Title

Improving data access democratizes and diversifies science

Journal

Proceedings of the National Academy of Sciences of the United States of America, 117(38)

ISSN

0027-8424

Authors

Nagaraj, Abhishek
Shears, Esther
de Vaan, Mathijs

Publication Date

2020-09-22

DOI

10.1073/pnas.2001682117

Peer reviewed

Improving data access democratizes and diversifies science

Abhishek Nagaraj[a,1], Esther Shears[b], and Mathijs de Vaan[a]

[a]Haas School of Business, University of California, Berkeley, CA 94720; and [b]Energy & Resources Group, University of California, Berkeley, CA 94720

Edited by Douglas S. Massey, Princeton University, Princeton, NJ, and approved July 28, 2020 (received for review January 30, 2020)

The foundation of the scientific method rests on access to data, and yet such access is often restricted or costly. We investigate how improved data access shifts the quantity, quality, and diversity of scientific research. We examine the impact of reductions in cost and sharing restrictions for satellite imagery data from NASA's Landsat program (the longest record of remote-sensing observations of the Earth) on academic science using a sample of about 24,000 Landsat publications by over 34,000 authors matched to almost 3,000 unique study locations. Analyses show that improved access had a substantial and positive effect on the quantity and quality of Landsat-enabled science. Improved data access also democratizes science by disproportionately helping scientists from the developing world and lower-ranked institutions to publish using Landsat data. This democratization in turn increases the geographic and topical diversity of Landsat-enabled research. Scientists who start using Landsat data after access is improved tend to focus on previously understudied regions close to their home location and introduce novel research topics. These findings suggest that policies that improve access to valuable scientific data may promote scientific progress, reduce inequality among scientists, and increase the diversity of scientific research.

data access | Landsat | inequality | diversity | science of science

How does improving access to data affect the rate and direction of scientific progress? Data are the lifeblood of modern empirical science and are used to both test and generate scientific theory. Yet, access to scientific data is often costly and difficult to obtain. Governments, private research institutions, and key individuals control access to critical data in fields as diverse as health (1), genomics and biology (2, 3), climate change (4, 5), ecology (6), astronomy (7), economics (8, 9), and meteorology (10). Many government and research organizations restrict access to their data and prevent data sharing, while others charge significant fees for data access in order to monetize this resource (11, 12). For example, the US government has recently considered whether to substantially increase fees for two widely used sources of remote-sensing imagery (13). Similar concerns are being raised about privately owned data. For example, in the ongoing crisis around COVID-19, commercial data on population mobility from cellphones is proving impactful (14), but access to such data remains largely restricted. In this paper, we study the effects of a steep decrease in the cost and sharing restrictions of satellite images collected via NASA's Landsat program on scientific research. Our evidence demonstrates that improving data access not only increases the quantity and quality of scientific research, it also democratizes and diversifies science.

Despite the salience of data access for scientific progress, research on the impact of limiting data access on the rate and direction of scientific inquiry is limited. Prior work that has looked at whether scientists who share their data are cited at higher rates finds mixed results (15, 16) and has also documented that data sharing among scientists is rare (17). Others have speculated that improved data access leads to "better science," but have not empirically examined this issue (18). In the context of satellite imagery (our focus), past work has provided some evidence that data costs affect the purchase of these data (19, 20) and that data access impacts firms relying on those data (21, 22). These studies, however, offer no insights on the effect of data access on the rate and direction of scientific progress.

While not focused on data, past research has looked at how scientific progress responds to improved access to other research inputs, especially in the life sciences. For example, intellectual property restrictions on genetic sequences decreased follow-on research and the development of genetic tests (23). Similarly, open access to biomaterials (24) increased their diffusion in follow-on research. More recent work has qualified these findings by showing that mere access might be insufficient to translate research inputs into publications; prior experience and resources could also be important (25). Whether and to what extent these results translate to fields outside of the life sciences and to the question of data access remains unknown.

Further, prior research has largely focused on the impact of improved access on overall levels of scientific output rather than on scientific inequality. The question of whether and how disadvantaged groups of scientists or less studied scientific topics benefit disproportionately as a result of improved data access remains underexplored. Important exceptions include recent work on the impact of open access to genetically engineered mice on the diversity of follow-on research (26) and work that links the impact of automation to the entry of outsiders in a field (27). While insightful, this research does not look at how open access may reduce inequality between scientists in environments that vary in terms of resources. Moreover, this work does not directly link the reduction in inequality to the diversification of science.

Significance

Data access is critical to empirical research, but past work on open access is largely restricted to the life sciences and has not directly analyzed the impact of data access restrictions. We analyze the impact of improved data access on the quantity, quality, and diversity of scientific research. We focus on the effects of a shift in the accessibility of satellite imagery data from Landsat, a NASA program that provides valuable remote-sensing data. Our results suggest that improved access to scientific data can lead to a large increase in the quantity and quality of scientific research. Further, better data access disproportionately enables the entry of scientists with fewer resources, and it promotes diversity of scientific research.

Author contributions: A.N., E.S., and M.d.V. designed research, performed research, analyzed data, and wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence may be addressed. Email: nagaraj@berkeley.edu.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2001682117/-/DCSupplemental.

First published September 8, 2020.

Overall, while "science of science" studies (28–30) suggest that access to research inputs shape science, further examination of the impact of improved data access on the quantity, quality, and diversity of scientific research is warranted.

In this study, we examine two main questions. First, we evaluate whether improved data access increases both the quantity and quality of science. Standard economic theory suggests that quantity should increase as a result of a reduction in data access restrictions. Better access should attract users with a lower willingness to pay, thereby expanding the pool of scientists who may exploit these data for scientific inquiry. With more researchers in the field, competition should also increase, boosting research quality (31). However, it is also possible that improved data access is accompanied by reductions in marketing and training efforts by the data provider, lowering awareness and reducing publications (32, 33). Further, even if quantity increases, it is possible that new projects are initiated by lower-quality researchers or on low-value projects, thereby lowering the quality of scientific output. Given contradictory theoretical possibilities, our quantitative examination sheds light on whether lowering data access restrictions increases or decreases the quantity and quality of science.

