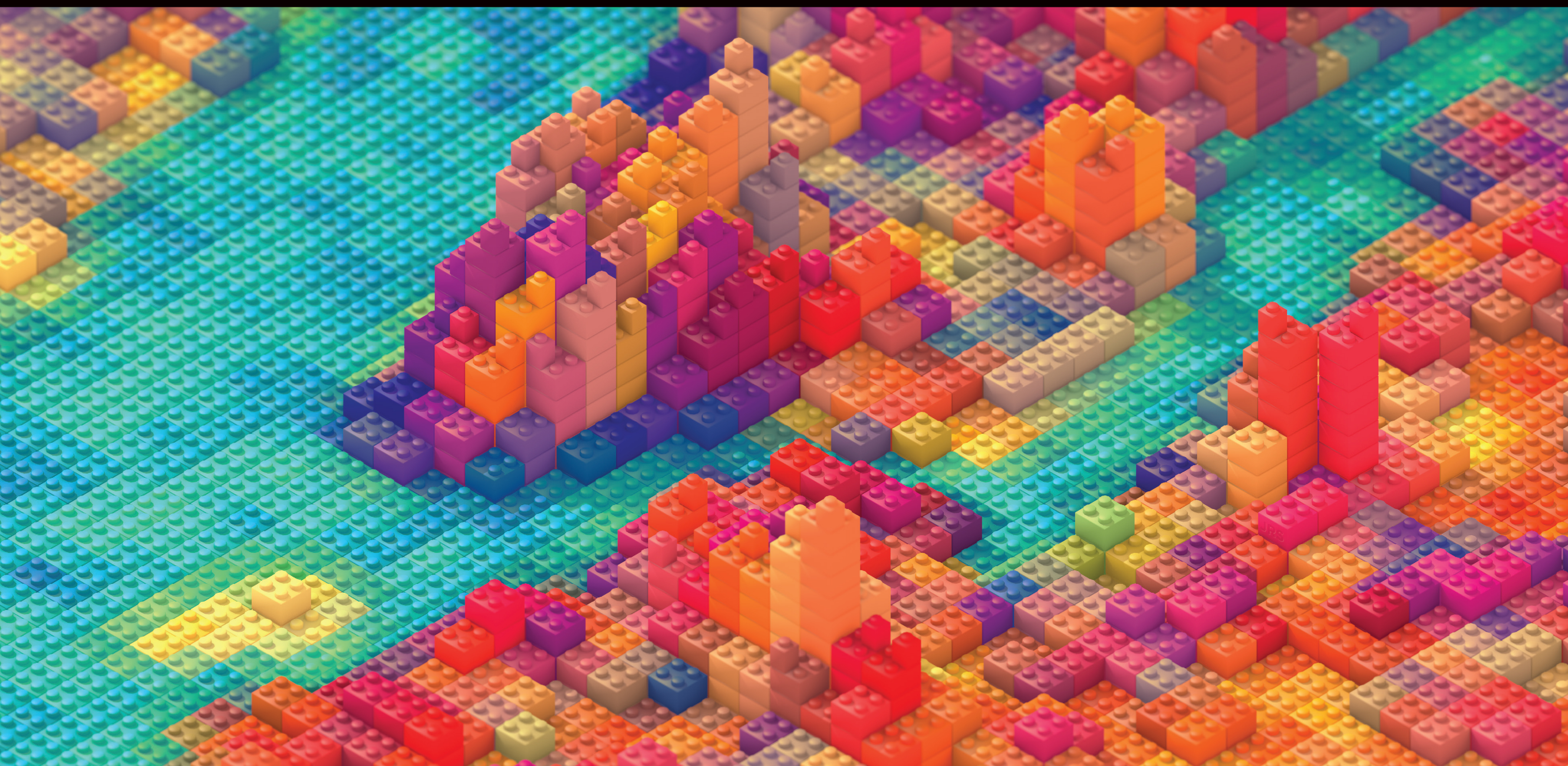
Atlas of Knowledge

Anyone Can Map

Katy Börner



For information about special quantity discounts, please e-mail special_sales@mitpress.mit.edu

This book was set in Adobe Caslon Pro by Tracey Theriault (graphic design and layout) and Katy Börner (concept), Cyberinfrastructure for Network Science Center, School of Informatics and Computing, Indiana University. Printed and bound in Malaysia.

Library of Congress Cataloging-in-Publication Data

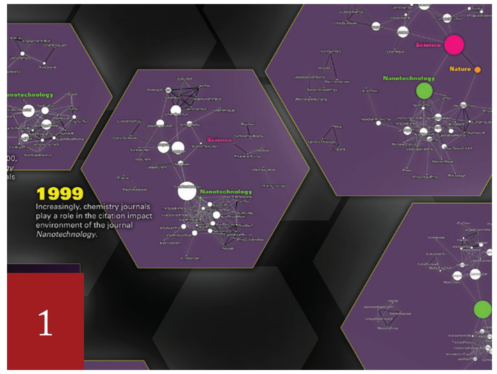
Börner, Katy.
Atlas of knowledge : anyone can map / Katy Börner.
pages cm
One of a series of three publications influenced by the travelling exhibit Places & Spaces: Mapping Science, curated by the Cyberinfrastructure for Network Science Center at Indiana University.
Includes bibliographical references and indexes.
ISBN 978-0-262-02881-3 (hardcover : alk. paper)
1. Information visualization. 2. Science—Atlases. 3. Statistics—Graphic methods. 4. Science—Study and teaching—Graphic methods. 5. Communication in science—Data processing. 6. Technical illustration. 7. Graph design. I. Title.
QA90.B6624 2015
501'.154—dc23

2014028219

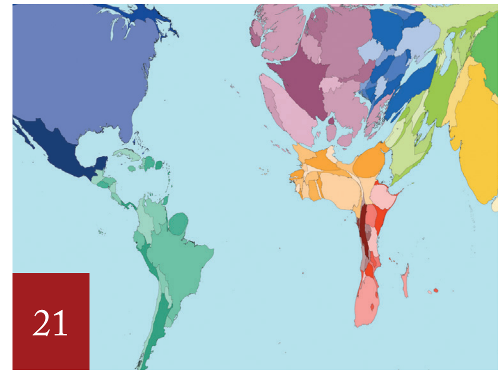
10 9 8 7 6 5 4 3 2 1

I dedicate the *Atlas of Knowledge* to my husband, Robert L. Goldstone.

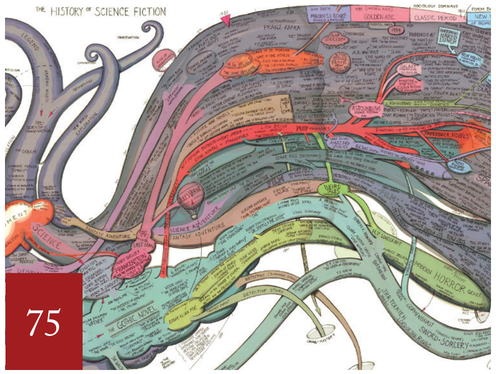
Contents



Part 1: Science and Technology Facts

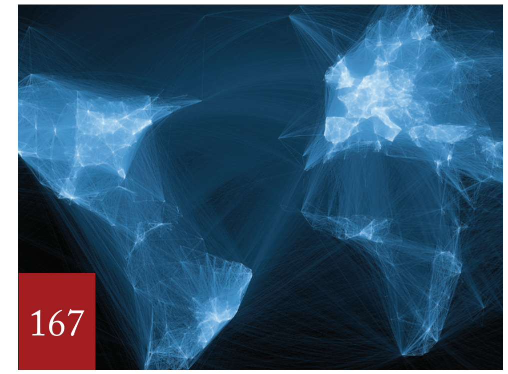


Part 2: Envisioning Science and Technology



75

Part 3: Science Maps in Action



167

Part 4: Outlook

Foreword

You could say it was Marco Polo who started it all when he returned from China and reported the distance he'd travelled east from Europe as a lot farther than it really was. So when the Italian hotshot mathematician Paolo Toscanelli used Polo's data to finalize a new map of the world and then Columbus got hold of a copy, the distance to China going the other way (west, straight across an empty ocean) looked quick and easy. Then, oops, America!

With the discovery of a new continent, there went the neighborhood. The definitive map of the world at the time was that crafted by Aristotle, who hadn't included America. What was the place doing there? And what about all the amazing stuff that began to pour in from the newfound world: new species, new minerals, new races, none of which were in Aristotle either.

In 1533, Dutch mathematician Gemma Frisius complicated matters with his idea for fixing a location by triangulation, thus making it easier for explorers to sail off into the blue; now at any point en route explorers could use the position of the last headland and the position of the next one to pinpoint where they were. Headland by headland, the more they advanced, turning the unknown into the known, the more unknown there was to explore. Discovery bred discovery, which left the other problem: What to do about their returning cargoes—that new stuff Aristotle hadn't mentioned—all of which seriously upset the comfortable medieval view of the world and everything in it.

Panic set in. If Aristotle could be that wrong, then which way was up? As contemporary worrier John Donne put it: "The new philosophy (aka the new discoveries) calls all in doubt." In the growing intellectual confusion, the search was on to generate data one could trust.

So thank you, René Descartes. In 1637, his methodical doubt and reductionism (double-check everything, down to the smallest detail) took the risk out of risk, and the West threw itself into intellectual and geographical exploration with all the abandon of an alcoholic in a brewery. The new mantra was "find useful knowledge." Armed with the sword of reductionism and protected by the shield of method, we boldly took scientific thinking where no minds had gone before. The aim: to learn more and more about less and less.

Faster than you could say "epistemology," the knowledge disciplines proliferated, generating niche studies (let's hear it for the PhD!) that in turn became disciplines generating their own niche studies. Silo-thinking was here to stay. And (to mix metaphors), inside every intellectual silo, blinkered specialists worked away, blissfully unaware of what might be going on in other silos.

Then the fun began. As products and ideas began to emerge from specialist silos, they would bump into each other with results that were more than the sum of the parts. One and one began to make three. Maybach brought together the perfume spray with gasoline and invented the carburetor. Electricity and magnetism made possible the telegraph. The discovery of the bacillus plus the invention of aniline dye added up to chemotherapy. As I have shown in my own work, innovation comes when ideas are linked in new ways. On the great web of knowledge, ultimately everything is linked to everything else. Innovation is the rule, not the exception.

As the specialists multiplied and communications technology made it easier for them to interact, the pace of innovation quickened, with unexpected results. Ripple effects could be unpredictable: The typewriter took women out of the kitchen into the office and boosted the divorce rate, refrigerators chilled food and punched a hole in the ozone layer, and X-rays bouncing off coal-crystal structures triggered the genetics industry. The sciences began to take on double, bump-together names: neurophysiology, molecular biology, astrophysics, and more. Gobbledygook was here to stay.

Then came the Internet, and suddenly it was Columbus and Frisius all over again. Today, we find ourselves in a vast, chaotic, interactive, constantly innovative, exponentially expanding world of data in which change is happening so fast that without the means to triangulate from one set of data to another, to see how the data relate, and what kind of innovation they may trigger we don't know where we are, where we're going, and, especially, what we're likely to find when we get there.

Accurate prediction is now more essential than ever, given above all the unimaginable potential social consequences of developments in different science and technology fields. Take, for example, nanotechnology: We have perhaps fifty years before the first nanofabricator, powered by photovoltaics, is able to manipulate material at the atomic level to create molecules and then turn those molecules into stuff and use that stuff to manufacture gold, food, bricks, water, and so on from primarily dirt, water, and air, making almost anything, almost free.

The first thing the first fabricator might do is make a copy of itself: one for everyone on the planet in a matter of months. Then live wherever your fancy takes you, entirely self-sufficient, with the means electronically to transmit yourself across the world as a three-dimensional hologram, a world not of 196 nations but of nine billion autonomous individuals with the freedom to do, and be, whatever they choose.

Chaos may follow. The free provision of every material need and behavior unfettered by community constraint may call into question every social institution from government to belief systems to the cultural values that unite us to the entire market economy.

Since leaving the caves, we have focused our full attention on dealing with scarcity. The finely honed skills we have developed in order to handle that millennial problem have left us totally unprepared for the radical abundance that lies down the road.

The journey from here to there is fraught with difficulties and perhaps even danger. We need to be able to identify when required that (as they would have said in medieval cartography) "Here there be dragons." We need maps to guide us, to show us where *not* to go, what innovations and new ideas *not* to espouse, to reveal the unknown unknowns so as to enable us to predict the outcome of our choices along the way.

This extraordinary *Atlas* is the first step on that road.

James Burke
Science historian, author, and television producer
London, United Kingdom

Preface

The *Atlas of Knowledge: Anyone Can Map* was written with the deep belief that just as "anyone can cook," it is also true that "anyone can map"—or at least learn to do either. The *Atlas* series is being written at a time when data literacy is becoming almost as important as language literacy. While the first of the series, *Atlas of Science: Visualizing What We Know*, provided a gentle introduction to the power of maps for the navigation, management, and utilization of knowledge spaces, the *Atlas of Knowledge* intends to empower anyone to map and make sense of science and technology (S&T) data to improve daily decision making.

Part 1 argues for a systems science approach in the study of S&T structure and dynamics. Drawing on research and teaching in data mining, information visualization, and science of science studies, it explains and exemplifies different levels and types of analysis and also reviews key facts at different levels of the S&T system.

Part 2 introduces a theoretical framework meant to guide readers through user and task analysis; data preparation, analysis, and visualization; visualization deployment; and the interpretation of S&T maps. It benefits from more than 10 years of tool development and feedback from many of the more than 150,000 tool users in academia, industry, and government.

Just like the *Atlas of Science*, this book accompanies the *Places & Spaces: Mapping Science* exhibit (**http://scimaps.org**). **Part 3** features maps from the fourth to the seventh iterations, designed for economic decision makers, science policy makers, and scholars as well as librarians and library users. The 40 large-scale, full-page maps are meant to exemplify data analysis workflows and visualization metaphors and to communicate key insights. The final 30 maps of this 10-year exhibit effort, comprising the eighth to the tenth iterations, will be included in the third volume of this series, the *Atlas of Forecasts: Predicting and Broadcasting Science, Technology, and Innovation*.

Part 4 examines S&T trends and discusses the possible impact of real-time data visualizations on practicing and steering S&T. It concludes with an outlook of expected developments that focus strongly on democratizing knowledge and participation as well as promoting the evolution of standards—in terminology, data sets, data mining and visualization algorithms, workflows, and interface design—toward higher replicability and utility.

To ease navigation and consumption, each major topic is presented solely on one double-page spread. References to other parts of the book interlink the different topics and sections, resulting in a whole that extends beyond the sum of its parts. The decision was made to compile the extensive number of references in the back matter of the *Atlas*, including more than 1,500 references, 350 image credits, 30 data credits, and 20 software credits on a page-by-page basis.

Although textbooks such as Nathan Yau's *Visualize This* or the IVMOOC book entitled *Visual Insights: A Practical Guide to Making Sense of Data* teach timely knowledge about tools and workflows, this *Atlas* series aims to present "timeless knowledge" that may still hold true many years from now—akin to Edward R. Tufte's notion of "forever knowledge" that involves information design principles that are indifferent to culture, gender, nationality, or history.

Analysis and visualization design require the many varied skills involved in data management, data analysis, design, communication, and technology. Depending on your background and expertise, different reading trajectories are proposed:

- If you are familiar with the science of science studies but not as well versed in science mapping, begin by perusing the maps in **Part 3**, then follow up by reading the **Part 2** text on how to design insightful visualizations.
- If you are a visualization expert interested in design principles and guides, go directly to **Part 2**.
- If you are a designer but not familiar with science visualizations, read **Part 1** and explore the maps in **Part 3** before consuming other parts.
- If you are a programmer interested in building tools for avid users, start by reading **Part 2**—which explains how to systematically render data into insights using algorithms and approaches from statistics, cartography, linguistics, network theory, and other areas of science. Then move on to **Parts 1** and **4** to learn about current and future user needs and applications.
- If you wish only to see the future of S&T mapping, go directly to **Part 4**.

Additional materials can be found at **http://scimaps.org/atlas2**, including high-resolution images that are available for closer examination; digital files of the more than 1,000 citations and source credits; access to data sets and tutorials on how to run specific workflows; and updates of essential materials in preparation for future editions.

I feel lucky to have had the luxury of being able to develop this *Atlas*—an attempt to organize and make accessible to many research on the analysis and visualization of S&T structure and dynamics. It is my hope that the knowledge and techniques presented in these books will not only live between the covers, online, or in the mind of each reader, but also will be applied to further our understanding and to improve both our personal and collective decision making.

Katy Börner
Cyberinfrastructure for Network Science Center
School of Informatics and Computing
Indiana University

August 11, 2014

Acknowledgments

It may seem unwise to devote a major part of one's research time to writing a series of books for readers who are unlikely to write papers or otherwise cite these books in academic circles. And yet it seems quite on target to enable those who finance science via tax dollars to benefit from the research results—forfeiting the maximization of citation counts via the production of research papers. Many others have taken this route, including the following luminaries who have inspired my own journey: Jacques-Yves Cousteau, the French explorer and researcher of the sea; David Attenborough, especially with his *Life on Earth* and *Living Planet* series; Paul Otlet, with his *Universal Atlas* or *Encyclopedia Universalis Mundaneum*; Stuart Brand, author of *The Whole World Catalog*; Richard Dawkins, famed for his "Growing Up in the Universe" lectures; Al Gore for his environmental efforts, as featured in the *An Inconvenient Truth* documentary; and Hans Rosling, whose Gapminder effort gave rise to the motto, "Let my dataset change your mindset."

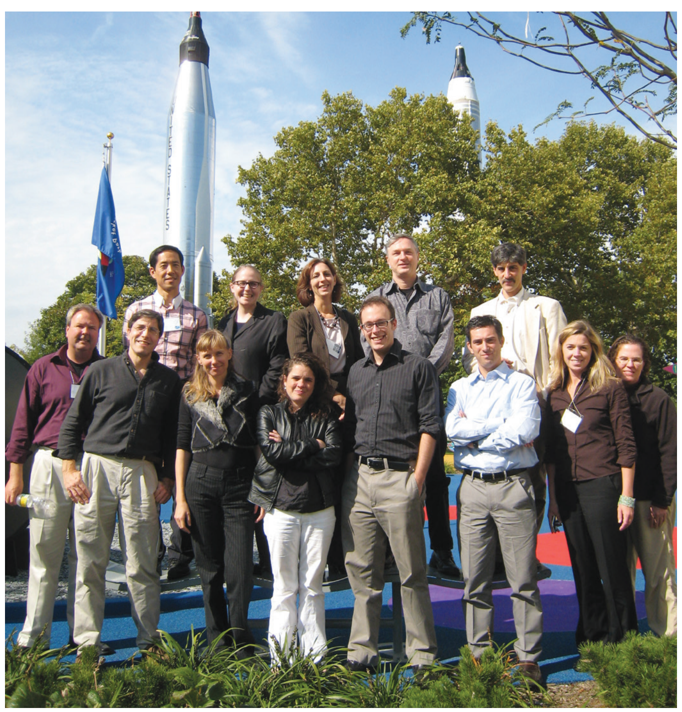
It is my hope that this *Atlas* series joins in giving both inspiration and encouragement to future science communicators.

I am deeply grateful to all those who helped to make possible this *Atlas* and the exhibit maps it features.

Part 2, Envisioning Science and Technology, benefited deeply from my teaching of relevant courses at Indiana University over the last 14 years, including teaching the Information Visualization MOOC (IVMOOC) to students from more than 100 countries in the spring of 2013.

The *Places & Spaces: Mapping Science* exhibit would not have been possible without the expertise and professional excellence of the more than 236 mapmakers and the 43 exhibit ambassadors around the globe. Exhibit advisers for the maps featured in this book include: Deborah MacPherson (Accuracy&Aesthetics), Kevin W. Boyack (SciTech Strategies, Inc.), Sara Irina Fabrikant (Geography Department, University of Zürich, Switzerland), Peter A. Hook (Law Librarian, Indiana University), André Skupin (Geography, San Diego State University), Bonnie DeVarco (BorderLink), and Dawn Wright (Geography and Oceanography, Oregon State University). External experts that reviewed iterations 4 through 7 included: John R. Hébert (Chief of the Geography and Map Division, Library of Congress), Thomas B. Hickey (OCLC), Michael Kurtz (Harvard-Smithsonian Center for Astrophysics), Denise A. Bedford (World Bank), William Ying (CIO ArtSTOR), Michael Krot (JSTOR), Carl Lagoze (Cornell University), Richard Furuta (Texas A&M University), Vincent Larivière (Université du Québec à Montréal, Canada), Adam Bly (CEO of SEED), Alex Wright (author of *Glut: Mastering Information Through The Ages*), and Mills Davis (Project10x.com).

Focused brainstorming workshops, organized with colleagues between 2008 and 2014, contributed greatly to the discussion of research and development work that is contained in these pages. A total of 25 such workshops were held on a range of topics, including "How to Measure, Map, and



October 1-2, 2009: NSF/JSMF Workshop on How to Measure, Map, and Dramatize Science, New York Hall of Science, NY

March 4-5, 2010: NSF/JSMF Workshop on Mapping of Science and Semantic Web, Indiana University, Bloomington, Indiana

October 9-10, 2010: Modeling Knowledge Dynamics, The Virtual Knowledge Studio, Amsterdam, The Netherlands

Dramatize Science," "Mapping the History and Philosophy of Science," "Modeling Knowledge Dynamics," "Artists Envision Science & Technology," and "Plug-and-Play Macroscopes" (see group photos).

A substantial part of the source review and initial writing was completed while I was a visiting professor at the Royal Netherlands Academy of Arts and Sciences (KNAW) in the spring of 2012. I would like to thank Paul Wouters of CWTS and Andrea Scharnhorst and Peter Doorn of DANS for their support.

Financial support came from the National Science Foundation under Grants No. DRL-1223698, OCI-0940824, SBE-0738111, and CBET-0831636; the National Institutes of Health under Grants No. U01-GM098959, R21-DA024259, and U24-RR029822; the James S. McDonnell Foundation; the Bill & Melinda Gates Foundation; Indiana's 21st Century Fund; Thomson Reuters; Elsevier; the Cyberinfrastructure for Network Science Center, University Information Technology Services, and the former School of Library and Information Science—all three at Indiana University. Some of the data used to generate the science maps is from the Web of Science by Thomson Reuters and Scopus by Elsevier. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

Copyediting of the *Atlas* was performed by Gordana Jelisijevic, Melinda Rankin, and Todd N. Theriault; *Atlas* layout and design by Tracey Theriault, with many of the images specifically created for this book by Perla Mateo-Lujan; reference checks and formatting by Todd N. Theriault; indexing by Amy Murphy; and copyright acquisition by Samantha Hale, Brianna Marshall, Joseph Shankweiler, David K. Kloster, and Michael P. Ginda.

Yong-Yeol Ahn, Kevin W. Boyack, Alberto Cairo, David Chavalarias, Joseph Cottam, Blaise Cronin, Vincent Delvaux, Scott Emmons, Yves Gingras, Daniel A. K. M. Halsey, Andrew J. Hanson, Peter A. Hook, Ketan K. Mane, Staša Milojević, Abel L. Packer, Roberto de Pinho, Bahador Saket, Ben Shneiderman, André Skupin, and Stephen M. Uzzo reviewed a penultimate draft of the book and their expert comments were instrumental in finalizing the *Atlas*. Other valued contributions are acknowledged in the **References & Credits (page 178)**.

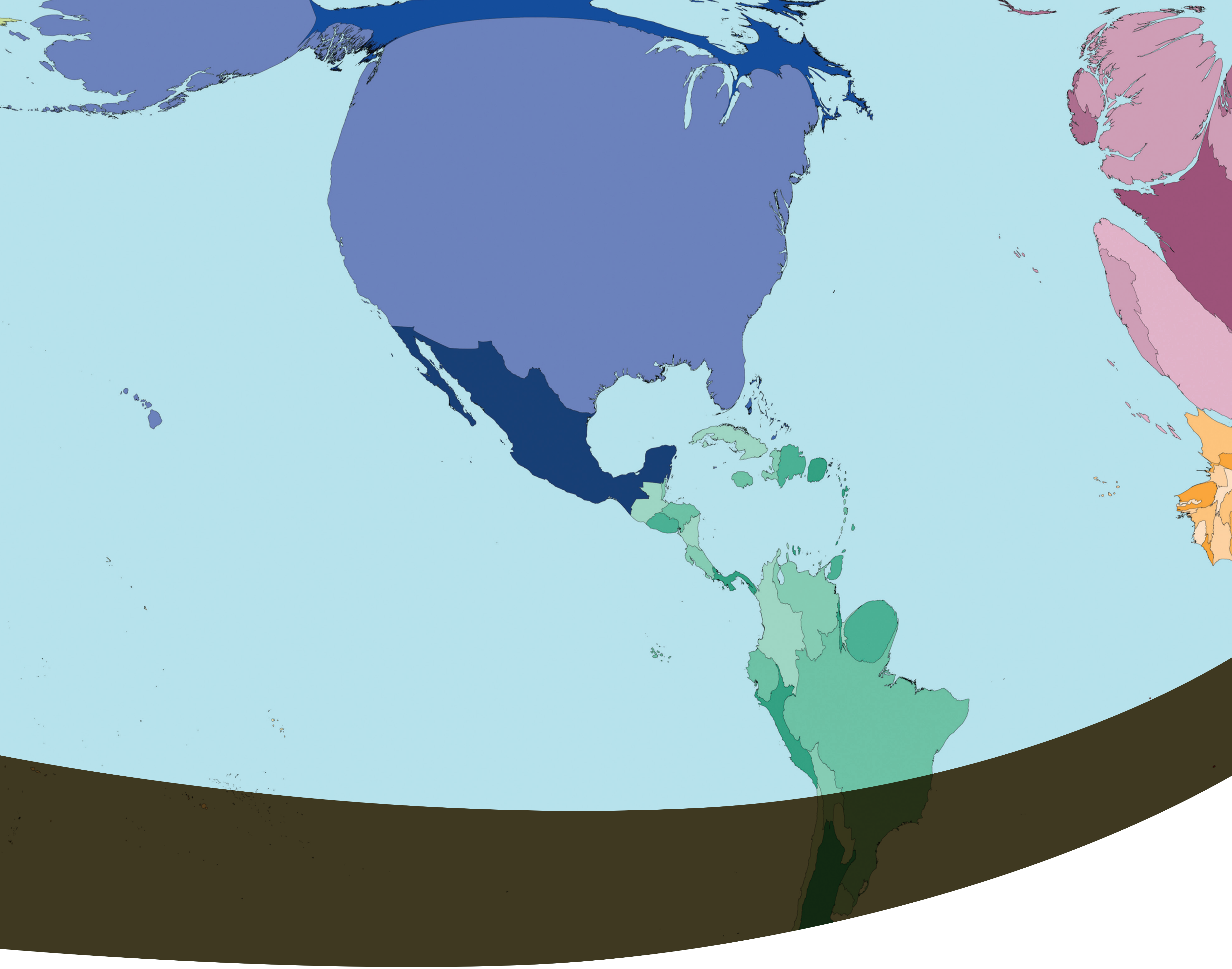
My sincere thanks go to Marguerite B. Avery, Katie Persons, and Katie Helke at MIT Press who ingeniously mastered the many complexities involved in publishing this *Atlas* series.

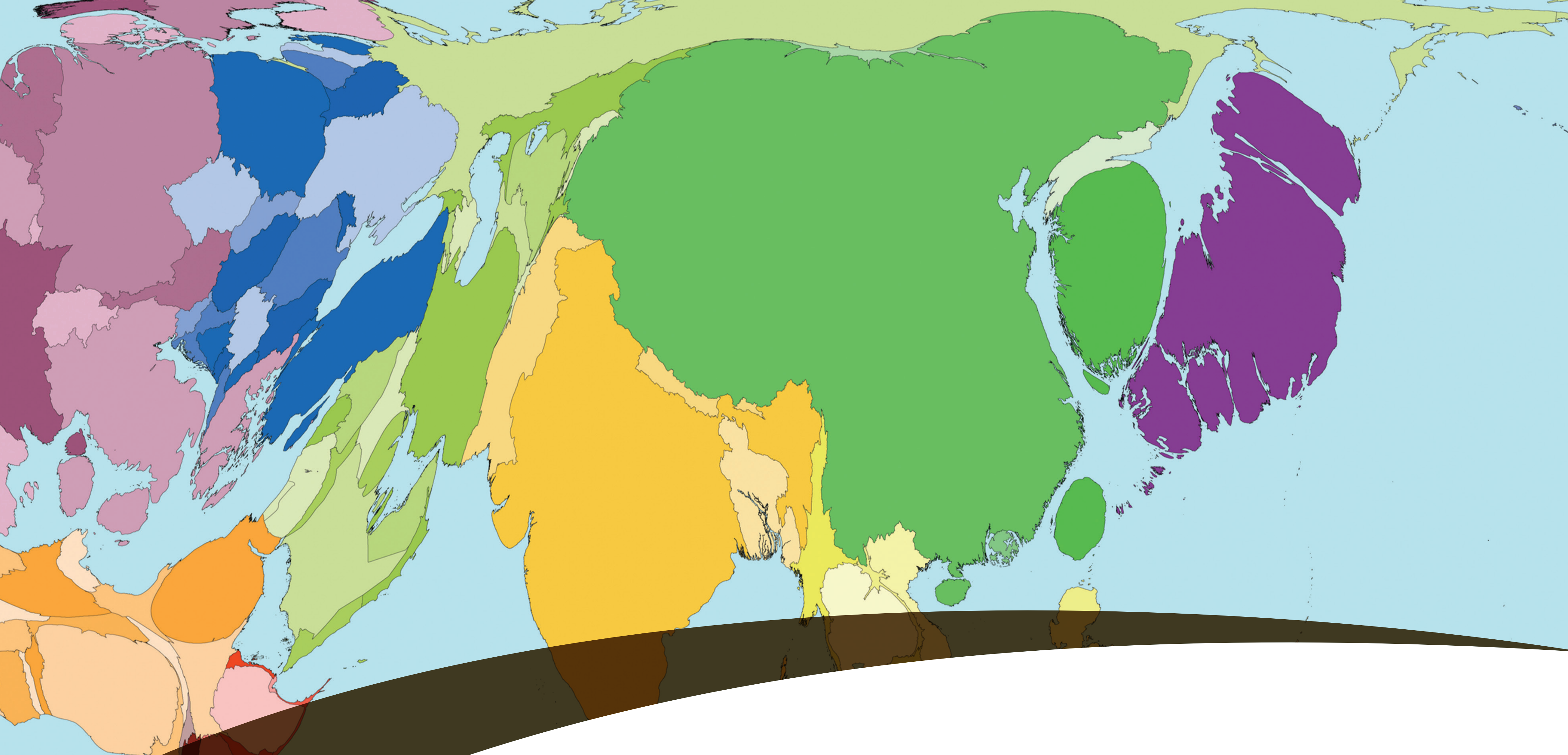
I am indebted to family and friends for providing much inspiration, energy, and loving support. This book benefited deeply from nurturing and thought-provoking family dinner discussions and empowering girls' nights out. My gratitude also rests with our cat, Jiji, who kept me company through the many long periods of writing.

August 11-12, 2011: JSMF Workshop on Standards for Science Metrics, Classifications, and Mapping, Indiana University, Bloomington, Indiana

March 25-26, 2013: Exploiting Big Data Semantics for Translational Medicine, Indiana University, Bloomington, Indiana

May 5, 2014: Researchers and Staff at the Cyberinfrastructure for Network Science Center, Indiana University, Bloomington, Indiana





Part 2: Envisioning Science and Technology

We thrive in information-thick worlds because of our marvelous and everyday capacity to select, edit, single out, structure, highlight, group, pair, merge, harmonize, synthesize, focus, organize, condense, reduce, boil down, choose, categorize, catalog, classify, list, abstract, scan, look into, idealize, isolate, discriminate, distinguish, screen, pigeonhole, pick over, sort, integrate, blend, inspect, filter, lump, skip, smooth, chunk, average, approximate, cluster, aggregate, outline, summarize, itemize, review, dip into, flip through, browse, glance into, leaf through, skim, refine, enumerate, glean, synopsize, winnow the wheat from the chaff and separate the sheep from the goats.

Edward R. Tufte

Motivation

Framework

Acquire

Analyze & Visualize

Deploy

Interpret

Foundations and Aspirations

Part 2 of this book introduces general data analysis and visualization techniques commonly used to study science and technology (S&T). Data analysis is an iterative process that cleans, filters, interlinks, mines, and augments data. Data visualization corresponds to an optimization of many different design decisions that relate not only to the layout and visual encoding of data but also to the interactivity and deployment of visualization. In this spread, foundations and aspirations for this *Atlas* are discussed, and the importance of empowering anyone to read and make visualizations is explained.

Maps, like speeches and paintings, are authored collections of information and also are subject to distortions arising from ignorance, greed, ideological blindness, or malice.
Mark Monmonier

Foundations

The structure and content of this part was inspired by scholarly works written over the last 250 years. Among them are William Playfair's *The Commercial and Political Atlas*; Jacques Bertin's *Semiology of Graphics*; John Tukey's practical epistemology; William Cleveland's combination of statistical and experimental evidence; Howard Wainer's work on history, statistics, and graphics; Edward Tufte's many examples of good design in *Beautiful Evidence*; Leland Wilkinson's codification of the structure of graphics in *The Grammar of Graphics*; and additional works from psychology, cartography, statistics, and other sciences that use data analysis and visualization, graphic design, and illustration to support decision making.

The process of creating insightful visualizations calls for the synergism of several disciplines: technology, to ensure that certain analyses can be run and designs produced; science, to provide correct and rigorous results; and art and design, to deliver aesthetically pleasing results that will attract and retain the attention of viewers so they may engage and gain valuable insights from those visualizations.

Setting Up Successful Projects

The design of insightful visualizations requires access to three essential ingredients: expertise, data, and resources. Expertise is traditionally provided by domain experts or clients that have specific insight needs (see **page 40**, **User Needs Acquisition**), are available to help with identifying and gaining access to relevant data sources (see **page 42**, **Data Acquisition**), and can interpret and evaluate results (see **page 72**, **Validation and Interpretation**). High quality and coverage of data is important. If faulty or incomplete data are used, visualizations, in turn, will also be faulty or incomplete. The problem of "garbage in, garbage out" could potentially escalate, as professionally rendered visualizations of incomplete or false data can easily lead to inappropriate decisions or the transmission of unverified information. Finally, resources include time and monetary investment or access to tools when performing the planned work. If any of these ingredients is not available, the visualization project is likely to fail.

Embracing the Power

Visualizations give form to either visible or invisible entities, making them tangible, understandable, and actionable. By thoughtfully representing high-quality, comprehensive data in an easy-to-read format, insightful renderings can change our view of the world. An example is Charles Darwin's 1837 *Tree of Life* drawing (see opposite page, top-left), which shows how species are purportedly related through evolutionary history and thereby reveals what may be life's common ancestry.

Visualizations have been instrumental in saving people's lives. One case in point is John Snow's *Cholera Map* of 1854 (see opposite page, lower-left), regarded as a key factor in the founding of the science of epidemiology. In the map, bars represent deaths caused by the 1854 London cholera epidemic. By showing them clustered around the water pump on Broad Street, the map enabled the recognition of cholera as a waterborne disease. Subsequent removal of the pump's handle led to the decreased incidence of cholera.

Another example is the "coxcomb" or polar-area diagram, first developed by Florence Nightingale. Her 1858 graphic on the *Causes of Mortality in the British Military during the Crimean War* (see opposite page, top-right) was critical in documenting that most soldiers had died of preventable or mitigable infectious diseases (blue) rather than of wounds sustained in battle (red) or other causes (black). The diagram presented vital statistical data in a way that persuaded Queen Victoria and others of the need to improve sanitary conditions in military hospitals, which substantially helped reduce death rates, profoundly influencing the subsequent course of the British military medical system.

David McCandless's *The Antibiotic Abacus: Adding up Drug Resistance* (opposite page, lower-right) uses data from the Centers for Disease Control and Prevention and the World Health Organization to communicate the increasing resistance of bacteria to antibiotics. Bacteria names are listed vertically on the left. Antibiotics and antibiotic families are plotted horizontally by date of introduction. Circles indicate the resistance of bacteria to different antibiotics (pink) and antibiotic families (purple): the larger the circle size, the higher the resistance. Note that many bacteria are "superbugs" that are resistant to multiple antibiotics. No major new antibiotics have been developed for the last 20 years—indicating a potentially fatal drug-development gap.

Visualizations have the power to help translate and cross-fertilize vital concepts across disciplinary boundaries—as did the discovery of the DNA structure by James D. Watson and Francis H.C. Crick in 1953 (see *Atlas of Science*, **page 121**). Visualizations may also serve to inspire and support future discoveries (see *The Visual Elements Periodic Table* in *Atlas of Science*, **page 115**).

Other visualizations raise our awareness of both human unity and fragility, such as the *Earthrise* picture, taken by astronaut William Anders during the *Apollo 8* mission in 1968.

In general, most people have a deep respect for facts and arguments expressed as numbers or visualizations. However, they often don't understand just how many different decisions need to be made in order to render data into insights. Information visualization designers play a key role in making that process more transparent. In addition to revealing data, analysis, and visualization details, they must provide pointers to supplemental information, as such details are vital for the proper interpretation of visualization results.

Doing It Yourself

Just as anybody can learn to cook, anybody can learn to analyze and visualize data. In a data-driven world, this is not only possible but also necessary for high productivity and intelligent decision making. This *Atlas* aims to teach general approaches and techniques that are independent of specific implementations and tools. Specifically, the subsequent double-page spread introduces a general workflow and a visualization framework that aim to guide the design of effective visualizations. As a new view of data will often also expose new data issues or inspire new questions, being able to rapidly generate and interpret results is an extremely powerful skill. As many data sets cannot be shared freely and the expertise of practitioners is invaluable for data selection and interpretation, it is desirable that as many individuals as possible acquire basic data visualization literacy. Those who master the basics can begin to find data visualization both fun and empowering while quickly advancing their skills.

Terminology

The following pages draw from many different areas of science, each with its own specific history, culture, and language. An algorithm cited in this section may have been originally developed in mathematics, physics, or biology; or a chart that appears here may be one used by engineers, economists, and statisticians alike, though each group will call it by a different name. This *Atlas* aims to introduce and exemplify an internally consistent approach and language for the design of insightful visualizations, which builds on and uses terminology from existing lines of research. Selecting key concepts and the best names for them posed a key challenge in the writing of this book. The ultimate choices were guided by the need for consistency within and universality across different conceptualizations and terminologies. References to original works as well as alternative names are given whenever new concepts and terminology are introduced (see **page 178**, **References & Credits**).

Disclaimer

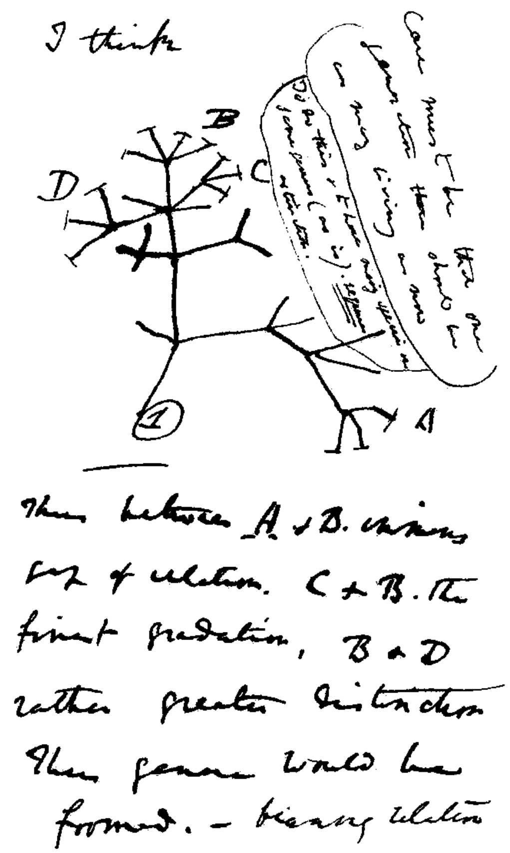
Part 2 reviews general "timeless" approaches and design principles. For "timely" step-by-step tutorials and practical design tips or reviews of specific tools, please see Katy Börner and David E. Polley's *Visual Insights*, Nathan Yau's *Visualize This*, Derek Hansen et al.'s *Analyzing Social Media Networks with NodeXL*, or Felice Frankel's *Visual Strategies*.

Visualizations are used to illustrate key concepts. See also **Part 3** (**page 75**) for detailed explanations of 40 large-scale maps; books by Edward R. Tufte for expert descriptions of hand-drawn visualizations; and recent books by David McCandless, Manuel Lima, and Sandra Rendgen for a rich assortment of highly innovative and colorful charts, graphs, and infographics.

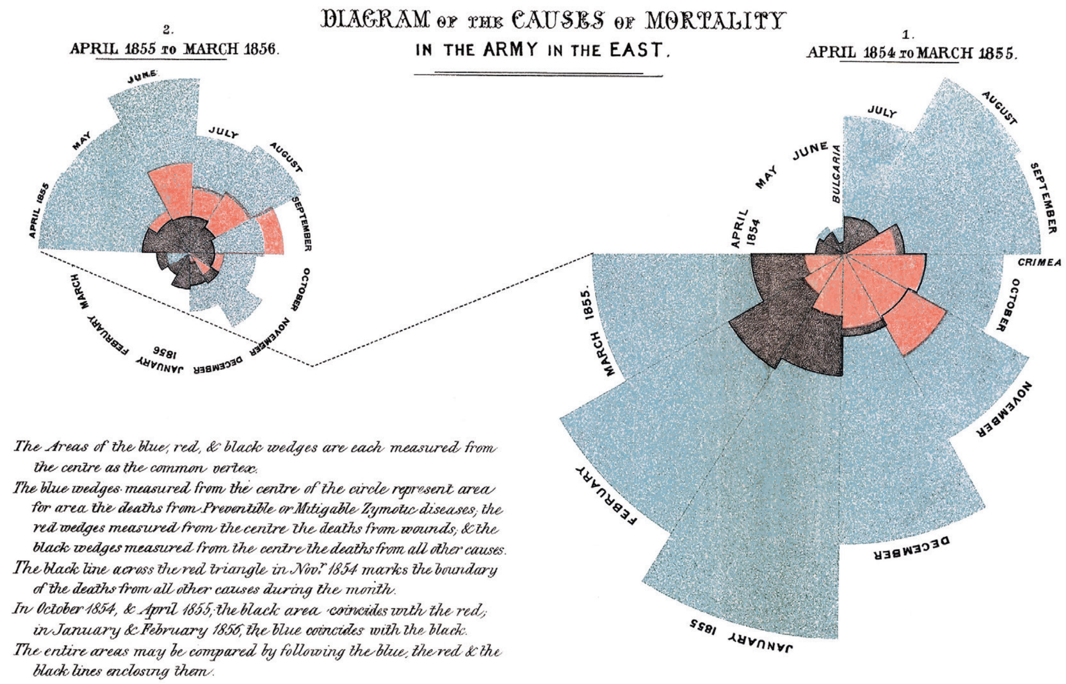
The *Atlas of Knowledge* focuses on the design and use of computer-generated (rather than hand-drawn) visualizations, which have the potential to empower anyone to make sense of big data. Toward that end, simple yet effective and validated visualizations are favored over complex visualizations designed primarily for experts.

Tree of Life

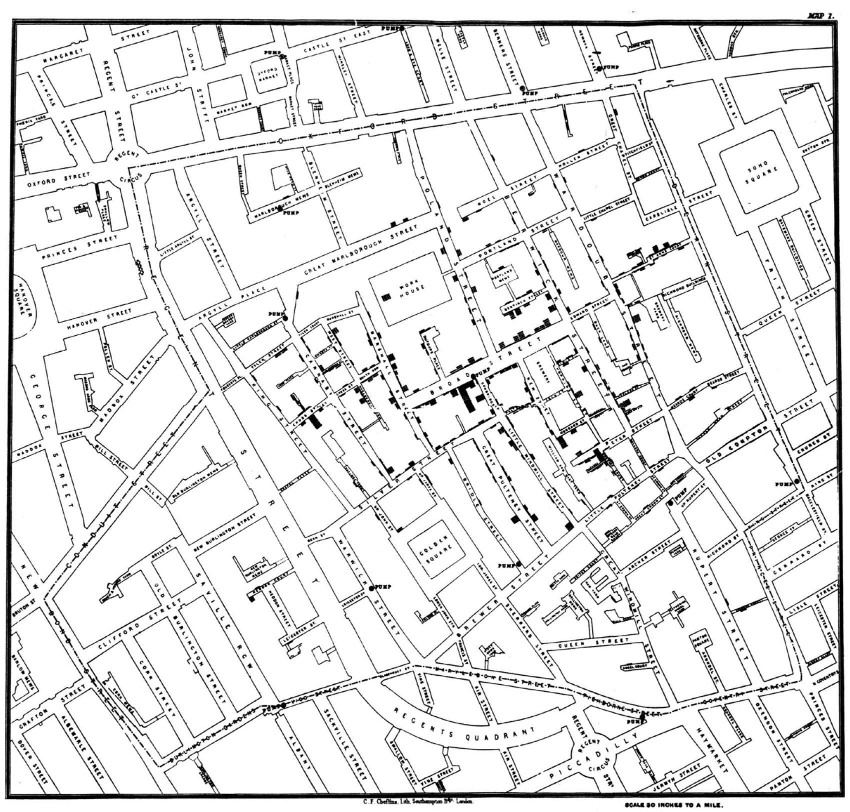
In this first sketch of an evolutionary tree (or branching diagram), Charles Darwin shows the tree's main trunk, labeled 1, as it divides and ends in leaf nodes, indicated by cross strokes. Major branches, labeled A through D, indicate living species. Twigs terminating abruptly and emerging at lower points along branches represent extinct species.



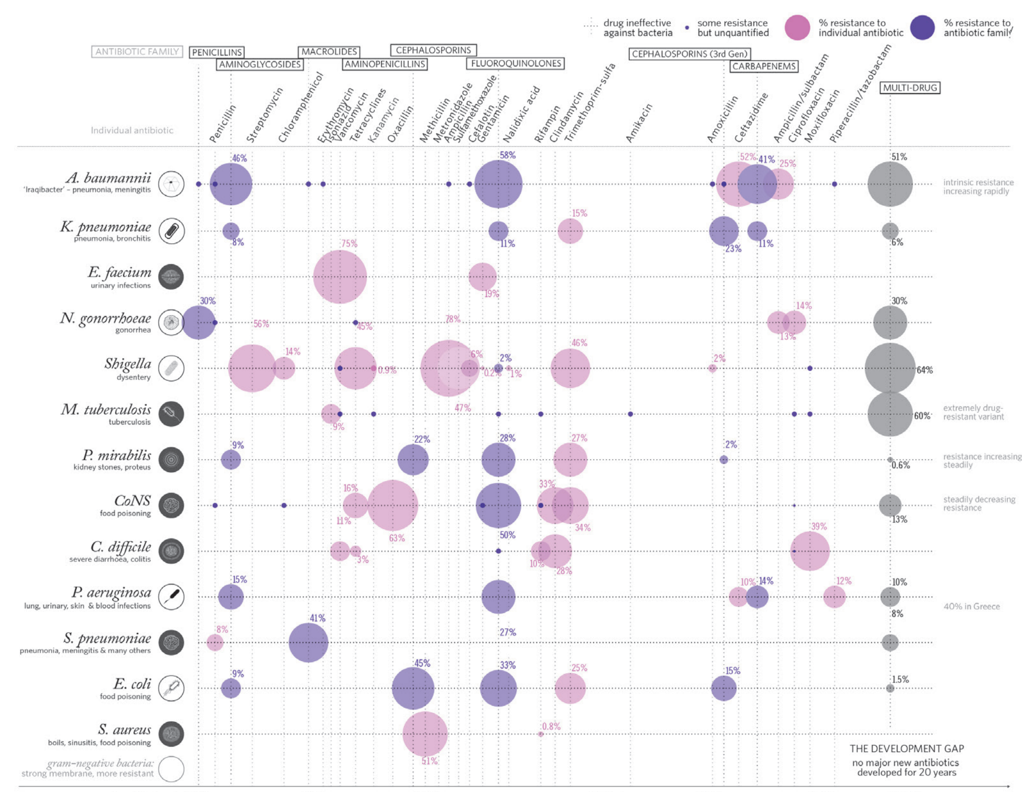
Causes of Mortality in the British Military during the Crimean War



Spot Map of the Golden Square Cholera Outbreak



The Antibiotic Abacus: Adding Up Drug Resistance



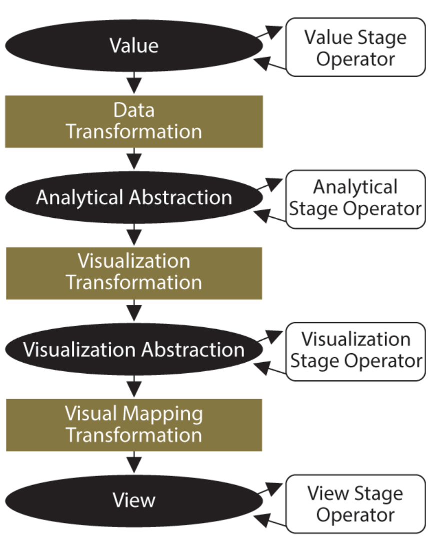
Needs-Driven Workflow Design

This double-page spread discusses the iterative design of data analysis and visualization workflows. The proposed workflow underscores the importance of having a deep understanding of user needs, expertise, and work environment. It groups and labels key processes in the data analysis and visualization workflow; emphasizes the sequential process of data reading and analysis as well as the parallel optimization of different visualization layers and deployment options; and stresses the importance of expert interpretation and validation. In addition, this spread introduces a theoretically grounded yet practically useful visualization framework that supports the design of effective visualizations.

Tell me, I forget. Show me, I remember. Involve me, I understand.
Benjamin Franklin

Visualization Taxonomies and Frameworks

Many visualization taxonomies and frameworks have been proposed (for key works, see **page 178**, **References & Credits**). Ed Chi's information visualization data-state reference model is exemplarily shown below. It identifies three transformations that convert the raw data values into a visualization view: The **Data Transformation** reads the raw data values and generates an analytical abstraction of the data, also called metadata. The **Visualization Transformation** takes that analytical data abstraction and reduces it to a visualization abstraction that can be visualized. The **Visual Mapping Transformation** reads that visualization abstraction and generates a static or interactive graphical view of the data.



Although Chi's model looks rather linear the overall process is typically very iterative and circular. Ideally, users are able to flexibly select the data that is used, the analytical abstraction that is run, and the visual mappings that are applied.

This *Atlas* series promotes (1) a needs-driven, highly iterative workflow design that combines sequential data analysis and parallel visualization design optimization; (2) argues for a clear separation of reference systems (also called base maps) and data overlays to ease the interpretation and generation of visualizations; and (3) introduces a visualization framework that distinguishes different types of insight needs (**page 26**), data scales (**page 28**), visualizations (**page 30**), graphic symbols (**page 32**), and graphic variables (**page 34**) in support of effective visualization design and transfer of visualization solutions across disciplinary boundaries. All three elements are discussed below.

Workflow Design

The *Atlas of Science* (**page 51**) discussed data acquisition, preprocessing, analysis, modeling, and visualization layout as the basic building blocks in data analysis workflows. The figure on the right shows the key elements and processes involved in the design of workflows. Starting with stakeholders in the top-left corner of the figure, workflow design involves four major tasks: **Acquire**, **Analyze & Visualize**, **Deploy**, and **Interpret**. Acquire comprises user needs analysis as well as data acquisition and preparation. **Analyze & Visualize** reads data and applies computational algorithms to convert data into visual insights. Deploy refers to the selection of output devices (e.g., paper printouts, online interactive interfaces) and the design of interactive user interfaces that might be interactive or feature combinations of multiple data views. The interpretation and validation of visualizations tend to inspire new hypotheses, insight needs, and future studies making the workflow design process highly iterative. The four tasks are used to organize Part 2—see section titles and page numbers given next to each task—effectively serving as a visual index to specific content. Subsequently, the importance of a detailed user and task analysis, access to high quality data, the sequential versus parallel nature of data acquisition, analysis, and visualization, and expert validation are discussed.

Users Are Central

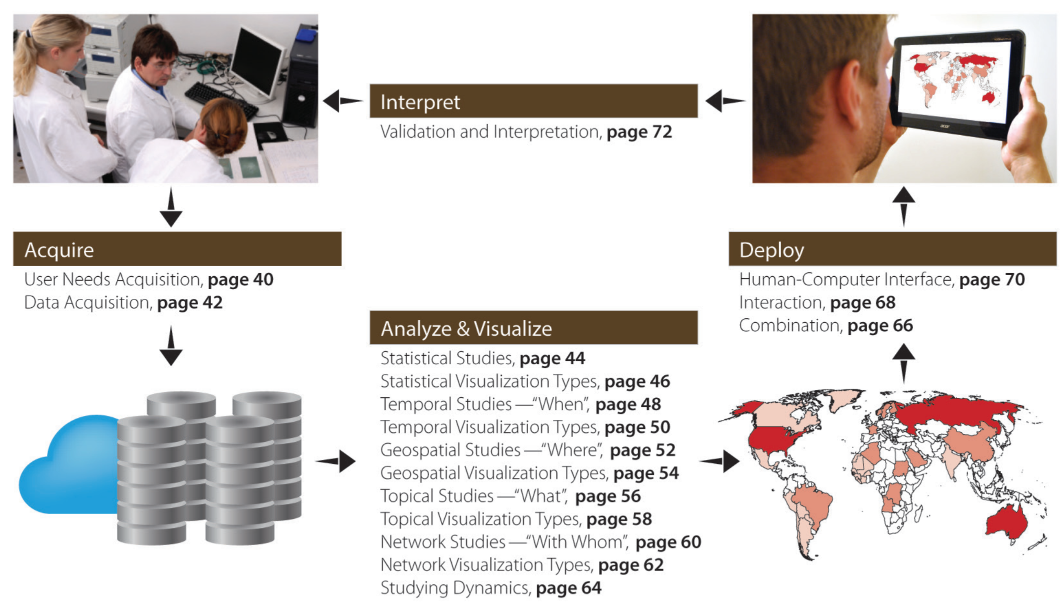
Detailed knowledge of user needs, expertise, and work environment is key for the design of successful visualizations. It is important to understand the type and level of analysis that users need (**see page 4**, **Systems Science Approach**); the insight needs they have (e.g., search versus comparison); the hardware-software combinations they use, as that affects deployment; and the level of data visualization literacy they currently have (e.g., what visualization types they can read and create). Involving users in data compilation, analysis, and visualization is the *only* way to ensure accuracy and relevance of results (see **page 40**, **User Needs Acquisition**).

Data Quality and Coverage

Data quality and coverage affect the type and level of analysis that can be performed. Answering "when" questions requires that data records have time stamps. Individual and global studies require data at the individual and global levels, respectively. Comparison tasks can only be supported if equivalent data on the entities to be compared is available. Data variables may be qualitative or quantitative (see **page 28**, **Data Scale Types**), influencing which visual encodings can be used (see **pages 30–39**). Data size will affect download speed and the display space that is required (see **pages 66–71** on deployment).

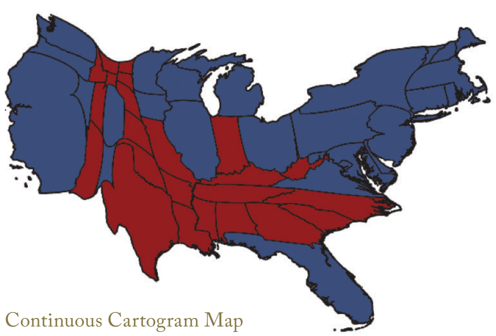
Sequential Data Acquisition and Analysis

The acquisition, cleaning, and analysis of data are commonly done using a sequence of steps that build on each other. For example, a data preprocessing step might delete existing data variables (e.g., by eliminating duplicates), merge them (e.g., by linking publication and funding data based on unique scholar names), or split them (e.g., by distinguishing male from female authors). Alternatively, a processing step can add new data variables (such as latitude and longitude information for postal addresses) or introduce linkages between data records (e.g., coauthor information on publication records can be used to extract coauthor networks). That is, the result of each processing step is a data set that may have different numbers and types of records and data variables. Similarly, different types of analysis might be applied to extract existing or calculate new data variables. For instance, publication year and title information might be used to identify topic trends and coauthor networks might be analyzed to identify backbones or clusters. Sequential application of different analyses ensures that all computed values are ready for use when generating the visualization—there is no need to combine the results from different parallel analyses.

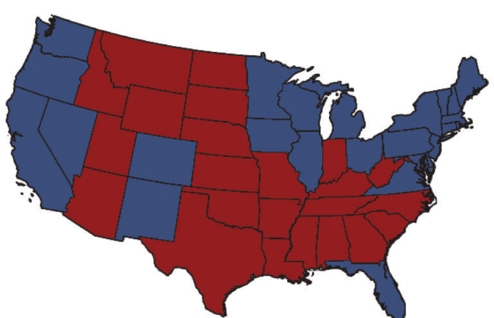


U.S. Map of Contiguous States

Disjoint Cartogram Map



Continuous Cartogram Map



Choropleth Map

Proportional Symbol Map with Line Overlays

Parallel Visualization Optimization

The *Atlas of Science* (**page 51**) introduced nine visualization layers, all of which can be grouped into visualization and deployment. Basically, visualization design comprises the selection of a base map reference system and the design of data overlays (see subsequent section). Deployment requires selecting an output medium and designing appropriate visual combinations and interactivity. Each of these subtasks or selections impacts all others. For example, selecting a small handheld device as preferred output medium considerably restricts the detail of the reference system (e.g., when using a world map, only general country outlines and few labels can be shown) and the number of data records that can be visualized; it also increases the need for effective interactivity design.

Expert Validation

It is absolutely mandatory to involve key stakeholders not only during user and task analysis, data acquisition, analysis, visualization, and deployment but also during the interpretation and validation of results. As data complexity and size increase and problems become more interdisciplinary in nature it might be necessary to involve experts with different knowledge and expertise. Different validation criteria and validation methods exist and can be applied to ensure visualizations are correct, readable, and actionable (see **page 72**, **Validation and Interpretation**).

Reference System versus Data Overlay

The *Atlas* series argues for a clear separation of reference systems (also called base maps) and data overlays. This separation makes it possible to cleanly separate reference systems (such as a Cartesian coordinate system, geospatial map, or anchoring background image of a brain) that are used in different scientific disciplines; it helps understand differences in how data is projected onto the different reference systems; and teach commonalities and differences in the design of data overlays for different visualization types.

In this *Atlas* series, a reference system defines the space onto which all data is projected. In order for users to read a visualization properly, the reference system must be well-defined and easy to understand. Data overlays are defined as a mapping of data record variables to proper **graphic symbol types** (e.g., circles or squares; see **page 32**) and **graphic variable types** (e.g., position, color or shape; see **page 34**). To give an example, a set of five maps is shown on the left. *The U.S. Map of Contiguous States* on the top is the reference system, or the base map. Below it, four data overlays are given. The *Disjoint Cartogram Map* plots data onto the size of each state by rescaling each state around its centroid, which preserves local shape but not topography. *The Continuous Cartogram Map* and the *Choropleth Map* both display 2012 U.S. presidential election results. States in red represent a majority vote for the Republican candidate, Mitt Romney; those in blue reflect a majority vote for the Democratic candidate, Barack Obama. The continuous cartogram sizes states according to their population size: the red areas are considerably reduced while blue areas are expanded providing a different view of the election results. The last map, entitled *Proportional Symbol Map with Line Overlays* shows a combination of data overlays: major U.S. airports are denoted by circles, which are size-coded by traffic data; atop are flights out of Chicago O'Hare International Airport, each represented by a line.

Reference system and data overlay together determine the resulting visualization type. For example, data variables (e.g., population counts, election results, or flight connections between geolocations) might be visualized by (1) distorting the size and/or shape of the base map, to produce what is called a cartogram; (2) visually encoding base map areas (e.g., color-coding them) in what is called a choropleth map; (3) modifying the Z dimension in a stepped relief map (see **page 53**, *In the Shadow of Foreclosures*); (4) visually encoding nodes in a proportional symbol map; or (5) visually encoding links in a linkage map.

Visualization Framework

The problem-solving space that needs to be traversed to arrive at a successful visualization solution is high-dimensional and inherently complex. Many different proposals exist on how to structure this space to make it easier to navigate and manage. The visualization framework proposed in this Atlas draws on work developed in different disciplines of science. Specifically, it distinguishes insight need types (**page 26**): sorting, trends, geospatial locations, relationships, etc.; data scale types (**page 28**): nominal, ordinal, interval, and ratio data; types of analysis (**page 4, Systems Science Approach**): temporal (when), geospatial (where), topical (what), and trees and networks (with whom); levels of analysis (**page 4, Systems Science Approach**): micro, meso, and macro; visualization types (**page 30**): table, chart, graph, map, and network layout; graphic symbol types (**page 32**): geometric symbols, linguistic symbols, and pictorial symbols; graphic variable types (**page 34**): position, form, color, texture, etc.; and, last but not least, interaction types (**page 26**): zoom, search, filter, etc., see below listing of all types discussed in Part 2. The framework creates a "periodic table" of reference systems and data overlays, which can help to identify promising visualization combinations. It is then applied to discuss data acquisition (**pages 40–43**); analysis and visualization of different types of data using approaches ranging from statistics to network science (**pages 44–65**); deployment (**pages 66–71**); and interpretation and validation (**pages 72–73**).

Visualization Framework

Insight Need Types page 26	Data Scale Types page 28	Visualization Types page 30	Graphic Symbol Types page 32	Graphic Variable Types page 34	Interaction Types page 26
• categorize/cluster • order/rank/sort • distributions (also outliers, gaps) • comparisons • trends (process and time) • geospatial • compositions (also of text) • correlations/relationships	• nominal • ordinal • interval • ratio	• table • chart • graph • map • network layout	• geometric symbols point line area surface volume • linguistic symbols text numerals punctuation marks • pictorial symbols images icons statistical glyphs	• spatial position • retinal form color optics motion	• overview • zoom • search and locate • filter • details-on-demand • history • extract • link and brush • projection • distortion

Insight Need Types

Visualizations commonly support either communication or exploration. While the former visualizations are mostly polished and static, the latter are less polished yet interactive. Jacques Bertin argues that a graphic representation might fulfill three functions: recording of information, communicating information, and processing information. Robert L. Harris distinguishes graphs for analyzing and planning; monitoring and controlling; and communicating, informing, and instructing. This spread reviews basic task and interactivity types and proposes a unifying naming scheme with descriptions and examples.

For a person to become deeply involved in any activity it is essential that he knows precisely what tasks he must accomplish, moment by moment.
Mihaly Csikszentmihalyi

Framework

This section defines a set of basic task types and a set of interactivity types. The former help guide the selection of visualization types (**page 30**), graphic symbol types (**page 32**), and graphic variable types (**page 34**). The latter guide interaction (**page 68**) and human–computer interface design (**page 70**).

For both types, i.e., basic task types (see table below) and interactivity types (see table in top-right), key approaches are discussed and a unified naming schema is proposed. Note that alignment in approaches is extremely difficult to attain and most likely imperfect, as most authors and tool developers do not provide a definition of the terms they use. Plus, the approaches were developed for very different purposes—from organizing materials in a book to helping users select appropriate visualizations.

Basic Task Types

A table of basic task types, identified by different scholars and tool developers, is shown below. Columns are sorted by time, left to right.

Jacques Bertin aims to identify tasks that can be mapped to graphic variable types, which he calls visual variable types (see **page 34**). Bertin identifies selection (whereby marks are perceived as different, forming families), order (whereby marks are perceived as ordered), association (or similarity, whereby marks are perceived as similar), and quantity (whereby marks are perceived as proportional to each other). While the first three task types are used to encode qualitative data, the last is relevant for quantitative data. Stephen Wehrend and Clayton Lewis distinguish ten general retrieval tasks, such as locate (search for a known object), identify (object is not necessarily known), distinguish, categorize, cluster, see distribution, rank, compare (within entities and between relations), associate, and correlate. Six of these ten tasks are relevant for data analysis and visualization and are given in the table. Stephen Few's Graph Selection Matrix was designed to help identify what graph type (point, line, bar, or box plot) is best for what task. It distinguishes different featured relationships, such as ranking, distribution, nominal comparison and deviation, time series, geospatial, part-to-whole, and correlation. Nathan Yau distinguishes five visualization types: patterns over time, proportions, relationships, differences, and spatial relations. Sandra Rendgen and Julius Wiedemann organize more than 400 visual graphics by location, time, category, and hierarchy. Felice Frankel distinguishes three major purposes of a visual graphic—form and structure, process and time, compare and contrast—and uses them to teach important visual design strategies. Diverse tools and online services exist that aim to empower users to generate different types of visualizations: IBM's Many Eyes site supports visualizations that reveal relationships among data points, compare data values, track rises and falls over time, see parts of a whole, analyze text, and generate maps. Chart Chooser helps users select the right graph by grouping the visuals via comparison, distribution, composition, trend, relationship, and table. The last column of the table shows the set of types that are used in this *Atlas* (see descriptions and examples on opposite page).

Basic Task Types

Bertin, 1967	Wehrend & Lewis, 1996	Few, 2004	Yau, 2011	Rendgen & Wiedemann, 2012	Frankel, 2012	Tool: Many Eyes	Tool: Chart Chooser	Börner, 2014
selection	categorize			category				categorize/cluster
order	rank	ranking					table	order/rank/sort
	distribution	distribution					distribution	distributions (also outliers, gaps)
	compare	nominal comparison & deviation	differences		compare and contrast	compare data values	comparison	comparisons
		time series	patterns over time	time	process and time	track rises and falls over time	trend	trends (process and time)
		geospatial	spatial relations	location		generate maps		geospatial
quantity		part-to-whole	proportions		form and structure	see parts of whole, analyze text	composition	compositions (also of text)
association	correlate	correlation	relationships	hierarchy		relations between data points	relationship	correlations/relationships

Interaction Types

Other scholars have identified interactivity types (see top-right table). For interactive data exploration, Ben Shneiderman cites overview (seeing the entire collection), zoom (zooming in on items of interest), filter (selecting interesting items), details-on-demand (selecting one or a group of items and getting details when needed), relate (viewing relationships among items; see basic task types in lower-left table), history (keeping a log of actions to support undo, replay, and progressive refinement), and extract (access subcollections and query parameters). Daniel Keim distinguishes major interaction techniques such as zoom, filter, and link and brush. The latter technique interlinks multiple visualizations of the same data—users can select data records via brushing in one view to highlight these records in all other views. Keim also lists projection and distortion techniques (e.g., hyperbolic and spherical spaces) as a means to provide focus and context. For additional reference, please see the discussion in **Interaction** (**page 68**).

Interactivity Types

Shneiderman, 1996	Keim, 2001	Börner, 2014
overview		overview
zoom	zoom	zoom
		search and locate
filter	filter	filter
details-on-demand		details-on-demand
history		history
extract		extract
	link and brush	link and brush
	projection	projection
	distortion	distortion

Naming Conventions

In this and all subsequent spreads, the following terminology will be used. Physical or virtual items will be called objects. Objects can be represented by a data record (also called a data point). A data record is an N-tuple (or vector) of data variables. Data variables (also called data properties, feature attributes, or parameters) may be qualitative or quantitative. The value of data variables may change over time. A data set (also called a data series) comprises one or more data records.

The example below shows the records of two scholars, each represented by a 6-tuple. Three data variables are qualitative (**ID**, **Name**, **Country**); all others are quantitative. The **Age** value will increase by one each year.

ID	Name	Age	Country	#Papers	#Citations
1	J. Smith	53	U.S.	101	367
2	J. Chen	45	China	59	150

In order to represent relationships between objects (e.g., scholars), a so-called linkage table can be used. Each link is represented by an M-tuple of data variables. The first two columns commonly represent the IDs of the objects that are linked. Other columns may represent additional attribute values. The table below exemplarily represents the coauthor links between the two scholars above, with **Weight** indicating the number of papers they authored together and **Begin** and **End** denoting the first and last years when a given joint paper was published.

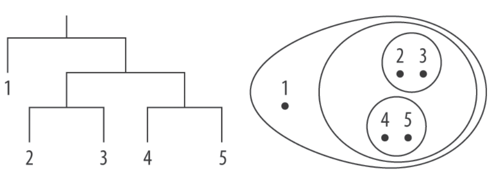
ID1	ID2	Type	Weight	Begin	End
1	2	Coauthor	3	1999	2005

Categorizing and Clustering

Categorization is the assignment of data records to a category (also called cluster, class, or group) of similar data records. Categories might be manually defined or computed using clustering techniques.

Clustering is the task of assigning a set of data records to groups (also called classes or categories) so that objects in the same cluster are more similar to each other than to those in other clusters. Cluster-defining properties may exist in the original raw data (e.g., publication year) or can be computed (e.g., the similarity of papers based on similar word usage. The result of clustering may be a hierarchy (below) or partition with disjoint or overlapping clusters.

