

## Avani P. Wildani née Gadani, Ph.D.

---

### CONTACT INFORMATION

470 Chelsea Ave NE  
Atlanta, GA, 30307

Voice: +1 909.437.8626  
E-mail: avani@mathcs.emory.edu

### SUMMARY

As a scientist and an engineer, I am fascinated with information: what it is, how it is stored, how it is accessed, and how it leads to decisions. My doctoral research focused on how patterns in disk accesses allow a system to predict what data is commonly accessed together along with how this knowledge of momentary grouping can be used to make systems more available and efficient.

My current interest is the dual of the storage problem. Whereas computer scientists have defined how to arrange storage to meet specific metrics such as fault tolerance and access speed, in neuroscience the metrics are observable but the system unknown. I am working to model information in the brain as a storage problem to better learn how we collect and interpret signals from our world.

### CURRENT RESEARCH

My lab uses machine-learning based prediction models to dynamically re-arrange data between storage devices for applications including power management, availability increase through fault isolation, and data compression. The focus is on exascale and cloud storage systems though we have applications in archival, high performance, and general enterprise storage.

#### *Workload Characterization*

To better apply our prediction models, we are also working to characterize and distill features from common workload types with a goal of building a characteristic database of workloads using learned features. Research from these projects has been published at ICDE, SYSTOR, PDSW, MASCOTS, HotDep, and SRDS.

#### *Software-Defined Networking for Wireless Mesh Networks*

We are exploring superlinearity in the scaling of fixed wireless mesh networks through the implementation of overlays also known as software-defined networks (SDNs). We've built an emulation environment through ns3 and plan to do a hardware deployment in the near future. Results from this work have been published in BDA.

#### *Imputation and Analysis for PM2.5 Concentration in Third-World Environments*

Combining multiple data sources with high proportions (50+%) of missing values, and building random forests and MCMC models to predict particulate content in the air with methods that are both interpretable and can leverage the domain expertise of our collaborators. Research from this project has been published in ISES and the Journal for Environmental Science & Technology.

#### *Computational Neuroscience*

There is a surprising synergy between the questions we ask as computer science researchers, such as What is the correct network configuration?, How should we balance power and performance?, or What features are most representative in this data?, and the questions we ask in neuroscience. My long-term goal in studying biological computation is to design models that explore familiar tradeoffs, such as power vs. reliability, in the context of our marvelously power efficient brains. This research will be bidirectional; ideally investigating the brain through systems methods will lead to a richer model of neural fault tolerance and computation while giving us power saving hints or a few extra nines of reliability as we build the complex cloud storage systems of the future.

### *Visual Classification*

Using topological classification to better understand the information network involved in biological vision. Current work involves characterizing populations of neurons in the primary visual cortex by deriving the persistent homology of the electrical spikes the neurons release for various stimuli. We use Python and Matlab for data analytics and simulations. Our preliminary results indicate the presence of distinct cell populations in the presence of natural stimuli. This project is funded by a T32 NIH research grant as well as partially published in ICML-Topology.

## EDUCATION

### **The Salk Institute of Biological Studies**

Postdoctoral Fellow (2013-2015)

Advisor: Tatyana Sharpee

### **University of California, Santa Cruz, CA USA**

Ph.D. Computer Science (September 2013)

Committee: Jiri Schindler, Ahmed Amer, Darrell D.E. Long, Ethan L. Miller (Advisor).

### **University of New Mexico, Albuquerque, NM USA**

M.S. Computer Science : Machine Learning (May 2007)

Advisor: Terran Lane

### **Harvey Mudd College, Claremont, CA USA**

B.S. Joint Math / Computer Science (May 2003)

Advisor: Mike Erlinger

## PROJECTS

### *Archival Economics*

We are studying the economics of storing data of unknown future value for an indeterminate amount of time. We simulate the costs, taking into account factors such as hardware, power, maintenance, and data migration, of different configurations of hardware and performance constraints for data stored and accessed over several human lifetimes. Our current focus is on building heterogeneous storage systems to capture the cost reductions in solid state devices and project future device configurations that would be ideally fit for a permanent archive. Research from this project has been published at MASCOTS and MSST.

### **Log Analysis**

Researched automatic classification of trouble tickets using information retrieval (variations of tf-idf) and statistical analysis. Also explored theoretical design for maximal log coverage for developers. Work published in SRDS and SLAML and submitted for a patent at IBM.

