERN). This provides an opportunity for LaRC to fill a void.

LaRC, while having massive amounts of data, faces several unique scientific and technological challenges. Some of these will involve analyzing data compiled by a small number of experts in a very specific area of interest. Drawing correlations and determining causation in these situations is likely to require different approaches and algorithms than those used by marketing professionals, social media experts, investment firms and medical professionals with access to data involving populations in the millions.

The OCIO and LaRC leadership has recognized the importance and potential of big data as a core capability and made this a key component of the CDT effort. This is an opportunity for LaRC to build a big data capability and position the organization to realize the benefits of these advanced technologies for mission enablement and success.

In February 2013, a kickoff meeting was held to formally begin the CDT strategy development process. The Big Data Team was formed April 2013. The CDT team outlined several principles to guide the strategy. One of those

principles called for a specific focus on applying and utilizing big data, deep analytics and machine intelligence as force multipliers for simulation-based engineering and science (SBES), systems innovation and scientific discovery. The intent was to provide LaRC competitive advantage, while leveraging external research and innovations. In pursuing that intent, the Big Data Team took several specific steps. These included:

- Forming a team with IT experts and end users/subject matter experts (SME) (e.g., researchers and engineers)
- Recognizing and understanding the possible value of big data, deep analytics and machine intelligence to other LaRC disciplines/missions by collaborating with discipline sub-team leads (aerosciences, structures, multiphysics, materials, high performance computing [HPC])
- Developing detailed use cases with many disciplines. The team also identified potential partnerships it could leverage to establish a scientifically-focused big data capability.
- Extensive knowledgeability collection via seminars, research, etc.
- Identifying and learning about significant research initiatives in order to shape a long-term vision and develop/establish potential partnerships

A1.3 Vision

LaRC faces a future where data intensive scientific discovery presents heretofore unforeseen opportunities to exploit tools and technologies for data mining, analysis, visualization, collaboration and dissemination. The data available from instruments, sensors and simulations and the ability to utilize that data to achieve entirely new levels of knowledge is unprecedented.

In the same way experimental science, exemplified by Galileo's descriptions of natural phenomena, led to theoretical science (i.e., Newton's laws and Maxwell's equations), the computational science of recent decades has evolved into data intensive science. This represents a fourth research paradigm, as outlined in *The Fourth Paradigm: Data-Intensive Scientific Discovery*, Microsoft Research, 2009.

In the future, researchers, engineers and project teams will have "virtual experts/colleagues" capable of answering specific questions, providing a level of human/intuitive cognition and machine cognition that will facilitate:

- Synthesizing and making sense of huge volumes of big data/ information and processes
- Providing discipline modeling and simulations (M&S) in real time
- The generation of predictions for new technologies and design configurations

A massive cloud-based data repository or data lake for LaRC could include all data collected by simulations and sensors. This collection of data would facilitate analysis not possible now since current approaches focus on a single dataset. In the future, such data collections could allow System-of-Systems (SoS) analysis, since the interaction of various model data sets could be looked at collectively.

Likewise cloud computing will need to be considered by LaRC in the future. Rising equipment costs, massive storage needs associated with big data and the need for adequate data security must be addressed as part of the discussion and, ultimately, decision process.

Parallel processing will also be a critical driver as LaRC tackles the future’s hardest problems. Artificial intelligence and Strong AI will further propel LaRC on its journey to gaining new knowledge and enable scientific discovery.

The vision for big data, deep analytics and machine intelligence is to enable LaRC to discover “unknowns” and deliver previously unimaginable capabilities by applying these transformational technologies as force multipliers for scientific and engineering discoveries and systems innovation and optimization. Achieving this vision will provide a number of tangible benefits. These include cost savings resulting from the use of more SBES and less physical testing to enable LaRC to be more competitive and innovate in providing transformational aerospace technologies. Another major benefit will be helping SMEs analyze more data, doing it faster and recognizing new patterns in data not feasible before. This will improve scientific discovery and engineering designs and allow scientists to spend significantly more time performing analysis rather than waiting on algorithms.

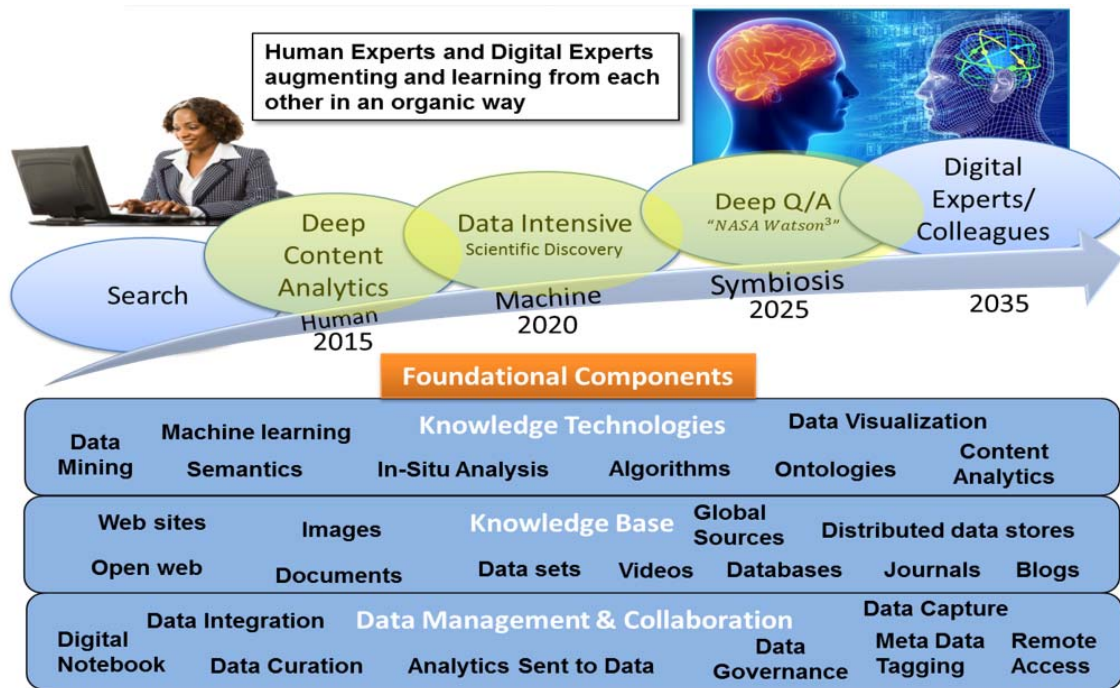


Figure 1 Big Data Analytics and Machine Intelligence Capability Vision

A1.4 Scope

The *Big Data Analytics and Machine Intelligence Strategy and Roadmap, 2014-2035* covers a 20-year period. Assumptions and challenges that bound the recommendations outlined in this document are listed below.

Specific assumptions include:

- **Missions and disciplines** – While mission focus areas may change, the disciplines that currently reside at LaRC will remain largely intact through 2035.
- **M&S** – The use of M&S at LaRC will expand in the future since it offers the potential for lower cost than physical experimentation.
- **Partnerships** – LaRC will increasingly rely on partnerships with other research centers, agencies and laboratories to build knowledge and mitigate funding constraints.
- **Agility and flexibility** – While technologies and tools available 3-5 years from now may be predicted with some degree of certainty, longer-range technologies and tools that will be developed may be unknown at this time. LaRC needs to maintain a high degree of agility and flexibility to quickly leverage new technologies and tools that directly impact its missions.

Specific challenges include:

- **Science is different** - Applying big data technologies for scientific applications is a more complex task and not as mature as other fields such as medicine and finance.
- **Data management and data policy** – Implementing big data will require the creation of data repositories that do not exist and development of architecture and meta data guidelines to support those repositories. There will also be a need to develop sufficient access/restriction controls.
- **Technology and techniques** – LaRC will need to deploy new technologies (e.g., storage, computing and analytical software) and techniques (i.e., types of analysis). A further challenge is that these new tools will need to be integrated with existing systems.
- **Staffing and talent** – Leveraging big data will require acquiring talent and skills not resident at LaRC today.
- **Cost** – LaRC has a limited investment budget. This necessitates making tradeoffs that facilitate access to state-of-the-art technology and research tools at a reasonable cost.
- **Cultural change** – The application of big data, deep analytics and machine intelligence will involve new ways of performing, researching and analyzing data that may not be easily embraced.

A2 Strategic Goals, Objectives, and Initiatives

To determine the pathway for developing a big data, deep analytics and machine intelligence capability, the Big Data Team focused on applying and utilizing force multipliers for SBES, systems innovation and scientific discovery; seeking out partnerships to build knowledge and offset constraints imposed by funding limitations; and leveraging the expertise and collaborating with of discipline SMEs, data scientists, IT specialists and leadership to ensure success.

The team also wanted to ensure integration was a priority. This included identifying potential implementation obstacles early in the process and ensuring resources for problem resolution were available. It also necessitated making recommendations that would allow enough flexibility during implementation to accommodate technological breakthroughs that propel big data, deep analytics and machine intelligence development forward more quickly than initially projected. The Big Data Team felt it was imperative to pursue pilots that are technologically mature, have measurable value and are consistent with available resources. The team also considered specific virtual capabilities being addressed by the CDT team in its strategy development process. Detailed descriptions of the virtual capabilities are in the *Comprehensive Digital Transformation Strategy & Roadmap*.

Figure 2 provides a high-level overview of the roadmap and associated timeline for development of the big data, deep analytics and machine intelligence capability at LaRC. The overview was generated based on the team's development of specific goals, objectives and initiatives, listed in Table 1. These were then distilled into a list of twelve key recommendations to showcase the most critical needs for developing the capability.

Big Data & Machine Intelligence Roadmap

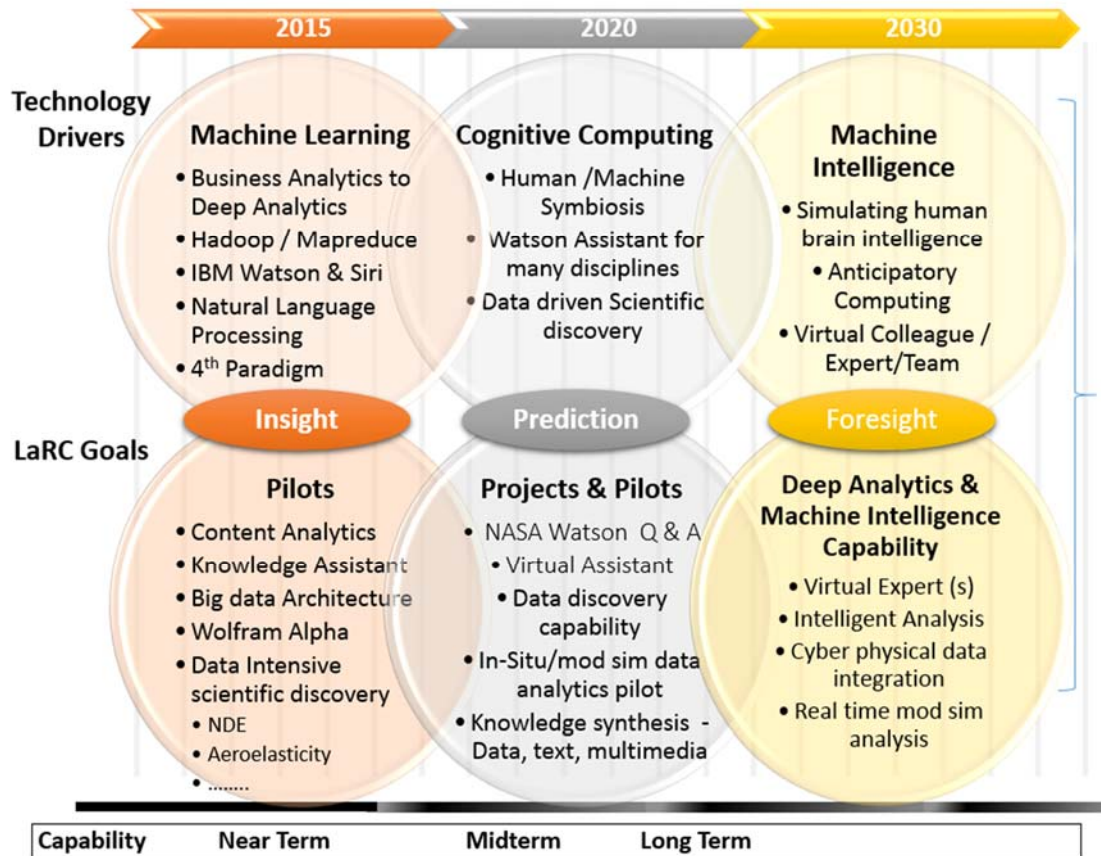


Figure 2 Overarching Big Data, Deep Analytics & Machine Intelligence Roadmap

- **Goal #1**—Keep up with big data, deep analytics and machine intelligence technologies and capabilities, and advance LaRC knowledge and utilization of them.
- **Goal #2**—Build a robust data intensive scientific discovery analytics capability and cognitive computing analytics to enable better science and engineering.
- **Goal #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.
- **Goal #4**—Ensure understanding and use of machine intelligence remains a long-term focus.

- **Goal #5**—Proactively pursue, utilize and leverage partnerships and collaborations with universities, federal research organizations, Department of Energy (DoE) labs and industry.
- **Goal #6**—Ensure buy-in at the grassroots level, resource availability and investment prioritization for building and enhancing a big data, deep analytics and machine intelligence capability.

Objectives are more narrowly scoped statements outlining general actions that, when achieved, contribute to accomplishing a goal. In effect, objectives can be thought of as sub-goals. Table 2 lists 17 objectives for big data, deep analytics and machine intelligence. Initiatives are specific action items that must be accomplished to achieve an objective. Table 2 also lists 54 initiatives for big data, deep analytics and machine intelligence.

Goal	Objective	Initiative
1 - Keep up with big data, deep analytics and machine intelligence technologies and capabilities, and advance LaRC knowledge and utilization of them	1.1 - Make big data part of existing LaRC technology considerations and research Impact: Chances of successfully leveraging big data increase as staff becomes aware of the capability being developed locally.	1.1.1 – Make big data a topic of consideration by the Chief Technologies and Chief Scientist offices
		1.1.2 – Make field part of strategic goal of either OCIO or LaRC
	1.2 - Educate LaRC staff on big data and capabilities and build a collaborative community Impact: Allows scientists to better understand how big data can assist them with their work and allow the capability to grow as staff becomes aware of latest improvements/technologies	1.2.1– Bring big data training and class on site or make it available through Satern online
		1.2.2 – Make big data a topic of monthly center briefings at LaRC and Virginia Air and Space Center (VASC), bringing in field experts as required
		1.2.3 – Identify internal users, SMEs to participate as initial core cadre and continuously seek to expand membership to other LaRC stakeholders

		1.2.4 - Build understanding and momentum of possibilities among scientists and engineers – Big Data Working Group; seminars; workshops
2 – Build a robust data intensive scientific discovery analytics capability and cognitive computing analytics to enable better science and engineering	<p>2.1 – Identify existing problems and opportunities that can be addressed</p> <p>Impact: Most effective way to show the potential impact big data will have at LaRC</p>	2.1.1 – Use existing use-case examples as a starting point for assessing new opportunities for pilots
		2.1.2 – Develop proposals for review, analysis and funding considerations, detailing the opportunities
		2.1.3 – Develop requirements documentation for each use case
	<p>2.2 – Assess big data tools and concepts and integrate them into future project analysis</p> <p>Impact: Development of more efficient algorithms and enables performance of better scientific research</p>	2.2.1 - Review and research applicable analysis tools and statistical and computational algorithms
		2.2.2 – Research how big data techniques can be part of the normal research and proposal process already being utilized
		2.2.3 – Work to modify the existing process to include big data techniques as part of the standard operating procedure (SOP).
	<p>2.3 – Begin pilot examples to assess results, demonstrate value and develop deeper understanding of needed techniques and architectures</p> <p>Impact: Shows the potential impact big data will have at LaRC</p>	2.3.1 – Develop Knowledge Assistant Pilot with IBM on the path to NASA Watson

		2.3.2 – Develop data mining pilots with nondestructive evaluation (NDE), Aeroelasticity and IT Security
		2.3.3 – Develop a data visualization and discovery pilot with strategic business opportunities
	2.4 - Establish an in-house big data, deep analytics and machine intelligence capability Impact: Enables the successful execution of big data pilots using a core team solely devoted to working the pilots	2.4.1 – Hire an initial big data scientist and big data programmer
		2.4.2 - Start building the expertise capability by hiring additional staff ~ one overall lead; additional data scientists; big data architects; machine learning /intelligence experts; advanced algorithm developer; students; establish points of contact (POC) in major mission organizations and disciplines
		2.4.3 – Develop new staff into a fully functional area of expertise
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	3.1 – Develop big data architecture and design blueprint to modularly scale to a data lake Impact: Facilitates successful, efficient achievement of a robust big data capability	3.1.1 – Have a Multi-Discipline Big Data Architecture Workshop facilitated by an architecture expert to determine the needs at LaRC
		3.1.2 - Investigate architecture platforms needed to make short-term goals a reality (Hadoop and other technologies, Amazon Web Services, etc.)
		3.1.3 –Develop design documentation to detail “how” each component will be implemented, what data sources will be supported, what methodologies will be used for transforming data to an acceptable format, how the components will be glued together, and performance and scalability of the application.

		3.1.4 – Consider each functionality as a milestone, and develop a timeline for each
		3.1.5 – Describe the basic functionalities the big data application will support, which at a minimum include: <ul style="list-style-type: none"> o Intelligent question answering capability o Predictive analysis o Categorization o Anomaly detection o Optimized parallel algorithm o Data mining, text analysis and data visualization o Features that support machine learning algorithms such as neural network, Bayesian network, regression analysis, K-means and Dirichlet algorithms
		3.1.6 – Define the following parameters for each functionality listed in 3.2.5: <ul style="list-style-type: none"> o Nature of user input o Function performed on user input o Desired output
	3.2 - Begin implementing new architectures and tools where possible Impact: Targets high impact areas first to show the value/impact of big data	3.2.1 – Begin purchases and integration of tools into the LaRC infrastructure
		3.2.2 – Convert existing systems and techniques (within IT infrastructure) to harness new capabilities

	<p>3.3 - Implement the big data knowledge base, access layer and user interface</p> <p>Impact: Operationalizes the capability</p>	<p>3.3.1 - Define knowledge base with a wide variety of data, including numeric, text, images, video, audio, structured, unstructured, semi-structured, human or machine generated</p>
		<p>3.3.2 – Begin work on implementing and integrating software components</p>
		<p>3.3.3 - Assess if the application is producing desired output stated in project specification documentation and make corrections/adjustments, as required</p>
		<p>3.3.4 – Develop and test a big data prototype and deploy it to a small group of beta testers to enhance software features</p>
	<p>3.4 – Begin developing an integrated cross-discipline data repository (data lake/knowledge base) that can be utilized for various analytics by different groups and for varied purposes</p> <p>Impact: Significantly improves cross-discipline collaboration and the ability to centrally store data and process algorithms without having to “move the data”</p>	<p>3.4.1 - Develop a long-term plan for a permanent data lake at LaRC</p>
		<p>3.4.2 - Understand various data sets and assorted formats and context – where, how much, value, etc.</p>
		<p>3.4.3 - Investigate architecture models available – demonstrations, discussions, consultants, etc.</p>
		<p>3.4.4 – Pursue agreements with scholarly information vendors to obtain meta data</p>

		3.4.5 – Begin large-scale implementation of a big data platform
	3.5 – Develop integrated environment to leverage/utilize data mining, HPC and M&S for better and faster science Impact: Aligns big data with major CDT focus areas/capabilities	3.5.1 - Engage and sustain relationships with HPC and M&S groups to understand the possibilities
		3.5.2 - Choose a pilot to model the required integration
4 - Ensure understanding and use of machine intelligence remains a long-term focus	4.1 - Keep up with long-term vision of machine learning and intelligence, with simulation of the human brain to identify possibilities for LaRC experimentation Impact: Facilitates continuous application of the best technology available to LaRC big data projects	4.1.1 – “Shape” the long-term vision by active participation and leadership of seminars and conferences (internal and external) focused on machine learning and intelligence
5 – Proactively pursue, utilize and leverage partnerships and collaborations with universities, federal research organizations, Department of Energy (DoE) labs and industry	5.1 - Identify, establish and leverage partnerships with universities, industry and government agencies Impact: Increases awareness of top of the latest technology/ research; enhances awareness of where to outsource work too difficult or expensive for LaRC to perform	5.1.1 - Establish partnership with Ames’ Data Sciences and Intelligent Systems Group. Build a formal relationship with identified areas/initiatives.

		5.1.2 – Utilize graduate students (Georgia Institute of Technology, Old Dominion University [ODU], William and Mary) to help with translational research and solutions to specific data mining use cases
		5.1.3 – Pursue partnerships with Defense Advanced Research Projects Agency (DARPA), National Science Foundation (NSF) and and Intelligence Advanced Research Projects Activity (IARPA). Actively participate to understand and leverage technologies being developed.
6 - Ensure buy-in at the grassroots level, resource availability and investment prioritization for building and enhancing a big data, deep analytics and machine intelligence capability	<p>6.1. - Develop a flexible and innovative funding strategy with inputs from stakeholders, users and Office of the Chief Financial Officer (OCFO)</p> <p>Impact: Provides leadership with a clear and detailed understanding of what steps are necessary to stand up a big data capability</p>	6.1.1 – Develop implementation plan, to include a funding approach for the first three years for big data, deep analytics and machine intelligence at LaRC
		6.1.2 - Demonstrate and communicate cost benefits/value to stakeholders on a regular basis via briefings and progress updates
		6.1.3 – Identify potential resource/infrastructure investments that can be made due to savings/value gained from leveraged partnerships
	<p>6.2 - Ensure tie-ins to programs are accounted for to gain long-term commitment</p> <p>Impact: Ensures scientists fully maximize the use of big data in their work</p>	6.2.1 - Engage with program managers to gain a full understanding of their needs and constraints
		6.2.2 - From early pilots , demonstrate return on investment (ROI) to programs
		6.2.3 - Investigate and follow up on potential funding sources at headquarters level

	<p>6.3 – Develop metrics to ensure performance standards for big data tools, techniques and systems are defined</p> <p>Impact: Allows leadership at LaRC to fully understand and gauge success</p>	<p>6.3.1 – Develop metrics that specify expected baseline performance for specific system components</p>
		<p>6.3.2 – Develop metrics that capture:</p> <ul style="list-style-type: none"> ○ Static historical measurements ○ Quantitative return measurements (e.g., cost savings) ○ Qualitative return measurements (e.g., value of science/discovery outcomes) ○ Quantitative performance measurements o Qualitative performance measurements
	<p>6.4 – Collect and disseminate performance data to users and leadership</p> <p>Impact: Allows leadership at LaRC to fully understand and gauge success</p>	<p>6.4.1 – Develop capability to collect data for each metric standard in near-real time</p>
		<p>6.4.2 – Ensure system can self-identify performance gaps/shortfalls in near-real time</p>
		<p>6.4.3 – Develop dashboard to keep LaRC leadership informed on big data/project status/performance on weekly/monthly basis</p>
		<p>6.4.4 – Produce annual white paper that provides a comprehensive overview of big data performance</p>
		<p>6.4.5 – Develop capability to produce tailored reports for projects/users of big data</p>

Table 2 Goals, Objectives, and Initiatives

A3 Use Cases

To determine the value of a big data, deep analytics and machine intelligence capability, the Big Data Team developed 10 initial use cases that show the applicability of these technologies to LaRC organizations. The Big Data

Team is continuing to develop more in areas like Atmospheric Sciences Data/Climate Change, Air Traffic Control Lab (ATOL), Flight Simulators and Multimedia. The team interviewed scientists and engineers in areas such as NDE; Entry, Descent and Landing (EDL); Aeroelasticity; Airframe Noise Reduction; Scientific and Technical Information (STI); and IT Security to determine how big data and machine learning technologies can be applied to LaRC technical challenges to improve mission success and research.. Table 3 provides a brief description of each use case.

Use Case	Description	Impact
Aeroelasticity: Predicting Flutter Conditions	Flutter occurs when an airplane wing begins to vibrate more and more rapidly until structural damage or catastrophic failure occurs. The Aeroelasticity department uses wind tunnels to test wing designs and determine the conditions under which flutter is most likely to occur of wing designs. The department, which has decades of wind tunnel data containing a wide variety of observational variables, is interested in collectively data mining all these data sets to discover new datagenerated predictors of aircraft flutter.	Potential new insights from analyzing multiple wind tunnel data sets simultaneously
NDE: Reconstruction of Material Computed Tomography (CT) Scans	NDE analyzes materials to identify flaws or anomalies. For more complex materials, only a fraction of the data available can be evaluated--roughly 10 percent in some cases. This severely limits the analysis. A single highly effective algorithm that could find anomalies for any material type would give researchers the ability to create a more complete picture and increase their understanding.	New forms of analysis and the potential for insights that are currently impossible from completely manual analysis techniques
Turbulence Modeling	Currently, physics models of turbulence contain significant discrepancies between predicted behavior and observed behavior from real-world testing. Dr. Karthik Duraisamy is working with other professors at the University of Michigan and private businesses to create neural networks that can predict turbulence. LaRC could undertake two initiatives relative to Dr. Duraisamy's work. The first would involve LaRC providing turbulence SMEs to assist with analysis of the neural network approach. The second would entail LaRC investigating a regression approach to add the capabilities of error estimation, confidence and power that neural networks lack.	Achieve a better understanding of the parameters that predict turbulence and decrease the gap between current physics based models and observed data

Airframe Noise Reduction Physics-Based Simulation and Design	Physics-based equations are used to model airframe noise. The current process is extremely time intensive and requires significant resources. The current algorithm does not store any of the 1.2 terabytes of data produced, which prevents any kind of regression-like predictive techniques. Additionally, the algorithm takes months to run and the impact of predictors on the response variable has not been fully assessed. Goals for the project include the ability to store data generated by the algorithm every time it is run, without an increase in runtime or processing requirements. The team would also like to determine which variables could be removed from the algorithm to speed up performance.	Be able to store all data generated from the algorithm in order to perform better analysis using data from multiple tests and run the algorithm more frequently and at a reduced cost
Entry, Descent and Landing	Monte Carlo analyses are run and data is passed to each of the different disciplines for further evaluation. Changes are then made by each discipline, which can affect the analysis results for other disciplines. EDL would like to have a more interconnected set of disciplines and shared data capabilities to create improved Monte Carlo simulations and better system-wide understanding of the impacts of changes made by each discipline. EDL would also like to improve analysis efficiency, connect different disciplines, facilitate sharing among the disciplines and enable disciplines to participate in the entire process, if desired.	Create better designs, improve analysis efficiency, connect disciplines, and facilitate sharing among the disciplines
IT Security-Real-Time Cyber-Threat Detection	IT Security would like to provide as robust a system as possible to protect NASA's network from intrusion attempts, maintain networks at peak operating performance, and maintain high levels of user access to the network. IT Security would like to: <ul style="list-style-type: none"> • Build a capability to use real-time streaming network traffic to search for anomalous behavior on the network and have increased protection from advanced persistent threat (APT) type attacks • Undertake a correlation analysis between vulnerable equipment and anomalous traffic to identify the most likely targets for infiltration into the network • Increase its data warehouse capabilities to store network traffic for 365 days rather than 180 days to be fully compliant with Federal mandates 	Significantly reduce the threat of state sponsored APT attacks on LaRC and significantly reduce the amount of data that is compromised from APT type attacks
Connecting Researchers and Fostering Innovation	Researchers find information by searching STI data. The current process lacks automation, machine intelligence and the ability to connect researchers. The desire is to build a knowledge base with machine intelligence that enables seamless, automated discovery and dissemination of STI data. This would facilitate sharing, enable delivery of content and create communities of researchers through alerts, recommendations and questions and answers to improve efficiency and foster innovation.	Improve researcher efficiency freeing up time to perform other tasks and connect researchers via communities

NASA Knowledge Assistant	A significant amount of time is spent mining for knowledge, usually in a manual fashion. Vast amounts of data sit unexplored, with potential scientifically significant connections and insights lost. A knowledge assistant would serve as a virtual colleague and help create communities among disciplines, resulting in inter-discipline and crossdiscipline innovation. The goal is to achieve a NASA Watson-like capability.	Quickly identify key trends, emerging experts and expert networks; summaries, alerts, recommendations, non-obvious relationships and intuitive visualization of results
Financial Systems Integration	Current work on the LaRC Information Technology Enhanced Services (LITES) contract has created a single repository of agency and LaRC budgeting tools with descriptions, a data dictionary and instructions on how the tools interface with one another. A system analysis revealed many tools were built in individual “silos” that needed better integration to provide greater efficiency. As a result, there is a need to build a data warehouse where all necessary financial information can be downloaded for daily analysis and queried. Users also need to perform “discovery analysis.” The desire is to do this for LaRC-specific tools, and then provide recommendations for Agency-level tools.	Significantly reduce time spent data mining various center and agency tools
Multivariate Data Analysis	Very few of LaRC’s research efforts or M&S is in small dimensions. However, the need to understand the relationships between variables or to demonstrate to others what is occurring between variables is still necessary. This creates a problem with traditional visualization techniques that can only display a few variables at once. New or better methods for visualizing relationships between large numbers of variables are needed.	Increased understanding of relationships between variables and improved ability to explain relationships to others

Table 3 Use Case Descriptions

A4 Knowledgeability

The Big Data Team gathered a large amount of research information during the development of the strategy and roadmap. This included whitepaper reports from Microsoft and Gartner as well as journal articles to shorter articles on the use of big data in social media, finance, medicine and science. The research revealed social media and finance are already heavily leveraging big data on a daily basis to predict consumer marketing potential or the impact of current events on the stock market. Likewise, the field of medicine is applying big data analytical techniques to a wealth of sensor and patient data. Many of these medical research efforts show significant potential to better predict patient complications resulting from surgery and adverse reactions to medications as well as proper treatment recommendations than doctors are currently able to do. Overall, this shows significant potential for LaRC to leverage similar capabilities for scientific discovery.

