Research in a Connected World

Collection Editors:
Alex Voss
Elizabeth Vander Meer
David Fergusson

Research in a Connected World

Collection Editors:

Alex Voss
Elizabeth Vander Meer
David Fergusson

Authors:

Malcolm Atkinson
Tobias Blanke
Ana Lucia DA COSTA
David De Roure
Stuart Dunn
Donal Fellows
David Fergusson
Paul Fisher
Jeremy Frey
Carole Goble
Sarah Harris
Mark Hedges
Andy Kerr
Erwin Laure
Steven Newhouse
Gergely Sipos
Martin Turner
Elizabeth Vander Meer
Alex Voss
Katy Wolstencroft
richard sinnott

Online:
< http://cnx.org/content/col10677/1.12/ >

C O N N E X I O N S

Rice University, Houston, Texas

Collection structure revised: November 22, 2009
PDF generated: October 26, 2012
For copyright and attribution information for the modules contained in this collection, see p. 94.

Table of Contents

Welcome[1]

This book is a very timely contribution. It will help researchers in every discipline grasp the opportunities brought about by the digital revolution. Never before has society undergone such rapid change in the ways in which it communicates and collaborates. This brings untold and, as yet, unimagined new avenues of research and new methods for pursuing research. It demands new thinking and changes in the ways we undertake research.

The digital revolution is probably the most dramatic revolution that humankind has ever experienced. Its foundation is the pervasive growth of digital communication and digital devices. The global reach of digital communication — the Internet, mobile phones and media distribution — means that it is arriving in every nation and reaching most parts of society simultaneously. By contrast, the industrialisation of economies continues to spread after 300 years and the invention and diffusion of printing, telephony and broadcasting is being absorbed and overtaken by the digital revolution.

The rapid changes in personal communication — mobile phones, texting, instant messaging, email, social networking, blogging and twittering — accelerate the propagation of ideas. Governments are opening their data for public scrutiny (President Obama's and Prime Minister Brown's declarations in 2009). Commerce is transformed with Web2.0 models of business: RFID tags in stores, eBay and Amazon trading and marketing wholly online, transactions for tax, payments, books and planning via web pages, and Google's advertising market. In healthcare digital scanning is becoming routine, and will soon be followed by treatment tailored for your genomic variations. Digital cameras, digital video, digital TV and computer games are an intensely competitive global market.

Such a welter of activity is transforming the context of research. Researchers can now find, and be expected to find, an immense number of documents and many sources of data. They create huge volumes of data with faster and higher resolution instruments and generate more precise and larger-scale experiments with laboratory automation. They can use computational notations to describe and refine models precisely. They can access digital replicas of artefacts from museums and libraries — far more than they could ever visit in a lifetime of industrious research — creating virtual assemblies of rarities that could not be contemplated with the original objects. Sensors can be widely deployed to study the atmosphere, oceans and land; they can be worn to study physiology, predation, migration, mating and recreation. Volunteer researchers can gather data worldwide and respond automatically to opportunities and incidents. Global consortia can curate data and make it available for researchers, allowing thousands to collaborate in assembling knowledge and developing models to guide future decisions.

These changes pose new opportunities, raise new questions about research mores and may suggest revision of ethical and legal frameworks. Consequently, researchers need information to exploit the new opportunities, to engage in searches never before possible and to join in worldwide collaborations enabled by the new forms of communication. They will need to be agile and adventurous to thrive in the global competition. Creativity and insight will be amplified by the new methods. Leadership and charisma will assemble complementary skills from around the world to tackle the immense intellectual and practical challenges that face humankind today. Researchers will access the products and by-products of the global revolution as sources of evidence about human behaviour, sampling on scales never previously imagined. Those who engage will shape the way in which future research is done.

[1] This content is available online at <http://cnx.org/content/m32854/1.1/>.

To do this, researchers must gain new skills in computational thinking and data-intensive research. This will be a dynamic process evolving as the pace of the digital revolution throws up new questions and delivers new capabilities. This book is an excellent snapshot; a launching pad from which to get started. Its readers will find key insights and authoritative references, but they must expect to move on to rapidly develop and shape the ideas needed for research in a connected world. They will be on the lookout for claims that appear to break the fundamental principles of distributed systems, but they will also enjoy the rewards of being at the forefront as new methods and technologies make significant advances in research possible.

