

The author and publisher have taken care in the preparation of this book, but make no expressed or implied warranty of any kind and assume no responsibility for errors or omissions. No liability is assumed for incidental or consequential damages in connection with or arising out of the use of the information or programs contained herein.

© Copyright 2010 by International Business Machines Corporation. All rights reserved.

Note to U.S. Government Users: Documentation related to restricted right. Use, duplication, or disclosure is subject to restrictions set forth in GSA ADP Schedule Contract with IBM Corporation.

IBM Press Program Managers: Steven M. Stansel, Ellice Uffer

Cover design: IBM Corporation

Associate Publisher: Greg Wiegand

Marketing Manager: Kourtnaye Sturgeon

Publicist: Heather Fox

Acquisitions Editor: Bernard Goodwin

Managing Editor: Kristy Hart

Designer: Alan Clements

Project Editor: Andy Beaster

Copy Editor: Keith Cline

Senior Indexer: Cheryl Lenser

Compositor: Nonie Ratcliff

Proofreader: Jennifer Gallant

Manufacturing Buyer: Dan Uhrig

Published by Pearson plc

Publishing as IBM Press

IBM Press offers excellent discounts on this book when ordered in quantity for bulk purchases or special sales, which may include electronic versions and/or custom covers and content particular to your business, training goals, marketing focus, and branding interests. For more information, please contact

U. S. Corporate and Government Sales

1-800-382-3419

corpsales@pearsontechgroup.com.

For sales outside the U. S., please contact

International Sales

international@pearson.com.

The following terms are trademarks or registered trademarks of International Business Machines Corporation in the United States, other countries, or both: IBM, the IBM logo, IBM Press, DB2, and WebSphere. Microsoft, Windows, Microsoft Excel, Microsoft PowerPoint, and the Windows logo are trademarks of Microsoft Corporation in the United States, other countries, or both. Googlebot is a registered trademark of Google Incorporated in the United States, other countries, or both. Apple, iPod, iTunes, and iPhone are trademarks of Apple Computer Incorporated in the United States, other countries, or both. Java and all Java-based trademarks are trademarks of Sun Microsystems, Inc. in the United States, other countries, or both. Other company, product, or service names may be trademarks or service marks of others.

Library of Congress Cataloging-in-Publication Data

Fishman, Neal.

Viral data in SOA : an enterprise pandemic / Neal A. Fishman.

p. cm.

Includes bibliographical references and index.

ISBN 978-0-13-700180-4 (pbk. : alk. paper) 1. Database management. 2. Business—Databases—Management. 3. Information storage and retrieval systems—Reliability. 4. Disinformation—Prevention. 5. Service-oriented architecture (Computer science)

I. Title.

QA76.9.D3F5835 2010

004.6'54—dc22

2009020221

All rights reserved. This publication is protected by copyright, and permission must be obtained from the publisher prior to any prohibited reproduction, storage in a retrieval system, or transmission in any form or by any means, electronic, mechanical, photocopying, recording, or likewise. For information regarding permissions, write to:

Pearson Education, Inc.

Rights and Contracts Department

501 Boylston Street, Suite 900

Boston, MA 02116

Fax (617) 671 3447

ISBN-13: 978-0-13-700180-4

ISBN-10: 0-13-700180-0

Text printed in the United States on recycled paper at R.R. Donnelley in Crawfordsville, Indiana.

First printing July 2009

Foreword



When Neal came to me to discuss his potential authoring of *Viral Data in SOA*, I was both very enthusiastic and concerned. I quietly wondered whether the broad audience that the book could potentially reach was really ready to come to terms with the true state of affairs with regard to business information? As IT professionals, do we have the means to stop the progression of *viral data*?

Look around you. A storm is brewing—a perfect storm of viral data, disinformation, and misinformation. The collapse of the world’s financial markets resulted in part from toxic assets that *appeared from nowhere* and humbled even the most mature and steadfast of institutions. During our careers, many of us have tried to invest a portion of our incomes using *trusted research* published by ratings agencies and from auditors’ reports. It’s all *trusted, unbiased information you can believe in*. Not!

As I reflect on the numerous projects that Neal and I have completed together at many of the top Global 2000 companies for both Ascential® and IBM®, it becomes clear to me that the trend and advance of viral data has now become pandemic. With a preponderance of evidence, I am still amazed that people don’t realize the effects that viral data has in their daily lives.

I am reminded of a recent data governance workshop Neal and I delivered to a major federal financial agency that had been caught up in the financial meltdown. Neal argued with the agency’s top architects about the imperative to establish a data quality and governance practice following the recent banking collapse to restore trust in their information. When one of the senior

architects retorted that they didn't need these things because they were already a "very conservative financial organization," Neal's response of "apparently not" said it all and brought a deadly silence to the meeting that lasted a full minute.

Within the agency, there had been an unawareness of how their IT practices contributed to viral data by delivering the wrong answers. When combined with other systems used by auditors, accountants, and rating agencies, the systemic problem arises as pan-enterprise and certainly pandemic. Of a greater concern is that a preponderance of corporate and government cultures remain satisfied with their *shared illusion* of regulatory compliance and transparency. Although many IT individuals are charged with fixing data issues, many inadvertently end up arguing against it!

Neal shows in this book: "What is believed to be true can be true, or false, or both true and false at the same time." One day your financial statement from Bernie Madoff says you are a millionaire with a 20 percent return, the next day you are the victim of a \$50 billion Ponzi scheme. This book strikes at the heart of the issues in SOA and the damaging approaches now being pursued by many vendors more interested in selling software than the long-term success of their enterprise customers.

Without a doubt, viral data in a service-oriented architecture has the capacity to become an enterprise pandemic and disable a company. SOA canonical data models present a fundamentally flawed approach by mapping malaligned data to a common form and promoting consumption of viral data without a data quality, standardization, alignment, and harmonization process. Service-oriented solutions through interoperability, reusability, layering of abstractions, and loose coupling can serve as perfect hosts to propagate misinformation: That is the knife's edge of SOA, *the perfect storm*. When the use of SOA achieves a high degree of interoperability within the corporate value chain, viral data can easily become a pandemic situation, becoming a perfect opportunity to miscommunicate with ubiquity and simultaneity.

One of my favorite parts in Neal's treatise is the discussion Neal shares about how PowerPoint® has potentially become responsible for destroying our ability to accurately communicate. However, Neal goes deeper, too, into other important areas such as governance, provenance, metadata, data modeling, data integration, master data management, trusted information, services-oriented architecture, and data chains.

So, are all of Neal's readers really ready to confront the true state of affairs in the enterprise today? Can we offer solutions for how to stop the progression of viral data? After reading this great book, I'm now thinking Neal should run for president and we should start our own political party, news channel, and religion. We can call it, "Trusted Unbiased Information You Can Believe In!"

Timothy G. Davis,
Boston, Massachusetts
April 2, 2009

Timothy G. Davis

Tim Davis is an Executive Director within the IBM Software Group. Tim's mission at IBM is to drive the Trusted Information Agenda through the advancement of industry-leading architectures for IBM customers worldwide.

Tim is also one of the leading content contributors for IBM curriculums, best practices, and methodologies. Tim is the founder of the IBM Center of Excellence for Data Integration, and recently led the development and launch of IBM's Information Grid, MDM Server Rapid Deployment, and IBM's SAP deployment accelerators.

Tim has more than 25 years of professional experience in large-scale systems integration, high-performance computing, SOA, MDM, ERP/SAP, CRM, data warehousing, analytics, banking risk management, and compliance. Tim has published numerous papers and holds a Master of Science degree in Electrical Engineering/Computer Science from USC and a Bachelor's degree in the same discipline from Clarkson University.

Foreword



The evolution of technology has allowed for a great many capabilities. It has also pointed at each phase to the next big problem. In many ways, the world has reached a point where integrated data between silos is the key to prevention, reaction, and compliance. The lack of integrated data leads to failures in both public and private environments. The data must be central to the problem at hand.

Public safety, environment monitoring, economic development, financial risk management, education, and other challenges that many nations are facing are quite literally manageable by *knowing better*. But no one holds all the pieces of the puzzle. Looking at the pieces individually and acting may actually make the problem *worse*.

Many of these capabilities are locked within vertical silos. A vertical silo can be an application, it can be a business unit, and it can be a company. We have written and rewritten systems to meet the needs of these vertical silos.

But no matter how wonderfully effective and efficient the silos have become, the power of understanding information *across the silos* has become the key to knowledge. As we struggle to free data from the walls of its silo, we realize that the way technology has evolved within an organization, and between organizations has been without standards for processes, controls, and data. Getting good and useful information in real time across silos is a real cross-boundary challenge. It takes a great deal of governance to orchestrate the *connecting of the dots*.

In January 2008, the Mayor of New York City, Michael R. Bloomberg, gave his annual State of the City address to the citizens of New York City,

launching a bold new initiative called HHS-Connect. The vision behind HHS-Connect is to *connect the dots* between the various agencies in New York that provide health and human services. Shortly thereafter, I had the pleasure of being announced by the mayor as his chief information officer for health and human services and as the executive director for HHS-Connect.

