Assessment of non-financial returns on cyberinfrastructure: A survey of current methods

Craig A. Stewart
stewart@iu.edu
orcid.org/0000-0003-2423-9019
Pervasive Technology Institute
Indiana University

Amy Apon
aapon@clemson.edu
orcid.org/0000-0001-5617-5334
Computer Science Division
Clemson University

David Y. Hancock
dyhancoc@iu.edu
orcid.org/0000-0001-8082-8980
Pervasive Technology Institute
Indiana University

Thomas Furlani
thomas.furlani@roswellpark.org
orcid.org/0000-0002-4683-0814
Roswell Park Comprehensive Cancer Center

Alan Sill
alan.sill@ttu.edu
orcid.org/0000-0003-2527-764X
High Performance Computing Center
Texas Tech University

Julie Wernert
jwernert@iu.edu
orcid.org/0000-0002-5705-9527
Pervasive Technology Institute
Indiana University

David Lifka
lifka@cornell.edu
orcid.org/0000-0003-0069-6530
Cornell University

Nicholas Berente
nberente@nd.edu
orcid.org/0000-0002-1403-4696
Mendoza College of Business
University of Notre Dame

Thomas Cheatham
tec3@utah.edu
orcid.org/0000-0003-0298-3904
University of Utah

Shawn D. Slavin
slavin@iu.edu
orcid.org/0000-0001-6602-4310
Pervasive Technology Institute
Indiana University

ABSTRACT

In recent years, considerable attention has been given to assessing the value of investments in cyberinfrastructure (CI). This paper focuses on assessment of value measured in ways other than financial benefits - what might well be termed impact or outcomes. This paper is a companion to a paper presented at the PEARC'19 conference, which focused on methods for assessing financial returns on investment. In this paper we focus on methods for assessing impacts such as effect on publication production, importance of publications, and assistance with major scientific accomplishments as signified by major awards. We in particular focus on the role of humans in the loop - humanware. This includes a brief description of the roles humans play in facilitating use of research cyberinfratructure - including clouds - and then a discussion of how those impacts have been assessed. Our conclusion overall is that there has been more progress in the past very few years in developing methods for the quantitative assessment of financial returns on investment than there has been in assessing non-quantitative impacts. There are a few clear actions that many research institutions could take to start better assessing the non-financial impacts of investment in cyberinfrastructure. However, there is a great need for assessment efforts to turn more attention to the assessment of non-financial benefits of investment in cyberinfrastructure, particularly the benefits of investing in humans and the benefits to humans who are involved in supporting and using cyberinfrastructure, including clouds.

CCS CONCEPTS

• **General and reference** → **Metrics**; • **Social and professional topics** → **Government technology policy**; • **Computer systems organization** → **Architectures**.

KEYWORDS

ROI; impact; publications; outreach; workforce development

ACM Reference Format:
Craig A. Stewart, Amy Apon, David Y. Hancock, Thomas Furlani, Alan Sill, Julie Wernert, David Lifka, Nicholas Berente, Thomas Cheatham, and Shawn D. Slavin. 2019. Assessment of non-financial returns on cyberinfrastructure: A survey of current methods. In *Humans in the Loop: Enabling and Facilitating Research on Cloud Computing (HARC '19), July 29, 2019, Chicago, IL, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3355738.3355749

HARC '19, July 29, 2019, Chicago, IL, USA

ACM ISBN 978-1-4503-7279-4/19/07…$15.00
https://doi.org/10.1145/3355738.3355749

1 INTRODUCTION

The original purpose of the development of computers was to enable research, to do calculations that were beyond the capability of humans to do sufficiently quickly and accurately. Today we think in terms of cyberinfrastructure, which is commonly defined as "…computing systems, data storage systems, advanced instruments and data repositories, visualization environments, and people, all linked together by software and high performance networks to improve research productivity and enable breakthroughs not otherwise possible" [37]. One of the distinctive aspects of this definition, as opposed to definitions for "information technology," "computational science," or "e-science" is that people are explicitly included as an essential component of cyberinfrastructure. This report is being prepared for inclusion in the workshop "Humans in the Loop: Enabling and Facilitating Research on Cloud Computing," and focuses on assessing the general non-financial impacts of cyberinfrastructure broadly considered, including the role of humans dealing with cyberinfrastructure in general, including clouds.

This report is a companion to a prior report that surveyed the current state of methods for assessing financial benefits: return on investments in cyberinfrastructure assessed strictly in monetary terms [35]. While there is more to research than money, when investments in any type of cyberinfrastructure are discussed within higher education organizations, the question inevitably arises *what do we get for this money?* And, as stated in [35], there are times when conversations are restricted to dollars and cents. Such conversations are necessary in the higher education community, particularly as that community looks to financial challenges and decreased enrollment in the future [16, 19].

Still, research organizations of all sorts - especially colleges and universities - are about much more than money. There are many other kinds of investment impacts in cyberinfrastructure, which either cannot be measured in financial terms or are measured poorly in financial terms. The impact of humans in the loop, for instance, as part of cyberinfrastructure enabling research, and the impact of cyberinfrastructure on enabling humans to develop in careers in research, are examples of important matters that are difficult or impossible to quantify in financial terms. How cyberinfrastructure, inclusive of humans, advances human knowledge and the quality of human life are other matters that are important but difficult to quantify in financial terms. A logic model-based view of organizational processes is helpful [15] in understanding the role of humans in cyberinfrastructure and the many non-financial impacts of cyberinfrastructure generally. Logic models, in general, offer a way to formalize an understanding of what one plans to do and what one intends as results. Figure 1 below is adapted from [15], and is also used in [35].

