

Cumulative Discovery and Moving Frontiers

Why can more knowledge make some problems easier while progress becomes harder to measure?

Complexity, Information, and Cross-Domain Transfer

How could improved methods accompany a falling success rate?

Imagine validated success falling from 0.66 to 0.56 after a new method arrives. A worse method, changed reporting, and more ambitious question choice could all produce that pattern.

If knowledge enables harder questions, the attempted problems are themselves an outcome. This suggests modeling question selection alongside capability.

Choose a measurement that can separate the explanations

Candidate measure	What it resolves—and misses
Current portfolio success	Actual yield; confounds capability with task mix
A fixed fresh task panel	Comparable capability; misses new questions
Stratified outcomes + task shares	Separates within-task gains from mixture change

Record attempted tasks and their outcomes, not just publications. A changing portfolio and a fixed benchmark answer complementary questions.

A minimal transition makes the feedback explicit

$$K_{t+1} = \max(K_t, V_t); \quad P(T_t=B \mid K_t) = 0.2 + 0.6K_t$$

Let V_t indicate validation of a reusable method. Its availability K switches from 0 to 1; success on hard tasks B rises from 0.1 to 0.5, while familiar tasks A stay at 0.8.

Suppose the share of B rises from 0.2 to 0.8 after validation. Measured success falls from 0.66 to 0.56 even though capability improves.

A fixed frontier and a moving frontier can disagree

Fixed mixture: $0.8 \times 0.8 + 0.2 \times 0.5 = 0.74$

In the opening transition model, success on hard tasks rises from 0.1 to 0.5, while their portfolio share rises from 0.2 to 0.8.

Hold the original mixture fixed: later success would be 0.74. Allow question choice to change: it is 0.56. Both should accompany the original 0.66.

A dynamic discovery model

$$\text{outcomes}_t \rightarrow K_{t+1} \rightarrow (\text{methods}_{t+1}, P_{t+1})$$

Let P_t be the task distribution and K_t the usable knowledge state. Outcomes at time t update knowledge, which can also alter the next task distribution.

This creates feedback between capability, question choice and measured difficulty.

Knowledge is more than a collection of statements

Reusable knowledge includes proved lemmas, tested software, material designs, measurement protocols and reliable negative results.

Its practical value depends on discoverability, compatibility and the cost of applying it to a new problem.

A reusable lemma changes the cost of a later proof

A previously proved statement can replace a repeated derivation, reducing search depth and verification effort for later results.

The benefit depends on whether the new problem can be expressed in terms of that lemma and whether the library is accessible.

Amortized cost separates creation and reuse

$$\text{average cost}(T) = B/T + u$$

Let B be the cost of building a validated component and u the cost of using it once. Across T uses, average cost is $B/T + u$.

Large reuse can make an initially expensive discovery highly valuable. Maintenance and adaptation add further terms.

DreamCoder demonstrates one reuse mechanism

Learning a library changes the program grammar, making useful compositions shorter and easier to search.

This provides a concrete mechanism for cumulative capability inside a specified task distribution. It does not establish accelerating returns for every future task.

Recombination expands the candidate space

$$\text{number of distinct pairs} = K(K-1)/2$$

With K reusable components, there are $K(K-1)/2$ unordered distinct pairs. New combinations can create opportunities unavailable from isolated components.

Most possible combinations may be useless or infeasible; combinatorial growth in possibilities is not a growth law for discoveries.

A larger space can increase both opportunity and burden

If only a fraction $q(K)$ of component pairs are useful, expected useful pairs are $q(K)K(K-1)/2$ under a simple common-probability model.

If $q(K)$ declines quickly enough, more components need not increase useful discoveries at the same rate.

Instruments can change identifiability

A new measurement technology can reveal a variable that separates previously indistinguishable mechanisms.

Its contribution is a change in experimental access, not merely an increase in sample count from the old channel.

