Data-Intensive Research Workshop Report

Authors: All of the workshop's participants, see Appendix A
Editors: Malcolm Atkinson, David De Roure, Jano van Hemert, Shantenu Jha,
Ruth McNally, Bob Mann, Stratis Viglas and Chris Williams

17 May 2010

Contents

List of Tables

Preface

Data-Intensive Research is any research in any discipline where careful thought about how to use data is essential for achieving success. Later chapters expand on this definition and demonstrate the diversity of ways in which handling and interpreting data may be challenging. Chapter 1 gives a fuller introduction and shows why it is timely to discuss this topic.

The Data-Intensive Research workshop was run by the e-Science Institute (http://esi.ed.ac.uk) at the University of Edinburgh for the week 15–19 March 2010. Over the course of the week the workshop involved approximately 100 participants who are the authors of this report (see Appendix A). The workshop was organised by those shown in Table 1 and followed the timetable given in Appendix B. Various web resources were built before, during and after the workshop as shown in Table 2.

This report is a first step in communicating the enthusiasm, understanding and sense of direction that was developed during the workshop. All participants contributed to this report in breakout groups, the final panel, and many informal discussions as well as via email lists, the wiki and tweeting. The input of the 30 speakers—see Table 3—is directly incorporated in the report, particularly in Chapters 1 to 4, which correspond approximately to the timetables of Monday's to Thursday's programme. These days viewed data-intensive research from the viewpoints of: (a) introduction to and the context of data-intensive research, (b) challenges emerging from the increasing volumes and sources of data, (c) challenges arising from the complexity of data, and (d) challenges in supporting researchers interacting with data. Friday's programme brought together all of the activities during the week to digest and summarise them, to consolidate and review our understanding, and to initiate the production of this report. It was a primary input into Chapter 5.

At the outset the organisers had planned to stimulate bridge-building between technical and discipline silos, by clustering challenges and disciplines into days and by setting up cross-cutting themes that ran throughout the week. The matrix thus formed, with an additional row to consider social and ethical issues that emerged during the workshop, is shown in Table 4. Ruth McNally kindly agreed to be co-opted into the editorial team to take care of that theme. There were two other nascent themes: (a) text-mining applications, particularly the integration of data from text with other data, and (b) training and ramps to better enable the adoption of data-intensive methods and the appreciation of data-intensive results. If anyone would wish to develop a theme section covering these, they will be added to Chapter 5.

Whilst the speakers and other participants produced most of the ideas, except where we explicitly quote a person or group, the editors take full responsibility for the selection and presentation of the text and figures in this report.

We would like to acknowledge the support of the e-Science Institute's events team and technical support staff who provided a smoothly run environment conducive to research discussions during the preparatory period and throughout the week. We also thank Jo Newman, School of Informatics, University of Edinburgh, for her help in populating the organisers' wiki and in communicating with the speakers and other participants. The e-Science Institute provided facilities and met costs with the help of EPSRC grants EP/D056314/1 and EP/F064152/1. The EPSRC senior research fellowship, EP/D079829/1, supported Malcolm Atkinson, and the European Union Framework Programme 7 ADMIRE project (ICT 215024) supported Atkinson,

Name	Affiliation	Roles & Responsibilities	See page
Malcolm Atkinson	School of Informatics, University of Edinburgh	Monday's and Friday's day organiser and overall co-ordination	1 & 37
David De Roure	School of Electronics and Computer Science, University of Southampton	Wednesday's organiser	24
Jano van Hemert	School of Informatics, University of Edinburgh	Thursday's day organiser	30
Shantenu Jha	Center for Computation and Technology, Louisiana State University	Programming-paradigms theme organiser	51
Ruth McNally	Department of Sociology, University of Lancaster	Sociological and ethical theme editor	56
Bob Mann	Institute for Astronomy, University of Edinburgh	Linking with Next Generation Sky Surveys theme	XXX
Stratis Viglas	School of Informatics, University of Edinburgh	Organiser of paradigms to structure data theme	50
Chris Williams	School of Informatics, University of Edinburgh	Organiser of the Analysis theme	49

Table 1: Data-Intensive Workshop organisers

Title	Description	URL
Organisers' wiki	Overview, detailed daily programmes, abstracts and speaker biographies	http://bit.ly/bQpu5h
Participant's wiki	Additional detail, records of discussion and open contributions	http://bit.ly/cygimA
Talks	Speakers' presentations, posters, etc.	http://bit.ly/9G8Juo
eSI event	Standard eSI event page	http://bit.ly/dAvKal
Twitter stream	Live commentary	http://bit.ly/dpKKtv
Tweets during workshop	snapshot	http://bit.ly/dpKKtv

Table 2: Web resources connected with the Workshop

van Hemert and Corcho, and several helper participants.

Speaker	Affiliation	Talk Title	See
Bernie A'cs	National Center for Supercomputing Applications, University of Illinois at Urbana-Champaign	Soaring through clouds with Meandre	15
Malcolm Atkinson	School of Informatics, University of Edinburgh	Setting the agenda for data-intensive research	1 & 38
Jim Austin	Department of Computing Science, University of York	Using search for engineering diagnostics and prognostics	29
Michael Batty	University College London	Challenges in Large Scale GeoSpatial Data Analysis: Mapping, 3D and the GeoSpatial Web	21
Mark Birkin	School of Geography, University of Leeds	Spatial microsimulation for city modelling, social forecasting and urban policy analysis	27
Peter Buneman	School of Informatics, University of Edinburgh	Curated databases	34
Mario Caccamo	BBSRC, The Gene Analysis Centre (TGAC)	Big Data Bioinformatics	24
James Cheney	School of Informatics, University of Edinburgh	Review of data-structuring theme	41
David De Roure	University of Southampton	Data-complexity challenges and strategies	40
Torild van Eck	The Royal Netherlands Meteorological Institute	Data challenges in Earthquake Seismology	17
Geoffrey Fox	Pervasive Systems Laboratory, Indiana University	Programming Paradigms Theme Opening	13
Hugh Glaser	Sem4 Ltd	Linked Data: Making things more accessible	28
Thore Graepel	Microsoft Research Cambridge	Learning from Data in Online Advertising and Games	7
Carole Goble	University of Manchester	Providing an environment where every data-driven researcher will thrive	22
Keith Haines	Reading e-Science Centre, University of Reading	Making the most of Earth-system data	18
Alan Heavens	Institute for Astronomy, University of Edinburgh	Dealing with large data sets in astronomy and medical imaging	17
Jano van Hemert	School of Informatics, University of Edinburgh	Data-interaction challenges and strategies	40
Shantenu Jha	Louisiana State University	Data-intensive programming theme review	44
Douglas Kell	Chief Executive, Biotechnology and Biological Sciences Research Council and Manchester University	Motivation and Strategies for data-intensive Biology	7
Martin Kersten	CWI	Scientific Databases: the story behind the scenes	19
Paul Lambert	Stirling University	Handling social science data: Challenges and responses	26
Xavier Llorá	National Center for Supercomputing Applications, University of Illinois at Urbana-Champaign	Soaring through clouds with Meandre	15
Andrew McCallum	University of Massachusetts	Discovering Patterns in Text and Relational Data with Bayesian Latent-Variable Models	35
Bill Michener	University of New Mexico	Building a virtual data centre for the biological, ecological and environmental sciences	33
Beth Plale	Indiana University	Earth-Systems data in real-time applications: low latency, metadata, and preservation	20
Jonty Rougier	University of Bristol	Model limitations: sequential data assimilation with uncertain static parameters	19
Andrey Rzhetsky	Department of Medicine, Department of Human Genetics, Computation Institute, Institute for Genomics and Systems Biology, University of Chicago	Extracting relevant biomedical information from text and presenting it well	32
Joel Saltz	Center for Comprehensive Informatics, Emory University	Medical-image processing and caBIG	31
Jason Swedlow	Wellcome Trust Centre for Gene Regulation and Expression, University of Dundee	The Open Microscopy Environment: Informatics and Quantitative Analysis for Biological Microscopy, HCAs, and Image Data Repositories	25
Stratis Viglas	School of Informatics, University of Edinburgh	Data-structuring paradigms Theme Opening	12 & 39
Paul Watson	University of Newcastle	Using Real-Time data to Understand and support Human Behaviour	35
Chris Williams	School of Informatics, University of Edinburgh	Data Analysis Theme Opening	12, 26 & 41
John Wood	Imperial College London	A Vision for Research in 2030	44

Table 3: Data-Intensive Research Workshop Speakers

	Monday	Tuesday	Wednesday	Thursday	Friday
day's topic	motivation	data volume	complexity	data interaction	future direction
application domains	astronomy, biology and business	environmental, astro- & geo-sciences	biology, medical images, social science & engineering	biomedical, environmental, corporate and behavioural data	multi-national research facilities
analysis	see 1.5 & 1.7	see 2.2 & 2.6	see 3.5 & 3.8	see 4.3 & 4.6	see 5.2.5
database	see 1.3 & 1.8	see 2.5	3.4	see 4.5	see 5.2.6
programming	see 1.9 & 1.10	see 2.9			see 5.2.7
	chapter 1	chapter 2	chapter 3	chapter 4	chapter 5

Table 4: Matrix of Challenges and Themes

Executive Summary

We met to develop our understanding of DIR. Approximately 100 participants (see Appendix A) worked together to develop their own understanding, and we are offering this report as the first step in communicating that to a wider community. We present this in turns of our developing/emerging understanding of "What is DIR?" and "Why it is important?". We then review the status of the field, report what the workshop achieved and what remain as open questions.

Why DIR?

The simplest definition of DIR is any research that only progresses well with serious attention to how data is used. Almost every speaker, had a refinement, for example, some chose the phrase "data-driven research" to denote the new paradigm of going from data to hypothesis, although the Arts & Humanities participants then pointed out that was their normal procedure. Some speakers emphasised the issue of scale, but more were concerned with the challenges of complexity in both data & analysis. Others recognised the challenges of data-quality and variability, the value of data citation and curation, and processes to select, describe, use and keep data.

The talks and demonstrations showed a wide range of examples where data-intensive methods were yielding new results in many disciplines. The research and its applications are going through a transition, as more data becomes available from cheaper and faster instruments, from new government policies, from industry & commerce and from wide community action. Today's challenges, such as: security, food, water, environment, health & economic stabilty, require that we become more adept at using data to inform policy and decision making.

The workshop took a data-centric view of the recurring concepts at the interface between data and computation. We introduce some of them here, by considering the actions and objectives of the researchers and the tools that facilitate them. In order to effectively exploit data, practitioners need to carry out some of the following: collecting, filtering & sampling, cleaning & normalising, fusing, linking & integrating, analysing & summarising, organising & structuring, curating & publishing, discovering, exploring, annotating, deriving, preserving and presenting. In data-driven research, researchers wish to discover patterns that are significant in their fields; in hypothesis-driven research, they seek evidence from data; whilst in applied science [2], they seek to present evidence which derive from data in a form that enables appropriate action.

New machine-learning methods enable the combination of domain knowledge and statistical approaches on large bodies of data [3]. Advances in data description are enabling extensive data-fusion, new tool-sets and services, which apply generic and domain-specific algorithms to data & provide research environments that can be used by a growing number of practitioners.

The growing power of data and the new methods raise ethical and social issues concerning the delivery of benefits and the protection of individuals.

The providers of research environments, whether they are established reference resources or local facilities, face new challenges. Current data-handling strategies will meet a power wall in a few years. Current policies on data-collection will make excessive demands. The potential for research and innovation will be frustrated, unless infrastructure providers can deliver: sufficient data throughput, wide accessibility and interfaces, that are readily usable by a large community.

Status of the Field

Participants reported many beacons of success, from the long-established reference resources (e.g. European Bioinformatics Institute and National Centre for Atmospheric Research), to refinining methods and delivering operational services (e.g. by Microsoft), to the individual practitioners who are composing data from multiple sources and extracting new knowledge.

Beacons of success show the huge potential of data-intensive research and methods, and in particular they show that in some cases they can be developed in one application domain and applied fluently to other areas & domains. The challenge today is that this almost always depends on establishing a talented team from and across different disciplines. But we recognised that societal challenges demand data-intensive applications in a huge number of contexts addressing a wide-variety of topics. It is infeasible to imagine that talented inter-disciplinary teams can be assembled in every case, so we need to develop the capacity and facilities which will enable the long-tail of practitioners to use the new methods correctly and effectively.

There was wide recognition that a large number of research practitioners would need training in data-centric thinking and in applying data-intensive methods. Others would need training in appreciating and evaluating the outcome of these methods. The workshop considered it important to develop curricula and engage educational institutions in a substantial programme, to develop and deliver the required training.

Examples of technical beacons of success, are Grays' Laws to steer communication between domain and technology experts, GrayWulf to deliver a high-data throughput, and sophisticated text-mining to extract detailed knowledge. Current challenges include the impending 'Power-Wall' barrier, the lack of conceptual frameworks for toolkits and data-composability, and the anticipated volumes of data from the latest instruments (e.g. new gene sequencing machines). Additional requirements are new data-analysis algorithms that enable effective and efficient operations and address new questions on large volumes of data. Finally, the impact of the social applications of DIR will not be determined solely by the design of technological systems, but will also crucially depend upon the configuration of the socio-economic and legal systems in which these new approaches to are intended to be an integral part.

What was Achieved?

A goal of the workshop was to bridge boundaries across and between domain and technological silos. This worked well, as evidenced by the number of alliances formed and participants rushing off to try new methods and tools.

Another goal, was to improve our understanding of data-intensive research. This took two forms: *a*) almost everyone discovered the availability of powerful new methods & technologies and came to appreciate their extensive applicability, and *b*) new pervasive issues became apparent, such as the challenge of understanding how to balance ethical and societal considerations as we adopt methods of unprecedented power. This latter topic became so pervasive, that we identified it as an additional cross-cutting theme, see §5.8.

A third goal, was to stimulate work in the field. Our impact here is harder to assess, but two eSI theme proposal have been submitted as a consequence of this workshop.

The workshop developed a shared understanding of the current landscape and challenges and has begun to articulate that understanding in this report (see §5.4 onwards).

What Remains to be Done?

A fundamental question is whether there are underlying principles of DIR, and if so, how should the exploration of DIR be structured so as to expose and exploit them.

We recognise that to approach this understanding, we have to classify the existing methods and applications of DIR. We did not anticipate a simple model because the field is so diverse and rich. Hints of this classification were beginning to emerge and we hope that by collating a broad range of DIR examples, we have provided a foundation for such a classification. We hope others will further this agenda by reporting useful clusters which admit consistent treatment.

Most participants felt the need for better composibility of data, and readily available and easy to use tool-sets; at present we see islands associated with the more mature data-sharing campaigns, and traditional computation-oriented tools & frameworks. Progress in this area depends on an understanding of how to trade between agility of independent researchers and standards.

Engaging the long tail of practitioners in DIR will require them to change their practices. This will require upstream integration of user preferences and their existing practices into the design of tools. Moreover, one size is unlikely to fit all, and tools may have to be tailored for different disciplines and contexts. Change is unlikely to be achieved solely by technical progress; it also requires social innovation in the systems of reward and recognition for good DIR practices.

When it comes to the applications of DIR as the basis for intervening in the lives of individuals, the challenges are not only technical because of the potential risks such interventions pose to individual privacy, autonomy, consent and confidentiality. There are also issues of equality and social justice, and the question of who will provide and pay for such services.

Because we focussed on the interface between applications and methods, we were unable to explore the requirements imposed by DIR on Cyber-infrastructure, tools & service provision. A need emerged for developments in analytic and visualisation algorithms, and for improved models of interaction with data — we look to future work to chart and address the requirements.

Data is growing in two dimensions: each collection grows as well as the number of collections grow, so there is always a long-tail of data resources, which are relatively small, but which may contain important information. It is an open question how best to accommodate this long-tail.

Chapter 1

Introduction

1.1 Overview

Maybe order logically and not necessarily chronologically? sj

Missing analysis currently. TR should have some. For example a table with the rows as the data challenges and the columns as the applications – so it is easy to see which challenges appear in which applications. Similarly, for tools/solutions, we have challenges as rows and tools/solutions as the columns telling us which challenges are met sj

For the brief summary/overview, I would like to recommend a tighter adherence to format. For example, in the DPA paper/book, I asked all contributors to adhere to: (i) 1 para for application description, (ii) why is it a distributed application (iii) how do you develop/execute as a distributed application and (iv) challenges/problems/success in doing so. Similarly here we could for data-intensive applications have (i) application description, (ii) tools uses (iii) which data-challenges are present and (iv) what is required. This will also help it to sync up with the tables suggested above sj

consistency needs to be imposed in write-up at many levels – tense, first versus third person.. sj

The first day's activities contained several stimulating talks and served as an introduction to and motivation for many aspects of data-intensive research. Professor Dave Robertson, Head of School of Informatics, University of Edinburgh, warmly welcomed the participants and then demonstrated the growing importance of data-intensive research in Edinburgh by citing examples of data-intensive activity in each of the School's seven institutes.

1.2 Introduction to DIR Workshop

Malcolm Atkinson, Professor of e-Science, Director of e-Science Institute and UK e-Science Envoy, School of Informatics, University of Edinburgh.

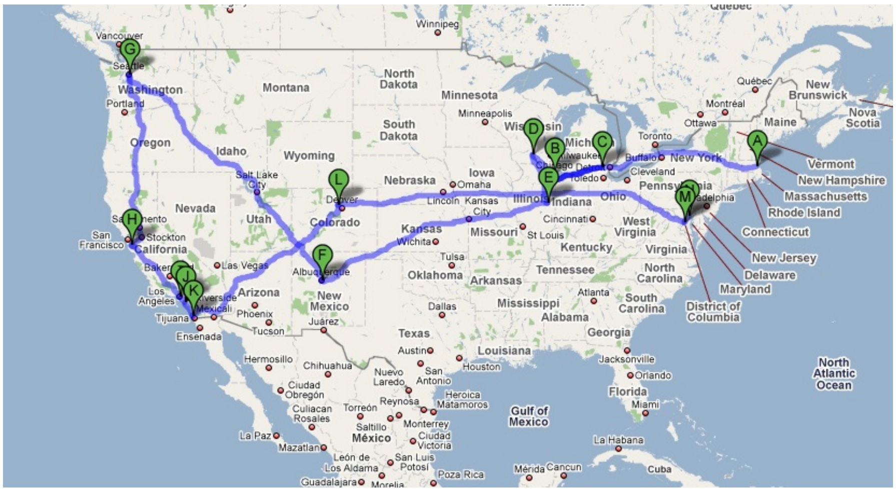

Figure 1.1: Route of fact-finding mission September 2009

The talk is available at: http://www.nesc.ac.uk/action/esi/download.cfm?index=4432.

The workshop was triggered by Atkinson and De Roure's recent (September 2009) tour of the USA (see Figure 1.1) to study how researchers were using data [1]. The workshop's goal is to stimulate thought about how to make the best use of data because: *a*) effective use of data is very powerful, and *b*) most data is not used or used ineffectively.

This workshop is intended to provoke thought about better use of data may be accomplished. It should start work on answering questions, such as:

1. What is data-intensive research?
2. What data-intensive applications and methods work today?
3. How can a researcher make a well-informed choice of data-intensive methods?
4. What impedes the successful use of today's growing wealth and diversity of data?
5. How can we organise research, e-Infrastructure and working practices to overcome these impediments?

And it should stimulate follow on collaborative activities that refine the questions and pursue theoretical and practical answers that facilitate data-intensive research.

A working definition of *data-intensive research* is:

> research where careful thought is needed about how to use data.

Where *use* of data may involve any of the following actions on data: collecting, filtering & sampling, cleaning & normalising, fusing, linking & integrating, analysing & summarising, organising & structuring, curating & publishing, discovering, exploring, annotating, deriving, preserving and presenting. The *careful thought* may be required to cope with issues such as: *a*) finding the right or sufficient data, *b*) coping with missing and poor quality data, *c*) integrating data from diverse sources with significant differences in form and semantics, *d*) understand and dealing with the complexity of the data, *e*) accommodating the ways in which the data and practices are changing, *f*) re-purposing or inventing and implementing analysis strategies that can extract the relevant signal, *g*) planning and composing multiple steps from raw data to presentable answers, *h*) engineering technology that can handle the data scale, the computational complexity and the demand created by many practitioners pursuing myriads of answers, and *i*) making tools and research e-Infrastructure that can be understood and easily used by researchers.

The talks and discussions during the week will mainly focus on how best to use data, and leading experts will be talking about their successful methods and their ways of thinking that work well in their domain. A key challenge is to assimilate the ideas and methods, and to understand how they may be translated to be used in new domains. This will require recognition of clusters of similar data-intensive problems and the clusters of methods that match their characteristics.

The workshop pays less attention to activities such as data creation, collection, storage, replication, archiving and curation so as to have maximum time devoted to data use. That is not to imply that these activities are not important – they are necessary for (future) data use. If the data to which those activities are applied is rarely or poorly used, those activities have little value.

The workshop takes place at a time when there is rapidly increasing interest in data-intensive research, but as Table 1.1 shows that this has been building up over the decades – space only permits us to show a few milestones.

The following conceptual metaphors may be helpful when thinking about data-intensive research:

1. *Digital revolution* A worlwide revolution is underway as individuals, society, business, government, services and education adopt and adapt the advances in digital technology for communications, image capture, sensors, computation and storage. This is more stressful and dramatic than previous revolutions, as it its own mechanisms make it a globally pervasive and extremely rapid process, in contrast with previous revolutions, e.g. the industrial or the advent of printing, which percolated incrementally across communities and countries. Research data use is a minute proportion of this activity, which IDC estimate will involve 1.8 zettabytes ($1.8 * 10^{21}$ bytes) of data in 2011 [27]. Data-intensive research activity may be influential out of all proportion to its size and spend, but it will be more likely to thrive in its niche if it adapts adroitly to the changes in the larger digital ecosystem (see [1] Chapter 5).

2. *Beacons of excellence* There are many centres and groups that are demonstrating repeated success in data-intensive research; e.g. the National Centre for Atmospheric Research (http://www.ncar.ucar.edu), the British Atmospheric Data Centre (http://www.badc.rl.ac.uk), the Institute for Data-Intensive Engineering and Science (see §1.3), the Reading e-Science Institute (see §2.4), the European Bioinformatics Institute (http://www.ebi.ac.uk), and many more. Their characteristic is that they have repeated successes as they apply data-intensive methods repeatedly to meet a succession of research challenges. The fact that some centres and individuals (see §4.3, §3.8 & §4.7 among others) can *repeatedly and rapidly* achieve DIR successes indicates that there are underlying principles that they have grasped. In most cases those principles are not articulated and formalised. Such exposition is necessary to enable wider communities to achieve DIR successes without having comparable requirements for exceptionally talented multi-disciplinary skills. This situation is comparable to the circumstances that triggered the study and practice of software engineering. We hypothesise that a similar study of data-intensive practices, with the corresponding follow through in: theory, empirical evaluation, education, tools and adopted methods, will be necessary for researchers and society at large to gain the full benefits of the growing wealth of data.

Date	**Event**	**Reference**
1971	Transatlantic agreement to share the Protein Data Bank (PDB)	[4]
1993	First Data-Intensive Research Centre	[5]
Feb. 1996	Human genome project (HUGO) Bermuda agreement on sharing data	[6]
Feb. 1997	HUGO Bermuda agreement on sharing data	[7]
1998	HUGO Bermuda agreement on sharing data	[8]
1999	Sloan Digital Sky Survey pioneers large-scale astronomic data publishing	[9]
2000	Stanford Linear Accelerator starts data taking for BaBaR	[10]
2001	Human genome sequence published with studies of 30 genes already published *using releases of pre-publication data*	[11]
2003	HUGO Fort Lauderdale agreement on sharing data	[12]
2006	First five years of SDSS analysed	[13]
2007	NSF solicitation for DataNet projects	bit.ly/aiEIzq
2007	Extremely Large Database (XLDB) workshop	[14]
Mar. 2008	Yahoo-hosted workshop on data-intensive research	bit.ly/dAbx1k
Mar. 2008	SciDB project requirements gathered	scidb.org
2008	Extremely Large Database (XLDB) workshop	[15]
Jan. 2009	Report on GrayWulf architecture and its use	[16, 17]
Jan. 2009	Report of Interagency Working Group on Digital Data to the Committee on Science of the National Science and Technology Council	[18]
Jan. 2009	US President commitment to open data	[19]
Mar. 2009	*Beyond the Data Deluge* paper published	[20]
Jun. 2009	Government commitment to linked data	[21]
2009	Extremely Large Database (XLDB) workshop	[22]
Sept. 2009	Toronto statement clarifying responsibilities on publishing and using biological data	[23]
Sept. 2009	DataNet projects: "DataOne" and "The Data Conservancy" start	dataone.org & www.dataconservancy.org
Sept. 2009	PARADE *Strategy for a European Data Infrastructure* published	[24]
Oct. 2009	*The Fourth Paradigm* book published	[25]
Oct. 2009	JISC call for Research Data Management projects	bit.ly/aEhVJ4
Nov. 2009	NSF CISE solicitation for data-intensive research	bit.ly/dcDsaM
Dec. 2009	European e-Infrastructure Reflection Group data report endorsed	[26]
Dec. 2009	First multi-national *Digging into Data* awards	bit.ly/chTqeC
Mar. 2010	Data-Intensive Research workshop in Edinburgh	bit.ly/bQpu5h
Apr. 2010	*Data-Intensive e-Science Workshop*, in Japan	bit.ly/aRe7lJ
Jun. 2010	Expected first alpha release of SciDB	scidb.org
Jun. 2010	*Third Data-Intensive Distributed-Computing Workshop* at HPDC'10	bit.ly/cJhtwX
Oct. 2010	Extremely Large Database (XLDB) workshop	

Table 1.1: Events shaping Data-Intensive Research (from [1] Chapter 1)

3. *Going the last mile* This term was used when we visited the Earth Data Analysis Centre (edac.unm.edu) to describe their expertise in deriving data from NASA data and presenting it in forms that could be used directly by decision makers in the private and public sectors. It is a key step in achieving useful effects from research – a necessary goal; in the words of USA's Energy Secretary Steven Chu, director of the Lawrence Berkeley National Laboratory: "We seek solutions. . . We don't seek—dare I say this—just scientific papers any more" [28].

