Innovation Under Resource Constraints: Supercomputing in Scientific Research

Justine Boudou* and John McKeon+

June 17, 2024

Abstract

Resource constraints significantly shape innovation by influencing both the rate and direction of output. In this paper, we study how constraints of a key input to innovation, computing power, impact innovative output. Our empirical setting is an NSF-funded initiative called XSEDE, which allows researchers to apply for access to a distributed network of high-performance computing resources, or "supercomputers," across the United States. Leveraging the unique resource allocation process of XSEDE, we are able to isolate the impact of computing resources on scientific production. We find that an increase in the share of computing resources allocated to a researcher leads to an increase in the number of papers published and alters the research direction. Specifically, additional computing resources lead researchers to study less popular and newer topics as well as to broaden the scope of their projects. Our findings underscore the intricate relationship between resources and innovative output, with consequences for optimizing resource allocation strategies and fostering innovation.

Keywords: *Computing Resources*, *Constraints*, *Science*, *Innovation*, *Allocation*

Authors contributed equally, listed alphabetically. *Boudou: Harvard Business School (jboudou@hbs.edu) +McKeon: Boston University (jmckeon@bu.edu)

1 Introduction

Resource constraints exert a significant influence on the trajectory of innovation, shaping both the rate and direction of output by impacting innovators' behaviors and incentives (Arrow 1962, Nelson 1959, Barney 1991, Agrawal et al. 2018). Constraints can influence firms' risk-taking (Krieger et al. 2022) and R&D investments (Mezzanotti & Simcoe 2023), along with investors' and firms' tolerance for failure (Nanda & Rhodes-Kropf 2017, Atanassov 2016). While much of this work focuses on firm-level output, we still know relatively little about how resource constraints impact the choices that innovators make. This is because variation in resource constraints that may cause changes in behaviors are usually difficult to observe or coincide with access to other resources. This paper aims to fill this gap by utilizing a novel empirical setting that allows us to study how constraints of a key input to innovation, computing power, impacts innovative output in the context of scientific researchers.

Computing power is an important input to modern innovation in its own right, particularly in the production of frontier knowledge and innovation (Thompson et al. 2022, Kim 2023, Tranchero 2023). Across many fields, access to high-performance computing resources has become indispensable for firms' competitive advantage and researchers striving to push the boundaries of knowledge. In the field of artificial intelligence for instance, firms increasingly rely on massive supercomputing clusters to support the training of their prediction models. For example, Microsoft has recently discussed investing $100 billion to build a new supercomputer, "Stargate."[1] In science, advanced computational resources facilitate the screening of potential drug candidates or the simulation of complex weather patterns, accelerating the development of new therapies and aiding in disaster preparedness. Despite its widespread importance, computing power remains a scarce resource for many innovators (Lohr 2019). These access inequalities further underline the importance of understanding innovators' choices when their computing power is constrained.

Our empirical setting allows us to isolate variation in supercomputing resources and to shed light on how they shape innovative behaviour. Our study has important implications for policymakers, R&D managers, constrained entrepreneurial firms, and scientific researchers. We document changes to the direction of science following variation in resource constraints, which has important downstream

[1] https://fortune.com/2024/04/02/microsoft-openai-stargate-100-billion-ai-supercomputer-star-wars-sdi/

implications for commercially-oriented innovation given recent patterns documented by Arora et al. (2018) and Marx & Fuegi (2020).

In this paper, we start by discussing researchers' strategies following the allocation of additional computing resources. We predict that as resource constraints are relaxed, researchers will produce additional output. Importantly, prior literature also suggests that the type of innovation produced may differ as constraints are relaxed. In our context, we consider how variations in computing resources affect researchers' choice to pursue less popular topics, research lines they haven't worked on before, and projects broader in scope.

We then empirically study our research question in the context of an NSF-funded initiative known as *The Extreme Science and Engineering Discovery Environment* (henceforth "XSEDE"), which allows researchers to apply for access to a distributed network of high-performance computing resources, or "supercomputers," across the United States. Each request is assessed by a committee whose role is to make a recommendation of the amount of computing resources the project would need if the total amount of XSEDE resources were unconstrained. The aggregate computing demand is then reconciled with available supply by reducing each recommendation based on an algebraic formula. Our empirical framework leverages this unique resource allocation process of XSEDE in order to create variation in the quantity of computing resources received by researchers.

We find that increased resource allocation leads to an increase in the number of papers published. Additionally, we find evidence consistent with the fact that as resource constraints are relaxed, researchers tend to explore newer and less studied research avenues. Further, researchers receiving more computing resources tend to broaden the scope of their research, pursuing projects which span a wider array of topics. These findings are consistent with additional resources allowing researchers to study projects with greater uncertainty that they did not explore when more constrained. They do this by studying newer topics that have not yet gained traction in the scientific community and recombining this with their existing work, creating broader papers.

Taken together, our study sheds light on the intricate relationship between resource allocation and innovative output, underscoring the importance of effective resource allocation strategies in promoting innovation among managers and policymakers. We emphasize the importance of understanding the changes in the attributes of output produced with additional resources. Our findings underscore the importance of considering not only the rate but also the direction of research output when evaluating resource allocation strategies at R&D intensive firms or publicly funded programs such as XSEDE, which must allocate scarce resources across projects. Our paper focuses empirically on outcomes most closely tied to the researchers decisions given computing constraints, rather than outcomes such as citations which blend these decisions with later reception in the scientific community.

This paper makes three distinct contributions. First, we contribute to the literature studying the impact of resource constraints on innovation, estimating the effect of computing power on research output. Much of this literature focuses on the effect of funding and financial constraints on firms' innovation output (Hoegl et al. 2008), showing for example that financing impacts firms' quantity and quality of innovations (Atanassov 2016), the type of firms that investors are willing to fund (Nanda & Rhodes-Kropf 2017) as well as the composition of their R&D investments (Mezzanotti & Simcoe 2023). Additional work studies the impact of other types of non-fungible constraints such as geography (Singh & Marx 2013, Catalini 2018), labor policy (Glennon 2023) or environmental regulation (Berrone et al. 2013) on innovation, as well as the role of other scarce inputs such as attention (Fini et al. 2022). We contribute to existing research by looking at the innovators themselves and the choices that they make as resource constraints vary. Our empirical context allows us to study a crucial input to innovation that is high-performance computing power.

Second, we contribute to the related literature on the marginal returns to resource allocation. Prior work studies the returns of R&D subsidies directly on firms' patenting and revenue (Bronzini & Piselli 2016, Howell 2017) or through the spillovers they generate (Azoulay et al. 2019, Myers & Lanahan 2022). In the same vein, our empirical setting allows us to explore the *intensive margin* of computing resource allocation. This is usually hard to estimate given that variation for access to non-fungible resources usually comes from shocks that vary *whether* firms get access or not to inputs or technology (Kong et al. 2022, Jacob & Lefgren 2011) but not necessarily *how much* they get access to. In our context, it is crucial to understand this margin as the relevant constraint to pursuing frontier research is not necessarily whether a researcher has access to *any* computing power, but rather*how much* computing power she has access to.

Third, we contribute to the literature on firms' and researchers' innovation strategy which explores risk-taking (Mandler 2017, Krieger et al. 2022, Franzoni & Stephan 2023) and other changes in innovation direction such as the departure from one's existing line of research (Myers 2020) or the expansion of existing projects (Myers & Tham 2023). Building upon this literature, our study investigates how high-performance computing resource allocation influences researchers' innovation strategies. By analyzing the impact of increased resource allocation on research output, research direction, and risk-taking behavior, we offer insights into the mechanisms driving innovation within the context of advanced computing resources. Our findings provide valuable contributions to understanding the complex interplay between resource allocation decisions, researchers' strategic choices, and the outcomes of scientific research.

Section 2 lays out a simple conceptual framework to ground our empirics. Section 3 describes the XSEDE setting and Section 4 discusses our empirical approach. Section 5 describes our data and Section 6 outlines our baseline findings and our exploration of potential mechanisms. Section 7 explores additional analyses and heterogeneity. Section 8 concludes.

2 Researcher Strategies Under Resource Constraints

2.1 Do researchers produce more output?

We first consider how an increase in the amount of computing resources allocated to researchers impacts the quantity of output they produce. In practice, it is not obvious that more resources necessarily leads to an increase in output – for instance, researchers could be "wasteful" with the additional resources, or may supplement with outside options. However, based on previous work and the specific nature of computing resources, we might expect to find a positive relationship.

First, existing literature tends to find a positive correlation between the availability of resources and scientific or innovation-related outputs, typically measured by publications or patents. For instance, access to technology such as Information Technology (IT) is associated with an increase in the number of publications (Ding et al. 2010). Similarly, Mezzanotti & Simcoe (2023) indicates that investments in basic research declines when funding cost increases. In the scientific context, work suggests that government funding leads to increased research output (Payne et al. 1999, Arora & Gambardella 2005).

Importantly, research has shown that the magnitude of the impact on funding on research output largely depends on the availability of outside options. For instance, Jacob & Lefgren (2011) find that being funded by the National Institute of Health (NIH) has a positive but small effect on the number of publications and citations and show that this is consistent with researchers being able to shift to another source of funding. In contrast, high-performance computing resources are more scarce and typically lack direct substitutes.[2]

As a result, we expect that an increase in the share of computing resources will have a positive impact on researchers' output.

2.2 Do researchers produce a different type of output?

While we establish that the quantity of output is likely to increase with additional computing resources, the type of innovation produced when resource constraints are relaxed might also differ. Prior literature suggests that spatial, institutional, and fungible constraints have meaningful consequences on innovation characteristics (Catalini 2018, Boyabatlı et al. 2016, Cerqueiro et al. 2017) through mechanisms such as risk-aversion and competition (Krieger et al. 2022, Mezzanotti & Simcoe 2023). Since computing resources represent a critical input for researchers and are subject to availability constraints, their increase could similarly impact the nature of the science pursued. First, additional computing resources may make researchers more tolerant of the risk associated with undertaking more uncertain or less "conventionally incentivized" projects (Azoulay et al. 2011). In science, where career incentives may lead researchers to prioritize more predictable projects, additional computing resources may increase scientists' willingness to take on projects with higher uncertainty as they are able to pursue both the high- and low-uncertainty projects within their research agenda. The specific ways in which computing resources shape the type of science pursued by researchers remain under-explored and warrants further investigation. We examine three potential avenues that researchers may reasonably pursue with more resources: (i) exploring less-studied research avenues (ii) venturing beyond their expertise (iii) broadening their research scope. While these strategies are not exhaustive, they offer a view into crucial ways in which computing resources might influence the type of science researchers

[2]For example, running large simulations often cannot be split up onto multiple resources but requires allocations within one environment.

engage in.

First, researchers may decide to use additional computing resources to pursue less-studied research avenues, either by studying topics that are inherently less popular or by focusing on topics that are younger and have appeared more recently in the scientific literature. The prevailing norms and incentives within the scientific community tend to prioritize the pursuit of topics that yield tangible returns, such as papers and citations (Dasgupta & David 1994). While studying less-studied areas might allow researchers to establish priority on a specific topic, it may also come at the cost of having a smaller audience and making the peer-review process more difficult (Krieger et al. 2023), ultimately increasing the perceived risk of pursuing these topics. As suggested by Azoulay et al. (2011), incentives such as the tolerance for early failure can lead researchers to change their research direction and to choose less-traveled scientific avenues. By reducing researchers' constraints and hence the associated risk of undertaking different types of projects, getting additional computing resources may similarly allow researchers to study less-traveled research avenues.

Second, the influx of additional resources can lead scientists to embark on research endeavors that are 'new-to-them' and to venture beyond their existing areas of expertise (Hill et al. 2021). Along these lines, a large body of innovation literature studies the exploration-exploitation trade-off, its underlying mechanisms, as well as its boundary conditions (Luger et al. 2018, Marino et al. 2015, Danneels & Sethi 2011, Sidhu et al. 2007, Choi et al. 2016, Fitzgerald et al. 2021, Zhuo 2022). While various incentives in the modern scientific community tend to encourage specialization (Jones 2009, Foster et al. 2015), researchers may be more willing to pursue projects outside their specialty - which are inherently more uncertain - as constraints are relaxed.[3] Hence, an increase in computing resources might increase scientists' exploration of new topics that are not part of their existing knowledge base.

Third, researchers may broaden their scope by pursuing more interdisciplinary or cross-cutting topics that traverse traditional disciplinary boundaries. Breadth, or scope, and its relationship with firm performance or innovation has been studied in various streams of literature, such as organizational learning (Dutt & Lawrence 2022, Ferguson & Carnabuci 2017). While prior research often uses knowledge breadth as a source of variation and studies its impact on innovation (Nagle & Teodoridis 2020,

[3]This mechanism is in line with Krieger et al. (2022) where, when firms receive cash windfalls, they invest in riskier drug candidates outside of their existing portfolio.

Zhou & Li 2012), the scope of research may also be endogenous to the constraints researchers face. However, our understanding of how resources constraints impact the breadth of researchers' projects remains limited. Because the scientific norms tend to reward researchers' expertise on a specific piece of knowledge (Jones 2009), researchers might not have strong incentives to pursue broad projects at first. In addition, broader projects might structurally require more resources to get completed. By shifting researchers' incentives towards undertaking more uncertain projects and/or allowing them to allocate more resources to a project, we posit that getting additional computing resources might increase the research space of researchers' projects.

Overall, we expect that an increase in the allocation of computing resources leads researchers to pursue a different type of science, the nature of which we explore empirically.