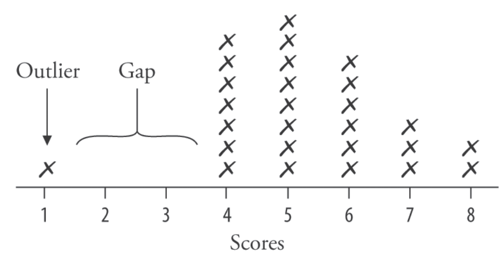
In addition, users may be able to manually explore clusters (see **page 68**, **Interaction**) and group data records. Clustering is frequently applied to make data patterns easier to see and to reduce visual complexity. For further reference, see **Clustering** (**pages 52** and **60**).



Distribution (also Outliers and Gaps)

Distributions capture how objects are dispersed in space. A statistical distribution is an arrangement of the values of a variable that shows their observed or theoretical frequency of occurrence. It supports the detection of outliers and gaps that are important for understanding data quality (uncertainty and missing or erroneous data) and data coverage (pedigree and scale).

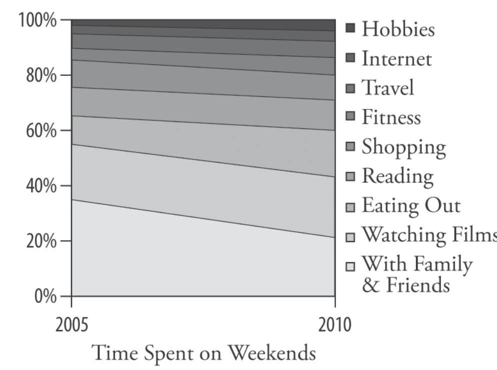
The example below shows the distribution of **Scores** for an imaginary exam. Each ✗ represents the score of one student, with most students achieving a score of 4 to 6. Five students scored higher, at 7 or 8. The single student who scored 1 is considered an **Outlier**; a **Gap** is shown between that student and the others. For further reference, see **Statistical Studies** (**page 44**).



Trends

A pattern of gradual change in the average or general tendency of data variables in a series of data records is called a trend. Trends can vary in length (from short-term, to intermediate, to long-term) and strength (in terms of the amount of change and the number of data variables and data records involved); see examples in **Temporal Studies—"When"** (**page 48**). Trends are commonly represented using a graph or map.

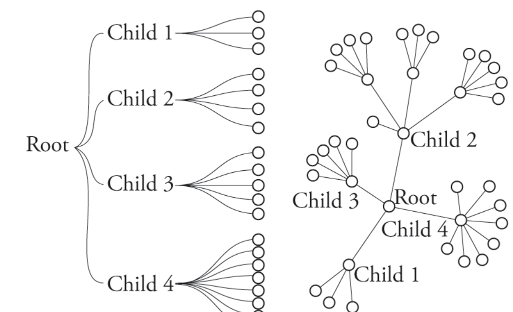
The comparison below of how people spent their weekend time in 2010 versus in 2005 shows a significant decreasing trend for spending time overall **With Family and Friends** and a milder increasing trend for specific activities such as **Eating Out**.



Composition (of Objects and of Text)

Composition refers to the way distinct parts or objects are arranged to form a whole. Part-to-whole relationships are important, as is the individual form and structure of the parts and the whole. Composition also refers to the process of putting words and sentences together to create text; see **Topical Studies—"What"** (**page 56**).

The two visualizations below show the number of directories and subdirectories in a file hierarchy as a tree view (left) and a force-directed layout (right); see **Network Studies—"With Whom"** (**page 60**).



Ordering, Ranking, and Sorting

Ordering (also called sorting) refers to the arrangement of objects in relation to one another according to a particular sequence, pattern, or method. The position in a sorted arrangement of objects is called a ranking.

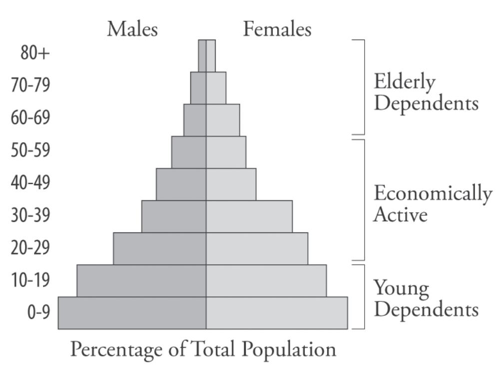
Shown below-left is an alphabetically sorted list of subsection titles, with the title in the fifth rank highlighted. Given on the right is a numerically sorted list of numbers. Items may also be sorted by size, speed, or other data properties.

Subsection Titles	Numbers
Categorizing and Clustering	3
Comparison	5
Composition (of Objects and of Text)	19
Correlations and Relationships	220
Distribution (also Outliers and Gaps)	23
Geospatial Location	29
Ordering, Ranking, and Sorting	101
Trends	1,000

Comparison

A comparison refers to the process of examining two or more objects to establish similarities and dissimilarities. Single data values, objects with many data values, object groups, or object interlinkages can be compared. Visual comparisons become easier if visualizations are shown side by side.

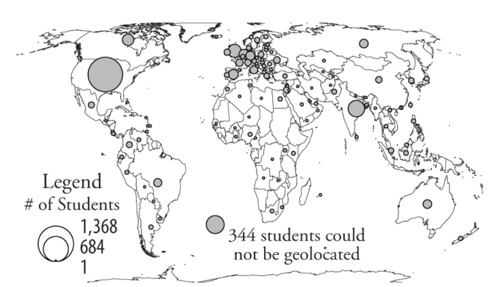
An example is the population pyramid below, which shows the number of male (left) and female (right) citizens per age group. Numbers decrease as age increases, with women shown to live slightly longer than men.



Geospatial Location

Geospatial location refers to a particular place or position. Two geometric objects can have diverse spatial relationships, defined by such "predicate" terms as equal, disjoint, intersects, touch, overlap, cross, within, or contain. A map is commonly used to show the locations, forms, sizes, and spatial relationships of objects; see description in **Geospatial Studies—"Where"** (**page 52**).

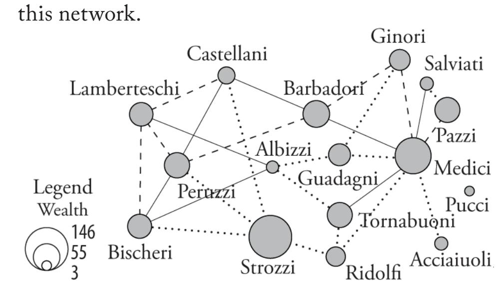
Shown here is a map of the world with a proportional symbol overlay that reveals the origin and number of students who registered for the spring 2014 Information Visualization MOOC course at Indiana University by the end of May 2014. Although 1,368 of the more than 3,600 students were based in the United States, students came from more than 200 countries.



Correlations and Relationships

Correlations express the relationship between two or more objects or attribute values. Relationships can have different cardinality: One-to-one relationships (e.g., position rank vs. income) are commonly represented by scatter plots and other graphs (see **page 44** and **47**, **Correlations**). One-to-many or many-to-many relationships are typically communicated using network visualization types; see **page 60**. Networks might have one or more node types and one or more link types. Links might be undirected or directed, unweighted or weighted.

The network below shows 16 nodes representing Italian families, size coded by wealth, and interlinked by marriage (dotted) and business (dashed) relationships, or both (solid). See **page 62**, **Radial Tree** for an alternative layout and a discussion of this network.



Data Scale Types

Data can be qualitative or quantitative. Qualitative data take on only specific values with no values in between and are frequently determined by counting. Examples are names or job types. Quantitative data may take on any value within a finite or infinite interval and are commonly acquired via measurement. Examples are time or counts. In 1946, Harvard psychologist Stanley S. Stevens coined the terms "nominal," "ordinal," "interval," and "ratio" to describe a hierarchy of data scales. This spread reviews existing works for the classification of data scale types. Specifically, it describes and exemplifies Stevens's data scale types and discusses their utility and limitations.

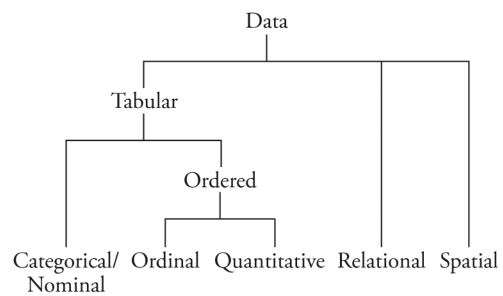
Not everything that counts can be counted, and not everything that can be counted counts.
Albert Einstein

Framework

Many different definitions exist for data scale types. Key works are shown in the table below. In his 1946 paper "On the Theory of Scales of Measurement," Stanley S. Stevens distinguished nominal, ordinal, interval, and ratio data based on the type of logical mathematical operations that are permissible (see section **Mathematical Operations** and table top-right). That is, the type of scale used depends on the mathematical transformations that can be performed on the data.

In 1967, Jacques Bertin argued for three data scale types: qualitative, ordered, and quantitative—which roughly corresponds to nominal, ordinal, and quantitative (also called numerical). His terminology was adopted by geographer Alan MacEachren, and many other cartographers and information visualization researchers. Robert Harris's *Classification of Scales* distinguishes the same three types as Bertin but calls them category, sequence, and quantitative.

Visualization researcher Tamara Munzner distinguishes tabular, relational, and spatial data; then further divides tabular into categorical/nominal and ordered; and finally subdivides ordered into ordinal and quantitative (see *Data Hierarchy*



Data Hierarchy

above). Using this classification, tabular visualizations such as GRIDL (**page 69**) or Gapminder (**pages 65** and **71**) may have categorical/nominal or ordered axes. Relational data refer to linkages between data records, which may be categorical (e.g., "marriage," "business"; see **page 27, Correlations and Relationships**) or weighted (quantitative), and are commonly represented using network visualizations (see **page 62, Network Visualization Types**). Spatial data (e.g., latitude and longitude information) is needed to geolocate records (see **page 54, Geospatial Visualization Types**).

Stevens's approach has been adopted here and is shown in the right-most column of the below table. The title was revised to *Data Scale Types* to match other terminology in the visualization framework. Descriptions and examples of the different data scale types can be found on the opposite page.

Data Scale Types

Stevens, 1946 *Scales of Measurement*	Bertin, 1967 *Level of Organization of the Components*	Harris, 1996 *Classification of Scales*	Munzner, 2011 *Visualization Principles*	Börner, 2014 *Data Scale Types*	
nominal	quantitative	category	categorical/nominal	nominal	More Qualitative
ordinal	ordered	sequence	ordinal	ordinal	
interval	quantitative	quantitative	quantitative	interval	More Quantitative
ratio	quantitative	quantitative	quantitative	ratio	

Conversions

Simple transformations can make real-world data more amenable to analyses and visualizations that truly satisfy users' needs. For example, quantitative data scale types can be converted into qualitative data scale types, or thresholds can be applied to convert interval data into ordinal data. Rankings (ordinal) are commonly converted to yes/no categorical decisions (e.g., with hiring or funding decisions). Typically, this is done in such a manner that equal groups result, and different approaches may be appropriate for different types of distributions (see **page 44, Statistical Studies**).

The reverse is possible as well: more qualitative data scale types can be converted into more quantitative data scale types. For example, Robert P. Abelson and John W. Tukey mapped ordinal scales onto interval scales and estimated the amount of error that resulted. Tukey also discussed situations in which interval scales (e.g., measurements from a miscalibrated scale) should be converted to a ratio scale that behaves more simply. Roger N. Shepard, Joseph B. Kruskal, and others developed multidimensional scaling methods to convert ordinal into ratio scales. See **page 178, References & Credits,** for details.

Mathematical Operations

Stevens distinguished types of scale based on the type of logical mathematical operations that are permissible. Major operations for all four types are given in the top-right table. Check marks indicate permitted operations, whereas cross-outs indicate that particular operations cannot be performed with the given data type. All types support determining equality and inequality (such as by identifying and categorizing the members of a numerical series). All but nominal types can be ordered (e.g., alphabetically or numerically). Only interval and ratio types support determining if differences are equal (e.g., 2 – 0 = 4 – 2). Ratio types also support operations that determine if aspects of objects (or numbers) are equal (e.g., 4/2 = 8/4). The bottom row shows the operations used to measure central tendency for the different data types (see also **page 44, Statistical Studies**).

Limitations

The four scale types do not account for all the data that one may encounter or measure. For example, percentages (which are bounded at both ends and

Data Scale Types		Nominal	Ordinal	Interval	Ratio
Logical Mathematical Operations	× ÷	×	×	×	√
	+ -	×	×	√	√
	< >	×	√	√	√
	= ≠	√	√	√	√
Measure of Central Tendency		mode	median	arithmetic mean	geometric mean

cannot tolerate even arbitrary scale shifts) cannot be classified in this system. In his seminal paper, Stevens argued for using the four data scale types for classifying and selecting permissible statistical procedures. A number of textbooks and analysis tools implemented his recommendation. However, given the fact that the four scale types are not able to capture all possible data and that scale types can be converted into other types, these automatic permissibility rules restrict the possible set of valuable analyses and could even lead to the selection of inaccurate analyses.

Applications

Data should never prescribe analyses or visualizations. Instead, user needs (translated into the questions asked of the data) should influence what data is collected and how it is used. For example, if a ranking of scholars is desired then nominal data variables are inappropriate but ordinal, interval, or ratio data variables are necessary (see example in section **Nominal Scale** on opposite page). If calculating the arithmetic mean of a variable is important then interval or ratio scale data has to be acquired.

Documentation

Psychologists emphasize the importance of documenting exactly what data scale has been used to acquire any given data, why that scale was developed (e.g., for intelligence tests), who should complete the scale, how the scale should be used and scored (including sample items and values), and the scale's characteristics. Without this information, data collected for specific purposes runs the risk of being inappropriately used in psychology and other fields of science.

Nominal Scale

A nominal scale (also called a categorical or category scale) is qualitative. Categories are assumed to be nonoverlapping in that each data variable is assigned to one category and no two variables are assigned to the same category.

Examples include dichotomous and nondichotomous data. A dichotomous (or dichotomized) example is an attribute that can be either "true" or "false." Nondichotomous examples (comprising multiple categories) are words or numbers constituting the names and descriptions of people, places, things, or events. Each word or number defines a distinct category that contains one or more entities. It is possible to have multiple assignments within a nominal category (e.g., a person can be bi-racial or have multiple nationalities or jobs).

Nominal data can be counted (e.g., the number of male/female scholars in an institution or the number of scholars per country). The results may then be displayed in frequency tables and graphs. Shown below is a fictive set of faculty members who work on an interdisciplinary research topic at a U.S. university and the counts of their departments, courses, books, and funding awards.

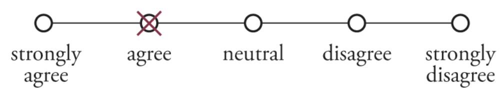
Entity Type	Count
Books	205
Courses	27
Departments	53
Faculty	55
Funding Awards	501

Mathematical qualitative operations such as equal and not equal can be performed (see the table on the opposite page, top-right). Although words and numbers that label or describe categories can be sorted alphabetically, they cannot be ranked or mathematically manipulated. No quantitative distinction can be drawn among them, as there is no intrinsic ranking or order. The mode, or the most common item, is allowed as the measure of central tendency for the nominal type. The median, or the middle-ranked item, makes no sense for the nominal type of data, because ranking is not allowed. Similarly, taking the mean on a nominal variable has no meaning.

Ordinal Scale

An ordinal scale (also called a sequence or ordered scale) is qualitative. It sorts or rank-orders values representing categories that are based on some intrinsic ranking but not at measurable intervals. That is, there is no information as to how close or distant values are from one another.

Examples include dichotomous and nondichotomous data. Dichotomous examples include "sick" versus "healthy" or "guilty" versus "innocent." Nondichotomous examples include days of the week or months in a year; job ranks within a workplace; degrees of satisfaction and preference rating scores (as with a Likert scale, offering **strongly agree**, **agree**, **neutral**, **disagree**, and **strongly disagree** choices that users can check; see below); or rankings such as low, medium, and high.



For ordinal string variables, alphabetical sorting might be applied (e.g., when listing index terms). However, that understanding cannot be applied when data follow a nonalphabetical order, as do the days of the week (see below; note that in the United States the week starts on a Sunday).

Days of the Week	Alphabetical Sorting
Sunday	Friday
Monday	Monday
Tuesday	Saturday
Wednesday	Sunday
Thursday	Thursday
Friday	Tuesday
Saturday	Wednesday

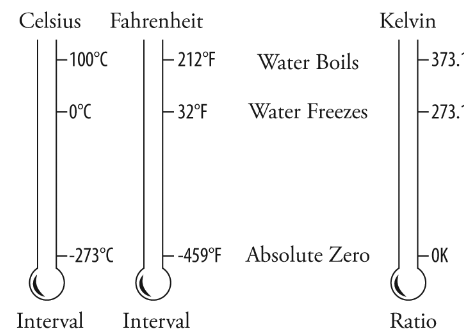
Mathematical qualitative operations, such as determining when figures are equal or not equal, can be performed; the mode and median (or middle-ranked item) but not the mean (or average) can be calculated (see **page 44, Statistical Studies**).

Note that most psychological measurements, such as of opinions or IQ scores, are ordinal. That is, the mean and standard deviations have no validity; only comparisons are valid. There exists no absolute zero, and a ten-point difference may carry different meanings at different points of the scale.

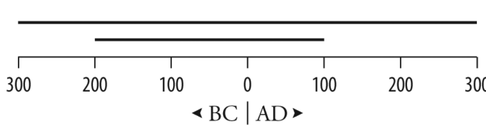
Interval Scale

An interval scale (also called a value or discrete scale) is a quantitative numerical scale of measurement, whereby the distance between any two adjacent values (or intervals) is equal, but the zero point is arbitrary. Interval-type variables are also called scaled variables or affine lines (in mathematics).

Examples are the Celsius and Fahrenheit temperature scales, which have an arbitrarily defined zero point; see the below comparison of both scales with the Kelvin ratio scale. Similarly, an interval scale is used to measure the distance between calendar dates within an arbitrary epoch (such as the AD year numbering system).



Scores on an interval scale can be added and subtracted; for example, the time interval between the first days of the years 1981 and 1982 is the same as that between 1983 and 1984—namely, 365 days. Interval scale values cannot be meaningfully multiplied or divided; for example, 20°C cannot be said to be "twice as hot" as 10°C. However, ratios of value differences can be expressed; for example, one difference can be twice another (see the bars for 600- and 300-year time durations in the figure below).

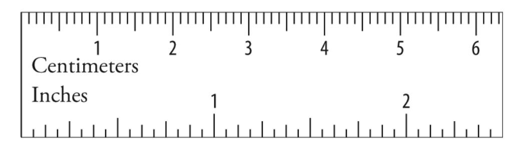


The mode, median, and arithmetic mean can be calculated to measure the central tendency of interval variables, whereas measures of statistical dispersion include range and standard deviation.

Ratio Scale

A ratio scale (also called a proportional or continuous scale) is a quantitative numerical scale. It represents values organized as an ordered sequence, with meaningful uniform spacing, and has a unique and nonarbitrary zero point.

Most physical measurements—including length (see ruler below), weight, height, mass, (reaction) time, energy, and intensity of light—are made on ratio scales. Periods of time can be measured on a ratio scale, and one period may be correctly defined as double another. The Kelvin temperature scale (see image at left) is a ratio scale because it has a unique, nonarbitrary zero point called absolute zero—even if that point is purely theoretical. Other examples of measurements would be the counts of any published papers, coauthors, or citations.



In physics, two types of ratio scales are distinguished: fundamental (e.g., length or weight) and derived (e.g., density or force). Examples are population counts (e.g., per city) and population density counts (i.e., population per unit area or unit volume), respectively. The former may be represented by proportional symbol maps that use size-coded geometric objects to represent the number of inhabitants per city. Population density is commonly represented by choropleth maps (see **page 54, Geospatial Visualization Types**).

A value of zero has special meaning; for example, with respect to age the actual zero point allows one to say that a ten-year-old is twice the age of a five-year-old. Qualitative operations such as addition, subtraction, multiplication, and division can be performed (e.g., length measurements can be converted from inches to feet or from feet to meters via multiplication with a constant). Statistical dispersion, standard deviation, and the interquartile range can all be calculated. In fact, all statistical measures are allowed because all necessary mathematical operations are defined for the ratio scale.

Visualization Types

Many different types of visualizations have been developed by scientists, engineers, designers, artists, and other scholars. Diverse proposals have also been made on how best to organize visualizations into different types—for instance, based on user task, data shown, reference system employed, data overlay provided, deployment used (hand-drawn versus computer-generated), or key insights gleaned. A pragmatic solution is presented here that uses the type of reference system employed as the main criterion. The final set of types selected comprises tables, charts, graphs, maps, and network graphs, as explained and exemplified in this double-page spread.

The best way to learn about visualizations is to make them.
Martin Wattenberg

Framework

The table below lists and compares major approaches to grouping and naming different types of visualizations. Jacques Bertin's *Semiology of Graphics* distinguishes diagrams, maps, and networks. Robert L. Harris distinguishes tables, charts (e.g., pie charts), graphs (e.g., scatter plots), maps, and diagrams (e.g., block diagrams, networks, Voronoi diagrams). Yuri Engelhardt distinguishes proportionally divided space, space with categorical or metric axes, map space, and text space. Ben Shneiderman's taxonomy organizes visualizations according to data types: linear (1D), planar (2D), volumetric (3D), multidimensional (nD), temporal, tree, network, and workspace. Microsoft Excel, a tool widely used, supports the creation of tables and diverse charts, including pie and doughnut charts as well as line and bar graphs. The set of visualization types adopted in this *Atlas* covers five types: table, chart, graph, map, and network graph (see descriptions and examples on right).

Conversions

Simple modifications can transform one visualization type into another. For example, changing the quantitative axes of a graph into categorical axes results in a table (see the *GRIDL* visualization on **page 69**). Interpolating discrete area topic maps as continuous, smooth-surface elevation maps makes them look like geospatial maps (see *In Terms of Geography* in *Atlas of Science*, **page 103** and **page 58, Isoline Map**).

Combinations

Most data sets can be visualized in a variety of ways (see examples on right and **page 66, Combination**). In some cases, the different views may be coupled to support data exploration (see **page 68, Interaction**). For example, human migration data may be depicted using a table of top-N migration flows and a world map with flow overlays; selecting a flow value in the table highlights the corresponding link in the map. That is, each visualization reveals a different aspect of the data set, which in turn leads to different insights (see **page 72, Validation and Interpretation**).

Visualization Types

Bertin, 1967	Harris, 1996	Shneiderman, 1996	Engelhardt, 2002	Tool: MS Excel	Börner, 2014
	table			tables	table
diagram	chart		proportionally divided space, random space	pie, doughnut	chart
diagram	graph	linear (1D), planar (2D), temporal, volumetric (3D), multidimensional (nD)	timeline (metric or ordered), metric axis, ordering axis, categorization axis	column, line, bar, area, surface, scatter, bubble, radar, stock	graph
map	map		map space (metric or ordered)		map
			text space		
network	diagram	tree, network			network layout (tree or network)
		workspace			

Tables

A table is an ordered arrangement of rows and columns in a grid. The space at which one row and column intersect is called a cell. Data values are stored in cells and can be indexed by the respective rows and columns. In most cases, each row holds one data record (see **page 26, Naming Conventions**). Columns are typically used to store data values for different data variables. The first row may be used as a header row, with column names consisting of a word, phrase, or numerical index. Meaningful header names help infer meaning about a dataset. Table elements can be color-coded or size-coded. They can also be sorted, grouped, and segmented in many different ways.

Score	Count
96-100	5
91-95	34
86-90	50
81-85	23
76-80	11
Below 75	1

Alternating Rows Table

Score	Count
96-100	5
91-95	34
86-90	50
81-85	23
76-80	11
Below 75	1

Groupings Table

Table types include frequency, percentage, summary, and quartile tables (see Robert Harris's *Information Graphics: A Comprehensive Illustrated Reference* for more types). Pivot tables are a data summarization that can be used to sort, count, total, average, or cross-tabulate data stored in one table.

Score	Count	Relative Count, %	Cumulative Count
96-100	5	3.85	5
91-95	34	26.15	39
86-90	50	38.46	89
81-85	23	17.69	112
76-80	11	8.46	123
Below 75	1	0.77	124

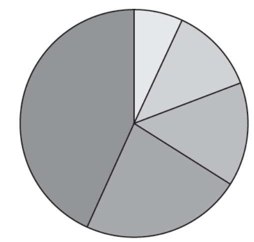
Frequency, Percentage, and Summary Table

Some tables support interactive selection and sorting of rows and columns as well as visual encoding. Cells may contain proportional symbols or small charts/graphs (see example on **page 66** in top-right). Line overlays can be used to show relations between table cells.

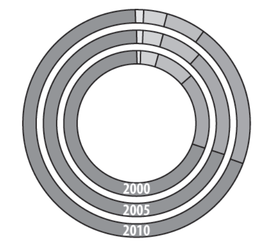
Charts

Charts visually depict quantitative and qualitative data without using a well-defined reference system. They are supported by many spreadsheet programs and are widely used in information graphics.

Examples are pie charts or doughnut charts. The sequence of "pie slices" and the overall size of a "pie" are arbitrary; the pie-slice angles and area sizes represent a percentage of the whole (i.e., the sum of all slices should be meaningful). Examples of a pie chart and doughnut chart with values for three years are shown below. Note that human comparisons made using angles or areas are less accurate than comparisons made using length (see **page 34, Graphic Variable Types**).

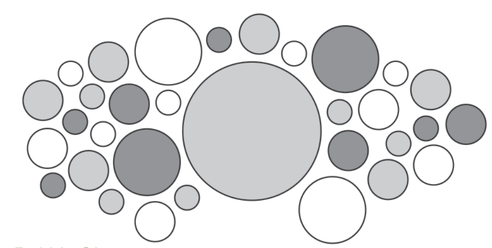


Pie Chart



Doughnut Chart

Bubble charts and tag clouds (also called word clouds) represent each data record with a randomly positioned geometric object or word (see below examples). However, to achieve the most effective use of space or to establish some discernible pattern, position may be specified. For instance, larger items (objects or words) may be set closer to the center, and/or words may be arranged to follow an alphabetical sequence.



Bubble Chart

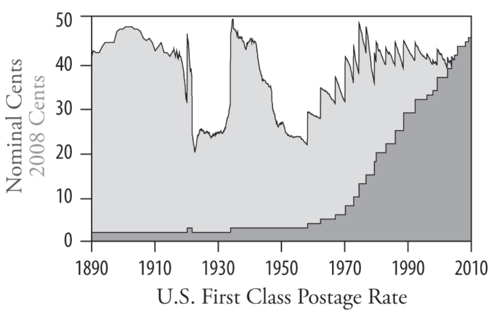
In these and other charts, graphic variable types such as area size, font size, and color may be used to encode additional properties (see **page 34**). Typically, quantitative data variables are used to size-code, whereas qualitative data variables are used to color- or shape-code.

Tag Cloud

Graphs

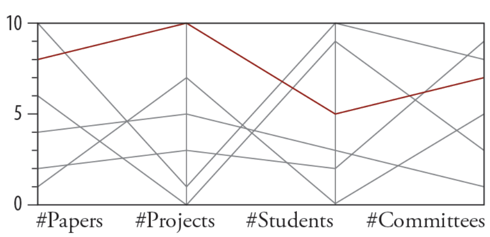
A graph plots quantitative and/or qualitative data variables using a well-defined reference system, such as coordinates on a horizontal or vertical axis. Binning, extrapolation, and smoothing can be applied to aggregate data so that larger data amounts can be more easily understood (see **page 44, Statistical Studies**, and **page 48, Temporal Studies—"When"**). Relationships between data records can be overlaid as links.

Many different graph types exist (see **page 46, Statistical Visualization Types**). Among them are line graphs (see below and discussion on **page 50**), bar graphs, and the stacked versions of each. Scatter plots and bubble graphs (see Gapminder visualizations on **pages 56** and **71**) are widely used.



Line Graph

Parallel coordinate graphs plot multiple data values per record using multiple axes. Links interconnect all values per record (see discussion of this graph on **page 47**).



Parallel Coordinate Graph

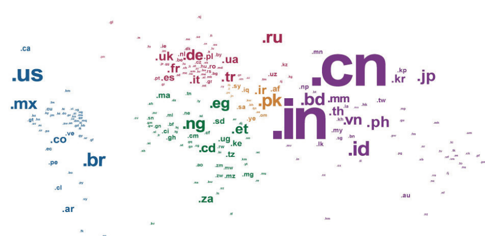
Crossmaps (**page 58, Topical Visualization Types**, and *Atlas of Science*, **page 94**) use a combination of quantitative and qualitative axes (e.g., topics versus time). Geometric symbols may be overlaid (e.g., circles might represent papers on different topics published in different years) and be sized according to some numerical property (e.g., the number of citations per paper). Symbols may also be hue-coded to indicate additional attribute values (e.g., red for review paper, green for research paper). Finally, linkages may be used to denote relations (e.g., citations between papers).

Maps

Maps display data records visually according to their physical (spatial) relationships and show how data are distributed geographically. They are used to show the location, proximity, and distribution of data records. The geolocation of a data record requires the existence of a data variable that defines a location, such as a postal address or a latitude/longitude data pair. Additional data variables can be visualized using graphic variable types (**page 34**) such as area size, font size, and color. Relationships between data records are commonly displayed using links.

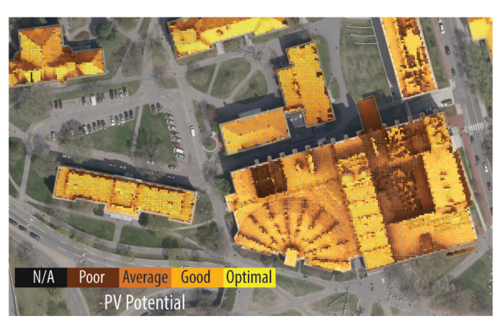
Major map types include cartograms, choropleth maps, relief maps, and proportional symbol maps (see **page 24, Needs-Driven Workflow Design**, and **page 54, Geospatial Visualization Types**).

The *Country Codes of the World* map below shows 245 country codes—the top-level domain codes or extensions used at the end of any internationally based URL or email address. Each two-digit country code is mapped according to the location of the country or territory it represents and color-coded by continent. It is also sized relative to the population of that region (with the exception of China and India, whose codes have been scaled at only 30 percent of their population size in order to fit the layout).



Proportional Symbol Map Showing Country Codes of the World

Data overlays may be either continuous or discrete and may display data for all areas or for selected areas only. Shown below is a choropleth map (**page 54**) that visualizes the potential of rooftop surface areas for solar energy generation. Dark brown denotes low potential; yellow indicates optimal potential.



Choropleth Map Using Roof Top Grid Layout

Network Layouts

Network layouts use nodes to represent sets of data records, and links connecting nodes to represent relationships. Different representations exist for tree and network structures.

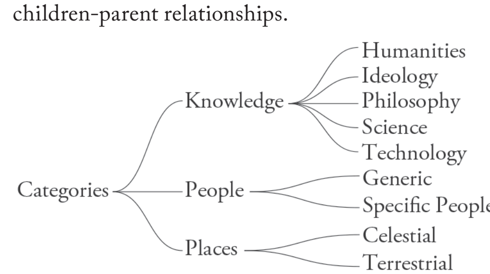
Nodes may be positioned in space according to their attribute values (e.g., publication year or geolocation), the relationships between records in terms of similarity or distance between attribute values (e.g., number of shared words), or a combination of both. Many different network layout algorithms exist (see **page 58, Network Visualization Types**). Node size or color value is used to encode additional quantitative variables, whereas shape, color hue, or pattern commonly represent qualitative data variables.

Edges may be weighted or unweighted, directed or undirected, symmetric (reciprocated), or asymmetric. They may be of different types and can have additional qualitative or quantitative variables. Edge shape, color hue, and pattern (e.g., dotted or dashed) may be used to encode qualitative data variables and directedness; size (line width) and color value are used to encode quantitative variables. In some cases, record relations are used exclusively to compute the position of nodes, though they are not directly visualized.

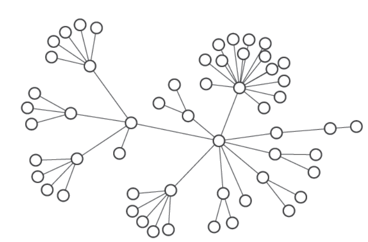
Trees

Tree layouts are used to display file directories, family trees, tournament trees, or classification hierarchies.

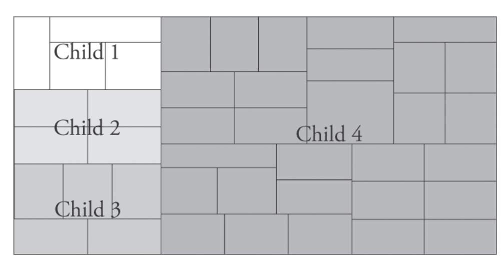
Trees may be represented as indented lists, dendrograms, node-link trees (see the tree view below and beneath that a force-directed layout of a different tree), circle packings (see **page 62, Enclosure Trees**), or treemaps (see below and **page 62**). The latter two use spatial nesting to represent children-parent relationships.



Tree View



Force-Directed Layout



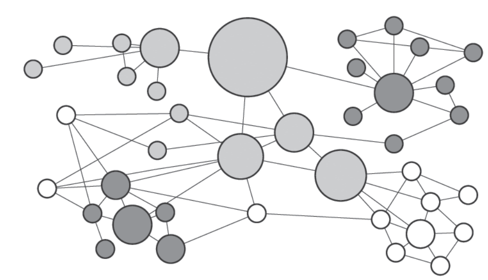
Treemap

Networks

Networks may depict social networks, concept or topic maps, food webs, or the interconnectivity of Internet servers, among others.

Networks may be represented by one-dimensional arc graphs (see below), tabular matrix diagrams, bimodal network visualizations, axis-based linear network layouts (see **page 63, Hive Graph**), or force-directed layouts (see below). The first four types use well-defined reference systems (e.g., nodes may be sorted by a node attribute), which means the axes are labeled and their value range is known. Force-directed layouts have no axes. In fact, the layout is unaffected by mirroring or rotation; only the distances between pairs of nodes matter (see also **page 62, Network Visualization Types**).

Arc Graph



Force-Directed Layout

Graphic Symbol Types

Cartographers, semioticians, statisticians, and others have worked to enumerate the basic, primary graphic symbols used to convey information on a map or visualization. The key types discussed here comprise geometric symbols (e.g., point, line, area equaling a bounded polygon, surface, volume), linguistic symbols (e.g., text and numerals), and pictorial symbols (e.g., images and statistical glyphs). They can designate location, convey qualitative and quantitative information, highlight specific information, help to identify and differentiate, depict form, represent multiple data variables via miniature graphs, or serve as enclosures. Each symbol has different graphic variables that can be used to encode additional quantitative and qualitative data; see the subsequent spreads in **Graphic Variable Types** (**page 34**) and the examples in the *Graphic Variable Types versus Graphic Symbol Types* table (**pages 36–39**).

In the final analysis, a drawing simply is no longer a drawing, no matter how self-sufficient its execution may be. It is a symbol, and the more profoundly the imaginary lines of projection meet higher dimensions, the better.
Paul Klee

Framework

Graphic symbols (also called geometric elements or geometric forms) are small graphic representations that are used to represent data records in a visualization. They encode different data variables via graphic variable types (**page 34**) such as spatial position, size, or color.

Different approaches to identifying and naming graphic symbol types are shown in the table below. The original titles are given in italics. Jacques Bertin's pioneering *Semiology of Graphics* identified and used three "Geometric Elements:" point, line, and area. Cartographer Alan MacEachren adopted Bertin's framework and successfully used it to explain how geospatial maps work.

Robert Harris expanded the set by adding volume and pictorial graphic symbol types to what he called "Symbol Types," of which his book *Information Graphics: A Comprehensive Illustrated Reference* provides detailed descriptions and numerous examples. He cleanly distinguishes two types of points: geometric and pictorial.

As part of his *Morphological Elements of Visual Language*, Robert E. Horn distinguishes three general types of graphic symbols: shapes, words, and images. He further lists different subtypes for each, as words can be "single words, phrases, sentences, [or] blocks of text." Horn distinguishes four types of shapes: point, line, abstract shape, and space between shapes. The latter type is not shown in the table below as it appears to be redundant when designing data visualizations—given the spatial position and visual encoding (e.g., size, of two graphic symbols, their distance can be computed).

Yuri Engelhardt—in his comparison and "translation" of numerous, discipline-specific approaches by key authors ranging from Edward Tufte, Jacques Bertin, and Stuart Card to Alan MacEachren and George Lakoff—identified what he called the "universal 'ingredients' of visual representations," consisting of (1) meaningful spaces—roughly equivalent with visualization types (**page 30**), (2) 'Visual Objects,' listed in the table below, and (3) visual properties (see **page 34**, **Graphic Variable Types**). Three of his visual objects were omitted from the table below, as they do not encode data variables: container—referring to the outer boundaries of a visualization; grid—used to improve readability of data values; and mark—used to highlight specific values.

In *The Grammar of Graphics*, Leland Wilkinson argued for the five "Geometric Forms" that include surface symbols but not linguistic and pictorial graphic symbol types.

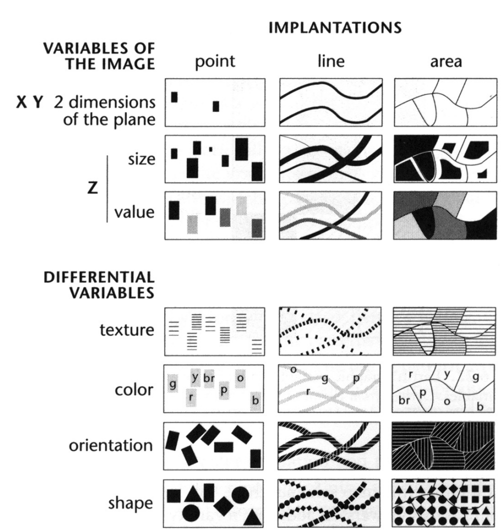
The final set of graphic symbol types that are used in this *Atlas* is given in the rightmost column of the table. Three general types of graphic symbols are distinguished: geometric, linguistic, and pictorial. Descriptions and examples are given on the opposite page. For more examples, see the *Graphic Variable Types versus Graphic Symbol Types* table (**pages 36–39**).

Graphic Symbol Types

Bertin, 1967 *Geometric Elements*	MacEachren, 1995 *Geometric Elements*	Harris, 1996 *Symbol Types*	Horn, 1998 *Morphological Elements of Visual Language*	Engelhardt, 2002 *Visual Objects*	Wilkinson, 2005 *Geometric Forms*	Börner, 2014 *Graphic Symbol Types*	
point	point	point: geometric	shapes: point	node	point	geometric	point
line	line	line	shapes: line	link, line locator	line	geometric	line
area	area	area	shapes: abstract shape	bar	area	geometric	area
				surface locator	surface	geometric	surface
		volume			solid	geometric	volume
			words: single words, phrases, sentences, blocks of text	label, character		linguistic	text, numerals, punctuation marks
		point: pictorial	images: objects in world	pictorial element		pictorial	images, icons, statistical glyphs

Instantiation

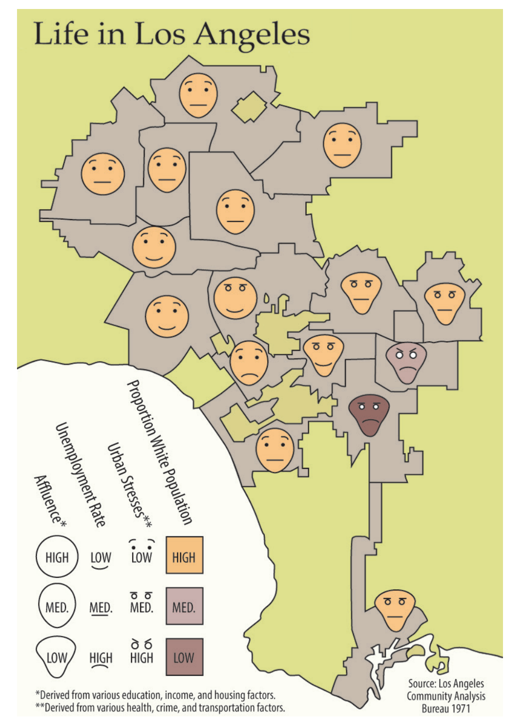
Each graphic symbol type has diverse attribute values, so-called graphic variable types (**page 34**), that can be used to encode additional data attribute values. MacEachren's instantiations (which he calls IMPLANTATIONS) of different graphic variable types for different symbol types are shown in the figure below. Columns represent the three graphic symbol types: **point**, **line**, and **area**. The rows represent different VARIABLES OF THE IMAGE such as **position**, **size**, and **value**; and DIFFERENTIAL VARIABLES such as **texture**, **color**, **orientation**, and **shape**. Instantiations of a substantially expanded set of graphic variable types and graphic symbol types can be found in the *Graphic Variable Types versus Graphic Symbol Types* table on **pages 36–39**.



Combinations

Multiple graphic symbol types can be combined. For example, a node in a network may be represented by a labeled circle—a combination of an area geometric symbol and a text linguistic symbol (see **page 53**, *The Debt Quake in the Eurozone*). Statistical glyphs such as pie charts can be combined with geometric lines to render the nodes and edges in a network graph (see **page 67**, *U.S. Healthcare Reform*). Gestalt principles such as proximity, continuity/connectedness, common region, or combinations thereof can be applied to visually interlink different graphic symbol types.

Analogously, different graphic variable types can be applied and combined. Exemplarily shown below is a geospatial map of Los Angeles with an overlay of statistical glyphs that resemble faces.



These so-called Chernoff faces (**page 33**) map different data variables onto facial expressions, such as head shape, mouth type, and eye type. Furthermore, a face can have different graphic variable types, here color hues. Each of the three facial expressions and the graphic variable type has three possible values resulting in 3 x 3 x 3 x 3 = 81 possible combinations.

Geometric Symbols

Geometric symbols are distinguished by the dimensionality they establish, involving points, lines, areas, surfaces, and volumes. They are easy to draw (to position, size, and color-code) using existing tools and easy to read and compare—even at very small sizes. Multiple symbols of the same type can be used, for example, to show data density. Disadvantages include the limited selection of symbols and the need to explain their usage in the legend.

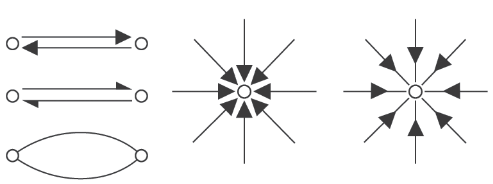
In traditional geometry, a point is nothing but a location in space, lacking size and any other visual encoding; a line has a given position and length but no width or color. In compliance with prior work that aims to define graphic symbol types and developed with the intention of using geometric symbols for encoding data variables, the framework presented here assumes that point, line, area, surface, and volume symbols can be size-, color-, and shape-coded; see examples in the *Graphic Variable Types versus Graphic Symbol Types* table (**pages 36–39**).

Points

A point symbol is commonly used to visualize data records that exist at a discrete point location, such as a postal address. Points are used to specify location and show density distribution. Additional data variables are encoded using graphic variable types (**page 34**).

Lines

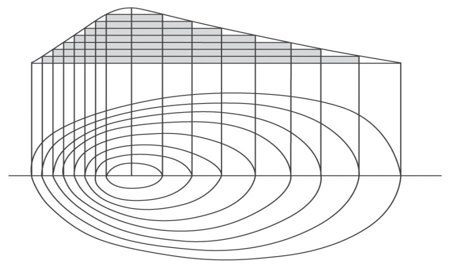
A line connects two points. Line symbols are applied to denote linear geographic objects such as streets, rivers, boundary lines, or geological faults as well as phenomena in motion, such as hurricane and tornado paths or ocean currents. Lines may be directed, as in network graph visualizations (**page 62**). This is commonly indicated through the use of arrows or line shapes, which may be read clockwise from source to target mode (see examples, below-left). When using arrowheads as line endpoints, nodes that have many incoming links may appear to have a larger size (see below-middle); this can be resolved by placing arrowheads at a distance from the destination nodes (see below-right).



Lines might be weighted and labeled and can be bundled (see **page 62**, **Network Visualization Types**).

Areas

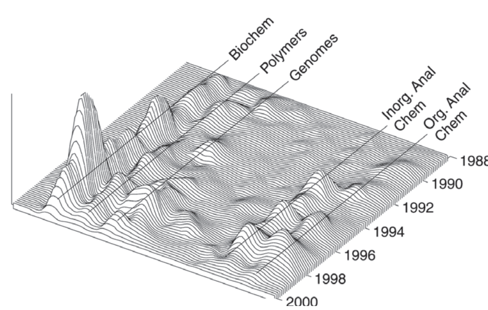
Area symbols include bounded polygons, used to represent country or state boundaries (see the *U.S. Map of Contiguous States* on **page 24**). Another type of area symbol is an isoline (also called an isopleth or isogram), which on a base map interconnects points that have the same value (e.g., places on a map registering the same amount or a given ratio of any given phenomenon, such as elevation or population density). More widely spaced lines indicate a gentle slope, whereas dense lines denote a steep slope (see below).



Areas can be qualitatively differentiated using graphic variables to show nominal differences (e.g., ethnic maps or vegetation and soil maps). Areas can be quantitatively differentiated using the choropleth, isoline, or cartogram methods (see **page 54**, **Geospatial Visualization Types**).

Surfaces

Surface symbols, such as surface plots, have a three-dimensional surface that connects a set of data points. An example is a surface plot of topics over time (see below and **page 58**, **Crossmap**).



Volumes

Volume symbols are also three-dimensional. They are used in bar graphs or **Stepped Relief Maps** (**page 54**). Examples include *In the Shadow of Foreclosures* (**page 53**) and *On Words—Concordance* (**page 57**).

Linguistic Symbols

Linguistic symbols, such as letters, numbers, or punctuation marks are widely used. One example is the use of chemical elements (i.e., symbols of the periodic table, such as Cu, Au, Zn, or Fe) or abbreviations for country names (e.g., CA, DE, FR, or US per the ISO two-letter code system), which most viewers would understand without the need of a legend (see **page 31**, *Country Codes of the World*).

The exact location and size of linguistic symbols tends to vary due to the differences in letter shapes; their proper placement can be aided by rendering linguistic symbols inside of geometric symbols (see **page 53**, *The Debt Quake in the Eurozone*).

Either serif (e.g., Cambria) or sans serif (e.g., Arial) typefaces may be used. Some type fonts (e.g., Caslon) have uppercase and lowercase numbers (see example below).

Cambria	A b c d e f 0 1 2 3 4 5 6 7
Arial	A b c d e f 0 1 2 3 4 5 6 7
Caslon	A b c d e f 0 1 2 3 4 5 6 7 8 9

A typeface can be proportional, containing glyphs of varying widths (e.g., Garamond), or monospaced, using a single standard width for all glyphs in the font (e.g., Courier). Using all uppercase letters in labels should be avoided, as reading all capitals takes more time than reading sentence-case text.

Garamond	Proportional Typeface
Courier	`Monospace Typeface`

Font families refer to groups of related fonts that vary in weight, orientation, and width, but not in design. For example, Times New Roman, *Times New Roman—Italic*, and **Times New Roman—Bold** are all members of the Times font family.

Fonts can be printed in different sizes or colors; formatted with underlining, outlining, or shading; and set in superscript or subscript positions (see **page 34**, **Graphic Variable Types**).

Some type fonts render pictorial symbols that can encode additional data variables via (partially) filled shapes (see examples below).

Webdings 

Conventional Dingbats ❏ ■ ① ❶ ❍ ● ➪ ➡ ✧ ✦

Text can be left or right aligned, centered, or justified. Numbers are commonly aligned vertically on the decimal point.

Pictorial Symbols

A pictorial symbol (also called an iconic symbol, sign, or pictogram) is an arbitrary or conventional mark used to represent complex notions, such as quantities, qualities, or relations. Pictorial symbols can be concrete reproductions of the objects they represent; specialized, such as statistical glyphs or the symbols used in weather maps; or abstract, composed of different geometric shapes. They can be shown from different perspectives, such as in profile or as a top view, and are typically positioned according to their centroid or mass point.

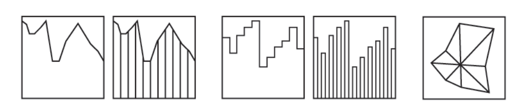
Images and Icons

Image symbols are drawn reproductions of the objects they represent. They tend to be easy to read and to understand. The larger their size and geometric complexity, the fewer that can be placed in a visualization.

Icons are specialized symbols designed to convey specific meaning. They are an efficient means of encoding information. Typically, a legend must be presented to signify what any given icon represents.

Statistical Glyphs

Statistical glyphs (also called miniature graphs) have no titles, labels, check marks, or grid lines (see **page 46**, **Statistical Visualization Types**). Examples are line graphs, profile graphs, histograms, bar graphs, and radar graphs (see below, from left to right), each of which can be used to encode 10 to 20 quantitative or qualitative variables. Glyphs are frequently used in combinations (**page 66**), small multiples (**pages 66**, **67**, and **69**), or matrix displays (**page 66**).



Two types of statistical glyphs that are more widely known and used are sparklines and Chernoff faces (see **page 46**, **Statistical Visualization Types**).

Sparklines are numerically dense, word-sized graphs that show data variation over time (see the miniature bar graph below).

Number of Tweets

Number of Downloads

Chernoff faces are pictorial symbols that map multiple data variables to facial expressions (see **page 32**, *Life in Los Angeles*). Most humans know how to read faces and can read data encoded in Chernoff faces.

Graphic Variable Types

The geometric, linguistic, and pictorial graphic symbol types discussed in the previous spread can be used to encode additional data variables using graphic variables. The key approaches to defining and grouping graphic variable types are compared here in an attempt to provide a Rosetta stone for interlinking different approaches and theories and to arrive at a set of well-defined and exemplified key types (see opposite page). Psychological results on the accuracy of graphic variable types are also discussed, as they help to guide the selection of graphic variable types that can be easily read and distinguished.

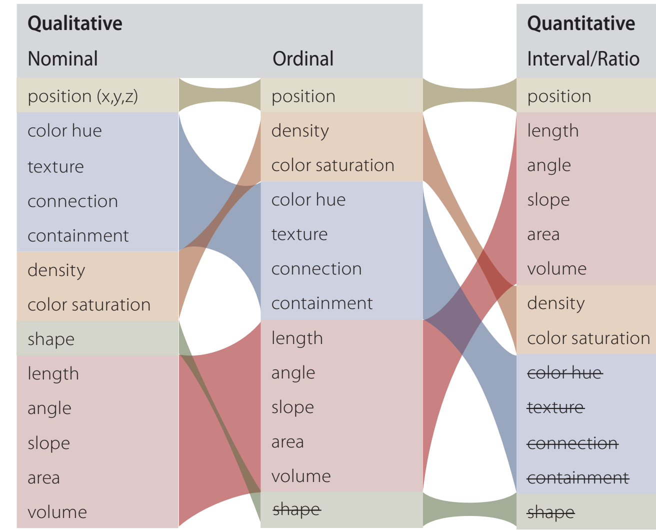
All the pieces are here—huge amounts of information, a great need to clearly and accurately display them, and the physical means for doing so. What is lacking is a deep understanding of how best to do it.
Howard Wainer

Framework

Various theories exist on how to identify and name graphic variable types. The table below lists the approaches proposed by leading experts. Cartographer and theorist Jacques Bertin conducted extensive landmark work as early as 1967 and later expanded on that research. Cartographer Alan MacEachren adopted Bertin's variable types, but also added clarity, which may be broken down into three subcomponents: crispness, resolution, and transparency. Crispness is the ability to selectively and dynamically filter for edges, fill, or both. Resolution defines how sharp or pixilated a given object appears and can be used to represent uncertainty in data. In his book *Visual Language*, political scientist Robert E. Horn added illumination and motion. In *The Grammar of Graphics*, Leland Wilkinson developed a complete grammar for the design of graphs and tables of graphs and introduced a hierarchical organizational schema for graphic variable types with superclasses form, color, texture, and optics. The rightmost column of the table shows the graphic variable types adopted in this *Atlas*. Spatial and retinal properties are distinguished. The former equate positioning in a three-dimensional space. The latter can be subdivided into form, color, texture, and optics—groupings that conform to Wilkinson's superclasses. Extending Wilkinson's schema, this table includes motion. It also adds a number of new graphic variable types, namely those that are preattentively processed even before attention is fully focused on it (e.g., curvature, angle, closure, stereoscopic depth) and those that conform to Gestalt principles (e.g., motion variables).

Graphic Variable Types

Bertin, 1967	Bertin, extended	MacEachren, 1995	Horn, 1998	Wilkinson, 2005	Börner, 2014		
location		location	location: in 2D or 3D	position	spatial	position	x y z
size (small vs. large)		size	size: area, thickness	form: size	retinal	form	size
shape (circle vs. triangle)		shape		form: shape	retinal	form	shape
orientation (up vs. down)		orientation	orientation	form: rotation	retinal	form	rotation curvature angle closure
color value (light vs. dark red)		color value	color: value	color: brightness	retinal	color	value
color hue (red vs. blue)		color hue	color: hue	color: hue	retinal	color	hue
	color intensity (saturated vs. dull)	color saturation		color: saturation	retinal	color	saturation
texture (spaced vs. dense)	pattern arrangement (striped vs. crossed)	texture	texture	texture: granularity, pattern, orientation	retinal	texture	spacing granularity pattern orientation gradient
		crispness		optics: blur	retinal	optics	blur
		resolution			retinal	optics	
		arrangement			retinal	optics	
	transparency	transparency		transparency	retinal	optics	transparency
			illumination		retinal	optics	shading stereoscopic depth
	animated: speed		motion		retinal	motion	speed velocity
	animated: rhythm				retinal	motion	rhythm



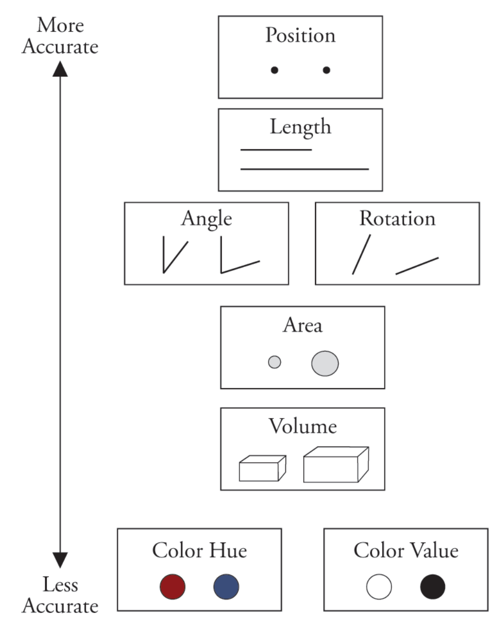
Combinations

In some cases, only one data variable is used to visually encode a graphic symbol, called a "univariate" symbol. Typically, multiple visual variables, or "multivariate" symbols, are mapped. The mapping of data variables to graphic symbols should be consistent per visualization. For instance, when data is identical, it should be consistently represented by the same chosen graphic symbol and its graphic variable encoding. Note that most attribute combinations are independent of each other (such as with shape and color hue); in some cases, combinations may be interdependent, such as when increases in symbol size conflict with position constraints (e.g., keeping all symbols on the canvas).

Perception Accuracy

In 1986, Jock D. Mackinlay published a ranking of perceptual tasks for different data scale types (**page 28**), as shown in the top-right figure. He ordered variables top-down according to how accurately humans perceive data at standard levels of measurement. The ranking was designed to help with the prioritization and matching of data scale types to graphic variable types. The six grayed-out graphical variable types are not relevant to the given data scale types. For all data scale types, **Position** is most accurately perceived. For **Nominal** data, **color hue** is second best. **Qualitative** data uses **density**; **Ordinal** data uses **length**.

Different studies have since been conducted to ascertain which graphic variable types most accurately convey quantitative data variables. William Cleveland and Robert McGill conducted a number of visual perception studies to determine what people can accurately decode. Robert Spence's visual summary of Cleveland and McGill's results is shown below. Note that only paired comparisons (e.g., **Position** versus **Length**) have been validated. Judging magnitudes differs from identifying outliers. The top of the image shows the tasks that are performed more accurately. A noticeable gap exists between the accuracy at which **Angle** or **Rotation** and **Area** can be judged. There is an even larger gap in accuracy when judging **Volume** and **Color Hue** or **Color Value**.



Descriptions and Examples

Spatial

Spatial position refers to the location of a record in a one- to three-dimensional space; see the **Spatial** rows in the *Graphic Variable Types versus Graphic Symbol Types* table (**pages 36–37**).

Retinal

Retinal variable types refer to all nonspatial properties; see the **Retinal** rows in the same table (**pages 36–39**).

Form

Form is defined as the visible shape or configuration of a graphic symbol.

Size refers to the scaling of graphic symbols and is commonly used to encode additional quantitative data variables, to attract attention, define importance, and support comparisons. Symbols can be size-coded by absolute data values, apparent magnitude values, or values that discriminate data ranges.

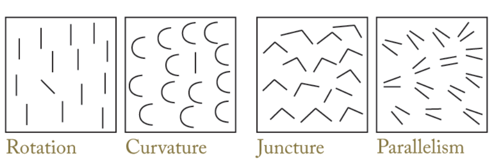
Shape comes in three basic types: geometric (e.g., triangles, squares, circles), natural (e.g., hands, trees, animals), and abstract (e.g., icons, glyphs). A legend must be provided to guide interpretation. Whenever possible, existing visual "grammar" systems should be used.

Rotation (also called angle or slope) refers to the orientation of graphical symbols (at any angle within the full rotation of 360 degrees, see below). It can be used to encode qualitative information (e.g., live, standing tree and dead, fallen tree, **page 37**) and quantitative information (e.g., clock face).

Curvature refers to the degree to which a graphic symbol is curved (see below).

Angle refers to the space between two intersecting graphic symbols at or close to the point at which they intersect. It is usually measured in degrees (see examples in the subsequent spread).

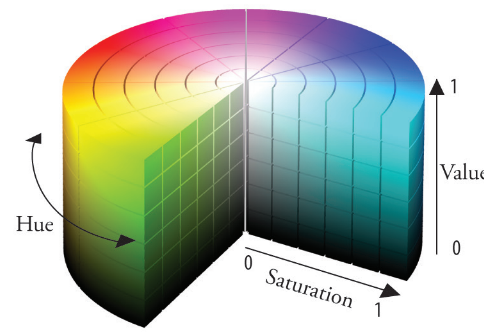
Closure is a graphic variable that indicates how much a circle or other geometric figure is closed.



All these six form attributes are preattentively processed; juncture and parallelism are not (see example above).

Color

The color of an object is determined by the measure of its value, hue, and the saturation of light being reflected from or emitted by it. An HSV (hue, saturation, value) color model is shown below.



Color is often used to convey importance or attract attention to specific symbols. It can help to alter the effects of camouflage (e.g., expose red cherries in a tree), develop an understanding of material properties (e.g., the condition of food or tools), and support comparisons. It can also be used to document nature (e.g., blue lakes in maps) and to generate or invoke emotions ranging from warm and active to cold and passive. Color is less effective in displaying how objects are positioned in space, how they are moving, or what their shapes are.

Value (also referred to as brightness, shade, tone, percent value, density, intensity, and luminance) relates to the amount of light coming from a source or being reflected by an object. It indicates how dark or light a color looks (see **page 36** for an example of a gradient that ranges from white to black). The ratio between the minimum and maximum brightness values in an image is also called a contrast ratio.

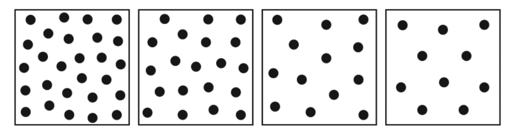
Hue (also called tint) refers to the dominant wavelength of a color stimulus. It is commonly used to represent qualitative data. However, if quantitative data (e.g., terrain heights) is being represented, the data should be carefully binned and a meaningful color sequence selected (e.g., blue lakes set against green forests or brown mountains set against the white of snow-covered mountaintops).

Saturation (also called intensity) refers to how much hue content is in the stimulus. Monochromatic hues are highly saturated. Completely desaturated colors constitute the grayscale, running from white to black, with all of the intermediate grays in between. More highly saturated (purer) colors appear in the foreground, whereas less saturated (duller) colors fade into the background.

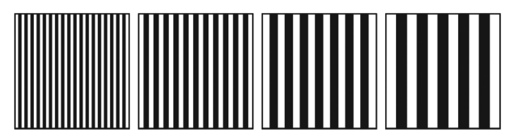
Texture

Texture relates to the surface or "look and feel" of an object. It adds depths and visual interest. Printed visualizations inherit the texture of the material on which they are printed. Those displayed onscreen have a designed texture that is made up of smaller graphic elements (lines, dots, shapes, etc.) set out in a consistent pattern. Texture properties comprise spacing, granularity, pattern, orientation, and gradient; these are explained and exemplified for different geometric symbol types on **pages 38–39**.

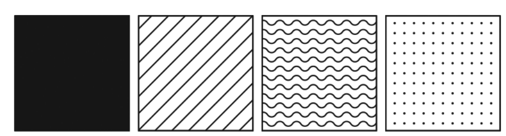
Spacing (also called density) refers to the amount of space between the graphic symbols that make up a texture (see below).



Granularity (also called coarseness) indicates the size of graphic symbols, while the ratio of figure to ground (or ratio of black symbols to white background) remains constant (see below).

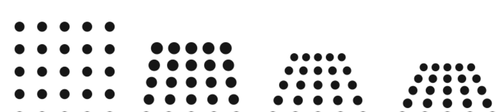


Pattern refers to the type of graphic symbols used (e.g., dots, lines, and solids as well as flags or data-generated symbols; see below). Textures with linear components (e.g., grids) are frequently used to reveal surface shapes. Background images (e.g., satellite images or aerial photographs) are used to provide context.



Orientation refers to the rotation or incline of graphic symbols. They may be perfectly horizontal or vertical, or diagonal at any angle within the full rotation of 360 degrees.

Gradient is used to indicate an increase or decrease in the magnitude of a property and also to show perspective (see below).



Optics

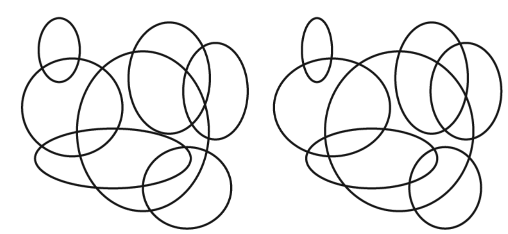
Optical properties can be used to indicate data uncertainty, deal with overlaps, emphasize structure, and attract attention.

Blur (also called crispness or resolution) is a measurement of discernable pixels. The fewer the pixels in any given visualization, the more blurred (or less clear) the image. Blur has been proposed by MacEachren as a means to depict data uncertainty.

Transparency (also called opacity or translucence) refers to the visibility of an object. Solid graphic symbols will stand out but may also overlap. Transparency can improve readability as it makes occlusions easier to detect.

Shading, related to illumination, refers to the darkened area or shape on a surface that is produced when a body comes between rays of light and that surface. It can be used to emphasize structure and to attract attention. It also helps to reinforce our perception of the location of light sources and objects. An even stronger effect is produced with motion (see discussion below). In fact, shadow motion can serve as a greater depth cue than a change in size due to perspective. Shadows are most effective when cast to a nearby surface. However, as shadows can interfere with other displayed information, they should be rendered with blurred edges.

Stereoscopic depth can be used to create or enhance the illusion of depth in a visualization. Two images are needed—one for each eye. The depth variance is encoded in the differences between the two views (see the example of intertwining rings below).



Motion

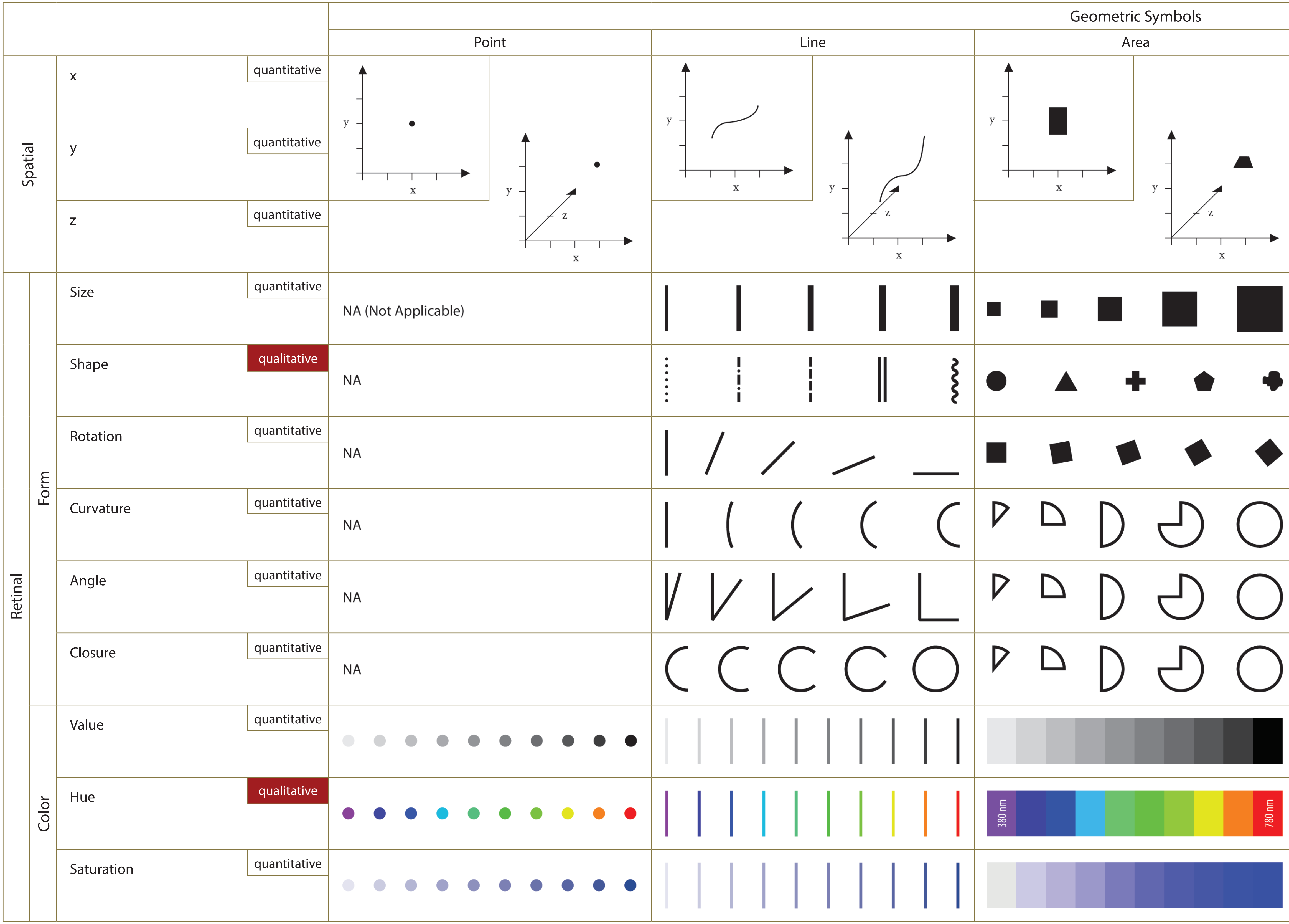
Graphic variable types that require moving objects are difficult to exemplify in print; yet they are highly effective in interactive visualizations.

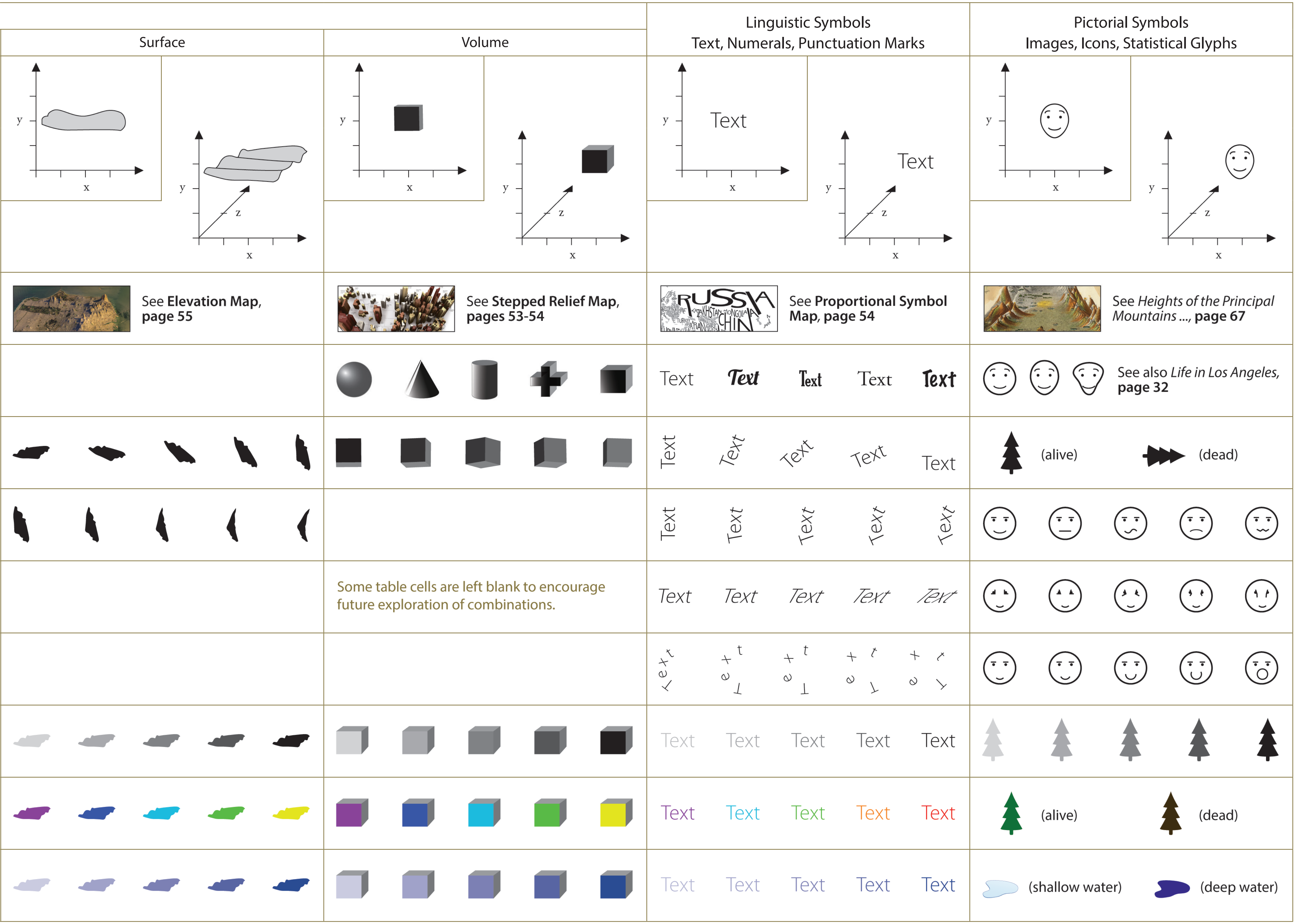
Speed refers to the rate at which a set of objects moves (but not the direction of movement).