4. *Datascopes* We are now at the 400th anniversary of the telescope; look at the way they have changed and how they have changed our view of the universe and our place in it. We need analogous devices to help us see the information in data—possibly to start a similarly adventurous journey (see [1] Chapter 6).

5. *Intellectual ramps* Researchers are usually engrossed in the challenges of their own domain; they focus all of their time and intellectual energy in the race to advance their own discipline. This leaves no time for exploring new technologies and methods on the off-chance that they may yield research benefits. But most of the methods, tools, technology and services for data-intensive research take substantial time and intellectual investment before they pay off. This activation energy and the risk of investing heavily in an endeavour that may have to be abandoned is a common impediment to the adoption of new practices. *Intellectual ramps* are proposed as carefully engineered environments where researchers can incrementally gain benefits as they learn about new methods and the tools that support them. Increasing the R&D investment in developing such ramps and the concomitant education would yield major benefits in uptake and advanced use across a wide range of disciplines—examples of such ramps are relatively rare (see [1] Chapter 7).

6. *Walking a path together* Data-intensive research is invariably interdisciplinary, e.g. mathematicians, statisticians and computer scientists may be working with a range of specialists in a given domain: e.g. they may be seeking to understand: the expansion, path and impact of an oil slick; the atmospheric transport of ash from a volcano and its impact on aviation, the evolutionary development of resistance in a pathogen and strategies for managing an epidemic, etc. A key requirement for such collaboration to succeed is for *sustained communication* across traditional boundaries, so that the ideas, methods and technologies of the partners on either side of each boundary co-evolve as they influence each other. Failure to do this often leads to frustrating experiences, where for example, a computing scientist interacts and believes he has "captured requirements" and then goes away to develop a "solution". By the time they return the others have developed new understanding that has changed their working practices and the "solution" is disregarded—the computing scientist may even be blamed for failing to deliver [29]. But excessive attempt to march in lock step also fails, as it is impossible for each discipline to progress at the same rate, so judiciously loose but still effective collaboration is needed as illustrated in Figure 1.2 (see [1] Chapter 8).

1.3 Strategies for exploiting large data

Alex Szalay, Johns Hopkins University
The talk is available at: http://www.nesc.ac.uk/action/esi/download.cfm?index=4429.

The volume of scientific data is doubling every year as a result of cheaper and faster digital

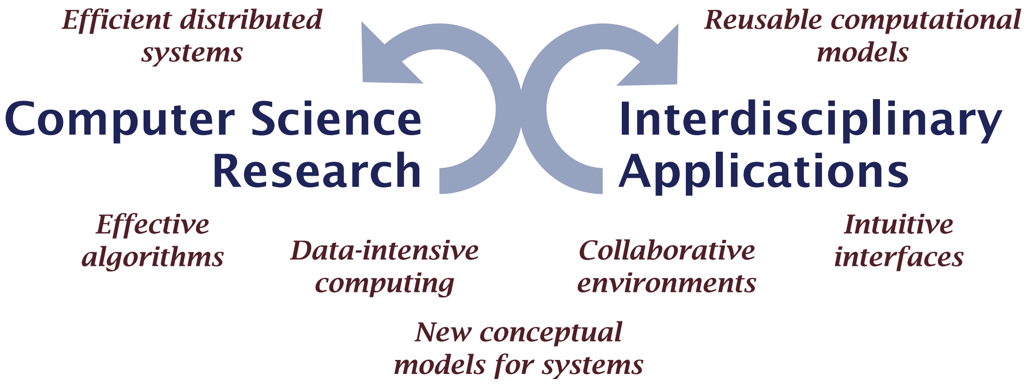
Efficient distributed systems
Reusable computational models
Computer Science Research
Interdisciplinary Applications
Effective algorithms
Data-intensive computing
Collaborative environments
Intuitive interfaces
New conceptual models for systems

Figure 1.2: Computer science and application research co-evolve to form data-intensive research whilst keeping their distinct identities—this is important for career development in the current framework for research credit.

detectors and sensors. This growth is two-dimensional – individual collections are growing and the number of collections is growing – reaching the long-tail of scientists using the many smaller data collections is a particular challenge. This wealth of data is transforming the way science is undertaken; a new paradigm of data-intensive science is emerging, cutting across all scientific disciplines [25].

Managing the data for Sloan Digital Sky Survey (SDSS) so that it was accessible to astronomers worldwide presented significant challenges—work undertaken with Jim Gray [30, 13]. The statistics are: 40 TB raw data, 5TB catalogues, 2.5 Terapixels, 930,000 distinct users (globally there are $\approx 10^4$ professional astronomers) and 10^8 rows of data delivered; but everything is a power law and demand doubles in every dimension. The SDSS supports GalaxyZoo where $2.5 * 10^5$ volunteers have classified $6 * 10^7$ galaxies (http://www.galaxyzoo.org).

The SDSS experience has led to two lines of research development at the Institute for Data-Intensive Engineering and Science (IDIES): *a*) *further applications* using and developing data-intensive methods, and *b*) *new computing architectures* that better support data-intensive applications.

The further applications include:

- the National Virtual Observatory (http://www.us-vo.org/),
- PanSTARRS (http://pan-starrs.ifa.hawaii.edu) [31],
- Immersive Turbulence simulation data (http://turbulence.pha.jhu.edu) [32, 33] and
- Sensor Networks (http://lifeunderyourfeet.org [34, 35].

The new computational architectures include:

- GrayWulf to balance IO performance with CPU capacity [16, 17], and
- Amdahl Blades to balance IO and IO operation rates with CPU capacity, and reduce energy use [36].

1.4 Motivation and Strategies for data-intensive Biology

Douglas Kell, Chief Executive, Biotechnology and Biological Sciences Research Council and Manchester University
The talk is available at: http://www.nesc.ac.uk/action/esi/download.cfm?index=4430.

Important relationships exist between the world of ideas and the world of data. Much of biology, especially molecular biology, has been hypothesis-driven (or hypothesis-dependent) in that this relation is seen as starting with an idea and seeing if data obtained are consistent with it – the Popperian hypothetico-deductive falsificationist agenda. By contrast, data-driven methods are inductive in character, start with the data and then find the idea (hypothesis) that best fits those data. Of course it is really an iterative cycle [37], and data-driven methods are becoming much more important in the post-genomic era. Omics data and Systems biology approaches provide particular examples of this [38, 39], including in collaborative knowledge generation [40].

Biology, especially genomics, is now the big data science, with many PBytes being produced annually. This brings biology into the era of data-intensive science [25]. Most problems in biology are combinatorial anyway.

BBSRC has long recognised these kinds of developments, and addressing them proactively is part of our recently launched Strategic Plan http://www.bbsrc.ac.uk/strategy/, including a £13 million plus investment in The Genome Analysis Centre (http://tgac.bbsrc.ac.uk/) in Norwich. We recognise, for instance, that we now need to bring the computing to the data, and not the other way round as was traditional. There are also considerable training and skills needs, for both curators and users.

BBSRC is also the lead for the UK node of the ELIXIR (European Life Sciences Infrastructure For Biological Information) project http://www.elixir-europe.org/page.php.

BBSRC pioneered principled data-sharing policies (making data publicly available is a standard Condition of Grant), and we have ring-fenced funding streams for data resources and their curation.

The peer-reviewed biomedical literature, a very particular and important kind of (textual) data resource, is increasing at the rate of some 2 papers per minute, and this presents special challenges and opportunities [41]. Integrating this kind of material remains challenging; synthesising the literature showing that unliganded iron is the main 'cause' of (or intermediary in) just about every degenerative disease [42] took me most of my 'spare' time for a year, and other literature-synthesising reviews such as those on pharmaceutical drug transporters [43, 44, 45, 46] are equally demanding.

1.5 Analysing and Modelling Large-Scale Enterprise Data

Thore Graepel, Microsoft Research Cambridge
The talk is available at: http://www.nesc.ac.uk/action/esi/download.cfm?index=4431.

This talk is concerned with the challenges that arise from analysing large-scale complex streams of data generated through online services at Microsoft. Our work aims at improving Microsoft's online services by developing and deploying machine-learning algorithms at scale. We work in the framework of graphical models for probabilistic modelling and perform approximate inference by message passing on very large factor graph models whose structure is determined by the data.

The interaction with data typically happens in two stages: In the first, static stage, we do data mining and try to find patterns in the data that can serve as predictive features for the machine learning algorithms. In the second stage, the algorithm is deployed into the production system and becomes part of a complex loop in which the predictions of the algorithm determine the future composition of the data it is trained on.

I will describe two algorithms that now operate at large scale within Microsoft's online services: TrueSkill$^{\mathrm{TM}}$ is the system responsible for skill estimation and matchmaking in the Xbox$^{\mathrm{TM}}$ Live online gaming system with over 20M users. AdPredictor$^{\mathrm{TM}}$ is the click-through rate (CTR) prediction system for Sponsored Search advertising in Microsoft's search engine Bing$^{\mathrm{TM}}$, and plays a key role in targeting ads to users.

1.6 Research Village

The research village allowed nine research groups to show their work in progress. The idea was that each research stall in the village set out its wares to attract the gathered throng; and every 15 minutes a gong indicated it was time to move to a new stall. This started conversations that continued throughout the week with many participants trying out things they had seen here for the first time.

The stalls were:

1. **Meandre**: The work at NCSA on Cloud, Meandre and Zigzag (http://seasr.org/ and http://seasr.org/meandre/) was demonstrated by Xavier Llorà and Bernard A'cs (see §1.10).

2. **Taverna, myExperiment and Biocatalogue** (http://www.taverna.org.uk/, http://www.myexperiment.org/ and http://www.biocatalogue.org) from projects led by Carole Goble and Dave De Roure, were presented by Katy Wolstencroft and Peter Li (see also §2.9).

   The Taverna Workbench is an open-source software tool for designing and executing workflows. Developed under the e-Science programme through myGrid and OMII-UK, Taverna enjoys very wide adoption and is currently used by over 350 academic and commercial organisations throughout the world. Workflows capture explicit methodologies for the systematic analysis of scientific data and facilitate communication of methods across research disciplines. Taverna has a substantial user base in bioinformatics, and is used by data-intensive researchers from chemistry to computational musicology.

   myExperiment is a social web site, which lets users discover, publish and share workflows. It has the largest public collection of scientific workflows (nearly 1000) and a membership of over 3000 researchers. The myExperiment web 2.0 service enables researchers to re-use and re-purpose workflows, reducing workflow re-invention whilst facilitating scientific collaborations and pronulgation of research expertise. Taverna tightly integrates with myExperiment to bring the full experience of workflow browsing into the Workbench.

   BioCatalogue is a registry of biological Web Services for use in workflows. It provides an open platform for biological Web Services registration, annotation and monitoring; providing a resource for community curation.

3. **The OMII-UK and Software Sustainability Institute** (http://www.omii.ac.uk/ and http://www.software.ac.uk) led by their director, Neil Chue Hong, was demonstrated by Ally Hume (OGSA-DAI), George Beckett / James Perry (DiGS), Terry Sloan (SPRINT).

   OMII-UK has cultivated many of the leading e-Science software tools, supporting their development through initiatives such as the Commissioned Software and ENGAGE programmes. Three pieces of software, of particular relevance to data-intensive research, were demonstrated: **OGSA-DAI** (http://www.ogsa-dai.org.uk) is an extensible framework for integrating data from heterogeneous sources; **DiGS**[1] is a distributed data management system featuring replication, validation and consistency checking; and **DataMINX** (http://www.dataminx.org) manages data transfers between all major grid data storage systems.

   The Software Sustainability Institute **SSI** has been created to work in partnership with research communities to identify key software that needs to be sustained, and makes software not just available but useful for researchers by improving usability, quality and maintenance. The **SPRINT Parallel R framework** (http://www.r-sprint.org), created by the Department of Pathway Medicine and EPCC was demonstrated. It enables statistical analyses written in R to be run on high-performance computing systems without specialist parallel-computing knowledge.

4. **Discovery Net and InforSense(IDBS)** (http://www.idbs.com/ and http://www.doc.ic.ac.uk/˜yg) led by Yike Guo and presented by Katie McMurray and Anthony Rowe.

   The Discovery Net project and InforSense were one of the early pioneers in using workflow technologies as a high-level programming framework for data-intensive scientific applications, as was highlighted by winning the Most Innovative Data Intensive Application at the 2002 Supercomputing conference. Now part of IDBS, on-going research and commercial success has shown that workflow is only one vital component of the Data-Intensive Application Stack. One of the significant trends in Translational and Biomarker research is the forming of large consortia of Academic, Biotechnology and Pharmaceutical companies with a common research aim. Structurally, these projects are data-intensive virtual organisations and this requires the ability for research software systems to not just process data for a single organisation, but also to store, analyse and visualise data in an architecture that enables these distributed consortia to work effectively on the same project data. The Discovery Cloud system, shown for the first time at this conference is an ongoing study in how a highly virtualised and elastic cloud-based architecture enables the rapid construction of virtual data-intensive application stacks to support these styles of project.

5. **Data-Intensive Research, Edinburgh** (http://research.nesc.ac.uk) led by Jano van Hemert with presenters Lianxiu Chan, Gagarine Yaikhom, Jos Koetsier and Rob Kitchen.

   They presented the group's research and technology; the group does interdisciplinary research to progress methods in computer science and tackle data-intensive challenges in diverse areas of science and business including the following:

   - Effective algorithms for data analysis, data mining and combinatorial optimisation.
   - Distributed and data-intensive systems for efficient orchestration of data and computation.

[1] http://ukqcd.epcc.ed.ac.uk/training/2008/talks/DiGS.pdf

- Reusable components and new conceptual models for systems that can be deployed across disciplines.
- Intuitive interfaces and collaboration environments to enable domain-specific researchers to make use of the above systems.

The work was illustrated with applications from seismology, brain imaging, computational chemistry, breast cancer and developmental biology.

6. **MonetDB** (http://monetdb.cwi.nl) led by Martin Kersten and presented by Milena Ivanova (see also §2.5).

   The data-intensity of modern sciences raises challenging data-management issues for the research community in terms of scalability, functionality and performance. The usage pattern of scientific data warehouses is characterised by long periods of analysis of large data volumes, intermixed with regular bulk-load of new data. Analytical applications are often disk-bound since extensive computations and aggregations may span the entire data set or large portions of it.

   Research on column-store databases has already indicated their advantages in comparison with traditional row-store systems. Vertical organisation provides for more efficient data access pattern for disk-bound queries, flexibility in the presence of changing workloads, and may also reduce data redundancy and storage needs.

   MonetDB is an open-source column-store database management system, developed at CWI. MonetDB is distinguished by several characteristics that together provide high performance for analytical workloads. The demonstrations showed the advantages MonetDB offers for analytical scientific applications.

   Besides the benefits stemming from the vertical organisation, the execution paradigm of MonetDB is based on full materialisation of intermediate results. This opens an opportunity to speed up query sequences with overlapping computations by careful preservation and re-use of common intermediates. Commonalities are often observed in logs of scientific activities, where collaborating or competing teams may perform similar, but slightly different analyses [47, 48, 49, 50].

7. **Centre for Advanced Spatial Analysis, UCL** (http://www.casa.ucl.ac.uk) led by Michael Batty and presented by Steven Gray.

   With the recent popularity of web-based mapping systems, the visualisation and analysis of spatial data is becoming increasingly important. In our MapTube (http://www.maptube.org) website we allow users to upload and compare thematic data on top of the regular Google Maps or OpenStreetMap layers. Extending this idea to data collection as well as visualisation, we demonstrate the concept of a “Mood Map” where members of the public answer a question which we ask via an online form. An example might be, “What single factor is affecting you most about the credit crunch?” with a single answer chosen from one of: “Mortgage or Rent”, “Petrol”, “Food Prices”, “Job Security”, “Utility Bills” or “Not Affected”. The first part of their postcode is also entered and it is this that is used to build a map of the responses by postcode district.

   The first mood maps were limited in that only we could set them up and that they only updated every half hour. With the ASK project, which is funded by the National e-Infrastructure for Social Simulation (NeISS), we aim to build a more flexible architecture for large-scale crowd sourcing of spatial data. This project will allow ordinary users to set up their own mood map surveys and see the results in real-time.

In addition to the online survey idea behind the mood maps, we have also looked at using other popular social networking sites to extract spatially tagged data. We demonstrate the "Tweetometer", which counts tweets on the Twitter site containing the keyword "LondonS and show how this can be graphed in real-time to show activity. Our system was used to crowd-source a map of snowfall in the UK during December and January, and it was also used by Carling for use in the Carling Cup Final, illustrating the wider use of e-Science research.

8. **Blackford Analysis P Instant 3D Registration** (http://www.blackfordanalysis.com) was led and presented by Alan Heveans and Ben Panter (see also §2.2).

   Blackford Analysis is a spinout group from the Institute for Astronomy at the University of Edinburgh, applying their MOPED technology to problems involving large datasets. MOPED was originally developed to interpret galaxy spectra, although is a general technique appropriate in many situations involving parametric modelling. The Blackford Analysis team demonstrated an application that provides real-time registration capability for medical imaging.

   Modern methods for diagnostic imaging result in large datasets. Data is acquired in 3D and typically comprises hundreds of slices, each at 512×512 resolution. The simple sounding task of calculating the affine transform between two scans involves the fitting of 12 parameters in over 32 million voxels. Although high performance computers can tackle the problem in a reasonable time (5-10 minutes) on today's datasets, as resolution increases the problem rapidly becomes intractable.

   Blackford Analysis use the MOPED approach to compress the data into a form that can be tackled quickly and efficiently – two medical scans can be aligned in less than a second, using a standard laptop. This step change in performance allows real-time registration of images, with many applications giving increases in patient throughput. Medical imaging is only one of the potential areas where MOPED is useful, and the Blackford team is very keen to discuss any problems of interest to the data-intensive research community.

9. **e-Research South – Lab blog book systems** (http://www.eresearchsouth.ac.uk) led by Anne Trefethen and Jeremy Frey and presented by David De Roure.

   The e-Research South Consortium is building a vibrant regional activity driven by specific application areas and building on existing e-Infrastructure technologies to enhance ease of use, uptake and 'accessibility', to develop know-how and tools for dealing with research data, to provide development of and access to advanced visualisation, and to provide opportunities for public engagement with science.

   "Southampton Smart Labs Systems" is one of these activities and provides a solution for data management for experimental and computational researchers for PublicationSource and Data on Demand. They demonstrated the concept of Blog-style systems as Laboratory Notebooks, using our current (Blog$_2$) systems and the new semantically rich system based on the semantic & web 2.0 Blog$_3$ together with the ORECChem Ontology of experiments.

   The blog book systems are linked to the flow of data from laboratories and experiments and we deliver fully traceable experimental data flowing from our laser surface experiments. These are self-describing data created by the Blog and Laboratory systems enables the use of semantic tools (such as the MIT Simile Software) to display and use the data.

1.7 Data Analysis Theme Opening

Chris Williams, School of Informatics, University of Edinburgh
The talk is available at: http://www.nesc.ac.uk/action/esi/download.cfm?index=4436.

Generally the most important reason for collecting data is in order to analyse it, either to understand its structure, or in order to make predictions. Although different disciplines will have different specific analysis goals, there are problems and solutions that arise with regard to the organisation and analysis of data across a wide range of application domains. *Data-centric thinking* allows us to understand and gain strength from these similarities. Such statistical and algorithmic analysis methods help us address the problem of "drowning in information, but starving for knowledge" (Naisbett, 1982 [51]).

The analysis of data typically begins with data "cleaning", dealing e.g. with missing or corrupted data. It is recognised that this step can take up 50-80% of the time in real-world data mining problems. Even if manual correction of the data is possible it is prohibitively expensive for large datasets, giving rise e.g. to challenges around tools to improve the automation of data cleaning, or the integration of data cleaning steps into the analysis itself.

Beyond data cleaning, data analysis can be divided into exploratory data analysis, descriptive modelling and predictive modelling. These techniques derive from statistics, machine learning and data mining. Exploratory data analysis refers mainly to visualisation of datasets, descriptive modelling to methods such as clustering where the aim is to understand the inter-relationships between measurements (unsupervised learning), and predictive modelling to the supervised task of predicting one or more variables on the basis of others.

There are several different dimensions to consider for these analysis tasks, e.g. with respect to *complexity*, *data quality*, the *incorporation of prior knowledge*, and *scale*. Here complexity can refer to the complexity of the model being used for analysis (e.g. network/circuit models in systems biology or social network analysis), or to complexity in the data, e.g. arising from the integration of multiple data sources. With regard to data quality there can be a trade-off between large amounts of "dirty" data, or smaller amounts of cleaner data, which might arise through data curation. The incorporation of prior knowledge is very important in scientific applications; one way that this can arise is through the use of structured probabilistic graphical models to encode (at least some of) the domain knowledge. With regard to data scale, there is often an abundance of raw data, and technological advances mean that this will grow rapidly. In some cases it is possible to deal with the volume simply by sampling the data rather than processing it all.

Over the course of the meeting we seek to gain an understanding of the data analysis challenges faced in different domains, and to find out effective ways to train personnel in data-centric analysis techniques in order to make scientific progress. Often this will require data experts working with domain experts in order to develop new techniques to address specific structure or features of the problem domain.

1.8 Data-structuring paradigms Theme Opening

Stratis Viglas, School of Informatics, University of Edinburgh
The talk is available at: http://www.nesc.ac.uk/action/esi/download.cfm?index=4433.

When dealing with structured data, the best known processing paradigm is that of a Relational Database Management System (RDBMS). RDBMSs have been around for some thirty years and are responsible for managing the vast majority of the world's structured data — and some of its unstructured data as well. Moreover, the architecture of RDBMSs has been adapted to allow high degrees or parallelism so that the system scales gracefully and linearly as the data volume increases, and as more computational resources in the form of processing nodes become available. However, an RDBMSs is not omnipresent in a large number of data-intensive research applications. Or, if it is there, it is merely acting as a backing store with its full querying capabilities not exploited. Finally, different processing paradigms have emerged that supposedly supercede the functionality offered by RDBMSs.

There has been a trend towards moving away from the use of RDBMs as the structured approach to dealing with data intensity. The arguments behind this trend is that most of the guarantees that a RDBMS gives may be too rigid for certain types of processing. For instance, eventual data consistency is preferred to the strong consistency an RDBMS provided; or systems need to scale mainly horizontally, *i.e.*, in terms of processing or storage nodes, and not so much in terms of data volume. However, most of these arguments have counter-arguments: the consistency guarantees of an RDBMS can be lifted, while they have been shown to be massively parallelisable both in terms of their data processing algorithms and in terms of their architecture.

Over the course of the workshop the aim is to answer questions like the following: Is there something fundamentally wrong with the architecture of a massively parallel RDBMS that fails to address the needs for data-intensive processing? Is it the case that RDBMSs have been optimised for a type of processing that is not the type of processing needed by data-intensive research? Are new processing paradigms like MapReduce fundamentally inapplicable on an RDBMS-like system architecture? And does the functionality offered by distributed file systems like HDFS or processing paradigms like MapReduce supercede the functionality of an RDBMS, or are they merely a subset?

1.9 Programming Paradigms Theme Opening

Geoffrey Fox, Indiana University
The talk is available at: http://www.nesc.ac.uk/action/esi/download.cfm?index=4437.

The different approaches to programming simulations are well understood although there is still much progress to be made in developing powerful high-level languages. Today OpenMP [52] and MPI [53] dominate the runtime used in large-scale simulations and the programming is typically performed at the same level in spite of intense research on sophisticated compilers. One also uses workflow to integrate multiple simulations and data sources together. This coarse grain programming level usually involves distributed systems with much research over last ten years on the appropriate protocols and runtime. In this regard Globus and SAGA represent important advances.

We can ask what the analogous programming paradigms and runtime are for data-intensive applications. We already know that many of the distributed system ideas will carry over as workflow has typically used dataflow concepts and integrated data and simulations. However as data processing becomes a larger part of the whole problem either in terms of data size or data-mining/processing/analytics, we can anticipate new paradigms becoming important. For example most data analytics involves (full matrix) linear algebra or graph algorithms (and packages like R) and not the particle dynamics and partial differential equation solvers characteristics of much supercomputer use. Further storage and access to the data naturally involves database and distributed file systems as an integral part of the problem. It has also been found that much data processing is less closely coupled than traditional simulations and is often suitable for dataflow at runtime and specification by functional languages. However we lack an authoritative analysis of data intensive applications in terms of issues like ease of programming, performance (real-time latency, CPU use), fault tolerance, and ease of implementation on dynamic distributed resources.

A lot of progress has been made with the MapReduce framework originally developed for information retrieval – a really enormous data intensive application. Initial research shows this is a really promising approach to much scientific data analysis. Here we see different choices to be explored with different distributed file systems (such as HDFS for Hadoop) supporting MapReduce variants and DryadLINQ offering an elegant database interface.

Note current supercomputing environments do not support HDFS but rather wide area file systems like LUSTRE – what is a possible resolution of this? MapReduce programming models offer better fault tolerance and dynamic flexibility than MPI and so should be used in loose coupling problems in preference to MPI. Parallel BLAST is a good example.

