Report of the ICTS program "Scientific discovery through intensive exploration of data"

Amit Apte (convener)[*], Ravi S. Nanjundiah (convener)[†], Vijay Chandru[‡], Roddam Narasimha[§], and Spenta R. Wadia[¶]

April 8, 2011

Executive Summary

There has been an explosion of data available for scientific investigations, coming from observations, new experiments, and numerical simulations. At the same time, there are sophisticated models of complex systems, based partially on physical principles, but increasingly also based on the data. The main aim of the meeting was to bring together researchers working on understanding the interplay between data – observational, numerical, and experimental – and the theories and models that need the data, and how this interplay illuminates the scientific questions being investigated – in short, *scientific discovery through intensive data exploration*.

The meeting was a programme of the International Centre for Theoretical Sciences, TIFR, held at the Jawaharlal Nehru Centre for Advanced Scientific Research (JNCASR), Bangalore, during 02-11 February 2011. There were four keynote presentations, and about twenty invited presentations, by eminent researchers such as Tim Palmer, Michael Mahoney, Tony Cass, Ian Foster, Alok Choudhary, Srinivas Aluru, Ravi Kannan, Umesh Waghmare, and many others. They represented the following fields of research which make essential use of large data sets.

1. Computer science and statistical methods
2. Life and health sciences
3. Earth sciences
4. Astronomy
5. High energy physics
6. Materials science and chemistry

There were two panel discussions on "Development and Deployment of Infrastructure for Scientific Computing in India," chaired by N. Balakrishnan (IISc) and on "Computational Genomics," chaired by Niranjan Nagarajan (Genome Institute of Singapore).

[*]TIFR Centre for Applicable Mathematics, Bangalore apte@math.tifrbng.res.in
[†]Centre for Atmospheric and Oceanic Sciences, IISc, Bangalore ravi@caos.iisc.ernet.in
[‡]Strand Life Sciences, Bangalore chandru@strandls.com
[§]Jawaharlal Nehru Centre for Advanced Scientific Research, Bangalore roddam@jncasr.ac.in
[¶]International Centre for Theoretical Sciences-TIFR, Mumbai/Bangalore wadia@theory.tifr.res.in

Recommendations for Implementation

There was a general consensus on (i) the inadequacy of the existing computational infrastructure for research using large data, and (ii) the necessity of a comprehensive plan for developing such an infrastructure, and most importantly, for developing trained manpower. The following recommendations emerged out of the discussions during the meeting.

1. Actively encourage new paradigms of doing science and new paradigms of collaborations between industry and government organisations.
   (a) There is need for computing infrastructure that can be shared and available for use by scientists across various institutes and universities.
   (b) Inter-disciplinary and inter-institutional collaborations on scientific problems using large data could be a focus area for agencies such as NSERB, DST, etc.
2. Encourage the growth of dedicated high-speed networks for the data intensive sciences.
   (a) Connectivity to international networks through collaborative efforts such as the GLORIAD should be pursued.
   (b) Strengthening of existing national academic and research network and creation of new ones should be a priority.
   (c) Local inter-institutional high-speed networks, coexisting with the above, need to be developed and existing campus networks need to be upgraded.
3. Innovation is required in the financial models under which academia and industry can work together.
4. Pilot projects, emulating the success of projects such as the Open Source Drug Discovery (OSDD), should be encouraged. Examples include, but are not limited to,
   (a) National Virtual Archive of data, including mirroring of datasets from around the world
   (b) A community-owned network for sharing large datasets
5. Capacity building and human resource development through short term workshops, and dedicated long term initiatives for masters and doctoral programs.
6. Setting-up of a task force consisting of national and international experts from academia, industry, and non-governmental organisations, to develop a plan to implement these suggestions on an urgent basis.

1 Introduction – the need for data based scientific investigations

The ever increasing computational capabilities, not only the computational speed but also the data storage capacities, have lead to ***an explosion of data available for scientific investigation***. One part of this data is ***observational*** in origin. For example, satellites and extensive observational networks now routinely provide huge quantities of real-time data about the earth system, many large and capable telescopes provide astrophysical observations, and medical records give data about human health and medical treatments of various kinds The other sources of large quantities of data are ***laboratory experiments***. The most notable examples are the Large Hadron Collider (LHC), the genetic and biological systems experiments, and experiments in material sciences and chemistry. The third source of data is relatively new in origin and less conventional than the previous two – it is data obtained by ***numerical experiments, through computer simulations*** of realistic and more complex models of physical systems being studied.

The other repercussion of the increased computational capabilities is that it is now possible to ***study sophisticated models of complex systems***, such as those mentioned above. There are two types of models for such systems – either based on previously well-established physical principles, or based purely on observations of the system. The former of the two types of models need to use data in order to be applicable in realistic situations. The latter of the two types of model are fundamentally dependent on data, without which their very construction will be impossible – the patterns and regularities, or lack thereof, in the observations lead to the models, which eventually are hoped to become part of a body of scientific theory. In both the above cases, ***data give fundamental insights into the properties of the underlying physical system, its functioning and its dynamics***.

