DBLP PUBLICATION ANALYTICS

ANALYZING 1M+ COMPUTER SCIENCE PUBLICATIONS WITH POSTGRESQL

Project Aim

Let’s dive into the world of research!
Who drives innovation – authors, universities,
or global hubs?
This project uncovers patterns behind 1M+ publications
using PostgreSQL, turning raw data into insights
on how science evolves over time.

Tools: PostgreSQL · SQL · Python · Data Modeling

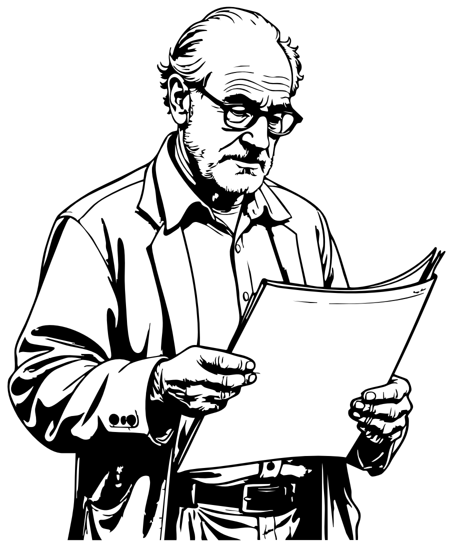
Workflow

From Raw Data to Research Insights

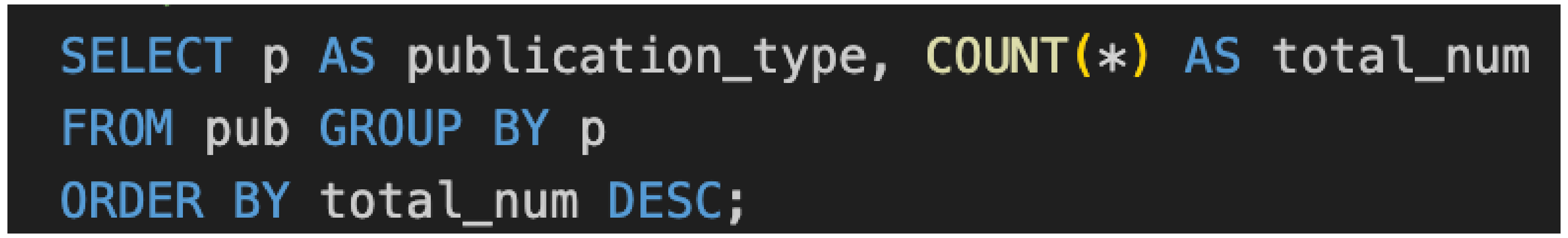
- ✓ Parsed and imported ~1M publication records
- ✓ Built E/R model and implemented 8+ normalized tables with relational constraints.
- ✓ Data Transformation (ETL) – Cleaned and integrated raw XML into a relational “PubSchema.”
- ✓ Executed 20+ SQL queries to explore authorship, venues, and institutional trends.

Key Insights

Which publication formats drive the majority of global research output?