1.9.1 Clouds: A Brief Overview

Cloud computing is at the peak of the Gartner technology hype curve [2] but there are good reasons to believe that as it matures that it will not disappear into their trough of disillusionment but rather move into the plateau of productivity as have for example service-oriented architectures. Clouds are driven by large commercial markets where IDC estimates that clouds will represent 14% of IT expenditure in 2012 and there is rapidly growing interest from government and industry. There are several reasons why clouds should be important for large-scale scientific computing

- Clouds are the largest scale computer centres and so they have the capacity to be important to large-scale science computations as well as those at smaller scales.
- Clouds exploit the economies of this scale and so can be expected to be a cost effective approach to computing. Their architecture explicitly addresses fault tolerance.
- Clouds are commercially supported and so one can expect reasonably robust software without the sustainability difficulties seen from the academic software systems critical to much current Cyberinfrastructure.
- There are 3 major vendors of clouds (Amazon, Google & Microsoft) and many other infrastructure and software cloud technology vendors including Eucalyptus Systems that spun off from UC Santa Barbara HPC research. This competition should ensure that clouds should develop in a healthy innovative fashion. Further attention is already being given to cloud standards
- There are many Cloud research, conferences and other activities with research cloud infrastructure efforts including Nimbus [54] OpenNebula [55], Sector/Sphere [56] and Eucalyptus [57].

[2] [2] Press Release Gartner's 2009 Hype Cycle Special Report Evaluates Maturity of 1,650 Technologies http://www.gartner.com/it/page.jsp?id=1124212.

- There are a growing number of academic and science cloud systems supporting users through NSF Programs for Google/IBM and Microsoft Azure systems. In NSF OCI, FutureGrid [58] will offer a Cloud testbed and Magellan [59] is a major DoE experimental cloud system. The EU framework 7 project VENUS-C [60] is just starting.
- Clouds offer "on-demand" and interactive computing that is more attractive than batch systems to many users.

Generalising from this, clouds are more important for data-intensive applications than classic simulations, as the latter are very sensitive to synchronisation costs, which are higher in clouds than traditional clusters.

1.10 Soaring through clouds with Meandre

Xavier Llorà and Bernie A'cs, NCSA
The talk is available at: http://www.nesc.ac.uk/action/esi/download.cfm?index=4435.

'Cloud computing' is a broad term with multiple interpretations. Hence, it is important to set the tone and the basic terminology used throughout the presentation. To achieve this goal and explain the current efforts on cloud computing at the National Center for Supercomputing Applications (NCSA) we divide the material into two sections. First we review how to build a conceptual and functional foundation for what is represented by computational cloud construction; then, we present a quick overview of the highly scalable, data-intensive infrastructure being developed at the centre and the current research efforts to tackle the challenges presented by big-data.

The research efforts highlighted include the experimental virtual-machine infrastructure and enterprise-cloud configuration that have lead to a conceptual exploration into developing cloud and a software infrastructure for data-intensive computing that is component-based and data-driven that is not bound to a particular development language but rather focused on the core functional requirements to provide flexible modular building blocks. These building blocks scale transparently, thanks to the abstractions that components provide, with no coding modifications when executed on a laptop, large computational grids, or clouds though the architectural models contrast significantly. These independent, but highly related, research efforts show that new approaches are needed for developing flexible, scalable, and functional applications that use diverse computational infrastructures for data-intensive research.

Meandre provides a data-intensive execution engine, component-based programming architecture, and data-driven execution paradigm that enable distributed data-flow designs so that processing can be transparently co-located with data sources. It also enables transparent scalability. The design of programming paradigms that promote new approaches to parallel processing for massive distributed data resources is a grand challenge for data-intensive research. The data-deluge in every domain of science and humanities demands a modular and transparently scalable application framework. Such a framework must be able to transcend physical system architectures to take advantage of a wide variety of functional features available in the cloud and those that are exposed in traditional HPC grids and other special purpose computational cluster configurations.

Chapter 2

Data-Volume's Challenge

2.1 Overview

Data and scale are overloaded terms and mean different things to different practitioners depending on their perspectives and goals.

In terms of data, the semantics is usually dependent on the application domain. It might refer to a corpus of unstructured text to someone working on text mining; or, it might refer to a time-series of sensor measurements to a geologist; or, a multitude of input/output parameters to a physicist; or, it might refer to a fully structured relational schema to a database researcher; the possibilities are as endless as the domains. At the end of the day, data is what all our systems process and produce.

Scale also comes in flavours. Is it the disk capacity needed to store the data? Is it the rate at which data is produced? Does it represent the number of complex interactions across different types of data? Or is it the number of concurrent parallel processes we can throw at a data-intensive problem? What we want is our systems to be scalable in a number of dimensions — some of which we might not have even identified yet, let alone comprehend.

There is a saying that "a craftsman is only as good as his tools." To some extent, we are all craftsmen and we have to pick the right tool for the job. Tools differ across crafts and no two jobs within the same craft are the same. The goal of the day is to be exposed to different "crafts" and listen to the jobs they include and the tools that are better suited for them. It is most likely the case that the saying will be verified once more. But what will hopefully come out of this exposition is the knowledge of some tool another craftsman has been using to do a job similar to what we have to deal with in our craft.

To that end, we had a series of talks in the morning:

- Astronomers produce vast quantities of data; Alan Heavens (University of Edinburgh) talked about the challenges in analysing large sky datasets, and how the statistical techniques developed for doing so are finding applications in other domains – see §2.2.
- However, we do not need to look to the endless sky for data challenges, when similar quantities and complex interactions can be found here on Earth; Torild van Eck (Royal

Netherlands Meteorology Institute) talked about the data challenges in earthquake seismology – see §2.3.

- Keith Haines (University of Reading) talked about less catastrophic elements of our planet: climate and oceanographic data and how these can be integrated into a single view – see §2.4.
- Arguably, relational databases are one of the *de facto* mechanisms to support data storage and manipulation. Martin Kersten (CWI) is the godfather of one of the most impressive open-source database systems around: MonetDB. He presented the challenges in developing database support for generic data processing – see §2.5.
- A collection of data usually adheres to a rudimentary data model. Is that always true? Jonty Rougier (University of Bristol) talked about the challenges of uncertainty in data modelling – see §2.6.
- In addition to long-term storage and long-running processing of data, we might be interested in real-time analysis. Beth Plale (Indiana University) presented the challenges and methods in this area – see §2.7.
- Finally, Michael Batty (University College London) talked about the use of complex geospatial data to understand geographically correlated data and to aid long-term planning – see §2.8.

2.2 Dealing with large data sets in astronomy and medical imaging

Alan Heavens, Institute for Astronomy, University of Edinburgh
The talk is available at: http://www.nesc.ac.uk/action/esi/download.cfm?index=4441.

Typical analysis problems involving large datasets are focussed on learning a relatively small number of pieces of key information. This imbalance between the size of the data set and the number of parameters to be extracted offers substantial opportunities for fast, accurate analysis via massive data compression techniques. In this talk, I will cover some such techniques, showing how by carefully-designed data compression, one can massively speed up many analysis problems without compromising accuracy, focussing on the patented MOPED algorithm [61].

2.3 Data challenges in Earthquake Seismology

Torild van Eck, The Royal Netherlands Meteorological Institute, Observatories and Research Facilities for European Seismology (ORFEUS) and the International Federation of Digital Seismograph Networks (FDSN)
The talk is available at: http://www.nesc.ac.uk/action/esi/download.cfm?index=4438.

Earthquake seismology research is largely based on data obtained from a distributed, widely diverse network of stations, permanently in operation, and temporary deployments on the sea-bottom and on land. Although the data ownership and the equipment are diverse, the data exchange format is well defined (SEED). This facilitates the data access for research. We do not have a comprehensive overview of all stations, however, a conservative estimate is about 5–6 thousand high quality permanent stations around the globe. Mobile stations, both on land and on the sea-bottom (OBS), account for another 5–6 thousand stations. Accelerometers, providing also earthquake engineering relevant data, account for another 5–10 thousand instruments. Within the FDSN, several data centres (IRIS in the US, ORFEUS data centres in Europe, JAMSTEC in Japan, etc) provide jointly open (real-time) access to about 20–30% of this data. Currently the number of openly available data is rising rapidly.

Earth tomography models, earthquake rupture process reconstructions, advanced seismic hazard models and warning systems are largely based on this data. For areas with dense networks impressive models of the upper mantle and crust can be obtained for example. However, as more data is becoming available and more details of the earth's internal structure are sought for, effective data management becomes an important bottleneck. Efficient data archiving, exchange, quality control, management, information mining, etc. require advanced techniques, not easily available to scientists not residing in one of the few excellently equipped labs. Data access procedures, currently supported by data centres, are mainly relying on robust simple data request routines. The fairly recent involvement of IT experts in seismology has resulted in experimenting with data provenance, integration of web services, new data mining techniques, workflow orchestration, for example. The current challenge in earthquake seismology is to create an efficient E-science environment in earthquake seismology that is robust, easy to handle and maintain, and providing advanced analysis techniques to a broad community of scientists and students.

Relevant organisations, projects and initiatives:

- ORFEUS P Observatories and Research Facilities for European Seismology (http://www.orfeus-eu.org)
- FDSN P International Federation of Digital Seismograph Networks (http://www.fdsn.org)
- NERIES P Network of Research Infrastructures for European Seismology (http://www.neries-eu.org) EC-infrastructure project
- EPOS P European Plate Observatory System (http://www.epos-eu.org) European Research Infrastructures proposal
- VEBSN P Virtual European Broadband Seismic Network (http://www.orfeus-eu.org/Data-info/vebsn.html)
- ESFRI P European Strategy Forum on Research Infrastructures (http://cordis.europa.eu/esfri/)

2.4 Making the most of Earth-systems' data

Keith Haines, Reading e-Science Centre, University of Reading
The talk is available at: http://www.nesc.ac.uk/action/esi/download.cfm?index=4439.

The challenge faced by the earth science community is to deal with both very large dataset, e.g. from numerical models and from satellites, as well as small but complicated datasets as from *in situ* monitoring instruments. Data need to be visualised efficiently, preferably directly from the data centres where they are stored, and brought together for quality control and intercomparison and validation of the modelling components of the earth system. In Reading we have developed a program aimed first and visualisation of large remote geophysical data, and we are in the process of extending this to exploration of multiple datasets simultaneously in order to achieve these aims. The aim is to push this integration as far as is practical using Open-Source tools and internet browser visualisation capabilities and to keep an interactive experience for the scientists exploring the data. Some of the successes and remaining challenges were discussed.

2.5 Scientific Databases: the story behind the scenes

Martin Kersten, CWI
The talk is available at: http://www.nesc.ac.uk/action/esi/download.cfm?index=4442.

In this presentation several stepping stones to address the science data stream are examined from the viewpoint of a database software architect. Starting from the experiences gained in building sustainable software to aid astronomy, and exploring the requirements for various fields, we reflect upon recent developments in the database community.

Topics reflected upon are: multimedia data, geospatial data, XML, RDF and the semantic crowd, sensors, hyped grid and map-reduce.

Although significant technological steps have been taken to accommodate the needs in science, there are still several major hurdles to be overcome when it comes to efficient and effective data management. The approach taken in the MonetDB project (monetdb.cwi.nl) [49] includes column-based storage, and recycling of intermediate results [62, 48, 50]. It has accommodated the SDSS data (see §1.3) [63] and is being used for a wide range of applications including: EMILI (http://emili-project.eu) on high-performance streaming database technology, TELEIOS on support for remote sensing, and LOD2 and PlanetData, both geared at scaleable Linked Open Data for the semantic web. Further work, in the area of on-the-fly, just-in-time, non-guided database summarisation is needed.

2.6 Model limitations: sequential data assimilation with uncertain static parameters

Jonathan Rougier, Department of Mathematics, University of Bristol
The talk is available at: http://www.nesc.ac.uk/action/esi/download.cfm?index=4440.

Many large datasets arise as a result of experiments and must be interpreted in the light of scientific theories. These theories are instantiated as models, which relate the observable quantities with additional unobserved or unobservable (because not operationally defined) quantities. In statistics these additional quantities are termed ‘parameters’, and the purpose of inference is to learn about them using the observables. The incompleteness of scientific theories, and in particular their often restricted domains, imply that the models have limitations, and these are often substantial, e.g. in environmental models.

If we account fully for model limitations then we must allow model parameters to be uncertain, and incorporate this uncertainty into the data assimilation process. Such parameters are termed ‘static’, since they do not evolve in time. We must also incorporate structural uncertainty, which enters as a stochastic component in the state equation. Learning about static parameters in a stochastic dynamical model is particularly challenging, as the uncertain state vector trajectory represents a large collection of uncertain quantities that sit between the observations and the static parameters we would like to learn about – in statistics these would be referred to as

Level	Name	What it means in e-Science
1	Intellectual & Technical Metadata	Ownership, intellectual property, copyright and domain-specific attributes
2	Structural Metadata	Data products and aggregations, also semantic information for interpreting metadata through CF vocabulary or ontology
3	Provenance	Lineage of data products as well as that of processes
4	Rendering Software	Domain-specific applications & dependency libraries
5	Processing Software	

Table 2.1: Levels of preservation information

nuisance parameters. 'Large' in this case is 10^2 or 10^3; only a few KB of data, but this is the curse of dimensionality that affects all forms of inference.

Recently developments in statistical computing have shown how to tackle these types of problem using "exact approximations"; which is to say approximations that can be embedded within a sequential inference to make them exact, if one is prepared to wait long enough. These calculations are ideally suited to computer clusters as they contain large chunks of embarrassingly-parallel code. These methods have been shown to perform well on toy problems, and are now been developed for much larger-scale problems. The original result for these developments was [64], with extensions and more thorough theoretical treatment in [65] and [66].

2.7 Earth-Systems data in real-time applications: low-latency, metadata, and preservation

Beth Plale, Indiana University
The talk is available at: http://www.nesc.ac.uk/action/esi/download.cfm?index=4443.

Beth and Keith's write up provide a simple effective model sj

Geoinformatics is having a growing impact on the sustainability of our planet as the amount of data available from environmental sensors, instruments, and satellites continues to increase, and tools become available that allow the integration of social data and field studies with observation sources. Our research advances geosciences through tools for accessing integrated atmospheric data, and for cyberinfrastructure that lowers the barriers to running complex weather-forecast scenarios.

The application space in which we have worked for many years is mesoscale meteorology [67], where weather forecasting requires the assimilation of large amounts of data from dozens of sources with as low a latency as possible between the time at which an observation is taken and the time at which it is used ingested for assimilation. We have explored alternate programming models that allow a researcher to plug real-time data sources into a workflow. In recent work, we preserve the graph abstraction that workflow systems support often through a graphical user interface from which a user can edit an existing workflow graph or construct a completely new workflow. The abstraction we introduce is that of new workflow node types that support input and output edges having time-bounded event streams instead of a single input [68].

Our second area of study is based around the belief of ours that *full reproducibility of a complex investigation, such as a workflow execution, is impractical and quickly becomes impossible.* Further, the workflows we have studied [69] are complicated graphs that yield a result or set of results that can only be understood after-the-fact by gathering information including: prior actions (lineage or provenance), input files, intermediate files, and configuration parameters. It is not practical to expect a scientist will carefully annotate a result with this information. Thus if we are to achieve a level of sharing of these research objects we must have tools that collect provenance and metadata automatically [70].

A research object, or set of objects, resulting from a workflow run may or may not have value, but if it is determined it does, the research object is used to generate a preservation object that can then be published to a long-term archive. We argue that the preservation object needs to include the information in levels 1–3 in Table 2.1[1] prior to being handed off to a long-term repository. This includes the information in the column entitled "What it means for e-Science", including attribution and distribution metadata, data products themselves, semantic information, provenance information, service descriptions, and perhaps even source code. This is explored in more detail in [72].

2.8 Challenges in Large-Scale GeoSpatial Data Analysis: Mapping, 3D and the GeoSpatial Web

Michael Batty, Centre for Advanced Spatial Analysis (CASA), University College London
The talk is available at: http://www.nesc.ac.uk/action/esi/download.cfm?index=4444.

Traditionally, scientists dealing with social and built environments have worked with comparatively small data sets, largely because our ability to collect data at the individual level has been confined to censuses that have been expensive to mount and whose data is personal in nature, relevant to the public task as set by government. Such data have been generally available at the level of its collection for research; in short, data sets based on individual data have only been available at aggregate level. This situation is changing for many reasons. More automated techniques of collection are in the vanguard of this change, but many more agencies in the public and private sectors are engaged in collection, as much for marketing purposes as for the public task. The web has clearly established a medium through which such data might be collected, while the advent of personal devices with locational awareness are generating very different kinds of data sets of a social nature which have different implications for privacy and confidentiality from those traditionally collected by government.

We are in fact poised at a threshold when massive data sets are likely to become available akin to those which we take for granted in the physical and biological sciences. We need dramatically improved tools for dealing with these, through statistical analysis, visualisation, data storage and access and so on. The talk presented some of the key challenges in this area with respect to two developments in geospatial data: the emergence of visualisation techniques for online mapping and multimedia for analytical and dissemination purposes, and second the importance of using contemporary ICT to enable the very collection of social data using techniques ranging from data mining of social networks to tracking mobile devices and crowd sourcing. The MapTube project (www.maptube.org), which is part of NCeSS and NeISS, illustrates the need for such techniques to truly open up data to wider constituencies of scientists. These have to fit in the evolving context of the geospatial web and initiatives in the UK such as the Making Public Data Public. These developments are providing new and rich data resources for social science and the built environment.

[1] The metadata categories (under "Name") are taken from [71]

2.9 Providing an environment where every data-driven researcher will thrive

Carole Goble, University of Manchester
The talk is available at: http://www.nesc.ac.uk/action/esi/download.cfm?index=4467.

I haven't found an abstract for this, and I think it might move to Dave De Roure's Chapter. mpa

2.10 Summary and Conclusions

The main outcome of the day's talks is that there is no "one-size fits all" approach to dealing with data volume, and, as expected, the architecture of a data-intensive system depends largely on the application domain.

The first possibility is to build massively distributed systems where each node is responsible for collecting and managing its own data; then, a well-defined format dictates the exchange of information between the various system's nodes so data can be shared and computation over larger data volumes can take place in a collaborative fashion. The per-node storage infrastructure is not dictated in any way, provided that data exchange is possible.

A potentially radically different approach is to have a centrally located data store. The term centrally might be misleading in this instance: it may well be distributed for performance and fault-tolerance reasons, but it presents a single unified view to the user. The problem then is one of identifying the interesting bits in the data. If it is a question of knowing the properties of the data one is interested in, there is standard (relational) database technology that can be leveraged and the data can be retrieved and through SQL. However, it may well be the case that the user is not looking for something specific, or cannot come up with a declarative specification of what s/he is looking for. In those case, there need to be mechanisms in place for a "higher-order" processing, perhaps through visualisation and simultaneous processing of multiple data sources.

The situation changes again if we are not interested in off-line analyses, but on-line ones. In those cases the definition of system response time radically changes: it is not how fast the data can be processed in some complex computation; rather, it is how fast the data is readily available to be processed. Moreover, it is not only the data itself that is important; the computation over it, or rather, the workflow of computation on it is equally —if not more— importance. Therefore, provenance of computation as well as data becomes a first-class citizen.

A different perspective is to accept the volume of data as intractable for certain types of computation, and then focus on processing summarisations of the data in order to identify interesting bits of information or test certain hypotheses. Then, it becomes an issue of throwing sufficient computing power to the problem as a small sample of the data can provide adequate approximations of certain patterns. Naturally, the larger the sample and the more the computing power, the better the approximation; so, scaling these methods to larger computing clusters and larger volumes of data can only yield better results.

As is evident from the previous discussion, the problem of dealing with large volumes of data does not have a simple solution. Various parameters apply: whether the data store is to be central or distributed; whether the type of processing is to be done remotely or in a cluster; whether computation takes place over off-line or on-line data; whether one processes the actual data or samples of data. Depending on the application at hand there are readily available approaches from multiple domains that might be the best options.

Chapter 3

Data-Complexity's Challenge

3.1 Preamble

It's not just the scale and volume of data that characterises data-intensive research, but also the complexity within and across datasets. Many of today's research challenges involve deep analysis of individual datasets and also analysis across multiple data sources. The nature of data collection — and perhaps the nature of how research is organised and conducted — involves increasing specialism in particular areas, which risks creating data silos. Hence our themes today could be described as "data-busting" and "silo busting".

One of the topics that arises today is the government initiatives to open up public data, and we will see examples of data being used for 'digital social research' and the mechanisms for linkage across space and time. The technological solutions to linkage — such as linked data, databases, workflows or mashups, was a particular focus in the breakouts as we aim to cut across technology silos too.

3.2 Big Data Bioinformatics

Mario Caccamo, The Genome Analysis Centre, Norwich, UK
The talk is available at: http://www.nesc.ac.uk/action/esi/download.cfm?index=4457.

Obtaining the sequence of the three-billion letters human genome [73, 11, 74] was hailed as one of the biggest enterprises in the history of science. And rightly so, it involved a remarkable organisation of scientists, policy-makers and funding agencies from across the globe. This feat has captured the imagination of many but beyond the scientific achievement it marked the beginning of a new era in molecular biology characterised by a revolution in data generation. As in other areas of science technological advances combined with the availability of high-performance computers have made possible to collect, process and analyse biological data at a rate that have transformed the life sciences. The advent of the next generation sequencing technologies five years ago has completely transformed the landscape of genomics. These technologies introduced massive parallel platforms that supported by advanced software systems generate in the order of billions of nucleotides (i.e. letters) per run of the instruments. Today, any laboratory equipped with the latest sequencing technologies can produce in few days as much sequence data as the Human Genome Project did and at a fraction of the cost.

More excitingly these new technologies have opened up the possibilities for novel applications that traditionally were outside the scope of sequencing. One example is the study of environmental samples such as microbiomes and soil; or the use of deep sequencing for the analysis of transcriptome. As sequencing becomes cheaper and more accurate we will soon be able to explore genetic information in real-time and at a single-cell level. The so-called third generation technologies will implement single-molecule sequencing directly reducing the operation time and increasing the accuracy of the instruments output. This unprecedented wealth of data, however, has come with new challenges. There is a growing gap between the capacity to generate genomic sequences and the ability to process and interpret the resulting data in what John McPherson has recently described as the "Next-Generation Gap" [75]. The sheer volume of genomic data requires a new level of software sophistication both to cope with the load, and to analyse it effectively.

Another important legacy from the Human Genome Project is the recognition that sharing and making the data available prior to publication can speed science and promote collaborations (Toronto Statement [23]). The time and resources required to distribute and store the volume of the genomic data generated by the current technologies, however, could be a barrier to the implementation of these principles. Novel compute hardware and software architectures are emerging to cope with the demands of big data. One of the leading concepts is that we should bring the analyses closer to the data and focus on the development of hardware architectures designed to support peta-scale data-intensive computations that are accessible to the scientific community. The prospects of a sustained increase in sequencing capacity combined with a well-established model for data sharing places genomics at the forefront of the data-driven science revolution.

3.3 The Open Microscopy Environment: Informatics and Quantitative Analysis for Biological Microscopy, HCAs, and Image Data Repositories

Jason Swedlow, University of Dundee, Wellcome Trust Centre for Gene Regulation and Expression
The talk is available at: http://www.nesc.ac.uk/action/esi/download.cfm?index=4455.

The Open Microscopy Environment (OME) is a multi-site collaborative effort among academic laboratories and a number of commercial entities that produces open tools to support data visualisation, management, and analysis for biological light microscopy and high content screening (HCS). Designed to interact with existing commercial software, all OME formats and software are free, and all OME source code is available under GNU public 'copyleft' licenses. With a strong foundation for biological light microscopy in place, OME has begun extending its coverage to other fields of biological imaging. We have successfully developed a data model for HCS data, and are currently releasing preview versions of software for HCS data management and analysis, with full releases scheduled for June 2010. In collaboration with Rockefeller University Press, the American society for Cell Biology, and the EBI, we have begun developing customised versions of OME technology to deliver scientific image repositories for distribution of multi-dimensional image data to the worldwide community.

OME develops and releases three different components:

1. The OME Data Model (http://ome-xml.org) provides a specification for saving metadata and exchanging metadata in microscopy and HCAs.
2. The OME-TIFF file format (http://ome-xml.org/wiki/OmeTiff) and the Bio-Formats file format library (http://openmicroscopy.org/site/products/bio-formats) provide an easy-to-use set of tools for converting data from proprietary file formats. These resources enable access to data by different processing and visualisation applications, and sharing of data between scientific collaborators. Extensive support for 3 major HCS formats has been added (InCell 1000, Opera, & MIAS).
3. The OMERO platform (http://openmicroscopy.org/site/products/omero) is a Java-based server and client application suite that combines an image metadata database, a binary image data repository and high performance visualisation and analysis. The current release of OMERO (Beta4.1) includes interfaces for Java, C/C++ and Python to support a wide variety of client applications and support for Matlab-based applications like Cellprofiler. For computational analysis of microscopy or HCS images, this standardised interface provides a single mechanism for accessing image data of all types – regardless of the original file format. Most recently, a new facility, OMEROTables, supports the management and analysis of tabular and region-of-interest data, using an HDF5-based file store. OMERO is used in the Columbus™ data management system from PerkinElmer, Inc., the softWoRx® DMS system from Applied Precision, Inc. and is the engine that runs the JCB DataViewer (http://jcb-dataviewer.rupress.org), the first publication system for original image data in the life sciences. More information is available at http://openmicroscopy.org.

3.4 Handling social science data: Challenges and responses

Paul Lambert, Stirling University
The talk is available at: http://www.nesc.ac.uk/action/esi/download.cfm?index=4454.

Social science data comes in many different varieties, but researchers across domains share similar challenges associated with organising, coordinating and adequately exploiting their data. These challenges do not ordinarily concern the sheer scale of the data, but rather involve navigating between, and adequately documenting, alternative permutations of related data and its analysis. Focusing on the example of complex social survey datasets, this talk will highlight well known challenges in working with social science data, and compare responses to those challenges which involve varying degrees of technological input. *This paragraph needs some attention* sj

3.5 The complexity dimension in data analysis

Chris Williams, School of Informatics, University of Edinburgh
The talk is available at: http://www.nesc.ac.uk/action/esi/download.cfm?index=4471.

