p(Persuasion | data, analysis, storytelling)

Scott Spencer

Last updated, 2020 March 07

Contents

Preface 7

1 *Introduction* 9

1.1 Communication throughout project progression 10

1.2 Building your persuasive portfolio 10

1.3 Receiving and providing criticism 11

2 *Communication for applied analytics, challenges and scope* 13

2.1 Introduction 13

2.2 Communication in data analytics 13

2.3 Communication begins with content 14

2.4 Workflow for credible communication 20

3 *An example, reproducible workflow* 23

3.1 Background and problem 23

3.2 The goal 24

3.3 The data 24

3.4 Modeling count data: Poisson distribution 26

3.5 Expanding the model: multiple predictors 33

3.6 Modeling count data with the negative binomial distribution 38

3.7 Hierarchical modeling 43

3.8 Time varying effects and structured priors 55

3.9 Using our model for decisions: Cost forecasts 60

3.10 Next steps 63

4 Purpose, audience, and craft 65
4.1 Communication structure 66
4.2 Narrative structure 73
4.3 Sentence structure 74
4.4 Layering and heirarchy 75
4.5 Audiences and purposes 76
4.6 Multiple or mixed audiences 79
4.7 Story 81
4.8 The importance of revision 83
4.9 Example memos 84

5 Persuasion and biases 87
5.1 Methods of persuasion 88
5.2 Statistical persuasion 94
5.3 Comparison through two numeric languages 98
5.4 Statistics and narrative 98
5.5 Comparison through metaphor, simile, analogy 99
5.6 Patterns that compare, organize, grab attention 100
5.7 Le mot juste — *the exact word* 102
5.8 Heuristics and biases 103
5.9 Brief proposals 105

6 Layout, hierarchy, and integration 109
6.1 Visual presentation is communication 109
6.2 Combined meaning of words and images 112
6.3 Visually integrating graphics and text 113

7 Visual design and perception 117
7.1 Why review data graphically? 117
7.2 Reasoning with images 119
7.3 Components of a graphic 119
7.4 Perceptions of visual data encodings 123
7.5 Maximize information in visual displays 130

8 Visually encoding data, common and xenographic 133

8.1 Encoding data-ink, common graphics 133

8.2 Graphics, layers and separation 134

8.3 Encoding data-ink, xenographics 135

8.4 Misleading encodings 139

9 Context, uncertainty, estimation, and prediction 141

9.1 Context 141

9.2 Uncertainty 141

9.3 Estimations and predictions from models 144

10 Utility and audience decisions 149

10.1 Explaining the so what—from models to decisions 149

11 Dynamic audience views of data, analyses, and story 151

11.1 Enabling audience exploration 151

About the author 153

Bibliography 155

Preface

What's the probability of persuasion given objectives, audience, data, analysis, design, and storytelling? In this book, I aim to explore the question. The literature of writing, and of persuasive communication is rarely taught alongside the literature for collecting, cleaning, visualizing, modeling, understanding, and communicating quantitative information. Pedagogy of persuasive communication in data science literature generally lacks the rigor of that taught elsewhere. I aim to provide coherency across various literature, and invite us to think, consider the perspectives of those authors, and question their opinions. I also encourage readers to go further, to be curious about what more they may learn from the cited references (I chose them with purpose). In short, I have *active learners* in mind:

> **An active learner** asks questions, considers alternatives, questions assumptions, and even questions the trustworthiness of the author or speaker. An active learner tries to generalize specific examples, and devise specific examples for generalities.
>
> An active learner doesn't passively sponge up information — that doesn't work! — but **uses the readings and lecturer's argument as a springboard for critical thought and deep understanding**.

To encourage adventure into ideas, I've included rabbit holes, like so:

> *Rabbit Hole.* Going *down the rabbit hole* is a phrase borrowed from *Alice's Adventures in Wonderland*[1], a story of adventure and exploration of new, interesting, and seemingly strange worlds. In these remarks, I encourage you to explore as Alice did.

[1] Lewis Carroll, *Alice's Adventures in Wonderland and Other Stories* (Canterbury Classics, 2013).

To help the active learner, I've provided examples and exercises throughout this text, like this:

> **Exercise 0.1** (An example exercise callout). In Lewis Carroll's beloved work, Alice encounters numerous characters during her adventure in wonderland. Does she listen to advice? Does she question advice?

1
Introduction

If storytelling seems out of place in this study, consider that the truly unique feature of human language is "the ability to transmit information about things that *do not exist* at all…"[1] Just as *Alice's Wonderland* does not exist, so too with data and algorithms — we cannot taste them, smell them, or touch them. They are mere concepts that represent things humans care about. Stories add a valuable communication tool to get others to act on what we learn from data. Storytelling with data is especially challenging. Maybe you've heard the advice, *write what you know*. You should. Commonly, data analytics projects require multiple skills and ideas, skills in creative problem solving and ideation, mathematics, probability, statistics, programming, subject matter or domain knowledge, and across all these, the ability to communicate well to all interested or involved.

We need to know all aspects of our current project well enough to explain them — to explain them well enough to persuade our audiences of our insights, and to act on them. And it is the goal of persuasion that proves most difficult: Harari writes, "the difficulty lies not in telling the story, but in convincing everyone else to believe it…" You will meet this challenge by studying well-crafted communication, *and practicing*! Indeed,

> Learners need to practice, to imitate well, to be highly motivated, and to have the ability to see likenesses between dissimilar things in [domains ranging from creative writing to mathematics].[2]

Along the way, I will guide you to additional resources, beyond the minimum. It's up to you to study them and share what you learn with your peers. And you should expect to learn from your peers because they will have studied ideas from these readings, as have you, and will bring their own understanding, opinions, and experiences with the material to enrich our discussions.

[1] Yuval Noah Harari, *Sapiens: A Brief History of Humankind* (London: Harvill Secker, 2014).

[2] Berys Gaut, "Educating for Creativity," in *The Philosophy of Creativity: New Essays*, ed. Elliot Samuel Paul (New York: Oxford University Press, 2014), 265–87.

1.1 *Communication throughout project progression*

With what projects will we apply communication skills?

In the course intended with this book, each of you will have your own project. Your project requires you to identify a company or organization in which you are interested. It requires you to identify a **problem** that entity faces, or an **opportunity** they may not have fully explored. The project requires you to locate **publicly available data**. Soon you will see exemplary data. You will **analyze that data**, through statistical summaries, visualizations, and maybe even some modeling. And you will **communicate your insights** using sentences — *messages* — and graphics in multiple forms, with varying goals, and **to different audiences**.

> **Exercise 1.1** (Research possible projects)**.** Research potential projects interesting to you. Identify a problem that entity faces, or an opportunity they may not have fully explored. Locate publicly available data. Consider for what questions the data are potentially relevent to the problem or opportunity.

You will use your deliverables to practice what we learn. These include, but are not limited to, a **memo** and **proposal**, written to the head of analytics; a **communication**, created for the head of marketing; an **infographic**, to inform a more general audience; a **presentation**, to persuade the chief executive officer to further invest in analytics. We will consider examples of these. Each of these deliverables is an opportunity to apply various forms of, and ideas in, communication, from the structure, meaning, and persuasive qualities of words and sentences, to visual organization, data encodings, and design of information. Here's a conceptual project timeline:

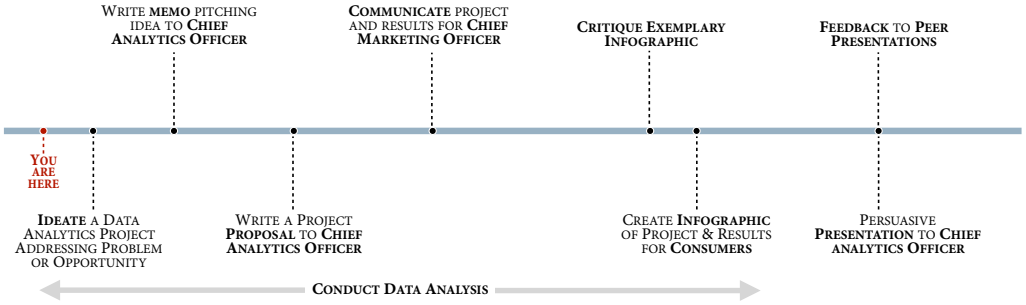


Figure 1.1: Conceptual project timeline, from ideating a project through analysis and completion that provides the basis for various communications.

1.2 *Building your persuasive portfolio*

This project progression should enable students to develop a portfolio of persuasive communication. The importance of such

a portfolio is well understood in various fields of study. "As your ultimate persuasive tool, your portfolio is the single most important design exercise of your academic and professional career."[3] If you have not already begun, start now. With this book.

[3] Margaret Fletcher, *Constructing the Persuasive Portfolio: The Only Primer You'll Ever Need* (New York: Routledge, Taylor & Francis Group, 2017).

1.3 *Receiving and providing criticism*

Criticism — the analysis of the merits and weaknesses of a work supported by well-developed theories — is vital for our improvement. In a course imagined for this material, your colleagues and I will try to help you with your *draft* work. They will do this by learning alongside you and practice providing criticism of your work, not of you. Criticism, here, is not to be understood as negative. Instead, it should be understood as *an opportunity for you **and** your colleague*.

Your colleague, in reviewing your draft work, will practice applying the concepts we've learned in class, and bring their own experiences into their review of your draft work. It is a time for them to practice, in that sense. If it is difficult to well apply the concepts we discuss to one's own work, it is yet more difficult to teach others through criticism. You will do this too — *you are them*. Meaning, you will be someone else's critic that helps them. I will be guiding each of you in learning to provide criticism.

On the flip side, you will gain most valuable *fresh eyes* and perspective to improve your drafts. Your critics will undoubtedly see problems you didn't, and consider solutions you haven't. When you have wowed your next client or employer with the persuasive portfolio you have developed, consider reaching out to thank your colleagues for their criticism. And pay it forward.

Figure 1.2: Jean-luc Doumont is an engineer from the Louvain School of Engineering. He earned his PhD in applied physics from Stanford University. He wrote his book to help engineers, scientists, and managers with business communication. The book succeeds both in its instruction by as exemplary communication.

Doumont[4] provides thoughtful advice on how to approach criticism. When reviewing someone else's document, center yourself on the purpose you both agree upon, such as clarity, accuracy, or correctness. Should this purpose be multiple, review one aspect at a time, focusing on content first. As a critique, focus less on typographic errors. *Typos* are usually more conspicuous than reasoning flaws, but also less important. As a critique, in your comments to the authors, strive to help, not to judge. Finally, structure the review. Provide a global assessment, to place further comments in proper perspective. As a rule, point out the weaknesses, to prompt improvements, but also the strengths, to increase the authors' willingness to revise the document and to learn. ***Your criticism should be in the form of applying the language and concepts we study***.

[4] Jean-Luc Doumont, *Trees, Maps, and Theorems*, Effective Communication for Rational Minds (Principiæ, 2009), *Reviewing documents of others*.

2

Communication for applied analytics, challenges and scope

2.1 Introduction

A noted psychologist once said, "One cannot *not* communicate."[1] We will consider how Watzlawick's concept reveals itself in all aspects of applied analytics, especially in today's collaborative environments, but even so when working solo: after all, your future self is also your audience. We plan to explore and test this idea throughout the course, throughout the life of your data analytics project.

[1] Paul Watzlawick, Janet Beavin Bavelas, and Don D Jackson, *Pragmatics of Human Communication: A Study of Interactional Patterns, Pathologies and Paradoxes* (W. W. Norton & Company, 2017).

2.2 Communication in data analytics

The qualities we need in an analytics team, writes Berinato[2], include project management, data wrangling, data analysis, subject expertise, design, and storytelling. For that team to create value, they must first ask smart questions, wrangle the relevant data, and uncover insights. Second, the team must figure out — and communicate — what those insights mean for the business.

These communications can be challenging, however, as an *interpretation gap* frequently exists between data scientists and the executive decision makers they support.[3]

How can we address such a gap?

Brady and his co-authors argue that *data translators* should bridge the gap, address data hubris and decision-making biases, and find linguistic common ground. Subject-matter experts should be taught the quantitative skills to bridge the gap because, they continue, it is easier to teach quantitative theory than practical, business experience.

Before delving into the above arguments, let's first consider from what perspective we're reading. Both perspectives are written for business executives, Berinato writes in the Harvard Business Review, Brady and his co-authors write from MIT Sloan Management Review.

[2] Scott Berinato, "Data Science & the Art of Persuasion," *Harvard Business Review*, December 2018, 1–13.

Figure 2.1: Scott Berinato is senior editor at Harvard Business Review.

[3] Chris Brady, Mike Forde, and Simon Chadwick, "Why Your Company Needs Data Translators," *MIT Sloan Management Review*, March 2017, 1–6.

Figure 2.2: Brady is a professor and consultant focusing on sports management.

According to HBR, their "readers have power, influence, and potential. They are senior business strategists who have achieved success and continue to strive for more. Independent thinkers who embrace new ideas. Rising stars who are aiming for the top."[4] Similarly, MIT Sloan Management Review reports their audience: "37% of MIT SMR readers work in top management, while 72% confirm that MIT SMR generates a conversation with friends or colleagues."[5] Further, all authors are in senior management. Berinato is senior editor. Brady and co-authors are consultants focusing on sports management. Why might it be important we know both an author's background and their intended audience?

[4] "HBR Advertising and Sales," *Harvard Business Review* (https://hbr.org/hbr-advertising-sales, n.d.).

[5] "Print Advertising Opportunities," Business, *MIT Sloan Management Review* (https://sloanreview.mit.edu/advertise/print/, 2020).

Perhaps it is not surprising for a senior executive to conclude that it would be easier to teach data science skills to a business expert than to teach the subject of a business or field to those already skilled in data science. Is this generally true? Might the background of a data translator depend upon the type of business or type of data science? Is it appropriate for this data translator be an individual? Berinato argues that data science work requires a team. Might the responsibility of a data translator be shared?

Bridging the gap requires **developing a common language**. Senior management do not all speak the language of analysts. Decision makers seek clear ways to receive complex insights. Plain language, aided by visuals, allow easier absorption of the meaning of data. Along with common language, data translators should foster better communication habits. **Begin with questions**, not assertions. Then, use **analogies and anecdotes** that resonate with decision makers. Finally, whomever fills this role, they must hone their skills, skills that include business and analytics knowledge, but also must learn to **speak the truth**, be constantly curious to learn, craft accessible questions and answers, keep high standards and attention to detail, be self-starters.

2.3 *Communication begins with content*

If a challenge in communicating is to *write what you know*, in the context of data analytics, we must understand what we mean by such a project. We begin with data.

2.3.1 *Data*

Implied in the phrase *data analytics project*, we need data for such a project. What, then, are data? Let's consider what Kelleher[6] has to say in the aptly titled chapter, *What are data, and what is a data set?*

[6] John D Kelleher and Brendan Tierney, *Data Science* (MIT Press, 2018).

Consider these definitions:

> **datum** : an abstraction of a real-world entity (person, object, or event). The terms variable, feature, and attribute are often used interchangeably to denote an individual abstraction.

Data are the plural of datum.

> **data set** : consists of the data relating to a collection of entities, with each entity described in terms of a set of attributes. In its most basic form, a data set is organized in an $n \cdot m$ data matrix called the analytics record, where n is the number of entities (rows) and m is the number of attributes (columns).

Data may be of different types, including **nominal**, **ordinal**, and **numeric.** These have subtypes as well. Nominal types are names for categories, classes, or states of things. Ordinal types are similar to nominal types, except that it is possible to rank or order categories of an ordinal type. Numeric types are measurable quantities we can represent using integer or real values. Numeric types can be measured on an **interval scale** or a **ratio scale**. The data attribute type is important as it affects our choice of analyses and visualizations.

Data can also be **structured** (like a table) or **unstructured** (like the words in this document). And data may be in a **raw** form such as an original count or measurement, or it may be **derived**, such as an average of multiple measurements. Normally, the real value of a data analytics project is in using statistics or modeling to derive one or more attributes that provide insight into a problem.

Finally, existing data originally for one purpose may be used in an **observational study**, or we may conduct **controlled experiments** to generate data.

2.3.2 Context

Data measurements never reveal all aspects relevant to their generation or impact upon our analysis[7]. Loukissas provides several interesting examples where the local information that generated the data matters greatly in whether we can fully understand the recorded, or measured data. His examples include plant data in an arboretum, artifact data in a museum, collection data at a library, information in the news as data, and real estate data. Using these examples, he convincingly argues we need to shift our thinking from data sets to *data settings*.

[7] Yanni A. Loukissas, *All Data Are Local: Thinking Critically in a Data-Driven Society* (Cambridge, Massachusetts: The MIT Press, 2019).

Let's consider another example, from baseball. In the game, a batter that hits the pitched ball over the outfield fence between the foul poles scores for his team — he hits a home run. But a batter's home

run count in a season does not tell us the whole story of their ability to hit home runs. Let's consider *some* of the context in which a home run occurs. Batters hit a home run pitched by a specific pitcher, in a specific stadium, in specific weather conditions. All of these circumstances contribute to the existence of a home run event, but that context isn't typically considered. Sometimes partly, rarely completely.

Perhaps obviously, all pitchers have different abilities to pitch a ball in a way that affects a batter's ability to hit the ball. Let's leave that aside for the moment, and consider more concrete context.

In Major League Baseball there are 30 teams, each with its own stadium. But each stadium's playing field is differently sized than the others, and each stadium's outfield fence has uneven heights and is different than other stadium fences! This context is made clear in an award-winning visualization[8]. So we can't fully appreciate any specific home run event without knowing where it occurred.

Further, the trajectory of a hit baseball depends heavily on characteristics of the air, including density and wind speed and direction[9]. The ball will not travel as far in cold, humid, dense air. And density depends on temperature, altitude, and humidity. A few stadiums have a roof with conditioned air protected somewhat from weather. But most are exposed. Thus, we would learn more about the qualities of a particular home run if understood in the context of these data.

Other aspects of this game are equally context-dependent. Consider each recorded *ball* or *strike*, an event made by the umpire when the batter does not swing at the ball. The umpire's call is intended to describe location of the ball as it crosses home plate. But error exists in that measurement. It depends on human perception. We have independent measurements by a radar system (as of 2008). But that too has measurement error we can't ignore. Firstly, there are 30 separate radar systems, one for each stadium. Secondly, those systems require periodic calibration. And calibration requires, again, human intervention. Moreover, the original radar systems installed in these stadiums in 2007 are no longer used. Different systems have been installed and used in their place. Thus, to fully understand the historical location of each pitched baseball means we must research and investigate these systems.

So when we really want to understand an event and compare among events (comparison is crucial for meaning), context matters. We've seen this in the baseball example, and in Loukissas's several fascinating case study examples in many types of data. When we communicate about data, we should consider context, *data settings*.

[8] Sam Vickars, "The Irregular Outfields of Baseball," Business, *The Data Face*, April 2019, winner Kantar Information is Beautiful Awards 2019.

[9] Robert K Adair, *The Physics of Baseball*, Third (HarperCollins, 2017).

2.3.3 *Scoping a data analytics project*

To communicate about a data analytics project, we first must understand the scope and breadth of this type of project before we communicate of it. On a high-level, it involves an iterative progression of the identification and understanding of decisions, goals and actions, methods of analysis, and data.

The framework of identifying goals and actions, and following with information and techniques gives us a structure not unlike having the outline of a story, beginning with why we are working on a problem and ending with how we expect to solve it. Just as stories sometimes evolve when retold, our ideas and structure of the problem may shift as we progress on the project. But like the well-posed story, once we have a well-scoped project, we should be able to discuss or write about its arc — purpose, problem, analysis and solution — in relevant detail specific to our audience.

Specificity in framing and answering basic questions is important: *What problem is to be solved? Is it important? Does it have impact? Do data play a role in solving the problem? Are the right data available? Is the organization ready to tackle the problem and take actions from insights?* These are the initial questions of a data analytics project. Project successes inevitably depend on our specificity of answers. Be specific.

2.3.4 *Defining goals, actions, and problems*

Identifying a specific problem is the first step in any project. And a well-defined problem illuminates its importance and impact. The problem should be solvable with identified resources. If the problem seems unsolvable, try focusing on one or more aspects of the problem. Think in terms of goals, actions, data, and analysis. Our objective is to take the outcome we want to achieve and turn it into a **measurable** and **optimizable** goal.

Consider what actions can be taken to achieve the identified goal. Such actions usually need to be specific. A well-specified project ideally has a set of actions that the organization is taking — or can take — that can now be better informed through data science. While improving on existing actions is a good general starting point in defining a project, the scope does not need to be so limited. New actions may be defined too. Conversely, if the problem stated and anticipated analyses does not inform an action, it is usually not helpful in achieving organizational goals. To optimize our goal, we need to define the **expected utility** of each possible action.

2.3.5 *Identifying accessible data*

Do data play a role in solving the problem? Before a project can move forward, data must be both accessible and relevant to the problem. Consider what variables each data source contributes. While some data are publicly available, other data are privately owned and permission becomes a prerequisite. To be sure, obtaining the right data is usually a top challenge: sometimes the variable is unmeasured or not recorded.

In cataloging the data, be specific. Identify where data are stored and in what form. Are data recorded on paper or electronically, such as in a database or on a website? Are the data structured — such as a `CSV` file — or unstructured, like comments on a twitter feed? Provenance is important[10]: how were the data recorded? By a human or by an instrument?

What quality are the data[11]? Measurement error? Are observations missing? How frequently is it collected? Is it available historically, or only in real-time? Do the data have documentation describing what it represents? These are but a few questions whose answers may impact your project or approach. By extension, it affects what and how you communicate.

[10] Luc Moreau et al., "The Provenance of Electronic Data," *Communications of the ACM* 51, no. 4 (April 2008): 52–58.

[11] Wenfei Fan, "Data Quality: From Theory to Practice," *SIGMOD Record* 44, no. 3 (September 2015): 7–18.

2.3.6 *Identifying the analyses and tools*

The workflow needed to bridge the gap between raw data and actions typically involves an iterative process of exploratory and confirmatory analysis[12], see Figure 2.3, which employs visualization, transformation, modeling, and testing.

[12] Xiaoying Pu and Matthew Kay, "The Garden of Forking Paths in Visualization: A Design Space for Reliable Exploratory Visual Analytics," in *BELIV Workshop 2018*, 2018, 1–9.

2.3.7 *Estimating constraints and finances*

Can the identified project be completed within constraints in time to support the relevant actions and decisions?

2.3.8 *Writing to clarify and communicate*

Writing is part and parcel to the analysis.

> I write entirely to find out what I'm thinking, what I'm looking at, what I see, and what it means.
>
> — Joan Didion, *What I Write*

We generally revise our written words and refine our thoughts together; the improvements made in our thinking and improvements made in our writing reinforce each other.[13] Clear writing signals

[13] Joshua Schimel, *Writing Science: How to Write Papers That Get Cited and Proposals That Get Funded* (Oxford ; New York: Oxford University Press, 2012).

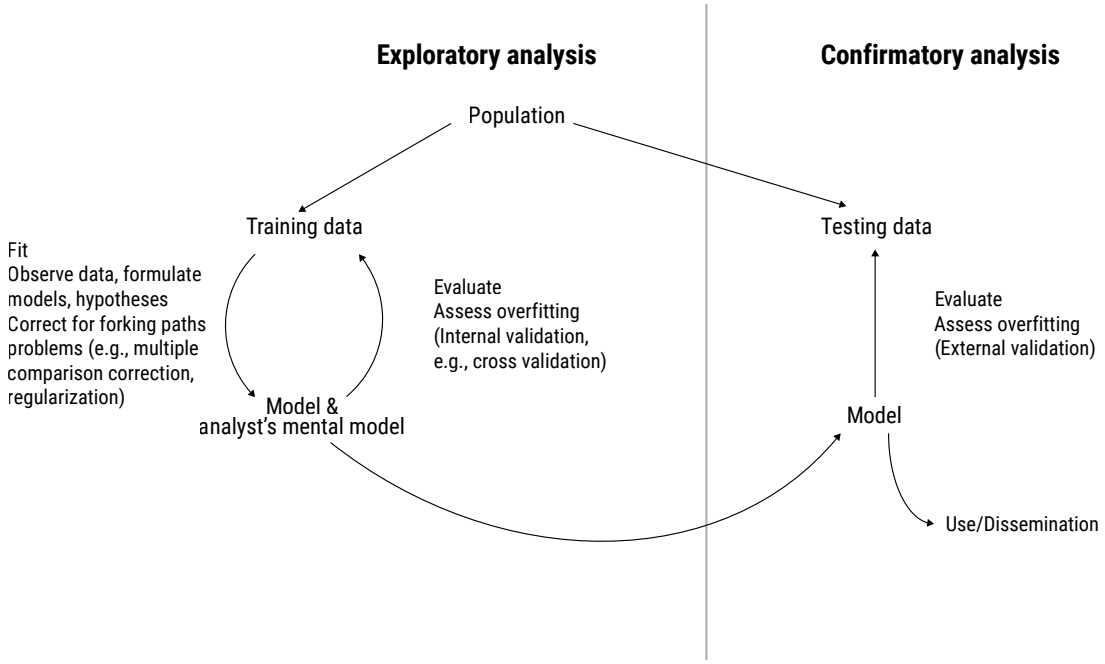


Figure 2.3: Analytic component of a general statistical workflow.

clear thinking. To test our project, then, we should clarify it in writing. Once it is clear, we can begin the processes of data collection, further clarify our understanding, begin technical work, again clarify our understanding, and continuing the iterative process until we converge on interesting answers that support actions and goals.

2.3.9 *Applying project scope: Citi Bike*

Let's develop the concept of project scope in the context of an example, one to help the bike share sponsored by Citi Bike.

You may have heard about, or even rented, a Citi Bike in New York City. Researching the history, we learn that in 2013, the New York City Department of Transportation sought to start a *bike share* to reduce emissions, road wear, congestion, and improve public health. After selecting an operator and sponsor, the Citi Bike bike share was established with a *bike fleet* distributed over a network of *docking stations* throughout the city. The bike share allows customers to unlock a bike at one station and return it at any other empty dock. Citi Bike spokeswoman Dani Simons has explained[14],

> Rebalancing is one of the biggest challenges of any bike share system, especially in a city like New York where residents don't all work a traditional 9-5 schedule, and though there is a Central Business District, it's a huge one and people work in a variety of other neighborhoods as well. At Citi Bike we've tried to be innovative in how we meet this

Figure 2.4: This Citi Bike docking station has available bikes and at least one available docking slot.

[14] Matthew Friedman, "Citi Bike Racks Continue to Go Empty Just When Upper West Siders Need Them," News, *West Side Rag*, August 2017.

challenge.

By *rebalancing*, she means ensuring continuous availability of both bikes and docking slots at each station. So we've just described business objectives, and a related problem or opportunity. Might this be a problem we can find available data and conduct analyses to *inform* the City's actions and further its goals?

> **Exercise 2.1** (Identify behaviors, events, data, sources, and context). Explore the availability of bikes and docking spots as depending on *users' patterns and behaviors*, *events* and locations at particular times, other forms of transportation, and on environmental context.
>
> What events may be correlated with or cause empty or full bike docking stations? What potential user behaviors or preferences may lead to these events? From what analogous things could we draw comparisons to provide context? How may these events and behaviors have been *measured* and *recorded*? What *data* are *available*? *Where* are it available? In what form?
>
> In what *contexts* are the data generated? In what ways may we find incomplete or missing data, or other *errors* in the stored measurements?
>
> May these data be *sufficient to find insights* through analysis, useful for decisions and goals?

Answers to questions as these provide necessary material for communication. Before digging into an analysis, let's discuss workflow.

2.4 *Workflow for credible communication*

> Truth is tough. It will not break, like a bubble, at a touch; nay, you may kick it about all day, like a football, and it will be round and full at evening.
>
> — Oliver W. Holmes, *The Professor at the Breakfast-Table*

Persuasive communication is credibly truthful, which means that our critics can test our language, our information, our methodologies, from start to finish. That others have not done so led to the reproducibility crisis noted in Nature[15]:

> More than 70% of researchers have tried and failed to reproduce another scientist's experiments, and more than half have failed to reproduce their own experiments.

[15] Monya Baker, "Is There a Reproducibility Crisis?" *Nature* 533, no. 26 (May 2016): 452–54.

By reproducibility, this meta-analysis considers whether replicating a study resulted in the same statistically significant finding (some have argued that reproducibility as a measure should compare, say, a p-value across trials, not whether the p-value crossed a given threshold in each trial). Regardless, we should **reproducibly** build our data analyses like Holmes's football, for our critics (later selves included)

to kick it about. What does this require? Ideally, our final product should include all components of our analysis from thoughts on our goals, to identification of — and code for — collection of data, visualization, modeling, reporting and explanations of insights. In short, the critic, with the touch of her finger, should be able to reproduce our results from our work. Perhaps that sounds daunting. But with some planning and use of modern tools, reproducibility is usually practical. Guidance on assessing reproducibility and a template for reproducible workflow is described by Kitzes and co-authors[16], along with a collection of more than 30 case studies. The authors identify three general practices that lead to reproducible work, to which I'll add a fourth:

[16] Justin Kitzes, Daniel Turek, and Fatma Deniz, *The Practice of Reproducible Research*, Case Studies and Lessons from the Data-Intensive Sciences (University of California Press, 2018).

1. Clearly separate, label, and document all data, files, and operations that occur on data and files.

2. Document all operations fully, automating them as much as possible, and avoiding manual intervention in the workflow when feasible.

3. Design a workflow as a sequence of small steps that are glued together, with intermediate outputs from one step feeding into the next step as inputs.

4. The workflow should track your history of changes.

Several authors[17] describe modern tools and approaches for creating a workflow that leads to reproducible research supporting credible communication.

[17] Christopher Gandrud, *Reproducible Research with R and RStudio*, Third edition, The R Series (Boca Raton, FL: CRC Press, 2020); Kieran Healy, "The Plain Person's Guide to Plain Text Social Science" (April 2018).

The workflow should include the communication. And the communication includes the code. What? Writing code to clean, transform, and analyze data may not generally be thought of as communicating. But yes! Code *is* language. And sometimes showing code is the most efficient way to express an idea. As such, we should strive for the most readable code possible. For our future selves. And for others. For code style advice, consult *The Art of Readable Code*[18] and *The Pragmatic Programmer*.[19]

[18] Dustin Boswell and Trevor Foucher, *The Art of Readable Code* (O'Reilly, 2011).

[19] David Thomas and Andrew Hunt, *The Pragmatic Programmer*, 20th Anniversary, Your Journey to Mastery (Addison-Wesley, 2020).

Next, we review an example, a reproducible workflow created from a few software tools (`R`, markdown, `Stan`), from identifying the business goals and problem, to considering data, performing analysis, and tying the analysis to decision making.

3
An example, reproducible workflow

Let's consider this example workflow of a project from beginning to end, starting with the (hypothetical) scenario and the observed (simulated) data, then fitting a series of Bayesian models in `Stan` and exploring them in `R`, and finally, using the models to enable decision-making. Our objective here is to consider an example, fully reproducible **workflow**. This workflow consists of including `R` and `Stan` code as blocks directly into the formatted-text narrative, of which the code and its executed results *knitted* together into the result you are reading. Of note, consider the primary audience is *not you*; the audiences are those relevant to that project.

> **Exercise 3.1** (Identify components of scope in workflow). As you review the example, reproducible workflow, consider whether, and with what detail, the example includes what we've discussed about the elements and scope of a data analytics project. What do you find? Any description of an iterative approach in the described analysis?

3.1 Background and problem

Our analysis responds to an owner of many residential buildings throughout New York City. The property manager explains that they are concerned about the number of complaints about loud music they receive from residents. Previously the company has offered monthly visits from a sound engineer to monitor decibel levels as a solution to this problem. While this is the default solution of many property managers in NYC, the residents are rarely home when the sound engineer visits, and so the manager reasoned that this is a relatively expensive solution that is currently not very effective.

One alternative is to deploy noise canceling devices — sound traps. In this alternative, they are installed throughout the building. The manufacturer of these sound traps provides some indication of

the device efficacy, but the manager suspects that this guidance was not calculated with NYC buildings in mind. In NYC, the manager rationalizes, sound carries more than elsewhere; and NYC buildings are built differently than other common residential buildings in the US. This is particularly important as the unit cost for each sound trap per year is high.

3.2 *The goal*

The manager has asked for help to find the optimal number of sound traps that should be placed in each of their buildings to minimize the number of noise complaints while also keeping expenditure on sound control affordable.

A subset of the company's buildings has been randomly selected for an experiment:

- At the beginning of each month, a sound engineer randomly places a number of sound traps throughout the building, without knowledge of the current decibel levels in the building.
- At the end of the month, the manager records the total number of complaints in that building.
- The manager would like to determine the optimal number of sound traps (traps) that balances the lost revenue (R) that complaints generate with the all-in cost of maintaining the traps (TC).

Bayesian data analysis provides a coherent framework for us to tackle this problem. Formally, we are interested in finding the number of traps that maximizes

$$\mathrm{E}(\mathrm{R}(\mathrm{complaints}(\mathrm{traps})) - \mathrm{TC}(\mathrm{traps})),$$

where the expectation averages over the distribution of complaints, conditional on the number of traps installed.

The property manager would also, if possible, like to learn how these results generalize to buildings they haven't setup the devices so they can understand the potential costs of sound control at buildings they are acquiring as well as for the rest of their building portfolio.

As the property manager has complete control over the number of traps set, the random variable contributing to this expectation is the number of complaints given the number of traps. We will model the number of complaints as a function of the number of traps.

3.3 *The data*

The owner has provided data for this problem representing data from 10 buildings in 12 successive months, thus 120 data points in total. Our initial, observed variables and their types are listed in table 3.1.

Table 3.1: Initial, observed variables for analysis.

variable	class	examples
mus	numeric	0.4, 0.4, 0.3, 0.1, ...
building_id	integer	37, 37, 37, 37, ...
wk_ind	integer	1, 2, 3, 4, ...
date	Date	"2017-01-15", "2017-02-14", "2017-03-16", "2017-04-15", ...
traps	integer	8, 8, 9, 10, ...
floors	integer	8, 8, 8, 8, ...
sq_footage_p_floor	numeric	5149, 5149, 5149, 5149, ...
live_in_super	integer	0, 0, 0, 0, ...
monthly_average_rent	numeric	3846.9, 3846.9, 3846.9, 3846.9, ...
average_tenant_age	numeric	53.9, 53.9, 53.9, 53.9, ...
age_of_building	integer	47, 47, 47, 47, ...
total_sq_foot	numeric	41192.1, 41192.1, 41192.1, 41192.1, ...
month	integer	1, 2, 3, 4, ...
complaints	integer	1, 3, 0, 1, ...

These are the variables we will be using:

`building_id`: The unique building identifier

`traps`: The number of sound traps used in the building in that month

`floors`: The number of floors in the building

`live_in_super`: An indicator for whether the building as a live-in superintendent

`monthly_average_rent`: The average monthly rent in the building

`average_tenant_age`: The average age of the tenants in the building

`age_of_building`: The age of the building

`total_sq_foot`: The total square footage in the building

`month`: Month of the year

`complaints`: Number of complaints in the building in that month

We have data for 10 buildings. Let's explore the data graphically. Firstly, we consider a histogram, figure 3.1, of the number of complaints in the 120 building-months in the data.

The pattern of this count data seems to be consistent with a Poisson distribution. Next, we graphically review complaints versus traps, shown in figure 3.2. Each dot represents a building-month, color-encoded for a live-in super and an off-premises super.

Graphing these variables over time, figure 3.3, we see no obvious patterns common to all locations.

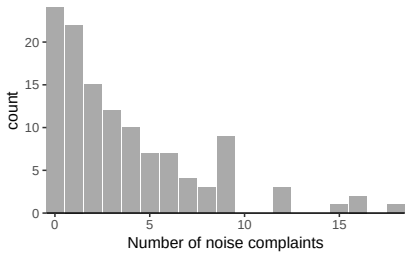


Figure 3.1: Frequency of the number of complaints.

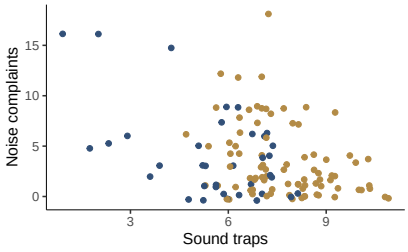


Figure 3.2: Number of complaints as a function of number of traps. Blue indicates a **live-in super** while brown indicates an **off-premises super**.

We will first analyze the association of the number of traps with the number of complaints, ignoring systematic variation over time and across buildings (we'll come back to those sources of variation later). That requires only two variables, complaints and traps. How should we model the number of complaints? We will demonstrate using a Bayesian workflow of model building, model checking, and model improvement.

3.4 *Modeling count data: Poisson distribution*

We already know some rudimentary information about what we should expect. The number of complaints over a month should be either zero or a positive integer. The property manager tells us that it is possible but unlikely that number of complaints in a given month is zero. Occasionally there are a large number of complaints in a single month. A common way of modeling this sort of skewed, single bounded count data is as a Poisson random variable. One concern about modeling the outcome variable as Poisson is that the data may be over-dispersed, but we'll start with the Poisson model and then check whether over-dispersion is a problem by comparing our model's predictions to the data.

