Sources of Inefficiency in Scientific Research

Maya Durvasula Zane Kashner*

September 11, 2024

Extended Abstract

Scientists focus nearly all effort on a small fraction of potential topics. Of the $\sim 20,000$ protein-coding human genes—the biological objects that determine, among other things, how diseases manifest and how well they respond to treatment—roughly ten percent receive nearly all scientific attention (Stoeger et al., 2018). More than half of these genes have never been the subject of a single focused scientific paper. Case studies of other basic building blocks of biomedical research—kinases (Grueneberg et al. 2008), proteins (Baker 2009), human embryonic stem cell lines (Scott et al. 2009), tumor cells (Edwards et al. 2011), model organisms (Kustatscher et al. 2022)—reveal similar trends.

Extreme skewness in the distribution of scientific effort across potential topics is not itself evidence of inefficiency. The true distribution of the social returns to research may be similarly skewed.[1] Empirically, however, various indicators of the scientific and medical importance of these unstudied topics indicate otherwise. Genes that have been linked repeatedly to cancer, for example, have received *zero* focused study. Similar patterns have been documented for genes and proteins associated with the COVID-19 virus. From a more fundamental perspective, if the value of basic science becomes apparent only with substantial delay, it is challenging to determine ex ante that a particular topic warrants no focused inquiry (Edwards et al., 2011).

These concerns are the starting point for this paper. We combine theory and empirics to investigate whether patterns such as these—highly sparse, concentrated, and uneven clustering of scientists across "idea space"—are indicative of inefficiencies in scientific investment. We identify three economic forces—knowledge spillovers, path dependence in scientist project choice,

*Durvasula: JD/PhD Candidate, Stanford Dept. of Economics & Stanford Law School (mdurvas@stanford.edu); Kashner: PhD Candidate, Stanford Graduate School of Business (zkashner@stanford.edu). We are very grateful to Nicholas Bloom, Matthew Brown, Bharat Chandar, Cody Cook, Jose Ignacio Cuesta, Ben Davies, Matthew Jackson, Ravi Jagadeesan, Charles Jones, Pete Klenow, Kyle Myers, David Ritzwoller, Brad Ross, Isaac Sorkin, Alessandra Voena, Ali Yurukoglu, seminar audiences, and especially our advisors, Rebecca Diamond, Liran Einav, Matthew Gentzkow, Neale Mahoney, Shoshana Vasserman, and Heidi Williams for detailed, thoughtful feedback. We thank Brad Ross for sharing code; Kate Barron and Ron Nakao for assistance with data access; and various officials at the NIH for guidance on data availability. We gratefully acknowledge support from the National Science Foundation under Grant No. DGE-1656518 to Stanford University and the Knight Hennessy Scholars Program.

[1]The idea that a small number of potential ideas are likely to yield socially valuable knowledge is standard in literatures studying growth and innovation. George Akerlof, as reported in Jones (2021), characterized this as the idea that, while the number of potential scientific ideas is extremely large, "most of them [are] like chicken ice cream."

and heterogeneity in scientist productivity—that, together, yield highly skewed distributions of research effort and distort investments away from the social optimum. Using newly-constructed data on the universe of investments in biomedical research, we document evidence of each source of inefficiency empirically. We draw on these descriptive facts to develop and estimate a quantitative model of investment. In ongoing work, we use estimated parameters to analyze counterfactual distributions of research effort under alternative policy designs.

Choices made by individual scientists, in the aggregate, determine the rate and direction of scientific progress. We, thus, develop a simple theoretical model of scientist project choice, which clarifies the conditions under which scientists choose to locate in close proximity to one another and, in turn, whether these co-location decisions differ from what is socially optimal. The scientist's problem is one of sequential investment. In each period, the scientist must choose from a set of potential projects by weighing costs against benefits. There are two key features in this model. First, scientists begin their independent careers with an initial stock of knowledge, formed during a period of apprenticeship to a senior scientist. Second, the costs to a scientist of selecting any potential project are determined both by its proximity to existing research projects and by its proximity to their own portfolio of past projects.

We show that, in the presence of knowledge spillovers, this dependence across periods in scientists' costs yields two distortions. Individual scientists select projects in close proximity to their initial stock of expertise, and aggregate investment is clustered around existing research domains. The extent of this distortion is determined by the distribution of scientist productivity. If the population includes only a small number of highly-productive scientists—who produce more projects, more students, and larger spillovers than their peers—the skewness of the distribution of research effort becomes more extreme.

