# Michael A. Schuh

*Research Statement*

## Overview

I thrive on research that is motivated by real-world problems discovered through interdisciplinary collaborations. Providing practical solutions to these problems tangibly affects the present world and paves the way for future advancement. A brighter future for the world is made possible through continued scientific progress, and I am eager to dedicate my career in Computer Science to these pursuits.

## Laboratory: Solar Big Data Mining

Much of my research revolves around information retrieval and applied data mining through interdisciplinary research in the domains of Solar Physics and Astroinformatics. As a senior member of the Data Mining Lab at Montana State University (MSU), I helped lead the research and development of several collaborative projects working with Big Data from NASA's Solar Dynamics Observatory (SDO) mission, which captures roughly 70,000 high-resolution images of the Sun (approx. 1.5TB of raw data) each day. Since traditional human-based analysis of this data is impractical, the SDO Feature Finding Team was created to produce a comprehensive set of automated feature recognition modules [1]. As a leader of the project and contributor of one of the 16 modules, our lab at MSU is building a trainable module for use in the first-ever Content-Based Image Retrieval (CBIR) system[1] for solar images [2]. The development of this module has led to many published works and several avenues of ongoing research [3].

Through this work over recent years, we have helped pioneer interdisciplinary research between Computer Science and Solar Physics – one of only a handful of groups doing so world-wide. From the perspective of solar physicists, we help solve problems they have not faced before, such as searching and cataloging images and regions of interest over vast data archives. As a computer scientist, we discover open research problems that underpin the solutions they require. Therefore, we are able to conduct novel research in both fields simultaneously, while providing practical solutions and applications to real problems.

Since the Fall of 2012, I have also taken on a mentoring role within our research lab, which usually has about ten (mostly non-U.S. resident) graduate students. I conducted weekly individual meetings with students, overseeing their progress on research and guiding their direction when needed. The ability to carry on multiple avenues of related research all aimed at larger outcomes outlined by grant funding was a satisfying experience, and I eagerly look forward to building my own research lab in the future.

## Thesis: High-Dimensional Data Retrieval

My thesis is focused on high-dimensional data indexing and $k$-nearest neighbor ($k$NN) retrieval. This is a critical component for fast and efficient large-scale CBIR systems, which typically employ upwards of 128-dimensional feature space for adequate characterization of data. Many typical database and information retrieval tasks are made much more difficult in higher dimensional spaces, a phenomenon referred to as the *curse of dimensionality*, dating back to 1957 [4]. While most historical research focused on the limits of the era (thousands of data points in 20-30 dimensions), modern data is massive and abundantly feature-rich (billions of data points in 100's of dimensions). In these real-life data mining applications, one of the most common searches in these spaces is the nearest neighbor query. This is useful in a wide variety of applications, including medical diagnosis, military and commercial surveillance, personalized recommendation systems, and many other data-driven decision support systems.

In initial research, we analyzed distance-based partitioning strategies of the iDistance indexing algorithm [5]. Our work showed that the algorithm stagnates in performance when an entire data partition must

[1] Publicly available at: http://cbsir.cs.montana.edu/sdocbir/

be searched to satisfy a given query, regardless of dataset size and dimensionality. While this is already significantly more efficient than sequential scan or other traditional low-dimensional methods, we achieve additional partitional segmentation with the creation of intuitive heuristics applied to a novel hybrid index called iDStar [6]. These extensions incorporate additional dataspace knowledge at the price of added algorithmic complexity and performance overhead, but show tremendous performance improvement possibilities in our results. Both algorithms and supporting research software we developed are free and open-source[2].

Clustering also plays an important role in many indexing techniques. We have started research towards developing novel clustering algorithms [7] for the explicit purpose of an indexing algorithm preprocessing step. Through empirical evaluation of indexing performance, we can customize an algorithm that delivers consistently better performance. This work is greatly exacerbated by the difficulties of high-dimensional spaces, and high-dimensional clustering is currently a very active and important research topic. In other related research, we addressed the idea of using a $k$NN-based index to satisfy range queries and outlined the general computational geometry problem of adapting such a system [8].

My current research aims to explore specific factors of poor index performance and the development of algorithmic extensions and optimizations to minimize their effects and occurrences. While testing individual performance factors, we will be able to establish more thorough algorithmic complexity bounds under various conditions, leading to a more robust index quality metric calculated during index creation. There also exists an abundance of follow-up research to investigate, including most interestingly: dynamic index maintenance through online partition tuning, parallelizing partition search and retrieval functionality, and facilitating more efficient multi-point collection queries for applications such as CBIR systems.

## Other: Machine Learning

In addition to my interdisciplinary data mining and high-dimensional data retrieval work, I also have a great interest in graph and text mining and related research. Following work of my advisor's prior students, I spent several months studying efficient graph indexing algorithms. I find the tremendous applicability of graphs fascinating and would greatly enjoy exploring more research opportunities in this direction.

