# Present and Future Needs for Scientific Computing at the CfA

A Memorandum for the Director and Chief Technology Officer Outlining Priorities for Investment in the CfA's Computational Infrastructure and Capabilities

*Submitted by the Scientific Computation Advisory Committee*
*June 2022*

Laura Brenneman, Chair
Daina Bouquin
Nancy Brickhouse, ex officio
Jason Eastman
Paul Edmon
Sylvain Korzennik
Sean Moran
Durai Ramadurai
Hossein Sadeghpour
Peter K. G. Williams

## Executive Summary

The Scientific Computation Advisory Committee (SCAC) is composed of a diverse group of representatives from across the CfA. The committee is charged with assessing the present and future scientific computing needs of the Center, creating prioritized recommendations for how to meet those needs, and advising the CfA Director's Office (DO) on how best to allocate resources in support of such recommendations.

The SCAC has traditionally conducted periodic surveys of the CfA community in order to assess the state of stakeholder satisfaction with the computational facilities, access, and support available. The most recent of these surveys, performed in 2019, is appended here (Appendix A). Recently, however, a confluence of circumstances has prompted the SCAC to undertake a more comprehensive review of the scientific computing landscape. Notably, the COVID-19 pandemic enforced mandatory remote work for the majority of CfA staff, particularly in Cambridge, MA, prompting many to reconsider the types of computational access, support and training they require. Transition of the CfA Directorship in 2022, anticipated reorganization within the DO, and new computational initiatives are also shaping how we envision the best path forward.

The recommendations summarized below derive from the results of the 2019 SCAC survey (Appendix A), as well as two years of committee meetings and discussions. Our findings fall into five principal categories: high performance computing, cloud computing, data management, IT support, and professional development. The recommendations made within each category aim to:

- **Increase our computational capabilities, robustness and support**, keeping us at the forefront of astrophysics;
- **Improve internal and external access to our scientific resources and products**, making us more visible and nimble for purposes of collaboration and dissemination of scientific knowledge;
- **Promote equity between Harvard and Smithsonian** components of our community;
- **Empower our workforce** to stay up to date with the latest innovations in programming, data science tools and techniques, collaboration platforms, and many other subjects.

Each category and its associated recommendations follow in subsequent sections. In total, we anticipate the required effort to total ~3-4 additional FTEs. The SCAC welcomes the opportunity to discuss these recommendations in greater detail with the DO.

## High Performance Computing

High performance computing (HPC) is an integral part of modern astrophysics, and will only increase in importance in the coming decades. The CfA's world-class reputation for excellence and leadership warrants a significant investment in HPC infrastructure and accessibility in order to keep pace with our expanding computational needs and to continue to foster innovation.  This is especially true as CfA researchers become more involved with large computational projects such as the Event Horizon Telescope or the Illustris project.

Currently, HCO staff have access to the FAS "Cannon" cluster (70,000 cores, 340 GPUs, and 40 PB of storage), while SAO staff have access to the Smithsonian "Hydra" cluster (5,000 cores, 6 GPUs, 4.5 PB of storage).  Hydra and Cannon also support separate software stacks, from using a different job scheduler to having a different suite of installed "packages" and licensed tools. This is in stark contrast to other CfA resources, such as telescopes, where there is a strong tradition of ensuring that HCO and SAO staff have equal access.  The largest current technical obstacle we have identified to achieving shared access is a differing interpretation of 2 CFR §200.468, the "Specialized Service Facilities" section of the Uniform Guidance for Federal Grants, between the two institutions.  Harvard interprets the language to mean that HPC facilities may be funded out of overhead, while SAO does not.  This is an issue because **Cannon operates in a cost-center model, while Hydra is funded by SI/OCIO and does not charge user fees.  Therefore, SAO cannot provide an SAO-wide fund to support its employees' potential use of Cannon, and there is no mechanism for HCO to compensate SI for its employees' potential use of Hydra.**  The first step to securing equal access to these HPC resources for all employees is to revisit this interpretation and determine if the two systems' cost models can be harmonized.  This conversation needs to be initiated at the DO level.

Computation is a constantly growing field; as such, sustained investment into both hardware and personnel is a necessity.  At present, SAO funds a person part-time (0.25 FTE) to support Hydra's use, while HCO, through the Institute for Theory and Computation (ITC), funds a person part-time (0.5 FTE) to support Cannon.  Expansion of Hydra is typically funded by OCIO's Office of Research Computing, while CfA-supported expansions of Cannon are done on an ad-hoc basis through the ITC budget.  A more sustainable strategy with better economies of scale would be for the CfA to help fund larger expansions of Cannon and Hydra for use by all researchers, regardless of institution.  Similarly, based on the size of the CfA and comparing against similar institutions (which on average have 1 facilitator per 100 researchers), this **committee recommends an increase in staffing to 2-3 full time HPC facilitators to support researchers using these resources.**  Duties of the facilitators would include guidance on best practices for general use cases, troubleshooting problems, and helping to plan future installations.

## Cloud Services

The use of cloud services at the CfA has been increasing in recent years.  The 2019 SCAC survey found that 77% of respondents desired access to cloud computing and 45% requested additional support and advice on using cloud services.  Cloud services can offer nearly immediate deployment of resources that can never be done on-premises on such a timescale. The cloud can also be an effective platform for developing proof-of-concept studies that require a large amount of resources for a short timescale or very specific resources not (yet) available on premises, as well as for moving well defined services off premises, like standard data processing pipelines, data backup, data sharing, visualization, and specific web services.

At present, it is quite tricky for individuals or small groups to understand if and when cloud services could or should be used; just because they are available does not mean that they will be cost effective, or comply with required regulations (e.g., ITAR). Some workloads make much more sense to execute on cloud infrastructure rather than locally, but **right now users are making individual arrangements with cloud providers without any expert assistance in terms of wayfinding, support, and problem solving**. Some projects have encountered unanticipated costs under this approach, and fear of such costs probably prevents other users from trying cloud approaches. Continuing in this ad-hoc manner exposes the CfA to other risks as well, such as failed long-term preservation of projects and web sites when funding sources are exhausted.

**We recommend that the CfA establish a dedicated position to provide cloud computing assistance and help develop guidelines for the use of cloud services at CfA.** The person in this role would:

- Offer advice to projects on whether cloud services would be appropriate for them;
- Recommend a cloud provider/service that offers a good match to their needs;
- Document some of the most commonly used/asked-for cloud services;
- Inform projects of factors they should consider before making a decision (ITAR restrictions, cost considerations, etc.).

Although this recommendation does not address broader strategic questions about the future of cloud computing at the CfA, we believe the above position will contribute to the knowledge needed to inform future decisions on the Center level.

## Data Management

CfA scientists acquire, analyze, and publish data for projects ranging in size from small individual investigations to large NASA observatories (i.e., *Chandra*). While projects at different size scales have vastly different data management needs and staff expertise, **there are essentially no CfA-wide policies, guidance, or support that bind together these disparate projects into a unified CfA effort toward data sharing, dissemination, and preservation.** Yet most proposals and contracts require an explicit data management policy.

The 2016 memo, “Data Repositories at CfA” (Appendix B), describes why such an effort would be desirable. In brief, making our data archives more easily accessible will increase the world-class science that can be done with CfA data. At the same time, offering dedicated professional support for such efforts will alleviate the ever-increasing time burden on CfA scientists to execute project data management plans entirely by themselves. This burden will become more acute because funding institutions are making their requirements around scientific data more stringent, as exemplified recently in NASA’s updates to their scientific information policy (SPD-41), and the NSF’s new requirements on data management plans.

Very recently, a new initiative to build a unified data portal and analysis platform for the CfA's public archives has been drafted, tentatively called the CfA Nexus (hereafter Nexus). While this effort is still in its early stages, **the SCAC supports ongoing efforts to further develop the Nexus concept.** The Nexus could play a large role in implementing the recommendations of the 2016 memo, which have grown more urgent in the intervening years. Estimates of the effort required for the Nexus project ultimately total ~10 FTEs, though a more precise timeline of the ramp-up to this level of effort will be important. Much of the Nexus work breaks down into software development and science support, along with funding to support connections with the Wolbach library (digital preservation), ADS (cross-referencing) and Systems (hardware expansion and cloud service interface). Some of these tasks align with separate recommendations made by the SCAC in this memo, e.g., building up hardware infrastructure and cloud computing expertise. In this case, it may be advantageous for the ~1-2 FTEs assigned for this effort through the Nexus to be "shared" in support of implementing the SCAC recommendations.

Some key needs in data management may end up outside the scope of the Nexus, however, and these needs should be addressed separately. Examples of needs that may not be addressed by the Nexus project could include digital preservation staff who would assist individual scientists with creating metadata and depositing research products in archival repositories to satisfy data management plan requirements for their grants, or professional development staff to provide computing support/training in the various software tools needed to prepare a data set for publication (see the Professional Development section below). Both the CF/Syshelp and the Wolbach Library could be suitable homes for these additional types of direct support.

## IT Support

The nature of modern research is evolving to blur the lines between scientific computing and what is traditionally considered "IT support." The 2019 SCAC survey (Appendix A) shows that in order to work effectively, researchers need additional support for tasks such as creating websites or sharing data with external collaborators. Further, about 55% of respondents self-manage their computers since they want to have administrator-level privileges to install specific software packages. A large fraction of these use laptops as their primary computer, while the CfA offers limited support for laptops, whether Mac, Windows or Chrome.

While larger scale projects often have dedicated IT support, including disaster recovery protocols, scientists working on machines outside of these frameworks are in most cases left to their own devices. One respondent to the 2019 survey commented, "Frankly, not having this type of support is embarrassing for [an] institution of this size and reputation." The three main issues facing individual users attempting to self-manage their computers are as follows:

- **Data backups (i.e., disaster recovery)**
  - Most users take on the responsibility for this themselves.
  - There is a general lack of coherent, easily accessible information on cloud backup options that are "free" to use and/or already supported at the CfA.

- **Software installation**
  - While the CF and Syshelp do provide installation and licensing of scientific software (e.g., HEASoft, IDL) on machines they manage (i.e., desktops at the CfA), this support does not extend to SAO/Harvard-purchased laptops or self-managed desktops.
  - One consequence is that users are often required to purchase individual licenses for software where CfA has already purchased a multi-user license.
- **Hardware support**
  - There is currently only limited hardware support for computers of any kind, and no in-house hardware support for SAO Mac users. This support exists for Harvard, but it is unclear whether SAO employees can make use of this service. As a result, SAO employees are currently obligated to take their machines to Apple Stores when hardware issues arise so that they do not void their AppleCare warranties, which are mandated by the CF/Syshelp for all SAO-purchased Macs.
  - This issue could be mitigated either by employing IT professionals trained/licensed in this type of hardware support, or by having readily available loaner machines in a variety of platforms (coupled with increased robustness of data backups). Users experiencing a hardware issue could then retain the ability to work while IT coordinates the repair of their machine with the appropriate vendor.

