Storytelling with data

03 | re-writing; (more) communication concepts; visual organization

Scott Spencer | Columbia University

course overview | main course deliverables

group re-writing exercise

group re-writing | revise analytics communication for new audience and purpose

**Improving traffic safety
through video analysis in Jakarta**

"We want this project to provide a template for others who hope to successfully deploy machine learning and data driven systems in the developing world. . . . These lessons should be invaluable to the many researchers and data scientists who wish to partner with NGOs, governments, and other entities that are working to use machine learning in the developing world."

In what ways are this audience and purpose similar to, and different from, the intended audience and purpose for the example memos?

Improving Traffic Safety Through Video Analysis in Jakarta, Indonesia

João Caldeira*
Department of Physics
University of Chicago
jcaldeira@uchicago.edu

Alex Fout*
Statistics
Colorado State University
alex.fout@colostate.edu

Aniket Kesari*
Jurisprudence & Social Policy
University of California, Berkeley
akesari@berkeley.edu

Raesetje Sefala*
Machine Learning
University of the Witwatersrand
raesetje.sefala@students.wits.ac.za

Joseph Walsh
Center for Data Science and Public Policy
University of Chicago

Katy Dupre
Center for Data Science and Public Policy
University of Chicago

Muhammad Rizal Khaefi
Pulse Lab Jakarta
muhammad.khaefi@un.or.id

Setiaji
Jakarta Smart City
setiaji@jakarta.go.id

George Hodge
Pulse Lab Jakarta
george.hodge@un.or.id

Zakiya Aryana Pramestri
Pulse Lab Jakarta

Muhammad Adib Imtiyazi
Jakarta Smart City

Abstract

This project presents the results of a partnership with Jakarta Smart City (JSC) and United Nations Global Pulse Jakarta (PLJ) to create a video analysis pipeline for the purpose of improving traffic safety in Jakarta. The pipeline transforms raw traffic video footage into databases. By analyzing these patterns, the city of Jakarta will better understand how human behavior and built infrastructure contribute to traffic challenges and safety risks. The results of this work should also be broadly applicable to smart city initiatives around the globe as they improve urban planning and sustainability.

1 Introduction

The World Health Organization's *Global status report on road safety 2015* estimates that over 1.2 million people die each year in traffic accidents [1]. Nearly 2000 such fatalities occur annually in the city of Jakarta, Indonesia. Many of these deaths are preventable through effective city planning. Jakarta has experienced rapid population growth over the last 50 years, from roughly two million people in 1970 to more than 10 million today. With this growth comes a rise in vehicle ownership and congestion, leading to an increase in the number of traffic incidents.

Juan Kanggrawan
Head of Data Analytics
Jakarta Smart City

"**Juan Intan Kanggrawan** is the current Head of Data & Analytics at Jakarta Smart City. His key role is to fully utilize data to formulate public policy and to improve quality of public services.

His main and foremost success metric is Jakarta citizen's satisfaction towards government. Juan is currently working on several city-scale strategic analytics initiatives.

He is actively analyzing complex, diverse and exciting urban data in daily basis: citizen complain/aspiration, transportation data from various sources, CCTV, global-regional-national Open Data, weather-flood-river bank, subsidy utilization for education & elderly, food commodities price elasticity, etc.

He is also developing and aligning strategic partnership framework between Jakarta Smart City with other government agencies, business enterprises, research agencies and universities"

Motivation

Message

Details

Appendix

?

Improving Traffic Safety Through Video Analysis: Pulse Lab Jakarta.

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.

(more) communication concepts

concepts, **start the communication on common ground — *from the mindset of your audience***

When you provide someone with new data, they

> quickly accept evidence that confirms their preconceived notions (what are known as prior beliefs) and

> assess counter evidence with a critical eye.

Focusing on what you and your audience have in common, rather than what you disagree about, enables change.

— Tali Sharot, *The Influential Mind*

Starting with common ground?

To **Michael Frumin** 2019 February 2
Director of Product and Data Science
for Transit, Bikes, and Scooters at Lyft

To inform the public on rebalancing, let's re-explore docking availability and bike usage with subway and weather

Let's re-explore station and ride data in the context of subway and weather information to gain insight for "rebalancing," broadening the factors our Simmons told the public: "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" (Friedman 2017).

