Open Datasets & Infrastructure at Scale

NBER Economics of Science Conference
September 2025

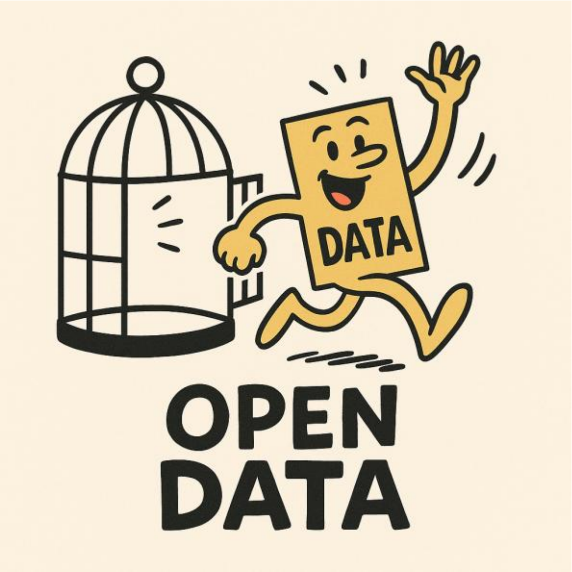
Dror Shvadron

University of Toronto

Matt Marx

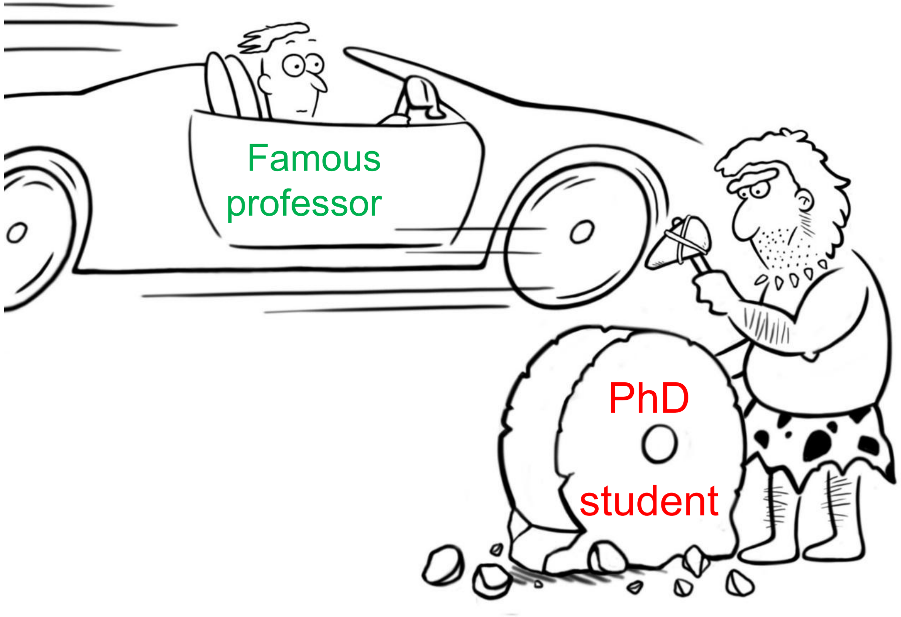
Cornell/NBER/Innovation Information Initiative (i3)

Why Open Data?

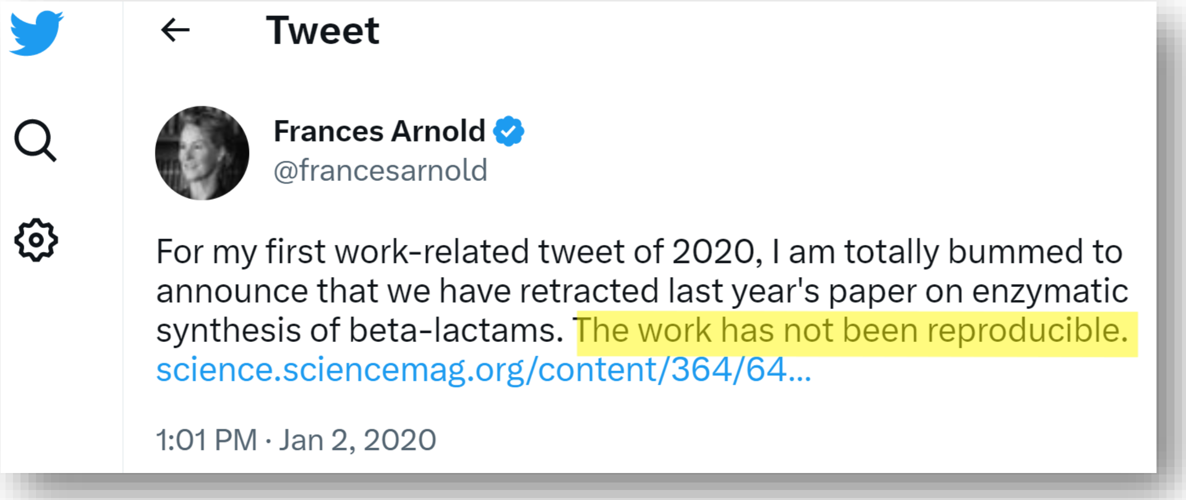


- Leveling
- Legitimacy
- Leverage

Open Data: Leveling



Open Data: Legitimacy

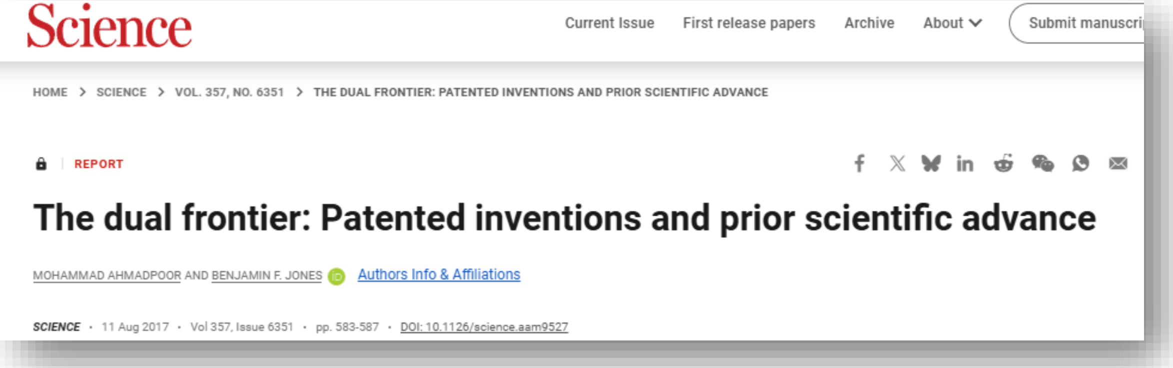


Frances Arnold
@francesarnold

For my first work-related tweet of 2020, I am totally bummed to announce that we have retracted last year's paper on enzymatic synthesis of beta-lactams. The work has not been reproducible. science.sciencemag.org/content/364/64...

