XDMoD Value Analytics: Visualizing the Impact of Internal IT Investments on External Funding, Publications, and Collaboration Networks

Olga Scrivner*[1], Gagandeep Singh[1], Sara E. Bouchard[1], Scott C. Hutcheson[1], Ben Fulton[2], Matthew R. Link[2], Katy Börner[1]

[1]Department of Intelligent Systems Engineering, School of Informatics, Computing and Engineering, Indiana University, Bloomington, Indiana, USA
[2]Pervasive Technology Institute, Indiana University, Bloomington, Indiana, USA

*** Correspondence:**
Corresponding Author
obscrivn@indiana.edu

Keywords: information visualization, scientometrics, impact analysis, grant income, return on investment, value analytics, high-performance computing

Abstract

Many universities invest substantial resources in the design, deployment, and maintenance of campus-based cyberinfrastructure. To justify the expense, it is important that university administrators and others understand and communicate the value of these internal investments in terms of scholarly impact. This paper introduces two visualizations and their usage in the Value Analytics (VA) module for Open XD Metrics on Demand (XDMoD), which enable analysis of external grant funding income, scholarly publications, and collaboration networks. The VA module was developed by Indiana University's (IU) Research Technologies division, Pervasive Technology Institute, and the Cyberinfrastructure for Network Science Center (CNS), in conjunction with the University at Buffalo's Center for Computational Research (CCR). It provides diverse visualizations of measures of information technology (IT) usage, external funding, and publications in support of IT strategic decision making. This paper details the data, analysis workflows, and visual mappings used in two VA visualizations that aim to communicate the value of different IT usage in terms of NSF and NIH funding, resulting publications, and associated research collaborations. To illustrate the feasibility of measuring IT values on research, we measured its financial and academic impact from the period between 2012 and 2017 for IU. The financial return on investment (ROI) is measured in terms of IU funding, totaling $339,013,365 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. In addition, the results show that *Medical Specialties*, *Brain Research*, and *Infectious Diseases* are the top three scientific disciplines ranked by the number of publications during the given time period.

1 Introduction

Access to high-performance computing (HPC) systems and advanced cyberinfrastructure generally is critical to advance research in many scholarly fields. Over the last thirty years, supercomputing has expanded from a few monolithic and extremely fast computing systems into a comprehensive "set of organizational practices, technical infrastructure and social norms that collectively provide for the

smooth operation of research and education work at a distance" (Towns et al. 2014). This new form of cyberinfrastructure (Stewart et al. 2017) is also used by a large number of researchers, scholars, and artists, and is more complicated, including high performance computing, storage systems, networks, and visualization systems. This requires a new suite of metrics that can help different stakeholders better understand the value of research cyberinfrastructure. In addition to job- and system-performance monitoring, metrics now available and needed to understand system usage and value include usage modality (e.g., central processing unit [CPU] usage per group or person, number of users, and wait time). All of these metrics are essential to better understand "why users do what they do and how they leverage multiple types and instances of cyberinfrastructure (CI) resources" (Katz et al. 2011). In recent years, there has also been growing interest in measuring the impact of CI on scientific research and publication outcome (e.g., Knepper and Börner 2016, Fulton et al. 2017, Madhavan et al. 2014, among others). Such insights are particularly relevant for campus CI that require significant investment and long-term strategic and financial planning. Visualizations help communicate the value of CI to diverse stakeholders ranging from domain experts to academic deans to financial administrators.

Currently, several CI frameworks enable on-demand rendering of impact metrics (e.g., funding, publications, and citations) that result from using HPC resources. While most monitoring tools for HPC are traditionally "largely passive and local in nature" (Furlani et al. 2013), these new frameworks, such as Deep Insight Anytime, Anywhere (DIA2) (Madhavan et al. 2014) and Extreme Science and Engineering Discovery Environment (XSEDE) Metrics on Demand (XDMoD VA) (Fulton et al. 2017), are open-source, customizable systems with "increased functionality, an improved interface, and high-level charting and analytical tools" (Palmer et al. 2015). DIA2 is a web-based visual analytics system aiming to assess research funding portfolios (Madhavan et al. 2014). Open XDMoD is one of the most widely used software systems in the U.S. and is currently employed by more than 200 institutions to evaluate their HPC usage (Palmer et al. 2015, Fulton et al. 2017). The tool has been developed as an open source software for "metrics, basic accounting, and visualization of CPU (central processing units) and storage usage" at the Center for Computational Research (CCR) of the University at Buffalo (Palmer et al. 2015).

Open XDMoD VA adds a value analytics module to Open XDMoD. The two visualizations presented in this paper contribute to XDMoD VA functionality as follows:

1. Financial and Intellectual Analytics: Analyze grant income and publications by researchers that use Information Technology (IT) and relate that income to use of local IT systems.
2. Co-PI collaboration networks: Analyze research collaborations by researchers that use local IT systems.

In this paper, we detail both types of visual analytics for XDMoD VA and demonstrate how they can help understand the IT impact on academic research. Specifically, we describe the data and visual analytics workflows used for the *Funding and Publication Impact* and the *Co-PI Collaboration Network* visualizations. The *Funding and Publication Impact* visualization uses a Sankey graph to interlink IT usage with funding and publication output. The *Co-PI Collaboration Network* uses NSF and NIH funding data to extract and depict Co-PI collaboration networks together with listings of scholars ranked by the total amount of grants. Both types of visualizations are interactive, supporting overview first, zoom and panning, and details on demand (Shneiderman 1996). The visualizations are rendered using the Web Visualization Framework (WVF) developed by CNS at IU that allows for an effective, highly customizable rendering of interactive visualizations.

