Value Analytics: A Financial Module for the Open XDMoD Project

Ben Fulton
Pervasive Technology Institute
Indiana University
Bloomington, Indiana 43017-6221
befulton@iu.edu

Steven Gallo
Center for Computational Research
State University of New York at Buffalo
Buffalo, New York
smgallo@buffalo.edu

Matt Link
Pervasive Technology Institute
Indiana University
Bloomington, Indiana
mrlink@iu.edu

Robert Henschel
Pervasive Technology Institute
Indiana University
Bloomington, Indiana
henschel@iu.edu

Tom Yearke
Center for Computational Research
State University of New York at Buffalo
Buffalo, New York
tyearke@buffalo.edu

Katy Börner
Information Visualization Laboratory
Indiana University
Bloomington, Indiana
katy@indiana.edu

Robert L. DeLeon
Center for Computational Research
State University of New York at Buffalo
Buffalo, New York
rldeleon@buffalo.edu

Thomas Furlani
Center for Computational Research
State University of New York at Buffalo
Buffalo, New York
furlani@ccr.buffalo.edu

Craig A. Stewart
Pervasive Technology Institute
Indiana University
Bloomington, Indiana
stewart@iu

ABSTRACT

Understanding the value of campus-based cyberinfrastructure (CI) to the institutions that invest in such CI is intrinsically difficult. Given today's financial pressures, administrative support for campus-based CI centers offering resources to local campus users is under constant budgetary pressure. This is partly due to the difficulty in obtaining quantitative metrics that clearly demonstrate the utility of investment in campus CI centers in enhancing scientific research and the financial aspects of enhanced competitive ability in seeking funding for research. We propose here the addition of a new realm of metrics to the standard cyberinfrastructure tool Open XDMoD (XD Metrics on Demand) that will allow us to correlate HPC usage with funding and publications. The modules to be added will allow CI centers to view metrics relevant to both scientific output in terms of in publications, and financial data in terms of awarded grants.

CCS CONCEPTS

•General and reference →Metrics; *Evaluation;* **•Human-centered computing →Visualization; •Social and professional topics →Management of computing and information systems;** *Funding;* **•Applied computing →Business intelligence;** *Decision analysis;* **•Information systems →***Data mining;*

PEARC17, New Orleans, Louisiana USA

DOI: 10.475/123_4

KEYWORDS

ACM proceedings, LaTeX, text tagging

ACM Reference format:
Ben Fulton, Steven Gallo, Matt Link, Robert Henschel, Tom Yearke, Katy Börner, Robert L. DeLeon, Thomas Furlani, and Craig A. Stewart. 2017. Value Analytics: A Financial Module for the Open XDMoD Project. In *Proceedings of Practice & Experience in Advanced Research Computing, New Orleans, Louisiana USA, July 2017 (PEARC17),* 6 pages.
DOI: 10.475/123_4

1 INTRODUCTION

Many universities debate the importance of investing in their own high-performance computing systems. These systems are a complex combination of hardware and software and require skilled personnel to maintain and administrate them. All this requires a significant investment, both in the short term as the hardware and supporting systems are purchased, and in the long term as maintenance, support, and software of the systems must be taken into account. How does a research center justify the cost to the financial officers of its institution?

Existing studies of return on investment (ROI) in campus cyberinfrastructure show that a steady and signifficant investment in high performance computing is very likely to lead to increases in publications and grant income [6]. The effect is demonstrated, for example, by Indiana University's Big Red II supercomputer, purchased in 2013 with a projected annual cost of approximately $3,000,000. Yet, according to an analysis performed at the university, facilities and administration monies coming to the university via grants awarded to users of the system over its first few years

of use, can be estimated at close to double the projected cost [13]. Thus, one can make a reasoned argument that investing in this supercomputer was a positive decision for that institution.

However, the calculations that come into play in performing this analysis are based on certain assumptions. What was the value of the supercomputer to the researcher? Could the research have been performed without the machine, and if so, at what cost? Were publications enabled or enhanced by the use of the computer? While there can never be complete answers to these questions, we propose here a suite of metrics to be added to the standard cyberinfrastructure analysis tool, Open XDMoD (XD Metrics on Demand). The metrics we propose encompass quantitative data on relevant publications, to better understand scientific value, and quantitative data on awarded grants, to better understand the return on investment in cyberinfrastructure.

