

Publishing Trends

On O'Reilly Learning Platform



Project Aim

Explore trending topics for tech books, analyzing books metadata over time (2018 - 2023).

Github: [giuseppevallarelli/publishing_trends](https://github.com/giuseppevallarelli/publishing_trends)

Giuseppe Vallarelli
Master in BI & BDA
Anno 2023-24

Software can be chaotic, but we make it work



Expert

Trying Stuff Until it Works

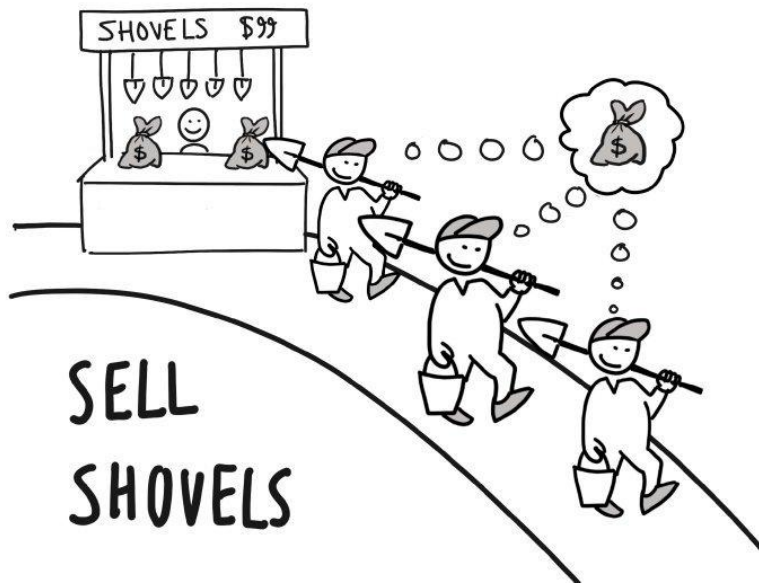
O RLY?

The Practical Developer
@ThePracticalDev

Gold Rush

<insert your favorite trending topic>

WHEN EVERYONE DIGS FOR GOLD



O'Reilly Website

Main page: oreilly.com/search

O'REILLY Topics Start Learning

Search 50,000+ courses, events, titles, and ...

What's New

Browse format

[All](#) [Books](#) [Courses](#) [Videos](#)

[Live Events](#) [Interactive](#) [Certifications](#)

[Audiobooks](#) [Playlists](#)

Filters

Topics

Publishers

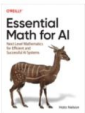
Rating

Publication date

Sort by: Relevance

61,363 Search results

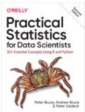
Book



Essential Math for AI

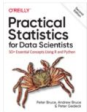
By [Hala Nelson](#)

O'Reilly Media, Inc. • January 2023

★★★★★ [2 reviews](#)  602 pages

More Info


Book



Practical Statistics for Data Scientists, 2nd Edition


By [Peter Bruce](#), [Andrew Bruce](#) and [Peter Gedeck](#)

O'Reilly Media, Inc. • May 2020

★★★★★ [9 reviews](#)  360 pages

More Info


Book





Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow, 3rd Edition


By [Aurélien Géron](#)

O'Reilly Media, Inc. • October 2022

★★★★★ [12 reviews](#)  861 pages

+ Add 

+ Add 

+ Add 

1 - 10 of 61,363

1 2 3 4 5

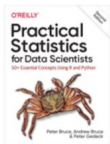
O'Reilly Website

Detail: /library/view/practical-statistics-for/{isbn}

▼ More Info

Book

+ Add  Read



Practical Statistics for Data Scientists, 2nd Edition

By [Peter Bruce](#), [Andrew Bruce](#) and [Peter Gedeck](#)

O'Reilly Media, Inc. • May 2020

★★★★★ 9 reviews  360 pages 1

Statistical methods are a key part of data science, yet few data scientists have formal statistical training. Courses and books on basic statistics rarely cover the topic from a data science perspective. The second edition of this popular guide adds comprehensive examples in Python, provides practical guidance on applying statistical methods to data science, tells you how to avoid their misuse, and gives you advice on what's important and what's not.

