A Snapshot of Computing at CU Boulder
In support of
Research Cloud Computing for the CU Boulder Campus
March, 2017

1. Charge

In September 2016, Vice Chancellor for Research Terri Fiez convened a group of subject matter experts from around the CU Boulder campus to look at how cloud computing might be included in the campus research computing structure. The charge to the group included:

(1) Create a plan for a cloud computing service model and cloud deployment model in support of research, especially research incorporating big data analytics. The plan will include requirements gathering to support the launching of any technical directions;
(2) Identify a service and deployment model in support of instruction and training;
(3) Generate a plan for an organization and staffing that provides open and equitable access to interested users;
(4) Explore a sustainable financial model for the operation and support of cloud resources; and
(5) Design guidelines and policies that are amenable to the adoption of cloud computing for researchers at CU Boulder in the future.

An initial report outlining a vision for research and education cloud computing, definition of success for the effort, results of a benchmarking exercise, identification of possible organizational structures, and recommendations for next steps was submitted in November, 2016.

In this report, we summarize a picture of current computing workflows, support used, and computing needs, based on survey and focus groups of CU Boulder researchers and research IT support personnel.

2. Working Group Members

Ken Anderson, Computer Science
Jim Dykes, IBS
Orrie Gartner, OIT
Dirk Grunwald, Computer Science
Sangtae Ha, Computer Science
David Hamrick, OIT
Thomas Hauser, OIT/RC
Brian Johnson, NSIDC

David Kohnke, Leeds
Larry Levine, OIT
Kurt Maute, AES
Joe McManus, ITP

Michael Paul, Information Science
Ben Shapiro, ATLAS/Comp. Science
Doug Smith, CEAS

Supported by: Ligea Ferraro, OIT
Facilitator: Emily CoBabe-Ammann, RIO

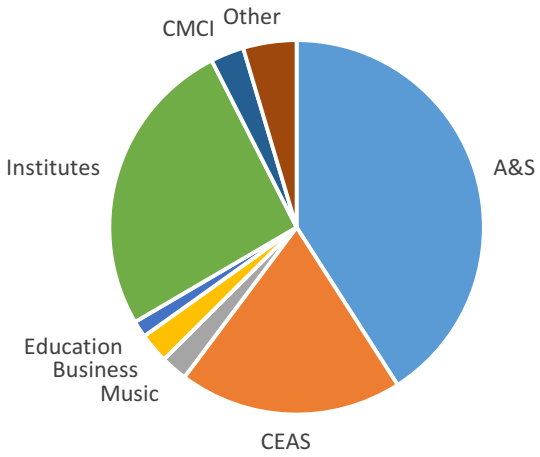
3. Capturing a Picture of Research Computing

The working group has taken a two-prong approach to assessing computing workflows, use of computing support and challenges users face. Initial information was captured through an 18-question survey fielded in December 2016. The survey was sent to all faculty and researchers on the CU Boulder campus. The second phase of information gathering consisted of a set of focus groups conducted in January and February 2017.

3.1. Demographics of the Survey Population

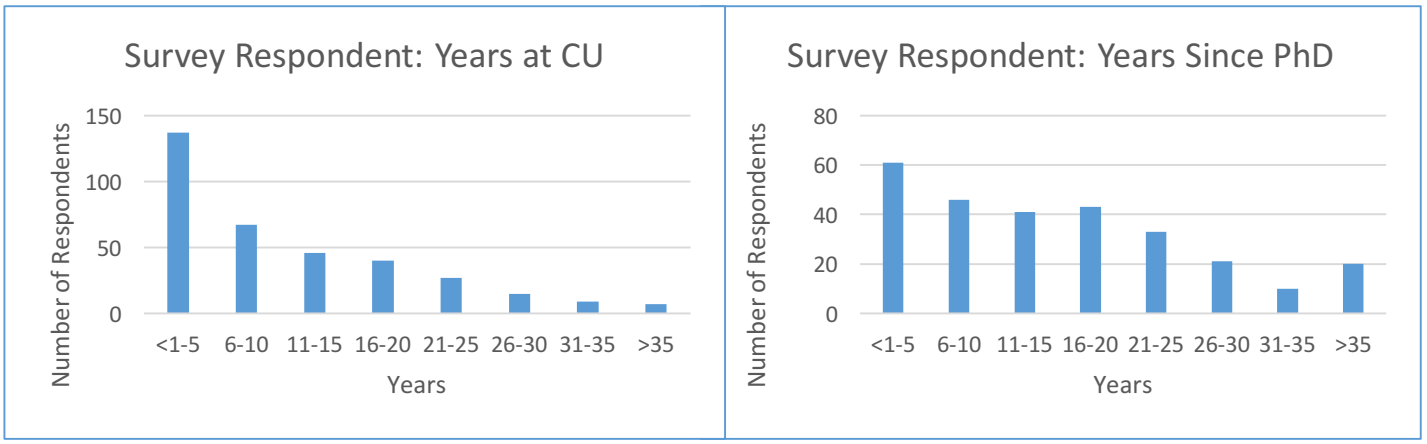
A total of 656 people had responded to the December survey. Roughly 40% of respondents affiliated with the College of Arts and Sciences, 25% from campus Institutes, and 20% from the College of Engineering and Applied Sciences (Fig. 1).

