# Cody C. Baker, Ph.D.

codycbakerphd.com
cody.c.baker.phd@gmail.com
cody-c-baker
github.com/CodyCBakerPhD



## SUMMARY

Experienced data scientist and software engineer with 12 years spent designing, conducting, and sharing results of complex research. Worked with faculty, staff, and students from dozens of top universities to achieve excellence in the preparation and performance of research-related tasks, including data management, hypothesis formulation, and both mathematical and statistical modeling.

Possesses a diverse set of software skills, including DevOps, full-stack app development, performance benchmarking, ETL, and cloud deployments. Enjoys hosting workshops and hackathons to train users in state-of-the-art technologies and best practices. Familiar with a large number of data formats, with a strong concentration on updating local file standards to more cloud-performant ones. Knowledgeable about various mathematical and statistical applications, including high-dimensional time series analysis, nonlinear dynamics, and machine learning.

## EDUCATION

2015 – 2020 **Ph.D.** | Applied and Computational Mathematics and Statistics | *University of Notre Dame*
Doctoral thesis: "Second-order moments of activity in large neural network models"

Relevant coursework: machine learning, Bayesian statistics, network science, stochastic analysis, time series analysis, partial differential equations, nonlinear dynamics

2011 – 2015 **B.S.** | Applied Mathematics and Cognitive Science | *University of Evansville*
Minors in Computer Science, Philosophy, Psychology, and Neuroscience
Senior thesis: "Crime prediction models applied to the city of Evansville"
Honors thesis: "Goodness of fit metrics for joint probability distributions"

Relevant coursework: calculus, linear algebra, probability, statistics, discrete & combinatorics, numerical analysis, real analysis, mathematical physics, algorithms & data structures, cryptography, machine learning, symbolic logic, cognitive psychology, neurophysiology, advanced neuroscience, philosophy of mind, philosophy of science

## EXPERIENCE

2025 – Present **Research Software Engineer** | *Center for Open Neuroscience* | *Dartmouth College* | Hybrid

- Maintaining and enhancing several NIH data archives that store neuroscientific data by standardizing dataset organization across multiple modalities and providing usage analytics, including geolocation, through parsing terabytes of raw S3 log files.
- Assisting with user support tickets and feature requests to ensure smooth operation of data submission and retrieval processes.
- Developing open source software for use in neurophysiology research with an emphasis on achieving reproducibility of scientific findings.
- Researching options for stable, cost-effective, long-term backup of petabyte-scale data storage. Working with a range of institutions and services to provide solutions and cost estimates to that effect.
- Providing software, consultation, and computing support for the Brain Behavior Quantification and Synchronization consortium (BBQS).

## EXPERIENCE (continued)

2020 – 2024 **Research Software Engineer** | *CatalystNeuro* | Remote

- Developed and maintained 5 repositories of open-source software by ensuring proper functionality of automated testing suites, documentation, tutorials, and demos.
- Created and maintained 12+ data processing pipelines for neuroscience labs, allowing their data to flow from acquisition to sharing in a seamless fashion.
- Curated a total of 256 TB of high-value datasets to NIH data archives on behalf of various research groups.
- Managed the company's cloud resources (Amazon Web Services; AWS), including storage, compute resources, and identity access management (IAM).
- Handled user interactions within the research community by offering technical support and resolving issues or feature requests in a timely manner.
- Facilitated user education across various platforms by running sessions at multiple conferences and workshops, increasing user adoption and effective system utilization.
- All software related in some way to the facilitation of terabyte-scale data management, analysis, and visualization for the field of neurophysiology.

2016 – 2020 **Research Assistant** for the Neural Computing Group at *University of Notre Dame*

- Reconciled the computational properties of biologically realistic neural networks with artificial machine learning models (such as those used across computer vision) through complex mathematical theory and stochastic biophysical simulations, with results communicated through 3 journal publications and a presentation at the high-impact COSYNE conference (top 4% of abstracts accepted).
- Collaborated with several top experimentalists in neuroscience as a trainee in the NeuroNex program, which focused on understanding how neural function emerges from underlying structure.

