Technological Breakthroughs and the Progress of Science: Evidence from Early Computers (1950-1970)

Pedro Aldighieri & Franco Malpassi

Department of Economics, Northwestern University

pedro.aldighieri@u.northwestern.edu francomalpassi@u.northwestern.edu

May 13, 2026

Technology as a Driver of Scientific Progress

- What drives scientific progress?
- Large literature on human capital (Jones [2009]; Wuchty et al. [2007]; Waldinger [2016]), institutions (Moser [2005]; Giorcelli and Moser [2020]; Moscona [2021]), and funding (Azoulay et al. [2019]; Myers [2020])
- Technology typically studied as an output of science (Jaffe [1989]; Mansfield [1991]), not as an input
    - Yet tools reshape what can be measured, computed, or inferred (Mokyr [2002]; Rosenberg [1992]; Krauss [2026])
- We study the impact of a general purpose technology — digital computers — on science itself (Bresnahan and Trajtenberg [1995])

This Paper

- Focus on the introduction of digital computers at US universities (1950–1970)
- Study how adoption impacted research direction, quality, and methods
  - Focus on the early and large mainframe computers
  - Computers housed in shared centers, access tied to institutional affiliation
- Construct the first comprehensive database of 2,200 computer installations across 184 universities
- Combine with publication data from OpenAlex and full-text analysis
- Exploit variation in timing of adoption across universities and pre-digital computational intensity across subjects

Related Literature

- **Factors Affecting Scientific Direction & Quality** (e.g., Nagaraj and Tranchero [2023]; Myers [2020]; Borjas and Doran [2012]; Boudou and Mckeon [2024])
  - Role of **technology itself** as a driver of scientific progress and direction.
  - Study introduction of a **GPT** with **large, discontinuous** jump in compute.
- **Technology and Science** (e.g., Ahmadpoor and Jones [2017]; Agrawal and Goldfarb [2008]; Gao and Wang [2023], Mokyr [1992, 2002, 2016])
  - Provide **causal evidence** on the reverse direction from technology to science.
- **Diffusion of General-Purpose Technologies** (e.g., David [1990]; Comin and Hobijn [2010]; Moser and Nicholas [2004]; Bresnahan and Trajtenberg [1995])
  - Trace the **diffusion of a GPT within science** where we can track where, when, and how computers entered the scientific record.

Preview of Results

- Computer-related papers appear immediately after campus installations
- Fields relying on manual calculation before computers adopt them more and earlier
- Computer papers are ∼20% more cited, ∼50% more likely to be top 1%, and 18–32% more novel
- Triple-difference university-level: computers shift research toward numerically intensive areas
  - Physical Sciences: +15% publications, +22% citations in exposed subfields
  - Social Sciences: +10% publications, +32% citations in exposed subfields
- Subject-level DiD: effects scale with pre-digital numerical intensity — 90th vs. 10th pctile: +31% vs. +17% pubs, +51% vs. +24% citations
- In the period, computers seem to make science relatively less empirical, shifting research towards simulations and methods
- No quantity-quality trade-off: more papers *and* better papers

Historical Background

“There will never be enough problems, enough work for more than one or two of these computers.”

– Howard Aiken, late 40s, quoted in Stern (1981)

“It would appear that we have reached the limits of what is possible to achieve with computer technology, although one should be careful with such statements, as they tend to sound pretty silly in five years.”

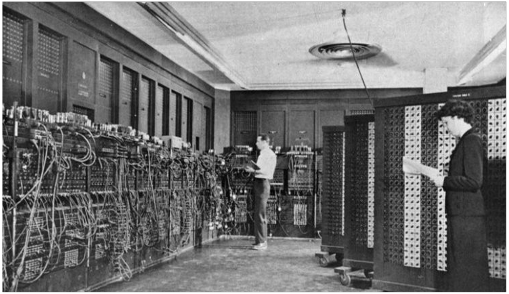
– John von Neumann (1949)

Revolution in Scientific Computing

- Pre-1945: Scientific research constrained by computational power
  - Manual computations and mechanical calculators, prone to error
  - Mechanical devices (e.g., differential analyzers) were special-purpose and limited
- Sharp transformation with digital computers (late 1940s) ▸ IAS
  - ENIAC (1946): First programmable computer, 1,000 times faster than predecessors
  - IBM 650 (1953): First widely adopted computer, first model for 27% of our sample
- ∼100,000-fold increase in computations per second between 1945 and 1970 (Nordhaus [2007])

Boom in Computer Innovation: From the ENIAC to IBM

The ENIAC: 1945



IBM 650: 1953

Unlocking Previously Infeasible Ideas

- Many applications were ideas people already had but lacked the computational power to execute
- Numerical Weather Prediction
  - Richardson [1922]: full method laid out in 1922, but infeasible by hand
  - Charney et al. [1950]: first successful numerical forecast on ENIAC (1950)
- Quantum Chemistry
  - Dirac (1929): QM provides the laws, but equations "too complicated to be soluble"
  - Boys (1950): first *ab initio* calculation on EDSAC; Roothaan (1951): matrix formulation for computers
- X-ray Crystallography
  - Fourier syntheses for 3D structures: feasible for small molecules, impossible for proteins by hand
  - Kendrew solved myoglobin on EDSAC (1958) — first 3D protein structure; Nobel Prize 1962 (Kendrew [1963])

Early Digital Computer Adoption by Universities

- High costs of computer installation led to staggered adoption
  - UNIVAC I cost $1.5M in 1951 (∼$15M in 2025 dollars), required 382 sq ft
  - Funding from manufacturer donations, NSF acquisition grants, dedicated fundraisers
- Mostly (50.5%) located in shared computer centers due to high cost and large size

  ▸ Where Installed
  - NSF conditioned funding on university-wide availability (National Research Council [1966])
- Researchers depended on universities for computer access
  - Limited remote access until the late 1960s ▸ Remote Access Example
- First universities got digital computers in 1951 (MIT, GWU)
- By 1969, all research-intensive universities had access to a computer

Data

Computer Installations Database

- First database of computer installations in US higher education, up to 1971
- Information obtained from historical surveys of universities: ▸ All Sources
  - Computers & Automation magazine Rosters of Organizations
  - University of Rochester Computer Center surveys ▸ Snapshot
  - Southern Education Board/NSF surveys by John Hamblen ▸ Snapshot
- In total, 24 survey sources and 82 survey-year pairs, totaling 18,282 computer snapshots ▸ Database Sample ▸ Timeline
- Covers 184 consolidated and verified US universities (2,200 installations)
- Covers all US research-intensive institutions (R1, R2) of the period ▸ List
- For 74% of installations, we have direct documentation of installation dates
- Computer-model mentions in published papers closely track installation dates ▸ Examples

Publication Data

- We retrieve publication and citation metadata from OpenAlex and SciSciNet
  - OpenAlex is widely used in prior work (Priem et al. [2022]; e.g. Azoulay and Greenblatt [2025]; Schmallenbach et al. [2024]; Shvadron et al. [2025])
- For each paper, we collect full metadata (e.g., citations)
- SciSciNet (Lin et al. [2023]) for science-of-science outcomes
- We add n-grams derived from full text via OA and publisher-supplied full text
- OA uses a 4-tier subject classification: domains, fields, subfields, and topics
- Covers 4 domains $\times$ 26 fields in Physical, Life, Health, and Social Sciences

▸ Publications over the years ▸ Sample

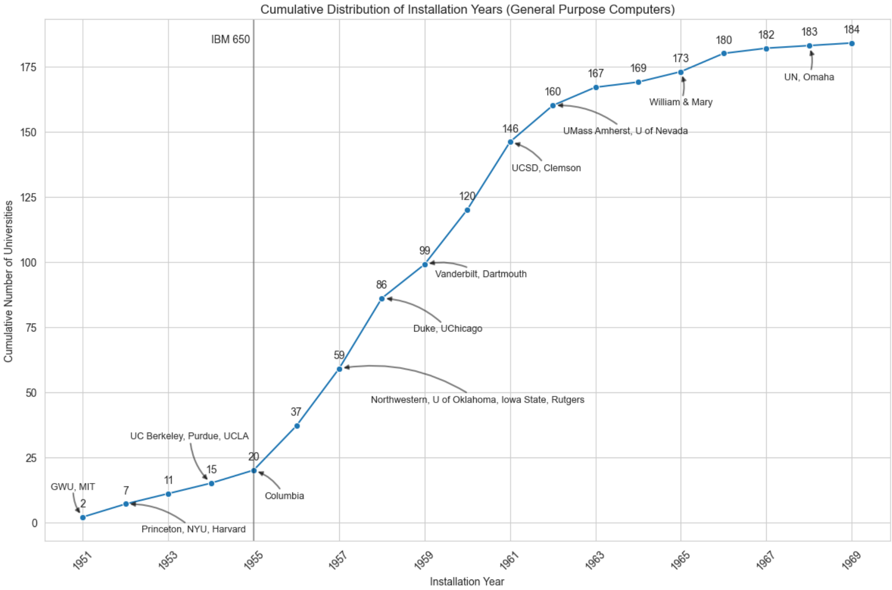
Identifying Computer Use in Articles

- N-gram search: search full-text for keywords like *digital computer*, *electronic computer*, *high-speed computing device* ▸ Keywords
  - Flags 2.27% of searchable papers (4.86% including “computer”)
  - Common for early users to explicitly state use of computers
- LLM classification from full text of 1.3M papers
  - Screen with extended keyword list, then classify with LLMs ▸ Extended List
  - Identifies 3.45% of papers as using/mentioning computers
- Both methods perform well on hand-coded validation of 200 papers ($F_1 = 0.928$ keywords, 0.967 LLMs)

Descriptive Analysis

Digital Computer Adoption by US Universities, 1950–1970

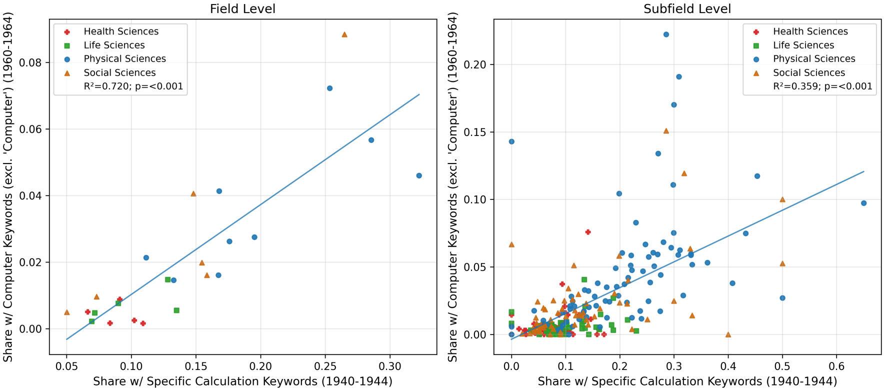
▸ Stats



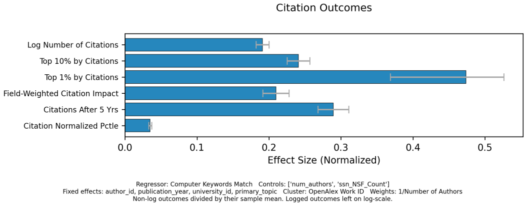
Correlates of Computer Diffusion

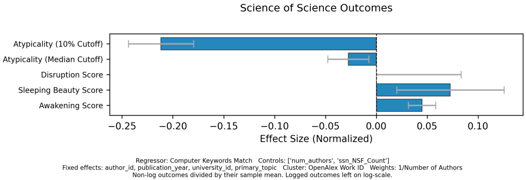
- Pre-1945 reliance on manual/mechanical calculation strongly predicts subsequent computer adoption ▸ Keywords
- $R^2 = 0.720$ at field level; $R^2 = 0.359$ at subfield level ▸ By Domain

Share w/ Specific Calculation Keywords (1940-1944) vs Share w/ Computer Keywords (excl. 'Computer') (1960-1964)



Citations and Research Impact



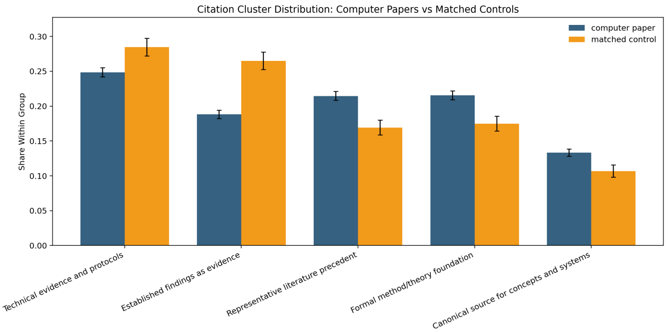


Notes: Weighted paper-author regressions with author, year, university, and topic fixed effects plus controls for number of authors and NSF grants. Standard errors clustered at the paper level (OpenAlex Work ID). Searchable-full-text sample with at least one in-sample university affiliation; SciSciNet adds the science-of-science outcomes. $N = 2.2\text{M}–3.8\text{M}$.

