DATA SCIENCE: STATISTICAL AND COMPUTATIONAL METHODOLOGY FOR NASA'S BIG DATA IN SCIENCE

A BDTF white paper with accompanying findings and recommendations

Ad-Hoc Task Force on Big Data

November 3, 2017

A BDFT White Paper

Data Science: Statistical and Computational Methodologies for NASA's Big Data in Science

An Aphorism

The scientist collects data in order to understand natural phenomena

The statistician helps the scientist acquire understanding from the data

*The computer scientist helps the scientist perform the calculations**

The individual proficient at all these tasks is a data scientist

** needed only for Big Data*

I. Executive Summary

Big Data and High Performance Computing deeply permeates NASA's Earth and space science endeavor. However, NASA science is not institutionally well-integrated with the large community of scholars and experts who develop the methodologies underlying modern Data Science: computer scientists, information technologists, applied mathematicians and statisticians. Technology transfer from and collaboration with these scholarly communities is insufficient to meet NASA's needs, both within NASA centers and in the wider U.S. space science research community.

This White Paper details some of these issues and recommends actions to improve methodologies used for current and future missions across all science domains. Examples include: compressed sensing for image and signal processing; Deep Learning neural networks for pattern recognition and classification; Virtual Reality visual exploration tools; and agile software development and other strategies from computer engineering.

Recommendations:

- NASA SMD should organize and fund professional development in statistics and informatics, both for its internal scientists and for the wider Earth and space science communities. This includes organizing training workshops, producing on-line training materials, attending methodology conferences, and hiring expert consultants.
- Specifications and performance reviews for mission science operations software development should include high standards for computational algorithms and statistical methods with evaluation by cross-disciplinary experts.
- NASA SMD should ensure modern software engineering (for example: agile, fast iteration processes for creation and modification of software systems) is applied to its sponsored development and maintenance projects. This may include active involvement of data science professionals as staff or consultants.

II. Introduction

NASA's culture and administrative mechanisms take full cognizance that major advances in space science are driven by improvements in instrumentation. But it is less well recognized that new instruments give rise to science questions so diverse and complex that traditional data analysis procedures are often inadequate. The knowledge and skills of the statistician, applied mathematician, and algorithmic computer scientist need to be incorporated into programs that currently emphasize engineering and physical science in order to fully achieve the scientific potential of satellite missions. These issues might be divided into two stages: data reduction through software pipelines developed within NASA mission centers; and science analysis that is performed by hundreds of space scientists dispersed through NASA, U.S. universities, and abroad. Both stages benefit from the latest statistical and computational methods; in some cases, the science result is completely inaccessible without modern methodology.

NASA-sponsored science is participating in a broad transformation towards the *Fourth Paradigm*[1] where scientific insights are extracted from peta-scale datasets using advanced statistical and computational methods. An example of this transition is occurring in astronomy where computations from cosmic microwave background and exoplanet detection satellites require large supercomputer allocations. The scientific interpretations involve diverse methodology such as supervised classification, spatial point processes, time series analysis, and Bayesian inference. In heliophysics, sophisticated image processing techniques are needed to identify classes of features from multi-petabyte solar datasets using statistical machine learning. New subfields like astrostatistics and astroinformatics are emerging with journals[2], scholarly societies[3], and conferences[4]. A critical problem is the education of space scientists in these new methodologies; recent STScI[5] and National Academy of Sciences[6] reports emphasize the need for professional training beyond long-standing formal education in the physical sciences.

The emerging term for such cross-disciplinary activities -- *Data Science* -- can be represented as the intersection of three skill areas: domain expertise (like space science), mathematics and statistics, and computational skills. The aphorism above encapsulates the relationship between these three domains. Data science is a rapidly emerging area of study and practice that has witnessed amazing acceleration in recent years with foundations in statistics and computer science over several decades[7]. However, NASA science is not a leader in this movement. Substantial

---

[1] *The Fourth Paradigm: Data-Intensive Scientific Discovery* (2009) T. Hey, S. Tansley & K. Tolle, Microsoft Research

[2] *Astronomy & Computing* (Elsevier, started 2013), AAS Journals Policy statement on software (2016)

