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.

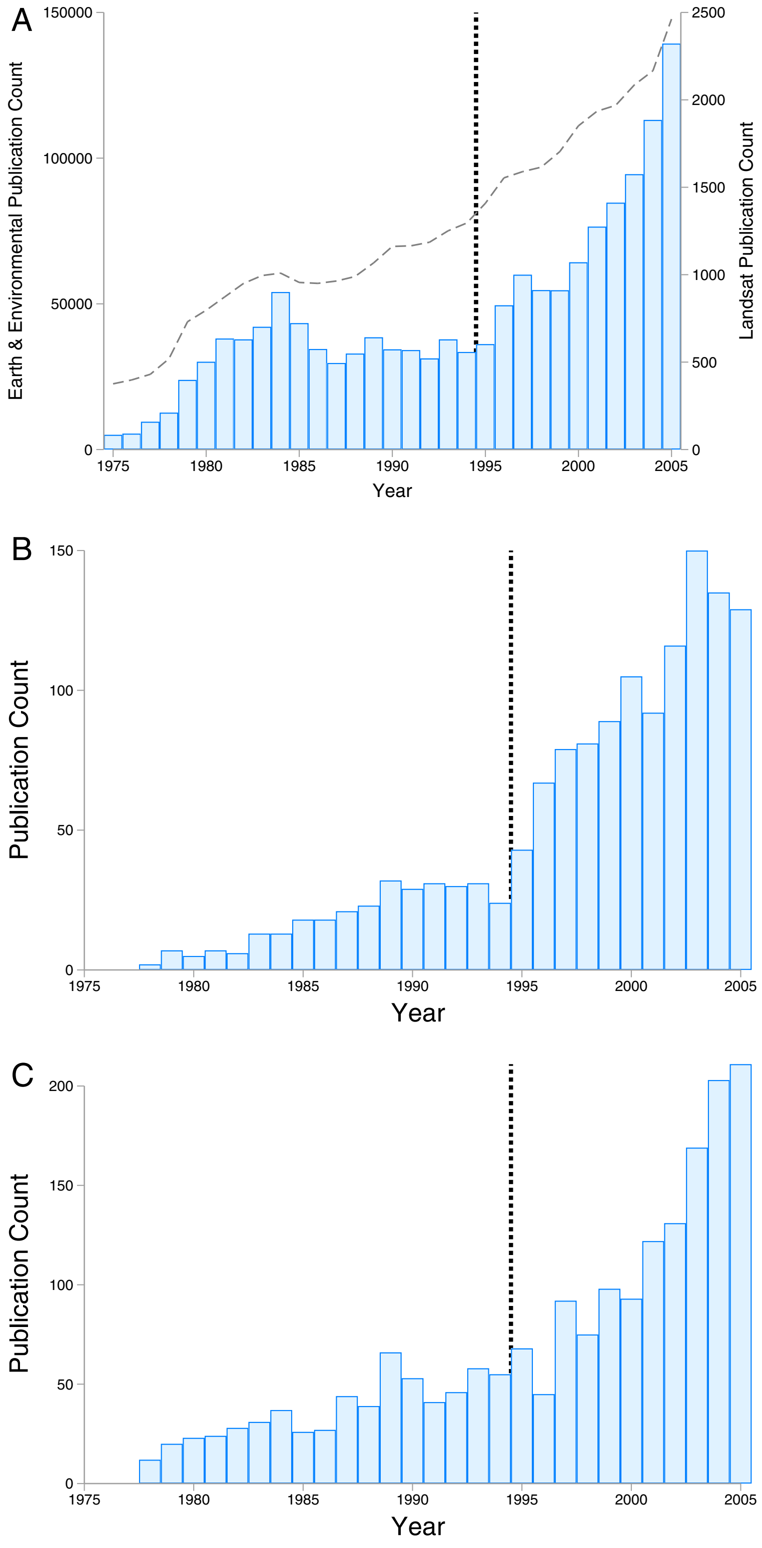


Fig. 1. Landsat-related publications before, during, and after the Landsat commercialization era. This figure shows the number of Landsat publications over time for three different types of publications. In all three panels, the bars in blue to the right of the vertical dashed line indicate publications 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.

SOCIAL SCIENCES

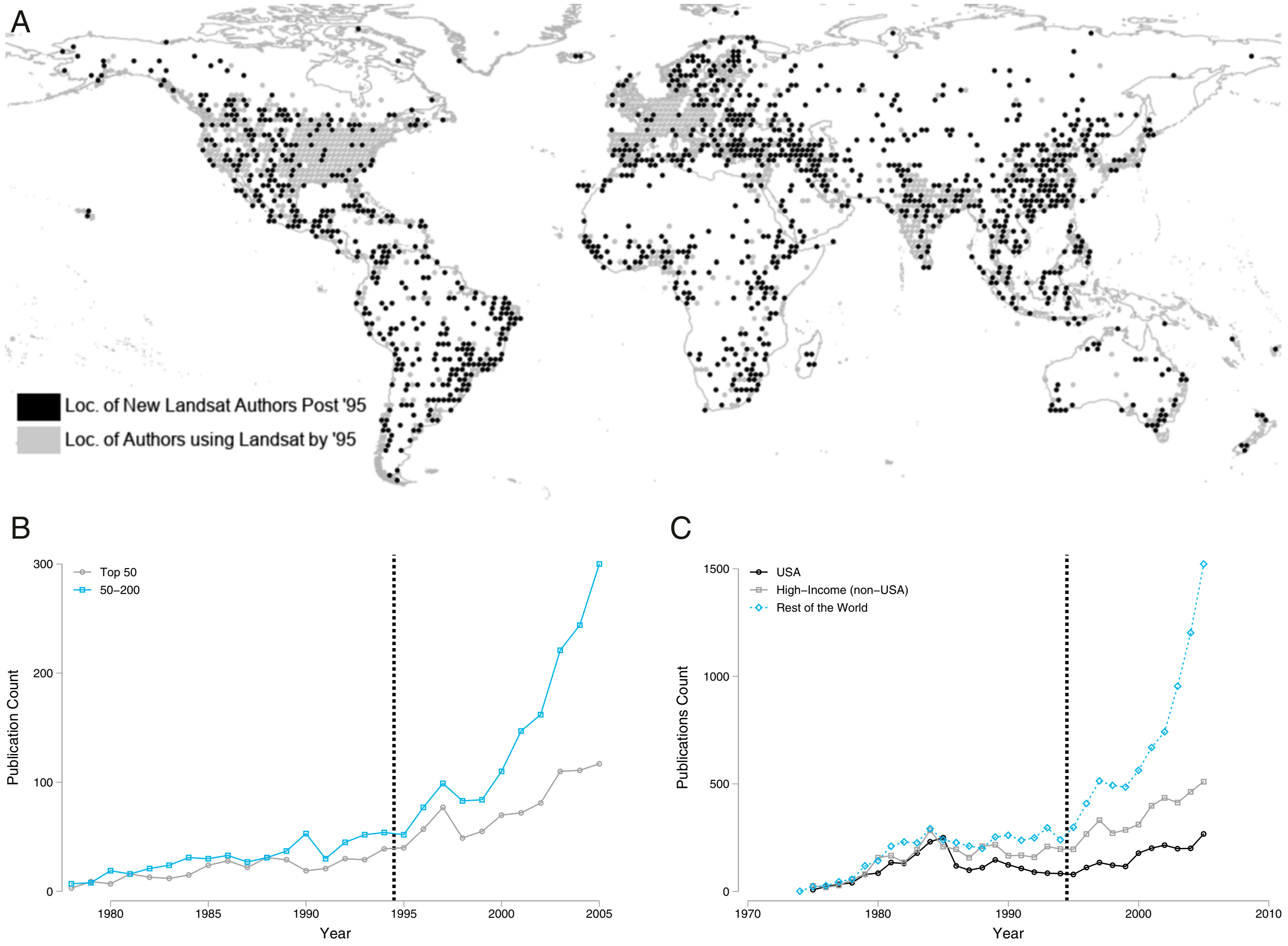


Fig. 2. How data access affects who participates in Landsat research. This figure explores the effects of lowering costs of data access on authors' locations. (*A*) A map where each light gray dot represents the presence of at least one author institution that has published a paper using Landsat data 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.

