Proposal for Gaining Access to Datasets in the IMEDS Lab for Studies on High Throughput Phenotyping from Electronic Health Records (EHRs)

Dr. Joydeep Ghosh, Schlumberger Centennial Chair Professor
Electrical & Computer Engineering Dept. at University of Texas, Austin

Dr. Joyce Ho, Assistant Professor
Dept. of Computer Science, Emory

- **Research Objectives:** The transformation of EHR data into concise clinical concepts, or phenotypes, is fundamental for healthcare studies as they can be used to obtain study cohorts, improve comparative effectiveness research, identify patients for screening tests and interventions, and offer interpretability of the data for medical practitioners [1]. However, EHR data are typically collected in operational settings for billing and management purposes, and are not designed for clinical research. Thus, they often do not readily map to simple, let alone more sophisticated and multifaceted, phenotypes. Traditional approaches for extracting phenotypes are typically slow, limited in scope, and manually intensive. In light of the above observations, our aim is to develop methods which will extract medical concepts from the data with minimal human supervision. The resulting phenotypes can be used for proactive patient management and prediction of progress (or deterioration) of patients.

- **Proposed Approaches:** We propose to represent and analyze the IMEDS data (specifically GE, MSLR and CCAE) as interconnected high-order relations; i.e., tensors (e.g. tuples of patient-medication-diagnosis, patient-lab, patient-diagnosis-procedure) and to develop novel computational algorithms for multi-tensor factorization with structure regularization. Many aspects of EHR data suggest they are better represented using tensors rather than a flat relational table (matrix), which is the default at the present time. Tensors succinctly capture the simultaneous interaction of multiple modes; for example, to identify a small subset of diagnoses and medications that are frequently observed together in a subgroup of patient EHRs. Therefore, such analysis can provide a powerful and data-driven approach to transform high–dimensional EHR data into medical concepts. Our preliminary research in this area has illustrated the promise of tensor factorization to uncover phenotypes which can enable numerous clinical applications such as cohort construction and predictive modeling of disease progression [2, 3, 4]. Moreover, such a framework would generalize several existing data mining methodologies, including dimensionality reduction, topic modeling and co-clustering, which all arise as limited special cases of analyzing second order tensors.

- **Impact:** There exists a paucity of methods to automatically extract medical concepts from high–dimensional EHR data without requiring human annotated samples. Moreover, black-

box algorithms that produce accurate predictive models have been met with resistance by medical professionals as the lack of interpretability fails to provide clinicians with actionable insights. Our goal is to produce a comprehensive computational framework that will substantially increase our ability to simultaneously identify existing as well as novel phenotypes from large–scale EHR data. The resulting phenotypes can also be used to develop tools for personalized health and well–being while also providing interpretability of high–dimensional EHR data. Our work will reflect algorithmic innovations as well as impact a broad range of health informatics settings. The proposed work will complement and enhance a current NSF funded project. This further ensures funding and interest in the project, and dissemination of the results to the scientific community.

- **Experience:** Professor Joydeep Ghosh (for bio and publications, please see http://ideal.ece.utexas.edu/ghosh/) will be the principal investigator for this project. He has over 400 refereed publications related to analysis of complex data and has given keynotes on health analytics. Two other individuals will be affiliated with this project. Joyce Ho, PhD, just graduated from the Electrical and Computer Engineering Department at UT-Austin and has joined as an assistant professor at Emory University. Jette Henderson is a third-year PhD student in the Institute for Computational Engineering and Sciences at UT-Austin.

- **Timeline:** We expect the research project to take 3 years and are requesting access beginning Feb 1, 2016 and concluding December o1, 2018.

- **Secure Data Control Plan:** Ideally both both UT and Emory will have access to the datasets requested. We have considered and agree with the policies and procedures described at http://imeds.reaganudall.org/On-Boarding. In addition, we have substantial prior experience with dealing with Category 1 data (strictest), HIPPAA compliant data, etc, and can make Secure Data Control Plans as need be.

References

[1] NIH Health Care Systems Research Collaboratory. Rethinking Clinical Trials: A Living Textbook of Pragmatic Clinical Trials. July 2014.

[2] Joyce C Ho, Joydeep Ghosh, Steve R Steinhubl, Walter F Stewart, Joshua C Denny, Bradley A Malin, and Jimeng Sun. Limestone: High-throughput candidate phenotype generation via tensor factorization. *Journal of Biomedical Informatics*, 52:199–211, December 2014.

[3] Joyce C Ho, Joydeep Ghosh, and Jimeng Sun. Marble: High-throughput phenotyping from electronic health records via sparse nonnegative tensor factorization. *KDD*, pages 115–124, 2014.

[4] Yichen Wang, Robert Chen, Joydeep Ghosh, Joshua C Denny, Abel N Kho, You Chen, Bradley A Malin, and Jimeng Sun. Rubik: Knowledge Guided Tensor Factorization and Completion for Health Data Analytics. *KDD*, pages 1265–1274, 2015.