[3] International Astrostatistics Association (founded 2012), American Astronomical Society Working Group in Astroinformatics and Astrostatistics (2014), American Statistical Association Astrostatistics Interest Group (2014), IEEE Astrominer Task Force (2014), International Astronomical Union Commissions on Computational Astrophysics and Astroinformatics & Astrostatistics (2015)

[4] Astrostatistics and Astroinformatics Portal – Meetings, https://asaip.psu.edu/meetings

[5] *Big Data @ STScI: Enhancing STScI's Astronomical Data Science Capabilities over the Next Five Years* (2016), http://archive.stsci.edu/reports/BigDataSDTReport_Final.pdf

[6] *Optimizing the U.S. Ground-Based Optical and Infrared Astronomy System* (2015), National Academy Press, https://www.nap.edu/login.php?record_id=21722

[7] While dozens of texts are available in data science, we highlight three well-respected and widely-read volumes that emphasize the mathematical and algorithmic foundations: C. M. Bishop, *Pattern Recognition and Machine Learning* (2007); T. Hastie, R. Tibshirani, J. Friedman, *The Elements of Statistical Learning: Data Mining,*

benefits can accrue with new analysis procedures at a fraction of the cost of the total satellite missions, and yet these techniques are not universally employed across every mission. Promising areas of methodological improvements for space science include:

- Statistical procedures including: False Discovery Rate for multiple hypothesis testing; Uncertainty Quantification for computationally expensive modeling; model selection and residual analysis following regression fits
- Machine learning, data mining and computational intelligence including: Deep Learning neural nets, pattern recognition, decision trees and Random Forests
- Data cleaning and curation including: data aggregation, normalization, synchronization, assimilation, and improved meta-data descriptions of datasets
- Data visualization to improve understanding including: visual exploration tools, sonification, haptic data sensing techniques, and virtual reality.

Some of the discussions in this white paper employ a classic NASA mission model where the mission's science team is responsible for processing of data products derived from the data gathered by the operational platform (satellite, aircraft, etc.) This processing often employs algorithm's employing physics models and mathematical techniques. In Earth science, observational data and higher-level derived products are managed and made available to the research communities via the Earth Observing System Data and Information System (EOSDIS - https://earthdata.nasa.gov/about.) "The EOSDIS science operations are performed within a distributed system of many interconnected nodes, the Science Investigator-led Processing Systems (SIPS), and distributed via discipline-specific Distributed Active Archive Centers (DAACs) with specific responsibilities for production, archiving, and [dissemination] of Earth science data products. The DAACs serve a large and diverse user community (…) by providing capabilities to search and access science data products and specialized services." The conclusions reached and recommendations made in this white paper are applicable not only to the mission model of science data processing but also to the extended model employed in the Earth sciences.

III. Statement of the Problem

The enormous strides in methodology from statistics, applied mathematics and computer science of recent decades are often not incorporated into NASA satellite data or science analysis programs. High standards for analysis methodology and algorithms are not set consistently for mission analysis pipelines within NASA mission centers, for science analysis software maintained by NASA archive centers, or for extramural science programs funded by NASA. The result is uneven quality in data and science analysis products of NASA science missions; the methods used in NASA software systems for data processing and science analysis are often obsolete.

The underlying cause of the problem involves human resources and academic culture. Space science has weak mechanisms for incorporating progress from the large scholarly community in statistics, applied mathematics, and computational algorithms. Cross-disciplinary training and collaborative cross-disciplinary research efforts are often inadequate. Space scientists, both within

Inference, and Prediction, 2nd ed. (2016); I. Goodfellow, Y, Bengio, A. Courville, *Deep Learning* (2016).

NASA and external universities and institutes, are often weakly trained in modern statistical and computational methodology.

The volume, variety and velocity of NASA science data is taxing established methods and technologies:

- The volume has become particularly acute in the Earth Sciences and Heliophysics divisions; the Astrophysics (with JWST and WFIRST) and Planetary divisions are soon entering the realm of Big Data.
- Planetary missions have instruments measuring very different aspects of their targets, and astrophysics has a bewildering variety of science questions addressed by NASA observatories.
- Computationally intensive simulations supporting these missions often generate even larger data volumes. This introduces new concerns such as optimal design of computer experiments and model validation against observation.
- Scientific progress can be impeded by architectural limitations on data movement, accessibility and data fusion, although progress is evident in systems like the NASA Astronomical Virtual Observatory.
- Access to and analysis of large databases sometimes lack of state-of-the-art statistical methodologies and computational algorithms designed for Big Data
- The state of documentation of data sets is inconsistent as are procedures for the development, testing, improvement, approval, documentation and promulgation of algorithms. These impede both current and future extraction of scientific results.