Fig. 1. Survey Respondent Affiliation



Respondents were asked both how long they had been at CU, as well as how long since their Ph.D. Responses show, unsurprisingly, that younger and newer employees responded at a greater rate than other researchers (Fig. 2).

Figure 2. Distribution of Respondents based on years at CU and since PhD



3.2 Demographics of Focus Group Participants

The focus groups interviewed a total of 25 people, largely from A&S, CEAS, and CU Institutes (Fig. 3.). The distribution of participants skews even more towards newer and younger researchers, with the majority of participants within 5 years of obtaining their Ph.D. (Fig. 4).

Fig. 3. Focus Group Affiliation

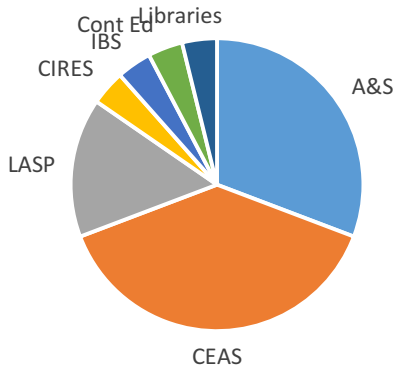
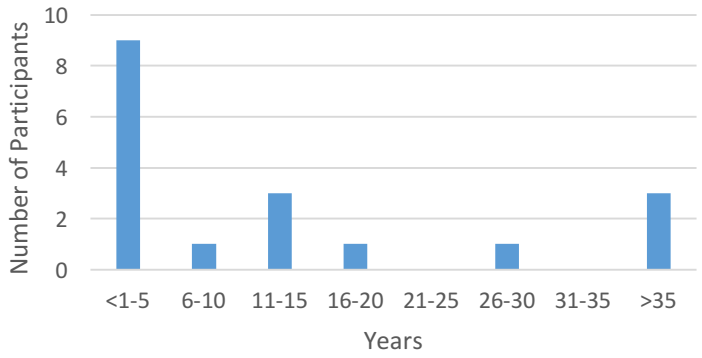


Fig.4. Focus Group: Years Since Ph.D.



4. Computer Usage in Research and Teaching

4.1. Computing Resources

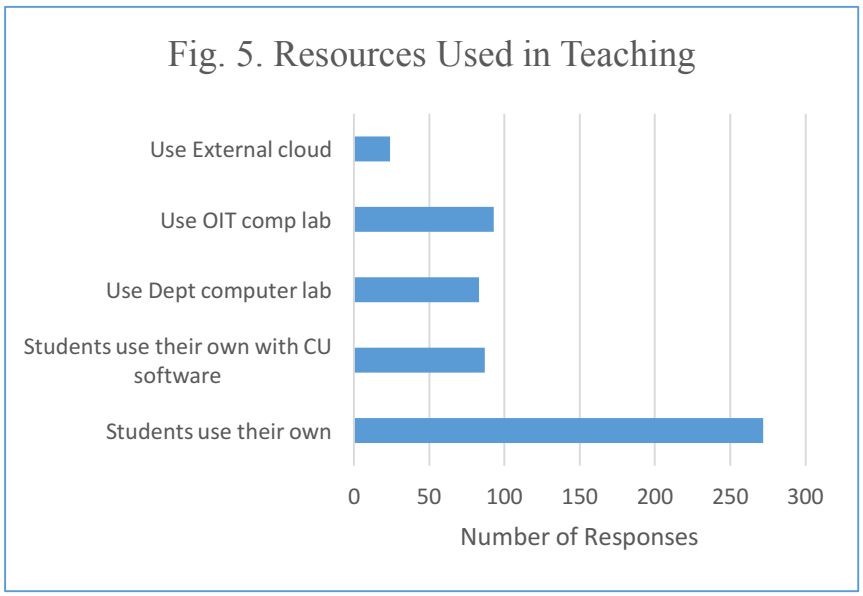
The survey asked whether respondents used computing resources on campus for their **research**. 78% of respondents indicated they do, while 22% do not. When asked what computing resources they use, the majority rely on their own computer and department/research group resources. Almost half take advantage of either Research Computing or other OIT resources, and almost a quarter utilize private cloud resources, either through OIT or their own department. In addition to these offerings, researchers are using Dropbox, commercial cloud, and federal computing resources (e.g., NOAA, NASA, NSF)(Table 1.)

