Returns to Science of Centrally Provisioned GPUs and Optimal Allocation

Connie Xu Ruby Zhang*
(Job Market Paper)

November 15, 2026

[Click Here for the Latest Version]

Abstract

Funders of research must consider where the marginal dollar is most productive and where it is least likely to displace the recipient's own investment. We study the allocation of federal computing grants to U.S. universities by estimating the returns to science of university-level computing infrastructure and modeling the crowd-out of universities' own investment. We construct an original dataset measuring Central Processing Unit (CPU) and Graphics Processing Unit (GPU) capacity over time and link it to research publications at U.S. R1 universities. Exploiting the large and discrete nature of GPU capacity increases to identify their effects, we find that AI and deep-learning publications increase more at less-resourced universities. We then estimate a structural production function where multiple scientific fields share a central computing cluster, and derive universities' investment responses to external grants. Compared to a publication-maximizing benchmark, the observed portfolio of National Science Foundation (NSF) cyberinfrastructure grants captures 39% of the publication gain attainable with the same budget. The publication-maximizing allocation shifts dollars from universities with medium-sized endowments toward both tails of the endowment distribution. However, the NSF's observed allocation is more geographically dispersed than the publication-maximizing allocation, highlighting tradeoffs in balancing efficiency versus equity.

Keywords: Computing, GPU, Multi-Product Production, Grant Allocation

*(**Corresponding Author**) **Zhang:** Harvard University. Email: rzhang15@g.harvard.edu. **Xu:** Harvard University. Email: conniexu@g.harvard.edu. We are deeply grateful to Shane Greenstein, Robin Lee, Ariel Pakes, and Jesse Shapiro for their guidance and support throughout this project. We thank Andrea Bonaccorsi, Allan Collard-Wexler, Felipe del Canto Monge, Michael Droste, Josh Feng, Avi Goldfarb, Jeff Gortmaker, Shao-Yu Jheng, Myrto Kalouptsidi, Andrew Kao, Kyle Myers, Lorenzo Palladini, Michael Rubens, Jungkyu (JK) Suh, Mariela Szejnfeld, Chloe Tanaka, Benjamin Vatter, Thomas Wollmann, Ryan Xia, Yixin Zhou and participants in the Harvard IO Seminar, Harvard Contracts Workshop, NBER Productivity Seminar, as well as DRUID, EARIE, and REER for valuable discussion and feedback. We gratefully acknowledge funding from the Lab for Economic Applications and Policy Award, the John and Kiendl Gordon Economics Support Fund, the Institute for Humane Studies (IHS020076; Zhang), the National Institute on Aging (T32-AG000186; Xu), and the National Science Foundation (DGE #2140743; Xu). We are grateful to Anton Tarasenko for outstanding research assistance. The computations in this paper were run on the FASRC Cannon cluster supported by the FAS Division of Science Research Computing Group at Harvard University.

1 Introduction

Economists have long recognized theoretical reasons why the private incentive to supply scientific research is lower than the social benefit (Nelson 1959, Arrow 1962). However, that does not provide guidance on how to allocate fixed federal budgets across multiple potential uses and users. One might think it would be simple to allocate to the researchers with the highest productivity. That intuition fails because researchers work within a university, and universities vary in their ability and willingness to fund research already. The funds might simply replace what the university would have funded on its own, reducing the incremental benefit of each dollar allocated.

These issues are especially salient in capital-intensive science, which depends on university-level infrastructure (Cockburn et al. 2019, Suh 2026), thereby raising the stakes of how and where expensive infrastructure is financed and allocated across universities. Among these capital inputs, graphics processing units (GPUs) have become essential not only in frontier artificial intelligence (AI) and deep-learning research but also in computational workflows across biology, chemistry, engineering, and the physical sciences. Therefore, universities increasingly operate expensive, on-premises high-performance computing (HPC) centers with GPU clusters. Unequal access to this infrastructure may in turn reinforce the "compute divide" between universities with and without the financial capacity to fund computing capital (Ahmed and Wahed 2020, Ahmed et al. 2023, Besiroglu et al. 2024).

The National Science Foundation's (NSF) Major Research Instrumentation (MRI) and Campus Cyberinfrastructure (CC*) programs are the major sources of external financing for university-level computing infrastructure, and they face this tradeoff directly. A well-resourced university may have more productive researchers, but it can also pay for a cluster from its own funds or turn to outside options such as cloud computing and national facilities, so a grant may largely replace spending that would have happened anyway. A less-resourced university may produce fewer publications from the same hardware, but it is less able to finance a cluster on its own, so a grant is more likely to add computing capacity that would not otherwise exist.

How would a funder with the goal of maximizing research output allocate a fixed computing budget across universities that differ in both productivity and financing constraints? First, we need to know the return to computing capital, defined as additional publications produced by additional CPUs and GPUs. Second, we need to know crowd-out, defined as the extent to which a university cuts its own computing investment when it receives a grant. The first object comes from a structural

production function for science publications, and the second object is derived from the same model. Together they deliver the allocation that maximizes publications from the existing NSF budget. This benchmark is informative about scientific output, but the NSF may have other goals, such as distributing resources equitably across institutions and states. The gap between the benchmark and the observed allocation measures how many publications the agency trades off in pursuing alternative goals.

The analysis rests on a new panel of computing capacity at 53 U.S. R1 universities from 2010 to 2025, which we construct using Wayback Machine snapshots of university research-computing websites. The panel records installed CPU and GPU hardware and is linked to publication and author histories from OpenAlex. Capacity is harmonized across generations of GPU hardware using a common measure of peak computing speed,[1] and papers are classified by their use of computational methods. The resulting dataset connects university-level computing capital to science publications over time and across STEM fields.

Three facts in the data motivate the model: less-endowed universities rely more on NSF funding for their clusters, publications and research labor both rise after large GPU upgrades, and universities buy less hardware of their own after receiving an NSF award. First, endowment per faculty, our proxy for financial resources, is correlated with reliance on external NSF funding for HPC clusters. Ordering universities by endowment per faculty, the average share of GPU capacity in NSF-funded clusters falls from 45% in the least-endowed fifth of our sample to 10% in the best-endowed fifth.[2] Second, large GPU upgrades are followed by more AI and deep-learning publications at less-resourced universities. The comparison uses large, discrete increases in GPU capacity in a staggered difference-in-differences design, estimated separately for universities below and above the median of endowment per faculty across all R1 universities. AI and deep-learning publications rise by about 10.6 percent at less-resourced universities and do not rise at more-resourced ones. Research labor also responds, though this varies by field. In computer science, the number of graduate students and postdocs rises after an upgrade while the faculty stock does not. Because output and inputs move together, the event studies cannot separate the contribution of the computing capital itself from that of the labor that arrives with it. Third, NSF infrastructure awards are followed by lower spending by

[1] The measure is FP16 TFLOPS: trillions of 16-bit ("half-precision") floating-point operations per second. It uses dense tensor-core throughput, a benchmark for deep-learning workloads. Section 3 describes the treatment of older hardware and alternative capacity measures.

[2] The rank correlation between endowment per faculty and total endowment across the 53 panel universities is 0.87, so the proxy orders universities much as total endowment would.

universities on their own computing hardware. In the year after an award, own hardware spending falls by about $0.54 million against an average award of $1.01 million. Awards are few and not randomly assigned, so this evidence is suggestive, and it motivates deriving crowd-out from the model.

To separate the contribution of capital from that of labor, and to evaluate counterfactual allocations, we model the university as a multi-product organization. The university maximizes publications across five scientific fields, from computer science to the life sciences, which share one central cluster, in the spirit of De Loecker et al. (2016). In each field, publications are produced from computing capital and research labor with a Cobb–Douglas technology and an unobserved productivity term. Capital combines CPUs and GPUs, and labor combines faculty with PhD students and postdocs, each with its own elasticity of substitution. These CES parameters are calibrated or estimated from wage bills, hardware purchases, and department-level usage data. Two obstacles arise in estimating the output elasticities. First, computing capacity is observed for the university as a whole and not by field. The university's first-order conditions imply how it divides the cluster across fields, which gives field-level capital as a function of observed data and the model's parameters. Second, universities invest more when they expect to be more productive. Productivity is modeled as a permanent university-by-field component plus a persistent shock, and identification relies on timing: capital and faculty are chosen before the current year's productivity shock is realized, so lagged inputs serve as instruments in a dynamic panel estimator (Arellano and Bond 1991, Blundell and Bond 1998). Because the field-level allocations depend on the elasticities being estimated, estimation iterates between the two until they agree.

The estimated capital elasticities are small but significant, ranging from 0.03 in engineering to 0.06 in computer science. Intuitively, this means doubling a university's computing capital raises its GPU-related publications by three to four percent. All fields exhibit decreasing returns to scale, so the marginal product of capital falls as a university's computing stock grows, which favors spreading grants across universities. Furthermore, estimated productivity is higher at more-endowed universities in all fields except engineering. The universities that use computing capital more productively are therefore often the ones able to self-finance, which is the tradeoff the NSF faces. The same parameters imply how a university responds to a grant. A university that is already buying computing hardware treats a grant like cash: it cuts its own hardware purchases and redirects the money, so only about ten cents of each grant dollar ends up as additional computing capital. A university that is not buying hardware of its own has no purchases to cut, so the entire grant becomes new capital. About

one-fifth of university-years are in this second group, and they are concentrated among less-endowed universities. Across R1 universities, a grant dollar displaces 72 cents of own spending on average. Hardware purchases around NSF awards are consistent with this prediction: the estimates are noisy and sensitive to how awards are dated, but in the baseline they reject full additionality and do not reject the model's implied crowd-out.

We use these elasticity and crowd-out estimates to simulate the reallocation of $30.5 million in FY2021–2025 NSF computing awards (MRI and CC*) received by universities in the panel. The benchmark maximizes total publications in GPU-related research areas across all fields: each dollar goes to the university where it produces the most additional papers after accounting for crowd-out. The observed NSF allocation captures 39% of the publication gain attainable with the same budget. The publication-maximizing allocation moves dollars from the middle of the endowment distribution toward both tails: universities between the 25th and 75th percentiles of endowment per faculty receive 74% of the observed budget but only 25% of the publication-maximizing allocation. Dollars move toward the bottom of the endowment per faculty distribution because less-endowed universities have small computing stocks and are more often not buying hardware of their own, so grants there add capacity, and toward the top because estimated productivity is high there.

The observed allocation is also far more geographically dispersed than the efficient one. Universities in states classified as historically underfunded under the Established Program to Stimulate Competitive Research (EPSCoR) receive 43% of the observed budget but 10% of the publication-maximizing allocation. A state currently qualifies when it receives less than 0.75 percent of total NSF research funding over the last five years. A proposal from 2017 would instead base eligibility on research funding per capita (Mervis 2017). The proposal does not move dollars in our exercise and instead changes which states count as underfunded. Under the per-capita definition, the EPSCoR share of the publication-maximizing allocation rises to 35%, because the newly eligible states are ones where the benchmark also allocates funding, while the observed share is essentially unchanged at 44%. Most of the tension between geographic dispersion and publication maximization disappears under the proposed rule.

There are two caveats to these findings. First, the benchmark computes gains based on changes only in capital and assumes labor and any freed-up funds to be fixed, abstracting away from researcher mobility between universities. The results understate the returns to external grants at universities that redirect funds towards labor. Second, the benchmark itself is static and assumes short-run

publication maximization. The NSF may have dynamic considerations beyond total research output since building up distributed scientific capacity generates local economic spillovers (Hausman 2022), trains the future scientific workforce (Gofman and Jin 2024), and increases the diversity of research directions (Hofstra et al. 2020). Even so, the comparison measures what these other objectives cost in short-run publications: the observed allocation produces 39% of the gain that the same budget could deliver.

Related Literature. This paper connects to four strands of literature spanning the returns to scientific inputs, concentration of computing capacity, structural production function estimation, and optimal resource allocation. First, work on the production of scientific knowledge has studied the impact of laboratory space, specialized equipment, rising capital intensity, undergraduate labor, and research grants on science (Waldinger 2016, Baruffaldi and Gaessler 2026, Suh 2026, Jacob and Lefgren 2011, Ganguli and Murciano-Goroff 2025, Azoulay et al. 2019). McKeon and Boudou (2026) use quasi-random allocations in NSF's national supercomputing program XSEDE (the predecessor of ACCESS) to estimate the intensive margin effect of supercomputing for individual scientists. Closest to our setting, Apon et al. (2015) compare the publication returns by field at universities with and without a Top500 HPC cluster.[3] The authors use 2005–2006 National Research Council data on departments in eight disciplines at over 200 U.S. universities, finding higher average returns at universities with local HPC in chemistry, civil engineering, physics, and history but not in computer science or economics. Smith et al. (2025) estimate a Cobb–Douglas-style production function of university-level publications, doctorates, and R&D expenditures in capital (TFLOPS of centrally operated computing) and labor (research computing center salaries) at five R1 universities over 2011–2023, and find positive returns to both inputs. This paper builds on research regarding returns to on-premises computing by using a decade-long computing capacity panel across 53 universities, and asks how differences across universities in estimated productivity and own investment affect grant allocation.

Second, research on the concentration of AI capacity documents the declining representation of non-elite universities and academic-only teams in compute-intensive research (Ahmed and Wahed 2020, Ahmed et al. 2023, Besiroglu et al. 2024), and the consequences of faculty moving to industry (Gofman and Jin 2024, Yue 2024). Related research shows that access to shared biological materials,

[3]The Top500 is a semiannual ranking of the world's 500 fastest supercomputers measured by theoretical peak performance. The list was dominated by U.S. universities and government facilities until the mid-2010s.

data, and research tools shapes who participates in certain kinds of science (Furman and Stern 2011, Nagaraj et al. 2020, Teodoridis 2018). This paper examines shared computing capital, whose upfront costs and economies of scale make university resources central to access for scientists.

Third, we adapt multi-product production function methods to a university's shared research infrastructure. As in De Loecker et al. (2016), aggregate inputs must be allocated across products whose outputs are observed separately. We combine static first-order conditions with dynamic panel moments (Arellano and Bond 1991, Blundell and Bond 1998, Rubens et al. 2026) to recover field-level capital allocations and output elasticities using a fixed-point procedure, and use the same parameters to derive investment responses to external funding. Unlike most production function estimation—which observes revenue rather than output—we observe both output (publications) and inputs (capital and labor) in physical units. We are able to estimate a quantity production function, and the productivity residual is not contaminated by demand shocks and markups (Foster et al. 2008, De Loecker and Goldberg 2014).

Finally, we build on work on the crowd-out and allocation of public research funding. Studies of R&D subsidies examine whether public money displaces private spending and reach mixed answers (David et al. 2000, Wallsten 2000, Lach 2002). At universities, federal research funding has led to crowd-in of private donations and state support (Payne 2001, Lanahan et al. 2016). The broader literature on misallocation provides the publication-maximizing benchmark of equalized marginal returns (Hsieh and Klenow 2009, Restuccia and Rogerson 2008). In our setting, we first derive crowd-out of the university's own investment from the model and test whether it is rejected in the data, then compute the publication-maximizing benchmark in the context of the NSF's allocation across universities.

Section 2 describes the institutional setting and Section 3 the data. Section 4 presents the empirical facts, including the event-study estimates. Section 5 models the university's problem and science production, with results presented in Section 6. Section 7 derives and tests crowd-out, Section 8 evaluates counterfactual NSF allocations, and Section 9 concludes.

2 High-Performance Computing at Universities

Universities centrally maintain high-performance computing (HPC) clusters, which are collections of servers whose interconnected computers run tasks in parallel and solve problems too large for a

single desktop computer. GPUs have become increasingly important alongside traditional CPUs. CPUs handle general-purpose and serial tasks while GPUs contain thousands of cores suited to parallel operations, including deep learning and scientific simulation. GPUs are expensive, and each generation has cost more than the last. Epoch AI puts the release price of NVIDIA's datacenter GPUs at $16,000 for the 2016 generation and $44,000 for the 2022 generation (Rahman 2024). So equipping a shared cluster with even a few dozen GPU servers is a multi-million-dollar investment. In addition to the cost of racks, power, and cooling, GPU hardware is also upgraded frequently, typically every five years due to depreciation and warranty expiration.

Universities provide access under three arrangements. A centrally funded cluster, supported by indirect costs or external infrastructure grants, gives faculty free access up to a quota. Under the *condo* model, faculty buy nodes with their own grants and the university hosts and maintains them. Owners receive priority access, while idle nodes serve other researchers. Under the *recharge* model, a university charges hourly rates to recover the cost of operating a service center or core facility,[4] reducing the reliance on indirect-cost funding. Most universities operate under a combination of a free quota alongside a condo model for faculty who wish to buy equipment.

On-premises clusters remain the major source of university research computing despite outside options. In practice, researchers can alternatively buy their own servers, purchase commercial cloud services, or obtain time at national facilities through NSF ACCESS or the Department of Energy (DOE). However, cloud services can be costly for sustained workloads and large datasets, while national facilities provide large-scale computing for the tail end of scientists running massive simulations.[5] In a 2021 Coalition for Academic Scientific Computing survey, on-premises clusters remained the most significant resource for many universities (Figure 1).

[4]The Great Lakes cluster at the University of Michigan is a prominent example.

[5]In 2011, the DOE's Magellan project found its computing centers typically 3 to 7 times less expensive than equivalent commercial cloud capacity (U.S. Department of Energy, Office of Science 2011). Commercial cloud could be cheaper with sporadic demand or facility capacity constraints. Emeras et al. (2019) reach a similar conclusion in a 2019 comparison of Amazon EC2 and an in-house HPC cluster.

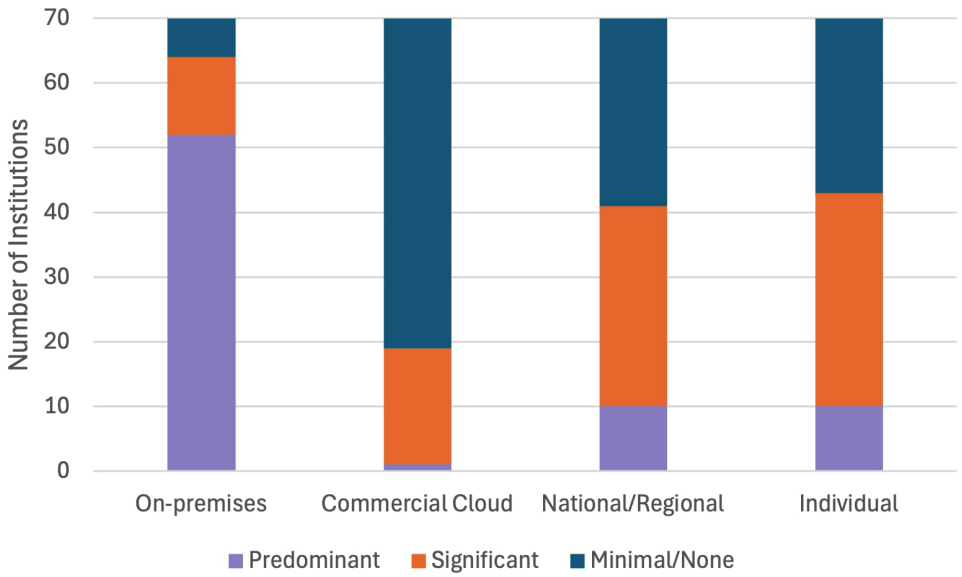


Figure 1: Survey from the Coalition for Academic Scientific Computing on Most Significant Computational Resource (2021). Bars report the number of surveyed institutions rating each type of computational resource as a Predominant, Significant, or Minimal/None part of their computing. There are four types of computational resource: on-premises clusters, commercial cloud, national/regional facilities, and individually owned hardware.

2.1 External Funding for Research Computing

The NSF supports both national computing facilities and university-owned clusters. Its Office of Advanced Cyberinfrastructure (OAC), within the Directorate for Computer and Information Science and Engineering (CISE), funds computing, data, software, and networking resources. CISE estimates that it provided 87% of federal funding for fundamental computer science research at U.S. academic institutions as of 2022 (National Science Foundation 2021b). Much of OAC's investment supports national systems accessed through programs such as ACCESS. We study clusters owned and operated on-premises by individual universities.

The NSF explicitly states broadening access to research computing as a goal. In its 2021 cyberinfrastructure blueprint, the NSF calls for "democratizing availability of and access to" cyberinfrastructure (National Science Foundation, Office of Advanced Cyberinfrastructure 2021). The ACCESS program's stated goal is "ensuring democratized and equitable access" to the national ecosystem (National Science Foundation 2022). In January 2024, the NSF launched the National Artificial Intelligence Research Resource (NAIRR) pilot to democratize access to AI-scale computing, data, and models (NAIRR Task Force 2023, National Science Foundation 2024b).

NSF Major Research Instrumentation (MRI) and Campus Cyberinfrastructure (CC*) are

the main federal programs that finance campus clusters. MRI is an NSF-wide program, run by the Office of Integrative Activities, that has funded shared research instruments of every kind since 1997; a computing cluster is one instrument type among many and competes with microscopes and spectrometers for the same budget. CC* is an OAC program that began in 2012 by funding campus network upgrades and added a campus computing area in 2019 (National Science Foundation 2012, 2019). We describe the features that matter for our analysis: who can apply and for how much, how awards are selected, and what the money may buy.

Any accredited U.S. college or university may apply to either program, though MRI limits each university to three proposals a year across all instrument types and caps requests at $4 million, so a cluster must first win an internal competition against the university's other instrument needs (National Science Foundation 2015, 2018, 2023). CC* sets no limit on proposals but caps a campus computing award at $400,000 in the FY2019–2022 solicitations, $500,000 in FY2023, and $700,000 in FY2024–2025. Cost sharing differs sharply between the programs. PhD-granting institutions must pay exactly 30 percent of the total project cost of an MRI award from non-federal sources, and the CHIPS and Science Act of 2022 waived the requirement for five years beginning with the FY2023 competition (CHIPS and Science Act 2022, National Science Foundation 2023).

Both programs select awards through the NSF's merit review, with program-specific criteria layered on intellectual merit and broader impacts. An MRI proposal must show that multiple research groups will share the instrument and that the university commits to operating and maintaining it (National Science Foundation 2018, 2023). Approximately 20 to 25 percent of MRI proposals are funded (Phelps 2019). A CC* computing proposal must present a campus-wide cyberinfrastructure plan, a table of the research projects that will use the cluster and their computing requirements, vendor quotes, and a plan for sustaining operations. Proposals serving a single discipline are not funded. The solicitations encourage proposals from under-resourced universities and, in FY2024, from minority-serving institutions and institutions in historically underfunded states (National Science Foundation 2021a, 2024a).

Both programs buy hardware that the university then operates as shared campus-wide infrastructure, which is the capacity our data measure. MRI requires at least 70 percent of an acquisition budget to be equipment, excluding instruments used primarily for teaching, but it does not require access for users outside the university (National Science Foundation 2023). CC* budgets are likewise "expected to be dominated by equipment," and a CC*-funded cluster must commit at least 20

percent of its time to researchers outside the campus, for example by joining the Open Science Grid (National Science Foundation 2021a).

These grants are discrete and substantial, often funding the purchase of an entire new cluster or a significant partition of GPUs. Since 1997, the MRI program has made 380 awards for HPC clusters totaling $243 million. Of these, 102 fund GPU clusters, the first in 2008 (Figure 2). Awards have also grown: the median HPC award rose from $0.40 million in 1997–2001 to $0.61 million in 2020–2024, in 2017 dollars.

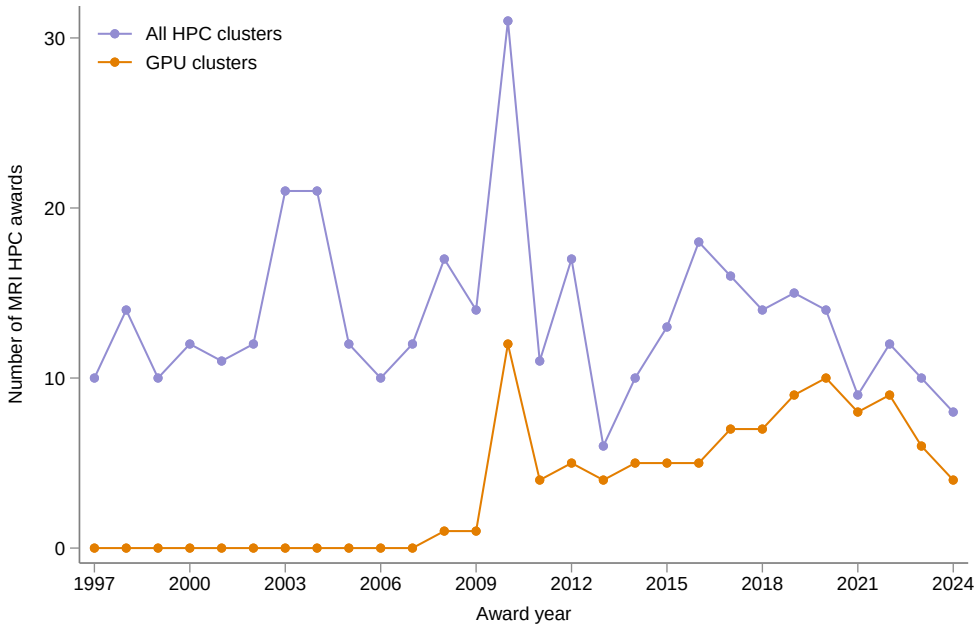


Figure 2: NSF Major Research Instrumentation Awards for High-Performance Computing Clusters, 1997–2024. Annual number of MRI awards for HPC clusters and the subset that fund GPU clusters, identified from award-abstract text. The 2010 spike reflects American Recovery and Reinvestment Act funding.

3 Data

In order to study the effects of university computing capacity on research, we construct a new panel of computing capacity at 53 universities and link it to data on publication outcomes and research labor. We build a panel at the university × field × year level from 2010 to 2025 across five major STEM fields: computer science, engineering, geosciences and physics, molecular and materials sciences, and life sciences. Data construction details are in Appendix A1.

3.1 Data Sources

We compile four kinds of data: computing capital, scientific output, research labor, and university characteristics such as endowment or total R&D.

3.1.1 University Computing Capacity Panel

We measure the CPU cores and GPUs installed in each university's shared cluster from archived hardware specifications on its research computing website. Annual Wayback Machine snapshots record CPU cores, GPU models and counts, access prices, and funding sources from 2010 or the earliest available year. The panel covers 53 R1 universities whose hardware histories we have manually verified. For ten universities, computing center staff also confirm the timing of major upgrades. Three research computing centers generously provided CPU and GPU core-hours by department,[6] which we use to calibrate the relative compute intensity of fields between CPU and GPU.

Harmonizing GPU Capacity. We put different generations of GPUs on a common performance scale by summing the reported top calculation speed. Our main performance measure is TFLOPS: trillions of floating-point operations per second. A floating-point operation (FLOP) is a single mathematical calculation, such as addition or multiplication, performed using decimal numbers. Floating-point arithmetic also uses different precisions. For example, scientific simulation often requires 64-bit operations, while neural-network training can use 16-bit operations. Our headline measure is dense FP16 tensor-core throughput, which captures capacity suited to deep learning. Cards without tensor cores contribute zero to this measure, so we also construct FP32 TFLOPS (32-bit floating-point operations), which covers every GPU model, and a dollar stock valued at acquisition-year prices. Specifications come from the Epoch AI Machine Learning Hardware database (Epoch AI 2026) and TechPowerUp (TechPowerUp 2026), cross-checked against manufacturer datasheets (NVIDIA Corporation 2017, 2021, Advanced Micro Devices 2021).

Hardware Price Ladder. We obtain prices for CPU and GPU hardware from university purchase orders in GovSpend, a database of public-records requests to state and local entities that covers 34 of the 53 panel universities. Appendix A1.4 describes the purchase-order prices and hardware reference table. Across 463 datacenter-GPU purchase lines at 133 public universities, the median price is $37

[6] These are the University of Pittsburgh, University of Vermont, and Ohio State University.

per FP16 TFLOPS, falling from $56 in 2018 to $17 in 2025 (Appendix Table A.3). Prices vary across GPU models and years rather than across buyers, with year and model effects explaining 52% of the variance of log price per TFLOPS, so a price ladder by model and year applies to the panel universities that GovSpend does not cover.

3.1.2 Scientific Output

We measure output from OpenAlex, an open index of scholarly works. We pull every article and preprint published 2005–2025 with at least one author affiliated with an R1 university. Each paper entry contains the title, abstract, authors, affiliations, OpenAlex topics, and citations. A paper's credit is divided equally across its authors and, within an author, across their listed affiliations. We use raw paper counts or a citation-weighted percentile within document type, publication year, and subfield. This measure adjusts for citation norms that differ across fields and for the shorter time recent papers have had to accumulate citations.

We classify papers by the computational methods their top three topics typically use. A large language model assigns each of OpenAlex's 4,516 topics to computational, experimental, theoretical, or mixed research, then classifies computational and mixed topics as GPU-heavy, CPU-heavy, or both. A paper is *computational* if any topic is computational or mixed, and *GPU-topic* if any topic is GPU-heavy or both. These labels describe the type of research, not the machine on which it ran. The event study uses this GPU-topic outcome and a separate keyword search of AI and deep-learning concepts in the abstract.[7] The production function baseline uses the same GPU-topic count to capture overall change in GPU-relevant research. Appendix A1.2 gives the classification rules and alternative credit measures.

3.1.3 Research Labor

The production function distinguishes two kinds of researchers. Flexible labor is PhD students and postdocs, which the university adjusts year to year. We measure it from the NSF Survey of Graduate Students and Postdoctorates in Science and Engineering (GSS). Because the GSS reports doctoral enrollment only from 2017, we extend the series backward using PhD graduates by university, field, and year from the ProQuest dissertation panel constructed by Shvadron et al. (2025).

Quasi-fixed labor is faculty and permanent research staff. No administrative count by field

[7] Only 70% of papers have a downloadable abstract.

exists for all R1 universities, so we construct it from the OpenAlex author records. We assign each author a main institution in each year and classify an author-year as quasi-fixed when it belongs to a spell of at least four years at that institution that began at least three years into the author's publishing career, or a spell of at least eight years. The first condition identifies persistence while the second separates hires from trainees, whose careers typically begin where they train. We count an author in every year of their spell whether or not they publish in that year, to reduce selection on publishing that year, and we adjust for spells that are censored at the end of the panel. Validated against the American Institute of Physics faculty survey, the measure ranks physics and astronomy departments across universities with a correlation of 0.93 and the within-university correlation is 0.31 over six-year changes (Appendix A1.3).

3.1.4 University Characteristics

University characteristics come from IPEDS, the NSF Higher Education Research and Development (HERD) Survey, and the Center for Measuring University Performance. IPEDS supplies faculty counts and financial data only at the university level, while NSF HERD supplies R&D expenditure by funding source at the field level. We use institutionally financed R&D to test model-implied crowd-out in Section 7. Endowment values use IPEDS supplemented with the Center for Measuring University Performance series where needed. We also collect publicly available NSF MRI and CC* awards and identify computing awards from their abstracts.

3.2 Summary Statistics

The panel resembles the R1 universe in median research spending and endowment, but underrepresents its wealthiest universities. Table 1 compares the 53 panel universities with all 187 Carnegie R1 universities. In 2024, median STEM R&D spending is $230 million in the panel and $226 million among all R1 universities, and median endowments are $958 million and $947 million, respectively (2017 dollars). The panel's endowment interquartile range is narrower, and median endowment per tenure-line faculty is $1,026 thousand, compared with $1,296 thousand across all R1 universities.

The panel is unbalanced because universities enter with their first available archived cluster snapshot. The median university is observed for 11 years between 2010 and 2025. In 2024, the median cluster has 14,392 CPU cores, 15,296 FP16 TFLOPS, and 1,911 FP32 TFLOPS. Capital investment is lumpy, as 55 percent of university-years add no FP16 GPU capacity.

Table 1: Summary statistics for the universities in the computing panel

	Median	P25	P75	N
Years in the panel	11	9	13	53
STEM R&D expenditure, 2024 ($ million)	230	143	537	53
All R1 universities	226	97	551	184
Endowment, 2024 ($ million)	958	703	1,681	53
All R1 universities	947	441	2,572	186
Endowment per tenure-line faculty, 2024 ($ thousand)	1,026	707	1,838	53
All R1 universities	1,296	697	2,745	186
CPU cores, 2024	14,392	7,732	27,136	53
GPU FP16 TFLOPS, 2024	15,296	6,240	36,672	53
GPU FP32 TFLOPS, 2024	1,911	1,124	4,150	53

Notes: 53 universities, 2005–2025; the panel is unbalanced because a university enters with its first archived cluster snapshot. Each row is the distribution across universities; the *All R1 universities* rows repeat the row above for all 187 Carnegie-2025 R1 universities with data in 2024 (N counts those with data), of which the panel universities are a subset. STEM R&D is total R&D expenditure in science and engineering fields from the NSF HERD survey. Endowment is the IPEDS end-of-year market value, completed with the CMUP series where IPEDS omits an affiliated foundation. Faculty are tenured and tenure-track instructional staff (IPEDS). Dollar amounts are in 2017 dollars (BEA GDP deflator). CPU cores and GPU TFLOPS are the installed stock summed over a university's central clusters. FP16 is the dense tensor-core rate and FP32 the single-precision rate available for every card.

Fields differ substantially in the share of research that uses GPU methods (Table 2). We group output into five STEM fields: computer science, engineering, geosciences and physics, molecular and materials sciences, and life sciences. Molecular and materials sciences comprise chemistry, materials science, and materials engineering. Life sciences include biological and biomedical sciences, bioengineering, health sciences, and agricultural and natural-resource sciences.

The median university in our panel produces between 28 and 62 GPU-topic papers per field per year, with equal fractional credit across authors. These papers account for 64 percent of computer science output and 51 percent of geosciences and physics output, compared with 13 percent in the life sciences. The life sciences have by far the most faculty, much of it engaged in clinical and laboratory work. Using pre-2019 publications to identify computational researchers, their share of faculty is 65 percent in computer science, 44 percent in engineering, 59 percent in geosciences and physics, and 11 percent in the life sciences.

Table 2: Research output and faculty by field group

	GPU-topic papers		Faculty	
Field group	Per year	Share of all	All	Computational
Computer science	52.0	64%	59	39
Engineering	56.3	37%	120	53
Geosciences and physics	41.9	51%	104	61
Molecular and materials sciences	28.5	38%	69	16
Life sciences	61.6	13%	623	69

Notes: Medians across the 53 universities of each university's average over its panel years 2013–2023. GPU-topic papers: OpenAlex works with at least one of their top three topics classified as benefiting from GPU computing, counted with equal fractional credit across a paper's authors; the share divides them by all of the university's papers in the field group. Field groups follow the NSF field crosswalk: engineering is aerospace, mechanical, electrical, chemical, civil, industrial, and other engineering; life sciences include biology, bioengineering, health, and agricultural sciences; molecular and materials sciences are chemistry, materials science, and materials engineering. Faculty: authors with four or more years at the university who arrived mid-career or stayed eight or more years, counted in every year of the spell; computational faculty: at least half of their 2010–2018 work in computational topics.