3 Setting

3.1 XSEDE

XSEDE was an initiative funded by the National Science Foundation (NSF) between 2009 and 2022 for a total of $257M.[4] XSEDE was one of the most advanced, powerful, and robust collection of integrated digital resources and services in the world for open scientific research and was designed to provide researchers across various disciplines with access to advanced computing resources, expertise, and support. XSEDE established a distributed network of high-performance computing (HPC) resources, often referred to as "supercomputers", strategically located across the United States. These supercomputers enabled researchers to tackle computationally intensive scientific challenges that were previously beyond reach.

Table 1 shows the 10 most used computing resources during our study period as well as their associated computing power in the common metric used by XSEDE called Service Unit (SU).[5] Figure 1 shows the location of these resources (in red) as well as the location of other resources used as part

[4]Following its success, XSEDE was replaced in 2022 by ACCESS (Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support)

[5]The XSEDE resources have various computational performance which requires the use of a common metrics to compare their usage. For the purpose of this paper, SU can be thought of as "core-hour". Other metrics which offer more nuance, such as Normalized Units (NU), were also used by XSEDE.

of the program.

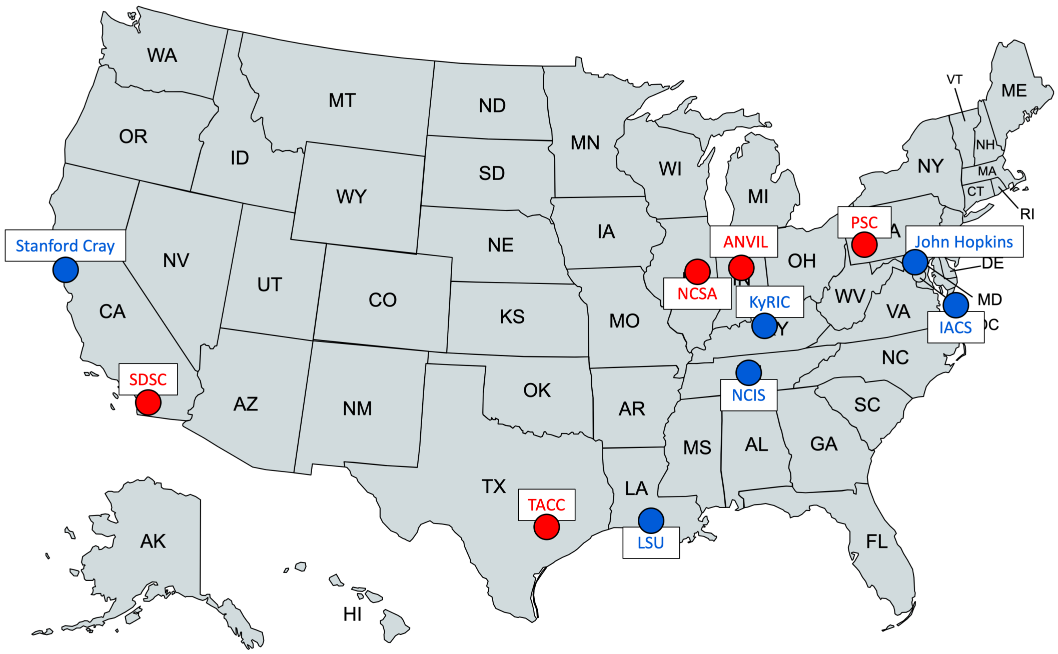
Table 1: Ten most used resources

Resource Name	SU (Millions)
TACC Dell/Intel Knights Landing (Stampede2)	27,454
SDSC Dell Cluster (Comet)	10,327
SDSC Expanse CPU	9,306
PSC Bridges-2 (Bridges-2)	8,493
TACC Dell PowerEdge (Stampede)	6,320
PSC Regular Memory (Bridges)	6,079
Purdue Anvil CPU	3,674
NCSA Delta GPU (Delta GPU)	2,340
IU/TACC (Jetstream)	2,264
PSC Bridges-2 GPU (Bridges-2 GPU)	1,965

Notes: This table shows the 10 most used resources over the period 2015-2022. Computing resources are all converted in Service Units (SU) which allows comparison across resources.

XSEDE played a pivotal role in facilitating groundbreaking discoveries across various scientific domains. One notable example of XSEDE's impact is its support for the Laser Interferometer Gravitational-Wave Observatory (LIGO) project. LIGO, a collaboration between Caltech and MIT, made history by detecting gravitational waves for the first time in 2015, confirming a key prediction of Einstein's theory of general relativity. XSEDE provided essential computational resources and expertise to process and analyze the vast amounts of data generated by LIGO's detectors, contributing to this discovery. Additionally, XSEDE supported research in astrophysics, enabled scientists to conduct simulations of cosmic phenomena, such as the formation of galaxies and the behavior of black holes, leading to deeper insights into the structure and evolution of the universe. In climate science, XSEDE-supported research improved climate models and weather forecasting capabilities, aiding in disaster preparedness and climate change mitigation efforts. Furthermore, XSEDE played a crucial role in drug discovery and molecular simulations, providing researchers with the computational power needed to screen potential drug candidates, simulate molecular interactions, and accelerate the development of new therapeutics for various diseases.

Figure 1: Location of XSEDE resources



3.2 Resource Allocation Process

Research allocations were available to researchers at a U.S.-based institution, where the vast majority of researchers are associated with universities or other research institutions.[6] Proposals were accepted every year on a quarterly schedule. PIs' proposals should include a scientific background and research objectives, as well as a resource usage plan containing code performance timings, resource usage details, scaling information and for highly competitive HPC resources, benchmark runs on the resource requested.[7]

Figure 2 shows an example of a proposal in the field of Astronomic Sciences, where the PI requested 476,372 SUs on the TACC machine Stampede 2. Figure 3 shows the summary table of the detailed calculations made by the PI to justify his computing resources' request. The proposal also includes a description of the reason why XSEDE is critical for the PI's research agenda and emphasizes the limited availability of outside options. In particular, it highlights that the amount of computing resources that

[6] In our sample, 94.1% of researchers are associated with a university, 5.6% with a non-university research institution, and 0.3% with a commercial entity.

[7] Such benchmarking is typically done through a small "startup" allocation that serves this purpose.

he needs is too high to be accommodated by his university.

Figure 2: Abstract and Scientific Background

Numerical Simulations of Protoplanetary Disks

Jacob B. Simon (PI), Philip J. Armitage

Summary

We request a total of **476,372 SUs** on the TACC machine STAMPEDE 2 SKX nodes to run and analyze numerical simulations of magnetohydrodynamic turbulence in the planet forming region of protoplanetary disks. We will use the well-tested ATHENA code to simulate a series of local, co-rotating radial regions of a model disk. To store the large datasets that will result from these calculations, we request $\mathbf{1.744 \times 10^4}$ **GB** of archival storage on the TACC RANCH system.

Figure 3: Requested Computing Resources

Table 1: Proposed Simulations

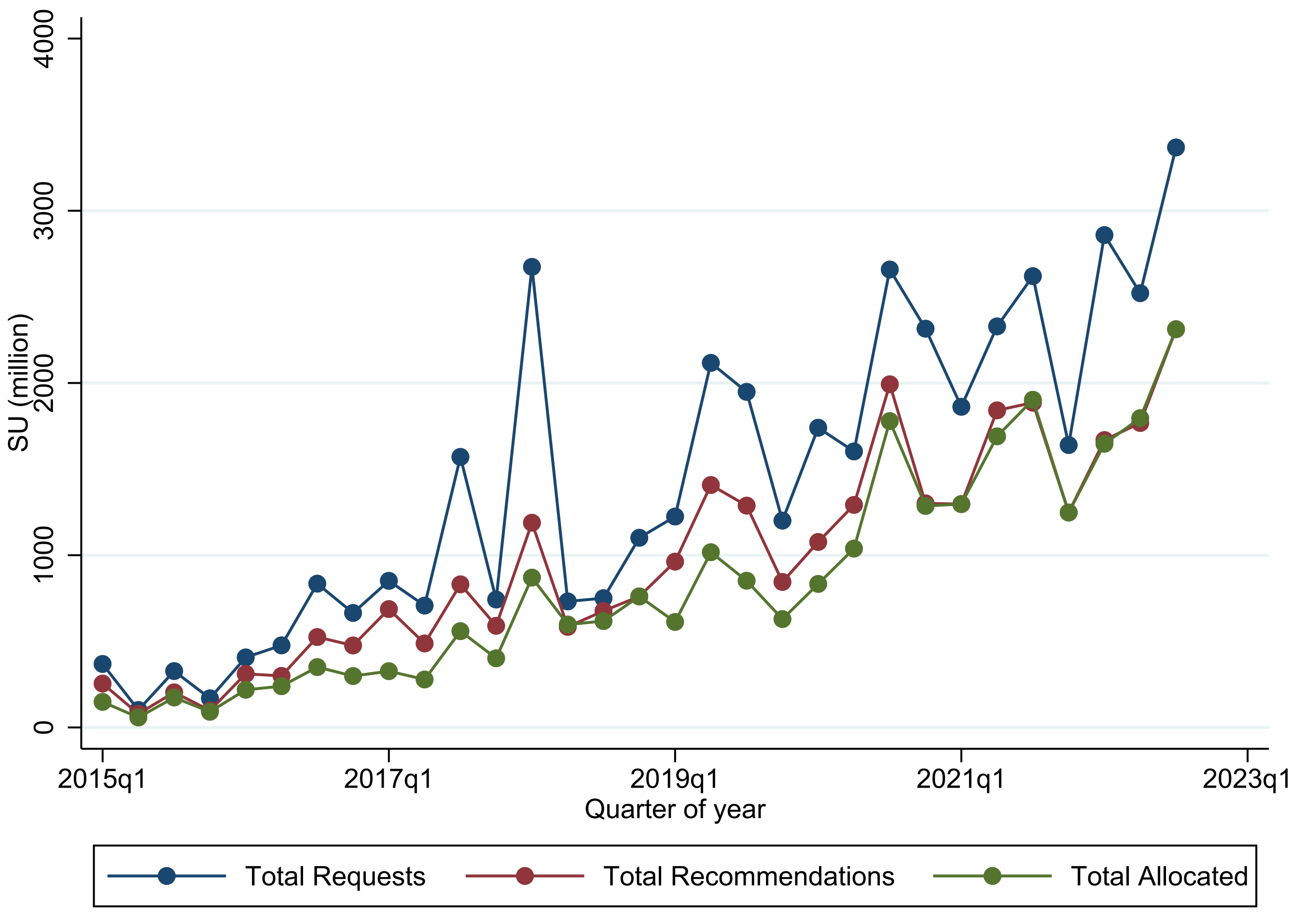
R (AU)	β_z	O_B	# zones	# cores	$\frac{\text{zones}}{\text{core}}$	N_t	Days	SUs
5	10^3	1	256×512^2	4,096	16×32^2	1.59×10^7	33.5	**69,147**
5	10^3	-1	256×512^2	4,096	16×32^2	1.59×10^7	33.5	**69,147**
5	10^4	1	256×512^2	4,096	16×32^2	1.59×10^7	33.5	**69,147**
5	10^4	-1	256×512^2	4,096	16×32^2	1.59×10^7	33.5	**69,147**
5	10^5	1	256×512^2	4,096	16×32^2	1.59×10^7	33.5	**69,147**
5	10^5	-1	256×512^2	4,096	16×32^2	1.59×10^7	33.5	**69,147**
1	10^4	1	256×512^2	4,096	16×32^2	1.59×10^7	33.5	**69,147**
10	10^4	1	256×512^2	4,096	16×32^2	1.59×10^7	33.5	**69,147**
							Total:	553,172 SUs
							Other resources:	-76,800 SUs
							Final Total:	**476,372 SUs**

5. Computational Resources

Access to the TACC system STAMPEDE 2 is *essential* for the success of our research project. Locally, our group does have access to a moderate sized high performance computing system, SUMMIT. This machine has 9,120 general computing CPUs (plus GPU capability on other nodes) and is operated through the University of Colorado, Boulder. However, using this machine for the simulations proposed here is problematic for two reasons. First, the number of cores required for a single shearing box simulation is a substantial fraction of the maximum number of cores on SUMMIT. Thus, we could never expect this system to permit regular access to the 4,096 cores required for our calculations. Second, in consulting with the Research Computing division at the University of Colorado, it has become clear that we can acquire at most one million CPU-hours on this machine per year *per research group*, significantly less than what we require to carry out our proposed simulations.

For each of the quarterly windows of submission, an XSEDE Resource Allocation Committee (XRAC) is responsible for reviewing requests. This committee includes about 60 computational scientists with expertise in various fields covering the diversity of fields of science encountered at XSEDE, though the most represented fields such as Material Research are more populated. Reviewers serve for a three-year term. The role of this team is to conduct an impartial and independent review of the resource allocation requests in order to make a *recommendation* regarding the amount of resources that the project (or "grant") would need to be completed. Importantly for our empirical design, reviewers are asked to gauge requests only on their own merit and are specifically instructed to not consider the amount of resources available. To that end, they are not informed about the total requests on the different systems while they complete their reviews. This approach is designed to ensure that reviews and recommendations are based purely on what the proposal would need to be completed.

Figure 4: Total requests vs supply over time



After each meeting, the Allocations team meets to reconcile the recommendations made by the XRAC with the overall supply of resources. This process starts with a "balancing" phase, where allocations from over-recommended resources are moved to architecturally equivalent under-subscribed resources until the excess demand is balanced equally across computing platforms. Next, this is followed by a "reconciliation" phase where recommendations are reduced through a formulaic solver in order to determine final *allocations*. The formula gives priority to NSF-funded projects, which represent about 50% of the proposals (see Appendix 3 for more details on the reconciliation formula). Importantly, the only project-related characteristics considered in the reduction phase are (i) the recommendation made by the XRAC and (ii) the share of the project funding that is supported by the NSF.

Figure 4 shows aggregate requests (i.e., demand), recommendations and final allocations (i.e., recommendations after considering supply available) for each quarter of our study period.