A designed molecule can expand experimental access

A validated binder may act as a measurement or perturbation tool. This can enable a new class of mechanistic tests.

The later discovery inherits both the capability and the uncertainties of the tool.

Negative knowledge can save future resources

A reliable failure can rule out a mechanism, expose an assay artifact or identify a representation's limitation.

Its value depends on recording conditions precisely enough that later researchers know when the failure applies.

Publication selection changes the visible success rate

$$P(\text{positive} \mid \text{published}) = ar / [ar + b(1-r)]$$

Let r be the conducted-study positive rate; a and b are publication probabilities for positive and null results. Toy values: $r=.20$, $a=1$, $b=.10$.

Among 100 studies, expect 20 positive and 80 null results, but 20 positive and 8 null publications. The published positive fraction is $20/28 \approx 71\%$; the conducted-study rate is 20%.

A publication reward can favor weaker evidence

Design	Studies	Power	True positives	False positives
Cost 2 each	100	.80	16	4
Cost 1 each	200	.40	16	8

Toy budget 200: both designs study a population with 20% real effects and false-positive probability .05. High-power studies cost twice as much. Rewarding each positive publication favors the cheap design: 24 versus 20 expected positives, with the same 16 true positives.

The burden of knowledge can increase training costs

Jones models how accumulated knowledge can increase the preparation needed to reach a frontier, encouraging specialization and teamwork.

Knowledge can therefore lower task-level execution costs while increasing the cost of becoming able to choose and solve frontier tasks.

Specialization trades expertise for coordination

A specialist can work deeply within a narrower area. Combining specialists can cover more knowledge but introduces communication and integration costs.

Team size is neither a direct measure of difficulty nor a guaranteed route to more original outcomes.

Research productivity needs an output denominator

$$\text{research productivity} = \text{measured progress} / \text{effective effort}$$

A simple accounting relation divides measured progress by effective research effort.

Changing the output definition, field composition or researcher-quality adjustment changes the estimated productivity trend.

Moore's law: output divided by research effort

Research productivity = chip-density growth / effective researchers

Bloom et al. treat chip density doubling roughly every two years as an output-growth rate near 35% per year. Effective semiconductor research effort rises about 18-fold, 1971–2014.

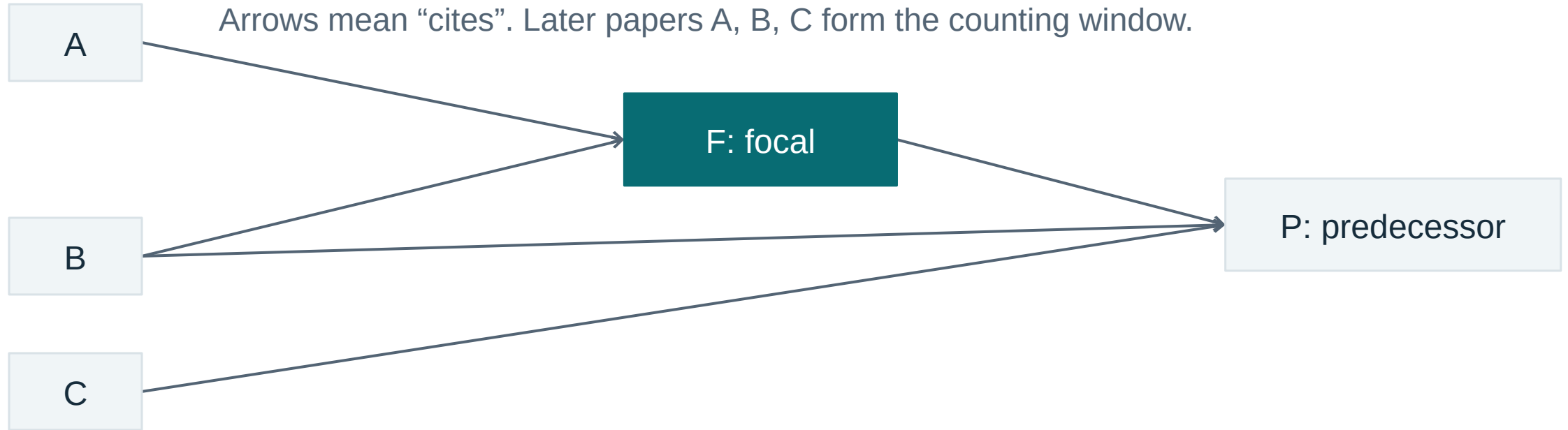
With approximately constant output growth, output growth per effective researcher falls by about 18-fold in this construction.

