PRINCIPLES AND PRACTICE OF BIG DATA

PREPARING, SHARING, AND ANALYZING COMPLEX INFORMATION

SECOND EDITION

JULES J. BERMAN

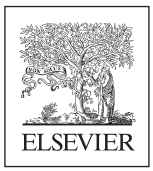
Principles and Practice of Big Data

Principles and Practice of Big Data

Preparing, Sharing, and Analyzing Complex Information

Second Edition

Jules J. Berman



ACADEMIC PRESS

An imprint of Elsevier

Academic Press is an imprint of Elsevier
125 London Wall, London EC2Y 5AS, United Kingdom
525 B Street, Suite 1650, San Diego, CA 92101, United States
50 Hampshire Street, 5th Floor, Cambridge, MA 02139, United States
The Boulevard, Langford Lane, Kidlington, Oxford OX5 1GB, United Kingdom

Notices
Knowledge and best practice in this field are constantly changing. As new research and experience broaden our understanding, changes in research methods, professional practices, or medical treatment may become necessary.

Practitioners and researchers must always rely on their own experience and knowledge in evaluating and using any information, methods, compounds, or experiments described herein. In using such information or methods they should be mindful of their own safety and the safety of others, including parties for whom they have a professional responsibility.

To the fullest extent of the law, neither the Publisher nor the authors, contributors, or editors, assume any liability for any injury and/or damage to persons or property as a matter of products liability, negligence or otherwise, or from any use or operation of any methods, products, instructions, or ideas contained in the material herein.

Library of Congress Cataloging-in-Publication Data
A catalog record for this book is available from the Library of Congress

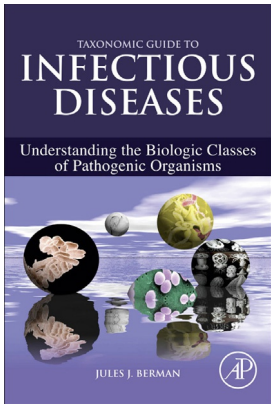
British Library Cataloguing-in-Publication Data
A catalogue record for this book is available from the British Library

ISBN: 978-0-12-815609-4

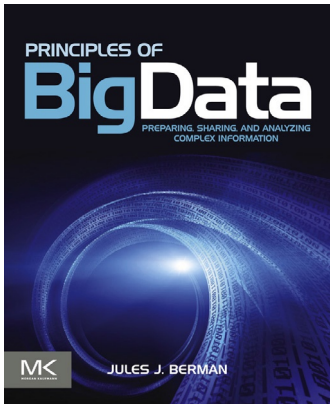
Publisher: Mara Conner
Acquisition Editor: Mara Conner
Editorial Project Manager: Mariana L. Kuhl
Production Project Manager: Punithavathy Govindaradjane
Cover Designer: Matthew Limbert

Typeset by SPi Global, India

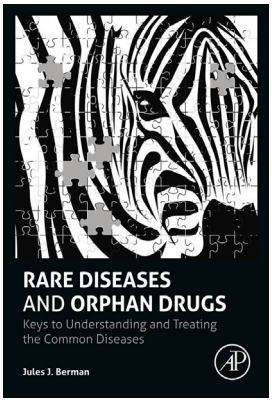
Other Books by Jules J. Berman



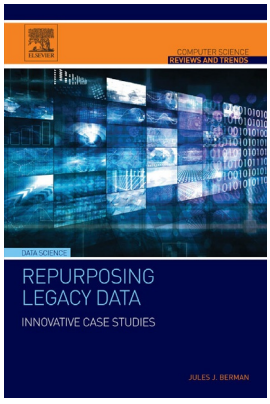
Taxonomic Guide to Infectious Diseases
Understanding the Biologic Classes of Pathogenic Organisms (2012)
9780124158955



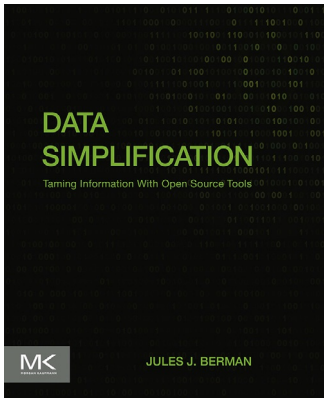
Principles of Big Data
Preparing, Sharing, and Analyzing Complex Information (2013)
9780124045767



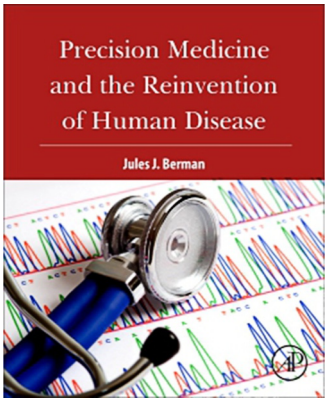
Rare Diseases and Orphan Drugs
Keys to Understanding and Treating the Common Diseases (2014)
9780124199880



Repurposing Legacy Data
Innovative Case Studies (2015)
9780128028827



Data Simplification
Taming Information with Open Source Tools (2016)
9780128037812



Precision Medicine and The Reinvention of Human Disease (2018)
9780128143933

Dedication

To my wife, Irene, who reads every day, and who understands why books are important.

About the Author

Jules J. Berman received two baccalaureate degrees from MIT; in Mathematics, and in Earth and Planetary Sciences. He holds a PhD from Temple University, and an MD, from the University of Miami. He was a graduate student researcher in the Fels Cancer Research Institute, at Temple University, and at the American Health Foundation in Valhalla, New York. His postdoctoral studies were completed at the US National Institutes of Health, and his residency was completed at the George Washington University Medical Center in Washington, DC. Dr. Berman served as Chief of Anatomic Pathology, Surgical Pathology, and Cytopathology at the Veterans Administration Medical Center in Baltimore, Maryland, where he held joint appointments at the University of Maryland Medical Center and at the Johns Hopkins Medical Institutions. In 1998, he transferred to the US National Institutes of Health, as a Medical Officer, and as the Program Director for Pathology Informatics in the Cancer Diagnosis Program at the National Cancer Institute. Dr. Berman is a past president of the Association for Pathology Informatics, and the 2011 recipient of the Association's Lifetime Achievement Award. He has first-authored over 100 scientific publications and has written more than a dozen books in the areas of data science and disease biology. Several of his most recent titles, published by Elsevier, include:

Taxonomic Guide to Infectious Diseases: Understanding the Biologic Classes of Pathogenic Organisms (2012)
Principles of Big Data: Preparing, Sharing, and Analyzing Complex Information (2013)
Rare Diseases and Orphan Drugs: Keys to Understanding and Treating the Common Diseases (2014)
Repurposing Legacy Data: Innovative Case Studies (2015)
Data Simplification: Taming Information with Open Source Tools (2016)
Precision Medicine and the Reinvention of Human Disease (2018)

Author's Preface to Second Edition

Everything has been said before, but since nobody listens we have to keep going back and beginning all over again.
Andre Gide

Good science writers will always jump at the chance to write a second edition of an earlier work. No matter how hard they try, that first edition will contain inaccuracies and misleading remarks. Sentences that seemed brilliant when first conceived will, with the passage of time, transform into examples of intellectual overreaching. Points too trivial to include in the original manuscript may now seem like profundities that demand a full explanation. A second edition provides rueful authors with an opportunity to correct the record.

