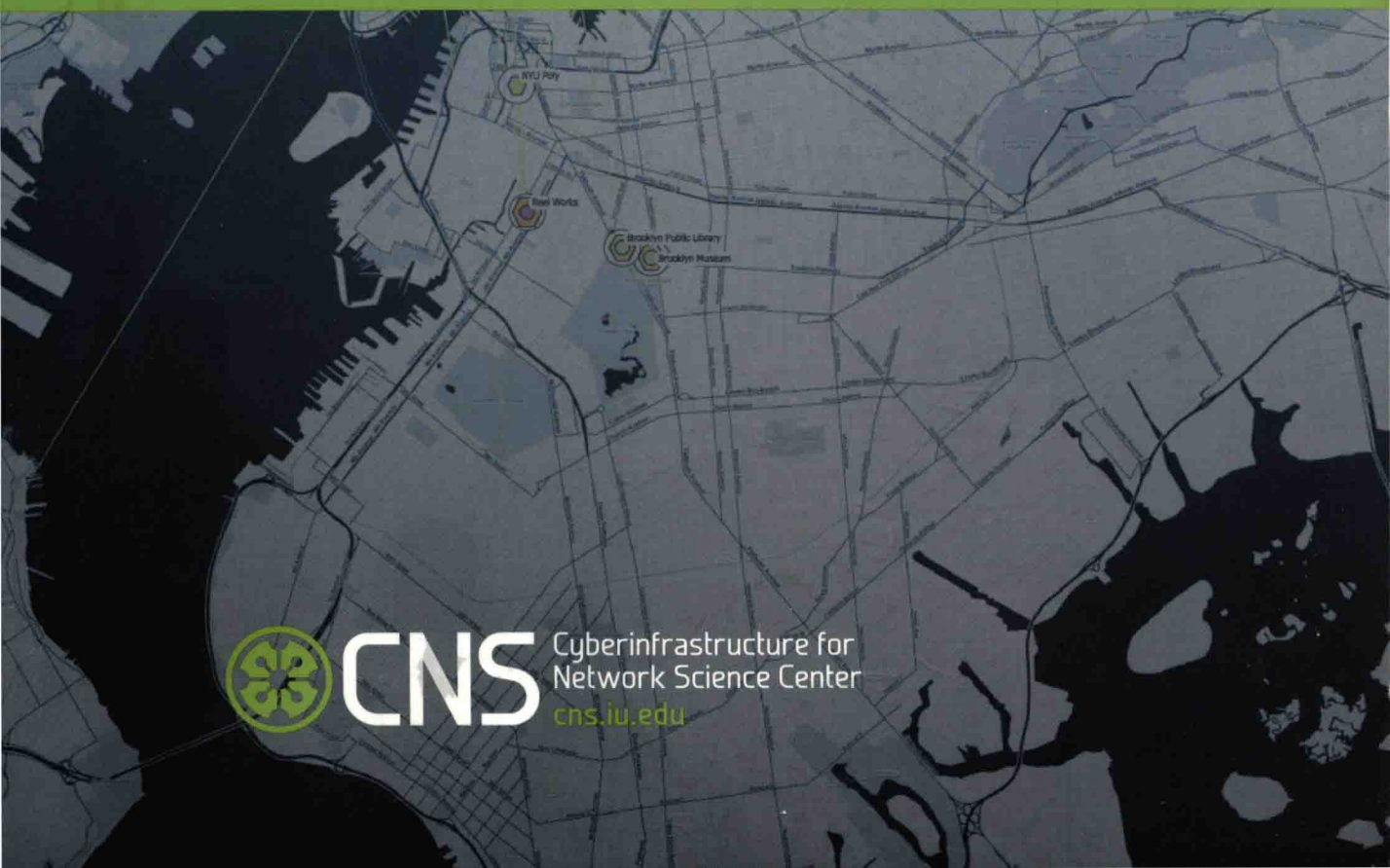
KATY BÖRNER & DAVID E. POLLEY

VISUAL INSIGHTS

A Practical Guide to Making Sense of Data



CNS Cyberinfrastructure for Network Science Center
cns.iu.edu

VISUAL INSIGHTS

A Practical Guide to Making Sense of Data

KATY BÖRNER & DAVID E. POLLEY

藏书章

The MIT Press
Cambridge, Massachussetts
London, England

© 2014 Massachusetts Institute of Technology

All rights reserved. No part of this book may be reproduced in any form by any electronic or mechanical means (including photocopying, recording, or information storage and retrieval) without permission in writing from the publisher.

