Developing Smarter Techniques to Deal with the Heliophysics Science Data Flood

Jon Vandegriff, Brent Smith, Kiley Yeakel, Sarah Vines, George Ho, George Clark, *(Johns Hopkins University Applied Physics Lab)*, Robert Candey *(Goddard Space Flight Center)*

As instrumentation improves, future generation Heliophysics science measurements will inevitably capture much larger data volumes. Instruments are already capable of producing significantly more data than can be sent back to Earth, even for Earth-orbiting missions. The Magnetospheric Multiscale Mission (MMS) only sends back 4% of its burst mode data (Galvez, 2019). Outer heliosphere missions, such as an interstellar probe (Brandt, 2019), will face severe restrictions on downlink data volume.

With increased data production in space, and in spite of telemetry constraints, there will be a dramatic increase in downlinked science data to be analyzed. Already, missions like the Solar Dynamics Observatory (SDO), are producing Petabyte-scale archives (Galvez, 2019). A typical event study using the highest resolution SDO data available data can require 4 months of download time (Angryk, 2018), or else require access to the computing facility where the data is stored (Barnes, 2019). Figure 1 shows data volume growth at the two primary Heliophysics archives, the Space Physics Data Facility (SPDF), and the Solar Data Analysis Center (SDAC). Projections to 2025 show archive volumes of 997 TB for SPDF and nearly 30 PB for SDAC. Simulations, another important Heliophysics analysis tool, are also producing large data volumes (~TB or more per execution) that require new techniques to analyze.

Because of the increase in both on-board data production and on-the-ground archives, dramatically smarter science data processing and analysis techniques are going to be needed. The two primary areas of need emerge: **1. smart filters on-board** future spacecraft that capture and preserve truly new measurements while filtering out unnecessary data. **2. automated, intelligent procedures for archival and model analysis on the ground** because science datasets that are so large they can no longer be fully examined by human attention. Both needs require adopting emerging techniques currently considered high risk, but also viewed as ultimately necessary, because otherwise the data becomes stuck behind bottlenecks of telemetry constraints (for on-board data) and human attention (for archived science data).

For improving on-board data selection, promising advances are being made in data filtering through the use of Artificial Intelligence (AI), or the subset of techniques known as Machine Learning (ML). Many scientists view these techniques with skepticism, and few have reached a Technology Readiness Level to be included in a competitive instrument proposal. The real need is for something more than just the latest AI approach. A well-engineering, science-centric, system-level approach that incorporates well-understood, first-principles based filtering (traditional approaches) with fast algorithms for event finding or pattern recognition (AI or ML approaches) should be developed incrementally to evolve risk into trust, allowing the community to explore what works. This kind of live, permanent science filtering is already happening in other fields, such as High Energy Physics and Radio Astronomy, where the full raw data stream can be orders of magnitude too large to retain. In these fields, a community process leads to the definition and approval of filters to decide how to reduce the data.

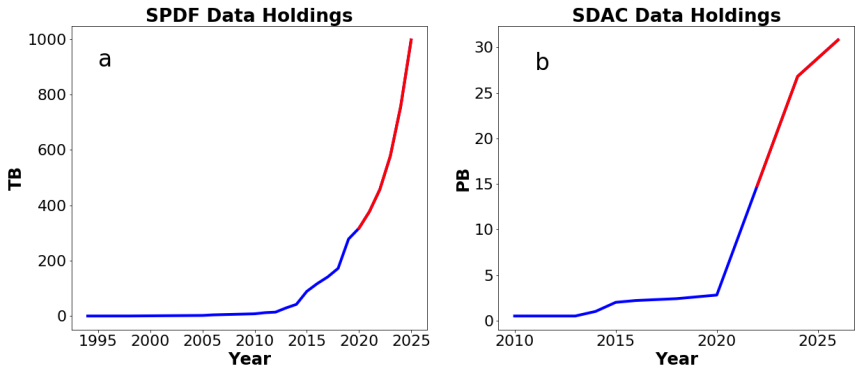

Figure 1: Data volumes for a) SPDF and b) SDAC as a function of time. Years 2021and beyond shown in red are projections based on known incoming data sources.