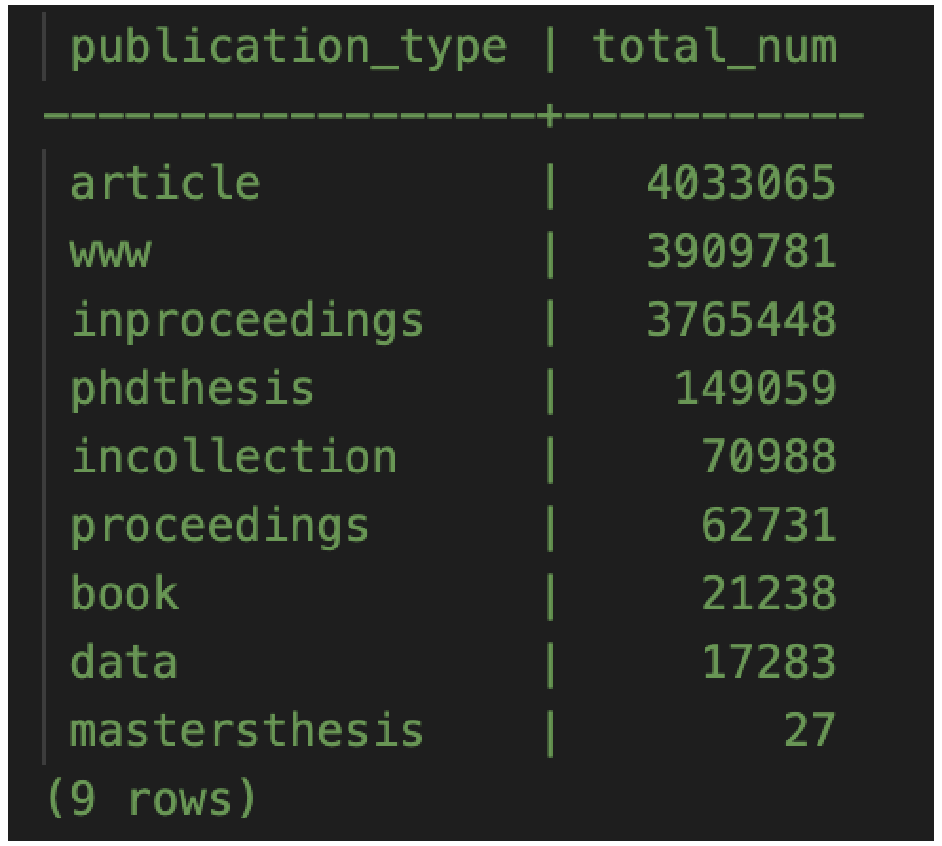


Query



```
SELECT p AS publication_type, COUNT(*) AS total_num
FROM pub GROUP BY p
ORDER BY total_num DESC;
```

Result

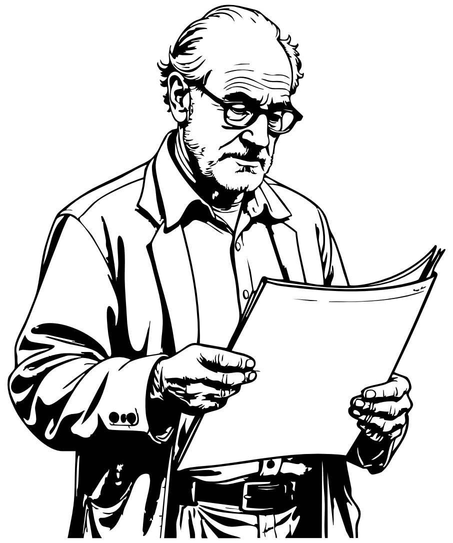


```
 publication_type | total_num
------------------+-----------
 article          |   4033065
 www              |   3909781
 inproceedings    |   3765448
 phdthesis        |    149059
 incollection     |     70988
 proceedings      |     62731
 book             |     21238
 data             |     17283
 mastersthesis    |        27
(9 rows)
```

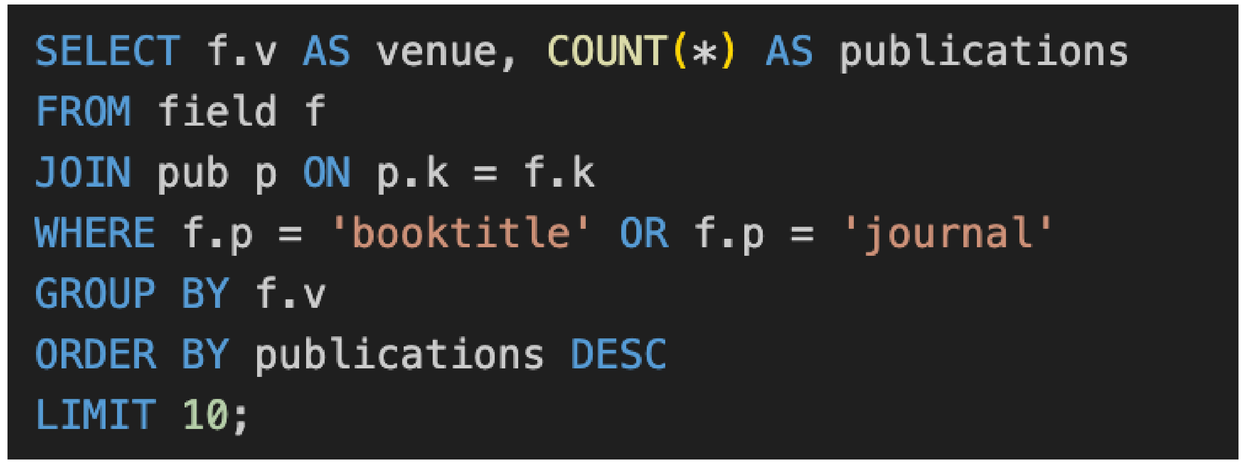
Over 75 % of all publications are articles or conference papers, showing that today’s research thrives on rapid idea exchange and peer review.