4 Empirical Strategy

Empirically, we want to estimate the impact of the amount of computing resources allocated to scientists on scientific output. In the ideal experiment, we would assign each grant a random number of computing hours and observe its subsequent output, which would allow us to estimate the effect of the marginal computing resource on each grant's productivity. While we cannot run such an experiment, we leverage the allocation process of XSEDE in order to get plausibly causal estimates.

In most similar settings one could study, the typical threat to a causal interpretation is that projects and researchers with a higher expected productivity will receive higher computing allocations, which would bias estimates upwards. We overcome this by exploiting the fact that, conditional on the recommendation from the XRAC and the percentage of project funding that is supported by the NSF, each grant's final allocation is a function of how busy the XSEDE's resources are, which is exogenous to the expected productivity of the grant. Said differently, conditional on the recommendation made by the XRAC and the percentage of the project funding financed by the NSF, the amount of resources recommended to other projects, which influences the degree to which the XSEDE resources are constrained, is orthogonal to any feature of the focal grant. Because the only grant-related parameters included in the reconciliation formula are the amount of computing resources recommended by the XRAC and the share of the project funding supported by the NSF, the amount of resources allocated to a grant can be considered exogenous once we control for these two parameters.

Our regressions of interest are of the form:

$$Y_i = \beta_0 + \beta_1 \% Allocated_i + \delta Recommended_i + \gamma \% NSF\_Funding_i + \mathbf{X_i} + \varepsilon_i \tag{1}$$

where i indexes grants. Our treatment variable is $\% Allocated_i = \frac{Allocated_i}{Recommended_i}$ which captures the fraction of resources allocated to a grant compared to what it would have objectively (as assessed by the XRAC) needed. Using this ratio as treatment variable rather than the level of resources allocated allows us to interpret the coefficient β_1 as the impact of receiving more or less computing resources compared to what a researcher would have needed in theory. Note that by definition of the XSEDE Allocation process, this ratio is less than or equal to 1. We control for the recommendation made by the XRAC ($Recommended_i$) either through a linear term or with decile bins. One can think of this variable as capturing a combination of project quality and scale. The variable $\% NSF\_Funding_i$ accounts for the share of the project funding that is supported by the NSF, since this is part of the reconciliation formula. We also include field of science fixed-effects, since some fields might apply for specific resources which are deferentially more or less busy and these fields might also have different research productivity norms (e.g., more papers). We also include grant start year fixed-effects in order to control for changes in overall supply over time and the fact that more recent projects had less time to generate publications. Depending on the specification, we will also include the interaction of fields of science and grant start year fixed-effects. We use robust standard errors clustered at the grant start year $\times$ field level. Table A3 shows that the main independent variable of interest $\% Allocated$ is uncorrelated with pre-determined covariates (PI seniority, institution, number of papers published before XSEDE and number of citations received before XSEDE).

Note that contrary to other settings such as the NIH, the XRAC has information about the amount of computing resources requested by researchers. It is therefore possible that recommendations are "anchored" on requests and that some researchers request more than what their projects need and hence get recommended more. In order for this to create endogeneity issues, asking for more should be correlated with both (i) the outcomes (i.e., PIs who ask for more should be either more or less productive than the average) and (ii) the main explanatory variable (i.e., PI asking for more should have a higher or lower ratio than the average). While we cannot observe which PIs asked for more than their true need, we create a proxy based on the difference between the recommendation and the request.[8] We then examine the correlation between this variable and the main outcomes and between this variable and the ratio variable (including the controls). Table A2 shows no significant correlation.

5 Data and Descriptive Statistics

5.1 Data

To study our question, we use data about allocation awards covering each quarter between 2014 and 2022.[9] We study proposals that get accepted, focusing on the intensive margin – around 70% of projects are accepted and thus are awarded some computing power. For each (accepted) proposal, the data contains information at the computing resource level about the initial request submitted by the PI, the recommended amount suggested by the XRAC and the final allocation that was awarded. We drop resources corresponding to storage and convert request, recommended, and allocated information associated with different resources into the standard unit of Service Unit. We also observe information about the proposal field of science, as well as PI first name, last name and organization. While computing resources get awarded for a year, PI can ask for renewals which go through the same allocation process as their initial requests. We call the combination of the initial request and its subsequent renewals an XSEDE *grant* and collapse information at the XSEDE grant level, keeping only grants for which we have complete information about their resources and for which we observe their first request year. Hence, each row of our dataset contains information about the total amount of requested resources made by the PI, the corresponding total amount recommended by the XRAC and the total amount of allocated resources approved by the XSEDE allocation team. In theory, we could have kept our dataset at the request level instead of collapsing it at the PI level. However, a major share of the scientific output that we observe can only be attributed to PIs, and not to individual requests, making *grants* the relevant unit of observation. This corresponds to 1,068 unique grants associated with 1,068 unique PIs.

In order to match each XSEDE grant to its scientific research production, we use two sources of data. We first perform a fuzzy match of our population of PIs to the Dimensions AI database[10] using first

[8] Larger gaps between the recommendation and request *may* suggest that the request was inflated by the researcher.
[9] We are missing complete information about proposals for the years before 2014.
[10] https://www.dimensions.ai/

name, last name and institution name. Once PIs are identified, we retrieve their papers and keep those that reference "XSEDE" in the acknowledgments section.[11] Dropping chapters, this gives us 3,856 papers. In order to complement this database, we also got access to the XSEDE internal publication database, where PIs can report the publications associated with each of their grants. From this, we retrieved 922 additional papers. Note that PIs' decision to report to the internal XSEDE database is voluntary. This is not an issue for our empirical design, as long as the decision to voluntarily disclose is not correlated with our treatment variable. In Appendix 4, we show that there is no correlation between our treatment variable and papers identified through the use of the internal database.

We then match these papers to their Dimension equivalent based on DOI number or title. Our exploration of underlying mechanisms relies heavily on "concepts" from the Dimensions database. The Dimensions AI database uses machine learning techniques to derive concepts in papers' abstracts and rank them based on their relevance on a scale from 0 (not relevant) to 1 (very relevant). In all analyses, we only consider concepts with a score above 0.5.[12]

Among the initial 1,068 grants, we observe scientific output for 725 of them. Among the 343 unmatched grants, 112 come from PIs we were unable to match or to uniquely match with Dimensions. For the remaining 231 grants, 1 of them has papers which acknowledged XSEDE but could not be uniquely matched to a grant because their PIs were associated with several grants, and 230 grants have PIs who did not report any paper to the internal XSEDE database nor acknowledge XSEDE in one of their publications. We also drop 12 grants whose ratio is strictly higher than 1 due to XSEDE data entry errors (see Table 2 for the successive sampling restrictions). Our final sample hence includes 713 unique grants. In order to assess whether matched grants are a selected sample, we construct an indicator equal to 1 if the grant is part of the 713 grants that constitute our main sample and 0 otherwise and we regress this measure on the main independent variable $\%Allocated$ and the main controls. Table A1 shows no statistically significant relationship.

[11] From our discussions with PIs who were part of the XSEDE program, acknowledging XSEDE in the papers is the norm. We may be missing papers that attempt to acknowledge a XSEDE without specifically mentioning a variation of "XSEDE" or "Extreme Science and Engineering Discovery Environment," but to be conservative we capture papers that we are confident originate from XSEDE computing allocations.

[12] Several papers use MeSH terms to analyze the research topics of scientific papers (Azoulay et al. 2011, Myers 2020). MeSH terms are used for publications in PubMed which covers the biomedical literature. The use of Dimensions AI concepts allows us to extend this analysis to all the fields of science in our dataset. Appendix 5 shows examples of papers' concepts and abstracts.

Table 2: Sample construction

Sample	Observations
Grants	1,068
Grants with identifiable PIs	956
Grants with a 1:1 correspondence with their PIs	955
Grants with at least one associated paper	725
Grants with a ratio≤1	713

Notes: This table shows the successive steps that lead to the construction of the final sample with 713 observations.

5.2 Descriptive Statistics

Table 3 shows some descriptive statistics about our sample of 713 grants.

Table 3: Summary statistics

	Mean	SD	Min	Max
Requested (Mil. SU)	62.4	144.8	0.0	1627.6
Recommended (Mil. SU)	42.4	91.9	0.1	930.7
Approved (Mil. SU)	36.4	80.8	0.1	862.9
% Allocated (Allocated / Recommended)	0.9	0.2	0.3	1.0
Grant start year	2017.9	2.0	2014.0	2022.0
Number of papers	6.6	10.8	1.0	161.0
Total cites	149.8	391.5	0.0	6697.0
Average number of unique concepts	24.7	9.8	1.0	66.7
R1 universities	0.8	0.4	0.0	1.0
Junior PI	0.2	0.4	0.0	1.0
Log(use + 1)	1.0	0.7	0.0	7.1
Mean concept year	1999.0	5.5	1980.0	2018.5
Share of new concepts (2yr)	0.6	0.1	0.0	0.7
Observations	713			

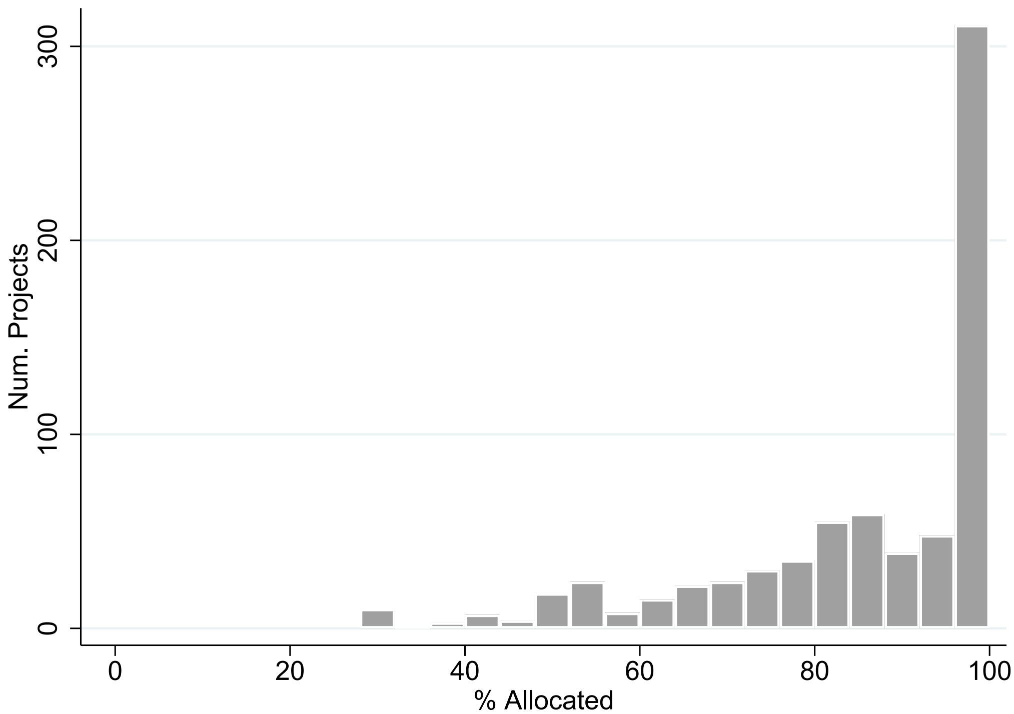
Notes: All summary statistics are at the grant level.

On average, researchers request 62.4 million service units (SU).[13] The XRAC makes lower recommendations on average, equal to 42.4 million SU and researchers gets a final average allocation of 36.4

[13]An SU can be thought of as a "core-hour." For perspective, on a quad core laptop, a project that is allocated 62.4 million SU would run for roughly 1,781 years.

million SU, with large variation in this number. On average, the share of allocated resources compared to the recommendation made by the review committee is 90%, but this variable is skewed. This can be seen in Figure 5 which shows the histogram of the treatment variable (*% Allocated*). The total numbers of papers associated with each grant (6.6 on average) is also skewed. A majority of PIs in our sample come from a R1 university, and 23% of them have less than 8 years of experience and are therefore classified as "Junior."[14]

Figure 5: Histogram of the share of resources allocated

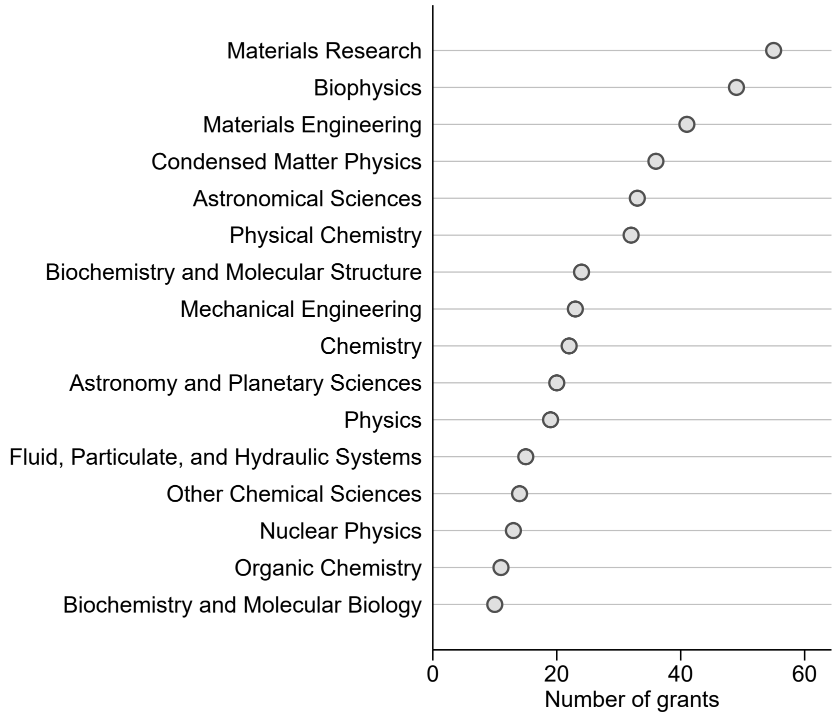


Notes: This figure reports the distribution of the treatment variable %Allocated, which represents the ratio of computing resources allocated to a grant compared to the amount of resources recommended by the XRAC.

