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** — Chicago, IL
  University of Chicago — *July 2020 - Present*
  - Designed architecture for data lake with first-class support for intermediate state storage and recomputation in a streaming setting built for online machine learning models
  - TA for ~ 100 student database and data science classes, which involved debugging and grading student projects; mentoring students through quarter long projects; and developing a database for educational purposes
- **Undergraduate Researcher** — Berkeley, CA
  University of California, Berkeley - RISE Lab — *August 2018 - May 2020*
  - Maintainer of Modin, a drop-in replacement for a distributed pandas library.
  - Designed data model and demonstrated the over 100× improvement over the current state-of-the-art in parallelizing dataframe operations (paper: doi.org/10.14778/3407790.3407807)
  - Designed and proved soundness of data model and type system for dataframes to facilitate 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 state-of-the-art approaches (report: williamma.me/reports/dataframe_partitioning.pdf)
  - Demonstrated 15× loss of revenue in GCP BigTable and introduced a new cost model to prevent this 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 SparkSQL TPC-H queries in a simulated serverless Spark for 2× improvement over the current state-of-the-art (report: williamma.me/reports/serverless_query_opt.pdf)
- **Undergraduate Researcher** — 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** — Berkeley, CA
  University of California, Berkeley - Art History Department — *June 2017 - May 2019*
  - Analyzed Roman imperial coinage to show trends of certain characteristics (e.g., "divus", "helmet", "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 visualize the findings of the Roman Imperial coinage (williamma12.github.io/roman_coinage/)
  - Created and managed a SQL database through SQLite containing the data from the British Museum, the American Numismatic Society, and OCRE website containing information for over 100k coins
- **Software Engineering Intern** — Berkeley, CA
  Starbutter AI — *September 2017 - January 2018*
  - Used Python, AWS, and PostgreSQL to create and maintain the analytics platform to analyze the product performance and lead weekly analytics meetings, which led to a 50% increase in users
  - Used Python, PostgreSQL, and AWS to create, clean, and manage relational databases to create reports for partnered companies and potential partnerships, which led to partnering with over 5 companies

Skills

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