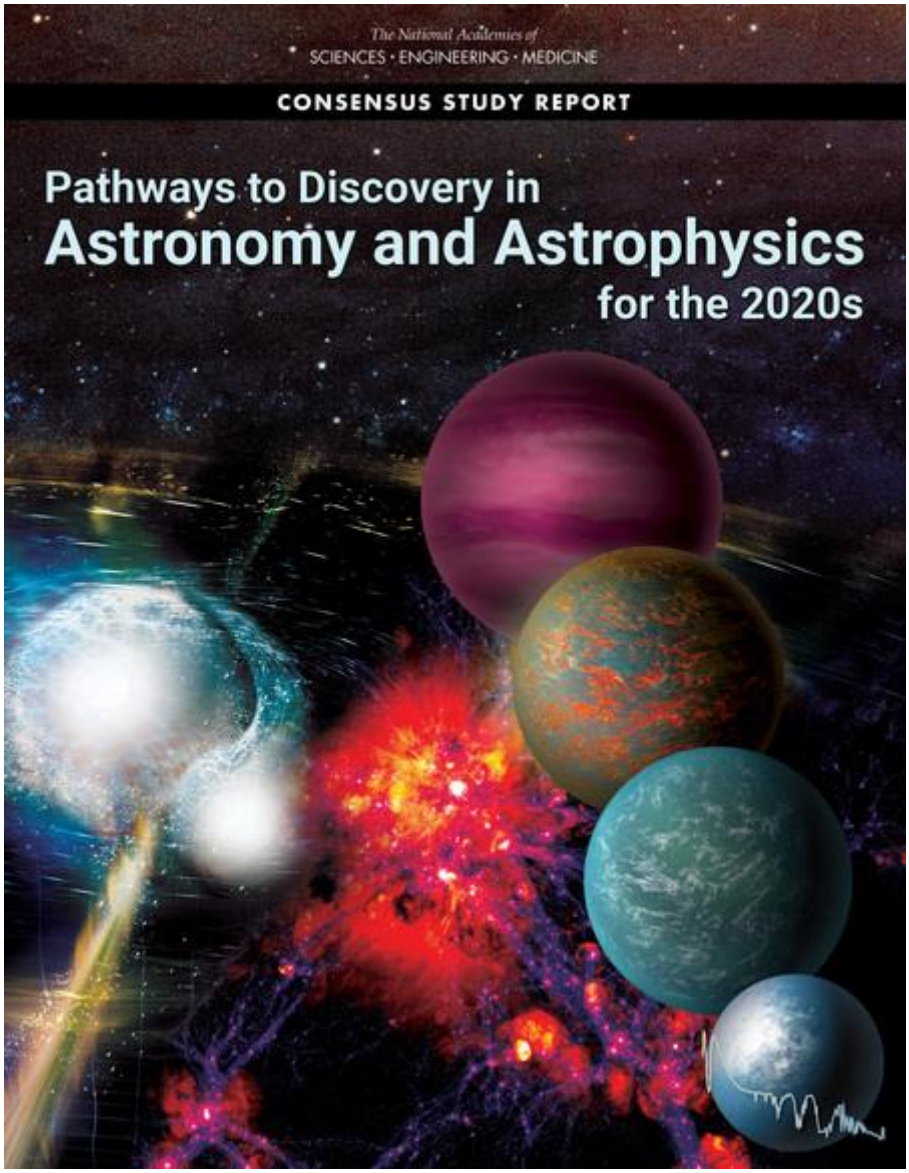
天文信息学

@

《2020s 年代天文学与天体物理学的发现之路》



张震、王培培、吴莹、杨嘉宁、于洋、张琦乾 译

2021 年 11 月

4.5 数据基础

在过去 150 年的大部分时间里，大多数天文观测都是由个人或机构进行的，保存在照相底片或数据磁带上。到 2020 年，这一局面已经完全改变。现代所有数据都是数字化的，其中很大一部分都存档在公众可以访问的在线数据库中。这些档案的科学重要性和影响至关重要。例如，在过去 15 年中，使用了哈勃太空望远镜数据档案的出版物的数量超过了最初提出提案的团队的出版物数量（图 4.7），而且引用数量相当，[9] 其他主要望远镜设施也出现了类似的趋势。经验证据表明，在组织良好的档案中管理科学数据可以实现多重用途，并延长数据的使用寿命（图 4.8）。

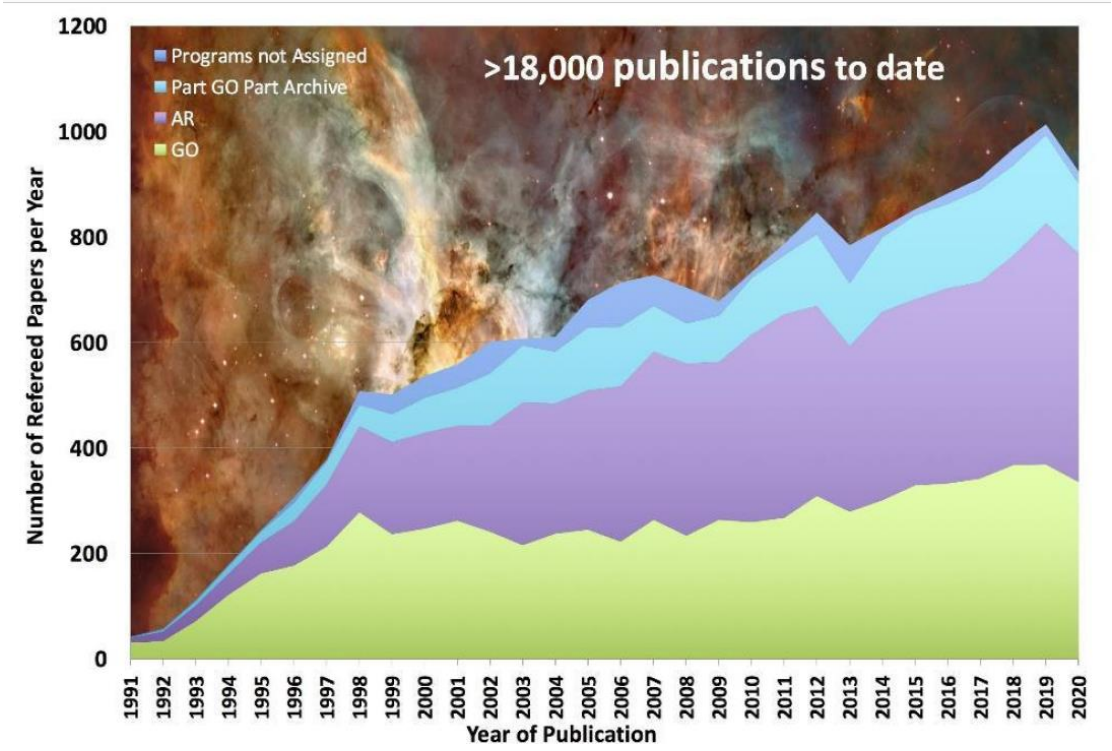


图 4.7 哈勃太空望远镜观测所得出版物的历史。绿色曲线显示了来自最初提出观测提案（GO）的团队的论文被引的趋势；紫色表示在最初的提案团队和论文作者之间没有重叠的论文，即使用纯数据档案（AR）研究；水蓝色表示作者同时包含 GO 和档案研究人员的论文。第四类是指无法分配到其他类别中的论文。，使用数据档案的论文，作为开放档案和数据处理管线的产物，从哈勃建立之初起，产生速度就超过了 GO 论文。资料来源：R.Osten，根据 STScI 提供的数据：https://archive.stsci.edu/hst/bibliography/pubstat.html.

