William Ma

wma1017@gmail.com

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

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*

Experience

- **Graduate Student in Data Systems** — Chicago, IL
  University of Chicago — *July 2020 - Present*
  - 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 (paper accepted to SIGMOD 2021)
  - Explored the trade offs of using a bloom filter at different storage media latencies and demonstrated the 20% performance improvement that can be had from using bloom filters (report: williamma.me/reports/bloom_filter.pdf)
  - 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, an open-source, drop-in replacement for a distributed pandas (i.e., dataframe) library.
  - Designed a data model and demonstrated Modin's over 100× improvement over the current state-of-the-art in dataframe operations (paper: doi.org/10.14778/3407790.3407807)
  - Designed and proved a sound data model and type system for dataframes, which facilitates future database-like optimizations within dataframes (report: williamma.me/reports/dataframe_type_system.pdf)
  - Developed an intelligent partitioning scheme for dataframes, which lead to a 50% improvement over the current state-of-the-art approach (report: williamma.me/reports/dataframe_partitioning.pdf)
  - 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 the current state-of-the-art (paper: doi.org/10.1145/3318464.3384410)
  - Developed cost-based optimizations for TPC-H queries in a simulated serverless SparkSQL for 2× improvement over the current state-of-the-art (report: williamma.me/reports/serverless_query_opt.pdf)
- **Undergraduate Researcher in Applied Statistics** — Berkeley, CA
  University of California, Berkeley - Statistics Department — *January 2018 - Present*
  - Implemented a distributed conjoint analysis, a commonly used survey technique, in Python using both multiprocessing and multithreading to have a 10× runtime reduction in estimating the preferences of survey respondents
  - Demonstrated that typical applications of conjoint analysis violated the underlying assumptions, which leads to erroneous conclusions and biased estimates of up to 40% off from the ground truth
- **Undergraduate Researcher in Digital Art History** — Berkeley, CA
  University of California, Berkeley - Art History Department — *June 2017 - May 2019*
  - Analyzed Roman Imperial coinage to show that trends of certain characteristics (e.g., "divus", "radiate") correlate to specific times in Roman history, such as 3rd century crisis and rule of Constantine (To be published in May 2021)
  - Used bokeh to build interactive visualizations the findings of the trends in Roman Imperial coinage (williamma12.github.io/roman_coinage/)
  - Created and managed a SQLite database containing the textual coinage data from the British Museum, the American Numismatic Society, and OCRE website containing information for over 100k coins

Skills

Languages: Python, SQL, Rust, Bash, C, Coq, R, LaTeX
Frameworks: Pandas, NumPy/SciPy, Jupyter, Matplotlib/Seaborn, Bokeh
Tools: AWS, GCP, Git, Linux, Spark, vim