Eugene Wu **ewu@cs.columbia.edu**
500 West 120 St, Room 421, NY, NY 10027

RESEARCH INTERESTS

I am broadly interested in the design of data management systems, and the application of database concepts, to greatly extend the ability for domain-experts and normal users to work with data. This combines research in the areas of database optimization, human-data-interaction, data provenance, visualization, and interface design.

EDUCATION

Winter 2014 **Massachusetts Institute of Technology**, Cambridge, MA
Ph.D., Electrical Engineering and Computer Science
Advisor: Samuel Madden
Dissertation: Explaining Data in Visual Analytic Systems

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
Co-director – Center for Data, Media & Society, Columbia Data Science Institute
Co-advisor – Columbia Journalism & Computer Science Dual Degree Program

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

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

2007-2008 **Google Inc.**, Mountain View, CA
Research Intern – Data Management Group

Summer 2005 **Microsoft Inc.**, Redmond, WA
Engineering Intern

Spring 2005 **IBM Extreme Blue.**, Almaden, CA
Engineering Intern

AWARDS

2018 VLDB 10 Year Test-of-Time Award
Google Faculty Award
Amazon Faculty Award
2016 SIGMOD best demo award

GRANTS

2018 Google Faculty Award Grant
Amazon Faculty Award Grant
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

CURRENT PROJECTS

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 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`

Explaining Machine Learning Models
Deep learning and neural networks have completely transformed the landscape of decision making in companies and society, including systems management, image processing, recommendations, speech recognition, and more. Our increased reliance of these models demands a deep understanding of *how* neural networks make their decisions, yet they are widely viewed as black boxes. Existing approaches to understand the internal behavior and logic in a neural network requires significant experites and is largely manual. DeepBase is a scalable system to quickly inspect the internal behavior of neural networks at the scale of hundreds or thousands of models, with the same ease as querying a relational database. In addition, our lab also develops tools to explain and interpret other machine learning models such as random forests.
`https://cudbg.github.io/lab/mlexplain`

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://cudbg.github.io/lab/cleaning`

SERVICE

2019	SIGMOD PC
	SIGMOD Student Research Competition Co-chair
2018	ICDE PC
	SIGMOD PC
	HILDA Co-chair
	SIGMOD New Researcher Symposium Co-chair
	SIGMOD Publicity Co-chair
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 2018 *Instructor, Database Topics in Research & Practice*
`http://columbiadb.github.io`

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

Spring 2017 *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*
`http://w4111.github.io`

Spring 2016 *Instructor, Big Data Systems*
`http://w4121.github.io`

Fall 2015 *Instructor, Introduction to Databases*
`http://w4111.github.io`

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, stickers, 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] Michael Cafarella et al. "Ten Years of Web Tables". In: *pVLDB (Invited Paper)*. 2018.

[2] Hamed Nilforoshan and Eugene Wu. "Leveraging Quality Prediction Models for Automatic Writing Feedback". In: *ICWSM*. 2018.

[3] Fotis Psallidas and Eugene Wu. "Demonstration of Smoke: A Deep Breath of Data-Intensive Lineage Applications". In: *SIGMOD (demo)*. 2018.

[4] Fotis Psallidas and Eugene Wu. "Provenance in Interactive Visualizations". In: *HILDA*. 2018.

[5] Fotis Psallidas and Eugene Wu. "Smoke: Fine-grained Lineage at Interactive Speeds". In: *VLDB*. 2018.

[6] Gabriel Ryan et al. "At a Glance: Approximate Entropy as a Measure of Line Chart Visualization Complexity". In: *InfoVIS*. 2018.

[7] Thibault Sellam et al. "DeepBase: Deep Inspection of Neural Networks". In: *ArXiv*. 2018.

[8] Thibault Sellam et al. ""I Like the Way You Think!" Inspecting the Internal Logic of Recurrent Neural Networks". In: *SysML*. 2018.

[9] Pei Wang et al. "Deeper: A Data Enrichment System Powered by Deep Web". In: *SIGMOD (demo)*. 2018.

[10] HaoCi Zhang et al. "Precision Interfaces for Different Modalities". In: *SIGMOD (demo)*. 2018.