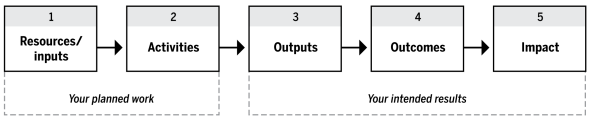

Figure 1: Logic Model of Organizational Processes

This logic model leads us to create a clear separation between *what we are doing* and *how we are doing it*. The primary objectives that a university or college has as for investment in CI include enabling outcomes from innovation, creativity, discovery, engineering, analyses, and research, along with providing education and training integral to 21st century workforce development. This logic model also shows that humans play important roles in both using and in supporting the use of cyberinfrastructure that are each required to achieve the benefits of such use. In the case of the development of STEM professionals, a person might play both of these roles at once: benefit in terms of skill and capability growth by working in a role supporting or implementing cyberinfrastructure, and learn skills and competencies that then aid that person's career development. Each of these roles enhances that person's value in both non-financial and financial ways to future employers. In other words, when considering the role of humans in cyberinfrastructure, it's hard to make a clear separation between aspects of *what we are doing* and *how we are doing it* because part of *how* people aid use of cyberinfrastructure becomes part and parcel of the *what*, which includes producing well trained STEM professionals.

Table 1 (based on a similar tables by Stewart, et al. [35, 36] enumerates benefits of investment in cyberinfrastructure in terms of outputs, outcomes, and impact, all of which are measured in ways other than financial. Note that impacts of humans are largely implicit, though the impact and influence on humans are explicit.

2 DATA SCIENCE APPROACHES TO THE STUDY OF ADVANCED CI IMPACT ON RESEARCH OUTPUT

The most basic unit of research output that persists over time is the peer-reviewed publication. Once upon a time, there were few journals in any given discipline of science, and a person well-versed in any field could offer a reasonably defensible ranking of all of the journals in a given area, and even offer a well informed (if biased) analysis of the most important publications in any given field. Those days are long gone. For example, Elsevier's SCOPUS database of citations, which Elsevier asserts is the largest such database in the world when considering recent publications [13] includes a total of 19 million records of citations to scientific and technical publications, the majority of them peer reviewed. Currently, SCOPUS includes indices of more than 22,000 journals. The other major citation database, Web of Science, covers more than 90,000 citations going back several decades in time prior to the earliest record in SCOPUS [21]. The distinct advantages of one over the other in any particular circumstance notwithstanding, it's clear that to understand anything about the scientific publication enterprise one must take a data science approach to the analysis of cyberinfrastructure impacts on the scientific endeavor.

2.1 Production and acceleration of research

Non-parametric methods and hypothesis testing with non-parametric efficiency estimators were applied by Apon and her co-workers to the analysis of the effect of locally-available supercomputing resources on university efficiency in producing research in [18]. Data from the National Research Council [29] was analyzed along with derived values from the Top 500 list to study the effect of locally-available supercomputing resources on six different academic areas. Interested readers are encouraged to see the original paper [18]

Table 1: Non-financial returns on investment in cyberinfrastructure, from an organizational logic model viewpoint

Output	Outcome	Nonfinancial measures of outcome	Impact
New discoveries reported in publications	Publications	Number of publications, citations of publications, impact factors of publications (and the journals in which publications appear	Improved quality of life for people
	Shorter time to publications	Time	Better management of natural resources
People trained in areas in which they would otherwise not have been trained	A better-trained STEM workforce		A better-trained workforce for the economy
			Improved global competitiveness for any given country
			Increased salary, greater employment security for the individual
Awards, press notices	Any award, e.g., Nobel Prize	Numbers, types of awards	Recognition of a particular invention's significance; reputational benefits for the people and organizations winning the award
Patents	An invention is legally protected by exclusive use of the patent holder or licensee	Number of patents	The invention may become a commercial product, or may be used in commercial products that improve people's quality of life, and the sustainability of human life on Earth

for details on the statistical methods. Results from this research found clear evidence that research departments in Chemistry, Civil and Environmental Engineering, Physics, and to a lesser extent, History, are more efficient at producing research in universities where supercomputing is readily available. That is, in the named departments, the presence of supercomputing increases research output. The results, however, suggest that research departments in Computer Science and Economics are less efficient at producing research in universities where supercomputing is readily available. The authors offer some hypotheses about why this might be the case, for example, that available supercomputing leads to more interdisciplinary research with fewer publications in core Computer Science venues. More research is needed to know if such research is less efficient in terms of scientific advancement.

2.2 Publications as metric and product

One approach used by multiple centers and projects is simply to tally up publications that were aided or in part enabled by services provided by that center. Examples of this are available for PTI [12], CyVerse [2], and Galaxy [7], including different styles of search filtering. These are, of course, relatively blunt measures, but they are measures, and they contribute to demonstrating the value of projects and facilities [34].

Apon and colleagues have taken more sophisticated approaches. Correlation and a two-stage least squares regression are used in [17] to analyze the research impacts of investments in supercomputing, which are measured using values derived from the Top 500 list of institutions that ranked in high or very high research categories according to the Carnegie Foundation classification. This paper looked at the sample of high and very high research institutions both with and without appearances on the Top 500 list. One model used publication counts as a dependent variable of NSF funding and Top 500 appearances. Apon and her co-authors found that investment in supercomputing yields statistically significant immediate returns in terms of increased academic publications relative to the institution's own past historical average. However, this effect suffered a fast depreciation over the two year horizon.

Open XDMoD (Metrics on Demand) is a tool that aids data collection and analyses related to impact of investments in CI [30]. In particular, Open XDMoD includes tools for analysis of publications and publication impact. Bibliometrics-based analysis is a commonly used method to evaluate the research impact of an individual, a research group, or even an organization. Publication count and

citation count based metrics provide an effective way to show the quantity and quality, and the impact of scientific research activities. For instance, it was used to evaluate the quality of research in the United Kingdom [32, 40].

Von Laszewski and colleagues have implemented software tools within OpenXDMoD to analyze the impact of infrastructure resources that goes well beyond just publication and citation counts [43]. This tool takes as input a list of publications created with the use of a particular piece of infrastructure, and generates tallies of publications, citations to those publications, and calculates h-index [28], g-index [6], and the m factor of h-index (which indicates the slope of the h-index [28]. This tool also enables peer analysis — comparing the impact of publications that use one particular piece of infrastructure vs. publications that did not by comparing citations to publications within peer journals. These methods were initially developed for analysis of the impact of cyberinfrastructure resources such as XSEDE (the eXtreme Science and Engineering Discovery Environment) [41] and have now been expanded to the analysis of other sorts of infrastructure, such as the facilities of the National Center for Atmospheric Research [8], as shown in Figure 4. These peer analyses showed that papers based on use of NCAR facilities achieved more citations than peer publications that had not; similar results prevail for XSEDE [44]. This is a strong indication that one of the impacts of use of cyberinfrastructure is increased impact of publications resulting from use of advanced cyberinfrastructure facilities, as opposed to publications that do not use such facilities. The methodology created and implemented by von Laszewski is one of the most thorough and quantitative tools in existence for analyzing the importance of cyberinfrastructure in influencing the impact of publications in a particular area of research.