MIT Press books may be purchased at special quantity discounts for business or sales promotional use. For information, please email special_sales@mitpress.mit.edu

This book was set in Open Sans by Samuel T. Mills (graphic design and layout), Cyberinfrastructure for Network Science Center, School of Informatics and Computing, Indiana University. Printed and bound in the United States of America.

Library of Congress Cataloging-in-Publication Data is available.

ISBN: 978-0-262-52619-7

10 9 8 7 6 5 4 3 2 1

VISUAL INSIGHTS

此为试读,需要完整PDF请访问: www.ertongbook.com

This book was written for all visual insight creators and consumers.

Note from the Authors

This book will be used as a companion resource for students taking the Information Visualization MOOC in January 2014 (http://ivmooc.cns.iu.edu). As such, we prioritized the timely availability of the text. While the text has been proof-read many times by several experts and two copy editors, and the workflows have been tested by many novice and power users on different operating platforms, we do appreciate error and bug reports sent to cns-sci2-help-l@iulist.indiana.edu so that remaining issues can be corrected in future editions.

There were many concepts we wished to cover in the book but were unable to, due to lack of space. Visual perception, cognitive processing, or how to perform human subject experiments are just a few facets out of many.

Finally, we acknowledge that the large size and interactive nature of some visuals contained in this book does not lend itself well to exploration via print format. Therefore, we have added a page to our website that contains links to high-resolution figures, conveniently organized by chapter and figure number (http://cns.iu.edu/ivmoocbook14). The green magnifying glass seen throughout the book indicates which figures are available online, and the associated links can be found in the figure titles.

Preface

In September 2012, I received a phone call from the deans of both iSchools at Indiana University (IU). Dean Robert B. Schnabel, School of Informatics and Computing, and Dean Debora "Ralf" Shaw, School of Library and Information Science, were interested in having me teach a massive online open course, or MOOC, in Spring 2013. I was immediately interested to explore this unique opportunity as the idea of "open education" fits extremely well with the "open data" and "open code" that the Cyberinfrastructure for Network Science Center (CNS), under my direction, is creating and promoting. I had been teaching open data and code workshops in many countries over the last ten years, and more than 100,000 users had downloaded our plug-and-play macroscope tools.[1] David E. Polley had recently joined our team, testing and documenting software, and teaching tool workshops at IU and international conferences. My PhD student Scott B. Weingart turned out not only to be a remarkable researcher and juggler but also an inspiring teacher. A January deadline seemed feasible—particularly with extensive support by IU—and I said yes to teach an Information Visualization MOOC (called IVMOOC) in the Spring 2013 semester.

We soon learned that Indiana University had decided to use the open source Google Course Builder (GCB)[2] platform for all MOOC development and teaching. At that moment in time, GCB had been used once—to teach *Power Searching with Google*[3] to more than 100,000 students. GCB had no support for sending out emails or grading work; setting up the course or assessments involved low-level coding and scripting. Interested to have IVMOOC students interact with me, others at CNS, each other, and external clients, we hired Mike Widner and Scott B. Weingart to implement a Drupal forum for GCB. To fill the need to grade work, Robert P. Light designed the IVMOOC database that captured not only students' scores in assessments but also who collaborated with whom, who watched what video for how long, etc. Ultimately, MOOC users need new techniques and tools to be most effective—teachers need to make sense of the activities of thousands of students, and students need to navigate learning materials and develop successful learning collaborations across disciplines and time zones—for example, to conduct client project work (see Chapter 9 on MOOC Visual Analytics).

In parallel to developing and recording materials for the IVMOOC, I was working on the *Atlas of Knowledge*, which has the subtitle "Anyone Can Map," inspired by Auguste Gusteau's catchphrase "Anyone Can Cook." The *Atlas* aims to feature timeless knowledge (Edward Tufte called it "forever knowledge"), or, principles that are indifferent to culture, gender, nationality, or history. In contrast, the IVMOOC features "timely knowledge," or, the most current data formats, tools, and workflows used to convert data into insights.