Table 1. Computing resources used in research.

Your own computer	**83%**
Departmental- or research-group owned	**70%**
Campus computing resources (RC or OIT)	**43%**
Cloud (private managed by OIT (SIS) or Dept)	**22%**

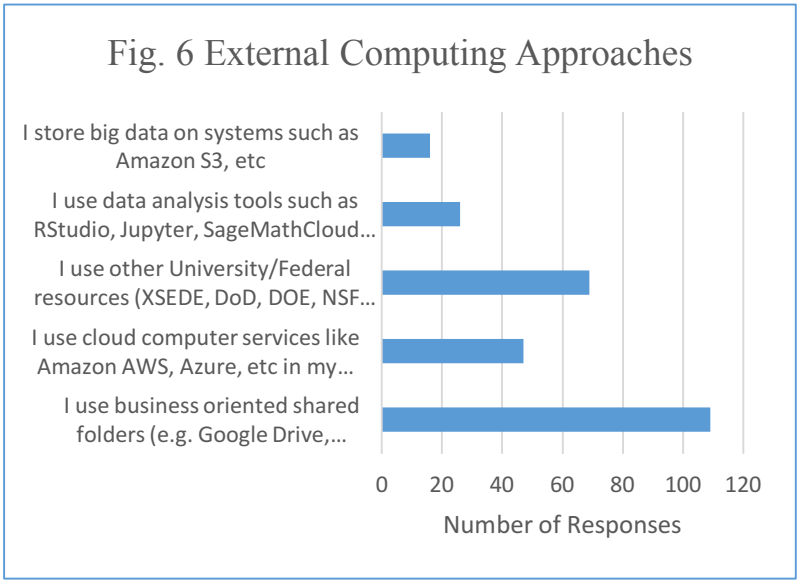
The survey went on to ask whether respondents used computing resources on campus for **education**. 77% of respondents indicated they do, while 23% do not (with more than a few going on to say they ban computers from their classroom). Of those that do, 54% ask students to use their own. About 15% ask students to use their own computer but provide them with software they need to use it.

Another 15% use a departmental lab, and about a fifth use an OIT computing lab. A small fraction (<5%) relies on external commercial cloud for their teaching (Fig. 5). In addition, responses indicate that respondents are using other resources, including mobile devices, freeware, ITLL and libraries, NSF and NASA resources, and other research servers. In other words, as more than one respondent said, 'whatever works'.



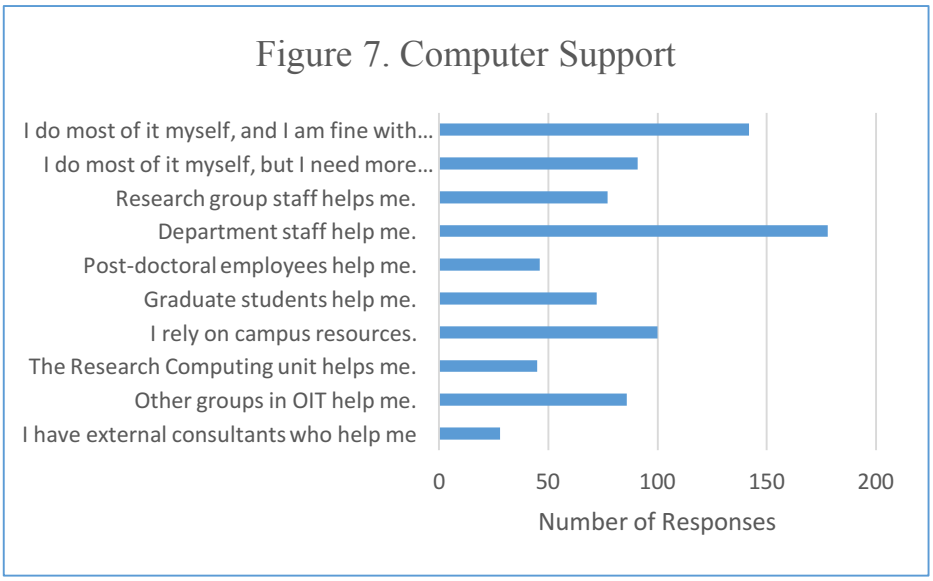
4.2 External Computing Resources

A more detailed picture emerges of how researchers are using external computing options. When asked, more than 30% of survey respondents indicate they use external computing. Most are using files sharing systems (e.g., Drop Box). More than a half use other university or federal resources (e.g., XSEDE, DoD, DoE), and almost 15% are using commercial cloud (Fig. 6).