When the first edition of Principles of Big Data was published in 2013 the field was very young and there were few scientists who knew what to do with Big Data. The data that kept pouring in was stored, like wheat in silos, throughout the planet. It was obvious to data managers that none of that stored data would have any scientific value unless it was properly annotated with metadata, identifiers, timestamps, and a set of basic descriptors. Under these conditions, the first edition of the Principles of Big Data stressed the proper and necessary methods for collecting, annotating, organizing, and curating Big Data. The process of preparing Big Data comes with its own unique set of challenges, and the First Edition was peppered with warnings and exhortations intended to steer readers clear of disaster.

It is now five years since the first edition was published and there have since been hundreds of books written on the subject of Big Data. As a scientist, it is disappointing to me that the bulk of Big Data, today, is focused on issues of marketing and predictive analytics (e.g., "Who is likely to buy product x, given that they bought product y two weeks previously?"); and machine learning (e.g., driverless cars, computer vision, speech recognition). Machine learning relies heavily on hyped up techniques such as neural networks and deep learning; neither of which are leading to fundamental laws and principles that simplify and broaden our understanding of the natural world and the physical universe. For the most part, these techniques use data that is relatively new (i.e., freshly collected), poorly annotated (i.e., provided with only the minimal information required for one particular analytic process), and not deposited for public evaluation or for re-use. In short, Big Data has followed the path of least resistance, avoiding most of the tough issues raised in the first edition of this book; such as the importance of sharing data with the public, the value of finding relationships (not similarities) among data objects, and the heavy, but inescapable, burden of creating robust, immortal, and well-annotated data.

It was certainly my hope that the greatest advances from Big Data would come as fundamental breakthroughs in the realms of medicine, biology, physics, engineering, and chemistry. Why has the focus of Big Data shifted from basic science over to machine learning? It may have something to do with the fact that no book, including the first edition of this book, has provided readers with the methods required to put the principles of Big Data into practice. In retrospect, it was not sufficient to describe a set of principles and then expect readers to invent their own methodologies.

Consequently, in this second edition, the publisher has changed the title of the book from "The Principles of Big Data," to "The Principles AND PRACTICE of Big Data." Henceforth and herein, recommendations are accompanied by the methods by which those recommendations can be implemented. The reader will find that all of the methods for implementing Big Data preparation and analysis are really quite simple. For the most part, computer methods require some basic familiarity with a programming language, and, despite misgivings, Python was chosen as the language of choice. The advantages of Python are:

- Python is a no-cost, open source, high-level programming language that is easy to acquire, install, learn, and use, and is available for every popular computer operating system.
- Python is extremely popular, at the present time, and its popularity seems to be increasing.
- Python distributions (such as Anaconda) come bundled with hundreds of highly useful modules (such as numpy, matplot, and scipy).
- Python has a large and active user group that has provided an extraordinary amount of documentation for Python methods and modules.
- Python supports some object-oriented techniques that will be discussed in this new edition

As everything in life, Python has its drawbacks:

- The most current versions of Python are not backwardly compatible with earlier versions. The scripts and code snippets included in this book should work for most versions of Python 3.x, but may not work with Python versions 2.x and earlier, unless the reader is prepared to devote some time to tweaking the code. Of course, these short scripts and snippets are intended as simplified demonstrations of concepts, and must not be construed as application-ready code.
- The built-in Python methods are sometimes maximized for speed by utilizing Random Access Memory (RAM) to hold data structures, including data structures built through iterative loops. Iterations through Big Data may exhaust available RAM, leading to the failure of Python scripts that functioned well with small data sets.
- Python's implementation of object orientation allows multiclass inheritance (i.e., a class can be the subclass of more than one parent class). We will describe why this is problematic, and the compensatory measures that we must take, whenever we use our Python programming skills to understand large and complex sets of data objects.

The core of every algorithm described in the book can be implemented in a few lines of code, using just about any popular programming language, under any operating system,

on any modern computer. Numerous Python snippets are provided, along with descriptions of free utilities that are widely available on every popular operating system. This book stresses the point that most data analyses conducted on large, complex data sets can be achieved with simple methods, bypassing specialized software systems (e.g., parallelization of computational processes) or hardware (e.g., supercomputers). Readers who are completely unacquainted with Python may find that they can read and understand Python code, if the snippets of code are brief, and accompanied by some explanation in the text. In any case, readers who are primarily concerned with mastering the principles of Big Data can skip the code snippets without losing the narrative thread of the book.

This second edition has been expanded to stress methodologies that have been overlooked by the authors of other books in the field of Big Data analysis. These would include:

– **Data preparation.**

How to annotate data with metadata and how to create data objects composed of triples. The concept of the triple, as the fundamental conveyor of meaning in the computational sciences, is fully explained.

– **Data structures of particular relevance to Big Data**

Concepts such as triplestores, distributed ledgers, unique identifiers, timestamps, concordances, indexes, dictionary objects, data persistence, and the roles of one-way hashes and encryption protocols for data storage and distribution are covered.

– **Classification of data objects**

How to assign data objects to classes based on their shared relationships, and the computational roles filled by classifications in the analysis of Big Data will be discussed at length.

– **Introspection**

How to create data objects that are self-describing, permitting the data analyst to group objects belonging to the same class and to apply methods to class objects that have been inherited from their ancestral classes.

– **Algorithms that have special utility in Big Data preparation and analysis**

How to use one-way hashes, unique identifier generators, cryptographic techniques, timing methods, and time stamping protocols to create unique data objects that are immutable (never changing), immortal, and private; and to create data structures that facilitate a host of useful functions that will be described (e.g., blockchains and distributed ledgers, protocols for safely sharing confidential information, and methods for reconciling identifiers across data collections without violating privacy).

– **Tips for Big Data analysis**

How to overcome many of the analytic limitations imposed by scale and dimensionality, using a range of simple techniques (e.g., approximations, so-called back-of-the-envelope

tricks, repeated sampling using a random number generator, Monte Carlo simulations, and data reduction methods).

– **Data reanalysis, data repurposing, and data sharing**

Why the first analysis of Big Data is almost always incorrect, misleading, or woefully incomplete, and why data reanalysis has become a crucial skill that every serious Big Data analyst must acquire. The process of data reanalysis often inspires repurposing of Big Data resources. Neither data reanalysis nor data repurposing can be achieved unless and until the obstacles to data sharing are overcome. The topics of data reanalysis, data repurposing, and data sharing are explored at length.

Comprehensive texts, such as the second edition of the Principles and Practice of Big Data, are never quite as comprehensive as they might strive to be; there simply is no way to fully describe every concept and method that is relevant to a multi-disciplinary field, such as Big Data. To compensate for such deficiencies, there is an extensive Glossary section for every chapter, that defines the terms introduced in the text, providing some explanation of the relevance of the terms for Big Data scientists. In addition, when techniques and methods are discussed, a list of references that the reader may find useful, for further reading on the subject, is provided. Altogether, the second edition contains about 600 citations to outside references, most of which are available as free downloads. There are over 300 glossary items, many of which contain short Python snippets that readers may find useful.

As a final note, this second edition uses case studies to show readers how the principles of Big Data are put into practice. Although case studies are drawn from many fields of science, including physics, economics, and astronomy, readers will notice an overabundance of examples drawn from the biological sciences (particularly medicine and zoology). The reason for this is that the taxonomy of all living terrestrial organisms is the oldest and best Big Data classification in existence. All of the classic errors in data organization, and in data analysis, have been committed in the field of biology. More importantly, these errors have been documented in excruciating detail and most of the documented errors have been corrected and published for public consumption. If you want to understand how Big Data can be used as a tool for scientific advancement, then you must look at case examples taken from the world of biology, a well-documented field where everything that can happen has happened, is happening, and will happen. Every effort has been made to limit Case Studies to the simplest examples of their type, and to provide as much background explanation as non-biologists may require.

