The i3 BigQuery Workspace:
Shared Infrastructure for Open Science

Matt Marx[*]
Cornell University & NBER

Dror Shvadron[†]
University of Toronto

February 2025

Abstract:

Large-scale open datasets have transformed empirical research on science and innovation, but their effective use remains constrained by persistent challenges, including computational barriers, provider dependence, reproducibility difficulties, limited transparency, and unequal access to research infrastructure. We describe the *i3 BigQuery Workspace*, a shared cloud-based platform that hosts curated datasets (including OpenAlex, PatentsView, and community-contributed resources like Reliance on Science) and allows researchers to query terabyte-scale data in seconds within a unified analytical environment. By combining execution-ready data, scalable computation, and transparent versioning, the workspace lowers fixed costs to entry and supports reproducible research. We document examples of research enabled by this infrastructure and discuss how community-owned data platforms can improve research efficiency and expand the set of feasible research questions in the economics of science and innovation.

JEL Classification: O31, O32, O33, O38, I28, C88

Keywords: Open science; Research data commons; Reproducible research; Economics of innovation; Cloud computing for research

[*] Cornell University, Department of Statistics and Data Science, Ithaca, NY, USA; NBER; email: mmarx@cornell.edu.

[†] University of Toronto, Rotman School of Management, Toronto, Canada; email: dror.shvadron@utoronto.edu.

1 Introduction

The availability of research data shapes both the rate and direction of scientific progress. Since Arrow (1962) and Nelson (1959), economists have recognized that scientific knowledge is a public good: non-rival in use and difficult to appropriate, leading to systematic underinvestment absent public support. Because new knowledge builds on existing knowledge, institutions that govern access to research inputs—data, materials, methods—have compounding effects on the production of science (Romer 1990, Dasgupta and David 1994, Furman and Stern 2011). The expanding frontier of knowledge has intensified these dynamics: as the burden of knowledge grows, researchers specialize more narrowly and increasingly produce knowledge in teams (Jones 2009, Wuchty et al. 2007), heightening the returns to shared infrastructure that facilitates collaboration across institutional and disciplinary boundaries. Conversely, restrictions on access can slow cumulative progress, narrow the pool of active researchers, and create inefficient duplication of effort (Heller and Eisenberg 1998, Williams 2013, Sampat and Williams 2019, Galasso and Schankerman 2015, Murray et al. 2016, Stephan 2012).

A growing empirical literature has documented these effects across scientific domains. Murray et al. (2016) show that when NIH reduced access costs for genetically engineered mice, new researchers entered and explored more diverse research paths. Nagaraj et al. (2020) find that making Landsat satellite imagery freely available increased scientific output, with researchers from developing countries and lower-ranked institutions benefiting disproportionately. Nagaraj (2022) demonstrates that public data encourages exploration and entry relative to private alternatives. Biasi and Moser (2021) show that weaker copyrights during World War II encouraged follow-on science, with the largest gains at less affluent institutions, and Bryan and Ozcan (2021) find that NIH open-access mandates increased patent citations to funded research by 12 to 27 percent. More broadly, surveys indicate that researchers increasingly perceive benefits from open data, even as they report lacking the resources, skills, and institutional rewards necessary to fully engage with open science practices (Tenopir et al. 2023, Ferguson et al. 2023).

The data to study these questions now exists at unprecedented scale. Bibliometric databases like OpenAlex (Priem et al. 2022) and Dimensions (Hook and Porter 2021) now cover the near-universe of scholarly output, while PatentsView, PATSTAT, and Google Patents have made the patent literature similarly tractable. The emerging "science of science" has been built on data resources like these (Fortunato et al. 2018), enabling large-scale empirical research on how knowledge is produced, disseminated, and built upon (Einav and Levin 2014, Varian 2014).

But making data available is not sufficient on its own. As datasets grow in volume and complexity, the costs of arranging access, processing raw files, and maintaining analysis-ready databases become prohibitive for all but the best-funded researchers. Downloading and storing terabyte-scale files requires infrastructure that few universities provide, and transforming those files into usable form demands skills in data storage, processing, and analysis that few Ph.D. programs teach (George et al. 2016). The bottleneck, in short, has shifted from data availability to data usability.

These barriers compound. When each research team must independently download, clean, and organize nominally identical datasets, the result is both massive duplication of effort and a quiet erosion of reproducibility. Researchers working with different versions of the same data, or following undocumented computational steps, cannot easily verify each other's results—and in practice, most do not try (Ankel-Peters et al. 2023). Studies suggest that only 30–40% of economics research is computationally reproducible (Christensen and Miguel 2018), and fewer than half of articles can be fully reproduced even from journals with data-sharing policies (Herbert et al. 2024). Yet when data and code are truly accessible, most results do reproduce (Fišar et al. 2024), suggesting that the problem lies less in the science than in the infrastructure surrounding it. Data commons—platforms that co-locate data, computing, and analytical tools on shared infrastructure—offer one model for closing this gap (Grossman 2023).

The Innovation Information Initiative

The Innovation Information Initiative (i3) was established in 2019 with the goal of providing open datasets and promoting research on the economics of science. As a scholarly collaborative contributing to an innovation data commons, i3 has curated open datasets and convened an annual Technical Working Group to discuss advances. The initiative has developed widely-used data products, all freely available: Reliance on Science linking patents to scientific papers (Marx and Fuegi 2020, 2022), Patent-Paper Pairs identifying individuals who both patent and publish (Marx 2025), Pasteur's Quadrant Researchers who both publish and patent (Scharfmann et al. 2025), and DISCERN matching patents and publications to U.S. publicly listed firms (Arora et al. 2024). Together, these efforts have lowered barriers to entry for researchers studying how science and technology interact.

