Alexandria3k: Researching the world's knowledge on your laptop

Diomidis Spinellis

Department of Management Science and Technology
Athens University of Economics and Business

Department of Software Technology
Delft University of Technology

www.spinellis.gr

@CoolSWEng
@CoolSWEng@mastodon.acm.org
Standing on shoulders or feet?
An extended study on the usage of the MSR data papers

Kotti, Kravvaritis, Dritsa, Spinellis

Empirical Software Engineering (2020)
DOI: 10.1007/s10664-020-09834-7

ACM SIGSOFT Distinguished Paper Award in MSR 2019
Impact of Software Engineering Research in Practice: A Patent and Author Survey Analysis

Kotti, Gousios, Spinellis

IEEE Transactions on Software Engineering (2022)

DOI: 10.1109/TSE.2022.3208210
Machine Learning for Software Engineering: A Tertiary Study

ZOE KOTTI, RAFAELA GALANOPOULOU, and DIOMIDIS SPINELLIS, Athens University of Economics and Business, Greece

Machine learning (ML) techniques increase the effectiveness of software engineering (SE) lifecycle activities. We systematically collected, quality-assessed, summarized, and categorized 83 reviews in ML for SE published between 2009-2022, covering 117 primary studies. The SE areas most tackled with ML are software quality and testing, while human-centered areas appear more challenging for ML. We propose a number of ML for SE research challenges and actions including: conducting further empirical validation and industrial studies on ML, reconsidering deficient SE methods, documenting and automating data collection and pipeline processes; reexamining how industrial practitioners distribute their proprietary data; and implementing incremental ML approaches.


Additional Key Words and Phrases: Tertiary study, machine learning, software engineering, systematic literature review

ACM Reference Format:

1 INTRODUCTION

Machine learning (ML) is a thriving discipline with various practical applications and active research topics, many of which nowadays entangle the discipline of software engineering (SE) [113]. Through ML we can address SE problems that cannot be completely algorithmically modeled, or for which existing solutions do not provide satisfactory results yet (e.g., defect/fault detection [16, 105, 106]). In addition, ML finds application in SE tasks where data cannot be easily analyzed with other algorithms (e.g., software requirements, code comments, code reviews, issues [9, 91, 174]). Another important aspect of ML is that it can significantly reduce manual effort in common SE tasks (e.g., automatic program repair [157], code suggestion [61], defect prediction [19], malware detection [147], feature location [80]) with great accuracy results [116, 117]. In fields such as health informatics ML and SE are considered complementary disciplines, where the growing scale and complexity of healthcare datasets have posed a challenge for clinical practice and medical research, requiring new engineering approaches from both fields [25].

In the early nineties, Huff and Selfridge [66] recognized the need for creating software systems that partially take some responsibility for their own evolution, offering the ability to implement, measure, and assess changes easily. These changes should also contribute to the overall improvement of the corresponding systems [142]. Around the same time, Brooks [29] prompted software practitioners to investigate evolutionary advancements rather than waiting for
Issues

- Lack of transparency, repeatability, reproducibility
- High latency, low bandwidth
- Rate limits
- Proprietary and restricted query languages
- Limited coverage
- Availability and cost
<table>
<thead>
<tr>
<th>Measure</th>
<th>Elliott 405</th>
<th>Raspberry Pi Zero</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Year</strong></td>
<td>1957</td>
<td>2015</td>
</tr>
<tr>
<td><strong>Price</strong></td>
<td>£85,000 (1957) — €2M (2018)</td>
<td>$5</td>
</tr>
<tr>
<td><strong>Instruction cycle time</strong></td>
<td>10.71–0.918 ms (93-1089 Hz)</td>
<td>1 ns (1 GHz clock)</td>
</tr>
<tr>
<td><strong>Main memory</strong></td>
<td>16 kB drum store</td>
<td>512 MB LPDDR2 SDRAM</td>
</tr>
<tr>
<td><strong>Fast memory</strong></td>
<td>1280 bytes (nickel delay lines)</td>
<td>32 kB (16 kB I + 16kB D L1 cache)</td>
</tr>
<tr>
<td><strong>Secondary memory</strong></td>
<td>1.2 MB (300,000 word magnetic film)</td>
<td>8 GB (typical micro SD flash card)</td>
</tr>
<tr>
<td><strong>Output bandwidth</strong></td>
<td>25 characters/s</td>
<td>373 MB/s (1080p60 HDMI)</td>
</tr>
<tr>
<td><strong>Weight</strong></td>
<td>3–6 tons</td>
<td>9 g</td>
</tr>
<tr>
<td><strong>Size</strong></td>
<td>21 cabinets, each 2m x 77cm x 77cm</td>
<td>65mm x 30mm x 5.4mm</td>
</tr>
<tr>
<td><strong>Operating power</strong></td>
<td>10 kW</td>
<td>0.7 W</td>
</tr>
</tbody>
</table>
Alexandria3k
Publication metadata analytics on the desktop

- Relational access to 1.9 TB of data
- 4.2 billion records in 74 tables
- Installed as a single Python module
- No (graph) database / cluster to install / maintain
- Efficient
  - Data sample queries run in minutes
  - Data building of full data slices in 5 h–2 days
  - Then queries run in seconds
  - Space requirements start at 157 GB for downloaded data
Agenda

- Data model and data
- Alexandria3k in practice
- Implementation
- Issues and limitations
- Way forward
Data schema
Crossref data in numbers

Number of elements

- Works: 134,048
- Works with a text mining link: 96,295
- Works with subject: 81,210
- Works with references: 52,907
- Works with affiliation: 36,390
- Works with an abstract: 15,368
- Works with funders: 7,519
- Author records: 359,557
  - Author records with ORCID: 16,746
  - Distinct authors with ORCID: 4,526
- Author affiliation records: 76,760
  - Distinct affiliation names: 19,453
- Work subject records: 182,858
  - Distinct subject names: 0
- Work funders: 15,492
  - Funder records with DOI: 10,811
  - Distinct funder DOIs: 30
  - Funder awards: 14,091
- References: 1,748,422