Principles and Practice of Big Data, Second Edition, is devoted to the intellectual conviction that the primary purpose of Big Data analysis is to permit us to ask and answer a wide range of questions that could not have been credibly approached with small sets of data. There is every reason to hope that the readers of this book will soon achieve scientific breakthroughs that were beyond the reach of prior generations of scientists. Good luck!

Author's Preface to First Edition

We can't solve problems by using the same kind of thinking we used when we created them.
Albert Einstein

Data pours into millions of computers every moment of every day. It is estimated that the total accumulated data stored on computers worldwide is about 300 exabytes (that's 300 billion gigabytes). Data storage increases at about 28% per year. The data stored is peanuts compared to data that is transmitted without storage. The annual transmission of data is estimated at about 1.9 zettabytes or 1,900 billion gigabytes [1]. From this growing tangle of digital information, the next generation of data resources will emerge.

As we broaden our data reach (i.e., the different kinds of data objects included in the resource), and our data timeline (i.e., accruing data from the future and the deep past), we need to find ways to fully describe each piece of data, so that we do not confuse one data item with another, and so that we can search and retrieve data items when we need them. Astute informaticians understand that if we fully describe everything in our universe, we would need to have an ancillary universe to hold all the information, and the ancillary universe would need to be much larger than our physical universe.

In the rush to acquire and analyze data, it is easy to overlook the topic of data preparation. If the data in our Big Data resources are not well organized, comprehensive, and fully described, then the resources will have no value. The primary purpose of this book is to explain the principles upon which serious Big Data resources are built. All of the data held in Big Data resources must have a form that supports search, retrieval, and analysis. The analytic methods must be available for review, and the analytic results must be available for validation.

Perhaps the greatest potential benefit of Big Data is its ability to link seemingly disparate disciplines, to develop and test hypothesis that cannot be approached within a single knowledge domain. Methods by which analysts can navigate through different Big Data resources to create new, merged data sets, will be reviewed.

What exactly, is Big Data? Big Data is characterized by the three V's: volume (large amounts of data), variety (includes different types of data), and velocity (constantly accumulating new data) [2]. Those of us who have worked on Big Data projects might suggest throwing a few more v's into the mix: vision (having a purpose and a plan), verification (ensuring that the data conforms to a set of specifications), and validation (checking that its purpose is fulfilled).

Many of the fundamental principles of Big Data organization have been described in the "metadata" literature. This literature deals with the formalisms of data description (i.e., how to describe data); the syntax of data description (e.g., markup languages such as eXtensible Markup Language, XML); semantics (i.e., how to make computer-parsable statements that convey meaning); the syntax of semantics (e.g., framework specifications such as Resource Description Framework, RDF, and Web Ontology Language, OWL); the creation of data objects that hold data values and self-descriptive information; and the deployment of ontologies, hierarchical class systems whose members are data objects.

The field of metadata may seem like a complete waste of time to professionals who have succeeded very well, in data-intensive fields, without resorting to metadata formalisms. Many computer scientists, statisticians, database managers, and network specialists have no trouble handling large amounts of data, and they may not see the need to create a strange new data model for Big Data resources. They might feel that all they really need is greater storage capacity, distributed over more powerful computers that work in parallel with one another. With this kind of computational power, they can store, retrieve, and analyze larger and larger quantities of data. These fantasies only apply to systems that use relatively simple data or data that can be represented in a uniform and standard format. When data is highly complex and diverse, as found in Big Data resources, the importance of metadata looms large. Metadata will be discussed, with a focus on those concepts that must be incorporated into the organization of Big Data resources. The emphasis will be on explaining the relevance and necessity of these concepts, without going into gritty details that are well covered in the metadata literature.

When data originates from many different sources, arrives in many different forms, grows in size, changes its values, and extends into the past and the future, the game shifts from data computation to data management. I hope that this book will persuade readers that faster, more powerful computers are nice to have, but these devices cannot compensate for deficiencies in data preparation. For the foreseeable future, universities, federal agencies, and corporations will pour money, time, and manpower into Big Data efforts. If they ignore the fundamentals, their projects are likely to fail. On the other hand, if they pay attention to Big Data fundamentals, they will discover that Big Data analyses can be performed on standard computers. The simple lesson, that data trumps computation, will be repeated throughout this book in examples drawn from well-documented events.

There are three crucial topics related to data preparation that are omitted from virtually every other Big Data book: identifiers, immutability, and introspection.

A thoughtful identifier system ensures that all of the data related to a particular data object will be attached to the correct object, through its identifier, and to no other object. It seems simple, and it is, but many Big Data resources assign identifiers promiscuously, with the end result that information related to a unique object is scattered throughout the resource, attached to other objects, and cannot be sensibly retrieved when needed. The concept of object identification is of such overriding importance that a Big Data resource can be usefully envisioned as a collection of unique identifiers to which complex data is attached.

Immutability is the principle that data collected in a Big Data resource is permanent, and can never be modified. At first thought, it would seem that immutability is a ridiculous and impossible constraint. In the real world, mistakes are made, information changes, and the methods for describing information changes. This is all true, but the astute Big Data manager knows how to accrue information into data objects without changing the pre-existing data. Methods for achieving this seemingly impossible trick will be described in detail.

Introspection is a term borrowed from object-oriented programming, not often found in the Big Data literature. It refers to the ability of data objects to describe themselves when interrogated. With introspection, users of a Big Data resource can quickly determine the content of data objects and the hierarchical organization of data objects within the Big Data resource. Introspection allows users to see the types of data relationships that can be analyzed within the resource and clarifies how disparate resources can interact with one another.

Another subject covered in this book, and often omitted from the literature on Big Data, is data indexing. Though there are many books written on the art of the science of so-called back-of-the-book indexes, scant attention has been paid to the process of preparing indexes for large and complex data resources. Consequently, most Big Data resources have nothing that could be called a serious index. They might have a Web page with a few links to explanatory documents, or they might have a short and crude "help" index, but it would be rare to find a Big Data resource with a comprehensive index containing a thoughtful and updated list of terms and links. Without a proper index, most Big Data resources have limited utility for any but a few cognoscenti. It seems odd to me that organizations willing to spend hundreds of millions of dollars on a Big Data resource will balk at investing a few thousand dollars more for a proper index.

Aside from these four topics, which readers would be hard-pressed to find in the existing Big Data literature, this book covers the usual topics relevant to Big Data design, construction, operation, and analysis. Some of these topics include data quality, providing structure to unstructured data, data deidentification, data standards and interoperability issues, legacy data, data reduction and transformation, data analysis, and software issues. For these topics, discussions focus on the underlying principles; programming code and mathematical equations are conspicuously inconspicuous. An extensive Glossary covers the technical or specialized terms and topics that appear throughout the text. As each Glossary term is "optional" reading, I took the liberty of expanding on technical or mathematical concepts that appeared in abbreviated form in the main text. The Glossary provides an explanation of the practical relevance of each term to Big Data, and some readers may enjoy browsing the Glossary as a stand-alone text.

The final four chapters are non-technical; all dealing in one way or another with the consequences of our exploitation of Big Data resources. These chapters will cover legal, social, and ethical issues. The book ends with my personal predictions for the future of Big Data, and its impending impact on our futures. When preparing this book, I debated whether these four chapters might best appear in the front of the book, to whet the reader's

appetite for the more technical chapters. I eventually decided that some readers would be unfamiliar with some of the technical language and concepts included in the final chapters, necessitating their placement near the end.