天文学正进入这场数据革命的第二次浪潮，越来越多的巡天设施从一开始就主要致力于制作档案数据集，这些数据集随后被成千上万的用户共享，用于无数个独立的科学项目。二十年来，斯隆数字巡天一直是这种新的巡天模式的开创性先驱。在太空中，美国宇航局的红外天文卫星和宽视场红外巡天探测器的全天观测创建了至今仍具有持久价值的数据集。最近，欧洲航天局（ESA）的 Gaia 正在测量 10 亿颗恒星的精确位置和自行，它彻底改变了银河系和恒星天体物理学（第 2 章）。尽管它完全由欧洲建造和支持，但它的数据档案在全世界都可以公开访问，并且在第一次数据发布后的 5 年内支持了美国天文学家的数百项研究。在未来十年里，2010 年十年调查中最优先的地面和空间项目 Vera Rubin 和 Nancy Grace Roman 天文台将分别提供丰富的数据集，有望彻底改变时域天文学，并有望在广泛的天体

物理学科中获得突破性发现。到 2030 年底，它们还将为所有天文台和任务带来前所未有的 500 PB（5 亿 GB）数据量，比人类历史上收集的天文数据多出几个数量级。当它与来自其他可用性日益增强的中等规模设施的数据相结合时，观测研究事业的本质正在发生改变。简言之，天文学的进步需要做好从目标观测到下一阶段的大型公共数据集的过渡的准备，以便最大限度地提高当前和未来设施的科学产出。

在巡天和大数据量日益重要的同时，计算天体物理学的一场相关革命也正在进行。数值模拟在行星、恒星、星际云和等离子体、星系以及宇宙本身的物理建模中扮演着越来越重要的作用。数值模拟也已经成为许多理论天体物理学家的基本技能。这些模拟的输出是一种宝贵的资源，但目前很少公开，并且将包含大量数据。尽管有许多供理论家和建模者使用公开可用的代码，但编写或维护代码的人却少得多。

数据革命也改变了许多天文学家进行研究的主要方式。许多观测天文学家很少亲自使用望远镜进行观测，而是花费大量时间开发对大型在线数据集进行复杂的分析的方法。今后，用于处理数据和创建结果的算法变得与基础观测值一样重要。共享软件的机制，如通过 Github 进行代码共享和修改，并通过 Jupyter notebooks 提供带工作示例的教程，可实现可复制性，并带来该领域可访问性的进一步提升和领域激励的提高。将论文与档案中包含的数据直接联系起来，也加强了科学结果与输入观测之间的联系。

这些革命当然不是天文学和天体物理学所独有的，而且跨越了许多领域。“利用数据革命”是 NSF 进行重大资助的十大举措（“十大理念”）之一。NASA 最近组建了一个“大数据工作组”[11]，并发布了《突破性科学的数据管理和计算策略（2019-2024）》[12]。本节提出的建立天文学和天体物理学数据基础的建议与这些工作很好地吻合。

4.5.1 数据的归档，管理，存储

美国航空航天局和美国国家科学基金会很早就认识到对这些丰富的数据集进行归档、管理、使用和分析的重要性，并且有多项计划支持这些领域的研究。美国航空航天局提出针对特定任务计划(图 4.8)，数据归档中心计划，以及个人项目研究计划，这些计划中包含的具体项目例如天体物理数据分析项目(ADAP)和系外行星研究项目(XRP)。美国国家科学基金会也支持美国天文台对天文数据管理进行研究，用于管理个人研究者项目中研究的数据以及管理个人研究出的成果。但数量庞大的数据具有更多变的归档方式和可访问性。因此我们遇到的问题是，目前现有的这些计划是否适合 2020 年及以后继续使用。所以在承认当前项目获得了许多研究成果的同时，这项为期十年的调查将重点关注那些适度的投资可以产生重大科学回报的领域。

事实上，望远镜收集到的所有数据都是一种宝贵的资源，都有可能为未来的发现做出贡献。而这种价值可以通过投资基础设施来实现最大化，因为基础设施能够以统一的方式收集和减少这些数据，并且存档数据易于检索，最终目标是使数据公开可用。由于空间和地面上仪器(例如积分场光谱仪和多目标光谱仪)的日益复杂，对数据处理的需求也日益增加。联合分析来自不同设施和波长的观测，以及利用相关科学平台工具进行复杂归档的重要性，将在未来十年急剧增长。一个主要的例子是在未来的十年中，暗能量和其他参数的宇宙学约束的测量，将严重依赖欧洲航天局、罗马和鲁宾天文台的数据联合处理和分析。正如前几章所详述的，对多波长、多信使和时域分析的巨大兴趣将在未来十年带来新的挑战，就像使用前所未有的大量数据开展任何科学项目一样。

虽然目前还不清楚这些数据能为科学研究起到什么样的作用，但总的趋势是数据在科学研究中的作用越来越大。在刚刚结束的十年中，我们看到了地面和太空中数据存档能力的扩

展(图 4.7、4.8、4.9、4.10、4.11)，并且预计在未来十年中，这一能力只会增长。由于太空设施距离地面遥远，所以从一开始就要求有效的数据存储。也许并不令人惊讶的是，管理良好的数据档案几乎适用于美国航空航天局的所有主要任务。这些数据具有长期的影响:哈勃太空望远镜早期采集的数据在最初采集后的近 30 年里仍能在论文中找到有效的用途。钱德拉 X 射线天文台早期保存的 70%的数据出现在四种以上的论文期刊中(图 4.8)。

而地面设施的情况要复杂得多。国际 ALMA 天文台和欧洲南方天文台(ESO)建造和运行的大型设施，以及斯隆数字巡天、全景巡天望远镜和快速响应系统(Pan-STARRS)等调查，建立了可与空间天文台质量媲美的数据档案。这些数据档案对多项科学研究做出了重大贡献。图 4.9 和 4.10 分别详细说明了 ESO 望远镜和 ALMA 天文台逐渐增加的数据档案;(注解 14)在此研究背景下，大约有三分之一的论文是使用至少一项数据产生的。与此同时，阿尔玛天文台在研究如何简化数据和校准数据存储，目前 95%的数据校准和成像都是由阿尔玛天文台完成的;阿尔玛天文台的研究成果除了加强了数据的实用性外,还大大减小了年轻科研人员进行科研工作的难度,并扩大了使用阿尔玛天文台数据的用户基础。美国国家科学基金会的国家光学红外天文研究实验室主办了一个天文学数据实验室，该实验室作为数据归档和传播美国双子座天文台观测数据的中央枢纽，并且该实验室研究重点是对大型数据的分析和开发数据处理工具。而美国航空航天局资助了凯克天文台，受资助凯克天文台负责的对美国航空航天局系外行星科学研究所的数据进行管理与存储。然而，以上的例子只是例外，而不是普遍现象。有几个因素可以解释这种情况。如很少有个人愿意出足够的资金支持研究简化数据和校准数据存储。并且美国地面天文台的资金也有限，很难给出足够的钱去研究简化数据和存储数据。而对于像使用复杂数据处理的甚大天线阵（JVLA）或超长基线阵列（VLBA），尽管尽了最大努力，但现有的数据存储和数据管理技术可能并不适用于所有数据(图 4.11 显示了甚大天线阵数据使用的增长趋势)。虽然一些机构将其数据存入公共数据库，但这些资源往往难以利用，最终的结果是失去了这些数据的价值。天文台的数据未得到充分利用主要是因为，科学家们希望可以很容易地访问数据和探索数据，而不是先花几个月时间去简化数据和校准数据存储。