HHS-Connect was established to link data from more than a dozen city agencies to enable caseworkers to share client information without compromising confidentiality. Mayor Bloomberg stated that “with HHS-Connect, caseworkers will now spend less time managing paperwork, and will be able to spend more time face to face with their clients. Building upon our online prescreening and eligibility tool, ACCESS NYC, beginning this year, in a first for any municipal government, we will link the computer systems of health and human service agencies, so that they can share client information without compromising confidentiality.”

The Deputy Mayor for Health and Human Services, Linda Gibbs, similarly commented that with “HHS-Connect, the City will fundamentally change how we provide services by connecting clients, agencies, and providers to ensure holistic and integrated services that wrap around a family. Data sharing will provide a more complete understanding of the clients’ needs that will ultimately lead to better client care.”

My team oversees technology strategy and architecture to ensure a coordinated approach to facilitate data integration and exchange between nine health and human services agencies under the direction of Deputy Mayor Linda Gibbs. The goal is also to exchange data with other New York City agencies, the state, federal agencies, and even third-party providers. In addition, giving clients new channels of digital access to their data and capabilities that allow for transactions online is critical to improving client experience.

New York City is one of the world’s largest and most vibrant cities, and being tasked with providing the all relevant stakeholders with an accurate and comprehensive single view of the New York resident is an exciting and rewarding challenge. My challenges range from leading the policy and legal effort of understanding what data can be shared with which parties under which circumstances to making sure each linked agency has access to the right cross-agency data in a timely manner to help ensure each citizen receives access to the services for which he or she is entitled.

HHS-Connect will not only increase the quality of life of the clients we serve in the city, but it will also help the city make better and quicker decisions, resulting in better outcomes for families. Collaboratively, I am part of a

broad team to make real progress toward the goal of integrated health and human services.

Helping to bridge all of these agencies exemplifies the importance and power of data—from seeing positive results attributable to accurate data, to the potential headaches associated with misinformation. The need to protect against viral data was a concept that resonated. We are building software services, we are using master data solutions, and we are using state-of-the-art techniques, hardware, and software. Ultimately, success will mean paying attention to the big picture, and building it piece by piece. *We are building the tracks while the train is in motion.*

In *Viral Data for SOA*, Neal positions the path for trustworthy data in the big picture, taking into account the organization, data needs, as well as software needs. Neal explores a number of topics with depth and insight, including data governance and data provenance.

While my job is fun, exciting, and full of daily challenges, my focus has to include making sure that I can reliably provision data to each agency that HHS-Connect supports. Achieving trustworthiness in data is a goal. Avoiding viral data is a mandate. Without good and consistent underlying data, the usefulness of *connecting the dots* will be lost.

Kamal Bherwani

Chief Information Officer Health and Human Services

Executive Director HHS-Connect

New York City

March 24, 2009

Preface

The overall quality of information in our organizations continues to be suspect and poor. Spurred by needs to eliminate duplication, cut costs, and provide *greener* solutions, organizations have renewed efforts to logically centralize IT systems through virtualization, federation, and shared services.

A shared data resource exposed through a layer of services can behave like a virus—unilaterally affecting all those who touch the data. This book addresses the treatment and prevention of harmful data in a service-oriented architecture.

Chapter 1, “Viral Data.” Chapter 1 explores the potential viral consequences of data in a services-oriented architecture is explored. The chapter provides examples of how different anomalies in data can have everything from mild to serious consequences across the organization.

Topics that Chapter 1 covers include the following: the silo structure of organizations; problems with trying to align IT with the business; eminent domain and management prerogative; different ways to view data quality; the interpretation of metrics; effects of positive and negative feedback mechanisms; SOA characteristics and how they equate in data design; and how data designs often force tight coupling.

Chapter 2, “Data Governance.” Many organizations have tried to manage their information quality, but as organizations centralize their services and data layers, a need has arisen to add a layer of governance to the data management framework. This chapter explains what data governance adds to the organization and provides several implementation models.

Topics that Chapter 2 covers include the following: the ubiquity of governance; the benefit of permitting deviations; oversight; basic tenets of business; rules of compliance; rules of behavior; self-governance; master data

becoming aggregation hubs; intragovernance versus intergovernance; communication styles; a metamodel for data governance; data governance bodies; a framework for data governance; enforcement; dialing as a means of success; and assessing risk.

Chapter 3, “Reference Model.” This chapter reviews a reference architecture that provides a context for how to assess, condition, and recondition shared data. The reference architecture becomes a backbone for implementing an enterprise-wide data quality methodology.

The methodology is described through the use of a flight-path metaphor and supplements publish and subscribe scenarios that might be used in association with a service bus. Opportunities to address (assess or condition) data quality are explained in the reference architecture using the flight-path stages of preflight, in-flight, and post-flight.

Topics that Chapter 3 covers include the following: transient and persistent data; metadata for structured, semistructured, and unstructured data; the life span of data; the Shannon-Weaver communication model; understanding the processes in which data moves through an organization; data flight paths; customizing data quality assessments to the data flight path; the differences between data alignment and data harmonization; types of work efforts on data quality; benefits of following a scenario-based analysis; using probabilistic and deterministic matching; and merging and surviving moved data.

Chapter 4, “Assessing the Damage.” Guided by the reference architecture, this chapter explains how to approach assessing the quality of information in an organization. The chapter reviews what aspects of data can be addressed: from modeling techniques, to the choices in hardware, to all the different types of data persisted in an enterprise.

Topics that Chapter 4 covers include the following: the impact of revisionism; hidden meanings embedded in business data and overloading concepts; keeping the persistence design separate from the presentation; working with abstract concepts; constructing definitions; targeting definitions to a defined audience; *isa* and *hasa* relationships; different techniques for rationalizing and interpreting data; normal forms; the visualization of data models; how to assess ownership of data; a taxonomy for classes of data; and data-value patterns.

Chapter 5, “Data Conditioning.” This chapter explains how to condition and recondition data. Conditioning is the augmenting of data to improve its level of quality either at the time of creation or soon thereafter. Reconditioning largely covers the subject of data decay and the augmenting of data previously conditioned. The chapter discusses data provenance as a

means to help manage and evaluate the condition of data. Examples are shown as to how metadata can be used to create reactive systems.

Topics that Chapter 5 covers include the following: reductionism; systems thinking; how to use data provenance; applying value chains and data chains; implications of early- and late-binding techniques; data lineage and conditional data lineage; metadata as a late-binding agent; a technique to evaluate and measure the stability of a business concept; data decay; and reclassifying business products.

Chapter 6, “Putting in Place.” This final chapter describes additional steps that enable an organization to implement a data quality initiative for managing and governing shared data used in an SOA.

Topics that Chapter 6 covers include the following: manipulating information for personal gain; creating software that has a continual ability to react to business needs; how to create abstract designs; metamodels; practical distinctions between conceptual, logical, and physical data models; data architecture; the use of stories to explain an abstraction; the importance of context; and master data management.

Who Should Read This Book

Viral Data in SOA: An Enterprise Pandemic has content that will appeal to a diverse business and technical audience. On the business side, subject matter experts, business liaisons, and managers who rely heavily on data from information technology departments to perform their functions are likely to benefit from the first half of the book.

Technical specialists of any *species* will find value in the holistic approach to gaining trusted information in a service-oriented environment. Therefore, this book will appeal to IT managers, analysts, programmers, data administrators, database administrators, enterprise architects, technical architects, system architects, SOA architects, and data architects.

This book covers a wide range of topics, including the following (alphabetically), and so will likewise appeal to a diverse audience with interest in any or all of the following:

Application design
Correcting data
Data abstractions
Data architecture

Data chains
Data governance
Data modeling
Data profiling

Data provenance	Reactive modeling
Data quality assessment	Semi-structured data
Database modeling	Service-oriented architecture
Eminent domain	Structured data
Information quality	Systems thinking
Master data management	Trusted information
Metadata	Unstructured data
Quality scorecards	XML

What You Will Learn

This book is a comprehensive guide to understanding:

- The importance of addressing data quality as part of a multidisciplinary effort
- The need to establish or reestablish data design principles
- That achieving trusted information requires a lifetime commitment

Introduction

In the book of Genesis, Adam and Eve both consume fruit from the tree of the knowledge of good and evil.¹ Although the type of fruit is not explicitly mentioned, for many centuries the apple has survived as a popular guess.

Potentially, the inference to an apple is sustained by the term given to the noticeable protrusion on the neck of many (human) males. The laryngeal prominence, commonly known as an Adam's apple, is the result of thyroid cartilage surrounding the larynx. The associative anatomical term, *Adam's apple*, gives further credence to the popular belief that the forbidden fruit was, in fact, an apple.

In the centuries that followed the death of Jesus Christ, the Old and New Testaments were translated into Latin from their original languages (Hebrew for the Old Testament and Greek for the New Testament). Then, in the early 1380s, Oxford Professor John Wycliffe (1330–1384) led a translation of the Old and New Testaments into English.²

Until that time, Latin had been the predominant language for these Testaments in the Christian world. In Latin, the word for evil (as in the tree of the knowledge of good and evil) is *malum*.³ As a Latin word, *malum* is a homonym and has other meanings besides evil. An alternative translation or meaning of the word *malum* is *apple*.⁴

Hebrew texts translated into Latin and then translated from Latin into English might be one reason how an unnamed fruit obtained an identity. Could that be true? Was the apple chosen as the result of a homonym? What is the actual provenance⁵ of the apple in the Adam and Eve story? How has the apple *persisted*⁶ over time?

