Frederick Hayes III Frederickhayes3@gmail.com | (330) 495-4713 | LinkedIn: frederickhayes3 | Website: fhayes3.com

---

Professional Summary

Highly analytical Data Scientist with over 4 years of experience in AI/ML solution development and data-driven strategy. Recently graduated with a

Master of Information and Data Science from UC Berkeley (May 2025), specializing in scalable ML and **Generative AI**. Passionate about leveraging data to enhance product reliability, drive informed decision-making, and create human-centric AI solutions. Proven ability in machine learning, statistical analysis, and data visualization, with a robust background in business analytics and technical development. Seeking to contribute to cutting-edge AI product development and research.

Research Interests

My research interests are driven by a core belief that the future of advanced AI lies not in merely scaling existing brute-force paradigms, but in a more thoughtful, nuanced, and sustainable approach. I am particularly compelled by the notion that nature, having iterated intelligence for millions of years, offers profound lessons that can guide us beyond the inherent limitations and unsustainable costs (monetary, ecological, data, and natural resources) of current development paths.

Therefore, my focus is on designing **efficient AI architectures**, especially **neuromorphic computing** and **Spiking Neural Networks (SNNs)**, to explore novel forms of **interaction with cognitive systems and the human body**. I aim to develop **lifelong learning** and **continuous learning paradigms**, leveraging **Reinforcement Learning (RL)**, that enable **seamless and intuitive human-AI collaboration and understanding**. This involves exploring architectures that can help us escape the "tyranny of the Transformer," which is currently the dominant paradigm for large-scale models, to find new, more efficient, and biologically inspired ways to build AI systems. Critically, this includes investigating how **linguistic symbiology**, as processed by the human brain, can inform the design of AI systems that achieve greater levels of **inference, adaptability, and nuanced understanding across diverse languages and cultures**.

Ultimately, my quest is for elegance, precision, transparency, and intellectual integrity in AI development, recognizing that true progress stems from owning our flaws and seeking fundamental advancements that benefit all of humanity.

Education

Master of Information and Data Science | University of California, Berkeley (May 2025)

- **Cumulative GPA: 3.900**
- Focused on leading innovative, ethically sound AI/ML projects and developing human-centered data solutions.
- **Relevant Coursework:** Machine Learning at Scale (A), Applied Machine Learning (A), Data Visualization (A), Causal Inference (A), and Experimental Design (A), Natural Language Processing (B+), and Generative AI (A).

Post Graduate Program in Data Science and Business Analytics | University of Texas at Austin - McCombs School of Business (March 2023)

- GPA: 4.14
- Gained expertise in advanced data visualization, predictive modeling, regression, and time series forecasting.
- Hands-on experience in applying data science methodologies to solve business problems.

Bachelor of Science in Computer Information Systems | University of the Incarnate Word, San Antonio, TX (May 2022)

- **Overall GPA: 3.98**
- Developed technical skills in computer systems, networking, and programming.
- Trained in various aspects of system migration, data analysis, and project management.

Associate's Degree in Computer Programming | Stark State College of Technology, North Canton, OH (May 2014)

- *Courses completed (transferred to UIW)*. Focused on database management, SQL, and programming fundamentals.

Transfer Credits: Kent State University (early 1990s).

Research Experience

1. Dosage Stability Modeling Phase 1

- **Role:** Lead Modeler & Analyst
- **Context:** Neogen Corporation Project, April 2023 – Present
- **Objective:** Developed a predictive framework to estimate when pharmaceutical dosages' % Label Claim would fall below the 95% regulatory failure threshold.