The i3 also supports junior researchers through a Fellows program. Limited to Ph.D students and postdocs, the Fellows program encourages development of open datasets via funding and mentorship. Current and former Fellows have produced open resources spanning global patent history, corporate innovation, bibliometrics, and machine learning, including efforts to digitize India's full patent record, construct firm-level patent panels, link Traditional Chinese Medicine with modern drug systems, and trace the transnational movement of scientists.

From Open Data to Shared Infrastructure

Open datasets have accelerated research on innovation, with thousands of articles citing those created by i3 and other entities. Drawing on our experience building and maintaining these resources, we identify five persistent challenges that limit who can work with large-scale innovation data and how reliably results can be produced.

1. *Computational barriers*: as datasets grow too large to process locally, access no longer equals ability to use, and getting started requires substantial infrastructure and expertise.
2. *Provider dependence*: when external providers change priorities or shut down (as with Microsoft Academic Graph and, more recently, USPTO's PatentsView), researchers who built on these resources face sudden disruption.
3. *Reproducibility barriers*: even when data is available, few replications occur in practice (Ankel-Peters et al. 2023), and computational burdens prevent verification of results (Christensen and Miguel 2018).
4. *Transparency gap*: when methods are black boxes, the community cannot verify quality, integrate datasets, or contribute improvements.
5. *Resource inequality*: infrastructure access varies dramatically by institution, stratifying who can pursue certain research questions.

Google BigQuery offers a promising foundation for addressing these challenges. As a cloud-based data warehouse, it allows researchers to manipulate massive datasets using standard SQL without managing infrastructure. Users pay only for computation consumed. The platform eliminates the need for local storage, handles computational scaling automatically, and provides a stable development environment. Beyond storage and queries, BigQuery supports advanced analytics directly in the cloud. BigQuery ML allows researchers to train machine learning models on millions of records without moving data to a separate environment. User-defined functions extend SQL with custom logic for text processing and entity extraction. Integration with Python and R through cloud notebooks enables seamless transitions between exploratory queries and statistical analysis.

This model has already enabled research at scale. Economists have used BigQuery to analyze millions of GitHub commits in real time (McDermott and Hansen 2021), to search the full text of

U.S. patents for trends in high-tech patenting (Webb et al. 2018), and to systematically identify financial innovations across decades of patent documents (Lerner et al. 2021). Workflows that once required high-performance computing clusters or months of local preprocessing can now be executed in minutes. Yet these capabilities remain underutilized in the economics of science, in part because the relevant datasets are scattered across formats and providers, and in part because the upfront costs of learning the platform fall on individual researchers.

The i3 BigQuery Workspace addresses these barriers by offering shared infrastructure for an open science innovation data commons.[1] Core datasets such as OpenAlex, PatentsView, and CrossRef are centrally maintained alongside community-contributed resources including Reliance on Science, Patent-Paper Pairs, Pasteur's Quadrant Researchers, and DISCERN. As i3 Fellows and other community members generate new data, these resources can be incorporated directly into the workspace. This design allows researchers to work with the data immediately, without incurring the fixed costs of data discovery, ingestion, and preprocessing.

Looking ahead, we are designing the workspace for long-term sustainability. Future investments include automated data pipelines that update at regular intervals, expanded training and documentation to ease adoption, and deeper integration with the i3 Fellows program so that early-career scholars both use the infrastructure and contribute to building it. To support these efforts, development and maintenance costs will be funded through community contributions, aligning the infrastructure's survival with the research community it serves.

In what follows, we detail the challenges facing researchers working with large-scale data, describe the i3 BigQuery Workspace and its design, document evidence from early adoption, and outline plans for future development. The broader aim is to show that shared infrastructure for research data is not only feasible but surprisingly affordable, and that the real obstacles are less technical than institutional: how to fund maintenance, how to govern shared resources, and how to align academic incentives with the unglamorous work of keeping data alive.

Section 2 elaborates on five major challenges facing researchers working with large-scale open data. Section 3 describes the i3 BigQuery workspace and explains how specific design choices address each challenge. Section 4 presents evidence on usage, including adoption statistics and examples of research enabled by the infrastructure. Section 5 examines future directions and open questions about sustaining research infrastructure.

[1] Technical details and documentation are available at https://i3open.org/bigquery.html.

2 Current Challenges

The growth in availabilty of large-scale, open datasets has transformed what empirical research can accomplish. Comprehensive bibliometric databases allow researchers to trace knowledge flows across the entire scientific enterprise. Full-text patent archives enable systematic analysis of technological language and citation patterns. Linked administrative records connect individual researchers to their publications, patents, funding, and career trajectories over decades. These capabilities have opened research questions that were simply not feasible a generation ago.

Yet realizing this potential comes with persistent barriers. Our experience building and maintaining open datasets revealed several recurring challenges that together illustrate the infrastructure gap a shared data workspace must address.

The most immediate barrier is *complexity*. A doctoral student trained in econometrics can estimate a regression but may never have encountered a dataset that cannot be opened in Stata. Yet OpenAlex exceeds 1.4 terabytes uncompressed and is moreover delivered in hundreds of subfiles that must be linked together. Navigating these resources—writing SQL queries, managing cloud storage, cleaning and linking records across sources—requires hard-won expertise that accumulates slowly and is rarely taught, including in Ph.D programs (Varian 2014). Without such skills, researchers reinvent wheels, write fragile code, and spend months on data preparation that others have already solved.

A second challenge is provider dependence. When external providers change priorities or shut down, researchers who built on these resources have their work interrupted. The discontinuation of Microsoft Academic Graph (MAG) in 2021 illustrates this vulnerability. MAG contained over 225 million publications with 2 billion citations, and researchers who had integrated it into their workflows faced abrupt obsolescence (Priem et al. 2022). OpenAlex launched as a successor, but multiple changes in schema definitions required substantial adaptation. More recently, USPTO's November 2025 announcement regarding PatentsView has created similar uncertainty.