4 Descriptive Facts

Three descriptive facts motivate our analysis. First, less-resourced universities rely more on external funding for their HPC clusters. Second, large GPU upgrades are followed by changes in both publications and research labor that vary by field and by financial resources. Finally, hardware purchases around NSF awards suggest that grants displace universities' own investment. Because publications and labor move together after an upgrade, the event studies cannot isolate the contribution of computing capital itself.

4.1 Financial Resources Matter for On-Premises Computing

Endowment per faculty is our measure of university financial resources. We average endowment per tenured and tenure-track faculty member over FY2012–15 (before all GPU capacity changes in the panel), and assign each university its percentile in the distribution of this measure across all 187 R1 universities. The universities in our sample range from the 1st to 96th percentile of that distribution. Endowment also tracks general research demand: regressing STEM R&D spending per faculty on endowment per faculty gives a slope of 0.34 (robust s.e. 0.05). This rank measures wealth and proxies for general financial resources since we do not observe the research computing budget.

Less-endowed universities rely more heavily on NSF-funded clusters. For each university, we compute the share of its 2024 total GPU capacity that is in clusters acknowledging an NSF award.

Figure 3 plots the average of this share across the universities in five equal-sized bins. The average share falls from 45 percent in the least-endowed bin to 10 percent in the best-endowed bin. Moving up ten percentiles in the R1 distribution is associated with an NSF share that is 5.9 percentage points lower (robust s.e. 2.3). The pattern is consistent with external grants financing capacity that less-resourced universities find harder to fund themselves.

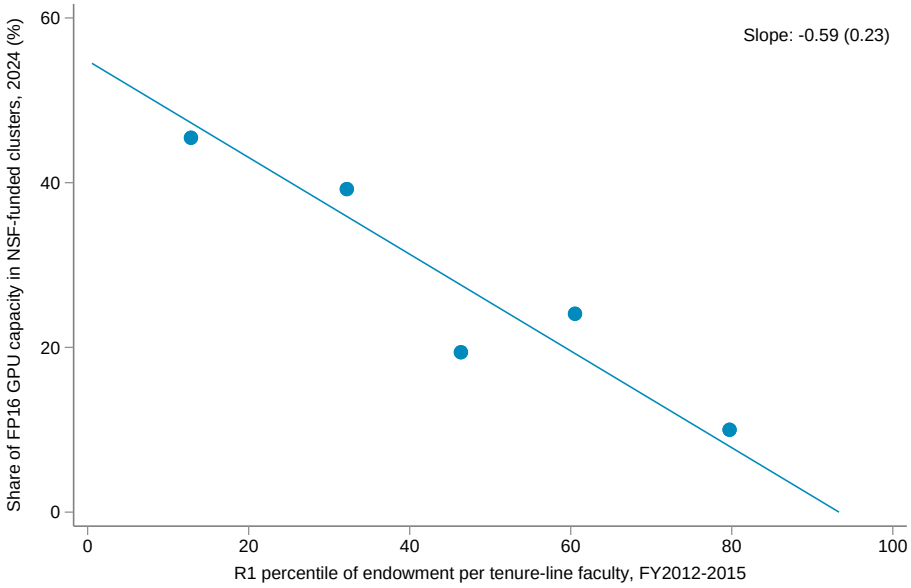


Figure 3: Reliance on NSF-Funded Clusters along the Endowment Distribution. Binned scatter plot of the share of a university's 2024 FP16 GPU capacity in clusters that acknowledge an NSF award against its percentile in the distribution of endowment per faculty (FY2012–15) across all R1 universities. Each point is the unweighted mean of the share across the universities in one of 5 equal-sized bins of the percentile. The line is the linear fit on the university-level data. Its slope, in percentage points per percentile, is reported with a heteroskedasticity-robust standard error in parentheses.

4.2 Publication and Labor Responses to GPU Upgrades by Resources and Field

We estimate annual responses to large GPU upgrades separately for universities with above- and below-median endowment per faculty (measured on all R1 universities). The treatment is a capacity upgrade, the outcomes are the university's GPU-related publications and its research labor, and controls are universities with no large upgrade inside the event window. We report university-wide and field-level responses, distinguishing AI and deep learning from GPU-topic research more broadly, and then ask whether labor inputs (i.e., PhD students, postdocs, and faculty) also move with the upgrade.

Outcomes. We use two outcomes. *GPU-topic* papers are those with at least one OpenAlex topic classified as GPU-heavy or as both GPU- and CPU-heavy (Section 3). This broader set includes the simulation, imaging, and data-analysis work that uses GPUs in the natural sciences and engineering as well as deep learning. *AI and deep-learning* papers are those whose title or abstract mentions machine learning, deep learning, neural networks, large language models or natural language processing, PyTorch or TensorFlow, or GPUs themselves (Appendix A1.2.4). Each paper is counted with equal fractional credit across its authors and summed to the university-year or university-field-year. Appendix A2 repeats every publication figure with output weighted by citation percentile.

We also use the two labor inputs as outcomes. *Flexible labor* is the number of full-time graduate students and postdoctoral researchers from the NSF Survey of Graduate Students and Postdoctorates (GSS). *Quasi-fixed faculty* is the stock of researchers with at least four years of publications at the university who arrived mid-career or stayed eight or more years, built from OpenAlex author histories as in Section 3.

Events and comparison groups. An event is a year in which a university's installed FP16 GPU capacity rises by at least the median upgrade: 5,000 TFLOPS (roughly sixteen NVIDIA A100 GPUs).[8] A university can have more than one event: we keep every increase above the threshold that happens at least 3 years after the university's previous event and we require at least one post-event year. This gives 48 events at 39 universities between 2018 and 2024, with 23 events at universities below the R1 median of endowment per faculty and 25 above it. Because adoption is staggered and most of the panel is eventually treated, we use a stacked design (Cengiz et al. 2019, Wing et al. 2024). For each event at university j in year k we build a stack containing the years $k-3$ to $k+3$ for that university and for every other university with no event inside that window. Thus, control universities either never have an event (never treated) or have an upgrade event later (not-yet treated). Stacks have between 11 and 33 controls, fewer in later years as more of the panel is treated.

Specification. Let $G_j \in \{B, A\}$ indicate whether university j is below or above the median of the endowment per faculty distribution across all R1 universities, and let z_j be its log endowment per faculty. Throughout this section, we call the two groups less- and more-resourced universities.

[8] Among university-years with a capacity increase, the median increase is 5,432 TFLOPS, with 25th and 75th percentiles of 1,979 and 13,000 TFLOPS (Appendix Figure A.1). Sixteen A100s are two standard eight-GPU servers and cost about $0.2 million at the median purchase-order price of $38 per FP16 TFLOPS.

Because the outcomes are counts (papers or people) with zeros, we estimate the stacked event study by Poisson pseudo-maximum likelihood (PPML) rather than in logs, following Chen and Roth (2024):

$$\mathbb{E}\left[y_{jst} \mid \cdot\right] = \exp\left(\alpha_{js} + \lambda_{st} + \tilde{\lambda}_{st}\, z_j + \sum_{g\in\{B,A\}} \sum_{\substack{k=-3 \\ k\neq -1}}^{3} \beta_k^g\, D_{jst}^k\, \mathbf{1}[G_j = g]\right) \tag{4.1}$$

where y_{jst} is the outcome of university j in stack s and year t. This is either paper counts or labor headcounts. D_{jst}^k equals one when j is the treated university of stack s and t is k years from its event, α_{js} are university-by-stack effects, λ_{st} are stack-by-year effects shared by all universities in the stack, and $\tilde{\lambda}_{st} z_j$ lets each stack-year effect vary linearly with resources, so more- and less-resourced universities may follow different aggregate trends. The year before the upgrade is the reference period. The coefficients β_k^B and β_k^A are the effects of an upgrade k years after the event at less- and more-resourced universities. A coefficient β is approximately a percent change for small values, so we report effects in percent. Standard errors are clustered by university.

Identification. The identifying assumption is that, absent the upgrade, a treated university would have followed the same annual publication path as the control universities in its stack, up to the university effects, the common year effects, and the linear resource trend. However, upgrade events are not random since universities decide to invest in compute if there are growing computational needs. Review, procurement, and construction delays separate an investment decision from installation, which usually takes 18 to 24 months. Identification requires the year of *installation* to be unrelated to otherwise different publication trends which spur the decision to invest.

4.2.1 AI and Deep-Learning Research Rises Where Resources Are Scarce

First, we restrict our analysis to AI and deep-learning publications, where we might expect to find the biggest response. We find that AI and deep-learning publications respond more at less-resourced universities. Figure 4 reports the estimates, pooling all fields in panel (a) and restricting to computer science in panel (b). For less-resourced universities, AI and deep-learning papers rise by 10.6 percent on average after the upgrade, with the effect growing over time. At more-resourced universities, the point estimates are negative but imprecise (-6.7 percent on average).

The pattern is sharper in computer science, where deep learning is a core use of GPUs. Less-resourced universities increase their AI and deep-learning publications by 15.5 percent, while the

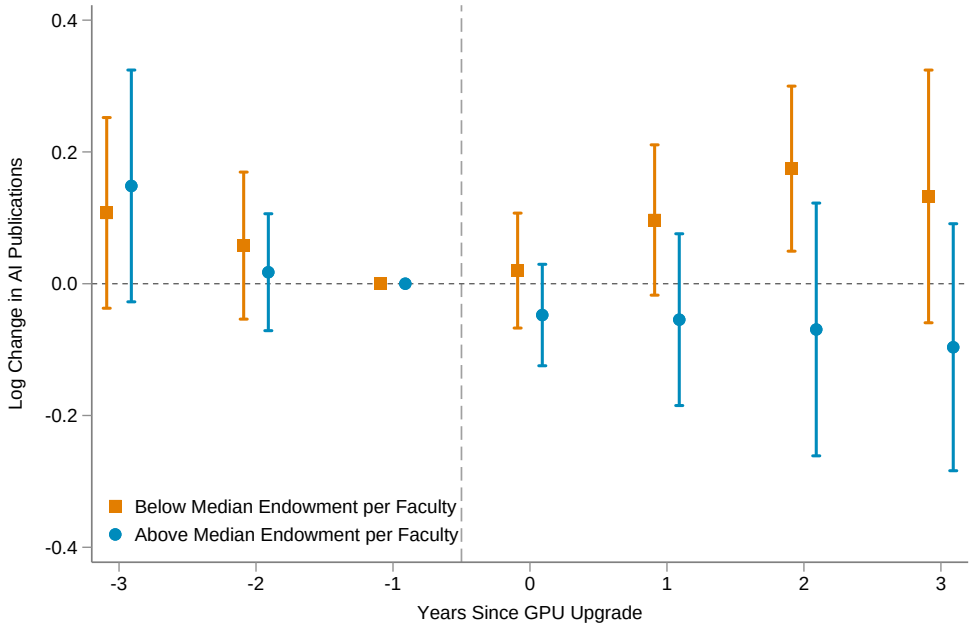
response at more-resourced universities is an imprecise 2.4 percent. Citation weighting sharpens the contrast (Appendix Figure A.2). Pooled across fields, citation-weighted AI and deep-learning output rises by 9.0 percent at less-resourced institutions and changes by −17.5 percent at more-resourced institutions. Thus, there is a decline in the quality of AI and deep-learning papers (as measured by citations) at more-resourced universities.

We interpret the response at less-resourced institutions as central upgrades relaxing binding access constraints. Researchers at more-resourced universities may already use department hardware, cloud credits, industry collaborations, or national facilities for deep-learning research, so a central upgrade adds little to the computing they can reach. The resource gradient is consistent with these alternatives reducing the marginal value of a central upgrade. The event study does not identify why AI output would fall at more-resourced universities after an upgrade, and the estimates are imprecise.

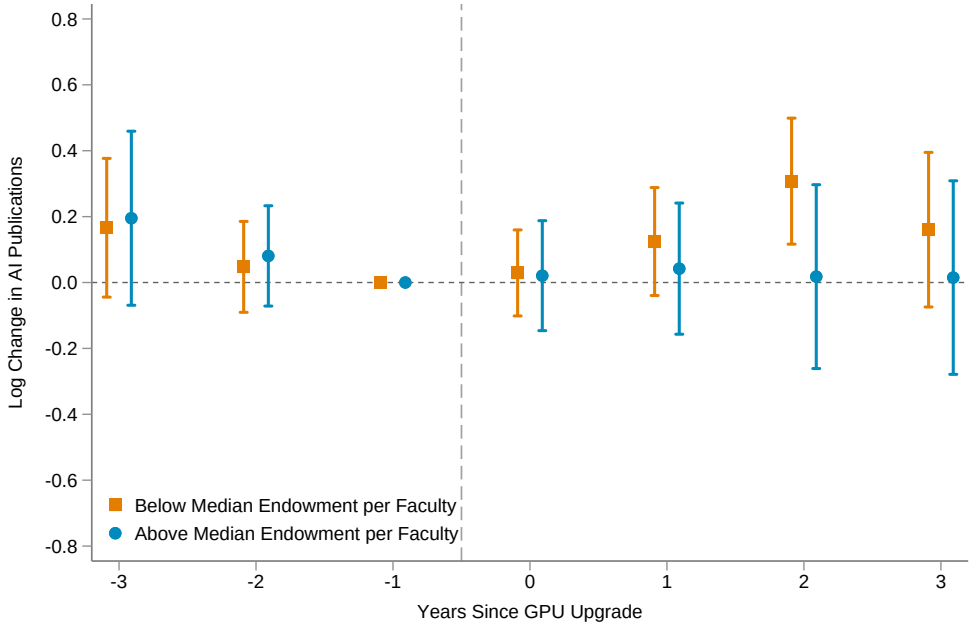
4.2.2 GPU-Topic Research Rises Modestly and Varies by Field

Now we examine a broader measure of output: all GPU-topic papers, which include scientific simulations that can be GPU-accelerated. Figure 5 shows that GPU-topic output rises after an upgrade by a similar, modest amount at both less- and more-resourced universities. Pooled across fields, panel (a) shows that the post-event means are 4.1 and 2.7 percent at less- and more-resourced universities. The response builds over the first two years in both groups, suggesting that capital usage occurs with a lag.

The field-level estimates are more heterogeneous and noisier. For less-resourced universities, the largest responses are in computer science (8.1 percent) and in the geosciences and physics (7.0 percent). For more-resourced universities, the point estimates are positive in the life sciences (6.1) and engineering (4.5) and negative in the molecular and materials sciences (−9.3). Citation weighting (Appendix Figure A.3) leaves the pooled response at less-resourced universities unchanged (4.0) and moves the response at more-resourced universities to zero. In the geosciences and physics, the citation-weighted response is 10.6 percent at less-resourced universities, but parallel trends do not hold, which limits a causal interpretation.

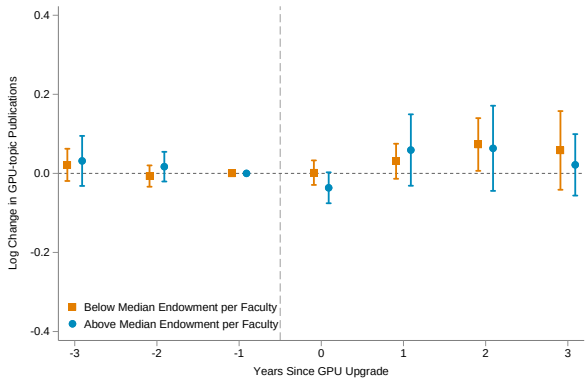


(a) All fields

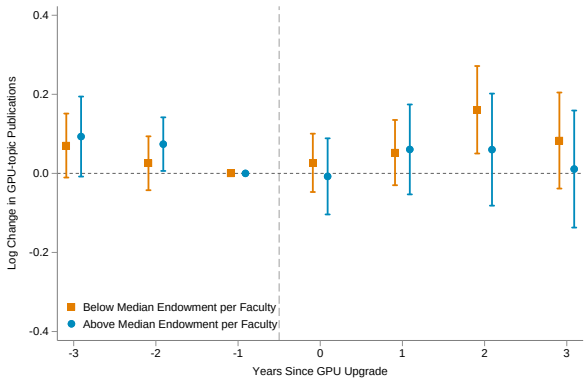


(b) Computer science

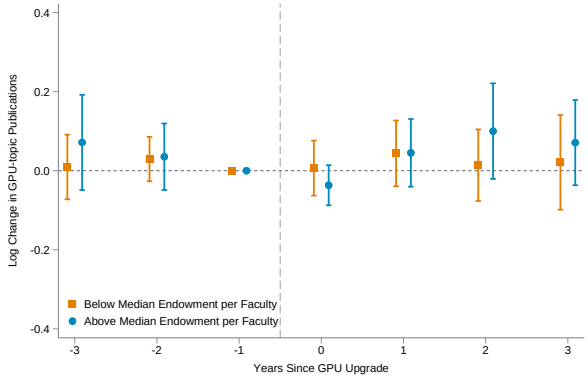
Figure 4: AI and Deep-Learning Papers around Large GPU Upgrades, by Financial Resources. Stacked event-study estimates of Equation (4.1) for the equal-author-weighted count of AI and deep-learning papers (title or abstract mentions deep learning, neural networks, large language models or natural language processing, PyTorch or TensorFlow, or GPUs), for all fields of the university (panel a) and for computer science (panel b). Points are PPML estimates of β_k^g in Equation (4.1), the proportional change in output relative to the year before the upgrade and to control universities in the same stack. Squares are universities below the R1 median of endowment per faculty and circles are those above. An event is a within-year increase in FP16 GPU capacity of at least 5,000 TFLOPS. Bars are 95% confidence intervals; standard errors are clustered by university.



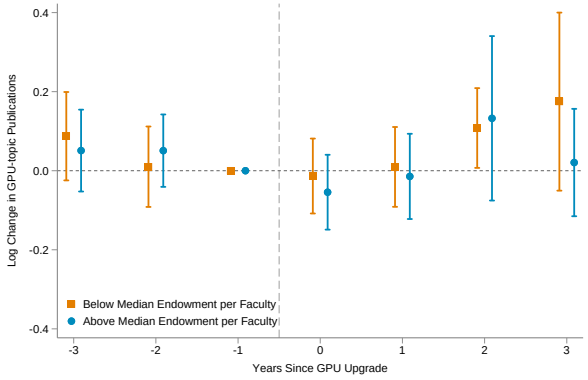
(a) All fields



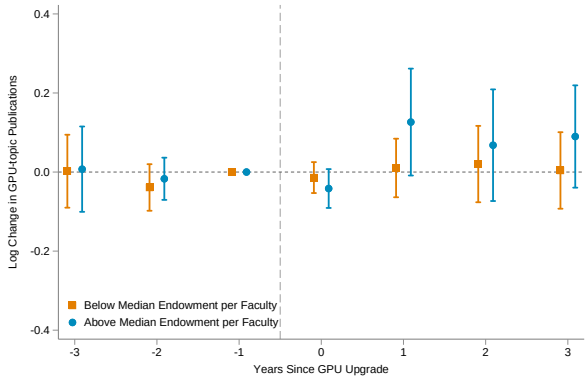
(b) Computer science



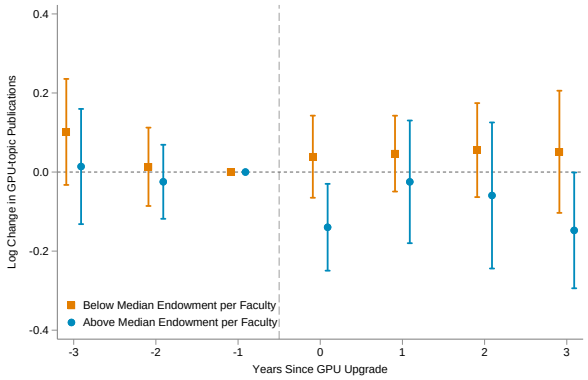
(c) Engineering



(d) Geosciences and physics



(e) Life sciences



(f) Molecular and materials sciences

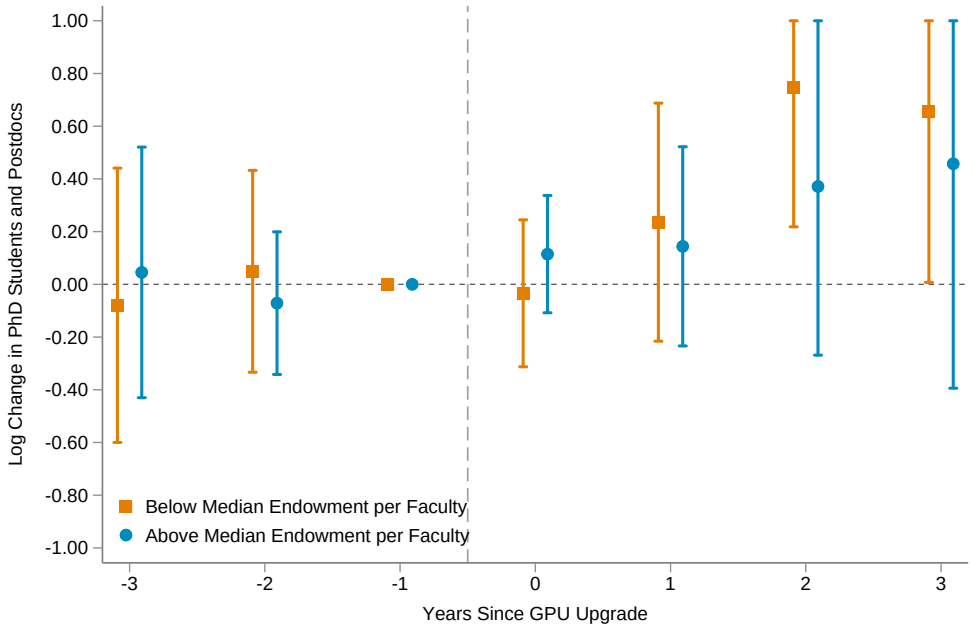
Figure 5: GPU-Topic Papers around Large GPU Upgrades, by Financial Resources. Stacked event-study estimates of Equation (4.1) for the equal-author-weighted count of GPU-topic papers (at least one OpenAlex topic classified as GPU-heavy or as both GPU- and CPU-heavy), for all fields of the university (panel a) and for each of five field groups (panels b–f). Points are PPML estimates of β_k^g in Equation (4.1), the proportional change in output relative to the year before the upgrade and to control universities in the same stack. Squares are universities below the R1 median of endowment per faculty and circles are those above. An event is a within-year increase in FP16 GPU capacity of at least 5,000 TFLOPS. Bars are 95% confidence intervals; standard errors are clustered by university.

4.2.3 Labor Responses to GPU Upgrades

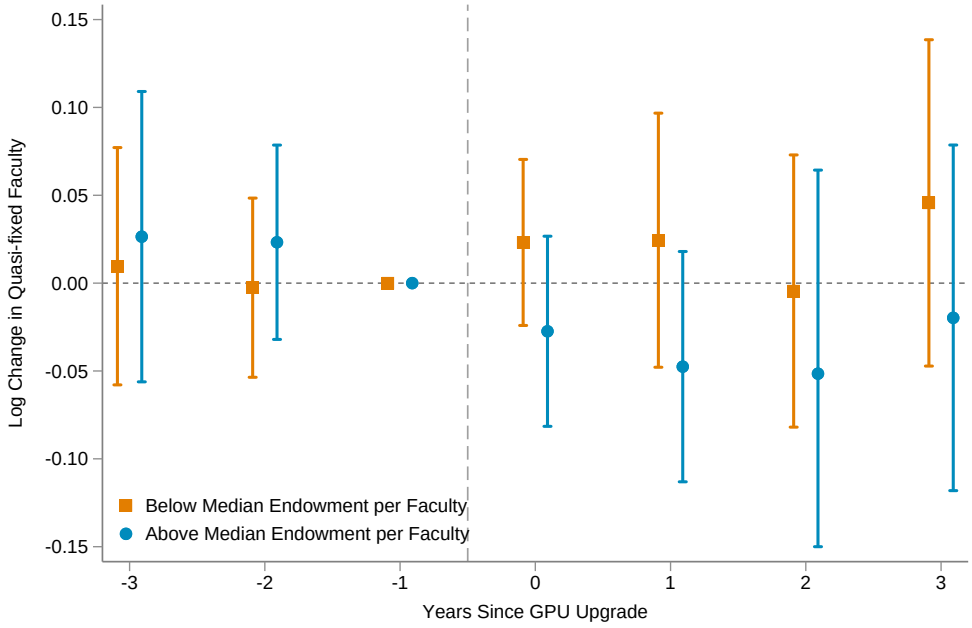
Publications are not the only margin that moves after an upgrade. Figure 6 reports the labor responses in computer science, where the publication response is also the clearest. The number of full-time PhD students and postdocs in computer science is 40.1 percent higher on average after the upgrade at less-resourced universities, and 27.2 percent higher at more-resourced universities. The estimates are imprecise, but the parallel trend assumption holds. The quasi-fixed faculty stock does not move significantly: 2.2 percent below the median and −3.7 above it. In computer science, an upgrade is accompanied by growth in the number of students and postdocs working with the same faculty.

The pattern differs across fields (Appendix Figures A.4 and A.5). Flexible labor rises in the molecular and materials sciences, though the parallel trend assumption fails for more-resourced universities. The remaining fields do not show a statistically significant change, though less-resourced universities show an increase in life science flexible labor (with failing pre-trends). Quasi-fixed faculty instead grows in engineering (3.5 and 5.3 percent) and drifts up slowly in the life sciences, while it is flat in computer science, the geosciences and physics, and the molecular and materials sciences. Pooled across the five groups, flexible labor is 8.1 percent higher below the median and unchanged above it (−0.3).

Taken together, the event studies show that large GPU upgrades are followed by changes in both output and inputs, and that which margin responds, students and postdocs, faculty, or neither, varies by field and by resources. A publication response therefore combines the direct effect of the new capital with the effect of the labor that arrives alongside it, and the reduced-form estimates cannot disentangle the two.



(a) Flexible labor: full-time graduate students and postdocs



(b) Quasi-fixed faculty

Figure 6: Labor Responses to Large GPU Upgrades in Computer Science, by Financial Resources. Stacked event-study estimates of Equation (4.1) for the number of full-time graduate students and postdocs in computer science from the NSF GSS (panel a) and for the quasi-fixed faculty stock in computer science from OpenAlex author histories (panel b). Points are PPML estimates of β_k^g in Equation (4.1), the proportional change in the labor input relative to the year before the upgrade and to control universities in the same stack. Squares are universities below the R1 median of endowment per faculty and circles are those above. An event is a within-year increase in FP16 GPU capacity of at least 5,000 TFLOPS. Bars are 95% confidence intervals; standard errors are clustered by university.

4.3 Evidence of Crowd-Out

NSF awards may add to a university's computing capital or replace purchases it would have made anyway. We examine this with the value of the hardware each panel university adds between consecutive capacity snapshots. GPU TFLOPS are valued at the median price per FP16 TFLOPS in university purchase orders (GovSpend) for datacenter GPUs in the year the hardware arrives, and CPU cores at the GovSpend median of $84 (Appendix Table A.3). Hardware added in flow year t is the change between the t and $t+1$ snapshots. The panel universities add $0.35 million of hardware per year on average and $0.07 million in the median year.

An event is an MRI or CC* computing award to a panel university, collapsed to the university and award year, with no other computing award in the preceding 3 years. There are 23 events at 22 universities with award years 2018–2023, worth $1.01 million on average and $23.1 million in total. Each event is matched to the three universities without a computing award within three years of the award year that are nearest in the average hardware added over the three pre-award years, log GPU capacity, and log endowment in the year before the award. The stacked regression is

$$H_{jst} = \alpha_{js} + \delta_{st} + \sum_{\substack{k=-3 \\ k\neq -1}}^{3} \beta_k \, \mathbf{1}\{j \text{ awarded in stack } s,\ t - s = k\} + e_{jst} \tag{4.2}$$

where H_{jst} is the hardware university j adds in flow year t within stack s, and the fixed effects are stack by university and stack by year. The coefficient β_k is the extra hardware, in millions of dollars per year, that an awarded university adds k years after the award relative to the year before the award and to its matched controls. Standard errors are clustered by university.

We estimate Equation (4.2) for two outcomes: total hardware added and the university's own spending. Own spending is calculated as total hardware added minus the event's award in the year its hardware arrives, so the two series coincide before the award.[9] Under full additionality, total hardware rises by the award and own spending does not move. Under complete crowd-out, total hardware does not move and own spending falls by the award.

Figure 7 shows both series. Before the award, total hardware and own spending are statistically indistinguishable from those of the controls. After the award, total hardware rises by $0.25 million in the following year, $0.26 million two years after, and $0.54 million three years after, compared to the average award of $1.01 million in the year most award hardware arrives. Own spending

[9] Award hardware is dated to the first year in which the cluster citing the award number appears in the capacity panel (5 awards), to the linked capacity jump (10 awards), or to two years after the award, the modal lag among linked awards (8 awards).

correspondingly falls by $0.54 million in the year after the award and recovers in the two years that follow. The evidence is suggestive of crowd-out in the year immediately after the award. Two features of the measurement limit the reading. First, awards are not randomly assigned, and the controls are universities with similar purchase histories and resources that did not win an award in the window. Second, the valuation only prices in GPU TFLOPS and CPU cores, while awards also fund storage, networking, and staff. Therefore, part of the shortfall per dollar is hardware the capacity measure does not capture.

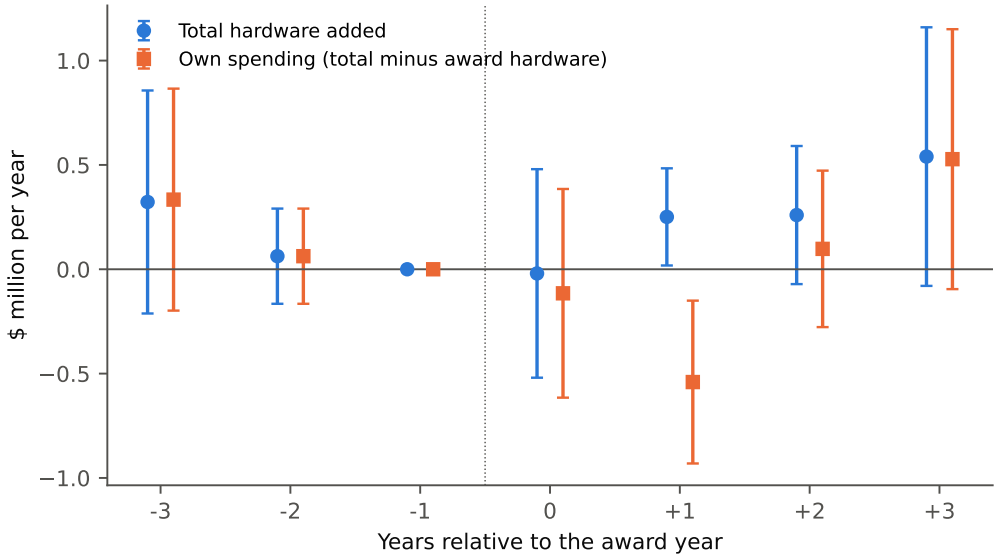


Figure 7: Hardware Purchases and Own Spending around NSF Computing Awards. Stacked event-study estimates of β_k in Equation (4.2), in millions of dollars per year, relative to the year before the award and to the 3 matched control universities. Circles are total hardware added, the CPU cores and GPU TFLOPS added between consecutive capacity snapshots, valued at GovSpend purchase-order prices. Squares are own spending, total hardware added minus the event's award in the year its hardware arrives. 23 MRI and CC* award events at 22 panel universities, award years 2018–2023. The average award is $1.01 million. Bars are 95% confidence intervals; standard errors are clustered by university.

5 Model

The return to a grant depends on both the output produced by additional capital and the university's investment response. Section 4 offers suggestive evidence that computing capital matters for GPU-relevant publications, and that external grants crowd out a university's own hardware purchases. We formalize this tradeoff with a stylized model in Section 5.1, then introduce a richer model with shared inputs and field differences in Section 5.2 (which is structurally estimated in Section 6).

5.1 Stylized Model

The return to an external grant separates into capital and labor channels before we impose a functional form or an optimizing model. Let N be an in-kind grant and I the university's own purchases, both in units of capital, so total capital is $K = N + I$. The university can adjust capital investment I and labor L in response to an external grant. For any differentiable production function in capital and labor $Y = f(K, L) = f(N + I(N), L(N))$, we have that:

$$\frac{dY}{dN} = \underbrace{\frac{\partial Y}{\partial K}}_{\substack{\text{marginal product} \\ \text{of capital}}} \times \underbrace{\left(1 + \frac{\partial I}{\partial N}\right)}_{\text{pass-through } m} + \underbrace{\frac{\partial Y}{\partial L} \times \frac{\partial L}{\partial N}}_{\text{complementary labor}} \tag{5.1}$$

where $MPK \equiv \partial Y/\partial K$ and $MPL \equiv \partial Y/\partial L$ are the marginal products of capital and labor, and the pass-through $m \equiv 1 + \partial I/\partial N$ is the share of a grant that becomes *new* capital rather than replacing the university's own purchases. The first term is the capital channel: the technical return to additional capital, scaled by the pass-through m. The second term captures any labor response. The following economic model determines these responses.

Setup. To characterize the responses, suppose the university combines capital K and labor L in a Cobb–Douglas production function with diminishing returns to scale:

$$Y = \Omega\, K^{\alpha_K} L^{\alpha_L}, \qquad \alpha_K + \alpha_L < 1 \tag{5.2}$$

where Ω is Hicks-neutral productivity. Own purchases cannot be negative ($I \geq 0$). The university chooses I and L to maximize output net of input costs, subject to a research budget B:

$$\max_{I \geq 0,\, L \geq 0} \; Y - pI - wL \qquad \text{s.t.} \qquad pI + wL \leq B \tag{5.3}$$

Universities differ in two primitives: productivity Ω, which determines output per unit of input, and the budget B, which determines how many inputs the university can finance. The budget is the amount available for these research inputs, rather than the university's total assets.

Productivity. Before a grant ($N = 0$), at a constrained interior solution, the marginal product of capital is

$$MPK \propto \frac{\Omega}{B^{1-(\alpha_K+\alpha_L)}} \tag{5.4}$$

Higher productivity raises the return to capital one-for-one. A larger budget lets the university operate at a larger scale, which lowers the return to additional units of input because of diminishing returns to scale. A well-resourced university therefore has a higher marginal product of capital only if its productivity advantage outweighs its larger existing input bundle.

Crowd-out. The pass-through m depends on which of three regimes the university is in, summarized in Table 3. Let $\chi \equiv \alpha_K/(\alpha_K + \alpha_L)$ denote the share of a marginal budget dollar that a constrained university spends on capital.

