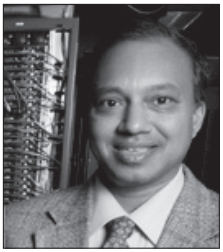
Big Data Workshop

Emerson Alumni Hall
University of Florida
Wednesday, June 19, 2013

Table of Contents



Big Data from Simulation of Extreme and Environmental Problems

S. Balachandar
Professor, Department of Mechanical & Aerospace Engineering
bala1s@ufl.edu

There are a wide class of scientific and engineering multiphysics problems, where large-scale simulations is the only option. The deep multiscale nature of these problems makes them computationally very expensive. On one hand, faithful simulations of these complex problems require advanced mathematical models and numerical algorithms that can perform efficiently at peta and exascale. On the other hand, the vast amount of data that will be generated from such simulations pose an even greater challenge. We must be concerned with the challenges of storing, retrieving, processing and interpreting such large data. In particular, we must focus is on advancing mathematical and computational techniques for addressing the challenges in data compression, low-dimensional projection, surrogate modeling, and so on.

Biology and Big Data

Brad Barbazuk
Associate Professor, Department of Biology and
The University of Florida Genetics Institute
bbarbazuk@ufl.edu

Next-generation sequence (NGS) technologies are recent advances in sequencing chemistries and instrumentation that are enabling rapid acquisition of high volumes of DNA sequence at greatly reduced cost. This paradigm change in DNA sequencing has ushered in an era where sequence acquisition frequently underpins investigation into biological phenomena. While NGS provides a powerful way to investigate important questions in the life-sciences, it is not without its drawbacks. NGS data is very high volume and consists of very of short DNA sequences. The big data challenges associated with NGS are numerous. The volume of the raw data makes transfer and storage difficult, subsequent manipulations amplify the storage footprint, and common analyses require multiple CPU and large memory. This presentation will introduce the use of NGS in medicine and life science, discuss current activities and project future needs, and present a simple analogy to illustrate how DNA sequencing has grown and help illustrate the data volume issue.

Proteomics and Metabolomics Data Analysis Toward Systems Biology

Sixue Chen
Associate Professor, Department of Biology, ICBR, PMCB, Genetics Institute
schen@ufl.edu

Proteomics and metabolomics data can reveal molecular mechanisms directly related to the phenotype and behavior of organisms. Because of rapid advancement in analytical chemistry including mass spectrometry, genome sequencing and computational biology, proteomics and metabolomics technologies have been implemented in more and more research projects to answer key scientific questions that cannot be addressed using other tools. More and more large scale data at protein and metabolite levels have been generated. The presentation will highlight recent advancement in data generation, integrative analysis, challenges and opportunities.

Opportunities for Big Data Medical Records

Michael Conlon
UF Clinical and Translational Research Institute
mconlon@ufl.edu

The future of health and healthcare will be determined in large part by our ability to collect, interpret and use medical records at the patient, provider and national level. Traditional medical records — notes from visits, disease, diagnostic, procedure and pharmaceutical records, along with medical images, are typically collected and managed by health care provider organizations. More recently, we are seeing a rapid two-dimensional expansion in data — 1) assembly of medical records across providers and systems, generating datasets of tens of millions of patients, and 2) deepening the data available on each patient through explosive growth in patient-reported outcomes, natural language processing, and availability of genetic and molecular information. In this talk I will characterize the explosion in available information and the resulting opportunities for speeding therapeutic development, identifying and curtailing adverse events and practices, with resulting improvement in healthcare practice, and health.

Hipergator And Infrastructure For Working With Data

Erik Deumens
Director, Research Computing, UFIT
deumens@ufl.edu

An overview will be given of the infrastructure that is now available to researchers at UF and their world-wide collaborators to perform analytics and modeling involving demanding computations on complex and large data sets and to share and move around these large data sets efficiently.