1:01 PM · Jan 2, 2020

Open Data: Leverage



HOME > SCIENCE > VOL. 357, NO. 6351 > THE DUAL FRONTIER: PATENTED INVENTIONS AND PRIOR SCIENTIFIC ADVANCE

REPORT

The dual frontier: Patented inventions and prior scientific advance

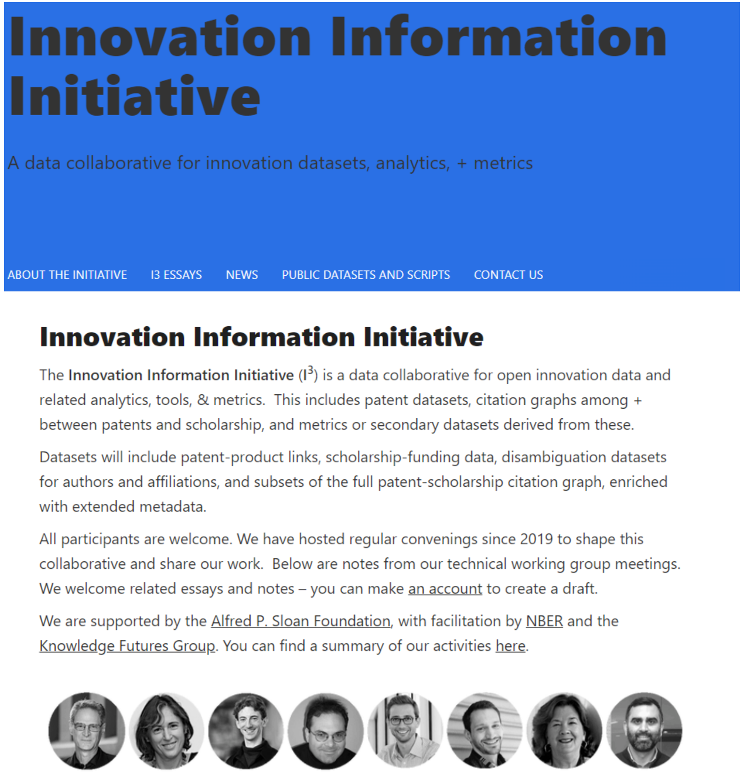
MOHAMMAD AHMADPOOR AND BENJAMIN F. JONES Authors Info & Affiliations

SCIENCE • 11 Aug 2017 • Vol 357, Issue 6351 • pp. 583-587 • DOI: 10.1126/science.aam9527

- Article cited 404 times
- But, dataset built using Web of Science, only accessible to NU faculty
- For comparison: 414 articles using i3 patent-to-paper citations (relianceonscience.org)

Open Data for the Economics of Science

ALFRED P. SLOAN
FOUNDATION



Innovation Information Initiative

A data collaborative for innovation datasets, analytics, + metrics

Innovation Information Initiative

The **Innovation Information Initiative** (I^3) is a data collaborative for open innovation data and related analytics, tools, & metrics. This includes patent datasets, citation graphs among + between patents and scholarship, and metrics or secondary datasets derived from these.

Datasets will include patent-product links, scholarship-funding data, disambiguation datasets for authors and affiliations, and subsets of the full patent-scholarship citation graph, enriched with extended metadata.

All participants are welcome. We have hosted regular convenings since 2019 to shape this collaborative and share our work. Below are notes from our technical working group meetings. We welcome related essays and notes – you can make an account to create a draft.

We are supported by the Alfred P. Sloan Foundation, with facilitation by NBER and the Knowledge Futures Group. You can find a summary of our activities here.

Open Data for the Economics of Science

I3 Open Innovation Dataset Index

>100 Open Datasets

Crossref

0.5 TB

Google Patents

2.6 TB

OpenAlex

1.5 TB

USPTO PatentsView

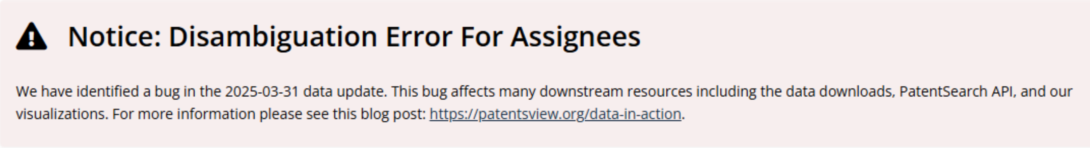
Dozens of files
>10GB each

New challenges?

Gap Between Open Data and Enabling of Research

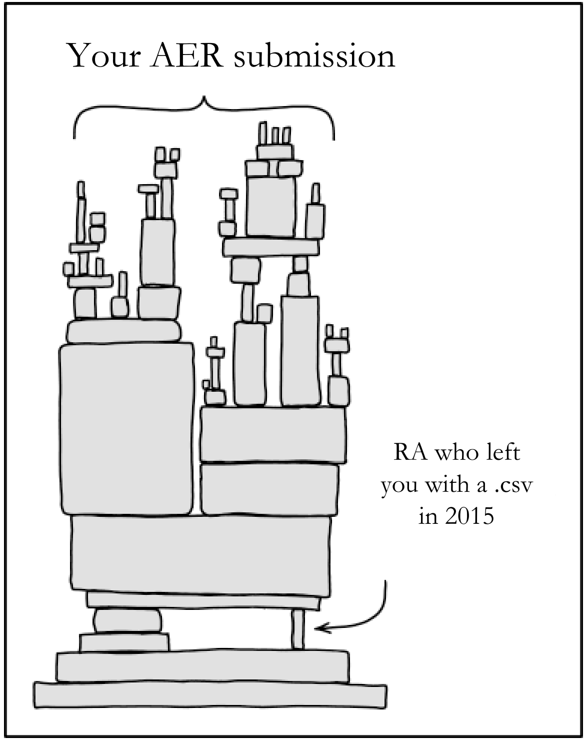
Large datasets are hard to work with

- Downloading and storing
- Merging with other datasets
- Keeping track of periodical updates
- Being aware of errors
- Imbalance in computational resources



⚠ **Notice: Disambiguation Error For Assignees**

We have identified a bug in the 2025-03-31 data update. This bug affects many downstream resources including the data downloads, PatentSearch API, and our visualizations. For more information please see this blog post: https://patentsview.org/data-in-action.



Proposed solution: i3 BigQuery Workspace

A shared platform for data storage and analysis

- Host a set of curated datasets
- Platform that enables analysis at scale on the cloud, integrated with advanced tools
- Cost-efficient, shared resources
- Industry standard methods and processes
- Sustainable through funding ecosystem

You might think: But we have Zenodo/Dataverse/Dropbox etc… What's the difference?

What is Google BigQuery?

Google Cloud’s data analysis platform:

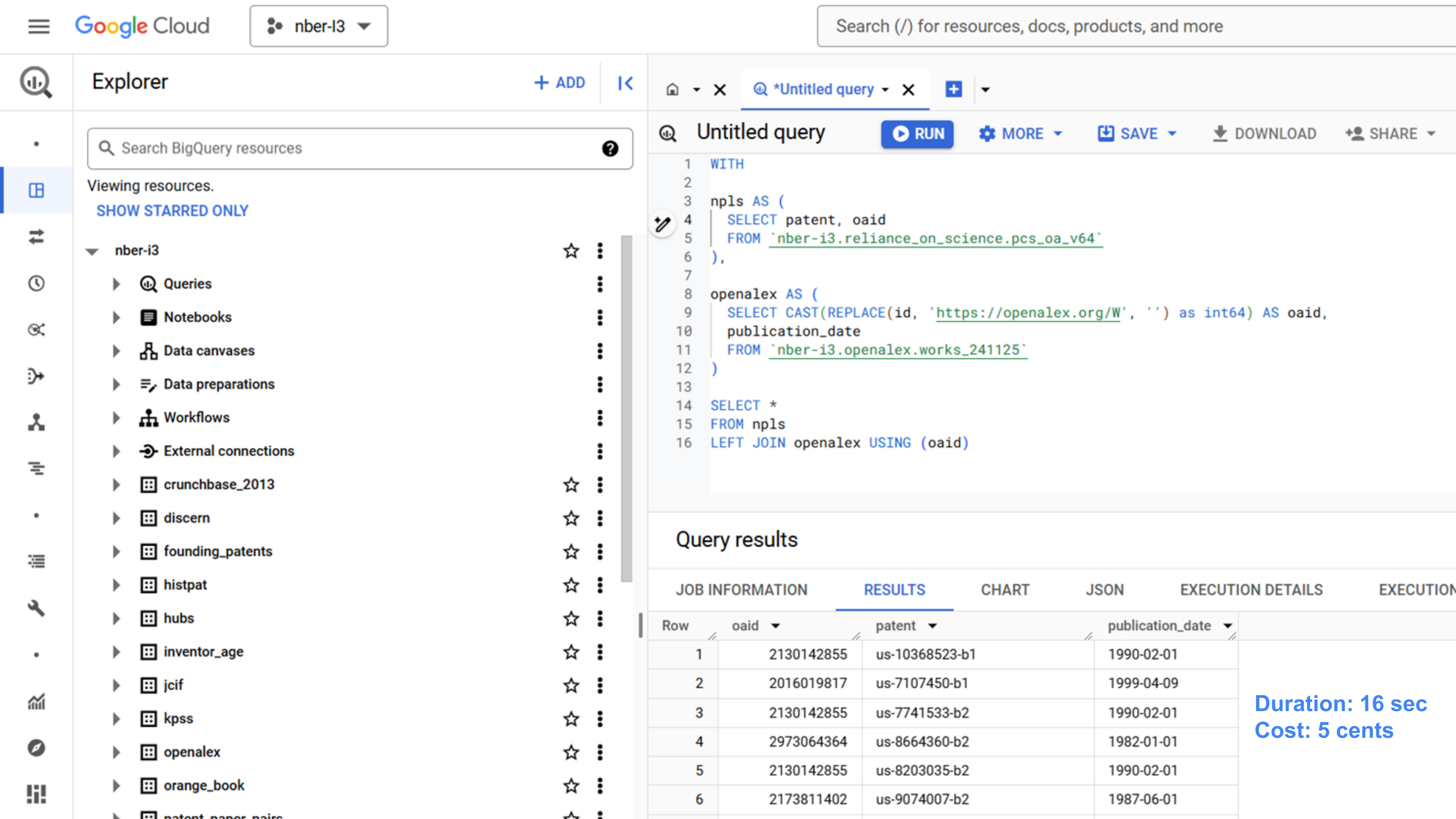
Scalable: Handles data from megabytes to terabytes effortlessly.

Interoperable: SQL-based, integrates seamlessly with Python, R, Julia (sorry Stata)

- Can always export data for local analysis

Collaborative: Shared code+data for coauthors, w/version control.

Requires Google Cloud billing account. (New users get free credits.)



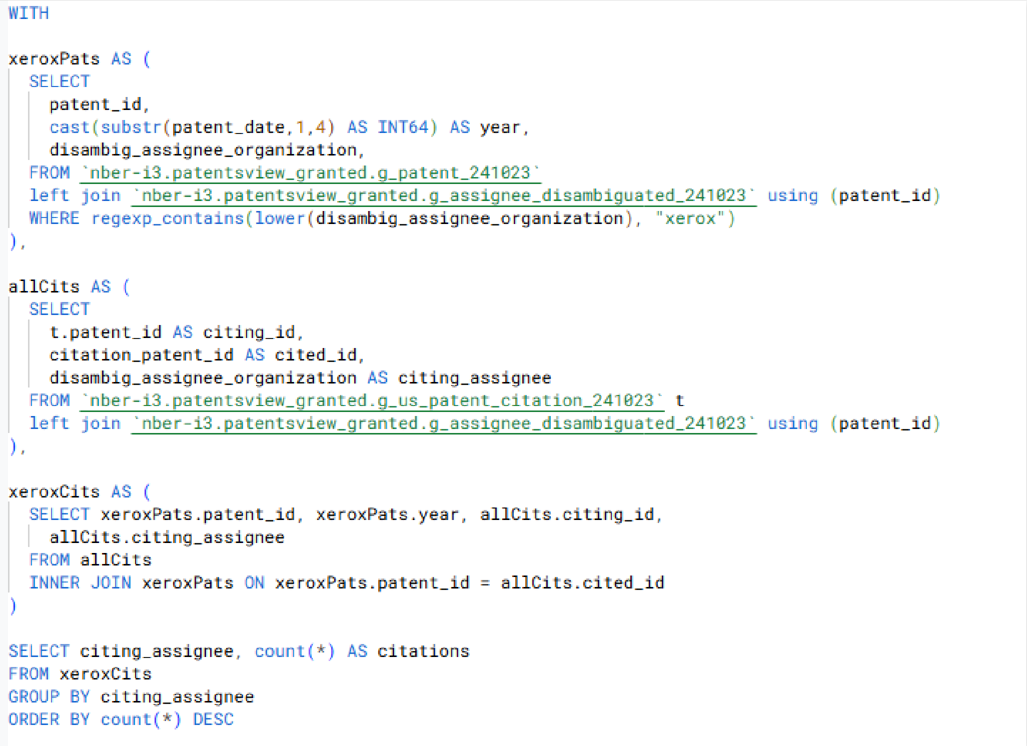
```
WITH

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)
```

Row	oaid	patent	publication_date
1	2130142855	us-10368523-b1	1990-02-01
2	2016019817	us-7107450-b1	1999-04-09
3	2130142855	us-7741533-b2	1990-02-01
4	2973064364	us-8664360-b2	1982-01-01
5	2130142855	us-8203035-b2	1990-02-01
6	2173811402	us-9074007-b2	1987-06-01

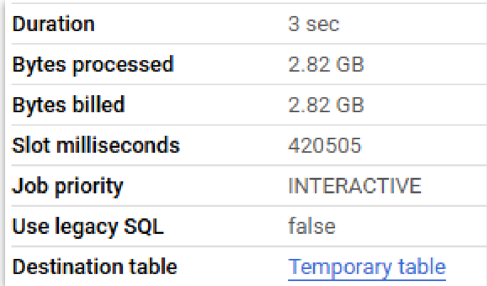
Duration: 16 sec
Cost: 5 cents

Example: Who cites Xerox patents?

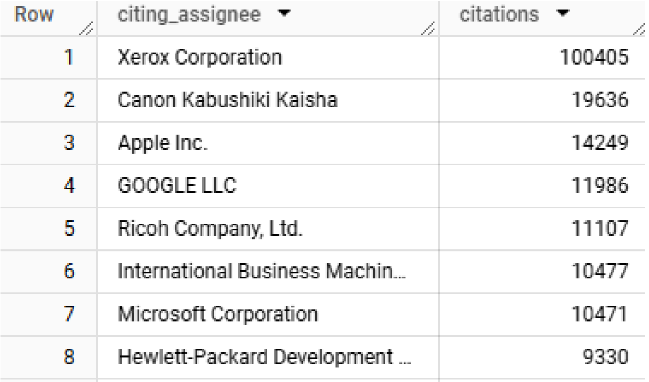
Cost: 1.5 cents



```sql
WITH

xeroxPats AS (
  SELECT
    patent_id,
    cast(substr(patent_date,1,4) AS INT64) AS year,
    disambig_assignee_organization,
  FROM `nber-i3.patentsview_granted.g_patent_241023`
  left join `nber-i3.patentsview_granted.g_assignee_disambiguated_241023` using (patent_id)
  WHERE regexp_contains(lower(disambig_assignee_organization), "xerox")
),

allCits AS (
  SELECT
    t.patent_id AS citing_id,
    citation_patent_id AS cited_id,
    disambig_assignee_organization AS citing_assignee
  FROM `nber-i3.patentsview_granted.g_us_patent_citation_241023` t
  left join `nber-i3.patentsview_granted.g_assignee_disambiguated_241023` using (patent_id)
),

xeroxCits AS (
  SELECT xeroxPats.patent_id, xeroxPats.year, allCits.citing_id,
    allCits.citing_assignee
  FROM allCits
  INNER JOIN xeroxPats ON xeroxPats.patent_id = allCits.cited_id
)

SELECT citing_assignee, count(*) AS citations
FROM xeroxCits
GROUP BY citing_assignee
ORDER BY count(*) DESC
```