This talk is concerned with complex models for the analysis of data, based on probabilistic graphical models. We consider two examples concerning (i) the reconstruction of gene regulatory networks, and (ii) condition monitoring of premature babies.

The first example is based on the paper entitled Reconstructing Gene Regulatory Networks with Bayesian Networks by Combining Expression Data with Multiple Sources of Prior Knowledge, by Adriano V. Werhli, Dirk Husmeier, *Statistical Applications in Genetics and Molecular Biology* 6(1), 2007. Here the goal is to learn the underlying structure of a regulatory network based on data from flow cytometry experiments (Sachs et al, *Science* 2005), and prior knowledge extracted from the KEGG pathways database. This can be expressed as wishing to sample from $p(G|D, B) \propto p(D|G)p(G|B)$ where G is the graph structure of the regulatory network, D is the data, and B is the background knowledge. Experiments by Werhli and Husmeier show that the incorporation of prior knowledge for the reconstruction of the Raf signalling network leads to improved performance compared to using either just the data term or prior expectations separately.

The second example concerns the application of a factorial switching linear dynamical model to monitoring the condition of a premature baby receiving intensive care (Quinn, Williams and McIntosh, IEEE PAMI 31(9) pp 1537-1551, 2009). This serves as an example for the wider use of graphical models for condition monitoring tasks.

The setup is that data drawn from an observed system is often usefully described by a number of hidden (or latent) factors. Given a sequence of observations, the task is to infer which latent factors are active at each time frame. For a premature baby receiving intensive care the observations are of heart rate, blood pressure and temperature etc, and the latent factors are the state of health of the baby, along with factors that describe particular patterns of artifact (e.g. taking a blood sample, probe recalibrations etc).

We demonstrate how to exploit knowledge of the structure of how the various latent factors interact so as to reduce the amount of training data needed for the system. A combination of domain knowledge engineering and learning is used to produce an effective solution. We use the model to infer the presence of two different types of factors: common, recognisable regimes (e.g. certain artifacts or common physiological phenomena), and novel patterns which are clinically significant but have unknown cause. Experimental results show the developed methods to be effective on real intensive care unit monitoring data.

3.6 Spatial microsimulation for city modelling, social forecasting and urban policy analysis

Mark Birkin, School of Geography, University of Leeds
The talk is available at: http://www.nesc.ac.uk/action/esi/download.cfm?index=4458.

The Moses project [76] is an exercise in social simulation which aims to produce data-intensive models of urban growth and dynamics, with special reference to social and demographic variations in space and time. The models are of interest to a wide constituency of 'users', from academics to policy-makers. In addition to baseline dynamic projections [77], the impact of scenarios in relation to planning options such as the provision of health-care facilities, new transportation infrastructure, or controls over housing and land use may all be explored through this technology.

Related developments in simulation modelling are emerging rapidly across the social sciences, many of them with an explicit focus on either microsimulation [78] or agent-based modelling [79, 80]. Examples are widespread in economics, health sciences, transportation research, sociology and geography. For instance, simulation of the diffusion of epidemics through a population is emerging as a litmus challenge to this community, which is intensive of both data and computational resources [81, 82, 83].

Researchers in social simulation are trying to mash up data across a tremendous wide variety of domains, providers, and scales. Thus Moses itself combines spatial data (Google maps, Ordnance Survey, OpenStreetMap), government local area statistics (Census Small Area Data, Special Migration Statistics, Special Workplace Statistics, ONS Vital Statistics), syndicated microdata (Census Anonymised Records – SAR; British Household Panel Survey; General Household Survey), vertical data (Hospital Episode Statistics, Health Survey for England, National Travel Survey), and specialised end-user data (e.g. East and South-East Leeds Housing Needs Study). This list is indicative and by no means exhaustive. Social scientists have a particular current interest in the exploitation of both commercial datasets and crowd-sourced data [84] which threatens to overwhelm existing capabilities for analysis, processing and storage.

The Digital Social Research community in the UK is responding to these challenges with an effort to begin construction of a National e-Infrastructure for Social Simulation [85]. NeISS aims to provide a framework for the seamless combination of multiple data sources together with appropriate mechanisms for analysis and visualisation, within a workflow-enabled portal architecture. The project combines computer scientists with social scientists from a variety of disciplinary backgrounds.

3.7 Linked Data: Making things more accessible

Hugh Glaser, Seme4 Ltd
The talk is available at: http://www.nesc.ac.uk/action/esi/download.cfm?index=4451.

In recent times the requirement to publish data and link it with other resources has become more acute. There are societal reasons for this, such as Freedom of Information and generating confidence in the accuracy of scientific predictions. There are also very practical reasons, as many scientific advances depend on being able to combine the data from multiple datasets to gain insights that could not be achieved by individual research groups.

The Linked Data Initiative (http://linkeddata.org/) addresses some of these issues, as it aims to provide a framework for linking datasets and other resources; this presentation looked at how researchers can begin to join the Linked Data world.

It began by considering the principles for Linked Data:

1. Use URIs as names for things.
2. Use HTTP URIs so that people can look up those names.
3. When someone looks up a URI, provide useful information, using the standards (RDF, SPARQL).
4. Include links to other URIs. so that they can discover more things.

Following this, the issue of co-reference was discussed, and the http:/sameas.org/ site that provides much of the glue was presented. Finally the RKBExplorer and other applications that have been built on top of the Linked Data were demonstrated.

Some references:
Hypertext Style: Cool URIs don't change – http://www.w3.org/Provider/Style/URI

Cool URIs for the Semantic Web – http://www.w3.org/TR/cooluris/
Linked Data – Design Issues – http://www.w3.org/DesignIssues/LinkedData.html
How to publish Linked Data on the Web – http://www4.wiwiss.fu-berlin.de/bizer/pub/LinkedDataTutorial/
Linked Data – http://linkeddata.org/
This presentation – http://eprints.ecs.soton.ac.uk/18691/

3.8 Using search for engineering diagnostics and prognostics

Jim Austin, Department of Computing Science, University of York
The talk is available at: http://www.nesc.ac.uk/action/esi/download.cfm?index=4456.

This work has brought out a number of challenges for the development and use of system that deal with large complex data. For some time Austin, his group and their collaborators have been developing scalable systems that tackle large data problems where the data is complex (difficult to analyse), large (100's Gb) and incomplete (data missing, added or changed). Examples of this are given in the DAME (http://www.cs.york.ac.uk/DAME) in the engineering domain and CARMEN (http://www.carmen.org.uk) in the scientific domain. The evidence from these projects and others is that data in both science and industry is growing, and there is a desperate need to develop methods to handle it. A major challenge is now the design of systems. We have a rapidly growing base of methods (note, methods do remain a challenge, in particular the development of methods that can provably scale to large data), but systems that tie the methods together are complex, difficult and thus expensive to build. We are now in a position to understand some of the principles that need to be followed. For example, it is vital to understand what will be done to the data before the systems and data repositories are built. If this is not done then the systems will be inefficient at best or just unusable at worst. We are at risk of copying the problems in industry where large complex IT systems often do not work. It is vital to make the systems flexible, as users always change their minds. Methods to develop large data systems must address these and other issues.

Many of the systems and technologies that Austin has developed have been commercialised. For example the Signal Data Explorer system (http://www.cybula.com) from the DAME project. The lessons just as well apply to the transfer of methods to everyday widespread academic use. A central issue is sustainability. There are many things that can be done in projects to help this. For example the design of systems that can be maintained, a challenge in academic research as itUs more about development than research. Another issue is understanding that systems support users and donUt replace them. Approaching systems that can help a user will be far more successful that one that replaces them! To ensure good transfer, working closely with users from the start with an expectation of widespread use is vital. This was exemplified by the UK eScience programme, but should be encouraged in new projects.

3.9 Summary and Conclusions

To be written by Dave De Roure

mpa

Chapter 4

Data-Interaction's Challenge

4.1 Overview

Should we say Data Interaction's Challenge or the Challenge of Data Interaction sj

Volume and even complexity are tangible concepts in that we can envisage measuring how much data needs to be processed, how complex the analysis patterns are or how complex the links between several data sets are. Interaction with data and analysis are much harder to pin down; analysing the sequence of interactions with a system will not necessarily reveal the intricacies of how humans and other systems interact with data and analysis. Nor will it suggest what technology is required to effectively and efficiently operate within a specific data-intensive environment.

The theme of this day is to identify the challenges that arise when solving problems that are data-intensive taking into account interaction. Interaction includes the ways in which researchers interact with running analyses and visualisations, and the longer-term succession of operations that researchers perform to progressively find, understand and use data. Many problems will involve geographically distributed teams. Collaboration across institutions and disciplines is becoming the norm, to gather skills and knowledge and to access sufficient information for statistical significance. This leads to privacy and ownership issues, especially where sensitive personal data is involved.

Challenges are not restricted to interaction with large data or even complex data. It is important to understand users have different skill sets and objectives, which lead to different patterns of interaction. To ensure an inclusive community, users should be able to access technology at different levels of intuitiveness. This calls for "learning ramps" to help everyone to progress as far as they wish towards expert usage modes.

Data-intensive environments are often highly dynamic. It is important to understand how in this context researchers build up their portfolio of data and analysis providers, and how they react as new data and tools become available. The social behaviour and networks involved in making and influencing choices will have their impact on what data and which tools become an established standard.

The following talks were given.

- Research communities are competitive, Joel Saltz (Emory University) talked about how enable distributed data access and tool provision in one such community—see §4.2.
- In data-intensive environment, how do we publish data? Bill Michener (University of New Mexico) talked about how to facilitate the full cycle of data acquisition in such environments—see §4.4.
- Peter Buneman (University of Edinburgh) provided an alternative viewpoint by drawing on his theoretical knowledge of databases to suggest strategies for thinking about and understanding data—see §4.5.
- Much data and knowledge is represented by unstructured text, Andrew McCallum (University of Massachusetts, Amherst) showed how machine-learning methods can provide us with actionable knowledge from such data—see §4.6.
- Sensor networks can include monitoring of human behaviour, Paul Watson (Newcastle University) introduced several projects where data of this kind is gathered and analysed – see §4.7.

The following topics motivated breakout sessions during the afternoon.

- Interacting via visualisations of large and complex data, e.g. large-image viewers, reducing complexity to find relevant patterns and structuring of data.
- Interacting via collaborative systems, e.g. scientific gateways, data and tool sharing, data repositories, ontologies, trust and peer-reviewing.
- Interacting with analysis: workflow systems, learning ramps, portals and service-based provisions.

4.1.1 Turing Award Lecture

Profesor Chris Bishop, Deputy Director, Microsoft Research, Cambridge
The talk is available at: http://tv.theiet.org/technology/infopro/turing-2010.cfm.

Professor Chris Bishop gave the Turing Lecture entitled "*Embracing Uncertainty: The new machine intelligence*". This prestigious lecture complemented our workshop and took place in a venue which was a short walk from the e-Science Institute.

> Abstract: Computers are traditionally viewed as logical machines that follow precise, deterministic instructions. The real world in which they operate, however, is full of complexity, ambiguity, and uncertainty. In this year's Turing Lecture, Professor Chris Bishop discusses the field of machine learning, and shows how uncertainty can be modelled and quantified using probabilities. Professor Bishop will look at the recent developments in probabilistic modelling that have greatly expanded the variety and scale of machine learning applications, and he explores the future potential for this technology.

4.2 Integrative Analysis of Pathology, Radiology and High Throughput Molecular Data

Joel Saltz, Center for Comprehensive Informatics, Emory University
The talk is available at: http://www.nesc.ac.uk/action/esi/download.cfm?index=4459.

Integrative analysis of multiple complementary types of information is playing an increasingly essential role in biomedical research. The Emory led caBIG^{TM} In Silico Brain Tumor Research Center is an excellent example of this. The effort is designed to leverage complementary molecular, pathology, radiology and outcome data, obtained by several large scale National Cancer Institute studies including the Cancer Genome Atlas (TCGA) study. This group is developing workflows consisting of novel image analysis algorithms and bioinformatics analyses to correlate imaging characteristics defined by Pathology and Radiology derived feature sets with underlying "omic" characteristics and with patient outcome. This effort is explicitly designed both to motivate advances in informatics and to generate useful brain tumour research results.

A major technical focus of the In Silico Brain Tumor research Center is high throughput digital microscopy, which combines image acquisition with generation of semantically annotated feature sets that describe tissue characteristics of disease at a cellular scale. This is a data intensive process; a high-power microscope can easily generate several terabytes of data each day. Tissue characterisation involves workflows that employ a sequence of steps: image segmentation, feature construction, feature selection, feature extraction, classification of features, and annotation of images and image regions. The sheer size of image datasets makes gleaning information from digital microscopy slides a data-intensive process requiring efficient system support. The large number of potentially useful micro-anatomic features and the large number of image preprocessing algorithms makes it essential to develop effective semantically oriented mechanisms to manage, query and curate features and to track provenance associated with the generation of each collection of features.

More information about the work is in [86, 87, 88, 89].

4.3 Text mining, computational biology and human disease

Andrey Rzhetsky, Department of Medicine, Department of Human Genetics, Computation Institute, Institute for Genomics and Systems Biology, University of Chicago
The talk is available at: http://www.nesc.ac.uk/action/esi/download.cfm?index=4472.

Imagine that a graduate student enters the Library of Congress with an assignment: she is to find and pull all texts relevant to protein glycosylation. Her problem is straightforward, known among text-miners as information retrieval (IR). If the student must not only find the books, but also to flag the most important concepts she encounters in each, she must perform named entity recognition (NER). Undaunted by her workload, imagine she decides to identify relations between concepts, such as protein BAD binds protein BAX, text-mining researchers call this information extraction (IE). Along with a group of additional tasks, such as question answering and text summarisation (QA and TS, respectively) which we outline below, problems of computational IR, NER and IE belong to a larger field of natural language processing (NLP)—human language understanding and generation by a computer. NLP is an integral part of artificial intelligence (AI), with its larger goal of recreating or surpassing the computational ability of the human brain. The graduate student in our example has a distinct advantage over the computer, of course—she is naturally gifted, as we all are, with the ability to parse and quickly extract meaning from language. Text mining, a new field that applies computational techniques to the problems faced by our graduate student above, can be thought of as a sub-field of both AI and NLP—but it also stakes a claim for "intellectual independence" with its emphasis on gaining new knowledge. Thus, while multiple definitions exist, text mining is typically associated with information retrieval, extraction, and synthesis, with a special stress on gaining new knowledge.

Moving beyond information retrieval and extraction, some proportion of published assertions can be repackaged to form synthetic ideas, that is, new compound concepts that are significantly more valuable to the scientific community than the sum of their original assertions. In 1986, researcher Don R. Swanson read a collection of scientific articles that spanned different fields of inquiry. Each of these disconnected communities had information valuable to the other, but due to the lack of crosstalk, did not appear to know it. Swanson noticed, in papers produced by researchers interested in Raynaud's disease, that its symptoms often included increased blood viscosity and rigidity of erythrocytes. Meanwhile, in the nutrition biology community several publications proposed that dietary fish oil could reduce blood viscosity, without explicitly connecting this finding to Raynaud's disease. In this early real-world example of synthetic idea generation, Swanson merged these findings to suggest that fish oil can be beneficial for Raynaud's patients and published his suggestion as a medical hypothesis. The hypothesis was confirmed some years later in an independent randomised clinical trial.

Even with current text mining capabilities, such synthetic ideas can be discovered automatically. A more distant but nonetheless realistic aim of the field is to trace and map more sophisticated ideas (idea isomorphisms) that are expressed differently in different scientific fields, yet represent identical problems or their solutions. Were such idea mappings made instantly available through an Internet interface, the result could be truly impressive. The diffusion of innovations across science could be markedly increased by making solutions developed in one area visible to specialists still desperately searching for them in a different field. Computationally pairing problems and solutions generated by different fields is a type of automated creativity (systematic search for synthetic ideas) that computers almost certainly will do for us in the not too distant future. Further information may be found in [90, 91, 92, 93, 94, 95]

4.4 Building a virtual data centre for the biological, ecological and environmental sciences

Bill Michener, University of New Mexico, Albuquerque
The talk is available at: http://www.nesc.ac.uk/action/esi/download.cfm?index=4466.

Addressing the Earth's environmental problems requires that we change the ways that we do science; harness the enormity of existing data; develop new methods to combine, analyse, and visualise diverse data resources; create new, long-lasting cyberinfrastructure; and re-envision many of our longstanding institutions. DataONE (Observation Network for Earth) represents a new virtual organisation whose goal is to enable new science and knowledge creation through universal access to data about life on earth and the environment that sustains it. DataONE is designed to be the foundation for new innovative environmental science through a distributed framework and sustainable cyberinfrastructure that meets the needs of science and society for open, persistent, robust, and secure access to well-described and easily discovered Earth observational data.

Supported by the U.S. National Science Foundation, DataONE will ensure the preservation and access to multi-scale, multi-discipline, and multi-national science data. DataONE is transdisciplinary, making biological data available from the genome to the ecosystem; making environmental data available from atmospheric, ecological, hydrological, and oceanographic sources; providing secure and long-term preservation and access; and engaging scientists, land-managers, policy makers, students, educators, and the public through logical access and intuitive visualisations. Most importantly, DataONE will serve a broader range of science domains both directly and through the interoperability with the DataONE distributed network. DataONE is a five-year project that began in 2009.

I identify key environmental scientific, cyberinfrastructure, and sociocultural challenges and provides a road map for how DataONE is addressing these challenges. Specific examples are based on a DataONE EVA (Exploration, Visualisation and Analysis) Working Group effort to create an integrated database that can be used for better understanding the ecology of bird migrations.

The important message is that to succeed in international preservation and access to data we need to recognise the importance of data from the perspective of the scientists to understand the requirements. Scientists want fast access in terms of data and tool discovery, automated/automatic metadata annotation and rapid analysis via appropriate visualisation. They want easy access in terms of data and metadata upload; integrated, linked, and synthesised databases and interoperable and intuitive scientific workflow systems. They want cheap solutions for data preservation that are free and open source, but also demand 'support' for tools they already routinely use, such as Excel.

The "elephant in the room" challenge is handling metadata as this needs tools and approaches that are fast, easy and cheap at the same time. Further information may be found in [96].

4.5 Curated databases

Peter Buneman, School of Informatics, University of Edinburgh
The talk is available at: http://www.nesc.ac.uk/action/esi/download.cfm?index=4464.

Curated databases are databases—or knowledge bases if you want to make a distinction—that are populated and updated with a great deal of human effort. Most reference works that one traditionally found on the reference shelves of libraries—dictionaries, encyclopaedias, gazetteers, etc.—are now curated databases. Since it is now easy to publish databases on the web, there has been an explosion in the number curated databases designed to support scientific research. The value of these databases lies in the organisation and the quality of the data they contain. Like the paper reference works they have replaced, they usually represent the efforts of a dedicated group of people to produce a definitive description of some subject area.

Curated databases present a number of challenges for database research. The topics of annotation, provenance, and citation are central, because curated databases are heavily cross-referenced with, and include data from, other databases, and much of the work of a curator is annotating existing data. Evolution of structure is important because these databases often evolve from semistructured representations, and because they have to accommodate new scientific discoveries. Much of the work in these areas is in its infancy, but it is beginning to have practical consequences and to suggest new areas of research for both theory and practice.

Many of the issues surrounding curated database are non technical. We need a good economic model for sustainability of such database. Open Access is a recent and successful model for journal papers. Is this model appropriate for curated databases that require long-term support? Moreover, people who write reference manuals sometimes expect to make money out of them. Furthermore, intellectual property in curated databases is a nightmare with legislation still largely based on the notion of copying. We need to bring curated databases into the scope of libraries and other archival institutions and organisations to identify and bring into practice sustainable models for curated databases.

4.6 Discovering Patterns in Text and Relational Data with Bayesian Latent-Variable Models

Andrew McCallum, University of Massachusetts
The talk is available at: http://www.nesc.ac.uk/action/esi/download.cfm?index=4480.

One way to make sense of large collections of data is to find a digestible number of latent components that summarise the original data. This talk presented recent advances in mixed-membership Bayesian networks (topic models) that provide presentable, interpretable views of various data, including text (in multiple languages), relations (of several types), time stamps and other attributes. It emphasises work in social network analysis. For example, the Author-Recipient-Topic model discovers role-similarity between entities by examining not only network connectivity, but also the words communicated on those edges; This method is demonstrated on a large corpus of email data subpoenaed as part of the Enron investigation. The "Group-Topic" model discovers groups of entities and the topical conditions under which different groupings arise; this is demonstrated on coalition discovery from many years worth of voting records in the U.S. Senate and the U.N. Further examples show Bayesian networks being successfully applied to various large text collections and relational data, leading to a discussion of their applicability to trend analysis, expert-finding and bibliometrics.

The challenge is to scale up the amount of beneficial interaction between humans. By mining data sets that capture human interaction, we can identify potentially beneficial relationships between two or more people. This first step to match making is important in areas where we need to form groups of mixed expertise or bring together people with the same expertise.

This is joint work with colleagues at UMass and Google: Xuerui Wang, Natasha Mohanty, Andres Corrada, Wei Li, David Mimno, Gideon Mann and Hanna Wallach.

4.7 Using Real-Time data to Understand and support Human Behaviour

Paul Watson, University of Newcastle
The talk is available at: http://www.nesc.ac.uk/action/esi/download.cfm?index=4465.

There has been a huge increase in the availability of real-time data. Examples of data sources include sensors, and also, "Software as a Service" applications which can capture every detail of users on-line behaviour. A key challenge is to extract value from this data by understanding the real-world behaviour it represents. The grand challenge is then to use this information to influence behaviour for the good.

This talk will examine issues in extracting value from this data, using examples from current projects within the "Social Inclusion through the Digital Economy" (SIDE) project: intelligent transport systems for older drivers, ambient kitchens for people suffering from dementia, as well as software as a service for scientists. A general architecture for processing this data will be described. We will argue that the key challenge is to find ways to use this growing body of real-time data to influence human behaviour.

4.8 Summary and Conclusions

From the talks and the breakouts it is evident, interaction can take several forms. Joel Saltz focusses on systems interacting with systems. These include computers communicating with other computers, methods and model integrating with other methods and models and data linking with other data. Bill Michener and Peter Buneman focussed more on humans interaction with systems. Buneman highlights the economic challenges around sustaining databases and Michener emphasises the importance of handling metadata as humans are inherently bad at providing detailed metadata when generating new data. Andrew McCallum's main contribution revolves around scaling human-human interaction by extracting better relationships between groups of people and using these relationships to form new ones or improve existing ones. Joel Saltz also touches on human-human interaction by explaining the difficulty of designing ontologies by committee in order integrate data. Paul Watson cuts across all these types of interaction where his main quest is to make good use of the data obtained by monitoring interactions, where these interactions can be between humans, humans interaction or systems communicating.

Irrespective of what form of interaction, one important aspect is about scaling up the interaction. In the context of above, how do we scale up useful human relationships in the light of more data about their interactions? How do we scale up the creation and use of metadata in the light of more and more different data? How do we scale up the usage of new and existing scientific methods? A vital instrument to achieve scaling will be in the form of "intellectual learning ramps". These ramps must be provided such that they facilitate people with different levels of experience to start using new methods in the form of tools running on ICT infrastructure. They must also allow anyone to increase their competence of using the methods as far as need to. As highlighted in the breakout groups, the key then is to deliver these ramps while ensuring the methods they facilitate are used correctly.

Another vital aspect of building intellectual learning ramps is the use a model of development that itself scales. It is unlikely we can afford or even train an army of software developers skilled in the provision of intellectual learning ramps for data-intensive research. Moreover, if we want to reach the researchers in the "long tail" of data-intensive research—the large set of researchers that do not have the largest data sets, but still cannot deal with their data—we need to provide methods and tools in the form of intellectual learning ramps that representative members of a community can utilise to build learning ramps appropriate for their own communities.

An example of how this can work is described in [97, 98]. Here, a intellectual learning ramp is developed to learn the next generation of chemists about computational chemistry. The ramp consists of a web portal to shield students from technical details and from the complexity of the chemistry tools. The web portal was developed by chemists, not by people from ICT, e-Science or computer science. This was made possible because they themselves used another intellectual learning ramp, Rapid [99, 100], that allows domain specialists to develop web portals without having to learn all the technology involved (Grid/Cloud/Cluster/HPC computing, web portal frameworks, security models, etc.). This model allows scaling the number of portals to be developed and the number of tools exposed through these portals even further.

Chapter 5

Themes Emerging in Data-Intensive Research

5.1 Overview

The overall goals of the workshop are:

1. To increase our understanding of what data-intensive research is already underway.
2. To develop a vision of the potential of data-intensive research in the next ten years.
3. To identify the challenges to achieving that data-intensive research potential.
4. To plan a strategy for achieving that vision by addressing those challenges.

The first seven talks summarised the primary outcomes of each day and each cross-cutting theme.

John Wood, of Imperial College, London, will give his vision of the future of data-intensive research. This has been developed through his long experience of scientific data, including the data challenges at the Rutherford-Appleton and Daresbury Laboratories when he was Chief Executive Council for the Central Laboratory of the Research Councils and his time as chair of the European Strategy Forum for Research Infrastructure. He is currently chair a body to advise on data for the European Union's Framework Programme 8. See §5.3.

5.2 Multifaceted views of DIR

As each day and theme organiser drew together the whole week's discussions, including the many energetic breakout sessions, many views of data-intensive research were emerging. It remains a challenge as to whether these were the result of different intellectual viewpoints observing *a common DIR scene* or views of different clusters of DIR activity in a larger space. More work is needed, studying DIR applications in detail, to characterise more precisely the applications, the methods and the challenges.

