Eugene Wu **eugenewu@mit.edu, (510) 499-5876**
32 Vassar St Room G930, Cambridge, MA 02139

RESEARCH INTERESTS

I am broadly interested in data management systems that extend data analysis capabilities to non-expert users. Relevant fields include core database optimization, data provenance, and interface design.

EDUCATION

Expected 2014 **Massachusetts Institute of Technology**, Cambridge, MA
Ph.D., Electrical Engineering and Computer Science
Advisor: Samuel Madden
Dissertation: Implementation and Applications of High Performance Provenance Systems for Data Analysis

May 2010 **Massachusetts Institute of Technology**, Cambridge, MA
M.S., Electrical Engineering and Computer Science
Advisor: Samuel Madden

Dissertation: Shinobi: Insert-aware Partitioning and Indexing Techniques For Skewed Database Workloads

Spring 2007 **UC Berkeley**, Berkeley, CA
B.S., Electrical Engineering and Computer Science

PROFESSIONAL EXPERIENCE

2007–2012 **Massachusetts Institute of Technology, Cambridge, MA**
Ph.D. Student – CSAIL

MIT Big Data Challenge
I developed and ran MIT's largest Big Data prediction and visualization challenge.
`http://bigdatachallenge.csail.mit.edu`

"Why" Analysis of SQL Aggregate Queries
I designed and implemented an analysis framework to explain outliers in the results of aggregation queries by constructing predicates on the input data. I formalized the concept of predicate influence and identified several operator properties to enable more efficient search algorithms on common statistical aggregates.

Efficient, Low Overhead Provenance
I designed, prototyped, and evaluated a low overhead provenance system for large-scale scientific workflow applications that process gigabytes of data per second.

Query Processing with Humans
This project pioneered the use of human computation platforms such as Mechanical Turk within a database query execution engine.

Index and Partitioning Techniques
I investigated the application of indexing and partitioning techniques for time-varying and skewed query workloads. Shinobi incrementally re-partitions and indexes database tables based on recent query access patterns. Our subsequent No Bits Left Behind paper proposed the use of unused space in B-tree index pages as a cache for heavily accessed tuples. This could improve the performance of skewed query workloads such as Wikipedia's access patterns by up to three orders of magnitude.

Trajectory Optimized Storage
I implemented the core storage system for TrajStore, a high performance data management system for storing and querying vehicle trajectory data by location and time. The system incrementally optimizes the storage layout as the query workload changes over time.

2007-2008 **Google Inc., Mountain View, CA**
Intern – Data Management Research
I worked in Alon Halevy's data management group on the WebTables project to mine the Google web corpus for tabular data. I developed the table extraction pipeline and extracted more than 125 million tables. In addition, I built a table search engine that lets users query over the structured data and automatically visualize attributes in graphs or maps.

Summer 2006 **Yahoo!, Santa Clara, CA**
Engineering Intern
I explored efficient implementations of RDF stores for an internal project.

Summer 2005 **Microsoft Inc., Redmond, WA**
Engineering Intern
I worked on efficient deep cloning and other internal features in Exchange Server

Spring 2005 **IBM Extreme Blue., Almaden, CA**
Engineering Intern
I developed a new software patch service for DB2 for z/OS team that reduced patch application times from the order of months to minutes.

2004–2006 **UC Berkeley, Berkeley, CA**
Undergraduate Researcher – Computer Science Department

High Performance Stream Processing
I designed and implemented one of the first high performance complex event processing systems for detecting high level events (e.g., shoplifting occured) from streams of raw sensor events (e.g., RFID tag XXX detected). Our results were published at SIGMOD, the premier database conference.

The HiFi Project
I implemented the RFID reader interface for extracting raw events from early RFID readers and the interactive dashboard for the VLDB demonstration. HiFi is a research project around cascading stream architectures for large-scale geo-distributed receptor-based networks.

TEACHING EXPERIENCE

Fall 2013 — *Instructor, Big Data Systems (MIT 6.885)*
I co-developed and instructed MIT's first Big Data course focused on large scale data analysis tools and techniques. Topics ranged from data cleaning and integration, large-scale systems like Hadoop, to scalable visualization techniques. We developed eight labs to give students hands-on experience with the systems covered in class. The course is freely available online at `http://github.com/mitdbg/asciiclass`

Spring 2012 — *Instructor, Introduction to Data Analysis*
I co-developed and taught an Introduction to Data Analysis course to approximately 20 students during MIT's Independent Activities Period in January. The course is freely available online at `http://dataiap.github.io`

2011 – 2012 — *Head of Curriculum, MEET*
MEET is a 3-year technology program and peace initiative that teaches Israeli and Palestinian high school students. I organized curriculum preparation for each year's incoming instructors. I also successfully migrated the organization from a Java-based curriculum to a Python-oriented one and developed the lesson plans for the transition.

Fall 2010 — *Teaching Assistant, Database Systems (MIT 6.830)*
I assisted in writing and grading the assignments and projects.

Summer 2010 — *Instructor, MEET*
I mentored a group of 30 Israeli and Palestinian high school students as part of the MIT MEET program, a peace initiative in the Middle East centered around teaching computer science.