Duration	3 sec
Bytes processed	2.82 GB
Bytes billed	2.82 GB
Slot milliseconds	420505
Job priority	INTERACTIVE
Use legacy SQL	false
Destination table	Temporary table

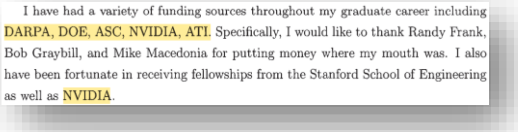


Row	citing_assignee ▾	citations ▾
1	Xerox Corporation	100405
2	Canon Kabushiki Kaisha	19636
3	Apple Inc.	14249
4	GOOGLE LLC	11986
5	Ricoh Company, Ltd.	11107
6	International Business Machin...	10477
7	Microsoft Corporation	10471
8	Hewlett-Packard Development ...	9330

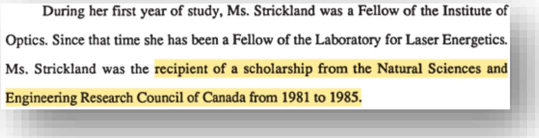
Funding the U.S. Scientific Training Ecosystem: New Data, Methods, and Evidence

Joint work with Hansen Zhang, Lee Fleming and Dan Gross

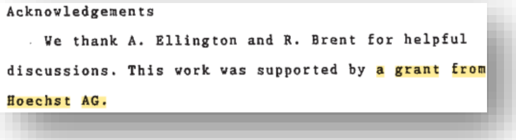
103 million sentences



I have had a variety of funding sources throughout my graduate career including DARPA, DOE, ASC, NVIDIA, ATI. Specifically, I would like to thank Randy Frank, Bob Graybill, and Mike Macedonia for putting money where my mouth was. I also have been fortunate in receiving fellowships from the Stanford School of Engineering as well as NVIDIA.



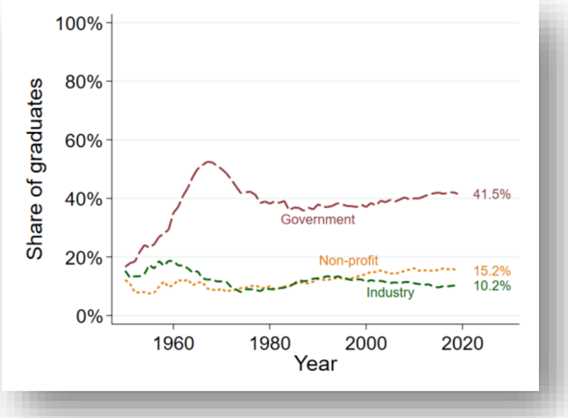
During her first year of study, Ms. Strickland was a Fellow of the Institute of Optics. Since that time she has been a Fellow of the Laboratory for Laser Energetics. Ms. Strickland was the recipient of a scholarship from the Natural Sciences and Engineering Research Council of Canada from 1981 to 1985.



Acknowledgements

We thank A. Ellington and R. Brent for helpful discussions. This work was supported by a grant from Hoechst AG.

Processing using LLMs

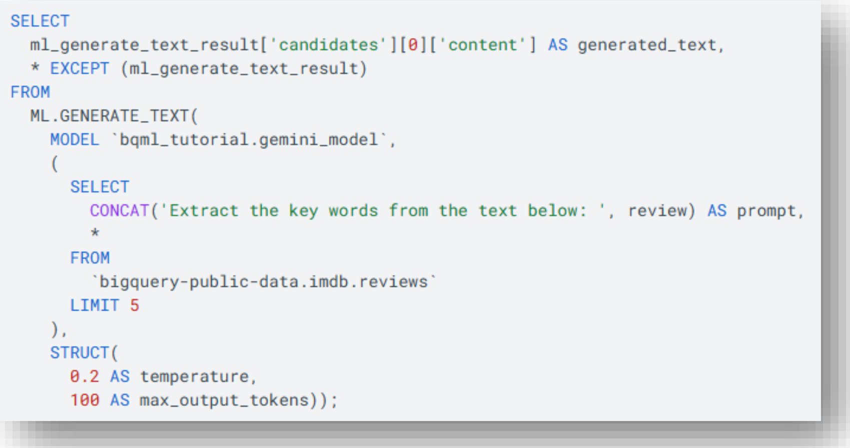


NBER working paper https://www.nber.org/papers/w33944

Aggregate data available through doi:10.7910/DVN/LSHUB1

Advanced features: Generative AI

- LLMs are integrated into Bigquery
- Run prompts at scale over large datasets
- Choose from a variety of models



```
SELECT
  ml_generate_text_result['candidates'][0]['content'] AS generated_text,
  * EXCEPT (ml_generate_text_result)
FROM
  ML.GENERATE_TEXT(
    MODEL `bqml_tutorial.gemini_model`,
    (
      SELECT
        CONCAT('Extract the key words from the text below: ', review) AS prompt,
        *
      FROM
        `bigquery-public-data.imdb.reviews`
      LIMIT 5
    ),
    STRUCT(
      0.2 AS temperature,
      100 AS max_output_tokens));
```

Gemini

ANTHROP\C

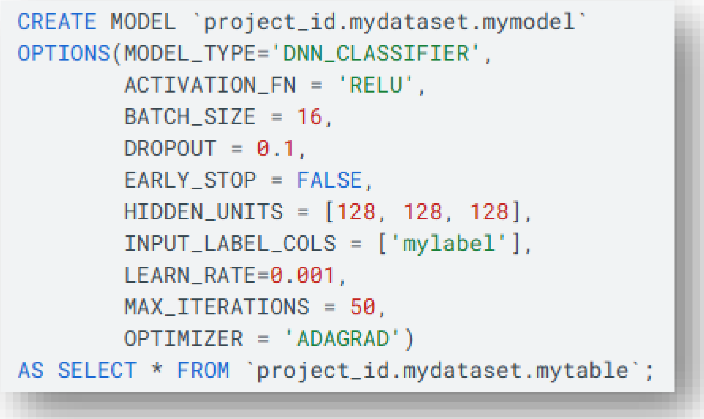
Meta

Hugging Face

Mistral AI

Advanced features: Machine Learning

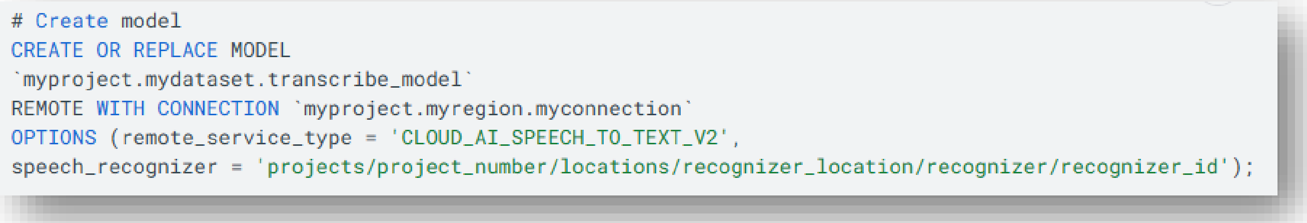
- Train classification models directly on the data
- Multiple types of models:
  - Regressions
  - Boosted tree
  - Random forest
  - Deep neural network
- Inference at scale