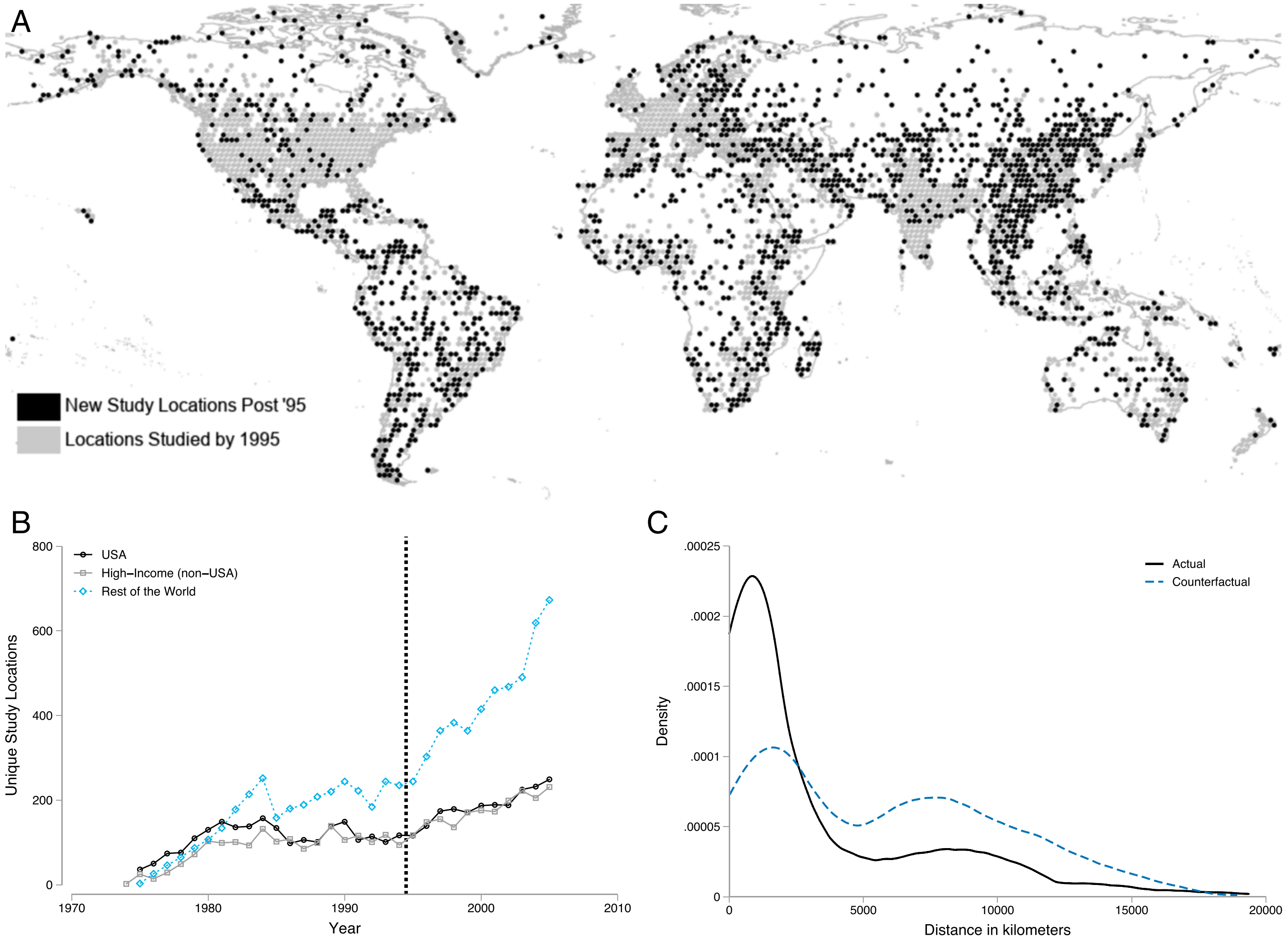


Fig. 3. How data access affects study locations. This figure explores the effects of lowering costs of data access on study locations. (*A*) A map where each light gray dot represents at least one Landsat publication that studies the region before data access costs were reduced. The dots in black represent 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.

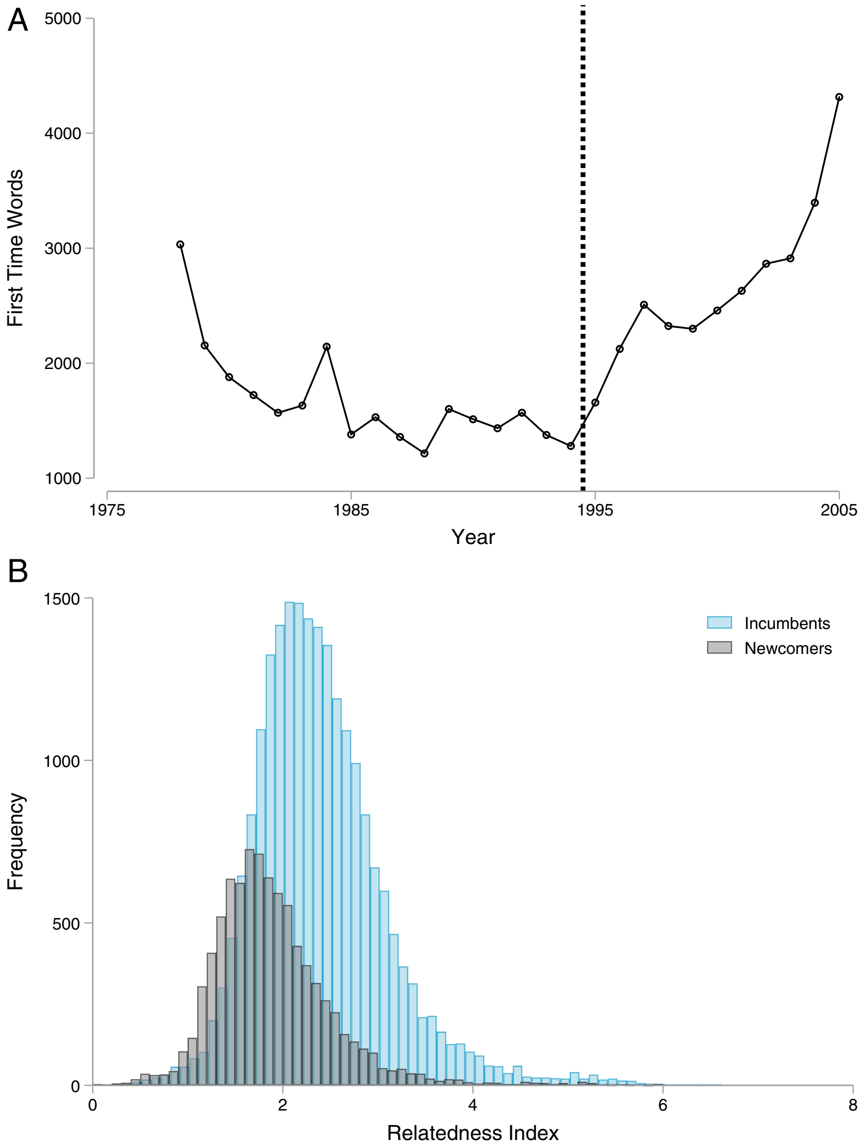


Fig. 4. Topical diversity in Landsat science. (*A*) The total number of first-time abstract words used in Landsat publications. (*B*) The distribution of the relatedness index by incumbent (dark) versus newcomer (light) authors. The higher the value of the index, the more related a focal word is to other 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).

Supporting Information (SI) Appendix: Improving Data Access Democratizes and Diversifies Science.

Abhishek Nagaraj[*], Esther Shears, and Mathijs de Vaan

University of California, Berkeley

[*] Corresponding Author: Please email nagaraj@berkeley.edu for any comments or suggestions.

www.pnas.org/cgi/doi/10.1073/pnas.2001682117

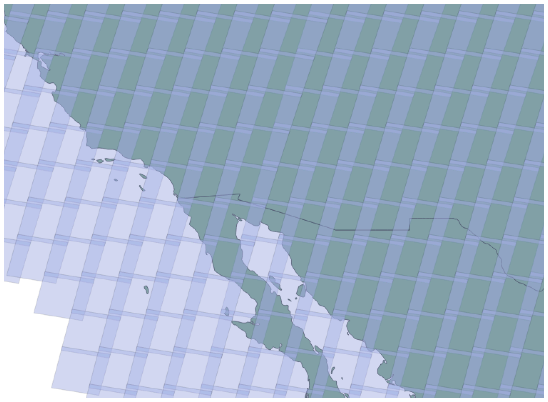
Appendix: Landsat Background

Landsat History and Applications

The first Landsat satellite launched in 1972 and since then, Landsat has grown to become the longest continuously running program for the collection of satellite imagery of the earth's surface. In 2011, the United Nations Educational, Scientific and Cultural Organization added the Landsat data archive to its Memory of the World Register, "a record of international documentary collections selected on the basis of world significance and outstanding universal value" ((*1*), p.xviii). As the world's population has doubled in the last fifty years, the Landsat program helps to "tell the tale, not only significant human impacts, but also of the effects wrought by natural processes" ((*1*), p.xix). The Landsat program consists of a series of satellites providing remote sensing imagery, and has had various funders, owners, and operators over its history. It is currently managed by the US Geological Survey (USGS) and the National Aeronautics and Space Administration (NASA). Other organizations that have controlled Landsat for a period of time include the National Oceanic and Atmospheric Administration (NOAA), a private company called Earth Observation Satellite Company (EOSAT, now Space Imaging) and the Department of Defense (DOD).

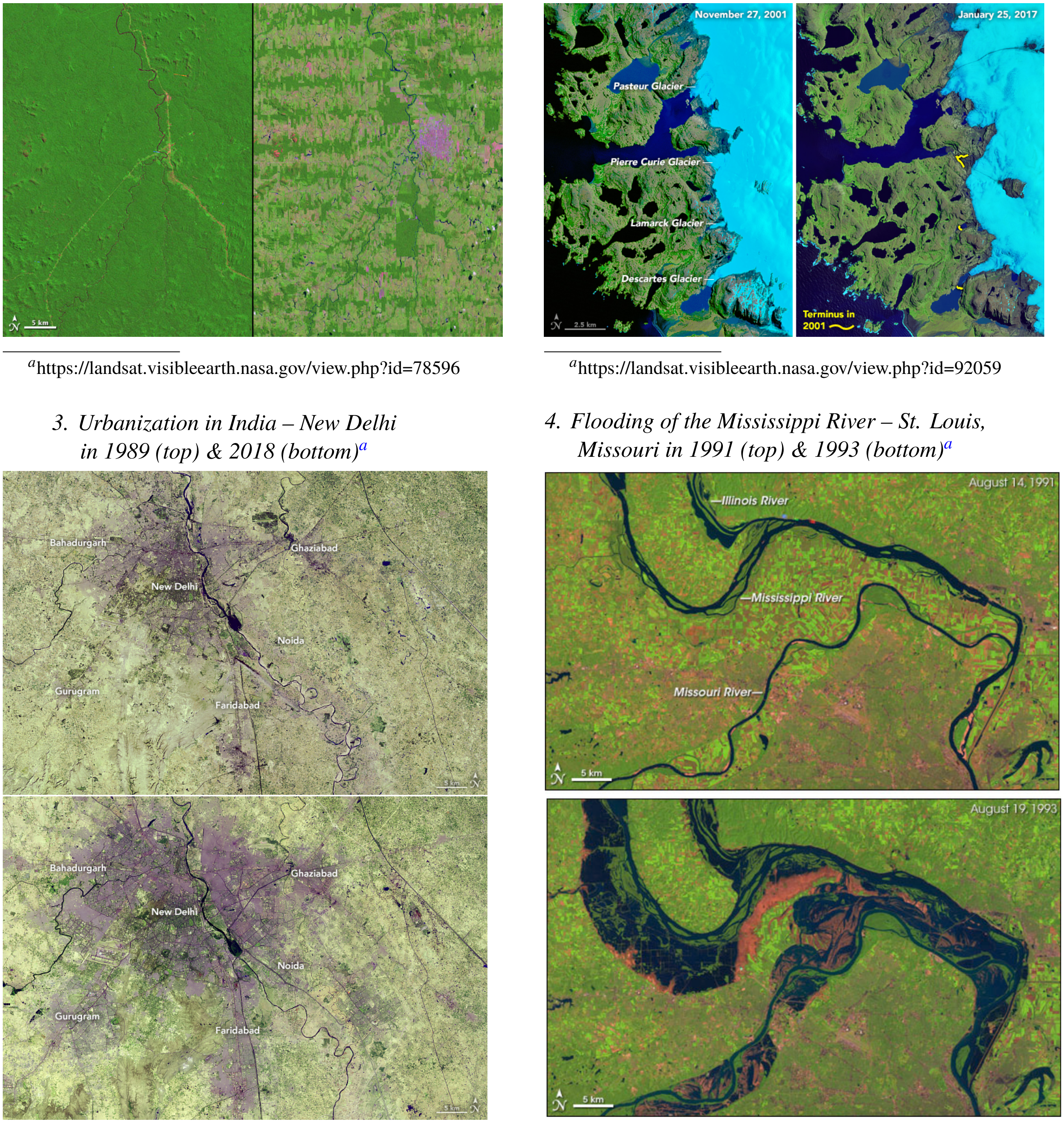
Landsat satellites orbit the Earth capturing images along a defined coordinate system. The Landsat World Reference System (WRS) Path and Row ID (henceforth referenced as PID) identifies one scene. The size of one scene is roughly 115 miles in length and 115 miles in width (around 13,200 square miles of coverage). In any given year, the active Landsat satellite may have captured many images of one scene, or not very many, or any. These images of a given PID scene may also contain high cloud cover, obscuring the details of the land or coast below. Images with high cloud cover are of little use to researchers requesting Landsat images. Figure S.1 provides a close-up view of Landsat scenes in the western United States.

