Welcome to “Data Science in CEDAR: Progress, Capacity-Building, and Traversing Disciplines”

We will begin promptly at 11 AM MT

Please mute yourself to prevent background noise

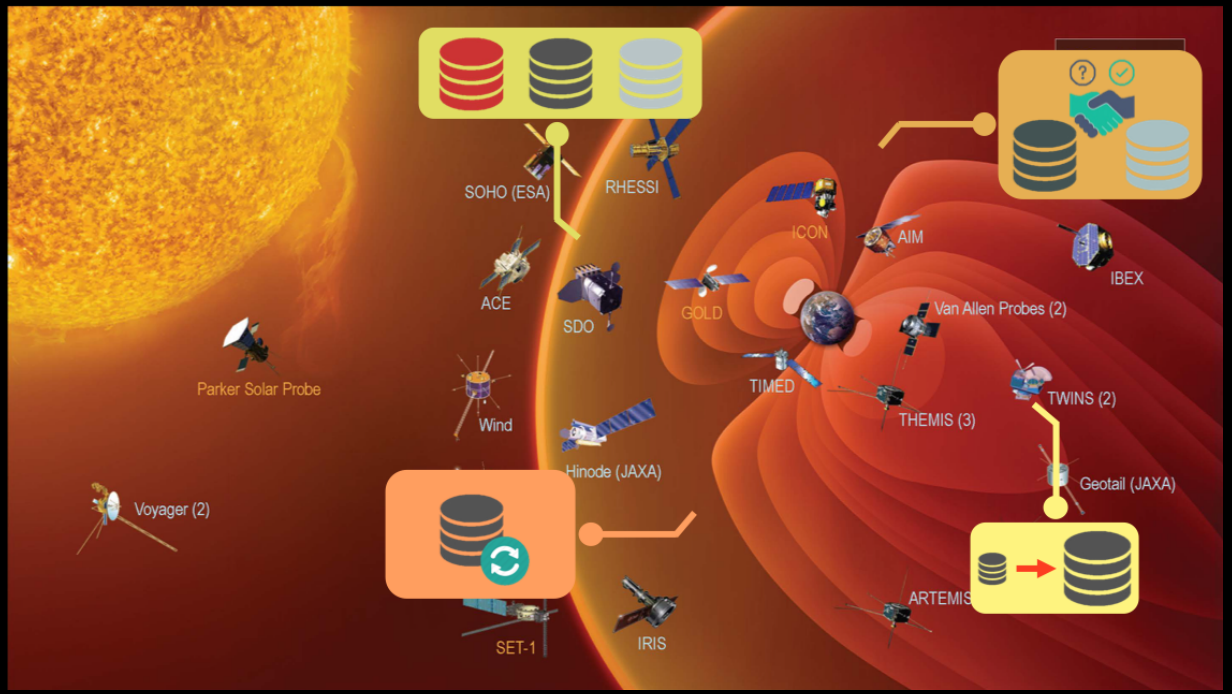
And

join the discussion on the CEDAR Slack channel #cedar-data-science

CEDAR DATA SCIENCE

Progress, Capacity-Building, and Traversing Disciplines

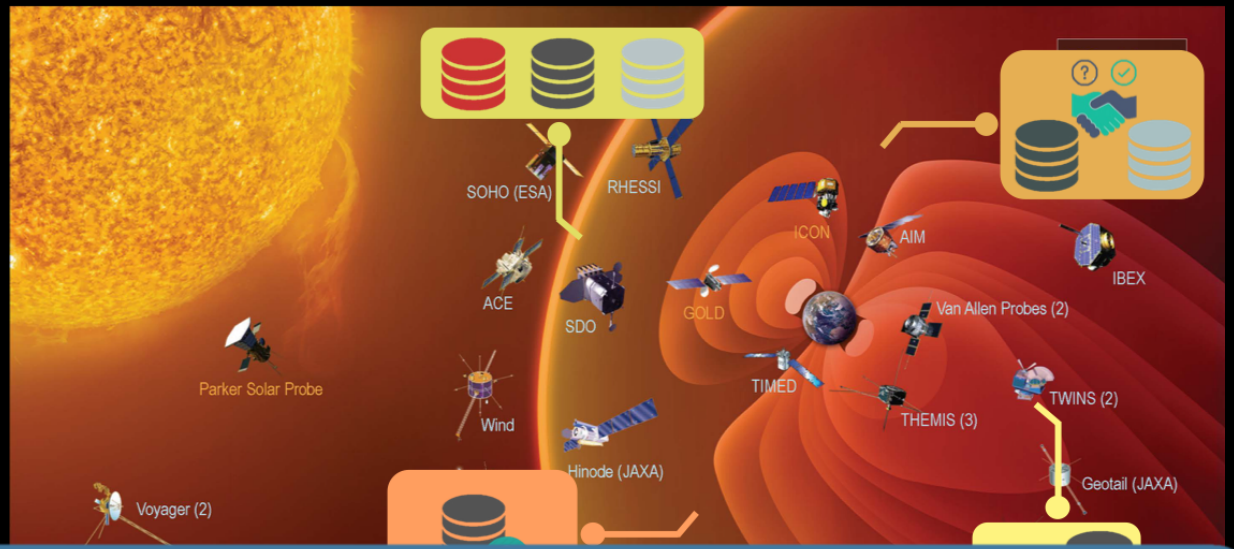
Ryan McGranaghan
Asti Bhatt
Bharat Kunduri
James Ahren
Marcin Pilinski
Enrico Camporeale



CEDAR DATA SCIENCE

Progress, Capacity-Building, and Traversing Disciplines

Ryan McGranaghan
Asti Bhatt
Bharat Kunduri
James Ahren
Marcin Pilinski
Enrico Camporeale



Data Science: Scalable architectural approaches, techniques, software and algorithms which alter the paradigm by which data are collected, managed and analyzed [and communicated].