With respect to automated data analysis of large-scale data sets on the ground, many ML approaches are already being tried, and so the barriers to using more advanced techniques end up being stubborn infrastructure issues such as data formatting and standardization. While progress has been made recently in standardizing Heliophysics data formats (Roberts, 2016) and access mechanisms (Vandegriff, 2020), the transition to large scale data stores will require new formats (Arendt, 2018). Because cloud-based storage differs from the file-based storage used at existing archives, new standards for cloud storage and analysis should be developed ahead of a transition to a new analysis paradigm, otherwise history will repeat itself as many diverse, incompatible mechanisms are created based on project-specific interests. The stakes are higher for "big data" collections, which are often now referred to as "immovable" because once they land in a particular archive, it is impractical to ever relocate or copy them. Another barrier to automated analyses is the skillset needed. It is now very difficult for one scientist to learn the emerging techniques for large-scale data processing in addition to the science domain knowledge that is the essential core of the analysis process. Interdisciplinary teams are required, and building these teams takes time and intentionality, i.e., funding for collaboration.

The following steps will help prepare the long-term Heliophysics instrument and mission pipeline for maximizing the science return from the expected flood of larger than Exabyte-scale data that will land on our doorstep in the coming decades.

1. Establish explicit ways for high-risk on-board data filtering techniques to advance to higher Technology Readiness Levels (TRLs). Try risky techniques using a retrospective approach on older mission data (what would AI have found in this dataset?), and then migrate this to live missions that are in an extended mission phase. This has been done in Planetary Science for the Mars Science Lab (Francis, 2018). Technology demonstration missions and cheaper cube-sat missions would provide opportunities to further mature and prove out such techniques.
2. Adopt community processes similar to High Energy Physics whereby the science validity of a science filtering mechanism is scrutinized and approved.
3. Encourage instrument and mission architectures that allow for substantial updates later in the mission such that a data filtering system could be installed in extended mission operations. Reviews of science filtering algorithms could be done via peer review or as part of existing, required mission preparation reviews. Explore mission concepts that include increases in computing capacity on-board to see what science benefits could result.

4. Leverage techniques across all NASA Science Mission Directorate (SMD) disciplines where similar techniques are being deployed. Many of the techniques being used will be recognizably similar: change and anomaly detection, hyper-spectral image analysis, structure tracking and evolution in images, time-series event detection. A cross-disciplinary AI/ML program element supported by all the SMD areas and from which each SMD discipline could draw expertise would help reduce the learning curve within each discipline.
5. Work now to establish data standards that will survive the transition to "big data" systems. A cross-SMD effort could be beneficial here too.

Finally, smarter on-board data processing would have direct benefits for strengthening research-to-operations goals within Heliophysics. Given a telemetry bottleneck, the most complete data for an operational decision is the full instrument output that is only available on the spacecraft. A science-savvy approach to on-board processing and analysis could greatly improve real-time alert systems for space weather. For the inverse problem of operations-to-research, more intelligent and efficient approaches to combing through large volumes of space weather data, model output, and science archives will be very useful in developing new models and forecast techniques. Space weather forecasting is indeed one area where ML has already made significant contributions (Camporeale, 2019).

Heliophysics studies in 2050 will be made by high throughput instruments that require intelligent processing on-board and on the ground in order to take full advantage of enormous data volumes. A pipeline that transitions emerging developments in AI and ML into trusted, science-focused techniques will help maximize the science return from these future measurements.

References

Angryk (2018) Solar Data Mining: Baby Steps, EarthCube Research Coordination Workshop, NJIT, Nov 14-16, Newark, NJ

Arendt et al (2018) Pangeo: Community tools for analysis of Earth Science Data in the Cloud, AGU Fall Meeting, IN54A-05, December 10-14

Barnes et al (2019) The Sun at Scale: Interactive Analysis of High Resolution EUV Imaging Data on HPC Platforms with Dask, AGU Fall Meeting, SH41C-3317, San Francisco, CA, December 9-13

Brandt et al (2019) Humanity's First Explicit Step in Reaching Another Star: The Interstellar Probe Mission, JBIS, Vol 72, 202–212

Camporeale (2019). The challenge of machine learning in Space Weather: Nowcasting and forecasting. Space Weather, 17, 1166–1207. https://doi. org/10.1029/2018SW002061

Francis et al (2018) Incorporating AEGIS autonomous science into Mars Science Laboratory rover mission operations, AIAA SpaceOps Conference, 2018-2576, Marseille, France, May 28 - June 1, https://doi.org/10.2514/6.2018-2576

Galvez et al (2019) ApJS, 242, 7, https://doi.org/10.3847/1538-4365/ab1005

Roberts et al (2016) The NASA Heliophysics Science Data Management Policy, https://hpde.gsfc.nasa.gov/Heliophysics_Data_Policy_v1.2_2016Oct04.html

Vandegriff et al (2019) Interoperability for Heliophysics and Planetary Time Series Data via HAPI, AGU Fall Meeting, IN11E-0698, San Francisco, CA, December 9-13, also see https://github.com/hapi-server/data-specification