Users of Open XDMoD will be able to load their financial funding data into the Open XDMoD database using a standardized input scheme. The Value Analytics realm will support fully integrated analysis of funding and usage. There will be options to enter the financial data in a customizable fashion that is most useful to the individual institution depending on what financial data they collect and what types of analyses thay want to do. The individual Open XDMoD instance with its financial data will be accessible only to the Open XDMoD user institutional users that have been granted permission. The remainder of this paper will provide details on this process and describe the progress that we have made towards constructing a fully functional value analytics module.

2 OPEN XDMOD

As described in the introduction, the financial analysis will be accomplished by the addition of a new module to the Open XDMoD HPC monitoring tool. This is the latest addition in the continuing development of the Open XDMoD technology. Originally in 2010, the State University of New York at Buffalo Center for Computational Research (CCR) began developing an open source tool to provide metrics, basic accounting and visualization of CPU and storage usage at that institution [11]. Initially titled UBMoD, the project quickly expanded into XDMoD for XSEDE, performing similar functions for the XSEDE cyberinfrastructure, a powerful collection of HPC compute, visualization, and storage resources. The expanded functionality for that project included an improved user interface, high-level charting, and additional analytical tools. The XDMoD project in turn led to an open source version of the tool, Open XDMoD [4], a version of XDMoD suitable for installation at an individual data center. Administrators can utilize Open XDMoD to gather a wide range of metrics on HPC resources (Figure 1), including resource utilization and performance, as well as monitor the jobs that are running on the system to determine their efficiency and resource consumption. This knowledge is beneficial in planning for future upgrades and acquisitions.

Open XDMoD is a web application suitable for installation on a Linux-based server at a data center. By parsing and processing the log files generated by the data center, Open XDMoD provides charts, graphs, and reports of relevant metrics over a customizable time period. Open XDMoD is able to accept data from multiple sources, including SLURM and Torque log files and campus LDAP

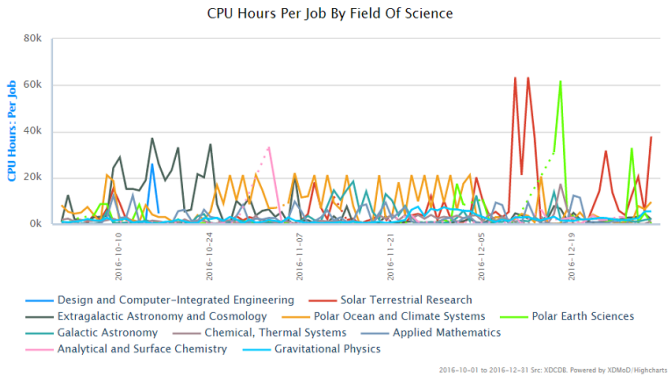

Figure 1: XDMoD allows administrators at HPC institutions to view metrics and reports in various modes and formats.

services. Through the highly customizable web interface, users can view summary charts and create dynamic charts and reports based on what they find pertinent to their particular needs.

3 IUPTI FUNDING ANALYTICS

Historically Indiana University's Pervasive Technology Institute (IUPTI) has been active in better understanding the data and usage from the university's cyberinfrastructure. Initially comprising a variety of university labs focusing on different types of advanced technology, IUPTI is currently responsible for various global initiatives such as Jetstream, the Science Gateway Group, the Advanced Visualization Lab, the Center for Applied Cybersecurity Research, and the National Center for Genome Analysis Support. The Research Technologies division has responsibility for supporting and maintaining the high-performance cyberinfrastructure provided to Indiana University researchers. As of 2016, RT supported three separate supercomputing clusters: Big Red II, Karst, and Mason. In addition, RT supports a high-speed scratch storage system and long-term tape storage for researchers at all eight of the university's statewide campuses.

The Research Technology Statistics (RT-STATS) project was begun with a similar goal as XDMoD: accounting, statistical and graphical analysis, and visualization of the usage and users of the university's supercomputers via a web browser. Given the size of the university, however, a plethora of other data sources were available for analysis. For example, when the purchase of Big Red II was proposed, IUPTI self-set a goal of having Big Red II used by adherents of at least 150 disciplines and sub-disciplines practiced at IU [12]. To determine the success or failure of this metric, users were asked to to self-select up to three disciplines at the time of account creation [13], and this information was incorporated into the RT-STATS project. From this information, administrator are able to view the relationships of discipline information to CPU usage, job times, and various other statistics. Similarly, utilization of long-term storage, environment module usage and other data sources were made available to users of the application.