3 MAP-BASED APPROACHES

Maps and analyses of networks of various kinds have been used for centuries to help us understand the world around us. Recent advances in data, algorithms, and computing infrastructures make it possible to model and visualize the structure and evolution of science and technology (S&T) systematically. Results can be used to evaluate and compare different scholars, institutions, regions, or countries; to identify emerging areas, track the diffusion of knowledge or innovations; or to visualize career trajectories. Network analysis can also be used to visualize the impact of investments, facilities, and research in various scientific disciplines. Figure 2 shows an example data visualization designed for policy analysis and investment decision-making at the level of individual academic institutions, to quantify and communicate the value of investments in cyberinfrastructure (CI). The interactive visualization allows users to explore the relations between IT resource usage at Indiana University (on left), funding awards aggregated by NIH institute and NSF (in middle), and publications that cite this funding aggregated by scientific discipline (on right).

Sankey graphs (the sort of graph depicted in Figure 2) take users on an exploratory quest, in this case moving from the IT resources via funding to papers published in diverse scientific disciplines. The financial return on investment in IT infrastructure is measured in terms of IU funding, totaling $339M for 885 NIH and NSF projects associated with IT usage, and the academic ROI constitutes 968 publications associated with 83 of these NSF and NIH awards. The visualization shows that Brain Research, Medical Specialties, and Infectious Diseases are the top three scientific disciplines ranked by the number of publications during the given time period. The visualization is freely available in the Value Analytics module that can be downloaded as a standard Open XDMoD extension package [33]. Hovering over a particular node will cause that node and all links emanating from it to be highlighted, whereas hovering over a particular link will highlight that link with the color of the node from which the link originated. In Figure 1, the user selected the link between NIH-NLM funding (2) and Brain Research from Scientific Disciplines (3) to understand what IT resources — answer: Karst (1) — were used for the 54 publications. The visualization featured in Figure 2 utilizes the UCSD map of science and classification system [20] to aggregate journals into subdisciplines of science that are further aggregated into 13 disciplines of science (e.g., mathematics or biology).

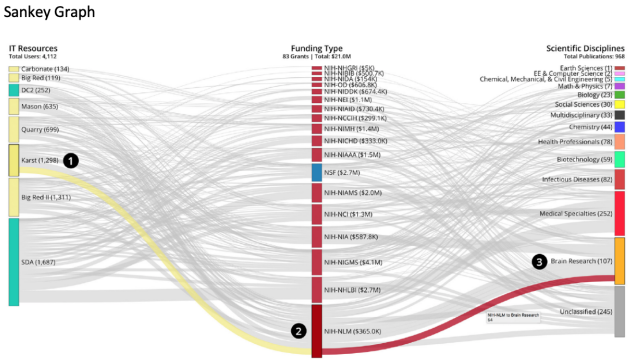

Figure 2: Example of a Sankey graph depicting relationships between CI system use, disciplines, and publications

The Sankey graph above deals in part with financial returns on investment (in the form of grants received); it also relates use of cyberinfrastructure to publications in a variety of fields and, in that way, can be helpful in understanding non-financial aspects of impact of cyberinfrastructure.

4 ASSESSING THE ROLE OF HUMANS THROUGH QUALITATIVE ANALYSES

4.1 Cyberinfrastructure Facility, Support, and Software Efficiencies

Typically, campus HPC center staff maintains many scientific code packages that are used by multiple research groups. Facility personnel will install, support, and optimize these code packages and keep them up to date, freeing the separate research groups from the responsibility to support these packages on their own. This allows researchers to focus more directly on their research, rather than consuming valuable time and effort installing and maintaining these packages. By offering training and access to a professionally

managed shared resource, this approach lowers the barrier to entry for new users of HPC and helps to maximize their research productivity.

The advantages of pursuing such an approach should be balanced against the desire and inclination of individual researchers or groups to want to maintain control over the software that they use. Recent advances in technologies such as containerization and virtualization as well as the development of cooperative working tools such as Jupyter notebooks [39] have improved researchers' ability to customize and share their working environments while using the infrastructure of a shared campus data center or shared access to cloud-based resources.

Quite rightly, the focus has turned in many ways to a discussion about the best ways to leverage such tools to improve the reproducibility of research results by sharing access to software, algorithms, and data used to reach a given result [22]. The balance between centrally supported research CI and portable user-controlled tools is optimized when common methods of supporting facilities emerge that minimize non-reproducible methods and maximize scientific interchange and sharing.

Examples of successful deployment of shared CI methods include both community-based and commercial software for central functionalities such as configuring, building, and maintaining clusters, central software storage methods, common standardized interfaces for data sharing, central maintenance of message-passing interface (MPI) software and hardware, and campus-wide or resource-wide access to software licenses. Deployment of such resources on a broader scale led, in the early 21st century, to the emergence of grid computing. It has now evolved to include a wide variety of community-based and commercial distributed computing methods.

Among many examples of successful deployment of CI to broadly distributed communities, we can cite the continued existence of the Open Science Grid (OSG) [10] and its relatives, including the European Open Science Cloud (EOSC) [5], European Grid Initiative (EGI) [4], Asia-Pacific Grid (APGrid) [1], European Data Initiative (EUDAT) [3], and a number of smaller and newer initiatives along these lines. Integration of institutional identity management systems with features needed to support collaboration across a variety of organizations, both real and virtual, is a key factor for success of distributed CI [23].

Staff support the management and use of CI resources, whether sited in individual departments, a common location local to an institution, or outsourced to a remote facility or the cloud. Many institutions have found that consolidating these human resources (at least in part) achieves economy of scale and enables sharing of expertise. Overall, including some centralization of CI capabilities allows the institutional support staff to act as a central knowledge base for high performance resource-intensive data analysis methods and techniques that would be inefficient to duplicate across the university, and provides a point of contact at each institution for collaboration with more widely distributed projects and resources. It also provides the technological base to make use of other methods of service delivery, such as use of national-scale supercomputing CI, academic grids, and clouds.