Reproducibility presents a third barrier. Recent evidence suggests that when researchers can access the original data and code, most published results do reproduce (Fišar et al. 2024). But few reproductions are even attempted. Obtaining the data is difficult, running the analysis is time-consuming, and there are few professional incentives to try (Christensen and Miguel 2018). Making reproduction easier by lowering the costs of data access and computation is an important step toward more reproductions actually occurring.

A related problem is the transparency gap between open access and open source. Many widely-used datasets are available for research but provide no visibility into how they were constructed or how accuracy varies across subpopulations. Databases like Dimensions and Google Patents offer substantial coverage, but their methods for creating those data remain proprietary. When users encounter apparent errors, they cannot verify whether the issue lies in the source data, the processing pipeline, or their own interpretation. Even documented methods can diverge from real-world performance: PatentsView's inventor disambiguation achieved nearly 100% precision on benchmark datasets, but independent evaluation found actual precision of only 87–91% (Binette et al. 2023). Cross-database comparisons reveal similar inconsistencies—patent counts vary substantially between PatentsView and PATSTAT even with standardized methodology (Mafata et al. 2024).

Finally, these barriers fall hardest on those with the fewest resources. Establishing proficiency with large-scale data takes time, mentorship, and access to infrastructure that not everyone has. PhD students and early-career researchers are most affected—they face the steepest learning curves while operating under the tightest constraints. At well-resourced universities, senior colleagues, dedicated data teams, and computing support can ease the path. At institutions without these resources, researchers are largely on their own.

2.1 Limits of Existing Research Infrastructure

A wide range of infrastructure solutions already exist to support open research, but each addresses only a subset of the constraints faced by large-scale empirical work. General-purpose data repositories, such as Zenodo and Dataverse, have substantially lowered barriers to data dissemination and long-run preservation. These platforms play a critical role in promoting open access and citation of research outputs. However, they are designed primarily for storage and download rather than execution. As datasets grow to hundreds of gigabytes or terabytes, the ability to formally access data does not imply the ability to work with them: users must still provision local storage, manage ingestion pipelines, and possess sufficient computational capacity to perform large-scale joins or text-based analyses. As a result, the fixed costs of using these data remain high.

By contrast, many universities maintain high-performance computing (HPC) systems. These systems address the need for scalable computation, but they are poorly suited to collaborative data-intensive research. Access to HPC clusters is typically restricted to specific groups, and usage often requires advance allocation, queueing, and specialized expertise. Moreover, HPC environments generally assume that users bring their own data, placing the burden of acquisition, cleaning, versioning, and storage on individual research teams. These features limit reproducibility and make it difficult for other researchers to rerun analyses without replicating substantial amounts of preparatory work. In practice, institutional HPC systems lower marginal computation costs but do little to reduce the fixed costs associated with discovering, preparing, and maintaining large datasets.

Taken together, existing solutions separate data access from computation and treat reproducibility as an afterthought rather than a design objective. The result is an infrastructure gap: researchers face persistent barriers even when data are nominally open and computational resources are available. The next section describes the i3 BigQuery Workspace as an attempt to address this gap by combining shared, execution-ready and versioned data with scalable computation under transparent and reproducible workflows.

3 BigQuery and the i3 Workspace

The i3 BigQuery workspace addresses each of the challenges outlined above by providing shared infrastructure built on Google BigQuery, a cloud-based data warehouse designed for large-scale analytics. BigQuery is widely used in industry, with companies like Spotify, Twitter, and The New York Times relying on it to manage petabytes of internal data. Oddly, this capable tool has seen limited adoption in academic research. Crucially for our purposes, BigQuery allows a centralized data repository to be shared publicly; anyone with a Google account can query the data without downloading files, configuring infrastructure, or requesting access. The platform offers capabilities well-suited to the scale and complexity of modern research data, and its pay-per-use pricing makes it accessible without institutional support.

3.1 Platform Capabilities

At its core, BigQuery is a serverless SQL engine. Researchers write standard SQL queries while the platform handles storage, computation, and scaling behind the scenes, with no infrastructure to provision or software to install. Queries run against datasets of any size, from a few thousand rows to hundreds of billions, and return results in seconds rather than hours. The SQL syntax will be familiar to anyone who has worked with Stata or R, and those with database experience will find the environment immediately productive.

The platform's architecture separates storage costs from computation charges. Data resides in Google's cloud infrastructure, organized in columnar format optimized for analytical queries. When a query is submitted, BigQuery allocates computational resources on demand, executes the query across distributed workers, and releases those resources when finished. Users pay only for the data scanned, currently $6.25 per terabyte, rather than for idle capacity. A free tier provides one terabyte of queries per month at no cost, enough for substantial exploratory analysis.

Beyond standard SQL, BigQuery offers advanced capabilities that extend what researchers can accomplish without leaving the platform. BigQuery ML allows users to train machine learning models directly in SQL, including linear and logistic regression, clustering, time series forecasting, and neural networks. A researcher can go from exploratory queries to predictive modeling without moving data to a separate environment or learning a new toolchain. User-defined functions extend SQL with custom logic useful for text processing, entity extraction, or domain-specific transformations. Native geospatial support enables spatial queries on patent inventor locations or institutional addresses. Integration with Python and R through cloud notebooks allows seamless transitions between SQL queries and statistical analysis. BigQuery also supports data versioning, allowing researchers to query historical versions of tables and access data exactly as it existed at a given point in time.

3.2 Current Resources in the i3 Workspace

The i3 workspace consists of a curated set of datasets uploaded to BigQuery and opened for public access. Our design choices respond directly to the five challenges identified earlier.

To address computational barriers, we pre-load core datasets including OpenAlex, PatentsView, CrossRef, and others so researchers can begin querying immediately. There is no need to download terabytes of data, configure local storage, or write ingestion pipelines. A researcher with a web browser can run their first query within minutes of creating an account. The same query that might take hours on a laptop executes in seconds in the cloud.

