

Using Computer Science to advance each of the e-Disciplines

Professor Jon Patrick and Associate Professor Sanjay Chawla,
School of Information Technologies,
University of Sydney.

Summary

This response to the NCRIS 2008 Roadmap exposure draft consultation is intended to contribute generally to the Roadmap's view of eResearch, and particularly to offer some answers to the questions (p 62) on priorities for investment in infrastructure to support ICT and computer science research.

Major points include:

- a need for greater understanding and recognition of the vital role of computational modelling as a unifying, cross-disciplinary approach which is best researched holistically rather than piecemeal in individual capabilities;
- the importance of reporting on and preserving the computational models which embody the underlying science far more than do the associated data;
- a proposal for a National Centre for Text and Data Mining as infrastructure both for fundamental information technology and computer science research, and to provide active support for NCRIS capabilities where data analysis is required.

Overview

E-Science is conceived of as a revolution in the way we conduct Science by exploiting the use of ICT to connect scientists to share knowledge including data, models and information. However the role and experience of the computer science discipline, as distinct from general purpose IT, has not been identified as a contributor to setting up a knowledge management infrastructure for e-science. This essay concentrates on issues around knowledge management in e-Science. The argument presented in this paper is that the contribution of computer science is under-represented in e-Science policy. As a result it is subsequently neglected in the role it should be given in the advance of e-Science as it is the principle source of expertise for the types of things scientists want done in computational thinking and modelling in the past and want to do in the future.

This essay addresses the issue of the appropriate support that is needed for ICT research to serve the needs of all eResearch disciplines. It argues that a highly generalised representation of eResearch to cover all disciplines gives insight into how ICT can give broad spectrum support to the computational modelling needs of the disciplines. It also argues that insufficient attention is given to the intrinsic characteristics of modern research methods, that is, the science is captured in the computation models not the data, and that this aspect is for the most part overlooked in the Draft Roadmap. Strategies for addressing this oversight and initiatives to support computational modelling across all disciplines are discussed.

Up to now the ICT needs of various large and influential project groups have been served for the most part by a piecemeal approach to ICT funding. If eResearch is to be supported in a systemic manner as indicated in the Draft Roadmap then the task has to be investigated through systemic criteria and not piecemeal criteria.

The 20 or so pages of overview in the Roadmap document comment on ICT needs on nearly every page (pg. 9-29). Hence if we are to view ICT as systemic throughout e-research disciplines we need systemic variables to organise our thinking about how to provide the required services and support. The individual discipline sections of the report show that all disciplines express the need for computational modelling without any real specificity. One example of this need is expressed in the Marine Environment section "Enhanced data management and analysis tools, including modelling capacity" (pg. 38). Hence we are left with the conundrum of determining how to support all players at once when they have perceptions of individual and idiosyncratic computational modelling needs.

Beyond the issue of supporting computational modelling there is the issue of equity of support. The synchrotron in Victoria is said to support the research of 300 scientists with the potential to support another 3000 secondary users. This suggests that a per capita investment of \$0.5 million has been made to benefit the research efforts of 300 people. It is unclear how many scientists we have in Australia but guesstimates from colleagues suggest something of the scale of 30,000. The arguments presented in this essay propose that the appropriate investment in ICT that is discipline neutral would potentially benefit the research of all scientists, that is an investment of, say \$30 million would provide a per capita cost of only \$1,000 constituting a very high value for that investment, and as well, give benefit to the greater part of the research community who will never benefit from NCRIS investments otherwise.

The draft roadmap makes extensive reference to issues of ICT support for research but it presents no organising variables applicable across all e-research disciplines for defining the nature of the ICT needs. This gap makes it much more difficult to identify and understand the infrastructural needs of ICT itself but also the valid generalisations we can make about ICT that are usable in defining a policy.

This paper offers a number of organising variables so as to construct an argument for systemic support for certain aspects of ICT research itself which at the same time support the ICT needs across all disciplines. We argue that support for certain types of ICT research is concomitantly supporting the ICT infrastructure needs of ALL eResearch disciplines. We argue that the problems are so significant we cannot provide ICT support for eResearch without doing research in ICT itself, which thereby makes that research infrastructural.

The aspect of ICT support we discuss in this essay is support for computational modelling as done by all eResearch disciplines. The organising variables that we have identified relevant for this exercise are:

- the type of research defined by its *a priori* knowledge on a continuum of black box modelling to white box modelling (see p7 below), with grey box modelling forming a continuum between the two ends of the scale.

