Making Data Visual
A Practical Guide to Using Visualization for Insight

Danyel Fisher and Miriah Meyer

O'REILLY®
Beijing • Boston • Farnham • Sebastopol • Tokyo
Making Data Visual

by Danyel Fisher and Miriah Meyer

Copyright © 2018 Miriah Meyer, Microsoft. All rights reserved.

Printed in the United States of America.

Published by O’Reilly Media, Inc., 1005 Gravenstein Highway North, Sebastopol, CA 95472.

O’Reilly books may be purchased for educational, business, or sales promotional use. Online editions are also available for most titles (http://oreilly.com/safari). For more information, contact our corporate/institutional sales department: 800-998-9938 or corporate@oreilly.com.

Revision History for the First Edition

- 2017-12-19: First Release


The O’Reilly logo is a registered trademark of O’Reilly Media, Inc. Making Data Visual, the cover image, and related trade dress are trademarks of O’Reilly Media, Inc.

While the publisher and the authors have used good faith efforts to ensure that the information and instructions contained in this work are accurate, the publisher and the authors disclaim all responsibility for errors or omissions, including without limitation responsibility for damages resulting from the use of or reliance on this work. Use of the information and instructions contained in
this work is at your own risk. If any code samples or other technology this work contains or describes is subject to open source licenses or the intellectual property rights of others, it is your responsibility to ensure that your use thereof complies with such licenses and/or rights.

978-1-491-92846-2

[LSI]
Preface

Visualization is a vital tool for understanding and sharing insights around data. The right visualization can help express a core idea or open a space to examination; it can get the world talking about a dataset or sharing an insight Figure P-1.

![Image of visualizations](image.png)

Figure P-1. Visualizations can take many forms, from views that support exploratory analysis (top left), to those that provide quick overviews in a dashboard (bottom), to an infographic about popular topics (top right).

Visualizations provide a direct and tangible representation of data. They allow people to confirm hypotheses and gain insights. When incorporated into the data analysis process early and often, visualizations can even fundamentally alter the questions that someone is asking.

Creating effective visualizations is hard. Not because a dataset requires an exotic
and bespoke visual representation—for many problems, standard statistical charts will suffice. And not because creating a visualization requires coding expertise in an unfamiliar programming language—off-the-shelf tools like Excel, Tableau, and R are ample enough to suffice.

Rather, creating effective visualizations is difficult because the problems that are best addressed by visualization are often complex and ill-formed. The task of figuring out what attributes of a dataset are important is often conflated with figuring out what type of visualization to use. Picking a chart type to represent specific attributes in a dataset is comparatively easy. Deciding on which data attributes will help answer a question, however, is a complex, poorly defined, and user-driven process that can require several rounds of visualization and exploration to resolve. In this book, we focus on the process of going from high-level questions to well-defined data analysis tasks, and on how to incorporate visualizations along the way to clarify understanding and gain insights.
Who Is This Book For?

This book is for people who have access to data and, perhaps, a suite of computational tools but who are less than sure how to turn that data into visual insights. We find that many data science books assume that you can figure out how to visualize the data once collected, and visualization books assume that you already have a well-defined question, ready to be visualized. If, like us, you would like to address these assumptions, then this book is for you.

This book does not cover how to clean and manage data in detail or how to write visualization code. There are already great books on these topics (and, when relevant, we point to some of them). Rather, this book speaks to why those processes are important. Similarly, this book does not address how to choose a beautiful colormap or select a typeface. Instead, we lay out a framework for how to think about data given the possibilities and constraints of visual exploration. Our goal is to show how to effectively use visualizations to make sense of data.

Who Are We?

The authors of this book have a combined three decades of experience in making sense of data through designing and using visualizations. We have worked with data from a broad range of fields: biology and urban transportation, business intelligence and scientific visualization, debugging code and building maps. We have worked with analysts from a variety of organizations, from small, academic science labs to teams of data analysts embedded in large companies. Some of the projects we have worked on have resulted in sophisticated, bespoke visualization systems designed collaboratively with domain specialists, and at other times we have pointed people to off-the-shelf visualization tools after a few conversations. We have taught university classes in visualization and have given lectures and tutorials. All in all, we have visualized hundreds of datasets.

We have found that our knowledge about visualization techniques, solutions, and systems shapes the way that we think and reason about data. Visualization is fundamentally about presenting data in a way that elicits human reasoning, makes room for individual interpretations, and supports exploration. We help our
collaborators make their questions and data reflect these values. The process we lay out in this book describes our method for doing this.

**Overview of Chapters**

Chapter 1 illustrates the process of making sense with visualizations through a quick example, exposing the role that a visual representation can play in data discovery.

Chapter 2 starts to get into details. It discusses a mechanism to help narrow a question from a broad task into something that can be addressed with an iterative visualization process. For example, the broad question “Who are the best movie directors?” does not necessarily suggest a specific visualization—but “Find movie directors who directed top-grossing movies using an IMDB dataset” can lead more directly to an answer by way of a visualization or two. This process creates an *operationalized* question, one that consists of particular tasks that can be directly addressed with data.

This process of narrowing a question down to actionable tasks requires input from multiple stakeholders. Chapter 3 lays out an iterative set of steps for getting to the operationalization, which we call *data counseling*. These steps include finding the right people to talk to, asking effective questions, and rapidly exploring the data through increasingly sophisticated prototypes.