2015 – 2017 **Teaching Assistant** for Applied Mathematical Methods at *University of Notre Dame*

- Ran 4 tutorial sections for 112 students, aiding their study of course concepts to achieve a 96% satisfaction rate.
- Graded homework assignments and exams. Filled in main lectures for the professor as needed.

2015 – 2016 **Data Analyst** for *JRM Environmental Inc*

- Analyzed geological data from water samples in conjunction with the Indiana Department of Environmental Management to issue compliance permits for Indianapolis regulation standards.

Summer 2015 **Technical Editor** for *Penguin Random House*

- Reviewed and corrected over 300 pages of "An Idiot's Guide to Algebra II".

Summer 2014 **Research Intern** for the BITLab at *Michigan State University*

- Explored statistical effects of algorithmic curation (the use of automated filtering mechanisms in the delivery and display of information) by measuring properties of simulated models of social networks, with results presented at the MIDSURE conference.

Summer 2013 **Research Intern** | Bioinformatics | *University of Kansas*

- Examined information diffusion through large-scale simulations of G-protein signaling mechanisms using a high-performance super-computing cluster (HPC), then presented results at an undergraduate symposium.

## KEY ACHIEVEMENTS

| | |
|---|---|
| Data | Published over **256** TB of neuroscience data to NIH archives, allowing anyone in the world to test scientific hypotheses. |
| Education | Trained **200+** professionals through **7** workshops and hackathons, disseminating the practices of open data. |
| Software | Developed software used by more than **50** research groups globally, integrating with daily data pipelines and practices. |
| Standardization | Standardized metadata of **40** data formats used by scientific devices, creating a universal language for comparing results. |
| Publications | Published **7** papers and **8** conference posters, spreading the word about open-source software and public data. |
| Pipelines | Updated **10** data analysis pipelines to leverage cloud streaming capabilities to improve speed and robustness. |

## SKILLS

| | |
|---|---|
| Coding languages | Python, MATLAB, Typescript, Javascript, HTML, LaTeX, R, C, C++, Java, CSS |
| Math, stats, & ML | Multivariate regression, classification, cross-validation, clustering, mixture models, dimensionality reduction, ensemble methods, deep learning, reinforcement learning, time series analysis, anomaly detection, computer vision |
| DevOps | Git, GitHub Actions, Docker, unit & integration testing, user & quality testing, Storybook, Chromatic, Puppeteer, Codecov |
| Libraries | NumPy, SciPy, pandas, polars, sklearn, OpenCV, FFmpeg, timeit, airspeed velocity, Matplotlib, Plotly, Pydantic, Boto3, concurrents, joblib, tqdm, requests, pytest, Flask, PyInstaller |
| Frameworks & APIs | AWS (IAM, S3, EC2, Batch, ECR, Route 53), Terraform, IPInfo, OpenCage, fsspec, HDF5 ROS3, remfile, h5wasm, obstore, Electron, Tk, Qt, React, pre-commit, Ruff, Black, flake8, pydocstyle, Mypy, Prettier, multiprocessing, threading, high-performance computing (HPC), Setuptools, Hatch, Twine, conda-forge, Vite, Node.js, Yarn |
| Databases | REDIS, DynamoDB, MongoDB, PostgresSQL, SQLite, NoSQL |
| Data formats | HDF5, Zarr, JSON, YAML, XML, Markdown, RST, Parquet, TSV, XLSX, MP4, AVI, TIFF, PNG, JPEG, MP3, WAV, PKL, MAT |
| Administration | GitHub, AWS, PyPI, Condaforge, DockerHub, ReadTheDocs |
| Business & team skills | Agile, scrum, research, mentoring, training, client consultations, drafting scopes of work, budgeting, grant writing |

## HIGHLIGHTED PROJECTS

Below are the most recent highlights of my major work. You can find the full listing at codycbakerphd.com/projects