Our second question is whether improved data access democratizes science by enabling the entry of scientists with more limited resources and whether it diversifies the topical focus of scientific research. Inequality in scientific funding is substantial (34–36), and monetary barriers to data access may exacerbate these inequalities. Therefore, improving data access may democratize science by allowing researchers with smaller research budgets (like those in lower-ranked universities or in the developing world) to enter the field and publish alongside better endowed researchers. Further, under a nonlinear model of science (37) where similar data can be used for a variety of different applications, the entry of less endowed researchers may also translate into a more diverse set of topics and research questions (38). The pursuit of research is partly a function of personal interests and local context of the researcher which implies that a more varied set of researchers is likely to pursue previously unexplored research questions in previously underexplored areas and research topics. In our context, for example, the entry of a researcher from an underrepresented country (China) could lead to an impactful publication that uses Landsat to research an understudied place (Sichuan province) and an underexplored topic (*Oncomelania* or freshwater snail-driven infectious disease spread) (39). In our analyses, we therefore test whether and to what extent data access democratizes and diversifies science.

Setting and Data

We focus on scientific applications of a government-provided data source that experienced a dramatic shift in access restrictions. Specifically, we study NASA's Landsat program which was launched in 1972 and is the longest-running enterprise for acquisition of satellite imagery of Earth. While Landsat images were relatively affordable at first launch, the program was commercialized, and access to imagery was substantially more expensive for almost a decade between 1985 and 1995, before restrictions and costs of data access were reduced again. The Landsat collection of moderate-resolution images of Earth over time provides valuable data for researchers interested in studying environmental and demographic change in a variety of fields, including geology, forestry, agriculture, regional planning, and climate change. In 1985, the entire program along with all of its data was transferred from the US government to a private agency. During this time, costs of data access were relatively high as users were charged $4,400 per image and data sharing was prohibited. However, the high cost of data access was accompanied by a substantial marketing enterprise that was responsible for popularizing and commercializing the data.

In 1995, the program was transferred back to the US government, and image prices dropped to $2,500 per image—a 43% price reduction. Significantly, data sharing policies were relaxed, allowing for free transfer of data between scientists, further reducing costs of data access.* These changes meant that scientists purchasing data were facing much lower costs and, perhaps more importantly, could legally share data for free with other scientists who did not yet have access. The Landsat program's preeminent role in environmental and climate science, combined with the dramatic variation in the cost of access and sharing constraints, provides a unique opportunity to test how data access restrictions affect both the rate of scientific progress as well as its diversity.† In this paper, we will refer to the period between 1985 and 1995 as the commercial era and to the period after 1995 as the open era.

Our data come from two main sources. The first is Landsat coverage data from the start of the program which details when and where images were taken, the number of images, and the image quality of each of those images (based on percentage of the image covered by clouds) along with a number of other technical details. Each image captures a fixed "block" on the surface of the Earth, and the size of one block is roughly 115 miles in length and 115 miles in width (around 13,200 square miles of coverage).

The outcome variables in this study come from Scopus, Elsevier's "abstract and citation database of peer-reviewed literature."‡ The results of a search for "Landsat" (and some related terms), up to 2005, yield a dataset of academic publications using or referencing Landsat from 1975 to 2005, composed of roughly 24,000 publications by over 34,000 authors (see *SI Appendix* for more details on our sampling strategy). Note that this strategy is conservative—we are less likely to include research using other types of satellite data, but might miss Landsat science that refers to the data source as "satellite imagery" or uses other generic terms.§ These publication titles, abstracts, and author affiliations were geoparsed, where we first detected words that represented place names (such as the "Columbia Glacier") using machine-learning entity-detection algorithms and then geocoded these place names to obtain a latitude and longitude. This allows us to match places studied in a paper as well as author locations to specific blocks on the surface of the Earth corresponding to a Landsat image location. Our data also include information on the publication itself (title, year, authors, publication source, abstract, etc.) as well as other metrics available from Scopus such as number of citations and journal quality measures. In a set of additional analyses we compare trends in Landsat publications to trends in non-Landsat publications, and we use the same strategy to geoparse these non-Landsat publications.

The Landsat data are freely accessible, while the Scopus data are only accessible with a subscription. We have created an Open Science Framework repository that includes links to the freely accessible data and query statements to extract the Scopus data.

*To put this shift in costs into perspective, the average study in our data focuses on three geographical areas. Assuming that the study examines change, one would need at least six images. In the commercial era, such a study would have cost at least $26,400, while the price would drop to $15,000 after the program was transferred back to the US government. Moreover, these costs could be lowered further as a result of data sharing opportunities.

†Note that there were several changes to Landsat data distribution following the transition in 1995. Our main focus in this paper is on the changes following the 1992 Land Remote Sensing Policy Act (which then affected the Landsat program in 1995), but we do provide several estimates of the effect of other changes in *SI Appendix*.

‡See www.scopus.com.

§Note that there were no other sources of satellite imagery until the early 1990s, and these less important alternate sources are not included in our sample, so they should not bias our results.

The repository also includes the code used to generate the results (https://osf.io/mw34x/).

Results: Quantity and Quality of Science

We first present evidence that demonstrates the effect of the transition of Landsat data from the commercial to the open era. Fig. 1*A* shows the number of Landsat-related publications over time. Fig. 1*A* shows that while the number of publications was growing rapidly in the period before commercialization (pre-1985), this growth was halted in the commercial era. Once Landsat data access improves after 1995, there is a strong and immediate growth in the number of Landsat-related publications. As a comparison, the dotted line in Fig. 1*A* shows the total number of publications classified by Scopus as being in the "Earth and environmental science" category during this period. For this broader set, we do not see a trend break around 1995, suggesting that the patterns we document are not driven by concurrent changes in the scientific interest toward environmental topics or the advent of the world wide web, an assertion we rigorously test and describe in the next section. Fig. 1 *B* and *C* show how quality is impacted by the easing of access restrictions to Landsat images. While the number of highly cited papers and papers in top journals remained flat during the commercial era, the start of the open era coincided with stark increases in both the number of publications that garner over 100 citations and those that are published in a top journal (defined as those in the top 2% of journals by Scopus' CiteScore metric).