Figure S.1: Close Up Look at PIDs



Landsat images have a wide variety of applications for scientific research. The research questions studied using Landsat cover a variety of topics like glacier retreat, urban expansion, population movements, deforestation, mining impacts (i.e. tar sands development, mountaintop removal), hydropower plant impacts, desert irrigation, agriculture expansion, shrinking of bodies of water (i.e. Aral Sea, Lake Urmia), fires (i.e. oil fires, forest fires), volcanic eruptions, hurricane flooding and other land use change studies ((*1*), p.xviii). Figure S.2 provides a few examples of the myriad different ways in which Landsat data have been applied.

Figure S.2: Examples of the use of Landsat Data for Detecting Environmental Change



1. Deforestation in the Amazon Rainforest – Rondonia, Brazil in 1975 (left) & 2012 (right)[a]

2. Glacier Melt in the Indian Ocean – Kerguelen Is. in 2001 (left) & 2017 (right)[a]

3. Urbanization in India – New Delhi in 1989 (top) & 2018 (bottom)[a]

4. Flooding of the Mississippi River – St. Louis, Missouri in 1991 (top) & 1993 (bottom)[a]

[a] https://landsat.visibleearth.nasa.gov/view.php?id=78596

[a] https://landsat.visibleearth.nasa.gov/view.php?id=92059

[a] https://landsat.visibleearth.nasa.gov/view.php?id=92813

[a] https://landsat.visibleearth.nasa.gov/view.php?id=5422

Commercialization Data Costs & Impact on Landsat Sales

The history of the Landsat program can be divided into three eras (2). The first began with NASA's initial development of Landsat in 1972 and continued through 1983. Satellites launched in this era, Landsat-1, -2, and -3 were largely similar, collecting multispectral data at a medium resolution of 30m. The second era (1982-1992) consisted of Landsat-4 and Landsat-5, launched in 1982 and 1984, respectively. These satellites included an additional Thematic Mapper (TM) sensor that gathered seven bands of data (as opposed to four in the previous sensors) and were particularly helpful for environmental monitoring and studying climate change. The key event in this phase, the Land Remote Sensing Commercialization Act of 1984, prompted the transition of operations and sales of Landsat images from the government to a private entity, EOSAT in 1985. The third era of Landsat began with the Land Remote Sensing Policy Act of 1992 which repealed the commercialization and started the process of moving all Landsat operations back under full control by the federal government. In particular, the Policy Act of 1992 mandated that the data policy for the pre-existing Landsat images should be renegotiated with EOSAT, before eventually transferring the management of pre-existing and new data to the government. It is this transition from the commercial second phase of the program to the open, third phase that is the focus of our paper.[1]

Of particular interest to our study is the cost of access of Landsat data under the second and third eras. The main problem with EOSAT's reign over Landsat was the cost. EOSAT served a wide variety of entities including commercial mining and oil exploration firms, the government as well as academics and researchers. Before EOSAT took over, Landsat photo products were around $10-$70 and the "computer-compatible tapes of the digital MSS data" cost $300 ((1), p.176). Under commercialization, these prices increased considerably. For example, in November 1991, the price of Landsat Thematic Mapper (TM) scenes increased to about $4,400 per scene. To put this in context, the cost to purchase one complete set of TM data covering the coterminus U.S. went from about $250,000 in 1982 to over $1.9 million in 1991 ((1), p.240). Further, EOSAT's restrictive policy did not allow data sharing. This was especially problematic for researchers who were interested in sharing data with collaborators and with the wider scientific community for replication and peer review. Anecdotal evidence supports the idea that these policies were particularly harmful for academic science. According to the Landsat Legacy Project Team: "From 1982-1990 the NASA Earth science program supported little Landsat-related research. There is no question that the expense of Landsat data severely impacted Landsat-based Earth science studies... Complaints about Landsat prices from the research community were not frivolous. Landsat data costs drained research budgets" ((1), p.202).

This situation began to change with the Policy Act of 1992. This wide-ranging piece of policy reshaped the future of the Landsat program extensively and mandated a dramatic lowering in the costs of data access. In practice, the implications of the Policy Act unfolder over the following decade and the specifics depended on the source of the underlying data (Landsat 4/5 vs. Landsat 7) and on the type of users in question. Since we are interested in the implications of this change for science, we focus on the implications of the law for academic, non-commercial users. For these users, this law established that, with some restrictions, unenhanced data from Landsat should be made available at the "cost of fulfilling user requests" (COFUR) and that this cost would exclude the fixed costs of designing, launching and maintaining the satellite system itself (3). Further, this Act constituted an independent entity, Landsat Program Management (LPM) that was tasked with implementing this goal and renegotiating access to Landsat data, especially for government, academic and non-commercial uses. This revised agreement

[1] Landsat-6 was planned but ultimately failed following an unsuccessful launch. Two more missions Landsat-7 and Landsat-8 were launched more recently (1999 and 2013).

between LPM and EOSAT on cost, processing and distribution rights for data was initialed in April, 1994 (*4*). Under this agreement, it was mandated that Landsat data be made available to educational institutions and non-profit organizations at reduced prices and with dramatically lower restrictions on data sharing. For example, this agreement mandated that the cost of scenes purchased after December 31, 1994 would go down to $2,500 (a reduction of 43%). Equally important, restrictions on data sharing and reuse were also lifted. For all data purchased in 1995 and beyond, could be shared without restrictions for non-commercial, academic and research uses ((*4*), p. 742). We therefore consider 1995 to be the year when data was available to researchers and academic institutions at reduced costs and with far fewer restrictions than before, and evaluate the effects of this change on academic science.

Appendix: Data Sources & Construction

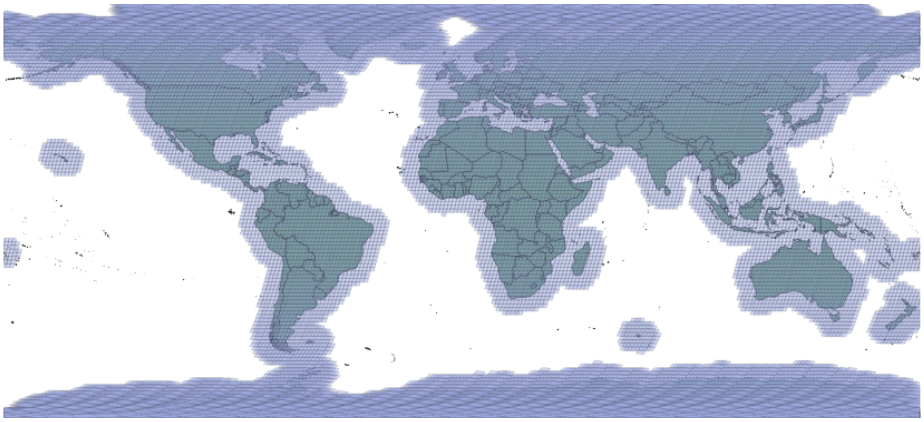
To evaluate the impact of data costs on the quantity, quality, and diversity of scientific research, we need data on the data source in question (Landsat images), a measure of scientific research (academic publications), and indicators of research diversity.

Landsat Data

We obtained Landsat coverage data from 1972-2013 from the United States Geological Survey (USGS) Earth Resources Observation and Science (EROS) data center metadata files. These data provide a list of all images collected by the Landsat program, including their location in the Landsat reference system, the image quality (based on percentage of the image covered by clouds), the specific Landsat satellite (Landsat 1-5), and types of imaging used to capture the scene (MSS or TM), and the center latitude and center longitude of the image. 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).[2] Many of these blocks do not cover the Earth's landmass. Therefore, to narrow the scope of Landsat images we consider relevant to our analyses, we created a risk-set of Landsat blocks by considering those blocks that intersected with land or ice mass, plus a buffer zone of blocks that border these masses. In total, this procedure leaves us with 12,577 blocks which we consider relevant for our analysis. This set is shown in Figure S.3.

