Research Computing on Campus - Application of a Production Function to the Value of Academic High-Performance Computing

Preston Smith
psmith@purdue.edu
Research Computing
Purdue University
West Lafayette, Indiana, USA

Stephen Lien Harrell
sharrell@tacc.utexas.edu
Texas Advanced Computing Center
The University of Texas
Austin, Texas, USA

ABSTRACT

High-performance computing is used by research institutions worldwide for managing research data, modeling, simulation, and big data analysis. This capability is critical for higher education institutions to attract top researchers and faculty, maximizing competitiveness for future funding, and training students in current data-intensive research methods.

As a corollary, procurement and operation of advanced cyberinfrastructure incurs substantial operating and capital expenses to an organization. In this paper we present an analysis of value metrics gathered at Purdue University to measure the return on investment (ROI) of institutional investment in cyberinfrastructure resources; and an application of an economic production function to measure the cyberinfrastructure's impact on the institution's academic, financial, and reputational output.

KEYWORDS

HPC, higher education, TCO, cost analysis, ROI, production function

ACM Reference Format:

Preston Smith and Stephen Lien Harrell. 2021. Research Computing on Campus - Application of a Production Function to the Value of Academic High-Performance Computing. In *Practice and Experience in Advanced Research Computing (PEARC '21)*, July 18–22, 2021, Boston, MA, USA. ACM, New York, NY, USA, 8 pages. <https://doi.org/10.1145/3437359.3465564>

1 INTRODUCTION

At many Universities high-performance computing is critical to attract research grants, recruit and retain top faculty, and to teach students cutting-edge data-intensive research methods. The NSF's *CI2030: Future Advanced Cyberinfrastructure* report examines the need and future of cyberinfrastructure (CI) in the US and argues that: "An integrated CI that reaches from university and college campuses to the national centers is needed; this will require coordinated investments by all of the stakeholders..." and that "... the vast majority of capacity-class computing activities will be carried out

on campuses...", it continues, "communication and collaboration across research silos offers the possibility of building and deploying an integrated cyberinfrastructure that effectively and efficiently supports a broad range of scientific and engineering research" [9]. With this direction, the nation will rely on university campuses to facilitate the bulk of computational research, which makes it critical that the value to university campuses for their providing of "capacity-class computing activities" is well understood by the entire leadership of the institutions.

The scale of investment required is significant enough that it will cross the desks of deans, CFOs, and CIOs, often all the way to the president and governing boards, making it all the more important that the value is well communicated in the language of administration, rather than IT.

The supercomputer systems used to do high-performance computing at research universities are physically large. They require substantial operating costs to power and cool, and regular capital expenditures to ensure sufficient capacity and performance for science and engineering applications. As with most scientific instruments, specialized staff are required to operate and support the use of these systems.

Alternatives to campus investment do exist, however. The National Science Foundation operates XSEDE - a national infrastructure that "integrates and coordinates the sharing of advanced digital services - including supercomputers and high-end visualization and data analysis resources - with researchers nationally to support science." [5] It is important to note that service providers within XSEDE are educational institutions with prior HPC infrastructure to support their local campus. Additionally, Commercial cloud computing offerings continue to grow more powerful, less expensive, and some even offer HPC-specific technologies like high-speed, low-latency interconnects. [2]

In the current climate of shrinking education budgets, limited grant funds, and cost-conscious campus administrators, it is critical for academic research computing facilities to continually evaluate the costs and value of their campus strategy for investing in the cyberinfrastructure to support discovery.

In this paper, we show four novel contributions for use by the cyberinfrastructure community:

- A description of representative input and output metrics for cyberinfrastructure and institutional productivity, with real-world data from Purdue University as an exemplar.
- An analysis of the financial return-on-investment of campus investment in cyberinfrastructure.

This work is licensed under a Creative Commons Attribution-NoDerivs International 4.0 License.

PEARC '21, July 18–22, 2021, Boston, MA, USA
© 2021 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-8292-2/21/07.
<https://doi.org/10.1145/3437359.3465564>

- An analysis of the statistical correlations between cyberinfrastructure expenditures and institutional outputs
- An analysis with the Cobb–Douglas production function using the institution’s outputs based on its cyberinfrastructure investment

1.1 Exemplar: Purdue Campus Cyberinfrastructure

Purdue University’s campus cyberinfrastructure is based around its Community Cluster program [15]. Since 2008, the community cluster program has annually refreshed the cyberinfrastructure available to campus, deploying a new 500-600 node cluster each year, and retiring the oldest one after a useful life of 5-7 years.