5.2.1 Highlights from Monday's programme

Malcolm Atkinson, School of Informatics, University of Edinburgh
The talk is available at: http://www.nesc.ac.uk/action/esi/download.cfm?index=4469.

Three talks on Monday delivered impetus to the week's agenda, demonstrating the power of data-intensive methods and challenging the community to do more with the growing wealth of data.

Alex Szalay (see §1.3) was the first to explain the origins of data growth, reminding us that this was growth in numbers of data collections as well as growth in size of collections, and hence presenting the challenge of reaching the 'long tail' of researchers who could now benefit from data-intensive methods and tools. Alex recounted his work with Jim Gray in their epic struggle to publish the Sloan Digital Sky Survey so that astronomers around the world could use it—more than 930,000 do! Their propagation of the methods to The National Virtual Observatory (http://www.us-vo.org/), to PanSTARRS (http://pan-starrs.ifa.hawaii.edu), to Immersive Turbulence simulation data (http://turbulence.pha.jhu.edu) and to Sensor Networks (http://lifeunderyourfeet.org most data comes from digital sensors – see §4.7 & §3.8); emphatically demonstrates that once data-intensive methods are mastered they quickly benefit many lines of research—a strong indicator of underlying principles. Alex went on to describe the new hardware architectures that they had built, first GrayWulf to balance IO performance with CPU capacity and then Amdahl Blades to reduce energy use. He emphasised the importance of software as the *new instrument*, but we also saw the impact of sustained intellectual investment in the data-intensive campaign.

Douglas Kell introduced us to the new biology, where data-driven hypothesis formation takes an increasingly significant role. He demonstrated, through a series of examples, how biological and ecological insights had been derived by combinations of machine learning and text mining over assembled collections of data from multiple sources. This new *modus operandi* for biologists involved large teams brought together by the complexity of the biological systems and to assemble the relevant data-intensive skills – as evidenced by the long lists of authors on the papers – the shared data being a means as well as a reason for woking together. Here again we heard how the techniques developed in one systems biology challenge could be quickly redeployed in many new ones. The opportunities were growing rapidly and new technology, such as faster sequencing machines (see §3.2), was delivering an increasing deluge of new data. So we were left with a fourfold challenge: *a*) to help the *long tail* of biologists to capitalise on the data-intensive opportunities for their research; *b*) to devise data-intensive technology that can handle the *growing volumes* of data; *c*) to deal with *demand* as more biologists use more data-intensive methods on more data sources; and *d*) to deal with the *complexity and uncertainty* of the full spectrum of biological data so that data-intensive methods are trusted.

Thore Graepel showed us many sources of complex and large volume data available within the Enterprise (Microsoft), ranging across: user clicks, traffic on services, game-playing records, fault reports and software production. The challenge is to use this data, within the constraints of privacy laws, to improve user experiences and profitability. The rates of data gathering were prodigious and sustained, which provides a key resource and challenge for the researchers. The goal of mining latent relationships from the data was addressed by a succession of steps: formulate a Baysian model using a factor-graph notation; transform this into a form suitable for distributed algorithms and evaluate the result on the Cosmos architecture – a data-streaming map-reduce system. A language PQL is used to declaratively describe the problem and transform it to run on Cosmos. This was demonstrated with three applications – all in production! *a*) TrueSkill$^{\mathrm{TM}}$

analysed records of interactive games (10^6 matches/day, $6 * 10^6$ players) to generate a global leader board and offer players matches with people of similar skill; *b*) AdPredictor$^{\text{TM}}$ analysed user clicks in all of the Microsoft's Bing$^{\text{TM}}$ search-engine requests to place the advertisements most likely to be clicked on the results page and to set advertising prices; *c*) Matchbox$^{\text{TM}}$ which estimates similarity between users' preferences in a 20–100 dimensional space and recognises affinities which are then used to make recommendations, e.g. for content download or virtual friendships. Once again we see data-analysis methods being pioneered for one application and then being quickly transferred to new applications. We also saw the construction of 'intellectual ramps' to make it much easier to adopt and deploy the methods.

Nine groups displayed their working data-intensive tools and research progress in the Research Village (see §1.6). As intended, this stimulated many good discussions and throughout the week participants were seen trying the technologies they had discovered there.

Three of the cross-cutting themes were introduced: data analysis (see §1.7), programming paradigms (see §1.9) and data-structuring paradigms (see §1.8). The work of each of these is reviewed below.

The day concluded with a talk by Xavier Llorà and Bernie A'cs (see §1.10) on Meandre, their workflow language and on their use of cloud and map-reduce technology for its enactment. As they also had a high-level language, ZigZag, and drove their transformations from high-level semantic descriptions, this was an opening salvo in the programming paradigms theme.

5.2.2 Highlights from Tuesday's programme: Data-volume challenges

Stratis Viglas, School of Informatics, University of Edinburgh
The talk is available at: http://www.nesc.ac.uk/action/esi/download.cfm?index=4475.

An easy way to define data intensity is to define it as data volume and keep it at that: we only need more efficient ways of storing and manipulating data. But that is nothing new. Perhaps the most important aspect of the day's programme was the fact that this substituting intensity for volume does not even begin to describe the problem, let alone define it.

A pattern that has emerged throughout the day's programme is the *need* for data, rather than specific cross-cutting architectures of dealing with data volume. This need has been recurring in all talks, where, depending on the application domain and its needs, different system architectures were employed. The main decision parameters for dealing with the volume of data can be highlighted as follows:

Filtering *vs* large-scale analytics. In some cases, the type of processing we need to do is isolating parts of the input —potentially through a declarative interface. This is tailored for systems built on database technology, which are optimised to do just that: quickly identify the relevant parts of the input and manipulate them to generate the result. On the other hand, we might want to run a large-scale analysis of the input, in which case setting up the system for fast filtering does not make much sense. In such cases, one is better off employing an architecture that is tailored towards the use of algorithms like MapReduce, which uniformly process the entire data set to perform a complex analysis on all its attributes.

Exhaustive *vs* sample-based methods. This represents a potential trade-off between data intensity and computational intensity. It is a well-known fact that certain algorithms, though efficient, will not scale well on large volumes of data, hence the exhaustive exploration of the data set becomes problematic. In those cases, we might be able to trade accuracy for speed by focusing on a sample of the data that will allow us to generelise the result of processing of that sample to the entire data set.

Stand-alone *vs* collaborative processing. The need for collaboration is dictated by the sheer volume of data research (and not only) organisations need to process. In some cases, one needs to either "outsource" the computation to the owner of the data and retrieve the results, or exchange data so that the same computation can be carried out over different data sets. Both these types of processing have been identified during the day's talks. Moreover, whenever one is dealing with a feedback loop in data generation, *i.e.*, the results of processing become new data than can be processed in the future, the problem of provenance becomes important: we need to explicitly record the transformations of the data to allow for reproducability of our inputs.

5.2.3 Highlights from Wednesday's programme: Data-complexity challenges

Dave De Roure, University of Southampton
The talk is available at: http://www.nesc.ac.uk/action/esi/download.cfm?index=4476.

To be Drafted by Dave

mpa

5.2.4 Highlights from Thursday's programme: Data-interaction challenges

Jano van Hemert, School of Informatics, University of Edinburgh
The talk is available at: No-Talk-Found.

As expected, Thursday's programme covered a broad range of types of interaction, including human-human, human-system and system-system interaction. Besides the obvious theme of how to scale up these interactions in an economically viable way the following themes were strongly present throughout talks and discussions.

Small data in the long tail is hard to get. All data is growing, not just the size of huge data sets, but also the number of sets both small and large. For the average scientists, small data sets are now getting large enough to break the technology to operate on them. This is a message that adds to the much easier to identify problem of huge data sets from very large projects that are breaking world-leading ICT infrastructures. However, small data sets are everywhere and are inherently difficult to share because of many reasons including privacy issues, misaligned standards and formats, competitiveness and lack of procedures that ensure the data is captured, maintained and curated.

If we assume we are able to effectively work with data both small and large, we then have to act in socially responsible ways. This is not new as the era of data mining already brought many discussions around data, but as electronically held data is now prevalent in every corner of science and indeed society and business it is more pressing as a single data set or a combination of several data sets can lead to events unacceptable to society.

One particular aspect of working with data in a responsible way is to ensure an understanding of the methods in use. ICT solutions enable the uptake of methods by making them more accessible, but can be harmful when they are then used in ways that lead to unreliable or even incorrect results. For instance, the use of statistical software packages requires knowledge of the methods they contain to understand whether these can be correctly applied to a given data set.

Making methods available to researchers in a way meaningful to them was highlighted as a critical component of data-intensive research. Here methods cover a broad spectrum of data access, analysis, modelling and processing steps. Methods are provided in the form of software. To allow meaningful access to these software researchers agree features unnecessary for a particular task must be hidden as well as any technological barriers should be removed. However, as users make more regular use of methods they are likely to want to explore more advanced features. It is therefore necessary to provide interfaces to methods that adapt to the users' needs and abilities. These learning ramps are an important way to scaling interaction with methods and the data they operate on.

Building learning ramps is a time consuming and expensive operation. To successfully develop a learning ramp it is important that researchers walk a path together with computer scientists to ensure they understand each other's domain well enough. Also, they must first agree about expectations before starting interdisciplinary projects as it takes long to understands each others jargon. Furthermore, people in different disciplines likely require particular achievements to ensure career progression that are not directly contributing to the path. Moreover, one discipline is often the enabler, while the other will be the user. The user often wants a solution as early as possible to maintain a competitive edge.

It is therefore important that the enabler has access to cost-effective and time-efficient tools to deliver an adequate solution. It is likely such a solution evolves into a more and more elaborate solution. In other words, the enabler uses tools that themselves can be seen as learning ramps.

If we assume such paths to data-intensive research environments can be successfully delivered, we then have to make sure the results produced in these environments are captured, shared and curated in a manner that is useful to the corresponding discipline. These results will include statistics about data, annotations of images, outcomes of scientific models, etc. If we want to prevent a knowledge deluge in five years we must make sure knowledge is kept in a context that allows integration with other knowledge and collaboration with the wider community.

5.2.5 Data-analysis theme review

Chris Williams, School of Informatics, University of Edinburgh
The talk is available at: http://www.nesc.ac.uk/action/esi/download.cfm?index=4477.
See §5.4.

5.2.6 Data-structuring theme review

James Cheney, School of Informatics, University of Edinburgh
The talk is available at: http://www.nesc.ac.uk/action/esi/download.cfm?index=4479.
See also Section §5.5.

Clearly, research on computer systems (databases, distributed/grid/cloud computing) and artificial intelligence (machine learning, text mining) already has found great applications to data-intensive research and there is potential for many more. These applications contribute directly to "primary" data-intensive research. However, a great deal of computer science research, particularly in areas such as programming languages, security, or theory could be considered "secondary" data-intensive research in that it is not immediately applicable to data-intensive researchers' day-to-day problems, but may have significant benefits further down the road.

What use, if any, can ideas or techniques from these areas be to data-intensive research? The title of my talk was a homage to an essay by Donald Good, "The Foundations of Computer Security: We Need Some"[1]. That essay identified the lack of clear definitions as a culprit in the widespread confusion and disagreement about research in computer security twenty years ago. I believe many of the problems with data-intensive research display this same lack of clarity and ensuing confusion.

One might expect computer scientists to help with this, and many of us are trying. But the goals of computer science researchers often diverge from those of data-intensive researchers. It appears that data-intensive research involves many problems that computer scientists could work on, but choose not to. Why not? A few possibilities include:

- Computer scientists don't know about the problems
- Computer scientists don't understand the problems
- The problems are too domain-specific to be marketable as computer science research
- It's not in their interest – the effort and opportunity cost of working on the problem might be disproportionate to the potential rewards
- The problem may sound too similar to an already-solved problem, or may actually already be solved in principle

Data-intensive research is largely being conducted by researchers whose primary training is not in computer science or software engineering, but are now responsible for developing and using seriously complicated computer systems in their daily work. It is easy for computer scientists to dismiss these efforts as lacking technical or conceptual sophistication, but this is a pointless exercise in reinforcing discipline boundaries. However, such efforts do often suffer from the fact that project managers (and supervisory funding agencies) do not know the "folklore" of computer science and software development: the basic sensibilities and principles that help guide good developers, programmers, system designers and researchers to do work of general and lasting value.

Hamming, an early pioneer of scientific computation, gave a widely-disseminated talk on how to do excellent research[2]. One of many great suggestions he gives is the following:

> I went home one Friday after finishing a problem, and curiously enough I wasn't happy; I was depressed. I could see life being a long sequence of one problem after another after another. After quite a while of thinking I decided, "No, I should be in the mass production of a variable product. I should be concerned with all of next year's problems, not just the one in front of my face."

Another significant obstacle is the lack of clarity about just what problem, or problems, data-intensive researchers need computer scientists to solve. This is exacerbated by discipline boundaries and communication gaps, the thanklessness of building bridges to cross such gaps, and the

[1] http://www.ieee-security.org/CSFWweb/goodessay.html

[2] http://www.cs.virginia.edu/r̃obins/YouAndYourResearch.html

absence of a clear conceptual foundation for data-intensive research.

Since giving the talk, I have run across a related rule for presenting research due to Lamport, called "state the problem before describing the solution"[3]. Quoting further:

> [Stating the problem in terms of the solution] makes the comparison of two different solutions rather difficult. With the second approach, one is forced to specify the precise problem to be solved independently of the method used in the solution. This can be a surprisingly difficult and enlightening task. It has on several occasions led me to discover that a "correct" algorithm did not really accomplish what I wanted it to.

Following Hamming's and Lamport's rules requires careful thought about what problem is actually being solved, not simply getting a numerical answer, or developing a solution and showing that it has some nice theoretical or performance properties. Lack of this clarity makes it incredibly frustrating for others to work on the same problem, since they first must try to intuit what the problem really was in the first place from extant (and often heavily discipline-specific) solutions. Lack of clarity about problems also makes it very difficult to compare solutions to similar problems.

In my talk I also wanted to critique some of the discussion that took place during the workshop concerning the challenges of scale. Menaces such as "big data" or "data deluges" or "data tsunamis" are often invoked as scare tactics for obtaining funding for data-intensive research. While these tactics have proved effective (and this is generally a good thing), they have also promulgated a myth that the problem is just a matter of building ever larger systems and and smarter algorithms to do relatively boring things with large amounts of uniform data. Data-intensive research, as illustrated by the range of talks in the workshop, is not just this. It also demands doing interesting things with large numbers of small, heterogeneous data sets, as in social science and bioinformatics curation settings. It would be a serious mistake for funding agencies to support only the former kind of data-intensive research.

Having argued that principles are needed to enable more computer scientists to contribute significantly to data-intensive research, what exactly do I mean by principles? There is plenty of scope for confusion here. I don't mean simply developing more high-level guidelines or bromides such as "one person's data is another's metadata". I think "Gray's Laws" and Hamming's and Lamport's rules are examples of useful guidelines, but not necessarily principles in the sense I mean. In fact I do not have a clear idea what a principle of data-intensive research should looks like.

Rather, what I am advocating is a principled approach, by analogy with settings where such approaches have already been fruitful, such as computational complexity theory, database theory, verification, programming languages, and security. Although such theoretical work in these areas does not always have immediate real-world applications, it contributes immensely to our understanding of what is possible.

A principled approach also opens the problems in these areas to scrutiny from people not working directly on a given system, encouraging research that may have unanticipated long-term impact. For example, models of concurrent processes originally developed by Hoare, Milner and many others to understand computer systems are now finding applications in systems biology.

Finally, a principled approach can help us separate the wheat from the chaff, that is, identify which aspects of data-intensive research are really new and which are old problems in a new

[3] http://research.microsoft.com/en-us/um/people/lamport/pubs/state-the-problem.pdf

disguise. Without doing this it is easy for computer scientists not working in this area to get the impression that it is all chaff, and dismiss it.

In a panel discussion, it was asked whether existing theoretical frameworks such as computational complexity (or data complexity in database query languages) could provide such a foundation. Such theories play an important role in understanding computer systems which data-intensive research will use. But I think we also need to develop a clearer understanding of what the key research problems arising in data-intensive research really are, particularly those that involve the interaction of data, computer systems and our scientific understanding of the natural world.

Echoing the last paragraph of Good's essay, I hope I have now said something to offend nearly everyone involved in data-intensive research. But the point of this was not merely to offend but to provoke discussion of how computer science could contribute more meaningfully to data-intensive research and what steps computer scientists and "primary" data-intensive researchers can take to make this more likely.

5.2.7 Data-intensive programming theme review

Shantenu Jha, Louisiana State University
The talk is available at: http://www.nesc.ac.uk/action/esi/download.cfm?index=4478.
See §5.6.

5.3 A Vision for Research in 2030

John Wood, Imperial College, London
The talk is available at: http://www.nesc.ac.uk/action/esi/download.cfm?index=4470.

I dread the school holidays. Why? Because getting to Imperial College front entrance in South Kensington, London is a nightmare. It is on the same street of several museums including the Science Museum that is invaded by the hordes each vacation. The nearby underground station is packed with parents and grandparents herding their youngsters together. Even if I go out onto the street I have to barge through crowds of people, all apparently happy to loiter around with time to spare. What is it that is so fascinating about the history of science that attracts them? Playing with magnets, seeing laser light shows, seeing how drugs are developed and marvelling at the ingenuity of engineers who built steam engines in the 19^{th} century are all part of the experience that most British children look back on fondly even in adult years. As we have just celebrated the 400^{th} anniversary of the telescope have all the major discoveries of science been found; what will the museum of the future record of the present? Are we at the limit of what we can observe both in the space-time domain and in energy? Just what is ahead of these young people as they are wowed by the discoveries of yesteryear? What will research be like for them as they start their independent research careers in 2030 and how can those of us making decisions now affect them? This is one of the roles of the newly created European Research Area Board (ERAB). Our prime objective is to ensure that the research environment in Europe remains as fertile as in the recent past.

The concept of Europe developing a unified science policy is an ambitious goal. The rich diversity of cultures and languages combined with very different approaches to education and the financing of research should not be jettisoned in favour of dumbing everything down to the lowest common denominator. We need to build on this diversity and let the different cultures contribute together to build the future. We need to learn from history rather than just starting with a blank sheet of paper. How can we really create a gourmet dish out of the mix of disciplines and approaches – this is the challenge for us all.

As a young research fellow in Cambridge I was free to conduct my own experiments and I would often decide on what to do while walking to the laboratory. I might discuss it with colleagues but generally I had an idea that I wished to pursue. The experiments and the analysis were difficult but I had mastered them during my Ph.D. and had read most of the major papers in my field. I was being invited to give reviews and keynote talks at international meetings. Will this freedom still remain in the future? While I am a great believer in the role and inspiration of the individual, much research in the future is going to be conducted in large international teams with many of the experiments and simulations being undertaken remotely. The research environment of the future will be much more like that which particle physicists and astronomers have created already. Yet there will be a difference since this scenario will involve a mix of disciplines including social scientists and philosophers, with little common overlap in their knowledge base initially. Indeed will we still train university students in single disciplines? While I detect little enthusiasm for moving away from investigating one area in depth, much more attention will be required to ensure that, for example, a mechanical engineer will be able to work together with an environmental biologist, an expert on standards and a legislator. Only in this more holistic way will the challenges that face us now be resolved. Today there are few enough teams of critical mass that look to the challenges as a whole body problem. In fact as the debate on climate change has shown, there are even segments of society that do not even acknowledge the problem and some of them are in charge of finance. Today, more than ever, we need to have firm evidence based research that informs political decision making in an objective way. Ill informed media, looking for a popular campaign to boost sales, can make a logical political decision be a vote looser in no time at all.

If I am to think of the environment for a researcher at the beginning of their independent career, what things will be different from my own experience? Indeed, has the day of the Ph.D. passed or will it be transformed? Just how will I become a member of a dispersed community yet retain my own identity. What will some of the issues I will have to face and how can the decision makers in Europe assist?

The first question ERAB has asked itself is whether universities and publicly funded research organisations are best suited for exploiting this new environment. The question for politicians is how best to fund whatever is needed without jeopardising their commitment to mass higher education. There is a tension between the support of excellence and the need for a cohesive European society. As the researcher will be using resources all around the world, does it make sense to be identified with a single institution? One can conceive of molecular structural data being undertaken remotely using NMRs in Japan, a neutron source in the USA and an X-ray source in France. All these will be accessible and possibly controlled using dedicated electronic infrastructure. Indeed today, Imperial College has a dedicated link to Georgia Tech. allowing researchers in London to operate equipment in real time in Atlanta.

As the research infrastructures become more complex so does the volume of data that will be produced. For example the European X-ray Free Electron Laser being built at DESY in Hamburg is estimated to producing 100 times more data onto the Grid than the LHC at CERN will release. At a recent users meeting at the European XFEL it was stated that the reason so little data is released onto the Grid by CERN is the fact that the scientists know what they are looking for and can remove the vast majority of data that is not needed at source. This is not possible at the XFEL because of the wide range of disciplines and experiments that will be undertaken from mechanisms of drug delivery to fundamental investigations of superconductors.

The whole problem of data storage, curation and authentication is a hot discussion topic at the moment. Just who do we trust to keep the data safe and who decides on the way in which metadata is chosen? With threats such as terrorism, protection of data from misuse is a high priority and the problem will become more severe in the future.

The young researcher in 2030 therefore will have to rely heavily on the work done by peers in different disciplines in various organisations around the world. However, in contrast to today, he or she might not even have to leave their own house to do this, being constantly connected with labs in all corners of the planet through the virtual world. As today the general public can marvel at the results from the Hubble Space Telescope from their home PC so this new world will enable the general public more access to living science. There is a great opportunity to engage the whole population in experiments and analysis.

Coupled with the data explosion will be the ever increasing power of super computers. In Europe there is a policy for a unified approach known as PRACE. As we proceed into the 100s of petaflops and beyond, the simulations that are being undertaken will become ever more complex and difficult to verify by reviewers. The mix of capacity alongside capability will be challenging. However as the large research infrastructures (both those which are large physical ones and also large complex networks) will be spewing out data at enormous speed so the simulations will be informing the experiment in a real time feed back situation. Just how will this all be controlled?

The European project "Lifewatch" (www.lifewatch.eu) is an excellent example of the e merging scenario. This infrastructure links together structural biology with environmental sensing from both land based and space detectors coordinated by the European Space Agency. The data are processed at CERN enabling biodiversity to be monitored and the models developed for a sustainable environment in order to inform policy and decision makers. This is an ambitious programme and the way it is conducted in the future may well be a model of the research environment of the future.

The challenges before us are well known and include climate change, health in an ageing population, and the long-term effects of the current economic crisis. The biggest challenge is likely to be how to achieve a good quality of life for all people around the globe, and to do this in a sustainable manner. If we want to have even a remote chance to tackle these problems, we need creative scientists, working together in networks of excellence.

What is the role which the European Research Area will play in this global, interconnected world, i n which researchers meet virtually just as often as, if not more frequently than face to face?

Research will be as global as the challenges it tries to address. Purely national research will continue but is currently ill prepared for tackling the global challenges. Initiatives such as the proposed Joint Programming which seeks to encourage funding bodies in different Member States to pool resources, are to be encouraged. Europe has the opportunity to draw together a critical mass of research infrastructure and human capital. To achieve its full potential, the ERA of 2030 has to be characterised by excellence, openness and innovation, interconnected with the rest of the world. In addition to this, science has to become a core part of European society and be owned by all citizens who will celebrate its achievements but who know how to sift out reliable evidence for policy decisions.

To counter the challenges we face in the world we have to concentrate on quality of research and allow excellence to flourish. The Europe of the future should not be afraid of supporting and demanding excellence. However, the way this excellence is identified and defined will change depending on the sector and geography. A clear distinction is necessary to delineate between world class excellent research and achieving European cohesion. The current European programmes do not make this distinction clear enough such that politicians can claim the concentration on excellence to be unfair. If the whole of European society has sufficient broadband access then the requirement for a more equitable distribution of facilities becomes less tenable. Indeed there is a good argument for co-location of facilities in a few well serviced hot spots. The creation of the 'Medicon' valley in Sweden is one such example.

The Internet gives everybody the chance to become a publisher. This is not only true for life style blogs but has an increasing impact on the scientific world. It is now possible for science to easily reach large audiences, with the potential to eliminate the role of established filters and gatekeepers, such as the traditional scientific journal. This also means that science can be easily reviewed, assessed, rated and commented upon by anybody, reinforcing scientific democracy. Poor research might thus be identified more quickly and taken off the market. The challenge here is how to create these open access systems and how to ensure that old gatekeepers are not simply replaced by new ones.

The public access to assessment criteria will mean that individual reputation will become central and new ways to assess excellence will be developed. To state that "I am a scientist and I know best" will be unacceptable. A Ph.D. from an internationally respectable university will be a hallmark for some. However, as access becomes easier for all how will we give a kite mark to people with non conventional backgrounds? There will be more room for cranks to fill the networks with crazy ideas. Over the years I have received my own fair share of letters accusing me or wasting public money on large international research projects. They are normally typed on thin paper with a list of those copied that normally include the Prime Minister. I was once discussing these epistles with the head of the Cavendish Laboratory in Cambridge. We both took the same approach to ignore the letters but he then laughed and said he was always afraid that one of the crazy theories might turn out to be true and he would have gone down in history as ignoring it. This problem will grow.

New approaches, to assess and validate research in the public eye which are open, flexible and transparent are emerging, based on implicit and explicit data (such as incoming links, page views and ratings). These developments will lead to the evolution of a new model of assessment, based on qualitative and quantitative data, which will contrast with the existing G"impact factorG• model. Yet this does not mean that an organic, bottom-up and peer-to-peer approach is the only scenario: it is quite possible that some third-party system could be developed, which assesses and rates researchers in a dynamic way, including qualitative and quantitative input, emphasis on network analysis, implicit and explicit data – call it the Google of reputation.