1. See Hertz (2001).
2. See McGrath (2001).
3. See Simpson (1977).
4. *Ibid.*
5. Provenance, specifically data provenance is a topic that is discussed in detail in Chapter 5.
6. In terms of viral data and data in general, *persistence* is the opposite of volatile or transient data and is used here as a play on words.

To venture into understanding meanings in communication or to delve into linguistics is to study semantics. What one person means (to say) and another person takes away (as being said) can represent a significant challenge with both verbal and written styles of communication. The discourse of viral data is concerned not only with verbal and written communication, but the communication and linguistics involved in the sharing of data between software programs—especially, those deployed as part of a service-oriented architecture (SOA).

The title of this book relies on you, the reader, forming a connection between terms more commonly used in physiology or medicine with those found in information technology (IT). The use of analogies and other explanatory devices such as metaphors and allegories can often assist in making a point or clarifying a point—even if only to provide a generality (a gist).

IT is no stranger to the concept of analogical association; many IT terms have been adopted because of their analogical ties, including the following:

- Software *engineer*
- Technical *architect*
- *Chief* technology *officer*
- *Deployment*
- Data *governance*
- *Service-oriented architecture*

The terms *engineer*, *architect*, *chief*, *officer*, *deployment*, *governance*, *service*, and *architecture* are all borrowed from other previously existing disciplines. The business side of an organization does the same thing—terms such as business *strategy*, *tactics*, and *mission* statement all carry nuances from their original intent into a different paradigm. In this case, the terms *strategy*, *tactics*, and *mission* are military words redeployed for a seemingly more peaceful use in commerce.

Regarding analogies, law professor Cass Sunstein (b. 1954) wrote,⁷ “Ordinary people make sense of the world by discerning patterns routed in analogical thinking.” As we are starting to see, analogies and other similar devices can assist the art of communication by helping to correlate what is already known or understood with other new or more complex concepts.

7. See Sunstein (1998).

In 1966, the cognitive psychologist Peter Wason (1924–2003) devised a logic puzzle known as a Wason test.⁸ The Wason test consists of four cards with a letter of the alphabet on one side and a number on the other. When placed on top of a table, the participant, initially, only gets to see one side of each card. If the visible side of each card reads L, A, 6, and 11, could you determine the minimum number of cards to turn over to prove the following rule as being either true or false: When a card shows an L, the flipside is always a 6.

Before the answer is revealed, if the letters and numbers are substituted for something else (for example, DB2[®], Java[™], database, and network) and then restate the question as follows: When a card shows DB2, the flipside is always a database. Deducing that the correct answer is DB2 and network is easier in this particular exercise—assuming that you are already familiar with the terms *DB2*, *Java*, *database*, and *network*.

Without some type of supporting information, a recognizable concept or term is often easier to deal with than an abstract concept or term. In both cases, the underlying problem remained the same; fundamentally, the only change that occurred was the potential familiarity or association with what was depicted on the cards and the way the question was phrased.

The answer to the first question is L and 11.

Associative linguistic devices such as analogies, metaphors, and allegories open a porthole by which an unknown or unfamiliar concept can readily become familiar by establishing a connection to something that is, hopefully, more familiar. Other forms of aids to help clarify meaning include innuendos, proverbs, and similes.

In a legal setting, innuendos can be used to provide an interpretation of words where the meaning is not too obvious. The use of proverbs and similes often provide just one side of the analogy or metaphor, requiring the reader to complete the association. The proverb, *you cannot make an omelet without breaking eggs*, speaks to the fact that progress often requires change of some manner—possibly irrevocably. What is or needs to *change* is left up to the reader to complete.

Viral data is a metaphor used to indicate that business-oriented data can exhibit qualities of a specific type of human pathogen: the virus.

In humans, the classification scheme for disease-causing organisms comprises five categories: viruses, bacteria, protozoa, fungi, and worms.

8. See Newstead (1995).

Interestingly, the virus is the most simplistic of the pathogens and parasites. In our human world, viruses and bacteria are the germ-causing agents of disease. Viruses and bacteria have been responsible for the great plagues and, yes, pandemics.

The term *virus* was coined by a Dutch botanist Martinus Beijerinck (1851–1931).⁹ The word *virus* is also Latin, literally meaning *poison* or *poisonous slime*. Now, data by itself is inert. Data requires software (or people) for the data to appear alive (or actionable) and cause a positive, neutral, or negative effect. The inert and actionable properties of data are covered throughout this book.

In biology, all living things are cellular with the exception of the virus. Absent in the virus are any of the structures within a cell necessary to perform the life (action) activities such as eating, energy production, and growth. Like data, a virus is inert.¹⁰ A virus is an inert particle—tiny and lifeless.

If a piece of data is never used by a computer program or seen by a person, the data cannot have a negative or infectious effect. When data gets inside a computer program or appears on a screen that is scanned by a person, that data has entered an actionable world. The inert biological virus is skilled at getting into a cell—the virus' version of an actionable world. Once the virus is inside a cell, the game of infection begins.

A virus outside of a cell is not dangerous; likewise, digital data outside of a computer program is not dangerous, and neither is printed data that is not viewed or referenced in any manner.

To be a pandemic requires more than just a widespread disease—even if the disease causes many unfortunate deaths. Cancer is a deadly widespread disease, but cancer is not a pandemic because the disease is, *generally believed*, not to be infectious. To be classified as a pandemic, the disease (or condition) needs to be widespread *and* infectious.

According to the World Health Organization, “Once a pandemic begins it will be too late to accomplish the many key activities required to minimize the impact.”¹¹ Within the boundary of an enterprise, viral data can be pandemic when a service-oriented architecture achieves high degrees of interoperability throughout the corporate value chain while leveraging synchronous data stores.

9. See Shors (2009).

10. See Gillen (2007).

11. See WHO (2004).

To paraphrase science writer Bernard Dixon,¹² directionally set magnetized ferromagnetic material is oh so small (the analogical 0s and 1s on a hard drive). By comparison, each Forbes Global 2000 company is absolutely huge, employing thousands of people, across numerous countries in numerous locations. The magnetized ferromagnetic material can kill a Forbes Global 2000 company.

A generation for humans is measured approximately by a 30-year interval. For some viruses, the equivalent measure is less than 5 days. Furthermore, for some bacteria, the time span from one generation to the next can be as little as 10 minutes. As for data traveling through the real-time enterprise, the generational time from one data store to the next can probably be measured in milliseconds.

Stock exchanges around the world now permit brokerages to co-locate their servers within an exchange's data center to help reduce latencies—if only by nanoseconds. “And latency affects more than execution; it also impacts prices and the distribution of data inside an enterprise.”¹³

The Spanish proverb “the beginning of health is to know the disease” infers that one must stand on a strong foundation if one is to move forward with success. A necessity for gaining the upper hand on viral data is to scrap some of the data-oriented dogma that has existed since the late 1960s and rethink many of the precepts associated with data.

Master of the modern short story, Anton Chekov (1872–1937) wrote, “When you're offered a thousand remedies you can be sure the disease is incurable.”¹⁴ Solutions for viral data in SOA may not be unilaterally preventative in all business situations, but the more one understands about the causes and what symptoms to look for, the better chance the enterprise has to control its pathogens.

Edward Tenner (b. 1943), coined the phrase *revenge theory* and has observed in a very Jerry Seinfeld (b. 1954) manner that automobiles move through London's congested streets at the same pace as horse-drawn carriages from the mid-1850s and that computers have no real effect on productivity because people learn to complicate and repeat tasks that have been made easier. On decision making, Tenner has remarked:

12. “A small bacterium weighs as little as 0.000000000001 [one-trillionth] grams. A blue whale weighs about 100,000,000 [one hundred million] grams. Yet a bacterium can kill a whale.” From Dixon (1996).
13. See Crosman (2008).
14. See Chekhov (1996).

From 1981 to the present, American businesses have spent hundreds of millions of dollars on software to help improve decision-making. Software producers are now disbursing millions of dollars on research to make their programs even better and easier to use. Spreadsheets and other decision-making aids are now ubiquitous in academia and the professions as well as in business. Thus it is all the more remarkable how little research has been devoted to the effects of computers on the quality of decision-making.¹⁵

The fact that Tenner's reference to the present was the year 1996 makes it all the more impressive that the argument still remains valid. In part, the progress of decision making has been hampered by inconsistent, inaccessible, and *nonexistent* data. Folklore has it that there are three degrees of falsehood: First there is the fib, then the lie, and finally, there is statistics.

Ordinarily, knowledge workers in a corporation do not seek to record or manufacturer information that deviates from the truth. However, misinformation can systemically happen in a business—the viral data can enter through sheer carelessness, a business rule, or a host of different ways.