Figure 6 shows the 15 most common fields of science in our sample. The most represented field is "Materials Research" with 55 grants, followed by Biophysics (49 grants) and Materials Engineering (41 grants).

[14]We take the first year for which we observe a publication associated with the PI as our best proxy for career start

Figure 6: Most common fields of science (top 15)



Notes: This figure reports the number of grants per field of science for the most common 15 fields of science.

6 Results

6.1 Quantity of Science

Our conceptual framework discusses the relationship between the amount of computing resources available to a researcher and the *quantity* and *type* of science she produces. In order to study the impact of additional resources on *quantity*, our first outcome of interest is the (log) total number of publications associated with each grant.[15] Table 4 shows the results of regressions of this outcome on the share of allocated resources received by grants. Column (1) includes controls for the amount of computing resources recommended by the XRAC using binned *Recommended* fixed effects[16] and column (2) uses a linear control. Both specifications include grant start year fixed effects, field of science fixed effects, grant start year $\times$ field fixed effects and control for the share of the project

[15]Since we only keep grants with a positive number of papers, this outcome is always strictly positive avoiding issues outlined in Chen & Roth (2024). We also confirm that results are robust to an inverse hyperbolic sine transformation.
[16]We create 10 bins of equal size.

funding that is supported by the NSF. Both columns show a strong positive relationship, implying that a 10 percentage point increase in the percentage of allocated resources leads to approximately a 7% increase in the number of papers produced. Equivalently, an increase from the 10^{th} to the 75^{th} percentile in the share of allocated resources leads to a 28% increase in the number of papers produced.[17] This result suggests evidence consistent with an increase in the number of papers as resources constraints are relaxed.

Table 4: Quantity of papers

	Binned	Linear
	(1)	(2)
$\%Allocated_i$	0.694***	0.693***
	(0.249)	(0.223)
%NSF $Funding_i$	-0.068	-0.066
	(0.082)	(0.081)
$\ln(Recommended_i)$		0.225***
		(0.029)
Field × Year FE	Yes	Yes
Rec Bin FE	Yes	
Lin Rec Ctrl		Yes
Dep Var Mean	1.7	1.7
R-Sq	0.48	0.47
Observations	522	522

Notes: The unit of observation is a grant. The outcome is the (log) number of papers. The observation count shown is the count with remaining variation after fixed effects are included. $\%Allocated$ is the share of resources allocated to a grant compared to the amount of resources recommended by the XRAC. $\%NSF\,Funding$ is the share of funding coming from the NSF. Column (1) bins the recommended amount into 10 bins of equal size and controls for the amount recommended by including bins fixed-effects. Column (2) controls for the amount recommended with a linear term. All regressions include grant start year fixed effects, grant field fixed effects and grant start year × grant field fixed effects. Standard errors are clustered at the grant start year × field level.

6.2 Type of Science

In what follows, we explore the consequences of an increase in the share of allocated computing resources on the type of science that researchers pursue. We successively discuss whether such increase leads to: (i) researchers pursuing less traveled research avenues (ii) researchers studying topics beyond

[17]The 10th percentile of $\%Allocated$ equals 59% and the 75th percentile of $\%Allocated$ equals 1.

their expertise (iii) researchers broadening the scope of their research. Importantly, the interest of this paper is to capture the characteristics of the science produced rather than the reception in the community. We aim to measure characteristics of the work at completion, rather than after the fact. Thus, we do not rely on citation-based measures which bundle together inherent quality with citation norms in the community.[18]

6.2.1 Less traveled research avenues

First, an increase in the share of computing resources could increase researchers' propensity to work on less-studied topics. For each paper, we calculate the number of times each concept it entails has been used by other PIs, excluding all papers from the focal PI.[19] We then average this measure at the grant level. The higher this outcome, the more the concepts used by the focal PI have been used by other researchers too – these are "more studied" concepts. The lower this outcome, the less well-studied the concept is. Results are presented in Columns (1) and (2) of Table 5. Column (1) controls for *Recommended* bins while column (2) uses a linear control. Results for both specifications are negative and statistically significant, showing that the higher the share of allocated resources, the less popular are the topics studied. The coefficients imply that a 10% increase in the share allocated leads to the study of concepts that are 2 to 3% less studied by other researchers.

This finding could be driven by (i) the study of topics that were initially less popular but rising in popularity (ii) the study of topics that were popular but losing attractiveness or (iii) the study of topics that have a persistently lower level of use by others. In order to investigate this, we calculate for each paper the number of times each concept it entails has been used by other PIs *prior* to the paper publication year and average this measure at the grant level. Results are presented in columns (3) and (4) and show a significantly negative coefficient, similar in magnitude to the previous results. Similarly, we calculate for each paper the number of times each concept it entails has been used by other PIs *after* the paper publication year and average this measure at the grant level. Results are presented in columns (5) and (6), which show negative and statistically significant coefficients with

[18]For completeness, we explore citation based measures in section 7.1.

[19]We standardize this number to account for publication year and field. This allows us to account for difference across fields and the fact that concepts used in more recent papers had less time to be used by other researchers.

Table 5: Use of concepts by others

	Log(1+Use)		Log(1+Use Before)		Log(1+Use After)	
	(1)	(2)	(3)	(4)	(5)	(6)
$\%Allocated_i$	-0.228**	-0.279***	-0.228**	-0.278***	-0.211**	-0.272***
	(0.097)	(0.098)	(0.099)	(0.100)	(0.096)	(0.097)
%NSF $Funding_i$	0.022	0.028	0.024	0.030	0.012	0.019
	(0.028)	(0.028)	(0.029)	(0.028)	(0.028)	(0.029)
$\ln(Recommended_i)$		0.014*		0.015*		0.009
		(0.008)		(0.008)		(0.008)
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Rec Bin FE	Yes		Yes		Yes	
Lin Rec Ctrl		Yes		Yes		Yes
Dep Var Mean	0.6	0.6	0.6	0.6	0.6	0.6
R-Sq	0.06	0.03	0.06	0.03	0.05	0.03
Observations	670	670	670	670	659	659

Notes: The unit of observation is a grant. The observation count shown is the count with remaining variation after fixed effects are included. The outcome in columns (1) and (2) is the (log) average number of citations received by concepts associated with a grant. Columns (3) and (4) focus on citations received prior to XSEDE entry and columns (5) and (6) focus on citations received after XSEDE entry. *%Allocated* is the share of resources allocated to a grant compared to the amount of resources recommended by the XRAC. *%NSF Funding* is the share of funding coming from the NSF. Columns (1), (3) and (5) control for the amount recommended with 10 bins of equal size. Columns (2), (4) and (6) control for the amount recommended with a linear term. All models include grant start year fixed effects. Standard errors are clustered at the grant start year × field level.

similar magnitudes to the previous four columns.[20] This suggests that when researchers are allocated a higher share of resources, they study topics that are persistently less popular.

We also study whether the concepts used are more or less recently introduced to the scientific literature as researchers get allocated additional computing resources. We take each concept associated with a paper and calculate the first year we see these concept appearing among XSEDE researchers.[21] We then average this measure at the grant level. This measure proxies how recent the concepts used by a grant are on average. Results are presented in Table 6. Columns (1) and (2) control for *Recommended* bins while (3) and (4) use linear controls. Columns (1) and (3) do not include grant start year fixed

[20] We conduct an t-test and confirm that these coefficients are statistically indistinguishable.

[21] Since we do not observe the *universe* of scientific articles, we cannot observe with certainty when the topic was initially studied overall. Though, the first time we observe a topic in XSEDE serves as a good proxy for its appearance in the computational scientific literature.

effects, and columns (2) and (4) do. All specifications find significantly positive results, suggesting that the higher the share allocated resources, the newer the topics being studied are. More precisely, a 10% increase in %*Allocated* is associated with the study of topics that are 5.5 to 6 years newer on average. Combining this results with Table 5 suggests that as computing resources increase, the concepts being studied are less common and newer-to-the-world. Such topics may not necessarily be inherently of lesser value (or "worse"), but rather are less mature than the average topic.

Table 6: Mean year of concepts

	Binned		Linear	
	(1)	(2)	(3)	(4)
$\%Allocated_i$	5.919***	5.566***	5.975***	5.774***
	(1.475)	(1.540)	(1.396)	(1.446)
%NSF $Funding_i$	-1.306***	-1.266***	-1.315***	-1.278***
	(0.461)	(0.480)	(0.454)	(0.476)
$\ln(Recommended_i)$			-0.476***	-0.490***
			(0.151)	(0.150)
Year FE		Yes		Yes
Rec Bin FE	Yes	Yes		
Lin Rec Ctrl			Yes	Yes
Dep Var Mean	1999.0	1999.0	1999.0	1999.0
R-Sq	0.09	0.09	0.07	0.07
Observations	670	670	670	670

Notes: The unit of observation is a grant. The observation count shown is the count with remaining variation after fixed effects are included. For each paper associated with a grant, we calculate the first time (measured in years) we observe this concept and average this measure at the grant level. %*Allocated* is the share of resources allocated to a grant compared to the amount of resources recommended by the XRAC. $\%NSF\ Funding$ is the share of funding coming from the NSF. Columns (1) and (2) control for the amount recommended with 10 bins of equal size. Columns (3) and (4) control for the amount recommended with a linear term. All models include grant start year fixed effects. Standard errors are clustered at the grant start year × field level.

Taken together, we find strong evidence consistent with the fact that researchers who receive a higher allocation choose to study topics that are less popular and are newer to the world.

6.2.2 Topics beyond expertise

Second, an increase in the share of computing resources could increase researchers' propensity to "pivot," i.e., to work on topics that they haven't studied previously. To analyze this, we calculate for each paper in our sample the share of concepts that have not been used by the focal PI in the years preceding their first entry into XSEDE. We then average this measure at the grant level. Conceptually, we are capturing the degree of "pivoting" or "exploration" by each researcher, studying topics that they are less familiar with.

Table 7: Share of new concepts

	Vs prev. 2y		Vs prev. 3y		Vs prev. 5y	
	(1)	(2)	(3)	(4)	(5)	(6)
$\%Allocated_i$	0.037	0.037	0.051	0.052	0.049	0.048
	(0.039)	(0.035)	(0.041)	(0.035)	(0.042)	(0.037)
%NSF $Funding_i$	0.003	0.005	0.000	0.001	-0.001	0.000
	(0.016)	(0.016)	(0.016)	(0.016)	(0.015)	(0.015)
$\ln(Recommended_i)$		-0.000		0.000		-0.002
		(0.005)		(0.005)		(0.005)
Field × Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Rec Bin FE	Yes		Yes		Yes	
Lin Rec Ctrl		Yes		Yes		Yes
Dep Var Mean	0.6	0.6	0.6	0.6	0.5	0.5
R-Sq	0.46	0.45	0.45	0.45	0.45	0.45
Observations	430	430	438	438	446	446

Notes: The unit of observation is a grant. The observation count shown is the count with remaining variation after fixed effects are included. The outcome is the (log) average number of new concepts associated with a grant. Columns (1) and (2) compare XSEDE papers with papers written in the 2 years that precede XSEDE entry. Columns (3) and (4) compare XSEDE papers with papers written in the 3 years that precede XSEDE entry. Columns (5) and (6) compare XSEDE papers with papers written in the 5 years that precede XSEDE entry. *%Allocated* is the share of resources allocated to a grant compared to the amount of resources recommended by the XRAC. *%NSF Funding* is the share of funding coming from the NSF. Columns (1), (3) and (5) control for the amount recommended with 10 bins of equal size. Columns (2), (4) and (6) control for the amount recommended with a linear term. All models include grant start year fixed effects, field of science fixed effects and grant start year fixed effects × field of science fixed effects. Standard errors are clustered at the grant start year × field level.

Table 7 shows the results of the regressions using the (log) share of new concepts as an outcome. Columns (1), (3), and (5) include *Recommended* bin controls and columns (2), (4), and (6) include linear recommended controls. Columns (1) and (2) use the pool of papers written by the researcher in the two years preceding her entry into XSEDE to define the initial pool of papers with which we compare papers written as part of XSEDE. Columns (3) and (4) expand this window to the three years prior to entry into XSEDE and columns (5) and (6) use a five year window. In all specifications, we find non-statistically significant coefficients. This suggests that getting a higher share of resources may not be associated with researcher pivoting.

6.2.3 Research breadth

Finally, an increase in the share of computing resources could increase the research breadth of projects pursued by researchers. To explore this, we study the effect of the share of allocated computing resources on the number of unique concepts cited by papers associated with a grant. We first calculate the number of unique concepts associated with each paper, which informs us about the diversity of ideas it embodies. Appendix 5 shows an example of this measure. We then average this measure at the grant level. We use this as a proxy for the breadth of research associated with a grant: the higher the average number of concepts, the broader the research.[22]

Table 8: Research Breadth

	Log(1+Unique Concepts)	
	(1)	(2)
$\%Allocated_i$	0.371**	0.333**
	(0.180)	(0.161)
%NSF $Funding_i$	0.039	0.050
	(0.052)	(0.049)
$\ln(Recommended_i)$		-0.015
		(0.021)
Field × Year FE	Yes	Yes
Rec Bin FE	Yes	
Lin Rec Ctrl		Yes
Dep Var Mean	3.1	3.1
R-Sq	0.36	0.35
Observations	487	487

Notes: The unit of observation is a grant. The observation count shown is the count with remaining variation after fixed effects are included. The outcome in columns (1) and (2) is the (log) average number of concepts associated with a grant. *%Allocated* is the share of resources allocated to a grant compared to the amount of resources recommended by the XRAC. *%NSF Funding* is the share of funding coming from the NSF. Column (1) controls for the amount recommended with 10 bins of equal size. Column (2) controls for the amount recommended with a linear term. All models include grant start year fixed effects, field of science fixed effects and grant start year fixed effects × field of science fixed effects. Standard errors are clustered at the grant start year × field level.