In this networked world, an informal collaboration of scientists (even amateur scientists) may well produce better results than well established bureaucracies of research. Some research organisations could become irrelevant and be replaced by flexible networks. This would be accompanied by a labour market for researchers which will become more fluid, with greater mobility and differences in salaries. The European Research Area will profit from initiatives like the European Researcher Pass, which will allow researchers to move freely between all member states, carrying health care and social security with them where ever they go without long-term disadvantage.

However, all of this does not mean that research institutions will disappear. In a world overloaded by information, institutions can guarantee standards of quality.

Where will these institutions be located? Could they simply be in Second Life in many cases? All this begs the question of how these communities will interact socially. In many physical institutions the coffee room is where ideas are formulated and debated. In the virtual research environment will the same passion be developed or will the e-generation evolve accordingly?

Openness does not only mean that research will be conducted by networks of excellent researchers moving according to opportunities. Scientific research also has to be more transparent and accountable. There is significant pressure to make scientific data available to the public, especially data collected with public funding or owned by the government (within the scope of Reuse of Public Sector Information). The mood is changing and there is stronger recognition of the immense value which could be generated from reusing such data, overcoming resistances from commercial interests and traditional gatekeeper positions. Initiatives such as Google Research Datasets will become a reality. In particular beta research material and rough data will increasingly be available and provide invaluable sources for research. Crucially, this can include digital versions of live meetings, such as audio/video recordings of physical meetings and conversations between researchers. These resources can be extremely valuable in blurring the boundaries between formal and informal knowledge exchange. While informational technology will not substitute physical contact, it will enhance its meaning.

This new approach to ownership of data, which is already discussed in the context of the open access movement, will have to lead to a totally new business models for research as proprietary data would change the fundamentals of revenue generation. It also signifies a new emphasis on the research process itself rather than the single, perfect new invention. Innovation, new ideas, key scientific findings increasingly come in unpredictable ways, through continuous exchanges of view between high-level researchers with similar interests, and between relevant people at different levels of the innovation cycle. Increasingly, in the new open innovation model, innovative products are made public before being finalised, through the so-called permanent beta approach, because large-scale deployment brings insights which could not realistically be replicated in a protected environment. A parallel change is likely to happen in science: results are no l onger solely delivered as a finalised product (the publication of an important journal; the book) but as draft products, i n order to enable wider feedback mechanisms, and continuous improvement, enabled by the sharing of rough data, facilitating serendipitous innovation. This also helps meeting the challenge for improved knowledge transfer.

This process approach to science means that we will have to move towards a “science-as-a-service mentality” (comparable to software development), where research is not only funded because it may be able to obtain a specific result, but as an on-going process that enables not only the achievement of great inventions *ex-ante*, but also the unplanned emergence of new ideas.

A key question is how will privately funded research interact with this more open and transparent approach to research. The current emphasis on “Open Innovation” will come under immense strain in the current economic climate. Likewise there is a need for a fundamental overhaul of the rules relating to State Aid and to Intellectual Property Protection in order to avoid the desire for protectionism. This will be the true test of European resolve in the coming months.

In this new interconnected, globalised world, why should there be a “European Research Area”? Europe’s strength has always also been its weakness: its diversity. Research lives of the diversity of points of view. The European Union has already created an inner market which allows people from all member states to interact and trade with each other more easily. If it fully realises the

Fifth Freedom – the freedom of knowledge, it has the huge potential to create a space where cultural diversity can unleash creativity and innovation.

In addition to diversity, Europe also has a strong common tradition of scientific and philosophical discovery. This European tradition has an holistic approach to education and research. It does not see education merely as knowledge transfer, but rather as the formation of a human being as a member of society, expressed in words such as Bildung or formation. The German word Wissenschaft describes research in natural sciences as well as the arts, humanities and social sciences and the early modern period has seen the continuous development of all these areas. The European approach to scientific research, therefore, should build on the strength of this tradition. At the moment, arts, humanities and social science often see themselves as being in contrast to the natural sciences and there is very little cross-fertilisation. The early years of the 21^{st} century have shown how fruitful the interchange of ideas between these disciplines can be and we should aim for a stronger dialogue between them, this could help to identify new areas of research, as well as the communication and acceptance of natural sciences in the society. Trying to understand how other disciplines work, particularly if those disciplines are somewhat alien to one's own specialisation can also promote thinking outside the box and unleash creativity. To enable this, students of the different disciplines will have to be educated so that they can communicate with each other more effectively. This has to be started long before students enter university and perhaps a greater emphasis on being a European Citizen should be included in schools' curricula. Building on the tradition of early modern renaissance people could be one of the great benefits Europe can bring to the globalised world of science.

I'm wondering about a new chapter here mpa

5.4 Analysis Paradigms

Over the week we saw a very interesting range of talks on a wide variety of topics. It was easy to be reminded of the story of the blind men and the elephant, in that DIR seemed to look very different to different people. However, there were many talks that described different kinds of data analysis, including those by Thore Graepel (Microsoft Research) on analyzing large-scale complex data streams from online services; Alan Heavens (University of Edinburgh) on modelling large datasets in astronomy and medical imaging; Jonty Rougier (University of Bristol) on incorporating parameter uncertainty into the data assimilation process for climate science; Mike Batty (University College London) on large scale geospatial data analysis; Paul Lambert (Stirling University) on e-Social Science data challenges; Chris Williams (University of Edinburgh) on the use of probabilistic graphical models in data analysis; Jim Austin, (University of York) on prognostics and diagnostics in engineering systems; Andrey Rzhetsky (University of Chicago) on biomedical text mining methods; and Andrew McCallum (University of Massachusetts Amherst) on discovering patterns in text and relational data with Bayesian latent-variable models.

The above talks covered the topics of exploratory data analysis, descriptive data analysis and predictive data analysis. The need for data cleaning and quality control was also highlighted. The talks also addressed the different dimensions of data scale, data complexity and model complexity.

Some notable issues that came up in the week were: Alex Szalay's observation on the power law for data set sizes, i.e. that there are far more small and medium sized datasets than large ones; that simulation data can also require data analysis (Szalay); the hard problem around inferring the structure of networks/circuits from observational and interventional data; the need to incorporate domain knowledge into models (e.g. using structured probabilistic graphical models).

Some take-home messages from the week are that

- There is considerable need for and opportunities for data analysis.
- This will require collaboration between domain and data experts.
- One useful way to facilitate this collaboration is via training for young (and not-so-young) researchers in data-centric thinking (e.g. around the organization and analysis of data).
- Examples of how this might be achieved are through courses in MSc or doctoral training programmes, or via summer schools.

5.5 Paradigms to structure data

The problem of data structure has gone through extensive approaches and transformations in the past. One thing has become even clearer from the workshop: when dealing with data structure, we are not referring to a data model or a data format. Rather, we desire mechanisms to define, describe, represent, process, and transform data; moreover any of those mechanisms needs to be explicitly recorded as well. Therefore, “structure” comes in many flavours and paradigms are necessary for structuring and manipulating any —or, better, all— aspects:

Data. The richness of the application domains is the main driver for increasing the complexity of data structuring paradigms. It is not as simple as for, say, organisational data, which tend to be easily describable in one of the existing data models like the relational or the XML one. Rather, there is a need for even richer data models that deal with specific domains and allow for the easier exchange of information. There needs to be a way of standardised representation of data, most likely per domain. This will allow both exchange and collaboration towards problem solving. In a way, it will establish a common language that all participating entitities speak.

Metadata. It is not always the case that metadata can simply be regarded as data and revert to one of the data structuring paradigms. Especially in experimental observations, one missing link that is crucial is the procedure by which an observation was captured. The instrument that produced it, different parameters of its operation, even worse enviromental parameters that might have influenced the observation and have made it non-reproducable. There is also an inherent “importance” of the observation that is hidden in this discussion: the less frequent, or the harder to capture an observation the more important it becomes.

Complexity. Computer scientists have well-defined and established ways of dealing with complexity, for instance in terms of data complexity or computational complexity. A formal model for experimental complexity, or observational complexity, or even interactive complexity of the data becomes a key player in data-intensive research. A principled approach towards developing notions of understanding the above is important in structuring the data in a way that lowers the corresponding complexity. To date, no such approach exists.

Provenance. This is perhaps the most important need that has emerged from the workshop and encompasses all the above notions of structure. It is often the case that data-intensive research deals with *second-level* data: data that has not emerged from primary observations, but rather data that has emerged from transformations of those primary observations. As such, a provenance model that tracks these transformations and can readily answer queries regarding their origin, or allow a researcher to “play back” the transformation or use one of its mid-points for further processing is crucial.

All of the issues touched upon above are merely the main axes of the data structuring paradigms that need to emerge if one aims to take data-intensive research at the next level and make it applicable across disciplines.

5.6 Programming Paradigms

When motivating the Programing Paradigms cross-cutting theme, the organisers framed the following motivational questions:

- Advantage and applicability of programmatic approaches over others.
- A mapping between existing approaches and application requirements. What is missing? How can these be met?
- Many approaches are tied to a specific infrastructure (e.g. Hadoop on HDFS). Is this lack of interoperability and extensibility a limitation and can it be overcome? Or does it reflect how applications are developed?
- How does the way we store and manage data (distributed versus local, structured versus table) influence our ability to process it?

Over the course of the workshop, many interesting talks referenced and reinforced challenges arising from the large-volumes of data. The aim was not to arrive at a comprehensive answer or solution to the above, we believe we gained insight into several of the questions above.

The primary focus however was on the data-management, integration and complexity of the data. The recurrence of linked data and analysis using linked data highlighted the need for ‘data analytics’ and scripting approaches to facilitate such analytics. In general however, both interestingly and surprisingly, limited emphasis was placed on the programmatically controlling and analysing large-volumes of data. We believe this was just a curious coincidence reflecting the talks that participants chose to present, as well as partially reflecting the fact that most talks focussed on “successes” and not “failures”. Additionally, it can be suggested that a large number of research problems that the community was addressing was more data-oriented, rather than *data-intensive.*

As the scale of data increases along several dimensions (volume, distributedness, complexity, etc.) a we need to rethink the way we programmatically handle different data at different stages (management, production and analysis). There have been several recent advances towards programmatically addressing these challenges, e.g., Sawzall [101], Pig [102], Dryad [103], not to ignore the many variants of MapReduce. Many of these approaches re-establish the primacy of data parallelism. Programming approaches will not be applicable in all cases, but where the rate of increase of ϵ with data size is greater than one, programming approaches will become critical.

As a specific contribution to the first point, we identified a classification scheme for data-intensive application – based upon (ϵ) defined as the ratio of compute requirements and data-processing (to have units of ops per byte [36]. In passing, we observed that there wasn’t a well defined or broadly agreed upon definition of ‘data-intensive’. As a corollary we also observed that there do not exist data-intensive benchmarks, and an explicit aim of a future DIR theme would be to identify and propose such a suite of benchmarks.

Additionally, our experience over the week long workshop led us to suggest:

- Related to the fact that there was limited discussion of the abstractions and programming paradigms for data-intensive, there was limited discussion of the challenges associated with the systems to support them. However, there was universal acknowledgment of the fact that as soon as the interoperability-integration-management aspects were resolved, the data-processing challenges would dominate.
- Much of the focus was on databases. But not surprisingly as data sizes get larger, both the CI (used as in contrast to Relational data-stores http://en.wikipedia.org/wiki/NoSQL) and programming requirements to process the data will emerge.
- The programming-paradigms breakout groups identified Clouds as part of the cyber-infrastructure that would be used and that Map-Reduce provided a good starting point to understand and analyse the role of data-intensive programming paradigms. (see sub-section on "Research Areas for Scientific Computing with MapReduce and Clouds").

Another summary statement, reflecting the boundary conditions of the participants and their interest, a large number of science applications that received extensive discussion were from the life-sciences, e.g. genomic and pathway data for example. The role of data-analytics in the life-science domain was well established.

5.6.1 Research areas for Scientific Computing with MapReduce and Clouds

There are multiple reasons to believe that clouds (see §1.9.1) will form an increasingly important component in the cyber-infrastructure used for data-intensive. This is in spite of well known deficiencies in the current cloud paradigm and offerings, namely, *a*) the centralised computing model for clouds runs counter to the concept of "bringing the computing to the data" and bringing the "data to a commercial cloud facility" may be slow and expensive, *b*) the virtualised networking currently used in the virtual machines in today's commercial clouds and jitter from complex operating system functions increases synchronisation/communication costs. This is especially serious in large-scale parallel computing and leads to significant overheads in many MPI applications. Indeed the usual (and attractive) fault tolerance model for clouds runs counter to the tight synchronisation needed in most MPI applications. There are other barriers to the whole-sale adoption of clouds, such as issues raised by security, legal and privacy, but we will not discuss them here as they are not typically determinants of performance for scientific data-intensive applications.

Some of these issues can be addressed with customised (private) clouds, however it seems likely that clouds will not supplant traditional approaches for very large scale parallel (MPI) jobs in the near future. It is natural to consider a hybrid model with jobs running on either classic HPC systems or clouds or in fact both as a given workflow could well have individual jobs suitable for different parts of this hybrid system.

Commercial clouds support "massively parallel" applications but only those that are loosely coupled and so insensitive to higher synchronisation costs. Let us focus on "massively parallel" or "many task" cloud applications as these most interestingly 'compete' with possible TeraGrid implementations. In this case, the programming model MapReduce describes problems suitable for clouds. One can compare MPI, MapReduce (with or without virtual machines) and different native cloud implementations and find comparable (within a range of 30%) performance on applications suitable for these paradigms [104]. However, MapReduce and its extensions offer the most user friendly environment.

One can describe the difference between MPI and MapReduce as follows. In MapReduce multiple map processes are formed – typically by a domain(data) decomposition familiar from MPI – these run asynchronously typically writing results to a file system that is consumed by a set of reduce tasks that merge parallel results in some fashion. This programming model implies straightforward and efficient fault tolerance by re-running failed map or reduce tasks. MPI addresses a more complicated problem architecture with iterative compute–communicate stages with synchronisation at the communication phase. This synchronisation means for example that all processes wait if one is delayed or failed. This inefficiency is not present in MapReduce where resources are released when individual map or reduce tasks complete. MPI of course supports general (built-in and user-defined) reductions so MPI could be used for applications of the MapReduce style. However the latter offers greater fault tolerance and a user-friendly, higher-level environment largely stemming from the coarse grain functional programming model implemented as side-effect free tasks. Over simplifying, MPI supports multiple Map-Reduce stages but MapReduce just one. Correspondingly clouds support applications that have the loose coupling supported by MapReduce, while classic HPC supports more tightly coupled applications. Research into extensions of MapReduce attempt to bridge these differences.

MapReduce covers many high throughput computing applications including “parameter searches”. Many data analysis applications, including information retrieval, fit the MapReduce paradigm. In LHC or similar accelerator data, map functions consist of Monte Carlo generation or analysis of events while the reduction function is the construction of histograms by merging distributions from different maps. In the SAR [4] data analysis of ice sheet observations, map functions consist of independent Matlab invocations on different data samples. Life Sciences have many natural candidates for MapReduce including sequence assembly and the use of BLAST and similar programs. On the other hand, partial differential equation solvers, particle dynamics and linear algebra require the full MPI model for high performance parallel implementation.

MapReduce and Clouds can be used for some of the applications that are most rapidly growing in importance. Their approach seems essential if one is to support large-scale, data-intensive applications. More generally a more careful analysis of clouds versus traditional environments is needed to quantify the simplistic analysis given above.

There is a clear algorithmic challenge to design more loosely coupled algorithms that are compatible with the map followed by reduce model of MapReduce or more generally with the structure of clouds. This could lead to generalisations of MapReduce which are still compatible with the cloud virtualisation and provide fault tolerance.

There are many software challenges including MapReduce itself; its extensions (both in functionality and higher-level abstractions); and improved workflow systems supporting MapReduce and the linking of clients, clouds and MPI engines. We note there is also active work in the preparation, management and deployment of program images (appliances) to be loaded into virtual machines. The intrinsic conflict between virtualisation and the issues around locality or affinity (between nodes in MPI or between computation and data) needs more research.

On the infrastructure side, we have already discussed the importance of high-quality networking

[4] `http://www.asf.alaska.edu/sardatacenter`

between MPI and cloud systems. Another critical area is file systems where clouds and MapReduce use new approaches that are not clearly compatible with traditional production Grid infrastructures. Support of novel databases such as Big Table across clouds and MPI clusters is probably important.

5.7 Dynamic Distributed Data Issues

There has been a lot of effort in managing & distributing tasks where computational loads are dominant. Such applications have after all, been historically the drivers of "grid" computing. There has however, been relatively less effort on tasks where the computational load is matched by the data-load, or even dominated by the data-load. For such tasks to be able to operate at scale, there are conceptually simple run-time trade-offs that need to be made, such as: *a*) determining whether to move data to computational nodes, versus *b*) keeping data localised and moving computational tasks to operate on the data *in situ*, or *c*) possibly neither, and just regenerate data *on the fly*. Due to fluctuating resource operating-points, it is essentially not possible to make such decisions upfront; currently, it is also very difficult to implement these *dynamic decisions* in a general-purpose and scalable fashion.

Although the increasing volumes and complexity of data, will make many problems dominated by the data-load, the computational-requirements will still be high. In practise, data-intensive applications will encompass data-driven applications — which are the basis for much of the *Fourth Paradigm* of science [25]. For example, many data-driven applications will involve computational activities triggered as a consequence of independent data creation; thus it is imperative for an application to be able to respond to unplanned changes in data load or content. Therefore, understanding how to support dynamic computations, is a fundamental, but currently a crucial missing element in data-intensive computing.

Data-intensive research must herefore operate at the triple point of *data-intensive*, *dynamic* and *distributed* (3D) attributes. More often than not, the focus will be on applications that represent the merger of the *Big-Data* problem **AND** with the need for supporting *Dynamic-Data* **AND** which may either be fundamentally distributed or need to be distributed.

5.7.1 Establishing the Fundamental Role of Dynamic-Data

We outline three simple, yet representative 3D applications that will help to highlight the fundamental role of Dynamic-Data in data-intensive computing:

Application 1: Sensor Data-driven Computation: For large-scale applications such as LEAD [5] the data-stream from the sensors drives the computational execution. Often in response to a predicted, or phenomenologically interesting event, the data-source and stream itself needs to be adapted, e.g. sampling rates (step-up/down), changing resolution, etc. Additional elements of dynamic adaptation – compute and data, arise from spatio-temporal variations in data generation (sensors).

Application 2: Dynamical Source and Analysis: A very nice example of dynamic distributed data-intensive applications is provided by http://ClimatePrediction.net [105], which has two intrinsic computational models, one for generating the data, and one for analysing the data.

[5] https://portal.leadproject.org/

Associated with the former is the concept of *phases*, where a client run will return a usable piece of the result part way through the computation of its work unit. The use of distributed computing in the data generation phase allows a collaborative processing, where the ClimatePrediction.net team can run a large set of models without having to own all of the computing and storage resources.

The model for the analysis of the data is less well known. Here there are a variable number of highly distributed computing resources and data stores. The data is distributed across the data stores in a regular way, with complete results from individual models always residing on the same server. The data is typically too big to transfer to a single location, so analysis must be performed *in situ*. It is necessary to provide some form of abstraction from the changing number of distributed resources. This was provided through a data parallel workflow language, Martlet [106], that contains a set of abstract constructs with which to build analysis functions. Other solutions, such as Parallel Haskell, are able to handle the changing numbers of computational resources, but a unique feature of this model of computing is that the *number of data sources is also changing*. There is definitely a gap here to be filled by further abstractions, as the current constructs (including Martlet) are just simple prototypes and their is a need for extensions with more powerful and useful ones.

Application 3: In-transit Adaptation: In many dynamic applications processing of data often needs to take place "in-flight" (e.g. via streaming) to meet boundary conditions, and/or data-volume reduction may be required to effectively store/manage data. An application example is provided by the coupled-fusion simulations that aim to provide integrated predictive plasma edge simulation to support next-generation plasma experiments, such as International Thermonuclear Experimental Reactor (ITER). Here data has to be transformed while it is being streamed using a mesh interpolation module to satisfy the different formulations – domain configurations and decompositions. Similar to the LEAD example, depending upon the specific components being connected (which is time dependent), the in-flight processing has to change.

The first application makes the case for agile execution and control of computing and data; the third reinforces the need to couple dynamic activity to data streams. As 3D applications become pervasive, the importance of dynamic placement, management and scheduling of data and data-sources/sinks will increase, as illustrated by all three examples. But there are several limitations in the current understanding and handling of Dynamic-Data, some of which carry-over from constraints in the way we handle the Big-Data challenge. For example, to address the challenge of Big-Data, several programming models, such as MapReduce & variants, have been developed. And even though the solution-space for Big-Data is not complete, most existing programming models (and associated tools and services) typically assume that the underlying data-set is "static", i.e. the work-load assigned to a worker does not change during execution. Thus, performance, deployment and execution decisions, once made, are typically assumed to be valid throughout the life-cycle/execution of the application. This situation is analogous and reminiscent of the first-generation of distributed applications that inherited the static execution models of legacy cluster applications. It is only with the right tools, abstractions and run-time support that a subsequent class of distributed applications have been able to break-free of the static (resource) usage model. For traditional distributed applications, the ability for dynamic resource utilisation and optimisation has lead to a concomitant performance enhancement. *Thus, any vision and plan for managing a data-intensive future should include a strategy for supporting such dynamic and distributed aspects of data-intensive computing.*

5.8 Social and Ethical Issues

5.8.1 Use of personal data in DIR

The growing power of data and the new methods raise social and ethical questions about the delivery of benefits and the protection of individuals. The use of personal data in DIR was discussed in the breakout session ‘sensors everywhere’. As citizens in modern industrial societies go about their daily lives, they leave digital data trails behind them. These real-time data are captured from sensors, digital communications and transactions with online software services. They can be used to model and predict behavioural patterns, detect anomalies, and prompt interventions. They can be used in applications of which the data ‘donors’ are unaware, and which do not necessarily benefit them personally, whilst being beneficial — profitable even — for others. Moreover, when data from separate databases are integrated, measures to anonymise the data ‘donors’ can be undermined.

The collection and use of personal data in DIR challenges conventional understandings of privacy, autonomy, consent and confidentiality. However, that fact that people volunteer personal data through Web 2.0 could indicate that norms of privacy are changing. Moreover, some DIR applications offer benefits for which privacy and autonomy may be considered worth sacrificing. Indeed, not maximising the potential of DIR to optimise social interventions could itself be considered unethical.

The Connected Home (see §4.7) could be a case in point. This project is developing sensors to monitor domestic activities as part of a system of care that prompts intervention when abnormal behaviour is detected. The motivation for the project is to enable frail and elderly people to live in their own homes for longer.

However, whether or not this project will have its intended impact is not just a matter of technological design. It also depends on the answers to questions like: Who will provide and pay for these services? How will vulnerable users / consumers be protected from unscrupulous and untrustworthy equipment, service providers and hackers? Who will monitor the signals and respond to the alerts? Who will be liable for faulty systems? Who guards the guardians?

As the Connected-Home project illustrates, the positive and negative impacts of the applications of DIR will not be determined solely by the design and operation of the technological systems, but will also crucially depend upon the configuration of the *socio-economic and legal* systems of which they are an integral part. Moreover, it is likely that what counts as privacy, confidentiality, reliability, trust and security, for example, will change depending on the setting. What is needed to explore these dynamics are detailed descriptions of the imagined future worlds envisaged for the operation of these new technological systems. These realistic scenarios — including worst case scenarios — could then be used for discussion and to inform policy decisions and legislation.

5.8.2 Changing practices in scientific research

As its name suggests, DIR is data intensive. Gathering the necessary volumes of data and analysing them means that DIR depends upon collaborations across borders and disciplines and amongst strangers, in some cases, including citizens (see for example, §1.3 & §4.4; . . .). The transition to DIR requires changes in research practices, in particular, the sharing and reuse of data and workflows. Infrastructures, software and standards are being designed to support and facilitate these changes. However, even the most marvellous technological or computational solution will not be adopted unless it is useable, and without their close co-operation upstream, what is designed and developed may not align with the needs and preferences of laboratory practitioners.

The SysMO project (see §2.9) illustrates what it means to integrate user preferences and needs, and their existing practices into the design and development of a web-based system to exchange, search, and disseminate data. DataONE (see §4.4) is another example where the socio-cultural challenges were recognised as ranking alongside the scientific and cyberinfrastructure challenges of building a virtual data centre.

SysMO illustrates another important lesson which is that technical and computational fixes alone will not suffice to overcome barriers to changing research practices. Changing practices will also depend upon social innovations, for example, systems of recognition and reward for data curation, sharing and reuse. The lesson here is that the socio-economic changes necessary for the implementation of DIR should be evaluated and co-developed alongside the computational and technological ones. Importantly for collaboration with strangers and potential competitors is the need to devise novel socio-technical ways of building trust. Moreover, one size is unlikely to fit all; there are likely to be differences in the factors that influence practices in different disciplines and international settings.

5.8.3 DIR and the sociology and philosophy of science and scientific knowledge

For the social sciences and humanities, DIR is not only a way of doing research (see §3.1), the very concept of DIR poses questions for research in the sociology and philosophy of science and scientific knowledge.

One manifestation of DIR is the ‘data-driven revolution’ (see §4.1.1), heralded as a new research paradigm which derives hypotheses from the data. However, the data never speak for themselves. What assumptions – about the data and about the world – underlie their use?

What is the relationship between DIR and hypothesis-driven research and theory? For example, Microsoft’s ‘Matchbox’ (see §1.5) uses large-scale customer purchase databases to model taste in films and make recommendations that are intended to influence individual purchasing behaviour. Yet it is not clear what, if any, traditional psychology theory this modelling incorporates. What is the discipline of this science of human behaviour? Is it a human science?