Dan Crichton, JPL

What do we hope to accomplish?

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 through embracing newer technologies and sound digital data scholarship; and

4. Grow methodology transfer to enhance CEDAR science.

What's needed from you?

1. Think in questions
    - How might we…
    - What if…
    - How to…

2. Follow those questions to new experiments and new interactions

3. Find ways to share your thoughts and sustain conversations (e.g., Slack, Miro, future Zoom calls)

4. *Act* on what you learn and become curious about today

How are we going to do it?

INSPIRATION

IDEATION

EXPERIMENTATION

The next hour…

A glimpse at progress in CEDAR data science:

- ~~(11:10-11:18) Janet Kozyra – Using the Cloud~~
- (11:18-11:26) Hyunju Connor – Where ML may be poised to impact CEDAR
- (11:26-11:34) Yang Pan – Advances with ML
- (11:34-11:42) Jenny Yang – Information content of our observational system
- (11:42-11:50) Garima Malhotra (FDL GNSS 2019) – Breakthroughs in domain-data science collaborations
- (11:50-11:58) Roxana Bujack – Data science at scale and across disciplines

We will reconvene for the panel and breakout sessions at **12:25 PM MT**

Please remember to mute yourself to prevent background noise

In the meantime, try out this challenge problem in the software of your choice…

Each new term in the Fibonacci sequence is generated by adding the previous two terms. By starting with 1 and 2, the first 10 terms will be:

1, 2, 3, 5, 8, 13, 21, 34, 55, 89,…

By considering the terms in the Fibonacci sequence whose values do not exceed four million, find the sum of the even-valued terms.

Post your solutions in the Slack channel #cedar-data-science

Data Science “Themes” Panelists

Shea Hess-Weber - Handling large volumes of data

Karthik Venkataramani - The data analysis methods that are changing (CEDAR) science

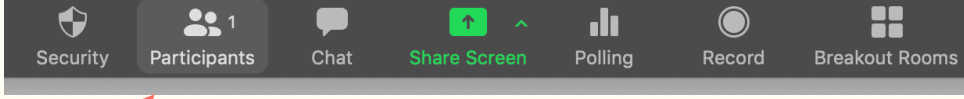
Janet Green - Creating platforms for visualizing and interacting with data

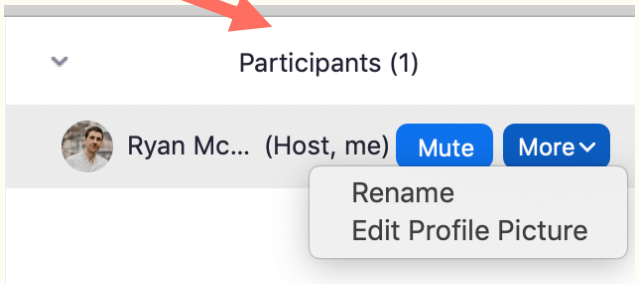
Adam Kellerman - Increasing the value of data

Abigail Azari - Broadening participation/Using data science to improve collaboration in the virtual setting

As you listen…

1. Decide which breakout room that you would like to join
2. Click on ‘Participants’ on the bottom of the zoom panel
3. Hover your cursor over your name and select ‘More’ and then ‘Rename’
4. Rename yourself to selected breakout room + name e.g., “Increasing data value - Ryan McGranaghan”





Breakout rooms:

- Data volumes
- Data analysis methods
- Creating platforms
- Increasing data value
- Broadening participation/collaboration

Report-backs

Next

Continue the question-generation and interaction

Follow-up and experiment

Breakout Session: Increasing the value of data

Most potent questions

What ongoing efforts are pushing the boundaries of data utilization?

What are the *fantastic use cases*?

Most potent questions

What are the new ideas for increasing the value of data that are creating impact in other fields?

How to leverage existing resources in other fields (The National Institute of Health, National Weather Prediction, computer science, applied mathematics) to advance our geospace data science?

Most potent questions

What are the challenges?

What are the limiting factors?

Where are we *not* effectively utilizing data?

What are we *not* considering?

Breakout Session: Collaboration and virtual interaction

Questions

- Sharing (large) datasets and corresponding metadata between and across teams: Google drive, RESTful APIs
- Analysis sharing: Jupyter notebooks/Jupyterlab, Streamlit etc
- Increasing communication between and across teams, I guess this session provides some excellent examples (Zoom, Slack, Slido etc)

Common Misconceptions

Most potent questions

What are the misconceptions around data science and machine learning?

Most potent questions

What are the misconceptions around data science and machine learning?

False **beliefs…**

Data Science = Machine Learning

Machine Learning is a black box

Machine Learning will solve all problems

Machine Learning will replace physics-based methods

The difficult part of Machine Learning is model training

Most potent questions

How do we better communicate machine learning and data science?

Most potent questions

What is required to increase adoption for data science in the physical sciences?

What is inhibiting adoption?