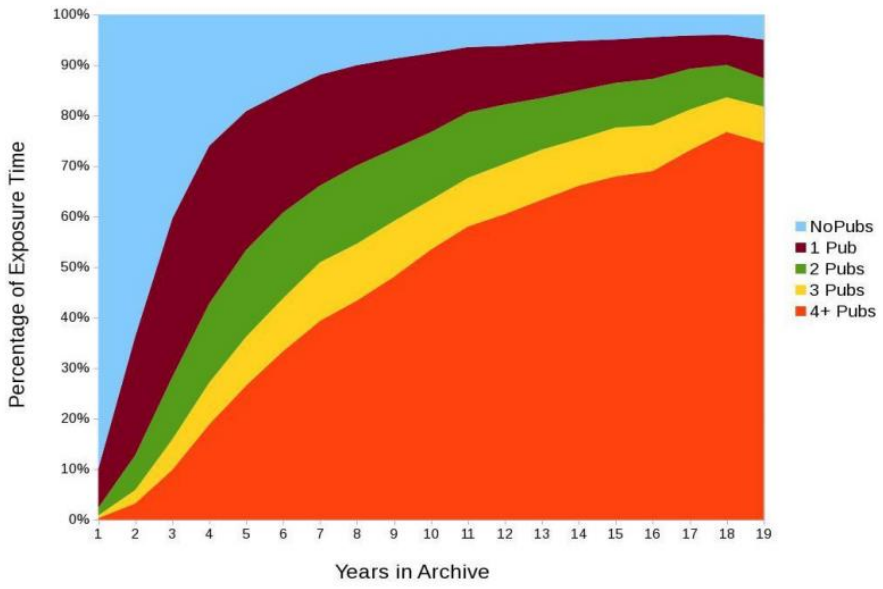


图 4.8 把从钱德拉 X 射线中获取的数据作为时间的函数，纵轴表示百分比。该图展示了整理的非常好数据对科学研究的影响。这里纵轴的数据被量化为暴露时间。在这里，70%的最古

老的数据集有四篇及以上使用该数据的出版物。资料来源:钱德拉数据运营团队。

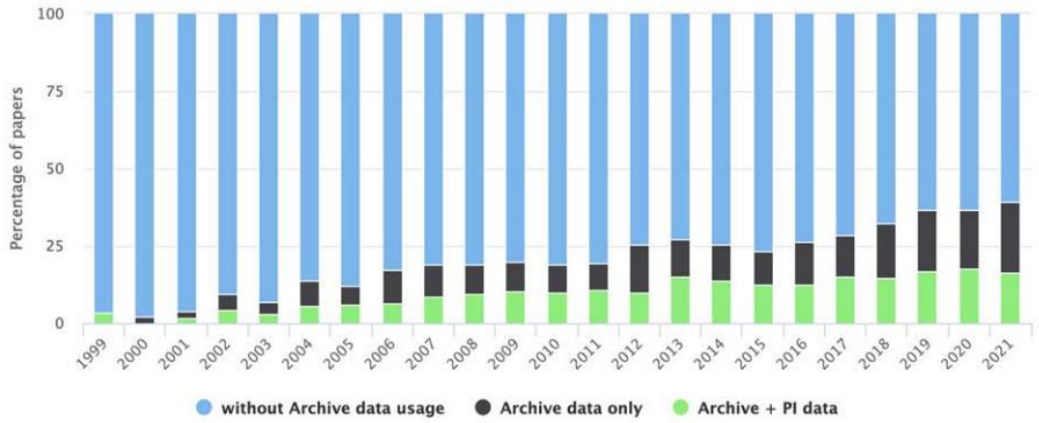


图 4.9 表示使用欧洲南方天文台数据的论文增长图，占当年发表论文总数的百分比。蓝色的表示论文作者与其他科研人员之间有重叠的论文，黑色的表示没有重叠(即纯归档使用)，绿色是中间值。在可获得统计数据的上一个完整的年份(2020 年)，占总数的三分之一以上的论文使用了某种格式的归档数据。资料来源:从 ESO 检索，由 ESO 图书馆的参考书目中提供。http://telbib.eso.org/index.php?boolany=or&boolaut=or&boolti=or&yearfrom=1996&yearto=2021&boolins=or&booltel=or&site=Paranal&search=Search。

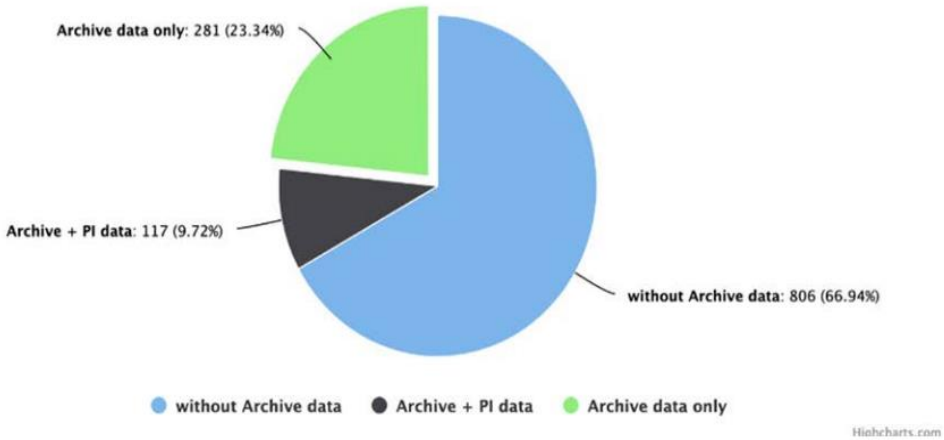


图 4.10 表示阿塔卡玛大型毫米波天线阵报告的引用论文中使用毫米/亚毫米阵列(ALMA)归档数据的百分比。大约有三分之一的论文使用了毫米/亚毫米阵列(ALMA)归档数据，使用方式为单独使用或结合其他数据使用。资料来源:从 ESO 检索，由 ESO 图书馆的参考书目中提供。
http://telbib.eso.org/statistics/archive.php?boolany=or&boolaut=or&boolti=or&yearfrom=2010&yearto=2021&boolins=or&telescope[]=%22ALMA%22&booltel=or&site=Chajnantor&fl=telescope,datastatus&stats=arc&query_stats=year%3A%5B2010+TO+2021%5D+and+site%3AChajnantor+and+%28telescope%3A%22ALMA%22%29。

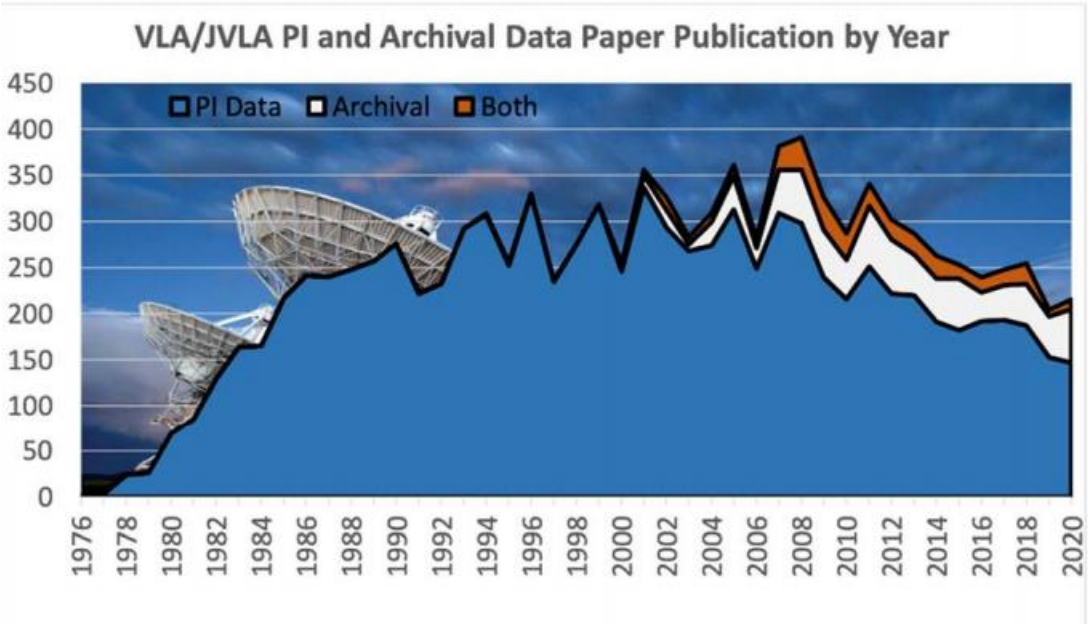


图 4.11 甚大天线阵（NRAO's Very Large Array/Jansky Very Large Array，简称 VLA/JVLA）的文章详细介绍了 PI 使用和引用文献中 VLA/JVLA 归档数据使用的演变。该文件不包括来自调查的论文，如美国国家射电天文台甚大天线阵天空调查（NRAO VLA Sky Survey，简称 NVSS）、甚大天线阵天空调查（Very Large Array Sky Survey，简称 VLASS）。
资料来源：https://public.nrao.edu/news/the-very-large-array-astronomical-shapeshifter/（背景图），R. Osten 基于 NRAO 的 L. Utley 的数据（主图）。