[22] In contrast to the journal breadth measure used in Azoulay et al. (2011), our measure more directly captures the researchers pursuit of broad topics separately from the impact in the scientific community.

Table 8 shows the results. Column (1) uses *Recommended* bin fixed effects and column (2) uses a linear control. In both specifications, we find a significantly positive coefficient, implying that a 10 percentage point increase in the share of allocated resources leads to a 3 to 4% increase in the number of unique concepts. This indicates that a higher share of allocated resources leads researchers to expand their research breadth.

7 Additional analyses

7.1 Scientific reception and impact

Overall, we find evidence consistent with the fact that, as researchers gets more computing resources, they write more papers. We also find that researchers with additional resources study newer, less popular topics and recombine more areas of knowledge. These findings allow us to closely capture the impacts of computing resources on researcher strategies as they are products of the research undertaken by the scientist. In contrast, the process of receiving citations in the scientific literature is not a perfect relationship to quality or impact (Azoulay et al. 2011) but is nonetheless interesting, so we briefly explore this in a post-hoc manner.

We study the impact of an increase in the share of computing resources allocated to researchers on citations. Our main outcome of interest is the average number of citations received by papers associated with each grant. In order to control for publication year and the field of science associated with each paper, we use an adjusted citation measure provided by Dimensions AI that calculates the scientific influence of a paper by dividing the number of citations it received by the average number of citations received by papers published in the same year and field of research. A value of 1 means that a paper receives exactly the average number of citations expected for its publication year and field. This essentially allows us to compare the citations received by papers in our sample relative to the *universe* of papers, while normalizing by publication year and field. We then collapse this measure at the grant level, taking the average across papers. This measure alleviates potential concerns about truncation issues when studying papers published relatively recently, since papers are compared within field and publication year. Results are presented in Table A4. All specifications control for the percentage of NSF funding received. Columns (1) and (2) control for the amount of resources recommended with bins, while columns (3) and (4) use a linear control. Columns (1) and (3) are our preferred specifications which omit grant start year fixed effects since the outcome is already normalized by year. Results show a negative relationship, significant at the 10% level. Columns (2) and (4) add grant start year fixed effect and show similar negative magnitudes though the coefficients are not statistically significant. Overall, results suggest that a 10 percentage point increase in the percentage of allocated resources is associated with a 3% to 4% decrease in the average normalized citation score received for the bundle of papers associated with a grant.[23]

At first glance, this result could be driven by a decrease in the inherent quality of the research or by a decrease in the scientific traction these projects receive. We find evidence suggesting that an increase in the share of allocated resources does not lead to a decrease in inherent quality by looking at the journals in which papers are published. Figure A1 exploits information about the SNIP (Source Normalised Impact per Paper) value of the journal associated with each paper of a grant. SNIP measures the average citation impact of the publications of a journal and is therefore a good proxy of quality. A higher share of allocated resources leads to a similar increase in the number of papers published in each journal quality bin, suggesting that an increase in computing resources has no significant impact on the *ex ante* quality distribution. Hence, the citation result rather suggests that the different research direction followed by researchers with more computing resources does not gain more traction in the scientific community.

7.2 Heterogeneity

We also analyze whether returns to more computing resources differ as a function of (i) access to outside options and (ii) researcher experience. To study (i), we differentiate between researchers affiliated with R1 vs non-R1 institutions. As R1 institution rely more heavily on research, we might expect researchers at these universities to have access to a greater range and/or better computing outside options. To study (ii), we use the cutoff of 8 years of experience to differentiate between 'junior' and 'senior' researchers. As junior researchers are less likely to be tenured, they might adopt different research strategies than senior when experiencing an increase in the share of computing resources allocated to

[23] In Table A5, we also construct an alternative field-year normalized citation measure that only uses our sample rather than the universe of papers. Results are qualitatively similar. Tables A6 and A7 repeat this exercise with the median instead of the mean and find similar results.

them.

Columns (1) and (2) of Tables A8, A9, A10 and A11 show no evidence of differential productivity by access to other sources of computing power, as proxied by if the researcher has an R1 university affiliation. This is consistent with the fact that the resources offered by XSEDE are not easily substituable.

Columns (3) and (4) of Table A8 finds no difference in the number of papers produced by junior and senior PIs. However, Columns (3) and (4) of Tables A9, A10 and A11 show interesting heterogeneity in the type of science produced by these two groups. As computing resources increase, junior PIs are less likely to pursue less traveled research avenues and topics beyond their expertise compared to more experienced researchers, but are more likely to pursue broader research projects.

8 Conclusion

In this paper, we build upon a conceptual framework that outlines the potential pathways through which variations in computing resources influence researchers' strategies and subsequent scientific outcomes. We posit that an increase in computing resource availability increases research output and impacts the type of science that researchers pursue. We then empirically investigate the relationship between the share of computing resources received by researchers and research output in the context of XSEDE, shedding light on the complex relationship between resource allocation and innovation outcomes. First, our findings indicate that an increase in the share of allocated resources leads to an increase in research output, as evidenced by the increase in the number of papers published. We then explore changes to research direction. Our findings indicate that a higher share of allocated resources leads researchers to work on newer and less popular topics. We also find that researchers who receive more computing resources tend to broaden the scope of their research.

Our findings suggest that computing resources are a key constrained input to the innovation process which significantly impact both the rate and direction of innovation. As constraints are relaxed, researchers tend to produce more papers, explore newer and understudied research areas and undertake more recombinatory work. Our study highlights the multifaceted nature of the relationship between resource allocation, research strategies, and outcomes within the XSEDE program. These findings underscore the importance of considering the nuanced impacts of resource allocation strategies on research outcomes and the need for continued exploration of the underlying mechanisms driving these dynamics.

References

Agrawal, A., Catalini, C., Goldfarb, A. & Luo, H. (2018), 'Slack time and innovation', *Organization Science* **29**(6), 1056–1073.

Arora, A., Belenzon, S. & Patacconi, A. (2018), 'The decline of science in corporate R&D', *Strategic Management Journal* **39**(1), 3–32.

Arora, A. & Gambardella, A. (2005), 'The impact of nsf support for basic research in economics', *Annales d'Economie et de Statistique* pp. 91–117.

Arrow, K. (1962), Economic Welfare and the Allocation of Resources for Invention, Princeton University Press, Princeton, pp. 609–626.

Atanassov, J. (2016), 'Arm's length financing and innovation: Evidence from publicly traded firms', *Management Science* **62**(1), 128–155.

Azoulay, P., Graff Zivin, J. S., Li, D. & Sampat, B. N. (2019), 'Public R&D Investments and Private-sector Patenting: Evidence from NIH Funding Rules', *Review of Economic Studies* **86**(1), 117–152.

Azoulay, P., Graff Zivin, J. S. & Manso, G. (2011), 'Incentives and creativity: evidence from the academic life sciences', *The RAND Journal of Economics* **42**(3), 527–554.

Barney, J. (1991), 'Firm resources and sustained competitive advantage', *Journal of management* **17**(1), 99–120.

Berrone, P., Fosfuri, A., Gelabert, L. & Gomez-Mejia, L. R. (2013), 'Necessity as the mother of 'green'inventions: Institutional pressures and environmental innovations', *Strategic Management Journal* **34**(8), 891–909.

Boudou, J. (2024), 'Are firms stealing talents? the sorting of scientists between industry and academia and consequences for science', *Working Paper* .

Boyabatlı, O., Leng, T. & Toktay, L. B. (2016), 'The impact of budget constraints on flexible vs. dedicated technology choice', *Management Science* **62**(1), 225–244.

Bronzini, R. & Piselli, P. (2016), 'The impact of r&d subsidies on firm innovation', *Research Policy* **45**(2), 442–457.

Catalini, C. (2018), 'Microgeography and the direction of inventive activity', *Management Science* **64**(9), 4348–4364.

Cerqueiro, G., Hegde, D., Penas, M. F. & Seamans, R. C. (2017), 'Debtor rights, credit supply, and innovation', *Management Science* **63**(10), 3311–3327.

Chen, J. & Roth, J. (2024), 'Logs with zeros? some problems and solutions', *The Quarterly Journal of Economics* **139**(2), 891–936.

Choi, B., Kumar, M. S. & Zambuto, F. (2016), 'Capital structure and innovation trajectory: The role of debt in balancing exploration and exploitation', *Organization Science* **27**(5), 1183–1201.

Danneels, E. & Sethi, R. (2011), 'New product exploration under environmental turbulence', *Organization Science* **22**(4), 1026–1039.

Dasgupta, P. & David, P. A. (1994), 'Toward a new economics of science', *Research Policy* **23**(5), 487–521.

Ding, W. W., Levin, S. G., Stephan, P. E. & Winkler, A. E. (2010), 'The impact of information technology on academic scientists' productivity and collaboration patterns', *Management Science* **56**(9), 1439–1461.

Dutt, N. & Lawrence, M. (2022), ‘Learning to manage breadth: experience as repetition and adaptation’, *Organization Science* **33**(4), 1300–1318.

Ferguson, J.-P. & Carnabuci, G. (2017), ‘Risky recombinations: Institutional gatekeeping in the innovation process’, *Organization Science* **28**(1), 133–151.

Fini, R., Perkmann, M. & Ross, J.-M. (2022), ‘Attention to exploration: The effect of academic entrepreneurship on the production of scientific knowledge’, *Organization Science* **33**(2), 688–715.

Fitzgerald, T., Balsmeier, B., Fleming, L. & Manso, G. (2021), ‘Innovation search strategy and predictable returns’, *Management science* **67**(2), 1109–1137.

Foster, J. G., Rzhetsky, A. & Evans, J. A. (2015), ‘Tradition and innovation in scientists’ research strategies’, *American Sociological Review* **80**, 875–908.

Franzoni, C. & Stephan, P. (2023), ‘Uncertainty and risk-taking in science: Meaning, measurement and management in peer review of research proposals’, *Research Policy* **52**(3), 104706.

Furman, J. L. & Stern, S. (2011), ‘Climbing atop the shoulders of giants: The impact of institutions on cumulative research’, *American Economic Review* **101**(5), 1933–1963.

Glennon, B. (2023), ‘How do restrictions on high-skilled immigration affect offshoring? evidence from the h-1b program’, *Management Science* .

Hill, R., Yin, Y., Stein, C., Wang, D. & Jones, B. (2021), ‘Adaptability and the pivot penalty in science’, *Working Paper* .

Hoegl, M., Gibbert, M. & Mazursky, D. (2008), ‘Financial constraints in innovation projects: When is less more?’, *Research Policy* **37**(8), 1382–1391.

Howell, S. T. (2017), ‘Financing innovation: Evidence from r&d grants’, *American economic review* **107**(4), 1136–1164.

Jacob, B. A. & Lefgren, L. (2011), ‘The impact of research grant funding on scientific productivity’, *Journal of public economics* **95**(9-10), 1168–1177.

Jones, B. F. (2009), ‘The Burden of Knowledge and the ”Death of the Renaissance Man”: Is Innovation Getting Harder?’, *Review of Economic Studies* **76**, 283–317.

Kim, S. (2023), ‘Shortcuts to Innovation: The Use of Analogies in Knowledge Production’, *Working Paper* .

Kong, D., Lin, C., Wei, L. & Zhang, J. (2022), ‘Information accessibility and corporate innovation’, *Management Science* **68**(11), 7837–7860.

Krieger, J. L., Myers, K. R. & Stern, A. D. (2023), ‘How important is editorial gatekeeping? evidence from top biomedical journals’, *Review of Economics and Statistics* pp. 1–33.

Krieger, J., Li, D. & Papanikolaou, D. (2022), ‘Missing novelty in drug development’, *The Review of Financial Studies* **35**(2), 636–679.

Lohr, S. (2019), At tech’s leading edge, worry about a concentration of power, Technical report.

Luger, J., Raisch, S. & Schimmer, M. (2018), ‘Dynamic balancing of exploration and exploitation: The contingent benefits of ambidexterity’, *Organization Science* **29**(3), 449–470.

Mandler, M. (2017), ‘The benefits of risky science’, *The Economic Journal* **127**(603), 1495–1526.

Marino, A., Aversa, P., Mesquita, L. & Anand, J. (2015), ‘Driving performance via exploration in changing environments: Evidence from formula one racing’, *Organization Science* **26**(4), 1079–1100.

Marx, M. & Fuegi, A. (2020), 'Reliance on science: Worldwide front-page patent citations to scientific articles', *Strategic Management Journal* **41**(9), 1572–1594.

Mezzanotti, F. & Simcoe, T. (2023), Research and/or development? financial frictions and innovation investment, Technical report, National Bureau of Economic Research.

Myers, K. (2020), 'The Elasticity of Science', *American Economic Journal: Applied Economics* **12**(4), 103–134.

Myers, K. & Lanahan, L. (2022), 'Estimating spillovers from publicly funded R&D: Evidence from the US Department of Energy', *American Economic Review* **Forthcoming**.
URL: *www.subsurfaceinsights.com*

Myers, K. & Tham, W. Y. (2023), 'Money, time, and grant design', *Working Paper* .
URL: *https://www.kylemyers.org/$_f$iles/ugd/73b908$_1$6ea1f7955fd472683781172c6c8bb0f.pdf*

Nagle, F. & Teodoridis, F. (2020), 'Jack of all trades and master of knowledge: The role of diversification in new distant knowledge integration', *Strategic management journal* **41**(1), 55–85.