Table 3: Crowd-Out Regimes in the Stylized Model

Regime	Condition	$\partial I/\partial N$	Pass-through m
Corner	No own purchases, $I = 0$	0	1
Constrained interior	Budget binds, $I > 0$	$-(1-\chi)$	χ
Unconstrained	Budget slack	-1	0

Notes: χ is the share of a marginal budget dollar a constrained university spends on capital. Derivations are in Appendix A3.

A university in the *corner* buys no capital of its own, so the grant displaces no purchases and pass-through is one. A *constrained interior* university treats the grant as cash: it reallocates its effective budget, spending the share χ on capital and the remainder on labor. Thus, low capital pass-through need not mean a low total output return; the freed funds support the labor term in (5.1). An *unconstrained* university already has its desired input bundle, so a grant below that capital stock changes who pays without changing the bundle. Holding productivity and the grant fixed, larger internal budgets move universities toward regimes with more crowd-out.

NSF's Tradeoff. High productivity can make capital grants valuable at a well-resourced university, while a small input stock and greater additionality can favor a less-resourced one. Appendix A3.4 shows that even the ranking of marginal grant returns can change with grant size. The dominance of either effect is an empirical question that depends on both productivity $\{\alpha_K, \alpha_L\}$ and crowd-out $\partial I/\partial N$ parameters.

5.2 University Science Production Function

To empirically measure the productivity and crowd-out parameters, we model the university's objective function and science production. We first focus on estimating α_K, for which there are two issues—endogeneity of capital stocks and no observed field-level capital allocations. Modeling the university's problem will allow us to identify instruments for the endogeneity issue and derive field-level capital allocations as a function of observed data. Concretely, universities choose to invest in capital in response to expected research productivity, which creates transmission bias in a naive regression of output on capital. We model research productivity (unobserved to the econometrician) as an AR(1) process and identify the production parameters using instruments based on the timing of input decisions. In addition to the endogeneity issue, we also observe CPU and GPU stocks only at the university level. Following the logic of multi-product production in De Loecker et al. (2016), we treat fields as separate products and use the university's optimization to frictionlessly allocate the cluster among them.

5.2.1 Production Technology

Let f index the five fields in the production function:

$$f \in \mathcal{F} = \{\text{computer science, engineering, geosciences and physics,}$$
$$\text{molecular and materials sciences, life sciences}\}$$

We exclude mathematics, the social sciences, psychology, and the humanities, where GPU-topic output is sparse during the main analysis period, 2013 to 2024. Output Y_{jft} is the fractional count of GPU-topic papers defined in Section 3, for field f at university j in year t:

$$Y_{jft} = \Omega_{jft}\, \mathcal{K}_{jft}^{\alpha_f^K}\, \mathcal{L}_{jft}^{\alpha_f^L} \tag{5.5}$$

where $\Omega_{jft} \equiv \exp(\omega_{jft})$ is Hicks-neutral productivity. Capital and labor are each CES aggregates:

$$\mathcal{K}_{jft} = \left[\gamma_f\, C_{jft}^{\frac{\sigma_K - 1}{\sigma_K}} + (1-\gamma_f)\, G_{jft}^{\frac{\sigma_K - 1}{\sigma_K}}\right]^{\frac{\sigma_K}{\sigma_K - 1}} \tag{5.6}$$

$$\mathcal{L}_{jft} = \left[\phi_f\, (L_{jft}^V)^{\frac{\sigma_L - 1}{\sigma_L}} + (1-\phi_f)\, (L_{jft}^F)^{\frac{\sigma_L - 1}{\sigma_L}}\right]^{\frac{\sigma_L}{\sigma_L - 1}} \tag{5.7}$$

C_{jft} and G_{jft} are the CPU cores and GPU TFLOPS allocated to field f. L_{jft}^V is flexible labor (PhD students and postdocs) and L_{jft}^F is quasi-fixed labor (faculty). The nested structure allows CPU–GPU substitution and trainee–faculty substitution to differ from substitution between capital and labor. Distribution parameters $\{\gamma_f, \phi_f\}_f$ and output elasticities $\{\alpha_f^K, \alpha_f^L\}_f$ vary by field. The substitution

elasticities (σ_K, σ_L) are common across fields. Intuitively, substitution elasticities (σ_K, σ_L) govern how easily the fields can substitute between inputs when prices shift (above one is substitutes, below one is complements). Distribution parameters $\{\gamma_f, \phi_f\}_f$ reflect the field's reliance on one input versus the other holding substitutability fixed.

Productivity is the sum of a permanent university-by-field effect and a persistent AR(1) shock:

$$\omega_{jft} = \bar{\omega}_{jf} + \nu_{jft}, \qquad \nu_{jft} = \varrho_f \, \nu_{jft-1} + \xi_{jft}, \qquad \xi_{jft} \text{ i.i.d.} \tag{5.8}$$

The permanent effect captures stable differences across universities, such as department reputation or quality. The persistent shock captures, for example, a field that hires a productive cohort or wins a large center grant. Its evolution is governed by the field-level parameters $\{\varrho_f\}_f$.

5.2.2 Timing and the University's Problem

The university chooses inputs to maximize the present discounted value of research output net of costs.[10] At the end of year $t-1$, before observing the next productivity shock, it chooses capital investment and faculty hiring. In year t, after observing ξ_{jft}, it chooses PhD and postdoctoral support and allocates the installed cluster across fields. Capital and faculty are therefore dynamic choices; flexible labor and within-year capital allocation are static choices.

The field-level capital allocation can be characterized without solving the full dynamic problem (Appendix A4), so we present the static problem for clarity. The within-year objective is:

$$\max_{\{L^V_{jft}\}_f, \{C_{jft}\}_f, \{G_{jft}\}_f} \underbrace{\sum_f \lambda_{jf} \, Y_{jft}}_{\substack{\text{weighted dollar} \\ \text{value of output}}} - \underbrace{\sum_f W^V_{jft} \, L^V_{jft}}_{\text{flexible labor cost}} - \underbrace{\Big(\sum_f W^F_{jft} \, L^F_{jft} + P^C_t \, I^C_{jt} + P^G_t \, I^G_{jt} \Big)}_{\substack{\text{faculty wages and capital investment} \\ \text{(dynamic decisions)}}} \tag{5.9}$$

Hiring and capital investment are chosen at the end of year t, with total faculty and capital stock becoming state variables for the next year $t+1$. These are dynamic decisions: W^F_{jft} is the faculty wage, and I^C_{jt} and I^G_{jt} are CPU and GPU investment at prices P^C_t and P^G_t (Appendix equation (A4.2)). They do not affect the static choices of field-level flexible labor and capital allocations. The weight λ_{jf} is the dollar value the university places on a paper in field f. It makes output and costs commensurable and allows valuations to differ across fields. These weights describe university preferences which the econometrician does not observe. The weights need not equal the weights in the NSF allocation benchmark, which counts papers equally when it maximizes total output.

[10] Research competes with other uses of university funds, including financial aid and teaching facilities.

Total faculty and capital are dynamic decisions, so the university hires PhD students and postdocs (L^V_{jft}) at wages W^V_{jft} subject to the budget constraint:

$$\sum_f W^V_{jft} L^V_{jft} \leq B_{jt} - \left(\sum_f W^F_{jft} L^F_{jft} + P^C_t I^C_{jt} + P^G_t I^G_{jt} \right) \tag{5.10}$$

where B_{jt} is the university's total budget. The university also allocates the cluster across fields subject to capacity constraints $\sum_f C_{jft} \leq C_{jt}$ and $\sum_f G_{jft} \leq G_{jt}$. Static choices do not affect next year's state and drop out of the continuation value, so we do not solve the full dynamic problem (refer to Appendix A4).

5.2.3 Allocating the Shared Cluster across Fields

The static first-order conditions deliver a rule that maps observed university-level capital into field-level inputs. Within each field, we take the ratio of the first-order conditions for capital and for flexible labor. The field's value weight and output cancel in the ratios and the common shadow values cancel when we normalize across fields. Assuming that PhD and postdoc wages are common across fields within a university-year and summing over fields gives

$$\begin{aligned} C_{jft} &= C_{jt} \cdot \frac{\psi^C_{jft} L^V_{jft}}{\sum_h \psi^C_{jht} L^V_{jht}} \\ G_{jft} &= G_{jt} \cdot \frac{\psi^G_{jft} L^V_{jft}}{\sum_h \psi^G_{jht} L^V_{jht}} \end{aligned} \tag{5.11}$$

where the compute intensities ψ^C_{jft} and ψ^G_{jft} are proportional to α^K_f / α^L_f and depend on the within-nest input mix (full derivation in Appendix A4.2). Intuitively, a field's share of the cluster is its share of PhD students and postdocs, tilted toward fields where compute is relatively more productive. Because the intensities depend on the allocation itself, we solve (5.11) as a fixed point.

The allocation rule in (5.11) rests on two assumptions. First, we assume flexible labor wages are common across fields within a university-year. Self-reported PhD stipends show small average differences across laboratory fields after accounting for university and year. This supports an approximation but does not establish equal effective labor costs in each university-year (Appendix A5). Second, we assume the cluster is fully allocated. HPC clusters do not run at full instantaneous utilization, but schedulers allocate the full stock of compute-hours across research groups who request usage. GPUs at Harvard run near 100% capacity and overall node occupancy (including CPUs) is stable at around 80% (Appendix Figure A.7). As long as utilization does not differ systematically across fields within a university-year, it cancels from the allocation shares.

6 Estimation

We calibrate substitution within the capital and labor nests using external data, then estimate output elasticities and productivity persistence from the panel. This separates parameters that output residuals alone identify poorly, especially when capital is observed only at the university level. Table 4 summarizes the three parameter blocks: labor $(\sigma_L, \{\phi_f\}_f)$, capital $(\sigma_K, \{\gamma_f\}_f)$, and the outer parameters $\{\alpha_f^K, \alpha_f^L, \varrho_f\}_f$. Appendix A6 gives the details.

Table 4: Overview of Parameter Estimation

Block	Parameters	Approach and Data
Labor nest	σ_L, ϕ_f	Calibrated: $\sigma_L = 1$; ϕ_f is the flexible-labor share of each field's total research wage bill
Capital nest	σ_K	Estimated from the CPU response to GPU price declines and GPU share of cluster user cost from purchase orders
	γ_f	Calibrated from field-level CPU and GPU usage at three universities
Outer parameters	α_f^K, α_f^L, ϱ_f	System GMM (Blundell and Bond 1998), iterated with the field-level allocation rule to a fixed point

6.1 Labor Nest

Since we do not have variation that differentially shifts flexible versus fixed labor wages, we set $\sigma_L = 1$, so that PhD students and postdocs combine with faculty in a Cobb–Douglas aggregate. The advantage of the Cobb–Douglas benchmark is that ϕ_f then equals the flexible-labor share of the field's research wage bill, which we can measure directly. The wage-share reading of ϕ_f is a long-run cost-minimization condition: faculty are fixed within the year, but a university that hires faculty optimally over time equates the two inputs' cost shares to their Cobb–Douglas weights.

We approximate the wage shares from self-reported PhD stipends, NIH NRSA postdoctoral rates, and 40% of all faculty salaries from IPEDS, allocating faculty to fields by the NSF HERD R&D share. Table 5 reports the result. Flexible labor accounts for about two-thirds of the research wage bill in computer science (0.68) and engineering (0.71), three-fifths in geosciences and physics (0.61), four-fifths in the molecular and materials sciences (0.80), and a little under three-fifths in the life sciences (0.57).

For robustness, we also calibrate $\sigma_L \in \{1.5, 2\}$ and $\phi_f = 0.5$ for all fields. Visa shocks move

the number of PhD students and postdocs but not their pay relative to faculty, so they identify the output elasticity of the labor aggregate and not the substitution within it. Both σ_L and ϕ_f are therefore calibrated rather than estimated.

Table 5: Flexible-Labor Share of the Research Wage Bill

	CS	Engineering	Geo./Phys.	Mol./Mat.	Life
ϕ_f	0.68	0.71	0.61	0.80	0.57

6.2 Capital Nest

Substitution between CPUs and GPUs. The ideal regression would relate the log ratio of GPU to CPU capacity to the log of their relative price, but we do not observe university-specific relative prices. GPU hardware becomes cheaper for every university at once, and year effects absorb a common price change. We instead exploit differences in when universities must replace their fleets, which turn a common price drop into university-specific changes in GPU quantities, and identify σ_K from how CPU capacity responds to those changes. If CPUs and GPUs are complements, universities that add GPUs also buy more CPUs to run alongside them, and vice versa if they are substitutes. Therefore, the ratio of the CPU response to the GPU response maps into σ_K.

To generate price variation that differs across universities, we interact the frontier GPU price with each university's exposure to replacement. Let p_t^G be the lowest price per FP16 TFLOPS among GPUs available in year t, and let s_{jt-1}^{rep} be the share of the university's GPU TFLOPS that is at least five years old. The instrument is

$$Z_{jt}^{\text{NVDA}} = \Delta \log p_t^G \times s_{jt-1}^{\text{rep}} \tag{6.1}$$

Universities whose fleets are due for replacement are the ones that respond to a new generation's price drop. We regress the change in log GPU TFLOPS and in log CPU cores between years $t-1$ and $t+1$ on Z_{jt}^{NVDA}, with university and year effects and year-specific slopes on each university's initial CPU–GPU mix and field composition (Appendix A6.2). Figure 8 shows the two responses. GPU capacity rises by 2.52 log points (s.e. 0.66) per unit of $-Z_{jt}^{\text{NVDA}}$, that is, per 100 log-point price drop reaching a fully exposed fleet; at the 2020 Ampere transition, a university whose whole fleet was due for replacement added about 1.6 log points more GPU capacity than one with a new fleet (first-stage F of 15). CPU capacity rises as well, by 0.44 (0.29): CPUs are not cut when GPUs become cheap, so the two inputs move together, which under the capital nest makes them complements. The ratio

of the responses is $\hat{\Theta} = \hat{\pi}_C/\hat{\pi}_G = 0.17$ (0.10). Mapping Θ into σ_K requires the GPU share of cluster spending, which we compute as the GPU share of the annual user cost of the installed stock at the median purchase-order prices of Appendix A1.4, \$38 per FP16 TFLOPS and \$84 per CPU core, and the inverse price elasticity of demand for the capital aggregate implied by the output elasticities, $\psi = 0.95$. The GPU share is 0.36, and the implied elasticity is $\hat{\sigma}_K = 0.67$ (0.16). We use the point estimate in the baseline, $\sigma_K = 0.67$, and report robustness results at 0.4 and 1 in Table A.9.

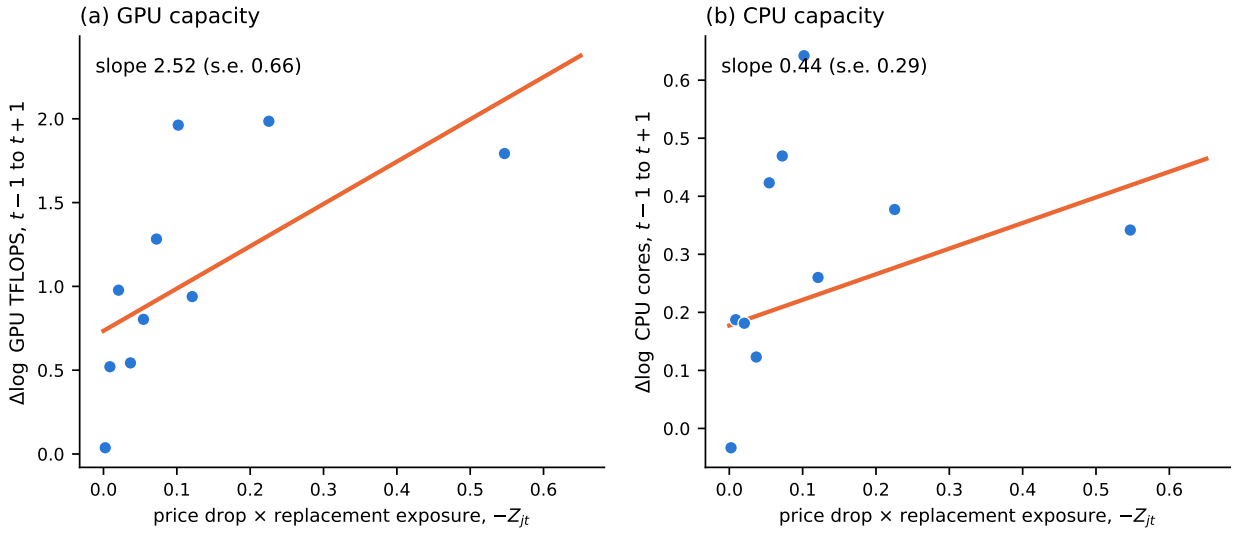


Figure 8: CPU and GPU Capacity Responses to the Replacement Instrument. Binned scatter plots of the change in log GPU TFLOPS (a) and log CPU cores (b) between years $t-1$ and $t+1$ against $-Z_{jt}^{\text{NVDA}}$, the frontier GPU price drop interacted with the share of the university's GPU fleet that is at least five years old. Dots are binned-scatter estimates according to Cattaneo et al. (2024): the piecewise-constant fit of the outcome on 10 quantile-spaced bins of the positive support of $-Z_{jt}^{\text{NVDA}}$ itself, with the mass point at zero (56% of the sample, university-years with no frontier transition or no fleet due for replacement) as an additional bin that enters the estimation but is not drawn. University and year effects are absorbed and year-specific slopes in the university's baseline log CPU/GPU ratio and baseline field shares enter as covariates and are evaluated at their sample means, rather than bin means of residualized variables. Lines are the linear fit of the same model, whose slopes are the first-stage coefficients, with standard errors clustered by university. The ratio of the slopes in panels (b) and (a) is $\hat{\Theta}$.

CPU weights. The level of γ_f is a choice of units, so only its ratios across fields matter, and we calibrate them from how fields actually split their computing between CPUs and GPUs. Three university research computing centers report annual CPU and GPU hours by department. Take computer science as the reference field. Then under the capital nest, the log ratio of CPU to GPU hours in a field differs from that in computer science by σ_K times the log odds ratio of the fields' CPU weights, plus a university-year term that absorbs each university's units. We regress the log ratio on field indicators and university-year fixed effects and invert the coefficients at $\hat{\sigma}_K$, normalizing

$\gamma_{\text{eng}} = 0.75$. Relative to computer science, the log ratio of CPU to GPU hours is 3.2 in engineering, 3.9 in geosciences and physics, 3.3 in molecular and materials sciences, and 2.4 in the life sciences, with standard errors of 0.2 to 0.4. Every other field uses more CPU hours relative to GPU hours than computer science at each center, and a uniform γ is rejected at each. The implied weights are $\gamma_f = (0.02, 0.75, 0.88, 0.75, 0.47)$ for computer science, engineering, geosciences and physics, molecular and materials sciences, and the life sciences. Hours of use are a flow while C_{jft} and G_{jft} are stocks, but the two coincide assuming a common utilization rate across fields, and only the cross-field ratio enters the calibration. Appendix A1.5 describes the usage data and reports the estimates by university. We also report results with uniform weights as a robustness check.

6.3 Output Elasticities

Given the nests and the allocation, let $k_{jft} \equiv \log \mathcal{K}_{jft}$ and $\ell_{jft} \equiv \log \mathcal{L}_{jft}$. We model capital entering with a one-year lag, which reflects the time between installation and adoption. This is motivated by the event study of Section 4.2 since the publication response to an upgrade builds over the first two years after hardware arrival. Quasi-differencing the production function with the productivity process (5.8) gives

$$y_{jft} = \varrho_f \, y_{jf,t-1} + \alpha_f^K \left[k_{jft-1} - \varrho_f \, k_{jft-2}\right] + \alpha_f^L \left[\ell_{jft} - \varrho_f \, \ell_{jft-1}\right] + (1 - \varrho_f)\, \bar{\omega}_{jf} + u_{jft} \tag{6.2}$$

where $u_{jft} \equiv \xi_{jft} + \varepsilon_{jft} - \varrho_f \, \varepsilon_{jf,t-1}$ combines the productivity innovation with measurement error ε_{jft} in output.

We estimate (6.2) by system GMM (Blundell and Bond 1998) separately for each field starting with a guess of field-level capital allocation. The differenced equation removes the permanent effect $\bar{\omega}_{jf}$ and uses lagged levels of inputs and output as instruments. Inputs chosen before this year's shock, such as capital, are uncorrelated with it. The levels equation uses lagged changes as instruments, under the assumption that changes in inputs are uncorrelated with the permanent effect. We supplement the instrument set with three sources of external variation. First, the GPU price shock (6.1) shifts capital. Second, visa-policy shocks (the 2017 travel ban, the 2020 restrictions on new visas, and national inflows of foreign postdocs), each scaled by the field's pre-period share of temporary-visa researchers from the NSF GSS, shift flexible labor. Finally, the leave-one-out labor of other fields shifts a field's share of the shared cluster through the allocation rule (5.11).

We recover $\{\alpha_f^K, \alpha_f^L, \varrho_f\}_f$ from the unrestricted coefficients by minimum distance. Because the allocation depends on α_f^K / α_f^L, we iterate between the field-level allocation and the GMM step

until the elasticities converge. Standard errors are clustered by university.

6.4 Results

The estimated capital elasticity is positive in every field and statistically significant at the 5% level in three (Table 6). All five fields exhibit decreasing returns to scale. The output elasticity of computing capital is 0.06 (0.02) in computer science, 0.05 (0.02) in geosciences and physics, and 0.05 (0.02) in the molecular and materials sciences. In the life sciences it is 0.05 (0.03), just short of significance at the 5% level, and in engineering 0.03 (0.02), not distinguishable from zero. The labor elasticity is 0.53 (0.17) in computer science, 0.95 (0.13) in engineering, 0.87 (0.22) in geosciences and physics, 0.56 (0.09) in the life sciences, and 0.44 (0.23) in the molecular and materials sciences. Productivity shocks are persistent in four fields ($\hat{\varrho}_f$ between 0.47 and 0.78) and are not significantly persistent in the life sciences (-0.14, with a standard error of 0.14). At the 5% level, the Hansen tests do not reject the overidentifying restrictions in any field, for the system (p-values from 0.10 to 0.50), for the differenced equation alone, or for the levels-equation instruments. Appendix Table A.8 splits the labor elasticity into its PhD-student-and-postdoc and faculty components since $\sigma_L \equiv 1$.

Capital elasticities are not very sensitive to the input calibrations (Appendix Table A.9). Across $\sigma_L = 1.5$ or 2, $\sigma_K = 0.4$ or 1, uniform labor weights, and fixed labor elasticities, capital elasticities stay between 0.02 and 0.07. Excluding papers with private-sector co-authors (due to company-provisioned computing resources) leaves the computer science estimates at 0.05. First- or last-author credit preserves the computer science, geosciences and physics, and life-science estimates, but weakens the molecular and materials estimate to 0.04 (0.03).

Weighting papers by their citation percentile raises the computer science elasticity to 0.09 but puts the geosciences and physics and life-science elasticities at zero. Outside of computer science, additional capital raises the number of GPU-topic papers but not their quality-adjusted volume, so the planner's gains in those fields rest on counts.

Two features of these estimates matter for the NSF's problem. First, the capital elasticities are small: doubling a university's computing stock raises GPU-topic output by three to four percent in the fields where capital matters. Thus, the marginal product of compute falls quickly as the installed stock grows, and returns to scale in capital and labor together are below one in every field (0.49 to 0.97). These diminishing returns create a reason to spread capital across universities, conditional on productivity and pass-through. Second, the elasticities differ across fields, so the return to a grant

depends on a university's field composition as well as its productivity.

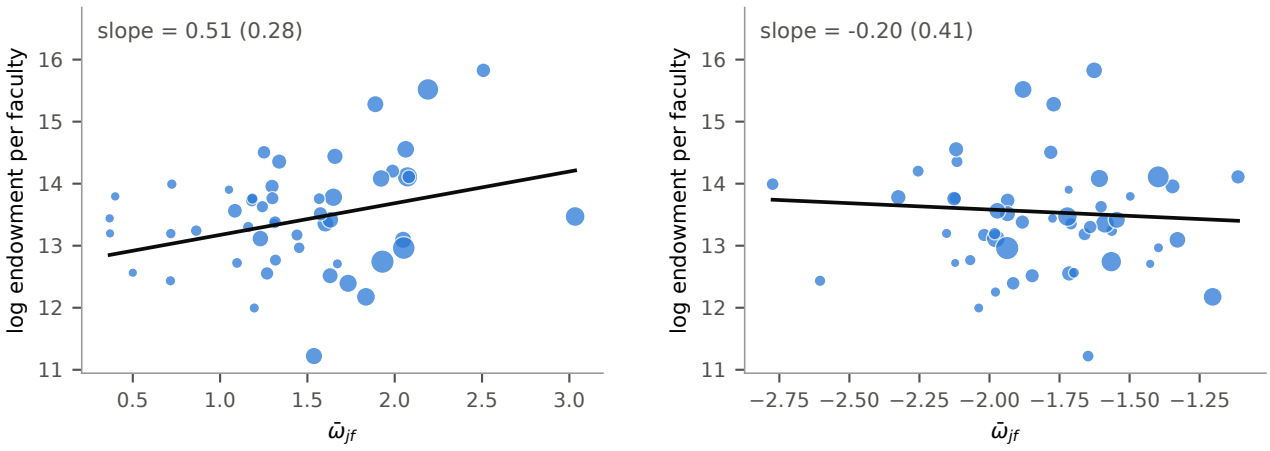
These elasticities and the event-study coefficients measure different objects. The production function gives an output response to capital conditional on labor and productivity. The event study follows universities after discrete upgrades, allowing complementary inputs to change. Neither by itself gives the return to a grant dollar, which also depends on crowd-out.

Table 6: Production Function Estimates by Field

	CS	Engineering	Geo./Phys.	Mol./Mat.	Life
$\hat{\alpha}_f^K$	**0.057**	0.027	**0.051**	**0.054**	0.053
	(0.020)	(0.023)	(0.021)	(0.025)	(0.028)
$\hat{\alpha}_f^L$	**0.531**	**0.945**	**0.871**	0.437	**0.562**
	(0.166)	(0.125)	(0.225)	(0.226)	(0.090)
$\hat{\varrho}_f$	**0.577**	**0.475**	**0.657**	**0.777**	−0.136
	(0.094)	(0.088)	(0.174)	(0.118)	(0.138)
Hansen J p-value	0.50	0.41	0.35	0.10	0.19
Differenced block p-value	0.34	0.19	0.36	0.07	0.20
Levels block p-value	0.66	0.82	0.36	0.37	0.29

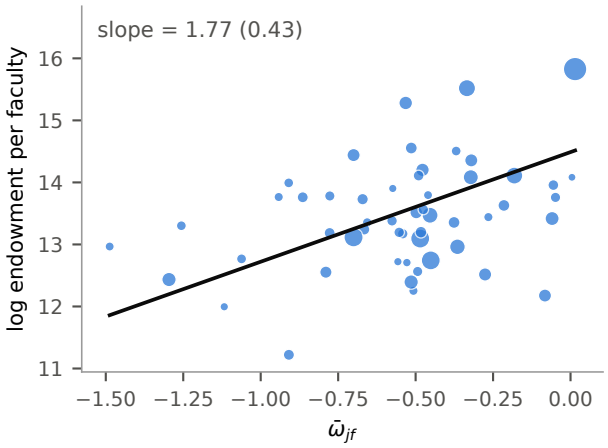
Notes: System GMM estimates of (6.2) by field, with $(\alpha_f^K, \alpha_f^L, \varrho_f)$ recovered by minimum distance from the unrestricted coefficients and the field-level capital allocation iterated to a fixed point. Output is the credit-weighted count of GPU-topic papers (Section 3). Calibration: $\sigma_K = 0.67$, $\sigma_L = 1$, $\gamma_f = (0.02, 0.75, 0.88, 0.75, 0.47)$ from field-level usage data (Table A.4), and $\phi_f = (0.68, 0.71, 0.61, 0.80, 0.57)$, the flexible-labor share of the research wage bill (Table 5). Standard errors in parentheses are clustered by university and condition on the calibrated nests and the allocation. Bold: at least 1.96 standard errors from zero. The elasticities are restricted to $[0, 1]$ and ϱ_f to $[-0.2, 1.05]$. Hansen J tests the overidentifying restrictions of the system and of the differenced equation alone. The difference-in-Hansen test is for the instruments of the levels equation.

Productivity and Financial Constraints. Estimated productivity is positively associated with endowment per faculty in all fields except engineering, with total endowment giving similar or stronger associations (Appendix Figure A.8). We recover $\bar{\omega}_{jf}$ as the mean of $y_{jft} - \hat{\alpha}_f^K k_{jf,t-1} - \hat{\alpha}_f^L \ell_{jft}$, so it absorbs everything that raises output per unit of measured input. This includes research quality and topic mix, as well as computing access outside the central cluster and measurement error in output and labor. Figure 9 plots log endowment per faculty against this residual (which is already log productivity), weighting universities by their number of authors in the field. The fitted slopes are 0.51 (0.28) in computer science, 1.77 (0.43) in geosciences and physics, 0.31 (0.25) in molecular and materials sciences, and 0.80 (0.23) in the life sciences. Engineering is the exception, at −0.20 (0.41). Only the geosciences and physics and life-science slopes are significant at 1%. This highlights the main tradeoff between allocating grants to more productive or less-resourced universities.

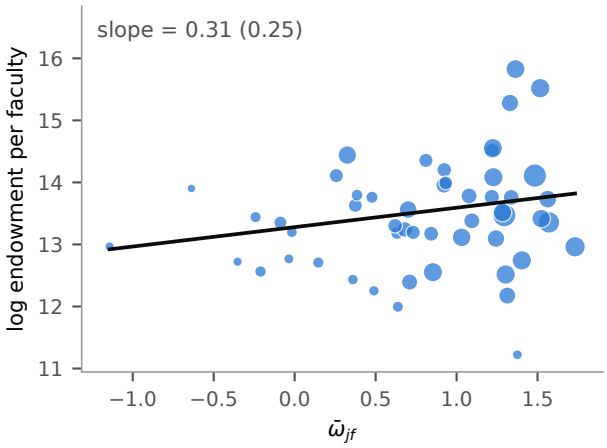


(a) Computer science

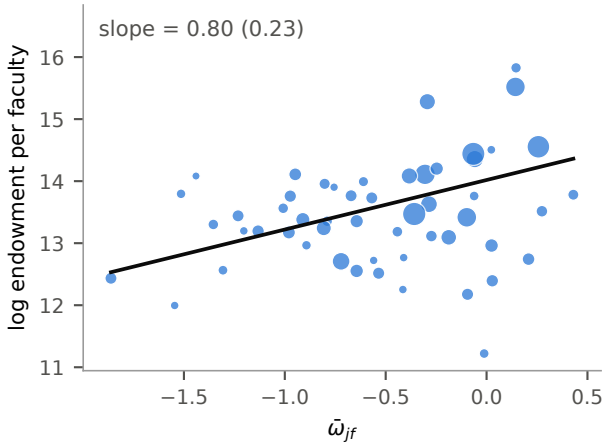
(b) Engineering



(c) Geosciences and physics



(d) Molecular and materials sciences



(e) Life sciences

Notes: Each point is a university. Its area is proportional to the university's mean number of authors publishing in the field over 2005–2023, with each author counted by the share of that year's papers that fall in the field. The horizontal axis is the university-field productivity effect $\bar{\omega}_{jf}$ implied by the estimates of Table 6: the university's mean over 2005–2023 of log GPU-topic output net of the estimated contributions of capital and labor, in log points. The vertical axis is the log of endowment per tenure-line faculty member (FY2012–15 mean). The line is the least-squares fit of log endowment per faculty on $\bar{\omega}_{jf}$ with the same weights as the points, and each panel reports its slope with the standard error in parentheses. Appendix Figure A.8 repeats the figure with total endowment.

Figure 9: Productivity fixed effects and endowment per faculty, by field

7 Crowd-Out of University Investment

A grant's pass-through, m_j, is the share that becomes additional computing capital after the university adjusts its own purchases. Rare and non-random NSF awards make this response difficult to estimate directly. We instead derive pass-through from the production model, infer investment regimes from observed behavior, and test the resulting predictions against hardware purchases and R&D spending.

7.1 Model-Implied Crowd-Out

Let I_{jt}^K denote university j's own spending on computing capital and N_{jt} its NSF computing award, both in dollars. We work with total computing capital (CPUs and GPUs together) rather than GPUs alone since computing infrastructure awards buy both together.

We apply the three regimes in Table 3 under the user-cost approximation derived in Appendix A7.1. A university that does not buy computing capital of its own is in the corner, and the grant is fully additional. An unconstrained university displaces the grant one-for-one. A constrained university that buys computing capital treats the grant as an addition to its budget, $\widetilde{B}_{jt} = B_{jt} + N_{jt}$, and spends a share χ_{jt}^K of the marginal dollar on computing capital, so that $\partial I_{jt}^K / \partial N_{jt} = -(1 - \chi_{jt}^K)$.

The same first-order conditions that deliver the allocation rule pin down χ_{jt}^K. Within a field, the university equates the return to a dollar spent on the cluster with the return to a dollar spent on PhD students and postdocs. Under Cobb–Douglas, this implies that spending on the two inputs is in the ratio of their output elasticities (scaling flexible labor by the labor-nest distribution parameter):

$$\frac{E_{jft}^K}{E_{jft}^L} = \frac{\alpha_f^K}{\phi_f \, \alpha_f^L} \tag{7.1}$$

where E_{jft}^K is the user cost of the cluster capital allocated to field f and $E_{jft}^L = W_{jt}^V L_{jft}^V$ is spending on flexible labor. The value weight λ_{jf} on the field, its productivity, and the tightness of the budget all cancel, exactly as in the allocation rule. We observe L_{jft}^V, so (7.1) reveals each field's cluster spending. Within a field, a constrained university therefore puts the share $\alpha_f^K / \left(\alpha_f^K + \phi_f \alpha_f^L\right)$ of a marginal dollar into the cluster. The condition is the stationary form of the university's dynamic investment problem. When the budget multiplier is stable over time, the first-order condition for capital collapses to a user-cost condition in which the annual cost of holding a unit of cluster, $(r + \delta)P^K$, plays the role of the wage (Appendix A7.1). The user cost is the interest forgone plus the depreciation on the purchase price, so (7.1) compares the annual cost of the field's cluster capital with its annual flexible-labor bill.