Many data science resources incorporate statistical methods but lack a deeper statistical perspective. If you're familiar with the R or Python programming languages and have some exposure to statistics, this quick reference bridges the gap in an accessible, readable format.

With this book, you'll learn:

- Why exploratory data analysis is a key preliminary step in data science
- How random sampling can reduce bias and yield a higher-quality dataset, even with big data
- How the principles of experimental design yield definitive answers to questions
- How to use regression to estimate outcomes and detect anomalies
- Key classification techniques for predicting which categories a record belongs to
- Statistical machine learning methods that "learn" from data
- Unsupervised learning methods for extracting meaning from unlabeled data

[Learn More](#)

Related topics

[Statistics](#)

[Data Science Tasks](#)

[Data Science](#)

[Data](#)

2

O'Reilly Website

Publishers

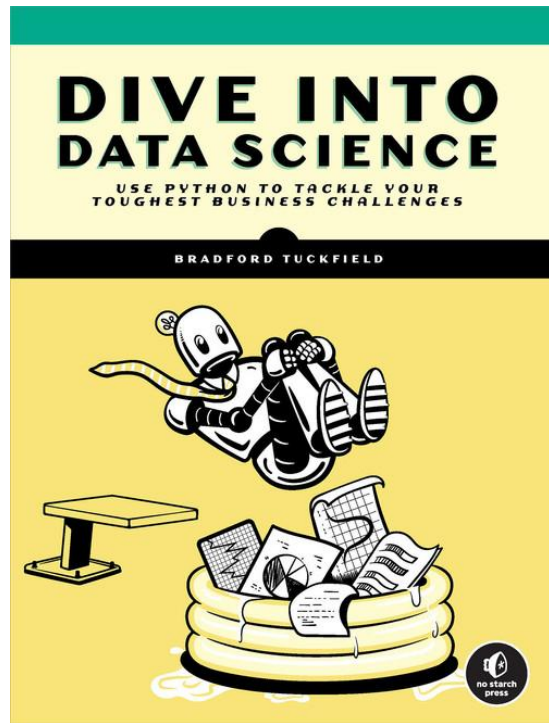


Dataset

JSON Metadata and Cover image

```
{
  "product_id": "9781098156879",
  "title": "Dive Into Data Science",
  "authors": ["Bradford Tuckfield"],
  "description": "...",
  "language": "en",
  "categories": [["Data", "Data Science"]],
  "url": "https://learning.oreilly.com/library/view/-/9781098156879/",
  "cover_image": "https://learning.oreilly.com/library/cover/9781098156879/",
  "publication_date": "2023-07-04",
  "publishers": ["No Starch Press"],
  "page_count": 288,
  "average_rating": null
}
```

5139 Books and related covers ≈ 500 MBs



BI Architecture

Tools adopted

Steps

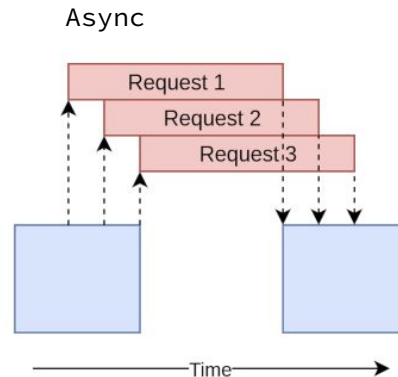
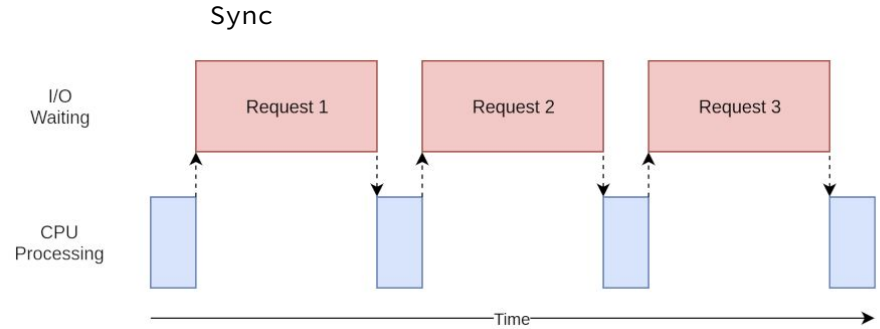
- Web scraping: Python (Asyncio) for books metadata and covers, taking advantage of internal REST API (< 5mins)
- ETL: Pandas and a pinch of Seaborn in a Jupyter Notebook
- DB: SQLite with a few tables (Book, Author and Category)
- Reporting and Graphs: Tableau



