

# Daisy Zhe Wang

Electrical Engineering and Computer Sciences  
University of California, Berkeley  
RAD Lab, 465 Soda Hall #1776  
Berkeley, CA 94720-1776  
(510) 734-1663  
[daisyw@cs.berkeley.edu](mailto:daisyw@cs.berkeley.edu)  
[www.eecs.berkeley.edu/~daisyw/](http://www.eecs.berkeley.edu/~daisyw/)

## Research Interests

---

Primary Interests: Probabilistic Data Management, Scalable Data Analysis, Statistical Machine Learning.  
General Interests: Large-scale Data Management Systems, Parallel and Cloud Computing, Data Mining.

## Education

---

*Ph.D. Electrical Engineering and Computer Sciences* *Expected May, 2011*  
University of California, Berkeley, Database Research Group.  
Advisors: Michael J. Franklin, Joseph M. Hellerstein, and Minos N. Garofalakis  
Thesis Topic: Probabilistic Data Management based on Graphical Models.

*B.A.Sc. Computer Engineering* *May, 2005*  
University of Toronto, Canada.  
Advisor: Hans-Arno Jacobsen

## Awards

---

UC Berkeley, Department of Electrical Engineering and Computer Sciences  
Stonebraker/Wong Fellowship 2009,  
Departmental Fellowship 2005.  
University of Toronto, Department of Electrical and Computer Engineering  
McAllister Research Award 2004,  
University of Toronto Scholar 2002, 2004,  
Dean's Honor List 2002-2004,  
Admission Scholarship 2002.

## Academic Experience

---

*Fall 2005-present* *Research Assistant, Database Group, UC Berkeley*  
Proposed, developed and evaluated BayesStore, a probabilistic database system for statistical machine learning models. Devised efficient ways to natively support graphical models and their inference algorithms. Invented algorithms to perform probabilistic querying and analysis over large-scale uncertain data and the associated models. Designed query execution strategies that optimize across relational and inference operators.

*Summer 2003-Spring 2005 Undergraduate Researcher, Middleware Systems Research Group, Uof T*  
Designed and developed a demo for "Scalable Location-Aware Publish Subscribe System" and presented it at IBM CASCON 2004 conference. Designed, implemented and evaluated algorithms for large-scale XPath and regular expression matching.

## **Industrial Experience**

---

*Summer 2008 Research Internship, Yahoo! Research. Santa Clara*  
Investigated techniques to compress and visualize lineage in the debug/feedback phase of PSOX, a web-scale information extraction (IE) system. Significantly improved data bandwidth of IE pipelines that involve both machine learning and database operations.

*Spring 2008 – Spring 2010 Research Collaborator, IBM Almaden Research Center*  
Developed and evaluated estimators for the cost and the output size of text extractors, such as dictionaries and regular expressions. Designed and evaluated different document synopses for more accurate estimation of various statistics over text corpora.

*Summer 2007 Software Engineer, Google Inc.*  
Explored the complex problem of scalable extraction of the metadata of the HTML tables from the entire Web. Developed statistical classifiers and rule-based detectors, which recovered millions of schemas from HTML tables.

*Summer 2006 Research Internship, Intel Research Berkeley*  
Designed and formalized the data model and developed the algorithms for relational operators over probabilistic data. This work formed the core of the BayesStore system.

*Summer 2003, 2004, 2005 Software Developer, DB2 UDB Compiler Group, IBM Toronto Lab*  
Worked with large and complex codebase of DB2 on AIX Unix. Developed and tested new functionalities for materialized tables. The code of this line item was shipped with DB2 8.4. Developed components for dbtop – a performance monitoring utility in IBM DB2.

## **Publications**

---

### **Refereed Conferences**

*C7. Hybrid In-Database Inference for Declarative Information Extraction*

**Daisy Zhe Wang**, Michael J. Franklin, Minos Garofalakis, and Joseph M. Hellerstein, Michael L. Wick

*To Appear, Proceedings of ACM SIGMOD International Conference on Management of Data, 2011.*

*C6. Selectivity Estimation for Extraction Operators over Text Data*

**Daisy Zhe Wang**, Long Wei, Yunyao Li, Frederick Reiss, and Shivakumar Vaithyanathan

*To Appear, Proceedings of 27<sup>th</sup> IEEE International Conference on Data Engineering (ICDE), 2011.*  
(Acceptance Rate: 19.8%)