- Computer papers are 18-32% more novel by the Uzzi et al. [2013] metric
- Premiums are entirely driven by papers using computers as tools ▸ Usage
- Display higher breadth and novelty as measured by topic combinations ▸ Topics
- But no more likely to be cited by patents ▸ Other

▸ Citations ▸ Cit. Emb. ▸ Sci-of-Sci ▸ Premiums

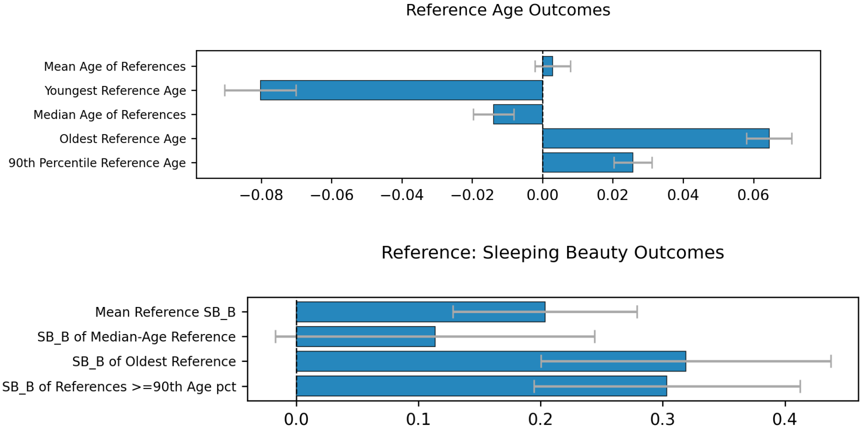
Why Are Computer Papers Cited More?



Notes: Shares of usable citation links by cluster in the 20,910-link sample (16,114 to computer papers; 4,796 to matched controls). Clusters from embeddings of short LLM descriptions of *why* the cited paper is used.

- **Composition:** computer papers cited more as canonical sources (highest-premium bucket), less as established findings (lowest-premium) ▸ Pipeline ▸ Specs ▸ Examples ▸ Regs
- **Accounting:** breadth + composition attenuate the FWCI premium by ∼10%; of that, ∼90% is composition, ∼10% excess breadth ▸ Decomposition

Reference Age and Sleeping Beauties



Notes: Bars report normalized treatment effects from pooled paper-level regressions. The regressor is an indicator for computer-related keywords in the paper's full text. Controls: number of authors and NSF grants citing the paper. Fixed effects: author, publication year, university, and primary topic. Standard errors are clustered at the paper level (OpenAlex Work ID). Observations are paper-author pairs weighted by the inverse number of authors. Sample restricted to papers with at least one in-sample university affiliation and searchable full text in OpenAlex. Non-logged outcomes are divided by their sample mean.

▶ Ref Age + SB Table ▶ Ref Age + SB Emb.

Author-Level Patterns

- Comparing authors mentioning computers ("adopters") vs. those who don't
- Computer adopters are positively selected, even after controlling for field, university, and cohort:
  - 4× publications, 4.3× citations, 64% higher H-index ▸ Table
  - Adopters are more experienced ▸ Figure and have 4.5× more top 1% papers ▸ Table
  - Advantages predate university adoption, though gap is smaller ▸ Table
  - Early vs. late adopters look similar within the adopter sample ▸ Table
- Intensive margin: ↑ computer papers → ↑ outcomes ▸ Table

Empirical Strategy

Difference-in-Differences Design

- Binary Treatment: Year of first digital computer installation (G_u)
- All institutions in sample had computers by 1969
- University-level DiD:

$$Y_{u,t} = \sum_{k \neq -1} \gamma_k \mathbf{1}\{t - G_u = k\} + \alpha_u + \lambda_t + \varepsilon_{ut}$$

  γ_k: effect of adopting computers k years ago on outcome Y

- Methods: Modern staggered-adoption estimators (Callaway and Sant'Anna [2021]; de Chaisemartin and D'Haultfoeuille [2020])
  - Standard TWFE biased with staggered adoption and heterogeneous effects ▸ Plots ▸ Bacon
- Identifying assumptions: Parallel trends, no anticipation, no spillovers
  - Limited remote access supports no spillovers
  - Equipment-specific funding channels support exogeneity

Triple Differences (DDD) Design

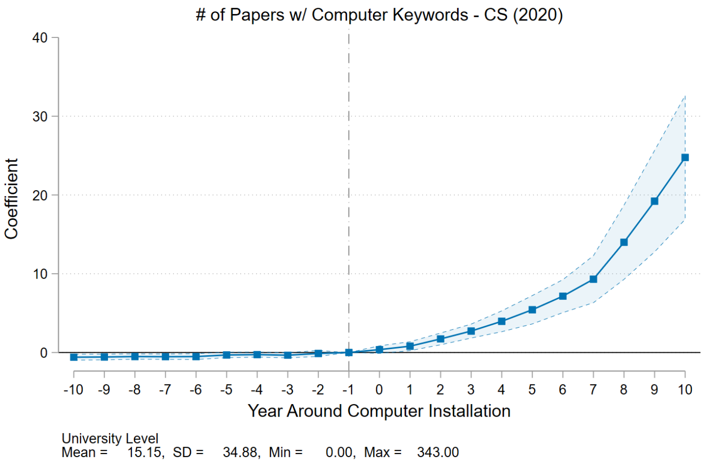
- Key idea: Different subjects have different computational demand
  - E.g., Numerical Analysis vs. Arts and Humanities
- Define exposure E_s from the 1940–1944 share of numerically intensive papers; classify subfields above vs. below the median
  - *Pre-digital* demand to avoid confounding from contemporaneous "hot field" shocks — inflows of researchers, funding, and quality that would also raise computational demand
- DDD regression specification:

$$Y_{ust} = \sum_{k \neq -1} \beta_k \mathbf{1}\{t - G_u = k\} \cdot E_s + \mu_{ut} + \eta_{st} + \delta_{us} + \varepsilon_{ust}$$

- β_k: within-university gap between exposed and unexposed subjects at event time k

Results

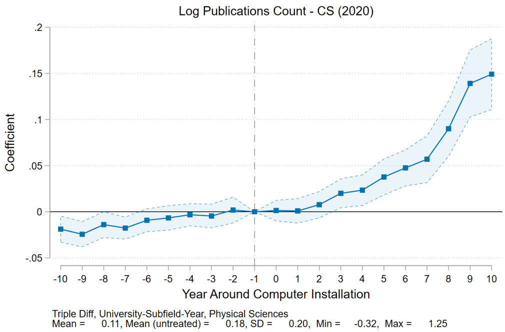
Computer Usage After Installation



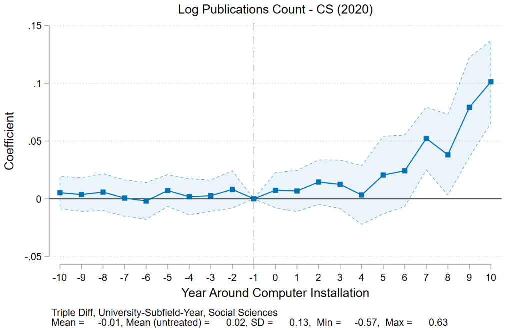
- ∼25 additional computer papers/year by year 10 (∼10% of mean annual output)

DDD: Direction of Science (Publication Counts)

Physical Sciences

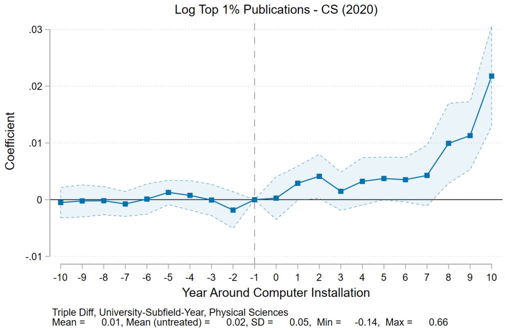


Social Sciences

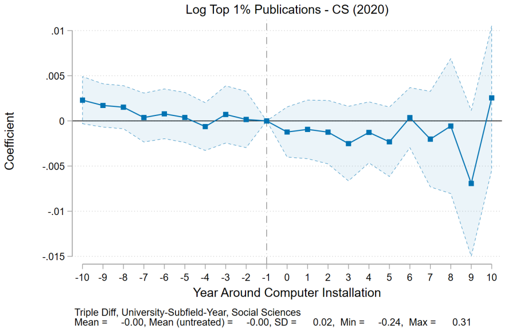


DDD: Quality of Science

Top 1% Papers: Physical Sciences



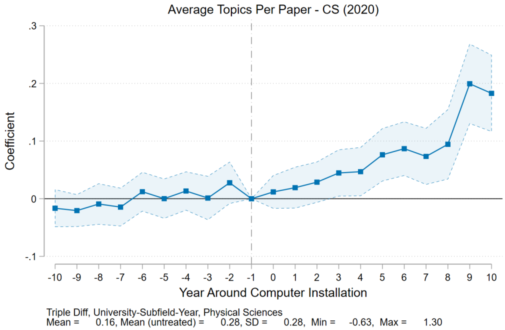
Top 1% Papers: Social Sciences



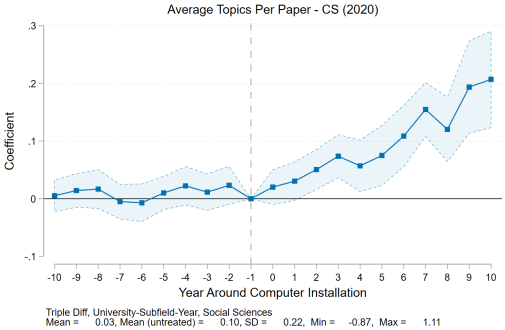
▸ Avg citations

DDD: Content of Science

Topics per Paper: Physical Sciences

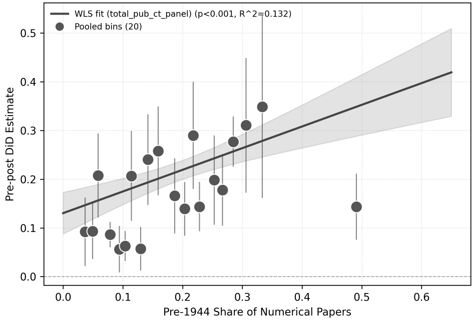


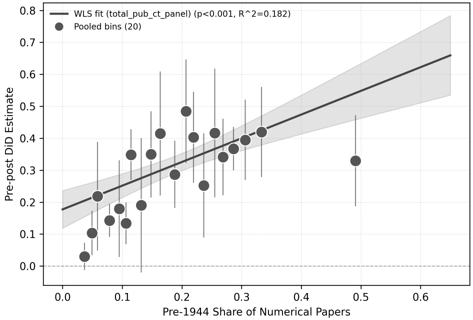
Topics per Paper: Social Sciences



▸ Topic novelty

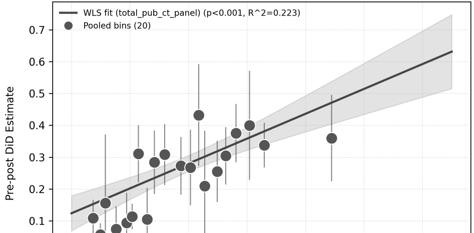
Effects by Subject Numerical Intensity

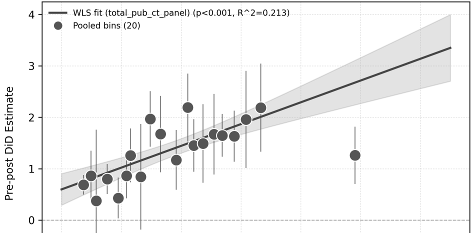




Quantity of Papers (log)

