Data-Intensive Research: making best use of research data (Draft 1)

Malcolm Atkinson[1] David De Roure[2]

[1] School of Informatics, University of Edinburgh, UK
[2] School of Electronics and Computer Science, University of Southampton, UK

14 December 2009

Foreword

a

Need bullet points for DK

Preface

This report focuses on the dramatic change in the ways in which research is undertaken, christened, "*The Fourth Paradigm*" by Jim Gray, that is sweeping the world. This is driven by and responding to an even greater global '*Digital Revolution*' that is overtaking business, government and society. The report calls for action now to grasp the new research opportunities and to address the new research challenges.

Box 1

Researchers will gain from rebalancing their existing efforts to increase their ability to make discoveries latent in data. Universities and research centres will benefit from equipping their institutions with skills and facilities to exploit the new opportunities. By investing in data use, research-funding organisations will accelerate intellectual progress and increase the value of their commitments to data generation, collection, curation and preservation; they will achieve a high return on investment. Many of today's research priorities are 'application science', with the goal of generating knowledge in time to avert disaster; this is particularly dependent on adroit use of existing data. Educators should equip graduates to thrive in a data-rich world; this will benefit their careers and employers.

The report is complementary to reports, such as the ESFRI Roadmap, that advocate investment in new methods of obtaining, generating or federating data. It complements the work of providers of research data, such as the European Bioinformatics Institute and the British Atmospheric Data Centre, and of digital librarians and the Digital Curation Centre.

This report is intended to initiate action and provoke discussion of how best to exploit the opportunities inherent in the Fourth Paradigm. Pursuing its goals will lead to evidence, discoveries and knowledge from the wealth of data that flow from the digital revolution and deliver new capabilities that will be applied throughout business, government and society. Increased use of data will be a persistent feature of future research, business, government and decision making.

Our fact-finding mission to the USA, described in appendices B to D, revealed to us many examples of routine and long-term use of data that would not be detailed in papers, because they are part and parcel of researchers' routine work. It also revealed many exciting innovations, which we capture in terms of three metaphors: 'datascopes' scanning data to reveal new insights, 'intellectual ramps' to enable researchers to incrementally adopt new methods of data analysis and 'going the last mile' to persuade people to act on the evidence in data.

The first part of the report, up to chapter 4, is intended to inform decision makers.
The *Executive summary* on pages ix and x summarises the principal messages.
Chapter 1 *Introduction* motivates and outlines the argument for improving data use.
Chapter 2 *Principles* lists six principles to inform decisions as the 'digital ecosystem' evolves.
Chapter 3 *Recommendations* delivers five high-level recommendations to improve data use.
Chapter 4 *Users* examines the behaviour and requirements of researchers using data.

The second part of the report, chapters 5 to 10, explores the issues and strategies for advancing data-intensive research in more detail. Each of its chapters begins with an accessible presentation and concludes with more technical information including references to sources of further ideas.
Chapter 5 *Digital-data ecosystem* considers research's role within the larger corporate domain.
Chapter 6 *Datascopes* sets directions for better tools to extract understanding from data.
Chapter 7 *Intellectual ramps* discusses how to make it easier for researchers to use data.
Chapter 8 *Solutions* proposes routes to sustainable facilities for data-intensive research.

MP Atkinson & D De Roure

Draft 1: 14 December 2009

Chapter 9 *Categories of data use* reviews the major variations in data use.
Chapter 10 *Actions* suggests a selection of actions relevant for UK institutions.

This document is not a prescription for data-intensive research; rather it is a provocation to discussion as well as initial action. Anchor points of the form, letter followed by digit, e.g. **R3**, are provided so that participants in that discussion can reference the focus of their comments (see appendix A for an index of these discussion points). We very much hope that a vigourous discussion will ensue.

We have been helped on this journey by a great many colleagues and much prior work, particularly our many kind hosts during our fact-finding mission in the USA. These are acknowledged in Appendix D. If we have omitted anyone, we apologise and will remedy our error in the next version; please prompt us!

Most of the good ideas in this report have arisen from the conversations with helpful colleagues, but we have synthesised these into a report, no doubt introducing errors and losing gems. The principles, recommendations and suggested actions in this report are wholly our own responsibility and they should not be attributed to those organisations that funded or hosted us.

Malcolm Atkinson and Dave De Roure

Notes for: The Four Paradigms

1) Thousands of years of *empirical* science involved observing things and testing ideas like Archimedes displacing water to measure volume [1, 2]. 2) For four hundred years *theoretical* science has formulated (mathematical) models that are abstractions and generalisations of observed phenomena and tested these against observation and experiment like James Clerk Maxwell describing light as electro-magnetic waves. 3) For 70 years *computational* science has described models using programs that may then be explored by simulation and tested against observation and experiment like Denis Noble modelling ion channels in muscle cells to understand the origin of heart beats [3, 4]. 4) For a decade or more *data-intensive* science has explored phenomena by analysing properties of data collected by observing those phenomena in the natural universe or in the laboratory like the Ventner and human genome project assembling shotgun sequence fragments.

Box 1: The Four Paradigms

Contents

Executive Summary

Today's challenges are both urgent and intellectually demanding; making the best use of the world's growing wealth of data is a crucial strategy for addressing them. Data are the catalysts in research, engineering and diagnosis. Data fuel analysis to produce key evidence for decisions and supply the information for compelling communication. Data connect computational systems and enable global collaborative endeavours.

The digital revolution is transforming global economies and societies with ever increasing flows of data, and a flood of faster, cheaper and higher-resolution digital devices. Research is being accelerated and enabled by the advances in automation, communication, sensing and computation. To reap the benefits, researchers need infrastructure and tools that are as convenient, pervasive and powerful as the Web 2.0 environment to enable them to combine and analyse enormous volumes of data from a wide range of sources. The questions they ask and the capabilities of these facilities will co-evolve as the new power stimulates new research strategies.

This will require radical changes in research behaviour and infrastructure provision. We advocate a collaborative endeavour to achieve this potential, exploring the new behaviours, methods, computational strategies and economies of provision.

Data should be used fluently in research, investigation, planning and policy formulation to equip those responsible with the necessary information, knowledge and wisdom. *The present cornucopia of data is under-exploited.* Effort and resources should be rallied to a *data-use initiative* that delivers a sustainable step change in our ability to harness the potential of data.

The data-use initiative will:
accelerate research and improve the quality of decisions, by
enabling *all* researchers to analyse data with improved methods
to 'go the last mile' to apply science and achieve influence.

An over-riding requirement is *to engage the next generation of talented minds in the creative processes of distilling information from data, establishing knowledge and developing wisdom.* This should rapidly lift our capacity to use data effectively and permanently transform the way research is undertaken. Better understanding of data use will inform data collection.

We propose the following framework for discussing the data-intensive research initiative.

Principles

P1 Support for research data should be in harmony with the evolving digital ecosystem.
P2 Attention to the analysis of data should be increased in order to make better use of data.
P3 Co-evolve research practices with new methods and their supporting software.

P4 Democratise research by improving education and access to data-intensive methods.
P5 Smooth the path from theoretical research through proof of concept to sustained use.
P6 The costs of data-intensive research should be visible to researchers.

Recommendations

R1 Stimulate new thinking in the next generation of researchers.
R2 Invest in creating and sharing methods for exploiting data.
R3 Increase data use by building 'intellectual ramps' and providing education.
R4 Conduct research to improve data-intensive research methods and their implementation.
R5 Align method creation with provision of data-intensive computational infrastructure.

Actions

A step change in our skills and capacity for data-intensive research is necessary to address current strategic challenges: (*a*) a resilient economy, (*b*) green technology and energy, (*c*) health and wellbeing in an ageing society, (*d*) living with environmental change, (*e*) global uncertainties and security, and (*f*) food security. Every one of these has to deal with growing quantities of data and take into account more interactions and complexities in order to assess progress, steer decisions, design solutions and optimise operations.

There will be a rapid growth and diversification of application sciences where data have to be analysed rapidly to support decisions. This will accelerate demand for data-intensive capabilities. A programme of actions to develop the capacity and leadership in addressing the research priorities will include:

A1 Rally the research community to action.
A2 Activate a data-intensive educational programme.
A3 Launch new research via an 'ideas factory'.
A4 Engage with and evaluate existing best practice.
A5 Stimulate researchers with immediate challenges.
A6 Establish data-intensive research facilities.
A7 Boost the facilities and capacity of the UK's reference data services.
A8 Initiate a programme of research from data-intensive foundations to applications.
A9 Develop greener data-intensive computation.
A10 Establish a framework for shared facilities, interdisciplinary data integration and maximal return on investment.

The data-use initiative will change the mind set of researchers and research users, and allocate a greater proportion of research effort and investment to using data than hitherto. There are already large quantities of under-used data. The available data are growing rapidly through research investment and as the by-product of many other activities. *A modest change in relative priorities will yield significant dividends.* Increased understanding of how to exploit data will lead to better informed data collection and preservation.

Survival in the digital revolution depends on rapid and appropriate adaptation.

There are many global and local challenges that will overwhelm society unless we improve the quality of our decisions. *Key to this is making the best use of available data.*

Chapter 1

Introduction

The world is in the early stages of a digital revolution; more stressful than the industrial revolution, as it is impacting virtually every nation simultaneously. Global access to data is changing the ways in which we think and behave. This is seeding change in global collaborations and businesses powered by shared data. Mastery in the exploitation of data is key to success.

Data are any digitally encoded information that can be stored, processed and transmitted by computers. They include:

- collections of data from instruments, observatories, surveys and simulations;
- results from previous research and earlier surveys;
- data from engineering and built-environment design, planning and production processes;
- data from diagnostic, laboratory, personal and mobile devices;
- streams of data from sensors in the built and natural environment,
- data from monitoring digital communications;
- the data transferred during the transactions that enable business, administration, healthcare and government;
- digital material produced by news feeds, publishing, broadcasting and entertainment;
- documents in collections and held privately, the texts and multi-media 'images' in web pages, wikis, blogs, emails and tweets; and
- digitised representations of diverse collections of objects, e.g. of museums' curated objects.

Data's accessibility may be controlled or open.

The digital revolution is transforming global economies and societies with ever increasing flows of data, and a flood of faster, cheaper and higher-resolution digital devices. Research is being accelerated and enabled by the advances in automation, communication, sensing and computation. To reap the benefits, researchers need infrastructure and tools that are as convenient, pervasive and powerful as the Web 2.0 environment to enable them to combine and analyse enormous volumes of data from a wide range of sources. The questions they ask and the capabilities of these facilities will co-evolve as the new power stimulates new research strategies.

This will require significant changes in research behaviour and infrastructure provision as illustrated by the following example. The growing collections of data, e.g. *a*) the streams of environmental-sensor data from a fully instrumented field, *b*) the set of all gene sequences for all organisms with their annotations, *c*) spatio-temporal images of biological systems from within the cell, via organs and individuals, to ecosystems, and *d*) meteorological records.

Scientists can, in principle, ask any questions of these data, for example one group might ask,

- Get the sequence data for gene X (including variations) for all organisms (implicitly using an ontology to deal with X's name variation).
- Prepare a 3D time sequence of the expression of genes X, Y and Z during the development of a *Drosophila melanogaster* brain.
- Partition the cohort for which we have fMRI scans according to which allele of gene X they have and synthesise a 3D map showing an aggregation of scans for each subgroup.

While another group of researchers might ask,

- Get the metagenomic data of soil samples from the patches of the field, which are in the top 5% for diurnal variability in CO_2 transport.
- For these patches, show how CO_2 transport varies as a consequence of recent rainfall.
- Get the aggregated data from automated laboratories of CO_2 transpiration of the plant species in each patch.
- Prepare map overlays showing soil properties and summaries of the last three questions.
- Prepare a national map showing the agronomy and population data for regions that are potentially similar to the selected patches studied above.

In practice today they can't ask such questions because the data are organised for one class of questions and each collection is separate. The advent of data-intensive computational strategies (discussed later), means that a much-larger class of questions about large or complex data can be answered economically. Today several large instruments, observatories and reference data collections are each beginning to use computational systems optimised for data-intensive operations. This will become a dominant data management strategy over the next decade.

We propose a path intended to lead to more collaborative behaviour where many data producing organisation would deposit data in '*shared data clouds*' that would also be populated with reference and legacy data. Researchers would then deposit data and pose questions that composed and compared data from any contributing source. Having a data cloud shared by environmental and earth-systems scientists, by economic and social scientists, by medical researchers and diagnosticians, by engineers and physical scientists and by biological and life scientists would facilitate the interdisciplinary compositions of data and research questions that would be infeasible otherwise.

The data-intensive cloud computing experts and scientific database experts have, in the last decade, become adept at handling more and more forms of computational query/analysis by using carefully chosen computing hardware and software architectures.

We propose that researchers embrace the opportunity brought about by these and other technical advances to blaze a path towards the vision of easily accessed and composed research data that they can fluently exploit for their fundamental research and society's urgent research applications. This will involve the development of new behaviours, frameworks, methods and computational systems. Each step will deliver new capabilities, boosting research and informing decisions. A collaborative approach is key to leadership in interdisciplinary research, to assembling and building skills, and to achieving the economies demonstrated by Web 2.0 companies.

The growing and diverse cornucopia of data presents many new opportunities. All research disciplines will benefit from improved use of data [5]. Bell *et al.* identify "data-intensive" as a new research paradigm [6] christened "*The Fourth Paradigm*" by Jim Gray [7]. International programmes, such as the Inter-governmental Panel on Climate Change (IPCC) [8] and the European Strategic Forum on Research Infrastructure (ESFRI) [9] propose the collection, generation and sharing of even larger reservoirs of data. The Web 2.0 business model recognises the key role of data [10]. Governments recognise the opportunity, as we see from the European INSPIRE directive [11], and recent announcements from the USA and UK governments [12, 13, 14] — see Table 1.1 for further examples. This growing wealth of data is a key resource for making better informed decisions, from the local and personal to global policies that shape the future.

Conversely, many decisions are made without using relevant data that is already available[1]. Building continues on coasts that will flood frequently as sea level rises and as storms have more energy. Aquifers are being polluted or depleted despite data on their limits. Energy production using fossil fuels continues to rise despite compelling evidence of the long-term danger to mankind [15]. Many *avoidable* errors occur daily, for example an excavator cutting a cable, a drill hitting a pipe, a misdiagnosis and incorrect responses to an emergency. *Avoiding such errors by better use of existing data would yield significant financial savings, reduce distress and save lives.*

Modest changes in the distribution of resources would leverage very large benefits.

This report launches a *data-use initiative* to increase rapidly the effective use of data in research. It will devote effort to the *full-path* of data use by rebalancing attention and investment. It emphasises '*going the last mile*' with data; not only uncovering the evidence that will support a scholarly observation, but also distilling and collating the evidence into forms that will be routinely used in research and decision making, so that the data ultimately increase the wisdom behind policy and action. Dozier and Gail introduce the term "*application science*" to denote this use of research to help people take appropriate action [16]. In their example, the fresh water supply for California is threatened by continuation of river management policies that have not adapted to changes induced by global warming. They use satellite and field data to measure the water stored as snow so that the changes in timing and volume of flow can be understood. They give other examples of failures to use environmental data; more evidence that a step change is needed. Application science is even more demanding on data-intensive technologies, as urgency forces the use of imperfect and incomplete data.

There are many beacons of good practice, but they are relatively rare and only a few cases have long-term support. The data-use initiative is proposed so that the *majority* of researchers in *most* disciplines benefit from the wealth of data. It will transform forever the way research is conducted, accelerating discovery, increasing the value of research decision making and seeding changes throughout the economy and society that are of value to all citizens.

The greatest change will be the increase in the numbers of people accessing data; this should be encouraged, e.g. engaging '*citizen scientists*' in collecting and annotating data[2]. When data are familiar the use of data is more like to figure in problem-solving strategies. When the data, services and tools the researchers need are well honed they will be used routinely.

The dynamics of human response to opportunities are hard to predict, especially as businesses, consumers' attitudes and technology are all changing rapidly. The data-use initiative must therefore be agile and must facilitate *co-evolution* of research thinking, methods, tools and services. Success depends on sustained and close working relationships between researchers addressing

[1] As D. Kell, writing about biology research, states, "Three months in the lab can save a whole afternoon on the computer", blogs.bbsrc.ac.uk/index.php/2009/02/the-miners-strike-again

[2] Public interest in classifying galaxies in the Sloan Digital Sky Survey is a good example www.galaxyzoo.org.

their domain's challenges and innovators of new methods (see chapter 4). Their research practices will need to evolve in harmony with the larger digital ecosystem (see chapter 5).