Additionally, the team brought in distinguished speakers from universities and colleges, including the University of Michigan, Georgia Institute of Technology, Iowa State University, William and Mary and ODU. Iowa State, in particular, is actively pursuing the use of the fourth paradigm to identify new types of materials, instead of simply seeking to improve upon the manufacturing process of known materials. Georgia Institute of Technology is working on advancing the visualization capabilities for M&S efforts. As a result, the Big Data Team is working with both of these universities, along with the likes of NASA Ames and Goddard Space Center to build some of its first partnerships in big data.

The team held discussions with corporations like Booz Allen Hamilton, in the area of data science and the notion of the data lake; IBM in the area of text analytics and content analytics; and AMA in more general areas of big data analytics. Team members also participated in the 2013 American Society of Engineering (ASE)/Institute of Electrical and Electronics Engineers (IEEE) International Conference on Big Data in Washington D.C. with speakers from Carnegie Mellon and Johns Hopkins University. Insights derived from this research include:

- Big data challenges are not the same across all domains; i.e., challenges in finance, social media and science are all different.
- Big data capabilities in the sciences are not as advanced as other fields like finance, medicine and social media; medicine is the area where several applications are evolving.
- The volume of big data is not what is most important. It is the analytics of big data that provides the most value to organizations.
- Big data is capable of providing new insights and understanding into the world of physics, the notion of the Fourth Paradigm.
- For science it is not just about the three Vs of big data, but also the fourth V (veracity); the complexity of science is the challenge for LaRC. In particular, engineering and physics often require solving very complicated mathematical equations, which means programmers need a deep understanding of how to solve such equations and do so efficiently.
- To show a value proposition, start with clear value-based goals, use cases and pilots.
- Do not throw away data; keep it in raw form to avoid losing potentially critical information.
- Big data analytics technologies and tools are increasing with many open-source options.
- Data capture and management are key pieces that cannot be ignored.

A5 Research and Partnerships

Big data, deep analytics, and machine intelligence technologies can be force multipliers on par with M&S to ensure and enhance LaRC mission success. LaRC seeks competitive advantage by being aggressive, innovative and early adopters of these transformational capabilities. To do that effectively, LaRC must leverage ongoing research initiatives and partnerships to be a smart adopter and implementer of these emerging technologies to best meet specific scientific and engineering data and knowledge mining challenges.

Research initiatives and potential partnerships LaRC should investigate are described below.

A5.1 Federal Research Initiatives

The Federal government, including the White House, is currently conducting research in data mining techniques, algorithms, image and video processing, architectures and large-scale dataset analytics. A comprehensive listing is in Appendix E. Following is a description of several key research initiatives with potential partnership value that LaRC will investigate in the next 1-2 years.

XDATA by DARPA: DARPA's XDATA program is designed to develop computational techniques and software tools for processing and analyzing the vast amount of mission-oriented information for defense activities. Funded for \$25M annually for four years beginning in 2012, the program seeks to develop computational techniques and software tools for analyzing large volumes of semi-structured and unstructured data.

BIGDATA by NSF: Big Data: Core Techniques and Technologies for Advancing Big Data Science & Engineering (BIGDATA) is a joint solicitation between the NSF and National Institutes of Health (NIH) that aims to advance the core scientific and technological means of managing, analyzing, visualizing and extracting useful information from large, diverse, distributed and heterogeneous data sets. The effort began in 2012 with funding for small projects (\$750,000 for three years) and mid-scale projects (\$1.25M to \$5M for five years). The main objectives are to promote research into data management, collection and storage and support new approaches to data analytics to gain knowledge from largescale databases.

SDAV Institute by DoE: Scalable Data Management, Analysis and Visualization (SDAV) Institute is a Department of Energy (DoE) effort to deliver end-to-end solutions that range from managing large datasets as they are being generated to creating new algorithms for analyzing the data on emerging architectures. The \$25M five-year initiative aims to develop a way to interact with data as it is being created in a simulation. This technique would allow researchers to monitor and steer the simulation, adjusting or even stopping it if there is a problem.

Artificial/Machine Intelligence by NSF: NSF awarded a \$25M, five-year grant to Harvard and MIT to study how the brain creates intelligence and how that process can be replicated in machines. The grant will be used to create a Center for Brains, Mind and Machines. The quest for the basis of intelligence is an ancient one, bolstered in recent years by the ability to create machines that have domain-specific abilities, such as [Google's self-driving car](#), IBM's [Watson](#) or Apple's [Siri](#). "But we are still some way from understanding the broad basis for human intelligence," notes Dr. L. Mahadevan, Professor of Mathematics, Evolutionary Biology and Physics at Harvard. "This new center will refocus our collective efforts at trying to solve this question from multiple perspectives."

A5.2 Partnerships

Innovation in industry and ongoing research within academia are major factors pushing technology forward. LaRC recognizes this and is working to position itself as a technology "adopter." This necessitates leveraging efforts underway in the universities, commercial sector and at other government agencies. LaRC has the opportunity to more rapidly develop future capabilities. In exchange, LaRC can provide partners unique intellectual capital,

including unique use cases and, world-class expertise in specific scientific disciplines. In an era when Federal budgets are increasingly vulnerable, developing partnerships is essential for LaRC to continue to remain relevant in a fiscally realistic manner. LaRC is currently pursuing partnerships the FUSE and XDATA programs as well as the IBM Watson group.

A6 Technology and Architecture

To get a clear picture of how LaRC can leverage big data, deep analytics and machine intelligence in the future, an understanding of tools, infrastructure and technological considerations shaping the future is necessary.

A6.1 Deep Analytics and Data Visualization

Deep analytics is an application of sophisticated natural language processing and machine learning algorithms to large corpus of knowledge to obtain insights, trends and answers to specific questions. Specific examples deep analytics include IBM Watson and the Wolfram Alpha computational knowledge engine. An outcome of the growth of big data is the need to represent data visually to ensure clear communication. Visualization can help explain analysis results, model the evaluation or assist in assessing data quality and selecting options. Data visualization displays can include modeling results of algorithms and techniques, volume of information, multi-resolution methods and interaction techniques and architectures. The growing need for visualization has led to the emergence on a new field--visualization analytics, which is the science of analytical reasoning facilitated by visual, interactive interfaces.

A6.2 Data Mining and Data Discovery

In the world of big data, data mining and data discovery are rapidly becoming a synonymous concept often referred to as the Fourth Paradigm, which is the progression from developing a hypothesis and then testing that hypothesis to letting the patterns and correlations in data suggest what the hypothesis should be.

A6.3 Machine Learning and Machine Intelligence

Machine learning is an artificial intelligence discipline geared toward the technological development of human knowledge. Machine learning allows computers to handle new situations via analysis, self-training, observation and experience. Machine learning facilitates the continuous advancement of computing through exposure to new scenarios, testing and adaptation, while employing pattern and trend detection for improved decisions in subsequent, though not identical, situations. Historically, machine learning is the development and advancement of systems that can learn from data. The goal is to build an artificially intelligent computer system capable of rivaling human capabilities, i.e., an intelligent machine. The most well-known example is IBM's Watson system. Watson is a complex analytics system that can process natural language input using parallel processing servers. The system has not yet reached the level of machine intelligence, but it is the closest any research effort has come to date. In the future, such a computer system would perform most of the work done by humans during the analysis process.

A6.4 Architecture and Infrastructure

Big data analysis requires a robust infrastructure and architecture (cloud computing). This necessitates having adequate resources, in terms of funding and personnel skill sets, to build the foundation for a big data capability. Organizations that lack those resources often look externally. This has allowed companies like Amazon to develop and grow a substantial capability to meet that need. Today, Amazon has a market share of nearly 80 percent. Cloud computing is a method of addressing the physical storage requirements of big data. The cloud is typically a remote computer system available through the Internet that provides a resource or capability for a fee. Cloud computing is a way to offload functions or needs to a remote system versus covering the cost and overhead locally.

Cloud services are expected to continue to become more affordable, and the use of the cloud will become a mainstay of business. Companies that use cloud resources are expected to grow from 27 percent today to 43 percent in five years. Cloud computing will help spur the implementation of increased security to address general public concerns; relational model data warehouses will become obsolete. A major idea for big data in the near future is the “data lake.” The idea is to store all data into a single collection, instead of silos of relational databases. In this model, data is tagged as it is ingested, instead of being deconstructed into a fixed relational model. When data is accessed, it is transformed into the format needed. This approach lends itself to analysis defined by need rather than collection requirements. Today, a data lake requires that data be in a single location; large data sets do not work well in a distributed configuration. However, this limitation could be mitigated in the future with faster data transfer rates.

Figures 3 and 4 show conceptual big data architectures. Key layers are described below:

- The lowest level, **Infrastructure**, harnesses the cloud to receive and store raw data in the various “Not Only Special Query Language” (NoSQL) and Hadoop architectures.
- As part of **Data Management**, each part of the data is tagged for identification as it is included in the data lake.
- Data is processed and analyzed in the **Analytics & Services** layer where various tools access, mine and analyze data using machine learning techniques and analytics tools.
- The layer seen by the end user is **Human Insights & Action**, where tools used to visualize present the data to the analyst and interfaces allow data manipulation.

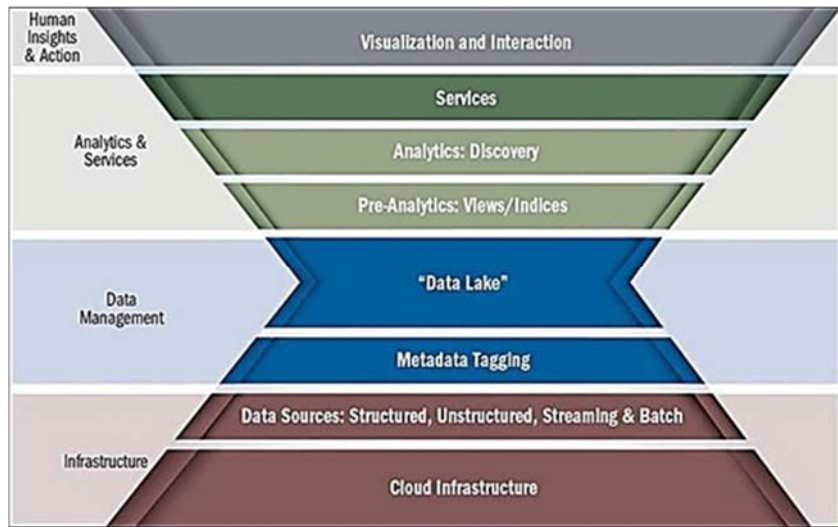


Figure 3 Big Data Architecture; courtesy Booz Allen Hamilton

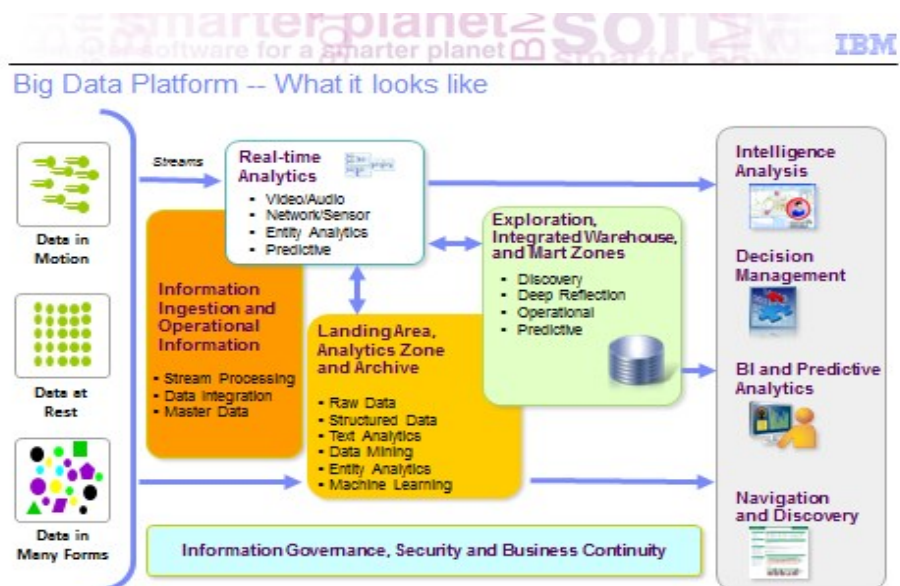


Figure 4 An alternate big data architecture; courtesy IBM-LaRC Knowledge Assistant Pilot.

A7 Workforce Skills

For LaRC to successfully implement big data, deep analytics, and machine intelligence as part of its overall CDT strategy, targeted investments in specific workforce skills will be necessary. Based on the Big Data Team's research, the following critical skills are needed:

- **Leadership** – Requires an innovative thinker and leader capable of driving change; brainstorming; formulating solutions; leading the program; ensuring value-based outputs; leveraging partnerships; and interacting effectively with scientists, engineers and LaRC leadership.
- **Big Data Architect** – Requires extensive experience in big data technologies and experience using multiple computer languages, building large scale distributed data processing systems/applications or large-scale Internet systems. Also requires the ability to work closely with scientists and LaRC leadership.
- **Analysis** – Requires the ability to interface with SMEs; formulate and define use cases, project scope and objectives; gather and analyze information; perform analytical extract, transform and load (ETL); standardize analytical data; foster automation initiatives.
- **Data Science** – Requires a solid foundation in computer science, modeling, statistics, analytics and mathematics coupled with strong business acumen and the ability to communicate findings to leadership, IT specialists and scientists.
- **Computer Science** – Requires the ability to develop algorithms and programs; use advanced analytics, quantitative analysis and data mining techniques; and develop customized big data and analytic solutions leveraging appropriate software tools
- **Machine Learning/Machine Intelligence** – Requires expertise in machine learning techniques like neural nets, support vector machines and artificial intelligence, as well as the ability to work closely with scientists to apply appropriate techniques to solve problems.
- **Statistics and Computational Mathematics** – Requires the ability to design statistical models; apply appropriate statistical tools and techniques; provide statistical, mathematical and data visualization capabilities; and work closely with scientists.
- **Application/System Administration** – Requires the ability to configure, monitor and administer the application software, knowledge base and IT infrastructure as well as address and ensure data security.

The range of required skills is diverse but also includes several common functions and expertise. These include:

- Familiarity with Hadoop and MapReduce
- Advanced algorithm development and statistical modeling
- Domain expertise
- Ability to parse data sets, mine data for patterns, filter and organize data and acquire new data sets

- Data modeling warehouse and unstructured data skills
- Knowledge of analytics, ETL, visualization and machine learning algorithms

A8 Findings, Key Recommendations, and Actions

While the goals, objectives, and initiatives outlined previously provided the framework for developing the 12 key recommendations, the recommendations were also informed by several internal and external findings. A recent decision by LaRC leadership to fund incubators dedicated to autonomy and nanomaterials also provided an avenue to develop big data capability in support of these efforts. This recent development also informed the 12 key recommendations.

A8.1 Findings

Several findings are shaping the way LaRC develops big data, deep analytics and machine intelligence capabilities in the future. External findings include the following:

- Analysis of big data is a combination of technologies, natural language processing, machine learning, etc. that have matured to the point that they enable new insights and allow better strategic decisions to be made.
- Big data provides an opportunity to find insights from data and content, and answer questions previously considered beyond reach.
- The challenges posed by big data will not be easy to resolve, but they are the next step in how to better understand the world and make better strategic decisions.
- A lack of sufficient numbers of big data scientists and machine learning/machine intelligence experts will result in their skills being highly sought in the marketplace.
- Eighty percent of the effort to implement a big data capability is in extracting, moving, cleaning and preparing the data, not actually analyzing it.

Internal findings including the following:

- LaRC, while possessing a huge amount of data, does not have an adequate archival retrieval capability.
- LaRC does not have sufficient numbers of trained staff with the skills required to successfully implement a big data, deep analytics and machine intelligence capability.
- Use cases are very heterogeneous, discipline specific, and require a deep understanding of the underlying physics of the problem, but can be efficiently tackled by development through pilots.
- There are two necessary big data capability paths for LaRC—1) a text and content analytics and deep question and answer (Q&A) capability, and 2) a data mining and analysis capability.

- LaRC faces an era of decreased budgets and increased competition for resources.
- The Big Data Team needs to have clear goals that provide high impact value to LaRC.
- LaRC must merge its physical testing capability with a more cost-effective M&S approach to science and engineering to maximize its scarce dollars and accelerate the development of missions from concept to flight. LaRC will need to aggressively collaborate internally (SMEs, Systems Analysis and Concept Directorate [SACD], Research Directorate [RD], Office of Strategic Analysis, Communication and Business Development [OSACB], Business Development [OSACB], Aeronautics Research Mission Directorate [ARMD], Chief Scientist, Chief Technologist, Chief Engineer, etc.) and externally (academia, industry, government, etc.) to build a capability that delivers the potential cost savings, accelerated development and discovery potential inherent in big data.

A8.2 Key Recommendations

LaRC must be prudent in how it implements big data, deep analytics and machine intelligence. The recommendations emphasize pilots, research and partnerships over capital intensive technology acquisition. Investment in personnel with specific big data skill sets will be required, but some of this cost will be offset by the use of contractors and student interns. Providing big data training to current LaRC workforce members may also mitigate costs. Table 4 lists the 12 key recommendations for developing a big data, deep analytics, and machine intelligence capability.

12 Key Recommendations	Near-Term	Mid-Term	Long-Term
	2014-2018	2019-2024	2025+
R 1 – Educate and promote the value of big data through seminars and workshops by experts and LaRC working group to foster the understanding of its value and use by mission organizations. (Links to Goal #1)			
R 2 – Understand incubator needs and incorporate them with deep analytics and machine learning pilots and capability; Demonstrate feasibility and add value to incubator success. (Links to Goal #1 , Goal #2 and Goal #6)			
R 3 –Build a big data and machine intelligence team, including data scientists, statisticians, algorithm developers, machine learning expert and comprised of civil service employees, contractors and students. (Links to Goal #2 , Goal # 4 and Goal #5)			

<p>R 4 – Develop and implement a data-driven scientific discovery capability; start with small-scale and highvalue pilots: NDE images, aeroelasticity data and cyber security. (Links to Goal #1, Goal #2 and Goal #3)</p>			
<p>R 5 – Develop an IBM Watson-like cognitive computing capability with deep analytics for research and question and answer Q&A for design; begin with pilots, including the Knowledge Assistant Pilot in progress. (Links to Goal #1, Goal #2 and Goal #3)</p>			
<p>R 6 – Identify and establish partnerships with universities, government and industry; leverage their expertise for LaRC’s big data capability and participate in research when possible. (Links to Goal #1 and Goal #5)</p>			
<p>R 7 – Develop a data capture and management capability for automatic capture of data with context, including meta data standards and tagging, real-time uploads and ingests; start with a pilot. (Links to Goal #2 and Goal #3)</p>			
<p>R 8 – Develop a big data architecture capability; Research and understand technologies, tools and architectures and incorporate learnings from the pilots. Start with Hadoop and cloud pilots. (Links to Goal #1 and Goal #3)</p>			
<p>R 9 – Keep machine intelligence as a North Star goal by actively researching state-of-the art developments, attending seminars /conferences; developing partnerships; and pursuing pilots with the Massachusetts Institute of Technology (MIT). (Links to Goal #4 and Goal #5)</p>			
<p>R 10 – Develop in-situ data analysis with M&S data and implement the capability of HPC, big data and M&S working together; start with pilots. (Links to Goal # 2 and Goal #3)</p>			

R 11 – Develop operational capability for virtual colleagues, experts and intelligent agents; start with pilots. (Links to Goal #1, Goal #2 and Goal #3)			
R 12 – Define and develop metrics for big data capabilities to demonstrate and communicate value to end users and leadership. (Links to Goal #1 and Goal #6)			

Table 4 Overall Roll-Up of 12 Key Recommendations and Projected Timeline

A8.3 Actions

An outline of the actions and outcomes for both Phase I (2014-2017) and Phase II (2018 – 2020) can be found in Figures 5 and 6, below.

Action	Outcome
Implement Deep Content Analytics to achieve a Knowledge Assistant: <ul style="list-style-type: none"> • Prototype with Incubator areas – Autonomy & Carbon Nano Tubes • Full capability with 1 discipline & Expand to 2-3 core disciplines • Establish Knowledge base architecture 	Researchers able to keep up, digest, and make sense of global knowledge quickly, enabling better ideation <ul style="list-style-type: none"> • Identify core ~20% of the body without reading thousands of documents and articles • Discover emerging trends, non-obvious relationships & experts
Implement Data Intensive Scientific Discovery capability for 3-4 areas. Start with pilots: <ul style="list-style-type: none"> • NDE Images, Aero elasticity Flutter experimental data, & Systems Analysis Simulations; Carbon Nano Tubes modeling data • Establish Data Lake architecture 	Researchers able to do more complete and faster data analysis not currently possible <ul style="list-style-type: none"> • Enables near real time processing of all scientific data and key variables using machine learning algorithms • Provides better innovations/science and saves SMEs time
Establish partnerships (Federal , Universities, Industry) <ul style="list-style-type: none"> • NASA LaRC (ASDC) , Ames and GSFC • UVA, ODU, Ga. Tech, Carnegie Mellon & MIT • DARPA , IARPA and NIST • Google and IBM 	NASA LaRC can leverage external expertise <ul style="list-style-type: none"> • Ames - data science; ASDC/GSFC - big data ; NIST standards • IARPA – scientific content analytics, DARPA- data analysis • UVA Data Science Institute; GT; CM – students and research • MIT, Google and IBM - cognitive computing, m. intelligence
Conduct education and outreach initiatives <ul style="list-style-type: none"> • Seminars and workshops (working with universities) • Cross organization and discipline working groups 	Maximize big data analytics value for mission <ul style="list-style-type: none"> • Enable propagation into organizations and projects • Share learning and best practices
Research machine intelligence <ul style="list-style-type: none"> • Begin pilots - Ex: Knowledge Bot for systems design simulations 	Develop deeper understanding of MI applications <ul style="list-style-type: none"> • Develop focused pilots for our use
Hire or obtain expertise (7 FTE) <ul style="list-style-type: none"> • Computer scientists; algorithm developers, machine learning,... 	Big data capabilities available for the center <ul style="list-style-type: none"> • Infuse capabilities into mission organizations

Figure 5 Phase I Actions

Action	Outcome
Data intensive scientific discovery capability in core disciplines	Ability to mine the both computational and mod sim data using discipline based algorithms for better science
Develop in-situ analytics of real time data to combine with mod-sim and high performance computing	Highly integrated real time modeling and simulation with big data and machine intelligence
Automate scientific data capture, tagging and integration for effective data management	-Rapidly & organically capture, tag and store newly generated data.
Fully linked knowledge base of global and NASA multimedia and multilingual information and data	Develops the critical foundation for collaborative data/knowledge ecosystem. Develop new research areas and systems significantly faster than before
NASA Watson & Intelligent Agent pilots Achieve Agency Level vision and buy-in	Ability to sift through millions of articles to find those of true value and able to answer design questions; Critical for the vision VRDP to be reality
Beyond: <ul style="list-style-type: none"> •Integrate an IBM-Watson-inspired “Automated Research/Design Assisatnt ” •Transform LaRC’ s ideation, innovation, and invention methods, to include machine-assisted techniques •Virtual Research/Design Partner 	New paradigms of scientific work. Enable NASA employees to achieve significantly greater scientific and engineering discoveries and systems innovation and optimization

Figure 6 Phase II Actions

A9 The Way Ahead

The next step for big data, deep analytics, and machine intelligence is to conduct small-scale pilots to demonstrate the value and need for big data and deep analytics capabilities. The approach must be congruent with CDT initiative. Current SMEs/disciplines interested in being early adopters of a big data capability include NDE, Aeroelasticity and Cyber Security. The Big Data Team plans to work on several pilots with the goal of using them to show leadership and scientists across LaRC the value a big data capability brings. The Big Data Team will also work with incubators and disciplines to develop larger-scale pilots that directly impact LaRC’s virtual capabilities and incubator research.

A9.1 Organizational and Cultural Changes

In the future, LaRC must adopt a more collaborative approach to research, experimentation, capability development and discovery. In the past, science was done within discipline specific lanes, with information often not being shared. A big data capability allows better M&S integration and provides an environment where data is available across all disciplines. Doing this will require meta data tagging up front to ensure research and experimentation information is readily available and sharable. Moving forward, the Big Data Team needs to be synchronized with HPC and M&S to ensure the big data capability is fully maximized.

A9.2 Funding Considerations

LaRC recognizes Federal budgets are not likely to grow in the future. The emphasis on pilots, research and partnerships over capital-intensive technology acquisition reflects this reality. The pilots will be designed to be implemented in the highest possible impact areas, while keeping the small budget in mind. Investment in personnel with specific big data skills sets will be required to successfully achieve the results early discipline adopters will expect. However, other needs can be met via contractors, who will come in for specific periods of time, student interns, and by providing big data training to the current LaRC workforce.

B: Progress Made to Date (2014-2017)

B1 Redefinition of CDT Strategy

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.

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.



Figure 7 Technology Progression across the Engineering Community.

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.

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.

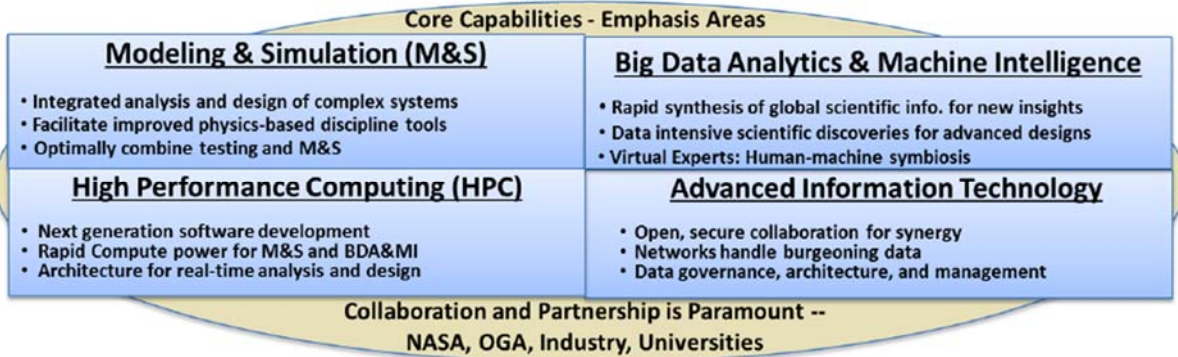


Figure 8 Comprehensive Digital Transformation Core Areas

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.

