AGU PUBLICATIONS

JGR

Journal of Geophysical Research: Space Physics

COMMENTARY

10.1002/2017JA024835

Key Points:
- Shifting data landscape of solar-terrestrial system creates opportunity for a new frontier in geospace research
- New frontier should be created at intersection of traditional approaches and state-of-the-art data-driven science and technology
- Suggest actions to sustain and expand momentum that exists to utilize emerging field of data science for geospace discovery

Correspondence to:
R. M. McGranaghan,
ryan.mcgranaghan@jpl.nasa.gov

Citation:
McGranaghan, R. M., Bhatt, A., Matsuo, T., Mannucci, A. J., Semeter, J. L., & Datta-Barua, S. (2017). Ushering in a new frontier in geospace through data science. *Journal of Geophysical Research: Space Physics*, *122*, 12,586–12,590. https://doi.org/10.1002/2017JA024835

Received 29 SEP 2017
Accepted 20 NOV 2017
Accepted article online 27 NOV 2017
Published online 16 DEC 2017

Ushering in a New Frontier in Geospace Through Data Science

Ryan M. McGranaghan[1,2], Asti Bhatt[3], Tomoko Matsuo[4], Anthony J. Mannucci[2], Joshua L. Semeter[5], and Seebany Datta-Barua[6]

[1]University Corporation for Atmospheric Research, Boulder, CO, USA, [2]NASA Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA, USA, [3]SRI International, Menlo Park, CA, USA, [4]Department of Aerospace Engineering Sciences, University of Colorado Boulder, Boulder, CO, USA, [5]Center for Space Physics, Boston University, Boston, MA, USA, [6]Department of Mechanical, Materials and Aerospace Engineering, Illinois Institute of Technology, Chicago, IL, USA

Abstract Our understanding and specification of solar-terrestrial interactions benefit from taking advantage of comprehensive data-intensive approaches. These data-driven methods are taking on new importance in light of the shifting data landscape of the geospace system, which extends from the near Earth space environment, through the magnetosphere and interplanetary space, to the Sun. The space physics community faces both an exciting opportunity and an important imperative to create a new frontier built at the intersection of traditional approaches and state-of-the-art data-driven sciences and technologies. This brief commentary addresses the current paradigm of geospace science and the emerging need for data science innovation, discusses the meaning of data science in the context of geospace, and highlights community efforts to respond to the changing landscape.

1. Introduction

We are at a crossroads in the study of geospace (i.e., the vast, coupled, complex region extending from the Sun, through interplanetary space, to the Earth's magnetosphere and upper atmosphere (Lotko, 2017)). On one hand we operate in the same paradigm that has guided the field over the past couple of decades, ruled by the triumvirate of models, data, and more recently, model-data fusion. On the other hand we are beginning to recognize that powerful new opportunities for scientific discovery are possible through increased data volume and sophisticated methods to explore these data. The emergence of the hyperconnected digital society and the massive quantities of data it generates has led to new analysis capabilities that scale well to the geospace environment. The geospace sciences are squarely positioned to benefit from the emerging field of data science, which enables the creation of new scientific insights from data through the union of statistics, computer science, applied mathematics, and visualization.

Data science does not simply provide off-the-shelf solutions. Rather, domain-specific knowledge is required to guide effective use of data science solutions. Therefore, it is important to first understand what data science means in the context of the geospace domain. We begin by describing the evolution of geospace science and use this background to discuss the current intersection of data science and geospace. Finally, outstanding questions are raised regarding the expansion of the relationship moving forward.

2. How Is Geospace Science Done Currently?

Geospace science has its origins in the nineteenth century but really began to take form when radio communication both necessitated and enabled the first scientific studies of the ionosphere (the charged region of the Earth's upper atmosphere between ~100 and 1,000 km) (Hargreaves, 1992). The space age and advent of capable computers beginning in the 1950s gave rise to in situ observation and numerical simulations of the space environment. Increasingly sophisticated computer simulations have followed over the decades. In parallel, geospace instrumentation has continually matured. Starting from the first radio observations by Edward Appleton, geospace remote sensing instrumentation in both radio and optical domains has benefitted from modern technological advances. The result is a preponderance of small- and large-scale distributed ground- and space-based systems sampling the geospace environment across multiple scales.