The remainder of the paper is organized as follows: Section 2 discusses prior work on analyzing and visualizing the impact of IT resources, particularly on scholarly output. In Section 3, we outline the data we used, as well as the preprocessing methods needed to render that data useful. Section 4 discusses the general methods applied to render data into visual insights. Section 5 details the results of using the data and methods to render the *Funding and Publication Impact* and *Co-PI Collaboration Network* visualizations. Key insights and planned developments are discussed in Section 6.

2 Related Work

In recent years, there has been growing interest in measuring the impact of high-performance computing (HPC) on scientific research and publication outcome. Advanced cyberinfrastructure resources require significant investment, particularly when implemented at the campus level, and insights about the value of such investments in financial and intellectual terms are essential for the strategic and financial planning of academic institutions (Fulton et al. 2017). A number of studies have examined the relationship between these values and CI resources usage. Three studies can be regarded as a starting point for introducing HPC user metrics: namely, Li et al. (2005), Iosup et al. (2006), and Lee et al. (2006). In addition to traditional metrics (e.g., job size, system utilization), these studies included user and group characteristics to analyze system performance. Hart (2011) extended this idea and highlighted the importance of usage and submission patterns to understand users and their behavior across HPC resources. Knepper (2011) further investigated the relation between users, their HPC usage behavior, and their field of science. Particularly, his study examined PIs, their network affiliation, and the scientific field and allocation size of their research projects from 2003 to 2011 using TeraGrid, a national (US) computing scientific infrastructure. The results revealed that PIs constituted 23% (3,334) of the total TeraGrid project users (14,474) and that molecular biosciences and chemistry are the two scientific fields with the highest number of projects involving TeraGrid usage, 2,292 and 1,828 respectively. Similarly, Furlani et al. (2012, 2013) observed that molecular biosciences recently joined physics at the top of the list of sciences with the greatest CPU usage of Extreme Science and Engineering Discovery Environment (XSEDE), a virtual system providing digital resources and computing services (Towns et al. 2014). In addition, Furlani's longitudinal study demonstrated a substantial increase in the number of PIs using XSEDE from ~500 in 2005 to ~1,600 in 2011. The authors noted, however, that often a project PI did not personally utilize the XSEDE resources, assigning computing tasks to graduate students or postdocs.

Other research focuses on various metrics for measuring academic performance with respect to HPC usage. Apon et al. (2010) suggest measuring the HPC investment in terms of "competitiveness," presented as a ranking system. Ranks are calculated using the Top 500 HPC list that reports on the fastest 500 computers in the world. An institution's rank is based on their investments in HPC according to the Top 500 HPC list. In their study, academic performance was characterized by publication counts and funding awards. Their funding showed that "consistent investments in HPC at even modest levels are strongly correlated to research competiveness." Apon et al.'s study also presented statistical evidence that NSF research funding and publication counts are good predictors of academic competitiveness. Subsequently, Knepper and Börner (2016) at the relationship between the CPU usage of XSEDE resources utilized by PIs and publication records. In addition, they mapped fields of science into HPC resources, creating a bipartite network. The results demonstrated that among the 27 top fields, physics, chemistry, astronomy, material research, and meteorology/climate modeling utilized XSEDE resources the most. They also observed that the type of HPC resources

plays an important role, as some systems link to almost all fields of science (e.g., NICS Kraken), whereas others serve only a small number of fields.

Newby et al. (2014) discussed various forms of "return on investment" in studies of research cyberinfrastructure. Their discussion included a number of different types of value derived from investment in advanced cyberinfrastructure, including scientific value, workforce development, economic value, and innovation. More recent work has adopted the financial definition of ROI: "a ratio that relates income generated … to the resources (or asset base) used to produce that income," calculated typically as "income or some other measure of return on investment." Values greater than 1.0 indicate that return is greater than investment (Kinney and Raiborn 2011). One of the challenges in measuring ROI in financial terms is that returns may take decades to materialize in some disciplines (Stewart et al. 2015). Recently, two studies have applied ROI metrics to study the impact of HPC usage at Indiana University. Thota et al. (2016) compared the annual cost of operating IU's Big Red II supercomputer with the funds brought into the university by the researchers that used Big Red II. The expected annual average cost for Big Red II is around $15 million dollars. Thota et al. found that for the year 2013, the total grant income to IU by PIs or Co-PIs who make use of this supercomputer was more than double the amount of the cost ($39.8 million). This is suggestive (although not conclusive) of a favorable financial ROI – only suggestive because the analysis of Thota et al. was not able to take into account how critical use of the Big Red II supercomputer was to the grants awarded to users of that system. Similarly suggestive, Fulton et al. (2017) showed a positive correlation between an increase in HPC usage and IU's award funding over a period of years. Stewart et al. (2015) analyzed the ROI for XSEDE (the eXtreme Science and Engineering Discovery Environment) and argued that the ROI of federal investment in this resource was greater than 1.