Nanda, R. & Rhodes-Kropf, M. (2017), 'Financing risk and innovation', *Management Science* **63**(4), 901–918.

Nelson, R. R. (1959), 'The economics of invention: A survey of the literature', *The Journal of Business* **32**(2), 101–127.

Payne, A. A., Siow, A. et al. (1999), *Does federal research funding increase university research output?*, Citeseer.

Perry, M. & Reny, P. J. (2016), 'How to count citations if you must', *American Economic Review* **106**(9), 2722–2741.

Sidhu, J. S., Commandeur, H. R. & Volberda, H. W. (2007), 'The multifaceted nature of exploration and exploitation: Value of supply, demand, and spatial search for innovation', *Organization Science* **18**(1), 20–38.

Singh, J. & Marx, M. (2013), 'Geographic constraints on knowledge spillovers: Political borders vs. spatial proximity', *Management Science* **59**(9), 2056–2078.

Thompson, N. C., Ge, S. & Manso, G. F. (2022), 'The importance of (exponentially more) computing power'.

Tranchero, M. (2023), 'Finding Diamonds in the Rough: Data-Driven Opportunities and Pharmaceutical Innovation', *Working Paper* .

Zhou, K. Z. & Li, C. B. (2012), 'How knowledge affects radical innovation: Knowledge base, market knowledge acquisition, and internal knowledge sharing', *Strategic management journal* **33**(9), 1090–1102.

Zhuo, R. (2022), 'Exploit or explore? an empirical study of resource allocation in scientific labs', *Working Paper* .

APPENDIX

1 Tables

Table A1: Matched vs Unmatched Grants

	Binned	Linear
	(1)	(2)
$\%Allocated_i$	0.060	0.055
	(0.128)	(0.125)
$\%\text{NSF}\,Funding_i$	0.070*	0.071**
	(0.036)	(0.036)
$\ln(Recommended_i)$		0.025*
		(0.014)
Field × Year FE	Yes	Yes
Rec Bin FE	Yes	
Lin Rec Ctrl		Yes
R-Sq	0.4	0.4
Observations	843.0	843.0

Notes: The unit of observation is a grant. The dependent variable is an indicator equal to 1 if the grants is part of our sample of 713 grants and 0 otherwise. *%NSF Funding* is the share of funding coming from the NSF. Columns (1) and (2) control for grant start year fixed effects × field fixed effects. Column (1) controls for the amount recommended with 10 bins of equal size. Column (2) controls for the amount recommended with a linear term. Standard errors are clustered at the grant start year × field level.

Table A2: Over-requesting test

	Ratio	Papers	Ratio	Citations
	(1)	(2)	(3)	(4)
$(Requested_i - Recommended_i)$	0.062	0.713	-0.012	-0.086
	(0.158)	(0.896)	(0.070)	(0.336)
%NSF $Funding_i$	0.100***	0.001	0.072***	-0.011
	(0.018)	(0.080)	(0.013)	(0.077)
Field × Year FE	Yes	Yes		
Year FE			Yes	Yes
Rec Bin FE	Yes	Yes	Yes	Yes
R-Sq	0.5	0.5	0.3	0.1
Observations	522.00	522.00	713.00	622.00

Notes: The unit of observation is a grant. For each grant, "diff" is the difference between the requested and the recommended amounts. *%NSF Funding* is the share of funding coming from the NSF. Columns (1) and (2) control for grant start year fixed effects × field fixed effects. Columns (3) and (4) control for grant start year fixed effects. All columns control for the the amount recommended with 10 bins of equal size. Standard errors are clustered at the grant start year and field level.

Table A3: Balance test

	%Allocated				
	(1)	(2)	(3)	(4)	(5)
Junior PI	0.003				-0.004
	(0.016)				(0.016)
R1 Univ		-0.019			-0.008
		(0.019)			(0.020)
Nb papers pre-XSEDE			-0.000		-0.000
			(0.000)		(0.000)
Nb cites pre-XSEDE				-0.000	0.000
				(0.000)	(0.000)
Field × Year FE	Yes	Yes	Yes	Yes	Yes
Rec Bin FE	Yes	Yes	Yes	Yes	Yes
R-Sq	1	1	1	1	1
Observations	500.0	522.0	493.0	493.0	493.0

Notes: The unit of observation is a grant. *%Allocated* is the share of resources allocated to a grant compared to the amount of resources recommended by the XRAC. $R1$ is an indicator equal to 1 if the PI is affiliated with an R1 institution. *Junior* is an indicator equal to 1 the PI is below the median of experience in our sample, equal to 13y. All columns control for the share of funding coming from the NSF, the amount recommended with 10 bins of equal size as well as grant start year × field fixed effects. Standard errors are clustered at the grant start year and field level.

Table A4: Number of normalized citations

	Binned		Linear	
	(1)	(2)	(3)	(4)
$\%Allocated_i$	-0.370*	-0.337	-0.366*	-0.337
	(0.209)	(0.221)	(0.202)	(0.205)
%NSF $Funding_i$	0.016	0.016	0.019	0.019
	(0.076)	(0.080)	(0.076)	(0.080)
$\ln(Recommended_i)$			0.109***	0.113***
			(0.020)	(0.021)
Year FE		Yes		Yes
Rec Bin FE	Yes	Yes		
Lin Rec Ctrl			Yes	Yes
Dep Var Mean	1.6	1.6	1.6	1.6
R-Sq	0.06	0.07	0.05	0.06
Observations	622	622	622	622

Notes: The unit of observation is a grant. The outcome is the (log) average number of citations associated with a grant. The observation count shown is the count with remaining variation after fixed effects are included. Citations associated with each paper are normalized to account for differences across fields and publication year. $\%Allocated$ is the share of resources allocated to a grant compared to the amount of resources recommended by the XRAC. $\%NSF\ Funding$ is the share of funding coming from the NSF. Columns (1) and (2) control for the amount recommended with 10 bins of equal size. Columns (3) and (4) control for the amount recommended with a linear term. Columns (2) and (4) control for grant start year fixed effects.

Table A5: Log(1+Average Citations) - Constructed

	Binned		Linear	
	(1)	(2)	(3)	(4)
$\%Allocated_i$	-0.146	-0.120	-0.155	-0.133
	(0.101)	(0.115)	(0.094)	(0.102)
%NSF $Funding_i$	-0.064**	-0.066*	-0.061*	-0.062*
	(0.033)	(0.034)	(0.033)	(0.034)
$\ln(Recommended_i)$			0.034***	0.035***
			(0.009)	(0.009)
Year FE		Yes		Yes
Rec Bin FE	Yes	Yes		
Lin Rec Ctrl			Yes	Yes
Dep Var Mean	0.6	0.6	0.6	0.6
R-Sq	0.04	0.04	0.03	0.03
Observations	713	713	713	713

Notes: The unit of observation is a grant. For each paper in our sample, we follow Perry & Reny (2016) and standardize each citation it received by dividing it by the mean number of citations received by papers in our sample published in the same year and field. We then average this measure at the grant level. $\%Allocated$ is the share of resources allocated to a grant compared to the amount of resources recommended by the XRAC. $\%NSF\,Funding$ is the share of funding coming from the NSF. Columns (1) and (2) control for the amount recommended with 10 bins of equal size. Columns (3) and (4) control for the amount recommended with a linear term. Columns (2) and (4) control for grant start year fixed effects.

Table A6: Log(1+Median Citations) - Dimensions

	Binned		Linear	
	(1)	(2)	(3)	(4)
$\%Allocated_i$	-0.415**	-0.493**	-0.415**	-0.479**
	(0.199)	(0.212)	(0.195)	(0.197)
%NSF $Funding_i$	0.000	0.009	0.004	0.010
	(0.080)	(0.082)	(0.079)	(0.082)
$\ln(Recommended_i)$			0.087***	0.090***
			(0.022)	(0.021)
Year FE		Yes		Yes
Rec Bin FE	Yes	Yes		
Lin Rec Ctrl			Yes	Yes
Dep Var Mean	1.5	1.5	1.5	1.5
R-Sq	0.05	0.06	0.04	0.05
Observations	622	622	622	622

Notes: The unit of observation is a grant. The outcome is the (log) median number of citations associated with a grant. *%Allocated* is the share of resources allocated to a grant compared to the amount of resources recommended by the XRAC. *%NSF Funding* is the share of funding coming from the NSF. Columns (1) and (2) control for the amount recommended with 10 bins of equal size. Columns (3) and (4) control for the amount recommended with a linear term. Columns (2) and (4) control for grant start year fixed effects.

Table A7: Log(1+Median Citations) - Constructed

	Binned		Linear	
	(1)	(2)	(3)	(4)
$\%Allocated_i$	-0.162*	-0.155	-0.175**	-0.170*
	(0.093)	(0.109)	(0.088)	(0.097)
%NSF $Funding_i$	-0.084***	-0.085**	-0.079**	-0.080**
	(0.032)	(0.033)	(0.032)	(0.033)
$\ln(Recommended_i)$			0.018**	0.018**
			(0.009)	(0.009)
Year FE		Yes		Yes
Rec Bin FE	Yes	Yes		
Lin Rec Ctrl			Yes	Yes
Dep Var Mean	0.5	0.5	0.5	0.5
R-Sq	0.03	0.03	0.02	0.03
Observations	713	713	713	713

Notes: The unit of observation is a grant. For each paper in our sample, we follow Perry & Reny (2016) and standardize each citation it received by dividing it by the mean number of citations received by papers in our sample published in the same year and field. We then take the median of this measure at the grant level. $\%Allocated$ is the share of resources allocated to a grant compared to the amount of resources recommended by the XRAC. $\%NSF\,Funding$ is the share of funding coming from the NSF. Columns (1) and (2) control for the amount recommended with 10 bins of equal size. Columns (3) and (4) control for the amount recommended with a linear term. Columns (2) and (4) control for grant start year fixed effects.

Table A8: Number of papers – differences by institution and experience

	Resources		Experience	
	(1)	(2)	(3)	(4)
$\%Allocated_i$	0.737***	0.734***	0.787***	0.759***
	(0.206)	(0.212)	(0.234)	(0.261)
Non-R1 Univ=1	0.120	0.078		
	(0.541)	(0.594)		
Non-R1 Univ=1 × $\%Allocated_i$	-0.164	-0.142		
	(0.605)	(0.661)		
$\ln(Recommended_i)$	0.223***		0.259***	
	(0.031)		(0.027)	
%NSF $Funding_i$	-0.067	-0.072	-0.110	-0.106
	(0.082)	(0.082)	(0.077)	(0.078)
Junior=1			-0.318	-0.369
			(0.296)	(0.308)
Junior=1 × $\%Allocated_i$			0.368	0.427
			(0.334)	(0.344)
Field × Year FE	Yes	Yes	Yes	Yes
Rec Bin FE		Yes		Yes
Lin Rec Ctrl	Yes		Yes	
Dep Var Mean	1.7	1.7	1.7	1.7
R-Sq	0.47	0.48	0.55	0.56
Observations	522	522	500	500

Notes: The unit of observation is a grant. The outcome is the (log) number of papers. The observation count shown is the count with remaining variation after fixed effects are included. *%Allocated* is the share of resources allocated to a grant compared to the amount of resources recommended by the XRAC. $R1$ is an indicator equal to 1 if the PI is affiliated with an R1 institution. *Junior* is an indicator equal to 1 the PI has less than 8 years of experience. $\%NSF\,Funding$ is the share of funding coming from the NSF. Columns (1) and (3) controls for the amount recommended with a linear term. Columns (2) and (4) bins the recommended amount into 10 bins of equal size and controls for the amount recommended by including bins fixed-effects. All regressions include grant start year fixed effects, grant field fixed effects and grant start year × grant field fixed effects.

Table A9: Use of concepts by others – differences by institution and experience

	Resources		Experience	
	(1)	(2)	(3)	(4)
$\%Allocated_i$	-0.264***	-0.219**	-0.377***	-0.328***
	(0.094)	(0.095)	(0.100)	(0.097)
Non-R1 Univ=1	0.081	0.050		
	(0.185)	(0.180)		
Non-R1 Univ=1 × $\%Allocated_i$	-0.083	-0.049		
	(0.208)	(0.203)		
$\ln(Recommended_i)$	0.014*		0.017*	
	(0.008)		(0.008)	
%NSF $Funding_i$	0.028	0.022	0.033	0.027
	(0.028)	(0.028)	(0.028)	(0.028)
Junior=1			-0.290*	-0.314*
			(0.164)	(0.166)
Junior=1 × $\%Allocated_i$			0.382**	0.408**
			(0.189)	(0.191)
Year FE	Yes	Yes	Yes	Yes
Rec Bin FE		Yes		Yes
Lin Rec Ctrl	Yes		Yes	
Dep Var Mean	0.6	0.6	0.6	0.6
R-Sq	0.03	0.06	0.04	0.07
Observations	670	670	651	651

Notes: The unit of observation is a grant. The observation count shown is the count with remaining variation after fixed effects are included. The outcome is the (log) average number of citations received by concepts associated with a grant. *%Allocated* is the share of resources allocated to a grant compared to the amount of resources recommended by the XRAC. *R*1 is an indicator equal to 1 if the PI is affiliated with an R1 institution. *Junior* is an indicator equal to 1 the PI has less than 8 years of experience. *%NSF Funding* is the share of funding coming from the NSF. Columns (1) and (3) controls for the amount recommended with a linear term. Columns (2) and (4) bins the recommended amount into 10 bins of equal size and controls for the amount recommended by including bins fixed-effects. All models include grant start year fixed effects.