Among the many sources incorporated were grants. Indiana University uses the Kuali Financial Services package [2] to track grants awarded to university researchers from the NSF, the NIH

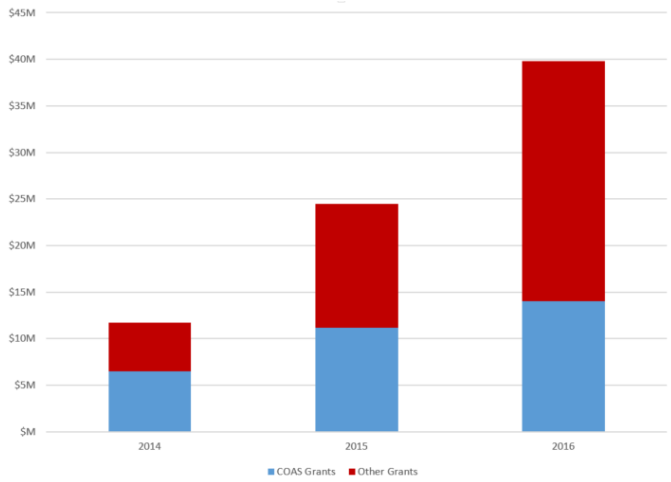

Figure 2: Example data available through Indiana University's RT-STATS project. Grant income to Principal Investigators with accounts on Indiana University's Big Red II cluster. A significant percentage of the grants were awarded to users in the College of Arts and Sciences (COAS).

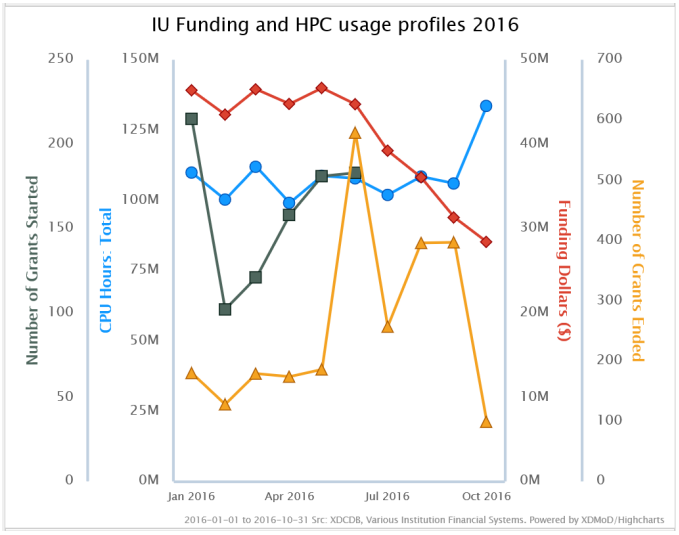

Figure 3: Screenshot of an XDMoD instance with the Value Analytics module enabled. The view shows Indiana University funding overlaid with HPC usage profiles in calendar year 2016.

and other sources. By incorporating this information into the RT-STATS system, IUPTI was able to demonstrate useful information relating grant dollars raised by university researchers to their usage of HPC resources, by developing a set of tools that links financial information – such as grant awards to IU faculty members – with usage of the supercomputers and HPC systems. Figure 2 shows an example: many users of Big Red II belong to the College of Arts and Sciences, and the figure shows grant dollars awarded to those users compared to the overall grant dollars awarded to users. Thus, the system facilitates quantification of financial income to the university in the form of grants and contracts, and cross-referencing and comparisons with usage of HPC and other research cyberinfrastructure systems.

4 XDMOD - VALUE ANALYTICS MODULE

To bring the capabilities of the RT-STATS financial analysis tools to the wider community, Indiana University and The University at Buffalo are collaborating to develop novel modules to be added to the existing Open XDMoD application. These modules are intended to provide a starting point for assessing the value of investment in campus-based CI both in scientific terms (number of publications) and in financial terms (grant income from researchers who use campus CI). The Value Analytics Module (XDMoD-VA) [5] will allow cyberinfrastructure centers and IT organizations to begin the process of quantifying the scientific and financial value of investments in HPC systems and supercomputers. By presenting a view of the institutional financial, collaboration, and publication data alongside the current HPC usage analysis in XDMoD, the module provides users with invaluable insight into the pros and cons of investing in cyberinfrastructure.