---

1 Börner, Katy. 2011. "Plug-and-Play Macroscopes." *Communications of the ACM* 54, 3: 60–69.

2 http://code.google.com/p/course-builder

3 http://www.google.com/insidesearch/landing/powersearching.html

Specifically, IVMOOC materials are structured into seven units to be taught over seven weeks (see Chapters 1–7 in this book). Each weekly unit features a theoretical component by me and a hands-on component by David E. Polley. The first theory unit introduces a theoretical visualization framework intended to help non-experts to assemble advanced analysis workflows and to design different visualization layers. The framework can also be applied to "dissect visualizations" for optimization or interpretation. The subsequent five units introduce workflows and visualizations that answer when, where, what, and with whom questions using temporal, geospatial, topical, and network analysis techniques. The final unit covers visualizations of dynamically changing data and the optimization of visualizations for different output media. The hands-on components feature in-depth instruction on how to navigate and operate several software programs used to visualize information. Furthermore, students learn the skills needed to visualize their very own data, allowing them to create unique visualizations. Pointers to the extensive Sci2 Online Tutorial[4] are provided where relevant. The theory component and the hands-on component are standalone. Participants can watch whichever section they are more interested in first, and then review the other section. After the theory videos there are self-assessments, and after the hands-on videos are short homework assignments.

Before, during, and after the course, students are encouraged to create and use Twitter and Flickr accounts and the tag "ivmooc" to share images as well as links to insightful visualizations, conferences and events, or relevant job openings to create a unique, real-time data stream of the best visualizations, experts, and companies that apply data mining and visualization techniques to answer real-world questions.

This graduate-level course is free and open to participants from around the world, and anyone who registers gains free access to the Scholarly Database[5] with 26 million paper, patent, and grant records and the Sci2 Tool[6] with 100+ algorithms and tools. Students also have the opportunity to work with actual clients on real-world visualization projects.

The IVMOOC final grade is based on results from the midterm exam (30%), final exam (40%), and projects/homework (30%). All participants that receive more than 80% of all available points will receive both a letter of accomplishment and badge.

Feel free to register for IVMOOC at http://ivmooc.cns.iu.edu and enjoy.

Katy Börner
Cyberinfrastructure for Network Science Center
School of Informatics and Computing
Indiana University
August 18, 2013

4 http://sci2.wiki.cns.iu.edu
5 http://sdb.cns.iu.edu
6 http://sci2.cns.iu.edu

Acknowledgments

I would like to thank Robert Schnabel, Dean of the School of Informatics and Computing, and Debora Shaw, then Dean of the School of Library and Information Science, Indiana University for inspiring the development of the IVMOOC.

This MOOC would not have been possible without the institutional support of Lauren K. Robel, Munirpallam A. Venkataramanan, Jennifer W. Adams, Barbara Anne Bichelmeyer, and Ilona M. Hajdu as diverse copyright, terms of service, and legal issues had to be resolved before any student could register.

We would like to thank Miguel Lara for extensive instructional design support throughout the development and teaching of the IVMOOC; Samuel T. Mills for designing the IVMOOC web pages; Robert P. Light and Thomas Smith for extending the GCB platform; Mike Widner, Scott B. Weingart, and Mike T. Gallant for adding a Drupal forum to GCB; Ralph A. Zuzolo and his team for recording the teaser video; and Rhonda Spencer, James P. Shea, and Tracey Theriault for marketing.

Many visualizations used in the IVMOOC and in this book come from the *Places & Spaces: Mapping Science* exhibit, online at http://scimaps.org, and from the *Atlas of Science: Visualizing What We Know* (MIT Press 2010). The Sci2 Tool and the Scholarly Database were developed by more than forty programmers and designers at CNS.

We would like to thank Samuel T. Mills for designing the book cover, redesigning many figures and tables featured here, and performing the complete book layout, Joseph J. Shankweiler for gathering the screenshots for the book, Tassie Gniady and Arul K. Jeyaseelan for testing the Sci2 workflows in the book, Scott R. Emmons and Simone L. Allen for providing feedback to a draft of the book, and Todd N. Theriault and Lisel G. Record for editing the book.

Marguerite B. Avery, Senior Acquisitions Editor, and Karie Kirkpatrick, Senior Publishing Technology Specialist, both at The MIT Press were instrumental in having the book edited, proofread, and published in time for the 2014 IVMOOC.

Last but not least, we thank all 2013 IVMOOC students for their feedback and comments, enthusiasm, and support.