Recalling the previous, public study by Columbia University Center for Spatial Research (Saldarriaga 2013), it identified trends in bike usage using heatmaps. As those visualizations did not combine dimensions of space and time, which the public would find helpful to see trends in bike and station availability by neighborhood throughout a day, we can begin our analysis there.

We'll use published data from NYC OpenData and The Open Bus Project, including date, time, station ID, and ride instances for all our docking stations and bikes since we began service. To begin, 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 current data.

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 and shows the public that we are, in Simmons's words, "innovative in how we meet this challenge." Let's draft a detailed proposal.

Sincerely,
Scott Spencer

Friedman, Matthew. "*Citi Bike Racks Continue to Go Empty Just When Upper West Siders Need Them*." News. West Side Rag (blog), August 19, 2017. https://www.westsiderag.com/2017/08/19/citi-bike-racks-continue-to-go-empty-just-when-upper-west-siders-need-them.

Saldarriaga, Juan Francisco. "*CitiBike Rebalancing Study*." Spatial Information Design Lab, Columbia University, 2013. https://c4sr.columbia.edu/projects/citibike-rebalancing-study.

Starting with common ground?

To **Scott Powers**
Director, Quantitative Analytics

2019 February 2

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.

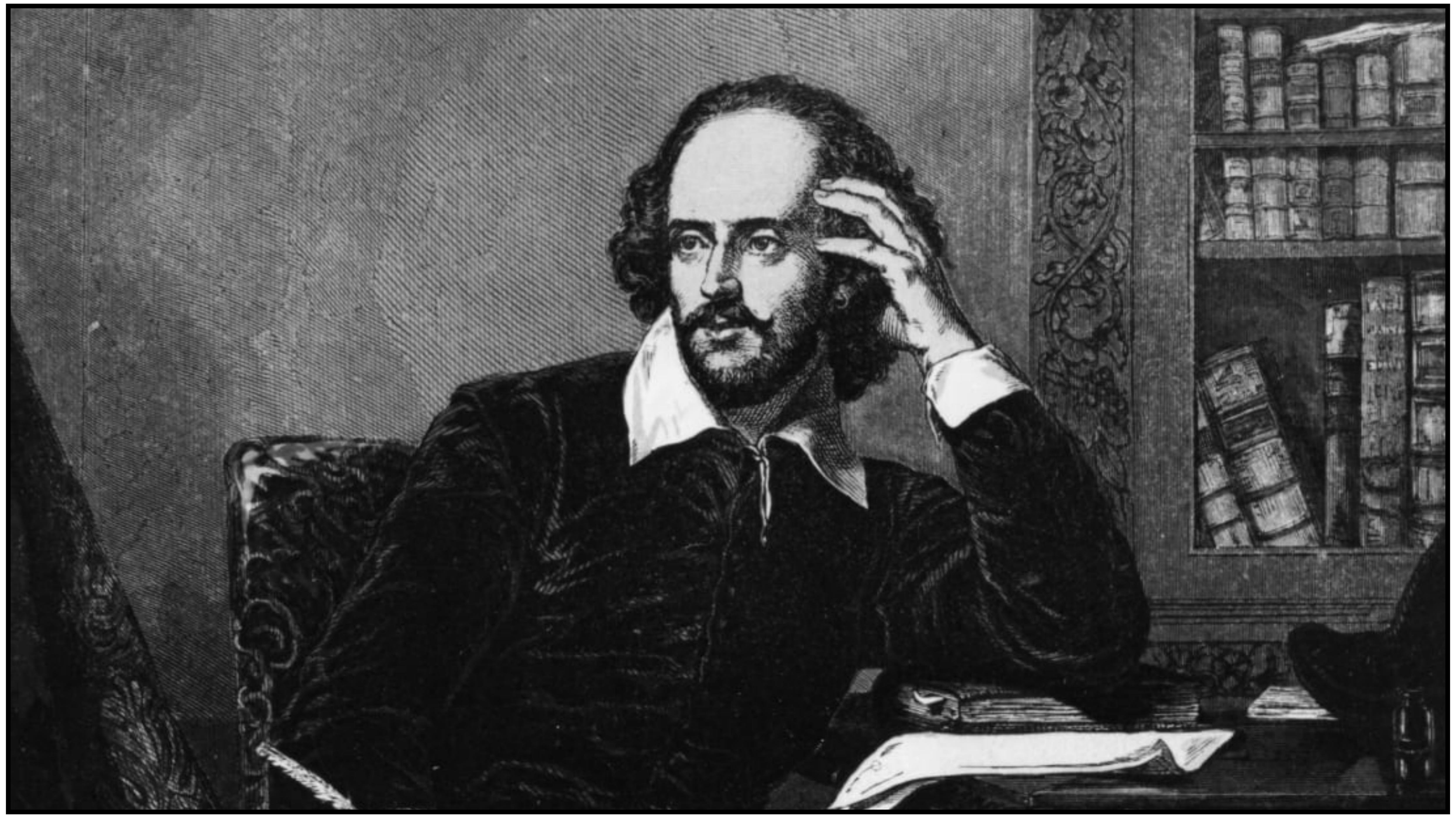
After defining our objective (*e.g.*, optimize expected runs) we will, from Statcast data, weight everything that could happen by its probability and accumulate these probability distributions. Joint distributions of all events, an eventual goal, will allow us to ask counterfactuals — "what if we do *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.

