

Johannes Roth

Machine Learning Engineer | Recommender Systems | Large-Scale ML Pipelines | Representation Learning

✉ johannes@roth24.de [in](#) LinkedIn [G](#)itHub [g](#)itHub jroth.space [📍](#) Leipzig, Germany

TECHNICAL SKILLS

ML	PyTorch, TensorFlow, scikit-learn, Transformers, CLIP, ViT, CNNs, GANs, Metric Learning, Uncertainty Estimation
Recommenders & Evaluation	Recommendation Systems, Ranking Metrics, Similarity Search, Embeddings, Bayesian Optimization, Offline Evaluation, A/B Testing
Engineering Languages	Python, SQL, Docker, AWS, Git, CI/CD, Flask, Redis, Linux, Bash, SLURM/HPC German (Native), English (Fluent)

EXPERIENCE

PhD Researcher – Large-Scale ML, Vision, and Human Data *May 2022 – Present*
Max Planck Institute for Human Cognitive and Brain Sciences & University of Gießen

- Built large-scale ML and data pipelines to curate **LAION-Natural**, a 120M-image subset of LAION-2B, including filtering, embedding-based analysis, quality control, and public dataset release workflows.
- Developed **Natural Controversial Stimuli**, an active learning framework that optimizes image selection to maximize disagreement between model-based similarity judgments.
- Used modern computer vision models and large feature stores for model comparison, embedding-space analysis, hard-example selection, and evaluation-set design.
- Built and managed the **AWS-based data infrastructure** for **re:vision**, including dataset hosting, access workflows, and the public challenge website.
- Simulated noisy sequential experimental designs to evaluate how timing, model assumptions, and measurement noise affect reliability and downstream statistical power.

Research Assistant – ML in Medicine *Jun 2021 – May 2022*
ScaDS.AI Dresden/Leipzig

- Engineered a multi-plane **UNet++ ensemble** for Glioblastoma segmentation (BraTS 2021), integrating Dice and Boundary loss to achieve competitive segmentation performance (Dice score: 0.90 for whole tumor).
- Developed attention-based mortality prediction models using **FT-Transformer** and **SAINT**, including epistemic uncertainty estimation for safer medical decision support (0.85 AUC-ROC).

Data Scientist / ML Engineer (Working Student) *Oct 2019 – May 2021*
CHECK24 (Travel Vertical)

- Designed, implemented and deployed an image-processing micro-service (**Flask + Redis**) enabling fast ML inference over **>20 million** hotel images.
- Built image-based systems for deduplication, retrieval, classification, and quality scoring, improving the handling and ranking of large-scale hotel image data.
- Used Bayesian hyperparameter optimization to improve ranking metrics and conversion-oriented performance of the hotel recommendation system; built Grafana dashboards to monitor API health, price stability, and data quality issues.

Full-stack Developer (Freelance) *Oct 2020 – May 2021*
Kimetric UG

- Implemented two academic websites using **Django**, Nginx, and Unicorn. Configured Linux hosting environments and automated deployment scripts (CI/CD).

Data Scientist (Working Student)
Webdata Solutions GmbH (now Vistex)

Oct 2018 – Oct 2019

- Revamped product-matching pipeline with a neural-network based approach trained on self-collected web-scraped datasets, increasing matching accuracy from **<50% to 92%**.
- Built data collection, training, evaluation, and interpretability workflows for product similarity and entity-resolution tasks.

PROJECTS & OPEN SOURCE

- **thingsvision** (Core Contributor) – Modular feature-extraction library for computer vision. Used to extract and compare features from **100+** modern vision models (CLIP, ViT, CNNs, etc.); **460k+ PyPI downloads**.
- **ReLAION-2B Natural** – Scored **2.1B images** for naturalness and released a large-scale naturalistic image subset with **167GB** of ViT-H/14 embeddings for visual similarity search.
- **re:vision Initiative** – Designed and implemented the public challenge website and AWS-backed dataset hosting infrastructure for a replication initiative around LAION-fMRI.

EDUCATION

PhD Candidate – Computational Cognitive Neuroscience **May 2022 – Present**
Max Planck Institute for Human Cognitive and Brain Sciences & University of Gießen

Dissertation on efficient experimental design, large-scale visual datasets, and model–brain alignment using modern computer vision models and high-resolution fMRI.

M. Sc. Computer Science **2017 – 2021**
Leipzig University

Grade **1.2** (Distinction). Focus: Data Analysis, Machine Learning, Medical Image Processing.

- **Master's Thesis (Grade 1.1)** – Used GANs to synthesize stimuli that maximally activate specific, targeted brain regions, recovering known category-selective areas in human brains.

B. Sc. Business Information Systems **2014 – 2017**
Leipzig University

Grade **1.5**. Focus: Distributed Systems, E-Commerce, Data Management, Economics.

SELECTED PUBLICATIONS & AWARDS

- **Award (2025):** CMBB Replication Award for contribution to reliable coding practices in neuroscience.
- **J. Roth, M. N. Hebart.** *How to sample the world for understanding the visual system.* CCN 2025 (Oral Presentation).
- **J. Roth et al.** *Ten principles for reliable, efficient, and adaptable coding.* Communications Psychology (2025, In Press).
- **J. Roth et al.** *Multi-plane UNet++ Ensemble for Glioblastoma Segmentation.* BraTS Challenge 2021.

LEADERSHIP & COMMUNITY

- **Technical leadership:** Led research-engineering workstreams spanning dataset curation, ML evaluation, experimental design, and reproducible analysis pipelines.
- **PhD Representative (2023–2024):** Elected to represent >180 doctoral researchers at MPI CBS.
- **Mentoring:** Supervised working students and interns in the research group.
- **Talks:** Presented research findings at international conferences (CCN), internal institute colloquia, and interdisciplinary research meetings.