**The SCAC recommends increased investment in CfA IT support**. It may be possible to increase the breadth of CfA IT support at a relatively low cost by evolving, or even eliminating, the distinction between the CF and Syshelp support groups, given the duplication of services currently present. New investments should be contingent on an **audit of the CfA IT "landscape"** by the CTO in order to identify a strategy that ensures that the unmet researcher needs identified above will gain an organizational home. **Key needs to prioritize are hardware support, individual software installation, and data backups.**

## Professional Development

Although professional development (PD) may seem out of scope for this document, the SCAC advances the topic because, without a sustainable PD program, the CfA community will continue to have a limited capacity to take advantage of advanced computational resources. In fact, the 2019 **SCAC survey confirmed that the CfA research community has a significant, unmet need for training and PD around scientific computation,** with 70% of survey respondents expressing interest in training in "open source software languages and modules." Considering the growing sophistication and centrality of computation in astrophysics, there is every reason to expect that this need will continue to grow. A robust CfA PD program would increase the impact of the Center's current researcher cohort and would act as a magnet for talented young scientists, since few organizations in astronomy have the breadth and depth to offer the range of PD opportunities that the CfA can provide.

Admittedly, the need for a coherent approach to PD at the CfA extends far beyond scientific computation. For this reason, the SCAC recommends that when considering CfA PD activities, the DO take into account both the full breadth of the CfA community (researchers, engineers, clerical staff, managers, etc.) and the full spectrum of professional skills that deserve to be cultivated at the CfA (scientific computation, scientific visualization/communication, mentoring, leading IDEA practices, etc.). To date, CfA PD efforts have been ad-hoc, as has their funding. A more deliberate approach with a more stable funding source would allow for greater organization in arranging training activities and a better overall ability to meet community demand. In the long term, **the SCAC recommends the establishment of a dedicated DO budget line for PD activities**, sufficient to support at least one full-time staff person to coordinate activities, monitor the efficacy of the program, and set direction. Wolbach Library, the current locus of the majority of CfA-internal PD activities, would be the natural unit to host dedicated PD staff.

To demonstrate the justification of this resource allocation, **SCAC recommends that the DO identify a CfA PD point person and invest in a PD pilot project of sufficient scope to allow for several experiments with different PD activities (approximately $20,000/yr × 3 years[1]).** For comparison, we estimate current ad-hoc PD spending at $15,000/yr. The pilot project would require the support of a staff person responsible for measuring interest in the activities and their impact. At the beginning of this project, the point person should work with the DO to determine success criteria that will be used to assess whether the project should convert to an open-ended program at its conclusion. Any conversion should carefully consider the long-term funding mechanism for such a program to ensure that the community's continuous need for computationally relevant PD opportunities is met.

[1] Library staff have envisioned a program of 8 professional development events per year; roughly 1/month from September-December and again from February-May. Based on past experience, about half of these events might be led by internal speakers, and the other half by external speakers. The average costs for such events are, respectively, $250 and $4500, resulting in a baseline budget of $19,000 per year.

# Appendix A: 2019 SCAC Survey Results

# Scientific Computing Advisory Committee Report

Submitted August, 2019

Paul Edmon (TA), Iouli Gordon (AMP) Chair,  Jim Babb (AMP), Daina Bouquin (LIB), Igor Chillingaryan (OIR), Nick Murphy (HEA), Matt Payne (SSP), Jonathan Weintroub (RG), Nancy Brickhouse DO (ex officio) Roger Brissenden DO (ex officio) Sylvain Korzennik SSP (ex officio) Van McGlasson CF (ex officio)

# Executive Summary

The intention of this report is to provide an overview of the Computational needs of the CfA and to provide recommendations for meeting both current and future needs.  To this end the Scientific Computing Advisory Committee (SCAC) put together a survey to poll the community as to their perception and needs.  Subsequently the SCAC held several meetings to discuss the results and make recommendations to the CfA DO and ADs.  The following report is divided up into five sections: Training and Professional Development, Data Visualization and New Tools, High Performance Computing, Data and Cloud, and CF/IT Services. Results of the survey are given at the end. It is planned to make these results available to everyone at the CfA, after the “individual comments” section is removed.

The overall response from the survey was favorable with over 100 participants.  The general sense is that there has been a lack of an organized response to Scientific Computing needs at the CfA.  While individual groups and labs have met those needs with resources at hand, the survey suggests that the CfA as a whole would benefit from a more concerted effort.  Special attention should be paid to Training and Professional Development as there was strong interest in learning about new tools and techniques.  High Performance Computing (HPC) continues to be the bread and butter of Scientific Computing with both the Hydra and Odyssey clusters serving the internal needs of the community.  Further investment in HPC will be needed to meet current and future demand.  There is a need for a long term storage and archive with concerted effort to both properly catalog and label the data but also ensure that the data is stored in a useful way for posterity.  The CF will require more resources to meet current internal core IT needs and, in addition, to be able to offer expertise on various computational topics.

All of these needs, if met, will allow the CfA to continue to be the premier astrophysical research facility in the world as well as aiding in recruitment of top talent.  As we have seen with recent discoveries, computation is an integral part of modern research.  Thus a concerted effort should be made to pool resources at the CfA and work towards creating a stable and cutting edge infrastructure and support for Scientific Computing so that researchers at the CfA can continue to make groundbreaking scientific discoveries.

# Table of Contents

# Training and Professional Development

There is a pronounced need to provide professional development and training opportunities to support collaborative computational work at the CfA. This is particularly true regarding open protocols and tools that enable scientists to share their work throughout their development processes (e.g. Open Source development practices, GitHub) and to work with remote resources like HPC. It would be highly valuable to the community to have access to training to support advanced skills in these areas as well as introductory training, especially regarding tools used in data analysis and visualization. 58.6% of the community also expressed their need to share web-based data products outside of the CfA, and the mechanisms scientists are currently using to share their work are diverse. It would therefore be valuable to focus on training opportunities on platforms that are not necessarily specific to the CfA, in addition to institutionally specific resources.

It would also be valuable to provide training to support archival preservation of data products (e.g. use of Zenodo rather than dropbox). 77.3% of respondents expressed favorable interest in a service that would provide long-term curation and archiving of data products, and considering the fact that more than half of the respondents also expected to make their data products available in perpetuity (Q25), training and other professional development opportunities related to working with computational resources should incorporate practices that support long term sustainability of those data products.

The committee recommends that the organization of seminars will be included in the responsibilities of the Wolbach library. The head librarian has been very successful in organizing of trainings of different computationally-related subjects in the past. The survey has identified that trainings in the following areas are highly desirable by the scientific community:

- Open source development and programming languages (Python, R, Julia)
- Code collaboration/version control tools (GitHub)
- Containers (Docker, etc.) and virtual environments (Conda)
- Cloud computing/storage/services
- Intro to/advanced topics in HPC (Hydra, Odyssey, Pleiades)
- GPU/CUDA/etc.
- Databases: SQL, NoSQL, other DB tools
- Intro to/advanced topics in Unix/Linux
- Field Programmable Gate Arrays
- Commercial software packages (Matlab, IDL)

# Data Visualization and New Tools

A substantial portion of respondents who answered questions related to data visualization and tools to support visualization were interested in taking advantage of infrastructure that are currently not available at the CfA. 40.3% of respondents to Q33 on data visualization resources expressed interest in using a video wall, while 32.3% were interested in VR and nearly a quarter of respondents would use 3D printing. Likely uses ranged from data analysis to public outreach and scientific presentations. Physical resources like a video wall could be made more widely accessible CfA scientists by installing one in the library rather than relying on ad hoc access to one managed by the solar group. It would also be valuable to make iterative investments in VR and 3D printing technologies to make sure scientists have the ability to try these tools more comprehensively to see how they could be incorporated into their work.

# High Performance Computing

The survey indicated that a large fraction of the respondents use HPC. The vast majority had no problems using Hydra or Odyssey. Most run primarily serial jobs, but also multi-threaded and MPI based jobs are a significant portion of the work. A small but growing group use GPUs or run hybrid jobs (both MPI and threaded applications). The types of HPC jobs are modelization and simulations, as well as data reduction, plus some visualization and machine learning. In 2018, 154 CfA users on Odyssey used 48 million CPU hours, 167 users on Hydra used 11 million CPU hours. Only 9% of respondents were interested in new technologies, with 10% wanting to use HPC for visualization. It is estimated, based on the survey responses, that to satisfy the needs of the CfA as a whole, a cluster of roughly 20,000 cores in size with about 10 PB of storage would be required, an investment well beyond the current CfA computing budget.

### *Reported Bottlenecks (in order of significance)*

- Processor speed
- Number of processors
- Amount of Memory per node
- Storage

### *Suggested Institutional Solutions: Infrastructure*

- Ramp up a CfA Budget Line Item for HPC Infrastructure purchases for both Odyssey and Hydra. Based on current Odyssey costs[1] and the above usage estimate this would run $3.5 million for compute and $500,000 for storage per year to lease. This is known to

[1] https://www.rc.fas.harvard.edu/billing-faq/ has $0.0205 per CPU/hour and $50/TB/year as of FY19

be a high estimate from discussions with FAS Research Computing Director Scott Yockel and could be lowered. Given this, it is recommended to ramp up to a budget of $1 million per year for HPC infrastructure costs at the CfA.  It would be best if it was a roll over budget item so savings for each year could be accumulated for larger bulk purchases

- Negotiate SAO access to Odyssey with FAS Research Computing

*Suggested Institutional Solutions: Training*

There is interest in more training (cloud computing, HPC and GPU).

- Users want more information on computational facilities (31% SI/SAO, 24% HU)
- 42% use/need/want to use cloud computing.
- Synergies with the 'training' recommendations: FAS Research Computing provides some training and could expand offerings if requested.

## Details & Discussion

111 responses on 35 questions, with 18 questions pertaining to HPC.

More Detail on Recommended Solutions:

- More cores per job, More storage, More memory, Better support, Faster storage, Different time allocation scheme, More cores per node, Faster network.

### HPC Usage in 2018

| | CPU [million hours] | % of machine | Equiv. cont'ly used cores | No of users (no. of PIs or no. of active) | No of cores avg/max | Storage [PB] |
|---|---|---|---|---|---|---|
| Odyssey | 48 | 10% | 5,500 | 154 (16) | 32/2048 | 3.00 |
| Hydra | 11 | 31% | 1,300 | 167 (40) | 12/512 | 0.25 |

### HPC Requirements to Completely Satisfy Survey Capacity Requests In-house

Looking at the numbers on the survey for Question 4 which asked about the scale of computation that was planned by respondents, a cluster of scale 20,000 cores would satisfy the needs of all users that wrote in.  The largest scale job requested was 10,000 cores but more typical sizes (in terms of bulk) was about 512-2048. Using the rule of thumb that it becomes difficult to schedule jobs that are larger than 10% of the cluster size we arrive at about 20,000 cores. For reference Odyssey's main public queue is 14,000 cores and FASRC is investing in new compute summer of 2019 which will about 30,000 cores and about 30 GPU’s. Given the growth in HPC needs and projected future needs, investment will need to be made in both Hydra and Odyssey. The CfA should establish a budget line item for HPC to fund infrastructure purchases for both Odyssey and Hydra hardware. CfA should also negotiate with FAS Research Computing to permit access for SAO employees to Odyssey, this may involve a fee for service which can be paid for by the HPC budget line item as per above.