What is believed to be true can be true, or false, or both true and false at the same time. Again, provenance is a concept that can prove very helpful in establishing truthfulness in data—trusted information.

The modern world of sports is big business, from American collegiate football stadiums that are able to accommodate well over one hundred thousand sitting fanatics;¹⁶ to the payment of \$894 million by NBC Universal to acquire the American television rights for the 2008 Olympic Games held in Beijing, China;¹⁷ and to the professional sports player, where contracts for high-caliber individuals like David Beckham (b. 1975) can extend into the hundreds of millions of dollars.¹⁸

Nearly everyone involved in sports is drawn to statistics—from the soccer mom hoping her child receives enough playing time, to workplace discussions around the water cooler that deliberate the minutest fractions of time that determined who stood on the winner's podium, to coaches, players, and fans.

15. See Tenner (1997).

16. Michigan Stadium, also known as the Big House, is home to the Wolverines, the football team of Michigan University. During a single game in the fall of 2003, the stadium accommodated more than 112,000 people. See Campbell (2007).

17. See Carter (2008).

18. See Bell (2007).

Baseball is a competitive sports game that is played professionally in a number of countries and on occasion has been a medal competition at the summer Olympic games.¹⁹ In a U.S. Major League Baseball (MLB) game held on April 28, 2008, the Baltimore Orioles beat the Chicago White Sox 4 to 3.²⁰ The winning run, finally, came in the fourteenth inning, and the Orioles pitcher, Alberto Castillo (b. 1975), earned his first Major League victory. In all MLB games, for historical and statistical purposes, one of the pitcher's on the winning side is given the victory.

Of particular note during that Orioles and White Sox game was that Castillo had yet to start his MLB career. In fact, Castillo and a handful of other players were not on either team's rosters on that date. Before the game could be concluded, rain halted play, and the teams had to wait nearly four months for the game to be restarted. The game was restarted and concluded on August 25 in Baltimore, Maryland. Officially, all the statistics for the game have been attributed to April 28 in Chicago, Illinois. You can extrapolate a lesson from this: Even in business, statistics can, on occasion, distort the truth.

Without the provenance to go along with the statistics, decisioning²¹ on the truth might become problematic. Statistics or other derived or aggregated data become part of the semantic landscape for communicating (whether verbally, in writing, or as a message exchanged between services).

Service-oriented architectures build on the concepts of distributed computing and modular programming. Services are often grouped as interoperable packages. In a general sense, interoperability in software is a "capability to communicate, execute programs, or transfer data among various functional units in a manner that requires the user to have little or no knowledge of the unique characteristics of those units."²² Essentially, interoperability refers to a capability of two or more programs to exchange and use exchanged data.

19. See Zoss (2004).

20. See Viera (2008).

21. *Decision* is a noun. However, by adding *-ing*, the word is made into a verb. This technique is known as a gerund. The use of the word decisioning as a verb rather than a noun such as decision or decision-making, helps to reinforce the mental picture of an action. Mechanisms that are capable of carrying out an action are the means by which a virus is propagated.

22. See TC (2003).

During each interchange, data can be further manipulated, persisted, or decisioned on like data in any traditional programming environment. Services that are orchestrated to interoperate in a real-time mode are on the one-hand extending the capabilities that IT departments can bring to the business and, on the other hand, are capable of causing unimpeded havoc.

The ability to turn disparate or stand-alone functionality into seamless-strung business processes or distribute the same data from a common data store to disparate processes creates for viral data a *perfect storm*,²³ a perfect opportunity to miscommunicate with ubiquity and simultaneity—a service-oriented pandemic reaching all corners of the enterprise.

A starting point in the search to control viral data in SOA is to begin with the semantics associated with the *lingua franca*. A subject matter expert from the business may know that preferred customers are offered a standard 15 percent discount, or that Federal Express delivers all plastic kumquats from suppliers located in Tennessee, or that excessive humidity can cause quality problems on the production line, without knowing exactly why these facts are true.

From an epistemological point, a subject matter expert can know that a *certain something* is true without fully understanding why that *certain something* is true in the first place. The functioning enterprise naturally allows conflicts in its vocabulary and is generally capable of putting into place exception criteria to manage undesirable behavior on a whim without needing to fully comprehend the ramifications. But then, sometimes, what we believe to be true turns out to be false.

Consider, for example, the history surrounding the creation of the Federal Reserve System. The Federal Reserve is an intricate part of monetary policy and the money system in the United States. Popular economics textbooks suggest that Congress created the Federal Reserve in 1913 as the central bank and monetary authority of the United States.²⁴

The economic crisis of 1907 had led the U.S. Congress to study the shortcomings of the American banking system and, eventually, to establish the Federal Reserve System.²⁵ The panic of 1907, with its more-than-usual

23. A *perfect storm* connotes a particularly bad situation arising from a large number of negative and, often, unpredictable contributory factors. As a metaphor, a perfect storm is normally attributed to a fierce storm arising from a rare combination of adverse meteorological factors. See OED (2008).

24. See Baumol (2009).

25. See McEachern (2009).

epidemic of bank failures, was the straw that broke the camel's back: The country was fed up once and for all with the anarchy of unstable private banking.²⁶

Despite these descriptions, the Federal Reserve System is not federal, and there are no reserves. Further, the Federal Reserve Banks are not even banks. The Federal Reserve System is a privately owned corporation, and its ownership, in the United States, has control over all things that deal with money.

The initial plan for the Federal Reserve was drafted at a secret meeting held in November 1910 at the private resort of J. P. Morgan (1837–1913) on Jekyll Island, located off the coast of Georgia. The seven meeting attendees represented an estimated one-fourth of the world's total financial wealth and included high-powered banking moguls, the leader of the Republican Party in the U.S. Senate, and the assistant secretary of the Treasury.

In attendance at Jekyll Island were Nelson Aldrich (1841–1915), Abram Piatt Andrew, Jr. (1873–1936), Frank Vanderlip (1864–1937), Henry Davison (1867–1922), Charles Norton (1870–1923), Benjamin Strong (1872–1928), and Paul Warburg (1868–1932).

At Jekyll Island, the meeting attendees conspired to:²⁷

- Stop the growing influence of small, rival banks, and to ensure that control over the nation's financial resources remained in the hands of those present
- Make the money supply more elastic (available) to reverse the trend of private capital formation and to recapture the industrial loan market
- Pool the meager reserves of the nation's banks into one large reserve so that all banks would be motivated to follow the same loan-to-deposit ratios (to protect at least some of the banks from currency drains and bank runs)
- Should this cartelization approach lead ultimately to collapse of the whole banking system, shift the losses from the owners of the banks to the taxpayers

Finally, in December 1913, President Woodrow Wilson (1856–1924) signed into law the Federal Reserve Act. The act that came into law did not

26. See Samuelson (2005).

27. See Griffin (2002).

include everything that the gang of seven had hoped, but in due course, with the inclusion of more than a hundred amendments, the cartel, known as the Federal Reserve System, achieved all of its initial objectives.

In business, some truths, for one reason or another, are hidden. Other truths are concealed because of a lack of information or because of the way the information is presented. For example, the Jekyll Island story, with its factual information about the formation of the Federal Reserve, is not as well known as some fictional (perhaps even conspiratorial) tales about its formation.

Implicit information hiding is why some people suggest that Microsoft® PowerPoint® has destroyed our ability to adequately communicate. For example, Edward Tufte (b. 1942), a professor at Yale University, believes that PowerPoint played a significant role in the Columbia space shuttle tragedy in 2003 because a Boeing representative did not, during a PowerPoint presentation to NASA,²⁸ fully convey the risks associated with the shuttle's reentry.

From school, to government, to work, speaking in abbreviated sound bites—bullet points—has become a natural way for us to communicate both simple and complex issues. At times, our speech, and not just our writing, becomes all too terse, as well. Over time, news about the Iraq war has been whittled down to small phrases: It is a civil war, it is a quagmire, America cannot be victorious, America needs to leave, America is winning, and America is fighting terrorism over there.

After repetitive use of what can be described as PowerPoint bullet points, some individuals discuss and debate the war by repeating bullet points and without exploring the topics further. Unexplored bullet points can become ingrained thoughts, beliefs, and accepted facts.

In a story about the Iraqi war, General Tommy Franks (b. 1945) remarked that, "It's quite frustrating the way this works, but the way we do things nowadays is combatant commanders brief their products in PowerPoint up in Washington to the Office of the Secretary of Defense (OSD) and the Secretary of Defense... In lieu of an order, or a fragmentary order, or plan, you get a set of PowerPoint slides... That is frustrating, because nobody wants to plan against PowerPoint slides."²⁹

Relying on PowerPoint slides rather than formal written orders has been thought by some military professionals to capture the amateurish approach to war planning. Franks has also been quoted as saying, "Here may be the

28. See Pigeon (2007).

29. See Ricks (2006).

clearest manifestation of OSD's contempt for the accumulated wisdom of the military profession and of the assumption among forward thinkers that technology—above all information technology—has rendered obsolete the conventions traditionally governing the preparation and conduct of war.”