Across fields, the university puts a marginal dollar where its return falls most slowly. Because returns to inputs are decreasing in every field, the university's cross-field problem has a unique solution. The share of the marginal dollar going to field f is proportional to the field's current spending scaled by $1/\left(1-\alpha_f^K-\phi_f\alpha_f^L\right)$. A field whose returns are closer to constant therefore absorbs more of the extra dollar. Combining the within-field and cross-field splits (details in Appendix A7),

$$\chi_{jt}^K = \frac{\sum_f \alpha_f^K\, w_{jft}}{\sum_f \left(\alpha_f^K+\phi_f\alpha_f^L\right)\, w_{jft}}, \qquad w_{jft} \equiv \frac{L_{jft}^V}{\phi_f\,\alpha_f^L\,\left(1-\alpha_f^K-\phi_f\alpha_f^L\right)} \tag{7.2}$$

Equation (7.2) averages the within-field capital shares, $\alpha_f^K/(\alpha_f^K+\phi_f\alpha_f^L)$, using each field's share of the marginal budget dollar. Fields with more flexible labor and more slowly diminishing returns receive greater weight. The rule depends on the output elasticities, calibrated labor shares, and university-specific field composition. Faculty remain fixed over the horizon used to measure crowd-out.

At the baseline production function estimates in Table 6, the within-field capital share ranges from 0.04 in engineering to 0.14 in the life sciences, with computer science at 0.14 (Appendix Table A.10). Engineering and the life sciences absorb about seven-tenths of a marginal dollar at the typical university, primarily because they support the most PhD students and postdocs. Engineering also has slowly diminishing returns and the lowest capital share. The capital share of a marginal dollar is therefore small: across R1 universities, χ_j^K averages 0.097, running from 0.073 at the 10th percentile to 0.128 at the 90th, and is relatively time-invariant. Because the flexible-labor counts come from the NSF GSS, which covers every research university, we can compute χ_{jt}^K for all R1 universities and not only for the capacity panel. A constrained university puts about ten cents of an extra dollar into its cluster and the rest into people. Differences in crowd-out across universities come almost entirely from which regime they are in.

7.2 Assigning Regimes

We do not observe a university's regime directly, so we infer it from its investment behavior. The three regimes differ in the capital investment response to an NSF award and the university's budget.

Regime	$\partial I^K/\partial N$	$\partial I^K/\partial B$	Adds capacity?
Corner	0	0	No
Constrained interior	$-(1-\chi_{jt}^K)$	χ_{jt}^K	Yes
Unconstrained	-1	0	Yes

We use recent capacity additions to proxy for whether a university buys computing capital of its own, or whether it is in the corner regime. The classification is at the university-year level: university j is in the corner in year t if it added no CPU cores and no GPU capacity in either of the two preceding years. Over 2019–2024, 21.4 percent of panel university-years are in the corner, and the corner is strongly related to resources. A logit of the corner indicator on log endowment per tenure-line faculty gives a coefficient of -0.757 (0.247), or an average marginal effect of -0.119 per log point (Table 7, Panel A). For the R1 universities outside the capacity panel, whose capacity we do not observe, we use this relationship to predict the probability $\widehat{P}_j$ that the university is in the corner from its endowment per faculty. We assume that no university is unconstrained (Appendix A7.3 reports the budget-response test behind this assumption).

University-level crowd-out. Combining the regimes, the implied response of a panel university j in year t is

$$\frac{\partial I^K_{jt}}{\partial N_{jt}} = \begin{cases} 0 & \text{if } j \text{ is in the corner in year } t \\ -\left(1-\chi^K_{jt}\right) & \text{otherwise} \end{cases} \tag{7.3}$$

where $m_{jt} \equiv 1 + \left(\partial I^K_{jt}/\partial N_{jt}\right)$ is the pass-through of a grant dollar into new computing capital. Corner status is proxied by capacity additions and χ^K_{jt} is computed from estimated and calibrated parameters. For an R1 university whose capacity we do not observe, the corner indicator is replaced by its predicted probability:

$$\frac{\partial I^K_{jt}}{\partial N_{jt}} = -\left(1-\widehat{P}_j\right)\left(1-\chi^K_{jt}\right), \qquad m_{jt} = \widehat{P}_j + \left(1-\widehat{P}_j\right)\chi^K_{jt} \tag{7.4}$$

A university that is investing passes through χ^K_{jt}, about ten cents of each grant dollar, and a university in the corner passes through the whole dollar. The expected pass-through of a university outside the panel weights the two by its corner probability.

Crowd-out rises with endowment, and the gradient comes entirely from the corner (Table 7, Panel B). Across the 187 R1 universities, the predicted corner probability runs from 0.055 at the 10th percentile to 0.352 at the 90th, so a grant dollar displaces between 58 and 85 cents of own spending, 72 cents on average. Pass-through averages 0.275 and falls by about 6 cents for every doubling of endowment per faculty (slope -0.059, standard error 0.002). This result highlights the second piece of the allocation tradeoff: while financial resources are positively associated with productivity in several fields, they are also negatively associated with predicted pass-through.

Table 7: Model-Implied Crowd-Out across Universities

Panel A: corner regime and endowment per faculty (logit, 53 panel universities, 2019–2024)				
Log endowment per tenure-line faculty	−0.757 (0.247)			
Average marginal effect on Pr(Corner)	−0.119 (0.035)			
Corner share in the sample	0.214			
Pseudo-R^2; university-years; universities	0.060; 290; 53			
Panel B: implied values across 187 R1 universities	Mean	10th pct.	Median	90th pct.
Capital share of marginal dollar, χ_j^K	0.097	0.073	0.095	0.128
Corner probability, $\widehat{P}_j$	0.197	0.055	0.190	0.352
Implied crowd-out, $\partial I^K/\partial N$	−0.725	−0.852	−0.734	−0.583
Pass-through, m_j	0.275	0.148	0.266	0.417
Crowd-out per doubling of endowment per faculty	−0.059 (0.002)			

Notes: Panel A: logit of the corner indicator (no CPU cores and no GPU capacity added in the two preceding years) on log endowment per tenured and tenure-track faculty member (FY2012–15 mean) over 2019–2024 university-years of the 53 panel universities; standard errors clustered by university. Panel B: χ_j^K from (7.2) with flexible labor by field from the NSF GSS, averaged over 2013–2023; $\widehat{P}_j$ is the Panel A prediction at the university's endowment per faculty; crowd-out and pass-through from (7.4). The last row is the slope of the implied crowd-out on log endowment per faculty across R1 universities, times $\ln 2$.

7.3 Testing the Implied Crowd-Out

The hardware regressions in Section 4.3 suggest that own spending declines after an award. The model predicts a university-year response, zero in the corner and $-(1-\chi_{jt}^K)$ otherwise. We combine this with each university's computing awards to generate a predicted investment response $\widehat{I}_{jt}$:

$$\widehat{I}_{jt} = \left(\frac{\partial I^K}{\partial N}\right)_{jt} N_{jt}$$

where N_{jt} is the dollar value of MRI and CC* computing awards made in year t. For years without a corner indicator, we substitute the university's 2019–2024 corner share. We then estimate

$$I_{jt} = \alpha_j + \delta_t + \sum_{\ell=0}^{2} \lambda_\ell \widehat{I}_{j,t-\ell} + \gamma \log(\text{Endowment})_{jt} + e_{jt} \tag{7.5}$$

where I_{jt} is the university's own hardware spending from Section 4.3: the dollar value of the CPU cores and GPU TFLOPS it adds between consecutive capacity snapshots, net of the award dollars whose hardware arrives in that year. If the model's crowd-out is right, the cumulative response $\sum_\ell \lambda_\ell$ equals one. If grants are fully additional, it equals zero. We sum over the award year and the next two, the period in which the money is spent.

When awards are dated by award year, the data do not reject the model's crowd-out and reject full additionality (Table 8; Figure 10 plots the estimates). The cumulative response to the implied crowd-out is 1.26 (0.42), which does not reject one but does reject zero. Replacing each university's observed corner status by the probability predicted from its endowment per faculty,

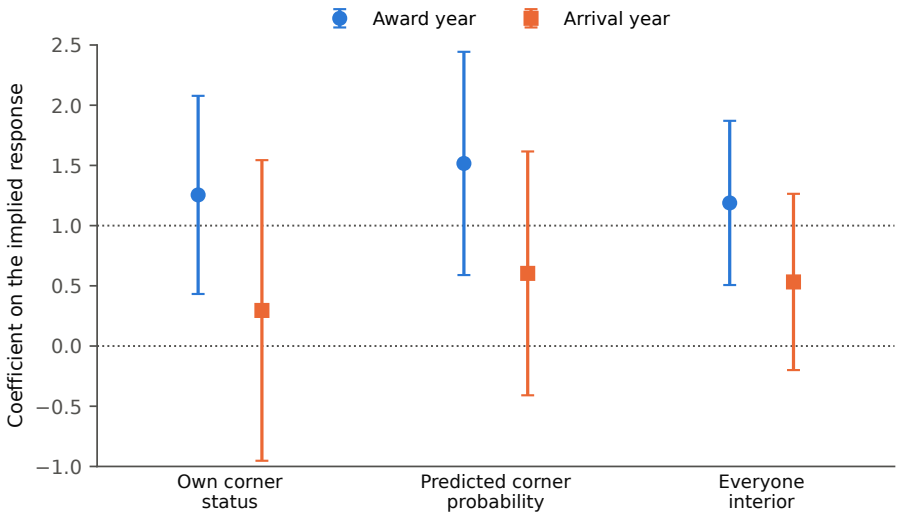
(7.4), gives 1.52 (0.47), and treating every university as interior gives 1.19 (0.35). However, three caveats apply. With 36 award-years, the test cannot separate the model from considerably smaller crowd-out, and the result depends on how the award is dated. If awards are dated by the arrival of the hardware rather than the award year, the response is 0.30 (0.64). And award dollars that never become computing hardware, such as staff, storage, and networking, push the measured response toward one mechanically.

Appendix A7.4 repeats the test on all 186 R1 universities, with institutionally financed R&D spending by field from the NSF HERD Survey as the outcome and the predicted corner probability in place of the observed status. There are four times the award dollars but a proxy outcome. The institutional-funds outcome does not reject the model: in computer science, the cumulative response of institutional funds is 0.73 (0.60), and the placebo estimates are indistinguishable from zero. Equipment outcomes give mixed results, including rejections in computer science and engineering.

Table 8: Crowd-Out in the Panel's Own Hardware Purchases

	Award year	Arrival year
Cumulative response of own hardware spending, $\partial I^K/\partial N$	−1.08 (0.32)	−0.48 (0.34)
Cumulative response of total hardware spending, m	−0.07 (0.27)	0.52 (0.34)
$\sum_\ell \hat{\lambda}_\ell$ on the implied response (H_0: $= 1$)		
Own corner status $\times\chi^K_{jt}$	1.26 (0.42)	0.30 (0.64)
Predicted corner probability $\times\chi^K_{jt}$	1.52 (0.47)	0.60 (0.52)
Everyone interior, $-(1-\chi^K_{jt})$	1.19 (0.35)	0.53 (0.37)
Mean implied crowd-out per award dollar	−0.74	−0.64
University-years; award-years; award dollars ($M)	345; 36; 33.8	345; 33; 26.1

Notes: The 53 panel universities, flow years 2017–2023. Own hardware spending is the dollar value of the CPU cores and GPU TFLOPS added between consecutive snapshots (GPUs at the GovSpend median purchase-order price per FP16 TFLOPS for datacenter GPUs in the arrival year, cores at the GovSpend median of $84) minus the NSF award dollars whose hardware arrives in that flow year (the award year plus one, or the linked capacity jump). Award dollars enter with lags 0–2, dated by the award (obligation) year in the first column, when the money is spent over the following two years, and by the arrival year in the second. Regressions include university and year fixed effects and log endowment; standard errors in parentheses are clustered by university. The implied response uses the university's own corner status (its 2019–24 corner share where the status is unobserved), the corner probability predicted from endowment per faculty, or no corner at all.



Notes: Estimates of $\sum_\ell \hat{\lambda}_\ell$ from Table 8, with 95% confidence intervals from standard errors clustered by university. Under the null hypothesis that the model's crowd-out is correct, the cumulative response equals one; under full additionality, it equals zero (dotted lines). Awards are dated by the award year (blue) or by the arrival of the hardware (orange). The three specifications of the implied response are those of Table 8.

Figure 10: Cumulative Response of Own Hardware Spending to the Implied Crowd-Out

8 Counterfactual Allocations of NSF Computing Grants

We combine capital productivity and grant pass-through to compare the observed NSF allocation with an output-maximizing benchmark. The benchmark holds labor and productivity fixed, so it measures publication gains only through additional computing capital. While the observed NSF allocation achieves lower gains than the output-maximizing benchmark, it is more geographically diverse, in line with the NSF's mandate of broad research access.

8.1 The Planner's Problem

We model the NSF as a planner allocating a fixed budget $\bar{B}_{\text{NSF}}$ of computing grants $N_j \geq 0$ across universities j. A university that receives N_j dollars adds $m_j N_j$ dollars of computing capital on net, where m_j is its pass-through from (7.3) evaluated at the baseline year. The university splits those dollars between CPU cores and GPU TFLOPS to maximize its output:

$$\mathcal{Y}_j(N_j) \equiv \max_{\Delta C_j,\, \Delta G_j} \widehat{Y}_{jt^*}(C_{jt^*} + \Delta C_j,\; G_{jt^*} + \Delta G_j) \quad \text{s.t.} \quad p_C\, \Delta C_j + p_G\, \Delta G_j = m_j\, N_j \tag{8.1}$$

where $\widehat{Y}_{jt^*}$ is predicted output summed across fields at baseline year t^*. The planner maximizes an Atkinson social welfare function over university output:

$$\max_{\{N_j\}} W(\rho) = \sum_{j=1}^{J} \frac{\mathcal{Y}_j(N_j)^{1-\rho}}{1-\rho} \qquad \text{s.t.} \qquad \sum_j N_j \leq \bar{B}_{\text{NSF}} \tag{8.2}$$

The parameter $\rho \geq 0$ governs inequality aversion. When $\rho = 0$, the planner maximizes total output. When $\rho = 1$, it maximizes the sum of log output and values proportional gains equally across universities. As $\rho \to \infty$, the planner maximizes the output of the lowest-producing university.

For every funded university, the planner's first-order condition is

$$\mathcal{Y}_j(N_j)^{-\rho} \cdot m_j \cdot MPK_j^{\$} = \mu, \qquad MPK_j^{\$} \equiv \max\left\{\frac{1}{p_G}\frac{\partial \widehat{Y}_j}{\partial G_j}, \frac{1}{p_C}\frac{\partial \widehat{Y}_j}{\partial C_j}\right\} \tag{8.3}$$

where μ is the shadow value of the NSF's budget and $MPK_j^{\$}$ is the output from one more dollar of computing capital, spent on whichever input has the higher return. Thus, an efficient planner ($\rho = 0$) equalizes $m_j \cdot MPK_j^{\$}$, the return to a grant dollar *after* crowd-out, across funded universities. A university with high productivity but low pass-through can receive less than a lower-productivity university where grants are more additional.

8.2 Reallocating the Observed Budget

We reallocate the MRI and CC* computing awards made to universities in our estimation panel in FY2021–2025, a total of $30.5 million across 22 universities, taking 2021 as the baseline year. We hold labor, field composition, and productivity fixed at their baseline values, and productivity is backed out so that predicted output equals observed output in 2021. Within each university, the new capital is reallocated across fields through the allocation rule (5.11). Prices are median purchase-order prices from GovSpend: $84 per CPU core and $38 per FP16 TFLOPS. We compute the efficient allocation by distributing grants until crowd-out-adjusted marginal products of capital ($m_j \cdot MPK_j^{\$}$) are equated across funded universities (implementation details in Appendix A8). To isolate the role of crowd-out, we report each exercise both with full ($m_j = 1$) and model-implied pass-through. The model-implied m_j is (7.3) evaluated in 2021: one for the 10 universities that were in the corner in that year, having added no computing capacity in the two preceding years, and χ_{jt}^K, about 0.1, for the rest.

Crowd-out lowers the gains under both allocations and reduces the share of the attainable gain captured by the observed allocation. Table 9 shows that without crowd-out, the observed allocation raises predicted GPU-topic output by 397 papers, while the efficient allocation of the same budget raises it by 832, so the observed allocation captures 48% of the attainable gain. Under the model-implied crowd-out, the observed allocation adds only 171 papers. The efficient allocation adds 440 papers, and the share of the attainable gain that the NSF captures falls to 39%. The efficient planner funds fewer universities under crowd-out and concentrates the budget on those in the corner, where a grant is fully additional: 9 of the 22 universities it funds are corners.

Table 9: Counterfactual NSF Allocation: Observed versus Efficient

	No crowd-out	Model-implied crowd-out
Gain, observed allocation (papers)	+397	+171
Gain, efficient allocation (papers)	+832	+440
Papers per $M (observed / efficient)	13.0 / 27.3	5.6 / 14.4
Share of attainable gain captured	48%	39%
Universities funded (observed / efficient)	22 / 44	22 / 22
GPU share of dollars (observed / efficient)	42% / 67%	53% / 65%

Notes: Reallocation of FY2021–2025 MRI and CC* computing awards at panel universities ($30.5M across 22 universities), baseline year 2021, predicted no-grant output of 14,224 GPU-topic papers at 53 universities. "No crowd-out" sets $m_j = 1$; "model-implied" uses m_j from (7.3) evaluated in 2021: one for the 10 universities in the corner and χ^K_{jt} for the rest. Each university splits its dollars between CPU cores and GPU TFLOPS to maximize its output. Labor, field composition, and productivity are held at baseline. Awards at universities outside the estimation panel ($59.8M) are excluded.

Figure 11 decomposes the gains by field. Computer science has the largest proportional gain, 10.4% under the efficient allocation without crowd-out and 5.2% with it, against 3.9% and 1.6% under the observed allocation. Its capital elasticity is the highest and its cluster use is almost entirely GPU-based. The life sciences follow at 5.9% and 3.3%, then the molecular and materials sciences (4.3% and 2.4%) and the geosciences and physics (3.3% and 1.9%). Engineering, whose capital elasticity is the smallest, gains 2.1% and 1.2%. Across all fields, the efficient allocation raises GPU-topic output by 5.9% without crowd-out and 3.1% with it, against 2.8% and 1.2% under the observed allocation (Appendix Table A.13 reports the levels).

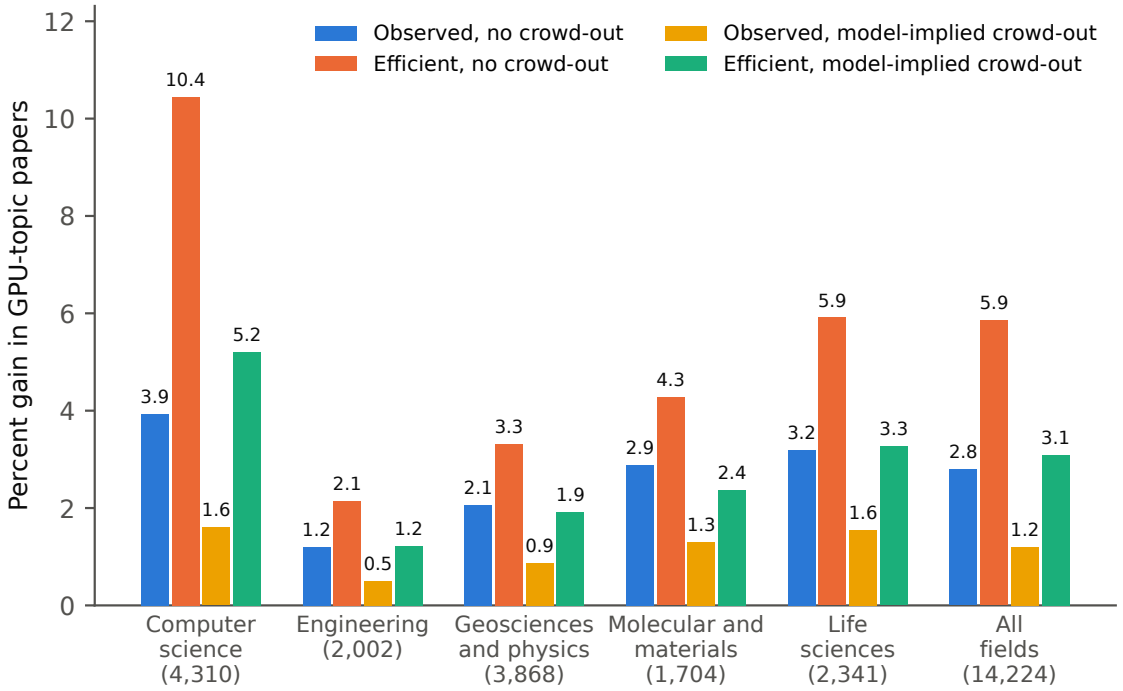


Figure 11: Gain in GPU-Topic Papers by Field. Percent change in predicted GPU-topic papers relative to the 2021 baseline under the observed NSF allocation and the efficient allocation, with full pass-through ($m_j = 1$) and with the model-implied pass-through from (7.3). Baseline counts in parentheses. Field output is re-solved through the allocation rule (5.11) at each university's counterfactual capital stock.

Figure 12 shows that the efficient allocation moves dollars from the middle of the endowment distribution toward both tails. It ranks the 53 panel universities by endowment per tenure-line faculty and plots the kernel density of grant dollars over the ranks under the observed allocation and the two efficient ones. The NSF's dollars sit in the middle of the distribution: the dollar-weighted mean rank of an award is the 50th percentile. 42% of the budget goes to universities below the median and 11% to the top quarter. Universities between the 25th and 75th percentiles receive 74% of the observed budget, against 35% and 25% under the efficient allocations without and with crowd-out. The efficient allocation without crowd-out moves dollars to both tails, 50% below the median and 35% in the top quarter, where small computing stocks and high productivity make the marginal product of capital high. The model-implied crowd-out tilts the allocation down: 60% of the budget goes below the median and the mean rank falls to percentile 43, because the universities in the corner, where a grant is fully additional, are concentrated among the least endowed. Ranking universities by total endowment instead (Appendix Figure A.9) paints a similar picture of shifting dollars from the middle of the distribution toward the tails.

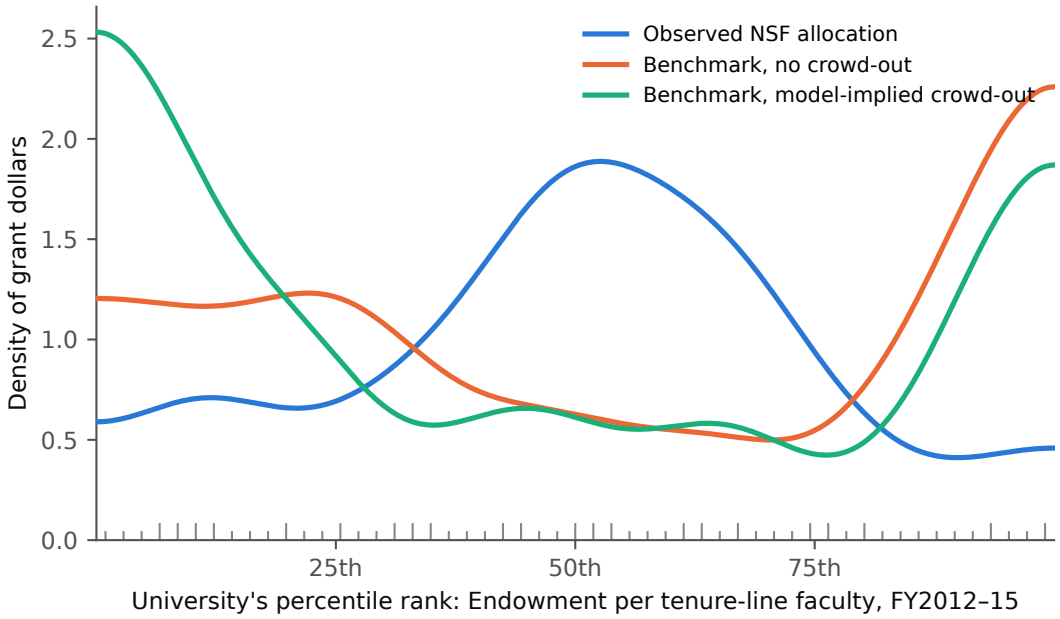


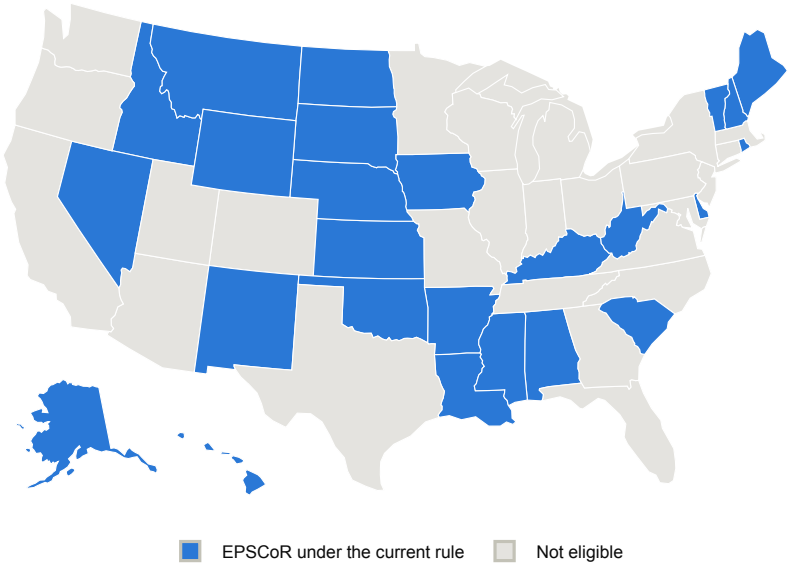
Figure 12: Where the Grant Dollars Go. Dollar-weighted kernel density of the universities' percentile rank on endowment per tenured and tenure-track faculty member (FY2012–15 mean) under the observed NSF allocation and the efficient allocations with full and with model-implied pass-through, FY2021–2025 MRI and CC* computing awards at the 53 panel universities ($30.5M). Ticks on the horizontal axis mark the universities, taller for those that received an award. Gaussian kernel with bandwidth 0.08 on the unit rank axis.

The output-maximizing allocation also favors universities with higher estimated productivity (Appendix Figure A.10). On a university-level index of the production function's productivity fixed effects, the dollar-weighted mean rank is percentile 46 under the observed allocation and percentile 70 under the efficient allocation with crowd-out, which puts 52% of the budget in the most productive quarter of universities against the observed 22%. Productivity is only modestly related to endowment per faculty in the panel (rank correlation 0.36), which is why the planner can favor productive universities and less-resourced ones at the same time.

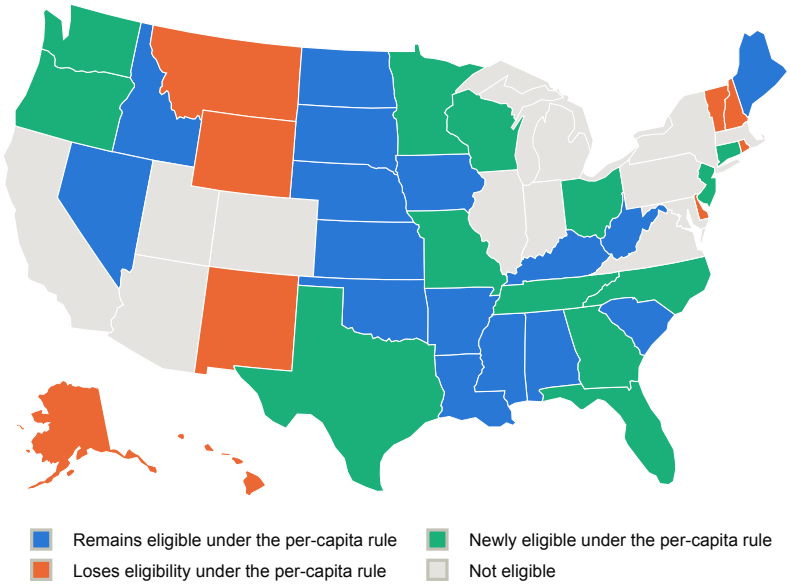
8.3 Geographic Dispersion of NSF Funding

The NSF has a statutory mandate to spread research funding across states. The Established Program to Stimulate Competitive Research (EPSCoR) was created at the NSF in 1978 in response to congressional concern that federal research dollars were concentrated in a few states (Harris 2017). A jurisdiction is eligible when it receives less than 0.75 percent of the NSF's research funding over the last five years (National Science Foundation 2026). Today, 25 states qualify (Figure 13, panel a). Because the criterion is a state's share of NSF funding, a populous state with many research universities cannot qualify however little the funding per capita. Mervis (2017) describes a proposal

by Representative Bill Foster to base eligibility on federal research funding per capita instead (Figure 13, panel b). The proposal has not been enacted, and we use it as a second, population-adjusted definition of the states the geographic mandate is meant to reach.



(a) Current rule: NSF EPSCoR jurisdictions, FY2022–2027



(b) Proposed per-capita rule

Figure 13: EPSCoR Jurisdictions under the Current Rule and the Per-Capita Proposal. Panel (a) shows the 28 jurisdictions eligible for the NSF's EPSCoR program in FY2022–2027. Panel (b) shows eligibility under the proposed rule that bases eligibility on federal research funding per capita (Mervis 2017). The map shows states that remain eligible (blue), the 13 jurisdictions that would become eligible (green), and the 10 that would lose eligibility (orange).

Figure 14 compares the share of the budget that reaches universities in EPSCoR states under the observed and the efficient allocations. Universities in current EPSCoR states are 16 of the 53 panel universities and produce 16% of baseline GPU-topic output. The observed NSF allocation sends them 43% of its dollars while the efficient allocation only sends them 10% without crowd-out and 10% with it. The observed allocation is thus about four times as concentrated in EPSCoR states as the output-maximizing one, and the dispersion is not confined to states at the margin of eligibility. Universities in the 12 states that qualify under both the current rule and the per-capita proposal, states with little federal research funding by either measure, receive 31% of the observed budget ($9.3M) against 9% under the efficient allocation. Appendix Figure A.11 shows that the pattern is weaker in counts than in dollars: the NSF funds EPSCoR-state universities at roughly their share of the panel, but gives them larger award amounts.

The per-capita definition changes the efficient allocations' EPSCoR shares by 21 and 25 percentage points, to 31% and 35%, while the observed share is essentially unchanged at 44%. The states that would gain eligibility hold 13 panel universities and 27% of baseline output, and the efficient allocations direct 22% and 26% of the budget to them.

There are caveats to keep in mind for these comparisons since our output measure is GPU-topic publications over a short horizon with labor and productivity held fixed. The geographic mandate reflects objectives that this measure does not capture: building research capacity over the long run, local economic spillovers (Hausman 2022), training scientists who remain in the region, and equitable access to federal research infrastructure. However, the comparison does highlight the publication cost of pursuing those objectives through computing grants. The observed allocation captures 39% to 48% of the attainable output gain. Under the current definition, it places about four times the efficient allocation's share of the budget in EPSCoR states. Under the per-capita definition, the gap narrows to 44% against 31% and 35%, because the states that would become eligible are ones where the efficient allocation also spends.

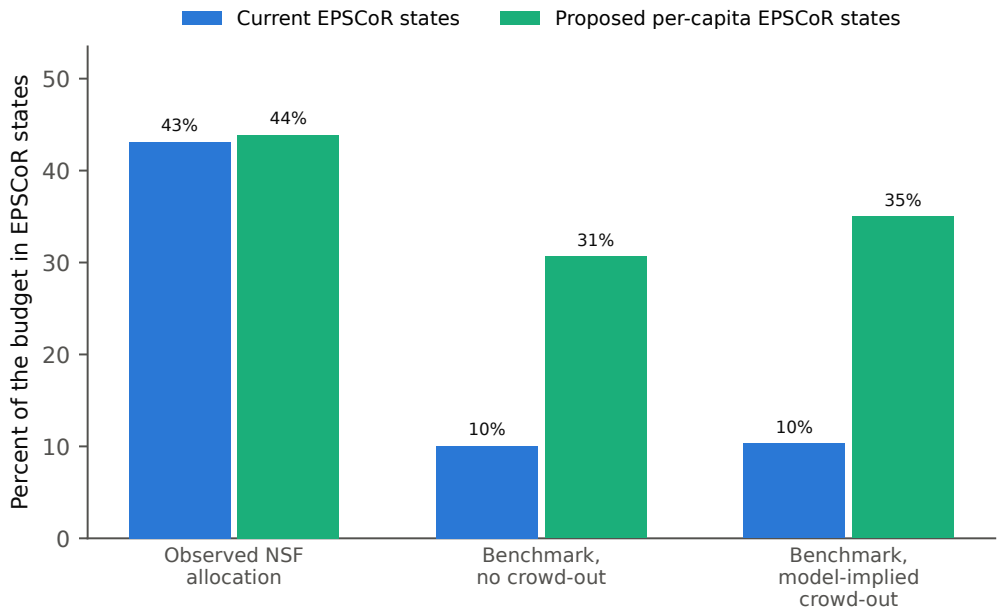


Figure 14: Share of Computing Grant Dollars Reaching EPSCoR States. Percent of the FY2021–2025 MRI and CC* budget at panel universities ($30.5M) that goes to universities in EPSCoR states under the observed NSF allocation and the efficient allocations with full and with model-implied pass-through, using the current eligibility list and the proposed per-capita rule of Figure 13. Appendix Figure A.11 shows the share of funded universities instead of dollars.

8.4 The NSF's Revealed Inequality Aversion

If the observed NSF allocation is optimal under some ρ, the first-order condition (8.3) holds at the observed allocation. Taking logs gives a revealed-preference condition for funded universities:

$$\ln\left(m_j \, MPK_j^{\$,\text{post}}\right) = a + \rho \, \ln\left(\widehat{Y}_j^{\text{base}}\right) + u_j \tag{8.4}$$

where $MPK_j^{\$,\text{post}}$ is the dollar marginal product after the observed grant and $\widehat{Y}_j^{\text{base}}$ is predicted output before it. The constant a is the log of the shadow value of the NSF's budget, $\ln \mu$, which is common to all funded universities. If the NSF maximizes output, returns are equalized and $\rho = 0$. If $\rho > 0$, the NSF accepts lower returns at lower-output universities.

Interpreting the fitted slope as a preference parameter requires the residual u_j to be uncorrelated with log baseline output among funded universities. In the model, u_j collects three things: objectives other than GPU-topic output that enter the NSF's decision, such as geography, training, and capacity building; frictions between the planner's problem and realized awards, including who applied and the noise of review; and error in our estimate of the marginal return, which rests on estimated elasticities, calibrated weights, and backed-out productivity. The first-order condition restricts none of them, so the slope is a preference parameter only if these components are unrelated to

a university's output. We therefore read (8.4) as a descriptive summary of the observed allocation. Baseline output stands in for post-grant output on the right-hand side, and the difference is below ten percent at every recipient.