B2 Vision, Roadmap, Recommendations, and Actions

The vision for big data, deep analytics and machine intelligence is to enable LaRC to discover “unknowns” and deliver previously unimaginable capabilities by applying these transformational technologies as force multipliers for scientific and engineering discoveries and systems innovation and optimization. Achieving this vision will provide a number of tangible benefits. These include cost savings resulting from the use of more SBES and less physical testing to enable LaRC to be more competitive and innovate in providing transformational aerospace technologies. Another major benefit will be helping SMEs analyze more data, doing it faster and recognizing new patterns in data not feasible before. This will improve scientific discovery and engineering designs and allow scientists to spend significantly more time performing analysis rather than waiting on algorithms.

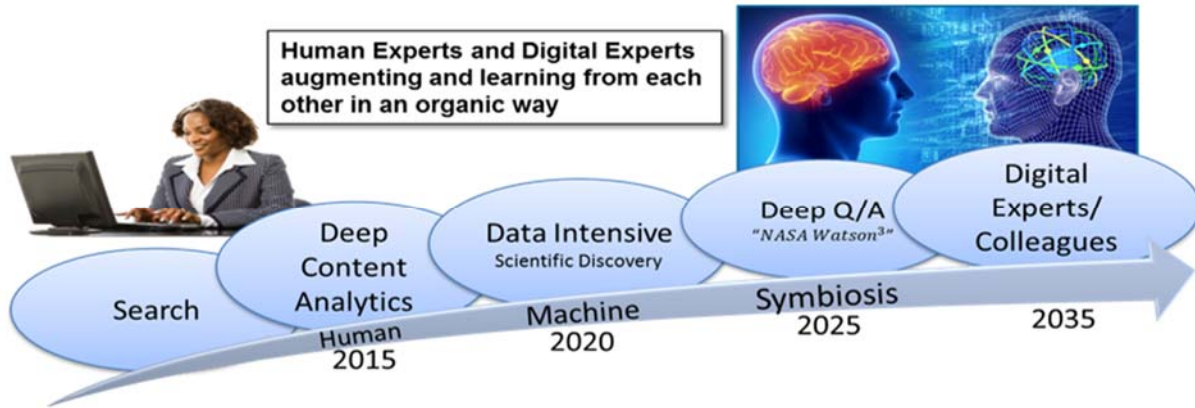


Figure 9 Vision for BDAMI at NASA Langley

A high-level overview of the roadmap and associated timeline for development of the big data, deep analytics and machine intelligence capability at LaRC is included below. The overview was generated based on the team's development of specific goals, objectives and initiatives, (as listed in Table 1 of Section A.) These were then distilled into a list of key recommendations to showcase the most critical needs for developing the capability.

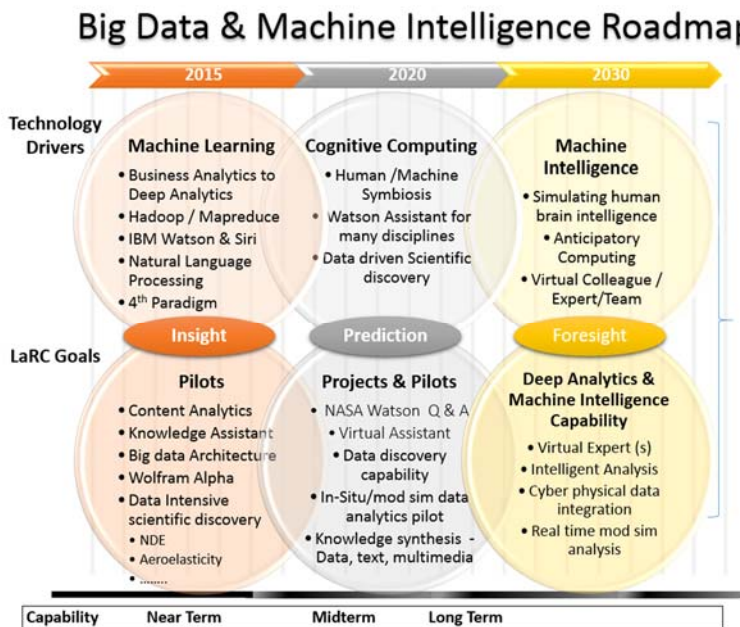


Figure 10 BDAMI Roadmap

Over the past two years, substantial progress has been made in each of the areas identified in the original vision. For reference, the key recommendations identified for completion in the Near-Term phase (2014-2017) included the following:

- **Key Recommendation 1:** Educate and promote the value of big data through seminars and workshops by experts and LaRC working group to foster the understanding of its value and use by mission organizations.
- **Key Recommendation 2:** Understand incubator needs and incorporate them with deep analytics and machine learning pilots and capability; Demonstrate feasibility and add value to incubator success.
- **Key Recommendation 3:** Build a big data and machine intelligence team, including data scientists, statisticians, algorithm developers, machine learning experts and comprised of civil service employees, contractors, and students.
- **Key Recommendation 4:** Develop and implement a data-driven scientific discovery capability; start with small-scale and high-value pilots: non-destructive evaluation (NDE) images, aeroelasticity data, and cybersecurity.
- **Key Recommendation 5:** Develop an IBM Watson-like cognitive capability with deep analytics for research and question and answer (Q&A) for design; being with pilots, including the Knowledge Assistant Pilot in progress.

Also, the proposed objectives from Phase I of the CDT Phased Actions and Strategy are included for reference in Figure 5. Each of the following sections discusses progress on these five actions.

Action	Outcome
Implement Deep Content Analytics to achieve a Knowledge Assistant: <ul style="list-style-type: none"> • Prototype with Incubator areas – Autonomy & Carbon Nano Tubes • Full capability with 1 discipline & Expand to 2-3 core disciplines • Establish Knowledge base architecture 	Researchers able to keep up, digest, and make sense of global knowledge quickly, enabling better ideation <ul style="list-style-type: none"> • Identify core ~20% of the body without reading thousands of documents and articles • Discover emerging trends, non-obvious relationships & experts
Implement Data Intensive Scientific Discovery capability for 3-4 areas. Start with pilots: <ul style="list-style-type: none"> • NDE Images, Aero elasticity Flutter experimental data, & Systems Analysis Simulations; Carbon Nano Tubes modeling data • Establish Data Lake architecture 	Researchers able to do more complete and faster data analysis not currently possible <ul style="list-style-type: none"> • Enables near real time processing of all scientific data and key variables using machine learning algorithms • Provides better innovations/science and saves SMEs time
Establish partnerships (Federal , Universities, Industry) <ul style="list-style-type: none"> • NASA LaRC (ASDC) , Ames and GSFC • UVA, ODU, Ga. Tech, Carnegie Mellon & MIT • DARPA , IARPA and NIST • Google and IBM 	NASA LaRC can leverage external expertise <ul style="list-style-type: none"> • Ames - data science; ASDC/GSFC - big data ; NIST standards • IARPA – scientific content analytics, DARPA- data analysis • UVA Data Science Institute; GT; CM – students and research • MIT, Google and IBM - cognitive computing, m. intelligence
Conduct education and outreach initiatives <ul style="list-style-type: none"> • Seminars and workshops (working with universities) • Cross organization and discipline working groups 	Maximize big data analytics value for mission <ul style="list-style-type: none"> • Enable propagation into organizations and projects • Share learning and best practices
Research machine intelligence <ul style="list-style-type: none"> • Begin pilots - Ex: Knowledge Bot for systems design simulations 	Develop deeper understanding of MI applications <ul style="list-style-type: none"> • Develop focused pilots for our use
Hire or obtain expertise (7 FTE) <ul style="list-style-type: none"> • Computer scientists; algorithm developers, machine learning,... 	Big data capabilities available for the center <ul style="list-style-type: none"> • Infuse capabilities into mission organizations

Figure 11 Actions and Outcomes for Phase 1 (2014-2017)

B3 Progress on Data Intensive Scientific Discovery (DISD) Projects

Implementing Data Intensive Scientific Discovery Capability

Each of the use cases identified in the 2014 strategy was assessed by the team, and substantial work occurred on four different domains. As part of the original strategy in 2014, the team worked to develop ten different use cases to show the applicability of these technologies to LaRC organizations. These initial cases were based on interviews with scientists and engineers across diverse disciplines including non-destructive evaluation, aeroelasticity, turbulence, and entry, descent and landing. Each case was selected in order to not only determine how big data and machine learning technologies might be applied to specific aerospace technical challenges, but also to better understand the correct methodologies to improve mission success. In the world of data science, data mining and data discovery form a synonymous concept often referred to as the Fourth Paradigm. Conceptually, this represents a shift from the traditional methodology of developing a hypothesis and then testing that hypothesis, to instead allowing patterns and correlations in data to suggest what a hypothesis should be.

B3.1 Nondestructive Evaluation: Automated Identification of Anomalies to Assess Structural Damage

In 2014, discussions with experts in Non-destructive Evaluation (NDE), which analyzes materials to identify flaws or anomalies, led to the identification that for many complex materials, only a fraction of the data available can be evaluated--roughly 10 percent in some cases. This severely limits the analysis. For less complex materials, there are algorithms in place. However, they are generally time intensive. In collaboration with SMEs, the team realized that these algorithms could be significantly improved by researching new approaches and looking for methods that would allow the algorithm to be parallelized.

B3.1.1 Overview

The current goal of this pilot is to develop techniques and algorithms to automatically detect anomalies during the non-destructive evaluation of materials including stainless steel, carbon fiber, and composites. This will significantly reduce SME analysis time and help experts discover additional anomalies that were previously undetected by visual analysis of images. The resulting tools will enable SMEs to design better material compositions and structures, and will help with innovative composite additive manufacturing using ISAAC. At Langley Research Center, researchers in the Nondestructive Evaluation Sciences Branch (NESB) use NDE techniques including ultrasound, thermography, and x-ray computed tomography (CT) to obtain information about the structures of various materials. Each of these techniques captures an abundance of data that currently must be analyzed by an expert human inspector to identify any internal anomalies (e.g. structural defects) of the material. This process is extremely labor intensive and automating this analysis would save inspectors significant time.

B3.1.2 Machine Learning & Statistical Techniques

In order to tackle this problem, we have utilized a number of different methods including novel statistical algorithms and state-of-the-art machine learning. Each method employed is explained briefly below.

Regression: Two-Dimensional (2D) Regression is designed for detecting anomalous pixels in an image with a 2D regression function. Each slice in the raw image was smoothed with a 41*41 average filter to remove noise. Each smoothed slice is fit into a 2D regression function, for which different slices could have different parameters. Then, the pixel values in each slice are replaced by residual values, which are the difference between regression value, and real value for each pixel. The histogram of residual values is calculated and all pixels that are far away from the center of Gaussian distribution are identified as the anomalous pixels. The crosshatch regression is also designed for identifying anomalous pixels. Each image slice is broken down into a series of one-dimensional signals, with each signal representing a single line of pixels in either the x or y direction. Every line of pixels in the image is represented in a signal. A regression-based analysis is applied to each individual signal to determine which pixels in the signal were outliers. A pixel that is found to be an outlier in both the x and y signal is considered to an anomalous pixel. A MATLAB application has been developed and delivered to the SMEs, which contains the 2D regression algorithm and the crosshatch regression algorithm. The 2D regression algorithm was implemented with C++ and MATLAB. A MATLAB wrapper was used for communication between the C++ functions and MATLAB. The crosshatch regression was written in R. Both the algorithms had been parallelized on image slices with the embarrassing parallel model.

Convolutional Neural Networks: A convolutional neural network (CNN) is a highly non-linear model that recently has achieved state-of-the-art results for many machine perception tasks including image recognition and membrane

segmentation. A CNN consists of several alternating layers that perform learned transformations to the input, ultimately mapping the input to a one-dimensional vector that can be classified with a traditional neural network. A combination Python, Caffe, and Lua/Torch were used to implement the methods for this work.

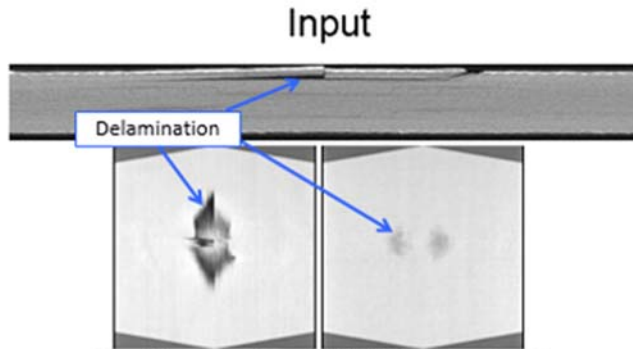


Fig. 1 Example slices of CFRP composite CT

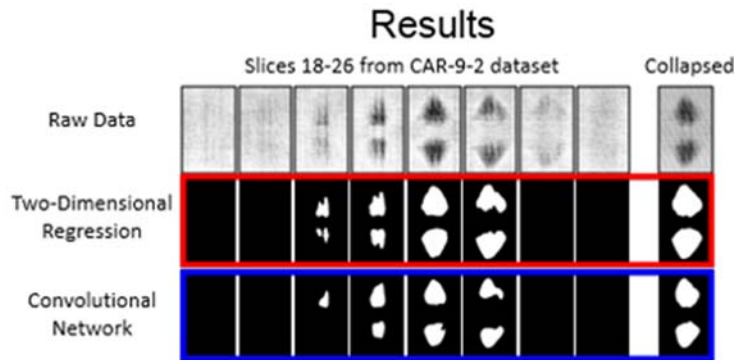


Figure 12 Example of a delamination in carbon fiber (top); results from both methods (bottom)

B3.1.3 Results

Both the 2D regression algorithm and crosshatch regression algorithm performed very well on the simulated and experimental data. Once trained, the CNNs were used to classify both simulated and experimental data. There are two major challenges for these two algorithms: increasing the accuracy on experimental data and screening anomalous pixels that are not delaminations. To solve the first challenge, we are working Gaussian fitting and a peak detection approach to dynamically set up the identification cutoff. For the later, we are developing a pre-processing routine to identify whether or not slices have delaminations with machine learning techniques. As for CNNs, we are currently investigating several approaches to increase the fidelity of the segmentations and to reduce the amount of computation required to produce segmentations. In particular, we are implementing an

encoder/decoder approach used to perform wound segmentation and incorporating information from multiple-scales in order to obtain more context when predicting each pixel.

B3.1.4 Next Steps

For this project the next steps include helping the SMEs to use the delivered algorithms for delamination detection on data sets and enhancing the algorithm with their feedback, implementing methods for matrix crack detection, fine tuning the CNN algorithm methodology, and leveraging work from experts at MIT.

B3.2 Aeroelasticity: Predicting Flutter from Aeroelasticity Data

In 2014, the team talked with the Aeroelasticity branch, which uses wind tunnels to test wing designs and determine the conditions under which aerodynamic events are most likely to occur for a large variety of wing designs. The branch has decades of wind tunnel data containing a wide variety of observational variables, and was interested in collectively data mining all these data sets to discover new data-generated predictors of aircraft flutter, a phenomenon that occurs when the wing of an airplane begins to vibrate more and more rapidly until structural damage or catastrophic failure of the wing occurs.

B3.2.1 Overview

The intent of this project is to use data science and machine learning methods to identify new ways of predicting aircraft flutter, by either forgoing the need to perform Fast Fourier Transforms (FFT) on accelerometer readings, or by identifying variables capable of predicting flutter other than traditional accelerometers along the tip of the wing of the aircraft and engine. The ultimate goal is to develop a real-time warning system for Subject Matter Experts (SMEs) conducting wind tunnel testing.

In the current state, SMEs rely on computational modeling to generate the predicted values of Mach (M) and dynamic pressure (q) at which flutter will occur. Wind tunnel testing is then conducted to determine the accuracy of these predications, with the intent of clearing the flutter envelope for a given aircraft model. During wind tunnel testing, a combination of SME expert observation and monitoring of sensor data is used to detect the onset of flutter in the model being testing. Regardless of configuration, the coalescence of modes in the frequency domain indicates the onset of flutter. Peak detection and tracking is a common method to observe this behavior. All aerodynamics that are observed in the frequency domain are also present in the time domain; thus, an approach that can identify flutter precursors in time domain signals would also be of value for SMEs. In the desired state, a “Flutter Assistant” based on cognitive computing will provide value-added insight into real-time system behavior. This will assist the SME in predicting the onset of flutter. Additionally, the tools incorporated in a Flutter Assistant will also have the capability of assisting with post-test analysis.

B3.2.2 Machine Learning and Statistical Techniques

When the Aeroelastic Flutter pilot began in the summer of 2014, initial efforts focused on applying frequency domain transforms to recorded accelerometer data and then using peak detection techniques to find salient modes of vibration. These initial peak detection techniques failed. In an attempt fit the accelerometer frequency data, high degree polynomials were fit directly to FFT information. This initial attempt also failed because the polynomial was often incapable of bending to fit sharp peaks in the data. Based on the lessons learned from these initial efforts at effective peak detection, the Piecewise Regression method was developed.

Piecewise Regression for Peak Detection: The Piecewise Regression Technique was initially developed during the Fall of 2014. The technique assumes that a predetermined number of mode peaks exist within the frequency domain for aircraft accelerometer signals. The Piecewise Regression attempts to fit $n + 1$ quadratics to FFT information where n is the total number of modes to be found. The quadratics fill the spaces bordering and between each mode. Furthermore, Piecewise Regression can be applied to multiple accelerometer signals at once using multivariate regression. The initial iteration of Piecewise Regression was written in the R Programming language using data exported from MATLAB. Given promising initial results and positive SME feedback, the Piecewise Regression was further developed and eventually fully converted to MATLAB beginning in 2015. The conversion to MATLAB did not change the underlying algorithm. It enabled dramatic increases in performance due to parallelization and elimination of redundant quadratic calculations. The conversion to MATLAB also enabled new visualization techniques. The reduced computation time enabled by the MATLAB implementation enabled batch processing and the evaluation of the technique across multiple tabs of wind tunnel test data. This, in turn, has enabled the quicker identification of areas where the algorithm could be further developed. Additional attempts at refining the algorithm included forcing quadratic minima into the space between the two modes which are being fit, and the development of a signal by signal technique which uses pruning and voting to determine salient modes. After more than a year of work, it was determined that there was no significant improvement in metrics that assessed the accuracy of the method when applied to both the S4T and TBW data. A robust methodology could not be established, and the performance of the algorithm on relatively simple data was less than desirable. In March 2016, the collaborative team of developers and SMEs determined that further work would be suspended, and lessons learned collected.

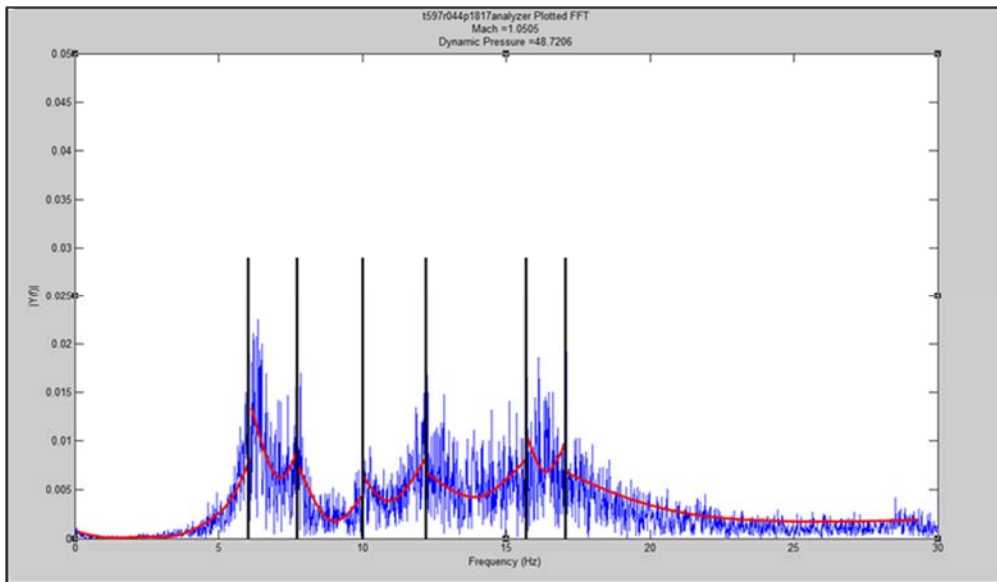


Figure 13 Piecewise Regression methods for detecting modal peaks in accelerometer signals

Time Series Motifs: Time series are an important class of temporal data for scientists at NASA Langley Research Center, as they represent one of the most common outputs from the various sensors and instruments used in

aerospace research. Characteristically, time series data is of a relatively large size, with a high dimensionality. Because the data is continuous, the series must be analyzed as whole, rather than being considered as an individual numeric field. This requirement means that similarity searches in time series are carried out via approximate methods. Pattern mining algorithms can identify repetitive subsequences or motifs in time series data collected from sensors, providing researchers with new insights into the dynamics of the systems they study. The Time Series Motif approach focuses on an unsupervised method of searching time series data from TBW testing. The methodology used in this pilot project belongs to a family of algorithms with the capability to mine unprocessed, time domain data in a fraction of the time required by many other methods. Regardless of the algorithm or technique selected, significant subsequences within the time series data are only of value to SMEs if they can be mapped to physics-related dynamics that occurred in the system being monitored by sensors or other instrumentation. By using such signatures for data classification, real-time streaming data and legacy datasets can be analyzed with machine learning algorithms to support the CDT goal of developing virtual helpers that will allow scientists and engineers to better focus on developing innovative solutions to complex technical challenges.

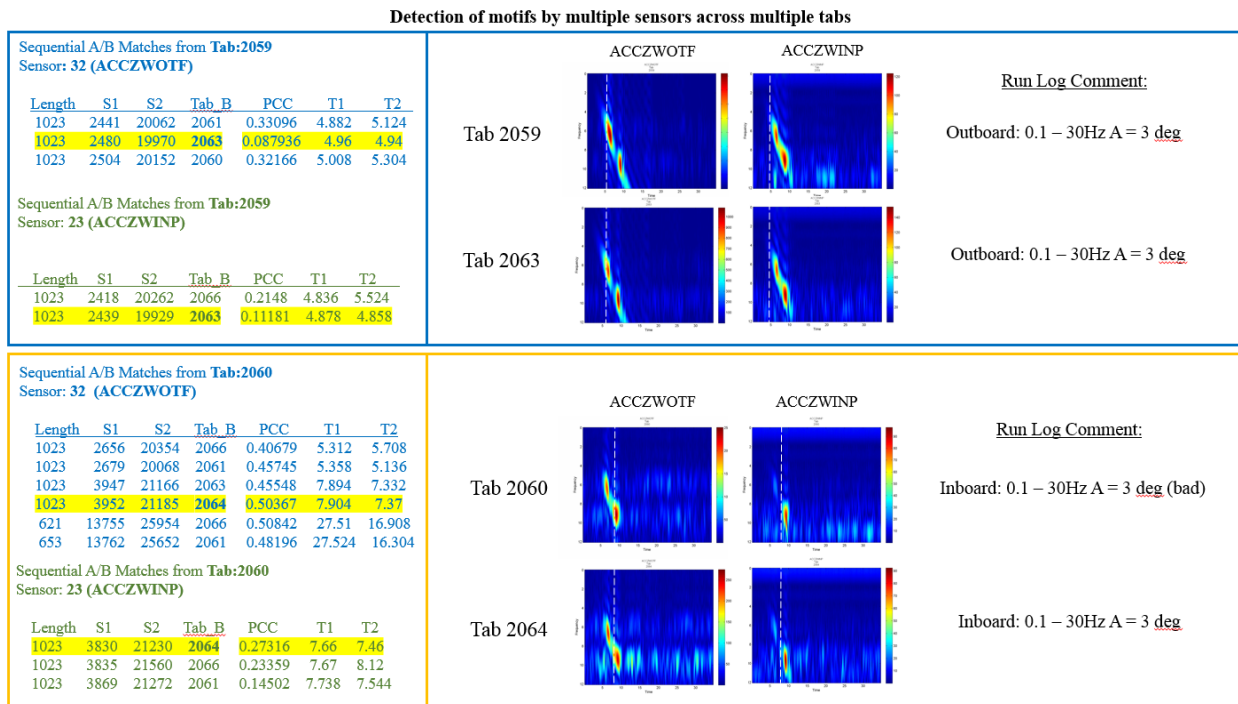


Figure 14 Detection of time series motifs in multiple sensors during wind tunnel testing

Regardless of the method selected, it is clear that the greatest benefit for SMEs will come from supervised machine learning tools that are trained on a valid dictionary of time series signatures which were efficiently mined from the time series data produced by wind tunnel sensors. While it is obvious that the validation of such motifs must come

from SMEs, this is not a trivial task. Because there has been no research on the analysis of flutter motifs, there is a complete lack of ground truth. The only way to overcome this obstacle is for SMEs to either review legacy data to confirm the presence of a flutter dynamic, or to conduct new wind tunnel tests in order to accurately label the aerodynamics occurring during testing.

B3.2.3 Results

In September 2016, the BDAMI team completed the first goal of the project and met with SMEs to hand off the tools that have been developed, along with documentation to ensure an effective knowledge transfer. The SMEs now have tools that provide them with the capability to mine unprocessed, time domain data for significant motifs (repeating patterns in the data) that could shed light on subtle characteristics that may be related to flutter precursors. Moving forward, the SMEs will take on the process of validating motifs in the data, and the BDAMI team will help to facilitate the task by providing training on the tools and methods.

B3.2.4 Next Steps

Additional paths toward the longer-term goal may include partnering with universities to refine and enhance the analytical methods. The team and SMEs will also be collaborating on a NASA TM that documents the methods, results, challenges, and lessons learned.

B3.3 Crew Cognitive State Monitoring and Detection

The goal of the Crew State Monitoring (CSM) project is to predict the cognitive state of pilots during simulated flight using machine learning models which have been trained from aircrew physiological features. Using physiological sensors such as electroencephalogram (EEG), electrocardiogram (ECG), respiration rate (RR), galvanic skin response (GSR), and eye tracking, our team, in conjunction with the CSM subject matter experts (SME), hopes to be able to identify unsafe cognitive states in aircrew real-time. The successful identification of these unsafe conditions could lead to more effective pilot training.

B3.3.1 Overview

The CSM SMEs have created a battery of experiments designed to engender specific cognitive states in aircrew test subjects while they complete a series of benchmark tests and fly a simulated aircraft. Each test subject is connected to the five physio sensors listed above. The data recorded by these sensors is used as input for machine learning models. The specific benchmark cognitive response is used as the predicted output of the models. The main cognitive states which our team focuses on predicting are low workload (normal state), channelized attention, diverted attention, and startle / surprise state. A wide range of features is extracted from the physio data and used in the machine learning models. For example, frequency domain transforms are applied to EEG channels, heart rate is calculated from ECG, and eyelid entropy are derived from eye tracking data.

B3.3.2 Machine Learning and Statistical Techniques

An ensemble of machine learning models has been developed for this project. Our team has trained deep neural network (DNN), gradient boosting, random forest, support vector machine (SVM), and decision tree models. Data preprocessing and machine learning training has been completed using MATLAB and open source Python software. Currently, we are in the process of training and testing 2 level ensemble models using 5 fold cross-validation. Multiple machine learning models are trained on the initial input features derived from original physio data. The output of these models are then used as meta-features: new learned features which are simply concatenated to the

original input feature set. A level 2 meta model is then trained off of the new dataset (the original features with meta features included). This is a standard machine learning technique utilized across many research domains for ensemble learners.

Figure 1. Ensemble model architecture.