XDMoD-VA is being developed so that it can be implemented with or without direct connections to a university or college's local financial systems, providing flexibility for data acquisition. Where it is permitted by policy and practice, XDMoD-VA will enable analysis of grants and contracts received by a particular institution (Figure 3). If it is not possible for a college or university's IT organization to have direct read access from the institution's financial systems, XDMoD-VA will provide the capability to download award data directly from NSF and NIH web services. Using this grant data, users may perform an analysis of institutional HPC data cross-referenced to the data from those funding agencies. It may be relatively common that institutional policies restrict access to internal financial management systems, and this capability will enable institutions to work with data from these major funding agencies, which are in many cases the most significant sources of grant income for an academic institution.

Many institutions that provide HPC resources have their users request an allocation of computing time and/or storage space. In this case, it becomes easier to associate a particular resource usage with a researcher's lab or center. However, in 1998, a strategic plan for information technology developed at Indiana University stated,

> Advances in computing and communication have created increased demands for data storage and management. And underpinning all of this is the need to provide researchers with good software tools and good support services.
>
> In support of research, UITS should provide broad support for basic collaboration technologies and…should provide advanced data storage and management services to researchers. The University should continue its commitment to high performance computing and computation [1].

To further this goal, the decision was made to operate under a principal of abundance, meaning that resources are available to

all users who are granted accounts. Actual usage of the system is granted on a "fair share" basis. While this scheme is popular with researchers, it makes it more difficult to associate a particular project with a particular usage of resources. In this case, XDMoD-VA's ability to import group information becomes paramount: if the system can associate a user with a particular lab or center, it can reasonably infer that the resources the user consumed will be associated with that lab or center. Strategies will be put into place to address cases where a user may be a member of multiple labs or centers.

4.1 Ingestion Strategies

University financial systems are as varied and unique as the universities themselves. While for the most accurate data it is recommended that administrators use data available from institutional financial systems, it is possible to work from NSF and NIH grant data that is publicly available. In either case, the data must be transformed into a format that is usable by XDMoD. Our intent is to provide standardized scripts to directly ingest data from major financial systems such as Peoplesoft, Ellucian, and Kuali; however, even users of these systems may want to customize the data they wish to make available to Open XDMoD. These customizations potentially make the use of a standard script difficult.

To facilitate the loading of data into Open XDMoD, we have defined schemas for several JSON [3] documents that specify the structure of the data that XDMoD is designed to handle: one for grant information and one for person information for people associated with a grant. XDMoD-VA scripts will output data conforming to the schema, and any custom modifications administrators make may be validated against the schema to verify that data still can be ingested successfully. For each grant, the schema includes fields for starting and ending dates, dollar amounts, the funding agency, and grant identifiers both for the organization and the agency. Additionally for each grant, the PI, Co-PI, and any key personnel are identified based on an ID code and organization. Standardized scripts will be available to users to translate data from their financial system into JSON matching the schemas. A forthcoming schema will be made available for publication data, but our experience is that there is little commonality between institutional systems used to internally track publication data, and there would be little advantage in attempting to provide a standard script. For this reason we anticipate that ingestion from NSF/NIH data sources will be sufficient in the majority of instances.

However, disambiguation of authorship may become an issue. As no authoritative identifier for researchers exists connecting their publications, grants, and other accomplishments, integrating these data may be difficult [10]. It is not always straightforward to discern a single author with potentially multiple attribution styles on multiple papers. To partially overcome this issue, we intend to support cross-referencing author information from the ORCID project [9], and in the future may support other disambiguation strategies.

At Indiana University, all faculty, staff, students and affiliates are assigned a unique identifier over all eight campuses which is used as the login name on any HPC systems, ID's related to any grants, and other systems. This unified namespace scheme greatly simplifies the task of disambiguation, but it may not be the case at all universities. In many cases, even the internal systems of a single institution may not have the ability to identify an HPC system user, researcher, grant awardee, or professor as a single individual. In this case additional work may need to be done to properly correlate an HPC user to a grant recipient. While in the end disambiguation must be left to the institution, the XDMoD-VA person information schema supports an arbitrary number of organizations and ID's for each individual in the hopes of simplifying the task of cross-referencing grant information to user information. The person information schema also includes title and department, as well as an association with any number of arbitrary groupings, for example labs or centers.