2026 – Present

**HEARTH** | *Lead Developer*

EMBER · BBQS · Video · Web Apps · Testing · TypeScript

- Built the Hub for Exploratory Analysis, Resource-sharing, Testing, and Hypothesis-generation (HEARTH), a suite of web apps for the Brain Behavior Quantification and Synchronization (BBQS) consortium and its data on the EMBER Archive.
- Includes an uploader for adding files to a Dandiset on the EMBER Archive through a simple drag-and-drop interface, a clip extractor for saving a snippet or single frame of a video, and an encoding helper for the education and comparison of video encodings.
- The latter two grew from prototypes created in collaboration with the Talmo Lab.
- Generalized the shell the apps share into a template for new single-page, backend-free apps and a set of common component packages, each covered by unit, integration, and visual snapshot tests.

## HIGHLIGHTED PROJECTS (continued)

2024 – Present

**S3 Log Extraction** | *Lead Developer*

DANDI · Data Visualization · Analytics · AWS · Python

- Created a Python package for the fast extraction of access summary data from Amazon S3 server logs, originally developed for the DANDI Archive and steadily being rolled out to others.
- Anonymizes each requester, maps it to a geographic region, and aggregates the requests into per-dataset summaries and archive-wide totals that are shared through a public data repository.
- Extended it for the DANDI Archive's own logs and built the DANDI usage page (usage.dandiarchive.org), which visualizes data usage across the archive.

2023 – 2024

**NWB GUIDE** | *Lead Developer*

NWB · Data Conversion · Validation · User Interface · User Experience · Testing · JavaScript

- Developed an intuitive user interface for file management using interactive validation and real-time suggestions, streamlining the process for data submission to NIH archives.
- Ensured a robust testing suite involving multiple levels of integration and user interactions emulated using Puppeteer to enhance reliability and long-term maintainability.
- Tracked and documented dozens of hands-on user tests to refine the user experience.

2020 – 2024

**NeuroConv** | *Lead Developer*

NWB · Data Conversion · API Design · Cloud Computing · AWS · Python

- Established an automated data conversion tool capable of reading more than 40 distinct data formats used by neurophysiology acquisition devices in order to automatically write to the NeurodataWithoutBorders (NWB) standard.
- Designed universal APIs which transparently handled each layer of complexity to simplify the tasks of tagging, grouping, metadata transcription, temporal alignment, asset linking, buffering, chunking, and compression.
- Implemented a distributed cloud deployment system to run large-scale, off-site, batched conversions through Amazon Web Services (AWS).

2024 – Present

**NWB Benchmarks** | *Core Developer*

NWB · HDF5 · Zarr · Benchmarking · Python

- Managed a collection of unit and integration tests whose performance metrics (time, memory, and web packets) are sampled against a range of environmental run configurations, machine specifications, and bandwidth limits with the results used to assess the performance of HDF5 and Zarr file packaging parameterization against cloud streaming methods.
- Run outputs are aggregated into a common public database and built-in visualization tools generate reports of summary findings.
- A manuscript discussing interpretations and recommendations is in preparation.
- Built as an extension on top of the popular Airspeed Velocity (ASV) benchmarking framework.

2025 – Present

**DANDI Compute** | *Lead Developer*

DANDI · Provenance · Spike Sorting · HPC · Data Pipelines · Python

- Built an orchestration system that runs processing pipelines over assets already published on the DANDI Archive and uploads the results back to the archive, such as spike sorting through the AIND ephys pipeline and LFP extraction from raw electrophysiology.
- Packages each unit of work as a self-describing job capsule carrying its own submission script, parameters, and provenance, runs it as part of a SLURM array job on the MIT Engaging cluster, and tracks every capsule through its lifecycle using nothing but the archive's own metadata.
- Automated job submission through a self-hosted runner on the cluster, and published a results page for reviewing, comparing, and curating the pipeline runs.