- **Problem & Innovation:** Initially conceived as a single global model, the project evolved into a **per-dosage modeling strategy** after exploratory analysis revealed distinct, dosage-specific degradation patterns and nonlinearities, significantly enhancing accuracy and interpretability.
- **Methodology:**
    - **Data Strategy:** Addressed limited ground truth data by generating **synthetic training data** that matched the statistical distribution and feature correlations of true datasets, enabling robust training and hyperparameter tuning.
    - **Model & Training:** Trained each dosage-specific model using a **scikit-learn pipeline**, applying feature preprocessing (standardization, one-hot encoding), target transformation (log transformation), and **GroupKFold cross-validation** to prevent data leakage.
    - **Optimization:** Employed **Optuna and Bayesian optimization (BayesSearchCV)** for hyperparameter tuning and model selection. Top-3 models (CatBoost, LightGBM, XGBoost) per dosage were **stacked using a Ridge regressor**. Final models were serialized using joblib for consistent deployment.
- **Overfitting Mitigation:** Implemented rigorous strategies including GroupKFold, restricted hyperparameter tuning, monitoring generalization metrics on ground truth data, and checks against synthetic label copying.
- **Evaluation & Results:** Models consistently outperformed baselines on ground truth test sets. Visual decay curve plots and tabular outputs demonstrated accurate month-to-month tracking of label degradation, with predictions falling within **1–2% of actual values** for most test cases.
- **Future Work & Impact:** Phase 1 concluded with successful, deployment-ready models. The project includes a clear roadmap for **Phase 2: Reinforcement Learning-based intervention modeling**, which plans to develop a Gym-based simulation and train PPO/DQN agents to recommend experimental design or formulation tweaks. This showcases a direct pathway from applied ML to advanced RL research.

2. NeuroBeacon (UC Berkeley Capstone Project) | *Team Member, Frederick Hayes III* |

Project Page

- **Context:** UC Berkeley School of Information, Spring 2025
- **Objective:** Developed an **adaptive, RL-driven, game-based cognitive training platform** to empower individuals to maintain and enhance cognitive function across the lifespan. It addresses the issue of uninspiring cognitive training by centering engagement and adaptability through real-time dynamic difficulty adjustment.
- **Methodology:** Leveraged **Deep Q-Learning** with pre-training on large, diverse datasets (EdNet, Duolingo's SLAM). Engineered domain-specific state variables, user embeddings, and a custom reward function. Implemented Primary/Target Networks and a Replay Buffer for continuous retraining and incremental learning. Designed a

full-stack prototype with cloud deployment (AWS SageMaker, Lambda, DynamoDB, ECS, Amplify).

- **Evaluation & Results:** Validated performance using simulated users (70 synthetic users across 7 profiles) and real beta test data. Demonstrated superior average cumulative reward and optimized difficulty distribution for various user types (e.g., beginners, those in cognitive decline) compared to baselines.
- **Key Learnings & Impact:** Showcased the flexibility of RL, the criticality of system architecture for continuous adaptation, and a framework for transferring learned policies to unseen environments. Aims to provide personalized, adaptive learning to improve everyday life and meet cognitive health needs.
- **Validation:** Selected as one of six projects for the **Spring 2025 Capstone Showcase**.

3. Retrieval-Augmented Generation for Audience-Specific Answering

- **Role:** Lead Researcher & Developer
- **Context:** UC Berkeley MIDS GenAI Course Project, April 2025
- **Objective:** Explored the construction and optimization of a dual-audience RAG system to serve both engineering researchers and marketing professionals with tailored LLM responses. Addressed the critical gap where LLM responses vary significantly by audience in enterprise settings.
- **Methodology:** Designed a flexible architecture using **LangGraph-based orchestration** for dynamic sub-question decomposition, advanced reranking, and hybrid scoring. Utilized interchangeable embedding models (mpnet-dot-v1, mixedbread), rerankers (MiniLM, ELECTRA, Cohere), and LLMs (Cohere, Mistral). Developed the **Fred Score**, a central, evolving composite metric balancing Fluency & Form, Semantic Match, Retrieval Alignment, and Responsiveness.
- **Rigorous Evaluation:** Rigorously tested against a labeled set of 75 questions with audience-specific gold answers. The Fred Score evolved to address hallucination detection (e.g., the "Paper Airplane Test").
- **Key Findings & Impact:** Achieved **Fred-R 0.82 and Fred-M 0.75** scores, significantly outperforming baselines. Prompting, reranker choice, chunk optimization, and LangGraph decomposition were key drivers of improvement, enhancing grounding and answer completeness.
- **Contribution:** Demonstrated that RAG systems can be tailored for divergent user groups through modular architecture, dynamic decomposition, and audience-aware prompting, with implications for rule-intensive domains. The unique benchmarking system developed is being added as an example to the course curriculum.