To test for and quantify these sources of inefficiency empirically, we construct new data on scientist careers and project choices. Using data extracted from the texts of scientific publications and methods from natural language processing, we introduce a metric that parameterizes the set of potential scientific topics in a high-dimensional space. We construct granular panel data for roughly $200,000$ scientists, which records potential investments, realized projects, costs, and payoffs over time.

Using these data, we collect empirical evidence consistent with the existence of each source of inefficiency identified in our theoretical model. In doing so, we document what are, to our knowledge, three new sets of descriptive facts on patterns of scientific investment. First, we find that scientists, with somewhat remarkable consistency, select projects by trading off the similarity of the project to their existing stock of knowledge with its expected payoffs. Put differently, scientists select projects that are more dissimilar from their previous work when expected payoffs are higher. We reproduce a version of this fact in Figure 1. Second, we find evidence of quantitatively significant path dependence in scientists' choices across periods. We identify a source of idiosyncratic variation in scientists' initial stocks of knowledge—generated by patterns of sorting into postdoctoral training groups—which allows us to distinguish path dependence in investment from persistent heterogeneity. We show that small differences in early training environments have substantial, persistent effects on specialization. Third, we collect evidence of significant heterogeneity across scientists both in their production of projects and in their "production" of students.[2]

[2]Since Lotka (1926), extreme inequality in the distribution of publications and citations across scientists has been a well-documented phenomenon, typically attributed to differences among scientists in the ability to do high-impact

These empirical patterns, interpreted against the results of our theoretical model, provide support for the idea that the clustering we observe in scientific topic space is, in fact, inefficient.

To quantify the extent of inefficiency and analyze how potential policy levers might affect the distribution of research effort, we develop and estimate a quantitative model of scientific investment. This model allows us to estimate, for each scientist in our sample, the following parameters: the cost of executing a project by its dissimilarity from their existing knowledge, preferences for novelty, ability to recover a project's latent value, and a measure of risk aversion.[3]

With these estimated parameters, we simulate the distribution of scientific project choices in counterfactual scenarios. Of particular interest—given our focus on path dependence stemming from scientists' initial training and heterogeneity in scientist ability—are scenarios that increase the extent of early career exploration. In ongoing work, we examine how changes in scientists' initial stocks of knowledge and in the costs of exploration affect the extent of clustering across the space of potential topics.

The idea that highly-skewed distributions of scientific research across topics reflect suboptimal allocations of effort has been a topic of discussion among scientists in recent years. In the scientific literature, researchers have characterized these patterns as evidence of "research bias" and have argued that misaligned incentives in academic research may be causing science, in the aggregate, to miss important discoveries (Grueneberg et al. 2008; Baker 2009; Edwards et al. 2011; Stoeger et al. 2018; Mullard 2019; Kustatscher et al. 2022). Efforts by governments, non-profit organizations, and scientific consortia have attempted to "illuminate" these understudied research areas (Mullard 2019; Kustatscher et al. 2022). To our knowledge, however, we offer the first theoretical and empirical investigation of these claims, as well as the first examination of counterfactual scenarios that may be facilitated with small changes to the design of existing institutions. We offer microfoundations for various determinants of scientific project choice—including knowledge spillovers, path dependence, and scientist co-location—and introduce an empirical framework for the study of sequential, cumulative investment.

This work relates to literatures that examine the direction of innovative effort (e.g., Budish et al. 2015), the role of education in shaping innovation (e.g., Jones 2009; Biasi and Ma 2022), and the trajectories of scientific careers (e.g., Myers 2020; Hill and Stein 2023, 2024).[4] Our formulation of scientific human capital derives, in part, from the model of "combinatorial" new ideas in Jones (2021), our investigation of inefficiency is based, loosely, on the idea of technological lock-in described in David (1985), and our empirical investigation of agglomeration in "idea space" draws on a long literature that investigates agglomeration in physical space (e.g., Ellison et al. 2010; Moretti 2021).

More generally, in advanced economies, long-run productivity growth is understood as originating from investment in *new* ideas – a concept that forms the basis for modern endogenous growth theory (e.g., Romer 1990; Aghion and Howitt 1992). In this paper, we investigate a source of stymied productivity growth in the context of scientific research: overinvestment in old ideas.