4.2 Extensions to the XDMoD Toolset

What a particular university or institution may find relevant is dependent on the objectives and goals of that institution. Since Open XDMoD supports the creation of arbitrary reports and charts, almost any data that is available in the system may be shown as a correlation. With the base Open XDMoD package, a number of metrics are available on HPC jobs: core utilization percentage, number of jobs utilizing multiple cores, total jobs, wait time, and others. The Value Analytics package makes additional metrics available based on financial and publication records for a given time period: total grants and total grant dollars, number of publications, etc. (Figure 4).

Items that might be of interest to users of the Value Analytics module include: (1) Grant dollars and users; (2) Sources and total amount of income; (3) Researchers and/or departments that are particularly effective at securing grants; and (4) grant money that is currently active and grants that will be soon be ending. For example, a user may want to view all grants brought in by users of a cluster for the previous financial year. A new feature will be developed for Open XDMoD that allows this to be expressed as a line graph, showing information for each month, or as a pie chart, splitting the users by department. Other options are also available.

One open question for each institution will be: should all grant data be imported into the XDMoD system? Or should it only consist of users of the HPC resources? This will have to be decided by each institution dependent upon the relevant policies of the institution's financial office or offices. However, at Indiana University we have found it useful to compare the grant data of users of HPC systems to non-users - not necessarily in absolute terms, but to show what percentage of grant dollars have come from users of the systems.

4.3 Publications

Many might argue that, rather than an analysis of external funding to demonstrate ROI, a more appropriate measure might be in terms of publications that researchers are able to create. Thus, leaders of local CI facilities need to be able to document the scholarly outputs of users of local CI facilities and the collaborative relationships among faculty users of such facilities, as individuals and as academic units. With the support of the NSF, Indiana University's Center for Network Science (CNS) has implemented a pair of custom tools: the Network Workbench Tool (NWB) for the study of large scale

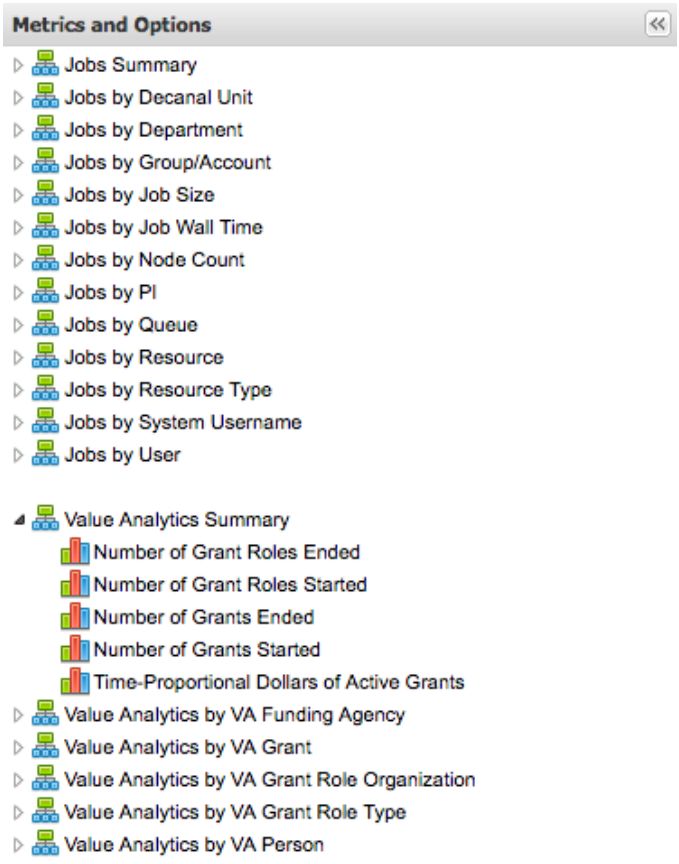

Figure 4: Metrics available to an Open XDMoD instance running with the Value Analytics module enabled.

networks and the Science of Science Tool (Sci2) optimized for science of science research and practice [8]. These tools enable novel algorithms and data sets to be combined in a plug-and-play fashion. NWB facilitates network analysis of author relationships and collaborations, and how science accomplishments relate to the funding and the evolution of scientific communities. In a strategic planning effort to better understand how local facilities enable and facilitate collaboration among units and their faculty and staff, NWB was utilized to create Figure 5, showing money flows between units that use IU's advanced CI. Similar examples will be integrated into XDMoD-VA, utilizing the tools and publication data made available by the NSF and other agencies, allowing administrators to leverage and cross-reference publication data, grant data, and the already existing HPC data made available by XDMoD.