Velocity is a vector quantity that captures the speed and direction of a set of moving objects.

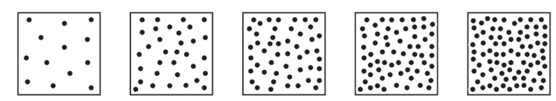
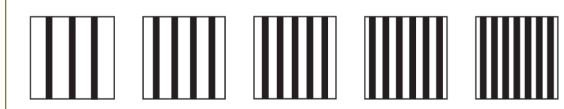
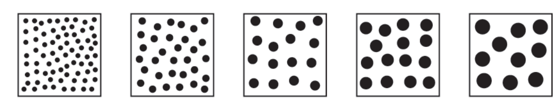
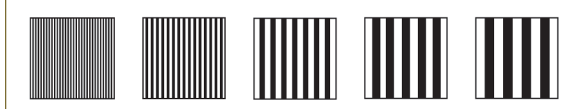
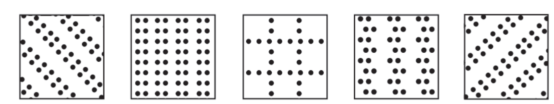
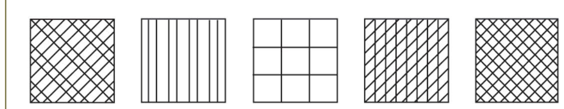
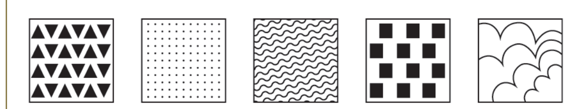
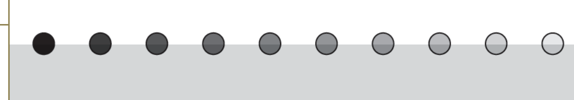
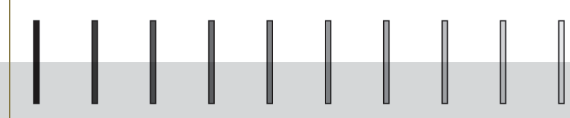
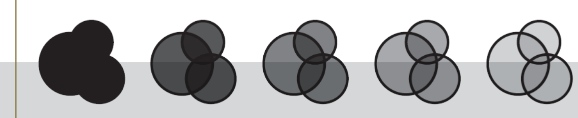
Rhythm (also called flicker) refers to regular, repeated pattern changes in spatial position or retinal variables. It is highly effective for attracting attention (e.g., to alert users of dangerous situations).

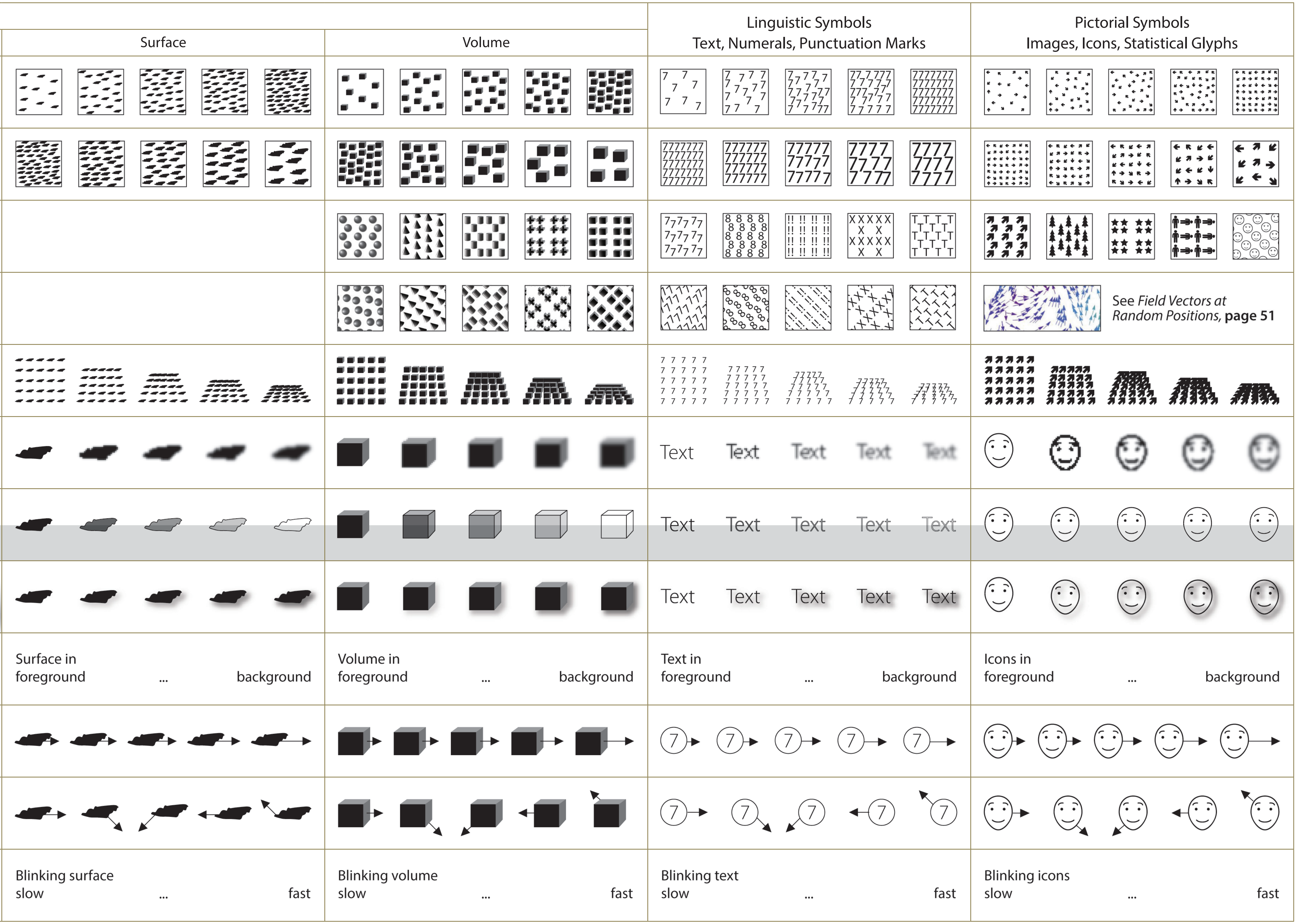
Graphic Variable Types Versus Graphic Symbol Types




Surface
Volume
Linguistic Symbols
Text, Numerals, Punctuation Marks
Pictorial Symbols
Images, Icons, Statistical Glyphs
See **Elevation Map, page 55**
See **Stepped Relief Map, pages 53-54**
See **Proportional Symbol Map, page 54**
See *Heights of the Principal Mountains ...,* **page 67**
See also *Life in Los Angeles,* **page 32**
(alive)
(dead)
Some table cells are left blank to encourage future exploration of combinations.
(alive)
(dead)
(shallow water)
(deep water)

Graphic Variable Types Versus Graphic Symbol Types (continued)

				Geometric Symbols		
				Point	Line	Area
Retinal	Texture	Spacing	quantitative			
Retinal	Texture	Granularity	quantitative			
Retinal	Texture	Pattern	qualitative			
Retinal	Texture	Orientation	quantitative	NA		
Retinal	Texture	Gradient	quantitative			
Retinal	Optics	Blur	quantitative			
Retinal	Optics	Transparency	quantitative			
Retinal	Optics	Shading	quantitative			
Retinal	Optics	Stereoscopic Depth	quantitative	Point in foreground ... background	Line in foreground ... background	Area in foreground ... background
Retinal	Motion	Speed	quantitative			
Retinal	Motion	Velocity	quantitative			
Retinal	Motion	Rhythm	quantitative	Blinking point slow ... fast	Blinking line slow ... fast	Blinking area slow ... fast


Linguistic Symbols
Text, Numerals, Punctuation Marks
Pictorial Symbols
Images, Icons, Statistical Glyphs
Surface
Volume
See *Field Vectors at Random Positions*, **page 51**
Text
Surface in foreground ... background
Volume in foreground ... background
Text in foreground ... background
Icons in foreground ... background
Blinking surface slow ... fast
Blinking volume slow ... fast
Blinking text slow ... fast
Blinking icons slow ... fast

User Needs Acquisition

Any good data analysis and visualization is driven by a deep care about the target user's needs. Users may wish to advance a theory (e.g., by testing a scientific hypothesis) or improve their daily decision making (e.g., by discovering which type of funding best supports an activity of their choice). In all cases, it is important to identify what keeps users up at night—for instance, determining the elements that may advance or thwart their careers, to ensure the final visualizations support the former while avoiding the latter. This spread details the first step in the **Needs-Driven Workflow Design, page 24**). Starting with a listing of general considerations, it reviews key user types and tasks, discusses the user needs acquisition process, and concludes with general advice on how to interlink user needs to the visualization framework discussed on **pages 24–39**. The end goal is to design insightful visualizations that truly match user needs and tasks and that would rank highly when validated using the criteria and methods discussed in **Validation and Interpretation** (page 72).

It is not the consumer's job to know what they want.
Steve Jobs

General Considerations

A complete and well-defined set of user needs and tasks is vastly important in the design of visualizations that are to make a true difference. The following discussion highlights the value of user needs in guiding analysis and visualization design. While exploring visualization production versus consumption, convention versus customization, and incremental prototyping and replication issues, the discussion also argues for the anticipation of change.

User Needs as Guides

There exist many data mining and visualization algorithms that can be used to render data into insights. Visualizations might be presented using static printouts or interactive displays. The problem-solving space created by diverse combinations of datasets, algorithms, and deployment options is large and complex. A detailed understanding of user needs helps navigate this space to select the best datasets, workflows, and deployment.

Datasets provided or demanded by users affect which data scale types are used (see the subsequent spread on data acquisition, and see also **page 28**). Reformulating user needs in terms of analysis types and levels (**page 5**) can help in identifying the most appropriate types of studies and selecting the best visualization types (**page 30**). Knowing which data variables are critical in which steps of the sense-making and decision-making process can help in selecting the most effective graphic symbol types (**page 32**) and graphic variable types (**page 34**).

Being informed about currently used (or potentially acceptable) hardware and software can help in designing effective human–computer interfaces (**page 70**) and interaction (**page 68**) that will be useful in daily practice.

Ultimately, each visualization design is an optimization of usability (i.e., effectiveness, efficiency, satisfaction, and accessibility), actability (in that it permits, promotes, or facilitates the performance of actions), productivity, enjoyability (also called pleasurability), and, last but not least, affordability.

Production Versus Consumption

Frequently, the visualization producer is different from the visualization user or consumer. Although the producer may have extensive knowledge on algorithms and tools, the consumer may be a domain expert with little or no expertise in visualization design. However, both types of expertise are needed to create the best visualization (i.e., to create effective workflows, adjust algorithm parameter values, improve mapping of data variables to visual variables, and select alternative data views). As the complexity of data, mining, and visualization designs increases, so too does the size of the teams that collectively produce and consume visualizations.

Convention Versus Customization

Conventional visualization designs generally become faster to make, cheaper, more reliable, and easier to use the more tested and widely used they are. Custom design solutions may prove necessary when new demands must be met, as long as the budget and allotted time support the development, evaluation, and user training required for them. Typically, standard solutions address general needs, whereas custom solutions address unique needs.

This is also true for visualization tools. Standard tools support the design of standard visualizations. Custom code is needed to render novel visualizations. Plug-and-play architectures (see **page 168, Plug-and-Play Macroscopes**) support the rapid development and dissemination of innovative custom code while making it easy to log, share, and rerun existing data analysis and visualization workflows. The best algorithmic and workflow solutions are born from solving specific, practical problems. Widespread adoption and refinement of solutions then leads to the creation of de facto standards.

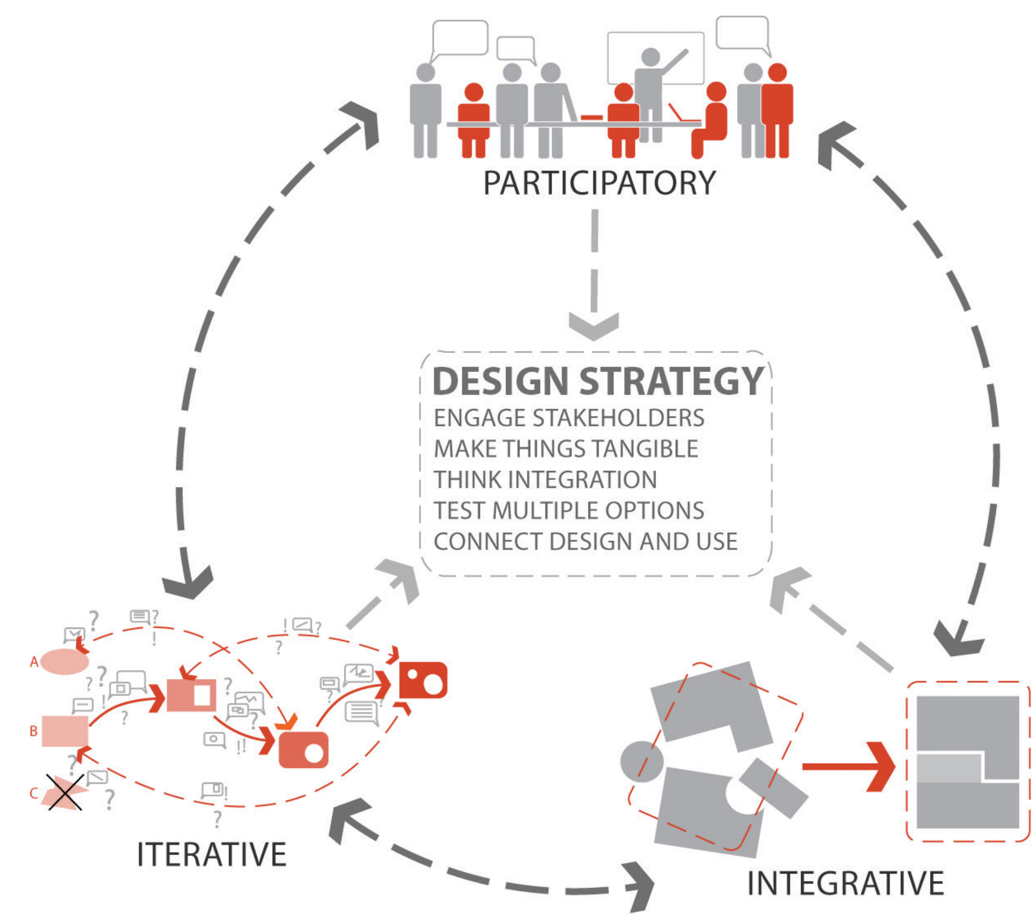
Iterative Prototyping and Replication

Most real-world applications require reliable and replicable workflows and effective, easy-to-use visualization design solutions. However, new data sets, algorithms, and tools are becoming available on a regular basis (see **page 168, Changes in the S&T Landscape**); novel workflows are invented every day; and very few standards exist.

To satisfy real-world demands, visualization designers and tool developers need to embrace the steadily increasing stream of data, algorithms, and tools; furthermore, they need to identify and standardize those algorithms and workflows that lead to superior results. Participatory, iterative prototyping is key (see figure below) and is accomplished through initial sketches (e.g., pencil drafts); early renderings of raw data, using existing tools, to see the data's coverage, patterns, and trends for the first time; the development of novel algorithms and workflows to optimize data analysis and visualization design; the comparison and validation of results (see **page 72, Validation and Interpretation**); and the detailed documentation and broad dissemination of validated visualizations, tools, and workflows. Integrative solutions that solve a problem holistically, with few or no trade-offs, are best.

Users and Needs

Deep knowledge about a user group and its decision-making process, the chosen subject matter, and the function a visualization is meant to serve must all be clear before the design process can begin. The ultimate goal is the detailed identification of user types, demographics, task types, conceptualization, work contexts, and priorities.



User Types
In establishing user types, one may ask: Is the visualization serving novices or experts, casual or power users? Will it be used by one individual (via a handheld device), a small group (using a larger display wall), or a large audience (in a theater-style setup)—face-to-face or online? Should the visualization therefore be understandable to a single individual, a small group, a larger organization, or an entire network of users?

Demographics
In researching demographics for any single user, one needs to acquire information on profession, location, gender, or age and to describe the user's range of abilities, accounting for vision, hearing, mobility, or cognitive impairments. One should also consider the user's level of technical and subject-matter expertise. Some user groups are text focused (lawyers) while others read three-dimensional structures easily (chemists). Finally, one must choose the visual language the user best understands for its function and/or content.

Task Types and Task Levels
In determining insight needs, one may ask: How do users currently do their work? What are their goals? What set of tasks supports these goals? One would then draw a diagram of the workflow and describe it in the words that users would use. Finally, one should prioritize tasks based on criteria such as the importance of the goal to the organization and the frequency of task performance. See also **Task Types** (i.e., statistical to network analysis; **page 5**) and **Task Levels** (i.e., micro to macro; **page 5**) for guidance on the selection of data analysis and visualization types.

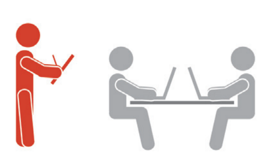
Conceptualization
When exploring conceptualization, one should ask: How do users conceptualize their work? What language do they use to describe what they do or what they wish to be able to do? What data do they use (see **page 28, Data Scale Types**), and which aspects of that data are most important to their decision making? What types of questions (see **page 26, Insight Need Types**) do they need to answer, at what level of abstraction, and with what accuracy? How do they define or evaluate the success of a visualization?

Work Environment
In researching work environment, one should aim to describe scenarios or circumstances in which the visualizations may be used. One should note environmental challenges, such as poor lighting or noise, as well as any technical challenges, including screen size or Internet bandwidth. One should also determine what hardware and (browser) software, monitors, and screen resolutions one's audience uses (see **page 70, Human-Computer Interface**). Do the users work mostly online or offline? Does their process result in printouts, or is it paperless and interactive? Is it static or dynamically evolving? What do they love or hate about the tools they use currently? Do they face any security restrictions?

Prioritization
Finally, one should aim to acquire information about priorities: Which pieces of information do users want first, second, and third? What information is indispensable in order to avoid disaster? How do users judge the result of their work or actions? What causes them to advance in their profession, or to get fired? The results of this type of analysis will guide the mapping of data variables to visual variables—high-priority features should be visually encoded using the most dominant visual representations (see **page 32, Graphic Symbol Types**, and **page 34, Graphic Variable Types**).

Needs Acquisition
Most users can easily propose quantitative changes to an existing practice, such as the need for faster response times, improved ease of use, and increased accuracy; yet few can envision qualitative new ways to navigate, manage, or make sense of data. They can, though, judge the potential value of novel visualizations—particularly if those visualizations show their own data in a new way. Access to detailed user and usage data, as well as to key stakeholders and leading experts (e.g., via participatory design), is necessary to characterize users and their tasks. This section reviews common methods, which may be applied independently or in combination. The methods differ in terms of cost, target population coverage, flexibility with regard to asking questions, respondents' willingness to participate, and response accuracy.

Interviews
In interviews, two or more people engage in a conversation, whereby a set of questions is asked by the interviewer to elicit answers from the participant(s). The researcher should listen carefully and patiently with an open mind—and also learn the necessary language, conceptualization, and metaphors—in order to design visualization and interfaces that match the users' needs and worldviews.



Observations
Using this method, experts observe target users in real-life situations to find out what they truly like or dislike, what they want or reject; what inspires them, and what confuses them; and what their ideal visualization (tool) might look like. Observations of how current products and services are used together with information on use context are particularly valuable.

Surveys
Surveys are a highly effective means of collecting quantitative information about products (such as visualizations) in a population. They can be conducted online or offline; using mail, email, phone, or online services; in face-to-face meetings; or through the use of questionnaires or focus groups (see below). Single-choice or multiple-choice questions might be asked. Likert scales (see **page 29, Ordinal Scale**) are widely used to scale responses. Open-ended, freeform questions may be acceptable, but results are harder to analyze.

Focus Groups
A small group of users is invited to join an interactive group setting and asked about their perceptions, opinions, beliefs, and attitudes toward a visualization product, service, concept, advertisement, or idea. Users are encouraged to talk with each other so that important patterns of interaction are revealed. For example, users might interpret a visualization differently; it is therefore important to be aware of each user's level of expertise and visualization literacy, and how the group's collective expertise can be best harnessed toward improving collaborative decision making by means of data visualizations.

Apprentice Model
Data mining and visualization experts are invited by users to serve as apprentices. They become intimately familiar with the work environment, key tasks and priorities, and what truly matters in users' daily decision making. The resulting knowledge is invaluable not only for the design of influential visualizations, but also for the introduction of qualitatively new conceptualizations and work practices into existing work environments.

Lead User Analysis
Users that face new needs months or years before the majority of a particular market segment encounters them are called lead users. Lead users benefit significantly by developing or otherwise obtaining a solution to those needs, making those users early adopters of new solutions (see Rogers's five types of adopters in *Atlas of Science*, **pages 58–59**). The methodology involves the identification of trends and general needs; seeking out lead people or organizations that are working on solutions to extreme versions of the general needs; and the identification and validation of potentially disruptive solutions.

Conjoint Analysis
This statistical technique can be used to understand the value of a limited number of product or service attributes. Ultimately, it aims to identify what combination of attributes is most influential on users' decision making. Via this methodology, users are asked to evaluate (e.g., select for purchase) a controlled set of potential products or services. (For example, a visualization may be static or interactive; black and white or colored; and shown on a different output device. Each variable has multiple attribute values. An output device may be a printout, a computer screen, or a handheld device.) An analysis of user preferences reveals the implicit evaluation of the individual attributes that make up a product or service.

User Mining and Modeling
As data on user demographics and behaviors (e.g., user profiles, purchasing data, website log data, or social network data) becomes available in digital form, data mining and visualization techniques can be applied to compute user preferences or reactions to new product offerings. The book and movie recommendation systems of Amazon and Netflix reflect but one way that customer behavior can be predicted and used to customize the visual display of information in order to increase sales. Other companies, such as Google and Facebook, mine massive amounts of news, social media, and other data to determine the reaction of customers to new product offerings. In S&T studies, diverse algorithms and approaches have been developed to understand if download counts, early citations, and other attributes can be used to predict the final citation count of a publication.

Data Acquisition

A tremendous increase in the number of papers, books, patents, experts, and funding has been seen over time, as shown in the *Atlas of Science* (see graphs on **pages 4–5**). Social media data sets such as blogs, tweets, and emails are becoming increasingly important for understanding S&T structures and dynamics (see **page 170, Data Monitoring and Analytics**). The *Atlas of Science* (**page 60**) explored data types, sizes, and formats; data quality and coverage; and data acquisition, preprocessing, augmentation, integration, and preservation. This spread discusses different variable types, formats, and aggregations, as well as the process of matching data variables to data analysis types (**page 5**) and to graphic variable types (**page 34**).

It is no longer enough to measure what we can—we need to measure what matters.
Robert Wells and Judith A. Whitworth

Data Variables

Naming conventions on **page 26** defined a data record as a N-tuple of data variables. The data variables may be qualitative or quantitative and have different data scale types (**page 28**). Data variables might exist in the original data or may be derived or computed from it. They may be dependent and independent as discussed below.

Original versus Derived Variables

Some data variables exist in the original data set (e.g., journal name and author address in publication data). Other data variables are derived (e.g., the author's address might be used to determine latitude and longitude data value pairs required to place a data record on a geospatial map). It is important to keep track of where each piece of data came from (also called data provenance) to ensure high data quality and to facilitate informed validation and interpretation (**page 72**).

Dependent versus Independent Variables

The variables used in an analysis or study can be divided into dependent variables, independent variables, and other variables. Dependent variables are expected to change whenever the independent variable is altered. Other variables such as covariates used to reduce the amount of variability might be recorded as well. There may be more than one variable of these three types. For example, the number of citations a paper acquires (here treated as a dependent variable) may depend on the number of authors, their reputations, and their geolocations (here treated as independent variables). During visualization, independent variables are commonly plotted on the horizontal x-axis; dependent variables are plotted on the vertical y-axis.

Data Format

Selecting or defining an appropriate data format is critical when acquiring and processing data. The selection of relevant data variables, together with their data scale types (**page 28**) and data formats, influences which analyses can be run and visualizations created. For example, when a paper becomes available, online or in print, the publication date can be recorded as either the full date or only the year. Different date formats may be used, and the chosen format must be documented (e.g., some U.S. foundations make their data sets available in the European date format: day, month, year). Values may be stated in different units (e.g., a salary may be stated in either U.S. or Canadian dollars). Author affiliations can be stated with or without explicit links to each author; the latter makes it impossible to geolocate all authors on a map, as correlations cannot be made between the authors' names and the addresses listed.

Data Aggregation

Aggregation of data (also called generalization or clustering) can be applied to optimize both data density and legibility (see also **page 52, Visual Generalization**). For example, temporal data can be aggregated by seconds, minutes, hours, days, weeks, months, years, decades, and so on. Geospatial data can be grouped by congressional district, ZIP code, county, state, country, or continent. Linguistic data, such as text characters, can be grouped into words, sentences, paragraphs, sections, chapters, books, and collections. Network data may feature individual nodes, subnetworks, or the entire network.

Matching Data Analysis Types

Data records have different data variables (e.g., publication title, year, and authors) that are each uniquely useful in different types of analysis (see **page 4, Systems Science Approach**). This is illustrated on the opposite page that displays a mélange of elements: commonly used data variables from the Web of Science publication database (top-right); a table with publication data in the Web of Science format, sorted by publication year (below); one paper in the table is highlighted in white and its cover page with all author names and journal title is shown (top-left); and different data views from the tabulated publications are displayed (at bottom). Data variables are grouped and color-coded by the types of questions they help to address.

Statistical Studies

When taking on a new data set, it is important to compute baseline statistics. That can be done by counting and plotting the annual numbers of records, unique authors, or citations; or by calculating distributions and correlations to ensure the data set has the desired coverage and quality (see **page 44, Statistical Studies**). As for the example on the opposite page, a table with unique journal titles and counts for number of papers or total **Times Cited** counts or a scatter plot (see **page 47**) of **Times Cited** counts and **Cited Reference Counts** might be computed.

Temporal Studies

If the user needs acquisition (**page 40**) identified the necessity for temporal analysis, or answering a **when** question, then the data must have one or more variables that represent time (see **page 48, Temporal Studies—"When"**). Time resolution is important. If a monthly resolution is necessary but only publication years are available for journal papers, then volume information can be used; it must be noted, however, that different journals publish different numbers of volumes per year. Bursts of activity can be identified (**page 48**) and plotted using horizontal bar graphs that show the beginning and end of a burst (i.e., the width of a bar represents the burst duration) and represent burst strengths by the height of the bar (see example on opposite page).

Geospatial Studies

If a geospatial question needs to be answered, there must be a way to geolocate records (see **page 52, Geospatial Studies—"Where"**). Address data can be used to identify latitude and longitude values; U.S. ZIP codes uniquely identify a geolocation. Again, resolution is important (e.g., if a U.S. congressional district needs to be identified, then a five-digit ZIP code will not suffice; the full nine-digit ZIP code is required to uniquely associate each ZIP code to exactly one district). The opposite page shows a world map with an overlay of proportionally area-sized circles that represent the number of lead authors per unique geolocation.

Topical Studies

If a topical or semantic question needs to be answered, there must be a way to determine the topical content of records. Text occurring in the title, abstract, keywords, full text, or subject category may be analyzed using linguistic techniques (see **page 56, Topical Studies—"What"**). Records can be clustered and labeled according to semantic similarity. Shown on the opposite page is an overlay of the tabulated publications on the *UCSD Map of Science and Classification System* (*TTURC NIH Funding Trends*, **page 65** and *Atlas of Science*, **page 13**).

Network Studies

If a network question needs to be answered, networks need to be extracted and analyzed (see **page 60, Network Studies—"With Whom"**). Relationships may exist between nodes of the same type (e.g., in unimodal coauthor networks or paper-citation networks; see example on opposite page) or may be of different types (e.g., in bimodal author–paper networks; see **page 63, Bimodal Graph**).

Matching Graphic Variable Types

When acquiring and formatting data, it is important to ensure that it corresponds to data scale types (**page 28**) that can be effectively mapped to graphic variable types (**page 34**).

Choosing the right type and number of data records and/or the appropriate level of aggregation is important. Showing too few records/classes will result in an information-poor visualization. Plotting too many may lead to visual clutter (e.g., graphic symbol occlusions, such as in "network hairballs"), making it difficult if not impossible to identify general trends and patterns. In addition, large numbers of classes may compromise legibility, as more classes require more graphic variables (e.g., colors) that become increasingly difficult to tell apart.

Ideally, there is a one-to-one mapping between the number of different data values (e.g., the number of classes) and the number of graphic variable values (e.g., color hues). However, in some circumstances, there may be a many-to-one mapping; that is, the number of qualitative data types may be larger than the number of distinct graphic variable values. For example, the map of the *Language Communities of Twitter* (see *Atlas of Forecasts*) shows language use in Twitter across Europe using more than 30 different colors that are optimized for maximum distinguishability.

Paper

OPEN ACCESS Freely available online PLoS one

Design and Update of a Classification System: The UCSD Map of Science

Katy Börner[1,2]*, Richard Klavans[3], Michael Patek[3], Angela M. Zoss[1], Joseph R. Biberstine[1], Robert P. Light[1], Vincent Larivière[1,4,5], Kevin W. Boyack[6]

1 Cyberinfrastructure for Network Science Center, School of Library and Information Science, Indiana University, Bloomington, Indiana, United States of America, **2** Royal Netherlands Academy of Arts and Sciences (KNAW), Amsterdam, The Netherlands, **3** SciTech Strategies, Inc., Berwyn, Pennsylvania, United States of America, **4** École de Bibliothéconomie et des Sciences de L'information, Université de Montréal, Montréal, Canada, **5** Observatoire des Sciences et des Technologies (OST), Centre Interuniversitaire de Recherche sur la Science et la Technologie (CIRST), Université du Québec à Montréal, Montréal, Canada, **6** SciTech Strategies, Inc., Albuquerque, New Mexico, United States of America

Abstract

Global maps of science can be used as a reference system to chart career trajectories, the location of emerging research frontiers, or the expertise profiles of institutes or nations. This paper details data preparation, analysis, and layout performed when designing and subsequently updating the UCSD map of science and classification system. The original classification and map use 7.2 million papers and their references from Elsevier's Scopus (about 15,000 source titles, 2001–2005) and Thomson Reuters' Web of Science (WoS) Science, Social Science, Arts & Humanities Citation Indexes (about 9,000 source titles, 2001–2004)–about 16,000 unique source titles. The updated map and classification adds six years (2005–2010) of WoS data and three years (2006–2008) from Scopus to the existing category structure–increasing the number of source titles to about 25,000. To our knowledge, this is the first time that a widely used map of science was updated. A comparison of the original 5-year and the new 10-year maps and classification system show (i) an increase in the total number of journals that can be mapped by 9,409 journals (social sciences had a 80% increase, humanities a 119% increase, medical (32%) and natural science (74%)), (ii) a simplification of the map by assigning all but five highly interdisciplinary journals to exactly one discipline, (iii) a more even distribution of journals over the 554 subdisciplines and 13 disciplines when calculating the coefficient of variation, and (iv) a better reflection of journal clusters when compared with paper-level citation data. When evaluating the map with a listing of desirable features for maps of science, the updated map is shown to have higher mapping accuracy, easier understandability as fewer journals are multiply classified, and higher usability for the generation of data overlays, among others.

Citation: Börner K, Klavans R, Patek M, Zoss AM, Biberstine JR, et al. (2012) Design and Update of a Classification System: The UCSD Map of Science. PLoS ONE 7(7): e39464. doi:10.1371/journal.pone.0039464

Editor: Neil R. Smalheiser, University of Illinois-Chicago, United States of America

Received February 7, 2012; **Accepted** May 21, 2012; **Published** July 12, 2012

Funding: The generation of the 2010 UCSD map of science and classification system is funded in part by the James S. McDonnell Foundation, the National Science Foundation under award SBE-0738111, and the National Institutes of Health under awards U24RR029822 and U01GM098959. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: K. Börner is a PLoS ONE Editorial Board member. Several authors are employed by a commercial company, SciTech Strategies, Inc. This does not alter the authors' adherence to all the PLoS ONE policies on sharing data and materials.

* E-mail: katy@indiana.edu

Introduction

Cartographic maps of physical places have guided humankind's explorations for centuries. In addition to supporting navigation, these maps are used to record national boundaries or mineral resources, to show flows of trade activity, or to communicate areas of political unrest. Science maps of abstract semantic spaces aim to serve today's explorers navigating the world of knowledge. These maps are generated through a scientific analysis of large-scale scholarly datasets in an effort to extract, connect, and make sense of the bits and pieces of knowledge they contain [1,2]. Science maps can be used to gain overviews of "all-of-science" or of a specific subdiscipline. Science maps in combination with a mapping process for new datasets can be used to visually depict and compare data overlays, e.g., of funding vs. publication data [3]. Science maps can help identify major research areas, experts, institutions, collections, grants, papers, journals, and ideas in a domain of interest. They can show homogeneity vs. heterogeneity, cause and effect, and relative speed of progress. They allow us to track the emergence, evolution, and disappearance of topics and help to identify the most promising areas of research.

1.1 Related Work

Reviews of science mapping efforts up until 2007 show more than 200 different maps [4]. The number, diversity, and sophistication of science mapping efforts has increased enormously since then due to the availability of scholarly data in digital format, algorithm development, and an increase in computing power, see Mapping Science exhibit maps (http://scimaps.org). Each science map depicts an abstract high-dimensional space using different datasets, reference systems, and graphic designs. Very few maps depict all major disciplines of scholarly activity–these are also called global maps of science [5]. Some of these maps are drawn by hand while others are computer generated. Some sketch the

PLoS ONE | www.plosone.org 1 July 2012 | Volume 7 | Issue 7 | e39464

Exemplary Web of Science Data Variables

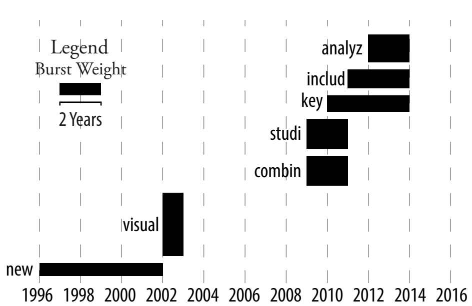
Grouped by type of study in which they are commonly used.

Statistical Variables	*Temporal Variables*	*Geospatial Variables*	*Topical Variables*	*Network Variables*
Cited Reference Count	Cited Year	City of Publisher	Abstract	Authors
Number of Pages	Publication Date	Conference Location	Funding Text	Cited Patent
Times Cited	Publication Year	Publisher Address	Journal Title (Full)	Cited References
		Reprint Address	New ISI Keywords	Editors
		Research Addresses	Original Keywords	Funding Agency & Grant No.
			Research Field	Publisher
			Subject Category	
			Title	

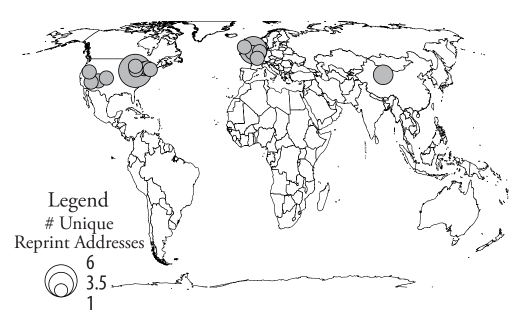
Times Cited	*Publication Year*	*Research Addresses*	*Journal Title (Full)*	*Title*	*Authors*
6	2013	[Mazloumian, Amin; Helbing...	SCIENTIFIC REPORTS	Global Multi-Level Analysis of the 'Scientific...	Mazloumian, A|Helbing...
7	2012	[Borner, Katy; Milojevic...	MODELS OF SCIENCE DYNAMICS...	An Introduction to Modeling Science: Basic...	Borner, K|Boyack...
5	2012	[Borner, Katy; Zoss, Angela...	PLOS ONE	Design and Update of a Classification...	Borner, K|Klavans...
51	2011	[Wagner, Caroline S...	JOURNAL OF INFORMETRICS	Approaches to understanding and...	Wagner, CS|Roessner...
15	2011	[Boyack, Kevin W.] SciTech...	PLOS ONE	Clustering More than Two Million Biomedical...	Boyack, KW|Newman...
13	2011	[Borner, Katy] Indiana...	COMMUNICATIONS OF THE...	Plug-and-Play Macroscopes	Borner, K
5	2011	[Guo, Hanning] Dalian Univ...	SCIENTOMETRICS	Mixed-indicators model for identifying...	Guo, HN|Weingart...
20	2010	[Falk-Krzesinski, Holly J....	CTS-CLINICAL AND...	Advancing the Science of Team Science...	Falk-Krzesinski...
16	2010	[Borner, Katy] Indiana Univ...	SCIENCE TRANSLATIONAL...	A Multi-Level Systems Perspective for the...	Borner, K|Contractor...
9	2010	[Borner, Katy; Huang...	SCIENTOMETRICS	Rete-netzwerk-red: analyzing and...	Borner, K|Huang...
17	2009	[Boyack, Kevin W.] Sandia...	SCIENTOMETRICS	Mapping the structure and evolution of...	Boyack, KW|Borner...
14	2009	[Borner, Katy] Indiana Univ...	JOURNAL OF INFORMETRICS	Visual conceptualizations and ...	Borner, K|Scharnhorst...
6	2009	Indiana Univ, Sch Lib & ...	SCIENTOMETRICS	The Scholarly Database and its ...	LaRowe, G|Ambre...
69	2008	[Mons, Barend; van s...	GENOME BIOLOGY	Calling on a million minds for community...	Mons, B|Ashburner...
55	2007	Indiana Univ, Sch Lib & ...	ANNUAL REVIEW OF...	Network science	Borner, K|Sanyal...
26	2007	Indiana Univ, Sch Lib & ...	COMPLEXITY	Analyzing and visualizing the semantic...	Holloway, T|Bozicevic...
81	2006	Arizona State Univ, Sch...	GLOBAL ENVIRONMENTAL...	Scholarly networks on resilience, vulner...	Janssen, MA|Schoon...
19	2006	Indiana Univ, Sch Lib & ...	SCIENTOMETRICS	Mapping the diffusion of scholarly knowle...	Borner, K|Penumarthy...
214	2005	Sandia Natl Labs...	SCIENTOMETRICS	Mapping the backbone of science	Boyack, KW|Klavans...
50	2005	Indiana Univ, SLIS...	COMPLEXITY	Studying the emerging global brain...	Borner, K|Dall'Asta...
9	2005	Indiana Univ, Dept...	ANIMAL BEHAVIOUR	Trends in animal behaviour research (1968...	Ord, TJ|Martins...
88	2004	Indiana Univ, Sch Lib & ...	PROCEEDINGS OF THE NATIONAL	The simultaneous evolution of author and...	Borner, K|Maru...
74	2004	Indiana Univ, Dept Psychol...	PROCEEDINGS OF THE NATIONAL	Mapping knowledge domains	Shiffrin, RM|Borner...
35	2004	Indiana Univ, Sch Lib & ...	PROCEEDINGS OF THE NATIONAL	Mapping topics and topic bursts in PNAS...	Mane, KK|Borner...
224	2003	Indiana Univ, Bloomington...	ANNUAL REVIEW OF...	Visualizing knowledge domains...	Borner, K|Chen...
41	2003	Sandia Natl Labs...	JOURNAL OF THE AMERICAN...	Indicator-assisted evaluation and funding...	Boyack, KW|Borner...
7	2002	Indiana Univ, Sch Lib &...	VISUAL INTERFACES TO DIGITAL ...	Visual interfaces to digital libraries:	Borner, K|Chen...
16	1996	HTWK Leipzig, FB	ADVANCES IN CASE-BASED...	Structural similarity and adaptation	Borner, K|Pippig...

Different Data Views

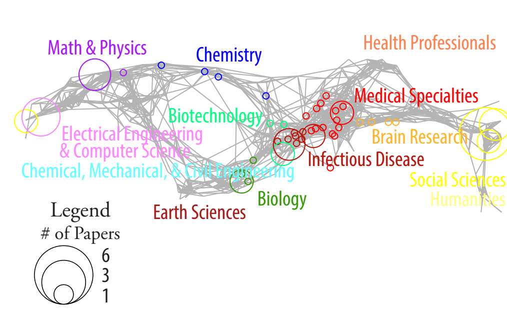
Bursts of Terms in Abstracts (see **page 48, Temporal Studies—"When"**)



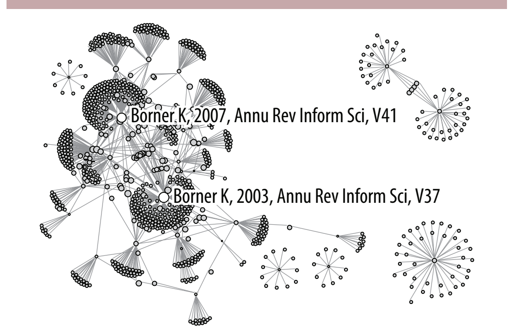
Geospatial Locations of Lead Authors (see **page 52, Geospatial Studies—"Where"**)



Topical Locations of Papers (see **page 56, Topical Studies—"What"**)



Paper-Citation Network (see **page 60, Network Studies—"With Whom"**)



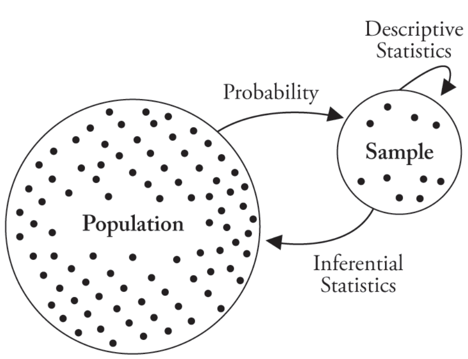
Statistical Studies

The field of statistics focuses on appropriate ways to collect, codify, analyze, and interpret numerical information. Standard analyses comprise data summaries, differences, averages, ratios, and distributions. This spread reviews common statistical analyses and presents exemplary visualizations on the opposite page. Special focus is given to insight need types (**page 26**), such as comparisons, correlations, distributions, and trends, together with sample analyses and visualizations that are particularly relevant for the study of S&T.

We all know that Americans love their statistics—in sport, obviously. And in finance too.
Evan Davis

Exploratory Versus Confirmatory

John W. Tukey made the important distinction between exploratory data analysis and confirmatory data analysis, believing that much statistical methodology placed too great an emphasis on the latter. Whereas confirmatory analysis aims to summarize data sets by computing their main characteristics, exploratory analysis uses statistical modeling and inference to predict data characteristics (see *Atlas of Forecasts*). The figure below shows the set of all data (called **Population**) and a limited subset (called **Sample**) of the data sampled from the population. Statistics can be used to determine which sample size is needed to answer a given question using **Probability** theory, to generate **Descriptive Statistics** for a **Sample**, and to use it to run **Inferential Statistics** to make generalizations from a **Sample** to a **Population**. The discussion ahead features key measures and approaches that are commonly used to describe, organize, and summarize the main characteristics of data.



Central Tendency Measures

These measures calculate the "center" around which data is distributed. The *mean* equals the arithmetic average, calculated by adding up all the values in a data set and then dividing that sum by the number of values in the data set. It is best for symmetric distributions without outliers but is less meaningful for scale-free distributions such as those characterizing the number of citations per paper or number of collaborators per author. The *median*, or mid-value, equals the middle value in an odd number of values and the average of the middle two points for an even number of values: That is, half of the data values are above the median, whereas the other half are below. It is commonly applied for skewed distributions or data with outliers. The *mode* is the most frequent value in a series. For example, in the sequence of values {1, 1, 2, 3, 5, 6}, the mean is 3, the median 2.5, and the mode is 1.

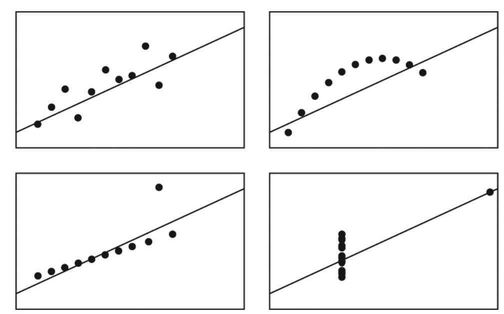
Data Distributions

Many real-valued random variables (e.g., highway distances between cities, **page 60**) cluster around a single mean value, and their distribution can be approximated by a bell-shaped continuous probability density function (also called normal or *Gaussian* distribution). Its cumulative distribution function is S-shaped (see adoption of innovations graph on **page 58** in *Atlas of Science*).

The distribution of a set of data is important as it affects data-sampling decisions and measurement of result confidence. *Variation* measures describe the "data spread" around the mean or expectation μ of a distribution. The standard deviation σ measures the amount of variation from μ. In the Gaussian normal distribution graph, roughly 68 percent of the observations (in the population) lie within one standard deviation of μ; about 95 percent lie within two σ; and 99.7 percent lie within three σ.

Many real-world scholarly data sets can be represented by a scale-free network whose degree distribution follows a power law (also called *Pareto* distribution), at least asymptotically. Examples are the number of citations per paper or scholar, the number of collaborators per scholar, and income earned or profits made (see **page 46, Statistical Visualization Types**, for plot; and **page 60, Network Studies—"With Whom,"** for details and additional examples, particularly the degree distribution of street versus airplane networks). The number of citations that papers attract over time (typically, many initially and then fewer over time) can best be approximated by a *Weibull* distribution.

Data sets that have similar statistical properties can look rather different when graphed. For example, Francis J. Anscombe created four data sets, each with 11 x-y value pairs, that have almost identical statistical properties: The mean of the *x* values is 9.0, the mean of the *y* values is 7.5, and there are nearly identical variances, correlations, and regression lines. However, when plotted they show a simple linear relationship (top-left in below figure), a curvilinear relationship (top-right), a linear relationship with one outlier (bottom-left), and a non-linear relationship with one outlier (bottom-right).

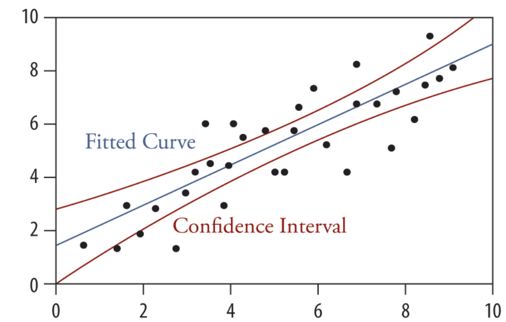


Curve Fitting

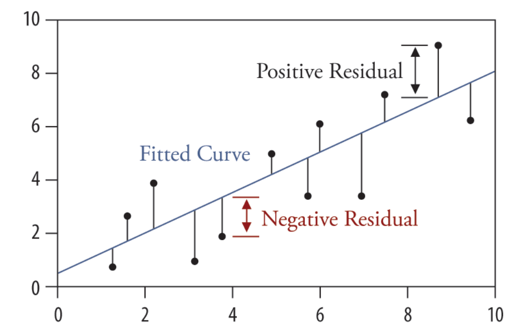
In any given data series, *curve fitting* (also called smoothing or regression) is a process that determines and superimposes a curve or surface that most closely approximates the data. The resulting analytic description of the data can be used to identify trends in the data; determine the types of relationship or correlation between variables (e.g., linear versus exponential); calculate the degree of variation of data points from a theoretical or expected curve; determine if data points vary randomly, uniformly, or otherwise from a theoretical or expected curve; or project future values.

Multiple types of curves can be used to fit the same data; then the type that most closely approximates the data and/or best fits the process that generated the data should be selected. Selecting the proper curve is particularly important when making projections; compare linear versus polynomial fit in the graphs in **Regressions, page 73**. Although the linear fit suggests continuous growth, the polynomial fit indicates a potential downturn in sales.

A *confidence interval* (also called a confidence band) describes the region in which the fitted curve would lie given a specific degree of confidence (e.g., 90 or 95 percent) if the entire family of data (i.e., the population) could be observed (see below).



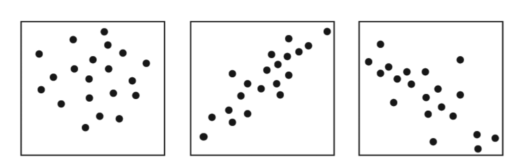
Residuals (also called fitting errors) are the distances between observed data points and the fitted curve. They are commonly plotted for the dependent variable and may be either positive or negative (see example below).



Statistical error (also called disturbance) is the amount by which an observation from a randomly chosen sample differs from its expected value (i.e., the whole population). An example is the difference between the age of each man in a sample and in the unobservable population mean.

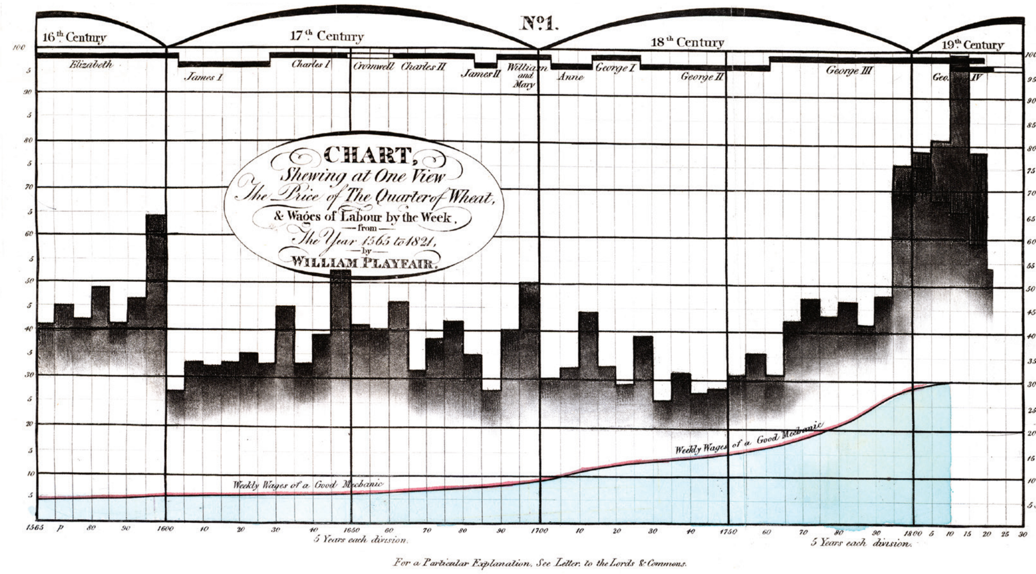
Correlations

A correlation is a mutual relationship or connection between two or more things (see also **page 47, Statistical Visualization Types**). A scatter plot with a "shotgun blast" pattern, or an alignment of points that is close to either the horizontal or vertical axis, indicates very low correlation (see below-left). If data points fall along a straight line, then a high degree of correlation exists—with a positive correlation if the high and low values of the two variables tend to coincide (below-middle) and a negative correlation if low values of one variable coincide with high values of the other variable (below-right). See subsequent spread for additional examples.



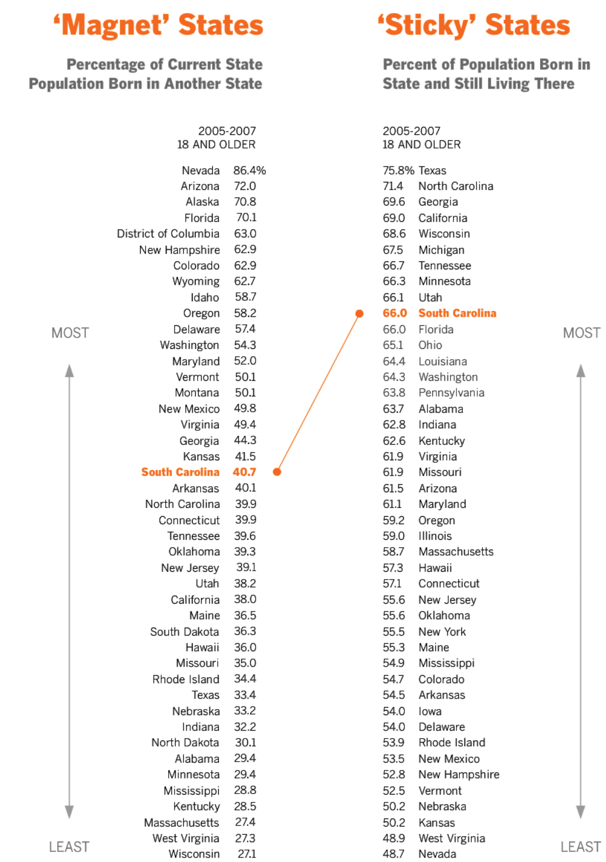
Wheat Prices Versus Wages

William Playfair (1759–1823), a Scottish engineer and political economist, was among the first who thought to use data not only to inform, but also to persuade, and even to campaign for causes. He developed line graphs to show changes in economic indicators (e.g., national debt, imports, exports) over time and across countries; comparative bar charts to show relations of discrete series; and pie charts and circle diagrams to show part–whole relations. This graph shows the weekly wages of a good mechanic as a red line, the price of a quarter of wheat as black-shaded bars, and the reigns of monarchs (displayed along the top) for the years 1565 to 1821. Major changes in wheat prices, the affordability of wheat, and the slowing increase in buying power are all clearly demonstrated here.



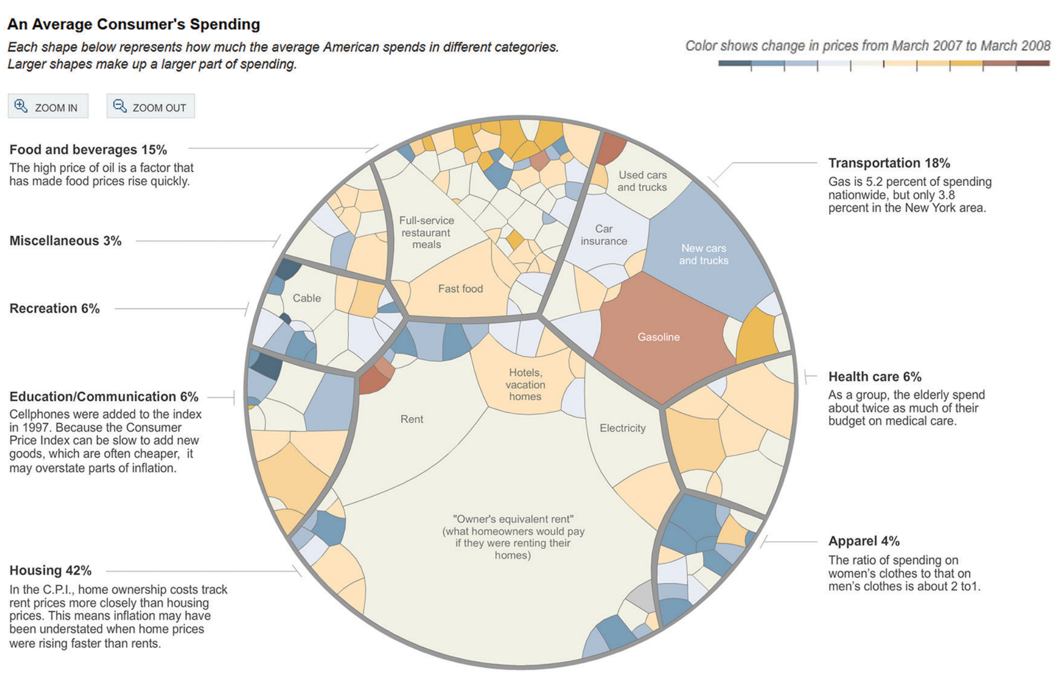
Magnet States Versus Sticky States

An American mobility study conducted by the Pew Research Center aims to answer which states are losing and gaining population and who moves from what state to what other state. Shown here are the top "magnet" states (left), which attract residents from other states, and the top "sticky" states (right), in which a high percentage of the native population still resides in the state. When an online user hovers his mouse cursor over a state for one list, its ranking in the other list also appears. See **page 178, References & Credits** for link to interactive visualization.



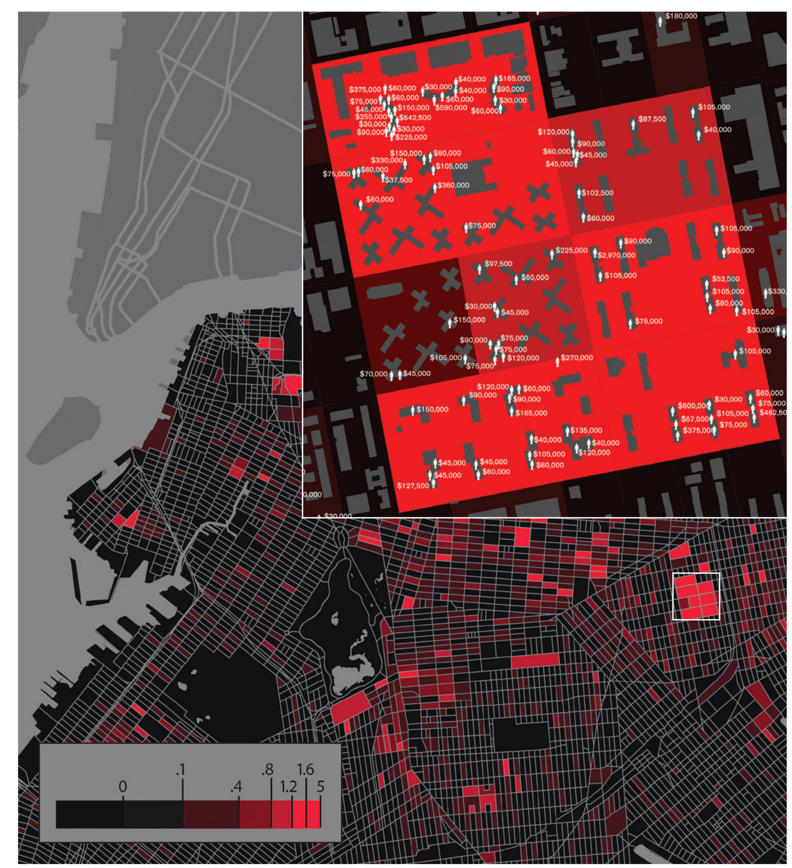
All of Inflation's Little Parts

This circular treemap from the Bureau of Labor Statistics, as published in *The New York Times*, shows the 200 product categories that are used to calculate the Consumer Price Index. The area size of each product category corresponds to an estimate of what the average American spends. Area color indicates price changes between March 2007 and March 2008.



Prison Expenditures for Brooklyn, New York City

The Spatial Information Design Lab at Columbia University studies the geography of incarceration. It aims to shift attention from punishment and rehabilitation to the conditions of neglected urban spaces. Investments into urban spaces from which prisoners often come and to which most return seem preferable to high prison expenditures (indicated by bright red in this map).



Statistical Visualization Types

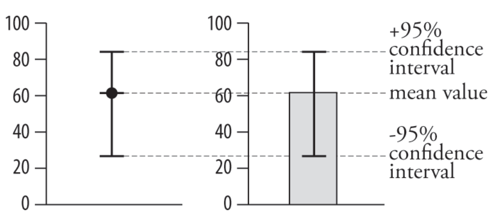
This spread discusses data visualization types that were specifically developed to depict statistical results. Some concisely encode several dimensions of data into a simple glyph such as a pictorial symbol that can be perceived as a single perceptual unit. Others graph data to satisfy key insight need types, such as comparisons or the identification and communication of correlations or distributions. For good measure, books by John W. Tukey, William S. Cleveland, Robert Harris, or Stephen Few may be consulted for detailed explanations and additional examples.

Glyphs

Different pictorial symbols have been developed to display key statistical features in a compact manner. Among them are error bars, box-and-whisker symbols, and sparklines, each of which is discussed ahead. For a discussion of Chernoff faces and other glyphs, see **Graphic Symbol Types** (**page 32**).

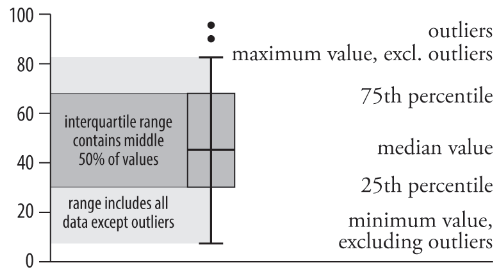
Error Bar

The error or uncertainty in a reported measurement can be depicted using error bars that may represent one standard deviation of uncertainty or a certain confidence interval (e.g., a 95 percent interval). The exact error measure used needs to be stated explicitly in the graph or legend. Error bars can be used with different graphic symbol types (see point and bar graph symbols in the figure below). They can help to determine whether differences are statistically significant, or they can suggest the goodness of fit of a given function.



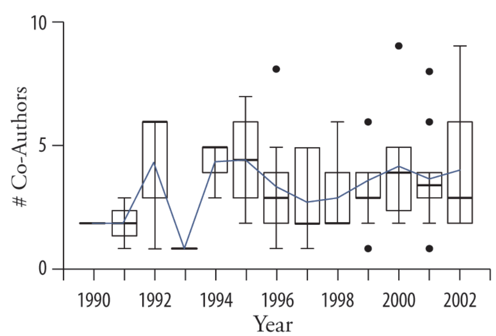
Box-and-Whisker Symbol

This symbol (also called a percentile plot, box diagram, box plot, or box-and-whisker plot) was introduced by John W. Tukey. It represents the key values, symmetry, and skewness of a data set using a rectangular box symbol with lines (whiskers) extending from both ends (see next column, top). The box is centered on the **median** data value (50th percentile). Both the median and mean may be denoted by a line. The box's ends designate the **25th** and **75th percentiles** of the data set (i.e., the range



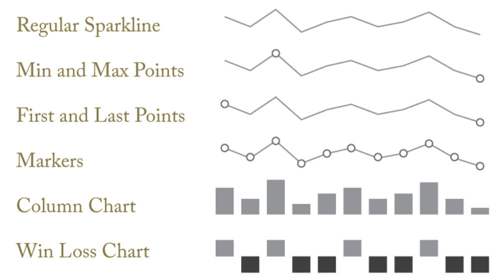
of quantitative values in which 50 percent of all data records fall). Whiskers typically start at the fifth or tenth percentiles and end at the 90th or 95th percentiles. **Outliers** (data points beyond the whisker ends) are denoted by dots.

Box symbols can be run vertically or horizontally. Data from multiple data distributions can be grouped. To ease comparison, they can be connected by a line that passes through the median or mean value. As an example, the graph below shows a box-and-whisker symbol for each publication year to indicate the number of coauthors for one scholar over 13 years. The bold horizontal line denotes the median; the blue line interconnects the means for each year; dots denote outliers.



Sparkline

Introduced by Edward Tufte, sparklines are numerically dense, word-sized glyphs that show data variation over time. They have a starting point and an endpoint; show data variation in between; typically highlight the minimum and maximum values; may show missing data; and may emphasize the area under the curve (see different types below).



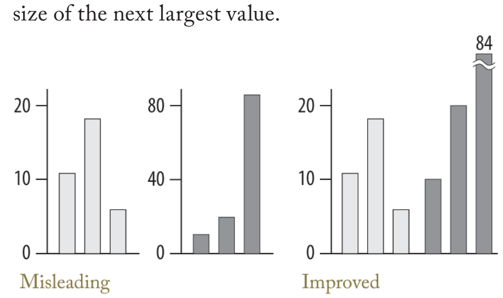
This shows the fluctuations of the Dow Jones industrial average over the course of February and March 2012, with the dramatic dip (red dot) indicating the March 16 panic surrounding the Fukushima Daiichi nuclear disaster in Japan.

Graphs

Among other uses, graphs can support comparisons and depict correlations, distributions, and trends. Exemplary visualizations that address these four insight need types (**page 26**) are discussed ahead (see additional examples in the previous spread).

Comparisons

The grouping or close proximity of data visualizations (e.g., the side-by-side or back-to-back placement of graphs, which is common in population pyramids; see **page 27**, **Comparison**) makes it easier to compare data sets. Consistent axes should be used to support such comparisons (see the **Misleading** and **Improved** examples below). However, broken bars and axes should be used sparingly to avoid misinterpretation (see **page 73**, **Distortions**). Broken bars like those below should only be used if the outlier is not a key part of the visualization and is at least three times or more the size of the next largest value.

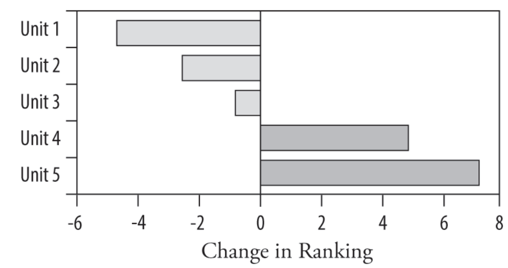


Line Graph

A line graph plots quantitative data as a series of points that are connected by lines; see **Validation and Interpretation** (**page 73**) for examples.

Bar Graph

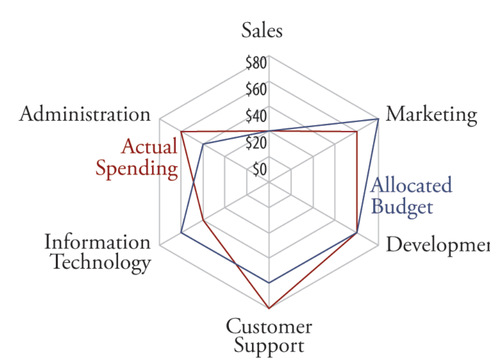
A bar graph (also called a column graph) displays quantitative data by means of a series of vertical or horizontal rectangles or bars. The bars commonly start at zero and end at the value of the data record that is represented by the bar. Positive and negative values can be plotted; records can be sorted by value (see graph below, which charts changes in ranking for five organizational units). Bars can also be stacked (see **page 50**). Typically, each bar represents one category (e.g., an institution, product, or year) and all the bars combined represent the data set. Bar graphs differ from histograms (see opposite page) that can be used to plot quantitative data, in that bars can be reordered and there are typically spaces between bars. A 100 percent stacked bar and column graph is known as a mosaic graph (**page 62**).



Radar Graph

The radar graph (also called a polygon graph, polar-area chart, radar plot, spider chart, or star chart) originated with André-Michel Guerry (1829) and Florence Nightingale (1858). It displays multivariate quantitative variables of different data records on axes starting from the same midpoint. The relative position and angle of the axes is typically uninformative, but axes can be reordered to minimize edge crossings. An example is the graph below, which plots **Allocated Budget** (in blue) versus **Actual Spending** (in red) in millions of dollars for an imaginary company; with the exception of **Administration** and **Customer Support,** most of the spending is on target.

Compare this type of graph with circular hive graphs (**page 63**) that use a radial coordinate system to display network data.



Parallel Coordinate Graph

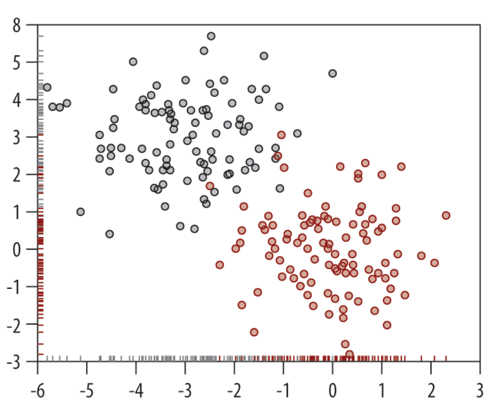
Like the radar graph, the parallel coordinate graph plots variables of different data records on quantitative or qualitative axes. In contrast, however, the axes are parallel to one another. The relative position of the axes is again uninformative, but axes can be reordered to minimize edge crossings. All values per record are then plotted on the given axes and interconnected by lines, the same as in a radar graph. Shown on **page 31** is a graph with four axes representing counts that may be used to judge a scholar's productivity over one year. Each of the six polylines represents all the values for a given scholar; one is highlighted in red.

