# Zachary Jacokes, Ph.D.

(615) 604-7498 | zjacokes@gmail.com | GitHub | Google Scholar | Website

## SUMMARY

Senior data scientist with 10+ years of hands-on experience building end-to-end analytical systems for large-scale, multi-source datasets. PhD in Data Science (UVA, 2026). Expertise spans the full stack: statistical modeling, machine learning, reproducible pipeline engineering, and data infrastructure in complex, regulated environments. Strong publication record and track record of turning ambiguous analytical problems into measurable, production-ready solutions.

## CORE COMPETENCIES

**Machine Learning & Modeling:** Supervised/unsupervised learning, classification, regression, clustering, dimensionality reduction, representation learning, time-series modeling, deep learning (PyTorch/TensorFlow), experimental design, A/B testing, causal inference

**Statistical Analysis:** Hypothesis testing, mixed effects models, power analysis, cross-validation, model evaluation and validation, harmonization of heterogeneous datasets

**Data Engineering & Infrastructure:** ETL/ELT pipeline design, relational database schema (SQL), data validation frameworks, HPC (Slurm), containerization (Docker/Singularity), AWS S3, Git/GitHub, reproducible scientific workflows

**Programming:** Python (advanced), R (advanced), SQL, Bash

## EDUCATION

**University of Virginia**: Ph.D. in Data Science (2026)

*Member of Inaugural School of Data Science Ph.D. Cohort*

**Emory University**: B.A. in Psychology (2013)

## EXPERIENCE

**Doctoral Researcher** | University of Virginia School of Data Science | 2021 – 2026

- Designed and validated machine learning experimentation pipelines for high-dimensional, multi-source datasets (500+ subjects), enabling reproducible cross-cohort evaluation of model generalization.
- Developed representation learning and spectral embedding frameworks to identify stable latent structure across heterogeneous, heteroskedastic data — applied to patient stratification and biomarker discovery (publication under review).
- Implemented nested cross-validation, power analysis, and dataset harmonization (ComBat) to ensure robust, generalizable findings across heterogeneous data collection environments — directly analogous to multi-site observational study design.
- Built scalable, containerized HPC pipelines (Slurm, Docker/Singularity) for large-scale data processing and model experimentation, with version-controlled, modular, documented code; reduced iteration cycles from days to hours.
- Developed interpretable multimodal models integrating behavioral and imaging-derived features, translating high-dimensional signal into actionable, audience-appropriate findings.

**Senior Data Specialist** | University of Virginia | 2019 – 2021

***Data Platform & Infrastructure***

- Architected and deployed a multi-site clinical data platform (REDCap) supporting 500+ participants across 5+ sites and 30+ standardized instruments, serving as the longitudinal backbone for a federally funded study.
- Designed relational database schema from scratch; built automated ETL pipelines computing derived metrics, scoring transformations, and endpoint aggregations across instruments, enabling downstream statistical and ML analysis.
- Implemented data validation and constraint logic (range checks, type enforcement, referential integrity) and led audit-response efforts to meet HIPAA and federal data standards.
- Managed role-based access controls for analysts, data engineers, and end users across distributed sites; authored documentation and delivered staff training.

***Analytical Pipeline Engineering***

- Designed and implemented an end-to-end automated pipeline for ingestion, de-identification, preprocessing, and structured output of multi-modal data across heterogeneous acquisition environments.

- Built parallelized HPC workflows enabling petabyte-scale dataset processing with reproducible, version-controlled outputs structured for downstream statistical and ML analysis.
- Applied harmonization methods to correct for site and instrument effects across multi-source data, enabling valid cross-cohort comparisons.

***Leadership & Collaboration***

- Coordinated cross-functional project teams spanning data engineering, quality control, and analytical workflows across multi-site research initiatives.
- Mentored two junior researchers in ML techniques, pipeline development, and research design; both contributed to published and conference work.

**Programmer / Analyst** | University of Southern California — Lab of Neuroimaging | 2015 – 2019

- Coordinated multi-site data collection, harmonization, and distribution pipelines for a multi-institution longitudinal consortium requiring standardized data governance across sites.
- Designed reproducible statistical analysis and quality control workflows, adopted as lab standards; developed data QC protocol using multivariate statistical methods for cross-site consistency.
- Published multiple first- and co-authored papers; presented work at international conferences (OHBM 2017, 2018, 2019).

**Research Assistant** | Yerkes National Primate Research Center | 2014 – 2015

- Developed Python-based data manipulation and analysis tools; contributed to experimental design and behavioral research methodology.

## SELECTED PUBLICATIONS

*First-authored and representative co-authored work:*

1. Jacokes Z, Beeler-Duden S, Lawson S, et al. [Sensory-behavior data integration]. MedRxiv (preprint). — Multimodal feature integration; topography-aware modeling.
2. Jacokes Z, Adoremos I, Hussain AR, et al. Unsupervised Dimensionality Reduction Techniques for the Assessment of Biomarkers. Biocomputing 2025. World Scientific; 2024:614–630. — Representation learning for high-dimensional feature discovery.
3. Jacokes Z, Jack A, Sullivan CAW, et al. Linear discriminant analysis of phenotypic data for classification by diagnosis and sex. Front Neurosci. 2022;16:1040085. — ML classification; generalization across heterogeneous populations.
4. Ressa HJ, Newman BT, Jacokes Z, et al. Widespread associations between behavioral metrics and structure suggest age mediates subtypes. Imaging Neuroscience. 2025;3. — Age-stratified subgroup analysis; phenotype-feature association.
5. Gupta R, Audhkhasi K, Jacokes Z, et al. Modeling Multiple Time Series Annotations as Noisy Distortions of the Ground Truth. IEEE Trans Affective Comput. 2018;9(1):76–89. — EM framework for noisy ground truth inference in longitudinal data.

*Full list: 14 peer-reviewed publications, 1 book chapter, 15+ conference abstracts (OHBM 2016–2023)*

## SERVICE & LEADERSHIP

- University of Virginia Raven Society — First inductee from the School of Data Science; member of Selection Committee.
- UVA Brain Institute — Consulted on research funding allocation and contributed to strategic direction discussions.
- Neurodata Interest Group — Founding member; led biweekly sessions on methods and publication discussions.