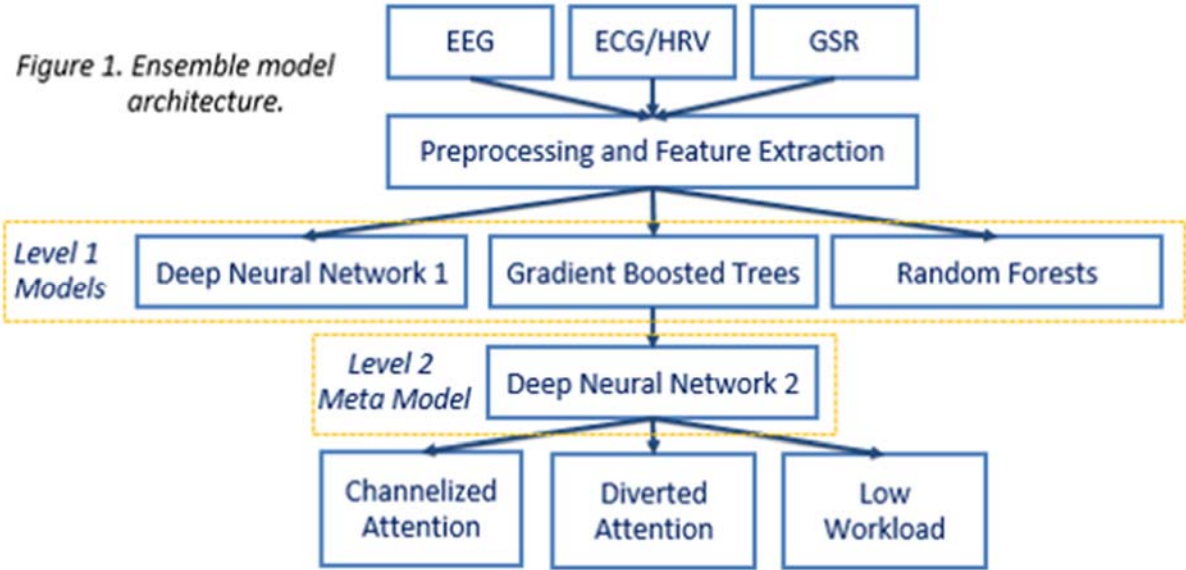




Figure 15 Methodology and techniques in use for CSM project (top); flight simulator training (bottom)

B3.3.3 Results

Our initial results so far look very promising. We are currently training separate machine learning models for each individual test subject. In the majority of subject models tested so far, our machine learning models correctly predict cognitive state with an accuracy of 90 % or higher. A major challenge to our team is that we have an abundance of physiological data which we could use to generate machine learning training features. We have currently only used a subset of all signals recorded. We would also like to explore the feasibility of training a single model which would work for all test subjects instead of multiple, individualized models. Finally, our team is preparing to participate in a new battery of experiments on test subjects. We would like to integrate our machine learning models into the CSM data collection equipment where cognitive state could be predicted and displayed real-time. Our next steps in this project will focus on hyperparameter tuning of models, integration of new signals, and preparation for real-time prediction during the next battery of tests.

B3.3.4 Next Steps

For this work, next steps will focus on extracting features from eye tracking data for incorporation into the machine learning models, evaluating software for automated feature extraction of key modalities in the data, incorporating low workload and nominal workload states for classification, and developing a Python prototype for demonstration to stakeholders.

B3.4 Rapid Exploration of Aerospace Designs

Modeling and simulation (mod-sim) implementations are routinely used in design space exploration as an early analysis tool for vehicle design. Many mod-sim codes, however, are computationally complex and incapable of rapid design space exploration. Machine learning (ML) offers a surrogate model capability which when used in conjunction with mod-sim, can facilitate a rapid exploration of the design space.

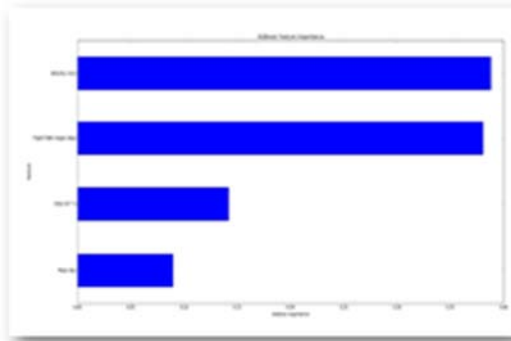
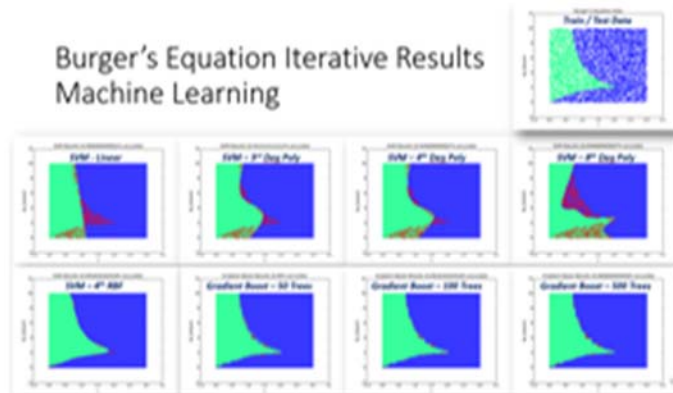
B3.4.1 Overview

The data for the design space could be from tests (e.g., wind tunnel) or modeling and simulations (e.g., CFD, finite-element analysis). Additionally, a READ web site serves as a mechanism for Langley staff to become familiar with machine learning algorithms, and contain links to other machine learning resources for additional information.

B3.4.2 Machine Learning and Statistical Techniques

Surrogate Modeling: As a proof of concept, surrogate modeling was conducted on NASA datasets to show the capabilities of ML in a surrogate modeling role. Burger's Equation is a differential equation often utilized for modeling complex systems. This dataset is a binary classification problem with only two inputs; multiple SVM and Gradient Boost models were used. The POST dataset was generated from mod-sim code utilized to simulate a Venus entry, descent and landing (EDL). It had a total of four continuous independent variables and was used to predict a single binary class indicating a successful landing; a gradient boost model with 100 trees was used with this data. The Multi Mission System Analysis for Planetary Entry (M-SAPE) dataset consisted of 6 continuous independent variables and 2 discrete independent variables. This model is used to predict Mars sample return (MSR) vehicle convergence; it has a binary discrete prediction value. Again, a gradient boost model was used. The prototype READ system has been implemented using PHP as a webfront-end, MySQL as a backend database, and Python scripts for data handling and machine learning. The design for this database schema is fairly light and flexible and can be altered during further development if required. The MySQL database will not hold any of the actual data used to train machine learning models. It will only be used to hold metadata on users, datasets, and trained models. The current READ web design is a prototype, which runs on local machines.

Burger's Equation (Gas Dynamics): The data presented a binary classification problem with only two inputs. Results for multiple SVM and Gradient Boost models are shown to the right.



POST: This dataset was generated from mod-sim code used to simulate a Venus entry, descent and landing (EDL). The plot on the left shows feature importance created during model training.

Figure 16 Examples of data analyzed with the methodology used in READ

Web Interface: The web interface is designed to serve as an intuitive means of uploading data, training machine learning models, and using the ML models for design space prediction visualization. The Welcome Page is the first page a user accesses when using the READ website. The Data Ingestions pages gives step by step directions on how to format data and then upload it to READ. After attempting to upload data, the end-user will be shown a page which details the success of the upload and then redirects the user back to the main menu (Welcome Page). The user will then be directed to a page with a radio button interface for selecting a dataset. All datasets which the user has uploaded into READ are displayed as possible selections. The user will select the dataset for which he or she would like to use for ML. Once a dataset has been selected, the user should then be brought to a model selection page. After selecting the dataset to use for training, the user can then select SVM hyperparameters for model training. Post training, a user will see the results of the model and can train a new model if he or she desires better results. Assuming the user is comfortable with the results, he or she can return to the Welcome Page and select visualization to use the new ML model for design space exploration. A user will first select the desired dataset, then the desired ML model, and then two continuous, independent variables to be used as the X and Y axes. If there are any remaining independent variables (continuous or discrete), the user will then select fixed values for these variables. Having selected inputs, the user can then click "Plot" to view ML visualization output. Users can return to the Visualization page to alter variable and axes settings in order to quickly explore the design space.

B3.4.3 Results

For the Burger's Equation data, the best performing models were SVM with radial basis function (RBF) and a gradient boost model with 100 trees. Both achieved an accuracy of 99.182 %. The POST dataset used a gradient boost model with 100 trees to achieve an accuracy of 100 %. Data from the M-SAPE model had performance of 99.520% using a gradient boost model. These results are encouraging, and provide a strong foundation for future development.

B3.4.4 Next Steps

Although machine learning has been employed for surrogate modeling by some scientist and engineers, many researchers remain unaware of its capabilities in this field. The primary goal of the web based Rapid Exploration of Aerospace Designs (READ) project is to provide a machine learning-based platform for rapid exploration of the design space with large number of attributes, and the web application provides an interface to socialize these concepts with SMEs across a variety of domains. The READ tool will be used as part of our outreach and use case development efforts.

B3.5 Turbulence Modeling

B3.5.1 Overview

Physics models of turbulence often contain significant discrepancies between predicted behavior and observed behavior from real-world testing. A number of new models have additional predictive parameters to correct these discrepancies; however these models have not performed as desired. In 2014, the team talked with Dr. Karthik Duraisamy at the University of Michigan, who was interested in creating a data-derived model that outperforms current physics-based models.

B3.5.2 Machine Learning and Statistical Techniques

Dr. Duraisamy worked with in collaboration with researchers at several universities and in industry to develop neural networks capable of informing turbulence closure models. His work involved aspects of experimental design, data decomposition, statistical inference, and machine learning with the goal of improving the predictive capabilities of turbulence models

B3.5.3 Results

Over the past two years, the Big Data team has followed Dr. Duraisamy's work, and has participated in several technical meetings with the Computational Aerosciences branch to assess the potential and future direction of his research. In May 2016 Dr. Duraisamy had detailed discussions with NASA Langley's Computational Fluid Dynamics (CFD) experts Dr. Mujeeb Malik, Dr. Chris Rumsey, Dr. Bala Ponnampakam, Dr. Gary Coleman, Dr. Stephen Woodruff and the Data Analytics and Machine Learning team about his machine learning work to improve turbulence modeling. These discussions resulted in specific ideas for applying Dr. Duraisamy work to NASA Langley CFD 2030 goals. These will be pursued as part of the Aeronautics Long-term Engagement in Authentic Research with NASA (LEARN) project work that is underway, with Dr. Duraisamy collaborating with NASA LaRC CFD and machine learning experts. In August 2016, Dr. Duraisamy presented his work as part of the NASA Langley Machine Learning Workshop.

B3.5.4 Next Steps

While initial consideration in 2014 was given to the possibility of the team developing a viable regression approach, the complexity of the aerodynamics does not readily lend itself to further investigation of that path. Instead, continued collaboration with our Langley experts and with researchers such as Dr. Duraisamy will help us accurately assess the potential for developing a use case that can demonstrate the potential of machine learning in this area.

B3.6 Entry, Descent, and Landing

B3.6.1 Overview

In 2014, discussions with Entry, Descent, and Landing (EDL) experts at Langley helped them team better understand how Monte Carlo analyses are being run and how data is passed to related disciplines for further evaluation. Because changes can be made during this process by one specific discipline, the analysis results can be subsequently affected when used by any other disciplines. EDL identified a desire to have a more interconnected set of disciplines and shared data capabilities to create improved Monte Carlo simulations and to gain a better system-wide understanding of the impacts of changes made by each discipline.

B3.6.2 Machine Learning and Statistical Techniques

Plans for a use case in EDL started with a focus on improving analysis efficiency, connecting different disciplines, and facilitating sharing of data among the disciplines. To improve the efficiency of data processing and subsequent analyses, data structures within MATLAB were applied as part of an exploratory data analysis proof of concept.

B3.6.3 Results

Collaborative discussions over the past two years led to the sharing of a representative data set, which provided a means to assess possible paths for a full-fledged use case. This included the development of enhanced methods of data handling and initial processing described above, and a short-term use case was developed in 2015.

B3.6.4 Next Steps

Subsequent to the 2015 effort, a more complex problem was identified and presented to the collaborative team at Georgia Tech; this work is being followed by CDT.

B3.7 Climate Science: Cloud Fraction Simulation

B3.7.1 Overview

Accurate simulation of cloud fraction in reanalysis models is a critical component in predicting the state of our future climate and atmosphere. Model uncertainty in this measurement poses a threat in representing the atmospheric and terrestrial heat budget. Various studies have looked into the evaluation of different atmospheric reanalysis products in this region, reporting across-the-board model bias.

B3.7.2 Machine Learning and Statistical Techniques

Methods focused on the use of Artificial Neural Networks (ANNs) to improve cloud fraction measurements from reanalysis models, since ANNs are able to resolve nonlinearity in the data. ANN regression models were developed

along with traditional linear regression to explore the correlation between Merra-2 averaged Arctic summer near-surface data such as atmospheric temperature and specific humidity, which are currently used in reanalysis models to formulate cloud fraction.

B3.7.3 Results

The results from the ANN regression model outperformed traditional linear regression. The ANN model shows an $R^2 = 0.88$, while the linear regression reports an $R^2 = 0.12$. This suggests that neural nets should be investigated further to improve relationships in climate models to explore further relationships with other atmospheric data.

B3.7.4 Next Steps

Future directions for this work may include training more models with different NN configurations, such that the connectivity of the nodes can be exploited. There are also plans to implement satellite variable data from VIRS, CloudSat, MODIS and Calipso.

B3.8 Space Launch System (SLS) Booster Separation Aerodynamics

B3.8.1 Overview

Our understanding of SLS Solid Rocket Booster aerodynamics during separation is limited by our ability to model this complex environment. No single model provides the needed accuracy, uncertainty quantification, and affordability needed. Multi-fidelity surrogate models, driven by machine learning techniques, can use our limited modeling resources more intelligently.

B3.8.2 Machine Learning and Statistical Techniques

Multi-fidelity methods use surrogates to provide a ‘correction’ between data sources of different origin and different fidelity. These surrogates combine the best aspects of their data sources: The speed of a low-cost CFD code can be combined with highly-detailed CFD to quickly cover the whole design space with the needed accuracy. Corrections can be performed using additive, proportional, or interlaced surfaces. Surrogate networks can be built for more than two sources as well. Artificial Neural Networks were evaluated for creation of the single-fidelity and correction surrogates. Networks were constructed using a feed-forward architecture, varying the number of hidden nodes and layers to fit the size of the training data set. The networks are trained to minimize response error in regions of interest to GN&C. The low-fidelity surrogates are trained against the Cart3D data. The difference between the low-fidelity prediction and the high-fidelity data is used to train the multi-fidelity correction.

B3.8.3 Results

Multi-fidelity surrogates demonstrated decreased variation and decreased mean error compared to the single-fidelity models. Surrogates built using deep Neural nets delivered lower variation and mean-squared error compared to baseline Kriging and lookup table databases.

B3.8.4 Next Steps

This multi-fidelity framework enables research into adaptive sampling, using surrogates to select the best data a priori, further reducing the amount of data required. Incorporation of uncertainty quantification into each surrogate, as well as creating surrogates to represent the uncertainty itself, will also be investigated. Ongoing work on this project is being pursued in collaboration with Georgia Tech.

B3.9 Hypersonic Inlet Performance Analysis

This new project for 2017 is investigating the use of machine learning methods for classifying safe operating conditions for hypersonic aircraft. Helping SMEs to better understand areas of high uncertainty in performance will reduce the computational/experimental effort required to assess the safety of vehicle components. The data derives from CFD modeling of scramjet engine inlets.

B3.10 Space Launch System (SLS) Additive Manufacturing Certification

This is a new project under the Additive Manufacturing Structural Integrity Initiative (AMSII), involving collaboration between LaRC, Ames, GRC, and MSFC. Work is anticipated to begin in early 2017, as data is being collected as part of a research effort to better understand additive manufacturing in NASA's domains. Langley's role is focused on investigating the origin, manifestation, and effects of defects in additively manufactured components. The data will be gathered from a multitude of sensors that monitor the manufacturing process as defects are intentionally introduced. SMEs are interested in tools that can process the data effectively and identify significant patterns in the data that characterize and/or predict defects.

B3.11 Flight Deck Analytics from Trajectory-Based Optimization

This is a new project planned for 2017. Data is collected from real-time simulations of national airspace trajectories and is stored in databases. SMEs hope to leverage this data to optimize simulation scenarios and to develop new scenarios to explore gaps.

B4 Progress on Deep Content Analytics (DCA) Projects

Implementing Deep Content Analytics to Achieve a Knowledge Assistant

Based on the original goals, pilot projects focused on autonomous flight and carbon nanotube research were successfully completed in 2015. These projects provided for the establishment of a knowledge base architecture, along with ample opportunity to explore the functionality and requirements of the IBM Watson Content Analytics software platform. Following these initial pilots, additional use cases have been developed for Space Radiation, Vehicle Design, Space Mission Analysis, Uncertainty Quantification, Model-Based Engineering, Human-Machine Teaming, and a limited investigation of the application of cognitive technologies with the content from the NASA Technical Reports Server.

Experts at NASA LaRC are facing a growing challenge – reading and digesting the volumes of technical literature that will allow them to make optimal decisions in a timely manner. Each year, nearly 450,000 papers are published in scientific journals, along with hundreds of thousands of technical reports, manuals, and patents. On average, it takes a scientist/expert 160 minutes to read a given paper/report. NASA experts must dedicate hundreds of hours of their time each year – which is not feasible or practical – to digest and leverage this information/knowledge. This results in a slow progression of research and actions due to the vast amount of information that must be reviewed to gain new insights and make new connections.

At LaRC, we are utilizing the power of IBM Watson Content Analytics (WCA) to analyze and digest large volumes of scientific information rapidly without ‘reading’, through the application of syntactical and semantic techniques so that SMEs can gain key insights and patterns, identify trends and connections, and visualize expert networks. Over the past two years, several use cases have been worked to support the goal of providing the technical community with natural language processing technologies that will quickly make sense of internal and global knowledge. These projects are outlined below.

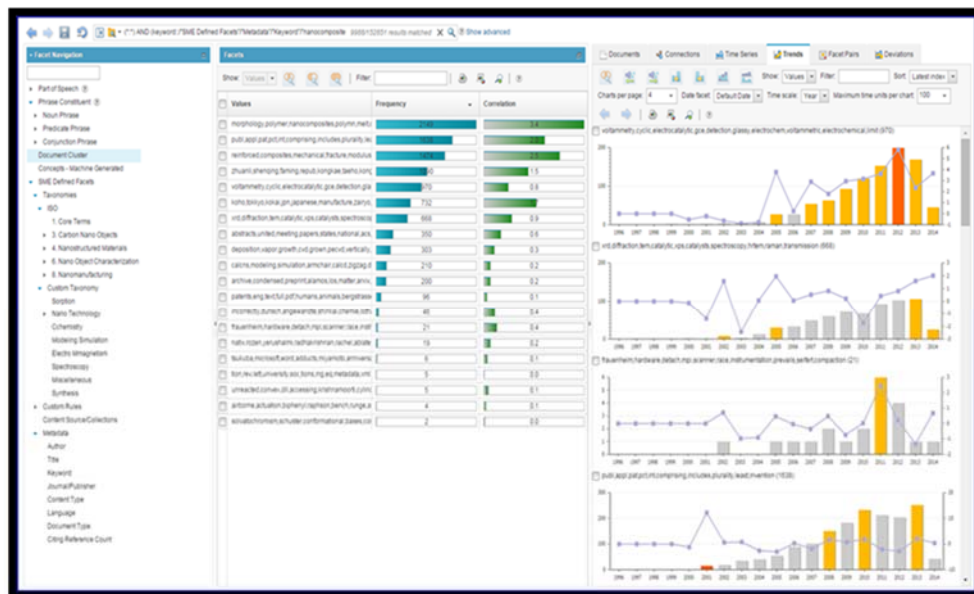


Figure 17 IBM Watson Content Analytics provides the basis for Knowledge Analytics at LaRC

WCA provides a platform to collect and analyze information in documents, databases, emails, websites, and other sources of information. It provides an interactive interface that can be used by users to discover relationships and anomalies between various topics and to retrieve documents that are relevant to a specified subject from a ranked list of results. There are two main ways that WCA can be used. First, it can help SMEs derive quick insights from large collections of documents. These insights usually operate on facets, which are characteristics of the documents derived from structured metadata (date, author, tags, etc.) or from concepts extracted from the text in the document. Secondly, the tool allows users to extract concepts for use by the WCA analytics view or other downstream solutions. Examples in other domains include the analysis of physician or lab reports to populate patient records, extracting relationships or named entities to feed investigation software in law enforcement, or defining sentiments expressed on social media to improve the statistical analysis of consumer behavior. Similar approaches can be devised to create value-added applications in aerospace domains. In the following sections, we outline the use cases that have been developed to investigate the potential of WCA.

B4.1 Carbon Nanotubes

Materials science researchers are looking beyond current applications of carbon-fiber composites in aircraft and spacecraft by investigating the use nanostructured materials such as carbon nanotubes. The WCA tool was used to assist the researchers by reading through more than 130,000 technical documents in a given collection and then identifying the most salient content without the need for a human operator to provide specific terms to start the search. In 2015, the tool was successfully demonstrated to SMEs and stakeholders.

B4.2 Autonomous Flight

In order to operate safely, autonomous flight systems must incorporate computer vision and image processing techniques to correctly identify and respond to potential hazards. Using the WCA tool along with web-crawling technologies, the IEEE database was searched to generate a corpus of publications related to the computer vision techniques used in autonomous flight. More than 4000 articles from IEEE and other selected websites were ingested into the WCA tool, along with expert-generated taxonomies and facets to enable deep analytics of the content. This use case also incorporated the development of web crawlers to mine text across a wide range of websites. In 2015, the tool was successfully demonstrated to SMEs and stakeholders.

B4.3 Space Radiation

The NASA Space Radiation (SR) Program Element has been working on a collaborative project with the team since November 2014, with the goal of creating a tool that allows researchers to search the NASA SR funded research corpus to help streamline research and maximize efficiency. To date, this ongoing project has ingested more than 300,000 documents, and the development of automated crawlers for the PubMed database will allow the collection to be updated with new content on a monthly basis. A diverse set of visualizations has been created to meet SME needs, including heat maps, bar charts, and pie charts. These visualizations are deployed in a manner that allows the SME to easily modify each visualization to suit their needs, increasing the power of the tool in providing new insights into this large corpus.

B4.4 Aerospace Vehicle Design

Working with SMEs in the Systems Analysis and Concepts Directorate (SACD), more than 20,000 reports from databases including AIAA, NTRS, SAWE, and hundreds of scanned historical documents are being ingested into the WCA tool. This will give SMEs the power to more easily find mass and weight properties used for various vehicle concepts and designs, while leveraging the corpus of research and design reports from the past fifty years. The SMEs envision a resource that will improve their ability to accelerate the design process, predict ‘what ifs’ of future vehicle designs, and better leverage the work they have done over the past few decades. To facilitate SME goals for the collection, a large taxonomy based on SME-provided technical content was built to facilitate efficient faceted search.

B4.9 NASA Technical Reports

Using a process developed in collaboration with STIP, this project resulted in the ingest of more than 8,000 abstracts from the NASA Technical Reports Servers (NTRS). Specifically focused on content from two subject categories – “Aeronautics” and “Composite Materials” – along with a taxonomy created from the NASA Scope and Subject Category Guide, the use case has provided insight into potential paths for future collaboration and development. As a next step, our team will be working to build additional dictionaries for the content, and will work with STIP to better understand the most effective path to pull data and expand the collection’s content.

B4.10 NASA Lessons Learned

This project was started to help gain insight, identify patterns and trends, and visualize LLIS data at a Center and/or Agency level. An initial collection of more than 2000 records provided by NESC and KSC has been pre-processed and ingested into Watson Content Analytics software and the BDAMI team has built metadata facets, custom parsing rules, and document clusters to help with this analysis. This prototype was demonstrated and feedback received was positive and useful. Going forward, plans include expanding the prototype with all NASA lessons learned, incorporating non-sensitive data from the Engineering Project and Task Technical Review, enhancing the analytics with more facets and visualization, and developing more complex parsing rules that will add context to the content.

B5 Progress on Deep Q&A Projects

B5.1 Cognitive Computing and the “NASA Watson” Vision

The vision for Deep Q&A and NASA Watson is to apply cognitive technologies and artificial intelligence to develop ‘NASA Watson’ that spans all aerospace disciplines and behaves as a trusted assistant to our researchers and engineers. At its core is an ability to ingest and understand scientific, engineering and other technical papers and publications, and associated data from multitude of sources and formats including multimedia. The key features of cognitive technology and knowledge-based artificial intelligence (AI) systems are:

- Cognitive-based systems are able to build knowledge and learn – through an understanding of natural language – and to reason and interact more naturally with human beings than traditional systems.
- Cognitive systems continue to evolve as they ingest new information, new scenarios, and new responses. They reason in a way that is similar to human thinking so conclusions are obvious, transparent, and useful.
- Experts train the systems, and it takes time to teach new and complex domains. Experts and AI systems work together doing what each does best.
- Cognitive and knowledge-based AI systems have the power to democratize knowledge, deep expertise spanning creativity and innovation.
- Cognitive systems amplify human cognition; this is the Power of Human-Machine symbiosis.

Currently, IBM Watson is a cognitive computing system that is amplifying and enhancing the abilities of experts in medicine, finance, banking, and consumer markets. The goal we have for ‘NASA Watson’ is to provide the NASA technical community with the capacity to ask and get relevant answers with evidence for scientific and engineering questions in all NASA core disciplines. This will require an integrated analysis and understanding of a large variety of data types and worldwide multimedia knowledge. NASA’s vision defines an Aerospace Research and Design Advisor that can understand mathematical equations and can mine chosen scientific and engineering information with associated data, images, and videos enabling our researchers and engineers to quickly develop, innovative, cost effective and technically feasible solutions and designs. We also see ‘NASA Watson’ helping our technical leadership to make better decisions from evidence-based choices, and helping to advocate investments to our stakeholders and founders.

NASA researchers, engineers and project teams can have a “Virtual Expert” or “Virtual Colleague” at their disposal that can answer specific questions, synthesize and make sense of volumes of big data or information, processes discipline modeling and simulation data in real time, and provide predictions for new technologies and design configurations. A “Design Assistant” would be able to extract relevant facts from documents and answer engineering questions using natural language during all phases of the design process. Other capabilities envisioned include this Design Assistant being able to make suggestions or provide feedback during design, based on a model of the design activity and relevant principles of physics. It will understand engineering and design documents in foreign languages, along with multimedia input, such as drawings, charts, and video.

Human intuitive cognition and machine cognition augment each other providing unimaginable new capabilities. Such a capability will benefit NASA Langley by helping to enable it to achieve its goal of less physical testing and reliable and cost effective simulation-based engineering and science. It can free up technical professionals’ time to be more creative and tackle harder challenges, enable NASA Langley to become more competitive and innovative, providing transformational aerospace technologies, while enhancing the nation’s ability to explore space affordably, inspiring the world and human kind with possibilities of living on other planets. Researchers and Engineers will have “Digital Advisors/Experts” enabling greater scientific discoveries, and innovative systems designs and complex operations including the abilities to:

- Quickly digest the latest research innovations by synthesizing a large volume of information rapidly, showing unobvious trends and paths.
- Analyze experimental, modeling and simulation and flight data in real time, helping with system configurations and design predictions/optimizations.
- Answer specific engineering questions in all aerospace disciplines by integrating data and knowledge and showing evidence and traceability.
- Conduct deep analysis and mining of multimedia scientific and engineering information with associated data, images and videos, comprehend numbers and mathematical equations resulting in actions, advice, and answers that augment, augment and replace human experts.