Correlations

The correlation between two variables (e.g., age and weight) can be plotted on a graph using Cartesian coordinates (see examples on **page 44**, lower-right).

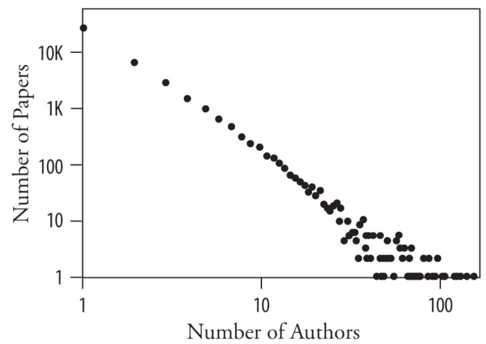
Scatter Plot

The scatter plot (also known as an x-y plot, or a dot, point, or symbol graph) displays quantitative information of data records, each represented by a graphical symbol type (e.g., a point; see **page 32**). The x-axis is commonly used to plot the independent variable, whereas the y-axis features the dependent variable. Scatter plots are used for investigating correlations between data variables or multiple data sets (which may be color-coded red and gray as in the example below).



Distributions

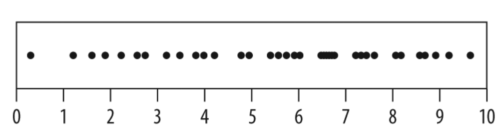
Frequency distribution graphs display every single data record. They can be used therefore to identify minimum and maximum values; how many data records have a certain value; if there are any outliers or unusual records; which value occurs most frequently; whether data records are distributed evenly or clustered; if the data is skewed, and if so in what direction; or how many clusters there are and where. An example is the log-log graph below, which shows the distribution of the number of authors per paper—revealing that the majority of papers have but one author (top-left dot), whereas very few have a large number of authors (lower-right dots).



Dot graphs, stripe graphs, stem-and-leaf graphs, and histograms are discussed ahead. Note that although histograms bin data (i.e., they do not display every single data record), they are widely used to display data distributions. For alternative visualizations such as tally charts and dot array charts, among other examples and detailed explanations, see Robert L. Harris's *Information Graphics: A Comprehensive Illustrated Reference*.

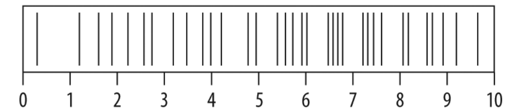
Dot Graph

In a dot graph (also called a dot chart), each data record is represented by a dot (see below). Jittering (i.e., adding small random values to the position values of graphic symbols so that the symbols are placed close to their real values) may be applied to avoid overlapping dots. Dots can encode other data variables via graphic variable types (**page 32**), such as the number of citations per publication (see **page 58, Crossmap**). Multiple dot graphs can be shown in one graph to compare the density of multiple data records or data sets. Data from multiple data distributions can be grouped.



Stripe Graph

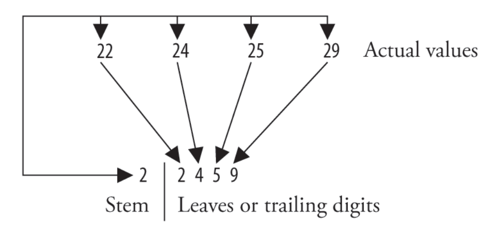
Stripe graphs (also called stripe charts) represent each data record with a stripe (see next column, top). The width of each stripe is uniform and does not encode a data variable. Jittering can be applied to avoid overlapping stripes. Stripes can encode other data variables via graphic variable types (**page 34**) such as topic areas (represented by color hue) of publication by a scholar. Multiple stripes can be shown to compare the density of multiple



data records/sets, and data from multiple data distributions can be grouped. Combining stripe graphs with a scatter plot results in what is known as a rug plot (see example in **Scatter Plot** section, left).

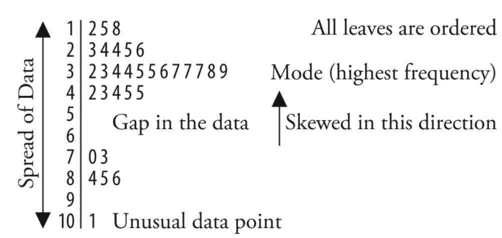
Stem and Leaf Graph

A stem-and-leaf graph (also called a stem-and-leaf chart or plot) shows the leading digit(s), or stem(s), of each data variable to the left of a vertical line. The other digits, or leaves, are plotted side-by-side to the right of the line, forming a sort of histogram. The stem-and-leaf graph for the numbers **22**, **24**, **25**, and **29** is shown below. Here, **2** is the stem or leading digit, whereas **2**, **4**, **5**, and **9** are the leaves or trailing digits placed to the right of the vertical line.



Stem-and-leaf graphs help to identify the spread of data; the mode, skew, and gaps; and also the outliers, or unusual data points. Leaves may be ordered by value to improve legibility and to help identify common values.

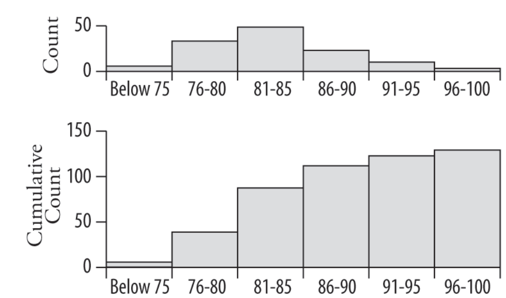
Shown below is an ordered stem-and-leaf graph of a larger data set ranging from 12 to 101. Twenty-five values are equal to or below 57, the so called midpoint. The other six values are larger. There are no values (i.e., there is a gap) between 45 and 70 and between 86 and 101; 101 is an unusual data point; and the data is skewed toward lower numbers.



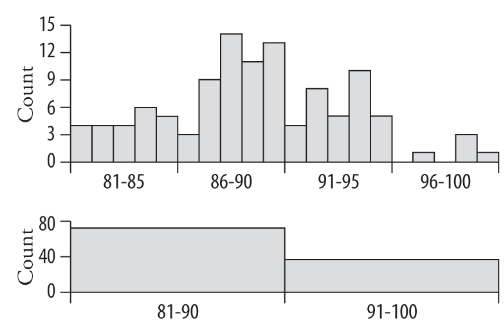
Multiple graphs can be displayed side-by-side or back-to-back (analogous to population pyramids, see **page 27, Comparison**) to support comparison.

Histogram

Histograms are used to show distributions of binned quantitative data variables. That is, they display bin-aggregated data rather than every single data record. Histograms can be plotted incrementally or cumulatively, as shown in the graphs below (top and bottom, respectively)—both of which feature the same distribution of class scores that were tabulated on **page 30, Tables**. A histogram differs from a **Bar Graph** (opposite page) in that there are no spaces between the bars, because there are no gaps between the bins. In addition, histogram bars cannot be reordered.



Histograms are widely used to visualize the distribution (shape, center, range, variation) of quantitative data variables. The bin size is important and exemplified in the histograms below, which show class scores for values above 80 and sample sizes 1 (top) and 10 (bottom).



Bilateral histograms (also called two-way histograms or paired bar graphs) are used to compare two frequency distributions back-to-back. They are called age pyramids or population pyramids when the binning is based on age. Typically, age intervals are plotted vertically, whereas the number of males or females per age interval is given horizontally (see figure and discussion on **page 27, Comparison**).

Bilateral histograms can be rendered as back-to-back stem-and-leaf charts, thus providing additional details on the data distribution.

Temporal Studies—"When"

Temporal analysis and visualization techniques are developed and applied to answer "when" questions. They aim to identify patterns, trends, bursts, or seasonality in a sequence of observations. This spread reviews major temporal analysis types and presents exemplary visualizations on the opposite page. Additional analysis types and visualizations that involve a temporal aspect are discussed in **Studying Dynamics** (**page 64**). Opportunities and challenges when analyzing and visualizing real-time data streams are discussed in **Data Monitoring and Analytics** (**page 170**) and **Real-Time Data Visualization** (**page 172**).

The price of anything is the amount of life you exchange for it.
Henry David Thoreau

Data Preprocessing

A time series is a sequence of events or observations that are ordered in one dimension: time. It can be continuous (i.e., there is an observation at every instant of time) or discrete (i.e., observations exist at regularly or irregularly spaced intervals). Time scales may be very short (i.e., events occur in a millisecond) or very long (i.e., events are recorded over years or centuries). Time-indexed information may be static (e.g., when analyzing historic data) or dynamically evolving (e.g., email or news data streams).

Resolution and Aggregation

Time may be given in milliseconds, seconds, minutes, hours, days, weeks, months, quarters, years, decades, or centuries.

Aggregation

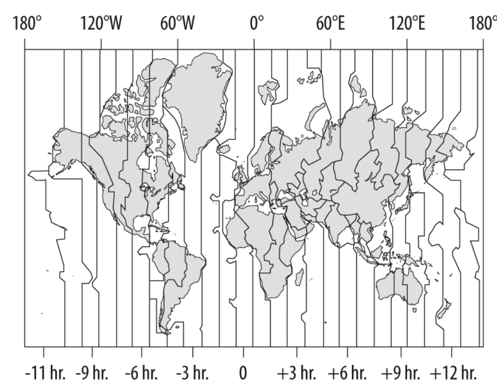
Temporal data can be clustered or aggregated by astronomical time (e.g., seconds, days, years) or cultural time (e.g., semesters or fiscal years). The higher the aggregation, the lower the resolution.

Time Zones

When dealing with global data, time zones need to be considered. Typically, time zones for individual countries match their international or state borders (see next column, top). The International Date Line roughly follows the 180° line of longitude, but zigzags around the borders of various countries. China, which crosses five time zones, has chosen to use only one.

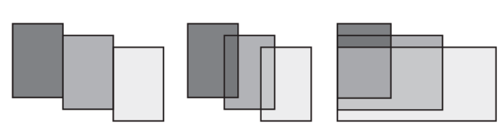
Outliers

It is important to identify and manage outliers (i.e., the minority of data points that are distant from most other data points). For example, if a web page gets "slashdotted" (when a popular website links to a smaller site), the massive increase in traffic is analogous to a denial-of-service attack. The resulting high number of download counts will affect all data statistics and may be better excluded from a general analysis and reported separately. Outliers may also be due to variability in the measurement, or they may indicate experimental error. Alternatively, outliers may be an indication that the "population" has a heavy-tailed distribution, as is true for paper-citation or coauthorship data sets in which few papers/authors have a large number of citations/coauthors, whereas most papers/authors have only a few (see **page 47**, **Distributions**).



Time Slicing

When generating animations, the data set needs to be divided into different time slices. Time frames can be **Disjoint**, **Overlapping**, or **Cumulative**; see below. When disjoint, every time-stamped row in the original table is in exactly one time slice. Overlapping means that selected rows are in multiple time slices. Cumulative signifies that every row in a time slice is in all later time slices.


Disjoint Overlapping Cumulative

Time frames can have either identical or different lengths (see *113 Years of Physical Review* in *Atlas of Science*, **page 159**). In some cases, the length of the time frames may be defined in a data-driven way based on key events (see Wikipedia edits in the *History Flow Visualization of the Wikipedia Entry "Abortion,"* in *Atlas of Science*, **page 125**) or career decision points, such as changes in affiliation. As a result, times with little or no activity are compressed to make space for the visual depiction of active phases.

Time frame duration is important. If it is too short, then too few data records are visible (in social network visualizations, many nodes may be unconnected). If it is too long, then too many data records appear (social network visualizations may be too dense to be legible).

In some cases, it is beneficial to align time slices with the calendar. For example, if the slices are weekly and aligned with the calendar, then the day the week starts is used to determine how they are aligned. In the United States, Canada, and Mexico, the week starts on Sunday; in much of the Middle East it starts on Saturday; and in most European countries it starts on Monday.

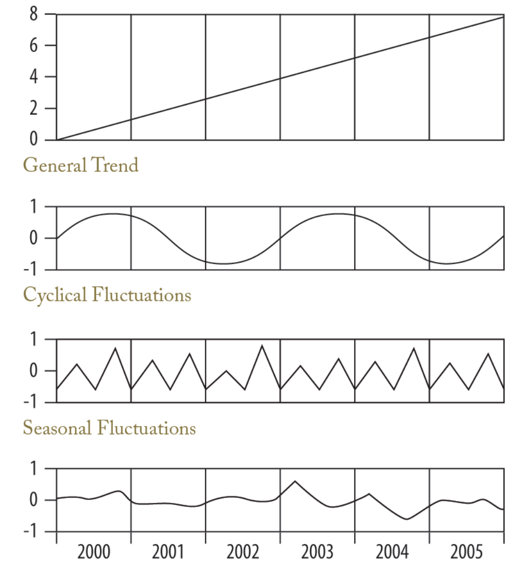
Trends

As shown (next column, right), a given time series can be decomposed into a **General Trend** component; a **Cyclical** component (e.g., day versus night, winter versus summer); a **Seasonal** component (e.g., summer vacation, holidays); and a **Random** component. Trends correspond to low frequency variations in the data. In order to identify trends, data can be smoothed using a so-called low-pass filter which reads the original time series and generates a time series in which spectral components at high frequencies are reduced. A common filter method is a *simple moving average* (also called rolling average or running average) of length *N*, where *N* is an odd integer, that computes a sample mean for each subset of *N* data values (see *New York City's Weather for 1980* on opposite page and **page 50**, *Household Power Consumption*).

To de-trend data, a so-called high-pass filter can be used that reduces low-frequency variations while high-frequency variations are unaffected. That is, if the fitted trend line tracks the lowest frequencies then all that remains after the high-pass filter is applied are the residuals from the trend line (see **page 44**, **Curve Fitting**).

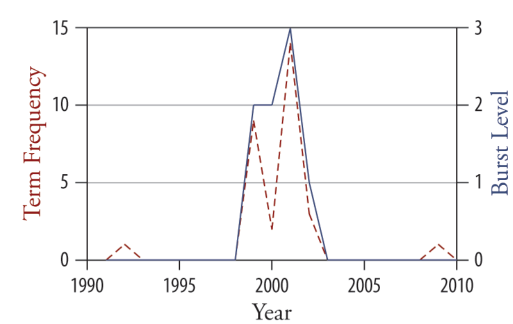
Bursts

A burst is defined as any sudden increase in activity (e.g., in the usage frequency of a certain word in a text stream or in the number of citations to a paper).



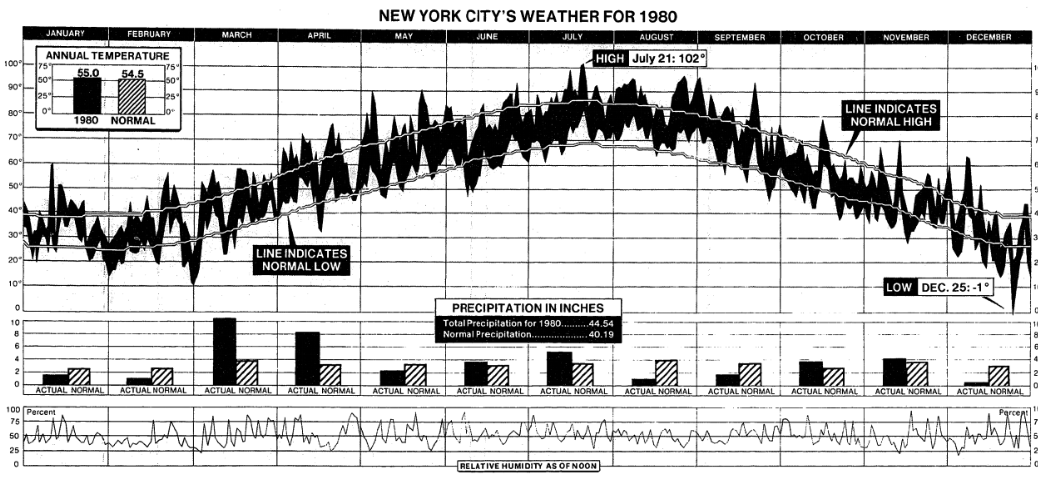
Jon Kleinberg's burst detection algorithm is commonly used to detect bursts. Given a set of time-stamped records, the algorithm identifies values (e.g., words) that occur with high intensity over a limited period of time. Rather than using plain frequencies, the algorithm employs a probabilistic automaton whose states correspond to the frequencies of individual words. State transitions correspond to points in time around which the frequency of the word changes significantly. For algorithm details, see original paper and textbook in **References & Credits** (**page 178**). The algorithm returns a ranked list of the most significant word bursts in the document stream together with the intervals of time in which they occurred. This can serve as a means of identifying topics or concepts that rose to prominence over the course of the stream, were discussed actively for a period of time, and then faded away.

Burst analysis does not require preprocessing of data; misspellings are too infrequent to cause bursts. Stopwords such as "the" typically have a high frequency throughout the time period; they do not burst. The graph below shows the number of times the stemmed term "magn" occurs in MEDLINE per publication year (dashed line) and the corresponding burst levels (solid line).

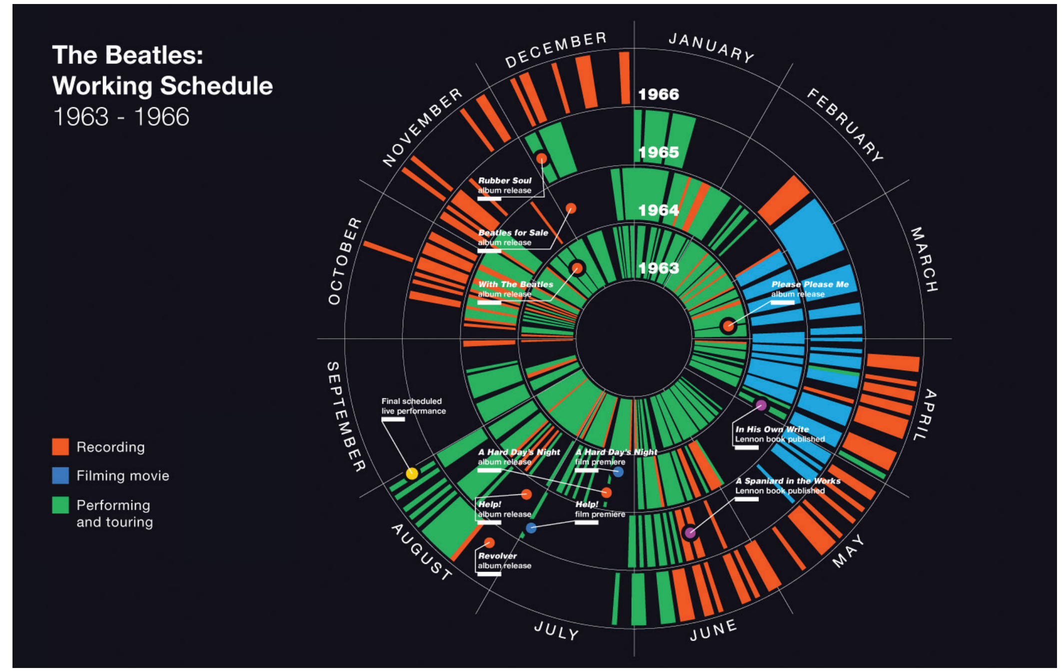


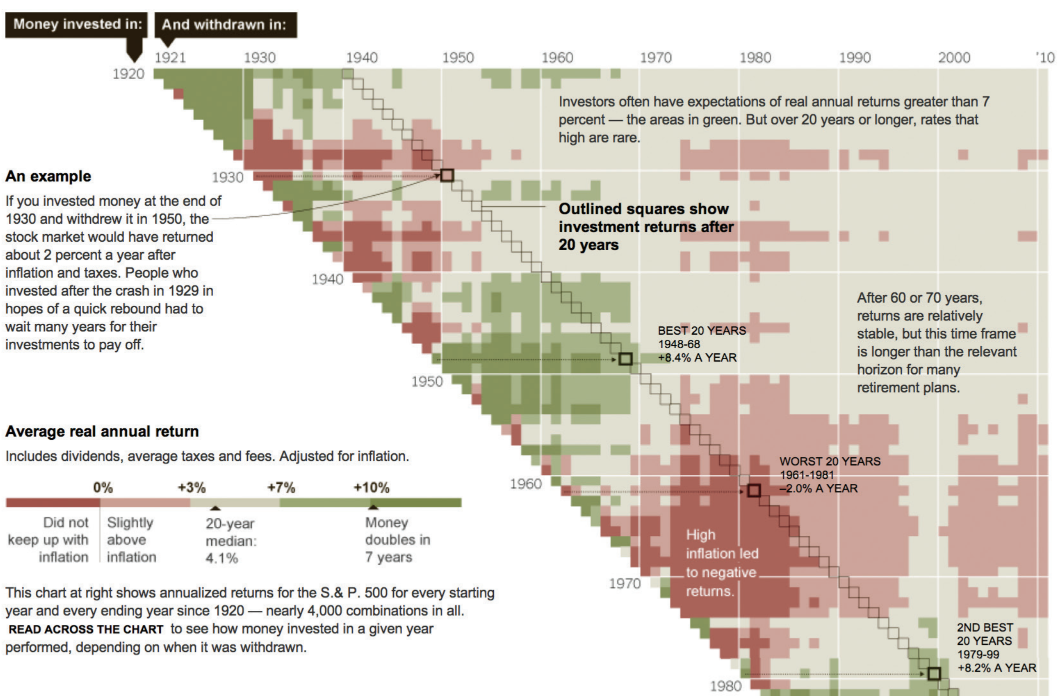
New York City's Weather for 1980

This graph of New York City's weather for 1980 shows the temperature values plus the normal high and low values for each day. Bar graphs in the middle section represent precipitation in inches for each month in 1980 as compared to normal monthly averages—with extremely high values indicated for March. Relative humidity for each day at noon is shown in the bottom section.



The Beatles: Working Schedule, 1963–1966



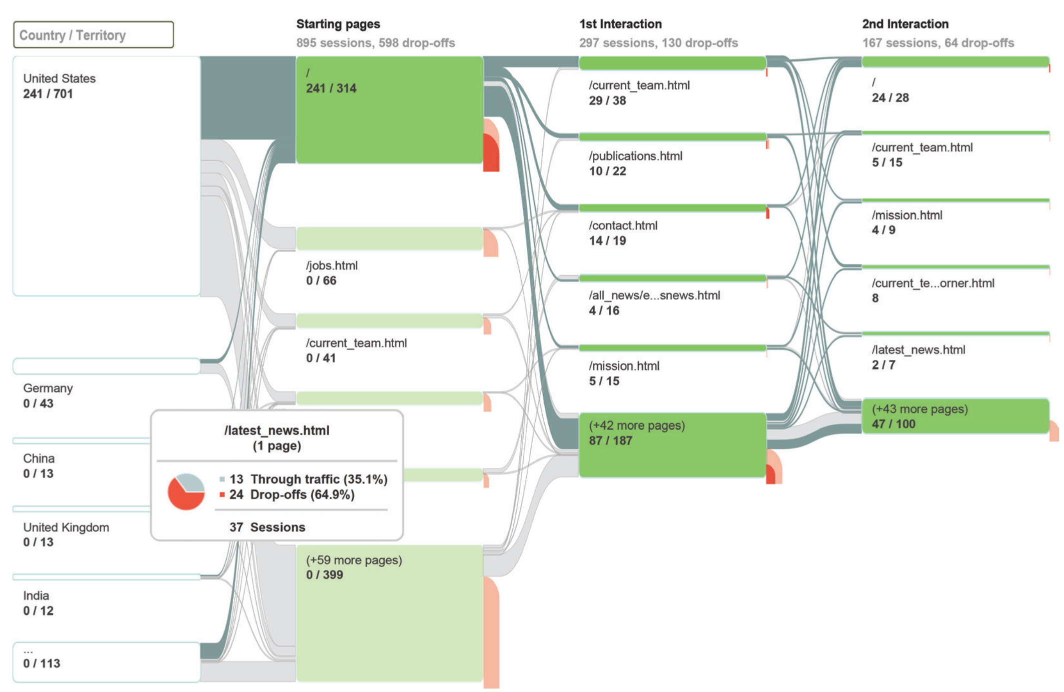


In Investing, It's When You Start and When You Finish

This graph shows annualized returns for the S&P 500 stocks actively traded in the United States over 91 years (1920–2011). Vertical years represent investments, and horizontal years represent withdrawals. The colors of the table cells signify how the money invested in a given year performed, depending on when it was withdrawn. Of the investment-withdraw-year combinations, few had money double within seven years (dark-green cells). Some combinations did not even keep up with inflation (dark red).

Sankey Graph of Google Analytics Data

Google Analytics uses Sankey graphs (see **page 63, Sankey Graph**) to show the flow of traffic across pages on a website. Depicted here is the traffic for **http://cns.iu.edu** from July 25 to August 24, 2014. A total of 895 sessions were recorded; most visitors came from the United States (**Country/Territory**). Exactly 314 users visited the home page (**Starting pages**). From the home page, visitors were most likely to go to the current team site, list of publications, information on how to contact CNS, news, and mission (**1st Interaction**). Red flows indicate drop-offs—visitors who idle or leave the site. As the number of interactions increases, the number of users decreases.



Temporal Visualization Types

Data changes over time can be represented using one static reference system with data overlays that communicate change (see **page 49**, *Sankey Graph of Google Analytics Data*); multiple static snapshots (see **page 65**, *TTURC NIH Funding Trends*); prerendered animations that can be started, stopped, fast-forwarded, or rewound (see **page 65**, *Gapminder Visualization*); or interactive services that can support changes in reference system, data overlay, and visual data encoding (see **page 67**, *U.S. Healthcare Reform*). This spread reviews major visualization types and discusses their utility to communicate trends, see distributions, perform comparisons, and identify correlations. Visualizations of temporal change involving geospatial and topic maps as well as network layouts are discussed in **Studying Dynamics** (**page 64**).

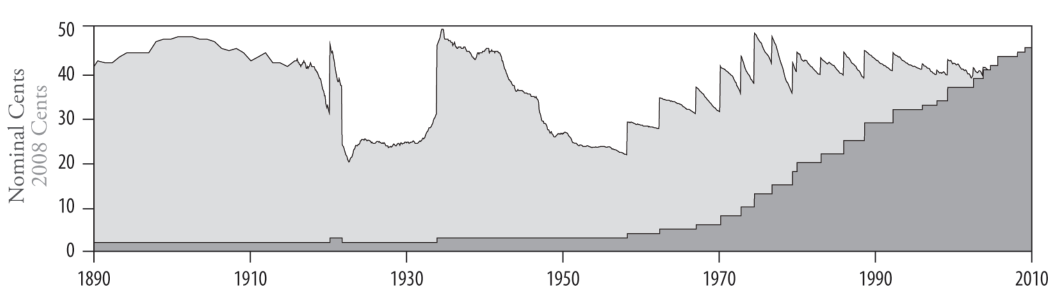
Trends and Distributions

A time-series graph (also called a timeline, chronological graph, or data-distribution graph) plots values over time, revealing the temporal distribution of a data set, such as the first and last time point, any absent values, outliers, trends, growths, peak latencies, and decay rates. If time is displayed on the horizontal axis, then it typically progresses from left to right.

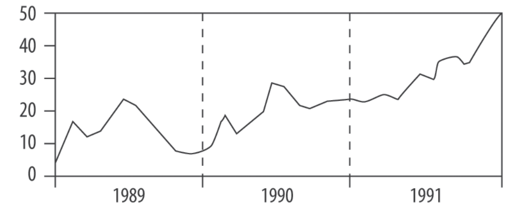
Time-series data may be discrete or continuous. The former is best represented by using discrete objects (e.g., bars in bar graphs). Continuous data often connects measurements (e.g., via lines) to highlight trends. An example of a discrete time series can be seen in the *Atlas of Science* on **page 16**, **Visionary Approaches**; the graph presented there plots the lives of famous people by using horizontal bars, each starting and ending with the dates of that individual's birth and death. Short and long lives can be quickly identified, as can lifetime overlaps, which indicate whether two people may have met. An example of a continuous time series is the graph below showing **U.S. first-class postage rates**. Nominal costs, plotted in dark gray, show the continuous increase from two cents in 1885 to 46 cents in 2013. Inflation-adjusted costs, plotted in light gray, show the relative stability of the cost of the stamp. The very same graph is given on **page 31** with an x-axis about half the size, substantially compressing time and making inflation-adjusted costs harder to read.

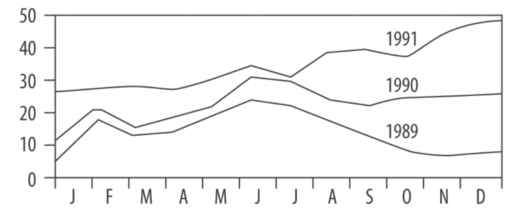
Data series can be plotted continuously (see third column, top), enabling long-term trends to be easily spotted. Repeated time scales (see below, bottom for a graph of the very same data) break data series into pieces (e.g., by year), making it easier to compare values but harder to spot overall trends. Color-coding can be used to highlight values above and below certain thresholds.

Bar graphs and line graphs can be stacked. For example, the four graphs on the right plot the same data. The top-left graph is a stacked bar graph; the bottom-left is a 100 percent stacked version; the top-right line graph is nonstacked, making the constant value 30 light-gray line stand out; the lower-right is a 100 percent stacked version of the line graph.

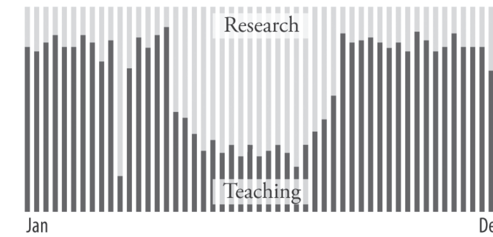


U.S. First Class Postage Rate



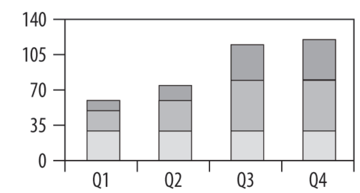


Another example of a 100 percent stacked bar graph is Nicholas Felton's 2005 *WORK VS. PLAY* report graph, which plots the amount of time spent working versus the amount of time spent playing. It inspired the graph on **Research** versus **Teaching** (see fourth column, top) that documents how much time an imaginary scholar may spend on research and teaching for each of the 52 weeks in a year. According to the graph, research (in light gray) is mostly conducted during spring break, summer, and the winter holidays, whereas teaching (in dark gray) consumes much of the spring and fall semester time.

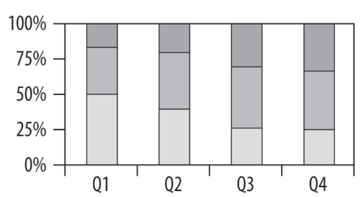


Circular line graphs display a time series on the circular axis, typically sorted in a clockwise direction. A full circle may represent 12 hours (as on a traditional, analog clock face), 24 hours (see below visualization of household power consumption), or one entire year (see **page 23**, *Causes of Mortality in the British Military During the Crimean War*, and **page 49**, *The Beatles: Working Schedule, 1963–1966*).

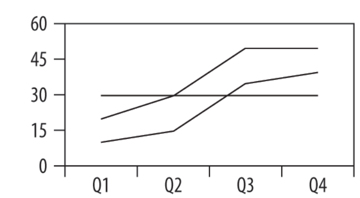
The visualization below plots energy usage per minute, using data collected in London, UK, during August 2009. The black line denotes average usage over more than four continuous days; the smoothed red line indicates the moving average of all usage data. Predictable surges in usage appear at breakfast time, dinnertime, and around 9 p.m.; unexpected usage is shown at 2 a.m.



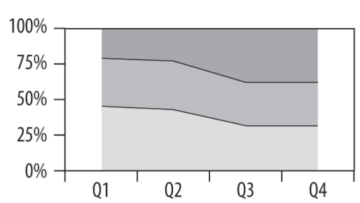
Stacked Bar Graph



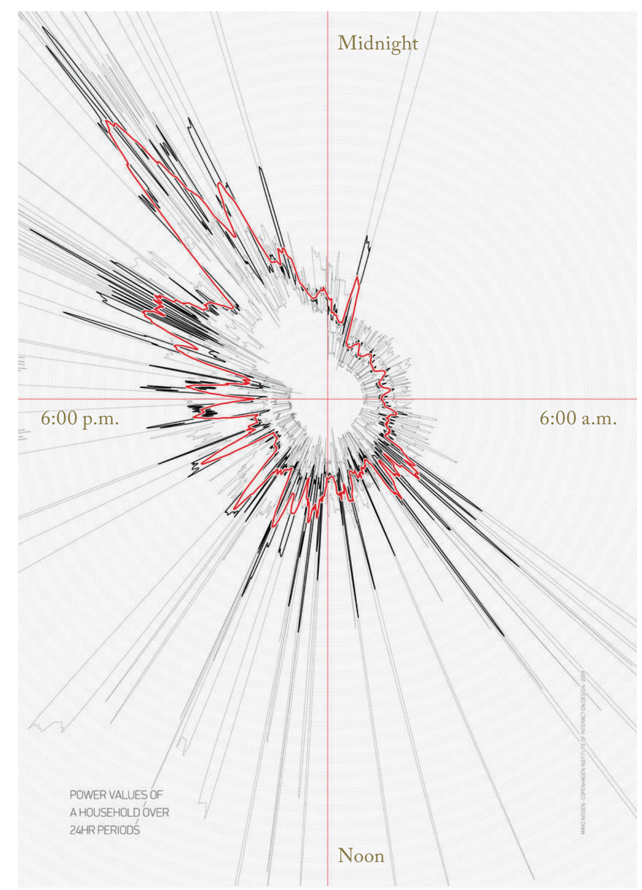
100% Stacked Bar Graph



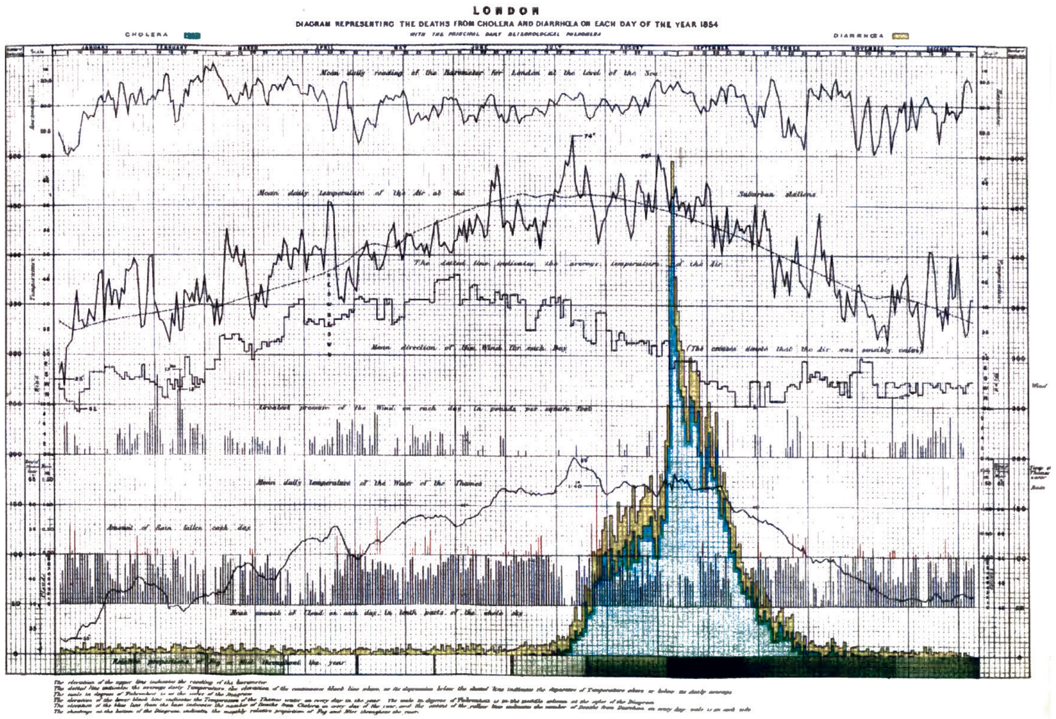
Line Graph



100% Stacked Line Graph



Household Power Consumption



Climatic Variables and Cholera and Diarrhea Cases in London, 1854

Comparison

Plotting data sets in the same reference system—graph, geospatial map, or network—supports comparisons (see also **page 66, Combination**). Examples include the *The Baby Name Wizard* (**page 69**) and the line graph above that plots climatic variables and incidences of different diseases in London for each day of the year 1854. Cholera deaths are shown in blue, with the epidemic peaking at the end of August; the maximum value for a day is about 450 incidences. Deaths from chronic diarrhea, rendered in yellow, occur throughout the entire year.

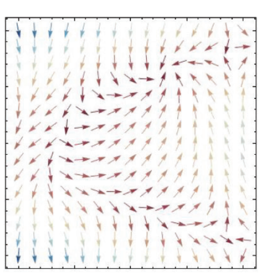
See also the *Spot Map of the Golden Square Cholera Outbreak* of 1854 that shows the geospatial distribution of death from cholera around a water pump on Broad Street, substantiating John Snow's theory that drinking water might be a potential cause of the disease (**page 23**).

Multiple static snapshots (also called small multiples) and glyphs (see **page 46, Statistical Visualization Types**) can be used to support comparisons (see *Icon Symbols on a Graph* and *Icon Symbols on a Map* on **page 66**).

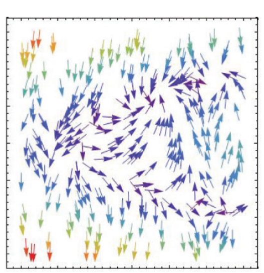
Alluvial graphs were designed to show change over time. They honor sequential ordering and can be used to compare networks and their structural changes over time (see *Evolving S&T Landscape* on **page 16**, top-right and **page 59, Alluvial Graph**). Whereas parallel coordinate graphs (**page 46**) focus on the legibility of attribute values over multiple axes, using links to interconnect all values per record, alluvial graphs focus on the legibility of linkages and use efficient sorting of arrow bundles to improve legibility.

Derivatives

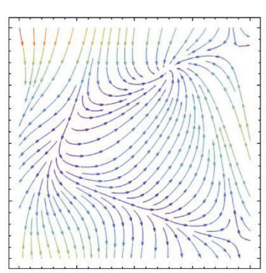
It is common for scientific visualizations to show temporal derivatives, such as velocity, rather than time-based indications. Such derivatives are valuable in trying to understand the speed of diffusion for tangible or intangible entities. Heat maps and glyphs are commonly used to represent fields. Four different visualizations of the wind velocity of Hurricane Gustav, the second most destructive hurricane of the 2008 Atlantic hurricane season, are shown below. Arrow direction indicates wind direction, whereas arrow length denotes wind speed. Similar visualizations can be used to depict the flow of tangible objects (e.g., people or goods) or intangible objects (e.g., virtual currency or innovations) over space and time (see maps by Waldo Tobler in *Atlas of Science*, **page 161**).



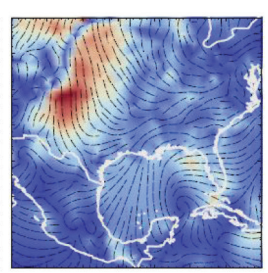
Field Vectors on a Regular Grid



Field Vectors at Random Positions



Streamlines



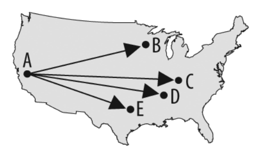
Streamlines Overlaid on a Geospatial Map

Flows over Time and Space

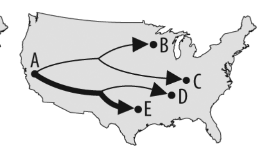
Different visual representations exist that depict the flow or movement of tangible or intangible objects from one location to another.

Flow Map

Flow maps are a combination of a (typically geospatial) base map and a network data overlay in which the flow quantity is represented by the width (or weight) of a directed link. A major characteristic is that the flows are bundled (see example below).



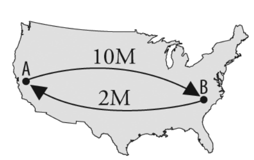
Unbundled Flow



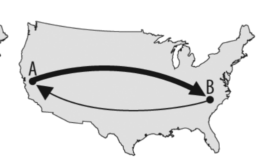
Bundled Flow

Some of the earliest known examples appear in the *Album de Statistique Graphique* by the Bureau de la Statistique Graphique of the Ministry of Public Works (1879–1899). Two other examples are *Europe Raw Cotton Imports in 1858, 1864, and 1865* (see **page 80**) and *Napoleon's March to Moscow* (see *Atlas of Science*, **page 84**), both by Charles Joseph Minard.

Links can be directed or undirected, weighted or not, and may have additional variable values. For example, the amount (quantity or value) of flow can be indicated with a link label but could also be depicted by link width (see below). If link width is proportional to a value, then a scale should be provided in the legend.



Link Label

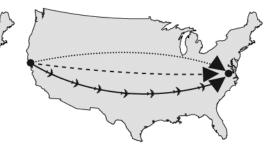


Link Width

Migration maps use arrows and line width to indicate the volume and direction of migration probabilities (see visualizations on **page 18**). The type of flow is commonly indicated in the title, on the map, or using graphic variables.

Link width can be approximate (see the links in the example below-left, indicating empty or loaded trucks). A legend is needed if different graphic symbols and/or variables are depicted. The entity or resource that has traveled (e.g., individual, company, water, or gas) and the type of travel made (e.g., via car, train, air, or pipe) may be indicated via link shape (see below-right) but the meaning of each link shape needs to be specified in a legend.

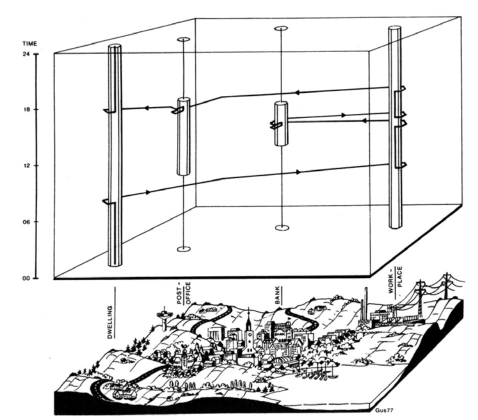
Link Width



Link Shape

Space-Time-Cube Map

Space-time-cube maps show movement in three dimensions using a two-dimensional terrain and a vertical time axis. Torsten Hägerstrand was among the first to map an individual's path in a space-time coordinate system (see below). Career trajectories and other movements over geospatial or topical space can be represented via space-time cubes (see **page 19**, *Nobelpreisträger für Physik*).



Animation

Change over time can also be depicted by having users watch a previously produced or manually steered animation showing such change over time. Although it is hard to focus on multiple changing objects at once, animation can be extremely effective in showing general trends (see **page 64, Studying Dynamics**).

Geospatial Studies—"Where"

Geospatial analysis (also called geostatistical analysis) has a long history in geography and cartography. It was developed to answer "where" questions by using statistical models and tools for spatial-data exploration and map generation. Specifically, it aims to answer where something happens and with what impact on neighboring areas. Given a limited number of data measurements, it supports the exploration of data variability, including unusual data values; the calculation of spatial relationships and global and local trends; the computation of statistically valid prediction surfaces, along with prediction uncertainties; the rendering of data as spatial animations to portray changes and flows (see *Impact of Air Travel on Global Spread of Infectious Diseases*, *Atlas of Science*, **page 150**); and the creation of reliable maps offering predictions, prediction errors, quantiles, and probabilities for improved decision making.

Everything is related to everything else, but near things are more related than distant things.
Waldo Tobler's first law of geography

Data Preprocessing

Geospatial data needs to be geocoded and georeferenced in order to be visualized. Distances and diffusion matrices may have to be calculated in support of geospatial analysis and visualization.

Geocoding

In order to place any data record on a map, its *geocode* (i.e., its location, as represented by an address, a census tract, a postal code, or geographic coordinates) must be determined. *Geographic coordinates* refer to locations on the Earth's surface that are expressed in degrees of latitude and longitude.

Gazetteers are used to maintain geographic name data. They contain lists of geographic places and their latitudinal and longitudinal coordinates, including other information such as area, population, and cultural statistics. For example, an author can be geolocated according to her affiliation—and her trajectory comprised of a sequence of geolocations, starting from her degree-granting institutions, extending to the places where she has worked, and culminating at the organization at which she retired (see **page 19**, *Nobelpreisträger für Physik*).

Alternatively, a Global Positioning System (GPS) can be used to acquire data on a person's geospatial position. Here, a constellation of 24 satellites orbiting Earth, at an altitude of 20,200 kilometers, transmits signals that allow a GPS receiver anywhere on Earth to calculate its own location (see **page 172**, *FourSquare Transportation Check-ins Showing Thanksgiving Travels*).

Reverse geocoding reads a point location (latitude and longitude) and returns an address or place name.

In order to place a data record, it needs to be *georeferenced* (i.e., coordinates from a known reference system, such as latitude and longitude, have to be assigned to the coordinates of an image or a planar map).

Distance

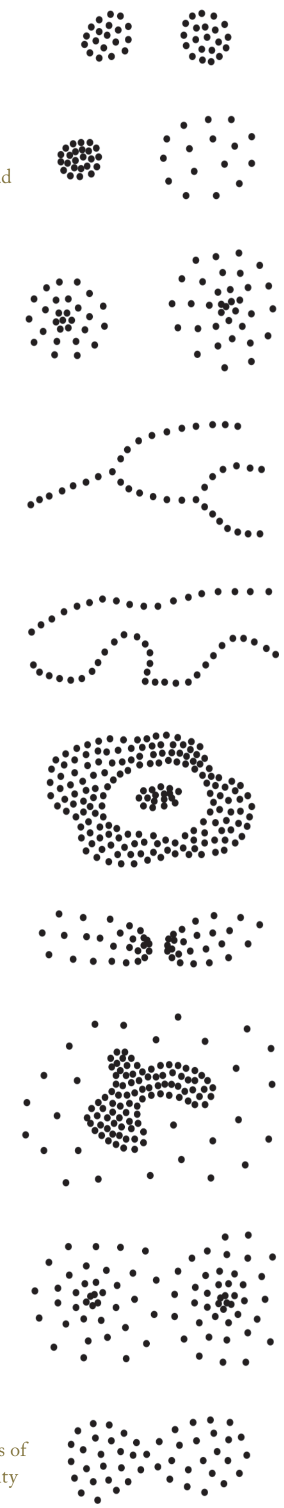
For any two points on the surface of the Earth, the shortest distance between them is always along a *great circle* (i.e., the Earth's circumference at its widest point). Although the Earth is in fact shaped as an oblique spheroid, great circle distance calculations tend to suffice for most applications.

Diffusion Matrix

Tangible objects (e.g., students, inventors, or money) and intangible objects (e.g., ideas, theories, or reputations) diffuse over time and space. To compute the diffusion of features, a *movement table* is used. A movement table is a square matrix indicating movement from every point to every other point per time duration. Sample tables may represent author movement based on affiliation data or knowledge-diffusion data based on citation linkages.

Clustering

Clustering (also called aggregation) of records may be driven by geospatial properties, by existing classifications, or in a data-driven way. A large number of clustering problems exist. Shown in the next column are ten sample layouts of dots in which dot proximity indicates similarity.


Clusters with Same Dot Density
Clusters with Different Dot Densities and Spread
Clusters with Varying Dot Densities
Dot Line Tree Structure
Two Unconnected Dot Lines
Cluster Enclosing Another Cluster
Clusters with Varying Dot Densities
High-Density Cluster in Low-Density Area
Bow Tie of Two Clusters with Varying Dot Densities
Bow Tie of Two Clusters of Homogenous Dot Density

Using Geometric Grids

Geospatial data can be aggregated by dividing geospatial areas into zones, such as regular grids, honeycomb patterns (see *In Terms of Geography* in *Atlas of Science*, **page 103**), or arbitrary patterns (e.g., those of a grid designed to match the shapes of building rooftops in order to show energy waste; see **page 31, Maps**).

Using Existing Classifications

Existing regional zonings and groupings can be used to aggregate geospatial data by neighborhoods, municipalities, sectors, states, regions, or nations.

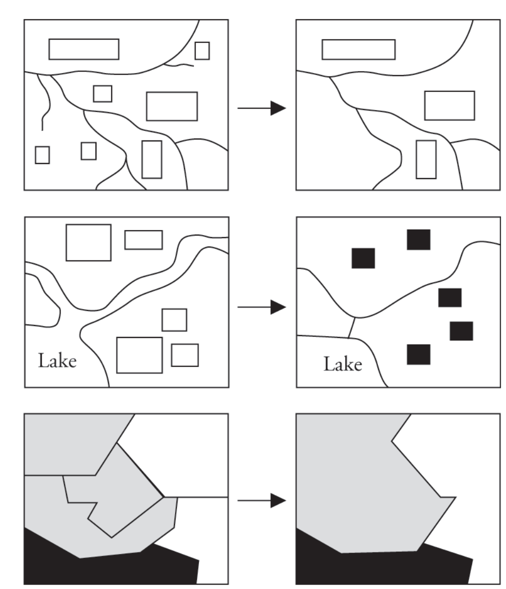
In the United States, a *census block* is the smallest geographic unit used by the U.S. Census Bureau for reporting census data. A *census tract* combines adjacent census blocks into a group of approximately 4,000 people. In Europe, the Nomenclature of Units for Territorial Statistics (NUTS) is a standard for referencing the subdivisions of countries for statistical purposes.

Data-Driven Clustering

Given a set of geolocated records, different algorithms can be applied to group them geospatially. K-means clustering can be run to create *k* groups of geospatially close records.

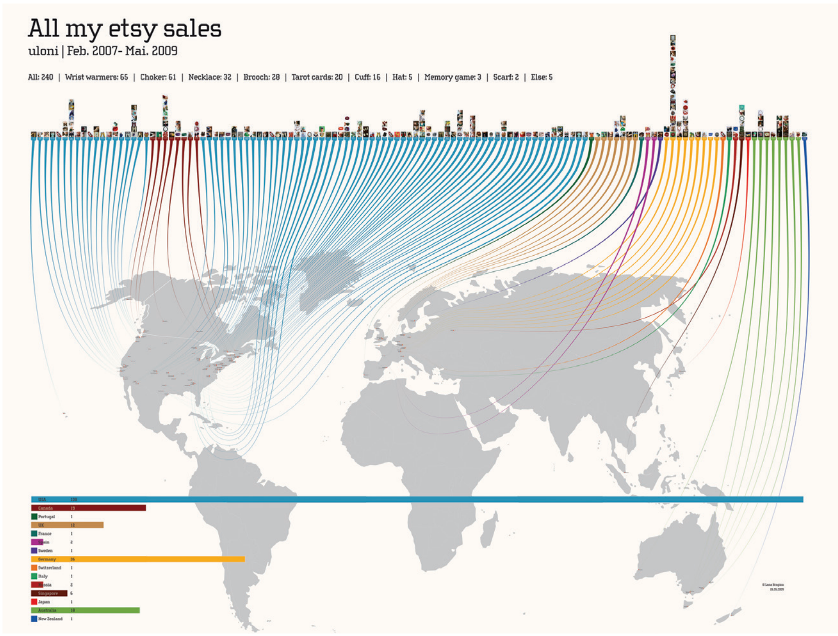
Visual Generalization

As the area that is being mapped becomes larger (or the density of items in an area increases), fewer individual features can be shown on a map. For example, individual houses can be depicted at a scale of 1:100 but not at a scale of 1:100,000. Decisions need to be made about what is important to retain (e.g., a specific selection of author affiliations or a general array that reflects the area they work in). Although generalization entails information loss, it should nevertheless be able to preserve the essence of the map while maintaining geometric and attribute accuracy, visual hierarchy, and aesthetic quality. To ease map reading and navigation between levels of generalization, key features in the original map should remain prominent (see examples below).



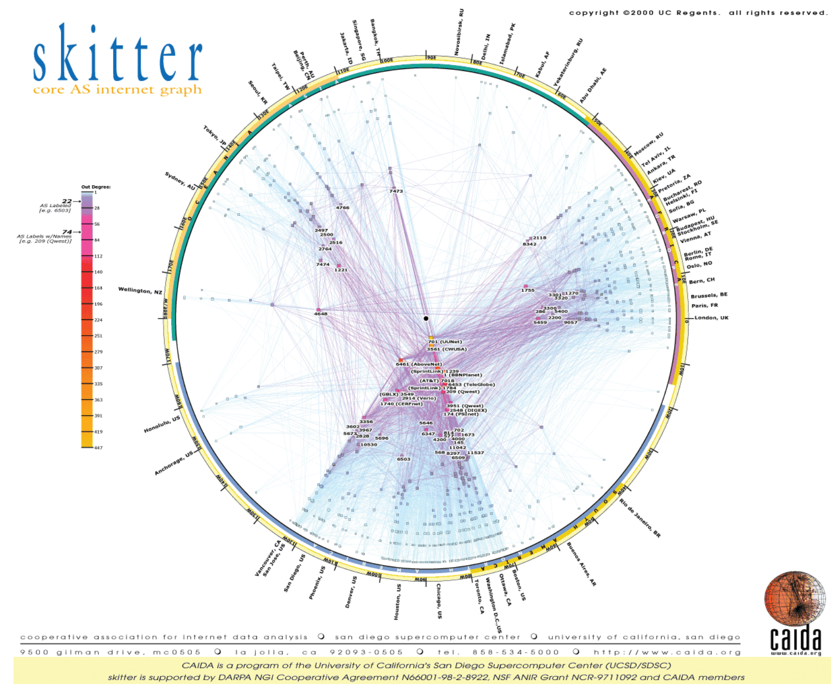
Etsy Sales Map

Ulani, an avid knitter and crocheter, sold 240 handmade items via Etsy between February 2007 and May 2009 to consumers around the globe: 138 from the United States, 36 from Germany, 19 from Canada, and 12 from the United Kingdom. She created this infographic to show where she sent each item. Vertical bar graphs at top feature images of items purchased per customer. Color-coded horizontal bar graphs depict the number of items purchased by consumers per country. Explore high resolution versions of all images at **http://scimaps.org/atlas2**.



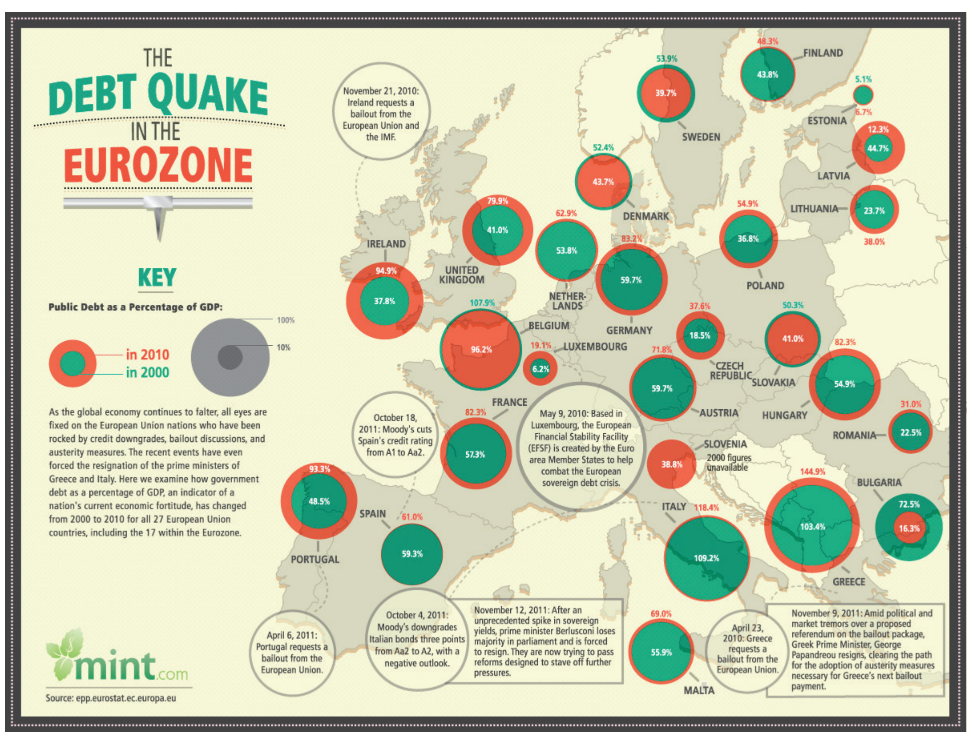
Skitter Internet Map

Using a circular layout, the network below represents 1,134,634 IP addresses and 2,434,073 IP links of topology data gathered from 25 monitors probing approximately 865,000 destinations spread across 76,000 globally routable network prefixes (62 percent of the total). IP links refer to immediately adjacent addresses in a traceroute-like path.



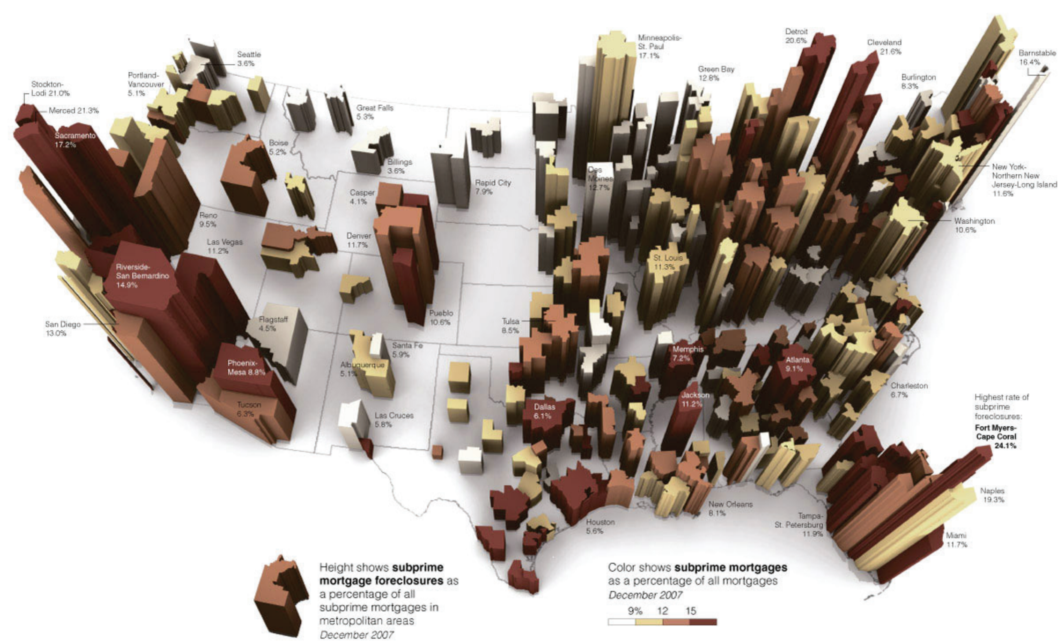
The Debt Quake in the Eurozone

This information graphic by ColumnFiveMedia and Mint shows change in government debt as a percentage of GDP from 2000 to 2010 for all 27 European Union countries, including the 17 within the Eurozone.



In the Shadow of Foreclosures

Hannah Fairfield created this stepped relief map for *The New York Times* online. It shows the unusually large number of delinquencies and foreclosures for subprime mortgages (a type of loan granted to individuals that would not qualify for conventional mortgages because of poor credit histories). The number of subprime mortgage foreclosures as a percentage of all subprime mortgages by geographic region are mapped to area height; **Fort Meyers-Cape Coral** has the highest value with 24.1%.



Geospatial Visualization Types

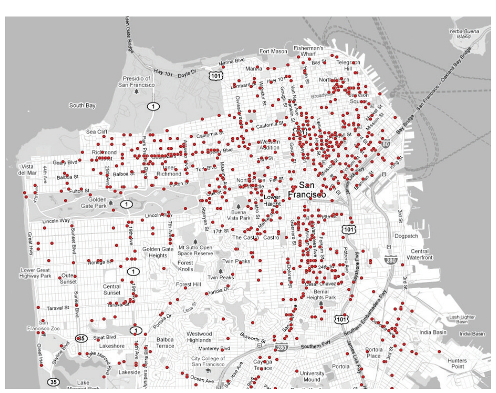
Different map types serve different purposes. Road maps help drivers to find their way. Weather maps show the temperature, air pressure, or rainfall in a given region. Geological maps show resources underground and are used to help plan building work or drilling for gas or oil. Here, the focus is on thematic maps that show a particular theme connected with a specific geographic area. Data portrayed may be physical, social, political, cultural, economic, sociological, agricultural, or technical; or it may reflect any other aspects of a city, state, region, nation, or continent to help viewers identify or compare spatial patterns. The maps are grouped by the space they represent: discrete versus continuous. Different types of line maps depicting flow are discussed as well; for information on space-time-cube maps, see the discussion in **Temporal Visualization Types** (**page 51**).

Discrete Space

Data can be raw or computed; it must be possible to aggregate the data via artificial collection units.

Dot Density Map

A dot density map depicts a set of data records by using dots to show the density, distribution, and skews of data. Each dot represents the same number of data records (see below map of San Francisco showing about 1,000 crimes recorded in 2009 and 2010). A dot may represent multiple data records; if one dot represented 100 records, an area with ten dots would then denote 1,000 records. In addition to indicating numbers of records, dot maps reveal the location and spatial distribution of those records and are frequently used to show population distribution. Dot maps can be misleading, however, as the spatial variation within aggregated regions cannot be represented, sharp gradients cannot be shown, and dots may be misread as point symbols.



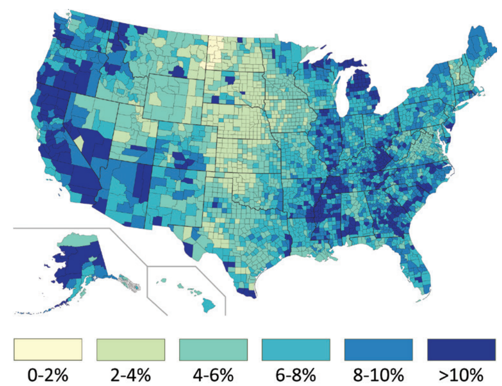
Proportional Symbol Map

A proportional symbol map (also called a graduated symbol map) plots the value of data variables proportionally to graphic symbol types (**page 32**) and their graphic variable types (**page 34**). This map type should be used if the data has absolute values and occurs at points, or can be aggregated at points, within geospatial areas. It should not be used to map densities, ratios, or rates (e.g., population densities); for such purposes, a choropleth map (discussed ahead) should be used instead. Circles are the most popular proportional symbol because they are easy to construct, scale, and read (e.g., overlaps are much easier to spot).

Examples include maps with bivariate circles, such as those used in *The Debt Quake in the Eurozone* (**page 53**); maps with pie-chart nodes coded by area size, as in *U.S. Healthcare Reform* (**page 67**); or the map below that uses size- and shape-coded linguistic symbols (country names) to visually represent each country.

Choropleth Map

A choropleth map represents data variables—such as densities, ratios, or rates aggregated over artificial collection units—through the coloring or shading of those areas. It is used when predefined statistical areas (e.g., census tracts, voting districts, or school districts) or administrative political subdivisions (e.g., townships, counties, or states) are important to visualize (e.g., when displaying population density or per-capita income per census track or country). Examples include *A Global Projection of Subjective Well-Being* (**page 98**) and *The Millennium Development Goals Map* (**page 120**). U.S. unemployment in 2009 is shown in the county-level map below; compare with the state-level choropleth map of U.S. election data on **page 24**.



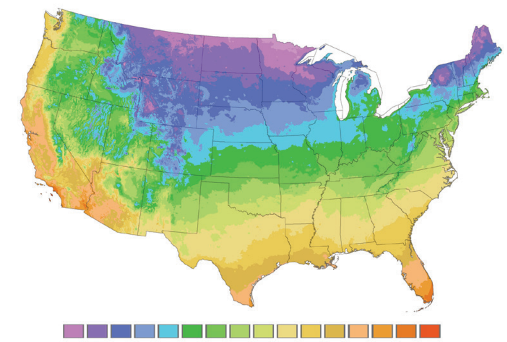
Phenomena that are continuous (e.g., average annual rainfall, temperature, or population distribution) should not be mapped via this method, because their distributions are not controlled by political or administrative boundaries. Instead, isarithmic maps (see opposite page) should be used, in which region boundaries are defined by data patterns and each isoline has a specific value. When using choropleth maps, a number of challenges arise, such as the illusion of sharp borders, vast regions appearing too homogeneous, or exceptionally small regions that prove too small to see.

Stepped Relief Map

The stepped relief map (also called a prism or block map) elevates areas proportionally to their data values. Examples include the visualization of energy usage by city block in New York City on the jacket of this *Atlas*, *In the Shadow of Foreclosures* (**page 53**), and *On Words—Concordance* (**page 57**).

Dasymetric Map

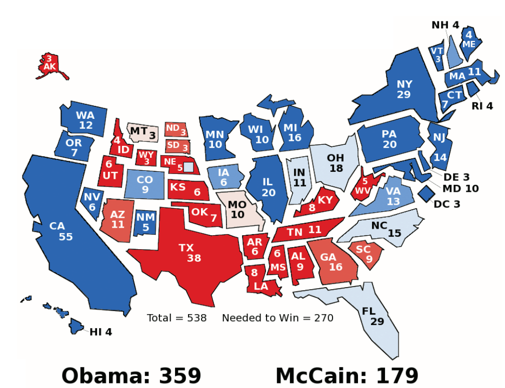
The dasymetric map is a hybrid of the choropleth and isarithmic maps (see opposite page). It utilizes standardized data but places aerial symbols by taking into consideration actual changing densities within the boundaries of the map. For example, the map below plots climate and plant hardiness zones (purple colors are coldest and orange is warmest) on top of a U.S. state boundary map for easy location referencing.



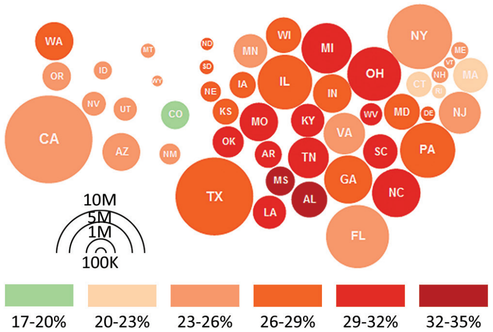
Cartogram Map

Cartograms (also called value-by-area maps) distort geographical areas in proportion to data values. For example, a cartogram of world population may show countries as being either larger or smaller in proportion to their populations while aiming to preserve the location of features insofar as possible. A cartogram is most effective if large areas have small values (i.e., the areas will appear smaller) and small areas have large values (i.e., the areas will appear enlarged). The cartogram is mostly used for world, continental, and country maps, as familiarity with nondistorted regions is necessary to read the map correctly. Three major types exist: disjoint, pseudo-continuous, and continuous cartograms.

Disjoint cartograms (also called noncontiguous cartograms) preserve the shape of size-coded regions and are noncontinuous. Each region is scaled according to a selected data variable (e.g., population) and positioned with relative accuracy in relation to neighboring regions. The map below shows a disjoint cartogram of the U.S. Electoral College.



Pseudocontinuous cartograms (also called Dorling cartograms) transform regions into geometric shapes (e.g., circles), which are sized proportional to the magnitude of a data variable and placed in approximate locations. Shown below is a U.S. map of all contiguous states, each of which is represented by a circle that is size-coded per the state's total rate of obesity and color-coded per the percentage (or prevalence rate) of obesity among the state's total population (see legend for scale used).



Continuous cartograms aim to preserve proximity and continuity. However, as they may distort shape extensively, they are more difficult to read and construct. Examples include *Venture Capital Disbursed* (**page 9**), *Ecological Footprint* (**page 90**), and the colored U.S. election maps (**page 25**).

Continuous Space

Qualitative or quantitative data can be mapped onto continuous space using different visualization types.

Elevation Map

An elevation map (also called an altitude map) plots discrete or continuous data to height values. This is done in relief maps, such as that of the San Francisco crime map below, which depicts higher crime rates by higher elevations.

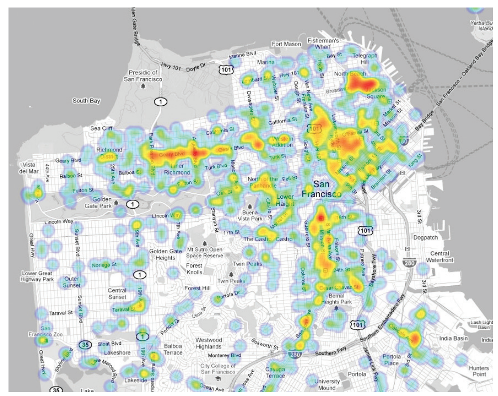
Isarithmic Map

Isarithmic maps (also called isogram maps) use isolines to map continuous data such as elevation or population density. An isoline is a line along which all points are of equal value. Types of isolines include isohypses, or contour lines used to represent elevation; isotherms, used for temperature; and isochrones, for travel time.

The smaller the distance between isolines, the steeper the slopes of maxima (e.g., hills) and minima (e.g., valleys). Areas between isolines can be shaded or colored, such as in a heat map, which is a type of isogram map; shadows can be cast using an imaginary light source (called shaded relief) to give the map a three-dimensional appearance. Minima and maxima can be indicated using hachures, and isolines can have values.

When computing isolines (e.g., for population density, which equals the population of a census district divided by the surface area of that district), each calculated value is presumed to be the value of the variable at the center of the area, and isolines are drawn by a process of interpolation.

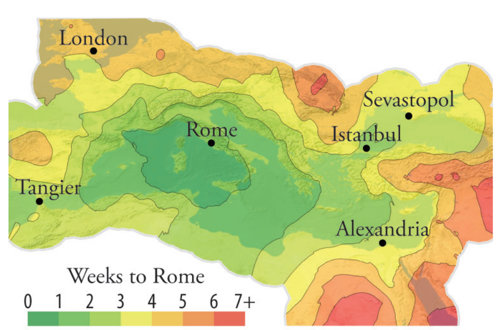
The below map of San Francisco renders the same 1,000 crimes as an isarithmic map with light blue denoting low crime and red indicating high crime areas.