New methods and tools are needed to make more effective use of data; we call these '*datascopes*' as they reveal previously hidden evidence to the 'naked' mind just as telescopes reveal the universe to the naked eye (see chapter 6). Building datascopes requires close collaboration between those who have intuitions about what might be hidden in a body of data with those with the skills to create new analytic methods.

Researchers are fully stretched understanding the progress and addressing the challenges in their own field. They are naturally focused and competitive as well as collaborative (see chapter 4). Researchers are therefore reluctant to be diverted into understanding a new method, particularly if they are unsure of its benefits. Datascopes become complex and harder to approach, as they evolve to have the power that satisfies their experienced users. Consequently, the data-use initiative must invest in '*intellectual ramps*' that allow researchers to engage incrementally, learning to do simple things easily and progressing to more sophisticated use when they need to (see chapter 7). This requires collaboration with new as well as experienced users, study of research practices and underpinning research into ramp design. Researchers may need to be induced to try a ramp in cases where benefits are deferred, e.g. until enough researchers adopt the practice or when benefits are conferred on the larger community.

Successful datascopes and ramps will lead to services that support their continued use. These will emerge from competition in the data ecosystem. It typically takes ten years from the initial effort to pioneer new research software to their emergence as a widely usable and reliable platforms that researchers may depend on. The data-use initiative should facilitate the transition from proof-of-concept demonstration to dependable research facility (see chapter 8).

'Going the final mile' frequently involves transforming results into carefully crafted forms so that a target community can *act on them* [17]. This requires understanding how the target community thinks and acts. A goal of the data-use initiative is: *to increase the impact of research on the quality of decisions, by using much improved methods to achieve influence.*

The data-use initiative will stimulate a step change in the way that research is undertaken. The increased attention to the use of data is already sweeping the research world. Those that engage in the initiative will be shaping the new mores and ethics of data-intensive research, as well as inventing new research methods and delivering influential information. It will trigger an explosion of new algorithms, software, tools, services and platforms to enable data-intensive research. Transformed curricula[3], 'intellectual ramps' and supported services will make it much easier for researchers to reap the potential of data. The initiative is necessaryto make the step change significant and to ensure that all aspects of the change are considered.

This report proposes principles and recommendations, analyses requirements and actions (chapters 2, 3, 9 and 10 respectively); these will need to be developed and revised regularly.

The data-use initiative will transform the ways that research is done. It will empower the next generation of researchers with skills, methods and tools that will accelerate their progress in extracting information, distilling knowledge and building wisdom from the ever growing wealth of data. With their leadership and these capabilities they will address hard intellectual problems and pressing global challenges with new vigour.

[3] To develop *computational thinking* [18, 19, 20] and expertise in using data.

Date	Event	Reference
1971	Transatlantic agreement to share the Protein Data Bank (PDB)	[56]
Feb. 1996	Human genome project (HUGO) Bermuda agreement on sharing data	[21]
Feb. 1997	HUGO Bermuda agreement on sharing data	[22]
1998	HUGO Bermuda agreement on sharing data	[23]
1999	Sloan Digital Sky Survey pioneers large-scale astronomic data publishing	[24]
2000	Stanford Linear Accelerator starts data taking for BaBaR	[25]
2001	Human genome sequence published with studies of 30 genes already published using pre-bublication data releases	[26]
2003	HUGO Fort Lauderdale agreement on sharing data	[27]
2006	First five years of SDSS analysed	[28]
2007	NSF solicitation for DataNet projects	www.nsf.gov/pubs/2007/nsf07601/nsf07601.htm
2007	Extremely Large Database (XLDB) workshop	[29]
Mar. 2008	Yahoo-hosted workshop on data-intensive research	research.yahoo.com/node/2104
Mar. 2008	SciDB project requirements gathered	scidb.org
2008	Extremely Large Database (XLDB) workshop	[30]
Jan. 2009	Report on GrayWulf architecture and its use	[70, 95]
Jan. 2009	Report of Interagency Working Group on Digital Data to the Committee on Science of the National Science and Technology Council	[5]
Jan. 2009	US President commitment to open data	[12]
Mar. 2009	*Beyond the Data Deluge* paper published	[6]
Jun. 2009	Government commitment to linked data	[31]
2009	Extremely Large Database (XLDB) workshop	[32]
Sept. 2009	Toronto statement on pre-publication publishing of biomedical research data	[33]
Sept. 2009	DataNet projects, DataOne and XXX start	
Oct. 2009	*The Fourth Paradigm* book published	[7]
Oct. 2009	JISC call for Research Data Management projects	www.jisc.ac.uk/whatwedo/programmes/mrd.aspx
Nov. 2009	NSF CISE solicitation for data-intensive research	www.nsf.gov/pubs/2009/nsf09558/nsf09558.htm
Dec. 2009	European e-Infrastructure Reflection Group report endorsed by ESFRI	[34]
Dec. 2009	First multi-national *Digging into Data* awards	www.diggingintodata.org
Mar. 2010	Expected first release of SciDB	scidb.org
Apr. 2010	*Data-Intensive e-Science Workshop*, in Japan	www.diew2010.org

Table 1.1: Events shaping Data-Intensive Research

Chapter 2

Principles

Principles are needed to guide the data-use initiative; the subsequent path for data-intensive research will be dynamically driven by the global research community.

P1 *Support for research data should be in harmony with the evolving digital ecosystem.* The total use of data, in government, business, health-care, engineering, media, entertainment, telecommunication and domestic commodities, dwarfs the anticipated use of data for research[1]. Research strategies should therefore work with this larger digital ecosystem, developing their niche and influencing the larger system[2].

P2 *Attention to the analysis of data should be increased in order to make the best use of data.* The arguments for collecting or generating data are well rehearsed [5, 9] as are those for preserving data [36, 37]. Investment in improving the use of data should be increased to meet research needs and to maximise the overall value of generating, collecting, preserving, curating and archiving data.

P3 *Co-evolve research practices with new methods and their supporting software.* Ensure that researchers developing new data analysis methods work closely with researchers using the methods to avoid wasteful bouts of technological determinism where technologists collect requirements and develop a system that is not used because the planned users' ideas have moved on.

P4 *Democratise research by improving education and access.* All members of the community will gain from a better understanding of the wealth of data and the new methods for using data. Consequently, education should seek to develop computational thinking [18] and advanced education should introduce data-intensive methods [7]. This should prepare the way for research and equip citizens to access data and perform analyses, so that they too may review the basis for decisions. Benefits to society will include: better understanding of the reasons for decisions, a workforce adept at deploying data-intensive methods and contributions by enthusiastic, skilled, volunteer 'citizen scientists'. The ethical and legal frameworks for research will need revision in response to emergent changes in research practices.

[1] The IDC estimate of world-wide digital data by 2011 is 1.8 zettabytes (1.8×10^{21} bytes) in more than 20×10^{15} containers — files, messages, images, DVDs, TV and radio broadcasts, films, etc.) [35].

[2] This larger digital ecosystem is changing society and generating new phenomena, which should be the subject of research. Its operational data are often valuable material for research and application science.

P5 *Smooth the path from foundational research via proof-of-concept to sustained use.* The growing wealth of data and the rapidly expanding global use of data is unprecedented. Consequently, it is essential to undertake new computing-science research to create the formal and algorithmic foundations of the new methods and to enable the new scales of data services. This will be most effective if it develops hand-in-hand with the 'field' experience of developing, using and supporting methods.

P6 *The costs of data-intensive research should be visible to researchers.* The IDC report [35] shows that the growth of digital data outstrips the world's ability to manage and store data. Continued growth of energy consumption and environmental costs of support for research data is unsustainable. Researchers who are well-informed about costs will become more adept at getting research done with less demand on resources — though it will always be a challenge that depends on good engineering of research facilities and ingenious strategies for finding the shortest path to the sought for knowledge.

Chapter 3

Recommendations

The recommendations in this chapter are generic; more specific actions are postponed until chapter 10 in order that they can refer to the intervening material. Implementation will benefit from the larger digital ecosystem (see chapter 5) and should build on the existing beacons of success in data-intensive research. While pursuing the data-use initiative, care must be taken to preserve successful practices and avoid disrupting productive researchers.

R1 *Stimulate new thinking in the next generation of researchers.* The dramatic changes characteristic of the digital revolution require quick and insightful thinking. Well-judged early moves will dominate the new niches in the evolving digital ecosystem. This requires the agility and energy of the next wave of research leaders. Incentives, programmes and events should be used to attract and prepare a strong cohort to the data-use challenges and to facilitate their ability to harvest the new opportunities. Elements of this are:

- international gatherings where high-flying, early career leaders combine their insights;
- centres that focus expertise and lead the new culture;
- programmes that stimulate productive combinations of communities, approaches and skills;
- collaboration with existing beacons of data-intensive research; and
- a global community committed to pooling its efforts to accelerate data-intensive research.

R2 *Invest in creating and sharing methods and software for exploiting data.* There are a myriad ways of exploiting data[1]; for each of these there may be dozens of algorithms each of which may have many implementations to suit different data and different computational contexts. They need to be calibrated, validated, measured and described. They often need modification, e.g. in urgent-computing contexts they need to trade accuracy with speed, whereas, in monitoring contexts, incremental algorithms are necessary to process continuous streams. For as long as there are active researchers, new methods of using data will be being invented with concomitant software requirements. This methods and tools research industry is a vital intellectual endeavour; made more challenging if every step from data to well-presented compelling evidence is to integrate well together and be easy to use. The data-use initiative must engender collaborative development of these new methods and software as *a*) it is costly in skilled human labour and *b*) trust in results will depend on shared experience of their efficacy. Forming communities with

[1] Such as: extracting statistics, discovering patterns, exposing anomalies, detecting delayed responses, comparing observations with model predictions, visualising extracted properties, collating observations, measurements and human judgements, or using them as a reference of agreed information.

shared interests and identifying what to share is part and parcel of the overall endeavour.

R3 *Increase data use by building 'intellectual ramps' and providing education.* There will be many kinds of 'intellectual ramp', examples include:

- best-practice guides for those planning data-intensive research;
- packages for generating metadata;
- automated metadata and provenance generation;
- progress blogging from devices, instruments and tools;
- visual interfaces to pre-integrated bodies of data;
- aids for extracting selected data and delivering them in standard forms;
- steerable visualisation frameworks;
- reference data for an application domain and
- facilities for data archiving and publishing.

The potential range is as wide as the combined fields of research. Many properties of a successful ramp will be transferable, and many ramps can be made by configuring a more generic component with the data and defaults of a particular community or methodology. The data-use initiative will need to develop ramp-making capabilities as there is insufficient capacity at present.

Education and ramps will develop hand-in-hand. Once a ramp has successfully engaged a critical mass of a community a tipping point is reached and the methods it supports become part of their culture. Teaching the methodological concepts and the judgement about when and how to employ a family of methods will often reveal a strategy for building their ramps. Successful ramps allow teachers and students to focus on those concepts and judgements[2].

R4 *Create and sustain the foundations for exploiting data.* Guiding theories are needed to give confidence in results, to support optimised implementations and to steer research towards effective and usable methods, i.e. as an underpinning for the practical and empirical exploration of data-intensive methods and access ramps. These theories must be translated into engineering that enables the economic accommodation of the scales of data and rates of data use. New performance criteria will focus on the rates at which information is gleaned from data and on the costs of supporting that full cycle of data use from capture, through impact to archive or discard. Measurements against these criteria will improve the cost-effectiveness of working with research data. For economic and environmental reasons certain functions will be supported by common services. Choosing and shaping these common services requires coordination across communities.

R5 *Provide a smooth path from proof of concept to production use.* New working practices, methods and tools will be created by collaborating pioneers from all disciplines and data-aware engineers. Where their value is proven these methods will be propagated to a growing circle of users, initially researchers in the originating domain and then in new domains. At each stage the implementations need to be improved to function efficiently in new computational contexts and to deliver extended functionality. For the methods that prove important for a significant number of researchers an obligation emerges to sustain the tools that deliver the methods, this requires continuous investment of effort to track evolving requirements and exploit the innovations in the larger computational context. This leads to obligations, on originators, users, service providers and funders akin to those described in the Toronto statement for data [33], since abrupt loss of tools would be as deleterious as abrupt loss of access to data. See chapter 8.

[2] research.nesc.ac.uk/node/483

Chapter 4

Research users

There are bold claims about changes to scientific practice resulting from the data revolution, such as Wired Magazine's "The End of Theory: The Data Deluge Makes the Scientific Method Obsolete" [38]. In reality data-intensive mehods complement the other research methods and the changes are pervasive but vary in extent and maturity across discipline and sub-discipline, with many areas experiencing their own 'computational turn'. As we scale up with increasing digital participation and increasing automation, the quantitative changes eventually yield qualitative change in research practice. Important research challenges are the driver for harnessing greater capability in data, computation and collaboration, many lying in inter-disciplinary teamwork and burgeoning new research areas.

Researchers are expert practitioners in the research methods of their discipline, and their practices deliver new insights and new results. The tooling to work digitally may improve the execution of these methods in quality, speed and scale, and hence accelerate the process of research. With this also comes a set of new digital artefacts — from data to descriptions of experiments — which stand to propagate know-how and enable reproducible research through sharing and reuse. These are beginning to be incorporated in the scholarly knowledge cycle, through supplemental materials associated with traditional publications, and in the repositories field we see the emergence of representations to bundle together the digital artefacts into shareable objects [39]. Where modern collaborative tools are used to share these artefacts and support scientific discourse we see the emergence of Science 2.0 [40].

Box 4.1

Box 4.2

In many areas we are witnessing something more profound than an increasing facility in existing practice. There is a growing volume and diversity of 'born digital' data arising from new experimental practice, sensor networks, digital media, citizen science or administrative information. This demands automation of existing methods to cope systematically with scale, and it also requires new methods to work with data that is already captured rather than designing experiments or surveys to capture the data. These opportunities both demand and encourage a shift in practice.

This perhaps is the space that gains most through collaboration with computer scientists, not just because new algorithmic approaches are needed but also because new ways of assembling

Notes for: Digital assembly of ancient texts

Digital imaging enables the virtual assembly of texts whose individual parts are held in museums and libraries around the world and that cannot be brought together physically. Three examples are: (a) the assembly of fragments of papyri that were then computationally matched to find torn parts that fitted together [AA], (b) the assembly of a very early version of the bible whose parts were held in different countries [BB], and (c) the digital imaging of Roman clay tablets to reveal hidden text [CC].

Box 4.1: Digital assembly of ancient texts

Notes for: Advent of statistical access

As Google digitises large collections of books, many from the warehouses and vaults of libraries, from all over the world new forms of access become possible. Books may be accessed and read by researchers who could never have found resources to visit the places that held they or have persuaded librariians to find the copies. Linguistic and literary researchers can now use natural-language processing techniques on a much increased statistical scale to discover trends and relationships in language, content and style. Literologists studying the the development of printing and publishing practices can text mine the publishing histories and the production of publishing houses in each period for comparison with company records.

Box 4.2: Advent of statistical access

the components are necessary. The assembly might be in the hands of the domain-specific computing support, or it may be directly in the hands of the researcher who can work with the 'software apparatus' with the same proficiency that they have previously worked with the experimental apparatus of their research field. A case in point is the use of scientific workflow systems, which provide a high-level description of a process without the researcher needing to do low-level programming. Ultimately this approach of empowering the researcher to be creative in the use of tools gives them the power to create new possibilities for themselves and others.

Some researchers gain tremendous benefit from new capabilities in data processing, simulation and modelling brought about by new hardware. While the heroic few are motivated to harness this power to conduct 'Big Science', it does not follow that the capabilities automatically benefit every researcher. The challenge is to expose an appropriate level of abstraction to the scientist to match their needs and to empower their research. We discuss this in chapter 7.

Significantly, important research questions also demand changing methods and practice, and lead to research conducted around problems rather than in silos. This is exemplified by climate change research, in which multiple disciplines come together and the overall research team is much larger than earlier endeavours. The change in practice is captured by Braman [41]: "Individual researchers have made the transition from working alone with only their own data, to working alone with everyone's data, to working in teams in one place with single-study data, to working in teams in one place with data from many studies, to working in distributed teams with massive data sets." Techniques in one field may influence others; for example, an approach such as ensemble methods[1] may have application in the social sciences.