Faculty members share the overall cost with the university, investing one-time grant or start-up funds to purchase access to nodes for the life of the cluster, while the university centrally funds network, storage, and staff. This allows for world-class systems to be acquired in a cost-effective manner.

2 PRIOR WORK

2.1 Value of Cyberinfrastructure

Research from Apon et al [6] shows that the presence of a Top 500 supercomputer on a campus positively impacts research output, both at the institutional level and at the departmental level. Additionally, a team from Indiana University [21] calculated the ROI of their campus supercomputer; the NSF-funded Jetstream cloud system; and XSEDE. As with [12], their methodology focused only on the acquisition costs, leaving out potential costs on data storage and transfer. Indiana’s work found that their campus cluster offered a cost savings of 2.5-3.7x over an equivalent amount of resources in Amazon EC2.

Chalker et al. [13] presented results of a survey of 31 academic centers, gathering data on the scale of the institution’s cyberinfrastructure and current practices for communicating ROI. This survey covered both on-premise and cloud CI, and reported data on outcomes correlated with categorical data about capital costs, scale, personnel costs.

At Purdue University the value of cyberinfrastructure has been an important topic for the continuation of strong funding. As such, there are numerous papers supporting this work. [19] Carlyle, Harrell and Smith [12] reported in 2010 that to use the cloud, individual investigators would spend 7x what they would on campus computing resources, and by 2019, still 2.7-5.5x more than campus resources. [17]. Smith [18] also reported very minor additional expense incurred to get substantial value in a computing grid from existing computing investment.

2.2 Identification of Input and Output Metrics

There have been multiple previous works aimed at defining output metrics for articulating the value of cyberinfrastructure. In a 2014 workshop report [11], Berente and collaborators identify four dimensions for CI value propositions: science, innovation, workforce, and economic. Berente, in a panel presentation at the 2014 Supercomputing conference [16], describes CI costs (inputs) as

including physical infrastructure, energy, capital computer equipment and software, and personnel. Products of the academy are described as Intellectual/Reputation ROI (publications, graduations, numbers of faculty and student users, and enhancement to reputation); and Financial ROI (grants and contracts, recharge fees, and patents/royalties).

Apon, [7] states that “Broadly speaking, universities combine faculty effort, physical plant, and other inputs to produce outputs such as research and education.”

Stewart et al. [20] describe a number of similar specific potential metrics to use to report return on investment in cyberinfrastructure. Our report on these metrics are noted in parentheses.

- Value of time saved
- Value of alternative investments (Within Section 2)
- Value of support from provider
- Value of training (Section 4.3)
- Grant income (Section 4.1)
- Products and patents (Section 4.5)
- Economic impact
- Research innovations (papers, citations) Section (4.2)

In this paper we utilize part of Stewart’s framework to assess the starred items above and introduce new metrics based on Recruitment and Retention in section 4.4.

2.3 Economics

In 1928, Charles Cobb and Paul Douglas published “A Theory of Production” [14], describing a production function model that predicts the output of an economy or business in terms of a function based on its capital and labor. The Cobb-Douglas production model includes as variables the output of production Y , input of capital C , and the input of labor L . Cobb and Douglas found the coefficients for their function to be .75 for L , and .25 for C , suggesting that labor accounts for 75% of the production function.

The model is represented as

$$Y = f(L, C) = 1.01 + L * .75C * .25 \quad (1)$$

Cobb and Douglas developed their model based on time series data on the volume of production from American manufacturing, which is conceptually very similar to our institutional data. Described in the terms of Cobb and Douglas - institutional products defined above such as R&D expenditures, publications, and doctorates awarded could be predicted by TeraFLOPS, cores, or the center’s annual capital costs (as capital), and staff costs (as labor).

2.4 Cost Models for Academic Cyberinfrastructure

In [15], McCartney noted that there is a continuum of models for funding research computing. Specifically, there are 3 major choices:

- Centrally funded by institution
- Full cost recovery
- Hybrid, shared cost model

At Purdue, a hybrid, shared cost model is used, sharing costs and governance between the university and the faculty.