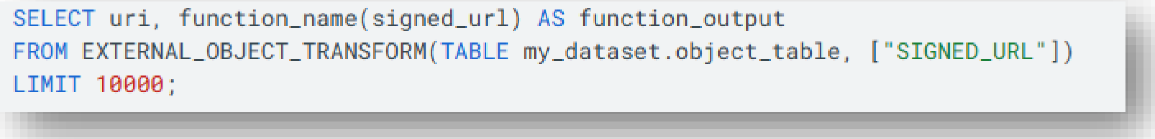
```
CREATE MODEL `project_id.mydataset.mymodel`
OPTIONS(MODEL_TYPE='DNN_CLASSIFIER',
        ACTIVATION_FN = 'RELU',
        BATCH_SIZE = 16,
        DROPOUT = 0.1,
        EARLY_STOP = FALSE,
        HIDDEN_UNITS = [128, 128, 128],
        INPUT_LABEL_COLS = ['mylabel'],
        LEARN_RATE=0.001,
        MAX_ITERATIONS = 50,
        OPTIMIZER = 'ADAGRAD')
AS SELECT * FROM `project_id.mydataset.mytable`;
```

Advanced features: Object Tables

- Create secure connections between Bigquery and unstructured data objects in the cloud (PDFs, images, etc…)
- Run advanced analyses using predefined models and custom cloud functions



```
# Create model
CREATE OR REPLACE MODEL
`myproject.mydataset.transcribe_model`
REMOTE WITH CONNECTION `myproject.myregion.myconnection`
OPTIONS (remote_service_type = 'CLOUD_AI_SPEECH_TO_TEXT_V2',
speech_recognizer = 'projects/project_number/locations/recognizer_location/recognizer/recognizer_id');
```



```
SELECT uri, function_name(signed_url) AS function_output
FROM EXTERNAL_OBJECT_TRANSFORM(TABLE my_dataset.object_table, ["SIGNED_URL"])
LIMIT 10000;
```

The i3 BigQuery workspace

Raw datasets

Openalex

USPTO PatentsView
Learn About Patents Around the World

crunchbase

(2013)

Derivative datasets

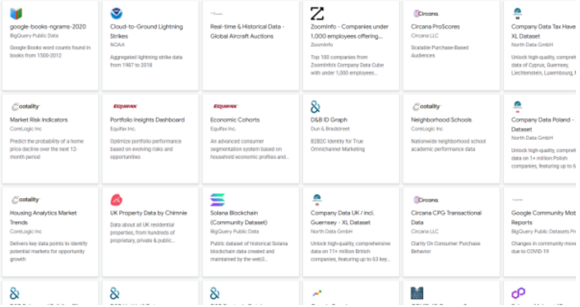
- Commercial Potential of Science
- DISCERN
- Founding Patents (assignee age)
- HistPat
- Hubs of Invention
- Inventor Age
- Journal Commercial Impact Factor
- KPSS patent value
- Patent Paper Pairs
- Patent Scope
- Reliance on Science

Seamless Integration with other BigQuery datasets

Google Patents

Dimensions
A **Digital Science** Solution

Marketplace



google-books-ngrams-2020
BigQuery Public Data
Google Books word counts found in books from 1500-2012

Cloud-to-Ground Lightning Strikes
NOAA
Aggregated lightning strike data from 1987 to 2018

Real-time & Historical Data - Global Aircraft Auctions

ZoomInfo - Companies under 1,000 employees offering...
ZoomInfo
Top 100 companies from ZoomInfo's Company Data Cube with under 1,000 employees...

Circana ProScores
Circana LLC
Scalable Purchase-Based Audiences

Company Data Tax Have
XL Dataset
North Data GmbH
Unlock high-quality, compreh
data of Cyprus, Guernsey,
Liechtenstein, Luxembourg,

Market Risk Indicators
CoreLogic Inc
Predict the probability of a home price decline over the next 12-month period

Portfolio Insights Dashboard
Equifax Inc.
Optimize portfolio performance based on evolving risks and opportunities

Economic Cohorts
Equifax Inc.
An advanced consumer segmentation system based on household economic profiles and...

D&B ID Graph
Dun & Bradstreet
B2B2C Identity for True Omnichannel Marketing

Neighborhood Schools
CoreLogic Inc
Nationwide neighborhood school academic performance data

Company Data Poland -
Dataset
North Data GmbH
Unlock high-quality, compreh
data on 1+ million Polish
companies, featuring up to 6

Housing Analytics Market Trends
CoreLogic Inc
Delivers key data points to identify potential markets for opportunity growth

UK Property Data by Chimnie
Data about all UK residential properties, from hundreds of proprietary, private & public...

Solana Blockchain (Community Dataset)
BigQuery Public Data
Public dataset of historical Solana blockchain data created and maintained by the web3...

Company Data UK / incl. Guernsey - XL Dataset
North Data GmbH
Unlock high-quality, comprehensive data on 11+ million British companies, featuring up to 63 key...

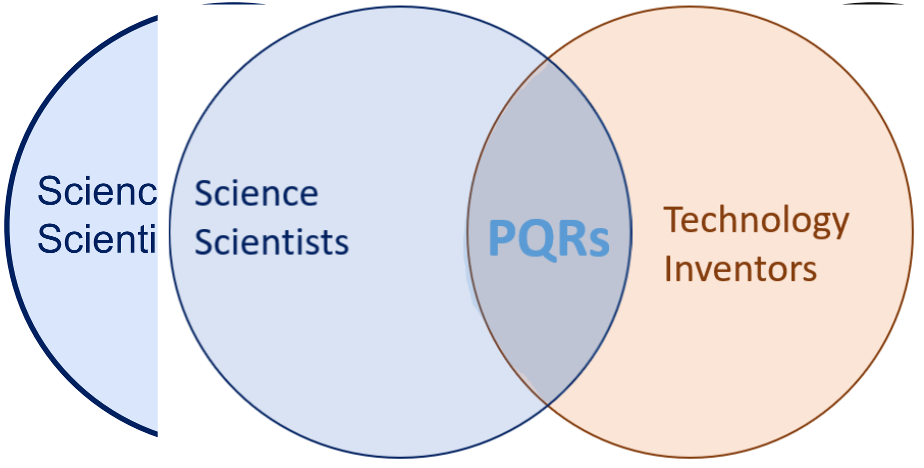
Circana CPG Transactional Data
Circana LLC
Clarity On Consumer Purchase Behavior