Spring 2010
Spring 2011 — *Instructor, Introduction to Java Course (MIT 6.S092)*
I instructed a class of 50 students in an introduction to the Java programming language. MIT does not have such an introductory course, so this course is taken by many MIT undergraduates to prepare them for 6.004, a core course that assumes proficiency in Java. The course is freely available online at `http://bit.ly/alvK9m`

Fall 2006 — *Teaching Assistant, Database Systems (UCB CS186)*
I taught approximately 30 students in weekly discussion sections. I assisted in writing and grading the assignments and projects.

PERSONAL

I love drawing and designing T-shirts and posters. I have created over 20 designs that have been printed and my shirts have been worn by thousands of people. The following link lists some of my designs.
`http://www.mit.edu/~eugenewu/gallery.html`

PUBLICATIONS

[1] M. J. Cafarella, A. Halevy, D. Z. Wang, E. Wu, and Y. Zhang. Webtables: exploring the power of tables on the web. *VLDB*, 1(1):538–549, 2008.

[2] M. J. Cafarella, A. Y. Halevy, Y. Zhang, D. Z. Wang, and E. Wu. Uncovering the relational web. *WebDB*, 2008.

[3] A. Cheung, L. Ravindranath, E. Wu, S. Madden, and H. Balakrishnan. Mobile applications need targeted micro-updates. In *APSys*, page 8. ACM, 2013.

[4] O. Cooper, A. Edakkunni, M. J. Franklin, W. Hong, S. R. Jeffery, S. Krishnamurthy, F. Reiss, S. Rizvi, and E. Wu. Hifi: A unified architecture for high fan-in systems. *VLDB*, pages 1357–1360, 2004.

[5] P. Cudre-Mauroux, E. Wu, and S. Madden. The case for rodentstore, an adaptive, declarative storage system. *CIDR*, 2009.

[6] P. Cudre-Mauroux, E. Wu, and S. Madden. Trajstore: An adaptive storage system for very large trajectory data sets. In *ICDE*, pages 109–120. IEEE, 2010.

[7] C. A. Curino, E. P. C. Jones, R. A. Popa, N. Malviya, E. Wu, S. R. Madden, H. Balakrishnan, N. Zeldovich, et al. Relational cloud: A database-as-a-service for the cloud. 2011.

[8] M. J. Franklin, S. R. Jeffery, S. Krishnamurthy, F. Reiss, S. Rizvi, E. Wu, O. Cooper, A. Edakkunni, and W. Hong. *Design considerations for high fan-in systems: The HiFi approach*, volume 5. CIDR, 2005.

[9] M. N. Garofalakis, K. P. Brown, M. J. Franklin, J. M. Hellerstein, D. Z. Wang, E. Michelakis, L. Tancau, E. Wu, S. R. Jeffery, and R. Aipperspach. Probabilistic data management for pervasive computing: The data furnace project. *IEEE Data Eng. Bull*, 29(1):57–63, 2006.

[10] S. Madden, E. Wu, S. Madden, Y. Zhang, E. Jones, C. Curino, et al. Relational cloud: The case for a database service. 2010.

[11] A. Marcus, E. Wu, D. Karger, S. Madden, and R. Miller. Human-powered sorts and joins. *VLDB*, 5(1):13–24, 2011.

[12] A. Marcus, E. Wu, D. Karger, S. Madden, and R. C. Miller. Demonstration of qurk: a query processor for human operators. In *SIGMOD*, pages 1315–1318. ACM, 2011.

[13] A. Marcus, E. Wu, D. R. Karger, S. R. Madden, R. C. Miller, et al. Crowdsourced databases: Query processing with people. 2011.

[14] A. Marcus, E. Wu, and S. Madden. Data in context: Aiding news consumers while taming dataspaces. *DBCrowd*, page 47, 2013.

[15] E. Wu. *Shinobi: Insert-aware partitioning and indexing techniques for skewed database workloads*. PhD thesis, Massachusetts Institute of Technology, Department of Electrical Engineering and Computer Science, 2010.

[16] E. Wu, P. Cudre-Mauroux, and S. Madden. Demonstration of the trajstore system. *VLDB*, 2(2):1554–1557, 2009.

[17] E. Wu, C. A. Curino, S. R. Madden, et al. No bits left behind. *CIDR*, 2011.

[18] E. Wu, Y. Diao, and S. Rizvi. High-performance complex event processing over streams. In *SIGMOD*, pages 407–418. ACM, 2006.

[19] E. Wu and S. Madden. Partitioning techniques for fine-grained indexing. In *ICDE*, pages 1127–1138. IEEE, 2011.

[20] E. Wu and S. Madden. Scorpion: Explaining away outliers in aggregate queries. *VLDB*, 2013.

[21] E. Wu, S. Madden, and M. Stonebraker. A demonstration of dbwipes: clean as you query. *VLDB*, 5(12):1894–1897, 2012.

[22] E. Wu, S. Madden, and M. Stonebraker. Subzero: a fine-grained lineage system for scientific databases. *ICDE*, 2013.