Keynote Abstract I

Science in Cyberspace

S. George Djorgovski
Professor of Astronomy,
California Institute of Technology

Like any other field of the modern human endeavor, science, scholarship, and education are being transformed by the advances in information and computation technology (ICT). Much of the scholarly work is now moving to virtual environments, typically accessible through the Web. Scientists now spend much of their time interacting with the data, theoretical models, published literature, and their colleagues through various Internet-based mechanisms. Scientific content now resides and develops in cyberspace.

The exponential growth of data volumes, and the simultaneous increase in the data complexity create new scientific opportunities, but also pose many challenges: first, in the arena of data gathering, preservation, management, and access; but more importantly, in the arena of data exploration and knowledge discovery. Techniques developed in the era of data poverty do not scale to the regime of exponentially growing data abundance. Equally transformative changes have occurred in the scientific publishing and knowledge dissemination through electronic media. Thus, the scientific method evolves, driven by the needs of the computationally enabled, data-intensive science for the 21st century.

This is particularly important, as many of the problems facing our civilization, e.g., climate change, sustainability, health, security, etc., are inherently multi-disciplinary, requiring large and complex data sets and simulations, and the effective means of their analysis; in other words, they are largely computational in nature. ICT can play an essential connecting role in the inter/multi-disciplinary approaches to such problems.

Scientific community has responded to these challenges with the rise of virtual scientific organizations (VSOs): distributed research environments connecting the data archives, and the web-based tools for their federation and analysis. We also see a rise of “X-Informatics”, where X = (astro, bio, geo…), as many sciences are incorporating modern computational tools to enable discovery in their domain. However, the uptake of these new methodologies for scientific research in the 21st century has been relatively slow in the academic community, and the reasons for that are largely sociological rather than technological.

Meanwhile, information technology also enables and empowers education and public outreach, activities whose importance is fundamental and growing. In fact, the current and upcoming generations of “digital natives” expect their education to be delivered through electronic media. Educational institutions must evolve in the ICT-dominated world.

Any technology that enables a better communication between scientists, scholars, and students, and also between scientists and their data, models, and literature is thus essential. One example of is the rise of immersive virtual reality environments and virtual worlds, which may become a next generation interface to the informational content now available through the Web, as well as human-friendly interaction environments for scientific collaboration and education.

Science, technology, culture, and society in general have always co-evolved each enabling progress in the others. This process has never been more intensive than today, with the ICT revolution enabling, driving, and connecting humans and their informational constructs.
Keynote Abstract II

Blue Waters: An Extraordinary Computer to Enable Extraordinary Research

Thom Dunning
National Center for Supercomputing Applications & Institute for Advanced Computing Applications and Technologies, University of Illinois at Urbana-Champaign, Urbana, Illinois

A dramatic increase in computing capability has the potential to create breakthrough advances in all fields of science and engineering, including predicting the behavior of complex biological systems, understanding the production of heavy elements in supernovae, designing catalysts and other materials at the atomic level, predicting changes in the earth’s climate and ecosystems, and designing complex engineered systems from chemical plants to airplanes. These breakthroughs require a computing resource capable of solving the most compute-, memory- and data-intensive research questions.

The Office of Cyberinfrastructure in the National Science Foundation is funding the acquisition and deployment of such an extraordinary new supercomputing resource at the National Center for Supercomputing Applications on the campus of the University of Illinois at Urbana-Champaign. This system, called Blue Waters, is based on the latest computing technology under development by IBM for DARPA’s High Productivity Computing Systems Program, including the Power7 processor with a peak performance of a quarter of a teraflop and a new communications hub chip capable of more than one terabyte/sec. Blue Waters will first come online in 2011.

Blue Waters will have more than 300,000 compute cores, 1000 terabytes of main memory, and 20 petabytes of disk storage, and up to 500 petabytes of archival storage. It will be connected to the nation’s research networks at 100+ Gbps. Blue Waters uses the most powerful microprocessors available today; an innovative low latency, high bandwidth communications network that facilitates scaling to large numbers of compute cores; and an I/O subsystem that will enable the solution of the most challenging data-intensive problems. Detailed performance projections indicate that Blue Waters will sustain 1 petaflops on many real-world science and engineering applications.

This presentation will provide information on Blue Waters and its proposed use by the science and engineering community.
Keynote Abstract III

Open Source Modeling as an Enabler of Transparent Decision Making

Ian T. Foster
Computation Institute, Argonne National Lab
and The University of Chicago

In an increasingly interconnected and human-modified world, decision makers face problems of unprecedented complexity. For example, world energy demand is projected to grow by a factor of four over the next century. During that same period, greenhouse gas emissions must be drastically curtailed if we are to avoid major economic and environmental damage from climate change. We will also have to adapt to climate change that is not avoided. Governments, companies, and individuals face what will be, in aggregate, multi-trillion-dollar decisions.