This discussion clearly points to a dire need to investigate the interplay between data, used in the broad sense described in the first paragraph above, and the theories and models that need the data, and how this interplay illuminates the scientific questions being investigated – in short, ***scientific discovery through intensive data exploration***. This was precisely the theme of the ICTS program with the same title, organized in JNCASR during 02-11 February 2011.

This report will summarize, in the following section, the proceedings of the meetings including various formal and informal discussions, and in the last section, point to specific ***implementable recommendations for developing infrastructure for scientific investigations based on the use of very large data and peta- and exa-scale computations***, which are fast emerging as a new paradigm in the scientific endeavor.

2 Organizing committee

The scientific organizing committee for this meeting consisted of the following researchers representing a broad spectrum of expertise in various scientific themes described above.

- Amit Apte, TIFR Centre for Applicable Mathematics, Bangalore
- Vivek Borkar, Tata Institute of Fundamental Research, Mumbai
- Vijay Chandru, Strand Life Sciences, Bangalore
- Ravi Kannan, Microsoft Research Labs, Bangalore

- Ravi S. Nanjundiah, Indian Institute of Sciences, Bangalore
- Roddam Narasimha, Jawaharlal Nehru Centre for Advanced Scientific Research, Bangalore
- J. Srinivasan, Indian Institute of Sciences, Bangalore
- Spenta R. Wadia, International Centre for Theoretical Sciences, Tata Institute of Fundamental Research, Bangalore/Mumbai

3 A summary of the proceedings of the meeting

3.1 Scientific themes

The following scientific themes were at the core of this program.

1. Technology and Statistical Methods
2. Life and health sciences
3. Earth Sciences
4. Astronomy
5. High Energy Physics
6. Materials Science and Chemistry

The first one focused on the mathematical, statistical, and technological methods and infrastructure necessary for and applicable to the analysis of vast data. As discussed in the previous section, the main focus of the remaining five themes was the investigation of scientific questions using large data sets and computational resources.

3.2 Keynote lectures

1. Tim Palmer, European Centre for Medium-Range Weather Forecasts and Oxford University, UK, "***Uncertainties in predicting our future weather and climate - from inadequate data to fundamental physics***"
   In this fascinating keynote lecture, Prof. Palmer described the fundamental mathematical and physical issues behind the uncertainties in climate and weather prediction problems. To quote the abstract of one of his lectures at The Isaac Newton Institute for Mathematical Sciences in Cambridge, UK:

   > Putting this new science into practice, however, is not straightforward, and ***will require new computing infrastructure hitherto unavailable*** to climate science. Hence, I will conclude with a plea to the governments of the world. Let's take the current stalemate of opinion as justifying a renewed effort to do all we humanly can to reduce existing uncertainties in predictions of climate change, globally and regionally, so we can move the argument forward, one way or the other, for the good of humanity. This will require a new sense of dedication both by scientists and by politicians around the world: by scientists to focus their efforts on the science needed to reduce uncertainties, and

by politicians to ***fund the technological infrastructure needed to enable this science*** to be done as effectively and speedily as possible.

(emphasis added, Ref:
`http://www.newton.ac.uk/programmes/CLP/seminars/120617001.html`)

2. Ian Foster, Argonne National Laboratory, USA, "***What the cloud really means for science***"
   Prof. Foster discussed how the biggest IT challenge facing science today is not volume but complexity. He argued that establishing and operating the processes required to collect, manage, analyze, share, archive, etc., that data is taking a large part of the scientists' time and killing creativity. With examples from the University of Chicago and ANL, he illustrated that in order to overcome this problem, we need to make it easy for ***providers to develop "applications" that encapsulate useful capabilities and for researchers to discover, customize, and apply these "apps" in their work***, which would have the effect of dramatically accelerating scientific discovery.
3. Tony Cass of CERN, Geneva, Switzerland, "***Worldwide data distribution, management and analysis for the LHC experiments***"
   In this opening keynote lecture, Tony Cass discussed how the Large Hadron Collider at CERN tries to address the key physics issues such as the origin of mass, the nature of Dark Matter in the Universe and precise details of the asymmetry between matter and anti-matter. The talk described the ***Worldwide LHC Computing Grid, designed to effectively distribute and analyze the unprecedented data volumes of 15-25 PB per year***, in particular reporting on the successes of the first year of LHC operation and on some possible future developments.
4. Michael W. Mahoney, Stanford University, USA, "***Algorithmic and statistical perspectives on large-scale data analysis***"
   Prof. Mahoney opened the lecture by comparing two complementary perspectives on data: one of computer scientists who tend to view the data as noiseless and focus on algorithms to speed up the computations, and another of natural scientists who often have, either explicitly or implicitly, an underlying statistical model in mind. He then described the ways in which, in recent years, ***ideas from statistics and computer science have begun to interact for solving large-scale scientific and Internet data analysis problems***, including examples of improved methods for (i) structure identification from large-scale DNA SNP data and (ii) selection of good clusters or communities from a data graph.