4. Rules Lawyer: An NLP-Based Question-Answering System for Board Game Rulebooks

- **Role:** Lead Developer
- **Context:** UC Berkeley MIDS NLP Course Project
- **Objective:** Developed Rules Lawyer, a QA system to provide quick, accurate rule clarifications for complex board game rulebooks. Addressed the challenge of multi-hop queries and nuanced rule interpretation in a niche domain.

- **Methodology:** Utilized retrieval-augmented methods and fine-tuned state-of-the-art NLP models including T5 (Base/Large), Flan-T5, BERT, and UL2. Involved dataset creation (500+ Q&A pairs), advanced preprocessing (sliding window splits, normalization), and semantic retrieval with reranking.
- **Evaluation:** Assessed performance using ROUGE (1, 2, L), BLEU, and BERTScore F1. Qualitative analysis provided insights into model strengths and limitations.
- **Key Findings & Impact:** The optimized **T5-Large model achieved best overall performance** (81.47 ROUGE-1, 74.19 ROUGE-2, 90.57 BERTScore F1), with UL2 excelling in semantic reasoning. Demonstrated the feasibility of applying RAG techniques to niche, rule-intensive domains, enhancing accessibility of complex systems.
- **Contribution:** Contributed a novel dataset, retrieval framework, and insights into hyperparameter tuning for rule-intensive QA tasks.

5. Predicting Emotional Wellness through Music Consumption

- **Role:** Co-Developer & Data Analyst
- **Context:** UC Berkeley MIDS Course Project
- **Objective:** Investigated whether the mental health outlook of a music listener could be predicted based on their music consumption behavior and self-reported well-being. This aimed to decode the link between genre preference and mental health.
- **Methodology:** Leveraged a low-data regime dataset (736 observations). Employed **6 different models** and **39 individual experiments** across **11 experiment groups**, utilizing **automated hyperparameter tuning (Optuna)** and **Stratified 5-fold Cross-Validation**. Models included Ensemble (soft voting) and SVM.
- **Key Findings & Impact:** Achieved **86.16% accuracy and 0.8680 F1 Score** on average test metrics. Demonstrated high precision and recall and effective class distinction. Identified a powerful link between genre preference and well-being, even with limited data.
- **Ethics & Fairness:** Addressed model bias from the dataset and focused on fair modeling.

6. Impact of Camera Use on Online Presentation Comprehension

- **Role:** Co-Researcher & Analyst
- **Context:** UC Berkeley MIDS Course Project
- **Objective:** Investigated whether having cameras on during online presentations improved comprehension, driven by the increasing shift to online learning and social presence theory.
- **Methodology:** Designed a **Treatment – Encouragement Design** experiment with **randomization** using Zoom breakout rooms, involving 223 Berkeley undergraduates. Outcome was a 10-point multiple-choice comprehension quiz. Employed ROXO Grammar (Randomization, Observation, Treatment, Observation).
- **Key Findings & Impact:** Demonstrated a positive, statistically significant effect (p-value: 0.07795726) of camera use on comprehension, with an estimated

CACE (Complier Average Causal Effect) of 1.189068 , offering insights for online education best practices.