The capability of infrastructure in data, computation and collaborative tooling is essential, but so is the agreement around interchange of data and metadata. Some disciplines are more mature in this digital practice. One beacon is the community of providers and users of earth observation data, a field which has been established for decades, is inherently international, where applications motivate a coherence in standards and where software tools are established and maintained. Dozier and Gail note that Earth and Environmental Science has gone through two phases and is entering a third: up to two decades ago, each discipline, e.g. geology, oceanography, ecology, etc., focused on its own challenges; in the last two decades a systems approach emerged where the interaction of these disciplines was necessary to understand the discipline. The new phase, just starting, is applied science — the attempt to use these integrated systems approaches to address practical problems [16]. No matter how mature and motivated the field researchers are always 'pushing the envelope' (see section 5.3). A complex field such as bioinformatics involves many domains of independent research, leading to a proliferation of data sources [42] and a rapidly evolving ecosystem of standards and tools.

The curators providing reference data collections (see **C6** & **C7** on page 36) are involved in setting up processes for the continuous ingest of new data, of the addition of professional and automated annotation, and of delivering data-intensive computational services based on the collected data. They have to balance the quality goals with rates of ingest and annotation. They decide when a pass over the collection is warranted in the light of new science or new standards, such as ontologies. They need to work with data suppliers and users to understand which aspects of existing services should continue and which should be upgraded. They introduce new information, methods, tools, standards and practices to the research communities. To do this, they build teams who combine knowledge at the research frontier of the disciplines they

[1] The use of a set of simulations with parameters chosen to explore a region of possibilities. Their results are then combined to yield the computational goal, e.g. a weather forecast with estimates of probability. Ensemble methods are widely used in engineering, physical and earth sciences.

serve, professional data curation and exploitation of computation. Such teams have grown in number, size and sophistication over the last three decades as the dependence of research on data has grown. They are now a vital and integral part of a research community and for many disciplines their roles have to be sustained for the discipline to progress.

In our data-centric world, the typical lifecycle of an investigation might start with discovery of resources followed by their acquisition, then the actual conduct of the work in collaboration with project members, followed by the publication of resulting papers, data and methods, all of which then benefit from curation. The open science approach has a stronger mandate to public visibility throughout the investigation, more akin to open source. However, data sharing also brings issues of privacy and of rights flow. When data is integrated from multiple sources, be it by distributed query, workflow or mashup, the question of the licence on the result is an issue requiring progress in legal understanding and practice.

We also observe two important trends which are brought about by changes in available hardware and associated business models. The first is increasing use of cloud computing, be it commercially-provided or institutional, which is particularly attractive where the research patterns vary their computational activity significantly — it is considerably more flexible than a local hardware alternative. On the other hand, it is also now entirely realistic for a research collaboration to establish a dedicated hardware resource for a particular dataset or application, particularly when it makes sense for the computation to occur close to the data.

Ultimately, researchers are motivated by their research problems and a desire to increase reputation, and how they manage their data must be understood in this light. In some cases data are placed back in the commons for public access, through community practice or as a condition of the award of funding. Current reward structures, based on citation, are largely oriented around academic publications: if data were given the same status then the incentive structures would favour publication of data in a re-usable fashion and encourage credit through re-use and citation. A similar argument could be applied to software and workflows. In the digital world we can automatically capture and make available the flow of IPR.

Chapter 5

Digital-data ecosystem

As a result of its use for communication, as a means of recording information and as the medium for preserving or transporting computational state, digital data are ubiquitous and increasing. Data underpin business, are widely used by government, are shaped by standards and conventions, and are an established research resource. Data-intensive research has to take account of the products and changes in this much larger, 'corporate' context, from which it draws relevant methods and technology and to which it will deliver new methods and technologies. This chapter reviews some of this corporate activity and the expected interactions.

5.1 Communication, business and entertainment

Numerous forms of communication use digital packets of data, e.g. RFID tag transmission, SMS messages, mobile-phone conversations, instant messaging, voice over Internet protocols (VOIP), web-page accesses, online transactions and business-to-business processes. Some, such as voicemail and email, store and forward messages with multiple cached copies *en route*.

Businesses record data to support business intelligence and business management. In the former, they use online analytic processing (OLAP), data mining and forecasting by modelling and fitting to aid business decisions such as procurement and marketing. In the latter, they undertake human resource management, budgeting and financial reporting, inventory control and operations. They develop workflows to capture business practices and use automation to increase precision, agility and repeatability.

Amazon[1] and eBay[2] are archetypal of businesses that conduct all of their commercial activity as data exchanges over the Internet. Globally accessed data sources, such as search engines and map services, deliver a predominantly one-way flow of information from provider to consumer. They attract a mass of consumers as a market for advertising that yields substantial revenues. Collected usage data enables providers to use business-intelligence to target advertising and shape products.

[1] www.amazon.com

[2] www.ebay.com

Shared-data services, such as Facebook[3], LinkedIn[4], Flickr[5], YouTube[6] and DropBox[7] extend the data-provider business model to include mechanisms for data to flow from consumer to provider. This mixes the provision of reliable storage with communication for groups with common interests. Services for collaborative working are provided, e.g. Google docs[8] and are integrated with communication, e.g. Google Wave[9]. The most successful collaborative authoring tool, a wiki[10], enabling global contributions and community organisation is exemplified by the collaborative encyclopaedia Wikipedia[11] with eleven million registered users and more than three million contributed articles in English constructed by nearly 350 million edits.

The film industry uses digital data for production and distribution; they were pioneers of computational farms and data grids to perform rendering and animation. For three decades newspapers and books have been composed and published using digital technology, music has been distributed in digital formats for two decades and television and radio are now broadcast digitally. In the UK the BBC has played a leading role in the Linked Data movement by publishing information about programmes and music making use of corresponding ontologies. The early adoption of digital techniques throughout the production and distribution chain provides a rehearsal for opportunities and challenges in other domains, notably the impact on business models. For example, Mendeley[12] lets researchers organise, share and discover research papers based on an iTunes-like model.

Over the past four decades engineering companies have transformed their businesses to model their products throughout design, testing, production and most recently operation. In the 70s the focus was largely on design, by the 90s computer-aided engineering used the data from design to drive production — leading to better precision and progressively more agile product cycles. Today, many products are instrumented to collect operational data to support maintenance, re-design and business improvement. In many industries, legislation requires that these data be preserved. This means that we now live in a built environment and use products that have copious collections of associated data — an opportunity, working with the owners of the data, to investigate and to inform decisions about the artefacts and their use.

5.2 Government

Hollerith pioneered the tabulating machine to process the USA census of 1890 and could therefore be the father of data-intensive research[13]. Every aspect of national and international information relevant to government and the services supporting society and its citizens, e.g. security, health care, agronomy, finance, law and defence, depends on large and complex collections of data; progressively, these collections are being made public. Recent trends include:

- the European Research Council statement on open access that includes "… essential that primary data … are deposited to the relevant databases as soon as possible …" [43];

[3] www.facebook.com
[4] www.linkedin.com
[5] www.flickr.com
[6] www.youtube.com
[7] www.getdropbox.com
[8] docs.google.com
[9] wave.google.com
[10] en.wikipedia.org/wiki/Wiki
[11] en.wikipedia.org
[12] www.mendeley.com
[13] www.columbia.edu/acis/history/census-tabulator.html

- the European INSPIRE Directive [11] with the adoption of over fifty standards covering, *inter alia*, geospatially located data[14];
- the e-Infrastructure Reflection Group report on the management of research data [34];
- the USA's inter-agency agreement on sharing data for science and society [5];
- the UK Cabinet Office set up the "*Power of Information Task Force*" in March 2008 [44, 45] which is primarily concerned with making information more accessible [46, 47]; and
- standards for presenting an integrated view of government data [31] and commitment to use them in the USA [12] and the UK [13, 48].

International bodies, such as: the UN[15], the Inter-governmental Panel on Climate Change (IPCC), OECD[16] [49] and World Bank[17], also gather and publish large collections of data that are typically public. National organisations, such as the Met Office[18], Ordnance Survey[19] and British Geological Survey[20] in the UK, and the National Geological Survey[21] and National Oceanic and Atmospheric Administration[22] in the USA manage data on behalf of their nations. In the USA their data are public, in the UK the trend is towards making their data more public[23].

As a result of these trends there are data on a wide variety of subjects, that originated for governmental purposes, which can now provide information for many avenues of research.

5.3 Standards

Every data-intensive researcher builds on a great many standards, from those for the Internet (IETF) and Web (W3C) to those specific to their community, e.g. NetCDF[24] for climate data endorsed by IPCC[25] or FITS[26] for astronomy data[27] endorsed by the International Virtual Observatory Alliance (IVOA). Most researchers are depending on many standards that are mostly hidden by the software, such as character-set representations[28], floating-point numbers[29], XML

[14] The INSPIRE Directive covers thematic data in 34 areas: Coordinate reference systems, Geographical grid systems, Geographical names, Administrative units, Addresses, Cadastral parcels, Transport networks, Hydrography, Protected sites, Elevation, Land cover, Orthoimagery, Geology, Statistical units, Buildings, Soil, Land use, Human health and safety, Utility and governmental services, Environmental monitor- ing facilities, Production and industrial facilities, Agricultural and aquaculture facilities, Population distribution P demography, Area man- agement/restriction/regulation zones and reporting units, Natural risk zones, Atmospheric conditions, Meteorological geographical features, Oceanographic geographical features, Sea regions, Bio-geographical regions, Habitats and biotopes, Species distributions, Energy resources, Mineral resources.

[15] www.un.org/esa/sustdev/natlinfo/indicators/idsd//methodologies/data.htm

[16] www.oecd.org/statsportal/0,3352,en_2825_293564_1_1_1_1_1,00.html

[17] tinyurl.com/dts98

[18] www.metoffice.gov.uk

[19] www.ordnancesurvey.co.uk

[20] www.bgs.ac.uk

[21] www.usgs.gov

[22] www.noaa.gov

[23] tinyurl.com/y8tqnlx

[24] www.unidata.ucar.edu/software/netcdf

[25] www.ipcc.ch & www.ipcc-data.org

[26] fits.gsfc.nasa.gov/documents.html

[27] www.ivoa.net/Documents

[28] e.g. the ISO standard for 8-bit coding of Latin character sets
www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=29505
or the Unicode unicode.org 16-bit encoding for globalisation.

[29] grouper.ieee.org/groups/754

for semi-structured data[30], RDF for semantic description[31], SNIA's standards for data-storage[32], OGC's standards for geospatial data[33] and and OGF's standards for grid interoperation[34]. Programmers and operations teams wrestle with these and many more standards to get software to interoperate and to enable researchers to combine or move between services.

In many cases, standards required by corporate users are also suitable for data-intensive research; adopting them brings advantages such as good implementations and tools. The choice of standards becomes more difficult when standards pertain to recent advances. Researchers often exceed the limits for which a standard was designed or explore technologies before consensus on pertinent standards has emerged in the corporate world. For example, today they have to choose how to access services via the Internet and whether to compose services using standard workflow systems, such as BPEL[35], one of the many workflow systems developed for scientific or engineering applications [50, 51] or the APIs and MashUps of the Web 2.0 world[36]. Research cannot wait, so later re-engineering when consensus emerges or a less comfortable match with the larger digital ecosystem is an ineluctable consequence.

Researchers always contend with a boundary where standards have been agreed for some aspects of their discipline but others have yet to be agreed. They have judge whether to invest time on a standard or continue with *ad hoc* arrangements. If they opt for standards they have two challenges *a*) to avoid being too ambitious and *b*) to gain consensus from enough of their community that the standard gets adopted. The AstroGrid project and its successors, are a good example, they adopted web service standards and developed standards such as FITS for inter-virtual-observatory communication, and other standards for tools on an astronomer's workstation to communicate. The result is that some 90% of the world's astronomy data is reachable from that workstation and real science can be done on it by linking the tools that are being built all around the world.

The effort to develop the Climate Systems Modelling Language illustrates well the other challenge [52, 53]. Several communities working to integrate ocean and atmospheric data recognised that despite the rich set of geo-spatial, satellite imaging and sensor standards developed by OGC, they required better representations of observational tracks and spatio-temporal data, e.g. a time series giving a grid of mean sea-surface temperatures for the North Atlantic for each January for the last 100 years. Several projects used CSML successfully for integration [54]. The researchers sought a wider community buy in to get all the major players, including companies and national meteorological offices, interested in adopting the standard. This co-evolution of consensus and a standard is necessary for the standard to work. But it is difficult, as new participants want new features, and then these features start interacting, making the standard complex and hard to define. The complexity deters adoption and people only commit to partial subsets. If this doesn't converge quickly the workarounds get set in stone and the researchers will have invested large amounts of time without any benefit.

[30] www.w3.org/XML
[31] www.w3.org/RDF
[32] www.snia.org
[33] www.opengeospatial.org
[34] www.ogf.org
[35] www.oasis-open.org/committees/wsbpel
[36] www.programmableweb.com

5.4 Established research data resources

Many researchers use existing research data services (see **C6** & **C7** on page 36), e.g. in 2008 there were 1170 molecular biology databases [42] and the Inter-University Consortium for Political and Social Research (ICPSR) identified 48,000 papers relating to its collection of data[37]. Many databases capture decades of expert effort in curating and evolving major collections, e.g. wwPDB [55], NASA, ESA, NCAR, BADC, MedLine, ... There are already many thousands of research data collections. For each of these, thousands of researchers have developed skills and software with large investments in modelling, analysis and visualisation tools. This investment enables productive research. It also establishes *de facto* standards that may, like the formal standards, be a mixed blessing: they lubricate research with ready-made solutions but it takes courage to leave their supported territory. The reference data services continuously invest in improving their product, balancing the requirements for new information against requirements for stability, e.g. ELIXIR[38], wwPDB [56, 57] and the preparations for the fifth IPCC report.

The challenges facing in the era of data-intensive research include: *a*) coping with the rapid growth in user numbers and submitted requests for their supported services; *b*) coping with the increasing volume and complexity of data, and the increasing rate of data or document deposition; and *c*) balancing the pressures to accommodate new requirements and information structures with the requirement to protect researchers' existing investment (see chapter 4).

This will be illustrated with examples. *Currently thinking of GenBank+EMBL+DDBJ sequence data with 1000-Genome Project Trace storage as scale. Look at wwPDB workload for usage growth. Look at new function, e.g. new forms of structure at PDB + new DOMs for more species, sustainability scale, skilled human labour annotating, annotation consortia. Prep for IPCC5. SDSS experience. Need to track down usage, size & complexity evidence in papers and from web sites.* mpa

5.5 Research in the digital ecosystem

There is a valuable two-way flow of ideas and technologies between data-intensive research and the corporate-data ecosystem, illustrated below.

I1 *Communication* is key to research; researchers need to find out what other researchers are doing, collaborate with scattered colleagues and publish their work. Search engines have been substantially refined in the corporate systems [58]; they are used routinely. Researchers appreciate more sophisticated retrieval tools, controlling the sources, improving relevance and tuning presentation to their needs e.g. the selection of the right kind of information in [59, 60]. Collections of open publications, such as Pub Med Central[39] and arXiv.org[40], and catalogues, such as DBLP[41] and CiteSeer[42], amplify the power of search and publication. This interplay between research-specific and corporate search tools will remain active for the foreseeable future.

The tools for collaboration are also evolving. The majority of operational and technical issues during the Sloan Digital Sky Survey (SDCC) will be captured by archiving emails [61]. Today's

[37] www.icpsr.umich.edu/icpsrweb/ICPSR/citations/methodology.jsp
[38] www.elixir-europe.org
[39] www.ncbi.nlm.nih.gov/pmc
[40] arxiv.org
[41] www.informatik.uni-trier.de/ ley/db
[42] citeseer.ist.psu.edu

projects use a plethora of communication methods, e.g.: computer-supported collaboration tools, social networking, version management, email, instant messaging, blogging, twittering, wiki, VOIP, web-seminar and web-conferencing. If an archive is intended to enable researchers to understand how discoveries were made, how should today's projects be archived?

Researchers also communicate with their equipment, e.g. they have arranged that their instruments blog, so that they can always find out what is happening in their laboratory [62]. Projects such as OptIPuter and AccessGrid[43] have coupled haptic devices and remotely steered visualisations, microscopes and telescopes [63].