3.3 Invited presentations

The other invited presentations of the meeting discussed scientific research that is directly related to or makes use of large data and computations. The talks by Sandeep Sirothia and Yashawant Gupta, discussed challenges in signal processing and computations for astronomical surveys and studies, especially with the advent of next generation telescopes. The talks by Vijay Natarajan, Vipin Chaudhary, Soumen Chakrabarty, and Ravi Kannan illustrated various techniques for visualizing, annotating, indexing, searching, large data sets, as well as specific mathematical methods and hardware platforms for dealing with them. Umesh Waghmare gave an illuminating introduction to the interplay between theory and computations in Chemistry, in the context of the exciting research in designing new materials. A series of talks, by Vijay

Chandru, Gyan Bhanot, Niranjan Nagarajan, Ramesh Hariharan, Rahul Raman, Srinivas Aluru, Andreas Dress, and Jayant Haritsa, spread over the whole duration of the meeting, discussed the development of computational tools and their use in biological research, in cancer biology, genetics, and computational genomics. The talk by G. Bala discussed the challenges in dealing with massive datasets generated by comprehensive numerical climate models.

The talk by Greg Cole of Center for International Networking Initiatives (CINI), University of Tennessee, USA, described in detail the GLORIAD advanced science internet network, which connects scientists in US, Russia, China, Korea, Canada, The Netherlands, India, Egypt, Singapore and Nordic Countries, in order to promote new opportunities for collaboration and cooperation among scientists, educators and students. He described the dynamic networks that change constantly with technology, the multitude of their uses in scientific endeavor, the partners in each of the countries, the possible way forward in Indo-US collaborations using this network, and community building efforts based on such a network. He also discussed the issues of the weakest link, metrics of performance, cyber security, monitoring networks, and the future network requirements for dealing with petabytes of data which is soon going to be very common in many scientific disciplines.

Most of these talks will be available in electronic format (the presentation files as well as videos of the talks) on the ICTS website for this program.

3.4 Panel discussion on Computational Genomics

There was a special panel discussion on "Computational Genomics" following the talk by Niranjan Nagarajan "Data integration in Computational Biology." It was organized to provide a forum for conference attendees and expert panelists to debate and discuss issues regarding computational challenges in biology and genomics and the unique role that Indian engineers and scientists can play in this increasingly data-intensive field. The panelists were Prof. Sowdhamini of National Centre for Biological Sciences of TIFR, Bangalore, Prof. Nagasuma Chandra of IISc, Dr. Hariharan of Strand Life Sciences, Bangalore, Dr. Nagarajan of Genomics Institute of Singapore, and Prof. Aluru of Iowa State University.

The discussions began with a brief introduction of the panelists and their areas of interest - Prof. Sowdhamini and Prof. Chandra being biologists who have increasingly relied on computational tools for their work; Dr. Hariharan providing an industrial perspective and Dr. Nagarajan and Prof. Aluru the perspective of academics who work extensively in computational genomics. The discussions were free-flowing with active audience interest. Some of the topics discussed included, modes of interaction for computer scientists and biologists in the field, role of multi-level modelling ideas in genomics, the lack of adoption of parallel algorithms as a solution to challenges in computational biology, and funding modes in Indian science. Overall, the panelists felt that the discussions were thought-provoking and made the session more interactive.

3.5 Panel discussion on "Development and Deployment of Infrastructure for Scientific Computing in India"

This panel discussion was part of the ICTS program "Scientific discovery through intensive exploration of data" which was held during 02-11 February 2011: Jawaharlal Nehru Centre for Advanced Scientific Research, Bangalore. The panel was convened by (and this report has been

written by) *Amit Apte, TIFR Centre for Applicable Mathematics, Bangalore, Leena Chandran-Wadia, ORF, Mumbai, and Ravi S. Nanjundiah, Centre for Atmospheric and Oceanic Sciences, IISc, Bangalore*.

The aim of the panel discussion was to explore the special scientific computing and infrastructure requirements of scientists in India working on the data intensive sciences. The broad areas include Atmospheric Sciences, Biological Sciences, Physics and Astrophysics. The infrastructure that is being referred to includes 1) hardware, software, and services for scientific computing, modelling and simulation, 2) management of large databases including mirroring of large public databases available in different countries around the world, and 3) dedicated high-speed networks to interconnect all of these.

The panelist are

1. Prof. N. Balakrishnan, IISc Bangalore: Panel Chair
2. Dr. Sharat Babu, CDAC Bangalore
3. Dr. Tony Cass, CERN Geneva
4. Dr. Leena Chandran-Wadia, ORF Mumbai
5. Dr. Gregory Cole, GLORIAD, Tennessee U.S.A
6. Dr. Vasant Jain, GE Bangalore
7. Prof. E.D. Jemmis, IISER Thiruvananthapuram
8. Prof. Nagasuma Chandra, IISc Bangalore
9. Dr. A. Paventhan, ERNET Bangalore
10. Dr. Satyendra Rana, TCRL Pune
11. Dr.Yogish Sabharwal, IBM Bangalore

The panelists and other participants discussed various aspects of computational and network infrastructure with specific emphasis on

- Shared and local Infrastructure
- Role of government bodies
- Role of industry
- Entrepreneurship
- Training and capacity building efforts

The specific recommendations arising out of this discussion are included in the executive summary and a detailed report of this panel discussion is in the appendix 1.

3.6 Participation from students, industry, and government labs

We must stress that there was enthusiastic participation from the industry as well as government labs. All major players involved in High Performance Computing (HPC) or Large Data research such as CDAC, CRL, GE, Infosys, IBM, Microsoft, Intel participated. There was a special session where they highlighted the work being done by the industry in the field of research using large data. A session was organized for the student participants to present the work they are doing in this field and the problems being faced by them. The participants represented a broad spectrum of problems ranging from bio-sciences to oceanography.