The denominator changes the magnitude

Research-effort definition	1971–2014 multiplier
Baseline broad	18×
Narrow	8×
Narrow, downweighted	11×
Broad, downweighted	26×

The study's alternative allocations produce the following research-effort multipliers over 1971–2014. These are robustness specifications, not confidence intervals.

A small citation graph makes disruption concrete



A focal paper F cites predecessor P. Later paper A cites F only; B cites both F and P; C cites P only. The disruption index is $(1-1)/(1+1+1)=0$. Add D citing F only and it becomes $(2-1)/4=0.25$.

A citation index is sensitive to its window

$$D = (n_i - n_j) / (n_i + n_j + n_k)$$

Start at $(n_i, n_j, n_k) = (1, 1, 1)$: $D=0$. Add one focal-only citation to get $(2, 1, 1)$: $D=1/4$.

From $(2, 1, 1)$, add five predecessor-only citations: $(2, 1, 6)$ gives $D=1/9$. Starting from $(1, 1, 1)$ instead would leave $D=0$.

Team size and disruption are associated

Wu, Wang and Evans find different patterns of development and disruption across team sizes in studied scientific and technological outputs.

The association does not show that assigning a smaller team to the same problem would necessarily cause a more disruptive result.

Scientific leadership as a quasi-experiment

Azoulay, Fons-Rosen and Graff Zivin study 452 deceased eminent life scientists, comparing their publication-defined subfields with matched control subfields.

The design compares publication changes around death, separating prior collaborators from non-collaborators. A raw before/after increase would confound the event with field trends.

Entry changes; the mechanism remains conditional

$$\text{Relative change} = 100 [\exp(0.082) - 1] \approx 8.55\%$$

Table 3 reports a non-collaborator coefficient 0.082 (SE 0.029):
 $100[\exp(0.082)-1] \approx 8.55\%$ more publications. Collaborator output falls about 20.7%.

This is a local entry response, not an 8.55% gain in total scientific progress. The study cannot make publication counts a measure of discovery value.

2026: individual impact and collective breadth can diverge

Hao and colleagues study AI use in scientific research and report associations with individual impact alongside a narrowing of collective topic coverage.

Output volume and citations can rise while the portfolio of questions becomes more concentrated.

Ten agreeing reports may share one source

Ten researchers or agents report the same result. If all rely on one dataset or copy one analysis, counting agreement as ten independent checks exaggerates the evidence.

The new direction is to trace how each claim was produced. Similar outputs can reflect independent corroboration, a common source or a shared measurement error.

Different checks address different shared failures

Candidate check	What it addresses	What can remain shared
Count votes	Individual idiosyncratic errors	Source, data and assumptions
Trace provenance	Repeated versions of one claim	Original measurement bias
Repeat independently	Sampling and execution variation	Instrument or model bias
Use another method	Method-specific failure	Common background assumptions

The audit changes the choice: copied reports call for source tracing; a shared assay bias calls for a different measurement method. Disagreement can reveal a useful fault line.

Common error limits the gain from agreement

$$\text{Var}(\bar{X}) = \tau^2 + \sigma^2/n$$

Let $X_i = \theta + U + \varepsilon_i$. The shared error U has variance τ^2 ; independent errors ε_i have variance σ^2 . All errors have mean zero, and U is independent of every ε_i .