Citations per Paper (log)





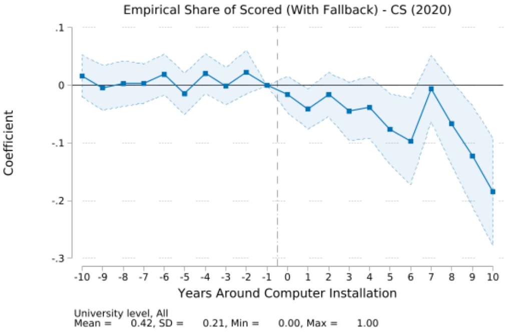
Classifying Research Methods

- We ask whether research methods shift inside universities after computer installation.
- Each paper is assigned a methodology type: empirical, theory, methods, simulation/computation, or other.
- We first build pseudo ground truth with Gemini 3.0 Flash full-text reads on 177,243 papers.
- We then train a ML gradient-boosted model so we can score papers even when local full text is unavailable.
- Main holdout benchmark: weighted F1 = 0.825 on a 10k full-text test set.

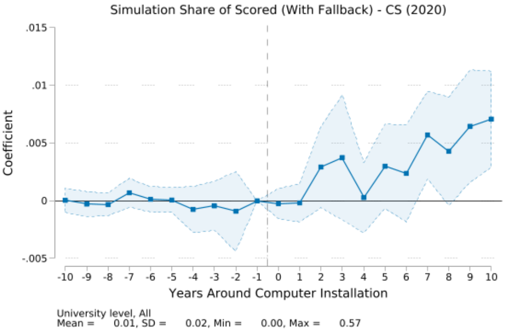
▸ Classifier details ▸ Paper types

Changes in Research Methodologies

Empirical Share of Scored Papers



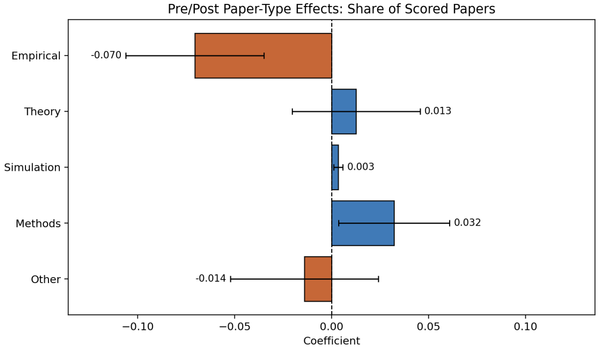
Simulation Share of Scored Papers



Notes: Callaway-Sant'Anna event-study estimates using university-year panel outcomes. Shares are measured among scored papers only ('ptype_shr_wf'), so they track reallocation within the classified subset rather than changes in scoring coverage.

▸ Theory + Methods

Within Scored Papers: Collapsed Pre/Post

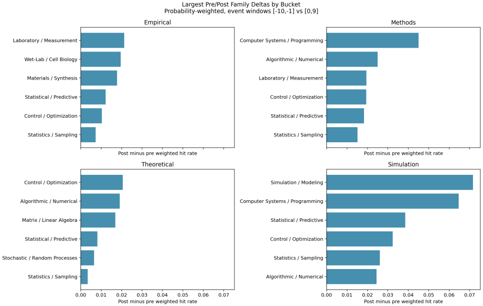


Notes: Bars show collapsed Callaway-Sant'Anna pre/post ATT estimates for shares among scored papers. This is best read as a composition shift: empirical counts rise in raw levels, but theory, methods, and simulation shares rise faster within the scored subset.

▸ Benchmark ▸ Class shares

- Raw empirical counts still increase after installation, so not an absolute collapse of empirical work.
- These bars come from separate CS share regressions, so they are not an exact adding-up decomposition.
- Paper-level regressions confirm computer papers are less likely to be empirical, even relative to theory ones. ▸ Paper-level regs

What Is Rising Inside Theory and Methods?



Notes: Transparent dictionary-hit frequencies, not another classifier. Sample: 270,134 predicted paper $\times$ role rows in 1951–1969; pre '[-10,-1]', post '[0,9]'. Lexicons were grounded in 120 manual full-text packets and 3,578 usable snippets.

▸ Families ▸ Title magnitudes ▸ Examples ▸ Keywords DiDs

Conclusion

Conclusion

- Computer usage appears immediately in the scientific record after installation
- Adoption follows pre-digital computational intensity: fields relying on manual calculation adopt first
- Computer papers are more cited, more novel, more broadly topical
- DDD: clear within-university reallocation toward compute-amenable subfields
  - Publication counts, quality, breadth, and novelty all increase
- Subject-level DiD: smooth gradient in pre-digital numerical intensity — 90th vs. 10th pctile: +31% vs. +17% pubs, +51% vs. +24% citations per paper
- Computers shift the methodology mix: simulation, methods, and computational theory rise; empirical share falls (raw counts still grow)
- Unlike standard GPT narratives: no delayed productivity gains; effects materialize within years
- No quantity-quality trade-off: more papers and better papers simultaneously

Thank you!

Pedro Aldighieri & Franco Malpassi

Department of Economics, Northwestern University

pedro.aldighieri@u.northwestern.edu
francomalpassi@u.northwestern.edu

Agrawal, A. and Goldfarb, A. (2008). Restructuring research: Communication costs and the democratization of university innovation. *American Economic Review*, 98(4):1578–1590.

Ahmadpoor, M. and Jones, B. F. (2017). The dual frontier: Patented inventions and prior scientific advance. *Science*, 357(6351):583–587.

Azoulay, P., Graff Zivin, J. S., Li, D., and Sampat, B. N. (2019). Public r&d investments and private-sector patenting: evidence from nih funding rules. *The Review of economic studies*, 86(1):117–152.

Azoulay, P. and Greenblatt, W. H. (2025). Does peer review penalize scientific risk taking? evidence from nih grant renewals. NBER Working Paper 33495, National Bureau of Economic Research.

Borjas, G. J. and Doran, K. B. (2012). The collapse of the soviet union and the productivity of american mathematicians. *The Quarterly Journal of Economics*, 127(3):1143–1203.

Boudou, J. and Mckeon, J. (2024). Innovation under resource constraints: Supercomputing in scientific research.

Bresnahan, T. F. and Trajtenberg, M. (1995). General purpose technologies 'engines of growth'? *Journal of Econometrics*, 65(1):83–108.

Callaway, B. and Sant'Anna, P. H. (2021). Difference-in-differences with multiple time periods. *Journal of econometrics*, 225(2):200–230.

Charney, J. G., Fjörtoft, R., and Neumann, J. v. (1950). Numerical integration of the barotropic vorticity equation. *Tellus*, 2(4):237–254.

Comin, D. and Hobijn, B. (2010). An exploration of technology diffusion. *American economic review*, 100(5):2031–2059.

David, P. A. (1990). The dynamo and the computer: An historical perspective on the modern productivity paradox. *American Economic Review*, 80:355–361.

de Chaisemartin, C. and D'Haultfoeuille, X. (2020). Two-way fixed effects estimators with heterogeneous treatment effects. *American Economic Review*, 110:2964–2996.

Gao, J. and Wang, D. (2023). Quantifying the benefit of artificial intelligence for scientific research. *arXiv preprint arXiv:2304.10578*.

Giorcelli, M. and Moser, P. (2020). Copyrights and creativity: Evidence from italian opera in the napoleonic age. *Journal of Political Economy*, 128(11):4163–4210.

Jaffe, A. B. (1989). Real effects of academic research. *The American economic review*, pages 957–970.

Jones, B. F. (2009). The burden of knowledge and the "death of the renaissance man": Is innovation getting harder? *The Review of Economic Studies*, 76(1):283–317.

Ke, Q., Ferrara, E., Radicchi, F., and Flammini, A. (2015). Defining and identifying sleeping beauties in science. *Proceedings of the National Academy of Sciences*, 112(24):7426–7431.

Kendrew, J. C. (1963). Myoglobin and the structure of proteins: Crystallographic analysis and data-processing techniques reveal the molecular architecture. *Science*, 139(3561):1259–1266.

References ii

Krauss, A. (2026). *The Engine of Scientific Discovery: How New Methods and Tools Spark Major Breakthroughs*. Oxford University Press, New York.

Lin, Z., Yin, Y., Liu, L., and Wang, D. (2023). Sciscinet: A large-scale open data lake for the science of science research. *Scientific Data*, 10:315.

Mansfield, E. (1991). Academic research and industrial innovation. *Research policy*, 20(1):1–12.

Mokyr, J. (1992). *The lever of riches: Technological creativity and economic progress*. Oxford University Press.

Mokyr, J. (2002). *The Gifts of Athena: Historical Origins of the Knowledge Economy*. Princeton University Press, Princeton, NJ.

Mokyr, J. (2016). *A Culture of Growth: The Origins of the Modern Economy*. Princeton University Press, Princeton, NJ.

Moscona, J. (2021). Flowers of invention: Patent protection and productivity growth in US agriculture. Working Paper, MIT.

Moser, P. (2005). How do patent laws influence innovation? evidence from nineteenth-century world's fairs. *American Economic Review*, 95(4):1214–1236.

Moser, P. and Nicholas, T. (2004). Was electricity a general purpose technology? evidence from historical patent citations. *American Economic Review*, 94(2):388–394.

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

Nagaraj, A. and Tranchero, M. (2023). How does data access shape science? evidence from the impact of u.s. census's research data centers on economics research.

National Research Council (1966). Digital computer needs in universities and colleges. Technical Report 1233, National Academy of Sciences – National Research Council, Washington, D.C.

Nordhaus, W. D. (2007). Two centuries of productivity growth in computing. *Journal of Economic History*, 67:128–159.

Priem, J., Piwowar, H., and Orr, R. (2022). Openalex: A fully-open index of scholarly works, authors, venues, institutions, and concepts. *arXiv preprint arXiv:2205.01833*.

Richardson, L. F. (1922). *Weather Prediction by Numerical Process*. Cambridge University Press, Cambridge.

Rosenberg, N. (1992). Scientific instrumentation and university research. *Research policy*, 21(4):381–390.

Schmallenbach, L., Bärnighausen, T. W., and Lerchenmueller, M. J. (2024). The global geography of artificial intelligence in life science research. *Nature Communications*, 15:7527.

Shvadron, D., Zhang, H., Fleming, L., and Gross, D. P. (2025). Funding the u.s. scientific training ecosystem: New data, methods, and evidence. NBER Working Paper 33944, National Bureau of Economic Research.

Tamkin, A., McCain, M., Handa, K., Durmus, E., Lovitt, L., Rathi, A., Huang, S., Mountfield, A., Hong, J., Ritchie, S., Stern, M., Clarke, B., Goldberg, L., Sumers, T. R., Mueller, J., McEachen, W., Mitchell, W., Carter, S., Clark, J., Kaplan, J., and Ganguli, D. (2024). Clio: Privacy-preserving insights into real-world ai use.