Furthermore, Colonel Andrew Bacevich (b. 1947) has quipped, “To imagine that PowerPoint slides can substitute for such means is really the height of recklessness.” As an analogy, imagine an automobile mechanic who uses a manufacturer's glossy sales brochure to figure out how to repair an engine.

Consider the trend in business intelligence that leverages knowledge through a series of dashboards. One wonders whether the brevity and semantic stumbling blocks of PowerPoint will be repeated in the use of management dashboards.

To quote physicist Richard Feynman (1918–1988), “You can know the name of a bird in all the languages of the world, but when you're finished, you'll know absolutely nothing whatever about the bird... I learned very early the difference between knowing the name of something and knowing something.”

What does this mean? What does that mean? Trying to determine what information is important and what information possibly represents excessive *noise*³⁰ exemplifies a culture facing information overload. “Everyday living is too fast, too busy, too complicated. More than at any time in history, it's important to have good information on just about every aspect of life, and there's more information available than ever. Too much, in fact. There's simply no time for people to gather and absorb the information they need.”³¹

The quote about everyday living too fast is not recent. That quote is attributed to Briton Hadden (1898–1929). Hadden's remark about excessive information was made around 1922, a year before he started *Time* magazine with Henry Luce (1898–1967).

Semantically, we can pose a simple question: What is a book? On the surface, this benign question may seem like it has a simple answer. However, some people might limit their definition of a *book* to editions that are presented in hardcover or paperback. Others may argue that even the story itself, is the *book*. In turn, are abridged books *real* books? What about books found on other media such as audiobooks?

30. The term *noise* infers limited, of little, or of no perceived value.

31. See Harper (2004).

Perhaps we can define a book as something that has an assigned International Standard Book Number (ISBN). Consider the popular 2003 novel *The Da Vinci Code* written by Dan Brown (b. 1964). Brown's book, like the majority of books sold commercially was given an ISBN, but not just one. The following is a small sampling of the ISBNs assigned to Brown's *The Da Vinci Code*:

ISBN-10: 0375432302	ISBN-13: 978-0375432309
ISBN-10: 0385504209	ISBN-13: 978-0385504201
ISBN-10: 0385504217	ISBN-13: 978-0385504218
ISBN-10: 0385513224	ISBN-13: 978-0385513227
ISBN-10: 0385513224	ISBN-13: 978-0385513227
ISBN-10: 0385513755	ISBN-13: 978-0385513753
ISBN-10: 0552154016	ISBN-13: 978-0552154017
ISBN-10: 1400079179	ISBN-13: 978-1400079179

What royalties are due to Mr. Brown? Answering this question can pose a difficult semantic question depending on how the question is asked, how and where the information is stored, who has access to what information, and what, if any, assumptions were implied in the way the question was initially phrased.

If establishing meaning can be difficult even when the information present is correct, the latency involved in uncovering incorrect information can be outright painful. The town of Valparaiso, Indiana, lies about 40 miles southeast of Chicago, Illinois. The town has a population approaching 30,000, and prior to 2006 had a township budget of \$21 million.

In 2004, Dennis Charnetzky (b. 1974) and Daelyn Charnetzky (b. 1974) started a series of repairs to their two-bedroom home located in Valparaiso. The Charnetzky's renovated their bathroom, refinished their hardwood floors, added a splash of paint, and put up some new wallpaper. These improvements caused the property value shoot up nearly \$400 million.³²

Because the property value increased, the county's computers automatically increased the Charnetzky's property tax liability. With an increase in the tax base, municipal budgets were raised accordingly. By the time the typo that was entered into the county computer system was discovered, the Valparaiso school district and government agencies faced a financial shortfall and were forced to cut budgets by \$3.1 million.

32. See Ruethling (2006).

Another type of semantic dilemma involves the ISO 4217 standard that is used to govern currency types and symbols. Over time, the ISO standard has transitioned beyond its intended scope. Examples of an ISO currency symbol include USD for United States dollars, BRL for the Brazilian real, and DKK for the Danish krone.

However, the symbol XXX is part of the standard, but does not represent the currency of any nation. Code XTS is reserved for testing purposes, while XAU and XAG are codes for precious metals. CLF, USN, USS, and a number of other codes are for bonds or other fund types. At the semantic level, ISO 4217 covers more than currencies.

Political boundaries provide an interesting dynamic in semantics. For example, many companies do business in England. However, England is not a sovereign country. The sovereign country of which England is a part is the United Kingdom of Great Britain and Northern Ireland. However, England as well as Scotland, Wales, and Northern Ireland are regarded as countries when participating in sporting events such as soccer, rugby, and cricket. This is not the case in all sporting events. For example, for the Olympics, the United Kingdom enters a team under the moniker GB (Great Britain).

The latitude afforded the United Kingdom would be equivalent to Germany entering teams from East Germany, West Germany, and Bavaria, or the United States of America entering teams from the Union, Confederacy, and Texas.

By way of a final example regarding political boundaries, since 1997, Hong Kong has been part of China. When analyzing trade with China, however, one has to know whether or not Hong Kong is included. Hong Kong operates with a degree of autonomy, at least until 2047 (50 years after the official transfer back to China from United Kingdom).

The viral affects of data are without bounds and can, under certain conditions, reach a pandemic state. For example, corporations can be affected by laws in other countries as well as the actions of other corporations—regardless of geographic distance. Terra Gruppen is an umbrella company located in Norway and consists of almost 80 local savings banks.

Four of the communities severed by Terra Gruppen are located in the Arctic Circle—Rana, Hemnes, Hattfjelldal, and Narvik. Each community faced bankruptcy as a ripple-effect of the subprime fiasco that started in the United States.³³ One of Terra Gruppen's company's knowingly violated a code of conduct put in place by Norwegian financial regulators resulting in many lives being seriously and negatively impacted.

33. See Werdigier (2007).

Index

Page numbers followed by *n* indicate topics found in footnotes.

Numerics

3NF (third-normal form), 121

A

abduction, 114

abstract concepts

 element coupling, 132

 modeling, 106-107

 providing definitions for, 107-113

abstract design patterns

 canonical data models, 185

 conceptual data models, 185

 context, considering, 188-194

 enterprise data models, 186-187

 logical data models, 180-186

 physical data models, 180-184, 188

abstraction, 178

 life span of services and, 178

 metamodels for, 178-181

 in OSAPI, 158

 provenance and, 194

 software abstraction, 178*n*

 views and, 194-201

abstraction layering, 31

 data structure design and, 42-43

accountability, provenance and, 146

accuracy, 167*n*

 in DQA, 86

action roles, oversight versus, 60

acronyms, specific to this book

 ASPECT, 79-82, 206

 CIDER, 88-97, 206

 DQA³, 85-88, 206

 ED-SODA, 65-66, 206

 FARMADE, 52-54, 206

 OSAPI, 157-160, 206

 QuIT CITeD, 130-132, 206

 SÉANCE, 89-97, 206

- Adam and Eve story, apple in, 1
- Adam's apple, 1
- adaptability of viral data
solutions, 177
- adaptation to change, 20-43
best practices, 36-38
decisioning environment
measurements, 33-37
eminent domain, 22-23
incentives and compensation, 32-33
metadata, 27-29
military, 29-30
quality measurement, 24-27
service-oriented architectures, 30-32,
38-43
viral data as symptom of, 24
- adjudication in SÉANCE
workflow, 96
- adjusting data governance, 64-70
- adoptability of insure controls, 57
- advancement (moving data), 80
- advocacy dial (data governance),
66, 69
- aggregation hub, 205
- Aldrich, Nelson, 9
- alignment
of business and information
technology, 15-20
CIA example, 16
multidivisional business structure,
16-20
viral data fostered by, 22
of data, 87*n*, 97-98
- American Express, 21
- American flag example (historical
revisionism), 103
- analogies, 2-3
- analysis (in DQA), 86
- Anderson, Laurie, 78
- Apple Computer, 21
- apple in Adam and Eve story, 1
- appraisal (in DQA), 87
- appropriation, 22-23
- arbitrator communication style, 53
- Ascential, xvii
- Ashley, 122
- ASPECT, 79-82, 206
- assessments, xxxvi
data assessments
in context, 134-138, 201-203
master data solutions, 205
multiple views for, 203-204
via redundancy, 204
scenarios for, 138
statistical sampling, 138
subjective and objective measures, 206
data quality assessment (DQA),
85-88
CIDER, 88-97
- assure controls (data governance
metamodel), 55-57
- atomic metadata, 74
- attributes
abstraction and, 188-194
naming, 182*n*
- audience, considering in writing
definitions, 112
- audits, 58
- authoritarian communication
style, 54
- B**
- Bacevich, Andrew, 11
- Baltimore Orioles, 7
- bandwidth, moving data, 72-73
- baseball, 7
- Basel II, 175*n*
- basic tenets, in business, 46-47