Readers may notice that many of the case examples described in this book come from the field of medical informatics. The healthcare informatics field is particularly ripe for discussion because every reader is affected, on economic and personal levels, by the Big Data policies and actions emanating from the field of medicine. Aside from that, there is a rich literature on Big Data projects related to healthcare. As much of this literature is controversial, I thought it important to select examples that I could document from reliable sources. Consequently, the reference section is large, with over 200 articles from journals, newspaper articles, and books. Most of these cited articles are available for free Web download.

Who should read this book? This book is written for professionals who manage Big Data resources and for students in the fields of computer science and informatics. Data management professionals would include the leadership within corporations and funding agencies who must commit resources to the project, the project directors who must determine a feasible set of goals and who must assemble a team of individuals who, in aggregate, hold the requisite skills for the task: network managers, data domain specialists, metadata specialists, software programmers, standards experts, interoperability experts, statisticians, data analysts, and representatives from the intended user community. Students of informatics, the computer sciences, and statistics will discover that the special challenges attached to Big Data, seldom discussed in university classes, are often surprising; sometimes shocking.

By mastering the fundamentals of Big Data design, maintenance, growth, and validation, readers will learn how to simplify the endless tasks engendered by Big Data resources. Adept analysts can find relationships among data objects held in disparate Big Data resources if the data is prepared properly. Readers will discover how integrating Big Data resources can deliver benefits far beyond anything attained from stand-alone databases.

References

[1] Martin Hilbert M, Lopez P. The world's technological capacity to store, communicate, and compute information. Science 2011;332:60–5.

[2] Schmidt S. Data is exploding: the 3V's of Big Data. Business Computing World; 2012. May 15.

1 Introduction

OUTLINE

Section 1.1. Definition of Big Data

It's the data, stupid.

Jim Gray

Back in the mid 1960s, my high school held pep rallies before big games. At one of these rallies, the head coach of the football team walked to the center of the stage carrying a large box of printed computer paper; each large sheet was folded flip-flop style against the next sheet and they were all held together by perforations. The coach announced that the athletic abilities of every member of our team had been entered into the school's computer (we were lucky enough to have our own IBM-360 mainframe). Likewise, data on our rival team had also been entered. The computer was instructed to digest all of this information and to produce the name of the team that would win the annual Thanksgiving Day showdown. The computer spewed forth the aforementioned box of computer paper; the very last output sheet revealed that we were the pre-ordained winners. The next day, we sallied forth to yet another ignominious defeat at the hands of our long-time rivals.

Fast-forward about 50 years to a conference room at the National Institutes of Health (NIH), in Bethesda, Maryland. A top-level science administrator is briefing me. She explains that disease research has grown in scale over the past decade. The very best research initiatives are now multi-institutional and data-intensive. Funded investigators are using high-throughput molecular methods that produce mountains of data for every tissue sample in a matter of minutes. There is only one solution; we must acquire supercomputers and a staff of talented programmers who can analyze all our data and tell us what it all means!

The NIH leadership believed, much as my high school coach believed, that if you have a really big computer and you feed it a huge amount of information, then you can answer almost any question.

Principles and Practice of Big Data. https://doi.org/10.1016/B978-0-12-815609-4.00001-7

That day, in the conference room at the NIH, circa 2003, I voiced my concerns, indicating that you cannot just throw data into a computer and expect answers to pop out. I pointed out that, historically, science has been a reductive process, moving from complex, descriptive data sets to simplified generalizations. The idea of developing an expensive supercomputer facility to work with increasing quantities of biological data, at higher and higher levels of complexity, seemed impractical and unnecessary. On that day, my concerns were not well received. High performance supercomputing was a very popular topic, and still is. [Glossary Science, Supercomputer]

Fifteen years have passed since the day that supercomputer-based cancer diagnosis was envisioned. The diagnostic supercomputer facility was never built. The primary diagnostic tool used in hospital laboratories is still the microscope, a tool invented circa 1590. Today, we augment microscopic findings with genetic tests for specific, key mutations; but we do not try to understand all of the complexities of human genetic variations. We know that it is hopeless to try. You can find a lot of computers in hospitals and medical offices, but the computers do not calculate your diagnosis. Computers in the medical workplace are relegated to the prosaic tasks of collecting, storing, retrieving, and delivering medical records. When those tasks are finished, the computer sends you the bill for services rendered.

Before we can take advantage of large and complex data sources, we need to think deeply about the meaning and destiny of Big Data.

Big Data is defined by the three V's:

1. Volume—large amounts of data;.
2. Variety—the data comes in different forms, including traditional databases, images, documents, and complex records;.
3. Velocity—the content of the data is constantly changing through the absorption of complementary data collections, the introduction of previously archived data or legacy collections, and from streamed data arriving from multiple sources.

It is important to distinguish Big Data from "lotsa data" or "massive data." In a Big Data Resource, all three V's must apply. It is the size, complexity, and restlessness of Big Data resources that account for the methods by which these resources are designed, operated, and analyzed. [Glossary Big Data resource, Data resource]

The term "lotsa data" is often applied to enormous collections of simple-format records. For example: every observed star, its magnitude and its location; the name and cell phone number of every person living in the United States; and the contents of the Web. These very large data sets are sometimes just glorified lists. Some "lotsa data" collections are spreadsheets (2-dimensional tables of columns and rows), so large that we may never see where they end.

Big Data resources are not equivalent to large spreadsheets, and a Big Data resource is never analyzed in its totality. Big Data analysis is a multi-step process whereby data is extracted, filtered, and transformed, with analysis often proceeding in a piecemeal, sometimes recursive, fashion. As you read this book, you will find that the gulf between "lotsa data" and Big Data is profound; the two subjects can seldom be discussed productively within the same venue.

Section 1.2. Big Data Versus Small Data

> *Actually, the main function of Big Science is to generate massive amounts of reliable and easily accessible data.... Insight, understanding, and scientific progress are generally achieved by 'small science.'*
>
> *Dan Graur, Yichen Zheng, Nicholas Price, Ricardo Azevedo, Rebecca Zufall, and Eran Elhaik [1].*

Big Data is not small data that has become bloated to the point that it can no longer fit on a spreadsheet, nor is it a database that happens to be very large. Nonetheless, some professionals who customarily work with relatively small data sets, harbor the false impression that they can apply their spreadsheet and database know-how directly to Big Data resources without attaining new skills or adjusting to new analytic paradigms. As they see things, when the data gets bigger, only the computer must adjust (by getting faster, acquiring more volatile memory, and increasing its storage capabilities); Big Data poses no special problems that a supercomputer could not solve. [Glossary Database]

This attitude, which seems to be prevalent among database managers, programmers, and statisticians, is highly counterproductive. It will lead to slow and ineffective software, huge investment losses, bad analyses, and the production of useless and irreversibly defective Big Data resources.

Let us look at a few of the general differences that can help distinguish Big Data and small data.

– **Goals**

small data—Usually designed to answer a specific question or serve a particular goal.

Big Data—Usually designed with a goal in mind, but the goal is flexible and the questions posed are protean. Here is a short, imaginary funding announcement for Big Data grants designed "to combine high quality data from fisheries, coast guard, commercial shipping, and coastal management agencies for a growing data collection that can be used to support a variety of governmental and commercial management studies in the Lower Peninsula." In this fictitious case, there is a vague goal, but it is obvious that there really is no way to completely specify what the Big Data resource will contain, how the various types of data held in the resource will be organized, connected to other data resources, or usefully analyzed. Nobody can specify, with any degree of confidence, the ultimate destiny of any Big Data project; it usually comes as a surprise.

– **Location**

small data—Typically, contained within one institution, often on one computer, sometimes in one file.