IV. Examples of approaches

Two promising new methods for NASA SMD science

Compressed sensing is a powerful lossy signal restoration method, related to wavelet and Fourier transforms, where a sparse representation is obtained that gives a near-optimal and highly efficient extraction of information from an image or signal[8]. Compressed sensing is providing enormous computational advantages in fields such as medical imaging and computer vision, but has only begun to be applied to problems in space science. Spatio-temporal movies of solar flare plasma structures from multiwavelength videos obtained by the Solar Dynamics Observatory shows the potential of this mathematical approach[9]. Compressed sensing also has high potential for improved planetary mapping with synthetic radar[10], weak lensing surveys for cosmology[11], and many other imaging applications.

---

[8] Candès, Emmanuel J.; Romberg, Justin K.; Tao, Terence, Stable signal recovery from incomplete and inaccurate measurements, *Comm. Pure Appl. Math.* 59 (8): 1207–1223, 2006; D. Donoho, Compressed Sensing, IEEE Trans. Information Theory, 52(4), 1289-1306, 2006; D. Donoho, U.S. Patent 8855431, 2014. Donoho and Candès have both received MacArthur Foundation 'genius' awards.

[9] Cheung, Mark C. M.; Boerner, P.; Schrijver, C. J.; Testa, P.; Chen, F.; Peter, H.; Malanushenko, A., Thermal Diagnostics with the Atmospheric Imaging Assembly on board the Solar Dynamics Observatory: A Validated Method for Differential Emission Measure Inversions, 2015, Astrophys. J. 807, #143

[10] M. A. Herman, Strohmer, T. 2009, High-resolution radar via compressed sensing, IEEE Trans. Signal Proc. 57(6), 2275-2284

[11] A. Leonard, F.-X. Dupe, J.-L. Starck, A compressed sensing approach to weak lensing, 2012, Astro. Astrophys., 539, #A85

Deep Learning, a type of multi-layer convolutional neural network, has recently achieved breakthroughs in the extraction of patterns and classification of very large collections of images, speech and other signals[12]. It is a computationally intensive machine learning technique that is mostly unsupervised, where the procedure learns to extract meaningful features from large and complex datasets. It has solved previously intractable problems such as accurate automated natural language translation from English to Japanese, labeling faces in images, playing the game Go, and improving efficiency in factory energy usage. Viewed as a critical advance in artificial intelligence, the potential of deep learning for space science is mostly unknown. Early results include mapping the proportion of ices and dust on the Martian poles from remote hyperspectral imaging[13], inference of electron densities from plasma wave measurements from the Van Allen Probes mission[14], and evaluating the evolutionary states of red giant stars using asteroseismological time series from the Kepler mission[15].

Upskilling and Education

Space scientists are typically well-trained in mathematics relating to physical processes, but not in mathematics relating to extraction of reliable information from complex noisy datasets. They are typically conversant with computer programming and processing on a moderate scale, but many are not prepared for the world of Big Data with challenges in data storage, access, and efficient analysis on high performance multicore computers.

The problem arises in the curriculum of physical scientists: courses in modern statistics, applied mathematics, and computer science are not in the required curriculum. This deficiency has been recently documented among astronomers: a survey of ~1100 scientists found that 90% write software but only 8% received substantial training in software development[16]. Informal on-the-job training is adequate for some purposes, but can lead to mediocrity, or even unnecessary failure, for more challenging problems with Big Data. In statistics, the methodology is so diverse that coursework and usage of statistical software environments is essential for many scientific projects.

