2018 Workshop: Next generation CEDAR science

Long title
Next generation CEDAR science: Addressing geospace system science in the age of data science
Conveners
Ryan McGranaghan
Tomoko Matsuo
Asti Bhatt
Steven Morley
Description

Data science can be defined as the use of scalable approaches and techniques to create new discovery from data through the union of domain-specific knowledge, statistics, computer science, applied mathematics, and visualization. Data science represents an evolutionary step in interdisciplinary research and is a critical component of the next generation of CEDAR science (see the CEDAR strategic plan). This session will explore the role of data science in the CEDAR community, including the cutting-edge developments and the intersection with established methods and models. We will emphasize the importance of a multi-disciplinary, system science approach to produce new geospace discovery and will provide a forum to discuss a more coordinated effort to implement data science innovation in CEDAR.

Agenda

We heard from an excellent set of speakers that helped progress the conversation around the use of data science in geospace. Ryan McGranaghan opened the session with a scene-setting presentation. These slides as well as contributed discussion topics from each of the session conveners can be found here: NextGenCEDAR Discussion Slides (pdf)

Central Points:

- Powerful use cases allow us to observe and begin to realize the potential of data science to progress geospace understanding

- Radically interdisciplinary collaboration is necessary to achieve success, and communication across interdisciplinary communities is a challenge
- Innovative new data science solutions should ***complement*** more traditional approaches in geospace
- Data science is a broad field, encapsulating the entire data lifecycle (from data storage and management through analysis and visualization). It is ***NOT*** simply machine learning
- We must invite the most capable/advanced people from data science community (specifically data curation, machine learning, and visualization) to participate in our community. This raises the question of how to get those people invested and interested in geospace science?

Speaker summaries:

First, we heard from Liz MacDonald about the 'The spark of crowdsourced opportunities & outcomes for Heliophysics'. She raised several though-provoking questions, including: 1) Is citizen science a new frontier for data science in geospace? 2) Why would we expect the same approaches to science that have guided progress for the past 10, 20, and 30 years (her point: we should focus on 'doing new things')? and 3) Can new discovery be created at the 'gaps' between fields (i.e., at the edges)? She presented compelling evidence about the potential of citizen science from the discovery of STEVE (Strong Thermal Emission Velocity Enhancement) aurora. Key questions from the subsequent discussion included: 1) How do we remove/avoid jargon for geospace research? and 2) Are there 'right' problems for citizen science? MacDonald NextGenCEDAR (pdf)

Next, Christine Gabrielse began her presentation by addressing the four V's of 'big data': 1) volume; 2) variety; 3) veracity (i.e., uncertainty); and 4) velocity. She illustrated that 'variety' may be the niche that geospace currently best aligns with. She highlighted the data variety challenge through an analysis of mesoscale structures in the magnetosphere-ionosphere system and the progress that heterogeneous data fusion can achieve. Based on her excellent research, she suggested that our community needs to embrace data variety in the near-term. Conversation after her talk centered around the challenge of making data 'usable'. Gabrielse NextGenCEDAR (pdf)

Chen Zhou gave an exciting overview of his research into using deep learning methods (specifically generative adversarial networks, GANs), achieving progress

toward the challenge of incomplete and sparse data in geospace as a parallel to image completion problems. Chen made a prescient point that collaboration with computer scientists is critical to successfully applying these advanced machine learning methods. Zhou NextGenCEDAR (pdf)

Liam Kilcommons followed Chen's presentation by providing a clear and compelling overview of an under-utilized, yet important, method for examining the robustness of data-driven discovery in geospace: cross-validation. He highlighted k-fold cross-validation using Super Dual Auroral Radar Network (SuperDARN) and Advanced Magnetospheric and Planetary Electrodynamics Response Experiment (AMPERE) data. His presentation was a glowing example of how new geophysical understanding can be obtained through data-driven methods; a testament to a theme of the session that innovative new data science solutions can complement more traditional approaches in geospace. Kilcommons NextGenCEDAR (pdf)

Piyush Mehta then used the ionosphere-thermosphere system as a case study to apply a new data assimilative framework. He raised important issues such as the distinction between theoretical, empirical, data-driven, and computational models and the idea of 'predictability' of the ionosphere. His presentation emphasized four points critical to data assimilation and areas of importance discussed throughout the CEDAR 2018 workshop: 1) model errors; 2) time step evolution; 3) initialization; and 4) calibration. The post-presentation discussion centered around the best approaches to capturing nonlinearities of geospace in the data assimilative model. Mehta NextGenCEDAR (pdf)