Staff (humans) supporting cyberinfrastructure do largely the same work whether they are supporting locally-sited CI resources, remote resources housed in a data center, or clouds. There are some clear differences in that staff don't have to actually do anything with hardware in remote data centers or commercial clouds. But systems administration, management of software, and support of users - the majority of what humans do in support of cyberinfrastructure - is largely the same regardless of where the physical CI systems are located. Indeed, the great success of national CI staff training, facilitation, and information exchange programs such as the XSEDE Campus Champions, and the Campus Research Computing Consortium (CaRCC) provide strong evidence of the similarity of CI support needs and challenges throughout the spectrum of CI activities.

4.2 Navigation, facilitation, and technology adoption choices

Current understandings of technology adoption choices based on social science research suggest that adoption is driven by performance expectancy (perceived value), effort expectancy (perceived ease of use), social influence, and facilitating conditions (including knowledge of a technology and the belief end users will find it accessible) [42]. As regards technology adoption choices in research, there is a tremendously important role for humans to play in all four of these areas, starting with what Venkatesh referred to as "facilitating conditions." The majority of people who use cyberinfrastructure for research purposes are researchers in some area of science, engineering, or scholarship other than computer science. Such people are generally intensely focused on the newest ideas, innovations, and trends in their own and related disciplines. They tend not to be focused on the newest trends in containers, communications libraries, CPUs, GPUs, FPGA, etc. Cyberinfrastructure support experts play a critical role in enabling researchers to stay informed about technology options, and such advice plays important roles in deciding what technologies are beneficial to begin with in the context of running an effective research program, and then making transitions. Technological changes happen at a pace that is too fast for most researchers to want to keep up with, though they would surely be intellectually capable of doing so. At the end of the day, a researcher has to do research and produce research results in their own field. One can't constantly be revamping research and data analysis processes to keep up with the latest trends and newest tools. One has to make (hopefully) wise choices about what technology to adopt, and how often it is appropriate to change and re-tool research processes. This creates a general and ongoing need for support, and groups such as Campus Champions and CaRCC Research Facilitators are, for example, particularly focused on these issues, and are adept at such support. Unfortunately, it's difficult to assess how and how much these activities affect research outcomes and speed.

4.3 Qualitative assessment of humans in the loop in supporting research generally

Qualitative analyses have been used to demonstrate the value of advanced cyberinfrastructure and, in fact, qualitative approaches may be the best way to measure the importance of humans in the loop. Some such studies are published, though many are not. For example, the University of Utah does annual surveys of satisfaction with the research facilitators that support researchers in their

use of local and national cyberinfrastructure facilities. While not published, these surveys are used within the university to help justify investments in CI support staff and facilities. Indiana University conducts a survey each year of its information technology service users, which includes specific questions on satisfaction with advanced CI facilities, and questions asking how important IU's CI facilities are in support of research and teaching. These survey results going back to 1992 (including full text of all comments made on the survey, with obscenities, vulgarities, and names of individual staff) are available openly and online at [14]. This forms a basis for "fact-based" discussions about the value of of what we once called IT systems and now call cyberinfrastructure facilities [31]. "Facts" in this case are largely carefully (and methodologically soundly) collected assessments of humans' opinions. Such assessments are valuable, though, and they are one of the best current approaches to assess the value of humans in the loop supporting use of cyberinfrastructure in general and research in clouds in particular.

Other approaches include solicitation of qualitative feedback from researchers through structured interviews done by assessment experts. One example of this sort of approach is in questions asked on surveys conducted by XSEDE, which have to do with the value of XSEDE overall and XSEDE Extended Collaborative Support Services. XSEDE's stated mission is to "substantially enhance the productivity of a growing community of scholars, researchers, and engineers through access to advanced digital services that support open research; and coordinate and add significant value to the leading cyberinfrastructure resources funded by the NSF and other agencies." The project offers online training and consulting services as part of that mission. Surveys have been used in two ways to assess the importance of the XSEDE cyberinfrastructure itself and of the humans that work as part of the project in supporting cyberinfrastructure, including three systems that are clouds or have cloud-like features (Jetstream, Comet, and Bridges). One particularly important part of XSEDE support services is termed by XSEDE the "Extended Collaborative Support Services" (ECSS).

ECSS sercvices are a sort of "humanware" activity that is requested and allocated via XSEDE just like use of an advanced supercomputer. ECSS services are offered by a group of experienced professionals, very many of whom have terminal degrees in a field of science and engineering; they are allocated to work with research teams to solve particularly daunting cyberinfrastructure challenges. This may involve actual coding of new programs, or implementation of programs with new libraries or in new environments. It is common for a project to be allocated 3 months of an ECSS consultant's time, so the level of support is intensive.

One of the survey questions asked of all recipients of ECSS services and a random subsample of XSEDE users overall is "how much time would it have taken you to do your research without the assistance provided by XSEDE?" Among the answers are statements such as "I could not have done this research without the support of XSEDE" and "I would not have tried this research without the support of XSEDE." While very much qualitative, such responses are indications of how CI system support makes research possible or practical when it might not otherwise have been undertaken.

Another approach to qualitative analyses involves structured interviews done by assessment experts. Such an approach was taken by IU in contracting for an assessment of the value of its CI systems. The resulting report is available online and extols the virtue of IU investment in facilities and expert humans to facilitate research, but it suffers from the fact that there is no basis for comparison. [24]. Another qualitative study looked at the factors that aided open source software projects in being successful and sustained over the long run [34]. Key characteristics such projects include good support mechanisms and some sort of annual meeting or conference that enabled direct human interaction between users of a software tool and the leaders of the project creating and supporting that tool.

It might be informative to perform a controlled experiment in which a randomly selected portion of the researchers at a given research institution received access to CI facilities and support, and another portion did not have access to these resources. This would be a powerful way to determine the impact of cyberinfrastructure, including people, on research productivity. It's also completely impractical and unethical. As a result, qualitative analyses remain the primary mechanism available to us.