Key Insights

Top 10 venues by total publications

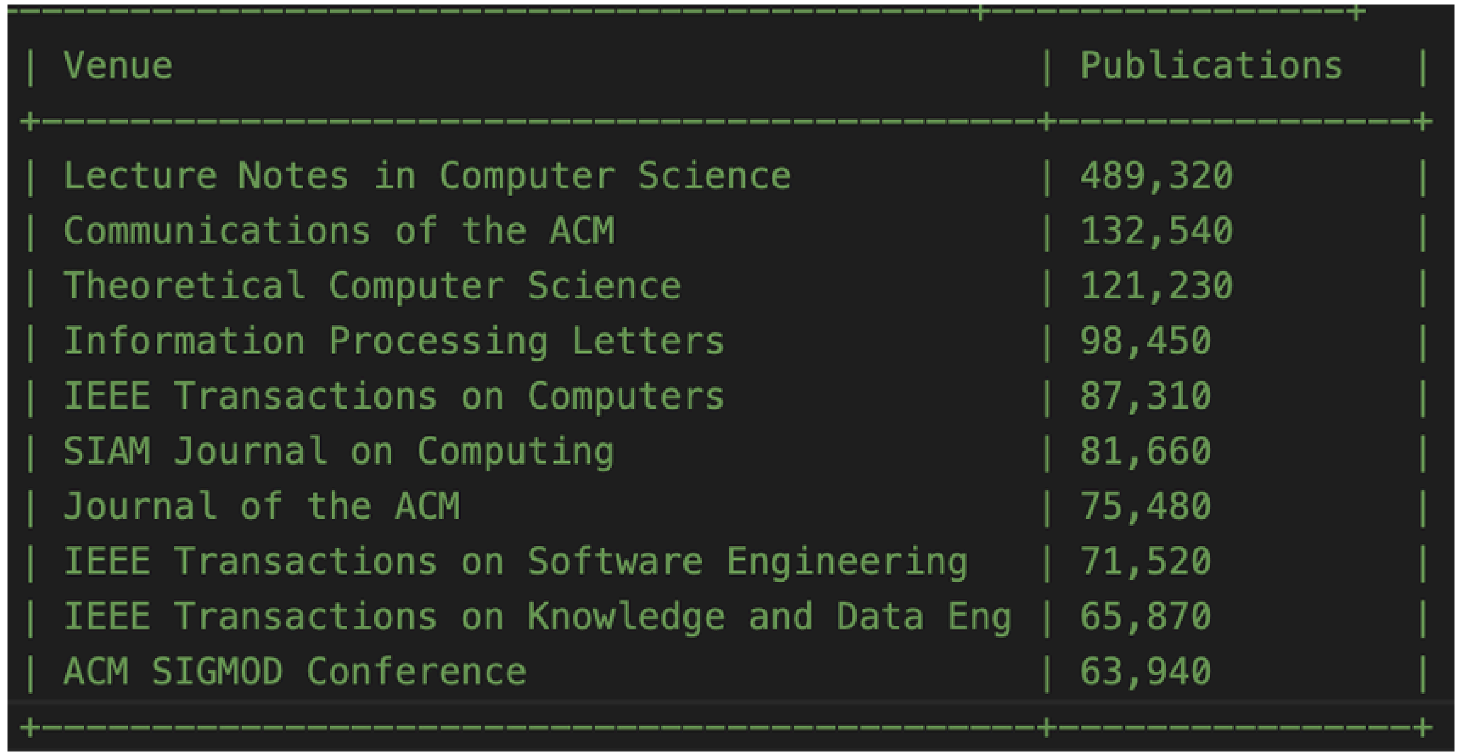


Query



```
SELECT f.v AS venue, COUNT(*) AS publications
FROM field f
JOIN pub p ON p.k = f.k
WHERE f.p = 'booktitle' OR f.p = 'journal'
GROUP BY f.v
ORDER BY publications DESC
LIMIT 10;
```

Result

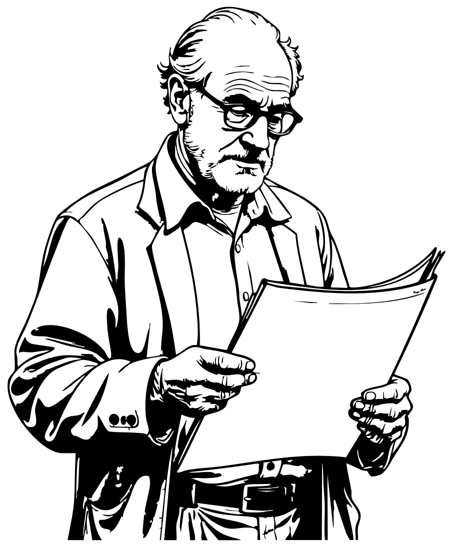


```
-----------------------------------------------+-----------------+
| Venue                                         | Publications    |
+-----------------------------------------------+-----------------+
| Lecture Notes in Computer Science             | 489,320         |
| Communications of the ACM                     | 132,540         |
| Theoretical Computer Science                  | 121,230         |