This descriptive analysis, while striking, is insufficient to fully establish the causal impact of access restrictions on science. Therefore, we complement this analysis by formally estimating the effect of the transition to the open Landsat era post-1995 in a regression framework. We present an identification strategy that effectively controls for a large number of alternative factors that could explain the patterns we describe and helps identify the causal role of data access restrictions in shaping scientific output. We exploit the fact that Landsat coverage at the block level was not uniform: technical errors and cloud cover in imagery caused wide variation in the amount of data available at the block level, even before Landsat data were commercialized. We argue that potential research on blocks with a greater amount of data should have been more affected by the privatization as compared to blocks that had fewer high-quality images.[¶] We consider the distribution of high-quality images in 1985, and we split the sample at the median into blocks with a higher level of coverage (treatment group) and those with a lower level of coverage (control group). In order for this comparison to be valid, it is important to check that above-median Landsat coverage areas are not likely to be those in which scientific exploration is more likely to occur. Our research design addresses this concern directly. Specifically, to control for any selection in terms of which blocks get better coverage, we control for the average number of publications in any given block (via block fixed effects) and examine whether treatment blocks have a greater increase in publications as compared to control blocks following the transition to the open era. If treatment blocks increase their publications more than control blocks, we can conclude that improved data access has a causal effect on scientific output. This framework is based on past research that has validated this approach (21).

Our estimates (*SI Appendix*, Table S1) from a difference-in-differences model with block and year fixed effects suggest that the number of published research articles at the block year increased by a factor of 3 (mean 0.15) as a result of improving access. Likewise, the number of highly cited publications increased by a factor of 6 (mean 0.0019), while the probability of any publication at the block year (mean 0.047) increased by about 50%. Note that these estimates indicate the relative increase in publications between treatment and control blocks and not the total global increase as indicated in Fig. 1.

Our baseline specification, while relatively robust, is vulnerable to two alternative explanations that could cloud the causal interpretation of our findings. First, we classify blocks into treatment and control groups based on the pre-1985 level of coverage. However, the Landsat project is constantly collecting new data, and if treatment blocks started receiving more data post-1995 as compared to control blocks, our estimates capture the effect of more data and not necessarily the effects of reduced costs of access. We collect information on the arrival of new images and show that this explanation cannot explain our findings (*SI Appendix*, Tables S8 and S9). Also, note that our research design relies on the control sample having the capacity to produce new science in the open era, an assumption that relies on a sufficient number of images being available. Accordingly, we present estimates limiting the control sample to only those blocks with five or more images and by comparing control blocks with above-median and above-90th percentile blocks in terms of image coverage pre-1985. These estimates (*SI Appendix*, Table S7) show that both exercises produce findings similar to our baseline estimates.

Second, as shown in Fig. 1*A*, global publications are increasing during the 1990s, especially in China and other countries around the world with previously limited participation in science. To make sure that our results are unaffected by these trends, we first provide estimates excluding Chinese blocks and show that our results are robust to their exclusion (*SI Appendix*, Table S6). We then conducted another analysis to account for global trends in publications. Rather than comparing Landsat publications in treatment and control blocks, we compare Landsat publications to a sample of over 50,000 geoparsed publications in the Earth and environmental sciences as identified through Scopus. Specifically, we compare the evolution of Landsat and non-Landsat publications at the block year level before and after 1995 (as shown in *SI Appendix*, Fig. S10). The regression estimates (*SI Appendix*, Table S10) indicate that even when using this completely different sample, Landsat publications increase disproportionately as compared to Earth and environmental sciences publications, indicating that our baseline results are not contaminated by an overall increase in scientific focus on certain blocks around the world.

Finally, in *SI Appendix* we included several additional analyses to show the robustness of our results. For example, in *SI Appendix*, Figs. S5 and S9 and Table S5, we show that it is unlikely that the results from our main research design are driven by unobserved differences in treatment and control blocks or by the overrepresentation of blocks in the United States. We also address the concern that our treatment effect is picking up on changes in data access that succeeded the 1995 change. In *SI Appendix*, Table S2, we show that while the 1995 change has a significant effect, later changes (in 1999 and 2001) matter as well, providing robustness for our main proposition that access costs have a meaningful effect on science.

Results: Democratization of Author Base

Improved data access is unlikely to benefit scientists equally. Specifically, scientists who are endowed with extensive financial resources are less likely to benefit from a transition to open data compared to less endowed scientists (35). Fig. 2*A* presents a map showing the locations of authors who use Landsat data in a scientific publication. A lighter, gray dot indicates locations with at least one researcher publishing a paper in the period from 1985 to 1995, i.e., when data access was costly and with limited sharing restrictions. A dark, black dot indicates

[¶]Although multiple number of images for the same block might seem redundant, typically, they are not. One feature that makes Landsat data valuable is the fact that it allows scientists to study change, such as urbanization or deforestation.