The key benefits of cognitive computing and knowledge- based AI to NASA include providing ways to be more innovative and competitive, and able to tackle harder challenges, facilitating research that leverages worldwide knowledge with optimal paths for discovery, facilitating designs that are optimal solutions for aeronautics and space vehicles, helping to optimize physical testing and modeling and simulation, and providing leadership with insight into decision-making about the most beneficial investments

Domain adoption and training is key; these Watson-based assistants may not be possible immediately. Breakthroughs in technology for understanding mathematics and tables, and mimicking human cognition and intuition are still required and are being worked. However, even within the limits of current technologies, we can start making progress by taking advantage of the power of a human-machine cognition symbiosis that helps NASA to better assess the path forward as these technologies mature. Examples of forward thinking projects could include:

- A pilot advisor that will help determine the best possible solution paths in distress situations, using a speech interface.
- A co-pilot, as cognitive computing technology adoption in our aerospace domain matures and we gain confidence in its performance
- Autonomous agents/robots/rovers for exploration of other planets, that can help us to do more things quickly and creatively
- Advisors and companions for long-duration space travel

Over the last year, in collaboration with IBM Watson experts, we have begun to work on proof of concepts in applying the Watson Discovery Advisor cognitive technology. Two cognitive applications – Pilot Advisor and Aerospace Innovation Advisor – are described below, along with progress made and next steps. The main objectives of each proof of concept are to understand how the application of cognitive technologies to aerospace domains will function, to identify the associated challenges with domain adoption, and to assess what is needed to develop usable applications for our scientists and engineers, in terms of both technology and resources. Using the Watson Discovery Advisor Platform, these prototypes will help us evaluate the application of cognitive technologies to our domains, and help us better understand how we should prepare to take advantage of these technologies for our mission challenges.

B5.1 Current Application: Pilot Advisor Proof of Concept

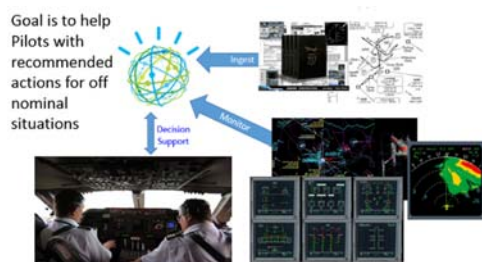


Figure 19 Pilot Advisor Proof of Concept

Researchers at NASA Langley have set the goal of developing an *Autonomous Pilot Advisor System*. This system, deployed on the flight deck, would replicate and augment the monitoring, assessing, and decision-making functions of a human (expert) pilot that will, during nominal and off-nominal situations, continuously identify risk, and determine/prioritize actions needed to mitigate risk, and continually and autonomously maintain safety-of-flight. The vision for this Autonomous Pilot Advisor System for an aircraft flight deck is to provide advice to the aircrew in both normal and abnormal (emergency) scenarios. As envisioned, this system will have three goals:

- Monitor and analyze internal (e.g., engines or hydraulics) aircraft system data and external factors (e.g., weather) data in real time
- When an unusual condition is detected, or an anomalous condition is input by aircrew, the cognitive system searches for technical and procedural advice from a very large corpus of diverse material of flight manuals and other identified related corpus.
- Develop and present recommended actions to the aircrew with background information and evidence to support them.

This proof of concept (POC) is focused on Goals 2 and 3 of the BDAMI Vision, and is helping to develop a long-term cognitive capability plan that will help to roadmap the implementation of the Autonomous Pilot Advisor System. The project was built around a capability of natural language processing, cognitive computing, and machine learning technologies that can absorb massive amounts of published information and experience to provide expert advice and probabilistic or deterministic situation analysis. The initial application for the POC was focused on root-cause analysis where the system would analyze (non-normal) aircraft states and system status information as inputs by aircrew, and provide diagnosis of the situation with recommended actions and showing evidence/analysis. This would provide the human expert/aircrew with a better understanding of the non-normal situation, as well as the effect of system faults on dependent systems and operations, and help with taking most effective decisions in non-normal situations.

Currently, the Watson Discovery Advisor (WDA) is designed to amplify and make effective use of inherently human expert abilities of experienced insight and intuition. The team is assessing IBM WDA, and working to demonstrate the capability of Watson Discovery Advisor to a root-cause analytic process for a commercial aircraft application. One could conjecture that the Watson Discovery Advisor - being used in medicine and life sciences as Oncology Advisor to help Oncologists to figure out the best treatments for the given patient – could be harnessed in this application to help Pilots and aircrew. The twelve-week proof of concept was completed by IBM Experts who worked closely with NASA flight crew systems subject matter experts. The POC used one specific aircraft accident use case and successfully demonstrated the potential of how the Pilot Advisor could help the flight crew, with positive feedback from SMEs. It has also uncovered gaps in the current Watson system, such as an understating of flight manual logic, deep domain adoption challenges and decision advice challenges in real time, and natural user interfaces needed for real time dialogue between expert and machine. IBM has also delivered a cognitive capability plan with a phased approach towards the vision, and SMEs who see the value and potential are working to obtain the funding to continue this development.

B5.2 Current Application: Aerospace Innovation Advisor Proof of Concept



Figure 20 Aerospace Advisor Proof of Concept

Researchers at NASA Langley Research Center are interested in the development of an *Aerospace Innovation Advisor* Proof of Concept (POC) and capability that demonstrates how natural language processing and machine learning technologies can be applied to aerospace research and development to accelerate the pace of discovery and innovation by analyzing and fully leveraging massive amounts of published information. This prototype will use select open-source documents from the NASA NTRS system, patents database, DOD DTIC, and web information from select aerospace web sites as a corpus. NASA Langley has been working with IBM Watson Content Analytics for the last two years with good results and is now ready to utilize the full functionality/power of IBM Watson Discovery Advisor (WDA) technology that is designed for examining a body of knowledge corpus to answer questions with traceability of evidence, to recognize patterns assisting researchers in the discovery of underlying causes behind effects, and to generate new insights and hypothesis that can be validated with evidence. This could help to cut the time from ideation and concept to innovation, and help to formulate reports that can shape decision and actions related to NASA R&D focus areas avoiding duplication and saving money.

Currently, Watson Discovery Advisor is being used in medicine and life sciences as Oncology Advisor to help Oncologists to figure out the best treatments for the given patient, and Life Sciences Discovery Advisor for optimal cancer research paths by mining the vast amounts of literature saving time and money. WDA is designed to amplify and make effective use of inherently human expert abilities of experienced insight and intuition. This Aerospace Innovation Advisor Prototype can help us to gain insight into using the advanced cognitive technologies in general (machines that learn from experts and assist experts to make better decision and discoveries) and can help us tremendously in evaluating their use for our domains.

NASA Langley and IBM Watson team have started to develop this POC demonstrating the capability of Watson Discovery Advisor to a chosen aerospace corpus, and are developing a Cognitive Capability Plan showing the roadmap to implement NASA Langley's vision of Aerospace Watson. The POC will demonstrate key features including answering questions with evidence, relationships within multiple disciplines, linkages of experts, and possible paths for innovation research. In developing this POC and cognitive capability plan, IBM will be collaborating with the NASA Langley team closely; the team consists of technical leaders, IT experts, computer scientists, and subject matter experts. The two subject areas being evaluated for this POC are hybrid electric propulsion, and aeronautics technology roadmaps for the NASA Aeronautics Research Mission Directorate (ARMD) over the last 10 years with associated technical papers and literature to leverage the work of experts for

decision making on funded research to advance ARMD technical and strategic goals. This POC is scheduled to start in February 2017.

B5.3 Next Steps

These proof of concepts will also spur the development of new base capabilities for the Watson cognitive system. This will include ingestion of content beyond simple text, including multimedia data and metadata, imbedded tables, and mathematical expressions. They will also drive further development of tools to enable easier domain adaptation, including support for processes such as lexicon development and inclusion as new sub-disciplines of a major discipline are added. The key takeaways are:

- Proof of Concepts underway will provide us with an excellent understanding of cognitive computing applications in our domains
- Proof of Concepts will help us to demonstrate functionality, potential, and gaps to NASA experts, leaders and stakeholders
- Cognitive systems amplify human/expert cognition; this is the power of Human-Machine symbiosis in action, and will allow us to learn more about how this works
- Cognitive capability plans, experience and feedback will help us to formulate Watson and Cognitive Computing plans for the next few years

B6 Collaboration

Establish Partnerships with Universities, Industry, and other Federal Agencies

As outlined in the original vision, collaborative partnerships were successfully developed with NASA Ames Research Center, NASA Glenn Research Center, Old Dominion University, Georgia Tech, IARPA, and IBM. In each instance, the team at LaRC has participated in technical discussions, conferences, and forums to better understand how our respective strengths can be leveraged to provide mutual benefit. A close working relationship with the Intelligent Systems Division at ARC has led to several opportunities for collaboration at both centers, as well as at workshops and conferences. The partnership with IBM provides the foundation for text analytics and natural language processing, while collaboration with Georgia Tech has accelerated the development machine learning applications at LaRC.

Collaborative partnerships have been pursued with other NASA centers, universities, federal agencies, and industry to better understand how our respective strengths can be leveraged to provide mutual benefit. These close working relationships incorporate elements from three pillars of the long-term vision – Data Intensive Scientific Discovery, Deep Content Analytics, and Deep Q&A – and have provided the opportunity for linking NASA scientists and engineers with the expertise needed to solve their complex challenges.

B6.1 Collaboration with other NASA Centers

Our team has developed a strong collaboration with the Intelligent System Division at Ames Research Center, which specializes in Autonomous Systems and Robotics, Collaborative & Assistant Systems, Discovery and Systems

Health, and Robust Software Engineering. Their deep research and algorithm development over the past 15 years is a complement to our younger team's more applied approach, providing for a mutually beneficial collaboration. As the only two core machine learning groups within NASA, working collaboratively will also provide significant benefit for the agency. Team members from both centers have attended workshops, conferences, and visits to universities together, and representatives from both teams actively participate in the agency's Big Data Working Group. In addition to Ames, we have also maintained a positive working relationship with Glenn Research Center, where initial efforts are being made to investigate applications of machine learning and data visualization in their aerospace domains. Efforts to share and collaborate with both centers will continue.

B6.2 Collaboration with Universities

Over the past three years, we have worked closely with leading universities to advance the goals of CDT, and maintain an awareness of novel research in the application of machine learning in scientific and engineering domains. Locally, we developed a Space Act Agreement (SAA) with Old Dominion University, providing the means to obtain expert consultation in areas such as signal processing and algorithm development, and to facilitate regular collaborative meetings to plan for future needs.

In early 2016, representatives from the CDT team visited Georgia Tech to discuss and learn about research areas and initiatives that can be leveraged for CDT goals in data analytics and machine intelligence, high performance computing, and advanced IT. An SAA is now in place to leverage expertise and technology that cuts across colleges/departments, and brings experts across the university together for the application of machine learning technologies. A close collaboration with domain/discipline experts from the Aerospace Systems Design Laboratory for consultation on specific applications to LaRC needs has supported the identification and development of additional use cases and grand challenges investigating the application of machine learning to climate science, EDL, and fluid dynamics.

In the same timeframe, CDT representatives visited the MIT Computer Science and Artificial Intelligence Lab (CSAIL) in January 2016 to learn more about their extensive research capabilities and pursue collaboration for furthering strategic goals. CSAIL has nearly 1000 researchers and faculty members, with more than 50 research groups working on hundreds of diverse projects, focusing on cutting edge research to develop novel ways to make systems and machines smarter, easier to use, more secure, and more efficient. The collaboration has focused on big data, machine learning, and cognitive computing areas, and how we can leverage their expertise and open source technologies for our aerospace data analytics. We have agreed to pursue a no cost, umbrella SAA to further this collaboration. As a part of these discussions we learned that CSAIL has deep research expertise capabilities in high performance computing, cybersecurity, robotics, autonomous systems, climate science, and software verification and validation, which we can also leverage. Currently, a membership in CSAIL has been procured, providing access for all NASA researchers to interface with CSAIL experts and develop more specific technical projects.

B6.3 Collaboration with Expertise in Industry and Other Federal Agencies

In order to further the objective of bringing advanced text analytics and cognitive computing technologies to Langley, representatives from the team visited the IBM Research Lab early in the effort to better understand Watson technologies and their potential for our domains. These initial visits provided the opportunity to see demonstrations

of Watson technologies under development, and for discussions with experts about their applications for NASA. Sharing our vision of Intelligent Agents/Experts that will help with NASA mission challenges resulted in the work that was previously outlined in sections on Deep Content Analytics and Deep Q&A. Through effective contract vehicles, Langley has been able to leverage not only tools such as Watson Content Analytics and Watson Discovery Advisor, but also the expertise of leading researchers at IBM, who are interested in understanding their customers' visions of cognitive technologies. Regular interactions with their scientists has helped to keep us abreast of cutting edge technologies, such as brain-inspired computing and intelligent image analysis for cancer detection.

The team has also worked to develop collaborative relationship with other federal agencies and research labs. Several conversations and in-person meetings with researchers at the Intelligence Advanced Research Projects Activity (IARPA) regarding work on analytics systems such as FUSE provided new insights into technology development, and helped to better evaluate the value of open source development approaches. Interactions with representatives from DoE labs has occurred on a regular basis, and the team has presented an overview of their efforts at workshops hosted by the Thomas Jefferson National Accelerator Facility in Newport News.

B7 Outreach and Education

Conduct Education and Outreach Initiatives

One of the primary objectives for the team, education and outreach has played an important role in the strategy for developing interest in data science applications, and in identifying new use cases. In early 2015, two center-wide seminars were conducted to introduce the team's work to a broad audience, and to encourage SMEs to contribute ideas for additional use cases. Based on the success of this approach, a larger-scale NASA Langley Machine Learning Workshop was planned. In the summer of 2016, this three-day event attracted more than 250 registrants. The workshop provided a forum for scientists and engineers to learn from 15 experts across academia and industry on advances in machine learning techniques and cognitive technologies, and their applications to NASA domains such as computational aerosciences, computational materials and structures, next generation airspace, autonomy, aerospace systems analysis and design, and climate science. Participation in this workshop helped to further develop the important work NASA Langley is doing in this area through investigating and applying emerging technologies in data analytics and machine learning to address NASA's technical challenges. In addition to these multi-day events, a regularly occurring CDT seminar series has brought leading researchers to the center to present their work and develop connections with Langley. Additional workshops on both open-source and licensed software platforms have helped SMEs become familiar with tools and techniques such as deep learning, natural language processing, and content analytics. Education and outreach events from 2014 to 2016 are outlined below.

Knowledge Analytics sessions/demonstrations and Machine Learning demonstrations included the following:

- Cognitive Computing Discovery Advisor (September 2014)
- NASA Watson Content Analytics Conference Presentation (October 2014)
- Watson Content Analytics for NASA HRP Research (September 2015)
- Watson Workshop: Chris Codella (August 2015)
- Deep Learning Talk/Workshop (March 2016)
- Natural Language Processing Talk/Workshop (August 2016)

Center-wide seminars from visiting experts included:

- Michael Krein: Applied Data Mining: Hype or Hallelujah (June 2015)
Nikunj Oza: Machine Learning Methods for Mining NASA Data (August 2015)
- Kalyan Veeramachaneni: Building Predictive Models (July 2015)
- Abdullah Mueen: Unsupervised Pattern Mining (August 2015)
- Rob High: Application of Watson Technologies for NASA Missions (October 2015)
- Bernd Chudoba: Application of Machine Intelligence for Aerospace Systems Design Decision Support

Presentations on the vision and work of the team included:

- Langley Big Data Analytics & Machine Intelligence Strategy (May 2014)
- Big Data Analytics and Machine Intelligence Seminar: Data Intensive Scientific Discovery (March 2015)
- Big Data in Aerospace Panel at AIAA SciTech (January 2016)

One of the primary objectives for the team, education and outreach has played an important role in the strategy for developing interest in data science applications, and in identifying new use cases. Activities related to this objective have included multi-day events hosted on center, a series of visiting lecturers, seminars and focus groups, participation in agency events, invitations to present at external events, and trainings on both open-source and licensed software platforms. All of these activities have helped SMEs to become more familiarized with the potential for machine learning in the aerospace domain.

B7.1 Seminars and Focus Groups

In March 2015, the team presented two center-wide seminars. Each session presented both an overview of the 20 year vision of a “Virtual Research and Design Partner” that will enable NASA employees to achieve greater scientific discoveries and system design optimizations, and the details of the team’s current work. The seminars also included a more in-depth presentation of current pilots the team is using to develop foundational expertise and capability in the application of big data analytics and machine learning technologies to data in Langley’s aerospace domain. Follow-up feedback sessions with SMEs were hosted in the week following each seminar, providing an opportunity for critical assessment of the work, and suggestions for potential new use cases. In March 2016, two focus group events were held with SMEs to revisit some of the topics discussed the year prior, and to gather input on strategy and next steps.

B7.2 Agency Events

In October 2014, the team hosted the first NASA “Big Data Big Think”, gathering over 20 leading thinkers in data science techniques & technologies from across the Agency, to begin building a community of technical excellence and kick off creation of a NASA Data Strategy. The event, sponsored by the Agency Chief Technology Officer for Information Technology, the Agency Executive for High Performance Computing, and LaRC’s CIO, served as a creative space for this team of futurists to begin shaping the way NASA can use big data as a powerful mission enabler. Following this event, the team has actively participated in additional Big Data Big Think workshops held at Johnson Space Center and Goddard Space Flight Center, presenting our work on projects and sharing our vision for future technology developments.

B7.3 Conferences

In January 2016, our work was presented as part of the Aerospace Panel at AIAA Sci Tech in San Diego, CA. An overview of our vision of having virtual experts and human experts working together allowing our researchers and engineers to focus on developing innovative solutions to our complex mission challenges was discussed, along with the mission focused use cases we are working on involving diverse aerospace data and information, and how the machine learning, statistical, data analytical techniques, and algorithms are being applied. The panel also included experts from NASA Ames, Rutgers University and Boeing R&T and Boeing Commercial.

In March 2016, the team participated in the Thomas Jefferson National Accelerator Facility (Jefferson Labs) Future Trends in Nuclear Physics computing workshop. At the invitation, they presented work as 'Big Data and Machine Learning in Aerospace', which was well-received with good discussion. Workshop attendees also included participants from other Department of Energy labs such as Lawrence Berkeley National Laboratory, Brookhaven National Laboratory, SLAC National Accelerator Laboratory, and the international lab of Conseil Européen pour la Recherche Nucléaire (CERN). Jefferson Lab high energy physics team and the NASA LaRC Chief Information Officer (CIO) team are very interested in collaboration in the areas of machine learning and scientific computing.

Team members regularly travel to conferences including the yearly Society for Industrial and Applied Mathematics (SIAM) International Conference on Data Mining, the International Conference on Learning Representations (ICLR), and many smaller conferences focused on artificial intelligence, deep learning, and cognitive computing.

B7.4 Center-wide Workshops

In July 2016, the Big Data Analytics and Machine Intelligence Capability Team hosted a four-day workshop on the IBM Watson Analytics and Explorer Software Technology tool. Participants included several Subject Matter Experts from the Research Directorate (RD), Engineering Directorate (ED), and Systems Analysis & Concepts Directorate (SACD), along with IT experts from CDT, Office of the Chief Information Officer (OCIO), and NASA Engineering and Safety Center (NESC). The workshop provided participants with hands-on labs to learn about how the features and functions of the tool can support NASA Langley Research Center (LaRC) research and NASA missions. Workshop participants learned about the methodology used to efficiently discover answers using global scientific information; such techniques will benefit SMEs by enhancing innovations and solutions to complex NASA challenges.

In August 2016, the team planned a workshop held at NASA Langley 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*

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. 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. For additional information, including overviews of each presenter's topic, please reference NASA TM 2016-219358, *Machine Learning Technologies and Their Applications for Science and Engineering Domains Workshop – Summary Report*.

B8 Research Machine Intelligence

The team has actively researched applications of data analytics and machine learning, and has used that knowledge to identify key partners for collaboration. Given the complexity of our NASA domains, much of the expertise required to successfully solve deep technical challenges still resides in academia. By developing a thorough foundational knowledge of the tools, techniques, and methods that are most applicable to aerospace research, the team has successfully developed partnerships that will help to advance our shared goals. Investigations into the potential use of Numenta in aerospace domains, human-machine teaming, and planning for a 2017 "blue sky" event focused on artificial intelligence have all contributed to this objective.

B9 Hire or Obtain Expertise

A core Big Data Analytics and Machine Intelligence Capability Team has been developed by recruiting young civil servants and contractors with skill sets in computer science, mathematics, statistics, and information technology, and pairing this team with experts in aerospace domains. To further the development of data analytics and machine intelligence in our NASA domains, we have adopted the philosophy that data science is a team effort, and cannot be effectively conducted by a lone data scientist.

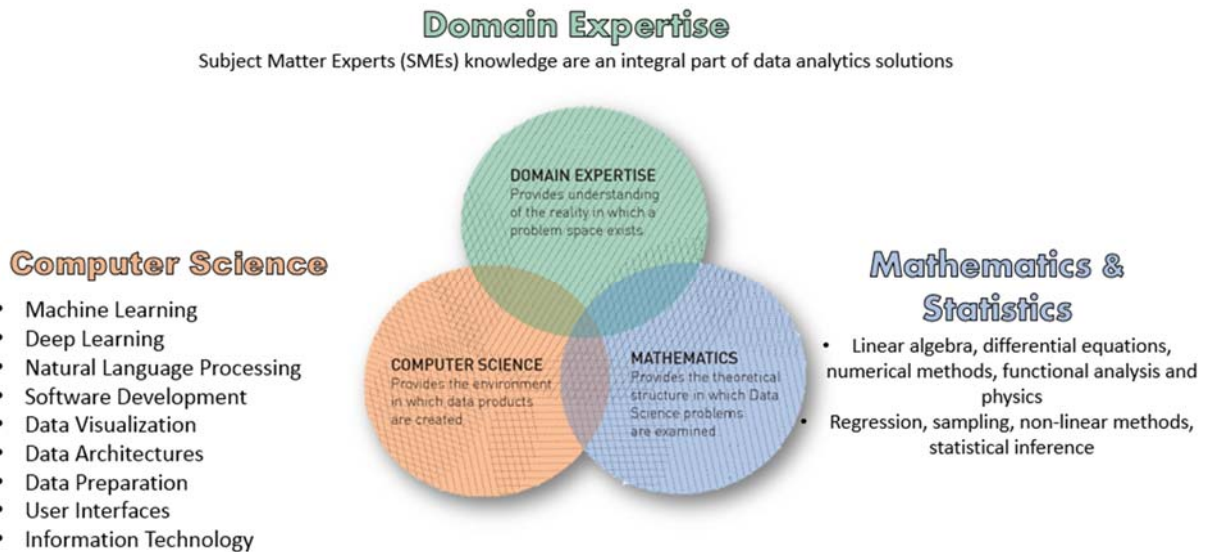


Figure 21 Data Science as a Team Effort

In FY15, two civil servants – one with a specialization in computer science, and the other with a specialization in applied mathematics – were hired, and an Associate CIO position was created to lead the Big Data Analytics and Machine Intelligence capability development. Over the following two years, three contractor positions were procured to provide additional expertise in statistics and computer science. In FY16, a computer science Pathways intern focusing on natural language processing and cognitive technologies was added to the team. We have leveraged the advantage of the NASA Internship, Fellowships, and Scholarships (NIFS) program by bringing in many young, motivated, and skilled interns from the fields of computer science, engineering, climate science, and cognitive science to expand our ability to investigate the application of new technologies. In the course of three years, we have mentored more than a dozen interns from many universities across the country. Additionally, collaboration with universities such as MIT, CMU, and Georgia Tech, and with IBM and NASA Ames Research Center has provided external expertise, particularly in instances where deep skills are needed to solve a complex domain challenge.

C: Moving Forward (2018 – 2020)

C1 Vision, Roadmap, Recommendations, and Actions

The vision for big data, deep analytics and machine intelligence is to enable LaRC to discover “unknowns” and deliver previously unimaginable capabilities by applying these transformational technologies as force multipliers for scientific and engineering discoveries and systems innovation and optimization. Achieving this vision will provide a number of tangible benefits. These include cost savings resulting from the use of more SBES and less physical testing to enable LaRC to be more competitive and innovate in providing transformational aerospace technologies. Another major benefit will be helping SMEs analyze more data, doing it faster and recognizing new patterns in data not feasible before. This will improve scientific discovery and engineering designs and allow scientists to spend significantly more time performing analysis rather than waiting on algorithms.

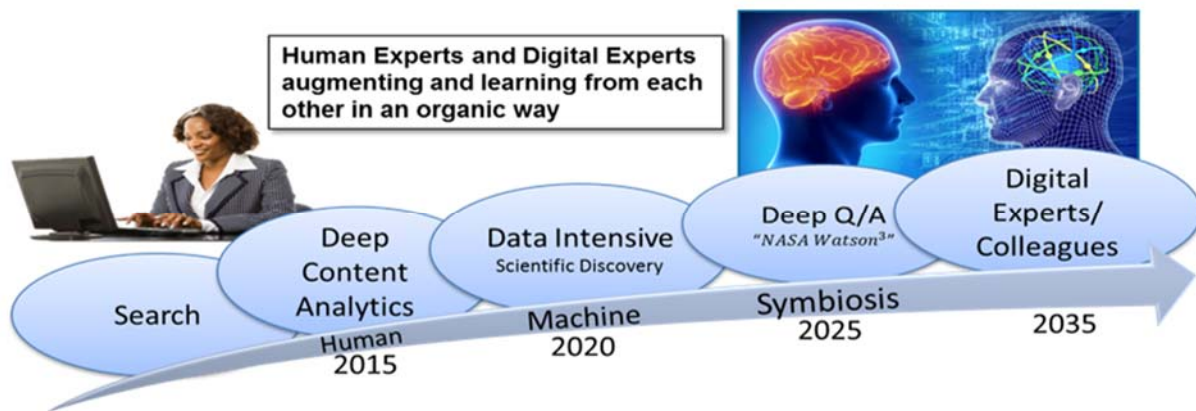


Figure 22 BDAMI Vision

A high-level overview of the roadmap and associated timeline for development of the big data, deep analytics and machine intelligence capability at LaRC is included below. The overview was generated based on the team’s development of specific goals, objectives and initiatives, (as listed in Table 1 of Section A.) These were then distilled into a list of key recommendations to showcase the most critical needs for developing the capability.

Big Data & Machine Intelligence Roadmap

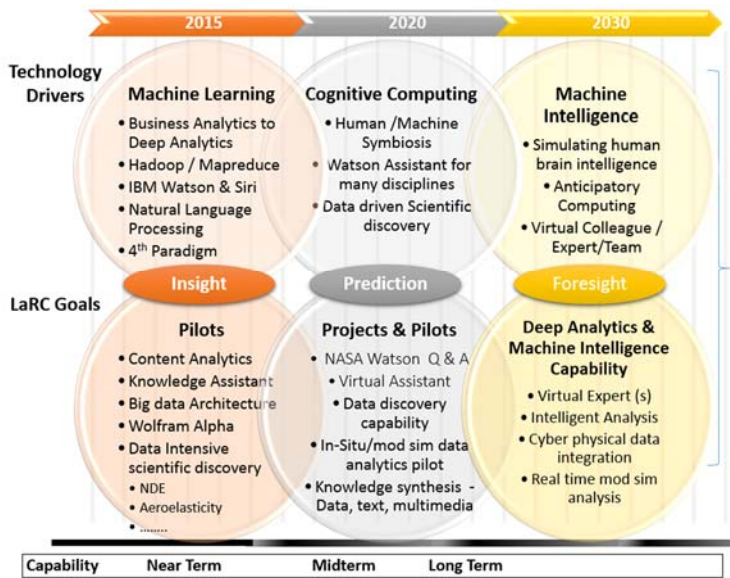


Figure 23 BDAMI Roadmap

In order to achieve the Mid-Term and Long-Term goals established in the *Big Data Analytics & Machine Intelligence Strategy and Roadmap*, work over the next three years will include additional focus on the remaining seven key recommendations, supported by the actions and outcomes outlined in Phase II of the CDT Phased Actions and Strategy. Those recommendations are listed below, and the Phase II actions/outcomes are described in Figure 24. We will also continue to work on the first five recommendations (near-term), that were outlined more extensively in Section B.

Near-Term Recommendations:

- **Key Recommendation 1:** Educate and promote the value of big data through seminars and workshops by experts and LaRC working group to foster the understanding of its value and use by mission organizations.
- **Key Recommendation 2:** Understand incubator needs and incorporate them with deep analytics and machine learning pilots and capability; Demonstrate feasibility and add value to incubator success.
- **Key Recommendation 3:** Build a big data and machine intelligence team, including data scientists, statisticians, algorithm developers, machine learning experts and comprised of civil service employees, contractors, and students.