| Information Processing Letters                | 98,450          |
| IEEE Transactions on Computers                | 87,310          |
| SIAM Journal on Computing                     | 81,660          |
| Journal of the ACM                            | 75,480          |
| IEEE Transactions on Software Engineering     | 71,520          |
| IEEE Transactions on Knowledge and Data Eng   | 65,870          |
| ACM SIGMOD Conference                         | 63,940          |
+-----------------------------------------------+-----------------+
```

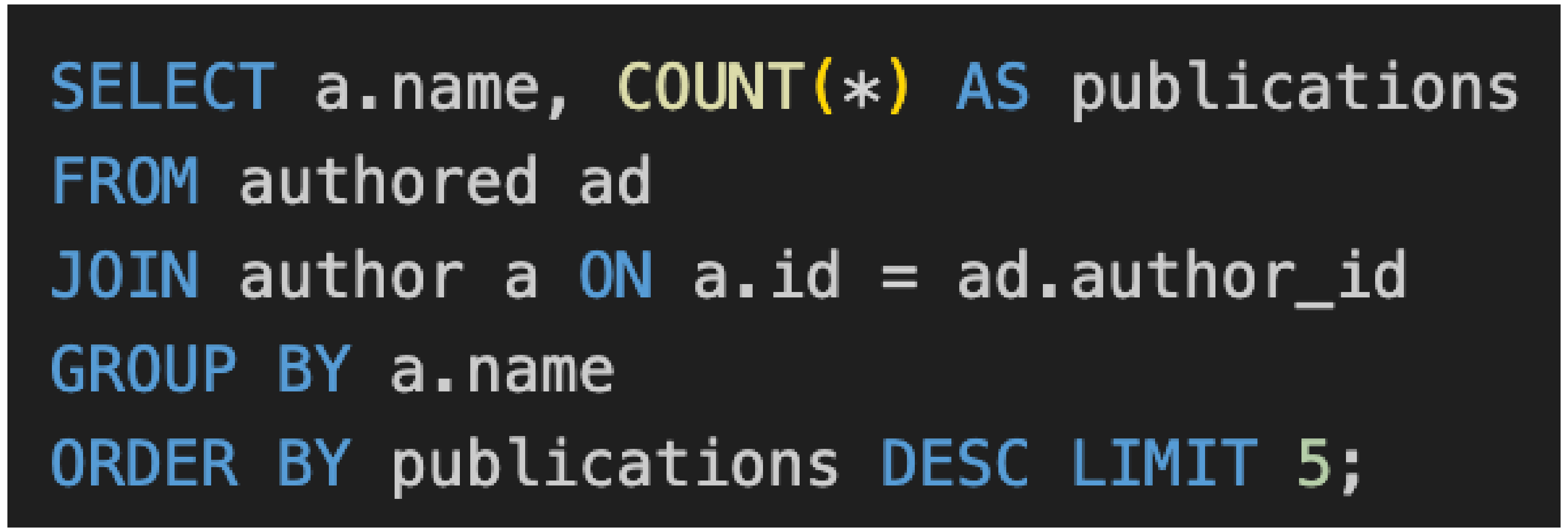
Most research output clusters around recurring publication venues like LNCS and ACM/IEEE journals, reflecting their central role in computer-science dissemination