### **Reliability**

Researched the trade-offs between power consumption, performance, and availability when reconstructing archival data using irregular erasure codes. Also showed that adding catastrophic, "large-stripe" parity is an inexpensive way to build a 3-fault tolerant system. Work published in HotDep and MASCOTS.

### **MIND Institute**

This research involved using fMRI images to detect interconnections within the brains of both healthy and schizophrenic patients. From this, we worked on deriving an efficient technique to perform clustering and hidden variable **Bayesian Structure Search** analyses. I have also used **Support Vector Machines** to obtain a baseline for our other classification methods, and worked extensively with **Spectral Graph Clustering** methods to handle the patients' categorical data.

### SLC Project

Research Assistant in the proposal stage of a large scale learning project involving gene expression data from rodents. I modeled the data using **Bayesian Networks** after doing basic clustering for manual parameter selection.

### Sandia National Laboratories Clinic

**Program Manager** of a team working with Dr. Kevin Boyack at Sandia to cluster discrete data points and produce a visualization tool to help with large data-set analysis. I **implemented density clustering** and also researched validity metrics to integrate with the tool we wrote in MATLAB.

REFEREED  
PUBLICATIONS

Juncheng Yang, Reza Karimi, Trausti Sæmundsson, **Avani Wildani**, Ymir Vigfusson, MITHRIL: *Mining Sporadic Associations for Cache Prefetching*, ACM Symposium on Cloud Computing 2017 (SoCC '17), October 2017.

Xuefei Hu, Jessica Hartmann Belle, Xia Meng, **Avani Wildani**, Lance Waller, Matthew Strickland, Yang Liu, *Estimating PM<sub>2.5</sub> Concentrations in the Conterminous United States Using the Random Forest Approach*, Environmental Science & Technology, June 2017.

**Avani Wildani**, *Fighting for a Niche: An Evolutionary Model of Storage*, 9th USENIX Workshop on Hot Topics in Storage and File Systems (HotStor'17), July 2017.

Wenxuan Wang, **Avani Wildani**, *Snapshot Judgements: Obtaining Data Insights without Tracing*, 9th USENIX Workshop on Hot Topics in Storage and File Systems (HotStor'17), July 2017.

Preeti Gupta, Darrell D. E. Long, Ethan L. Miller, David S.H. Rosenthal, **Avani Wildani**, *Effects of Prolonged Media Usage and Long-term Planning on Archival Systems*, 32nd International Conference on Massive Storage Systems and Technologies (MSST2016), May 2016.

**Avani Wildani**, Ian Adams, *A Case for Rigorous Workload Classification*, 23rd IEEE International Symposium on Modeling, Analysis, and Simulation of Computer and Telecommunication Systems (MASCOTS 2015), October 2015.

**Avani Wildani**, Ethan L. Miller, *Can We Group Storage?: Statistical Techniques to Identify Predictive Groupings in Storage System Accesses*. ACM Transactions on Storage, September 2015.

**Avani Wildani**, Ethan L. Miller, Ian Adams, Darrell D. E. Long, *PERSES: Data Layout for Low Impact Failures*, 22th IEEE International Symposium on Modeling, Analysis, and Simulation of Computer and Telecommunication Systems (MASCOTS 2014), September 2014.

Preeti Gupta, **Avani Wildani**, Daniel Rosenthal, Ethan L. Miller, Ian Adams, Christina Strong, Andy Hospodor, *An Economic Perspective of Disk vs. Flash Media in Archival Storage*, Proceedings of the 22th IEEE International Symposium on Modeling, Analysis, and Simulation of Computer and Telecommunication Systems (MASCOTS 2014), September 2014.

**Avani Wildani**, Tatyana Sharpee, *Persistent Homology for Characterizing Stimuli Response in the Primary Visual Cortex*, Proceedings of the 31st International Conference on Machine Learning – Workshop on Topological Methods for Machine Learning. (ICML-Topology 2014), June 2014.

**Avani Wildani**, Ian Adams, Ethan L. Miller, *Single-Snapshot File System Analysis*, Proceedings of the 21st IEEE International Symposium on Modeling, Analysis, and Simulation of Computer and Telecommunication Systems (MASCOTS 2013), August 2013.

**Ian Adams, Mark W. Storer, Avani Wildani, Ethan L. Miller, Brian Madden**, *Validating Storage System Instrumentation*, **Proceedings of the 21st IEEE International Symposium on Modeling, Analysis, and Simulation of Computer and Telecommunication Systems (MASCOTS 2013)**, August 2013.