Uzzi, B., Mukherjee, S., Stringer, M., and Jones, B. (2013). Atypical combinations and scientific impact. *Science*, 342:468–472.

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

Wuchty, S., Jones, B. F., and Uzzi, B. (2007). The increasing dominance of teams in production of knowledge. *Science*, 316(5827):1036–1039.

Remote Access at Oregon State

This need is partially alleviated in a somewhat unsatisfactory manner by computational facilities provided through the IBM 7094 at Western Data Processing Center (WDPC) on the UCLA campus. ... While this facility theoretically provides the capability for solution of large problems, ... the time delay and cost in sending and receiving data, limited transmission time (only up to 1-1/2 hours per day) and lack of direct access to the computer make this arrangement unsatisfactory. ... Several faculty members have spent considerable time and money traveling to WDPC to debug programs.

Computer Facility Grant Proposal of Oregon State University to NSF, June 1965

▸ Back

Testing Princeton's IAS Computer

"During the testing of the arithmetic unit [of the MANIAC] in 1948, the team tested it against von Neumann himself. As they entered in more and more complicated terms, von Neumann finally erred, proving to their collective satisfaction 'the power of matter over mind.'"

– Bigelow (1980)

▸ Back

Petition for IBM 650 at NU

Back

Economics:

1. The acceleration principle and other determinants of investment: an econometric analysis of capital expenditures, capital expenditure plans, sales expectations, sales changes, profits and other related data collected in the McGraw-Hill capital expenditure surveys.

2. The trade cycle model with some empirically derived coefficients for high order difference equations.

3. Empirical demand functions, from cross sectional and time series price and income data.

Professors: R. L. Basmann, R. Eisner

Proposed Uses of Computer by Economics Department at Northwestern, 1957
Source: Northwestern University Archives

Database Sample Snapshot

department	computer	manufactu	year_insta	month_ins	year_deco	month_de	average_h	lowest_sn	lowest_sn	highest_sr	highest_sr	source
Vogelback Computing Center	CDC 3400/8090	CDC	1964	january			273	1965	january	1966	september	hamblen (1966, 1968)
Vogelback Computing Center	EAI PACE Analog	EAI						1962	september	1964	february	edp (1962); dpy (1964)
Vogelback Computing Center	IBM 1401	IBM	1962					1961	july	1965	january	nrc (1963, 1965); dpy (
Vogelback Computing Center	IBM 1401	IBM						1965	january	1965	january	nrc (1965)
Medical School	IBM 1620/1710	IBM						1964		1965	january	dpy (1964); nrc (1965)
Administrative Data Processing	IBM 360/30	IBM	1966					1968	may	1968	may	hamblen (1968)
Vogelback Computing Center	IBM 650	IBM	1958					1957	june	1962	may	amsn (1960); datamat
Vogelback Computing Center	IBM 709	IBM	1961	july	1964	august	273	1960	july	1966	september	hamblen (1966); nrc (1
Vogelback Computing Center	LGP-30	Librascope						1963	january	1965	january	nrc (1963, 1965); dpy (

Figure 1: Installations of Computers at Northwestern University. Some columns removed for readability.

Back

Hamblen Survey, 1966

Back

1964-65 COMPUTER SURVEY--SOUTHERN REGIONAL EDUCATION BOARD COMPUTER SCIENCES PROJECT
CONTRACT NSF C465

ITEM I-A-4,5,6 COMPUTERS INSTALLED AND ON ORDER FOR RESEARCH AND INSTRUCTIONAL USES

CTL 1 TYPE 1 LEVEL 4

INSTITUTION		COMPUTER SYST.	YEAR INST	TO BE REPLACED	LEASE	PURCH	BOTH	1964-65 AVG.USE HRS/MO
OKLA STATE UNIVERSITY STILLWATER OKLAHOMA	74074	IBM 1410	64	X	*			288
		IBM 1620	63			*		450
		IBM 7040	65					
UNIVERSITY OF OKLAHOMA NORMAN OKLAHOMA	73069	IBM 1410	62	X	*			492
		IBM 1620	62		*			300
		IBM 360/40	67					
		IBM 360/65	68					
OREGON STATE UNIVERSITY CORVALLIS OREGON	97331	ALW III-E	57			*		200
		IBM 1620	61				*	200
		IBM 1410	64	X	*			100
		CDC 3300	66					
		PDP 8	00					
UNIVERSITY OF OREGON EUGENE, OREGON	97403	IBM 1620	60			*		
		IBM 360/50	66					
		PDP 7	66					
PENNSYLVANIA STATE UNIVERSITY UNIVERSITY PARK PA	16802	IBM 7074	61	X		*		720
		IBM 7074	62	X	*			240
		IBM 1401	62	X		*		650
		IBM 1410	64	X	*			650
		IBM 1620	63	X	*			80
		IBM 1620	62		*			150
		IBM 360/67	68					
		IBM 360/50	66					

Keenan Survey, 1963

Back

1. NAME OF UNIVERSITY University of Illinois
2. MAILING ADDRESS Urbana, Illinois (service branch)

PART I – GENERAL INFORMATION

1. DIRECTOR OR PERSON IN CHARGE J.N. Snyder
   (A) DEGREE AND ACADEMIC FIELD Ph.D.-Physics
   (B) ACADEMIC POSITION Res. Prof.
   (C) REPORTS TO Head of Digital Computer Lab.
2. DATE CENTER OR LABORATORY ESTABLISHED 1948
   (A) TAX SUPPORTED yes
   (B) APPROX. FLOOR AREA - MACHINE ROOM 1000 SQ. FT. CLASSROOM SQ. FT.
   OFFICES 1500 SQ. FT. USER WORKROOM 1000 SQ. FT.
   LIBRARY SQ. FT. (OTHER) SQ. FT.
   STORAGE SQ. FT.
   (C) COMPUTERS IBM 7094 HRS./MONTH USED 500
   IBM 1401 (2) 500
   (Illiac II, IBM 1401) (100 100)
   (D) PERCENT OF EQUIP. OWNED 100 LEASED
   (E) MAJOR EQUIPMENT ON ORDER none EXPECTED DELIVERY

PART II – PERSONNEL INFORMATION

1. NUMBER OF STAFF MEMBERS DESIRABLE
   (A) ANALYSTS 4 (C) OPERATORS 18
   (B) PROGRAMMERS 10 (D) CLERICAL 6
2. REGULAR POSITIONS

REGULAR POSITIONS	DEGREE-SUBJ. AREA	COMPUT. EXPER.	HRS/WK.	MO/YR.	JOINT APPT.
Dir.	Ph.D.-Physics	12 yrs.	40	9	Physics
	Ph.D.-Physics	6 yrs.	40	9	Physics

Survey Sources

1. Computers & Automation Rosters
2. Inventory of Computers in U.S. Higher Education (Hamblen)
3. Rochester annual computing-center survey (Keenan)
4. Digital Needs in Universities and Colleges (Roesser)
5. AMS Notices survey of high-speed computers
6. Datamation university survey
7. Educational Programs and Facilities in Nuclear Science and Engineering
8. Data Processing Yearbooks / Computer Yearbook
9. Business Electronics Reference Guide
10. Hearings on H.R. 4845
11. Hydrologic Computer Programs
12. Digital Computer Newsletter
13. ONR Survey of Automatic Digital Computers
14. BRL Survey of Domestic Electronic Digital Computing Systems (Weik)
15. Florida State University administrative survey
16. IBM 650 installation data (IBM Archives)
17. Mathematics in Education
18. Business Automation university survey
19. Data Processing for Management
20. Research Centers Directory
21. Survey of Numerical Weather Prediction
22. Use of Electronic Data Processing Equipment hearings
23. AEC Authorizing Legislation hearings
24. Datamation installation news

Back

Survey Coverage Timeline

▸ Back

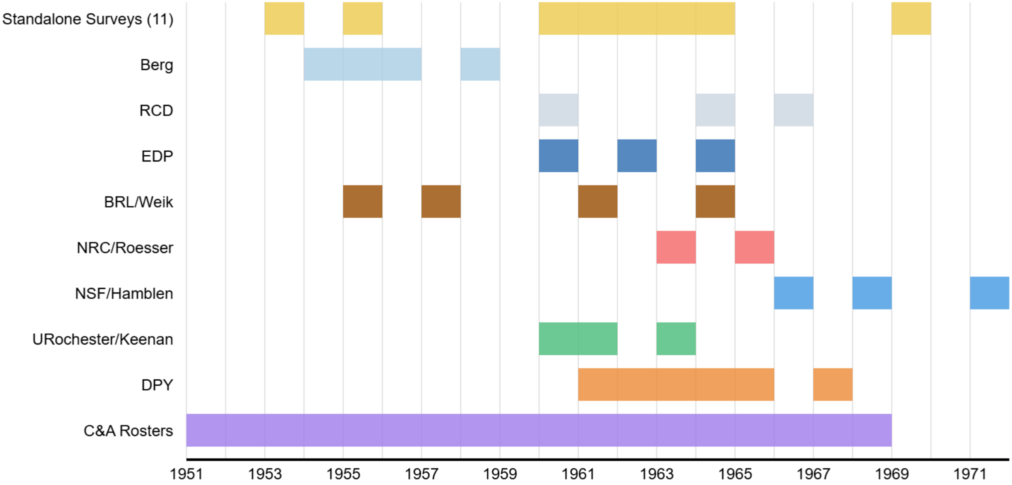


Figure 2: Yearly coverage of survey sources in the database.

Universities In Sample

▸ Back

Abilene Christian College
American University
Arizona State University
Auburn University
Boston College
Boston University
Brandeis University
Brigham Young University
Brown University
Caltech
Carnegie Mellon University
Case Western Reserve
Columbia University
Cornell University
Dartmouth College
Duke University
Emory University
Florida State University
George Washington University
Georgetown University
Georgia Institute Of Technology
Harvard University
Howard University
Illinois Institute Of Technology
Indiana University
Iowa State University
Johns Hopkins University
Kansas State University
Lehigh University
Louisiana State University
MIT
Michigan State University
Mississippi State University
Montana State University
New York University
North Carolina State University
Northwestern University
Ohio State University
Oklahoma State University
Oregon State University
Penn State University
Princeton University
Purdue University
Rensselaer Polytechnic Institute
Rice University
Rutgers University
Stanford University
SUNY Buffalo
Syracuse University
Texas A&M University
Tufts University
Tulane University
UC Berkeley
UC Los Angeles
UC San Diego
University of Chicago
University of Colorado Boulder
University of Florida
University of Illinois UC
University of Iowa
University of Kansas
University of Maryland
University of Michigan
University of Minnesota
University of Missouri
University of Nebraska
University of North Carolina
University of Notre Dame
University of Oklahoma
University of Oregon
University of Pennsylvania
University of Pittsburgh
University of Rochester
University of Southern California
University of Tennessee
University of Texas Austin
University of Utah
University of Virginia
University of Washington
University of Wisconsin Madison
Vanderbilt University
Virginia Tech
Washington University StL
Wayne State University
West Virginia University
Yale University
+ 124 more

Computer Installations Descriptive Statistics

Back

- IBM dominated with 58% of installations, DEC followed at 9%
- The IBM 650 was first computer for 49 universities (27%)
- Pre-1955: 12/16 universities (75%) built own computers
- 27 universities built 44 computers internally, mostly IAS-based
- Analog computers: 171 installations (8%)
- Computer centers housed 50.5% of all installations Locations

	# Computers per university
mean	11.6
std	11.2
min	1
25%	5
50%	8
75%	15
max	83

Where Were Computers Installed?