Asynchronous retrieval

Some processes are CPU-bound: they consist of a series of instructions which need to be executed one after another until the result has been computed. All of the time they are running is time that they are making full use of the computer's facilities (give or take).

Other processes, however, are IO-bound: they spend a lot of time sending and receiving data from external devices or processes, and hence often need to start an operation and then wait for it to complete before carrying on. During the waiting they aren't doing very much.



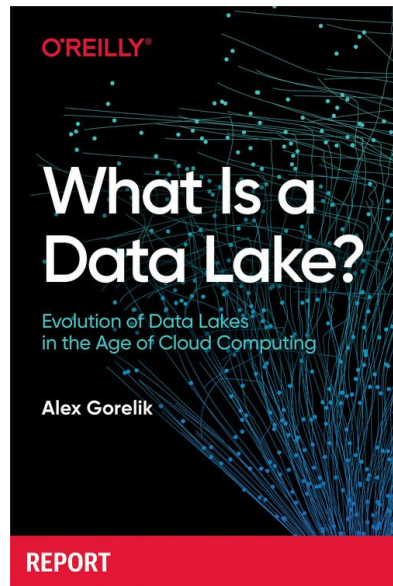
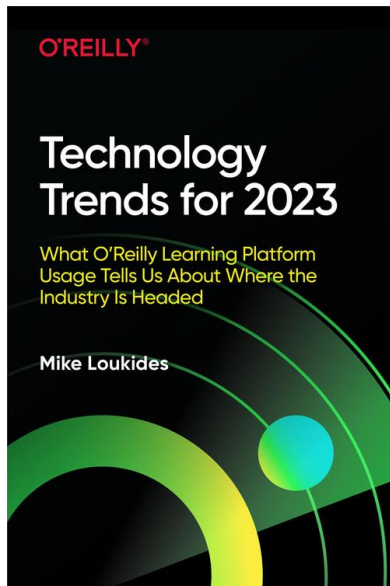
A cleaning tale

Removing O'Reilly Reports from Dataset

Some entries don't belong to the dataset, because they're report publications (O'Reilly), they usually have a low *page_count*.

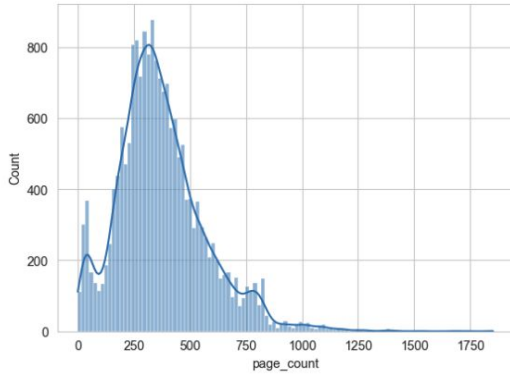
We filter them out with 3 different strategies:

1. Category (*Radar*)
2. Outliers (*page_count*)
3. OCR scan (*image text content*)



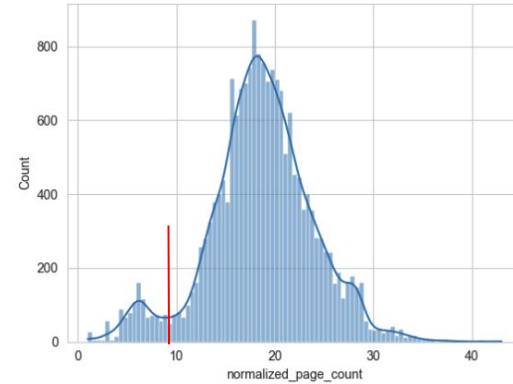
A cleaning tale

Identifying Outliers: IQR Method



skew ≈ 1.0

$\sqrt{\text{page_count}} \Rightarrow$



skew ≈ -0.20

$\text{IQR} = Q3 - Q1$

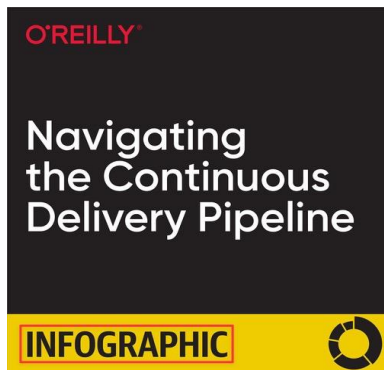
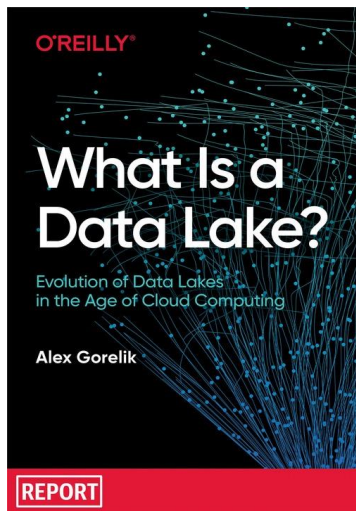
Red bar (lower fence) = $Q1 - 1.5 (\text{IQR})$

We drop all the entries below the red bar

<https://online.stat.psu.edu/stat200/lesson/3/3.2>

A cleaning tale

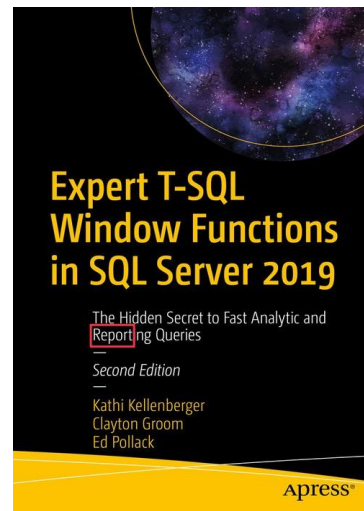
OCR: looking for Report / Infographic in the Cover text



Watch out ! Report / Infographic might be part of the text :-)

We can still check:

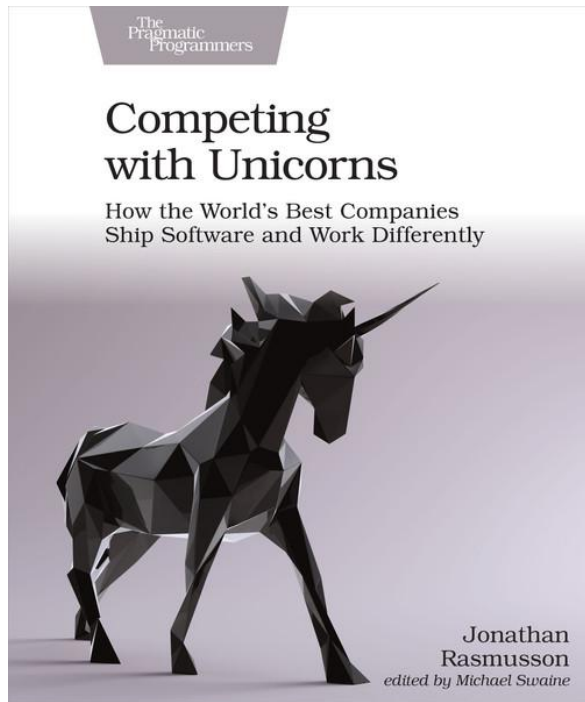
1. title attribute
2. publisher
3. report/infographic is the last word in the cover