The numerical nitty-gritty of the book follows. Chapter 4 discusses types and relations of data, and defines terms like *dimensions*, *measures*, *categorical*, and *quantitative*. Chapter 5 then organizes common visualization types by the tasks they fulfill and the data they use. Then, Chapter 6 explores powerful visualization techniques that use multiple views and interaction to support analysis of large, complex datasets. These three chapters are meant to provide an overview of some of the most effective and commonly used ideas for supporting sensemaking with visualizations, and are framed using the operationalization and data counseling process to help guide decision-making about which visualizations to choose.

With this understanding of getting to insight—from questions to data to visualizations—the remainder of the book illustrates two examples of carrying out these steps. The case study in Chapter 7 describes the creation of a business
intelligence dashboard in collaboration with a team of developers and analysts at Microsoft. The one in Chapter 8 draws from science, presenting an example with a team of scientists who work with biological data. These case studies illustrate the flexibility of the process laid out in this book, as well as the diverse types of outcomes that are possible.

This book is accompanied by a companion website. From this site you can download the code and interactive versions of the visualizations presented in Chapters 5 and 6, as well as other code and supplementary material.

**Acknowledgments**

Danyel and Miriah would like to thank Danyel’s colleagues at Microsoft, including Steven Drucker, Mary Czerwinski, and Sue Dumais, for their enthusiasm and encouragement. We also thank Miriah’s research group, the Visualization Design Lab at the University of Utah, including Alex Lex, for helping the project to evolve and providing feedback on ideas. We are both deeply appreciative of our work organizations for supporting the time and energy required by projects like this one, and for seeing the value in communicating our research broadly. At O’Reilly Media, we thank Mike Loukides for encouraging us to start this work, and Shannon Cutt and Rachel Roumeliotis for guiding it from start to finish.

Portions of this work were presented at the IPAM Workshop on Culture Analytics of 2016, Microsoft Data Insights Summit of 2016, University of Illinois Urbana-Champaign HCI Seminar Series of 2016, University of British Columbia HCI Seminar Series of 2016, Women in Data Science Conference at Stanford University in 2017, and O’Reilly Velocity Conference in 2017. Our thanks to the organizers of those events, and to participants who gave us critical feedback and helped clarify our thoughts.

Early feedback on the operationalization process came from Christian Canton of Microsoft. Michael Twidale and Andrea Thomer, both of UIUC, helped inform the discussion of data counseling with their insights on how reference librarians do their work.

We are grateful to Dominik Mortiz and Kanit “Ham” Wongsuphasawat for putting together the examples used in Chapters 5 and 6. Their work, as well as
that of the rest of the Vega-Lite team, is helping shape the future of data visualization. We also thank Alex Bigelow for supplying the skateboarding visualization figure in the Preface.

We thank Jacqueline Richards for her review and discussion of the case study in Chapter 7. Similarly, the collaboration with Angela DePace and her group at the Harvard Medical School for the case study in Chapter 8 provided valuable and rich insights into the process of designing visualizations for domain experts. The projects described in both of these chapters were deeply influential in our work practices.

Our technical reviewers, Michael Freeman, Jeff Heer, and Jerry Overton, helped clarify and strengthen the arguments we make.

Finally, Miriah thanks Brian Price for his endless support and encouragement, without which she could never do the things she does.

O’Reilly Safari

Safari (formerly Safari Books Online) is a membership-based training and reference platform for enterprise, government, educators, and individuals.


For more information, please visit http://oreilly.com/safari.

How to Contact Us

Please address comments and questions concerning this book to the publisher:

O’Reilly Media, Inc.
1005 Gravenstein Highway North

Sebastopol, CA 95472

800-998-9938 (in the United States or Canada)

707-829-0515 (international or local)

707-829-0104 (fax)

We have a web page for this book, where we list errata, examples, and any additional information. You can access this page at http://bit.ly/making-data-visual.

To comment or ask technical questions about this book, send email to bookquestions@oreilly.com.

For more information about our books, courses, conferences, and news, see our website at http://www.oreilly.com.

Find us on Facebook: http://facebook.com/oreilly

Follow us on Twitter: http://twitter.com/oreillymedia

Watch us on YouTube: http://www.youtube.com/oreillymedia
Chapter 1. Getting to an Effective Visualization

Choosing or designing a good visualization is rarely a straightforward process. It is tempting to believe that there is one beautiful visualization that will show all the critical aspects of a dataset. That the right visual representation will reveal hidden insights. That a perfect, simple, and elegant visualization—perhaps just a line chart or a well-chosen scatterplot—will show precisely what the important variable was and how it varied in precisely the way to illustrate a critical lesson.

This is often the impression that we, at least, are left with after reading data science case studies. But in our experience, this does not match the reality of visual data analysis. It takes hard work, and trial and error, to get to an insightful visualization. We start by thinking about what we want to know, and we refine fuzzy questions into actionable, concrete tasks. We clean, reshape, and restructure the data into forms that we can put into a visualization. We work around limitations in the data, and we try to understand what the user wants to learn. We have to consider which visual representations to use and what interaction mechanisms to support. Along the way, we find other variables that tell us more about the dataset and that help clarify our thinking. And no single visualization is ever quite able to show all of the important aspects of our data at once—there just are not enough visual encoding channels.

Designing effective visualizations presents a paradox. On the one hand, visualizations are intended to help users learn about parts of their data that they don’t know about. On the other hand, the more we know about the users’ needs and the context of their data, the better we can design a visualization to serve them. The process described in this book embraces this paradox: it leverages the knowledge users have of their datasets, the context the data lives in, and the ways it was collected—including its likely flaws, challenges, and errors—in order to figure out the aspects of it that matter.