I2 *Collaboration* is key to research and corporate success. Researchers exhibit a wide variety of collaboration behaviours, from working alone to closely knit teams. They may draw on a wide variety of collaboration aids, e.g. Web 2.0 services, such as those for scheduling meetings, conducting quick surveys and authoring shared documents. In doing so, they are influencing the corporate tools even though research use is a tiny fraction of their market, e.g. Microsoft is developing mechanisms for typesetting chemical formulae and for embedding data in their documents [64].

Research communities of software developers and users sometimes adopt predefined cooperative models, e.g. by using the Apache Software Foundation[44]. The myExperiment[45] project is investigating whether a less formalised 'social computing' *modus operandi* will be productive, inspired by the success of Wikipedia. They hope to see an increase in reuse and repurposing of the combined work of a creative community for workflows in any language [65, 66].

The power of computational frameworks, such as: workflows, database queries and Pig Latin, is crucially dependent on a rich library of computational components, called 'user-defined functions' (UDFs). If ways could be found to pool this resource so that UDFs worked in a wide variety of the contexts, their value to researchers and the corporate digital systems would be a greatly increased.

I3 *Analytical queries* Research in academia and industry followed by three decades of joint research and industrial engineering has led to optimising relational query engines that are extremely powerful. They utilise cluster and distributed computing and allow very sophisticated queries to support OLAP and many specialised forms of data, such as geospatial. Jacobs notes that many scientific questions, e.g. discovering delayed responses, canot be handled with OLAP and that existing tools can be pathalogically slow [67]. User-defined functions extend the power of query systems almost indefinitely and can be combined with tailored indexing strategies [68]. These are then used by researchers, for example, to access the bioinformatics data at NCBI and EBI [AA], to manage social science data in CESSDA [BB] and to access the Sloan Digital Sky survey astronomic data [24, 28, 69].

This led to joint research with Microsoft into data-intensive architectures [70], to enhancements to SQLServer to better accommodate scientific processing [71] and investment in data-intensive cloud computing [72]. Work on extremely large databases, such as, XLDB [29, 30, 32], the planning for LSST [73] and SciDB [74, 75, 76, 77], is engaging both industry and academia, while GrayWulf is beginning production work for Pan-STARRS[46] [78]. Here again we see sustained interchange of ideas, technologies and strategies between the research and corporate domains.

[43] www.accessgrid.org
[44] www.apache.org
[45] www.myexperiment.org
[46] pan-starrs.ifa.hawaii.edu/public

Chapter 6

Datascopes

New methods and tools are needed to make more effective use of data; we call these '*datascopes*' as they reveal previously hidden evidence to the 'naked' mind just as telescopes reveal the universe to the naked eye.

Telescopes have yielded a succession of views of the universe as their technology has co-evolved with the advances in understanding that they provoked. Similar progress in our ability to comprehend data is needed. Early astronomers were involved in details of the technology, but today much of the design, construction and operation is delegated to specialists. Today, researchers extracting evidence from data are often heavily engaged in the technology, just like early astronomers. An archetypical example is the large numbers of physicists who have been engaged in preparing software to extract evidence from the LHC data[1].

The goal of data-intensive research is to extract information from data, interpret information to acquire wisdom and evaluate wisdom to develop trust. Information extraction may be a direct test of a hypothesis or a search for patterns and anomalies that may inspire hypotheses. Wisdom can be developed by independently formulating models and comparing their predictions with the information. Repeated success, e.g. in predicting planetary orbits and then other planets from perturbations in those orbits, develops confidence in the model; Newtonian mechanics and gravity in this example. Correct selection and use of models can be interpreted as wisdom. When experience enables people to be confident about when each model will work they trust their formulation. This depends on independent evaluation by re-walking the path of evidence.

An *application science* is under urgent pressure to deliver knowledge to support action in time to avert disasters by deriving adequate and tailored information using imperfect data; there is little opportunity to learn from experience [16]. Even without application pressure, far better technology is needed for extracting information from data in order to increase the "intellectual velocity" of research [79].

We cannot possibly anticipate where building better datascopes data will lead. The early astronomers using *optical* hand-steered telescopes could not have imagined the Sloan Digital Sky Survey, the Hubble telescope[2], the Cosmic Microwave Background Explorer (COBE)[3], the

[1] lcg.web.cern.ch/LCG/public
[2] hubblesite.org
[3] lambda.gsfc.nasa.gov/product/cobe

Herschel telescope[4], the Square Kilometre Array radio telescope[5] and the neutrino detector in Antarctica[6], nor their impact on conception of our place in the universe and the laws of physics.

We need a campaign to develop skills and tools for data-intensive research. Datascopes and the understanding they yield will co-evolve as dramatically as telescopes and cosmology — but at Internet speed!

D1 *Frameworks* Astronomical telescopes have a computationally steered framework on which the components and experiments are mounted. Datascopes require computationally controlled software frameworks interconnecting their components; today's candidates are workflow systems, query engines, map-reduce engines and desk-top tools. A good framework will monitor operations, generate metadata and provenance data, and deliver aids for detecting and diagnosing malfunctions.

D2 *Gathering* Astronomers steer mirrors, parabolic radio dishes and arrays of detectors to gather the photons they want. Data-intensive researchers need mechanisms to gather the data they want; from instruments, experiments, scattered sensors, mobile devices, existing data collections, runs of simulations or a (by-)product of any activity in the digital ecosystem. Researchers will select, deploy and control these data gathers. Whilst the sky is open access, data-intensive researchers often have to gain permission to gather data.

D3 *Capture* Astronomers use photography, digital detectors and data storage to record incoming signals. Data-intensive researchers capture their data in digital storage. A process that may include encryption, optimised representations [80], compression and redundant copies. This is the locus for curating *primary* data, for example of the images in a sky survey [24, 81, 73] and the traces from DNA sequencers [82, 83, 84]. Such archives have to support revisits by researchers who have developed better methods for the subsequent stages of the datascope or who wish to explore phenomena the next stages hide.

D4 *Preparation* Variations in the atmosphere, instruments and background noise distort the raw astronomical signal. To reduce the impact of these effects on subsequent processing a series of computational 'cooking' operations transform the raw signal into standardised forms. Similar requirements occur for almost all sources of data; there will be missing data, erroneous values, distortions caused by variations in the collecting processes and so on. Data cleaning processes are needed to validate data quality, deal with repairable data, reject irredeemable data, normalise data against reference data and transform it to standard forms. Data popularity obeys a Zipf distribution, 90% of requests being satisfied by 10% of the data, much data is rarely or never used. Consequently, it may be more economic to clean data as it is requested than to clean all of it when it is deposited, particularly as cleaning for a specific task may be less onerous.

There are already a great many tools for data cleaning; often developed for specific kinds of data or error, or for specific sub-domains of data-intensive research. Many require considerable skilled work by researchers, e.g. specifying rules or building training sets. A more automated approach has been developed by Wenfei Fan [85, 86, 87]. Sampling, image processing, text mining and pseudomisation[7] are important data preparation processes with their own R&D programmes. Whatever form data preparation takes, automation, e.g. using scientific workflows, is necessary

[4] sci.esa.int/herschel/

[5] www.skatelescope.org

[6] icecube.wisc.edu

[7] Removing information that may enable recognition while retaining features important for the research, e.g. PrivacyGuard (research.nesc.ac.uk/node/486).

to handle the incoming data flow [88, 89, 78].

D5 *Cataloguing* Collections of astronomical plates, images or astronomical objects are catalogued. Various forms of automated and human annotation are used to build catalogues. Data-intensive researchers have to recognise which features are important and decide how to classify phenomena, just as astronomers had to decide on stars, planets, satellites, asteroids, galaxies, pulsars and quasars; an iterative research process in any field. Researchers also recognise (and mark 'disregarded') spurious categories, e.g. for astronomers: aircraft, artificial satellites and optical aberrations. Clustering algorithms can suggest categories and machine-learning algorithms may implement classification. Catalogues often contain a sufficient abstraction of the phenomena for many research purposes and can be well indexed to accelerate their use.

D6 *Integration* Research is frequently concerned with relationships, e.g. to detect causality, infer constants or deliver representative aggregations. This often requires the combination of data from different sources, which may first need translation to common representative forms, e.g. using the same: vocabulary, units, coordinate axes, sampling mesh, etc. Some of these transformations are demanding, e.g. warping 3D tissue images to a 'standard' anatomy. The resulting data then need to be brought together so that matching and composition algorithms — often defined by researchers — can do their work.

D7 *Derivation* Data are too numerous to comprehend — often they could not be read in a lifetime — therefore humans require derivatives that represent aspects of the phenomena they are trying to understand. These take many forms, e.g. the statistical characteristics of a population in a healthcare district, a matrix of correlation coefficients, a time series of matrices of atmospheric, oceanic and surface parameters in a climate model, a matrix of absorption properties per voxel[8] of a medical image, surfaces corresponding to a sequence of threshold values, a time series of fitted surfaces, a graph of social interactions, and so on. Data has to be fed into derivation algorithms and their results are grist for further data mills. Much research goes into inventing these derivatives and algorithms, but once invented they are used repeatedly.

D8 *Presentation* This is the 'eye-piece' of the datascope; derivations require presentation to be comprehensible, typically using visualisation with viewing controls. For example, a graph of human interactions or the tree of life can be distorted through a 'fish-eye' lens to bring the part of the graph of interest into the foreground and to give it context with abstracted views of the rest of the graph [90]. Data that is geographically related will be oriented by layering on familiar maps, e.g. using Unidata or Google maps [91]. Properties within a volume may be shown using translucent colouring mapped to thresholds in the parameters of interest — control of the thresholds colours and haptic feedback can reveal the items of interest, as demonstrated by Ynnerman [92]. Feature depiction, annotation, orientation control and slicing let the researcher explore the data and see relationships, as exploited in developmental embryology atlases [93] and in Ellisman's use of computer-game software to present neuroinformatics data from the whole brain to individual proteins in one framework [YY]. Much research is about understanding processes; this is assisted by computationally constructing 'videos' to depict the successive states of the process using the presentation techniques for each state and 'time travel' controlled by researchers.

Many examples exist today, where computationally adept teams of researchers have assembled their own datascope, steered it and extracted impressive results — e.g. [94]. Because there is much domain related insight and intellectual creativity in deciding where to look, how to prepare, combine and classify data, and how to derive and present useful representations, this

[8] A volume element of a 3D image.

will never be an easy task. There will never be a generic datascope that works with all data for all research. However, datascopes that are easy to assemble and control from extensible libraries of data analysis elements are urgently needed to enable the majority of researchers to extend the gains made by pioneers without having computing specialists in their teams.

Existing data-scopes are specific to forms of data, e.g. MatLab[9] expects matrices, R[10] handles sequences of records; both are oriented to data in memory. Query systems can be combined with OLAP tools or extended with functions and provide a framework for data that is, in most cases, relational. These are being extended to handle scientific data, such as matrices, e.g. SciDB [77, 75], but will remain oriented to warehoused data and restricted forms of analysis. Other analytic stages are needed, e.g. discovering correlations with a time lag or clustered in space, assessing the fit between model and observations or fitting surfaces. The text mining, data mining, image processing and domain specific tools need integrating in generic frameworks to deliver more powerful and multi-purpose datascopes. Crosscutting issues, illustrated below, are expected to affect the development of datascopes.

X1 *Efficiency* If computational costs are high experts are required to reduce them or access specialised resources. Moving large volumes of data is intrinsically expensive, this means that computation should often be shipped to data[11], redundant data movement should be avoided and processing platforms should be balanced for data movement [70, 95, 96]. Jacobs shows the pathologies of processing data without using linear access patterns [67]. This strongly suggests devising algorithms that sweep data, which is the rationale for the scanning read and append only file systems and map/reduce processing [97, 98, 99, 100, 101].

Research is needed to transform data-analysis algorithms into incremental algorithms that use the map/reduce pattern and adapt well to memory parameters. Random accesses may be so much slower that it is better to sweep all of the data, particularly if multiple requests are computed in one pass [25]. This poses the challenges of recognising the cases where a sweep will be faster and, for $O(n^2)$ tasks, finding partial results that can be completed during a second scan, and so on for higher-order tasks. Running a data sweep on a regular schedule, e.g. every hour, can deliver predictable response times for complex requests. Indexing and hashing schemes have to deliver very large reductions in data accesses before they provide benefits [102].

Delivering datascopes with sufficent performance poses several research questions, as shown by the following examples.

X1.1 *Platform architecture* The case for a balanced hardware platform has been well made [70] and the latest investigation by Szalay is using much lower power processors and solid-state disks [xx]. These architectures will evolve further with an impetus from corporate R&D, e.g. Windows Azure[12].

X1.2 *Software architecture* Taking into account corporate R&D there are two front runners: those that exploit the context of query engines, e.g. GrayWulf [95] and SciDB [77], and those that exploit the Map/Reduce frameworks, e.g. Hadoop[13] [100] and Yahoo's Pig Latin [103].

X1.3 *Analysis algorithms* It is as yet unclear how many of the required algorithms can be recast in incremental forms that fit in one of these frameworks and use linear access patterns as far as

[9] www.mathworks.com

[10] www.r-project.org

[11] Data grows as a result of digital technology, but programs are created by brain power, which does not achieve comparable performance gains, therefore, programs do not grow as fast as data.

[12] www.microsoft.com/windowsazure

[13] hadoop.apache.org

possible. A concomitant questions is how the existing investment in functions capturing aspects of the natural world and methods for analysing particular forms of data can be carried forward and reused in this algorithmic context. There are clear benefits from automatic transformation but is it feasible? Exploring such questions will also investigate the extent to which the platform and the algorithm can be considered independently. There would be significant advantages if algorithms were formulated so that they could move between platforms.

X1.4 *Approximation* It is possible to use algorithms with lower cost that yield an approximate answer. Investigation is necessary into how to characterise, present and control this trade off, and into researchers' responses to its availability.

X2 *Plugability* Data needs to pass between stages of a datascope and these stages may be repeated or composed of many data analytic steps. Plugging the components together should be made easy for the research user. Research is needed to automate the data transformations necessary between components, the choice of data transport and the coordination of the components. The leading mechanisms for specifying the composition are workflow systems and query languages. An investigation is needed into how these notations can be mapped to the evaluation architectures that emerge from X1.

X3 *Dependability* Researchers have to be able to trust a datascope. This will be based on a range of issues, such as: the evidence for the validity of the algorithms used by the stages, the monitoring techniques that help them review how their data is treated, the provenance records and the security and privacy mechanisms.

X4 *Sustainability* Investment in developing and using a datascope will be predicated on researchers' assumptions about its continuing availability. A major challenge is to find the ways in which software complexity can be managed as the sophistication and capability of a datascope grows. This requires a software architecture that balances efficiency with validation of the components and their composition. Successful datascopes will depend on a critical mass of experts contributing to their design, implementation and use. This is likely to be organised as an open project under which it is hard to control the evolution of a software architecture — this is the reason why SciDB is not an open project until its first release in March 2010[14].

[14] scidb.org

Chapter 7

Intellectual ramps

We introduce the notion of an ‘intellectual ramp’ that lets researchers engage incrementally with the tools, techniques and methods of data-intensive research, so that they can meet their own needs when they choose. A ramp is a safe and supportive means for researchers to advance to more sophisticated data use, without encountering a demanding learning barrier before new methods can be applied successfully.

Ramps come in many forms: best-practice guides, assisted provenance and metadata generation, visual interfaces to pre-integrated bodies of data, aids for extracting selected data and delivering them in standard forms, and facilities for data archiving and publishing. Some ramps can be made by configuring a more generic component with the data and defaults of a particular community, method or user – a task-specific interface for the workflow of the researcher.

The ramp is a powerful socio-technical metaphor: once people see the ramp as an object in its own right we can look at its shape, how and why researchers move around it, how it’s built and how well it works. Many data-intensive research projects and services have built ramps to meet the varied needs of researchers in different communities. Some find compelling the notion of an ‘on ramp’ to data-intensive research; for others, a ramp is somewhere that researchers can dwell, or just visit occasionally in the conduct of their research — to ascend is an opportunity not an obligation.

In the US the Science Gateways endeavour to be ramps, exemplified by NanoHub[1], which is designed to be a resource for the entire nanotechnology discovery and learning community. The Unidata activity delivers not just data but associates software tools for use in geoscience education and research[2]. It is interesting to note that these successful ramps have ‘education’ in their mission statements. Some ramps are human, e.g. NCAR’s Earth Observing Laboratory brings a team of scientists, software engineers and programmers to deliver data-management services throughout a project lifecycle[3].