To guard against provider dependence, we maintain versions of datasets indefinitely, even after providers discontinue them. If a provider changes its data model or shuts down entirely, as happened with Microsoft Academic Graph, researchers can still access historic versions for reproducibility and independence. Looking ahead, we plan to take on maintenance of core datasets ourselves when providers withdraw support. When we do so, the processing code will be open source, reproducible, community-maintained, and automatically uploaded to BigQuery.

To support reproducibility, we version all datasets with dated suffixes. A researcher can specify the exact version used in their analysis, and any other researcher can run the same query against the same data years later. Updating an analysis to use newer data is equally simple, requiring only a change to the date suffix. This addresses a persistent problem in computational research, namely the inability to precisely replicate an analysis when the underlying data has changed.

To ensure transparency, we prioritize open-source datasets with documented methods. We open-source the ingestion processes used to obtain and load each dataset, allowing users to verify how data was processed and to reproduce or modify the pipeline for their own purposes.

To reduce resource inequality, we host the data at no cost to users, eliminating the need for local storage or dedicated hardware that large-scale analysis traditionally required. Researchers pay only for the queries they run, and these costs are reasonable for standard research workloads. A graduate student at a teaching-focused institution has the same access as a researcher at a major research university. We provide documentation, example queries, and tutorials to lower the learning curve, recognizing that infrastructure access means little without the knowledge to use it effectively.

The workspace currently hosts two categories of datasets, summarized in Table 1. Core datasets include major open resources covering scholarly publications, patents, and related metadata. These provide the primary data sources for research on science and innovation.

Community-contributed datasets extend this foundation with specialized resources developed by i3 fellows and collaborators. These include Reliance on Science, which links patents to the scientific papers they cite (Marx and Fuegi 2022), DISCERN, which matches patents and publications to U.S. publicly listed firms (Arora et al. 2024), and KPSS patent valuations based on stock market responses (Kogan et al. 2017). The workspace grows as community members contribute new datasets, creating a shared resource that compounds over time.

4 Evidence from Usage

4.1 Examples of Research Enabled by the i3 BigQuery Workspace

Many recent studies use the Google Patents Public Dataset, which is provided by IFI Claims and hosted on BigQuery.[2] For example, prior works have used these data to study the evolution of technologies, characterize patterns of patenting activity across firms and inventors, and reconstruct historical patent records at scale (Webb et al. 2018, Lerner et al. 2021, Gross and Sampat 2022). Abi Younes and de Rassenfosse (2024) provide a broad set of replicable patent indicators constructed directly from the Google Patents Public Datasets using BigQuery.

In our own work, BigQuery serves as the computational backbone for longitudinal analysis of the U.S. scientific workforce. Shvadron et al. (2025a) build a dataset of 1.2 million U.S. STEM PhD dissertations spanning 1950–2022, combining ProQuest metadata with text from titles, abstracts, and acknowledgments. BigQuery stores intermediate outputs from LLM-based pipelines that classify and process over 100 million sentences from the dissertation texts, and is later used to match funders mentioned in acknowledgments to standardized organization identifiers. Shvadron et al. (2025b) use

[2] IFI Claims was recently acquired by Digital Science (IFI CLAIMS Patent Services and Digital Science 2021)

Table 1: Datasets Available in the i3 BigQuery Workspace

Panel A: Core Datasets

Dataset	Description
Crossref	Scholarly metadata and DOI registration records
CrunchBase (2013)	Startup ecosystem and venture capital data
OpenAlex	Catalog of scholarly works, authors, institutions, and citation networks
Orange Book	FDA-approved drug products with patent and exclusivity info
PatentsView	Complete patent data from the USPTO
Retraction Watch	Database of retracted scientific publications
USPTO Patent Assignment	Patent ownership transfers and assignment records

Panel B: Community-Developed Datasets

Dataset	Description	Citation
British Historic Patents	Historical UK patent records	Berkes et al. (2026)
Commercial Potential of Science	Measures of commercial potential of science	Masclans et al. (2025a)
DISCERN	Patenting and scientific publications by U.S. publicly listed firms	Arora et al. (2024)
Founding Patents	Assignee age at patent grant	Ewens and Marx (2023)
Geocoding of Worldwide Patent Data	Geographic coordinates for inventors and assignees	de Rassenfosse et al. (2019)
HistPat	Historical patents with harmonized geography	Petralia et al. (2016)
Inventor Age	Inventor demographic and career-stage information	Kaltenberg et al. (2023)
JCIF	Journal Commercial Impact Factor	Bikard and Marx (2020)
KPSS Patent Value	Stock market-based patent valuations	Kogan et al. (2017)
Paper Twins	Matched paper pairs	Bikard and Marx (2020)
Patent Hubs	Geographic clustering of innovation	Bikard and Marx (2020)
Patent Paper Pairs	Linked patent-publication data	Marx (2026)
Patent Scope	Patent breadth and technological coverage measures	Kuhn and Thompson (2019)
PQRS	Inventors who publish and scientists who patent	Scharfmann et al. (2025)
Reliance on Science	Patent citations to scientific literature	Marx and Fuegi (2022)

BigQuery ML to train a predictive model for linking graduates to their publication records, then run inference on the complete sample of dissertations joined to candidate authors in OpenAlex. BigQuery handles the large-scale joins required to track migration through affiliation changes over time and to compute country-level shares of patent citations to graduates' research.

Early adopters of the i3 BigQuery workspace have used it to integrate multiple datasets at scale. Arts and Melluso (2025) trace knowledge flows from science to technology by linking publication data from OpenAlex with patent data from PatentsView, along with supplementary datasets including KPSS patent valuations and Reliance on Science. Their analytical pipeline runs entirely on BigQuery, matching publication text to patent full text across tables with billions of rows and constructing derived datasets at the inventor-scientist, patent, and idea levels without moving data between environments.