4.3 Use of Computing Support Across Campus

An important area of exploration in the survey was to look at how researchers are accessing the computing support they need. When asked if they had someone who helped them with their computing needs, only 53% said "Yes". 65% suggest they do most of their computing themselves and they are fine with that. 40% say they do most of their own computing themselves, but they need more help. A third rely on help from OIT, while fifth rely on Research Computing. Most have departmental staff that help, while the majority rely on a myriad of graduate students, post-docs and research group staff for support. Perhaps most surprisingly, almost 15% rely on external consultants to help them. (Fig. 7)



5. Computing Workflow

5.1. Overview

The most important aspect of both the survey and particularly the focus groups was to look at research computing workflows on campus. In the survey, respondents were asked what best described their computing approach in several areas of storage, hardware, software and compute (Table 2). Responses are in Fig 8.

Compute: Perhaps most intriguing, almost half of respondents are currently using cloud computing, such as Amazon AWS, Azure and Google, as part of their workflow. A third of respondents describe their computing as using "multitudes of computers and primarily care about floating point or computational performance". About half are using large data collections, either generated here at CU or elsewhere.

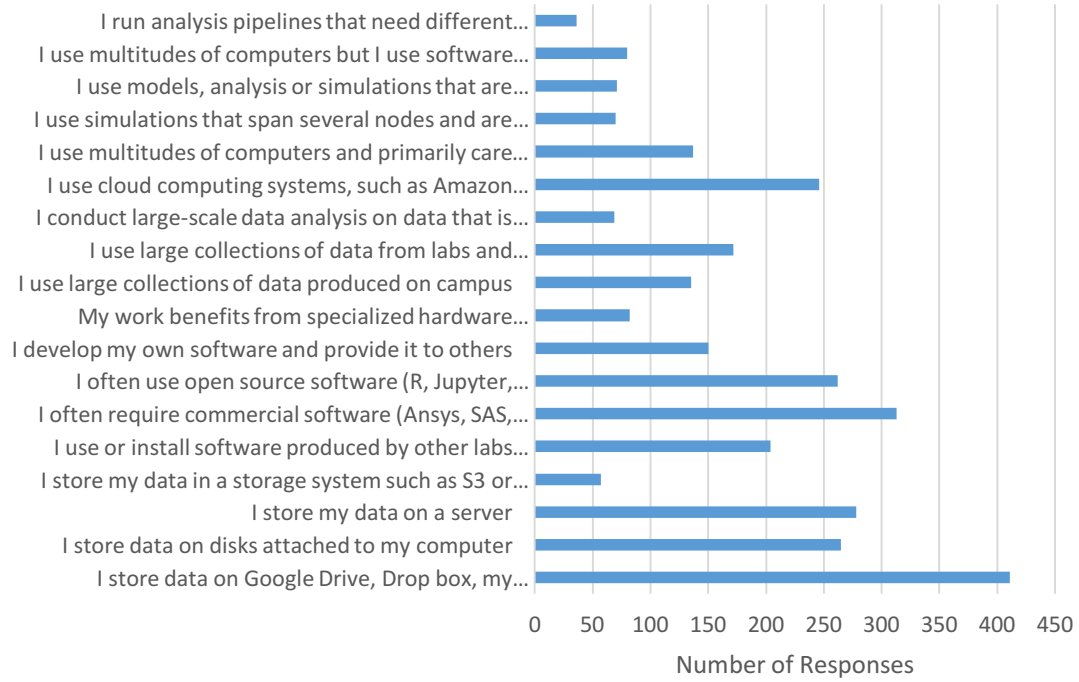
Hardware/Software: 15% identified their need for specialized hardware (NVIDIA GPU). Two thirds are using commercial software (e.g., Ansys, SAS), and more than half are using open source software (e.g., R, Python, Jupyter).

Storage: Unsurprisingly, most respondents are using Google Drive, Drop Box, and their computer for storage. Half indicate they are storing data on disks attached to their computer, and half say they store data on a server.