Big Data—Spread throughout electronic space and typically parceled onto multiple Internet servers, located anywhere on earth.

– **Data structure and content**

small data—Ordinarily contains highly structured data. The data domain is restricted to a single discipline or sub-discipline. The data often comes in the form of uniform records in an ordered spreadsheet.

Big Data—Must be capable of absorbing unstructured data (e.g., such as free-text documents, images, motion pictures, sound recordings, physical objects). The subject matter of the resource may cross multiple disciplines, and the individual data objects in the resource may link to data contained in other, seemingly unrelated, Big Data resources. [Glossary Data object]

– **Data preparation**

small data—In many cases, the data user prepares her own data, for her own purposes.

Big Data—The data comes from many diverse sources, and it is prepared by many people. The people who use the data are seldom the people who have prepared the data.

– **Longevity**

small data—When the data project ends, the data is kept for a limited time (seldom longer than 7 years, the traditional academic life-span for research data); and then discarded.

Big Data—Big Data projects typically contain data that must be stored in perpetuity. Ideally, the data stored in a Big Data resource will be absorbed into other data resources. Many Big Data projects extend into the future and the past (e.g., legacy data), accruing data prospectively and retrospectively. [Glossary Legacy data]

– **Measurements**

small data—Typically, the data is measured using one experimental protocol, and the data can be represented using one set of standard units. [Glossary Protocol]

Big Data—Many different types of data are delivered in many different electronic formats. Measurements, when present, may be obtained by many different protocols. Verifying the quality of Big Data is one of the most difficult tasks for data managers. [Glossary Data Quality Act]

– **Reproducibility**

small data—Projects are typically reproducible. If there is some question about the quality of the data, the reproducibility of the data, or the validity of the conclusions drawn from the data, the entire project can be repeated, yielding a new data set. [Glossary Conclusions]

Big Data—Replication of a Big Data project is seldom feasible. In general, the most that anyone can hope for is that bad data in a Big Data resource will be found and flagged as such.

– **Stakes**

small data—Project costs are limited. Laboratories and institutions can usually recover from the occasional small data failure.

Big Data—Big Data projects can be obscenely expensive [2,3]. A failed Big Data effort can lead to bankruptcy, institutional collapse, mass firings, and the sudden disintegration

of all the data held in the resource. As an example, a United States National Institutes of Health Big Data project known as the "NCI cancer biomedical informatics grid" cost at least $350 million for fiscal years 2004–10. An ad hoc committee reviewing the resource found that despite the intense efforts of hundreds of cancer researchers and information specialists, it had accomplished so little and at so great an expense that a project moratorium was called [4]. Soon thereafter, the resource was terminated [5]. Though the costs of failure can be high, in terms of money, time, and labor, Big Data failures may have some redeeming value. Each failed effort lives on as intellectual remnants consumed by the next Big Data effort. [Glossary Grid]

– **Introspection**

small data—Individual data points are identified by their row and column location within a spreadsheet or database table. If you know the row and column headers, you can find and specify all of the data points contained within. [Glossary Data point]

Big Data—Unless the Big Data resource is exceptionally well designed, the contents and organization of the resource can be inscrutable, even to the data managers. Complete access to data, information about the data values, and information about the organization of the data is achieved through a technique herein referred to as introspection. Introspection will be discussed at length in Chapter 6. [Glossary Data manager, Introspection]

– **Analysis**

small data—In most instances, all of the data contained in the data project can be analyzed together, and all at once.

Big Data—With few exceptions, such as those conducted on supercomputers or in parallel on multiple computers, Big Data is ordinarily analyzed in incremental steps. The data are extracted, reviewed, reduced, normalized, transformed, visualized, interpreted, and re-analyzed using a collection of specialized methods. [Glossary Parallel computing, MapReduce]

Section 1.3. Whence Comest Big Data?

All I ever wanted to do was to paint sunlight on the side of a house.

Edward Hopper

Often, the impetus for Big Data is entirely ad hoc. Companies and agencies are forced to store and retrieve huge amounts of collected data (whether they want to or not). Generally, Big Data come into existence through any of several different mechanisms:

– An entity has collected a lot of data in the course of its normal activities and seeks to organize the data so that materials can be retrieved, as needed.

The Big Data effort is intended to streamline the regular activities of the entity. In this case, the data is just waiting to be used. The entity is not looking to discover anything or to do anything new. It simply wants to use the data to accomplish what it has always been doing;

only better. The typical medical center is a good example of an "accidental" Big Data resource. The day-to-day activities of caring for patients and recording data into hospital information systems results in terabytes of collected data, in forms such as laboratory reports, pharmacy orders, clinical encounters, and billing data. Most of this information is generated for a one-time specific use (e.g., supporting a clinical decision, collecting payment for a procedure). It occurs to the administrative staff that the collected data can be used, in its totality, to achieve mandated goals: improving quality of service, increasing staff efficiency, and reducing operational costs. [Glossary Binary units for Big Data, Binary atom count of universe]

– An entity has collected a lot of data in the course of its normal activities and decides that there are many new activities that could be supported by their data.

Consider modern corporations; these entities do not restrict themselves to one manufacturing process or one target audience. They are constantly looking for new opportunities. Their collected data may enable them to develop new products based on the preferences of their loyal customers, to reach new markets, or to market and distribute items via the Web. These entities will become hybrid Big Data/manufacturing enterprises.

– An entity plans a business model based on a Big Data resource.

Unlike the previous examples, this entity starts with Big Data and adds a physical component secondarily. Amazon and FedEx may fall into this category, as they began with a plan for providing a data-intense service (e.g., the Amazon Web catalog and the FedEx package tracking system). The traditional tasks of warehousing, inventory, pick-up, and delivery, had been available all along, but lacked the novelty and efficiency afforded by Big Data.

– An entity is part of a group of entities that have large data resources, all of whom understand that it would be to their mutual advantage to federate their data resources [6].

An example of a federated Big Data resource would be hospital databases that share electronic medical health records [7].

– An entity with skills and vision develops a project wherein large amounts of data are collected and organized, to the benefit of themselves and their user-clients.

An example would be a massive online library service, such as the U.S. National Library of Medicine's PubMed catalog, or the Google Books collection.

– An entity has no data and has no particular expertise in Big Data technologies, but it has money and vision.

The entity seeks to fund and coordinate a group of data creators and data holders, who will build a Big Data resource that can be used by others. Government agencies have been the major benefactors. These Big Data projects are justified if they lead to important discoveries that could not be attained at a lesser cost with smaller data resources.

Section 1.4. The Most Common Purpose of Big Data Is to Produce Small Data

> *If I had known what it would be like to have it all, I might have been willing to settle for less.*
>
> *Lily Tomlin*

Imagine using a restaurant locater on your smartphone. With a few taps, it lists the Italian restaurants located within a 10-block radius of your current location. The database being queried is big and complex (a map database, a collection of all the restaurants in the world, their longitudes and latitudes, their street addresses, and a set of ratings provided by patrons, updated continuously), but the data that it yields is small (e.g., five restaurants, marked on a street map, with pop-ups indicating their exact address, telephone number, and ratings). Your task comes down to selecting one restaurant from among the five, and dining thereat.