Put another way, this book is about the path from “I have some data…” to “We know this because of these clear, concise, and insightful visualizations.” We
believe that creating effective visualizations is itself a process of exploration and discovery. A good visualization design requires a deep understanding of the problem, data, and users.

**Getting to Insight**

We most often work with other people that have a dataset they are trying to make sense of. The process of designing a visualization usually starts when people walk into our office.

*CLIENT: I have some data that I’d like to visualize. How should I draw it?*

The client seems to expect us to pull a visualization off the shelf, to sculpt that perfect visualization. We almost always frustrate them by asking what they hope to see.

*Q: What is it about the data that you would like to visualize?*

*CLIENT: I want to see how profitable our stores are.*

*Q: What in your data indicates a store being profitable?*

*CLIENT: It means that the store has lots of sales of high-profit items.*

*Q: How does profit vary by store?*

And so on.

By the end of this process, we often find that the clients do not have a visualization problem, but an operationalization one. Their struggles to choose a visualization stem from a lack of clarity about which attributes of the data are most important and how those attributes relate to one another. Once they can describe how the data attributes relate to the question they are trying to answer, finding an appropriate visualization becomes much easier.

We have learned over the years that designing effective visualizations to make sense of data is not an art—it is a systematic and repeatable process. We have systematized this process into what we believe are reproducible and clear steps. This process tracks our understanding of four components:
Data

What data is available, and what does it mean? What does the data look like, and what are its important aspects? Where did it come from, and why was it originally collected?

Tasks

What needs to happen with the data? What are the low-level questions and tasks that will support high-level goals?

Stakeholders

Who is involved with the data, the problem, and the goals? What can they say about the problem to help design an effective visualization? Who will view the final visualization, and what sorts of things do we expect them to learn from it? What domain knowledge do they bring to the table? What answers would they find satisfying?

Visualization

How does the understanding of data, tasks, and stakeholders come together? What representations of this data will fulfill the tasks for the users?

Regardless of the visualization outcome, this process will almost certainly lead to new discoveries and insights. These discoveries help to inform the operationalization, but they will also likely steer the process down new and unexpected paths. The guidance and framework in this book are meant to help identify opportunities for discovering new knowledge and to make an otherwise messy process a bit more structured.

Hotmap: Making Decisions with Data

As an example of how visualizations can help you to better understand a problem, and help an organization make decisions, we can look back to 2006. Microsoft was rolling out its new mapping tool, Virtual Earth, a zoomable world map. The team behind Virtual Earth had lots of questions about how their users were using this new tool, so they collected usage data.

The usage data was based on traditional telemetry: it had great information on
what cities were most viewed, how many viewers were in “street” mode versus “photograph” mode, and even information about viewers’ displays. They instrumented search and navigation, and they collected counts for the number of times that users looked at certain sentinel regions. And because Virtual Earth was built on top of a set of progressively higher-resolution image tiles, the team was also archiving server logs that tracked how often individual tiles were downloaded.

Interviews with team members suggested that they did not have an intuitive notion of how their tool was being used. In conversation, one team member argued that people were likely to look at their own homes; another thought that the overhead photography would mostly be used over mountains. The goals were varied: they included seeing whether the user experience was well balanced across user needs and deciding how and where to invest in future rounds of photography.

We addressed these questions with a visualization tool called Hotmap. Figure 1-1 shows a screen capture from the visualization tool, focusing on the central United States. Hotmap uses a heatmap encoding of the tile access values. This is a visualization technique that uses a colormap to encode the access values at the geospatial locations of the tiles. Colored spots on the map are places where more users have accessed image tiles. The colormap is a logarithmic color scale, so bright spots have many more accesses than dim ones.

![Figure 1-1. Hotmap, looking at the central United States. The white box surrounds an anomaly in South Dakota.](image)
Some of the brightest areas correspond to major population centers—Chicago and Minneapolis on the right, Denver and Salt Lake City in the middle, and West Coast cities on the left. Near the center, though, is an anomalous shape: a bright spot where no big city exists. There is a star shape around the bright spot, and an arc of bright colors nearby. The spot is in a sparsely populated bit of South Dakota—there was no obvious reason to the team why users might zoom in there.

That point is, however, very close to the center of a map of the continental US. In fact, the team learned that the center of the star corresponds to the center of the default placement of the map in many browsers. The bright spot with the star most likely corresponds to users sliding around after inadvertently zooming in, trying to figure out where they’ve landed; the arc seems to correspond to variations in monitor proportions.

As a result of this usability challenge, many mapping tools—including Bing Maps (the successor product to Virtual Earth)—no longer offer a zoom slider, which keeps users from accidentally zooming all the way in on a single click.

A second screen capture, shown in Figure 1-2, reveals a bright spot off the coast of Ghana. This spot exhibits the same star pattern created by users scrolling around to try to figure out what part of the map they are viewing. This spot is likely only bright because it is at 0 degrees latitude, 0 degrees longitude, a point that GIS tools run into often. While computers might find (0,0) appealing, it is unlikely that there is much there for the typical Virtual Earth user to find interesting.¹
This second bright spot inspired a hunt for bugs. The team rapidly learned that Virtual Earth’s search facility would sometimes fail, and instead of returning an error message, typos and erroneous searches would sometimes redirect the user to (0,0). Interestingly, the bug had been on the backlog for some time because the team had decided that it was not likely to surface often. Seeing this image made it clear that some users really were being confused by the error, so the team prioritized the bug.

