

Open Source as a Research Method

From the research paper to billions of devices

Daniel Lemire, professor
Université du Québec (TÉLUQ)

blog: <https://lemire.me>

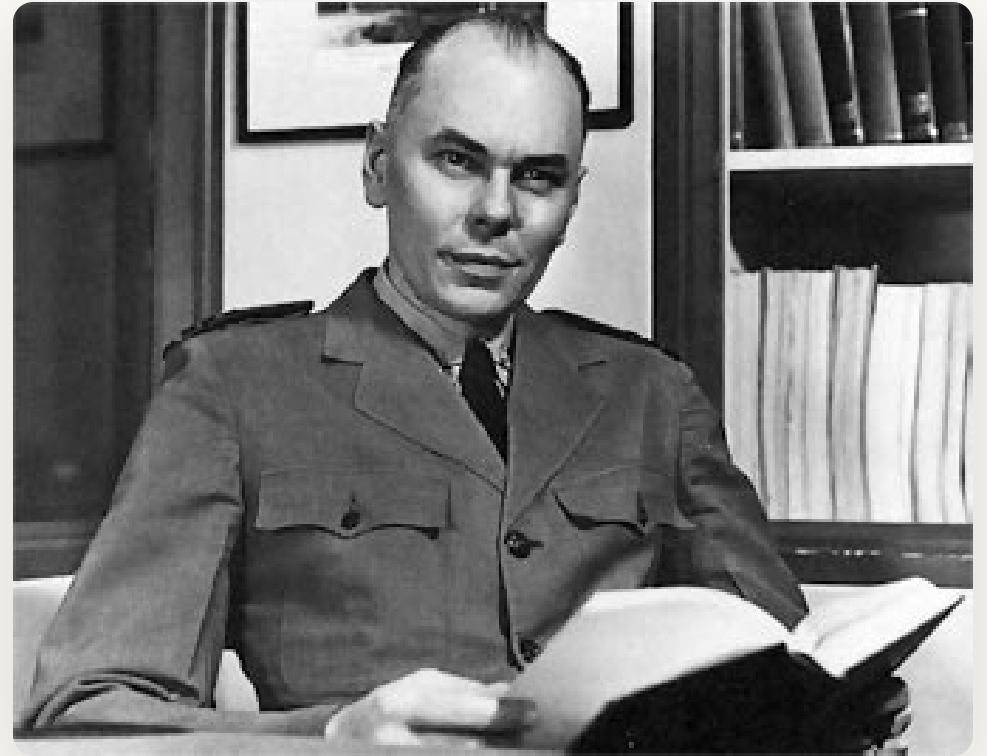
X: [@lemire](#) · GitHub: github.com/lemire

“

The problem in this business isn't to keep people from stealing your ideas; it's making them steal your ideas.