Table 2. Survey Responses on Computing

Storage	I store data on Google Drive, Drop box, my computer
	I store data on disks attached to my computer
	I store my data on a server
	I store my data in a storage system such as S3 or Petalibrary
Software	I use or install software produced by other labs and research groups
	I often require commercial software (Ansys, SAS, MATLAB, etc.)
	I often use open source software (R, Jupyter, Python, etc.)
	I develop my own software and provide it to others
Hardware	My work benefits from specialized hardware (NVIDIA GPU)
Data	I use large collections of data produced on campus
	I use large collections of data from labs and research groups elsewhere
	I conduct large-scale data analysis on data that is not at CU
Compute	I use cloud computing systems, such as Amazon AWS, Azure, Google
	I use multitudes of computers and primarily care about floating point or computational performance (e.g., using MPI, fortran, etc.)
	I use simulations that span several nodes and are tightly coupled
	I use models, analysis or simulations that are independent from each other and require large (>1TB) shared datasets
	I use multitudes of computers but I use software that operates on a lot of data (e.g., Hadoop, Spark, genomics software)
	I run analysis pipelines that need different compute hardware at each stage

Figure 8. How I Computer



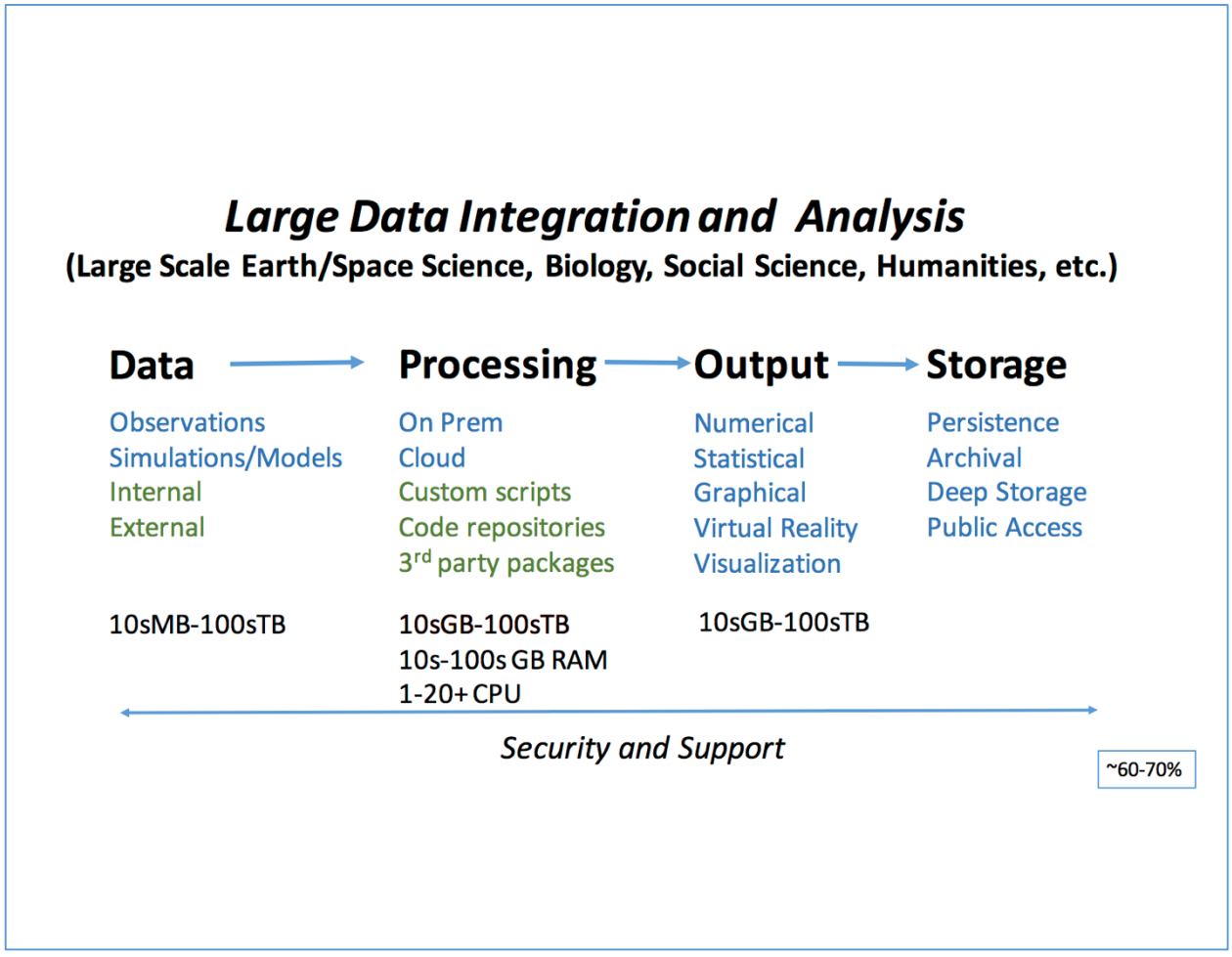
5.2. Drawing the Computing Workflow

In addition to the survey questions, focus group participants were asked to outline or draw their workflow. Two primary groupings of workflow emerged in these discussions. The first can be described as "Large Data Integration and Analysis" (Fig. 9). Observational data, as well as modeling data, from Earth and Space Sciences, as well as Biology, Social Sciences and Humanities most often require this workflow. It can also be viewed as a research production line, where analyses are routine, robust, and repeating.

This work involves multiple data sets, mostly derived from external sources but including internally generated data. Data sizes can be enormous, with researchers describing needing routinely to move 100s terabytes of data. Processing can use 100s GB RAM and 20+ CPU. It can require custom scripts and relies on code repositories, as well as 3rd party packages. Output also runs into the 100s TB.

Our best estimate, based on assessment of survey responses, is that 60-70% of current computing workflow at CU Boulder falls into this category.