- the scale within the given model, of the variable space (small to massive) and the data points space (small to massive)
- the roles of staff for the tasks to be performed: user, data collector, data modeller, analytics methods modeller, software engineer.
- data verses computational models as representing the science.

This essay presents a definition of a taxonomy of computational modelling for understanding the different roles that computer science has played in the past. While many arguments for the value of computer science are presented based on the scientific use of black-box modelling such as machine learning, the scope of the arguments is applicable to the continuum of modelling strategies encompassed by the complete taxonomy, for all sciences and other disciplines. This proposal also argues that the emphasis on “data preservation” both in the UK and Australian e-Science proposals is severely imbalanced and gives entirely inadequate attention to the more important software processes of preserving data models and managing experimental programs. We argue that the Science is in the models, even more than it is in the data, and this is where the majority of the focus of e-Science ICT R&D should be concentrated. The proper development of the two software processes in e-Science software suites require the dual skills of computational thinking (algorithmics) and software engineering, expertises held at their most sophisticated, principled and competent level by computer scientists.

The essay makes a number of tangible proposals, of which the major ones are:

- Setting up a National Centre for Text and Data Mining, with the purpose of doing research and development into methods for the efficient harnessing of black-box computational modelling techniques to support all eDisciplines.
- The Centre to use a 5-role model of participant functions in black-box computational modelling communities.
- Design and implement active learning methods as an integral part of black-box computational modelling.
- A virtual repository for scientific models and data be created composed of local repositories assembled by the organisations participating in public research funding programs.
- All disciplines tackle their workforce supply problems by promoting their undergraduate population to do either major or minor study programmes in computer science concurrently with their discipline of interest.

Introduction

The UK Picture

Some of the broader ideas in this essay are present in different ways in literature already produced in the UK to describe e-Science, namely the E-Infrastructure Strategy For Research especially in the report of the Data and Information Creation Working Group (Bicarregui, 2006, pp13-14,17). However this essay goes further than the high level generalisations in these reports and we propose specific methods for knowledge management of e-Science content. This proposal does not cut across the developments in other infrastructure building strategies such as grid computing, which is concerned at solving the classical problem of complexity, that is, the time-space trade-off for large scale problems. Rather we propose methods to support a

higher level human processing for science.

These reports also pay a great deal of attention to the topic of preservation and retention of digital records but treat it from a Library Science perspective seeing their service to science as preserving their “data”. The only concession that can be found in recognising other computational artefacts in e-science is the off-handed remark

“Similarly, for complex experiments, retaining the model of the experimental system may be as important as the results.” (Beagrie, 2007, pg. 8)

In the same report the section on Data Sharing is reproduced here *in toto*, indicating no appreciation of the importance and role of computational models as one of the preservation needs and experiment management systems for process control and knowledge sharing. This comment is establishing a position for the collectors of irreproducible data and applying it inappropriately to all of science of which it is the smallest portion:

“3.5 Data sharing and citation

Citation of data will have become mainstream alongside citation of literature, leading to much more data-led research and new types of science. The academic reward system will provide appropriate credit and recognition for data contributions. A much improved understanding of the real requirements of different disciplines will lead to a cultural change in the attitude towards data sharing, licensing and automating access rights, which will lead to fruitful interactions within and between various disciplines and sub-disciplines. In addition, a developed and interoperable infrastructure will be in place, nationally and internationally, which focuses on access and re-use of data.

Much larger scale interoperation of data resources will be available, easily discovered and seamlessly used – across data types – across the lifecycle of data – across silos of data – and in the context of the broader scholarly knowledge cycle. Automatic tools for semantic information import and export, autonomic curation (e.g. agents) and provenance capture will be deployed. All types of multimedia will be more easily indexed and searched than today.”

(Beagrie, 2007, pg12)

The Australian Picture

The picture in Australia is similar to that in the UK. There are some references to data management and computational modelling but little appreciation of the importance of experiment process management and more generally of computer science to broad scale success in e-Science from many of the separate sub-disciplines. We advocate recognizing the common themes, and expanding the vision to focus on the value to science of sharing, and jointly analysing, the computational models as well as the data. Relevant comments from some of the NCRIS Expert Working Groups (EWG) in the Roadmap discussion paper were as follows.