McCartney described “This joint investment/joint ownership approach self-consciously engages the central computer group with

individual research groups whose work requires cyberinfrastructure. The essential characteristic of this approach is that it is a “coalition of the willing” By design, there is no institutional coercion — the arrangement is so attractive that it would be unwise for researchers not to participate.”

Purdue’s hybrid model is executed with annual investments in high-performance computing, beneficial to the institution which is in line with the conclusion in Apon et al. [7] that “consistent investments in HPC at even modest levels are strongly correlated to research competitiveness”

3 PURDUE COSTS

As described above, the financial arrangement of the Purdue Community Cluster Program is, in the simplest terms:

- Faculty purchase capacity of compute nodes
- The University centrally funds shared infrastructure (network, storage, facilities)
- And generally-funded central IT staff operate and support the computing resources.

In FY 2020, it cost Purdue a total of \$5.88 million dollars to staff, support ongoing operations, and to execute the annual lifecycle replacement of the campus cyberinfrastructure. A breakdown of the annual total costs of doing business are shown in Figure 1. These costs will represent capital and labor inputs to the institution’s production function.

Figure 1: Expense Analysis - FY21 Purdue Campus Cyberinfrastructure

4 OUTPUT METRIC ANALYSIS

To measure the return of investment, we first identified the key products of the university: knowledge generation (publication, grants), and graduates.

Secondary measures include data on student use of HPC for STEM education, and metrics on the community clusters’ role as a tool for faculty recruitment. Many of these are identified by Stewart [20], and Berente [16].

4.1 Sponsored Research

Universities nationwide report their research expenditures as a standard metric for research productivity. US Universities report various breakdowns of their R&D expenses to the NSF via its Higher Education Research and Development “HERD” survey [4]. (Purdue ranks #37 with \$631.6M of R&D expenditures in 2018, the most recent year with published data.)

Table 1 and Figure 2 report the annual sponsored research expenditures recorded by PIs which have invested in Purdue research computing resources (“RCAC”), out of the total sponsor expenditures at Purdue (“Purdue Total”). “Purdue Total” Sponsor Expenditures is a subset of total R&D expenditures, which reports all expenditures, including costs borne by the institution.

Table 1: Sponsor Expenditures as a Share of Purdue Total (\$M)

FY	RCAC	Non-RCAC	Purdue Total	Percent RCAC
2015	\$ 98.24	\$ 257.87	\$ 356.11	28%
2016	\$ 124.42	\$ 261.45	\$ 385.87	32%
2017	\$ 159.26	\$ 262.47	\$ 421.73	38%
2018	\$ 276.15	\$ 142.40	\$ 418.55	66%
2019	\$ 214.50	\$ 197.80	\$ 412.30	62%
2020	\$ 242.80	\$ 200.70	\$ 443.50	55%

Figure 2: Awards and Expenditures

4.1.1 Awards and Expenditures. In Figure 2 it is shown that in FY 2020, 55% of all research expenditures at Purdue were spent by faculty who participate in the community cluster program. Additionally, 57% of all F&A (overhead) expenses recovered in FY 2020 were from those same faculty members. Half of the top 100 funded faculty at Purdue are partners in the community cluster program.

All categories of expenditures and awards will be used as institutional output metrics for the production function.

4.1.2 Proposals. Data from the sponsored research office suggests significant importance of research computing facilities for grant submission. Since 2015 the number of proposals crediting research computing has increased over 4-fold, with dollar amounts doubling in 4 years.

We recognize that this likely under-reports the true scope of proposals actually requiring research computing resources. Some of the increase in reporting is due to work with the pre-award office to understand the impacts of proposals upon the cyberinfrastructure.

4.2 Publications

Analyzing publication output and the impact research facilities have on it is notoriously difficult to track, especially when relying on self-reporting.

To gather data, we generated a list of all research computing partners, and used the API provided by Scopus [1], to gather unique identifiers for those researchers' identities. R scripts using the Bibliometrix package [8] package were developed to query Scopus for a given calendar year, first tallying up publications by all Purdue affiliates, and then further reducing that set of data to subsets 1) representing the research partners identified above, and 2) publications with Purdue author affiliations from the list of 82 high-impact science journals tracked in the Nature Index. [3]

From 2010-2020, the Scopus data showed that research computing partners published 18,061 articles, of 70,107 total published by Purdue affiliates, with 17,989 of these total articles were published in high-impact Nature Index journals. With the net cost to the institution known, this allows us to measure the net cost to Purdue for each article published. 1,736 articles were published by HPC-using faculty in calendar year 2020, at a net cost to Purdue of \$3,387 per each HPC paper published. Distributing the \$5.88 million total cost of the cyberinfrastructure across the 6,712 articles published by Purdue authors in 2020 yields a net CI cost per paper of \$876 dollars.

