Eugene Wu **ewu@cs.columbia.edu**

500 West 120 St, Room 421, NY, NY 10027

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

Winter 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

2015– **Columbia University, NY, NY**
Assistant Professor – Computer Science

2015 **UC Berkeley, Berkeley, CA**
Visiting Scholar – AMPLab

2008–2014 **Massachusetts Institute of Technology, Cambridge, MA**
Ph.D. Student – CSAIL

CURRENT PROJECTS

Explaining Machine Learning Models
Machine learning models are increasingly used in critial real-world applications such as self-driving cars, loan processing, fake news detection, and more. However these models are highly complex and have a reputation for being black boxes when they make a prediction, it is unclear how the decision was made. Similarly, it is not clear what the model is using to make a prediction and how changes in the data would affect its predictions.
To this end, our lab develops algorithms to interpret complex machine learning models (e.g., deep neural networks, random forests, etc) by identifying training data that affected a prediction, describing what parts of the model are learning, and how user generated inputs can be improved to better help the model.
`https://cudbg.github.io/lab/mlexplain`

Data Visualization Management Systems
A Data Visualization Management System (DVMS) integrates visualizations and databases, by compiling a declarative visualization language into an end-to-end relational operator pipeline that renders the visualization and is amenable to database-style optimizations. Thus the DVMS can be both expressive via the visualization language, and performant by leveraging traditional and visualization-specific optimizations to scale interactive visualizations to massive datasets.
`https://cudbg.github.io/lab/dvms`

Perceptual Functions far Data Visualization
Increasing data sizes has made it more difficult to build highly responsive interactive visualization tools due to the enormous quantity of input data and results that must be computed. Sampling-based approximation query processing is a promising research direction however over and under-sampling can easily lead to wasted resources or incorrect visualizations. We are modeling human perceptual limitations and using those models to automatically help visualization systems generate approximate but perceptually accurate visualizations.
`http://perceptvis.github.io/`

Data Cleaning for Machine Learning
Data cleaning is fundamentally challenging because there does not exist a pre-existing notion of correctness for the cleaned data. In addition, it is unclear whether data cleaning improves the downstream applications. For example, it is possible that data cleaning can make machine learning models worse than no cleaning at all. We are exploring semi and fully-automated data cleaning techniques for machine learning applications
`https://activeclean.github.io`

Data and Query Explanation
Data analysis is rarely a one-off linear process it requires performing analyses, and carefully studying and understanding the results. The latter process is particularly challenging because analysts lack tools to help understand why analysis results look strange, contain outliers, or have patterns that differ from their expectations. Our lab develops tools and algorithms to provide user-understandable explanations.
`https://cudbg.github.io/lab/dbexplain`

PAST PROJECTS AND JOBS

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.

AWARDS

2016	SIGMOD best demo award

GRANTS

2016	ACM SIGMOD Conference 2016: Student Activities and Travel Support
	IIS: Medium: Collaborative Research: Composing Interactive Data Visualizations
	Columbia Alliance: Perceptual Functions for Faster Interactive Visualizations
	REU: Development of Graphical Perception as a Service
2015	III: Small: Collaborative Research: Towards Interactive Data Visualization Management Systems

SERVICE

2017	ICDE Area Chair
	WWW PC
	SIGMOD Demo PC
	SIGMOD PC
	VLDB PC
	HILDA PC
	SSDBM PC
	HCOMP PC
2016	InfoVis Reviewer
	HILDA PC
	NEDBDay Co-Chair
	SIGMOD travel award committee
2015	SIGMOD travel award committee
2014	DATA4U PC

TEACHING EXPERIENCE

Spring 2016 *Instructor, Interactive Data Exploration Systems*
`http://columbiaviz.github.io`

Instructor, Computing Systems for Data Science
`http://w4121.github.io`

Fall 2016 *Instructor, Introduction to Databases*

Spring 2016 *Instructor, Big Data Systems*

Fall 2015 *Instructor, Introduction to Databases*

Fall 2013 *Instructor, From Ascii To Answers (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 *Instructor, Introduction to Java Course (MIT 6.S092)*
Spring 2011 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://eugenewu.net/gallery.html`

*

References

[1] Sanjay Krishnan and Eugene Wu. "PALM: Machine Learning Explanations For Iterative Debugging". In: *HILDA*. 2017.