3 Data Acquisition and Preparation

Four data sources are used to analyze and visualize the impact of internal IT usage on external funding, associated publications, and collaboration networks: IT usage data for faculty, staff, and students working at an institution; IU award database; NIH and NSF award data for the same institution together with publications that list these awards in the acknowledgements. These datasets, as well as their matching, cleaning, and preparation for visualization, are detailed subsequently. The overall process is illustrated in Figure 1. Please note that all public data and all code is available online at http://cns.iu.edu/2017-Value-Analytics.html.

** Place Figure 1 here

3.1 IT Usage Data

XDMoD HPC resource usage log data is used to extract IT usage information. Among others, the logs contain five elements: IT system type, IT system name, units used (CPU hours for computing and Gigabytes for storage), user name, first name, and last name of IT user. An IT system consists of two types, namely storage and compute node, and several systems within each type.

From January 2012 to October 2017, there were two major storage and six computing systems utilized by (Co-)PIs for NSF and NIH grants at Indiana University.

Storage:

1. Scholarly Data Archive (SDA),
2. Data Capacitator 2 (DC2)

Computing:

1. Big Red, a supercomputer,
2. Big Red II, a supercomputer,
3. Karst, a cluster for serial jobs,
4. Mason, designed for data-intensive, high-performance computing tasks (Thota et al. 2016),
5. Quarry, a computing cluster, and
6. Carbonate, designed for data-intensive computing, particularly for genome and phylogenic software.

Among computing systems, Big Red was decommissioned in 2013 and replaced by Big Red II, Quarry was decommissioned in January 2015, and Carbonate became available in July 2017 to replace Mason, scheduled to retire on January 1, 2018.

HPC log files do not differentiate between a group and a single user account. Using unique IT user names and the fields with last and first names, 1,187 instances of group accounts were identified and removed. The log files were then filtered by the year > 2012-01-01 with the storage and computer resources > 0. Tables 1 and 2 provide a summary of IT resource usage at IU from January 2012 to October 2017 with a total of 65,495,233,153 CPU job hours run, 114,893 GB stored, and 4,112 unique users active.

Table 1. Summary of HPC computing jobs at IU by individual users (Jan 2012 - Oct 2017)

Big Red	Carbonate	Mason	Quarry	Karst	Big Red II	Total CPU-Hours	Users
119	134	635	699	1,298	1,311	65,495,233,153	4,197

Table 2. Summary of HPC storage usage at IU by individual users (Jan 2012 - Oct 2017)

DC2	SDA	Total Gigabytes	Users
252	1,687	114,893	1,939

3.2 Internal IU Award Database

IU internal grant data is generated from Kuali Financial Services. This database imports user ID, the name of award agency, grant ID, and total amount. For this paper, the following query was specified: a) the start date is 2012-01-01 and b) the grant amount is greater than zero. Out of the total of 28,965 awards between 2012-01-01 and 2018-01-01, 597 grants are from NSF and 2,677 grants are from NIH agencies. The number of unique (Co-)PI users are 425 for NSF and 690 for NIH.

3.3 NIH Grant and Publication Data

NIH grant award numbers and linkages to publications that cite award numbers can be downloaded in bulk using ExPORTER Data Catalog[1] or as a data extract using Research Portfolio Online Reporting Tools (RePORTER).[2] The NIH data provides access to both intramural and extramural NIH-funded research projects from the past 25 years and publications since 1980 (NIH 2017). For this study, we have used the RePORTER; however, our methodology is applicable to ExPORTER. When extracting data from the ExPORTER, NIH files are downloaded separately for each year and then merged, whereas the RePORTER output is already merged. The RePORTER query form also allows for generating data by means of query elements, such as keywords, organization names, publications, and others. By combining these elements, the user is able to create highly customized searches of this NIH funding database. In order to obtain the grant total and the number of publications for XDMoD, the following query filters were applied: project year (2012-2017), organizations listed in the NIH lookup (IU Bloomington, IU South Bend, and IUPUI), state (Indiana), and publication year (2012-2017), as illustrated in Figure 2. The *Agency/Institute/Center* field is kept with its default set to "admin," and subprojects are set to be excluded.

** Place Figure 2 here

The query was run on August 22, 2017, 9am EST and the query results were exported in CSV format with relevant fields: namely, *Project Number*, *Contact PI / Project Leader*, *Other PI or Project Leader(s), FY Total Cost by IC*, and *Funding IC*. Publication data was exported with the following fields: *Core Project Number*, *ISSN*, *Journal*, *PMID*, *PUB Year*, and *Title*. The results comprised 933 grants and 9,838 unique publications that acknowledge funding by these grants.

3.4 NSF Grant and Publication

NSF grants and associated publications were downloaded using the NSF Award Search Web API.[3] The API supports highly customized queries. For this project, a query was run on October 31, 2017, 2pm EST, using the following filters:

1) *awardeeName*="Indiana University",
2) *startDateStart*=01/01/2012 and
3) *printFields*=id,publicationResearch,agency,startDate,expDate,fundProgramName,title,piFirstName,piLastName,estimatedTotalAmt,coPDPI,primaryProgram, awardeeCity,awardeeName.[4]

The results comprised 565 unique awards and 245 unique publications. Data was retrieved in JSON format and converted to a CSV format.