- **Key Recommendation 4:** Develop and implement a data-driven scientific discovery capability; start with small-scale and high-value pilots: non-destructive evaluation (NDE) images, aeroelasticity data, and cybersecurity.

Mid-Term and Long-Term Key Recommendations:

- **Key Recommendation 6:** Identify and establish partnerships with universities, government and industry; leverage their expertise for LaRC's big data capability and participate in research when possible.
- **Key Recommendation 7:** Develop a data capture and management capability for automatic capture of data with context, including metadata standards and tagging, real-time uploads and ingests; start with a pilot.
- **Key Recommendation 8:** Develop a big data architecture capability; Research and understand technologies, tools and architectures and incorporate learnings from the pilots. Start with Hadoop and cloud pilots.
- **Key Recommendation 9:** Keep machine intelligence as a North Star goal by actively researching state-of-the-art developments, attending seminars /conferences; developing partnerships; and pursuing pilots with the Massachusetts Institute of Technology (MIT).
- **Key Recommendation 10:** Develop in-situ data analysis with M&S data and implement the capability of HPC, big data and M&S working together; start with pilots.
- **Key Recommendation 11:** Develop operational capability for virtual colleagues, experts and intelligent agents; start with pilots.
- **Key Recommendation 12:** Define and develop metrics for big data capabilities to demonstrate and communicate value to end users and leadership.

Phase II Actions (2018-2020)

Action	Outcome
Action 1: Data intensive scientific discovery capability in core disciplines	<ul style="list-style-type: none"> Ability to mine the both computational and mod sim data using discipline based algorithms for better science
Action 2: Develop in-situ analytics of real time data to combine with mod-sim and high performance computing	<ul style="list-style-type: none"> Highly integrated real time modeling and simulation with big data and machine intelligence
Action 3: Automate scientific data capture, tagging and integration for effective data management and architecture	<ul style="list-style-type: none"> Rapidly & organically capture, tag and store newly generated data
Action 4: Fully linked knowledge base of global and NASA multimedia and multilingual information and data	<ul style="list-style-type: none"> Develops the critical foundation for collaborative data/knowledge ecosystem Develop new research areas and systems significantly faster than before
Action 5: NASA Watson & Intelligent Agent pilots	<ul style="list-style-type: none"> Ability to sift through millions of articles and vast amount of data to find those of true value, and able to answer design questions with confidence
Action 6: Achieve Agency Level vision and buy-in	<ul style="list-style-type: none"> Critical for the funding and the vision VRDP to be reality NASA community can benefit from VRDP
Beyond: <ul style="list-style-type: none"> Integrate an IBM-Watson-inspired “Automated Research/Design Assistant” Transform LaRC’s ideation, innovation, and invention methods, to include machine-assisted techniques Virtual Research/Design Partner 	<ul style="list-style-type: none"> New paradigms of scientific work Enable NASA employees to achieve significantly greater scientific and engineering discoveries and systems innovation and optimization

Figure 24 Phase II Actions and Outcomes, 2018-2020

C2 Focus on NASA Langley Product Lines and Linkage to NASA Projects

As BDAMI capability development moves forward, ensuring that value-added deliverables support the NASA Langley Product Lines (Atmospheric Characterization; Systems Analysis and Concepts; Advanced Materials and Structural Systems; Aerosciences; Entry, Descent, & Landing; Measurement Systems; Intelligent Flight Systems) will be of primary importance in determining the actions and objectives that will best support the Center and the Agency; it will be equally important for the team to take part in planning discussions with mission directorates and project/program managers. The goal of participating in such meetings will be to identify key opportunities for the application of machine learning technologies to various project and product line goals. Supporting these projects will provide for significantly greater scientific and engineering innovation and optimization; developing effective collaborative relationships will help us to efficiently target product line needs, and define strategies to solve each domain’s complex challenges. To date, BDAMI has demonstrated significant progress in linking to product lines and programs, and efforts will be made to continue this effort. Current and upcoming projects have connected the CDT/BDAMI effort with the Advanced Composites Project, Convergent Aeronautics Solutions Project, Smart-NAS

Test Bed for Safe Trajectory-Based Operations Project, Airspace Technology Demonstrations Project, Space Launch System, and Safe Autonomous Systems Operations Project.

C3 Focus on Mission Support Functions

Acting as a catalyst for the development of machine learning across the Center will remain a priority. To date, much of the work has focused on missions. In the coming years, collaboration with mission support organizations will help to support and enhance centralized services, enabling employees to become more efficient and make better-informed decisions. Working to achieve agency-level buy in will be critical for the long-term vision, and can be supported by the cross-center collaborations that have been built over the past few years.

C4 Enhancing Data Intensive Scientific Discovery (DISD)

Developing machine learning tools that support SMEs in their everyday work will open paths to new scientific discovery by allowing scientists and engineers to gain new insights from their data, accelerating research and development. In the next three years, the team will continue efforts to build this capability, including work on the projects discussed in Section B, and through the identification of new use cases in scientific and engineering domains including computational materials, computational fluid dynamics, autonomous systems, trajectory-based optimization, system-wide safety, and air traffic control. Moving forward, main goals in this area include:

- Delivering solutions that are actively used by SMEs, and gathering feedback for improving those tools
- Leveraging knowledge from current work, identifying new areas where value can be demonstrated, and starting use cases in those areas
- Creating general-purpose machine learning toolsets for multiple types of data (time series, images, etc.) that will enable SMEs to independently explore applications of common methods

C5 Enhancing Deep Content Analytics (DCA)

New technologies in natural language processing and text analytics will result in significant time savings for SMEs, who will be able to more easily digest and navigate large volumes of scholarly and technical information. This, in turn, will provide them with the means to identify research trends, gain new insights, and connect with other researchers. Work on the use cases and projects outlined in Section B will continue, with the expectation that the capabilities of the Watson Content Analytics platform will be expanded in 2018. Capability enhancements will also come through the more intensive development of existing functions within the software, such as data visualization and faceted search improvements. As we move toward the longer-term objectives of the strategy, an investigation of cloud-based applications could be leveraged to support LaRC becoming a centralized Agency provider of content analytics services. Moving forward, main goals in this area include:

- Utilizing the deep functionality of Watson Content Analytics to provide customizable user interfaces, visualizations, faceted search, taxonomies, and ontologies to make information and analytics more usable and digestible

- Focusing on the development of collections built from open-source repositories such as the NASA Technical Reports Server, NASA Lessons Learned Information System, and arXiv.org, that will encourage use by a broad audience within the NASA technical community
- Working toward the establishment of a NASA-wide capability for Watson Content Analytics hosted at LaRC, and providing technical capability as a service

C6 Enhancing Deep Q&A and “NASA Watson”

Gaining a better understanding of the potential uses of cognitive assistants in NASA’s aerospace domains through proof of concept demonstrations will help to more accurately define requirements and scope deliverables for investments in these technologies. By concurrently leveraging capability development in DISD, DCA, and Deep Q&A, the value of the individual technology gains can be merged, advancing us toward the goal of implementing a “Virtual Research/Design Partner” that will support SMEs in their everyday tasks. Linking with other programs and directorates that have an interest in such technologies will provide the opportunity for additional proof of concepts. Currently, we have been exploring the use of IBM Watson applications including Watson Content Analytics and Watson Discovery Advisor. Taking advantage of similar technologies such from leading-edge developers at Google and Amazon will allow for experimentation with low-cost platforms that will help SMEs define how more advanced applications could be used in their domain. Moving forward, the main goals in this area include:

- Identifying opportunities for applying these tools in additional product lines and programs
- Collaborating with other NASA centers to develop a more comprehensive solution that can support Agency goals
- Developing additional generalized cognitive agents that can help all NASA employees with daily tasks such as search, scheduling, and project management

C7 Enhancing Collaboration

Leveraging external expertise from universities, industry, and other federal agencies will create a satellite capability with linkage back to the LaRC core team. This can act as a force multiplier for staying abreast of technology developments in cognitive computing and artificial intelligence. And, a fully-linked knowledge base of global and NASA multimedia and multilingual information and data will develop the foundation for a knowledge ecosystem that can support research synthesis by identifying gaps and overlaps. Moving forward, the main goals in this area include:

- Connecting resources from MIT CSAIL with NASA SMEs to take advantage of cutting-edge research and technology development for our complex challenges
- Establishing a strong relationship with Georgia Tech machine learning experts, for active collaboration on projects in both DISD and DCA
- Developing a relationship with CMU machine learning and cognitive science groups, and leveraging their expertise

C8 Artificial Intelligence

Artificial Intelligence, machine learning, deep learning, and cognitive technologies have made significant advancements and now are increasingly useful, pervasive, and capable. Recent reports from the Office of Science and Technology Policy (OSTP) and the White House indicate the potential of these technologies to further our society's economy and innovation. Staying abreast of research in these areas will not only further CDT and BDAMI goals, but could also facilitate a larger role the Center could take at the Agency and Aerospace ecosystem levels in infusing these transformational technologies for NASA mission challenges. Moving forward, the main goals in this area include:

- Helping to educate, engage and inspire all mission, mission support, and program organizations
- Developing a Blue Sky workshop to encourage employees in all areas of work at the Center – including mission, mission support, and infrastructure – to brainstorm potential applications to their challenges, and identify strategies for near, mid, and long-term actions
- Monitoring technology developments and initiating proof of concept efforts when appropriate

C9 Enhancing Outreach and Education

To date, several seminars, workshops, and Center-wide training events have provided value by helping SMEs gain fundamental skills in applying machine learning tools to their work. The events also create a networking opportunity to encourage cross-domain collaboration. In the coming years, outreach and education will continue to be supported by the CDT Seminar Series, training events hosted by the BDAMI team, and by bringing experts from industry and academia to the Center for short tutorials, multi-day workshops, and longer-term collaboration. Moving forward, the main goals in this area include:

- Propagating an understanding of machine learning and its applications in aerospace domains through workshops and seminars so that these technologies can be infused into all organizations
- Developing and advancing center-wide skills through a practical machine learning curriculum
- Hosting lectures and tutorials from experts in academia and industry

C10 Integration with High Performance Computing and Modeling & Simulation

Identifying challenges that will offer an opportunity to integrate the success of Big Data Analytics and Machine Intelligence capability development with High Performance Computing and Modeling & Simulation will help to more fully realize the long-term goals of the Comprehensive Digital Transformation (CDT). Domains such as In-Situ Space Assembly or Computational Materials may provide an avenue for collaboration and the advancement of technologies and capabilities in these three important pillars of CDT.

Concluding Remarks

Big Data analytics and machine learning technologies are maturing with a significant potential to help address NASA challenges. The overall vision of BDAMI is to have a Virtual Research and Design Partner enabling NASA to achieve greater scientific discoveries and system design optimizations; the multi-organizational team has embarked on working select projects and pilots in order to develop foundational capability in applying big data analytics and machine learning technologies to Langley's aerospace domain data and information, and will continue to pursue the objectives outlined in the previous sections as we move forward.

- *Data Intensive Scientific Discovery (DISD)* is developing a machine learning and data mining capability that will enable our SMEs to save time and make new discoveries using their complex experimental and computational data.
- *Deep Content Analytics (DCA)* is providing the technical community with natural language processing technologies that will quickly make sense of internal and global knowledge by identifying trends and experts, aiding in discovery, and finding answers to specific questions with evidence.
- *Cognitive Assistant Prototypes in Deep Q&A* are helping our Center evaluate the application of cognitive technology to our domains, and increasing our understanding of how we should prepare to take advantage of these technologies for our mission challenges.

Key successes from 2014 – 2017 include:

- Developing a broad and far-reaching vision with detailed goals, recommendations, and a roadmap, and making substantial progress by following these plans.
- Success in building the Center's expertise in data analytics and machine learning by recruiting civil servants, contractors, and interns with skills in computer science, mathematics, statistics, and information technology.
- Collaboration between a young, multi-skilled team and subject matter experts from nondestructive evaluation, aeroelastic flutter, crew safety monitoring, aerospace design, autonomous systems, human-machine teaming, space radiation, vehicle design, and computational fluid dynamics has helped many researchers see that there is a potential for machine learning technologies to enable them to do better science and engineering.
- Projects and pilots have cut across many disciplines and product lines areas, with linkage to Agency and Langley projects in the Aeronautics Research Mission Directorate (ARMD), with significant funding for projects provided to augment Center Management and Operation (CMO) funding.
- Projects and pilots in Data Intensive Scientific Discovery (DISD) have focused on automated identification of anomalies to assess structural damage in nondestructive evaluation; predicting flutter from aeroelasticity data; crew cognitive state monitoring and detection; rapid exploration of aerospace designs; turbulence modeling; entry, descent, and landing; climate science; Space Launch System (SLS) booster separation aerodynamics; SLS additive manufacturing; hypersonic inlet performance; flight deck analytics from trajectory-based optimization.
- Projects and pilots in Deep Content Analytics (DCA) have focused on carbon nanotube research, autonomous flight, space radiation, aerospace vehicle design, uncertainty quantification, space mission

analysis, model-based engineering, human-machine teaming, NASA Technical Reports, and NASA Lessons Learned.

- Proof of concepts in Deep Q&A have included both a Pilot Advisor and Aerospace Innovation Advisor.
- Leveraging expertise through strong collaboration with NASA Ames Research Center, Georgia Tech, MIT and IBM; this helped to build an ecosystem to enhance our local expertise in transformational data analytics and machine learning technologies.
- Engaged and educated both the BDAMI team and the technical community with more than a dozen Center-wide seminars, workshops, and tutorials.
- Coordinated the *Machine Learning Technologies and Their Applications for Science and Engineering Domains Workshop*, a three-day event on Center, with lectures from 15 leading experts.
- Gaining recognition both internally and externally as a capability team that is developing the expertise to work with NASA mission challenges through presentations at the AIAA Big Data Session, World of Watson cognitive computing event, AIAA symposium, and as co-partners in Small Business Technology Transfer (STTR) Artificial Intelligence topic.

Moving forward, the key challenges are:

- Machine learning technologies applications in scientific and engineering domains expertise are in nascent stages, and require smart investigation and experimentation to develop workable solutions. This is being addressed with a scientific methodology and by leveraging open-source machine learning techniques and university expertise.
- Data and information in our domains is tied to physics, which needs to be both an implicit and explicit part of algorithms; most machine learning techniques require substantial amounts of labeled data, which is not easily available in our domains. This is being worked by influencing subject matter experts to consider the importance of data capture and labeling during the design of experiments.
- Obtaining and keeping core skills in computer science, machine learning, and natural language processing requires creativity; using a small core team at the Center with access to expertise from universities and industry, along with strong collaboration from SMEs is critical. Some personnel challenges may have to be addressed with non-traditional approaches to be successful.

Opportunities for moving forward are many, as we are poised to take advantage of our experience over the last three years. Many NASA projects can benefit from machine learning and artificial intelligence technologies, and as these capabilities mature, we can make significant contributions to mission success. Data Intensive Scientific Discovery projects and pilots are helping immensely in understanding and applying cutting-edge machine learning and deep learning techniques to our aerospace data for knowledge extraction and discoveries. The natural language processing analytics and cognitive technology projects using Watson software are helping us to develop a similar deep understanding and application of these technologies to aerospace domains. The expertise and experience we are gaining from working in these areas will provide us with the insight to architect solutions for NASA challenges using machine learning, cognitive computing, and artificial intelligence technologies. As the Comprehensive Digital Transformation starts to work toward the goals of vehicle flight prediction and digital twin, the need for integrating machine learning with high performance computing and modeling & simulation is critical, and the Big Data Analytics and Machine Intelligence Capability can make a significant contribution toward this progress.

Appendix A: References for Section A

Austin, Tom; Blau, Brian; McGee, Ken; Prentice, Stephen; and Ghubril, Adib. "Carl Austin Predicts 2014: The Emerging Smart Machine Era," Gartner, November 21, 2013.

"Big Data – FCW Custom Report," 2013, Brocade.

"Big Data Now – Current Perspectives from O'Reilly Media," October 24, 2012, O'Reilly Media.

"Frontiers in Massive Data Analysis" 2013 National Research Council

Gaudin, Sharon. "Google on new path, developing self-driving cars," computerworld.com, October 11, 2010.

Gaudin, Sharon. "IBM's Watson's ability to converse is a huge advance for AI research," computerworld.com, February 15, 2011.

Gaudin, Sharon. "Look out Siri! Google Now taking a bite out of Apple," computerworld.com, April 29, 2013.

Hardesty, Larry. "Artificial-intelligence research revives its old ambitions," MIT News Office, 2013.

Hey, Tony, Tansley, Stewart, and Tolle, Kristin (editors). *The Fourth Paradigm: Data-Intensive Scientific Discovery*, Microsoft Research, 2009.

Honigsbaum, Mark. "Human Brain Project: Henry Markram plans to spend 1bn building a perfect model of the human brain," theguardian.com, October 12, 2013

Lapkin, Anne. "Hype Cycle for Big Data, 2012," Gartner, July 2012.

Laney, Doug. "Innovating with Information: Art of the Possible," Gartner, 2012.

Liebowitz, Jay. *Big Data and Business Analytics*, CRC Press, 2013.

Merlin. "Managing Big Data," FCW.com, 2012.

Mervis, Jeffrey. "Agencies Rally to Tackle Big Data," *Science*, pg. 22, 6 April 2012.

National Research Council. *Frontiers in Massive Data Analysis*. Washington, D.C.: The National Academies Press, 2013

Office of The President, Fact Sheet: Data to Knowledge to Action, Washington D.C., November 12, 2013.

Provost, Foster, and Fawcett, Tom. *Data Science for Business – What you Need to Know About Data Mining and Data Analytic Thinking*, July 2013, O'Reilly Media

Zikopoulos, Paul, Deroos, Dirk, Parasuraman, Krishnan, Deutsch, Thomas, Corrigan, David, and Giles, James. *Harness the Power of Big Data – The IBM Big Data Platform*, 2013, IBM.

Appendix B: References for Section B

Ciresan, Dan, et al. (2012) Deep neural networks segment neuronal membranes in electron microscopy images. *Advances in neural information processing systems*.

Harrivel, A. R., Liles, C. A., Stephens, C. L., Ellis, K. K., Prinzel, L. J., and Pope, A. T. (2016). *Psychophysiological Sensing and State Classification for Attention Management in Commercial Aviation*. AIAA SciTech.

Keogh, E., Lin, J., Fu, A. (2005). Hot SAX: Efficiently finding the most unusual time series subsequence. In *Data mining, Fifth IEEE International conference on*, pages 8-16.

Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. (2012). *Imagenet classification with deep convolutional neural networks*. *Advances in neural information processing systems*.

Lin, J., Keogh, E., Lonardi, S., and Patel, P. (2002) Finding motifs in time series. In *Proc. of the 2nd Workshop on Temporal Data Mining*, pages 53-68.

Minnen, D., Starner, T., Essa, I., & Isbell, C. (2006, October). Discovering characteristic actions from on-body sensor data. In *Wearable computers, 2006 10th IEEE international symposium on* (pp. 11-18). IEEE.

Mueen, A. (2013). Enumeration of time series motifs of all lengths. In *Data Mining (ICDM), 2013 IEEE 13th International Conference on*, pages 547-556. IEEE.

Mueen, A., Keogh, E. (2010). Online discovery and maintenance of time series motifs. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1089-1098. ACM.

Mueen, A., et al. (2009). Exact discovery of time series motifs. In *SDM*, pages 473-484. SIAM

Patnaik, D., Marwah, M., Sharma, R., & Ramakrishnan, N. (2009, June). Sustainable operation and management of data center chillers using temporal data mining. In Proceedings of the 15th ACM SIGKDD international conference on knowledge discovery and data mining (pp. 1305-1314). ACM.

Schwabacher, M., Oza, N., & Matthews, B. (2009). Unsupervised anomaly detection for liquid-fueled rocket propulsion health monitoring. *Journal of Aerospace Computing, Information, and Communication*, 6(7):464-482.

Scott, Robert C., et al. "Aeroservoelastic Wind-Tunnel Test of the SUGAR Truss Braced Wing Wind-Tunnel Model." (2015). 56th AIAA/ASCE/AHS/ASC Structures, Structural Dynamics, and Materials Conference.

Silva, Walter A., et al. "An Overview of the Semi-Span Super-Sonic Transport (S4T) Wind-Tunnel Model Program" (2012). 53rd AIAA/ASME/ASCE/AHS/ASC Structures, Structural Dynamics and Materials Conference.

Syed, Z., Stultz, C., Kellis, M., Indyk, P., & Guttag, J. (2010). Motif discovery in physiological datasets: a methodology for inferring predictive elements. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 4(1), 2.

Wang, Chaghan, et al. (2015). A unified framework for automatic wound segmentation and analysis with deep convolutional neural networks." *Engineering in Medicine and Biology Society (EMBC), 2015 37th Annual International Conference of the IEEE*.

Zhang, G., Xu, R., Wang, W., Pepe, A. A., Li, F., Li, J. McKenzie, F., Schnell, T., Anderson, N., and Heitkamp, D. (2012). *Model Individualization for Real-Time Operator Functional State Assessment*. CRC Press.

Appendix C: Key Presentations

The slide decks included in this section are intended to provide a high-level overview of the many aspects of the Big Data Analytics and Machine Intelligence Capability Development at NASA Langley Research Center.

Big Data Analytics and Machine Intelligence Capability Algorithms and Software: AIAA SciTech Conference (January 2016)



Comprehensive Digital Transformation

Enable Innovative Solutions to Complex NASA Mission Challenges

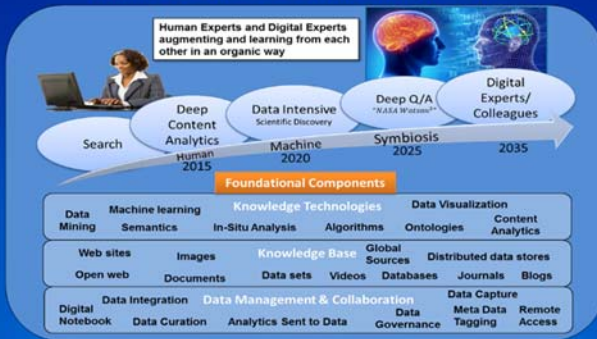
- > Open ideation and innovation
- > Focused, relevant research
- > Intelligent and rapid system designs
- > Agile response to emerging missions

Modeling & Simulation <ul style="list-style-type: none"> • Physics-based understanding and simulation – improved discipline tools • Integrated analysis and design of complex systems • Optimally combine testing and M&S 	Data Analytics & Machine Intelligence <ul style="list-style-type: none"> • Rapid synthesis and digestion of global scientific information for knowledge extraction, insights and answers • Mining of diverse computational and experimental data sets for new correlations, discoveries and advanced designs • Virtual/Digital Experts – Human Machine symbiosis
High Performance Computing <ul style="list-style-type: none"> • Next generation code development • Rapid Compute power for M&S and BDA&MI • Architecture for real-time analysis and design 	Advanced IT <ul style="list-style-type: none"> • Open, secure collaboration with NASA & partners • Networks handle burgeoning data • Data governance, architecture, and management

External collaboration is of paramount importance



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



Projects and Pilots
Towards
Virtual Partners

Two Key Areas for Virtual Partner – Data Intensive Scientific Discovery

Data Intensive Scientific Discovery (DISD)

Deriving new insights, correlations, and discoveries not otherwise possible from our diverse experimental and computational data sets
The Fourth Paradigm

Projects & Pilots

- Anomaly Detection in the Nondestructive Evaluation of Materials images
- Predicting Flutter from Aeroelasticity Data
- Cognitive Assessment of Crew/Pilot State
- Knowledge Bot for Complex Simulation Software Optimization
- Rapid Exploration of Aerospace Designs
- Entry, Descent, and Landing Trajectory Data Analysis



The variety of techniques and diverse datasets used in these projects and pilots represents a cross-cutting approach to solving complex, physics-based problems in multiple aerospace domains.



Data Intensive Scientific Discovery Projects - 1

Anomaly Detection in Non-Destructive Evaluation of Materials Images



Develop techniques and algorithms to automatically detect anomalies during the nondestructive evaluation of materials

Goals

- Significantly reduce SME analysis time and assist experts in discovering additional anomalies
- Help to design better material compositions and structures

Techniques

- Two-Dimensional Regression designed to detect anomalous pixels
- Convolutional Neural Networks to classify the image data

Accomplishments & Next Steps

- Algorithms are validated with real data sets and further enhanced
- Deliver a tool with a good UI for SMEs to use as an "Assistant" for anomaly detection of composite materials analysis in March

Predicting Flutter from Aeroelasticity Data



Develop methods to automatically detect the onset of flutter during wind tunnel testing

Goals

- Find new ways of predicting flutter in the time domain
- Identify non-traditional predictor variables and unseen patterns
- Better understand precursors to flutter and improve configurations

Techniques

- Piecewise Regression to locate peaks, track coalescence of structural modes
- Time-Series Motifs to identify signatures in the data that could represent precursors to flutter

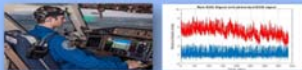
Accomplishments & Next Steps

- Peak detection tested with multiple datasets
- Several significant time series motifs detected
- Generating synthetic data for validation of algorithms



Data Intensive Scientific Discovery Projects - 2

Cognitive Assessment of Crew State Monitoring



Build classification models for predicting cognitive state using physiological data collected during flight simulations

Goals

- Identify unsafe cognitive states in aircrew real-time
- Apply results for more effective pilot training

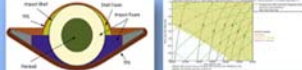
Techniques

- Ensemble of machine learning tools (deep neural network, gradient boosting, random forest, support vector machine, decision tree)
- Data pre-processing using detrending and power spectral density

Accomplishments & Next Steps

- Initial data mapping, statistical analysis, and signals processing
- Support classification efforts on single modalities
- Explore combining multiple signal models using ensembling

Rapid Exploration of Aerospace Designs



Develop a generalized machine learning platform to be used for analyzing mod-sim data for design optimization

Goals

- Provide surrogate modeling to explore the trade space of aerospace vehicle designs with easy-to-use web interface
- Use fast machine learning models instead of computationally-intensive code for rapid exploration and optimization

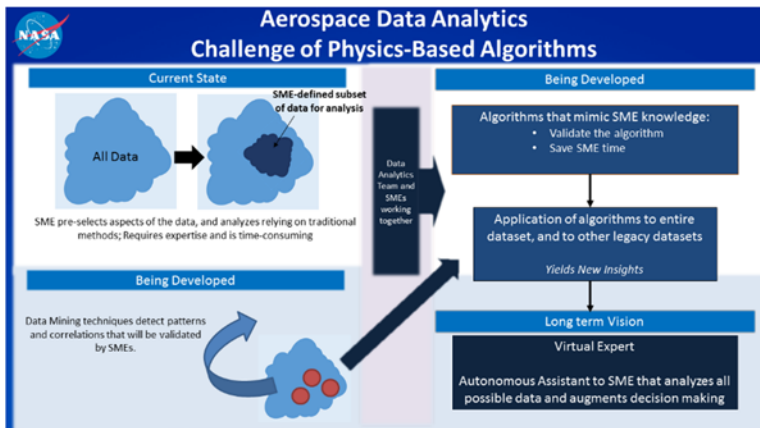
Techniques

- Supervised machine learning algorithms, SVM, and Neural Networks will be trained on labeled data

Accomplishments & Next Steps

- Python 2.7 with SKLearn algorithms are being used
- Windows Server 2012 with PHP set up for web interface





Two Key Areas for Virtual Partner- Knowledge Analytics

Knowledge Analytics (KA)

Obtaining insights, identifying trends, aiding in discovery, and finding answers to specific questions by mining knowledge from scholarly, web, and multimedia content

Cognitive Computing

Knowledge Assistants
Using Watson Content Analytics

- Carbon Nanotubes Research
- Autonomous Flight Research
- Space Radiation Research

Aerospace Innovation Advisor POC
Using Watson Discovery Advisor

Example Topics:

- Hybrid Electric Propulsion
- On Demand Mobility

Cognitive-based systems are able to build knowledge and learn, through understanding natural language, to reason and interact more naturally with human beings than traditional systems. They are also able to put content into context with confidence-weighted responses and supporting evidence. Uses Natural language processing, machine learning, and speech recognition technologies.