# Data and Cloud

## Summary of Findings

The relevant survey results are pasted below. About 60% of respondents have indicated that distributing data products are essential to their research and grant activity. Some are satisfied with distribution of static files of small to medium size for which they use CfA ftp or services like Zenodo or archives available through the publishers. About 25% of respondents have encountered problems that involve inability of providing dynamic pages and services on institutional servers. Their solutions were diverse from buying cloud space on AWS, Azure or Google to purchasing and maintaining their own servers. Some of these servers are placed at CDP but are not being supported by the CF. Some scientists ask their colleagues at the institutions with more friendly policies to host their data for them.
In order to address data needs of the CfA community we feel that following steps are needed:

**Data**

- Create institutional data archive on dedicated servers. Mentioning institutional infrastructure for data distribution will make proposals of our scientists stronger.
- General purpose storage needs total to roughly 10 PB and need to be available for purchase by grant funds but also be durably funded when grants end.
- Institutional policies have to be checked and if possible revise to allow outgoing browsable data.
- Need to ensure ability to move large amounts of data in a timely manner
- CF does have Google space that have been allocated to virtual machines of some groups free of charge. Unfortunately, this is not a common knowledge. This has to be on an institutional scale with survey results serving as a guide on size.

**Cloud**

- There is significant interest in Cloud solutions especially for shared storage, collaborative office software (like Google Docs), and code development. Some groups pay for GitHub accounts that are not subject to limitations. It is not expensive but institutional access maybe worthwhile.

## Expanded Notes and Discussion

**Data**

- Significant interest from respondents (~75%) in a service at the CfA to support data archiving and data product curation (Q23).
- >70% of people want to store data for >5 year timescales (Q25), but the majority just use local machines and/or websites (Q24).
- There should be tools in place moving large amounts of data between institutions. We already have Internet2 for instance but scientists will need to be educated about that.

**Cloud**

- Caution needs to be taken in Cloud applications to right size expenditures, especially with respect to any computation. Bottlenecks for Cloud computation come down to data locality problems.
- Need to enable people who get grants that have AWS and Azure credits attached.
- There is significant interest in Cloud solutions for distribution of data & results (Q20) especially for shared storage (Q21 & Q30) , archiving (Q24), collaborative office software (like Google Docs: Q21), and code development

# CF/IT Services

A pleasing outcome of the survey is that there does seem to be general appreciation for and use of, the services offered by CF. This is seen in the generally high interest in the survey, and the large numbers of responses to questions which relate to CF services. While most users self-manage their machines, more than 1/4 of respondents depend on CF-managed computers. For those machines that are self-managed there is implicit endorsement of the network, printer, and storage infrastructure provided by CF. (One respondent reported their machine managed by “my son” which proves that the CfA is not entirely devoid of a sense-of-humor.)

It is notable, if the survey is representative, that at SAO by far the majority of machines are MacOS or Linux, roughly equal shares, with MacOS having a slight edge. Among the science and technical community at CfA only 5% of machines at SAO are Windows OS. Note however that CF does support a very large base of Windows OS machines among administrative staff, and also at the Chandra Operations Control Center uses Windows OS, communities who were not expected to respond strongly to this survey.

Question 26 asks "What services would you like to see more information about on the CF website?”. This was heavily answered by respondents, and the dominant four answers were SAO Computational Facilities, Harvard Computational Facilities, Eduroam, and WiFi. There was only a single respondent to each of the six other options. It seems reasonable to consider that this question doesn’t necessarily signal a lack of information on these topics on the CF website, but rather, keen interest in these services from the community.

There are services not heavily supported by CF in which there seems to be a lot of interest. Question 23 asked "Would you be interested in a service at the CfA to support data archiving and data product curation?”, with only binary response options, yes, or no. 95% of respondents responded to this question, and 77% of those responded in the affirmative. This would suggest a data product curation service would be very popular if offered by CF. Perhaps related, there was a 77% response to the question on Cloud services, with the overwhelming majority desiring access to Cloud Storage, or Cloud computing, or both.

Question 1 on training sessions is also notable. This question was of the “select all that apply” type, with 16 options, which means the response can be more than 100%. The result was a response of 422%, by far the most popular question in the survey, with substantial interest in a broad swath of topics, with Python and open source languages leading the pack, and Github/version control a close second. Fairly arcane seeming topics had some response, for example, training in GPUs and FPGAs. Of course it isn’t inexpensive to offer training, and if CF is tasked with this new service, it would drive their budget. Maybe CF coordinating peer training is a possibility?

While CF services are clearly valued, it is equally clear that the demands by the CfA research staff go well beyond what the CF can provide at its current staffing levels. Two core CF service areas, in particular, are in urgent need of additional staffing: (1) the Linux-based scientific computing support group, and (2) the web services design and management group. With the revamping of the CfA website and branding it seems illogical not to increase web services support. Both areas have had, over the last decade, staffing losses through attrition. The Linux group has today only half the personnel it did ten years ago. As a result, system build certification and migrations to new versions of the OS lag behind the pace of new releases. These positions provide core and foundational CF functions, and they ought to be permanent, and overhead funded, as the whole of the CF currently is.

Once the core staffing problems have been addressed, the SCAC discussed the provision of additional specialized services, like training, data curation, visualization, cloud services, HPC, etc., services that go beyond the core CF mission and functions, and are not currently budgeted. In the case of these specialized (and expensive) services, the Committee discussed the idea of establishing a cost-center mechanism to support them, or at least to evaluate this option. Thus, a viable CF model which provides both core and expanded/specialized services, would increase the CF's overall value, and could be funded with a mixed overhead and cost-center model, if overhead alone can't match the needs. In such a model the provision of value through specialized services would be more directly and clearly linked to the science enabled, as documented in funding proposals for science using these services.

By contrast to CE, an expanded/specialized CF support model funded by a cost center is unlikely to depend on a single or just a few large projects. These services will most likely be used by a large swath of the CfA and if the price-point is right, funded by a large number of relatively small contributions, whose cost for the PIs should be easily integrated in future funding requests. A base support of these services might have to be funded by federal, overhead or other discretionary funds to help those who may lack external funding (like graduate students, pre-docs, junior post-docs, etc.)

# Appendix: Survey Results

**1 Would you be interested in having training sessions offered at the CfA on any of following subjects (select all that apply):**

| | | | |
|---|---|---|---|
| Open source software languages and modules (Python, R, Julia): | 78 | 16.7% | ( 70.3%) |
| Code collaboration/version control tools (GitHub) : | 53 | 11.3% | ( 47.7%) |
| Containers (Docker etc) and virtual environments (Conda) : | 44 | 9.4% | ( 39.6%) |
| Cloud computing/storage/services : | 44 | 9.4% | ( 39.6%) |
| Intro to/advanced topics in HPC (Hydra, Odyssey, Pleiades) : | 42 | 9.0% | ( 37.8%) |
| GPU/CUDA/etc. : | 40 | 8.5% | ( 36.0%) |
| Databases: SQL, NoSQL, other DB tools : | 38 | 8.1% | ( 34.2%) |
| Intro to/advanced topics in Unix/Linux : | 34 | 7.3% | ( 30.6%) |
| Open source development practices : | 33 | 7.1% | ( 29.7%) |
| Commercial software packages (Matlab, IDL, etc.) : | 31 | 6.6% | ( 27.9%) |
| Field Programmable Gate Arrays : | 16 | 3.4% | ( 14.4%) |
| none : | 11 | 2.4% | ( 9.9%) |
| Need webcast as we are in Hawaii : | 1 | 0.2% | ( 0.9%) |
| web design for dissemination of results and data : | 1 | 0.2% | ( 0.9%) |
| citing code, archiving code : | 1 | 0.2% | ( 0.9%) |
| idl - graph plotting : | 1 | 0.2% | ( 0.9%) |

Answered = 468 out of 111 or 421.6%

----------------------------------------------------------------------

**2 Do you presently use or anticipate using high-performance computing resources (e.g. Hydra, Odyssey, Pleiades, Cloud computing, etc.) during the next three years?**

| | | | |
|---|---|---|---|
| Yes : | 52 | 46.8% | ( 46.8%) |
| no : | 34 | 30.6% | ( 30.6%) |
| Maybe : | 25 | 22.5% | ( 22.5%) |

Answered = 111 out of 111 or 100.0%

----------------------------------------------------------------------

**3 What type of code do you typically run or foresee running?**

| | | | |
|---|---|---|---|
| Serial : | 45 | 31.2% | ( 40.5%) |
| Multi-Threaded (MT: OpenMP, etc.) : | 35 | 24.3% | ( 31.5%) |
| Distributed w/ message passing (MPI: MPICH/MVAPICH/OpenMPI) : | 32 | 22.2% | ( 28.8%) |
| GPU (CUDA/OpenACC) : | 21 | 14.6% | ( 18.9%) |
| Hybrid (both MT and MPI) : | 11 | 7.6% | ( 9.9%) |

Answered = 144 out of 111 or 129.7%

---------------------------------------------------------------------

## 4 Do you have an estimate on how many CPU cores, and for how long, you would typically need?

| | | | |
|---|---|---|---|
| no | : | 7 | 15.9% ( 6.3%) |
| 1,000-10,000 CPU, 24 hour wall times | : | 1 | 2.3% ( 0.9%) |
| no. Still exploring possibilities | : | 1 | 2.3% ( 0.9%) |
| 16, typically about several weeks | : | 1 | 2.3% ( 0.9%) |
| 1024 for a month | : | 1 | 2.3% ( 0.9%) |
| 4-8 cores, 4-12 hours | : | 1 | 2.3% ( 0.9%) |
| 100 cores for a week or so at a time | : | 1 | 2.3% ( 0.9%) |
| 170,000 cpu hours | : | 1 | 2.3% ( 0.9%) |
| 12 - 24 for several hours per job | : | 1 | 2.3% ( 0.9%) |
| I'm a heavy user and typically use something like 500-100 cores) | : | 1 | 2.3% ( 0.9%) |
| 2000 cores for 40 days | : | 1 | 2.3% ( 0.9%) |
| ~10,000 core-hours | : | 1 | 2.3% ( 0.9%) |
| 32 cores, 100 hours | : | 1 | 2.3% ( 0.9%) |
| Varies a lot. | : | 1 | 2.3% ( 0.9%) |
| 8 cores. 3 - 6 days. | : | 1 | 2.3% ( 0.9%) |
| Unclear. In what span? Typically 5-10 at a time | : | 1 | 2.3% ( 0.9%) |
| 64 on timescale of a week to a month or so | : | 1 | 2.3% ( 0.9%) |
| 4-6 cores, upto 1 day/run | : | 1 | 2.3% ( 0.9%) |
| typically 500 for ~3 days | : | 1 | 2.3% ( 0.9%) |
| 500 cores for 6 hours | : | 1 | 2.3% ( 0.9%) |
| a few hundred cores for a few days | : | 1 | 2.3% ( 0.9%) |
| 500, 3-6 months | : | 1 | 2.3% ( 0.9%) |
| 1? | : | 1 | 2.3% ( 0.9%) |
| 500 CPU / year with 2300 hours each over the next 5 years | : | 1 | 2.3% ( 0.9%) |
| 36 cores 5 days | : | 1 | 2.3% ( 0.9%) |
| 500 Cpus for a week | : | 1 | 2.3% ( 0.9%) |
| 100 for 2 days every 3 months | : | 1 | 2.3% ( 0.9%) |
| Don't know | : | 1 | 2.3% ( 0.9%) |
| 250-500 cpu-days/day | : | 1 | 2.3% ( 0.9%) |
| 50 | : | 1 | 2.3% ( 0.9%) |
| probably 100 cores for a day | : | 1 | 2.3% ( 0.9%) |
| Would need less than 5 for sure. | : | 1 | 2.3% ( 0.9%) |
| Typically, perhaps ~1000 cores over a week for a production | : | 1 | 2.3% ( 0.9%) |
| 1 million CPU hours | : | 1 | 2.3% ( 0.9%) |
| 1000 cpus / 8h | : | 1 | 2.3% ( 0.9%) |
| 2048 for 12 hours per run | : | 1 | 2.3% ( 0.9%) |
| 100 cores/1-3 months | : | 1 | 2.3% ( 0.9%) |
| I think the most that I've used on Hydra is 64 cores | : | 1 | 2.3% ( 0.9%) |

