Based in the Seattle area
206–607–7302
jamarq@uw.edu
www.linkedin.com/in/james-marquardt

James Marquardt

Experienced leader passionate about building products using modern machine learning techniques (LLMs, generative AI, reinforcement learning optimization) and software engineering skills.

Experience

2018–2023 **Senior Machine Learning Engineer and Data Scientist**, *98point6*.

98point6 provides a platform for healthcare providers to offer convenient and efficient visits via a mobile app. My role was to discover and implement machine learning solutions geared towards time and cost saving opportunities. Highlights:

- Improved team performance by creating a program for structured physician shadows, developing a framework for deploying beta features, and maintaining an internal blog of product focused applied research.
- Initiated and oversaw the development of features that saved 20–40% of doctors' time in projects such as diagnosis suggestion, prescription suggestion, and clinical best practice automation.
- Reduced prescription related safety incidents by 50% by incorporating medical literature context into search services and implementing a RLHF ranking system based on incident responses.
- Employed MLOps methods to ensure stable product development. This included using general technologies such as Terraform, DVC, and Jenkins. It also included AWS-specific technologies such as IAM, ECS, Lambda, S3, CDK, API Gateway, and Sagemaker.
- Bootstrapped training data collection by implementing Slack bots to capture previously unstructured and discarded information, saving roughly $10k in specialized annotator costs per project.
- Created doctor-patient matching optimization system with reinforcement learning methods and a large language model (LLM) based observation space. Observed 15% patient throughput improvement in beta test.
- Directly supervised feature teams consisting of engineers, researchers, and domain experts.

Tech used: Python, Haskell, Java, React, PyTorch, Tensorflow, Ray, Elasticsearch, Spark, Redshift, AWS

2016–2018 **Platform Technical Lead and Data Scientist**, *KenSci*.

KenSci provided healthcare specific machine learning solutions for payers and providers with a goal of improving safety and reducing inefficiencies in various care and insurance settings. My role was to oversee client engagements as the lead data scientist and to manage the platform engineering team
Highlights:

- Identified previously unknown community acquired sepsis and end of life care issues in a large hospital system that contributed 5-15% in excess total inpatient costs, and delivered machine learning models for early patient level intervention
- Partnered with one of the largest healthcare providers in Singapore to modernize their P&L analysis using PowerBI and deployed a machine learning system to save 3-4% of annual costs by flagging instances of potential fraud, waste, and abuse
- Led a team of five engineers to develop a scalable and flexible platform for deploying machine learning models on Azure using technologies as AKS, Functions, Active Directory, and Azure Machine Learning
- Developed ETL pipelines for processing millions of records in HL7 and other EHR data formats stored in on premise data centers and relocating them in Azure Storage using Airflow. Improved distributed compute components of these pipelines leading to processing times dropping from $>$ 2 hours to $<$ 5 minutes

Tech used: Python, R, C#, Tensorflow, Spark, SQL Server, Azure, Airflow

2015–2016 **Data Scientist**, *Microsoft*.

- Found empirical evidence that countered a company wide belief that increased general Azure expenditures in the first month of a contract was a good predictor of churn avoidance. Demonstrated that more granular user behaviors were were better predictors and evangelized a plan for technical program managers to help customers better utilize Azure services
- Developed machine learning models to predict outages in on premises installations based on service logs and configuration settings that enabled a 20% reduction in Active Directory related incidents.

Tech used: Python, R, C#, Hadoop (Hive), SQL Server, Azure Cosmos

2014–2015 **Data Scientist**, *University of Washington*.

- Conducted research on developing machine learning algorithms for analyzing large-scale medical data sets
- Worked with a team of researchers to develop predictive models for identifying patients at high risk for developing chronic diseases
- Presented research findings at multiple conferences and published papers in peer-reviewed journals

Tech used: R, Python, SQL Server

2004–2010 **Infantry Sergeant**, *United States Army*.