4 Disussion and implementable recommendations

From the opening keynote lecture by Tony Cass describing the worldwide computing infrastructure developed by CERN for the data from Large Hadron Collider experiment in Geneva to be available to scientists across the globe, to the talk by Greg Cole describing the GLORIAD network and its utility and potential in connecting the global scientific community, until the last talk by G. Bala describing the peta- and exa-scale computing required for scientific use of the numerical data generated by the next generation of climate models, the discussions during the meeting made it clear that a new paradigm of scientific research based on the use of very large data and complex numerical models needs ***a computational infrastructure on a scale which is much larger than currently available in the country***, as well as development of ***skilled manpower in order to make its effective use.***

Specific implementable recommendations:

1. There is need for computing infrastructure that can be shared and be available for use by scientists across various institutes and universities. At least some of this must be of the highest quality, and it needs to be upgraded frequently to keep up with the fast changes.
2. The access to the infrastructure must be through a unified interface which is easy to use. There must be ways for researchers affiliated to organizations which by themselves may not have adequate computational resources to access a "shared" infrastructure.
3. Networking is a key area that needs immense focus and major thrust. End-to-end connectivity issues need to be tackled, since the weakest (or the slowest) link in a network ultimately determines its performance from the users' perspective.
4. A unified solution needs to be found for dealing with (non-)availability of licensed software or to find open source solutions, for a large section of the scientific community, especially those sharing a common hardware infrastructure.
5. We also suggest development of 'national virtual archives' for data on various topics including but not limited to biology, earth sciences, engineering, social sciences, astronomy, and other sciences. These could be archives where all data generated within the country with public funding could be deposited (after a statutory period during which the researcher who has developed the dataset gets exclusive usage). This could also include data generated by various public agencies through public funding (such as those generated by GSI/IMD etc). This along with good connectivity through NKN could go a long way in improving the quality of research in the country which at times is stunted due to lack of access to good data/connectivity.
6. Development of virtual communities and setting up infrastructure for it could be encouraged, a prime example being the Open Source Drug Discovery (OSDD) project, which very effectively uses the high-bandwidth connectivity to bring together scientists, students, etc. A pilot project that could be taken up is the Earth Sciences Virtual Community, which shares observational, experimental, and numerical data, using common computational infrastructure, to work on inter-related scientific questions.
7. Masters and/or Doctorate courses in "applied computer science" could be developed at various universities, in order to deal with the issue of lack of trained human resources.
8. New methods have to be devised for recognizing the contributions of scientists who create or archive databases of great scientific value. Such methods may include new practices in citation procedures and modification of current evaluation criterion to include

credit for human resource development as well as creating public goods, which include novel computational infrastructure development, new methods of providing access to data, etc. This gives recognition to the time invested by the researchers in these activities.

9. There is a immediate need to set up a dedicated research centre focusing on computational approaches to questions from several basic and applied sciences. This centre could be located in an institution which has expertise in various engineering and scientific disciplines. This could be along the lines of the Computation Institute (a joint institute of the University of Chicago and Argonne National Laboratory), which is "an intellectual nexus and resource center for scholars from multiple disciplines building and applying computational platforms for science," but needs to be tailored to the needs of the country.
10. There needs to be a conscious effort from the scientists as well as funding agencies to promote interactions between computer scientists, mathematicians, and those working in other disciplines, through frequent workshops or working meetings and also through calls for collaborative research proposals. A new multi-disciplinary journal or bulletin, focusing on data intensive scientific discovery, could be started in an effort to disseminate the results of such interactions.
11. The immediate next step could be the setting-up of a task force consisting of national and international experts from academia, industry, non-governmental organisations, to develop a plan to implement these suggestions on an urgent basis.

Appendix 1

Report of the Panel Discussion on “Development and Deployment of Infrastructure for Scientific Computing in India”, February 4, 2011

This panel discussion was part of the ICTS program “Scientific discovery through intensive exploration of data” which was held during 02-11 February 2011: Jawaharlal Nehru Centre for Advanced Scientific Research, Bangalore. The panel was convened by (and this report has been written by) *Amit Apte, TIFR Centre for Applicable Mathematics, Bangalore, Leena Chandran-Wadia, ORF, Mumbai, and Ravi S. Nanjundiah, Centre for Atmospheric and Oceanic Sciences, IISc, Bangalore*.

The aim of the panel discussion was to explore the special scientific computing and infrastructure requirements of scientists in India working on the data intensive sciences. The broad areas include Atmospheric Sciences, Biological Sciences, Physics and Astrophysics. The infrastructure that is being referred to includes 1) hardware, software, and services for scientific computing, modelling and simulation, 2) management of large databases including mirroring of large public databases available in different countries around the world, and 3) dedicated high-speed networks to interconnect all of these.