3.4.1 *Model*

Given that we have chosen a Poisson regression, we define the likelihood to be the Poisson probability mass function over the num-

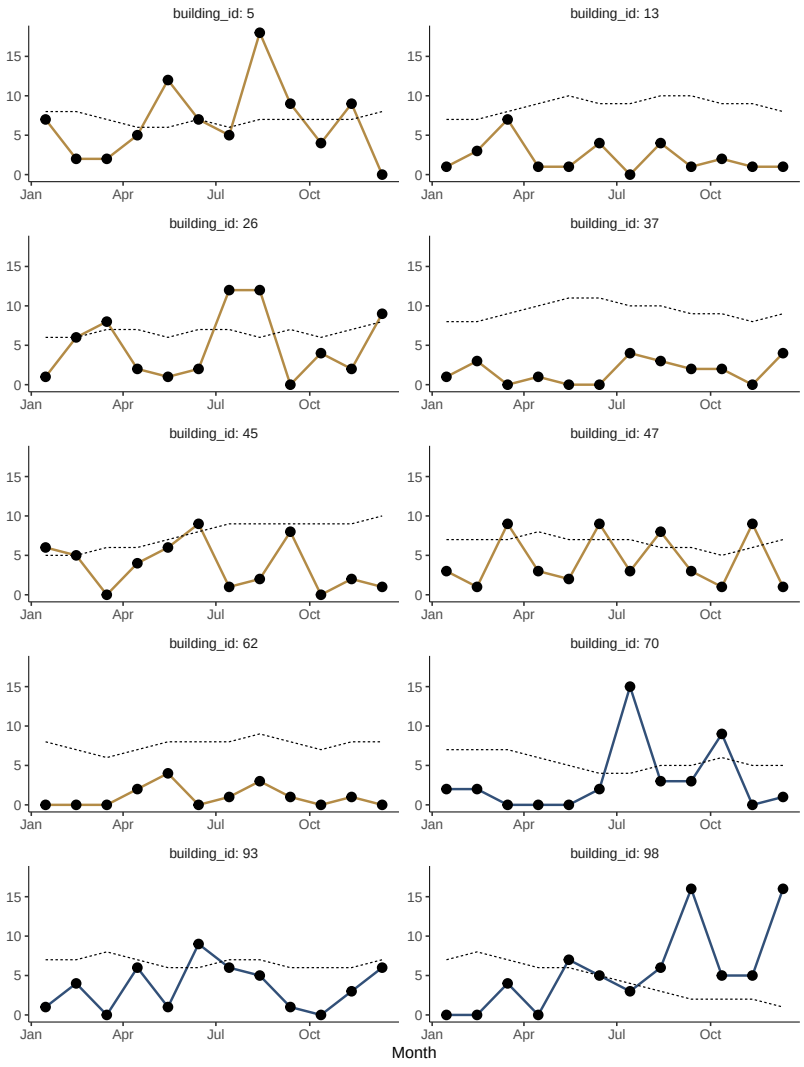


Figure 3.3: Time series of traps and complaints for each building in figure 3.3, again coloring **live-in supers** and **off-premises supers**. These data are graphed alongside the number of traps, shown as a dashed (– – –) line.

ber of traps placed in the building, denoted below as `traps`. This model assumes that the mean and variance of the outcome variable `complaints` (number of complaints) is the same. We'll investigate whether this is a good assumption after we fit the model. For building $b = 1, \ldots, 10$ at time (month) $t = 1, \ldots, 12$, we have

$$\begin{aligned} \text{complaints}_{b,t} &\sim \text{Poisson}(\lambda_{b,t}) \\ \lambda_{b,t} &= \exp\left(\eta_{b,t}\right) \\ \eta_{b,t} &= \alpha + \beta \, \text{traps}_{b,t} \end{aligned}$$

Let's encode this probability model in a Stan program (compiled into R object `comp_model_P`). Stan code.

```
functions {
  int poisson_log_safe_rng(real eta) {
    real pois_rate = exp(eta);
    if (pois_rate >= exp(20.79))
      return -9;
    return poisson_rng(pois_rate);
  }
}
data {
  int<lower=1> N;
  int<lower=0> complaints[N];
  vector<lower=0>[N] traps;
}
parameters {
  real alpha;
  real beta;
}
model {
  beta ~ normal(-0.25, 1);
  alpha ~ normal(log(4), 1);

  complaints ~ poisson_log(alpha + beta * traps);
}
generated quantities {
  int y_rep[N];
  for (n in 1:N)
    y_rep[n] = poisson_log_safe_rng(alpha + beta * traps[n]);
}
```

Before we fit the model to the data that have been given to us, we should check that our model works well with data that we have simulated ourselves. We'll simulate data according to the model (compiled into R object `comp_dgp_simple`) and then check that we can sufficiently recover the parameter values used in the simulation. Stan code.

```
data {
  int<lower=1> N;
  real<lower=0> mean_traps;
}
model {
```

```
}
generated quantities {
  int traps[N];
  int complaints[N];
  real alpha = normal_rng(log(4), 0.1);
  real beta = normal_rng(-0.25, 0.1);

  for (n in 1:N)  {
    traps[n] = poisson_rng(mean_traps);
    complaints[n] = poisson_log_rng(alpha + beta * traps[n]);
  }
}
```

For the compiled Stan program `comp_dgp_simple`, We simulate fake data by calling the `sampling()` function. R code.

```
fitted_model_dgp <-
  sampling(comp_dgp_simple,
           data = list(N = nrow(noise_data),
                       mean_traps = mean(noise_data$traps)),
           chains = 1, iter = 1,
           algorithm = 'Fixed_param',
           seed = 123)
```

We now extract the sampled data and look at its structure in R: R code.

```
sims_dgp <- rstan::extract(fitted_model_dgp)
str(sims_dgp)

List of 5
 $ traps     : num [1, 1:120] 7 5 8 11 9 6 5 6 8 9 ...
  ..- attr(*, "dimnames")=List of 2
  .. ..$ iterations: NULL
  .. ..$           : NULL
 $ complaints: num [1, 1:120] 0 1 0 0 0 0 0 0 1 0 ...
  ..- attr(*, "dimnames")=List of 2
  .. ..$ iterations: NULL
  .. ..$           : NULL
 $ alpha     : num [1(1d)] 1.29
  ..- attr(*, "dimnames")=List of 1
  .. ..$ iterations: NULL
 $ beta      : num [1(1d)] -0.283
  ..- attr(*, "dimnames")=List of 1
  .. ..$ iterations: NULL
 $ lp__      : num [1(1d)] 0
  ..- attr(*, "dimnames")=List of 1
  .. ..$ iterations: NULL
```

To pass the fake data to our Stan program using `RStan`, we need to arrange the data into a named list. The names must match the names used in the `data` block of the Stan program. R code.

```
stan_dat_fake <- list(N = nrow(noise_data),
                      traps = sims_dgp$traps[1, ],
                      complaints = sims_dgp$complaints[1, ])
```

Now that we have the simulated data we fit the model to see if we can recover the `alpha` and `beta` parameters used in the simulation. R code.

```
fit_model_P <- sampling(comp_model_P,
                        data = stan_dat_fake,
                        seed = 123,
                        chains = 4, cores = 4)

posterior_alpha_beta <- as.matrix(fit_model_P, pars = c('alpha','beta'))
```

```
          parameters
iterations    alpha       beta
      [1,] 1.857192 -0.3925038
      [2,] 1.752724 -0.3919860
      [3,] 1.745682 -0.3960594
      [4,] 1.790196 -0.3756891
      [5,] 1.805990 -0.4062681
      [6,] 1.891166 -0.3893186
```

3.4.2 *Assess parameter recovery*

We graphically compare, in figure 3.4, the **known (simulated) values** of the parameters to their estimated posterior distributions.

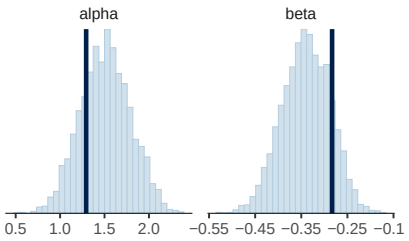


Figure 3.4: Compare known values to posterior distributions.

The posterior uncertainties are large here, but the true values are well within the inferential ranges. If we did the simulation with many more observations the parameters would be estimated much more precisely while still including the true values (assuming the model has been programmed correctly and the simulations have converged).

We also check if the `y_rep` datasets (in-sample predictions) that we coded in the `generated quantities` block are similar to the `y` (complaints) values we conditioned on when fitting the model.

Figure 3.5 is a plot of the density estimate of the **observed data** compared with 200 estimates of the data.

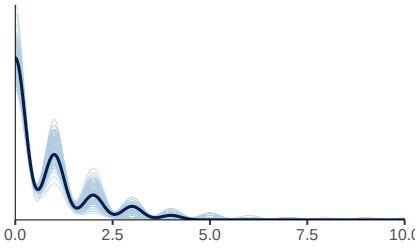


Figure 3.5: Density estimate of observed data compared with 200 simulated datasets.

In the plot above we have the kernel density estimate of the **observed data** (y, thicker curve) and 200 simulated data sets (y_{rep}, thin curves) from the posterior predictive distribution. If the model fits the data well, as it does, there will be little difference between the observed dataset and the simulated datasets.

In figure 3.6, we use a rootogram[1] to graph the **expected counts (continuous line)** versus the observed counts. The observed histogram matches the expected counts relatively well.

[1] Christian Kleiber and Achim Zeileis, "Visualizing Count Data Regressions Using Rootograms," *The American Statistician* 70, no. 3 (July 2016): 296–303.

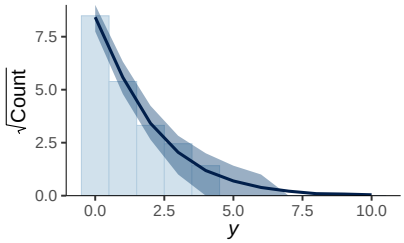


Figure 3.6: **Expected counts (continuous line)** versus observed counts.

3.4.3 *Fit with real data*

To fit the model to the data given to us, we first code a list to pass to Stan using the variables in the `noise_data` dataframe:

R code.

```
stan_dat_simple <- list(N = nrow(noise_data),
                        complaints = noise_data$complaints,
                        traps = noise_data$traps)
```

As we have compiled the model, we next sample from it.

R code.

```
fit_P_real_data <- sampling(comp_model_P,
                            data = stan_dat_simple,
                            chains = 4, cores = 4)
```

Here are the parameters:

```
Inference for Stan model: 876057cb647dda742cb72f0ee9007ee2.
4 chains, each with iter=2000; warmup=1000; thin=1;
post-warmup draws per chain=1000, total post-warmup draws=4000.

       mean se_mean   sd  2.5%   25%   50%   75% 97.5% n_eff Rhat
alpha  2.59    0.01 0.16  2.28  2.48  2.59  2.69  2.89   823    1
beta  -0.19    0.00 0.02 -0.24 -0.21 -0.19 -0.18 -0.14   756    1

Samples were drawn using NUTS(diag_e) at Mon Mar  2 11:34:21 2020.
For each parameter, n_eff is a crude measure of effective sample size,
and Rhat is the potential scale reduction factor on split chains (at
convergence, Rhat=1).
```

The coefficient β is estimated to be negative, implying that a higher number of traps set in a building appears to be associated with fewer complaints about noise in the following month. But we still need to consider how well the model fits.

3.4.4 *Posterior predictive checking*

The replicated datasets are not as dispersed as the **observed data** (figure 3.7) and don't seem to capture the observed rate of zeroes,

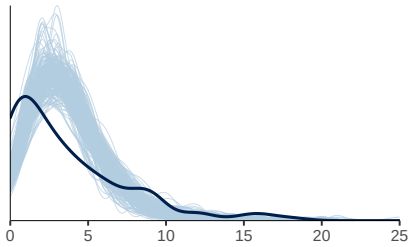


Figure 3.7: The replicated datasets are not as dispersed as the observed data.

The Poisson model may not be a good fit for these data. Let's explore this further by considering, in figure 3.8, the proportion of zeroes in the **real data** and predicted data.

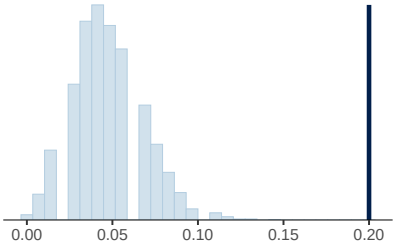


Figure 3.8: Compare the proportion of zeroes in the real data and predictions.

Figure 3.8 shows the **observed proportion of zeroes (thick vertical line)** and a histogram of the proportion of zeroes in each of the simulated data sets. It is clear that the model does not capture this feature of the data well at all. Let's consider, in figure 3.9, the standardized residuals of the observed vs predicted number of complaints.

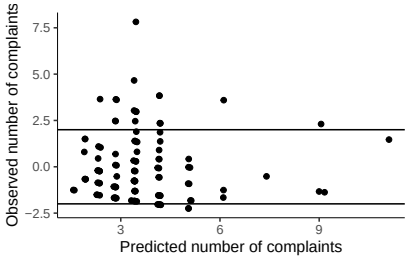


Figure 3.9: Comparing standardized residuals of the observed vs predicted number of complaints indicates more positive residuals than negative.

It looks as though we have more positive residuals than negative, which indicates that the model tends to underestimate the number of complaints that will be received.

We again graphically compare **expected counts (continuous line)** with observed counts (histogram) in the rootogram, figure 3.10.

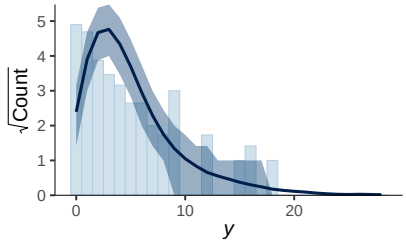


Figure 3.10: Expected counts (continuous line) versus the observed counts (blue histogram).

If the model was fitting well these would be relatively similar, but this figure shows that the number of complaints is underestimated if there are few complaints, over-estimated for medium numbers of complaints, and underestimated for large numbers of complaints.

We also view how the predicted number of complaints varies with the number of traps. The model doesn't seem to fully capture the **observed data**.

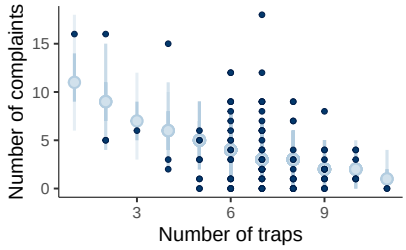


Figure 3.11: Predicted number of complaints varies with the number of sound traps.

The model doesn't estimate the tails of the **observed data** well.

3.5 *Expanding the model: multiple predictors*

Modeling the relationship between complaints and traps is the simplest model. We can expand the model, however, in a few ways that will be beneficial for our client. Moreover, the manager has told us that they expect there are a number of other reasons that one building might have more complaints of noise than another.

3.5.1 *Interpretability*

Currently, our model's mean parameter is a rate of complaints per 30 days, but we're modeling a process that occurs over an area as well as

over time. We have the square footage of each building, so if we add that information into the model, we can interpret our parameters as a rate of complaints per square foot per 30 days.

$$\begin{aligned} \text{complaints}_{b,t} &\sim \text{Poisson}(\text{sq\_foot}_b \, \lambda_{b,t}) \\ \lambda_{b,t} &= \exp\left(\eta_{b,t}\right) \\ \eta_{b,t} &= \alpha + \beta \, \text{traps}_{b,t} \end{aligned}$$

The term sq_foot is called an exposure term. If we log the term, we can put it in $\eta_{b,t}$:

$$\begin{aligned} \text{complaints}_{b,t} &\sim \text{Poisson}(\lambda_{b,t}) \\ \lambda_{b,t} &= \exp\left(\eta_{b,t}\right) \\ \eta_{b,t} &= \alpha + \beta \, \text{traps}_{b,t} + \text{log\_sq\_foot}_b \end{aligned}$$

A quick check in figure 3.12 suggests a relationship between the building square footage and the number of noise complaints.

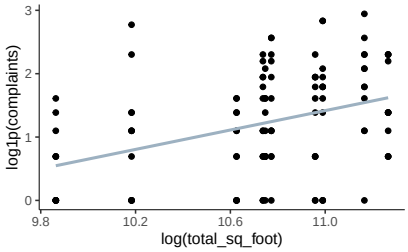


Figure 3.12: Compare the square footage of the building with the number of noise complaints.

Using the property manager's intuition, we include two extra pieces of information we know about the building — the (log of the) square floor space and whether there is a live in super or not — into both the simulated and real data.

R code.

```
stan_dat_simple$log_sq_foot <- log(noise_data$total_sq_foot/1e4)
stan_dat_simple$live_in_super <- noise_data$live_in_super
```

3.5.2 *Stan program for Poisson multiple regression*

Now we code a new Stan model that uses multiple predictors.

Stan code.

```
functions {
  int poisson_log_safe_rng(real eta) {
    real pois_rate = exp(eta);
    if (pois_rate >= exp(20.79))
      return -9;
    return poisson_rng(pois_rate);
  }
}
data {
```

```
  int<lower=1> N;
  int<lower=0> complaints[N];
  vector<lower=0>[N] traps;
  vector<lower=0,upper=1>[N] live_in_super;
  vector[N] log_sq_foot;
}
parameters {
  real alpha;
  real beta;
  real beta_super;
}
model {
  beta ~ normal(-0.25, 1);
  beta_super ~ normal(-0.5, 1);
  alpha ~ normal(log(4), 1);
  complaints ~ poisson_log(alpha +
                           beta * traps +
                           beta_super * live_in_super +
                           log_sq_foot);
}
generated quantities {
  int y_rep[N];
  for (n in 1:N)
    y_rep[n] = poisson_log_safe_rng(alpha +
                beta * traps[n] +
                beta_super * live_in_super[n] +
                log_sq_foot[n]);
}
```

3.5.3 *Simulate fake data with multiple predictors*

As before, we check the model using simulated data. We use Stan to simulate data, compiling the program into R object `comp_dgp_multiple`. Stan code.

```
data {
  int<lower=1> N;
}
model {
}
generated quantities {
  vector[N] log_sq_foot;
  int live_in_super[N];
  int traps[N];
  int complaints[N];
  real alpha = normal_rng(log(4), 0.1);
  real beta = normal_rng(-0.25, 0.1);
  real beta_super = normal_rng(-0.5, 0.1);
  for (n in 1:N) {
    log_sq_foot[n] = normal_rng(1.5, 0.1);
    live_in_super[n] = bernoulli_rng(0.5);
    traps[n] = poisson_rng(8);
    complaints[n] = poisson_log_rng(alpha + log_sq_foot[n]
      + beta * traps[n] + beta_super * live_in_super[n]);
  }
}
```

Next, we sample the compiled model to get simulated data. R code.

```
fitted_model_dgp <- sampling(comp_dgp_multiple,
                             data = list(N = nrow(noise_data)),
```

```
                              chains = 1, cores = 1,
                              iter = 1, algorithm = 'Fixed_param',
                              seed = 123)
sims_dgp <- rstan::extract(fitted_model_dgp)
```

We'll push the simulated data as a list into Stan. R code.

```
stan_dat_fake <- list(N = nrow(noise_data),
                      log_sq_foot = sims_dgp$log_sq_foot[1, ],
                      live_in_super = sims_dgp$live_in_super[1, ],
                      traps = sims_dgp$traps[1, ],
                      complaints = sims_dgp$complaints[1, ])
```

We compile the model into R object `comp_model_P_mult`: Stan code.

```
functions {
  int poisson_log_safe_rng(real eta) {
    real pois_rate = exp(eta);
    if (pois_rate >= exp(20.79))
      return -9;
    return poisson_rng(pois_rate);
  }
}
data {
  int<lower=1> N;
  int<lower=0> complaints[N];
  vector<lower=0>[N] traps;
  vector<lower=0,upper=1>[N] live_in_super;
  vector[N] log_sq_foot;
}
parameters {
  real alpha;
  real beta;
  real beta_super;
}
model {
  beta ~ normal(-0.25, 1);
  beta_super ~ normal(-0.5, 1);
  alpha ~ normal(log(4), 1);
  complaints ~ poisson_log(alpha +
                           beta * traps +
                           beta_super * live_in_super +
                           log_sq_foot);
}
generated quantities {
  int y_rep[N];
  for (n in 1:N)
    y_rep[n] = poisson_log_safe_rng(alpha +
                beta * traps[n] +
                beta_super * live_in_super[n] +
                log_sq_foot[n]);
}
```

And we sample from the model: R code.

```
fit_model_P_mult <- sampling(comp_model_P_mult,
                              data = stan_dat_fake,
                              chains = 4, cores = 4)
```

Then compare these parameters to the **true parameters**:

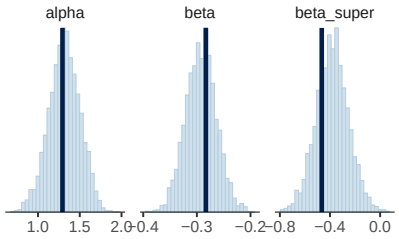


Figure 3.13: Compare model parameters to the true parameters.

Now that we've recovered the parameters from the simulated data, we're ready to fit the measured data that were given to us.

3.5.4 *Fit the measured (observed) data*

We explore the fit by comparing the data to posterior predictive simulations:

R code.

```
fit_model_P_mult_real <- sampling(comp_model_P_mult,
                                  data = stan_dat_simple,
                                  chains = 4, cores = 4)

y_rep <- as.matrix(fit_model_P_mult_real,
                   pars = "y_rep")
```

As we see in figure 3.14, This again looks like we haven't estimated the **observed smaller counts** well, nor have we estimated the **observed larger counts**.

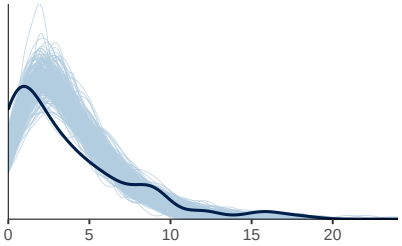


Figure 3.14: Density overlay.

We're still underestimating the proportion of zeroes in the **observed data** (figure 3.15). Ideally this **vertical line** would fall somewhere within the histogram. We also, in figure 3.16, graph uncertainty intervals for the predicted complaints for different numbers of traps.

We've increased the tails a bit more at the larger numbers of traps, but we still have some large observed numbers of complaints that the model would consider extremely unlikely events.

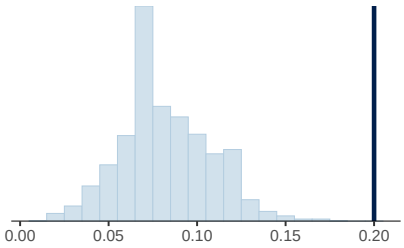


Figure 3.15: Compare proportion of zeros estimated to the data.

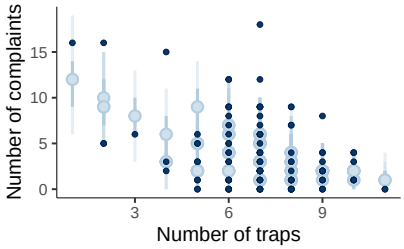


Figure 3.16: Uncertainty intervals for the predicted complaints for different numbers of traps.

3.6 *Modeling count data with the negative binomial distribution*

When we considered modelling the data using a Poisson, we saw that the model didn't appear to fit as well to the data as we would like. In particular the model under-predicted low and high numbers of complaints, and over-predicted the medium number of complaints. This is one indication of overdispersion, where the variance is larger than the mean. A Poisson model doesn't fit overdispersed count data well because the same parameter λ, controls both the expected counts and the variance of these counts. The natural alternative to this is the negative binomial model:

$$
\begin{aligned}
\text{complaints}_{b,t} &\sim \text{Neg-Binomial}(\lambda_{b,t}, \phi) \\
\lambda_{b,t} &= \exp\left(\eta_{b,t}\right) \\
\eta_{b,t} &= \alpha + \beta\,\text{traps}_{b,t} + \beta_{\text{super}}\,\text{super}_b + \text{log\_sq\_foot}_b
\end{aligned}
$$

In Stan the negative binomial mass function we'll use is called `neg_binomial_2_log`$(\text{ints}\, y, \text{reals}\, \eta, \text{reals}\, \phi)$ in Stan. Like the `poisson_log` function, this negative binomial mass function that is parameterized in terms of its log-mean, η, but it also has a precision ϕ such that

$$
\text{E}[y] = \lambda = \exp(\eta)
$$

$$
\text{Var}[y] = \lambda + \lambda^2/\phi = \exp(\eta) + \exp(\eta)^2/\phi.
$$

As ϕ gets larger the term λ^2/ϕ approaches zero and so the variance of the negative-binomial approaches λ; that is, the negative-binomial gets closer and closer to the Poisson.

3.6.1 *Stan program for negative-binomial regression*

Let's code a model using the negative-binomial, compiled into R object `comp_dgp_multiple_NB`. Stan code.

```
data {
  int<lower=1> N;
}
model {
}
generated quantities {
  vector[N] log_sq_foot;
  int live_in_super[N];
  int traps[N];
  int complaints[N];
  real alpha = normal_rng(log(4), 0.1);
  real beta = normal_rng(-0.25, 0.1);
  real beta_super = normal_rng(-0.5, 0.1);
  real inv_phi = fabs(normal_rng(0, 1));

  for (n in 1:N) {
    log_sq_foot[n] = normal_rng(1.5, 0.1);
    live_in_super[n] = bernoulli_rng(0.5);
    traps[n] = poisson_rng(8);
    complaints[n] = neg_binomial_2_log_rng(alpha + log_sq_foot[n]
      + beta * traps[n] + beta_super * live_in_super[n], inv(inv_phi));
  }
}
```

3.6.2 *Fake data fit: Multiple negative-binomial regression*

Next, we generate one draw from the fake data model so we can use the data to fit our model and compare the known values of the parameters to the posterior density of the parameters. R code.

```
fitted_model_dgp_NB <- sampling(comp_dgp_multiple_NB,
                                data = list(N = nrow(noise_data)),
                                chains = 1, cores = 1, iter = 1,
                                algorithm = 'Fixed_param', seed = 123)
sims_dgp_NB <- rstan::extract(fitted_model_dgp_NB)
```

Here's our dataset to feed into this Stan model. R code.

```
stan_dat_fake_NB <- list(N = nrow(noise_data),
                         log_sq_foot = sims_dgp_NB$log_sq_foot[1, ],
                         live_in_super = sims_dgp_NB$live_in_super[1, ],
                         traps = sims_dgp_NB$traps[1, ],
                         complaints = sims_dgp_NB$complaints[1, ]
                         )
```

After compiling the inferential model into R object `comp_model_NB`, Stan code.

```
functions {
```

```
  int neg_binomial_2_log_safe_rng(real eta, real phi) {
    real gamma_rate = gamma_rng(phi, phi / exp(eta));
    if (gamma_rate >= exp(20.79))
      return -9;
    return poisson_rng(gamma_rate);
  }
}
data {
  int<lower=1> N;
  vector<lower=0>[N] traps;
  vector<lower=0,upper=1>[N] live_in_super;
  vector[N] log_sq_foot;
  int<lower=0> complaints[N];
}
parameters {
  real alpha;
  real beta;
  real beta_super;
  real<lower=0> inv_phi;
}
transformed parameters {
  real phi = inv(inv_phi);
}
model {
  alpha ~ normal(log(4), 1);
  beta ~ normal(-0.25, 1);
  beta_super ~ normal(-0.5, 1);
  inv_phi ~ normal(0, 1);
  complaints ~ neg_binomial_2_log(alpha +
                                  beta * traps +
                                  beta_super * live_in_super +
                                  log_sq_foot, phi);
}
generated quantities {
  int y_rep[N];
  for (n in 1:N)
    y_rep[n] = neg_binomial_2_log_safe_rng(alpha + beta * traps[n] +
      beta_super * live_in_super[n] + log_sq_foot[n], phi);

}
```

we run our NB regression with the fake data and extract samples to examine posterior predictive checks and check whether we've sufficiently recovered our known parameters, alpha `beta`,

R code.

```
fitted_model_NB <- sampling(comp_model_NB,
                            data = stan_dat_fake_NB,
                            chains = 4, cores = 4)
posterior_alpha_beta_NB <-
  as.matrix(fitted_model_NB,
            pars = c('alpha', 'beta', 'beta_super', 'inv_phi'))
```

We construct the vector of **true values** from the simulated dataset and compare with the recovered parameters in figure 3.17.

R code.

```
true_alpha_beta_NB <- c(sims_dgp_NB$alpha,
                        sims_dgp_NB$beta,
                        sims_dgp_NB$beta_super,
                        sims_dgp_NB$inv_phi)
```

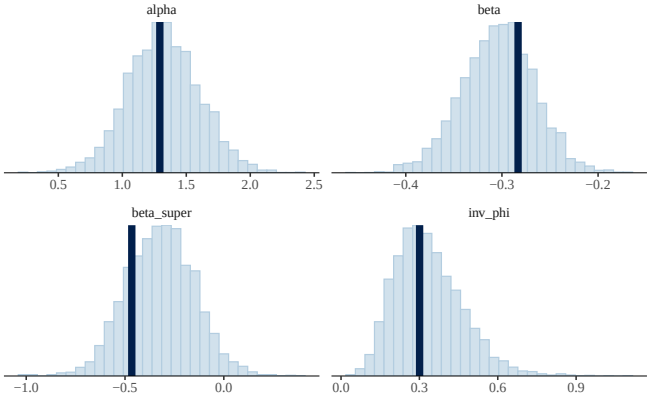


Figure 3.17: Compare simulated dataset with recovered parameters.

3.6.3 *Fit to measured data and check our fit*

R code.

```
fitted_model_NB <- sampling(comp_model_NB,
                            data = stan_dat_simple,
                            chains = 4, cores = 4)
sims_NB <- rstan::extract(fitted_model_NB)
```

Let's compare our predictions with the **observed data** in figure 3.18.

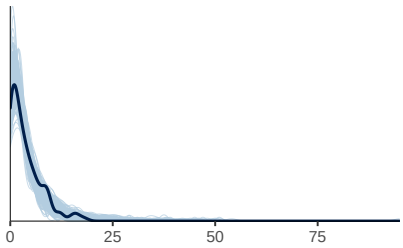


Figure 3.18: Compare predictions with observations.

It appears that our model now estimates both the number of small counts better as well as the tails. Let's check if the negative binomial model does a better job estimating the **observed number of zeroes** in figure 3.19.

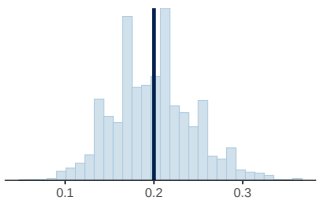


Figure 3.19: Negative binomial model comparison with data.

These look OK, but let's look at the standardized residual plot.

The standardized residuals look fair in figure 3.20, but we still have some large *standardized* residuals. This might be because we are

currently ignoring that the data are clustered by buildings, and that the probability of noise issue may vary substantially across buildings.

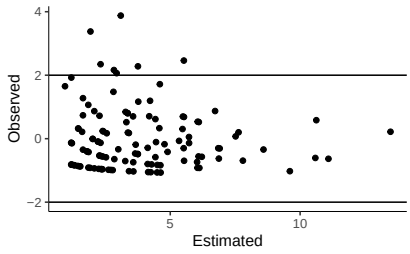


Figure 3.20: Standardized residual plot.

In the rootogram in figure 3.21, the estimates seem more plausible.

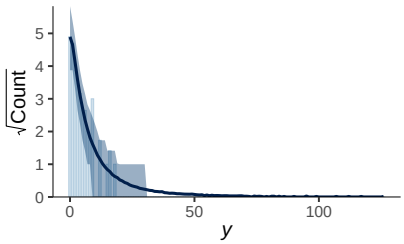


Figure 3.21: The expected number of complaints matches closer to the observed number of complaints.

We can tell this because now the expected number of complaints matches much closer to the **observed number of complaints** for a given number of sound traps, see figure 3.22. However, we still have some larger counts that appear to be outliers for the model.

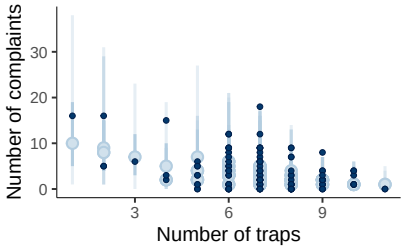


Figure 3.22: Compare predictions with number of sound traps.

We haven't considered that the observed data are clustered by building yet. Let's check posterior predictions to see whether we should include building information into the model.

Figure 3.23 suggests that we're getting plausible predictions for most building means, but some are estimated better than others and some have larger uncertainties than we might expect. If we explicitly model the variation across buildings we may be able to get better estimates.

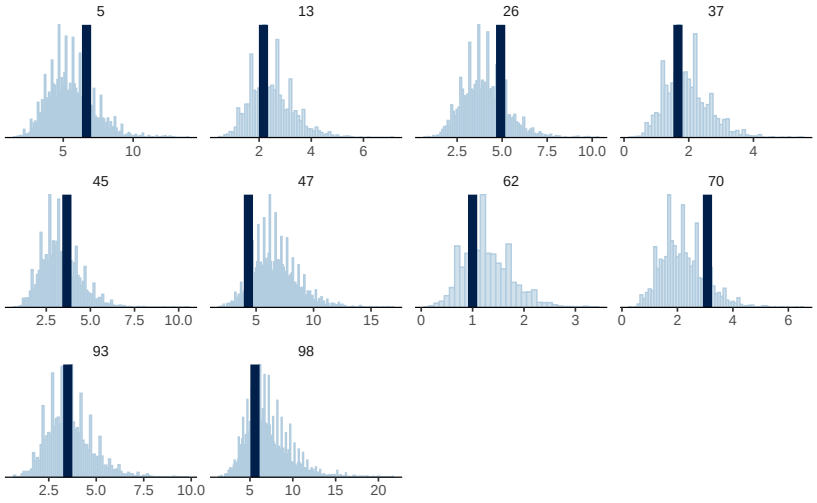


Figure 3.23: Compare posterior means to observations.

3.7 *Hierarchical modeling*

3.7.1 *Modeling varying intercepts for each building*

Let's add a hierarchical intercept parameter, α_b at the building level to our model.

$$\begin{aligned} \text{complaints}_{b,t} &\sim \text{Neg-Binomial}(\lambda_{b,t}, \phi) \\ \lambda_{b,t} &= \exp\left(\eta_{b,t}\right) \\ \eta_{b,t} = \mu_b + \beta\,\text{traps}_{b,t} + \beta_{\text{super}}\,\text{super}_b + \text{log\_sq\_foot}_b \\ \mu_b &\sim \text{normal}(\alpha, \sigma_\mu) \end{aligned}$$

In our Stan model, μ_b is the b-th element of the vector `mu` which has one element per building.

One of our predictors varies only by building, so we can rewrite the above model more efficiently like so:

$$\begin{aligned} \eta_{b,t} &= \mu_b + \beta\,\text{traps}_{b,t} + \text{log\_sq\_foot}_b \\ \mu_b &\sim \text{normal}(\alpha + \beta_{\text{super}}\,\text{super}_b, \sigma_\mu) \end{aligned}$$

We have more information at the building level as well, like the average age of the residents, the average age of the buildings, and the average per-apartment monthly rent so we can add that data into a matrix called `building_data`, which will have one row per building and four columns:

`live_in_super`

`age_of_building`

`average_tentant_age`

`monthly_average_rent`

We'll write the Stan model like:

$$\eta_{b,t} = \alpha_b + \beta \, \text{traps} + \text{log\_sq\_foot}$$
$$\mu \sim \text{normal}(\alpha + \texttt{building\_data}\,\zeta, \, \sigma_\mu)$$

3.7.2 *Prepare building data for hierarchical model*

We'll need to do some more data prep before we can fit our models. Firstly, to use the building variable in Stan, we transform it from a factor variable to an integer variable.

R code.

```
N_months <- length(unique(noise_data$date))

## Add some IDs for building and month
noise_data <- noise_data %>%
  mutate(
    building_fac = factor(building_id, levels = unique(building_id)),
    building_idx = as.integer(building_fac),
    ids = rep(1:N_months, N_buildings),
    mo_idx = lubridate::month(date)
  )

## Center and rescale the building specific data
building_data <- noise_data %>%
    select(building_idx, live_in_super, age_of_building,
      total_sq_foot, average_tenant_age, monthly_average_rent) %>%
    unique() %>%
    arrange(building_idx) %>%
    select(-building_idx) %>%
    scale(scale=FALSE) %>%
    as.data.frame() %>%
    mutate( ## scale by constants
      age_of_building = age_of_building / 10,
      total_sq_foot = total_sq_foot / 10000,
      average_tenant_age = average_tenant_age / 10,
      monthly_average_rent = monthly_average_rent / 1000
    ) %>%
    as.matrix()

## Make data list for Stan
stan_dat_hier <- with(noise_data,
  list(complaints = complaints,
       traps = traps,
       N = length(traps),
       J = N_buildings,
       M = N_months,
       log_sq_foot = log(noise_data$total_sq_foot/1e4),
       building_data = building_data[,-3],
       mo_idx = as.integer(as.factor(date)),
       K = 4, building_idx = building_idx))
```

3.7.3 *Compile and fit the hierarchical model*

Let's compile the model into R object `comp_model_NB_hier`. Stan code.

```
functions {
  int neg_binomial_2_log_safe_rng(real eta, real phi) {
    real gamma_rate = gamma_rng(phi, phi / exp(eta));
    if (gamma_rate >= exp(20.79))
      return -9;
    return poisson_rng(gamma_rate);
  }
}
data {
  int<lower=1> N;
  int<lower=0> complaints[N];
  vector<lower=0>[N] traps;
  // 'exposure'
  vector[N] log_sq_foot;
  // building-level data
  int<lower=1> K;
  int<lower=1> J;
  int<lower=1, upper=J> building_idx[N];
  matrix[J,K] building_data;
}
parameters {
  real<lower=0> inv_phi;
  real beta;
  vector[J] mu;
  real<lower=0> sigma_mu;
  real alpha;
  vector[K] zeta;
}
transformed parameters {
  real phi = inv(inv_phi);
}
model {
  mu ~ normal(alpha + building_data * zeta, sigma_mu);
  sigma_mu ~ normal(0, 1);
  alpha ~ normal(log(4), 1);
  zeta ~ normal(0, 1);
  beta ~ normal(-0.25, 1);
  inv_phi ~ normal(0, 1);
  complaints ~ neg_binomial_2_log(mu[building_idx] +
               beta * traps + log_sq_foot, phi);
}
generated quantities {
  int y_rep[N];
  for (n in 1:N) {
    real eta_n = mu[building_idx[n]] +
    beta * traps[n] + log_sq_foot[n];
    y_rep[n] = neg_binomial_2_log_safe_rng(eta_n, phi);
  }
}
```

And fit the model to data. R code.

```
fitted_model_NB_hier <- sampling(comp_model_NB_hier,
                                 data = stan_dat_hier,
                                 chains = 4, cores = 4,
                                 iter = 4000)
```

3.7.4 *Diagnostics*

We get warnings from Stan about divergent transitions,

```
Divergences:

490 of 8000 iterations ended with a divergence (6.125%).
Try increasing 'adapt_delta' to remove the divergences.


Tree depth:

0 of 8000 iterations saturated the maximum tree depth of 10.


Energy:

E-BFMI indicated no pathological behavior.
```

which indicates that there may be regions of the posterior that have not been explored by the Markov chains.