BCNF (Boyce-Codd Normal Form), 121

Beckham, David, 6

behavior, rules of, 47

Beijerinck, Martinus, 4

best practices, 36-38

Bible translations, 1

big bang theory, 143*n*

BigCo, 140*n*, 141-142

blind overloading, 107-108

Bloomberg, Michael R., *xxi*

books, definition of, 11-12

Boulware, Lemuel, 175

bounded codified data, 131

Boyce-Codd Normal Form (BCNF), 121

Brown, Dan, 12

business

adaptation to change, 20-43

best practices, 36-38

decisioning environment measurements, 33-37

eminent domain, 22-23

incentives and compensation, 32-33

metadata, 27-29

military, 29-30

quality measurement, 24-27

service-oriented architectures, 30-32, 38-43

viral data as symptom of, 24

alignment with information

technology, 15-20

CIA example, 16

multidivisional business structure, 16-20

viral data fostered by, 22

laws of, 46-47

business data

abstract concepts, providing definitions for, 107-113

blind overloading, 107-108

data models, 122-126

misinterpretation of, Social Security

Number example, 104-106

modeling abstract concepts, 106-107

reasoning

case-based reasoning, 115

classification, reasoning through, 116-117

consistency-based reasoning, 115

generalization, reasoning through, 116

heuristic reasoning, 115

imprecise reasoning, 115

logical inference, 114

nonmonotonic reasoning, 116

parametric reasoning, 116

probabilistic reasoning, 115

semantic reasoning, 116

specialization, reasoning through, 117-122

types of, 113-114

semantic evaluations, 122

business metadata, 123

business needs, failure to identify, 29

business rules, data misalignment with, 126-127

business transactions, IT transactions versus, 176*n*

business-speak, 186

C

Callan, Erin, 174

Canby, George, 103

canonical data models, 185

- Captain Scarlet, 76*n*
 Metcalfe, Paul, 108
- cardinality trait, evaluating, 133
- Caruso-Cabrera, Michelle, 174
- case-based reasoning, 115
- Castillo, Alberto, 7
- Central Intelligence Agency (CIA), 15
- certification, 58
- chain of custody, 145
- change, adaptation to, 20-43
 best practices, 36-38
 decisioning environment
 measurements, 33-37
 eminent domain, 22-23
 incentives and compensation, 32-33
 metadata, 27-29
 military, 29-30
 quality measurement, 24-27
 service-oriented architectures, 30-32, 38-43
 viral data as symptom of, 24
- Charnetzky, Dennis and Daelyn, 12
- chart of accounts in data
 governance, 64
- Chekov, Anton, 5
- Chicago White Sox, 7
- CIA (Central Intelligence Agency), 15
- CIDER, 88-97, 206
- circular definitions, 111
- classification
 data class taxonomy, 130-132
 reasoning through, 116-117
- codified data, 131
- column traits, 132
 cardinality, 133
 domain, 133
 format, 133
 precision, 132
 size, 132
- communication
 semantics and, 1-13
Adam and Eve story, 1
analogies, 2-3
books, definition of, 11-12
controlling viral data, 8
Federal Reserve System example, 8-10
information overload, 11-12
ISO 4217 currency standard, 13
political boundaries, 13
PowerPoint example, 10-11
property value mistake example, 12
sports statistics, 6-7
viral data as metaphor, 3-5
- Shannon-Weaver communication
 model, 206
- communication models
 for data movement, 77-79
 Shannon-Weaver, 206
- communication traits of
 governance, 52-54
- compensation plans, 32-33
- complex metadata, 74
- compliance, rules of, 47
- composite metadata, 74
- compression of data, 71-72
- compulsory purchase, 22-23
- concept-relationship-concept
 abstract metamodel, 178-181
- conceptual data models, 185-186
- concreteness, 178
- conditioning in structural data
 design, 155-170
- conflicts of interest, 49
- consistency-based reasoning, 115
- consolidation (moving data), 81
- constraints as information loss, 28

construction (CIDER), 88
context, xxxvi-xxxvii, 209
 in abstract design patterns,
 considering, 188-194
 for column/element traits, 132
 data assessment in, 201-203
**continuity controls (data
 governance metamodel), 58-59**
controlling viral data, 8
corporate governance, 46
course of action in SÉANCE
 workflow, 96
CRUD, 157*n*
cubism, 106
cultural name example
 (specialization reasoning),
 117-122

D

The Da Vinci Code (Brown), 12
daedal data classes, 132
data
 moving
 ASPECT, 79-82
 bandwidth considerations, 72-73
 communication model, 77-79
 data access, 76-77
 flight analogy, 82-88
 metadata, 73-76
 role in data governance, 61
 semistructured, 71, 73-76, 79, 93,
 101, 131-132, 206
 structured, 71, 73-74, 76, 79, 101,
 131-132, 155-156, 206
 unstructured, 28, 36, 71, 73, 76, 79,
 101, 132, 206
data access, 76-77

data assessment
 in context, 134-138, 201-203
 master data solutions, 205
 multiple views for, 203-204
 via redundancy, 204
 scenarios for, 138
 statistical sampling, 138
 subjective and objective
 measures, 206
data chains, value chains versus,
 144-145
data class taxonomy, 130-132
data compression, 71-72
**data conditioning in structural data
 design, 155-170**
data decay, 170-171
data governance, 46-48
 data, role of, 61
 data quality versus, 59
 dialing (adjusting), 64-70
 framework, 61-64
 government regulations concerning,
 60-61
 metamodel, 54-59
 oversight versus action roles, 60
 provenance and, 147
 role of, 175
data lineage, provenance and,
 149-151
data misalignment
 with business rules, 126-127
 default entries, 129
 mistaken entries, 126-129
data modeling example
 (governance), 49-51
data models, 122-126
 canonical data models, 185
 conceptual data models, 185
 enterprise data models, 186-187

logical data models, 180-186
 physical data models, 180-184, 188

data ownership, 129-130

data persistence, 71-72

data provenance, 6, 143-154
 abstraction and, 194
 accountability, 146
 data chains versus value chains,
 144-145
 data lineage and, 149-151
 decisions concerning, 150-152
 diplomatics, 146-147
 late binding versus early
 binding, 148
 statistics and, 7
 Zachman Framework for Enterprise
 Architecture, 152-154

data quality. *See also* quality
 measurement
 alignment and harmonization, 97-98
 data governance versus, 59
 matching, 98-100
 merging, 100
 survivorship, 100

data quality assessment (DQA),
 85-88, 206
 CIDER, 88-97

data structure design, effect on
 service-oriented architecture
 characteristics, 41-43

data synchronization, 171

Davison, Henry, 9

decayed data, 170-171

decision-making, 6

decisioning environment
 defined, 33*n*
 measurements, 33-37

decisions dial (data governance),
 66-69

deductive reasoning, 114

default values, 129

definitions
 for abstract concepts, providing,
 107-113
 qualities of, 108
 scope of, 109-111

delta move (ASPECT), 81

Descartes, René, 139

design patterns, abstract
 canonical data models, 185
 conceptual data models, 185
 context, considering, 188-194
 enterprise data models, 186-187
 logical data models, 180-186
 physical data models, 180-184, 188

deviation, governance and, 45-46

dialing data governance, 64-70

differentia in definitions, 108

diplomatics, provenance and,
 146-147

directives dial (data governance),
 66-68

directives in data governance, 62-63

director communication style, 54

discovery (CIDER), 89

Dixon, Bernard, 5

domain trait, evaluating, 133

**DQA³ (data quality [analysis]
 [accuracy][appraisal])**,
 85-88, 206
 CIDER, 88-97

Dupont, 16

E

early binding
 metadata for, 149
 provenance and, 148

Eastwood, Clint, 29

ED-SODA, 65-66, 206
 Electronic Data Systems Corporation, 21
 element coupling, 132
 element traits, 132
 cardinality, 133
 domain, 133
 format, 133
 precision, 132
 size, 132
 eminent domain, 22-23
 Empire Today, 21
 enforcement in data governance, 63
 engagement in SÉANCE workflow, 96
 enlightenment (CIDER), 89
 enrichment (moving data), 81
 ensure controls (data governance metamodel), 55-56
 enterprise data models, 186-187
 enterprise resource planning (ERP), 140*n*
 envoy communication style, 54
 ERP (enterprise resource planning), 140*n*
 essential attributes in definitions, documenting, 110-111
 ETL (extract, transform, and load), 82-83
 evidentiary controls (data governance metamodel), 56
 examination in SÉANCE workflow, 90-96
 experiences dial (data governance), 66-67
 expropriation, 22-23
 extensibility, 31
 data structure design and, 42
 external governance, 49-52

F

facilitator communication style, 53-54
 failure to identify business needs, reasons for, 29
 false negatives, 99
 false positives, 99
 FARMADÉ technique (governance communication), 52-54, 206
 Fastow, Andrew, 175
 fat fingering, 126-129
 Federal Reserve System, creation of, 8-10
 feedback loops, 34
 linear, 34
 negative, 35
 nonlinear, 35
 positive, 34
 viral data and, 35-37
 Feynman, Richard, 11
 field lengths in structural data design, 155-170
 final move (ASPECT), 81
 flat organizational structures, 17
 flexibility of viral data solutions, 177
 flight analogy (moving data), 82-88
 in-flight, 82-88, 97, 101, 147, 179, 200
 post flight, 82-88, 91, 97-98, 101, 147, 179, 200
 preflight, 82-88, 97, 98*n*, 101, 147, 179, 200
 FLWOR expressions, 198-199
 Foreign Corrupt Practices Act of 1977, 60
 forensic accounting, 174
 format trait, evaluating, 133

