OMB No. 0925-0001 and 0925-0002 (Rev. 10/15 Approved Through 10/31/2018)

BIOGRAPHICAL SKETCH

Provide the following information for the Senior/key personnel and other significant contributors.  
Follow this format for each person. **DO NOT EXCEED FIVE PAGES.**

NAME: Bill Howe

eRA COMMONS USER NAME (credential, e.g., agency login): N/A

POSITION TITLE: Associate Director and Senior Data Science Fellow, UW eScience Institute  
Affiliate Associate Professor, Computer Science & Engineering

EDUCATION/TRAINING (Begin with baccalaureate or other initial professional education, such as nursing, include postdoctoral training and residency training if applicable. Add/delete rows as necessary.)

| INSTITUTION AND LOCATION | DEGREE  (if applicable) | Completion Date  MM/YYYY | FIELD OF STUDY |
| --- | --- | --- | --- |
| Georgia Institute of Technology, School of Industrial and Systems Engineering, Atlanta, GA | BS | 03/1999 | Industrial and Systems Engineering |
| Portland State University, Department of Computer Science, Portland, OR | Phd | 2007 | Computer Science |

# A. Personal Statement

I am uniquely positioned to assist this project’s research goals. I am the Associate Director and Senior Data Science Fellow at the UW eScience Institute and Affiiate Associate Professor in Computer Science & Engineering. I am a co-founder of Urban@UW, and with support from the MacArthur Foundation and Microsoft, I lead UW's participation in the MetroLab Network. I created a first MOOC on Data Science through Coursera that has attracted over 200,000 students, and I led the creation of the UW Data Science Masters Degree where I serve as its first Program Director and Faculty Chair. I am a member of the UW Graduate Faculty. My group's research aims to make the techniques and technologies of data science more accessible, particularly at scale. Our methods are rooted in database models and languages, though we sometimes work in machine learning, visualization, HCI, and high-performance computing. We are an applied, systems-oriented group, frequently sourcing projects through collaborations in the physical, life, and social sciences. I will add considerable breadth of experience to the computational core of this work.

# B. Positions and Honors

Positions and Employment

|  |  |
| --- | --- |
| 1999-2000 | Systems Analyst, Project for EXCEL Communications with Deloitte Consulting, Atlanta, GA |
| 2000 | Systems Analyst, Project for Verizon Wireless with Deloitte Consulting, Seattle, WA |
| 2000 | Lead Developer, Schlumberger Oil Services with Siebel Systems Consulting, Seattle, WA |
| 2000 – 2001 | Technical Lead, Project for Microsoft Consulting Services, Redmond, WA |
| 2001 – 2006 | Graduate Research Assistant, Oregon Health and Science University, Beaverton OR |
| 2006 – 2009 | Senior Research Associate Oregon Health and Science University, Beaverton OR |
| 2009–2013 | Senior Research Scientist, University of Washington, Seattle, WA |
| 2009–present | Affiliate Faculty in Computer Science & Engineering, University of Washington, Seattle, WA |
| 2013–present | Principal Research Scientist, University of Washington, Seattle, WA |
| 2013–present  2015–present | Associate Director and Senior Data Science Fellow, eScience Institute, Seattle, WA  Program Director and Faculty Lead, Data Science Masters Degree, University of Washington, Seattle, WA |

Other Experience and Professional Memberships

|  |  |
| --- | --- |
| 2006 | Teaching Portland State University |
| 2010-Present | Teaching University of Washington |
| 2012 | Program Committee, ScienceCloud |
| 2012 | Demo Co-chair, SSDBM 2012 |
| 2012 | Chair, Workshop on HPC meets Databases, co-located with Supercomputing |
| 2012-2013 | Program Committee, PVLDB |
| 2013 | Program Committee, ICDE |
| 2013 | Program Committee, LDAV |
| 2013 | Program Committee, SSDBM |
| 2014 | Program Committee, eScience |
| 2014 | Program Committee, HotCloud |
| 2014 | Program Chair, eScience |