The UK Virtual Research Environment projects[4] are ramps: myExperiment provides a ramp into the use of scientific workflows, important because workflows are easy to run but tricky to write [66]. Some of the efforts to hide complex infrastructure behind simple APIs are ramps

[1] nanohub.org

[2] www.unidata.ucar.edu

[3] www.eol.ucar.edu

[4] www.jisc.ac.uk/whatwedo/programmes/vre.aspx

for developers, like SAGA (Simple API for Grid Applications)[5] [104] and other offerings from the Open Middleware Infrastructure Institute (OMII-UK)[6], whose business is ramp engineering. For the scientist, a simple drag-and-drop interface to running e-Science computations is a gentle ramp — elegantly demonstrated by the Drop-and-Compute interface to Condor job submission developed in the Manchester Interdisciplinary Biocentre [XX].

Some ramps have facilitators to guide researchers up the ramp — think of librarians assisting researchers who are then able to help themselves. Sometimes people deliver services as intermediaries rather than facilitators; there is then a concern that this may hide the computational thinking from researchers, discouraging them from recognising the full gamut of opportunities and from contributing to method innovation. Intermediaries, as 'power users', may inhibit the development of ramps suitable for direct use by their clients. The first steps on a ramp should conceal complexity and provide the illusion of unbounded resource. These realities may be incrementally exposed as a researcher develops understanding of the computational thinking behind the methods a ramp supports.

For the provider of the data-intensive science superhighway the ramp metaphor may conjur a compelling image of the on-ramp for their users. But our intellectual access ramps are also places where a researcher might happily live in one place and not be obliged to travel further – a place of comfort and safety. A researcher might use a ramp for just a few minutes regularly or even intermittently as part of their research routine, to achieve some incidental task, and have no interest in greater understanding – we can liken this to catching a bus.

It seems that successful ramps are characterised by an effective alignment of community, data and software, and they have a role in developing research skills as well as in conducting research. It follows that ramp construction needs an alignment of interests and funding encompassing a community of users in research and education, their data and service providers, and the developers of their software tools. This combination might not be in the remit of any one funder but it is in the interests of all, because everyone stands to gain from researchers ascending the intellectual access ramps to achieve new outcomes and build new know-how.

M1 *A specialised environment for safe play* One approach involves provision of a customised, task-specific user interfaces to facilitate the work of a specific group or community of users. Behind the interfaces may be databases, computational infrastructure, instruments, workflows and visualisation tools, but the complexity of using and assembling them is hidden from their user. The resources may be community-provided, as in a Science Gateway, with tools, applications and data that are integrated via a portal or as a suite of application programs. They are usually presented via a graphical user interface, that is customised for the work of the targeted community[7]. While that definition is couched in terms of the technology, the related notion of a collaboratory takes a more social stance: A collaboratory is an organisational entity that spans distance, supports rich and recurring human interaction oriented to a common research area, and provides access to data sources, artifacts and tools required to accomplish research tasks [105]. This is closely allied to the UK notion of the Virtual Research Environment. The Nanohub Science Gateway and myExperiment VRE [66] exemplify this kind of ramp.

M2 *Augmenting the familiar research environment* Here the researcher remains in their familiar work environment which is augmented with new tools, components or services in order to access the capability of data intensive science. Classical examples of this include the R project for

[5] saga.cct.lsu.edu
[6] www.omii.ac.uk
[7] http://www.teragrid.org/gateways/

statistical computing and the MATLAB language for technical computing, both of which have enjoyed various parallel extensions. Scientific workflow tools, such as Kepler[8] and Taverna [106], augment the environment with automation and make it easy to bring through additional components and services. They can be bundled with kits of components and services for particular user communities. Applications may themselves be enhanced, and developers are facilitated in this by an approach such as the SAGA (Simple API for Grid Applications) API which insulates application developers from middleware providing a ramp for developers.

M3 *The methods ramp* A UK study reported e-Infrastructures are often seen as complex and challenging by their users (both current and potential). It is clear that current users often experience frustrations, while potential users may be unaware of its benefits and of how to take the first steps towards exploiting them [107]. One approach to this is to equip researchers for data intensive research through the methods in which they are trained. This training takes many forms and occurs at different times in the development of research skills.

[8] www.kepler-project.org

Chapter 8

Path to Sustainable Solutions

There is no single recipe for achieving research breakthroughs or discovering and enabling new data-intensive methods. However, there are ingredients that are particularly powerful and some impediments that should be avoided. Every successful strategy has a means of 'walking a path together' so that thinking on the research journey adapts as understanding grows, as lines of enquiry are followed, as new techniques and resources are discovered and as the larger digital ecosystem changes. The thinking and tools of collaborating researchers have to co-evolve for success to be achieved and maintained.

Long-term research relationships combining computational thinking with expertise in any research domain encourage cross-infection of the modes of thinking as well as comfort with each other's research goals and vocabulary. Partnerships can be complex, bringing in many disciplines, such as mathematics, statistics and engineering, as well as engaging in consortia with other groups to pool effort to assemble major research resources. Leadership is therefore crucial. So too are funding regimes that value and support sustained interdisciplinary collaboration, whether through project funding or institutes. In-depth research relationships take time to build and pay off. Funding that generates an imbalance between the disciplines can actually be harmful.

Leading teams invent new techniques and pioneer their application and implementation. These can be key competitive assets combined with a team's well-honed skills and collected knowledge. Even within such a group, each new technique that proves its worth will transition to an in-house product that they plan to support and use repeatedly. It may undergo re-engineering, its interfaces may be improved and its scope widened. As its value is appreciated it may be shared more widely, e.g. with collaborators in new projects, and hence moved to new computational contexts. This may require investment in 'productising' the software[1] so that it is more transportable, re-factoring so it is more maintainable, re-engineering for performance and writing better documentation — the originators knew how to drive it and understood the limits of its 'safe flying envelope'. If it thrives in a larger community its scope may be widened for use in other disciplines, with a requirement to handle new vocabularies and formats, and to inter-work with other tools. This may require extensions, which are much facilitated if the re-factoring has defined software 'sockets' for user-defined functions that experts in the new communities can write.

[1] Software and databases are common notations that let team members capture and refine their understanding and thus perfect a research method. They are often key to a method's implementation and propagation.

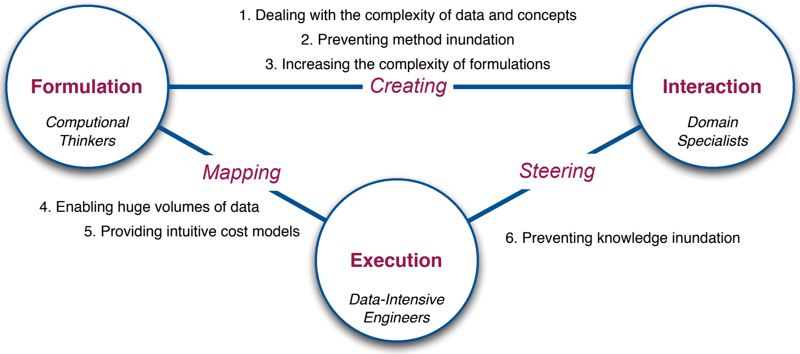
1. Dealing with the complexity of data and concepts
2. Preventing method inundation
3. Increasing the complexity of formulations
Creating
Formulation
Computional Thinkers
Interaction
Domain Specialists
Mapping
Steering
4. Enabling huge volumes of data
5. Providing intuitive cost models
6. Preventing knowledge inundation
Execution
Data-Intensive Engineers

Figure 8.1: The anatomy of data-intensive innovation teams

The whole path from invention to wide use is typical of innovation: many proto-products fail along the way for technical reasons, because of rivals or lack of enthusiastic backers, and each successive step along the path requires new skills and increased investment. The originating researchers are unlikely to take their inventions far along this path — they need to focus on their own research, haven't the skills and cannot muster the resources — they will already be hard at work on the next stage of their research. Consequently, long-term benefit to the larger research community depends on aligning and balancing provisions to carry methods along this path when that is deemed justified by their value to research. Stages on the path are illustrtaed below.

S1 *Creative teams* Figure 8.1 shows relationships in interdisciplinary teams.

The *computational theoreticians* involved in *formulation* draw on computing theory to enable more sophisticated descriptions of phenomena while retaining computable logic. They invent and refine algorithms to carry out more sensitive searches, more precise simulations, more representative derivations and more subtle presentations. They examine the whole machinery to spot the need for underpinning theory to improve the validity and utility of results. They become expert in the challenges in the team's domain; and transfer their thinking to meet similar challenges in other domains — through publication and multiple engagements.

The *data-intensive engineers* who make the *execution* of the required computations feasible are computer scientists, systems architects and engineers who design strategies to map data and computations onto resources. They draw on database, data management and distributed systems research and adapt products from the corporate digital world. They develop algorithms to encode and place data, to accelerate processes and to partition and assign computational tasks. They become expert in the data and patterns of access that their group is addressing and in the behaviour of the research community and the operational systems. This expertise is similarly transferable.

The *domain specialists* deliver *interaction* with a research domain. They know how to obtain relevant data; understand it and their domain's models. They are masters of the research questions and application-science challenges posed by the domain. They may be specialist intermediaries, such as statisticians and bio-, chemo-, and geo-informaticians, or researchers in the domain who have become adept at computational thinking. They work to convert research questions into computational strategies and to test, refine and apply the resulting methods. Their insights can

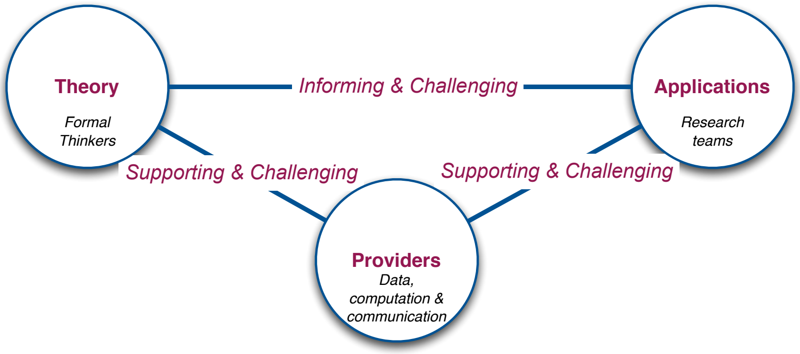

Figure 8.2: The research triangle

be transferred to similar challenges in their domain.

These three elements of a group *jointly* devise research methods. This gains from mutual recognition of the experience each brings and of the value of their research disciplines. It may be undermined by inappropriate expectations.

S2 *Transition* transfers methods and their implementations from their originating group carrying out 'productisation'. This benefits from software architects and engineers, from numerical analysts and statisticians, and from more data-intensive engineering to increase the scope, improve performance and deliver easier code mobility. Improvements to documentation, user-interfaces, parameterisation, outreach, training and consultancy are needed to help adoption. Rallying this effort can be difficult as it does not benefit a specific group — it depends on sustained leadership raising interest, rallying resources and attracting commitment. It can be organised as a foundation set up for the purpose, such as OMII-UK, as part of the work of an institute, such as EBI or NCAR, or as a community contribution via an open-project model, such as project R. During transition the 'product' is particularly vulnerable to diverging from research requirements; an excellent way to prevent this is to embed members of the transition team with active research groups. Success depends on reaching a critical mass of contributors and users in the relevant research niche; so that the research community throws its weight behind the winner, adopting it as a *de facto* standard. These 'mind-share' niches are often worldwide and can spread across subject domains.

S3 *Sustained support* Once there is a substantial community that depends on a set of techniques and the tools that support them, there are pressing reasons to sustain the technology. This means more than fixing bugs and moving the software onto new platforms; it means organising requirements gathering from researchers and responding to opportunities emerging from the corporate digital world. It also recruits and guides effort to implement, validate and package improvements. Whether this is done by a company, an institute or an open-source project, it depends on authoratative leadership with 'staying power'. Chue Hong debunks myths about short-cut solutions [108].

Sustained data-intensive research requires balanced provision for and good working relationships between the three communities shown in Figure 8.2.

S0 *Theory* Theoreticians draw on their research disciplines's theory, mathematics, statistics and computational theory to develop abstract and generic models that address the complexity arising throughout data-intensive research. Providers support theoreticians as they computationally test their models and challenge them with the problems they encounter operationally. Theory is essential to evaluate the safety of conclusions delivered by data-intensive research — no one can check results by hand — this is particularly important where society will act on application science.

S4 *Applications* Research that engages with data from the real or planned world can be termed 'applications'. It is undertaken by the innovating groups pioneering new methods as described above. But it is also undertaken by much larger numbers of small groups and individual researchers pressing on with their research. They constitute the majority of users and as 'the long tail' they are the largest market for the data-intensive methods [28]. Delivering data-intensive methods to these researchers, so that they can easily accomplish routine tasks with relative modest datascopes will yield the greatest return on investment — accelerating their work will yield a massive number of benefits for society and business — hence the importance of stages **S2** and **S3** of the innovation pipeline. The challenge for providers is to simultaneously serve this large community needing convenient services, *as easy as catching the bus to work*, and the demanding community of pioneering groups.

S5 *Providers* The providers range from researchers using their own laptop, via systems administrators running machines and software for a small research group, through departmental, institutional and national provision, to global consortia and multi-national companies. They may specialise in providing a specific resource or provide an integrated service. They help researchers by providing tools, resources and services, and stimulate them to adopt new methods by communicating what is successful.

Data-intensive research will thrive if these three communities are supported and collaborate. They will support a flood of new enthusiasts engaging in data-intensive research; bring with them tomorrow's research leaders. This will be helped by changes in education preparing their data and computational thinking skills.

S6 *Reflection and measurement* In a period of rapid change it is incumbent on the research community to observe, measure and record what is going on. They should access data from the corporate digital world to understand changes in society, health care, business, government and so on. Many branches of research will benefit from accessing that data. But scientists, particularly computer scientists should be collecting data about this research digital world for future researchers to examine how research responded to the digital revolution [28]. They should use that data to understand how the new data-intensive methods and collaboration techniques are affecting the way research is done — measuring whether the "intellectual velocity" is increasing, whether the methods do accelerate innovation, whether they make researchers more productive and whether they improve the quality of results. The methodology of empirical software engineering [109, 110, 111] may suggest good measurement strategies.

Chapter 9

Categories of data use

The world of data intensive-research is complex with many different requirements. A characterisation of this variation is introduced to facilitate discussion on where similarities may be expected and where differences should be respected; it is based on three features of data-intensive research:

1. the typical activities of the researchers and data curators;
2. the structural and organisational models employed; and
3. the nature of the phenomena that are being studied.

All three of these features show two forms of correlation: *a*) where the phenomena being studied have a scale or complexity that requires extensive collaboration, their community develops agreed goals, standards and practices to enable sharing and inter-working, and *b*) where the community has been using data for longer, there is greater understanding of their value and of the need to invest in pooled data facilities. For example, the scale of the LHC's data has led 2000 particle physicists to form a global data-management consortium. The intrinsically global nature and complexity of climate systems has led to a worldwide and multi-disciplinary collaboration. The 1971 agreement to share protein-structure data across the Atlantic has grown to a global consortium, wwPDB[1], curating and sharing many forms of data about more than 57,000 structures [55, 112, 57].

The following list characterises research-data collections by aspects of data use at any stage in the data life cycle. As we progress down the list the structure and organisation is increased and the ease with which researchers can change the data's structure and organisation is reduced. Two conflicting requirements have to be balanced: *a*) researchers want change to reflect new understanding and to support new research methods, and *b*) they want stability so that their large investment in software and procedures is not invalidated.

C1 *Informal local data* The majority of research data sits in this category, as many individuals and research groups build up data as they need it. It is often in files, informally organised and minimally described. It grows in complexity as each new research requirement emerges.

C2 *Structured local data* Researchers structure their data to use standard tools, to gain consistency and efficiency. They may use a standard data model, such as relational, XML or RDF, or semi-structured data [113].

[1] www.wwpdb.org

C3 *Informally published data* As the investment in data grows an individual or group may decide to make their research data available to others, either openly to any users or to specific collaborators. This is more valuable when the researchers are continuing to collect and organise the data than when they deposit it at the conclusion of their work. If the data is complex, large and well used this publication can be technically onerous, as illustrated by the Sloan Digital Sky Survey [28].

C4 *Community published data* There are many collections of data that are built by community effort, e.g. the records produced by ornithologists who walk the same route once a month for decades counting birds, the botanists who periodically count species in metre squares, and so on.