In this example, your data selection was drawn from a large data set, but your ultimate analysis was confined to a small data set (i.e., five restaurants meeting your search criteria). The purpose of the Big Data resource was to proffer the small data set. No analytic work was performed on the Big Data resource; just search and retrieval. The real labor of the Big Data resource involved collecting and organizing complex data, so that the resource would be ready for your query. Along the way, the data creators had many decisions to make (e.g., Should bars be counted as restaurants? What about take-away only shops? What data should be collected? How should missing data be handled? How will data be kept current? [Glossary Query, Missing data]

Big Data is seldom, if ever, analyzed *in toto*. There is almost always a drastic filtering process that reduces Big Data into smaller data. This rule applies to scientific analyses. The Australian Square Kilometre Array of radio telescopes [8], WorldWide Telescope, CERN's Large Hadron Collider and the Pan-STARRS (Panoramic Survey Telescope and Rapid Response System) array of telescopes produce petabytes of data every day. Researchers use these raw data sources to produce much smaller data sets for analysis [9]. [Glossary Raw data, Square Kilometer Array, Large Hadron Collider, WorldWide Telescope]

Here is an example showing how workable subsets of data are prepared from Big Data resources. Blazars are rare super-massive black holes that release jets of energy that move at near-light speeds. Cosmologists want to know as much as they can about these strange objects. A first step to studying blazars is to locate as many of these objects as possible. Afterwards, various measurements on all of the collected blazars can be compared, and their general characteristics can be determined. Blazars seem to have a gamma ray signature that is not present in other celestial objects. The WISE survey collected infrared data on the entire observable universe. Researchers extracted from the Wise data every celestial body associated with an infrared signature in the gamma ray range that was suggestive of blazars; about 300 objects. Further research on these 300 objects led the researchers to

believe that about half were blazars [10]. This is how Big Data research often works; by constructing small data sets that can be productively analyzed.

Because a common role of Big Data is to produce small data, a question that data managers must ask themselves is: "Have I prepared my Big Data resource in a manner that helps it become a useful source of small data?"

Section 1.5. Big Data Sits at the Center of the Research Universe

Physics is the universe's operating system.

Steven R Garman

In the past, scientists followed a well-trodden path toward truth: hypothesis, then experiment, then data, then analysis, then publication. The manner in which a scientist analyzed his or her data was crucial because other scientists would not have access to the same data and could not re-analyze the data for themselves. Basically, the results and conclusions described in the manuscript was the scientific product. The primary data upon which the results and conclusion were based (other than one or two summarizing tables) were not made available for review. Scientific knowledge was built on trust. Customarily, the data would be held for 7 years, and then discarded. [Glossary Results]

In the Big data paradigm the concept of a final manuscript has little meaning. Big Data resources are permanent, and the data within the resource is immutable (See Chapter 6). Any scientist's analysis of the data does not need to be the final word; another scientist can access and re-analyze the same data over and over again. Original conclusions can be validated or discredited. New conclusions can be developed. The centerpiece of science has moved from the manuscript, whose conclusions are tentative until validated, to the Big Data resource, whose data will be tapped repeatedly to validate old manuscripts and spawn new manuscripts. [Glossary Immutability, Mutability]

Today, hundreds or thousands of individuals might contribute to a Big Data resource. The data in the resource might inspire dozens of major scientific projects, hundreds of manuscripts, thousands of analytic efforts, and millions or billions of search and retrieval operations. The Big Data resource has become the central, massive object around which universities, research laboratories, corporations, and federal agencies orbit. These orbiting objects draw information from the Big Data resource, and they use the information to support analytic studies and to publish manuscripts. Because Big Data resources are permanent, any analysis can be critically examined using the same set of data, or re-analyzed anytime in the future. Because Big Data resources are constantly growing forward in time (i.e., accruing new information) and backward in time (i.e., absorbing legacy data sets), the value of the data is constantly increasing.

Big Data resources are the stars of the modern information universe. All matter in the physical universe comes from heavy elements created inside stars, from lighter elements. All data in the informational universe is complex data built from simple data. Just as stars

can exhaust themselves, explode, or even collapse under their own weight to become black holes; Big Data resources can lose funding and die, release their contents and burst into nothingness, or collapse under their own weight, sucking everything around them into a dark void. It is an interesting metaphor. In the following chapters, we will see how a Big Data resource can be designed and operated to ensure stability, utility, growth, and permanence; features you might expect to find in a massive object located in the center of the information universe.

Glossary

Big Data resource A Big Data collection that is accessible for analysis. Readers should understand that there are collections of Big Data (i.e., data sources that are large, complex, and actively growing) that are not designed to support analysis; hence, not Big Data resources. Such Big Data collections might include some of the older hospital information systems, which were designed to deliver individual patient records upon request; but could not support projects wherein all of the data contained in all of the records were opened for selection and analysis. Aside from privacy and security issues, opening a hospital information system to these kinds of analyses would place enormous computational stress on the systems (i.e., produce system crashes). In the late 1990s and the early 2000s data warehousing was popular. Large organizations would collect all of the digital information created within their institutions, and these data were stored as Big Data collections, called data warehouses. If an authorized person within the institution needed some specific set of information (e.g., emails sent or received in February, 2003; all of the bills paid in November, 1999), it could be found somewhere within the warehouse. For the most part, these data warehouses were not true Big Data resources because they were not organized to support a full analysis of all of the contained data. Another type of Big Data collection that may or may not be considered a Big Data resource are compilations of scientific data that are accessible for analysis by private concerns, but closed for analysis by the public. In this case a scientist may make a discovery based on her analysis of a private Big Data collection, but the research data is not open for critical review. In the opinion of some scientists, including myself, if the results of a data analysis are not available for review, then the analysis is illegitimate. Of course, this opinion is not universally shared, and Big Data professionals hold various definitions for a Big Data resource.

Binary atom count of universe There are estimated to be about 10^80 atoms in the universe. Log2(10) is 3.32192809, so the number of atoms in the universe is 2^80*3.32192809 or 2^266 atoms.

Binary units for Big Data Binary sizes are named in 1000-fold intervals: 1 bit = binary digit (0 or 1); 1 byte = 8 bits (the number of bits required to express an ascii character); 1000 bytes = 1 kilobyte; 1000 kilobytes = 1 megabyte; 1000 megabytes = 1 gigabyte; 1000 gigabytes = 1 terabyte; 1000 terabytes = 1 petabyte; 1000 petabytes = 1 exabyte; 1000 exabytes = 1 zettabyte; 1000 zettabytes = 1 yottabyte.

Conclusions Conclusions are the interpretations made by studying the results of an experiment or a set of observations. The term "results" should never be used interchangeably with the term "conclusions." **Remember, results are verified. Conclusions are validated** [11].

Data Quality Act In the United States the data upon which public policy is based must have quality and must be available for review by the public. Simply put, public policy must be based on verifiable data. The Data Quality Act of 2002 requires the Office of Management and Budget to develop government-wide standards for data quality [12].

Data manager This book uses "data manager" as a catchall term, without attaching any specific meaning to the name. Depending on the institutional and cultural milieu, synonyms and plesionyms (i.e., near-synonyms) for data manager would include: technical lead, team liaison, data quality manager, chief curator, chief of operations, project manager, group supervisor, and so on.

Data object As used in this book, a data object consists of a unique object identifier along with all of the data/metadata pairs that rightly belong to the object identifier, and that includes one data/metadata pair that tells us the object's class.

```
75898039563441
  name                G. Willikers
  gender              male
  age                 35
  is_a_class_member   cowboy
```

In this example, the object identifier, 75898039563441, is followed by its data/metadata pairs, including the one pair that tells us that the object (a 35-year-old man named G. Willikers) belongs to the class of individuals known as "cowboy."

The utility of data objects, in the field of Big Data, is discussed in Section 6.2.

Data point The singular form of data is datum. Strictly speaking, the term should be datum point or datumpoint. Most information scientists, myself included, have abandoned consistent usage rules for the word "data." In this book, the term "data" always refers collectively to information, numeric or textual, structured or unstructured, in any quantity.