4.4 A cloud computing example: the importance of humans in enabling use of the Jetstream cloud system

Jetstream is a first-of-its-kind system in the sense that it is the first cloud system funded by the NSF accessible to, and designed to be a resource for, the national research community. Jetstream is also in a real sense a pilot project because the NSF has never before funded the creation of a production cloud system. Proposing and implementing Jetstream was at times very much a process of putting a square peg into a round hole. Jetstream was funded through an NSF high performance computing (HPC) system acquisition solicitation, but the actual acquisition of the hardware from the vendor (Dell) was a small portion of the effort needed to put a cloud system into production.

The system software for Jetstream is mostly open source — it runs the widely used OpenStack cloud software system [11], as well as many other open software components. It has been in production operation since June 2016. More than 2,000 individuals currently have accounts on the system, and more than 20,000 users have run jobs on Jetstream via science gateways and workflow systems that operate on the system.

Jetstream is the first US-based cloud system designed specifically for production use by a national community of researchers from many disciplines; it is thus a harbinger of things to come. Certainly one can do scientific research on commercial cloud services, but research activities are at best a minor component of the commercial cloud industry's activities and business. Jetstream is designed from the hardware to the user interface for the purpose of supporting academic research. A particularly important aspect of the interface design is a library of "pre-built" VMs that can be instantiated by any user, customized, and then stored in a way that is either private or publicly available to the community.

The name Jetstream is rooted in the reason for the system's existence - it responds to a need expressed by the US National Science Foundation to expand the diversity and size of the community of researchers and students that make use of NSF-funded cyberinfrastructure. It was this call to which the IU Pervasive Technology Institute responded when proposing Jetstream, and it is this sense

that was the inspiration for the name Jetstream. In the upper atmosphere, the Jetstream is a zone of rapid air movement that lies at the boundary of two large masses of air. In the US cyberinfrastructure ecosystem, our goal was for Jetstream to function as a rapidly available, responsive cyberinfrastructure facility at the boundary of existing NSF-funded cyberinfrastructure.

What we discovered, however, was that even with an interactive cloud system designed to be easily used and effective in supporting many fields of research [25–27, 38], it took considerable outreach and consulting assistance to develop a community of users. Jetstream staff have now given hundreds of talks to tens of thousands of audience members about Jetstream's value. The fall and spring of 2016-2017 was particularly important in increasing community awareness. We worked with researchers unfamiliar with both Jetstream and with NSF allocation processes to help them submit high quality proposals for allocations on Jetstream, get those allocations, and then make good use of those allocations. The XSEDE Campus Champions were particularly important in promoting understanding of Jetstream's existence and capabilities. The importance of humans in the loop was such that the NSF approved and funded a supplemental request for additional staff to do outreach, training, and consulting about Jetstream. Unlike in a famous baseball movie, "build it and they will come" was not the path to the widespread use that Jetstream now enjoys. Build it, tell people about it, tell people about it more, and do intensive consulting to get some early success stories was the path to the widespread positive impression that the US research community now has of Jetstream. Were it not for humans in the loop promoting the value of Jetstream and facilitating its research, it would have served nowhere near the number of researchers and students that it has served thus far.

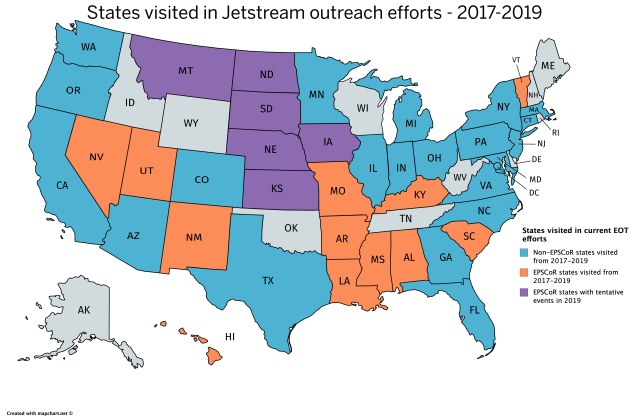

Figure 3: States visited during Jetstream EOT events

4.5 Developing outreach efforts for Jetstream

Because Jetstream is very different in nature from the other resources provided in the XSEDE infrastructure, efforts needed to be made to educate potential users and communities as to what Jetstream is, how it is different from traditional HPC, and how researchers might best make use of the resource. In addition, one charge from the NSF in the award solicitation was in reaching new communities not served by traditional HPC or current infrastructure efforts [XXXXXX 14536-Citation]. This in itself required looking beyond the communities already engaged with the XD program and into under served and under represented communities and institutions.

The goals for Jetstream outreach efforts were along two primary lines. The first was to make large communities of researchers aware, which meant speaking at larger venues. The second was engaging under served groups more directly, which entailed doing more hands on training events in smaller settings. The first step to both paths was to identify potential audiences. While getting the word out was important in general, finding the events where the most potential Jetstream users would be found was crucial. The Jetstream education, outreach, and training (EOT) team started identifying conferences, both national and regional, where user communities that required capabilities outside of normal HPC limits such as long run times, root access, or on-demand computing might gather. In addition, the team also began working with XSEDE Campus Champions, the National Center for Genome Analysis Support (NCGAS) and contacts at various academic and research institutions across the United States, looking especially at institutions in Established Program to Stimulate Competitive Research (EPSCoR) jurisdictions. [25]

Early engagements in 2015 and early 2016 were entirely theoretical in nature as the production system was not online and available to users. These efforts were solely to start making researchers aware that the resource was coming. Once the system was online and the first allocations awarded, it was possible to show work actual users were doing and to adequately demonstrate how researchers and communities could utilize Jetstream for their own efforts. The Jetstream team then made efforts to engage well-known researchers that conducted large scale workshops to train other researchers and students. This exposure yielded excellent results as it gave practical, first-hand experience in using Jetstream for their specific domain science as led by respected researchers in those areas. In addition, the EOT team felt it was important to encourage both research and education efforts using Jetstream by helping allocation applicants with proofreading and review efforts to help ensure successful allocation submissions.