[2] Hamed Nilforoshan, Jiannan Wang, and Eugene Wu. "PreCog: Improving Crowdsourced Data Quality Before Acquisition". In: *ArXiv*. 2017.

[3] Hamed Nilforoshan et al. "Dialectic: Enhancing Text Input Fields with Automatic Feedback to Improve Social Content Writing Quality". In: *arXiv preprint arXiv:1701.06718* (2017).

[4] Hamed Nilforoshan et al. "Segment-Predict-Explain for Automatic Writing Feedback". In: *Collective Intelligence*. 2017.

[5] Xiaolan Wang, Alexandra Meliou, and Eugene Wu. "QFix: Diagnosing errors through query histories". In: *SIGMOD* (2017).

[6] Eugene Wu. "CIDR: Chat-oriented Innovations in Database Research". In: *CIDR*. 2017.

[7] Eugene Wu et al. "Combining Design and Performance in a Data Visualization Management System". In: *CIDR*. 2017.

[8] Haoci Zhang, Thibault Sellam, and Eugene Wu. "Precision Interfaces". In: *HILDA*. 2017.

[9] Daniel Alabi and Eugene Wu. "PFunk-H: Approximate Query Processing using Perceptual Models". In: *HILDA* (2016).

[10] Sanjay Krishnan et al. "Activeclean: An interactive data cleaning framework for modern machine learning". In: *Proceedings of the 2016 International Conference on Management of Data*. ACM. 2016, pp. 2117–2120.

[11] Sanjay Krishnan et al. "ActiveClean: interactive data cleaning for statistical modeling". In: *Proceedings of the VLDB Endowment* 9.12 (2016), pp. 948–959.

[12] Sanjay Krishnan et al. "Towards Reliable Interactive Data Cleaning: A User Survey and Recommendations". In: *HILDA*. 2016.

[13] Liwen Sun et al. "Skipping-oriented partitioning for columnar layouts". In: *Proceedings of the VLDB Endowment* 10.4 (2016), pp. 421–432.

[14] Xiaolan Wang, Alexandra Meliou, and Eugene Wu. "QFix: Demonstrating error diagnosis in query histories". In: *SIGMOD* (2016).

[15] Eugene Wu et al. "Graphical Perception in Animated Bar Charts". In: *arXiv preprint arXiv:1604.00080* (2016).

[16] Yifan Wu, Joseph M Hellerstein, and Eugene Wu. "A DeVIL-ish Approach to Inconsistency in Interactive Visualizations". In: *HILDA* (2016).

[17] Anant Bhardwaj et al. "Collaborative data analytics with DataHub". In: *Proceedings of the VLDB Endowment* 8.12 (2015), pp. 1916–1919.

[18] Arka A Bhattacharya et al. "Automated metadata construction to support portable building applications". In: *Proceedings of the 2nd ACM International Conference on Embedded Systems for Energy-Efficient Built Environments*. ACM. 2015.

[19] Daniel Haas et al. "CLAMShell: speeding up crowds for low-latency data labeling". In: *VLDB* (2015).

[20] Daniel Haas et al. "Wisteria: Nurturing scalable data cleaning infrastructure". In: *Proceedings of the VLDB Endowment* 8.12 (2015), pp. 2004–2007.

[21] Sanjay Krishnan et al. "SampleClean: Fast and Reliable Analytics on Dirty Data". In: (2015).

[22] Eugene Wu. "Data Visualization Management Systems." In: *CIDR*. 2015.

[23] Eugene Wu et al. "Explaining data in visual analytic systems". PhD thesis. Massachusetts Institute of Technology, 2015.

[24] Leilani Battle et al. "Indexing Cost Sensitive Prediction". In: *arXiv preprint arXiv:1408.4072* (2014).

[25] Alekh Jindal et al. "Vertexica: your relational friend for graph analytics!" In: *Proceedings of the VLDB Endowment* 7.13 (2014), pp. 1669–1672.

[26] Eugene Wu, Leilani Battle, and Samuel R Madden. "The case for data visualization management systems: Vision paper". In: *Proceedings of the VLDB Endowment* 7.10 (2014), pp. 903–906.