Publications - impactful ones more so - is a key metric to be used as an output for the production function.

4.3 STEM Education

Increasingly, curricular demands that students regardless of major should be data literate, or capable of using the computational tools relevant to their domains. From data science to meteorology, bioinformatics, or CFD, undergraduate and graduate instructors increasingly require access to modern cyberinfrastructure to teach students.

To support these needs, we operate an HPC cluster dedicated to instructional use, to allow instructors to easily integrate HPC and data science technology in their curricula. Figure 3 shows the growth in the use of the *Scholar* cluster [10] in classes.

In the Fall 2020 semester, 115 courses used the campus cyberinfrastructure in their course, serving a total of 4,362 individual students, primarily undergraduates.

While data on classroom use of data is a valuable metric for multiple Purdue stakeholders, it does not fit into the category of being either an input to from the research computing center, or an institutional output.

4.3.1 CI-using Graduates. One of the fundamental products of a research university is students receiving degrees. To measure the relationship between research computing usage and Purdue graduates, we generated a list of all users who have run a research cluster job since 2012 (the earliest date with data readily available), and the institutional research team in the Office of the Provost

Figure 3: HPC Usage in the classroom

reported the number, type, and semester for degrees awarded to those list of people. This list explicitly excludes students using *Scholar*, the academic cluster. In total, HPC users have received 2,140 degrees from Purdue since 2009, with 132 baccalaureate, 142 masters, and 178 PhD degrees awarded in academic year 2019.

This data shows a dramatic increase in the number of students using HPC resources for sponsored research beginning in approximately 2015. This suggests that HPC is a scientific instrument that is key for educating students in the 21st century.

4.3.2 Earned Doctorates. Since 1957, the National Center for Science and Engineering Statistics (NCSES) conducts an annual census of all doctoral degrees issued by accredited U.S. institutions. We cross-referenced this number with the number of PhD graduates who ran HPC jobs during their PhD career.

Purdue awarded 808 doctoral degrees in 2020, with 29.1% of that 2020 class of earned doctorates running jobs on the campus cyberinfrastructure during the course of their PhD. In 2016, less than 5% of the PhD cohort of 727 graduates had used the campus cyberinfrastructure during their career.

Trained scientists is a key output of the institution and will be used as an output metric for the production function.

4.4 Recruitment and Retention

Anecdotally, research faculty tell us that the quality of research computing facilities figures into their decision when choosing to work at Purdue as a professor, vs other offers.

A metric identified during data exploration analyzed the rank of cluster partners at the time of the partner's purchase. We observed a dramatic shift, shown in Figure 4, from 2012 to 2017 - in 2012, just 5% of Carter's 45 initial partners were assistant professors (45% were full professors), and by 2020, Bell saw 52% of its 40 initial investors as assistant professors, and just 21% at the full professor rank.

This indicates that new faculty begin their career at Purdue already knowing how to use advanced cyberinfrastructure, expecting it as a service at the university, and using their Purdue-provided startup funds to invest into community clusters.

While faculty rank data is a potentially valuable metric for multiple Purdue stakeholders, it does not fit into the category of being

either an input to from the research computing center, or institutional output.

Figure 4: Rank of Faculty Investors - Bell (2020) vs Carter (2012)

4.5 Technology Commercialization

The final group of metrics identified centers around commercialization of IP developed at Purdue. Similarly to degree data, we provided a list of research computing partners to Purdue's office of technology commercialization, and received a list of those faculty members copyright disclosures, issued patents, and licensing agreements for the least 5 fiscal years. In 2020, HPC-using researchers filed 74 disclosures, were issued 32 patents, and executed 56 technology licenses.

5 SUMMARY OF METRICS, ROI, AND UNIT COSTS

Given a known annual cost for operating a campus research computing center, we attempt to define a set of metrics for the outputs of the university that are enabled by the investment in cyberinfrastructure.