Over time, the engagement efforts grew by adding additional members to the EOT team and also via organizations like NCGAS embracing Jetstream for their educational efforts. Also, the types of outreach efforts have evolved from basic talks to hands on tutorials using Jetstream's Atmosphere interface to Jetstream command line interface (CLI) hands on tutorials to various data science workshops that utilize Jetstream as the interactive computational resource. The quest for additional venues to discuss Jetstream is an on-going effort. There seems to be no shortage of communities and researchers that would benefit from a freely available research cloud resource. The Jetstream EOT team will continue over the course of the grant to make an effort to reach these communities and help enable science for researchers from all over the US.

4.6 Human support of remote systems: Cornell Center for Advanced Computing

Cornell University Center for Advanced Computing (CAC) provides human support for a number of HPC and Cloud-based compute resources that are located at Cornell or at other research computing centers. These activities are based around creating documentation modules, called Cornell Virtual Workshops, that can be accessed by system users, with topics from basic "getting started materials" to more advanced topics. The Cornell Virtual Workshop (CVW) is a set of web-based, asynchronous learning modules on advanced computing topics ranging from high-performance parallel computing to data analysis and visualization. Begun in 1994, the Cornell Virtual Workshop has many advantages common to online training, including (a) they are always available as a 24x7 option for users who want to study a topic on demand and at their own pace, (b) google searches include CVW pages, (c) they are updated in-place, i.e. there is no need to get a "new copy", (d) they can provide an information-rich environment, e.g. the built-in HPC glossary, and (e) they provide an experience-rich environment, with videos, simulations, exercises, and quizzes. The Cornell CAC has received numerous grants from the National Science Foundation (NSF), the Department of Defense (DOD), and private industry to develop and deploy Cornell Virtual Workshops. In order to provide CVW modules, CAC works closely with the service provider or client requesting a new module to identify requirements for what should be described. To maintain consistency across CVW, CAC uses a standard style for all types of text that conforms closely to technical documentation from other sources. Finally, to ensure virtual workshop materials are solid for publication, new materials go through a multi-part review process, including peer, supervisor, and external reviews, in order to identify and fix any issues.

Today, CAC is developing and deploying online training for Jetstream, Stampede2, and XSEDE resources. Over the past ten years, there have been over a million content page visits. The most recent XSEDE Annual Survey results showed that "Consistent with previous years, all training methods are rated highly (all but one is above 3.5), but XSEDE users continue to express clear preferences for self-serve, self-paced, just-in-time options", with online training averaging 4.19 out of 5. Unfortunately, understanding how online training availability affects future success has been very difficult to track, due to both the long-term nature of tracking training through education through career, and to the anonymous use of online materials; when materials are password-protected, fewer people choose to cross that small barrier, and search tools often cannot deliver the same pointers to the needed materials.

5 WELL-SKILLED HUMANS AS OUTPUT: EFFECT OF TRAINING AND CAREER DEVELOPMENT

In any discussion about humans and cyberinfrastructure there is widespread agreement that there is more demand for people well-trained and well-versed in the use of cyberinfrastructure than there is a supply of such people. There is also widespread agreement that knowledge of cyberinfrastructure can aid the career of a graduate student, an early career academic, or any research professional. However, we have yet to find any studies that are specific to enhancement of career opportunities or acceleration of career growth based on the presence (or lack thereof) of cyberinfrastructure skills.

6 PRIZES AND PATENTS

"The Nobel Prize is considered the most prestigious award in the world" [9]. Nobel prizes are awarded in scientific areas including physics, chemistry, economics, and physiology or medicine. They are very much a "trailing indicator" but when a CI facility has contributed to such an event, it is indeed, a big deal. XSEDE has supported research work that contributed to three Nobel Prizes (one in Chemistry, two in Physics). The Open Science Grid contributed to the same two Nobel Prizes in physics. And Indiana University, which has been involved in XSEDE and the OSG, tallies all three as accomplishments to which it contributed.

Contributions to patents seems like another indicator that could be explored, just like contributions to work that results in a Nobel Prize. However, we can find no published or online listings of Patent awards associated with use of advanced CI facilities.

7 DISCUSSION

Looking at the state of tools and processes for quantitative analyses of financial returns on investment in cyberinfrastructure [35], one can see common themes emerging regarding how such analyses are done. Additionally, there are a number of peer-reviewed and other technical reports focusing on cost effectiveness and financial ROI of investments in cyberinfrastructure. Going back to a logic model of organizational processes, the available work on financial ROI for investments in cyberinfrastructure tells us collectively a great deal about how we use cyberinfrastructure, and in particular how financially effectively we make technology choices about cyberinfrastructure hardware to support research activities.

There are, at present, far fewer published works that have to do with assessing the impact of cyberinfrastructure in non-financial terms, which is of course the more important aspect of the use of cyberinfrastructure: the "what is done" question as opposed to the "how is it done" question. Some of the lack of studies is simply the result of common sense: recent accomplishments such as the Nobel Prizes cited above, or the recent visualization of a black hole, simply and clearly could not have been done without use of cyberinfrastructure. On the other hand, methods developed by developed by von Laszewski, Apon, Boerner, Furlani, and their colleagues could be used to assess quantitatively the value of cyberinfrastructure in accelerating the speed of research accomplishments and supporting research accomplishments that are more important (or at least more widely cited) than research done without the benefit of advanced cyberinfrastructure facilities.

Assessing the role of humans in the loop in supporting cyberinfrastructure in general, and cloud computing in particular, remains challenging. We have a number of narratives, a handful of qualitative studies, and few quantitative studies that investigate this area, and it is critically important that the cyberinfrastructure community engage in more research in this area. The experiences cited here relative to the NSF-funded Jetstream cloud system, for example, make two strong assertions: 1) because that system is developed from the hardware up through all layers of the software stack specifically to

support academic research, the system is significantly better suited to supporting academic research than commercial cloud systems; 2) even with a system designed to support academic research, humans played and continue to play a critical role in the system's use and adoption. Current sales pitches for the use of commercial cloud computing systems, from some sources at least, can be read as minimizing the need for humans facilitating research in the cloud. In order for the cyberinfrastructure community to facilitate research productivity as more and more computing is done in the cloud, it is important that we be able to articulate the value of humans enabling research in the cloud. Without being able to demonstrate the value of humans in supporting research in the cloud, we run the risk of watching experiments in which researchers are expected to use cloud facilities without essential support.