Accounting for crowd-out lowers the point estimate of the NSF's revealed inequality aversion, though the two estimates are not statistically distinguishable. Estimating (8.4) gives $\hat{\rho} = 0.76$ (0.26) without crowd-out and $\hat{\rho} = 0.65$ (0.40) with the model-implied pass-through. Thus, part of what looks like a preference for spreading grants across lower-output universities is explained by the fact that grants are more additional there. Note that the estimate with crowd-out is imprecise and does not reject $\rho = 0$, an output-maximizing NSF. We interpret ρ as a reduced-form summary of the agency's allocation rule rather than a structural welfare weight. The NSF may value geographic dispersion (Section 8.3), university capacity building, and researcher training, or face political constraints, none of which our output measure captures.

9 Conclusion

Given the upfront costs of computing capital and the growing demand for it in science production, federal funding plays an important role. However, it is unclear ex ante where federal infrastructure grants have the highest marginal return. Universities with higher research productivity also tend to be the ones that can fund capital through other channels. We study this tradeoff between productivity and financial constraints by constructing a new panel of university central CPU and GPU capacity linked to publications and researchers. Using large, discrete GPU capacity increases in an event-study framework, we find that AI and deep-learning output rises after large GPU upgrades at less-resourced universities. Furthermore, we find that research labor also moves with capacity upgrades: graduate students and postdocs increase in computer science while the faculty stock does not.

Since both capital and labor contribute to research output, we decompose the effect of computing capital versus labor in a structural production function that treats fields as separate research outputs using a shared university cluster. Estimated capital elasticities range from 0.03 to 0.06. Under the model's investment rule, a constrained university allocates about ten cents of an additional research-budget dollar to computing capital. Grant pass-through is therefore higher at universities that are not buying computing capital of their own. We infer this regime from the lack of recent capacity additions and find greater prevalence among less-endowed universities.

Armed with both productivity estimates and model-implied crowd-out parameters, we can simulate counterfactual allocations for a federal funder of infrastructure, in this case the NSF. We hold labor and productivity fixed to determine the publication output-maximizing allocation. The observed NSF portfolio captures 48% of the output gain attainable with the same budget under no crowd-out and 39% under model-implied crowd-out. The output-maximizing allocation moves funding from the middle of the endowment per faculty distribution toward both tails. Accounting for crowd-out shifts more funding toward universities not already investing in computing, most of them among the least endowed. Productive research capacity and grant additionality can thus favor the same recipients. The observed portfolio is also far more geographically dispersed than either efficient allocation: universities in EPSCoR states receive 43% of the observed allocation dollars against 10% of efficient allocation dollars. However, changing the criteria for EPSCoR to a per-capita threshold increases the share of the efficient allocation in EPSCoR states to 35%, which is much closer to that of the observed allocation.

There are three main limitations to this empirical exercise. First, grants can release funds for hiring and other research inputs, which are held fixed in this counterfactual. The event studies show that students, postdocs, and in some fields faculty move with an upgrade, so a reduced-form publication response bundles the capital with the labor that arrives alongside it. This is why the production function, which treats computing capital, flexible labor, and quasi-fixed faculty as separate inputs, is needed to recover the return to capital alone. Second, crowd-out tests have limited power and depend on award timing and expenditure measurement. The investment regimes are inferred proxies rather than solutions to the full dynamic problem. Finally, publication counts over a short horizon omit the long-run objectives of infrastructure funding: building long-term research capabilities, generating local economic spillovers, and educating the next generation of scientists. These are important considerations that will be left to future work.

References

Advanced Micro Devices (2021). AMD Instinct MI200 series accelerator datasheet. https://www.amd.com/content/dam/amd/en/documents/instinct-tech-docs/instinct-mi200-datasheet.pdf.

Ahmed, N. and M. Wahed (2020). The de-democratization of AI: Deep learning and the compute divide in artificial intelligence research. *arXiv:2010.15581*.

Ahmed, N., M. Wahed, and N. C. Thompson (2023). The growing influence of industry in AI research. *Science 379*(6635), 884–886.

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

Arellano, M. and S. Bond (1991). Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations. *Review of Economic Studies 58*(2), 277–297.

Arrow, K. J. (1962). Economic welfare and the allocation of resources for invention. In R. R. Nelson (Ed.), *The Rate and Direction of Inventive Activity: Economic and Social Factors*, pp. 609–626. Princeton, NJ: Princeton University Press.

Azoulay, P., J. S. Graff Zivin, D. Li, and B. N. Sampat (2019). Public R&D investments and private-sector patenting: Evidence from NIH funding rules. *The Review of Economic Studies 86*(1), 117–152.

Baruffaldi, S. and F. Gaessler (2026, January). The returns to physical capital in knowledge production: Evidence from lab disasters. *American Economic Journal: Applied Economics 18*(1), 248–285.

Besiroglu, T., S. A. Bergerson, A. Michael, L. Heim, X. Luo, and N. Thompson (2024). The compute divide in machine learning: A threat to academic contribution and scrutiny? *arXiv:2401.02452*.

Blundell, R. and S. Bond (1998). Initial conditions and moment restrictions in dynamic panel data models. *Journal of Econometrics 87*(1), 115–143.

Cattaneo, M. D., R. K. Crump, M. H. Farrell, and Y. Feng (2024). On binscatter. *American Economic Review 114*(5), 1488–1514.

Cengiz, D., A. Dube, A. Lindner, and B. Zipperer (2019). The effect of minimum wages on low-wage jobs. *Quarterly Journal of Economics 134*(3), 1405–1454.

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

CHIPS and Science Act (2022). Public Law 117-167, Section 10320, waiving cost sharing in the Major Research Instrumentation program for five years; codified at 42 U.S.C. 18998. https://www.law.cornell.edu/uscode/text/42/18998.

Cockburn, I. M., R. Henderson, and S. Stern (2019). The impact of artificial intelligence on innovation: An exploratory analysis. In A. Agrawal, J. Gans, and A. Goldfarb (Eds.), *The Economics of Artificial Intelligence: An Agenda*, pp. 115–146. University of Chicago Press.

David, P. A., B. H. Hall, and A. A. Toole (2000). Is public R&D a complement or substitute for private R&D? A review of the econometric evidence. *Research Policy 29*(4–5), 497–529.

De Loecker, J. and P. K. Goldberg (2014). Firm performance in a global market. *Annual Review of Economics 6*(Volume 6, 2014), 201–227.

De Loecker, J., P. K. Goldberg, A. K. Khandelwal, and N. Pavcnik (2016). Prices, markups, and trade reform. *Econometrica 84*(2), 445–510.

Emeras, J., S. Varrette, V. Plugaru, and P. Bouvry (2019). Amazon elastic compute cloud (EC2) versus in-house HPC platform: A cost analysis. *IEEE Transactions on Cloud Computing 7*(2), 456–468.

Epoch AI (2026). Data on machine learning hardware. https://epoch.ai/data/

machine-learning-hardware. CC-BY 4.0. Accessed July 2026.

Foster, L., J. Haltiwanger, and C. Syverson (2008, March). Reallocation, firm turnover, and efficiency: Selection on productivity or profitability? *American Economic Review 98*(1), 394–425.

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

Ganguli, I. and R. Murciano-Goroff (2025, September). Exposure to science and scientific careers: Evidence from minimum wage increases and university lab employment. Working Paper 34244, National Bureau of Economic Research.

Gofman, M. and Z. Jin (2024). Artificial intelligence, education, and entrepreneurship. *The Journal of Finance 79*(1), 631–667.

Harris, L. A. (2017). Established program to stimulate competitive research (EPSCoR): Background and selected issues. CRS Report R44689, Congressional Research Service.

Hausman, N. (2022). University innovation and local economic growth. *Review of Economics and Statistics 104*(4), 718–735.

Hofstra, B., V. V. Kulkarni, S. Munoz-Najar Galvez, B. He, D. Jurafsky, and D. A. McFarland (2020). The diversity–innovation paradox in science. *Proceedings of the National Academy of Sciences 117*(17), 9284–9291.

Hsieh, C.-T. and P. J. Klenow (2009). Misallocation and manufacturing TFP in China and India. *The Quarterly Journal of Economics 124*(4), 1403–1448.

Jacob, B. A. and L. Lefgren (2011). The impact of research grant funding on scientific productivity. *Journal of Public Economics 95*(9–10), 1168–1177.

Lach, S. (2002). Do R&D subsidies stimulate or displace private R&D? Evidence from Israel. *Journal of Industrial Economics 50*(4), 369–390.

Lanahan, L., A. Graddy-Reed, and M. P. Feldman (2016). The domino effects of federal research funding. *PLoS ONE 11*(6), e0157325.

McKeon, J. and J. Boudou (2026, May). Innovation under resource constraints: Supercomputing in scientific research. SSRN working paper, Harvard Business School. Revised May 2026. Available at SSRN: https://ssrn.com/abstract=4868123.

Mervis, J. (2017). Does your state get its fair share of federal research dollars? *Science*, News, 5 October 2017. Accessed 17 October 2026.

Nagaraj, A., E. Shears, and M. de Vaan (2020). Improving data access democratizes and diversifies science. *Proceedings of the National Academy of Sciences 117*(38), 23490–23498.

NAIRR Task Force (2023). Strengthening and democratizing the U.S. artificial intelligence innovation ecosystem: An implementation plan for a national artificial intelligence research resource. Technical report, National Artificial Intelligence Research Resource Task Force.

National Science Foundation (2012). Campus cyberinfrastructure–network infrastructure and engineering program (CC-NIE). Program Solicitation NSF 12-541, National Science Foundation. https://www.nsf.gov/funding/opportunities/cc-campus-cyberinfrastructure/nsf12-541/solicitation.

National Science Foundation (2015). Major research instrumentation program (MRI): Instrument acquisition or development. Program Solicitation NSF 15-504, National Science Foundation. https://www.nsf.gov/funding/opportunities/mri-major-research-instrumentation-program/nsf15-504/solicitation.

National Science Foundation (2018). Major research instrumentation program (MRI): Instrument acquisition or development. Program Solicitation NSF 18-513, National Science Foundation. https://www.nsf.gov/funding/opportunities/mri-major-research-instrumentation-program/nsf18-513/solicitation.

National Science Foundation (2019). Campus cyberinfrastructure (CC*). Program Solicitation NSF 19-533, National Science Foundation. https://www.nsf.gov/funding/opportunities/

cc-campus-cyberinfrastructure/nsf19-533/solicitation.

National Science Foundation (2021a). Campus cyberinfrastructure (CC*). Program Solicitation NSF 21-528, National Science Foundation. https://www.nsf.gov/funding/opportunities/cc-campus-cyberinfrastructure/nsf21-528/solicitation.

National Science Foundation (2021b). FY 2022 budget request to congress: Directorate for computer and information science and engineering. Technical report, National Science Foundation. https://www.nsf.gov/about/budget/fy2022.

National Science Foundation (2022). NSF ACCESS awardees will advance innovations in cyberinfrastructure accessibility, user support and integration services. Press release. https://www.nsf.gov/news/nsf-access-awardees-will-advance-innovations.

National Science Foundation (2023). Major research instrumentation program: Instrument acquisition or development. Program Solicitation NSF 23-519, National Science Foundation. https://www.nsf.gov/funding/opportunities/mri-major-research-instrumentation-program/nsf23-519/solicitation.

National Science Foundation (2024a). Campus cyberinfrastructure (CC*). Program Solicitation NSF 24-530, National Science Foundation. https://www.nsf.gov/funding/opportunities/cc-campus-cyberinfrastructure.

National Science Foundation (2024b). Democratizing the future of AI R&D: NSF to launch national AI research resource pilot. Press release, January 24, 2024. https://www.nsf.gov/news/democratizing-future-ai-rd-nsf-launch-national-ai.

National Science Foundation (2026). EPSCoR criteria for eligibility. https://www.nsf.gov/funding/initiatives/epscor/epscor-criteria-eligibility. Web page. Accessed 17 October 2026.

National Science Foundation, Office of Advanced Cyberinfrastructure (2021). Transforming science through cyberinfrastructure: NSF's blueprint for a national cyberinfrastructure ecosystem for science and engineering in the 21st century. Technical report, National Science Foundation. https://web.archive.org/web/20240125142754/https://www.nsf.gov/cise/oac/vision/blueprint-2019/.

Nelson, R. R. (1959). The simple economics of basic scientific research. *Journal of Political Economy 67*(3), 297–306.

NVIDIA Corporation (2017). NVIDIA Tesla V100 GPU architecture. Whitepaper WP-08608-001. https://images.nvidia.com/content/volta-architecture/pdf/volta-architecture-whitepaper.pdf.

NVIDIA Corporation (2021). NVIDIA Ampere GA102 GPU architecture. Whitepaper, v2.1. https://www.nvidia.com/content/PDF/nvidia-ampere-ga-102-gpu-architecture-whitepaper-v2.1.pdf.

Payne, A. A. (2001). Measuring the effect of federal research funding on private donations at research universities: Is federal research funding more than a substitute for private donations? *International Tax and Public Finance 8*(5–6), 731–751.

Phelps, R. L. (2019). Research infrastructure: Major research instrumentation (MRI) and mid-scale research infrastructure. Presentation at the NSF Grants Conference, Los Angeles, May 20–21, 2019. National Science Foundation, Office of Integrative Activities. https://www.jsums.edu/research/files/2019/05/MRI_spr19.pdf.

Rahman, R. (2024). Performance per dollar improves around 30% each year. Epoch AI, Data Insights, October 23, 2024. https://epoch.ai/data-insights/price-performance-hardware.

Restuccia, D. and R. Rogerson (2008). Policy distortions and aggregate productivity with heterogeneous establishments. *Review of Economic Dynamics 11*(4), 707–720.

Rubens, M., Y. Wu, and M. J. Xu (2026, March). Exploiting or augmenting labor? *American Economic Review: Insights 8*(1), 72–89.

Shvadron, D., H. Zhang, L. Fleming, and D. P. Gross (2025). Measuring the production of scientific human capital: New data, methods, and evidence on the U.S. scientific training ecosystem. NBER Working Paper 33944, National Bureau of Economic Research.

Smith, P. M., J. Gemmill, D. Y. Hancock, B. W. O'Shea, W. Snapp-Childs, and J. Wilgenbusch (2025). Application of the cyberinfrastructure production function model to R1 institutions. *Frontiers in Research Metrics and Analytics 10*, 1449996.

Suh, J. (2026, June). Instrumental burden: Rising capital intensity in science and consequences for innovation. SSRN working paper, New York University, Stern School of Business. Available at SSRN: https://ssrn.com/abstract=7017658.

TechPowerUp (2026). GPU specs database. https://www.techpowerup.com/gpu-specs/. Accessed July 2026.

Teodoridis, F. (2018). Understanding team knowledge production: The interrelated roles of technology and expertise. *Management Science 64*(8), 3625–3648.

U.S. Department of Energy, Office of Science (2011). The Magellan report on cloud computing for science. Technical report, U.S. Department of Energy, Office of Advanced Scientific Computing Research. https://science.osti.gov/-/media/ascr/pdf/program-documents/docs/Magellan_Final_Report.pdf.

Waldinger, F. (2016). Bombs, brains, and science: The role of human and physical capital for the creation of scientific knowledge. *The Review of Economics and Statistics 98*(5), 811–831.

Wallsten, S. J. (2000). The effects of government-industry R&D programs on private R&D: The case of the Small Business Innovation Research program. *RAND Journal of Economics 31*(1), 82–100.

Wing, C., S. M. Freedman, and A. Hollingsworth (2024). Stacked difference-in-differences. Working Paper 32054, National Bureau of Economic Research.

Yue, D. (2024, November). I, Google: Estimating the impact of corporate involvement on AI research. Research Paper 5033334, Georgia Tech Scheller College of Business.

A1 Data Construction

This appendix documents computing capacity (Section A1.1), scientific output (Section A1.2), quasi-fixed research labor (Section A1.3), GPU acquisition costs (Section A1.4), and field-level cluster usage (Section A1.5). Output and quasi-fixed labor come from OpenAlex. Flexible labor comes from the NSF Survey of Graduate Students and Postdoctorates and the ProQuest dissertation panel from Shvadron et al. (2025), as described in Section 3.

A1.1 Computing Capacity

For each university, we locate the hardware specifications for its central cluster and retrieve annual Wayback Machine snapshots back to 2010 or the earliest available year. When the hardware page moves, we search archived university sites to link its successive addresses. A Claude Code LLM agent (Opus 4.6) extracts CPU cores and nodes, GPU models and counts, pricing tiers, and funding information, and is instructed to check its output against the source page. We manually verify the hardware histories for the 53 universities in the panel and spot-check extracted records. For ten universities, research computing centers also confirm major upgrade dates directly.

A1.2 Scientific Output

We pull every OpenAlex work of type *article* or *preprint* published 2005–2025 with at least one authorship affiliated with an R1 university, querying the OpenAlex works endpoint by institution. Each work is stored as one row per author $\times$ listed affiliation, with the author's position in the byline, the paper's OpenAlex topics, and its citation statistics as of the pull date. Works pulled through several universities are deduplicated on the OpenAlex work id, retracted works are dropped, and OpenAlex caps authorships at the first 100 authors of a paper. The corpus contains roughly 8.4 million works. The estimation sample is the inner join of this corpus with the 53 universities and years for which computing capacity is observed.

A1.2.1 Apportioning Credit across Authors and Institutions

A paper with n authors is divided equally among them, so each author holds $1/n$ of the paper. An author who lists m affiliations on that paper has their $1/n$ divided equally across the m institutions. A university's credit for a paper is the sum of the shares of the author–affiliation rows that name it. We call this *equal credit*; it is the baseline credit rule, with each paper summing to one across all author–affiliation shares, including unresolved affiliations. Authors with no resolvable affiliation keep their share, which is therefore assigned to no university.

We also construct alternative credit rules for robustness. *First-author* and *last-author* credit give the whole paper to the first or the last author (split across their affiliations as above); *first-or-last* credit gives half to each, or all to a solo author. These reflect the convention in laboratory sciences that the last author is the principal investigator. *Institution credit* divides the paper equally across its distinct institutions regardless of how many authors each contributes, which guards against very large author lists. A *one-institution* variant assigns each author's $1/n$ entirely to their main institution in

that year (defined in Section A1.3.1) rather than splitting it across listed affiliations. This recovers credit for rows with no affiliation and removes the dilution that arises when a university and its hospital are both listed, and it raises total credit in the estimation sample by about nine percent. Equal credit is used for the baseline results, and re-estimating the production function on the one-institution panel leaves every capital elasticity within a standard error of the baseline.

Credit is summed to the university × OpenAlex subfield × publication year, the subfield being that of the paper's primary topic. Subfields map to the paper's field taxonomy through a crosswalk from the 252 OpenAlex subfields to NSF fields.

A1.2.2 Citation Measures

Each credit-weighted count is computed for several valuations of a paper. *Papers* values every paper at one. *Citation percentile* values a paper at its OpenAlex citation-normalized percentile, which ranks the paper's citation count against works of the same document type, publication year, and subfield; it lies in $[0, 1]$. *Tail* values a paper at one if its percentile is at least 0.95 and zero otherwise, so the credit-weighted sum counts top-five-percent papers. *Field-weighted citation impact* (FWCI) is the paper's citation count divided by the mean for its type, year, and subfield, capped at the 99th percentile of the corpus; an indicator for FWCI at least one is also carried. Citation counts are cumulative as of the pull, not a fixed window, so recent papers have had less time to accumulate citations; the percentile and FWCI measures are the ones normalized for this. A paper whose percentile OpenAlex does not report is valued at zero. The main text uses papers and the citation percentile; the tail measure is zero in a large share of university–field–years and is used only descriptively.

A1.2.3 Classification of Papers as Computational and GPU-Intensive

The classification works at the level of OpenAlex *topics*, of which there are 4,516, each with a name, a field, a domain, keywords, and a one-paragraph description. We classified every topic with a large language model, supplying it with these fields.[11] The model answers two questions. First, is research in this topic typically *computational* (reproducing it would normally require cluster computing), *experimental*, *theoretical*, or genuinely *mixed*? The result is 451 computational, 500 mixed, 2,702 experimental, and 863 theoretical topics. Second, for the 951 computational or mixed topics, is the computation typically *GPU-heavy*, *CPU-heavy*, or *both*? This yields 119 GPU-heavy and 239 "both" topics, so 358 topics are classed as GPU-topic.

OpenAlex assigns each paper a primary topic and up to two secondary topics. A paper is flagged as GPU-topic if any of the topics are classified as such, and this serves as the main output in the production function. These flags are propensities: they say what kind of research the paper reports, not whether a particular machine was used. Therefore, the production function relates the shared stock of CPU and GPU capital to publications in topics likely to benefit from it, not to verified use by individual papers.

[11] The classification was run once, manually inspected, and 30 topics were corrected by hand afterwards.

A1.2.4 Abstract- and Full-Text-Based Measures

A second family of measures reads the paper's title, abstract, and full text (when available) with regular expressions and is used as a text-based indicator of GPU and AI work in the event-study analysis. A *generative-AI* tier fires on GPU, large language model, natural language processing, PyTorch, TensorFlow, machine learning, artificial intelligence, deep learning, or neural network; its deep-learning term list also covers transformer and convolutional architectures, generative models, graph neural networks, deep reinforcement learning, and the classic machine-learning terms random forest and gradient boosting, so the outcome the main text calls "AI and deep learning" is machine learning broadly.

Two limitations of the text measures should be kept in mind. First, abstracts are available for about 70 percent of works, a share that is stable over the sample period, so title-and-abstract measures undercount papers whose relevant terms appear only in an unavailable abstract. A stable coverage share need not imply stable composition. Second, OpenAlex indexes the full text only of open-access works whose publishers permit it, so a search of the full text is selected on open-access status and trends upward over the sample. We therefore keep a full-text keyword search only as a robustness exercise and never as a primary outcome. The rules are guarded against review and survey articles and against domain-specific false positives.

A1.2.5 Excluding Industry Co-Authors

Papers co-authored with researchers at firms may have been computed on those organizations' resources rather than the university's. This may be especially prevalent in computer science due to the computing available at hyperscalers (Yue 2024). We rebuild every output panel after dropping papers with any author affiliated with an institution that OpenAlex classifies as a *company*. The flag is defined over every affiliation of every author, so a university author who also lists a company triggers the drop. In the corpus, seven percent of works have a company co-author.

A1.3 Quasi-Fixed Research Labor

The production function distinguishes quasi-fixed labor, the stock of researchers attached to a university with high adjustment costs (faculty and permanent research staff), from flexible labor (PhD students and postdocs). No administrative count of faculty by field exists for the universe of R1 universities, so quasi-fixed labor is measured from the publication records themselves: an author is counted as quasi-fixed at a university in a year when their publication history shows a long, persistent attachment to that institution that began after their career was under way. This section gives the rule and then validates it against the one external field-level count we could obtain, the American Institute of Physics survey of physics and astronomy departments.

A1.3.1 Author Panel

From the same corpus (articles and preprints, 2005–2025) we build a panel of every author with at least one affiliation at an R1 university. In each year an author publishes, their *main institution* is

the institution listed on the largest number of their papers that year. Ties are broken in favor of persistence: the main institution of the nearest earlier year with a single top institution, then of the nearest later such year, then the institution with the most papers over their whole career. Years in which they do not publish are filled with the institution they held on both sides of the gap, provided the gap is at most three years; gaps between two different institutions are not filled at this stage. *Career age* is the number of years since the author's first publication in the corpus. Each author-year is also placed in a single OpenAlex subfield, the modal subfield of that year's papers, which is how field cells are formed.

A1.3.2 The Quasi-Fixed Classification

A *spell* is a run of consecutive years at the same main institution. An author-year is classified as quasi-fixed under the main (*composite*) definition when

$$\text{spell length} \geq 4 \quad \text{and} \quad \big(\text{career age at the start of the spell} \geq 3 \;\; \text{or} \;\; \text{spell length} \geq 8\big).$$

The spell-length condition identifies persistence. The arrival condition separates hires from trainees: a researcher who arrives at an institution at least three years into their publishing career is more likely to be a faculty or staff hire, whereas a PhD student's career typically begins at the institution where they train. The second branch admits very long spells regardless of arrival age, using a long stay as evidence of persistent attachment. Both conditions are constant within a spell, so an author's classification does not flip from year to year. Two single-signal definitions are carried for robustness: *persistence* (spell length of at least four years alone) and *position* (the author's share of first-or-last authorships held as last author exceeds the median of their subfield, a seniority proxy observable within a single year). Everyone not quasi-fixed is counted on the flexible side of the OpenAlex panel, which is used only for diagnostics; flexible labor in the production function comes from the NSF surveys.

A1.3.3 Flow and Stock Counts

The *flow* count of a cell (university $\times$ subfield $\times$ year) is the number of quasi-fixed authors who publish in that cell-year. It understates labor in years when an active researcher happens not to publish, and in small cells this measurement error moves with the output shock. The production function therefore uses the *stock* count: an author is counted at their institution in every year their spell covers, whether or not they publish, with their subfield carried forward from their most recent active year. Two conventions complete the stock. A gap of at most three years between spells at two different institutions is split at its midpoint, the first half credited to the earlier institution and the second to the later; longer gaps are left uncounted. And presence extends one year past an author's last publication, so that researchers whose next paper has not yet appeared are not read as having left; one-paper authors are never extended. A fractional-credit version of the flow count, which splits an author across subfields in proportion to their credit-weighted papers, is carried for robustness.

A1.3.4 Panel-Edge Adjustment

Because spells are observed only within 2005–2025, the classification is censored at both ends of the panel. At the right edge an author whose spell is still open in 2025 and has not yet reached the qualifying length cannot be classified: they may be a transient or a stayer observed too early. Left uncorrected, this mechanically depresses the quasi-fixed count in the final years, exactly when GPU capacity grows fastest, and biases the capital elasticity upward. We flag these censored authors (an open spell at the panel end, shorter than four years for mid-career arrivals or eight years for career-start arrivals) and impute them to qualify at the rate observed among the classifiable authors of the same cell-year:

$$L^F_{\text{adj}} \;=\; L^F \times \frac{N}{N - N_{\text{censored}}}$$

where N is the cell's total author count. The adjustment is computed after summing to the estimation cell, since it is a ratio of sums. It is applied to the stock composite count, and the resulting *edge-adjusted composite stock* is the measure of quasi-fixed labor in every production-function specification. The adjustment assumes that censored authors qualify at the same rate as classifiable authors in the cell; the raw and adjusted counts show sensitivity to that assumption. At the left edge, career age is censored for authors already publishing in 2005, which routes early spells through the eight-year branch and produces a rising quasi-fixed share over 2005–2012 in the full R1 panel. This left edge is not adjusted. Its effect is confined to the earliest panel years, which contribute few moments. In the estimation universities the ramp is complete by 2012 (Table A.2).

A1.3.5 Aggregation to the Estimation Cell

Subfield counts are summed to the university × field × year cells of the production function through a manually constructed OpenAlex-to-NSF field crosswalk. We drop mathematics, psychology, the social sciences, and the humanities. A cell with positive flexible labor but no quasi-fixed author is floored at one so that its logarithm is defined.

A1.3.6 Validation against the AIP Physics and Astronomy Faculty Survey

The American Institute of Physics (AIP) surveys every U.S. physics and astronomy department biennially about its faculty. We obtained, for all departments with an IPEDS ID, the full-time-equivalent (FTE) faculty positions involving teaching or research in physics or astronomy as of March 1 of each survey year, 2010–2026, split into tenured, tenure-track, permanent non-tenure-track, and temporary positions (the last including soft-money research professors), and excluding postdocs and graduate students.[12] We use *tenured plus tenure-track* FTE as a benchmark for the persistent research workforce. Tenure-line positions are 71 to 83 percent of total physics and astronomy FTE in the estimation universities, a share that falls over the period.

[12] Department rows were coalesced where the file carried a recent and an older record of the same department; departments of the same university (Physics, Astronomy, Applied Physics, Planetary Science) were summed, and a university-year is used only when every department it ever reports is present; a reported zero in the 2026 wave with no components is a not-yet-returned survey and is treated as missing. The 2014 wave's temporary and non-tenure-track components exceed the FTE total in 86 departments, but its tenured and tenure-track counts are consistent with the totals.

Of the 187 R1 universities, 175 appear in the AIP file and 171 can be matched to OpenAlex physics and astronomy output; all 53 estimation universities match. On the OpenAlex side we sum the labor counts over the physics and astronomy subfields to the university-year. Because the AIP figure is a March snapshot, OpenAlex year $t-1$ is the most recent complete publication year, and we compare AIP year t with OpenAlex year $t-1$; this alignment fits better than the contemporaneous one. The comparison asks three questions of each OpenAlex measure: does it rank departments correctly across universities (between-university correlation of log counts); does it track movements within a university net of common trends (two-way demeaned correlation); and does it track changes (correlation of log differences over two, four, and six years, the survey's horizons).

Table A.1: OpenAlex Quasi-Fixed Labor Measures against AIP Tenure-Line Faculty, Physics and Astronomy

	Between	Within	Two-way	Δ_2 log	Δ_6 log	Level ratio
Composite stock, edge-adjusted (main)	0.93	0.28	0.26	0.07	0.31	1.82
Composite stock, raw	0.93	0.27	0.25	0.06	0.30	1.80
Composite flow	0.93	0.28	0.26	0.09	0.29	1.37
Composite flow, fractional credit	0.94	0.28	0.26	0.10	0.28	1.57
Persistence stock	0.94	0.28	0.27	0.05	0.35	2.89
Position stock	0.95	0.21	0.22	0.06	0.24	1.56
All authors, stock	0.95	0.26	0.25	0.09	0.28	4.70

Notes: 369 university-years. OpenAlex counts summed over physics and astronomy subfields, year $t-1$, against AIP tenured plus tenure-track FTE in all physics and astronomy departments, March of year t. Between: correlation of university-mean logs. Within: correlation of university-demeaned logs. Two-way: university- and year-demeaned. Δ_k log: correlation of k-year log differences. Level ratio: median OpenAlex count over AIP FTE.

Table A.1 reports the results. The measure validates as a cross-sectional stock: between-university correlations are 0.93 to 0.95 for every definition, and the edge-adjusted composite stock is about 1.8 times tenure-line FTE, consistent with permanent research staff and long-stay postdocs qualifying as quasi-fixed in addition to faculty. Within universities the correlations are positive but modest, about 0.26 in the two-way demeaned data and 0.3 over six-year changes, and the composite definitions outperform the position definition on every dimension. The OpenAlex measure should accordingly be read as a reliable measure of the relative size of departments and a noisier measure of their year-to-year movement.

Table A.2: Trends in Tenure-Line Faculty and OpenAlex Measures, 31 Estimation Universities Observed in Every AIP Wave, Index 2012 = 1

	2010	2012	2014	2016	2018	2020	2022	2024
AIP tenure-line FTE	0.95	1.00	0.95	1.04	1.03	1.08	1.08	0.97
Composite stock, edge-adjusted (main)	0.83	1.00	0.99	1.12	1.21	1.20	1.16	0.96
Composite stock, raw	0.83	1.00	0.99	1.12	1.21	1.18	1.08	0.76
Position stock	0.99	1.00	0.96	1.04	1.03	1.04	1.02	0.81
All authors, stock	0.92	1.00	0.96	1.10	1.16	1.25	1.22	1.01

Table A.2 shows what the edge adjustment does. Tenure-line physics and astronomy faculty in the estimation universities is flat over 2012–2024. The raw composite stock falls to 0.76 of its 2012 level by 2024, the right-edge censoring described above; the adjusted measure ends at 0.96, in line

with the survey. The adjustment is a near-proportional rescaling of the final years, and restores the aggregate trend while changing the within-university correlations little.

A1.4 GPU Capital in Dollars

The capacity panel records, for each university cluster and year, the GPU models installed and their counts, taken from research-computing web pages (archived snapshots, extracted and hand-checked). We convert this stock into dollars by pricing every GPU at the price prevailing when the university acquired it. The acquisition year of a GPU model is the first year it appears in its cluster, or the cluster's stated deployment year when that is earlier and the model is part of the cluster's first observed configuration; no GPU is dated before its launch. The capital stock of university j in year t is the sum over installed GPUs of count times acquisition-year price, expressed in 2017 dollars with the BEA GDP implicit price deflator. It is a historical-cost measure: no depreciation is applied, and a card bought in 2018 keeps its 2018 price in every later year. Dividing the stock by its FP16 (or FP32) TFLOPS gives the vintage-weighted acquisition cost per TFLOPS of each university's fleet.

Prices come from the following sources, in order of preference. For 70 percent of GPUs (count-weighted) the price is the median transaction price paid by U.S. higher-education institutions for that model in that year; a further 18 percent use the adjacent year when the same year has fewer than three purchase-order lines. Ten percent, mostly models that universities rarely buy through itemized orders, use the manufacturer's launch price. One percent of GPUs are priced at a same-generation sibling because no price exists for the exact model, and one percent (AMD Instinct accelerators and Intel Xeon Phi coprocessors) remain unpriced. A launch-price-only series is built in parallel as a robustness check; the two agree closely in levels and move together over time.

GovSpend. GovSpend is a commercial database of government purchase orders assembled from public-records requests to state and local entities, including public universities. We pull all purchase-order line items from 1,072 higher-education agencies for 2015–2026 that match a list of computing-hardware keywords (about 179,000 lines), discard the roughly one third that are false keyword matches, and parse the GPU model, unit count, and unit price on the remaining itemized lines. The model-by-year median price per GPU, weighted by units, is the transaction price used above. Model and year fixed effects explain 83 percent of the variance in log unit prices for datacenter cards, and the price paid for the same model in the same year does not vary systematically with buyer size or wealth, so the medians can be read as market prices rather than as university-specific bargains. Thirty-four of the 53 panel universities appear in GovSpend; private universities and a few publics do not, which is why GovSpend informs the price series but not the capacity measure itself.

GPU reference table. A hand-built table of 65 GPU and accelerator models gives each model's release date, launch price, and throughput at FP64, FP32, and dense FP16 (tensor-core) precision. Specifications come from NVIDIA and AMD datasheets and architecture whitepapers, from the Epoch AI Machine Learning Hardware dataset, and, for cards released before 2017, from the TechPowerUp GPU database (a 3,203-card export). All FP16 figures are dense rates; the "with sparsity" figures that NVIDIA has marketed since Ampere are halved. Launch prices are list prices at release; the

H100 has no published list price and takes Epoch AI's per-GPU cost imputed from DGX server pricing.

Epoch AI. Epoch AI's *Data on Machine Learning Hardware* and *Data on AI Chip Sales* (CC-BY 4.0) supply throughput specifications for recent accelerators and an approximate cost per chip, used as the last resort when neither a transaction nor a launch price exists.

TechPowerUp. The TechPowerUp GPU database provides release dates, launch prices, FP32 shader throughput, and core counts for the full history of consumer and professional cards. It is used for cards released before the deep-learning era, which are absent from Epoch AI, and for the tensor-core counts behind a few derived FP16 figures. Its own FP16 field is not used because it reports the shader path rather than tensor-core throughput.

Price deflator. All dollar values are deflated to 2017 with the BEA implicit price deflator for gross domestic product, carried forward at its last published value for the most recent fiscal years.