4.4 Visualizations

The CNS is a leader in the generation of scholarly visualizations such as science maps, which are generated through a scientific analysis of large-scale scholarly datasets in an effort to extract, connect, and make sense of the bits and pieces of knowledge they contain [7]. The tools that they have made available for network analysis have proved invaluable in understanding the structure and dynamics of science, and facilitate network analysis of such topics as what

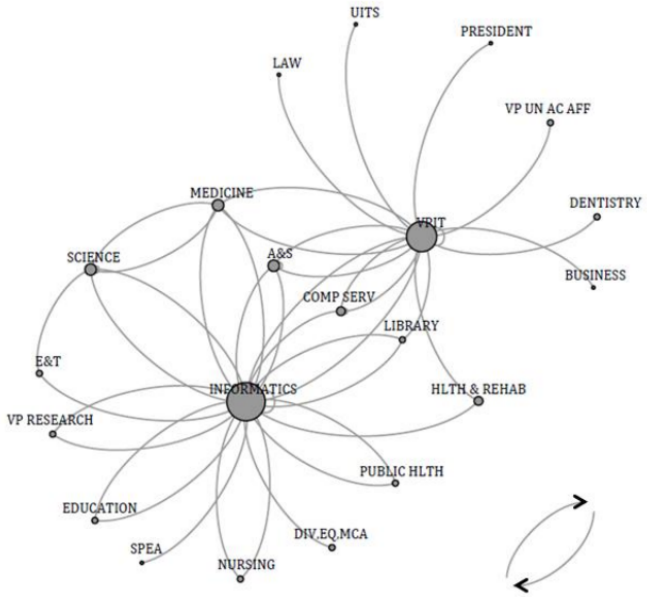

Figure 5: An Indiana University analysis performed with the Network Workbench Tool demonstrates a layout of collaborative relationships. The analysis shows money flows between units that use IUfis advanced cyberinfrastructure.

influence relationships are, who collaborates with whom, and how science accomplishments relate to the funding and the evolution of scientific communities. Similar issues arise in attempting to demonstrate the relationships between funding, publications, and cyberinfrastructure, which are complex and do not lend themselves easily to simple charts. XDMoD-VA will leverage the experience of the CNS center by providing innovative visualizations demonstrating the relationships between these multivariate dimensions. Figure 6 shows a Sankey diagram relating the usage of Indiana University cyberinfrastructure systems to funding agencies and publications.

5 CONCLUSIONS

The Open XDMoD - Value Analytics module will allow users to quantify the Return on Investment (ROI) of their investment in cyberinfrastructure resources. It will do this by allowing administrators to import financial and publication data into the same tools that are used to analyze statistical information that can be gleaned from the cyberinfrastructure itself. Once the financial information and HPC information are correlated, administrators should more easily be able to create projections, anticipate objections, and have raw data allowing them to make more informed decisions on the ROI of their cyberinfrastructure investment. The modules are anticipated to be available for general usage in October of 2017.

6 ACKNOWLEDGMENTS

This material is based, in part, upon work supported by the National Science Foundation under Grant No. 1053575 and 1566393 and also

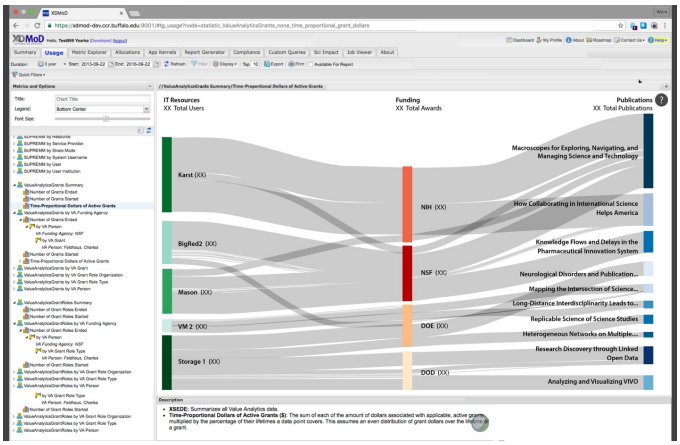

Figure 6: Sample visualization to be made available in XDMoD-VA: Display of a multivariate analysis of the relationship between IT resources, funding agencies, and publications. The width of each line represents grant dollars awarded to researchers.