8 CONCLUSIONS

This paper has discussed tools useful in understanding the impact of cyberinfrastructure, particularly people, on research outcomes in terms other than financial terms. The published literature - particularly peer-reviewed literature - regarding assessment of value of cyberinfrastructure in financial terms is well ahead of the literature on non-financial aspects. That means that there is more careful science about the *how* of how cyberinfrastructure aids research processes than there is about the *what* of what cyberinfrastructure helps produce. Furthermore, there are tools available that would aid in such assessments that are not yet widely used in analysis of the impact of cyberinfrastructure. Finally, there is considerable and pressing need for more quantitative and qualitative studies of the importance of humans in facilitating research in the cloud in particular. Without such continued work, there is a risk that research organization leaders may fail to appreciate the level of investment in humans that is required to facilitate and accelerate research based in use of cyberinfrastructure facilities, both physical and virtual.

REFERENCES

[1] [n. d.]. Asia-Pacific Grid. http://www.apgridpma.org/. Accessed: 2019-07-01.
[2] [n. d.]. CyVerse Publications. http://www.cyverse.org/about/. Accessed: 2019-07-01.
[3] [n. d.]. European Data Initiative. https://eudat.eu/. Accessed: 2019-07-01.
[4] [n. d.]. European Grid Initiative. https://www.egi.eu/. Accessed: 2019-07-01.
[5] [n. d.]. European Open Science Cloud. https://ec.europa.eu/research/openscience/index.cfm?pg=open-science-cloud. Accessed: 2019-07-01.
[6] [n. d.]. G-Index. https://en.wikipedia.org/wiki/G-index. Accessed: 2019-07-01.
[7] [n. d.]. Galaxy Publications. https://galaxyproject.org/publication-library/. Accessed: 2019-07-01.
[8] [n. d.]. National Center for Atmospheric Research. https://ncar.ucar.edu. Accessed: 2019-07-01.
[9] [n. d.]. The Nobel Prize. https://sweden.se/society/the-nobel-prize-awarding-great-minds/. Accessed: 2019-07-01.
[10] [n. d.]. Open Science Grid. https://opensciencegrid.org/. Accessed: 2019-07-01.
[11] [n. d.]. The OpenStack Foundation. https://www.openstack.org. Accessed: 2019-07-01.
[12] [n. d.]. Pervasive Technology Institute Publications. https://bibbase.org/service/mendeley/42d295c0-0737-38d6-8b43-508cab6ea85d. Accessed: 2019-07-01.
[13] [n. d.]. Scopus. https://www.scopus.com/. Accessed: 2019-07-01.
[14] [n. d.]. UITS User Satisfaction Survey. http://www.iu.edu/~uitssur/. Accessed: 2019-07-01.
[15] 2004. W.K. Kellog Foundation Logic Model Development Guide.
[16] 2018. Colleges With the Greatest Percentage Decreases in Full-Time Undergraduates. *Chronicle of Higher Education* 65, 15 (12 2018). https://www.chronicle.com/article/Colleges-With-the-Greatest/245285
[17] Amy Apon, Stanley Ahalt, Vijay Dantuluri, Constantin Gurdgiev, Moez Limayem, Linh Ngo, and Michael Stealey. 2010. High Performance Computing Instrumentation and Research Productivity in U.S. Universities. *Journal of Information Technology Impact* 10, 2 (9 2010), 87–98. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1679248
[18] Amy W. Apon, Linh B. Ngo, Michael E. Payne, and Paul W. Wilson. 2015. Assessing the effect of high performance computing capabilities on academic research output. *Empirical Economics* 48, 1 (01 Feb 2015), 283–312. https://doi.org/10.1007/s00181-014-0833-7
[19] Jill Barshay. 2018. College students predicted to fall by more than 15% after the year 2025. *The Hechinger Report* (9 2018). https://hechingerreport.org/college-students-predicted-to-fall-by-more-than-15-after-the-year-2025/
[20] Katy Börner, Richard Klavans, Michael Patek, Angela M Zoss, Joseph R Biberstine, Robert P Light, Vincent Larivière, and Kevin W Boyack. 2012. Design and update of a classification system: The UCSD map of science. *PloS one* 7, 7 (2012), e39464.
[21] Clarivate. 2019. Web of Science Kernel. https://clarivate.com/products/web-of-science/
[22] C. Collberg and T.A. Proebsting. 2016. Repeatability in Computer Science Systems Research. *Commun. ACM* 59, 3 (2016), 62–69.
[23] Open Geospatial Consortium. 2019. OGC Testbed-14: Federated Clouds Engineering Report. http://www.opengis.net/doc/PER/t14-D023.
[24] Lizanne DeStefano and Lorna Rivera. 2015. *Cyberinfrastructure Value Assessment Report.* Technical Report.
[25] Jeremy Fischer, Brian W Beck, Sanjana Sudarshan, George Turner, Winona Snapp-Childs, Craig A Stewart, and David Y Hancock. 2018. Methodologies and practices for adoption of a novel national research environment. *Proceedings of the Practice and Experience on Advanced Research Computing-PEARC* 18 (2018).
[26] Jeremy Fischer, David Y Hancock, John Michael Lowe, George Turner, Winona Snapp-Childs, and Craig A Stewart. 2017. Jetstream: A cloud system enabling learning in higher education communities. In *Proceedings of the 2017 ACM Annual Conference on SIGUCCS*. ACM, 67–72.
[27] David Y Hancock, Craig A Stewart, Matthew Vaughn, Jeremy Fischer, John Michael Lowe, George Turner, Tyson L Swetnam, Tyler K Chafin, Enis Afgan, Marlon E Pierce, et al. 2017. JetstreamâĂŤEarly operations performance, adoption, and impacts. *Concurrency and Computation: Practice and Experience* (2017), e4683.
[28] Jorge E Hirsch. 2005. An index to quantify an individual's scientific research output. *Proceedings of the National academy of Sciences* 102, 46 (2005), 16569–16572.
[29] Jeremiah P. Ostriker, Paul W. Holland, Charlotte V. Kuh, and James A. Voytuk. 2011. A Data-Based Assessment of Research-Doctorate Programs in the United States (2011). Available online at http://www.nap.edu/rdp/.
[30] Jeffrey T. Palmer, M. Gallo, Steven, Thomas R. Furlani, Matthew D. Jones, Robert L. DeLeon, Joseph P. White, Nikolay Simakov, Abani K. Patra, Jeanette Sperhac, Thomas Yearke, Ryan Rathsam, Martins Innus, Cynthia Cornelius, James Browne, William L. Barth, and T. Evans, Richard. 2015. Open XDMoD: A Tool for the Comprehensive Management of High Performance Computing Resources. *Computing in Science and Engineering* 17, 4 (2015), 52–62.
[31] Christopher Peebles, Craig Stewart, Brian Voss, and Sue Workman. 2001. Measuring quality, cost, and value of it services. EDUCAUSE Conference Proceedings, Indianapolis IN.
[32] Teresa Penfield, Matthew J Baker, Rosa Scoble, and Michael C Wykes. 2014. Assessment, evaluations, and definitions of research impact: A review. *Research Evaluation* 23, 1 (2014), 21–32.
[33] Olga Scrivener, Gangadeep Singh, Sara E. Bouchard, Scott C. Hutcheson, Ben Fulton, Matthew R. Link, and Katy Borner. 2018. XD Metrics on Demand Value Analytics: Visualizing the Impact of Internal Information Technology Investments on External Funding, Publications, and Collaboration Networks. *Frontiers in Research Metrics and Analytics* (2018). https://doi.org/10.3389/frma.2017.00010
[34] Craig A Stewart, William K Barnett, Eric A Wernert, Julie A Wernert, Von Welch, and Richard Knepper. 2015. Sustained Software for Cyberinfrastructure: Analyses of Successful Efforts with a Focus on NSF-funded Software. In *Proceedings of the 1st Workshop on The Science of Cyberinfrastructure: Research, Experience, Applications and Models.* ACM, 63–72.
[35] Craig A Stewart, David Y Hancock, Julie Wernert, Thomas Furlani, David Lifka, Alan Sill, Nicholas Berente, Donald F McMullen, Thomas Cheatham, Amy Apon, et al. 2019. Assessment of financial returns on investments in cyberinfrastructure facilities: A survey of current methods. (2019).
[36] Craig A. Stewart, David Y. Hancock, Julie Wernert, Matthew R. Link, Nancy Wilkins-Diehr, Therese Miller, Kelly Gaither, and Winona Snapp-Childs. 2018. Return on Investment for Three Cyberinfrastructure Facilities: A Local Campus Supercomputer, the NSF-Funded Jetstream Cloud System, and XSEDE (the eXtreme Science and Engineering Discovery Environment). In *2018 IEEE/ACM 11th International Conference on Utility and Cloud Computing (UCC)*. IEEE, 223–236. https://doi.org/10.1109/UCC.2018.00031