Finally, Eric Sutton concluded the contributed talks with a fascinating look at new approaches to thermospheric neutral density specification. He framed his new approach through a clarification between chaotic (e.g., the troposphere) and strongly-driven systems (e.g., the ionosphere-thermosphere). His new assimilative approach demonstrated accuracy with respect to the Challenging Minisatellite Payload (CHAMP) and Gravity Recovery and Climate Experiment (GRACE) observations. In the discussion-oriented ethos of the session, Eric raised intriguing questions: 1) How do estimated drivers represent reality? and 2) How to disentangle solar and geomagnetic influences? Sutton NextGenCEDAR (pdf)

Justification

Geospace is experiencing a massive growth of the volume of data that can be used for discovery, from diverse observing platforms to inexpensive sensors to simulation output, firmly entrenching the discipline in the realm of big data. The growth of data is accompanied by challenges for its efficient use and analysis. However, the emergence of the hyper connected digital society and high-performance computing (e.g., cloud-computing) has led to new technologies and analysis capabilities that scale well to the geospace environment. These developments have revolutionized the creation of new scientific insights from data through the union of statistics, computer science, applied mathematics, and visualization (i.e., data science). Both the National Research Council (NRC) Decadal Survey and the CEDAR Strategic Plan emphasize the importance of system science approaches to unraveling the multi-scale complexity and coupling of geospace. The changing data landscape and growth of data science squarely position the CEDAR community to usher in the next generation of geospace system science, taking advantage of growing amounts of data and new analysis tools

We aim to bring together a multi-disciplinary group from across the disciplines of space physics, statistical analysis, and computer and data sciences to:

1. Discuss the application of cutting-edge data science approaches (e.g., machine learning) to geospace system science;
2. Provide a forum to navigate the intersection between innovative data science tools and established methods and models; and
3. Outline the paths from methodology to new fundamental understanding.

To accomplish these goals this workshop will focus on innovation to address the complexities of system science research. We most directly respond to the Key Science Goal 2 of the NRC Decadal Survey: Determine the dynamics and coupling of Earth's magnetosphere, ionosphere, and atmosphere and their response to solar and terrestrial inputs.

Additionally, this session supports progress towards several of the CEDAR Strategic Plan Thrusts:

1. Strategic Thrust #1: Encourage and Undertake a Systems Perspective to Geospace
    - We will coordinate methods capable of providing new understanding about complex processes in the geospace system, including multi-scale specification, cross-scale feedback, and nonlinearity.
2. Strategic Thrust #5: Fuse the Knowledge Base across Disciplines
    - We will start a conversation to identify synergies between the field of space science and the quickly emerging fields of data fusion, data science and machine learning.
    - We hope to create a multi-disciplinary community dedicated to the objectives outlined above to promote sessions at future workshops and conferences.
3. Strategic Thrust #6: Manage, Mine, and Manipulate Geoscience Data and Models
    - Innovative methods will contribute to more effective utilization of the geospace observational system.

Finally, to coordinate the integration of data science into CEDAR research, we propose to lay the foundation for a CEDAR Data Science Working Group (CDSWG). We envision that the CDSWG would provide recommendations to the CEDAR Science Steering Committee about breadth, depth, strategic plans, partnerships, and organization of data science efforts within CEDAR. We believe the CEDAR summer workshop is an ideal forum to coordinate such an effort.

Our formal objectives with this workshop session are:

1. Evaluate the role of state-of-the-art data science methods in the next generation of geospace system science research
2. Understand the intersection between innovative data science tools and established methods and models
3. Inspire the community to embrace data science and plan to sustain momentum at future workshops and conferences
    - Fall AGU 2018: Prepare for formal discussions and to involve the larger community
    - New Geospace Environment Modeling Focus Group (GEM FG): Identify strong links between GEM and CEDAR communities to address geospace complexity

This workshop will serve as a critical component in the long-term goal to promote stronger collaboration within (lower atmosphere, upper atmosphere, magnetosphere) and outside of (space physics, computer science, applied math) the space sciences discipline.

We will leverage the progress and sustain the momentum created through previous CEDAR workshops, specifically "Next generation systems science: Embracing data fusion and data science methods to understand geospace complexities" convened in 2017.

Summary

We heard from an excellent set of speakers that helped progress the conversation around the use of data science in geospace. Ryan McGranaghan opened the session with a scene-setting presentation. These slides as well as contributed discussion topics from each of the session conveners can be found here: NextGenCEDAR Discussion Slides (pdf)

Central Points:

- Powerful use cases allow us to observe and begin to realize the potential of data science to progress geospace understanding
- Radically interdisciplinary collaboration is necessary to achieve success, and communication across interdisciplinary communities is a challenge
- Innovative new data science solutions should ***complement*** more traditional approaches in geospace
- Data science is a broad field, encapsulating the entire data lifecycle (from data storage and management through analysis and visualization). It is ***NOT*** simply machine learning
- We must invite the most capable/advanced people from data science community (specifically data curation, machine learning, and visualization) to participate in our community. This raises the question of how to get those people invested and interested in geospace science?

View PDF