Knowledge Assistants *Using Watson Content Analytics*

<p>Carbon Nanotubes research</p> <p>130,000 articles metadata of ~20 year literature analyzed. Identified experts, trends, insights, and connections. Buying scholarly content is a challenge.</p>		<p style="text-align: center;">Key Capabilities</p> <ul style="list-style-type: none"> • Digest and analyze thousands of articles without reading with ability to dissect the content interactively • Automatically identify subsets of documents from a large corpus and provide gists • Provide a means to rapidly identify trends and connections • Identify experts and connections among them at all levels and affiliations • Explore technology gaps that could be leveraged • Replace traditional methods of SMEs manually reviewing and tracking research • Help to identify cross-domain leverages and research
<p>Autonomous Flight</p> <p>4000 metadata and full-text articles analyzed. Integrate analysis of scholarly and informal web content to identify experts and new partnerships</p>		
<p>Space Radiation Research</p> <p>1000 metadata and full text articles analyzed. Using identified possible duplications, connections, and technology gaps. In the process of analyzing all six elements of Human Research Program</p>		

Successfully demonstrated value and developed robust expertise
Making it available as a Center capability to the technical community

Watson Discovery Advisor



Accelerate the discovery of new insights by synthesizing information in seconds

- Take advantage of massive sources of data
- Find answers to questions that have not been asked yet or answered before
- Find insights into hidden relationships and dig deeper
- Generate leads to hard questions and provide evidence to substantiate new claims

Cognitive Technologies for Aerospace

Aerospace Innovation Advisor

Proof of Concept (March – June 2016)

Apply cognitive computing technologies that 'understand' massive amounts of information and enhance experts' abilities
Example Topics: Hybrid Electric Propulsion, On-Demand Mobility

Ames Research Center:

Aircraft Dispatcher Assistant
Feasibility Study

Armstrong Flight Research Center:

Pilot Assistant
Investigation

Johnson Space Center:

Astronaut Health
Investigation(s)

LoRC is connected with all of these efforts



Algorithms and Techniques



Linear regression

Application 1:

Non-Destructive Evaluation (NDE) Image analysis

Goal:

Automate delamination detection

Method: Fit data with linear regression and detect outlier regions. Regression performed on 1D and 2D signals

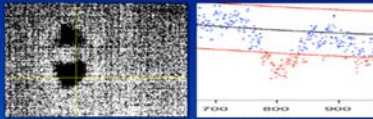
Application 2:

Aeroelastic Flutter Data Analytics

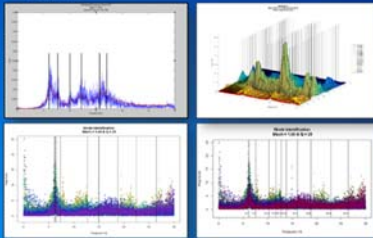
Goal:


Detect precursors and onset of aeroelastic flutter

Method: Fit best quadratics between structural modes to detect mode coalescence



Top: Linear regression of 1D-signals for anomaly detection in carbon fiber; Bottom: Mode identification in flutter time-series data using linear regression






Gaussian process

Application: Knowledge Bot for Simulated Software


Goal: Emulate simulation to predict convergence/divergence

Method: Gaussian Process for emulator and to find next best point to maximum knowledge

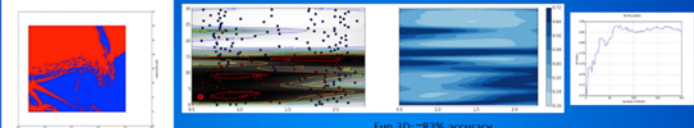
Justification: Approach followed in literature for weather simulation emulation




Methodology



Finding boundary of two circles: ~99% accuracy



Fun 3D: ~83% accuracy



Time Series Motifs


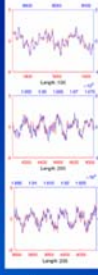
Application: Pattern mining of time domain Aeroelastic Flutter Data

Goal: Identify flutter precursors to:

- Create a dictionary of motifs for a given configuration
- Classify data for use with machine learning algorithms that will support a real-time 'Flutter Assistant'

Method: Application of the Motif Enumeration (MOEN) algorithm created by Dr. Abdullah Mueen; Open-source framework


Justification: MOEN has been successfully applied to research problems in other scientific domains including robotics, biology, and seismology

In: Robert C. et al. "Nonparametric, Model-Free, Test of the 300th Test Bed Wing with Sensor-Based 'IoT' (NASA/DARPA/NSA) Structures, Structural Dynamics, and Materials Conference 2015.

In order to detect motifs across the various sensor signals, a given sensor's output (Signal A) is compared to another sensor (Signal B) by creating a composite signal (Signal A/B).

The algorithm is then applied to the composite signal to detect the motifs (above right) common to both sensors. Significant motifs are identified by a physics-based selection process and then validated by SMEs.




Deep learning: convolutional neural network (CNN)


Application: NDE Image Analysis to segment delaminations

Method: Convolutional encoder/decoder neural network; end-to-end training to map raw data to segmentation; Using Caffe


Justification: Very successful in medical image analysis such as wound segmentation (top right)



From Wang, et al. "Transfer learning for composite image defect and analysis: An end-to-end, multi-scale, deep learning architecture" (BMC) 2015. The International Conference on the IEEE, 2015.



Results on Simulated Data




Results on Experimental Data

Artificial neural networks (ANN)

Application 1: Crew State Monitoring

Goal: Build classification models capable of accurate, real-time prediction of aircrew cognitive state using physio data collected during flight simulations

Method: ANN trained to classify cognitive state



EEG | ECG | Respiration Rate | Galvanic Skin Response | Eye Tracking

Feature Generation

Input Layer

Hidden Layer

Output Layer / Classification

"Normal" State | Channelized Attention | Diverted Attention | Startle / Surprise

Application 2: Rapid Exploration of Aerospace Designs (READ)

Goal: Build classification / regression models on user-uploaded data for aerospace designs

Method: train ANNs on labeled data, use trained models for prediction and visualization

Ensemble of Machine Learning Techniques

Application 1: Non-Destructive Evaluation (NDE) Image analysis

Goal: Automate delamination detection

Method: Combine several machine learning models into overall prediction using regression to determine if sample contains a delamination; Using MatLab

Random forests	Extremely random forests	Ada Boost	Gradient boosting	k-nearest neighbors
Fully grow k independent classification trees and combine predictions	Similar to random forests but split per node is also randomized	Fit consecutive weak learners based on classification tree stumps and combine predictions	Similar to Ada Boost with different loss function	Identify k closed points to test sample based on distance metric

EEG | ECG | Respiration Rate | Galvanic Skin Response | Eye Tracking

Feature Generation

Level 1 Models

Artificial Neural Network | Gradient Boost Classifier | Random Forest

Level 2 Meta Model

Artificial Neural Network

"Normal" State | Channelized Attn | Diverted Attn | Startle / Surprise

Comprehensive Digital Transformation: An Overview (February 2016)



Comprehensive Digital Transformation (CDT)

An Overview

Damodar Ambur

Executive Lead for CDT, Office of the Center Director

Joseph Morrison (M&S Lead)

Manjula Ambur (BDA & MI Lead)

Dana Hammond (HPC Lead)

Edward McLarney (Advanced IT Lead)

NASA Langley Research Center

February 10, 2016

1



Outline

- | | |
|---|---------------|
| • CDT Overview | Damodar Ambur |
| • Modeling and Simulation | Joe Morrison |
| • Big Data Analytics and Machine Intelligence | Manjula Ambur |
| • High Performance Computing | Dana Hammond |
| • Advanced Information Technologies | Ed McLarney |
| Q&A | All |

2

NASA Vision, Mission and Strategic Goals and Langley Strategic Focus Areas

Deliver on Today's Commitments **Create Tomorrow's Opportunities**

NASA VISION
We reach for new heights and reveal the unknown for the benefit of humankind

NASA MISSION
Drive advances in science, technology, aeronautics, and space exploration to enhance knowledge, education, innovation, economic vitality, and stewardship of Earth

NASA Strategic Goals

GOAL 1
Expand the frontiers of knowledge, capability, and opportunity in space.

GOAL 2
Advance understanding of Earth and develop technologies to improve the quality of life on our home planet.

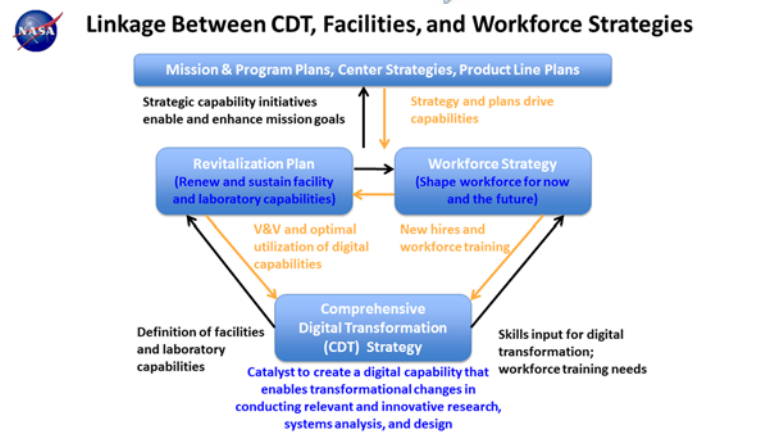
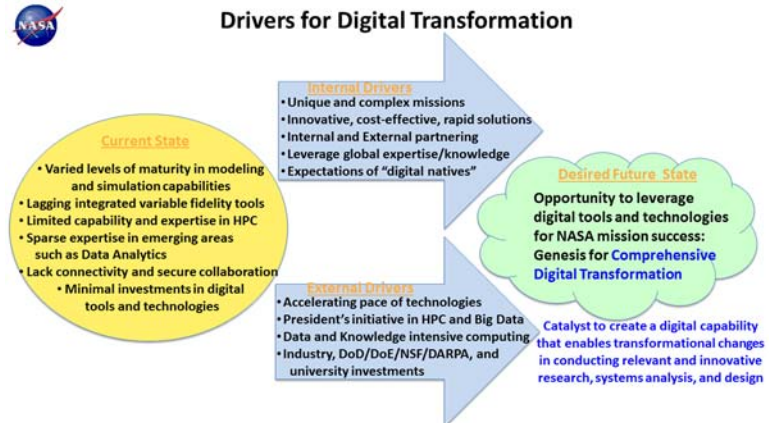
GOAL 3
Serve the American public and accomplish our Mission by effectively managing our people, technical capabilities, and infrastructure.

Langley Strategic Thrusts

- LARC viewed as a HELIOS/ STMD Strategic Investment
- Apollo-Earth Experiment - Terrestrial RAD Critical Research
- Key player in the Asteroid Redirect Mission
- Development and Use of Smart, Smarter Sensors
- Innovative Content in Airspace Ops and Safety Program
- Radical Concepts in the Advanced Air Vehicles Program
- Develop Aeronautics Research Showcase Vehicles
- Create Next Science Mission/Instrument
- Alternative/Affordable Platforms for Earth Science
- Revitalization Plan
- Workforce Strategy
- Comprehensive Digital Transformation

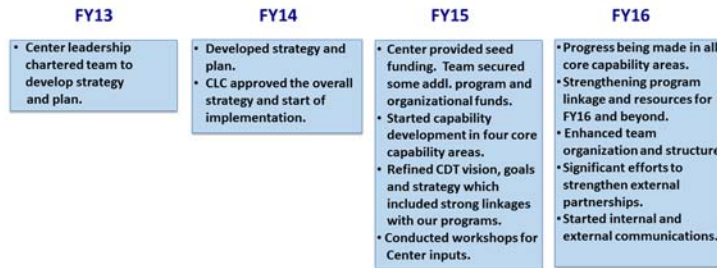
LANGLEY VISION

- On-Demand Air Transportation
- Understanding, Adapting to, and Mitigating the Earth's Climate System
- Human on Mars in the 2030's



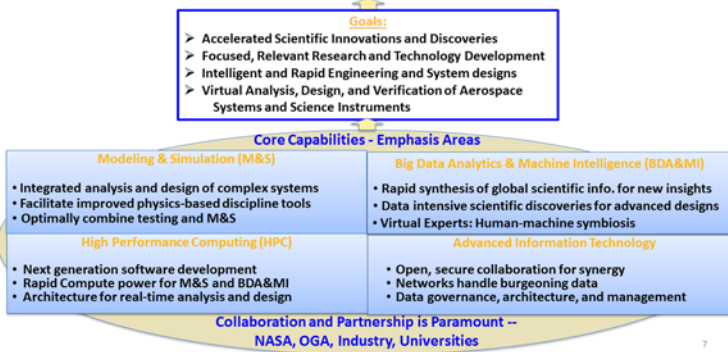


Summary Progress

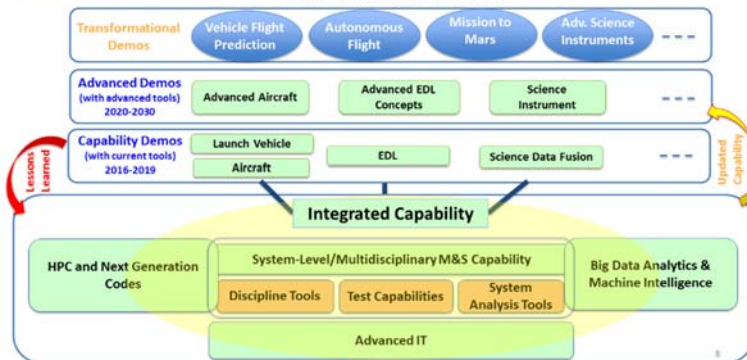


Comprehensive Digital Transformation

Vision: Catalyst to Enable Transformative Solutions to NASA Mission Challenges



Virtual Analysis and Design of Aerospace Systems and Science Instruments





CDT Organizational Structure and Team



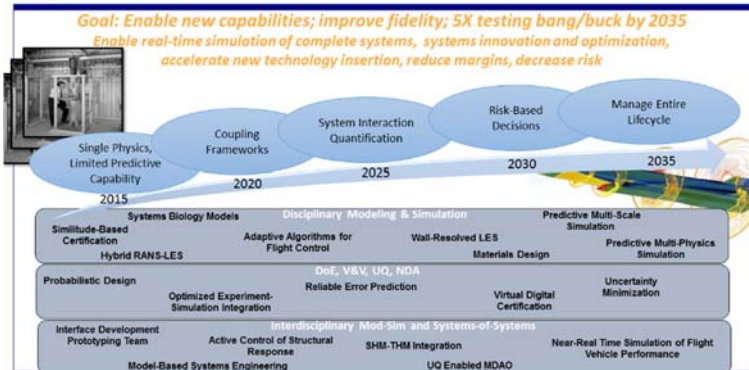
Summary of Strategic Approach

- Agency needs digital technologies
 - > To develop transformative solutions to complex mission challenges
 - > To conduct relevant and innovative research and rapid systems analysis and design
 - > In partnership within and external to the Center/Agency
- CDT initiative is a catalyst to create and apply digital capabilities to benefit Agency mission
 - > Leverage advancements from externals in all digital tools and technologies
 - > Utilize seed investments to demonstrate capabilities through pilots
 - > Leverage current program work to demonstrate benefits
 - > Advocate needed capability advancements in alignment with program goals
 - > Facilitate capability demonstrations to enable transformational solutions
- CDT is an important journey into the future
 - > Current and advanced technologies to enable transformation
 - > Pervasive collaboration for the "greater good"

10



Vision: Modeling, Simulation and Systems Analysis Capabilities





Modeling, Simulation and Systems Analysis Capabilities

- **Enable capabilities to analyze, understand, predict, and measure performance of complex physical behavior and phenomena with high confidence**
 - High fidelity simulations to explore discipline physics, e.g., turbulence, transition, fracture mechanics, radiation
- **Enable design of new aerospace concepts with reduced margins, improved performance, reduced emissions, greater reliability, longer life**
 - Rapidly optimize multi-functional materials
 - Extended vehicle life and reduced maintenance costs
- **Enable design of aerospace systems of systems to accomplish complex, long term, high impact mission goals**
 - Design integrated radiation protection system for long-term missions integrating multiple technologies to mitigate risk (improved materials, active shielding, radiation effect mitigation through medical advances)
 - Large-scale, live, virtual, constructive simulation of airspace architecture
- **Fusion of modeling, simulation, and experimental data with big data analytics to evaluate impact and benefits of technology infusion and the latest research innovations**
 - Link Langley data with Agency, national and worldwide models to develop climate mitigation solutions
 - Verified, validated M&S with quantified uncertainty for intended application
- **Enable risk-based decisions**

Leading to accelerated ideation and design of revolutionary aerospace systems



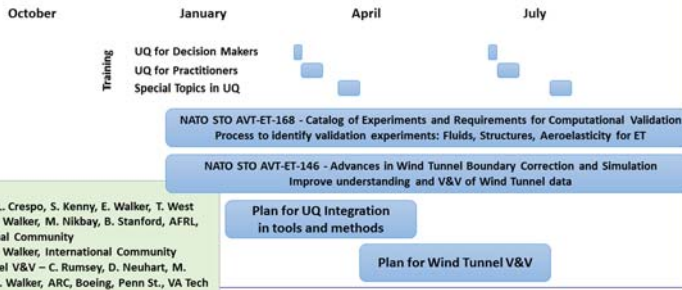
Summary of Modeling & Simulation Accomplishments - FY15

- Led comprehensive re-write of the Agency Modeling & Simulation & Information Technology Technical Area Roadmap (TA11) – Glaessgen
- Experimented with a 4 x 10% whitespace M&S leadership model for CDT – found it did not allow for sufficient focus
- Research Directorate provided M&S leadership focus by establishing a new Associate Director position for M&S leadership – Morrison
- Presented Phase II options to CLC, resulting in Vehicle Flight Prediction's selection as the first Phase II capability – Bauer
 - Presented Vehicle Flight Prediction planning to Aerodynamics Technical Working Group at AIAA Aviation Conference (June 2015)
 - Aerodynamics Technical Working Group and multiple commercial companies endorsed one or more Flight Prediction Workshops
- Provided seed funding for three CDT-aligned M&S integration projects:
 - Cyber-physical UAS linkage experiment – cloud-based prototype to link UAS GPS and health data between real world systems and simulated systems (Glaab, Chung)
 - Multi-disciplinary integration – developing and validating codes to link damage, sensor, and maintenance modules (Hochhalter, Leckey, Warner)
 - NASA-USAF-DoE microstructure interoperability – link LaRC microstructures code with AFRL statistical model; partner with AFRL, Sandia, LaRC



Verification, Validation and Uncertainty Quantification

Verified and validated M&S with quantified uncertainty to enable risk-based decisions



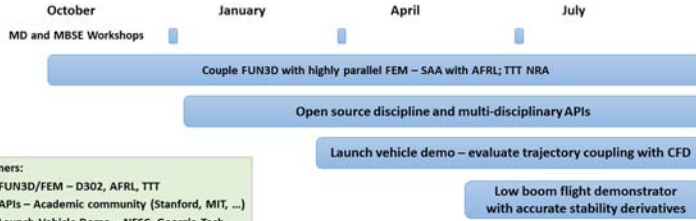
Partners:

- Training – L. Crespo, S. Kenny, E. Walker, T. West
- ET-168 – E. Walker, M. Nikbay, B. Stanford, AFRL, International Community
- ET-146 – E. Walker, International Community
- Wind Tunnel V&V – C. Rumsey, D. Neuhart, M. Kegerise, E. Walker, ARC, Boeing, Penn St., VA Tech



Multi-Disciplinary Framework and API

- Enable capabilities to analyze, understand, predict, and measure performance of complex physical behavior and phenomena with high confidence
- Enable design of new aerospace concepts with reduced margins, improved performance, reduced emissions, greater reliability, longer life
- Enable design of aerospace systems-of-systems to accomplish complex, long term, high impact mission goals
- Leverage multiple multi-disciplinary activities across LaRC, the Agency, academia, OGA, and external partners



Partners:

- FUN3D/FEM – D302, AFRL, TTT
- APIs – Academic community (Stanford, MIT, ...)
- Launch Vehicle Demo – NESC, Georgia Tech
- Low Boom – M. Fremaux



Vehicle Flight Prediction

- Enable Certification/Qualification by Analysis
- Reduce flight testing
- Enable system analysis capability that is overarching from concept-to-flight
- Enable improved confidence in models
- Enable physics-based models of every system/subsystem



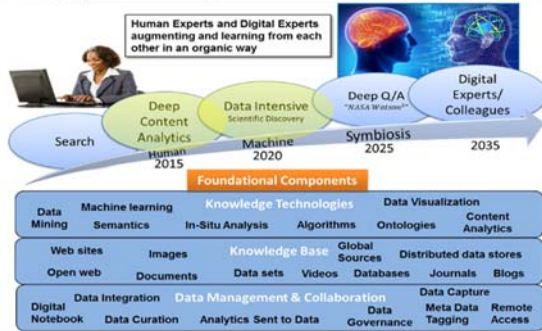
Partners:

- VFP – S. Bauer, A. Washburn
- FPW – NESC (D. Schuster), S. Bauer, A. Washburn, Glenn, Armstrong, Ames, Johnson, Army, Navy, AF, Corporations



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

Enable NASA employees to achieve greater scientific discoveries and systems innovations





Data Intensive Scientific Discovery: Aerospace Data Assistants



Data Intensive Scientific Discovery (DISD)/Fourth Paradigm

Deriving new insights, correlations, and discoveries from diverse experimental and computational data sets

Anomaly Detection in the Non-Destructive Evaluation of Materials Images

- **Goal:** Save SME time and design better material compositions and structures
- **FY16 Outcome:** SMEs to use as an 'Assistant' in composite panels analysis and evaluation
- **Algorithms & Tools:** Convolutional Neural Networks, 2D Regression, Random Forest, Python, Torch
- **Program Linkage:** ACP/ARMD
- **Organizational Linkage:** NESB/RD



Predicting Flutter from Aeroelasticity Data

- **Goal:** Detect the onset of flutter with non-traditional predictor variables and improve configurations
- **FY16 Outcome:** SMEs Test algorithms in TDT testing and enhance algorithms to predict flutter
- **Algorithms & Tools:** Piecewise Regression, Time Series Motifs, MATLAB, R, MOEN
- **Program Linkage:** CST and TTT/ARMD
- **Organizational Linkage:** AEB/RD



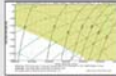
Aerospace Data Assistants (Cont.)



Data Intensive Scientific Discovery (DISD)/Fourth Paradigm

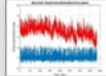
Rapid Exploration of Aerospace Designs

- **Goal:** Provide a platform that analyzes modeling and simulation data for design optimization
- **FY16 Outcome:** Web-based system with algorithms that could be used for MSAPE, EXAMINE and IDEAS data
- **Algorithms & Tools:** Support Vector Machines, Gradient Boost, k-means clustering, Python
- **Program Linkage:** STMD, ARMD
- **Organizational Linkage:** VAB, SADC



Cognitive Assessment of Crew State Monitoring

- **Goal:** Predict cognitive state and identify unsafe conditions leading to better pilot training
- **FY16 Outcome:** Use of multi-modality classifiers/models for real time simulator data analysis
- **Algorithms & Tools:** Random Forrest, Deep Learning, Gradient Boost; Python, Theano, MATLAB
- **Program Linkage:** ATD/TASA/ARMD
- **Organizational Linkage:** CSAOB/RD



Aerospace Data Assistants (Concluded)



Data Intensive Scientific Discovery (DISD)/Fourth Paradigm

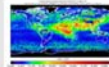
Entry Descent Landing Trajectory Analysis

- **Goal:** Apply machine learning algorithms for trajectory data analysis, helping to integrate data from various disciplines for better design
- **FY16 Outcome:** Compare current analysis results from MSL runs with machine learning analysis methods
- **Algorithms & Tools:** To be determined
- **Program Linkage:** STMD
- **Organizational Linkage:** EDLB/ED





Climate Science Data Analysis

- **Goal:** Apply machine learning for enhanced climate data fusion, visualization, and analysis for new insights
- **FY16 Outcome:** Develop a AIST proposal for data fusion in the Cloud use case, in collaboration with Ames
- **Algorithms & Tools:** To be determined
- **Program Linkage:** AIST/SMD
- **Organizational Linkage:** ASDC/SD

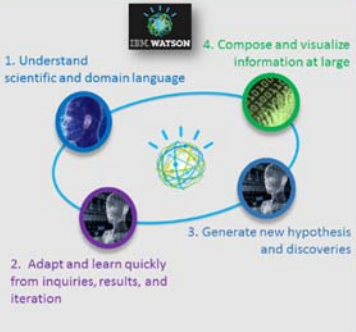


Deep Knowledge Analytics: Knowledge Assistants

Goal: Provide the capability to obtain insights and identify experts and trends quickly by mining knowledge from technical articles/documents, web, and multimedia content
FY16 Outcome: Mine at least four collections; Launch content analytics as robust center-wide capability

Knowledge Corpus Domains		Key Capabilities
<p>FY15</p> <ul style="list-style-type: none"> Carbon Nanotubes Research Autonomous Flight Research Space Radiation Research 		<ul style="list-style-type: none"> Digest and analyze thousands of articles without reading Explore technology gaps that could be leveraged Identify cross-domain leverages and research <p>Tool: Using Watson Content Analytics</p>
<p>FY16</p> <ul style="list-style-type: none"> Space Radiation Research Uncertainty Quantification Human-Machine Teaming Vehicle Design Analysis Entry Descent Landing <p><i>Program Linkages: Incubators; HRP; SASO/ARMD</i> <i>Organizational Linkage: RD; SACD; ED</i></p>		

Deep Knowledge Analytics: Virtual Advisors
Using Watson Discovery Advisor

Cognitive Technologies for Aerospace	
	<p>Goal: Accelerate the discovery of new insights by synthesizing information in seconds, and providing answers with evidence</p> <p>FY16 Outcomes: Develop and Demonstrate two Proof of Concepts - Application to our aerospace domains:</p> <ul style="list-style-type: none"> Aerospace Innovation Advisor Proof of Concept Program Linkage: CAS <i>Example Topics: Hybrid Electric Propulsion; On Demand Mobility</i> Pilot Advisor Proof of Concept Program Linkage: SASO <i>Flight deck expert system for Root Cause Analysis and advise</i> <p><i>ARC, AFRC, and JSC are also investigating use; LaRC is connected with those efforts</i></p>

Partnerships and Education

Partnerships	Education & Outreach
<p>Universities</p> <ul style="list-style-type: none"> GA Tech: Systems Mod Sim & Data Analytics / NIA University of Michigan: Confluence of Mod Sim, HPC, & Big Data University of Washington: Big Data in Aerospace Program MIT: Computer Science and AI Lab / SAA ODU: Machine Learning Carnegie Mellon: Machine Learning <p>Agencies and Industry</p> <ul style="list-style-type: none"> NASA Ames: Data Science & Machine Learning IBM: Cognitive Computing Technologies NASA HQ: Big Data Group; Data Strategy <p>FY16 Outcome: Leverage technology, developments, and expertise for NASA goals</p>	<ul style="list-style-type: none"> Seminars & Workshops MATLAB Courses Lunch & Learn Machine Learning Algorithms and Platforms Websites: <ul style="list-style-type: none"> Big Data Machine Learning Toolbox Knowledge Analytics <p>FY16 Outcome: Build skills for at least 100 SMEs through courses, seminars, and websites</p>



Acknowledgements – Big Data Analytics Team

Data Analytics and Machine Learning Expertise:

Manjula Ambur, Lin Chen, Christina Heinich, Charles Liles, Robert Milletich, Daniel Sammons, Ted Sidehamer, and Jeremy Yagle

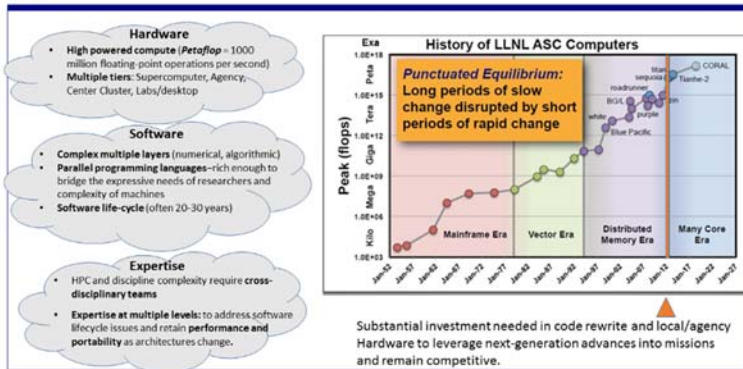
Subject Matter Expertise:

Danette Allen, Damodar Ambur, Dale Arney, Trey Arthur, Randy Bailey, Eric Burke, Jeff Cerro, Kyle Ellis, Christie Funk, Dana Hammond, Angela Harrivel, Jeff Herath, Jon Holbrook, Patty Howell, Lisa Le Vie, Constantine Lukashin, Alan Pope, Brandi Quam, Cheryl Rose, Jamshid Samareh, Mark Sanetrik, Rob Scott, Lisa Scott-Carnell, Steve Scotti, Walt Silva, Mia Siochi, Chad Stephens, Scott Striepe, Marty Waszak, Bill Winfree, and Kristopher Wise

8



High Performance Computing - Essential Elements



State of HPC at LaRC

"HPC is a pacing item for much of science and for many technological developments on the horizon."