- Trained and mentored a team of five infantryman in tactics, operational security, and cross cultural relations
- Served as an Arabic translator in high impact meetings between military and local officials

Selected projects

2022– **Epana, a fast re-ranker for AI generated text**.
Modern text generators by LLMs require expensive fine tuning procedures in order to adapt them to specific policies. This tool provides a way for generated content options to be re-ranked using a configurable and fast metric learning based model.

2022– **Sophos, reasonably sized generative models**.
Extremely large language models demonstrate many amazing capabilities, but their size makes using them problematic for both inference and training. This project creates a reusable pipeline for reducing the size of GPT style models through methods including quantization and domain directed pruning (iterative pruning and fine-tuning).

2022–2023 **OpenAI evaluations**.
Contributed code to the OpenAI evaluators project on GitHub (openai/evals) to be used in the improvement of large language models. Notable additions included evaluations of model ability to interpret ASCII art and perform geospatial reasoning

2021 **LogicCLIP, an early fusion multi modal interpretation framework**.
Interpretability of machine learning models frequently becomes problematic when operating in the high dimensional feature space of image data. This system provided a way to extract features from images using the Google CLIP model and use them along with structured data in order to build models whose predictions could be easily interpreted using the understandable feature space

2018 **ELMER, a large language model for emergency room triage**.
Many ER triage systems rely solely on manually extracted features in order to determine patient urgency, with much of the context provided in the patient's chief complaint being discarded for decision making purposes. This system utilized the ELMo large language model along with additional structured features to achieve state of the art urgency categorization.

2014 **High Performance Computing (HPC) celestial body categorization**.
As part of a NSF funded residence with the AURA observatory in Chile, developed models for celestial body categorization and operationalized them on a HPC cluster using LSF

Selected publications

Eckert, C., Nieves-Robbins, N., Spieker, E., Louwers, T., Hazel, D., **Marquardt, J.**, ... & Teredesai, A. (2019). *Development and prospective validation of a machine learning-based risk of readmission model in a large military hospital. Applied Clinical Informatics, 10(02), 316-325.*

Sushmita, S, Newman, S., **Marquardt, J.**, Ram, P., Prasad, V., De Cock, M., & Teredesai, A. (2015). *Population cost prediction on public healthcare datasets. Proceedings of ACM Digital Health, 87–94.*

Chin, S.-C., **Marquardt, J.**, Liu, R., De Cock, M. (2014). *Prediction of Hospitalization Cost for Childbirth Proceedings of HI-KDD 2014 (ACM SIGKDD Workshop on Health Informatics), workshop at KDD2014 (20th ACM SIGKDD Conference on Knowledge Discovery and Data Mining)*

De Cock, M., Teredesai A., Ram, P., Hazel, D., **Marquardt, J.**, Velamur, R., Basu Roy, S., Prasad, V. (2014). *A Distributed Machine Learning Framework for Multi-Factor Healthcare Cost Prediction Proceedings of W3PHI-2014 (the 1st AAAI Workshop on World Wide Web and Public Health Intelligence), workshop at AAAI2014*

Marquardt, J., Farnadi, G., Vasudevan, G., Moens, M. F., Davalos, S., Teredesai, A., & De Cock, M. (2014). *Age and gender identification in social media. Proceedings of CLEF 2014 Evaluation Labs, 1180, 1129-1136.*

Marquardt, J., Newman, S., Hattarki, D., Srinivasan, R., Sushmita, S., Ram, P., ... & Teredesai, A. (2014). *Healthscope: An interactive distributed data mining framework for scalable prediction of healthcare costs. In 2014 IEEE International Conference on Data Mining Workshop, 1227-1230.*

Education

2013–2014 **University of Washington**, *M.S. in Computer and Science*.
Thesis on distributed topic modeling (Distributed Diverging Topic Models: A Novel Algorithm for Large Scale Topic Modeling in Spark)

2010–2013 **University of Washington**, *B.S. in Computer Science and Systems*.

Service

2020– **Youth sports coach**.
Serving as head coach of youth basketball and football teams

2021– **PTA student performances coordinator**.
Responsible for scheduling and organizing student performances including talent shows and culture festivals

2022– **Veterans mentoring**.
Providing mentoring for military veterans using the Veterati platform