## HIGHLIGHTED PROJECTS (continued)

2026 – Present

**DANDI Cache** | *Lead Developer*

NWB · DANDI · Provenance · GitHub Actions · Automation · Data Pipelines · Python

- Created a living collection of up-to-date, pre-computed mappings commonly used by developers working with the DANDI Archive, such as which assets are NWB files, which of those are valid according to the NWB Inspector, and measures of their internal structure.
- Each cache is the same pipeline around a different operation: read one or more upstream caches, work out what is not yet recorded, process each new item, and publish JSON Lines with full DataLad provenance.
- Every cache shares a central library, container images, GitHub Actions, and a template, gathered into one DataLad superdataset with a dashboard of the current state of each.

2022 – 2024

**NWB Inspector** | *Lead Developer*

NWB · DANDI · Metadata · Validation · Command Line · Python

- Built a command line tool used by the NIH data archive to validate all data uploads, which ensures all submissions are provided automated suggestions for metadata improvements that enhance data findability and reuse.
- Mirrored the design, style, and functionality of linting tools such as flake8, pydocstyle, and ruff.

2026 – Present

**Historia** | *Lead Developer*

Command Line · GitHub Actions · Automation · Python

- Created a command line tool for backing up records of GitHub activity (the pull requests and issues a person opened, was assigned, or reviewed) for use in creating dashboard reports of work history.
- Paired with a GitHub Action that runs the whole scheduled update of a work history data repository: fetching recent activity, committing it, populating a GitHub Projects board, and publishing a compressed archive of the data.

2023 – Present

**Neurosift** | *UX Consultant and Contributor*

NWB · HDF5 · DANDI · Data Visualization · Web Apps · User Experience · TypeScript

- *Contribution:* Added on-the-fly analyses computed in the browser: spectrograms of LFP streams with adjustable hyperparameters, and event-related signals that align any time series to a table of intervals.
- Aided in the design of a browser-based application for visualizing the contents of terabyte-scale neurophysiology datasets.
- Supports time series, images, videos, pose, tabular, annotated events, and all associated metadata through an intuitive exploration interface.
- Utilizes a specialized HDF5 web assembly implementation for increased efficiency when streaming data from the S3-backed archives.

2025 – Present

**nwb2bids** | *Lead Developer*

NWB · BIDS · Data Standards · Metadata · FAIR Data · Python

- Created a metadata extraction tool for organizing datasets of NeurodataWithoutBorders (NWB) files following the Brain Imaging Data Structure (BIDS) structure.
- Encourages best practices in data curation, such as self-describing entities, FAIR principles, and rich Hierarchical Event Descriptor (HED) annotations.
- Developed in conjunction with the BIDS Enhancement Proposal (BEP) 032 for micro-electrode neurophysiology data.

## HIGHLIGHTED PROJECTS (continued)

2021 – 2025 **Hierarchical Data Modeling Framework (HDMF)** | *Contributor*

HDF5 · Zarr · Data Standards · Parallel Computing · Python

- *Contribution:* Created the GenericDataChunkIterator for more efficient lazy, buffered writing of terabyte-scale datasets (existing defaults overlapped chunks and had poor default chunking heuristics), later extending it with serialization, parallel orchestration, and progress bar control.
- Enhanced a generic schema language for defining structured data standards intended to be written to group-like backend file types (such as HDF5 and Zarr) by generalizing, parallelizing, and optimizing lazy data buffering for terabyte-scale datasets.

2021 – 2023 **NWB Widgets (Legacy)** | *Developer*

NWB · Data Visualization · Python

- Drafted Jupyter widgets-based tools for visualizing the contents of terabyte-scale neurophysiology datasets.
- Supports many custom plugins for specific neural data types.
- Has some remote streaming capability, but is more intended for reading from local disks.
- Largely replaced by the Neurosift browser-based application.