[37] Craig A. Stewart, R. Knepper, M.R. Link, M. Pierce, E.A. Wernert, and N. Wilkins-Diehr. 2017. Cyberinfrastructure, Cloud Computing, Science Gateways, Visualization, and Cyberinfrastructure Ease of Use. In *Encyclopedia of Information Science and Technology, Fourth Edition* (fourth ed.), M. Khosrow-Pour (Ed.). IGI Global, Hershey, PA, 1063–1074.

[38] D.Y. Hancock*-T. Miller J. Fischer L. Liming G. Turner J.M. Lowe S. Gregory E. Skidmore M. Vaughn D. Stanzione N. Merchant I. Foster J. Taylor P. Rad V. Brendel E. Afgan M. Packard W. Snapp-Childs. Stewart*, C.A. 2018. Jetstream - a novel cloud system for science. In *Contemporary High Performance Computing: From Petascale toward Exascale, Volume Three*, J. Vetter (Ed.). Oxford University Press, Oxford, 189–222.

[39] F. Pérez B. Granger M. Bussonier J. Frederic K. Kelley J. Hamrick J. Grout S. Corlay P. Ivanov D. Avila S. Abdalla C. Willing T. Kluyver, B. Ragan-Kelley and Jupyter Development Team. 2016. Jupyter NotebooksâĂŤa publishing format for reproducible computational workflows. https://dx.doi.org/10.3233/978-1-61499-649-1-87.

[40] PR Thomas and David S Watkins. 1998. Institutional research rankings via bibliometric analysis and direct peer review: A comparative case study with policy implications. *Scientometrics* 41, 3 (1998), 335–355.

[41] John Towns, Timothy Cockerill, Maytal Dahan, Ian Foster, Kelly Gaither, Andrew Grimshaw, Victor Hazlewood, Scott Lathrop, Dave Lifka, Gregory D Peterson, et al. 2014. XSEDE: accelerating scientific discovery. *Computing in Science & Engineering* 16, 5 (2014), 62–74.

[42] Viswanath Venkatesh, Michael G Morris, Gordon B Davis, and Fred D Davis. 2003. User acceptance of information technology: Toward a unified view. *MIS quarterly* (2003), 425–478.

[43] Gregor von Laszewski, Fugang Wang, Geoffrey C Fox, David L Hart, Thomas R Furlani, Robert L DeLeon, and Steven M Gallo. 2015. Peer comparison of xsede and ncar publication data. In *2015 IEEE International Conference on Cluster Computing*. IEEE, 531–532.

[44] G. von Laszewski, F. Wang, G. C. Fox, D. L. Hart, T. R. Furlani, R. L. DeLeon, and S. M. Gallo. 2015. Peer Comparison of XSEDE and NCAR Publication Data. In *2015 IEEE International Conference on Cluster Computing*. 531–532. https://doi.org/10.1109/CLUSTER.2015.98

ACKNOWLEDGMENTS

This report was supported by the Indiana University Pervasive Technology Institute and by NSF Awards 1445604, 1053575, 1548562, 1362134, 1405767, 1243436, 1148996 and 0946726. Thanks to our many collaborators, particularly R. Eigenmann who suggested this line of research. Thanks to Harmony Jankowski for editing, Lauren Huber for graphics, and Monica Shannon for logistical support.

CAS and DYH contributed equally to this report. CAS bears responsibility for any and all mistakes.