The Frontier Technologies EWG states that it “has as one of its goals improved data management and has made the comments that the PfC capability has a focus on infrastructure and supporting services to allow the research community to collect, share, analyse, store and retrieve information , and acknowledges the importance of

this capability." This has no reference to the fundamental processes of e-Science, that is computational modelling, experimental processes and knowledge management. They even go on to state that these things are not required in saying: "With regard to the interactions between PfC and the other NCRIS capabilities the Group is of the view that the 'generic' ICT services that are broadly applicable across the disciplines have so far been adequately considered." Reference is made to the need for computational modelling in the area of nanostructures, but without showing any appreciation for the type of mechanisms and skills by which this can be accomplished.

The Humanities, Arts & Social Sciences EWG recognize the need to use "high performance computing to model and simulate the complexity of historical and cultural phenomena", but show no assessment of their modelling and knowledge management needs.

The Promoting and Maintaining Good Health EWG refer to the need for "large scale physical and ICT infrastructure for storage management and retrieval of specimens, biochemical and genetic analysis of specimens and exchange and analysis of complex genetic and health information." Like the Frontier Technologies EWG they make a general assertion about the need for ICT without identifying their specific needs in saying "Any bioinformatics infrastructure will be dependent on the enabling ICT platforms - networks, high performance computing, data services, and collaborative services." **We would also draw your attention to the fact that their data from clinical contexts is more than 90% text and their science is embedded in the published literature so that they have a significant need for Natural Language Processing (NLP) systems. The UK has responded to these needs by creating the National Center of Text Mining (www.nactem.ac.uk) with an emphasis on studying biomedical texts.**

The Safeguarding Australia EWG have got closer to expressing specific needs than any of the other groups in saying "there is a strongly related need for a national Repository to collect, validate and standardise the various models developed across Australia that could be reusable in different contexts and provide opportunities for these to be used in a collaborative workspace environment."

The ICT EWG who might be expected to recognise and assert the importance of computer science and software engineering to e-Science only makes half the case and fails to assert the importance of the critical issues of computational modelling and experiment process management in their submission to the NCRIS review. They focus mainly on data in stressing the need for "the development of a national strategy for the management, description and preservation of research data". They do give attention to the manpower problem of the future : "the missing cadre of professionals with deep research and deep ICT skills."

The National e-Research Architecture Taskforce (NeAT) forms part of the existing NCRIS PfC 51.16 component. Within the documents of NeAT there are many substantial statements asserting the need to create "tools and services" for e-Research including very supportive extracts from the PfC Investment Plan, and arguments against piecemeal approaches to developing NCRIS capabilities.

However the Minutes of NeAT meetings show that so far, the ICT needs are being viewed through a lens for each specialty in the NCRIS plan, rather than through a computer science lens that can see across the needs of individual groups. We advocate thinking about the nature of computational modelling as a technology applicable across multiple disciplines or the classes of white-box, grey-box and black-box modelling and the consequences that brings to e-Research planning. **Computer Science brings the key viewpoint that computational modelling can be researched and engineered across disciplines in a number of dimensions in a way that makes a greater contribution than supporting individual disciplines doing it *in vacuo*.**

With the emergence of computational modelling making a major contribution to many sciences in processing both text and numerical data, overlooking computer science's contribution has the potential to create deficits so significant that at best e-Science will be a success only for the most aggressive and ICT-expert scientists and leave the rest of science and the social scientists on the sidelines.

e-Science's need for Computer Science

Computer Science brings to the knowledge management task of creating an e-Science revolution two important expertises: computational thinking or algorithmics, and software engineering. Computational thinking is needed so that scientists can express their scientific knowledge in operational ways in order to explore the concomitant behaviours of their models and data with a rich world of readily available tools. Software engineering is needed to build the e-science experiment management environment so that the processes of experimentation run efficiently and satisfy all-comers. We distinguish the difference between software engineering — the discipline of researching and crafting the optimal computational solution to a complex problem — and the trade of programming, where the expertise is in using a programming language and not in engineering the solution.