Table A10: Share of new concepts – differences by institution and experience

	Resources		Experience	
	(1)	(2)	(3)	(4)
$\%Allocated_i$	0.066	0.068	0.122***	0.129**
	(0.047)	(0.053)	(0.046)	(0.050)
Non-R1 Univ=1	0.061	0.079		
	(0.061)	(0.059)		
Non-R1 Univ=1 × $\%Allocated_i$	-0.010	-0.029		
	(0.073)	(0.070)		
$\ln(Recommended_i)$	0.004		0.001	
	(0.005)		(0.005)	
%NSF $Funding_i$	0.007	0.009	0.007	0.008
	(0.016)	(0.016)	(0.017)	(0.017)
Junior PI=1			0.136***	0.154***
			(0.051)	(0.054)
Junior PI=1 × $\%Allocated_i$			-0.118**	-0.137**
			(0.059)	(0.063)
Field × Year FE	Yes	Yes	Yes	Yes
Rec Bin FE		Yes		Yes
Lin Rec Ctrl	Yes		Yes	
Dep Var Mean	0.6	0.6	0.6	0.6
R-Sq	0.5	0.5	0.5	0.5
Observations	405.0	405.0	405.0	405.0

Notes: The unit of observation is a grant. The outcome is the (log) average number of new concepts associated with a grant, compare XSEDE papers with papers written in the 5 years that precede XSEDE entry. The observation count shown is the count with remaining variation after fixed effects are included. *%Allocated* is the share of resources allocated to a grant compared to the amount of resources recommended by the XRAC. *R*1 is an indicator equal to 1 if the PI is affiliated with an R1 institution. *Junior* is an indicator equal to 1 if the PI has less than 8 years of experience. $\%NSF\ Funding$ is the share of funding coming from the NSF. Columns (1) and (3) controls for the amount recommended with a linear term. Columns (2) and (4) bins the recommended amount into 10 bins of equal size and controls for the amount recommended by including bins fixed-effects. All regressions include grant start year fixed effects, grant field fixed effects and grant start year × grant field fixed effects.

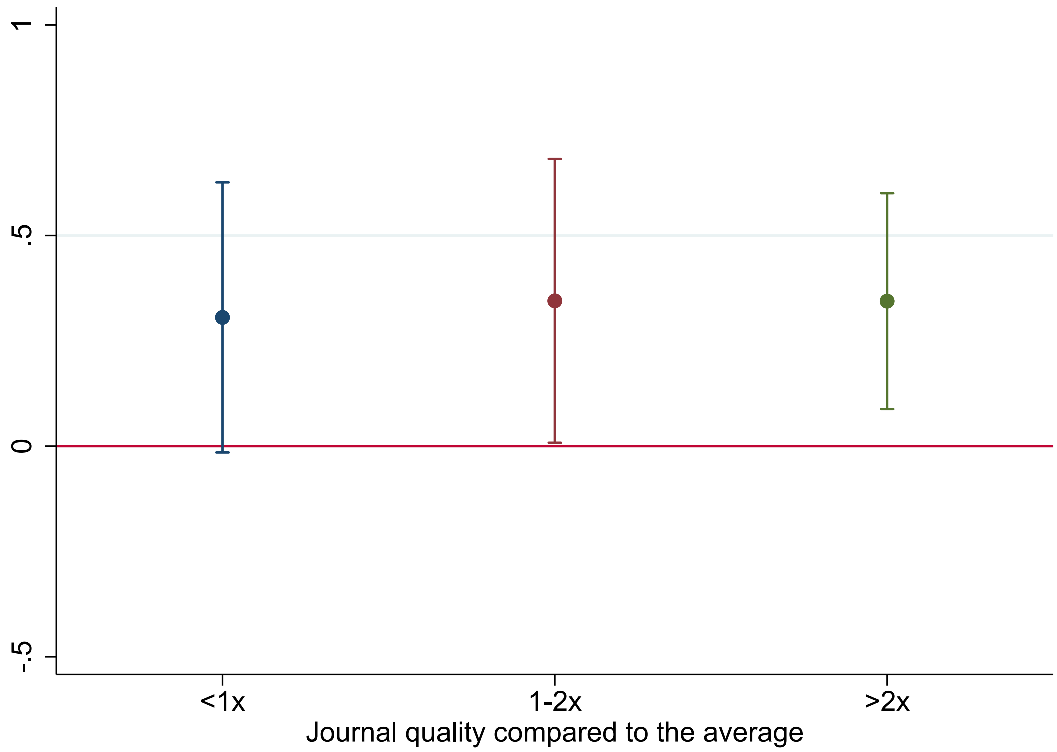
Table A11: Research Breadth – differences by institution and experience

	Resources		Experience	
	(1)	(2)	(3)	(4)
$\%Allocated_i$	0.474**	0.505**	0.211	0.250
	(0.182)	(0.196)	(0.170)	(0.175)
Non-R1 Univ=1	0.399	0.378		
	(0.409)	(0.395)		
Non-R1 Univ=1 × $\%Allocated_i$	-0.529	-0.515		
	(0.438)	(0.425)		
$\ln(Recommended_i)$	-0.021		-0.007	
	(0.021)		(0.021)	
%NSF $Funding_i$	0.045	0.032	0.033	0.023
	(0.049)	(0.052)	(0.045)	(0.049)
Junior=1			-0.620*	-0.615*
			(0.352)	(0.349)
Junior=1 × $\%Allocated_i$			0.707*	0.701*
			(0.404)	(0.405)
Field × Year FE	Yes	Yes	Yes	Yes
Rec Bin FE		Yes		Yes
Lin Rec Ctrl	Yes		Yes	
Dep Var Mean	3.1	3.1	3.1	3.1
R-Sq	0.35	0.36	0.36	0.36
Observations	487	487	470	470

Notes: The unit of observation is a grant. The outcome is the (log) average number of concepts associated with a grant. The observation count shown is the count with remaining variation after fixed effects are included. *%Allocated* is the share of resources allocated to a grant compared to the amount of resources recommended by the XRAC. $R1$ is an indicator equal to 1 if the PI is affiliated with an R1 institution. *Junior* is an indicator equal to 1 if the PI has less than 8 years of experience. $\%NSF\,Funding$ is the share of funding coming from the NSF. Columns (1) and (3) controls for the amount recommended with a linear term. Columns (2) and (4) bins the recommended amount into 10 bins of equal size and controls for the amount recommended by including bins fixed-effects. All regressions include grant start year fixed effects, grant field fixed effects and grant start year × grant field fixed effects.

2 Figures

Figure A1: Number of papers by bins - Journals



We assign each paper in our dataset to one of 3 mutually exclusive categories that represent journal quality. A SNIP value of 1 indicates that the journal has the average number of citations for journals in that field.Papers assigned to the first bin are published in a journal with a SNPI value below 1. Papers assigned to the middle bin are published in a journal with a SNPI value between 1 and 2. Papers assigned to the last bin are published in a journal with a SNPI value above 2. We then calculate for each grant the number of papers that fall into each of these categories and regress the log of this number on our treatment variable, controlling for the amount recommended and grant start year fixed effects. Each dot shows the treatment effect for each bin, with the associated 90% confidence interval.

3 Reconciliation formula

After recommendations have been made by the XRAC for each proposal, the Allocations team meets to "reconcile" recommendations with the resources available. The first phase is called a "balancing" phase, where allocations from over-recommended resources are moved to architecturally equivalent under-subscribed resources until the excess demand is balanced equally across computing platforms. If there is still over-subscription after this phase, the balancing phase is followed by a "reconciliation" phase where recommendations are reduced through a fomulaic solver in order to determine allocations. The formula gives priority to NSF-funded projects.

The reconciliation formula is the following:

$$\text{Award} = [(1 - G) \times R_n + (1 - G) \times F \times R_o] \times Size(R, S) \tag{1}$$

with G the global scaling factor that is solved for by non-linear optimization, R the Recommended amount, R_n the Recommended amount supported by NSF-funding, R_o the Recommended amount non-supported by NSF-funding, F the funding priority factor and S a size-scaling factor (smaller awards are reduced less).

4 Data Construction

Table A12: Correlation between papers retrieved through the internal database and ratio

	(1) Linear Rec	(2) Linear Rec	(3) Binned Rec	(4) Binned Rec
$\%Allocated_i$	0.000	-0.001**	0.000	0.000
	(0.000)	(0.000)	(0.000)	(0.000)
$\ln(Recommended_i)$	0.041***		0.026***	
	(0.015)		(0.009)	
Field × Year FE			Yes	Yes
Rec Bin FE		Yes		Yes
Lin Rec Ctrl	Yes		Yes	
Dep Var Mean	0.2	0.2	0.2	0.2
R-Sq	0.0	0.0	0.3	0.3
Observations	5075.0	5075.0	4980.0	4980.0

5 Concepts

We present two examples to illustrate the use of concepts. The first paper is relatively narrow in scope, with only four concepts. The second paper is broader, with 22 concepts.

Paper 1: Defect-engineered graphene chemical sensors with ultrahigh sensitivity

- Abstract: *We report defect-engineered graphene chemical sensors with ultrahigh sensitivity (e.g., 33% improvement in NO2 sensing and 614% improvement in NH3 sensing). A conventional reactive ion etching system was used to introduce the defects in a controlled manner. The sensitivity of graphene-based chemical sensors increased with increasing defect density until the vacancy-dominant region was reached. In addition, the mechanism of gas sensing was systematically investigated via experiments and density functional theory calculations, which indicated that the vacancy defect is a major contributing factor to the enhanced sensitivity. This study revealed that defect engineering in graphene has significant potential for fabricating ultra-sensitive graphene chemical sensors.*
- https://doi.org/10.1039/C5CP04422G
- Concepts: chemical sensors, conventional reactive ion, density functional theory calculations, ultrahigh sensitivity
- Breadth (as calculated with our measure): 4

Paper 2: Hybrid massively parallel fast sweeping method for static Hamilton–Jacobi equations

- Abstract: *The fast sweeping method is a popular algorithm for solving a variety of static Hamilton–Jacobi equations. Fast sweeping algorithms for parallel computing have been developed, but are severely limited. In this work, we present a multilevel, hybrid parallel algorithm that combines the desirable traits of two distinct parallel methods. The fine and coarse grained components of the algorithm take advantage of heterogeneous computer architecture common in high performance computing facilities. We present the algorithm and demonstrate its effectiveness on a set of example problems including optimal control, dynamic games, and seismic wave propagation. We give results for convergence, parallel scaling, and show state-of-the-art speedup values for the fast sweeping method.*
- https://doi.org/10.1016/j.jcp.2016.06.023

- Concepts: Hamilton-Jacobi equation, algorithm, computer architecture, dynamic game, equations, example problems, fast sweeping method, heterogeneous computer architectures, high performance, hybrid parallel algorithm, optimal control, parallel algorithm, parallel computing, parallel method, parallel scaling, popular algorithms, seismic wave propagation, speedup values, static Hamilton–Jacobi equations, sweeping algorithm, sweeping method, wave propagation
- Breadth (as calculated with our measure): 22

6 Back of the envelope policy implications

In this section, we briefly consider the policy implications of our findings. First, we consider the net return on investing in supercomputing infrastructure. Since we have the ability to place dollar estimates on them, we focus here on the quantity of papers and citations rather than our directional outcomes which must be considered on a case-by-case basis from a policy perspective. Second, we test for heterogeneous productivity effects by some key observable factors.

If we can assign a value to a publication and a citation, we might be able to speak to the trade-off of getting additional computing resources on average. We can call p the value of a publication and c the value of a citation. From our results, a 10% increase in allocated resources is associated with a 38% increase in papers and a 10% decrease in citations, representing 2.4 additional papers and 0.6 fewer citations. The net value of a 10% increase in allocated resources is therefore:

$$2.4 \times p - 0.6 * c * m$$

with m the mean number of citations since our measure demeans citation by field/publication year.[24] To the best of our knowledge, there is no clear estimate of the value of a publication. Myers (2020) suggests a value between \$344k and \$665k, estimated from RFA grants and open grants. We take \$350k as a lower bound. For the value of a citation, we take the estimate of \$3.8k.[25] This implies that giving the average project 10% higher allocation yields a net return on output of \$794k.[26] A 10% higher allocation is equivalent to roughly 2.9M

[24] In our sample, m equals 20.

[25] For this calculation, we use the estimates provided by Boudou (2024). The value of a citation comes from Furman & Stern (2011) which reports a value of \$2,400 in 1996 USD which we adjust to 2018 USD.

[26] Given the current estimates for the additional number of papers and the drop in citations, the net return would become negative if the ratio c/p is higher than 0.2, meaning that the value of a citation is one-fifth or more of the value of a publication – e.g., $\frac{c}{p} > \frac{2.4}{0.6*20} = 0.2$

SUs. The cost of adding an additional SU of capacity to a supercomputer is approximately \$0.15.[27] Under the simple assumption that this additional capacity can be added by expanding within existing computing centers, this suggests that the cost to providing an additional 10% is 435k, resulting in a net return of \$359k.

We also investigate whether some researchers would benefit more than others, which can shed light on the scientific returns as a function of researcher characteristics and yield potential implications for providing allocations as a function of researcher observables. We focus on two sources of heterogeneity: (i) access to other sources of computing power and (ii) researcher experience. We proxy for access to other sources of computing power by creating an indicator equal to 1 if the PI is affiliated with a R1 university and 0 otherwise. Researchers at R1 universities may have access to superior computing clusters on campus, relative to researchers at lower tier institutions, that could potentially serve as partial substitutes for XSEDE resources.[28] We proxy for researcher experience by creating an indicator equal to 1 if the PI is below the median of experience in the sample (13 years) and 0 otherwise. Less experienced PIs have stronger career incentives and may also be operating closer to the frontier in their field.