The most important factor in the success of *Homo sapiens* is their ability to communicate and collaborate. The connected world enables this as never before, as both the speed and scale of collaboration have experienced a step change. Those with the knowledge, enthusiasm and agility to exploit this transformation will pioneer new forms of global behaviour. It is vital that researchers draw on this new resource for combining human talent to address the world's most pressing challenges before it is too late.

When Sir John Taylor launched e-Science, he said, "**e-Science** is about global collaboration in key areas of **science** and the next generation of infrastructure that will enable it". This book shows that Taylor's assertion was a serious understatement. It shows that the new capabilities delivered by the connected world empower new kinds of human collaboration *for all forms of thinking and doing*. Research has a two-fold role: to pioneer these new ways of thinking and doing wherever it will achieve intellectual and practical advances, and to reflect on the deep changes that are underway in global society by recording the massive changes of the digital revolution and better understanding how they shape, and are shaped by, society. This book provides a window into research transformed by the digital revolution, revealing its benefits across disciplines and the added responsibilities that come with these new methods of working. It calls on researchers to observe, record and analyse the digital revolution. It is a valuable resource for researchers as they seize the opportunities brought by the digital age.

Malcolm Atkinson *UK e-Science Envoy* and Director of the e-Science Institute
David De Roure *National Strategic Director for e-Social Science*

Editor's Introduction to Research in a Connected World[2]

The massive availability of networked information and communications technologies today allows us to change the ways we go about our daily working lives as well as the way we spend out leisure time. New ways of shopping, of staying in touch with colleagues and friends, of learning or of navigating places have emerged that are enabled by the ubiquitous electronic devices and networked services that have become available over the past few years. Similarly, as researchers we are today utilising computers in many ways, be it through the use of basic services such as email or the utilisation of the most advanced digital technologies enabling new research methods. No matter what discipline we work in, there are legitimate questions about what potential use we might make of these technologies and what the implications of such use might be.

Over the past decade, funding organisations such as the UK's research councils have funded efforts to make the most advanced information and communications technologies available to researchers and investments are made to develop persistent and sustainable infrastructures to underpin a widespread uptake of digital methods – the development of e-Research. What has been lacking, however, is the development of appropriate learning material such as textbooks that would teach the basics of advanced information systems and digital methods in a way that is accessible to researchers from a wide range of disciplines. This book is an attempt to fill this gap. Its aim is to fill the gap between the initial interest generated by presentations of the potential of e-Research and the various training courses that convey the skills necessary to use specific technologies.

Chapter Outline

The book is divided into four main sections. The first two chapters provide a general introduction to the principles behing e-Research and introduce distributed systems, showing how they differ from single-user desktop systems. The second section discusses a number of different examples of e-Research from a range of disciplines, demonstrating how research can benefit from and be driven forward by the use of advanced information and communications technologies. The third section outlines a number of infrastructures for research that are available to researchers today and discusses the strategies behind the development of European grid initiatives that aim to provide a sustainable environment for the development of e-Research practices. Next, we discuss the role of data and its management over the research lifecycle as well as a number of relevant technologies. The fifth section discusses different ways that researchers can access infrastructure services and the ways they can be factored into actual everyday research practices. Finally, we conclude the book with a collection of resources that we hope will help the reader explore the field of e-Research further and make informed choices about the adoption of the technologies and methods described in this book.

[2]This content is available online at <http://cnx.org/content/m32855/1.2/>.

Acknowledgements

First of all, we would like to thank our colleagues who have contributed chapters to this collection. They have given generously of their time and the essential input of expertise without which this book could not have come into existence. We would also like to thank the organisations that have provided support in cash or in kind:

The UK's JISC has provided financial support through the funding for the e-Infrastructure Use cases and service usage models project.

The Scottish Informatics and Computer Science Alliance has supported the editing process by funding contributions made by Alex Voss.

The National e-Science Centre has supported the editing process by funding the contributions made by David Fergusson and Elizabeth Vander Meer.

The Manchester e-Research Centre has supported the editing process by funding contributions made by Alex Voss and by administering the production process of the first edition of the book.

Research in a Connected World[3]

Key Concepts

- data-rich science
- e-Research

Introduction

Research today is often critically dependent on computation and data handling. The practice has become known under various terms such as e-Science, e-Research, and cyberscience. We would like to avoid using these terms, but when it is unavoidable, in the interests of brevity, we use the term e-Research in a broad sense to include all information processing support for research. Irrespective of the name, many researchers acknowledge that the use of computational methods and data handling is central to their work.