Another example comes from Masclans et al. (2025b), who develop ex-ante measures of the commercial potential of scientific research with the goal of identifying promising discoveries prior to any observed commercialization activity. The paper trains a sequence of machine learning models that map textual features of scientific publications to downstream commercial outcomes, measured using patent-to-publication citation linkages. Model training is conducted on more than one million publications, with large-scale inference subsequently run on over 30 million publications, all accessed directly from the i3 BigQuery repository. Training and inference are carried out on Vertex AI within Google Cloud, while publication metadata and patent citation data are fetched from i3-hosted BigQuery tables. In turn, the resulting commercial potential measures are made available for use by other researchers, and are hosted within the i3 BigQuery workspace.

4.2 Example Query

Consider what should be a very simple data-analysis task: the researcher would like to use the Reliance on Science paper-to-patent citations, analyzing time trends in terms of which papers are cited. Because the Reliance on Science dataset contains only the paper and patent IDs, but not additional information, the dates of paper publication must be merged on. Although conceptually straightforward, doing so would involve downloading the entirety of OpenAlex (1.4T), requiring more disk space than the average laptop user tends to have available. Once downloaded, one must uncompress and merge together hundreds of OpenAlex subfiles and then extract the date field. If feasible at all, this operation could take a day or two to carry out.

In the i3 BigQuery Workspace, this operation can accomplished in less than one minute and nearly costlessly. The following query joins patent-to-paper citations from Reliance on Science with publication metadata from OpenAlex:

```
WITH npls AS (
  SELECT patent, oaid
  FROM ‘nber-i3.reliance_on_science.pcs_oa_v64‘
),
openalex AS (
  SELECT CAST(REPLACE(id, ’https://openalex.org/W’, ’’) AS INT64) AS oaid,
         publication_date
  FROM ‘nber-i3.openalex.works_241125‘
)
SELECT *
FROM npls
LEFT JOIN openalex USING (oaid)
```

Listing 1: BigQuery query linking patents to OpenAlex works

Because both of these datasets are preloaded in the i3 BigQuery Workspace, this query runs in 16 seconds, costs 5 cents, and returns over 100 million rows linking patents to the scientific papers they cite.

4.3 Reproducibility

BigQuery queries are standard SQL, which means they integrate directly into replication packages. Authors can include query scripts alongside their analysis code, and any researcher with access to the i3 workspace can execute the same queries on the same data. This addresses a persistent barrier to reproduction: even when code is available, obtaining and configuring the underlying data often requires substantial effort that discourages verification.

The platform also supports reproducibility during the journal review process. Authors can grant temporary access to data editors, allowing them to run analyses directly rather than relying on submitted outputs. This lowers the cost of verification and makes it practical for journals to enforce data availability requirements for large-scale empirical work.

5 Future Directions

5.1 Sustainability and Governance of Shared Infrastructure

The challenges outlined in Section 2—including provider dependence, reproducibility barriers, and resource inequality—are not solely technical. Rather, they reflect a broader institutional problem:

maintaining large-scale research datasets is a collective-action challenge. The benefits of high-quality, well-maintained data accrue widely across the research community, while the costs of curation, updating, documentation, and user support tend to fall on a small number of contributors. As a result, many valuable data resources remain fragile, dependent on individual projects, short-term funding, or the continued priorities of external providers.

Existing infrastructure models address this problem only imperfectly. Commercial platforms integrate data and computation, but retain control over data models, update cycles, and long-run availability, exposing researchers to shifts in provider priorities and limiting transparency. Foundation-based infrastructure provides stable public goods, but is typically limited to narrow functions and does not support execution-ready analytical workflows. Institution-specific solutions, such as local computing environments or bespoke data pipelines, reduce some barriers but reproduce inequalities in access and are difficult to sustain or reuse beyond their original context.

The i3 BigQuery Workspace is being developed as an alternative approach centered on community stewardship of shared research infrastructure. The intended governance model combines voluntary contributions with compensated activities supported through grants, fellowships, and related funding mechanisms. This approach is designed to distribute responsibility for dataset curation, pipeline maintenance, documentation, and user support across contributors, while creating capacity to sustain labor-intensive or time-sensitive work as the workspace grows. By avoiding reliance on a single institution or provider, the goal is to support durable, cumulative research while limiting exposure to disruptions arising from institutional or commercial changes.

More broadly, the experience of building the i3 workspace highlights the feasibility of community-owned data infrastructure in fields characterized by large, shared datasets. The underlying economics are favorable: once data are hosted in a shared cloud environment, the marginal costs of storage and computation are modest relative to the fixed costs borne repeatedly by individual institutions through licensing fees, duplicated data pipelines, and fragmented infrastructure. The primary obstacles are therefore institutional rather than technical—aligning incentives for maintenance, recognizing contributions to shared data resources, and establishing governance arrangements that can persist over time.

5.2 Large Language Models and Data Infrastructure

Recent advances in large language models (LLMs) further increase the importance of shared, execution-ready research infrastructure. Many emerging applications in the economics of science and innovation rely on large-scale text analysis, including embedding generation, document classification, entity extraction, and retrieval-augmented analysis over corpora containing millions of scientific papers, patents, or administrative records. At this scale, the primary constraint is no longer model availability, but the ability to apply models efficiently to large, versioned datasets.

These workflows place a premium on compute locality—the co-location of data, models, and computation. Moving large corpora between storage systems, local machines, and external model endpoints quickly becomes infeasible, both technically and financially. Infrastructure that allows models to be trained or applied directly where the data reside lowers fixed costs and reduces the need for bespoke data pipelines. As a result, the traditional separation between data repositories and computational environments becomes increasingly untenable for LLM-based research.