C5 *New shared repositories* When large research facilities are being built, e.g. LHC, Pan-STARRS, LSST, SKA, XFEL[2], a corresponding design for the data is normally undertaken, i.e. they build their data strategy on a 'green field' site. They may carry forward the use of legacy tools, but even these will be revamped for the new scales of data. This permits substantial advances in the methods of handling research data [73].

C6 *Reference repositories* Many research communities have established repositories of reference data; gathered from multiple sources and often professionally curated [114] and carefully structured. They develop a deposition and validation model to provide agreed assurances as well as standards compliance.

C7 *Federated reference repositories* Federated repositories are collaborations of reference repositories that conform to common standards to help researchers combine data from several of their resources. They often hold copies of each other's data to ensure reliability. wwPDB is a well established example, and CESSDA[3], CLARIN[4] and DARIAH[5] are projects to initiate such federations for social science, language and arts and humanities research respectively.

An almost orthogonal characteristic is the form of data recorded; as illustrated by the following examples.

F1 *Lists* of data, each about one instance of a class of entities. e.g. list of literature references that are to be examined, of sources of statistical that have been summarised, of recordings of dialogue with the place, time, participants and recording arrangements, or of medical patients involved in a research programme with associated genotypical, phenotypical and social context data. These lists allow human iteration and may support automated iteration. Their order may be significant.

F2 *Sets of independent events*, e.g. the records from detectors for each collision in a particle accelerator; order is not significant and query and iteration are supported over the data or over abstractions of each event with much less data per event can be used as proxies [25]. Some events, e.g. volcanic eruptions, earthquakes and extreme weather events, may be related through time and so temporal ordering may be significant for those that pertain to the same location.

F3 *Catalogues* of similar objects, e.g. catalogues of astronomic objects, of recorded musical performances, of known bio-molecules, of bio-informatic databases, etc. Catalogues may provide enough data for some researchers but they refer to the primary data to allow independent analysis. Gazetteers are often used to link names with geo-spatial data, e.g. EDINA's version of the

[2] www.xfel.eu
[3] www.cessda.org
[4] www.clarin.eu
[5] www.dariah.eu

Ordnance Survey 1:50,000 Gazatteer[6], which is combined with services for dealing with historical variation of names and boundaries, so that a document referring to a place can be automatically linked with it[7].

F4 *Time series* where data represent successive states of a phenomenon, such as the Antarctic ozone hole, a supernova, the positions of atoms during a chemical reaction, the languages spoken by a community or medical provision in a particular region and so on. When a time series is shorter than the life-time of the data or research project, e.g. a chemical reaction, then data are collected in the same form by the same mechanisms. When a time series exceeds the duration of a research programme, e.g. stellar evolution, climate change, development of language, evolution of life on earth, etc., researchers use a variety of proxies to reconstruct the historical record, e.g. for climate models: observation by meteorologists, ships logs, dendrochronology, sediment samples and ice cores. Arranging data collected in radically different ways in a coherent framework is difficult because the interpretation of data depends on models that may be revised.

F5 *Spatial data* where data represent information about spatially distributed phenomena, such as: our galaxy, earth systems or a living heart. Spatial relationships may fit in regular arrays or be irregular because of varying opportunities to collect data.

F6 *Spatio-temporal data* capture temporal variations of spatial phenomena, e.g. the propagation of seismic waves, the eruption of a volcano, the collision of two galaxies, the changes in temperature of an ocean, the electrical signals in a heart, and so on. An idea of the growing complexity of data to cover spatio-temporal phenomena is the requirement for eight data structures in addition to the extensive ISO geo-spatial standards to integrate ocean and climate research data [52].

F7 *Linked data* is data that is published according to a set of principals which greatly facilitate interpretation, re-use and integration, i.e. using URIs as names for things, making these HTTP URIs so that people can look them up, providing useful information in response to a lookup and including links to other URIs in order to find more information. The Linking Open Data project[8] aims to publish open data sets as RDF on the Web and set RDF links between data items from different data sources, so as to extend the Web with a data commons: the number of datasets us growing and many see this as an appropriate approach for open government data. The use of Linked Data for publishing scientific data is not yet widely established.

Many other aspects of a research community's use of data, e.g. their adoption of standards, their commitment to metadata and their interest in earlier data, need to be understood before commonalities can be recognised and essential differences registered.

[6] edina.ac.uk/digimap/description/products/gazetteer.shtml
[7] www.jisc.ac.uk/whatwedo/programmes/infrastructure/geocrosswalk.aspx
[8] esw.w3.org/topic/SweoIG/TaskForces/CommunityProjects/LinkingOpenData

Chapter 10

Proposed actions

For the UK, the EU and global society a step change in our skills and capacity for data-intensive research is necessary to address current strategic challenges: (*a*) a resilient economy, (*b*) green technology and energy, (*c*) health and wellbeing in an ageing society, (*d*) living with environmental change, (*e*) global uncertainties and security, and (*f*) food security. Every one of these has to deal with growing quantities of data and take into account more interactions and complexities in order to assess progress, steer decisions, design solutions and optimise operations. While driven by their own priorities they will quickly exploit cross-cutting improvements in our skill at using data. The same issues reappear at the local, regional and global scales in business, engineering, planing, health care, administration and entertainment. There will be a rapid growth and diversification of application science where data have to be analysed rapidly to support decisions. This will accelerate demand for data-intensive capabilities. A programme of actions is needed to trigger the necessary research, education and investment.

The long-term development of data-intensive research will be dynamically driven by international research competition, by interaction with the larger digital ecosystem and by normal governance. However, the data-use initiative calls for strong and decisive action to focus attention and resources. The following actions are indicative of the stimulus required and are put forward as a basis for further debate. They are oriented to UK research strategy but can easily be translated to more local or other national contexts.

A1 *Organise a workshop on data-intensive research in early 2011.* This workshop would bring together the UK's key players and stimulating international data-intensive research leaders; they would include: (*a*) digital curation experts (engaged with the help of the Digital Curation Centre), (*b*) data-driven research communities, including the UK's partners in ESFRI projects, (*c*) leading data-intensive research practitioners and (*d*) researchers addressing the foundations of data-intensive research. This two-day meeting would be held in a central UK location, e.g. The Royal Society, in 2011. The proceedings would identify UK requirements, capabilities and enthusiasm. Restoring the momentum of the UK's previous leadership in this field will be a crucial outcome.

Box 10.1

A2 *Activate a data-intensive educational programme.* A series of actions would stimulate curriculum change and encourage new thinking among educators.

- Initiate international and national summer schools and data-intensive ‘boot camps’ to stimulate networking and innovative thinking among early career, high-flying researchers.
- Enquire of each existing doctoral training centre (DTC) how its curriculum prepares its graduates for computational thinking and data-intensive work? High-light and fund the propagation of exemplars of best practice.
- Establish three new DTCs to pioneer data-intensive research, from its foundations to its application, that is of relevance to a broad range of disciplines. These DTCs should involve computer science, mathematical sciences, information science and researchers expert in the use of data. Between them, the three centres should link with a nearly all disciplines and deliver expertise relevant to all of the strategic challenges.
- Establish a working party representative of higher-education and employers to plan extensive injection of computational thinking and data-intensive skills in university courses.

A3 *Run an immersive ‘ideas factory’ event in late 2010 or early 2011* bringing together leaders in data-intensive research, mixing: fundamental research, pioneering ‘field’ experience and advances in data-oriented computational architectures, with commercial, industrial and governmental ‘customers’ for and ‘suppliers’ of data-intensive methods to identify and launch high-priority research projects.

A4 *Engage with current best practice.* This will use current advanced systems from research and commerce to rapidly develop (*a*) precise knowledge about their capabilities and limitations; (*b*) an injection of the skills in their use into the research and education communities; (*c*) clarity about what research is needed before they can be fully exploited; and (*d*) an entry into international collaborations that are developing the methods and technologies. The programme of engagements should develop knowledge of platform architectures, algorithms and data management, high-level languages, methods, operations and logistics for a representative spectrum of early adopters populating a matrix such as that shown below.

	Scientific DBs	MapReduce models	Commercial Clouds
Earth systems	Exploration of examples from earth-systems research including time series and spatio-temporal data	The same or different earth systems evaluations	The same or different earth systems evaluations
Biomedical	Exploration of handling sequence data, 2 or 3D images annotations linked with published literature and spatio-temporal data	The same or different biomedical research challenges	The same or different biomedical research challenges
Social science	Exploration of research using surveys, geo-spatial and network linkage graphs	The same or different social science challenges	The same or different social science challenges

Initial experiments should have specific capability goals, e.g. to handle feature extraction and derivation of target integrated data from a large set of functional MRI scans, using a specific technological strategy, such as Hadoop [100] and Pig Latin [103] or SciDB [77], so that the addition to our knowledge and interaction with the international collaborations can be rapidly achieved. Any group proposing to undertake such an experiment must already have access to the large body of data that will be used.

A5 *Stimulate researchers with Challenges* New data-intensive research should be triggered by a number of sharply defined and relevant calls; this might correspond to immediate challenges in a variety of disciplines or to cross-cutting issues. They should be formulated by the communities with challenges in close consultation with computing, mathematical and information scientists to ensure they are feasible for immediate action, if they are not they should feed into **A8**. With this advice calls should be specific about the mix of expertise required. The idea is illustrated with examples:

- Calls to the IPR legal experts to define a practical legal framework that encompasses all forms of data and delivers clarity about IPR flow as data and documents are combined from multiple sources.
- Calls to stimulate collaboration in discovering particular kinds of information from multi-sourced data, e.g. calls made by international consortia along similar lines to the *Digging into data challenge*[1] made jointly by the Joint Information Systems Committee (JISC) from the United Kingdom, the National Endowment for the Humanities (NEH)[2] from the United States, the National Science Foundation (NSF) from the United States, and the Social Sciences and Humanities Research Council (SSHRC)[3] from Canada.
- Calls to demonstrate the power of specific analytic techniques on currently pressing challenges. *Work in progress* mpa

These stimuli will help build capacity for larger and more demanding projects, help researchers build teams with effective mixes of skills and experience, and provide performance information for the selection of strategies, technologies and teams for larger projects.

A6 *Establish data-intensive research facilities.* The universities and organisations hosting research, the bodies funding research and JISC should establish data-intensive research facilities that complement provision for data generation, collection, curation and archiving and which will work well with existing data and computation services. These will include the following.

- Data storage and computation facilities that are close together and ideally integrated, c.f. GrayWulf [70] — balancing bandwidth and instruction rates [96].
- Standardised services for moving data between facilities, including large and long-distance transfers to connect major experimental and observational facilities, e.g. connecting ESFRI facilities for UK researchers [9], reference data services, international distributed computation services and specialised computation facilities.
- Standardised identity, authority, accounting and privacy services.
- Standardised services for archival storage providing virtual data spaces in which researchers and digital curators can organise and preserve data.
- Services for sustaining selected data-intensive software.
- Consultancy and training services to help researchers adopt data-intensive methods.

Some of these services already exist but may have to be adapted for data-intensive research and sustained to achieve researchers' trust.

A7 *Boost capacity of UK's reference data services* There are many existing reference data services (**C6** & **C7**) that are crucial for data-intensive research. They will need greater capacity and an increased range of facilities to drive and support the anticipated surge in data-intensive research,

[1] www.diggingintodata.org/Home/tabid/149/Default.aspx
[2] www.neh.gov/
[3] www.sshrc.ca/

as is already recognised in ESFRI projects, such as ELIXIR. This will require direct investment to build capacity and skills, and R&D to deliver the increased range of facilities.

A8 *Initiate a programme of research into data-intensive methods.* Build a collaborating and leading community in the UK by funding a network of computing and computational researchers, working with statisticians, mathematicians and numerical analysts to develop a theoretical and engineering platform that will ensure the long-term fidelity and capability of data-intensive research. Ensure that this community works closely with practitioners who are pioneering or applying data-intensive research methods and that they take into account relevant developments in the larger digital ecosystem, e.g. [58, 100, 103]. The research should cover the design and optimisation of data-intensive computational frameworks (**D1** & **X1**), the design and validation of data analysis algorithms (**D2**–**D8** & **X1.3**) and automated translation of these to incremental algorithms matched to the frameworks, the co-optimisation of data placement and task assignment, pioneering datascopes (chapter 6) and intellectual ramps (chapter 7), high-level languages, quality assurance (**X3**) and software composition methods (**X2**). This research may be facilitated by a common shared data-intensive research platform. Calls for such research have already been made in other countries, e.g. the NSF solicitations for DataNet projects[4] and the CISE cross-cutting programme on data-intensive computing[5].

A9 *Greener data-intensive computation* Today's digital communications, data storage and computers consume significant amounts of energy. Work in the corporate digital ecosystem will develop more energy efficient technology, software and operating practices. Many of the global data-intensive companies, driving the Web2.0 businesses, have demonstrated that they can reduce energy costs and move their data centres where the electricity is locally produced from renewable resources[6]. The data-intensive research community should engage in optimising their work, algorithms, software and operational practices so that they consume less energy. The research-services should data centres close to renewable energy sources, pooling the computational infrastructures to to achieve the energy and management economies achievable through cloud-computing practices [58]. The data transport and interaction with remote facilities should be based on all-optical networks, to reduce their energy requirement.

A10 *Establish a UK coordinating group.* Its principle role would be to improve communication and thereby accelerate developments and reduce unnecessary duplication of effort, e.g. it would identify opportunities to make substantial cost and environmental impact savings by combined provision (**A6**, **A7** & **A9**). It would act as a focus for UK interactions with international data-intensive research organisations arranging collaborative provision actions and agreeing inter-working practices. It would more than recover its costs by the savings achieved by encouraging the use of common infrastructure. It would improve researcher productivity and mobility by reducing unnecessary variation in provisions and by recognising common priority requirements.

Acknowledgements

The UK Engineering and Physical Sciences Research Council grant EP/D079829/1 has supported this work.

[4] www.nsf.gov/pubs/2007/nsf07601/nsf07601.htm

[5] www.nsf.gov/pubs/2009/nsf09558/nsf09558.htm

[6] These centres are operated with very few staff at the site.

NOTES FOR: PREPARING FOR THE DATA-INTENSIVE EVENT

This would build on past data-intensive workshops, including: *Data-Intensive Computing Symposium* — March 26, 2008 (research.yahoo.com/node/2104) and *Microsoft eScience Workshop* 15–17 October 2009 at which [7] was launched (research.microsoft.com/en-us/events/escience2009); on contemporaneous workshops: including: *Data-Intensive eScience Workshop* (DIEW) — 1–4 April 2010 in Tsukuba, Japan (www.diew2010.org) and on a series of UK preparatory workshops, including: *Data-Intensive Research: how should we improve our ability to use data* — 15–19 March 2010 e-Science Institute Edinburgh (XXX).

Box 10.1: Preparing for the Data-Intensive Event

Appendix A

Index of discussion points

This report is intended to encourage discussion, either to challenge, clarify or refine the suggestions and understanding so far developed, or to initiate R&D into a topic that has been introduced. To encourage such discussion points with potential for discussion have been identified with flags of the form **A1**. To aid readers in locating these discussion points, the following table is provided.

