1

The Digital-Data Challenge

Malcolm Atkinson
School of Informatics, University of Edinburgh, Edinburgh, UK

Mark Parsons
EPCC, University of Edinburgh, Edinburgh, UK

This chapter is intended to interest all readers and to set the scene for the rest of the book. It reviews the current, rapidly growing wealth of data and our abilities to take best advantage of the potential knowledge that this data bonanza offers.

1.1 THE DIGITAL REVOLUTION

There is evidence all around us that the world is going through the early stages of a digital revolution; it is more stressful than the Industrial Revolution, as it is impacting virtually every nation simultaneously. From digital photographs to digital music and from electronic tax returns to electronic health records, this revolution is affecting all of us in many aspects of our lives. 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.

The DATA Bonanza: Improving Knowledge Discovery in Science, Engineering, and Business, First Edition.
Edited by Malcolm Atkinson, Rob Baxter, Michelle Galea, Mark Parsons, Peter Brezany, Oscar Corcho, Jano van Hemert, and David Snelling.

This is a far more stressful revolution than those brought about by the previous advances in communication: speech, writing, telecommunications, and broadcasting, and by the Industrial Revolution, because it is so rapid and, unlike previous revolutions, it is penetrating every corner of the globe simultaneously, thanks to the Internet. Survival in a revolution depends on rapid and appropriate adaptation to the changes it brings. The successful survivors will be those people, organizations, and nations who are most adept at adapting to change.

Today's global challenges are 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, business, and diagnosis. Data underpin scholarship and understanding. Data fuel analysis to produce key evidence for business decisions and supply the information for compelling communication. Data connect computational systems and enable global collaborative endeavors.

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, the public and private sectors need novel infrastructure and tools that are as convenient, pervasive, and powerful as the Web 2.0 environment. While the Web facilitates communication for business and personal use, the new data-intensive infrastructure will facilitate the analysis of enormous volumes of data from a wide range of sources. This will empower new ways of thinking about the decisions that individuals and society face. The questions that the users will ask of data and the capabilities of the data-intensive facilities will coevolve as the new power stimulates new approaches to the challenges—both large and small—we face in our lives. Enabling that widespread access for every individual, group, and organization to explore data in the search for better knowledge will have profound effects.

These changes in the use of data will require radical changes in our working behavior and infrastructure provision. We advocate a collaborative endeavor to achieve this potential, exploring the new types of behavior, methods, computational strategies, and economies of provision. Data should be used fluently in research (be it business or academic), investigation, planning, and policy formulation to equip those responsible with the necessary information, knowledge, and wisdom. *The present cornucopia of data is underexploited.* Many people are aware of this; few understand what to do. The purpose of this book is to help the readers understand how to navigate this digital revolution and benefit from it in all aspects of their lives.

1.2 CHANGING HOW WE THINK AND BEHAVE

Ever since humans invented writing at the beginning of the Bronze Age, around 7000 years ago, we have created persistent data. Before the advent of the computer, data were transferred from generation to generation in the form of written, or latterly printed, books or other documents. Since the advent of the digital computer, and, in particular, since the start of the third millennium, the amount of recorded data has exploded. All of the books and documents written by man over the past 7000 years represent a tiny fraction of the data stored on computers at present.

Digital data are now a universal glue; they carry radio-frequency identification (RFID) messages, represent cartographic and satellite images, form SMS and email messages, encode mobile phone traffic, facilitate social networks, such as Facebook and Twitter, and enable the transactions of everyday life: booking a hotel, paying an account, planning a journey, and arranging to board an aircraft. Indeed, the rapid flow of data between individuals and organizations has changed the form of these interactions; at present, people expect to make detailed choices from offers shaped for them and get almost instantaneous responses when they interact with utilities, businesses, and government. Data capture our personal images in digital cameras and videos, carry films through each stage of production and transmission to market, represent books and newspapers, and are increasingly the medium of broadcasting. Data are the primary product of medical-imaging systems, satellite observations, astronomic surveys, microscopes, and experiments in nearly every field of science. Data capture inventories, accounts, plans, and conducted processes. Data are used to represent documents from initial drafting to final publication and production; today more and more are published in digital form. Data are the result of all simulation model runs and of many design studies. Data are the medium and product of social computation.

Managing the data deluge and using it to our advantage, whether that be for geopolitical planning purposes or a more mundane task such as buying a new car, requires significant changes in our behavior and in the information and communication technology (ICT) infrastructures that support the curation, management, and provision of the data we need. As data volumes increase exponentially, incremental changes will no longer be sufficient. We need to establish new working practices and new models of computing and infrastructure provision if we are to see the benefits of these huge quantities of data and the opportunities they bring.