3.5 Data Preparation

NIH Award–Publication Linkage. NIH grants and NIH publications are linked via the project number. An additional preprocessing step is required for this linkage. The project number from the grant file is given in a 14-digit format (e.g., 1R01HS022681-01), whereas the publication file is

[1] https://exporter.nih.gov/ExPORTER_Catalog.aspx

[2] https://projectreporter.nih.gov/reporter.cfm

[3] https://www.nsf.gov/developer/

[4] http://api.nsf.gov/services/v1/awards.json?awardeeName=%22Indiana+University%22&offset=26&startDateStart=01/01/2012&printFields=id,publicationResearch,agency,startDate,expDate,fundProgramName,title,piFirstName,piLastName,estimatedTotalAmt,coPDPI,primaryProgram, awardeeCity,awardeeName

assigned an 11-digit format (e.g., R01HS022681). Combining these two files results in 2,046 unique publication records linked to 293 grants (11-digit format). *Funding and Publication Impact* data then consist of IT resources, funding agencies, publications, journals, and grant total, merged by (Co-)PI and project number. The *Co-PI Collaboration Network* includes the names of PIs and Co-PIs, grant funding total, and the number of grants awarded.

NSF Award–Publication Linkage. For each award, the NSF API data retrieves publications in the form of a list, which is then split into three fields: authors, publication title, and publication journal. PIs' first and last name fields were merged, yielding 318 authors, 565 awards, and 245 publications. Among these awards, 55 awards are associated with publications, and 48 (Co-)PIs are associated with these 55 awards.

IT User–IU Award Linkage between IT usage data and IU award data was performed using user IDs. Next, IU awards were linked to NIH and NSF award-publication linkages via project numbers. The result is a table that links 61 IT users to the very same number of (Co-)PIs with 83 project awards, and 968 associated publications based on the unique PMID identifier in the case of NIH award or the unique publication title for NSF awards, as the NSF API data does not provide publication-unique identifiers. As a result of this merge, the *Funding and Publication Impact* data include IT resources, funding agencies, publications, journals, and grant total, merged by PI and project number. The total number of awards associated with IT resources is 657 for NIH and 228 for NSF awards, totaling $339,013,365. It should be noted that our main objective is to measure both financial and academic impacts. We have excluded the IU awards without publications and the awards for which the (Co-)PI did not use IT resources. As a result, the number of awards and their publications is lower than the total number of IU awards, thus totaling $ 21,016,055 for 83 NSF and NIH awards.

Data Aggregation was performed to determine the number of users per IT resource, the total award amount per NIH Institute or Center (IC), and the total number of publications per discipline of science. Table 3 exhibits the total NIH and NSF award amount per IT storage and IT resource for unique project IDs. NIH ICs identify which Center for Scientific Review (CSR) reviewed the grant application for a funding decision (NIH Research Portfolio Online Reporting Tools 2017). In contrast, NSF API does not provide such a field. Table 3 shows the list of NIH IC and NSF together with the number of awards, publications, and total award amount for the IU dataset.

Table 3. NIH ICs and NSF funding grants with IU (Co-)PIs that use IT resources

Funding Agencies	Number of funding awards	Number of publications	Sum of FY Total Cost by IC in $
NIH-NIGMS	14	156	4,057,619
NIH-NHLBI	6	183	2,655,533
NIH-NIAMS	6	116	1,992,701
NIH-NIAAA	9	61	1,523,761
NIH-NIMH	6	37	1,423,969
NIH-NCI	7	97	1,338,995
NIH-NEI	3	33	1,057,979

NIH-NIAID	2	10	730,446
NIH-NIDDK	3	9	674,441
NIH-OD	2	2	606,848
NIH-NIA	4	49	587,810
NIH-NIBIB	1	NA	500,696
NIH-NLM	3	86	364,995
NIH-NICHD	1	56	332,956
NIH-NCCIH	1	18	299,149
NIH-NIDA	1	5	154,000
NIH-NHGRI	1	1	5,000
NIH Total	70	919	18,306,898
NSF	13	49	2,709,157
Total	83	968	21,016,055

Publication records were aggregated using the UCSD map of science (Börner et al. 2012), a classification system that assigns each journal to one or more subdisciplines of science that are further aggregated into 13 disciplines of science (e.g., mathematics or biology).[5] It should be noted that some journal names retrieved from NIH and NSF vary considerably from the UCSD classification system; several preprocessing steps are necessary, such as lowering cases and normalizing punctuation. There were 245 cases where publication records did not match the UCSD map of science dictionary, and 33 publications were associated with more than one discipline with the same relative association proportion. For these publications, we created two additional categories, "Unclassified" and "Multidisciplinary." As a result, each publication is associated with one of the 13 disciplines of science. The number of publications per discipline is given in Table 4.

Table 4. Number of publication per discipline associated with the NIH and NSF awards

Discipline	**Number of Publications**
Medical Specialties	252
Unclassified	245
Brain Research	107
Infectious Diseases	82
Health Professionals	78
Biotechnology	59
Chemistry	44
Multidisciplinary	33
Social Sciences	30
Biology	23

[5] http://cns.iu.edu/2012-UCSDMap.html

Math & Physics	7
Chemical, Mechanical, & Civil Engineering	5
Electrical Engineering & Computer Science	2
Earth Sciences	1
Total	**968**

Finally, all aggregated data and linkage tables were converted to JSON format as required by the visualization plugin in the XDMoD VA portal. The conversion script is available at http://cns.iu.edu/2017-Value-Analytics.html. The JSON format specification for the *Funding and Publication Impact* and *Co-PI Collaboration Network* visualizations is illustrated in Figures 3 and 4.