Isochrone Map

Isochrone maps (also called travel time maps or anamorphic maps) utilize isolines to show equal travel time. They were first used in the *Album de Statistique Graphique*, by the Bureau de la Statistique Graphique of the Ministry of Public Works, led by Émile Cheysson (1879–1899).

The isochrone map below documents travel time to Rome in July, circa 200 A.D., via sailing ship, civilian riverboat, and walking. Contours indicate time in days, ranging from 7 to 42 days. This map and others can be explored interactively online to understand how travel times and transport prices structured the Roman world.



Vector Fields

Diffusion potentials and gradients can be visualized as continuous spatial gravity models. For an example that is based on such a model, see the central image on **page 161** of the *Atlas of Science* that depicts the pressure to move in the United States.

Given the distribution of particular features over geographic space as raster data, vectors can be determined by using the density gradient to compute the "pressure field" exerted by these points. The pressure field can then be used to predict the tendency for (outward) diffusion and (inward) absorption/adoption at a certain point in geospatial or topic space.

Vector fields can be visualized using glyphs (e.g., length- and width-coded arrows) to indicate the potential and gradient (force and direction) of the field. See the wind velocity visualization for Hurricane Gustav on **page 51**, bottom-left.

Line Map

Line maps (also called linkage maps) show the paths that either tangible or intangible objects take to get from one geospatial place to another. Ernest George Ravenstein's map of the *Currents of Migration* (1885) is one of the first line maps ever created.

Route Map

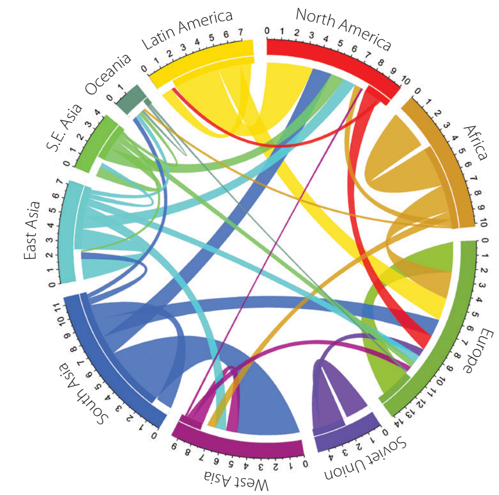
Route maps depict public transportation systems, such as walkways, railroad tracks, streets, and air traffic corridors, but also show distribution networks, comprising water pipes or electric cables. Variables such as the number of street lanes or maximum speed, pipe diameter, or cable voltage can be encoded. Note that route maps encode only information about the paths themselves, whereas flow maps (**page 51**) encode details on flow content and volume.

Subway Map

Subway maps aim to simplify the route map layout by optimizing a number of desirable properties such as symmetry, evenly distributed nodes, uniform edge lengths, minimized edge crossings, orthogonal drawings, and minimized areas, bends, slopes, and angles. These criteria may be relaxed to speed up the layout process. An example is the *PhD Thesis Map* in the *Atlas of Science* (**page 90**).

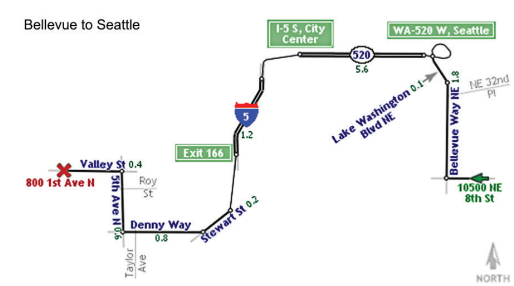
Flow Map

Flow maps (**page 51**) represent any matter that flows, moves, and migrates together with the direction and amount of such flow. They are used to visualize the trajectories of tangible objects (e.g., scholars or physical goods) and intangible objects (e.g., ideas, expertise, or digital documents). In contrast to route maps, they show little of the concrete paths that connect one point to another. Examples are the chord graph of human migration flows during 2005 to 2010 in below figure and the visualizations on **page 18**.



Strip Map

Strip maps (also called diagrammatic maps) distort geospatial and other types of space to improve legibility. They focus on the sequential relationships of items shown (e.g., landmarks on a road trip) instead of on directional and geographic relationships. An early version of the strip map is the Peutinger map (*Atlas of Science*, **page 10**), which shows major Roman Empire travel routes on an approximately 20' x 1' (60 m x 30 cm) scroll. A more recent implementation is Line Draw, which generates abstract route maps when given a travel starting point and destination. The sample route below shows directions from Bellevue (right) to Seattle (left). Line Draw details best travel routes for leaving a city and major highways between cities; plus, it can zoom in to a specific address, when needed. Selected information such as street names and highway numbers are provided. Distances and travel times may also be noted.



Space-Time-Cube Map

This map type shows space (x-y plane) with the third spatial dimension to represent time (the z-axis); see discussion and example on **page 51** and *Nobelpreisträger für Physik* on **page 19**).

Topical Studies—"What"

A linguistic analysis of text is commonly applied to answer "what?" questions. Large-scale text corpora (e.g., titles, abstracts, or full texts) of papers, patents, grants, job applications, or email data streams are semantically preprocessed, analyzed, and visualized to identify term frequency distributions or temporal dynamics inherent in the text. This spread discusses different data processing steps together with approaches that aim to identify topic distribution and topical change over time. Results are visualized using lists, charts, graphs, maps, and network layouts (see sample visualizations on the opposite page and science maps in **Part 3** of this *Atlas*). Different topical visualization types are discussed on the subsequent spread.

We are what we write, we are what we read, and we are what we make of what we read.
Martin Bloomer, Phil Hodkinson, and Stephen Billett

Data Preprocessing

Preprocessing typically entails stemming, stopword removal, and identification of unique terms as well as extraction of any single or compound terms that are to be used in the semantic analysis. The result is a matrix that documents how often a sequence of terms appears in each record; thousands of unique term combinations times thousands or even millions of records proves to be a very high-dimensional semantic space. To ease navigation and processing of this space, different similarity measures and dimensionality-reduction approaches can be applied to generate a much lower dimensional space, which preserves the main structure that is inherent in the original data.

Fielding

Typically, the first step in data processing is fielding. Although some text may be semantically tagged (i.e., it is known what part of a text file represents the title, author name(s), address field, abstract, different sections, and references), most text tends not to be fielded. Manual or algorithmic methods have to be applied to parse the text and to identify what parts of the text represent what semantic content.

Text Selection

Next, a decision has to be made as to what part of the text should be used in a study. Titles (e.g., from scientific publications) are typically short and therefore have comparatively few words to reveal the topic of a given paper. For example, in the title "All you ever wanted to know about X," only "X" is relevant to the meaning of that paper. Abstracts and keywords are commonly used in semantic analyses. Full text is required for citation-context analyses but is generally large in size, and more disk space and processing power is required to process full text. Furthermore, care must be taken to normalize for different text lengths, because the probability of a term occurring in a record rises as text size increases.

Stemming and Stopword Removal

Stemming is used to reduce terms to their stem or root form (e.g., "scientific" and "scientifically" are reduced to the root form, "scientif"). As a practice, stemming considerably reduces the number of unique terms. Stopword lists are applied to exclude common (and therefore dispensable) words or phrases such as "the" or "a" from a textual analysis. Standard stopword lists exist, but users can add additional terms as needed for a specific analysis. All text may be converted to lowercase to greatly reduce the number of unique terms; however, terms like "IT" and "it" then become identical in the process.

Tokenization

Tokenization breaks up text into words, phrases, symbols, or other meaningful elements, called tokens. Special attention is paid to punctuation, including hyphens. Delimiters are used to separate tokens (e.g., the string "science and technology" would be split into three tokens: "science|and|technology"). Words or phrases composed of multiple terms to communicate one concept (e.g., "bibliographic coupling") can be extracted together in order to preserve the intended meaning.

Sequences of *n* items occurring in text are called n-grams. They may be characters, syllables, or words. For example, "science and technology" can be subdivided into three unigrams (science, and, technology), two bigrams (science and, and technology), and one trigram (science and technology). During the n-gram construction process, punctuation marks are typically treated as a separate term except for currency symbols, decimal components of numbers, and apostrophes indicating possessive case. Case is frequently ignored, with some negative implications for search specificity (see above "IT" and "it" example). The number of n-grams that can be extracted from a corpus greatly exceeds the number of terms in that corpus. The Google Labs' Ngram Viewer supports the quantitative study of trends based on n-grams appearing in more than five million books published between 1800 and 2000. Shown below is the search result for each of the following terms: science, technology, art, design, and poetry. The most dramatic changes appear for "design," as use of the term starts high, then wanes, then waxes again. In contrast, the term "technology" is rarely used before 1920; after 1960, however, a surge in the rate of use occurs, likely caused by the space race during the Cold War that made people aware of technological advancements.

Normalizations

Normalizations are often necessary in text comparisons. For example, when comparing texts across years, the n-gram frequencies for each year should be divided by the total number of words that appear in the corpus for that year. The same normalization also works for comparing texts of different lengths.

Descriptive Term Identification

High- and low-frequency terms (e.g., extremely common terms or misspelled words) may be excluded from a semantic analysis as they contribute little to the understanding of the textual similarity of text records.

Gerard Salton's term frequency/inverse document frequency (TF/IDF) weight can be calculated to identify the most descriptive terms. The weight of a term t equals the product of the term frequency (TF) and the inverse document frequency (IDF). IDF is calculated by taking the logarithm of the total number of records divided by the number of records that contain the term t. That is, the TF/IDF value increases proportionally to the number of times a term appears but is offset by the frequency of the term in the corpus. Other approaches such as Latent Semantic Analysis (LSA) or topic detection can be applied to compute latent terms or topics from unstructured collection of text. The similarity of two text records is then computed based on the most descriptive terms or topics. For details, see publications listed in **page 178**, **References & Credits**.

Tagging

Grammatical tagging, such as part-of-speech (POS) tagging, identifies if a word is a noun, verb, or adjective, singular or plural, and so on. Lookup tables and more advanced linguistic analyses are used to identify publication titles, author names, or author addresses in a publication record (see **Fielding**).

Distributions

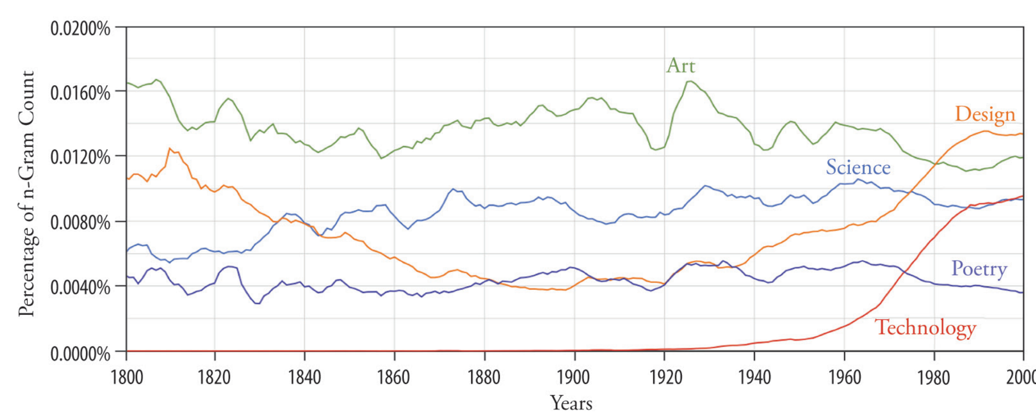
Understanding the topical distribution of text and its change over time is a major goal of topical studies.

Term Frequency and Distributions

The number of times a term occurs in a text corpus is called term frequency. It is often equated with the term's level of importance or relevance. Raw frequency counts can be deceiving because they do not account for text length or change in the number of text records published per year. Term frequencies can be graphed using a line graph (see below) or mapped onto the graphic variables of graphic symbols (see *On Words—Concordance* on opposite page, which maps the frequency of major terms used in two books on American politics onto the height of each term).

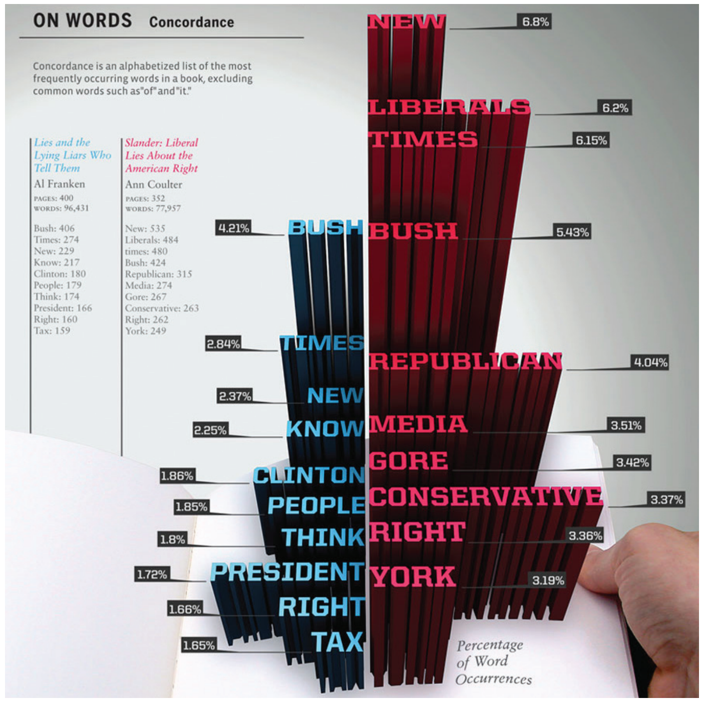
Temporal Dynamics

Term frequency in a corpus may change over time. Selecting those terms, or n-grams, that have the most absolute change over time can lead to new insights. Burst detection, discussed in **Temporal Studies—"When"** (**page 48**), is frequently applied to identify sudden changes in the frequency of terms, author names, or citation reference strings.



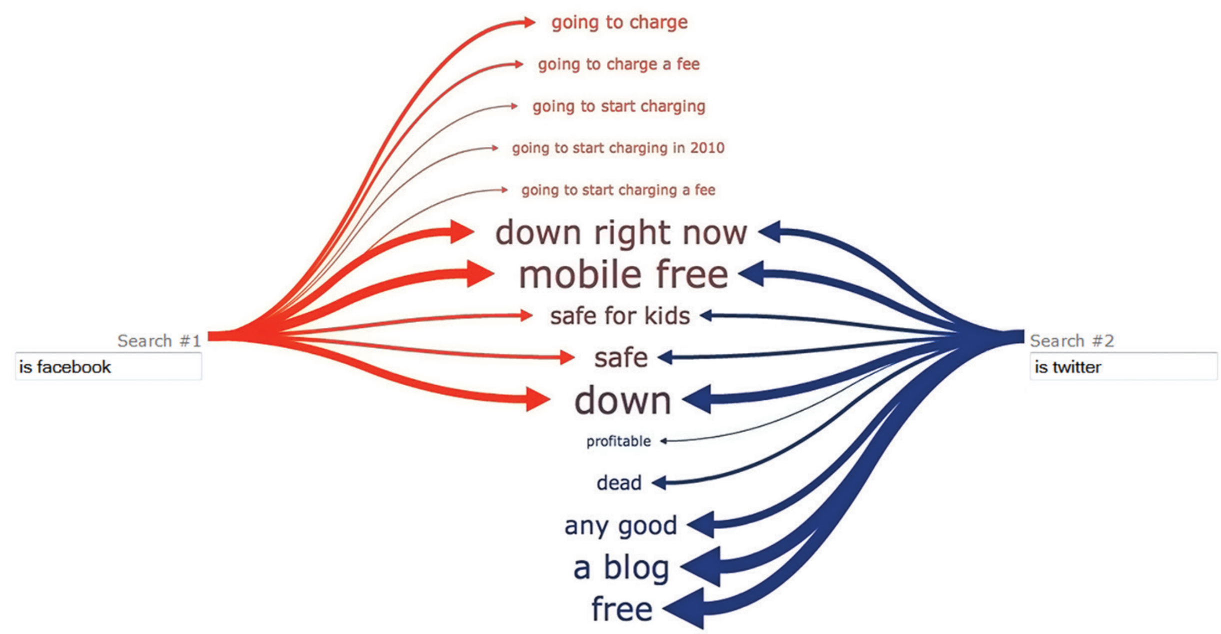
On Words—Concordance

This visualization shows the top ten words that occur most frequently in two books on American politics: *Lies and the Lying Liars Who Tell Them* (in blue) and *Slander: Liberal Lies About the American Right* (in red).



Is Facebook-Is Twitter Phrase Graph

This graph shows what people search for when using Google's autocomplete search function. One can enter two phrases to see how they are commonly completed. In this instance, apparently, both online services are frequently co-occurring with "mobile free" and "down," but only "is twitter" is associated with "free" while the phrase "is facebook" co-occurs with "going to charge."



Sentiment Analysis of the Bible

Several Bible translations were analyzed to compute a composite sentiment average for each given verse. Phrases like "I like X" were flagged as positive and colored in black, whereas phrases like "I hate Y" were flagged as negative and colored in red.

Editions of Darwin's *On the Origin of Species*

Six editions of Darwin's classic book are shown here. Each edition is color-coded, and text new to that edition is overlaid upon the book's chapters I through XV, arranged from left to right. The book size appears to have changed considerably—from approximately 150,000 words in the 1859 first English edition to about 190,000 words in the 1872 sixth edition. New phrases were also introduced (e.g., "survival of the fittest," introduced by British philosopher Herbert Spencer, didn't appear until the fifth edition).

Topical Visualization Types

Visualizations of text may address different insight need types (**page 26**), such as the topical structure and composition of text, trends, or relationships. Diverse visualization types (**page 30**) are used. Temporal visualization types, such as Alluvial graphs (**page 51**), can be used to communicate topical trends. Geospatial metaphors are frequently applied to depict topic spaces (see **page 52, Geospatial Studies—"Where"**). Network analysis and visualization are employed to communicate topical relationships (see **page 60, Network Studies—"With Whom"**).

Composition and Frequency

Topical composition and term frequency may be represented with lists, tag clouds, or pie charts.

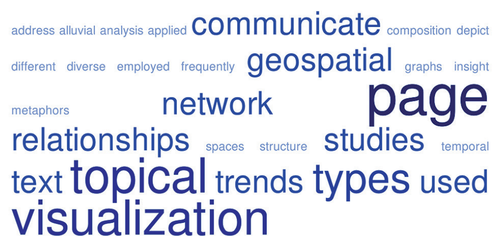
Lists

Texts such as book chapters or software code can be plotted, revealing the frequency of terms or the number and length of chapters or program modules (see **page 57**, *Sentiment Analysis of the Bible*). Jonathan Harris's *Wordcount* below arranges the 86,800 most frequently used English words side by side as one very long sentence, starting with the number one most frequently used word, "the," on the left. The red number below each term indicates its frequency rank.

Ben Fry's rendering of *Editions of Darwin's* On the Origin of Species (**page 57**) uses color-coding to highlight changes over time. Stephen Eick's highly interactive Seesoft interface (**page 69**) uses color, size, and other coding to highlight software code passages that match a search query, were modified, or were written by a certain programmer. Given a textual listing, linkages and networks can be overlaid; see the *Taxonomy Visualization of Patent Data* in the *Atlas of Science* (**pages 132–135**), which shows references and citations between patents and patent classes.

Tag Cloud

Tag clouds (also called word clouds) lay out words in a given area to provide a quick overview of content coverage. Words may be color- and size-coded by additional attributes, such as quantitative term frequency or qualitative type (e.g., part of speech, singular/plural, etc.; see **page 56, Tagging**). They may also be sorted alphabetically or by frequency. Shown below are the top 27 unique words that make up the opening paragraph of this page. Sorted alphabetically, they are also size-coded by the number of occurrences; "page" is thus rendered as the largest word, because it appears the most often (five times).



Words can be positioned according to size, with larger words set near the center and smaller words grouped closely around them for optimal use of space. The word cloud below, created with Wordle, shows the same text as the previous word cloud; however, the terms have not been alphabetized, and certain words have been capitalized per the original content.

Tag clouds can have arbitrary shapes (see *Visualizing Trends and Dynamics: 30 Years of Scientific Development* in *Atlas of Forecasts*). Words can also be hyperlinked, so that one may click on a word to highlight or retrieve all data records containing that word, in support of content navigation (see **page 68, Interaction**).

Graphs

The topical composition of a data record (e.g., a paper) or data set (e.g., all publications by an author, institution, or country) can be represented using bar graphs, radar graphs, or parallel coordinate graphs (see **page 46, Graphs**).

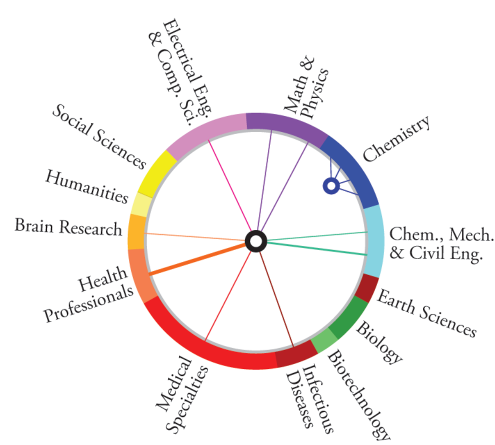
Structure

The topical structure of text can be represented using different reference systems that range from circular to geospatial to network graph.

Circular Graph

Raw text, categories, or classification hierarchies can be plotted in a circle or oval, creating a semantic space that can be used to plot other text. For example, the SciVal rendering of the *UCSD Map of Science and Classification System* (*Atlas of Science*, **page 13**) plots the 554 subdisciplines by grouping them within 13 color-coded disciplines, all arranged in a circle (see below). Institutions or countries that have many publications in a single discipline, such as **Chemistry**, will be plotted close to the periphery (see blue circle), whereas those that have broad coverage will fall in the middle (see the black circle in center). Size- and color-coding of nodes and edges can be used to depict other properties, such as the number of publications per subdiscipline (see **page 96**, *The Scientific Roots of Technology*, and **page 118**, *U.S. Vulnerabilities in Science*).

A network layout of the *UCSD Map of Science and Classification System* with data overlay can be seen in **Topical Locations of Papers** (**page 43**).



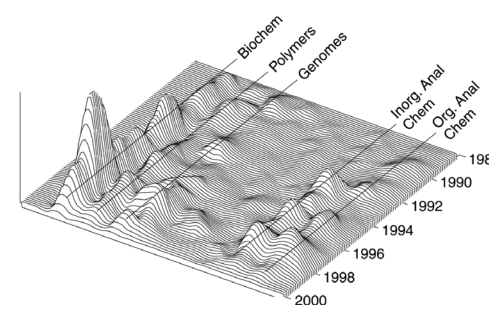
W. Bradford Paley's *TextArc Visualization of "The History of Science"* (*Atlas of Science*, **pages 129–131**) lays out text in a circular manner, creating a semantic space in which words can be placed.

GRIDL

The GRaphical Interface for Digital Libraries (GRIDL), developed by Ben Shneiderman and colleagues, uses a tabular reference system to plot text records such as publications (**page 69**). The x-axis is commonly used to plot temporal or spatial properties (e.g., years or author locations, respectively). The y-axis can display topic hierarchies; tree branches can be extended or collapsed. Drilling down into a subtopic corresponds to categorical zooming. Text records (e.g., papers) are placed in relevant grid cell(s)—either as a dot or, if there are too many, as a bar. Dots and bars can be color-coded to represent additional attributes. GRIDL is highly interactive (see **page 68, Interaction**).

Crossmap

Crossmaps plot papers, authors, and other scholarly entities in a two-dimensional space. They were developed by Stephen Morris and colleagues for technology forecasting. Axes can use categorical data (e.g., author names) or ratio data (e.g., publication years). Some are sortable (e.g., alphabetized or sorted by year) or organized according to a topical cluster hierarchy (see below). Once the axes have been created, data variables (e.g., scholarly papers) can be placed, and geometric symbols (e.g., circles) may be size- and color-coded (e.g., by the number of recent citations, as in the *Timeline of 60 Years of Anthrax Research Literature*; see *Atlas of Science*, **pages 94–97**).

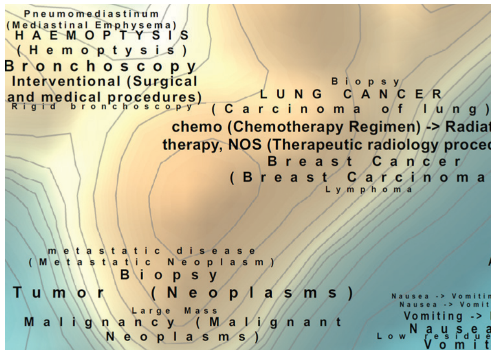


Jittering (i.e., adding small random values to the position values of graphic symbols) can be applied to reduce overlap. Geometric symbols (e.g., circles representing papers that are published on a given topic in a certain year) can be interlinked (e.g., to represent citation linkages). Additional graphical elements can be used to indicate cluster boundaries, research fronts, bursts, or external events.

Crossmaps can also be rendered as **Isoline Maps** to show the growth and decline of different topic areas over time.

Isoline Map

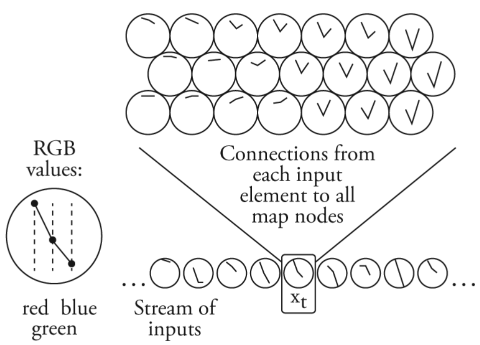
Isoline maps (**page 54, Geospatial Visualization Types**) can be used to depict the structure of topic spaces (see next page top-left or example on **page 134**, *Knowledge Cartography*). In this type of map, continuous lines join existing points that are of the



same value. Color-coding or height-coding can be employed to ease legibility (see figure in **Crossmap** section or the *GIS Map of White and McCain* in *Atlas of Science*, **page 34**).

Self-Organizing Map

Self-organizing maps (SOMs) use a two-dimensional output space to represent the main structure of a much higher dimensional semantic space. Shown below is a stream of input vectors of red, green, and blue (RGB) value triples. After training the SOM, each map node in the output space is represented by a model vector that is similar to the input vectors it represents; similar model vectors appear close to one another. The output space can be a grid of any size and shape, possibly wrapping around the edges. It may also be colored (e.g., to indicate the number of input vectors per output node) and clustered (see *In Terms of Geography* in *Atlas of Science*, **page 102**).



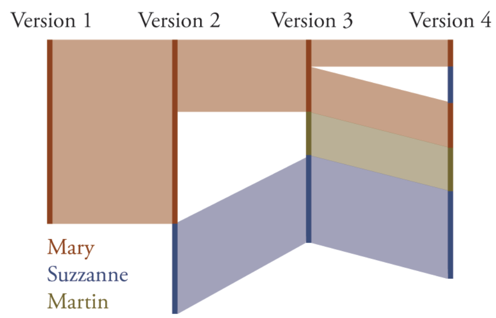
Shown on top-right is a cutout of a large-scale topic map generated using natural language processing, topic modeling, a self-organizing map, and GIS rendering of more than 11,000+ clinical admission records. Color-coding is used to indicate how focused different areas are (i.e., the degree to which a particular region in the display space is dominated by a limited number of topics as opposed to representing a broad mixture of topics). Green indicates a very low focus, yellow a medium focus, and brown a very high focus. The **Cancer** region, including lung and breast cancer, is highly focused, whereas the **Nausea** region in the bottom-right is less focused.

Trends

Different types of graphs are used to show the evolution of topics over time (see **page 48, Trends**).

History Flow

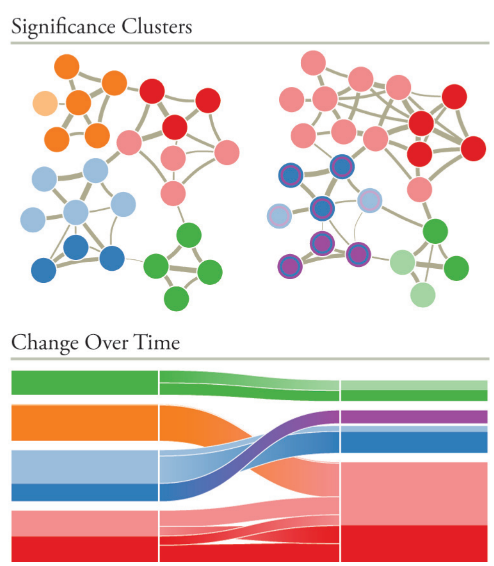
A history flow visualization depicts the revision history of a given text. An upside-down stacked line graph is used to show the existence and length of text chunks over time (see below). Time points or revision numbers run from left to right. Character or word counts determine the y-axis placement. Bands, representing text chunks, can be color-coded to indicate author, title, topic, and other subjects.

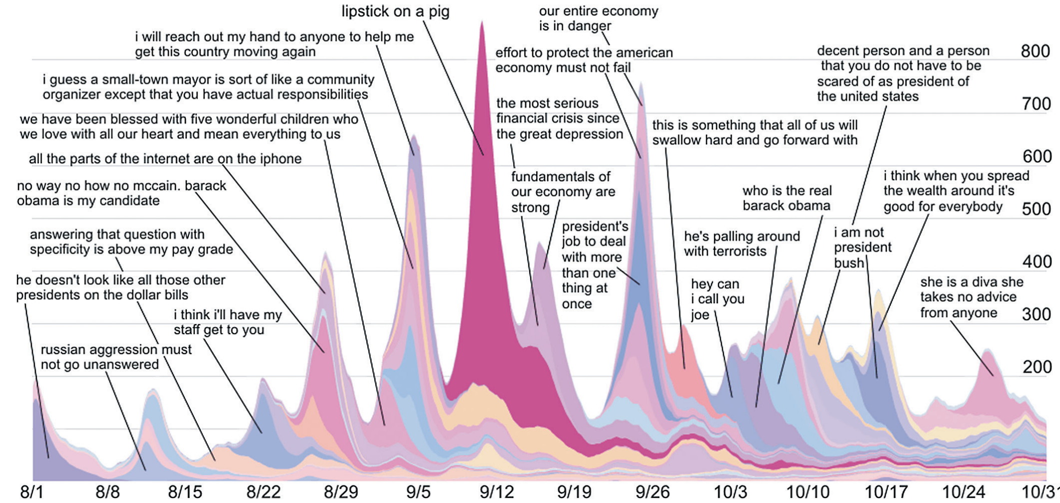


The *History Flow Visualization of the Wikipedia Entry "Abortion"* depicts multiple versions of the same Wikipedia entry (see *Atlas of Science*, **pages 124–127**).

Alluvial Graph

Alluvial graphs (also called alluvial diagrams) can be used to show the merge and split of topics over time. One example is the *Evolving S&T Landscape* (**page 16**, top-right). Another example appears below. At top, a network is shown for an earlier time (left) and a later time (right). Major topic clusters are color-coded. Changes in cluster size and composition

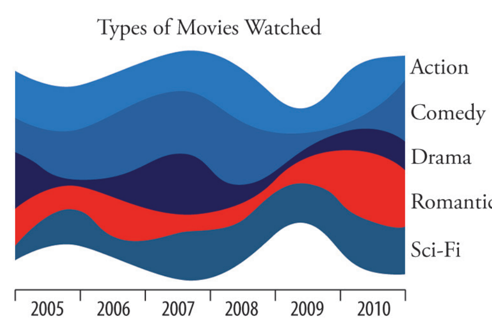




over time are then shown at bottom, using an Alluvial graph with colored bands that correspond to major colored clusters.

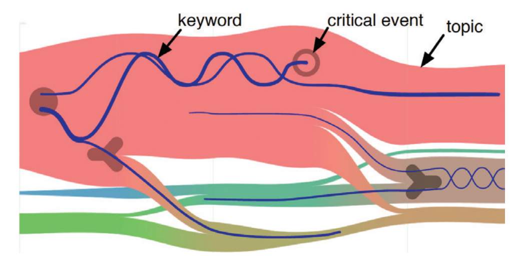
Stream Graph

A stream graph (also called a theme river) describes thematic changes in a set of texts. The metaphor of a river, flowing left to right through time, is used to indicate such change. Each topic is represented as a "stream" that narrows or widens to indicate how the strength of a topic decreases or increases in associated documents at any given point in time. The streams can be color-coded and labeled (see below); some are even interactive (i.e., by hovering over a stream online, one can bring up information on the texts that it represents).



Stream graphs can also be rendered as stacked line graphs for the purpose of improving legibility and easing comparisons. Text labels can be added to ease interpretation (see example at top-right of this page, which shows the top 50 most dominant news threads between August 1, 2008, and October 31, 2008; the thickness of each strand corresponds to each thread's volume over time).

Spaces between streams help to separate major topics. Additional text and symbols can be added to help interpret the evolution of topics (see the *TextFlow* visualization next column, top).



Relationships

Associations and dependencies between texts are frequently represented by links.

Arc Graph

Arc graphs (also called arc diagrams) can be used to represent structures in text strings, such as patterns of repetition (see also **page 31, Networks**). Examples include *Visualizing Bible Cross-References* (**page 150**) and *The Shape of Song* that uses arcs to connect repeated sections of music with translucent arcs as shown below. The height of an arc can be used to represent attributes other than distance. Color- and weight-coding of linkages (above and/or below a vertical or horizontal line) can be employed to communicate additional attributes.

Networks

Network layouts are widely used to depict topic spaces. The *Map of Information Flow* (**page 9**, lower-right) uses directed, size-coded linkages to depict citations between major areas of science.

Network Studies—“With Whom”

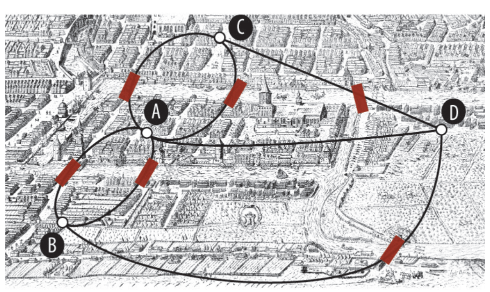
Network analysis and visualization techniques answer “with whom” questions, such as “Who collaborates, likes, or competes with whom?” When using network science approaches, the first step is to represent data by using nodes and edges. Nodes may have diverse attributes; edges may have labels, weight (e.g., to signify similarity or strength), and direction. Next, network analysis can be applied to identify clusters of similar nodes or backbones over which major traffic flows. In the process of visualizing a network, nodes, edges, and attributes must be mapped to graphic symbol types (**page 32**) and their graphic variable types (**page 34**).

Network: Any thing reticulated or decussated, at equal distances, with interstices between the intersections.

Samuel Johnson

Seeing Networks

In 1735, mathematician Leonhard Euler solved the Königsberg bridge problem using a network approach that is now considered to be the first theorem of graph theory. He reformulated the problem as a network graph in which unconnected land masses in the city of Königsberg are visualized as nodes (labeled A to D in below map). Those nodes are linked by edges that represent the seven bridges of Königsberg (gray bars). Using this approach, Euler proved that there is no continuous walking path (i.e., in order for all seven bridges to be crossed, some paths must be retraced).



In social network analysis, network nodes commonly denote people. Diverse relationships (ties) are studied, including similarities (e.g., share same spatial space or temporal space); memberships (e.g., same group or activity); attributes (e.g., gender or attitude); social relations (e.g., family membership, friendship); affective ties (e.g., loves or hates); cognitive ties (e.g., one knows the individual or knows about him/her); interactions (e.g., has talked to, helped, or collaborated with individual); and flows (e.g., knowledge or resources).

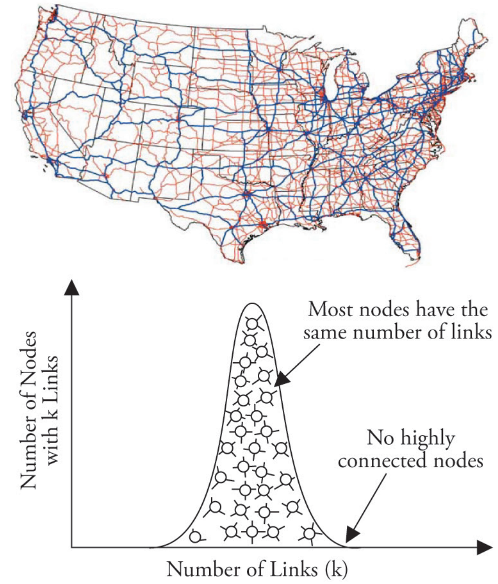
In S&T studies, different network relations between individuals, institutions, countries, etc., are studied, such as communication relations (e.g., who talks to whom), instrumental relations (e.g., who asks whom for expert advice), boundary penetration relations (e.g., who is on whose board of directors), sentiment relations (e.g., friendship cliques), power relations (e.g., who follows whom), kinship relations (e.g., who is related to whom), and transaction relations (e.g., who gives gifts to whom). A listing of books that provide a general introduction to network science and examples can be found in **References & Credits (page 178)**.

Network Extraction

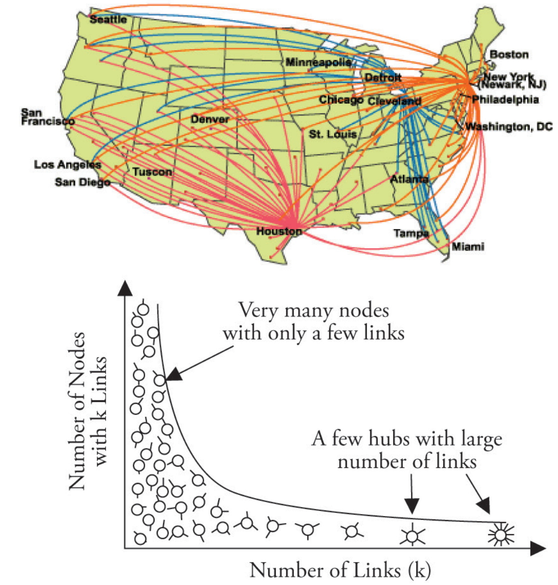
In the study of S&T, common network nodes (or units of analysis) are authors, institutions, and countries as well as words, papers, journals, patents, and funding awards (see *Atlas of Science*, **page 54**, **Conceptualizing Science**). Nodes of the same type can be interlinked via different link types (e.g., papers based on topical similarity or on citation linkages, such as co-citation or bibliographic coupling). Nodes of different types can also be interlinked (e.g., author–paper or paper-funding networks). The resulting networks may be either directed or undirected, weighted or unweighted, labeled or unlabeled.

Network Types

Different types of networks exist, all with markedly different properties. Key types are reviewed here.



U.S. Street Network with Gaussian Distribution



U.S. Airline Network with Power-Law Distribution

Tree Graph

A *tree graph* (also called a connected forest) is a simple, connected, undirected, and acyclic graph. A tree with n nodes has $n\text{-}1$ edges. In rooted trees, all nodes except for the root node have only one parent node. Nodes that have no children are called leaf nodes. All other nodes are referred to as intermediate nodes. Organizational charts and classification hierarchies have a tree structure.

Network Graph

Three different types of network graphs are commonly distinguished: random, scale free, and small world.

Random networks are formed by taking a set of isolated nodes and randomly adding successive edges between them (see the *U.S. Street Network with Gaussian Distribution*, above-left).

Scale-free networks have an uneven distribution of connectedness, whereby most nodes have few connections and few nodes are “highly connected” hubs (see the *U.S. Airline Network with Power-Law Distribution*, above-right).

Small-world networks have a high local clustering coefficient and a low average path length. Many scholarly networks (e.g., coauthor and paper-citation networks) have small-world properties.

Network Analysis

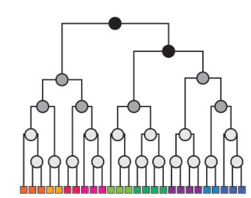
Many different network analysis approaches exist (see **page 178**, **References & Credits**).

Node and Edge Properties

Major node properties comprise node degree and reachability (e.g., as measured by betweenness centrality). Major edge properties include durability (how long they last), reciprocity (whether a relationship is mutual), intensity (whether edges are weak or strong), and quality (reliability or certainty).

Clustering

Clustering of network graphs (also called community detection or graph partitioning) is used to identify clusters of nodes that maximize both within-group homogeneity and cross-group heterogeneity (see example below). When clustering networks, it is important to note that clusters have a high internal density of links, whereas the number of links between clusters is comparatively low.

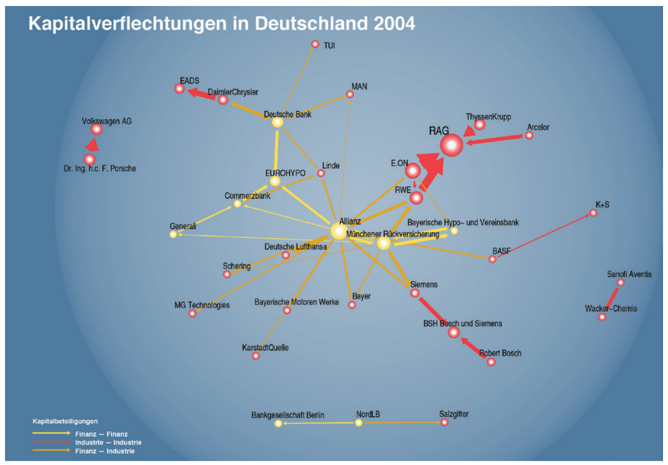


Backbone Identification

Many real-world networks are dense, which means there is a high ratio of the sum of all existing edges to the sum of all potential edges (i.e., when all nodes are interlinked). That makes it difficult to identify main “traffic highways.” Backbone identification algorithms use node, link, or network attributes to identify those parts of a network that handle the major traffic and/or have the highest-speed transmission paths. The algorithms identify and delete superfluous edges, keep the highest weight edges per node, or calculate the minimum spanning tree (see the bold edges in the example on **page 63**, lower-right).

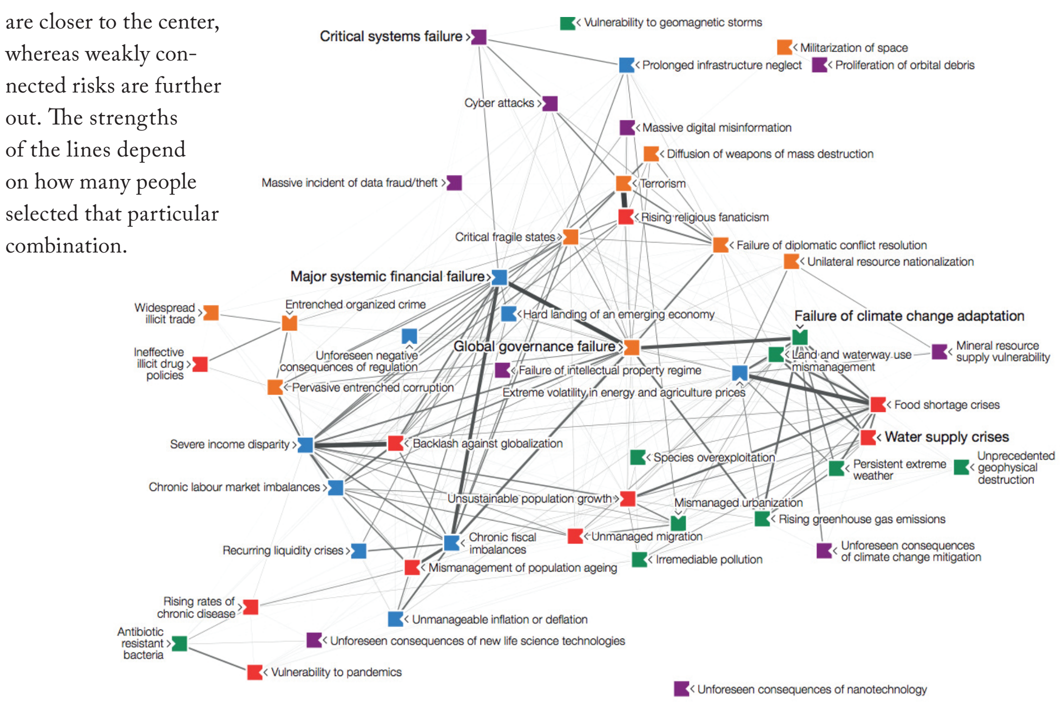
Kapitalverflechtungen in Deutschland

This network visualization by Lothar Krempel of the Max Planck Society in Germany shows the evolving network of leading companies in Germany. Yellow nodes signify banking and insurance companies; red nodes signify industrial companies (e.g., airline, automotive, and manufacturing firms). Node size denotes the volume of shared linkages. Yellow lines are used to link financial companies; red lines, industrial firms; and orange lines, financial companies to industrial firms. It reveals how during that time frame a drastic reduction was observed in the number and volume of linkages.



The Risk Interconnection Map, 2013

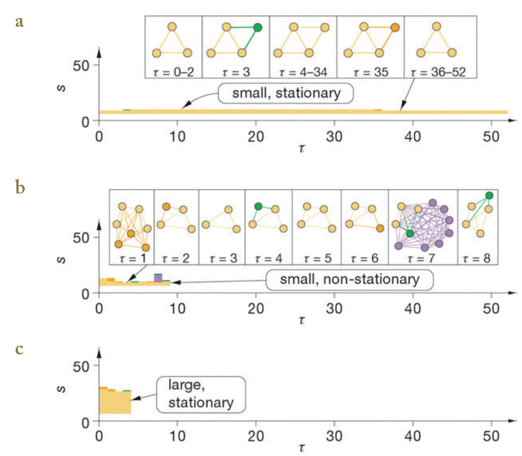
The World Economic Forum conducted a survey that asked experts to identify a minimum of three and a maximum of ten connections between major risks. The resulting network includes 529 paired connections. More connected risks are closer to the center, whereas weakly connected risks are further out. The strengths of the lines depend on how many people selected that particular combination.

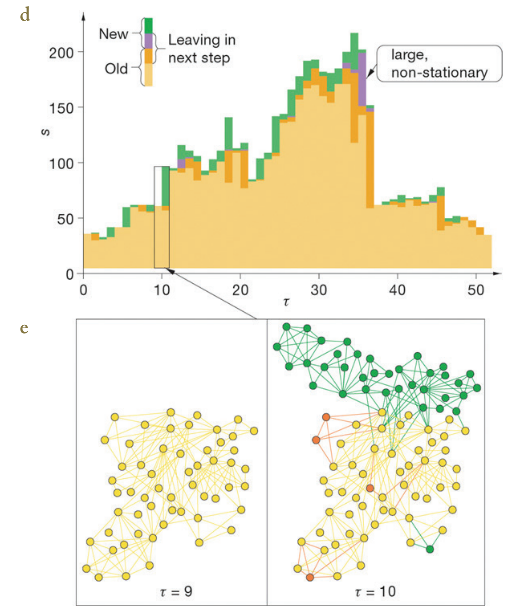


Quantifying Social Group Evolution

Gergely Palla and colleagues studied coauthor and phone-call "collaboration events." Shown are coauthorship events extracted from publications in the arXiv e-Print condensed matter archive, published over 142 months by over 30,000 authors. The stacked bar graphs show community composition per time step. Four types of author nodes are distinguished: those who joined in a previous time step (yellow), current newcomers (green), those who joined previously but will leave in the next time step (orange), and those who joined for this one time step only (purple). Collectively, the number of all nodes is represented by bar height.

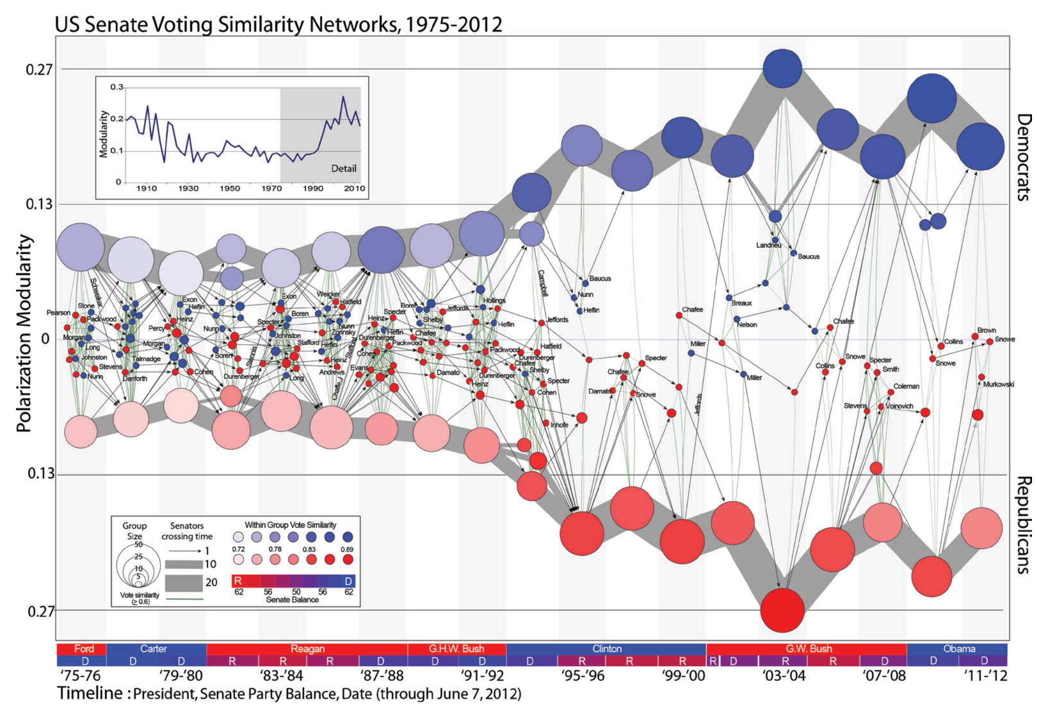
Shown at left is the evolution of three communities: (a) small and stationary; (b) small and nonstationary; and (c) large and stationary. Shown at right is (d) a large, nonstationary community and (e) network structures for two time steps.





U.S. Senate Voting Similarity Networks, 1975–2012

Using U.S. Senate voting data from 1975 to 2012, this timeline by James Moody and Peter Mucha shows the increasing political polarization in America. Over time, fewer and fewer senators occupy a middle ground outside of their party's camp.



Network Visualization Types

Network visualization algorithms should be selected according to layout optimality criteria. These include the visibility of all nodes, their links, and their labels; the countability of every node's degree; the ability of every link to be followed from source to destination; minimal numbers of link crossings; links having more or less the same length; large angles between incident or crossing lines; observable outliers, clusters (subnetworks), and backbones; and easy navigation/interaction. Furthermore, in a given sequence of networks (e.g., when animating change over time) the layout should ease comparison with respect to the layout of the previous network in the sequence. Data-driven criteria comprising layout distances between node pairs should reflect the similarity/distance values between those node pairs; variations in node density should reflect varying structural network cohesion; and geometric symmetries should reflect structural symmetries. Note that in network graphs, however, empty space does not signify the absence of phenomena.

Tree Layout

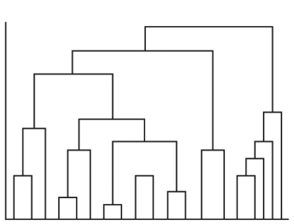
Many data sets, such as hierarchies (e.g., file systems, organizational charts), branching processes (e.g., genealogy, phylogenetic trees), and decision processes (e.g., search trees), have a tree structure. When depicting a hierarchy or tree, it is important to show the number of children per node, the tree depth (i.e., the number of edges from the root node to the leaf nodes), and the overall tree size.

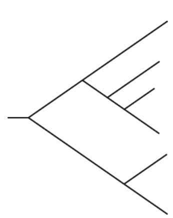
Tree View

Trees can be represented as lists of tree node labels interconnected by curved lines (see examples on **page 31** and **page 57**, *Is Facebook–Is Twitter Phrase Graph*).

Dendrogram

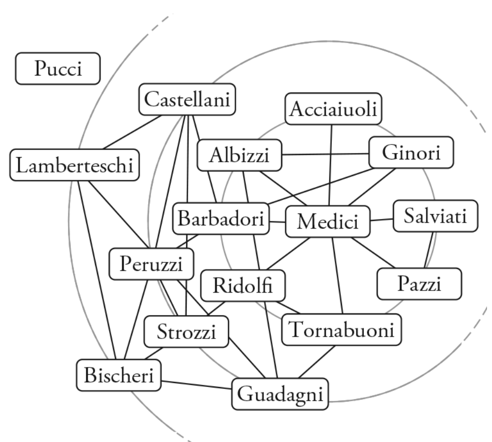
Tree structures can also be represented by dendrograms, which may be displayed vertically or horizontally and may be rectangular (see below-left) or slanted (see below-right). The *Timeline of 60 Years of Anthrax Research Literature* crossmap (*Atlas of Science*, **pages 94–97**) uses a dendrogram to depict the hierarchical clustering of topics.





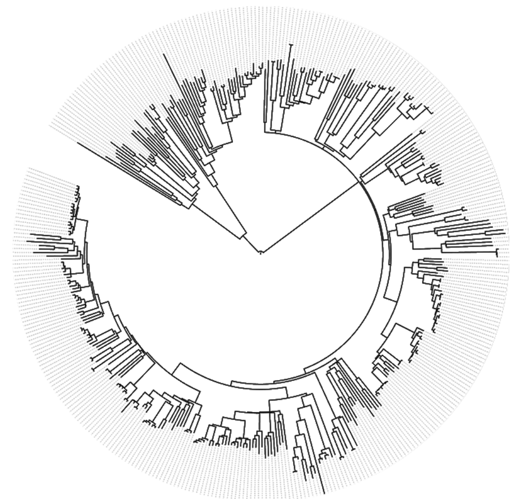
Radial Tree

A radial tree places all nodes in concentric circles, which are focused in the center of the screen. The nodes are evenly distributed, and the branches of the tree do not overlap. Shown below is the network of marriage and business ties of elite fifteenth-century Florentine families. The layout focuses on the **Medici** family, which had a uniquely central place in the network. Other families are placed in three concentric circles according to the number of links needed to reach the center node. **Pucci** is unconnected; see force-directed layout of the very same network on **page 27**, lower-right.



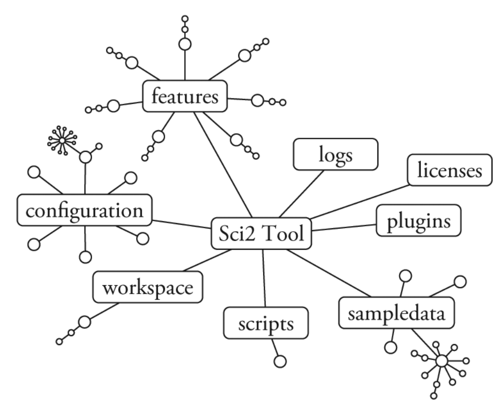
Link Tree

Link trees place the root node in the middle of a circle and intermediate and leaf nodes in concentric circles (see the tree ahead, extracted from the *Tree of Life* phylogeny on **page 124**).



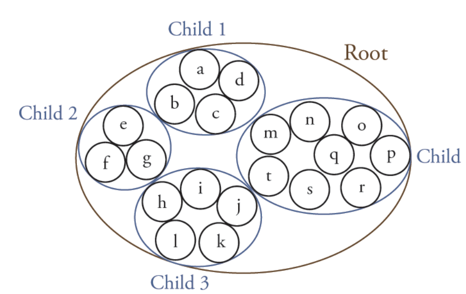
Balloon Tree

A balloon tree places child nodes in concentric circles around their parent nodes. Node size is adjusted as needed to reduce visual clutter. The graph below shows the directory structure of the Sci2 Tool, which was used to generate many of the network visualizations featured in this spread. The main directory can be seen in the middle; it is linked to labeled subdirectories, which are further linked to sub-subdirectories.



Enclosure Trees

Enclosure trees (also called circle packings) show the nesting of nodes using ellipsoids (see below).

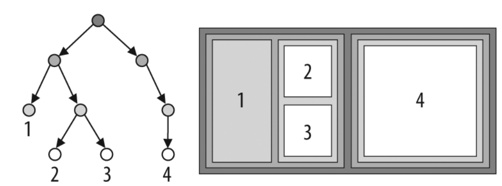


Mosaic Graph

Mosaic graphs were used as early as 1874 to represent contingency tables. They are a combination of 100 percent stacked column graphs and 100 percent stacked bar graphs (see **page 46, Statistical Visualization Types**), which make groupings and relative sizes visible.

Treemap

Treemaps extend mosaic graphs to represent deeper tree structures. They are a space-filling form of enclosure trees. Given an area, they use a space-filling recursive subdivision to lay out a tree structure without producing holes or overlaps (see the example below and also the *Map of the Market* in *Atlas of Science*, **page 200**). Area sizes may correspond to the attributes of the subtrees they represent. The same area size can have different manifestations in terms of aspect ratio. Areas may be labeled, color-coded, and shaded (see **page 34, Graphic Variable Types**). Originally developed for rectangular areas, treemaps can also be generated for circular or arbitrarily-shaped areas using Voronoi tessellations (see **page 45**, *All of Inflation's Little Parts*).



Network Layouts

Many different types of network layouts exist. Some are deterministic (i.e., each run of an algorithm results in the same layout); others are nondeterministic (i.e., running an algorithm on the same data twice tends to result in different layouts). Some have a well-defined reference system (e.g., nodes are sorted and plotted according to a given attribute value). Other layouts optimize node distances according to similarity relationships between nodes while minimizing edge crossings.

Adjacency Matrix

An adjacency matrix (also called a matrix diagram; a reorderable matrix; or a sociomatrix in social network analysis) represents which nodes in a network are adjacent to other nodes. It is a matrix with rows and columns labeled by nodes and with each cell representing the value of a dyadic variable or link. Values in the diagonal denote self-links. For an undirected graph, the adjacency matrix is symmetric (i.e., it is sufficient to display only values above or below the diagonal, also called the upper or lower triangle; see

page 49, *In Investing, It's When You Start and When you Finish*).

The layout is deterministic and easy to read. The matrix rows and columns may be reordered to improve legibility. Blockmodeling reorders the matrix so that the elements of a block are made contiguous (i.e., cohesive groups form contiguous intervals); a special case is that of the partition of a network in a cohesive core and a loosely connected periphery.

Different graphic symbol types (**page 32**) and their graphic variable types (**page 34**) can be used to encode additional link attribute values. For example, in a paper-citation adjacency matrix that shows inter-institution citations and represents the number of citations via height, highly cited and citing institutions are easily identified as vertical and horizontal bands. Similarly, the large number of self-citations (the diagonal) is easily spotted.

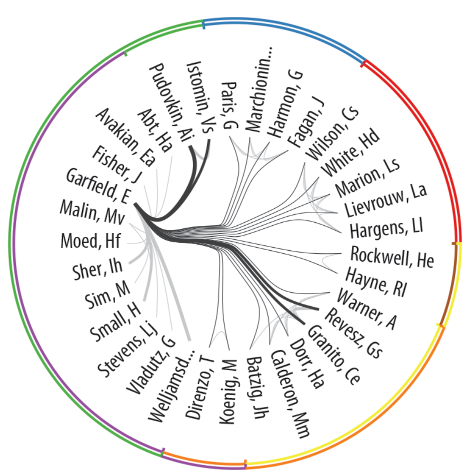
Arc Graph

See **Arc Graph** (**page 59**).

Circular Graph

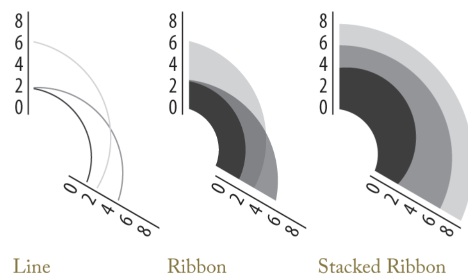
Circular layouts place all nodes in a circle, typically with spacing that is equidistant or driven by data (e.g., more similar nodes might be in closer proximity). To help find nodes and reveal structure, nodes can be sorted by node attributes (e.g., alphabetically for labels or numerically by quantitative attribute values). Node and link attributes can be represented via different graphic symbol types (**page 32**) and their graphic variable types (**page 34**). Examples include *Europe's Who Owes What to Whom* (**page 10**, lower-right); the *Skitter Internet Map* (**page 53**); and *Inter-Institutional Collaboration Explorer* (**page 61**).

Circular layouts can also be used to visualize hierarchical networks. In the coauthor network for Eugene Garfield (see below), all author nodes are placed on a circle and connected by coauthor links that run through the circle's interior. Edge bundling was applied to improve the legibility of links (see *Atlas of Science*, **page 161** on **Flow Maps**). The colored arcs indicate how author nodes cluster hierarchically according to Blondel community detection (**page 60**, **Clustering**).

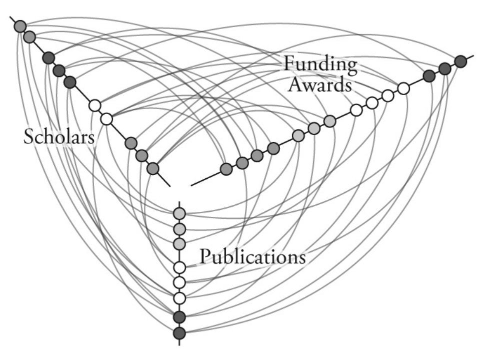


Hive Graph

Hive graphs (also called hive plots) resemble parallel coordinate graphs (**page 47**) but use a radial reference system to place nodes on axes according to their attribute values. Line, ribbon, and stacked ribbon hive graphs are shown below.

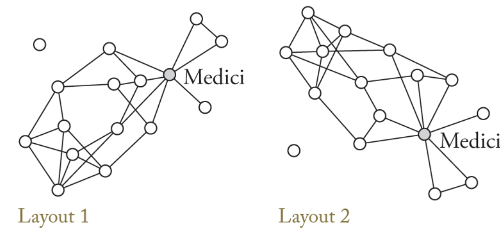


In circular hive graphs, each of the axes supports two graphs (on either side). The three axis-line graph below depicts a set of **Scholars** who have **Publications** or **Funding Awards** and how these **Publications** acknowledge the **Funding Awards**. Additional node and edge attributes (e.g., color or size) can be used to encode attribute values.



Node-Link Graph

A node-link graph (also called a network diagram, structure plot, or sociogram) uses nodes and edges to represent a network. Nodes and edges may have additional attribute values that can be encoded using graphic variable types (**page 34**). Network layout may be random, circular, orthogonal (as in subway maps, see **page 55**), hierarchical, sorted by time (see the *HistCite Visualization of DNA Development* in *Atlas of Science*, **pages 120–123**), sorted by node properties (e.g., node degree or betweenness centrality; see circular graph on left), radial (see **Radial Tree** on opposite page), or force directed. The latter places nodes according to their similarities or the distances between them, aiming to minimize edge crossings while still maintaining their relative positions. Most force-directed layout algorithms aim to reduce the inherent stress, but they are nondeterministic (i.e., each layout results in a slightly different solution with a similar placement of nodes; see different layouts of the Florentine network below).



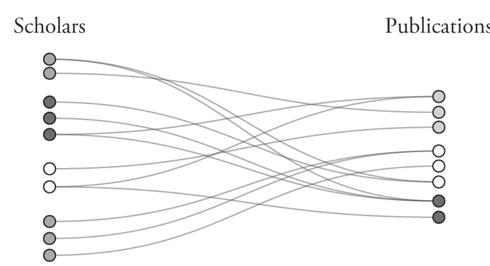
When working with large networks, it is beneficial to extract the most important nodes and edges and to identify and visually highlight important "landmark" nodes, subnetwork communities, and backbones.

Sankey Graph

Sankey graphs (also called Sankey diagrams) show the flow of resources between nodes in a network with line width representing flow magnitude (**page 49**). Like flow maps (**page 51**) and Alluvial graphs (**page 59**), Sankey graphs bundle lines to reduce visual clutter. Sankey graphs differ from Alluvial graphs in that they ignore temporal ordering.

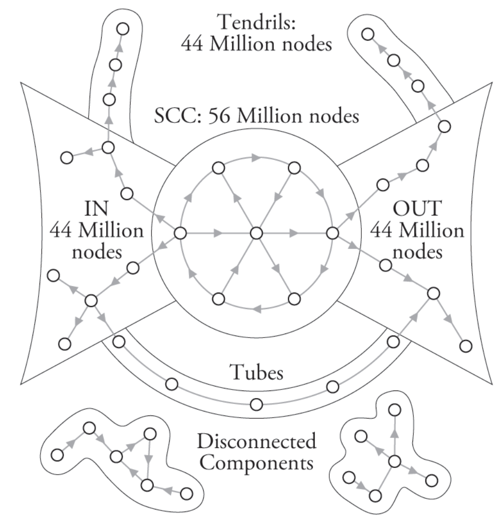
Bimodal Graph

Bimodal network graphs (also called two-mode networks) contain two types of nodes and are commonly represented by two (sorted) lists, which are interconnected by linkages. For example, the **Scholars** and their **Publications** shown in the circular hive graphs on left can also be depicted in the bimodal network graph below.



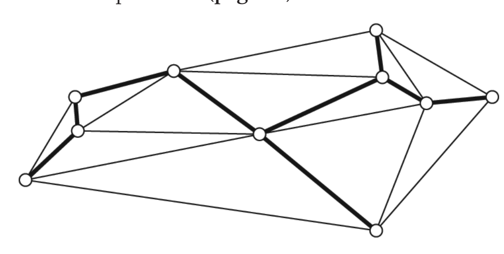
Conceptual Drawings

Large-scale networks are commonly composed of one giant, weakly connected component and other disconnected components. The core-periphery structure of the giant component resembles a bow tie. The original drawing from 2000 shown below, by Andrei Broder and colleagues, identified that the World Wide Web has about 44 million **IN** nodes, 56 million nodes in the strongly connected component (**SSC**), and 44 million **OUT** nodes. One can pass from any node of **IN**, through **SCC**, to any node of **OUT**. Hanging from **IN** and **OUT** are **Tendrils** containing nodes that are reachable from portions of **IN** or that can reach portions of **OUT** without passage through **SCC**. **Tubes** refer to passages from a portion of **IN** to a portion of **OUT** without traversing **SCC**.



Network Overlays

Network overlay maps (also called substance-based layouts) overlay networks on existing reference systems, such as graphs (see **page 59**, **Arc Graph**), geospatial maps (see **page 19**, right), topic maps (see *Taxonomy Visualization of Patent Data* in *Atlas of Science*, **pages 132–135**), or images (e.g., a photograph of a brain cross-section with names of neuroscience authors in the brain sections they study and interlinked by coauthor relations or a satellite image in which the names of key institutions are placed and linked). They can also be used to highlight the backbone, i.e., major structure of a graph (see below). Using interactivity, it becomes possible to refocus on different nodes via zoom and selection operations (**page 68**).



Studying Dynamics

Dynamic analysis and visualization can be applied to detect change over time, but change in other attribute values can be studied as well. When using data from longer time spans, evolving data formats may have to be harmonized. Major changes in the number of data records per time slice may require adjustments in parameter values to maintain legibility. Note that change may also affect the target user group and their insight needs (see **page 40, User Needs Acquisition**). Plus, tools and workflows change over time. This spread features different types of dynamics together with dynamic visualizations that communicate change via modification of the reference system as well as evolving data overlays, graphic symbol types (**page 32**), and/or graphic variable types (**page 34**).

May you have the hindsight to know where you've been, the foresight to know where you are going, and the insight to know when you have gone too far.
Irish Blessing

Types of Dynamics

Dynamics may come from changes in data variable values, as well as from changes in the types and numbers of measured or derived data values, variables, or linkages. In addition, reference systems may evolve. Four types of dynamics are discussed below.

Data Values

Most data variable values examined in S&T studies change as time progresses. Examples are the number of papers per authors, the number of citations per paper, or the number of faculty at a university. Over time, data values may increase beyond expectation, making an adjustment of the reference system necessary. For example, in a timeline graph, the time axis may need to be expanded or its scale changed to display all the values that come into existence over time (see **Reference Systems** section).

Data Variables

New data sets and updated data formats come into existence on a daily basis. The alignment of data formats (or taxonomies, classification systems, ontologies) that changed over time can be time consuming or even impossible. Yet, the availability of a more precise time stamp or geolocation, unique author and institution identifiers, and/or linkages to other data sets is likely not only to increase the quality of existing visualizations but also to make novel analyses and visualizations possible.

Data Records and Linkages

As time passes, new data records and linkages are published (e.g., new papers are published or new coauthorship relations develop). This raises the question of whether they should be visualized separately or cumulatively for each time slice, or if a sliding time window should be used (see **page 48, Time Slicing**).