7. Internet Use and Age

- **Role:** Data Analyst & Modeler
- **Context:** UC Berkeley MIDS Course Project
- **Objective:** Explored the relationship between age and internet use, aiming to identify potential divides across age groups and generations in an increasingly online world.
- **Methodology:** Utilized data from the **General Social Survey (GSS)** (2,069 observations) , carefully evaluating LLM assumptions like Independence of Observations (IID) and No Perfect Collinearity (BLP). Employed
  Linear and Quadratic Regression models , demonstrating how transformations improved model fit.
- **Key Findings & Impact:** Identified a clear negative relationship between age and internet use, where for each increase of 1 year in age, internet usage decreased by approximately 0.32 hours. This highlighted important divides across age groups and generations in internet access and use , underscoring critical societal implications.

8. AGM's New Horizons: Enhanced Delivery Network

- **Role:** Team Member, Data Architect & Modeler
- **Context:** UC Berkeley MIDS Course Project (DATASCI 205: Fund. Data Eng), Spring 2024
- **Objective:** Designed a new, enhanced delivery network for "AGM," aiming to expand its footprint, increase reliability and consistency of deliveries, and achieve a 1-hour delivery promise. This required integrating new technologies and upgrading technology infrastructure.
- **Methodology & System Design:** Proposed a solution integrating multi-modal transportation (vans, **BART system**, and **drones**). Leveraged a suite of modern data tools:
  Neo4j (graph database for route calculation, centrality, dynamic networks), **Redis** (in-memory database for geospatial indexing, ML feature storage, message brokering), and **MongoDB** (NoSQL for geospatial indexing, real-time monitoring via Change Streams, scalability).
- **Implications:** Showcased the design of a complex, **real-time, data-driven logistical network** focused on efficiency and dynamic adaptability in a spatial domain.

Professional Experience

Data Scientist | Neogen Corporation (April 2023 – Present, Remote)

- Developed and deployed an AI model (1D CNN, TensorFlow) for LAMP primer design optimization, significantly reducing false positives and enhancing the specificity and reliability of molecular diagnostic assays.
- Built and validated per-dosage machine learning models (CatBoost, LightGBM, XGBoost) to predict pharmaceutical dosage stability, accurately estimating time-to-failure below regulatory thresholds.
- This project lays groundwork for Phase 2: Reinforcement Learning-based intervention optimization.
- Created RAG models to enhance institutional knowledge availability and developed methods to extract user interaction metrics to assist in employee training initiatives.
- Leveraged advanced data analytics and statistical analysis to support regulatory compliance initiatives.
- Collaborated with business and lab-based stakeholders to design and deploy proprietary ML models and neural networks to amplify the effectiveness of lab processes and equipment.

Web Platform Specialist | University of the Incarnate Word (July 2018 – April 2023) – San Antonio, TX

- Designed and executed the university's web platform, incorporating accessibility standards and workflow improvements for over 5,000 pages.
- Managed CMS migrations, content updates, and optimizations to enhance efficiency and web accessibility.
- Provided training and support to over 300 users from diverse backgrounds on CMS and accessibility.

Applications Developer | Case Western Reserve University (October 2014 – July 2018) – Cleveland, OH

- Created and managed CMS-based content assets for the university's digital platform.
- Directed site-wide migration of over 1,000 pages between CMS products, optimizing content load times and improving user engagement.
- Acted as a liaison between stakeholders and technical teams to meet content and operational needs.

Publications / Writings

- **Hayes, F.** (2025). *Let's Talk About the Work: When AI Dreams in Waves; Exploring the Unexpected Behaviors of Continuous Thought Machines*. Medium. When AI Dreams in Waves
- **Hayes, F.** (2025). *Let's Talk About the Work: How AI Can Help Us Spot the Fakes in DNA Testing*. Medium. How AI Can Help Us Spot the Fakes in DNA Testing

