日球物理数据分析科学平台

Monica G. Bobra（斯坦福大学）、Will T. Barnes（常驻海军研究实验室的 NRC 博士后）、Thomas Y. Chen（哥伦比亚大学）、Mark C. M. Cheung（洛克希德·马丁太阳与天体物理实验室）、Laura A. Hayes（美国国家航空航天局戈达德太空飞行中心）、Jack Ireland（美国国家航空航天局戈达德太空飞行中心）、Miho Janvier（空间天体物理研究所）、Michael S. F. Kirk（美国国家航空航天局戈达德太空飞行中心和 ASTRA llc.）、James P. Mason（科罗拉多大学博尔德分校）、Stuart J. Mumford（谢菲尔德大学和 Aperio Software）、Paul J. Wright（斯坦福大学）

引言：当今日球物理学数据分析面临什么问题?

2020年，我们生活在一个数据丰富的时代。用于研究日球层的美国国家航空航天局仪器所采集的数据量不断增加，采集速度也前所未有。例如，太阳动力学观测台（SDO）的任务数据集目前总量已达 19 拍字节。日球物理学界还会开展协同观测活动，即许多仪器在同一时段观测同一目标，从而获得种类繁多的数据——图像数据、光谱数据、时间序列数据，以及关于局部粒子和场条件的原位数据。日球物理学天然是一门多信使学科，许多研究都依赖于将这些不同的数据集结合起来使用。

然而，科学家难以自由探索这些数据。网络速度和有限的存储空间使获取大型数据集变得困难。计算能力不足，使得高效分析大型数据集变得困难。尽管加速且可扩展的计算资源（例如图形处理器、云计算）已广泛普及，大多数日球物理学家仍使用笔记本电脑或台式电脑来运行数据分析工作流程，而这些设备的内存和磁盘空间都十分有限。这些障碍使整个领域无法充分发挥科学发现和科学投资回报的潜力。

为了利用这些数据取得新的科学突破，我们需要采用现代化的科学工作流程。在当今的主流范式中，日球物理学家会将数据下载到本地计算机上进行分析。这种工作流程限制了依赖大型数据集的科学研究范围。科学家只能开展符合本地网络速度、存储能力、软件工具和计算能力限制的研究。

为了支持规模大得多的研究，我们建议采用另一种范式。在这种工作流程中，日球物理学家在外部计算机上开展科学研究，该计算机配备多个任务数据集、充足的计算能力和软件工具。这些开放访问的科学平台目前已应用于天体物理学（例如 Bauer等，2019年）和地球科学（例如 Robinson等，2019年）领域，使科学家能够快速分析拍字节级的数据。日球物理学家不再受网络速度、存储、软件和计算资源的限制，便可以自由开展引人入胜的科学研究，而这些研究过去实际上难以实现。

面向其他科学界的科学平台

其他科学界群体已经认识到这种方法的优势。例如，

天文学家可以使用美国国家科学基金会资助的科学平台，分析来自薇拉·鲁宾天文台的巡天数据。该平台将巡天数据与美国国家超级计算应用中心的高性能计算资源部署在一起（Dubois-Felsmann 等，2019）。实际操作中，这意味着天文学家可以打开终端窗口或 Jupyter Lab，登录外部计算平台，编写代码，并在任意巡天数据上运行代码。其他类似案例也存在，例如美国国家科学基金会资助的科学平台——美国国家光学天文台数据实验室（Fitzpatrick 等，2014；Taghizadeh-Popp 等，2020；Thomas 等，2020）。

Pangeo 数据科学平台（Odaka 等，2020年）由包括美国国家科学基金会、美国国家航空航天局、美国国家大气研究中心、斯隆基金会和英国气象局在内的多个机构资助，彻底改变了许多地球科学家对数据科学工作流的看法。地球科学家利用 Pangeo 分析来自各种空间天文台的拍字节级数据。

科学平台的四项必要特征

虽然各个平台的技术设计要素各不相同，但都包含四项基本特征：数据存储、计算能力、软件工具和开放访问。科学数据应采用标准且灵活的格式，并进行版本控制，存储在灵活数据库中。与数据存放在同一位置的计算资源应支持在中央处理器和图形处理器上进行并行处理。科学平台还应提供相应环境，以便轻松安装和使用开源、开放开发且经过版本控制的科学软件。最后，平台应为整个社群提供可公开访问的协作工作空间。