** Place Figure 3 here

** Place Figure 4 here

4 Methods

4.1 Collaboration Network Extraction

The *Co-PI Collaboration Network* was extracted from NIH grant data as described in Section 3. Co-PIs were calculated by first matching the IU grant database's 'Agency Award Number' to the NIH 'Core Project Number', and then splitting the NIH 'Other PI or Project Leader(s)' field by semicolon and counting the number of grants and total grants for each co-author pair to compute the weight for collaboration edges. For NSF, the Co-PI information was extracted from the field 'coPDPI,' which also includes Co-PI's IDs. These IDs were removed and the field was split by comma.

4.2 Sankey Graphs and Force Network Layout in WVF

Sankey Graphs show the magnitude of flow between nodes in a network as well as the relationship between flows and their transformation (Riehmann et al. 2017). For the *Funding and Publication Impact* graph, we used the D3 Sankey API.[6] It reads input nodes and weighted links and computes positions using the Gauss–Seidel iterative method (Barrett et al. 1994). First, the horizontal position of the left-most nodes is computed; then, nodes further on right are positioned while minimizing link distance. After all the nodes are positioned, a reverse pass is made from right-to-left, and overlapping nodes are moved to minimize collision. The entire process is repeated several times to optimize the layout.

The final visualization features three types of nodes, namely IT resources on left, funding (e.g., NIH institutes and NSF) in middle, and publication disciplines on right. The height of a bar, or node, is proportional to the maximum of the weighted sum of incoming links and the weighted sum of outgoing links. The nodes are placed in ascending order by their heights.

Force-Directed Graphs are used to display the relationship between objects by calculating the position of each node based on their shared edges. The D3 force-directed graph[7] applies three primary forces upon the nodes: namely, the sum of all forces, a force between two linked nodes, and a central force using the layout algorithm by Dwyer et al. (2006). The WVF applies *linkStrength* as a

[6] https://bost.ocks.org/mike/sankey/

[7] https://bl.ocks.org/mbostock/4062045

parameter to calculate node positions and it is constant for all nodes.[8] In the *Co-PI Collaboration Network* visualization, the nodes are Co-PIs and edges represent their research collaborations. Node size corresponds to the number of grants and node color denotes the total funding amount in U.S. dollars per PI. Edge thickness indicates the number of co-authored grants.

The Web Visualization Framework (WVF) was used to build both visualizations.[9] WVF is a highly configurable packaging of several industry-standard web libraries (Angular, D3, HeadJS, Bootstrap, and many others) that allows its users to quickly build visualization applications. Existing WVF visualizations support rendering interactive horizontal bar graphs, geospatial maps, network graphs, bimodal graphs, science maps, and others. The WVF provides lightweight in-browser aggregation and analysis, but it relies on web services or external data sources to provide primary analyses. Applications built using WVF plugins allow each visualization within a page to use data from and interact with other visualization plugin elements, but through loosely coupled data connections. This allows for both the visualization elements and the aggregation and filtering methods to be replaced or removed without affecting other elements.

5 Results

This section explains how the data and methods discussed above are applied to IU institutional data and what insights were gained. Specifically, we describe the *Funding and Publication Impact* and *Co-PI Collaboration Network* visualizations, as well as the portal that supports easy access to both. The visualizations aim to help stakeholders understand the financial ROI measured in terms of total acquired funding and academic ROI measured by publications associated with these awards.

5.1 XDMoD VA Portal

The XDMoD portal is an interactive dashboard with an intuitive graphical interface to XDMoD metrics such as number of jobs, service units charged, CPUs used, or wait time (Furlani et al. 2012). XDMoD metrics can be broken down by field of science, institution, job size, principal investigator, and resource. Academic metrics (e.g., publications, citations, and external funding) can be uploaded by users or incorporated by institutions via Open XDMoD (Fulton et al. 2017). XDMoD Value Analytics adds new functionality by offering metrics on financial and scientific impact via visualization plugins, as illustrated in Figure 5. Key features of the VA interface include the ability to interact and drill-down, allowing users to access additional related information simply by clicking inside edges and nodes or selecting the desired filters.

** Place Figure 5 here

5.2 XDMoD Funding and Publication Impact Visualization

The *Funding and Publication Impact* visualization allows users to interactively explore the relations between IT resource usage (on left in Figure 6), funding awards aggregated by NIH institute and NSF (in middle), and publications that cite this funding aggregated by scientific discipline (on right). Sankey graph links take users on an exploratory quest, moving from the IT resources via funding to papers published in diverse scientific disciplines.