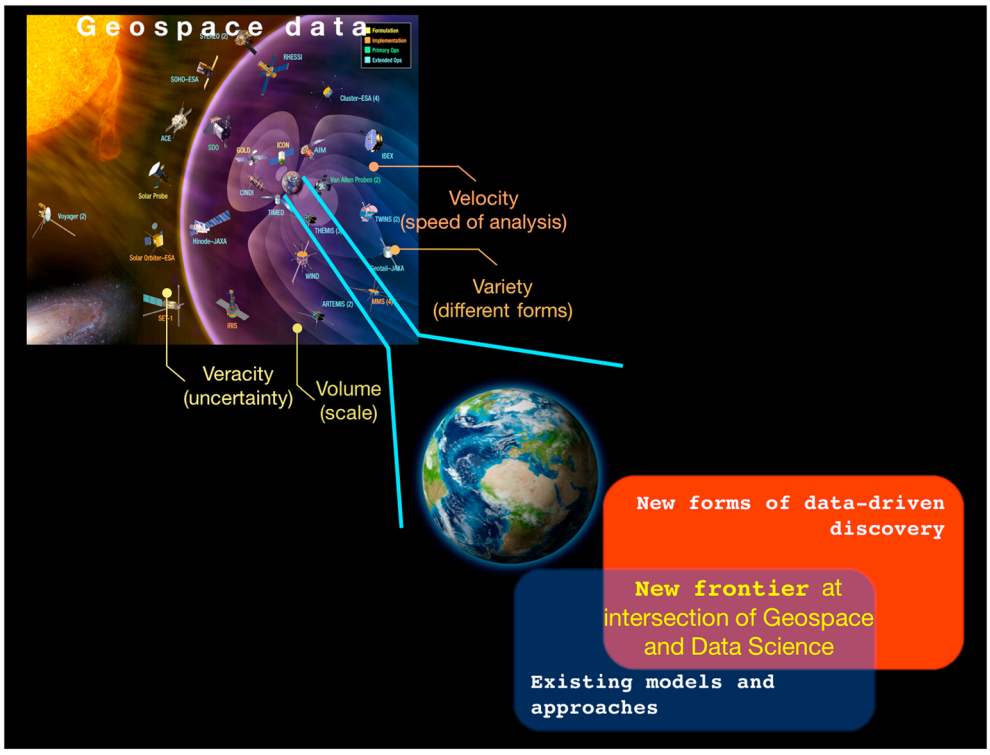

Figure 1. The landscape of geospace science data and opportunity for a new frontier. Illustrated are the four Vs of data: volume (growing number of contributing observing platforms), variety (diversity of observing platforms), veracity (uncertainty of data), and velocity (the speed at which data are observed and analyzed). *NASA Heliophysics Systems Observatory fleet image credit: NASA* — https://www.nasa.gov/content/goddard/heliophysics-system-observatory-hso.

Now, instruments observe each component of geospace yet lack the spatiotemporal coverage to produce a unique Sun-to-Earth solution. This ushered in an era of model-data fusion, and data assimilation (following its success for terrestrial weather prediction) became the cutting-edge. With a growing volume of data, and the prospect of, perhaps, an exponential trend (National Research Council, 2002), computation-intensive data-driven techniques (i.e., machine learning) are gaining prominence. Computer-aided data mining for discovery has indeed proliferated in the modern era of geospace science (Lundstedt, 2005; Pankratius et al., 2016).

The importance of data-intensive techniques will only grow as 21st century advances increasingly entrench geospace science in the realm of "big data," grappling with the four Vs: volume, variety, veracity, and velocity (Bell et al., 2009) (see Figure 1).

With increased data come challenges and complexities relating to their curation, analysis, and visualization (the data life cycle). In geospace analyses, opportunities to improve the data life cycle are ubiquitous. These include more efficient methods to process large data sets, enhanced ability to take advantage of diverse data sets (Lotko, 2017), and effective means for data and technology sharing and digital scholarship (Gil et al., 2016) that will permit new efforts to build on previous work and facilitate multidisciplinary, collaborative research. Improvements in these areas have the potential to bring about a *new frontier* in geospace.

