

ABOUT ME

Data scientist and data science lead with 11+ years across industry and research. Build and evaluate LLM and machine-learning systems end to end, from benchmark design, rubric methodology and LLM-as-judge to the engineering needed to run them in production. Background in Bayesian statistics, causal inference and computational cognitive neuroscience informs an approach to LLM evaluation as a measurement problem: building evaluations that are reliable, interpretable and statistically sound. Track record of uncovering business-critical issues, leading cross-organisational technical work and developing data scientists.

SKILLS

General - Data science, research, software engineering, data engineering, stakeholder management

Programming - Python, SQL

Statistics - Bayesian & frequentist statistics, causal modeling, DAGs

Machine Learning - Deep learning, (non-)linear regression and classification, unsupervised learning

LLMs - LLM-as-judge, evaluation and benchmark design, rubric development, structured output, multilingual evaluation

Software - NumPy, Pandas, Scikit-Learn, Pytorch, Airflow, Docker, Google Cloud Project, PyMC

EDUCATION

Ph.D. Candidate in Computational Cognitive Neuroscience (ABD)

University of Oldenburg
Completed all research requirements; 2 peer-reviewed first author publications; transitioned to industry. 2016 - 2021

M.Sc. in Neurocognitive Psychology

University of Oldenburg
2013 - 2016

EXPERIENCE

Fab AI

7/2025–present

Head of Data Science

8/2026–present

- Lead data scientist for the Quality Assurance Facility, Fab AI's programme evaluating frontier and small language models for education in low- and middle-income countries.
- Designed with education experts a benchmark suite testing whether LLMs can generate reading material matched to a learner's phonics level, in English and Kiswahili; tech lead for two of them, directing a five-person team across two organisations.

Senior Data Scientist

7/2025–8/2026

- Led design, implementation and deployment (GCP) of an LLM-as-judge system scoring lesson plans, storybooks and textbooks against pedagogical rubrics. The rubrics and workflow are now used by external organisations to assess education technology.
- Led rubric development with education specialists: binary criteria with fail conditions reached 87% human–LLM agreement against a 91.8% human–human ceiling, across GPT-5, Claude and Gemini.
- Line-managed two data scientists: mentored one to a first-author publication, and designed a GCP benchmarking pipeline, built by the other, that automated previously manual model onboarding.
- Introduced engineering rigour across the data science function: code review, project standards, Jira, and a move back to GCP.

eyeo

3/2021–3/2025

Staff Data Scientist

3/2024–3/2025

- Uncovered critical double-digit percentage user loss hidden by telemetry issues, directly impacting company strategy. Developed Bayesian hierarchical models to interpolate missing data, enabling accurate forecasting despite severe data quality issues.
- Became go-to data scientist for user retention and revenue analysis, presenting directly to C-level executives. Built causal models using DAGs to quantify impact of retention initiatives.
- Established company-wide data quality standards and analytical blueprints for causal inference. Aligned user metrics across three subsidiaries, creating first unified view across programmatic advertising, browser, and B2B revenue streams.

Senior Data Scientist

6/2022–3/2024

- Designed ML-powered monitoring system that increased problematic ad detection by 580% while reducing manual review by 90%, improving experience for 250M users.
- Built data pipelines using Airflow, Docker, and GCP. Mentored 4 team members in modern data practices. Led training sessions on Airflow, functional data engineering, and Bayesian modeling.

Data Scientist

3/2021–5/2022

- Designed and productionized ML model that increased problematic ad identification by 60%.
- Built Bayesian MRP model for estimating internet traffic.
- Conducted strategic research and analysis for key business decisions.

University of Oldenburg - Researcher

11/2016–2/2021

- Conducted research and published articles on the neural basis of speech processing using machine learning and human neural data with a focus on interpretability and sparse coding.
- Built research software (using Python and Docker) for machine learning in neuroimaging that is still in use.

LINKS

 [github](#)
 [linkedin](#)
 [website](#)

EXTRAS

Mentoring - I mentored junior data scientists at the University of Sussex.

Academic exchange grant (2019) - I won a grant for developing a deep learning model for automatic speech recognition that incorporates neuronal data.

Communication - I presented my work at five conferences.

- Supervised 5 Master's research projects (3-6 months each) on machine learning applications in neuroscience.

Columbia University, New York - [Visiting Researcher](#) 6/2019–1/2020

- Developed an interpretable Deep Learning model for linking audio data with brain activity based on self attention in recurrent neuronal networks.

Self-employed / U. Oldenburg - [Research Data Scientist](#) 11/2013–11/2015

- Statistical analyses in Python and R for Medical School Hanover and University of Göttingen; modelling of non-linearities in human invasive neural recordings at Oldenburg.

SELECTED PUBLICATIONS

Preprint 2026	Visual Reasoning Benchmark: Evaluating Multimodal LLMs on Classroom-Authentic Visual Problems from Primary Education.  , M. Huti, A. Mackintosh, A. Waldock, D. Andrews, M. Lelièvre, M. Boos , et al., <i>arXiv:2602.12196</i> .
Journal Article 2021	Generalizable dimensions of human cortical auditory processing of speech in natural soundscapes: A data-driven ultra high field fMRI approach.  , M. Boos , J. Lücke, and J. W. Rieger, <i>Neuroimage</i> .
Journal Article 2016	Probabilistic inference: Task dependency and individual differences of probability weighting revealed by hierarchical Bayesian modelling.  , M. Boos , C. Seer, F. Lange, and B. Kopp., <i>Frontiers in Psychology</i> .

SELECTED WORK

Evaluating educational materials with LLM-as-judge 	2026
• Rubric methodology and results behind the Fab AI evaluation pipeline.	
Deep learning for auditory encoding 	2022
• Python library for self-attention RNNs in neuroscience.	
Voxel-wise encoding 	2020
• Scalable encoding models for neuroimaging with Docker.	
Open Source contributions	
• Nilearn  (neuroimaging ML), Skrub  - machine learning with dirty categorical data	