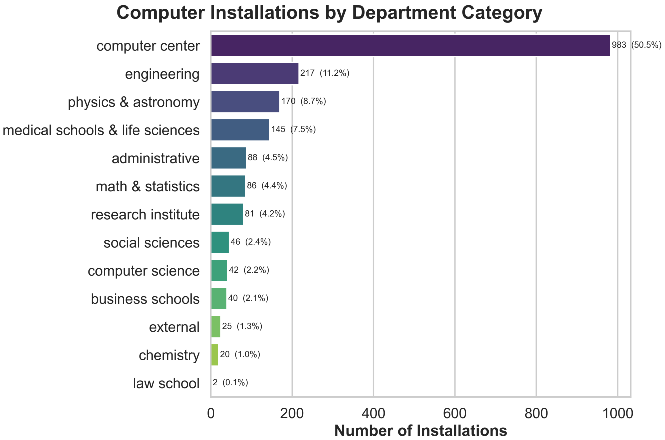


Figure 3: Shared computer centers account for 50.5% of installations with known location.

▸ Adoption ▸ D Stats

Computer Keywords

▸ Back

Core keyword list

- (computer), electronic computer
- digital computer, automatic computer
- high-speed computer, mainframe computer
- high-speed computing device
- electronic brain
- data processing equipment
- computer program
- computer algorithm
- programming language
- FORTRAN, COBOL

Extended LLM screen

- calculator, computing, computational
- punch card, punched card
- Monte Carlo, simulation
- ALGOL, algorithm, program, programming
- electronic machine
- analog computer
- electromechanical computer
- automatic equipment

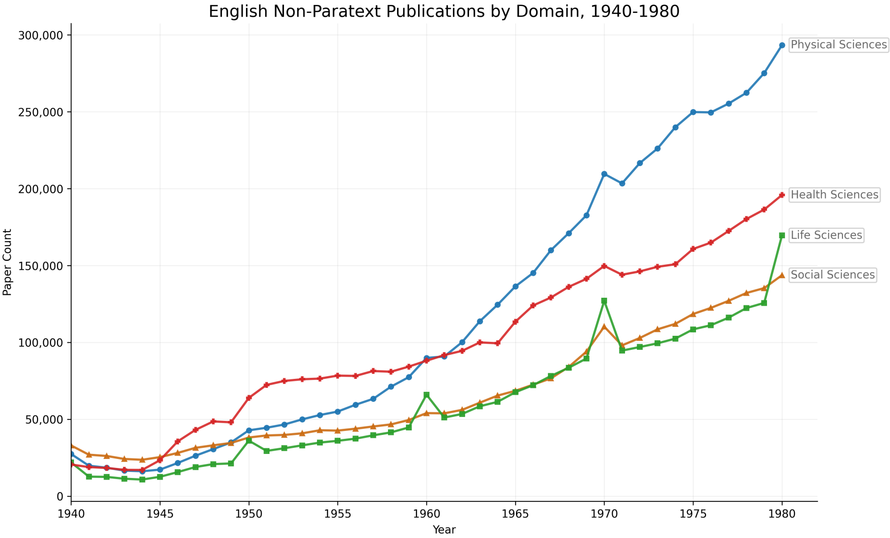
Extended LLM screen (cont.)

- data processing, EDP, ADP
- IBM, UNIVAC, Burroughs

Rates in draft

- Keywords: 2.27% of searchable papers
- 4.86% including "computer"
- LLMs: 3.45% of papers

Publications by Field



Back

Publication Sample Selection

Back

- Restrict all samples to English-language papers published after 1940
- Paper-level descriptives: all OpenAlex papers, with searchable n-grams for roughly 40% of works
  - 8.9M papers before 1970; 16.7M before 1980
- Author- and university-level analyses: papers with at least one affiliation in our university sample (1940–1970)
  - About 650,000 papers; 73% with searchable full text

Calculation-Intensity Keywords

Back

Manual / mechanical calculation

- manual calculation, computed by hand
- hand computation, longhand calculation
- checked by hand
- punched card, Hollerith, keypunch
- desk calculator, mechanical calculator
- adding machine, comptometer
- Friden, Marchant, Brunsviga, Burroughs
- tabulating department, machine accounting

Linear algebra / numerical methods

- matrix inversion, matrix multiplication
- Gaussian elimination, Gauss-Jordan
- normal equations, eigenvalue, eigenvector
- Runge-Kutta, Newton-Raphson
- finite difference, difference equation
- differential equation, simplex method

Analog instruments

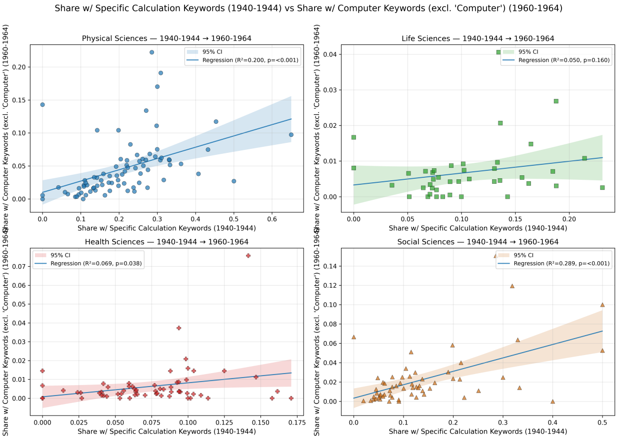
- differential analyzer, network analyzer
- harmonic analyzer, slide rule
- nomogram, integrator
- analog simulator, model board
- patch board, A-C network analyzer

Statistics / new computational uses

- least squares, maximum likelihood
- log-likelihood, regression analysis
- ANOVA, principal component analysis
- probit, logit, sample size, data analysis
- numerical simulation, stochastic simulation
- numerical experiment, random number table
- pseudo-random, Monte Carlo

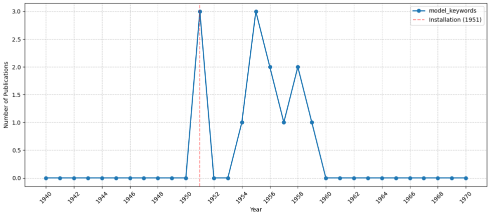
Computer Diffusion by Domain

Back

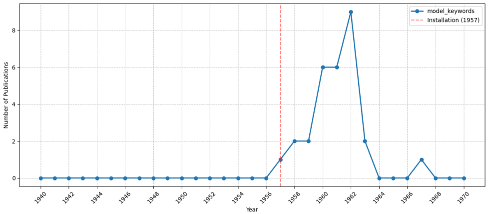


Computer Model Mentions Across Universities

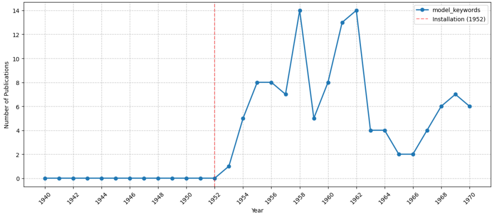
Back



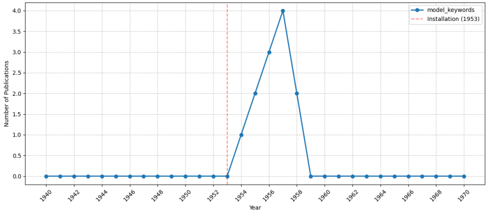
Massachusetts Institute of Technology



Northwestern University



University of Illinois, Urbana-Champaign



University of Michigan, Ann Arbor

Research Impact by Usage Type

▸ Back

	Log Cites	Top 1%	# Cncpts	Cncpt Max Nov	Atp-Z (10%)
Computer as a Tool	0.226***	0.010***	0.140***	0.340***	-5.321**
	(0.014)	(0.002)	(0.024)	(0.076)	(2.122)
Computer as an Object of Study	-0.032	-0.004	-0.132	-0.782*	-7.187
	(0.080)	(0.008)	(0.122)	(0.466)	(9.215)
Computer (Hardware/Software context)	-0.133**	-0.010*	-0.057	-0.107	17.834
	(0.055)	(0.006)	(0.087)	(0.313)	(14.974)
Mention of Computer (other)	-0.071***	-0.000	-0.221***	-0.497***	-2.970
	(0.025)	(0.003)	(0.040)	(0.148)	(6.913)
Number of Authors	0.092***	0.003***	0.059***	0.140***	-0.947***
	(0.004)	(0.000)	(0.004)	(0.013)	(0.260)
NSF Grants (paper)	0.040	0.043	0.206	-0.141	-12.219
	(0.184)	(0.042)	(0.357)	(1.015)	(8.918)
Observations	1,144,696	1,144,645	1,144,696	1,144,696	608,277
R^2	0.714	0.406	0.614	0.577	0.583
Mean of Dep Var	1.534	0.020	2.853	10.264	37.643
Author/Year/Univ FE	Yes	Yes	Yes	Yes	Yes
Subject FE	Topic	Topic	Topic	Topic	Topic

Notes: Paper-author observations, weighted by the inverse number of authors. Categories were generated by passing paper full text to an LLM. Sample restricted to papers with at least one in-sample university affiliation and searchable full text in OpenAlex. Standard errors clustered at the paper level (OpenAlex Work ID). Citation and novelty premia are concentrated in papers using computers as tools.

Citation Outcomes ▸ Back

	Log cites	Top 10%	Top 1%	FWCI	C5	Cite pct.
Computer-Keyword Flag	0.191***	0.046***	0.009***	0.409***	1.122***	0.035***
	(0.004)	(0.001)	(0.001)	(0.018)	(0.042)	(0.001)
Number of Authors	0.114***	0.021***	0.002***	0.343***	0.834***	0.021***
	(0.002)	(0.000)	(0.000)	(0.007)	(0.020)	(0.000)
NSF Grants (paper)	0.294***	0.091***	0.007	0.583***	1.001***	0.049***
	(0.026)	(0.012)	(0.005)	(0.100)	(0.213)	(0.004)
Observations	3,903,691	3,903,526	3,903,526	3,902,497	3,903,691	3,903,526
R^2	0.610	0.428	0.291	0.385	0.435	0.612
Mean of Dep Var	1.827	0.191	0.019	1.950	3.880	0.568
Author/Year/Univ FE	Yes	Yes	Yes	Yes	Yes	Yes
Subject FE	Topic	Topic	Topic	Topic	Topic	Topic

Citation Outcomes (Emb.)

▸ Back

	Log cites	Top 10%	Top 1%	FWCI	C5	Cite pct.
Computer-Keyword Flag	0.196***	0.034***	0.005***	0.337***	0.530***	0.031***
	(0.005)	(0.001)	(0.001)	(0.019)	(0.042)	(0.001)
Observations	1,882,763	1,882,654	1,882,654	1,882,004	1,882,763	1,882,654
R^2	0.002	0.000	0.000	0.000	0.000	0.001
Mean of Dep Var	-0.016	0.046	0.001	0.171	0.644	0.003

Notes: Second-stage OLS regresses the residualized outcome on the residualized computer-keyword indicator. In the first-stage DML residualization, both treatment and outcome are partialled out on author, publication year, university, and primary-topic fixed effects; controls for number of authors and NSF grants; and concatenated abstract/title embeddings plus SPECTER embeddings. Residuals are estimated with 3-fold cross-fitting and SGD. Second-stage standard errors are clustered at the paper level (OpenAlex Work ID) and paper-author observations are weighted by the inverse number of authors. Sample restricted to searchable-full-text observations from 1947–1975.