by the Indiana University Pervasive Technology Institute (founded with funding from Indiana University and the Lilly Endowment, Inc.). 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, Indiana University, or the Lilly Endowment. We would like to thank our vendor partners Cray and DDN for their collaboration in the implementation of Big Red II.

REFERENCES

[1] 1998. Indiana University Information Technology Strategic Plan: Architecture for the 21st Century. http://hdl.handle.net/2022/471. (May 1998). Accessed: 2017-03-08.

[2] 2005. Kuali Financial Services. http://www.kuali.org/. (2005). Accessed: 2017-03-02.

[3] 2009. JSON Schema: A Media Type for Describing JSON Documents. http://json-schema.org/latest/json-schema-core.html. (2009). Accessed: 2017-02-24.

[4] 2010. XDMoD (XD Metrics on Demand). https://github.com/ubccr/xdmod. (2010). Accessed: 2017-03-02.

[5] 2016. XDMoD (XD Metrics on Demand) Value Analytics Module. http://indiana.edu/~xdmodva/. (2016). Accessed: 2017-03-02.

[6] A. W. Apon, L. B. Ngo, M. E. Payne, and P. W. Wilson. 2015. Assessing the efiect of high performance computing capabilities on academic research output. *Empirical Economics* 48, 1 (2015), 283–312.

[7] Katy Borner, Richard Klavans, Michael Patek, Angela M. Zoss, Joseph R. Biberstine, Robert P. Light, Vincent Lariviere, and Kevin W. Boyack. 2012. Design and Update of a Classification System. *PLoS ONE* 7 (2012).

[8] K. Borner, N. Ma, J.R. Biberstine, R.M. Wagner, R. Berhane, and H. Jiang. 2010. Introducing the science of science (Sci2) tool to the reporting branch, office of extramural research/office of the director, national institutes of health.. In *Workshop on the Science of Science Measurement.*

[9] Editorial. 2009. Credit where credit is due. *Nature* 462 (2009).

[10] Stephen B. Johnson, Michael E. Bales, Daniel Dine, Suzanne Bakken, Paul J. Albert, and Chunhua Weng. 2014. Automatic generation of investigator bibliographies for institutional research networking systems. *Journal of Biomedical Informatics* 51 (October 2014), 8–14.

[11] JT Palmer, SM Gallo, TR Furlani, MD Jones, DeLeon RL, JP White, N Simakov, AK Patra, J Sperhac, T Yearke, R Rathsam, M Innus, CD Cornelius, JC Browne, WL Barth, and RT Evans. 2015. Open XDMoD: A Tool for the Comprehensive Management of High-Performance Computing Resources. *Computing in Science & Engineering* 17 (July/August 2015). DOI:http://dx.doi.org/10.1109/MCSE.2015.68

[12] Craig A. Stewart, Beth Plale, Von Welch, Matthew R. Link, Therese Miller, Eric A. Wernert, Michael J. Boyles, Ben Fulton, David Y. Hancock, Robert Henschel, Scott A. Michael, Marlon Pierce, Robert J. Ping, Tassie Gniady, Geoffrey C. Fox, and Gary Miksik. 2015. Pervasive Technology Institute Annual Report: Research Innovations and Advanced Cyberinfrastructure Services in Support of IU Strategic Goals During FY 2015. (2015). http://hdl.handle.net/2022/20566

[13] Abhinav Thota, Ben Fulton, Le Mai Weakley, Robert Henschel, David Y. Hancock, Matthew Allen, Jenett Tillotson, Matthew Link, and Craig A. Stewart. 2016. A PetaFLOPS Supercomputer as a Campus Resource: Innovation, Impact, and Models for Locally-Owned High Performance Computing at Research Colleges and Universities. In *SIGUCCS 16,*. ACM.