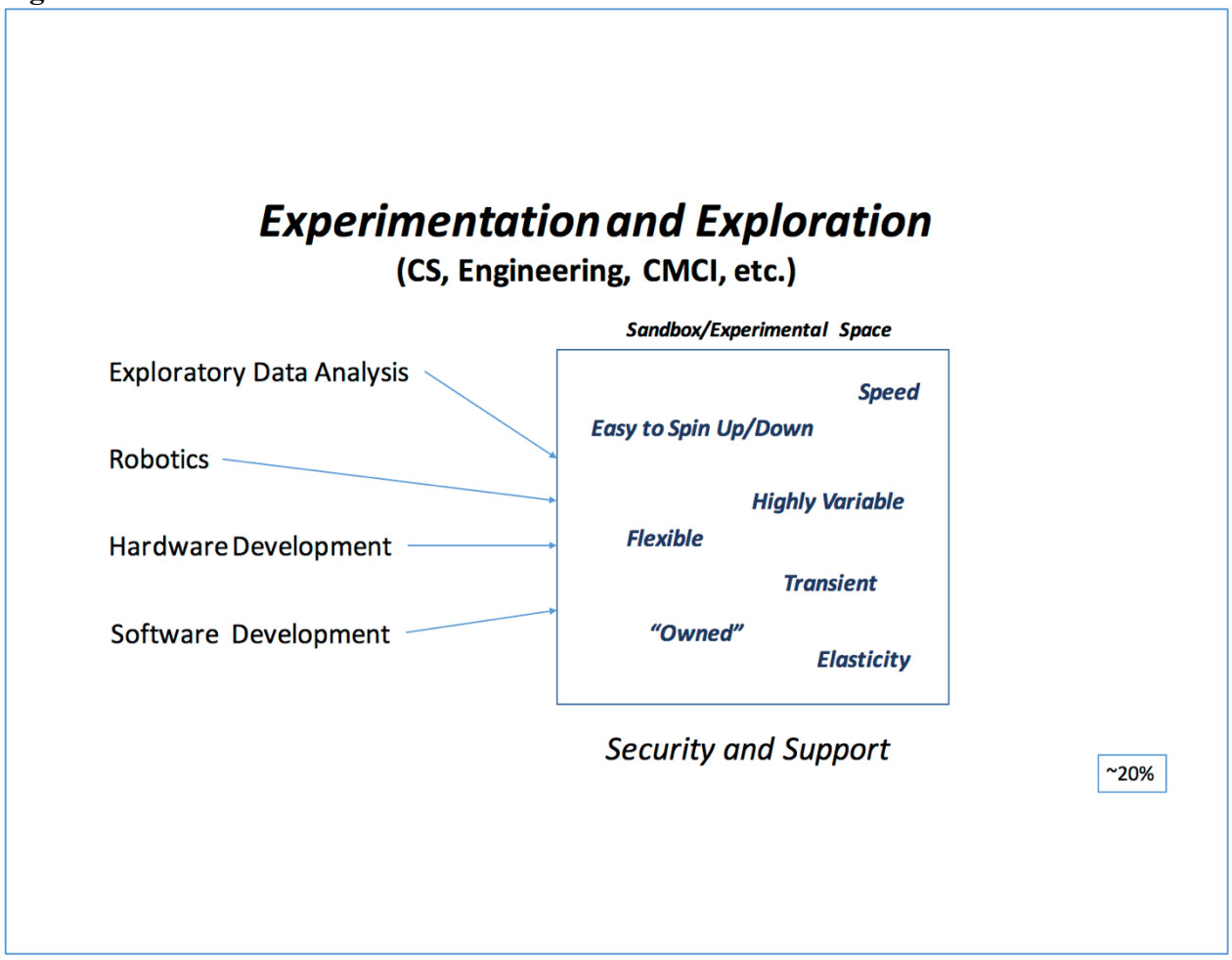
Fig. 9



The second workflow can be described as "Experimentation and Exploration", research that requires sandbox or experimental space that is flexible, highly variable and transient. Experimental work being done in software and hardware development, and robotics and instrumentation development. (Fig. 10)

Our best estimate, based on assessment of survey responses, is that 20% of current computing workflow at CU Boulder falls into this category.

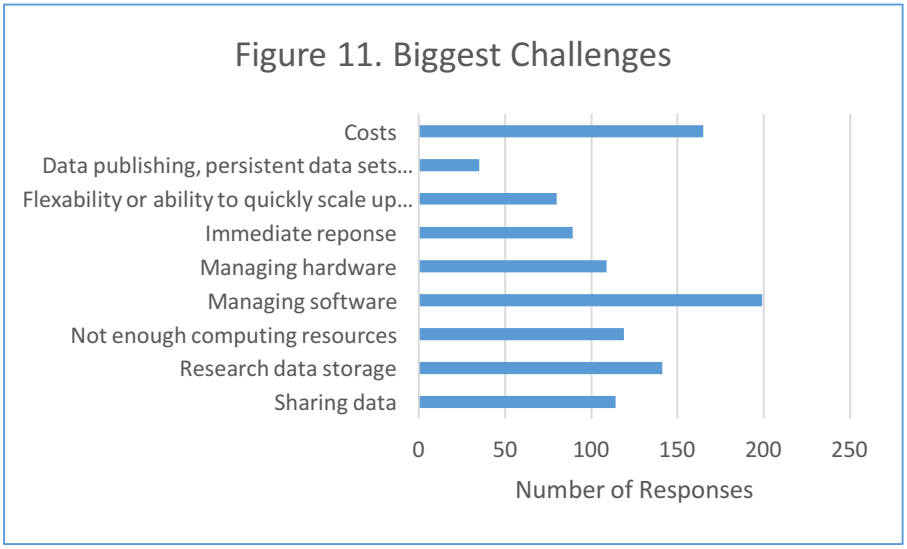
Figure 10.



6. Computing Challenges

In both the survey and the focus groups, we asked participants to describe the challenges they are facing in the computing world. Survey respondents identified managing software and cost as their biggest challenges, followed closely by

research storage managing hardware, sharing data, and not having enough computing resources. (Fig. 11).



In the focus groups, a different tack was taken. Participants were asked to identify three pain points and were given a 'magic wand' to solve their most pressing issues. Responses were binned into "Infrastructure" (compute and storage) and "Support". Key responses are listed in Table 3.

Table 3. Key Computing Challenges/Needs

Infrastructure			
	Compute		
		Flexibility	
			Non-proposal allocation for exploratory research
			Flexibility in experimentation, balanced with fair use
			AWS or similar to make VMs, use as they need, then trash it
		Scalability	
			Fast parallelable visualization capabilities for massive data sets
			Remote Linux VMs similar to iPlant and AWS infrastructure with good X11 forwarding
			Binder/Docker for spinning up scalable instances
	Storage		
			Access to the cloud, uploading scripts and codes to the cloud, run on data where they are originally archived - eliminate need to transfer data. Run analysis where data lives
			Lots of small data with lots of read/write creates I/O bottleneck
			Storage NFS is slow
Support			
			Code writing support
			Installing and running experimental software
			Help with data publication, ranging from MB to 5+ TB
			Coherent sys admin support (move away from ad hoc or reliance on graduate students)

7. What does the Future Look Like?

Focus group participants were asked to paint a picture of what their computing future looks like on a 3-5year time frame. Not surprisingly, the answer was MORE…More data, both in terms of size and in terms of heterogeneity…. And MORE (And FASTER) compute, storage and support needs. In addition, an increasing emphasis on and requirements for reproducible science is driving a need for more externally available space for data and algorithms. Participants identifies several key areas of emphasis for managing what's coming, including faster pipelines (especially graphic and image), increase remote access and mobility for easier collaboration, and a focus on identity management, DOIs, and persistent data URLs for better deployment of public scholarship. ***In general, there was consensus that researchers would need an order of magnitude more storage and compute within 5 years.***