The same purchase orders supply the prices used in Sections 6, 4.3, and 8. Table A.3 reports them by purchase year. Across 463 datacenter-GPU purchase lines at 133 public universities, the median price is \$37 per FP16 TFLOPS, falling from \$56 in 2018 to \$17 in 2025; the median CPU price is \$76 per core. Averaging the yearly medians gives \$38 per FP16 TFLOPS and \$84 per core, the prices used in Sections 6 and 8. Prices vary mainly across GPU models and years rather than across buyers: year and model effects explain 52% of the variance of log price per TFLOPS, adding university effects raises the adjusted R^2 from 0.50 to 0.60, and the median coefficient of variation within a model-year is 0.32. The hardware flows of Section 4.3 value GPU additions at the median of the arrival year and CPU additions at \$84 per core.

Table A.3: Hardware Prices in University Purchase Orders (GovSpend)

	Datacenter GPUs: $ per FP16 TFLOPS					CPUs
Purchase year	Lines	Universities	Models	Median	IQR	Median $ per core
2015	–	–	–	–	–	110
2016	–	–	–	–	–	74
2017	–	–	–	–	–	72
2018	11	8	2	56	51–63	72
2019	47	27	2	49	36–63	65
2020	51	26	3	40	32–60	83
2021	63	38	5	36	32–44	104
2022	61	38	7	39	34–57	82
2023	74	41	11	44	27–56	69
2024	70	36	13	38	25–57	72
2025	86	42	14	17	15–32	141
Pooled	463	133	16	37	26–54	76
Mean of yearly medians				38		84

Notes: Purchase-order lines from GovSpend for datacenter-class GPUs (2018–2025) and CPUs (2015–2025) at public universities, restricted to lines with a plausible unit price. A line's price per FP16 TFLOPS divides its price by the dense tensor-core throughput of the GPUs it contains. The last row is the geometric mean of the yearly medians; these are the prices used in Sections 6 and 8. In a regression of log price per FP16 TFLOPS on fixed effects, year effects explain 23% of the variance, GPU-model effects 50%, year and model effects together 52%, and adding university effects raises the adjusted R^2 from 0.50 to 0.60; the median coefficient of variation of price within a model-year is 0.32.

A1.5 Field-Level CPU and GPU Usage

The weights γ_f in the capital nest are calibrated from the realized allocation of compute across fields at three university research computing centers. Each center reports annual CPU and GPU usage by department or field of science: University A reports core-hours on its CPU clusters and on its GPU cluster for the top departments on each cluster (2017–2026); University B reports CPU hours and GPU hours by department from its XDMoD accounting portal (August 2022–August 2026); University C reports CPU core-hours and GPU-hours by field for all users of its statewide center (2015–2026), of which the university accounts for roughly four-fifths. We assign each department or field to a fine NSF field using the crosswalk of Section 3 and sum to the five field groups; administrative units, course accounts, and catch-all categories such as "other natural sciences" are left unassigned and excluded from the field-level calculations. All three centers meter usage at the job level, so the hours reflect what fields actually consume.

Usage is measured in each center's own units: University A's GPU figure is core-hours on GPU nodes while Universities B and C report GPU-hours directly, and no center reports a common exchange rate between CPU and GPU time. We therefore use only within-center-year contrasts across fields. Under the capital nest of Section 5, the log ratio of CPU to GPU hours in a field differs from that in computer science by $\sigma_K \log\left[\frac{\gamma_f/(1-\gamma_f)}{\gamma_{\text{CS}}/(1-\gamma_{\text{CS}})}\right]$ plus a center-year term that captures the center's relative price and units of the two kinds of compute. We regress the log ratio on field-group indicators and center-year fixed effects, using field-by-year cells with at least one hour of each type of compute, and invert the coefficients at the estimated σ_K, normalizing $\gamma_{\text{eng}} = 0.75$ because the level of γ_f is a choice of units. Table A.4 reports the tilts and implied weights for each center and pooled; all

four other fields have higher CPU-to-GPU usage ratios than computer science at every center, and a uniform γ is rejected at each. Hours of use are a flow and the model's C_{jft} and G_{jft} are stocks allocated to a field; the two coincide when fields run their allocated capacity at a common utilization rate (Appendix A4.2.1), and only the cross-field ratio, not the level, enters the calibration.

Table A.4: CPU Weights Calibrated from Field-Level CPU and GPU Usage

	Computer science	Engineering	Geosciences & physics	Molecular & materials	Life sciences
Panel A: log(*CPU hours*/*GPU hours*) *relative to computer science*					
University A	0	3.60 (0.40)	4.59 (0.76)	4.34 (0.39)	2.93 (0.53)
University B	0	3.13 (0.40)	3.30 (0.68)	3.55 (0.84)	3.17 (0.53)
University C	0	2.98 (0.43)	3.94 (0.41)	2.22 (0.50)	1.65 (0.29)
Pooled	0	3.25 (0.25)	3.87 (0.37)	3.25 (0.35)	2.42 (0.28)
Panel B: implied γ_f *(engineering normalized to 0.75)*					
University A	0.01	0.75	0.93	0.90	0.53
University B	0.03	0.75	0.79	0.85	0.76
University C	0.03	0.75	0.93	0.49	0.29
Pooled	0.02	0.75	0.88	0.75	0.47

Notes: Panel A reports OLS coefficients on field-group indicators in a regression of the log ratio of CPU to GPU hours on university-year fixed effects, estimated on field-by-year cells with at least one hour of each type of compute; standard errors in parentheses are clustered by university-year. The university-year effects absorb each center's units (core-hours on GPU nodes at University A, GPU-hours at Universities B and C), so only within-university-year contrasts across fields are used. Under the capital nest of Section 5, $b_f - b_{\text{CS}} = \sigma_K \log\left[\frac{\gamma_f/(1-\gamma_f)}{\gamma_{\text{CS}}/(1-\gamma_{\text{CS}})}\right]$. Panel B inverts this at $\sigma_K = 0.67$ and normalizes $\gamma_{\text{eng}} = 0.75$ (the level of γ_f is a choice of units). Field groups follow the NSF field crosswalk (Table 2).

A2 Additional Event-Study Results

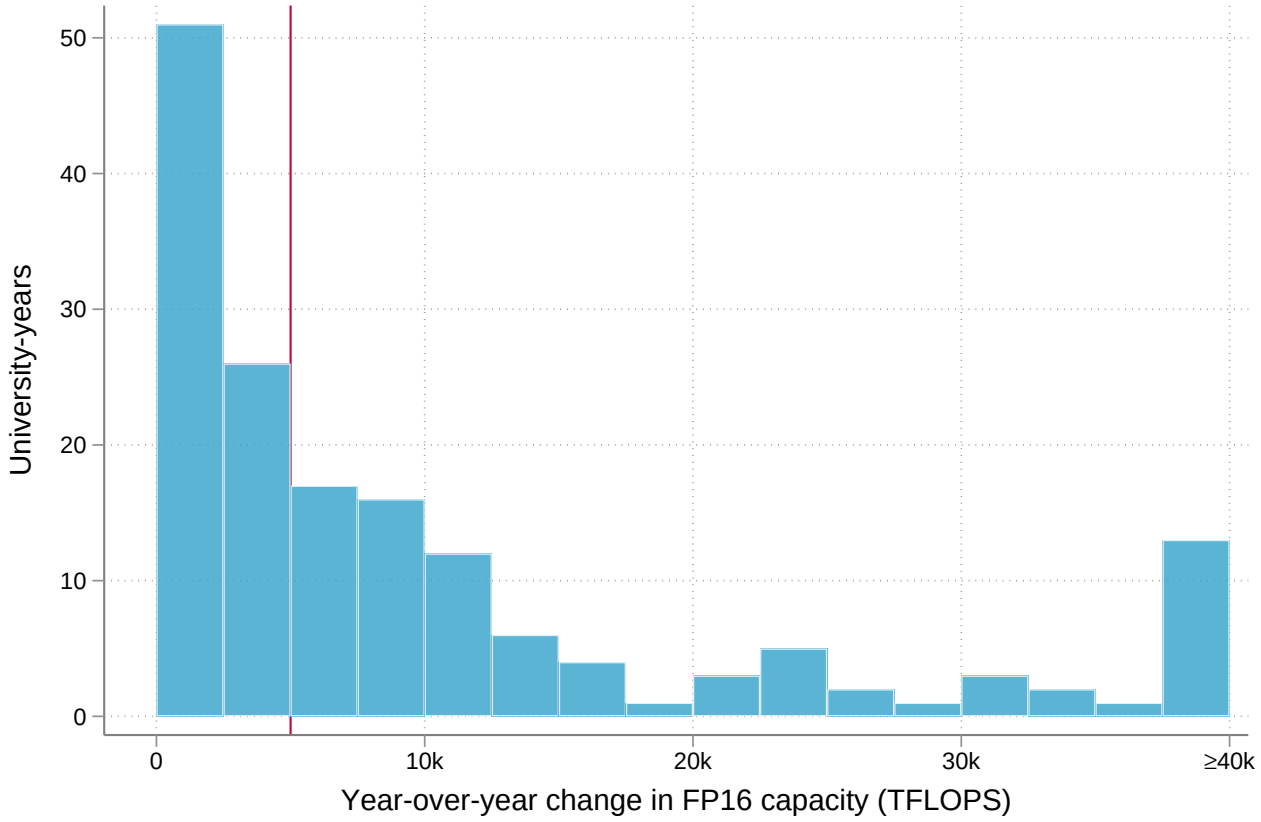
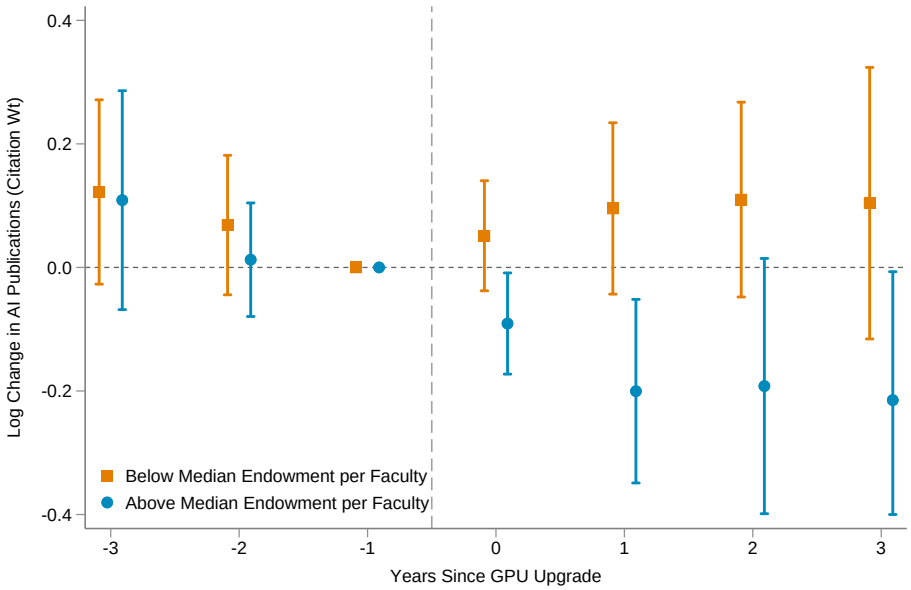
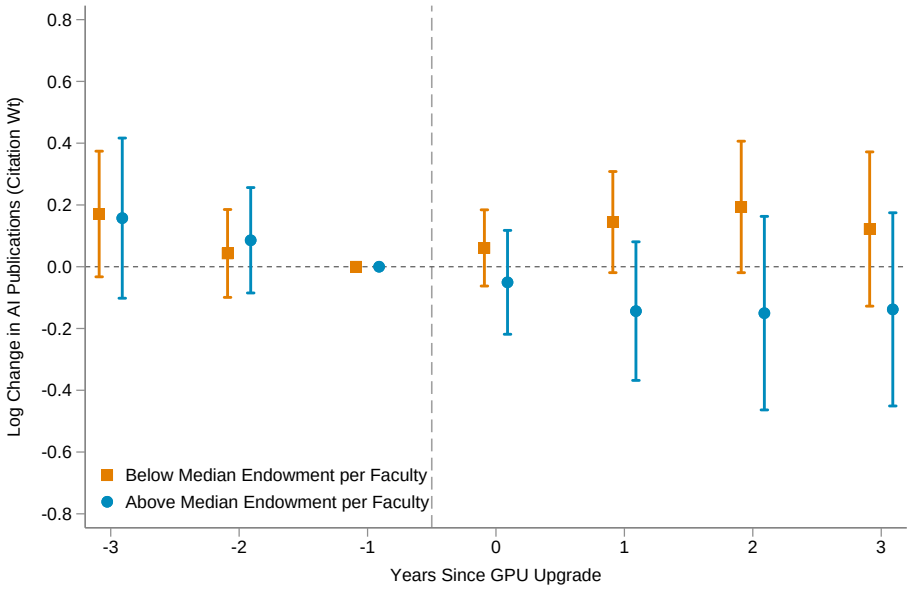


Figure A.1: Size of GPU Capacity Upgrades. The distribution of year-over-year increases in installed FP16 GPU capacity across the 163 university-years with an increase, 2010–2025. The solid line is the event threshold of 5,000 TFLOPS.

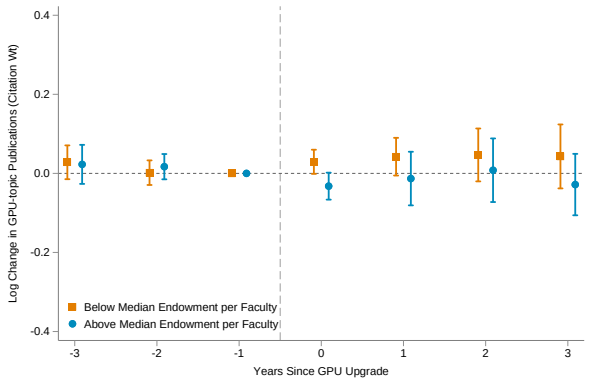


(a) All fields

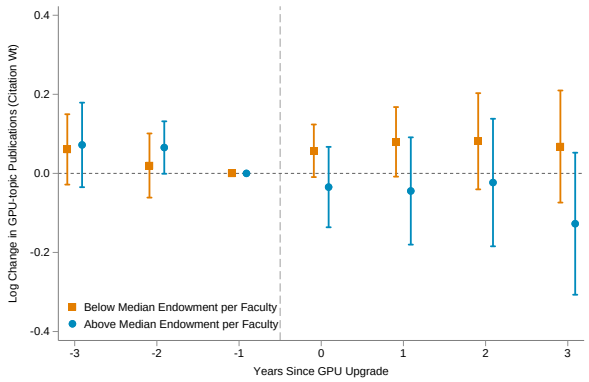


(b) Computer science

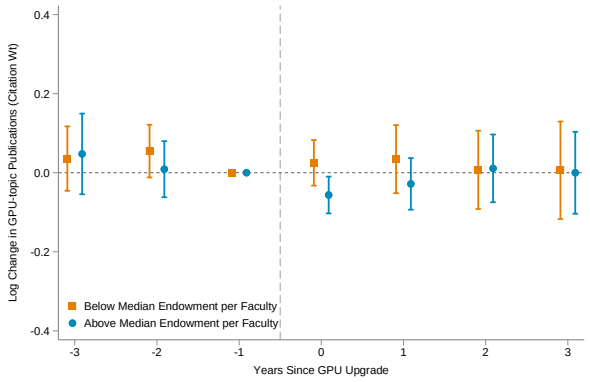
Figure A.2: Citation-Weighted AI and Deep-Learning Papers around Large GPU Upgrades, by Financial Resources. As in Figure 4, with output weighted by citation percentile. Each paper's equal-author credit is weighted by its citation percentile within document type, publication year, and subfield. Points are PPML estimates of β_k^g in Equation (4.1), the proportional change in output relative to the year before the upgrade and to control universities in the same stack. Squares are universities below the R1 median of endowment per faculty and circles are those above. An event is a within-year increase in FP16 GPU capacity of at least 5,000 TFLOPS. Bars are 95% confidence intervals; standard errors are clustered by university.



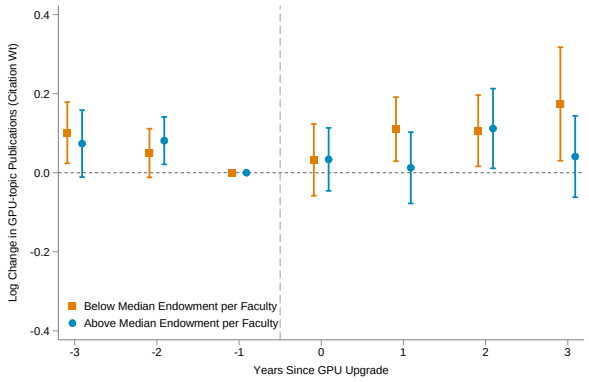
(a) All fields



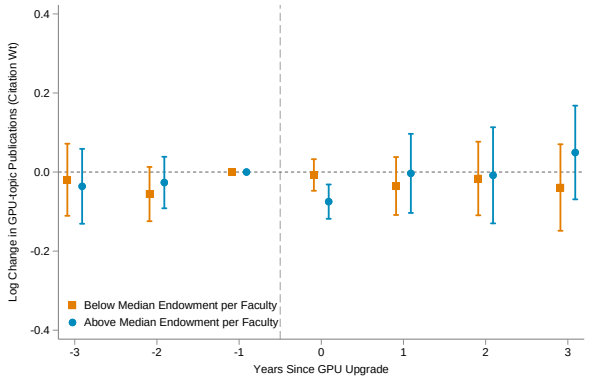
(b) Computer science



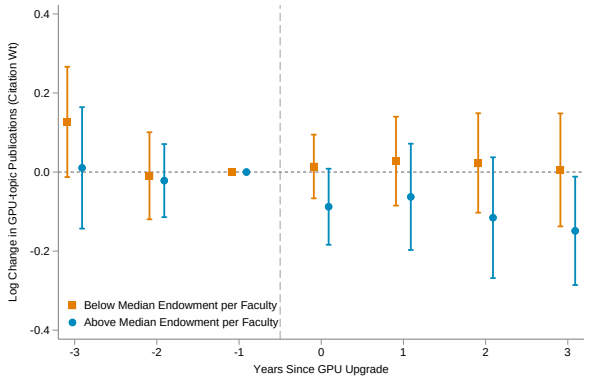
(c) Engineering



(d) Geosciences and physics

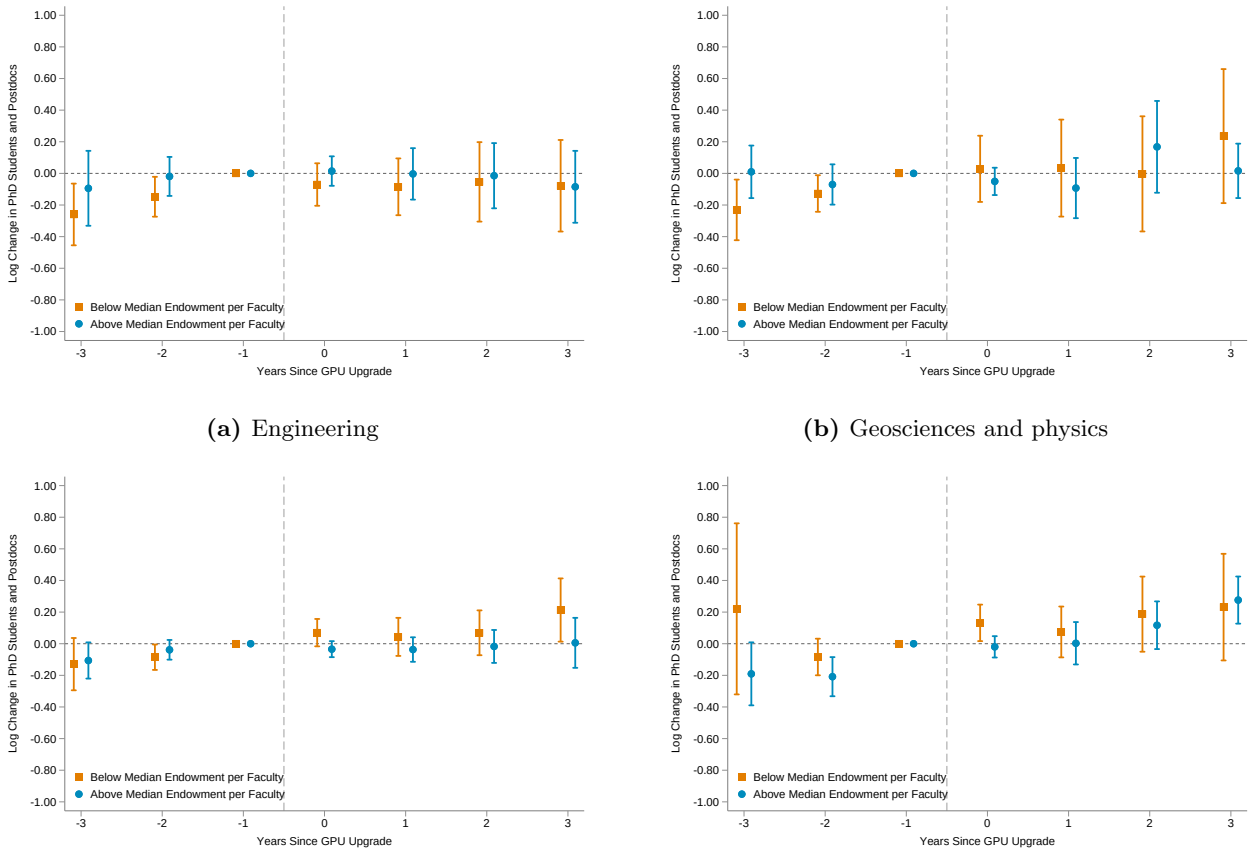


(e) Life sciences



(f) Molecular and materials sciences

Figure A.3: Citation-Weighted GPU-Topic Papers around Large GPU Upgrades, by Financial Resources. As in Figure 5, with output weighted by citation percentile. Each paper's equal-author credit is weighted by its citation percentile within document type, publication year, and subfield. Points are PPML estimates of β_k^g in Equation (4.1), the proportional change in output relative to the year before the upgrade and to control universities in the same stack. Squares are universities below the R1 median of endowment per faculty and circles are those above. An event is a within-year increase in FP16 GPU capacity of at least 5,000 TFLOPS. Bars are 95% confidence intervals; standard errors are clustered by university.



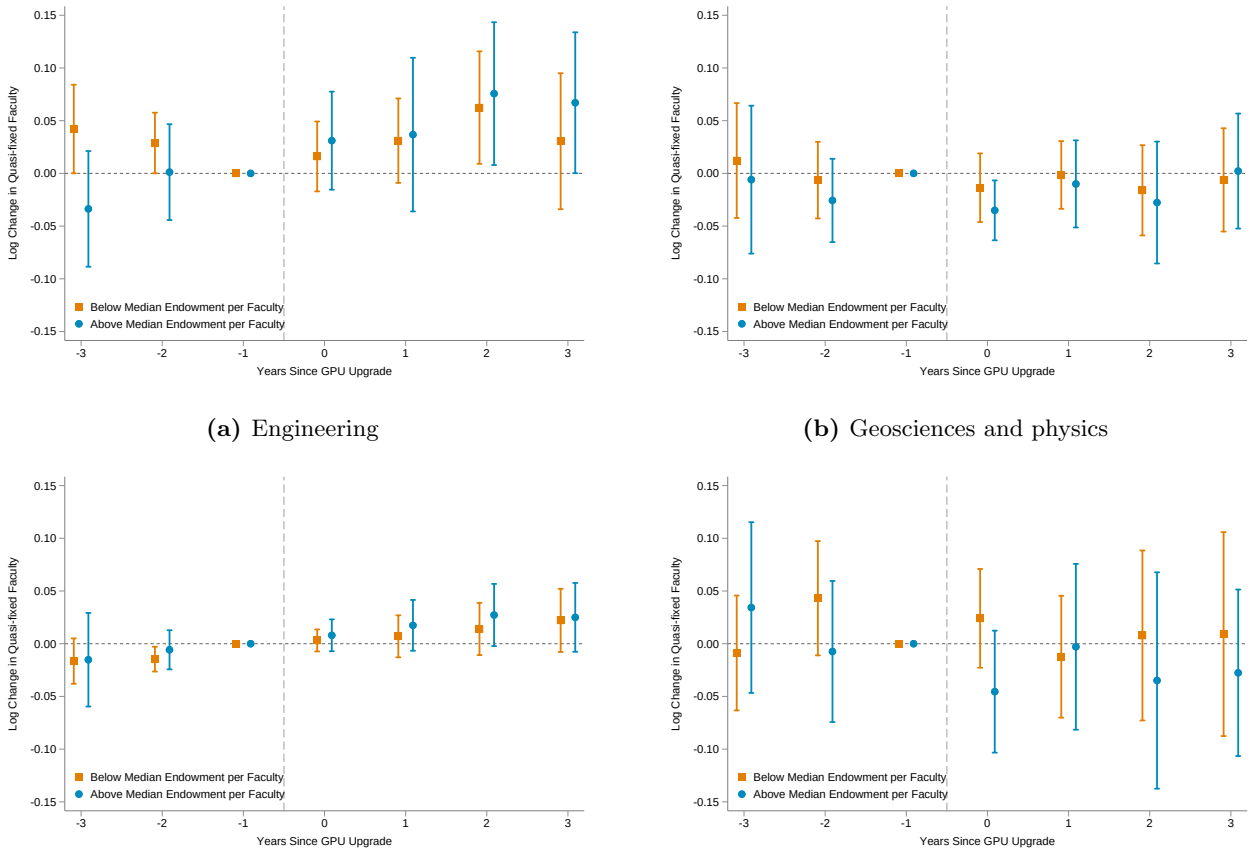
(a) Engineering

(b) Geosciences and physics

(c) Life sciences

(d) Molecular and materials sciences

Figure A.4: Flexible Labor around Large GPU Upgrades, Remaining Fields. As in panel (a) of Figure 6, for the number of full-time graduate students and postdocs (NSF GSS) in the four field groups not shown there. Points are PPML estimates of β_k^g in Equation (4.1), the proportional change in the labor input relative to the year before the upgrade and to control universities in the same stack. Squares are universities below the R1 median of endowment per faculty and circles are those above. An event is a within-year increase in FP16 GPU capacity of at least 5,000 TFLOPS. Bars are 95% confidence intervals; standard errors are clustered by university.



(a) Engineering

(b) Geosciences and physics

(c) Life sciences

(d) Molecular and materials sciences

Figure A.5: Quasi-Fixed Faculty around Large GPU Upgrades, Remaining Fields. As in panel (b) of Figure 6, for the quasi-fixed faculty stock (OpenAlex author histories) in the four field groups not shown there. Points are PPML estimates of β_k^g in Equation (4.1), the proportional change in the labor input relative to the year before the upgrade and to control universities in the same stack. Squares are universities below the R1 median of endowment per faculty and circles are those above. An event is a within-year increase in FP16 GPU capacity of at least 5,000 TFLOPS. Bars are 95% confidence intervals; standard errors are clustered by university.

A3 Stylized Model: Derivations and Simulation

This appendix derives the results in Section 5.1. Let $s \equiv \alpha_K + \alpha_L < 1$ denote returns to scale and let $\eta \geq 0$ denote the multiplier on the budget constraint in (5.3).

A3.1 First-Order Conditions

At an interior solution, the first-order conditions are

$$MPK \equiv \frac{\partial Y}{\partial K} = p\,(1+\eta) \tag{A3.1}$$

$$MPL \equiv \frac{\partial Y}{\partial L} = w\,(1+\eta) \tag{A3.2}$$

When $\eta = 0$, the university finances its privately optimal bundle. When $\eta > 0$, the financing constraint binds and the marginal return to both inputs exceeds their price. Dividing the two conditions gives the input ratio $L/K = \alpha_L p/(\alpha_K w)$, which is common across universities under Cobb–Douglas production.

A3.2 Marginal Product under a Binding Budget

With $N = 0$ and a binding budget, the university spends all of B, splitting it in proportion to the output elasticities:

$$K^c = \frac{\alpha_K}{s}\frac{B}{p}, \qquad L^c = \frac{\alpha_L}{s}\frac{B}{w} \tag{A3.3}$$

Substituting into the marginal product of capital,

$$MPK = \mathcal{C}\,\Omega\,B^{s-1} = p\,(1+\eta), \qquad \mathcal{C} \equiv \alpha_K \left(\frac{\alpha_K}{s\,p}\right)^{\alpha_K - 1}\left(\frac{\alpha_L}{s\,w}\right)^{\alpha_L} \tag{A3.4}$$

which is Equation (5.4). Holding productivity fixed, a larger budget relaxes the financing constraint, since $\partial \log(1+\eta)/\partial \log B = s - 1 < 0$. Holding the budget fixed, higher productivity raises the shadow value of internal funds one-for-one in logs. Comparing a high-resource university H and a low-resource university L with $B_H > B_L$ and $\Omega_H > \Omega_L$, both constrained,

$$\frac{MPK_H}{MPK_L} = \frac{\Omega_H}{\Omega_L}\left(\frac{B_H}{B_L}\right)^{s-1} \tag{A3.5}$$

so the high-resource university has the larger technical return to capital if and only if $\Omega_H/\Omega_L > (B_H/B_L)^{1-s}$.

A3.3 Crowd-Out Regimes

Measure the grant N in units of capital.

Constrained interior. If the budget binds and the university still buys capital ($I > 0$), the constraint can be written as $pK + wL = B + pN$: the grant is equivalent to cash. The university's choices depend only on the effective budget,

$$K^c(N) = \frac{\alpha_K}{s}\frac{B+pN}{p}, \qquad L^c(N) = \frac{\alpha_L}{s}\frac{B+pN}{w}, \qquad I^c(N) = \frac{\alpha_K B - \alpha_L\, pN}{s\,p}. \tag{A3.6}$$

Differentiating,

$$\frac{\partial I^c}{\partial N} = -\frac{\alpha_L}{s} = -(1-\chi), \qquad \frac{\partial K^c}{\partial N} = \chi, \qquad \frac{\partial L^c}{\partial N} = \frac{\alpha_L\, p}{s\,w}, \qquad \chi \equiv \frac{\alpha_K}{s}. \tag{A3.7}$$

Combining these with the first-order conditions, the full return to the grant in this regime is $dY/dN = p\,(1+\eta(N))$, where $1+\eta(N) = (\mathcal{C}/p)\,\Omega\,(B+pN)^{s-1}$. The return is larger where the financing wedge is larger, and it falls as the grant itself relaxes the wedge.

Corner. From (A3.6), a constrained university buys no capital of its own when $pN \geq (\alpha_K/\alpha_L)B$, that is, when the grant already exceeds the capital it would choose given its effective budget. Then $I = 0$, $K = N$, and the university spends its entire budget on labor, so $\partial I/\partial N = 0$ and $\partial K/\partial N = 1$. Corners are more likely when the budget is small relative to the grant.

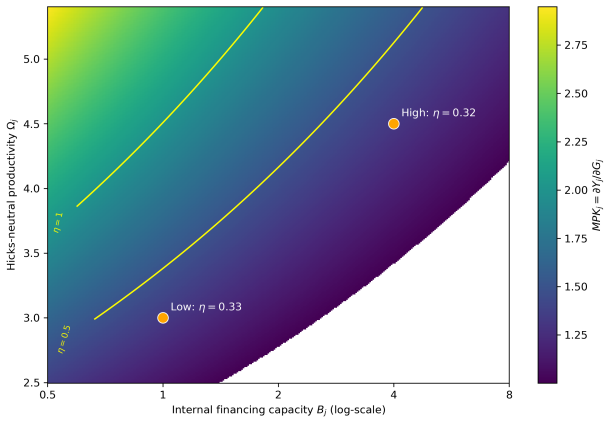
Unconstrained. If $\eta = 0$, the university buys inputs until $MPK = p$ and $MPL = w$. The optimal bundle (K^*, L^*) does not depend on the grant, so as long as the grant is smaller than K^*, the university reduces its purchases one-for-one: $\partial I/\partial N = -1$ and $\partial K/\partial N = \partial L/\partial N = 0$. Under Cobb–Douglas, spending at the unconstrained optimum is $s\,Y^*$, and Y^* is increasing in Ω, so the university is unconstrained when $B \geq s\,Y^* - pN$. Larger budgets make this regime more likely, while higher productivity raises the university's desired scale and makes it less likely.

A3.4 Illustrative Simulation

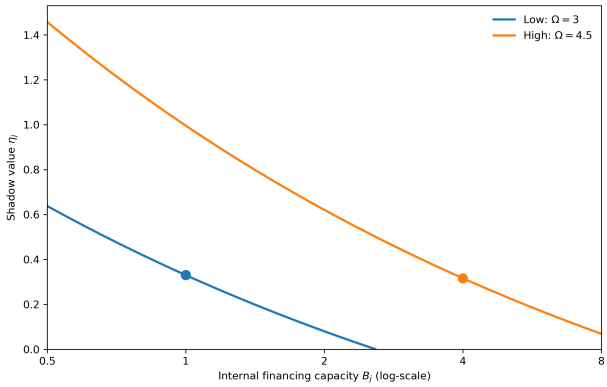
We illustrate the tradeoff with a two-university calibration in which both universities remain in the constrained-interior regime. We set $\alpha_K = 0.25$, $\alpha_L = 0.45$, and normalize $p = w = 1$. The low-resource university has $(B_L, \Omega_L) = (1, 3)$ and the high-resource university has $(B_H, \Omega_H) = (4, 4.5)$.

Figure A.6a plots MPK over combinations of (B, Ω) for which the budget binds. Moving up raises productivity and the return to capital. Moving right raises the budget and lowers the return through diminishing returns. Figure A.6b plots the financing wedge η against the budget, holding productivity fixed. At the calibrated points, $\eta_L = 0.331$ and $\eta_H = 0.317$: despite having four times the budget, the high-resource university is nearly as constrained because its higher productivity raises its desired scale.