Table A8 shows the results for the paper outcome.

Columns (1) and (2) explore heterogeneity with respect to access to other resources and find no significant effect – researchers from R1 and non-R1 universities have statistically equivalent returns to computing allocation with respect to quantity of papers. Columns (3) and (4) explore heterogeneity by PI experience. While both senior and junior PI write more paper when getting more resources, the treatment effect for more junior PIs is more than twice as large as the one of more senior researchers (1.27 vs 0.55). This implies that in an environment where computing resources are constrained, providing additional resources to junior researchers yields a higher return on the quantity dimension.

The fact that we find no heterogeneity regarding R1 vs non-R1 universities is consistent with the fact that outside options providing the same type of high-performance computing resources are not easily available, so that researchers from both types of institutions benefit on average. Our findings on junior PIs points to the fact that these researchers are largely contributing to our baseline results, and warrants further inquiry.

[27] https://rescale.com/blog/the-real-cost-of-high-performance-computing/

[28] More specifically, these resources may be able to substitute for a small gap between the allocation requested from XSEDE and the one received.

7 Shape of the production function

R&D managers and policymakers are also interested in whether there is decreasing or increasing marginal returns to computing resources, which informs about whether they should give more computing resources per individual or rather expand the pool of beneficiaries. In what follows, we dig deeper into the linearity assumption we have used so far.

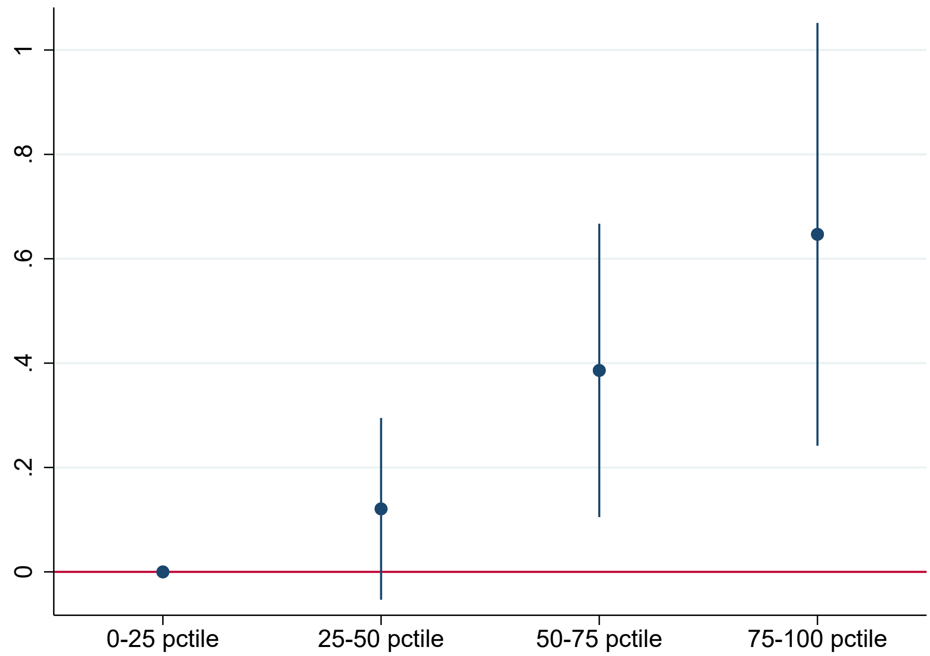
To that end, we first create 8 quantiles of the amount of allocated resources. We then estimate similar regressions with one indicator variable for each quantile in order to flexibly estimate the impact of getting more resources on papers. Figure A2 shows the treatment effect for each bin when the outcome is the log number of papers. Interestingly, we do not find evidence in our sample of strong decreasing nor increasing marginal returns. Rather, the marginal output of papers is roughly linearly related to the allocation of each project. This analysis does not point strongly toward a case of increasing nor decreasing marginal returns. This allows us to interpret our findings at face value, without imposing nuance about the shape of the production function.

8 Additional Mechanisms: Hits & Misses

We explore this mechanism in various ways. First, we calculate for each grant the number of papers that fall into different bins of the citation distribution. This allows us to explore, in a relatively non-parametric manner, where in the citation distribution the largest increase in quantity come from. For each grant, we calculate the number of papers that fall into 10 mutually exclusive bins: papers that receive less than half the expected average number of citations given their publication year and field, papers that receive between half and 1 times the expected average number of citations etc. Bins are constructed so that they are roughly of the same size with 10% of the observations. The bins are represented in the x-axis of Figure A3. We then run 10 regressions (one for each bin) with the log number of papers as an outcome, controlling for the amount recommended and grant start year fixed effects. Figure A3 shows the coefficient on the treatment variable for each bin. A higher percentage of allocated resources leads to an increase across the distribution of citation, except for extremely cited papers (more than 10 times the average) and papers cited between 2 and 3 times the average[29]. In addition, the biggest increase is concentrated among papers with less than twice

[29] Note that the figure suggests a lower increase in the number of papers than what we found in Table 4. This is due to the fact that we use a different set of fixed effects - the table includes grant start year, field and grant start year × field fixed effects while the figure includes only start year fixed effects.

Figure A2: Returns to scale - Papers



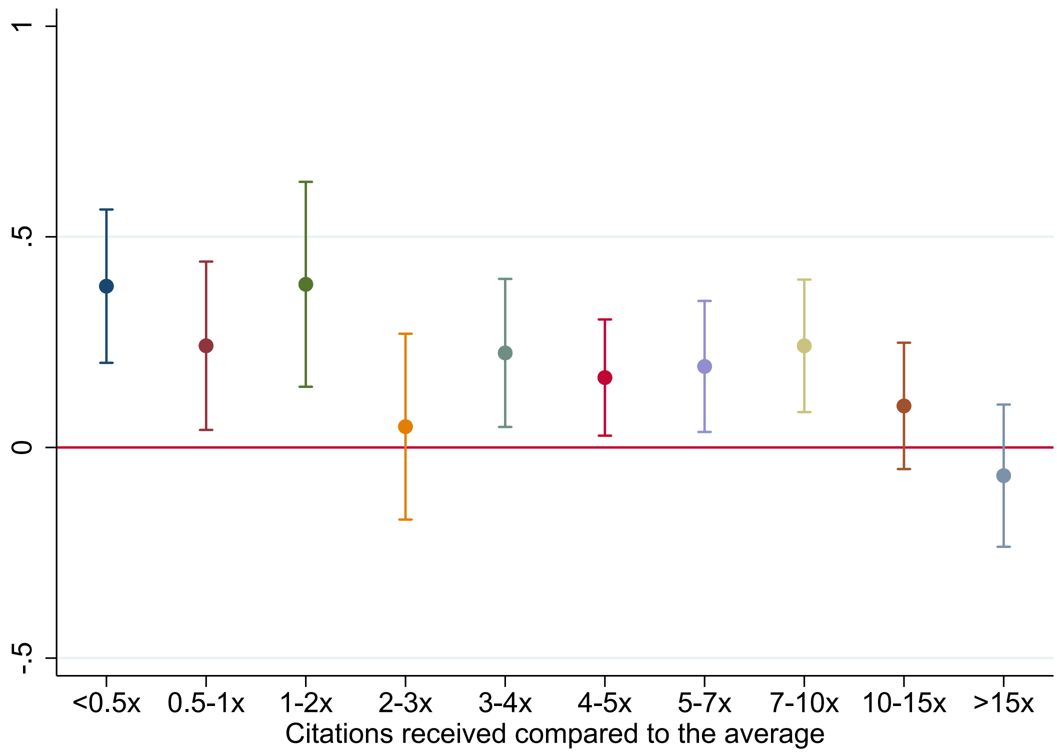
Notes: We create 8 quantiles for the amount of computing resources allocated to researchers. Each dot shows the coefficient of a regression of the (log) number of papers associated with a grant on 8 indicators, one per quantile. We control for the amount recommended with a linear term. We also control for the share of funding coming from the NSF, grant start year fixed effects, grant field fixed effects and grant start year × grant field fixed effects. Robust standard errors clustered at the grant start year × field of science level. The bars represent the 90% confidence interval.

the expected number of citations, and we do not observe a similar magnitude for highly cited papers.

We repeat the same exercise by creating bins based on the citation distribution considering *papers in our sample only*. Figure A4 in Appendix shows the treatment effect for each bin. We find similar results: a higher percentage of allocated resources is associated with a higher number of papers across the distribution, except for papers with a number of citations falling between the 50^{th} and the 75^{th} percentile.

Table A13 shows regressions where outcomes are successively the minimum number of citations received by a grant, the maximum number of citations received by a grant and the standard deviation of citations received by a grant. A higher share of allocated resources is associated with a decrease in the minimum number of citations, but we do not see any impact on the maximum number of citations nor on the standard deviation of citations.

Figure A3: Log(1+Number of papers) - Citations bins



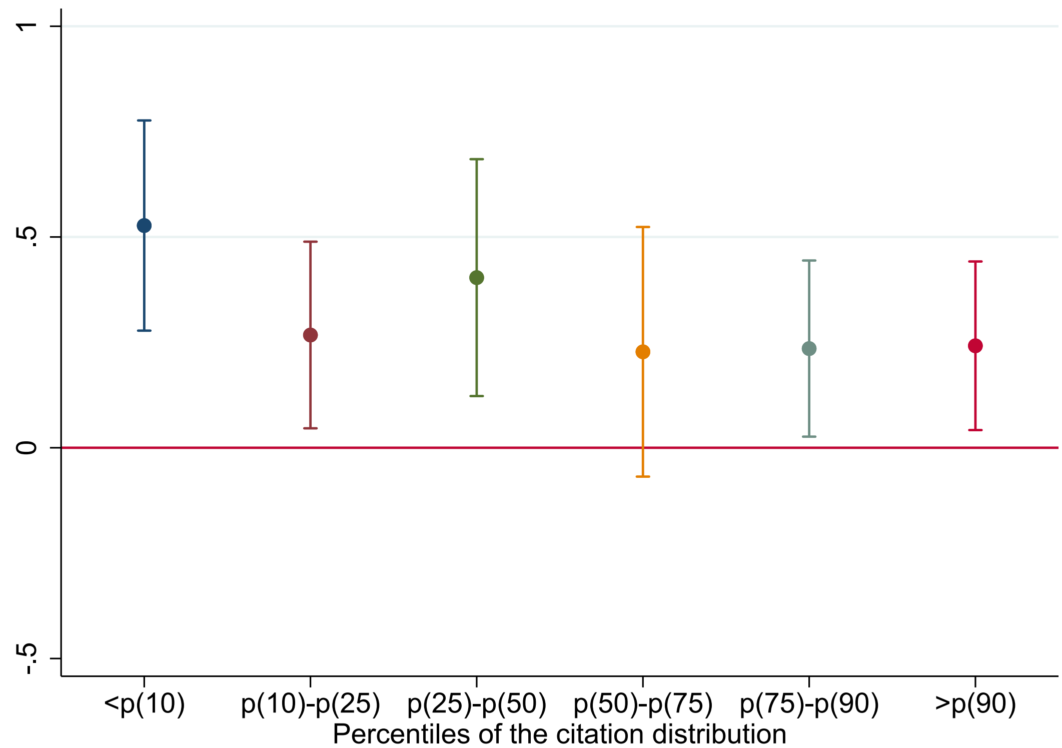
We assign each paper in our dataset to one of 10 mutually exclusive categories that represent the average number of citation it received compared to papers published the same year and in the same field. We then calculate for each grant the number of papers that fall into each of these categories and regress the log of this number on our treatment variable, controlling for the amount recommended and grant start year fixed effects. Each dot shows the treatment effect for each bin, with the associated 90% confidence interval.

Table A13: Minimum, maximum and standard deviation of citations - Dimensions

	Log(1+Min)		Log(1+Max)		Log(1+sd)	
	(1)	(2)	(3)	(4)	(5)	(6)
$\%Allocated_i$	-0.630***	-0.649***	-0.031	-0.022	-0.280	-0.269
	(0.225)	(0.214)	(0.269)	(0.248)	(0.296)	(0.275)
$\%\text{NSF } Funding_i$	0.028	0.029	0.015	0.016	0.005	0.003
	(0.082)	(0.083)	(0.092)	(0.091)	(0.086)	(0.085)
$\ln(Recommended_i)$		-0.016		0.202***		0.103***
		(0.024)		(0.023)		(0.023)
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Rec Bin FE	Yes		Yes		Yes	
Lin Rec Ctrl		Yes		Yes		Yes
Dep Var Mean	0.9	0.9	2.1	2.1	1.4	1.4
R-Sq	0.04	0.03	0.11	0.11	0.07	0.07
Observations	622	622	622	622	462	462

Notes: The unit of observation is a grant. The outcome is the (log) minimum number of citations associated with a grant, the maximum number of citations associated with a grant and the standard deviation of citations associated with a grant. *%Allocated* is the share of resources allocated to a grant compared to the amount of resources recommended by the XRAC. *%NSF Funding* is the share of funding coming from the NSF. Columns (1), (3) and (5) control for the amount recommended with 10 bins of equal size. Columns (2), (4) and (6) control for the amount recommended with a linear term. All columns include grant start year fixed effects.

Figure A4: Log(1+Number of papers) - Citations bins (within sample)



We consider only papers that are part of our sample. We create 6 mutually exclusive categories and attribute each paper to one of these bins based on where they fall in the citation distribution compared to other papers published the same year and in the same field. We then calculate for each grant the number of papers that fall into each of these categories and regress the log of this number on our treatment variable, controlling for the amount recommended and start year fixed effects. Each dot shows the treatment effect for each bin, with the associated 90% confidence interval.