---

research. Lotka's "law" states the following: if there are k scientists who publish one paper, then there are $\frac{k}{n^2}$ publishing n papers (Lotka 1926).

[3] We interpret this risk aversion parameter as capturing a mixture of the scientist's preference for uncertainty and what we term "clairvoyance" or "taste"—their ability to identify projects that are likely to be especially successful.

[4] Our survey of related literature here is intentionally brief.

References

Aghion, Philippe and Peter Howitt, "A Model of Growth Through Creative Destruction," *Econometrica*, 1992, *60* (2), 323–351. Publisher: [Wiley, Econometric Society].

Baker, Monya, "Human embryonic stem cell research stuck on two early lines," *Nature Reports Stem Cells*, August 2009, pp. 1–1. Publisher: Nature Publishing Group.

Biasi, Barbara and Song Ma, "The Education-Innovation Gap," March 2022.

Budish, Eric, Benjamin N. Roin, and Heidi Williams, "Do Firms Underinvest in Long-Term Research? Evidence from Cancer Clinical Trials," *American Economic Review*, July 2015, *105* (7), 2044–2085.

David, Paul A., "Clio and the Economics of QWERTY," *The American Economic Review*, 1985, *75* (2), 332–337. Publisher: American Economic Association.

Edwards, Aled M., Ruth Isserlin, Gary D. Bader, Stephen V. Frye, Timothy M. Willson, and Frank H. Yu, "Too many roads not taken," *Nature*, February 2011, *470* (7333), 163–165. Number: 7333 Publisher: Nature Publishing Group.

Ellison, Glenn, Edward L Glaeser, and William R Kerr, "What causes industry agglomeration? Evidence from coagglomeration patterns," *American Economic Review*, 2010, *100* (3), 1195–1213.

Grueneberg, Dorre A., Sebastien Degot, Joseph Pearlberg, Wenliang Li, Joan E. Davies, Amy Baldwin, Wilson Endege, John Doench, Jacqueline Sawyer, Yanhui Hu, Frederick Boyce, Jun Xian, Karl Munger, and Ed Harlow, "Kinase requirements in human cells: I. Comparing kinase requirements across various cell types," *Proceedings of the National Academy of Sciences*, October 2008, *105* (43), 16472–16477. Publisher: Proceedings of the National Academy of Sciences.

Hill, Ryan and Carolyn Stein, "Race to the bottom: Competition and quality in science," *Mimeo*, 2023.

_ and _, "Scooped! Estimating rewards for priority in science," *Journal of Political Economy*, 2024.

Jones, Benjamin F., "The Burden of Knowledge and the Death of the Renaissance Man: Is Innovation Getting Harder?," *Review of Economic Studies*, January 2009, *76* (1), 283–317.

Jones, Charles I., "Recipes and Economic Growth: A Combinatorial March Down an Exponential Tail," Working Paper 28340, National Bureau of Economic Research January 2021. Series: Working Paper Series.

Kustatscher, Georg, Tom Collins, Anne-Claude Gingras, Tiannan Guo, Henning Hermjakob, Trey Ideker, Kathryn S. Lilley, Emma Lundberg, Edward M. Marcotte, Markus Ralser, and Juri Rappsilber, "An open invitation to the Understudied Proteins Initiative," *Nature Biotechnology*, June 2022, *40* (6), 815–817. Number: 6 Publisher: Nature Publishing Group.

Lotka, A.J., "The frequency distribution of scientific productivity," *Journal of Washington Academy Sciences*, 1926, *16*, 317–323.

Moretti, Enrico, "The effect of high-tech clusters on the productivity of top inventors," *American Economic Review*, 2021, *111* (10), 3328–3375.

Mullard, Asher, "A probe for every protein," *Nature Reviews Drug Discovery*, September 2019, *18* (10), 733–736. Bandiera_abtest: a Cg_type: News Number: 10 Publisher: Nature Publishing Group.