** Place Figure 6 here

[8] https://github.com/d3/d3-3.x-api-reference/blob/master/Force-Layout.md

[9] https://github.com/cns-iu/WVF

In this interactive visualization, users are provided with various functionalities on demand, such as mouse-over and selection; a legend explaining color and size coding can be viewed on demand. For example, 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, as illustrated in Figure 6. In this example, the user wishes to explore the connections between IT resource use, grant funding, and the number of publications in the field of *Brain Research* from *Scientific Discipline* (3). After hovering over one of the links connected to the *Brain Research* node from the *Scientific Discipline* category on the right, the common link will connect *Brain Research* publications with the funding agency from the *Funding* category (2) and *IT Resources* (1). Hovering over a link also brings up additional information relevant to the node (e.g., number of papers, here 54). To explore other links and nodes, the user can simply double-click on them to reset. The example in Figure 6 also illustrates how the user may gain insights from the visual data. In particular, the user's selection reveals that 54 out of 107 papers in *Brain Research* acknowledge *NIH-NLM* funding and that the grants associated with these papers utilized *Karst*, as a computing IT resource.

Interpretation: The visualization shows that in 2012-2017, a total of 114,894 GB and 65,495,118,259 CPU hours were used by externally funded projects that had associated publications. The grant income to IU by PIs and Co-PIs which use IU HPC resources—for researchers who had both awards and publications during the period analyzed— is $21,016,055. A majority of this funding comes from NIH projects that total $18,306,898 (87%) with the top three ICs being NIH-NIGMS, NIH-NHLBI, and NIH-NIAMS. In terms of publications, IT resource usage, via grants, links to 968 publications. *Brain Research, Medical Specialties, and Infectious Diseases* have the largest number of publications.

5.3 XDMoD VA Co-PI Network Visualization

The *Co-PI Collaboration Network* visualization is shown in Figure 7. It features a force-directed network layout on the left and a sorted horizontal bar graph on the right. Both visualizations are coupled so that hovering over an investigator in the network highlights that same investigator in the bar graph. The node size for each investigator indicates the number of grants received, while the node color indicates the total award amount; see legend in interactive visualization for details. If two investigators collaborated (i.e., their names are listed together on a grant), there exists an edge between them. The thickness of this edge represents the number of times they collaborated together. To see collaborations in the network, simply hover over the node of a particular investigator. This will highlight the selected investigator node and all emanating edges leading to other collaborator nodes in that investigator's network. Similarly, hovering over a bar in the bar graph highlights all corresponding entries and renders other bars opaque for easy viewing of the selection. The range filter on the top (1) can be used to increase or decrease the number of node labels in the network visualization. The plus and minus buttons in the top left can be used to zoom in and out. During zooming, the legend is automatically updated to ensure that node values and edge thickness remain accurate.

** Place Figure 7 here

Interpretation: The visualization helps identify three key elements of the academic and financial impact: namely, the number of awards, their total dollar amount, and research collaborations. Collaborations are rendered as a network with nodes representing researchers and edges denoting their Co-PI relationships. Given the rather short time frame, there are many, relatively small

collaboration clusters. Most links are thin, indicating a one-time collaboration; there are few instances of multiple collaborations denoted by a thicker edge between nodes. A slider (1) filters labels by the number of grants. The legend (4) provides additional insights on the number of grants, their total amount, and the number of co-authored grants. The *Total Amount* column on the right (2) shows researchers sorted by total funding during the years 2012-2017. By selecting a (Co-)PI bar (2), the collaboration network for that (Co-)PI is highlighted (3).

6 Discussion and Outlook

The work presented in this paper aims to help researchers, administrators, and funders understand and communicate the impact of campus cyberinfrastructure investments on scholarly productivity in terms of funding intake, publication output, and scholarly networks. The visualizations enable academic institutions to better understand return on investment (ROI) on advanced cyberinfrastructure for different types of research (e.g., as expressed by NIH ICs and disciplines of science). As part of the work, we demonstrated different methods for collecting and processing publicly available data from NIH and NSF official sites and from institutional production systems that advance the functionality of the XDMoD VA portal.

Expanding on the work by Knepper and Börner (2016), we primarily focused on the relationship between storage and computing resources utilized by (Co-)PIs and associated funding and publication records. Grant income to the university by (Co-)PIs who used IT resources during the period analyzed was $339,013,365 for 885 NIH and NSF projects and grant income from (Co-)PIs who used IT resources and had both grant awards and publications was $21,016,055. A total of 968 publications were associated with 83 of these NSF and NIH awards. In addition, the results show that *Brain Research*, *Medical Specialties*, and *Infectious Diseases* are the top three scientific disciplines ranked by their publication records during the given time period. Note that only awards associated with publications and IT resources are displayed; and only funding from two agencies, namely NIH and NSF, is shown.

In the future, we plan to advance the presented work as follows:

- Institutions will be able to upload not only IT compute cycles and storage usage counts but also information on storage size and number of compute cycles to provide additional insights into usage patterns across scientific disciplines.
- Funding data will be automatically retrieved via NIH RePORTER and NSF APIs, reducing the amount of manual work involved. Both online resources can be queried periodically to update award and publication data. Data from other funding agencies might be added as well.
- Fuzzy matching algorithms will be implemented to increase the number of journals mapped to scientific disciplines. This will help reduce the number of publications designated as *Unclassified*.

As Fulton et al. (2017) state, “measuring intellectual outcomes is difficult, particularly since the results of intellectual accomplishments may take years or decades to be fully realized.” To evaluate data quality and data matching (e.g., by PI name), we are working on a comparison of data retrieved from NSF/NIH versus data available via IU’s Sponsored Research production databases. Results will help understand data issues and optimize matching algorithms. Understanding the value of investment in cyberinfrastructure is challenging, as the impact of such investments has many dimensions, including intellectual contributions and financial impact. The XDMoD VA modules facilitate understanding of the role cyberinfrastructure by analyzing a number of metrics and allowing visualization of the diverse ways in which they impact institutional planning and strategy as well as the development of human knowledge.