The i3 BigQuery Workspace provides a foundation for this emerging mode of empirical work by combining shared, versioned data with scalable computation in a single environment. Intermediate outputs from LLM pipelines—such as embeddings, classifications, or extracted entities—can be stored alongside source data and reused by other researchers, reducing duplication of effort and improving reproducibility. By lowering the fixed costs of applying LLM-based methods at scale, shared infrastructure broadens access to these tools and helps ensure that advances in machine learning translate into cumulative scientific progress rather than isolated, institution-specific applications.

5.3 Open Questions

Several questions remain unresolved. First, sustainability: while the costs of maintaining shared infrastructure are modest by research standards, they still require committed, long-term support. Identifying funding arrangements that can sustain ongoing maintenance without reintroducing dependence on a small number of institutions remains an open issue. Second, governance: as shared resources grow, decisions about priorities, standards, and scope require collective input. Determining appropriate structures for decision-making and responsibility allocation is an area of ongoing development. The tools required to store, process, and analyze large datasets at scale already exist. The remaining questions concern how shared infrastructure should be funded and governed as it matures.

6 Conclusion

Large-scale datasets have expanded the scope of empirical research on science and innovation, but realizing their potential requires more than open access alone. Researchers continue to face computational barriers, provider dependence, reproducibility challenges, transparency gaps, and substantial inequality in access to infrastructure—constraints that individual research teams are poorly positioned to address on their own.

The i3 BigQuery Workspace illustrates one approach to organizing shared research infrastructure that mitigates these constraints. By hosting widely used datasets in a common cloud environment, the workspace lowers fixed costs to entry and reduces reliance on local computing resources. By maintaining independent, versioned copies of core datasets and open ingestion pipelines, it limits exposure to disruptions arising from changes in commercial or institutional providers. And by enabling researchers to execute identical queries on identical data, it supports replication and cumulative research at scale.

More broadly, the experience of the i3 community highlights the importance of institutional design in sustaining open research infrastructure. Technical tools for large-scale data storage and computation are now widely available, but their research value depends on governance arrangements that support shared maintenance, transparency, and long-run accessibility. When data pipelines and processing choices are open and collectively stewarded, maintaining research-grade datasets becomes a community activity rather than a private burden.

Looking ahead, these considerations will become increasingly salient as large language models and other data-intensive methods become central to empirical research. As compute locality and execution-ready data grow more important, infrastructure choices will shape not only the cost of research, but also who can participate and which questions can be studied at scale. Shared, community-owned infrastructure offers a path toward ensuring that advances in data and machine learning translate into cumulative scientific progress rather than fragmented, institution-specific capabilities.

References

George Abi Younes and Gaétan de Rassenfosse. Replicable Patent Indicators Using the Google Patents Public Datasets. *Australian Economic Review*, 57(1):102–113, 2024. doi: 10.1111/1467-8462.12545.

Jörg Ankel-Peters, Nathan Fiala, and Florian Neubauer. Do economists replicate? *Journal of Economic Behavior & Organization*, 212:219–232, August 2023. ISSN 0167-2681. doi: 10.1016/j.jebo.2023.05.009.

A Arora, S Belenzon, L Cioaca, L Sheer, HM Shin, and D Shvadron. Discern 2.0: Duke innovation & scientific enterprises research network [dataset]. *Zenodo (CERN European Organization for Nuclear Research). https://doi. org/10.5281/zenodo*, 3594642, 2024.

Kenneth J. Arrow. Economic welfare and the allocation of resources for invention. In *The Rate and Direction of Inventive Activity: Economic and Social Factors*, pages 609–626. Princeton University Press, 1962.

Sam Arts and Nicola Melluso. Reliance on scientific ideas in patenting. Working paper, 2025.

Enrico Berkes, Matthew Lee Chen, and Matteo Tranchero. 300 years of British patents. *Research Policy*, 55: 105347, 2026. doi: 10.1016/j.respol.2025.105347.

Barbara Biasi and Petra Moser. Effects of copyrights on science: Evidence from the WWII book republication program. *American Economic Journal: Microeconomics*, 13(4):218–260, 2021. doi: 10.1257/mic.20190113.

Michaël Bikard and Matt Marx. Bridging academia and industry: How geographic hubs connect university science and corporate technology. *Management Science*, 66:3425–3443, 2020. doi: 10.1287/mnsc.2019.3385.

Olivier Binette, Sokhna A. York, Emma Hickerson, Youngsoo Baek, Sarvo Madhav, and Christina Jones. Estimating the performance of entity resolution algorithms: Lessons learned through PatentsView. *The American Statistician*, 77(3):254–268, 2023. doi: 10.1080/00031305.2022.2153674.

Kevin A. Bryan and Yasin Ozcan. The impact of open access mandates on invention. Working paper, 2021.

Garret Christensen and Edward Miguel. Transparency, Reproducibility, and the Credibility of Economics Research. *Journal of Economic Literature*, 56(3):920–980, September 2018. ISSN 0022-0515. doi: 10.1257/jel.20171350.

Partha Dasgupta and Paul A. David. Toward a new economics of science. *Research Policy*, 23(5):487–521, 1994. doi: 10.1016/0048-7333(94)01002-1.

Gaëtan de Rassenfosse, Jan Kozak, and Florian Seliger. Geocoding of worldwide patent data. *Scientific Data*, 6:260, 2019. doi: 10.1038/s41597-019-0264-6.

Liran Einav and Jonathan Levin. Economics in the age of big data. *Science*, 346(6210):1243089, November 2014. doi: 10.1126/science.1243089.

Michael Ewens and Matt Marx. Firm age and invention: An open access dataset. Technical report, Working Paper, 2023. URL https://foundingpatents.com/. Available at: https://foundingpatents.com/.

Joel Ferguson, Rebecca Littman, Garret Christensen, Elizabeth Levy Paluck, Nicholas Swanson, Zenan Wang, Edward Miguel, David Birke, and John-Henry Pezzuto. Survey of open science practices and attitudes in the social sciences. *Nature Communications*, 14(1):5401, September 2023. ISSN 2041-1723. doi: 10.1038/s41467-023-41111-1.