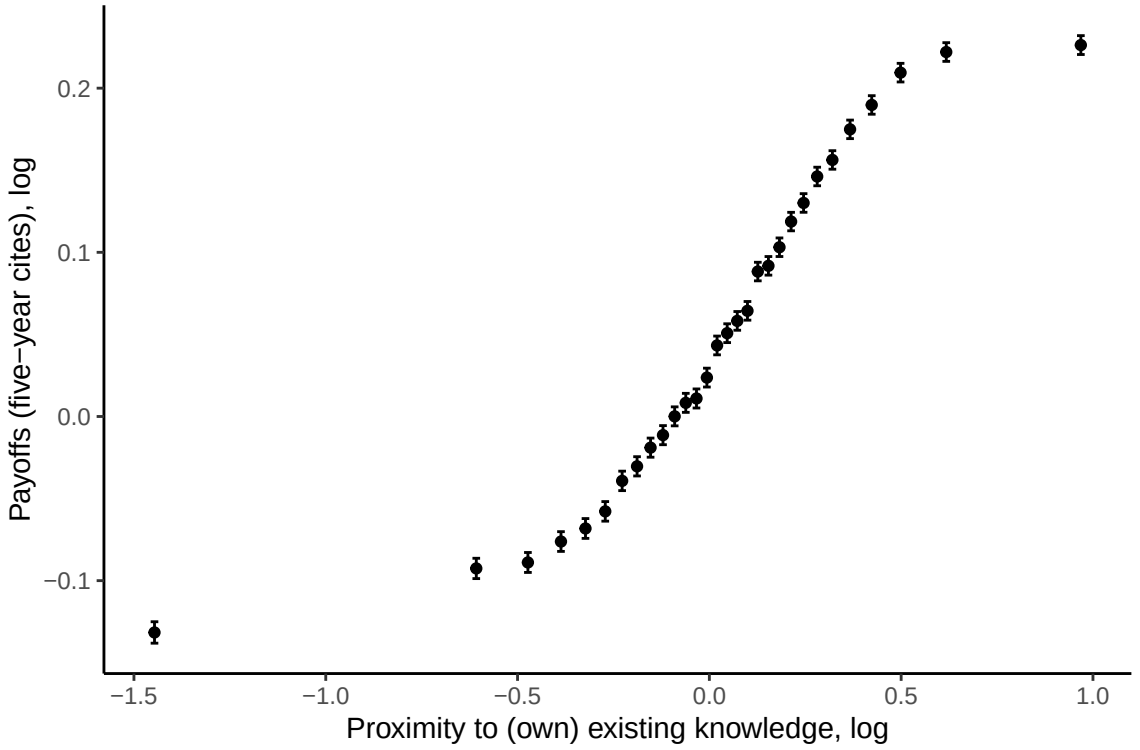
Myers, Kyle, "The Elasticity of Science," *American Economic Journal: Applied Economics*, October 2020, *12* (4), 103–134.

Romer, Paul M., "Endogenous Technological Change," *Journal of Political Economy*, October 1990, *98* (5, Part 2), S71–S102.

Scott, Christopher Thomas, Jennifer B. McCormick, and Jason Owen-Smith, "And then there were two: use of hESC lines," *Nature Biotechnology*, August 2009, *27* (8), 696–697. Number: 8 Publisher: Nature Publishing Group.

Stoeger, Thomas, Martin Gerlach, Richard I. Morimoto, and Luís A. Nunes Amaral, "Large-scale investigation of the reasons why potentially important genes are ignored," *PLOS Biology*, September 2018, *16* (9), e2006643.

Figure 1: Scientific Proximity as a Determinant of Investment



Note: We construct data on the project choices of 199,605 scientists. Each paper is represented as a vector in the roughly 30,000-dimensional space of scientific topics. In our preferred specifications, the scientist's stock of existing knowledge is defined, in this space, as the convex hull of the papers that the scientist has published to date. We measure the distance between a potential project and the scientist's stock of knowledge, then, by generating an approximate distance between the project and this convex hull measure. For each project, we collect a measure of realized citations. We apply a scaling procedure, which yields a measure of the citations that the project was expected to receive at the time it was selected. The **x-axis** plots the measure of proximity to expertise described above. The **y-axis** plots the measure of expected payoffs described above. Both are log-transformed, within-scientist, leave-one-out means. The log transformation accounts for heterogeneity in both measures across scientists. Within scientist leave-one-out-means facilitate interpretation. This figure is a binned scatterplot, where bins are determined as ventiles of the x-axis proximity measure.