Although the Virtual Earth team started out using the Hotmap visualization expecting to find out about how users interacted with maps, they gleaned much more than just a characterization of usage patterns. Like many—dare we say most?—new visualizations, the most interesting insights were those that the viewers were not anticipating to find.
Where Visualization Is Useful

Is visualization the silver bullet to help us make sense of data? Not always. There are two questions to consider to help you decide if your data analysis problem is a good candidate for a visualization solution.

First, *could the analysis tasks be supported with an algorithm?* A crisp task such as “I want to know the total number of users who looked at Seattle” suggests that an algorithm, statistical test, or even a table of numbers might be the best way to answer the question. On the other hand, “How do users explore the map?” is much fuzzier. Fuzzy tasks are great candidates for a visualization solution because they require you to look at the data from different angles and perspectives, and to be able to make decisions and inferences based on your own knowledge and understanding.

The second question to consider is “Is all the necessary information contained in the dataset?” If there is information about the problem that is not in the dataset which requires an expert to interpret the data that is there, then visualization is a great solution. Going back to our fuzzy question about exploring a map, we can imagine that it is unlikely that there will be an explicit attribute in the data that classifies a user’s exploration style. Instead, answering this question requires someone to interpret other aspects of the data to bring knowledge to bear about what aspects of the data imply an exploration style. Again, visualization enables this sort of flexible and user-centric analysis.

For all but the crispest questions about explicitly measured phenomena, visualization is probably a good tool to throw at a problem. In our experience, we have almost never come up against a problem that cannot benefit from some amount of visualization.
Further Reading

The Hotmap project is discussed in:


---

1 So many datasets have references to (0,0) that GIS practitioners refer to that location as “null island.”

2 See “Further Reading” for other stories of how Hotmap has been used.
Chapter 2. From Questions to Tasks

All visualization begins with a question about data. An analyst wants to know something about a phenomenon in the world, or wants to share their knowledge about it with someone else. She believes the phenomenon they wish to examine is represented somehow in the data.

The challenge in this process is that the question the analyst wishes to address can seem far from the data. The analyst might be working on a broad goal: say, “Are high-salary employees more productive than less well-paid ones?” This leads to a process of making the question measurable. What does the analyst mean by high-salary, and productive? What visualization or set of visualizations would demonstrate the relationship between these variables?

The process of breaking down these questions into something that can actually be computed from the data is iterative, exploratory, and sometimes surprising. This chapter describes how to refine high-level questions into specific, data-driven tasks. The outcome of that process is a set of concise design requirements for a visualization tool that supports finding answers to those questions.

The general concept of refining questions into tasks appears across all of the sciences. In many fields, the process is called operationalization, and refers to the process of reducing a complex set of factors to a single metric. The field of visualization takes on that goal more broadly: rather than attempting to identify a single metric, the analyst instead tries to look more holistically across the data to get a usable, actionable answer. Arriving at that answer might involve exploring multiple attributes, and using a number of views that allow the ideas to come together. Thus, operationalization in the context of visualization is the process of identifying tasks to be performed over the dataset that are a reasonable approximation of the high-level question of interest.

A visualization is not the inevitable outcome of operationalization. Exploring the data might show that the goal is best achieved with a statistical analysis or with machine learning. Similarly, the outcome of the process might show that a
cluster analysis across multiple attributes is more useful than a plot. We find that more often than not, visualization is a vital component of getting to a successful operationalization.

This chapter emphasizes the data aspects of this process. The next chapter moves to the human side of the process: how to get the information necessary to effectively operationalize the high-level questions. Later chapters then look at how to translate the operationalized questions into specific visualizations.

**Example: Identifying Good Movie Directors**

To guide the process through operationalization, this chapter examines an exemplar question: “Who are the best movie directors?”

Nonspecific questions like this are how many data explorations start. Answering a question like this requires a much more specific task that can be precisely addressed with a dataset. Before we can be more specific, we first need to take a step back: who needs to know the answer to this question? The use case might be a film student trying to assert that his dissertation is about one of the most influential directors, or a hiring manager looking to hire a director for an upcoming project, or a journalist putting together a splashy article that will feature a top list.

Each of these users needs suggests different interpretations for the notion of *best* director. The film student is looking for a way to quantify and defend a notion of influence, whereas the hiring manager might want to limit themself to people working today who are less accomplished and thus more affordable. For this example, though, the user will be a journalist who is putting together an article about a new movie and wants to include a list of the best directors.

The goal of operationalization is to refine and clarify the question until the analyst can forge an explicit link between the data that they *can* find and the questions they *would like* to answer. For this example, the dataset at hand contains a list of movies rated by the film-aficionado community. Each movie is associated with a director, a number of raters, and an average rating score.
The dataset used for this chapter is comprised of two of IMDB’s downloadable lists, `directors.list` and `ratings.list`. There is a copy of the Jupyter notebook that parses them into cleaner CSVs on the book’s companion website. The script cleans the data to remove entries that the database refers to as not being movies, such as video games and TV shows. The analysis and visualizations in this chapter are carried out in Python and recorded in a second Jupyter notebook available at the same site.

With both data and a high-level question in hand, the visualization work can begin. Data alone is not enough to dictate a set of design requirements for constructing a visualization. What is missing here is a translation of the high-level question “Who are the best movie directors?” into a set of concrete tasks over the data.