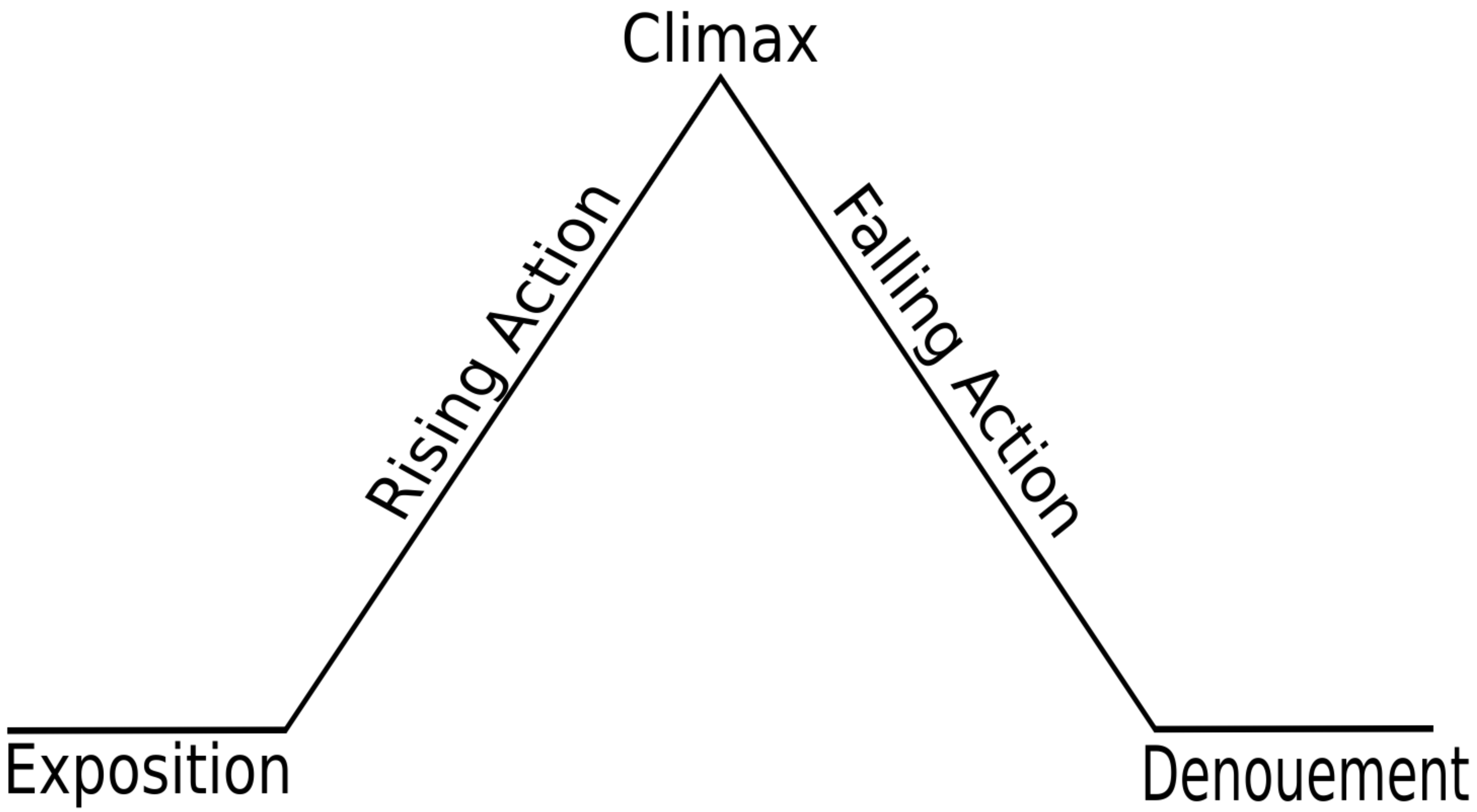
Computational issues are being addressed in Stan, a tool that enables inferences through advanced simulations. This open-source software is free but teaching its applications will require time. To shorten our learning curve, we can start with Stan interfaces that use familiar syntax (like lme4) but return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