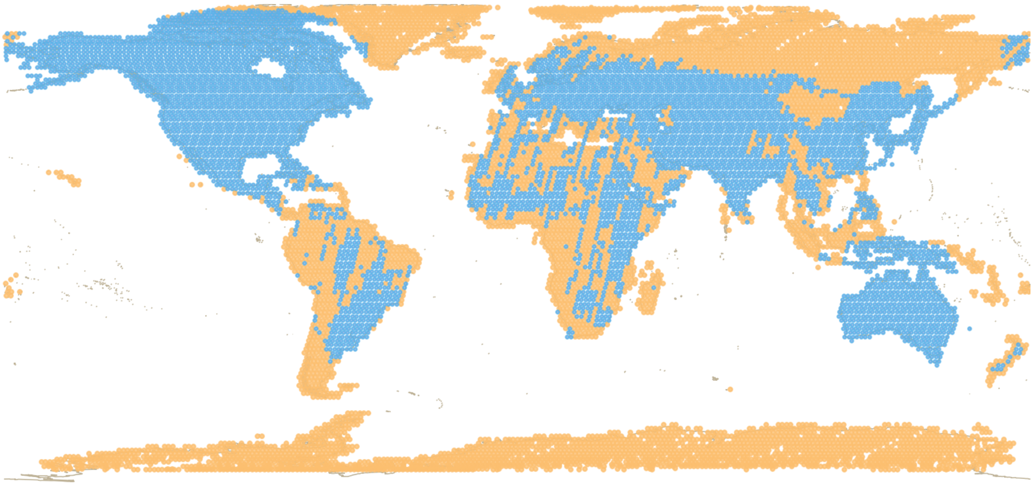
[2] Because the different Landsat satellites use a different coordinate system (Landsats 1 through 3 use World Reference System 1 (WRS1), Landsats 4 and 5 use WRS2), we translate WRS2 blocks to WRS1 blocks in the Landsat data. To do this, in QGIS we intersect WRS2 block centroids with WRS1 block polygons to create a translation set.

Figure S.3: Risk Set of PIDs (with World Map embedded below for comparison)



Compiling the risk-set of blocks allows us to characterize the availability of data for scientists to use for follow-on research and metadata allows us to identify considerable variation in the number and quality of images for each Landsat block. One useful way to measure this variation is by counting the total number of scenes available for a given block before Landsat is commercialized. The higher the number of images, the more affected a block should be by commercialization, since the value of blocks with few images is limited to begin with. We use this variation to characterize the impact of commercialization on science. In particular, we classify blocks into one of two groups depending on whether they were above or below the median (i.e. 18 images taken before commercialization) in terms of the number of images in 1985. We then analyze the differential impact of changes in the cost of Landsat data on research across these two sets of blocks. Figure S.4 provides a map showing the blocks with above-median number of images (blue) as compared to those with below-median coverage (yellow).

Figure S.4: Map of PIDs Separating Blocks With Above or Below Median Number (18) of Landsat Images by 1985



Landsat data was largely collected by the central Earth Resources Observation and Science (Center), but some data was also sent to international ground stations at various locations around the world. In theory, scientists could access data from either the EROS center or one of the ground stations. While the data from the international ground stations has been now centralized at the EROS center (*5*), it is possible that we are missing information on the type of data collected at the international ground stations in the metadata files we analyzed. We were not able to identify to what extent our data fail to capture international ground station collection efforts, which is a potential limitation of our analysis.

Scopus Data

The measures to capture scientific research come from SCOPUS, Elsevier's "abstract and citation database of peer-reviewed literature."[3] The results of a search for "Landsat" (and some related terms) in All Fields, up to 2005 (executed in 2017), yielded a data set of academic publications using or referencing Landsat from 1974 to 2005, comprised of roughly 24,000 publications by over 34,000 authors. These publication titles, abstracts, and author affiliations were geoparsed, i.e. we first detected words that represented place names (using machine learning entity-detection algorithms), and then geocoded these place names to obtain a latitude and longitude for these places. This allows us to match places studied in a paper as well as author addresses to specific blocks on the surface of the earth corresponding to a Landsat imaging location. For example, consider publication "A mini-surge on the Ryder Glacier, Greenland, observed by satellite radar interferometry" which was published in the journal *Science* in October 1996 (*6*). Our algorithm parses through the title and abstract of this publication and extracts the fact that it is using Landsat data to study Greenland, which corresponds with a particular block of Landsat imagery. We use a similar procedure to match the author affiliation locations to Landsat blocks. For example, the authors of this study are based in Pasadena, CA (at the Jet Propulsion Lab) and College Park, MD (and the University of Maryland), and we match this publication to the corresponding Landsat blocks. In addition to this procedure that matches publications and authors with specific locations, our publication data set also includes information on the publication itself (title, year, authors, publication source, abstract, keywords, etc.) as well other metrics available from SCOPUS such as number of citations and journal quality measures. In particular, journal quality metrics from SCOPUS include SCOPUS' CiteScore, CiteScore percentile, Source Normalized Impact per Paper (SNIP), and Scimago Journal and Country Ran (SJR).

Measuring Democratization and Diversity

In addition to the data on Landsat image coverage and Landsat-related academic publications, we also need to incorporate variables that can capture democratization and diversity of scientific research. To construct these measures, we focus both on authors and on research topics. For authors, we measure the economic status of the author's country of residence (e.g U.S., High Income, or Rest of the World) as well as the rank of their academic institution according to the QS World top university ranking. We consider academic research to be more inclusive if it is produced by authors from countries with GDP per capita less than $30,000/year or if it is produced by authors affiliated with lower ranked academic institutions. We choose the $30,000 cutoff since this represents the mean income among the blocks with at least one study in our sample. Our country income tiers are coded based on World Bank data which matches the country to its GDP per capita for both the year of the publication and for 2016. Diversity in topic space is measured by matching the geographic focus of a study to the country's income tier (U.S., High Income, or Rest of the World). We consider academic research to be more diverse in topic space if it focuses on a wide range of new geographic locations. For both the author and publication diversity measures, a country is categorized as high-income if its GDP per capita is greater than $30,000/year.

[3] See www.scopus.com.

Summary Statistics

As mentioned before, we have 12,577 unique blocks over which we track Landsat coverage and publications. Of these blocks, 5.72% are in the U.S. (719 blocks), 39.56% are in other high-income countries (4,975 blocks), and 54.73% are in the rest of the world (6,883 blocks).

In total, we consider all publications between 1975-2005 that refer to Landsat in our analysis, which leaves 23,939 publications in our sample. Of these publications, 1,396 are cited more than a 100 times by 2016, and we classify these publications as "highly cited" (5.8%). Further, of all the publications, 1,937 (8.1%) were published in a top journal, i.e. those published in the top 2 percentile of journals by SCOPUS' CiteScore. We use these measures to visually evaluate the effect of data access on publication output and quality. Further, in our regression analysis, we use a sample of 10,292 blocks that received at least one image by 1985. Within this sample, the median block receives 18 images by 1985, although this number varies between 1 and 302.

When it comes to authors, the publications in our sample were written by a total of 34,323 authors and on average, a publication has 2.8 authors (minimum author count is zero, maximum author count is 56). Further, our geoparsing algorithm detects over a total of 85,879 locations associated with authors (9,205 before 1985, 18,193 in the commercial era, 58,481 in the open era). Further, we match each of these locations to blocks, and find that as of 2005, about 1,656 blocks had at least one participating Landsat author. Of these blocks, 359 are in the U.S., 391 are in other high-income countries, and 935 in the rest of the world. Further, of the 23,939 publications in our sample, 5% of the publications included an author from a Top 50 Institution (1,195) and 8.8% included an author from the Top 51-200 Institutions (2,109).

Finally, when we look at the topics under study, i.e. the locations of the study detected from our geoparsing exercise, we find a total of 48,264 location names studied in the Landsat literature between 1975-2005. These locations match to 2,994 unique Landsat blocks that have been studied during this time. In other words, about 23% of the blocks on land have at least one paper about them by 2005. About 468 of these blocks are in the US, 664 in other High Income countries, and the rest of the world has about 1906 blocks. Further, on average, a paper has 2.9 studied locations. In Era 2, this average is 2.6 and in Era 3, each publication has an average of 3.2 studied locations.

Appendix: Regression Estimates & Supplementary Figures

Effect of Data Access on Quantity and Quality

We now turn to describing the main regressions we use to supplement the figures in the main article and provide estimates of the impact of reducing data access restrictions on scientific output. Our first set of specifications estimates the effect of reducing data access restrictions on academic science by comparing the number of publications in blocks with above-median level of Landsat coverage to blocks with below-median coverage as described in Figure S.4. We use Ordinary Least Squares (OLS) to estimate the following regression specification using a balanced panel at the block-year level:

$$Y_{it} = \alpha + \beta_1 \times Above - Median_i \times Post - 95_t + \gamma_i + \delta_t + \epsilon_{it}$$

where $Post - 95_t$ equals one for all years after 1995 and $Above - Median_i$ equals one for all blocks with above-median level of data as defined before and γ_i and δ_t represent block and time fixed effects respectively for block i and year t. These block-level fixed effects help to control for any stable cross-sectional differences between above-median and below-median blocks in a flexible way. Therefore, even though above-median and below-median blocks are not randomly distributed across the world, our estimates are able to control for the fact that some regions will always have more scientific interest – and thus more publications – irrespective of the cost of data access. Similarly, we also control for year fixed effects which help to control for a general increase in environmental science across the world. Therefore, the key assumption underlying our difference-in-differences specification is the "parallel trends" assumption, or the assumption that above-median and below-median blocks are evolving similarly in terms of their likelihood of being studied in a scientific paper. As we will show, our quantitative analysis finds support for this assumption.

The main outcome variables Y_{it} are the total number of publications, the total number of highly cited publications (defined as publication with more than 100 cites) and the likelihood of at least one publication in a given year in a given block. The estimate of β_1 can be interpreted as the change in the level of Y_{it} between above-median blocks as compared to below-median blocks after costs of data access have been reduced, and provides the main estimate of the impact of reduced costs of access on scientific publishing.