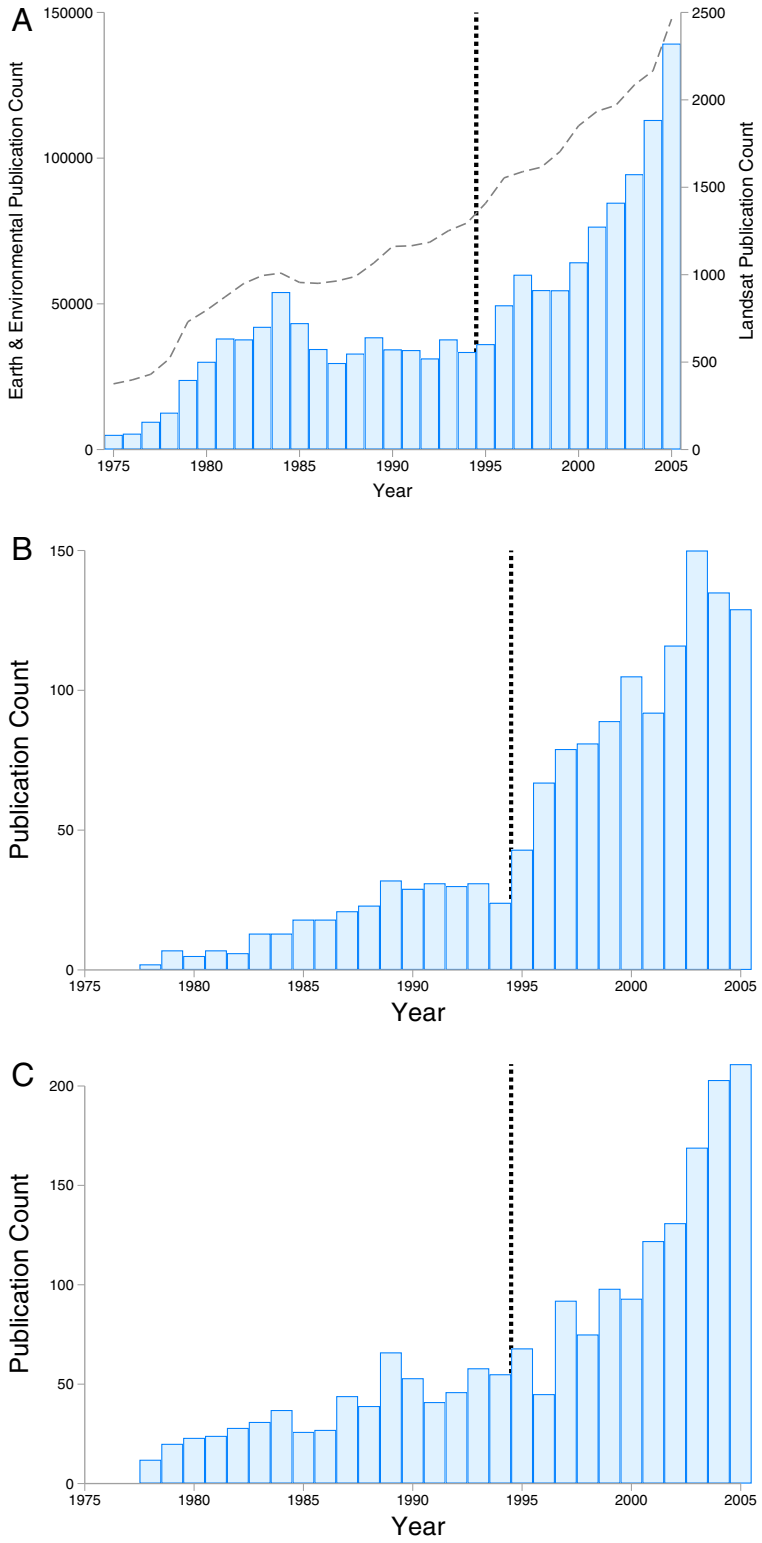

Fig. 1. Landsat-related publications before, during, and after the Landsat commercialization era. This figure shows the number of Landsat publications over time for three different types of publications. In all three panels, the bars in blue to the right of the vertical dashed line indicate publications after the Landsat program was transferred back to the US government. (*A*) All publications, (*B*) publications with 100 or more cites as of 2017, and (*C*) all publications in about 80 journals that represent the top two percentiles of journals ranked by citation score metrics. In *A*, the dashed line shows the general trend for all earth and environmental science publications. In all three panels, note that trends in the number of publications are mostly steady during the commercial era, after which there is a rapid increase in publications in the open data era.

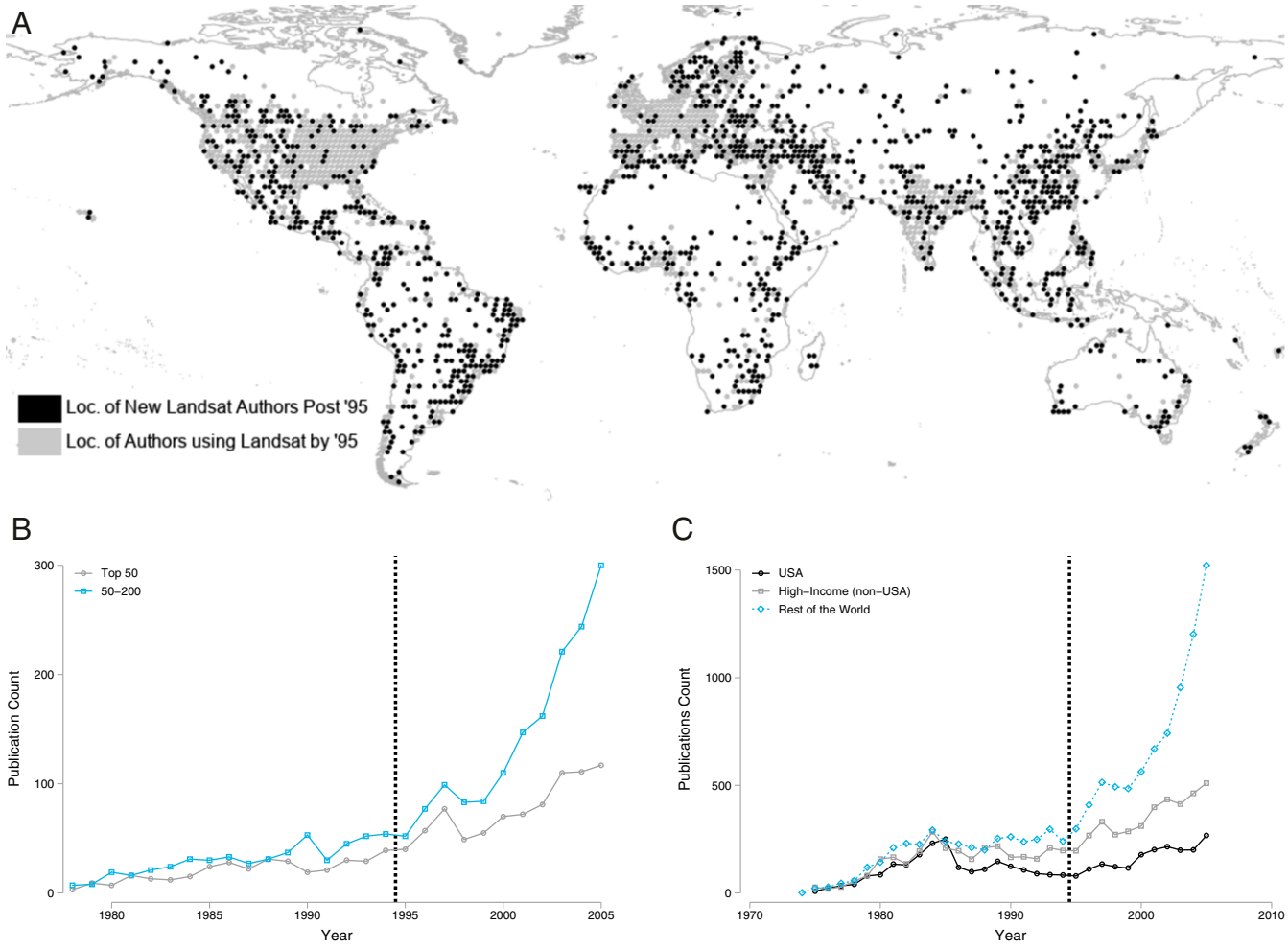

Fig. 2. How data access affects who participates in Landsat research. This figure explores the effects of lowering costs of data access on authors' locations. (*A*) A map where each light gray dot represents the presence of at least one author institution that has published a paper using Landsat data before data access costs were reduced. The dots in black represent locations where an author institution published a paper using Landsat data only after data access costs were reduced. A graph depicting this change is found in *SI Appendix*, Fig. S6. (*B*) Total number of Landsat publications separated by institutional rank (top 50 vs. 50 to 200) as per the Quacquarelli Symonds (QS) World top university rankings. (*C*) Total number of publications separated by the authors' country income categories. For publications with authors from different country income groups, we sort the publication based on the minimum country income group. Overall, the data suggest that lowering costs of data access was particularly helpful for authors in lower-ranked institutions and in non–high-income countries.