In this analysis, we see that we have divergent transitions because we need to reparameterize our model. We retain the overall structure of the model but transform some of the parameters so that it is easier for Stan to sample from the parameter space. Before we reparameterizing, we first consider how reparameterizing may resolve the issue. We examine: the fitted parameter values, including the effective sample size, and traceplots and scatterplots that reveal particular patterns in locations of the divergences. First let's extract the fits from the model.

R code.

```
sims_hier_NB <- rstan::extract(fitted_model_NB_hier)
```

Here are the parameter estimates that are of most interest.

```
Inference for Stan model: 2c1ee83c97acb05da1f2461f9097ddd4.
4 chains, each with iter=4000; warmup=2000; thin=1;
post-warmup draws per chain=2000, total post-warmup draws=8000.

          mean se_mean   sd  2.5%   25%   50%   75% 97.5% n_eff Rhat
sigma_mu  0.25    0.01 0.17  0.06  0.13  0.21  0.33  0.68   672    1
beta     -0.23    0.00 0.06 -0.35 -0.27 -0.23 -0.19 -0.11   709    1
alpha     1.26    0.02 0.43  0.39  0.98  1.27  1.55  2.13   725    1
phi       1.58    0.01 0.35  1.03  1.33  1.54  1.79  2.39  1108    1
mu[1]     1.28    0.02 0.55  0.16  0.92  1.29  1.63  2.37   802    1
mu[2]     1.23    0.02 0.53  0.17  0.89  1.23  1.57  2.28   859    1
mu[3]     1.41    0.02 0.49  0.43  1.10  1.42  1.73  2.37   841    1
mu[4]     1.45    0.02 0.48  0.51  1.12  1.44  1.76  2.44   851    1
mu[5]     1.08    0.01 0.42  0.24  0.81  1.10  1.35  1.93   946    1
mu[6]     1.17    0.02 0.48  0.19  0.85  1.19  1.49  2.13   748    1
mu[7]     1.47    0.02 0.52  0.45  1.13  1.47  1.81  2.48   829    1
mu[8]     1.26    0.01 0.42  0.41  0.98  1.27  1.54  2.09   869    1
mu[9]     1.42    0.02 0.57  0.29  1.04  1.44  1.80  2.53   669    1
mu[10]    0.86    0.01 0.37  0.16  0.62  0.86  1.09  1.61  1008    1

Samples were drawn using NUTS(diag_e) at Sat Mar  7 22:34:04 2020.
For each parameter, n_eff is a crude measure of effective sample size,
and Rhat is the potential scale reduction factor on split chains (at
convergence, Rhat=1).
```

The effective samples seem a little low for many of the parameters relative to the total number of samples. This alone doesn't indicate the need to reparameterize, but it does indicate that we should look further at the trace plots and pairs plots. First let's look at the traceplots (figure 3.24) to see if the divergent transitions form a pattern.

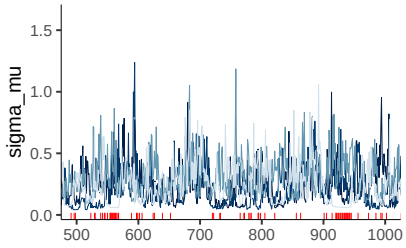


Figure 3.24: Traceplots show divergences bunching in patterns.

Looks as if the divergent parameters, the rug plot of red bars underneath the traceplots corresponds to samples where the sampler gets stuck at one parameter value for σ_μ.

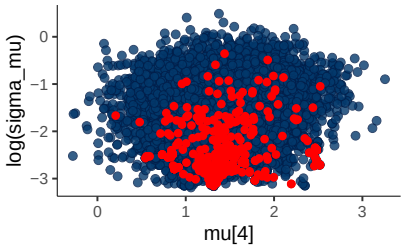


Figure 3.25: Scatter of MCMC draws from parameters.

In figure 3.25, we find cloud-like shape shows most of the divergences clustering towards the bottom. We'll see a bit later that we actually want this to look more like a funnel than a cloud, but the divergences are indicating that the sampler can't explore the narrowing neck of the funnel.

One way to see why we should expect some version of a funnel is to look at some simulations from the prior, which we can do without MCMC and thus with no risk of sampling problems:

R code.

```
N_sims    <- 1000
log_sigma <- rep(NA, N_sims)
theta     <- rep(NA, N_sims)

for (j in 1:N_sims) {
  log_sigma[j] <- rnorm(1, mean = 0, sd = 1)
  theta[j] <- rnorm(1, mean = 0, sd = exp(log_sigma[j]))
}

draws <- cbind("mu" = theta, "log(sigma_mu)" = log_sigma)
```

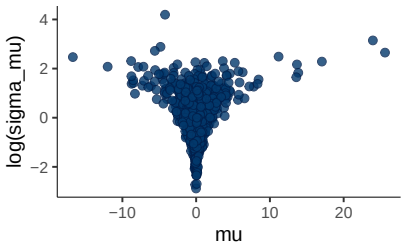


Figure 3.26: Scatter of MCMC draws from the prior form a funnel.

If the data are at all informative we shouldn't expect the posterior to look exactly like the prior. But unless the data are highly informative about the parameters and the posterior concentrates away from the narrow neck of the funnel, the sampler will have to confront the funnel geometry. (See the Visual MCMC Diagnostics.)

We consider another view of the divergences using a parallel coordinates plot:

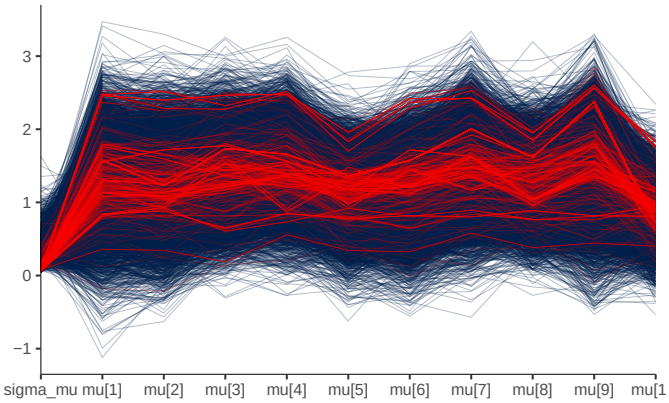


Figure 3.27: Parallel coordinate plot shows divergences concentrating when `sigma_mu` is small.

In figure 3.27, too, we see evidence that our problems concentrate when `sigma_mu` is small.

3.7.5 *Reparameterize and recheck diagnostics*

Instead, we should use the non-centered parameterization for μ_b. We define a vector of auxiliary variables in the parameters block, `mu_raw` that is given a normal(0,1) prior in the model block. We then make `mu` a transformed parameter: We can reparameterize the random intercept μ_b, which is distributed:

$$\mu_b \sim \text{normal}(\alpha + \texttt{building\_data}, \zeta, \sigma_\mu)$$

In Stan code,

Stan code.

```
transformed parameters {
  vector[J] mu;
  mu = alpha + building_data * zeta + sigma_mu * mu_raw;
}
```

This gives `mu` a normal$(\alpha + \texttt{building_data}, \zeta, \sigma_\mu)$ distribution, but it decouples the dependence of the density of each element of `mu` from `sigma_mu` (σ_μ). `hier_NB_regression_ncp.stan` uses the non-centered parameterization for `mu`. We will examine the effective sample size of the fitted model to see whether we've fixed the problem with our reparameterization.

Compile the model into R object `comp_model_NB_hier_ncp`.

Stan code.

```
functions {
  int neg_binomial_2_log_safe_rng(real eta, real phi) {
    real gamma_rate = gamma_rng(phi, phi / exp(eta));
    if (gamma_rate >= exp(20.79))
      return -9;
    return poisson_rng(gamma_rate);
  }
}
data {
  int<lower=1> N;
  int<lower=0> complaints[N];
  vector<lower=0>[N] traps;
  vector[N] log_sq_foot;
  int<lower=1> K;
  int<lower=1> J;
  int<lower=1, upper=J> building_idx[N];
  matrix[J,K] building_data;
}
parameters {
  real<lower=0> inv_phi;
  real beta;
  vector[J] mu_raw;
  real<lower=0> sigma_mu;
  real alpha;
  vector[K] zeta;
}
transformed parameters {
  real phi = inv(inv_phi);

  vector[J] mu = alpha +
                 building_data * zeta +
                 sigma_mu * mu_raw;
}
model {
  mu_raw ~ normal(0, 1);
  sigma_mu ~ normal(0, 1);
  alpha ~ normal(log(4), 1);
  zeta ~ normal(0, 1);
  beta ~ normal(-0.25, 1);
  inv_phi ~ normal(0, 1);
  complaints ~ neg_binomial_2_log(mu[building_idx] +
                                  beta * traps +
                                  log_sq_foot, phi);
}
generated quantities {
```

```
  int y_rep[N];
  for (n in 1:N) {
    real eta_n = mu[building_idx[n]] +
                 beta * traps[n] +
                 log_sq_foot[n];

    y_rep[n] = neg_binomial_2_log_safe_rng(eta_n, phi);
  }
}
```

Fit the model to the data. R code.

```
fitted_model_NB_hier_ncp <- sampling(comp_model_NB_hier_ncp,
                                     data = stan_dat_hier,
                                     chains = 4, cores = 4,
                                     control = list(adapt_delta = 0.95))
```

Our parameter estimates and effective sample sizes for the new model follows,

```
Inference for Stan model: 041aec96a5df841e1548e3c3343d99dd.
4 chains, each with iter=2000; warmup=1000; thin=1;
post-warmup draws per chain=1000, total post-warmup draws=4000.

          mean se_mean   sd  2.5%   25%   50%   75% 97.5% n_eff Rhat
sigma_mu  0.24    0.01 0.18  0.01  0.10  0.20  0.34  0.69  1266    1
beta     -0.23    0.00 0.06 -0.35 -0.27 -0.23 -0.19 -0.11  2381    1
alpha     1.25    0.01 0.43  0.41  0.97  1.25  1.54  2.10  2344    1
phi       1.60    0.01 0.36  1.03  1.34  1.55  1.80  2.45  4265    1
mu[1]     1.27    0.01 0.55  0.21  0.91  1.26  1.63  2.37  2283    1
mu[2]     1.22    0.01 0.53  0.18  0.87  1.20  1.57  2.27  2435    1
mu[3]     1.39    0.01 0.50  0.44  1.05  1.39  1.72  2.38  2888    1
mu[4]     1.43    0.01 0.48  0.47  1.10  1.43  1.75  2.41  2485    1
mu[5]     1.07    0.01 0.42  0.27  0.79  1.07  1.35  1.90  2936    1
mu[6]     1.17    0.01 0.49  0.18  0.85  1.16  1.50  2.13  2627    1
mu[7]     1.44    0.01 0.51  0.44  1.10  1.44  1.77  2.47  2951    1
mu[8]     1.24    0.01 0.43  0.41  0.96  1.24  1.52  2.09  2773    1
mu[9]     1.41    0.01 0.56  0.30  1.03  1.41  1.78  2.47  2657    1
mu[10]    0.86    0.01 0.37  0.14  0.61  0.85  1.10  1.62  3070    1

Samples were drawn using NUTS(diag_e) at Sat Mar  7 22:34:15 2020.
For each parameter, n_eff is a crude measure of effective sample size,
and Rhat is the potential scale reduction factor on split chains (at
convergence, Rhat=1).
```

The model has improved its effective sample sizes for `mu`. We extract the parameters for running our posterior predictive checks.

Compare, in figures 3.28 and 3.29, our earlier mcmc information (left, top) showing divergences with our new model (right, bottom).

We review the marginal plot with our estimates against **observed values**, again, in figure 3.30.

Our new model looks good. If we've estimated the building-level means well, then the posterior distribution of means by building

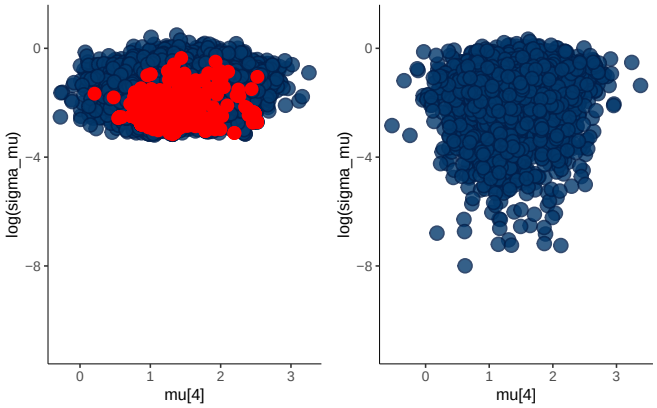


Figure 3.28: The new model (right) improves effective samples sizes over the prior model (left).

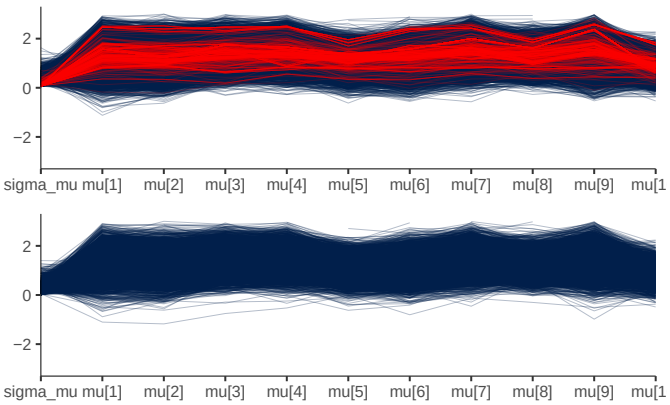


Figure 3.29: Compare both models: the new model has no divergences.

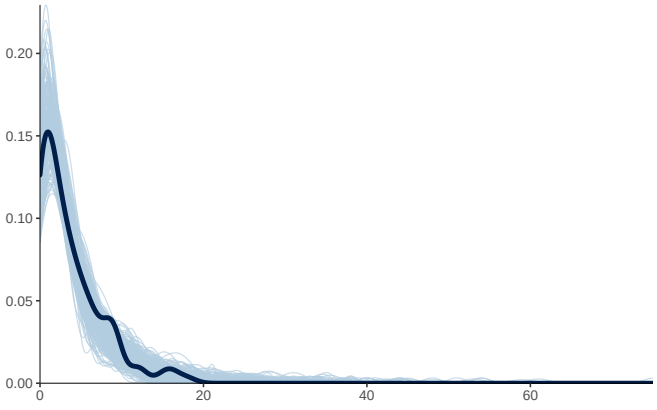


Figure 3.30: Posterior predictions have improved.

should match well with the **observed means** of the quantity of building complaints by month.

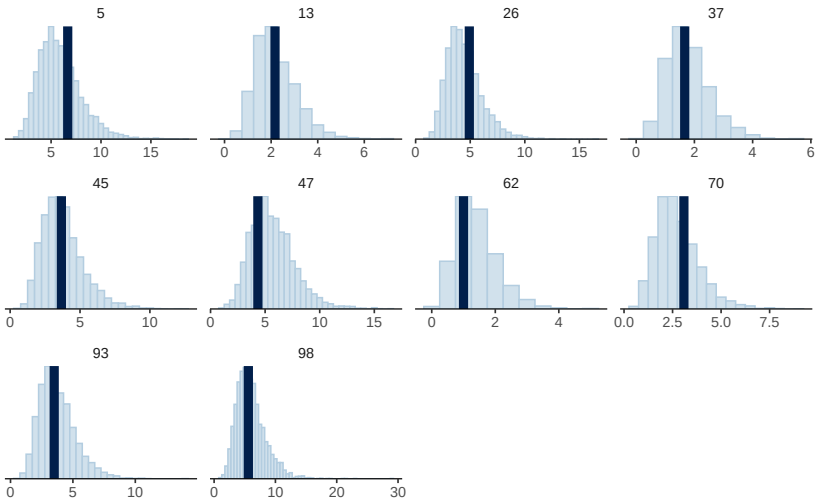


Figure 3.31: Posterior checks by building also have improved.

We weren't terribly off with estimates of the building-specific means before, but now figure 3.31 suggests we have well captured the **observed means** with our model. The model is also able to do a decent job estimating within-building variability, as shown in figure 3.32.

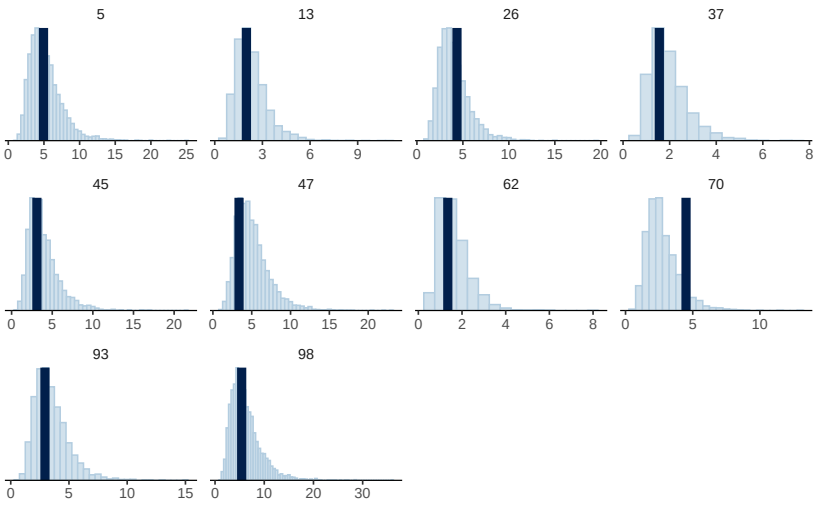


Figure 3.32: The model estimates within-building variability.

Again, we compare predictions to **observed complaints** by number of traps (figure 3.33):

The standardized residuals, figure 3.34, have also shrunk (improved).

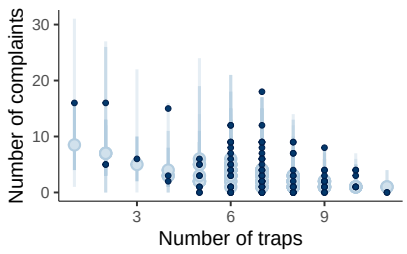


Figure 3.33: Posterior predictive intervals.

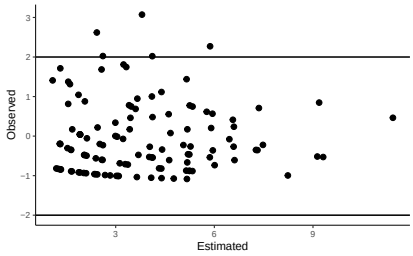


Figure 3.34: Standardized residuals have improved.

Finally, we compare the expected and **measured** values through a rootogram in figure 3.35.

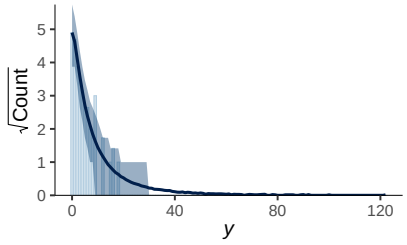


Figure 3.35: Compare expected and measured values.

3.7.6 *Varying intercepts* and *varying slopes*

We have some new data that extends the number of time points of observations for each building. This lets us explore how to expand the model a bit more with varying *slopes* in addition to the varying intercepts and also, later, also model temporal variation.

Perhaps if the levels of complaints differ by building, so does the coefficient for the effect of traps. We can add these varying coefficients to our model and observe the fit.

$$
\begin{aligned}
\text{complaints}_{b,t} &\sim \text{Neg-Binomial}(\lambda_{b,t}, \phi) \\
\lambda_{b,t} &= \exp\left(\eta_{b,t}\right) \\
\eta_{b,t} &= \mu_b + \kappa_b \, \texttt{traps}_{b,t} + \log\_\text{sq\_foot}_b \\
\mu_b &\sim \text{normal}(\alpha + \texttt{building\_data}\,\zeta, \sigma_\mu) \\
\kappa_b &\sim \text{normal}(\beta + \texttt{building\_data}\,\gamma, \sigma_\kappa)
\end{aligned}
$$

Let's compile the model.

R code.

```
comp_model_NB_hier_slopes <-
  stan_model('stan_programs/hier_NB_regression_ncp_slopes_mod.stan')
```

Fit the model to data and extract the posterior draws needed for our posterior predictive checks.

R code.

```
fitted_model_NB_hier_slopes <-
  sampling(comp_model_NB_hier_slopes,
           data = stan_dat_hier,
           chains = 4, cores = 4,
           control = list(adapt_delta = 0.95))
```

To see if the model infers inter-building differences, we can plot a histogram of our marginal posterior distribution for `sigma_kappa`.

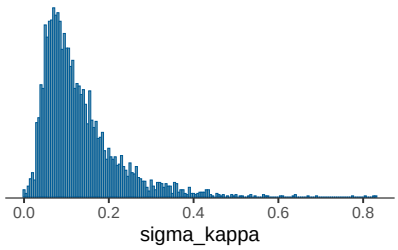


Figure 3.36: Marginal posterior distribution for `sigma_kappa`.

```
Inference for Stan model: hier_NB_regression_ncp_slopes_mod.
4 chains, each with iter=2000; warmup=1000; thin=1;
post-warmup draws per chain=1000, total post-warmup draws=4000.

             mean se_mean   sd  2.5%   25%   50%   75% 97.5% n_eff Rhat
kappa[1]    -0.01    0.00 0.08 -0.14 -0.07 -0.03  0.03  0.17   770 1.01
kappa[2]    -0.42    0.00 0.10 -0.64 -0.48 -0.42 -0.35 -0.25  1661 1.00
kappa[3]    -0.59    0.00 0.11 -0.81 -0.66 -0.59 -0.51 -0.39  5266 1.00
kappa[4]    -0.22    0.00 0.07 -0.36 -0.27 -0.22 -0.18 -0.09  3892 1.00
kappa[5]    -0.60    0.00 0.09 -0.78 -0.66 -0.60 -0.54 -0.42  5201 1.00
kappa[6]    -0.44    0.00 0.10 -0.67 -0.50 -0.43 -0.37 -0.25  2823 1.00
kappa[7]    -0.31    0.00 0.07 -0.44 -0.36 -0.31 -0.26 -0.18  5364 1.00
kappa[8]    -0.23    0.00 0.15 -0.56 -0.32 -0.22 -0.13  0.04  2269 1.00
kappa[9]     0.08    0.00 0.06 -0.03  0.04  0.08  0.12  0.20  5536 1.00
kappa[10]   -0.72    0.00 0.16 -1.00 -0.82 -0.73 -0.62 -0.38  1140 1.00
beta        -0.35    0.00 0.07 -0.48 -0.38 -0.35 -0.31 -0.22  2193 1.00
alpha        1.41    0.01 0.32  0.73  1.22  1.42  1.61  2.00  2139 1.00
phi          1.61    0.00 0.19  1.27  1.48  1.60  1.73  2.02  4198 1.00
sigma_mu     0.52    0.02 0.44  0.02  0.18  0.41  0.74  1.62   493 1.01
sigma_kappa  0.13    0.00 0.09  0.03  0.07  0.11  0.16  0.37   476 1.01
mu[1]        0.26    0.03 0.76 -1.54 -0.15  0.37  0.78  1.47   724 1.01
mu[2]        1.67    0.01 0.53  0.72  1.30  1.63  1.99  2.82  1567 1.00
mu[3]        2.13    0.00 0.33  1.52  1.91  2.13  2.35  2.81  4959 1.00
mu[4]        1.50    0.01 0.51  0.50  1.18  1.50  1.80  2.56  3883 1.00
mu[5]        2.39    0.01 0.42  1.60  2.11  2.38  2.67  3.22  5774 1.00
mu[6]        1.91    0.01 0.38  1.21  1.67  1.88  2.13  2.79  2666 1.00
mu[7]        2.68    0.00 0.26  2.19  2.50  2.66  2.85  3.20  4816 1.00
mu[8]       -0.51    0.02 0.96 -2.25 -1.15 -0.58  0.06  1.54  2538 1.00
mu[9]        0.21    0.01 0.57 -0.92 -0.16  0.21  0.59  1.31  5593 1.00
mu[10]       1.79    0.04 1.13 -0.85  1.20  1.96  2.56  3.52   837 1.01

Samples were drawn using NUTS(diag_e) at Sat Mar  7 22:34:54 2020.
```

```
For each parameter, n_eff is a crude measure of effective sample size,
and Rhat is the potential scale reduction factor on split chains (at
convergence, Rhat=1).
```

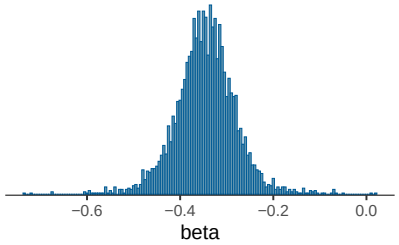


Figure 3.37: Marginal posterior distribution for `beta`.

While the model can't specifically rule out zero from the posterior, it does have mass at small non-zero numbers, so we should leave in the hierarchy over `kappa`. Plotting the marginal data density again, the model still looks well calibrated.

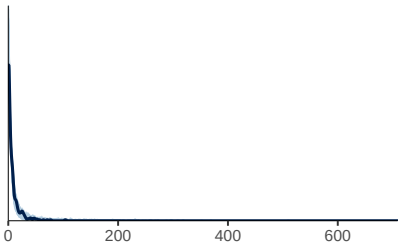


Figure 3.38: Marginal density estimates compared with observations.

3.8 *Time varying effects and structured priors*

We haven't looked at how complaints change over time. Let's look at whether there's any pattern in our estimates to **observed** values for each month (over time). We show that in figure 3.39.

We might augment our model with a log-additive monthly effect, $\texttt{mo}_t$.

$$\eta_{b,t} = \mu_b + \kappa_b \, \texttt{traps}_{b,t} + \texttt{mo}_t + \text{log\_sq\_foot}_b$$

We have complete freedom over how to specify the prior for $\texttt{mo}_t$. There are several competing factors for how the number of complaints might change over time. It makes sense that there might be more noise in the environment during the summer, but we might also expect that there is more noise control in the summer as well. Given that we're modeling complaints, maybe after the first sighting

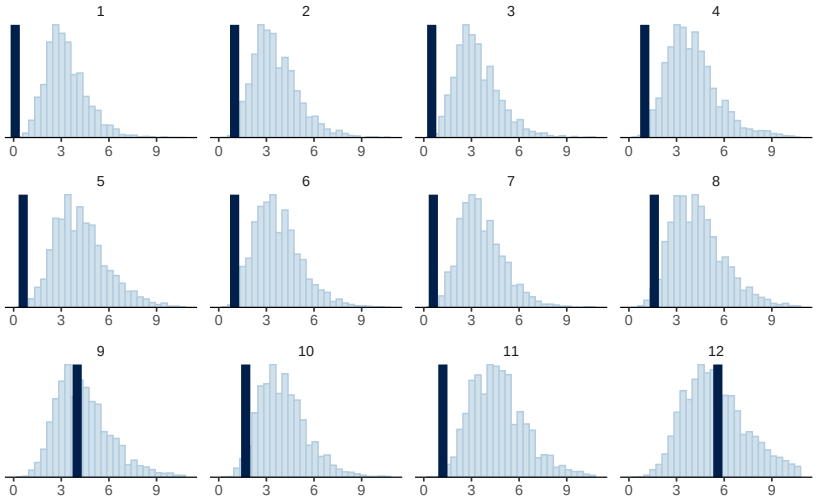


Figure 3.39: Estimates grouped by building.

of noise in a building, residents are more vigilant, and thus complaints of noise would increase. This can be a motivation for using an autoregressive prior for our monthly effects. The model is as follows:

$$\begin{aligned} \texttt{mo}_t &\sim \text{normal}(\rho\, \texttt{mo}_{t-1}, \sigma_{\texttt{mo}}) \\ &\equiv \\ \texttt{mo}_t = \rho\, \texttt{mo}_{t-1} + \epsilon_t, \quad \epsilon_t &\sim \text{normal}(0, \sigma_{\texttt{mo}}) \\ \rho &\in [-1, 1] \end{aligned}$$

This equation says that the monthly effect in month t is directly related to the last month's monthly effect. Given the description of the process above, it seems like there could be either positive or negative associations between the months, but there should be a bit more weight placed on positive ρs, so we'll put an informative prior that pushes the parameter ρ towards 0.5.

Before we write our prior, however, we have a problem: Stan doesn't implement any densities that have support on $[-1, 1]$. We can use variable transformation of a raw variable defined on $[0, 1]$ to give us a density on $[-1, 1]$. Specifically,

$$\begin{aligned} \rho_{\text{raw}} &\in [0, 1] \\ \rho &= 2 \cdot \rho_{\text{raw}} - 1 \end{aligned}$$

Then we put a beta prior on ρ_{raw} to push our estimate near 0.5.

One further wrinkle is that we have a prior for $\texttt{mo}_t$ that depends on $\texttt{mo}_{t-1}$. That is, we are working with the *conditional* distribution of $\texttt{mo}_t$ given $\texttt{mo}_{t-1}$. But what should we do about the prior for $\texttt{mo}_1$, for

which we don't have a previous time period in the data?

We need to work out the *marginal* distribution of the first observation. Thankfully we consider that AR models are stationary, so $\text{Var}(\texttt{mo}_t) = \text{Var}(\texttt{mo}_{t-1})$ and $\text{E}(\texttt{mo}_t) = \text{E}(\texttt{mo}_{t-1})$ for all t. Therefore the marginal distribution of $\texttt{mo}_1$ is the same as the marginal distribution of any $\texttt{mo}_t$.

First we derive the marginal variance of $\texttt{mo}_t$.

$$\begin{aligned}\text{Var}(\texttt{mo}_t) &= \text{Var}(\rho \texttt{mo}_{t-1} + \epsilon_t) \\ \text{Var}(\texttt{mo}_t) &= \text{Var}(\rho \texttt{mo}_{t-1}) + \text{Var}(\epsilon_t)\end{aligned}$$

where the second line holds by independence of ϵ_t and ϵ_{t-1}). Then, using the fact that $Var(cX) = c^2 Var(X)$ for a constant c and the fact that, by stationarity, $\text{Var}(\texttt{mo}_{t-1}) = \text{Var}(\texttt{mo}_t)$, we then obtain:

$$\begin{aligned}\text{Var}(\texttt{mo}_t) &= \rho^2 \text{Var}(\texttt{mo}_t) + \sigma^2_{\texttt{mo}} \\ \text{Var}(\texttt{mo}_t) &= \frac{\sigma^2_{\texttt{mo}}}{1 - \rho^2}\end{aligned}$$

For the mean of $\texttt{mo}_t$ things are a bit simpler:

$$\begin{aligned}\text{E}(\texttt{mo}_t) &= \text{E}(\rho\, \texttt{mo}_{t-1} + \epsilon_t) \\ \text{E}(\texttt{mo}_t) &= \text{E}(\rho\, \texttt{mo}_{t-1}) + \text{E}(\epsilon_t)\end{aligned}$$

Since $\text{E}(\epsilon_t) = 0$ by assumption we have

$$\begin{aligned}\text{E}(\texttt{mo}_t) &= \text{E}(\rho\, \texttt{mo}_{t-1}) + 0 \\ \text{E}(\texttt{mo}_t) &= \rho\, \text{E}(\texttt{mo}_t) \\ \text{E}(\texttt{mo}_t) - \rho \text{E}(\texttt{mo}_t) &= 0 \\ \text{E}(\texttt{mo}_t) &= 0/(1 - \rho)\end{aligned}$$

which for $\rho \neq 1$ yields $\text{E}(\texttt{mo}_t) = 0$.

We now have the marginal distribution for $\texttt{mo}_t$, which, in our case, we will use for $\texttt{mo}_1$. The full AR(1) specification is then:

$$\begin{aligned}\texttt{mo}_1 &\sim \text{normal}\left(0, \frac{\sigma_{\texttt{mo}}}{\sqrt{1 - \rho^2}}\right) \\ \texttt{mo}_t &\sim \text{normal}\left(\rho\, \texttt{mo}_{t-1}, \sigma_{\texttt{mo}}\right) \forall t > 1\end{aligned}$$

We compile the model: R code.

```
comp_model_NB_hier_mos <-
  stan_model('stan_programs/hier_NB_regression_ncp_slopes_mod_mos.stan')
```

```
fitted_model_NB_hier_mos <-
  sampling(comp_model_NB_hier_mos,
           data = stan_dat_hier,
           chains = 4, cores = 4,
           control = list(adapt_delta = 0.9))
```

Given time constraints, we won't go on expanding the model for now. Questions remain: what other information would help us understand the data generating process better? What other aspects of the data generating process might we still want to capture? As usual, we run through our posterior predictive checks, comparing the density of our estimates to **observed** values.

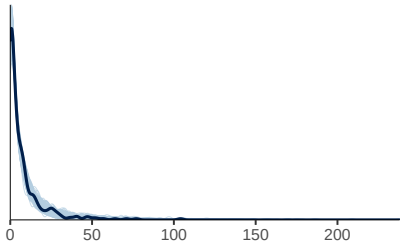


Figure 3.40: Posterior predictive checks.

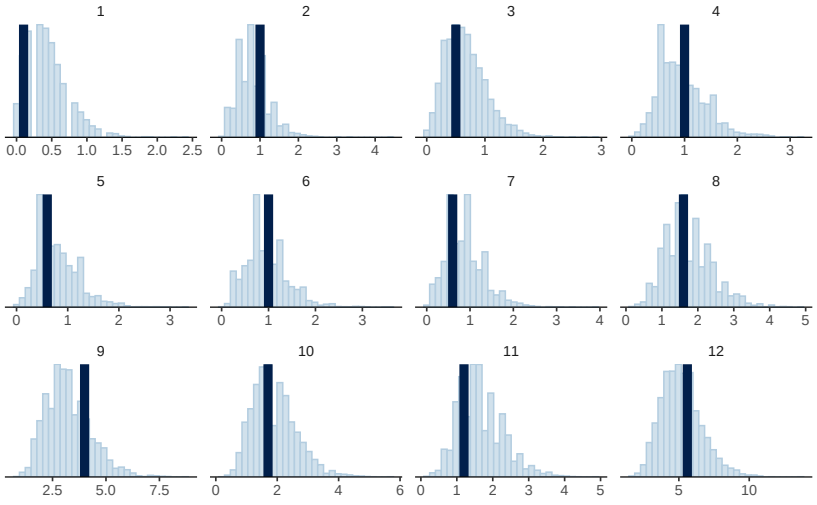


Figure 3.41: Posterior predictive checks.

Our monthly random intercept has better estimate a monthly pattern across all the buildings (figure 3.41). We can also compare the prior and posterior for the autoregressive parameter to see how much we've learned. Figures 3.42 and 3.43 shows two different ways of comparing the prior and posterior visually.

Our paramter estimates include,

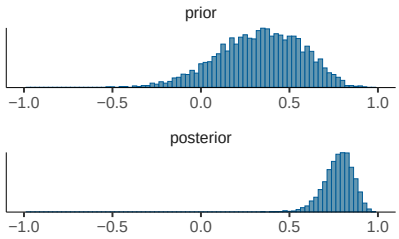


Figure 3.42: Compare draws from prior and draws from posterior.

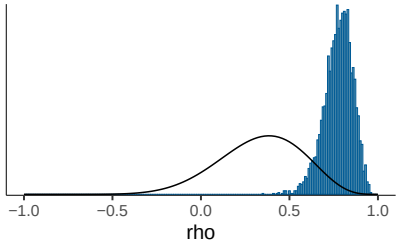


Figure 3.43: Overlay prior density curve on posterior draws

```
Inference for Stan model: hier_NB_regression_ncp_slopes_mod_mos.
4 chains, each with iter=2000; warmup=1000; thin=1;
post-warmup draws per chain=1000, total post-warmup draws=4000.

             mean se_mean   sd  2.5%   25%   50%   75% 97.5% n_eff Rhat
rho          0.78    0.00 0.08  0.60  0.73  0.78  0.83  0.91  1545    1
sigma_mu     0.32    0.01 0.24  0.02  0.13  0.27  0.45  0.90  1314    1
sigma_kappa  0.09    0.00 0.06  0.01  0.05  0.08  0.11  0.24   979    1
gamma[1]    -0.18    0.00 0.10 -0.38 -0.25 -0.18 -0.12  0.03  2250    1
gamma[2]     0.12    0.00 0.08 -0.03  0.07  0.11  0.16  0.28  1668    1
gamma[3]     0.11    0.00 0.15 -0.20  0.02  0.11  0.20  0.42  1704    1
gamma[4]    -0.01    0.00 0.07 -0.15 -0.04  0.00  0.03  0.12  1182    1

Samples were drawn using NUTS(diag_e) at Sat Mar  7 22:35:37 2020.
For each parameter, n_eff is a crude measure of effective sample size,
and Rhat is the potential scale reduction factor on split chains (at
convergence, Rhat=1).
```

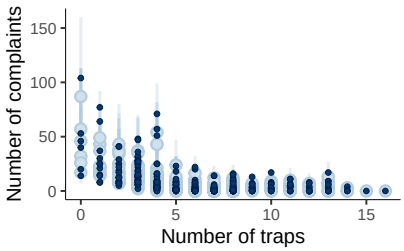


Figure 3.44: Intervals for posterior predictive checks.