Results from this analysis are presented in Table S.2. In all three models, the estimate on β_1 is positive and significant implying that reducing cost of access greatly increasing the amount of publications, the number of highly cited publications and the likelihood of any publication in blocks with significant Landsat imagery as compared to those without. Since these are OLS estimates, they can be interpreted by comparing the estimates with the average level of the outcome variables before 1995. In particular, the estimates imply that total publications increase by a factor of three (mean 0.15), the number of highly cited publications by a factor of six (mean 0.0019) and the likelihood of publication increase by about 50% (mean .047).

Effect of Data Access on Democratization: Authors

Having established that reducing data access restrictions (lowering costs of data) leads to increased publications, we now turn to estimating whether this change serves to improve the diversity of science. We examine the effects of reduced restrictions to data by rank of the author(s') institution(s) and the income level of the country (or countries) of their institution(s). In particular, we use the same data set used to produce Figure 2 to estimate the impact of data access on certain groups of authors. Note that our analysis here is much simpler than the analysis for the baseline analysis. Our goal is not to provide a causal estimate via a quasi-experimental design, but rather to provide quantitative estimates underlying Figure 2, and establish whether differences are statistically significant. Since our analysis is not at the block level, we do not rely on the sample used for Table S.1, but rather on the much smaller samples based on group and year averages.

In Table S.3, Panel A we compare the impact of improved data access separately for institutions ranked in the top 50 as compared to those in the category 50-200. Before 1995, these two groups of institutions produced about 20 and 30 papers per year using Landsat data. After data access becomes relatively cheaper this number increases by about 49 (column 1) for all institutions, but this increase is more than double for institutions ranked 50-200. In other words, these institutions increase their total annual publications by about 95 publications, while those in the top 50 increase publications by only about 49. These estimates do not change much when including year fixed effects (column 2).

Table S.3, Panel B provides estimates using a similar strategy except by focusing on country income category rather than institutional rank. Country income is divided into three groups, the US, other non-US high-income countries (including western Europe), and rest of the world. Before 1995 non-US high-income countries and the rest of the world had a similar number of annual publications on average. The non-US high-income countries and the rest of the world published about 240 and 188 papers yearly, on average, while the US published about 121. However, as indicated in Panel B, column 2, the increase in publications post-1995 is about 44 publications for high-income countries, which increases by 428 more publications for non high-income countries.

Overall, Table S.3 confirms the intuition behind Figure 2 that lowering the costs of data access is particularly beneficial for authors in lower-ranked academic institutions and those in non high-income countries.

Effect of Data Access on Diversity: Research Locations

Finally, we now turn to assessing the impact of lowering costs of data access on the types of topics studied in the focal paper – notably the geographical location of the study. Does lower cost of data access also diversity the geographical focus of papers? We employ a similar specification to the regressions measuring author diversity, but instead of conducting the analysis at the publication-level, we are looking at the effect of reduced data access restrictions on the unique locations studied, classified by income of the studied location country. We do this because there might be multiple locations, around the world, studied within one publication. For example, if an author based in the United States now published a study focused on Egypt and California, we will count this as two unique locations, one as a high-income location and one as a non high-income location. Table S.4 provides estimates from this analysis. After restrictions to data access is eased, the number of locations studied goes up by about 70 blocks, but this increase is much greater in non-high income regions of the world, which see a further increase of about 91 unique locations. Compared to a baseline of about 200 locations, this represents a near doubling of the number of non high-income blocks studied per year in the literature. Next, we examine the number of blocks which are studied using Landsat data for the first-time in the literature (columns 3 and 4). Here too the results are similar. Non high-income regions see about 19 additional blocks being studied for the first time in the literature every year after data access is made easier, while the baseline effects for high-income regions is insignificant and close to zero.

Overall, this analysis suggests that lowering the cost of data access greatly increased the diversity of the regions studied using Landsat data by increasing the representation of non high-income regions as the topic of study.

Methodological explanation for Figure 3, Panel c - Distance between Authors and Study Locations: In an ideal type experiment one would sample all blocks, identify all scientists who could potentially incorporate Landsat data in their research, and then randomly give these scientists access. One could then examine whether having a treated scientist close by increases the chances of a block being studied. Several issues – including the fact that sampling all scientists whose work could possibly build on Landsat data is infeasible – prevent us from conducting such an analysis. We have, however, designed a counterfactual thought experiment that should address the puzzle too. The question informing the thought experiment is: If the studies on previously unexplored blocks in the open era had been conducted by authors with data access in the commercial era, would the distance between study location and author location have been greater than what is actually observed? To answer this question, we need to assign scientists active in the commercial era to new geographic blocks studied in the open era. Doing so randomly would overestimate the counterfactual distance, but doing so based purely on closeness will underestimate distance because other factors affect the locations one studies. We therefore establish the "preference" for closeness (i.e. distance between author and study location) among scientists in the commercial era. We do so by matching all locations studied in the commercial era with all authors active in that era and by determining where the observed pair (i.e. the location and the scientist who actually studied that location) ranks in the distance distribution. We then create a list of all combinations of commercial era authors and new study locations in the open era and use the percentiles obtained in the previous step to select our counterfactual matches. We then compare the actual and counterfactual distributions. In contrast to Figure 3, Panel c, Figure S.8 below includes new study locations in the U.S. The average distance between actual authors and study location is 4,704 kilometers, while the distance between counterfactual author and study location is 5,430 (t = 3.8669).

Robustness and Additional Results

Finally, we also present additional figures and results that complement the baseline analysis in order to provide robustness for the overall research design.

Excluding the US: Note that in the our baseline analysis we compare blocks with above-median coverage, with blocks with below-median coverage. Importantly, the Landsat program focused on the U.S. and therefore the entire continental U.S. has good coverage and is in the above-median category. If U.S. science has a surge in publications post-1995, then this pattern could confound our results. We therefore examine the robustness of our results to excluding U.S.-focused publications. Table S.5 presents our baseline results excluding all U.S. blocks. The results remain robust and large. Further, Figure S.9 presents figures similar to the ones in the main article, but exclude all U.S. blocks. These figures further validate that our baseline results are not U.S.-specific and do not depend on U.S. science only. In sum, when we drop U.S. blocks from our analysis, the overall conclusion that lowering the cost of data access improves the quantity, quality and diversity of scientific research holds.

Excluding China: Like U.S. scientists, Chinese scientists are responsible for a large volume of publications in the Earth and Environmental Sciences. If the rise of Chinese science started around the same time that Landsat access improved, one might be worried that this trend presents an important confound for our baseline estimates. To address this issue, we estimated our main regressions on data excluding blocks in China. The results of these regressions are presented in Table S.6 and show that the main results are qualitatively similar to earlier estimates on the whole sample. We therefore conclude that the rise of Chinese science, while important in its own right, is unlikely to have driven the main effect of data access on the increase of Landsat science.

Time-trends: An important assumption for the validity of the baseline results is that treatment and control blocks would have followed a similar trend in terms of their outcomes had data access not changed in 1995. In other words, if control blocks were already gaining publications (perhaps because of the increasingly global nature of academic science), then the difference in the change in publications between treated and control groups is simply a continuation of a pre-existing differential trend between the two groups. On the other hand, if the two groups were changing their rate of publications at a similar rate, then our estimation strategy is valid.

Here we examine this “pre-trends” assumption which is important for the robustness of difference-in-differences specifications. Specifically, we estimate $Y_{it} = \alpha + \Sigma_z \beta_t \times Above - Median_i \times 1(z)_t + \gamma_i + \delta_t + \epsilon_{it}$, where γ_i and δ_t represent block and time fixed effects, respectively, for block i and year t, and z_t represents the “lag,” or the number of years that have elapsed since 1995, i.e. the year in which data access restrictions were reduced. Figure S.5 presents estimates of β_t from this regression for the total publications outcome. These estimates measure the difference between treated and control blocks for year before and after 1995 in terms of total publications.

This resulting figure makes two points. First, there are no pre-existing differences in trends between treatment and control blocks. This is reassuring because it shows that even though treated and control regions are quite different, their rate of change in terms of publications is similar. Second, there is an immediate increase in publication activity after data access is made cheaper, and this impact becomes even larger about four to five years after data access made made possible. This points to the dynamic role of data access policy on science.

Testing robustness to alternate treatment/control definitions: Our research design relies on comparing blocks with higher coverage with blocks with a lower level of coverage. If the control blocks are blocks with few images to do research with, why would we expect scientists to study these blocks at all? Note that our intention is not to compare blocks with many images to blocks with zero or a very limited number of images. Instead, our strategy is to compare blocks with more images to those with fewer images. In other words, although below-median blocks have fewer images, the number of images available, on average, is 6.49. This implies that there were multiple images in the control blocks that were collected even before 1985 that could be used to do research. In fact, in the area of remote sensing, the first image of any location is very valuable and the value of additional images is significant, but lower. Past research shows, when researchers are interested in stable environmental features (such as in geology), the arrival of even a single image can encourage follow-on research and private-sector discovery in the area of gold mining (*7*).

Regardless of this point, one might be hesitant about a test that compares blocks with a large number of images to those with very few or zero images. Therefore, we test the robustness of our estimates to two alternative specifications. First, we excludes all blocks in the control group with less than five images to ensure that control blocks have a minimum level of capacity to produce research. Second, rather than comparing treatment blocks with control blocks by splitting along the median, we split blocks into three groups: "low" number of images (below median), "medium" number of images (median-90th percentile) and a "high" number of images (above 90th percentile). The medium and high category are subsets of the treatment group and so we are effectively comparing between blocks with sufficient coverage. Estimates from both these tests are presented in Table S.7. Both tests confirm our baseline predictions. First, we find positive and significant results when excluding blocks with fewer than 5 images. This suggests that blocks with zero or very few images are not driving our results. Further, the more the number of images the greater the positive effect. In fact, the baseline estimate jumps to 1.46 for total publications in the "high" as compared to about 0.195 for blocks in the "medium" category, suggesting that our effect is not dependent on the particular way in which the control group is constructed.