locations with researchers who started publishing Landsat research only after data access restrictions were reduced. The locations with black dots therefore represent new author locations, potentially enabled by the reduced cost of access to Landsat data after 1995. This map shows that while many authors in the United States and Western Europe were already leveraging Landsat data when access restrictions were high, many researchers from regions such as South America, Africa, Eastern Europe, the Middle East, and China started exploiting Landsat information only when access restrictions were reduced. A graphical depiction of this change is in *SI Appendix*, Fig. S6.

This pattern, where the proportion of authors from less developed regions and scientific institutions with lower endowments benefit from lowering the costs of data access, can be clearly seen in Fig. 2 *B* and *C*. Fig. 2*B* charts the number of publications from authors in top 50[#] ranked institutions (in gray) as compared to those from institutions ranked 50 to 200 (in blue), while Fig. 2*C* shows the number of publications by income level of the authors' country. As is clear from Fig. 2 *B* and *C*, growth in number of publications is mostly driven by scientists in contexts with fewer resources. In *SI Appendix*, Table S3, we quantitatively examine the differential impact of lowering data access restrictions for authors in lower-ranked institutions and those from lower-income regions. Overall, these estimates suggest a statistically significant difference between the increase in total publications for authors from lower-ranked institutions and lower-income countries.

Results: Diversity of Scientific Focus

Having shown that the open era democratized Landsat research by allowing the entry of new authors, we now turn to examining the question of whether this change also resulted in increased diversity in scientific focus. Since the types of research questions studied by scientists are likely to be influenced by their local contexts, democratizing who participates in science might diversify science itself. We consider two approaches to measuring this diversity: the geographic focus of the study and the research

[#]We classified every publication as belonging to a top 50 institution if at least one author was affiliated with an institution in the top 50 universities in the world according to QS World university rankings.

topic as captured by the words used in the abstract of a paper (37). Specifically, we explore whether improved data access facilitated research on previously unexplored 1) study locations and 2) topics as indicated by words used in abstracts.

Geographic Focus. We first examine the impact of improved data access on the geographic focus of the research. Analogous to the map we presented for authors, Fig. 3*A* presents a map that demonstrates the change in the locations examined using Landsat data. The dots in gray represent locations that had already been studied by 1995, while the dots in black represent new locations that were studied for the first time after data access had improved. The map shows that after data access improved, new study locations emerged mainly in middle- and low-income regions of the world. To show this pattern more directly, Fig. 3*B* plots the cumulative number of unique study locations in the United States, in other high-income countries, and in the rest of the world. As is clear from Fig. 3*B*, improving data access is associated with an increase in study locations, especially in lower-income countries. Simple regression versions of Fig. 3*B* described in *SI Appendix*, Table S4, confirm that these differences are statistically significant.[‖] *SI Appendix*, Fig. S7, explores these patterns further and shows the increase in the number of unique study locations in a given year and the number of first-time locations by country income.

We have shown that improved data access led to the entry of new scientists as well as a focus on new study locations, but it is not clear whether the two patterns are related. We therefore conducted additional analyses. First, we split the sample of publications in the open era into those with at least one author who had used Landsat data during the commercial era (incumbents) and those without any authors who had previously used the data (newcomers). We then calculate whether new study locations were introduced mostly in newcomer or incumbent publications. We find that newcomer publications introduce 3,982 new study locations, while incumbents introduce 1,965 new locations. The difference is partly driven by publication volume, but even if one adjusts for this difference, newcomer publications are 15% more likely to introduce a new study location.

Since new locations may have been studied by incumbent authors in the absence of newcomers, our next analysis aims to provide an estimate of how much incumbent authors would have to expand their horizon to cover the new study locations introduced during the open era. To calculate this estimate, we assign incumbents to new study locations, measure their distance from these study locations, and then compare this counterfactual distribution of distances to the realized distribution of distances between the actual authors (i.e., newcomers) and the new study locations.[**] Fig. 3*C* shows the distribution of actual and counterfactual distances between authors and first-time study locations in the open era for non-US study locations. As is clear from this chart, the observed distances between authors and study locations are significantly lower than the counterfactual distances. In fact, the average observed distance is 3,196 km, while the average counterfactual distance is 5,799 km (t = 9.2963), a difference of over 2,500 km. These patterns hold when considering study locations within the United States, but the differences are less pronounced. In *SI Appendix*, we present more details on this analysis as well the full distribution of distances that includes both US and non-US study locations. Overall, this result suggests that newly entering scientists played a prominent role in expanding the geographic focus of Landsat research in the open era.

Topical Focus. While it is clear from Fig. 3 that the democratization of the author base diversified the geographic focus of Landsat science, we also investigate the extent to which the topical focus in the literature expanded. If new scientists are more likely to be from different parts of the world and have a variety of different research interests, it is possible that they use Landsat data to examine previously unexplored topics. To reprise the example we used before, a Chinese researcher using Landsat is not only more likely to study a region in China, he or she is also more likely to use it to focus on questions of relevance to the local context: infectious disease spread from a local freshwater snail (39). Western scientists in the past might have ignored this topic.

Our analysis is based on the text in the abstracts of publications using Landsat data. We first preprocessed the data by removing stop words, punctuation, and other textual information in the abstract field that is not part of the abstract (e.g., publisher information). We then tokenized the abstract by identifying the unique words used in those abstracts. These words serve as indicators of its topical focus and will form the basis of our textual analysis. Fig. 4*A* plots the introduction of these novel words in our data by calendar year. The graph shows that while the introduction of novel words was decreasing when data sharing restrictions were in place, there is a large increase in the number of unique words in the literature after 1995. This trend is suggestive evidence of an expansion in scientific focus toward a more diverse set of topics and fields.