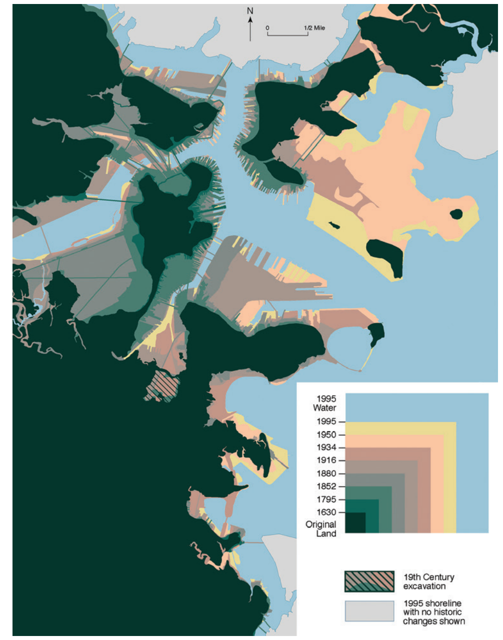
Reference Systems

Reference systems may need to be updated in response to increases in data values (e.g., a highly cited paper may require extending the data range of a graph). Data density may rise to the point where too much occlusion renders the visualization illegible. Applying alternative analysis algorithms may then be required to extract and highlight key structures and trends. In addition, the reference system may need to be distorted to make more space for densely populated areas. Geospatial maps may evolve in response to external events (e.g., see **page 16**, *Political Borders of Europe from 1519 to 2006*). The map on this page shows the substantial increase of Boston's land area and change in shoreline due to landmaking projects between 1630 (dark green) and 1999. The more recent the changes are, the brighter the colors become. Land added from 1950 to 1995 is shown in bright yellow.

Network graphs, when used as reference systems for multiple time frames, should correctly represent the structure that is inherent in the data. However, maps from consecutive time frames should provide, as much as possible, a "stable reference framework"—a nontrivial goal, because network layouts tend to change drastically over time.

Presentation Types

The preceeding four different types of dynamics can be presented using four general approaches,



effectively generating a four-by-four matrix. Combinations of types are possible.

One Static Image

Static images (e.g., those printed in newspapers or scientific journals) are a common format to share visualizations (see **page 70, Device Options**, for the advantages and disadvantages of using this format). Temporal graphs are used to show changing properties or derivative statistics (see **page 50, Temporal Visualization Types**). If location data is two-dimensional, then a simple arrow or trail can be used to show change over time (see the *Gapminder Visualization* on the opposite page as well as *Hurricanes & Tropical Storms—Locations and Intensities since 1851* in *Atlas of Forecasts*). Proportional symbol encoding can also be used, for instance, to show the amount of time a user is idle in the virtual world (see idle circles in ***Virtual World User Activity*** on opposite page).

Multiple Static Images

Dynamic change can be represented using multiple static images; see evolving activity patterns and flow and network overlays in the following geospatial maps: *Europe Raw Cotton Imports in 1858, 1864 and 1865* (**page 80**); *Mobile Landscapes: Using Location Data from Cell Phones for Urban Analysis* (**page 108**); and *Literary Empires: Mapping Temporal and Spatial Settings of Victorian Poetry* (**page 136**). Alternatively, evolving networks (see *Maps of Science: Forecasting Large Trends in Science* in *Atlas of Science*, **pages 170–173**) or small multiple displays (**page 66**) can be used.

Different visualization panels are typically arranged in proper (temporal) sequence: from left to right or from top to bottom. Visual pathways may be suggested by using arrows or narratives. If the visualizations share a common reference system and the same mapping of data variables to graphic symbol types (**page 32**) and graphic variable types (**page 34**), then one legend suffices; if not, then multiple legends are needed.

Animations

The rapid display of any sequence of images can be used to create an illusion of continuous movement known as animation. When combined with continuously updated legends and accompanied by text, this can be an effective way to communicate dynamic data. Most animations can be started, stopped, fast-forwarded, or rewound interactively. They may use either a fixed or evolving reference system, as discussed below.

Fixed Reference System

An animation that uses a static reference system is comparable to a flipbook of small multiple displays. Usually, the animation frames are generated by first mapping the full data set, then saving the record/node positions, and finally using those positions for all earlier time frames. An example are the four science map overlays in *TTURC NIH Funding Trends* on opposite page. Using the *UCSD Map of Science and Classification System*, publications from four cumulative time frames were overlaid to communicate the topical focus and number of TTURC publications. Care must be taken to ensure that the visual encoding, specifically the size-coding of graphic symbol types, does not extend beyond the available canvas.

Evolving Reference System

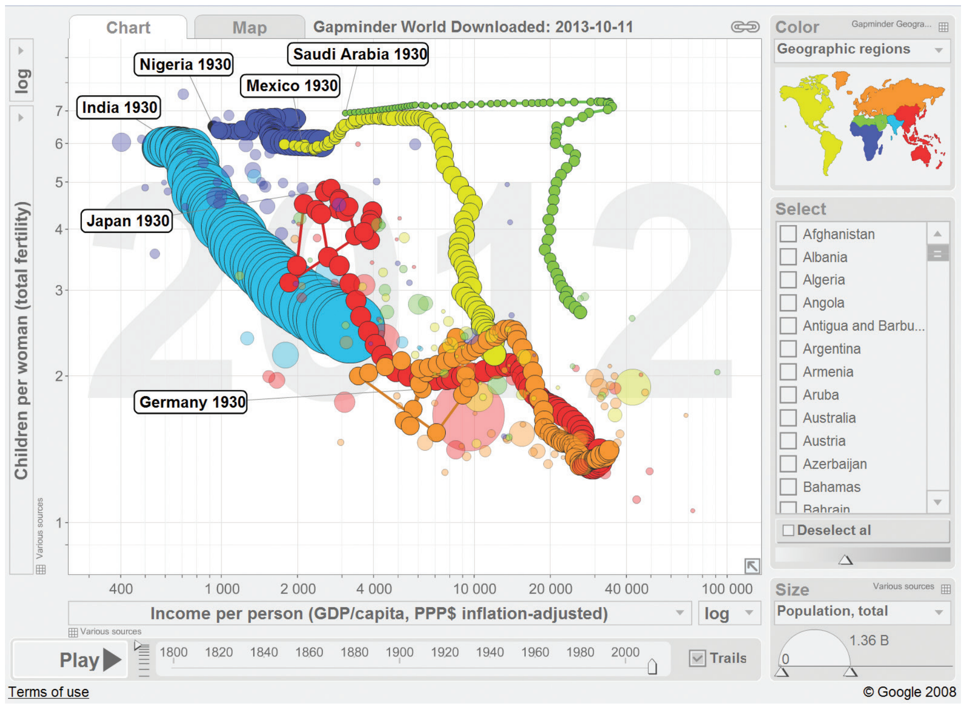
Some animations, such as network layouts, readjust the reference system so that it is correct or optimal for each time point. Examples are the evolving coauthor networks in *TTURC NIH Funding Trends* (opposite page) and the evolving journal citation maps in *The Emergence of Nanoscience & Technology* (**page 138**). Whereas geospatial maps evolve due to external events—such as wars that change country boundaries or droughts that dry up lakes—changes in network structures are typically data-driven.

Interactive Visualizations

Dynamic visualizations can be explored via desktop or online interfaces that support data exploration (see **page 26, Interaction Types** and **page 68, Interaction**). Different reference systems and views of the data may be selected, and overview, filter, and details-on-demand functionality may be provided.

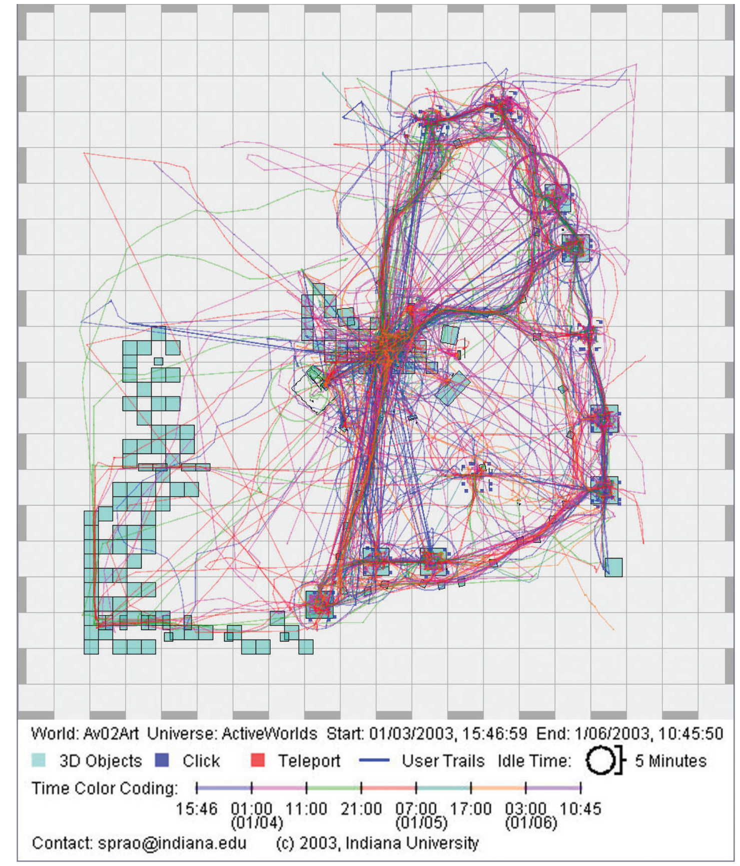
Gapminder Visualization

Hans Rosling's Gapminder visualizations aim to communicate socioeconomic data to a general audience (see also **page 71**, *200 Countries, 200 Years, 4 Minutes*). The below **Children per Woman** over **Income per Person** scatter plot maps countries—represented by a circle that is size-coded by **Total Population**—for the years 1930 to 2012.



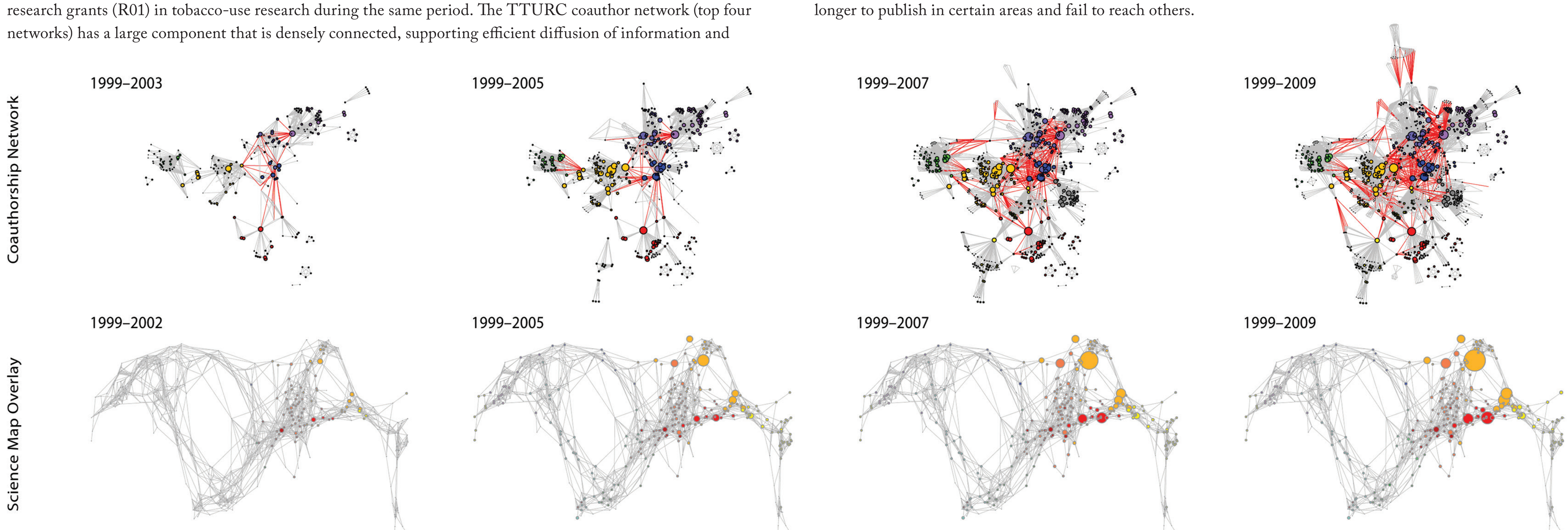
Virtual World User Activity

Shown here is a square educational world from the Active World universe of virtual worlds. Buildings are cyan; user trails are color-coded by time. Stationary users are represented by circles that grow in size as idle time increases.



TTURC NIH Funding Trends

This study examined the impact of different funding strategies by the National Institutes of Health (NIH) comparing transdisciplinary tobacco-use research centers (TTURC) versus traditional investigator-initiated research grants (R01) in tobacco-use research during the same period. The TTURC coauthor network (top four networks) has a large component that is densely connected, supporting efficient diffusion of information and expertise. R01 networks (not shown) are sparsely connected or not at all. TTURC research publications (bottom science maps) quickly cover all major areas of science, whereas comparable R01 publications (not shown) take longer to publish in certain areas and fail to reach others.



Motivation
Framework
Acquire
Analyze & Visualize
Deploy
Interpret

Combination

Most data sets are complex, and different types of analysis and visualization may be required to make sense of them. Presenting the same data using different reference systems, such as temporal, geospatial, and topic space, that are coupled (i.e., selecting a data record in one view highlights the same record in all other views) makes it possible to examine different aspects simultaneously. Showing different data in the same reference system, such as by using population pyramids or small multiples, eases comparisons. Multilayer visualizations can be employed to provide focus and context or to support navigation across multiple levels of abstraction—from micro to macro.

The whole is greater than the sum of its parts.
Aristotle

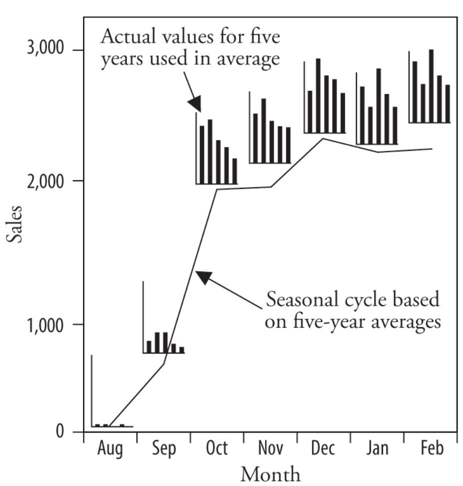
Multiple Views

Many old maps use multiple panels to tell a comprehensive and often global story (see *Heights of the Principal Mountains in the World, Lengths of the Principal Rivers in the World* and *Zoological Geography* on the opposite page). The spatial attributes of real-world objects (e.g., the concrete positions of mountains or the trajectories of rivers) can be distorted to make specific properties easier to read and compare (e.g., the height of mountains or the length of rivers).

Computer-generated graphs may plot data points and their distributions in one visualization. For example, scatter plots may be combined with stripe graphs to create so-called rug plots (**page 47**, lower-left). The *Ecological Footprint* (**page 90**) features a world map and several graphs to communicate consumption and pollution for different countries. Interactive online visualizations feature multiple windows that provide different views of scholarly data. For instance, *Knowledge Cartography* (**page 134**) provides a timeline, a geographic map, a thematic map of disciplines, and a map of collaborations. For a discussion of tightly coupled windows that support interactive data exploration, see **page 68, Interaction**.

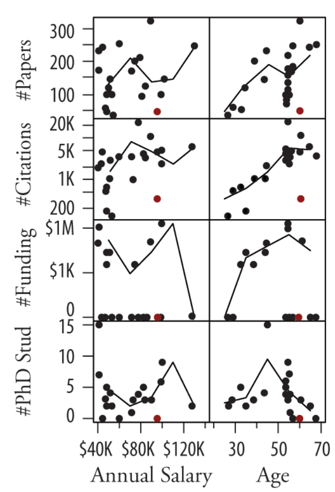
Small Multiples

Small multiples are sets of thumbnail-sized graphics of multivariate data that are frequently used in comparisons. They use the same measures and scale and may be used in a tabular display, a graph, a map, or any other visualization type (**page 30**). An example is the line graph in the lower-left from Harris's *Information Graphics* book, which shows sales from August to February using five-year averages. For each month, the values of all five years are plotted using a miniature bar graph. The map on the right of it shows small multiple graphs on a map. *The Baby Name Wizard* (**page 69**) and the table in the top-right show miniature maps in a row.

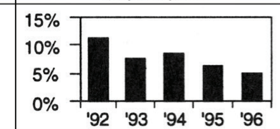
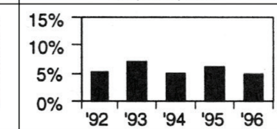
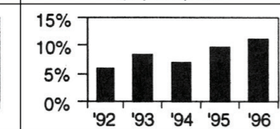
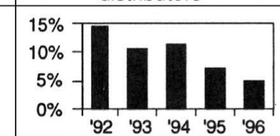
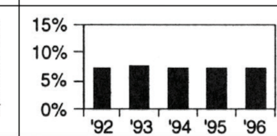
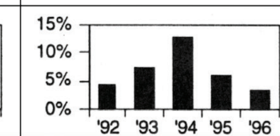


Icon Symbols on a Graph

Icon Symbols on a Map



Matrix Display

	Company A	Company B	Company C
Headquarters	Dallas, Texas	Paris, France	Tokyo, Japan
Ownership	Public	Public	Private
1996 Sales	$56,000,000	$76,000,000	$87,000,000
1996 Net profit	$4,000,000	$5,000,000	$6,000,000
Percent administrative expenses			
Type sales force	Manufacturers' Representatives	Direct and agents	All direct
Distributed to retailers by	Wholesalers and distributors	Wholesalers	Direct from factory
Percent marketing expenses			
Product strategy	Emphasize new products	Emphasize service	Emphasize improved existing products
Area where major sales efforts are focused			
Trend in market share			

Tabular Display

Tables can be used to effectively organize multiple visualizations of different data sets, using a combination of words, numbers, and visualizations as a means of comparison. The example above, also taken from *Information Graphics*, shows the profiles of three companies for five years (1992–1996). Text and numbers, as well as bar graphs, geospatial maps, and arrows indicating general trends, are employed to render a holistic picture. Tables of any size and any visualization types (**page 30**) may be used.

Matrix Display

Matrix displays plot visualizations of multivariate data in a tabular or matrix-like fashion. Each matrix cell displays one combination of attribute elements in the given rows and columns. Shown on the left is a four-by-two matrix that shows the numerical correlations between age and salary (columns) and the number of papers, citations, funding dollars, and doctoral students for 20 faculty members in a fictional department. Correlations can be easily spotted (e.g., the positive correlation between age and citations due to the time it takes before publications accumulate citation counts). Outliers can be spotted as well (e.g., the red dot indicating a faculty member who, despite high age and salary, has a low number of citations, possibly due to extensive teaching or service duties).

Multilevel Display

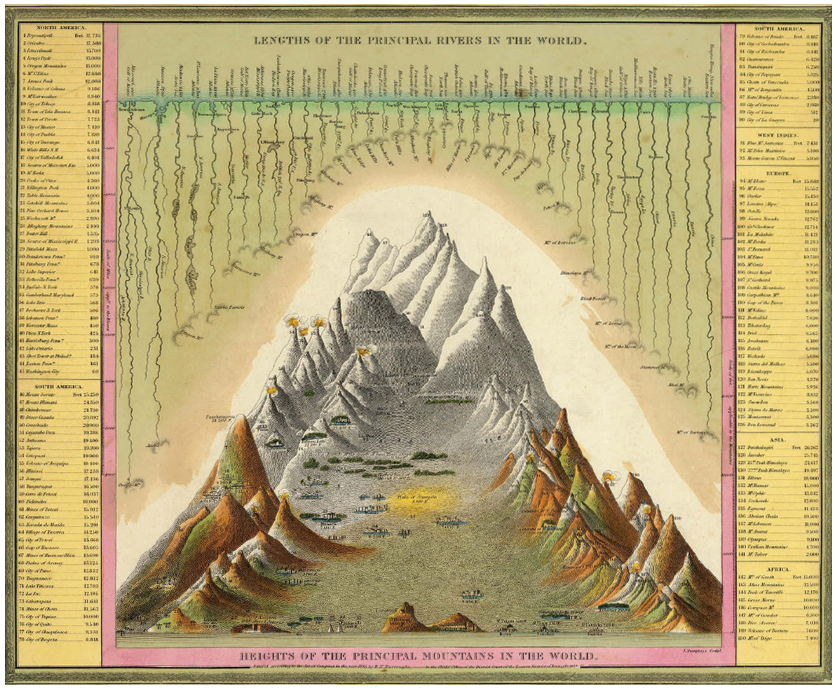
Many visualizations are composed of multiple data layers. Simultaneously showing each of the separate layers can help to improve legibility of the single layers and their overall composition. Interactive visualizations (**page 68**) commonly support the selection and display of specific layers (e.g., publications, patents, and/or funding data overlaid on a geospatial or science map), making it possible to focus on and compare a smaller subset of the data. The two visualizations discussed in the lower part of **page 67** feature multiple interactive, coupled windows (see **page 178, References & Credits** for links to websites).

Exploded Diagram

Exploded diagrams are common in engineering and anatomical drawings. They show how different parts (e.g., of an information visualization) relate to one another, the underlying data, or the planned decision making. They may also reveal the design process or assembly.

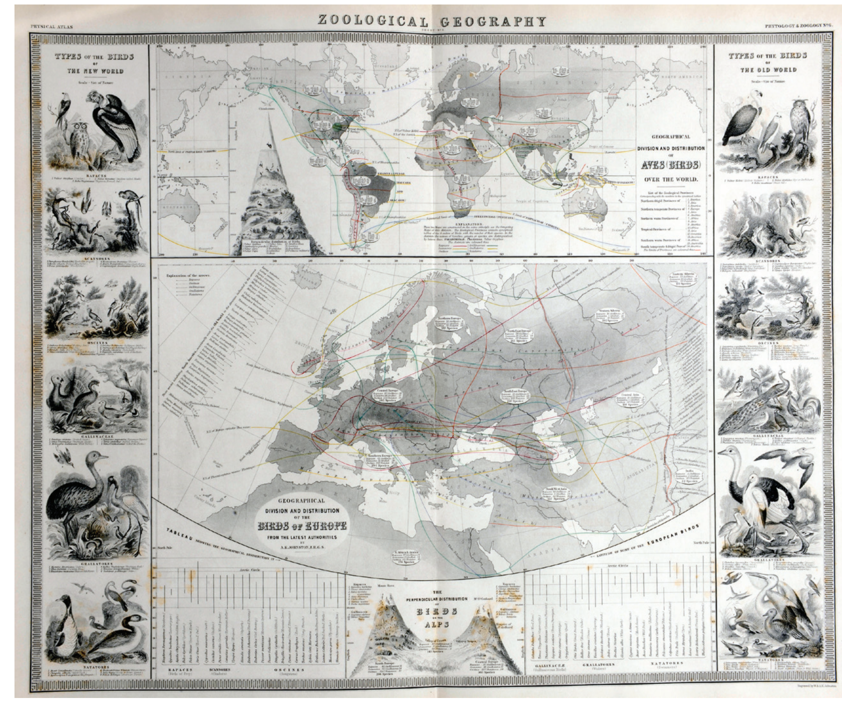
Heights of the Principal Mountains in the World, Lengths of the Principal Rivers in the World

This map from 1846 shows the heights and lengths, respectively, of the world's most prominent mountains and rivers (see high-resolution version of this map at **http://scimaps.org/atlas2**).



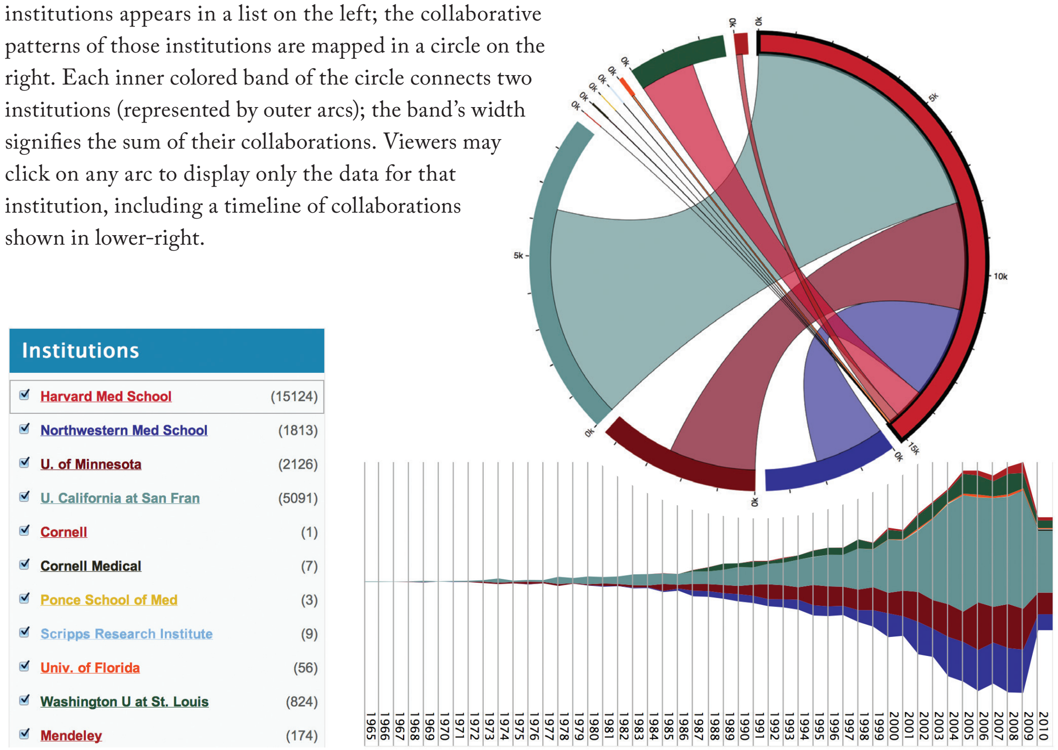
Zoological Geography

This 1856 lithograph from Alexander Keith Johnston's *Physical Atlas* shows different types of birds of the "new" and "old" worlds (on the left and right, respectively). Districts and migration paths are color-coded by type. The perpendicular distribution of birds in the Alps is also given in the lower part of the map.



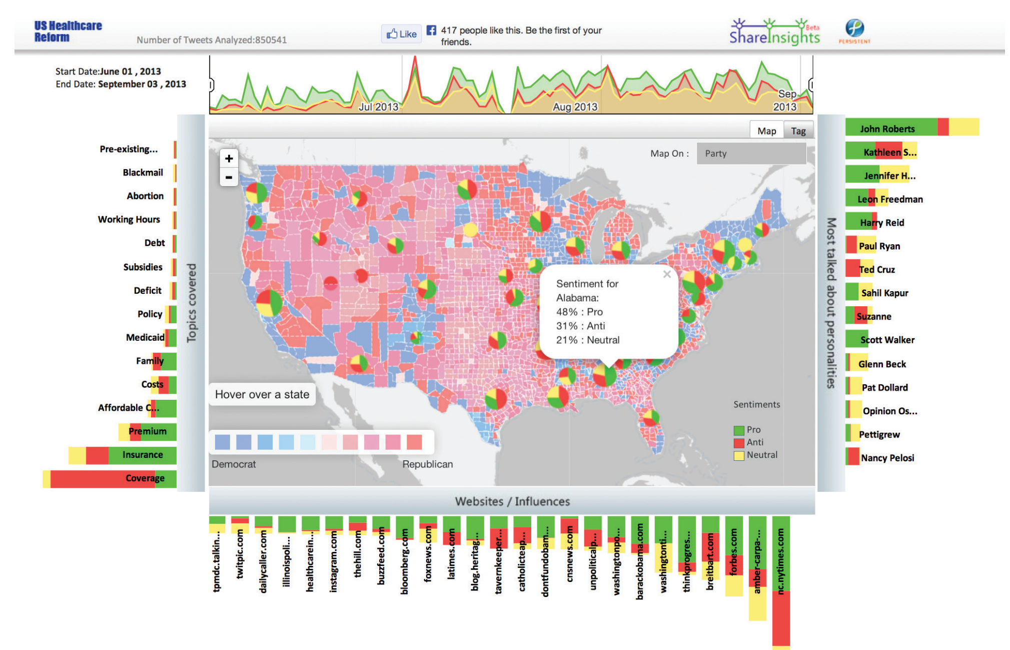
Inter-Institutional Collaboration Explorer

Developed by Nick Benik and Griffin Weber at Harvard, this interactive explorer maps Inter-Institutional Collaboration Explorer from 1987 to 2010. The total of collaborative publications produced by each of 11 U.S. institutions appears in a list on the left; the collaborative patterns of those institutions are mapped in a circle on the right. Each inner colored band of the circle connects two institutions (represented by outer arcs); the band's width signifies the sum of their collaborations. Viewers may click on any arc to display only the data for that institution, including a timeline of collaborations shown in lower-right.



U.S. Healthcare Reform

This map by Persistent Systems (via their ShareInsights platform) shows sentiments extracted from tweets concerning the U.S. healthcare reform. Sentiments are color-coded to signify **Pro** (green), **Anti** (red), and **Neutral** (yellow) expressions. They are overlaid on a geospatial map, with pie chart glyphs indicating the sentiment for each state and a timeline graph (at top) showing the number of tweets over time. Bar graphs chart **Topics covered** (at left); major **Websites/Influences** (at bottom); and **Most talked about personalities** (at right).



Interaction

Many data sets are too large to fit on one screen or printout. Interaction permits the user to first gain a global overview of all the data and then to zoom in to that data, search for and filter out relevant records, and/or retrieve details on demand. The structure and dynamics of data can be explored at multiple orders of magnitude. In principle, any part of the analysis workflow and any layer of the visualization design can be modified via user input. For example, users can select the (real-time) data sets that are shown; the preprocessing, analysis, modeling, or layout that is performed; the specific data that is on display; the visual encoding of different data variables; the aggregation and clustering that is applied and visualized; the combination of visual views that are shown; or the legend that is presented, including the way it was compiled.

Graphing data needs to be iterative because we often do not know what to expect of the data: A graph can help discover unknown aspects of the data, and once the unknown is known, we frequently find ourselves formulating new questions about the data.
William S. Cleveland

Interaction Types

Different deployment (e.g., print versus digital) supports different types of interactions that are appropriate for answering specific questions. Two-dimensional or three-dimensional printouts facilitate the detailed examination of static visualizations at a supremely high resolution. Digital devices support animations and interactivity but typically at a lower resolution (see **page 70**, **Device Options**). Although interactivity is particularly beneficial during data exploration, it can also be highly effective during data communication (see **page 71**, *200 Countries, 200 Years, 4 Minutes*).

User interactions can be grouped according to the transformation(s) they effect: **Data Transformations** that process and analyze raw data and compute data formats that can be visualized; **Visualization Transformations** that define the visual encoding of data records; and **Visual View Transformations** that manipulate the final views of the data. All three are discussed in this section.

Data Transformations

Data slicing (e.g., by time), filtering, and querying allow users to quickly find and access relevant information. Diverse data preprocessing and analysis methods, as discussed on **pages 44–60**, help extract important patterns and trends. Filtering, resampling, aggregation, or dimensionality reduction are commonly used to reduce visual clutter.

Visualization Transformations

These transformations refer to the decisions made about the reference system (**page 24**), projection or distortion, and data overlay that should be used. Designing the data overlay, specifically, requires deciding which data variables should be mapped to which visual elements (see **page 32**, **Graphic Symbol Types**, and **page 34**, **Graphic Variable Types**). The strong interdependence between data analysis and visualization is worth noting, as derivative data variables (e.g., node degree, bursts of activity, clusters, or backbones) can only be used during visual mapping if they have been previously computed.

To achieve higher update speeds when manipulating large data sets, data records can be rendered initially either at a lower resolution (e.g., as wire-frame models) or without textual labels. Only truly important items might be rendered on the screen while computationally expensive queries are performed. Users may also be able to select which labels, gridlines, and background imagery should be visible during interactive exploration.

Visual View Manipulations

The display of millions of data records often leads to visual clutter (i.e., data visualizations with many overlapping or occluding data records that are difficult if not impossible to read). Manipulations such as overview, (semantic) zoom, search, filter, and details on demand are applied to visualize multiple scales of time and geospatial, semantic, and network space. Ideally, different levels of resolution have the same information density. Rapid, incremental, and reversible actions combined with immediate and continuous feedback help to reduce errors while encouraging exploration.

Overview

Just as there is no better way to first see a new city than from the top of its highest tower or nearby mountain, the most desirable way to first see a data set is from above, before zooming in to examine intricacies.

Zoom

Zoomable user interfaces empower users to explore very large information spaces. Zooming coupled with damping makes it possible to navigate effectively by starting to zoom slowly before accelerating and then finally slowing down gracefully when approaching the desired destination.

Filter

Diverse interface elements have been developed to support dynamic queries. Among them are range sliders (a variant of scroll bars), which support dynamic pruning from both sides (see the *London Travel-Time Map* on the opposite page), and alphasliders, which support rapid, serial, and visual filtering by reducing the range of alphanumerically sorted data that is displayed.

Visualization tabs support navigation between multiple windows. For example, *The Baby Name Wizard* on the opposite page contains tabs for a timeline and geospatial maps.

Detail on Demand

Being able to access raw data is essential for many applications. By clicking on a graphic symbol, a user can bring up a listing of all the data records it represents. Selecting data records would then bring up summary information or lead to the raw data that is being locally hosted or retrieved from third parties. This focus and context support is important, because it helps the user to make decisions regarding detailed data records in the context of a larger data set.

Search

Visual search for a specific data record using the naked human eye can be extremely time-consuming. In contrast, automatic search supports the rapid selection of data records based on either primary or derived data variables (e.g., name or node degree, respectively).

Sorting

Sorting by value or category is especially helpful when trying to understand minimum and maximum values or the general distribution of a data set. Missing or top n values can be easily identified, and thresholds can be applied.

Extraction

Many users need to run further analyses of the final set(s) of data records. That is, they need a way to save these records for further processing. Frequently, only a subset of the data variables is needed (i.e., it is desirable to support the download of custom data formats).

History

Users will be more likely to explore novel workflows if they can "undo" previous actions. Log files of user actions can be used to share and rerun workflows in support of result replications; they are also valuable for submitting bug reports.

Information Density

When supporting different interaction types, it is important to provide an appropriate information density. The process of determining the best information density depends strongly on data, analyses, and user characteristics. Typically, interactive information visualizations simultaneously show results from a lower bound of 500 data records to an upper bound of 100,000 data records. Homogeneous data can be more densely represented than highly multi-dimensional, loosely correlated data. Important data patterns may only be evident within specific scale ranges. Therefore, visualization designers need to be sensitive to relationships between data sampling, analysis, and visual resolution. Casual users (who use a visualization infrequently and for a short time) need a simple, less dense visualization, as compared to expert analysts (who may use the same visualization extensively on a fairly regular basis).

Interaction Support

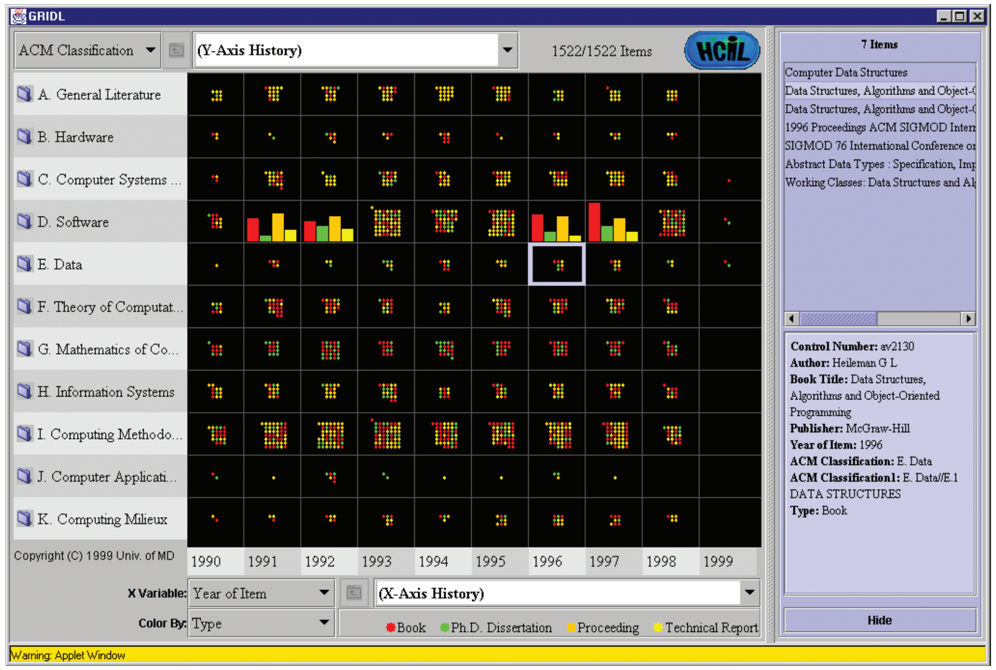
Interacting with large-scale information spaces can easily lead to a feeling of "being lost."

Three major paradigms are used to support navigation: (1) spatial navigation that mimics our experiences in the physical world; (2) semantic navigation, which is driven by semantic relationships or underlying logic; and (3) social navigation, which takes advantage of the behavior of like-minded people.

In addition, there exist three forms of user guidance: (1) manipulation support (e.g., constraining user manipulation by having objects snap to a grid, or having objects repel each other to avoid obfuscation); (2) coordination support (e.g., using tightly coupled windows—also called tightly coordinated windows or brushing and linking—to identify a set of data records in one window and see them highlighted in all other views of the same data set); and (3) self-evaluation support (e.g., status displays, commonly used in computer games, to communicate users' progress and accuracy during data exploration).

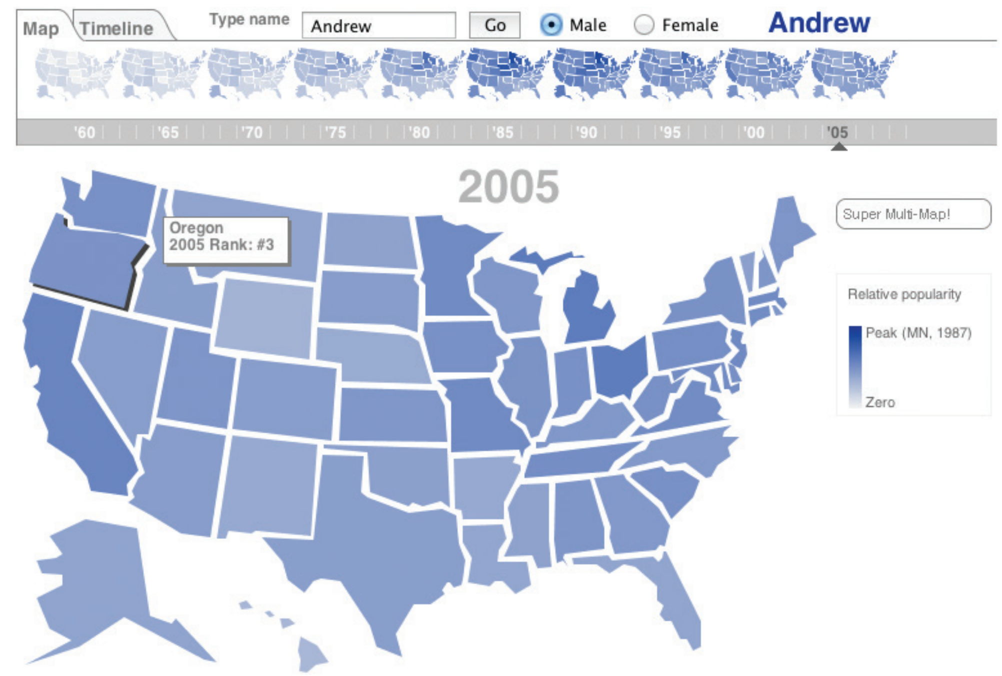
GRIDL

The GRaphical Interface for Digital Libraries (GRIDL) visualization, introduced on **page 58**, is highly interactive. At the highest zoom level, patterns and distributions for 100 to 10,000 records can be easily recognized. Selecting a table cell brings up a listing of all relevant documents (see top-right). Any document can be selected to explore document details (see bottom-right). The mapping of data attributes to axes as well as to size and color can be readily changed.



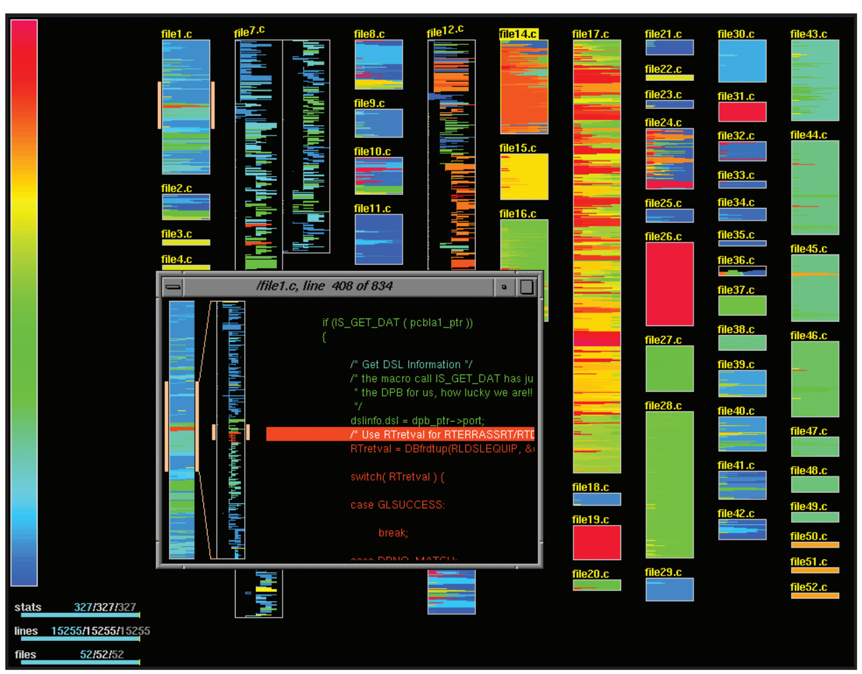
The Baby Name Wizard

This online service lets anyone explore the prevalence of baby names in each U.S. state over time by simply entering a name (or a sequence of letters) and then selecting the **Map** or **Timeline** tab. Running a query for **Andrew** and selecting the Map tab results in the visualization below: a small multiple display of U.S. maps that show the steadily increasing number of male babies named Andrew per state.



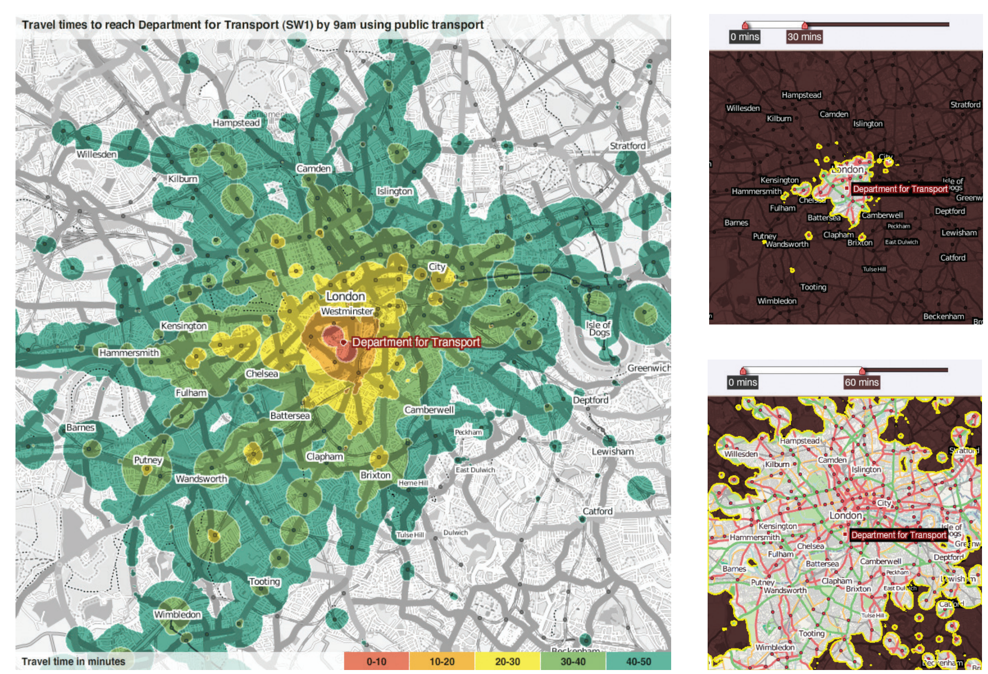
Seesoft: A Tool for Visualizing Line Oriented Software Statistics

In 1992, Stephen G. Eick and colleagues at AT&T Bell Labs published a software visualization system that visualizes up to 50,000 lines of code simultaneously in support of discovery, project management, code tuning, and analysis of development methodologies. The highly interactive interface represents each line of code with a thin line color-coded by data variables of interest, such as age, programmer, and the purpose of the code.



London Travel-Time Map

This interactive map by mySociety lets users specify minimum and maximum travel times in London to help them explore the accessibility of different areas from the Department for Transport in Pimlico if they were to start their trip at 9:00 a.m. Orange represents the shortest travel times, of 0 to 10 minutes; dark green represents the longest travel times, of 40 to 50 minutes. The interactive version of the interface is shown at right.



Human-Computer Interface

Various hardware and software combinations support a wide range of user input and computer output. They also have a wide range of price tags and can lead to vastly different user experiences. Large-size paper printouts are most affordable yet static. Mobile devices that are an integral part of users' lives support real-time data access and interactivity. Larger audiences benefit from displays that are visible to many and potentially support multiuser interaction. Virtual reality setups that emulate a three-dimensional visual, audio, and haptic space, akin to our real-world environment, are expensive and often reserved for expert domain applications or gaming. Internet access and speed determine the feasibility of online services. This spread provides an overview of hardware and software properties and interface affordances that need to be taken into account when deploying visualizations.

The real voyage of discovery consists not in seeking new landscapes, but in having new eyes.
Marcel Proust

Needs and Affordances

Different user groups have very different insight needs, learning objectives, or monitoring goals. A detailed understanding of user needs (see **page 40, User Needs Acquisition**) helps answer questions such as: Which human senses should be engaged, and in what way, to support effective navigation, access, manipulation, and insight making? Which information is best communicated via text, visualizations, audio, haptic feedback, or combinations thereof? Plus, what user input is necessary to effectively steer data navigation and exploration? Highly sophisticated setups that are difficult to learn or maintain, or simply too innovative for their time (see *Morton L. Heilig's Sensorama* on opposite page), tend not to succeed. Interfaces that utilize (or blend into) the fabric of their users' daily lives and offer immediate tangible benefits are likely to have faster and higher adoption rates.

Device Properties

Different input devices (e.g., a camera or scanner) and output devices (e.g., a printer or screen) can vary greatly in terms of resolution, brightness, and color range. They may also support different viewing angles, update frequencies, and write/read different file formats.

Resolution

The resolution of a camera, scanner, printer, or monitor is commonly measured in dots per inch (DPI)—the number of dots in a one-inch line. Pixel, short for picture element, is the smallest dot that a device can read or write. Voxel, or volume element, is the smallest volume that a device can read or write. In 2014, the preferred values were 72 DPI for the web and 300 DPI for printouts. Laser printers support a resolution of up to 1200 DPI to support anti-aliasing, different gray levels, and superacuities.

Smaller screens require more zooming and panning to view the same amount of information. Effective interaction design can help with navigating diverse windows, opening and closing palettes, or switching between detail and overview.

File Size

The more pixels/voxels per inch, the higher the resolution, and the larger the file size. A photo taken with a 16:9 aspect ratio camera—at a resolution of 2981 x 1677 = 5,000,000 pixels or five megapixels (MP)—can be printed in 300 DPI at a size of 9.9" x 5.6" (25.2 x 14.2 cm). It would require 2.5 HD screens (each with a resolution of 1920 x 1080 = 2,073,600 pixels) to be viewed in full resolution. Downsampling the image to 72 DPI web resolution reduces the file size by a factor of about 16.

File size is also affected by color depth and color mode. Color depth (also called bit depth) depends on the mode the image was captured in (8 bit, 16 bit, or higher). Common color modes are grayscale, RGB, or CMYK. An RGB image has three channels (red, green blue), CMYK has four (cyan, magenta, yellow, black), and a grayscale image has one (black). In 8-bit RGB color mode, the file size of a 5 MP image is 14.3 MB; in the more common 16-bit, it is 28.6 MB; and in 32-bit, it is 57.2 MB.

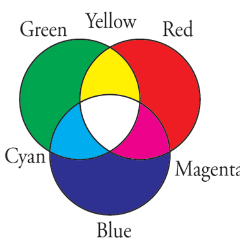
Many sciences produce super-high-resolution images by combining multiple images. For example, the Photopic Sky Survey is a 5,000 megapixels (MP) photograph of the entire night sky, stitched together from 37,440 exposures—it would require 1,000 times more space to print or display than a 5 MP image.

Brightness

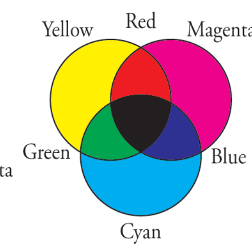
Paper printouts require illumination by external sources (e.g., the sun) to be seen. The brightness of digital displays is indicated in lumens, a photometric measure for the perceived brightness of a light source. A standard 100-watt incandescent light bulb emits approximately 1,700 lumens. Laptop and TV screens as well as data projectors have 500 to 12,000 lumens.

Color

A color space is a mathematical model for describing color. Diverse color models exist that organize colors according to their properties. Examples include RGB (red, green, blue), which is an **Additive Model** used in computer displays; CMYK (cyan, magenta, yellow, black), a **Subtractive Model** widely used in printing (see below); and the use of HSV (hue, saturation, value; see **page 35, Color**).



Additive Model (RGB)



Subtractive Model (CMYK)

Viewing Angle

The field of view is the angular extent of the observable world. Humans have an almost 180-degree forward-facing field of view. Their binocular vision, which is important for stereo and depth perception, covers 140 degrees. A smaller distance to a screen or printout equals a larger field of view taken up by that visual.

Handheld devices stimulate about 5 to 10 percent of the visual field, whereas large display walls may cover the entire field of view.

Update Frequency

The update frequency (also called refresh rate) denotes the number of times per second that a display screen is redrawn. Higher update frequencies decrease flickering, thereby reducing eyestrain. Standard computer displays use a 60 Hz refresh rate (i.e., the screen is refreshed 60 times per second). Stereo displays need to render a separate picture for each eye and require a frequency of at least 120 Hz. TV screens use 60 Hz (NTSC) and 50 Hz (PAL/SECAM) frequencies.

Data Format

There are two main format types for storing images. **Vector** formats such as PostScript (PS) or Scalable Vector Graphics (SVG) store a geometric description that can be rendered at any size. **Raster** formats such as JPG, TIFF, GIF, BMP, and PNG store data as grids of pixels. Examples are shown below.


Vector Raster

Device Options

Visualizations can be printed on paper or in three dimensions, projected onto screens, or displayed on handheld devices or in virtual reality setups. The advantages and disadvantages of different options are discussed ahead.

Printouts

Paper comes in different sizes, weights, surface textures, and colors. Paper printouts are cheap and fast (no boot-up time required); they are also easy to transport, deploy (no outlet needed), and annotate (e.g., by using a pen). Prints offer the highest resolution [a map the size of a 4' x 6' (1.2 x 1.8 m) dining table in print quality can display 420 MP] and help to minimize changes in eye focusing and head or body movements. Plus, they can be easily explored by multiple users. Archival paper prints stored in a dry, dark room are likely to be readable in 500 years.

Three-dimensional prints can be created manually (e.g., using same-size bricks to render three-dimensional bar graphs) or by using 3D printers that create structures out of plastics, resins, and metals in different resolutions, using one or more colors.

Digital Displays

Computer, laptop, tablet, and phone displays come in different sizes and resolutions, with varied interactivity, at a wide range of prices. Online services such as Zoom.it or GigaPan.org support the sharing of large visuals and their interactive explorations via zooming and panning functions. In 2014, ultra high-resolution television displays support 33 MP. If more pixels are needed, multiple displays or projectors can be combined into a tiled wall or globe (see the *Giant Geo-Cosmos OLED Display* and *Indiana University's Virtual Reality Theater* on the opposite page).

Standard user input comprises text input via keyboard, click and selection via mouse-like devices, audio input (e.g., voice recognition), and touch-sensitive surfaces. Touchscreen tables support the identification of multiple fingers (and users); cameras support recognition of gestures and eye movements. Handheld devices may have eye-tracking, heart rate, temperature, and other sensors.

Stereo Displays

The exploration of 3D structures can benefit from stereoscopic displays such as 3D computer and TV screens. Devices such as the Responsive Workbench and ImmersaDesk use a horizontal screen to project stereoscopic images, which makes them well-suited for tasks that in the real world would be performed on a table. Multiple users wearing shutter glasses can view high-resolution, head-tracked images, and stereo sound.

CAVE systems are multi-person theaters that use rear projection of images on all four walls, the floor, and the ceiling. Some can be reconfigured; that is, the position of walls can be modified, such as in **Indiana University's Virtual Reality Theater** (see bottom-middle). Early CAVE Systems used two projectors with a resolution of 1024 x 768 pixels to illuminate each wall; printed in 300 DPI, that 0.8 MP resolution produced a 3.8" x 2.9" (9.7 x 7.4 cm) image—the size of half a postcard.

Illuminated Diagram Display

This display combines the high data density of large paper printouts with the flexibility of an interactive program driving a touch-panel display and two projectors that illuminate the maps (see bottom-middle picture of setup on **page 76** and *Atlas of Science*, **pages 180–185**).

Augmented Reality and Wearables

Hardware miniaturization and advanced software development support a deeper integration of physical and virtual worlds. Augmented reality refers to the embedding of virtual information in the physical world, using see-through displays or clever camera setups (see the still image from Rosling's BBC documentary *200 Countries, 200 Years, 4 Minutes* on the bottom-right). Wearable (mobile) user interfaces may soon allow information to be available anytime, anywhere, as part of our clothing and the gadgets we carry.

Morton L. Heilig's Sensorama

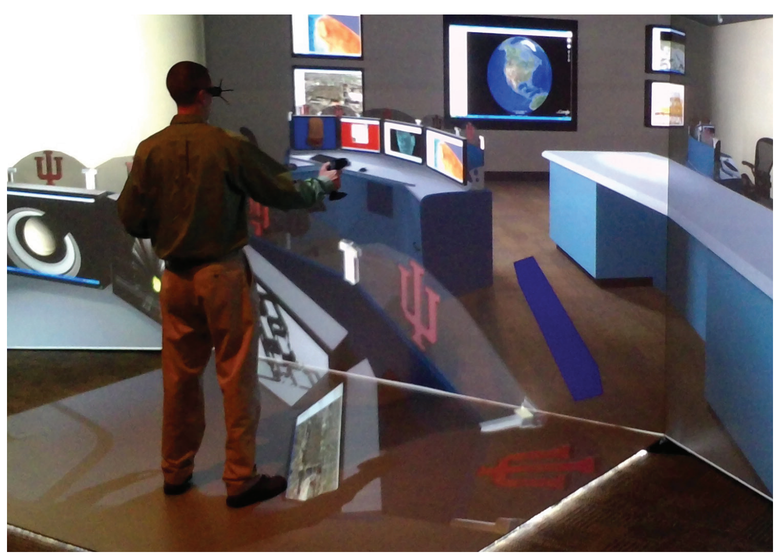
Patented in 1962 by American cinematographer and inventor Morton L. Heilig, the Sensorama 3D movies featured stereo images, wide vision, motion, color, stereo sound, aromas, wind, and vibrations to provide full sensual vividness and dynamic vitality.

Giant Geo-Cosmos OLED Display

Dozens of Ingo Günther's Worldprocessor Globe designs (see *Atlas of Science*, **pages 140–163**) come to new animated life accompanied with a data-driven soundtrack on the geo-cosmos display, the emblematic heart of the Museum of Emerging Science and Innovation in Tokyo, Japan. The 20' (6m) diameter display features 10,362 palm-sized, organic light-emitting diode (OLED) panels, for a total of more than 10 MP.

Indiana University's Virtual Reality Theater

Immersive environments such as this reconfigurable CAVE system make it possible to virtually experience product designs or architectural solutions before they exist physically. Shown here is an interactive walkthrough of a proposed furniture layout for an IT control room at Indiana University.



200 Countries, 200 Years, 4 Minutes

This BBC documentary, featuring Hans Rosling, captures the development of 200 countries over 200 years. Specifically, it uses effective visualizations, persuasive argumentation, and an innovative camera setup to communicate the immense changes over time in lifespan and in the income per person (GDP per capita) rates, adjusted for inflation and differences in cost of living (purchasing power) across countries.

Validation and Interpretation

There now exists a rich variety of algorithims, tools, and services that turn data into visualizations. While some are designed for use by experts, a growing number of easy-to-use tools is widely used by non-experts. Most datasets can be analyzed and visualized in many different ways. The majority of the possible algorithm and visualization design combinations is incorrect or imperfect; only a select few combinations result in readable, informative, and actionable visualizations. This spread reviews the criteria and methods for validating (alternative) visualizations and for estimating their value for sound decision making. Examples of good and bad visualizations are used to illustrate common problems and potential solutions (see opposite page).

Human judgment without automated data mining is blind; automated data mining without human judgment is empty.
Colin Allen

Validation Criteria

Visualizations are commonly optimized and evaluated according to three qualities: function (utility, usability, effectiveness, and scalability), aesthetics (quality and appeal), and integrity (accuracy and replicability); for details, see works by Edward Tufte, David McCandless, and Bradford W. Paley (**page 178**, **References & Credits**). Some metrics can be observed or computed (e.g., in terms of speed, accuracy, or scalability). Others (e.g., beauty or relevance) require expert evaluation.

Function

A visualization should display the most important information in clear and accessible form. Relevant questions for consideration can be broken down into function-specific categories.

Utility

Does the visualization satisfy the technological, contextual, and business insight needs of the target audience? What is the decision-making value—that is, which major insight does the visualization provide, and why does it matter? Does it inspire viewers to learn more or to act differently? Does it support asking questions, making future explorations, or generating hypotheses? How generic is the solution? What range of questions can be answered? Do people continue to use it in practice? Do they buy it or purchase upgrades? Is the creator invited to continue producing similar visualizations?

Usability

Is the visualization easy to read and use by the target audience? Is its purpose clear? Does it use a common yet sufficiently expressive reference system? Is the mapping, from data scale types to graphic variable types, easy to understand? Is the provided interactivity easy to use, and is it sufficient?

Effectiveness

For each visualization, one should clearly state the user needs and then show the rationale behind the selection of certain reference systems, metaphors, color-coding, interactivity design, etc. Questions to be addressed comprise: Is the display space used effectively? Is the number of data points and the data density appropriate? Is all relevant data visible, or are there occlusions? Are the key findings dominantly represented? Is the given story told in a consistent fashion? Does it allow easy access to additionally needed data?

Scalability

Most visualizations work well at the micro and meso levels; few scale to the macro-level, big-data studies that have millions or even billions of data points. Does the visualization degrade gracefully as the amount of data increases (e.g., are data analysis techniques used to help derive insights from dense networks that are initially illegible or visually akin to spaghetti balls)? How responsive is the visualization to user interaction?

Aesthetics

Visualizations need to attract the attention of viewers to communicate. Visual aesthetics (i.e., well-composed, high-quality data renderings) are important.

Design Quality

Visual aesthetics comprise design quality, the originality of the underlying idea, and international and/or interdisciplinary appeal. Carefully selected and easy-to-read image compositions, color palettes, shapes, and forms help to improve quality.

Appeal

Ideally, viewers will be attracted by a visualization and have fun interacting with it. The visualization will have even higher mass appeal if it has been featured in news channels, popular blogs, social media, on the cover of a major journal or magazine, or as part of a prominent museum exhibit.

Integrity

A visualization should present data in the most objective way. It should be generated using the most accurate and highest coverage data and the best methods available. All of these factors add to the creator's credibility.

Accuracy

The quality of the data, analysis, and design is key for the creation of accurate visualizations. If uncertainty exists in either the data or in the analysis and visualization workflow, then it should be stated unambiguously. Subjective choices or manual data modifications need to be clearly documented.

Replicability

Any visualization should come with sufficient documentation to recreate it. Documentation should comprise information on the original data (including source and baseline statistics); details about how data was cleaned or preprocessed; the analysis and visualization algorithms that were applied; and the parameter values that were used. One should list all authors, ideally with brief information on their expertise and specific contributions, and mention all funders, as commercial interests are likely to influence visualization design and description. A detailed documentation of work will improve consistency and ease future studies.

Validation Methods

When designing visualizations, it is beneficial to validate results early and often. Different qualitative and quantitative methods exist to (obtrusively or nonobtrusively) evaluate visualizations.

Field studies are employed to understand how users interact with a visualization or tool in the real world—with their own data and tasks. Longitudinal field studies work with users over extended periods of time. Field experiments design user tasks to simulate real analyses and recruit groups of users for one-on-one sessions that test the visualization or software (not the users), encourage thinking aloud, and record top usability issues. Both emphasize real-world context and learning through observation (not just opinion).

User Studies

User studies are commonly employed to evaluate or compare design alternatives. Evaluation metrics such as task-time completion and error counts shed light on the usability and effectiveness of visualizations. Users may be asked to think aloud so that evaluators can capture their thought processes and insights. Eye-tracking devices help researchers understand how interactive visualizations guide users' eyes as well as their navigation and processing of information spaces. Longitudinal studies (i.e., repeated observations over long periods of time) are used to study the adoption of novel visualizations among existing ones.

Human (Expert) Validation

An open-ended protocol, a qualitative insight analysis, and an emphasis on domain relevance may all benefit the identification of those visualization features that can help users achieve insight and those that may prove problematic—directly informing visualization refinement and improvement. For example, human experts may be asked to draw a domain map, and this map would then be compared to visualizations automatically constructed according to domain data. Experts may also be consulted in classification and labeling studies, in which participants are asked to freely explore given visualizations and then to identify major domains and prevalent topics (e.g., by drawing cluster boundaries around similar objects and assigning a label to each cluster). In utilization studies, participants use visualizations to make sense of data, and the results are compared to those derived by automatic means.

Controlled Experiments on Benchmark Tasks

For rigorously evaluating visualizations, many scientific communities have compiled data repositories and synthetic data sets that support the given experiments. In general, benchmark tasks must be predefined by test administrators, and users must precisely follow specific instructions during the experiments. Each task has a definitive completion time that is fairly short (typically under one minute), in support of a large number of task repetitions. Each task has definitive answers that are used to measure accuracy. Answers are often simple (e.g., multiple choice in support of objective mechanical or automated scoring).

Crowdsourcing Evaluation

Amazon's Mechanical Turk and similar platforms can be used to crowdsource evaluation (see **page 174, Democratizing Knowledge and Participation**). For example, Jeffrey Heer and Michael Bostock crowd sourced graphical perception experiments by replicating prior studies of spatial encoding and luminance contrast; conducting new experiments on rectangular area perception (as in treemaps or cartograms) and on chart size and gridline spacing; and analyzing the impact of reward (payment) levels on completion time and result quality finding that higher rewards lead to faster completion rates.

Interpretation

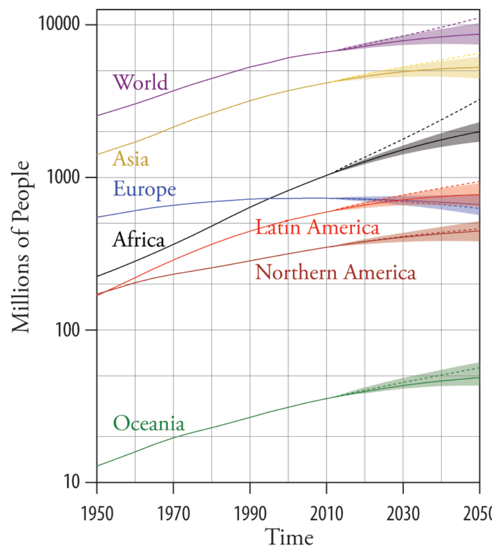
Data analysis and visualization create a "formalized representation" of data, which needs to be interpreted to inform sensemaking and actions. When reading a visualization, it is important to detect any omissions, errors, and biases.

Errors are easily made in any step of the analysis and visualization workflow. Critical data can be left out; algorithm and parameter selections can have a major impact on visualization layout and design; and visual encoding choices will affect the interpretation of results. John Brian Harley's theory of cartographic silence distinguishes two types of silences: intentional silences, which are specific acts of censorship, and unintentional silences, which are unconscious omissions. Examples of misleading visualizations are given on the right. When interpreting a visualization, it is important to understand both its power and its limitations.

When using visualizations in decision making, it is important to distinguish (1) the true question or issue from (2) the data and methods applied to answer it and (3) the potential impact of planned decisions. Frequently, decisions influence future actions and the resulting data. For example, funding a new area of research will lead to new hires; newly hired scholars will then publish or perish; and each publication will cite other papers—most likely within the funded area. That is, there is a strong correlation between the amount of funding an area of science enjoys and the number of citations papers in that area receive. If future funding is based on the number of existing citations, then "rich areas" become even richer over time—which might not be intended.

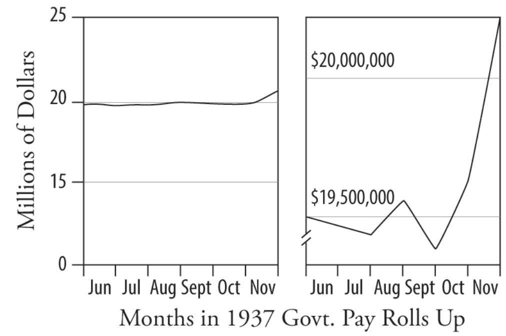
Scales

The same data plotted on a linear scale will appear quite different when plotted on a logarithmic scale. Data that grows exponentially (e.g., the increase in world population from 1 billion in 1800 to 7 billion in 2011; see graph on **pages 2–3** in *Atlas of Science*) will look like a straight line in a logarithmic plot (see the United Nations population estimates below for different continents between 1950 and 2050).



Distortions

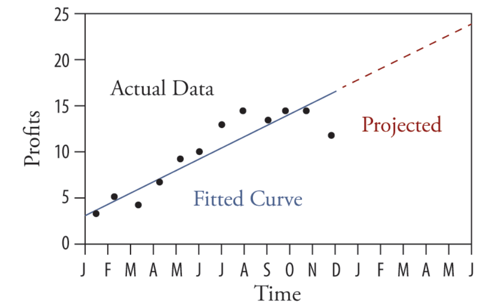
Visualizations can be distorted in many different ways, making them difficult or impossible to interpret correctly. Two renderings of the same data—government payrolls in 1937—are shown here; the left image with the broken y-axis scale is meant to suggest an increase in payrolls, whereas the right image confirms payroll stability.

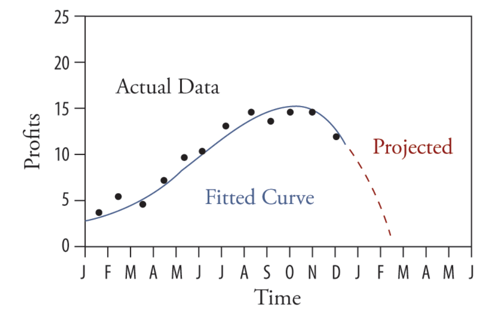


Not only elements of the reference system (e.g., axes) but also data overlay (e.g., graphic symbol types such as bars; see **page 46, Comparisons**) may be broken.

Regressions

As discussed in **Statistical Studies** (**page 44**), the selection of different curve fittings strongly influences the prediction of future values. Shown here are a linear (top) and polynomial (bottom) fitting of the same data; notice the vastly different projections that appear for the month of March.

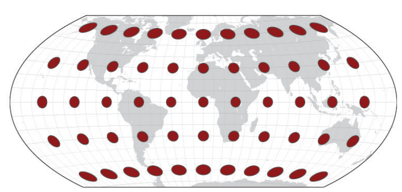




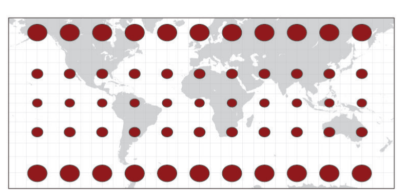
Projections

Changes made in geospatial projections have a major impact on area sizes and the distances between data points. Shown below are three common projections, with Tissot's indicatrices placed at the same geospatial position to illustrate the different distortion at these points for each of the various projections.

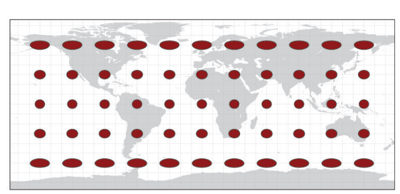
Winleel Tripel



Mercador

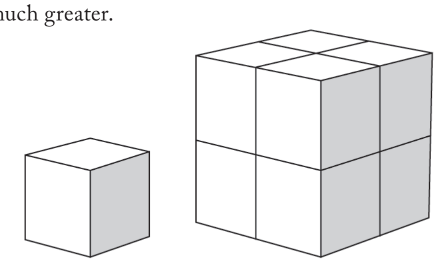


Lambert



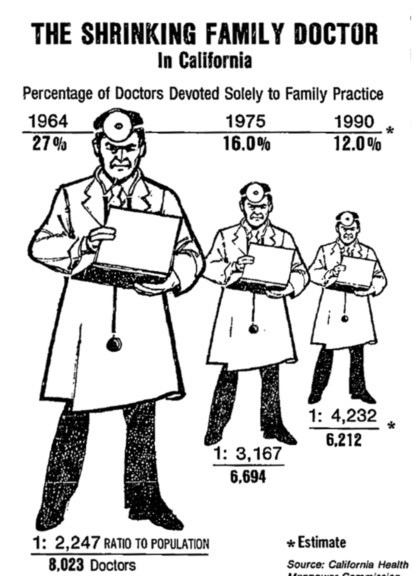
Dimensions

Representing data using three-dimensional objects tends to lead to confusion in interpretation. For example, changing the height of a 3D object (e.g., doubling the height of a 1" x 1" x 1" cube) changes its width and depth proportionally, effectively increasing its volume eight times (so that it becomes a 2" x 2" x 2" cube), see below. Another example can be found in Darrell Huff's *How to Lie with Statistics* that uses three-dimensional drawings of two moneybags to show how the weekly salary for a carpenter from the fictional country of Rotundia differs from that of a U.S. carpenter. According to the fictional data, U.S. carpenters earn twice as much, and the U.S. moneybag is about twice the height—however the impression of the difference is much greater.



