For several decades now, the amount of data available to us has been growing at a pace far higher than our ability to process it; this trend has accelerated many-fold in recent years with the emergence of efficient and mass-produced scientific instruments, increasing ease of generating and publishing data, and proliferation of Internet-connected devices. In this talk, I will present an overview of our ongoing work on building a platform for enabling collaborative data science, where teams of data scientists can simultaneously analyze, modify, and share datasets, to understand trends and to extract actionable insights. While numerous solutions exist for specific data analysis tasks, underlying infrastructure and data management capabilities for supporting ad hoc collaboration pipelines are still largely missing. I will present our vision for a unified, dataset-centric platform for addressing these challenges, and present our recent work on: (a) efficiently managing a large number versioned datasets, (b) designing and supporting a unified query language to seamlessly query versioning and provenance information, and (c) lifecycle management of complex machine learning models like deep neural networks.

Bio: Amol Deshpande is a Professor in the Department of Computer Science at the University of Maryland with a joint appointment in the University of Maryland Institute for Advanced Computer Studies (UMIACS). He received his Ph.D. from University of California at Berkeley in 2004. His research interests include uncertain data management, adaptive query processing, data streams, graph analytics, and sensor networks. He is a recipient of an NSF Career award, and has received best paper awards at the VLDB 2004, EWSN 2008, and VLDB 2009 conferences.