Some progress in education has been made. Textbooks on statistical methodology and data analysis (with computer codes) for remote sensing, atmospheric science and astrophysics are available and taught in some universities[17]. More successful have been the informal and short-form trainings, including: Summer Schools, Hack Days, and bootcamps. Programs like the

[12] Hinton, G. et al. Deep neural networks for acoustic modeling in speech recognition. IEEE Signal Proc. Mag. 29, 82–97, 2012; Y. LeCun, Y. Bengio, G. Hinton, Deep learning, Nature, 521, 436-444, 2015

[13] A. Deleforge, F. Forbes, S. Ba, R. Horaud, 2015 Hyper-spectral image analysis with partially latent regression and spatial Markov dependencies, IEEE J. Selected Topics Signal Processing, 9(6), 1037-1048

[14] I. S Zhelavskaya, M. Spasojevic, Y. Y. Shprits, W. S. Kurth, 2016, Automated determination of electron density from electric field measurements on the Van Allen Probes spacecraft, J. Geophys. Res. Space Physics, 121, 4611-4625.

[15] Hon, M., Stello, D., Yu, J., Deep learning classification in asteroseismology, 2017, Mon. Not. Royal Astro. Soc., in press, arxiv/1705.06405

[16] I. Momcheva & E. Tollerud, Software Use in Astronomy: An Informal Survey (2015) https://arxiv.org/abs/1507.03989

[17] Texts include D. S. Wilks, *Statistical Methods in the Atmospheric* Sciences, 3rd ed. (2011); E. D. Feigelson & G. J. Babu, *Modern Statistical Methods for Astronomy with R Applications* (2012), Z. Ivezic, A. J. Connolly, J. T. VanderPlas, A. Gray, *Statistics, Data Mining, and Machine Learning in Astronomy: A Practical Guide for the Analysis of Survey* Data (2014); M. J. Canty, *Image Analysis, Classification and Change Detection in Remote Sensing with Algorithms in ENVI/IDL and Python*, 3rd ed. (2014); C. Bailer-Jones *Practical Bayesian Inference for Physical Scientists* (2017). For a complete list in astrostatistics, see https://asaip.psu.edu/resources/recent-books/methodology-books-for-astronomy

Software and Data Carpentries[18] and programming workshops like AstroHackWeek[19], are positive examples of how the community can upskill itself. But, except for these instances and perhaps a disorganized Facebook-based discussion forum[20], these resources are touching relatively few space scientists. University curricula are not renovating fast enough to match the needs of methodological education for future space scientists, and professional development resources are insufficiently funded or organized to meet the needs of current space scientists.

The role of scholarly societies

The broader research communities have recognized the importance of these issues with burgeoning activities. The AGU Fall meetings with >20,000 Earth and space scientists has dozens of sessions on informatics: exploiting Big Data, improving data infrastructure and access, coordinating data frameworks, facing supercomputing challenges, and developing methodology for data analytics and Deep Learning to accelerate scientific discovery[21]. A once-in-a-century reorganization of the IAU in 2015 led to the addition of 'Data Science' to the name of a Division and the initiation of new Commissions on Computational Astrophysics and on Astroinformatics and Astrostatistics[22]. Cross-disciplinary working groups for advanced methodology in astronomy also arose in the AAS, ASA, and the IEEE engineering society in the past few years. In 2016, the AAS Journals that publish 1/3 of the world's research literature created a new Statistical Scientific Editor position to improve the methodology in its articles[23].

V. Recommended approaches

Training workshops

Several organizations, particularly in astrophysics, have organized informal training workshops in statistics and informatics. These training workshops receive little or no funding from NASA.

- Since 2001, the European *Astronomical Data Analysis* conference series have been accompanied by hands-on tutorial sessions on advanced data processing[24].
- Since 2005, Penn State University has organized a week-long *Summer School in Statistics for Astronomers,* partially supported by the National Science Foundation [25]. Similar tutorials have been offered upon request at ~20 research institutions worldwide.
- Since 2014, ESA has held week-long *Data Analysis and Statistics Workshops* at the ESAC facility in Madrid ES[26].
- In recent years, American Astronomical Society meetings are preceded by a selection of professional development workshops that include training in computer programming (Python, R, software carpentry) and statistical methodology[27].