Miloš Fišar, Ben Greiner, Christoph Huber, Elena Katok, and Ali I. Ozkes. Reproducibility in Management Science. *Management Science*, 70(3):1343–1356, March 2024. ISSN 0025-1909. doi: 10.1287/mnsc.2023.03556.

Santo Fortunato, Carl T. Bergstrom, Katy Börner, James A. Evans, Dirk Helbing, Staša Milojević, Alexander M. Petersen, Filippo Radicchi, Roberta Sinatra, Brian Uzzi, Alessandro Vespignani, Ludo Waltman, Dashun Wang, and Albert-László Barabási. Science of science. *Science*, 359(6379):eaao0185, 2018. doi: 10.1126/science.aao0185.

Jeffrey L. Furman and Scott Stern. Climbing atop the shoulders of giants: The impact of institutions on cumulative research. *American Economic Review*, 101(5):1933–1963, 2011. doi: 10.1257/aer.101.5.1933.

Alberto Galasso and Mark Schankerman. Patents and cumulative innovation: Causal evidence from the courts. *Quarterly Journal of Economics*, 130(1):317–369, 2015. doi: 10.1093/qje/qju029.

Gerard George, Ernst C. Osinga, Dovev Lavie, and Brent A. Scott. Big Data and Data Science Methods for Management Research. *Academy of Management Journal*, 59(5):1493–1507, October 2016. ISSN 0001-4273. doi: 10.5465/amj.2016.4005.

Daniel P. Gross and Bhaven N. Sampat. The economics of knowledge production: Evidence from World War II scientists and engineers. *American Economic Review*, 112(10):3244–3282, 2022.

Robert L. Grossman. Ten lessons for data sharing with a data commons. *Scientific Data*, 10:120, 2023. doi: 10.1038/s41597-023-02029-x.

Michael A. Heller and Rebecca S. Eisenberg. Can patents deter innovation? The anticommons in biomedical research. *Science*, 280(5364):698–701, 1998.

Sylvérie Herbert, Hautahi Kingi, Flavio Stanchi, and Lars Vilhuber. Reproduce to validate: A comprehensive study on the reproducibility of economics research. *Canadian Journal of Economics*, 57(3):961–988, 2024. doi: 10.1111/caje.12728.

Daniel W. Hook and Simon J. Porter. Scaling Scientometrics: Dimensions on Google BigQuery as an Infrastructure for Large-Scale Analysis. *Frontiers in Research Metrics and Analytics*, 6, April 2021. ISSN 2504-0537. doi: 10.3389/frma.2021.656233.

IFI CLAIMS Patent Services and Digital Science. Leading patent data platform ifi CLAIMS joins Digital Science. https://www.ificlaims.com/news/leading-patent-data-platform-ifi-claims-joins-digital-science/, 2021. Accessed 2025-12-29.

Benjamin F. Jones. The burden of knowledge and the "death of the renaissance man": Is innovation getting harder? *Review of Economic Studies*, 76(1):283–317, 2009. doi: 10.1111/j.1467-937X.2008.00531.x.

Mary Kaltenberg, Adam B. Jaffe, and Margie E. Lachman. Invention and the life course: Age differences in patenting. *Research Policy*, 52(1):104629, 2023. doi: 10.1016/j.respol.2022.104629. URL https://doi.org/10.1016/j.respol.2022.104629.

Leonid Kogan, Dimitris Papanikolaou, Amit Seru, and Noah Stoffman. Technological innovation, resource allocation, and growth. *The Quarterly Journal of Economics*, 132(2):665–712, 2017. doi: 10.1093/qje/qjw040.

Jeffrey M. Kuhn and Neil Thompson. How to measure and draw causal inferences with patent scope. *International Journal of the Economics of Business*, 26(1):5–38, 2019. doi: 10.1080/13571516.2018.1553284. URL https://doi.org/10.1080/13571516.2018.1553284.

Josh Lerner, Amit Seru, Nick Short, and Yuan Sun. Financial innovation in the 21st century: Evidence from U.S. patents. Working Paper 28980, National Bureau of Economic Research, July 2021.

Mpho Mafata, Ian D. van der Linde, Jos J. Winnink, and Robert J.W. Tijssen. Comparison of the coverage of the USPTO's PatentsView and the EPO's PATSTAT patent databases: A reproducibility case study of the USPTO general patent statistics reports. Preprint, Research Square, 2024.

Matt Marx. Connecting undamental science to commercial innovation: a public dataset of patent-paper pairs. Technical report, Working Paper, 2025.

Matt Marx. Connecting fundamental science to commercial innovation: A public dataset of patent-paper pairs. Working paper, 2026.

Matt Marx and Aaron Fuegi. Reliance on science: Worldwide front-page patent citations to scientific articles. *Strategic Management Journal*, n/a(n/a), 2020. ISSN 1097-0266. doi: 10.1002/smj.3145.

Matt Marx and Aaron Fuegi. Reliance on science by inventors: Hybrid extraction of in-Text patent-to-article citations. *Journal of Economics & Management Strategy*, 31(2):369–392, 2022. ISSN 1530-9134. doi: 10.1111/jems.12455.

Roger Masclans, Sharique Hasan, and Wesley M. Cohen. Measuring the commercial potential of science. *Strategic Management Journal*, 46:2199–2236, 2025a. doi: 10.1002/smj.3720.

Roger Masclans, Sharique Hasan, and Wesley M Cohen. Measuring the commercial potential of science. *Strategic Management Journal*, 46(9):2199–2236, 2025b.

Grant R. McDermott and Benjamin Hansen. Labor reallocation and remote work during COVID-19: Real-time evidence from GitHub. Working Paper 29598, National Bureau of Economic Research, December 2021.