Flag	Topic	Section	Page
P1	Research in harmony with the digital ecosystem	2	7
P2	Increasing investment in using data	2	7
P3	Co-evolving practice, methods and software	2	7
P4	Education and access to data-intensive methods	2	7
P5	From theory to proof of concept to sustained facility	2	8
P6	Make researchers aware of costs	2	8
R1	Stimulate new thinking in next generation	3	9
R2	Creating and sharing methods and software	3	9
R3	Building intellectual ramps and education	3	10
R4	Theoretical research to underpin data-intensive research	3	10
R5	Path from proof of concept to production use	3	10
I1	Human communication in corporate and research worlds	5.5	19
I2	Human collaboration in corporate and research worlds	5.5	20
I3	Data analysis in corporate and research worlds	5.5	20
D1	Computational frameworks for datascopes	6	22
D2	Gathering relevant data from multiple sources	6	22
D3	Capturing relevant data in storage	6	22
D4	Preparing and cleaning data	6	22
D5	Cataloguing data	6	23
D6	Integrating and combining data	6	23
D7	Deriving data	6	23
D8	Presenting data	6	23
X1	Efficient data-analysis methods	6	24

continued on next page

continued from previous page

Flag	Topic	Section	Page
X1.1	Data-intensive hardware architectures	6	24
X1.2	Data-intensive software architectures	6	24
X1.3	Data-analysis algorithms	6	24
X1.4	Approximation	6	25
X2	Plugability of data-analysis elements	6	25
X3	Dependability of data-analysis facilities	6	25
X4	Sustainability of data-analysis facilities	6	25
M1	A specialised environment for safe play	7	28
M2	Augmenting a familiar research environment	7	28
M3	The methods ramp	7	29
S0	Theoretical research as a context for method innovation	8	33
S1	Creation of methods by research groups	8	32
S2	Transition from invention to production use	8	33
S3	Achieving sustained support for widespread use	8	33
S4	Applying data-intensive methods	8	34
S5	Computational contexts for data-intensive applications	8	34
S6	Observing and measuring the ways research is changing	8	34
C1	Informal local data: *ad hoc* and individual arrangements	9	35
C2	Structured local data: imposed quasi-regular structures	9	35
C3	Informally published data: sharing with collaborators	9	36
C4	Community published data: volunteers collect data	9	36
C5	New shared repositories: ‘green field’ data design	9	36
C6	Reference repositories: established & often curated	9	36
C7	Federated reference repositories: for scope & durability	9	36
F1	Lists of any kind of item	9	36
F2	Sets of independent events: regular for iteration	9	36
F3	Catalogues: summary data + references to primaries	9	36
F4	Time series: capturing sequence of states	9	37
F5	Spatial data: spatially distributed values	9	37
F6	Spatio-temporal data: combining F4 & F5	9	37
F7	Linked data standard describing & referencing content	9	37
A1	Data-intensive research workshops	10	39
A2	Education in data-intensive and computational thinking	10	39
A3	Immersive ideas creation factory	10	40
A4	Engagement with contemporary best practice	10	40
A5	Stimulate researchers with Challenges	10	41
A6	Data-intensive research facilities	10	41
A7	Boost UK provision for reference data services	10	41
A8	Theoretical foundations of data-intensive research	10	42
A9	Greener data-intensive research	10	42
A10	Coordination of data-intensive facilities	10	42

Appendix B

Mission statement

The following is the fact sheet sent to each potential host in preparation for our visits.

UK e-SCIENCE ENVOY MISSION TO THE US, SEPTEMBER 2009

Delegation

- Professor Malcolm Atkinson, UK e-Science Envoy[1]
- Professor David De Roure, University of Southampton[2].

Mission aim

The aim of the mission is to develop, refine and articulate a better understanding of the use and provision of research data, with a particular focus on how to improve the experience and success of the researchers who use research data in a cost-effective way. We hope to influence research-data strategies in the UK and beyond.

Objectives

The objectives during the fact-finding tour are:

1. To gather information on the developing practices and policies for exploiting research data with particular concern for the requirements of researchers who use or generate the data.
2. To enter into dialogue both with experts in the field of research data and with users of research data in order to refine our understanding and develop international connections.
3. To review strategies with researchers, experts from industry and policy makers to compare visions for the future of research data and to share ideas.

[1] The e-Science Envoy is an EPSRC appointment to develop advice on e-Science, engage in international e-infrastructure negotiations, and publicise and champion e-Science research covering all Research Council remitstinyurl.com/y9nb7w9.

[2] www.soton.ac.uk/d̃der/

4. To give seminars on the subject when requested by the groups being visited in order to stimulate dialogue and discussions.

Remit

Our remit covers the use of research data across all disciplines and application domains and we are concerned with the full gamut of users and producers, from those with small quantities of data of short-term and local interest to the very long-lived and very large- scale reference resources. We aspire to understand both the variation in requirements and the requirements common across large communities.

Actions following the mission

In preparation for the fact-finding tour we are preparing a draft of a position paper. After the tour, we will:

1. Complete and publish the paper in a respected and widely read publication.
2. Produce and make publicly available a full report
3. Make a presentation of our conclusions to the EPSRC Research Infrastructure Strategic Advisory Team (this body advises on provision of research infrastructure for EPSRC and influences computational provision across all of the RCUK).
4. Make a presentation to the JISC Support for Research (JSR) Committee, which steers provision of widely accessible research facilities across UK universities.
5. Make other presentations and targeted reports if requested, e.g. to other Research Councils and charitable bodies who support research.

We undertake to properly acknowledge all who are kind enough to find time for discussions with us and the organisations that host those discussions. We will also notify those who have helped when the paper and report become available.

Appendix C

Questions

The following list of questions evolved during the tour and was used to start discussions; copies were left behind with an invitation to send further answers.

1. Is the digital-data revolution beyond influence? If not, in what direction should we be trying to steer it? How should we do this?
2. Many more researchers could benefit from adroit use of data. How should we help them?
3. For success three factors must align: a) the users must find the new methods and technology useful, b) they must offer an intellectual access ramp P– easy to start and then acquire skill incrementally leading to the possibility of doing complex things, and c) have a persistent, affordable and feasible operational model.
   (a) How do you deliver that alignment for your community, technology or service?
   (b) What intellectual access ramps do you have in use?
   (c) What intellectual access ramps do you require?
4. How do you characterise your community's requirements? How much do they have in common with others? Where they are different, why are they different?
5. How many people in your community use data-intensive methods today? Do they share data and methods? How many could benefit from those methods? What limits the adoption rate?
6. In what ways are your communityUs digital data changing? What impact is that having on research and the methods used? What is planned? How does this differ from what should be done?
7. We cannot afford independent software stacks for every community.
   (a) What components and services do they share today?
   (b) What are good target components or services for shared provision?
   (c) With what processes and resources can these be built and sustained?
8. To what extent are you engaged in international collaboration over the use or provision of data (software)? Do you see collaborative opportunities that are being missed?
9. What are your plans for C_0 data services, i.e. with zero carbon footprint?
10. Do you see any requirements for changes in policy regarding data?

Appendix D

Visits and Contributions

The evidence for this report was collected over many years of working in applications of computers, during the e-Science programme, during the preparation of the Century of Information report [115] and in a fact-finding tour of the USA in September 2009. The latter provided the primary information and is reported below.

The following text is from Dave's blog http://blog.openwetware.org/deroure/ as a place holder for a fuller log. mpa

Boston

Tuesday 8th MIT. Hosted by Eric PrudUhommeaux and Philippe Le Hegaret in W3C, we caught up on standards including RDFa and HTML5 in the morning, gave a lunchtime talk to Carlo RattiUs SENSEable Cities lab and met with Sam Madden and Michael Stonebraker in the afternoon to learn about SciDB P •a project in serious danger of succeedingS. In the evening I went to a Semantic Web gathering, where Oshani Seneviratne presented her study of Creative Commons attribution violations.

Wednesday 9th Started the day at the British Consulate with Jacqueline Ashborne, Science and Innovation officer, and learnt about their help for visits and collaboration in research. John Willbanks of Science Commons kindly gave us a ride to Harvard, alerting us to the legal issues of derivative works in the context of Web and database queries. The rest of the day was hosted by Alyssa Goodman and Roslin Reid: we had a really interesting mix of meetings, including Pepi Fabbiano who participated in the excellent •Harnessing the Power of Digital Data for Science and SocietyS document. We inaugurated our mission and this yearUs IIC seminar series simulataneously with our first talk: •The DataQuestS.

Chicago

Thursday 10th Chicago. Met with Ian Foster and his Computation Institute colleagues in University of Chicago. Great new building and amazing seminar room which we also inaugurated

(IUve never seen so many projectors, pointing in different ways and even at each other!) and enjoyed a round table discussion. Great people and projects, from analysing news to systems biology. Our meeting over dinner with Ian and Steve Tuecke recalled the early days of the UK e-Science programme.

Friday 11th Chicago. Spent the morning at University of Illinois at Chicago with Bob Grossman, a whiteboard and caffeine talking about the Open Cloud Consortium and testbed. I applaud BobUs principle of using the minimum software necessary! Northwestern in the afternoon to meet with David Martin, Noshir Contractor and Jim Chen, where we also enjoyed a tour of the Starlight facility in all its amazing technicolour connectivity.

University of Michigan, Ann Arbor

Sunday 12th PMonday 13th. University of Michigan, hosted by Dan Atkins. Lots of really useful meetings, with particular relevance from an e-Social Science viewpoint. We had a roundtable discussion over pizza in the Daniel Atkins conference Room (in the same building that Arpanet was conceived!) and gave the next version of our talk. The day shone with interdisciplinarity and sophistication in e-Science thinking, from collaboratories to socio-technical design.

University of Wisconsin, Madison

Tuesday 14th. University of Wisconsin - Madison, hosted by Miron Livny, the creator of Condor. Miron shared his many insights into technology adoption with a clarity for which he is famed. Met the team, learned about HDFS in Condor as a SAN-alternative, and demoed Ian CottamUs very compelling Condor and Dropbox integration.

University of Illinois at Urbana-Champaign

Wednesday 15th. Early start at the University of Illinois at Urbana-Champaign where we met a group of scholars in the Center for Informatics Research in Science and Scholarship with a great understanding of the people side of the picture, and then to NCSA to give our talk and enjoy round tables on e-humanities (including eDream) and and e-science, all hosted by Jim Myers. We debated the Semantic Web! It was also a great chance to catch up on HASTAC.

University of New Mexico at Albuquerque

Thursday 16th- Friday 17th University of New Mexico at Albuquerque, hosted by Bill Michener. The first visit where everyone we met was completely focused on data! Our talk and discussion were in the library P we soon overcame our instinct to talk quietly, as more and more rows of seats were added at the back :-) This was a visit with an emphasis on production research data and it was impressive to see the balance of skills involved in its delivery.

Microsoft, Redmond

Monday 20th Microsoft. Hosted by Tony Hey, we spent the day in the Executive Briefing Centre being briefed executively and then briefly executing our talk. Significantly we intersected with another expedition P Prof Doug Kell of BBSRC and his officers, who were at the end of a similar tour. Check out the open source Word Add-in For Ontology Recognition and Creative Commons Add-in for Microsoft Office 2007 And check out DougUs blog too.

Stanford Linear Accelerator Center

Tuesday 21st SLAC. In the Stanford Linear Accelerator Center, Jacek Becla introduced us to his world of eXtremely Large DataBases and the XLDB events (the 3rd workshop was held recently). We enjoyed a demo of SciDb (see Tuesday 8th in a previous post!)

ISI, University of Southern California, Los Angeles

Wednesday 22nd ISI. A dynamic day of meetings (the meetings were dynamic and so was the schedule!) at the Information Sciences Institute with Yolanda Gil, Ewa Deelman, Ann Chervenak, Carl Kesselman and members of his team. Lots of examples of Computer Science coming to the aid of real users.

University of California, Irvine

Thursday 23rd Irvine and UCSD. Hosted by Paul Dourish at Irvine, we had a fascinating meeting with Gary Olson, collaboratory guru and one of the editors of Scientific Collaboration on the Internet. Richard Taylor and the work of Hazel Asuncion in traceability, workflows, and software architectures (software provenance is important too!);

National Center for Microscopy and Imaging Research, UCSD

Mark Ellisman of the National Center for Microscopy and Imaging Research at UCSD, where he spotted many microscopes and multiple ramps.

National Center for Atmospheric Research, Boulder

Friday 24th NCAR. Hosted by Don Middleton, we had a super day with his team in the National Center for Atmospheric Research in Boulder (at 5400 feet!) With serious attention to data management through its lifecycle, and delivery of tools as well as data, the day was full of examples of best practice P not just in data but in teamwork.

Washington DC

National Science Foundation

Bibliography

[1] Archimedes (collected by Thomas Venatorius — originally Thomas Gechauff). *Αρςημιμα̊ϵδους του Σψρακουσιου, Τα μϵςηρι νυν ςα̊οζομϵνα, ηαπαντα (The surviving complete works of Archimedes of Syracuse).* Basel; I. Hervagius, 1544.

[2] Archimedes. *The works of Archimedes, ed. in modern notation, with introductory chapters, by T. L. Heath.* Cambridge: University press, 1897.

[3] Dennis Noble. Cardiac action and pacemaker potentials based on the Hodgkin-Huxley equations. *Nature*, 188:495–7, November 1960.

[4] Otto Fred Hutter and Dennis Noble. Rectifying properties of heart muscle. *Nature*, 188:495, 1960.

[5] 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.

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

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

[8] IPCC. Fourth Assessment Report: Climate Change 2007 (AR4). Technical report, Intergovernmental Panel on Climate Change, 2007.

[9] ESFRI. European roadmap for research infrastructures: Roadmap 2008. Technical report, Office for Official Publications of the European Communities, 2008.

[10] Tim. O'Reilly. *What is Web 2.0: Design Patterns abd Business Models for the Next Generation of Software.* O'Reilly, 2005.

[11] EU Parliament. Directive 2007/2/EC of the European Parliament and of the Council of 14 March 2007 establishing an Infrastructure for Spatial Information in the European Community (INSPIRE). *Official Journal of the European Union*, 50(L108), April 2007.

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

[13] Cabinet Press Office, HM Government. PM welcomes Sir Tim Berners-Lee to Downing Street. 10 Downing Street News, September 2009.

[14] Allegra Stratton. Ordnance Survey maps to go free online. Newspaper item, Guardian, November 2009.

[15] David JC MacKay. *Sustainable Energy — Without the Hot Air*. UIT Cambridge Ltd, 2009.

[16] 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.

[17] Hans Rosling. Let my dataset change your mindset. YouTube, June 2009.

[18] Jeannette M. Wing. Computational thinking. *Commun. ACM*, 49(3):33–35, 2006.

[19] Peter B. Henderson, Thomas J. Cortina, and Jeannette M. Wing. Computational thinking. In Ingrid Russell, Susan M. Haller, J. D. Dougherty, and Susan H. Rodger, editors, *SIGCSE*, pages 195–196. ACM, 2007.

[20] Jeannette M. Wing. Computational thinking and thinking about computing. In *IPDPS*, page 1. IEEE, 2008.

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

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

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

[24] 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.

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

[26] 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.

[27] 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.

[28] 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.

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

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

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

[32] Jacek Becla and Kian-Tat Lim. Report from the third workshop on extremely large databases. Technical report, SLAC National Accelerator Laboratory, Menlo Park, CA 94025, USA, 2009 (in preparation).

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

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

[35] 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.

[36] DCC. The DCC Curation Life-Cycle Model. Technical report, Digital Curation Centre, 2008.

[37] Swan Alma and Brown Sheridan. To share or not to share: Publication and quality assurance of research data outputs. Technical report, Research Information Network (RIN), June 2008.

[38] Chris Anderson. The end of theory: The data deluge makes the scientific method obsolete. *Wired*, 16(07), June 2008.

[39] Carl Lagoze and Herbert Van de Sompel. Compound information objects: The oai-ore perspective, May 2007.

[40] Ben Shneiderman. Computer Science: Science 2.0. *Science*, 319(5868):1349–1350, 2008.

[41] Sandra Braman. What Do Researchers Need? Higher Education IT from the Researcher's Perspective. Technical Report ECP0601, ECAR, University of Wisconsin-Milwaukee, 2006.

[42] Michael Y. Galperin and Guy R. Cochrane. Nucleic Acids Research annual Database Issue and the NAR online Molecular Biology Database Collection in 2009. *Nucleic Acids Research*, 37(D1–D4), November 2009.

[43] ERC Scientific Council. Statement on Open Access. erc.europa.eu, December 2007.

[44] Power of Information Task Force. Power of Information Task Force Report. Technical report, Cabinet Office, February 2009.

[45] Cabinet Office. Interim Progress Report on implementing the Government's Response to the Power of Information Review. Technical Report Cm7157, UK Government, 2009.

[46] Gus O'Donnell. Information matters: building government's capability in managing knowledge and information. Technical report, HM Government, Cabinet Office, 200x.

[47] Power of Information Task Force. Modernise data publishing and reuse. Technical report, HM Government, Cabinet Office, 2009.

[48] Bryan Glick. Tim Berners-Lee gives first insight into government data plan. *Computing*, June 2009.

[49] OECD Follow Up Group on Issues of Access to Publicly Funded Research Data. Promoting Access to Public Research Data for Scientific, Economic, and Social Development. Technical report, OECD, 2003.

[50] Ewa Deelman, Dennis Gannon, Matthew S. Shields, and Ian Taylor. Workflows and e-science: An overview of workflow system features and capabilities. *Future Generation Comp. Syst.*, 25(5):528–540, 2009.

[51] Ewa Deelman and Ian Taylor, editors. *Proceedings of the 4th Workshop on Workflows in Support of Large-Scale Science, WORKS 2009, November 16, 2009, Portland, Oregon, USA*. ACM, 2009.