A cleaning tale

Lessons learned

- You may not know in advance which data might prove to be useful (e.g. cover images), so retrieve as much data as possible.
- Start with the simplest transformation that might work (e.g. Radar category) and move up to more complicated stuff (comput. expensive).
- Math is a great ally (missing data, outliers, etc.), use it !-)



Publishers

Number of books per Publisher (2018-2023)



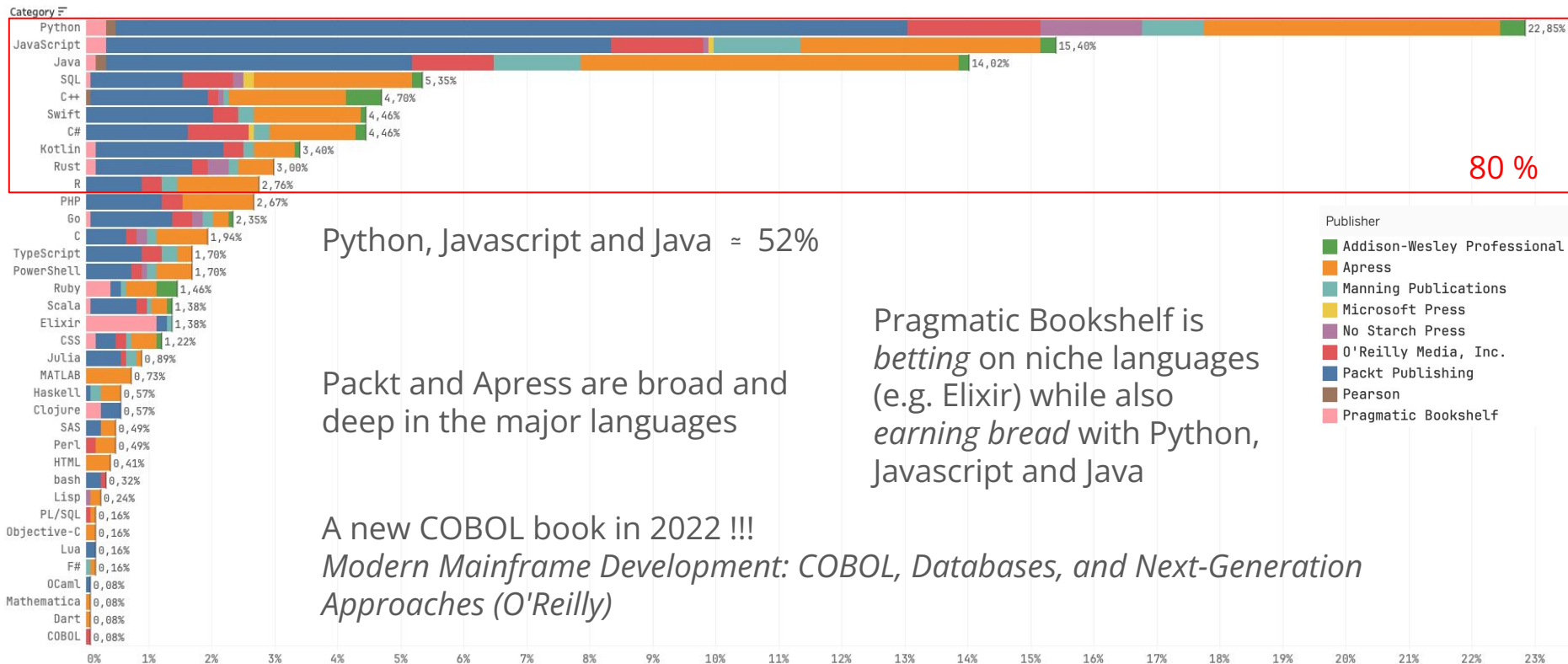
Tier 1: Packt Publishing, Apress

Tier 2: O'Reilly, Manning

Tier 3: Addison-Wesley, No Starch Press, Microsoft Press, Pragmatic Bookshelf, Pearson