New images added post-1995: We identify our treatment and control blocks based on coverage data prior to 1985. However the number of images at the block-level is not fixed in time – images are constantly being collected in the commercial and open eras. This fact might lead to at least two concerns about the images added post-1985. First, treatment blocks may have seen a stronger increase in high quality images that started in 1995. If that is the case, the main effect of data access on Landsat science presented in our paper might be smaller or even zero. Second, one may argue that if new images mostly covered control blocks and this change coincided with the improvement of data access, we may have underestimated the true effect of data access. To evaluate these concerns, we collect data on the number and quality of images collected post-1985 and estimated several regressions in which we include control variables for the number of post-1985 images.

Specifically, in Table S.8, we add in Total Images (the number of Landsat images taken of that block in that year); Image Group Fixed Effects (image group buckets: 0, 1-9, 10-30, 31+); and Total Good

Images (count of images with less than 30 percent cloud cover, taken of that block in that year). The results presented in this table suggest that after controlling for these variable the estimated effect size of our *Post-95 X Above Media* interaction on Total Publications and Any Publications is virtually unchanged.

Next, in Table S.9, instead of adding variables that capture added images in that year, we add variables that capture the cumulative number of images. Specifically, we add Cumulative Images (cumulative image count post-1995); Cumulative Image Group Fixed Effects (cumulative image group buckets: 0, 1-9, 10-30, 31+); and Cumulative Good Images (cumulative good image count post-1995, good image defined as an image with less than 30 percent cloud cover). While we do see that the effect size attenuates to some extent, especially when we control for the cumulative number of images, it remains strong, positive and significantly different from zero.

Comparing a Broader Set of Non-Landsat Publications to the Landsat Sample

Finally, note that an ideal experiment analyzing the impact of data access costs on science might consider two observationally equivalent and costly data sources and eliminate the costs to one but not the other. Our research design is different than this ideal, although it approximates it to some extent. Specifically, rather than comparing Landsat publications to publications based on another data source, we compare Landsat publications in high coverage blocks with Landsat publications in low coverage blocks. The main assumption is that it is possible to publish science based on Landsat data in both types of blocks – except that the effective benefits from cost reductions are higher in treatment blocks since they have a greater number of images. By focusing on this intra-Landsat variation we are able to compare very similar types of scientific output and isolate the effect of cost variation. However, in addition to this variation, we might also want to compare Landsat publications to another comparable set of non-Landsat publications. Among other benefits, such a comparison would help account for the changing global trends in terms of scientific publishing (for example, the growing focus on Asia and the developing world).

Accordingly, we developed an empirical strategy as an alternative to our treatment/control distinction based on Landsat image coverage. Specifically, we collected a 5% sample of publications from SCOPUS (a total of 99,454 publications) that are labeled as Earth and Environmental Science publications.[4] We then geoparsed these publications and are able to assign 51,976 publications to one or more blocks. We use the GeoPy and Mordecai Python libraries for this exercise.

Armed with these data, we aggregate by year the number of geoparsed Landsat publications and the number of publications in earth and environmental sciences. These trends are presented in Figure S.10. While similar in spirit to Figure 1 in the main text, this figure is constructed using geoparsed publications at the block level. While both graphs show an increase in publications over time, it is clear that the increasing trend in publication volume of Landsat science starts in 1996 whereas the increasing trend in Earth and Environmental Science is relatively constant over time. The graphs also show that the relative growth in Landsat is much stronger than the growth in Earth and Environmental Science at any point in time. This is our first hint that Landsat publications increase at a greater rate post-1995 than do publications in the broader Earth and Environmental Sciences field.

Next, we used these data to more formally show the same pattern using a regression model. We build a

[4] The full sample of Earth and Environmental Science publications is very large and extracting publications from SCOPUS requires substantial manual labor. We therefore randomly sample 5% of all Earth and Environmental Science publications in a given year. Sampling publications by year is important to prevent interference with the time trend.

dataset that includes two observations for each block-year, one that includes that year's count of Landsat publications while the other includes that year's count of Earth and Environmental Science publications. These data are restricted to the set of blocks with at least one Landsat publication. We also restrict our attention to the "treatment" group (blocks with above-median images) in the baseline analysis, since this is where we expect Landsat publications to increase. Using these data, we estimate the following specification.

$$Y_{ilt} = \alpha + \beta_1 \times Landsat_l + \beta_2 \times Landsat_l \times Post_t + \beta_3 \times CumSum_{ilt} + \gamma_i + \delta_t + \epsilon_{ilt}$$

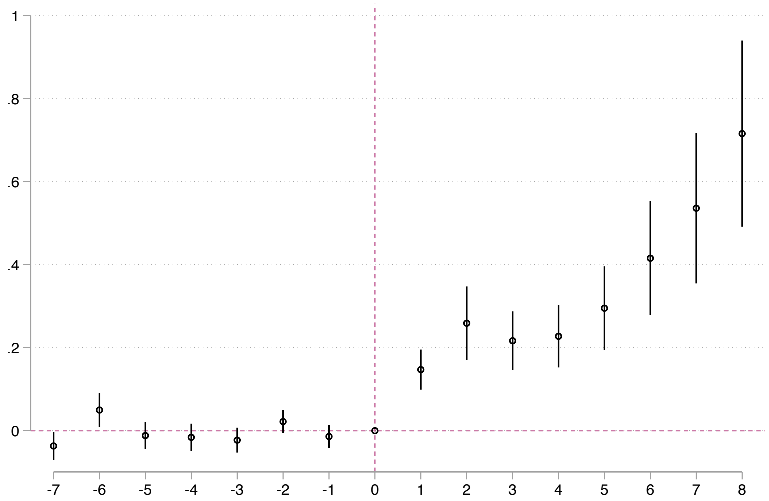
for block i in year t for group type l that equals one for Landsat publications and zero otherwise. Our outcome variable Y_{ilt} is the number of publications in block i in year t by group l. $Landsat_l$ is a dummy variable that equals one if the dependent variable indicates Landsat publications. We control for the cumulative sum of publications in block i until year t by group l. This control is needed to account for the differences in level[5] between Landsat and Earth and Environmental Sciences. This specification allows us to compare publication outcomes within blocks between Landsat science and Earth and Environmental Science, while our main analyses compare Landsat publications between blocks.

The results are presented in Table S.10 and show that using this completely different sample, we are able to replicate the main results from our earlier analyses. Specifically, we find that Landsat publications increase following 1995 after controlling for block and year fixed effects. Note that the coefficient on the $Landsat_l$ dummy variable is negative indicating the smaller size of this publication group. We take these results as evidence suggesting that it is unlikely that increasing trends in research activity in the Earth and Environmental Sciences are responsible for the treatment effect of data access presented in this paper.

[5] i.e. even the 5% sample is substantially larger than the sample of Landsat publications.

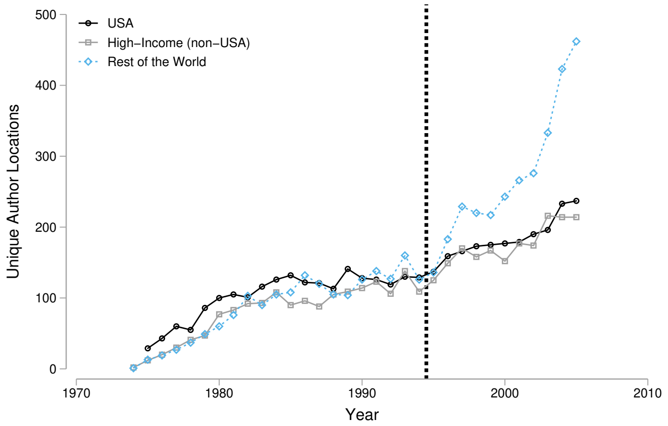
Additional Figures and Tables

Figure S.5: Yearly Estimates of the Impact of Data Access on Follow-on Publications



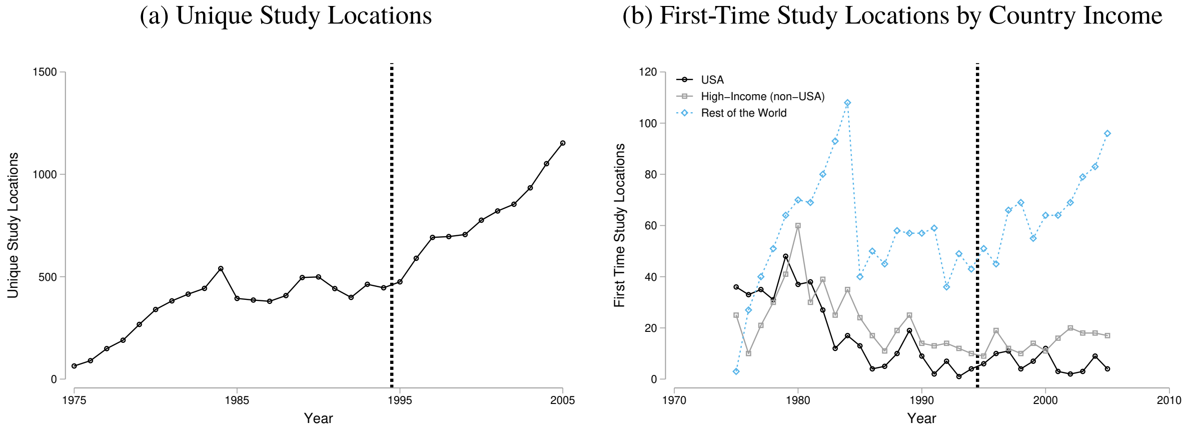
Note: The figure describes the impact of lower costs of data access on total Landsat publications over time. The vertical line represents the year 1995 in which data access costs were reduced. The y-axis plots estimates (and 95 percent confidence intervals) of β_t from the event study specification specified in Appendix C. This figure describes the estimated difference between treatment and control blocks for years relative to 1995.

Figure S.6: Unique Author Locations by Country Income



Note: This figure plots total number of unique author locations separated by country income categories.