Thousands
<table>
<thead>
<tr>
<th>Record Type</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Journal Article</td>
<td>93,491</td>
</tr>
<tr>
<td>Book Chapter</td>
<td>18,144</td>
</tr>
<tr>
<td>Proceedings Article</td>
<td>7,208</td>
</tr>
<tr>
<td>Component</td>
<td>5,573</td>
</tr>
<tr>
<td>Dataset</td>
<td>2,317</td>
</tr>
<tr>
<td>Reference Entry</td>
<td>1,085</td>
</tr>
<tr>
<td>Book</td>
<td>964</td>
</tr>
<tr>
<td>Journal Issue</td>
<td>928</td>
</tr>
<tr>
<td>Monograph</td>
<td>548</td>
</tr>
<tr>
<td>Book</td>
<td>964</td>
</tr>
<tr>
<td>Report</td>
<td>693</td>
</tr>
<tr>
<td>Standard</td>
<td>348</td>
</tr>
<tr>
<td>Other</td>
<td>1,310</td>
</tr>
<tr>
<td>Dissertation</td>
<td>503</td>
</tr>
<tr>
<td>Posted Content</td>
<td>894</td>
</tr>
<tr>
<td>Reference Entry</td>
<td>1,085</td>
</tr>
<tr>
<td>Dataset</td>
<td>2,317</td>
</tr>
<tr>
<td>Component</td>
<td>5,573</td>
</tr>
<tr>
<td>Proceedings Article</td>
<td>7,208</td>
</tr>
<tr>
<td>Book-Chapter</td>
<td>18,144</td>
</tr>
<tr>
<td>Journal Article</td>
<td>93,491</td>
</tr>
</tbody>
</table>
Crossref publications per year
One more thing...
Journals, Funders, Open Access

• Crossref journal names (109k records)
• Crossref funder names (21k records)
• DOAJ open access journal metadata (19k records)
Alexandria3k in practice
CLI usage

usage: a3k [-h] [-d DEBUG] [-v]
    {help,populate,process,query,list-processes,list-complete-schema,list-source-schema,list-process-schema,list-sources,version} ...

a3k: Relational interface to publication metadata

positional arguments:
    {help,populate,process,query,list-processes,list-complete-schema,list-source-schema,list-process-schema,list-sources,version}
        Name of the a3k operation to perform.
        help                Show top-level help message.
        populate            Populate an SQLite database.
        process             Run a processing step on the specified database.
        query               Run a query directly on a data source.
        list-processes      List available data processes.
        list-complete-schema List all data source and process schemas.
        list-source-schema  List all data source schemas (default) or the specified one.
        list-process-schema List the schema of all processes (default) or of the specified one.
        list-sources        List available data sources
        version             Report program version

optional arguments:
    -h, --help            show this help message and exit
    -d DEBUG, --debug DEBUG
CLI invocation example

a3k populate covid.db \\crossref 'April 2022 Public Data File from Crossref' \\
--row-selection "title like '%COVID%' OR abstract like '%COVID%' "
Python module example

```python
from alexandria3k.crossref import Crossref

crossref_instance = Crossref('April 2022 Public Data File from Crossref')

crossref_instance.populate("covid.db", condition="title like '%COVID%' OR abstract like '%COVID%'")
```
Typical workflow

1. Download data
   - < 3h
   - 156 GiB

2. Run EDA queries directly on sample
   - 2’ on 1%
   - 8 records / s

3. Populate database
   - 4–20 h
   - 4–190 GiB

4. Develop, test, refine analysis queries
   - 1’–7h
   - ≤ 5GiB
Main use cases

• Run ad hoc SQL queries
• Populate SQLite databases
  – Select elements horizontally
    • SQL expression
    • Sampling
  – Select elements vertically
    • Table.Column
  – Building takes minutes, hours, or a couple of days
  – Then, SQLite database queries often run in seconds
Crossref publications by year

a3k query crossref 'April 2022 Public Data File from Crossref' \--query 'SELECT published_year AS year, Count(*) AS number FROM works GROUP by published_year' >results.csv
Crossref sampling

time alexandria3k query crossref 'April 2022 Public Data File from Crossref' \ 
  --query 'SELECT works.abstract is not null AS have_abstract, Count(*) FROM works GROUP BY have_abstract' \ 
  --sample 'random.random() < 0.01'

0  1218383
1  156617

real  2m6.488s
user  1m58.878s
sys   0m6.920s
Crossref population metrics

alexandria3k populate crossref 'April 2022 Public Data File from Crossref' graph.db \
--columns works.doi work_references.work_id work_references.doi work_funders.id \n    work_funders.work_id work_funders.doi funder_awards.funder_id funder_awards.name \n    author_affiliations.author_id author_affiliations.name work_subjects.work_id work_subjects.name \n    work_authors.id work_authors.work_id work_authors.orcid

SELECT COUNT(*) FROM works;
SELECT COUNT(*) FROM (SELECT DISTINCT work_id FROM works_subjects);
SELECT COUNT(*) FROM (SELECT DISTINCT work_id FROM work_references);
SELECT COUNT(*) FROM affiliations_works;
SELECT COUNT(*) FROM (SELECT DISTINCT work_id FROM work_funders);

SELECT COUNT(*) FROM work_authors;
SELECT COUNT(*) FROM work_authors WHERE orcid is not null;
SELECT COUNT(*) FROM (SELECT DISTINCT orcid FROM work_authors);

SELECT COUNT(*) FROM authors_affiliations;
SELECT COUNT(*) FROM affiliation_names;

SELECT COUNT(*) FROM works_subjects;
SELECT COUNT(*) FROM subject_names;

SELECT COUNT(*) FROM work_funders;
SELECT COUNT(*) FROM funder_awards;

SELECT COUNT(*) FROM work_references;
Number of ORCID elements (for chart)

SELECT "persons" AS type, (SELECT COUNT(*) FROM persons) AS records,
(SELECT COUNT(*) FROM (SELECT DISTINCT orcid FROM persons)) AS persons UNION

SELECT "researcher_urls" AS type, (SELECT COUNT(*) FROM researcher_urls) AS records,
(SELECT COUNT(*) FROM (SELECT DISTINCT orcid FROM researcher_urls)) AS persons UNION

SELECT "person_countries" AS type, (SELECT COUNT(*) FROM person_countries) AS records,
(SELECT COUNT(*) FROM (SELECT DISTINCT orcid FROM person_countries)) AS persons UNION

SELECT "person_keywords" AS type, (SELECT COUNT(*) FROM person_keywords) AS records,
(SELECT COUNT(*) FROM (SELECT DISTINCT orcid FROM person_keywords)) AS persons UNION