Next we examine the contribution of newcomer scientists to this growth in the diversity of topics post-1995. We leverage the set of incumbent and newcomer authors and examine whether there are differences in the topics studied by both groups. As a first step, we simply compared words exclusively used by newcomers and words exclusively used by incumbents in the open era. We find that newcomers used 26,632 words that had not yet been used in the commercial era, while incumbents used 13,348 novel words. This gap is partly driven by the larger number of newcomer publications, but even when we consider the average number of new words per publication, newcomers use 38% more novel words per paper than incumbents (2.49 versus 1.73 per publication).

While the data do suggest that newcomers introduce more novel words than incumbent scientists, it is not obvious that these words represent meaningful new research topics. To address this concern, we measure the semantic relationships between newly introduced words and examine the internal consistency of those words. We use word embedding models (40) to examine the vectors of words introduced by newcomers and incumbents. Word embeddings are locations in a multidimensional space that can be used to measure symantic relationships between words. For each word, we identified the five words[††] closest in embedding space and computed the average distance between them. For example, if we observe the term "tree", our method classifies it as being more related in word-embedding space to "forest" than another word like "glacier." The computed average distance is a measure of how related a newly introduced word is to other newly introduced words. We log-transformed this measure to produce a relatedness index, where a larger number represents a word that is more internally consistent and is more likely to be part of a broader topical discussion. We plot the distribution of this index separately for the vector of new words introduced by newcomers and incumbents in Fig. 4*B*. The graph clearly shows that the distribution of words introduced by newcomers is shifted

[‖] Note that these estimates do not adopt the quasi-experimental research design like in Fig. 1 and represent descriptive (rather than causal) estimates of the impact of data access restrictions on diversity.

[**] The method we used to assign incumbents to new study locations is detailed in *SI Appendix*.

[††] Results are robust to different cutoffs: 10, 20, and 50 words.

SOCIAL SCIENCES

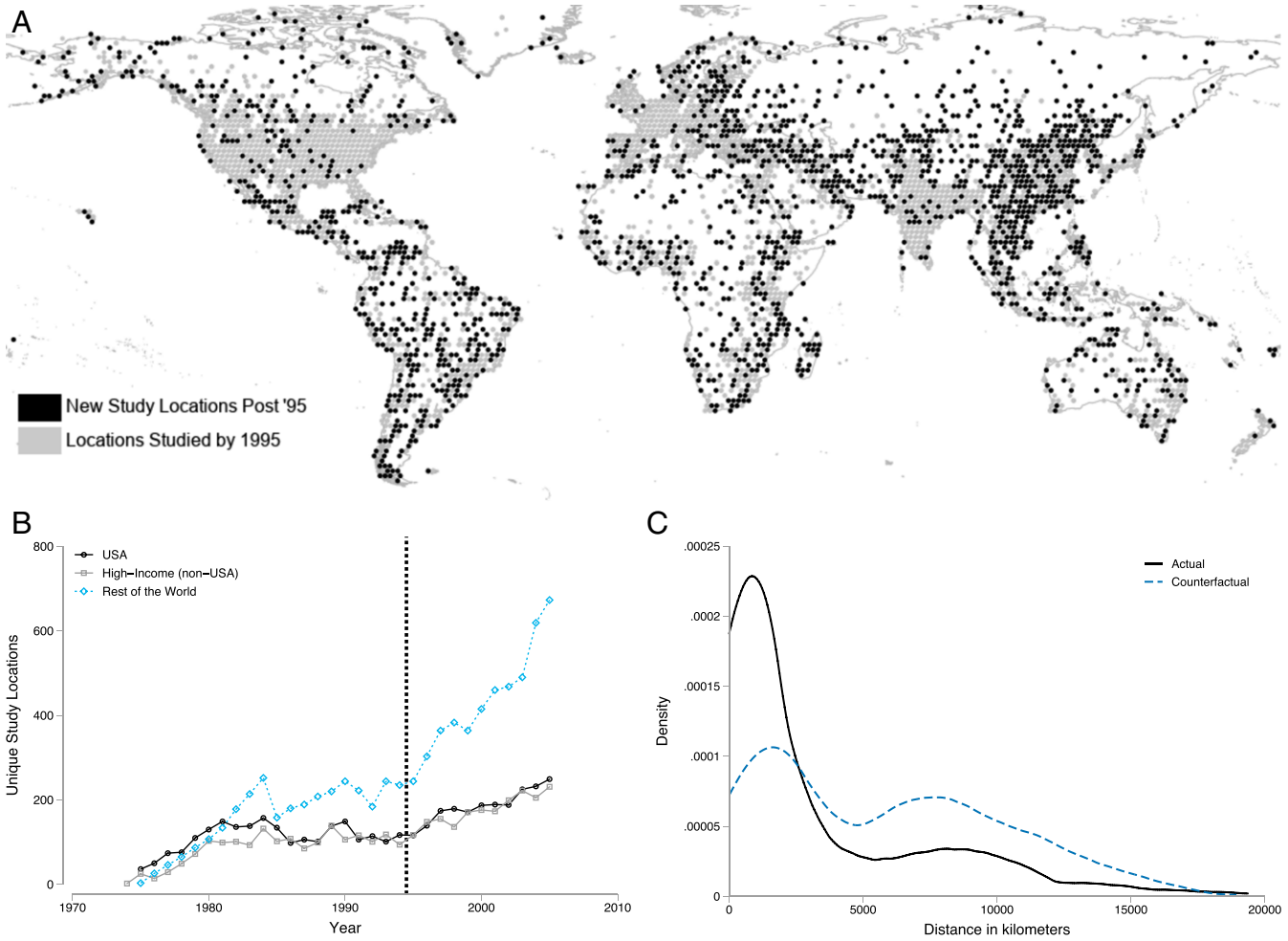

Fig. 3. How data access affects study locations. This figure explores the effects of lowering costs of data access on study locations. (*A*) A map where each light gray dot represents at least one Landsat publication that studies the region before data access costs were reduced. The dots in black represent at least one Landsat publication that studies the region only after data access costs were reduced. (*B*) Total number of unique locations studied by Landsat publications separated by country income groups. (*C*) The distribution of distances between authors and study locations for all study locations that had not been explored in the commercial era and are located outside the United States. A version of *C* that includes US locations is in *SI Appendix*, Fig. S8. Overall, these findings suggest that easing data access restrictions particularly helped increase the number and range of study locations.

to the right. Therefore, newcomers not only introduce more new words to the literature, but these words are also more internally consistent, suggesting that they may capture a new topic or sets of topics. One example of the set of internally consistent terms that are introduced by newcomers includes *Oncomelania* (the genus of freshwater snail discussed before) along with related terms such as infection, transmission, snail, and schistosomiasis (a type of infectious disease).

