John Hoffman

Contact Information		📞 +1 (847) 757 5840
		@ johnhoffmanchicago@gmail.com
		🔲 github.com/johnh2o2
		🔲 linkedin.com/in/johnh2o2

Profile

Research-oriented data scientist with deep expertise in Bayesian and probabilistic modeling, leading high-impact, ambiguous research projects from conception to production deployment. PhD in Astrophysics with extensive publications and proven ability to integrate state-of-the-art AI techniques for novel applications in large-scale data analysis.

Education

Ph.D., Astrophysics, **Princeton University** — 2018
Graduate Certificate in Computational and Information Science
Thesis: *Bayesian techniques for finding and characterizing variable stars*
B.S., Physics; **B.S.**, Mathematics, **University of Illinois at Urbana-Champaign** — 2013

Skills

Programming: Python (10+ years, pandas, PyTorch, sklearn, Stan), CUDA, C, SQL, HPC (SLURM)
ML & Statistical Methods: Bayesian/probabilistic modeling, hierarchical models, time-series analysis, statistical inference, optimization, Gaussian processes, mixed-effects models
Tools/Platforms: AWS/cloud platforms, Git, Linux/Unix, distributed computing, data pipelines, LLM integration for data curation

Work Experience

Staff Data Scientist at FAIR (Meta) — 2022 to present
- Led end-to-end human evaluation research initiatives for flagship AI models: Seamless M4T (Nature), NLLB (Nature), MovieGen, SAM Audio (core contributor)
- Architected FAIR-wide compute analytics infrastructure (co-authored HPCA paper); spearheaded video-data strategy for multi-billion parameter models

Senior Data Scientist at Strong Analytics — 2021 to 2022
- Developed probabilistic models for revenue optimization using mixed-effects frameworks and gradient boosting, delivering measurable ROI

Data Scientist at Copilot AI (Xaxis, GroupM) — 2019 to 2021
- Built production ML systems for real-time advertising optimization, experimenting with reinforcement learning and high dimensional clustering

Selected Projects

cuvarbase: GPU-accelerated time-series analysis library (github) — 2017
- Architected GPU-based astronomical time-series library with custom CUDA kernels, achieving orders-of-magnitude performance improvements
- Enabled discovery of record-setting 6.91-minute eclipsing binary (Nature paper); adopted by NASA's Transiting Exoplanet Survey Satellite (TESS QLP) for a 40x speedup of planet searches

Fast Template Periodogram (github, arXiv paper) — 2016-2018
- Developed novel algorithmic approach reducing computational complexity from $\mathcal{O}(N^2)$ to $\mathcal{O}(N \log N)$, achieving 140x runtime reduction

Publications

Google Scholar: scholar.google.com/citations?user=kUl81GQAAAAJ
Featured: "No Language Left Behind" (Nature, 2024) — Led statistical analysis for breakthrough multilingual translation model; "Fast Template Periodogram" (ApJ, 2021) — Novel algorithmic contribution

Awards

- DOE Computational Science Graduate Fellowship finalist — 2014
- Graduated with highest distinction (UIUC) — 2013
- Chancellor's Scholarship, top 1% of students (UIUC) — 2009
- Presented at SciPy Conference — 2017