Forrester, Jay, 140
 framework for data governance,
 61-64
 Franks, Tommy, 10
 fruit in Adam and Eve story, 1
 functional dependencies, 121

G

General Motors, 21
 generalization, reasoning
 through, 116
 generations, 5
 Genesis, apple in, 1
 genus in definitions, 108
 Gibbs, Linda, xxii
 Glass-Steagall Act, 22
 Goldman Sachs, 22
 Google, 36
 Googlebot, 36
 governance
 business laws, 46-47
 communication traits, 52-54
 conflicts of interest, 49
 data governance, 48
 data, role of, 61
 data quality versus, 59
 dialing (adjusting), 64-70
 framework, 61-64
 government regulations concerning,
 60-61
 metamodel, 54-59
 oversight versus action roles, 60
 data modeling example, 49-51
 deviation and, 45-46
 external governance, 49-52
 intersituations, 52
 intrasituations, 52
 layers of, 46
 management governance, 48

management versus, 52
 punting, 51
 rules of behavior, 47
 rules of compliance, 47
 self-governance, 47-48
 government regulations, data
 governance and, 60-61
 Gramm-Leach-Bliley Act
 of 1999, 60
 granularity of provenance, 149-151
 Grotius, Hugo, 22
 guidelines in data governance, 63

H

Hadden, Briton, 11
 harmonization (of data), 87*n*, 97-98
 hasa (relationship term), 108
 Health Insurance Portability and
 Accountability Act (HIPAA), 60
Heartbreak Ridge (film), 29
 heuristic reasoning, 115
 HHS-Connect, xxii-xxiii
 hierarchical organizational
 structure, 17
 HIPAA (Health Insurance
 Portability and Accountability
 Act), 60
 historical revisionism, 103
 Hobson's choice, 30
 Hopkinson, Francis, 103
 Hoyle, Fred, 143
 Hughes Aircraft, 21

I

identifiers, 130
 imprecise reasoning, 115
 improvement (CIDER), 88
 incentive plans, 32-33

incorrect information
 Federal Reserve System example, 8-10
 latency and, 12
 PowerPoint example, 10-11
independence of viral data
 solutions, 177
indicator data, 131
induction, 114
in-flight, 82-88, 97, 101, 147, 179, 200
information loss as constraint, 28
information overload, 11-12
information technology
 adaptation to change, 20-43
best practices, 36-38
decisioning environment measurements, 33-37
eminent domain, 22-23
incentives and compensation, 32-33
metadata, 27-29
military, 29-30
quality measurement, 24-27
service-oriented architectures, 30-32, 38-43
viral data as symptom of, 24
 alignment with business, 15-20
CIA example, 16
multidivisional business structure, 16-20
viral data fostered by, 22
initial seeding (ASPECT), 81
innuendos, 3
inspections, 58
insure controls (data governance metamodel), 55-57
integration forensics, 174
interfaced in OSAPI, 158

interoperability, 7, 30-31
 data structure design and, 42
interrogatives, 152
intersituations of governance, 52
intrasituations of governance, 52
IPO, 190*n*
Iraq war, miscommunication during, 10-11
isa (relationship term), 108
ISO 4217 currency standard, 13
IT transactions, business transactions versus, 176*n*

J-K

Java Message Service (JMS), 140*n*
Jekyll Island, 9-10
Jett, Joseph, 32
Jingle Heimer-Schmidt, John Jacob, 120-121
JMS (Java Message Service), 140*n*
Joyce, James, 73-74, 76, 138
judgment in SÉANCE workflow, 96
Kidder Peabody, 32

L

language translations
 of Bible, 1
 in definitions, 112
Lao-Tze, 24
late binding
 metadata for, 149
 provenance and, 148
latency, 5
 misinformation and, 12
laws of business, 46-47
layering of abstraction, 31
 data structure design and, 42-43
layers of governance, 46

Legislative Bill 1, 111
 Legislative Bill 157, 110
 Lehman Brothers, 21, 33
 length of fields in structural data
 design, 155-170
 life events abstraction example,
 189-193
 life span
 of data, 76
 of services, 178
 Lincoln, Abraham, 106-107
 lineage, provenance and, 149-151
 linear feedback loops, 34
 Logan's Run, 160*n*
 logic, Wason test, 3
 logical data models, 49-52, 67, 121,
 182-188, 190, 192, 194
 physical data models and, 180-184
 logical inference, 114
 Loro Piana, 145
 loose coupling, 32
 in three-tier architectures, 38-43
 Luce, Henry, 11

M

Madoff, Bernard, xviii, 173
 major transformations, 84
 Malaysia, 119
 malum, 1
 management, governance versus, 52
 management governance, 48
 mandates in data governance, 63
 Marathon Oil, 21
 Mars Climate Orbiter, 175
 master data solutions, 205
 matching (data), 98-100
 matrix organizational structure, 17
 measurement controls (data
 governance metamodel), 57

measurement points in assure
 controls (data governance
 metamodel), 57
 measurements in decisioning
 environment, 33-37
 measuring quality, 24-27
 mediator communication style,
 53-54
 memory-resident databases, 72
 merging (data), 100
 message-service-message abstract
 metamodel, 178
 metadata, 27-29
 business metadata, 123
 for early/late binding, 149
 provenance and, 148
 structural metadata, 123
 structure of, 164-166
 technical metadata, 123
 types of, 73-76
 metamodels
 for abstraction, 178-181
 of data governance, 54-59
 metaphors, viral data as, 3-5
 military, adaptation to change,
 29-30
 minor transformations, 84
 "miracle on the Hudson," 191
 misalignment of data
 with business rules, 126-127
 default entries, 129
 mistaken entries, 126-129
 miscommunication, Iraq war
 example, 10-11
 misinformation
 Federal Reserve System example,
 8-10
 latency and, 12
 PowerPoint example, 10-11

misinterpretation of business data,
 Social Security Number
 example, 104-106

Mizuho Securities, 127-128, 175

modeling
 abstract concepts, 106-107
 data, 122-126
canonical data models, 185
conceptual data models, 185
enterprise data models, 186-187
logical data models, 180-186
physical data models, 180-184, 188

Morgan Stanley, 22

Morgan, J. P., 9

moving data
 ASPECT, 79-82
 bandwidth considerations, 72-73
 communication model, 77-79
 data access, 76-77
 flight analogy, 82-88
 metadata, 73-76

Mudd, Daniel, 174

multidivisional business structure,
 16-20

multiple views for data assessment,
 203-204

N

names example (specialization
 reasoning), 117-122

naming attributes, 182*n*

narrowing data for parametric
 reasoning, 116

natural identifiers, 130

NBC Universal, 6

Nebraska safe-haven law example
 (documenting essential
 attributes), 110

negative feedback loops, 35

negative terminology in
 definitions, 111

network bandwidth, moving data,
 72-73

New Testament translations, 1

Nikkei, 127

noise, 11*n*
 in communication model, 78

nonlinear feedback loops, 35

nonmonotonic reasoning, 116

Norton, Charles, 9

null values, 133*n*

O

observation in SÉANCE workflow,
 90-96

Old Testament translations, 1

Olympic Games, 6, 7

operating controls (data governance
 metamodel), 55-56

Operation Urgent Fury, 30

operational metadata, 149

organic nature of insure controls, 57

organigrams, 17

organizational structures,
 types of, 17

OSAPI, 157-160, 206

overloading, blind, 107-108

oversight, action roles versus, 60

oversight dial (data governance),
 66-68

ownership
 of data, 129-130
 in OSAPI, 158

P

pandemics, 4

parametric reasoning, 116

Paterson, David, 191*n*

- pattern-matching evaluations, 122
- perfect storm, *8n*
- performance in OSAPI, 158
- performing controls (data governance metamodel), 56-57
- persistence, *1n*
 - of data, 71-72
 - of viral data in communication model, 78
- persistence tier, 40
- physical data models, 188
 - logical data models and, 180-184
- pi, 132
- Piatt Andrew Jr., Abram, 9
- Picasso, Pablo, 106, 178
- plausible reasoning, 115
- political aspects of problem-solving,
 - technical aspects versus, 140-142
- political boundaries, 13
- Ponzi, Charles, 173
- Ponzi scheme, xviii, 173
- Porter, Michael, 144
- positive feedback loops, 34
- positive terminology in
 - definitions, 111
- post flight, 82-88, 91, 97-98, 101, 147, 179, 200
- PowerPoint in communication, 10-11
- precision trait, evaluating, 132
- preflight, 82-88, 97, *98n*, 101, 147, 179, 200
- presentation tier, 40
- preservation (moving data), 80
- proactive controls (data governance metamodel), 56
- proactive data governance, 63
- probabilistic comparison (of data), 98-100
- probabilistic reasoning, 115
- problem-solving
 - provenance in, 143-154
 - accountability*, 146
 - data chains versus value chains*, 144-145
 - data lineage and*, 149-151
 - decisions concerning*, 150-152
 - diplomats*, 146-147
 - late binding versus early binding*, 148
 - Zachman Framework for Enterprise Architecture*, 152-154
 - reactive traits for, 177-178
 - reductionism versus systems thinking, 139-140
 - technical versus political aspects, 140-142
- procedural controls (data governance metamodel), 57
- process tier, 41
- propagation of viral data, 206-208
- property value mistake example (latency), 12
- provenance, 6, 143-154
 - abstraction and, 194
 - accountability, 146
 - data chains versus value chains, 144-145
 - data lineage and, 149-151
 - decisions concerning, 150-152
 - diplomats, 146-147
 - late binding versus early binding, 148
 - statistics and, 7
 - Zachman Framework for Enterprise Architecture, 152-154

pseudo real time, 72*n*
 punting governance
 opportunities, 51

Q

qualitative values, reporting viral
 data, 142
 quality measurement, 24-27
 quality of data
 alignment and harmonization, 97-98
 data governance versus, 59
 matching, 98-100
 merging, 100
 survivorship, 100
 quantitative values, reporting viral
 data, 142
 quantity data, 131
 QuIT CITeD (data class taxonomy),
 130-132, 206