Fiona Murray, Philippe Aghion, Mathias Dewatripont, Julian Kolev, and Scott Stern. Of mice and academics: Examining the effect of openness on innovation. *American Economic Journal: Economic Policy*, 8(1): 212–252, 2016. doi: 10.1257/pol.20140062.

Abhishek Nagaraj. The Private Impact of Public Data: Landsat Satellite Maps Increased Gold Discoveries and Encouraged Entry. *Management Science*, 68(1):564–582, January 2022. ISSN 0025-1909. doi: 10.1287/mnsc.2020.3878.

Abhishek Nagaraj, Esther Shears, and Mathijs de Vaan. Improving data access democratizes and diversifies science. *Proceedings of the National Academy of Sciences*, 117(38):23490–23498, 2020. doi: 10.1073/pnas.2001682117.

Richard R. Nelson. The simple economics of basic scientific research. *Journal of Political Economy*, 67(3): 297–306, 1959.

Sergio Petralia, Pierre-Alexandre Balland, and David Rigby. HistPat: Historical patent data, 2016.

Jason Priem, Heather Piwowar, and Richard Orr. OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts, May 2022.

Paul M. Romer. Endogenous technological change. *Journal of Political Economy*, 98(5):S71–S102, 1990.

Bhaven Sampat and Heidi L. Williams. How do patents affect follow-on innovation? Evidence from the human genome. *American Economic Review*, 109(1):203–236, 2019. doi: 10.1257/aer.20151398.

E Scharfmann, M Marx, and L Fleming. Pasteur's quadrant researchers bring novelty, impact to publishing, and patenting. *Science*, 390(6776):891–893, 2025.

Dror Shvadron, Hansen Zhang, Lee Fleming, and Daniel P. Gross. Funding the U.S. Scientific Training Ecosystem: New Data, Methods, and Evidence, 2025a.

Dror Shvadron, Hansen Zhang, Lee Fleming, and Daniel P. Gross. A Quarter of US-Trained Scientists Eventually Leave. Is the US Giving Away Its Edge?, 2025b.

Paula E. Stephan. How economics shapes science. 2012.

Carol Tenopir, Natalie M. Rice, Suzie Allard, Lynn Baird, Josh Bober, et al. Perceived benefits of open data are improving but scientists still lack resources, skills, and rewards. *Humanities and Social Sciences Communications*, 10:431, 2023. doi: 10.1057/s41599-023-01831-7.

Hal R. Varian. Big Data: New Tricks for Econometrics. *Journal of Economic Perspectives*, 28(2):3–28, May 2014. ISSN 0895-3309. doi: 10.1257/jep.28.2.3.

Michael Webb, Nick Short, Nicholas Bloom, and Josh Lerner. Some facts of high-tech patenting. Working Paper 24793, National Bureau of Economic Research, July 2018.

Heidi L. Williams. Intellectual property rights and innovation: Evidence from the human genome. *Journal of Political Economy*, 121(1):1–27, 2013. doi: 10.1086/669706.

Stefan Wuchty, Benjamin F. Jones, and Brian Uzzi. The increasing dominance of teams in production of knowledge. *Science*, 316(5827):1036–1039, 2007. doi: 10.1126/science.1136099.

Appendices

A Accessing the i3 BigQuery Workspace

This appendix provides a brief technical overview of the i3 BigQuery Workspace, including access instructions, available datasets, cost structure, and example queries. Full documentation is maintained at https://i3open.org/bigquery.html.

Access and Setup

The workspace is hosted on Google BigQuery under the project identifier `nber-i3`. To access:

1. Create or sign into a Google Cloud Console account with billing enabled.
2. Open BigQuery from the navigation menu.
3. Click **Add** → **Star a project by name** and enter `nber-i3`.

New Google Cloud users receive complimentary credits sufficient for initial analysis.

Pricing

BigQuery uses a consumption-based pricing model. Users are charged based on the volume of data scanned by each query (approximately $6.25 per terabyte processed), not the size of results returned. Storage costs for datasets hosted in the workspace are covered by the i3 initiative. Several practices help minimize costs:

- Select only the columns needed rather than using `SELECT *`, since BigQuery's columnar storage charges per column accessed.
- Filter on partitioned columns (e.g., year or date) to limit the data scanned.
- Use the query validator to preview estimated costs before execution.
- Identical queries run within 24 hours return cached results at no additional cost.
- Use `TABLESAMPLE SYSTEM (10 PERCENT)` to sample data during exploration.

Example Query

The following SQL query retrieves patent citations where the cited assignee is Xerox, scanning approximately 10 GB at a cost of roughly $0.015:

```
SELECT
    citing_patent_id,
    cited_patent_id,
    citation_date
FROM
    ‘nber-i3.patentsview.citations‘
WHERE
    cited_assignee = ’Xerox’
```

Integration with Python and R

BigQuery can be accessed programmatically. In Python:

```
from google.cloud import bigquery
client = bigquery.Client(project="your-project-id")
query = """
    SELECT * FROM ‘nber-i3.openalex.works‘
    WHERE publication_year = 2023
    LIMIT 1000
"""
df = client.query(query).to_dataframe()
```

In R:

```
library(bigrquery)
sql <- "SELECT * FROM ‘nber-i3.openalex.works‘
        WHERE publication_year = 2023
        LIMIT 1000"
results <- bq_project_query("your-project-id", sql)
df <- bq_table_download(results)
```

Native Jupyter notebook support is also available directly within the Google Cloud Console.

Contributing Datasets

Community members can contribute datasets to the workspace by uploading data to their own Google Cloud project, sharing it publicly, and contacting the i3 team with a description, project identifier, dataset name, intended applications, and relevant citations. Accepted datasets are reviewed and cloned into the main repository. The i3 community discussion group is available at https://groups.google.com/g/i3-bigquery.