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

[12] Sanjay Krishnan et al. "AlphaClean: Data Cleaning With Distributed Search and Machine Learning". In: *arXiv*. 2017.

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

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

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

[16] Marianne Procopio et al. "Load-n-Go: Fast Approximate Join Visualizations That Improve Over Time". In: *DSIA*. 2017.

[17] Gabriel Ryan et al. "Approximate Entropy as a Measure of Line Chart Complexity". In: *InfoVIS Poster*. 2017.

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

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

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

[21] Yifan Wu et al. "Towards a Bayesian Model of Data Visualization Cognition". In: *DECISIVE*. 2017.

[22] Haoci Zhang, Thibault Sellam, and Eugene Wu. "Mining Precision Interfaces From Query Logs". In: *ArXiv*. 2017.

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

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

[25] Sanjay Krishnan et al. "Activeclean: An interactive data cleaning framework for modern machine learning". In: *SIGMOD*. 2016.

[26] Sanjay Krishnan et al. "ActiveClean: interactive data cleaning for statistical modeling". In: *PVLDB* (2016).

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

[28] Liwen Sun et al. "Skipping-oriented partitioning for columnar layouts". In: *PVLDB* (2016).

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

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

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

[32] Anant Bhardwaj et al. "Collaborative data analytics with DataHub". In: *PVLDB* (2015).

[33] Arka A Bhattacharya et al. "Automated metadata construction to support portable building applications". In: *BuildSys*. ACM. 2015.

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

[35] Daniel Haas et al. "Wisteria: Nurturing scalable data cleaning infrastructure". In: *PVLDB Demo* (2015).

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

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

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

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

[40] Alekh Jindal et al. "Vertexica: your relational friend for graph analytics!" In: *PVLDB* (2014).

[41] Eugene Wu, Leilani Battle, and Samuel R Madden. "The case for data visualization management systems: Vision paper". In: *PVLDB* (2014).

[42] Alvin Cheung et al. "Mobile applications need targeted micro-updates". In: *APSys*. 2013.

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

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

[45] Eugene Wu, Steve Madden, and Michael Stonebraker. "Subzero: a fine-grained lineage system for scientific databases". In: *ICDE*. 2013.

[46] Eugene Wu, Samuel Madden, and Michael Stonebraker. "A demonstration of DBWipes: clean as you query". In: *PVLDB* (2012).

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

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

[49] Adam Marcus et al. "Demonstration of qurk: a query processor for humanoperators". In: *SIGMOD*. 2011.

[50] Adam Marcus et al. "Human-powered sorts and joins". In: *PVLDB* (2011).

[51] Adam Marcus et al. "Platform considerations in human computation". In: *CSCW* (2011).

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

[53] Eugene Wu and Samuel Madden. "Partitioning techniques for fine-grained indexing". In: *ICDE*. 2011.

[54] Philippe Cudre-Mauroux, Eugene Wu, and Samuel Madden. "Trajstore: An adaptive storage system for very large trajectory data sets". In: *ICDE*. 2010.

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

[56] 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.

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

[58] Eugene Wu, Philippe Cudre-Mauroux, and Samuel Madden. "Demonstration of the trajstore system". In: *PVLDB* (2009).

[59] Michael J Cafarella et al. "Uncovering the relational web". In: *VLDB* (2008).

[60] Michael J Cafarella et al. "Webtables: exploring the power of tables on the web". In: *PVLDB* (2008).

[61] Minos N Garofalakis et al. "Probabilistic Data Management for Pervasive Computing: The Data Furnace Project." In: *IEEE Data Eng. Bull.* (2006).

[62] Daniel Gyllstrom et al. "SASE: Complex event processing over streams". In: *arXiv* (2006).

[63] Eugene Wu, Yanlei Diao, and Shariq Rizvi. "High-performance complex event processing over streams". In: *SIGMOD*. 2006.

[64] Michael J Franklin et al. "Design considerations for high fan-in systems: The HiFi approach". In: *CIDR* (2005).

[65] Owen Cooper et al. "Hifi: A unified architecture for high fan-in systems". In: *VLDB* (2004).