Perspective

Linear perspective has parallel lines converging to a single point; that is, objects of the same size that are placed further away appear smaller than nearby objects. This can cause confusion in data visualizations. For example, the doctors in this example appear to be proportionally the same size, contrary to the data values they represent.



Hausmann, Ricardo, César A. Hidalgo, Sebastián Bustos, Michele Coscia, Sarah Chung, Juan Jimenez, Alexander Simoes, Muhammed A. Yildirim. 2011. *The Atlas of Economic Complexity*. Boston, MA: Harvard Kennedy School and MIT Media Lab. Accessed August 28, 2013. **http://www.cid.harvard.edu/documents/complexityatlas.pdf**.

Hidalgo, César A., Bailey Klinger, Albert-László Barabási, and Ricardo Hausmann. 2007. "The Product Space Conditions the Development of Nations." *Science* 317 (5837): 482–487.

Mager, Christoph. 2012. "Heidelberg Nobel Prize Winners." In *Wissenschaftsatlas of Heidelberg University: Spatio-Temporal Relations of Academic Knowledge Production*, edited by Peter Meusburger and Thomas Schuch, 250–253. Knittlingen, Germany: Bibliotheca Palatina.

Skupin, André, and Ron Hagelman. 2005. "Visualizing Demographic Trajectories with Self-Organizing Maps." *GeoInformatica* 9 (2): 159–179.

Image Credits

The Atlantic Slave Trade by Philip D. Curtin © 1969 by the Board of Regents of the University of Wisconsin System. Reprinted by permission of the University of Wisconsin Press.

Nobelpreisträger, für Physik, Chemie, und Medizin from Mager 2012.

Brain Circulation

References

Johnson, Jean M., and Mark C. Regets. 1998. *International Mobility of Scientists and Engineers to the United States: Brain Drain or Brain Circulation*, National Science Foundation (NSF 98–316).

Salmi, Jamil. 2012. "Attracting Talent in a Global Academic World: How Emerging Research Universities Can Benefit from Brain Circulation." *Brain Circulation* 2 (1). Accessed September 5, 2014. **http://academicexecutives.elsevier.com/articles/attracting-talent-global-academic-world-how-emerging-research-universities-can-benefit**.

Saxenian, Anna Lee. 2002. "Brain Circulation: How High Skilled Immigration Makes Everyone Better Off." *The Brookings Review* 20 (1): 28–31.

Productivity

References

Department of Business, Innovation and Skills. *International Comparative Performance of the UK Research Base – 2011*. Accessed November 5, 2013. **http://www.bis.gov.uk/assets/biscore/science/docs/i/11-p123-international-comparative-performance-uk-research-base-2011**.

Image Credits

Image © Crown 2011, licensed under the Open Government Licence v2.0.

Trade Networks

Global Trade Ecology

References

Minard, Charles Joseph. 1866. *Europe Raw Cotton Imports in 1858, 1864 and 1865*. Paris, France. Courtesy of the Library of Congress, Geography and Maps Division. In "4th Iteration (2008): Science Maps for Economic Decision Makers," *Places & Spaces: Mapping Science*, edited by Katy Börner and Elisha F. Hardy. **http://scimaps.org**.

R&D Investment Dependencies

References

Chartered Institute of Management Accountants. 2012. *Managing Inter-Firm Interdependencies in R&D Investment: Insights from the Semiconductor Industry* 8 (3). Accessed November 5, 2013. **http://www.cimaglobal.com/Documents/Thought_leadership_docs/Organisational%20management/RD-invest-report.pdf**.

Shifts in Technology's Center of Gravity

References

Quah, Danny. 2011. "World's Center of Economic Gravity Shifts East." CNN Online. Accessed November 5, 2013. **http://globalpublicsquare.blogs.cnn.com/2011/04/07/worlds-center-of-economic-gravity-shifts-east**.

Thibodeau, Patrick. 2012. "U.S. Sees Tech's 'Center of Gravity' Shifting to Asia." *Computerworld*. Accessed November 5, 2013. **http://www.computerworld.com/s/article/9234640/U.S._sees_tech_s_center_of_gravity_shifting_to_Asia?pageNumber=1**.

Diffusion of Knowledge

Geospatial and Topical Diffusion

References

Chen, Chaomei, Weizhong Zhu, Brian Tomaszewski, and Alan MacEachren. 2007. "Tracing Conceptual and Geospatial Diffusion of Knowledge." In *Proceedings of HCI International 2007, Beijing, China, July 22–27. Lecture Notes in Computer Science* 4564: 265–274. Berlin: Springer-Verlag.

Viral Marketing

References

Hinz, Oliver, Bernd Skiera, Christian Barrot, and Jan U. Becker. 2011. "Seeding Strategies for Viral Marketing: An Empirical Comparison." *Journal of Marketing* 75 (6).

Diffusion of Reputation

References

Hauke, Sascha, Martin Pyka, Markus Borschbach, Dominik Heider. 2010. "Reputation-Based Trust Diffusion in Complex Socio-Economic Networks." In *Information Retrieval and Mining in Distributed Environments*, edited by Alessandro Soro, Eloisa Vargiu, Giuliano Armano, Gavino Paddeu, 21–40. Berlin: Springer Verlag.

Radicchi, Fillipo, Santo Fortunato, Benjamin Markines, and Alessandro Vespignani. 2009. "Diffusion of Scientific Credits and the Ranking of Scientists." *Physical Review* 80 (5).

Feedback Cycles and Science Models

References

Barabási, Albert-László, and Réka Albert. 1999. "Emergence of Scaling in Random Networks." *Science* 286 (5349): 509–512.

Börner, Katy. 2010. *Atlas of Science: Visualizing What We Know*. Cambridge, MA: The MIT Press.

Börner, Katy. 2017. *Atlas of Forecasts: Predicting and Broadcasting Science*. Cambridge, MA: The MIT Press.

Council for Chemical Research. 2009. *Chemical R&D Powers the U.S. Innovation Engine*. Washington, DC. Courtesy of the Council for Chemical Research. In "5th Iteration (2009): Science Maps for Science Policy Makers," *Places & Spaces: Mapping Science*, edited by Katy Börner and Elisha F. Hardy. **http://scimaps.org**.

Latour, Bruno, and Steve Woolgar. 1982. "The Cycle of Credibility." In *Science in Context: Readings in the Sociology of Science*, edited by Barry Barnes and David Edge, 35–43. Cambridge, MA: The MIT Press.

Martino, Joseph P. 1969. *Science and Society in Equilibrium*. Holloman Air Force Base, NM. Courtesy of AAAS. In "5th Iteration (2009): Science Maps for Science Policy Makers," *Places & Spaces: Mapping Science,* edited by Katy Börner and Elisha F. Hardy. **http://scimaps.org**.

Martino, Joseph P. 1969. "Science and Society in Equilibrium." *Science* 165 (3895): 769–772.

Whitley, Richard. 1984. *The Intellectual and Social Organization of the Sciences*. New York: Oxford University Press.

20 Part 2: Envisioning Science and Technology

References

Dorling, Danny, Mark E. J. Newman, Graham Allsopp, Anna Barford, Ben Wheeler, John Pritchard, and David Dorling. 2006. *Ecological Footprint*. Sheffield, UK and Ann Arbor, MI. Courtesy of the Universities of Sheffield and Michigan. In "4th Iteration (2008): Science Maps for Economic Decision Makers," *Places & Spaces: Mapping Science,* edited by Katy Börner and Elisha F. Hardy. **http://scimaps.org**.

Tufte, Edward R. 1997. *Visual Explanations: Images and Quantities, Evidence and Narrative*. Cheshire, CT: Graphics Press. 50.

Image Credits

Extracted from: Dorling et al. 2006. © Copyright Sasi Group (University of Sheffield) and Mark E. J. Newman (University of Michigan).

22 Foundations and Aspirations

References

Monmonier, Mark. 1996. *How to Lie with Maps*. 2nd ed. Chicago, IL: University of Chicago Press. 2.

Foundations

References

Bertin, Jacques. 1983. *Semiology of Graphics*. Madison, WI: University of Wisconsin Press.

Cleveland, William S. 1993. *Visualizing Data*. Summit, NJ: Hobart Press.

Cleveland, William S. 1994. *The Elements of Graphing Data*. Summit, NJ: Hobart Press.

Playfair, William. 2005 (1786). *The Commercial and Political Atlas and Statistical Breviary*. Edited by Howard Wainer and Ian Spence. New York: Cambridge University Press.

Tufte, Edward R. 1990. *Envisioning Information*. Cheshire, CT: Graphics Press.

Tufte, Edward R. 1997. *Visual Explanations: Images and Quantities, Evidence and Narrative*. Cheshire, CT: Graphics Press.

Tufte, Edward R. 2001. *The Visual Display of Quantitative Information*. 2nd ed. Cheshire, CT: Graphics Press.

Tukey, John W. 1977. *Exploratory Data Analysis*. Reading, MA: Addison-Wesley.

Wainer, Howard. 1997. *Visual Revelations: Graphical Tales of Fate and Deception from Napoleon Bonaparte to Ross Perot*. New York: Copernicus.

Wainer, Howard. 2005. *Graphic Discovery: A Trout in the Milk and Other Visual Adventures*. Princeton, NJ: Princeton University Press.

Wainer, Howard. 2009. *Picturing the Uncertain World: How to Understand, Communicate, and Control Uncertainty through Graphical Display*. Princeton, NJ: Princeton University Press.

Wilkinson, Leland. 2005. *The Grammar of Graphics*. New York: Springer.

Embracing the Power

References

Börner, Katy. 2010. *Atlas of Science: Visualizing What We Know*. Cambridge, MA: The MIT Press.

Darwin, Charles. 1837. "Evolutionary Tree Sketch from *First Notebook on Transmutation of Species*." *Wikipedia, the Free Encyclopedia*. Accessed December 17, 2013. **http://en.wikipedia.org/wiki/File:Darwin_tree.png**.

Darwin, Charles. 1859. *On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life*. London: John Murray.

Friendly, Michael. 2008. "The Golden Age of Statistical Graphics." DataVis.ca. Accessed February 8, 2014. **http://www.datavis.ca/papers/ssc/GoldenAge-SSC-2x2.pdf**.

Koch, Tom. 2005. *Cartographies of Disease: Maps, Mapping, and Medicine*. Redlands, CA: ESRI Press.

ims25. 2008. "Nightingale's 'Coxcombs.'" Understanding Uncertainty, May 11. Accessed February 8, 2014. **http://understandinguncertainty.org/coxcombs**.

Institute for Health Metrics and Evaluation. 2013. "GBD Compare." Accessed December 17, 2013. **http://viz.healthmetricsandevaluation.org/gbd-compare**.

NASA. 2008. "Earthrise at Christmas" (Taken in 1968). Last modified March 23. **http://www.nasa.gov/multimedia/imagegallery/image_feature_102.html**.

Nightingale, Florence. 1858. "Diagram of the Causes of Mortality in the Army in the East" from *Notes on Matters Affecting the Health, Efficiency, and Hospital Administration of the British Army*. Wikimedia Commons. Accessed December 17, 2013. **http://commons.wikimedia.org/wiki/File:Nightingale-mortality.jpg**.

Nightingale, Florence. 1858. *Mortality of the British Army*. London: Harrison and Sons.

Porostocky, Thomas. 2013. "Causes of Untimely Death" (Infographic). In "Want to Save Lives? You Need a Map of What's Doing Us In" by Lee Simmons. *Wired*. Accessed December 17, 2013. **http://www.wired.com/wiredscience/2013/11/infoporn-causes-of-death**.

Rehmeyer, Julie. 2008. "Florence Nightingale: The Passionate Statistician." *Science News*, November 26. Accessed February 7, 2014. **https://www.sciencenews.org/article/florence-nightingale-passionate-statistician**.

Robertson, Murray, and John Emsley. 2005. *Visual Elements Periodic Table*. London, United Kingdom. Courtesy of the Royal Society of Chemistry Images, © 1999-2006 by Murray Robertson. In "2nd Iteration (2006): The Power of Reference Systems," *Places & Spaces: Mapping Science*, edited by Katy Börner and Deborah MacPherson. **http://scimaps.org**.

Wikimedia Foundation. 2014. "Earthrise." *Wikipedia, the Free Encyclopedia*. Accessed February 8, 2014. **http://en.wikipedia.org/wiki/Earthrise**.

Wikimedia Foundation. 2014. "Florence Nightingale." *Wikipedia, the Free Encyclopedia*. Accessed February 8, 2014. **http://en.wikipedia.org/wiki/Florence_Nightingale**.

Wikimedia Foundation. 2014. "John Snow." Wikipedia, the Free Encyclopedia. Accessed February 8, 2014. **http://en.wikipedia.org/wiki/John_Snow_%28physician%29.**

Disclaimer

References

Börner, Katy, and David E. Polley. 2014. *Visual Insights: A Practical Guide to Making Sense of Data*. Cambridge, MA: The MIT Press.

Frankel, Felice C., and Angela H. DePace. 2012. *Visual Strategies: A Practical Guide to Graphics for Scientists and Engineers*. New Haven, CT: Yale University Press.

Hansen, Derek, Ben Shneiderman, and Marc A. Smith. 2010. *Analyzing Social Media Networks with NodeXL: Insights from a Connected World*. Burlington, MA: Morgan Kaufmann.

Lima, Manuel. 2011. *Visual Complexity: Mapping Patterns of Information*. New York: Princeton Architectural Press.

McCandless, David. 2009. *The Visual Miscellaneum: A Colorful Guide to the World's Most Consequential Trivia*. New York: Harper Design.

Rendgen, Sandra. 2012. *Information Graphics*. Edited by Julius Wiedemann. Cologne, Germany: Taschen.

Tufte, Edward R. 1990. *Envisioning Information*. Cheshire, CT: Graphics Press.

Tufte, Edward R. 1997. *Visual Explanations: Images and Quantities, Evidence and Narrative*. Cheshire, CT: Graphics Press.

Tufte, Edward R. 2001. *The Visual Display of Quantitative Information*. 2nd ed. Cheshire, CT: Graphics Press.

Weissman, Jerry. 2009. *Presenting to Win: The Art of Telling Your Story*. Upper Saddle River, NJ: Pearson Education, Inc.

Yau, Nathan. 2011. *Visualize This: The FlowingData Guide to Design, Visualization, and Statistics*. Indianapolis, IN: Wiley.

Tree of Life

Image Credits

Tree of Life reproduced and altered with the kind permission of the Syndics of Cambridge University Library.

Causes of Mortality in the British Military during the Crimean War

References

Wikimedia Foundation. 2010. "Diagram of the Causes of Mortality in the Army in the East." Wikipedia, the Free Encyclopedia. Accessed September 18, 2010. **http://en.wikipedia.org/wiki/File:Nightingale-mortality.jpg**.

Image Credits

See Wikimedia 2010.

Spot Map of the Golden Square Cholera Outbreak

References

Wikimedia Foundation. 2005. "Spot Map of the Golden Square Cholera Outbreak." Wikibooks. Accessed September 18, 2014. **http://nl.wikibooks.org/wiki/Bestand:Snow-cholera-map.jpg**

Image Credits

See Wikimedia Foundation 2005.

The Antibiotic Abacus: Adding Up Drug Resistance

References

McCandless, David. 2014. *Knowledge is Beautiful*. London: HarperCollins.

Image Credits

The Antibiotic Abacus from **http://www.informationisbeautiful.net/visualizations/antibiotic-resistance**. Accessed September 18, 2014.

Data Credits

Foundational data for The Antibiotic Abacus is available at **http://bit.ly/KIB_Antibiotics**. Accessed September 14, 2014.

24 Needs-Driven Workflow Design

Visualization Taxonomies and Frameworks

References

Card, Stuart K., and Jock D. Mackinlay. 1997. "The Structure of the Information Visualization Design Space." In *Proceedings of the IEEE Symposium on Information Visualization*, 92–99. Los Alamitos, CA: IEEE Computer Society.

Chi, Ed H. 2000. "A Taxonomy of Visualization Techniques Using the Data State Reference Model." In *Proceedings of the IEEE Symposium on Information Visualization*, 69–75. Los Alamitos, CA: IEEE Computer Society.

Keim, Daniel A. 2001. "Visual Exploration of Large Data Sets." *Communications of the ACM* 44 (8): 38–44.

Mackinlay, Jock D. 1986. "Automating the Design of Graphical Presentations of Relational Information." *ACM Transactions on Graphics* 5 (2): 110–141.

Pfitzner, Darius, Vaughan Hobbs, and David Powers. 2003. "A Unified Taxonomic Framework for Information Visualization." In *Proceedings of the Asia-Pacific Symposium on Information Visualisation*, 24: 57–66. Darlinghurst, Australia: Australian Computer Society, Inc.

Shneiderman, Ben. 1996. "The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations." In *Proceedings of the IEEE Symposium on Visual Languages*, 336–343. Los Alamitos, CA: IEEE Computer Society.

Image Credits

Redesigned by Perla Mateo-Lujan based on an image from Chi 2000.

Workflow Design

References

Börner, Katy. 2010. *Atlas of Science: Visualizing What We Know*. Cambridge, MA: The MIT Press.

Reference System Versus Data Overlay

References

Bostock, Michael. 2012. "Non-Contiguous Cartogram." *Mbostock's Blocks* (blog), November 11. Accessed March 1, 2014. **http://bl.ocks.org/mbostock/4055908**. © Copyright 2013 Mike Bostock.

Bostock, Michael. 2013. "Voroni Arc Map." *Mbostock's Blocks* (blog), November 22. Accessed March 1, 2014. **http://bl.ocks.org/mbostock/7608400**.

Ittelson, William H. 1996. "Visual Perception of Markings." *Psychonomic Bulletin & Review* 3 (2): 171–187.

Image Credits

2012 Political Election Choropleth Map © 2012 M. E. J. Newman.

2012 Political Election Contiguous Cartogram © 2012 M. E. J. Newman.

See Bostock 2012.

See Bostock 2013.

U.S. Map of Contiguous States created by Samuel T. Mills.

Visualization Framework image created by Perla Mateo-Lujan and Tracey Theriault based on a concept by Katy Börner.

Software Credits

U.S. Map of Contiguous States rendered using the Sci2 Tool. **http://sci2.cns.iu.edu**. Accessed September 18, 2014.

26 Insight Need Types

References

Bertin, Jacques. 1983. *Semiology of Graphics*. Madison, WI: University of Wisconsin Press. 12.

Burkhard, Remo A. 2004. "Learning from Architects: The Difference between Knowledge Visualization and Information Visualization." In *Proceedings of the Eighth International Conference on Information Visualization (IV'04)*, 519–524.

Burkhard, Remo A. 2006. "Learning from Architects: Complementary Concept Mapping Approaches." Information Visualization 5: 225–234.

Card, Stuart K., and Jock D. Mackinlay. 1997. "The Structure of the Information Visualization Design Space." In *Proceedings of the IEEE Symposium on Information Visualization*, 92-99. Los Alamitos, CA: IEEE Computer Society.

Harris, Robert L. 1999. *Information Graphics: A Comprehensive Illustrated Reference*. New York: Oxford University Press.

Pirolli, Peter, and Stuart Card. 2005. "The Sensemaking Process and Leverage Points for Analyst Technology as Identified through Cognitive Task Analysis." In *Proceedings of the International Conference on Intelligence Analysis*, 2–4. McLean, VA: MITRE.

Framework

Basic Task Types

References

Bertin, Jacques. 1983. *Semiology of Graphics*. Madison, WI: University of Wisconsin Press.

Few, Stephen C. 2012. *Show Me The Numbers: Designing Tables and Graphs to Enlighten*. Burlingame, CA: Analytics Press.

Frankel, Felice C., and Angela H. DePace. 2012. *Visual Strategies: A Practical Guide to Graphics for Scientists and Engineers*. New Haven, CT: Yale University Press.

IBM. 2013. Many Eyes. Accessed December 18, 2013. **http://www-958.ibm.com/software/analytics/manyeyes**.

Juice Labs. 2013. Chart Chooser. Accessed December 18, 2013. **http://labs.juiceanalytics.com/chartchooser**.

Rendgen, Sandra. 2012. *Information Graphics*. Edited by Julius Wiedemann. Cologne, Germany: Taschen.

Wehrend, Stephen C., and Clayton Lewis. 1990. "A Problem-Oriented Classification of Visualization Techniques. In *Proceedings of the 1st Conference on Visualization '90*, 139–143. Los Alamitos, CA: IEEE Computer Society.

Yau, Nathan. 2011. *Visualize This: The FlowingData Guide to Design, Visualization, and Statistics*. Indianapolis, IN: Wiley.

Image Credits

Created by Perla Mateo-Lujan, with data provided by Katy Börner.

Interaction Types

References

Keim, Daniel A. 2001. "Visual Exploration of Large Data Sets." *Communications of the ACM* 44 (8): 38–44.

Shneiderman, Ben. 1996. "The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations." In *Proceedings of the IEEE Symposium on Visual Languages*, 336–343. Los Alamitos, CA: IEEE Computer Society.

Descriptions and Examples

Categorizing and Clustering

Image Credits

Image created by Perla Mateo-Lujan.

Distribution (also Outliers and Gaps)

Image Credits

Image created by Perla Mateo-Lujan.

Trends

References

Harris, Robert L. 1999. *Information Graphics: A Comprehensive Illustrated Reference*. New York: Oxford University Press.

Peltier, Jon. 2011. "Chart Busters: Pie Charts Can't Show Trendlines." *Peltier Tech Blog*, May 5. Accessed March 1, 2014. **http://peltiertech.com/WordPress/chart-busters-pie-charts-cant-show-trendlines.**

Image Credits

Time Spent on Weekends adapted by Perla Mateo-Lujan from Harris 1999 © Oxford University Press, 2000; *see* also Peltier 2011.

Composition (of Objects and of Text)

Image Credits

Image created by Perla Mateo-Lujan.

Software Credits

Network layout rendered using the Sci2 Tool. **http://sci2.cns.iu.edu**. Accessed September 14, 2014.

Comparison

References

Harris, Robert L. 1999. *Information Graphics: A Comprehensive Illustrated Reference*. New York: Oxford University Press.

Image Credits

Demographic Pyramid adapted by Perla Mateo-Lujan from Harris 1999 © Oxford University Press, 2000.

Geospatial Location

References

ESRI. 2013. "Within." Accessed December 18, 2013. **http://edndoc.esri.com/arcsde/9.1/general_topics/understand_spatial_relations.htm#Within**.

Data Credits

IVMOOC 2014 data as of May 29, 2014. **http://ivmooc.cns.iu.edu**. Accessed September 18, 2014.

Software Credits

Network layout rendered using the Sci2 Tool. **http://sci2.cns.iu.edu**. Accessed September 18, 2014.

Contributors

Data compiled by Robert P. Light; map rendered by Scott Emmons and redesigned by Perla Mateo-Lujan.

Correlations and Relationships

References

Padgett, John F. 1986. "Florentine Families Dataset." Accessed March 1, 2014. **http://www.casos.cs.cmu.edu/computational_tools/datasets/sets/padgett**.

Image Credits

Image created by Katy Börner and Perla Mateo-Lujan, using data from Padgett 1986.

Data Credits

See Padgett 1986.

Software Credits

Network layout rendered using the Sci2 Tool. **http://sci2.cns.iu.edu**. Accessed September 14, 2014.

28 Data Scale Types

References

Cameron, William Bruce. 1963. *Informal Sociology: A Casual Introduction to Sociological Thinking*. New York: Random House.

Stevens, Stanley S. 1946. "On the Theory of Scales of Measurement." *Science* 103 (2684): 677–680.

Velleman, Paul F., and Leland Wilkinson. 1993. "Nominal, Ordinal, Interval, and Ratio Typologies Are Misleading." *The American Statistician* 47 (1): 65–72.

Framework

References

Bertin, Jacques. 1983. *Semiology of Graphics*. Madison, WI: University of Wisconsin Press. 34.

Dunn, Dana S. 2000. *Statistics and Data Analysis for the Behavioral Sciences*. New York: McGraw-Hill

Harris, Robert L. 1999. *Information Graphics: A Comprehensive Illustrated Reference*. New York: Oxford University Press.

MacEachren, Alan M. 2004. *How Maps Work: Representation, Visualization, and Design*. New York: Guilford.

Munzner, Tamara. 2014. *Information Visualization: Principles, Techniques, and Practice*. Natick, MA: AK Peters.

Stevens, Stanley S. 1946. "On the Theory of Scales of Measurement." *Science* 103 (2684): 677–680.

Image Credits

© Tamara Munzner, "Visualization Principles," 2011.

Conversions

References

Abelson, Robert P., and John W. Tukey. 1963. "Efficient Utilization of Non-Numerical Information in Quantitative Analysis: General Theory and the Case of Simple Order." *The Annals of Mathematical Statistics* 34 (4): 1347–1369.

Kruskal, Joseph B. 1964. "Nonmetric Multidimensional Scaling: A Numerical Method." *Psychometrika* 29 (2): 115–129.

Tukey, John W. (1961) 1986. "Data Analysis and Behavioral Science, or Learning to Bear the Quantitative Man's Burden by Shunning Badmandments." In *The Collected Works of John W. Tukey*, vol. III, edited by Lyle V. Jones, 391–484. Belmont, CA: Wadsworth, Inc.

Velleman, Paul F., and Leland Wilkinson. 1993. "Nominal, Ordinal, Interval, and Ratio Typologies Are Misleading." *The American Statistician* 47 (1): 65–72.

Mathematical Operations

References

Stevens, Stanley S. 1946. "On the Theory of Scales of Measurement." *Science* 103 (2684): 677–680.

Image Credits

Image adapted from Stevens 1946.

Descriptions and Examples

References

Chrisman, Nicholas R. 1998. "Rethinking Levels of Measurement for Cartography." *Cartography and Geographic Information Science* 25 (4): 231–242.

Nominal Scale

References

VIVO Project. 2014. "NetSci: Index of Contents." Accessed March 1, 2014. **http://vivo-netsci.cns.iu.edu/vivo12/browse.**

Data Credits

See VIVO Project 2014.

Ordinal Scale

References

Wikimedia Foundation. 2014. "Example Likert Scale." *Wikipedia, the Free Encyclopedia*. Accessed March 1, 2014. **http://en.wikipedia.org/wiki/File:Example_Likert_Scale.svg**.

Image Credits

Redesigned with permission under a Creative Commons license from Nicholas Smith, © 2012.

Interval Scale

Image Credits

Image rendered by Perla Mateo-Lujan based on a concept by Katy Börner.

Ratio Scale

References

Stevens, Stanley S. 1946. "On the Theory of Scales of Measurement." *Science* 103 (2684): 677–680.

30 Visualization Types

References

Bertin, Jacques. 1983. *Semiology of Graphics*. Madison, WI: University of Wisconsin Press.

Harris, Robert L. 1999. *Information Graphics: A Comprehensive Illustrated Reference*. New York: Oxford University Press.

Heer, Jeffrey, Michael Bostock, and Vadim Ogievetsky. 2010. "A Tour Through the Visualization Zoo." *Communications of the ACM* 53 (6): 59–67.

Wattenberg, Martin. 2010. Interview. In *Journalism in the Age of Data* (Video). Accessed January 31, 2014. **http://datajournalism.stanford.edu**.

Framework

References

Bertin, Jacques. 1983. *Semiology of Graphics*. Madison, WI: University of Wisconsin Press.

Engelhardt, Yuri. 2002. "The Language of Graphics: A Framework for the Analysis of Syntax and Meaning in Maps, Charts, and Diagrams." PhD diss., University of Amsterdam.

Harris, Robert L. 1999. *Information Graphics: A Comprehensive Illustrated Reference*. New York: Oxford University Press.

Microsoft. 2013. Excel. Accessed December 20, 2013. **http://office.microsoft.com/en-us/excel**.

Shneiderman, Ben. 1996. "The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations." In *Proceedings of the IEEE Symposium on Visual Languages*, 336–343. Los Alamitos, CA: IEEE Computer Society.

Image Credits

Visualization Types table designed by Perla Mateo-Lujan.

Descriptions and Examples

References

Harris, Robert L. 1999. *Information Graphics: A Comprehensive Illustrated Reference*. New York: Oxford University Press.

Tables

Image Credits

Image created by Perla Mateo-Lujan.

Charts

Image Credits

Pie Chart, Doughnut Chart, and Bubble Chart examples by Samuel T. Mills.

Tag Cloud example by Samuel T. Mills and Perla Mateo-Lujan.

Graphs

References

Wikimedia Foundation. 2014. "U.S. First Class Postage Rate." *Wikipedia, the Free Encyclopedia*. September 21, 2014. **http://en.wikipedia.org/wiki/History_of_United_States_postage_rates#mediaviewer/File:US_Postage_History.svg**.

Image Credits

See Wikimedia Foundation 2014.

Parallel Coordinate Graph designed by Perla Mateo-Lujan.

Maps

References

Yunker, Jon. 2007. "Country Codes of the World." Accessed March 1, 2014. **http://bytelevel.com/map/ccTLD.html**.

Image Credits

Adapted from Yunker 2007.

Solar Light Map of Cambridge, MA © Mapdwell LLC. All rights reserved.

Network Layouts

Trees

Image Credits

Tree View and Force-Directed Layout images by Samuel T. Mills.

Treemap image by Perla Mateo-Lujan.

Networks

References

Krzywinski, Martin. 2011. Hive Plots. Accessed January 30, 2014. **http://www.hiveplot.net**.

Krzywinski, Martin, Inanc Birol, Steven J. M. Jones, and Marco A. Marra. 2011. "Hive Plots: Rational Approach to Visualizing Networks." *Briefings in Bioinformatics* 13 (5): 627–644.

Lima, Manuel. 2014. *The Book of Trees: Visualizing Branches of Knowledge*. New York: Princeton Architectural Press.

Wattenberg, Martin. 2014. The Shape of Song. Accessed January 30, 2014. **http://www.turbulence.org/Works/song**.

Image Credits

Arc Graph and Force-Directed Layout by Samuel T. Mills.

32 Graphic Symbol Types

References

Harris, Robert L. 1999. *Information Graphics: A Comprehensive Illustrated Reference*. New York: Oxford University Press. 379.

Klee, Paul. 1964. *The Diaries of Paul Klee*. Berkeley, CA: University of California Press. [Quotation, p. 183].

Framework

References

Bertin, Jacques. 1983. *Semiology of Graphics*. Madison, WI: University of Wisconsin Press.

Börner, Katy. 2017. *Atlas of Forecasts: Predicting and Broadcasting Science*. Cambridge, MA: The MIT Press.

Engelhardt, Yuri. 2002. "The Language of Graphics: A Framework for the Analysis of Syntax and Meaning in Maps, Charts, and Diagrams." PhD diss., University of Amsterdam.

Harris, Robert L. 1999. *Information Graphics: A Comprehensive Illustrated Reference*. New York: Oxford University Press. 231.

Horn, Robert E. 1998. *Visual Language: Global Communication for the 21st Century*. Bainbridge Island, WA: MacroVU, Inc.

MacEachren, Alan M. 2004. *How Maps Work: Representation, Visualization, and Design*. New York: Guilford. 271.

Tufte, Edward R. 1990. *Envisioning Information*. Cheshire, CT: Graphics Press.

Tufte, Edward R. 1997. *Visual Explanations: Images and Quantities, Evidence and Narrative*. Cheshire, CT: Graphics Press.

Tufte, Edward R. 2001. *The Visual Display of Quantitative Information*. 2nd ed. Cheshire, CT: Graphics Press.

Tufte, Edward R. 2007. *Beautiful Evidence*. Cheshire, CT: Graphics Press.

Wilkinson, Leland. 2005. *The Grammar of Graphics*. New York: Springer.

Instantiation

References

Bertin, Jacques. 1981. *Graphics and Graphics Information Processing*. Berlin: Walter de Gruyter.

MacEachren, Alan M. 2004. *How Maps Work: Representation, Visualization, and Design*. New York: Guilford. 271.

Image Credits

Image from Bertin 1981, © De Gruyter 1981.

Combinations

References

Turner, Eugene. 1977. *Life in Los Angeles*. Accessed March 1, 2014. **http://media-cache-ak0.pinimg.com/originals/11/14/9b/11149b5830447ed2c51c68c61fce1285.jpg**.

Wertheimer, Max. 1923. "Untersuchungen zur Lehre von der Gestalt. II." *Psychologische Forschung* 4 (1): 301–350.

Image Credits

Eugene Turner, 1977. "Life in Los Angeles" California State University Northridge.

Descriptions and Examples

References

Elmer, Martin. 2013. "The Trouble with Chernoff." *Map Hugger* (blog). Accessed December 20, 2013. **http://maphugger.com/post/44499755749/the-trouble-with-chernoff**.

Holten, Danny and Jarke J. van Wijk. "A User Study on Visualizing Directed Edges in Graphs." In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2299-2308. New York: ACM.

Kimerling, A. Jon, Aileen R. Buckley, and Phillip C. Muehrcke. 2009. *Map Use: Reading and Analysis*. 6th ed. Redlands, CA: ESRI.

Geometric Symbols

Lines

Image Credits

Image created by Perla Mateo-Lujan.

Areas

Image Credits

Image created by Perla Mateo-Lujan.

Surfaces

References

Morris, Steven, Camille DeYong, Zheng Wu, Sinan Salman, Dagmawi Yemenu. 2002. "DIVA: A Visualization System for Exploring Document Databases for Technology Forecasting." *Computers and Industrial Engineering* 43 (4): 841–862.

Image Credits

Reprinted from Morris et al. 2002 © National Academy of Sciences, U.S.A.

Linguistic Symbols

Image Credits

Typeface examples: Microsoft (Cambria); Monotype (Arial); Adobe (Adobe Caslon).

Proportional and Monospace typeface examples: Adobe (Adobe Garamond); Courier (public domain).

Dingbat examples: Microsoft (Webdings); Adobe (Conventional Dingbats).

Pictorial Symbols

Statistical Glyphs

References

Harris, Robert L. 1999. *Information Graphics: A Comprehensive Illustrated Reference*. New York: Oxford University Press.

Vande Moere, Andrew. 2013. *Information Aesthetics* (blog). Accessed December 20, 2013. **http://infosthetics.com**.

Image Credits

Statistical Glyphs adapted by Perla Mateo-Lujan from Harris 1999 © Oxford University Press, 2000.

Word Count Sparklines courtesy of Andrew Vande Moere. **http://infosthetics.com**. Accessed September 18, 2014.

Contributors

André Skupin provided expert comments.

34 Graphic Variable Types

References

Bertin, Jacques. 1981. *Graphics and Graphics Information Processing*. Berlin: Walter de Gruyter.

Card, Stuart K., and Jock D. Mackinlay. 1997. "The Structure of the Information Visualization Design Space." In *Proceedings of the IEEE Symposium on Information Visualization*, 92–99. Los Alamitos, CA: IEEE Computer Society.

DEVise Development Group. 2010. Home Page. Accessed December 20, 2013. **http://pages.cs.wisc.edu/~devise**.

Green, Marc. 1998. "Towards a Perceptual Science of Multidimensional Data Visualization: Bertin and Beyond." *ERGO/GERO Human Factors Science*. Accessed December 20, 2013. **http://graphics.stanford.edu/courses/cs448b-06-winter/papers/Green_Towards.pdf**.

Kimerling, A. Jon, Aileen R. Buckley, Phillip C. Muehrcke, Juliana O. Muehrcke. 2011. *Map Use: Reading, Analysis, Interpretation*. 7th ed. Redlands, CA: Esri Press Academic.

Livny, Miron, Raghu Ramakrishnan, Kevin S. Beyer, Guangshun Chen, Donko Donjerkovic, Shilpa Lawande, Jussi P. Myllymaki, and Kent Wenger. 1997. "DEVise: Integrated Querying and Visual Exploration of Large Datasets." In *Proceedings of the 1997 ACM SIGMOD International Conference on Management of Data*, 301–312. New York: ACM.

MacEachren, Alan M. 2004. *How Maps Work: Representation, Visualization, and Design*. New York: Guilford.

Mackinlay, Jock D. 1986. "Automating the Design of Graphical Presentations of Relational Information." *ACM Transactions on Graphics* 5 (2): 110–141.

Roth, Steven F., and Mattis, Joe. 1990. "Data Characterization for Intelligent Graphics Presentation." In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 193–200. New York: ACM.

SciencesPo. 2010. "La Graphique, Jacques Bertin 2000." Accessed December 20, 2013. **http://cartographie.sciences-po.fr/fr/la_graphique_jacques_bertin2**.

Shneiderman, Ben. 1996. "The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations." In *Proceedings of the IEEE Symposium on Visual Languages*, 336–343. Los Alamitos, CA: IEEE Computer Society.

Stevens, Stanley S. 1946. "On the Theory of Scales of Measurement." *Science* 103 (2684): 677–680.

Stolte, Chris, and Pat Hanrahan. 2002. "Polaris: A System for Query, Analysis and Visualization of Multi-Dimensional Relational Databases." *IEEE Transactions on Visualization and Computer Graphics* 8 (1): 62–65.

Treinish, Lloyd A. 1999. "A Function-Based Data Model for Visualization." In *Proceedings of IEEE Visualization '99: Late Breaking Hot Topics*, 73–76. Los Alamitos, CA: IEEE Computer Society.

Wainer, Howard. 1983. Foreword to *Semiology of Graphics* by Jacques Bertin. Madison, WI: University of Wisconsin Press. ix.

Wilkinson, Leland. 2005. *The Grammar of Graphics*. New York: Springer.

Framework

References

Bertin, Jacques. 1983. *Semiology of Graphics*. Madison, WI: University of Wisconsin Press.

Bertin, Jacques. 1981. *Graphics and Graphics Information Processing*. Berlin: Walter de Gruyter.

Healey, Christopher G., and James T. Enns. 2012. "Attention and Visual Memory in Visualization and Computer Graphics." *IEEE Transactions on Visualization and Computer Graphics* 18 (7): 1170–1188.

Horn, Robert E. 1998. *Visual Language: Global Communication for the 21st Century*. Bainbridge Island, WA: MacroVU, Inc.

Jacobson, Robert E., ed. 1999. *Information Design*. Cambridge, MA: MIT Press.

Kosslyn, Stephen M. 1994. *Elements of Graph Design*. New York: W.H. Freeman and Co.

MacEachren, Alan M. 2004. *How Maps Work: Representation, Visualization, and Design*. New York: Guilford. 275–276.

Marriott, Kim, and Bernd Meyer, eds. 1998. *Visual Language Theory*. New York: Springer.

Mijksenaar, Paul. 1997. *Visual Function: An Introduction to Information Design*. New York: Princeton Architectural Press.

Mollerup, Per. 1997. *Marks of Excellence*. London: Phaidon Press.

Wainer, Howard. 1997. *Visual Revelations: Graphical Tales of Fate and Deception from Napoleon Bonaparte to Ross Perot*. New York: Copernicus.

Ware, Colin. 2000. *Information Visualization: Perception for Design*. Burlington, MA: Morgan-Kaufman.

Wertheimer, Max. 1923. "Untersuchungen zur Lehre von der Gestalt. II." *Psychologische Forschung* 4 (1): 301–350.

Wilkinson, Leland. 2005. *The Grammar of Graphics*. New York: Springer.

Wurman, Richard Saul. 1997. *Information Architects*. New York: Graphis Inc.

Combinations

References

Wilkinson, Leland. 2005. *The Grammar of Graphics*. New York: Springer.

Perception Accuracy

References

Cleveland, William S., and Robert McGill. 1984. "Graphical Perception: Theory, Experimentation, and Application to the Development of Graphical Methods." *Journal of the American Statistical Association* 79 (387): 531–554.

Mackinlay, Jock D. 1986. "Automating the Design of Graphical Presentations of Relational Information." *ACM Transactions on Graphics* 5 (2): 110–141.

Spence, Robert. 2007. *Book Information Visualization: Design for Interaction*. 2nd ed. Harlow, UK: Pearson/Prentice.

Image Credits

Image adapted with permission of Jock Mackinlay.

Image courtesy of Robert Spence.

Descriptions and Examples

References

Harris, Robert L. 1999. *Information Graphics: A Comprehensive Illustrated Reference*. New York: Oxford University Press.

Paper Leaf Design. 2011. "Elements of Design: A Quick Reference Sheet." *Paper Leaf* (blog). Accessed December 20, 2013. **http://www.paper-leaf.com/blog/wp-content/uploads/2011/02/EoD_White_1440.jpg**.

Retinal

References

Wilkinson, Leland. 2005. *The Grammar of Graphics*. New York: Springer. 317.

Form

References

Börner, Katy. 2017. *Atlas of Forecasts: Predicting and Broadcasting Science*. Cambridge, MA: The MIT Press.

Image Credits

Images from Harris 1999 © Oxford University Press, 1999.

Color

References

Brewer, Cynthia A. 1994. "Color Use Guidelines for Mapping and Visualization." In *Visualization in Modern Cartography*, edited by Alan M. MacEachren and D. R. Fraser Taylor, 123–147. Oxford, UK: Pergamon.

Brewer, Cynthia A. 1999. "Color Use Guidelines for Data Representation." In *Proceedings of the Section on Statistical Graphics, American Statistical Association*, 55–60. Alexandria, VA: ASA.

Brewer, Cynthia A., and Mark Harrower. 2013. "ColorBrewer 2.0: Color Advice for Cartography." Accessed December 20, 2013. **http://colorbrewer2.org**.

Fairchild, Mark D. 1998. *Color Appearance Models*. Reading, MA: Addison-Wesley.

Mersey, Janet E. 1990. "Color and Thematic Map Design: The Role of Colour Scheme and Map Complexity in Choropleth Map Communication." *Cartographica* 27 (3): 1–167.

Paper Leaf Design. 2014. "Color Theory: Quick Reference Sheet for Designers." Accessed February 10, 2014. **http://www.paper-leaf.com/samples/designfreebies/ColorTheory_Print.pdf**.

Travis, David. 1991. *Effective Color Displays: Theory and Practice*. London: Academic Press.

Tufte, Edward R. 1990. "Color and Information." In *Envisioning Information*. Cheshire, CT: Graphics Press.

Ware, Colin. 2000. *Information Visualization: Perception for Design*. Burlington, MA: Morgan-Kaufman.

From Wikimedia Foundation. 2014. "HSV Color Solid Cylinder Alpha Lowgamma." *Wikipedia, the Free Encyclopedia*. Accessed March 1, 2014. **http://en.wikipedia.org/wiki/File:HSV_color_solid_cylinder_alpha_lowgamma.png**.

Image Credits

See Wikimedia Foundation 2014.

Texture

References

Caivano, Jose Luis. 1990. "Visual Texture as a Semiotic System." *Semiotica* 80 (3-4): 239–252.

MacEachren, Alan M. 2004. *How Maps Work: Representation, Visualization, and Design*. New York: Guilford.

Paper Leaf Design. 2011. "Elements of Design: A Quick Reference Sheet." *Paper Leaf* (blog). Accessed December 20, 2013. **http://www.paper-leaf.com/blog/wp-content/uploads/2011/02/EoD_White_1440.jpg**.

Image Credits

Images from Harris 1999 © Oxford University Press, 1999.

Optics

References

Frankel, Felice C., and Angela H. DePace. 2012. *Visual Strategies: A Practical Guide to Graphics for Scientists and Engineers*. New Haven, CT: Yale University Press. 106.

Wikimedia Foundation. 2013. "Stereoscopic Depth Rendition." *Wikipedia, the Free Encyclopedia*. Accessed December 20, 2013. **http://en.wikipedia.org/wiki/Stereoscopic_depth_rendition**.

Image Credits

See Wikimedia Foundation 2013.

Graphic Variable Types Versus Graphic Symbol Types

Image Credits

Four-page table created by Perla Mateo-Lujan and Katy Börner.

40 User Needs Acquisition

References

Markoff, John. 2011. "Steven P. Jobs, 1955–2011: Apple's Visionary Redefined Digital Age." *The New York Times*, October 5. Accessed December 20, 2013. **http://www.nytimes.com/2011/10/06/business/steve-jobs-of-apple-dies-at-56.html?pagewanted=all**.

General Considerations

Production Versus Consumption

References

Hook, Peter A., and Katy Börner. 2005. "Educational Knowledge Domain Visualizations: Tools to Navigate, Understand, and Internalize the Structure of Scholarly Knowledge and Expertise." In *New Directions in Cognitive Information Retrieval*, edited by Amanda Spink and Charles Cole, 187–208. Dordrecht, Netherlands: Springer-Verlag.

Iterative Prototyping and Replication

References

Felix, Elliot. 2010. "Design Strategy." Accessed March 1, 2014. **http://elliotfelix.files.wordpress.com/2010/03/elliot-felix-design-strategy1.jpg**.

Image Credits

Design Strategy Chart © Elliot Felix 2010.

Users and Needs

User Types

References

Burkhard, Remo A. 2005. *Knowledge Visualization: The Use of Complementary Visual Representations for the Transfer of Knowledge. A Model, a Framework, and Four New Approaches*. PhD Thesis, Eidgenössische Technische Hochschule ETH Zürich.

Demographics

References

Börner, Katy. 2017. *Atlas of Forecasts: Predicting and Broadcasting Science*. Cambridge, MA: The MIT Press.

Needs Acquisition

Surveys

References

SurveyMonkey. 2014. SurveyMonkey Home Page. Accessed January 31, 2014. **https://www.surveymonkey.com**.

Apprentice Model

Contributors

Bradford W. Paley introduced the concept of an apprentice.

Lead User Analysis

References

Börner, Katy. 2010. *Atlas of Science: Visualizing What We Know*. Cambridge, MA: The MIT Press.

Lead User Infosite. 2014. Leaduser.com Home Page. Accessed March 1, 2014. **http://www.leaduser.com**.

Von Hippel, Eric. 1986. "Lead Users: A Source of Novel Product Concepts." *Management Science* 32 (7): 791-805.

Wikimedia Foundation. 2014. "Lead User." *Wikipedia, the Free Encyclopedia*. Accessed March 1, 2014. **http://en.wikipedia.org/wiki/Lead_user**.

Conjoint Analysis

References

Orme, Bryan K. 2005. *Getting Started with Conjoint Analysis.* Madison, WI: Research Publishers LLC.

Wikimedia Foundation. 2014. "Conjoint Analysis (Marketing)." *Wikipedia, the Free Encyclopedia*. Accessed March 1, 2014. **http://en.wikipedia.org/wiki/Conjoint_analysis_%28marketing%29**.

User Mining and Modeling

References

Brody, Tim, Stevan Harnad, and Leslie Carr. 2006. "Earlier Web Usage Statistics as Predictors of Later Citation Impact." *JASIST* 57 (8): 1060–1072.

Fu, Lawrence, and Constantin Aliferis. 2009. Method for Predicting Citation Counts. US Patent 20090157585 A1, filed November 7, 2008, and issued June 18, 2009.

Meho, Lokman I. 2006. "The Rise and Rise of Citation Analysis." arXiv. Accessed March 1, 2014. **http://arxiv.org/abs/physics/0701012**.

42 Data Acquisition

References

Börner, Katy. 2010. *Atlas of Science: Visualizing What We Know*. Cambridge, MA: The MIT Press.

Wells, Robert, and Judith A. Whitworth. 2007. "Assessing Outcomes of Health and Medical Research: Do We Measure What Counts or Count What We Can Measure?" *Australia and New Zealand Health Policy*, 4:14. Accessed December 20, 2013. **http://www.anzhealthpolicy.com/content/4/1/14**.

Data Aggregation

References

Harris, Robert L. 1999. *Information Graphics: A Comprehensive Illustrated Reference*. New York: Oxford University Press.

Matching Data Analysis Types

References

Börner, Katy. 2010. *Atlas of Science: Visualizing What We Know*. Cambridge, MA: The MIT Press.

Börner, Katy, Richard Klavans, Michael Patek, Angela Zoss, Joseph R. Biberstine, Robert Light, Vincent Larivière, and Kevin W. Boyack. 2012. "Design and Update of a Classification System: The UCSD Map of Science." *PLoS One* 7 (7): e39464. Accessed October 31, 2013. **http://sci.cns.iu.edu/ucsdmap**.

Image Credits

Image created by Katy Börner and Perla Mateo-Lujan.

Software Credits

Images rendered using the Sci2 Tool. **http://sci2.cns.iu.edu**. Accessed September 14, 2014.

Matching Graphic Variable Types

References

Bertin, Jacques. 1983. *Semiology of Graphics*. Madison, WI: University of Wisconsin Press. 33, 357.

Börner, Katy. 2017. *Atlas of Forecasts: Predicting and Broadcasting Science*. Cambridge, MA: The MIT Press.

Brewer, Cynthia A., and Mark Harrower. 2013. "ColorBrewer 2.0: Color Advice for Cartography." Accessed December 20, 2013. **http://colorbrewer2.org**.

Fischer, Eric. 2012. *Language Communities of Twitter*. Oakland, CA. Courtesy of Eric Fischer. In "8th Iteration (2012): Science Maps for Kids," *Places & Spaces: Mapping Science*, edited by Katy Börner and Michael J. Stamper. **http://scimaps.org**.

Harrower, Mark, and Cynthia A. Brewer. 2003. "ColorBrewer.org: An Online Tool for Selecting Colour Schemes for Maps." *The Cartographic Journal* 40 (1): 27–37.

Slocum, Terry A., Robert B. McMaster, Fritz C. Kessler, and Hugh H. Howard. 1999. *Thematic Cartography and Geovisualization*. Upper Saddle River, NJ: Prentice Hall.

44 Statistical Studies

References

BookRags Media Network. 2014. "Evan Davis Quotes." Accessed January 21, 2014. **http://www.brainyquote.com/quotes/quotes/e/evandavis481959.html**.

Cleveland, William S. 1993. *Visualizing Data*. Summit, NJ: Hobart Press.

Cleveland, William S. 1994. *The Elements of Graphing Data*. Summit, NJ: Hobart Press.

Few, Stephen C. 2012. *Show Me The Numbers: Designing Tables and Graphs to Enlighten*. Burlingame, CA: Analytics Press.

Tukey, John W. 1977. *Exploratory Data Analysis*. Reading, MA: Addison-Wesley.

Exploratory Versus Confirmatory

References

Börner, Katy. 2017. *Atlas of Forecasts: Predicting and Broadcasting Science*. Cambridge, MA: The MIT Press.

Tukey, John W. 1977. *Exploratory Data Analysis*. Reading, MA: Addison-Wesley.

Data Distributions

References

Anscombe, Francis J. 1973. "Graphs in Statistical Analysis." *American Statistician* 27 (1): 17–21.

Börner, Katy. 2010. *Atlas of Science: Visualizing What We Know*. Cambridge, MA: The MIT Press.

Börner, Katy, Jeegar T. Maru, and Robert L. Goldstone. 2004. "The Simultaneous Evolution of Author and Paper Networks." *PNAS* 101 (Suppl. 1): 5266–73.

Redner, Sidney. 1998. "How Popular is Your Paper? An Empirical Study of the Citation Distribution." *The European Physical Journal B* 4 (2):131–134.

Wikimedia Foundation. 2013. "Frank Anscombe." *Wikipedia, the Free Encyclopedia*. Accessed December 20, 2013. **http://en.wikipedia.org/wiki/Francis_Anscombe**.

Wikimedia Foundation. 2013. "Normal Distribution." *Wikipedia, the Free Encyclopedia*. Accessed December 20, 2013. **http://en.wikipedia.org/wiki/Normal_distribution**.

Wikimedia Foundation. 2014. "Standard Deviation Diagram." *Wikipedia, the Free Encyclopedia*. Accessed January 31, 2014. **http://en.wikipedia.org/wiki/File:Standard_deviation_diagram.svg**.

Image Credits

Image created by Katy Börner and Perla Mateo-Lujan.

Curve Fitting

References

Harris, Robert L. 1999. *Information Graphics: A Comprehensive Illustrated Reference*. New York: Oxford University Press.

Image Credits

Images adapted from Harris 1999 © Oxford University Press, 1999.

Correlations

References

Playfair, William. 2005 (1786). *The Commercial and Political Atlas and Statistical Breviary*. Edited by Howard Wainer and Ian Spence. New York: Cambridge University Press.

Staff. 2007. "Worth a Thousand Words." The Economist, December 19. Accessed January 6, 2014. **http://www.economist.com/node/10278643?story_id=10278643**.

Wikimedia Foundation. 2014. "William Playfair." *Wikipedia, the Free Encyclopedia*. Accessed January 6, 2014. **http://en.wikipedia.org/wiki/William_Playfair**.

Wheat Prices Versus Wages

Image Credits

Chart: Showing at One View the Price of the Quarter of Wheat, and Wages of Labour by the Week, from 1565 to 1821 courtesy of Beinecke Rare Book and Manuscript Library, Yale University.

Magnet States Versus Sticky States

References

Pew Research Center. 2008. "American Mobility: Who Moves? Who Stays Put? Where's Home?" Pew Research: Social & Demographic Trends. Accessed January 31, 2014. **http://pewsocialtrends.org/files/2011/04/American-Mobility-Report-updated-12-29-08.pdf**.

Pew Research Center. 2014. "Map: U.S. Migration Flows." Pew Research: Social & Demographic Trends. Accessed January 6, 2014. **http://www.pewsocialtrends.org/2008/12/17/u-s-migration-flows**.

Data Credits

U.S. Census Bureau Data.

All of Inflation's Little Parts

References

Bloch, Matthew, Shan Carter, and Amanda Cox. 2008. "All of Inflation's Little Parts." *The New York Times*, May 3. Accessed January 6, 2014. **http://www.nytimes.com/interactive/2008/05/03/business/20080403_SPENDING_GRAPHIC.html**.

Data Credits

Bureau of Labor Statistics

Michael Balzer, University of Konstanz (Germany)

Prison Expenditures for Brooklyn, New York City

References

Spatial Information Design Lab. 2008. *The Pattern*. Published by the Graduate School of Architecture, Planning and Preservation of Columbia University, New York, NY.

Image Credits

Prison Expenditure by Census Block, Brooklyn, New York, 2003 by Laura Kurgan, Eric Cadora, David Reinfurt, and Sarah Williams. Million Dollar Blocks Project, 2006, Spatial Information Design Lab, GSAPP, Columbia University.

46 Statistical Visualization Types

References

Carr, Daniel B., and Sarah M. Nusser. 1995. "Converting Tables to Plots: A Challenge from Iowa State." *Statistical Computing & Statistical Graphics Newsletter* 6: 11–18.

Cleveland, William S. 1993. *Visualizing Data*. Summit, NJ: Hobart Press.

Cleveland, William S. 1994. *The Elements of Graphing Data*. Summit, NJ: Hobart Press.

Few, Stephen C. 2012. *Show Me The Numbers: Designing Tables and Graphs to Enlighten*. Burlingame, CA: Analytics Press.

Friendly, Michael. 2008. "The Golden Age of Statistical Graphics." *Statistical Science* 23 (4): 502–535.

Gelman, Andrew, Cristian Pasarica, and Rahul Dodhia. 2002. "Let's Practice What We Preach: Turning Tables into Graphs." *The American Statistician* 56 (2): 121–130.

Harris, Robert L. 1999. *Information Graphics: A Comprehensive Illustrated Reference*. New York: Oxford University Press.

Tukey, John W. 1977. *Exploratory Data Analysis*. Reading, MA: Addison-Wesley.

Wainer, Howard. 1984. "How to Display Data Badly." *The American Statistician* 38 (2): 137–147.

Glyphs

Error Bar

References

NuMBerS Project Team. 2014. "Graphs and Charts." Accessed January 7, 2014. **http://web.anglia.ac.uk/numbers/graphsCharts.html**.

Wikimedia Foundation. 2014. "Error Bar." *Wikipedia, the Free Encyclopedia*. Accessed January 7, 2014. **http://en.wikipedia.org/wiki/Error_bar**.

Image Credits

Image rerendered by Perla Mateo-Lujan from an image by NuMBerS Project Team 2014.

Box-and-Whisker Symbol

References

Harris, Robert L. 1999. *Information Graphics: A Comprehensive Illustrated Reference*. New York: Oxford University Press.

NuMBerS Project Team. 2014. "Graphs and Charts." Accessed January 7, 2014. **http://web.anglia.ac.uk/numbers/graphsCharts.html**.

Tukey, John W. 1977. *Exploratory Data Analysis*. Reading, MA: Addison-Wesley.

Image Credits

Image rerendered by Perla Mateo-Lujan from an image by NuMBerS Project Team 2014.

Number of Co-Authors per Year graph: data compiled by Katy Börner; graph rendered by Robert P. Light; design by Perla Mateo-Lujan.

Data Credits

Albert-László Barabási Web of Science data available at **http://wiki.cns.iu.edu/display/SCI2TUTORIAL/2.5+Sample+Datasets.** Accessed September 18, 2014.

Software Credits

Rendered using the Sci2 Tool. **http://sci2.cns.iu.edu**. Accessed September 18, 2014.

Sparkline

References

Duggirala, Purna. 2010. "What are Excel Sparklines & How to Use Them." Accessed January 7, 2014. **http://chandoo.org/wp/2010/05/18/excel-sparklines-tutorial**.

Tufte, Edward R. 2007. *Beautiful Evidence*. Cheshire, CT: Graphics Press.

Yaffa, Joshua. 2011. "The Information Sage." *The Washington Monthly*, May/June.

Image Credits

Image from Duggirala 2010.

Graphs

Comparisons

References

Harris, Robert L. 1999. *Information Graphics: A Comprehensive Illustrated Reference*. New York: Oxford University Press.

Wong, Dona M. 2010. *The Wall Street Journal Guide to Information Graphics: The Dos and Don'ts of Presenting Data, Facts, and Figures*. New York: W.W. Norton & Company. 69.

Image Credits

Image adapted from Harris 1999 © Oxford University Press, 1999.

Bar Graph

References

Harris, Robert L. 1999. *Information Graphics: A Comprehensive Illustrated Reference*. New York: Oxford University Press.

Image Credits

Image adapted from Harris 1999 © Oxford University Press, 1999.

Radar Graph

References

Bertin, Jacques. 1983. *Semiology of Graphics*. Madison, WI: University of Wisconsin Press. 50.

Guerry, André-Michel. 1829. "Tableau des variations météorologique comparées aux phénomènes physiologiques, d'aprés les observations faites à l'obervatoire royal, et les recherches statistique les plus récentes." *Annales d'Hygiène Publique et de Médecine Légale* 1: 228.

Nightingale, Florence. 1858. *Mortality of the British Army*. London: Harrison and Sons.

Wikimedia Foundation. 2014. "Spider Chart." *Wikipedia, the Free Encyclopedia*. Accessed January 31, 2014. **http://en.wikipedia.org/wiki/File:Spider_Chart.jpg**.

Image Credits

See Wikimedia Foundation 2014.

Correlations

Scatter Plot

References

Campbell, Rob. 2010. "Rug Plots." MATLAB Central. Accessed January 17, 2014. **http://www.mathworks.com/matlabcentral/fileexchange/27582-rug-plots**.

Friendly, Michael. 2008. "The Golden Age of Statistical Graphics." *Statistical Science* 23 (4): 502–535.

Herschel, John F. W. 1833. "On the Investigation of the Orbits of Revolving Double Stars." *Memoirs of the Royal Astronomical Society* 5: 171–222.

Image Credits

Image from Campbell 2010.

Distributions

References

Harris, Robert L. 1999. *Information Graphics: A Comprehensive Illustrated Reference*. New York: Oxford University Press. 124.

Image Credits

Image adapted from Harris 1999 © Oxford University Press, 1999.

Data Credits

MEDLINE data can be linked at **http://cnets.indiana.edu/groups/nan/webtraffic/websci14-data**. Accessed September 18, 2014.

Dot Graph

References

Harris, Robert L. 1999. *Information Graphics: A Comprehensive Illustrated Reference*. New York: Oxford University Press.

Image Credits

Image adapted from Harris 1999 © Oxford University Press, 1999.

Stripe Graph

References

Harris, Robert L. 1999. *Information Graphics: A Comprehensive Illustrated Reference*. New York: Oxford University Press.

Image Credits

Image adapted from Harris 1999 © Oxford University Press, 1999.

Contributors

Andrea Scharnhorst uses a stripe graph to visualize author publications on different topics.

Stem and Leaf Graph

References

Harris, Robert L. 1999. *Information Graphics: A Comprehensive Illustrated Reference*. New York: Oxford University Press. 370.

Image Credits

Image adapted from Harris 1999 © Oxford University Press, 1999.

Histogram

References

Harris, Robert L. 1999. *Information Graphics: A Comprehensive Illustrated Reference*. New York: Oxford University Press.

Image Credits

Image adapted from Harris 1999 © Oxford University Press, 1999.

48 Temporal Studies—"When"

References

BookRags Media Network. 2014. "Henry David Thoreau Quotes." Accessed January 7, 2014. **http://www.brainyquote.com/quotes/authors/h/henry_david_thoreau.html**.

Monroe, Megan., Rongjian Lan, Juan Morales del Olmo, Ben Shneiderman, Catherine Plaisant, and Jeff Millstein. 2013. "The Challenges of Specifying Intervals and Absences in Temporal Queries: A Graphical Language Approach." In *Proceedings of the SIGCHI Conference on Human Factors in Computing*, 2349–2358. New York: ACM Press.

Data Preprocessing

Time Zones

References

Dunn, Steve. 2010. "Dealing with Timezones in a Global Environment." Razorleaf. Accessed January 7, 2014. **http://razorleaf.com/2010/04/timezones-in-global-plm**.

Wikimedia Foundation. 2014. "Time Zones 2008." *Wikipedia, the Free Encyclopedia*. Accessed September 21, 2014. **http://en.wikipedia.org/wiki/File:Timezones2008.png**.

Image Credits

See Wikimedia Foundation 2014.

Time Slicing

References

Börner, Katy. 2010. *Atlas of Science: Visualizing What We Know*. Cambridge, MA: The MIT Press.

Börner, Katy, and David E. Polley. 2014. *Visual Insights: A Practical Guide to Making Sense of Data*. Cambridge, MA: The MIT Press.

Herr II, Bruce W., Russell Jackson Duhon, Elisha F. Hardy, Shashikant Penumarthy, and Katy Börner. 2007. *113 Years of* Physical Review. Bloomington, IN. Courtesy of Indiana University. In "3rd Iteration (2007): The Power of Forecasts," *Places & Spaces: Mapping Science*, edited by Katy Börner and Julie M. Davis. **http://scimaps.org**.

Ke, Weimao, Lalitha Visvanath, and Katy Börner. 2004. "Mapping the Evolution of Co-Authorship Networks." Accessed January 9, 2014. **http://scimaps.org/maps/map/mapping_the_evolutio_81**.

Wattenberg, Martin, and Fernanda B. Viégas. 2006. *History Flow Visualization of the Wikipedia Entry "Abortion."* Cambridge, Massachusetts. Courtesy of Martin Wattenberg, Fernanda B. Viégas, and IBM Research. In "2nd Iteration (2006): The Power of Reference Systems," *Places & Spaces: Mapping Science*, edited by Katy Börner and Deborah MacPherson. **http://scimaps.org**.

Wikimedia Foundation. 2014. "Seven-Day Week." *Wikipedia, the Free Encyclopedia*. Accessed January 31, 2014. **http://en.wikipedia.org/wiki/Seven-day_week**.

Image Credits

See Börner and Polley 2014.

Trends

References

Australian Bureau of Statistics. 2008. "Time Series Analysis: Seasonal Adjustment Methods." Accessed January 7, 2014. **http://www.abs.gov.au/websitedbs/d3310114.nsf/51c9a3d36edfd0dfca256acb00118404/c890aa8e65957397ca256ce10018c9d8!opendocument**.

Harris, Robert L. 1999. *Information Graphics: A Comprehensive Illustrated Reference*. New York: Oxford University Press.

Image Credits

Adapted from Harris 1999 © Oxford University Press, 1999.

Bursts

References

Börner, Katy, and David E. Polley. 2014. *Visual Insights: A Practical Guide to Making Sense of Data*. Cambridge, MA: The MIT Press.

Cyberinfrastructure for Network Science Center. 2010. Scholarly Database. Accessed March 1, 2014. **http://sdb.cns.iu.edu**.

Kleinberg, Jon M. 2002. "Bursty and Hierarchical Structure in Streams." In *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 91–101. New York: ACM Press.

Kleinberg, Jon M. 2002. "Sample Results from a Burst Detection Algorithm." Accessed Jan 9, 2014. **http://www.cs.cornell.edu/home/kleinber/kdd02.html**.

United States National Library of Medicine. 2014. PubMed. Accessed March 1, 2014. **http://www.ncbi.nlm.nih.gov/pubmed**.

Data Credits

See Cyberinfrastructure for Network Science Center 2010.

See United States National Library of Medicine 2014.

Contributors

Robert P. Light generated the burst graph; design by Perla Mateo-Lujan.

New York City's Weather for 1980

References

The New York Times Staff. 1981. "New York City's Weather for 1980." *The New York Times*, January 11. Accessed January 9, 2014. **http://www.datavis.ca/gallery/images/NYweather.jpg**.

Image Credits

From *The New York Times* © 1980. *The New York Times*. All rights reserved. Used by permission and protected by the Copyright Laws of the United States. The printing, copying, redistribution, or retransmission of this Content without express written permission is prohibited.

The Beatles: Working Schedule, 1963–1966

References

Deal, Michael. 2011. "The Beatles: Working Schedule 1963–1966." Personal Home Page. Accessed March 7, 2014. **http://www.mikemake.com/Charting-the-Beatles**.

In Investing, It's When You Start and When You Finish

References

The New York Times Staff. 2011. "In Investing, It's When You Start and When You Finish." *The New York Times*, January 2. Accessed January 31, 2014. **http://www.nytimes.com/interactive/2011/01/02/business/20110102-metrics-graphic.html?_r=0**.

Yau, Nathan. 2011. "In Investing, Timing is Everything." *FlowingData* (blog), January 13. Accessed January 31, 2014. **http://flowingdata.com/2011/01/13/in-investing-timing-is-everything**.

Image Credits

From *The New York Times*, January 2, 2011. © 2011 *The New York Times*. All rights reserved. Used by permission and protected by the Copyright Laws of the United States. The printing, copying, redistribution, or retransmission of this Content without express written permission is prohibited.

Sankey Graph of Google Analytics Data

References

Yau, Nathan. 2012. "How to Make a Sankey Diagram to Show Flow." *FlowingData* (blog), April 26. Accessed January 31, 2014. **http://flowingdata.com/2012/04/26/how-to-make-a-sankey-diagram-to-show-flow**.

Software Credits

© 2012 Google Inc. All rights reserved. Google and Google Analytics are registered trademarks of Google Inc., used with permission.

Contributors

Michael P. Ginda compiled the data and rendered the visualization.

50 Temporal Visualization Types

Trends and Distributions

References

Börner, Katy. 2010. *Atlas of Science: Visualizing What We Know*. Cambridge, MA: The MIT Press.

Felton, Nicholas. 2010. "The 2010 Feltron Annual Report." *Feltron* (blog). Accessed January 9, 2014. **http://feltron.com/ar10_10.html**.

Harris, Robert L. 1999. *Information Graphics: A Comprehensive Illustrated Reference*. New York: Oxford University Press. 417.

Nissen, Mayo. 2009. "Visualising Household Power Consumption." MOMA: Talk to Me. Accessed March 1, 2014. **http://www.moma.org/interactives/exhibitions/2011/talktome/objects/145529**.

Wikimedia Foundation. 2014. "U.S. First Class Postage Rate." *Wikipedia, the Free Encyclopedia*. September 21, 2014. **http://en.wikipedia.org/wiki/History_of_United_States_postage_rates#mediaviewer/File:US_Postage_History.svg**.

Image Credits

U.S. First Class Postage Rate, *see* Wikimedia Foundation 2014.

Images adapted from Harris 1999 © Oxford University Press, 1999.

See Nissen 2009.

Comparison

References

Frumin, Michael. 2009. "Spark It Up." *Frumination* (blog), May 7. Accessed January 31, 2014. **http://frumin.net/ation/2009/05/spark_it_up.html**.

Koch, Tom. 2005. *Cartographies of Disease: Maps, Mapping, and Medicine*. Redlands, CA: ESRI Press.

Derivatives

References

Baly, William. 1854. *Reports on Epidemic Cholera. Drawn Up at the Desire of the Cholera Committee of the Royal College of Physicians*. London: J. Churchill.

Börner, Katy. 2010. *Atlas of Science: Visualizing What We Know*. Cambridge, MA: The MIT Press.

The Weather Channel. 2013. "Hurricane Gustav." Accessed January 31, 2014. **http://www.weather.com/newscenter/hurricanecentral/2008/gustav.html**.

Wolfram. 2014. "Visualize Wind Velocity Data." Accessed January 31, 2014. **http://www.wolfram.com/products/mathematica/newin7/content/VectorAndFieldVisualization/VisualizeWindVelocityData.html**.

Image Credits

See Baly 1854.

Vector and Field Visualizations © 2014 Wolfram Alpha LLC. **http://www.wolframalpha.com**. Accessed September 18, 2014.