Science of Science Outcomes

▸ Back

	Atp-Z (10%)	Atp-Z (Med)	Disrupt	SB	Awak
Computer-Keyword Flag	-10.175***	-2.895***	0.000	1.432***	0.308***
	(0.783)	(1.082)	(0.000)	(0.531)	(0.047)
# Authors	-1.134***	-1.176***	0.000***	0.119	0.065***
	(0.141)	(0.193)	(0.000)	(0.120)	(0.010)
NSF grants (paper)	-2.925	-2.278	-0.006***	-1.146	-0.140
	(2.480)	(3.385)	(0.001)	(1.282)	(0.339)
Observations	2,434,402	2,428,263	3,903,691	3,353,525	3,353,525
R^2	0.479	0.503	0.377	0.282	0.404
Mean of Dep Var	48.063	105.177	0.012	19.686	6.856
Author/Year/Univ FE	Yes	Yes	Yes	Yes	Yes
Subject FE	Topic	Topic	Topic	Topic	Topic

Topic and Concept Outcomes

▸ Back

	Topics			Concepts		
	# Tpks	Tpk HHI	Pair Avg Nov	# Cncpts	Cncpt HHI	Pair Avg Nov
Computer-Keyword Flag	0.022***	-0.004***	0.064***	0.086***	-0.007***	0.203***
	(0.002)	(0.001)	(0.014)	(0.008)	(0.000)	(0.016)
Number of Authors	0.023***	-0.005***	0.097***	0.072***	-0.002***	0.060***
	(0.001)	(0.000)	(0.004)	(0.002)	(0.000)	(0.004)
NSF Grants (paper)	0.108***	-0.028***	0.429***	0.565***	-0.016***	0.448***
	(0.015)	(0.004)	(0.081)	(0.062)	(0.002)	(0.087)
Observations	3,903,691	3,903,691	3,903,691	3,903,691	3,903,691	3,903,691
R^2	0.593	0.519	0.506	0.510	0.507	0.462
Mean of Dep Var	2.095	0.491	7.392	3.485	0.197	9.168
Author/Year/Univ FE	Yes	Yes	Yes	Yes	Yes	Yes
Subject FE	Topic	Topic	Topic	Topic	Topic	Topic

Other Outcomes ▸ Back

	# Patent Cites	# Countries	# Institutions	# Refs
Computer-Keyword Flag	0.244	0.003***	0.013***	1.967***
	(0.197)	(0.001)	(0.002)	(0.042)
# Authors	0.013	0.039***	0.140***	0.080***
	(0.012)	(0.001)	(0.002)	(0.010)
NSF grants (paper)	-0.084	-0.006	0.009	4.132***
	(0.154)	(0.004)	(0.013)	(0.297)
Observations	3,903,691	3,903,691	3,903,691	3,903,691
R^2	0.289	0.490	0.537	0.454
Mean of Dep Var	0.259	0.605	0.643	5.492
Author/Year/Univ FE	Yes	Yes	Yes	Yes
Subject FE	Topic	Topic	Topic	Topic

Reference Age and Sleeping Beauty Outcomes

Back

Panel A: Reference age				
	Mean	Min	Median	Max
Computer-Keyword Flag	0.025	-0.217***	-0.099***	1.300***
	(0.022)	(0.014)	(0.021)	(0.066)
Number of Authors	-0.188***	-0.100***	-0.186***	-0.314***
	(0.005)	(0.003)	(0.005)	(0.015)
NSF Grants (paper)	0.450**	-0.261**	0.190	2.944***
	(0.202)	(0.114)	(0.206)	(0.703)
Observations	2,912,619	2,912,619	2,912,619	2,912,619
R^2	0.564	0.545	0.551	0.490
Mean of Dep Var	8.554	2.705	7.134	20.159
Author/Year/Univ FE	Yes	Yes	Yes	Yes
Subject FE	Topic	Topic	Topic	Topic
Panel B: Reference dormancy (Sleeping Beauty scores)				
	Mean	Median Ref.	Oldest	Age90 Mean
Computer-Keyword Flag	14748.189***	5949.303*	47497.033***	42901.972***
	(2791.209)	(3497.185)	(9005.499)	(7858.461)

Reference Age and Sleeping Beauty Outcomes (Emb.)

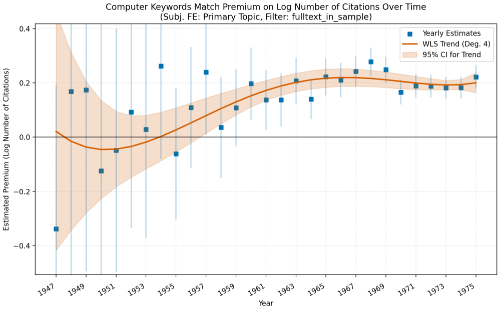
▸ Back

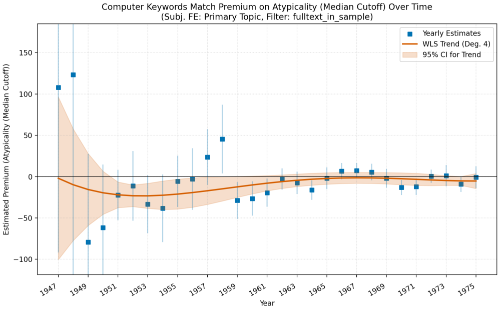
Panel A: Reference age				
	Mean	Min	Median	Max
Computer-Keyword Flag	-0.048*	-0.121***	-0.086***	0.851***
	(0.023)	(0.015)	(0.022)	(0.068)
Observations	1,267,437	1,267,437	1,267,437	1,267,437
R^2	0.000	0.000	0.000	0.000
Mean of Dep Var	-0.405	0.062	0.220	-0.255
Panel B: Reference dormancy (Sleeping Beauty scores)				
	Mean	Median Ref.	Oldest	Age90 Mean
Computer-Keyword Flag	19494.863***	16492.462***	48145.559***	38288.376***
	(3049.162)	(3818.570)	(9517.431)	(8363.198)
Observations	1,266,675	1,262,071	1,263,557	1,264,247
R^2	0.000	0.000	0.000	0.000
Mean of Dep Var	10413.798	-61481.215	52900.751	36815.254

Notes: Second-stage OLS regresses the residualized outcome on the residualized computer-keyword indicator. In the first-stage DML residualization, both treatment and outcome are partialled out on author, publication year, university, and primary-topic fixed effects; controls for number of authors and NSF grants; and concatenated abstract/title embeddings plus SPECTER embeddings. Residuals are estimated with 3-fold cross-fitting and SGD. Second-stage standard errors are clustered at the paper level (OpenAlex Work ID) and paper-author observations are weighted by the inverse number of authors. Sample restricted to searchable-full-text observations from 1947–1975.

Extended Computer Premiums Over Time

▶ Back

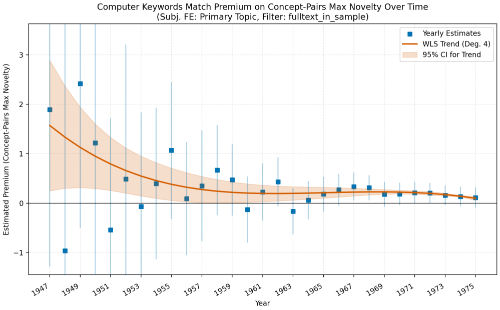


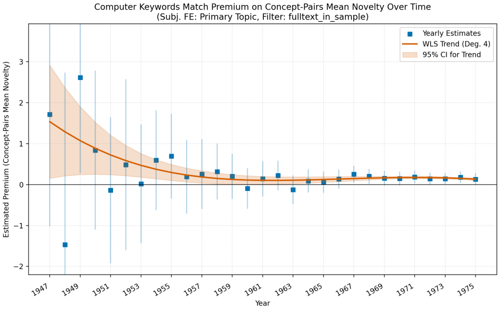


Notes: Yearly coefficients from regressing each outcome on computer-keyword $\times$ year dummies, controlling for university, author, and primary-topic fixed effects and number of authors. Shaded bands show 95% confidence intervals.

▶ Next

Extended Computer Premiums Over Time (II) Back





Notes: Yearly coefficients from regressing each outcome on computer-keyword $\times$ year dummies, controlling for university, author, and primary-topic fixed effects and number of authors. Shaded bands show 95% confidence intervals. Previous

Computer Usage Clustering

▸ Back

- Start from papers flagged as using or mentioning computers via keywords and LLM screening
- First-pass LLM classification separates four broad roles: tool, object of study, hardware/software, and other mentions
- Restrict to research articles using computers, then follow Tamkin et al. [2024]: Gemini 2.5 Flash descriptions, `gemini-embedding-001` embeddings, and K-means with $K = 61$
- Label lower-level clusters with Gemini 2.5 Pro using centroid and neighboring examples, then manually aggregate them into 11 high-level usage categories

Citation-Function Pipeline

Back

- Start from computer papers and matched controls built with embedding-based nearest neighbors under tight year/domain restrictions
- Download open-content citing papers from OpenAlex
- Recover inline citation context with GROBID-style tags or string/pattern matches
- Follow the usage-taxonomy workflow to write, embed, and cluster short descriptions of *why* the cited paper is used
- Current sample: 98,051 citation-context descriptions; 20,910 usable clustered links; 16,114 links to computer papers; 4,796 links to matched controls; 16,120 cited papers in the paper-level panel

Tamkin-style clustering

Citation-Function Examples ▸ Back

- Precedent = “people have talked about this before”;
- Established findings = “this paper shows the fact I’m invoking”
- Technical protocols = “this paper tells me how to do/measure it”;
- Formal method = “this paper gives the framework I’m using”;
- Canonical source = “this paper is the recognized origin of the concept/system”

Computer-paper example   Matched-control example

- **Technical evidence / protocols**: “as previously described” means the cited paper supplies a lab protocol or benchmark input; Close and Kidd (1969) plays the same empirical-benchmark role for a control paper.
- **Established findings as evidence**: Waterman and Horch (1966) or Daley et al. (1971) are cited as accepted evidence or factual support the new paper relies on.
- **Representative precedent**: Siegel (1974) or Posner, Nissen, and Klein (1976) appear as one example in a list showing the phenomenon or debate already existed in the literature.
- **Formal method / theory foundation**: Eisenthal and Cornish-Bowden (1974) or Wilkinson (1961) are cited as the estimator, derivation, or framework the analysis actually uses.
- **Canonical source for concepts / systems**: Hoare on monitors or the TAXIS data model are cited for provenance: the original concept, system, or language being invoked.
- Full brief with centroid-nearest inside examples, nearby outside examples, and matched computer-vs-control comparisons is in the outputs folder.

Citation Regressions ▸ Back Reg Results

- **Unit of analysis:** cited paper. Main 'FWCI' sample: $N = 15{,}938$ cited papers with at least one classified citation.
- **Overall bucket premium scores** (right panel on "Why Are Computer Papers Cited?"):

$$FWCI_p = \alpha + X_p'\gamma + \lambda_{\text{year}(p)} + \delta_{\text{domain}(p)} + \sum_{k=1}^{4} \theta_k s_{pk} + u_p$$

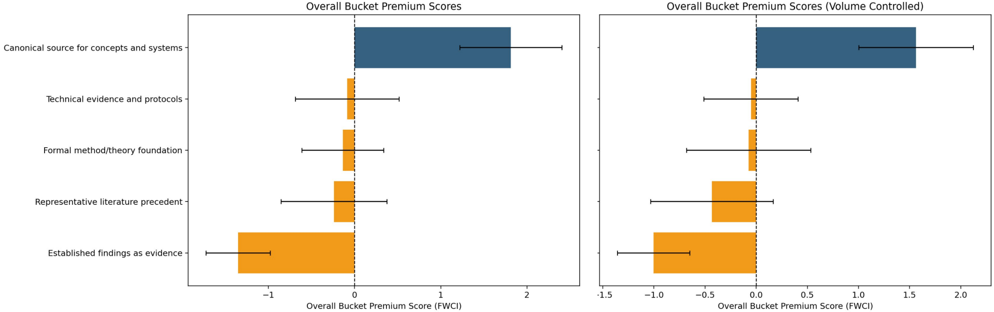
- Here s_{pk} is the share of paper p's classified incoming citations in cluster k. We omit cluster 0 in estimation, then re-center the fitted cluster coefficients so the five reported bucket scores average to zero.
- **Premium decomposition regression** ("Does Citation Function Help Explain the Premium?"):

$$FWCI_p = \alpha + \beta T_p + X_p'\gamma + \lambda_{\text{year}(p)} + \delta_{\text{domain}(p)} + B_p^* + \sum_{k=1}^{4} \theta_k s_{pk} + u_p$$

- T_p is the computer-paper indicator. B_p^* is excess breadth: residual effective clusters after partialling out classified-citation-count bins, controls, and fixed effects.
- The attenuation logic is: compare the baseline treated coefficient to the treated coefficient after adding B_p^* and the cluster-share block. That difference is the part of the computer-paper premium accounted for by citation-function structure.