3. Data Science in the Context of the Geospace Community

Constructively organizing these efforts is difficult and is made more difficult by a lack of consistent terminology. Data science can provide solutions in both regards. Broadly defined, data science refers to scalable architectural approaches, techniques, software, and algorithms that alter the paradigm by which data is collected, managed, analyzed, and communicated. It is a combination of statistics, computer science, and domain knowledge. Data science is ubiquitous in our society (Kalil, 2016), impacting virtually every arena, from politics to medicine, economics to culture. Data science-driven transformation in fields similar to geospace, such as Earth science (Yue et al., 2016) and climate research (Carleton & Hsiang, 2016), is emblematic of the immense

potential, and data science is, accordingly, beginning to gain attention in the geospace community. Here we discuss new and ongoing data science-related efforts across the geospace community, including both federally driven initiatives and community-driven grassroots efforts.

The National Science Foundation (NSF) EarthCube Project (https://earthcube.org/) is a joint effort between the NSF Directorate for Geosciences and the Division of Advanced Cyberinfrastructure that brings together scientists across the geoscience domain with geoinformatics researchers and data scientists to (1) influence how data will be collected, accessed, analyzed, visualized, shared and archived; (2) participate in interdisciplinary and collaborative research; and (3) contribute to the transformation of geoscience research through the emerging practices of digital scholarship, data and software stewardship, and open science.

Two geospace-specific NSF EarthCube projects are underway.

The Integrated Geosciences Observatory (InGeO—https://ingeo.datatransport.org/app/) is a pilot program that aims to cut across the entire Sun-to-Earth system by integrating separate data sets and technologies into a single, unified online platform, using contemporary open-source, community-created solutions. InGeO embraces the transdisciplinary ethos of data science, offers increased access to diverse data sets, analysis tools, and assimilative models and represents a scalable approach to geospace science (i.e., offering online, rather than local, computing). An important goal of the InGeO project is to improve the process of publishing data and recording geospace-specific ontology. As part of the InGeo program, a Python implementation of the procedures that solve the problem of Assimilative Mapping of Ionospheric Electrodynamics (Richmond & Kamide, 1988) as described in Matsuo et al. (2005) is being made available to the public, exemplifying a new culture of technology sharing.

While InGeO signals a shift in the manner in which data are accessed, described, and shared, the Magnetosphere-Ionosphere-Atmosphere Coupling (MIAC—https://www.earthcube.org/group/magnetosphere-ionosphere-atmosphere-coupling) NSF EarthCube project is developing cyber-based tools to better use diverse and voluminous data sets across the MIA system. MIAC will specifically enhance the utility of the next generation of the Active Magnetosphere and Planetary Electrodynamics Response Experiment and the SuperMAG global magnetometer and Super Dual Auroral Radar networks by unifying the information from each. While the MIAC project illustrates a set of data science techniques to exploit a specific set of geospace data, other data sets may very well require a different set of techniques. Exploration is critical.

The expansive EarthCube projects signal that innovation in geospace is focused on data science solutions and that agencies like the NSF recognize the importance of such efforts. Additionally, pockets of the larger geospace community are separately realizing the power of data science solutions. Machine learning methods are now being applied for prediction throughout the Sun-to-Earth system, including for solar flares and coronal mass ejections (e.g., Bobra & Couvidat, 2015, 2016), the magnetosphere (e.g., Bortnik et al., 2016), the ionosphere (e.g., Uwamahoro & Habarulema, 2015), and for solar and geomagnetic indices (e.g., Rezende et al., 2010; Lazzus et al., 2017).

Bringing together the software-centric and cyberinfrastructure focus of EarthCube and the growing number of specific applications of data science in geospace, projects such as *Computer-Aided Discovery in Geoscience* (https://esto.nasa.gov/files/solicitations/AIST_16/ROSES2016_AIST_A41_awards.html#victor), a NASA Advanced Information Systems Technology Program-funded project, aim to create new online, scalable algorithms, tools, and data fusion capabilities to bridge the entire process from raw data collection to new scientific discovery (Pankratius et al., 2016).

Data science goes beyond methods and analyses, encompassing effective communication of data as well. Effective communication requires clear, concise explanation and powerful visualization. New tools, such as Data-Driven Documents (D3.js—https://d3js.org/), allow data to be represented interactively, and the geoscience community is embracing new presentation formats to promote more dynamic data visualization and communication options (e.g., the American Geophysical Union (AGU) eLightning poster talks—http://fallmeeting.agu.org/2017/elightning-sessions-presentations/).