这些尚未被归档的观测数据，可以被视为从美国地基系统中获取更多科学信息的巨大机会。通过适当的战略规划和适度的财政投资，创建和归档可供科学使用的数据产品可以在地基设施的科学领域提供数倍的回报。科学家们有限的时间和资金支持可以完全集中在分析和发现上——望远镜观测的影响可以跨越数十年岁月，因为光子在科学方面的再利用方案，可能是在收集光子的时候无法想象的。提升观测档案的访问便利程度和质量也可以作为扩大专业参与度的有力手段，因为它们为任何可以上网的人（甚至是通过公民科学倡议的公众）带来了最先进的数据，同时最大限度地降低了进入活跃的研究社区的门槛。这种通过档案访问实现的科学民主化将在未来十年继续下去。

发现：正如空间任务证明的那样，在欧洲南方天文台（European Southern Observatory，简称 ESO）和阿塔卡马大型毫米波/亚毫米波天线阵（Atacama Large Millimeter/submillimeter Array，简称 ALMA）的存档工作的支持下，来自地基望远镜的原始数据和简化数据可以极大地增加其本身的科学影响，如果有通道可以提供处理后的数据，效果会进一步提升。

随着薇拉·鲁宾天文台等大型观测设施的出现，多波段和多信使天体物理学以及时域天文学在这十年中的发展，对涉及多个档案的数据发现和分析的需求更大了，这也促使我们对这些档案进行协调处理。扶持基金会小组和委员会考虑了如何最佳地满足这一需求。小组建议，建立一个跨机构的伞式组织，称为天文数据归档系统（Astronomical Data Archiving System，简称 ADAS），用以协调现有天文数据中心的活动，并为新的投资确定优先次序。国家虚拟天文台（National Virtual Observatory，简称 NVO）在 2007 年至 2014 年间的工作

目标类似于加强档案的互操作性，但其结构和实施方式与此处提议的 ADAS 有显著差异。通过与类似的国际虚拟天文台组织合作，NVO 取得了一系列重要成果，包括创建了一套通用的数据格式和元数据标准，实现了第一代数据检索和探索工具，加强了美国和世界各地许多单独的数据中心之间的交流。然而，2014 年，美国国家科学基金会（National Science Foundation，简称 NSF）的支持结束了，建设任何此类新的组织都需要从 NVO 的经验中获得的教训，以及在过去 20 年间许多独立运作的档案中心的经验。通过为档案科学家和软件开发人员提供灵活和稳定的职业道路，继而保存现有数据中心的专业知识和资源是非常重要的。同样也很重要的，是提供一个集中的渠道，让美国天文学界对数据归档工作的优先次序提出意见，从而使档案中心的工作与社区的科学需求保持一致。扶持基金会小组所设想的系统还可以解决软件开发、高性能/高吞吐量计算、理论模拟数据的归档和整理、相关领域的社区培训等跨机构战略规划。

创建有效档案的一个重要组成部分是与跨机构和国际档案服务机构进行协调，以发展最佳实践和互操作性。虽然国际虚拟天文台联盟仍在继续运作，但由于美国的国家协调工作的欠缺，各个国家和国际资助机构以及制作和存档天文数据的机构之间的有效沟通遭遇了阻碍。对此进行改善的工作将来自于一个端到端的方法，其考虑到数据从仪器到档案、到分析和出版的整个流程。提高科学数据产品和有效档案普及的最好途径是携手并进——公开代码有助于最大限度地减少冗余，鼓励采用通用标准，促进使用多个数据集的应用。调查委员会赞同扶持基金会小组提出的 ADAS 目标的重要性，但认为这项工作的范围和形式在细节方面需要进一步研究，研究由美国国家航空航天局（National Aeronautics and Space Administration，简称 NASA）和 NSF 牵头，美国能源部（Department of Energy，简称 DOE）亦可能参与。

建议：NASA 和国家科学基金会应该探索各种机制，以改善美国档案中心之间的协调，并建立一个集中的纽带，与国际档案社区进行互动。这项工作的目标应由天文学界广泛的科学需求决定。

美国的地基光学和红外系统分布在公共和私人设施中，因此需要特别考虑。30 年来，这一领域的许多需求是由图像还原和分析设施（Image Reduction and Analysis Facility，简称 IRAF）提供的，这是一个由多个联邦机构资助的免费软件系统。它最初是由国家光学和红外天文台（National Optical and Infrared Observatories，简称 NOAO）在 20 世纪 80 年代开发和维护的，由于缺乏对软件进行现代化改造的资金，它演变成了 GitHub 上的一个社区支持的平台。这种以社区为基础的方法自此成为在地基光学和红外天文学中提供通用数据通道和分析工具的主流模式。主要的例子包括 Python 光谱还原通道（Python Spectroscopic Reduction Pipeline，简称 PypeIt），它已被至少 11 个天文台的多个望远镜和仪器所采用，还有 Astropy 项目收集的大量基于 Python 的应用软件套件。

尽管这些基于社区的努力在填补最新通道和软件的空白方面做了很多工作，但因为缺乏足够和可靠的资金以及稳定的贡献者队伍，有些项目可能不得不放弃他们的支持，就像 IRAF 项目最终被迫放弃一样。这些例子说明，在构建旨在支撑大多数现代天文软件的软件基础设施的背景下，正常的资助流程是如何运作不良的——任何建立这种基础设施的努力都依赖于短期的筹资效果，没有任何途径来获得长期规划和稳定所需的长期承诺。国家科学基金会可以为这些努力提供基础的支持，例如，通过要求减少和分配部分仪器设备资金的计划，来激励通道开发和数据归档，以及开放针对多个研究者或天文台的联合提案，以对现有管道的调整和运行进行资助。如果天文台愿意向公共档案馆分发其原始数据和可用于科学的数据产品，就可以向其提供用于定制和运行这些通道的资金。这种投资的成本只是已经投入这些设

施的所有资金中的一小部分，但其科学回报可能极其可观。

建议：国家科学基金会和利益相关方应该制定一个计划，解决如何设计、建设、部署、维持通道的课题，以便在所有通用的地基天文台（包括联邦政府的和私人资助的）上生成科学就绪的数据。提供资金，以确保所有通道观测数据以标准格式归档，最终供公众使用。

4.5.2 软件开发

天文学已经进入了一个新时代——对于项目的成功而言，设计和构造精良的软件同硬件一样重要。一些高度复杂的软件包括：望远镜的数据归算处理管线（例如 Astropy），数据分析包，物理过程模拟程序，例如恒星演化（例如，Modules for Experiments in Stellar Astrophysics [MESA]），N 体模拟（例如 Galaxies with Dark Matter and Gas Interact [GADGET]），流体动力学（例如 Enzo）。此外，先进的统计技术和机器学习在物理学和天文学的大量数据归算中扮演着越来越重要的角色，它们同样需要复杂的编码。未来的趋势是：很多软件包更多地被大型团队开发，而且必须能够充分利用各类硬件平台——从通用笔记本的 CPU 到使用了大量并行和图形处理单元（GPU）的多核计算集群。