The history of science shows that adoption of new research technologies and infrastructures affects the types of knowledge produced. What difference will the technologies and cyberinfrastructures that are integral to DIR make to the direction of research and the kinds of questions that are (and are not) addressed?

DIR is not just about knowing the world; it is oriented towards changing the world by making recommendations and interventions, correcting malfunctions, optimising outcomes. Search engines, for example, personalise search returns. What kinds of worlds are being promoted through the selective interventions DIR supports? And what alternatives are being systematically excluded?

In practice, some of the data used in DIR may be incorrect and incomplete, the models imperfect, the simulations inaccurate, and the interventions inappropriate. And some of the time this may not even be noticed. But what happens when DIR disagrees with small data science, or when DIR comes into conflict with everyday experience? Who has the expertise and power to challenge knowledge and interventions based on the power of numbers? There are lessons here from the UK National DNA Database (NDNAD). In arguably one of earliest data-driven applications in public policy, the NDNAD can lawfully be used to generate suspects for unsolved crimes. The intervention for the suspect can be arrest, prosecution, trial and even conviction. However, false positive database matches can occur, especially when the crime scene profile is partial or mixed. The difficulty in challenging a database match with other kinds of evidence illustrates the extraordinary power of data-intensive forensic science. Is this likely to be the case for other DIR-based interventions? How will DIR be challenged by other forms of knowledge? What happens when there is a difference of opinion about whether an intervention is warranted and what that intervention should be? Are there lessons here from evidence-based medicine?

Chapter 6

Summary, Recommendations and Future Work

This chapter is in a preliminary state. It will (possibly in a second edition) hold a summary and analysis (§6.1), which contains tables showing the relationship between the talks and other activities during the workshop, with Data-Intensive challenges, methods and tools. This will support a 'gap analysis' that leads to recommendations (§6.2) and to identification of future research (§6.3).

6.1 Summary and Analysis

The week's activities and messages are summarised in four tables:

1. the data-intensive challenges posed by each talk, e.g. coping with existing/anticipated data volume, are shown in Table 6.1;
2. the data-intensive methods demonstrated as effective, e.g. sampling to make analyses feasible, are shown in Table 6.2;
3. the data-intensive tools that are already available for others to use, e.g. parallel R, are shown in Table 6.3; and
4. other messages that do not fit in one of the above tables, e.g. the requirement for data-intensive education, are shown in Table 6.4.

Challenges raised during the workshop

Talk		Volume	Complexity	Analytics	Demand	Interaction
Szalay	§1.3	Sky-survey data: SDSS, NVO & PanSTARRS; repeated analysis of data from simulations; and sensor networks	SDSS schema 200 pages (§2.5)	Queries and analytic derivations over preceding data	930,000 distinct users of SkyServer	

continued on next page

Challenges raised during the workshop (continued)

Talk		**Volume**	**Complexity**	**Analytics**	**Demand**	**Interaction**
Kell	§1.4	New sequencing data, metagenomic and diagnostic applications	Composing many sources of data from literature plus multiple laboratory and field sources	data-driven hypothesis generation & model complexity exceeding human analytic capacity	NG sequencing will be widely used	Collaborative multi-disciplinary knowledge building
Graepel	§1.5	Streams of global commercial data, e.g. game-playing records, clicks in a search engine & users' choices of media downloads	Many variables potentially significant in models of human behaviour	Large Bayesian models	Large flows of new data plus complex models	Iteration towards operational decision support
Llorà & A’cs	§1.10		Handling social-science and Arts & Humanities applications			

Table 6.1: Challenges raised during the DIR workshop

Methods introduced at the workshop

Talk		**Volume**	**Complexity**	**Analytics**	**Demand**	**Interaction**
Szalay	§1.3	IO-balanced architectures, RDB optimisers as analysis framework, 3D-clustering indexes	Mapping to relational models	user-defined functions in SQL-server	SQL-server on clusters & repeated analysis of simulation data	Web portals accessing canned queries
Kell	§1.4	Investment in TGAC & Elixir	Text-mining, data-driven science	Text mining & machine learning	Cloud computing investment at EBI	*in silico* biology, data-driven hypothesis generation & collaborative shared-knowledge building
Graepel	§1.5	Approximate inference, parallelised with MPI	Factor graph probabilistic models	Machine-learning in large-scale production systems	Cloud provisioning for machine-learning	

continued on next page

Methods introduced at the workshop (continued)

Talk		**Volume**	**Complexity**	**Analytics**	**Demand**	**Interaction**
Llorà & A'cs	§1.10		Meandre semantically driven workflows & higher-level programming abstraction in Zigzag			

Table 6.2: Methods introduced at the DIR workshop

Tools presented at the workshop

Talk		**Volume**	**Complexity**	**Analytics**	**Demand**	**Interaction**
Szalay	§1.3	GrayWulf & Amdahl Blade servers				
Graepel	§1.5	Cosmos	PQL & graphic tools for Bayesian networks	TrueSkill^{TM}, AdPredictor^{TM} & Matchbox^{TM}	Cosmos	
Llorà & A'cs	§1.10		Meandre & Zigzag		NCSA cloud	
Stall 2	§1.6		Taverna & Biocatalogue			myExperiment
Stall 3	§1.6	SPRINT (parallel R), DiGS global data federation & DATAMINX data interchange	OGSA-DAI distributed data fusion	SPRINT	DiGS	
Stall 4	§1.6		Inforsense		Discovery cloud	
Stall 5	§1.6		DISPEL	biomedical image data-mining		RAPID
Stall 6	§1.6	MonetDB: column-based data store	MonetDB: relational, XML & RDF models	MonetDB: OLAP data cube		
Stall 7	§1.6		MapTube		Crowd-sourced data	MapTube & Tweetometer
Stall 8	§1.6	MOPED: sampling		MOPED: focused approximation		

continued on next page

Tools presented at the workshop (continued)

Talk		Volume	Complexity	Analytics	Demand	Interaction
Stall 9	§1.6					Southampton Smart Labs

Table 6.3: Tools presented at the DIR workshop

Other ideas and issues developed at the DIR workshop

Talk		Description
Szalay	§1.3	Imminent power-wall addressed by low-energy balanced architectures
Kell	§1.4	Challenge of tracking and recognising credit and responsibility in the collaborative integration and assembly of data
Kell	§1.4	Careers and respect for the technical experts who enable data-intensive research
Graepel	§1.5	Determining the acceptable limits of machine-learning over people's behaviour

Table 6.4: Other ideas and issues raised at the DIR workshop

6.2 Recommendations

The following are draft and preliminary recommendations, pending discussion with other editors mpa

The following recommendations are in arbitrary order, i.e. order does not imply priority or importance; some may depend on progress against other recommendations.

Application analysis A study should be conducted of the applications identified here and extended to cover a representative sample, in order to characterise and cluster the set of challenges applications exhibit. These characterisations and clusters should be examined to identify common patterns which may be addressed by similar solutions.

Benchmarks and measurement Characteristic examples, including data, workload patterns, required queries and analyses should be defined and made publicly accessible. These should lead to agreed methods for measuring progress. The measurements should include factors such as: *a*) response time, *b*) throughput or rate of delivering results, *c*) energy used to deliver results, *d*) quality of results – accuracy, stability & ease of interpretation, *e*) speed of new users in gaining first useful results, and *f*) speed of experienced users in setting up new/revised sophisticated analyses.

Ethical, social and legal The advent of high rates of flow of multiple sources of personal data, coupled with widespread availability of governmental data transforms opportunities and expectations. The ethical, social and economic issues of both using and failing to use these sources of data are changing. Consideration should be given to understanding the appropriate new balances to strike, at various geographical scales. These balances may need to be reflected in revised legislation at those scales.

Education The DIR workshop recounted many new opportunities and methods, and it was clear that we are in a time of rapid change (the 'digital revolution' – see §1.2) when many more opportunities will arise. It is therefore necessary to devise educational programmes for all ages to better equip people for life and research in a data-abundant world.

Ramps The concept of an 'Intellectual ramp' was introduced by Atkinson (§1.2) and examples appeared in several talks. There is as yet not classification of ramps, no systematic approach to deciding when ramps are appropriate, designing and implementing them and evaluating their efficacy. Such R&D into ramps would pay dividends in accelerating the adoption of data-intensive methods, in facilitating the education and in reaching significant parts of the ‘long tail’.

6.3 Future work

Appendix A

Participants

Data-Intensive Research Workshop Participants

Name	**Department**	**Organisation**
Mr Bernhard A'cs	NCSA	University of Illinois at Urbana-Champaign
Mr Asif Akram	Oxford University Computing Service	University of Oxford
Dr Rosalind Allen	School of Physics	University of Edinburgh
Ms Sheila Anderson	Centre for e-Research	King's College London
Prof Malcolm Atkinson	Director	National e-Science Centre
Prof Jim Austin	Department of Computing Science	University of York
Dr Jerome Avondo	Cell and Developmental Biology	John Innes Centre
Prof Mark Baker	SSE	University of Reading
Dr Richard Baldock	Comparative and Developmental Genetics	MRC Human Genetics Unit
Prof Michael Batty	Centre for Advanced Spatial Analysis	University College London
M Beckett	EPCC	University of Edinburgh
Dr Nigel Binns	Genomic Technology & Informatics	University of Edinburgh
Dr Mark Birkin	Geography	University of Leeds
Prof Peter Buneman	School of Informatics	University of Edinburgh
Dr Mario Caccamo	Bioinformatics	The Genome Analysis Centre
Mr Archie Campbell		Generation Scotland
Miss Yin Chen		Edinburgh University
Dr James Cheney	School of Informatics	University of Edinburgh
Mr Neil Chue Hong		OMII-UK (Southampton)
Mr Jeremy Cohen	Computing	London e-Science Centre
Dr Iain Coleman		National e-Science Centre
Dr Oscar Corcho		Universidad Politécnica de Madrid
Prof David De Roure		University of Southampton
Dr Gerard Devine		University of Reading
Prof Simon Dobson	School of Computer Science	University of St Andrews
Mr Matthew Dovey	Development	JISC
Dr Alastair Droop	York Centre for Complex Systems Analysis	University of York
Dr Torild van Eck		ORFEUS

continued on next page

Data-Intensive Research Workshop Participants (continued)

Name	Department	Organisation
Ms Sonia Evans	Computer Science	University of St Andrews
Dr Geoffrey Fox	Computer Science, Physics & Informatics	Indiana University
Mr Hugh Glaser		Seme4 Limited
Prof Carole Goble	School of Computer Science	University of Manchester
Mr Steven Gray		University College London
Mr Thore Graepel	OSA Research Group	Microsoft Research Cambridge
Dr Magnus Hagdorn	School of GeoSciences	University of Edinburgh
Prof Keith Haines		University of Reading
Dr Liangxiu Han	School of Informatics	Edinburgh University
Dr Andy Heath	Earth & Ocean Sciences	Liverpool University
Dr Alan Heavens	Physics and Astronomy	University of Edinburgh
Dr Jano van Hemert	National e-Science Centre	University of Edinburgh
Mr Chris Higgins		University of Edinburgh
Dr Simon Hodson	JISC Executive, Innovation Group, E-Research	JISC
Mr Adam Huffman		University of Manchester
Dr Lorna Hughes		King's College London
Mr Ally Hume	EPCC	University of Edinburgh
Mr Kashif Iqbal		Irish Center for High End Computing (ICHEC)
Dr Milena Ivanova	INS	Centrum voor Wiskunde en Informatica
Prof Paul Jeffreys	Office of the Director of IT	University of Oxford
Mr Scott Jensen	Computer Science	Indiana University
Dr Shantenu Jha	CCT	Louisiana State University, Baton Rouge
Dr Daniel Katz		University of Chicago
Dr Graham Kemp	Computer Science and Engineering	Chalmers University of Technology
Ms Alison Kennedy		EPCC
Prof Jessie Kennedy	School of Computing	Napier University
Prof Martin Kersten		Centrum voor Wiskunde en Informatica
Mr Rob Kitchen		National e-Science Centre
Dr Jos Koetsier		National e-Science Centre
Dr Paul Lambert		University of Stirling
Dr Peter Li	Computer Science	Manchester Interdisciplinary Biocentre
Mr Chee Sun Liew	Informatics	National e-Science Centre
Dr Dave Liewald	Generation Scotland	University of Edinburgh
Mr Xavier Llorà	NCSA	University of Illinois at Urbana-Champaign
Dr Elizabeth Lyon	UKOLN	UKOLN
Mr Stuart Macdonald	EDINA	University of Edinburgh
Dr Adrian Mackenzie	cesagen	Lancaster University
Dr Robert Mann	Institute for Astronomy	University of Edinburgh
Dr David McAllister	Research, Innovation & Skills	BBSRC
Dr Andrew McCallum		UMass
Dr Alistair McGowan	Geographical and Earth Sciences	University of Glasgow
Miss Katie McMurray		IDBS
Dr Ruth McNally	ESRC Cesagen	Lancaster University
Dr William Michener	Biology Department & LTER Network Office	University of New Mexico

continued on next page

Data-Intensive Research Workshop Participants (continued)

Name	**Department**	**Organisation**
Prof Andrew Millar		University of Edinburgh
Dr Mike Mineter	Centre for Earth System Dynamics	School of Geosciences
Mr Benjamin Panter	Institute for Astronomy	Royal Observatory
Dr Mark Parsons	NeSC & EPCC	University of Edinburgh
Mr James Perry	EPCC	University of Edinburgh
Mr Savvas Petrou	EPCC	University of Edinburgh
Mr Chris Place		School of Geosciences
Dr Beth Plale	Computer Science	Indiana University
Dr Omer Rana	Computer Science	Cardiff University
Prof David Robertson	Informatics	University of Edinburgh
Dr Keith Rochford		Dublin Institute for Advanced Studies
Dr David Rodriguez Gonzalez	National e-Science Centre	School of Physics
Dr Jonathan Rougier	Mathematics	University of Bristol
Dr Anthony Rowe	Dept of Computing	Imperial College
Mr Chris Rusbridge	School of Informatics	University of Edinburgh
Prof Andrey Rzhetsky	Medicine & Human Genetics	University of Chicago
Dr Joel Saltz	Center for Comprehensive Informatics	Emory University
Dr Alessandro Spinuso	Seismology	KNMI
Dr Amos Storkey	Informatics	University of Edinburgh
Dr Charles Sutton	ANC	School of Informatics
Prof Jason Swedlow	College of Life Sciences	University of Dundee
Prof Alex Szalay	Department of Physics & Astronomy	Johns Hopkins University
Dr Claire Tansley		EPSRC
Dr Costantino Thanos	ISTI	CNR
Dr Stratis Viglas		University of Edinburgh
Dr Paul Watson		University of Newcastle upon Tyne
Dr Angus Whyte	Digital Curation Centre	University of Edinburgh
Dr Max Wilkinson	Digital Library Technology	British Library
Dr Chris Williams	Informatics	University of Edinburgh
Prof Robin Williams	ISSTI	University of Edinburgh
Dr Katy Wolstencroft	Computer Science	University of Manchester
Dr Simon Wong		Irish Center for High End Computing (ICHEC)
Prof John Wood	Faculty Office	Imperial College
Dr Gagarine Yaikhom	School of Informatics	School of Informatics
Mr Fan Zhu		School of Informatics

Table A.1: Participants at the Data-Intensive Research Workshop

Appendix B

Timetable

Timetable

Time	**Title**	**Speaker**	**See**
	Monday 15th March 2010		
10:30	Welcome	Dave Robertson	§1.1
10:35	Introduction	Malcolm Atkinson	§1.2
10:45	Strategies for exploiting large data	Alex Szalay	§1.3
11:30	Motivation and strategies for data-intensive biology	Douglas Kell	§1.4
12:15	Analysing and Modelling Large-Scale Enterprise Data	Thore Graepel	§1.5
13:30	Research Village		§1.6
15:45	Introduction to the Data-Analysis theme	Chris Williams	§1.7
16:05	Introduction to the Database Paradigms theme	Stratis Viglas	§1.8
16:25	Introduction to the Programming Paradigms theme	Geoffrey Fox	§1.9
16:45	Soaring through clouds with Meandre	Xavier Llorà & Bernie A'cs	§1.10
	Tuesday 16th March 2010		
09:00	Dealing with large data sets in astronomy and medical imaging (by throwing almost everything away)	Alan Heavens	§2.2
09:30	Data challenges in Earthquake Seismology	Torild van Eck	§2.3
10:00	Making the most of Earth-system data	Keith Haines	§2.4
11:00	Scientific Databases: the story behind the scene	Martin Kersten	§2.5
11:30	Model limitations: sequential data assimilation with uncertain static parameters	Jonty Rougier	§2.6
12:00	Earth-Systems data in real time applications: low latency, metadata, and preservation	Beth Plale	§2.7
12:30	Challenges in Large Scale GeoSpatial Data Analysis: Mapping, 3D and the GeoSpatial Web	Michael Batty	§2.8
17:00	Providing an environment where every data-driven researcher will thrive	Carole Goble	§2.9

continued on next page

Timetable (continued)

Time	**Title**	**Speaker**	**See**
	Wednesday 17th March 2010		
09:00	Big Data Bioinformatics	Mario Caccamo	§3.2
09:30	The Open Microscopy Environment: Informatics and Quantitative Analysis for Biological Microscopy, HCAs, and Image Data Repositories	Jason Swedlow	§3.3
10:00	Handling social science data: Challenges and responses	Paul Lambert	§3.4
11:00	The complexity dimension in data analysis	Chris Williams	§3.5
11:30	Spatial microsimulation for city modelling, social forecasting and urban policy analysis	Mark Birkin	§3.6
12:00	Linked Data: Making things more accessible	Hugh Glaser	§3.7
12:30	Using search for engineering diagnostics and prognostics	Jim Austin	§3.8
	Thursday 18th March 2010		
09:00	Medical image processing & caBIG	Joel Saltz	§4.2
09:30	Text mining, pathways and disease	Andrey Rzhetsky	§4.3
10:00	Building a virtual data center for the biological, ecological and environmental sciences	Bill Michener	§4.4
11:00	Curated databases	Peter Buneman	§4.5
11:30	Discovering Patterns in Text and Relational Data with Bayesian Latent-Variable Models	Andrew McCallum	§4.6
12:00	Using Real Time data to Understand and support Human Behaviour	Paul Watson	§4.7
17:30	Turing Award Lecture: *Embracing Uncertainty: The new machine intelligence*	Chris Bishop	§4.1.1
	Friday 19th March 2010		
09:00	Highlights from Monday's introduction	Malcolm Atkinson	§5.2.1
09:15	Highlights from Tuesday: A question of scale	Stratis Viglas	§5.2.2
09:30	Highlights from Wednesday: Complexity in the Big Data house	David De Roure	§5.2.3
09:45	Discussion of Data Analysis theme	Chris Williams	§5.2.5
10:00	Discussion of Programming Paradigms theme	Shantenu Jha	§5.2.7
10:45	Discussion of Database Paradigms theme	James Cheney	§5.2.6
11:00	Highlights from Thursday: Interacting with Data	Jano van Hemert	§5.2.4
11:15	A vision for Research in 2030	John Wood	§5.3

Table B.1: Summary Timetable of Data-Intensive Research

Bibliography

[1] Malcolm Atkinson and David De Roure. Realising the power of data-intensive research. Technical report, National e-Science Centre, 2010.

[2] Jeff Dozier and William B Gail. The emerging science of environmental applications. In Tony Hey, Stewart Tansley, and Kristin Tolle (Editors), editors, *The Fourth Paradigm: Data-Intensive Scientific Discovery*, pages 13–19. Microsoft, 2009.

[3] Christopher M Bishop. *Pattern Recognition and Machine Learning*. 2010.

[4] Helen M Berman. The Protein Data Bank: a historical perspective. *Acta Crystallographica Section A: Foundations of Crystallography*, A64(1):88–95, 2008.

[5] David Maier, Lois Delcambre, Calton Pu, and Jon Walpole. Database Research at the Data-Intensive Systems Center. *SIGMOD Record*, 22(4):81–86, 1993.

[6] HUGO. Summary of Principles Agreed at the First International Strategy Meeting on Human Genome Sequencing (Bermuda, 25-28 February 1996).

[7] HUGO. Summary of the Report of the Second International Strategy Meeting on Human Genome Sequencing (Bermuda, 27th February - 2nd March, 1997).

[8] Mark Guyer. Statement on the Rapid Release of Genomic DNA Sequence. *Genome Research*, 8(5):413–413, 1998.

[9] Alexander S. Szalay, Peter Z Kunszt, Aniruddha R Thakar, Jim Gray, and Don Slutz. The Sloan Digital Sky Survey and its Archive. In *Proceedings of the ADASS'99 conference*, 1999.

[10] Jacek Becla and Daniel L. Wang. Lessons learned from managing a petabyte. In *CIDR*, pages 70–83, 2005.

[11] Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, Funke R, Gage D, Harris K, Heaford A, Howland J, Kann L, Lehoczky J, LeVine R, McEwan P, McKernan K, Meldrim J, Mesirov JP, Miranda C, Morris W, Naylor J, Raymond C, Rosetti M, Santos R, Sheridan A, Sougnez C, Stange-Thomann N, Stojanovic N, Subramanian A, Wyman D, Rogers J, Sulston J, Ainscough R, Beck S, Bentley D, Burton J, Clee C, Carter N, Coulson A, Deadman R, Deloukas P, Dunham A, Dunham I, Durbin R, French L, Grafham D, Gregory S, Hubbard T, Humphray S, Hunt A, Jones M, Lloyd C, McMurray A, Matthews L, Mercer S, Milne S, Mullikin JC, Mungall A, Plumb R, Ross M, Shownkeen R, Sims S, Waterston RH, Wilson RK, Hillier LW, McPherson JD, Marra MA, Mardis ER, Fulton LA, Chinwalla AT, Pepin KH, Gish WR, Chissoe SL, Wendl MC, Delehaunty KD, Miner TL, Delehaunty A, Kramer

JB, Cook LL, Fulton RS, Johnson DL, Minx PJ, Clifton SW, Hawkins T, Branscomb E, Predki P, Richardson P, Wenning S, Slezak T, Doggett N, Cheng JF, Olsen A, Lucas S, Elkin C, Uberbacher E, Frazier M, Gibbs RA, Muzny DM, Scherer SE, Bouck JB, Sodergren EJ, Worley KC, Rives CM, Gorrell JH, Metzker ML, Naylor SL, Kucherlapati RS, Nelson DL, Weinstock GM, Sakaki Y, Fujiyama A, Hattori M, Yada T, Toyoda A, Itoh T, Kawagoe C, Watanabe H, Totoki Y, Taylor T, Weissenbach J, Heilig R, Saurin W, Artiguenave F, Brottier P, Bruls T, Pelletier E, Robert C, Wincker P, Smith DR, Doucette-Stamm L, Rubenfield M, Weinstock K, Lee HM, Dubois J, Rosenthal A, Platzer M, Nyakatura G, Taudien S, Rump A, Yang H, Yu J, Wang J, Huang G, Gu J, Hood L, Rowen L, Madan A, Qin S, Davis RW, Federspiel NA, Abola AP, Proctor MJ, Myers RM, Schmutz J, Dickson M, Grimwood J, Cox DR, Olson MV, Kaul R, Raymond C, Shimizu N, Kawasaki K, Minoshima S, Evans GA, Athanasiou M, Schultz R, Roe BA, Chen F, Pan H, Ramser J, Lehrach H, Reinhardt R, McCombie WR, de la Bastide M, Dedhia N, Blöcker H, Hornischer K, Nordsiek G, Agarwala R, Aravind L, Bailey JA, Bateman A, Batzoglou S, Birney E, Bork P, Brown DG, Burge CB, Cerutti L, Chen HC, Church D, Clamp M, Copley RR, Doerks T, Eddy SR, Eichler EE, Furey TS, Galagan J, Gilbert JG, Harmon C, Hayashizaki Y, Haussler D, Hermjakob H, Hokamp K, Jang W, Johnson LS, Jones TA, Kasif S, Kaspryzk A, Kennedy S, Kent WJ, Kitts P, Koonin EV, Korf I, Kulp D, Lancet D, Lowe TM, McLysaght A, Mikkelsen T, Moran JV, Mulder N, Pollara VJ, Ponting CP, Schuler G, Schultz J, Slater G, Smit AF, Stupka E, Szustakowki J, Thierry-Mieg D, Thierry-Mieg J, Wagner L, Wallis J, Wheeler R, Williams A, Wolf YI, Wolfe KH, Yang SP, Yeh RF, Collins F, Guyer MS, Peterson J, Felsenfeld A, Wetterstrand KA, Morgan MJ Patrinos A, de Jong P, Catanese JJ, Osoegawa K, Shizuya H, Choi S, and Chen YJ; International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, Feb 2001.

[12] Cech TR, Eddy SR, Eisenberg D, Hersey K, Holtzman SH, Poste GH, Raikhel NV, Scheller RH, Singer DB, and Waltham MC; National Academics Committee on Responsibilities of Authorship in the Biological Sciences. Sharing publication-related data and materials: responsibilities of authorship in the life sciences. *Plant Physiology*, 132(1):19–24, May 2003.

[13] Vik Singh, Jim Gray, Aniruddha R. Thakar, Alexander S. Szalay, Jordan Raddick, Bill Boroski, Svetlana Lebedeva, and Brian Yanny. SkyServer Traffic Report — The First Five Years. Technical Report MSR-TR-2006-190, Microsoft Research, December 2006.

[14] Jacek Becla and Kian-Tat Lim. Report from the first workshop on extremely large databases. Technical report, SLAC National Accelerator Laboratory, 2007.

[15] Jacek Becla and Kian-Tat Lim. Report from the 2nd workshop on extremely large databases. *Data Science Journal*, 7:196–208, 2008.