There is no question that scientific research over the past twenty years has undergone a transformation. This transformation has occurred as a result of various factors. New technologies, leading to new methods of working, have accelerated the pace of discovery and knowledge accumulation not only in the natural sciences but also in the social sciences and arts and humanities. Advances in scientific and other knowledge have generated vast amounts of data which need to be managed well so that they can be analysed, stored and preserved for future re-use. Larger scale science enabled by the Internet, and other information and communication technologies (ICTs), scientific instrumentation and automation of research processes has resulted in the emergence of new research paradigms that are often summarised as **'data-rich science'**. A feature of this new kind of research is an unprecedented increase in complexity, in terms of the sophistication of research methods used, in terms of the scale of phenomena considered as well as the granularity of investigation.

e-Research involves the use of computer-enabled methods to achieve new, better, faster or more efficient research and innovation in any discipline. It draws on developments in computing science, computation, automation and digital communications. Such computer-enabled methods are invaluable within this context of rapid change, accumulation of knowledge and increased collaboration. They can be used by the researcher throughout the research cycle, from research design, data collection, and analysis to the dissemination of results. This is unlike other technological "equipment" which often only proves useful at certain stages of research. Researchers from all disciplines can benefit from the use of e-Research approaches, from the physical sciences to arts and humanities and the social sciences.

The following sections in this introduction will elaborate on these transformations in research and the role played by ICT, describing research collaborations, “big research” in a globalised world and participation in research.

Research Collaborations

Today’s research into social and scientific issues and problems often involves increased sharing of resources – because individual research institutions cannot afford having these resources or because they are inherently

[3]This content is available online at <http://cnx.org/content/m20834/1.3/>.

distributed (for example in the case of linked radio telescopes). The research community has changed, so that more work is done in international collaborations and these collaborations have become increasingly multi- or interdisciplinary.

Tackling the grand challenges of many disciplines today requires the coordinated effort of groups of researchers working on different aspects of a problem. Also, individual researchers can more rapidly increase their knowledge in a particular field if they are able to become part of an international and interdisciplinary collaborative network. Instead of working on their own or only with colleagues within their own institutions, researchers now often work in collaborations with colleagues in other institutions, who can provide specialist knowledge, skills or access to resources.

e-Research provides researchers with an environment for sharing resources and facilitates collaborations by making large, distributed data sets accessible, through enabling synchronous or asynchronous collaboration across geographical distances and providing access to resources regardless of location. This opening up of research means that researchers need not be held back by their own resource constraints and can more freely participate in cutting-edge projects.

e-Research Technologies Supporting Collaboration

e-Research technologies support the research collaborations described above by introducing a model for resource sharing based on the notions of "resources" that are accessed through "services". Resources can be computational resources such as high-performance computers, storage resources such as storage resource brokers or repositories, datasets held by data archives or even remote instruments such as radio telescopes. In order to make resources available to collaborating researchers, their owners provide services that provide a well-described interface specifying the operations that can be performed on or with a resource, e.g., submitting a compute job or accessing a set of data.

This simple underlying model of collaboration is complemented by additional functionality such as authentication and authorisation to regulate access to a resource or management functions such as resource reservation. It is important to note that the underlying model is kept simple and that any additional functionality layered on top of it is also formulated in terms of resources and services wherever possible. Using these general principles, it is possible to build a vast range of tools and applications that support collaborative research.

Computer-enabled methods of collaboration for research take many forms, including use of video conferencing, wikis, social networking websites and distributed computing itself. For example, researchers might use Access Grid[4] for video conferencing to hold virtual meetings to discuss their projects. Access Grid and virtual research environments provide simultaneous viewing of participating groups as well as software to allow participants to interact with data on-screen. Wikis have also become a valuable collaborative tool. This is perhaps best demonstrated by the OpenWetWare[5] website, which promotes the sharing of information between researchers working in biology, biomedical research and bioengineering using the concept of a virtual Lab Notebook. This allows researchers to publish research protocols and document experiments. It also provides information about laboratories and research groups around the world as well as courses and events of interest to the community.

Social networking sites have been used or created for research purposes. The myExperiment[6] social website is becoming an indispensible collaboration tool for sharing scientific workflows and building communities. Such sharing cuts down on the repetition of research work, saving time and effort and leading to advances and innovation more rapidly than if researchers were on their own, without access to similar work (for comparison to their own). Other social networking sites such as Facebook have been adopted by researchers and extensions have been built to allow them to be used as portal to access research information. For example, content in the ICEAGE Digital Library[7] can be accessed within Facebook.