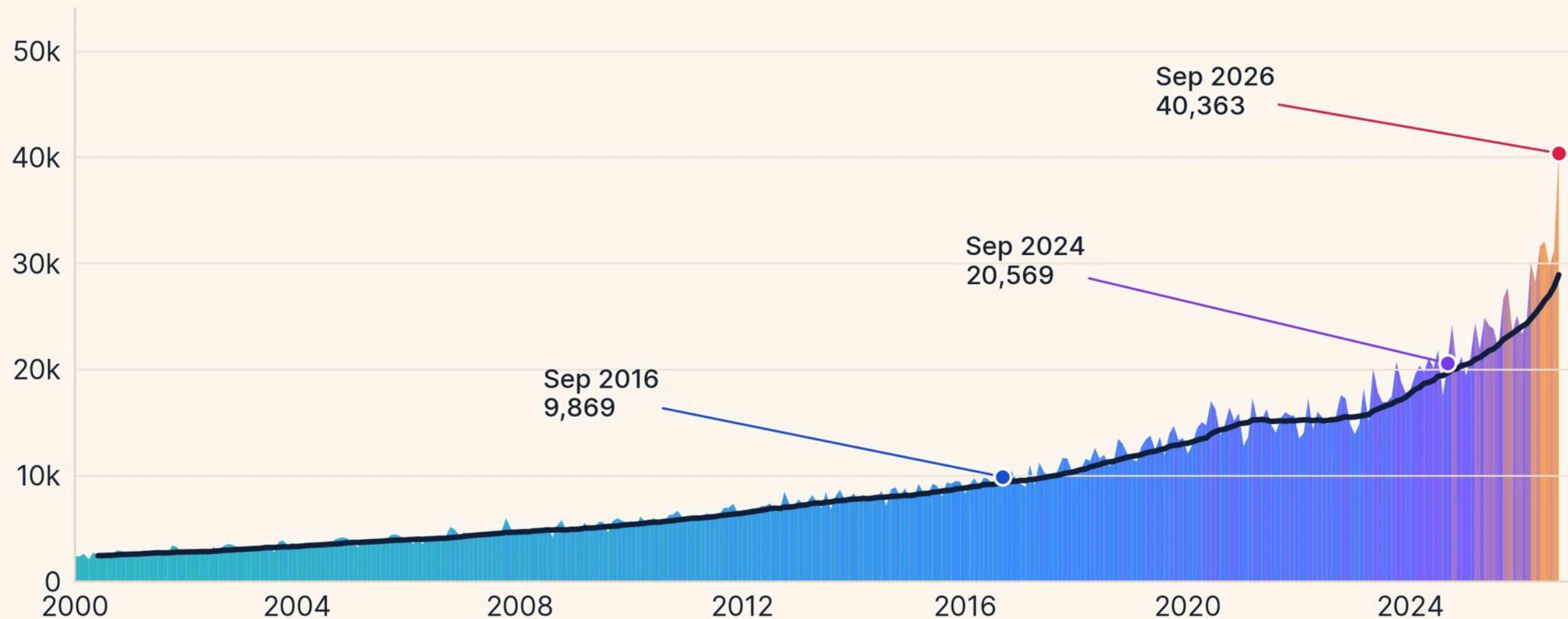
Howard Aiken

quoted by David Patterson, 2017 Turing Award



January 2000 to September 2026. October 2026 left out: the month is not finished.

Papers posted to arXiv each month



Source: arXiv monthly submissions. Color: the month. Dark line: 12-month average.

The linear model of innovation

- The professor thinks. The professor publishes.
- The engineer reads, designs.
- The customer consumes.

“

The industrial revolution was driven by demand for cheap and better clothing for the masses.



David Monniaux, CC BY-SA 4.0

“

In effect, our customer has vanished and been replaced by either the whales or a vacuum. This loss of our customer has resulted in a collection of bad effects. First, there are no “real world” clients to keep us focused.

Michael Stonebraker — “Where We Have Failed”
Making Databases Work, 2018



David Monniaux, CC BY-SA 4.0

A story: simdjson

Step 1 — parsing JSON

2019

- JSON is everywhere, and parsing it is **slow**

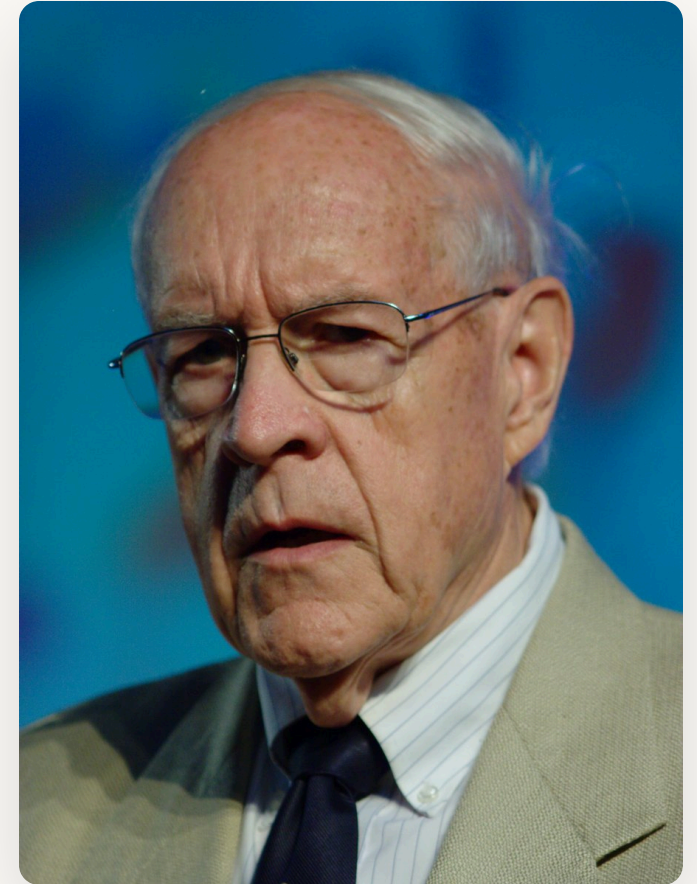
```
{
  "Image": {
    "Width": 800,
    "Height": 600,
    "Title": "View from 15th Floor",
    "Thumbnail": {
      "Url": "http://www.example.com/81989943",
      "Height": 125,
      "Width": 100
    }
  }
}
```