I have been involved in several research projects in the areas of machine learning and artificial intelligence. For one of my first research appointments, we created an application that facilitates improved knowledge discovery from aircraft maintenance data by transforming transactional database records into ontology-based event graphs, and then providing a filterable visualization of event sequences through time [9]. The development of these ontological models and software tools aided future work in diagnostic model maturation and knowledge discovery from data (KDD) through graph-based data mining.

During my graduate studies in machine learning, I developed evolutionary algorithms using particle swarm optimization and genetic programming to derive custom kernel functions that improved performance of support vector machine classification [10]. I have since been contacted many times regarding this work and would be interested in continuing similar research with evolutionary algorithms in the future. My main interest is their ability to find novel solutions "outside the box" of typical human thinking, opening the doors to possibly revolutionary new concepts and conclusions. This is also very practical, as often times researchers might know what merits a good solution, but not necessarily how to go about creating one.

## Conclusions

An appointment in academia suites me as a person and professional. I have always been inspired by the late Dr. Carl Sagan, who once said, "We make our world significant by the courage of our questions and by the depth of our answers." Finding answers through research ignites my inner drive of discovery that sent me to graduate school in the first place and that will have me continuing to ask questions for the rest of my life. While we might not know all of the questions to ask yet, I look forward to the journey that lies ahead.

[2] Publicly available at: http://code.google.com/p/idistance/

## References

[1] P. Martens, G. Attrill, A. Davey, A. Engell, S. Farid, P. Grigis, *et al.*, "Computer vision for the solar dynamics observatory (SDO)," in *The Solar Dynamics Observatory* (P. Chamberlin, W. D. Pesnell, and B. Thompson, eds.), pp. 79–113, Springer, 2012.

[2] M. Schuh, J. Banda, R. Angryk, and P. Martens, "Introducing the first publicly available content-based image-retrieval system for the solar dynamics observatory mission," in *AAS/SPD Meeting*, vol. 44 of *Solar Physics Division Meeting*, p. #100.97, July 2013.

[3] M. A. Schuh, J. M. Banda, T. Wylie, P. McInerney, K. Ganesan Pillai, and R. A. Angryk, "On visualization techniques for solar data mining," *Astronomy and Computing (accepted to appear)*, 2015.

[4] R. Bellman, *Dynamic Programming*. Princeton University Press, 1957.

[5] M. A. Schuh, T. Wylie, J. M. Banda, and R. A. Angryk, "A comprehensive study of iDistance partitioning strategies for kNN queries and high-dimensional data indexing," in *Proc. of the 29th British National Conference on Databases (BNCOD), Big Data* (G. Gottlob, G. Grasso, D. Olteanu, and C. Schallhart, eds.), vol. 7968 of *Lecture Notes in Computer Science*, pp. 238–252, Springer Berlin Heidelberg, July 2013.

[6] M. A. Schuh, T. Wylie, and R. A. Angryk, "Improving the performance of high-dimensional kNN retrieval through localized dataspace segmentation and hybrid indexing," in *Proc. of the 17th East European Conference on Advances in Databases and Information Systems (ADBIS)* (B. Catania, G. Guerrini, and J. Pokorný, eds.), vol. 8133 of *Lecture Notes in Computer Science*, pp. 344–357, Springer Berlin Heidelberg, Sept. 2013.

[7] T. Wylie, M. A. Schuh, J. W. Sheppard, and R. A. Angryk, "Cluster analysis for optimal indexing," in *Proc. of the 26th International Florida Artificial Intelligence Research Society Conference (FLAIRS)*, pp. 166–171, AAAI, May 2013.

[8] M. A. Schuh, T. Wylie, C. Liu, and R. A. Angryk, "Approximating high-dimensional range queries with kNN indexing techniques," in *Computing and Combinatorics, the Proc. of the 20th International Conference (COCOON '14)* (Z. Cai, A. Zelikovsky, and A. Bourgeois, eds.), vol. 8591 of *Lecture Notes in Computer Science*, pp. 369–380, Springer International Publishing, Aug. 2014.

[9] M. Schuh, J. Sheppard, S. Strasser, R. Angryk, and C. Izurieta, "Ontology-guided knowledge discovery of event sequences in maintenance data," in *Proc. of the IEEE AUTOTESTCON Conference*, pp. 279–285, IEEE, Sept. 2011. Best Student Paper Award Winner.

[10] M. A. Schuh, R. A. Angryk, and J. W. Sheppard, "Evolving kernel functions with particle swarms and genetic programming," in *Proc. of the 25th International Florida Artificial Intelligence Research Society Conference (FLAIRS)*, pp. 80–85, AAAI, May 2012.