Figure S.7: How Data Access Affects Study Locations

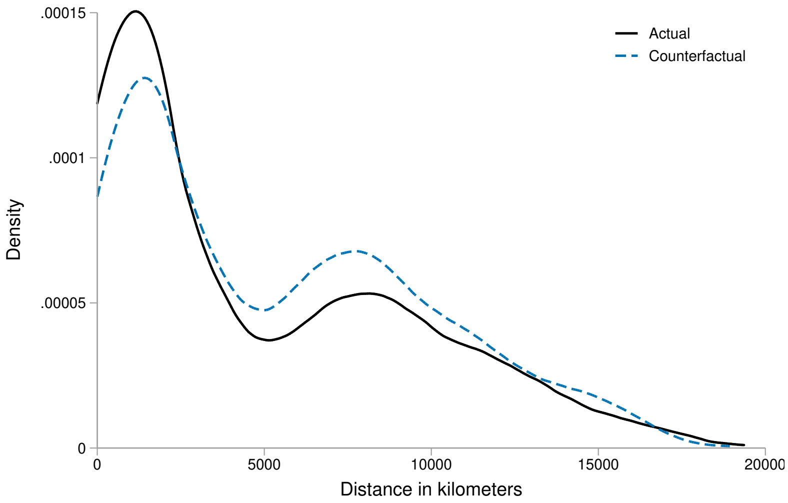


(a) Unique Study Locations

(b) First-Time Study Locations by Country Income

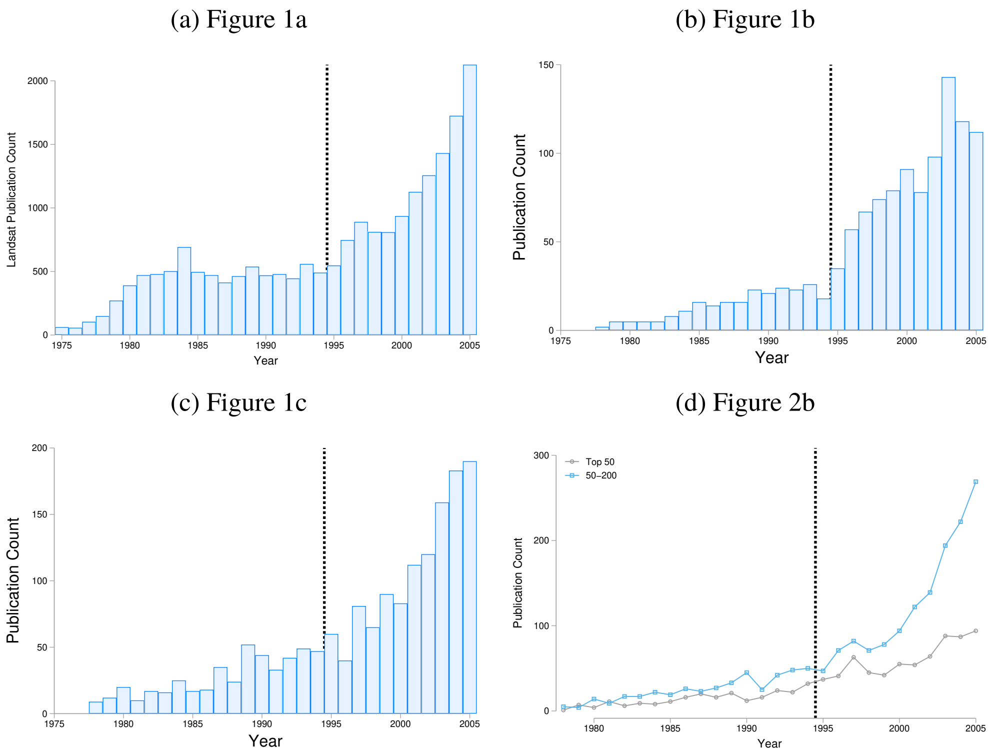
Note: Panel A plots total number of unique locations studied by Landsat publications. Panel B plots total number of first-time locations studied in Landsat publications separated by country income categories. Overall, the data suggest that lowering costs of data access particularly helped increase the number of Landsat publications about lower-income regions around the world.

Figure S.8: Distance between Author and Study Locations, including U.S. locations



Note: Distribution of actual and simulated distances between author locations and study locations, in kilometers.

Figure S.9: Figures Omitting USA Observations



(a) Figure 1a

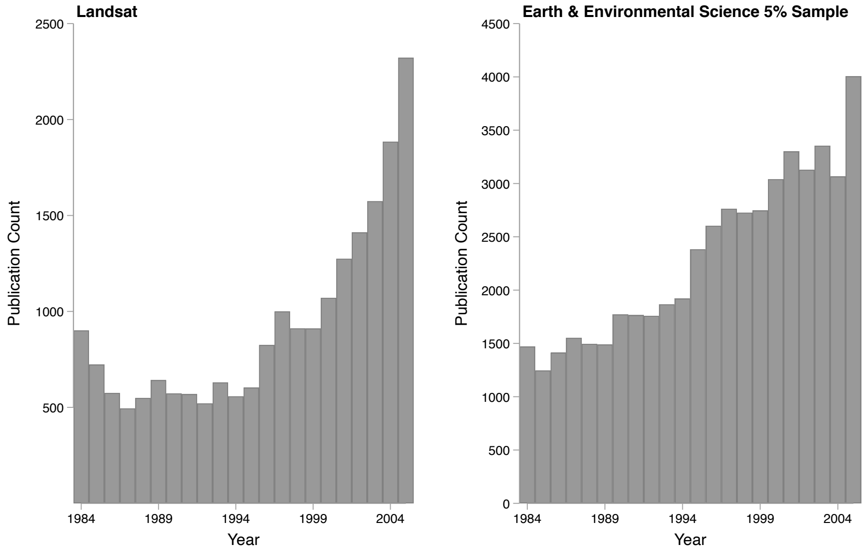
(b) Figure 1b

(c) Figure 1c

(d) Figure 2b

Note: The panel labels refer to which figure in the main text these graphs are mirroring, while excluding any US observations.

Figure S.10: Comparing Landsat and Non-Landsat Publications



Note: This figure compares the yearly publication trend of our Landsat-related sample to a 5 percent sample of all Earth and Environmental Science publications (excluding Landsat- and remote sensing-related papers) taken from SCOPUS.

Table S.1: Baseline Estimates for the Impact of Data Access on Publications

	Total Pubs	Highly Cited Pubs.	Any Pubs.
Post-95 X Above Median	0.437^{***}	0.0121^{***}	0.0251^{***}
	(0.0644)	(0.00183)	(0.00241)
Block FE	Yes	Yes	Yes
Year FE	Yes	Yes	Yes
adj. R^2	0.0101	0.00448	0.0216
N	216132	216132	216132
Clusters	10292	10292	10292

*+:p<0.15; *:p<0.10; **:p<0.05; ***:p<0.01, Standard errors clustered at block-level shown in parentheses. Note:* This table presents estimates that correspond to Figure 1. The unit of analysis is at the block-year for 10,292 blocks over 21 years from 1985 to 2005, for a total of 216,132 observations. Estimates are presented from the specification $Y_{it} = \alpha + \beta Post_t \times Above\ Median_i + \gamma_i + \delta_t + \epsilon_{it}$, where γ_i represents block fixed effects and δ_t represents year fixed effects. $Above\ Median_i$=0/1, equals 1 if a block received 18 images or greater from the Landsat program before 1985. These blocks are the ones that are most likely to benefit after 1995. $Post_t$=0/1, equals 1 for all years after 1995, when Landsat data were available at lower-cost.

Table S.2: Additional Estimates for the Impact of Data Access on Publications

	Total Pubs	Highly Cited Pubs.	Any Pubs.
Post-95 X Above Median	0.157^{***}	0.00564^{***}	0.00856^{***}
	(0.0236)	(0.00131)	(0.00240)
Post-99 X Above Median	0.106^{***}	0.00400^{**}	0.0120^{***}
	(0.0217)	(0.00175)	(0.00319)
Post-01 X Above Median	0.468^{***}	0.00866^{***}	0.0196^{***}
	(0.0745)	(0.00241)	(0.00344)
Block FE	Yes	Yes	Yes
Year FE	Yes	Yes	Yes
adj. R^2	0.0113	0.00484	0.0223
N	216132	216132	216132
Clusters	10292	10292	10292

*+:p<0.15; *:p<0.10; **:p<0.05; ***:p<0.01, Standard errors clustered at block-level shown in parentheses. Note:* This table presents a version of the estimates in Table S.1. The main difference is that rather than estimate one $Post_t$ variable around the year 1995, we include three interactions with the variables $Post_95$, $Post_99$ and $Post_01$ where $Post_95$ is one for years 1996-1999, $Post_99$ is one for the years 2000-2001 and $Post_01$ is one for years 2002-2005. Otherwise, the specification is identical to the baseline specification.

Table S.3: Panel A: Baseline Estimates for the Impact of Data Access by Authors' Institutional Rank

	Total Pubs	Total Pubs
Post-95	49.07^{***}	
	(8.190)	
Rank-50-200	12.00^{***}	12.00^{***}
	(3.859)	(3.199)
Post-95 X Rank-50-200	55.27^{**}	55.27^{***}
	(25.76)	(17.00)
Year FE	No	Yes
adj. R^2	0.511	0.787
N	42	42

*+:p<0.15; *:p<0.10; **:p<0.05; ***:p<0.01, Standard errors clustered at the rank-group level shown in parentheses. Note:* This table presents estimates that correspond to Figure 2, Panel b. The unit of analysis is at the Top 50-200 institutional rank category, over 21 years from 1985 to 2005, for a total of 42 observations. Estimates are presented from the specification $Y_{it} = \alpha + \beta Post_t \times Rank50 - 200_i + \delta_t + \epsilon_{it}$, where δ_t represents year fixed effects. $Rank50 - 200_i$=0/1, equals 1 if an author institution was ranked in the Top 200 to Top 50. These blocks are the ones that are most likely to benefit after 1995. $Post_t$=0/1, equals 1 for all years after 1995, when Landsat data were available at lower-cost.

