JONAH SMITH

+1 (412) 225-7356 — jonahsmith211@comcast.net — Pittsburgh, PA, USA — linkedin.com/in/jonah-smith-10846b21b

EDUCATION

University of Pittsburgh Aug 2022 – May 2026
Bachelor's, Computer Science
Bachelor's, Data Science
Minor, Applied Statistics GPA: 3.915

PROFESSIONAL EXPERIENCE

Federal Specialized Services, Pittsburgh, PA May 2024 – July 2025
Software Developer and Data Analyst Intern

- Developed and deployed web-based software applications to streamline internal company processes, enhancing efficiency and productivity.
- Utilizing Python for analysis of large datasets, providing real-time insights via custom tools.
- Led software development from concept to implementation, delivering fully functional web solutions.
- Implemented algorithms to optimize data processing and ensure scalability for large data handling.
- Collaborated with lawyers to conditionally sift and extract data from legal documents, databases, and internal networks.

SKILLS

- **Programming:** Python, Java, JavaScript, SQL
- **Data & ML:** Pandas, NumPy, Scikit-learn, PyTorch, TensorFlow, Neural Networks
- **Analysis & Viz:** Statistics, Risk Management, Tableau, Data Analysis
- **Tools & Dev:** Git, HTML/CSS

SELECTED PROJECTS & PORTFOLIO

Concept Extraction Pipeline (Python, NLP, GPT API)
Capstone Project — University of Pittsburgh
Developing a large-scale concept extraction system for educational content. Built a pipeline to process and annotate 4,000+ lecture slides using natural language processing, LLM-based annotation scripts, and a structured codebook for semantic concept identification and evaluation. Used the annotation pipeline to tune and train the overarching model.

Machine Learning Regression & Classification – PPG Industries (R)
University of Pittsburgh, Data Science Coursework
Applied ML to proprietary RGB/HSL paint color data for both continuous and binary targets. Conducted EDA, feature engineering, and logit transforms; developed, tuned, and evaluated linear, regularized, Bayesian, neural network, random forest, and gradient boosting models; analyzed feature importance and uncertainty.

Synthetic Data Modeling (Python)
Personal Project
Engineered a synthetic data generator using block bootstrapping to replicate volatility clustering and heavy-tailed distributions. Tuned this overarching framework to fit various models long-term models, enabling scalable risk evaluation of time-series data sets.

LLM Compression & Bias Analysis (Python, Hugging Face, scikit-learn)
Research Replication Project
Replicated and extended Gonçalves & Strubell (2023) to analyze how quantization and knowledge distillation affect social bias in large language models. Evaluated BERT, DistilBERT, and RoBERTa using CrowS-Pairs and StereoSet datasets, applying regression-based bias metrics and statistical testing for fairness evaluation.