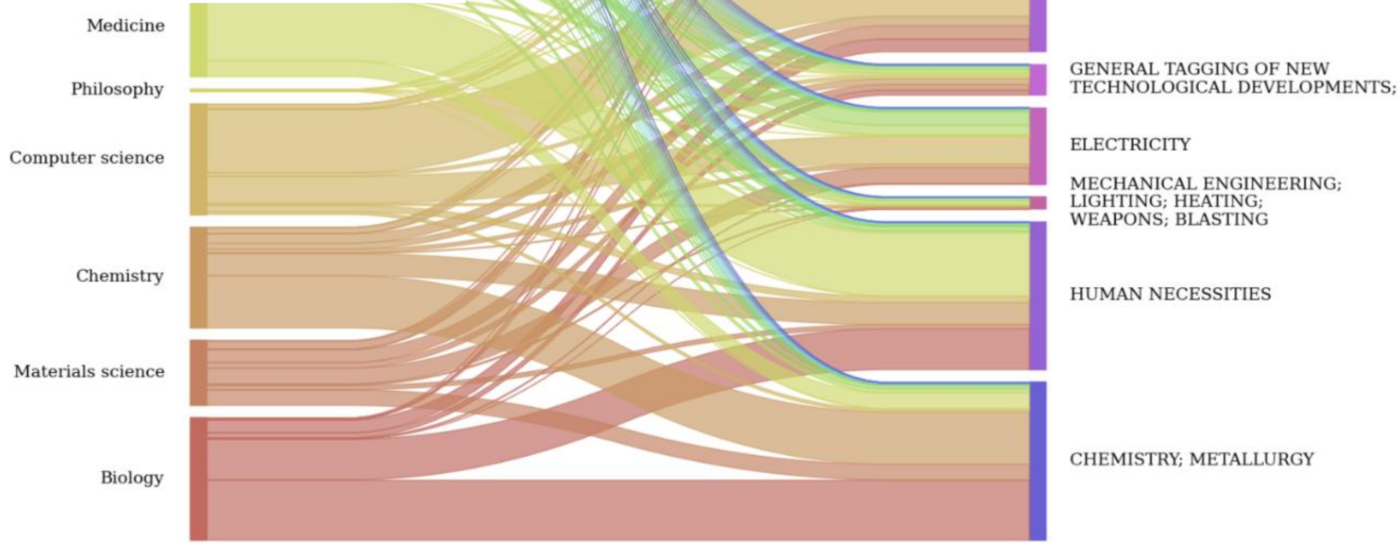
Google Community Mob
Reports
BigQuery Public Datasets Pr
Changes in community mov
due to COVID-19

Pasteur's Quadrant Researchers (PQRs)

Joint work with Emma Scharfmann and Lee Fleming

ALFRED P. SLOAN FOUNDATION

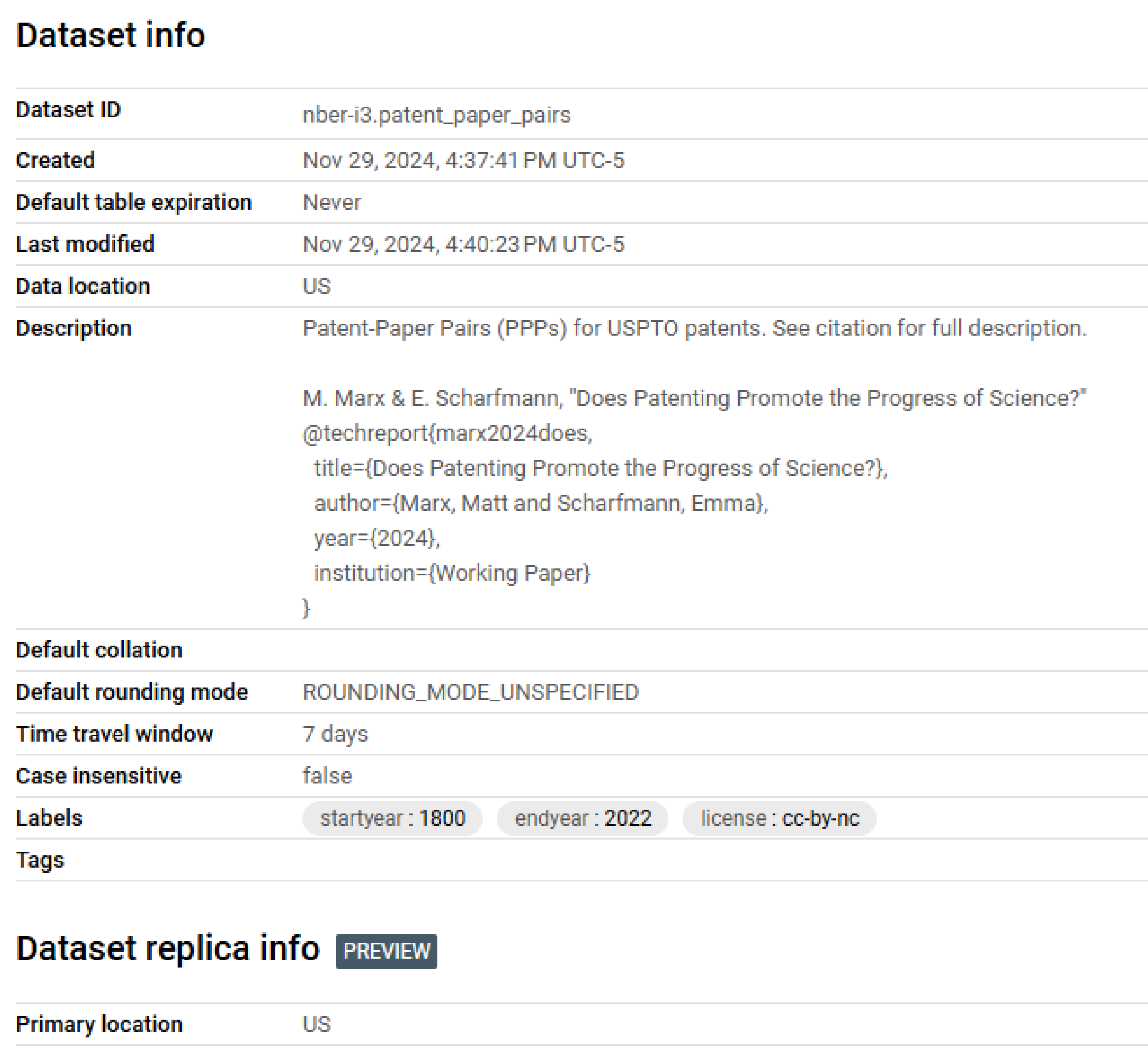




- 582,199 PQRs
  from 1976-2023
- Confidence %
  for each PQR
- Also download:
  relianceonscience.org

Notes

- Datasets include variable definitions, year ranges, etc.
- Pls use provided cites, observe license type
- “i3-bigquery” Google Group for updates



Dataset info

Dataset ID	nber-i3.patent_paper_pairs
Created	Nov 29, 2024, 4:37:41 PM UTC-5
Default table expiration	Never
Last modified	Nov 29, 2024, 4:40:23 PM UTC-5
Data location	US

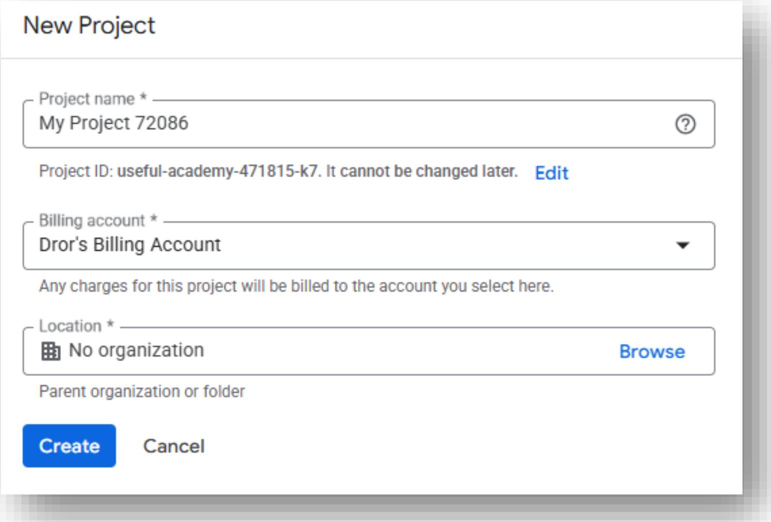
Description	Patent-Paper Pairs (PPPs) for USPTO patents. See citation for full description.
	