Panel B: Baseline Estimates for the Impact of Data Access by Authors' Country Income

	Total Pubs	Total Pubs
Post X Rest-of-World	428.0^{***}	428.0^{***}
	(115.0)	(88.92)
Post X High-Income (non-US)	119.9^{***}	119.9^{**}
	(38.15)	(56.35)
Rest-of-World	120.1^{***}	120.1^{***}
	(17.67)	(16.53)
High-Income (non-US)	67.30^{***}	67.30^{***}
	(17.22)	(14.67)
Post-1995	44.24^{*}	
	(23.15)	
Year FE	No	Yes
adj. R^2	0.594	0.745
N	63	63

*+:p<0.15; *:p<0.10; **:p<0.05; ***:p<0.01, Standard errors clustered at country-group level shown in parentheses. Note:* This table presents estimates that correspond to Figure 2, Panel c. The unit of analysis is at the country-group category of non-US-High-Income and the Rest-of-World, over 21 years from 1985 to 2005, for a total of 63 observations. Estimates are presented from the specification $Y_{it} = \alpha + \beta Post_t \times Rest - of - World_i + \delta_t + \epsilon_{it}$, where δ_t represents year fixed effects. $Rest - of - World_i$=0/1, equals 1 if if the author country is from a country other than the US or non-US high income countries. These blocks are the ones that are most likely to benefit after 1995. $Post_t$=0/1, equals 1 for all years after 1995, when Landsat data were available at lower-cost.

Table S.4: Baseline Estimates for the Impact of Data Access by Study Location

	Unique Loc.	Unique Loc.	First-time Loc.	First-time Loc
Post-1995	69.38***		-0.945	
	(13.31)		(2.047)	
High-Income (non-US)	-12.60*	-12.60	8.500***	8.500***
	(7.404)	(7.718)	(2.405)	(1.817)
Rest-of-World	93.90***	93.90***	42.00***	42.00***
	(11.16)	(9.835)	(3.127)	(2.677)
Post X High-Income (non-US)	-3.764	-3.764	-0.0455	-0.0455
	(17.64)	(18.57)	(2.884)	(3.529)
Post X Rest-of-World	160.8***	160.8***	18.91***	18.91***
	(41.96)	(27.31)	(5.530)	(5.164)
Year FE	No	Yes	No	Yes
adj. R^2	0.780	0.892	0.898	0.906
N	63	63	63	63

*+:p<0.15; *:p<0.10; **:p<0.05; ***:p<0.01, Standard errors clustered at country-group level shown in parentheses. Note:* This table presents estimates that correspond to Figures 3, Panel b and SI, Figure SI. S.7(b). The unit of analysis is at the country-group category of non-US-High-Income and the Rest-of-World, over 21 years from 1985 to 2005, for a total of 63 observations. Estimates are presented from the specification $Y_{it} = \alpha + \beta Post_t \times Rest-of-World_i + \delta_t + \epsilon_{it}$, where δ_t represents year fixed effects. $Rest-of-World_i$=0/1, equals 1 if if the studied location is from a country other than the US or non-US high income countries. These blocks are the ones that are most likely to benefit after 1995. $Post_t$=0/1, equals 1 for all years after 1995, when Landsat data were available at lower-cost.

Table S.5: Baseline Estimates - Excluding USA Observations

	Total Pubs	Highly Cited Pubs.	Any Pubs.
Post-95 X Above Median	0.238***	0.00473***	0.0187***
	(0.0429)	(0.00129)	(0.00241)
Block FE	Yes	Yes	Yes
Year FE	Yes	Yes	Yes
adj. R^2	0.0104	0.00265	0.0195
N	201117	201117	201117
Clusters	9577	9577	9577

*+:p<0.15; *:p<0.10; **:p<0.05; ***:p<0.01, Standard errors clustered at block-level shown in parentheses. Note:* This table presents estimates that mirror Table S.1, with all USA observations excluded.

Table S.6: Baseline Estimates - Excluding China Observations

	Total Pubs	Highly Cited Pubs	Any Pubs
Post-95 X Above Median	0.426***	0.0122***	0.0231***
	(0.0658)	(0.00187)	(0.00245)
Block FE	Yes	Yes	Yes
Year FE	Yes	Yes	Yes
adj. R^2	0.0103	0.00444	0.0195
N	205989	205989	205989
Clusters	9809	9809	9809

*+:p<0.15; *:p<0.10; **:p<0.05; ***:p<0.01, Standard errors clustered at block-level shown in parentheses. Note:* Excluded any PIDs (blocks) in China from baseline regression.

Table S.7: Examining Robustness using Alternate Treatment/Control Groups

	If Total Pre-1985 Images 5+		New Treatment Var: 1 if 17+, 2 if 149+	
	Total Pubs	Any Pubs	Total Pubs	Any Pubs
Post-95 X Above Median	0.349***	0.00745**	0.195***	0.0163***
	(0.0685)	(0.00319)	(0.0413)	(0.00244)
Post-95 X Above 90th Percentile			1.465***	0.0623***
			(0.283)	(0.00564)
Block FE	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes
adj. R^2	0.0113	0.0254	0.0145	0.0228
N	166992	166992	216132	216132
Clusters	7952	7952	10292	10292

*+:p<0.15; *:p<0.10; **:p<0.05; ***:p<0.01, Standard errors clustered at block-level shown in parentheses. Note:* The first modification restricts the data to keep observations if the total pre-1985 image count is 5 or greater. This ensures the control group is not all zero-image blocks. The second modification updates the treatment variable to be categorical: 0 if below median, 1 if above median (17+), and 2 if above 90th percentile (149+) image counts.

Table S.8: Baseline Estimates Controlling for Total Images, Image Group Fixed Effects, and Total Good Images

	Original Main Reg		with Total Images		with Image Group FE		with Total Good Images	
	Total Pubs	Any Pubs	Total Pubs	Any Pubs	Total Pubs	Any Pubs	Total Pubs	Any Pubs
Post-95 X Above Median	0.437***	0.0251***	0.431***	0.0265***	0.386***	0.0238***	0.448***	0.0270***
	(0.0644)	(0.00241)	(0.0616)	(0.00247)	(0.0557)	(0.00242)	(0.0661)	(0.00251)
Total Images			-0.000968	0.000209***				
			(0.00136)	(0.0000769)				
Total Good Images							0.00191	0.000334***
							(0.00209)	(0.000108)
Block FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
adj. R^2	0.0101	0.0216	0.0101	0.0216	0.0113	0.0219	0.0101	0.0216
N	216132	216132	216132	216132	216132	216132	216132	216132
Clusters	10292	10292	10292	10292	10292	10292	10292	10292

+:p<0.15; *:p<0.10; **:p<0.05; ***:p<0.01, Standard errors clustered at block-level shown in parentheses. Note: Image Group Buckets: 0, 1-9, 10-30, 31+

Table S.9: Baseline Estimates Controlling for Cumulative Images (post-1995), Cumulative Image Group Fixed Effects, and Cumulative Good Images (post-1995)

	Original Main Reg		with C. Images		with C. Image Group FE		with C. Good Images	
	Total Pubs	Any Pubs	Total Pubs	Any Pubs	Total Pubs	Any Pubs	Total Pubs	Any Pubs
Post-95 X Above Median	0.437***	0.0251***	0.229***	0.00992***	0.428***	0.0251***	0.314***	0.0158***
	(0.0644)	(0.00241)	(0.0417)	(0.00261)	(0.0632)	(0.00240)	(0.0483)	(0.00265)
Cumulative Images			0.00220***	0.000160***				
			(0.000497)	(0.0000134)				
Cumulative Good Images							0.00187***	0.000140***
							(0.000617)	(0.0000186)
Block FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
adj. R^2	0.0101	0.0216	0.0119	0.0236	0.0101	0.0217	0.0106	0.0223
N	216132	216132	216132	216132	216132	216132	216132	216132
Clusters	10292	10292	10292	10292	10292	10292	10292	10292

+:p<0.15; *:p<0.10; **:p<0.05; ***:p<0.01, Standard errors clustered at block-level shown in parentheses. Note: Cumulative Image Group Buckets: 0, 1-9, 10-30, 31+

Table S.10: Estimates Using New Treatment & Control Specification: Landsat versus Earth & Environmental Science Publications

	Total Pubs	Any Pubs
Post-95 X Landsat Pub	0.164**	0.0358***
	(0.0776)	(0.00505)
Landsat Publication	-0.291***	-0.101***
	(0.0562)	(0.00365)
Cumulative Sum	0.0483***	0.000496***
	(0.000279)	(0.0000182)
Block FE	Yes	Yes
Year FE	Yes	Yes
adj. R^2	0.566	0.390
N	87738	87738
Clusters	2089	2089

*+:p<0.15; *:p<0.10; **:p<0.05; ***:p<0.01, Standard errors clustered at block-level shown in parentheses. Note:* Control group is Earth and Environmental Sciences publications (excluding Landsat- and remote sensing-related papers) and the Treatment is Landsat-related publications. The Cumulative Sum is the cumulative sum of images per block, per treatment/control group. Sample includes all blocks with at least one Landsat publication and in the Treatment group in terms of image coverage as per the baseline analysis.

References

(*1*) Samuel N Goward et al. “Landsat’s Enduring Legacy: Pioneering Global Land Observations from Space”. In: (2017).

(*2*) Joanne Irene Gabrynowicz. “The perils of Landsat from grassroots to globalization: a comprehensive review of US remote sensing law with a few thoughts for the future”. In: *Chi. J. Int’l L.* 6 (2005), p. 45.

(*3*) Peter Folger. “Landsat: Overview and Issues for Congress”. In: (2014).

(*4*) E.J. Sheffner. “The Landsat Program – Recent History and Prospects”. In: *Photogrammetric Engineering and Remote Sensing* 60.6 (1994), pp. 735–744.

(*5*) Michael A Wulder et al. “The Global Landsat archive: Status, consolidation, and direction”. In: *Remote Sensing of Environment* 185 (2016), pp. 271–283.

(*6*) Ian Joughin et al. “A mini-surge on the Ryder Glacier, Greenland, observed by satellite radar interferometry”. In: *Science* 274.5285 (1996), pp. 228–230.

(*7*) Abhishek Nagaraj. *The Private Pmpact of Public Data–Landsat Satellite Maps and Gold Exploration*. 2020.