**Aleatha Parker-Wood, Brian Madden, Michael McThrow, Darrell D. E. Long, Ian Adams, Avani Wildani**, *Examining Extended and Scientific Metadata for Scalable Index Designs*, **Proceedings of the 6th Annual International Systems and Storage Conference (SYSTOR 2013)**, June 2013.

**Avani Wildani, Ethan L. Miller, Ohad Rodeh**, *HANDS: A Heuristically Arranged Non-Backup In-line Deduplication System*, **Proceedings of the 29th IEEE International Conference on Data Engineering (ICDE 2013)**, April 2013.

**Avani Wildani, Lee Ward, Ethan L. Miller**, *Efficiently Identifying Working Sets in Block I/O Streams*, **Proceedings of the 4th Annual International Systems and Storage Conference (SYSTOR 2011)**, May 2011.

**Avani Wildani, Ethan L. Miller**, *Semantic Data Placement for Power Management in Archival Storage*, **Proceedings of the 5th International Workshop on Petascale Data Storage (PDSW 2010)**, held in conjunction with SC2010, November 2010.

**Wendy Belluomini, Binny Gill, Avani Wildani, Pin Zhou**, *GAUL: Gestalt Analysis of Unstructured Logs for Diagnosing Recurring Problem in Large Enterprise Storage Systems*, **29th IEEE International Symposium on Reliable Distributed Systems (SRDS 2010)**, November 2010.

**Ari Rabkin, Wei Xu, Avani Wildani, Armando Fox, Dave Patterson, Randy Katz**, *A Graphical Representation for Identifier Structure in Logs*, **Workshop on Managing Systems via Log Analysis and Machine Learning Techniques (SLAML 2010)**, October 2010.

**Avani Wildani, Thomas Schwarz, Ethan L. Miller, Darrell D. E. Long**, *Protecting Against Rare Event Failures in Archival Systems*, **Proceedings of the 17th IEEE International Symposium on Modeling, Analysis, and Simulation of Computer and Telecommunication Systems (MASCOTS 2009)**, September 2009.

**Kevin Greenan, Darrell D. E. Long, Ethan L. Miller, Thomas Schwarz, Avani Wildani**, *Building Flexible, Fault-Tolerant Flash-based Storage Systems*, **Proceedings of the Fifth Workshop on Hot Topics in System Dependability (HotDep 2009)**, June 2009.

#### OTHER PUBLICATIONS

**Avani Wildani**, *The Promise of Data Grouping in Large Scale Systems*, Ph.D. Thesis. September 2013.

**Avani Wildani**, *FAST 2011 Work-In-Progress Summaries*, USENIX ;login: Magazine, 2011.

**Avani Wildani**, *FAST 2010 Work-In-Progress Summaries*, USENIX ;login: Magazine, 2010.

**Avani Gadani, Eric Wu, Daniel Lowd, Brian Roney, Belinda Thom**, *Implementation of and Experimentation with a Clustering Tool*, Harvey Mudd Clinic Final Report, June 2003.

#### ADDITIONAL TALKS

*Skeletons in Your Data: Why and How Should We Characterize Storage Workloads?*, Invited Talk: IBM Almaden Research Center, October 2017.

*New Interfaces in Neural Computing*, Keynote presentation, 2017 ACM Richard Tapia Celebration of Diversity in Computing, September 2017.

*Skeletons in Your Data: Why and How Should We Characterize Storage Workloads?*, Invited Talk: Intel Labs, August 2017.

*Towards music process understanding through deep neural network song feature identification in mice and song birds*, Invited Talk: Pandora December 2016.

*Feedback Mechanisms in Neural Network Validation*, Invited Talk: University of Reykjavik. June 2016.

*Can We Group Storage?*, Invited Talk: Seagate. February 2015.

*Topological Characterization of Visual Invariants*, Salk Science Retreat: Lightning Talk. October 2014.

*Emerging Topics in Computational Biology: Beyond Gene Sequencing*, Technical Talk (Data Science Track): Grace Hopper Celebration of Women in Computing. October 2014.

*Data Grouping and Vision*, Invited Talk: Vicarious. September 2014.

*The Promise of Data Grouping*, Invited Talk: IBM Research. August 2013.

PERSES, Invited Talk: NetApp. November 2012.

