Jacob Cutter

San Ramon, CA | 510-861-0706 | jecutter92@gmail.com | [LinkedIn](#) | **[Website](#)** | **[GitHub](#)**

Scrappy and creative data scientist specializing in ML model development, data pipelining, and mining actionable insights from noisy data. PhD in Experimental Physics with a practical approach to problem solving, looking to contribute to impactful AI products built to scale.

SKILLS

Programming/Databases: Python, C++, BASH/shell, SQL (MySQL, PostgreSQL, SQLite), Github/GitLab
Techniques (Tools): ML/AI Modeling (Scikit-Learn, PyTorch), ETL (SQL, Pandas, NumPy), Visualization (Matplotlib, Seaborn, Jupyter, Bokeh), Apps (Flask, Docker), NLP (NLTK, spaCy), Statistical Modeling and Analysis, Signal Processing, Data Processing/Management (Grid Computing, SGE, SLURM, RAID)

EXPERIENCE

Deepgram, San Francisco Bay Area, CA
Data Scientist, Team Lead of Product Development   Oct 2022 - Present
- Coordinate cross-functional efforts across Product, DataOps, and Engineering departments to optimize AI model R&D and expand our ASR languages and product offerings to customers
- Set team priorities and tasking based on feedback from the Sales and Product teams
- Employ the team's latest AI architectures and tooling to train and productionalize new ASR models
- Continue IC work while managing a team of researchers to execute on technical projects that have a direct impact on the revenue stream (i.e. customer acquisition, upsell opportunities)

Data Scientist   Oct 2020 - Oct 2022
- Trained and maintained dozens of iterations of core E2E ASR models using PyTorch frameworks
- Devise and report KPIs to internally evaluate ASR and diarization models (core speech products)
- Evaluate novel next-generation deep learning architectures to optimize ASR accuracy/performance
- Spearhead project to improve ASR punctuation by building out multilingual pipelines for text cleaning, data preparation, and experimental ML modeling with PyTorch
- Assist with the development of Flask applications for NLU products such as text summarization
- Sourced and preprocessed TBs (1000s of hours) of speech data for model training and evaluation, including curation for curriculum training and transfer learning domain-specific models
- Leverage customer metadata and mine information from production PostgreSQL databases to construct representative training/evaluation datasets and improve product R&D
- Build modular, parallelized Python tools for model training, testing, and data analysis

Insight Data Science, *Data Science Fellow,* San Francisco CA   May 2020 - July 2020
- Created a music classification app for listeners and content creators to filter Spotify playlists by emotion
- Combined emotional labels mined from Last.fm SQLite databases, Spotify audio features, and song lyric sentiment to build classifiers that predict the emotional charge of songs with up to 71% accuracy
- Deployed the classification models on AWS in an interactive Streamlit web application

UC Davis Physics Department, *Graduate Student Researcher and T.A.*   Sept 2014 - August 2020
- Designed local R&D particle detection experiments, using statistical analysis and modeling to characterize important nuclear processes and inform major dark matter searches
- Developed end-to-end C++/Python pipelines for signal processing, data reduction and visualization, and synthesizing TBs of noisy binary waveform data into physical measurements
- Used MySQL database replication and Flask to remotely monitor lab operations via web interface
- Managed large-scale data processing pipelines for an international multi-million dollar experiment
- Taught discussion/lab sections for multiple undergraduate physics courses, receiving excellent student evaluations averaging higher than 4.5/5 across the board
- Mentored ~75 high school students over 3 summers as part of the COSMOS program, working with students to execute intensive data-driven astrophysics projects

PROJECTS

NBA Basketball Analytics | github.com/jecutter/nba-data-models, jecutter.github.io/blog/   Jan 2020
- Web-scraped many seasons of NBA player, lineup, and play-by-play data using Scrapy and Selenium
- Created an interactive visualization dashboard using Bokeh to explore player and lineup data
- Built a Random Forest classifier model to perform useful player comparisons with 92% accuracy
- Developed an RAPM model using ridge regression on matchup results harvested from play-by-play data, providing a lineup-independent impact metric for scouting undervalued players

EDUCATION

University of California, Davis   August 2020
Ph.D. in Experimental Particle Physics, Designated Emphasis in Nuclear Science

University of California, Davis   June 2014
B.S. in Particle Physics