尽管软件开发和开发人员对该领域的发展越来越重要，但他们没有得到足够的资金或现有体制的支持。此外，在天文学中，拥有专业的软件开发技能的人是至关重要的，但他们很可能在天文学之外有更多就业机会。事实上，整个物理学领域都是如此。对于那些选择专门开发科学软件基础设施的科学家来说，在这条路上可能很难得到有助于他们个人保留和发展的终身教职。这些职位可以由国家实验室、科学中心、天文台或大学的研究职位提供，而资助机构却认为，资助软件基础设施的提议可能更像“仪器仪表”提议，而不是标准的 PI 拨款。正如《NASA 地球和空间科学的开源软件政策选项》中所讨论的，要为软件维护和开放源代码软件项目提供资金，这些项目在过去十年中使天文科学产生了变革，在未来可能会产生更大收益。[19]

发现：软件开发已经成为天文学各个子领域的重要组成部分。然而，软件开发人员和大型软件开发工作并没有得到现有结构的充分资助或支持。

4.5.3 高性能和高吞吐量计算

从物理过程的理论模拟到复杂的数据分析，计算力在天文学和天体物理学中扮演着越来越普遍的角色。因此，专业计算设备和其在专业领域上的应用已日益成为科学进程的组成部分，因此需要在今后十年不断进行投资和培训。高性能和高吞吐量计算资源(HPC 和 HTC)在天体物理研究中发挥着越来越重要的作用(图 4.12)，前者对于理论模拟至关重要，后者则是大数据分析的关键。HPC 覆盖了计算天文学家主要的研究方向，因此推动了高性能选项的使用。专业的 HTC 目前通过云计算实现，通常这也是满足天文需求的经济有效的解决方案。然而，随着数据分析问题规模的扩大，计算成本可能会成为阻碍，基于公有云的方案会比基于私有云的方案更具成本优势。资助项目应该适应这种快速变化的交易模式，同时要保证将资金投入到云计算中，而不是更多传统计算硬件的购买。

美国能源部(United States Department of Energy，简称 DOE)和美国国家科学基金会(National Science Foundation，United States，简称 NSF)宣布计划在未来十年大幅扩展他们的 HPC/HTC 能力，美国宇航局(National Aeronautics and Space Administration，简称 NASA)则计划进行相对温和的策略来扩展相应的能力。NSF 的计算资源可以在没有 NSF 支持的情况下使用，但是 NASA 的计算资源不能在脱离 NASA 的情况下使用。正如 4.3 节强调的那

样，有必要通过跨机构的资助来打破机构之间的界限，对于每个机构来说，为横向项目提供 HPC/HTC 计算资源是极其重要的。在大型国家计算设备上开发具有竞争力的专业代码，需要同时具备计算机科学和天体物理学方面的专业知识，以及用于代码开发和测试的设施访问权限。这些要求对于无法获得这些专业知识或设施的机构的科学家是一个重大阻碍。提供培训（来自 NSF、NASA 或国家实验室）可以有效帮助营造公平的竞争环境。这种资助可以采取多种形式，例如通过对个人的小额资助、对培训研讨会或学校的资助，或通过 NSF 和/或 NASA 中心提供的培训机会。

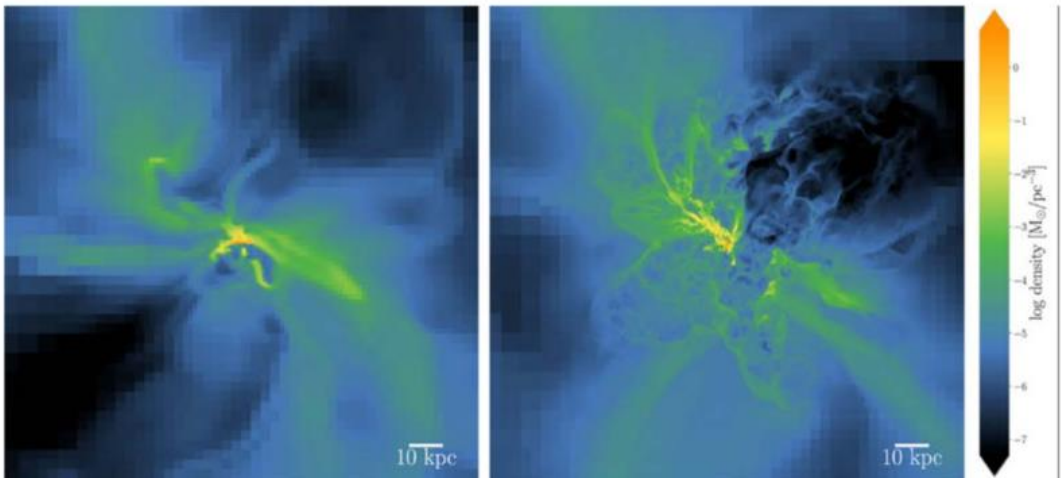


图 4.12 通过对比不同分辨率时大多数恒星正在形成时的红移的模拟，展示了高性能计算模拟推动理解复杂过程，如影响星系形成和演化的因素。来源:改编自 Molly S. Peeples et al 2019, “Figuring Out Gas & Galaxies in Enzo (FOGGIE). I. Resolving Simulated Circumgalactic Absorption at 2 ⩽ z ⩽ 2.5,” The Astrophysical Journal, 873 129. © AAS. 复制经过许可。doi:10.3847/1538-4357/ab0654.

4.5.4 数据科学与机器学习

在过去的十年里，数据科学取得了巨大的进步。机器学习技术在天体物理学中发挥着越来越重要的作用，而且这种趋势很可能会持续下去。过去几年，大学已经设立了多个数据科学/天体物理学的联合教员职位，并不断增设新的课程。本科生和研究生都在攻读 2010 年时尚不存在的联合学位，国家科学基金不断加大对各个子领域的“大数据”的研究投入。

天文数据为数据科学研究提供了许多机会。例如，摩尔基金会数据驱动发现倡议中的一篇关键论文将斯隆数字巡天（SDSS）列为数据驱动发现领域第六大最具影响力的工作，仅次于香农的经典信息论。天文数据对于数据科学来说很有价值，原因包括：数据集丰富、公开可用、结构良好且建模良好。这已经引发了许多无似然推理的新技术的发展，并促进了密度估计、隐式生成模型和概率编程等方面的进步。这些技术现在被广泛应用于各个领域（例如，粒子物理、化学和神经科学），并成为了跨机器学习和物理科学的新兴领域的一部分。

数据科学为研究天文数据和天体物理系统提供了强大的新工具。在提供识别数据异常现象的工具方面，机器学习已经取得了巨大的成功，并且可以通过重要因素加速大数据集中的参数估计（图 4.13）。这些技术可能会从本世纪二十年代可用的新数据集中带来变革性的发现。通过对复杂的非线性现象和仪器效应进行建模，机器学习有可能增加从天文数据集中获得的信息量。如果它能够成功用于模拟多尺度现象，它将能够更精确地模拟从行星形成到星系形成的各种天文过程。

发现：未来十年，包括机器学习的应用在内的数据科学将在天文研究中发挥越来越大的作用。无论研究人员日后从事天体物理学还是其他 STEM 领域的职业，将数据科学领域的培训纳入研究生或更高领域研究人员的培训之中，能够更好地为他们做好准备。

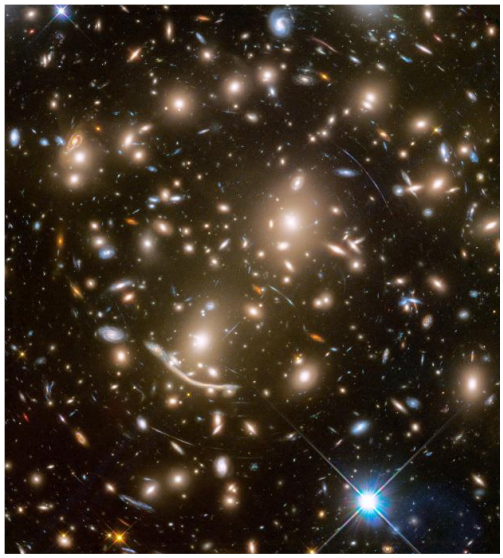


图 4.13 哈勃前沿场拍摄的 Abell 370 星系团，展示了由于背景星系的强引力透镜作用产生的无数弧线，因为它们的光在经过大质量星系团时被扭曲。机器学习已经证明了其识别强透镜弧的能力比当前技术水平快好几个数量级（参考 Ntampaka 等人，2019）[21]，这是在大数据集中深度学习技术对模型参数估计带来的影响的一个例子。资料来源：太空望远镜科学研究所，https://frontierfields.org/. NASA, ESA, J. Lotz 和 HFF 团队(STScI)。