Support for the IVMOOC development comes from the Cyberinfrastructure for Network Science Center, the Center for Innovative Teaching and Learning, the School of Informatics and Computing (SoIC), the former School of Library and Information Science—now the Department of Information and Library Science at SoIC, the Trustees of Indiana University, and Google. Open data and open code development work is supported in part by the National Science Foundation under Grants No. SBE-0738111, DRL-1223698, and IIS-0513650, the U.S. Department of Agriculture, the National Institutes of Health under Grant No. U01 GM098959, and the James S. McDonnell Foundation.

此为试读,需要完整PDF请访问: www.ertongbook.com

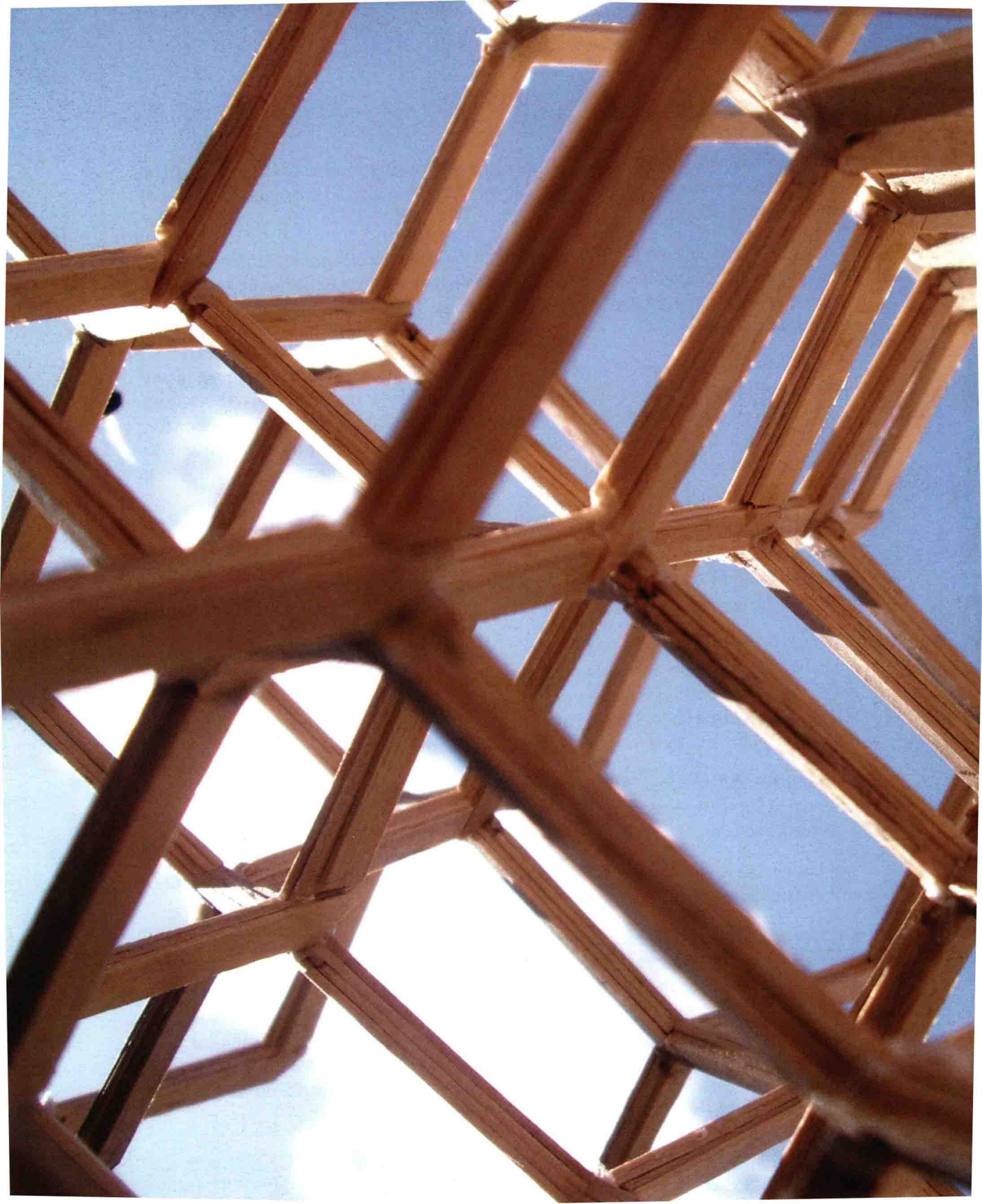


Table of Contents

Chapter One

Visualization Framework and Workflow Design

Chapter 1: Theory Section

Welcome to the Information Age, where each one of us receives more information via tweets, emails, news, and other data streams each day than can humanly be processed in 24 hours; and anyone with an Internet connection has access to a majority of humankind's knowledge. Our offices are filling up and our email inboxes are overflowing (see Figure 1.1, left). We urgently need more effective ways to make sense of this massive amount of data—to navigate and manage information, to identify collaborators and friends, or to notice patterns and trends (see Figure 1.1, right).