Answered = 44 out of 111 or 39.6%

----------------------------------------------------------------------

### 5 Do you have an estimate on how many GPU/CUDA cores, and for how long, you would typically need?

| | | | |
|---|---|---|---|
| no | : | 7 | 25.0% ( 6.3%) |
| n/a | : | 5 | 17.9% ( 4.5%) |
| not sure | : | 2 | 7.1% ( 1.8%) |
| 0 | : | 2 | 7.1% ( 1.8%) |
| no, I am figuring out the requirements as I learn more about GPUs | : | 1 | 3.6% ( 0.9%) |
| 4-GPU nodes, for 10K hours | : | 1 | 3.6% ( 0.9%) |
| 1 gpu / 48h | : | 1 | 3.6% ( 0.9%) |
| Typically 5-10 at a time, for a week or so each | : | 1 | 3.6% ( 0.9%) |
| Never used GPUs | : | 1 | 3.6% ( 0.9%) |
| As many as possible,~ hour wall times | : | 1 | 3.6% ( 0.9%) |
| I've never used GPU/CUDA. I'm not sure if I will want to | : | 1 | 3.6% ( 0.9%) |
| 100sims x 4cores x 2hours = 800 GPU hours | : | 1 | 3.6% ( 0.9%) |
| ? | : | 1 | 3.6% ( 0.9%) |
| 16 for 12 hours per run | : | 1 | 3.6% ( 0.9%) |
| a few cores for of order a week or so at a time | : | 1 | 3.6% ( 0.9%) |
| no plans, but interested | : | 1 | 3.6% ( 0.9%) |

Answered = 28 out of 111 or 25.2%

----------------------------------------------------------------------

### 6 Do you have an estimate on how much active disk storage, and for how long, you may need?

| | | | |
|---|---|---|---|
| no | : | 5 | 13.5% ( 4.5%) |
| 30-50TB, permanent (purchased w/ own funds) | : | 1 | 2.7% ( 0.9%) |
| Tens of TB | : | 1 | 2.7% ( 0.9%) |
| 5Tb | : | 1 | 2.7% ( 0.9%) |
| Minimal (a few dozen Gigabytes) | : | 1 | 2.7% ( 0.9%) |
| 3-6 months | : | 1 | 2.7% ( 0.9%) |
| More than 50 TB forever (the outputs from weather and chemical models) | : | 1 | 2.7% ( 0.9%) |
| up to 20 TB | : | 1 | 2.7% ( 0.9%) |
| a few terabytes | : | 1 | 2.7% ( 0.9%) |
| As much as I can get - calculations can require up to a TB per dump | : | 1 | 2.7% ( 0.9%) |
| 10 TB of node-local storage | : | 1 | 2.7% ( 0.9%) |
| ~1TB, 2 weeks | : | 1 | 2.7% ( 0.9%) |
| ~1TB | : | 1 | 2.7% ( 0.9%) |
| 10 TB for 2 years | : | 1 | 2.7% ( 0.9%) |
| 100sims x 100snaps x 10GB = 10 TB | : | 1 | 2.7% ( 0.9%) |
| 500 GB for a few years | : | 1 | 2.7% ( 0.9%) |
| I currently have 400 TB | : | 1 | 2.7% ( 0.9%) |
| 5 Tby, for years | : | 1 | 2.7% ( 0.9%) |

| | | | |
|---|---|---|---|
| 1T | : | 1 | 2.7% ( 0.9%) |
| 5 TB would be adequate, I think for my work. | : | 1 | 2.7% ( 0.9%) |
| 200 GB, few weeks | : | 1 | 2.7% ( 0.9%) |
| 10 TB for a week or two at a time | : | 1 | 2.7% ( 0.9%) |
| 1TB, upto 2days | : | 1 | 2.7% ( 0.9%) |
| Just as a ballpark estimate, I'd say something like 1 TB | : | 1 | 2.7% ( 0.9%) |
| 2TB for two months | : | 1 | 2.7% ( 0.9%) |
| ~180TB 5 years | : | 1 | 2.7% ( 0.9%) |
| 20 TB | : | 1 | 2.7% ( 0.9%) |
| ~1 Tb | : | 1 | 2.7% ( 0.9%) |
| ~ 1 TB for a month or so | : | 1 | 2.7% ( 0.9%) |
| 1TB for short term. data removed to local disk after. | : | 1 | 2.7% ( 0.9%) |
| ~4 TB | : | 1 | 2.7% ( 0.9%) |
| 1-2 TB, 5-7 days | : | 1 | 2.7% ( 0.9%) |
| 25-50 TB | : | 1 | 2.7% ( 0.9%) |

Answered = 37 out of 111 or 33.3%

----------------------------------------------------------------------

### 7 Will you want/need to archive the results? If so, how much and for how long?

| | | | |
|---|---|---|---|
| no | : | 5 | 14.3% ( 4.5%) |
| forever | : | 2 | 5.7% ( 1.8%) |
| yes | : | 1 | 2.9% ( 0.9%) |
| yes, permanent | : | 1 | 2.9% ( 0.9%) |
| 3 years | : | 1 | 2.9% ( 0.9%) |
| Up to now, I've archived my data on my own external disks | : | 1 | 2.9% ( 0.9%) |
| currently archive 25TB in Amazon S3, flushing to Amazon Glacier | : | 1 | 2.9% ( 0.9%) |
| not sure | : | 1 | 2.9% ( 0.9%) |
| the current archive is a SQL database | : | 1 | 2.9% ( 0.9%) |
| 100 TB for 5-10 years | : | 1 | 2.9% ( 0.9%) |
| ~10 TB | : | 1 | 2.9% ( 0.9%) |
| not necessary currently | : | 1 | 2.9% ( 0.9%) |
| Indefinitely | : | 1 | 2.9% ( 0.9%) |
| no archiving | : | 1 | 2.9% ( 0.9%) |
| n/a | : | 1 | 2.9% ( 0.9%) |
| Mass compression on final results, so not much to store long term | : | 1 | 2.9% ( 0.9%) |
| only locally | : | 1 | 2.9% ( 0.9%) |
| As long as possible | : | 1 | 2.9% ( 0.9%) |
| no. Results should be compact. | : | 1 | 2.9% ( 0.9%) |
| yes, for as long as my phd endures | : | 1 | 2.9% ( 0.9%) |
| ~5 years | : | 1 | 2.9% ( 0.9%) |
| yes, just for a few days | : | 1 | 2.9% ( 0.9%) |
| Results archived back to CfA on tape/USB disk/NetApp storage | : | 1 | 2.9% ( 0.9%) |
| no need to archive | : | 1 | 2.9% ( 0.9%) |

| | | |
|---|---|---|
| Less than one year. | 1 | 2.9% ( 0.9%) |
| Probably. Results from these sims could be used for 3-5 year | 1 | 2.9% ( 0.9%) |
| ~1-5 Tb, long term storage. | 1 | 2.9% ( 0.9%) |
| Yes, less than a year | 1 | 2.9% ( 0.9%) |
| <1 year | 1 | 2.9% ( 0.9%) |
| 14TB 3 years | 1 | 2.9% ( 0.9%) |

Answered = 35 out of 111 or 31.5%

--------------------------------------------------------------------

### 8 What class of jobs do you typically run, or foresee running?

| | | |
|---|---|---|
| Data Reduction or Analysis/Image Processing/etc.. | 48 | 37.5% ( 43.2%) |
| Modelization with parameter sweeps | 33 | 25.8% ( 29.7%) |
| Tightly Coupled Parallel Jobs/Simulations | 27 | 21.1% ( 24.3%) |
| Visualization | 17 | 13.3% ( 15.3%) |
| MCMC codes for model parameter inference | 1 | 0.8% ( 0.9%) |
| Quantum chemical calculations | 1 | 0.8% ( 0.9%) |
| Catalog generation | 1 | 0.8% ( 0.9%) |

Answered = 128 out of 111 or 115.3%

--------------------------------------------------------------------

### 9 What do you see as typical bottlenecks or challenges?

| | | |
|---|---|---|
| Processor Speed | 29 | 17.0% ( 26.1%) |
| Number of Processors | 28 | 16.4% ( 25.2%) |
| Memory Size | 28 | 16.4% ( 25.2%) |
| Storage: available space | 28 | 16.4% ( 25.2%) |
| Storage: IO speed/bandwidth (between nodes and disk) | 25 | 14.6% ( 22.5%) |
| Poor Accessibility/Usability | 7 | 4.1% ( 6.3%) |
| Insufficient Resource Allocation/Resource Contention | 7 | 4.1% ( 6.3%) |
| Network: IO speed/bandwidth between nodes | 6 | 3.5% ( 5.4%) |
| Processor Type | 5 | 2.9% ( 4.5%) |
| Network: IO speed/bandwidth between CfA and the cluster | 5 | 2.9% ( 4.5%) |
| manpower | 1 | 0.6% ( 0.9%) |
| Slow network connections outside CfA | 1 | 0.6% ( 0.9%) |
| user's fear of failure | 1 | 0.6% ( 0.9%) |

Answered = 171 out of 111 or 154.1%

--------------------------------------------------------------------

### 10 Do you wish to use unique or novel hardware/technologies?

| | | |
|---|---|---|
| no | 60 | 87.0% ( 54.1%) |
| Yes | 9 | 13.0% ( 8.1%) |

Answered = 69 out of 111 or 62.2%

---------------------------------------------------------------------

**11 If yes, which one(s)?**

| | | | |
|---|---|---|---|
| Advanced GPU and/or MIC offloading | 1 | 20.0% | ( 0.9%) |
| Autodesk Maya distributed rendering | 1 | 20.0% | ( 0.9%) |
| n/a | 1 | 20.0% | ( 0.9%) |
| distributed databases | 1 | 20.0% | ( 0.9%) |
| GPU enhanced code, which isn't really novel anymore | 1 | 20.0% | ( 0.9%) |