附录 H：研究扶持基金会小组报告

H.1 引言

天文学和天体物理学十年调查 2020（Decadal Survey on Astronomy and Astrophysics 2020，简称 Astro2020）指导委员会委托本小组完成以下任务：(1) 总结目前的资源和支持状况，(2) 确定主要挑战，(3) 向 Astro2020 委员会提出关于计算、模拟、数据收集、数据处理，资助模式和方案，实验室天体物理学，以及一般技术发展计划等方面的建议。

为了完成相关任务，小组参考了科学界提交的许多有价值的白皮书，在三次小组会议上的相关陈述，国家科学、工程、医学研究院以前的研究，与其他科学和项目小组的互动，以及成员的专业知识。关于探索者和中型项目的部分基于小组间工作队的工作，这些工作队来自其他相关的优先小组。

H.2 提高在启动基金上的投资

H.2.1 数据存档的投资和关键分析

数据存储中心将继续为收集、策划、记录、提供社区培训和使数据集可访问提供关键的

支持基础设施。在今后十年中这些任务将更加重要。

可靠地对天球中的事件和天体进行存档一直是天文研究的基础。在现代，这些数据存储在数据库中以数字方式进行策划，过去、目前正在进行中，到 2030 年底，来自空间和地面的所有波段的未来任务将产生大约 500 PB 的天文数据，比人类历史上收集的天文数据多出几个数量级。解释这些数据所需的模拟将生成大小相当的数据集。虽然数据量以少数主要任务为主，但中小型数据集在规模和复杂性方面具有足够的挑战性，也需要特别注意其存档需求。在过去十年中，大多数基于大型任务和巡天的科学论文都是数据存储分析。在今后十年中，这些数据存储分析将变得更加重要，其技术实施也将更具挑战性。

随着 2020 年及以后的任务和巡天计划的实施，其中存在着巨大的机会，以大大增加科学回报，使其超越其核心目标，利用它们解决一系列更广泛的预见和不可预见的问题。这些机会的说明性示例包括:

- 以像素水平结合地面和空间的成像制作星图的天球图像，包括但不限于 Rubin 天文台、Euclid 卫星、Roman 望远镜 、eROSITA 任务和未来的宇宙微波背景辐射（CMB）实验，这将允许基于宇宙和星系尺度引力透镜的强大约束， 太阳系和银河系的光度红移、星系运动和演化。这一工作将需要单独的项目之外的投资，以实现增强的科学回报。还需要高质量的宇宙模拟来实现这些联合分析。

- 融合多波段不同口径不同时间的观测结果。如果访问方法、格式和元数据的集成足够紧密，能够高效、正确地分析联合数据集，则将允许对暂现和可变现象设置新的窗口。

- 及时分析多信使现象（电磁波、引力波和粒子探测器），将会使我们能够快速、有力地进行基础物理、黑洞和宇宙学的测试，并触发后续观测。

该小组概述了构建美国天文数据系统的框架，以便解决天文学最引人注目的科学目标。

现有的联邦资助的数据中心至关重要。由空间望远镜科学研究所（STScI）、NOIRLab、红外处理和分析中心（IPAC）、美国国家射电天文台（NRAO）、史密松天体物理台（SAO）、引力波开放科学中心（GWOSC）等地点运行，专注于不同类型的数据，并发展了独特的专业知识。他们的数据、元数据、文档保存、他们开发的获取工具的组合，最重要的是，他们的科技人员对于最大限度地提高该专业对新望远镜的投资的科学回报至关重要。然而，该小组的使用案例表明，按波段、任务和资助机构划分限制了科学成果的产出。由于它们由不同的机构资助，包括美国宇航局、NSF 和 DOE，具有不同的政策和科学目标，因此阻碍了实现新的联合能力所需的协调。

商业硬件和软件行业的技术发展也可以增强天文数据的处理能力。今天，大多数天文数据中心都使用自己的硬件：然而，云服务（包括计算和存储）是处理天文数据的一种越来越经济实惠和灵活的方式。十年前，尖端的数据库允许复杂的服务器端查询（例如，结构化查询语言 [SQL]），通常由最终用户下载数据和进行本地分析。今天，大多数中心都存在或正在开发科学平台（通常依赖于 Jupyter 的笔记本电脑部署），以提供更完整的服务器端分析工具，以便几乎所有分析都可以在服务器上执行，以减少下载到最终用户的数据（如果有任何数据的话）。这些科学平台背后的一个关键技术工具是"容器"，它概括了系统可以在任何硬件上复制和操作的完整操作系统。

开源软件和软件开发资源的激增扩大了天文学家可用于构建、访问和分析数据存储数据的工具套件。过去十年中软件和数据科学的突飞猛进开始影响到天文学领域，并且天文数据中心已经准备对这些现有的和未来即将出现的新技术加以利用。然而，数据中心以协调和协

同的方式采用这些新技术发展的能力有限。资源不足的重点是很少的资金用于建设它们之间的共享基础设施，部分原因是它们在 NASA、NSF 和 DOE 之间分离。

美国天文数据存储系统面临的一个主要挑战是需要协调其联邦数据存储，以满足 2020 之后 10 年的科学需求。在过去 15 年中，制定执行这些议定书的标准议定书和工具是必要的，但不足以促成这种协调。该小组建议，需要采取更积极和有力的方法，以最大限度地提高该专业对仪器、望远镜和卫星的投资的科学回报。

为了应对这一挑战，该小组设想建立一个天文数据档案系统（ADAS），作为一个伞式组织，协调现有数据存储中心的活动，目标是提高它们在完成任务方面的效力，并开辟本来不可能的新机会。地球科学、国际天文界以及美国以前和正在进行的虚拟天文台项目等其他领域的模型可以为解决天文数据存储中日益增长的需求和仍然存在的挑战提供经验。设计这一新的协调系统时的关键考虑包括：

- 保留定义了当前中心和任务的角色、责任、专业知识和资金流。
- 为数据存储科学家提供职业道路，使他们能够在 ADAS"家庭"内流动。
- 提供资源和使命，发起和领导全社会的努力，重点是通过教育、培训、公民科学和课程开发扩大参与计算技能、软件开发和数据科学。
- 提供资源和任务，以寻求机会，共同开发资源，并在各中心之间共享专业知识。
- 使科学驱动的努力能够跨任务和中心的边界进行分析。
- 为用户提供支持，以访问解释数据所需的模拟：
- 有能力将新的中心设在 ADAS 的保护伞下，并与国际合作伙伴进行协调/合作。
- 整合缺乏主要中心支持的小型项目的数据，并为这些项目的数据存档/保存提供支持。
- 为美国天文界提供机制，为 ADAS 活动的规划和优先安排提供投入。
- 继续支持期刊和访问期刊。美国宇航局天体物理学数据系统为 1300 万份出版物提供免费书目记录，在天文研究中发挥重要作用，并将继续在未来十年内发挥重要作用。
- 制定数据存储和传输规范的常见策略。

在现有单个中心之外，ADAS 可以提供资源，以完成其使命，使所有中心更加互联互通，并实现跨中心科学分析和协调补充培训计划。

为了在 2020 及之后的 10 年实现成果产出，该小组建议 ADAS、数据存储中心、资助机构、美国宇航局资助的主要任务以及 NSF 资助和 DOE 资助的项目通过资助一些与数据相关的工作来应对许多其他挑战：