The four major issues of science research relevant to maximally exploiting ICT are:

- Data Collection,
- Computational Modelling,
- Complexity,
- Knowledge Sharing, and
- Declarative Languages

Computer science has a very important role to play in these latter three spaces if they are to achieve broad spectrum large scale success. Data collection is a well developed and well used field for any science. Computational modelling is in a variable state with significant maturity around engineering disciplines and knowledge rich problems, but in its infancy for problems with poorly developed models with a large variable space, where machine learning methods are the only mechanisms available for modelling. Complexity, on the other hand, is being enhanced by advances in processor architecture and distributed computing and is beyond the subject of this essay. Electronic knowledge sharing and management are well established but radically under performing for their potential.

Science as with most Research & Development disciplines relies on IT to support the collection of data. However IT is far more important in the two stages of: *analysing data*, i.e. discovery to reveal previously unknown structures in data, and in *modelling data* where representation of processes, whether natural or engineered, are understood to such a degree that scientific hypothesis testing is performed as the norm. Computer Science is an important discipline for improving the general availability of data analytics & modelling through its development of computational thinking methods or algorithmics in a properly engineered environment. Furthermore, it should take the lead in promoting the wide scale development, delivery and adoption of this technology as the backbone of IT in e-Science. However, it is not sufficient for Computer Science to be merely a passive provider of techniques and algorithms for reuse. Rather it has to directly support the construction of a complete edifice that engages scientists and supports their processes of knowledge management, sharing and discovery. It can do this by constructing software workbenches for creating and managing permanent repositories that truly enable the sharing and reviewing of research data, and **more importantly do likewise for theoretical models, in ways that provide high levels of access and interactivity for the scientists and increase productivity and the auditability of the science.**

One argument of this essay is that the computational models created by scientists are their actual science, and their data is often only ancillary and in many cases discardable. The major focus for preservation and sharing should be on the creation and testing of computational models and the management of the computational aspect of the experimental process through systems that are properly designed and implemented by experienced and well-trained software engineers.

Most scientists are familiar enough with programming to understand the nature of the task of programming by writing in a procedural language like Java or C. However few have an understanding of the difference between a Procedural Language and a Declarative Language and the important efficiencies the latter provides for the task of investigating structures in collections of data. A good example of the power of a declarative language can be seen in our work with the Intensive Care Unit at the Royal Prince Alfred Hospital. The ICU records all data collected about a patient in a commercial clinical information system named CareVue. It has simple retrieval functions and reporting suited to retrieving data about a single patient. That is, it serves the point of care work satisfactorily, but it fails to provide any support for investigating research questions about a collection of patients or to compare individual patients. We have designed and implemented a declarative language for asking ANY ad hoc question of the ICU's data stores, including data held in the prose of the patients' notes. The language is called the Clinical Data Analytics Language (CliniDAL). It enables staff to use their own medical terminology in their own (restricted) natural language to ask a question about a group of patients or express an hypothesis between two groups of patients and have the answer computed without any intervening manual process. The strength of CliniDAL is not just in its support for the ICU but the fact that it is a completely generalised Declarative Clinical Language and it can be bolted on to ANY clinical information system and provide the same investigative power. To achieve such processing using a procedural language is beyond any sensible limits of effort and needless to say unnecessary given the declarative language paradigm is available for future

utilisation. Constructing specialist declarative languages for classes of e-Disciplines represents a major way forward for increased scale and variety of research that will be accessible to all e-Disciplines.

Computational Modelling as the most important role for ICT in e-Science

The concept of computational modelling has been present in computer science since its earliest inception. In its modern form it can be seen as a continuum stretching from modelling in the context of a knowledge rich expertise such as engineering simulation to modelling data sets drawn from environments with absolutely no *a priori* knowledge where scientists can only use data mining and general purpose machine learners. We wish to characterise the former as **white-box modelling** which is well represented by the work of simulations of electro-mechanical systems with a highly defined knowledge base such as aircraft design. An example slightly further towards a world of greater uncertainty is given by the disciplines based around the built environment where engineers and architects wish to build scale prototypes to model the effects of design under the stresses of nature, say for example assessing environment control in a building.