These and other questions (e.g., relating to food security and epidemic response) are challenging because they depend on interactions within and between physical and human systems that are not well understood. Furthermore, we need to understand these systems and their interactions during a time of rapid change that is likely to lead us to states for which we have limited or no experience. In these contexts, human intuition is suspect. Thus, computer models are used increasingly to both study possible futures and identify decision strategies that are robust to the often large uncertainties.

The growing importance of computer models raises many challenging issues for scientists, engineers, decision makers, and ultimately the public at large. If decisions are to be based (at least in part) on model output, we must be concerned that the computer codes that implement numerical models are correct; that the assumptions that underpin models are communicated clearly; that models are carefully validated; and that the conclusions claimed on the basis of model output do not exceed the information content of that output. Similar concerns apply to the data on which models are based. Given the considerable public interest in these issues, we should demand the most transparent evaluation process possible.

I argue that these considerations motivate a strong open source policy for the modeling of issues of broad societal importance. Our goal should be that every piece of data used in decision making, every line of code used for data analysis and simulation, and all model output should be broadly accessible. Furthermore, the organization of this code and data should be such that any interested party can easily modify code to evaluate the implications of alternative assumptions or model formulations, to integrate additional data, or to generate new derived data products. Such a policy will, I believe, tend to increase the quality of decision making and, by enhancing transparency, also increase confidence in decision making.

I discuss the practical implications of such a policy, illustrating my discussion with examples from the climate, economics, and integrated assessment communities. I also introduce the use of open source modeling with the University of Chicago's new Center on Robust Decision making for Climate and Energy Policy (RDCEP), recently funded by the US National Science Foundation.
The human genome is a marvel of genetic programming. It comprises two copies of approximately 3 billion pairs of its 4 constituent units (“bases”) of DNA (called A, G, T, and C) distributed among 23 chromosomes. It programs the development, from a single fertilized cell, of an exquisitely sculpted organism with 100 trillion cells arranged into a myriad of organs, muscles and bones. It also programs a brain with 100 billion neurons, each with an average of 10,000 connections, with the capacity to learn, to think, to talk, and to remember. Amazing.

Almost as amazing is the pace of change in deciphering this genetic software. It is just 57 years since James Watson and Francis Crick solved the structure of DNA, giving us the iconic double helix. The following decade showed that genes encode proteins, the key functional analog components of cells, through a temporary intermediate called RNA, which became the “central dogma” of the new science of molecular biology.

The field gathered speed in the mid-1970s with the introduction of DNA cloning and sequencing, which enabled decoding of the information, much of which, surprisingly, lay outside conventional genes and was assumed to be junk. These technologies became progressively more sophisticated over the ensuing years, to the point that by 1990 it was feasible to consider the decoding the entire genome. The first draft sequence of the human genome, a composite, was completed ten years later. It involved thousands of DNA sequencing machines and cost several billion dollars. It was a tour-de-force, the attempt of which, let alone achievement, was inconceivable just 20 years earlier.

However, this was just the beginning, and the pace of change since has been dizzying. Over the past few years a beautiful intersection of nanotechnologies, optical technologies and DNA technologies has revolutionized DNA sequencing. The volume of data has skyrocketed, and the cost is dropping like a stone. The latest generation of automated machines can generate 200 billion bases (“gigabases”) of DNA sequence per run (over 60 human genome equivalents), in just a few days.

The cost of sequencing a human genome, done “shotgun”-style on fragmented DNA and assembled by computer, is now down to about $10,000. The amount of DNA sequence being generated and deposited in databases is growing at around 10-fold per annum, which makes Moore’s Law look positively lethargic. New technologies are coming, such as reading sequence by electrical charge disturbance as molecules pass through membrane nanopores, and it is widely expected that the cost of sequencing will fall by at least an order of magnitude in the near future.

The information obtained will transform our understanding of biology, medicine and evolution. There are astounding projects afoot. These include the “1,000 Genomes Project”, the “Human Variome Project”, the “International Cancer Genome Consortium Project”, all of which will contribute to cataloguing the genetic variation that underlies the diversity of the human family and our susceptibilities to common diseases. These projects also include the “10,000 Vertebrate Genomes Project” to sequence every known vertebrate species, or at least a representative of every genus. There will be a fleet of other studies to catalogue the genetic diversity of life in all its forms. Sooner or later, and sooner rather than later, every genome of scientific or practical interest will be sequenced. This will be expanded by data concerning the epigenome (contextual chemical changes to DNA and the proteins around which it is wrapped), which varies during development, environmental exposure and brain development.
function, and the transcriptome (the repertoire of RNAs produced from the genome which are modified in many different ways in different cells), both of which are far more complex that the genome itself.

This is creating huge computational challenges and will cause seismic sociological shifts, although the latter is not yet well appreciated. Increasingly individuals will commission their own genome sequence, along with analysis of this information, which is already being supplied by companies with strong connections to the leading edges of the internet. This is likely to blindside the medical establishment, and overturn many longstanding structures and practices.