“

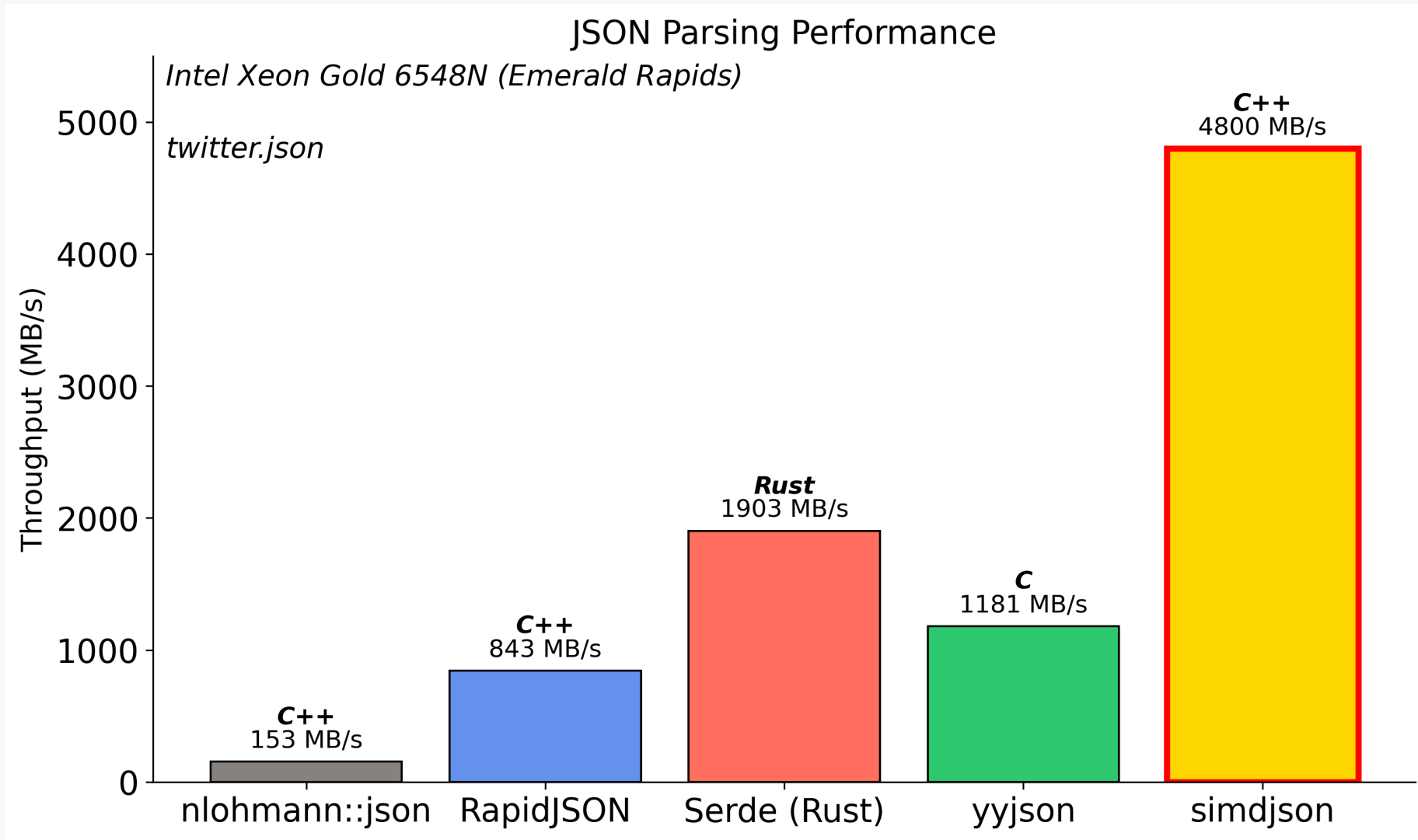
The scientist builds in order to study; the engineer studies in order to build.