Black-box modelling is the situation we have when a large number of observations have been collected with no knowledge of the generating model behind the data, e.g. collection of financial scam documents, or a large population of medical disease data. This is true of a lot of data mining tasks in fields such as the social sciences, and more specialised areas such as natural language processing. In these problem spaces there are a very large number of variables where the scientists have poor understanding of how they relate and they rely on having to interpret very low correlations which are nevertheless important (this is very true of language data). “Machine learners” best represent this type of modelling, which is a field that has emerged from the broader discipline of Artificial Intelligence. Machine learners are algorithms that exploit the statistical distributions observed in the data to infer a structure of a basic classification process, e.g. a decision tree. The chosen machine learner type (e.g. decision tree, or support vector machine, etc.) is firstly *trained* with a set of data items for which the true classes are known, then *tested* on unseen data or data not used in the training process, and then improved by repeating many times the test and train process in a feedback cycle based on recognising and remodelling associations made by the learner. In this way the machine learner process mimics the scientific process itself to create and later improve the representation of the underlying processes.

In between black-box modelling and white-box modelling we use the term **grey-box modelling** to represent those disciplines that have a certain amount of knowledge but still a large number of variables of unknown relationships. A good example is models of chemical pathways in biology and their relationships with drugs for disease treatment.

The issue for ICT support for this continuum of computational modelling is the question of what exists for experiment management and what can be created to

improve the productivity of this wide range of activities. There is a very good example of knowledge management in the world of white-box modelling, with a long and effective campaign generated from high levels of computer science and software engineering research and development. This is the environment created in the OpenModelica project (www.openmodelica.org) project at Linköping University, Sweden, based on the *Modelica* language (www.modelica.org/) over 10 years. This software environment contains a very large amount of knowledge about electrical and mechanical engineering and represents an improvement in the speed and sophistication of the modelling these engineers can do in their daily work. Even greater success is in the work of John Pople who won the Nobel Prize for Chemistry in 1998 for the chemistry simulation suite called GAUSSIAN. Although the algorithm does provide for an *ab initio* starting point for its calculations, hence implying it is model free as in black-box computational modelling, this is a ruse as the system uses a great deal of chemistry knowledge to progress its calculations. It is quite unclear as to the extent which these knowledge management and simulation systems support the management of large scale programs of experiments.

In the arena of grey-box modelling there is nothing to compare with the Modelica environment but there is significant work being pursued by a team from the Universities of Manchester and Southampton, UK, on developing generic models of workflow to support experimental scientists particularly in the biological sciences. The best example is their MyGrid (www.mygrid.org.uk) open source project in the UK aimed at significantly enhancing the exploitation of grid computing. Taverna is a major sub-component of MyGrid: it is "an open-source workflow tool which provides a workflow language (scufl - Simple Conceptual Unified Flow Language) and graphical interface to facilitate the easy building, running and editing of workflows over distributed computer resources. When we talk about 'workflow' in myGrid we mean the specification and execution of ad-hoc in-silico experiments using bioinformatics resources. A workflow-based approach allows the e-Scientist to describe and enact their experimental processes in a structured, repeatable and verifiable way". It includes the spin-off system myExperiment which is a Virtual Research Environment that enables "colleagues to share digital items associated with your research — in particular it enables you to share and execute scientific workflows." Declarative Languages represent one of the tools useful in the continuous feedback pathway of oscillating between black-box and white-box computational modelling as is performed by the most sophisticated e-Researchers.

In the area of black-box modelling there are even fewer attempts to tackle the management of experimental processes. Black-box modelling is computational work with little contribution from models about the source of the data. Individual scientists and computer scientists working in grey-box and white-box disciplines often come into possession of some "interesting" data and use black-box modelling to get some basic understanding of its behaviour. However computational linguistics uses black-box modelling as its major investigative paradigm particularly as in the last 15 years the discipline has moved from compiling rule based knowledge to exploiting statistical properties of language. Their experimental processes require the continuous shaping and honing of their attribute sets for input into a machine learner.

We know of only one attempt to build a software environment to manage the process of large scale continuous experimentation using machine learners to deliver an industrially robust NLP system for end users, and that is the Scamseek system (www.acs.org.au/news/100805.htm). Scamseek was developed for a commercial user, but it does not matter whether the user is a commercial enterprise or a learned scientist they demand the same requirements from the computer scientist and software engineer. No open source system is available that delivers the wide range of services needed to manage a large regime of black-box modelling experiments, and we know of no such proprietary product either. The community of black-box modelling scientists needs the computational thinking of computer scientists and the design and creative skills of the software engineers to build a *Modelica* equivalent for their work processes.