Answered = 5 out of 111 or 4.5%

---------------------------------------------------------------------

**12 Do you use a HPC cluster?**

| | | | |
|---|---|---|---|
| Yes | 44 | 57.1% | ( 39.6%) |
| no | 33 | 42.9% | ( 29.7%) |

Answered = 77 out of 111 or 69.4%

---------------------------------------------------------------------

**13 If yes, which one(s)?**

| | | | |
|---|---|---|---|
| Hydra | 18 | 40.9% | ( 16.2%) |
| Odyssey | 10 | 22.7% | ( 9.0%) |
| Pleiades | 2 | 4.5% | ( 1.8%) |
| outside of CfA (NASA Pleiades) | 1 | 2.3% | ( 0.9%) |
| Hydra, Pleiades | 1 | 2.3% | ( 0.9%) |
| Tycho at Paris Observatory | 1 | 2.3% | ( 0.9%) |
| El Gato (Arizona), Odyssey, Summit (DOE) | 1 | 2.3% | ( 0.9%) |
| NERSC | 1 | 2.3% | ( 0.9%) |
| NASA discover | 1 | 2.3% | ( 0.9%) |
| It's a cluster being built for a specific project | 1 | 2.3% | ( 0.9%) |
| AWS | 1 | 2.3% | ( 0.9%) |
| n/a | 1 | 2.3% | ( 0.9%) |
| Hydra, Odyssey | 1 | 2.3% | ( 0.9%) |
| Pleiades, Grace (Yale HPC), Wiluna | 1 | 2.3% | ( 0.9%) |
| Hydra, Pleiades, XSEDE | 1 | 2.3% | ( 0.9%) |
| Hydra and Pleiades | 1 | 2.3% | ( 0.9%) |
| Odyssey, BRC/savio, TACC/Stampede | 1 | 2.3% | ( 0.9%) |

Answered = 44 out of 111 or 39.6%

---------------------------------------------------------------------

**14 Have you tried using a HPC cluster and could you use it?**

| | | | |
|---|---|---|---|
| Yes, I used Hydra and/or Odyssey with no problem | 31 | 70.5% | ( 27.9%) |
| no, I never tried to use either Hydra or Odyssey | 6 | 13.6% | ( 5.4%) |
| Yes, I tried using Hydra but could not make use of it | 5 | 11.4% | ( 4.5%) |
| Yes, I tried using Odyssey but could not make use of it | 2 | 4.5% | ( 1.8%) |

Answered = 44 out of 111 or 39.6%

----------------------------------------------------------------------

### 15 What would be needed to make that/these HPC cluster(s) more useful for you? (check all that apply)

| | | | |
|---|---|---|---|
| more cores per job (MPI) | 11 | 14.3% | ( 9.9%) |
| more storage (scratch disk space) | 10 | 13.0% | ( 9.0%) |
| more memory (per node or per job) | 10 | 13.0% | ( 9.0%) |
| better support (staff to help run or install s/w) | 10 | 13.0% | ( 9.0%) |
| faster storage (bandwidth between nodes and disk) | 8 | 10.4% | ( 7.2%) |
| different time allocation scheme | 5 | 6.5% | ( 4.5%) |
| more cores per node (MT), | 5 | 6.5% | ( 4.5%) |
| faster network (bandwidth between nodes) | 5 | 6.5% | ( 4.5%) |
| faster network (bandwidth between CfA and cluster) | 3 | 3.9% | ( 2.7%) |
| better documentation | 1 | 1.3% | ( 0.9%) |
| I did not use Hydra as my time on Pleiades was sufficient | 1 | 1.3% | ( 0.9%) |
| It works well. Skylake and newer Intel cores are preferable | 1 | 1.3% | ( 0.9%) |
| Support staff to handle hardware issues | 1 | 1.3% | ( 0.9%) |
| Remote visualization for large data with e.g. Paraview | 1 | 1.3% | ( 0.9%) |
| More GPUs | 1 | 1.3% | ( 0.9%) |
| Node-local NVMes and GPUs | 1 | 1.3% | ( 0.9%) |
| Interactive jobs on Hydra | 1 | 1.3% | ( 0.9%) |
| Outdated libraries/dependencies | 1 | 1.3% | ( 0.9%) |
| Hydra needs to be better configured for MPI jobs like simulations | 1 | 1.3% | ( 0.9%) |

Answered = 77 out of 111 or 69.4%

----------------------------------------------------------------------

### 16 Does your research require that you distribute web-based data products or services to people outside of CfA?

| | | | |
|---|---|---|---|
| Yes | 65 | 58.6% | ( 58.6%) |
| no | 46 | 41.4% | ( 41.4%) |

Answered = 111 out of 111 or 100.0%

----------------------------------------------------------------------

### 17 If yes, what do you need to distribute your data products? (check all that apply):

| | | | |
|---|---|---|---|
| Static pages and server (from a set of files or web pages) | 48 | 39.7% | ( 43.2%) |
| Dynamic pages running client side scripts (e.g. javascript) | 26 | 21.5% | ( 23.4%) |

| | | | |
|---|---|---|---|
| Other tools or services (e.g. Zenodo, GitHub, etc) | : 23 | 19.0% | ( 20.7%) |
| Dynamic service running server side scripts (e.g. php, etc.) | : 23 | 19.0% | ( 20.7%) |
| not my research, but CXC distributes multiple types of software | : 1 | 0.8% | ( 0.9%) |

Answered = 121 out of 111 or 109.0%

---------------------------------------------------------------------

**18 Have you encountered problems making your data products available to people outside the CfA via one of our CfA-based websites?**

| | | | |
|---|---|---|---|
| no | : 81 | 76.4% | ( 73.0%) |
| Yes | : 25 | 23.6% | ( 22.5%) |

Answered = 106 out of 111 or 95.5%

---------------------------------------------------------------------

**19 If you answered 'Yes' to the previous questions please describe**

- rsync to Odyssey ITC space is clumsy with two-factor authentication.
- Issues with IT security - challenge to meet requirements without support.
- cannot run python scripts
- Need to be static (no js/ph allowed)
- intermittent problems with ftp. ftp size limit
- Understanding and managing a web app directly from Odyssey.
- Limited disk space on public server, not allowed to set up private server
- Using cloud server to distribute public data archives
- Cumbersome web sites and emailing (via gmail)
- I haven't even tried in general.
- Security issues having to deal with web hosting data
- n/a
- OIR TDC distributes raw and processed images and spectra to all users of our instruments
- restricted CF rules -- have to use cloud hosting
- Storage space, restrictions on use of scripting
- Difficult to share large (>10 GB) amounts of data with collaborators.
- storage size
- not enough disk space, concerns about archival reliability
- Usually for security check reasons
- I wanted to make a page with server side scripting to dynamically retrieve data but I could not figure out how best to set that up.
- I have been using the CF twiki to collaborate with people outside the cfa. It works, but it somewhat clumsy and took a substantial amount of time to coordinate.
- Web servers move behind firewalls from time to time.
- The current rules do not allow hosting browsable data on CF managed servers. We have to buy our own server put it in DMZ and maintain it. Some groups were given access to

Google space bought by the institute. This is not a common knowledge and is unclear what are the limits if one wants to put a virtual machine for their project.

----------------------------------------------------------------------

**20 If you distribute web-based data products or services to people outside of CfA by other means than the CfA-based websites, what are you using?**

| | | |
|---|---|---|
| Cloud-based services or external repositories (e.g. AWS) | : 22 | 42.3% ( 19.8%) |
| Asked colleagues in other institutions to host it for you | : 9 | 17.3% ( 8.1%) |
| Dedicated project server at CfA (in the DMZ) | : 7 | 13.5% ( 6.3%) |
| Odyssey community webserver | : 1 | 1.9% ( 0.9%) |
| Odyssey ITC disk space | : 1 | 1.9% ( 0.9%) |
| anonymous ftp site | : 1 | 1.9% ( 0.9%) |
| Personally paid-for website | : 1 | 1.9% ( 0.9%) |
| both cloud-based services and colleagues at other institutions | : 1 | 1.9% ( 0.9%) |
| attachments | : 1 | 1.9% ( 0.9%) |
| Don't so distribute | : 1 | 1.9% ( 0.9%) |
| CDAWeb | : 1 | 1.9% ( 0.9%) |
| hosted in Hawaii with UH network distribution | : 1 | 1.9% ( 0.9%) |
| VM on Odyssey. Paul Edmon helped. | : 1 | 1.9% ( 0.9%) |
| Harvard Dataverse | : 1 | 1.9% ( 0.9%) |
| NASA data archives | : 1 | 1.9% ( 0.9%) |
| Google Drive | : 1 | 1.9% ( 0.9%) |
| Paid for hosting, github | : 1 | 1.9% ( 0.9%) |

Answered = 52 out of 111 or 46.8%

----------------------------------------------------------------------

**21 Which cloud-based services or external repositories do you use, if any?**

| | | |
|---|---|---|
| Dropbox | : 4 | 11.4% ( 3.6%) |
| GitHub | : 2 | 5.7% ( 1.8%) |
| google | : 1 | 2.9% ( 0.9%) |
| google drive, drop-box | : 1 | 2.9% ( 0.9%) |
| Dropbox, Amazon WS, Google Drive | : 1 | 2.9% ( 0.9%) |
| none | : 1 | 2.9% ( 0.9%) |
| none that I know of | : 1 | 2.9% ( 0.9%) |
| Zenodo | : 1 | 2.9% ( 0.9%) |
| google drive, dropbox | : 1 | 2.9% ( 0.9%) |
| Linode | : 1 | 2.9% ( 0.9%) |
| Google cloud | : 1 | 2.9% ( 0.9%) |
| aws | : 1 | 2.9% ( 0.9%) |
| Google drive, dropbox, onedrive | : 1 | 2.9% ( 0.9%) |
| Dropbox, Google Cloud | : 1 | 2.9% ( 0.9%) |
| Google Cloud, external php supported server | : 1 | 2.9% ( 0.9%) |
| Dropbox, Google Drive/Docs | : 1 | 2.9% ( 0.9%) |

| | | | |
|---|---|---|---|
| WAPS at CfA | : | 1 | 2.9% ( 0.9%) |
| Google Drive | : | 1 | 2.9% ( 0.9%) |
| AWS | : | 1 | 2.9% ( 0.9%) |
| Github, Google Drive, Dropbox | : | 1 | 2.9% ( 0.9%) |
| Private server at another institution | : | 1 | 2.9% ( 0.9%) |
| Github | : | 1 | 2.9% ( 0.9%) |
| none. | : | 1 | 2.9% ( 0.9%) |
| NASA's GES DISC | : | 1 | 2.9% ( 0.9%) |
| Google drive. | : | 1 | 2.9% ( 0.9%) |
| github and bitbucket for code. | : | 1 | 2.9% ( 0.9%) |
| Dropbox, GitHub, Google Cloud | : | 1 | 2.9% ( 0.9%) |
| Before switching to our own server we used Microsoft Azure | : | 1 | 2.9% ( 0.9%) |
| AWS S3(IA) | : | 1 | 2.9% ( 0.9%) |
| github, dropbox, want to use AWS | : | 1 | 2.9% ( 0.9%) |
| Google Drive | : | 1 | 2.9% ( 0.9%) |