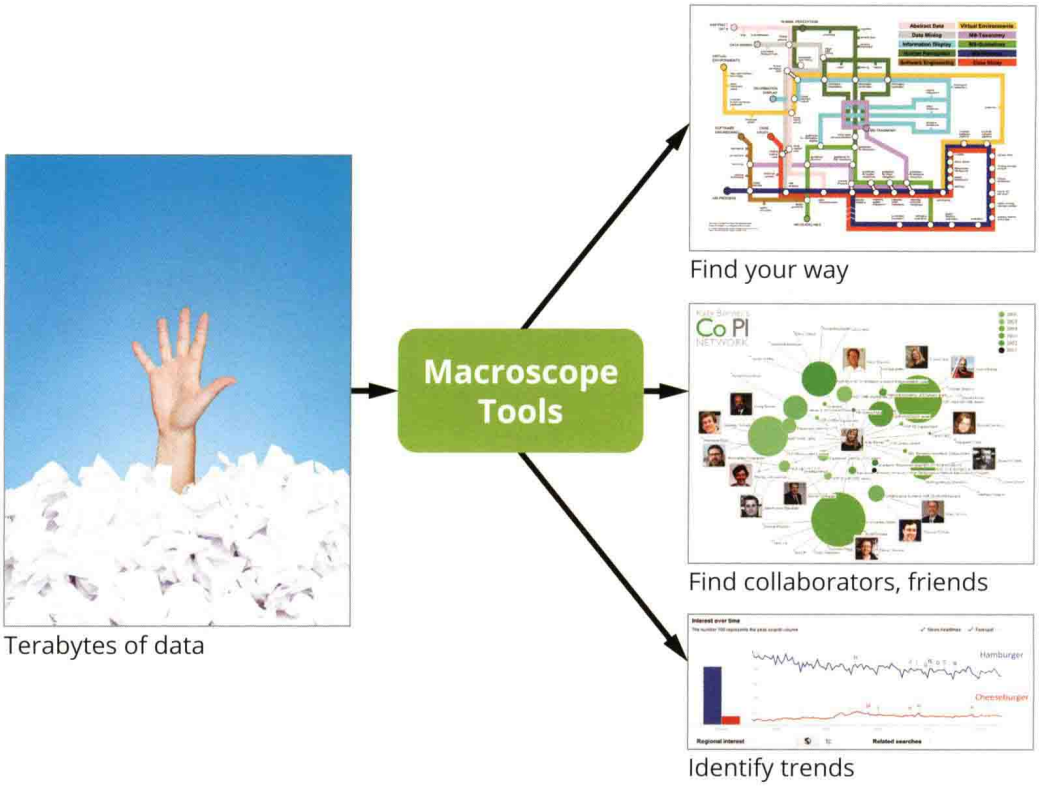


Figure 1.1 Converting data into actionable insights

This book teaches you how to use advanced data mining and visualization techniques to convert data into insights. Each chapter has a theory part on white background paper and a hands-on section on gold. The theory part comes with self-assessments while the hands-on part contains homework assignments.

This first chapter presents a theoretical visualization framework that helps to select the most appropriate algorithms and to assemble them into effective workflows. Its hands-on

section introduces so-called macroscope tools[1] that empower anyone to read, process, analyze, and visualize data.

1.1 VISUALIZATION FRAMEWORK

This section introduces a framework that helps organize and group visualizations and supports the identification of what type and what level of analysis is best to address specific user needs.

The use of grouping to develop organizational frameworks has a long history outside of information visualization. In fact, science often begins by grouping or classifying things. For example, zoologists classify animals so that tigers and lions and jaguars end up in the family Felidae (big cats). Dmitri Mendeleev grouped chemical elements in the periodic table according to chemical properties and atomic weights, leaving "holes" in the table for elements yet to be discovered. Similarly, the systematic analysis and grouping of information visualizations helps in the design of new visualizations, and also in interpreting visualizations encountered in journals, newspapers, books, and other publications.