Honors

|  |  |
| --- | --- |
| 1999 | Graduated with Honors, Georgia Tech |
| 2003 | Distinction in Research Proficiency Exam |
| 2004 | Top paper selected for publication in VLDB Journal, 14(4), Nov. 2005 |
| 2007 | Department Dissertation Award, Portland State University |
| 2010 | Best of VLDB 2010, selected for publication in VLDB Journal special issue |

# C. Contributions to Science

**Most relevant to the current application**

1. Bioinformatics is increasingly a data-driven science. Future success for life scientists will depend upon the ability to leverage the large-scale data. By adopting the advances in information technology made by fields that have already faced the type of inflection point bioinformatics face, such as cloud computing, we believe bioinformatics can weave the computational environments that exist today into a solution for our data problems. We have developed systems and approaches to dramatically reduce the barrier to entry for large-scale data management and analytics, and provide greater access to the frontier of technology in computer science among researchers in the life, physical, and social sciences.

* Barga, Roger, **Bill Howe**, David Beck, Stuart Bowers, William Dobyns, Winston Haynes, Roger Higdon, Chris Howard, Christian Roth, Elizabeth Stewart and others. 2011. Bioinformatics and data-intensive scientific discovery in the beginning of the 21st century, Omics: A Journal of IntegrativeBiology, 15, 4 (2011), pages 199–201
* Bu, YingYi, **Bill Howe**, Magdalena Balazinska, and Michael Ernst. 2012. The HaLoop approach to large-scale iterative data analysis. Special Issue: Best of VLDB 2010 VLDB Journal, 21, 2 (2012).
* Alawini, Abdussalam, David Maier, Kristin Tufte, and Bill Howe. 2014. Helping scientists reconnect their datasets. Scientific and Statistical Database Management (SSDBM) (2014).
* Shrainik Jain, Dominik Moritz, **Bill Howe**, Ed Lazowska, SQLShare: Results from a Multi-Year SQL-as-a-Service Experiment. SIGMOD (2016)

1. We have done extensive work in systems, methods, and algorithms for large-scale analysis of biological and environmental data. In recent work, we evaluate Gaussian mixture model (GMM) classification algorithms commonly used for medical applications against large-scale, continuous environmental flow cytometry datasets by implementing them in Hadoop. Our approach outperforms current state-of-the-art cytometry classification algorithms in accuracy, and can be coupled with manual or automatic partitioning of data into homogeneous sections for further classification gains. We propose the GMM with partitioning approach for classification of large-scale, high-frequency flow cytometry data. In other recent work, we developed new algorithms for scalable graph clustering with application to protein-protein interaction networks.

* Hyrkas, Jeremy, Sophie Clayton, Francois Ribalet, Daniel Halperin, E. Virginia Armbrust, **Bill Howe**. Scalable clustering algorithms for continuous environmental flow cytometry Bioinformatics 32(3) (2016)
* Hyrkas, Jeremy, Daniel Halperin, and **Bill Howe**. 2014. Time-Varying Clusters in Large-Scale Flow Cytometry, Innovative Applications of Artificial Intelligence (2014)
* Seung-Hee Bae, Daniel Halperin, Jevin West, Martin Rosvall, **Bill Howe**. Scalable and Efficient Flow-Based Community Detection for Large-Scale Graph Analysis. Transactions on Knowledge Discovery from Data. (2016)
* Kwon, YongChul, Magdalena Balazinska, **Bill Howe**, and Jerome Rolia. 2010. Skew-resistant parallel processing of feature-extracting scientific userdefined functions. Proc. of the ACM Symposium on Cloud Computing (SOCC) (2010).

