2022 Workshop: Data Science and Open Science in CEDAR

Long title
Data Science and Open Science in CEDAR: Data Science and Open Science in action in CEDAR
Conveners
Ryan McGranaghan
Asti Bhatt
Dogacan Ozturk
Bharat Kunduri
Ryan.Mcgranaghan@colorado.edu
Description

Our specific objectives will be to:

- Build on a many year foundation of CEDAR Data Science advances, establishing CEDAR as a leader among the scientific domains in unifying data and domain science;
    - Promote interaction and collaboration between the CEDAR community and related disciplines (e.g., Earth Science);
    - Improve agility and capability within CEDAR science through embracing newer technologies and sound digital data scholarship;
    - Grow methodology transfer to enhance CEDAR science; and
    - Create materials to lead to a 2023 CEDAR Grand Challenge.
- This year, our session will target a draft document to become a proposal for a 2023 Grand Challenge around data science transforming CEDAR science
- A new dimension in 2022 will be to center Open Science in the discussion

Outcomes: Progress toward these objectives will prepare us to contribute to the Heliophysics Decadal Survey (outlining the future of our broader science domain) and to cultivate a community that is knowledgeable of data science and open science over the coming decade. Generative discussions will increase our community's competitiveness in the NSF big ideas and ultimately will advance the New Frontier of CEDAR research [McGranaghan et al., 2017

] that this series of “Data Science in CEDAR” workshops have helped create. Additional outcomes will include:

- Identify the powerful use cases to advance data science capabilities within CEDAR;
- Sustain and amplify earlier data science efforts for CEDAR science applications;
- Encourage and facilitate the adoption of data science in the CEDAR community;
- Curate a community and the corresponding capacities for a more structured foundation for data science in CEDAR science; and
- Define and imagine Open Science in CEDAR and how it will connect us to other sciences

Agenda

1. Data science in CEDAR - Introductory remarks
2. Rebecca Bishop (5-7 min)
3. Yue Deng (5-7 min)
4. Yang Pan (5-7 min)
5. Pablo Reyes (5-7 min)
6. Open forum for discussion on data science in CEDAR (10-15 min)
7. Open science in CEDAR - Introductory remarks
8. Ryan McGranaghan (5-10 min)
9. Open discussion

Justification

Characterizing the geospace environment requires measurements from several regions within the geospace. Fortunately, data to advance the scientific understanding of the geospace environment are growing across the four V’s of ‘big data’: 1) Volume; 2) Variety; 3) Veracity (i.e., uncertainty); and 4) Velocity. This growth represents both a challenge to efficiently and comprehensively utilize these data, and an opportunity for new discovery by embracing 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.

Specifically in 2022, we will highlight the theme of open science, defined broadly as: "Open science is transparent and accessible knowledge that is shared and

developed through collaborative networks." - Vicente-Saez & Martinez-Fuentes [2018]

We will sustain a focus on data science as addressing the full data lifecycle:

- Data collection: use of data science to more intelligently collect data
- Data management: use of data science to more intelligently structure data (e.g., linking data and knowledge graphs)
- Data analysis: use of data science to more intelligently relate input to output (e.g., machine learning)
- Data communication: use of data science to more intelligently visualize and relate data.

There has now been a series of devoted CEDAR Data Science sessions dating back to 2017, that have continuously supported our community in unifying data and domain sciences and evolved to meet the new demands/challenges.The progress our community has made sets the stage for a new session that will not only continue to share the latest progress, but will also solidify the CEDAR community as a guiding example as we outline the next decade of Heliophysics.

Therefore, the proposed workshop is a timely effort to sustain and amplify the momentum from what is now a long legacy of advancing CEDAR science through data science, including the following selected workshops that the conveners have planned or been central contributors to:

- Next Generation System Science (2017)
- Digital Geospace (2017)
- Grand Challenge: Multi-scale I-T System Dynamics (started in 2018 with multiple sessions - see, specifically, Ryan McGranaghan's introduction to our Grand Challenge from the data perspective)
- Next Generation CEDAR Science (2018)
- The challenge, opportunity, and art of data science for geospace (2019)
- Data Science in CEDAR: Progress, Capacity-Building, and Traversing Disciplines (2020)
- Data Science in CEDAR: CEDAR Data Science as a guide for the Heliophysics Decadal Survey (2021)

This session will respond to several thrusts of the Decadal Survey:

- Determine the origins of the Sun’s activity and predict the variations of the space environment, - Enable effective space weather and climatology capabilities, and - The need to establish a space weather research program to effectively transition research to operations;

and the CEDAR Strategic Plan:

- Strategic Thrust 6 : Manage, Mine, and Manipulate Geoscience Data and Models,
- Strategic Thrust 1 : Encourage and Undertake a Systems Perspective to Geospace;

which collectively emphasize a need to embrace data science.

Finally, this session responds directly to NASA's Strategy for Data Management and Computing for Groundbreaking Science

, the strategic vision for 2019-2024.

Additionally, the National Science Foundation announced new investments that will be made toward their 10 ‘big ideas’, particularly focusing on two ideas that together objectify radically interdisciplinary work and data science across the scientific landscape:

- Harnessing data revolution

Convergence research

The members of the CEDAR community are making valuable strides to embrace and create a structure for data science and NSF big ideas. Therefore, this session will extend the conversation around increasing capability to address data challenges and opportunities and growing convergence in the CEDAR community.

Workshop format
Short Presentations
View PDF