Data resource A collection of data made available for data retrieval. The data can be distributed over servers located anywhere on earth or in space. The resource can be static (i.e., having a fixed set of data), or in flux. Plesionyms for data resource are: data warehouse, data repository, data archive, and data store.

Database A software application designed specifically to create and retrieve large numbers of data records (e.g., millions or billions). The data records of a database are persistent, meaning that the application can be turned off, then on, and all the collected data will be available to the user.

Grid A collection of computers and computer resources (typically networked servers) that is coordinated to provide a desired functionality. In the most advanced Grid computing architecture, requests can be broken into computational tasks that are processed in parallel on multiple computers and transparently (from the client's perspective) assembled and returned. The Grid is the intellectual predecessor of Cloud computing. Cloud computing is less physically and administratively restricted than Grid computing.

Immutability Immutability is the principle that data collected in a Big Data resource is permanent and can never be modified. At first thought, it would seem that immutability is a ridiculous and impossible constraint. In the real world, mistakes are made, information changes, and the methods for describing information changes. This is all true, but the astute Big Data manager knows how to accrue information into data objects without changing the pre-existing data. Methods for achieving this seemingly impossible trick are described in Chapter 8.

Introspection Well-designed Big Data resources support introspection, a method whereby data objects within the resource can be interrogated to yield their properties, values, and class membership. Through introspection the relationships among the data objects in the Big Data resource can be examined and the structure of the resource can be determined. Introspection is the method by which a data user can find everything there is to know about a Big Data resource without downloading the complete resource.

Large Hadron Collider The Large Hadron Collider is the world's largest and most powerful particle accelerator and is expected to produce about 15 petabytes (15 million gigabytes) of data annually [13].

Legacy data Data collected by an information system that has been replaced by a newer system, and which cannot be immediately integrated into the newer system's database. For example, hospitals regularly replace their hospital information systems with new systems that promise greater efficiencies, expanded services, or improved interoperability with other information systems. In many cases, the new system cannot readily integrate the data collected from the older system. The previously collected

data becomes a legacy to the new system. In such cases, legacy data is simply "stored" for some arbitrary period of time in case someone actually needs to retrieve any of the legacy data. After a decade or so the hospital may find itself without any staff members who are capable of locating the storage site of the legacy data, or moving the data into a modern operating system, or interpreting the stored data, or retrieving appropriate data records, or producing a usable query output.

MapReduce A method by which computationally intensive problems can be processed on multiple computers, in parallel. The method can be divided into a mapping step and a reducing step. In the mapping step a master computer divides a problem into smaller problems that are distributed to other computers. In the reducing step the master computer collects the output from the other computers. Although MapReduce is intended for Big Data resources, and can hold petabytes of data, most Big Data problems do not require MapReduce.

Missing data Most complex data sets have missing data values. Somewhere along the line data elements were not entered, records were lost, or some systemic error produced empty data fields. Big Data, being large, complex, and composed of data objects collected from diverse sources, is almost certain to have missing data. Various mathematical approaches to missing data have been developed; commonly involving assigning values on a statistical basis; so-called imputation methods. The underlying assumption for such methods is that missing data arises at random. When missing data arises nonrandomly, there is no satisfactory statistical fix. The Big Data curator must track down the source of the errors and somehow rectify the situation. In either case the issue of missing data introduces a potential bias and it is crucial to fully document the method by which missing data is handled. In the realm of clinical trials, only a minority of data analyses bothers to describe their chosen method for handling missing data [14].

Mutability Mutability refers to the ability to alter the data held in a data object or to change the identity of a data object. Serious Big Data is not mutable. Data can be added, but data cannot be erased or altered. Big Data resources that are mutable cannot establish a sensible data identification system, and cannot support verification and validation activities. The legitimate ways in which we can record the changes that occur in unique data objects (e.g., humans) over time, without ever changing the key/value data attached to the unique object, is discussed in Section 8.2.

For programmers, it is important to distinguish data mutability from object mutability, as it applies in Python and other object-oriented programming languages. Python has two immutable objects: strings and tuples. Intuitively, we would probably guess that the contents of a string object cannot be changed, and the contents of a tuple object cannot be changed. This is not the case. Immutability, for programmers, means that there are no methods available to the object by which the contents of the object can be altered. Specifically, a Python tuple object would have no methods it could call to change its own contents. However, a tuple may contain a list, and lists are mutable. For example, a list may have an append method that will add an item to the list object. You can change the contents of a list contained in a tuple object without violating the tuple's immutability.

Parallel computing Some computational tasks can be broken down and distributed to other computers, to be calculated "in parallel." The method of parallel programming allows a collection of desktop computers to complete intensive calculations of the sort that would ordinarily require the aid of a supercomputer. Parallel programming has been studied as a practical way to deal with the higher computational demands brought by Big Data. Although there are many important problems that require parallel computing, the vast majority of Big Data analyses can be easily accomplished with a single, off-the-shelf personal computer.

Protocol A set of instructions, policies, or fully described procedures for accomplishing a service, operation, or task. Protocols are fundamental to Big Data. Data is generated and collected according to protocols. There are protocols for conducting experiments, and there are protocols for measuring the results. There are protocols for choosing the human subjects included in a clinical trial, and there are protocols for interacting with the human subjects during the course of the trial. All network

communications are conducted via protocols; the Internet operates under a protocol (TCP-IP, Transmission Control Protocol-Internet Protocol).

Query The term "query" usually refers to a request, sent to a database, for information (e.g., Web pages, documents, lines of text, images) that matches a provided word or phrase (i.e., the query term). More generally a query is a parameter or set of parameters that are submitted as input to a computer program that searches a data collection for items that match or bear some relationship to the query parameters. In the context of Big Data the user may need to find classes of objects that have properties relevant to a particular area of interest. In this case, the query is basically introspective, and the output may yield metadata describing individual objects, classes of objects, or the relationships among objects that share particular properties. For example, "weight" may be a property, and this property may fall into the domain of several different classes of data objects. The user might want to know the names of the classes of objects that have the "weight" property and the numbers of object instances in each class. Eventually the user might want to select several of these classes (e.g., including dogs and cats, but excluding microwave ovens) along with the data object instances whose weights fall within a specified range (e.g., 20–30 pound). This approach to querying could work with any data set that has been well specified with metadata, but it is particularly important when using Big Data resources.

Raw data Raw data is the unprocessed, original data measurement, coming straight from the instrument to the database with no intervening interference or modification. In reality, scientists seldom, if ever, work with raw data. When an instrument registers the amount of fluorescence emitted by a hybridization spot on a gene array, or the concentration of sodium in the blood, or virtually any of the measurements that we receive as numeric quantities, the output is produced by an algorithm executed by the measurement instrument. Pre-processing of data is commonplace in the universe of Big Data, and data managers should not labor under the false impression that the data received is "raw," simply because the data has not been modified by the person who submits the data.

Results The term "results" is often confused with the term "conclusions." Interchanging the two concepts is a source of confusion among data scientists. In the strictest sense, "results" consist of the full set of experimental data collected by measurements. In practice, "results" are provided as a small subset of data distilled from the raw, original data. In a typical journal article, selected data subsets are packaged as a chart or graph that emphasizes some point of interest. Hence, the term "results" may refer, erroneously, to subsets of the original data, or to visual graphics intended to summarize the original data. Conclusions are the inferences drawn from the results. Results are verified; conclusions are validated.