[4] http://www.ja.net/services/video/agsc/AGSCHome/whatisaccessgrid.html
[5] http://openwetware.org/
[6] http://www.myexperiment.org/
[7] http://library.iceage-eu.org/

Systems Research in a Globalised World

Many researchers now devote a significant amount of their attention to global issues, which previously could not be addressed due to technological and informational limitations. These global issues include, for instance, climate change, pandemics, rainforest destruction and biodiversity. Such "big research" problems fall under wider contemporary concerns about living sustainably and understanding human biology and health (including the aetiology of diseases and the search for cures).

This ubiquitous global perspective has in large part emerged because of a worldwide exchange of information and the availability of data resulting from use of ICT, coupled with the use of ICT to organise that data. For example, the earth is seen as a system or as systems within systems, which necessitates the need for cross-scale research. Earth system science in geosciences provides a useful example of this change to "systems research". ICT is used to model and simulate integrations of geology, oceanography and environmental sciences, generating a more complex, holistic view than was possible prior to the increased use of computer enabled methods. There has also been a recent concerted development of systems biology, which involves integration of mathematics, engineering and computer science to manage the data deluge in biology in order to answer big questions concerning sustainable living and human health on a global level.

A significant number of researchers in the social sciences and arts and humanities have also taken up this global view. For the social sciences, this perspective is clear, for instance, in the idea of "global knowledge" and attempts to solve social issues relating to sustainable living through large-scale data gathering and analysis. In the arts and humanities, a global perspective is evident in the development of the Global Performing Arts Consortium[8] , an international database of performing arts resources, and in global cultural and international studies research which often relies on/requires access to large amounts of cross-culturally derived data to adequately substantiate conclusions.

Participation in Research – Democratising "Big Science"

e-Research not only enables scientists to tackle "big" questions, but it has also allowed for wider participation in research. Volunteer computing allows members of the public to support and take part in research conducted by teams of professional researchers by providing compute resources or by performing specific tasks that are part of the research process. For example, the SETI@home project[9] makes use of volunteers' desktop computers to search for extraterrestrial life while Folding@home[10] uses the compute power provided by volunteers to study protein folding. In the case of climateprediction.net[11] , any member of the public with appropriate computer equipment can contribute to the study of climate change. In each of these cases, tasks and data are shared across a network of dispersed computers, thus increasing the compute power and storage capacity available far beyond the capabilities of a single computer. Several of the examples of inspiring e-Research projects we will introduce here have been successful as a result of using volunteer computing.

Open Source Science is not just about direct public participation. It is also about transparency, so that the public has access to and can observe the research process. Open Notebook Science enables better collaboration among researchers at the same time that it makes research project records available online for perusal by the lay public. In this way, "big science" is democratised, no longer purely the product and tool of a cloistered research elite but an activity within a wider societal context that society members can take part in.

[8] http://www.glopac.org/
[9] http://setiathome.ssl.berkeley.edu/
[10] http://folding.stanford.edu/
[11] http://www.climateprediction.net/

Research in a Connected World - Fundamental Concepts and Inspiring Examples

Preceding sections in this introduction have presented a strong argument for the uptake of e-Research methods by illustrating their importance in a multitude of research endeavors. The Research in a Connected World brochure serves as an introduction to e-Research for those unfamiliar with such methods, revealing it's potential and promise for all disciplines. The brochure consists of individual modules that give researchers a grounding in fundamental concepts and a taste of what is possible when using computer-enabled methods.

We provide an introduction to distributed systems, contrasting them to desktop PCs, and then move on to detailed discussion of inspiring examples of e-Research, looking at projects in many different fields. These examples are followed by examples that show the wider impact of e-Research and explore the unique collaborations that have developed not only among other academic researchers but also between researchers and the wider public. The subsequent section of the brochure describes elements of and issues relating to distributed systems, beginning with a short history of distributed computing and including modules on the taxonomy of research computation problems, distributed computing architectures, issues concerning managing complex data, visualisation, use of portals and virtual research environments. A final module contains a list of relevant services and contacts.

We hope this resource will not only inform you but also inspire you to begin to use computer-enabled methods to further your research. If you already consider yourself an e-Researcher, we hope to have introduced you to new tools that you can begin to apply in your own work.