SELECT "person_external_identifiers" AS type, (SELECT COUNT(*) FROM person_external_identifiers) AS records,
(SELECT COUNT(*) FROM (SELECT DISTINCT orcid FROM person_external_identifiers)) AS persons UNION

SELECT "distinctions" AS type, (SELECT COUNT(*) FROM distinctions) AS records,
(SELECT COUNT(*) FROM (SELECT DISTINCT orcid FROM distinctions)) AS persons UNION

SELECT "educations" AS type, (SELECT COUNT(*) FROM educations) AS records,
(SELECT COUNT(*) FROM (SELECT DISTINCT orcid FROM educations)) AS persons UNION

SELECT "employments" AS type, (SELECT COUNT(*) FROM employments) AS records,
(SELECT COUNT(*) FROM (SELECT DISTINCT orcid FROM employments)) AS persons UNION

SELECT "invited_positions" AS type, (SELECT COUNT(*) FROM invited_positions) AS records,
(SELECT COUNT(*) FROM (SELECT DISTINCT orcid FROM invited_positions)) AS persons UNION

SELECT "memberships" AS type, (SELECT COUNT(*) FROM memberships) AS records,
(SELECT COUNT(*) FROM (SELECT DISTINCT orcid FROM memberships)) AS persons UNION

SELECT "qualifications" AS type, (SELECT COUNT(*) FROM qualifications) AS records,
(SELECT COUNT(*) FROM (SELECT DISTINCT orcid FROM qualifications)) AS persons UNION

SELECT "services" AS type, (SELECT COUNT(*) FROM services) AS records,
(SELECT COUNT(*) FROM (SELECT DISTINCT orcid FROM services)) AS persons UNION

SELECT "fundings" AS type, (SELECT COUNT(*) FROM fundings) AS records,
(SELECT COUNT(*) FROM (SELECT DISTINCT orcid FROM fundings)) AS persons UNION

SELECT "peer_reviews" AS type, (SELECT COUNT(*) FROM peer_reviews) AS records,
(SELECT COUNT(*) FROM (SELECT DISTINCT orcid FROM peer_reviews)) AS persons UNION

SELECT "research_resources" AS type, (SELECT COUNT(*) FROM research_resources) AS records,
(SELECT COUNT(*) FROM (SELECT DISTINCT orcid FROM research_resources)) AS persons;
Consolidation / Disruption index

**Self-Consistent Equations Including Exchange and Correlation Effects**

W. Kohn and L. J. Sham
University of California, San Diego, La Jolla, California

(Accepted 21 June 1965)

From a theory of Hohenberg and Kohn, approximate methods for treating an inhomogeneous system of interacting electrons have been developed. These methods are based on the Hohenberg-Kohn theorem, which states that a given physical property, such as the energy of an electron gas, can be expressed in terms of the properties of a corresponding system of independent particles. The methods proposed here use the idea of a molecular orbital model to determine the exchange and correlation energy of a system.

A. INTRODUCTION

No more a great deal of attention has been given to the problem of a homogeneous gas of interacting electrons and it is expected that the results will be considered with a considerable degree of confidence over a wide range of densities. Of course, such a homogeneous gas represents only a mathematical model, since in all real systems (anions, neutral, molecules, solids) the electron density is nonuniform.

It is then a matter of interest to see how properties of the homogeneous gas can be verified in theoretical studies of inhomogeneous systems. The well-known methods of Thomas-Fermi and the Slater exchange give a step in this direction. In the present paper we use the formation of Hohenberg and Kohn to carry this study further. In this paper we are concerned with systems which have a large number of electrons and some of the results which include, in an approximate way, exchange and correlation effects. They imply the importance of the total energy of an inhomogeneous electron gas as a function of the density.

We define two alternative sets of equations [Eqs. (2.8) and (2.22)] which are analogous, respectively, to the conventional Hartree and Hartree-Fock equations, and, although they also include correlation effects, they are more difficult to solve.

The local effective potentials in these equations are unique in a sense which is described in Sec. II. In particular, we find that the Slater exchange homoatomic potential, which contains the correlation effect, is too large by a factor of 1.5.

The methods proposed here are based on the general correlation energy of the homogeneous electron gas, which has been found to be less than the correlation energy of the analogous system of independent particles. The methods proposed here are similar to the ideas used by Hohenberg and Kohn in their treatment of the free electron gas. The methods proposed here involve a number of simplifications which are analogous to the simplifications used by Hohenberg and Kohn in their treatment of the free electron gas.

B. THE GROUND STATE

1. Local Effective Potential

It has been shown that the ground state of an inhomogeneous electron gas is a static potential function which can be written in the form

\[ \rho(x) = \frac{1}{2\pi} \int e^{-i(k_x x + k_y y + k_z z)} d^3k \rho(k) \]

where \( \rho(x) \) is the density and \( \rho(k) \) is a universal function of the density. This expression, in turn, is a solution to the correct density functions. In this section we present a first approximation for \( \rho(k) \), which leads to a simple scheme in which the methods proposed here are similar to the ideas used by Hohenberg and Kohn in their treatment of the free electron gas.

We define a set of equations [Eqs. (2.8) and (2.22)] which are analogous, respectively, to the conventional Hartree and Hartree-Fock equations, and, although they also include correlation effects, they are more difficult to solve.

The local effective potentials in these equations are unique in a sense which is described in Sec. II. In particular, we find that the Slater exchange homoatomic potential, which contains the correlation effect, is too large by a factor of 1.5.

The methods proposed here are based on the general correlation energy of the homogeneous electron gas, which has been found to be less than the correlation energy of the analogous system of independent particles. The methods proposed here are similar to the ideas used by Hohenberg and Kohn in their treatment of the free electron gas. The methods proposed here involve a number of simplifications which are analogous to the simplifications used by Hohenberg and Kohn in their treatment of the free electron gas.

