Doing investigative journalism with Python at The Marshall Project

What is The Marshall Project?

“The Marshall Project is a nonpartisan, nonprofit news organization that seeks to create and sustain a sense of national urgency about the U.S. criminal justice system. We have an impact on the system through journalism, rendering it more fair, effective, transparent and humane.”

What is this?

This is a document that accompanies our keynote chat at the 2025 PyCon US conference that shares some of the stories we talked about and reinforces some of the themes we discussed.

Journalism that uses Python can take different forms

- Bespoke databases create data from disparate sources that isn’t already being synthesized comprehensively
    - The Next to Die
- Data-driven investigations use administrative data or turn documents into data in order to identify patterns or quantify how often something occurs
    - In New York Prisons, Guards Who Brutalize Prisoners Rarely Get Fired

Data can have a long tail

- In terms of both reporting and data maintenance, we have to weigh continuing to follow certain topics against focussing on urgent, emerging stories.
    - What 120 Executions Tell Us About Criminal Justice in America is an analysis of the entire The Next to Die database published when we decided to stop tracking executions.
- Sometimes a topic can become newly relevant and data gathered and analysed for one story can be useful to cover a topic that has reemerged.
    - New York Prison Nurses Face Choice Between Patients and Abusive Guards includes a new analysis of the New York correctional officer discipline data that was a big part of the original investigation.
    - At other times, a particular data set, type of data, or analytic path may be a slow-burning fascination across the newsroom. In California, Climate Chaos Looms Over Prisons — and Thousands of Prisoners — in a Lakebed, used prison location data that had long been a fascination, and compelled an update to the pipeline to consume the data.

Even with code, data journalism at an “everywhere” scale is tricky

- In the U.S., there isn’t really one criminal justice system. There are thousands of systems shaped by local laws and practices.
- This means that both conceptually, and in terms of data availability, making comparisons between jurisdictions, or even providing a comprehensive analysis is challenging.
- This isn’t just a challenge for news organizations. The Marshall Project has covered issues with the federal government’s own collection of crime data from local agencies, such as in this 2024 story: Crime Rates and the 2024 Election: What You Need to Know.
- Even when we set out to collect data nationwide, fighting public records battles in multiple states, gaps created by agency data management practices, and different capabilities of agency records management systems can mean that nationwide data reporting doesn’t go as planned. This was the case for In Chicago, Shootings Go Unsolved as Gun Possession Arrests Rise, a story for which reporters collected arrest data for cities across the U.S., but ultimately focussed on Chicago. The story’s methodology highlights some of the work required to gain insights from even one city’s data.

Engaging with our work

- All of our work is available on the web at https://www.themarshallproject.org/, without a paywall. Our newsletters are a great entry point to our journalism. Opening Statement is a daily overview of reporting about the criminal justice system both from The Marshall Project and other outlets. Closing Argument is a weekly newsletter that dives into one timely topic. Geoff has occasionally written pieces for Closing Argument, focussing on unhoused people and their experience with the justice system.
- We are a nonprofit organization, so donations are very important for sustaining our work.
- Investigate This! Is an initiative to make some criminal justice data more accessible, particularly to local reporters and researchers.
- If you work with criminal justice data, whether at a government agency, as a researcher or developing government technology, we’d love to talk to you! Please reach out.

Engaging with data journalism, broadly

- NICAR is the largest conference for people doing the kind of work that we do. The schedule for the 2025 conference gives a good sense of the topics and skills highlighted by the conference.
    - Joining IRE, the organization that puts on the conference, gets you access to decades of resources for working with particular data sets, or computational

    methods used by journalists. The tip sheets collected for the 2025 NICAR conference give a good sense of what these resources look like.
    - IRE also sponsors the Philip Meyer Awards, which highlights investigative data reporting. Reading about the award winners and their methodologies is a great resource that reflects both analytical methods as well as technical choices.
- The Sigma Awards is another great place to find robust data journalism and the methods behind it.
- News Nerdery is a Slack instance where a lot of data journalists hang out. The #python, #artificial-intelligence and #machine-learning channels have a lot of overlap with topics at PyCon.

Python software we use

There's lot's! This is just a quick list:

- Altair
- Camelot
- Dagster
- Klaxon Cloud: This software had its origins as an open source Rails app maintained by The Marshall Project before being adopted and ported to Python by MuckRock.
- pdfplumber
- Marimo
- Pandas
- VisiData

Get in touch

Tom Meagher
Signal: tfmfm.13
Bluesky: @tfmfm.bsky.social
Email: tmeagher@themarshallproject.org

Geoff Hing
Signal: ghing.16
Bluesky: @geoffhing.bsky.social
Email: ghing@themarshallproject.org