[27] Alvin Cheung et al. "Mobile applications need targeted micro-updates". In: *Proceedings of the 4th Asia-Pacific Workshop on Systems.* ACM. 2013, p. 8.

[28] Adam Marcus, Eugene Wu, and Sam Madden. "Data In Context: Aiding News Consumers while Taming Dataspaces". In: *DBCrowd 2013* 47 (2013).

[29] Eugene Wu and Samuel Madden. "Scorpion: Explaining Away Outliers in Aggregate Queries". In: *VLDB* (2013).

[30] Eugene Wu, Steve Madden, and Michael Stonebraker. "Subzero: a fine-grained lineage system for scientific databases". In: *Data Engineering (ICDE), 2013 IEEE 29th International Conference on.* IEEE. 2013, pp. 865–876.

[31] Eugene Wu, Samuel Madden, and Michael Stonebraker. "A demonstration of DBWipes: clean as you query". In: *Proceedings of the VLDB Endowment* 5.12 (2012), pp. 1894–1897.

[32] Carlo Curino et al. "Relational cloud: A database-as-a-service for the cloud". In: (2011).

[33] Adam Marcus et al. "Crowdsourced databases: Query processing with people". In: (2011).

[34] Adam Marcus et al. "Demonstration of qurk: a query processor for humanoperators". In: *Proceedings of the 2011 ACM SIGMOD International Conference on Management of data.* ACM. 2011, pp. 1315–1318.

[35] Adam Marcus et al. "Human-powered sorts and joins". In: *Proceedings of the VLDB Endowment* 5.1 (2011), pp. 13–24.

[36] Adam Marcus et al. "Platform considerations in human computation". In: *Workshop on crowdsourcing and human computation* (2011).

[37] Eugene Wu, Carlo Curino, and Samuel Madden. "No bits left behind". In: (2011).

[38] Eugene Wu and Samuel Madden. "Partitioning techniques for fine-grained indexing". In: *Data Engineering (ICDE), 2011 IEEE 27th International Conference on.* IEEE. 2011, pp. 1127–1138.

[39] Philippe Cudre-Mauroux, Eugene Wu, and Samuel Madden. "Trajstore: An adaptive storage system for very large trajectory data sets". In: *Data Engineering (ICDE), 2010 IEEE 26th International Conference on.* IEEE. 2010, pp. 109–120.

[40] Sam Madden et al. "Relational Cloud: The Case for a Database Service". In: (2010).

[41] Eugene 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.

[42] Philippe Cudre-Mauroux, Eugene Wu, and Sam Madden. "The Case for RodentStore, an Adaptive, Declarative Storage System". In: *arXiv preprint arXiv:0909.1779* (2009).

[43] Eugene Wu, Philippe Cudre-Mauroux, and Samuel Madden. "Demonstration of the trajstore system". In: *Proceedings of the VLDB Endowment* 2.2 (2009), pp. 1554–1557.

[44] Michael J Cafarella et al. "Uncovering the relational web". In: *under review* (2008).

[45] Michael J Cafarella et al. "Webtables: exploring the power of tables on the web". In: *Proceedings of the VLDB Endowment* 1.1 (2008), pp. 538–549.

[46] Minos N Garofalakis et al. "Probabilistic Data Management for Pervasive Computing: The Data Furnace Project." In: *IEEE Data Eng. Bull.* 29.1 (2006), pp. 57–63.

[47] Daniel Gyllstrom et al. "SASE: Complex event processing over streams". In: *arXiv preprint cs/0612128* (2006).

[48] Eugene Wu, Yanlei Diao, and Shariq Rizvi. "High-performance complex event processing over streams". In: *Proceedings of the 2006 ACM SIGMOD international conference on Management of data*. ACM. 2006, pp. 407–418.

[49] Michael J Franklin et al. *Design considerations for high fan-in systems: The HiFi approach*. Vol. 5. CIDR, 2005.

[50] Owen Cooper et al. "Hifi: A unified architecture for high fan-in systems". In: *PROCEEDINGS OF THE INTERNATIONAL CONFERENCE ON VERY LARGE DATA BASES* (2004), pp. 1357–1360.