We define a set of equations [Eqs. (2.8) and (2.22)] which are analogous, respectively, to the conventional Hartree and Hartree-Fock equations, and, although they also include correlation effects, they are more difficult to solve.
Evolution of scientific publishing

- Authors per work: 1.4 → 4.4
- Works per author: 1.99 → 1.59
- References per work: 13 → 46
- Pages per work: 6.3 → 12.8
- CD's index: 0.134 → 0.001
- Works published: 81k → 7.0M (log)
- Citations per work: 0.009 → 1.14 (log)
- Journals: 958 → 68k (log)
- Works cited at least once: 0.3% → 20% (log)
- 2-year IF: 0.06 → 1.5 (log)
- 20-year IF: 0.02 → 1.3 (log)
WITH ranked_countries AS (
    SELECT
        SUBSTRING(date_published, 1, 4) AS year,
        usp_applicants.country AS country,
        COUNT(*) AS patent_count,
        ROW_NUMBER() OVER(PARTITION BY SUBSTRING(date_published, 1, 4) ORDER BY COUNT(*) DESC) AS country_rank
    FROM us_patents
    INNER JOIN usp_applicants
    ON us_patents.container_id = usp_applicants.patent_id
    GROUP BY year, usp_applicants.country
),

  top_5_2022 AS (
    SELECT country
    FROM ranked_countries
    WHERE year = '2022' AND country_rank <= 5
  )

SELECT rc.year,
        rc.country,
        rc.patent_count
FROM ranked_countries rc
JOIN top_5_2022 t5
ON rc.country = t5.country
ORDER BY rc.year, rc.country;
**Abstract**

The development of statistical software is research has transformed the way scientists and researchers conduct their statistical analysis. Despite these advancements, it was not clear which statistical software is mainly used for which research design thereby creating confusion and anxiety in choosing the right statistical tools. Therefore, this study aimed to review the trend of statistical software usage and their associated study designs in articles published in health sciences research.

**Methods**

This bibliometric analysis study reviewed 453 articles published in PubMed in three 10-year intervals (1991, 2007, and 2017). The data were collected through Google sheet and were analyzed using SPSS software. This study described the trend and usage of currently available statistical tools and the different study designs that are associated with them.

**Results**

Of the statistical software mentioned in the retrieved articles, SPSS was the most common statistical tool used (32.1%) in the three-time periods followed by SAS (15.3%) and Stata (13.6%). SPSS was the last used statistical software with only 0.6% of the total articles. SPSS was mostly associated with observational (66.6%) and experimental (25.8%) study designs. On the other hand, Review Manager (45.7%) and Stata (35.3%) were the most statistical software associated with systematic reviews and meta-analyses.

**Conclusion**

In this study, SPSS was found to be the most widely used statistical software in the selected study periods. Observational studies were the most common health science research design. SPSS was associated with observational and experimental studies while Review Manager and Stata were mostly used for systematic reviews and meta-analysis.

---

**Trends in the Usage of Statistical Software and Their Associated Study Designs in Health Sciences Research: A Bibliometric Analysis**

Email: Manuol 1, Abdulrahman Al-Shanori 2, Abdulrahman Al-Shanori 3, Abdulrahman Al-Shanori 4, Abdulrahman Al-Shanori 5, Abdulrahman Al-Shanori 6

1. Research Unit Biostatistics, King Saud University, Health Sciences, College of Medicine King Abdulaziz International Medical Research Center, Riyadh, KAU
2. Research Unit Epidemiology, King Saud University, Health Sciences, College of Medicine, Riyadh, KAU
3. King Saud University, Health Sciences, College of Medicine, Riyadh, KAU

**Keywords:** statistical software, study design, healthcare publications, meta, stats, sa, published

---

**Introduction**

With the evolution of open access in the publishing world, access to empirical research has never been more widespread than it is now. For most of the researchers, however, the key feature of their articles is the robustness and replicability of their methods section particularly the design of the study and the type of statistical tests to employ. The emergence of statistical software has transformed the way scientists and researchers conducting their statistical analysis. Therefore, performing complex and at times intensive statistical analysis manually has become thing of the past [1].

Statistical software has many useful applications for researchers in the healthcare sciences. Furthermore, the researchers conveniently read their data by representing their data as visual aid using charts and graphs [2]. It also helps the researchers to easily calculate their results using statistical tests by accounting for their variables either numerical, categorical or both [2]. However, in the past few decades, statistical software usage went through different stages based on their development and applications [3]. Although some software are more dedicated to a specific field, the degree of usage of specific software may depend on the preference of the investigators or the type of study design that is selected in their research.

---

**Software usage percentage (All Years Combined)**

<table>
<thead>
<tr>
<th>Software</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>SPSS</td>
<td>32.1%</td>
</tr>
<tr>
<td>SAS</td>
<td>15.3%</td>
</tr>
<tr>
<td>Stata</td>
<td>13.6%</td>
</tr>
<tr>
<td>R project</td>
<td>7.7%</td>
</tr>
<tr>
<td>Excel</td>
<td>6.9%</td>
</tr>
<tr>
<td>GraphPad</td>
<td>6.1%</td>
</tr>
<tr>
<td>Review Manager</td>
<td>5.4%</td>
</tr>
<tr>
<td>Epi Info</td>
<td>4.6%</td>
</tr>
<tr>
<td>Statistica</td>
<td>3.8%</td>
</tr>
<tr>
<td>Lisrel</td>
<td>3.6%</td>
</tr>
<tr>
<td>JMP</td>
<td>3.2%</td>
</tr>
<tr>
<td>WinBUGS</td>
<td>3.1%</td>
</tr>
<tr>
<td>Minitab</td>
<td>2.9%</td>
</tr>
<tr>
<td>MedCalc</td>
<td>2.8%</td>
</tr>
<tr>
<td>Alexandria3K</td>
<td>2.6%</td>
</tr>
<tr>
<td>Original</td>
<td>2.5%</td>
</tr>
</tbody>
</table>
A data set of COVID research

alexandria3k populate covid.db \
crossref 'April 2022 Public Data File from Crossref' \
--row-selection "title like '%COVID%' OR abstract like '%COVID%'"

- 9:06:23 elapsed time
- 2.9 GB data, 3.6 GB fully indexed
COVID research topics

```
SELECT rank() OVER (ORDER BY count(*) DESC), count(*), name
FROM work_subjects
GROUP BY name;
```
## COVID research funding