C5. *Querying Probabilistic Information Extraction.*

**Daisy Zhe Wang**, Michael J. Franklin, Minos Garofalakis, and Joseph M. Hellerstein  
*Proceedings of 36<sup>th</sup> Very Large Data Base Endowment (VLDB), 2010, Vol.3: p1057-1067. (Acceptance Rate: 25%)*

C4. *Probabilistic Declarative Information Extraction.*

**Daisy Zhe Wang**, Eirinaios Michelakis, Michael J. Franklin, Minos Garofalakis, and Joseph M. Hellerstein  
*Proceedings of 26<sup>th</sup> IEEE International Conference on Data Engineering (ICDE), 2010 short paper: p173-176. (Acceptance Rate: 20%)*

C3. *BayesStore: Managing Large, Uncertain Data Repositories with Probabilistic Graphical Models.*

**Daisy Zhe Wang**, Eirinaios Michelakis, Minos N. Garofalakis, Joseph M. Hellerstein.  
*Proceedings of 34<sup>th</sup> Very Large Data Base Endowment (VLDB), 2008, Vol1: p340-351. (Acceptance rate: 16.5%)*

C2. *WebTables: Exploring the Power of Tables on the Web.*

Michael J. Cafarella, Alon Halevy, **Daisy Zhe Wang**, Eugene Wu, Yang Zhang.  
*Proceedings of 34<sup>th</sup> Very Large Data Base Endowment (VLDB), 2008, Vol1: p538-549. (Acceptance rate: 16.5%)*

C1. *Bonsai: Exploration and Cultivation of Machine Learning Models*

David Purdy, **Daisy Zhe Wang**  
*Proceedings of Joint Statistical Meetings (JSM), 2008.*

### **Refereed Workshops**

W3. *Functional Dependency Generation and Applications in Pay-as-you-go Data Integration Systems.*

**Daisy Zhe Wang**, Luna Dong, Anish Das Sarma, Michael J. Franklin, Alon Halevy  
*Proceedings of the ACM SIGMOD WebDB, 2009.*

W2. *Uncovering the Relational Web.*

Michael Cafarella, Alon Halevy, Yang Zhang, **Daisy Zhe Wang**, Eugene Wu  
*Proceedings of the ACM SIGMOD WebDB, 2008.*

W1. *Granularity Conscious Modeling for Probabilistic Databases*

Eirinaios Michelakis, **Daisy Zhe Wang**, Minos N. Garofalakis, Joseph M. Hellerstein  
*Proceedings of ICDM DUNE, 2007: p501-506.*

### **In Submission, Non-Refereed and Technical Reports**

M3. *Probabilistic Data Management for Pervasive Computing: The Data Furnace Project.*

Minos N. Garofalakis, Kurt P. Brown, Michael J. Franklin, Joseph M. Hellerstein,  
**Daisy Zhe Wang**, Eirinaios Michelakis, Liviu Tancau, Eugene Wu, Shawn R. Jeffery, Ryan Aipperspach.  
*IEEE Data Engineering Bulletin, 29, No. 1: p57-63, 2006.*



## References

---

### **Michael J. Franklin**

Professor, Department of Electrical Engineering and Computer Sciences, Computer Science Division  
University of California, Berkeley  
Soda Hall #1776  
Berkeley, CA 94720-1776  
(510) 642-1662  
franklin@cs.berkeley.edu

### **Joseph M. Hellerstein**

Professor, Department of Electrical Engineering and Computer Sciences, Computer Science Division  
University of California, Berkeley  
Soda Hall #1776  
Berkeley, CA 94720-1776  
(510) 643-4011  
hellerstein@cs.berkeley.edu

### **Minos Garofalakis**

Professor, Department of Electronic & Computer Engineering  
Technical University of Crete  
University Campus -- Kounoupidiana  
73100 Chania, Hellas (Greece)  
+30-28210-37211  
minos@acm.org

### **Alon Halevy**

Head, Structured Data Research Group  
Google Inc.  
1600 Amphitheatre Parkway  
Mountain View, CA, 94043  
(650) 253-2574  
halevy@google.com

### **Renee J. Miller**

Professor, Department of Computer Science  
University of Toronto  
Bahen Center for Information Technology  
40 St. George Street, Room BA7270  
Toronto ON M5S 2E4  
(416) 946-3621  
miller@cs.toronto.edu