Metabolomics at UF

Arthur S. Edison
Professor, Biochemistry & Molecular Biology and National High Magnetic Field Laboratory
aedison@ufl.edu

Metabolomics is the measurement and quantitative characterization of small molecule metabolites in living systems. It represents the most functional level of 'omics data (i.e. genomics, transcriptomics, proteomics, metabolomics), because metabolites change rapidly with environmental conditions and can often be used as biomarkers of disease. When combined with measurements of phenotype and analyzed with multivariate tools of modern data mining, metabolomics and other 'omics data form the basis of systems biology. I will summarize the types of datasets and data mining approaches in metabolomics studies. I will also introduce the Southeast Center for Integrated Metabolomics (SECIM), a new resource at UF to provide metabolomics services.

A Role For Glue People In Big Data Research

William Farmerie
Interdisciplinary Center for Biotechnology Research (ICBR)
wgf2@ufl.edu

The industrialization of very large-scale DNA sequencing is transforming genetic data acquisition from an academic-centered domain, to a commercially viable commodity service. This is the good news: the cost and pace of data generation is not the limiting factor in gene-based science. Our challenge as a research-centered organization is converting data into value by enabling and coordinating data management, analysis, and interpretation. And we must do so without obscuring the portal of entry, or unnecessarily complicating systems navigation. In order to facilitate academic research in this area, a class of scientist-specialists, project-liaisons, or "glue people" will become essential resources, reducing the entry and navigation barriers for faculty-scientists engaged in big data research.

Big Data for Environmental Monitoring

Paul Gader
Professor and Chair, Computer and Information Science and Engineering
pgader@ufl.edu

Accurate measurements of the state of our global environment are essential for informed decision-making in precision agriculture, prediction of natural disasters, and climate change. NASA and the NSF are on the verge of producing data sets using new sensors, specifically imaging spectrometers and LiDARs. The sizes of these data sets are unprecedented. Furthermore, estimating the parameters describing the environment is an ill-posed inverse problem and requires regularization using ancillary measurements as well as structured and unstructured knowledge. Computational methods including Machine Learning, Data Science, High-Performance Computing, and Probabilistic Reasoning need to be coupled with scientific and remote sensing expertise to provide the accurate information necessary for informed debate on crucial global issues.

Automated Analysis of Traffic Simulation

David K. Hale
Department of Civil Engineering, Transportation Group
david@ce.ufl.edu

Usage of traffic simulation has increased significantly over the past two decades; and this high fidelity modeling, along with moving vehicle animation, has allowed important transportation decisions to be made with better confidence. During this time, traffic engineers have typically been encouraged to embrace the process of calibration, in which steps are taken to reconcile simulated and field-observed traffic performance. According to international surveys, top experts, and conventional wisdom, existing (non-automated) methods of calibration have been difficult and/or inadequate. There has been a significant amount of research in the area of automated calibration techniques, for traffic simulation. However, many of these research projects and papers have not provided the level of flexibility and practicality that are typically required by real-world engineers. With this in mind, the self-calibration features within TSIS-CORSIM were designed with an eye on maximizing practicality, flexibility, and ease-of-use. TSIS-CORSIM is the flagship software tool for traffic simulation at the University of Florida. I applied for patent protection on this automated calibration process; in part due to the significant cross-disciplinary potential, as the process appears to be compatible with non-simulation and non-transportation models. This automated process is specifically designed to accommodate any model or process that has 1) a significant number of input parameters, 2) a significant number of output parameters, and 3) an unpredictable running time on the computer. In addition to my interest in pursuing self-calibration for various models in various fields, I would also like to investigate rigorous output analysis techniques, in order to automate many of the transportation decisions typically relegated to simple engineering judgment.