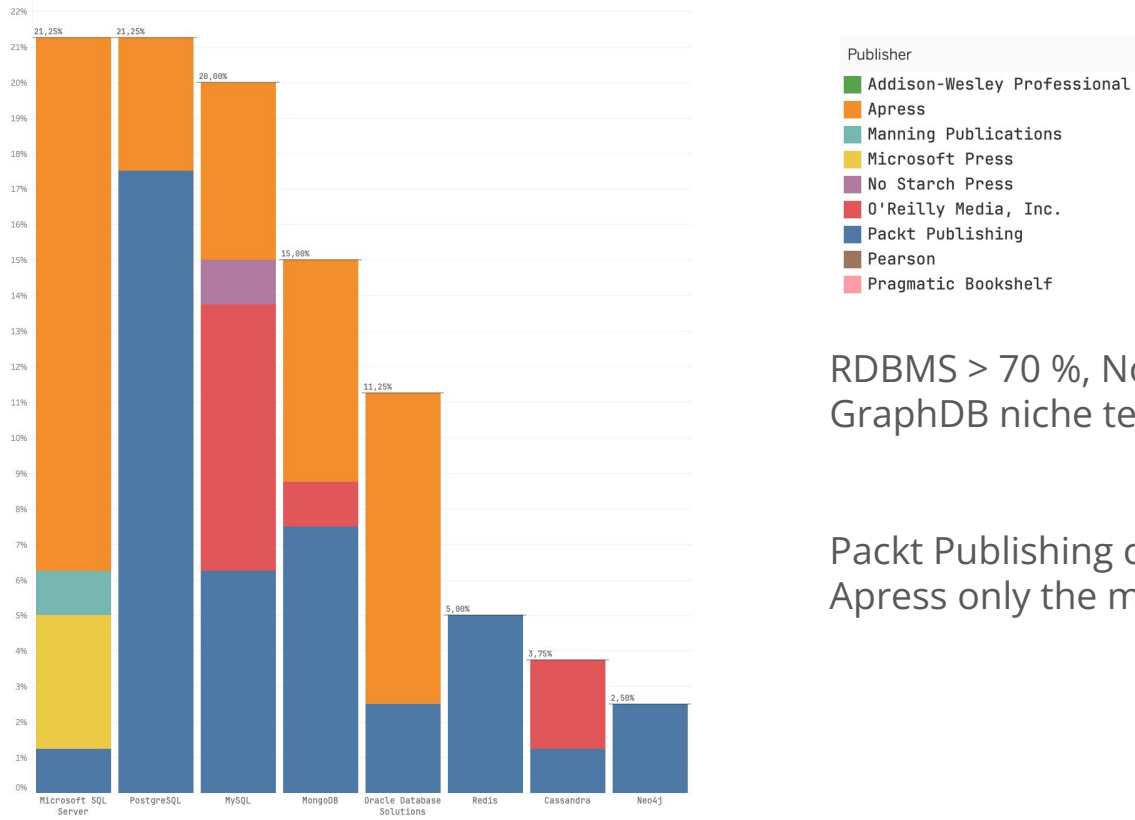
Publishers

Programming, scripting and markup languages



Publishers

Databases

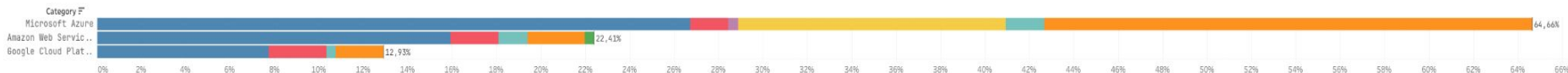


RDBMS > 70 %, NoSQL for specialized workloads, GraphDB niche technology.

Packt Publishing covers all database solutions, while Apress only the major ones in terms of market share.