Each of these efforts provides a foundation on which to build a new frontier at the intersection of data science and geospace science. Conversations across the community are already attempting to formally define this new frontier, albeit often referring to data science obliquely. For example, the *Exploring the Geospace Frontier: Quo Vadis?* conference held in Boulder, Colorado, in May 2016 brought together scientists from across the geospace domain to discuss cross-cutting solutions to address challenges in

space weather prediction (workshop report at https://www2.hao.ucar.edu/sites/default/files/users/sheryls/QuoVadisWorkshopReport.pdf). At the 2017 Coupling, Energetics, and Dynamics of Atmospheric Regions (CEDAR) summer workshop three workshop sessions aligned well with the theme of embracing data science in space physics, particularly in the context of the near Earth space environment: (1) geospace science in the digital age: new tools and methods (primary convener: Asti Bhatt; http://cedarweb.vsp.ucar.edu/wiki/index.php/2017_Workshop:Digital_Geospace); (2) next generation systems science: embracing data fusion and data science methods to understand geospace complexities (primary convener: Ryan McGranaghan; http://cedarweb.vsp.ucar.edu/wiki/index.php/2017_Workshop:Next_generation_systems_science); and (3) the high-latitude geospace system: frontiers in science and sensing (primary convener: Josh Semeter; http://cedarweb.vsp.ucar.edu/wiki/index.php/2017_Workshop:High_Latitude_System_Frontiers). Each of these sessions explored ideas for innovation in the CEDAR community covering all aspects of the data life cycle: capture, analysis, and understanding. The major points that emerged during these sessions were as follows: (1) Given the expanse of the geospace system, we must rely on distributed, diverse observations to provide observational support for new understanding; (2) data science-based innovation is required to gain the most utility from the diverse observational system; and (3) data science-driven efforts allow new scientific understanding to emerge.

A key component to acting on these ideas is the "democratization" of technology. The Python for Space Science session at CEDAR (primary convener: Russell Stoneback; http://cedarweb.vsp.ucar.edu/wiki/index.php/2017_Workshop:Python_for_Space_Science) illustrated progress toward better software sharing practices for the study of geospace.

4. What Is Next?

We must ask how to maintain and extend the momentum that has been generated to use data science for discovery in the geospace sciences. We suggest that three actions are paramount:

1. A tight integration among solar, magnetospheric, ionospheric, and thermospheric communities must be targeted.
2. To optimally realize the potential of the modern data and information sciences in the geosciences, we must continue to reduce barriers (both organizational and language-based) between geoscientists, computer scientists, and engineers. Funding agencies should continue to promote interdisciplinary programs to foster such collaborations.
3. Partnerships between NASA, the Department of Defense, and the NSF are needed to realize the full potential of experiments involving collaborative measurements from ground and space.

Community commitment to these paths will catalyze more effective utilization of the data sciences for the benefit of geospace science. Opportunities for individuals and teams to get involved exist at all levels, including initiating interdisciplinary projects at their own institutions, joining open informatics groups (e.g., AGU Earth and Space Science Informatics—http://essi.agu.org/), and taking advantage of existing open-source tools made available by the data science community. Gil (2016) is a useful resource for getting started, providing references to specific tools and practices. Data science and machine learning conferences and journals offer opportunities for increased interaction with data scientists and will be important to spark fruitful collaboration. Pragmatically, improvements to the data life cycle will produce new actionable knowledge to inform decisions regarding space weather, an urgent need for social and economic well being and security nationally (National Science and Technology Council, 2015a, 2015b) and internationally (Guhathakurta et al., 2013). It is clear that the timing is right for our community to embrace data science in a unified way.

Acknowledgments
This research was supported by the NASA Living With a Star Jack Eddy Postdoctoral Fellowship Program, administered by the University Corporation for Atmospheric Research and coordinated through the Cooperative Programs for the Advancement of Earth System Science (CPAESS). Portions of this research were carried out at the Jet Propulsion Laboratory, California Institute of Technology, under a contract with the National Aeronautics and Space Administration. We gratefully acknowledge John Bosco Habarulema for his valuable input. This paper is conceptual and neither used nor generated data to be made available.

References

Bell, G., Hey, T., & Szalay, A. (2009). Beyond the data deluge. *Science*, *323*(5919), 1297–1298. https://doi.org/10.1126/science.1170411

Bobra, M. G., & Couvidat, S. (2015). Solar flare prediction using SDO/HMI vector magnetic field data with a machine-learning algorithm. *The Astrophysical Journal*, *798*(2), 135.

Bobra, M. G., & Ilonidis, S. (2016). Predicting coronal mass ejections using machine learning methods. *The Astrophysical Journal*, *821*(2), 127.

