TT Presenters for 2012

Schedule for 2012:

2012 Spring topics will primarily focus on "Big Data" and we intersperse that with social coding, experimental reproducibility talks.

Date	Presenter	Contact	Title	Host	URL/Link	Abstract/Notes
Feb 8	Anjul Bhambhri, VP for Big Data , IBM		Big Data and Better Business Outcomes	Nirav Merchant		Organizations today want to tap into the wealth of information hidden in the data around them to improve competitiveness, efficiency and profitability. Huge volumes of data are created every day from a variety of sources including: sensors, smart devices, social media and billions of Internet and smart phone users worldwide. The challenge is storing, managing and deriving just-in-time insights from this data, while preserving and using existing information management investments. The Big Data challenge is pervasive across the majority of industries including: finance, government, telecommunications, retail, healthcare, energy and utilities. This presentation will look at emerging technologies (Watson), evolving roles (data scientists) and how technology can drive better business outcomes. Only through Integrating Big Data and applying context, patterns and intelligence will drive new business efficiencies.
Feb 22	Zachary Simpson		Happy Science Coding	Nirav Merchant	http://happysciencecoding.com/	Zack Booth Simpson, Fellow of the Institute for Cellular and Molecular Biology UT Austin, will demonstrate a new online development platform aimed at scientists called "Happy Science Coding". The web-based platform permits social coding with online documents editable from the browser, simple automatic version control, and a open, server-based execution model that permits any computer in the cloud to serve as an execution machine. A model for maintaining development state will be proposed that would enable scientists to mark their code and dependency state for publication ensuring that others would be able to run the code in perpetuity.

March 14	Michael Schatz , CSHL		Entering the era of mega-genomics	Nirav Merchant	http://schatzlab.cshl.edu/research/	The continuing revolution in DNA sequencing and biological sensor technologies is driving a digital transformation to our approaches for observation, experimentation, and interpretation that form the foundation of modern biology and genomics. Whereas classical experiments were limited to thousands of hand-collected observations, today's improved sensors allow billions of digital observations and are improving at an exponential rate that exceeds Moore's law. These improvements have made it possible to sequence new genomes and monitor the dynamics of biological processes on an unprecedented "mega-scale", but have brought proportionally greater quantitative and computational requirements. The growing digital demands have motivated extensive research into computational algorithms and parallel systems for analysis. Recently a great deal of research has been focused on applying emerging scalable computing systems to genomic research. One of the most promising is the Hadoop open-source implementation of MapReduce: it is specifically designed to scale to very large datasets, its intuitive design supports rich parallel algorithms, and is naturally applied to analysis of many biological assays. During my presentation, I will describe some recent innovations using these and other technologies for large-scale genome assembly, variation detection, and transcription analysis. These are promising early results but continued research is essential in the coming years, especially as we hope to model and mine these data to uncover genotype-to-phenotype relations that can only be detected across very large populations.

March 28	Paul Brown		SciDB: Large Scale Array Data Management	Nirav Merchant	http://www.scidb.org/	In this talk you will learn about the SciDB big data storage and analytic platform. In contrast to big data approaches associated with Map/Reduce technologies and inspired by the requirements of web log analysis, SciDB builds on ideas and methods developed over many years to cope with the challenges of large scale scientific data analysis. SciDB is an transactional DBMS that provides its users with a declarative query language build on top of an array data model. Building on its extensibility and MPP foundations, SciDB supports a wide range of statistical and data processing functionality in a similar fashion to ScaLAPACK. SciDB is a completely new implementation that takes advantage of the intrinsic ordering of array data to deliver superior scalability and performance over complex, real-world workloads.

April 11	Jeff Kantor, LSST		An Overview of the LSST Data Management System	Andrew Lenards	http://www.lsst.org/lsst/about	The LSST Data Management System (DMS) processes the incoming stream of images that the camera system generates to produce transient alerts and to archive the raw images, periodically creates new calibration data products that other processing functions will use, creates and archives an annual Data Release (a static self-consistent collection of data products generated from all survey data taken from the date of survey initiation to the cutoff date for the Data Release), and makes all LSST data available through an interface that uses community-based standards and facilitates user data analysis and production of user-defined data products with supercomputing-scale resources. This paper discusses DMS distributed processing and data, and DMS architecture and design, with an emphasis on the particular technical challenges that must be met. The DMS publishes transient alerts in community-standard formats (e.g. VOEvent) within 60 seconds of detection. The DMS processes and archives over 50 petabytes of exposures (over the 10-year survey). Data Releases, include catalogs of tens of trillions of detected sources and tens of billions of astronomical objects, 2000-deep co-added exposures, and calibration products accurate to standards not achieved in wide-field survey instruments to date. These Data Releases grow in size to tens of petabytes over the survey period. The expected data access patterns drive the design of the database and data access services. Finally, the DMS permits interactive analysis and provides nightly summary statistics describing DMS output quality and performance.

April 25	Philip Guo, Stanford		CDE: A tool for creating portable experimental software packages	Nirav Merchant	http://www.stanford.edu/~pgbovine/cde.html	One technical barrier to reproducible computational science is that it is hard to distribute scientific code in a form that other researchers can easily execute on their own computers. Before your colleagues can run your experiments, they must first obtain, install, and configure compatible versions of the appropriate software and their myriad of dependencies. To eliminate this technical barrier, I have created a tool called CDE that automatically packages up all of the software dependencies required to re-run your computational experiments on another computer. CDE is easy to use: All you need to do is execute the commands for your experiment under its supervision, and CDE packages up all of the Code, Data, and Environment that your commands accessed. When you send that self-contained package to your colleagues, they can re-run those exact commands on their computers without first installing or configuring anything. CDE is free and open source, available at http://www.pgbovine.net/cde.html	
May 9	Wes Turner	Kitware	Open Source Software for Scientific Data Analysis and Presentation	Eric Lyons		Scientific software development often consists of searching for a useful algorithm and then spending additional weeks decoding the text of a paper to actually reach a usable piece of code. However, much of what is produced is derivative and can be thought of as commodity. Kitware, Inc. is a company founded to give away commodity software allowing scientists and Kitware developers to focus on the science of generating new answers instead of implementing old solutions. In this talk I will present an open source toolkit for scientific data visualization and exploration along with some of the open source applications that Kitware has derived. All of the code I will present is freely available with liberal, Apache 2 licensing terms.	
May 23	Amazon AWS		TBD->Large scale database infrastructure: DynamoDB	Nirav Merchant	http://aws.amazon.com/dynamodb/		