The choice of dataset and operationalization is fundamentally a specific perspective on a problem; they stand in for what the analyst wishes to understand. In this example, there are other ways to frame the inquiry and other types of data that could be collected. This is a large part of why visualization is so important for answering questions like these: it allows an analyst’s experience and knowledge to layer directly on top of the data that is ultimately shown. The analyst’s skills and experience allow them to make inferences about the more abstract questions they are really interested in.

**Making a Question Concrete**

The process of operationalization winds its way from a general goal or a broad question to specific tasks, and then to visualizations that support those specific tasks based on concrete data.

To achieve this, the analyst searches for proxies. Proxies are partial and imperfect representations of the abstract thing that the analyst is really interested in. For example, *high movie ratings* may be a reasonable proxy for *best* in our movie example. Selecting and interpreting proxies requires judgment and expertise to assess how well, and with what sorts of limitations, they represent the abstract concept.

In operationalization, there are two important types of proxies:
• A proxy *task* is a lower-level task that stands in for the original. The result of a proxy task reflects on the answer to the original question, but the proxy task itself is more closely related to the data; it can be accomplished with quantitative tools, such as a visualization or a statistical analysis.

• A proxy *value* is an attribute in the data that stands in for a more abstract concept. This can be an existing attribute, or it can be derived from the data.

Operationalizing a question often results in more questions, which require further articulation of proxies. One step in this process is to find places where a question is underspecified or does not directly reference the data on hand, in order to identify where proxies are necessary.

Collaboration with stakeholders crucially informs the process of operationalizing questions. It helps to learn what data is available and how the results will be used. Interviews help to identify the questions and goals of the stakeholders with respect to the data and to further understand what data is available or can be made available. Throughout the process, an analyst translates questions and goals into a description of the problem that is amenable to a data solution. Interview techniques and prototyping are discussed in more detail in Chapter 3.

In this book we advocate an approach of systematic operationalization in order to bolster explicit acknowledgment, validation, and support of the range of possible proxy decisions for a question. This systematic approach leaves open future possibilities and provides guidance for making downstream decisions. The start to this process is getting to understand the question and what is available in the data—and appreciating the gaps between them.

This chapter both describes and illustrates the operationalization process. It uses the movie director example to show how to refine a question into detailed, specific tasks. It discusses the four components that we use to describe an operationalized task.

**A Concrete Movie Question**

The example started with the high-level question “Who are the best directors?” The dataset is a list of directors and a list of movies. The first task is to
operationalize *best director*. As a rough definition, a good director has directed many good movies. But *many good movies* is also ill-defined, and thus a proxy for *good movie* might in turn be based on its rating on IMDB.

These decisions replaced one bit of ambiguity with three more. How many of these *best* directors need to appear in the results? What counts as *good* IMDB ratings, and what are *many* of them? For that matter, a quick glance at the IMDB data reveals that there are short films, TV episodes, video games, and so on—so what counts as a *movie*?

It is possible to choose the measures arbitrarily: “More than five movies with IMDB ratings greater than 9.8,” or “average movie rating higher than 8.2,” or “no movies with a rating less than 5.” While it is not uncommon to make these sorts of decisions based on rough knowledge of the data, or even based on choosing nice, round numbers, looking at the actual data is important. The top-rated items on IMDB might turn out to have very high ratings but only one review. Great directors might direct a few stinkers, so just looking at the average rating might turn out to be a poor choice. The only way to learn what the data says is to start digging into it.

Choosing a proxy allows the analyst to sanity-check their decisions; it can be valuable to do this iteratively at each step, checking both the quality of the data and of the proxy.

A quick glance at the first five data items in the dataset reveals non-mainstream movies (*Table 2-1*). The alphabetical first movie in the dataset is called *#1*, with a total of 12 raters; the second is the similarly obscure *#1 Serial Killer*. Since the scenario targets a general audience, it should probably focus on movies that most people are likely to know. A different scenario could suggest very different proxies.
A quick glance at the first data items in the movie dataset (which is sorted alphabetically) reveals that there could be movies with positive ratings that have very few raters, implying an obscure (but decent) movie.

<table>
<thead>
<tr>
<th>ID</th>
<th>Raters</th>
<th>Score</th>
<th>Title</th>
<th>Director</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>12</td>
<td>6.4</td>
<td>#1 (2005)</td>
<td>Breen, James (V)</td>
</tr>
<tr>
<td>1</td>
<td>35</td>
<td>6.0</td>
<td>#1 Serial Killer (2013)</td>
<td>Yung, Stanley (I)</td>
</tr>
<tr>
<td>2</td>
<td>5</td>
<td>5.8</td>
<td>#137 (2011)</td>
<td>Elliott, Frances</td>
</tr>
<tr>
<td>3</td>
<td>11</td>
<td>7.4</td>
<td>#140Characters: A Documentary About Twitter (2…</td>
<td>Beasley, Bryan (I)</td>
</tr>
<tr>
<td>4</td>
<td>23</td>
<td>6.7</td>
<td>#30 (2013)</td>
<td>Wilde, Timothy</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

The decision to stick with mainstream movies suggests a need for a proxy for popularity. One choice could be the number of ratings for each movie. By plotting the distribution of the number of ratings by movies (Figure 2-1), we see that the vast majority of movies in the dataset actually have very few ratings.
This first plot shows that the number of ratings is heavily skewed. One way to make this distribution more interpretable is to plot it on a logarithmic scale. In Figure 2-2, the data has been bucketed; a film with 1,000 ratings now appears in the bucket for $\log_{10}(1000) = 3$. Taking the log of the number of ratings smooths the distribution, more effectively showing its shape.
Distribution of the logarithm (base 10) of the number of ratings. The peak is under 2: most films have under 100 ratings.