Bortnik, J., Li, W., Thorne, R. M., & Angelopoulos, V. (2016). A unified approach to inner magnetospheric state prediction. *Journal of Geophysical Research: Space Physics*, *121*, 2423–2430. https://doi.org/10.1002/2015JA021733

Carleton, T. A., & Hsiang, S. M. (2016). Social and economic impacts of climate. *Science*, *353*(6304), aad9837. https://doi.org/10.1126/science.aad9837

Gil, Y. (2016). The scientific paper of the future: OntoSoft training. https://doi.org/10.5281/zenodo.159206, http://www.scientificpaperofthefuture.org

Gil, Y., David, C. H., Demir, I., Essawy, B. T., Fulweiler, R. W., Goodall, J. L., … Yu, X. (2016). Toward the geoscience paper of the future: Best practices for documenting and sharing research from data to software to provenance. *Earth and Space Science*, *3*, 388–415. https://doi.org/10.1002/2015EA000136

Guhathakurta, M., Davila, J. M., & Gopalswamy, N. (2013). The International Space Weather Initiative (ISWI). *Space Weather*, *11*, 327–329. https://doi.org/10.1002/swe.20048

Hargreaves, J. (1992). *The solar-terrestrial environment: An introduction to geospace—The science of the terrestrial upper atmosphere, ionosphere, and magnetosphere*. Cambridge Atmospheric and Space Science Series: Cambridge University Press.

Kalil, T. (2016). Charter of the Data Science Interagency Working Group Committee of Technology National Science and Technology Council.

Lazzus, J. A., Vega, P., Rojas, P., & Salfate, I. (2017). Forecasting the *Dst* index using a swarm-optimized neural network. *Space Weather*, *15*, 1068–1089. https://doi.org/10.1002/2017SW001608

Lotko, W. (2017). The unifying principle of coordinated measurements in geospace science. *Space Weather*, *15*, 553–557. https://doi.org/10.1002/2017SW001634

Lundstedt, H. (2005). Progress in space weather predictions and applications. *Advances in Space Research*, *36*(12), 2516–2523. https://doi.org/10.1016/j.asr.2003.09.072

Matsuo, T., Richmond, A. D., & Lu, G. (2005). Optimal interpolation analysis of high-latitude ionospheric electrodynamics using empirical orthogonal functions: Estimation of dominant modes of variability and temporal scales of large-scale electric fields. *Journal of Geophysical Research*, *110*, A06301. https://doi.org/10.1029/2004JA010531

National Research Council (2002). *Assessment of the usefulness and availability of NASA's Earth and space science mission data*. Washington, DC: The National Academies Press. https://doi.org/10.17226/10363

National Science and Technology Council (2015a). *National Space Weather Action Plan*. USA: Executive Office of the President (EOP).

National Science and Technology Council (2015b). *National Space Weather Strategy*. USA: Executive Office of the President (EOP).

Pankratius, V., Li, J., Gowanlock, M., Blair, D. M., Rude, C., Herring, T., … Lonsdale, C. (2016). Computer-aided discovery: Toward scientific insight generation with machine support. *IEEE Intelligent Systems*, *31*(4), 3–10. https://doi.org/10.1109/MIS.2016.60

Rezende, L. F. C., de Paula, E. R., Stephany, S., Kantor, I. J., Muella, M. T. A. H., de Siqueira, P. M., & Correa, K. S. (2010). Survey and prediction of the ionospheric scintillation using data mining techniques. *Space Weather*, *8*, S06D09. https://doi.org/10.1029/2009SW000532

Richmond, A. D., & Kamide, Y. (1988). Mapping electrodynamic features of the high-latitude ionosphere from localized observations: Technique. *Journal of Geophysical Research*, *93*(A6), 5741–5759. https://doi.org/10.1029/JA093iA06p05741

Uwamahoro, J. C., & Habarulema, J. B. (2015). Modelling total electron content during geomagnetic storm conditions using empirical orthogonal functions and neural networks. *Journal of Geophysical Research: Space Physics*, *120*, 11,000–11,012. https://doi.org/10.1002/2015JA021961

Yue, P., Ramachandran, R., Baumann, P., Khalsa, S. J. S., Deng, M., & Jiang, L. (2016). Recent activities in Earth data science [technical committees]. *IEEE Geoscience and Remote Sensing Magazine*, *4*(4), 84–89. https://doi.org/10.1109/MGRS.2016.2600528