It looks as if, per figure 3.44, our model finally estimates a reasonable posterior predictive distribution compared with **observed values** for all given numbers of sound traps, and appropriately captures the tails of the data generating process.

3.9 *Using our model for decisions: Cost forecasts*

OUR MODEL seems to be fitting well, so now we use the model to help us make decisions about how many traps to put in our buildings. We'll make a forecast for 6 months forward. First we code the model and compile it into R object `comp_rev`.

Stan code.

```
functions {
  int neg_binomial_2_log_safe_rng(real eta, real phi) {
    real gamma_rate = gamma_rng(phi, phi / exp(eta));
    if (gamma_rate >= exp(20.79))
      return -9;
    return poisson_rng(gamma_rate);
  }
}
data {
  int<lower=1> N;
  int<lower=0> complaints[N];
  vector<lower=0>[N] traps;

  vector[N] log_sq_foot;

  int<lower=1> K;
  int<lower=1> J;
  int<lower=1, upper=J> building_idx[N];
  matrix[J,K] building_data;

  int<lower=1> M;
  int<lower=1,upper=M> mo_idx[N];

  int<lower=1> M_forward;
  vector[J] log_sq_foot_pred;
}
transformed data {
  int N_hypo_traps = 21;
  int hypo_traps[N_hypo_traps];
  for (i in 1:N_hypo_traps)
    hypo_traps[i] = i - 1;
}
parameters {
  real<lower=0> inv_phi;
  vector[J] mu_raw;
  real<lower=0> sigma_mu;
  real alpha;
  vector[K] zeta;
  vector[J] kappa_raw;
  real<lower=0> sigma_kappa;
  real beta;
  vector[K] gamma;
  vector[M] mo_raw;
  real<lower=0> sigma_mo;
  real<lower=0,upper=1> rho_raw;
}
transformed parameters {
  real phi = inv(inv_phi);

  vector[J] mu = alpha +
                 building_data * zeta +
```

```
                sigma_mu * mu_raw;

  vector[J] kappa = beta +
                    building_data * gamma +
                    sigma_kappa * kappa_raw;

  real rho = 2 * rho_raw - 1;
  vector[M] mo = sigma_mo * mo_raw;
  mo[1] /= sqrt(1 - rho^2);
  for (m in 2:M) {
    mo[m] += rho * mo[m-1];
  }
}
model {
  inv_phi ~ normal(0, 1);
  kappa_raw ~ normal(0,1) ;
  sigma_kappa ~ normal(0, 1);
  beta ~ normal(-0.25, 1);
  gamma ~ normal(0, 1);
  mu_raw ~ normal(0,1) ;
  sigma_mu ~ normal(0, 1);
  alpha ~ normal(log(4), 1);
  zeta ~ normal(0, 1);
  mo_raw ~ normal(0,1);
  sigma_mo ~ normal(0, 1);
  rho_raw ~ beta(10, 5);

  {
  vector[N] eta = mu[building_idx] +
                  kappa[building_idx] .* traps +
                  mo[mo_idx] +
                  log_sq_foot;

  complaints ~ neg_binomial_2_log(eta, phi);
  }
}
generated quantities {
  int y_pred[J,N_hypo_traps];
  matrix[J,N_hypo_traps] rev_pred;

  for (j in 1:J) {
    for (i in 1:N_hypo_traps) {
      int y_pred_by_month[M_forward];
      vector[M_forward] mo_forward;

      mo_forward[1] = normal_rng(rho * mo[M], sigma_mo);

      for (m in 2:M_forward)
        mo_forward[m] = normal_rng(rho * mo_forward[m-1], sigma_mo);

      for (m in 1:M_forward) {
        real eta = mu[j] +
                   kappa[j] * hypo_traps[i] +
                   mo_forward[m] +
                   log_sq_foot_pred[j];

        y_pred_by_month[m] = neg_binomial_2_log_safe_rng(eta, phi);
      }
      y_pred[j,i] = sum(y_pred_by_month);
      rev_pred[j,i] = -10 * y_pred[j,i];
    }
```

```
  }
}
```

An important input to the revenue model is how much revenue is lost due to each complaint. The client has a policy that for every 10 complaints, they'll call an exterminator costing the client $100, so that'll amount to $10 per complaint.

R code.

```
rev_model <-
  sampling(comp_rev,
           data = stan_dat_hier,
           cores = 4, chains = 4,
           control = list(adapt_delta = 0.9))
```

Below we've generated revenue curves for the buildings. These charts give us precise quantification of our uncertainty around our revenue projections at any number of traps for each building.

A key input to our analysis will be the cost of installing traps. We're simulating the number of complaints we receive over the course of a year, so we need to understand the cost associated with maintaining each trap over the course of a year. There's the cost attributed to the raw trap and related material. The cost of maintaining one trap for a year plus monthly replenishment of the material is about $20.

R code.

```
N_traps <- 20
costs <- 10 * (0:N_traps)
```

We'll also need labor for maintaining the traps, which need to be serviced every two months. If there are fewer than five traps, our in-house maintenance staff can manage the stations (about one hour of work every two months at $20/hour), but above five traps we need to hire outside noise control to help out. They're a bit more expensive, so we've put their cost at $30 / hour. Each five traps should require an extra person-hour of work, so that's factored in as well. The marginal person-person hours above five traps are at the higher noise-control labor rate.

R code.

```
N_months_forward <- 12
N_months_labor <- N_months_forward / 2
hourly_rate_low <- 20
hourly_rate_high <- 30
costs <- costs +
  (0:N_traps < 5 & 0:N_traps > 0) *
  (N_months_labor * hourly_rate_low) +
  (0:N_traps >= 5 & 0:N_traps < 10) *
  (N_months_labor * (hourly_rate_low + 1 * hourly_rate_high)) +
  (0:N_traps >= 10 & 0:N_traps < 15) *
  (N_months_labor * (hourly_rate_low + 2 * hourly_rate_high)) +
  (0:N_traps >= 15) *
  (N_months_labor * (hourly_rate_low + 3 * hourly_rate_high))
```

Figure 3.45 provides with number of traps on the x-axis and profit/loss

forecasts and uncertainty intervals on the y-axis.

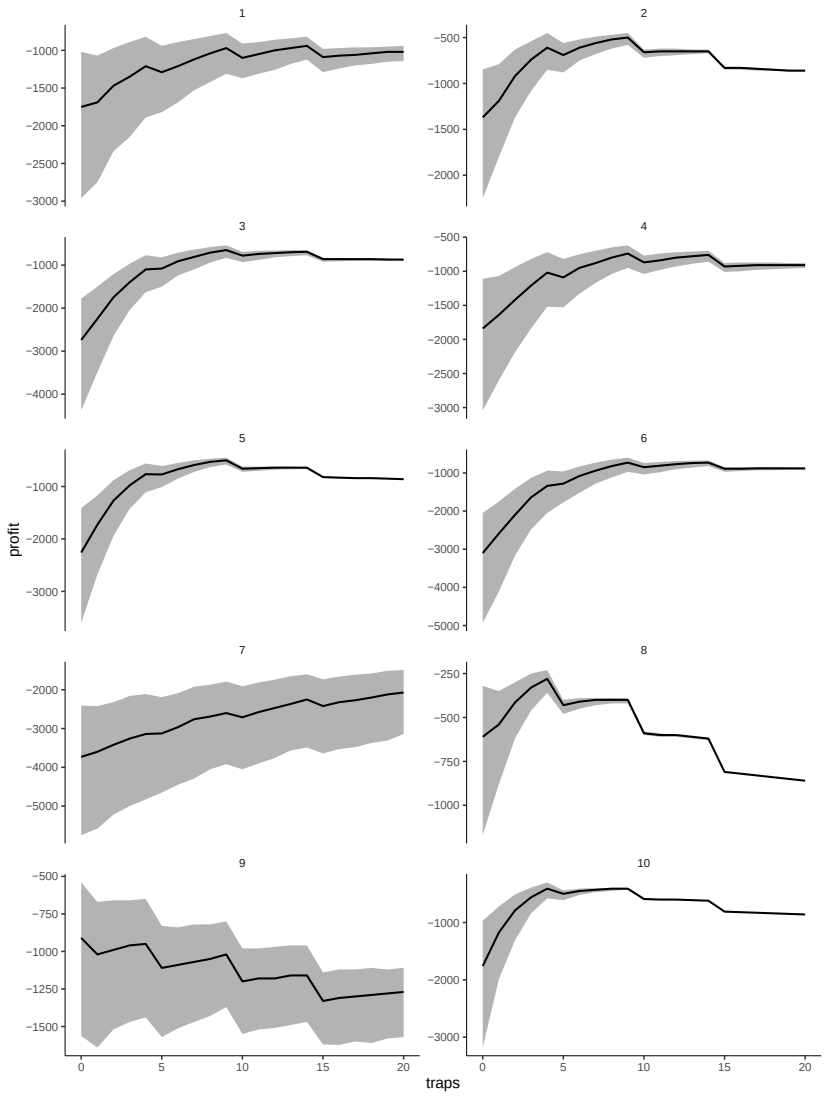


Figure 3.45: Each curve suggests the optimal number of sound traps in each building.

The optimal number of traps differs by building.

3.10 Next steps

Questions we may want to consider: how would we build a revenue curve for a new building? Let's say our utility function is revenue. If we wanted to maximize expected revenue, we can take expectations at each station count for each building and choose the trap numbers that maximizes expected revenue. This will be called

a maximum revenue strategy. How can we generate the distribution of portfolio revenue (the sum of revenue across all the buildings) under the maximum revenue strategy from the draws of `rev_pred` we already have?

4
Purpose, audience, and craft

As we prepare to scope, and work through, a data analytics project, we must communicate variously if it is to have value. The reproducible workflow shown in the last chapter at least provides value as a communication to its immediate audience — its authors, as a reference for what they accomplished — and to those with similar needs[1] to understand and critique the logical progression and scope of the project. That form of communiation, however, will be less valuable than other communication forms for different audiences and purposes. Given the information compiled from our project — the content — we now consider communicating various aspects for other purposes and audiences.

[1] Those with similar needs are similar to the intended audience for Joao Caldeira et al., "Improving Traffic Safety Through Video Analysis in Jakarta, Indonesia," in *Nd Conference on Neural Information Processing Systems NeurIPS*, 2018, 1–5.

The importance of adjusting communication is not unique to data analytics. Let's consider, say, how communication form differs when for a news story versus an op-ed. Long-time editor of the op-ed at the *New York Times* explains,

> The approach to argument that I learned in classes at Berkeley was much more similar to an op-ed than the inverted pyramid of daily journalism or the slow, anecdotal flow of feature stories that had dominated my professional life[2].

[2] Trish Hall, *Writing to Persuade: How to Bring People over to Your Side*, First edition (New York: Liveright Publishing Corporation, a division of W.W. Norton & Company, 2019).

The qualities of an op-ed piece must be, she writes: "surprising, concrete, and persuasive." These qualities are similar to that we need in business communication, which generally drive decisions. All business writing begins with a) identifying the purpose for communicating and b) understanding your audiences' scopes of knowledge and responsibilities in the problem context. Neither is trivial; both require research. To motivate this discussion, let's consider and deconstruct two example memos — one for Citi Bike, the other for the Dodgers — written for data science projects. It will be helpful for this exercise to place both memos side-by-side for comparison as we work through them below.

4.1 Communication structure

Let's begin discussing the communication structure from several perspectives: purpose, narrative structure, sentence structure, and effective redundancy through heirarchy. Then, we consider audiences, story, and the importance of revision.

4.1.1 Purpose and audience

IN THE FIRST EXAMPLE, we return to Citi Bike. After project ideation, and scoping, we want to ask Citi Bike's head of data analytics to let us write a more detailed proposal to conduct the data analytics project. We accomplish this in 250 words. The title and body of the fully composed memo, section 4.9, is in example 4.1:

> **Example 4.1** (Citi Bike 250-word memo)**. To inform rebalancing, let's explore docking and bike availability in the context of subway and weather information.**
>
> We should explore station and ride data in the context of subway and weather information to gain insight for "rebalancing," what our Dani Simmons explains is "one of the biggest challenges of any bike share system, especially in … New York where residents don't all work a traditional 9-5 schedule, and though there is a Central Business District, it's a huge one and people work in a variety of other neighborhoods as well."
>
> A rebalancing study by Columbia University Center for Spatial Research[3] previously identified trends in bike usage using heatmaps. As those visualizations did not combine dimensions of space and time, which will be helpful to uncover trends in bike and station availability by neighborhood throughout a day, we can begin our analysis there.
>
> NYC OpenData and The Open Bus Project provide published date, time, station ID, and ride instances for all our docking stations and bikes since we began service. To begin our project, we can visually explore the intersection of trends in both time and location with this data to understand problematic neighborhoods and, even, individual stations, using up-to-date information.
>
> Then, we will build upon the initial work, exploring causal factors such as the availability of alternative transportation (e.g., subway stations near docking stations) and weather. Both of which, we have available data that can be joined using timestamps.
>
> The project aligns with our goals to, in Simmons's words, "be innovative in how we meet this challenge." Let's draft a detailed proposal.

[3] Juan Francisco Saldarriaga, "CitiBike Rebalancing Study" (Spatial Information Design Lab, Columbia University, 2013).

It begins, in the title of this memo, with our purpose of writing, to conduct data analysis on specifically identified data to inform the issue of rebalancing, one of Citi Bike's goals:

> To inform rebalancing, let's explore docking and bike availability in the context of subway and weather information.

This is what Doumont[4] calls a **message**. We should craft communica-

[4] Doumont, *Trees, Maps, and Theorems*.

tions with messages, *not* merely information. Doumont explains that a message differs from raw information in that it presents "intelligent added value," that is, something to understand about the information. A message *interprets* the information for a specific audience and for a specific purpose. It conveys the *so what*, whereas information merely conveys the what. What makes our title a message? Before answering this, let's compare one of Doumont's examples of information to that of a message. This sentence is mere information:

> A concentration of 175 μg per m^3 has been observed in urban areas.

A message, in contrast to information, would be the *so what*:

> The concentration in urban areas (175 $\mu g/m^3$) is unacceptably high.

In our title, we request an action, approval for the *exploratory analysis on specified data*, for a particular purpose, *to inform rebalancing*. This purpose also implies the *so what*: unavailable bikes or docking slots, unbalanced stations, are bad. We're asking to help remedy the issue.

This beginning, if effective, is only because we wrote it for a particular audience. Our audience is head of data analytics at Citi Bike, and presumably knows the problem of rebalancing; it is well-known in, and beyond, the organization. Thus, our sentence implicitly refers back to information our audience already knows[5]. Relying on his or her knowledge means we do not need to first explain what rebalancing is or why it is a problem.

Let's introduce a second example before digging further into the structure of the *Citi Bike* memo. Having multiple examples to analyze has the added benefit of allowing us to induce some general, but effective, writing principles.

Professional teams in the sport of baseball, including the *Los Angeles Dodgers*, make strategic decisions within the boundaries of the sport's rules for the purpose of winning games. One of those rules involves *stealing bases*, as in figure 4.1. This concept is part of our next example, written to the Los Angeles Dodgers's Director of Quantitative Analytics, Scott Powers. The title and body of the fully formatted memo, section 4.9, is shown in example 4.2.

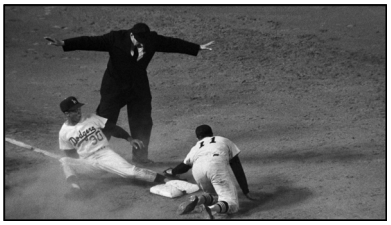


Figure 4.1: In a close call, the baseball umpire spread his arms, signaling that the Dodgers baserunner successfully ran to the base faster than the catcher could throw the ball to the base to get him out — the runner stole the base. Knowing when to try to steal is strategic and depends on other events that are at least partly measured as data.

[5] We will discuss this structure — old before new — in detail later.

Figure 4.2: The intended audience for this memo, Dr. Scott Powers, is Director of Quantitative Analytics at the Los Angeles Dodgers. He earned his doctor of philosophy in statistics from Stanford, has authored publications in machine learning, knows R programming, and as an employee of the Dodgers, knows their history. Powers manages a team of data scientists; their responsibilities include assessing player and team performance.

Example 4.2 (Dodgers 250-word memo)**. Our game decisions should optimize expectations. Let's test the concept by modeling decisions to steal.**

Our Sandy Koufax pitched a perfect game, the most likely event sequence, only once: those, we do not expect or plan. Since our decisions based on other most likely events don't align with expected outcomes, we leave wins unclaimed. To claim them, let's base decisions on expectations flowing from decision theory and probability models. A joint model of all events works best, but we can start small with, say, decisions to steal second base.

> After defining our objective (e.g. optimize expected runs) we will, from Statcast data, weight everything that could happen by its probability and accumulate these probability distributions. Joint distributions of all events, an eventual goal, will allow us to ask counterfactuals — "what if we do *this*" or "what if our opponent does *that*" — and simulate games to learn how decisions change win probability. It enables optimal strategy.
>
> Rational and optimal, this approach is more efficient for gaining wins. For perspective, each added win from the free-agent market costs 10 million, give or take, and the league salary cap prevents unlimited spend on talent. There is no cap, however, on investing in rational decision processes.
>
> Computational issues are being addressed in `Stan`, a tool that enables inferences through advanced simulations. This open-source software is free but teaching its applications will require time. To shorten our learning curve, we can start with Stan interfaces that use familiar syntax (like `lme4`) but return joint probability distributions: `R` packages `rethinking`, `brms`, or `rstanarm`. Perfect games aside, we can test the concept with decisions to steal.

The beginning of this memo reads:

> Our game decisions should optimize expectations. Let's test the concept by modeling decisions to steal.

In the first three words, the subject of this sentence, signals to our audience something they have experience with: *our game decisions*. It's familiar. Then, we provide a call to action: our game decisions *should optimize expectations. Let's test the concept by modeling decisions to steal*. As with the Citi Bike memo, the Dodgers memo begins with a message and stated purpose. And as with Citi Bike, the information in the title message of the Dodgers memo is shared knowledge with our audience. In fact, we rely on the educational background of our audience, who we know has earned a doctor of philosophy in statistics, when including the concept to "optimize expectations"[6] without first explaining what that is because we know, or from the audience's background, can assume the audience understands the concept.

So in both cases, we have begun with language and topics already *familiar to the audience*, which follows the more general writing advice from Doumont[7], who instructs us to

> put ourselves in the shoes of the audience, anticipating their situation, their needs, their expectations. Structure the story along their line of reasoning, recognizing the constraints they might bring: their familiarity with the topic, their mastery of the language, the time they can free for us.

What else do we know about chief analytics officers in general? Their jobs require them to be efficient with their time. Thus, by starting with our purpose, letting them know what we want them to do, we are considerate of their "constraints" and "time they can free for us."

[6] An *expectation* is specifically defined in probability theory. To *optimize* is also a specific mathematical concept.

[7] Doumont, *Trees, Maps, and Theorems*.

Beginning with a purpose and call-to-action also allow the executive to understand the memo's relevance *to them*, in terms of their decision-making, immediately; they have a reason to continue reading.

The persuasive power of beginning with the main message for your audience, or issue relevant to your audience, is nearly as timeless as it is true. Cicero, the Roman philosopher with a treatise on rhetoric, explained that we must not begin with details because "it forms no part of the question, and men are at first desirous to learn the very point that is to come under their judgment."[8]

Next, let's review the structure of these memos to see whether we've "structur[ed] the story along their line of reasoning."

[8] Marcus Tullius Cicero and J. S. Watson, *Cicero on oratory and orators*, Landmarks in rhetoric and public address (Carbondale: Southern Illinois University Press, 1986).

4.1.2 Common ground

Let's compare the first sentences of the body of both examples. The Citi Bike memo begins,

> We should explore station and ride data in the context of subway and weather information to gain insight for "rebalancing," what our Dani Simmons explains is "one of the biggest challenges of any bike share system, especially in … New York where residents don't all work a traditional 9-5 schedule, and though there is a Central Business District, it's a huge one and people work in a variety of other neighborhoods as well."

This sentence starts with the title request, and then ties the purpose — rebalancing — to corporate goals. It does so by quoting the company's spokesperson, which serves as both evidence of the *so what*. Offering, and accepting, Simmons's quote serves a second purpose in writing: it helps to establish **common ground** with our audience.

If we want to affect the behaviors and beliefs of the person in front of us, we need to first understand what goes on inside their head and establish common ground. Why? When you provide someone with new data, they quickly accept evidence that confirms their preconceived notions (what are known as prior beliefs) and assess counterevidence with a critical eye[9]. Four factors come into play when people form a new belief: our old belief (this is technically known as the "prior"), our confidence in that old belief, the new evidence, and our confidence in that evidence. Focusing on what you and your audience have in common, rather than what you disagree about, enables change. Let's check for common ground in the Dodgers memo. The first sentence of the body begins,

[9] Tali Sharot, *The Influential Mind*, What the Brain Reveals About Our Power to Change Others (Henry Holt and Company, 2017).

> Our Sandy Koufax pitched a perfect game, the most likely event sequence, only once: those, we do not expect or plan.

Sandy Koufax is one of the most successful Dodgers players in the history of the franchise. He is one of less than 20 pitchers in the history of baseball to pitch a *perfect game*, something extraordinary. Our audience, as an employee of the Dodgers, will be familiar with this history. It is also something very positive — and shared — between author and audience. It helps to establish common ground, in two ways. Along with that positive, shared history, it sets up an example of a statistical mode, one that we know the audience would agree is unhelpful for planning game strategy because it is too rare, even if it is a statistical *mode*. It helps to create common ground or agreement that it may not be best to use statistical modes for making decisions.

In both memos, we are also trying to use an interesting fact that may be unexpected or surprising in this context (Sandy Koufax, Dani Simmons) to grab our audience's attention. In journalism, this is one way to create *the lead*. William Zinsser[10] explains that the most important sentence in any communication is the first one. If it doesn't induce the reader to proceed to the second sentence, your communication is dead. And if the second sentence doesn't induce the audience to continue to the third sentence, it's equally dead. Readers want to know — very soon — what's in it for them.

Your lead must capture the audience immediately cajoling them with freshness, or novelty, or paradox, or humor, or surprise, or with an unusual idea, or an interesting fact, or a question. Next, it must provide hard details that tell the audience why the piece was written and why they ought to read it.

[10] William Zinsser, *On Writing Well*, Sixth, The Classic Guide to Writing Nonfiction (Harper Resource, 2001).

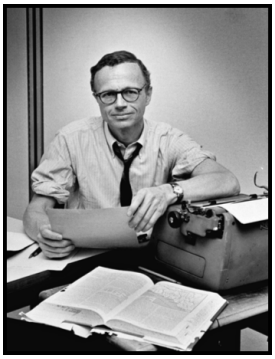


Figure 4.3: William Zinsser: A longtime teacher of writing at Columbia and Yale, the late professor and journalist is well-known for putting pen to paper, or finger to key, as the case may be.

4.1.3 Details

At this point in both memos, we have begun our memo with information familiar to our audience, relevant to their job in decision-making, and established our purpose. We have also started with information they would agree with. We've created common ground. The stage is set. What's next? Here's the next two sentences in the body of the Citi Bike memo:

> A rebalancing study[11] by Columbia University Center for Spatial Research previously identified trends in bike usage using heatmaps. As those visualizations did not combine dimensions of space and time, which will be helpful to uncover trends in bike and station availability by neighborhood throughout a day, we can begin our analysis there.

[11] Saldarriaga, "CitiBike Rebalancing Study."

The first sentence introduces previous work — background — on rebalancing studies and its limitations, and we proposed to start where the prior work stopped. This accomplishes two objectives. First, it helps our audience understand beginning details of our proposed project. Second, it helps the audience see that our proposed work is

not redundant to what we already know. Thus, we began the details of our proposed solution. What is described in the Dodgers memo at a similar point? This:

> To claim them, let's base decisions on expectations flowing from decision theory and probability models. A joint model of all events works best, but we can start small with, say, decisions to steal second base.

As with Citi Bike, the next two sentences start introducing details of the proposed project.

After introducing the nature of the proposed project in both memos, we identify data that makes the proposed project feasible. In the Citi Bike memo we identify specific categories of data and the publicly available source of those data:

> **NYC OpenData** and **The Open Bus Project** provide published **date**, **time**, **station ID**, and **ride instances** for all our **docking stations** and **bikes** since we **began service**.

Similarly, in the Dodgers memo,

> After defining our objective (e.g. optimize expected runs) we will, from **Statcast** data, weight everything that could happen by its probability and accumulate these probability distributions.

It may seem we are less descriptive of the data than in the Citi Bike memo, but the label "Statcast" signals to our particular audience a group of specific, publicly available variables collected by the Statcast system[12]. After identifying data, we explain *how* we plan its analysis.

[12] Daren Willman, "Statcast Search CSV Documentation" (MLB Advanced Media, n.d.); Daren Willman, "Standard Statistics" (MLB Advanced Media, 2020).

Having identified data, both memos then describe more details of our proposed methodology. In Citi Bike, we discuss two stages. We plan to graphically explore specific variables in search of specific trends first.

> To begin our project, we can visually explore the intersection of trends in both time and location with this data to understand problematic neighborhoods and, even, individual stations, using up-to-date information.

Then, we specifically identify additional data we plan to join and explore as causal factors for problem areas:

> Then, we will build upon the initial work, exploring causal factors such as the availability of alternative transportation (e.g., subway stations near docking stations) and weather. Both of which, we have available data that can be joined using timestamps.

Similarly, in the Dodgers memo, go into the planned methodology. We plan to model expectations from the data:

> . . . from Statcast data, weight everything that could happen by its probability and accumulate these probability distributions.

4.1.4 Benefits

Having described our data and methodology in both memos, we now describe some benefits. In the Citi Bike memo,

> The project aligns with our goals to, in Simmons's words, "be innovative in how we meet this challenge."

And in the Dodgers memo, perhaps because we believe the benefits are comparatively less obvious, or less proven, we further develop them:

> Joint distributions of all events, an eventual goal, will allow us to ask counterfactuals — "what if we do *this*" or "what if our opponent does *that*" — and simulate games to learn how decisions change win probability. It enables optimal strategy.
>
> Rational and optimal, this approach is more efficient for gaining wins. For perspective, each added win from the free-agent market costs 10 million, give or take, and the league salary cap prevents unlimited spend on talent. There is no cap, however, on investing in rational decision processes.

4.1.5 Limitations

In the Citi Bike memo, we didn't identify limitations. Should we?

In the Dodgers memo, we do, while also explaining how we plan to overcome those limitations:

> Computational issues are being addressed in `Stan`, a tool that enables inferences through advanced simulations. This open-source software is free but teaching its applications will require time. To shorten our learning curve, we can start with Stan interfaces that use familiar syntax (like `lme4`) but return joint probability distributions: `R` packages `rethinking`, `brms`, or `rstanarm`.

4.1.6 Conclusion

Finally, we wrap up in both memos. in the Citi Bike memo, after echoing the quote from Simmons, we state,

> Let's draft a detailed proposal.

Again, the Dodgers memo is similar. There, we circle back to our introduction to Sandy Koufax and his perfect game, then conclude,

> Perfect games aside, we can test the concept with decisions to steal.

Again, this idea of echoing something from where we began is journalism's complement to the lead.

Zinsser explains that, ideally, the ending should encapsulate the idea of the piece and conclude with a sentence that jolts us with its fitness or unexpectedness. Consider bringing the story full circle — to strike at the end an echo of a note that was sounded at the beginning. It gratifies a sense of symmetry.

Executives' lines of reasoning commonly, but do not always, follow the general document structure described above. If we don't have information otherwise, this is a good start.

4.2 *Narrative structure*

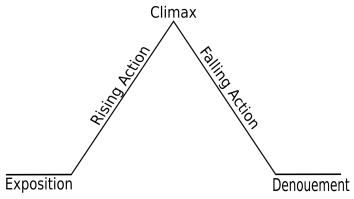


Figure 4.4: The five components of narratives create tension that helps hold audiences' attention.

The above ideas — tools — are helpful in structuring and writing persuasive memos, and longer communications for that matter. And as writing lengthens, the next couple of related tools can be especially helpful in refining the narrative structure in a way that holds our audience's interest by creating *tension*. German dramatist Gustav Freytag in the late 19th century illustrated a narrative arc used in Shakespearean dramas, shown in figure 4.4[13].

[13] This form of narrative has dominated since Aristotle's *Poetics*, but narrative is broader. See Rick Altman, *A Theory of Narrative* (New York: Columbia University Press, 2008).

The primary elements of an applied analytics project may be thought of as a well-articulated business problem, a data science solution, and a measurable outcome to produce value for the organization. The analytics project may thus be conceptualized as a narrative arc, with a beginning (problem), middle (analytics), and end (overcoming of the problem), along with characters (analysts, colleagues, clients) who play important roles.

Nancy Duarte[14] used the narrative arc to conceptualize an interesting alternative way to think about structure that creates tension: alternating *what is* with *what may be*, as in figure 4.5.

[14] Nancy Duarte, *Resonate: Present Visual Stories That Transform Audiences* (Wiley, 2010).

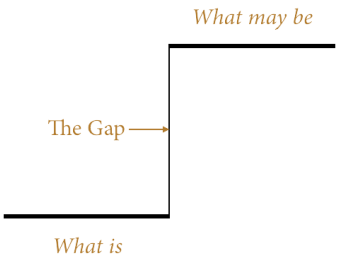


Figure 4.5: Describing *what may be* after *what is* creates a contrast, which we find interesting.

We can repeat this approach, see figure 4.6, switching between *what is* and *what may be* to maintain a sense of tension or interest throughout a narrative arc. Once you become aware, you may be surprised how

much you find writing in this form.

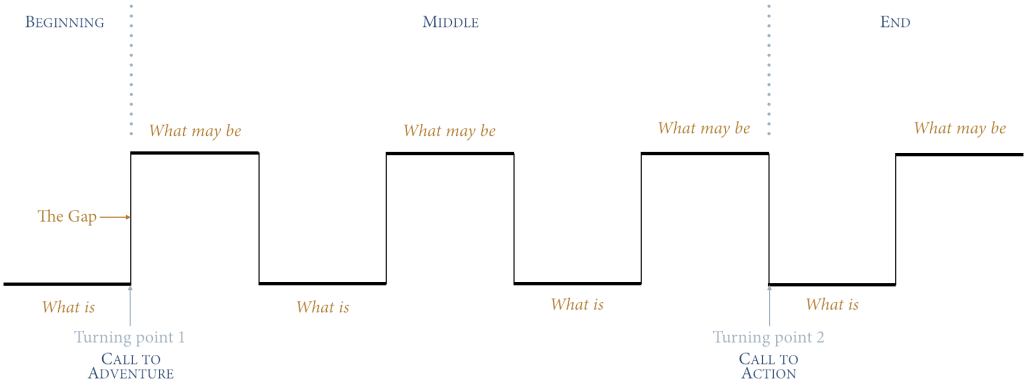


Figure 4.6: Duarte illustrates the repeated switching between *what is* and *what may be*, which helps to hold audience interest throughout a narrative.

This juxtaposition of two states for creating tension is *another form of comparison*. Reconsider the narrative in *The Next Rembrandt*. Try to label each unit of information as what was known before starting the project, and what was knew information, learned in the planning or work in the project.

> **Exercise 4.1** (Identify what is, what could be gaps). Revisit the two example memos. Identify sentences or paired sentences that shift focus from *what is* to *what could be*, creating a contrast.

Let's look closer, now, to sentence structure.

4.3 *Sentence structure*

When we describe old before new, using sentence structure, it generally improves understanding. The concept has also been described as an *information unit*. "The information unit is what its name implies: a unit of information. Information, in this technical grammatical sense, is the tension between what is already known or predictable and what is new or unpredictable."[15] As a general principle, "readers follow a story most easily if they can begin each sentence with a character or idea that is familiar to them, either because it was already mentioned or because it comes from the context."[16]

Consider an alternative flow of information. Put new information before old information. Reversing the information flow will likely confuse your audience. This point was clearly demonstrated in a classic movie, *Memento*, where Director Christopher Nolan tells the story of a man with anterograde amnesia (the inability to form new memories) searching for someone who attacked him and killed his

[15] M. A. K. Halliday and Christian M. I. M. Matthiessen, *An Introduction to Functional Grammar*, 3rd ed (London : New York: Arnold ; Distributed in the United States of America by Oxford University Press, 2004).

[16] Joseph M Williams, Joseph Bizup, and William T Fitzgerald, "17. Revising Style," in *The Craft of Research* (University of Chicago Press, 2016), 248–67.

Figure 4.7: Poster for Memento, a movie designed to purposefully confuse the audience by narrating the story (partly) in reverse.

wife, using an intricate system of Polaroid photographs and tattoos to track information he cannot remember. The story is presented as two different sequences of scenes interspersed during the film: a series in black-and-white that is shown chronologically, and a series of color sequences shown in reverse order (simulating for the audience the mental state of the protagonist). The two sequences meet at the end of the film, producing one complete and cohesive narrative. Yet, the reversed order is (effectively) designed to hold the audience in confusion so that they may get a sense of the confusion experienced by someone with this illness. Indeed, that we demonstrate this with film implies that ordering in visual representation matters too, and it does. As such, we revisit this in the context of images.

For reasons similar, explain complex information last. This is particularly important in three contexts: introducing a new technical term, presenting a long or complex unit of information, introducing a concept before developing its details. And just as the old—new paradigm helps to convey messages, so too does expressing crucial actions in verbs. Make your central characters the subjects of those verbs; keep those subjects short, concrete, and specific.

Exercise 4.2 (Identify sentence structure). Revisit the Dodgers memo again. This time, for each sentence and words within the sentence, try to identify whether the word or phrase is new or old. When determining this, consider both the words, phrases, and sentences preceding the one under analysis. Of note, you may also consider the audience's background knowledge as a form of information.

4.4 *Layering and heirarchy*

Most communications benefit by providing multiple levels in which the narrative may be read. Even emails and memos — concise communications — enable two layers, the title and the main body. Thus, the title should inform the audience of the relevance of the communication: what is it the author wants them to do or know. It should also, or at least, invite the audience to learn more through the details of the main body. As the communication lengthens, more layers may be used. The title's purpose remains the same, as does the main body. But we may add middle layers, *headers* and *subheaders*.[17] These should not be generic. Instead, the author should be able to read just these and understand the gist of the communication. This concept is well established where we intend persuasive communication. A well-known instructor of legal writing, for example, explains how to draft this middle layer:[18]

> Strong headings are like a good headline for a newspaper article: they give [the audience] the gist of what [they] need to know, draw [them]

[17] Doumont, *Trees, Maps, and Theorems*.

[18] Ross Guberman, *Point Made: How to Write Like the Nation's Top Advocates*, Second edition (Oxford ; New York, NY: Oxford University Press, 2014).

> into text [they] might otherwise skip, and even allow … a splash of creativity and flair.
>
> The old test is still the best. Could [the audience] skim your headings and subheadings and know why [they should act]?
>
> A good way to provide these "signposts" is to make your headings complete thoughts that, if true, would push you toward the finish line.

In accord:[19]

> Since clarity is the all-important objective, it helps to let the reader know in advance what topic you're to discuss. Headings are most effective if they're full sentences announcing not just the topic but your position on that topic.

[19] Antonin Scalia and Bryan A Garner, *Making Your Case*, Limited, The Art of Persuading Judges (Thomson West, 2008).

In short, headings should be what Doumont calls *messages*. Headings provide "effective redundancy." The redundancy gained from headers may be two fold. They, first, introduce your message before the detailed paragraphs and, second, may be collected up front, as a table of contents. Even short communication benefit from headers, and communications of at least several pages will likely benefit from such a table of contents along with headers.

4.5 Audiences and purposes

In the two memos we wrote to the head of analytics with the purpose to persuade our audience to approve continued work on our project. Let's compare that communication with the 124-word blog post[20] in example 4.3 describing a forthcoming (when published) data analytics project.

[20] Joao Caldeira et al., "Improving Traffic Safety Through Video Analysis: Pulse Lab Jakarta," *Data Science for Social Good*, 2018.

> **Example 4.3** (124-word Jakarta blog post)**. Improving Traffic Safety Through Video Analysis**. Nearly 2,000 people die annually as a result of being involved in traffic-related accidents in Jakarta, Indonesia. The city government has invested resources in thousands of traffic cameras to help identify potential short-term (e.g. vendor carts in a hazardous location) and long-term (e.g. poorly engineered intersections) safety risks. However, manually analysing the available footage is an overwhelming task for the city's Transportation Agency. In support of the Jakarta Smart City initiative, our team hopes to build a video-processing pipeline to extract structured information from raw traffic footage. This information can be integrated with collision, weather, and other data in order to build models which can help public officials quickly identify and assess traffic risks with the goal of reducing traffic-related fatalities and severe injuries.

Figure 4.8: Traffic congestion in Jakarta is among the worst globally. The project aimed to inform solutions.

The authors are from the Data Science for Social Good group at the University of Chicago, who partnered with Jakarta Smart City and UN Global Pulse, Jakarta.

Who might be their primary audience? You may find a clue in their post-project, award-winning paper[21]. May the blog post have more than one intended audience? We'll address this soon. For what purpose may the blog have been written? What details are included in the blog post? How specific are these details?

[21] Caldeira et al., "Improving Traffic Safety Through Video Analysis in Jakarta, Indonesia."

Exercise 4.3 (Analyze Jakarta blog post)**.** Compare the intended audience and structure and details of the *Jakarta* blog post to that in their award-winning, post-project paper, written within about 1,500 words. Identify an overall structure to the blog post. Does information in each sentence, or portion thereof, refer back to earlier sentences?

How is the structure of the blog post similar to, and different from, the structure of the *Citi Bike* and *Dodgers* memos?