We can also compute some basic summarizing statistics about the number of ratings: the median movie in the dataset has just 26 ratings while the 75th percentile is at 132 ratings. By looking up the number of ratings for a sample of blockbusters, we note that movies that anyone can name offhand have tens of thousands of ratings. These are useful observations; perhaps it would be valuable to trim to a slimmer set of movies to ensure that most are ones that a reasonable number of people have seen.

We want to choose a number, though, that’s fair to good movies, even if they are not very popular—in this case, we pick, somewhat arbitrarily, the most-rated 25% of movies. This amounts to around 70,000 films with more than 132 ratings.

We next pivot and look at the distribution of ratings for the slimmed-down set of movies, shown in Figure 2-3. This distribution shows a distinct curve with a clear peak and noticeable drop-off: ratings above 7.5 seem different from lower ratings. (This distribution has a median score of 6.6, and a 75th percentile of 7.4.)
Step back from our dive into the data, we can observe that we have proceeded some distance along the operationalization. We have defined a good director and decided that it is based on their movies; we have focused on movies and chosen a set that are popular enough to be part of the analysis. But there are still unanswered questions: How will we rank directors against each other? What makes for a “best” director?

A systematic approach to operationalization allows an analyst to see the full range of decisions and helps in pulling together the set of proxies that can inform a final answer. Ultimately, an interactive visualization tool can enable exploration of multiple proxies to allow for a set of justified, and validated, answers. For our running example, we will continue with the operationalization after describing a framework for making decisions explicit throughout the process.

**Breaking Down a Task**
Throughout the operationalization, we need to identify where in a question or task there is a need for a more refined proxy. Doing so systematically can make it easier to validate those decisions, as well as to produce a road map of the process. This allows the analyst to effectively revisit decisions once a better understanding of the problem is gained.

An analyst can refine a task by first breaking it down into four specific components. Identifying these components and how they do or do not directly reference the data becomes a template for choosing more specific tasks. The components are:

**Objects**

Things or events that exist in the world: in our example, a director and a movie are both objects. In other contexts, objects might be a user or a sale of a single item. When a task is specific enough, each object will be something that can be represented in or computed from, the data. Fairly often, when the task is at its most specific, an object will correspond to a single row in a database.

**Measures**

The outcome variables that will be measured for the objects. Quality of a director, happiness of a user, and sales of a store are all measures. In a sufficiently specific task, the measure is either an existing attribute in the dataset or one that can be directly computed from the data. A measure is sometimes aggregated across many items of data. In our example, a number of movies are aggregated together to get a score for a single director.

**Groupings (or partitions)**

Attributes or characteristics of the data that separate the data items into groups. For example, groupings might include store region (western versus eastern), start date of players, whether users have purchased an upgrade, or sales by year. In a specific task, partitions are attributes of the objects or can be calculated directly from those attributes. When the visualization is created, partitions will often manifest as groupings, separations across charts, or filters.

**Actions**
Words that articulate the specific thing being done with the data, such as *compare, identify, characterize*, etc. Actions guide the process of choosing appropriate visualizations.

The action is useful for identifying the other components. Take this task: 
*Compare the amount of money spent in-game by players who play more hours versus those who play fewer hours.* The action is *compare*. What is compared? The *players* (the object). What is it about players that we want to compare? The *money spent* (the measure). Finally, there is a specific partition on the objects. They will be broken into two groups: those that play many hours and those that play few hours.

The following components are the heart of an iterative process:

1. Refine the question into one or more tasks that, individually or together, address the general question.

2. For each task:
   
   a. Identify the components of the task.

   b. Look for ambiguous components—namely, components that are not directly addressable by the dataset.

   c. For each ambiguous component, define a proxy by creating a new question that addresses the component, and return to step 1 with those questions.

   d. If there are no ambiguous components then the task is deemed actionable, and thus can be addressed with a visualization or other computational technique.

Next, we’ll explicate some of the questions from the movie example to illustrate how the components work in practice, beginning with Example 2-1.
Example 2-1. Breaking down the task to find good directors

**Task:** Identify the top directors who have directed many good, popular movies

**Action:** Identify

**Object:** Director

**Measure:** Number of good, popular movies

**Grouping:** Filter out non-movies

Identifying top directors implies that there’s a meaningful sort on the directors so that the top can be found (Example 2-2). Thus, we can further refine the action to specify an ordering. Also, our first look at the data showed that many movies are unpopular, which implied a grouping to filter out unpopular movies.

Example 2-2. Refined task for good directors

**Task:** Rank order of directors by those who have directed many good, popular movies

**Action:** Rank order

**Object:** Director

**Measure:** Number of good movies

**Grouping:** Filter out non-movies and unpopular movies

Filtering out unpopular movies is a subtask (Example 2-3), which we addressed with a histogram of the number of ratings for movies. The visualization of the distribution allowed us to determine a good cut point for popular versus unpopular—namely, popular movies were those in the top 25% of movies with the highest number of ratings.
Example 2-3. Subtask for filtering unpopular movies

**Task:** Filter out movies with significantly fewer ratings

**Action:** Filter

**Object:** Movie

**Measure:** Number of ratings

**Grouping:** Separate into most popular and least popular movies

This subtask can be brought back into Example 2-2 as a proxy for unpopular movies.