Answered = 35 out of 111 or 31.5%

----------------------------------------------------------------------

**22 Does your research going forward require additional support (staffing, etc.) from that currently provided for distributing web-based data products?**

| | | | |
|---|---|---|---|
| no | : | 58 | 59.8% ( 52.3%) |
| Yes - hardware like server and disk space | : | 15 | 15.5% ( 13.5%) |
| Yes - staffing to implement/code service(s) | : | 12 | 12.4% ( 10.8%) |
| Yes - software not currently available | : | 7 | 7.2% ( 6.3%) |
| Relax rules on server side scripting | : | 1 | 1.0% ( 0.9%) |
| visualization wall & associated software | : | 1 | 1.0% ( 0.9%) |
| not sure | : | 1 | 1.0% ( 0.9%) |
| Security of project servers | : | 1 | 1.0% ( 0.9%) |
| SQL server, php | : | 1 | 1.0% ( 0.9%) |

Answered = 97 out of 111 or 87.4%

----------------------------------------------------------------------

**23 Would you be interested in a service at the CfA to support data archiving and data product curation?**

| | | | |
|---|---|---|---|
| Yes | : | 77 | 73.3% ( 69.4%) |
| no | : | 28 | 26.7% ( 25.2%) |

Answered = 105 out of 111 or 94.6%

----------------------------------------------------------------------

**24 What do you currently use for archiving and curating data products like datasets and code? (check all that apply)**

| | | | |
|---|---|---|---|
| Store final products locally | 72 | 39.1% | ( 64.9%) |
| Store final products on a different cloud-based platform | 45 | 24.5% | ( 40.5%) |
| Deposit final products and code releases in external repo | 35 | 19.0% | ( 31.5%) |
| Host final products on a self-managed website | 32 | 17.4% | ( 28.8%) |

Answered = 184 out of 111 or 165.8%

---------------------------------------------------------------------

### 25 How long do you currently plan to make your data products available?

| | | | |
|---|---|---|---|
| in perpetuity | 53 | 52.0% | ( 47.7%) |
| 1-5 years | 28 | 27.5% | ( 25.2%) |
| 6-10 years | 21 | 20.6% | ( 18.9%) |

Answered = 102 out of 111 or 91.9%

---------------------------------------------------------------------

### 26 What services would you like to see more information about on the CF website? (check all that apply)

| | | | |
|---|---|---|---|
| Smithsonian computational facilities and access to them | 58 | 30.5% | ( 52.3%) |
| Harvard computational facilities and access to them | 46 | 24.2% | ( 41.4%) |
| Eduroam | 45 | 23.7% | ( 40.5%) |
| CfA wifi | 35 | 18.4% | ( 31.5%) |
| none | 1 | 0.5% | ( 0.9%) |
| really UPDATED CF help page about ALL IT services | 1 | 0.5% | ( 0.9%) |
| VPN setup | 1 | 0.5% | ( 0.9%) |
| Remote desktop software | 1 | 0.5% | ( 0.9%) |
| Google cloud platform supported software packages | 1 | 0.5% | ( 0.9%) |
| Project specific CF support policy | 1 | 0.5% | ( 0.9%) |

Answered = 190 out of 111 or 171.2%

---------------------------------------------------------------------

### 27 What is your main computer OS?

| | | | |
|---|---|---|---|
| MacOS | 53 | 47.7% | ( 47.7%) |
| Linux | 51 | 45.9% | ( 45.9%) |
| MS Windows | 6 | 5.4% | ( 5.4%) |
| Fortran | 1 | 0.9% | ( 0.9%) |

Answered = 111 out of 111 or 100.0%

---------------------------------------------------------------------

### 28 Is your main computer

| | | | |
|---|---|---|---|
| Self-managed | 63 | 56.8% | ( 56.8%) |
| Managed by the CF | 29 | 26.1% | ( 26.1%) |

```
Managed by HEA-sys                    :  16  14.4% ( 14.4%)
Managed by the SSXG in HEAD           :   1   0.9% (  0.9%)
HUIT                                  :   1   0.9% (  0.9%)
My son                                :   1   0.9% (  0.9%)

Answered = 111 out of 111 or 100.0%
-------------------------------------------------------------------
```

**29 If self-managed, why?**

- I think Mac laptops are all self-managed.
- Most efficient way
- Faster to sudo everything, not deal with downtime, unlimited customizability, etc.
- Because the CF won't manage Macs
- Graduate student using personal laptop
- no other option
- Laptop
- Mac not supported so far as I know; sometimes need root access to install software.
- Because I don't need someone to manage my OS.
- ... why not? I don't understand the question.
- I have found it very difficult to manage my CfA machine, particularly when adding new software, due to all the extra security.
- Macbook bought on grant money.
- Simplicity of having complete control over my system
- CF could not provide the service and flexibility I need
- There weren't funds for me to have a new computer.
- control of software etc
- I occasionally need software and services not supported by the CF
- in Hawaii, no CF managed available
- have complete and immediate control
- Self-owned therefore self-managed
- Why would I want to spend valuable research time managing my own machine?
- not supported by CF
- Mac
- I think that's the only choice with MacOS
- I didn't know there was any other option
- Flexibility, convenience
- It is easier as transportable.
- I like to be able to install new programs and tools as I discover them
- I use macOS with a package manager (MacPorts) for open-source software. It's easy to install anything I need right away.
- flexibility
- because it better fits my projects / pattern of usage
- Laptop
- macOS

- To allow for self management of software
- Need root access to install many necessary codes
- Was told it had to be
- I have a personal OSx machine and understand it, and prefer to not have external manipulation.
- Mac
- Didn't know about CF management.
- what do you mean... why not manage my own laptop
- Ability to try out new software quickly
- HEAD Macs are all self-managed.
- I often need root and am an experienced sysadmin
- Laptop
- want to be able to take it home, easier to set up the code that I want, much easier to get and maintain access, familiarity with MacOS
- historically always been so, though CF assists when crisis mode occurs
- It's a Mac
- I always have from an era when macs were officially unsupported. For all I know, they still are.
- Mac laptop
- laptop
- easiest
- more flexible to configure without CF process or gatekeeping
- Control, stability.

----------------------------------------------------------------------

**30 Do you use/need/want to use:**

| | | | |
|---|---|---|---|
| cloud storage (on large scale) | : 48 | 56.5% | ( 43.2%) |
| cloud computing | : 36 | 42.4% | ( 32.4%) |
| databases | : 1 | 1.2% | ( 0.9%) |

Answered = 85 out of 111 or 76.6%

----------------------------------------------------------------------

**31 What software do you use to do visualization?**

| | | | |
|---|---|---|---|
| Python modules (e.g. yt, seaborn, bokeh, etc.) | : 73 | 40.8% | ( 65.8%) |
| IDL | : 34 | 19.0% | ( 30.6%) |
| PGPlot | : 13 | 7.3% | ( 11.7%) |
| Javascript libraries (D3, React, Plotly, etc) | : 12 | 6.7% | ( 10.8%) |
| Matlab | : 9 | 5.0% | ( 8.1%) |
| Mathematica | : 6 | 3.4% | ( 5.4%) |
| R libraries (e.g. ggplot2, shiny, etc.) | : 2 | 1.1% | ( 1.8%) |
| Origin | : 2 | 1.1% | ( 1.8%) |
| ds9 | : 2 | 1.1% | ( 1.8%) |
| gnuplot, etc... | : 1 | 0.6% | ( 0.9%) |

| | | |
|---|---|---|
| sm, ds9 | : 1 | 0.6% ( 0.9%) |
| ds9 and self-written code | : 1 | 0.6% ( 0.9%) |
| TecPlot, Paraview, ViSiT | : 1 | 0.6% ( 0.9%) |
| gnuplot | : 1 | 0.6% ( 0.9%) |
| Tecplot | : 1 | 0.6% ( 0.9%) |
| Panoply (from NASA GISS), Gnuplot | : 1 | 0.6% ( 0.9%) |
| Julia | : 1 | 0.6% ( 0.9%) |
| Custom tools | : 1 | 0.6% ( 0.9%) |
| CIAO, ds9 | : 1 | 0.6% ( 0.9%) |
| PlPlot | : 1 | 0.6% ( 0.9%) |
| Adobe's Dreamweaver | : 1 | 0.6% ( 0.9%) |
| older, out of date packages | : 1 | 0.6% ( 0.9%) |
| Autodesk Maya | : 1 | 0.6% ( 0.9%) |
| TOPCAT / STILTS / ds9 / CDS Aladin | : 1 | 0.6% ( 0.9%) |
| sm, VisIt | : 1 | 0.6% ( 0.9%) |
| mathcad | : 1 | 0.6% ( 0.9%) |
| chips, sm | : 1 | 0.6% ( 0.9%) |
| Don't know name | : 1 | 0.6% ( 0.9%) |
| own software | : 1 | 0.6% ( 0.9%) |
| CASA | : 1 | 0.6% ( 0.9%) |
| super mongo | : 1 | 0.6% ( 0.9%) |
| ROOT | : 1 | 0.6% ( 0.9%) |
| Fortran | : 1 | 0.6% ( 0.9%) |
| ncview | : 1 | 0.6% ( 0.9%) |
| supermongo | : 1 | 0.6% ( 0.9%) |

Answered = 179 out of 111 or 161.3%

--------------------------------------------------------------------

### 32 Where do you do your visualization?

| | | |
|---|---|---|
| Desktop/laptop | : 102 | 85.0% ( 91.9%) |
| HPC resource | : 12 | 10.0% ( 10.8%) |
| Cloud | : 4 | 3.3% ( 3.6%) |
| self-managed server, desktop too old and slow for most jobs | : 1 | 0.8% ( 0.9%) |
| explicit virtual machines for data reduction onsite | : 1 | 0.8% ( 0.9%) |

Answered = 120 out of 111 or 108.1%

--------------------------------------------------------------------

### 33 Which of the following visualization tools would you be most likely to use?

| | | |
|---|---|---|
| Video wall | : 25 | 40.3% ( 22.5%) |
| Virtual reality | : 20 | 32.3% ( 18.0%) |
| 3D printing | : 14 | 22.6% ( 12.6%) |
| Provider of ordinary 2-D images | : 1 | 1.6% ( 0.9%) |

| | | | |
|---|---|---|---|
| personal laptop | : | 1 | 1.6% ( 0.9%) |
| web based visualization for outside collaborators | : | 1 | 1.6% ( 0.9%) |