- 供科学使用的数据产品和 API 可利用 ADAS 和/或存档中心现有的共同基础设施和协议，用于资助所有新任务和项目。
- 专门支持天体物理学的软件基础设施工作，包括社区主导和数据存储中心主导的工作。
- 启动新的方案和/或支持社区主导的培训和教育努力，使更广泛的科学家能够做出贡献。
- 保存来自小型或非联邦项目的数据。

- 保存元数据和文档的标准、机器可读方法。
- 为必要的软件工程工作提供必要的支持，确保某些重要的分析数据具有可复制性。

理论预测的天体物理模拟的数据存储值得在此特别提及，这是数据中心应该面对的重要的挑战，模拟包括天文数据输入（例如星空图像）的仪器模拟和天体物理学定律的模拟，这些物理定律可能会在某种理论框架下产生星空的特定图像。在这里，重点是模拟，这些模拟对于观测结果的统计分析至关重要。该小组建议数据中心继续开发软件和工具，以便能够让仪器来模拟仪器的行为。随着模拟和观测规模的扩大，与天体物理模拟共同定位观测数据的能力将被证明是必要的，该小组建议数据存储与提供这项服务所需的高性能计算中心和模拟组建立伙伴关系。该小组建议，模拟数据存储确保所有精心策划的模拟软件都进行过版本化、可追溯，并在必要时可用于复制模拟。

2020 年代，来自各种规模天文设施的新数据集将带来许多新发现和天体物理学突破性的成果。这一新科学的广度和深度、社区对它的贡献程度和包容性，以及该专业利用新机会和意外机会的能力，将取决于为未来十年及以后设计的资金充足和协调良好的数据存储系统。

这一建议符合 美国宇航局科学任务局《2019-2024 年开创性科学数据管理和计算战略》中的建议。事实上，该小组设想，除了数据存储现代化领域所包括的项目外，这个新系统还将涉及开放数据/开放软件、高端计算和高级功能领域。通过更加强调社区教育的方式，还可以加强美国宇航局的战略部署。

需要对天文学界和现有的研究中心的投入进行详细研究，以适当地确定和定义这个系统。该小组设想，该系统将需要大约 50 名全职员工（FTEs），其中包括天文学家、软件工程师和其他未来需要实施重点研究的工作人员。该系统每年将增加 1000 万英镑的运营成本，高于现有数据存储的费用。该小组建议这可以作为对现有计划的补充。

这一初步估计是基于在主要现有和计划设施从事数据管理、处理和分发的 FTEs 数量来考虑的，这些设施加起来有数百个 FTEs。要有效地协调这些系统，需要中心投入足够的努力，这促使这里设想的大量投资（类似于地球科学的系统，尽管这一投资比地球科学数据和信息系统（ESDIS）要小）。至关重要的是，无论上述目标如何，任何类似 ADAS 系统的最终活动范围和期望都应适当调整其现有资源。虽然一些 ADAS FTEs 将位于现有的大型中心，但有些可能最好与较小的项目同处一个中心。 如下所述，小组建议考虑一种方法，即 NASA 作为机构间支持数据存储计划的牵头机构，而 NSF 和 DOE 是向更广泛的国家社区提供高性能计算资源的牵头机构。

H2.2 软件

软件开发是天文学几乎所有方面的重要组成部分，而软件开发人员，也许更好地称为“软件仪器构建人员”，是天文学社区的重要组成部分。然而，两者都没有得到足够的资金或现有体制的支持。这个行业已经进入了一个时代，在这个时代里，项目最终的成功与否和影响力大小同等地依赖于软件和硬件的发展。

因此，软件开发需要包括在预算、计划和职业发展中。在这份小组报告中，软件包括数据归算处理管线（如 Astropy 和大型项目开发的分析包），大型软件项目（如 Modules for Experiments in Stellar Astrophysics (MESA)10 或 Enzo11），以及个人或小组在发表论文中产生结果所使用的代码。在过去的几十年里，软件开发有了显著的发展。大型团队的人经常一

起工作来编写和开发软件。

对于高性能计算来说，大型团队是必不可少的，一方面是由于正在解决的问题越来越复杂，另一方面是由于计算机硬件越来越复杂。今天的天文软件必须充分利用现有的各类计算系统的优势，从图形处理单元(GPU)，到运行在数千个节点集合中的多核处理器，再到运行在笔记本电脑上的标准通用 CPU。对于理论模型和数据分析来说，系统建模的复杂性要求具有广泛专业知识的团队开发大型代码。作为团队开发软件的实践与作为个人开发软件是完全不同的。该小组设想，未来的天文培训将包括大型团队代码开发和各类计算机硬件的最佳实践。

图像归算和分析设施(IRAF)12 在联邦资助下通过多种途径开发和维护了 30 年，它展示出一个免费可用的主流软件系统如何极大地增强社区的能力，使得来自多种仪器和望远镜的所有类型的数据可以使用。然而，由于对这一中坚力量进行现代化缺乏联邦政府的支持，导致这个用来进行数据归算和分析的软件基础设施出现巨大的漏洞。这个漏洞在一定程度上被志愿者运营的 Astropy 项目和 Astropy 包填补。然而，同样由于缺乏足够的和可靠的资金支持，Astropy 可能要和 IRAF 最终被迫所做一样，放弃它对社区软件的支持和开发。这些例子说明了在构建支持大多数现代天文软件的软件基础设施时，传统的拨款结构的不足。建立这样的基础设施的任何努力都取决于短期内持续的筹资努力，而有任何途径能获得长期规划和稳定所必需的长期承诺。这些基础设施开发工作需要可靠的联邦资金，以及软件和维护人员。

通过在整个天文学社区投资于软件培训和天文学软件开发人员，这些人可以在未来十年中为天文学转型构建工具。这将需要支持代码开源，并维护由个人以及广大社区开发的大型代码。这些投资将提高再现性，减少不必要的代码重复。

天文软件开发正在培育一代人，但他们正在天文学之外寻找更多的机会。如果天文学没有更多的资助机会和职业路径，很难留住这些软件开发者。

2018 年美国国家科学院报告的《NASA 地球和空间科学的开源软件政策选项》13 进行了重要论述：

SMD[科学使命理事会]需要培育一种新的开放文化和鼓励共享、合作的社会规范，可以部分地通过使用有针对性的赠款、研究金和奖金，鼓励学术界开发开放源码软件。向开放地转型也可以通过建立和使用开放源代码库（程序员编写软件时使用的代码和工具），来收集和传播社区软件。

2020 年 NASA 发布的《地球与空间科学研究机遇》（ROSE）14 呼吁通过资助开源软件和开源工具、框架和库，来响应 2018 年美国国家科学院报告的这一部分。然而，由于资金水平有限，这仅仅只是支持这些项目的很小一步。

美国国家科学院的报告提出了对所有提案要求软件管理计划的可能性。该报告进一步建议：涉及发布软件的提案应包括软件开发、文档、分发、支持、出版物和维护的预算说明。在提案中增加软件维护计划能鼓励代码作者自己进行开放代码和长期代码维护，并允许同行评审过程设定文化变化的速度。尽管维护会带来相应的成本，但高成本开发的代码在持续可用性方面带来成本显著下降。

硬件和软件环境在不断发展，面对这种不断变化的环境，天文软件界需要采用可持续的软件解决方案。例如，“容器化”是使代码保持可操作性的有效解决方案。容器化是对整个操作系统的封装，因此代码可以在任何硬件上操作。我们小组已将实施上述的容器化确定为 ADAS 的目标之一。

可复制性和再现性是科学过程的重要组成部分。对于软件而言，这需要支持、激励和教育社区关于保留软件和需要的输入文件的最佳实践，来使得分析和结果可复制、可再现。这可能需要调查人员、合作机构、科学图书馆、出版商、档案馆、存储库和联邦机构之间的协调努力。美国天文学会（AAS）期刊关于软件论文的出版、与开源软件学报的合作、Github 存储库的可用性，以及欧洲 Zenodo 存储库都是令人鼓舞的发展。改进软件的引用标准 15 有助于对工作给予适当的评价，记录分析中使用的成分，这对于实现结果的可复制性和可再现性至关重要。