Publishers

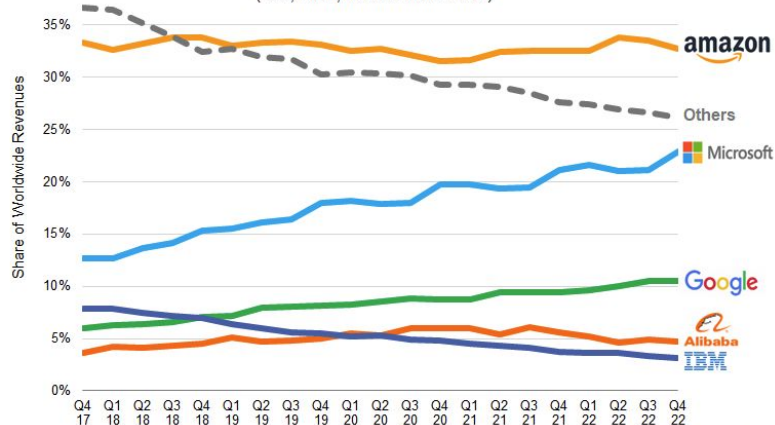
Cloud Providers



Publisher

- Addison-Wesley Professional
- Apress
- Manning Publications
- Microsoft Press
- No Starch Press
- O'Reilly Media, Inc.
- Packt Publishing
- Pearson
- Pragmatic Bookshelf

Cloud Provider Market Share Trend
(IaaS, PaaS, Hosted Private Cloud)



Source: Synergy Research Group

For our publishers (dataset), Microsoft Azure is already dominant.

<https://techcrunch.com/2023/02/06/even-as-cloud-infrastructure-market-growth-slows-microsoft-continues-to-gain-on-amazon/>

Publishers

Themes

Category	F	Year of Publication Date				
		2018	2019	2020	2021	2022
ML & AI		15,97%	17,51%	18,30%	16,64%	12,29%
Web Development		14,43%	12,62%	12,26%	9,87%	8,64%
Data Engineering		10,78%	9,79%	10,19%	9,69%	10,30%
Cloud Computing		8,68%	8,47%	9,62%	9,14%	12,29%
Business		9,38%	9,60%	9,25%	8,23%	10,63%
Security		5,88%	7,91%	8,30%	8,78%	9,80%
Software Architecture		4,06%	5,65%	4,72%	6,95%	7,48%
Design		3,64%	3,95%	4,53%	4,02%	6,31%
Internet of Things (IoT)		5,74%	3,58%	4,91%	5,30%	3,32%
DevOps		4,90%	3,20%	3,21%	4,75%	5,32%
Game Development		3,78%	3,58%	3,58%	3,11%	2,99%
Blockchain / Decentralized Apps		3,36%	4,14%	2,64%	1,10%	1,50%
Mobile Development		3,08%	3,20%	2,08%	2,93%	1,83%
Data Visualization		2,38%	2,07%	1,51%	2,38%	1,33%
Agile		1,68%	1,32%	1,89%	2,01%	1,33%
QA / Testing		1,26%	1,13%	0,94%	2,38%	1,66%
Math, Science, Engineering		0,84%	1,51%	0,75%	1,28%	1,00%
User Experience (UX)		0,98%	1,32%	1,13%	0,73%	0,33%
Cryptocurrency		1,40%	1,69%		0,18%	0,66%
Robotics		1,40%	0,38%	1,13%	0,91%	0,33%
Quantum Computing		0,14%	0,38%	0,75%	0,91%	1,00%
Soft Skills		0,42%	0,75%	0,19%	0,91%	1,00%
IT Certifications			0,38%	0,57%	0,55%	1,00%
Encryption / Cryptography			0,38%	0,19%	0,55%	0,33%
Edge Computing			0,19%	0,19%		0,66%

Data related to 2023 not complete

"A new threat to financial stability lurks in the cloud" (Financial Times)