Finally, if new authors introduce new topics, we should also expect them to publish their work in a wider set of academic journals. Compared to the set of 982 unique journals in the commercial era, there were 486 new journals that published work by incumbent authors and 1,275 new journals that published work by the new authors in open era.[‡‡]

Overall, the results from Figs. 3 and 4 are clear: not only did the opening of Landsat data lead to the entry of a more diverse author base, but these newcomers also diversified the scientific discourse itself.

Conclusion

This study examines the role of data access on science. When data access barriers are relaxed, it is much more likely to be exploited by scientists, leading to a greater quantity and quality of scientific output. Further, ease of data access democratizes science by allowing authors with fewer financial resources to participate in the scientific process. This process of democratization also increases the diversity of scientific research itself.

Our results come from a comparison of high- and low-coverage areas for a single dataset in the area of Earth and environmental science research. Future work could generalize these findings by comparing across multiple datasets and research fields with varying levels of data access costs. Despite our results coming from a single case study, we believe that they may generalize and be relevant to other fields where data access is important. As stated in the introduction, the question of data access is central to virtually every scientific field that relies on empirical measurement. In each of these fields, the scientific labor force is divided into a few elites, who have access to resources and are able to leverage them to access data, while others must rely on poorer quality data or engage in primary data collection. As scientific norms change with many journals now requiring researchers to make their

[‡‡] Journal field in our Scopus publications data includes various document types (journal articles, books, conference proceedings, and editorials). We did not restrict to only journals and treated different years of a conference as a different unique journal.

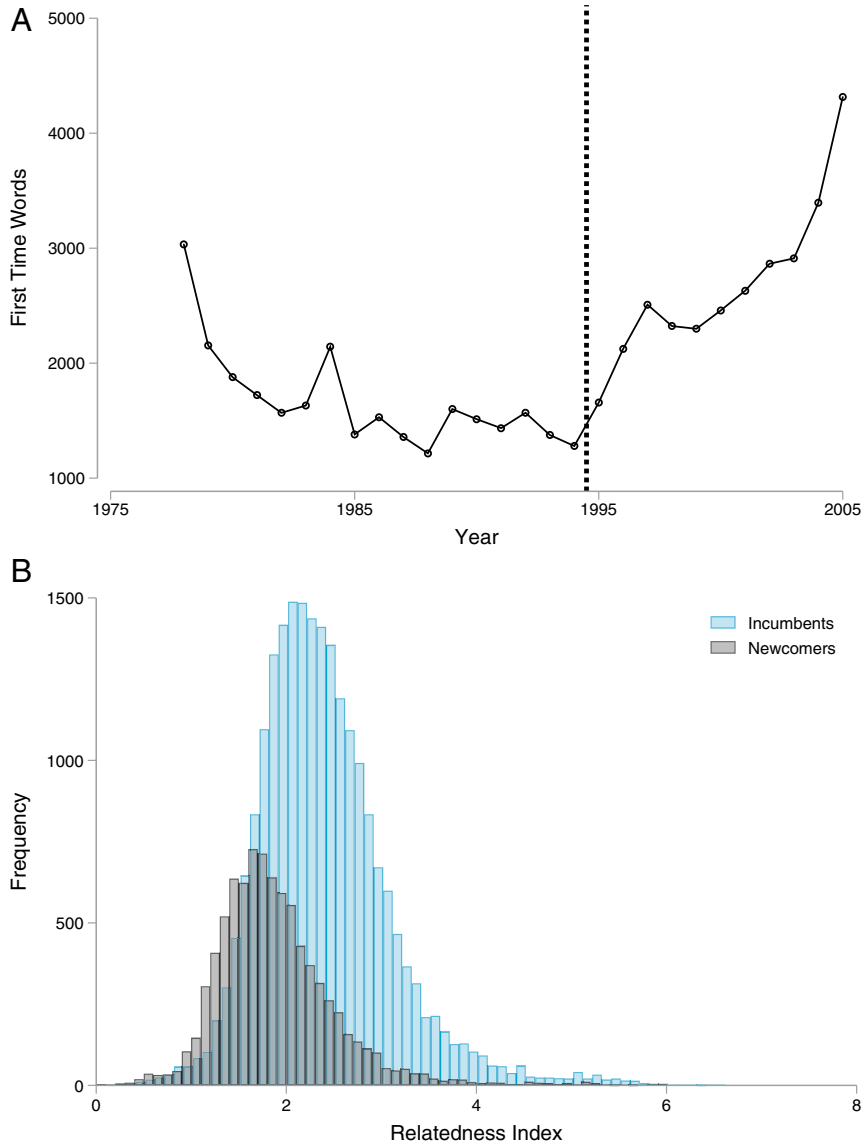

Fig. 4. Topical diversity in Landsat science. (*A*) The total number of first-time abstract words used in Landsat publications. (*B*) The distribution of the relatedness index by incumbent (dark) versus newcomer (light) authors. The higher the value of the index, the more related a focal word is to other newly introduced words. The distribution of newcomer words is clearly shifted to the right, which implies that new words introduced by newcomers are more likely to be related to other words introduced by newcomers (compared to new words introduced by incumbents).

data available and many funding agencies (in particular, NIH and NSF) requiring data from funded projects be made available, many fields are seeing an abundance of data being made available to a wider set of researchers. Our research suggests that not only will such improvements in data access affect the distribution of scientific credit across a wider and more diverse pool of researchers, they could also shift the topical focus of scientific research toward a broader set of research questions.

Ultimately, data are the life blood of scientific research. While recouping the cost of data generation and maintenance might sometimes be necessary, our research suggests that policies to restrict access to important data sources should consider the costs of such measures on the quantity, quality, and diversity of science before they are implemented.

Data Availability. All data and code required to generate the results are publicly accessible and have been deposited in the Open Science Framework (https://osf.io/mw34x/).

ACKNOWLEDGMENTS. We are grateful to Sameer Srivastava and seminar participants at the Academy of Management and participants of the 19th Annual Roundtable for Engineering Entrepreneurship Research for their advice. We also thank Sukwoong Choi and Sahiba Chopra for excellent research assistance.