<table>
<thead>
<tr>
<th>Rank</th>
<th>Publications</th>
<th>Funding body</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3506</td>
<td>National Natural Science Foundation of China</td>
</tr>
<tr>
<td>2</td>
<td>2316</td>
<td>National Institutes of Health</td>
</tr>
<tr>
<td>3</td>
<td>1022</td>
<td>National Science Foundation</td>
</tr>
<tr>
<td>4</td>
<td>914</td>
<td>Wellcome Trust</td>
</tr>
<tr>
<td>5</td>
<td>661</td>
<td>National Institute for Health Research</td>
</tr>
<tr>
<td>6</td>
<td>615</td>
<td>Medical Research Council</td>
</tr>
<tr>
<td>7</td>
<td>588</td>
<td>National Institute of Allergy and Infectious Diseases</td>
</tr>
<tr>
<td>8</td>
<td>541</td>
<td>Canadian Institutes of Health Research</td>
</tr>
<tr>
<td>9</td>
<td>520</td>
<td>Deutsche Forschungsgemeinschaft</td>
</tr>
<tr>
<td>10</td>
<td>503</td>
<td>Conselho Nacional de Desenvolvimento Científico e Tecnológico</td>
</tr>
<tr>
<td>11</td>
<td>495</td>
<td>Bill and Melinda Gates Foundation</td>
</tr>
<tr>
<td>12</td>
<td>483</td>
<td>National Research Foundation of Korea</td>
</tr>
<tr>
<td>13</td>
<td>481</td>
<td>Japan Society for the Promotion of Science</td>
</tr>
<tr>
<td>14</td>
<td>439</td>
<td>National Heart, Lung, and Blood Institute</td>
</tr>
<tr>
<td>15</td>
<td>430</td>
<td>National Key Research and Development Program of China</td>
</tr>
<tr>
<td>16</td>
<td>422</td>
<td>National Center for Advancing Translational Sciences</td>
</tr>
<tr>
<td>17</td>
<td>417</td>
<td>Instituto de Salud Carlos III</td>
</tr>
<tr>
<td>18</td>
<td>394</td>
<td>National Institute on Aging</td>
</tr>
<tr>
<td>19</td>
<td>382</td>
<td>Coordenação de Aperfeiçoamento de Pessoal de Nível Superior</td>
</tr>
<tr>
<td>20</td>
<td>365</td>
<td>National Cancer Institute</td>
</tr>
</tbody>
</table>

**SQL Query**

```sql
SELECT rank() OVER (ORDER BY count(*) DESC), count(*), name
FROM work_funders
GROUP BY name
LIMIT 20;
```
### Affiliations of COVID publications

<table>
<thead>
<tr>
<th>Rank</th>
<th>Works</th>
<th>Affiliation (top parent)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1465</td>
<td>Government of the United States of America</td>
</tr>
<tr>
<td>2</td>
<td>925</td>
<td>University of California System</td>
</tr>
<tr>
<td>3</td>
<td>910</td>
<td>University of Toronto</td>
</tr>
<tr>
<td>4</td>
<td>824</td>
<td>University of London</td>
</tr>
<tr>
<td>5</td>
<td>660</td>
<td>University of Oxford</td>
</tr>
<tr>
<td>6</td>
<td>654</td>
<td>Istituti di Ricovero e Cura a Carattere Scientifico</td>
</tr>
<tr>
<td>7</td>
<td>632</td>
<td>Mount Sinai Health System</td>
</tr>
<tr>
<td>8</td>
<td>592</td>
<td>Tehran University of Medical Sciences</td>
</tr>
<tr>
<td>9</td>
<td>587</td>
<td>University of North Carolina System</td>
</tr>
<tr>
<td>10</td>
<td>501</td>
<td>University of Melbourne</td>
</tr>
<tr>
<td>11</td>
<td>437</td>
<td>The University of Texas System</td>
</tr>
<tr>
<td>12</td>
<td>434</td>
<td>National University of Singapore</td>
</tr>
<tr>
<td>13</td>
<td>428</td>
<td>University of Cambridge</td>
</tr>
<tr>
<td>14</td>
<td>425</td>
<td>French National Centre for Scientific Research</td>
</tr>
<tr>
<td>15</td>
<td>400</td>
<td>Yale University</td>
</tr>
<tr>
<td>16</td>
<td>371</td>
<td>UNSW Sydney</td>
</tr>
<tr>
<td>17</td>
<td>369</td>
<td>Government of India</td>
</tr>
<tr>
<td>18</td>
<td>369</td>
<td>Shahid Beheshti University of Medical Sciences</td>
</tr>
<tr>
<td>19</td>
<td>366</td>
<td>Raymond and Ruth Perelman School of Medicine at the University of Pennsylvania</td>
</tr>
<tr>
<td>20</td>
<td>361</td>
<td>Cornell University</td>
</tr>
</tbody>
</table>

---

Match works with identified authors’ affiliations

```sql
WITH work_rors AS
(SELECT DISTINCT work_id, ror_id
FROM work_authors_rors
LEFT JOIN work_authors
ON work_authors_rors.work_author_id = work_authors.id),

ror_work_counts AS
(SELECT ror_id,
Count(*) AS number
FROM work_rors
GROUP BY ror_id),

ror_name_work_counts AS
(SELECT name,
number
from ror_work_counts
INNER JOIN research_organizations
ON ror_work_counts.ror_id = research_organizations.id),

unmatched_work_affiliations AS
(SELECT DISTINCT work_id, author_affiliations.name
FROM work_authors
INNER JOIN author_affiliations
ON work_authors.id = author_affiliations.author_id
LEFT JOIN work_authors_rors
ON work_authors_rors.work_author_id = work_authors.id
WHERE work_authors_rors.ror_id is null),

unmatched_affiliation_work_counts AS
(SELECT name, Count(*) AS number
FROM unmatched_work_affiliations
GROUP BY name),

all_work_counts AS
(SELECT *
FROM ror_name_work_counts
UNION
SELECT * FROM unmatched_affiliation_work_counts)