Data and Research for Sustainable Food Security

James W. Jones
Distinguished Professor, Florida Climate Institute/Agricultural & Biological Engineering Department
jimj@ufl.edu

Food insecurity is emerging as a major global challenge due to various factors, including an increasing human population, changing diets, climate shocks, competition between uses of agricultural lands for food vs. energy, and limited water for expanding irrigation. Global prices of food crops have spiked in recent years and are projected to increase as the above factors are combined with a warmer climate that is projected to decrease food productivity over much of the Earth's surface. We can no longer consider food insecurity as a problem only in developing or unstable countries; it is critical that we better understand the real threats to food security at home and abroad. How will agriculture feed 9 billion people on Earth in 2050? This is a complex question with global and regional manifestations; insight into solutions requires science-based biophysical and socioeconomic models of food production, distribution, and use along with large and diverse databases. Global databases are required on historical and future climate conditions, including its seasonal and annual variations and extreme events, soil, water resources, socioeconomic conditions, crops, production technologies, policies, agronomic experiments, and agricultural statistics. In this talk, I will describe examples of recent advances in integrated assessment of food production and food security at regional and global scales in the AgMIP program. I will also describe databases required, which when taken together, are examples of the big data that is needed for strategically assessing the future of food and analyzing likely benefits of alternative policies and technologies at local to global scales.

Topological Data Analysis

Kevin Knudson
Department of Mathematics
kknudson@honors.ufl.edu

Methods from Algebraic Topology may be employed to reveal geometric structures inside large data sets. I will briefly discuss two ideas — persistent homology and discrete Morse theory.
Persistent homology measures geometric features in data sets at a variety of scales, allowing one to determine which are significant and which are noise. Discrete Morse theory is a mechanism to find critical points and gradient-like flows in data sets. Some examples will be presented as well.

Big Data Meets High-Performance Reconfigurable Computing

Herman Lam
Associate Professor, NSF Center for High-Performance Reconfigurable Computing (CHREC)
hlam@ufl.edu

In conventional computing, an application must conform to the predefined, fixed hardware of the target architecture (such as a cluster of CPUs and GPUs) using fixed cores, memory, interconnect, and I/O structures. Reconfigurable computing (RC), on the other hand, can be defined as computing with hardware-reconfigurable circuits, devices, systems where the architecture can be adapted to match unique needs of each application. In this presentation, we will describe the research and technology of CHREC (NSF Center on High-Performance Reconfigurable Computing), which are well positioned to strongly support the unique demands of Big Data in domains in which the Big-Data problems are not well served by conventional computing. Examples include Big-Data challenges in computational biology (e.g., data generated by new-generation sequencers) and Big-Data analytics in computational finance. Signal, image, and video processing is another demanding Big-Data area in which the escalation in sensor fidelity and diversity is creating an explosion in data to be processed in limited time, power, space, and cost (e.g., emerging on-board processing requirements in space computing). Using Novo-G, a reconfigurable supercomputer developed and deployed at CHREC, we have achieved orders of magnitude improvement in performance for applications in these domains with O(1000) of times less cost, size, power, and cooling than massive conventional supercomputers.

Empowering Nonlinear And Stochastic Optimization For Large-Scale Data Analysis

Guanghui (George) Lan
Professor, Department of Industrial and Systems Engineering
glan@ise.ufl.edu

We will discuss the big-data challenges in the design and analysis of optimization algorithms. The last several years have seen an unprecedented growth in the amount of available data. While nonlinear, especially convex programming models are critically important to extract useful knowledge from raw data, high problem dimensionality, large data volumes and inherent uncertainty present significant challenges to the design of optimization algorithms. We briefly introduce some of our recent work in the design and analysis of nonlinear and stochastic optimization algorithms, which provide robust and scalable solutions to many machine learning and imaging processing problems.

Big Data, Big Noise, And Big Simulations

Joanna R. Long
Associate Professor, Biochemistry & Molecular Biology and National High Magnetic Field Laboratory (NHMFL)
jrlong@mbi.ufl.edu

The AMRIS facility at the University of Florida is a premier facility for developing MRI and NMR techniques through its affiliation with the NHMFL. These techniques are unique in their ability to non-invasively characterize systems from an atomic level to a macroscopic level, and in fact magnetic resonance encompasses a large universe of experiments for characterizing the chemical and physical properties of nuclear spins. However, the low energy characteristics of MR techniques which make them attractive also render them to be inherently insensitive. Fortunately, most experiments involve nuclear spin systems which can be fully described using a finite parameter sets. I will discuss how sparse data sampling, statistical analyses and real time simulation of data could assist in on-the-fly decision optimization of data collection to maximize MR results in real time.