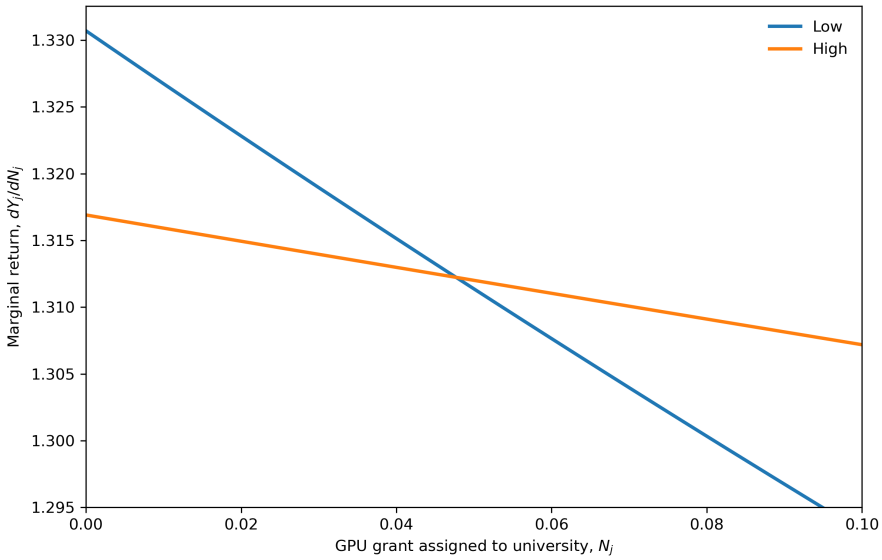
Figure A.6c assigns a grant to one university at a time and plots dY/dN over $N \in [0, 0.1]$. For small grants, the low-resource university has the higher return because its constraint is tighter. But the same grant is large relative to its budget, so its wedge closes quickly and its return falls steeply. The two returns cross at $N \simeq 0.048$, after which the high-resource university's productivity advantage dominates. The preferred recipient therefore depends on productivity, internal resources, and grant size.



(a) Marginal product of capital



(b) Shadow value of the research budget



(c) Marginal output from a grant

Figure A.6: Productivity, Resources, and the Return to a Grant. Ω is productivity, B is the internal research budget, and η is its shadow value. The low-resource university has $(B_L, \Omega_L) = (1, 3)$ and the high-resource university has $(B_H, \Omega_H) = (4, 4.5)$. Panel (a) plots the marginal product of capital over budget and productivity, with contours for η. Panel (b) holds each university's productivity fixed and varies its budget. Panel (c) assigns a grant to one university at a time: the low-resource university has the higher return for small grants, but its return falls faster as its constraint relaxes. Returns cross at $N \simeq 0.048$. All quantities are in normalized units, with capital and labor prices equal to one.

A4 The University's Problem and the Allocation Rule

A4.1 Full Problem

University j enters year t with CPU capital C_{jt}, GPU capital G_{jt}, faculty $\mathbf{L}^F_{jt} = \{L^F_{jft}\}_f$, and productivity $\boldsymbol{\omega}_{jt} = \{\omega_{jft}\}_f$. It values a paper in field f at λ_{jf} dollars and maximizes the value of research output net of input costs. Its value function is

$$V(C_{jt}, G_{jt}, \mathbf{L}^F_{jt}, \boldsymbol{\omega}_{jt}) = \max_{\substack{I^C_{jt}, I^G_{jt}, \{H_{jft}\}_f, \\ \{L^V_{jft}\}_f, \{C_{jft}\}_f, \{G_{jft}\}_f}} \Big\{ \sum_f \lambda_{jf}\, Y_{jft} - \text{Cost}_{jt} + \beta\, \mathbb{E}\big[V\big(C_{jt+1}, G_{jt+1}, \mathbf{L}^F_{jt+1}, \boldsymbol{\omega}_{jt+1}\big) \,\big|\, \boldsymbol{\omega}_{jt}\big] \Big\} \tag{A4.1}$$

where spending on inputs is

$$\text{Cost}_{jt} \equiv \sum_f W^V_{jft}\, L^V_{jft} + \sum_f W^F_{jft}\, L^F_{jft} + P^C_t\, I^C_{jt} + P^G_t\, I^G_{jt} \tag{A4.2}$$

with W^V and W^F the wages for flexible and quasi-fixed labor and P^C_t and P^G_t the capital prices. The university maximizes subject to the laws of motion

$$C_{jt+1} = (1-\delta_C)\, C_{jt} + I^C_{jt}, \qquad G_{jt+1} = (1-\delta_G)\, G_{jt} + I^G_{jt}, \qquad L^F_{jft+1} = (1-\delta_F)\, L^F_{jft} + H_{jft} \tag{A4.3}$$

the productivity process (5.8), the capacity constraints

$$\sum_f C_{jft} \le C_{jt}, \qquad \sum_f G_{jft} \le G_{jt} \tag{A4.4}$$

and the budget constraint

$$\text{Cost}_{jt} \le B_{jt} \tag{A4.5}$$

where B_{jt} is total available funding. The budget constraint implies that the university cannot borrow against future research output. Because the objective is net of costs, the constraint need not bind: a university with a large enough budget stops hiring and investing where the value of the marginal product equals the input price, and spends the remainder outside the research enterprise.

The budget B_{jt} is assumed to be exogenous: it is set by endowment payout rules, appropriations, and indirect-cost recovery, which the university takes as given, so it enters the problem as a known sequence rather than as a state variable with a law of motion of its own. Investment I^C_{jt} and I^G_{jt} chosen in year t become capital in year $t+1$. In the main text, this is the choice made at the end of year $t-1$ for year t, and output in year t uses the stock installed by $t-1$ because of the publication lag.

A4.2 Deriving the Allocation Rule

Let μ^C_{jt}, μ^G_{jt}, and η_{jt} be the multipliers on the CPU, GPU, and budget constraints. The continuation value depends only on next year's state, which is determined by the dynamic choices $(I^C_{jt}, I^G_{jt}, H_{jft})$ and the current state. It therefore does not depend on the static choices $(C_{jft}, G_{jft}, L^V_{jft})$ and drops out of their first-order conditions. Multiplying each first-order condition by its choice variable,

$$\mu^C_{jt}\, C_{jft} = \alpha^K_f\, s^{\mathcal{K}}_C(jft)\, \lambda_{jf} Y_{jft}, \tag{A4.6}$$

$$\mu^G_{jt}\, G_{jft} = \alpha^K_f\, s^{\mathcal{K}}_G(jft)\, \lambda_{jf} Y_{jft}, \tag{A4.7}$$

$$(1+\eta_{jt})\, W^V_{jft}\, L^V_{jft} = \alpha^L_f\, s^{\mathcal{L}}_V(jft)\, \lambda_{jf} Y_{jft}, \tag{A4.8}$$

with within-nest value shares

$$s^{\mathcal{K}}_{C}(jft) = \gamma_f \Big(\frac{C_{jft}}{\mathcal{K}_{jft}}\Big)^{\rho_K} = 1 - s^{\mathcal{K}}_{G}(jft), \qquad s^{\mathcal{L}}_{V}(jft) = \phi_f \Big(\frac{L^V_{jft}}{\mathcal{L}_{jft}}\Big)^{\rho_L} = 1 - s^{\mathcal{L}}_{F}(jft) \tag{A4.9}$$

where $\rho_K \equiv (\sigma_K - 1)/\sigma_K$ and $\rho_L \equiv (\sigma_L - 1)/\sigma_L$. Capital allocation carries no cost in the static problem because the cluster is already installed, so costs enter only the labor condition, through the factor $1 + \eta_{jt}$: a dollar spent on PhD labor costs its face value plus the shadow value of the budget. When the budget is slack ($\eta_{jt} = 0$), the university hires until the value of the marginal product equals the wage. This is the unconstrained regime of Section 7. Dividing (A4.6) and (A4.7) by (A4.8) cancels $\lambda_{jf} Y_{jft}$. Under common within-university wages, the multiplier ratio $\mu^C_{jt}/[(1 + \eta_{jt})W^V_{jft}]$ does not depend on the field. Summing the result across fields and imposing that the capacity constraints bind cancels the multipliers, which gives

$$C_{jft} = C_{jt} \cdot \frac{\psi^C_{jft} W^V_{jft} L^V_{jft}}{\sum_h \psi^C_{jht} W^V_{jht} L^V_{jht}}, \qquad \psi^C_{jft} \equiv \frac{\alpha^K_f}{\alpha^L_f} \cdot \frac{s^{\mathcal{K}}_{C}(jft)}{s^{\mathcal{L}}_{V}(jft)} \tag{A4.10}$$

and analogously for G_{jft} with $\psi^G_{jft} \equiv (\alpha^K_f/\alpha^L_f)\, s^{\mathcal{K}}_{G}(jft)/s^{\mathcal{L}}_{V}(jft)$. Under the common-wage assumption $W^V_{jft} = W^V_{jt}$, wages cancel and we obtain (5.11). The rule is not closed-form because the shares depend on the allocation. Given parameters, we solve it by iterating with a starting guess of the Cobb–Douglas allocation ($\sigma_K = \sigma_L = 1$, where shares are the constants γ_f and ϕ_f) until allocations change by less than 10^{-8}. With $\sigma_L = 1$, as in our baseline, $s^{\mathcal{L}}_{V}(jft) = \phi_f$.

A4.2.1 Utilization

Let $u_{jt} \in (0, 1]$ be the utilization rate of the cluster, so that effective capital is $u_{jt} C_{jt}$. If u_{jt} does not vary across fields within a university-year, it cancels from the shares, $C_{jft}/C_{jt} = u_{jt} C_{jft}/(u_{jt} C_{jt})$. Common utilization changes the scale of effective capital. Treating it as measurement error also requires the orthogonality conditions used in estimation.

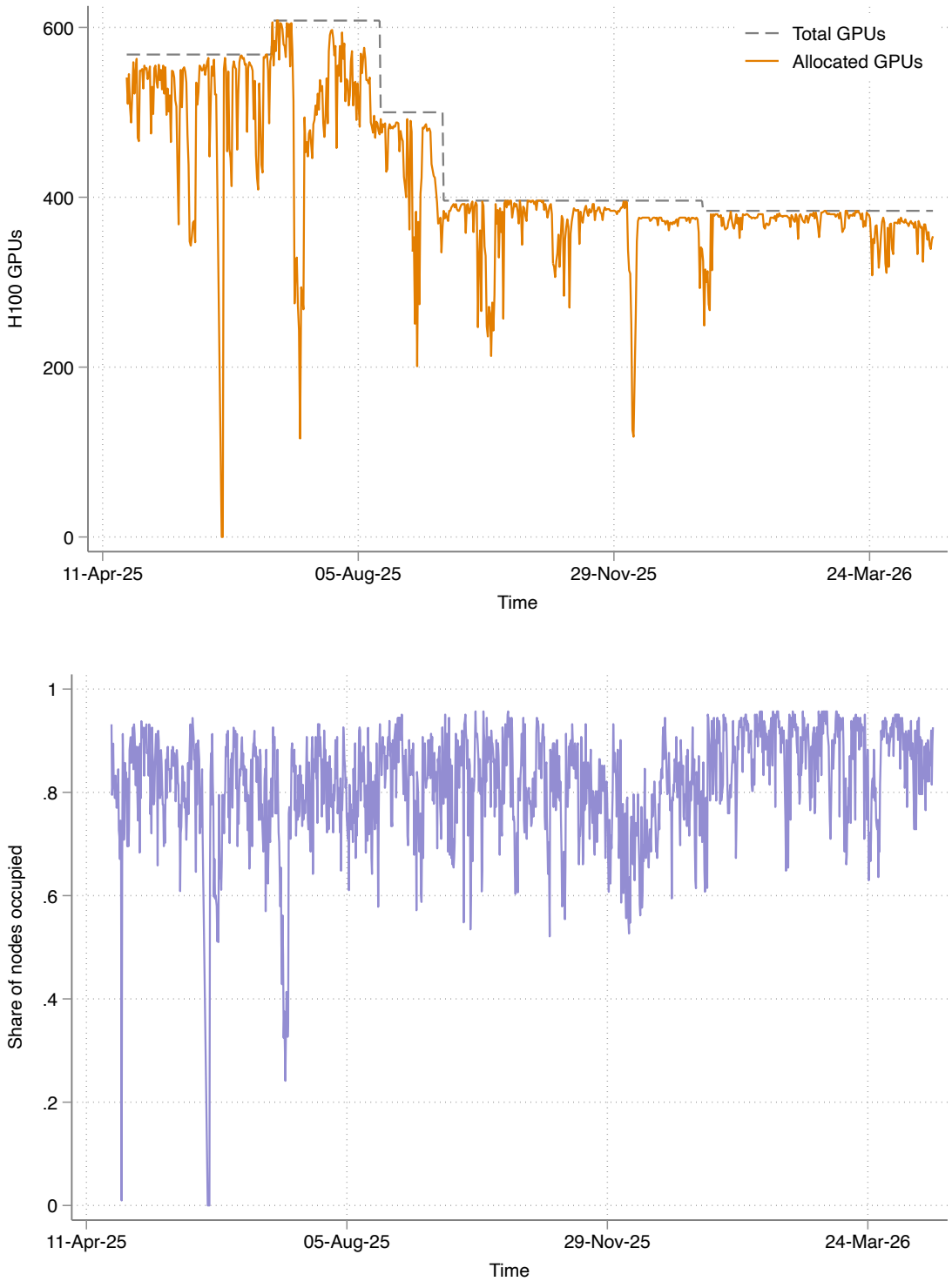


Figure A.7: Harvard Research Computing Total Utilization for NVIDIA H100 GPUs and All Nodes. High-frequency utilization on Harvard's FASRC cluster, April 2025–March 2026. Top panel: total installed NVIDIA H100 GPUs (dashed) versus GPUs allocated to jobs (solid); allocation tracks the available stock closely, and the total steps down as the pool is resized. Bottom panel: share of all compute nodes occupied, fluctuating around 0.8. The series describe one university and motivate the capacity assumptions in Section 5.2.3; they do not establish common utilization across fields or universities.

A5 Evidence on Within-University Wage Differences

The model assumes that the wage W_{jt}^V of flexible research labor (PhD students and postdocs) varies by university × year but stays common across all fields f. The evidence below concerns average field differences in PhD stipends. Small average differences do not establish equal wages within each university-year since university-specific differences could offset one another.

We use self-reported PhD stipends from `phdstipends.com`, a crowd-sourced site on which students report their university, department, academic year, and annual stipend. The raw file contains 17,933 reports spanning 2002 to 2026. We keep reports from 2011 onward (before which coverage is thin), drop exact duplicates, and truncate pay at the within-year 1st and 99th percentiles to remove data-entry errors. We string-match university names to the IPEDS database and obtain 15,350 self-reports across 334 U.S. universities from 2011 to 2025. Department names are matched to eight broad fields: computer science, engineering, geosciences, life sciences, mathematics and statistics, physical sciences, psychology, and social sciences. The final sample consists of 11,513 self-reports across 314 universities.

School identity explains stipend variation. Table A.5 reports the R^2 (and adjusted R^2) from regressing log stipend on combinations of year, university, and field fixed effects. University identity alone explains 24% of the variance in log stipends, and the broad field alone explains only 2%. Adding field fixed effects on top of university × year effects raises the R^2 by just 1.2 percentage points. In other words, once we know the university and the year, the broad field a student is in is nearly uninformative about their stipend.

A natural concern is that the high university R^2 is mechanical with 314 university dummies. Two checks in Table A.5 address this concern. First, the adjusted R^2 (0.246) on the additive university specification is similar to the raw R^2 (0.267). Second, column (3) re-runs the decomposition on a random sample of 20 universities drawn from the 100 most-reported institutions (for a total of 2,068 observations). The adjusted and raw R^2 are nearly identical (0.172 vs. 0.179), and university still has two to three times the explanatory power that field does. Column (2) additionally drops the roughly 3% of surviving reports below $8,000, which are concentrated in monthly or partial-year figures. These reports inflate the residual variance and mechanically depress every R^2 level, but they leave the university-versus-field *ranking* intact. The university explains 48% of the variance, and field, added on top of university × year, still adds only 4 percentage points.

Within-university laboratory-science premia. Within university × year cells, we estimate the gap in stipend across fields:

$$\log w_r \;=\; \alpha_{s(r)\times t(r)} \;+\; \sum_g \beta_g \,\mathbf{1}\{f(r) \in g\} \;+\; \varepsilon_r, \tag{A5.1}$$

where r is a self-report, $s(r)$ its university, $t(r)$ its year, and g its broad field, with life sciences as the omitted base. β_g measures the mean log stipend gap of field g relative to life sciences in the *same* university and year. Standard errors are clustered by university. Table A.6 reports all seven premia. Laboratory sciences (engineering, physical sciences, geosciences, and computer science) have stipend gaps of about four log points or less relative to life sciences once the sub-$8,000 reports are removed

(column 2). The systematic departures are in the non-laboratory fields: mathematics and statistics, psychology, and social sciences are paid 14 to 19 log points less than life sciences within the same university × year across both samples.

Table A.5: Variance decomposition of log PhD stipends

Fixed effects	(1) All	(2) ≥ $8k	(3) 20 of top 100
Year	0.026 (0.025)	0.084 (0.083)	0.044 (0.038)
University	0.238 (0.217)	0.479 (0.464)	0.179 (0.172)
Broad field	0.021 (0.021)	0.068 (0.068)	0.058 (0.055)
Year + University	0.267 (0.246)	0.575 (0.563)	0.220 (0.208)
+ Broad field	0.284 (0.262)	0.630 (0.619)	0.267 (0.253)
University × Year	0.403 (0.253)	0.679 (0.597)	0.314 (0.222)
+ Broad field	0.415 (0.268)	0.720 (0.649)	0.352 (0.262)
Observations	11,513	11,164	2,068
Universities	314	309	20

Notes: Each cell reports the R^2 (and, in parentheses, the adjusted R^2) from a regression of log stipend on the indicated fixed effects. "Broad field" is the eight-way field classification. Column (1) is the full sample; column (2) additionally drops the ≈ 3% of reports below $8,000 (largely monthly or partial-year amounts); column (3) is a random draw of 20 universities from the 100 most-reported institutions. The sample is from `phdstipends.com`, 2011–2025.

Table A.6: Within university-year field premia in log PhD stipends

Field (vs. life sciences)	(1) All S&E	(2) S&E, ≥ $8k
Engineering	-0.052^{***}	0.010
	(0.020)	(0.011)
Physical sciences	−0.020	0.004
	(0.018)	(0.009)
Geosciences	−0.050	−0.027
	(0.032)	(0.018)
Computer science	-0.109^{***}	-0.036^{**}
	(0.035)	(0.015)
Mathematics & statistics	-0.141^{***}	-0.142^{***}
	(0.027)	(0.016)
Psychology	-0.190^{***}	-0.162^{***}
	(0.022)	(0.014)
Social sciences	-0.155^{***}	-0.164^{***}
	(0.018)	(0.014)
Observations	10,324	9,998
University × year cells	1,408	1,385

Notes: Coefficients from equation (A5.1): log stipend on broad-field indicators, absorbing university × year fixed effects, with life sciences the omitted category. Estimation uses only university-year cells containing at least two distinct broad fields. Standard errors clustered by university in parentheses. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

A6 Estimation Details

A6.1 Labor Nest

With $\sigma_L = 1$, the labor aggregate is $\mathcal{L}_{jft} = (L^V_{jft})^{\phi_f}(L^F_{jft})^{1-\phi_f}$, and cost minimization implies that ϕ_f equals the flexible-labor share of the field's research wage bill. We compute flexible-labor costs from PhD stipends and NIH NRSA postdoctoral stipend rates, and faculty research costs as 40% of faculty salaries, allocating faculty to fields by their share of R&D expenditure.

A6.2 CPU–GPU Substitution

Mapping the price response into σ_K. Let π_C and π_G be the responses of log CPU and log GPU capacity to the instrument Z^{NVDA}_{jt}, and $\Theta \equiv \pi_C/\pi_G$. With CES capital and GPU and CPU spending shares s_G and $s_C = 1 - s_G$, the ratio of the cross-price to own-price response is

$$\Theta(\sigma_K) = \frac{s_G\,(1/\sigma_K - \psi)}{s_G/\sigma_K + \psi\, s_C}, \qquad \text{so} \qquad \hat{\sigma}_K = \frac{s_G\,(1-\hat{\Theta})}{\psi\,(s_G + \hat{\Theta}\, s_C)} \tag{A6.1}$$

where ψ is the inverse of the price elasticity of demand for the capital aggregate. With the Cobb–Douglas outer nest, a university that equates the marginal value of its capital aggregate to the user cost has a demand elasticity of $1/(1-\alpha^K_f)$, so $\psi = 1 - \bar{\alpha}^K$, where $\bar{\alpha}^K = 0.05$ averages the baseline estimates of Table 6 with the fields' shares of GPU-topic output; this gives $\psi = 0.95$. Complements ($\sigma_K < 1/\psi = 1.05$) imply $\Theta > 0$: when GPUs become cheaper, universities buy more CPUs as well. Θ exceeds $\Theta(\sigma_K{=}1) = 0.02$ exactly when $\sigma_K < 1$.

GPU share of cluster spending. We measure s_G as the GPU share of the annual user cost of the installed stock:

$$s_G = \frac{u_G\,\bar{G}}{u_G\,\bar{G} + u_C\,\bar{C}}, \qquad u_X = (r + \delta_X - \pi_X)\,p_X \tag{A6.2}$$

with $r = 0.04$ and $\delta_G = \delta_C = 0.20$, the depreciation rate implied by a five-year replacement cycle. Prices p_X are the geometric means of the yearly median prices in GovSpend purchase orders at public universities (Table A.3): \$38 per FP16 TFLOPS (2018–2025) and \$84 per CPU core (2015–2025). Price trends π_X are the mean annual log changes in the yearly medians between the first and last year (-0.17 for GPUs and $+0.02$ for CPUs), so the user cost of a GPU includes the expected decline in its value. Quantities $\bar{G}$ and $\bar{C}$ are geometric means of installed FP16 TFLOPS and CPU cores over university-years with both kinds of capacity. This gives $u_G = \$15.5$ per TFLOPS-year, $u_C = \$18.0$ per core-year, and $s_G = 0.36$.

A6.3 System GMM

For each field, we estimate the differenced and levels versions of the quasi-differenced production function with five unrestricted coefficients:

$$\Delta y_{jft} = \pi_1 \Delta y_{jf,t-1} + \pi_2 \Delta k_{jf,t-1} + \pi_3 \Delta k_{jf,t-2} + \pi_4 \Delta \ell_{jft} + \pi_5 \Delta \ell_{jf,t-1} + \Delta u_{jft}, \tag{A6.3}$$

$$y_{jft} = \pi_1 y_{jf,t-1} + \pi_2 k_{jf,t-1} + \pi_3 k_{jf,t-2} + \pi_4 \ell_{jft} + \pi_5 \ell_{jf,t-1} + (1-\varrho_f)\bar{\omega}_{jf} + u_{jft}. \tag{A6.4}$$

The common-factor restrictions are $\pi_1 = \varrho_f$, $\pi_2 = \alpha^K_f$, $\pi_3 = -\varrho_f \alpha^K_f$, $\pi_4 = \alpha^L_f$, and $\pi_5 = -\varrho_f \alpha^L_f$. We stack the moments from both equations (21 moments for 5 coefficients), estimate by two-step

GMM with university-clustered weights, and recover $(\alpha_f^K, \alpha_f^L, \varrho_f)$ by minimum distance. Standard errors use the delta method, clustered by university, and condition on the calibrated nests and the allocation.

Instruments. Table A.7 lists the instruments. The differenced equation uses lagged levels, which are valid if inputs are chosen before the current productivity shock. The levels equation uses lagged changes, which are valid if changes in CPU capital and flexible labor are uncorrelated with the permanent effect. We do not impose this assumption for GPU capital. The leave-one-out instrument for capital, $z^{\text{loo}}_{jf,t-s} = \log \sum_{h \neq f} L^V_{jh,t-s}$ for $s \in \{2, 3\}$, hinges on the allocation rule: more flexible labor in other fields lowers field f's share of the cluster.

Table A.7: Moment Conditions of the System GMM Estimator

Equation	Instrument	Dates	Moments
Differenced equation, $E[z_{jf,t-s}\,\Delta u_{jft}] = 0$			
	$\log C_{j,t-s}$ (CPU cores)	$s = 1, 2, 3$	3
	$\log G_{j,t-s}$ (GPU TFLOPS)	$s = 1, 2, 3$	3
	$\log L^V_{jf,t-s}$	$s = 2, 3$	2
	$\log \sum_{h \neq f} L^V_{jh,t-s}$ (other fields' flexible labor)	$s = 2, 3$	2
	$y_{jf,t-3}$	$s = 3$	1
	$\Delta Z^{\text{NVDA}}_{jt}$	t	1
	$\Delta z^{\text{TB}}_{jft},\ \Delta z^{\text{PP}}_{jft},\ \Delta z^{\text{FP}}_{jft}$	t	3
Levels equation, $E[\Delta z_{jf,t-1}\,(u_{jft} + (1-\varrho_f)\bar{\omega}_{jf})] = 0$			
	$\Delta \log C_{j,t-1}$	$t-1$	1
	$\Delta \log L^V_{jf,t-1}$	$t-1$	1
	Z^{NVDA}_{jt}	t	1
	$z^{\text{TB}}_{jft},\ z^{\text{PP}}_{jft},\ z^{\text{FP}}_{jft}$	t	3
Total			21

Notes: Lagged levels are valid in the differenced equation if inputs are chosen before the current innovation and output measurement error is serially uncorrelated. Quasi-differencing then makes u_{jft} an MA(1) process, and $y_{jf,t-3}$ is the first valid output lag. Lagged changes are valid in the levels equation if changes in cores and flexible labor are uncorrelated with the permanent effect. The current change in cores and all changes in GPU capacity are excluded. The weight matrix is the Arellano–Bond first-step matrix in step one and the university-clustered moment covariance in step two. The common-factor restrictions are imposed by minimum distance with the inverse of the clustered covariance of the unrestricted coefficients as weights. The visa-policy shocks z^{TB}, z^{PP}, and z^{FP} are defined below.

Visa-policy instruments. From the NSF GSS microdata, we observe temporary-visa counts of full-time graduate students and postdocs by university, field, and year. Exposure is the field's pre-period temporary-visa share,

$$s^{\text{grad}}_{jf} = \frac{\text{temporary-visa graduate students}}{\text{graduate students}}\Big|_{2012\text{–}16}, \qquad s^{\text{pd}}_{jf} = \frac{\text{temporary-visa postdocs}}{\text{postdocs}}\Big|_{2005\text{–}09}. \tag{A.6.5}$$

The shocks are

$$z^{\text{TB}}_{jft} = s^{\text{grad}}_{jf} \times \mathbf{1}[t \geq 2017] \qquad \text{(travel ban, EO 13769)}, \tag{A.6.6}$$

$$z^{\text{PP}}_{jft} = s^{\text{grad}}_{jf} \times \mathbf{1}[t \geq 2020] \qquad \text{(Proclamation 10043)}, \tag{A.6.7}$$

$$z^{\text{FP}}_{jft} = s^{\text{pd}}_{jf} \times \left[\log N^{-j}_{ft} - \overline{\log N^{-j}_{f}}\Big|_{2005\text{–}09}\right] \qquad \text{(national foreign-postdoc inflow)} \tag{A.6.8}$$

where N_{ft}^{-j} is the number of temporary-visa postdocs in field f at all GSS institutions other than j. Universities with more foreign students or postdocs before the shocks are hit harder after them. Levels instrument the levels equation, and first differences instrument the differenced equation. The sample starts in 2010 so the postdoc exposure is measured in 2005–2009 and graduate exposure in 2012–2016.

A6.4 Additional Results

Table A.8: Output Elasticities by Input

	CS	Engineering	Geo./Phys.	Mol./Mat.	Life
Computing capital	**0.057** (0.020)	0.027 (0.023)	**0.051** (0.021)	**0.054** (0.025)	0.053 (0.028)
PhD students and postdocs	**0.359** (0.112)	**0.668** (0.089)	**0.535** (0.138)	0.349 (0.180)	**0.319** (0.051)
Faculty	**0.172** (0.054)	**0.277** (0.037)	**0.335** (0.087)	0.088 (0.045)	**0.242** (0.039)
Returns to scale	0.59	0.97	0.92	0.49	0.61

Notes: With $\sigma_L = 1$, output is $Y = \Omega \mathcal{K}^{\alpha^K} (L^V)^{\phi_f \alpha^L} (L^F)^{(1-\phi_f)\alpha^L}$, so the labor elasticity splits between PhD students and postdocs (L^V) and faculty (L^F) by the calibrated share ϕ_f; only the total labor elasticity is estimated. Returns to scale is $\alpha_f^K + \alpha_f^L$. Estimates from Table 6; standard errors in parentheses, clustered by university. Bold: at least 1.96 standard errors from zero.

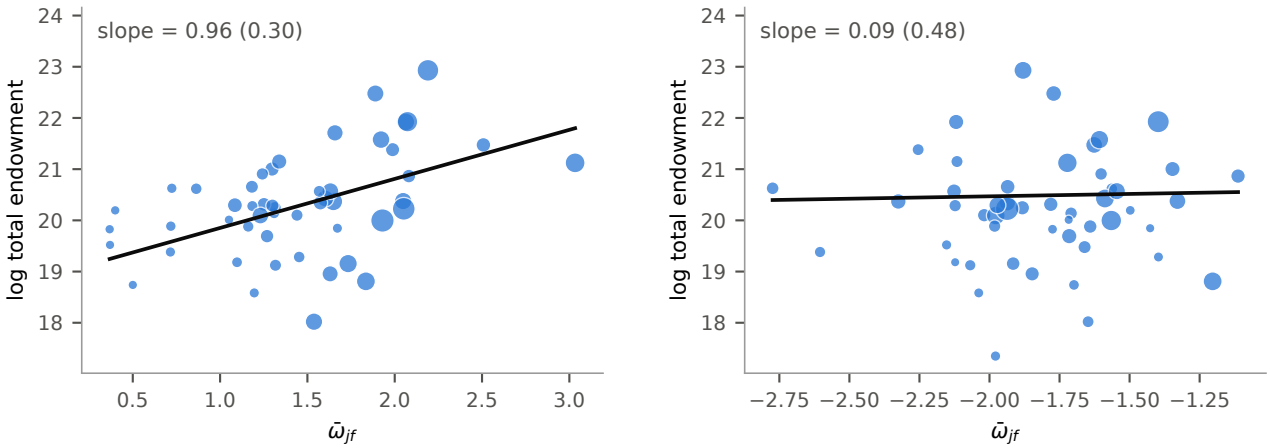
Table A.9: Robustness of the Capital Elasticity $\hat{\alpha}_f^K$

	CS	Engineering	Geo./Phys.	Mol./Mat.	Life
Baseline	**0.057** (0.020)	0.027 (0.023)	**0.051** (0.021)	**0.054** (0.025)	0.053 (0.028)
$\sigma_L = 1.5$	**0.048** (0.018)	0.027 (0.023)	**0.054** (0.021)	**0.056** (0.024)	0.044 (0.029)
$\sigma_L = 2$	**0.040** (0.019)	0.027 (0.023)	**0.055** (0.021)	**0.056** (0.024)	0.038 (0.030)
$\sigma_K = 0.4$	**0.053** (0.017)	0.022 (0.021)	**0.038** (0.018)	0.040 (0.023)	0.034 (0.027)
$\sigma_K = 1$	**0.064** (0.020)	0.031 (0.025)	**0.053** (0.023)	**0.063** (0.025)	**0.064** (0.029)
Uniform $\gamma_f = 0.5$	**0.068** (0.017)	0.024 (0.022)	0.023 (0.016)	0.040 (0.024)	**0.055** (0.027)
Uniform $\phi_f = 0.5$	**0.048** (0.016)	0.035 (0.022)	**0.048** (0.022)	0.032 (0.025)	**0.059** (0.027)
α_f^L fixed at 0.25	**0.054** (0.018)	0.029 (0.023)	**0.053** (0.023)	**0.049** (0.024)	0.053 (0.028)
α_f^L fixed at 0.50	**0.059** (0.020)	0.029 (0.023)	**0.053** (0.022)	**0.056** (0.025)	0.053 (0.028)
α_f^L fixed at 0.75	**0.053** (0.021)	0.028 (0.023)	**0.052** (0.021)	**0.064** (0.026)	0.037 (0.029)
Output: no private-sector co-authors	**0.048** (0.022)	0.029 (0.021)	0.033 (0.023)	0.026 (0.023)	0.055 (0.030)
Output: first- or last-author credit	**0.061** (0.020)	0.014 (0.029)	**0.061** (0.024)	0.038 (0.031)	0.043 (0.025)
Capital: FP32 TFLOPS	**0.041** (0.017)	0.041 (0.023)	**0.032** (0.015)	**0.106** (0.025)	0.000 (0.077)
Capital: dollars at transaction price, all cards	**0.033** (0.010)	0.022 (0.022)	0.029 (0.016)	**0.047** (0.017)	0.000 (0.022)
Year effects	**0.045** (0.021)	0.000 (0.019)	0.032 (0.017)	**0.048** (0.023)	0.006 (0.039)
Output: citation-percentile weighted	**0.088** (0.022)	0.030 (0.029)	0.000 (0.027)	0.031 (0.030)	0.000 (0.036)

Notes: Each row re-estimates the production function of Table 6 under the stated change and reports $\hat{\alpha}_f^K$ with its standard error in parentheses (clustered by university). Rows that change σ_K hold the CPU weights γ_f at the baseline calibration. Rows that fix α_f^L hold the labor elasticity at the stated value in every field and re-solve the allocation. Citation-percentile weighting replaces each paper's credit by its citation percentile within field and year. "No private-sector co-authors" drops papers with a co-author at a firm. "First- or last-author credit" counts a paper for a university only when one of its authors is the first or the last author, instead of splitting credit equally across authors. Dollar capital values each installed GPU at its transaction price in the year of purchase. Bold: at least 1.96 standard errors from zero; an estimate of 0.000 lies on the boundary of the parameter space.

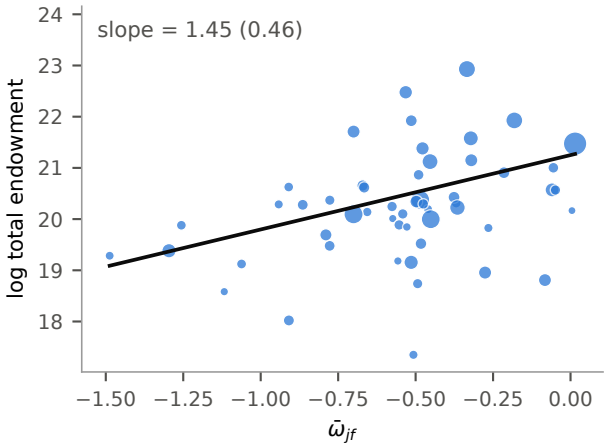
Figure A.8 repeats Figure 9 with the log of total endowment in place of endowment per faculty member, with the same weights. The slopes are steeper in computer science, the molecular and materials sciences, and the life sciences (0.96 (0.30), 0.64 (0.29), and 1.10 (0.28)), similar in geosciences

and physics (1.45 (0.46)), and the slope in engineering, 0.09 (0.48), remains indistinguishable from zero. Part of the difference is scale: with returns to scale below one, $\bar{\omega}_{jf}$ rises with the size of the university at given output per unit of input, and total endowment rises with size, so the per-faculty measure in the text is the cleaner comparison of financial access.

