Management and Analytic of Biomedical Big Data with Cloud-based In-Memory Database and Dynamic Querying

A Hands-on Experience with Real-world Data

Mengling Feng
Massachusetts Institute of Technology
mfeng@mit.edu

Mohammad Ghassemi
Massachusetts Institute of Technology
ghassemi@mit.edu

Thomas Brennan
Massachusetts Institute of Technology
tpb@mit.edu

John Ellenberger
SAP Research
john.ellenberger@sap.com

Ishrar Hussain
SAP Research
ishrar.hussain@sap.com

Roger Mark
Massachusetts Institute of Technology
rgmark@mit.edu

ABSTRACT

Analyzing Biomedical Big Data (BBD) is computationally expensive due to high dimensionality and large data volume. Performance and scalability issues of traditional database management systems (DBMS) often limit the usage of more sophisticated and complex data queries and analytic models. Moreover, in the conventional setting, data management and analysis use separate software platforms. Exporting and importing large amounts of data across platforms require a significant amount of computational and I/O resources, as well as potentially putting sensitive data at a security risk. In this tutorial, the participants will learn the difference between in-memory DBMS and traditional DBMS through hands-on exercises using SAP's cloud-based HANA in-memory DBMS in conjunction with the Multi-parameter Intelligent Monitoring in Intensive Care (MIMIC) dataset. MIMIC is an open-access critical care EHR archive (over 4TB in size) and consists of structured, unstructured and waveform data. Furthermore, this tutorial will seek to educate the participants on how a combination of dynamic querying, and in-memory DBMS may enhance the management and analysis of complex clinical data.

Instructors

Mengling Feng is currently a visiting scholar at MIT in Harvard-MIT Health Sciences and Technology Division. Dr. Feng was awarded the Ministry of Education Scholarship for his undergraduate studies and the A*STAR Graduate Scholarship for his PhD study. His research was recognized with the *Bi-annual Best Paper Award* from the Institute for Infocomm Research. Dr. Feng's research focus is to develop data mining and machine learning methods to discover or infer casual phenomenon among real-life practices.

Mohammad Ghassemi is a PhD student in Electrical and Computer Engineering at the Massachusetts Institute of Technology with a research interest in statistical signal processing and medical informatics. In 2010, Mohammad received the Gates-Cambridge Scholarship to fund his MPhil at the University of Cambridge in Information Engineering. He was also awarded the Goldwater scholarship while pursuing two undergraduate degrees in Electrical Engineering and Applied Mathematics.

Thomas Brennan is currently a post-doctoral Research Engineer in the Laboratory of Computation Physiology, MIT. With a background in electrical and computer engineering, Thomas Brennan was awarded the Rhodes Scholarship from South Africa in 2004. In 2009 he completed his D.Phil. in Biomedical Engineering from the University of Oxford. In 2010, he accepted a Wellcome Trust post-doctoral research fellowship at Oxford.

John Ellenberger currently works in the "Chairman's Special Projects" group at SAP Labs where he serves as the liaison to MIT where he focuses on Big Data Topic including medical analytics, machine learning and data privacy and architecture. John has worked in SAP's research function for 10 years in both strategic and technical research roles.

Ishrar Hussain is working as a Machine Learning Researcher at SAP Research Labs, Montreal, and developing state-of-the-art machine learning solutions for Big Data applications using the SAP HANA's in-memory DBMS. He received his Master' degree in Computer Science in 2007 from Concordia University in Montreal, Canada. He is now also a Ph.D. Candidate at the same university and is expecting to graduate in 2014.

Roger G. Mark earned the SB and PhD degrees in electrical engineering from Massachusetts Institute of Technology and the MD degree from Harvard Medical School. At present Dr. Mark is Distinguished Professor of Health Sciences and Technology, and Professor of Electrical Engineering at MIT. His current research activities include *Integrating Data, Models, and Reasoning in Critical Care* and *PhysioNet* (http://www.physionet.org), both of which involve the development, use and open distribution of large physiological and clinical databases.

KDD'14, August 24–27, 2014, New York, NY, USA.
ACM 978-1-4503-2956-9/14/08.
http://dx.doi.org/10.1145/2623330.2630806.