Big Geographic Data and GISciences

Liang Mao
Assistant Professor, Department of Geography
liangmao@ufl.edu

Spatial-temporally resolved data are big, and are widely used in geo-analysis, geo-simulation, and geo-visualization. In this talk, I'll give a bit of background about what geographic data is, why it is big, and how GISciences deal with it. I'll also describe a couple of "Big Data" research projects we've been doing that are especially relevant to medical geography, including the use of satellite images to predict air pollution, massive agent-based simulation for influenza diffusion, and the transmission of vector-borne diseases through air travel, etc. At last, I will present my own perspective on the challenging research problems it brings up.

Imaging Structure And Function In Biology With Magnetic Resonance And Electroencephalography

Thomas H. Mareci
Department of Biochemistry and Molecular Biology, College of Medicine
Advanced Magnetic Resonance Imaging and Spectroscopy Facility,
McKnight Brain Institute
National High Magnetic Field Laboratory
thmareci@ufl.edu

Magnetic resonance imaging can provide non-invasive information about both biological structure and function in the brain, with complimentary functional information available from electroencephalography, from a single subject. Typically magnetic resonance measurements view the subject in various ways that reflect important characteristics of the subject's anatomy, physiology, and function. A large amount of digital image data is acquired at a relatively high rate, then extensive analysis is performed on the data to provide the information from the various views of the subject to relate structure to function. For example, the structure of the entire human brain of a single individual can be measured with several gigabytes of digital image data in less than 30 minutes. In a similar timeframe, the function of the brain can be measures in a similar data size with both magnetic resonance and electroencephalography. The analysis of these data require files many times the size of the acquired data (usually hundreds of gigabytes) to provide the information needed to relate structure to function in a single human brain. Therefore, magnetic resonance imaging and electroencephalography can generate a large amount of data at such a rapid rate that data handling and processing can become a significant challenge.

Big Data Support for Scientific Disciplines through Information Technology Engineering

Andréa Matsunaga, Maurício Tsugawa
ACIS Laboratory, Department of Electrical and Computer Engineering, UF
ammatsun@acis.ufl.edu; tsugawa@acis.ufl.edu

The Advanced Computing and Information Systems (ACIS) Laboratory conducts fundamental and applied research on systems that integrate computing and information processing. This includes research on the management and processing of data of different types and scales as exemplified by the following ACIS projects, currently underway: iDigBio is making data and images of millions of biological specimens available in electronic format, representing the long-tail and variety of big data; Disaster Mitigation and Recovery through the use of virtual machine migration technologies exemplifies real-time big data; and FutureGrid conducts research and development of cloud technologies, and enables the deployment of distributed testbeds for big data analysis.

Novel Methods For In Situ Measurement And Simulation Of Hurricanes

Kamran Mohseni
Professor, Mechanical and Aerospace Engineering and Electrical and Computer Engineering
mohseni@ufl.edu

Big DATA: Research and Education

Sanjay Ranka
Professor, Computer Information Science and Engineering
ranka@cise.ufl.edu

The Power of Multiple Data Streams and Big Data in Health and Health Care

Betsy Shenkman
Professor and Chair, Department of Health Outcomes and Policy,
College of Medicine
Director, Institute for Child Health Policy
eshenkman@ufl.edu

This presentation will focus on the value of large datasets in health care to: compare the effectiveness of different interventions, examine the safety, quality, and outcomes of care, and conduct longitudinal studies designed to address heterogeneity of treatment effects (what works best for whom under what circumstances). Key issues related to data linkage, the use of computational capabilities to build clinical and patient-reported data infrastructures to support research and ongoing quality improvement in health care will be addressed. Exemplars related to current studies leveraging large datasets, including stakeholder engagement, will be provided. Data linkages and partnerships in development will also be presented.