日球物理学科学平台

日球物理学界将从科学平台中获益良多。日球物理观测站产生了海量数据集，但这些数据尚未得到充分利用。根据 SunPy 项目对太阳物理学界开展的一项调查（Bobra 等，2020年），82% 的太阳物理学家使用观测数据开展研究。

尽管如此，大多数日球物理学家仍无法使用计算设施。在同一项调查中，SunPy 项目发现，太阳物理学界有 14% 的人使用本地或区域集群，9% 使用图形处理器，5% 使用商业云开展研究。约三分之一的业内人士只使用笔记本电脑或台式电脑。这也意味着，尽管大规模并行计算是提升计算性能的最大机遇，大多数业内人士却没有充分利用它（Robinson 等，2020）。

开源科学软件已经提供了功能强大且易于使用的工具，可扩展地提升计算性能（例如 Dask；Rocklin 等，2015）。此外，将开源、版本控制的仪器校准软件（如 AIAPy（Barnes 等，2020））与开放获取、版本控制的数据结合起来，有助于学界创建经过校准、可复现且可共享的数据集。

如今，日球物理学界尚不存在科学平台。太阳科学平台原型（Barnes 等，2019年）提供了对 SDO 数据的交互式访问，并

借助美国国家航空航天局昴宿星团超级计算机实现高性能计算，展示了用户如何快速分析大型数据集。此类科学平台将为太阳与空间物理学的数据分析开启可扩展、交互式超级计算的新时代。

建议

我们建议美国国家航空航天局维持并资助科学平台，以支持交互式、可扩展的数据分析，从而最大限度地发挥太空仪器采集数据的科学回报[1]。为实现这一愿景，我们建议在未来 10 年内达成以下一系列短期目标：(1) 支持科学软件的协调式开放开发，兼顾交互性与可扩展性；(2) 为社区提供有关高性能计算和云计算数据分析的教育与培训；(3) 建立一种资助模式，使研究人员能够使用资助经费向商业云服务商购买计算时间。

我们还建议在未来 20 年内达成以下一系列长期目标：(1) 在美国国家航空航天局高性能计算中心（例如戈达德的 ADAPT、艾姆斯的昴宿星团）建立专用基础设施，以支持交互式、可扩展的数据分析；(2) 将当前和最终的任务数据档案库与这类专用基础设施设在同一地点，确保在任务生命周期结束很久之后，仍能充分发挥这些数据的科学潜力。这些建议将共同帮助科学家加快研究工作流、开展可重复研究，并最大限度地发挥美国国家航空航天局数据集的科学回报。

参考文献

Barnes W. T. 等。Zenodo 2020年，doi：10.5281/zenodo.401698。Barnes W. T. 等。AGU秋季会议 2019，bibcode：2019AGUFMSH41C3317B。Bauer A. E. 等。2019年，arxiv：1905.05116。Bobra M. G. 等。《太阳物理学》295 2020年，doi：10.1007/s11207-020-01622-2。Dubois-Felsmann G. 等。2019年，网址。Fitzpatrick M. J. 等。国际光学工程学会 2014，doi：10.1117/12.2057445。Momcheva I. 等。2015年，arxiv：1507.03989。美国国家科学院、工程院和医学院。2020年，doi：10.17226/25668。Odaka T. E. 等。《计算机与信息科学通讯》1190 2020年，doi：10.1007/978-3-030-44728-1_12。Robinson N. H. 等。2019年，arxiv：1908.03356。Rocklin M. 第14届 Python 科学会议论文集。126，2015年，网址。Rollin T. 等。Zenodo 2020年，doi：10.25080/Majora-342d178e-01f。Taghizadeh-Popp M. 等。2020年，arxiv：2001.08619。

[1] 这一愿景以及这些短期和长期目标与2020年美国国家科学院报告《太阳与空间物理学*2013*年十年调查实施进展：中期评估》中的建议3.2.4一致。该建议指出：“美国国家航空航天局和美国国家科学基金会应通过以下方式，最大限度地提高大型复杂数据集的科学回报：(1) 提供现代统计学和计算技术方面的培训机会；(2) 提供科学平台，使用通用标准存储、检索和处理数据；(3) 提供支持跨学科合作的资助机会；(4) 推动开源软件的开发。”