Flows over Time and Space

Flow Map

References

Börner, Katy. 2010. *Atlas of Science: Visualizing What We Know*. Cambridge, MA: The MIT Press.

Harris, Robert L. 1999. *Information Graphics: A Comprehensive Illustrated Reference*. New York: Oxford University Press.

Minard, Charles Joseph. 1866. *Europe Raw Cotton Imports in 1858, 1864 and 1865*. Paris, France. Courtesy of the Library of Congress, Geography and Maps Division. In "4th Iteration (2008): Science Maps for Economic Decision Makers," *Places & Spaces: Mapping Science*, edited by Katy Börner and Elisha F. Hardy. **http://scimaps.org**.

Minard, Charles Joseph. 1869. *Napoleon's March to Moscow*. Paris, France. Courtesy of Edward Tufte, Graphics Press, Cheshire, CT. In "1st Iteration (2005): The Power of Maps," *Places & Spaces: Mapping Science*, edited by Katy Börner and Deborah MacPherson. **http://scimaps.org**.

Ministère des Travaux Publics. 1879–1897. *Album de Statistique Graphique*. Paris: Imprimerie Nationale.

Wikimedia Foundation. 2014. "Flow Map." *Wikipedia, the Free Encyclopedia*. Accessed January 7, 2014. **http://en.wikipedia.org/wiki/Flow_map**.

Image Credits

Images from Harris 1999 © Oxford University Press, 1999.

Space-Time-Cube Map

References

Carlstein, Tommy, Don Parkes, and Nigel J. Thrift. 1978. *Human Activity and Time Geography*. New York: John Wiley & Sons.

Leibniz-Institut für Länderkunde. 2011. "Lebensläufe der Heidelberger Nobelpreisträger für Physik, Chemie und Medizin." Accessed January 9, 2014. **http://www.uni-heidelberg.de/md/zentral/universitaet/geschichte/nobel_phys_chem_med.pdf**.

Neumann, Andreas. 2005. "Thematic Navigation in Space and Time." Paper presented at the SVG Open 2005 Conference. Accessed January 9, 2014. **http://www.svgopen.org/2005/papers/abstract_neumann_thematic_navigation_in_space_and_time**.

Image Credits

Space Time Cube Map created by Torsten Hägerstrand from Carlstein, Parkes, Thrift 1978.

52 Geospatial Studies—"Where"

References

Börner, Katy. 2010. *Atlas of Science: Visualizing What We Know*. Cambridge, MA: The MIT Press.

Colizza, Vittoria, Alessandro Vespignani, and Elisha F. Hardy. 2007. *Impact of Air Travel on Global Spread of Infectious Diseases*. Bloomington, IN. Courtesy of Indiana University. In "3rd Iteration (2007): The Power of Forecasts," *Places & Spaces: Mapping Science*, edited by Katy Börner and Julie M. Davis. **http://scimaps.org**.

Tobler, Waldo R. 1970. "A Computer Movie Simulating Urban Growth in the Detroit Region." *Economic Geography* 46 (2): 234–240.

Data Preprocessing

References

iTouchMap. 2014. "Latitude and Longitude of a Point." Accessed January 9, 2014. **http://itouchmap.com/latlong.html**.

Clustering

References

Börner, Katy. 2010. *Atlas of Science: Visualizing What We Know*. Cambridge, MA: The MIT Press.

Skupin, André. 2004. *In Terms of Geography*. New Orleans, Louisiana, 2005. Courtesy of André Skupin, San Diego State University, San Diego, CA. In "1st Iteration (2005): The Power of Maps," *Places & Spaces: Mapping Science*, edited by Katy Börner and Deborah MacPherson. **http://scimaps.org**.

Zahn, Charles T. 1971. "Graph-Theoretical Methods for Detecting and Describing Gestalt Clusters." *IEEE Transactions on Computers*, 20 (1): 68–86.

Image Credits

Images from Zahn 1971. © 1971 IEEE. Reprinted, with permission, from *IEEE Transactions on Computers*.

Using Geometric Grids

References

Börner, Katy. 2010. *Atlas of Science: Visualizing What We Know*. Cambridge, MA: The MIT Press.

Skupin, André. 2004. "The World of Geography: Visualizing a Knowledge Domain with Cartographic Means." *PNAS* 101 (Suppl. 1): 5274–5278.

Visual Generalization

References

Strebe, Daniel. 2012. "The Impotence of Maps: Or, Deconstructing the Deconstruction of Their Construction." Presentation given at NACIS 2012. Accessed January 9, 2014. **http://www.mapthematics.com/Downloads/Impotence%20of%20Maps.pdf**.

Töpfer, Friedrich. 1962. "Das Wurzelgesetz und seine Anwendung bei der Reliefgeneralisierung." *Vermessungstechnik* 10 (2): 37–42.

Töpfer, Friedrich. 1974. *Kartographische Generalisierung*. Gotha/Leipzig: VEB Herrmann Haack/Geographisch-Kartographische Anstalt.

Töpfer, Friedrich, and Wolfgang Pillewizer. 1966. "The Principles of Selection." *The Cartographic Journal* 3: 10–16.

Tobler, Waldo A. 1970. "A Computer Movie Simulating Urban Growth in the Detroit Region." *Economic Geography* 46 (2): 234–240.

Tobler, Waldo A. 1973. "A Continuous Transformation Useful for Districting." *Annals of the New York Academy of Sciences*, 219: 215–220.

Image Credits

Visual Generalization Map Examples © 1996 Esri. All rights reserved. Published in the United States of America.

Etsy Sales Map

References

Bragina, Lana (ulaniulani). 2009. "All My Etsy Sales." Flickr. Accessed January 31, 2014. **http://www.flickr.com/photos/madame_ulani/3569828622/sizes/l**.

Bragina, Lana (Uloni). 2009. "Infographics: One Seller's Visualization." *The Etsy Blog*, June 18. Accessed January 9, 2014. **http://www.etsy.com/blog/en/2009/infographics-one-sellers-visualization**.

Bragina, Lana (ulaniulani). 2009. "3 Month Crocheting." Flickr. Accessed January 31, 2014. **http://www.flickr.com/photos/madame_ulani/3455161546**.

Image Credits

Etsy Sales Map courtesy of Lana Bagina, **http://www.ulani.de**. Accessed September 18, 2014.

Skitter Internet Map

References

Lima, Manuel. 2002. "AS Internet Graph." Visual Complexity. Accessed January 31, 2014. **http://www.visualcomplexity.com/vc/project.cfm?id=20**.

CAIDA. 2013. "IPv4 and IPv6 AS Core: Visualizing IPv4 and IPv6 Internet Topology at a Macroscopic Scale in 2013." Accessed January 31, 2014. **http://www.caida.org/research/topology/as_core_network**.

Image Credits

Skitter Internet Map © 2000 The Regents of the University of California. All Rights Reserved.

The Debt Quake in the Eurozone

Image Credits

The Debt Quake in the Eurozone courtesy of Morgan Quinn—Intuit Inc., **http://mint.com**. Accessed September 18, 2014.

In the Shadow of Foreclosures

References

Fairfield, Hannah. 2008. "In the Shadow of Foreclosures." *The New York Times*, April 5. Accessed February 21, 2014. **http://www.nytimes.com/imagepages/2008/04/05/business/20080406_METRICS.html.**

Image Credits

From *The New York Times*. © 2008 *The New York Times*. All rights reserved. Used by permission and protected by the Copyright Laws of the United States. The printing, copying, redistribution, or retransmission of this Content without express written permission is prohibited.

54 Geospatial Visualization Types

Contributors

André Skupin provided expert comments.

Discrete Space

Dot Density Map

References

McCune, Doug. 2011. "Ethics and the Use of DUI Data." *Doug McCune* (blog), March 23. Accessed February 16, 2014. **http://dougmccune.com/blog/tag/datasf**.

Snow, John. 1855. *On the Mode of Communication of Cholera*. 2nd ed. London: John Churchill.

Image Credits

See McCune 2011.

Proportional Symbol Map

References

Goldsmith, Andrew. 2014. "World Map." Accessed January 10, 2014. **http://www.flickr.com/photos/andrewgoldsmith/5076593587/in/photostream**.

Starr, Benjamin. 2011. "A Typographical Map of the World." Visual News. Accessed January 10, 2014. **http://www.visualnews.com/category/design/maps-infographics/page/10**.

Image Credits

See Goldsmith 2014.

Choropleth Map

References

White, Adrian and the National Geographic *EarthPulse* Team. 2008. *A Global Projection of Subjective Well-Being*. Washington, DC. Courtesy of National Geographic. In "4th Iteration (2008): Science Maps for Economic Decision Makers," *Places & Spaces: Mapping Science,* edited by Katy Börner and Elisha F. Hardy. **http://scimaps.org**.

The World Bank and The National Geographic Society. 2006. *The Millennium Development Goals Map: A Global Agenda to End Poverty*. Washington, DC. Courtesy of The World Bank and The National Geographic Society. In "5th Iteration (2009): Science Maps for Science Policy Makers," *Places & Spaces: Mapping Science,* edited by Katy Börner and Elisha F. Hardy. **http://scimaps.org**.

Yau, Nathan. 2009. "How to Make a US County Thematic Map Using Free Tools." *FlowingData* (blog), November 12. Accessed January 31, 2014. **http://flowingdata.com/2009/11/12/how-to-make-a-us-county-thematic-map-using-free-tools**.

Image Credits

Image created by Michael P. Ginda using code provided by Nathan Yau of *FlowingData*. Code is from the Bureau of Labor and Statistics, map from Wikipedia (public domain), and the colors were generated using ColorBrewer.

Dasymetric Map

References

USDA Agricultural Research Service. 2012. "USDA Plant Hardiness Zone Map." Accessed March 1, 2014. **http://planthardiness.ars.usda.gov/PHZMWeb**.

Image Credits

Image created by the Prism Climate Group at Oregon State; USDA.

Cartogram Map

References

Heer, Jeffrey, Michael Bostock, and Vadim Ogievetsky. 2010. "A Tour Through the Visualization Zoo." *Communications of the ACM* 53 (6): 59–67.

Schlarmann, James. 2012. "Candidate Obama May Turn November into a Landslide Victory." Political Garbage Chute. Accessed February 16, 2014. **http://www.politicalgarbagechute.com/candidateobama**.

Image Credits

2012 Electoral Map Cartogram © 2012, Frontloading HQ.

Dorling Cartogram Example created by Jeffrey Heer, Michael Bostock, and Vadim Ogievetsyky.

2012 Political Election Map and Cartogram © 2012 M. E. J. Newman.

Continuous Space

Elevation Map

References

DataSF. 2014. "Terms of Use." Accessed March 1, 2014. **http://www.datasf.org/page.php?page=tou&return_url=/datafiles/index.php?dir=Police&by=name&order=asc**.

McCune, Doug. 2011. "Ethics and the Use of DUI Data." *Doug McCune* (blog), March 23. Accessed February 16, 2014. **http://dougmccune.com/blog/tag/datasf**.

Image Credits

Map from McCune 2011.

Data Credits

See Data SF 2014.

Isarithmic Map

References

McCune, Doug. 2011. "Ethics and the Use of DUI Data." *Doug McCune* (blog), March 23. Accessed February 16, 2014. **http://dougmccune.com/blog/tag/datasf**.

Image Credits

Map from McCune 2011.

Isochrone Map

References

Friendly, Michael. 2007. "The Golden Age of Statistical Maps & Diagrams." DataVis.ca. Accessed January 10, 2014. **http://www.datavis.ca/papers/maps/GoldenAge2x2.pdf**.

Ministère des Travaux Publics. 1879–1897. *Album de Statistique Graphique*. Paris: Imprimerie Nationale.

Scheidel, Walter, and Elijah Meeks. 2014. "Travel Time to Rome in July." ORBIS: The Stanford Geospatial Network Model of the Roman World. Accessed March 1, 2014. **http://orbis.stanford.edu/images/gallery/ttr_800.png**.

Image Credits

See Scheidel and Meeks 2014.

Data Credits

ORBIS: The Stanford Geospatial Network Model of the Roman World reconstructs the time, cost, and financial expense associated with a wide range of different types of travel in antiquity around 200 AD. **http://orbis.stanford.edu/#introducing**. Accessed September 18, 2014.

Vector Fields

References

Börner, Katy. 2010. *Atlas of Science: Visualizing What We Know*. Cambridge, MA: The MIT Press.

Tobler, Waldo R. 1979. "Smooth Pycnophylactic Interpolation for Geographical Regions." *American Statistical Association* 74 (367): 519–536.

Tobler, Waldo R. 1981. "A Model of Geographic Movement." *Geographical Analysis* 13 (1): 1–20.

Tobler, Waldo R. 1983. "Push Pull Migration Laws." *Annals of the Association of American Geographers* 73 (1): 1–17.

Tobler, Waldo R. 1987. "Experiments in Migration Mapping by Computer." *The American Cartographer* 14 (2): 155–163.

Tobler, Waldo R. 1995. "Migration: Ravenstein, Thornthwaite, and Beyond." *Urban Geography* 16 (4): 327–343.

Line Map

References

Corbett, John. 2014. "Ernest George Ravenstein: The Laws of Migration, 1885. Center for Spatially Integrated Social Science. Accessed January 10, 2014. **http://www.csiss.org/classics/content/90**.

Ravenstein, Ernest George. 1885. "The Laws of Migration." *Journal of the Statistical Society of London* 48 (2): 167–235.

Subway Map

References

Börner, Katy. 2010. *Atlas of Science: Visualizing What We Know*. Cambridge, MA: The MIT Press.

Nesbitt, Keith V. 2004. *PhD Thesis Map*. Newcastle, Australia. Courtesy of IEEE and Keith V. Nesbitt, Charles Sturt University, Australia, © 2004 by IEEE. In "1st Iteration (2005): The Power of Maps," *Places & Spaces: Mapping Science*, edited by Katy Börner and Deborah MacPherson. **http://scimaps.org**.

Flow Map

References

Abel, Guy. 2014. "Circular Migration Flow Plots in R." Personal Home Page. Accessed September 14, 2014. **http://gjabel.wordpress.com/2014/03/28/circular-migration-flow-plots-in-r**.

Abel, Guy J., and Nikola Sander. 2014. "Quantifying Global International Migration Flows." *Science* 343 (6178): 1520–1522.

Image Credits

Figure from Abel 2014.

Software Credits

See Abel and Sander 2014.

Strip Map

References

Agrawala, Maneesh, and Christ Stolte. 2001. "Rendering Effective Route Maps: Improving Usability through Generalization." In *Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques*, 241–249. New York: ACM.

Börner, Katy. 2010. *Atlas of Science: Visualizing What We Know*. Cambridge, MA: The MIT Press.

Harris, Robert L. 1999. *Information Graphics: A Comprehensive Illustrated Reference*. New York: Oxford University Press.

Microsoft Corporation. 2014. "MapPoint." Accessed January 10, 2014. **http://mappoint.msn.com**.

Image Credits

See Agrawala and Stolte © 2001 Association for Computing Machinery, Inc. Reprinted by permission.

56 Topical Studies—"What"

References

Bloomer, Martin, Phil Hodkinson, and Stephen Billett. 2004. "The Significance of Ontogeny and Habitus in Constructing Theories of Learning." *Studies in Continuing Education* 26 (1): 19–43.

Porter, Alan L., and Scott W. Cunningham. 2005. *Tech Mining: Exploring New Technologies for Competitive Advantage*. Hoboken, NJ: John Wiley & Sons.

Sinclair, John. 1991. *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Data Preprocessing

Stemming and Stopword Removal

References

Porter, Martin. 2006. "The Porter Stemming Algorithm." Accessed January 14, 2014. **http://tartarus.org/~martin/PorterStemmer**.

Software Credits

See Porter 2006.

Tokenization

References

Michel, Jean-Baptiste, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K. Gray, The Google Books Team, Joseph P. Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, Steven Pinker, Martin A. Nowak, Erez Lieberman Aiden. 2010. "Quantitative Analysis of Culture Using Millions of Digitized Books." *Science* 331 (6014): 176–182.

Image Credits

Image courtesy of Google Books Ngram Viewer, **http://books.google.com/ngrams**. Accessed September 18, 2014.

Descriptive Term Identification

References

Blei, David M., Andrew Y. Ng, and Michael I. Jordan. 2003. "Latent Dirichlet Allocation." *Journal of Machine Learning Research* 3 (4–5): 993–1022.

Deerwester, Scott, Susan T. Dumais, George Furnas, Thomas K. Landauer, and Richard Harshman. 1990. "Indexing by Latent Semantic Analysis." *Journal of the American Society for Information Science* 41 (6): 391-407.

Landauer, Thomas K., Peter W. Foltz, and Darrell Laham, 1998. "Introduction to Latent Semantic Analysis." *Discourse Processes* 25 (2-3): 259-284.

Salton, Gerald, and C. S. Yang. 1973. "On the Specification of Term Values in Automatic Indexing." *Journal of Documentation* 29 (4): 351–372.

Distributions

Term Frequency and Distributions

References

Levy, Amun. 2008. "On Words." *Good* 9 (March/April). Accessed January 14, 2014. **http://awesome.good.is/transparency/009/trans009onwords.html**.

On Words Concordance

References

Coulter, Ann. 2002. *Slander: Liberal Lies About the American Right*. New York: Crown Publishers.

Franken, Al. 2003. *Lies: And the Lying Liars Who Tell Them*. New York: Dutton.

Levy, Amun. 2008. "On Words." *Good* 9 (March/April). Accessed January 14, 2014. **http://awesome.good.is/transparency/009/trans009onwords.html**.

Image Credits

On Words Concordance originally published in *Good Magazine*.

Is Facebook-Is Twitter Phrase Graph

References

Viégas, Fernanda, and Martin Wattenberg. 2009. "Web Seer." Home Page. **http://hint.fm/projects/seer**.

Image Credits

Image from Viégas and Wattenberg 2009.

Software Credits

http://hint.fm/seer. Accessed September 18, 2014.

Sentiment Analysis of the Bible

References

OpenBible. 2014. Home Page. Accessed January 31, 2014. **http://www.openbible.info**.

Smith, Stephen. 2011. "Applying Sentiment Analysis to the Bible." OpenBible. Accessed January 31, 2014. **http://www.openbible.info/blog/2011/10/applying-sentiment-analysis-to-the-bible**.

Smith, Stephen. 2011. "Sentiment Analysis of The Bible." OpenBible. Accessed January 31, 2014. **http://a.openbible.info/blog/2011-10-sentiment-full.png**.

Image Credits

Sentiment Analysis of the Bible courtesy of **http://www.openbible.info**. Accessed September 18, 2014.

Data Credits

Raw data available at **http://a.openbible.info/blog/2011-10-sentiment-data.zip**. Accessed September 18, 2014.

Software Credits

Viralheat Sentiment API: **https://app.viralheat.com/developer/sentiment_api**. Accessed September 18, 2014.

Editions of Darwin's *On the Origin of Species*

References

Darwin, Charles. 1859. *On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life*. London: John Murray.

Fry, Ben. 2009. "On the Origin of Species: The Preservation of Favoured Traces." Personal Website. Accessed January 14, 2014. **http://benfry.com/traces**.

Fry, Ben. 2009. "Watching the Evolution of the 'Origin of the Species'." Personal Website. Accessed January 14, 2014. **http://benfry.com/writing/archives/529**.

Reas, Casey, and Ben Fry. 2007. *Processing: A Programming Handbook for Visual Designers and Artists*. Cambridge, MA: The MIT Press.

van Wyhe, John, ed. 2002. *The Complete Work of Charles Darwin Online*. Accessed January 14, 2014. **http://darwin-online.org.uk**.

Image Credits

Editions of Darwin's *The Origin of the Species* © 2009 Ben Fry.

Data Credits

See van Wyhe 2002.

Software Credits

Built with Processing; *see* Reas and Fry 2007.

58 Topical Visualization Types

Composition and Frequency

Lists

References

Börner, Katy. 2010. *Atlas of Science: Visualizing What We Know*. Cambridge, MA: The MIT Press.

Börner, Katy, Elisha F. Hardy, Bruce W. Herr II, Todd M. Holloway, and W. Bradford Paley. 2006. T*axonomy Visualization of Patent Dat*a. Bloomington, Indiana and New York, New York. Courtesy of Indiana University and W. Bradford Paley. In "2nd Iteration (2006): The Power of Reference Systems," *Places & Spaces: Mapping Science*, edited by Katy Börner and Deborah MacPherson. **http://scimaps.org**.

British National Corpus. 2014. Home Page. Accessed March 1, 2014. **http://www.natcorp.ox.ac.uk**.

Eick, Stephen G., Joseph L. Steffen, and Eric E. Sumner Jr. 1992. "Seesoft—A Tool for Visualizing Line Oriented Software Statistics." *IEEE Transactions on Software Engineering* 18 (11): 957–968.

Fry, Ben. 2003. "Revisionist: Visualizing the Evolution of Software Projects." Personal Website. Accessed January 14, 2014. **http://benfry.com/revisionist**.

Harris, Jonathan. 2003. WordCount. Accessed January 14, 2014. **http://www.wordcount.org**.

Data Credits

See British National Corpus 2014.

WordCount underlying word frequency list at **http://www.number27.org/assets/misc/words.txt**. Accessed September 18, 2014.

Tag Cloud

References

Börner, Katy. 2017. *Atlas of Forecasts: Predicting and Broadcasting Science*. Cambridge, MA: The MIT Press.

Castellani, Brian. 2013. *Map of Complexity Science*. Cleveland, OH. Courtesy of Arts and Science Factory, LLC. In "9th Iteration (2013): Science Maps Showing Trends and Dynamics," *Places & Spaces: Mapping Science*, edited by Katy Börner and Todd N. Theriault. **http://scimaps.org**.

Steinbock, Daniel. 2014. TagCrowd. Accessed January 14, 2014. **http://tagcrowd.com**.

Tagul. 2014. Home Page. Accessed January 14, 2014. **http://tagul.com**.

Wordle. 2014. Home Page. Accessed January 14, 2014. **http://www.wordle.net**.

Image Credits

Wordcount © 2003 Jonathan Harris.

Tag cloud example courtesy of TagCrowd. **http://tagcrowd.com**. Accessed September 18, 2014.

Data Credits

Wordcount rendered using British National Corpus, **http://www.natcorp.ox.ac.uk**. © 2010 University of Oxford. Accessed September 18, 2014.

Software Credits

Visualization of the opening paragraph computed using TagCrowd; *see* Steinbock 2014.

Structure

Circular Graph

References

Börner, Katy. 2010. *Atlas of Science: Visualizing What We Know*. Cambridge, MA: The MIT Press.

Börner, Katy, Richard Klavans, Michael Patek, Angela Zoss, Joseph R. Biberstine, Robert Light, Vincent Lariviére, and Kevin W. Boyack. 2012. "Design and Update of a Classification System: The UCSD Map of Science." *PLoS One* 7 (7): e39464. Accessed October 31, 2013. **http://sci.cns.iu.edu/ucsdmap**.

Boyack, Kevin W., and Richard Klavans. 2008. *The Scientific Roots of Technology*. Albuquerque, NM and Berwyn, PA. Courtesy of SciTech Strategies, Inc. In "4th Iteration (2008): Science Maps for Economic Decision Makers," *Places & Spaces: Mapping Science,* edited by Katy Börner and Elisha F. Hardy. **http://scimaps.org**.

Boyack, Kevin W. and Richard Klavans. 2008. *U.S. Vulnerabilities in Science*. Berwyn, PA and Albuquerque, NM. Courtesy of SciTech Strategies. In "5th Iteration (2009): Science Maps for Science Policy Makers," *Places & Spaces: Mapping Science,* edited by Katy Börner and Elisha F. Hardy. **http://scimaps.org**.

Paley, W. Bradford. 2006. *TextArc Visualization of* The History of Science. New York, NY. Courtesy of W. Bradford Paley. In "2nd Iteration (2006): The Power of Reference Systems," *Places & Spaces: Mapping Science*, edited by Katy Börner and Deborah MacPherson. **http://scimaps.org**.

Image Credits

Designed by Perla Mateo-Lujan, inspired by the UCSD Map of Science, Richard Klavans and Kevin W. Boyack, SciTech Strategies, Inc. **http://www.mapofscience.com**. Accessed September 18, 2014.

GRIDL

References

Shneiderman, Ben, David Feldman, Anne Rose, Xavier Ferré. 2000. "Visualizing Digital Library Search Results with Categorical and Hierarchical Axes." In *Proceedings of the Fifth ACM Conference on Digital Libraries*, 57–66. New York: ACM.

Crossmap

References

Börner, Katy. 2010. *Atlas of Science: Visualizing What We Know*. Cambridge, MA: The MIT Press.

Morris, Steven A. 2005. *Visualizing 60 Years of Anthrax Research*. Stillwater, Oklahoma, 2005. Courtesy of Steven A. Morris, Oklahoma State University, Stillwater. In "1st Iteration (2005): The Power of Maps," *Places & Spaces: Mapping Science*, edited by Katy Börner and Deborah MacPherson. **http://scimaps.org**.

Morris, Steven, Camille DeYong, Zheng Wu, Sinan Salman, Dagmawi Yemenu. 2002. "DIVA: A Visualization System for Exploring Document Databases for Technology Forecasting." *Computers and Industrial Engineering* 43 (4): 841–862.

Image Credits

Reprinted from Morris, Steven, Camille DeYong, Zheng Wu, Sinan Salman, Dagmawi Yemenu. 2002. "DIVA: A Visualization System for Exploring Document Databases for Technology Forecasting." *Computers and Industrial Engineering* 43 (4): 841–862. Copyright 2002 National Academy of Sciences, U.S.A.

Isoline Map

References

Börner, Katy. 2010. *Atlas of Science: Visualizing What We Know*. Cambridge, MA: The MIT Press.

Quaggiotto, Marco. 2008. *Knowledge Cartography.* Milano, Italy. Courtesy of the Department of Industrial Design, Art, Communication and Fashion (INDACO), Politecnico di Milano, Italy, and Complex Networks and Systems Group, ISI Foundation, Turin, Italy. In "6th Iteration (2009): Science Maps for Scholars," *Places & Spaces: Mapping Science,* edited by Katy Börner and Elisha F. Hardy. **http://scimaps.org**.

Skupin, André, Mike Conway, Wei Wei, Brian Chapman, and Wendy Chapman. 2014. "In Terms of MIMIC."

Image Credits

Courtesy of André Skupin.

Data Credits

MIMIC-II; *see* Saeed et al. 2011.

Self-Organizing Map

References

Börner, Katy. 2010. *Atlas of Science: Visualizing What We Know.* Cambridge, MA: The MIT Press.

Honkela, Timo. 2005. "Von Foerster meets Kohonen: Approaches to Artificial Intelligence, Cognitive Science and Information Systems Development." *Kybernetes* 34 (1/2): 40-53.

Saeed, Mohammed. Mauricio Villarroel, Andrew T. Reisner, Gari Clifford, Li-Wei Lehman, George Moody, Thomas Heldt, Tin H. Kyaw, Benjamin Moody, and Roger G. Mark. 2011. "Multiparameter Intelligent Monitoring in Intensive Care II (MIMIC-II): A Public-Access Intensive Care Unit Database." *Critical Care Medicine* 39 (5): 952–960.

Skupin, André. 2004. *In Terms of Geography*. New Orleans, Louisiana, 2005. Courtesy of André Skupin, San Diego State University, San Diego, CA. In "1st Iteration (2005): The Power of Maps," *Places & Spaces: Mapping Science*, edited by Katy Börner and Deborah MacPherson. **http://scimaps.org**.

Image Credits

Image from Honkela 2005. © Emerald Group Publishing Limited all rights reserved.

Trends

History Flow

References

Börner, Katy. 2010. *Atlas of Science: Visualizing What We Know*. Cambridge, MA: The MIT Press.

Wattenberg, Martin, and Fernanda B. Viégas. 2006. *History Flow Visualization of the Wikipedia Entry "Abortion."* Courtesy of Martin Wattenberg, Fernanda B. Viégas, and IBM Research. In "2nd Iteration (2006): The Power of Reference Systems," *Places & Spaces: Mapping Science*, edited by Katy Börner and Deborah MacPherson. **http://scimaps.org**.

Alluvial Graph

References

Blondel, Vincent D., Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. 2008. "Fast Unfolding of Communities in Large Networks." *Journal of Statistical Mechanics* P10008.

MapEquation. 2014. Home Page. Accessed February 16, 2014. **http://www.mapequation.org**.

Rosvall, Martin, and Carl T. Bergstrom. 2008. "Maps of Random Walks on Complex Networks Reveal Community Structure." *PNAS* 105 (4): 1118–1123.

Yau, Nathan. 2012. "How to Make a Sankey Diagram to Show Flow." *FlowingData* (blog), April 26. Accessed January 31, 2014. **http://flowingdata.com/2012/04/26/how-to-make-a-sankey-diagram-to-show-flow**.

Image Credits

Image from Rosvall 2008 © 2008 National Academy of Sciences, U.S.A.

Software Credits

See MapEquation 2014.

Stream Graph

References

Byron, Lee, and Martin Wattenberg. 2008. "Stacked Graphs—Geometry and Aesthetics." *IEEE Transactions on Visualization and Computer Graphics* 14 (6): 1245–1252.

Cui, Weiwei, Shixia Liu, Li Tan, Conglei Shi, Yangqiu Song, Zekai Gao, Xin Tong, and Huamin Qu. 2011."TextFlow: Towards Better Understanding of Evolving Topics in Text." *IEEE Transactions on Visualization and Computer Graphics* 17 (12): 2412–2421.

Havre, Susan, Beth Hetzler, and Lucy Nowell. 2000. "ThemeRiver: Visualizing Theme Changes over Time." In *Proceedings of the IEEE Symposium on Information Visualization*, 115–123. Washington, DC: IEEE Computer Society.

Leskovec, Jure. 2014. Stanford Network Analysis Project. Accessed March 1, 2014. **http://memetracker.org.**

Leskovec, Jure, Lars Backstrom, and Jon Kleinberg. 2009. "Meme-tracking and the Dynamics of the News Cycle." In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 497–506. New York: ACM.

Image Credits

Image from Leskovec 2009. © 2009 Association for Computing Machinery, Inc. Reprinted by permission.

Image from Cui 2011. © 2011 IEEE. Reprinted, with permission, from *IEEE Transactions on Visualization and Computer Graphics*.

Image from Havre 2000. © 2000 IEEE. Reprinted, with permission, from *IEEE Symposium on Information Visualization*.

Image from Byron 2008. © 2008 IEEE. Reprinted, with permission, from *IEEE Transactions on Visualization and Computer Graphics*.

Software Credits

See Leskovec 2014.

Relationships

Arc Graph

References

Dittus, Martin. 2006. "IRC Arcs." Accessed January 15, 2010. **http://www.visualcomplexity.com/vc/project.cfm?id=403**.

Harrison, Chris, and Christoph Römhild. 2008. *Visualizing Bible Cross-References*. Pittsburgh, PA. Courtesy of Chris Harrison and Christoph Römhild. In "7th Iteration (2010): Science Maps as Visual Interfaces to Digital Libraries," *Places & Spaces: Mapping Science*, edited by Katy Börner and Michael J. Stamper. **http://scimaps.org**.

Wattenberg, Martin. 2002. "Arc Diagrams: Visualizing Structure in Strings." In *Proceedings of the IEEE Symposium on Information Visualization*, 110–116. Washington, DC: IEEE Computer Society.

Wattenberg, Martin. 2012. "The Shape of Song." Visual.ly. Accessed March 1, 2014. **http://visual.ly/shape-song**.

Image Credits

See Wattenberg 2012.

60 Network Studies—"With Whom"

References

Johnson, Samuel. 1755. *A Dictionary of the English Language*. London: W. Strahan.

Newman, Mark E. J. 2010. *Networks: An Introduction*. New York: Oxford University Press.

Newman, Mark E. J., Albert-László Barabási, and Duncan J. Watts. 2006. *The Structure and Dynamics of Networks*. Princeton, NJ: Princeton University Press.

Rainie, Lee, and Barry Wellman. 2012. *Networked: The New Social Operating System*. Cambridge, MA: The MIT Press.

Seeing Networks

References

Barabási, Albert-László. 2003. *Linked: How Everything Is Connected to Everything Else and What It Means for Business, Science, and Everyday Life*. New York: Plume.

Borgatti, Stephen P., Ajay Mehra, Daniel J. Brass, and Giuseppe Labianca. 2009. "Network Analysis in the Social Sciences." *Science* 323 (5916): 892–895.

Christakis, Nicholas A., and James H. Fowler. 2009. *Connected: The Surprising Power of Our Social Networks and How They Shape Our Lives*. New York: Little, Brown and Company.

Euler, Leonhard. 1741. "Solutio Problematis ad Geometriam Situs Pertinentis." In *Commentarii Academiae Scientiarum Petropolitanae*, 128–140. Accessed January 15, 2014. **http://www.math.dartmouth.edu/~euler/docs/originals/E053.pdf**.

Watts, Duncan J. 2003. *Six Degrees: The Science of a Connected Age*. New York: Norton.

Image Credits

Euler network graph adapted by Perla Mateo-Lujan from **http://physics.weber.edu/carroll/honors/konigsberg.htm**. Accessed September 18, 2014.

Network Extraction

References

Börner, Katy. 2010. *Atlas of Science: Visualizing What We Know*. Cambridge, MA: The MIT Press.

Börner, Katy, and David E. Polley. 2014. *Visual Insights: A Practical Guide to Making Sense of Data*. Cambridge, MA: The MIT Press.

Börner, Katy, Jeegar T. Maru, and Robert L. Goldstone. 2004. "The Simultaneous Evolution of Author and Paper Networks." *PNAS* 101 (Suppl. 1): 5266–73.

De Nooy, Wouter, Andrej Mrvar, and Vladimir Batageli. 2011. *Exploratory Social Network Analysis with Pajek*. New York: Cambridge University Press.

Network Types

Network Graph

References

Barabási, Albert-László. 2002. *Linked: The New Science of Networks*. Cambridge, MA: Perseus.

Kochen, Manfred, ed. 1989. *The Small World*. Norwood, NJ: Ablex.

Lee, Muh-Tian. 2001. "IMAGE." Accessed January 16, 2014. **http://www.learner.org/courses/mathilluminated/units/11/textbook/05.php**.

Milgram, Stanley. 1967. "The Small World Problem." *Psychology Today* 1 (1): 61–67.

Nolte, Nick. 2006. "Map of U.S. Route 6." Wikimedia Commons. Accessed January 16, 2013. **http://commons.wikimedia.org/wiki/File:US_6_map.png**.

United States Department of Transportation. 2005. "National Highway System." Wikimedia Commons. Accessed March 1, 2014. **http://commons.wikimedia.org/wiki/File:National_Highway_System.jpg**.

Watts, Duncan J. 1999. *Small Worlds: The Dynamics of Networks between Order and Randomness*. Princeton, NJ: Princeton University Press.

Watts, Duncan J., and Steven H. Strogatz. 1998. "Collective Dynamics of 'Small-World' Networks." *Nature* 393: 440–442.

White, Harrison C. 1970. "Search Parameters for the Small World Problem." *Social Forces* 49 (2): 259–264.

Image Credits

Airline Routes courtesy of NASA Ames Education Division.

See United States Department of Transportation 2005.

See Nolte 2006.

Network Analysis

Clustering

References

Clauset, Aaron, Christopher Moore, and Mark E. J. Newman. 2008. "Hierarchical Structure and the Prediction of Missing Links in Networks." *Nature* 453 (7191): 98–101.

Clauset Aaron, Mark E. J. Newman, and Christopher Moore. 2004. "Finding Community Structure in Very Large Networks." *Physical Review E* 70 (6): 066111.

Fortunato, Santo. 2010. "Community Detection in Graphs." *Physics Reports* 486:75–174.

Image Credits

Reprinted by permission from Macmillan Publishers Ltd: *Nature* © 2008.

Kapitalverflechtungen in Deutschland

References

Krempel, Lothar. 2006. "Die Deutschland AG 1996–2004 und die Entflechtung der Kapitalbeziehungen der 100 grössten deutschen Unternehmen." Accessed January 16, 2014. **http://www.socio.ethz.ch/modsim/tagungen/plenum06/krempel06slides.pdf**.

Quantifying Social Group Evolution

References

Palla, Gergely, Albert-László Barabási, and Tamás Vicsek. 2007. "Quantifying Social Group Evolution." *Nature* 446 (7136): 664–667.

Image Credits

Reprinted by permission from Macmillan Publishers Ltd: *Nature* © 2007.

The Risk Interconnection Map, 2013

References

World Economic Forum. 2013. *Global Risks 2013*. Cologny, Switzerland: World Economic Forum. 53.

U.S. Senate Voting Similarity Networks, 1975–2012

References

Moody, James, and Peter J. Mucha. 2013. "Portrait of Political Party Polarization." *Network Science* 1 (1): 119–121.

Staff. 2013. "U.S. Political Polarization Charted in New Study." Duke Today, May 20. Accessed February 16, 2014. **http://today.duke.edu/2013/05/us-political-polarization-charted-new-study#video.**

Image Credits

U.S. Senate Voting Similarity Networks, 1975–2012 available at **http://www.soc.duke.edu/~jmoody77/congress/NetScience_pubfinal.pdf**.

62 Network Visualization Types

References

Bertin, Jacques. 1983. *Semiology of Graphics*. Madison, WI: University of Wisconsin Press. 46.

Tree Layout

References

Hanrahan, Pat. 2001. "To Draw a Tree." Accessed January 16, 2014. **http://www-graphics.stanford.edu/~hanrahan/talks/todrawatree**.

Hanrahan, Pat. 2004. "Trees and Graphs." Accessed January 16, 2014. **http://graphics.stanford.edu/courses/cs448b-04-winter/lectures/treesgraphs/tree.graph.pdf**.

Tree View

References

Heer, Jeffrey. 2014. "Treeview." Prefuse. Accessed January 16, 2014. **http://prefuse.org/gallery/treeview**.

Software Credits

See Heer 2014.

Dendogram

References

Börner, Katy. 2010. *Atlas of Science: Visualizing What We Know*. Cambridge, MA: The MIT Press.

Morris, Steven A. 2005. *Visualizing 60 Years of Anthrax Research*. Stillwater, Oklahoma, 2005. Courtesy of Steven A. Morris, Oklahoma State University, Stillwater. In "1st Iteration (2005): The Power of Maps," *Places & Spaces: Mapping Science*, edited by Katy Börner and Deborah MacPherson. **http://scimaps.org**.

Radial Tree

References

Padgett, John F. 1986. "Florentine Families Dataset." Accessed March 1, 2014. **http://www.casos.cs.cmu.edu/computational_tools/datasets/sets/padgett**.

Data Credits

See Padgett 1986.

Software Credits

Network layout rendered using the Sci2 Tool. **http://sci2.cns.iu.edu**. Accessed September 18, 2014.

Link Tree

References

Ciccarelli, Francesca, Tobias Doerks, Christian Von Mering, Christopher J. Creevey, Berend Snel, Peer Bork. 2006. "Toward Automatic Reconstruction of a Highly Resolved Tree of Life." *Science* 311 (5765): 1283–1287.

Balloon Tree

Software Credits

Network layout rendered using the Sci2 Tool. **http://sci2.cns.iu.edu**. Accessed September 18, 2014.

Mosaic Graph

References

Friendly, Michael. 2002. "A Brief History of the Mosaic Display." *Journal of Computational and Graphical Statistics* 11 (1): 89–107.

Harris, Robert L. 1999. *Information Graphics: A Comprehensive Illustrated Reference*. New York: Oxford University Press.

Mayr, George von. 1874. *Gutachten über die Anwendung der Graphischen und Geographischen Methoden in der Statistik*. München: J. Gotteswinter.

Treemap

References

Bederson, Ben, Ben Shneiderman and Martin Wattenberg. 2002. "Ordered and Quantum Treemaps: Making Effective Use of 2D Space to Display Hierarchies." *ACM Transactions on Graphics* 21 (4): 833-854.

Bloch, Matthew, Shan Carter, and Amanda Cox. 2008. "All of Inflation's Little Parts." *The New York Times*, May 3. Accessed January 16, 2014. **http://www.nytimes.com/interactive/2008/05/03/business/20080403_SPENDING_GRAPHIC.html**.

Börner, Katy. 2010. *Atlas of Science: Visualizing What We Know*. Cambridge, MA: The MIT Press.

MarketWatch. 2014. "Map of the Market." Accessed March 1, 2014. **http://www.marketwatch.com/tools/stockresearch/marketmap**.

Shneiderman, Ben. 1992. "Tree Visualization with Tree-Maps: A 2-Dimensional Space Filling Approach." *ACM Transactions on Graphics* 11 (1): 92–99.

Network Layouts

Adjacency Matrix

References

Bertin, Jacques. 1983. *Semiology of Graphics*. Madison, WI: University of Wisconsin Press.

Börner, Katy, Shashikant Penumarthy, Mark Meiss, and Weimao Ke. 2006. "Mapping the Diffusion of Information Among Major U.S. Research Institutions." *Scientometrics* 68 (3): 415–426.

Doreian, Patrick, Vladimir Batagelj, and Anuska Ferligoj. 2005. *Generalized Blockmodeling*. Cambridge: Cambridge University Press.

Harris, Robert L. 1999. *Information Graphics: A Comprehensive Illustrated Reference*. New York: Oxford University Press.

Circular Graph

References

Börner, Katy. 2010. *Atlas of Science: Visualizing What We Know*. Cambridge, MA: The MIT Press.

Holten, Danny. 2006. "Hierarchical Edge Bundles: Visualization of Adjacency Relations in Hierarchical Data." *IEEE Transactions on Visualization and Computer Graphics* 12 (5): 741-748.

Image Credits

Image from Holten 2006. © 2006 IEEE. Reprinted, with permission, from *IEEE Transactions on Visualization and Computer Graphics*.

Data Credits

Eugene Garfield Web of Science data available at **http://wiki.cns.iu.edu/display/SCI2TUTORIAL/2.5+Sample+Datasets**. Accessed September 18, 2014.

Software Credits

Circular layout rendered using the Sci2 Tool. **http://sci2.cns.iu.edu**. Accessed September 18, 2014.

Hive Graph

References

Krzywinski, Martin. 2011. Hive Plots. Accessed January 30, 2014. **http://www.hiveplot.net**.

Krzywinski, Martin, Inanc Birol, Steven J. M. Jones, and Marco A. Marra. 2011. "Hive Plots: Rational Approach to Visualizing Networks." *Briefings in Bioinformatics* 13 (5): 627–644.

Image Credits

Hive graph type examples courtesy of Martin Krzywinski, Canada's Michael Smith Genome Sciences Center.

Node-Link Graph

References

Börner, Katy. 2010. *Atlas of Science: Visualizing What We Know*. Cambridge, MA: The MIT Press.

Garfield, Eugene, Elisha F. Hardy, Katy Börner, Ludmila Pollock, and Jan Witkowski. *HistCite Visualization of DNA Development*. Philadelphia, PA. Courtesy of Eugene Garfield, Thomson Reuters, Indiana University, and Cold Spring Harbor Laboratory. In "2nd Iteration (2006): The Power of Reference Systems," *Places & Spaces: Mapping Science*, edited by Katy Börner and Deborah MacPherson. **http://scimaps.org**.

Padgett, John F. 1986. "Florentine Families Dataset." Accessed March 1, 2014. **http://www.casos.cs.cmu.edu/computational_tools/datasets/sets/padgett**.

Data Credits

See Padgett 1986.

Software Credits

Network layout rendered using the Sci2 Tool. **http://sci2.cns.iu.edu**. Accessed September 18, 2014.

Sankey Graph

References

Phan, Doantam, Ling Xiao, Ron Yeh, Pat Hanrahan, and Terry Winograd. 2005. "Flow Map Layout." In *Proceedings of the IEEE Symposium on Information Visualization*, 219–224. Washington, DC: IEEE Computer Society.

Conceptual Drawings

References

Broder, Andrei, Ravi Kumar, Farzin Maghoul, Prabhakar Raghavan, Sridhar Rajagopalan, Raymie Stata, Andrew Tomkins, and Janet Wiener. 2000. "Graph Structure in the Web." *Computer Networks: The International Journal of Computer and Telecommunications Networking* 33 (1–6): 309–320.

Network Overlays

References

Börner, Katy. 2010. *Atlas of Science: Visualizing What We Know.* Cambridge, MA: The MIT Press.

Börner, Katy, Elisha F. Hardy, Bruce W. Herr II, Todd M. Holloway, and W. Bradford Paley. 2006. *Taxonomy Visualization of Patent Data*. Bloomington, Indiana and New York, New York. Courtesy of Indiana University and W. Bradford Paley. In "2nd Iteration (2006): The Power of Reference Systems," *Places & Spaces: Mapping Science*, edited by Katy Börner and Deborah MacPherson. **http://scimaps.org**.

Hennig, Marina, Ulrik Brandes, Jurgen Pfeffer and Ines Mergel. 2013. *Studying Social Networks: A Guide to Empirical Research*. Frankfurt: Campus Verlag.

Rafols, Ismael, Alan L. Porter, and Loet Leydesdorff. 2010. "Science Overlay Maps: A New Tool for Research Policy and Library Management." *JASIST* 61 (9): 1971–1887.

64 Studying Dynamics

References

Arijon, Daniel. 1976. *Grammar of the Film Language*. Los Angeles: Silman-James Press.

Cutting, James E. 2002. "Representing Motion in a Static Image: Constraints and Parallels in Art, Science, and Popular Culture." *Perception* 31: 1165–1194.

Lasseter, John. 1987. "Principles of Traditional Animation Applied to Computer Animation." ACM SIGGRAPH Computer Graphics 21 (4): 35–44, July 1987.

McCloud, Scott. 2003. *Understanding Comics*. Northampton, MA: Tundra.

ThinkExist. 2014. "Irish Blessings." Accessed January 17, 2014. **http://thinkexist.com/quotation/may_you_have_the_hindsight_to_know_where_you-ve/172094.html**.

Types of Dynamics

Image Credits

Growth of Boston Map reproduction courtesy of the Norman B. Leventhal Map Center at the Boston Public Library.

Presentation Types

One Static Image

References

Börner, Katy. 2017. *Atlas of Forecasts: Predicting and Broadcasting Science*. Cambridge, MA: The MIT Press.

Nelson, John. 2012. *Hurricanes & Tropical Storms—Locations and Intensities since 1851*. Lansing, MI. Courtesy of IDV Solutions. In "9th Iteration (2013): Science Maps Showing Trends and Dynamics," *Places & Spaces: Mapping Science*, edited by Katy Börner and Todd N. Theriault. **http://scimaps.org**.

Multiple Static Images

References

Börner, Katy. 2010. *Atlas of Science: Visualizing What We Know*. Cambridge, MA: The MIT Press.

Fairfield, Hannah. 2009. "Labor Lost, and Found." *The New York Times*, June 7. Accessed January 17, 2014. **http://www.nytimes.com/imagepages/2009/06/07/business/economy/20090607_metrics.html**.

Klavans, Richard, and Kevin W. Boyack. 2007. *Maps of Science: Forecasting Large Trends in Science*. Berwyn, PA and Albuquerque, NM. Courtesy of Richard Klavans, SciTech Strategies, Inc. In "3rd Iteration (2007): The Power of Forecasts," *Places & Spaces: Mapping Science*, edited by Katy Börner and Julie M. Davis. **http://scimaps.org**.

Minard, Charles Joseph. 1866. *Europe Raw Cotton Imports in 1858, 1864 and 1865*. 1866. Paris, France. Courtesy of the Library of Congress, Geography and Maps Division. In "4th Iteration (2008): Science Maps for Economic Decision Makers," *Places & Spaces: Mapping Science,* edited by Katy Börner and Elisha F. Hardy. **http://scimaps.org**.

Walsh, John A., David Becker, Bradford Demarest, Theodora Michaelidou, Laura Pence, and Jonathan Tweedy. *Literary Empires: Mapping Temporal and Spatial Settings of Victorian Poetry.* Bloomington, IN. Courtesy of Indiana University, with content provided by the David Rumsey Historical Map Collection. In "6th Iteration (2009): Science Maps for Scholars," *Places & Spaces: Mapping Science,* edited by Katy Börner and Elisha F. Hardy. **http://scimaps.org**.

Williams, Sarah, Carlo Ratti and Riccardo Maria Pulselli. 2006. *Mobile Landscapes: Using Location Data from Cell Phones for Urban Analysis*. Cambridge, MA. Courtesy of MIT SENSEable City Laboratory. In "5th Iteration (2009): Science Maps for Science Policy Makers," *Places & Spaces: Mapping Science,* edited by Katy Börner and Elisha F. Hardy. **http://scimaps.org**.

Evolving Reference System

References

Leydesdorff, Loet. 2010. *The Emergence of Nanoscience & Technology*. Amsterdam, Netherlands. Courtesy of Loet Leydesdorff, Thomas Schank, and *JASIST*. In "6th Iteration (2009): Science Maps for Scholars," *Places & Spaces: Mapping Science,* edited by Katy Börner and Elisha F. Hardy. **http://scimaps.org**.

Interactive Visualizations

References

Shneiderman, Ben. 1996. "The Eyes Have IT: A Task by Data Type Taxonomy for Information Visualizations." In *Proceedings of the IEEE Symposium on Visual Languages*, 336–343. Washington, DC: IEEE Computer Society.

Gapminder Visualization

Image Credits

Static image of data visualization from Gapminder World, powered by Trendanalyzer from **http://**

www.gapminder.org. Accessed September 18, 2014. "Google" is the registered trademark of Google Inc. Used with permission.

Virtual World User Activity

References

Börner, Katy, and Shashikant Penumarthy. 2003. "Social Diffusion Patterns in Three-Dimensional Virtual Worlds." *Information Visualization* 2 (3): 182–198.

Börner, Katy, and Shashikant Penumarthy. 2004. "Mapping Virtual Worlds and Their Inhabitants." Accessed March 7, 2014. **http://cns.iu.edu/images/pres/2004-borner-mapvwrld-poster.pdf**.

Image Credits

Extracted from Börner and Penumarthy 2004.

TTURC NIH Funding Trends

References

Stipelman, Brooke A., Kara L. Hall, Angela Zoss, Janet Okamoto, Daniel Stokols, and Katy Börner. 2014. "Mapping the Impact of Transdisciplinary Research: A Visual Comparison of Investigator-Initiated and Team-Based Tobacco Use Research Publications." *The Journal of Translational Medicine and Epidemiology*. 2 (2): 1033.

Image Credits

This research was supported in part by the National Cancer Institute at the National Institute of Health (contract number HHSN26100800812).

66 Combination

References

Goodreads. 2014. "Aristotle Quotes." Accessed January 17, 2014. **http://www.goodreads.com/author/quotes/2192.Aristotle**.

Image Credits

Images from Harris 1999 © Oxford University Press, 1999.

Multiple Views

References

Berghaus, Heinrich Karl Wilhelm. 1852. *Physikalischer Atlas: Geographisches Jahrbuch zur Mittheilung aller wichtigern neüen Enforschungen*. Gotha, Germany: Justus Perthes.

Small Multiples

References

Harris, Robert L. 1999. *Information Graphics: A Comprehensive Illustrated Reference*. New York: Oxford University Press. 386.

Tufte, Edward R. 1990. *Envisioning Information*. Cheshire, CT: Graphics Press. 28, 78.

Tabular Display

References

Harris, Robert L. 1999. *Information Graphics: A Comprehensive Illustrated Reference.* New York: Oxford University Press. 395.

Matrix Display

References

Harris, Robert L. 1999. *Information Graphics: A Comprehensive Illustrated Reference*. New York: Oxford University Press. 239.

Multilevel Display

References

Town of Bernalillo. 2014. "Geographic Information System." Accessed January 17, 2014. **http://www.townofbernalillo.org/depts/gis.htm**.

Heights of the Principal Mountains in the World, Lengths of the Principal Rivers in the World

References

Mitchell, Samuel Augustus. 1859. *A New Universal Atlas Containing Maps of the Various Empires, Kingdoms, States, and Republics of the World*. Philadelphia, PA: Charles Desilver.

Image Credits

Image reproduced from Mitchell 1859, courtesy of the David Rumsey Collection. **http://www.davidrumsey.com**. Accessed September 18, 2014.

Zoological Geography

References

Johnston, Alexander Keith. 1849. "Geographical Division and Distribution of Aves (Birds) over the World" and "Geographical Division and Distribution of the Birds of Europe." Accessed January 17, 2014. **http://libweb5.princeton.edu/visual_materials/maps/websites/thematic-maps/landmark-thematic-atlases/landmark-thematic-atlases.html#Johnston**.

Image Credits

Zoological Geography courtesy of the Historic Map Collection, Department of Rare Books and Special Collections, Princeton University Library.

Inter-Institutional Collaboration Explorer

References

Börner, Katy, Michael Conlon, Jon Corson-Rikert, and Ying Ding. 2012. *VIVO: A Semantic Approach to Scholarly Networking and Discovery*. San Rafael, CA: Morgan & Claypool.

U.S. Healthcare Reform

Image Credits

The U.S. Healthcare Reform map was created by Persistent Systems using their ShareInsights Big Data Analytics Platform, which analyzes conversations from unstructured and structured data sources.

68 Interaction

References

Becker, Richard A., and William S. Cleveland. 1987. "Brushing Scatterplots." *Technometrics* 29 (2): 127–142.

Cleveland, William S., and Marylyn E. McGill, eds. 1988. *Dynamic Graphics for Statistics*. Belmont, CA: Wadsworth & Brooks/Cole.

Eick, Stephen G. 1994. "Data Visualization Sliders." In *Proceedings of the 7th Annual ACM Symposium on User Interface Software and Technology*, 119–120. New York: ACM.

Hanrahan, Pat. 2004. "Trees and Graphs." Accessed January 16, 2014. **http://graphics.stanford.edu/courses/cs448b-04-winter/lectures/treesgraphs/tree.graph.pdf**.

Heer, Jeff, and Ben Shneiderman. 2012. "Interactive Dynamics for Visual Analysis." *Communications of the ACM* 55 (4): 45-54.

Shneiderman, Ben. 1994. "Dynamic Queries for Visual Information Seeking." *IEEE Software* 11 (6): 70–77.

Ward, Matthew O., Georges Grinstein, and Daniel Keim. 2010. *Interactive Data Visualization: Foundations, Techniques, and Applications*. Natick, MA: A. K. Peters.

Interaction Types

References

Chi, Ed. 1999. "A Framework for Visualizing Information." PhD thesis, University of Minnesota.

Visual View Manipulations

Filter

References

Ahlberg, Christopher, and Ben Shneiderman. 1994. "The Alphaslider: A Compact and Rapid Selector." In *Proceedings of the SIGCHI Conference on Human Factors in Computing*, 365–371. New York: ACM.

Ahlberg, Christopher, and Ben Shneiderman. 1994. "Visual Information Seeking: Tight Coupling of Dynamic Query Filters with Starfield Displays." In *Proceedings of the SIGCHI Conference on Human Factors in Computing*, 313–317. New York: ACM.

Shneiderman, Ben. 2007. "Dynamic Queries, Starfield Displays and the Path to Spotfire." Accessed January 17, 2014. **http://www.cs.umd.edu/hcil/spotfire**.

Detail on Demand

References

Furnas, G. W. 1986. "Generalized Fisheye Views." In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 16–23. New York: ACM.

Lamping, John, Ramana Rao, and Peter Pifolli. 1995. "A Focus +Context Technique Based on Hyperbolic Geometry for Visualizing Large Hierarchies." In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 401–408. New York: ACM/Addison-Wesley.

Roberts, Jonathan C. 2007. "State of the Art: Coordinated and Multiple Views in Exploratory Visualization." In *Proceedings of the Fifth International Conference on Coordinated and Multiple Views in Exploratory Visualization*, 61–71. Washington, DC: IEEE Computer Society.

Spence, Robert. 2007. *Book Information Visualization: Design for Interaction*. 2nd ed. Harlow, UK: Pearson/Prentice.

Interaction Support

References

Roberts, Jonathan C. 2007. "State of the Art: Coordinated and Multiple Views in Exploratory Visualization." In *Proceedings of the Fifth International Conference on Coordinated and Multiple Views in Exploratory Visualization*, 61–71. Washington, DC: IEEE Computer Society.

GRIDL

References

Shneiderman, Ben, David Feldman, Anne Rose, Xavier Ferré. 2000. "Visualizing Digital Library Search Results with Categorical and Hierarchical Axes." In *Proceedings of the Fifth ACM Conference on Digital Libraries*, 57–66. New York: ACM.

Image Credits

GRIDL image from the Human-Computer Interaction Lab, University of Maryland, **http://www.cs.umd.edu/hcil/west-legal/gridl/images/acm.gif**. September 18, 2014.

The Baby Name Wizard

References

Generation Grownup, LLC. 2014. "The Baby Name Wizard." Accessed January 20, 2014. **http://www.babynamewizard.com**.

Image Credits

Image courtesy of BabyNameWizard.com.

Seesoft: A Tool for Visualizing Line Oriented Software Statistics

References

Eick, Stephen G., Joseph L. Steffan, and Eric E. Sumner, Jr. 1992. "Seesoft—A Tool for Visualizing Line Oriented Software Statistics." *IEEE Transactions on Software Engineering* 18 (11): 957–968.

Image Credits

Image reproduced with permission of Alcatel-Lucent, © 1992

London Travel-Time Map

References

mySociety. 2014. "More Travel-Time Maps and Their Uses." Accessed January 20, 2014. **http://www.mysociety.org/2007/more-travel-maps**.

Image Credits

Travel time map created by mySociety as part of the Mapumental project, funded and supported by the Department for Transport. © Crown. All rights reserved. Department for Transport 100020237 2006.

70 Human–Computer Interface

Needs and Affordances

References

Heilig, Morton L. Sensorama Simulator. US Patent US 3050870 A, filed January 10, 1961, and issued August 28, 1962. Accessed January 20, 2014. **http://www.google.com/patents/US3050870**.

Payatagool, Chris. 2008. "Theory and Research in HCI: Morton Heilig, Pioneer in Virtual Reality Research." Telepresence Options: Your Guide to Visual Collaboration. Accessed January 20, 2014.

http://www.telepresenceoptions.com/2008/09/theory_and_research_in_hci_mor.

Proust, Marcel. 1929. *Remembrance of Things Past*. Translated by C.K. Scott Moncrieff. New York: Random House.

Robinett, Warren. 1994. "Interactivity and Individual Viewpoint in Shared Virtual Worlds: The Big Screen vs. Networked Personal Displays." *ACM SIGGRAPH Computer Graphics* 28 (2): 127–130.

Device Properties

Resolution

References

Wikimedia Foundation. 2014. "List of Displays by Pixel Density." *Wikipedia, the Free Encyclopedia*. Accessed January 20, 2014. **http://en.wikipedia.org/wiki/List_of_displays_by_pixel_density**.

Apple. 2014. iPad Home Page. Accessed January 20, 2014. **http://www.apple.com/ipad**.

File Size

References

Risinger, Nick. 2014. Sky Survey. Accessed January 20, 2014. **http://skysurvey.org**.

Brightness

References

Wikimedia Foundation. 2014. "Lumen (Unit)." *Wikipedia, the Free Encyclopedia*. Accessed January 20, 2014. **http://en.wikipedia.org/wiki/Lumen_%28unit%29**.

Device Options

Digital Displays

References

EBU. 2012. "4K and 8K UHDTV defined." Accessed August 20, 2014. **https://tech.ebu.ch/news/4k-and-8k-uhdtv-defined-16may12**.

Stereo Displays

References

Cruz-Neira, Carolina, Daniel J. Sandin, Thomas A. DeFanti. 1993. "Surround-Screen Projection-Based Virtual Reality: The Design and Implementation of the CAVE." In *Proceedings of SIGGRAPH '93*, 135–142. New York: ACM.

Cruz-Neira, Carolina, Daniel J. Sandin, Thomas A. DeFanti, Robert V. Kenyon, and John C Hart. 1992. "The CAVE: Audio Visual Experience Automatic Virtual Environment." *Communications of the ACM* 35 (6): 65–72.

Czernuszenko, Marek, Dave Pape, Daniel Sandin, Tom DeFanti, Gregory L. Dawe, and Maxine D. Brown. 1997. "The ImmersaDesk and Infinity Wall Projection-Based Virtual Reality Displays." *ACM SIGGRAPH Computer Graphics* 31 (2): 46 – 49.

Krüger, Wolfgang, and Bernd Fröhlich. "The Responsive Workbench." *IEEE Computer Graphics and Applications* 14 (3): 12–15.

Illuminated Diagram Display

References

Börner, Katy. 2010. *Atlas of Science: Visualizing What We Know.* Cambridge, MA: The MIT Press.

Paley, W. Bradford. 2002. "Illuminated Diagrams: Using Light and Print to Comparative Advantage." TextArc. Accessed January 20, 2014. **http://www.textarc.org/appearances/InfoVis02/InfoVis02_IlluminatedDiagrams.pdf**.

Augmented Reality and Wearables

References

Azuma, Ronald, Yohan Baillot, Reinhold Behringer, Steven Feiner, Simon Julier, and Blair MacIntyre. 2001. "Recent Advances in Augmented Reality." *IEEE Computer Graphics and Applications* 21 (6): 34–47.

Morton L. Heilig's Sensorama

References

Heilig, Morton L. Sensorama Simulator. US Patent US 3050870 A, filed January 10, 1961, and issued August 28, 1962. Accessed January 20, 2014. **http://www.google.com/patents/US3050870**.

Payatagool, Chris. 2008. "Theory and Research in HCI: Morton Heilig, Pioneer in Virtual Reality Research." Telepresence Options: Your Guide to Visual Collaboration. Accessed January 20, 2014. **http://www.telepresenceoptions.com/2008/09/theory_and_research_in_hci_mor**.

Sakane, Itsuo. 2011. "Morton Heilig's Sensorama (Interview).mov." YouTube. Accessed August 20, 2014. **https://www.youtube.com/watch?v=vSINEBZNCks**.

Image Credits

Image from the Morton Heilig website: "Inventor in the Field of Virtual Reality." Accessed March 1, 2014. **http://www.mortonheilig.com/InventorVR.html**.

Indiana University's Virtual Reality Theater

Image Credits

Chauncey Frend, the Advanced Visualization, a unit of the Research Technologies division of University Information Technology Services, © the Trustees of Indiana University. This content is released under the Creative Commons Attribution 3.0 Unported license. **http://creativecommons.org/licenses/by/3.0**. Accessed September 18, 2014.

Software Credits

Interactive walkthrough application was created at Indiana University using a combination of Rhino 3D and 3DVia Virtools.

Contributors

Mike Boyles kindly provided this image.

Giant Geo-Cosmos OLED Display

References

Yau, Nathan. 2012. "Giant Globe Display." FlowingData (blog), February 24. Accessed March 1, 2014. **http://flowingdata.com/2012/02/24/giant-globe-display**.

Image Credits

Image courtesy of Ingo Günther.

Contributors

Ingo Günther provided expert comments.

200 Countries, 200 Years, 4 Minutes

Image Credits

The Joy of Stats © Wingspan Productions Ltd 2011.

72 Validation and Interpretation

References

Kant, Immanuel. 1781. *Critik der reinen Vernunft*. Riga: Johann Friedrich Hartknoch.

Kimerling, A. Jon, Aileen R. Buckley, Phillip C. Muehrcke, and Juliana O. Muehrcke. 2012. *Map Use: Reading, Analysis, Interpretation*. 7th ed. Redlands, CA: ESRI Press.

Contributors

Colin Allen's quote is an adaptation of Kant's famous dictum (1781): "Thoughts without content are empty; intuitions without conceptions are blind."

Validation Criteria

References

McCandless, David. 2009. *The Visual Miscellaneum: A Colorful Guide to the World's Most Consequential Trivia.* New York: Harper Design.

Tufte, Edward R. 1990. *Envisioning Information*. Cheshire, CT: Graphics Press.

Tufte, Edward R. 1997. *Visual Explanations: Images and Quantities, Evidence and Narrative*. Cheshire, CT: Graphics Press.

Tufte, Edward R. 2001. *The Visual Display of Quantitative Information*. 2nd ed. Cheshire, CT: Graphics Press.

Contributors

W. Bradford Paley provided expert comments.

Function

Utility

References

Bateson, Gregory. 1973. *Steps to an Ecology of Mind*. Frogmore, St. Albans: Paladin. 428.

MacKay, Donald M. 1969. *Information, Mechanism and Meaning*. Cambridge, MA: The MIT Press.

Shannon, Claude E. 1948. "A Mathematical Theory of Communication." *The Bell System Technical Journal* 27: 379–423, 623–656.

Shannon, Claude E., and Warren Weaver. 1949. *The Mathematical Theory of Communication*. Urbana, IL: University of Illinois Press.

Effectiveness

References

Frankel, Felice C., and Angela H. DePace. 2012. *Visual Strategies: A Practical Guide to Graphics for Scientists and Engineers*. New Haven, CT: Yale University Press.

Tukey, John W. 1990. "Data-Based Graphics: Visual Display in the Decades to Come." *Statistical Science* 5:327–339.

Scalability

References

Eick, Stephen G., and Alan F. Karr. 2002. "Visual Scalability." *Journal of Computational and Graphical Statistics* 1 (11): 22–43.

Light, Robert P., David E. Polley, and Katy Börner. 2014. "Open Data and Open Code for Big Science of Science Studies." *Scientometrics*, February 19. Accessed March 7, 2014. **http://link.springer.com/article/10.1007/s11192-014-1238-2**.

Aesthetics

Accuracy

References

MacEachren, Alan M. 1994. *SOME Truth with Maps: A Primer in Symbolization and Design*. Washington, DC: Association of American Geographers.

Sebrechts, Marc M., John V. Cugini, Sharon J. Laskowski, Joanna Vasilakis, and Michael S. Miller. 1999. "Visualization of Search Results: A Comparative Evaluation of Text, 2D, and 3D Interfaces." In the *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 3–10. New York: ACM.

Validation Methods

References

Carpendale, Sheelagh. 2008. "Evaluating Information Visualizations." In *Information Visualization: Human-Centered Issues and Perspectives*, edited by Andreas Kerren, John T. Stasko, Jean-Daniel Fekete, Chris North, 19–45. Berlin: Springer-Verlag.

User Studies

References

Nielsen, Jakob. 2006. "F-Shaped Pattern for Reading Web Content." Nielsen Norman Group. Accessed January 20, 2014. **http://www.nngroup.com/articles/f-shaped-pattern-reading-web-content**.

Human (Expert) Validation

References

North, Chris. 2006. "Toward Measuring Visualization Insight." *IEEE Computer Graphics and Applications* 26 (3): 6–9.

Skupin, André, Joseph R. Biberstine, and Katy Börner. 2013. "Visualizing the Topical Structure of the Medical Sciences: A Self-Organizing Map Approach." *PLoS One* 8 (3): e58779. Accessed March 6, 2014. **http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0058779**.

Controlled Experiments on Benchmark Tasks

References

Chen, Chaomei, and Yue Yu. 2000. "Empirical Studies of Information Visualization: A Meta-Analysis," *International Journal of Human-Computer Studies* 53 (5): 851–866.

Crowdsourcing Evaluation

References

Heer, Jeffrey, and Michael Bostock. 2010. "Crowdsourcing Graphical Perception: Using Mechanical Turk to Assess Visualization Design." In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 203–212. New York: ACM.

Interpretation

References

Harley, John B. 1988. "Silences and Secrecy: The Hidden Agenda of Cartography in Early Modern Europe." *Imago Mundi* 40: 57–76.

Harley, John B. 1989. "Deconstructing the Map." *Cartographica* 26 (2): 1–20.

Harley, John B. 1990. "Cartography, Ethics, and Social Theory." *Cartographica* 27 (2): 1–23.

Index

T

U

V

W

Y

Z

"Finally, visualization reaches maturity as Katy Börner shows readers the powerful, but rarely seen, forces that shape our lives. Drawing on the work of thousands of creative visual designers, this book is a tribute to human ingenuity in creating our new world of visual thinking."

Ben Shneiderman, University of Maryland, author of *Designing the User Interface: Strategies for Effective Human-Computer Interaction*

"There are many books that now tell us how to use the plethora of new methods that let us visualize complex data. But Atlas of Knowledge *goes well beyond this, showing us in step-by-step fashion with key exemplars how we can harness this power of visualization. A wonderful guide for anyone who wishes to exploit the power of contemporary scientific visualization in their own work."*

Michael Batty, University College London, author of *The New Science of Cities*

"This isn't just a wonderfully illustrated coffee-table book, but one of the best references ever on how to design information graphics and visualizations. Just the sections on how to choose the best way to visually represent data depending on your goals are worth the price of the entire book."

Alberto Cairo, University of Miami, author of *The Functional Art: An Introduction to Information Graphics and Visualization*

"A brilliant synthesis of knowledge visualization practices and a timely contribution as the next wave of Big Data scholarship materializes. Börner's book is a 'must have.'"

Richard Rogers, University of Amsterdam, author of *Digital Methods*

"The Atlas *is a comprehensive and systematic compendium of state-of-the-art methods for knowledge communication and science evaluation. Using practical and relevant case studies, the* Atlas *illustrates trends in scientific knowledge management and mapping."*

Abel L. Packer, Federal University of São Paulo Foundation, Director of SciELO / FAPESP Program

science/design

90000
9 780262 028813

The MIT Press
Massachusetts Institute of Technology
Cambridge, Massachusetts 02142
http://mitpress.mit.edu

978-0-262-02881-3