Answered = 62 out of 111 or 55.9%

-------------------------------------------------------------------

**<u>34 What would you use these tools for?</u>**

- Outreach and scientific conferences presentations
- Outreach
- Walkthroughs of molecular clouds and star forming regions using modern simulations
- Presentations
- Explore and visualize complex data sets
- Research
- Generating effective ways to visualize large datasets
- I'm working with Yale HPC on making a virtual reality showing of my cosmological simulation. It would be great to present that in Cambridge as well!
- Data visualization
- Presentations
- work with non-SI collaborators in realtime
- viewing and analyzing observations
- Public exhibitions
- visualizing and analyzing output from simulations and images from large telescopes such as Subaru
- Supernova remnants
- data exploration, outreach
- navigation in multi-dimensional datasets
- Visualizing large data sets
- high resolution observations
- dissemination of science and projects (satellites etc)
- TBD.
- Presentations
- Talks and Outreach
- displaying large, high resolution imaging
- EPO
- If I do a 3D simulation, I could use virtual reality to view the simulation from inside and look around. However, a large computer monitor is usually sufficient.
- public outreach
- We have some large (gigapixel) 2D images to view and 3D images to fly through
- Educational outreach
- mechanical prototyping
- simulations

-------------------------------------------------------------------

**35 If you have any comments you want to make regarding Computational Services please write them below**

- Didn't the HPCship at CFA sail a long time ago?
- More services will require more resources (h/w, s/w and ppl)
- need better support for desktop computing. In particular, better access to CF disks such as via NFS mount.
- There is a great need to move some information from behind the CfA firewall. For example, much of the content on the CF webpage cannot be accessed unless you are on a CfA machine or accessed through the VPN. The issue is that to understand how to get on the VPN, you need to read the firewall protected webpage. Also, the CF password reset rate is way to short. I think 90% of the people I know in my research group just allow the password to expire. Then when they need to access a CF machine, they have to go down to the CF help desk and get a temp password. This cannot be the most efficient way to secure the CF network.
- We rely mainly on the CF for support, they are doing a great job but are resource-limited. It would be great to have more services from them, such as assistance in installing specialized software for data reduction/analysis that is open source but not currently installed on CF systems. Currently if we need to do that, we have to set up a self-managed system where we have root access and can customize it, but then don't have access to our data on the netapp disks.
- It would be nice to have people who could consult on more issues other than hardware or networking. For instance, someone who can answer simple questions about python usage. As an example, the person who wrote 'ds9' is in the building, and every couple of years or so I ask him a specific question, and get an immediate answer, thus saving me lots of time.
- Access to CfA machines and those in the DMZ is terribly slow from outside, using the VPN it may take 2 or 3 seconds for a single keystroke to get through.
- I'm very happy with the CF services that I receive from Tom R. and Bob B., and from Van himself!
- would be great if any educational opportunities are recorded for use in our timezone, or scheduled after 2PM EST
- Wishlist: backup of laptops; implementation of room scheduling in Google Calendar (not strictly scientific, I know).
- The SAO Computation Facility is well run and currently meets all my needs.
- It would be great to have IDL licenses available! Some data reduction software institute telescopes is in IDL.
- aside from the ugly new logo, I think the general services are ok. The availability of HPC at the CfA is really pathetic. One of the reasons I do not use Hydra is that I cannot get enough cpu cycles - on the NASA discover system, I have excellent throughput with little or no wait time.
- CF has been helpful when I have hardware or internet questions, but if I have any kind of software questions (including, but not limited to, various installed libraries, help installing additional libraries, and Python installation/interaction problems) I'm almost always told that they can't help me. It would be really useful to have *software* experts on hand to

help with these things, as well as access to multiple updated options for libraries/compilers.

- While I can see that the amount of storage allocated to each user on Hydra needs to be limited, the current limitations on the amount and length of time of storage are very confining. In particular the 180 day limit on data under /pool seems unnecessary.
- Would like to have institutional knowledge and support for cloud computing
- What is SI's future regarding desktops and software licensing? Adobe cloud site license would be nice (stand-alone so laptops can use them).
- none.
- CF managed linux machines have network mounted home directories. This leads to poor performance for custom analysis software that I need to install locally in my home directory. I don't think I have write access on local drives on each CF machine. Perhaps I am wrong about that.
- It'd be great to have the Julia language as a module on Hydra and other clusters, or perhaps some documentation on websites on how to use it.
- It is imperative that scientists at the CfA are provided with tools and support for distribution of their data products to the outside world in modern ways. It aligns with the Smithsonian mission.  Frankly not having this type of support is embarrassing for the institution of this size and reputation.
- biggest bottleneck by far for all my work is transferring files to Hydra.

# Appendix B: Data Repositories at the CfA (2016)

# Data Repositories at CfA

Pepi Fabbiano, Raffaele D’Abrusco, Arnold Rots, Tom Dame, Eric Keto, Warren Brown, Sean Moran, Ed DeLuca, Mark Weber, Trey Winter, Kathy Reeves, Alberto Accomazzi, Larry Rothman, Youli Gordon, Roman Kochanov, Josh Grindlay, Sylvain Korzennik, Daina Bouquin, Matt Ashby, Randall Smith, Ian Evans

## Introduction

The Harvard-Smithsonian Center for Astrophysics (CfA) is the world’s largest astronomical institution and produces a stream of world-class science. An indispensable part of this science is the production of an invaluable store of astronomical data across the spectrum. These data present CfA with both a scientific challenge and a technical one.

*Scientifically*, how do we maximize the world-leading science extracted from these data?
We believe that the entirety of CfA data holdings should be brought to world-leading standards (IVOA, SVO), to the benefit  of CfA astronomers, astronomers at large, and the health and reputation of CfA as a whole.

*Technically*, the amount and complexity of CfA data has grown dramatically. Industry recognizes that data management has become a complicated and specialized field that requires dedicated professional efforts to be a world-leader. The level of specialization required is a drag on individual CfA researchers, and so on the scientific performance of CfA.
We believe that the preservation of CfA-held data in permanent, accessible, and searchable archives can solve this challenge, and can do so for a modest investment of resources. This document explains how this can be done.

## Why Data Archives Are Important for 21st Century Astrophysics

For CfA to stay a world-leading institution, it needs to keep upping its game in the "data experience" provided to researchers, as well as to funding institutions, and to posterity. That upping of game requires a quality of support that is no longer reasonable nor desirable to shift onto researchers (or the research quality gets dragged down), and in fact requires more of dedicated professional support.

## 1. Enabling Great Science

CfA provides one of the premier locations for staff and visitors to do research, in part because of its richness of topics, data, researchers, and other resources. An important way to make the most of this environment is to make CfA's datasets as widely available and easily accessible as possible, and to ease the development and integration of new projects. Furthermore, CfA can be more than just the physical location from which independent teams host data archives and services. CfA can enable better science for the rest of the community, and more efficiently, as a well-managed portal for the missions and projects with which it is associated. But to do that better requires dedicated support.

## 2. Enabling Data Re-Use to Multiply the Value of each Dataset

Studies of the use of *Chandra* and *STScI* archives show that the value of data is increased greatly when it is made available to a wide variety of astrophysicists. At present most data at CfA are not easily discoverable and cross-searchable outside the nearest area of expertise. Yet data preservation and access is now mandated by our national funding agencies, NASA, NSF, as well as by the Smithsonian Institution. Making data more readily available enhances the reputation of the institutions that do so. In fact statistics on the use of such data archives is increasingly being used as a measure of the success of funded research.

## 3. Preserving Unique Datasets for Posterity

For both solar and astrophysics missions, NASA has an open data policy, and currently keeps an archive of post-mission datasets. But it is not part of NASA's vision to keep such archives indefinitely. Yet DASCH, for example, shows how century-old data sets can have great value in astronomy. Instead long-term preservation for posterity does fit under the Smithsonian's mandate. As CfA is a world leader in its involvement with a large number of astrophysical observatories and projects of all types and sizes, and as part of the Smithsonian, it would seem that our role in preserving and serving such datasets indefinitely for the benefit of the U.S. and all humanity is practically thrust upon us. But to do that better requires dedicated support.

## 4. Access to analysis and data mining tools

If data are available through established interfaces (e.g. IVOA, SVO), existing tools that conform to these interfaces can be employed in their scientific analysis, thus increasing the productivity of CfA scientists. We also advocate that the CfA, as a major astronomical data center, take a proactive stance regarding the exploitations of these data in the near and more distant future. This means that **data science** expertise at CfA should be considered and actively encouraged, to optimize the statistical exploration, visualization and analysis of possibly massive, complex and heterogeneous datasets. Data science requires a spectrum of cross-disciplinary skills and methodologies. Various groups at the CfA have developed some aspects of this expertise to tackle successfully specific data challenges concerning their field of research, e.g.:

- Time-resolved astronomy (DASCH)
- Spatially-resolved dataset (solar activity, solar features and their structures, etc)
- Techniques for multi-wavelength classification (SEDs)...

We advocate support for expertise in this area and for coordinating these efforts, to maximize the scientific return of astronomical research at the CfA and everywhere CfA data are used. This includes the development of sophisticated data mining and statistical tools for the analysis of increasing larger surveys, in which CfA scientists are and will be involved.

# Areas of Support

There is currently no CfA-wide policy, model, or support for the **development**, **maintenance**, and **preservation** of our data archives. Instead they are scattered and are often hard to find, or query. This is largely due to their widely varying levels of financial and technical support. To remedy this situation and to allow all CfA projects to comply with SI and other data policy mandates, we identify three main areas where support is urgently needed and should be included in the Strategic Plan.

## 1. Discoverability and Access

Like most prominent astronomical institutions, CfA needs an **archive portal** that provides the community access to all our data repositories. The objective of this portal is to provide a comprehensive overview of all data available from CfA and to make the data products searchable and discoverable. As a simple first step the website should contain a prominent page with links to the various data repository access pages that are already in existence. The next step would be the design of a common portal based on standard access protocols to the repositories, such as those provided by the IVOA.

## 2. Project Data Management Plans (PDMP)

As PDMPs are mandated for many if not all of our research projects, support for developing such plans is crucial. This may not be an issue for experienced project staff, but it can be overwhelming to those who have never prepared one before. CfA needs to provide support for PDMP development in the form of guidance, templates, and consultations. This is not a huge job, but such support is essential.

## 3. Project Support

There is considerable variety among the data repositories at CfA, in their scope, objectives, target audiences, size, resources, funding, and the expertise of their staff:

- **Large projects** generally have the resources and expertise to manage the development and maintenance of their archives. This is especially true for space-based missions, like Chandra, since they operate under NASA data management mandates. Still, there is the issue of who will take care of a project's archive after it has ended.
- **Small projects** with limited funding may need more support, but it will often be consultative in nature.
- A separate category are **user-contributed high-level scientific data products**, such as those that are included in peer-reviewed publications. The data volume is generally fairly small, but allowing them to be discovered is challenging.

One cannot expect most scientists to be able to fully specify and design a data repository; but targeted consultations on hardware configuration, system design, and curation principles go a long way to point the PI in the right direction. In general, the types of support fall in a number of distinct categories; which of these are needed by any specific project depends on the requirements and the expertise of its staff. It is mostly about **assistance**, not about "implementing systems." These categories are:

1. System design
2. Hardware configuration
3. Hardware acquisition
4. Enterprise level storage
5. Disaster recovery plan
6. Cloud computing expertise
7. Virtualization
8. Software tools needed
9. Databases
10. Interfaces
11. Policy and institutional requirements
12. Networking and firewall issues
13. General consultation
14. Disposition after project termination

Establishing **links between related datasets and between datasets and the literature** (the ADS being one of our repositories) is something that has gained prominence and will continue to gain in importance in a future where data mining and handling of Big Data are becoming part of standard research tools, as well as providing meaningful performance metrics. The principles and tools for this type of data repository tools will need to be detailed, too.