[52] Keiran Millar, Rob Atkinson, Andrew Woolf, Kristin Stock, Roger Longhorn, Chris Higgins, Mark Small, Sander Hulst, Torill Hamre, Maria Ferreira, Irene Lucius, Peter Breger, John Pepper, Dominic Lowe, Quillon Harphen, and Stephen Wells. Developing Feature Types and Related Catalogues for the Marine Community — Lessons from the MOTIIVE project. *International Journal of Spatial Data Infrastructures Research*, 2:132–162, 2007.

[53] J.D. Blower, S.C. Hankin, R. Keeley, S. Pouliquen, J. de la Beaujardière, E. Vanden Berghe, G. Reed, F. Blanc, Conkright, M. Gregg, J. Fredericks, and D. Snowden. Ocean data dissemination: New challenges for data integration. In J. Hall, D.E. Harrison, and D. Stam-

mer, editors, *Proceedings of OceanObs'09: Sustained Ocean Observations and Information for Society (Vol. 1), Venice, Italy, 21-25 September 2009*, number WPP-306. ESA, 2010.

[54] JD Blower, D Lowe, AL Gemmell, A Woolf, A Shaon, S Pascoe, K Millard, Q Harpham, and T Loubrieu. Harmonization of environmental data using the climate science modelling language. e-Science Centre, University of Reading, 2009.

[55] Helen M Berman, Kim Henrick, and Haruki Nakamura. Announcing the Worldwide Protein Data Bank. *Nature Structural Biology*, 10(12), December 2003.

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

[57] Henrick Kim, Feng Zukang, Bluhm Wolfgang F, Dimitropoulos Dimiyris, Doreleijers Jurgen F, Dutta Shuchismita, Flippen-Anderson Judith L, Ionides John MC, Kamada Chisa, Krissinel Eugene, Lawson CatherineL, Markley John L, Nakamura Haruki, Newman Richard, Shimizu Yukiko, Swaminathan Jawahar, Velankar Sameer, Ory Jeramia, Ulrich Eldon L, Vranken Wim, Westbrook John D, Yamashita Reiko, Yang Huanwang, Young Jasmine, Yousufuddin Muhammed, and Berman Helen M. Remediation of the protein data bank archive. *Nucleic Acids Research*, 36 (Database issue):D426–33, January 2008.

[58] Jeff Dean. Designs, lessons and advice from building large distributed systems. Keynote at Large-Scale Distributed Systems and Middleware (LADIS), October 2009.

[59] 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.

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

[61] Alexander S. Szalay. Observation that emails contain most of the design discussion during the SDSS. Personal communication, September 2009.

[62] Jeremy G. Frey. Logs, blogs and pods: smart electronic laboratory notebooks, 2009.

[63] Mark H. Ellisman, T. Hutton, A. Kirkland, Abel W. Lin, C. Lin, T. Molina, Steven. Peltier, Rajvikram. Singh, K. Tang, Anne.E. Trefethen, David.C.H. Wallom, and 12. X. Xiong1. The OptIPuter microscopy demonstrator: enabling science through a transatlantic lightpath. *Philisophical Transactions of the Royal Society A*, 367(1898):2645–2653, July 2009.

[64] Herbert Van de Sompel and Carl Lagoze. All Aboard: Toward a Machine-Friendly Scholarly Communication System. In Tony Hey, Stewart Tansley, and Kristin Tolle (Editors), editors, *The Fourth Paradigm: Data-Intensive Scientific Discovery*, pages 193–199. Microsoft Research, October 2009.

[65] David De Roure and Carole Goble. Research Objects for Data Intensive Research. In *Proceedings of the Microsoft e-Science Conference*. Microsoft, October 2009.

[66] D. De Roure, C. Goble, and R. Stevens. The Design and Realisation of the myExperiment Virtual Research Environment for Social Sharing of Workflows. *Future Generation Computer Systems*, 25:561–567, 2009.

[67] Adam Jacobs. The pathologies of big data. *Commun. ACM*, 52(8):36–44, 2009.

[68] Peter Z Kunszt, Alexander S Szalay, and Anniruddha R Thakar. The Hierarchical Triangular Mesh. In *ESO Astrophysics Symposia: Mining the Sky*, pages 631–637. Springer, 2001.

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

[70] 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* [116], pages 1–10.

[71] Uwe Röhm and José A. Blakeley. Data management for high-throughput genomics. In *CIDR* [117].

[72] Michael Armbrust, Armando Fox, David A. Patterson, Nick Lanham, Beth Trushkowsky, Jesse Trutna, and Haruki Oh. Scads: Scale-independent storage for social computing applications. In *CIDR* [117].

[73] Jacek Becla, Andrew Hanushevsky, Sergei Nikolaev, Ghaleb Abdulla, Alexander S. Szalay, María A. Nieto-Santisteban, Aniruddha R Thakar, and Jim Gray. Designing a multi-petabyte database for LSST. *CoRR*, abs/cs/0604112, 2006.

[74] Jacek Becla and Kian-Tat Lim. Report from the SciDB workshop. Technical report, Stanford Linear Accelerator Center, Menlo Park, CA 94025, USA, 2008.

[75] Andrew Pavlo, Erik Paulson, Alexander Rasin, Daniel J. Abadi, David J. Dewitt, Samuel Madden, and Michael Stonebraker. A comparison of approaches to large-scale data analysis. In *SIGMOD '09: Proceedings of the 2009 ACM SIGMOD International Conference*. ACM, June 2009.

[76] Philippe Cudré-Mauroux, Hideaki Kimura, Kian-Tat Lim, Jennie Rogers, Roman Simakov, Emad Soroush, Pavel Velikhov, Daniel Wang, Magdalena Balazinska, Jacek Becla, David J. DeWitt, Bobbi Heath, David Maier, Samuel Madden, Jignesh M. Patel, Michael Stonebraker, and Stanley B. Zdonik. A demonstration of scidb: A science-oriented dbms. *PVLDB*, 2(2):1534–1537, 2009.

[77] Michael Stonebraker, Jacek Becla, David J. DeWitt, Kian-Tat Lim, David Maier, Oliver Ratzesberger, and Stanley B. Zdonik. Requirements for science data bases and scidb. In *CIDR* [117].

[78] Yogesh Simmhan, Catharine van Ingen, Roger Barga, Alexander S. Szalay, and Jim Heasley. Reliable Management of Community Data Pipelines using Scientific Workflows. Technical Report MSR-TR-2009-125, Microsoft Research, September 2009.

[79] Jim Gray. Jim Gray on eScience: A Transformed Scientific Method. In Tony Hey, Stewart Tansley, and Kristin Tolle (Editors), editors, *The Fourth Paradigm: Data-Intensive Scientific Discovery*, pages xix–xxxiii. Microsoft, 2009.

[80] R Venkatesh, Dennis Y Altudov, B Sezgin, and JA Blakeley. Partial deserialization of complex type objects. Technical Report US7,555,506 B2, US Patents Office, June 2009.

[81] 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.

[82] Guy Cochrane, Ruth Akhtar, Philippe Aldebert, Nicola Althorpe, Alastair Baldwin, Kirsty Bates, Sumit Bhattacharyya, James Bonfield, Lawrence Bower, Paul Browne, Matias Castro, Tony Cox, Fehmi Demiralp, Ruth Eberhardt, Nadeem Faruque, Gemma Hoad, Mikyung Jang, Tamara Kulikova, Alberto Labarga, Rasko Leinonen, Steven Leonard, Quan Lin, Rodrigo Lopez, Dariusz Lorenc, Hamish McWilliam, Gaurab Mukherjee, Francesco Nardone, Sheila Plaister, Stephen Robinson, Siamak Sobhany, Robert Vaughan, Dan Wu, Weimin Zhu, Rolf Apweiler, Tim Hubbard, and Ewan Birney. Priorities for nucleotide trace, sequence and annotation data capture at the Ensembl Trace Archive and the EMBL Nucleotide Sequence Database. *Nucl. Acids Res.*, page gkm1018, 2007.

[83] Guy Cochrane, Ruth Akhtar, Philippe Aldebert, Nicola Althorpe, Alastair Baldwin, Kirsty Bates, Sumit Bhattacharyya, James Bonfield, Lawrence Bower, Paul Browne, Matias Castro, Tony Cox, Fehmi Demiralp, Ruth Eberhardt, Nadeem Faruque, Gemma Hoad, Mikyung Jang, Tamara Kulikova, Alberto Labarga, Rasko Leinonen, Steven Leonard, Quan Lin, Rodrigo Lopez, Dariusz Lorenc, Hamish McWilliam, Gaurab Mukherjee, Francesco Nardone, Sheila Plaister, Stephen Robinson, Siamak Sobhany, Robert Vaughan, Dan Wu, Weimin Zhu, Rolf Apweiler, Tim Hubbard, and Ewan Birney. Priorities for nucleotide trace, sequence and annotation data capture at the Ensembl Trace Archive and the EMBL Nucleotide Sequence Database. *Nucleic Acids Research*, 36(Database Issue):D5–D12, January 2008.

[84] Cochrane G, Akhtar R, Bonfield J, Bower L, Demiralp F, Faruque N, Gibson R, Hoad G, Hubbard T, Hunter C, Jang M, Juhos S, Leinonen R, Leonard S, Lin Q, Lopez R, Lorenc D, McWilliam H, Mukherjee G, Plaister S, Radhakrishnan R, Robinson S, Sobhany S, Hoopen PT, Vaughan R, Zalunin V, and Birney E. Petabyte-scale innovations at the European Nucleotide Archive. *Nucleic Acids Research*, 37(Database Issue):D19–25, January 2009.

[85] Wenfei Fan, Floris Geerts, Xibei Jia, and Anastasios Kementsietsidis. Conditional functional dependencies for capturing data inconsistencies. *ACM Trans. Database Syst.*, 33(2), 2008.

[86] Wenfei Fan. Dependencies revisited for improving data quality. In Maurizio Lenzerini and Domenico Lembo, editors, *PODS*, pages 159–170. ACM, 2008.

[87] Wenfei Fan, Floris Geerts, and Xibei Jia. Conditional dependencies: A principled approach to improving data quality. In Alan P. Sexton, editor, *BNCOD*, volume 5588 of *Lecture Notes in Computer Science*, pages 8–20. Springer, 2009.

[88] Yogesh Simmhan, Catharine van Ingen, Roger Barga, Alex Szalay, and Jim Heasley. Building Reliable Data Pipelines for Managing Community Data using Scientific Workflows. In *Proceedings of the IEEE e-Science Conference*, Oxford, December 2009.

[89] 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.

[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] J.D Blower, K Haines, A Santokhee, and C.L Liu. GODIVA2: interactive visualization of environmental data on the Web. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 367(1890):1035–1039, 2009.

[92] Lundin Karljohan E, Cooper Matthew, and Ynnerman Anders. Haptic interaction with dynamic volumetric data. *IEEE Transactions on Visualization and Computer Graphics*, 14(2):263–276, 2008.

[93] Lorna Richardson, Shanmugasundaram Venkataraman, Peter Stevenson, Yiya Yang, Nicholas Burton, Jianguo Rao, Malcolm Fisher, Richard A. Baldock, Duncan R. Davidson, and Jeffrey H. Christiansen. EMAGE mouse embryo spatial gene expression database: 2010 update. *Nucl. Acids Res.*, page gkp763, 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] 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* [116], pages 1–10.

[96] Gordon Bell, Jim Gray, and Alexander S Szalay. Petascale computational systems: balanced cyberinfrastructure in a data-centric world. *IEEE Computer*, 39(1):110–12, 2006.

[97] Sanjay Ghemawat, Howard Gobioff, and Shun-Tak Leung. The google file system. In Michael L. Scott and Larry L. Peterson, editors, *SOSP*, pages 29–43. ACM, 2003.

[98] Jeffrey Dean and Sanjay Ghemawat. Mapreduce: simplified data processing on large clusters. *Commun. ACM*, 51(1):107–113, 2008.

[99] Fay Chang, Jeffrey Dean, Sanjay Ghemawat, Wilson C. Hsieh, Deborah A. Wallach, Michael Burrows, Tushar Chandra, Andrew Fikes, and Robert E. Gruber. Bigtable: A distributed storage system for structured data. *ACM Trans. Comput. Syst.*, 26(2), 2008.

[100] Tom White. *Hadoop: The Definitive Guide*. O'Reilly, 2009.

[101] Robert L. Grossman and Yunhong Gu. Data mining using high performance data clouds: experimental studies using sector and sphere. In Ying Li, Bing Liu, and Sunita Sarawagi, editors, *KDD*, pages 920–927. ACM, 2008.

[102] Jim Gray, María A. Nieto-Santisteban, and Alexander S. Szalay. The zones algorithm for finding points-near-a-point or cross-matching spatial datasets. *CoRR*, abs/cs/0701171, 2007.

[103] Christopher Olston, Benjamin Reed, Utkarsh Srivastava, Ravi Kumar, and Andrew Tomkins. Pig latin: a not-so-foreign language for data processing. In Jason Tsong-Li Wang, editor, *SIGMOD Conference*, pages 1099–1110. ACM, 2008.

[104] Chris Miceli, Michael Miceli, Shantenu Jha, Hartmut Kaiser, and Andre Merzky. Programming Abstractions for Data Intensive Computing on Clouds and Grids. In *Proceedings of CCGrid2009*, 2009.

[105] Gary M. Olson, Ann Zimmerman, and Nathan Bos, editors. *Scientific Collaboration on the Internet*. MIT Press, November 2008.

[106] Peter Li, Tom Oinn, Stian Soiland, and Douglas B. Kell. Automated manipulation of systems biology models using libsbml within taverna workflows. *Bioinformatics*, 24(2):287–289, 2008.

[107] Elizabeth van der Meer, Malcolm Atkinson, David Fergusson, Lorna Hughes, Elphini. Fragkouli, Alexander Voss, S. Anderson, M. Asgari-Targhi, Rob Procter, and Peter Halfpenny. Adoption of e-infrastructure services: findings, opportunities and issues. In *5th International Conference on e-Social Science*, Cologne, Germany, 06/2009 2009.

[108] Neil Chue Hong. Software Sustainability: Looking Past the Myths. NeSC Public Lecture, November 2009.

[109] Barry W Boehm, H Dieter Rombach, and Marvin V Zelkowitz, editors. *Foundations of Empirical Software Engineering: The Legacy of Victor R. Basili*. Springer, 1998.

[110] Forrest Shull, Janice. Singer, and Dag IK Sjøberg, editors. *Guide to Advanced Empirical Software Engineering*. Springer, 2007.

[111] Dag I. K. Sjøberg, Tore Dybå, and Magne Jørgensen. The future of empirical methods in software engineering research. In Lionel C. Briand and Alexander L. Wolf, editors, *FOSE*, pages 358–378, 2007.

[112] Markley John L, Ulrich Eldon L, Berman Helen M, Henrick Kim, Nakamura Haruki, and Akutsu Hideo. BioMagResBank (BMRB) as a partner in the Worldwide Protein Data Bank (wwPDB): new policies affecting biomolecular NMR depositions. *J Biomol NMR*, 40(3):153–5, 2008.

[113] Serge Abiteboul, Peter Buneman, and Dan Suciu. *Data on the Web: From Relations to Semistructured Data and XML*. Morgan Kaufmann, 1999.

[114] Constantopoulos, Panos and Dallas, Costis and Androutsopoulos, Ion and Angelis, Stavros and Gavrilis, Deligiannakis, Antonios and Gavrilis, Dimitris and Kotidis, Yannis and papatheodorou, Christos. DCC and U: An extended digital curation lifecycle model. *The International Journal of Digital Curation*, 4(1), 2009.

[115] Atkinson Malcolm, Britton David, Coveney Peter, De Roure David, Garnett Ned, Geddes Neil, Gurney Robert, Haines Keith, Hughes Lorna., Ingram David, Jeffreys Paul, Lyon Liz, Osborne Ian, Perrott Ron, Procter Rob, Rusbridge Chris, Trefethen Anne, and Watson Paul. Century-of-information research (cir): A strategy for research and innovation in the century of information. Technical report, The e-Science Directors' Forum Strategy Working Group, 2008.

[116] *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.

[117] *CIDR 2009, Fourth Biennial Conference on Innovative Data Systems Research, Asilomar, CA, USA, January 4-7, 2009, Online Proceedings*. www.crdrdb.org, 2009.