Bucket Premium Scores

Back



Notes: Left: cited-paper regression of FWCI on cluster shares, controlling for number of authors, NSF grants, cited-paper year FEs, and domain FEs; the treated dummy is omitted so plotted coefficients summarize the overall association of each bucket with citation impact. Scores re-centered so all five buckets are directly comparable; whiskers are 95% CIs. Right: same specification adding $\log(1 + \text{classified citation count})$ to partial out citation volume. Ranking is stable across both specifications: canonical source remains the highest-premium bucket, established findings the lowest.

Citation Premium Decomposition

▸ Back

Outcome: FWCI

Spec	Treat. coef.	Atten.	p-value
Baseline	1.193	–	0.000
+ Excess breadth	1.179	1.2%	0.000
+ Composition	1.086	9.0%	0.000
+ Breadth + composition	1.077	9.8%	0.000
+ Volume (robustness)	0.773	35.2%	0.000
+ Volume + breadth + composition	0.657	45.0%	0.001

Notes: Unit = cited paper ($N = 15{,}938$ with at least one classified citation). Left panel: FWCI regressed on the computer-paper indicator, number of authors, NSF grants, cited-paper year FEs, and domain FEs. Rows add excess breadth, composition (cluster shares), or both. Composition controls are citation-cluster shares: the fraction of a paper's classified incoming citation links falling in each of the five function clusters. Excess breadth is constructed in two steps: (1) regress the number of effective clusters (i.e. the inverse Herfindahl across citation-function categories) on flexible classified-citation-count bins, the baseline controls, and fixed effects; (2) take the residual as the excess-breadth measure and enter it as an additional control. This isolates breadth of citation reasons beyond what is mechanically explained by having more classified citations. Right panel: Shapley decomposition splits the explained attenuation (0.117) into shares attributable to each block, averaging across both orderings (breadth-then-composition and composition-then-breadth).

- Baseline premium: $\hat{\beta}_0 = 1.193$
- After breadth + composition: $\hat{\beta}_{B*+C} = 1.077$
- Explained part: $1.193 - 1.077 = 0.117$
- Shapley decomposition of explained part:
  - Composition (mix of citation reasons): $\sim$90%
  - Excess breadth: $\sim$10%
- Excess breadth = residual effective clusters after partialling out citation-count bins, controls, and FEs
- Caveat: breadth and citations jointly determined — descriptive accounting, not causal mediation

Author-Level Patterns

Back

	(1) Log Works	(2) Log Cites	(3) H-Index	(4) # Topics	(5) # Affiliations
Computer Adopter	1.364***	1.463***	7.752***	5.247***	1.820***
	(0.0165)	(0.120)	(1.254)	(0.294)	(0.166)
Number of Works		0.00536**	0.0577**	0.0120**	0.00617**
		(0.00176)	(0.0185)	(0.00399)	(0.00204)
Observations	316,970	316,970	316,970	316,970	316,970
R^2	0.309	0.437	0.533	0.292	0.327
Mean of Dep Var	2.732	5.158	12.16	16.77	3.525
Affiliation FE	Yes	Yes	Yes	Yes	Yes
Cohort FE	Yes	Yes	Yes	Yes	Yes
Subject FE	Topic	Topic	Topic	Topic	Topic

Notes: Standard errors clustered at the affiliation level. Outcomes are measured over the author's entire career. Cohort is defined by the year of an author's first publication in the dataset. Affiliation is the modal one across the author's publications. Field (area) is the modal topic across the author's publications. $^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$.

Author Patterns: Early vs. Late Adopters

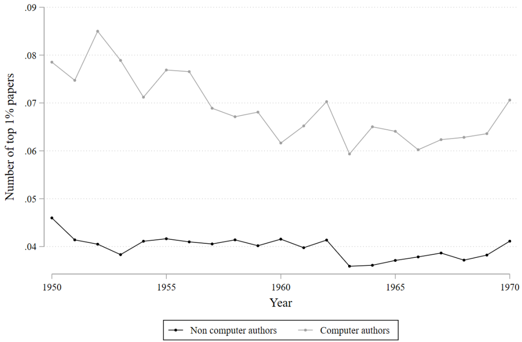
▸ Back

	(1) Log Works	(2) Log Cites	(3) H-Index	(4) # Topics	(5) # Affiliations
Adoption Lag (Freq)	0.00817	0.0109	-0.0391	0.0713***	-0.0405*
	(0.00440)	(0.00628)	(0.0368)	(0.0179)	(0.0177)
Number of Works		0.00515***	0.0782***	0.00283***	0.0118***
		(0.000364)	(0.00427)	(0.000380)	(0.000983)
Observations	6,141	6,141	6,141	6,141	6,141
R^2	0.373	0.601	0.759	0.341	0.524
Mean of Dep Var	4.358	7.271	23.65	24.01	6.636
Affiliation/Cohort/Topic FE	Yes	Yes	Yes	Yes	Yes

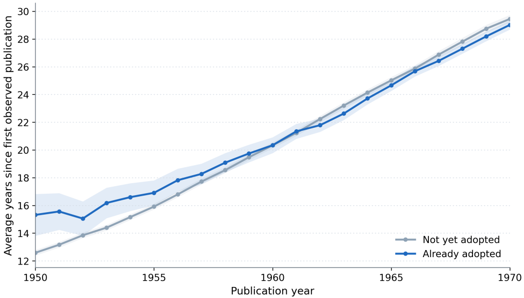
SE clustered at affiliation level. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Author-Level Patterns: Top Citations & Experience

Back



More Top 1% Cited Papers



Longer Publishing Careers (Experience)

Author Patterns: Outcomes at University Computer Adoption Year

Back

	(1) Log Works	(2) Log Cites	(3) H-Index	(4) Top 1%	(5) Top 10%
Computer Adopter	0.227***	0.592***	1.652***	0.576**	0.0967***
	(0.0182)	(0.0364)	(0.161)	(0.177)	(0.0217)
Number of Works		0.000731*	0.00350	0.00347	0.000412
		(0.000369)	(0.00184)	(0.00204)	(0.000242)
Observations	122,159	134,521	134,521	122,159	122,159
R^2	0.344	0.251	0.246	0.169	0.0964
Mean of Dep Var	1.530	3.460	4.438	2.581	0.267
Affiliation/Cohort/Topic FE	Yes	Yes	Yes	Yes	Yes

SE clustered at affiliation level. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

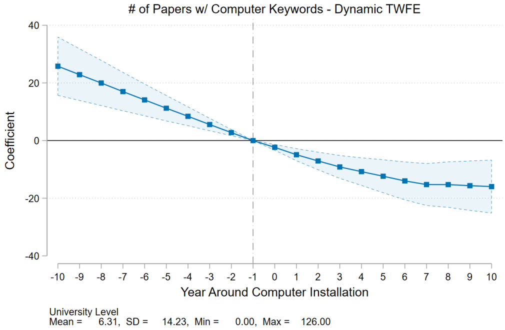
Author Patterns: Intensive Margin

Back

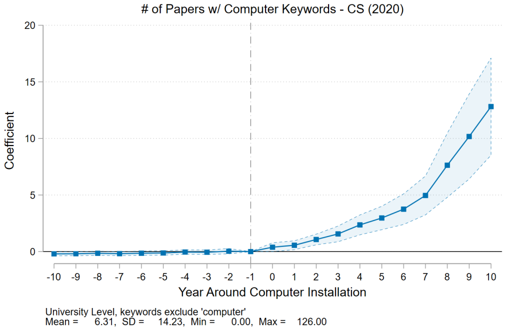
	(1) Log Works	(2) Log Cites	(3) H-Index	(4) Topics	(5) Affiliations
Computer Paper Count	0.188***	0.181***	1.166***	0.553***	0.258***
	(0.00790)	(0.0256)	(0.250)	(0.0639)	(0.0294)
Number of Works		0.00553**	0.0582**	0.0128**	0.00631**
		(0.00185)	(0.0189)	(0.00435)	(0.00208)
Observations	316,970	316,970	316,970	316,970	316,970
R^2	0.280	0.417	0.524	0.269	0.317
Mean of Dep Var	2.732	5.158	12.16	16.77	3.525
Affiliation/Cohort/Topic FE	Yes	Yes	Yes	Yes	Yes

SE clustered at affiliation level. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Why TWFE Fails



TWFE event study

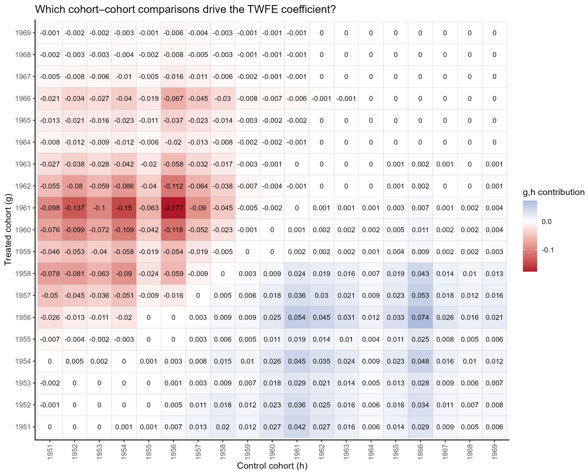


Callaway-Sant'Anna event study

Same sample and treatment definition; TWFE flips the sign because staggered timing and heterogeneous effects induce forbidden comparisons.

Back to design Back to usage Bacon

Why TWFE Fails: Bacon Decomposition



Goodman-Bacon decomposition of a simple pre/post TWFE

Back to design Back to usage

Model	Estimate	SE	p-value	N
TWFE	-2.036	0.489	0.000	5 704
Callaway-Sant'Anna (2021)	5.417	0.716	0.000	-
de Chaisemartin-D'Haultfoeuille (2020)	5.417	0.663	0.000	3 312

Simple pre/post DiD: TWFE flips the sign, while modern estimators remain positive.

Classifier Details

- LightGBM is a gradient-boosting model that combines many small decision trees by training new trees on the residuals of the ensemble prediction.
- It is useful here because it can learn nonlinear interactions between titles, metadata, and embeddings without us hand-coding them.
- The output is a probability vector over paper types, which we can aggregate into university-year methodology shares.

Benchmark	Accuracy	Weighted F1	Macro F1
Main LightGBM (validation)	0.752	0.749	0.657
10k full-text benchmark	0.824	0.825	0.760
Fallback no-lexicon model	0.814	0.814	0.751

Universe	N
Unique papers in 1951–1969 panel slice	470,395
Classified unique papers	268,887
Keyword-matched computer papers in panel slice	25,665
Keyword-matched computer papers among classified	17,667
Classified papers with local full text	100,378
Already in Gemini first-pass set	56,650

Notes: The main production model is LightGBM. The appendix keeps only the benchmark numbers we actually need here: the rebuilt validation score, the 10k full-text benchmark, and the no-lexicon fallback.

▸ Back

What the Classifier Measures ▸ Back

- Each paper is assigned a probability distribution over labels
- For the panels, we aggregate the probability mass to account for uncertainty of the classifier
- 57% of papers are assigned scores in our panel sample

Bucket	Mass share	What it mainly captures
Empirical	48.3%	Data, experiments, field evidence, or observational measurement.
Theory	18.3%	Formal derivations, conceptual models, or theoretical argument.
Methods	11.1%	Algorithms, procedures, software, and measurement techniques.
Simulation	1.1%	Simulation, computational modeling, or numerical experimentation.
Other	21.3%	Reviews, bibliographies, editorials, and reference-type material.