```

Output the top-20 affiliations according to number of published works

```sql
SELECT Rank() OVER (ORDER BY number DESC) AS rank, number, name
FROM all_work_counts
LIMIT 20;```
SELECT original_works.published_year, original_works.published_month, count(*)
FROM works AS original_works
INNER JOIN work_references ON work_references.work_doi = original_works.doi
INNER JOIN works AS cited_works ON work_references.doi = cited_works.doi
GROUP BY original_works.published_year, original_works.published_month
ORDER BY original_works.published_year, original_works.published_month;
# Extreme collaboration under COVID

<table>
<thead>
<tr>
<th>Rank</th>
<th>Author records</th>
<th>Affiliation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2352</td>
<td>Writing Committee for the REMAP-CAP Investigators for the Society of Critical Care Medicine Discovery Viral Infection and Respiratory Illness Universal Study (VIRUS): COVID-19</td>
</tr>
<tr>
<td>2</td>
<td>1731</td>
<td>REMAP-CAP Writing Committee for the REMAP-CAP Investigators</td>
</tr>
<tr>
<td>3</td>
<td>734</td>
<td>Registry Investigator Group</td>
</tr>
<tr>
<td>4</td>
<td>729</td>
<td>for the COVID-19 Phase 3 Prevention Trial Team</td>
</tr>
<tr>
<td>5</td>
<td>604</td>
<td>for the COVID-19 and Cancer Consortium</td>
</tr>
<tr>
<td>6</td>
<td>587</td>
<td>for the CORIMUNO-19 Collaborative Group</td>
</tr>
<tr>
<td>7</td>
<td>555</td>
<td>for the COVID-19 and Cancer Consortium (CCC19)</td>
</tr>
<tr>
<td>8</td>
<td>536</td>
<td>Shiraz University of Medical Sciences</td>
</tr>
<tr>
<td>9</td>
<td>412</td>
<td>for the PREP-IT Investigators</td>
</tr>
<tr>
<td>10</td>
<td>375</td>
<td>University of Oxford</td>
</tr>
<tr>
<td>11</td>
<td>369</td>
<td>for the RECOVERY-RS Collaborators</td>
</tr>
<tr>
<td>12</td>
<td>364</td>
<td>Universidade de Sào Paulo, Brazil</td>
</tr>
<tr>
<td>13</td>
<td>351</td>
<td>National Institute for Infectious Diseases “L. Spallanzani” IRCCS, Rome, Italy</td>
</tr>
<tr>
<td>14</td>
<td>336</td>
<td>ФКУЗ Российский научно-исследовательский противочумный институт «Микроб» Роспотребнадзора, Саратов, Российская Федерация</td>
</tr>
<tr>
<td>15</td>
<td>331</td>
<td>Tehran University of Medical Sciences</td>
</tr>
<tr>
<td>16</td>
<td>321</td>
<td>Hamad Medical Corporation</td>
</tr>
<tr>
<td>17</td>
<td>305</td>
<td>for the STOP-COVID Investigators</td>
</tr>
<tr>
<td>18</td>
<td>298</td>
<td>Fundação Oswaldo Cruz, Brazil</td>
</tr>
<tr>
<td>19</td>
<td>285</td>
<td>The WHO Rapid Evidence Appraisal for COVID-19 Therapies (REACT) Working Group for the Psoriasis Patient Registry for Outcomes, Therapy and Epidemiology of COVID-19 Infection (PsOProtect); the Secure Epidemiology of Coronavirus Under Research Exclusion for Inflammatory Bowel Disease (SECURE-IBD); and the COVID-19 Global Rheumatology Alliance (GRA)</td>
</tr>
<tr>
<td>20</td>
<td>276</td>
<td>Rheumatology Alliance (GRA)</td>
</tr>
</tbody>
</table>
SELECT Avg(author_number), Max(author_number) FROM ( 
    SELECT Count(*) AS author_number FROM works 
    LEFT JOIN work_authors ON works.doi = work_authors.work_doi 
    GROUP BY works.doi 
); 

5.47 7194
Convalescent plasma in patients admitted to hospital with COVID-19 (RECOVERY): a randomised controlled, open-label, platform trial

RECOVERY Collaborative Group†
Not an isolated case

```
SELECT works.doi, Count(*) AS author_number FROM works
LEFT JOIN work_authors
  ON works.doi = work_authors.work_doi
GROUP BY works.doi
ORDER BY Count(*) DESC
LIMIT 20;
```