Panelists:

Prof. N. Balakrishnan – IISc Bangalore: Panel Chair
Dr. Sharat Babu – CDAC Bangalore
Dr. Tony Cass – CERN, Geneva
Dr. Leena Chandran-Wadia – ORF, Mumbai
Dr. Gregory Cole – GLORIAD, Tennessee, U.S.A
Dr. Vasant Jain – GE, Bangalore
Prof. E.D. Jemmis – IISER-Thiruvananthapuram
Prof. Nagasuma Chandra – IISc Bangalore
Dr. A.Paventhan – ERNET Bangalore
Dr. Satyendra Rana – TCRL, Pune
Dr.Yogish Sabharwal – IBM, Bangalore

Summary of the Discussion:

The data intensive sciences require inter-disciplinary and multi-disciplinary skills and therefore need new paradigms of collaborative and cooperative research. Scientists and technologists must work closely together to develop and deploy special purpose infrastructure, just as in the LCG (the LHC Computing Grid) which was created to support the physicists working on the LHC (Large Hadron Collider) in CERN, Geneva.

The technologists at CERN designed, created and deployed the LCG to specifications provided by the scientists and they now also maintain it as a production quality service, available 24/7, leaving the scientists free to focus on the data analysis. What is even more important is that they were able to generalise the software and services created in the LCG into the 'European Grid Infrastructure' and make it available for use by scientists in other disciplines all across Europe (http://www.egi.eu/). Therefore, this separation of 'providers' and 'users' is extremely important since it allows for huge savings in cost through re-use of the knowhow and the creation of shared infrastructure across disciplines.

However, the key difference in India is that unlike in CERN, Indian scientific institutions do not have the skilled technologists on campus who can build and run such production services for the scientists. Therefore we need to be able to compensate for this, 1) by finding innovative ways to work with the technology industry in the short term, and 2) by making sure that all new multi-disciplinary projects have very large training and capacity building components for the longer term. One or more pilot projects would be invaluable for understanding the full scope of these requirements. Also, it may be necessary to create a consortium or management council that can take a holistic view of these activities and provide strategic planning for the future.

Shared and local Infrastructure

The case for shared infrastructure comes from not just the requirements of the disciplines or even the cost savings, but also from exploiting the economies of scale in terms of reducing power usage for a low carbon footprint solution. Adequate local infrastructure must also be provided to scientists who require them. The design of the high-speed networks that connect these together must be dictated by the requirements of the sciences. No generic network connectivity plan, like that of the NKN will be adequate for the purpose. All the software must be written using Free and Open Source, with the active involvement of scientists. Many panelists also suggested putting together a large training and capacity building effort simultaneously, in order to create the manpower that can work on these infrastructure projects.

Role of Industry

The role of private sector industry came in for a lot of discussion. In the past there has been little interaction between academia and industry on infrastructure. However, in the new climate of need and shortage, we must explore and find unique models of working together. Industry can not only help create and manage facilities, but they can also help with training requirements.

The industry participants discussed how they are able to deliver 'computing as a service' to users within or outside their organization. Thus users in GE who need high performance computing are always presented with a uniform interface across the GE research centres around the world, irrespective of the scale of resources they require. The 'back-end' of this service is either managed by a

dedicated team or is outsourced to service providers. In another example, Tata CRL in Pune not only offers compute time on their cluster but also offers services of configuring and smoothly running the scientific applications on their clusters.

Role of government bodies

CDAC is in the process of making their GARUDA computing grid available to the scientific community in India. This is presently a best-effort service. *Unless CDAC is willing to be accountable to the user community, in terms of the quality of service they provide and their responsiveness to feature requests by the scientific community etc., GARUDA may not see much uptake among the scientists, particularly for use in cutting-edge research.*

ERNET has been providing Internet connectivity and some of the associated services to educational institutions for over 25 years now. The new National Knowledge Network will connect about 1500 of the country's top research and educational institutions to each other at speeds of upwards of 100 Mbps. However, the network is only as good as its weakest link and campus networks in even premier institutions such as the IISc Bangalore are not fast enough to allow researchers to make optimal use of the high-speed NKN connectivity.

Entrepreneurship

This kind of high-end infrastructure and cutting edge research has the potential to provide opportunities for many start-ups to emerge. These will help establish the missing linkages from the laboratories to the industry and must be encouraged. Communities of scientists, across disciplines, must be incentivised to come forward and work with industry partners towards this, with flexible funding options from the government. *The present system of giving credit to scientists for only research publications must be replaced with a more broad based evaluation system which will allow them to take up such nationally important projects and deliver on them.*

At the end of the panel discussion Prof. Balakrishnan, the chair of the panel, summarized the discussions by saying that 'One size fits all' types of solutions are not viable options any more. *It is necessary for us to build heterogeneous environments of computing services, as per the requirements of the various sciences, and connect them to each other by high-speed networks.*

An immediate Outcome:

A pilot project for creating a community-owned network for sharing large datasets: An urgent need has been expressed by more than one group of researchers (Atmospheric Sciences and researchers in the OSDD project) to have access to large datasets being held in publicly available databases in the US. Since these databases are very large, typically over 1 Petabyte, they require: 1) high speed international connectivity to download and mirror the data, and 2) high-speed dedicated network connectivity between the mirror site and the users of the data who may be located at various institutions throughout the country.