In conclusion, collaboration between the IT experts of the three communities of modelling will bring a great deal of efficiency to the larger tasks facing the grey-box and black-box modellers, but they still have their own challenges to confront that are not part of the world of white-box modelling. No matter what happens in the future, without the combined skills of computer scientists bringing their computational thinking and the software engineers their expert praxis of design and construction the future of ICT contribution to the national infrastructure will be entirely underdone. The greatest losers will be the wider advanced industrial, scientific and R&D communities.

Proposal 1

e-Science requires a general source of expertise for black-box computational modelling as part of its IT infrastructure. In the UK a [National Centre for Text Mining](http://www.nactem.ac.uk), (www.nactem.ac.uk) has been set up with a narrower brief. We propose that a National Centre for Text and Data Mining (NCTDM) be created to cover the full spectrum of disciplines needing to use black-box computational modelling while at the same time leaving white-box and grey-box modelling to be driven by their individual communities of expertise, but able to call on the developments of the NCTDM by collaborating on their projects. Its role would be to act as the lead organisation in creating:

a programme of open source software development

- **for the research, development and adaptation of new methods for the analysis of text and data.**
- **for the care and sharing of scientific data and models of both text and data in the national repository**

The Computational Processes of Experimentation using Black-Box Modelling

The preceding argument is applicable to the broad issues of management of the national e-Science initiative and to other disciplines searching for the same knowledge management gains. However consideration must be given to wider issues of black-box computational modelling for scientific praxis.

Current practice in the use of machine learning generally involves scientists

producing a data set, extricating a feature set usable by a machine learner of choice, then repeatedly varying the feature set until an optimal learner is identified. The process leads generally to weak science and even poorer reproducibility of results.

The science is weak because there are a number of transformations, often under the guise of “normalisation”, of the data that create a distance between the theory under scrutiny and the testing of the theory thus weakening the generalisability of the results and even the meaningfulness of the testing. The science has poor reproducibility because of at least three reasons. Firstly, often a good deal of pre-processing goes into the preparation of the data prior to use in the computational learning stage that is not reported because it is (unjustifiably) considered unnecessary to reveal, a mere background detail. Secondly, the choice of machine learner is varied depending on the researcher’s desire to find the “optimal” solution, thus ensuring that the diversity of statistical tests is not properly factored into the scientific conclusions. Thirdly, the results between experiments and experimenters computed with different machine learners, even though they might be the same algorithm, are not necessarily comparable.

A process that facilitates greater comparability between researchers’ outputs would appear easy to achieve by just requiring scientists to use the same machine learners and collect the same data, but this would stifle innovation in the name of standardisation, which is necessary for rolling out technology (unless you are as powerful as Microsoft) but not for discovery and invention. Instead the essential ingredient for ensuring comparability is the standardisation of reporting. In this case the reporting should **not** be as it has been in the past, as an academic publication, but rather the provision of computational models executable on a standardised computer platform using the research data collected (not massaged) by the researcher.

Implementation of such an environment requires a non-intuitive change in procedures for publishing scientific results. Researchers would produce industrial strength software that will execute their models in a standardised environment. It is unlikely that most researchers will have the software engineering skills to achieve this task so it must be provided by the national body for the preservation and dissemination of computationally modelled science.

This proposal does not suggest that there is a standard machine learner or computational model for all of scientific research, but rather that the researchers in a field need to contribute the software in which their models execute into open-source projects that will bring them to a professional software engineering standard and then loaded into a repository, along with the concomitant data, for dissemination and reuse.

A 5-role Organisational Model for Black-box Computational Modelling in e-Science

An organisational structure for putting machine learning to work in knowledge discovery and modelling for science consists of four distinct expertises, the

interaction of which creates the coherence of the ICT contribution to a research programme. These categories are not necessarily mutually exclusive in terms of the capabilities of individuals although rarely will one person be found to be proficient in all four, that is, they are roles for individuals first and foremost before they are occupations.

Data-centric Users are those users whose interest is in the collection of the data and manipulation of the data and its consequences for explaining the natural or engineered world.

Model-centric Users are those users who have a specific domain of expertise but who specialise in the computational modelling of that domain.

Machine Learner Researcher-Engineers are responsible for the development of algorithmic statistics embedded in software that others use and for applying those models to appropriate science questions.

Software Engineers are the experts who understand the technical issues and have the skills to make the modelling systems serve the needs of the external community, by designing them to be the most generalised and the most computationally efficient and by making them readily usable by the User categories.