2025 – Present **duct** | *Contributor*

Command Line · Continuous Integration · Resource Monitoring · Python

- *Contribution:* Added macOS support and CI runners, a basic structure for using duct as a library, demonstrated self-contained executable binary builds, and clearer errors on unsupported systems.
- Maintained a simple command line tool for monitoring and maintaining records of process-specific system resource utilization of any other command line invocation.
- Tracks wall clock time, CPU time, RAM consumption, and more!

## DATA CURATION

Below are among the more notable of the dozens of datasets I have helped manage, curate, and publish. You can find the full listing at codycbakerphd.com/data-curation

**Visual Coding (2-Photon Calcium Imaging)** | *Primary Engineer*

Neural Dynamics and Computing, Allen Institute for Brain Science · Microscopy

Dataset · Jupyter Notebook · Conversion Code

- Wrote custom code to standardize the mixed JSON/HDF5 metadata records for thousands of microscopy experiments and their associated visual stimuli.
- Published a total of **93.4 TB** of visual cortex microscopy data on the associated NIH data archive.
- Adapted a Jupyter Notebook tutorial to demonstrate how to remotely stream and analyze the dataset.
- Significant portions of this tutorial later became the flagship scientific example of the DANDI archive indexed by the Open Data Program (AWS) landing page.

**Brain Wide Map** | *Primary Engineer*

International Brain Lab · Electrophysiology

Manuscript · Dataset · Jupyter Notebook · Conversion Code

- Wrote conversion code which interfaced with PostgreSQL metadata database to standardize the data structure of thousands of extracellular electrophysiology experiments with complex behavior.
- Uploaded a total of **66 TB** of synchronized electrophysiology, audio, and behavioral event data on the associated NIH data archive.

## DATA CURATION (continued)

**Neural signal propagation atlas of Caenorhabditis elegans** | *Primary Engineer*
Leifer Lab, Princeton University · Microscopy
Manuscript · Dataset · Jupyter Notebook · Conversion Code

- Wrote custom code for handling the output of a unique microscopy rig, including an electrically tunable lens capable of acquiring images at variable (but precisely measured) depths.
- Published a total of **10 TB** of C. elegans microscopy data on the associated NIH data archive.
- Showcased reproduction of key figures from the associated manuscript through an interactive Jupyter Notebook.

**Glia Accumulate Evidence that Actions Are Futile and Suppress Unsuccessful Behavior** | *Primary Engineer*
Ahrens Lab, Janelia Research Campus (HHMI) · Microscopy
Manuscript · Dataset · Conversion Code

- Wrote custom code for handling the output of a unique microscopy rig coupled with virtual reality environment and some limited electrophysiology responses from peripheral motor nerves.
- Published a total of **23 TB** of volumetric whole-brain zebrafish microscopy data on the associated NIH data archive.

**MICrONS** | *Supervisor*
Tolias Lab, Baylor College of Medicine · Microscopy
Manuscript · Dataset · Jupyter Notebook · Conversion Code

- Supervised the conversion and publication of a high-value dataset which combined large-scale two-photon calcium imaging with associated electron microscopy reconstructions of the same tissue.
- Resulted in the publication of a total **1.2 TB** of functional imaging data on the NIH data archive, with region IDs coregistered with the associated electron microscopy datasets.

**Miscellaneous (Buzsaki Lab)** | *Primary Engineer*
Buzsaki Lab, New York University · Electrophysiology
Conversion Code

- Wrote custom conversion code for a dozen previous datasets.
- Unique aspects of these data collections include temperature-modulating probes, long-term passive recordings of spontaneous behavior, and paired optogenetic stimulation with extracellular electrophysiology.
- Published a total of **39 TB** of mouse electrophysiology and behavioral data on the associated NIH data archive, including the first TB-scale dataset on DANDI.