Exercise 4.4 (Write 250-word Jakarta Memo)**.** Consider re-writing the blog post using 250-words. Re-write it to persuade the head of analytics at the Data Science for Social Good to get approval to move forward with a full proposal. You can use additional details from the post-write up paper, just ignore the described results of the project. When writing, consider how your audience, head of data analytics, and purpose may differ from the original material.

Let's consider how another data science project has been described, on a website and in a video. *The Next Rembrandt*[22] involved creating an original painting that most people would find indistinguishable from the late Rembrandt's actual work.

[22] www.nextrembrandt.com

Exercise 4.5 (Analyze The Next Rembrandt)**.** Review *The Next Rembrandt*. What problem were they trying to solve? What were the data the analysts worked with? With what details did they explain the project scope and methods? Be specific. Who may have been their audience? Assuming the audience, do you believe their choice of details was appropriate? Do you feel this is a story? Why or why not? Do you recognize any logical structure to the narrative? As we questioned Knaflic's example, do you believe this description of a data analytics project uses language common to author and audience? What leads you to believe this?

Exercise 4.6 (Write 250-word The Next Rembrandt Memo)**.** You have the benefit of knowing about the *The Next Rembrandt* project. but imagine it had not yet begun or approved. Apply the structure from the two memos above when writing a short paragraph or two, persuading the head of analytics at one of the project's sponsors to approve for you to write a proposal for conducting such a data analytics project. Weave the ideas for scoping that project with appropriate detail into your writeup. Use whatever details you like from the project website, just ignoring results.

When writing, consider how your audience, head of data analytics, may differ from the original material.

We've focused on identifying our purpose for communicating and developing structure for our communication. We must center the communication, as mentioned, on our audiences' scopes of knowledge and responsibilities in our problem context. Words and narrative must adapt to our audience.

Communication with any c-suite executive, if effective, begins with relevance to that audience's responsibilities and decision-making.

4.5.1 Chief Analytics Officer

Our discussions, communications, and exercises so far have focused on an analytics audience. More formally, the Chief Analytics Officer leads an organization's data analytics strategy, driving data-related business changes to transform company into a more analytics-driven one[23]. Thus, your communications to this audience should begin with relevance to "head up a company's data analytics operations, transforming data into business value, and drives data-related business change." Only after establishing that message, should you delve into your narrative and details.

[23] Minda Zetlin, "What Is a Chief Analytics Officer? The Exec Who Turns Data into Decisions," *CIO*, November 2017.

4.5.2 Chief Marketing Officer

A Chief Marketing Officer shares some responsibilities with the analytics officer and other executives, while other responsibilities are primarily her own. Broadly, he or she leads responses to changing circumstances; shapes products, sales strategies, and marketing ideas, collaborating across the company.

To dig deeper into the background and motivations of a marketing executive, we are guided by David Carr, who is Director of Marketing Strategy and Analysis at the London office of Digitas, a global marketing agency.

Carr[24] describes three main types of value that marketing drives:

1. **business value**: long and near-term growth, greater efficiency and enhanced productivity
2. **consumer value**: attitudes and behaviors that effect brand choice, frequency and loyalty
3. **cultural value**: shared beliefs that create a favorable environment in which to operate and influence

[24] David J Carr, "What Value Do You Create? Marketings 3 Types of Value," *Medium | Marketing*, January 2019.

He illustrates his research of, and experience with, these values graphically, as a central circle, and in concentric rings identifies various characteristics and details related to these values.

> **Exercise 4.7** (Compare and contrast executives)**.** Review Carr's graphic describing details and characteristics of the three types of value that marketing drives. Identify which of these are primarily the responsibility of marketing, and which of these responsibilities are shared with an analytics executive.

Relatedly, Carr[25] has mapped out the details for designing and managing a brand, and explained its interconnections:

[25] David J Carr, "A Map of Modern Brand Building," *Medium | David J Carr*, November 2016.

> The brand strategy should be influenced by the business strategy and should reflect the same strategic vision and corporate culture. In addi-

> tion, the brand identity should not promise what the strategy cannot or will not deliver. There is nothing more wasteful and damaging than developing a brand identity or vision based on strategic imperative that will not get funded. An empty promise is worse than no promise.

We can tie many aspects of brand building and marketing value to measurements and data. Carr explains how marketing does — and should — work with data.[26] His article suggests how we should craft data-driven messages for marketing executives.

[26] David J Carr, "Data Is the New Oil: Dirty, Misunderstood, Polluting the World & Pulled from All the Wrong Places," *Medium | Redwhale*, January 2018.

4.5.3 Chief Executive Officer

Typically, the chief analytics and marketing officers report directly to the CEO, who has ultimate responsibility to drive the business. Bertrand[27] reviews empirical studies on the characteristics of CEOs. They write, while "modern-day CEOs are more likely to be generalists," more than one quarter of those running fortune 500 companies have earned an MBA. The core educational components of the MBA program at Columbia, for example, include managerial statistics, business analytics, strategy formulation, marketing, financial accounting, corporate finance, managerial economics, global economic environment, and operations management.[28] This type of curricula suggests the CEO's vocabulary intersects with both analytics and marketing. Indeed, Bertrand explains that "current-day CEOs may require a broader set of skills as they directly interact with a larger set of employees within their organization." If they are fluent in the basics of analytics and marketing, their responsibilities are both broader and more focused on leading the drive for creating business value. Our communications with the CEO should begin with and remained focused on how the content of our communication helps the CEO with their responsibilities.

[27] Marianne Bertrand, "CEOs," *Annual Review of Economics* 1 (2009): 121–49.

[28] Columbia University, "MBA Core Curriculum," *Columbia Business School* (https://www8.gsb.columbia.edu/programs/mba/acade curriculum, 2020).

4.6 Multiple or mixed audiences

We should keep in mind that audiences[29] have a continuum of knowledge. Everyone is a **specialist** on some subjects and a **non-specialist** on others. Moreover, even a group of all specialists could be subdivided into more specialized and less specialized readers. Specialists want details. Specialists want more detail because they can understand the technical aspects, can often use these in their own work, and require them anyway to be convinced. Non-specialists need you to bridge the gap. The less specialized your audience, the more basic information is required to bridge the gap between what they know and what the document discusses: more background at

[29] Note the plural. While we have identified a single person in the example memo, that memo may be passed to others on his team — it may have secondary audiences.

the beginning, to understand the need for and importance of the work; more interpretation at the end, to understand the relevance and implications of the findings.

Frequently we encounter **mixed audiences**. Audiences are multiple, for each reader is unique. Still, readers can usefully be classified in broad categories on the basis of their proximity both to the subject matter (the content) and to the overall writing situation (the context). Primary readers are close to the situation in time and space. Uncertainty of the knowledge of a reader is like having a mixed audience, one knowing more than the other. Writing for a mixed audience is, thus, quite challenging. That challenge to write for a mixed audience is to give secondary readers information that we assume the primary readers know already while keeping the primary reader interested. The solution, conceptually, is simple: just ensure that each sentence makes an interesting statement, one that is new to all readers — even if it includes information that is new to secondary readers only. Thus, make each sentence interesting for all audiences. Let's consider another of Doumont's examples. The first sentence in the example,

> We worked with IR.

may not work because IR may be unfamiliar to some in the audience. One might try to fix the issue by defining the word or, in this case, the acronym:

> We worked with IR. IR stands for information Resources and is a new department.

But that isn't ideal either because those who already know the meaning aren't given new information. It is, in fact, pedantic. The better approach is to weave additional information, like a definition, into the information that the specialist also finds interesting, like so:

> We worked with the recently launched Information Resources (IR) department.

Looking back at the Dodgers memo in example 4.2, consider the difference between

> After defining our objective (e.g. optimize expected runs) we will, from Statcast data, weight everything that could happen by its probability and accumulate these probability distributions.

and

> After defining our objective (e.g. optimize expected runs) we will, from Statcast data, compute expectations. Expectations are computed by weighting everything that could happen by its probability and accumulate these probability distributions.

Notice the difference? The former sentence weaves the definition of an expectation into the sentence to help any secondary audience less familiar. The latter sentence explicitly defines expectations, which the Director of Quantitative Analytics may find patronizing, a reaction we want to avoid, especially when trying to persuade.

Word choice, and what we emphasize, can be subtle. What if we had started the opening sentence with

> The most likely sequence of events on defense is a perfect game — occurring just 23 times in major-league baseball, once by our own Sandy Koufax.

instead of the actual sentence used,

> Our Sandy Koufax pitched a perfect game, the most likely event sequence, only once: those, we do not expect or plan.

In the first, unused, version, we start and emphasize the aspect of a perfect game as being the most likely sequence of events. In the second, we begin with a more personal and positive tone, more subtly adding a parenthetical about the part of a perfect game we want to emphasize and build from.

What about communicating details of the statistical analysis? This is no different. The details of data analysis can be explained well to audiences not specializing in data science or statistics, as demonstrated beautifully in *The Art of Statistics*[30]. The text should be studied, then, for exemplary practices in communicating about these *technical* concepts.

[30] D. J. Spiegelhalter, *The Art of Statistics: How to Learn from Data*, First US edition (New York: Basic Books, 2019).

4.7 Story

At this point, we've identified a problem or opportunity upon which our entity may decide to act. We've found data and considered how we might uncover insights to inform decisions. We've scoped an analytics project. In beginning to write, we've considered document structure, sentence structure, and narrative. We've also begun to consider our audience. Let's now focus on how we may directly employ *story*. A little research on story, though, reveals differences in use of the term. The Oxford English Dictionary defines story[31] generally as a *narrative*:

[31] ""Story, N.","* *Oxford English Dictionary*, 2015.

> An oral or written narrative account of events that occurred or are believed to have occurred in the past. . .

Distinguished novelist E.M. Forster famously described story as "a series of events arranged in their time sequence." Of note, he also

compares and distinguishes story from plot: "plot is also a narrative of events, the emphasis falling on causality: 'The king died and then the queen died' is a story. But 'the king died and then the queen died of grief' is a plot. The time sequence is preserved, but the sense of causality overshadows it."[32] But not just any narrative works for moving our audiences to act[33]. Let's consider other points of view.

[32] E. M. Forster, *Aspects of the Novel* (United Kingdom: Edward Arnold, 1927).

[33] Harari, *Sapiens*.

4.7.1 Unexpected change and information gaps

To understand the narrative arc of successful stories, John Yorke studied numerous stories, and from those induced general principles[34]. A journalist and author, too, has studied narrative structure but, with a different approach — he focuses on the cognitive science and psychology of how our mind works and relates those characteristics to story[35]. Story, writes Storr, typically begins with "unexpected change", the "opening of an an information gap", or both. Humans naturally want to understand the change, or close that gap; it becomes their goal. Language of messages and information that close the gap, then, form what we may think of as narrative's plot.

[34] John Yorke, *Into the Woods: A Five-Act Journey into Story* (The Overlook Press, 2015).

[35] Will Storr, *Science of Storytelling* (New York, NY: Abrams Books, 2020).

Indeed, Storr suggests that the varying so-called designs for plot structure[36] are all really different approaches to describing change:

> But I suspect that none of these plot designs is actually the 'right' one. Beyond the basic three acts of Western storytelling, the only plot fundamental is that there must be regular change, much of which should preferably be driven by the protagonist, who changes along with it. It's change that obsesses brains. The challenge that storytellers face is creating a plot that has enough unexpected change to hold attention over the course of an entire novel or film. This isn't easy. For me, these different plot designs represent different methods of solving that complex problem. Each one is a unique recipe for a plot that moves relentlessly forwards, builds in intrigue and tension and never stops changing.

[36] For example, Blake Snyder, *Save the Cat!: The Last Book on Screenwriting You'll Ever Need* (S.l.: Michael Wiese, 2013); Christopher Booker, *The Seven Basic Plots: Why We Tell Stories* (London ; New York: Continuum, 2004).

A quantitative analysis of over 100,000 narratives suggests this too[37].

[37] David Robinson, "Examining the Arc of 100,000 Stories: A Tidy Analysis," *Variance Explained*, April 2017.

We evolved for recognizing change, and for *cause and effect*. Thus, a narrative or story is driven forward by linking together change after change as instances of cause and effect. Indeed, to create initial interest, we only need to *foreshadow* change.

The need to show change extends to graphics stories, too. Science graphics editor at The New York Times, Jonathan Corum, explained the importance of change.[38]

[38] Jonathan Corum, "See, Think, Design, Produce 3," *13pt Information Design* (http://style.org/stdp3/, March 2016).

...

4.7.2 *Examples*

Let's consider a couple of narratives in data science. The short narratives in Howard Wainer's excellent book, each about a data science concept that people frequently misunderstand[39], are exemplary. He begins each of these by setting up a contrast or information gap. In chapter 1, for example, he teaches the "Rule of 72" as a heuristic to think about compounding quantities by posing a question:

> Great news! You have won a lottery and you can choose between one of two prizes. You can opt for either:
>
> 1. $10,000 every day for a month, or
> 2. One penny on the first day of the month, two on the second, four on the third, and continued doubling every day thereafter for the entire month.
>
> Which option would you prefer?

[39] Howard Wainer, *Truth or Truthiness*, Distinguishing Fact from Fiction by Learning to Think Like a Data Scientist (Cambridge: Cambridge University Press, 2016).

Similarly, in chapter 2, Wainer teaches us implications of the law of large numbers by exposing an information gap. Again, he uses a question:

> "Virtuosos becoming a dime a dozen," exclaimed Anthony Tommasini, chief music critic of the New York Times in his column in the arts section of that newspaper on Sunday, August 14, 2011.
>
> ...
>
> But why?

Once he has setup his narratives, he bridges the gap. Let's keep in mind that his purpose in these stories are for audience *awareness*. To teach. We can adapt these narrative concepts, though, in communications for other purposes.

> **Exercise 4.8** (Review example memos for story). By this discussion, are the *Citi Bike* and *Dodgers* memos a story? If not, what they may lack? If so, explain what structure or form makes them a story. Do the story elements — or would they if used — add persuasive effect?

> *Rabbit Hole* (Inner workings of narrative). For a detailed understanding of narrative, consult seminal and recent work.[40]

[40] Altman, *A Theory of Narrative*; Mieke Bal, *Narratology: Introduction to the Theory of Narrative* (Toronto; Buffalo; London: University of Toronto Press, 2017); Paul Ricoeur, *Time and Narrative. Vol. 1: ...*, trans. Kathleen McLaughlin, Repr (Chicago, Ill.: Univ. of Chicago Press, 1984); Paul Ricoeur, *Time and Narrative. Vol. 2: ...*, trans. Kathleen McLaughlin and David Pellauer, Repr (Chicago, Ill.: Univ. of Chicago Press, 1985); Paul Ricoeur, *Time and Narrative. Vol. 3: ...*, trans. Kathleen Blamey and David Pellauer, Repr (Chicago: Univ. of Chicago Pr, 1988).

4.8 *The importance of revision*

"We write a first draft for ourselves; the drafts thereafter increasingly for the reader."[41] Revision lets us switch from writing to understand to writing to explain. Switching audience is critical, and

[41] Joseph Williams and Gregory Colomb, *Style: Toward Clarity and Grace*, Toward Clarity and Grace (University of Chicago Press, 1990).

not doing so is a common mistake. Schimel explains one manifestation of the error:[42]

> Using an opening that explains a widely held schema is a flaw common with inexperienced writers. Developing scholars are still learning the material and assimilating it into their schemas. It isn't yet ingrained knowledge, and the process of laying out the information and arguments, step by step, is part of what ingrains it to form the schema. Many developing scholars, therefore, have a hard time jumping over this material by assuming that their readers take it for granted. Rather, they are collecting their own thoughts and putting them down. There is nothing wrong with explaining things for yourself in a first draft. Many authors aren't sure where they are going when they start, and it is not until the second or third paragraph that they get into the meat of the story. If you do this, though, when you revise, figure out where the real story starts and delete everything before that.

[42] Schimel, *Writing Science*.

Revision gives us opportunity to focus on our audience once we understand what we have learned. This benefit alone is worth revision.

But it does more, especially when we allow time to pass between revisions: "If you start your project early, you'll have time to let your revised draft cool. What seems good one day often looks different the next."[43] As you revise, read aloud. While normal conversations do not typically follow grammatically correct language, well-written communications should smoothly flow when read aloud. Try reading this sentence aloud, following the punctuation:

> When we read prose, we hear it… it's variable sound. It's sound with — pauses. With *emphasis*. With, well, you know, a certain rhythm.[44]

[43] Wayne C Booth et al., "13. Organizing Your Argument," in *The Craft of Research*, Fourth (University of Chicago Press, 2016).

[44] Richard Goodman, *The Soul of Creative Writing* (Routledge, 2008).

And when revising, consider each word and phrase, and test whether removing that word or phrase changes the context or meaning of the sentence *for your audience*. If not, remove it. In a similar manner, when choosing between two words with equally precise meaning, it is generally best to use the word with fewer syllables or that flows more naturally when read aloud.

Exercise 4.9 (Revise a colleague's memo). Exchange a draft memo, and suggest a few revisions by applying the concepts we've covered.

4.9 *Example memos*

We will revisit the fully-formatted example Citi Bike and Dodgers memos that follow:

To **CitiBike**
Director of Analytics

2019 February 2

To inform rebalancing, let's explore docking and bike availability in the context of subway and weather information.

We should explore station and ride data in the context of subway and weather information to gain insight for "rebalancing," what our Dani Simmons explains is "one of the biggest challenges of any bike share system, especially in … New York where residents don't all work a traditional 9-5 schedule, and though there is a Central Business District, it's a huge one and people work in a variety of other neighborhoods as well."

A rebalancing study (Saldarriaga, 2013) by Columbia University Center for Spatial Research previously identified trends in bike usage using heatmaps. As those visualizations did not combine dimensions of space and time, which will be helpful to uncover trends in bike and station availability by neighborhood throughout a day, we can begin our analysis there.

NYC OpenData and The Open Bus Project provide published date, time, station ID, and ride instances for all our docking stations and bikes since we began service. To begin our project, we can visually explore the intersection of trends in both time and location with this data to understand problematic neighborhoods and, even, individual stations, using up-to-date information.

Then, we will build upon the initial work, exploring causal factors such as the availability of alternative transportation (e.g., subway stations near docking stations) and weather. Both of which, we have available data that can be joined using timestamps.

The project aligns with our goals to, in Simmons's words, "be innovative in how we meet this challenge." Let's draft a detailed proposal.

Sincerely,
Scott Spencer

Saldarriaga, Juan Francisco. *CitiBike Rebalancing Study*. Spatial Information Design Lab, Columbia University, 2013. http://spatialinformationdesignlab.org/projects/citibike-rebalancing-study.

To **Scott Powers** 2019 February 2
Director, Quantitative Analytics

Our game decisions should optimize expectations. Let's test the concept by modeling decisions to steal.

Our Sandy Koufax pitched a perfect game, the most likely event sequence, only once: those, we do not expect or plan. Since our decisions based on other most likely events don't align with expected outcomes, we leave wins unclaimed. To claim them, let's base decisions on expectations flowing from decision theory and probability models. A joint model of all events works best, but we can start small with, say, decisions to steal second base.

After defining our objective (e.g. optimize expected runs) we will, from Statcast data, weight everything that could happen by its probability and accumulate these probability distributions. Joint distributions of all events, an eventual goal, will allow us to ask counterfactuals — "what if we do *this*" or "what if our opponent does *that*" — and simulate games to learn how decisions change win probability. It enables optimal strategy.

Rational and optimal, this approach is more efficient for gaining wins. For perspective, each added win from the free-agent market costs 10 million, give or take, and the league salary cap prevents unlimited spend on talent. There is no cap, however, on investing in rational decision processes.

Computational issues are being addressed in Stan, a tool that enables inferences through advanced simulations. This open-source software is free but teaching its applications will require time. To shorten our learning curve, we can start with Stan interfaces that use familiar syntax (like `lme4`) but return joint probability distributions: `R` packages `rethinking`, `brms`, or `rstanarm`. Perfect games aside, we can test the concept with decisions to steal.

Sincerely,
Scott Spencer

5
Persuasion and biases

Should we use data science to persuade others? The late Robert Abelson thought so when he published *Statistics as Principled Argument* in 1995.[1] But since then, this question has been under the public eye as we try to correct the replication crisis we mentioned in section 2.4. A special interest group has formed in service of this correction.[2] They explain,

> we propose to refer to *transparent statistics* as a *philosophy of statistical reporting whose purpose is to advance scientific knowledge rather than to persuade*. Although transparent statistics recognizes that rhetoric plays a major role in scientific writing [citing Abelson], it dictates that when persuasion is at odds with the dissemination of clear and complete knowledge, the latter should prevail.

[1] Robert P Abelson, *Statistics as Principled Argument* (Psychology Press, 1995).

[2] Chat Wacharamanotham et al., "Special Interest Group on Transparent Statistics Guidelines," *The 2018 CHI Conference*, April 2018, 1–441.

Andrew Gelman poses the question, too:[3]

> Consider this paradox: statistics is the science of uncertainty and variation, but data-based claims in the scientific literature tend to be stated deterministically (e.g. "We have discovered ... the effect of X on Y is ... hypothesis H is rejected"). Is statistical communication about exploration and discovery of the unexpected, or is it about making a persuasive, data-based case to back up an argument?

[3] Andrew Gelman, "Ethics in Statistical Practice and Communication: Five Recommendations," *Significance* 15, no. 5 (October 2018): 40–43.

Only to answer:

> The answer to this question is necessarily each at different times, and sometimes both at the same time.
>
> Just as you write in part in order to figure out what you are trying to say, so you do statistics not just to learn from data but also to learn what you can learn from data, and to decide how to gather future data to help resolve key uncertainties.
>
> Traditional advice on statistics and ethics focuses on professional integrity, accountability, and responsibility to collaborators and research subjects.

All these are important, but when considering ethics, statisticians must also wrestle with fundamental dilemmas regarding the analysis and communication of uncertainty and variation.

Exercise 5.1 (Discuss persuasion's (mis)applications). Gelman seems to place persuasion with deterministic statements and constrasts that with the communication of uncertainty. How do you interepret Gelman's statement? Must we trade uncertainty for persuasive arguments? Discuss these issues and the role of persuasion, if any, in the context of a data analytics project.

5.1 *Methods of persuasion*

A MEANS OF PERSUASION "is a sort of demonstration (for we are most persuaded when we take something to have been demonstrated)," writes Aristotle.[4] Consider, first, appropriateness of timing and setting, *Kairos*. Can the entity act upon the insights from your data analytics project, for example? What affect may acting at another time of place mean for the audience? Second, arguments should be based on building common ground between listener and speaker, or listener and third-party actor. Common ground may emerge from shared emotions, values, beliefs, ideologies, or anything else of substance. Aristotle referred to this as *pathos*. Third, Arguments relying on the knowledge, experience, credibility, integrity, or trustworthiness of the speaker — *ethos* — may emerge from the character of the advocate or from the character of another within the argument, or from the sources used in the argument. Fourth, the argument from common ground to solution or decision should be based on the syllogism or the syllogistic form, including those of enthymemes and analogies. Called *logos*, this is the logical component of persuasion, which may reason by framing arguments with metaphor, analogy, and story that the audience would find familiar and recognizable. Persuasion, then, can be understood as researching the perspectives of our audience about the topic of communication, and moving from their point of view "step by step to a solution, helping them appreciate why the advocated position solves the problem best."[5] The success of this approach is affected by our accuracy and transparency.

[4] Aristotle and C. D. C. Reeve, *Rhetoric* (Indianapolis ; Cambridge: Hackett Publishing Company, Inc, 2018).

[5] Richard M. Perloff, *The Dynamics of Persuasion: Communication and Attitudes in the 21st Century*, Sixth edition (New York: Routledge, Taylor & Francis Group, 2017).

Exercise 5.2 (Kairos, pathos, ethos, logos in Citi Bike memo). In the *Citi Bike* memo, example 4.1, identify possible audience perspectives of the communicated topic. In what ways, if at all, did the communication seek to start with common ground? Do you see any appeals to credibility of the author or sources? What forms of logic were used in trying to persuade the audience to approve of the request? Consider whether other or additional approaches to kairos, pathos, ethos, and logos could improve the persuasive effect of the communication.

Exercise 5.3 (Kairos, pathos, ethos, logos in Dodgers memo)**.** In the *Dodgers* memo, example 4.2, identify possible audience perspectives of the communicated topic. In what ways, if at all, did the communication seek to start with common ground? Do you see any appeals to credibility of the author or sources? What forms of logic were used in trying to persuade the audience to take action? Consider whether other or additional approaches to kairos, pathos, ethos, and logos could improve the persuasive effect of the communication.

Exercise 5.4 (Kairos, pathos, ethos, logos in Dodgers proposal)**.** In the second *Dodgers* example — the draft proposal at the end of this chapter, section 5.9 — is the communication approach identical to that in the Dodgers memo? If not, in what ways, if at all, did the communication seek to start with common ground? Do you see any appeals to credibility of the author or sources? What forms of logic were used in trying to persuade the audience to take action? Consider whether other or additional approaches to kairos, pathos, ethos, and logos could improve the persuasive effect of the communication.

Exercise 5.5 (Kairos, pathos, ethos, logos in student memo)**.** As with the above exercises, examine *your* draft data analytics memo. Identify how the audience may view the current circumstances and solution to the problem or opportunity you have described. Remember that it tends to be very difficult to see through our biases, so ask a colleague to help provide perspective on your audience's viewpoint. Have you effectively framed the communication using common ground? Explain.

5.1.1 Accuracy

Narrative arguments must avoid any temptation for overstatement. Strunk and White[6] warn:

> A single overstatement, wherever or however it occurs, diminishes the whole, and a carefree superlative has the power to destroy, for readers, the object of your enthusiasm.

[6] William Strunk and E B White, *The Elements of Style*, Fourth (Allyn & Bacon, 2000).

Two prominent legal scholars, one a former United States Supreme Court Justice, agree[7]:

> Scrupulous accuracy consists not merely in never making a statement you know to be incorrect (that is mere honesty), but also in never making a statement you are not *certain* is correct. So err, if you must, on the side of understatement, and flee hyperbole. . . Inaccuracies can result from either deliberate misstatement or carelessness. Either way, the advocate suffers a grave loss of credibility from which it is difficult to recover.

[7] Scalia and Garner, *Making Your Case*.

As in law, so too in the context of arguments supporting research:

> But in a research argument, we are expected to show readers why our claims are important and then to support our claims with good reasons and evidence, as if our readers were asking us, quite reasonably, Why should I believe that?... Instead, you start where your readers do, with

> their predictable questions about why they should accept your claim, questions they ask not to sabotage your argument but to test it, to help both of you find and understand a truth worth sharing (p. 109).... Limit your claims to what your argument can actually support by qualifying their scope and certainty (p. 129)[8].

[8] Wayne C Booth et al., *The Craft of Research*, Fourth (University of Chicago Press, 2016).

5.1.2 *Transparency*

Edward Tufte[9] explains, "The credibility of an evidence presentation depends significantly on the quality and integrity of the authors and their data sources."

Be accurate. Be transparent.

[9] Edward R. Tufte, *Beautiful Evidence* (Graphics Press, 2006).

5.1.3 *Syllogism and enthymeme*

> Leaving aside emotional appeals [for the moment], persuasion is possible only because all human beings are born with a capacity for logical thought. It is something we all have in common. The most rigorous form of logic, and hence the most persuasive, is the syllogism.
>
> — Garner & Scalia, *Making Your Case*.

Syllogisms are one of the most basic tools of logical reasoning and argumentation. They are structured argument, constructed with a major premise, a minor premise, and a conclusion. Formally, the structure is of the form,

> All A are B.
>
> C is A.
>
> Therefore, C is B.

Such rigid use of "all" and "therefore" isn't necessary, what's necessary is the meaning of each premise and conclusion.

We may sometimes abbreviate the syllogism, leaving one of the premises implied (*enthymeme*). The effectiveness of this approach depends upon whether your audience will, from their knowledge and experience, naturally fill in the implied gap in logic.

Syllogism and enthymeme are a powerful tool for persuasion. But the persuasive effect may be compromised — as tested experimentally[10] — by various audience *biases* and perceptions of credibility, discussed above. Logic also serves as a building block for a rhetoric of narrative, i.e., a narrative that *convinces* the audience.

[10] David E. Copeland, Kris Gunawan, and Nicole J. Bies-Hernandez, "Source Credibility and Syllogistic Reasoning," *Memory & Cognition* 39, no. 1 (January 2011): 117–27; J. St. B. T. Evans, Julie L. Barston, and Paul Pollard, "On the Conflict Between Logic and Belief in Syllogistic Reasoning," *Memory & Cognition* 11, no. 3 (May 1983): 295–306.

5.1.4 *Narrative as argument*

A rhetoric of narrative is logical, but also emotive and ethical.[11] It may seem surprising to find argument common in fiction[12], and

[11] John Rodden, "How Do Stories Convince Us? Notes Towards a Rhetoric of Narrative," *College Literature* 35, no. 1 (2008): 148–73.

[12] George Orwell, *1984* (New York: Haughton Mifflin Harcourt, 2017) The novel argues against political tyrrany.

its value grows with non-fiction and communication for business purposes. A rhetorical narrative functions, if effective, by adjusting its ideas to its audience, and its audience to its ideas. The idea, in this sense, includes the sequence of events that demonstrate change or contrast, introduced in section 4.7. To enable action on an issue, in Aristotle's words, *dispositio*, it was essential to state the case through description — writing imaginable pictures — and narration (telling stories).[13]

[13] Aristotle and Reeve, *Rhetoric*.

Exercise 5.6 (Imaginable pictures in example memos)**.** Consider the memo examples 4.1 and 4.2. Do either elicit images in the narratives? Explain. In the *Citi Bike* memo, what might be a reason for quoting Dani Simmons? Does that reason compare with or differ from how you perceive possible reasons for referencing Sandy Koufax in example 4.2.

5.1.5 *Priming and emotion*

An introductory story can *prime* an audience for our main message:

> priming is what happens when our interpretation of new information is influenced by what we saw, read, or heard just prior to receiving that new information. Our brains evaluate new information by, among other things, trying to fit it into familiar, known categories. But our brains have many known categories, and judgments about new information need to be made quickly and efficiently. One of the "shortcuts" our brains use to process new information quickly is to check the new information first against the most recently accessed categories. Priming is a way of influencing the categories that are at the forefront of our brains.[14]

[14] Linda L. Berger and Kathryn M. Stanchi, *Legal Persuasion: A Rhetorical Approach to the Science*, Law, Language and Communication (Milton Park, Abingdon, Oxon ; New York, NY: Routledge, 2018).

As we make decisions based on emotion[15], and we may even start with emotion and back into supporting logic[16], we can introduce our messages with emotional priming, too. Yet we should be careful with this approach as audiences may feel manipulated and become resistant — or even opposed — to our message.

[15] Antonio R. Damasio, *Descartes' Error: Emotion, Reason, and the Human Brain* (New York: Putnam, 1994).

[16] Jonathan Haidt, "The Emotional Dog and Its Rational Tail: A Social Intuitionist Approach to Moral Judgment." *Psychological Review* 108, no. 4 (2001): 814–34.

5.1.6 *Tone of an argument*

When trying to persuade, authors sometimes approach changing minds too directly:

> Many of us view persuasion in macho terms. Persuaders are seen as tough-talking salespeople, strongly stating their position, hitting people over the head with arguments, and pushing the deal to a close. But this oversimplifies matters. It assumes that persuasion is a boxing match, won by the fiercest competitor. In fact, persuasion is different. It's more like teaching than boxing. Think of a persuader as a teacher, moving people step by step to a solution, helping them appreciate why the advocated position solves the problem best. Persuasion, in short, is a process.[17]

[17] Perloff, *The Dynamics of Persuasion*.

Try gradually leading audiences to act, framing your message as more reasonable among options, compromising, or any combination of these. And about those other *options* for decisions. Showing our audience that our message is more reasonable among options requires discussing those other options. If we do not discuss alternatives, and our audience knows of them or learns of them, they may find our approach less credible, and thus less persuasive, because we did not consider them in advocating our message.

5.1.7 Narrative patterns

Stories are built upon narrative patterns[18]. These include patterns for argumentation, the action or process of reasoning systematically in support of an idea, action, or theory. Patterns for argumentation serve the intent of persuading and convincing audiences. Let's consider three such patterns: **comparison**, **concretize**, and **repetition**.

[18] Nathalie Henry Riche et al., *Data-Driven Storytelling* (CRC Press, 2018).

Comparison allows the narrator to show equality of both data sets, to explicitly highlight differences and similarities, and to give reasons for their difference. We have already seen various forms of graphical comparison used for understanding. In *Storytelling with Data*[19], the author offers an example showing graphical comparison to support a call to action, see figure 5.1.

[19] Cole Nussbaumer Knaflic, *Storytelling with Data*, A Data Visualization Guide for Business Professionals (Wiley, 2015).

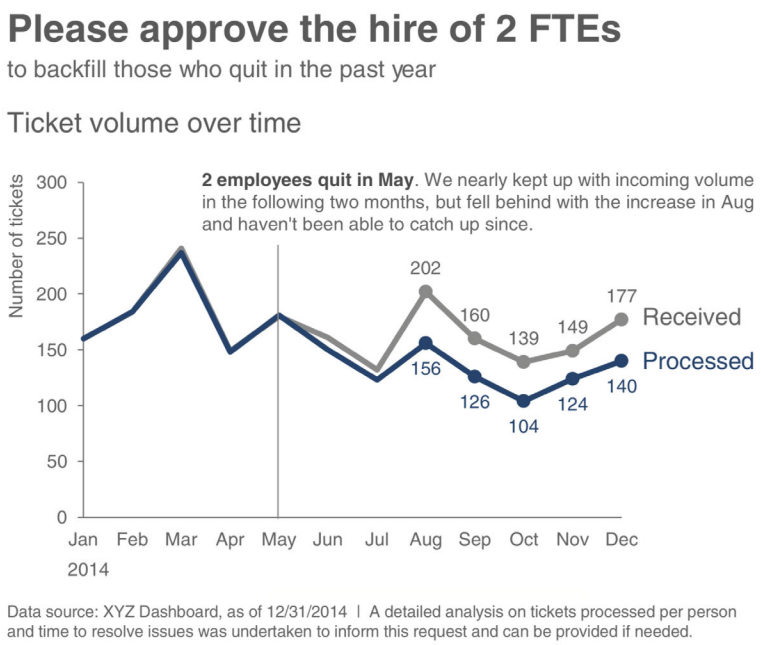


Figure 5.1: Knaflic's example uses comparison to persuade its audience to hire employees.

Concretizing, another type of pattern useful in argumentation,

shows abstract concepts with concrete objects. This pattern usually implies that each data point is represented by an individual visual object (*e.g.*, a point or shape), making them less abstract than aggregated statistics. Let's consider, first, an example from Reuters. In their article *Drowning in Plastic*[20], the authors encode data as individual images of plastic bottles collecting over time, figure 5.2, also making comparisons between the collections and familiar references, to demonstrate the severity of plastic misuse.

[20] Simon Scarr and Marco Hernandez, "Drowning in Plastic," *Reuters Graphics*, September 2019.

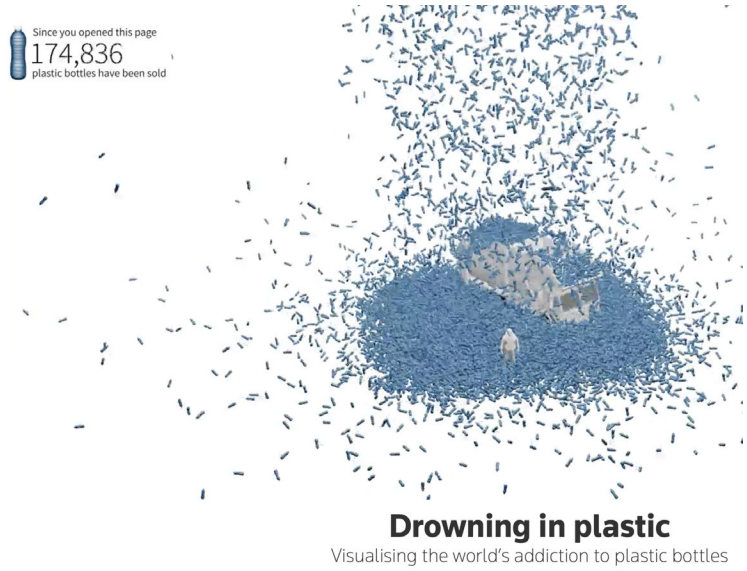


Figure 5.2: Authors use individual images of bottles to concretize the problem with plastic.

From a persuasive point of view, how does this form of data encoding compare with their secondary graphic, see figure 5.3, in the same article:

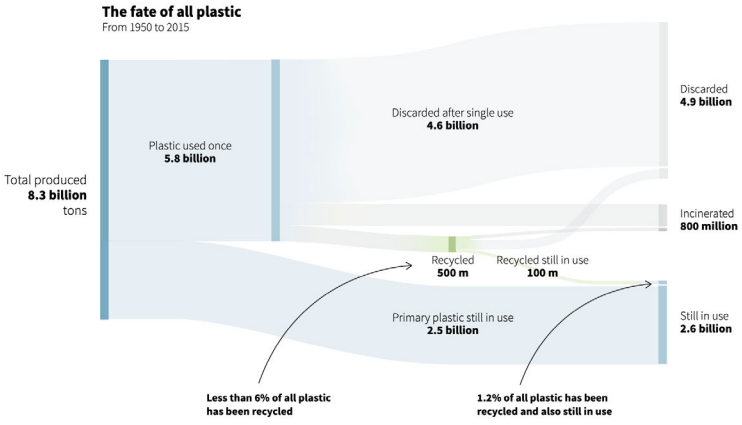


Figure 5.3: This graphic reports plastic (mis)use graphically and through annotation.