Rick Stevens (Associated Laboratory Director) for Argonne National Laboratory before U.S. House of Representatives. May 22, 2013.

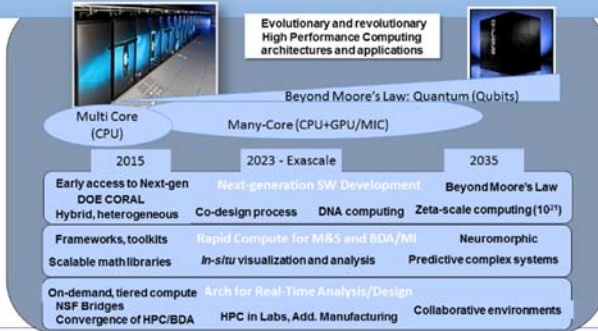
LaRC needs critical mass in HPC hardware, software, and workforce/expertise - [Findings from CDT HPC early deep-divide analysis](#)

- HPC workforce/expertise at all levels**
 - To efficiently leveraging current HPC technologies
 - To prepare for the HPC paradigm shift (that started 5 years ago)
 - Leadership class machines already demand these skills for entry (we are unprepared to partner)
 - Be trained in evolving HPC hardware architectures and programming models
- HPC hardware on all tiers**
 - Our pace of research is being constrained by HPC hardware resources
 - Mission and project requirements have been gradually reduced to meet decreased computational resources.



Vision: HPC Community of Practice

Goal: Enable Rapid Scientific and Systems Level Computing
Enable real-time simulation of complete systems, systems innovation and optimization, accelerate new technology insertion, reduce design margins, decrease risk



CDT HPC Activities in FY 16 Rapid Compute Power for M&S and BDA/ML

Key Activities

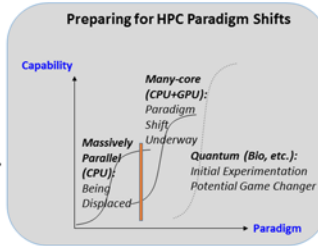
- Evolutionary architectures: Enable M&S and BDA/ML with rapid HPC compute power
- Revolutionary Architectures: Evaluate the applicability of quantum computing to LaRC project

Technology and Capability Advancements

- Prepare for Emerging Technologies (HPC Paradigm shifts)
- Demonstrate rapid compute power as alternate environments for robustness, reliability, and stability of SMART NAS concepts, algorithms, and technologies. Precursor to HPC.

Specific Use Cases:

- SMART NAS – adapting a SMART NAS component to run in the HPC Linux environment. Goal: demonstrate added capabilities.
- Quantum Computing – Early exploratory projects in carbon nanostructures on a quantum annealing platforms. Goal: position LaRC to leverage HPC “Beyond Moore’s Law” for NASA’s unique problems.



8



CDT HPC Activities in FY 16 Architecture for Real-time Analysis and Design

Key Activities:

- Ability to infuse multiple concurrent analyses (and data sets) into an ongoing real-time simulation.

Technology and Capability Advancements:

- Provide computational steering. Namely, embed and control live (or recorded) analyses, and possible projected scenarios, into a real-time analysis
- Ability to integrate multi-fidelity simulations together (including live, delayed, or recorded real data)

Specific Use Case:

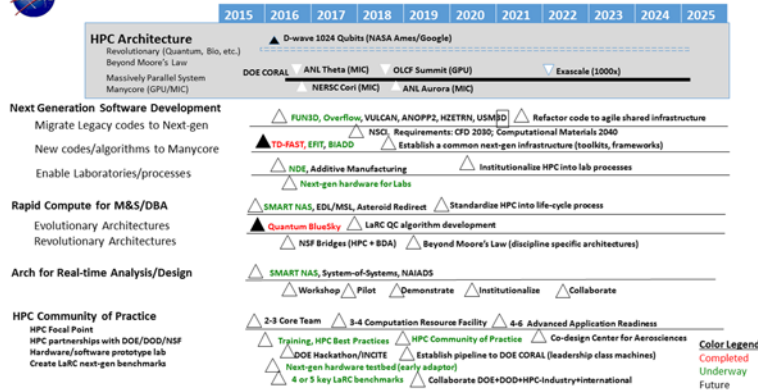
- SMART NAS – Faster than real-time multi-track shadow analysis of robustness, reliability, and stability of SMART NAS concepts, algorithms, and technologies.

SMART NAS – Reduce risk. Extend real-time shadow analysis of robustness, reliability, and stability of NAS concepts, algorithms, and technologies.

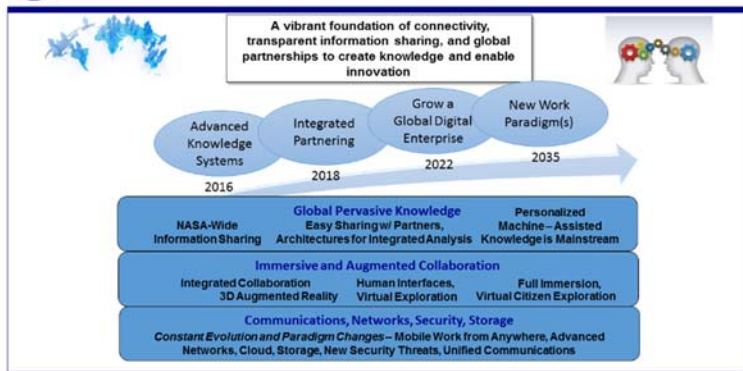




CDT High Performance Computing Vision FY16-FY18



Vision: Advanced Information Technology (IT)



Advanced Information Technology Accomplishments - FY15





Summary Activities for CDT Advanced IT - FY16

Secure Collaboration within and outside NASA

- Secure collaboration with internal and external partners
- Hyperwalls for Multi-center Aeronautics collaboration
- Collaborative Problem Solving and Education – Collaboratory with C. Camarda.
- Software release streamlining
- Contribute to Agency collaboration

Network Optimization and Network Trust

- NASA-wide network trust
- Network optimization w/ ASDC

Integration Architecture for Digital Transformation

Other Areas of Work

- Training, education, seminars, and workshops
- Cloud (OCIO)
- Enhanced knowledge systems (unfunded; pursuing alternatives)

Team: Ed McLarney, Tony Arviola, Jeff Brandt



Secure Collaboration with Internal & External Partners



Multi-Center Aeronautics Collaboration

Use multi-display touch screen powerwalls to collaborate between research centers on Aeronautics missions

- GRC is leading the effort; LaRC is a customer site
- Funded by Aero CAS Augmentation

Specific use cases: Aeronautics, shared with partners

Specific outcomes and benefits:

Implementation of 4 to 12 fully interconnected powerwalls located at GRC, AFRC, ARC, and Langley supporting immersive collaboration environment





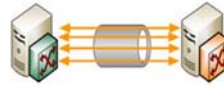
Software Release Process Streamlining

- Research and optimize methods to streamline software release among NASA centers and other trusted partners
 - Conduct SWOT analysis; benchmark similar organizations; identify existing choke points
- Specific outcomes and benefits:**
 Deliver improved software release processes that meet legal and policy guidelines



Network Trust and Optimization

- Establish Improved trust across firewalls among all willing NASA centers
 - Technical approach, cultural approach, advocate to HQ & other centers; work with willing centers
 - Increase center's effective data throughput to external sources and customers
- Specific outcomes and benefits:**
 Establish baseline of trusted communication ports that will allow ~80% of collaboration communications to bypass center perimeter firewalls
 Decrease transfer time of huge files by a factor of 10x (ASDC).
 Advocate for Agency-funded wide area network enhancements up to 4x current connection bandwidth



Advanced IT Architecture

- Develop a plan for growing an advanced IT architecture to support integration of M&S, Big Data Analytics, Machine Learning, HPC, and Advanced IT
 - Focus on integrating via interfaces and standards; consider enterprise service bus; leverage external best practices
 - Choose mission-enabling use cases w/ other CDT leads and customers
- Specific outcomes and benefits:**
 Deliver a service architecture that ensures that Advanced IT is supporting the capabilities required by the other CDT focus areas

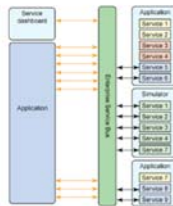



Figure 26 CDT Overview (February 2016)

Knowledge Analytics and Data Analytics (August 2016)

Knowledge Analytics Using Watson Technologies

Manjula Ambur, Ted Sidehamer and Jeremy Yagle
Big Data Analytics and Machine Intelligence Team
Comprehensive Digital Transformation Initiative
NASA Langley Research Center
Aug 2016



NASA Langley Comprehensive Digital Transformation

Vision: Catalyst to Enable Transformative Solutions to NASA Mission Challenges

Goals:

- Accelerated Scientific Innovations and Discoveries
- Focused, Relevant Research and Technology Development
- Intelligent and Rapid Engineering and System designs
- Virtual Analysis, Design, and Verification of Aerospace Systems and Science Instruments

Core Capabilities - Emphasis Areas


<p style="text-align: center; color: orange;">Modeling & Simulation (M&S)</p> <ul style="list-style-type: none"> • Integrated analysis and design of complex systems • Facilitate improved physics-based discipline tools • Optimally combine testing and M&S 	<p style="text-align: center; color: orange;">Big Data Analytics & Machine Intelligence (BDA&MI)</p> <ul style="list-style-type: none"> • Rapid synthesis of global scientific info. for new insights • Data intensive scientific discoveries for advanced designs • Virtual Experts: Human-machine symbiosis
<p style="text-align: center; color: orange;">High Performance Computing (HPC)</p> <ul style="list-style-type: none"> • Next generation software development • Rapid Compute power for M&S and BDA&MI • Architecture for real-time analysis and design 	<p style="text-align: center; color: orange;">Advanced Information Technology</p> <ul style="list-style-type: none"> • Open, secure collaboration for synergy • Networks handle burgeoning data • Data governance, architecture, and management

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

Big Data Analytics and Machine Intelligence


Vision: Virtual Research & Design Partner

Human Experts and Digital Experts augmenting and learning from each other in an organic way



Foundational Components

Data Mining	Machine learning Semantics	Knowledge Technologies In-Situ Analysis Algorithms	Global Sources Databases	Data Visualization Ontologies Content Analytics	Distributed data stores Journals Blogs
Web sites	Images	Knowledge Base	Global Sources	Distributed data stores	Journals Blogs
Open web	Documents	Data sets	Videos	Databases	Journals Blogs
Digital Notebook	Data Integration	Data Management & Collaboration	Data Governance	Data Capture	Meta Data Remote Access
	Data Curation	Analytics Sent to Data		Tagging	



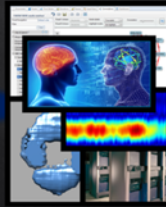
Two Key Areas for Virtual Partner

Data Intensive Scientific Discovery (DISD)

Deriving new insights, correlations, and discoveries not otherwise possible from our diverse experimental and computational data sets – ***The Fourth Paradigm***

Deep Knowledge Analytics (DKA)

Obtaining insights, identifying trends, aiding in discovery, and finding answers to specific questions by mining knowledge from scholarly, web, and multimedia content – ***Cognitive Computing***



What is Watson Content Analytics?

Description:

Provides the ability to search and analyze large volumes of unstructured information from multiple sources, to quickly understand and deliver relevant insight.

Value:

- Uncovers the meaning and context of human language within unstructured information
- Enables interactive visualization of data to reveal trends, patterns and correlations
- Provides customizable natural language processing technology to extract facts, concepts and relationships from information

User: Research Analysts, Data Scientists, Analysts/Investigators



The Power of Watson Content Analytics: Knowledge Assistants Pilots

Key Capabilities

- Digest and analyze thousands of articles without reading
- Provide a means to rapidly identify trends
- Identify connections among experts at all levels and affiliations
- Use the power of taxonomies to enable deep analytics
- Explore technology gaps that could be leveraged
- Replace traditional methods of SMEs manually reviewing and tracking research
- Help to identify cross-domain research

Examples

Autonomous Flight



Vehicle Design



Carbon Nanotubes

Space Radiation



Knowledge Engineering Overview Machine & Expert Generated Facets

Machine Generated Facets/NLP

SME Defined Facets

Documents

The Power of Analytics comes from natural language processing and a combination of machine and expert generated facets enabling deep analysis of content to provide insights and connections



Knowledge Assistant: Vehicle Analysis

Goals

- Develop vehicle analysis knowledge base that could be leveraged by SACD with the help of analytics
- Rapidly and efficiently identify experts, insights and connections

Facet Pairs - Joining Machine Generated Concept to Authors Looking for Experts and Connections



Corpus

- 450 documents from SME collections and added metadata for about 100.
- SME's are working to expand corpus

Accomplishments and Next Steps

- Analytics: NLP; partial metadata; SME provided taxonomy
- Demonstrated to SMEs: sees the value and potential
- Adding new documents/metadata, refining dictionaries, taxonomies and annotators where possible.

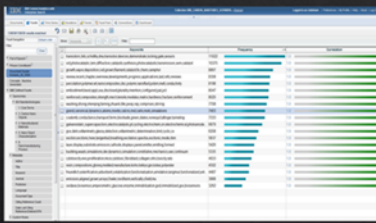


Knowledge Assistant: Carbon Nanotubes

Goals

- Automatically identify a subset of documents from large corpus and provide 'gists'
- Rapidly and efficiently identify experts, trends, and insights, connections and technology gaps

Automated Document Clustering - Content Analysis Output



Approach

- Using Watson Content Analytics (WCA) software
- 130,000 articles metadata were used
- Use of natural language processing, statistical algorithms and semantic techniques

Accomplishments and Next Steps

- Successfully demonstrated to SMEs and stakeholders
- Center-wide capability to mine individual researchers collections, and open literature (NASA, patents)



Knowledge Assistant: Autonomous Flight

Goal

- Use the WCA software to integrate analysis of scholarly and informal web content to identify new partnerships

Approach

- Uses 4000 articles selected from IEEE and select web sites
- The machine-generated and expert-generated taxonomies facets enable the deep analytics of the content

Accomplishments and Next Steps

- Successfully demonstrated to SMEs and stakeholders
- Our Production System will be using web crawlers to mine text data across a wide range of websites



Example of Author and Topics Relationships View



Knowledge Assistant: Space Radiation

Goals

- Identify leverage points and possible duplication
- Rapidly and efficiently identify experts, trends, and insights, connections and technology gaps

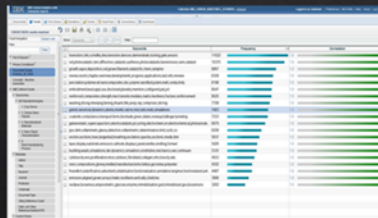
Approach

- Using Watson Content Analytics (WCA) software
- 1000 articles and metadata of all the research was used

Accomplishments and Next Steps

- Demonstrated to SMEs: sees the value and potential
- Planned demonstration to Human Research Program on July 16th to show the potential for possible funding
- Pursuing Watson Space Radiation Discovery Advisor

Automated Document Clustering - Content Analysis Output



Methodology

- TEAM APPROACH
 - Analytics Experts (Manjula Ambur, Ted Sidehamer, Jeremy Yagke)
 - Digital Librarians (Dorothy Notarnicola)
 - IBM content analytics experts
 - NASA SMEs (Jeff Cerro, Dale Arney, Lisa Scott-Carnell, Mia Stochi, Christopher Witt, Danette Allen)
- FOCUS ON CONTENT & KNOWLEDGE ENGINEERING
 - Having the rich and comprehensive corpus is
 - Close collaboration with SMEs is critical – taxonomies, ontologies...
 - Understand all possible sources and formats for data capture and ingest
 - Use researcher's collections, open-source publications, and internal content
- ITERATIVE PROCESS
 - Understand SME goals
 - Build Facets and Content Rules and Annotators with WCA Studio
 - Regular meetings throughout the development process

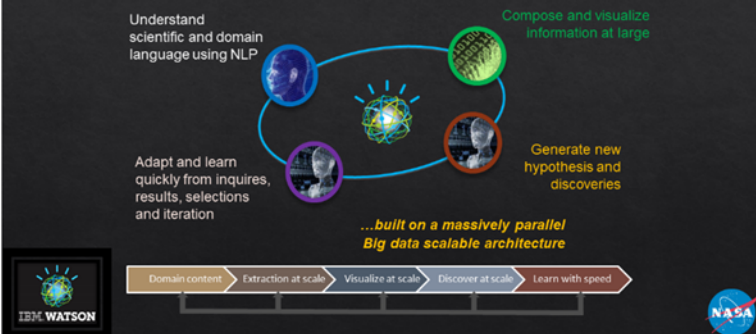


Demonstrated Value to Researchers and Engineers

- ◆ Saves months of reading
- ◆ Identifies new collaborators
- ◆ Replaces manual methods of analyzing technical publications
- ◆ Provides insight into research overlap
- ◆ Highlights technology gaps
- ◆ Advanced analytics for new insights



IBM Watson combines transformational technologies



Moving Forward



Center-wide Capability for Knowledge Analytics

Focused on using researcher's collections, internal content, and open-source resources

Proof of Concept: Aerospace Innovation Advisor and Pilot Advisor

Proof of Concepts using the full capability of IBM Watson technologies



Aerospace Innovation Advisor and Pilot Advisor

Based on Watson Discovery Advisor

Accelerate the discovery of new insights by synthesizing information in seconds

- Take advantage of massive sources of data
- Find answers to questions that have not been asked yet or answered before
- Find insights into hidden relationships and dig deeper
- Generate leads to hard questions and provide evidence to substantiate new claims





Funding obtained for Aerospace Innovation Advisor Proof of Concept with NASA-published research to demonstrate the value to our mission challenges

Knowledge Assistants

Using Watson Content Analytics

Analyze and digest large volume of scientific information rapidly without 'reading' to

- Gain key insights and patterns
- Identify trends and connections
- Visualize experts networks

Pilots Examples:

- Carbon Nanotubes
- Autonomous Flight
- Space Radiation
- Vehicle Design

Aerospace Innovation Advisor and Pilot Advisor POC

Using Watson Discovery Advisor

Apply cognitive computing technologies that 'understand' massive amounts of information and enhance experts' abilities by

- Answer complex questions in seconds with confidence levels and traceability of evidence
- Show unobvious relationships within disciplines
- Show best possible paths for moving research forward and solutions for complex problems
- Enable rapid ideation and innovation

Example Topics for POC :

- Hybrid Electric Propulsion

For More Information, Please Visit our Websites:

Big Data Analytics and Machine Intelligence Capability
<http://bigdata.larc.nasa.gov>
 Provides vision, strategy and roadmap, and a high-level overview of current use cases in both Data Intensive Scientific Discovery (DISD) area that focuses on mining mission data sets for discoveries, and Knowledge Analytics (KA) area that focuses on extracting insights from technical literature and documents

Machine Learning Toolbox
<http://machinelearning.larc.nasa.gov>
 A guide for those who are interested in learning more about various machine learning techniques for mining experimental and computational data sets for DISD, and details of our use cases.

Knowledge Analytics Utilizing Watson Content Analytics (WCA)
<http://knowledgeanalytics.larc.nasa.gov>
 Introduces the capabilities of WCA for synthesis and extraction of trends and insights rapidly and using as Knowledge Assistant for a specific technical area. Also offers users to explore current use cases or request their own.



Figure 27 Knowledge Analytics and Data Assistants: August 2016

Cognitive Computing Vision and Watson Applications at NASA (April 2016)



IBM

Cognitive Computing Vision and Watson Applications at NASA



Manjula Ambur, Associate CIO for Big Data Analytics and Machine Intelligence,
NASA Langley Research Center
April 2016



© 2016 IBM Corporation



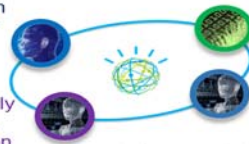
what is Cognitive Computing

- Cognitive-based systems are able to build knowledge and learn, through understanding of natural language, to reason and interact more naturally with human beings than traditional systems.
- Cognitive systems continue to evolve as they ingest new information, new scenarios, and new responses. They reason in a way that is similar to human thinking so conclusions are obvious, transparent, and useful
- Cognitive systems amplify human cognition – Power of Human Machine symbiosis
 - Experts train system – takes time to teach new domains
 - Experts and system work together doing what they do best
 - Democratization of knowledge, expertise, and innovation
- Watson is a cognitive computing system



How Does Cognitive Computing work?

1. Understand scientific and domain language using NLP
2. Adapt and learn quickly from inquires, results, selections and iteration
3. Generate new hypothesis, discoveries, and answers
4. Compose and visualize discoveries and answers



...built on a massively parallel
Big data scalable architecture

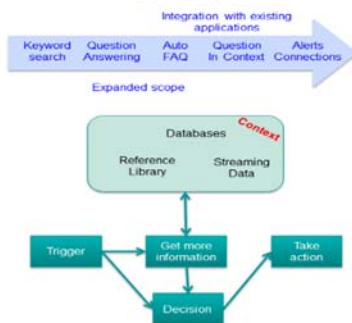




Cognitive Computing Vision for NASA

- ‘Digital Advisors/Experts’ enabling greater aerospace scientific discoveries, innovative systems designs and complex operations.
 - Able to quickly digest latest research innovations by synthesizing large volume of information rapidly showing unobvious trends, and paths
 - Analyze experimental, modeling and simulation and flight data in real time helping with system configurations and design predictions/optimizations
 - Answer specific engineering questions in all aerospace disciplines by integrating data and knowledge and showing evidence and traceability
 - Deep analysis and mining of multimedia scientific and engineering information with associated data ,images and videos, that comprehends numbers and mathematical equations resulting in actions, advise and answers that augment, augment and replace human experts.

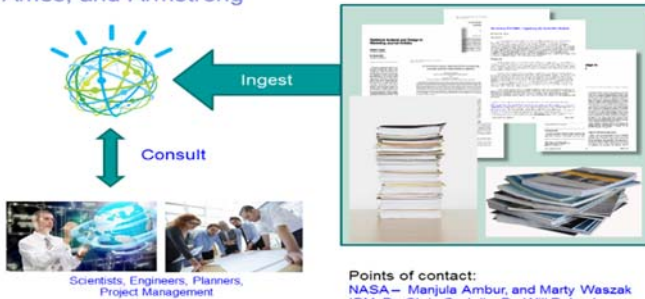
Watson Flight Operations Advisor – Ames, Armstrong



Points of contact:
 NASA – Dr. Richard Mogford, Dr. Jack Ryan
 IBM: Dr. Chris Codella, Dr. Will Dubyak

© 2010 IBM Corporation

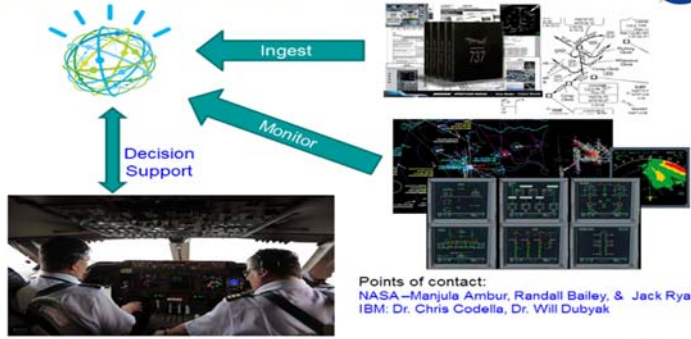
Watson Aerospace Innovation Advisor – Langley, Glenn, Ames, and Armstrong



Points of contact:
 NASA – Manjula Ambur, and Marty Waszak
 IBM: Dr. Chris Codella, Dr. Will Dubyak

© 2010 IBM Corporation

Watson Pilot Advisor – Langley, Armstrong



Points of contact:
NASA – Manjula Ambur, Randall Bailey, & Jack Ryan
IBM: Dr. Chris Codella, Dr. Will Dubyak

© 2010 IBM Corporation

Watson Astronaut Health and Robotics - JSC



© 2010 IBM Corporation



Cognitive Computing for NASA – A few thoughts

- As a research and design advisor using Watson Discovery Advisor – domain adoption and training is key
- As a Pilot Advisor to help with best possible solution paths in distress situations – can have speech interface like Siri
- Could be a co-pilot as cognitive computing technology adoption to our domain matures and we gain confidence
- More autonomous agents/robots/rovers for long Space travel and on other planets that can help us to do more things quickly and creatively
- As Advisors and/or companions for long duration space travel to Mars and to 'live' on Mars
- The Digital Advisor/Expert may not be possible immediately. Breakthroughs in understanding mathematics and tables, and mimicking human cognition, intuition are still required.



Comments and Next Steps

- Proof of Concepts underway will provide us an excellent understanding of cognitive computing application to our domains
- Proof of concepts will help us to demonstrate functionality, potential and gaps & limitations to NASA experts, leaders & stakeholders
- Cognitive systems amplify human/expert cognition – Power of Human Machine symbiosis in Action – We will learn how this works!!
 - Experts train system – takes time to teach new domains
 - Experts and system work together doing what they do best
- Cognitive capability plans, experience and feedback will help us to formulate Watson and Cognitive computing plans for next few years

Figure 28 Cognitive Computing Vision and Watson Applications at NASA (April 2016)

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

The public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing the burden, to Department of Defense, Washington Headquarters Services, Directorate for Information Operations and Reports (0704-0188), 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.
PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.

1. REPORT DATE (DD-MM-YYYY) 01-12-2016		2. REPORT TYPE Technical Memorandum		3. DATES COVERED (From - To)	
4. TITLE AND SUBTITLE Big Data Analytics and Machine Intelligence Capability Development at NASA Langley Research Center: Strategy, Roadmap, and Progress				5a. CONTRACT NUMBER	
				5b. GRANT NUMBER	
				5c. PROGRAM ELEMENT NUMBER	
6. AUTHOR(S) Ambur, Manjula Y.; Yagle, Jeremy J.; Reith, William; McLarney, Edward L.				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-20775	
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-2016-219361	
12. DISTRIBUTION/AVAILABILITY STATEMENT Unclassified - Unlimited Subject Category 82 Availability: NASA STI Program (757) 864-9658					
13. SUPPLEMENTARY NOTES					
14. ABSTRACT In 2014, a team of researchers, engineers and information technology specialists at NASA Langley Research Center developed a Big Data Analytics and Machine Intelligence Strategy and Roadmap as part of Langley's Comprehensive Digital Transformation Initiative, with the goal of identifying the goals, objectives, initiatives, and recommendations need to develop near-, mid- and long-term capabilities for data analytics and machine intelligence in aerospace domains. Since that time, significant progress has been made in developing pilots and projects in several research, engineering, and scientific domains by following the original strategy of collaboration between mission support organizations, mission organizations, and external partners from universities and industry. This report summarizes the work to date in Data Intensive Scientific Discovery, Deep Content Analytics, and Deep Q&A projects, as well as the progress made in collaboration, outreach, and education. Recommendations for continuing this success into future phases of the initiative are also made.					
15. SUBJECT TERMS Artificial intelligence; Big data; Cognitive computing; Data; Data analysis; Data science; Machine intelligence; Machine learning					
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	121	19b. TELEPHONE NUMBER (Include area code) (757) 864-9658