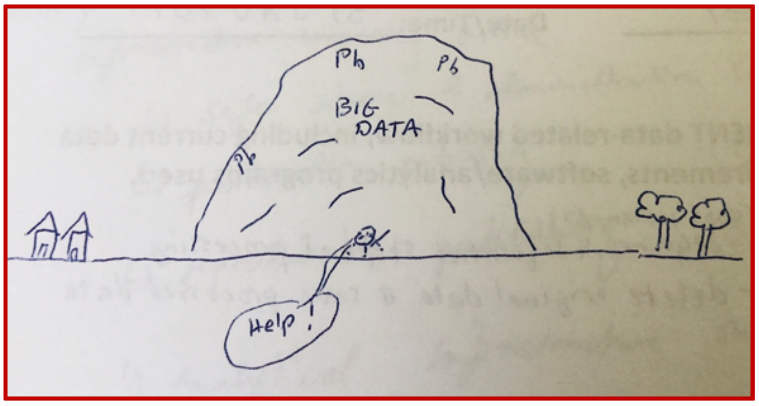
One participant summed up the situation:

> ***"As scale increases, the current environment is unsustainable"***

Another stated:

> ***"Data is quickly becoming too large to move."***

And finally, when asked to draw her future, one participant presented the following:



8. Conclusion

Both the surveys and the focus groups paint a picture of a vibrant academic community supported by a significant research and educational computing effort. At the same time, there are signs that the future is already here. Researchers are already hitting walls, in terms of ability to manage their computing workflow, because of the size of datasets and output. Those working in the experimental space need a more flexible, adaptable approach. Researchers and educators have already moved into the cloud for their work, both for ease and out of necessity. It is incumbent upon us to look for architectural solutions that will support their efforts.

9. Next Steps

This effort began, and remains, an approach to the use of "the cloud" for CU Boulder researchers. Along the way, we've learned that the requirements for researchers have already lead to cloud solutions and that "cloud" for researchers is both on and off premise solutions. The mission of the effort is not to say "the cloud's the answer, what's the question?" We do, however, know that, based on researchers' and educators' requirements, part of the answer is the cloud. CU Boulder's need for cloud solutions is growing, and many researchers have already sought cloud solutions on their own. It is clearly time for OIT, RIO, and other entities to coordinate and optimize cloud needs.

To begin to develop models of how campus might address these emerging computing needs, we recommend convening a small group of OIT and financial specialists. The group will analyze scenarios for organizational impact and costing, with the goal of creating a campus-wide umbrella of support and services that provides the best solutions for meeting faculty research and education requirements. The group will call upon subject matter experts (e.g. users, OCG, legal, etc.) to address specific area of work.

The group will look at current and future needs, looking at both gaps and opportunities. The group will focus on several key areas:

- Services and Infrastructure: including key partners, data security, legal and contractual processes, networks, data curation and accessibility, and tools.
- Support: including classes, training, networks.
- Incentives: identifying ways to engage and facilitate both PIs, as well as IT staff across campus.
- Policy and Pathways: including a formal communication plan, as well as clear guidelines that create smooth on ramps for researchers, educators and IT professionals.

- Key partnerships and resources to help optimize the balance between autonomy and centralized support.

Outcomes of the effort should include a picture of current and future states for research computing, strategies for engagement and support, as well as a financial model that promotes sustainability. These efforts should also include:

1. Define a core set of cloud services to be offered through OIT, with other campus entities. (OIT, Institutes, RIO, campus representation)
2. Define a core set of education and training to support those services, offered through OIT, in partnership with other campus entities. (OIT, Institutes, Departments, RIO, campus representation).
3. Develop an organization and staffing plan to support research cloud computing.
4. Develop an efficacious costing model for cloud research computing. Begin by assessing campus research needs for computing and storage (and for networking, which will be influenced by how much needs to be stored where and analyze how). Consider not just how might RC and OIT adjust how resources are deployed, but also: a) what should the campus investment be overall in support of research for computing, storage, and networking as a combination of cloud and on prem services; and b) how can that investment be met by current resources, incremental resources from the campus, and a billing model, likely subsidized, so that faculty are incentivized to work together to build collaborative resources. The driving factor is what best serves faculty research achievement overall?
5. Begin to identify the campus risk posture for external service providers, particularly in the realm of public verses private cloud. (OIT, Campus Leadership)

Once these efforts are completed, by Fall 2017, a campus-level working group can develop a deployment plan, with timetable and costs.