# Ben Wortman

### Lead Data Scientist

814.594.4981 | ben.wortman@gmail.com | ben-wortman.com | State College, PA

### Experience

### Associate Director of Data Science | FINRA

#### Associate Director of Data Science

- Currently scaling my data science team which is at the center of FINRA's enterprise AI strategy of driving regulatory efficiency through GenAI applications.
- Led the development of agentic orchestration APIs for high-volume, cross-unit tasks including fact extraction, summarization, and question answering. These are projected to deliver \$1.4M in internal savings and \$2M in savings for member firms.
- Spearheaded the assessment and integration of third-party chatbot solutions to streamline HR operations.
- Directing two R&D initiatives leveraging traditional ML for proactive risk identification and monitoring, supporting FINRA's mission of market integrity.

### Lead Data Scientist | FINRA

- Promoted via skip-level to a hybrid leadership role, managing a team of 4 data scientists while serving as senior individual contributor.
- Delivered an internal NLP product for automated topic modeling of public comments, cutting analysis time from weeks to minutes and generating \$250K in annual savings; enabled real-time executive visibility into emerging public sentiment.
- Designed and deployed document classification and key fact extraction pipelines, eliminating 10,000+ hours of manual review and accelerating regulatory response times.

### Data Scientist | FINRA

- Developed and deployed a document classification and triage pipeline, improving identification of risky filings by 130%.
- Built an internal application for detecting duplicate semantic content as part of FINRA's guidance modernization initiative.
- Created an interpretable NER benchmarking library to evaluate LLM and transformer model performance on internal FINRA datasets, supporting informed model selection for regulatory use cases.

### Data Science Consultant | Profound

• Consulted for an early-stage marketing technology startup, designing predictive models and data pipelines for the launch of their <u>conversation explorer tool.</u>

### Research Assistant | Penn State University

- Developed a novel taxonomy of emotion by clustering word embeddings for advanced emotion classification research.
- Led the technical operations of the group's motion capture lab.
- Delivered coursework on high-performance computing and deep learning frameworks (TensorFlow, PyTorch), including emerging ML topics and applications.

### Machine Learning Intern | Carnegie Mellon SEI Al Division

- Developed a prototype interface for probabilistic object detection and localization in autonomous vehicles.
- Authored a comprehensive review on satellite object detection methods for the National Geospatial Agency.

### Data Engineer Intern | Impact Radius

- Built enterprise data solutions for media spend attribution from walled garden platforms (Facebook, Instagram).
- Retained post-internship as a part-time data engineer to extend and scale solutions during the academic year.

### Oct 2024 - Dec 2024

2020 - 2022

### 2019 - 2020

Summer 2021

### 2022 - Present April 2025 - Present

### 2022 - 2023

2023 - 2025

### Education

Masters, Informatics | The Pennsylvania State University | 3.84

B.S. Data Science, Math Minor | The Pennsylvania State University | 3.75

### Skills

Languages - Python, SQL, C, Scala, Matlab/Octave, R, HTML/CSS

Technologies and Select Libraries - AWS, Google Cloud, Langchain, Langgraph, Pytorch, NLP, Generative AI, LLMs, DSPy, Dash, Streamlit, FastAPI, Scikit-learn, mTurk, Multiprocessing, Cluster Computing, Pyspark, Data Visualization, XAI, Shapely/LIME, Time Series Analysis, Pillow, OpenCV, Certificate in Machine Learning

Other Skills - Research, Leadership, Communication, Project Management, Public Speaking, Technical Writing

### Select Project Work

### **Generic Key Fact Extraction**

- Led development of an agentic orchestration pipeline for key fact extraction and question answering, integrating end-to-end OCR, multiple context retrieval methods, attribution, error handling, and logging.
- Applied across multiple use cases, delivering \$1.4M in internal savings and \$2M in projected savings for member firms.

#### **Risk Review for Audit Submissions**

- Built an automated question answering pipeline addressing 26 audit submission review questions using ML classifiers, Retrieval-Augmented Generation (RAG), and DSPy prompt optimization.
- Achieved an estimated savings of 10,000+ analyst hours annually.

#### **Automated Comment Review for Rule Proposals**

- Developed a topic modeling and aggregation solution for public comments, reducing review times from weeks to minutes.
- Leveraged LLMs to assess topic alignment, identify commenter profiles, detect toxic language, and generate narrative summaries via a custom dashboard, enabling rapid insights into comment trends and feedback

#### **Bespoke Knowledge Graph Creation for Legal Documents**

- Led development of an internal R&D application that generated knowledge graphs from document folders, identifying entities and relationships using Claude 3 Sonnet.
- Built an interactive UI for visualizing the graph and drilling down into source documents for context validation.

#### **Tarot-Chat**

- Personal project: Designed and deployed a Streamlit app on AWS Lightsail using GPT-3.5 to deliver interactive tarot card readings with LLM tool calls.
- Integrated Amazon affiliate recommendations based on reading topics.
- You can play with it <u>here</u>.

#### **SEC Filing Risk Triaging**

- Developed a document classification pipeline using a fine-tuned Longformer model and custom preprocessing, improving F1 score from 0.46 to 0.94.
- Delivered a 65% reduction in review time and a 130% increase in detection of potential violations.

Other Projects: Probabilistic Attribution for Walled Gardens, Hidden Patch Attacks on Optical Flow, Interface for Quantifying Uncertainty in Autonomous Vehicles, Identifying Gaps in Emotion Models Using Word Embeddings, Image De-fencing with Synthetic Data and Adversarial Loss

### May 2024 - Present

#### Jan 2024 - November 2024

### Mar 2024 - Present

Jan 2022 - Dec 2023

### Feb 2025 - Present

## Aug 2024 - Feb 2025

December 2022

May 2020

### **Publications**

Wortman, B. (2024). Models of Human Emotion and Artificial Emotional Intelligence. In: Wang, J.Z., Adams, Jr., R.B. (eds) Modeling Visual Aesthetics, Emotion, and Artistic Style. Springer, Cham. <u>https://doi.org/10.1007/978-3-031-50269-9\_1</u>

**Wortman, B**., & Wang, J. Z. (2023). HICEM: A high-coverage emotion model for artificial emotional intelligence. *IEEE Transactions on Affective Computing*.

**Wortman, B**. (2021). Hidden Patch Attacks for Optical Flow. ICML 2021 Workshop on Adversarial Machine Learning. <u>https://openreview.net/forum?id=7Zc8KHNadih</u>

### Grants

Amazon - \$300k - Affective and Social Interaction between Human and Intelligent MachinesFY 2021 - 2023PI: Dr. James WangFi Dr. James Wang

**XSEDE Allocation** - \$40k - Advancing Computational and Image Understanding Technologies **FY 2021 - 2022** PI: Dr. James Wang