There are various ways to group visualizations, such as those shown in Figure 1.2. Visualizations can be grouped by user insight needs, by user task types, or by the data to be visualized. They can also be grouped based on what data mining techniques are used, what interactivity is supported, or by the type of deployment (e.g., whether the visualizations are printed on paper, animated, or presented on interactive displays). Details and references to existing taxonomies and frameworks can be found in the *Atlas of Knowledge.*[2]

Here, a pragmatic approach is applied to teach anyone how to design meaningful visualizations. Starting with the types of questions users have, the framework supports the selection of data mining and visualization workflows as well as deployment options that answer these user questions.

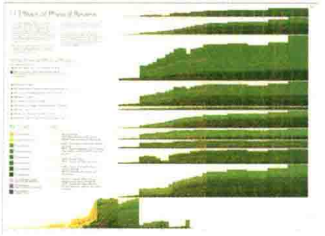
Levels of Analysis and Types of Analysis

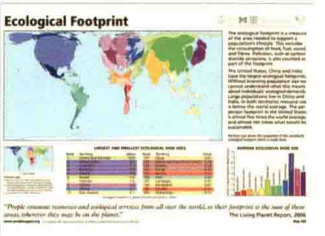
The visualization framework distinguishes three levels of analysis: micro, meso, and macro. The **micro** level, or the individual level, consists of small datasets, typically between 1 and 100 records—for example, one person and all of his/her friends. The next level is the **meso**, or the group level, about 101 to 10,000 records. An example might include information about researchers working at a single university or on a certain research topic. Finally, the broadest level of analysis is the **macro**, sometimes referred to as the global or population level. Datasets for projects at this level of analysis typically exceed 10,000 records, such as data pertaining to an entire country or all of science.

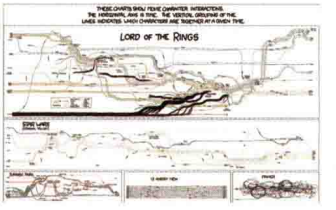
---

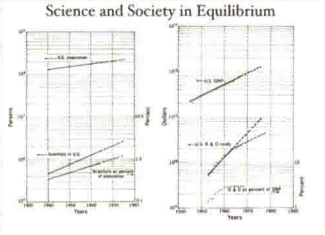
1 Börner, Katy. 2011. "Plug-and-Play Macroscopes." *Communications of the ACM* 54, 3: 60–69.

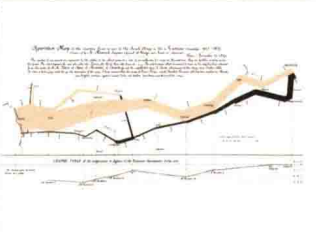
2 Börner, Katy. 2014. *Atlas of Knowledge: Anyone Can Map*. Cambridge, MA: The MIT Press.

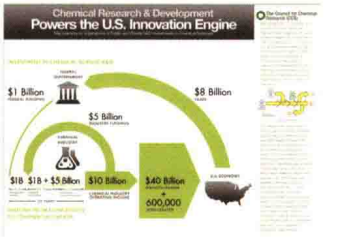


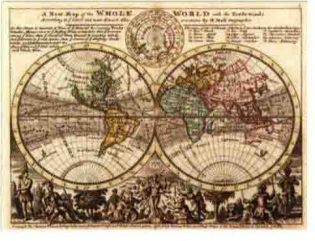


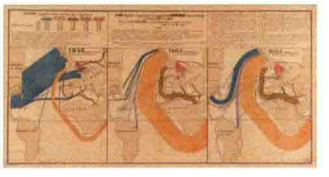


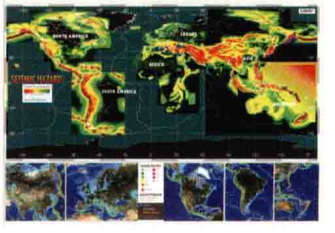


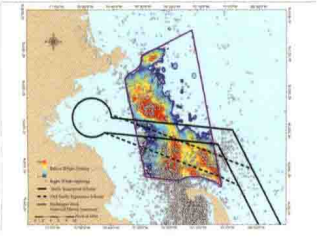


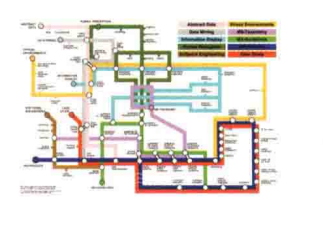












此为试读,需要完整PDF请访问: www.ertongbook.com