R

Radio frequency identification
 (RFID), 145
 Rallying Point, 152-153
 reaction to change, 20-43
 best practices, 36-38
 decisioning environment
 measurements, 33-37
 eminent domain, 22-23
 incentives and compensation, 32-33
 metadata, 27-29
 military, 29-30
 quality measurement, 24-27
 service-oriented architectures, 30-32,
 38-43
 viral data as symptom of, 24
 reactive controls (data governance
 metamodel), 56

reactive data governance, 62-63
 reactive traits (for problem-solving),
 177-178
 reasoning, types of, 113-114
 case-based reasoning, 115
 classification, reasoning through,
 116-117
 consistency-based reasoning, 115
 generalization, reasoning
 through, 116
 heuristic reasoning, 115
 imprecise reasoning, 115
 logical inference, 114
 nonmonotonic reasoning, 116
 parametric reasoning, 116
 probabilistic reasoning, 115
 semantic reasoning, 116
 specialization, reasoning through,
 117-122
 reassure controls (data governance
 metamodel), 55, 58-59
 reductionism
 defined, 139
 systems thinking versus, 139-140
 redundancy, data assessment
 via, 204
 refactoring, 123*n*
 reference model, 100-101
 data persistence, 71-72
 data quality
 alignment and harmonization, 97-98
 matching, 98-100
 merging, 100
 survivorship, 100
 data quality assessment (DQA),
 CIDER, 88-97

moving data
ASPECT, 79-82
bandwidth considerations, 72-73
communication model, 77-79
data access, 76-77
flight analogy, 82-88
metadata, 73-76
 regulations, data governance and,
 60-61
 relationship terms in
 definitions, 108
 remediation (CIDER), 89
 reporting viral data, quantitative
 versus qualitative values, 142
 representative communication
 style, 53
 reusability, 31
 data structure design and, 42
 revenge theory, 5
 revisionism
 historical revisionism, 103
 misinterpretation of business data,
 104-106
**RFID (Radio frequency
 identification)**, 145
 rich metadata, 76
 risk management, data governance
 and, 66
 rogue applications, viral data
 from, 24
 root-cause analysis, 105
 Ross, Betsy, 103
 rules of behavior, 47
 rules of compliance, 47

S

Saarinen, Eliel, 188
 Safeguarding Customer Information
 rule, 60

Safeway Stores, 21
 sanctions dial (data governance),
 66-68
 Sarbanes-Oxley Act of 2002,
 60, 175*n*
 scenarios for data assessments, 138
 Schwartz, Alan, 174
 scope
 of definitions, 109-111
 in SÉANCE workflow, 89
 Securities Exchange Act of 1934, 60
 self-governance, 47-48
 semantic disintegrity, 202
 semantic evaluations, 122
 semantic reasoning, 116
 semantics, communication and,
 1-13
 Adam and Eve story, 1
 analogies, 2-3
 books, definition of, 11-12
 controlling viral data, 8
 Federal Reserve System example,
 8-10
 information overload, 11-12
 ISO 4217 currency standard
 example, 13
 political boundaries, 13
 PowerPoint example, 10-11
 property value mistake example, 12
 sports statistics, 6-7
 viral data as metaphor, 3-5
 semistructured data, 71, 73-76, 79,
 93, 101, 131-132, 206
 Seinfeld, Jerry, 5
 Senge, Peter, 140
 service-oriented architectures
 adaptation to change, 30-32
 interoperability, 7

- loose coupling in three-tier architectures, 38-43
- technical characteristics of, 30-32
- SÉANCE workflow, 89-97, 206
- shades of governance, 48
- Shannon, Claude, 77
- Shannon-Weaver communication model, 77, 206
- Shinseki, Eric, 24
- Siekaczek, Reinhard, 173
- silo-oriented organizational structures, 17-20
- silos, 203
- simple metadata, 74
- size trait, evaluating, 132
- Smith, Henry, 106
- SOA, propagation of viral data, 206-208
- Social Security Administration (SSA), 104
- Social Security Number example (misinterpretation of business data), 104-106
- software abstraction, 178*n*
- solutions
 - provenance in, 143-154
 - accountability*, 146
 - data chains versus value chains*, 144-145
 - data lineage and*, 149-151
 - decisions concerning*, 150-152
 - diplomats*, 146-147
 - late binding versus early binding*, 148
 - Zachman Framework for Enterprise Architecture*, 152-154
 - reactive traits for, 177-178
 - reductionism versus systems thinking, 139-140
 - technical versus political aspects, 140-142
- South Florida Sun-Sentinel* (newspaper), 36
- specialization, reasoning through, 117-122
- sports statistics, 6-7
- SSA (Social Security Administration), 104
- stability in OSAPI, 158
- staging areas, 95
- standardization, 57
- standards in data governance, 63
- Stars and Stripes, 103
- statistical sampling for data assessments, 138
- statistics
 - provenance and, 7
 - in sports, 6-7
- Streisand, Barbra, 175
- Strong, Benjamin, 9
- structural data design, data conditioning in, 155-170
- structural metadata, 123
- structure of metadata, 164-166
- structured data, 71, 73-74, 76, 79, 101, 131-132, 155-156, 206
- Sullenberger, Chelsey, 191
- sunsetting, 178*n*
- Sunstein, Cass, 2
- surrogate values, 130
- survivorship (data), 100
- sustaining controls (data governance metamodel), 57
- Swansea, 112
- symbiosis (moving data), 80
- synchronization of data, 171

systems thinking, 206

defined, 140

reductionism versus, 139-140

T

taxonomy of data classes, 130-132

technical aspects of problem-solving, political aspects versus, 140-142

technical metadata, 123, 149

temporal data, 131

Tenner, Edward, 5

Terra Gruppen, 13

TETLT (transform, extract, transform, load, and transform), 82-84

textual data classes, 131

thing modeling, 43*n*

third-normal form (3NF), 121

three-tier architectures, loose coupling in, 38-43

tight coupling, problems with, 39-43

time in data governance, 64

trans-enterprise, 145*n*

transformations in TETLT patterns, 83-84

transition (moving data), 81

translations in definitions, 112

Truman, Harry S, 118*n*

truth tables, 116

trustworthiness of data

in context, 134-138, 201-203

master data solutions, 205

multiple views for, 203-204

via redundancy, 204

scenarios for, 138

statistical sampling, 138

subjective and objective measures, 206

Tufte, Edward, 10

U

U.S. Department of Homeland

Security, hierarchical

organizational structure, 17-20

unbounded codified data, 131

uncertainty, 99

United Airlines, 36

United States Steel, 21

unstructured data, 28, 36, 71, 73, 76, 79, 101, 132, 206

V

validity, 167*n*

value chains, 203

data chains versus, 144-145

Vanderlip, Frank, 9

vehicle identification number (VIN)

example (structural data design), 158-169

views

abstraction and, 194-201

multiple views for data assessment, 203-204

VIN example (structural data design), 158-169

violet assessment example, xxxv-xxxvii, 209

viral data

in business and information

technology alignment, 22

controlling, 8

defined, xxxiii

feedback loops and, 35-37

forms of, 175-176
host for, 176
as metaphor, 3-5
perfect storm for, 8
persistence in communication
 model, 78
propagation of, 206-208
quality measurement and, 24-27
reporting, quantitative versus
 qualitative values, 142
from rogue applications, 24
solutions, list of, 205
symptom of adaptation to change, 24
virus, origin of term, 4

W-Z

walkie-talkie, 128-129
Walsh, Mark, 33
Warburg, Paul, 9
Washington, George, 103
Wason test, 3
Wason, Peter, 3
Weaver, Warren, 77
Welsh, language, 112
Western Reminiscences (Smith), 106
Whitaker, Edmund, 45
Wilson, Woodrow, 9
WMI codes, 164
World Health Organization, 4
Wycliffe, John, 1
XM Radio, 21
Zachman Framework for Enterprise
 Architecture, 152-154