Big Data in Biodiversity Studies

Pamela S. Soltis
Curator, Florida Museum of Natural History
psoltis@ufl.edu

Earth today supports approximately 1.8 million named species of organisms, approximately 1% of all life that has ever existed on the planet, with the total number of living species, including those unknown to science, estimated to be 10-100 million. Understanding how these species are related evolutionarily and distributed ecologically is fundamental to a comprehensive view of the biosphere — from its origins to predicted responses to global change. Synthetic analyses of data — genes, genomes, morphology, natural history collections information, geological information, climate, land use, etc. — require new computational environments that allow facile integration of data, software, and high-performance computing. Analyses of the ~4000 species of Florida plants provide a model for this type of computational environment, and a prototype is being developed at iDigBio, the National Resource for Biodiversity Collections, housed at UF & FSU.

Operations Research in the Big Data Realm: Challenges and Opportunities

Cole Smith
Professor and Interim Chair, Industrial and Systems Engineering
cole@ise.ufl.edu

The modern field of operations research (OR) has traditionally suffered from a lack of data. Models exist to solve many variations of vital problems arising in computational biology, transportation, energy, communications, defense, and so on. With the explosion of the amount of data that is routinely collected and stored in contemporary settings, a rapid shift has occurred in the OR field: How to best exploit the overwhelming amount of data that is now present to support OR investigations. These challenges include how to parse through and optimize over tremendously large data sets, and how to guide the investigation (or sampling) of new data sets. The goal of this talk is to describe the role of OR within Big Data, and to promote collaboration among our researchers and those outside the ISE department and the College of Engineering.

Big Data And The Search For A Microbial Cause For Disease

Eric Triplett
Professor and Chair, Microbiology and Cell Sciences, IFAS
ewt@ufl.edu

We are part of a large community trying to understand the role of the human gut microbiome in the development of autoimmunity for type 1 diabetes and celiac disease. As part of an international collaboration on type 1 diabetes called TEDDY (The Environmental Determinants of Diabetes in the Young), we are trying to learn those environmental variables that are correlated with the development of autoimmunity with a particular focus on those factors that regulate the microbial community in the human gut including antibiotic use, infectious episodes, and diet. We also are generating large microbiome datasets to correlate microbial taxa with these variables. This is a large complex problem where we need to be careful to separate signal from noise. We need to be sure that various confounders are not leading us down false paths. And we need to be aware that disease may be caused by different factors in different subjects. Our datasets are a complex mixture of numerical and categorical data where new methods of analyses will likely need to be developed.

Big Data Systems for Knowledge Base Construction from Text, Images and Crowds

Daisy Zhe Wang
Assistant Professor, Computer and Information Science and Engineering
daisyw@cise.ufl.edu

Keyword search engines have been the state-of-the-art information retrieval tool over large text corpora for two decades. To date, most search engines have little understanding that keywords and documents refer to entities and relations in real-life. Better search results and experience can be achieved by understanding entities and relations in documents as well as in queries. A knowledge base (KB) or a knowledge graph (KG) containing relevant entities (nodes) and relations (edges) should be the backbone of any application that is fueled by text. Given a large amount of text data, Big Data systems are needed that can automatically (1) construct knowledge bases using machine learning models and statistical inference algorithms, (2) manage the uncertainty inherent in the extracted knowledge, and (3) maintain them over time.

I will talk about our current experience in the TREC Knowledge Base Acceleration (KBA) challenge organized by NIST dealing with 4.5 terabytes of compressed web and social media data. I will talk about my current effort to build a probabilistic knowledge base (ProbKB) system with a deep integration of the statistical machine learning (SML) methods with scalable data processing frameworks. I will also talk about other projects at UF Data Science Research Lab (http://dsr.cise.ufl.edu/) including Big Data systems for large-scale image retrieval and Big Data analytic systems leveraging crowd sourcing and human intelligence.

Notes