REFEREED POSTER  
PRESENTATIONS

Juncheng Yang, Reza Karimi, **Avani Wildani**, Ymir Vigfusson, *A Simple Cache Prefetching Layer Based on Block Correlation*, FAST 2017 Work-in-Progress. (Talk + Poster)

J. Bi, B. Vu, **A. Wildani**, Y. Wang, A. Lyapustin, Y. Liu, *Citywide Validation and Improvement of the MAIAC Aerosol Product in Lima, Peru*, 27th meeting of the International Society of Exposure Science (ISES 2017), Oct 2017.

Clarissa Tuxen, Sérgio Gramacho, **Avani Wildani**, *Scaling in Socially-Driven Computer Networks*, Grace Hopper Celebration (GHC 2016), Oct 2016. **Best Undergraduate Poster**

Clarissa Tuxen, Sérgio Gramacho, **Avani Wildani**, *Socially Driven Computer Networks*, Biological Distributed Algorithms (BDA 2016), Aug 2016.

**Avani Wildani**, Ian F. Adams, *A Case for Rigorous Workload Classification*, FAST 2015 Work-in-Progress.

Preeti Gupta, **Avani Wildani**, Ethan L. Miller, Darrell D.E. Long, David S.H. Rosenthal, *Which Media Should Be Used In Long-term*, FAST 2015.

**Avani Wildani**, Tatyana Sharpee, *Characterizing Vision with Persistent Homology*, Joint Symposium on Neural Computation, May 2014.

Zhike Zhang, Preeti Gupta, **Avani Wildani**, Ignacio Corderi, and Darrell D.E. Long, *Reverse Deduplication: Optimizing for Fast Restore*, FAST 2013 Work-in-Progress. (Talk + Poster)

Ian F. Adams, Ethan L. Miller, Mark W. Storer, **Avani Wildani**, and Yangwook Kang, *Improved Analysis and Trace Validation Using Metadata Snapshots*, FAST 2013.

Yan Li, Christina Strong, Ignacio Corderi, **Avani Wildani**, Aleatha Parker-Wood, Andy Hospodor, Thomas M. Kroeger, and Darrell D. E. Long, *Energy-Aware Storage*. FAST 2013.

**Avani Wildani**, Ethan L. Miller, *Grouping Data for Faster Rebuilds: The Art of Failing Silently*, FAST 2012 Work-in-Progress. (Talk+Poster)

**Avani Wildani**, Ethan L. Miller, *Probabilistic Reputation for Personal Trust Networks*, FAST 2009 Work-in-Progress. (Talk+Poster)

WORK  
EXPERIENCE

**Emory University:** *Assistant Professor, Math and Computer Science*  
See Current Research for research details.

July 2015 - present

**The Salk Institute:** *Postdoctoral Research Associate*

August 2013 - July 2015

Postdoctoral training in the CNL-T lab under Dr. Tatyana Sharpee. See Current Research for project details.

**IBM Almaden Research Center:** *Research Intern*

Summer 2011 (continuing)

Worked with a small team addressing the disk bottleneck problem in data de-duplication for primary storage by using my thesis work in data grouping. Involved analyzing traces of 100GB+ in short amounts of time as well as writing Python libraries to simulate de-duplication. This work appeared in ICDE 2013.

**Sandia Labs - Computer Science Research Institute:** *Research Intern*

Summer 2010

Analyzed block I/O data from a multi-use production system in order to define and identify similarity between particular data blocks. Looked into creating access groupings based on the similarity data. Research on similarity was published in SYSTOR 2011.

**Google:** *Intern*

Summer 2009

Implemented a PID-based control system in Python to automatically balance customers for Bigtable (a key/value store). Extensively researched reward functions and SLA/SLO management techniques and wrote out a detailed experiment plan to investigate the tradeoffs between economic and standard scheduling techniques. Worked with the Census team to expand their SLO framework to include SLAs.

**IBM Almaden Research Center:** *Research Intern*

Summer 2008 (continued through fall)

Worked with a team of 4 researchers to develop a highly accurate method to automatically classify log excerpts from commercially deployed storage systems to speed up problem isolation. This involved coding in Java, particularly the Lucene search engine, and Python for the backend and data-manipulation respectively. I also did a significant portion of the system design and analysis. This work was published in SRDS 2009.

