David S. Roth

davidroth.me ⓘ
davidsamuelroth@gmail.com ✉

EDUCATION

NYU Center for Data Science, *M.S. Data Science*   *2021 - 2023*
Computational Linear Algebra, Causal Inference, Inference and Representation, Deep Learning, Natural Language Understanding, Probabilistic Time Series Analysis, Responsible Data Science

NYU Leonard Stern School of Business, *B.S. Business / Statistics*   *2014 - 2018*
Probability & Statistics, Linear Regression and Forecasting Models, Regression and Multivariate Data Analysis, Stochastic Processes, Theory of Probability

NYU Courant Institute of Mathematical Sciences, *B.A. Computer Science*   *2014 - 2018*
Algorithms & Data Structures, Discrete Math, Linear Algebra, Computer Systems Organization, Large Scale Web Applications, Machine Learning, Computer Vision, High Performance Machine Learning

EXPERIENCE

⊘ Edgi Learning   *February 2023 - September 2023*
NLP Engineer

Principal engineer on Edgi Librarian, a generative AI tutoring, educational content recommendation and learning assessment chatbot. Contributions include conversation state management, training data curation, prompt engineering and vector database content search.

⌂ Zillow   *September 2022 - December 2022*
Research Fellow

Built training and inference pipelines for multi-modal (image+text) graph neural networks using PyTorch and Deep Graph Library. Utilized message-passing and learned gating functions to compute dense node representations for property listings and predict attributes and node links over graphs containing 1M+ nodes.

◉ Spotify   *June 2022 - September 2022*
Data Science / Machine Learning Engineering Intern

Built inference pipeline for entity disambiguation pipeline used to match recordings to registered musical works with PyBeam, Luigi and Tensorflow. Delivered pipeline capable of scoring ~35M match pairs across 50 machines in 20 minutes wall time.

Assessed error rates on lyric content across languages and markets to identify opportunities for improvement in Spotify's lyric ingestion pipeline. Designed and oversaw annotation trial to evaluate heuristic and statistical approaches to quality control.

KPMG Data & Analytics   *January 2019 - August 2021*
Data Scientist, NLP

Built data processing pipelines for greenfield credit card offering at major NYC financial institution using Airflow, Amazon SQS and PySpark. Worked on small, agile team to rapidly build and scale fault-resistant systems to handle 20+ million daily transaction records. Worked closely with anomaly detection and fraud teams to develop and deploy models to detect fraudulent transactions.

Fine-tuned generative LSTM language model for extractive question answering over regulatory documentation. Delivered model with top-5 accuracy of 82%.

SKILLS

- Python (10 yrs), SQL (8 yrs), Scala (3 yr), Rust (1 yr)
- PyTorch, TensorFlow, ONNX, GPT, BERT, HuggingFace,
- Spark, Beam, Airflow,
- Google Cloud Platform, AWS

RESEARCH

- *Variations and Relaxations of Normalizing Flows (preprint)*, arXiv:2309.04433
- *Speech Transcription and Disfluency Removal with Acoustic-Language Model Pretraining (preprint)* arXiv:2309.04516

OTHER

Dollar-a-Day Fund   *June 2023 - Present*
Web Developer (Volunteer)

INTERESTS

Multi-instrumentalist and full-stack musician (writer/performer/engineer) ◉ ☁