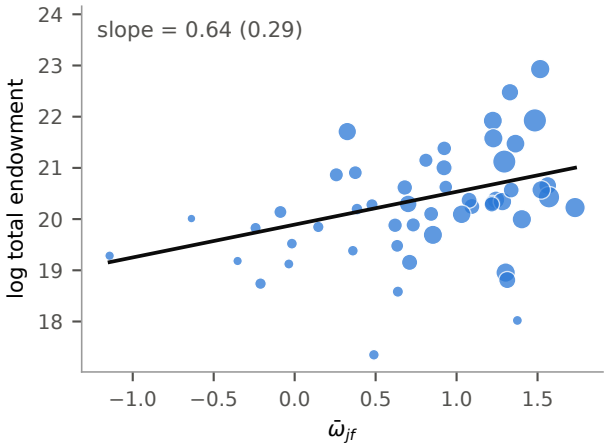


(a) Computer science

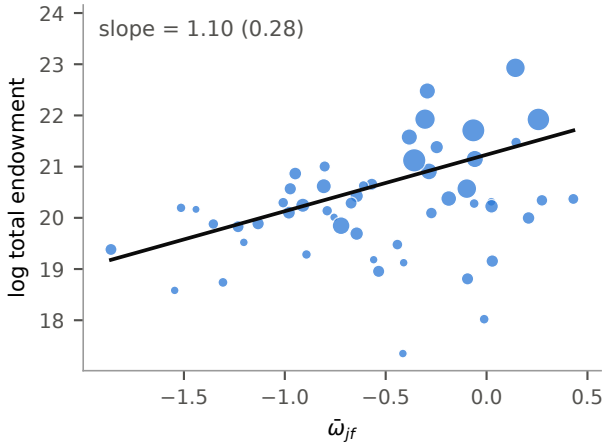
(b) Engineering



(c) Geosciences and physics



(d) Molecular and materials sciences



(e) Life sciences

Notes: As in Figure 9, with the log of total endowment (FY2012–15 mean, 2017 dollars) on the vertical axis. Points are sized and the line is weighted by the university's mean number of full-time-equivalent authors in the field over 2005–2023. Each panel reports the slope of log endowment on $\bar{\omega}_{jf}$ with the standard error in parentheses.

Figure A.8: Productivity fixed effects and total endowment, by field

A7 Crowd-Out Details

A7.1 From the Dynamic Problem to a Static Spending Rule

The university chooses computing investment at the end of year $t-1$, so the relevant first-order condition involves the continuation value. When the university buys capital of its own,

$$\beta\,\mathbb{E}\left[\frac{\partial V_{t+1}}{\partial K_{t+1}}\right]\frac{1}{P^K} = 1+\eta_{jt} \tag{A7.1}$$

where P^K is the price of a unit of computing capital. The left side is the discounted value of one more dollar of cluster, and it depends on the value weights λ_{jf} through the value of future output. The right side is the cost of a dollar: its face value plus the shadow value of the budget.

The envelope condition for the capital stock is

$$\partial V_t/\partial K_t = \sum_f \lambda_{jf}\,\partial Y_{jft}/\partial K_t + \beta(1-\delta)\,\mathbb{E}[\partial V_{t+1}/\partial K_{t+1}]$$

If the budget multiplier is stable over time, so that $\eta_{jt} \approx \eta_j$, substituting (A7.1) into the envelope condition and using $\beta = 1/(1+r)$ gives

$$\sum_f \lambda_{jf}\,\frac{\partial Y_{jft}}{\partial K_t} = (1+\eta_j)\,(r+\delta)\,P^K \tag{A7.2}$$

In a stationary regime, the dynamic condition collapses to a user-cost condition: the university sets the value of the marginal unit of cluster equal to its rental cost, $(r+\delta)P^K$, marked up by the budget multiplier. This is the sense in which the continuation value drops out. What it leaves behind has the same form as the static first-order condition for flexible labor, $\lambda_{jf}\,\partial Y_{jft}/\partial L^V_{jft} = (1+\eta_{jt})W^V_{jt}$, with the rental cost of capital in place of the wage. When the budget is slack ($\eta_j = 0$), (A7.2) determines a finite target stock that does not depend on the grant; this is the unconstrained regime.

A7.2 Deriving the Capital Share of a Marginal Dollar

Consider a constrained university that buys computing capital of its own. Because the in-kind award can be used only for hardware the university would otherwise have bought, it is equivalent to cash: the university's choices depend only on the effective budget $\widetilde{B}_{jt} = B_{jt} + N_{jt}$. Faculty are fixed within the year, so the adjustable inputs are computing capital and flexible labor. With $\sigma_L = 1$, field output is

$$Y_{jft} = \Omega_{jft}\,\mathcal{K}_{jft}^{\alpha^K_f}\,(L^V_{jft})^{\phi_f\alpha^L_f}\,(L^F_{jft})^{(1-\phi_f)\alpha^L_f} \tag{A7.3}$$

and the cluster is allocated across fields by the rule (5.11), so the value of cluster spending in field f has elasticity α^K_f regardless of the CPU–GPU mix.

Within a field. Let E^K_{jft} denote the user cost of the cluster capital allocated to field f, $E^L_{jft} = W^V_{jt}L^V_{jft}$ spending on flexible labor, and $E_{jft} = E^K_{jft} + E^L_{jft}$ the field's adjustable spending. Multiplying the capital condition (A7.2), applied field by field through the allocation rule, and the labor condition by their respective inputs gives

$$(1+\eta_{jt})\,E^K_{jft} = \alpha^K_f\,\lambda_{jf}Y_{jft}, \qquad (1+\eta_{jt})\,E^L_{jft} = \phi_f\alpha^L_f\,\lambda_{jf}Y_{jft} \tag{A7.4}$$

Taking the ratio cancels $\lambda_{jf}Y_{jft}$ and the multiplier, which gives (7.1). Hence $E^K_{jft} = (\alpha^K_f/\tilde{\nu}_f)\,E_{jft}$: within a field, the capital share of adjustable spending is constant at $\alpha^K_f/\tilde{\nu}_f$, with $\tilde{\nu}_f \equiv \alpha^K_f + \phi_f\alpha^L_f$,

so the same share applies to a marginal dollar. Given the within-field split, field output is $Y_{jft} = A_{jft}\, E_{jft}^{\tilde{\nu}_f}$, where A_{jft} collects productivity, fixed faculty, prices, and the elasticities. Summing (A7.4) across the two inputs, observed spending reveals the dollar value the university places on the field's output: $E_{jft} = \tilde{\nu}_f\, \lambda_{jf} Y_{jft}/(1+\eta_{jt})$, and from the labor condition alone, $E_{jft} = \left[\tilde{\nu}_f/(\phi_f \alpha_f^L)\right] W_{jt}^V L_{jft}^V$.

Across fields. The constrained university chooses $\{E_{jft}\}_f$ to maximize $\sum_f \lambda_{jf} A_{jft} E_{jft}^{\tilde{\nu}_f}$ subject to $\sum_f E_{jft} = \widetilde{B}_{jt}$. The first-order condition is $\lambda_{jf}\tilde{\nu}_f Y_{jft}/E_{jft} = 1 + \eta_{jt}$ for every field. Differentiating with respect to $\widetilde{B}_{jt}$ and using the first-order condition to eliminate $\lambda_{jf} A_{jft}$,

$$(\tilde{\nu}_f - 1)\,\frac{dE_{jft}}{E_{jft}} = \frac{d\eta_{jt}}{1+\eta_{jt}} \quad \Longrightarrow \quad \frac{\partial E_{jft}}{\partial \widetilde{B}_{jt}} = \frac{E_{jft}/(1-\tilde{\nu}_f)}{\sum_h E_{jht}/(1-\tilde{\nu}_h)} \tag{A7.5}$$

The marginal dollar is spread across fields in proportion to current spending scaled by $1/(1-\tilde{\nu}_f)$. A field whose returns are closer to constant has a flatter marginal return, so it absorbs more of the extra dollar before its return falls to the common level. This solution requires $\tilde{\nu}_f < 1$ in every field, which is the second-order condition of the cross-field problem; it holds whenever returns to scale in the production function are decreasing, since $\tilde{\nu}_f < \alpha_f^K + \alpha_f^L$. At the estimates in Table 6, $\tilde{\nu}_f$ ranges from 0.37 in the life sciences to 0.70 in engineering (Table A.10), so the condition holds in every field.

Combining. The capital share of a marginal dollar is the within-field capital share weighted by each field's share of the marginal dollar:

$$\chi_{jt}^K = \sum_f \frac{\alpha_f^K}{\tilde{\nu}_f} \cdot \frac{\partial E_{jft}}{\partial \widetilde{B}_{jt}} = \frac{\sum_f \dfrac{\alpha_f^K}{\tilde{\nu}_f} \cdot \dfrac{E_{jft}}{1-\tilde{\nu}_f}}{\sum_f \dfrac{E_{jft}}{1-\tilde{\nu}_f}} = \frac{\sum_f \alpha_f^K\, w_{jft}}{\sum_f \tilde{\nu}_f\, w_{jft}}, \qquad w_{jft} \equiv \frac{L_{jft}^V}{\phi_f \alpha_f^L (1-\tilde{\nu}_f)} \tag{A7.6}$$

where the last step substitutes $E_{jft} = [\tilde{\nu}_f/(\phi_f \alpha_f^L)] W_{jt}^V L_{jft}^V$ and cancels the common wage. This is (7.2). The constrained-regime crowd-out then follows as in the stylized model: own spending on capital is $\chi_{jt}^K \widetilde{B}_{jt} - N_{jt}$, so $\partial I_{jt}^K/\partial N_{jt} = -(1-\chi_{jt}^K)$. The response to the university's own budget in this regime is $\partial I_{jt}^K/\partial B_{jt} = \chi_{jt}^K$, which is the basis of the budget-shock test in Appendix A7.3.

Because χ_{jt}^K is the capital share of the marginal dollar, the expression $\chi_{jt}^K \widetilde{B}_{jt} - N_{jt}$ for own capital spending holds exactly only if the marginal share is constant along the path. The within-field shares are constant under Cobb–Douglas and the cross-field weights move with spending, so χ_{jt}^K changes little within a university over time; we use the derivative locally, at the observed allocation.

Assumptions. Three assumptions underlie (7.2). First, the outer technology is Cobb–Douglas, which makes the within-field capital share constant and lets the value weights, productivity, and the budget multiplier cancel. Second, the budget multiplier is stable over time, so that the dynamic investment condition reduces to the user-cost condition (A7.2). Third, faculty are fixed over the horizon at which crowd-out is measured, so the adjustable inputs are the cluster and flexible labor; this matches the timing of the production function.

Table A.10 lists the ingredients of (7.2) field by field. Engineering has the lowest within-field capital share, 0.04, because its labor elasticity is close to one, and together with the life sciences it receives about seven-tenths of a marginal dollar: the two fields have the largest flexible-labor bills and, in engineering, the flattest returns. Computer science, where the within-field share is 0.14, receives a

small share of the marginal dollar. This is why χ_{jt}^K is close to 0.1 at every university.

Table A.10: Ingredients of the Capital Share of a Marginal Dollar

	CS	Engineering	Geo./Phys.	Mol./Mat.	Life
Capital elasticity α_f^K	0.057	0.027	0.051	0.054	0.053
Flexible-labor elasticity $\phi_f \alpha_f^L$	0.359	0.668	0.535	0.349	0.319
Adjustable-input returns $\tilde{\nu}_f = \alpha_f^K + \phi_f \alpha_f^L$	0.416	0.696	0.587	0.402	0.372
Within-field capital share $\alpha_f^K / \tilde{\nu}_f$	0.137	0.039	0.087	0.133	0.142
Marginal-dollar weight $1/(1-\tilde{\nu}_f)$	1.71	3.29	2.42	1.67	1.59
Share of the marginal dollar, mean over universities	0.06	0.35	0.14	0.10	0.35

Notes: Elasticities from Table 6 and ϕ_f from Table 5. The last row is the share of a marginal budget dollar that goes to each field, $\tilde{\nu}_f w_{jft} / \sum_h \tilde{\nu}_h w_{jht}$ from (A7.5), averaged over panel university-years 2019–2023.

A7.3 Constrained versus Unconstrained Universities

Section 7.2 assumes that no investing university is unconstrained; this appendix reports the test behind that assumption. Both the constrained and the unconstrained regime invest, so the capacity panel cannot separate them. What does separate them is the response to the university's own budget. A constrained university whose budget increases by a dollar puts χ_{jt}^K of it into the cluster, and the rest into PhD students and postdocs. An unconstrained university is already at the stock where the value of the marginal unit equals its price, so an extra dollar changes nothing. Among investing universities in a group τ, if a fraction ω_τ is unconstrained, the average response of own research spending to the budget is $(1-\omega_\tau)$ times the constrained response. We estimate this response for all R1 universities using the budget variables in IPEDS Finance, allowing it to differ across terciles of endowment per faculty:

$$I_{jt} = \alpha_j + \delta_t + \sum_\tau \mathbf{1}[\tau(j) = \tau] \left(\beta_\tau^B B_{jt} + \beta_\tau^A A_{jt} \right) + e_{jt} \tag{A7.7}$$

where I_{jt} is institutionally financed R&D expenditure or non-federal capitalized equipment expenditure from the HERD Survey, B_{jt} is the sustainable endowment payout (the average effective spending rate, 0.045, times the trailing three-year endowment), and A_{jt} is state and local appropriations, all in dollars. The university fixed effects absorb the level of resources and the coefficients are identified by changes in a university's budget over time, which is dominated by market returns for the endowment. The unconstrained regime predicts $\beta_\tau^B = \beta_\tau^A = 0$ while the constrained regime predicts a positive response. The level of β_τ^B identifies ω_τ only under the assumption that the whole marginal dollar is research spending.

Table A.11 reports that the estimated research-spending response is positive in every endowment tercile. Institutionally financed R&D rises by 0.274 dollars (0.085) per dollar of sustainable payout in the top tercile and by 0.275 (0.145) in the middle tercile; the bottom-tercile response is larger but imprecise. Non-federal equipment spending responds in every tercile (joint p-value 0.000), and in the top tercile it also responds to appropriations. The response to market returns alone (Panel B), which isolates the exogenous part of the payout, is too imprecise to be informative. The equipment share of the research response (Panel D) brackets the model's χ^K of about 0.097: it is 0.190 (0.081) in the bottom tercile, 0.286 (0.178) in the middle, and 0.059 (0.020) at the top, and it is

larger within the computing fields, where equipment takes a fifth to three-fifths of a marginal institutional research dollar. These results reject a zero average spending response, but do not identify the fraction of unconstrained universities. Thus, we assume there are no unconstrained universities. The equipment share is of the same order as the model prediction, though lower at the best-endowed universities and higher elsewhere. Moving an investing university from constrained to unconstrained reduces pass-through from χ^K to zero, a difference of about ten cents per grant dollar.

Table A.11: Response of Institutionally Financed Spending to University Budgets, by Tier

	Institutional R&D funds		Non-federal capitalized equipment	
	All fields	Computing fields	All fields	Computing fields
Mean ($M/year)	92.2	24.0	5.6	2.7
Panel A: endowment payout, $ per $ (spending rate × trailing three-year endowment)				
Low	1.287 (0.804)	0.220 (0.384)	0.245 (0.091)	0.107 (0.055)
Medium	0.275 (0.145)	0.120 (0.046)	0.079 (0.016)	0.072 (0.023)
High	0.274 (0.085)	0.042 (0.015)	0.016 (0.005)	0.009 (0.003)
Joint p-value, no response in any tier	0.004	0.005	0.000	0.000
Panel B: endowment return shock, $ per $ (lagged endowment × common return, cumulative lags 1–3)				
Low	0.090 (1.243)	−0.072 (0.347)	−0.116 (0.098)	0.075 (0.078)
Medium	0.369 (0.400)	0.036 (0.100)	−0.070 (0.033)	0.006 (0.050)
High	0.092 (0.195)	0.021 (0.027)	0.007 (0.012)	0.003 (0.006)
Joint p-value, no response in any tier	0.806	0.839	0.052	0.756
Panel C: state and local appropriations, $ per $				
Low	−0.008 (0.072)	0.005 (0.020)	−0.003 (0.006)	−0.002 (0.003)
Medium	0.073 (0.040)	−0.007 (0.013)	0.010 (0.006)	0.000 (0.003)
High	0.064 (0.112)	0.003 (0.036)	0.027 (0.005)	0.012 (0.006)
Joint p-value, no response in any tier	0.280	0.944	0.000	0.205
Panel D: equipment share of the payout response, $\hat{\beta}_\tau^{equip}/\hat{\beta}_\tau^{R\&D}$ (model: $\bar{\chi}_\tau^K$)				
Low	0.19 (0.08)		0.50 (0.74)	
Medium	0.29 (0.18)		0.60 (0.33)	
High	0.06 (0.02)		0.21 (0.07)	
Observations	2,729	2,642	2,729	2,621
Universities	184	179	184	179

Notes: Outcomes are HERD institutionally financed R&D expenditure and non-federal capitalized equipment expenditure ($M) at R1 universities, FY2010–2024, for all fields and for computer science, engineering, physical sciences and geosciences combined. Budget variables from IPEDS Finance, in $M. Panel A: the sustainable payout, 0.045 (the average effective spending rate) times the mean of the endowment at the end of the three preceding fiscal years; Panel C: state and local appropriations, in the same regression. Panel B: a separate regression on the spending rate times the beginning-of-year endowment times the median return across R1 endowments in that year, cumulated over lags one to three, controlling for the three lagged endowment levels (each by tier) and for the cumulative change in appropriations, so that the response is identified by market returns rather than endowment growth. All regressions include university and year fixed effects and interact every budget variable with tier; standard errors in parentheses are clustered by university. Panel D is the ratio of the Panel A equipment and R&D coefficients in the same tier (delta-method standard errors); for a constrained university the model predicts $\bar{\chi}_\tau^K$ whatever share of a budget dollar reaches research. Tiers are terciles of endowment per tenure-line faculty among R1 universities.

A7.4 Test on Institutionally Financed R&D Spending

The capacity panel covers 53 universities and 36 award-years. The NSF HERD Survey covers every R1 university, so it offers a test with four times the award dollars, at the cost of an outcome that is a proxy for own computing investment. The survey provides institutionally financed R&D expenditure and non-federal capitalized equipment expenditure by field for 186 R1 universities from 2013 to 2023.

For each university and year, we construct the model-implied response to its awards,

$$\widehat{I}_{jt} = \left(\frac{\partial I^K}{\partial N} \right)_{jt} N_{jt}$$

where N_{jt} is the dollar value of MRI and CC* computing awards made in year t and the response is (7.4), evaluated at the university's own χ^K_{jt} and its predicted corner probability.

We then estimate the field-level analogue of (7.5),

$$I_{jft} = \alpha_j + \delta_t + \sum_{\ell=0}^{2} \lambda_\ell \, \widehat{I}_{j,t-\ell} + \gamma \log(\text{Endowment})_{jt} + e_{jft} \tag{A7.8}$$

where I_{jft} is institutionally financed R&D spending, or non-federal capitalized equipment spending, in field f. If the model's crowd-out is correct, the cumulative response $\sum_\ell \lambda_\ell$ equals one in fields that use computing capital. In fields that do not, such as the social sciences, computing awards should not displace spending, and $\sum_\ell \lambda_\ell = 0$. We date awards by the year they are made and sum the response over that year and the next two, because the money is spent over the following budget cycles and a university that receives hardware in kind can spread the offsetting cut in its own purchases over the same period, so the response in the award year alone understates crowd-out. The model's prediction is for the university's total own spending on computing capital, which HERD does not report as a line. The field outcomes test where the offset shows up: institutional R&D funds in a field capture any reduction in the university's own support for that field, while non-federal equipment captures only reductions in capitalized purchases.

The data do not reject the implied crowd-out in institutionally financed research spending, and the placebo fields show no response, but the test has limited power (Table A.12). In computer science, the cumulative response of institutional funds is 0.73 (0.60), consistent with one; it is also consistent with zero. The estimates for engineering and the physical sciences are centered above one, and the life sciences below zero, all with standard errors several times larger. In the social sciences the response is −0.28 (0.59), and in mathematics 0.43 (0.29); neither rejects zero. Equipment tells a more mixed story. In the physical sciences and geosciences, non-federal equipment falls by about half the implied amount, 0.55 (0.25), which rejects zero but not one. In computer science and engineering, equipment does not move, and the response of zero rejects the model. These rejections should be read with the scale of the outcome in mind: non-federal equipment spending in computer science averages 0.30 million dollars a year at an R1 university, less than a typical award, so it cannot fall by the implied amount. The offset could occur elsewhere in the budget, but these estimates do not establish where it occurs.

Table A.12: Test of Model-Implied Crowd-Out in HERD R&D Spending

	Institutional R&D funds		Non-federal equipment	
Field	Mean ($M/yr)	$\sum_\ell \hat{\lambda}_\ell$	Mean ($M/yr)	$\sum_\ell \hat{\lambda}_\ell$
Fields that use computing capital (H_0: $\sum_\ell \lambda_\ell = 1$)				
Computer science	2.65	0.73 (0.60)	0.30	−0.01 **(0.27)**
Engineering	14.15	3.09 (5.10)	1.63	−0.03 **(0.37)**
CS + engineering	17.37	3.90 (5.15)	2.16	−0.03 **(0.39)**
Physical sciences + geosciences	9.37	1.31 (1.51)	1.16	0.55 (0.25)
Life sciences	50.75	−3.37 (5.53)	2.64	0.31 (0.50)
Placebo fields (model predicts $\sum_\ell \lambda_\ell = 0$)				
Social sciences	5.43	−0.28 (0.59)	0.10	0.03 (0.07)
Mathematics	1.05	0.43 (0.29)	0.06	0.01 (0.03)
Non-S&E	11.49	1.69 (2.12)	0.30	−0.08 (0.14)

Notes: Estimates of (A7.8) with the university-specific implied response $\widehat{I}_{jt} = (\partial I^K/\partial N)_{jt} N_{jt}$ from (7.4). Outcomes are institutionally financed R&D expenditure and non-federal capitalized equipment expenditure by field from the NSF HERD Survey, 186 R1 universities, 2013–2023 (725–2,034 university-years). Regressions include university and year fixed effects and log endowment. Standard errors in parentheses are clustered by university. Bold: rejection at the 5% level of $\sum_\ell \lambda_\ell = 1$ in fields that use computing capital, or of $\sum_\ell \lambda_\ell = 0$ in placebo fields.

A8 Counterfactual Details

Marginal products. Holding field shares fixed, with $Y_{jf} = \Omega_{jf} \mathcal{K}_{jf}^{\alpha_f^K} \mathcal{L}_{jf}^{\alpha_f^L}$ and G_{jf} equal to a fixed share of G_j, the marginal products of university-level capital are

$$\frac{\partial \widehat{Y}_j}{\partial G_j} = \frac{1}{G_j} \sum_f \alpha_f^K \, s_G^{\mathcal{K}}(jf) \, Y_{jf}, \qquad \frac{\partial \widehat{Y}_j}{\partial C_j} = \frac{1}{C_j} \sum_f \alpha_f^K \, s_C^{\mathcal{K}}(jf) \, Y_{jf} \tag{A8.1}$$

In the counterfactual, field shares respond to the new stock. We therefore compute marginal products as finite differences that re-solve the allocation rule (5.11) at the new (C_j, G_j). Dollar marginal products divide by p_G and p_C.

Productivity. We back out Ω_{jf} at the baseline year t^* so that predicted output equals observed output, and hold it fixed.

Efficient allocation. With $\rho = 0$, the planner's problem is to maximize $\sum_j \mathcal{Y}_j(N_j)$ subject to the budget. Because each $\mathcal{Y}_j$ is concave in N_j, we allocate the budget in small increments, each time to the university with the highest pass-through-weighted dollar marginal product $m_j \, MPK_j^{\$}$, and within that university to the input with the higher dollar return. We stop when the budget is exhausted. At the solution, $m_j \, MPK_j^{\$}$ is equalized across funded universities up to the increment size, and is below the common value at unfunded universities.

Additional results. Table A.13 reports the field-level gains behind Figure 11 in levels and percent. Table A.14 gives the budget shares by endowment tercile, the mean pass-through, and the number of corner universities in each tier. Figure A.9 repeats Figure 12 with universities ranked on total endowment rather than endowment per faculty; the rank correlation between the two measures is 0.87. Figure A.11 repeats Figure 14 with the share of funded universities in place of the share of dollars.

Ranking by productivity. Figure A.10 ranks universities on productivity. The production function delivers a productivity fixed effect for each university and field, $\bar{\omega}_{jf}$, the mean over the sample years of the productivity residual ω_{jft} recovered from the estimated system (Section 6). Fields differ in the level of $\bar{\omega}_{jf}$, so we first subtract the field's mean across universities, which leaves each university's productivity relative to other universities in the same field. We then average these field-demeaned values over the university's fields, weighting each field by the university's predicted GPU-topic output in that field in 2021, to summarize productivity in the fields that contribute most to baseline output. The index summarizes estimated productivity; it is not a marginal product, which also depends on field elasticities, input shares, and capital stocks.

Table A.13: Predicted GPU-Topic Papers by Field (% Change from Baseline)

		No crowd-out		Model-implied crowd-out	
	Baseline	Observed	Efficient	Observed	Efficient
Computer science	4,310	+3.9	+10.4	+1.6	+5.2
Engineering	2,002	+1.2	+2.1	+0.5	+1.2
Geosciences and physics	3,868	+2.1	+3.3	+0.9	+1.9
Molecular and materials	1,704	+2.9	+4.3	+1.3	+2.4
Life sciences	2,341	+3.2	+5.9	+1.6	+3.3
Total	14,224	+2.8	+5.9	+1.2	+3.1

Notes: Field-level output re-solved through the allocation rule (5.11) at each university's counterfactual capital stock, baseline year 2021. Capital elasticities from Table 6. GPU-topic papers are those with at least one topic classified as GPU-heavy or as both GPU- and CPU-heavy, as in the production function.

Table A.14: Budget Shares and Pass-Through by Endowment Tier

	Low	Medium	High
Share of the budget			
Observed NSF portfolio	25%	48%	27%
Efficient, no crowd-out	43%	18%	39%
Efficient, model-implied crowd-out	49%	21%	30%
Pass-through in 2021			
Mean m_j	0.39	0.20	0.20
Universities in the corner	6	2	2
Gain in papers, model-implied crowd-out (observed / efficient)			
	+34 / +213	+40 / +78	+97 / +149

Notes: Share of the reallocated budget by financial-constraint tier, baseline year 2021. Pass-through m_j from (7.3): one for a university in the corner in 2021 (no CPU or GPU capacity added in 2019 or 2020), χ^K_{jt} otherwise. Tier sizes 18 / 17 / 18 universities.

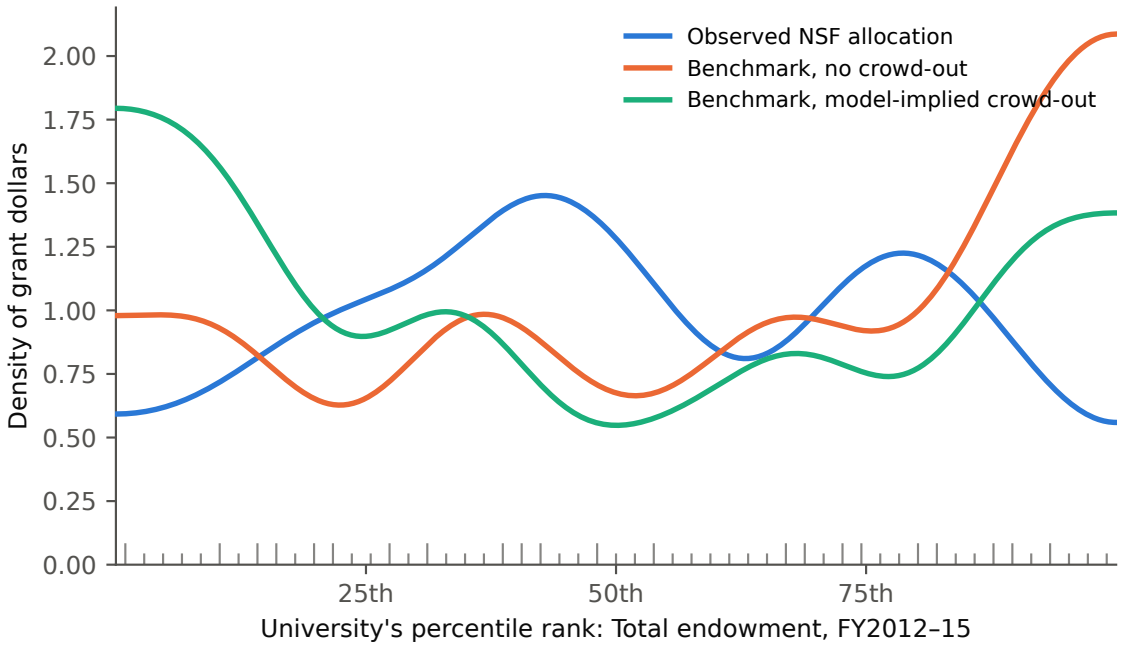


Figure A.9: Where the Grant Dollars Go, Ranked by Total Endowment. As in Figure 12, with universities ranked on total endowment, FY2012–15 mean.

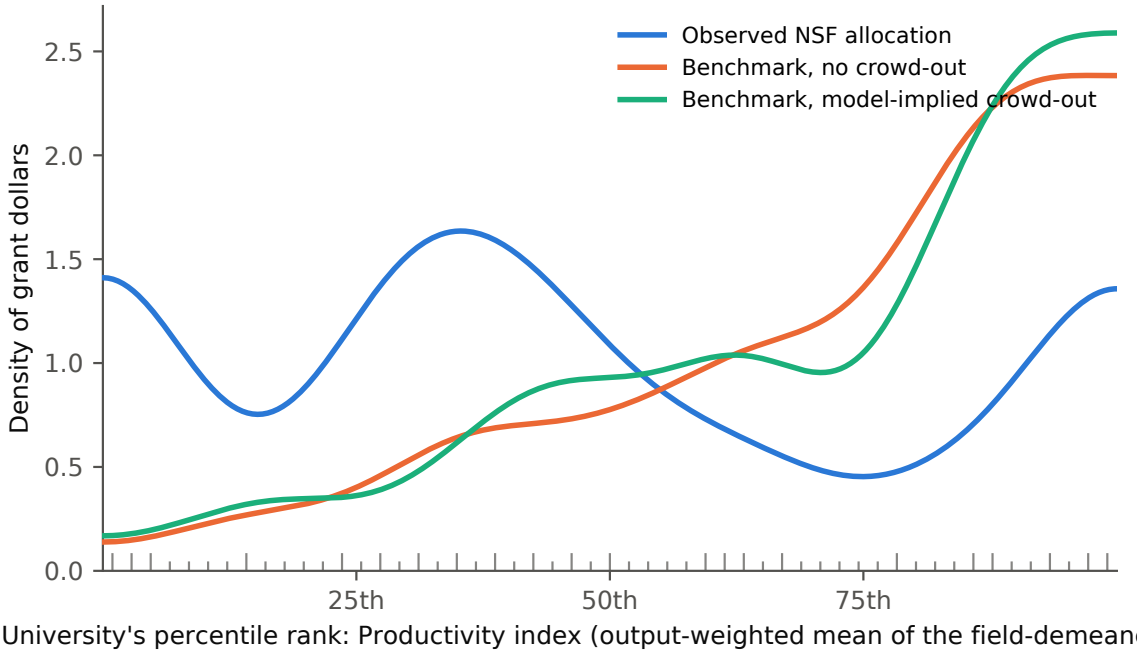


Figure A.10: Where the Grant Dollars Go, Ranked by Productivity. As in Figure 12, with universities ranked on the productivity index described above: the output-weighted mean over the university's fields of the field-demeaned productivity fixed effect $\bar{\omega}_{jf}$, with 2021 predicted GPU-topic output as weights.

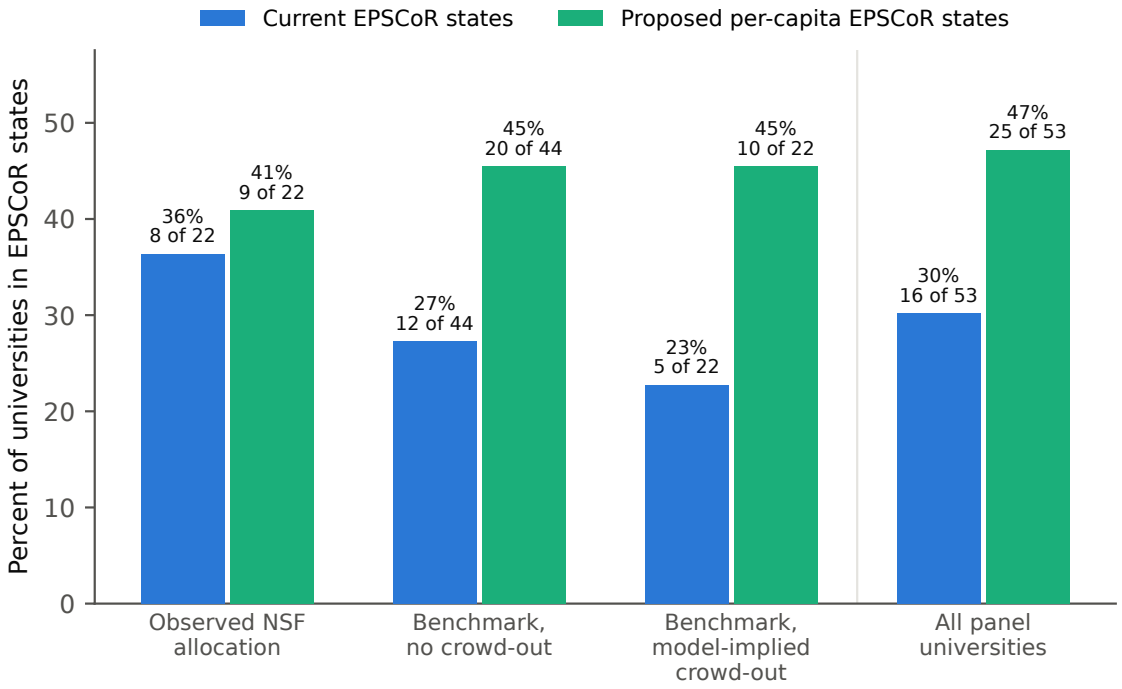


Figure A.11: Share of Funded Universities in EPSCoR States. As in Figure 14, with the share of funded universities located in EPSCoR states in place of the share of dollars; counts below each bar. The right-hand group is the share of all 53 panel universities in EPSCoR states.

Appendix References

Shvadron, D., H. Zhang, L. Fleming, and D. P. Gross (2025). Measuring the production of scientific human capital: New data, methods, and evidence on the U.S. scientific training ecosystem. NBER Working Paper 33944, National Bureau of Economic Research.

Yue, D. (2024, November). I, Google: Estimating the impact of corporate involvement on AI research. Research Paper 5033334, Georgia Tech Scheller College of Business.