joshua.f.allen@gmail.com
GitHub
LinkedIn

Joshua Allen, PhD

Summary Data Scientist experienced in designing and building scalable systems for experimentation and analytics from the ground up. Skilled in applying causal modeling and Bayesian inference to inform business decisions, drive value, and communicate uncertainty to stakeholders. Adept at explaining and simplifying complex concepts for colleagues, stakeholders, and executives.

Skills

R: Tidyverse(ggplot2, dplyr, etc), tidymodels, brms, Shiny, Targets, sf

Python: PyMC, scikit-learn, Polars, PySpark ,NumPy, Pandas, Matplotlib, Seaborn, Selenium, Plotnine, FastAPI

SQL, Docker, Quarto, React, D3

Quasi-Experimental Design | Data Cleaning | Data visualization | Machine Learning

Reproducible Pipelines | Exploratory Analysis | Written & Verbal Communication

Experience

Georgia State University **Aug. 2019 - July 2025**
Research Scientist|PhD Candidate:

- Designed and implemented end-to-end data pipelines covering all phases of the data lifecycle using R and Quarto. Pipeline **improved data deliverable time from 3 days to less than 2 hours**.
- Designed causal analysis using Double Machine Learning and Sensitivity tests to uncover novel insights into long-term political behavior.
- Implemented reproducible and scalable Python workflows to collect, ingest, and clean 3 terabytes of data locally for natural language processing.
- Developed 18+ production-grade web scrapers in R/Python for large-scale data collection.
- Provided **internal R training** curriculum development, training sessions, and presentations **using Git version control, Quarto, and deployed using CI/CD workflow for 200 researchers**.
- Led development of the **ecdata R and Python packages** to deliver the largest open source political executive communication text dataset.
- Designed undergraduate courses in causal inference for social science research covering A/B tests, Difference-in-Differences, and other canonical quasi-experimental methods.
- Made extensive use of data visualization tools in R and Python to distill and communicate complex topics to technical and non-technical stakeholders.
- Mentored students and advised colleagues on statistical analysis, scientific programming, and best practices for reproducible research.

Projects

Football Power Rankings

Toolsets used: PyMC, FastApi, React, D3, and Polars

- Built a Bayesian Bradley-Terry model to rank NFL teams by latent ability and incorporate uncertainty of the estimation.
- Deployed a FastApi pipeline to update rankings in real time.
- Designed a React + D3 dashboard for intuitive visualization of results.
- Translated ranking uncertainty into insights for consumer preference modeling and competitive benchmarking. Live Demo

Surpristing YAC Players

Toolsets used: PyMC and Polars

- Modeled player performance based on beyond expected outcomes using Bayesian hierarchical models.
- Adjusted for contextual factors like: score differential, yards to go, and time remaining using splines and Gaussian processes.
- Highlighted undervalued players to pick up in Fantasy Football, increasing average points by 3-5 points per game.

Memorial to The Deported Jews of France

Toolsets used: R Shiny, sf, Selenium

- Geocoded over 70,000 addresses using OSM and other free APIs.
- Built production grade scrapers to accurately scrape information.
- Used R Shiny and Mapbox GL to show the distribution of deportations across France.
- For a demo see this repository

Education

Doctor of Philosophy, Political Science: Georgia State University (2019 - 2025)
Masters of Arts, Political Science: Georgia State University (2017 - 2019)
Bachelors of Arts, Political Science: Sonoma State University (2013 - 2017)