What is a Distributed System?[12]

Key Concepts

- distributed systems

Introduction

Over the past decades, as we have begun to explain, we have moved from processing the data that we can hold in a lab notebook to working with many thousands of terabytes of information. (For reference, a terabyte is a million megabytes, and a megabyte is a million letters. A plain textbook might be a few megabytes in size, as might a high-quality photograph — a terabyte is like a huge library.) And yet we keep striving to work with ever more: more genomic data; more high-energy physics data; ever more detailed astronomical photographs; ever richer seismographic measurements; ever more layers of interpretation of artistic details; ever greater volumes of financial data; ever more complex and realistic simulations. We drill down ever deeper into the details. How are we coping with this?

We are in the middle of a huge revolution in information processing, driven by the fact that our tool of choice for working with information — the computer — has been getting exponentially better ever since their invention during the Second World War. We live in the middle of an age of wonder. And yet, despite now being able to hold immense quantities of computation and storage in our hands, our desire to work with ever more has grown even faster.

Thankfully we have been living through another revolution at the same time; the telecommunications revolution. The telecommunications revolution started with the invention of the telegraph, but accelerated with the convergence of computers and telecoms to create the Internet. This not only allows people to share information, but also computers, and it has transformed the world. The first indication of just how amazing this would be came with the WorldWideWeb (WWW), the first internet system to really reflect everything that people do throughout society [link/reference here to history chapter]. But it will not be the last; the ripples from the second wave are now being felt, and it is the global research community that are in the lead. This second wave is Distributed Computing.

The Way Distributed Computing Works

Simply put, distributed computing is allowing computers to work together in groups to solve a single problem too large for any one of them to perform on its own. However, to claim that this is all there is to it massively misses the point.

Distributed computing is not a simple matter of just sticking the computers together, throwing the data at them and then saying "Get on with it"! For a distributed computation to work effectively, those systems must cooperate, and must do so without lots of manual intervention by people. This is usually done by splitting problems into smaller pieces, each of which can be tackled more simply than the whole problem. The results of doing each piece are then reassembled into the full solution.

[12]This content is available online at <http://cnx.org/content/m31661/1.1/>.

The power of distributed computing can clearly be seen in some of the most ubiquitous of modern applications: the Internet search engines. These use massive amounts of distributed computing to discover and index as much of the Web as possible. Then when they receive your query, they split it up into fast searches for each of the words in the query. The results of the search are then combined in the twinkling of an eye into your results. What about locating computers on which to execute the web search? That is itself a distributed computing problem, both in the process of looking up computer addresses and also in finding an actual computer to respond to the message sent on that address.

Early distributed systems worked over short distances, perhaps only within a single room, and all they could really do was to share a very few values at set points of the computation. Since then, things have evolved: networks have got faster, numbers of computers have got larger and the distances between the systems have got larger too.

The speeding up of the networks (from the telecommunications revolution) has been extremely beneficial as it has allowed many more values to be shared effectively, and more often. The larger number of computers has only partially helped; while it has meant that it is possible to use more total computation and to split the problems into smaller pieces (allowing a larger overall problem), it has also increased the amount of time and effort that needs to be spent on communication between the computers, since the number of ways to communicate can increase (see Figure 1).

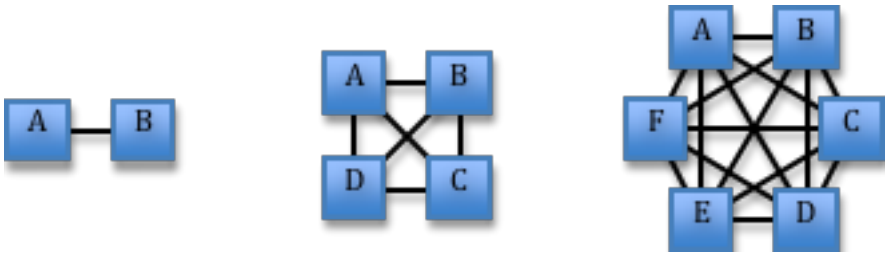

Figure 1: The growth in the number of links as the number of computers goes up.

There are, of course, ways to improve communication efficiency, for instance by having a few computers specialize in handling the communications (like a post office) and letting all others focus on the work, but this does not always succeed when the overall task requires much communication.