Acknowledgements

We would like to thank Tom Furlani and his team at the University at Buffalo for their leadership in developing XDMoD; Winona Snapp-Childs, and Robert Henschel of UITS at IU for their expert input to the design and implementation of XDMoD VA; and Todd Theriault and Craig A. Stewart for edits and comments on an earlier draft of this paper. This work was partially funded by the National Science Foundation under grants 1053575 and 1566393, and also supported by the IU Pervasive Technology Institute. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

Conflict of Interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Author Contributions

Performed data preparation: OS, BF. Analysed data: OS, KB. Designed visualization: GS, SB, KB. Wrote the paper: OS, KB, GS, SB, SH, MRL.

References

Apon, Amy, 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): 87–98. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1679248.

Barrett, Richard, Michael Berry, Tony F Chan, James Demmel, June M Donato, Jack Dongarra, Victor Eijkhout, Roldan Pozo, Charles Romine, and Henk Van Der Vorst. 1994. *Templates for the Solution of Linear Systems: Building Blocks for Iterative Methods*. 2nd ed. Philadelphia, PA: SIAM. http://www.netlib.org/templates/templates.pdf.

Börner, Katy, 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." Edited by Neil R. Smalheiser. *PLoS ONE* 7 (7). Public Library of Science: e39464. doi:10.1371/journal.pone.0039464.

Dwyer, Tim, Kim Marriott, and Michael Wybrow. 2006. "Integrating Edge Routing into Force-Directed Layout." Edited by Kaufmann, Michael and Wagner, Dorothea. In *Graph Drawing*, 8–19. Berlin: Springer. doi:10.1007/978-3-540-70904-6_3.

Fulton, Ben, Steven Gallo, Robert Henschel, Tom Yearke, Katy Börner, Robert L. DeLeon, Thomas R. Furlani, Craig A. Stewart, and Matthew R. Link. 2017. "XDMoD Value Analytics." In *Proceedings of the Practice and Experience in Advanced Research Computing 2017 on Sustainability, Success and Impact - PEARC17*, 1–7. doi:10.1145/3093338.3093358.

Furlani, Thomas R., Ryan J. Gentner, Abani K. Patra, Gregor von Laszewski, Fugang Wang, Jeffrey T. Palmer, Nikolay Simakov, et al. 2013. "Using XDMoD to Facilitate XSEDE Operations, Planning and Analysis." In *Proceedings of the Conference on Extreme Science and Engineering Discovery Environment Gateway to Discovery - XSEDE '13*. New York, NY: ACM Press. doi:10.1145/2484762.2484763.

Furlani, Thomas R., Matthew D. Jones, Steven M. Gallo, Andrew E. Bruno, Charng-Da Lu, Amin Ghadersohi, Ryan J. Gentner, et al. 2013. "Performance Metrics and Auditing Framework Using Application Kernels for High-Performance Computer Systems." *Concurrency and*

Computation: Practice and Experience 25 (7): 918–31. doi:10.1002/cpe.2871.

Furlani, Thomas R., Barry I Schneider, Matthew D Jones, John Towns, David L Hart, Abani K Patra, Robert L Deleon, et al. 2012. "Data Analytics Driven Cyberinfrastructure Operations, Planning and Analysis Using XDMoD." In *Proceedings of the SC12 Conference*, *Salt Lake City, Utah*. https://laszewski.github.io/papers/vonLaszewski-draft-data-analytics-planing.pdf.

Hart, David L. 2011. "Measuring TeraGrid: Workload Characterization for a High-Performance Computing Federation." The International Journal of High Performance Computing Applications 25 (4): 451–65. doi:10.1177/1094342010394382.

Iosup, Alexandru, Catalin Dumitrescu, Dick Epema, Hui Li, and Lex Wolters. 2006. "How Are Real Grids Used? The Analysis of Four Grid Traces and Its Implications." In *2006 7th IEEE/ACM International Conference on Grid Computing*, 262–69. Washington, DC: IEEE Computer Society Press. doi:10.1109/ICGRID.2006.311024.

Katz, Daniel S., David Hart, Chris Jordan, Amit Majumdar, J.P. Navarro, Warren Smith, John Towns, Von Welch, and Nancy Wilkins-Diehr. 2011. "Cyberinfrastructure Usage Modalities on the TeraGrid." In *Proceedings of the 2011 IEEE International Symposium on Parallel and Distributed Processing Workshops and Phd Forum*, 932–39. Washington, DC: IEEE Computer Society Press. doi:10.1109/IPDPS.2011.239.

Kinney, M.F. and C.A. Raiborn. 2011. Cost Accounting. South-Western, Mason, OH. ISBN: 978-1-111-97172.

Knepper, Richard. 2011. "The shape of the TeraGrid." In *Proceedings of the 2011 TeraGrid Conference on Extreme Digital Discovery - TG '11*. New York, NY: ACM Press.

Knepper, Richard, and Katy Börner. 2016. "Comparing the Consumption of CPU Hours with Scientific Output for the Extreme Science and Engineering Discovery Environment (XSEDE)." *PloS ONE* 11 (6). doi:10.1371/journal.pone.0157628.