Digital data are changing the way we think and how we behave. With information always available at our fingertips, for instance, through Google or Wikipedia, we are not necessarily better informed. We live in a world of instant information—far too much information for us to cope with. Many people believe the information that is presented without questioning its veracity; they only use a minuscule percentage of the data available when making decisions. Two issues must be addressed: better presentation of high quality aggregations of the data and accessible clarity about that presentation's origins and validity.

The goal is to accelerate and facilitate *knowledge discovery* from the growing wealth of data. For this, every step from finding and gaining access to the data, through information extraction to knowledge delivery, must be technically well supported. These steps should be as accessible to the individuals as Web 2.0 searches are at present. But the precision must be trustworthy and well understood, and the knowledge must be delivered to each person in a form that suits their needs. There are still many other issues to consider in any knowledge discovery project, the largest of these almost certainly being the social issues related to data ownership and how people react to requests for projects to use "their" data. Such socioeconomic considerations are not within the scope of this book, but readers should be aware of the complexity of the challenges posed by them—see, for example, Section 1.6 of [1]. Effective security mechanisms are needed to underpin these social contracts.

1.3 MOVING ADROITLY IN THIS FAST-CHANGING FIELD

The future will contain many innovations in our day-to-day behavior and the technology that supports it. Some data will remain specific to a community, business, or facility; some will be constrained by ethical and legal considerations. Many individuals and organizations will make use of information produced by large swathes of society.

The fundamental challenge that this book addresses is how to succeed in an age where individual people or organizations can no longer cope using manual methods of transforming data into information, to understand issues and to make decisions. Those who learn to move adroitly and embrace the new technologies that are being developed to support the analysis of data are those who will thrive in the digital age.

In addition to showing how these new technologies can benefit individual people or organizations, in this book, we propose a path intended to lead to more collaborative behavior, which will also drive and draw on innovation in other data-intensive contexts. For example, many data producers, be they separate organizations or separate divisions within a large organization, could deposit data in *shared data clouds* that would also be populated with reference and legacy data. When undertaking research to inform scientific enquiry or making business decisions, users will then deposit data and pose questions that compose and compare data from any contributing source. Legal and ethical regimes will become even more important to guide and control this use, balancing the concern for privacy with society's and the individual's needs for the best-quality decisions.

1.4 DIGITAL-DATA CHALLENGES EXIST EVERYWHERE

It is a common misconception that we are in the middle of the digital revolution. We are almost certainly only at the beginning. The sudden growth of digital-data collections is outstripping our ability to manually cope with the opportunities the creation of these collections bring. Obvious examples of large data collections are the enormous indexes of Web search and derived data created by online search engines such as Google. Companies such as eBay and Amazon handle prodigious volumes of data to support their global business with very large numbers of companies and individuals dependent on their services, but they do not present external users with opportunities to extract business information from their wealth of data, which is an in-house asset. Some of the largest data collections are being created and being made widely available in the scientific domain. Examples of types of scientific data collections include the following:

- the streams of sensor data from an extensively instrumented natural environment,
- the set of all gene sequences for all organisms with their annotations,
- spatiotemporal images of biological systems from within the cell, via organs and individuals, to ecosystems, and
- meteorological and oceanographic records.

Some datasets are already very large and contain multiple petabytes of data.[1] Many are interconnected, as data in one collection references related items in other collections.

In the business context, although to date the quantities of data are usually somewhat smaller than those addressed by science, the digital-data challenges are often more complex, involving the management of complex hierarchies of trust and security in order to access, integrate, and derive business information. For instance, in the coming few years, a prerequisite for any large business will be the creation of a data cloud shared across the organization from sources as diverse as operations, manufacturing, human resources, finance, and sales. The use of this data cloud in conjunction with the growing bodies of public and purchasable data will be a key business tool driving decision making. A better grasp of the changing business context will allow companies to respond to new opportunities and difficult trading conditions, enabling them to survive downturns and be ready to respond when economies improve. As the value of data sharing is seen to drive business success, it will extend beyond individual organizations and include trusted third parties such as key suppliers. This will be a profound change in the way that businesses respond to digital-data-driven information.

Likewise, in the scientific research space, having a data cloud shared by environmental and Earth-systems scientists, economic and social scientists, medical researchers and diagnosticians, engineers and physical scientists, and biological and life scientist's will facilitate the interdisciplinary compositions of data and open up research questions that would be infeasible otherwise.