However, we still have some work to do on Example 2-2: the measure *number of good movies* is ill-defined with respect to the data. We need to refine this component by developing a proxy for a *good* movie. Once we do that, we can then examine what it means to have *directed many* of them.

These proxies require further elaboration. How *many* high-scoring movies are required from directors? Do low-scoring movies count against them? This process of identifying reasonable proxies is often iterative. For example, in exploring and validating a proxy with the data, it might become obvious that the effects of filtering by the number of ratings was a mistaken approach.

At this point, we can recognize that we need a proxy measure for *good* (Example 2-4). There are a variety of proxies that we can try here, with various visualizations. The process continues onward.

Example 2-4. Subtask for good movies

**Task:** Quantify “a good movie”

**Action:** Quantify

**Object:** Movie

**Measure:** Goodness
Breaking down a task into components helps in guiding refinement of a task into one that can be addressed with the data. The most direct way to do so is to consider the question “Are the object, measure, and grouping each directly described in the data?” For each of these three components, is it clear which aspects of the data are important or how to derive what we need from the data? If not, repeat the process of formulating a subquestion in order to derive a more specific answer.

Let’s take a look at a very different example—this time, from a gameplay metrics scenario (Example 2-5).

**Example 2-5. Exemplar task for analyzing a game**

**Task:** Compare the amount of money spent in-game by players who play more hours versus those who play fewer hours.

**Action:** Compare

**Object:** Players

**Measure:** Money spent

**Grouping:** Players who play many hours; players who play few hours

In Example 2-5, the partition divides between *many* and *few* hours. This component needs to be refined further, which leads to a new question: “In the game, how many is ‘many’ hours for a player?” The analyst might take a series of steps. They might look at the distribution of hours played, or they might choose to filter out players who have played zero hour or those who haven’t made it past the tutorial, or they might look at other metrics that are important to the game. These steps would help the analyst figure out good proxies for *many* and *few* hours.

**When Tasks Lead to New Questions**

There are four broad categories of new lines of inquiry that can emerge from
refining a question. First, as in the movie example, the refinement process often reveals that a new analysis is needed to answer these questions.

Second, operationalizing can also lead in new directions. In the process of exploring who the best directors are, the analyst might notice that some directors stick to a single genre; they might decide that this analysis might be interesting divided across multiple genres. They might also notice that both IMDB and Rotten Tomatoes have scores on movies, and want to see how these results vary based on Rotten Tomatoes scores instead of IMDB.

Third, the data itself can lead to new questions too. In exploratory data analysis (EDA), for example, the data analyst discovers new questions based on the data. The process of looking at the data to address some of these questions generates incidental visualizations—odd patterns, outliers, or surprising correlations that are worth looking into further.

Finally, doing some analysis often leads to doing a round of data cleaning. While data cleaning is largely out of the scope of this book, odd outliers and surprising trends are, as often as not, the result of dirty data.

**Returning to the Example: Exploring Different Definitions**

There are several different possible definitions of best director.

Here is one: the best director has the most movies with more than 134 ratings. Table 2-2 shows the top scorers. The most prolific directors in our dataset are Chuck Jones and Fritz Feleng (who directed classic Looney Tunes animations), William Hanna (who directed *Tom and Jerry* and other classic Hanna-Barbera cartoons), and George Méliès (an early inventor of special effects and shorts).

<table>
<thead>
<tr>
<th>Director</th>
<th>Avg. raters</th>
<th>Avg. score</th>
<th>Count</th>
<th>Total raters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jones, Chuck (I)</td>
<td>719</td>
<td>7.4</td>
<td>148</td>
<td>106,397</td>
</tr>
<tr>
<td>Freleng, Fritz</td>
<td>402</td>
<td>7.2</td>
<td>141</td>
<td>56,730</td>
</tr>
</tbody>
</table>

*Table 2-2. Top five directors by number of films over threshold*
Georges Méliès has 526 films on his IMDB page; only 114, however, made it over the threshold of raters. The huge number of films is explained by the fact that the films are shorts—more familiar on television now, but once also shown in theaters. This should be an opportunity to do more data cleaning to join in another table that will tell us whether a film is a short or not, and filter those out. IMDB has a film duration data table; in a typical analysis process, the next step would be to merge in this table, adding a new proxy for what makes for a short film.

We might explore other definitions of best directors. For example, the best directors might make the movies that people want to rate the most. Table 2-3 is a list of the directors whose movies have, in total, the most ratings.
the threshold: a Mongolian movie from 2016 with 624 raters and an average score of 9.7. This measure of popularity returns a very different set of results than the previous measure: ten thousand times more people rated Quentin Tarantino’s movies than Uranchimeg Urtnasan’s work.
Table 2-4. Top five directors by average score

<table>
<thead>
<tr>
<th>Director</th>
<th>Avg. raters</th>
<th>Avg. score</th>
<th>Count</th>
<th>Total raters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Urtnasan, Uranchimeg</td>
<td>624</td>
<td>9.7</td>
<td>1</td>
<td>624</td>
</tr>
<tr>
<td>Miller, George (XXXVII)</td>
<td>394</td>
<td>9.6</td>
<td>2</td>
<td>787</td>
</tr>
<tr>
<td>Chowdhury, Amitabh Reza</td>
<td>14,628</td>
<td>96</td>
<td>1</td>
<td>14,628</td>
</tr>
<tr>
<td>Biebert, Aaron</td>
<td>12,040</td>
<td>9.6</td>
<td>1</td>
<td>1,204</td>
</tr>
<tr>
<td>Arsyn, Ken</td>
<td>619</td>
<td>9.5</td>
<td>6</td>
<td>3,712</td>
</tr>
</tbody>
</table>

But can the quality of a director be measured based on just one or two movies? Each step of data exploration leads to another step of refining the question. Is it more important to have many raters, a high average score, or a high minimum score?