Lee, Bu-Sung, Ming Tang, Junwei Zhang, Ong Yew Soon, C. Zheng, P. Arzberger, and D. Abramson. 2006. "Analysis of Jobs in a Multi-Organizational Grid Test-Bed." In *Proceedings of the Sixth IEEE International Symposium on Cluster Computing and the Grid (CCGRID'06)*, 59–59. Washington, DC: IEEE Computer Society Press. doi:10.1109/CCGRID.2006.1630950.

Li, Hui, David Groep, and Lex Wolters. 2005. "Workload Characteristics of a Multi-Cluster Supercomputer." In *Lecture Notes in Computer Science*, edited by D Feitelson, L Rudolph, and U Schwiegelshohn, 176–93. Berlin: Springer. doi:10.1007/11407522_10.

Madhavan, Krishna, Niklas Elmqvist, Mihaela Vorvoreanu, Xin Chen, Yuetling Wong, Hanjun Xian, Zhihua Dong, and Aditya Johri. 2014. "DIA2: Web-Based Cyberinfrastructure for Visual Analysis of Funding Portfolios." *IEEE Transactions on Visualization and Computer Graphics* 20 (12): 1823–32. doi:10.1109/TVCG.2014.2346747.

Newby, Greg, Amy Apon, Nick Berente, Rudolph Eigenmann, Susan Fratkin, David Lifka, and Craig A. Stewart. 2014. "Return on Investment from Academic Supercomputing:" In *Presentation at SC14 (Supercomputing 2014)*. https://scholarworks.iu.edu/dspace/handle/2022/19242.

NIH. 2017. "RePORTER User Manual." Research Portfolio Online Reporting Tools. Retrieved from https://projectreporter.nih.gov/RePORTER_Manual_files/RePORTERManual.pdf

Palmer, Jeffrey T, Steven M Gallo, 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 D. Cornelius, James C. Browne, William L. Barth, and Richard T. Evans. 2015. "Open XDmoD: A Tool for the Comprehensive Management of High-Performance Computing Resources." *Computing in Science & Engineering*. http://www.buffalo.edu/content/dam/www/ccr/pdfs/OpenXDMoD_preprint.pdf.

Riehmann, Patrick, Manfred Hanfler, and Bernd Froehlich. 2017. "Interactive Sankey Diagrams." In IEEE Symposium on Information Visualization, 2005. INFOVIS 2005., 233–40. IEEE. Accessed August 24. doi:10.1109/INFVIS.2005.1532152.

Shneiderman, Ben. 1996. "The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations." In *Proceedings of the IEEE Symposium on Visual Language*, 336–43. Washington, DC: IEEE Computer Society Press. doi:10.1109/VL.1996.545307.

Stewart, Craig A., Ralph Roskies, Richard Knepper, Richard L. Moore, Justin Whitt, and Timothy M. Cockerill. 2015. "XSEDE Value Added, Cost Avoidance, and Return on Investment." In *Proceedings of the 2015 XSEDE Conference on Scientific Advancements Enabled by Enhanced Cyberinfrastructure - XSEDE '15*, 1–8. New York, NY: ACM Press. doi:10.1145/2792745.2792768.

Stewart, Craig A., Richard D. Knepper, Matthew R. Link, Marlon Pierce, Eric A. Wernert, and Nancy Wilkins-Diehr. 2017. Cyberinfrastructure, Science Gateways, Campus Bridging, and Cloud Computing. In *Encyclopedia of Information Science and Technology*, Third Edition. Vol. IX. 2014. Hershey. PA. http://www.igi-global.com. Available from: http://hdl.handle.net/2022/18608

Thota, Abhinav S., Ben Fulton, Le Mai Weakley Weakley, Robert Henschel, David Y. Hancock, Matt Allen, Jenett Tillotson, Matt Link, and Craig A. Stewart. 2016. "A PetaFLOPS Supercomputer as a Campus Resource." In *Proceedings of the 2016 ACM on SIGUCCS Annual Conference - SIGUCCS '16*, 61–68. New York, NY: ACM Press. doi:10.1145/2974927.2974956.

Towns, John, Timothy Cockerill, Maytal Dahan, Ian Foster, Kelly Gaither, Andrew Grimshaw, Victor Hazlewood, et al. 2014. "XSEDE: Accelerating Scientific Discovery." *Computing in Science and Engineering* 16 (5): 62–74. Washington, DC: IEEE Computer Society Press. doi:10.1109/MCSE.2014.80.

Figure Captions

Figure 1. Data sources, data linkage, data preparation for XDMoD VA visualizations

Figure 2. NIH RePORTER online search query

Figure 3. JSON data schema for the *Funding and Publication Impact* visualization plugin

Figure 4. JSON data schema for the *Co-PI Collaboration Network* visualization plugin

Figure 5. XDMoD VA Portal, see interactive version at http://demo.cns.iu.edu/xdmod-p/portal.html

Figure 6. XDMoD VA *Funding and Publication Impact Visualization*, see interactive version at http://demo.cns.iu.edu/xdmod-p/impact.html

Figure 7. XDMoD VA *Co-PI Network Visualization*, see interactive version at http://demo.cns.iu.edu/xdmod-p/co-pi.html