In practice, at present, those who own or can access data cannot ask such questions because the data sources are organized for one class of questions and each collection is isolated, that is, without explicit and reliable links to support comparison or augmentation from other sources.

1.5 CHANGING HOW WE WORK

We face many challenges in this new era of data-intensive computing; the following examples often reinforce one another:

1. coping with the increasing volume and complexity of data, and the increasing rate of data or document deposition;

[1] A petabyte is 10^{15} or 1,000,000,000,000,000 bytes. See Appendix A for the glossary of terms.

2. accommodating the rapid growth in the number of people who want to make use of emergent data services and their rising expectations;
3. balancing the pressures to accommodate new requirements and information structures with the requirement to protect existing investment and limiting the costs.

With the advent of data-intensive computational strategies (discussed later), a much larger class of questions about large or complex data can be answered economically. The scientific domain has been the first domain to see the opportunities that the analysis of large or complex datasets can bring. At present, several large instruments, observatories, and reference data collections are each beginning to use computational systems optimized for data-intensive operations. These changes in working practices are being driven by the scientific research community. Just as the Internet was initially developed by the scientific community, business and other user communities will rapidly identify the benefits and adapt data-intensive technology for their own purposes. This will become a dominant data management strategy used by both business and science over the next decade. Succeeding will mean changing how we work, moving from manual to automatic methods, and accepting that we can no longer survive by doing it all ourselves.

The data-intensive cloud computing experts and database experts have, in the past decade, become adept at handling more and more forms of computational query and analysis by using carefully chosen computing hardware and software architectures. The latest strategies include new inventory-based forms of consistency between data centers [2] and use human input to fill gaps needed to answer a query [3]. We propose that researchers embrace the opportunity brought about by these and other technical advances to blaze a path toward the vision of easily accessed and composed data that they can fluently exploit for their work. This will involve the development of new types of behavior, frameworks, methods, and computational systems. Each step will deliver new capabilities, boosting knowledge and informing decisions.

However, producing new technology and adopting new methods of working are not enough. To succeed, we need to develop the skills needed in every walk of life to extract knowledge from data. Viewed as a whole, these problems are daunting, perhaps insurmountable. To succeed, we must adopt a *divide-and-conquer* approach and partition the challenges we face.

1.6 DIVIDE AND CONQUER OFFERS THE SOLUTION

This book explains how knowledge discovery can be simplified by partitioning the conceptual and design challenges—an approach that works well in virtually all data-intensive contexts. We show how each data-intensive community can be partitioned into three groups of experts. First, the *domain experts* who try to better understand how to use the data in their domain. Second, the *data analysis experts* who try to develop ways of representing and visualizing data and who create and refine the algorithms that extract information from that data. Third, the *data-intensive engineers* who develop the tools and systems required for data-intensive computing.

1. *Domain experts*: These experts work in the particular domain where the knowledge discovery from their data takes place. They pose the questions that they want to see their data answer, such as, “How can I increase sales in this retail sector?”, “What are the correlations in these gene expression patterns?”, or “Which geographic localities will suffer most in times of flooding?”. They are aware of the meaning of the data that they use. They are presented with data-intensive tools, optimized and tailored to their domain by the following two groups of experts. Many of the technical and operational issues are hidden from them.
2. *Data analysis experts*: These experts understand the algorithms and methods that are used in knowledge discovery. They may specialize in supporting a particular application domain. They will be supported with tool sets, components, and workbenches, which they will use to develop new methods or refactor, compose, and tune existing methods for their chosen domain experts. They are aware of the structures, representations, and type systems of the data they manipulate.
3. *Data-intensive engineers*: These experts focus on engineering the implementation of the data-intensive computing platform. Their work is concerned with engineering the software that supports data-intensive process enactment, resource management, data-intensive platform operations, and language implementations. They deliver libraries of components, tools, and interfaces that data analysis experts will use. A data-intensive engineer’s role includes organizing the dynamic deployment, configuration and optimization of the data’s movement, and storage and processing as automatically as possible. They also provide well-organized libraries of common data-handling and data-processing components.

We return to the task of understanding how to support all three categories of data-intensive expert in Section 3.2. Partitioning users is only one aspect of dividing the problem. We must also divide each data-intensive task so that computational issues can be conquered. All too often, knowledge discovery projects run into difficulties because the focus of the project team is only on one aspect of the computational issues and others are ignored. In this book, we describe how the computation should be partitioned into three levels of abstraction in order to deliver effective user experiences. These levels are introduced here and developed in Chapter 3.