Do the two graphics intend to persuade in different ways? Explain. Here's another example from news, the New York Times[21], which

[21] Farhad Manjoo, "I Visited 47 Sites. Hundreds of Trackers Followed Me." *New York Times*, August 2019.

represents each instance of tracking an individual who browsed various websites. Figure 5.4 represents a snippet from the full information graphic. The full graphic concretizes each instance of being tracked. Notice each colored dot is timestamped and labeled with a location. The intended effect is to convey an overwhelming sense to the audience that online readers are being watched — a lot.

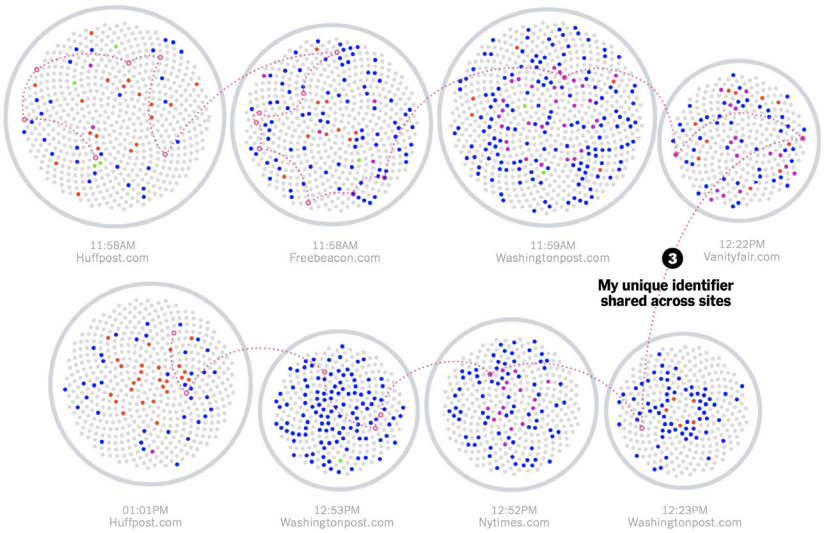


Figure 5.4: This snippet of the information graphic shows concretizing each instance of tracking someone's every browser click online to create an overwhelming sense of being watched.

Review the full infographic and consider whether the use of concretizing each timestamped instance, labeled by location, heightens the realization of being tracked more than just reading the more abstract statement that "hundreds of trackers followed me."

Like concretizing, repetition is an established pattern for argumentation. Repetition can increase a message's importance and memorability, and can help tie together different arguments about a given data set. Repetition can be employed as a means to search for an answer in the data. Let's consider another information graphic, which exemplifies this approach. An article by Roston[22] uses several rhetorical devices intended to persuade the audience that greenhouse gasses cause global warming. A few of the repeated graphics and questions are shown in figure 5.5, reproduced from the article.

[22] Eric Roston and Blacki Migliozzi, "What's Really Warming the World?" *Bloomberg*, June 2015.

5.2 *Statistical persuasion*

Let's consider, now, how statistics informs persuasion.

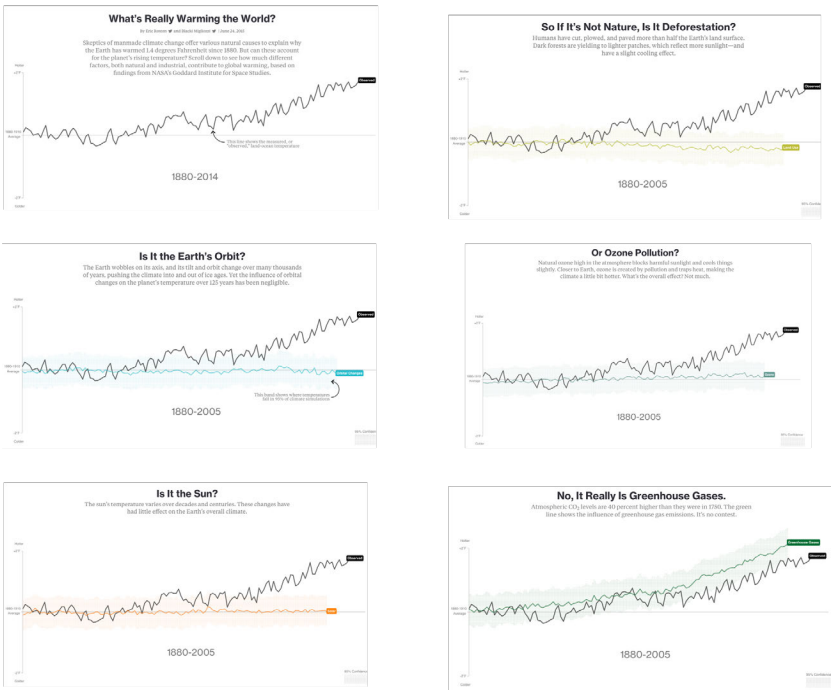


Figure 5.5: Repetition is used in several ways in this graphic-based news story.

5.2.1 *Comparison is crucial*

We've touched upon the importance of comparison. Tufte[23] explains the centrality of comparison, "The fundamental analytical act in statistical reasoning is to answer the question 'Compared with what?' "

Abelson, too, forcefully argues that comparison is central: "The idea of comparison is crucial. To make a point that is at all meaningful, statistical presentations must refer to differences between observation and expectation, or differences among observations." Abelson tests his argument through a statistical example,

> The average life expectancy of famous orchestral conductors is 73.4 years.

He asks: Why is this important; how unusual is this? Would you agree that answering his question requires some **standards of comparison**? For example, should we compare with orchestra players? With non-famous conductors? With the public? With other males in the United States, whose average life expectancy was 68.5 at the time of the study reported by Abelson? With other males who have already reached the age of 32, the average age of appointment to a first conducting post, almost all of whom are male? This group's average life expectancy was 72.0.

[23] Tufte, *Beautiful Evidence*.

5.2.2 *Elements of statistical persuasion*

Several properties of data, and its analysis and presentation, govern its persuasive force. Abelson describes these as magnitude of effects, articulation of results, generality of effects, interestingness of argument, and credibility of argument: *MAGIC*.

Magnitude of effects. The strength of a statistical argument is enhanced in accord with the quantitative magnitude of support for its qualitative claim. Consider describing effect sizes like the difference between means, not dichotomous tests. The information yield from null hypothesis tests is ordinarily quite modest, because all one carries away is a possibly misleading accept-reject decision. To drive home this point, let's model a realization from a linear relationship between two independent, random variables normal$(x \mid 0, 1)$ and normal$(y \mid 1, 1)$ by simulating them in R as follows:

```
set.seed(9)
y <- rnorm(n = 1000, mean = 1, sd = 1)
x <- rnorm(n = 1000, mean = 0, sd = 1)
```

And model them using a linear regression,

```
model_fit <- lm(y ~ x)
```

Results in a statistically significant p-value:

```
Call:
lm(formula = y ~ x)

Residuals:
     Min       1Q   Median       3Q      Max
-3.10551 -0.65170  0.02839  0.64702  2.74517

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.00517    0.03031  33.159   <2e-16 ***
x           -0.06278    0.03059  -2.052   0.0404 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9585 on 998 degrees of freedom
Multiple R-squared:  0.004202,  Adjusted R-squared:  0.003204
F-statistic: 4.211 on 1 and 998 DF,  p-value: 0.04041
```

Yet we know there is no actual relationship between the two variables. p-values say little, and can mislead. Here's what a p-value of less than, say, 0.01 means: If it were true that there were no systematic difference between the means in the populations from which the samples came, then the probability that the observed means would have been as different as they were, or more different, is less than one in a hundred. This being strong grounds for doubting the viability of the null hypothesis, the null hypothesis is rejected.

More succinctly we might say it is the probability of getting the data given the null hypothesis is true: mathematically, $P(\text{Data} \mid \text{Hypothesis})$. There are two issues with this. First, and most problematic, the threshold for what we've decided is significant is arbitrary, based entirely upon convention pulled from a historical context not relevant to much of modern analysis.

Secondly, a p-value is not what we usually want to know. Instead, we want to know the probability that our hypothesis is true, given the data, $P(\text{Hypothesis} \mid \text{Data})$, or better yet, we want to know the possible **range** of the **magnitude of effect** we are estimating. To get the probability that our hypothesis is true, we also need to know the probability of getting the data if the hypothesis were not true:

$$P(\text{H} \mid \text{D}) = \frac{P(\text{D} \mid \text{H})P(\text{H})}{P(\text{D} \mid \text{H})P(\text{H}) + P(\text{D} \mid \neg\text{H})P(\neg\text{H})}$$

Consider an example by Dragicevic[24]. He describes the statistical information for four diet pills. Of the four pills, shown in figure 5.6, those interested should probably prefer pill 2, the data of which shows *no* statistical significance, instead of pill 1, the data of which does show statistical significance.

[24] Pierre Dragicevic, "Fair Statistical Communication in HCI," in *Modern Statistical Methods for HCI*, ed. Judy Robertson and Maurits Kaptein (Springer International Publishing, 2016), 291–330.

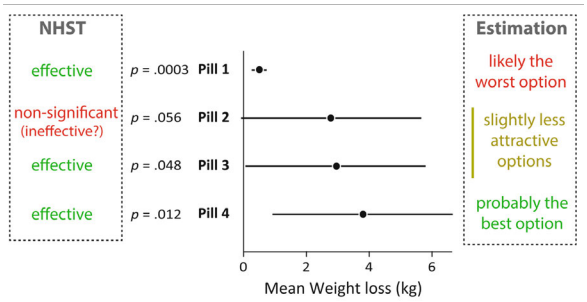


Figure 5.6: While the p-value of pill 2 is not significant, users should choose it instead of pill 1.

Decisions are better informed by comparing effect sizes and intervals. Whether exploring or confirming analyses, show results using an estimation approach — use graphs to show effect sizes and interval estimates, and offer nuanced interpretations of results. Avoid the pitfalls of dichotomous tests[25] and p-values. Dragicevic writes, "The notion of binary significance testing is a terrible idea for those who want to achieve fair statistical communication." In short, p-values alone do not typically provide strong support for a persuasive argument. Favor estimating and using magnitude of effects. Let's briefly consider the remaining characteristics that Abelson describes of statistical persuasion. These:

[25] Indeed, we are being warned to abandon significance tests. Blakeley B. McShane et al., "Abandon Statistical Significance," *The American Statistician* 73, no. sup1 (March 2019): 235–45.

Articulation of results. The degree of comprehensible detail in which conclusions are phrased. This is a form of specificity. We want to honestly describe and frame our results to maximize clarity (mini-

mizing exceptions or limitations to the result) and parsimony (focusing on consistent, connected claims).

Generality of effects. This is the breadth of applicability of the conclusions. Over what context can the results be replicated?

Interestingness of argument. For a statistical story to be theoretically interesting, it must have the potential, through empirical analysis, to change what people believe about an important issue.

Credibility of argument. Refers to believability of a research claim, requiring both methodological soundness and theoretical coherence.

Let's get back to the ever-important concept of comparison.

5.3 *Comparison through two numeric languages*

In language describing quantities, we have two main ways to compare. One form is additive or subtractive. The other is multiplicative. We humans perceive or process these comparisons differently. Let's consider an example from *Info We Trust*[26]:

> The Apollo program crew had **one more** astronaut than Project Gemini. Apollo's Saturn V rocket had about **seventeen times more** thrust than the Gemini-Titan II.

[26] R J Andrews, *Info We Trust: How to Inspire the World with Data* (Wiley, 2019).

We process the comparative language of "seventeen times more" differently than "1,700 percent more" or "33 versus 1.9". Add and subtract comparisons are easier for people to understand, especially with small numbers. Relative to additive comparisons, multiplying or dividing are more difficult. This includes comparisons expressed as ratios: a few times more, a few times less. People generally try to interpret multiplying operations through pooling, or repeat addition.

In this example, it may be better to show a graphical comparison,

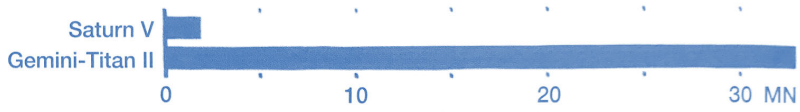


Figure 5.7: A bar chart allows relative comparisons between quantities that may be generally more useful than merely displaying numbers.

5.4 *Statistics and narrative*

We've discussed narrative and statistics as forms of persuasion. And we've seen examples of their combination. Is the combination more persuasive than either individual form? Some researchers have

claimed that the persuasive effect depends on the data and statistics.[27] They argue from an empirical study that narrative can improve less convincing data or statistics, but may actually detract from strong numerical evidence. Their study involved survey responses from participants that reviewed a_1) less favorable data (a phone that was relatively heavy and shatter-tested in a 3-foot drop) in the form of a list and a_2) the same data embedded within a narrative. The data was then changed to be more favorable (a phone that was relatively light and shatter-tested in a 30-foot drop) and b_1) placed into a list and b_2) the same, more favorable data was embedded within the same narrative. When comparing responses involving the less-favorable data, the researchers found that the narrative form positively influenced participants relative to presenting the data alone. But when comparing responses involving the more favorable data, the relationship reversed. Respondents was more swayed by the data alone than when presented with it embedded within the narrative. Of note, there was no practical (or significant) difference in responses between narratives with either data. They conclude, from the study, that narratives operate by taking the focus off the data, which may either help or harm a claim, depending on the strength of the data.

But a review of the actual narrative created for the study reveals that the narrative was not about the thing generating the data (a phone and its properties). Instead, the narrative was about a couple hiking that encountered an emergency and used the phone during the emergency. In other words, the data of the phone characteristics amounted to what the advertising industry might call a "product placement." Product placements, of course, have been found to be effective in transferring sentiment about the narrative to sentiment about the product. But it would be dangerous to generalize from this empirical study to potential effects and operations of other forms of narrative. Instead of choosing between listing convincing data on its own or embedding it as a product placement, we should consider *providing narrative context focused on the data and thing that generated it*. In other words, we can create a narrative that emphasizes the data, instead of shifting our audiences' focus from the data. And we can create that narrative context using metaphor, simile, and analogy, discussed next.

[27] Rebecca J. Krause and Derek D. Rucker, "Strategic Storytelling: When Narratives Help Versus Hurt the Persuasive Power of Facts," *Personality and Social Psychology Bulletin* 46, no. 2 (February 2020): 216–27.

5.5 *Comparison through metaphor, simile, analogy*

Metaphor adds to persuasiveness by reforming abstract concepts into something more familiar to our senses, signaling particular aspects of importance, memorializing the concept, or providing co-

herence throughout a writing.[28] The abstract concepts we need help explaining, ideas we need to make important, or the multiple ideas we need to link, we call the target domain. Common source domains include the human body, animals, plants, buildings and constructions, machines and tools, games and Sport, money, cooking and food, heat and cold, light and darkness, and movement and direction. Let's consider some examples. We begin with short example 5.1, excerpts from *The Next Rembrandt*.

[28] Ward Farnsworth, *Farnsworth's Classical English Metaphor* (David R. Godine Publisher, 2016); Zoltán Kövecses, *Metaphor: A Practical Introduction*, Second (Oxford University Press, 2010); Paul Ricoeur, *The Rule of Metaphor: Multi-Disciplinary Studies of the Creation of Meaning in Language*, trans. Robert Czerny and Kathleen McLaughlin (Toronto; Buffalo; London: University of Toronto Press, 1993); George Lakoff and Mark Johnson, *Metaphors We Live by* (Chicago: University of Chicago Press, 1980).

> **Example 5.1** (Excerpt from The Next Rembrandt)**.** To bring [Rembrandt] back, we distilled the artistic DNA from his work and used it to create The Next Rembrandt. ... To create new artwork using data from Rembrandt's paintings, we had to maximize the data pool from which to pull information. ... We created a height map using two different algorithms that found texture patterns of canvas surfaces and layers of paint. That information was transformed into height data, allowing us to mimic the brushstrokes used by Rembrandt.

Try to identify the target and source domains. Do you believe these source domains relate the target domains to something more familiar or concrete? For our second example, we return to Andrews's book, *Info We Trust*[29]. Andrews has more space to build the metaphor in example 5.2, using borrowing from the source domain of music.

[29] Andrews, *Info We Trust*.

> **Example 5.2** (Excerpt from Info We Trust)**.** How do we think about the albums we love? A lonely microphone in a smoky recording studio? A needle's press into hot wax? A rotating can of magnetic tape? A button that clicks before the first note drops? No!
>
> The mechanical ephemera of music's recording, storage, and playback may cue nostalgia, but they are not where the magic lies. The magic is in the music. The magic is in the information that the apparatuses capture, preserve, and make accessible. It is the same with all information.

After setting up this metaphor, he repeatedly refers back to it (example 5.3) as a form of shorthand each time:

> **Example 5.3** (References back to the music metaphor)**.** When you envision data, do not get stuck in encoding and storage. Instead, try to see the music. ... Looking at tables of any substantial size is a little like looking at the grooves of a record with a magnifying glass. You can see the data but you will not hear the music. ... Then, we can see data for what it is, whispers from a past world waiting for its music to be heard again.

What, if anything, do you think use of this source domain adds to the audiences understanding of data and information?

5.6 Patterns that compare, organize, grab attention

We can use patterns to "make the words they arrange more emphatic or memorable or otherwise effective."[30] In *Classical English Rhetoric*, Farnsworth provides a wealth of examples, categorized. Un-

[30] Ward Farnsworth, *Farnsworth's Classical English Rhetoric* (David R. Godine Publisher, 2011).

expected word placement calls attention to them, creates emphasis by coming earlier than expected or violating the reader's expectations. Note that, to violate expectations necessarily means reserving a technique like inversion for just the point to be made, lest the reader come to expect it — more is less, less is more. Secondly, it can create an attractive rhythm. Thirdly, when the words that bring full meaning come later, it can add suspense, and finish more climactic.

These patterns can be the most effective and efficient ways to show comparisons and contrasts. While Farnsworth provides a great source of these rhetorical patterns in more classical texts, we can find plenty of usage in something more relevant to data science. In fact, we have already considered a visual form of repetition in section 5.1.7. Let's consider this structure used in another example text for data science, found in *Observation and Experiment*[31].

[31] Paul Rosenbaum, *Observation and Experiment: An Introduction to Causal Inference* (Harvard University Press, 2017).

> **Example 5.4** (Reversal of structure, repetition at the end)**. A covariate is** a quantity determined prior to treatment assignment. In the Pro-CESS Trial, the age of the patient at the time of admission to the emergency room **was a covariate**. The gender of the patient **was a covariate**. Whether the patient was admitted from a nursing home **was a covariate**.

The first sentence begins "A covariate is . . . " Then, the next three sentences reverse this sentence structure, and repeat to create emphasis and nuance to the reader's understanding of a covariate. Here's another pattern from Rosenbaum's excellent book:

> **Example 5.5** (Repetition at the start, parallel structure)**. One might hope** that panel (a) of Figure 7.3 is analogous to a simple randomized experiment in which one child in each of 33 matched pairs was picked at random for exposure. **One might hope** that panel (b) of Figure 7.3 is analogous to a different simple randomized experiment in which levels of exposure were assigned to pairs at random. **One might hope** that panels (a) and (b) are jointly analogous to a randomized experiment in which both randomizations were done, within and among pairs. **All three of these hopes** may fail to be realized: there might be bias in treatment assignment within pairs or bias in assignment of levels of exposure to pairs.

Repetition and parallel structure are especially useful where, as in these examples, the related sentences are complex or relatively long. Let's consider yet another pattern:

> **Example 5.6** (Asking questions and answering them)**.** Where did Fisher's null distribution come from? From the coin in Fisher's hand.

Rhetorical questions or those the author answers are a great way to create interest when used sparingly. Seeing just a few examples invites direct imitation of them, which tends to be clumsy. Immersion in many examples allows them to do their work by way of a subtler process of influence, with a gentler and happier effect on the resulting style.

5.7 Le mot juste — *the exact word*

Writing poetically, Goodman[32] explains the importance of finding the exact word. *Le mot juste*, in French, is how it's expressed.

In our search we must also keep in mind, and use, words with the appropriate precision, as Alice explains:

> "When *I* use a word," Humpty Dumpty said in rather a scornful tone, "it means just what I choose it to mean—nothing more nor less."
>
> "The question is," said Alice, "whether you *can* make words mean so many different things."
>
> — Carroll, Lewis. *Alice's Adventures in Wonderland*.

Yet empirical studies suggest variation in our understanding of words that express quantity. For words meant to convey quantity, their meanings vary more than Alice would like. A researcher[33] reports survey responses from 23 NATO military officers who were asked to assign probabilities to particular phrases if found in an intelligence report. Another, online survey[34] of 46 individuals provided responses to the question: *What [probability/number] would you assign to the phrase [phrase]?* where the phrases matched those of the NATO study. The combined responses in figure 5.8 show wide variation in what probabilities individuals associate with words, although some ordering or ranking is evident.

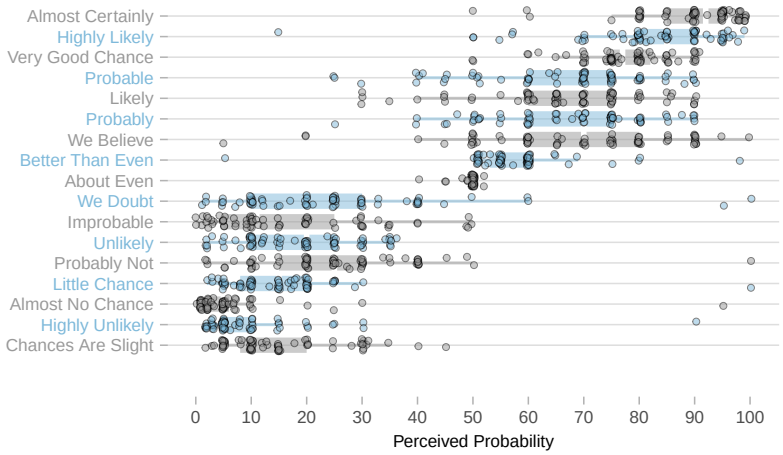


Figure 5.8: Results from the combined studies reflect uncertainty in the probability that people associate with words.

As with variation in probabilities assigned to words about uncertainty, the empirical study suggests variation in amounts assigned to words about size, shown in Figure 5.9.

[32] Goodman, *The Soul of Creative Writing*.

[33] Scott Barclay et al., "Handbook for Decision Analysis" (Decisions and Designs, Inc., 1977).

[34] Zonination, "Perceptions of Probability and Numbers," August 2015.

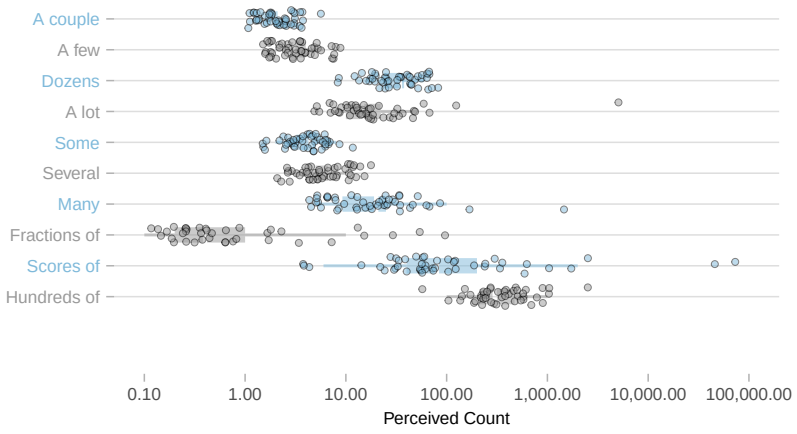


Figure 5.9: Even words whose definitions refer to counts have significant variation in perceived meaning.

Variance in perception of the meaning of such words does not imply we should avoid them altogether. It does mean, however, we should be aware of the meaning others may impart and complement them with numerals or graphic displays.

5.8 Heuristics and biases

Humans have two separate processes for understanding information, which Kahneman[35] labels as system one and system two. If we are to find common ground, and move our audience to a new understanding for decisionmaking, we must understand how they think. Intuitive (system one) thinking — impressions, associations, feelings, intentions, and preparations for actions — flow effortlessly. This system mostly guides our thoughts.

System one uses heuristics, biases. Reflective (system two) thinking, in contrast, is slow, effortful, and deliberate. Both systems are continuous, but system two typically monitors things, and only steps in when stakes are high, we detect an obvious error, or rule-based reasoning is required. For a sense of this difference, Kahneman provides exemplaray information that we process using system one, as in figure 5.10, and system two, as in mentally calculating 17 x 24. For other examples, consider figure 5.2 (system one) and figure 5.3 (processing may depend on familiarity with the graphic — a alluvial diagram — and which comparisons are of focus within the graphic).

We have decades of empircal and theoretical research available[36], while theoretical foundations have long been in place.[37]

[35] Daniel Kahneman, *Thinking, Fast and Slow* (Farrar, Straus and Giroux, 2013).

Figure 5.10: Most of us immediately sense emotion from this face, system one processing, but would need to work hard to mentally calculate 17 x 24, system two processing.

[36] Thomas Gilovich, Dale Griffin, and Daniel Kahnman, *Heuristics and Biases*, ed. Thomas Gilovich, Dale Griffin, and Daniel Kahneman, The Psychology of Intuitive Judgment (Cambridge: Cambridge University Press, 2009).

[37] Joshua B Miller and Andrew Gelman, "Laplaces Theories of Cognitive Illusions, Heuristics, and Biases" (December 2018).

Kahneman gives executives ways to guard against some biases by asking questions and recommending actions:[38]

self-interested biases | Is there any reason to suspect the team making the recommendation of errors motivated by self-interest? Review the proposal with extra care, especially for over optimism.

the affect heuristic | Has the team fallen in love with its proposal? Rigorously apply all the quality controls on the checklist.

groupthink | Were there dissenting opinions within the team? Were they explored adequately? Solicit dissenting views, discreetly if necessary.

saliency bias | Could the diagnosis be overly influenced by an analogy to a memorable success? Ask for more analogies, and rigorously analyze their similarity to the current situation.

confirmation bias | Are credible alternatives included along with the recommendation? Request additional options.

availability bias | If you had to make this decision in a year's time, what inform-ation would you want, and can you get more of it now? Use checklists of the data needed for each kind of decision.

anchoring bias | Where are the numbers from? Can there be . . . unsubstantiated numbers? . . . extrapolation from history? . . . a motivation to use a certain anchor? Re-anchor with data generated by other models or benchmarks, and request a new analysis.

halo effect | Is the team assuming that a person, organization, or approach that is successful in one area will be just as successful in another? Eliminate false inferences, and ask the team to seek additional comparable examples.

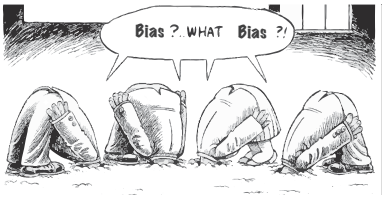
sunk-cost fallacy, endowment effect | Are the recommenders overly attached to past decisions? Consider the issue as if you are a new executive.

overconfidence, optimistic biases, competitor neglect | Is the base case overly optimistic? Have a team build a case taking an outside view: use war games.

disaster neglect | Is the worst case bad enough? Have the team conduct a premortem: imaging that the worst has happened, and develop a story about the causes.

loss aversion | Is the recommending team overly cautious? Align incentives to share responsibility for the risk or to remove risk.

[38] Daniel Kahneman, Dan Lovallo, and Olivier Sibony, "Before You Make That Big Decision ..." *Harvard Business Review* 89, no. 6 (June 2011): 50–60.



We increase persuasion by addressing these issues in anticipation that our audience will want to know. It's very hard to remain aware

of our own biases, so we need to develop processes that identify them and, most importantly, get feedback from others to help protect against them. Get colleagues to help us: Present ideas from a neutral perspective. Becoming too emotional suggests bias. Make analogies and examples comparable to the proposal. Genuinely admit uncertainty in the proposal, and recognize multiple options. Identify additional data that may provide new insight. Consider multiple anchors in the proposal.

5.9 *Brief proposals*

We've covered a lot of material in these last two lectures, from business writing, to visual components of communication and now different forms of persuasion. We can use all these techniques to help in writing a brief proposal to a chief analytics officer, asking him or her to approve your analytics project. Recall the example Dodgers memo? Let's continue that example with a 750-word brief proposal.[39] To assess whether the example proposal accomplishes its goals, note the audience. His background includes a doctor of philosophy in Statistics, and experience with machine learning and statistical programming in R. The example follows:

[39] Scott Spencer, "Proposal for Exploring Game Decisions Informed by Expectations of Joint Probability Distributions," Proposal, February 2019.

Exercise 5.7 (Deconstruct the Dodgers proposal)**.** Try to identify the document structure. Does it identify problems and goals? Data? Methods? Compare the structure, specificity and level of detail to both the memo, and to the *Jakarta* writeups, which were written for different purposes and audiences. Next, consider the tools we've covered in business writing, starting with messages and goals, applying typographic best practices, aligning information with grids, integrating graphics within paragraphs, linking words and graphics, annotation, and use of comparison, metaphor, patterns, and examples or analogies to persuade. How many can you find? If you were the director would you be persuaded to approve of the project? Why or why not? How might you edit the proposal to make it more persuasive?

Proposal for exploring game decisions informed by expectations of joint probability distributions

To: Scott Powers, Director of Quantitative Analysis, Los Angeles Dodgers
From: Scott Spencer, Faculty and Lecturer, Columbia University

14 February 2019

Our game decisions based on current modeling do not maximize spend per win. We witnessed the mid-market Astros use analytics to overtake us in the 2017 World Series (Luhnow 2018ab). Our efforts also do not maximize expected wins. But we can. To do so, we need to jointly model probabilities of all game events and base decisions on *expectations* of those distributions. With adequate computing emerging, we can be first using the probabilistic programming language Stan and parallel processing. To demonstrate the concept, consider a probability model for decisions to steal second base, below, which suggests teams are too conservative, leaving wins unclaimed. This model allows us to ask, for example—*should Sanchez steal against Sabathia? Or against Pineda?*

1 Our current analyses do not optimize expected wins

Seven terabytes of uncompressed data generated per game overshadow the lack of situational data needed for decision-making that maximizes expected utility. Consider that pitchers, on average, only face10 percent of major league batters regardless of game state; the reverse is true, too. Or when deciding whether a base runner should attempt to steal against a specific pitcher and catcher in a state of play, say, we are lucky to have any data. Common analyses and heuristics for these situations are inadequate: they not only overfit the data (if any exist), but also offer no manner of estimating changes in probabilities for maximizing *expected* utility (winning the game).

Accurately quantifying probabilities, and changes thereof, in a given context enable us to answer counterfactuals, from which we can build strategies that maximize our objectives (Parmigiani 2002). This approach is possible at scale using Stan (Carpenter et al. 2017). It's time to jointly model probabilities of all events.

2 Modeling probabilities for steal success illustrates a broader benefit

To see the potential of implementing probability models, let's consider, again, the decision to steal bases, given a specific counterfactual:

> In a game against New York Yankees, should Milwaukee Brewers's Lorenzo Cain attempt to steal second base with no one else on base and two outs before the seventh inning, against Gary Sanchez as catcher and Michael Pineda as pitcher? What if against Sanchez and CC Sabathia as pitcher?

More specifically, how can we know the *expectation* that Cain's attempt in each situation increases the probability of expected runs that inning and by how much? Using Stan, I've coded a generative model that along with play outcomes considers various information (runner foot-speed, catcher pop-time) and player characteristics, like pitcher handedness. With the model, we have an answer that also shows the uncertainty. Given 2017 data, this model suggests Cain should steal against Pineda, not Sabathia:

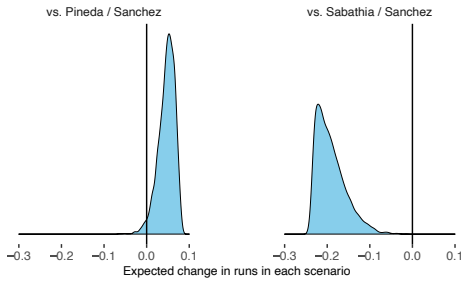


Figure 1. Of the two scenarios, Cain should only attempt to steal against the Sanchez–Pineda duo.

Notably, we get these expectations without multiple trials of either scenario. More generally, this model suggests that on average team managers are too conservative, leaving runs unrealized:

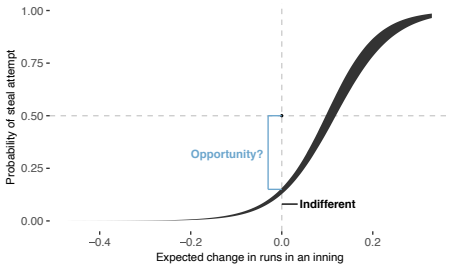


Figure 2. When the change in expected runs is zero, managers should be indifferent to attempted steals, saying go half the time.

The **black band** represents the range of variation across managers' decisions. At the intersection of **indifference**, managers tend to say steal only **10 percent** of the time, leaving opportunity.

The above is but one example of a more general approach that weighs probabilities of all possible outcomes to maximize expected utility. With broad implementation—jointly modeling the conditional probabilities of all relevant events—we can optimize decisions.

3 For value, compare an investment to free-agent costs

A fully-realized model will require significant effort from a team with deep experience in baseball, generative modeling, and Stan. To get the talent, we should compare cost to acquiring expected wins from free-agents. Each win above a *replacement-level* player costs about 10 million per year (Swartz 2017). As with free-agent value over replacement player, game-time decisions informed from more accurate probabilities should add wins over a season. The scope of what we can answer, moreover, goes beyond in-game strategy (player acquisitions, salary arbitration). More immediately, however, we can begin to implement this approach for specific events, with a scope closer to the example above, being mindful that information learnt are conditional upon unmodeled context.

4 For accuracy, compare model results to betting market odds

Measuring performance of a fully-realized model may seem tricky: *we only see the outcome of our decisions*. But we can, say, compare the accuracy of our estimates against the betting market where interested investors are trying to forecast game outcomes.

5 Conclusion

The mid-market Astros show teams can do more with information. Millions in additional revenue—and more wins—await discovery through a joint, probability model of all events from which we can maximize conditional expectations. Let's discuss how to draw the talent for a title worth our spend.

6 References

Carpenter, Bob, et. al. 2017. "Stan: A Probabilistic Programming Language." *Journal of Statistical Software* 76 (1): 1–32.

Luhnow, Jeff. 2018a. "How the Houston Astros are winning through advanced analytics." McKinsey Quarterly 13 June 2018: 1–9.

———. 2018b. "A view from the front lines of baseball's data-analytics revolution." McKinsey Quarterly 5 July 2018: 1–8.

Parmigiani, G. 2002. "Decision Theory: Bayesian." In *International Encyclopedia of the Social Behavioral Sciences*, 3327–34.

Swartz, Matt. 2017. "The Recent History of Free-Agent Pricing." https://www.fangraphs.com/blogs/the-recent-history-of-free-agent-pricing/.

6
Layout, hierarchy, and integration

6.1 Visual presentation is communication

6.1.1 Typography

When we consider the visual presentation of communication, we may first think about a data graphic. But consider this paragraph from *Elements of Style*[1], white space removed:

Vigorouswritingisconcise.Asentenceshouldcontainnounnecessarywords,aparagraphnounnecessarysentences,forthesamereasonthatad

[1] Strunk and White, *The Elements of Style*.

The visual presentation of communication involves all best practices in **typography** and **design**. Adding white space between words, just one of many components of typography, is an obvious decision. It makes the advice from Strunk and White[2] more readable, more understandable:

> Vigorous writing is concise. A sentence should contain no unnecessary words, a paragraph no unnecessary sentences, for the same reason that a drawing should have no unnecessary lines and a machine no unnecessary parts. This requires not that the writer make all his sentences short, or avoid all detail and treat subjects only in outline, but that every word tell. A single overstatement, wherever or however it occurs, diminishes the whole, and a carefree superlative has the power to destroy, for readers, the object of your enthusiasm.

[2] Their very-short, classic book on writing would not be in its 50th Edition were it not still valuable. Leading by example, this tiny book provides dos and don'ts with examples of each. Re-read.

Best practices in visual presentation of communication go well beyond spacing between words. Matthew Butterick[3] explains best practices, well, best. Typography is the visual component of the written word. "Typography is for the benefit of the reader":

> Most readers are looking for reasons to stop reading. . . . Readers have other demands on their time. . . . The goal of most professional writing is persuasion, and attention is a prerequisite for persuasion. Good typography can help your reader devote less attention to the mechanics of reading and more attention to your message.

[3] Matthew Butterick, "Butterick's Practical Typography" (https://practicaltypography.com/, 2018); Butterick credits a great deal to, among others, Robert Bringhurst, *The Elements of Typographic Style*, Third (Hartley & Marks, 2004).

The typographic choices in the example memos, section 4.9, and proposal, section 5.9, follow Butterick's advice:

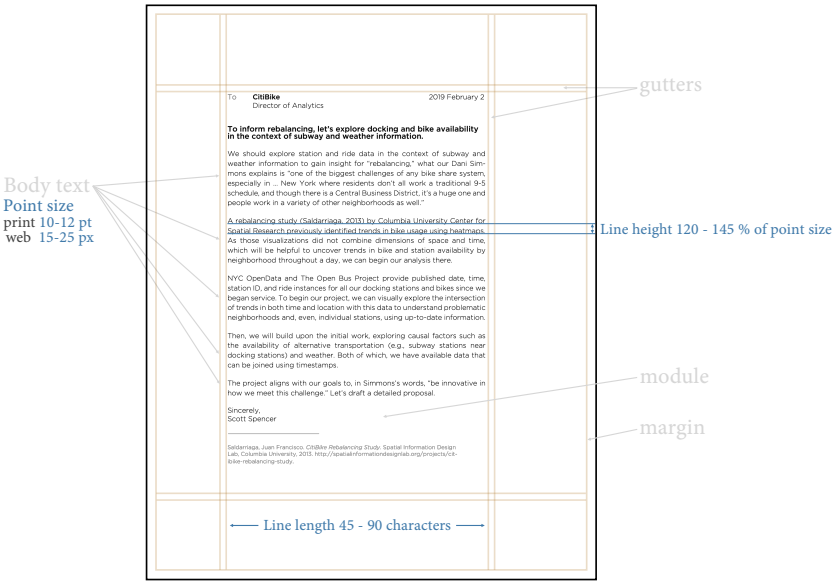


Figure 6.1: Basic typographic guidelines as implemented in examples.