High performance computing is a critical piece of research infrastructure both for a given R1 university and nationally, as the research community executes the NSF's Cyberinfrastructure plan. It cost Purdue \$5.88M in FY2020 to operate community clusters at Purdue at current levels. From those investments these research facilities have a direct and positive return on investment, according to:

- Research expenditures (at a 49x multiple of the investment to support research computing in FY 2020), and F&A expenditures recovered by Purdue (57% of all F&A recovered in 2020 was by HPC-using faculty)
- Faculty recruitment and retention (Investment by junior faculty has grown from 5% of a cluster's partners in 2012 to 52% today) signaling that HPC is becoming an important factor in recruitment.
- Training future scientists (2142 degrees have been awarded to HPC-using students since 2009, at a cost of \$11.9k per degree awarded)

- Faculty output (1,736 articles were published by HPC-using faculty in calendar year 2020, at a cost of \$3,387 per paper published)

Table 2: Direct Financial ROI Metrics, FY 2020

2020 Metric	Amount (\$M)
Cost of Doing Business	\$5.88
Awards to RCAC-using Faculty	\$289.90
Direct Expenditures Enabled	\$242.80
F&A Expenditures Enabled	\$52.70
Return on Investment	
Awards Return	49.30x
Direct Expenditure Return	41.29x
F&A Expenditure Return	8.96x

6 ANALYSIS

With costs to support an academic cyberinfrastructure facility defined as inputs, and the outputs of the institution defined and measured, we have demonstrated a clear value for the institutional investment in cyberinfrastructure facilities. Next we seek to apply a statistical analysis to the input and output metrics to understand the relationships between them, and identify which specific drivers of costs or measures of the institutional CI facility's capabilities contribute to the institution's outputs.

6.1 Methods

To study the various inputs and outputs in such a way to understand their relationships, a longitudinal data set was created, containing a wide variety of CI metrics as inputs; and Purdue's financial, scholarly, and reputational metrics as outputs.

Based on metrics identified in the literature described in Section 2.2, we created a set of key institutional data points based on 21 years (1999-2020) of data gathered from Purdue's sponsored programs office, NSF survey reports, research computing financial statements, center user demographics, center capacity, center usage reports, degrees awarded, and publication information. This data was organized into a data table and loaded into R for correlation analysis. The data correlated are described in Table 3.

The inputs and outputs exhibiting the strongest correlations were then further analyzed by fitting a linear regression model to the data set, to identify input metrics that predict institutional outputs, and finally combined into a production function. (Section 6.3)

6.2 Correlation Analysis

A visualization of the correlation matrix of the input and output data is depicted in Figure ??.

Of the 4 potential input factors, we focus on the raw capacity of the campus cyberinfrastructure, measured in aggregate TeraFLOPS available to campus, as the capital variable in the production function; and on the salary costs of the of the staff supporting the campus cyberinfrastructure, measured in dollars as the labor variable in the production function.

Figure 5: Visualization of Correlation Matrix of Purdue CI Inputs to Institutional Outputs

Table 3: Selected Purdue Input and Output Metrics

Input	Output
Annual net cost of the center	Purdue direct sponsor expenditures
Annual salary costs	Purdue F&A expenditures
Total TF operated	HERD reported R&D Expenditures
# of cluster faculty PIs	Purdue sponsored research awards
Center capital costs	Earned doctorates reported to NSF
	Purdue Publications
	US News ranking

Table 4: Correlation Table - Inputs to Selected Institutional Outputs

Output	TeraFLOPS r	Salary r
1 Total Sponsor Expenditures	0.77	0.89
2 HERD R&D Expenditures	0.75	0.93
3 Research Awards	0.77	0.78
4 F&A expenditures	0.80	0.84
5 Purdue Earned Doctorates	0.81	0.87
6 US News Ranking	0.47	0.35
7 Purdue Publications	0.78	0.90

6.2.1 Correlation Results. Both the TeraFLOPS available in the campus cyberinfrastructure and the salary cost of the research computing center exhibited strong correlations to every output metric, with the exception of *US News and World Reports* university rankings and technology commercialization.

The correlation coefficients of all outputs against these two input variables are summarized in Table 4.

6.3 Regression Analysis

6.3.1 Bivariate Models - Capacity. As the institution's computing capacity in TeraFLOPS demonstrates a strong degree of correlation for every output except US News rankings and technology commercialization, we now focus on performing a regression of each output on TeraFLOPS to establish if TeraFLOPS by itself can predict the identified institutional outputs.