```
SELECT Count(*) FROM (  
  SELECT Count(*) AS author_number FROM works
  LEFT JOIN work_authors
    ON works.doi = work_authors.work_doi
  GROUP BY works.doi
  HAVING author_number > 100
);
```

<table>
<thead>
<tr>
<th>DOI</th>
<th>Authors</th>
</tr>
</thead>
<tbody>
<tr>
<td>10.1016/s0140-6736(21)00897-7</td>
<td>7,194</td>
</tr>
<tr>
<td>10.1016/s0140-6736(21)00676-0</td>
<td>6,349</td>
</tr>
<tr>
<td>10.1016/s0140-6736(21)00163-5</td>
<td>6,303</td>
</tr>
<tr>
<td>10.1016/s0140-6736(21)01825-0</td>
<td>6,215</td>
</tr>
<tr>
<td>10.1093/bjs/znab336</td>
<td>5,549</td>
</tr>
<tr>
<td>10.1016/s0140-6736(21)00149-5</td>
<td>5,370</td>
</tr>
<tr>
<td>10.1016/s1470-2045(21)00493-9</td>
<td>5,203</td>
</tr>
<tr>
<td>10.1093/bjs/znab183</td>
<td>4,870</td>
</tr>
<tr>
<td>10.1038/s41586-021-03767-x</td>
<td>3,903</td>
</tr>
<tr>
<td>10.1200/jco.20.01933</td>
<td>3,647</td>
</tr>
<tr>
<td>10.1093/bjs/znaa051</td>
<td>3,608</td>
</tr>
<tr>
<td>10.1007/s00134-021-06448-5</td>
<td>2,013</td>
</tr>
<tr>
<td>10.1001/jama.2022.2910</td>
<td>1,805</td>
</tr>
<tr>
<td>10.1007/s00439-021-02397-7</td>
<td>1,577</td>
</tr>
<tr>
<td>10.1016/s2352-3018(21)00151-x</td>
<td>1,574</td>
</tr>
<tr>
<td>10.1016/s2214-109x(21)00289-8</td>
<td>1,555</td>
</tr>
<tr>
<td>10.1503/cjs.021321</td>
<td>1,431</td>
</tr>
<tr>
<td>10.1093/bjs/znab307</td>
<td>1,295</td>
</tr>
<tr>
<td>10.1186/s12967-021-03094-9</td>
<td>1,295</td>
</tr>
</tbody>
</table>
The dreaded Journal Impact Factor
\[ \text{IF}_y = \frac{\text{Citations}_y}{\text{Publications}_{y-1} + \text{Publications}_{y-2}}. \]
Journal Impact Factor

alexandria3k populate impact_data.db crossref 'April 2022 Public Data File from Crossref'
  --row-selection 'works.published_year BETWEEN 2019 AND 2021'
  --columns works.doi works.issn_print works.issn_electronic works.published_year \
    work_references.work_doi work_references.doi
alexandria3k populate impact_data.db journal-names

ATTACH 'impact_data.db' AS impact_data;

CREATE TABLE works_issn AS
  SELECT doi AS doi, published_year
    Coalesce(issn_print, issn_electronic) AS issn
  FROM impact_data.works
  WHERE issn is not null;

CREATE index works_issn_doi_idx ON works_issn(doi);

CREATE TABLE citations AS
  SELECT cited_work.issn, COUNT(*) AS citations_number
  FROM impact_data.work_references
    INNER JOIN works_issn AS published_work
      ON work_references.work_doi = published_work.doi
    INNER JOIN works_issn AS cited_work
      ON work_references.doi = cited_work.doi
  WHERE published_work.published_year = 2021
    AND cited_work.published_year BETWEEN 2019 AND 2020
  GROUP BY cited_work.issn;

CREATE TABLE publications AS
  SELECT issn, COUNT(*) AS publications_number
  FROM works_issn
  WHERE published_year BETWEEN 2019 AND 2020
  GROUP BY issn;

CREATE TABLE impact_factor AS
  SELECT publications.issn, citations_number, publications_number
  FROM publications
    LEFT JOIN citations
      ON citations.issn = publications.issn
  WHERE publications_number > 0;
  Cast(Coalesce(citations_number, 0) AS FLOAT) / publications_number AS impact_factor
  FROM publications
  WHERE publications_number > 0;
# Results

<table>
<thead>
<tr>
<th>ISSN</th>
<th>Title</th>
<th>IF</th>
</tr>
</thead>
<tbody>
<tr>
<td>0007-9235</td>
<td>CA A Cancer Journal for Clinicians</td>
<td>103.3</td>
</tr>
<tr>
<td>2092-6413</td>
<td>Experimental &amp; Molecular Medicine</td>
<td>86.0</td>
</tr>
<tr>
<td>0009-2665</td>
<td>Chemical Reviews</td>
<td>48.2</td>
</tr>
<tr>
<td>1546-0738</td>
<td>MMWR Surveillance Summaries</td>
<td>46.6</td>
</tr>
<tr>
<td>0092-8674</td>
<td>Cell</td>
<td>45.8</td>
</tr>
<tr>
<td>0028-4793</td>
<td>New England Journal of Medicine</td>
<td>45.6</td>
</tr>
<tr>
<td>0034-6861</td>
<td>Reviews of Modern Physics</td>
<td>44.7</td>
</tr>
<tr>
<td>0031-9333</td>
<td>Physiological Reviews</td>
<td>42.8</td>
</tr>
<tr>
<td>0306-0012</td>
<td>Chemical Society Reviews</td>
<td>40.7</td>
</tr>
<tr>
<td>2333-4436</td>
<td>Journal of Materials Physics and Chemistry</td>
<td>39.0</td>
</tr>
<tr>
<td>2058-8437</td>
<td>Nature Reviews Materials</td>
<td>38.9</td>
</tr>
<tr>
<td>1471-0072</td>
<td>Nature Reviews Molecular Cell Biology</td>
<td>38.5</td>
</tr>
<tr>
<td>2589-7780</td>
<td>EnergyChem</td>
<td>36.2</td>
</tr>
<tr>
<td>0079-6425</td>
<td>Progress in Materials Science</td>
<td>35.7</td>
</tr>
<tr>
<td>1078-8956</td>
<td>Nature Medicine</td>
<td>35.4</td>
</tr>
<tr>
<td>2333-8628</td>
<td>International Journal of Environmental Bioremediation &amp; Biodegradation</td>
<td>35.0</td>
</tr>
<tr>
<td>2367-3613</td>
<td>Living Reviews in Relativity</td>
<td>34.9</td>
</tr>
<tr>
<td>0066-4146</td>
<td>Annual Review of Astronomy and Astrophysics</td>
<td>34.2</td>
</tr>
<tr>
<td>0935-4956</td>
<td>The Astronomy and Astrophysics Review</td>
<td>32.9</td>
</tr>
<tr>
<td>1476-4598</td>
<td>Molecular Cancer</td>
<td>31.8</td>
</tr>
<tr>
<td>1474-1733</td>
<td></td>
<td>31.7</td>
</tr>
<tr>
<td>1057-5987</td>
<td>MMWR Recommendations and Reports</td>
<td>31.2</td>
</tr>
<tr>
<td>0732-0582</td>
<td>Annual Review of Immunology</td>
<td>30.5</td>
</tr>
<tr>
<td>1754-5692</td>
<td>Energy &amp; Environmental Science</td>
<td>30.0</td>
</tr>
<tr>
<td>1553-4006</td>
<td>Annual Review of Pathology Mechanisms of Disease</td>
<td>29.5</td>
</tr>
<tr>
<td>2058-7546</td>
<td>Nature Energy</td>
<td>28.4</td>
</tr>
<tr>
<td>2542-4351</td>
<td>Joule</td>
<td>28.2</td>
</tr>
<tr>
<td>1543-5008</td>
<td>Annual Review of Plant Biology</td>
<td>28.1</td>
</tr>
<tr>
<td>2520-8489</td>
<td>Electrochemical Energy Reviews</td>
<td>27.9</td>
</tr>
<tr>
<td>1074-7613</td>
<td>Immunity</td>
<td>27.5</td>
</tr>
</tbody>
</table>
SELECT doi, Count(*) 
FROM work_references 
GROUP BY doi 
ORDER BY count(*) DESC 
LIMIT 10;

31" elapsed time

39 715 citations
alexandria3k query crossref 'April 2022 Public Data File from Crossref' --partition \
--query "SELECT title FROM work_references
    LEFT JOIN works
    ON work_references.work_doi = works.doi
    WHERE work_references.doi = '10.1103/physrevlett.77.3865''"

"Solid-liquid density and spin crossovers in (Mg, Fe)O system at deep mantle conditions"
Two-Dimensional BAs/InTe: A Promising Tandem Solar Cell with High Power Conversion Efficiency
Fatigue of graphene
Energetics of paramagnetic oxide clusters: the Fe(<scp>iii</scp>) oxyhydroxy Keggin ion
Stochastic many-body perturbation theory for Moiré states in twisted bilayer phosphorene
Dual-hybrid direct random phase approximation and second-order screened exchange with nonlocal van der Waals correlations for noncovalent interactions
Prediction on temperature dependent elastic constants of “soft” metal Al by AIMD and QHA
Triple VTe2/graphene/VTe2 heterostructures as perspective magnetic tunnel junctions
On the nature of homo- and hetero-dinuclear metal–metal quadruple bonds — Analysis of the bonding situation and benchmarking DFT against wave function methods
The extraordinary stability imparted to silver monolayers by chloride
Efficient Band Gap Prediction for Solids
Importance of Electronic Relaxation for Inter-Coulombic Decay in Aqueous Systems
Prediction of Reorganization Free Energies for Biological Electron Transfer: A Comparative Study of Ru-Modified Cytochromes and a 4-Helix Bundle Protein
...
--- Find the most cited articles in the period 2019-2021 --- published within that period