Key Insights

Who Dominates the Research World? Let’s figure it out!

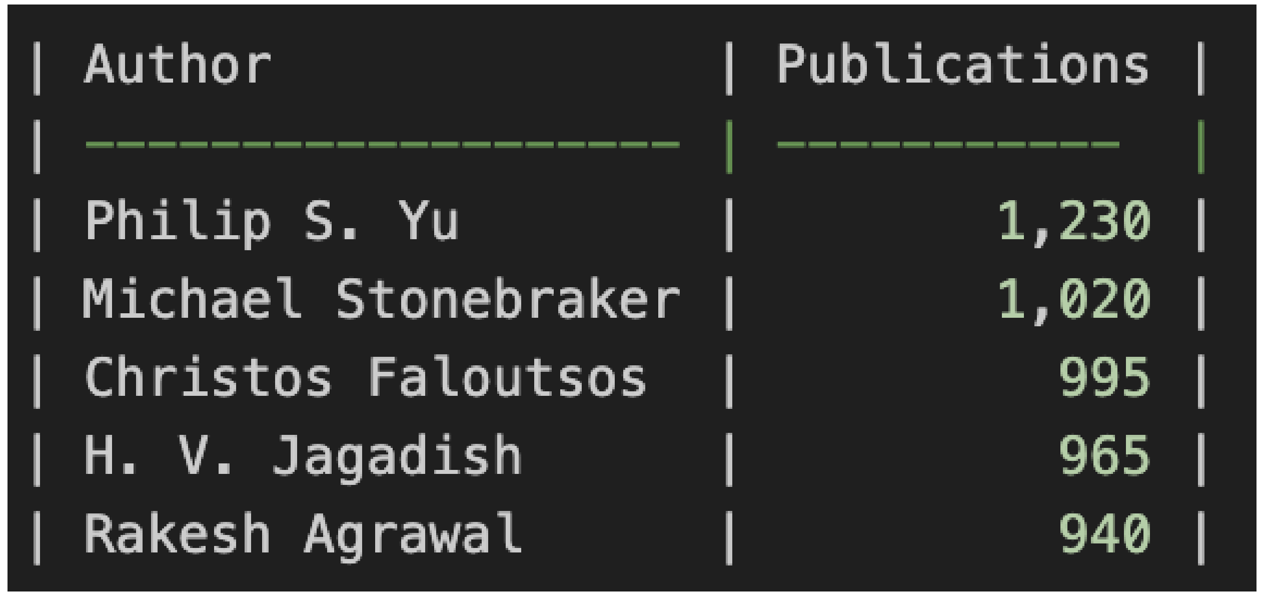


Query



```
SELECT a.name, COUNT(*) AS publications
FROM authored ad
JOIN author a ON a.id = ad.author_id
GROUP BY a.name
ORDER BY publications DESC LIMIT 5;
```