	M. Marx & E. Scharfmann, "Does Patenting Promote the Progress of Science?"
	@techreport{marx2024does,
	title={Does Patenting Promote the Progress of Science?},
	author={Marx, Matt and Scharfmann, Emma},
	year={2024},
	institution={Working Paper}
	}

Default collation	
Default rounding mode	ROUNDING_MODE_UNSPECIFIED
Time travel window	7 days
Case insensitive	false
Labels	startyear : 1800 endyear : 2022 license : cc-by-nc
Tags	

Dataset replica info PREVIEW

Primary location	US

Create your own workspace, integrate with i3 and others



New Project

Project name * My Project 72086

Project ID: useful-academy-471815-k7. It cannot be changed later. Edit

Billing account * Dror's Billing Account

Any charges for this project will be billed to the account you select here.

Location * No organization

Parent organization or folder

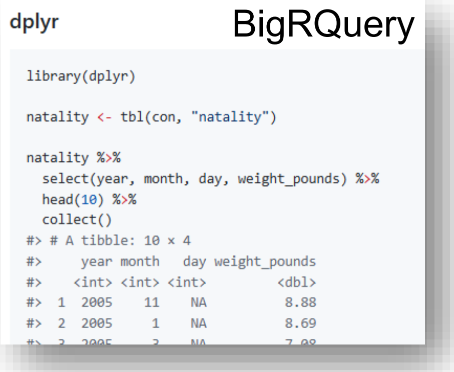
- Users set up their own projects (workspaces)
- Pay for analyses based on use
- Pay for personal storage of datasets
- Full, seamless integration with i3 and other datasets

What about replication packages?

- Nothing beats local self-contained replication packages…

  … unless huge datasets are involved!

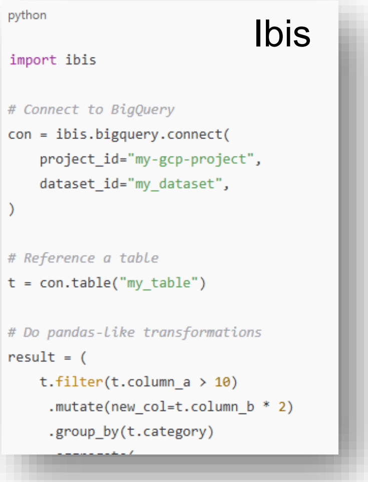
- Multiple ways to integrate BigQuery queries into standard workflows.
- Potential to make replication packages easier to follow and share with editors.
- We discussed these ideas with the data editor at AEA



dplyr

BigRQuery

```
library(dplyr)

natality <- tbl(con, "natality")

natality %>%
  select(year, month, day, weight_pounds) %>%
  head(10) %>%
  collect()
#> # A tibble: 10 × 4
#>     year month   day weight_pounds
#>    <int> <int> <int>         <dbl>
#>  1  2005    11    NA          8.88
#>  2  2005     1    NA          8.69
#>  3  2005     3    NA          7.08
```



python

Ibis

```
import ibis

# Connect to BigQuery
con = ibis.bigquery.connect(
    project_id="my-gcp-project",
    dataset_id="my_dataset",
)

# Reference a table
t = con.table("my_table")

# Do pandas-like transformations
result = (
    t.filter(t.column_a > 10)
     .mutate(new_col=t.column_b * 2)
     .group_by(t.category)
     .aggregate(
```

Reproducibility → Reproductions?

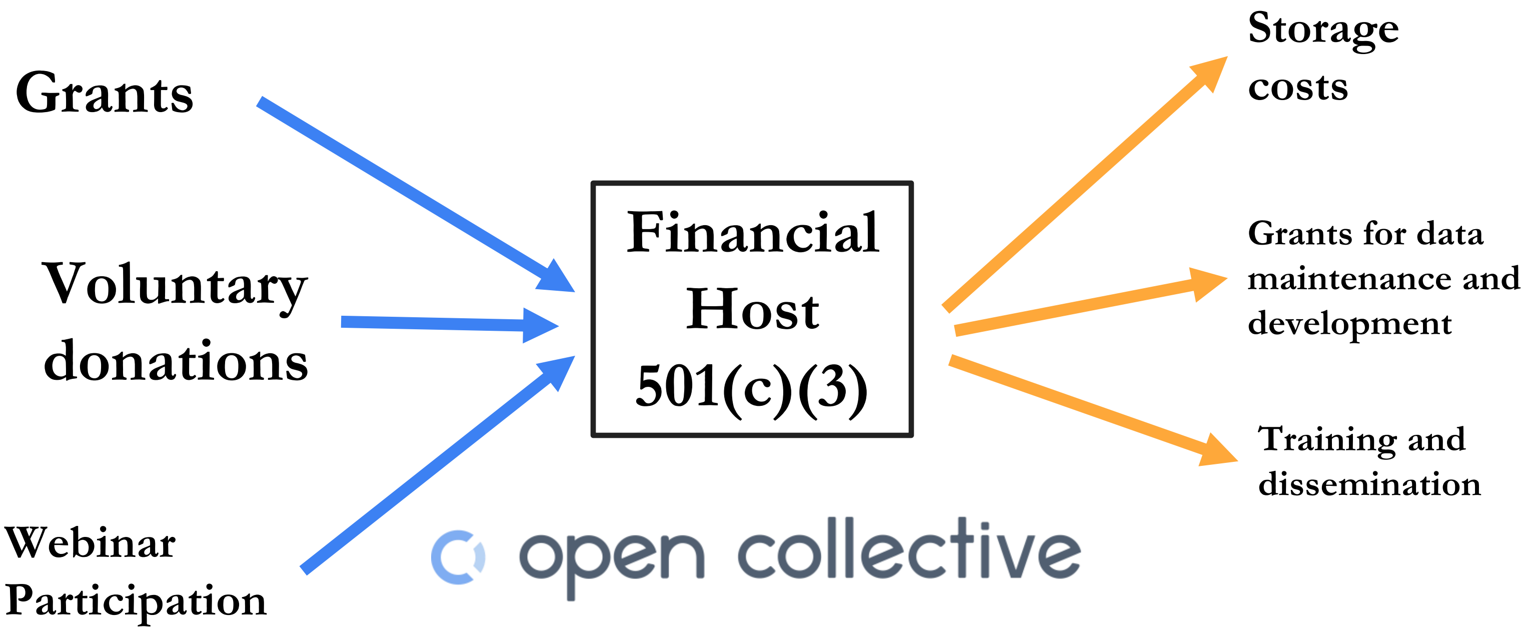
- “reproducible” research, but not reproduced
- Standardized, open, maintained datasets are a prerequisite for reproduct**ion**
- Research projects based on BigQuery can be developed through version-controlled code repositories.
  - Can be built for reproductions from raw-data to final analyses regardless of scale
  - New data versions can be integrated automatically for reproductions
- **Field-level research dashboards?**

Project Sustainability

Costs

- Data storage (currently about 30 USD/month, increases with scale)
- Updates to current datasets
- Upload additional datasets
- Support training and usage
- Support the creation of new datasets