```sql
SELECT works.doi, Count(*)
FROM work_references
LEFT JOIN works ON work_references.doi = works.doi
WHERE published_year BETWEEN 2019 AND 2021
GROUP BY works.doi
ORDER BY Count(*) DESC
LIMIT 10;
```

---

48” elapsed time

21 424 citations
Author h5-index

• Zhanhu Guo = 76 (15 papers / year)
• 12 authors > 60
• 100 > 38
Thousands of scientists publish a paper every five days

To highlight uncertain norms in authorship, John P. A. Ioannidis, Richard Klavans and Kevin W. Boyack identified the most prolific scientists of recent years.
How is this possible?

• Clustering coefficient of distance 2 citations
• **Significantly** different from other highly-cited papers
  – For h5 > 50: median 0.05
  – For random sample: median 0.03
  – Mann-Whitney U test $U_M=781$, $p$-value 0.0006
Implementation
Plugin-based architecture

- CLI
- Python API
- Data sources: Crossref, PubMed, ORCID, USPTO, ROR, DOAJ, ...
- Processes: Work topics, Affiliation ROR, USPTO DOI, ...
- Plugin API
Crossref key implementation ideas

• SQLite + virtual tables
• Database partitioning, partition index
• Query tracing
• Realized vertical slices of partitions for queries
• PK, FK table with matched population query records
How to run Crossref query on 1 TB (simple case)

alexandria3k query crossref 'April 2022 Public Data File from Crossref'
--query "SELECT doi FROM work_references where doi is not null"

CREATE VIRTUAL TABLE work_references USING filesource();

SELECT doi FROM work_references where doi is not null;
How to run Crossref query on 1 TB

```
SELECT title FROM work_references
LEFT JOIN works ON work_references.work_doi = works.doi
WHERE work_references.doi = '10.1103/physrevlett.77.3865';

ATTACH DATABASE 'file:virtual?mode=memory&cache=shared' AS virtual;

CREATE TABLE works AS SELECT title, doi
FROM virtual.works WHERE virtual.works.container_id=1453;

CREATE TABLE work_references AS SELECT doi, work_doi
FROM virtual.work_references WHERE virtual.work_references.container_id= 1453;

SELECT title FROM work_references
LEFT JOIN works ON work_references.work_doi = works.doi
WHERE work_references.doi = '10.1103/physrevlett.77.3865';
```
Crossref population: simple case

```
INSERT INTO populated.works
    SELECT works.title, works.doi FROM works
    WHERE works.container_id = 0;

INSERT INTO populated.work_authors
    SELECT work_authors.* FROM work_authors
    WHERE work_authors.container_id = 0;

INSERT INTO populated.works
    SELECT works.title, works.doi FROM works
    WHERE works.container_id = 1;

[...]```
Conditional Crossref population 1/2

alexandria3k populate lis.db crossref ... \
--row-selection "work_subjects.name = 'Library and Information Sciences' "
--columns works.title works.doi work_authors.orcid work_subjects.*

ATTACH DATABASE 'lis.db' AS populated;

SELECT DISTINCT 1 FROM works, work_authors, author_affiliations, ...
WHERE work_subjects.name = 'Library and Information Sciences';

CREATE TABLE populated.works(doi, container_id, title, ...);
[...]

CREATE TEMP TABLE temp_works AS
  SELECT id, rowid FROM works WHERE container_id = 0;

CREATE TEMP TABLE temp_work_subjects AS
  SELECT rowid, name, work_id FROM work_subjects WHERE container_id = 0;

CREATE TEMP TABLE temp_work_authors AS
  SELECT rowid, work_id FROM work_authors WHERE container_id = 0;
[...]

Tables with PKs, FKs and query fields

Traced query & query trace results

Populated tables
CREATE TEMP TABLE temp_matched AS
SELECT works.id, works.rowid
FROM temp_works AS works
LEFT JOIN temp_work_subjects AS work_subjects
ON works.id = work_subjects.work_id
WHERE (work_subjects.name = 'Library and Information Sciences');

INSERT INTO populated.work_authors
SELECT work_authors.orcid
FROM work_authors
WHERE work_authors.container_id = 0
AND EXISTS (SELECT 1
FROM temp_matched AS temp_works
LEFT JOIN temp_work_authors
ON temp_works.id = temp_work_authors.work_id
AND work_authors.rowid = temp_work_authors.rowid);

Key to all partition records matching the specified condition

Topologically ordered table JOINs

Populate tables with partition’s data based on matched records
ORCID/USPTO key implementation ideas

• Stream-based
  – Web fetch
  – Decompress
  – Tar records

• Skip XML parsing where possible
Issues and limitations

• Low ORCID coverage:
  – Only 17/360 million author records
• Affiliations missing / appear in diverse forms
• Only 11% of Crossref records have an abstract
• Subjects cover only Scopus-indexed journals
• Difficulty of determining “citable items”
Way forward

• Help community to conduct studies
• Integrate more OA data
  – arXiv, DBLP, MESH, PLoS taxonomy, ... 
• Improve processes
  – Author & org disambiguation, topic classification, ... 
• Evangelize more and better data availability
  – ORCID
  – Publication metadata improvements
Thank you!

github.com/dspinellis/alexandria3k
🌐 www.spinellis.gr
𝕏 @CoolSWEng
✉️ @CoolSWEng@mastodon.acm.org
✉️ dds@aueb.gr
Unix Tools: Data, Software and Production Engineering

Grow from being a Unix novice to Unix wizard status! Process big data, analyze software code, run DevOps tasks and excel in your everyday job through the amazing power of the Unix shell and command-line tools.

6 weeks
4-6 hours per week
Self-paced
Progress at your own speed
Free
Optional upgrade available

There is one session available:
5,685 already enrolled! After a course session ends, it will be archived.

Starts Sep 20
Enroll