Notes: Percentages are average predicted probability mass across the 268,887 classified papers in the 1951–1969 diagnostic slice, not argmax shares.

▸ Raw classes

Raw 7-Class Taxonomy

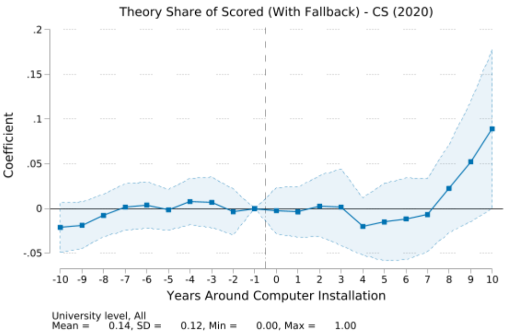
Raw class	Mass share	Maps to
Quantitative empirical	31.1%	Empirical
Qualitative empirical	17.2%	Empirical
Formal theoretical	12.9%	Theory
Discursive theoretical	5.4%	Theory
Methods	11.1%	Methods
Computational / Simulation	1.1%	Simulation
Other	21.3%	Other

Notes: Shares are average predicted mass across the 268,887 classified papers in the 1951–1969 panel slice. Raw labels collapse into the five panel buckets in the main text. Computational papers are rare in this period; other is mostly reviews, bibliographies, editorials, and similar reference material rather than a generic model-failure bucket.

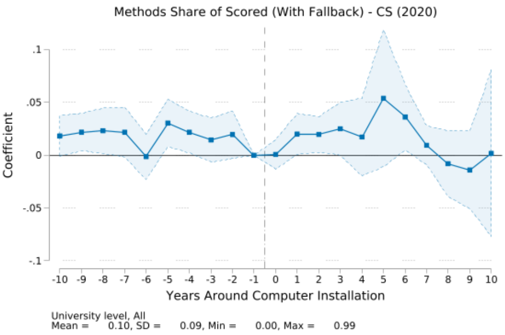
▸ Back

Changes in Research Methodologies: Theory and Methods

Theory Share of Scored Papers



Methods Share of Scored Papers



Notes: These are the same 'ptype_shr_wf' event-study objects as in the main text, shown here for the two categories that appear to absorb most of the reallocation inside the scored subset.

▶ Back

Paper-Level Method Regressions

Back

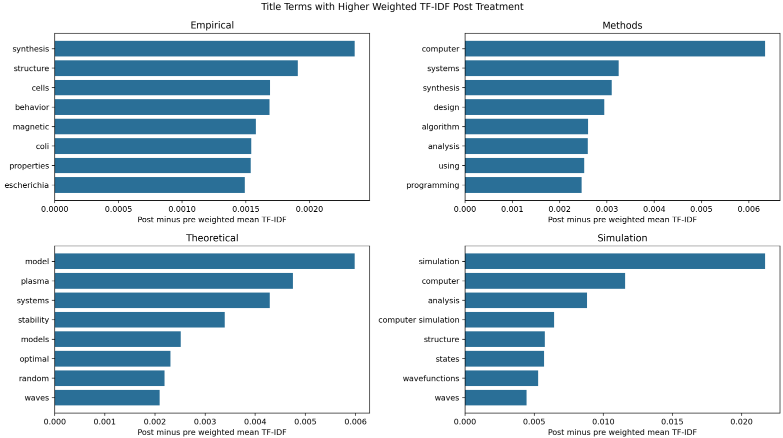
	Empirical	Theory	Simulation	Methods	log(Emp./Theory)
Computer-Keyword Flag	-0.043***	-0.018***	0.028***	0.036***	-0.034***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.002)
# Authors	0.028***	-0.015***	0.000	0.000	0.049***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.001)
NSF grants (paper)	0.029***	-0.005*	0.001	-0.006*	0.038***
	(0.005)	(0.002)	(0.001)	(0.003)	(0.007)
Observations	2,128,533	2,128,533	2,128,533	2,128,533	2,128,533
R^2	0.798	0.816	0.572	0.667	0.841
Mean of Dep Var	0.466	0.146	0.009	0.124	0.375
Author/Year/Univ FE	Yes	Yes	Yes	Yes	Yes
Subject FE	Topic	Topic	Topic	Topic	Topic

Notes: Paper-author observations, weighted by the inverse number of authors. Treatment indicates computer-related keywords in the paper's full text. Controls: number of authors and NSF grants citing the work. Fixed effects: author, publication year, university, and primary topic. Standard errors clustered at the paper level (OpenAlex Work ID). Sample restricted to papers with at least one in-sample university affiliation and searchable full text in OpenAlex. Negative values in log(Emp./Theory) indicate relatively more theoretical than empirical mass.

Transparent Family Dictionaries

Family	Definition and example language
Computer systems / programming	Explicit computer, compiler, programming-language, assembler, time-sharing, or computer-center language. Example terms: computer, compiler, ALGOL, program, time-sharing.
Algorithmic / numerical	Algorithms, numerical procedures, matrix computation, finite-difference, and finite-element schemes. Example terms: algorithm, numerical, matrix, finite element.
Simulation / modeling	Simulation, simulation models, numerical forecasting, or model-based experimentation. Example terms: simulation, model, modeling, dynamics.
Control / optimization	Feedback, control, queueing, encoding, and related systems or optimization problems. Example terms: control, optimal, feedback, queueing, encoding.
Laboratory / measurement	Experimental apparatus, assay, protocol, or measurement language that often accompanies computer-assisted recording and monitoring. Example terms: apparatus, assay, procedure, perfusion, measurement.
Statistical / predictive	Quantitative processing of observed data. Example terms: regression, statistical, variance, correlation, prediction.
Biomedical signal / detection	Empirical signal-processing and diagnostic language. Example terms: electrocardiogram, heart sounds, body-surface potential, detection, spectral analysis.
Statistics / sampling	Statistical inference, regression, variance analysis, or survey/sampling design. Example terms: statistical, regression, sampling, inference.
Differential equations	PDE/ODE and boundary-value or continuum-model problems. Example terms: differential equation, boundary value, PDE, ODE.
Materials / synthesis	Chemistry and materials language used as a placebo-style comparison family rather than a claim about the main computer mechanism. Example terms: chemical synthesis, enzymatic synthesis, polymer alloy.

Title Terms Behind the Shift



Notes: Title terms that rise most, post minus pre, by methodology bucket (1951–1969 classified slice, 268,887 papers; 10-year pre/post window around each university's installation year). Within each bucket row, bars rank title tokens by the difference between post- and pre-treatment probability-weighted mean TF-IDF. Computed over the full title vocabulary, not a hand-picked lexicon.

▸ Back

Examples Behind the Labels

Empirical computer use

- *Study of High Frequency Components in Electrocardiogram by Power Spectrum Analysis* (W2003343840): measured biomedical signals processed with spectral analysis.
- *The Detection of Heart Disease in Children* (W2974997846): tape-recorded heart sounds used for computer-assisted screening and classification.
- *The Prediction of Flow Patterns, Liquid Holdup and Pressure Losses Occurring During Continuous Two-Phase Flow In Horizontal Pipelines* (W1975029220): numerical prediction built on observed engineering measurements.

Methods

- *The structure of yet another ALGOL compiler* (W2050572856): compiler architecture and implementation.
- *Programming Technique: An improved hash code for scatter storage* (W2003248512): programming-method paper on data storage and retrieval.
- *The simplex method of linear programming using LU decomposition* (W2071877138): numerical optimization method built for computation.

Computer-adjacent theory

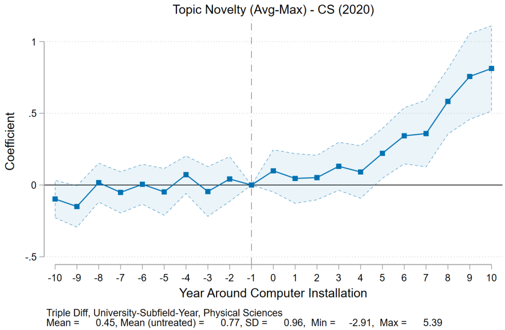
- *Feedback Queueing Models for Time-Shared Systems* (W2074065133): queueing/control theory aimed at time-sharing computer systems.
- *Levels of computer systems* (W1983161084): formal systems-theory discussion explicitly about computer architecture.
- *Source encoding in the presence of random disturbance* (W1584278176): information/encoding theory that sits on the computation-communications interface.

Simulation

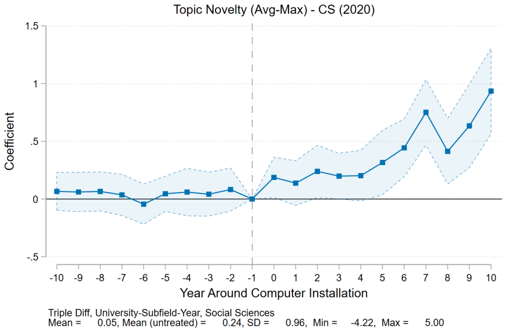
DDD: Content of Science, Topic Novelty

Back

Topic Novelty: Physical Sciences



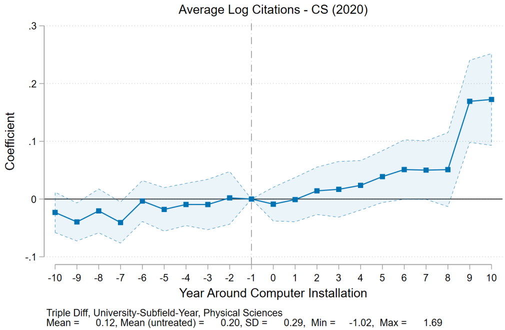
Topic Novelty: Social Sciences



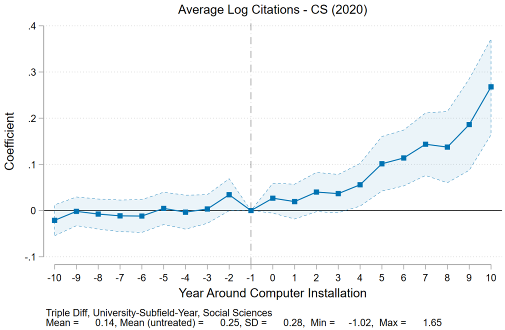
DDD: Quality of Science, Avg Citations

Back

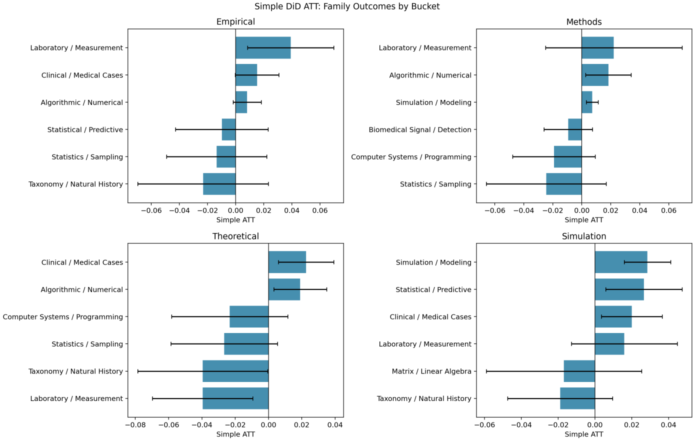
Avg Citations: Physical Sciences



Avg Citations: Social Sciences



Keyword-Family DiDs



Notes: Callaway-Sant'Anna ATTs from running the same classifier-bucket DiD on the probability-weighted hit rate of each transparent keyword family. Rows are methodology buckets and columns are families. These are robustness objects: they show that the strongest classifier-based family shifts also show up when we run a DiD directly on the family hit rates, not just in pre/post comparisons.

▶ Back