---

[18] https://software-carpentry.org/about/ & http://www.datacarpentry.org/about/
[19] https://arxiv.org/pdf/1711.00028.pdf
[20] Facebook groups: https://www.facebook.com/groups/remotesensing/ (Remote sensing and GIS, ~15,000 members); https://www.facebook.com/groups/astro.r/ (astrostatistics, ~3000 members)
[21] https://fallmeeting.agu.org/2016/scientific-program-guidelines/
[22] https://www.iau.org/science/scientific_bodies/divisions/B/info/commissions/
[23] http://journals.aas.org/editorial.html
[24] http://ada8.cosmostat.org
[25] http://astrostatistics.psu.edu/su17/
[26] https://www.cosmos.esa.int/web/esac-science-faculty/esac-faculty-events
[27] https://files.aas.org/aas227/AAS_227_Block_Schedule.pdf (January 2016 AAS meeting workshops)

- Seven Solar Information Processing Workshops, combining solar astronomers with statisticians and computer scientists, were held prior to 2014 with partial NASA support.

Communities of Practice

The National Science Foundation has funded a system of Big Data Regional Innovation Hubs (BDHubs), networks of organizations working on technologies and applications around Big Data, data science, and science applications. One regional consortium fosters workforce development, another builds public-private multi-sector partnerships, and others focus on societal challenges addressed through Big Data. Each Hub collectively responds to "Spoke" funding programs that provide the main research funding for the BDHubs effort. Spokes are organized thematically, for instance in urban science, transportation, agriculture, and medicine, and other efforts of interest to society and to NSF directorates. The system provides a scaffold for Big Data innovation, promoting more impactful proposals that leverage Big Data technologies "to support basic research and people to create knowledge that transforms the future". Key to this effort is participation of industry and nonprofit partners, though the leads for all hubs are university-based faculty.

An excellent example of a similar effort focused on building connections within a campus is the Moore-Sloan Data Science Environments (DSEs), a partnership of three-universities and two private foundations created in 2013 to address the lack of academic support for research oriented data scientists at universities throughout the US[28]. The Gordon and Betty Moore Foundation and Alfred P. Sloan Foundation fund it at a level of $37.8 million over five years. The DSE premise is that changes in academia are required to more fully support data scientists who enable new data-driven discoveries. Data scientists are in high demand in industry but lack similar support in academia due to their inherently interdisciplinary nature. DSEs support the academic careers of data scientists, promote education and training in data-intensive discovery, build an ecosystem of tools and software, and support reproducibility and open science. These themes are further augmented by additional working groups around topics such as Cross-Domain Image Analysis and Text as Data[29].

Software engineering

Software engineering is a major field that specializes in developing methods for software architecture and design, parallel and distributed computing, performance optimization, database access, debugging and testing, maintenance and reuse. For three decades, the IEEE society has been running international conferences on software engineering, architecture, reliability, and comprehension[30]. Some of the issues that can benefit NASA programming efforts include:

- Software design patterns emphasize broad principles in strategy, structure and implementation of major software development efforts. Software design can be based on components (such as object-oriented programming), linear pipelines, or other principles. Each design approach has different implications for parallelization, efficiency, maintenance and reuse.
- Software development approaches include: top-down strategies for larger systems and bottom-up strategies for smaller systems; agile software development and evolutionary prototyping when the requirements change as the project progresses; and the traditional waterfall model of

[28] http://msdse.org
[29] ImageXD: https://bids.berkeley.edu/research/image-xd, TextXD: https://bids.berkeley.edu/research/textxd
[30] http://icse2017.gatech.edu

sequential software design and development. NASA has a strong agile development program for flight software[31] but less prominently in space science applications.

- Parallel and distributed computing architectures require special programming designs. The MapReduce model, for example, separates data filtering from analysis functions and is effective in multithreaded applications on many processors. Effective use of hybrid CPU-GPU clusters requires particular attention to intra-core caching and inter-core communication.
- Software maintenance and reuse (beyond well-established libraries) are important elements of continuing or replacing legacy programs. Staff time devoted to fixing bugs and adapting code to changing system environments can be reduced with high-level debuggers and impeded by earlier weak programming practices.

Distribution of space science software

In the past, it was often difficult to promulgate software with advanced methods from an expert research group to the wider scientific community. NASA SMD's Applied Information System Research Program, along with other research programs, encountered this problem during the 1990s. But software dissemination has greatly improved mechanisms today. Source code can be deposited and updated on repositories like GitHub[32] that currently hosts tens-of-millions of open-source software projects. Well-organized software libraries and packages implementing advanced analysis for space science are also be provided through language-specific repositories. The Python Package Index currently has over 100,000 packages, and the public domain R statistical software environment has over 10,000 packages with millions of users[33]. NASA scientists played a central role in developing the important Astropy toolbox[34].