1. Before scientists can analyze, publish, or share their data, they often need to determine how their datasets are related. Determining relationships helps scientists identify the most complete version of a dataset, detect versions of datasets that complement each other, and determine multiple datasets that overlap. In previous work, we showed how observable relationships between two datasets help scientists recall their original derivation connection. While that work helped with identifying relationships between two datasets, it is infeasible for scientists to use it for finding relationships between all possible pairs in a large collection of datasets. In order to deal with larger numbers of datasets, we are extending our methodology with a relationship-prediction system, ReDiscover, a tool to identify pairs from a collection of datasets that are most likely related and the relationship between them. We report on the initial design of ReDiscover, which uses machine-learning methods such as Conditional Random Fields and Support Vector Machines to the relationship-discovery problem. Our preliminarily evaluation shows that ReDiscover predicted relationships with an average accuracy of 87%.

* Abdussalam Alawini, David Maier, Kristin Tufte, **Bill Howe**, Rashmi Nandikur. Towards automated prediction of relationships among scientific datasets. Scientific and Statistical Database Management (SSDBM) (2015)
* **Bill Howe**, David Maier, Laura Bright, Smoothing the ROI Curve for Scientific Data Management Applications. CIDR 2007

**Additional recent publications of importance to the field (in chronological order)**

* **Howe, Bill** David Maier, and Antonio Baptista. 2003. A language for spatial data manipulation. Journal of Environmental Informatics. 2, 2 (December 2003).
* **Howe, Bill** David Maier. “Algebraic Manipulation of Scientific Datasets,” International Conference on

Very Large Databases, 2004, Toronto, Canada. (VLDB 2004)

* **Howe**, **Bill** David Maier. “Smoothing the ROI Curve for Scientific Data Management Applications,” Third Biennial Conference on Innovative Data Systems Research, January 2007, Asilomar, CA.
* Vo, Huy, Brian Summa, Joao Comba, Juliana Freire, **Bill Howe**, Claudio Silva, and Valerio Pascucci. 2011. Parallel visualization on large clusters using mapReduce. IEEE Symposium on Large-Scale Data Analysis and Visualization, LDAV (2011)
* Balazinska, Magdalena, **Bill Howe**, and Dan Suciu, Data Markets in the Cloud: An Opportunity for the Database Community, Proceedings of the VLDB Endowment PVLDB, 4, 12, 2011. (Also in conference)
* Perry, Daniel, **Bill Howe**, Alicia M. F. Key, Cecilia Aragon. 2013. VizDeck: Streamlining Exploratory Visual Analytics of Scientific Data. iConference 2013 Proceedings (2013) pages 338–350. doi:10.9776/13206

# D. ****Ongoing Research Support****

2015-2018: Co-PI, NSF Big Data Hub, Western Region, $1,200,000, with Ed Lazowska (UW), Airel Rokem (UW), Mike Franklin (UC Berkeley), and Mike Norman (SDSC)

2015-2016: Lead PI, PNNL Subcontract, Unified Services for Federated Big Data Environment

2014-2016: Lead PI, Macarthur Foundation, Towards a Seattle Metro Consortium for Urban Data Science

2014-2019: Co-PI, Gordon and Betty Moore Foundation and the Alfred P. Sloan Foundation, Data Science Environments with NYU and UC Berkeley

2014-2016: Co-PI, Intel Science and Technology Center for Big Data, one part of large distributed group across MIT, Brown, UW, and Portland State

2013-2016: Co-PI, NSF Grant IIS-1247469, BIGDATA: A Formal Foundation for Big Data Management

**Completed Research Support (within last three years)**

2014-2015: Lead PI, PNNL Subcontract Hybrid Architectures for Graph Query

2013-2014: Lead PI, PNNL Subcontract Hybrid Languages for Graph Query

2012-2014: Co-PI, NSF Grant ACI-1216879, Collaborative Research: Conceptualizing An Institute for Empowering Long Tail Research

2011-2013: Lead PI, NSF Grant IIS-1064505, Collaborative Research: Database-As-A-Service for Long Tail Science

2012-2014: Co-PI, EMC Grant Large-Scale Visual Analytics

2011-2013: Co-PI, NSF Grant CCF-1047815, Relational Data Markets in the Cloud,

2010-2013: Co-Lead, Gordon and Betty Moore Foundation, Data-Intensive Science in the Long Tail