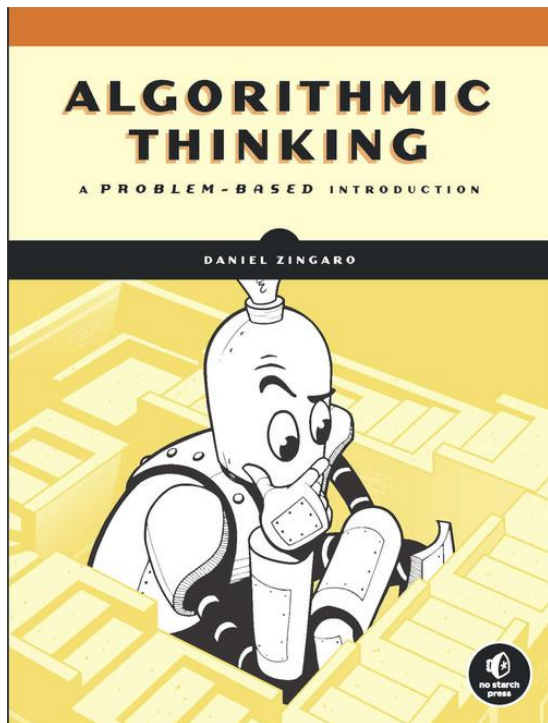
...more than 90 per cent of the members of the American Bankers Association are shifting activity on to the cloud, although more than 80 per cent say this is at an early stage...

Cloud computing echoes this so-called Spof (single point of failure) issue, as Michael Hsu, acting Comptroller of the Currency, told the BIS earlier this year. Most notably — and as Brussels often complains — the cloud is dominated by an oligopoly of Amazon, Microsoft and Google. If one of those players suffered a big cyber attack, weather-linked disruption or simply went bankrupt, that would rock the system...

Chinese hackers breached US government emails via Microsoft Cloud exploit

Conclusions

Publishers, books and related themes



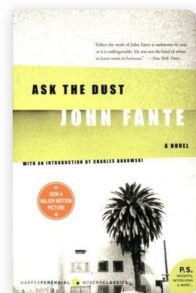
The industry doesn't evolve much year over year as many people think. Being competent at fundamentals is extremely important.

We may still be in the early stages of Cloud adoption, Security will become a central theme in the future.

Next steps 1/2

Publishers comparison

goodreads



Want to read

Buy on Amazon



The Saga of Arturo Bandini #3

Ask the Dust

John Fante, Charles Bukowski (Introduction)

★★★★★ 4.11

33,566 ratings · 2,421 reviews

"Fante was my god." —Charles Bukowski, in his introduction to *Ask the Dust*

Arturo Bandini, a young, struggling Italian-American writer living in a seedy hotel in 1930s Los Angeles, falls hard for the elusive, mocking, unstable Camilla Lopez, a Mexican waitress. The pair embark on a strange and strained love-hate relationship, which slowly, but inexorably, descends into the realm of madness.

Show more

Genres [Fiction](#) [Classics](#) [Novels](#) [American](#) [Literature](#) [The United States Of America](#) [Americana](#) [...more](#)

165 pages, Paperback

First published January 1, 1939

Unfortunately it was not possible to compare publishers, having lots of undefined values for the average book rating, moreover no data was available on the number of users that rated each book, but we could enrich the dataset with data from Goodreads!

Measuring scores spread with book ratings could give a measure of publisher inherent quality.

Next steps 2/2

Book grading or simply helping Nate

goodreads

Pattern Recognition and Machine Learning



Nate

91 reviews

5 followers

Follow



September 25, 2009

Even with the help of a nuclear physicist turned neurophysiology data analyst, I couldn't work beyond the first four chapters, and perhaps only a percentage of those. However, the efforts are rewarding. If you have read the entirety of this book, and understand it, then I would very much like to replace part of my brain with yours.

54 likes

 Like  Comment ...

We could use goodreads reviews and book descriptions to classify books into levels of complexity: *beginner*, *intermediate* and *advanced*.