This network must be managed by the scientists themselves in order to have the necessary control on access to the data as well as its security. Such community built networks for research and education have many precedents within the United States (the National Lambda Rail, the ESNET, the Ultra Science Net among many others). In the International arena, the GLORIAD network, built by communities of scientists in various countries, has been providing support for collaborative research to millions of users from universities, national labs and science facilities in many countries. GLORIAD is a true community built and owned network, funded in part by the respective countries and in small part (10%) by the NSF in the US.

India has recently been offered the opportunity to connect to GLORIAD through the generous donation of two 1 Gbps links for a period of one year (one each to Amsterdam and Hong Kong and thence to the US), by Tata communications. We propose to make this 1 year the duration of a special pilot project, in which we propose to use the free international connectivity to download and mirror databases in the two subjects and make the data available, as a service, to all researchers in the respective fields. The equipment necessary to connect the GLORIAD network to institutions in India is already in India, having been provided by GLORIAD for use in this 'Taj' expansion of GLORIAD into India[1], which was triggered by the donation from Tata communications. All that remains is to provide for the local infrastructure at the participating institutions and for interconnecting them with high-speed networks. For this, we intend to apply for funding to the newly constituted NSERB.

The funding required by the pilot project must cover: 1) the servers, software and hosting costs required to mirror these two large databases, 2) the costs of the dedicated high-speed network connectivity to this data by users at their host institutions, 3) the local infrastructure (hardware and software) at each individual institution, and 4) the costs for a suitable, sizeable, training and capacity building effort. Such a pilot project will help clarify the requirements for future projects.

Recommendations:

Actively encourage new paradigms of doing science and new paradigms of working with industry and with government organisations. This includes collectively building out the extensive infrastructure that is required for this and making them available as shared infrastructure.

Encourage the growth of dedicated high-speed networks for the data intensive sciences. For this, the unused dark fibre in the country must be made available for use by scientists in their special science projects. In the US it was the advent of the 'dark fibre law' that gave universities the opportunity to lease unused fibre and become Gigabit campuses overnight. We must do the same in India, where there is over a million Kms of unused 'dark' fibre. The development

1http://articles.timesofindia.indiatimes.com/2011-03-29/chennai/29357309_1_iit-m-m-s-ananth-iit-madras

of such networks will also provide the NKN with some much needed competition which will help keep its services responsive and nimble, very much like what the advent of the National Lambda Rail did for Internet2 in the US.

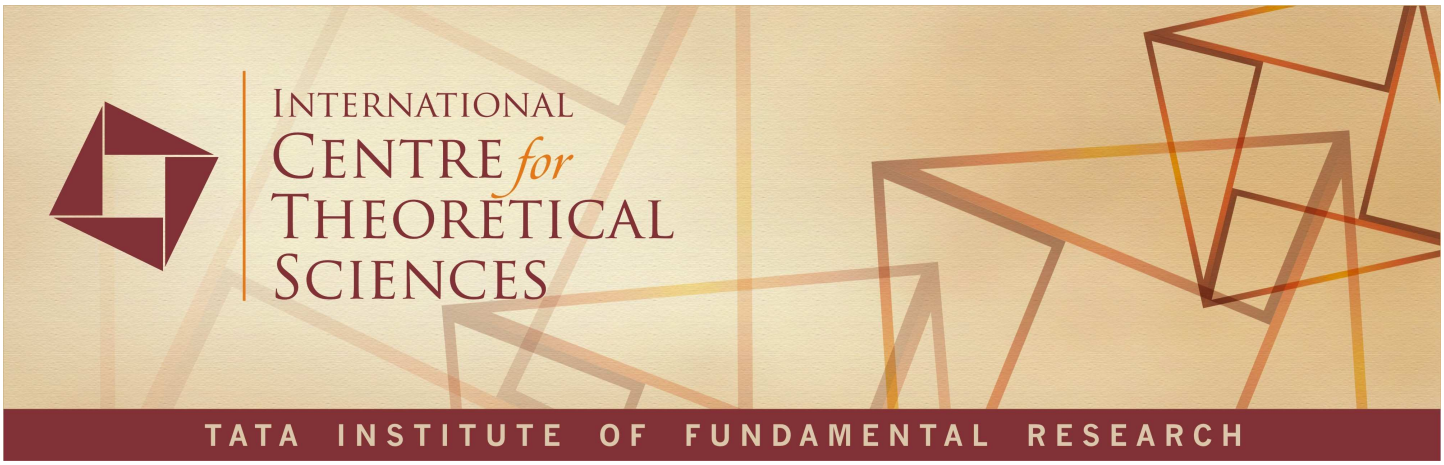
Another way in which unused fibre can be made available to the Research and Education community is to extend the use of the Universal Services Obligation Funds of the Indian Telecom providers (presently used to reach mobile phone connectivity to rural areas) to educational institutions. This would be similar to the E-Rate[2] program in the US through which schools and libraries get Internet access at affordable costs. These are strategic decisions that we must make as a country and they require debate and discussion with the Telecom Ministry and other stakeholders. ***The move towards Green IT and low carbon footprint solutions for technology is relatively recent in the world today and India has the strategic opportunity to bypass the older technologies and move to the front of the line, provided we are willing to take it***.