Frederick P. Brooks, Jr.

“The Computer Scientist as Toolsmith II,” CACM, 1996



David Monniaux, CC BY-SA 3.0



The real publication

“
An article about computational science in a scientific publication is not the scholarship itself, it is merely advertising of the scholarship. The actual scholarship is the complete software development environment and the complete set of instructions which generated the figures.

Jon Claerbout — as recorded by Buckheit and Donoho, 1995

Step 2 — bottleneck: numbers

REVEALED BY USE

- Once everything else is fast, parsing **floating-point numbers** becomes the bottleneck.
- Result: the **Eisel-Lemire** algorithm, exact and fast (*Number Parsing at a Gigabyte per Second*).
- Adoption: Rust, Go, .NET, C++ — the standard libraries themselves.

Where do ideas come from?

“

I never come up with anything by going off to a mountaintop to think. Instead, my ideas come from two sources: talking to real users with real problems and then trying to solve them. This ensures I come up with ideas that somebody cares about and the rubber meets the road and not the sky.

Michael Stonebraker — 2014 Turing Award · Ingres, PostgreSQL, Vertica

Step 3 — another bottleneck: UTF-8

REVEALED BY USE

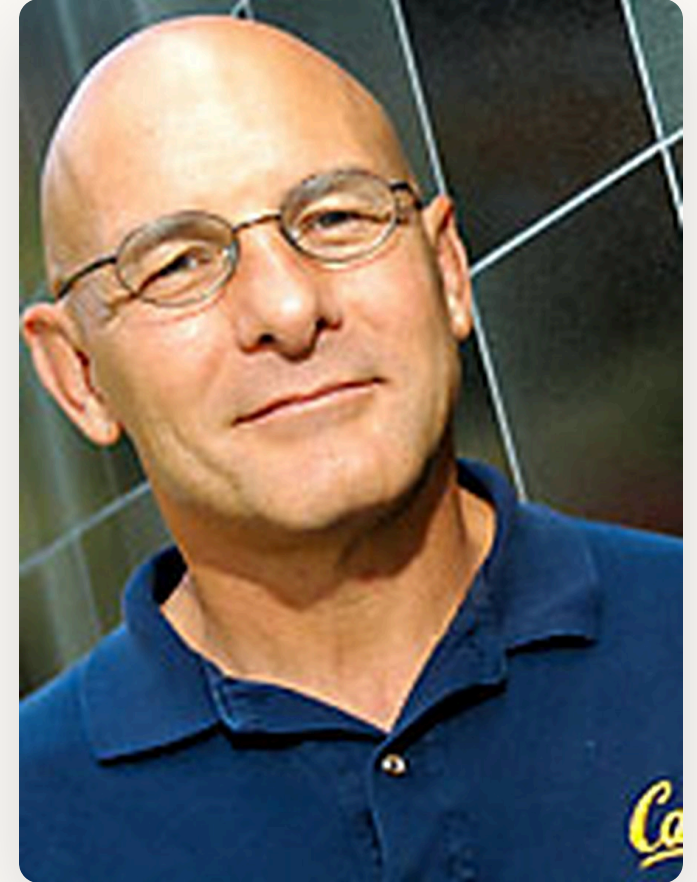
- Parsing JSON also means **validating** UTF-8: security depends on it.
- Goal reached: validation in **less than one instruction per byte**.
- That piece broke off to become a library of its own: **simdutf**.
- Adoption: Node.js, Bun, Deno, browsers — Unicode validation and transcoding for the web.

“

For my successful projects, the early indicators were not papers or awards, but non-researchers trying to use our ideas.

David Patterson

“How to Build a Bad Research Center,” CACM, 2014



Peg Skorpinski, CC BY-SA 3.0

Thank you

Daniel Lemire · Université du Québec (TÉLUQ)

blog: <https://lemire.me>

X: [@lemire](#) · GitHub: github.com/lemire

simdjson · fast_float · simdutf · Roaring Bitmaps