The distance between computers has increased for different reasons. Computers consume power and produce heat. A single PC normally only consumes a small amount of power and produces a tiny amount of heat; it is typically doing nearly nothing, waiting for its users to tell it to take an action. With a computational task, it would be far busier and will be consuming electrical energy in the process; the busier it is, the more it consumes and produces heat. Ten busy PCs in a room can produce as much heat as a powerful domestic electric heater. With thousands in one place, very powerful cooling is required to prevent the systems from literally going up in smoke. Distributing the power consumption and heat production reduces that problem dramatically, but at a cost of more communications delay due to the greater distances that the data must travel.

There are many ways that a distributed system can be built. You can do it by federating traditional supercomputers (themselves the heirs to the original distributed computing experiments) to produce systems that are expensive but able to communicate within themselves very rapidly; this remains favoured for dealing with problems where the degree of internal communication is very high, such as weather modelling or fluid flow simulations. You can also make custom clusters of more traditional PCs that are still dedicated to being high-capability computers; these have slower internal communications but are cheaper, and are suited for many "somewhat-parallel" problems, such as statistical analysis or searching a database for matches (e.g., searching the web). And you can even build them by , in effect, scavenging spare computer cycles from across

a whole organization through a special screen saver (e.g., Condor, BOINC); this is used by many scientific projects to analyse large amounts of data where each piece is fairly small and unrelated to the others (e.g., Folding@Home, SETI@Home, Malaria Control).

Special Challenges for Distributed Computing

We are now moving from having the number of computers working on a problem being small enough to fit in a building or two to having tens of thousands of computers working on single problems. This brings many special challenges that mark distributed computing as being a vastly more complex enterprise than what has gone before.

The first of these special problems is security. The first aspect of security is the security of the computers themselves, since few people feel like giving some wannabe digital mobster a free pass to misuse their computers. The second aspect is the security of the data being processed, much of which may be highly confidential or a trade secret (e.g., individual patients' medical data, or the designs for products under development). The third aspect of security is the security of the messages used to control the other computers, which are often important in themselves and could be used to conduct a wide range of other mischief if intercepted.

The second special problem of distributed computing is due to the use of systems owned by others, either other people or other organizations. The issues here are to do with the fact that people ultimately retain control over their own systems; they do not like to cede it to others. This behaviour could be considered just a matter of human nature, but it does mean that it is extremely difficult to trust others in this space. The key worries relate to either the computer owner lying about what actions were taken on their systems (for pride, for financial gain, for spite, or just out of straight ignorance) or the distributed computer user using the system for purposes other than those that the computer owner wants to permit.

Interoperability presents the third challenge for distributed computing. Because the systems that people use to provide large-scale computing capabilities have grown over many years, the ways in which they are accessed are quite diverse. This does mean that a lot of effort has been put into finding out access methods that balance the need for efficiency with those of security and flexibility, but it also means that frequently it is extremely difficult to make these systems work together as one larger system. Past attempts by hardware and software vendors to lock people in to specific solutions have not been helpful here either; researchers and practitioners want to solve a far more diverse collection of challenges than the vendors have imagined there to be. After all, the number of things that people wish to do is limited only by the human imagination.

These challenges can be surmounted though, even if the final form of the solutions is not yet clear. We know that the demands of security can be met through a combination of encryption, digital signatures, firewalls, and placing careful constraints on what can be done by any program. The second challenge is being met through the use of techniques from digital commerce like formal contracts, service level agreements, and appropriate audit and provenance trails. The third, which will become ever more important as the size of problems people wish to tackle expands, is primarily dealt with through standardization of both the access mechanisms (whether for computation or for data) and the formal understanding of the systems being accessed by common models, lexicons and ontologies.

The final major challenge of distributed computing is managing the fact that neither the data nor the computations are open to relocation without bounds. Many datasets are highly restricted in where they can be placed, whether this is through legal constraints (such as on patient data) or because of the sheer size of the data; moving a terabyte of data across the world can take a long time, and in no time the most efficient technique becomes sending disks by courier, despite the large quantity of very high capacity networks that exist out there.

This would seem to indicate that it makes sense to move the computations to the location of the data, but that is not wholly practical either. Many applications are not easy to relocate: they require particular system environments (such as specialized hardware) or direct access to other data artefacts (specialized databases are a classic example of this) or are dependent on highly restricted software licenses (e.g., Matlab, Fluent, Mathematica, SAS, etc.; the list is enormous). This problem does not go away even when the