Then $\text{Var}(\bar{X}) = \tau^2 + \sigma^2/n$. With $\tau^2 = \sigma^2 = 1$ and $n = 10$, it is 1.1, versus 0.2 for ten independent measurements of variance 2. More copies cannot average away the shared component.

A dynamic cost model needs several components

$$\text{total cost}_t = B_t + S_t + V_t + A_t$$

Let B_t be preparation cost, S_t search cost, V_t validation cost and A_t adaptation cost. Knowledge can change each in a different direction.

The total can rise while one component falls sharply.

The four examples after accumulation

Domain	Reusable asset	New frontier or burden
Mathematics	Lemma and construction library	Harder statements; library search
Programs	Validated primitives	Larger specifications; integration
Molecules	Scaffolds and assay platforms	Harder targets; context validation
Mechanisms	Intervention tools and causal facts	Richer models; new ambiguities

Knowledge changes both the method and the set of worthwhile questions.

Conclusion: progress changes its own conditions

Accumulation can simplify fixed tasks, expand experimental access and move researchers toward harder questions. It can also increase preparation and coordination costs.

Next: a predictive theory must model these opposing effects without claiming a universal law prematurely.

References

Thomas Hubert et al. (2025). Olympiad-level formal mathematical reasoning with reinforcement learning. Nature; journal article, 2025

Bogdan Georgiev, Javier Gómez-Serrano, Terence Tao and Adam Zsolt Wagner (2025). Mathematical exploration and discovery at scale. arXiv:2511.02864v1; preprint, 3 November 2025

Kevin Ellis et al. (2021). DreamCoder: Bootstrapping Inductive Program Synthesis with Wake-Sleep Library Learning. PLDI 2021; conference paper; preprint has earlier title

Benjamin F. Jones (2009). The Burden of Knowledge and the Death of the Renaissance Man: Is Innovation Getting Harder?. Review of Economic Studies 76, 283–317; journal article

References · continued

Martin L. Weitzman (1998). Recombinant Growth. Quarterly Journal of Economics 113, 331–360; journal article

Judea Pearl and Elias Bareinboim (2014). External Validity: From Do-Calculus to Transportability Across Populations. Statistical Science 29, 579–595; journal article; arXiv posted 2015

Jonas Peters, Peter Bühlmann and Nicolai Meinshausen (2016). Causal inference by using invariant prediction: identification and confidence intervals. Journal of the Royal Statistical Society B 78, 947–1012; journal article

Martin Pacesa et al. (2025). One-shot design of functional protein binders with BindCraft. Nature 646, 483–492; online 27 August 2025

References · continued

Annie Franco, Neil Malhotra and Gabor Simonovits (2014). Publication bias in the social sciences: Unlocking the file drawer. Science 345(6203), 1502–1505; published primary research

Lingfei Wu, Dashun Wang and James A. Evans (2019). Large teams develop and small teams disrupt science and technology. Nature 566, 378–382; journal article

Nicholas Bloom, Charles I. Jones, John Van Reenen and Michael Webb (2020). Are Ideas Getting Harder to Find?. American Economic Review 110, 1104–1144; journal article

Michael Park, Erin Leahey and Russell J. Funk (2023). Papers and patents are becoming less disruptive over time. Nature 613, 138–144; journal article

References · continued

Pierre Azoulay, Christian Fons-Rosen and Joshua S. Graff Zivin (2019). Does Science Advance One Funeral at a Time?. American Economic Review 109(8), 2889–2920; journal article

Qianyue Hao, Fengli Xu, Yong Li and James Evans (2026). Artificial intelligence tools expand scientists' impact but contract science's focus. Nature 649, 1237–1243; journal article

UCL Science and Technology Studies (2025–26). HPSC0162: Social Epistemology of Science. UCL course syllabus