Only bivariate regressions on TeraFLOPS for US News rankings, and IP commercialization fail to indicate that TeraFLOPS are a statistically significant ($p < .05$) independent variable.

Table 5 summarizes the regression analysis of these outputs on TeraFLOPS.

The regression equation for US News rankings on TeraFLOPS fails to reject the null hypothesis, suggesting that there is no relationship between TeraFLOPS and US News rankings. At this time, there is insufficient data for IP commercialization to establish any meaningful relationships.

Among academic outputs, earned doctorates increased by 6.1 doctoral degrees awarded, and publication output increased by 72.3 articles, for each 100 TeraFLOPS of capacity.

Next, we turn to focus on financial outputs, which showed the strongest Pearson's correlation with TeraFLOPS.

HERD R&D expenditures increased \$6.7 million, direct sponsor expenditures increased \$4.2 million, F&A expenditures increased \$.7 million, and new research awards increased by \$4.9 million for each 100 TeraFLOPs of capacity.

6.3.2 Bivariate Models - Staffing Investment. The correlation matrix of the research computing center's salary costs with institutional output shows multiple strong correlations. Table 6 summarizes the regression analysis of these outputs on salary costs.

Table 5: Regression - Purdue Institutional Outputs on TeraFLOPS in Campus Cyberinfrastructure

	Awards	HERD Exp	Sponsor Exp	F&A Exp	Earned Doctorates	Pubs
(Intercept)	285.64*** (19.23)	400.52*** (26.10)	264.35*** (16.22)	50.57*** (2.96)	548.50*** (19.47)	4356.89*** (261.04)
‘TF ‘	0.05*** (0.01)	0.07*** (0.01)	0.04*** (0.01)	0.01*** (0.00)	0.06*** (0.01)	0.72*** (0.13)
R ²	0.59	0.57	0.60	0.65	0.66	0.61
Adj. R ²	0.57	0.54	0.57	0.62	0.65	0.58
Num. obs.	20	21	20	17	21	21

*** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$

Among academic outputs, earned doctorates increased by 8 doctoral degrees awarded, and publication output increased by 106.4 articles, for each \$100,000 spent on research computing salaries.

For financial outputs, HERD R&D expenditures increased \$10.5 million, direct sponsor expenditures increased \$6.3 million, F&A expenditures increased \$1.1 million, and new research awards increased by \$7.2 million for each \$100,000 spent on research computing salaries.

6.3.3 Production Function Models. To build upon the simple bivariate models, we combine both predictors (TeraFLOPS capacity and salary costs) into a single model to predict institutional outputs, based on the Cobb-Douglas production function.

$$Y = f(L, K) = L_{\text{staff}} K_{\text{flops}} \quad (2)$$

Multiple regression analyses were performed to predict HERD R&D expenditures, publications and earned doctorates based on TeraFLOPS and salary costs (capital and labor costs in the production model, respectively). We enter TeraFLOPS capacity first, acknowledging Apon’s findings that the presence of a supercomputer on campus positively impacts research output. The summary of the multiple regression production models is shown in Table 7.

Using the production function, HERD R&D Expenditures increased \$8.25 million for each \$100,000 spent on research computing salaries, and \$2.35 million for each 100 TeraFLOPS of capacity.

The production function model indicates that the number earned doctorates awarded increased by 5.27 PhDs for each \$100,000 spent on research computing salaries, and 3.34 PhDs for each 100 TeraFLOPS of capacity.

Finally, the production function shows publication output increases by 76.3 articles for each \$100,000 spent on research computing salaries, and 32.2 articles for each 100 TeraFLOPS of capacity.

7 CONCLUSIONS

Here we have demonstrated a quantifiable ROI-based approach to show the costs around a campus research computing center and a set of outputs that the center enables, enabling a clear presentation of the center’s financial ROI and the per-unit costs of several outputs. Further statistical analysis shows strong, statistically significant correlations between multiple inputs and institutional outputs, and describes statistically significant regression models to predict the impact of Purdue University’s CI capabilities based on the scale of systems and staff available to researchers.