As an example, instances of these categories taken from our own work on the Scamseek project, a project to identify financial scams on the Internet, are as follows:

Data-Centric Users: Linguists

Model-centric Users: Computational Linguists

Machine Learner Researchers/Engineers: Collaborating academics/developers of open source machine learners

Software engineers: Software engineers

Another class of user can be added to this model as *Service Users* who could be considered as the industrial organisations who want to use the systems for operational purposes. They are not relevant to the scope of this essay, although the demands for better engineered access to research outputs could lead to greater uptake of research results by industrial interests.

The data-centric user is the important determiner in the organisational model. This user needs to have delivered to them an operational environment or workbench, where they can see their data and how it interacts within the theoretical model, and have the ability to introduce new data and see the variations in the model and perform error analysis on the resultants. Their operational environment is in fact the same as the service user. In essence they interact with the manner in which the model and data coalesce and require a system to manage a large scale program of experiments.

On the other hand the model-centric user is concerned with the variations in the process of constructing the model, particularly the way in which features are derived from raw data and the selection and parameterisation of the models, whether they be machine learners or grey or white-box computational models.

The machine learner researchers/engineers are responsible for inventing methods

that satisfy the needs of the user groups and capture appropriate statistical characteristics of the domain tasks. This is the work of creating algorithmic statistics for machine learning whether for knowledge discovery and/or knowledge verification. This is vital work typically done by academics and their research students.

The software engineers have the task of pulling all this mélange of knowledge together to create operational systems for all user groups, and so they form a closed loop back to the Service Users. They have to adapt to other systems, make efficient, and install the creations of the modelling scientists into an experiment management workbench that is truly usable by the users.

Proposal 2

This paper advocates an organisational structure that recognises the 5-role organisational model and centres the operational task for the NCTDM as delivering advanced-level software engineering of experiment management workbenches as the appropriate form of black-box computational modelling needed for e-Science.

Revealing Structure through Iterative Learning

The dominant paradigm of machine learning is supervised learning. That is, a data set is collected and part of it at least is scrutinised manually by experts and classified for its meaning and role in the scientific field under study. This expertly labelled data set then stands as a "gold standard" against which new experimental results are judged and so it forms a type of nascent theory. This is a good strategy for beginning to understand the nature of novel data. It becomes weaker as more data is collected and the natural diversity of the real world comes into play. Ideally one would want to add to the gold standard as frequently as possible but that is a manually intensive activity. The improvements can only be achieved with serious attention to the computational support that can be provided to it. (This approach is consistent with a view of Science, that it is after all about identifying structure in data and then revising that structure as it is better understood with more data of greater detail becoming available).

Support for this type of continuous revision needs to be provided through a mechanism called **active learning**, that is, the feedback of human review of experimental results as an integrated process within the machine learning software. To achieve active learning as a matter of course for all machine learning and computational modelling is a software engineering task of significant intellectual demand. **However active learning is a key foundation of the enhancement of scientific endeavour through the systematic and intensive use of ICT.**

Active learning is performed by all user roles — service-centric, data-centric and domain-centric — in the model revision process, and so it has to be designed into each of their processes. It is an intellectual challenge as to how to do what is ostensibly the same process for the different roles in the theory producing process.

Proposal 3

This paper advocates that the key work program of the National Centre of Text and Data Mining is the design and implementation of active learning methods

to be integrated with machine learning, and other grey-box and white-box computational modelling where appropriate.

Public Policy on Reuse of Models and Data

A new paradigm for e-Science will be created when Science does more than just share scientific knowledge in the manner we do today. Rather it needs to exploit IT to add a whole new level of co-operation, verification and accountability. Hence it should support the creation of a properly engineered "repository of scientific knowledge" of the **data and models** emanating from publicly funded scientific research.

This involves delivering all three major components of any IT service to the Science community: hardware, software tools, and a management environment. In Australia the past public initiatives have operated to principally provide hardware & network deliverables. Initiatives that focus dominantly on hardware & networks only build the railway tracks of IT. Building a computer network is to ignore the needs for trains (software) and for train drivers (managed services), so that only those groups with existing strong competencies and resources in IT will gain from having the train tracks. The European Community and the USA appear to have shown more enlightened approaches to supporting the development of appropriate software infrastructure.