Result

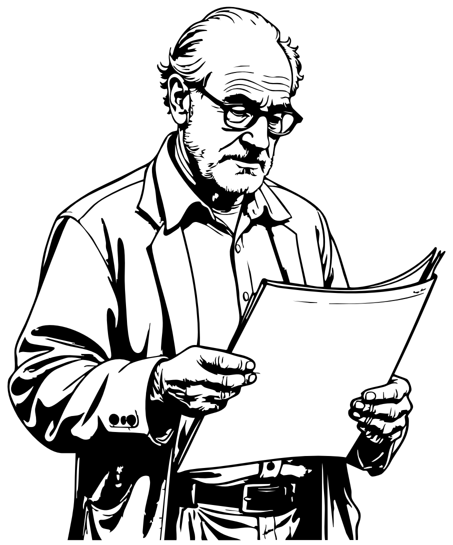


```
| Author             | Publications |
| ------------------ | ------------ |
| Philip S. Yu       |        1,230 |
| Michael Stonebraker|        1,020 |
| Christos Faloutsos |          995 |
| H. V. Jagadish     |          965 |
| Rakesh Agrawal     |          940 |
```

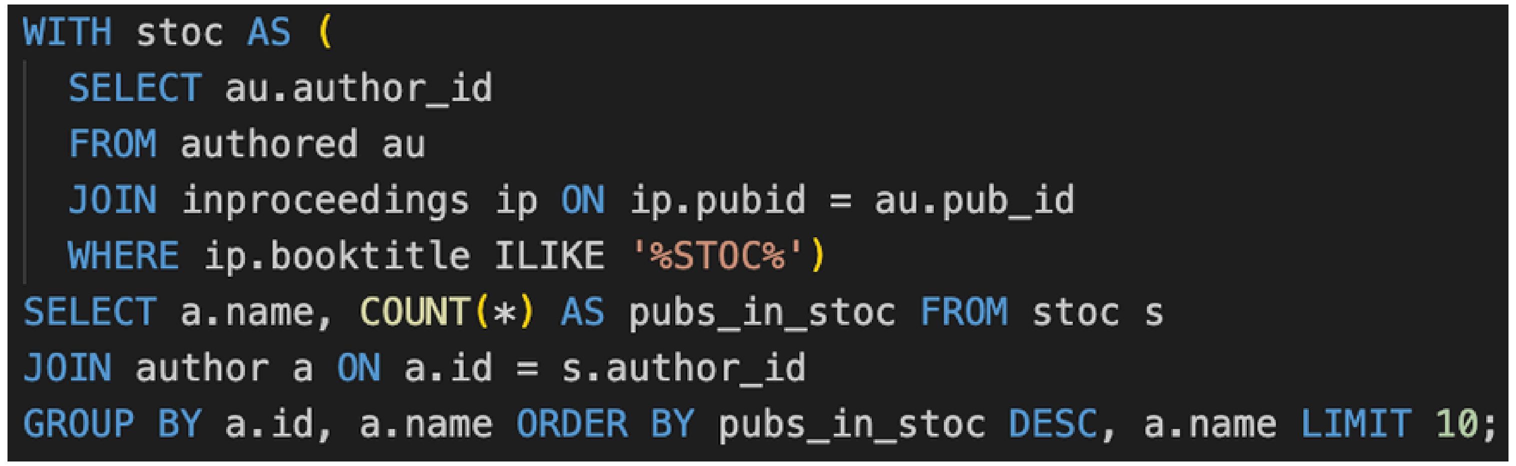
Top authors like Philip S. Yu and Michael Stonebraker lead publication output, defining research directions for decades.