Regression models show that each 100 TeraFLOPS of capacity predicts:

- \$4.9 million of research awards
- \$4.2 million of direct and \$0.7 million of F&A expenditures
- \$6.7 million of HERD R&D expenditures
- 6.1 earned PhDs, and 72.3 publications

In a simple bivariate model, each \$100,000 of salary investment predicts:

- \$7.2 million of research awards
- \$6.3 million of direct and \$1.1 million of F&A expenditures
- \$10.2 million of HERD R&D expenditures
- 8 earned PhDs, and 106.4 publications

Utilizing the Cobb-Douglas production function as a conceptual framework to build a model, we have created a multiple regression model with TeraFLOPS of capacity and salary investment as predictors, representing capital and labor in the production function.

In these production function models, each \$100,000 spent on research computing salaries predicts

- \$8.25 million of HERD R&D expenditures
- 5.27 earned PhDs, and
- 76.3 publications

and each 100 TeraFLOPS of capacity predicts

- \$2.35 million of HERD R&D expenditures
- 3.34 earned PhDs, and
- 32.2 publications.

8 FUTURE WORK

Future work will explore production models in greater detail in order to build high quality models that reliably predict the institution’s academic, and reputational outcomes. Further analysis across institutions will test if the general conclusions identified here for Purdue that TeraFLOPS predict research expenditures and earned doctorates holds true for other institutions. Finally, a detailed study at the departmental and individual PI level will be performed, analyzing the impact of CI use on the individual faculty member or academic department’s productivity.

ACKNOWLEDGMENTS

The authors would like to thank Purdue CIO Karl Browning and Professors Gerhard Klimeck, Thomas Hacker, Norbert Neumeister, and Matt Huber from our community cluster governance board for their guidance and input in defining our value proposition metrics.

Table 6: Regression - Purdue Institutional Outputs on Salary Costs for Campus Cyberinfrastructure

	Awards	HERD Exp	Sponsor Exp	F&A Exp	Earned Doctorates	Pubs
(Intercept)	151.52*** (30.49)	210.37*** (26.18)	149.61*** (21.67)	27.23*** (6.01)	417.09*** (28.50)	2469.40*** (324.09)
‘Salary’	71.85*** (10.98)	104.61*** (9.42)	63.07*** (7.46)	11.49*** (1.92)	79.91*** (10.26)	1064.06*** (116.66)
R ²	0.68	0.86	0.80	0.70	0.75	0.81
Adj. R ²	0.67	0.85	0.79	0.68	0.74	0.80
Num. obs.	22	22	20	17	22	22

*** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$ **Table 7: Regression - Production Model Outputs on TeraFLOPS, Salary Costs**

	HERD Exp	Doctorates	Pubs
(Intercept)	241.40*** (26.24)	446.75*** (25.86)	2884.00*** (310.89)
‘TF’	0.02* (0.01)	0.03** (0.01)	0.32** (0.11)
‘Salary cost’	82.46*** (11.59)	52.73*** (11.42)	763.32*** (137.29)
R ²	0.89	0.85	0.85
Adj. R ²	0.87	0.83	0.84
Num. obs.	21	21	21

*** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$

REFERENCES

- [1] [n. d.]. Data | Curated. Connected. Complete. <https://www.elsevier.com/solutions/scopus>
- [2] [n. d.]. Introducing the new HB and HC Azure VM sizes for HPC | Blog | Microsoft Azure. <https://azure.microsoft.com/en-us/blog/introducing-the-new-hb-and-hc-azure-vm-sizes-for-hpc/>
- [3] [n. d.]. Nature Index. <https://www.natureindex.com/>
- [4] 2020. Higher Education Research and Development Survey (HERD) | NCSES | NSF. <https://www.nsf.gov/statistics/srvyherd> [Online; accessed 17. Feb. 2020].
- [5] 2020. What we do - XSEDE. <https://www.xsede.org/about/what-we-do> [Online; accessed 17. Feb. 2020].
- [6] 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 (2010), 87–98.
- [7] 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 (feb 2015), 283–312. <https://doi.org/10.1007/s00181-014-0833-7>
- [8] Massimo Aria and Corrado Cuccurullo. 2017. bibliometrix: An R-tool for comprehensive science mapping analysis. *Journal of Informetrics* 11, 4 (2017), 959–975. <https://doi.org/10.1016/j.joi.2017.08.007>
- [9] David A Bader, Georgia Tech, Helen M Berman, Manish Parashar, Thom H Dunning, Gwen Jacobs, Tilak Agerwala, Michael Hildreth, Rich Loft, and Steven Ruggles. 2018. *CI2030: Future Advanced Cyberinfrastructure A report of the NSF Advisory Committee for Cyberinfrastructure Transmittal letter from the NSF Office of Advanced Cyberinfrastructure Report CI2030: Future Advanced Cyberinfrastructure NSF Advisory Committee for Cyberinfrastructure 1 Cyberinfrastructure Ecosystem Working Group*. Technical Report. <https://www.nsf.gov/cise/oac/ci2030/>.
- [10] M. E. Baldwin, X. Zhu, P. M. Smith, S. L. Harrell, R. Skeel, and A. Maji. 2016. Scholar: A Campus HPC Resource to Enable Computational Literacy. In *2016 Workshop on Education for High-Performance Computing (EduHPC)*, 25–31. <https://doi.org/10.1109/EduHPC.2016.009>
- [11] Nicholas Berente, James Howison, John L. King, Joel Cutcher-Gershenfeld, and Robert Pennington. 2014. Leading Cyberinfrastructure Enterprise: Value Propositions, Stakeholders, and Measurement. *SSRN Electronic Journal* (apr 2014). <https://doi.org/10.2139/ssrn.2416247>
- [12] Adam G Carlyle, Stephen L Harrell, and Preston M Smith. 2010. Cost-Effective HPC: The Community or the Cloud. *IEEE Conference on Cloud Computing Technology and Science* (dec 2010).
- [13] Alan Chalker, Curtis W. Hillegas, Alan Sill, Sharon Broude Geva, and Craig A. Stewart. 2020. Cloud and On-Premises Data Center Usage, Expenditures, and Approaches to Return on Investment: A Survey of Academic Research Computing Organizations. In *Practice and Experience in Advanced Research Computing (PEARC '20)*. Association for Computing Machinery, New York, NY, USA, 26–33. <https://doi.org/10.1145/3311790.3396642>
- [14] Charles W. Cobb and Paul H. Douglas. 1928. A Theory of Production. *The American Economic Review* 18, 1 (1928), 139–165. <http://www.jstor.org/stable/1811556>
- [15] Gerry McCartney, Thomas Hacker, and Baijin Yang. 2014. Empowering Faculty: A Campus Cyberinfrastructure Strategy for Research Communities. *Educause Review* (2014).
- [16] Greg Newby, Amy Apon, Nick Berente, Rudolph Eigenmann, Susan Fratkin, David Lifka, and Craig Stewart. 2014. *Return on Investment from Academic Supercomputing: SC14 Panel*. Technical Report.
- [17] Preston Smith, Stephen Lien Harrell, Alex Younts, and Xiao Zhu. 2019. Community Clusters or the Cloud: Continuing Cost Assessment of On-premises and Cloud HPC in Higher Education. In *Proceedings of the Practice and Experience in Advanced Research Computing on Rise of the Machines (Learning) (PEARC '19)*. ACM, New York, NY, USA, 97:1–97:4. <https://doi.org/10.1145/3332186.3333155>
- [18] Preston M Smith. 2011. A cost-benefit analysis of a campus computing grid. (2011).
- [19] Preston M Smith, Thomas J Hacker, and Carol X Song. 2008. Implementing an industrial-strength academic cyberinfrastructure at Purdue University. In *IEEE International Parallel & Distributed Processing Symposium*. 1–7.
- [20] Craig A. Stewart, Amy Apon, Ron Payne, Shawn D. Slavin, David Y. Hancock, Julie Wernert, Thomas Furlani, David Lifka, Alan Sill, Nicholas Berente, Donald F. McMullen, and Thomas Cheatham. 2019. Assessment of financial returns on investments in cyberinfrastructure facilities. In *Proceedings of the Practice and Experience in Advanced Research Computing on Rise of the Machines (learning) - PEARC '19*. ACM Press, New York, New York, USA, 1–8. <https://doi.org/10.1145/3332186.3332228>
- [21] Craig A Stewart, David Y Hancock, Julie Wernert, Matthew R Link, Nancy Wilkins-Diehr, Therese Miller, Kelly Gaither, and Winona Snapp-Childs. [n. d.]. 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). ([n. d.]). <https://doi.org/10.1109/UCC.2018.00031>