Algorithms, Machine Learning & Optimization

New approaches to data analysis are advancing quickly in the fields of machine learning and optimization, which if applied across the NASA science missions, would accelerate discoveries and support the generation of new science currently not possible. Though machine learning is at the peak of inflated expectations according to a major consulting firm[35], scientists in the astronomy and physics community are reaping tangible benefits from its application in an array of uses: real-time decision making from instruments and sensors, high performance categorization of survey data, advanced pattern recognition in high energy physics to name a few.

One key factor in utilizing machine learning approaches to science problems is knowing which method to apply on one's datasets and at what stage of the research program. Some methods are better for hypothesis generation, while others are better for model creation and understanding. The tools and reference implementations of these new methods require a familiarity with the mathematics and computer science theory if they are to be used properly. Powerful methods such as Classification Trees with Random Forests, Support Vector Machines, and Deep Learning neural networks have considerable flexibility. The user's choices of algorithmic procedure may, or may not, significantly affect the scientific conclusion. Scientists who have advanced training in

[31] https://www.nasa.gov/sites/default/files/files/FlightSoftware.pdf
[32] http://github.com
[33] http://pypi.python.org; https://cran.r-project.org/web/packages/
[34] Astropy Collaboration, 2013. "Astropy: A community Python package for astronomy". *Astronomy & Astrophysics*. **558**: A33
[35] http://www.gartner.com/newsroom/id/3412017

machine learning, or statisticians/computer scientists with cross-training in the domain of science the data sets come from, are the ones best suited to make these choices and apply machine learning appropriately to the problem at hand.

Machine learning can mean an array of different approaches to extracting insights from data. Supervised learning uses human curated / validated / tagged datasets to help a program construct a model of new input data. Unsupervised learning attempts to categorize datasets based on qualities within the data, and is used for anomaly detection, neural networks, and other clustering applications. Deep Learning which can be either supervised or unsupervised is a layered network of models corresponding to different abstractions or hierarchies in the data. Deep Learning has made significant progress towards image and scene recognition – in some cases on par with exceeding human accuracy, at a fraction of the speed. There are many other sub-areas within machine learning that could prove invaluable to NASA SMD-sponsored scientists.

Visualization

Data visualization helps scientists in a variety of stages of discovery, from exploration of new data, to real-time analysis and decision making, to interactive communication of results. Modern scale scientific data have challenges with traditional visualization techniques. Large datasets often cannot be completely loaded into memory, hindering real-time exploration. Highly complex datasets have too many variables for pair-wise comparison, hindering the discovery of correlations or anomalies.

New techniques are beginning to be available to manage these, and other similar challenges with big data visualization. By using machine learning on high dimensional datasets, recommender systems and group theoretic 'scagnostics' are able to suggest reasonable alternative to standard pair-wise comparisons, alerting researchers to combinations of variables that could be of interest[36]. For large data sets, techniques such as pre-fetching the likely next subsets of data to be visualized can speed up real time exploration of big data.

Further, the language of data visualization has become significantly more advanced in the last few years with more compelling, user-friendly tools for building and sharing graphical representations of data. For example, new languages such as Vega[37] provide a standard and expressive way to program data visualizations for both exploration and analysis. These tools take into account the ways in which humans understand information, and help researchers gain new insights into their data.

Finally, Augmented Reality and Virtual Reality technologies offer entirely new ways to visualize digital information. Though currently too expensive and early stage to be generally useful to scientists, the future of these techniques will offer immersive and infinite resolution visuals of data that will allow researchers the ability to see both the 'forest' and the 'trees' of the phenomena being

[36] C. Richthammer et al., Interactive visualization of recommender systems data, SHCIS '17, Proceedings of the 4th Workshop on Security in Highly Connected IT Systems, 19-24; Wilkinson, L. & Wills, G. (2007) Scagnostics distributions. J, Computational & Graphical Statistics, 17, 473-491.

[37] Vega, A visualization grammar, https://vega.github.io/vega/

studied. As price and size/bulk come down, every researcher will be able to have these capabilities in their lab or office.