NASA 科学任务理事会的《突破性科学的数据管理和计算策略（2019-2024）》在这方面提出了前瞻性的设想。国家科学基金会天文科学司（AST）可以与计算机和信息科学与工程司（CISE）合作，共同制定支撑天体物理学软件开发和数据管理的战略计划。类似地，DOE 宇宙学前沿可以与高级科学计算研究（ASCR）项目合作，为其天体物理学项目制定战略。这些战略可以包括为天文软件开发社区进行软件工程、计算机科学和编程实践方面的培训。

H.2.4 高性能和高吞吐量计算

在理论科学和数据科学中，计算已经成为天文学和天体物理学中几乎每个主题的基础。提高计算基础设施的性能对于科学进步至关重要。DOE 和 NSF 将在未来十年大幅提高其高性能计算能力，而 NASA 将以较慢的速度提高其高性能计算能力。

高性能计算 (HPC) 涉及使用大型超级计算机进行高速详细计算。HPC 任务需要大量的计算能力来解决复杂问题（例如，物理过程的模拟）。

高吞吐量计算 (HTC) 能够以高效的方式执行相对简单的计算任务（例如，处理和分析非常大的数据集）。

在过去的十年中，理论和数据科学中的计算无疑已成为天文学和天体物理学中几乎每个主题的基础。数值模拟和大数据分析变得越来越精密，同时它们在天体物理学中的作用也有所增长。尽管接近摩尔定律的极限，但计算能力也一直在稳步增长，百亿亿级超级计算机最早有望在 2022 年公开可用，并且到 2030 年计算能力有可能大幅扩展。随着软件、分析和计算能力的日益成熟，在未来的十年里，科学发现有巨大的潜力和机会。

HPC 支持通过模拟恒星、行星、星系、宇宙和引力波事件的形成和演化等过程来进行发现。HTC 支持使用大型数据集进行发现，包括使用存档观测数据的调查、互补观测的联合像素处理以及对大型模拟数据集的分析。提高在 HPC 和 HTC 方面基础设施的性能及其使用人员的专业知识，对此有持续并且不断扩大的支持对于科学进步来说至关重要。

随着时间的推移，观测和合成数据集的数据量不断增长，从 TB 级到 PB 级，并且很快将到达 EB 级。在这个大数据时代，有机会采用基于公有云的云计算作为具有成本效益的解决方案，而不是托管庞大的硬件资源和众多专有设施。目前，HTC 的云计算效用比 HPC 更清晰，尽管这种情况可能会发生变化。专家小组建议资助机构继续探索云计算的潜力，并在适当的情况下为利用云计算的项目提供支持。

在接下来的十年，DOE 和 NSF 计划大幅增加他们在 HPC 方面的能力，而 NASA 计划以较慢的速度扩张。为了确保为任务提供足够的 HPC 资源，并确保在未来十年内社区有足够的机会使用 HPC 设施，需要协调各资助机构，或者扩建 NASA 的 HPC 设施。然而，支持横向项目的资金很少。如上文 H.2.3 节所述，Astro2010 十年调查建议发展一项 TCAN 计划，该计划旨在成为 NASA 和 DOE 之间的空间天文学合作以及 NSF 和 DOE 地基天文学合作。一项 TCAN 计划由 NASA 和 NSF 发起，此后也成为 NASA 的唯一计划。专家小组认为，重振三个资助机构之间所有的重点合作，将能够最有效地利用资源，促进目前由于缺乏支持而进展缓慢的关键前沿计算领域的快速发展。

在过去十年，HPC 模拟已成为理论建模、预测和巡天构想以及观测数据的最终分析和解释不可或缺的一部分。开发软件并利用软件来承担这些任务不仅需要拥有大量计算和存储资源的专业设施，还需要在计算机科学和天体物理学方面具有广泛专业知识的人员。对于个体研究人员在获取 HPC 资源时的潜在障碍，社区非常关注，并且提交了几份白皮书（特别是为那些与大型 HPC 和 HTC 设施没有设备权限的机构）来强调这些问题。

专家小组建议各机构增加对 HPC 和 HTC 设施使用培训的投资。为了确保进行 HPC 和 HTC 计划的适当培训以及更公平地使用 HPC 和 HTC 设施，在本科和研究生阶段提供广泛的培训非常重要。此外，专家小组建议直接从设施中为培训提供更多支持，包括针对所有职业级别的讲习班和实习计划。

资助机构之间的协调将有利于更广泛的国家计划。该小组提出了一种方法，即 NASA 牵头支持归档，DOE 和 NSF 牵头向天体物理学领域的科学家提供 HPC 超级计算设施。

H2.5 数据科学与机器学习

天文学和数据科学之间的互动是富有成效的双向交流。数据科学的进步使天体物理学有了新的见解。丰富的天文数据集具有潜在的物理对称性，可以推动数据科学技术的发展。对新数据科学技术培训的支持将丰富从大数据集中获得的回报，并推进数据科学和天文学。

在过去的十年里，数据科学取得了巨大的进步。机器学习技术在天体物理学中扮演着越来越重要的角色，而且这种趋势很可能会持续下去。过去几年，已经有多个数据科学/天体物理学的联合教员任命。许多大学都在设立这一领域的新课程。本科生和研究生都在攻读 2010 年时尚不存在的联合学位。

天文数据为数据科学研究提供了许多机会，并已被证明是一个有价值的数据集。例如，Stalzer 和 Mentzel17 将斯隆数字巡天（SDSS）列为大数据领域第六大最具影响力的论文，仅次于香农的经典信息论。天文数据对数据科学很有价值，原因有很多：

- 这些数据是开放的，没有商业价值，不存在与其他类型的图像数据相关的道德伦理问题。相比之下，从互联网上抓取的人脸图像通常没有获得许可，通常被用于照片监控，并且通常是带有种族偏见的样本。
- 天文数据丰富，涉及范围从图像、表格、图表、不均匀的时间序列到多维网格。
- 物理科学中的数据由单个粒子、行星和恒星以特定方式相互作用构成，并具有易于理解的对称性——这一点不同于其他领域广泛研究的图像和序列的数据结构。这种丰富的结构已经启发了图神经网络和几何深度学习的早期工作。
- 天体物理学家有高保真的模拟器，可以捕捉描述天文现象（例如，大尺度结构的演化）和天文过程（例如，鲁宾望远镜对引力透镜星系的观察）的机械因果模型。近年来，天体物理学家和数据科学家已经开发了许多新的技术来进行无似然推断，在密度估计、隐式生成模型和概率编程方面取得了进展。这些技术现在被广泛应用于各个领域（例如，粒子物理、化学和神经科学），并且是跨机器学习和物理科学的新兴领域的一部分。
- 因为数据可以模拟，所以可以查询模型是否过拟合。由于许多天体物理数据集中蕴含的物理知识是已知的，因此可以得知人工智能（AI）是否正在学习真正的基础规则。这是比交叉验证更重要的模型测试，对于使模型安全并真正提高对科学的理解非常重要。
- 与使用人工智能对文本或图像进行分类相比，为物理世界构建理论潜在的道德问题更少。虽然数据科学中的大部分工作都聚焦于图像，因为它们不需要领域知识并且

对在线广告和客户分析很有用，但物理系统可能是通向通用人工智能（AGI）的更好途径。

数据科学为研究天文数据和天体物理系统提供了强大的新工具。作为一种识别数据异常现象的工具，机器学习已经取得了巨大的成功。这些技术可能会从预计在二十世纪二十年代可用的新数据集中带来变革性的发现。通过对复杂的非线性现象和仪器效应进行建模，机器学习有可能增加从天文数据集获得的信息量。如果机器学习可以成功地用于模拟多尺度现象，它将可以更准确地模拟从行星形成到星系形成的各种天文过程。

由于这是一个迅速发展的领域，该小组建议资助机构利用拨款计划和现有的数据中心来发起和支持对天文学家进行持续培训，并提供广泛的机会，使不同的科学家群体能够应用和教授这些技术。