**Google:** *Intern*

Summer 2006, Summer 2007

Documented Internal Build System Components using HTML/CSS and Wiki. Studied and documented Code Dependencies in Python, UNIX Shell, and C++ scripts. Used Python to implement statistical techniques to analyze web data. **CENIC:** *Systems Administrator* Oct 2003 - May 2004 Configured and supported a variety of **internal solutions** for CENIC's network management. This included limited exposure to **Cisco optical equipment** (15808 and 15540), and the deployment of a complete Cisco VoIP solution for the office. Also **administered 5-7 OS X and Solaris servers** including mail, news, Request Tracker, and Ciscoworks.

**University of California, Santa Cruz, CA USA:**

*Research Assistant*

March 2008 - August 2013

**Storage Systems** research under Dr. Ethan Miller. (see Projects)

**University of New Mexico, Albuquerque, NM USA:** *Research Assistant*

Jan 2005 - May 2007

**Machine Learning** research under Dr. Terran Lane. (see Projects)

**Harvey Mudd College, Claremont, CA USA** *Research Assistant*

May 2002 - Jan 2008

IDXP/BEEP **Intrusion Detection** project at Harvey Mudd College, with Prof. Mike Erlinger, Head of the IDWG. Over Summer 2002, we **implemented a functional IDS** using Perl and Java with help from Aerospace and Silicon Defense. We also monitored traffic using a modification of the snort tool and NFR and sent these alerts into a **MySQL database**.

## TEACHING

### *Instructor of Record*

**Intermediate Java Programming.** Object-oriented concepts, data structures, and basic algorithms for sophomore computer science majors.

**Algorithms.** Theory of algorithms, dynamic programming, graph algorithms, and basic theory of computation for junior and senior computer science majors.

**Advanced Computer Systems (Graduate)** Covered major topics in distributed systems including virtualization, concurrency, security, and reliability, along with some current research in field.

**Neural Computing (Graduate)** Covered current research in neural computing and relationships between biological, computational, and thermodynamic neural models.

### *Teaching Assistant*

Lead TA for **Introduction to Programming in Python.** Wrote most course material including lecture slides and assignments, taught labs, and coordinated other TAs and graders. 72% of reviews "above average."

### *Teaching Assistant*

Lead TA for **Introduction to Java Programming.** Wrote and graded assignments, Taught Labs, Managed and coordinated with 5 TAs, and maintained the class web site.

*Grader/Tutor (Harvey Mudd College)* Graded assignments and held tutoring sessions for **Algorithms, Discrete Mathematics, Linear Algebra, Differential Equations, Architecture and Operating Systems, and Introduction to Computer Science for Majors.**

## SERVICE

### *Conference Co-chair:*

SAC 2016; Grace Hopper Celebration of Women in Computer Science 2016, 2017 (Systems Track)

### *PC Member:*

NAS 2015, Storage Track; NVMSA 2015, 2016; Grace Hopper 2015, Data Science Track, MSST 2017

### *Reviewer:*

Mass Storage 2013, ACM Transactions on Storage, IEEE Transactions on Services Computing, PLoS Computational Biology

### *Session Chair:*

MASCOTS 2013, 2014, 2015, MSST 2017

### *SciChat Lecturer:*

Teaching science to middle school students, 2014 (Winner of "Scientist Idol" award for best lecture)

### *WitsOn Mentor:*

Mentoring women in science (Piazza group)

## GRANTS AND AWARDS

International Computation in Modeling, Meritorious

National Science Foundation CSEMS Initiative Computer Science and Mathematics (CSAM) Scholarship

Grace Hopper Scholarship (2009, 2014)

File and Storage Technologies (FAST) Student Grant Recipient (2010, 2012)

CRA-W Workshop Fellowship

Google FAST Grant

NIH T32 Training Grant (Trainee; PI: Dr. Tom Albright: \$50,000 for training in vision research)

Pioneer Postdoctoral Fellowship (\$31,250)

Emory PERS Grant (\$5,000)

Various small grants from Google, Microsoft, and NVIDIA

## TOOLS

**Programming and Markup Languages** I've used for large projects:

Python (including NumPy/Numeric, SciPy, and SimPy), C++, Java, Prolog, MATLAB, L<sup>A</sup>T<sub>E</sub>X,  
Shell Scripting, XML, HTML

**Programming Languages** I have used for smaller projects:

R, C, SML, Lisp, Ciao, Rex, ns2 simulator, Tcl

**Software and Operating Systems** I can use proficiently:

UNIX Utilities, MATLAB, Weka  
Solaris, Mac OS X / BSD, Windows, Linux (Ubuntu, Redhat, Debian, etc.)