Science Of course, there are many different definitions of science, and inquisitive students should be encouraged to find a conceptualization of science that suits their own intellectual development. For me, science is all about finding general relationships among objects. In the so-called physical sciences the most important relationships are expressed as mathematical equations (e.g., the relationship between force, mass and acceleration; the relationship between voltage, current and resistance). In the so-called natural sciences, relationships are often expressed through classifications (e.g., the classification of living organisms). Scientific advancement is the discovery of new relationships or the discovery of a generalization that applies to objects hitherto confined within disparate scientific realms (e.g., evolutionary theory arising from observations of organisms and geologic strata). Engineering would be the area of science wherein scientific relationships are exploited to build new technology.

Square Kilometer Array The Square Kilometer Array is designed to collect data from millions of connected radio telescopes and is expected to produce more than one exabyte (1 billion gigabytes) every day [8].

Supercomputer Computers that can perform many times faster than a desktop personal computer. In 2015 the top supercomputers operate at about 30 petaflops. A petaflop is 10 to the 15 power floating point operations per second. By my calculations a 1 petaflop computer performs about 250,000 operations in the time required for my laptop to finish one operation.

WorldWide Telescope A Big Data effort from the Microsoft Corporation bringing astronomical maps, imagery, data, analytic methods, and visualization technology to standard Web browsers. More information is available at: http://www.worldwidetelescope.org/Home.aspx

References

[1] Graur D, Zheng Y, Price N, Azevedo RB, Zufall RA, Elhaik E. On the immortality of television sets: "function" in the human genome according to the evolution-free gospel of ENCODE. Genome Biol Evol 2013;5:578–90.

[2] Whittaker Z. UK's delayed national health IT programme officially scrapped. ZDNet, September 22, 2011.

[3] Kappelman LA, McKeeman R, Lixuan Zhang L. Early warning signs of IT project failure: the dominant dozen. Information Systems Management 2006;23:31–6.

[4] An assessment of the impact of the NCI cancer Biomedical Informatics Grid (caBIG). Report of the Board of Scientific Advisors Ad Hoc Working Group. National Cancer Institute; March 2011.

[5] Komatsoulis GA. Program announcement to the CaBIG community. National Cancer Institute. https://cabig.nci.nih.gov/program_announcement [viewed August 31, 2012].

[6] Freitas A, Curry E, Oliveira JG, O'Riain S. Querying heterogeneous datasets on the linked data web: challenges, approaches, and trends. IEEE Internet Computing 2012;16:24–33.

[7] Drake TA, Braun J, Marchevsky A, Kohane IS, Fletcher C, Chueh H, et al. A system for sharing routine surgical pathology specimens across institutions: the Shared Pathology Informatics Network (SPIN). Hum Pathol 2007;38:1212–25.

[8] Francis M. Future telescope array drives development of exabyte processing. Ars Technica; 2012. April 2.

[9] Markoff J. A deluge of data shapes a new era in computing. The New York Times; 2009. December 15.

[10] Harrington JD, Clavin W. NASA's WISE mission sees skies ablaze with blazars. NASA Release 12-109; 2002. April 12.

[11] Committee on Mathematical Foundations of Verification, Validation, and Uncertainty Quantification; Board on Mathematical Sciences and Their Applications, Division on Engineering and Physical Sciences, National Research Council. Assessing the reliability of complex models: mathematical and statistical foundations of verification, validation, and uncertainty quantification. National Academy Press; 2012. Available from: http://www.nap.edu/catalog.php?record_id=13395. [viewed January 1, 2015].

[12] Data Quality Act. 67 Fed. Reg. 8,452, February 22, 2002, addition to FY 2001 Consolidated Appropriations Act (Pub. L. No. 106-554. Codified at 44 U.S.C. 3516).

[13] Worldwide LHC Computing Grid. European Organization for Nuclear Research. Available from: http://public.web.cern.ch/public/en/lhc/Computing-en.html; 2008 [viewed September 19, 2012].

[14] Carpenter JR, Kenward MG. Missing data in randomised control trials: a practical guide. November 21. Available from:http://www.hta.nhs.uk/nihrmethodology/reports/1589.pdf; 2007.

2

Providing Structure to Unstructured Data

OUTLINE

Section 2.1. Nearly All Data Is Unstructured and Unusable in Its Raw Form

> *I was working on the proof of one of my poems all the morning, and took out a comma. In the afternoon I put it back again.*
>
> *Oscar Wilde*

In the early days of computing, data was always highly structured. All data was divided into fields, the fields had a fixed length, and the data entered into each field was constrained to a pre-determined set of allowed values. Data was entered into punch cards with pre-configured rows and columns. Depending on the intended use of the cards, various entry and read-out methods were chosen to express binary data, numeric data, fixed-size text, or programming instructions. Key-punch operators produced mountains of punch cards. For many analytic purposes, card-encoded data sets were analyzed without the assistance of a computer; all that was needed was a punch card sorter. If you wanted the data card on all males, over the age of 18, who had graduated high school, and had passed their physical exam, then the sorter would need to make 4 passes. The sorter would pull every card listing a male, then from the male cards it would pull all the cards of people over the age of 18, and from this double-sorted sub-stack, it would pull cards that met the next criterion, and so on.

Principles and Practice of Big Data. https://doi.org/10.1016/B978-0-12-815609-4.00002-9

As a high school student in the 1960s, I loved playing with the card sorters. Back then, all data was structured data, and it seemed to me, at the time, that a punch-card sorter was all that anyone would ever need to analyze large sets of data. [Glossary Binary data]

How wrong I was! Today, most data entered by humans is unstructured in the form of free-text. The free-text comes in email messages, tweets, and documents. Structured data has not disappeared, but it sits in the shadows cast by mountains of unstructured text. Free-text may be more interesting to read than punch cards, but the venerable punch card, in its heyday, was much easier to analyze than its free-text descendant. To get much informational value from free-text, it is necessary to impose some structure. This may involve translating the text to a preferred language; parsing the text into sentences; extracting and normalizing the conceptual terms contained in the sentences; mapping terms to a standard nomenclature; annotating the terms with codes from one or more standard nomenclatures; extracting and standardizing data values from the text; assigning data values to specific classes of data belonging to a classification system; assigning the classified data to a storage and retrieval system (e.g., a database); and indexing the data in the system. All of these activities are difficult to do on a small scale and virtually impossible to do on a large scale. Nonetheless, every Big Data project that uses unstructured data must deal with these tasks to yield the best possible results with the resources available. [Glossary Parsing, Nomenclature, Nomenclature mapping, Thesaurus, Indexes, Plain-text]

Section 2.2. Concordances

> *The limits of my language are the limits of my mind. All I know is what I have words for. (Die Grenzen meiner Sprache bedeuten die Grenzen meiner Welt.)*
>
> *Ludwig Wittgenstein*

A concordance is a list of all the different words contained in a text with the locations in the text where each word appears. Concordances have been around for a very long time, painstakingly constructed from holy scriptures thought to be of such immense value that every word deserved special attention. Creating a concordance has always been a straightforward operation. You take the first word in the text and you note its location (i.e., word 1, page 1); then onto the second word (word 2 page 1), and so on. When you come to a word that has been included in the nascent concordance, you add its location to the existing entry for the word. Continuing thusly, for a few months or so, you end up with a concordance that you can be proud of. Today a concordance for the Bible can be constructed in a small fraction of a second. [Glossary Concordance]

Without the benefit of any special analyses, skimming through a book's concordance provides a fairly good idea of the following:

– The topic of the text based on the words appearing in the concordance. For example, a concordance listing multiple locations for "begat" and "anointed" and "thy" is most likely to be the Old Testament.