1. *Tools* are used to create the data-intensive processes to analyze the data and to create prepackaged solutions to particular data-intensive knowledge discovery tasks.
2. *Language* is needed to clearly express the queries and manipulations required to explore the various data sources during knowledge discovery.

3. *Enactment* sits below the language and tools and is controlled using the tools and programmed using the language.

Of crucial importance is the realization that partitioning the computational challenges into these layers and partitioning those involved in knowledge discovery projects into the three roles described above allows the overall data-intensive challenge to be broken down into manageable pieces. These data-intensive partitions, both people focused and computation focused, mean that each application's domain, for example, business, commerce, healthcare, or biology, can innovate and respond independently with relatively little knowledge of the underlying methods and technology.

Solutions in each application's domain build on a common expanding infrastructure being built by the data analysis experts and the data-intensive engineers, who work to increase the power of the libraries of compatible data-processing elements, of the predefined patterns for composing common solution elements, and of the data-intensive platforms that are available to all application domains, obviating the need for bespoke solutions on every occasion. This means that data analysis developers get data-handling power from data-intensive engineering, and data-intensive engineers are able to invest in good, well-optimized solutions that are used multiple times. This achieves a transition, moving from the present practice of large and highly skilled teams making heroic efforts to extract knowledge from data, to an environment where individuals skilled in their domain can use knowledge discovery methods directly.

1.7 ENGINEERING DATA-TO-KNOWLEDGE HIGHWAYS

As the digital revolution progresses, people will need to make data journeys; getting data from wherever it may be, in whatever form it is, combining, transforming, and analyzing the data, to produce information-bearing derivatives. Those derivatives have to be transformed into just the right form to communicate the relevant knowledge and be delivered to the right person at the right time, so that person may act on the new knowledge. To facilitate many people making many such journeys, we need data-to-knowledge highways. Traditional highway engineers have expertise in identifying what new highways should be built, working out where they should go to and how they should connect with the existing network of transport infrastructure, how they should be built with minimum cost and disruption, and how they should be operated sustainably. At present, we need to nurture a new breed of *data-to-knowledge highway engineers*. They will develop professional methods and skills including the following:

1. analysis of existing data-to-knowledge journeys and future requirements to see where and when new data-to-knowledge highways should be built;

2. design of those highways so that they serve the maximum number of useful data paths, from identified clusters of source data to identified clusters of knowledge delivery destinations;
3. organization of data-highway construction, calling on the data equivalent of civil engineers, namely, database providers, network providers, storage providers, computation providers; visualization services; and interconnections with all manner of other data and communication services;
4. establishment of the guidance and control mechanisms, "data GPS and traffic lights," to enable data-to-knowledge journeys to start and finish wherever they need, traveling effortlessly by taking full advantage of the evolving data-to-knowledge infrastructure.

All of the three foci of expert interest (domain expertise, data analysis expertise, and data-intensive engineering) have to be satisfied by a data-to-knowledge highway. Consequently, data-to-knowledge highway engineers have to understand all three points of view, take a holistic view, and balance the interests of these three categories of professional data-to-knowledge highway user. An operational data-to-knowledge highway should have the following properties:

1. Domain users are happy because the on and off ramps match their journeys and extra lanes prevent congestion.
2. The data analysis experts are impressed by and exploit the way bridges and bypasses optimize their regular routes.
3. The data-intensive engineers are engaged in optimizing the total traffic, building and scheduling the vehicles, and reshaping the surrounding infrastructure to take maximum advantage—they may also be pioneering better technology for the next data-to-knowledge highway.

This book is taking the first steps toward understanding just how a data-to-knowledge highway should be built, and the first steps in shaping the methods and skills of the profession that will build those highways. We envisage a future where such data-to-knowledge highways support millions of knowledge discoveries a day that exploit the full wealth of available data.

REFERENCES

1. I. H. Witten, E. Frank, and M. A. Hall, *Data Mining: Practical Machine Learning Tools and Techniques* (Third Edition). Morgan Kauffman, 2011.
2. D. Kossmann, T. Kraska, and S. Loesing, "An evaluation of alternative architectures for transaction processing in the Cloud," in *SIGMOD Conference*, pp. 579–590. ACM, 2010.
3. A. Feng, M. J. Franklin, D. Kossmann, T. Kraska, S. Madden, S. Ramesh, A. Wang, and R. Xin, "CrowdDB: query processing with the VLDB crowd," *Proceedings of the VLDB*, vol. 4, no. 12, pp. 1387–1390, 2011.