Those best practices do more than aid readability. Experiments have demonstrated that "high quality typography can improve mood [of the reader]"[4], and the better their mood, the more likely they are to agree with you.

Butterick's recommendations, and as implemented in the example memo, are designed functionally. When designing communications for the interwebs, also consult *Web Typography*[5]. There will be occasions, however, when more creativity can be used in combination with function. Information graphics are an example. You may find inspiration in *Explorations in Typography*[6], which studies the creative placement of text.

[4] Kevin Larson and Rosalind Picard, "The Aesthetics of Reading," *MIT Affective Computing Lab*, January 2005, 1–12.

[5] Richard Rutter, *Web Typography*, A Handbook for Designing Beautiful and Effective Responsive Typography (Ampersand Type, 2017).

[6] Carolina de Bartolo, Stephen Coles, and Erik Spiekermann, *Explorations in Typography*, Second (101Editions, 2019).

6.1.2 *Grid systems and narrative layout*

Another aspect of typography and design rely on **grid systems**. A very basic grid is shown in figure 6.1, some of its components drawn in brown and labeled in gray: *gutters*, *module*, and *margin*. The gutters between the gridlines create *white space* that separate information placed into *columns*, *rows*, *modules*, or *spatial zones* (a spatial zone comprises multiple modules or rows or columns). Of course, the grid lines are not part of the final communication; we create them temporarily to layout and align information. That layout is informed by visual perception and the way we process information in a given

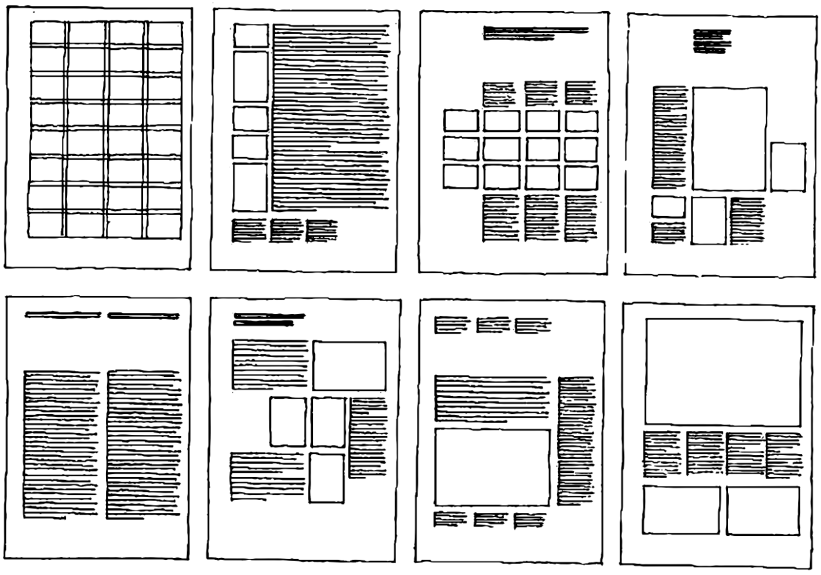
culture. When reading English, for example, we generally start processing the information from the top, left, our eyes scanning to the right, and then returning left and down, in a repeating zig-zag pattern. Hebrew is right to left. We call this type of narrtive structure *linear*.[7] And various graphic design choices can purposefully *or inadvertently* guide the reader through the material in other ways. Images, unlike sentences, create an *open* narrative structure, allowing us to reason differently.[8] We'll come back to this concept.

[7] Juuso Koponen and Jonatan Hildén, *Data Visualization Handbook*, First (Finland: Aalto Art Books, 2019).

[8] Koponen and Hildén, *Data Visualization Handbook*.

Grid systems can be much more complex. We are guided by Muller-Brockmann in his seminal reference, "Arranging surfaces and spaces into a grid creates conformity among texts, images and diagrams. The size of each implies its importance. Reducing elements in a grid suggests planning, intelligibility, clarity, and orderliness of design. One grid allows many creative ways to show relationships."[9] A grid with 8 rows by 4 columns and *gutter* spacing between the blocks, for example, can lead to numerous arrangements of disparate, but related, information:

[9] Josef Müller-Brockmann, *Grid Systems in Graphic Design*, A Visual Communication Manual for Graphic Designers, Typographers, and Three Dimensional Designers (ARTHUR NIGGLI LTD., 1996).



Yet the commonly aligned sides of word blocks, images, and data graphics can help connect related information. By connect, we mean the layout creates or enables a path that the audience's eye follows, a *scan* path. In this paragraph of text, you started reading at its beginning and followed horizontally until the end of the line, then scanned to the left beginning of the line below and repeated the process. In strip comics, the sequentially arranged images encourage a similar linear narrative. But other layouts enable an open narrative. These include radial layouts in which the order we scan relies on *focal points*,

which are prominent components due to, say, their size or color in relation to the surrounding information. Of note, in some circumstances we my intend a serial narrative within an open narrative. Consider labeling or numbering the features , using gestault principles, or both, to guide the audience.

Thus, as Muller-Brockmann explained, grids enable orderliness, adds credibility to the information, and induces confidence. Information presented with clear and logically set out titles, subtitles, texts, illustrations and captions will not only be read more quickly and easily but the information will also be better understood.

Exercise 6.1 (Identify grids for text). Try to identify placement of the (invisible) grid lines used to align information in the *Dodgers* proposal, end of Chapter 5, which is primarily text.

Exercise 6.2 (Identify grids for graphics). Consider the poster version of the information graphic *The Top 2000 loves the 70s & 80s*.[10] Try to identify placement of the (invisible) grid lines used for alignment.

[10] Nadieh Bremer, "The Top 2000 Loves the 70s & 80s," Personal, *Visual Cinnamon* (https://www.visualcinnamon.com/portfolio/top2000, December 2016).

6.2 *Combined meaning of words and images*

Words, graphics, and images — when combined — can provide to some extent what Doumont prescribed: *effective redundancy*. This is sometimes called dual coding. And to maximize their combination, we first consider that we process languages and images differently.[11] Words are read, and processed in linear fashion, *serially*, one after the other. Images, on the other hand, can be processed or understood as a whole, *in parallel*.

[11] Colin Ware, *Information Visualization: Perception for Design*, Fourth (Philadelphia: Elsevier, Inc, 2020).

Secondly, each type of medium conveys meaning differently; neither exactly overlap: *a description of an image never actually represents the image. Rather, . . . it is a representation of thinking about having seen a picture — it's already formulated in its own terms*.[12] Each is better at conveying certain types of messages. Sousanis puts it: "while image is, text is always about." Text is usually better for expressing abstract concepts, and procedure, such as logic or programming. Diagrams help when explaining structural relationships.

[12] Nick Sousanis, *Unflattening* (Cambridge, Massachusetts: Harvard University Press, 2015); paraphrasing Michael Baxandall, *Patterns of Intention: On the Historical Explanation of Pictures* (New Haven: Yale University Press, 1985).

We can benefit from various studies into the interplay of words and images are found in comics,[13], and extrapolate those concepts into information visualization. Done right, each informs and enriches the other.

[13] Neil Cohn, *The Visual Narrative Reader*, ed. Neil Cohn (Bloomsbury Academic, 2016); Sousanis, *Unflattening*; Scott McCloud, *Understanding Comics: The Invisible Art* (Kitchen Sink Press, 1993).

Images and graphics also enable a unique form of comparison, juxtaposing one image or encoding to another — or to the absence of another — to form meaning.

6.3 *Visually integrating graphics and text*

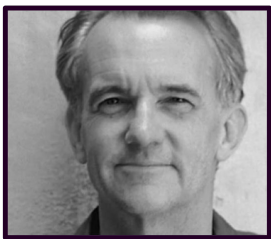


Figure 6.2: Edward Tufte has been hailed "The Leonardo da Vinci of data" by the New York Times. He is professor emeritus of Political Science, Statistics, and Computer Science at Yale University. His publications include four seminal textbooks in information design.

[14] Edward R. Tufte, *The Visual Display of Quantitative Information*, Second (Graphics Press, 2001).

Good design and typography also enable visual connections between words and sentences to, say, data graphics. Tufte[14] explains, at their best, graphics are instruments for reasoning about quantitative information. Often the most effective way to describe, explore, and summarize a set of numbers—even a very large set—is to look at pictures of those numbers. Furthermore, of all methods for analyzing and communicating statistical information, well-designed data graphics are usually the simplest and at the same time the most powerful. And if "a means of persuasion is a sort of demonstration," and we now agree with Aristotle that it is, then graphics are frequently the most effective way to demonstrate things, especially for understanding patterns and comparisons.

But it isn't a *Hobson's choice*, words or graphics. Instead, we should use both. Tufte explains how they work together: "The principle of data/text integration is: data graphics are paragraphs about data and should be treated as such."

Visual displays may be integrated directly within the text. Tufte's book is a living example, and explains the approach:

> We were able to integrate graphics right into the text, sometimes into the middle of a sentence, eliminating the usual separation of text and image — one of the ideas *Visual Display* advocated.

Experiments support Tufte's advice. The *Data Visualization Handbook* summarizes an experiment of eye-tracking movements and comprehension when reading communications in various layouts[15], we learned that layouts that integrate images within text columns improve communication over both radial layouts and layouts that separte text from images. The integrated approach promoted careful reading of the text between images while layouts separating text from images promoted the reading of a title, skipping the body text, and focusing on the images. Radial layouts were reviewed more quickly than linear, integrated text-image layouts, and less information was retained.

[15] Jana Holsanova, Henrik Rahm, and Kenneth Holmqvist, "Entry Points and Reading Paths on Newspaper Spreads: Comparing a Semiotic Analysis with Eye-Tracking Measurements," *Visual Communication* 5, no. 1 (February 2006): 65–93.

For effective integration, visual display need only be large enough to clearly convey the information as intended for our audience in the manner to be consumed. To make the point, consider the word-sized graphics Tufte[16] calls **sparklines**:  . Also note that when the graphic is large enough to include annotation,

[16] Tufte, *Beautiful Evidence*.

> The principle of text/graphic/table integration also suggests that the same typeface be used for text and graphic and, further, that ruled lines separating different types of information be avoided.

Exercise 6.3 (Identify exemplary data graphic paragraph). Locate two or three narratives with data graphics as paragraphs that you believe the graphic helped persuade audiences of the point of the narrative. Explain why the graphic explained better than words as used.

6.3.1 *Annotating data graphics with words*

Annotations add explanations and descriptions to introduce the graph's context, which is important for almost any audience. Annotation plays a crucial role in asynchronous data storytelling as the surrogate for the storyteller. They can also explain how to read the graph, which helps readers unfamiliar with the graph — whether a simple line chart or an advanced technique like a treemap or scatterplot. When done right, the annotation layer will not get in the way for experienced users. Consider, for example, figure 6.3.

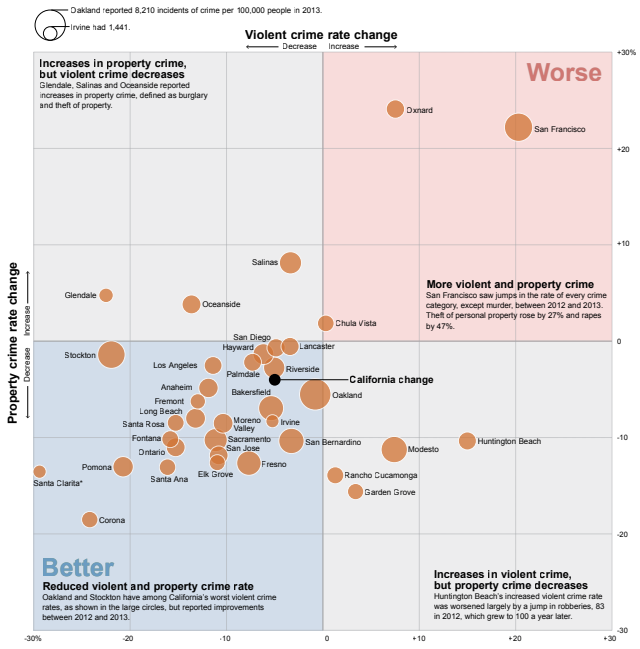


Figure 6.3: Example of data graphic containing annotation to assist the audience.

From a cognitive perspective, Ware writes that "plac[ing] explanatory text as close as possible to the related parts of a diagram, and us[ing] a graphical linking method" will "reduce [the] need to store information temporarily while switching back and forth between locations."[17] This example, published in a newspaper article[18], displays a scatter plot that encodes the rate change of crime on the x-axis, change of property crime on the y-axis, and rate of crimes as size of the location or point. Note the plot is segmented into quadrants,

[17] Ware, *Information Visualization*.

[18] Jon Schleuss and Rong-Cong Lin II, "California Crime 2013," *Los Angeles Times*, 2013.

color-coded to indicate better and worse conditions, and annotations are overlain that explain how to interpret the meaning of a data point located within quadrants of the graphic. The various annotations greatly assist its general audience in decoding the data and considering insights.

6.3.2 *Visually linking words with graphics*

Placement of data graphics within words and annotating graphics with words are the first step in integrating the information. Another best practice includes using color encodings and other explicit markings, linking words to encodings, such as adding lines connecting related information[19]:

> The link between the narrative and the visualization helps the reader discern what item in the visualization the author is referencing in the text. Create links with annotation, color, luminosity, or lines.

[19] Riche et al., *Data-Driven Storytelling*.

For example, color words in annotations on a data graphic and in the paragraphs surrounding that graphic with the same hue as used in the data encodings of the graphic. Academic Matthew Kay[20] provides example uses of color for linking words to data encodings.

We find another great example of linking paragraphs with illustrations in Byrne's revision of Euclid's first six books.[21]

[20] Matthew Kay, "Figures" (www.mjskay.com/figures/, August 2015).

[21] Oliver Byrne, *The first six books of the elements of Euclid in which coloured diagrams and symbols are used instead of letters for the greater ease of learners*, Bibliotheca universalis (Köln: TASCHEN, 2017).

6.3.3 *Linking multiple graphics*

If individual graphs reveal information and structure from the data, an ensemble of graphs can multiply the effect. By ensemble, we mean multiple graphs simultaneously displayed, each containing different views of the data with common information linked together by various techniques. And while William Cleveland[22] describes "brushing and linking" — where items selected in on one visual display highlights the same subset of observations in another visual display — as an interactive tool, he effectively shows the technique by highlighting the same data across static displays. Another author[23] provides a nice example, walking through use of ensembles in exploring data quality, comparing models, and presenting results. As the authors explain,

[22] William S Cleveland, *The Elements of Graphing Data* (Wadsworth, 1985).

[23] Antony Unwin and Pedro Valero-Mora, "Ensemble Graphics," *Journal of Computational and Graphical Statistics* 27, no. 1 (December 2018): 157–65.

> Coherence in effective ensembles covers many different aspects: a coherent theme, a coherent look, consistent scales, formatting, and alignment. Coherence facilitates understanding.

The additional effort for coherence "are more design than statistics, but they are driven by the statistical information to be conveyed, and it is therefore essential that statisticians concern themselves with

them." Along with using the same theme styles, their choice of **placement** is informed by best practices in graphic design, which apply a **grid system**, already discussed.

7
Visual design and perception

7.1 Why review data graphically?

The value of data graphics can be grasp from a brief analysis of the following four datasets (1-4) of (x, y) data in table 7.1 from a famous data set:

Table 7.1: These four simple datasets are known as Anscombe's Quartet.

1		2		3		4	
x	y	x	y	x	y	x	y
10	8.04	10	9.14	10	7.46	8	6.58
8	6.95	8	8.14	8	6.77	8	5.76
13	7.58	13	8.74	13	12.74	8	7.71
9	8.81	9	8.77	9	7.11	8	8.84
11	8.33	11	9.26	11	7.81	8	8.47
14	9.96	14	8.10	14	8.84	8	7.04
6	7.24	6	6.13	6	6.08	8	5.25
4	4.26	4	3.10	4	5.39	19	12.50
12	10.84	12	9.13	12	8.15	8	5.56
7	4.82	7	7.26	7	6.42	8	7.91
5	5.68	5	4.74	5	5.73	8	6.89

For most of us, reviewing the table for comparing the four datasets[1] is cognitively taxing, and especially when scanning for differences in the relationships between x and y across datasets. Processing the data this way occurs sequentially; we review data pairs with focused attention. And, here, summary statistics do not differentiate the datasets. All x variables share the same mean and standard deviation (table 7.2). So do all y variables.

[1] F J Anscombe, "Graphs in Statistical Analysis," *The American Statistician* 27, no. 1 (February 1973): 17–21.

Table 7.2: The mean and standard deviation per dataset are identical.

	1		2		3		4	
	x	y	x	y	x	y	x	y
mean	9.00	7.50	9.00	7.50	9.00	7.50	9.00	7.50
sd	3.32	2.03	3.32	2.03	3.32	2.03	3.32	2.03

Further, the linear regression on each dataset (table 7.3) suggests

that the (x, y) relationship across datasets are the same. Are they?

Table 7.3: Linear regression coefficients across datasets are practically identical.

Parameter	Mean	Std Err	t-val	p-val
Dataset 1				
(Intercept)	3.000	1.125	2.667	0.026
x	0.500	0.118	4.241	0.002
Dataset 2				
(Intercept)	3.001	1.125	2.667	0.026
x	0.500	0.118	4.239	0.002
Dataset 3				
(Intercept)	3.002	1.124	2.670	0.026
x	0.500	0.118	4.239	0.002
Dataset 4				
(Intercept)	3.002	1.124	2.671	0.026
x	0.500	0.118	4.243	0.002

A well-crafted visual display, however, can instantly illuminate any differing (x, y) relationships among the datasets. To demonstrate, we arrange four scatterplots in figure 7.1 showing the relationships between (x,y), one for each dataset. Overlain on each, we show the linear regression calculated above.

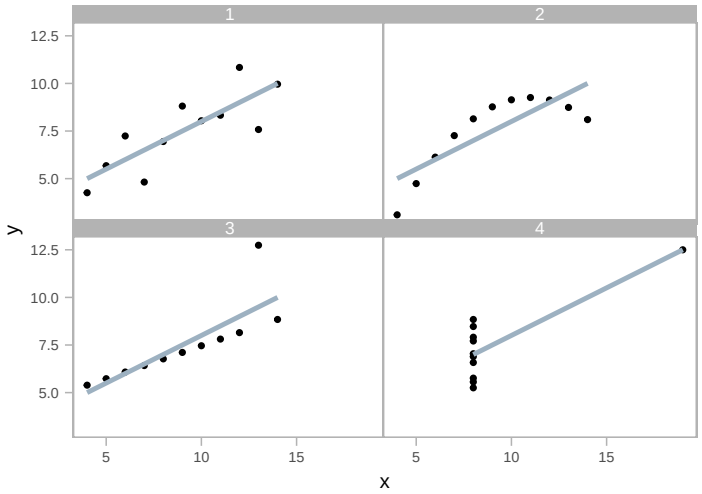


Figure 7.1: The differing (x, y) relationships among datasets in Anscombe's Quartet become instantly clear when visualized.

As the example shows, exploratory data analysis using visual and spatial representations add understanding. It allows us to find patterns in data, *detecting* or recognizing the geometry that encodes the values, *assembling* or grouping these detected elements, and *estimating* the relative differences between two or more quantities.[2] In estimating, we first *discriminate* between data: we judge whether **a** is equal to **b**. Then we *rank* the values, judging whether **a** is greater than, less

[2] Cleveland, *The Elements of Graphing Data*; William S Cleveland, *Visualizing Data* (Hobart Press, 1993).

than, or equal to **b**. Finally, we consider the *ratio* between them using encoded geometries (*e.g.*, relative distance from a common line). Unlike with sequential processing required for table lookups, pattern recognition — and outliers from those patterns — seem to occur in parallel, quickly because we are attuned to *preattentive* attributes.[3]

[3] Ware, *Information Visualization*.

7.2 *Reasoning with images*

We previously mentioned how, unlike processing text in linear fashion, images enable an open narrative, which we process differently.[4]

We may also combine linear and open narrative structures in various ways.[5]

...

[4] Koponen and Hildén, *Data Visualization Handbook*; Sousanis, *Unflattening*; Stephen Michael Kosslyn, William L. Thompson, and Giorgio Ganis, *The Case for Mental Imagery*, Oxford Psychology Series 39 (New York: Oxford University Press, 2006); Baxandall, *Patterns of Intention*.

[5] E Segel and J Heer, "Narrative Visualization: Telling Stories with Data," *IEEE Transactions on Visualization and Computer Graphics* 16, no. 6 (November 2010): 1139–48.

7.3 *Components of a graphic*

Graphics include a coordinate system, arranged spatially, and have numerous attributes that we may make visible in some way, if it helps users understand the graphic. These components can be understood in two categories. Those encoding data (*data-ink*) and all the rest (*non-data-ink*).

7.3.1 *Non-data-ink*

We'll use an `R/ggplot` implementation of graphics to discuss these components[6]. Figure 7.2 shows the names for most of the non-data-ink components of a visual display.

[6] Other implementations of graphics will typically name the components of a graphic similarly.

Most of the aesthetics of each labeled component can be set, modified, or removed using the `ggplot` function `theme()`, which takes plot components as parameters. We set parameters equal to other formatting functions like, say, `element_text()` for formatting its typography, `element_rect()` for formatting its various shape or coloring information, or `element_blank()` to remove entirely the element. In Figure 7.2, for example, we set the panel border attribute linetype and color using `theme(panel.border = element_rect(color = "gray60", linetype = "dashed", fill = NA))`. We can use the `ggplot` function `annotate()` to include words or draw directly onto the plotting area. Figure 7.3 shows the basic code structure.

In the pseudocode of figure 7.3, we map variables in the data to aesthetic characteristics of a plot that we see through `mapping = aes(<aesthetic> = <variable>)`[7]. Particular aethetics depend on the type of geometric encoding we choose. A scatter plot, say,

[7] Note that <...> is not part of the actual code. It represents, for purposes of discussion, a placeholder that the coder would replace with appropriate information.

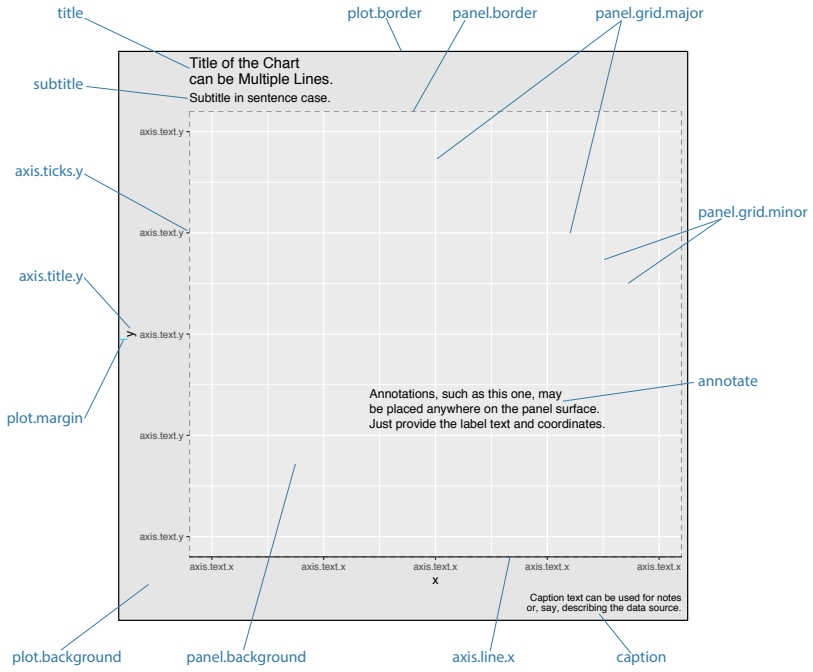


Figure 7.2: Edward Tufte advocates maximizing the data-ink ratio within reason. Some non-data-ink can be critical to understanding the data. For each element's marking, coloring, size, shape, orientation, or transparency setting, ask whether it maximizes our audience's understanding of the intended insight.

```
# load grammar of graphics
library(ggplot2)

p <-

  # functions for data ink

  ggplot(data = <data>,
         mapping = aes(<aesthetic> = <variable>,
                       <aesthetic> = <variable>,
                       <...> = <...>) +
  geom_<type>(<...>) +
  scale_<mapping>_<type>(<...>) +
  coord_<type>(<...>) +
  facet_<type>(<...>) +
  <...> +

  # functions for non-data ink

  labs(<...>) +
  theme(<...> = <...>) +
  annotate(<...>) +
  <...>
```

Figure 7.3: GGplot's functions are set up as layers. We may use more than one geometry or annotation.

would at least include `x` and `y` aesthetics. The geometric encodings are created through functions named for their geometries: *e.g.*, `geom_point(<...>)` for the scatter plot, which we generalize to `geom_<type>(<...>)`. The geometry is then mapped onto a particular coordinate system and scale: `coord_<type>(<...>)` and `scale_<mapping>_<type>(<...>)`, respectively. Finally, we annotate and label the graph. These can be thought as layers that are added (+) over each previous layer.

The remaining markings of a graphic are the data-ink, the data encodings, discussed next.

7.3.2 Data-ink

Encodings depend on data type, which we introduced in section 2.3.1. As Andrews[8] explains, "value types define how data is stored and impact the ways we turn numbers into information." To recap, these types are either qualitative (nominal or ordered) or quantitative (interval or ratio scale).

[8] Andrews, *Info We Trust*.

"A component is **qualitative**" and nominal, Bertin explains, "when its categories are not ordered in a universal manner. As a result, they can be reordered arbitrarily, for purposes of information processing."[9] The qualitative categories are equidistant, of equal importance. Considering Citi Bike, labeled things such bikes and docking stations are qualitative at the nominal level.

[9] Jacques Bertin, *Semiology of Graphics: Diagrams Networks Maps* (Redlands: ESRI Press, 2010).

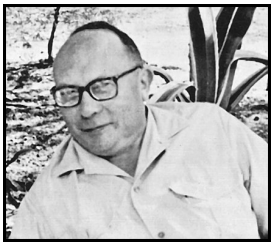


Figure 7.4: Jacques Bertin was a French cartographer and theorist, trained at the Sorbonne, and a world renowned authority on the subject of information visualization. He later assumed various leadership positions in research and academic institutions in Paris. *Semiology of Graphics*, originally published in French in 1967, is internationally recognized as a foundational work in the fields of design and cartography.

"A component is **ordered**, and only ordered, when its categories are ordered in a single and universal manner" and "when its categories are defined as equidistant." Ordered categories cannot be reordered. The bases in baseball are ordinal, or ordered: first, second, third, and home. Examples of qualitative ordering may be, say, temporal: morning, noon, night; one comes before the other, but we would not conceptually combine morning and night into a group of units.

When we have countable units on the **interval level**, the data of these counts are **quantitative**. A series of numbers is quantitative when its object is to specify the variation in distance among the categories. We represent these numerically as integers. The number of bike rides are countable units. The number of stolen bases in baseball are countable units. We represent these as integers.

Finally, **ratio-level**, **quantitative** values represent countable units per countable units of something else. The number of bike rides per minute and the number of strike outs per batter would be two examples, represented as fractions, real numbers.

The first and most influential structural theory of statistical graphics is found the seminal reference, *Semiology of Graphics*[10].

[10] Jacques Bertin, *Semiology of Graphics* (University of Wisconsin Press, 1983).

Based on Bertin's practical experience as a cartographer, part one of this work is an unprecedented attempt to synthesize principles of graphic communication with the logic of standard rules applied to writing and topography.

Part two brings Bertin's theory to life, presenting a close study of graphic techniques, including shape, orientation, color, texture, volume, and size, in an array of more than 1,000 maps and diagrams. Here are those encoding types:

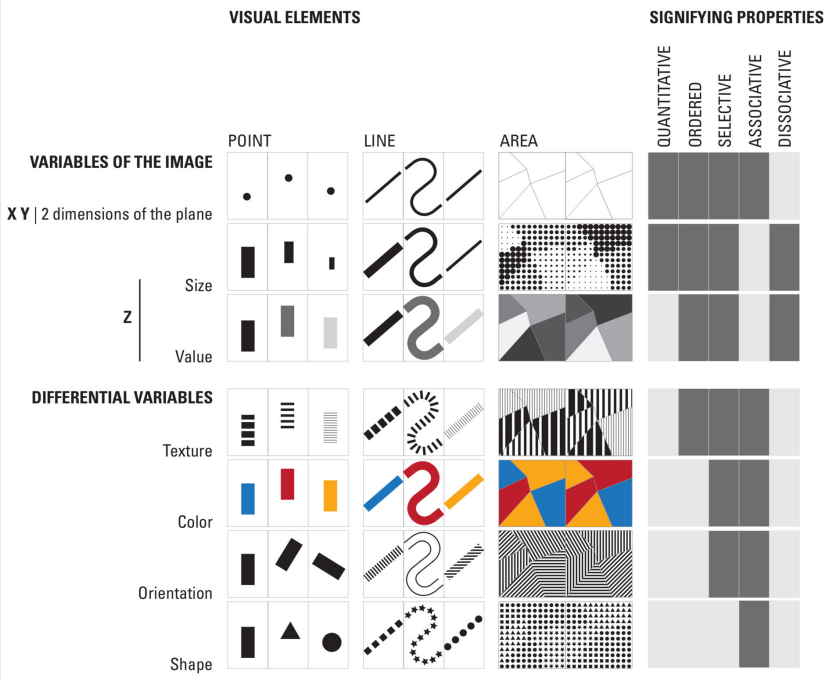


Figure 7.5: Bertin's illustration of the possible encoding forms for data.

Less commonly discussed is Bertin's update[11] to his original work. In the update, after defining terms he reviews the natural properties of a graphic image. The essence of the graphic image is described in three dimensions. The first two describe spatial properties (e.g. x and y axes) while the third dimension (denoted z) encodes the characteristics of each mark — e.g. size, value, texture, color, orientation, shape — at their particular spatial (x, y) locations.

Bertin's ideas, over 50-years old, have proven reliable and robust.[12]

[11] Bertin, *Semiology of Graphics*.

[12] Alan M. MacEachren, "(Re)Considering Bertin in the Age of Big Data and Visual Analytics," *Cartography and Geographic Information Science* 46, no. 2 (March 2019): 101–18, `https://doi.org/10.1080/15230406.2018.1507758`; **???**

7.3.3 *Grammar*

Graphics are not charts, explains Wilkinson[13]:

> We often call graphics charts. There are pie charts, bar charts, line charts, and so on. [We should] shun chart typologies. Charts are usually instances of much more general objects. Once we understand

[13] Leland Wilkinson, *The Grammar of Graphics*, Second (Springer, 2005).

> that a pie is a divided bar in polar coordinates, we can construct other polar graphics that are less well known. We will also come to realize why a histogram is not a bar chart and why many other graphics that look similar nevertheless have different grammars. . . . Elegant design requires us to think about a theory of graphics, not charts.

We should think of chart names only as a shorthand for what they do. To broaden our ability to represent comparisons and insights into data, we should instead consider their representation as types of measurement: length along a common baseline, for example, or encoding data as color to create Gestalt groupings.

In Leland Wilkinson's influential work, he develops a grammar of graphics. That grammar respects a fundamental limitation, a difference from pictures and other visual arts:

> We have only a few rules and tools. We cannot change the location of a point or the color of an object (assuming these are data-representing attributes) without lying about our data and violating the purpose of the statistical graphic — to represent data accurately and appropriately.

Leland categorizes his grammar:

> **Algebra** comprises the operations that allow us to combine variables and specify dimensions of graphs. **Scales** involves the representation of variables on measured dimensions. **Statistics** covers the functions that allow graphs to change their appearance and representation schemes. **Geometry** covers the creation of geometric graphs from variables. **Coordinates** covers coordinate systems, from polar coordinates to more complex map projections and general transformations. Finally, **Aesthetics** covers the sensory attributes used to rep- resent graphics.

He discusses these components of graphics grammar in the context of data and its extraction into variables. He also extends the discussion with **facets** and **guides**.

How do we perceive data encoded in this grammar?

7.4 Perceptions of visual data encodings

We assemble mental models of grouping through differences in similarity, proximity, enclosure, size, color, shading, and hue, to name a few. In figure 7.1, for example, we recognize dataset three as having a grouped linear relationship with one outlier based on proximity. Using shading, for example, we can separate groups of data. In the left panel of figure 7.6, we naturally see two groups, one gray, and the other black, which has an outlier. We could even enclose the outlier to further call attention to it, as shown on the right panel.

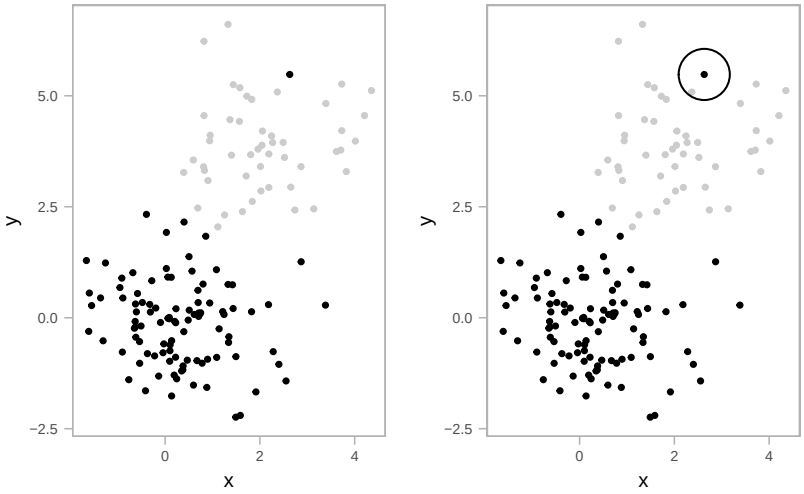


Figure 7.6: We can use preattentive attributes to separate data categorically and call attention to particular aspects of that data.

Several authors[14] provide in-depth reviews of these ideas. We can, and should, use these ideas to assist us in understanding and communicating data through graphical displays.

Graphical interpretation, however, comes with its own limitations. Our accuracy in estimating the quantities represented in visual encoding depends on the geometries used for encoding. In other words, it can be easy for us, and less familiar readers, to misinterpret a graph. Consider the example in Figure 7.7 where the slope of the trend changes rapidly. Considering the left panel alone, it may seem deviations from the fitted line decrease as x increases. But the residuals encoded in the right panel show no difference.

[14] Ware, *Information Visualization*; Bertin, *Semiology of Graphics*; Isabel Meirelles, *Design for Information*, An Introduction to the Histories, Theories, and Best Practices Behind Effective Information Visualizations (Rockport, 2013); C G Healey and J T Enns, "Attention and Visual Memory in Visualization and Computer Graphics," *IEEE Transactions on Visualization and Computer Graphics* 18, no. 7 (May 2012): 1170–88.

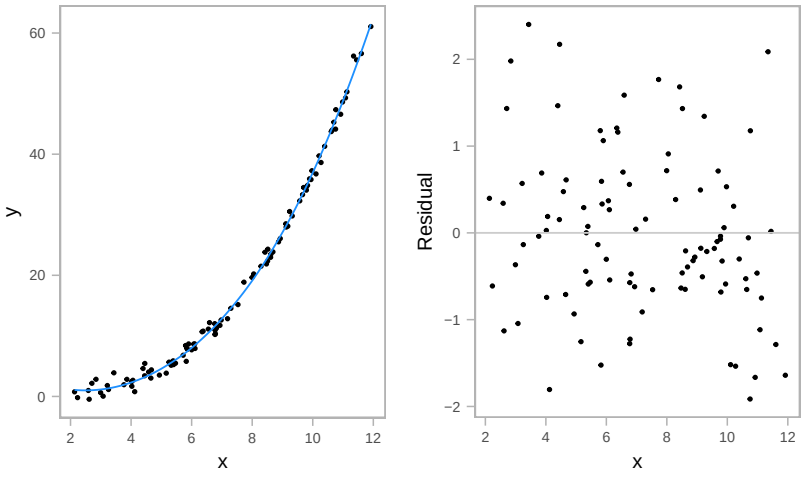


Figure 7.7: Without careful inspection, it may seem that deviations from the fitted line decrease as x increases. The plot of residuals, however, shows the reverse is true.

The misperception arises if we mistakenly compare the minimal distance from each point to the fitted line instead of comparing the vertical distance to the fitted line. Cleveland[15] has thoroughly reviewed our perceptions when decoding quantities in two or more curves, color encoding (hues, saturations, and lightnesses for both categorical and quantitative variables), texture symbols, use of visual reference grids, correlation between two variables, and position along a common scale. Empirical studies[16] have quantified our accuracy and uncertainty when judging quantity in a variety of encodings.

The broader point is to become aware of issues in perception and consider multiple representations to overcome them. Several references mentioned in the literature review delve into visual perception and best practices for choosing appropriate visualizations. The *Data Visualization Handbook*[17], for example, usefully arranges data types within visual variables and orders them by our accuracy in decoding, shown in figure 7.8:

[15] Cleveland, *The Elements of Graphing Data*.

[16] William S Cleveland and Robert McGill, "Graphical Perception: Theory, Experimentation, and Application to the Development of Graphical Methods," *Journal of the American Statistical Association* 79, no. 387 (September 1984): 531–54; Jeffrey Heer and Michael Bostock, "Crowdsourcing Graphical Perception: Using Mechanical Turk to Assess Visualization Design," in *Proceedings of the Sigchi Conference on Human Factors in Computing Systems*, 2010, 203–12.