1. J. T. Wilbanks, E. J. Topol, Stop the privatization of health data. *Nature* **535**, 345–348 (2016).
2. J. Kaye, C. Heeney, N. Hawkins, J. De Vries, P. Boddington, Data sharing in genomics—Re-shaping scientific practice. *Nat. Rev. Genet.* **10**, 331–335 (2009).
3. V. Marx, Biology: The big challenges of big data. *Nature* **498**, 255–260 (2013).
4. M. A. Wulder, N. C. Coops, Satellites: Make Earth observations open access. *Nature* **513**, 30–31 (2014).
5. J. T. Overpeck, G. A. Meehl, S. Bony, D. R. Easterling, Climate data challenges in the 21st century. *Science* **331**, 700–702 (2011).
6. O. J. Reichman, M. B. Jones, M. P. Schildhauer, Challenges and opportunities of open data in ecology. *Science* **331**, 703–705 (2011).

7. K. N. Abazajian *et al.*, The seventh data release of the Sloan Digital Sky Survey. *Astrophys. J. Suppl.* **182**, 543–558 (2009).
8. D. Card, R. Chetty, M. S. Feldstein, E. Saez, "Expanding access to administrative data for research in the United States" in *Ten Years and Beyond: Economists Answer NSF's Call for Long-Term Research Agendas*, C. L. Schultze, D. H. Newlon, Eds. (American Economic Association, 2010), pp. 81–84.
9. R. Hill, C. Stein, H. Williams, Internalizing externalities: Designing effective data policies. *AEA Pap. Proc.* **110**, 49–54 (2020).
10. Private weather data should not replace basic research. *Nature* **542**, 5–6 (2017).
11. E. G. Campbell, E. Bendavid, Data-sharing and data-withholding in genetics and the life sciences: Results of a national survey of technology transfer officers. *J. Health Care Law Policy* **6**, 241–255 (2002).
12. G. King, Ensuring the data-rich future of the social sciences. *Science* **331**, 719–721 (2011).
13. G. Popkin, US government considers charging for popular Earth-observing data. *Nature* **556**, 417–418 (2018).
14. D. Holtz *et al.*, Interdependence and the cost of uncoordinated responses to COVID-19. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 19837–19843 (2020).
15. H. A. Piwowar, R. S. Day, D. B. Fridsma, Sharing detailed research data is associated with increased citation rate. *PLoS One* **2**, e308 (2007).
16. M. J. McCabe, F. Mueller-Langer, Does data disclosure increase citations? Empirical evidence from a natural experiment in leading economics journals (2019). https://ssrn.com/abstract=3329272. Accessed 15 August 2020.
17. C. L. Borgman, The conundrum of sharing research data. *J. Am. Soc. Inf. Sci. Technol.* **63**, 1059–1078 (2012).
18. J. C. Molloy, The Open Knowledge Foundation: Open data means better science. *PLoS Biol.* **9**, e1001195 (2011).
19. M. A. Wulder, J. G. Masek, W. B. Cohen, T. R. Loveland, C. E. Woodcock, Opening the archive: How free data has enabled the science and monitoring promise of Landsat. *Remote Sens. Environ.* **122**, 2–10 (2012).
20. M. Borowitz, Government data, commercial cloud: Will public access suffer?*Science* **363**, 588–589 (2019).
21. A. Nagaraj, The private impact of public data—Landsat satellite maps and gold exploration (2020).
22. A. Nagaraj, S. Stern, The economics of maps. *J. Econ. Perspect.* **34**, 196–221 (2020).
23. H. L. Williams, Intellectual property rights and innovation: Evidence from the human genome. *J. Polit. Econ.* **121**, 1–27 (2013).
24. J. L. Furman, S. Stern, Climbing atop the shoulders of giants: The impact of institutions on cumulative research. *Am. Econ. Rev.* **101**, 1933–1963 (2011).
25. S. Zyontz, N. Thompson, "Who tries (and who succeeds) in staying at the forefront of science: Evidence from crispr" in *Academy of Management Proceedings* (Academy of Management, Briarcliff Manor, NY), vol. 2018, p. 15258 (2018).
26. F. Murray, P. Aghion, M. Dewatripont, J. Kolev, S. Stern, Of mice and academics: Examining the effect of openness on innovation. *Am. Econ. J. Econ. Pol.* **8**, 212–252 (2016).
27. J. L. Furman, F. Teodoridis, Automation, research technology, and researchers' trajectories: Evidence from computer science and electrical engineering. *Organ. Sci.* **31**, 330–354 (2020).
28. P. Azoulay *et al.*, Toward a more scientific science. *Science* **361**, 1194–1197 (2018).
29. R. Sinatra, D. Wang, P. Deville, C. Song, A. L. Barabási, Quantifying the evolution of individual scientific impact. *Science* **354**, aaf5239 (2016).
30. S. Fortunato *et al.*, Science of science. *Science* **359**, eaao0185 (2018).
31. P. Aghion, C. Harris, P. Howitt, J. Vickers, Competition, imitation and growth with step-by-step innovation. *Rev. Econ. Stud.* **68**, 467–492 (2001).
32. J. West, S. Gallagher, Challenges of open innovation: The paradox of firm investment in open-source software. *R&D Manag.* **36**, 319–331 (2006).
33. E. W. Kitch, The nature and function of the patent system. *J. Law Econ.* **20**, 265–290 (1977).
34. R. K. Merton, The Matthew effect in science: The reward and communication systems of science are considered. *Science* **159**, 56–63 (1968).
35. T. Bol, M. d. Vaan, A. v. d. Rijt, The Matthew effect in science funding. *Proc. Natl. Acad. Sci. U.S.A.* **115**, 4887–4890 (2018).
36. P. Azoulay, T. Stuart, Y. Wang, Matthew: Effect or fable?*Manag. Sci.* **60**, 92–109 (2014).
37. P. Aghion, M. Dewatripont, J. C. Stein, Academic freedom, private-sector focus, and the process of innovation. *Rand J. Econ.* **39**, 617–635 (2008).
38. J. Volmink, L. Dare, Addressing inequalities in research capacity in Africa. *BMJ* **331**, 705–706 (2005).
39. E. Seto *et al.*, The use of remote sensing for predictive modeling of schistosomiasis in China. *Photogramm. Eng. Rem. Sens.* **68**, 167–174 (2002).
40. V. Tshitoyan *et al.*, Unsupervised word embeddings capture latent knowledge from materials science literature. *Nature* **571**, 95–98 (2019).