VI. Conclusions

Though state-of-the-art techniques and practices are in practice in many places across the NASA enterprise, advanced methodologies of data science are not being uniformly integrated into data and science analysis, either within NASA or in its funded programs. NASA has not taken leadership in the development, promulgation, or community education of statistical and computational methodologies for space science. NASA has played an insufficient role in the development of data science as a profession through scholarship, investment, employment or education.

To promote improved computational and statistical methodology in the service of NASA science, the Task Force recommends:

- NASA SMD should fund and encourage its scientists and engineers to pursue professional development in statistics and informatics. Informatics is very broad including efficient algorithms, database management, Deep Learning, parallel programming, software engineering, statistical methodology, and other topics. This can include in-service workshops to promulgate best practices, attending external meetings on advances in methodology, and hiring expert consultants to bring state-of-the-art procedures to M&A efforts.
- NASA SMD should contribute to efforts to educate the wider Earth and space science communities in recent developments in statistics and informatics. This can include hosting training workshops and developing online training materials using modern platforms such as Jupyter notebooks.
- Specifications for mission science operations software development should include proposal, documentation and evaluation of computational algorithms and statistical methods when they are crucial for the software performance and science outcomes. The goal is to establish high standards for analysis methodology and algorithms within mission analysis pipelines and archive software services. Cross-disciplinary experts should participate in software performance reviews.
- NASA SMD should ensure modern software engineering (for example: agile, fast iteration processes for creation and modification of software systems) is applied to its sponsored development and maintenance projects. This may include active involvement of data science professionals as staff or consultants.

VII. Acknowledgements

Thanks to Victor Pankratius (MIT) for insights into software engineering, and to Aneta Sieminigowska (Smithsonian Astrophysical Observatory) for material from the AAS Working Group on Astroinformatics and Astrostatistics.

VIII. List of Acronyms

AAS	American Astronomical Society
AGU	American Geophysical Union

ASA	American Statistical Association
BDTF	Big Data Task Force
BDHubs	Big Data Regional Innovation Hubs (NSF)
CPU	Central processing unit
DSE	Data Science Environment (Moore-Sloan Foundation)
ESA	European Space Agency
IEEE	Institute of Electrical and Electronic Engineers
GPU	Graphical processing unit
IAU	International Astronomical Union
JWST	James Webb Space Telescope
M&A	Mission and archives
SMD	Science Mission Directorate
SQL	Structured Query Language
STScI	Space Telescope Science Institute
WFIRST	Wide Field Infrared Survey Telescope

BDTF Findings & Recommendation: **Methodologies**

Finding:
The volume, variety and velocity of NASA science data is taxing established methods and technologies. The problem arises both from data generated by science missions and from computationally intensive simulations supporting these missions.

Finding:
The enormous strides in methodology from statistics, applied mathematics and computer science of recent decades are often not incorporated into NASA satellite data or science analysis programs. High standards for analysis methodology and algorithms are not set consistently for analysis pipelines within NASA mission centers, for science analysis software maintained by NASA archive centers, or for extramural science programs funded by NASA.

Recommendation:
The BDTF recommends that NASA SMD make the necessary changes in training, proposal and mission reviews, and implementation of the critical capabilities that data science algorithms provide.

- NASA SMD should organize and fund professional development in statistics and informatics, both for its internal scientists and for the wider Earth and space science communities. This includes organizing training workshops, producing on-line training materials, attending methodology conferences, and hiring expert consultants.
- Specifications and performance reviews for mission science operations software development should include high standards for computational algorithms and statistical methods with evaluation by cross-disciplinary experts.
- NASA SMD should ensure modern software engineering (for example: agile, fast iteration processes for creation and modification of software systems) is applied to its sponsored development and maintenance projects. This may include active involvement of data science professionals as staff or consultants.

Background: see the BDTF white paper: Data Science: Statistical and Computational Methodologies for NASA's Big Data in Science.

Rationale
Data science and modern software engineering methods provide powerful insights and allow for fully leveraging the large, real-time, and complex datasets coming from NASA SMD missions and its research programs. These methods often come from adjacent fields from the normal SMD focus areas, and therefore require cross-disciplinary training, collaborations, and other technology transfer.

Consequences for not adopting the recommendation
NASA science missions will not adequately leverage data to extract the full value from the datasets made available. Inadequate methodologies will result in weak recovery of the science potential, lower quality results, and higher costs for analysis. Furthermore, there is potential for inefficient software maintenance practices.