Project Sustainability: plans ahead



Grants

Voluntary donations

Webinar Participation

Financial Host 501(c)(3)

open collective

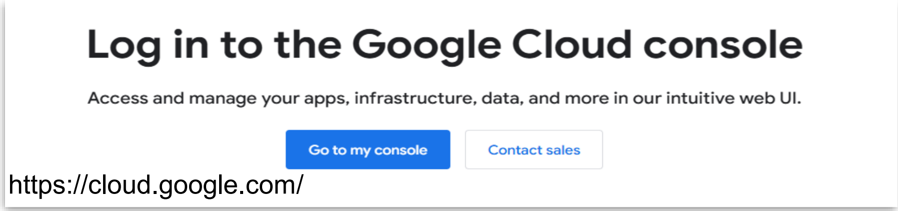
Storage costs

Grants for data maintenance and development

Training and dissemination

Getting started

1



Log in to the Google Cloud console

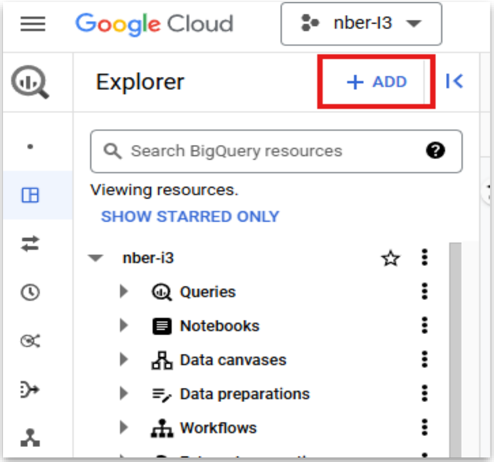
Access and manage your apps, infrastructure, data, and more in our intuitive web UI.

Go to my console

Contact sales

https://cloud.google.com/

2



Google Cloud

nber-I3

Explorer

+ ADD

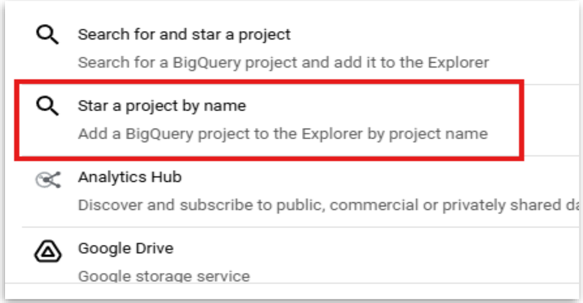
Search BigQuery resources

Viewing resources.

SHOW STARRED ONLY

- nber-i3
  - Queries
  - Notebooks
  - Data canvases
  - Data preparations
  - Workflows

3



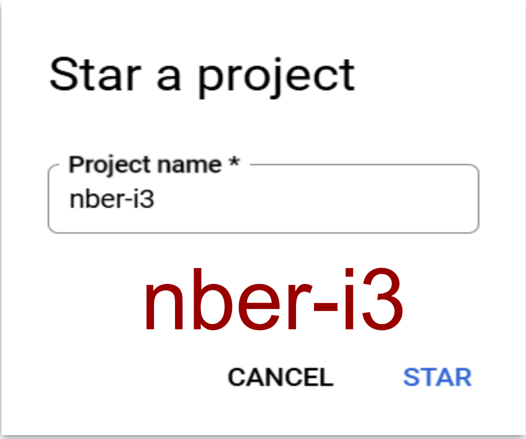
Search for and star a project
Search for a BigQuery project and add it to the Explorer

Star a project by name
Add a BigQuery project to the Explorer by project name

Analytics Hub
Discover and subscribe to public, commercial or privately shared da

Google Drive
Google storage service

4



Star a project

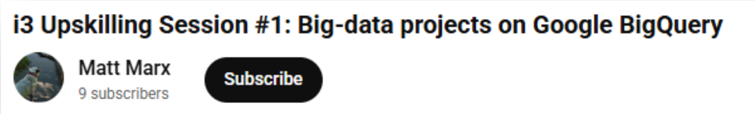
Project name *
nber-i3

nber-i3

CANCEL

STAR

i3 “upskilling” sessions (https://is.gd/i3upskilling)



i3 Upskilling Session #1: Big-data projects on Google BigQuery

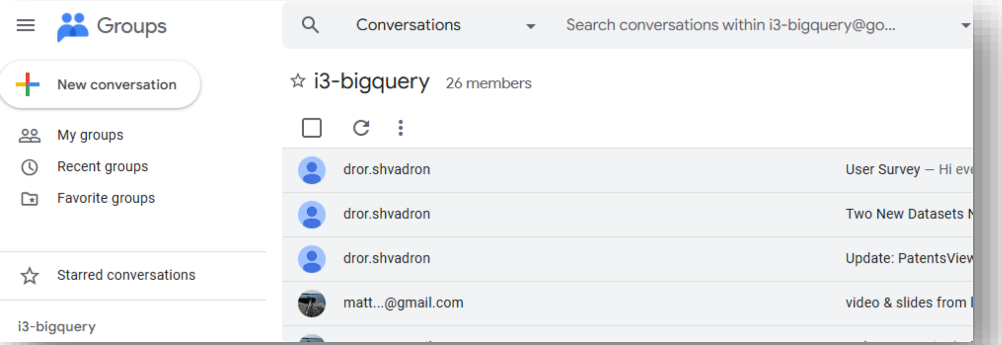
Matt Marx

9 subscribers

Join us!

- Discussion group
- Ideas
- Creation and uploading of datasets
- Maintenance
- Workshops, “upskilling”
- Funding

https://groups.google.com/g/i3-bigquery



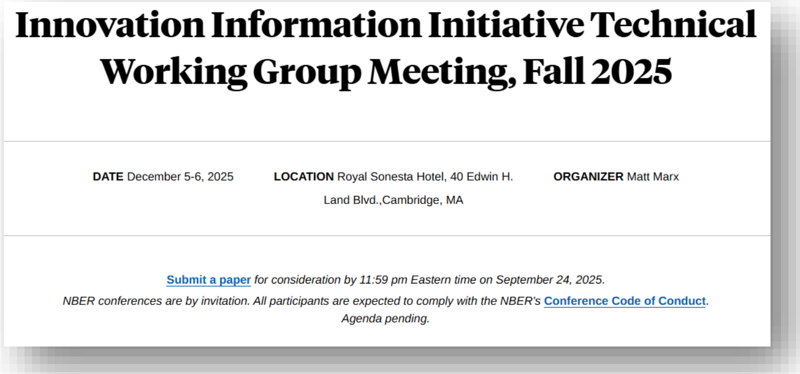
☆ i3-bigquery 26 members

dror.shvadron	User Survey – Hi eve
dror.shvadron	Two New Datasets N
dror.shvadron	Update: PatentsView
matt...@gmail.com	video & slides from I

Questions?

Thank you!

dror.shvadron@rotman.utoronto.ca
mmarx@cornell.edu



Innovation Information Initiative Technical Working Group Meeting, Fall 2025

DATE December 5-6, 2025 **LOCATION** Royal Sonesta Hotel, 40 Edwin H. Land Blvd.,Cambridge, MA **ORGANIZER** Matt Marx

Submit a paper for consideration by 11:59 pm Eastern time on September 24, 2025.
NBER conferences are by invitation. All participants are expected to comply with the NBER's Conference Code of Conduct.
Agenda pending.

Useful links

BiqQuery repository: nber-i3

Discussion Group:
https://groups.google.com/g/i3-bigquery

Reliance on Science:
https://relianceonscience.org/