Since the early 1990’s, the ubiquity of personal computing technology has produced an abundance of staggeringly large data sets—it is estimated that Facebook alone logs over 25 terabytes of data per day and large bioinformatics data sets that integrate microarrays, sequences, and ontology annotations continue to grow. To compound this fact, these data sets are populated from disparate, often unknown, sources and are in a wide-range of formats. There is a great need for systems by which one can elucidate the similarity among and between groups in these data sets and produce easy-to-understand visualizations of the results. In this talk, I will discuss a method for efficiently and accurately approximating the solution of the kernel c-means clustering algorithm, specifically focusing on the fuzzy variant. Kernel clustering has been shown to be effective for data sets where the groups are not linearly separable in the input space or are high-dimensional. However, kernel fuzzy c-means (kFCM) presents computation and storage requirement challenges: clustering 500,000 objects requires 1 terabyte of main memory. I will show that on medium scale data (~50,000 objects) the approximate kFCM (akFCM) algorithm gives up to three orders of magnitude speed-up and a constant factor reduction in memory footprint with little-to-no degradation in performance, as compared to literal kFCM. I also demonstrate that akFCM performs well on large-scale data (>500,000 objects), including magnetic resonance imaging volumes. Last, I will apply the clustering method to bioinformatics data composed of genes described by Gene Ontology annotations to show how akFCM can be used for comparative genomics.

Short Bio: Dr. Tim Havens is an NSF / CRA Computing Innovation Fellow in the Department of Computer Science and Engineering at Michigan State University. He received an M.S. in electrical engineering from Michigan Tech University in 2000 and a Ph.D. in electrical and computer engineering from the University of Missouri in 2010. Prior to his Ph.D. work, he was an Associate Technical Staff at MIT Lincoln Laboratory where he specialized in the simulation and modeling of directed energy and global positioning systems. He has published over 40 journal articles, conference papers, and book chapters and received a best paper award from the Midwest Nursing Research Society in 2009. He is a senior member of the IEEE and an accomplished jazz bassist. You may learn more about Tim and his work at his website: http://www.timhavens.com

All students, faculty, and public are welcome.