## PUBLICATIONS

### Journal Articles

2025 C. J. Gillon, **C. Baker**, R. Ly, E. Balzani, B. W. Brunton, M. Schottdorf, S. Ghosh, and N. Dehghani. "Open data in neurophysiology: Advancements, solutions & challenges." *eNeuro, vol. 12, no. 11.* doi:10.1523/ENEURO.0486-24.2025

K. Gunalan, R. Choudhury, et al. "DANDI: Shaping the future of neurophysiology and neuroimage data publishing and exploration through collaboration with the community, standards, and analytic platforms." *In preparation*.

R. Ly, **C. Baker**, S. Prince, and O. Rübel. "A practical evaluation of remote access methods for neurophysiology data in the cloud." *In preparation*.

H. Mayorquin, **C. Baker**, P. Adkisson-Floro, S. Weigl, A. Trappani, L. Tauffer, O. Rübel, and B. Dichter. "NeuroConv: Streamlining neurophysiology data conversion to the NWB standard." *Python in Science Conference, 2025.* doi:10.25080/cehj4257

2024 J. Magland, J. Soules, **C. Baker**, and B. Dichter. "Neurosift: DANDI exploration and NWB visualization in the browser." *Journal of Open Source Software, vol. 9, no. 97, p. 6590, 2024.*

## PUBLICATIONS (continued)

2023 M. Hawrylycz, M. E. Martone, et al. "A guide to the BRAIN Initiative Cell Census Network data ecosystem." *PLoS biology, vol. 21, no. 6, e3002133, 2023.*

2020 **C. Baker**, E. Froudarakis, D. Yatsenko, A. S. Tolias, and R. Rosenbaum. "Inference of synaptic connectivity and external variability in neural microcircuits." *Journal of Computational Neuroscience, vol. 48, pp. 123–147, 2020.*

**C. Baker**, V. Zhu, and R. Rosenbaum. "Nonlinear stimulus representations in neural circuits with approximate excitatory-inhibitory balance." *PLoS computational biology, vol. 16, no. 9, e1008192, 2020.*

2019 **C. Baker**, C. Ebsch, I. Lampl, and R. Rosenbaum. "Correlated states in balanced neuronal networks." *Physical Review E, vol. 99, no. 5, p. 052 414, 2019.*

### Conference Presentations

2024 **C. Baker**, G. Flynn, R. Ly, O. Rübel, and B. Dichter. "NWB GUIDE: Simplifying the conversion of neurophysiology data to NWB format." *Society for Neuroscience (SfN) 2024.*

2023 **C. Baker**, S. Weigl, H. Mayorquin, and B. Dichter. "NeuroConv: Automated conversion of neurophysiology data to NWB format." *Society for Neuroscience (SfN) 2023.*

2020 **C. Baker**, V. Zhu, and R. Rosenbaum. "Nonlinear computations in semi-balanced networks." *Computational and Systems Neuroscience (COSYNE) 2020.* Contributed talk (top 4% of abstracts).

2018 **C. Baker** and R. Rosenbaum. "Inferring connectivity and latent input covariance from spike train correlations." *Computational and Systems Neuroscience (COSYNE) 2018.* Poster (top 43% of abstracts).

### Workshops

2024 **C. Baker**, R. Ly, and B. Dichter. "Neurodata Without Borders and DANDI Archive Tutorial." *NWB and DANDI Workshop, Baylor College of Medicine.*

2023 **C. Baker**. "Neuroconv and NWB GUIDE tutorial." *NWB User Days, Janelia Research Campus.*

**C. Baker** and R. Ly. "Neurodata Without Borders and DANDI Archive Tutorial." *Pre-conference, COSYNE.*

**C. Baker**, O. Rübel, and S. Ghosh. "Neurodata Without Borders and DANDI Archive Hackathon." *NeuroDataShare, Sainsbury Wellcome Centre.*

**C. Baker** and G. Viejo. "Artificial intelligence, machine learning, computing, and visualization in neuroinformatics." *Open Data in Neuroscience (ODIN), Massachusetts Institute of Technology.*

2022 **C. Baker**. "NeuroConv tutorial." *NWB User Days, Janelia Research Campus.*