William Ma

william@williamma.org

LinkedIn: www.linkedin.com/in/williamma12
Github: www.github.com/williamma12

Experience

- **Senior Machine Learning Engineer, Apple Search AI/ML** — Seattle, WA
  Apple — *July 2021 - Present*
  - Tech lead for a team of 6 Sr./Staff level engineers building features for and maintaining the Siri Knowledge Q&A system (**>70% of all Siri traffic goes through our system at ~15k queries per second**)
  - Owned entity linking pipeline and shipped features like landmark identification, video question answering disambiguation, and Safari highlights which were **critical features for the release of Apple Intelligence**
  - Improved entity scoring model by building features for and training a custom BERT-based model using TF keras resulting in top-1 accuracy improving from 45% to 96% when correct entity is available
  - Redesigned entity linking pipeline that resulted in a simplified deployment and usage of server-side entity linking by generalizing the models and pipeline deployment while keeping low-latency below 150ms and improving accuracy by 5%
  - Fine-tuned and deployed in-house, 30b parameter LLMs for mention generation and entity reranking in entity linking pipelines that detected dominant entities in **10+ million websites representing 10TB weekly** (shipped 4× faster than comparable features)
  - Built and maintained entity generation for entity linking, via sparse and dense retrieval using Approximate Neural Network, which included automated evaluation, data staging, data gating, and performance improvements to update ~1TB of data every day
  - Managed customer onboarding, led the development of product owner requests from PRDs to releases, and am primary on-call for all high profile events such as Ocars, Grammy's, and 2024 Election Day
- **Graduate Student in Data Systems** — Chicago, IL
  University of Chicago — *July 2020 - June 2021*
  - Designed an architecture for a new data lake with first-class support for intermediate state storage and recomputation in a streaming setting to replace lambda architecture in both machine learning model serving and data analytics workflows
  - Showed using SparkSQL and Kafka that sharing query subplans in an incremental batch execution engine can lead to over 6× runtime reduction over the state-of-the-art (doi.org/10.1145/3448016.3457282)
  - Demonstrated a 20% performance improvement from using a bloom filter on stored intermediate state generated from maintaining materialized views
  - TA for database and data science classes (~ 100 students), which involved debugging and grading student projects; mentoring students through quarter long projects; and developing a database for educational purposes
- **Undergraduate Researcher in Data Systems** — Berkeley, CA
  University of California, Berkeley - RISE Lab — *August 2018 - May 2020*
  - Maintainer of Modin, a popular (~10k stars) open-source library for distributed dataframes
  - Designed a data model and demonstrated Modin's >100× improvement over SOTA in dataframe operations
  - Designed and proved a sound data model and type system for dataframes, which facilitates future database-like optimizations within dataframes (arXiv:2001.00888)
  - Developed an intelligent partitioning scheme for dataframes, which lead to a 50% improvement over SOTA approach
  - Demonstrated a 15× loss of revenue in GCP BigTable and introduced a new cost model to prevent this loss and provide users with 50% faster queries over SOTA
  - Developed cost-based optimizations for TPC-H in a simulated serverless SparkSQL for 2× improvement over SOTA (doi.org/10.1145/3318464.3384410)

Education

- **University of Chicago** — Chicago, IL
  MSc in Computer Science; GPA: 3.9; Focus: Databases — *September 2020 – June 2021*
- **University of California, Berkeley** — Berkeley, CA
  BA in Computer Science w/ High Distinction in General Scholarship; GPA: 3.8 — *August 2016 – May 2020*

Skills

Languages: Python, SQL, Golang, Rust, Bash, C, Coq, R, LaTeX
Frameworks: Spark, TensorFlow, keras, Pandas, NumPy/SciPy, Jupyter, Matplotlib/Seaborn, Bokeh
Tools: AWS (including EMR), Docker, Kubernetes, Git, Spark, vim