The computational challenges include a massive increase in storage, which will soon rise to exabytes, and then zettabytes, as well as requirements for data repositories that are scalable in both disk capacity and I/O performance. There will be equal challenges (and opportunities) associated with development and integration of new software and visualization tools to store and interrogate this data deluge - the “Fourth Paradigm” enunciated by Jim Gray.

The biggest challenge posed by genomics is conceptual. The central dogma of molecular biology, which was framed within a mechanical age, is beginning to look primitive. For 50 years it has been assumed that most genetic information is transacted by proteins, despite the fact that only a tiny fraction (1.2%) of the human genome does so. The major surprise from gene cloning was the discovery that genes are mosaics of protein-coding and non-coding information, the latter of which was considered irrelevant, despite the fact that it is copied into RNA. Just as shockingly, the genome projects showed that we have about the same number of protein-coding genes (~20,000) as a worm with only 1,000 cells, most of which have similar functions, a perplexing finding that has been waved away by invoking combinatorial control.

Where is the information that programs our autopoeitic development and cognitive capacity? Presumably the answer lies in the peripheral information that regulates gene expression in increasingly sophisticated ways and patterns. This raises the more fundamental question of how regulation scales with increased complexity. The answer, which is predicted by one of several models of how operationally connected networks might scale, is quadratic: genome sequencing empirically shows that the number of regulatory genes in bacteria scales as a square function of the total number of genes. Therefore, as a functionally integrated system increases in size, the proportion of the embodied information required to program its assembly and function (to contain and channel the combinatorial explosion) increases. This ultimately imposes a limit on the complexity that such systems can attain, which can only be raised (but not escaped) by changing the nature and architecture of their regulatory structures.

Life addressed this problem, well before we had the capacity to appreciate it. The extent of the non-protein-coding sequences in genomes increases with developmental and cognitive complexity in the animals and, far from being inert, is copied into RNA in dynamic patterns in different cells and tissues. Recent evidence suggests that these RNAs comprise a massive digital (sequence-specific) regulatory system that controls many levels of gene regulation, including epigenetic regulation, by intersecting with generic proteins that interpret and act upon these signals. This system far more genomically efficient, developmentally programmable and evolutionarily flexible, and enabled the increasing sophistication and diversification of complex organisms and the colonization of new niches in the land and air.

Evolution then superimposed plasticity on this system to enable the rise of cognition, which confers the most powerful selective advantage of all, by adaptively inventing biochemical mechanisms to modify RNA in response to environmental signals - which appears to be the molecular basis of learning and memory. It appears that biology discovered the power of contextually adaptable digital control systems a billion years before us, and learnt how to learn. Thus, what was dismissed as junk may well hold the key to understanding human evolution and cognition, and hold important potential lessons for information structures and programming in other domains.
Keynote Abstract V

Advanced Applications of Virtual Reality

Jurgen P. Schulze
Research Scientist, CalIT2,
University of California at San Diego

The focus of the research at the Immersive Visualization Laboratory at the California Institute for Telecommunications and Information Technology is on virtual reality (VR) software applications which leverage the unique features of such systems. Those features include: user tracking, 3D input devices, immersive displays, multi-user capability, high-end graphics hardware. In 2007, we built the StarCAVE, still one of the most advanced virtual reality systems in the world. Since then, we created a tiled autostereoscopic display wall, and a tiled LCD panel based virtual reality system, the NexCAVE. We also made advances in interaction techniques and interaction hardware, like two-handed interaction, multi-user tracking, and brain-computer interfaces.

This keynote will present some of the latest virtual reality software applications we have been working on. This includes the visualization of power consumption and temperature in a mobile data center; photorealistic display of places and objects in VR; multi-spectral image visualization; surface-from-motion with stereo images; and a new user interface for 3D television.
Over the last half-century we have learned a great deal about how to design and use programming languages. One might think that we could design "the perfect programming language" once and for all, but there are good reasons why that is not possible: the concerns of programmers vary, not only by application area but also because of technological progress in the design and construction of computers.

One simple way to categorize programming languages is to divide them into the procedural, functional, object-oriented, and declarative styles. Historically the procedural style came first, resembling cooking recipes with step-by-step instructions; then the functional style, resembling mathematical expressions.

Object-oriented programming became widespread in the 1980s and 1990s with the adoption of the C++ and Java languages; in this style, data objects often reflect or represent real-world entities that communicate and have a kind of independent existence or state.

Declarative programming relies more on statements of relationships to be maintained, more or less automatically, than on descriptions of how to do so; it has been successful in some specialized areas but is not yet mainstream.

For decades scientific applications were predominantly programmed in Fortran, an early (and stunningly successful) procedural language that has gained some functional and object-oriented features but remains fundamentally procedural in practice. More recently, object-oriented languages have also been widely adopted for scientific applications, perhaps more because of the widely available and well-supported infrastructure that supports such languages for business purposes than because of any inherent suitability of the object-oriented style for scientific applications.

A few functional languages have been designed with scientific applications in mind but have not yet gained much traction.

We will argue that two important changes in computer hardware technology (one present and one future), as well as recent progress in software technology, will make the functional programming style especially and increasingly attractive for scientific applications.