The choice of metrics leads to very different outcomes. A slight tweak determines whether you find directors of animated cartoons, blockbuster directors, or a very diverse set of international directors.

How Specific Does the Process Get?

This process of refinement leads to a scary scenario. In Disney’s *Fantasia*, in the Sorcerer’s Apprentice sequence, Mickey Mouse attempts to stop an enchanted broom by chopping it in half and instead produces two half-size enchanted brooms. Will our analysis subtasks forever multiply?

The operationalization process is an iterative one and the end point is not precisely defined. The answer to the question of how far to go is, simply, far enough. The process is done when the task is directly actionable, using the data at hand. The analyst knows how to describe the objects, measures, and groupings in terms of the data—where to find it, how to compute, and how to aggregate it. At this point, they know what the question will look like and they know what they can do to get the answer.

An actionable task means that it is possible to act on its result. That action might
be to present a useful result to a decision maker or to proceed to a next step in a different result. An answer is actionable when it no longer needs further work to make sense of it.

Low-level objects are ready to be interpreted from the data. Sometimes they can be read directly off the data table, but more often it is more indirect; the analyst may need to carry out transformations on the data, whether mathematical transformations or database joins. For instance, in the movie example, the object is the director; the proxy for the director is the result of aggregating multiple movies together. Partitions and measures at the lowest level will resolve to concrete manipulations of the objects.

The process ends when all the tools needed to answer a question are in place—whether as a number, a visualization, or even as an interaction across multiple visualizations representing multiple proxies. The analyst might decide that the right cutoff for many hours of gameplay is six hours—a number—or the hours played by the top 10% of players—a formula—or above the logical breakpoint, which might be represented by a distribution. These results get propagated back into any other tasks that depend on them.

**Making Use of Results**

This process of propagating results back into higher-level questions is flexible. Sometimes the low-level question does not have an exact answer but instead resolves in its own visualization or interaction. That visualization might help an analyst in making a decision, but it might also imply parameters on the data. For example, the journalist might realize that there are several possible cutoffs for defining what it means to be a good movie. Rather than simply picking a specific threshold, an analyst might instead propagate a mechanism for dynamically determining cutoffs into higher-level tasks. Seeing a variable propagated like this can be a cue that an interactive visualization—rather than a static image—might be helpful.

Visualization is also important for supporting the operationalization process, even if the end result is not an interactive visualization. In the movie example, visualization helped us to understand the nature and distribution of the data. Visualization can be more prominent with more complex analysis tasks. If the
analyst wanted to compare ratings against popularity, it would be difficult to display that on a list; if they wanted to explore hypotheses about how the popularity of directors changes over time, more visual representations would help them explore the data.

**Conclusion: A Well-Operationalized Task**

A well-operationalized task, relative to the underlying data, fulfills the following criteria:

- Can be computed based on the data
- Makes specific reference to the attributes of the data
- Has a traceable path from the high-level abstract questions to a set of concrete, actionable tasks

A well-operationalized task is a first step toward creating a visualization. Chapter 4 begins to describe the ways in which the objects, measures, and partitions can be shaped into aspects of a visualization. Chapters 5 and 6 construct visualizations based on them.

Written out in detail, this process can seem tedious, but in practice, it is abbreviated and simplified. There are two important uses for this systematization. First, the process of explicitly looking at components can help untangle knotty problems, decomposing places where the analyst has made assumptions about the data. Explaining precisely why the number of IMDB ratings is a proxy for popularity forces the analyst to explore whether it is a good choice—and, perhaps, to revise that choice later.

The process also helps guide questions and interviews. Chapter 3 explains how to carry out operationalization with domain experts. Recognizing the need to make decisions about proxies helps guide these conversations. Every dataset has subtleties; it can be far too easy to slip down rabbit holes of complications. Being systematic about the operationalization can help focus our conversations with experts, only introducing complications when needed.
Further Reading

The process outlined here is similar—and in many ways parallel—to the Goal, Question, Metric (GQM) process found in the software engineering space. GQM refines from a general goal to a specific metric, usually oriented around process improvement so that the consumer can have a single number that helps them know whether they are succeeding in improving that process.

Our process is more exploratory and often comes earlier in the cycle. A GQM analysis might choose a goal like “improve user retention.” In contrast, exploratory operationalization might start with a question like “Do users come back to our site?” with the awareness that the problem is multifaceted and complex, and might require a variety of different metrics to describe. For more on GQM, see:


The data visualization field has spent a great deal of effort trying to understand the tasks that can be accomplished in a visualization. Amar and Stasko, for example, explore a low-level analysis of tasks carried out on a specific visualization. At the other end of the spectrum, Brehmer and Munzner explore high-level tasks for visualization, starting with comparing presentation and exploration:


1 Median and percentile are ways of characterizing a distribution of numbers. If one were to sort the numbers, the 75th percentile would be 75% of the way down
in the list. The median would be at the halfway point.