## 4 Data Science at CfA

While data science is not astrophysics, data science is and will be more and more needed to fully exploit the CfA data. We advocate that the CfA invest in a coordinated effort for developing and making available the more advanced  tools needed for the scientific exploitment of the data.

# Realizing the Support

The obvious question is where this support is supposed to come from and how it is to be funded. The good news is that much of the needed expertise and resources is already present at CfA:

1. The CF clearly has the technical expertise and could shift its focus to that type of support; this may also include access to SI resources
2. The library's expertise in organizing and disseminating information and resources is an obvious match
3. The existing community of data repositories, especially the better supported ones including the ADS, are well positioned to share their experience and expertise.

Some funding will be required. But we do not need to start from scratch and some resources are already available. For one thing, *improved communication and sharing within the CfA data repository community will help tremendously*. **But it has to be clear that responsible care for our data is a core part of the CfA mission, an essential part of the scientific endeavor, and recognized as such in the Strategic Plan.**

# Appendix
# Existing CfA Repositories

## Chandra Data Archive

The CDA is the mission archive of the Chandra X-ray Observatory, one of NASA's Great Observatory, launched in 1999. It is part of the Chandra X-ray Center (CXC), operated by SAO under contract with NASA. As such it is required to comply with NASA requirements which includes building and maintaining the data archive which, at the end of the mission, is to be transferred to the custody of HEASARC. The CDA encompasses the entire lifecycle of the Chandra observations, from proposals, through mission planning, to archived products. In general, observations remain proprietary for one year, then become public. After four reprocessing runs, the default version of the data products is the most recent one. The archive includes a source catalog (version 2 is in preparation) and the most complete bibliographic database currently in existence. That database links publications to datasets at high granularity, which allows us to generate innovative performance metrics.
The CDA provides access through a variety of (web) user interfaces and through standard IVOA-compliant protocols.
The current data volume is 27 TB. Two copies are kept, one at MIT (primary for operations), the other at CDP (primary for our users). We are pursuing the possibility of storing one "glacial"

backup copy in the cloud. The database management system is Sybase. The archive operations team consists of 7 people, the archive development team of 5 people.
The website is http://cxc.cfa.harvard.edu/cda

## Submillimeter Array Data Archive

The volume of data collected by the SMA since its commissioning in 2002 is 13 TB. Roughly half of these data were obtained in 2016 following commissioning of a new wide-band correlator. Going forward we expect a data rate of 30-50 TB/year. Complete copies of the archive reside in Hilo, Hawaii, and on the computers of the Radio Telescope Data Center (RTDC) in Cambridge. New data are transferred from the summit observatory site to Hilo to Cambridge twice per day. Non-proprietary data (older than 15 months) are directly downloadable using an RTDC search tool; proprietary data must be requested using an online form. Since 2008 the RTDC has also maintained an archive of data calibrated using the MIR software package and ported into MIRIAD format. As with most interferometric observatories, providing visibility data is not typically a smooth process for VO-compliant programs which tend to work with final image products. Especially for a sparse array such as the SMA, imaging is often a complicated process with tuning for specific science goals leading to very different outcomes with the same data. Whether and what VO-compliant data can be provided will be a subject for future consideration.

## OIR Telescope Data Center

The OIR Telescope Data Center maintains a repository of approximately 50 TB of raw and reduced data on CF-managed storage in Cambridge, with backups to Herndon. This grows nightly with automated transfers from MMT and Whipple Observatories; we accumulate data at a rate of a few TB per year. We also distribute data to PIs after it has been taken, and utilize an opening in the CF firewall to stage data and email download links to PIs. This system is cumbersome and due for replacement, but developing a replacement system conforming to modern standards is a challenge for our group. In 2016 we launched the OIR Science Archive website (oirsa.cfa.harvard.edu), a VO-compliant database which contains several hundred thousand spectra. This is hosted on a google cloud server set up by the CF; as we expand the public data on the site, we would welcome assistance in learning about performance scaling, best practices for site management, and how to handle our expanding need for cloud disk space.

## Solar Group (including NASA Missions such as SDO, Hinode, IRIS, Solar Probe Plus, etc.)

The Solar / Stellar X-Ray Group is involved (as the principal institution or as a partner) with instrument projects on several NASA missions. These projects' data solutions share some amounts of oversight, pooling of resources, and partial autonomy from CF and SysHelp, while also having unique needs and creating a heterogeneous environment. Across all mission projects currently there is about 400 TB of data storage space, including backups. SDO, by far, has the largest data flow: CfA ingests data from the SDOC at Stanford at a mean rate of about 0.5–0.6 TB/day, with peaks of up to 3 TB/day. At times, this same flow has been outgoing to other centers, as CfA is part of the global data distribution chain for SDO. CfA also acts as an SDO data center for the Virtual Solar Observatory (VSO), currently serving about 1–3 TB/month. As more typical examples of the other missions, Hinode/XRT peaks at around 25 GB/month incoming and 50–75 GB/month outgoing, and Hi-C data serves about 25 GB/month max. The primary long-term archive for SDO will be handled by the Stanford SDOC and NASA.  For other solar missions, the typical post-mission archival strategy is to provide final calibrated data files to NASA's Solar Data Analysis Center, with no further support to CfA to maintain an archive.

In addition to NASA missions, there are other notable data flows and data services provided by the Solar Group. Over the past couple of years, we have been involved in working with museums and other public sites to provide daily solar data to video wall exhibits. Other examples are projects to provide automatic feature-finding data products for SDO data, providing VSO data service (without a local archive) for the SWAP mission, and various catalogs of metadata and derived data products. These sorts of activities do not always fall under the umbrellas of either the main mission projects nor individual research grants.

The Solar Group has not experimented much with running archives or services in the cloud, although it is planned that future mission projects will host calibration software development and distribution on GitHub. Because the Solar Group has been largely permitted to control and handle its own data needs, the main logistical concerns when dealing with CF have been about firewalls, overly constraining SI security policies, and ample space/power for servers.

## NASA Astrophysics Data System (ADS)

The ADS is a disciplinary repository for bibliographic content in Astronomy and Astrophysics, although it also finds substantial use in the Physics and Geophysics communities.  In addition to its search capabilities, ADS tracks citations between papers, links to datasets associated with the publications, provides article-level metrics, and features personalized notification services.

The ADS literature database contains entries for more than eleven million items and includes the full-text for approximately four million of them.  Nearly every modern refereed article in Physics, Astronomy, or Geophysics is in the ADS full-text collection.  The ADS is more than just a prototypical bibliographic database, but rather it acts as an aggregator of scholarly resources relevant to Astrophysics research.  In addition to indexing content from a variety of publishers, the ADS provides a number of value-added services which enrich its collections in many different ways.  ADS collects and exposes links to external resources such as publishers and NASA data archives.  ADS enriches article metadata via text-mining of full-text sources (extracting references, acknowledgments, keywords). ADS incorporates bibliographies from institutes and archives. ADS generates and maintains citation and co-readership networks to support discovery and bibliometric services.

While the total amount of data stored by the ADS is modest in terms of bytes (approximately 2TB for the data served to the public), the complexity of its dataset far greater, since it is the result of the curated aggregation of a variety of resources from the literature and data archives. The total amount of data currently stored on our local servers is approximately 80TB. During the last three years, the ADS has moved the new generation of its public services to the cloud (Amazon Web Services), and plans to continue to grow its use of cloud-based computing over time given the flexibility of the platform and the freedom it provides to its developers.

## HITRAN Program

HITRAN is an acronym for *hi*gh-resolution *tran*smission molecular absorption database. HITRAN is a compilation of spectroscopic parameters that a variety of computer codes uses to predict and simulate the transmission and emission of light in gaseous media. HITRAN has a great many applications: remote sensing of the terrestrial atmosphere, characterization of planetary, exoplanet, and stellar atmospheres, monitoring of environmental quality, industrial control, etc.

We have developed several tools to use HITRAN. HITRAN*online* is a web-based interface that enables users to extract, filter, and plot relevant portions of the database for their application. It also includes display of original sources and bibliography. Almost 5000 users throughout the world have accessed HITRAN*online* since its appearance just over a year ago. For more advanced features, we have developed HAPI (HITRAN Application Programming Interface). HAPI allows for sophisticated processes on the HITRAN data, including radiative-transfer calculations and implementation of complex line-shape functions.

We have created our own server system at the SAO to handle the large traffic that HITRAN*online* has required. Around 320 unique users visit the HITRAN*online* per day.

## DASCH Data Repository

The Digital Access to a Sky Century @ Harvard (DASCH) project is digitizing all the Harvard College Observatory (HCO) glass plate images, approximately 450,000 in estimated total number. An initial overview of DASCH is given in http://adsabs.harvard.edu/abs/2012IAUS..285...29G.
The project enables Time Domain Astronomy (TDA) on all classes of stellar objects, from M dwarfs to supernovae, transients and quasars, over the 1888-1991 duration of the plate taking with ~15 principal telescopes (and ~60 minor series) in both hemispheres for full-sky coverage. The estimated ~100,000 plates recording astronomical spectra are not being digitized for DASCH. Most of the principal series plates are 8x10in (a small fraction are 4x5in) with plate scales ranging from 60-600 arcsec/mm, with some of the "patrol series" plates covering very

large fields of view with 1200 arcsec/mm. The plates are digitized (2 at a time) on a very fast (90s for 2 plates) high precision (0.5um positional accuracy) custom-built digital scanner into 11um pixels, as recorded by a 4Kx4K CCD with ~7ms exposures every 0.5s of overlapping image tiles (40 x 40mm). A single 8 x 10in plate is mosaiced from 60 image tiles into a fits file image (~1GB), with full header information. Approximately 30% of the sky has now been digitized from ~150,000 plates producing ~9.5 billion stellar magnitude measurements. The scanning is being done and data released in a series of 12 Data Releases (DRs), starting from the North Galactic Pole and with each DR being a 15deg bin in Galactic latitude, b. DR5 was released on August 4, 2016, covering the latitude band b = 30-15deg. All data in DR1-DR5, as well as data from 5 initial "Development Fields" can now be accessed from the DASCH website at dasch.rc.fas.edu. The current data total is ~300TB on spinning disk (mostly image .fits files); a light curve (LC) on any given object can be rapidly accessed (<10s, typically) from the DASCH servers. Images can be pulled up for visual examination of any given point in a LC in ~10s. The digital data for LC construction and download are a much smaller data total, ~2TB currently. The image data and derived data products are backed up on LTO5 tape cartridges (~8TB each; "guaranteed" for ~30y) and stored off site at secure locations. All data for immediate access are hosted on DASCH servers and disks on the Harvard research computing cluster, Odyssey. At the current rate of scanning, the archive should be complete (~1PB) in late 2019. The Odyssey servers and data distribution will be maintained for at least 10y, and longer term plans for permanent access as well as possible mirror sites are being developed.