Management services created around the systematic development of software tools ensure that collected data is available to other researchers in the field. Currently this is an uncommon occurrence. Most data is being preserved on a site local to the researcher sourcing the data. Furthermore the models developed by the researcher are not available to other researchers for testing with new data or for mutual challenges between researchers. A national repository for data and computational models would allow for a leap in the degree of sharing of data and importantly models. Further enhancement of sharing would be achieved with a clear policy that the national repository develop and adapt machine learning software for general research analytics with an open source policy. Support would be provided for interested parties to move into open source communities for general software and domain specific modelling projects and machine learning research.

Greater enhancement of this approach would be achieved by incorporating the functions of the national repository into the national grant allocations systems. As grants are contributed from public monies it is not unreasonable to have the IT components of their budgets defined to be contributions to the national repository. Grantees should be required to contribute their software developments to the national repository both by managing them as open source projects and by conforming to the submission standards defined by the national body. Such a policy would not only release a huge amount of data and software trapped in local computer systems but also allow for unprecedented auditing of the national research outputs, and position Australia well in the international development of this area.

Further advantage can be gained from a national repository for models and data by creating an outlet for "grey science" and the many "grey" data sets and models sitting on every scientist's desktop. These are data sets that have produced either

negative results and therefore are not usually publishable or their significance has not shown the level of impact necessary to gain publication. This data is important for very long term research as it partly fills in gaps in tenuous connections across highly complex yet diaphanous webs of relationships. The biomedical field is taking the lead in collecting together software but has not yet extended that to computational models unless they are intrinsically embodied in the software.

A National Repository

The national repository can be constructed as a virtual system, extant on the computers of every participating organisation. The contents would be managed and curated by the local managers through the management software created by the National Centre for Text and Data Mining. Administrative strategies for getting scientists to lodge their content could be readily legislated for and/or be the conditions of scientific grants. IP rights with commercial value of participating organisations should be established within 6 months of the end of a project or else automatically released to the public through the repository.

Proposal 4

A national e-Science repository be constructed as a virtual repository composed of local repositories assembled by the participating organisations from the projects they participate in through public funding mechanisms.

Workforce Needs

This essay presents a model of team organisation for conducting blackbox modelling research. That organisation identifies that a certain amount of computer science skills are invaluable to the eResearcher no matter the discipline that they specialise in. If discipline specialities are serious about tackling their ICT competency skills then they need to ensure their graduates complete at least a minor programme in their degree, and preferably a major programme in computer science. This requires initiatives to be taken by the disciplines to engage with computer science departments and to encourage their students to take the computer science options. The early adopters of this policy will make the gains in terms of up-skilling their workforce and produce the early gains in research achievements. The longer disciplines take to make this engagement, the further they will fall behind in their research achievements. Minimally NCRIS can take a role of disseminating the need for this strategy amongst the eResearch communities.

International Collaborations

The School of IT at the University of Sydney has close contacts with the various activities at the University of Manchester which is the centre of e-Research activities in the UK. The group under Professor Carol Goble have been successful at seizing the high ground with workflow systems for e-science, but also with the work of Professors Iain Buchan and Alan Rector in Health Informatics and Professor Sophia Ananiadou the Director of the National Centre for Text Mining, are all groups supported by the UK e-Science programs. Furthermore we have established relationships with Professor Junichi Tsujii of the Department of Computer Science, University of Tokyo, who has significant collaborations with the UK groups and leads machine learning and language technology in Japan.

Bibliography

National Collaborative Research Infrastructure Strategy 2008 Review of the NCRIS Roadmap, Discussion Paper.

National Collaborative Research Infrastructure Strategy: 2008 Strategic Roadmap for Australian Research Infrastructure — Exposure Draft.

Neil Beagrie, E-Infrastructure Strategy For Research: Final Report From The Osi Preservation And Curation Working Group, 2007

<http://www.nesc.ac.uk/documents/OSI/preservation.pdf>, accessed 11th Jan, 2007.

Juan Bicarregui, Richard Boulderstone, Lorraine Estelle, Jeremy Frey, Neil Jacobs, William Kilbride, Brian Matthews, Robert McGreevy. 20/20 Vision: an e-Infrastructure for the next decade. Report of the Data and Information Creation Working Group to the e-Infrastructure Steering Group. 2006,

<http://www.nesc.ac.uk/documents/OSI/data.pdf>, accessed 11th Jan, 2007.