2019 Workshop: Geospace Data Science

Long title
The challenge, opportunity, and art of data science for geospace
Conveners
Ryan McGranaghan
Bharat Kunduri
Jade Morton
Eric Donovan
Asti Bhatt
Description

Timing is ripe for the CEDAR community to embrace data science and the NSF big ideas. Therefore, this session will create a new conversation around increasing capability to address data challenges and opportunities and growing convergence in the CEDAR community.

Our specific objectives will be to:

1. Identify problems and challenges that can immediately be addressed using data science tools (i.e., the compelling and transformational 'use cases');
2. Promote interaction and collaboration between the CEDAR community and related disciplines (e.g., Earth Science);
3. Improve agility and capability within CEDAR science; and
4. Grow methodology transfer to enhance CEDAR science.

Outcomes:

- Progress toward these objectives will increase our community's competitiveness in the NSF big ideas and usher in a New Frontier of CEDAR research McGranaghan et al., 2017. 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; and

- Curate a community to develop a new CEDAR Grand Challenge Workshop in 2020, including targeted objectives, roadmap, and a draft proposal.

Agenda

Thursday 20 June 2019 10 AM - 12 PM (Foundations of data science in CEDAR science)

- 10:00 - 11:40 AM: Talks (12-minute talks, 2-minute discussions)

- Tomoko Matsuo (data assimilation; what it takes to fuse observations in geospace)
- Farzad Kamalabadi (what is data science and how has it evolved in CEDAR science) (pdf)
- Steve Morley (machine learning and relationship to traditional statistical approaches) (pdf)
- Kristina Lynch (prediction versus 'interpolation'/'filling in the blanks' approaches and what we need to know) (pdf)
- Russell Stoneback (the data wrangling side of data science; Pysat) (pdf)
- Jim Ahrens LANL "Data Science at Scale" (big data in CEDAR/space science and tools to navigate it) (pdf)
- Yun-Ju Chen UT Dallas (applied data science in CEDAR applications) (pdf)

- 11:40 AM - 12:00 PM: Contributed Talks (5 minutes - focused on provocation) & Open Discussion:

- Foci:
  - Emerge questions and topics for the afternoon panel session
  - Illustrate a concrete application or use case of data science in CEDAR science
- Jenny Yang - Data and Machine Learning Challenges via an analysis of GNSS Network Position Errors during the March 2015 St. Patrick Storm (pdf)
- Muhammad Rafiq - Google Summer of Code and benefits of non-traditional partnerships (pdf)
- Asti Bhatt: Machine learning results from Frontier Development Laboratory (pdf)
- Gonzalo Cucho-Padin - "Optical tomography in CEDAR science and the data challenges and solutions" (pdf)

Thursday 20 June 2019 1:30 - 3:30 PM (Emerging the trends and gaps for

data science in CEDAR science and creating the needed new connections)

- 1:30 - 2:15 PM Panel

  - Short introduction by Ryan McGranaghan followed by 2-minute introduction by each panel member
    - Questions will be solicited from the audience
  - Nathaniel Frissel - (citizen science)
  - Seebany Datta-Barua (CEDAR science at intersection of physics and engineering)
  - Susan Skone (advanced instruments and intelligent operation for CEDAR science; Transition Region Explorer (TREx))
  - Enrico Camporeale (trends in machine learning)
  - Laura Mazzaro (Descartes Labs - the utility of data science for the geosciences)

- 2:15 - 3:00 PM Breakout groups

  - *Each moderator responsible to come up with a set of provocative questions that drive the topical conversation to the session goals; the more visual and concrete, the better
  - Machine learning applications in geospace (success stories, lessons learned, and trends)
    - Moderator: Bharat Kunduri
  - Data provenance; Modernization of geospace science workflows using community recommended best practices (e.g., the use of open source software and cloud computing)
    - Moderator: Asti Bhatt
  - Interdisciplinary efforts (best practices, potential applications)
    - Moderator: Eric Donovan
  - Intersection of physics-based and data-driven methods; Validation
    - Moderator: Jade Morton
  - Common misconceptions about data science, machine learning, and artificial intelligence & ML adoption
    - Moderator: Ryan McGranaghan
  - *Potential*: ‘Going beyond accuracy’: robust evaluation of ML models
    - Moderator: TBD

Justification

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).

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, and
- Strategic Thrust 1 : Encourage and Undertake a Systems Perspective to Geospace;

which collectively emphasize a need to embrace data science.

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

View PDF