Adequately and flexibly finance a number of pilot projects **which will seek to:**

- Identify the detailed requirements for technology support for different types of computational sciences.
- Identify and launch several capacity building efforts in collaboration with industry.
- Generalize the infrastructure requirements across multiple disciplines and help build communities of scientists.

Innovation is required also in the financial models under which academia and industry can work together. The pilot projects must seek to establish these too.

2http://www.universalservice.org/sl/about/overview-program.aspx



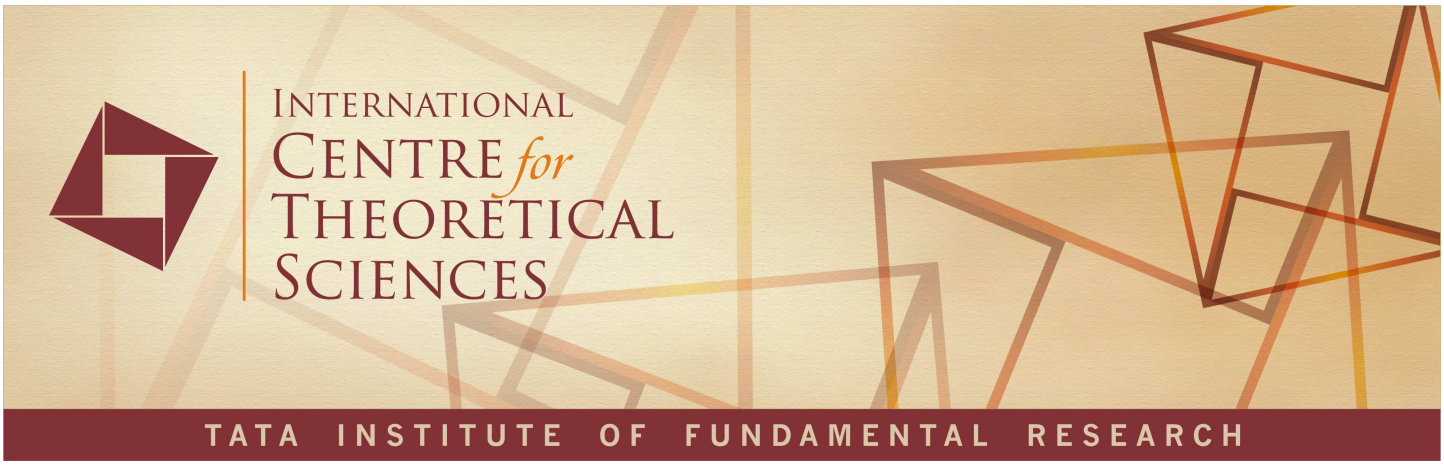
Appendix 2

Schedule for the meeting on

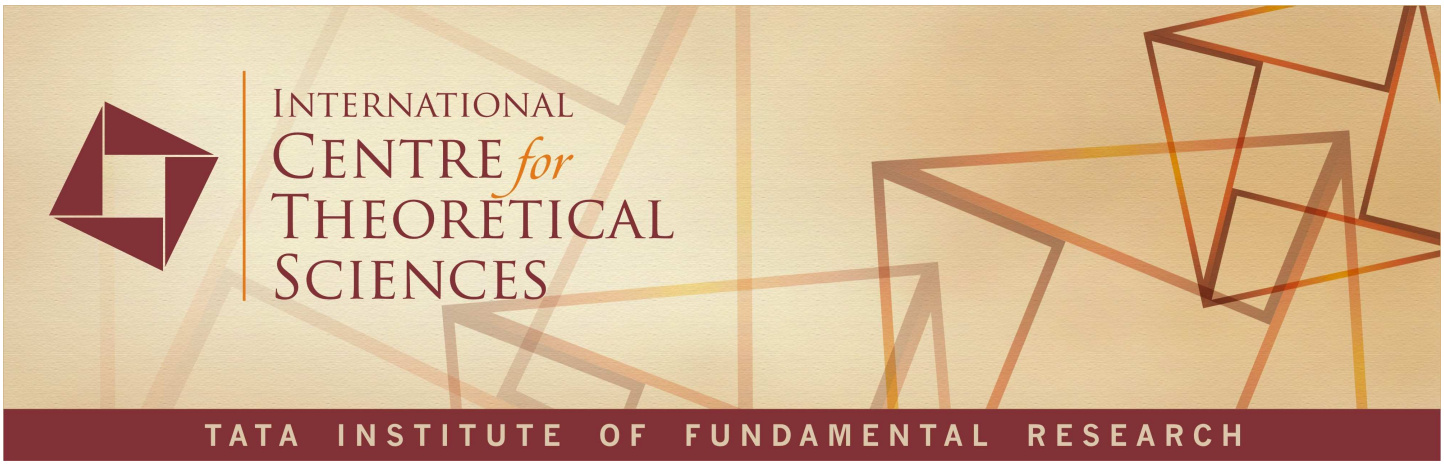
Scientific discovery through intensive exploration of data

02-11 February 2011: Jawaharlal Nehru Centre for Advanced Scientific Research, Bangalore