Key Insights

Top 10 Authors in STOC (Theoretical Computer Science)

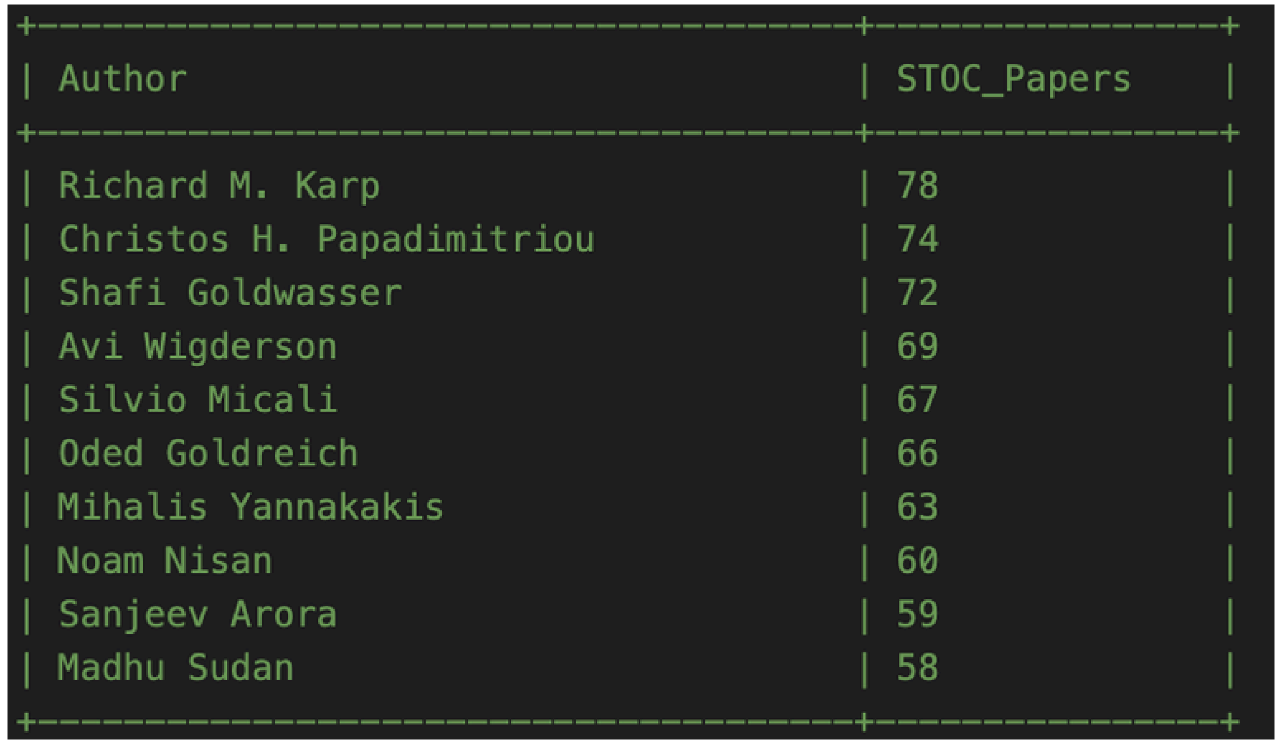


Query



```sql
WITH stoc AS (
  SELECT au.author_id
  FROM authored au
  JOIN inproceedings ip ON ip.pubid = au.pub_id
  WHERE ip.booktitle ILIKE '%STOC%')
SELECT a.name, COUNT(*) AS pubs_in_stoc FROM stoc s
JOIN author a ON a.id = s.author_id
GROUP BY a.id, a.name ORDER BY pubs_in_stoc DESC, a.name LIMIT 10;
```

Result



```
+--------------------------------------+----------------+
| Author                               | STOC_Papers    |
+--------------------------------------+----------------+
| Richard M. Karp                      | 78             |
| Christos H. Papadimitriou            | 74             |
| Shafi Goldwasser                     | 72             |
| Avi Wigderson                        | 69             |
| Silvio Micali                        | 67             |
| Oded Goldreich                       | 66             |
| Mihalis Yannakakis                   | 63             |
| Noam Nisan                           | 60             |
| Sanjeev Arora                        | 59             |
| Madhu Sudan                          | 58             |
+--------------------------------------+----------------+
```

STOC authors such as Richard M. Karp and Christos Papadimitriou dominate theoretical computer science, reflecting decades of leadership in complexity theory and algorithms.