- **Hayes, F.** (2025). *Retrieval-Augmented Generation for Audience-Specific Answering: Final Report*. UC Berkeley MIDS Course Project Report.
- **Hayes, F.** (2024). *Rules Lawyer: An NLP-Based Question-Answering System for Board Game Rulebooks*. UC Berkeley MIDS Course Project Report.
- **Hayes, F., Chong, Q., & Hernandez, J.** (2024). *Decoding the Mental Health Melody: Predicting Wellness Through Music Consumption*. UC Berkeley MIDS Course Project Presentation.
- **Hayes, F., Amer, Z., & Baugh, R.** (2024). *Impact of Camera Use on Online Presentation Comprehension*. UC Berkeley MIDS Course Project Presentation.
- **Hayes, F., Chan, J., & Polovina, N.** (2024). *Internet Use and Age*. UC Berkeley MIDS Course Project Presentation.
- **Hayes, F., Cho, A., Solow, D., & Chan, J.** (2024). *AGM's New Horizons: Enhanced Delivery Network*. UC Berkeley MIDS Course Project Presentation.
- **Hayes, F.** (2025). *Dosage Stability Modeling Phase 1: Final Report*. Neogen Corporation Project Report.
- **Hayes, F.** (2025). *Moshi: A Speech-Text Foundation Model for Real-Time Dialogue - A Critical Analysis*. UC Berkeley Reading Group Presentation.

Awards & Honors

- **UC Berkeley MIDS Capstone Showcase Selection** | Spring 2025
    - NeuroBeacon project selected as one of six capstone projects for presentation at the highly competitive Spring 2025 Capstone Showcase.

Relevant Coursework

University of California, Berkeley – Master of Information and Data Science

- **DATASCI 267: Generative AI** (A) – Comprehensive exploration of generative models, including architectures and applications.
- **DATASCI 261: Machine Learning at Scale** (A) – Focused on developing expertise in ML solutions for large-scale and complex problems.
- **DATASCI 266: NLP Deep Learning** (B+) – Covered advanced topics in Natural Language Processing and deep learning techniques.
- **DATASCI 207: Applied Machine Learning** (A) – Practical application of machine learning algorithms and methodologies.
- **DATASCI 241: Experimentation & Causal Inference** (A) – Rigorous training in designing experiments and drawing causal conclusions from data.
- **DATASCI 205: Fundamental Data Engineering** (A) – Foundations of data engineering principles and practices.

- **DATASCI 203: Statistics for Data Science** (A-) – Core statistical concepts essential for data analysis.
- **DATASCI 201: Research Design & Data Analysis** (A) – Principles of research design and analytical methods for data.
- **DATASCI 200: Data Science Programming** (A) – Foundational programming skills for data science.

Teaching / Mentoring Experience

- **Technical Trainer & Educator** | University of the Incarnate Word (July 2018 – April 2023)
    - Trained over 500 users of all skill levels, from incoming freshmen to tenured faculty and non-technical staff, in diverse technical and digital literacy skills.
    - Conducted training sessions covering web design principles, basic HTML, web accessibility standards (**WCAG**), mobile responsiveness, CSS, Content Management System (CMS) usage, Microsoft Office Suite, and other essential IT skills.
    - Adapted teaching methods to accommodate various learning styles and technical proficiencies, ensuring effective knowledge transfer across a broad audience.

Technical Skills

- **Languages:** Python, R, SQL, HTML, JavaScript, CSS
- **ML & AI Frameworks:** TensorFlow, PyTorch, Keras, scikit-learn, NumPy, Pandas, CatBoost, LightGBM, XGBoost, Random Forest, Extra Trees, LangChain, LangGraph
- **Data Analysis & Visualization:** Predictive Modeling, Statistical Analysis, Data Visualization (Matplotlib, Tableau), Data Cleaning, Feature Engineering, Graph Analysis, Experimentation Design, A/B Testing, Neo4j
- **AI/ML Specializations:** Reinforcement Learning (RL), Natural Language Processing (NLP), Computer Vision, Generative AI, Deep Learning, 1D Convolutional Neural Networks (CNNs)
- **Cloud Platforms:** AWS, Azure, GCP, Google Cloud Storage
- **Tools & Methodologies:** Git, Keras Tuner, Optuna, scikit-learn pipelines, joblib, Adobe Creative Cloud, Microsoft Office Suite, Red Hat Linux, Cascade CMS, CRM, Agile, Project Management