Wed 02 Feb		
10:00-11:00	Opening Session	
	Amit Apte	Welcome remarks
	R. Narasimha	Conference overview
	Spenta Wadia	ICTS: A new initiative in Indian science
11:00-11:30	Tea/Coffee break	
11:30-12:30	*Tony Cass, CERN – Keynote lecture*	Worldwide Data Distribution, Management and Analysis for the LHC Experiments
12:30-14:00	Lunch	
14:00-15:30	Vijay Natarajan, Indian Institute of Science	Topology-Based Methods for Visualization
15:30-16:30	Tea/Coffee break	
16:00-17:00	Vijay Chandru, Strand Life Sciences	Intelligence in the Era of Data Intensive Life Sciences
Thu 03 Feb		
10:00-11:00	*Michael Mahoney, Stanford University – Keynote lecture*	Algorithmic and statistical perspectives on large-scale data analysis
11:00-11:30	Tea/Coffee break	
11:30-12:30	Soumen Chakrabarti, IIT Bombay	Annotating, indexing and searching the Web of objects, attributes and relations
12:30-14:00	Lunch	
14:00-15:00	Gyan Bhanot, Rutgers University	How the availability of large data sets is enabling discovery in clinical cancer biology and population genetics: two case studies
15:00-15:30	Tea/Coffee break	
15:30-17:00	Soumen Chakrabarti, IIT Bombay	Learning to rank in vector spaces and social networks



Fri 04 Feb		
10:00-11:00	Ravi Kannan, Microsoft Research	Sampling on the fly
11:00-11:30	Tea/Coffee break	
11:30-12:30	Umesh Waghamare, JNCASR	Multi-scale modelling and simulations of ferroelectric phase transitions
12:30-13:45	Lunch	
13:45-14:45	Yashawant Gupta, NCRA, Pune	Some Challenges in Signal Processing and Computing in Astrophysics
14:45-15:00	Tea/Coffee break	
15:00-17:00	A Panel Discussion	Development of Computational Infrastructure in India: A Discussion
17:00-18:00	Reception	
18:00-19:00	P. P. Divakaran	Yuktibhasha and the Origins of Calculus
Sat 05 Feb		
Sun 06 Feb	Free days (possible excursion to Mysore)	
Mon 07 Feb		
10:00-11:00	Alok Choudhary, Northwestern University	High performance data mining: an essential paradigm for knowledge discovery
11:00-11:30	Tea/Coffee break	
11:30-12:30	Niranjan Nagarajan, Genome Institute of Singapore	Biology as a data-driven science: from a trickle to a flood
12:30-14:00	Lunch	
14:00-15:30	Umesh Waghamare, JNCASR	Multi-scale modelling and simulations of ferroelectric phase transitions
15:30-16:00	Tea/Coffee break	
16:00-17:00	*Tim Palmer, University of Oxford and European Center for Medium-range Weather Forecast – Keynote lecture*	Uncertainties in Predicting our Future Weather and Climate - From Inadequate Data to Fundamental Physics
Tue 08 Feb		
09:00-10:00	*Ian Foster, Argonne National Lab – Keynote lecture*	What the cloud *really* means for science
10:00-10:30	Tea/Coffee break	
10:30-11:30	Ramesh Hariharan, Strand Life Sciences	Sequencing the Genome and What it tells us
11:30-12:30	Free time	
12:30-14:00	Lunch	
14:00 onward	Visit to the ISRO Satellite Centre (ISAC)	



Wed 09 Feb		
10:00-11:00	*Ram Sasisekharan/Rahul Raman, MIT – Keynote lecture*	Data Management, Integration and Mining Strategies for Glycomics - A Rapidly Evolving Paradigm in the Post-Genomics Era
11:00-11:30	Tea/Coffee break	
11:30-12:30	Sandeep Sirothia, NCRA, Pune	Astronomy: A multidimensional view
12:30-14:00	Lunch	
14:00-15:00	Niranjan Nagarajan, Genome Institute of Singapore	Data integration in computational biology - can we be hypothesis free?
15:00-15:30	A Panel Discussion	Future trends in computational genomics
15:30-16:00	Tea/Coffee break	
16:00-17:00	Srinivas Aluru, Iowa State University and IIT Bombay	Next-gen sequencing: Data intensive computing in Biosciences
17:15-18:15	Research presentations from IBM, Intel, GE, Infosys	
19:00 onward	Conference dinner	Venue: Jawahar Visitors House of JNCASR, Gymkhana Campus of Indian Institute of Science, Bangalore 560 012, Phone: 2293-2499
Thu 10 Feb		
10:00-11:00	Andreas Dress, Shanghai Institutes for Biological Sciences	Topological Proteomics
11:00-11:30	Tea/Coffee break	
11:30-12:30	Jayant Haritsa, IISc	Mutating Database Engines to be Bio-friendly
12:30-14:00	Lunch	
14:00-15:30	Sandeep Sirothia, NCRA, Pune	TIFR GMRT Sky Survey - A Case Study
15:30-16:00	Tea/Coffee break	
16:00-17:00	Vipin Chaudhary, Computational Research Lab Pune	Data Intensive Computing Architecture and Discovery Initiative
Fri 11 Feb		
10:00-11:00	Jayant Haritsa, IISc	Indexing Techniques for Biological Data
11:00-11:30	Tea/Coffee break	
11:30-12:30	G. Bala, CAOS, IISc	Climate modelling and the challenges in dealing with massive datasets
12:30-14:00	Lunch	