[16] Alexander S. Szalay, Gordon Bell, Jan vandenBerg, Alainna Wonders, Randal C. Burns, Dan Fay, Jim Heasley, Tony Hey, María A. Nieto-Santisteban, Aniruddha R Thakar, Catharine van Ingen, and Richard Wilton. Graywulf: Scalable clustered architecture for data intensive computing. In *HICSS* [107], pages 1–10.

[17] Yogesh Simmhan, Roger S. Barga, Catharine van Ingen, María A. Nieto-Santisteban, Laszlo Dobos, Nolan Li, Michael Shipway, Alexander S. Szalay, Sue Werner, and Jim Heasley. Graywulf: Scalable software architecture for data intensive computing. In *HICSS* [107], pages 1–10.

[18] Interagency Working Group on Digital Data. Harnessing the power of digital data for science and society: report to the committee on science of the national science and technology council. Technical report, Executive office of the President, Office of Science and Technology, Washington D.C. 20502 USA, January 2009.

[19] Barack Obama. Transparency and open Government. Memorandum for Executive Departments and Agencies, January 2009.

[20] Gordon. Bell, Tony. Hey, and Alexander S.. Szalay. Beyond the data deluge. *Science*, 323(5919):1297–1298, March 2009.

[21] Tim Berners-Lee. Putting government data online. Technical report, W3C, june 2009.

[22] Jacek Becla, Kian-Tat Lim, and Daniel Liwei Wang. Report from the third workshop on extremely large databases. *Data Science Journal*, 2009 (to appear).

[23] Toronto International Data Release Workshop Authors. Prepublication data sharing. *Nature*, 461:168–170, 09 2009.

[24] Kimmo Koski, Claudio Gheller, Stefan Heinzel, Alison Kennedy, Achim Streit, and Peter Wittenburg. Strategy for a European Data Infrastructure: White Paper. Technical report, Partnership for Advanced Data in Europe (PARADE), September 2009.

[25] Tony Hey, Stewart Tansley, and Kristin Tolle (Editors). *The Fourth Paradigm: Data-Intensive Scientific Discovery*. Microsoft, October 2009.

[26] e-IRG Data Management Task Force. Report on Data Management. Technical report, e-Infrastructure Reflection Group, November 2009.

[27] John F Gantz, Chute Christopher, Manfrediz Alex, Minton Stephen, Reinsel David, Schlichting Wolfgang, and Toncheva Anna. The diverse and exploding digital universe. Technical report, IDC, March 2008.

[28] R DelVecchio. UC Berkeley: Panel looks at control of emissions. S.F. Chronicle, 22 March 2007.

[29] David De Roure and Carole A. Goble. Software design for empowering scientists. *IEEE Software*, 26(1):88–95, 2009.

[30] Alexander S. Szalay. The sloan digital sky survey and beyond. *SIGMOD Rec.*, 37(2):61–66, 2008.

[31] Maria Nieto-Santisteban, Yogesh Simmhan, Roger Barga, Laszlo Dobos, Jim Heasley, Conrad Holmberg, Nolan Li, Michael Shipway, Alexander S. Szalay, Catharine van Ingen, and Sue Werner. Pan-STARRS: Learning to Ride the Data Tsunami. In *Proceedings of the Microsoft e-Science Workshop*. Microsoft Research, December 2008.

[32] E. Perlman, R. Burns, Y. Li, and C. Meneveau. Data Exploration of Turbulence Simulations using a Database Cluster. In *Supercomputing (SC07)*. ACM, IEEE, 2007.

[33] E. Perlman Y. Li, M. Wang, Y. Yang, C. Meneveau, R. Burns, S. Chen, A. Szalay, and G. Eyink. A Public Turbulence Database Cluster and Applications to Study Lagrangian Evolution of Velocity Increments in Turbulence. *Journal of Turbulence*, 9(31):1–29, 2008.

[34] Răzvan Musăloiu-E., Andreas Terzis, Katalin Szlavecz, Alex Szalay, Joshua Cogan, and Jim Gray. Life Under your Feet: A Wireless Soil Ecology Sensor Network. In *Proceedings*

of the Third Workshop on Embedded Networked Sensors (EmNets 2006), pages 30–31. Harvard University, Cambridge, Massachusetts, May 2006.

[35] Katalin Szlavecz, Andreas Terzis, Stuart Ozer, Razvan Musaloiu-Elefteri, Joshua Cogan, Sam Small, Randal C. Burns, Jim Gray, and Alexander S. Szalay. Life under your feet: An end-to-end soil ecology sensor network, database, web server, and analysis service. *CoRR*, abs/cs/0701170, 2007.

[36] Alexander S. Szalay, Gordon C. Bell, H. Howie Huang, Andreas Terzis, and Alainna White. Low-Power Amdahl-Balanced Blades for Data Intensive Computing. *ACM Operating Systems Review*, 2010.

[37] Douglas B. Kell and Steven Oliver. Here is the evidence, now what is the hypothesis? the complementary roles of inductive and hypothesis-driven science in the post-genomic era. *Bioessays*, 26:99–105, 2004.

[38] Douglas B. Kell. Metabolomics, modelling and machine learning in systems biology: towards an understanding of the languages of cells. The 2005 Theodor Bücher lecture. *FEBS J*, 273:873–894, 2006.

[39] Douglas B. Kell. Systems biology, metabolic modelling and metabolomics in drug discovery and development. *Drug Disc. Today*, 11:1085–1092, 2006.

[40] M.J. Herrgå rd, {32 others}, and Douglas B. Kell. A consensus yeast metabolic network obtained from a community approach to systems biology. *Nature Biotechnol.*, 26:1155–1160, 2008.

[41] D. Hull, S.R. Pettifer, and Douglas B. Kell. Defrosting the digital library: bibliographic tools for the next generation web. *PLoS Comput Biol*, 4(e1000204. doi:10.1371/journal.pcbi.1000204), 2008.

[42] Douglas B. Kell. Iron behaving badly: inappropriate iron chelation as a major contributor to the aetiology of vascular and other progressive inflammatory and degenerative diseases. *BMC Medical Genomics*, 2, 2009.

[43] P.D. Dobson and Douglas B. Kell. Carrier-mediated cellular uptake of pharmaceutical drugs: an exception or the rule? *Nat Rev Drug Discov*, 7:205–220, 2008.

[44] P.D. Dobson, K. Lanthaler, Steven Oliver, and Douglas B. Kell. Implications of the dominant role of cellular transporters in drug uptake. *Curr Top Med Chem*, 9:163–184, 2009.

[45] P.D. Dobson, Y Patel, and Douglas B. Kell. "metabolite-likeness" as a criterion in the design and selection of pharmaceutical drug libraries. *Drug Disc. Today*, 14:31–40, 2009.

[46] Douglas B. Kell and P.D. Dobson. The cellular uptake of pharmaceutical drugs is mainly carrier-mediated and is thus an issue not so much of biophysics but of systems biology. In *Proc Int Beilstein Symposium on Systems Chemistry*, number 149–168, 2009.

[47] Milena Ivanova, Martin L. Kersten, and Niels Nes. Adaptive segmentation for scientific databases. In *ICDE*, pages 1412–1414. IEEE, 2008.

[48] Milena Ivanova, Martin L. Kersten, Niels J. Nes, and Romulo Goncalves. An architecture for recycling intermediates in a column-store. In Ugur Çetintemel, Stanley B. Zdonik, Donald Kossmann, and Nesime Tatbul, editors, *SIGMOD Conference*, pages 309–320. ACM, 2009.

[49] Peter A. Boncz, Martin L. Kersten, and Stefan Manegold. Breaking the memory wall in monetdb. *Commun. ACM*, 51(12):77–85, 2008.

[50] S. Manegold, M. L. Kersten, and P. A. Boncz. Database Architecture Evolution: Mammals Flourished long before Dinosaurs became Extinct. In *Proceedings of the International Conference on Very Large Data Bases (VLDB)*, Lyon France, August 2009. 10-year Best Paper Award for Database Architecture Optimized for the New Bottleneck: Memory Access.

[51] Naisbett. *Megatrends*. Warner Books, 1982.

[52] L. Dagum, R. Menon, and S.G. Inc. OpenMP: an industry standard API for shared-memory programming. *IEEE Computational Science & Engineering*, 5(1):46–55, 1998.

[53] M. Snir and S. Otto. *MPI-The Complete Reference: The MPI Core*. MIT Press Cambridge, MA, USA, 1998.

[54] Nimbus Cloud Computing for Science http://www.nimbusproject.org/.

[55] OpenNebula Open Source Toolkit for Cloud Computing http://www.opennebula.org/.

[56] Sector and Sphere Data Intensive Cloud Computing Platform http://sector.sourceforge.net/doc.html.

[57] Eucalyptus Open Source Cloud Software http://open.eucalyptus.com/.

[58] FutureGrid Grid Testbed http://www.futuregrid.org.

[59] Magellan Cloud for Science http://magellan.alcf.anl.gov/ , http://www.nersc.gov/nusers/systems/magellan/.

[60] European Framework 7 project starting June 1 2010 VENUS-C Virtual multidisciplinary EnviroNments USing Cloud infrastructu re.

[61] Alan Heavens, Raul Jimenez, and Ofer Lahav. Massive Lossless Data Compression and Multiple Parameter Estimation from Galaxy Spectra. *Mon.Not.Roy.Astron.Soc.*, 317:965, 2000.

[62] Milena Ivanova, Martin L. Kersten, and Niels Nes. Self-organizing strategies for a column-store database. In Alfons Kemper, Patrick Valduriez, Noureddine Mouaddib, Jens Teubner, Mokrane Bouzeghoub, Volker Markl, Laurent Amsaleg, and Ioana Manolescu, editors, *EDBT*, volume 261 of *ACM International Conference Proceeding Series*, pages 157–168. ACM, 2008.

[63] Milena Ivanova, Niels Nes, Romulo Goncalves, and Martin L. Kersten. Monetdb/sql meets skyserver: the challenges of a scientific database. In *SSDBM*, page 13. IEEE Computer Society, 2007.

[64] M.A. Beaumont. Estimation of population growth or decline in genetically monitored populations. *Genetics*, 164:1139–1160, 2003.

[65] C. Andrieu and G.O. Roberts. The pseudo-marginal approach for efficient Monte Carlo computations. *Annals of Statistics*, 37(2):697–725, 2009.

[66] C. Andrieu, A. Doucet, and R. Holenstein. Particle Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society, Series B*, 72(2), 2010.

[67] Kelvin K. Droegemeier, K. Brewster, M. Xue, D. Weber, D. Gannon, B. Plale, D. Reed, L. Ramakrishnan, J. Alameda, R. Wilhelmson, T. Baltzer, B. Domenico, D. Murray,

M. Ramamurthy, A. Wilson, R. Clark, S. Yalda, S. Graves, R. Ramachandra, J. Rushing, E. Joseph, and V. Morris. Service-oriented environments for dynamically interacting with mesoscale weather. *Computing in Science and Engineering*, 7:12–27, 2005.

[68] C. Herath and Beth Plale. Streamflow — Programming Model for Data Streaming in Scientific Workflows. In *Proceedings of CCGrid, Melbourne, Australia*, May 2010.

[69] Lavanya Ramakrishnan and Beth Plale. A Multi-Dimensional Classification Model for Scientific Workflow Characteristics. *under review*, 2010.

[70] Beth Plale, You-Wei Cheah, and Yiming Sun. Towards Quantification of Limits in Automated Curation of e-Science Data. Poster at Microsoft e-Science Conference, December 2008.

[71] Stacey Kowalczyk. Towards a Preservable Object. A Qualifying Examination, Department of Computing Science, University of Indiana, October 2007.

[72] Yiming Sun, You-Wei Cheah, and Beth Plale. Provenance for Preservation. under review, 2010.

[73] J. Craig Venter, Mark D. Adams, Eugene W. Myers, Peter W. Li, Richard J. Mural, Granger G. Sutton, Hamilton O. Smith, Mark Yandell, Cheryl A. Evans, Robert A. Holt, Jeannine D. Gocayne, Peter Amanatides, Richard M. Ballew, Daniel H. Huson, Jennifer Russo Wortman, Qing Zhang, Chinnappa D. Kodira, Xiangqun H. Zheng, Lin Chen, Marian Skupski, Gangadharan Subramanian, Paul D. Thomas, Jinghui Zhang, George L. Gabor Miklos, Catherine Nelson, Samuel Broder, Andrew G. Clark, Joe Nadeau, Victor A. McKusick, Norton Zinder, Arnold J. Levine, Richard J. Roberts, Mel Simon, Carolyn Slayman, Michael Hunkapiller, Randall Bolanos, Arthur Delcher, Ian Dew, Daniel Fasulo, Michael Flanigan, Liliana Florea, Aaron Halpern, Sridhar Hannenhalli, Saul Kravitz, Samuel Levy, Clark Mobarry, Knut Reinert, Karin Remington, Jane Abu-Threideh, Ellen Beasley, Kendra Biddick, Vivien Bonazzi, Rhonda Brandon, Michele Cargill, Ishwar Chandramouliswaran, Rosane Charlab, Kabir Chaturvedi, Zuoming Deng, Valentina Di Francesco, Patrick Dunn, Karen Eilbeck, Carlos Evangelista, Andrei E. Gabrielian, Weiniu Gan, Wangmao Ge, Fangcheng Gong, Zhiping Gu, Ping Guan, Thomas J. Heiman, Maureen E. Higgins, Rui-Ru Ji, Zhaoxi Ke, Karen A. Ketchum, Zhongwu Lai, Yiding Lei, Zhenya Li, Jiayin Li, Yong Liang, Xiaoying Lin, Fu Lu, Gennady V. Merkulov, Natalia Milshina, Helen M. Moore, Ashwinikumar K Naik, Vaibhav A. Narayan, Beena Neelam, Deborah Nusskern, Douglas B. Rusch, Steven Salzberg, Wei Shao, Bixiong Shue, Jingtao Sun, Zhen Yuan Wang, Aihui Wang, Xin Wang, Jian Wang, Ming-Hui Wei, Ron Wides, Chunlin Xiao, Chunhua Yan, Alison Yao, Jane Ye, Ming Zhan, Weiqing Zhang, Hongyu Zhang, Qi Zhao, Liansheng Zheng, Fei Zhong, Wenyan Zhong, Shiaoping C. Zhu, Shaying Zhao, Dennis Gilbert, Suzanna Baumhueter, Gene Spier, Christine Carter, Anibal Cravchik, Trevor Woodage, Feroze Ali, Huijin An, Aderonke Awe, Danita Baldwin, Holly Baden, Mary Barnstead, Ian Barrow, Karen Beeson, Dana Busam, Amy Carver, Angela Center, Ming Lai Cheng, Liz Curry, Steve Danaher, Lionel Davenport, Raymond Desilets, Susanne Dietz, Kristina Dodson, Lisa Doup, Steven Ferriera, Neha Garg, Andres Gluecksmann, Brit Hart, Jason Haynes, Charle Haynes, Cheryl Heiner, Suzanne Hladun, Damon Hostin, Jarrett Houck, Timothy Howland, Chinyere Ibegwam, Jeffery Johnson, Francis Kalush, Lesley Kline, Shashi Koduru, Amy Love, Felecia Mann, David May, Steven McCawley, Tina McIntosh, Ivy McMullen, Mee Moy, Linda Moy, Brian Murphy, Keith Nelson, Cynthia Pfannkoch, Eric Pratts, Vinita Puri, Hina Qureshi, Matthew Reardon, Robert Rodriguez, Yu-Hui Rogers, Deanna Romblad, Bob Ruhfel, Richard Scott, Cynthia Sitter, Michelle

Smallwood, Erin Stewart, Renee Strong, Ellen Suh, Reginald Thomas, Ni Ni Tint, Sukyee Tse, Claire Vech, Gary Wang, Jeremy Wetter, Sherita Williams, Monica Williams, Sandra Windsor, Emily Winn-Deen, Keriellen Wolfe, Jayshree Zaveri, Karena Zaveri, Josep F. Abril, Roderic Guigó, Michael J. Campbell, Kimmen V. Sjolander, Brian Karlak, Anish Kejariwal, Huaiyu Mi, Betty Lazareva, Thomas Hatton, Apurva Narechania, Karen Diemer, Anushya Muruganujan, Nan Guo, Shinji Sato, Vineet Bafna, Sorin Istrail, Ross Lippert, Russell Schwartz, Brian Walenz, Shibu Yooseph, David Allen, Anand Basu, James Baxendale, Louis Blick, Marcelo Caminha, John Carnes-Stine, Parris Caulk, Yen-Hui Chiang, My Coyne, Carl Dahlke, Anne Deslattes Mays, Maria Dombroski, Michael Donnelly, Dale Ely, Shiva Esparham, Carl Fosler, Harold Gire, Stephen Glanowski, Kenneth Glasser, Anna Glodek, Mark Gorokhov, Ken Graham, Barry Gropman, Michael Harris, Jeremy Heil, Scott Henderson, Jeffrey Hoover, Donald Jennings, Catherine Jordan, James Jordan, John Kasha, Leonid Kagan, Cheryl Kraft, Alexander Levitsky, Mark Lewis, Xiangjun Liu, John Lopez, Daniel Ma, William Majoros, Joe McDaniel, Sean Murphy, Matthew Newman, Trung Nguyen, Ngoc Nguyen, Marc Nodell, Sue Pan, Jim Peck, Marshall Peterson, William Rowe, Robert Sanders, John Scott, Michael Simpson, Thomas Smith, Arlan Sprague, Timothy Stockwell, Russell Turner, Eli Venter, Mei Wang, Meiyuan Wen, David Wu, Mitchell Wu, Ashley Xia, Ali Zandieh, and Xiaohong Zhu. Human Genome Special Issue. *Science*, 291:1145–1434, 2001.

[74] Collins F.S., Green E.D., Guttmacher A.E., and Guyer M.S. A vision for the future of genomics research. *Nature*, 431:835–847, 2003.

[75] John McPherson. Next-Generation Gap. *Nature Methods*, 6:6–11, 2009.

[76] Mark Birkin, P. Townend, A. Turner, B. Wu, and Jie Xu. MoSeS: A Grid-enabled spatial decision support system. *Social Science Computing Review*, 27:493–508, 2009.

[77] B. Wu, Mark Birkin, and P. Rees. A spatial microsimulation model with student agents. *Computers Environment and Urban Systems*, 32:440–453, 2008.

[78] D Ballas and G Clarke. Spatial Microsimulation. In AS Fotheringham and P Rogerson, editor, *The Sage Handbook of Spatial Analysis*. Sage, 2009.

[79] JM Epstein and R Axtell. Growing artificial societies from the bottom up. Technical report, The Brookings Institution, 1996.

[80] N Gilbert and K Troitzsch. *Simulation for the Social Scientist, Second Edition*. Open University Press, 2005.

[81] S. Eubank, H. Guclu, V. S. A. Kumar, M. V. Marathe, A. Srinivasan, Z. Toroczkai, and N. Wang. Modeling Disease Outbreaks in Realistic Urban Social Networks. *Nature*, 429:180, 2004.

[82] N. M. Ferguson, A. T. Cummings, S. Cauchemez, C. Fraser, S. Riley, A. Meeyai, and *et al.* Strategies for containing an emerging influenza pandemic in Southeast Asia. *Nature*, 427:7056, 2005.

[83] JM Epstein. Modelling to contain pandemics. *Nature*, 460:687, 2009.

[84] M Savage and R Burrows. The Coming Crisis of Empirical Sociology. *Sociology*, 41:885–899, 2007.

[85] Mark Birkin, Rob Procter, Rob Allan, S Bechhofer, I Buchan, Carole A. Goble, A Hudson-Smith, Paul Lambert, David De Roure, and Richard Sinnott. The Elements of a Computa-

tional Infrastructure for Social Simulation. *Philisophical Transactions of the Royal Society A*, to appear 2010.

[86] Vijay S. Kumar, P. Sadayappan, Gaurang Mehta, Karan Vahi, Ewa Deelman, Varun Ratnakar, Yolanda Gil Jihie Kim, Mary W. Hall, Tahsin M. Kurç, and Joel H. Saltz. An integrated framework for performance-based optimization of scientific workflows. In *High-Performance and Distributed Computing*, pages 177–186, 2009.

[87] Vijay S. Kumar, Sivaramakrishnan Narayanan, Tahsin M. Kurç, Jun Kong, Metin N. Gurcan, and Joel H. Saltz. Analysis and semantic querying in large biomedical image datasets. *IEEE Computer*, 41(4):52–59, 2008.

[88] Joel H. Saltz, Tahsin M. Kurç, Shannon Hastings, Stephen Langella, Scott Oster, David Ervin, Ashish Sharma, Tony Pan, Metin N. Gurcan, Justin Permar, Renato Ferreira, Philip R. O. Payne, Ümit V. Çatalyürek, Enrico Caserta, Gustavo Leone, Michael C. Ostrowski, Ravi K. Madduri, Ian T. Foster, Subhasree Madhavan, Kenneth H. Buetow, Krishnakant Shanbhag, and Eliot L. Siegel. e-science, cagrid, and translational biomedical research. *IEEE Computer*, 41(11):58–66, 2009.

[89] Oster S, Langella S, Hastings S, Ervin D, Madduri R, Phillips J, Kurc T, Siebenlist F, Covitz P, Shanbhag K, Foster I, and Saltz J. cagrid 1.0: an enterprise grid infrastructure for biomedical research. *J Am Med Inform Assoc.*, 15(2):138–49, 2008.

[90] Andrey Rzhetsky, Ivan Iossifov, Tomohiro Koike, Michael Krauthammer, Pauline Kra, Mitzi Morris, Hong Yu, Pablo Ariel Duboué, Wubin Weng, and W. John Wilbur. Geneways: a system for extracting, analyzing, visualizing, and integrating molecular pathway data. *Journal of Biomedical Informatics*, 37(1):43–53, 2004.

[91] Hagit Shatkay, Fengxia Pan, Andrey Rzhetsky, and W. John Wilbur. Multi-dimensional classification of biomedical text: Toward automated, practical provision of high-utility text to diverse users. *Bioinformatics*, 24(18):2086–2093, 2008.

[92] Lixia Yao and Andrey Rzhetsky. Quantitative systems-level determinants of drug targets. *BMC Bioinformatics*, 9(S-10), 2008.

[93] Liu J, Ghanim M, Xue L, Brown CD, Iossifov I, Angeletti C, Hua S, Negre N, Ludwig M, Stricker T, Al-Ahmadie HA, Tretiakova M, Camp RL, Perera-Alberto M, Rimm DL, Xu T, Rzhetsky A, and White KP. Analysis of Drosophila segmentation network identifies a JNK pathway factor overexpressed in kidney cancer. *Science*, 323(5918):1218–1222, 2009.

[94] Ivan Iossifov, Tian Zheng, Miron Baron, T.Conrad Gilliam, and Andrey Rzhetsky. Genetic-linkage mapping of complex hereditary disorders to a whole-genome molecular-interaction network. *Genome Research*, 18:1150–1162, 2009.

[95] Andrey. Rzhetsky, Michael. Seringhaus, and Mark. Gerstein. Seeking a new biology through text mining. *Cell*, 134(1):9–13, 2008.

[96] William K Michener, James H Beach, Matthew B Jones, Bertram Ludäscher, Deana D Pennington, Ricardo S Pereira, Arcot Rajasekar, and Mark Schildhauer. A knowledge environment for the biodiversity and ecological sciences. *Intelligent Information Systems*, 29:111–126, 2007.

[97] C.A. Morrison, N. Robertson, A. Turner, J. van Hemert, and J. Koetsier. Molecular orbital calculations of inorganic compounds. In *Inorganic Experiments*, pages 261–267. Wiley-VCH, 2010.

[98] J. Koetsier, A. Turner, P. Richardson, and J.I. van Hemert. Rapid chemistry portals through engaging researchers. In A Trefethen and D De Roure, editors, *Fifth IEEE International Conference on e-Science*, pages 284–291, 2009.

[99] J.I. van Hemert and J. Koetsier. Giving computational science a friendly face. *Zero-In*, 1(3):12–13, 2009.

[100] J. Koetsier and J.I. van Hemert. Rapid development of computational science portals. In S. Gesing and J.I. van Hemert, editors, *Proceedings of the IWPLS09 International Workshop on Portals for Life Sciences*, CEUR Workshop Proceedings, Edinburgh, September 2009.

[101] Interpreting the data: Parallel analysis with Sawzall, Rob Pike, Sean Dorward, Robert Griesemer, Sean Quinlan, Volume 13, Issue 4, Scientific Programming.

[102] Pig Latin: A Not-So-Foreign Language for Data Processing C. Olston, B. Reed, U. Srivastava, R. Kumar and A. Tomkins. ACM SIGMOD 2008 International Conference on Management of Data, Vancouver, Canada, June 2008.

[103] Dryad: Distributed Data-Parallel Programs from Sequential Building Blocks Michael Isard, Mihai Budiu, Yuan Yu, Andrew Birrell, and Dennis Fetterly European Conference on Computer Systems (EuroSys), Lisbon, Portugal, March 21-23, 2007.

[104] Thilina Gunarathne, Tak-Lon Wu, Judy Qiu, and Geoffrey Fox. Cloud Computing Paradigms for Pleasingly Parallel Biomedical Applications. In *Proceedings of Emerging Computational Methods for the Life Sciences Workshop of ACM HPDC 2010 conference*, 2010.

[105] C. Christensen, T. Aina, and D. Stainforth. The challenge of volunteer computing with lengthy climate model simulations. In *e-Science and Grid Computing, 2005. First International Conference on*, page 8, 2005.

[106] D.J. Goodman. Introduction and evaluation of martlet: A scientific workflow language for abstracted parallelisation. In *Proceedings of the 16th international conference on World Wide Web*, page 992. ACM, 2007.

[107] *42st Hawaii International International Conference on Systems Science (HICSS-42 2009), Proceedings (CD-ROM and online), 5-8 January 2009, Waikoloa, Big Island, HI, USA*. IEEE Computer Society, 2009.