[17] Koponen and Hildén, *Data Visualization Handbook*.

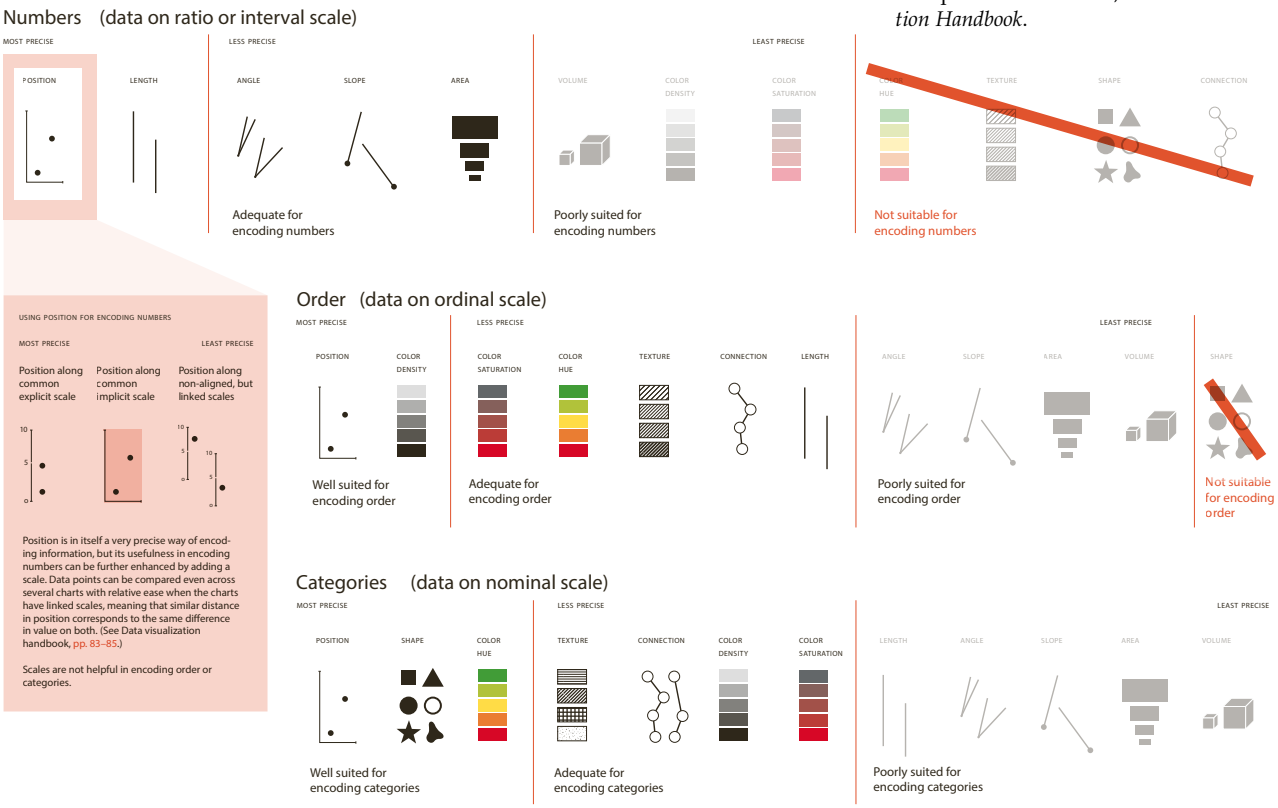
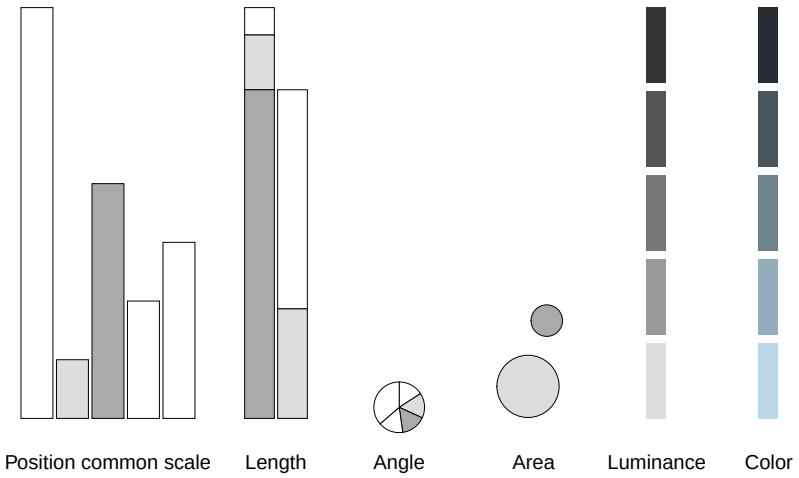


Figure 7.8: Visual variables, organized by how well they are suited for representing data measured on each type of scale.

Placing encodings in the context of chart types, figure 7.9, we decode them from more to less accurate, position encoding along common scales (*e.g.*, bar charts, scatter plots), length encodings (*e.g.*,

stacked bars), angles (*e.g.*, pie charts), circular areas (*e.g.*, bubble charts), luminance, and color[18]:



[18] Tamara Munzner, *Visualization Analysis and Design* (CRC Press, 2014).

Figure 7.9: We gage position along a common scale more accurately than length, which we gauge more accurately than angle or area. Luminance or color are typically reserved for encoding quantities in a third dimension.

A thorough visual analysis may require multiple graphical representations of the data, and each require inspection to be sure our interpretation is correct.

7.4.1 Color

As mentioned, We can encode data using color spaces, which are mathematical models. The common color model RGB has three dimensions — red, green, and blue, each having a value between 0 and 255 (2^8) — where those hues are mixed to produce a specific color.

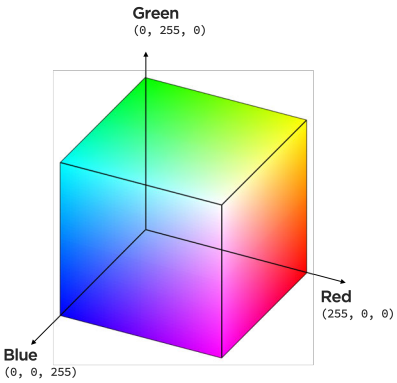


Figure 7.10: The RGB colorspace represented in three dimensions, with values from 0-255. Of note, the curious value of 255 originates from when computers, which store in formation in bits. $2^8 = 256$ values, starting at 0.

Notice the hue, chroma, and luminance of this colorspace, figure 7.11, seems to have uneven distances and brightness along wavelength.

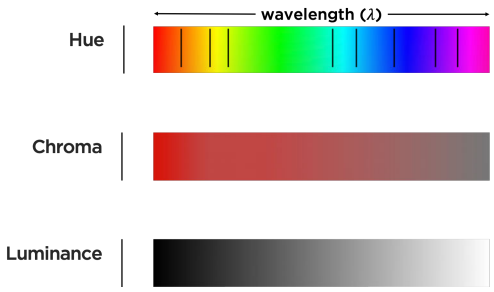


Figure 7.11: The RGB colorspace represented as hue, chroma, and luminance. The uneven distances in wavelength between hues creates misperceptions in data encodings in this color space.

Let's consider how we might, as illustrated in figure 7.12, map data to these characteristics of color.

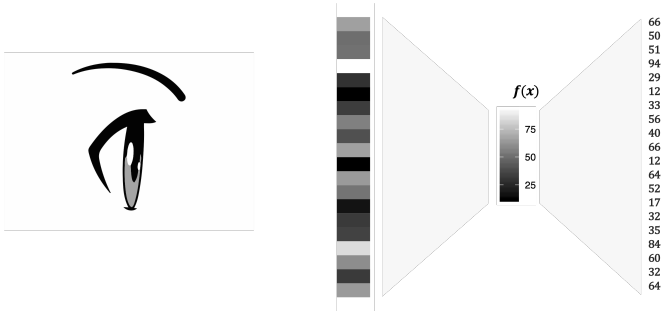


Figure 7.12: How can we map data to light, whether using its hue, chroma, or luminance?

Luminance is the measured amount of light coming from some region of space. Brightness is the perceived amount of light coming from that region of space. Perceived brightness is a very nonlinear function of the amount of light emitted. That function follows the power law:

$$\text{perceived brightness} = \text{luminance}^n \tag{7.1}$$

where the value of n depends on the size of the patch of light. Colin Ware[19] reports that, for circular patches of light subtending 5 degrees of visual angle, n is 0.333, whereas for point sources of light n is close to 0.5. Let's think about this graphically. Visual perception of an arithmetical progression depends upon a physical geometric progression.[20] In a simplification shown in figure 7.13, this means: if the first 2 steps measure 1 and 2 units in rise, then step 3 is not only 1 unit more (that, is, 3 in an arithmetical proportion), but is twice as much (that is, 4 in a geometric proportion. The successive steps then measure 8, 16, 32, 64 units.

[19] Ware, *Information Visualization*.

[20] Josef Albers, *Interaction of Color* (Yale University Press, 2006).

Color intervals are the distance in light intensity between one color and another, analogous to musical intervals (the relationship between notes of different pitches).

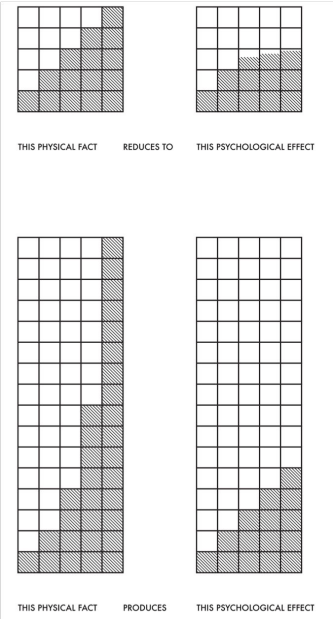


Figure 7.13: As Albers illustrates, Weber's law, applied to creating color steps we perceive as evenly spaced requires we convert from an arithmetical progress to a geometric progression.

Uneven wavelengths between what we perceive as colors, as we saw in the RGB color space, results in, for example, almost identical hues of green across a range of its values while our perception of blues change more rapidly across the same change in values. We also perceive a lot of variation in the lightness of the colors here, with the cyan colors in the middle looking brighter than the blue colors.

Figure 7.14: As with luminance, hue values in the RGB color space fail to uniformly scale across values.

Other color spaces show changes in color we perceive as uniform. Humans compute color signals from our retina cones via an opponent process model, which makes it impossible to see reddish-green or yellowish-blue colors. The International Commission on Illumination (CIE) studied human perception and re-mapped color into a space where we perceive color changes uniformly. Their CIELuv color model has two dimensions — u and v — that represent color scales from red to green and yellow to blue.

More modern color spaces improve upon CIELuv by mapping colors

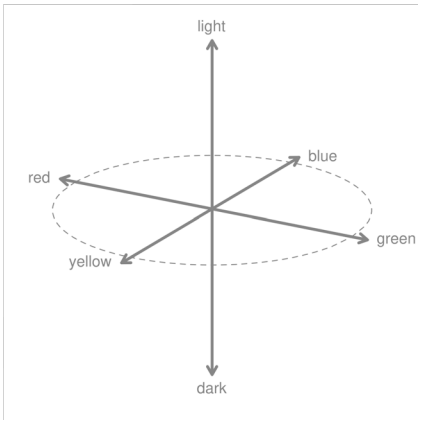


Figure 7.15: CIELuv color model has two dimensions — u and v — that represent color scales from red to green and yellow to blue.

as perceived into the familiar and intuitive Hue-Chroma-Luminance dimensions. Several modern color spaces, along with modification to accommodate colorblindness, are explained in the expansive *Data Visualization Handbook*[21]. In contrast with the perceptual change shown with an RGB colorspace of figure 7.14, the change in value shown in figure 7.16 our green-to-blue hues in 10 equal steps using the HCL model are now perceptually uniform.

[21] Koponen and Hildén, *Data Visualization Handbook*.

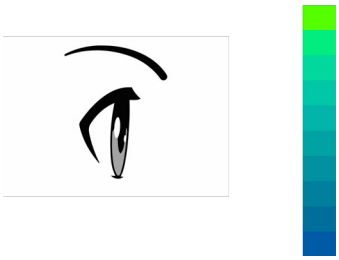


Figure 7.16: In a perceptually uniform colorspace creates an even gradient.

With categorical data, we do not want one color value to appear brighter than another. Instead, we want to choose colors that both separate categories while holding their brightness level equal (7.17). For an implementation of perceptually uniform color spaces in `R`, review the package `colorspace`.

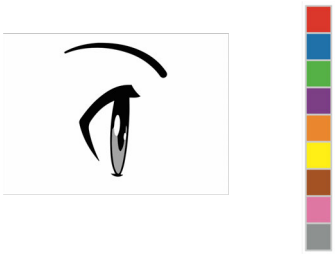


Figure 7.17: Perceptually uniform color spaces also help in distinguishing categorical data.

7.4.2 *Relativity of color*

Notice, by the way, that each of the 10 values shown in figure 7.16 appear to show a gradient in hue. That isn't the case, the hue is uniform. Our eyes, however, perceive a gradient because the adjacent values create an edge contrast. Humans have evolved to see edge contrasts, as in figure 7.18.

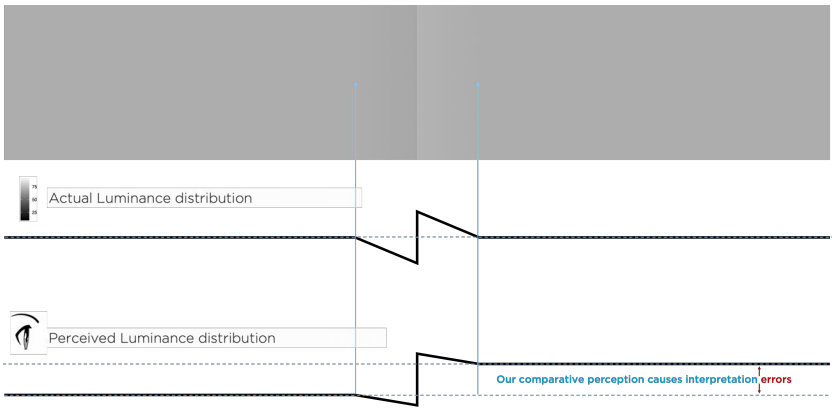


Figure 7.18: We see comparative — not absolute — luminance value. Adjacent data encoded by color may cause us to misperceive the value we're inspecting.

We see comparative — not absolute — luminance value. The edge between the left and right gray rectangles, created by a luminance adjustment tricks us into seeing each rectangle as uniform, though the outer portions have the same luminance. Need proof? Cover the edge portion between them.

Similarly, our comparative perception has implications for how to accurately represent data using luminance. Background or adjacent luminance — or hue or saturation — can influence how our audience perceives our data's encoded luminance value. The small rectangles in the top row all have the same luminance, though they appear to change. This misperception is due to the background gradient.

> **Exercise 7.1** (Identify graphics data-ink encodings.). Locate two or three graphics on the internet, each with different types of data-ink encodings you believe are *well-designed*. Be adventurous. Describe those encodings without using names of charts.
>
> Now locate two graphics, each with different types of data-ink encodings you believe are *problematic*. Describe the encodings, what makes them problematic, and suggest a more appropriate encoding.

7.5 *Maximize information in visual displays*

MAXIMIZE THE INFORMATION in visual displays **within reason**. Tufte[22] measures this as the **data-ink ratio**:

[22] Edward R. Tufte, "Data-Ink Maximization and Graphical Design," in *The Visual Display of Quantitative Information* (Graphics Press, 2001), 1–15.

Figure 7.19: Background information causes us to misperceive that each row of small rectangles are encoded with identical luminance values.

$$\begin{aligned}\text{data-ink ratio} &= \frac{\text{data-ink}}{\text{total ink used to print the graphic}} \\ &= \text{proportion of a graphic's ink devoted to the non-redundant display of data-information} \\ &= 1.0 - \text{proportion of a graphic that can be erased without loss of data-information}\end{aligned} \tag{7.2}$$

That means identifying and removing non-data ink. And identifying and removing redundant data-ink. Both within reason. Just how much requires experimentation, which is arguably the most valuable lesson[23] from Tufte's classic book, *The Visual Display of Quantitative Information*. In it, he systematically redesigns a series of graphics, at each step considering what helps and what may not. Tufte, of course, offers his own view of which versions are an improvement. His views are that of a designer and statistician, based on his experience and theory of graphic design.

Some of his approaches have also been subject to experiments[24], which we should consider within the context and limitations of those experiments. More generally, for any important data graphic for which we do not have reliable information on its interpretability, we should perform tests on those with a similar background to our intended audiences.

Let's reconsider the example figure from Knaflic:

[23] Indeed, most criticisms of Tufte's work misses the point by focusing on the most extreme cases of graphic representation within his process of experimentation, completely losing what we should learn — how to reason and experiment with data graphics. Focus on learning the reasoning and experimentation process.

[24] E. W. Anderson et al., "A User Study of Visualization Effectiveness Using EEG and Cognitive Load," *Computer Graphics Forum* 30, no. 3 (June 2011): 791–800, https://doi.org/10.1111/j.1467-8659.2011.01928.x.

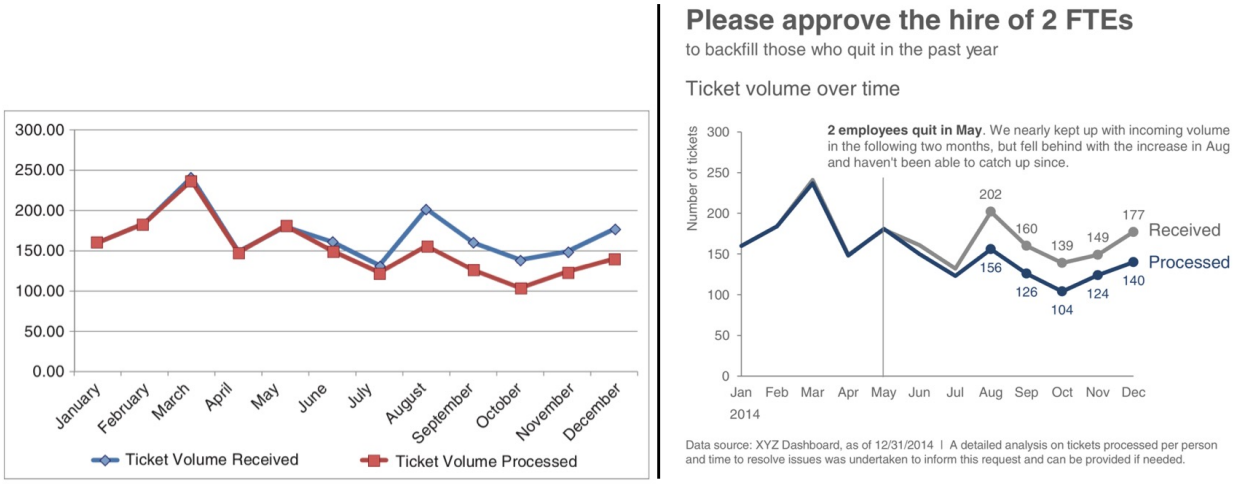


Figure 7.20: Knaflic systematically changes a graphic, beginning with the original, default graph on the left, and finishing with the graphic on the right.

Exercise 7.2 (Identify non- and redundant-data ink removed.). Compare Knaflic's before-and-after example. Try to articulate all differences. Consider whether her changes follow Tufte's principles, and whether each of her changes would improve her audience's understanding of the intended narrative and supporting evidence.

8
Visually encoding data, common and xenographic

8.1 Encoding data-ink, common graphics

Resources abound for encoding and coding common graphics. In an award-winning graphic form[1] we find taxonomies for common graphics, and an analysis of basic charts. Available elsewhere[2]. Again consulting the *Data Visualization Handbook* will explain common statistical graphics, including *bar charts*, *dot plots*, *line charts* and their variants, like slopegraphs, streamgraphs, bumps charts, cycle plots, sparklines, pie and donut charts, scatterplots (scatter or x-y, strip plot, beeswarm plot), bubble charts, heatmaps, box plots, violin plots, and many more. We should not try to memorize each type. Instead, we should understand how they work using the language and ideas from section 7.4. Apply the the advice about studying metaphor and rhetorical figures (section 5.6) when constructing graphics, too:

> Seeing just a few examples invites direct imitation of them, which tends to be clumsy. Immersion in many examples allows them to do their work by way of a subtler process of influence, with a gentler and happier effect on the resulting style.

[1] Yan Holtz and Conor Healy, "From Data to Viz," 2018.

[2] Kieran Healy, *Data Visualization* (Princeton University Press, 2018); Knaflic, *Storytelling with Data*; Cleveland, *Visualizing Data*; Cleveland, *The Elements of Graphing Data*.

And we have already used many common graphics in previous chapters. We encountered bar charts (*e.g.*, figures 5.7 and 7.9), and other instances of graphics in their natural environments, like histograms (*e.g.*, figure 3.1), which are very similar to bar charts, scatterplots (*e.g.*, figures 3.2, 3.9), linecharts (*e.g.*, figure 3.3), ribbon charts (*e.g.*, figure 3.45). The rootogram, which we've seen several times (*e.g.*, figure 3.6) is merely a line and ribbon chart overlain onto a histogram. We've seen density plots, which appear as a line chart, except the line represents the density of values on the y-axis, given the value on the x-axis (*e.g.*, figure 3.5). We've even used a so-called parallel coordinates plot, (*e.g.*, figure 3.27), which uses a line chart encoding where the x-axis represents categories, instead of a continuous

quantity. We have even seen a more stylized form of a boxplot[3] (with dots of observed values overlain), which we referred to simply as an uncertainty interval (*e.g.*, figure 3.16).

These types of charts, and those similar, are common because they effectvely encode quantities measured from a common base line (the x or y axis, or both), which, as we have learned, is generally less error prone in decoding values, see section 7.4. Our particular implementations have also generally tried to maximize the data-ink ratio. Even their sizes have been reduced to just large enough to consider the patterns they reveal. In some cases, however, common encodings are not optimal. Instead, an encoding more unique will better inform our audiences. These, we may categorically call xenographics.

Of note, the difference between common graphics and what has been called xenographics is somewhat arbitrary. The more important point is not the name we use — chart names are just short-hand to convey instances of graphics — but that we anticipate what encodings our audience already understands how to decode and what encodings our audience needs explaination on how to decode.

[3] Boxplots were introduced by Tukey, and explained in John W Tukey, *Exploratory Data Analysis* (Addison-Wesley, 1977); Tukey's boxplot was dissected in Tufte, "Data-Ink Maximization and Graphical Design" as an exercise to maximize the data-ink using alternate encodings.

8.2 Graphics, layers and separation

When designing graphics, and especially when comparing encodings or annotating them, we must visually layer and separate types of information or encodings. As Tufte[4] explains, "visually stratifying various aspects of the data." By layering or stratifying, we mean placing one type of information over the top of a second type of information. The grammar of graphics, discussed earlier, accomplishes such a layering. To visually separate the layered information, we can assign, say, a hue or luminance, for a particular layer. Many of the graphics of chapter 3 separate types of data through layering. Figure 3.5, for example, includes two data types: observed and simulated data. Two hundred simulations of data are placed in the background by adding it to the graphic before the observed value, and by using a light color to contrast with the dark color of the observed data encoding.

These design choices may also create a sense of near and far.[5] We may create a sense of depth, of foreground and background, using any of size, overlapping the forms or encodings, the encodings relative values (lightness, opacity). "The seeming nearness or distance of each form will also contribute to the viewer's sense of its importance and, therefore, its meaning relative to other forms presented within the same space." Ultimately we are trying to achieve a visual hierarchy for the audience to understand at each level.

[4] Edward R. Tufte, "Layers and Separation," in *Envisioning Information* (Graphics Press, 1990).

[5] Timothy Samara, *Design Elements: A Graphic Style Manual*, Understanding the Rules and Knowing When to Break Them (Rockport, 2014).

8.3 Encoding data-ink, xenographics

For a growing collection of interesting approaches to visualizing data in uncommon ways, consult the website *Xenographcs*[6]. But we have already seen a few less common data encodings. Recall, for example, instances of tracking information encoded as dots within circles in figure 5.4. Let's consider a couple more. Getting back to our example Citi Bike project, we identified various data visuals used in earlier exploratory work.[7] In that earlier study, researchers visualized bike and docking station activity data in the form of heatmaps overlaying maps, and heatmaps as a grid wherein the x-axis encoded time of day, the y-axis encoded docking station names as categorical data, hue at a given time and docking station encoded the imbalance between incoming and outgoing bikes, and luminocity at the same location encoded activity level, as shown in figure 8.1.

[6] Maarten Lambrechts, "Xenographics: Weird but (Sometimes) Useful Charts," *Xenographics* (https://xeno.graphics, 2020).

[7] Saldarriaga, "CitiBike Rebalancing Study."

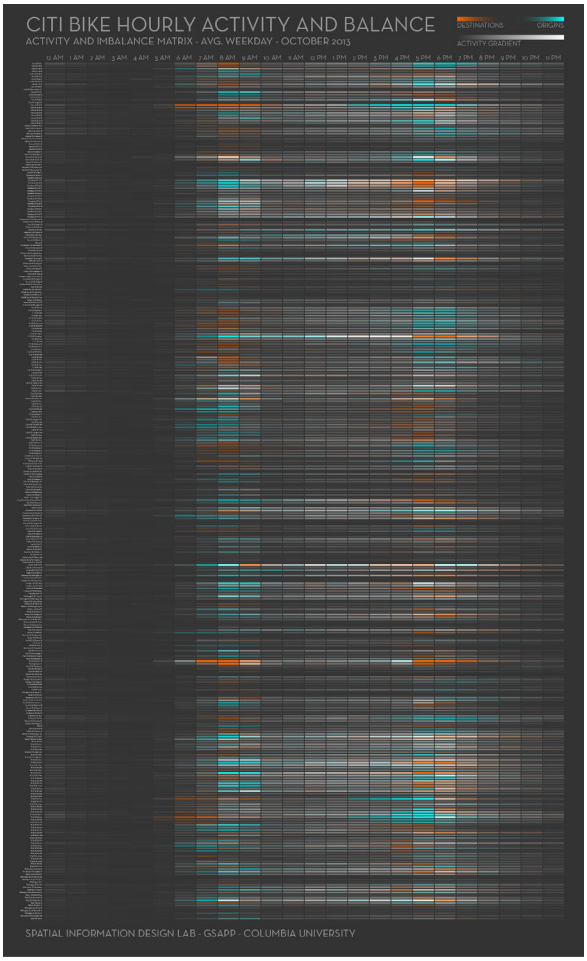


Figure 8.1: Researchers visualized bike and docking station activity data in the form of heatmaps overlaying maps, and heatmaps as a grid wherein the x-axis encoded time of day, the y-axis encoded docking station names as categorical data, hue at a given time and docking station encoded the imbalance between incoming and outgoing bikes, and luminocity at the same location encoded activity level.

The more interesting aspect of this graphic is that the dual hue, luminance encoding enables markings to disappear if either a) incoming and outgoing activity is balanced or b) the activity level is very low. The limitations of the overall encoding, however, include an unfamiliar listing of docking stations by name on the y-axis. As we proposed in the memo, example 4.1, let's try encoding these variables differently. Let's try addressing the admitted challenge of encoding geographic location with time in a way that allows further, meaningful encodings. We will do this in stages. First, we consider activity level, which we naturally think of as a daily pattern. Other graphics[8] have explored daily patterns of activity, and encode that activity level using polar coordinates. We borrow from that work, encoding bike activity level the way we think about time — circular, think of a 24-hour clock. Our first graphic is in figure 8.2. We read the graphic as reflecting activity level over time, which is encoded circular, with midnight at the top, 6am to the right, noon at the bottom, 18 hours (6pm) to the left. To help visualize time of day, we label sunrise and sunset, and shade areas before and after sunrise as dark and light.

[8] Zan Armstrong and Nadieh Bremer, "Why Are so Many Babies Born Around 8:00 A.M.?" *Scientific American*, June 2017; Nadieh Bremer, "The Baby Spike," Portfolio, *Visual Cinnamon* (https://www.visualcinnamon.com/portfolio/baby-spike, April 2017) explains how the graphic was made.

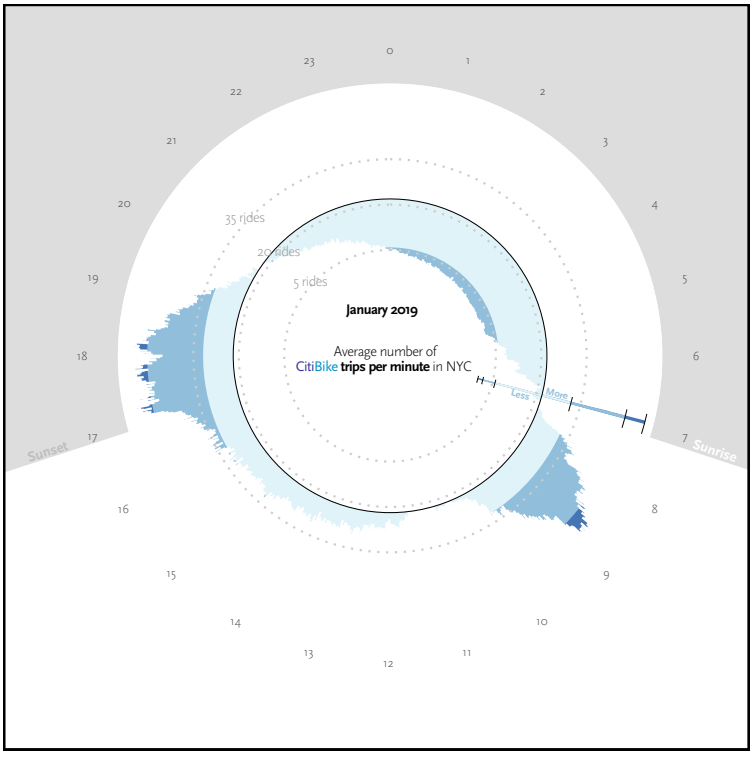


Figure 8.2: This graphic encodes bike activity levels throughout a 24-hour day, where time is encoded as polar coordinates.

As did Nadieh, we encode an average activity level along the **black line**, activity level at a given time as the distance from that average. And the color within that distance from average activity level encodes the quantiles (think boxplot) of activity. As with encod-

ing average activity level, we annotate with reference activity levels: 5, 20, and 35 rides per minute. What is remarkable is the observed magnitude of change from average (black circle) ride rates that exist throughout the day, which reflects this rebalancing problem. Minutes in only light blue show when 50 percent of the ride rates exist. Minutes that include dark blue show when the highest (outside black circle) or lowest (inside black circle) rate of rides happen. Finally, the remaining minutes with medium blue show when the rest of the rates of rides occur.

We now address the limitation of the prior work. In this regard, we can learn from the famous graphic by Minard of Napoleon's march, see figure 8.3.

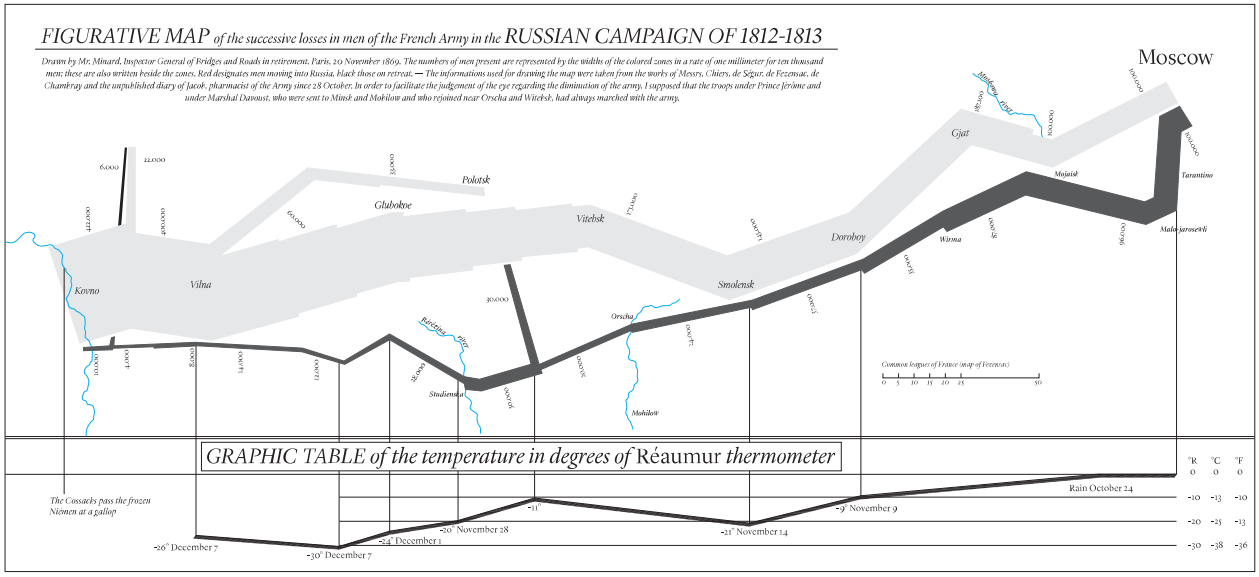


Figure 8.3: Minard's Napoleon graphic redrawn and translated into English.

In Minard's graphic, as Tufte explains, he overlays the path of Napoleon's march onto a map in the form of a ribbon.[9] While the middle of that ribbon may accurately reflect geographic location, the width of that ribbon does not. Instead, the width of the ribbon encodes the number of solders at that location, wherein time is also encoded as coinciding with longitude. That encoding gives a sense of where the solders were at a given time, while also encoding number of solders. We try a similar approach with Citi Bike, shown in figure 8.4. We place each docking station the a black **dot** (·) overlaying a geographic map of New York City. At each station, we encode using color an **empty** or **full** station as a line segment (|) starting at the station **dot** and extending towards time of day, the length of a unit circle. The line segments are partly transparent so that an indi-

[9] Tufte, *The Visual Display of Quantitative Information* analyses Minard's graphic, declaring it, perhaps, the greatest ever created.

vidual empty or full station won't stand out, but repeated problems at that time of day over the three weeks of the data (January 2019) would be more vivid and noticeable. Finally, we annotate the graphic with a narrative and a key that explains these encodings, along with encoding the general activity levels of the graphic in figure 8.2.

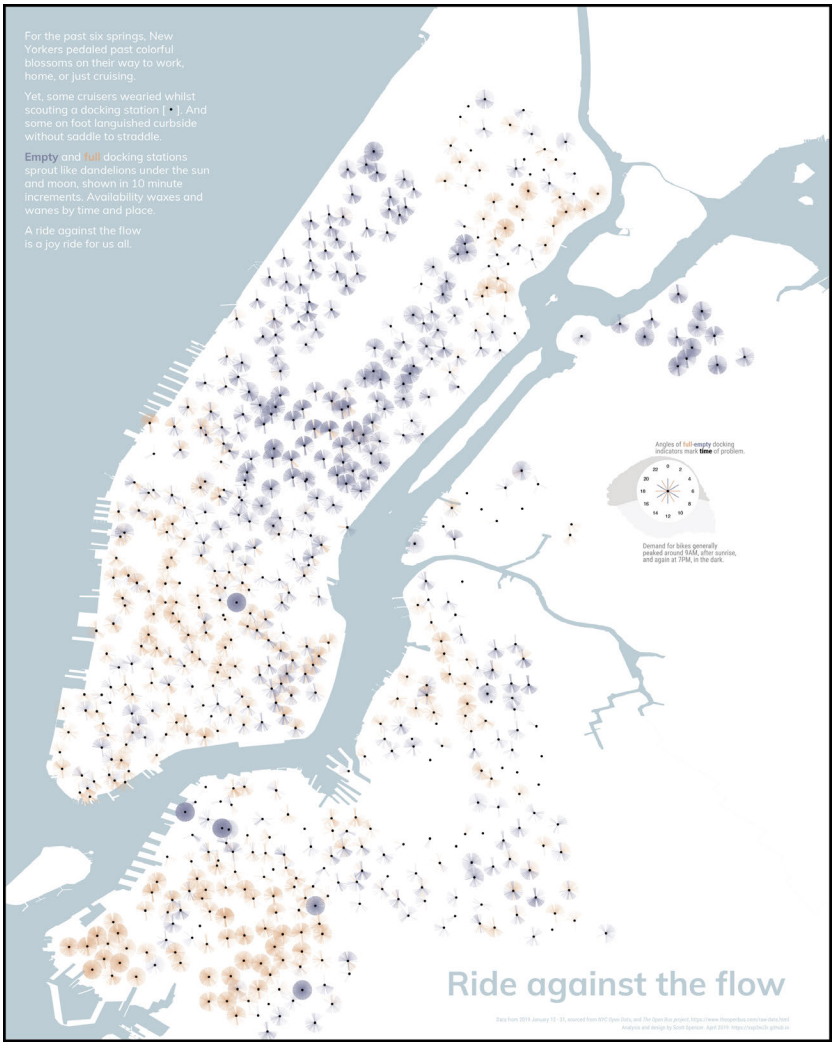


Figure 8.4: The visualization — a *xenographic* — invites riders to explore bike and docking station availability for encouraging re-distribution for the NYC bike share. The data on trips and station availability are encoded in seven dimensions: space, time, bike and dock availability, rate of new rides per minute, and whether unavailability at a given time of day occurred multiple times. I used the metaphor of unavailability as dandelions among flowers that riders travel through each spring, weeds that need fixing and a request: by riding against the flow—redistributing bikes—those riders are helping us all.

The infographic adds, as its title, a call to action: *Ride Against the Flow*[10]. When encoding custom graphics, basic math can come in handy. The encodings (colored line segments) for empty and full docking stations at each station were created by mapping the hour of a day to the angle in degrees/radians of a unit circle, and calculating the end of the line segments as an offset from the docking station geolocation using basic trigonometry,

[10] Scott Spencer, "Ride Against the Flow," 2019, longlisted Kantar Information is Beautiful Awards.