Sincerely,
Scott Spencer

communication structure, **story or narrative**

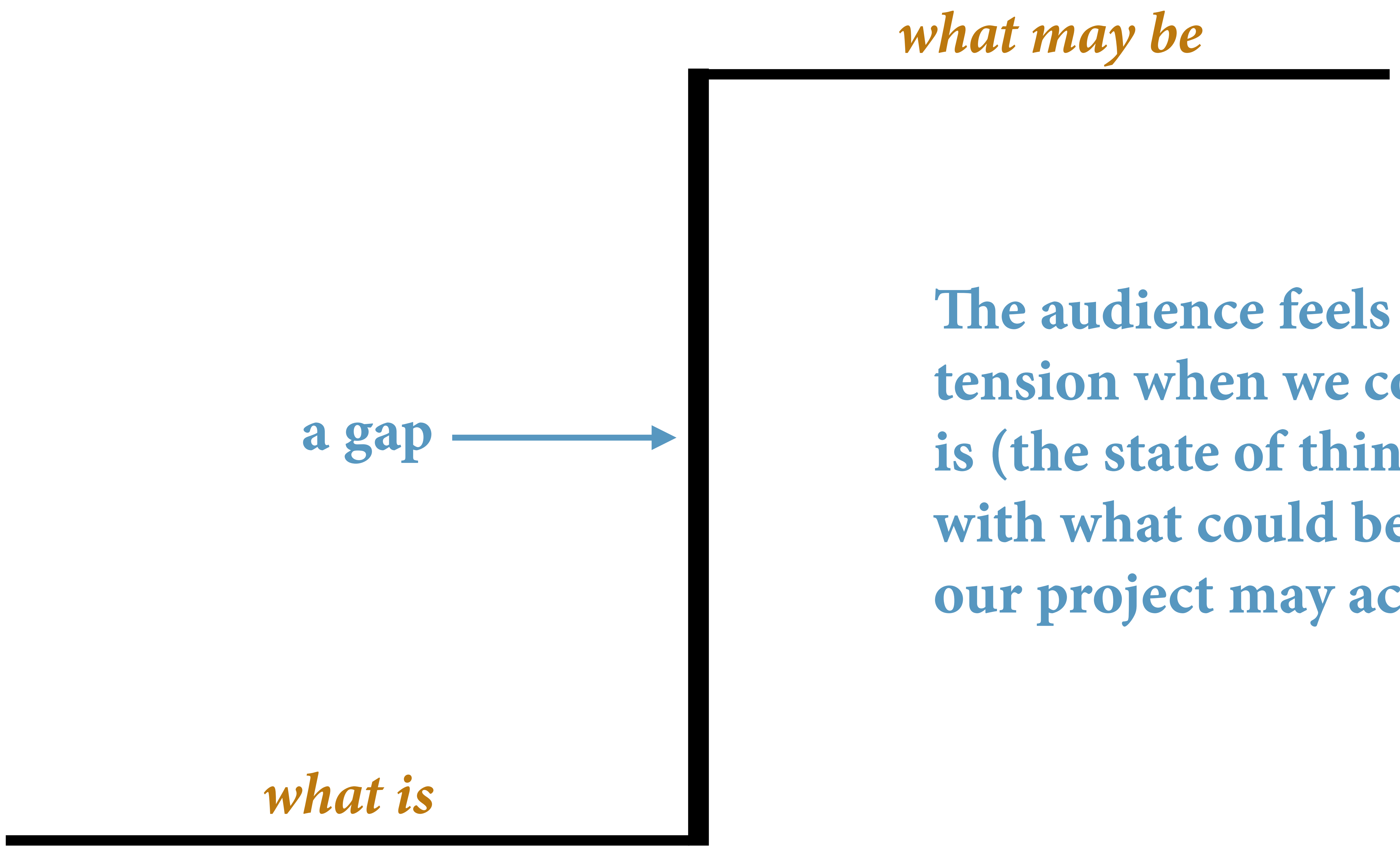


communication structure, story — from Shakespeare to data science?!



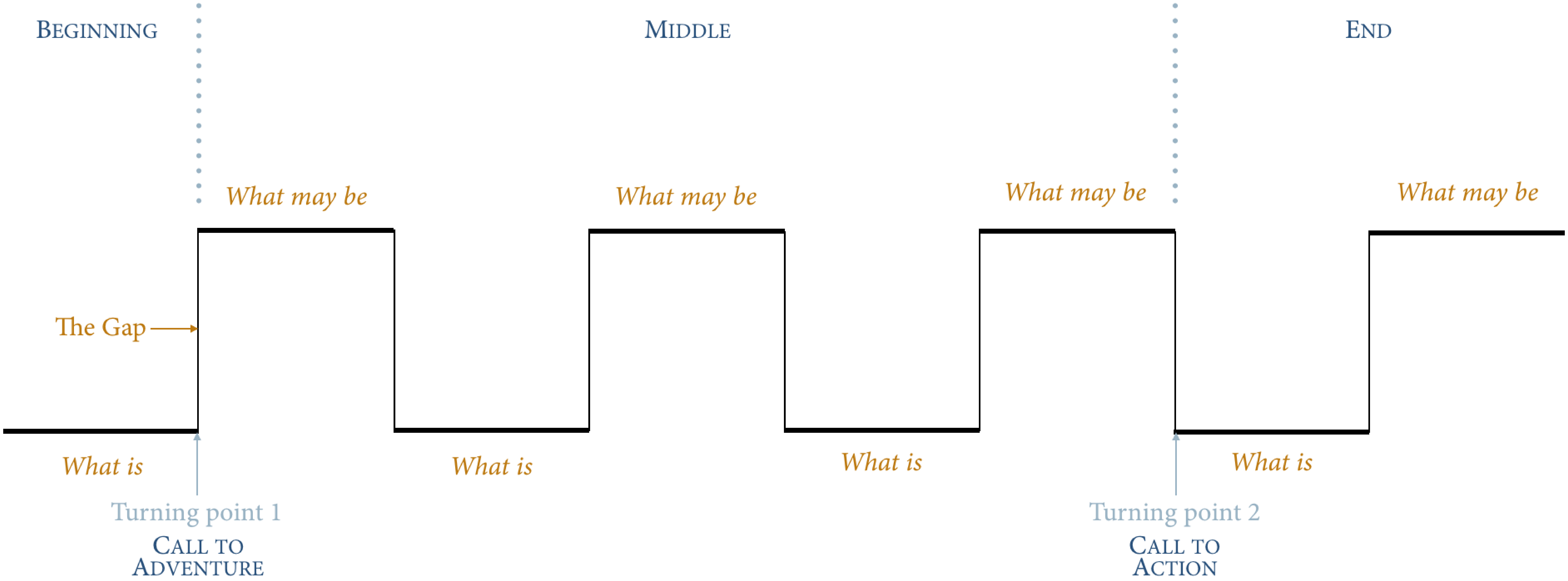
unexpected change

opening of an information gap



The audience feels dramatic tension when we contrast what is (the state of things now) with what could be (*e.g.*, what our project may achieve).

communication structure, **keeping audience interest throughout a communication**



the *lead* and the *ending*

sentence structure

sentence structure, **old before new**

old **new**

Scott Spencer / https://ssp3nc3r.github.io scott.spencer@columbia.edu

Booth, section 17.3, example 10(a)

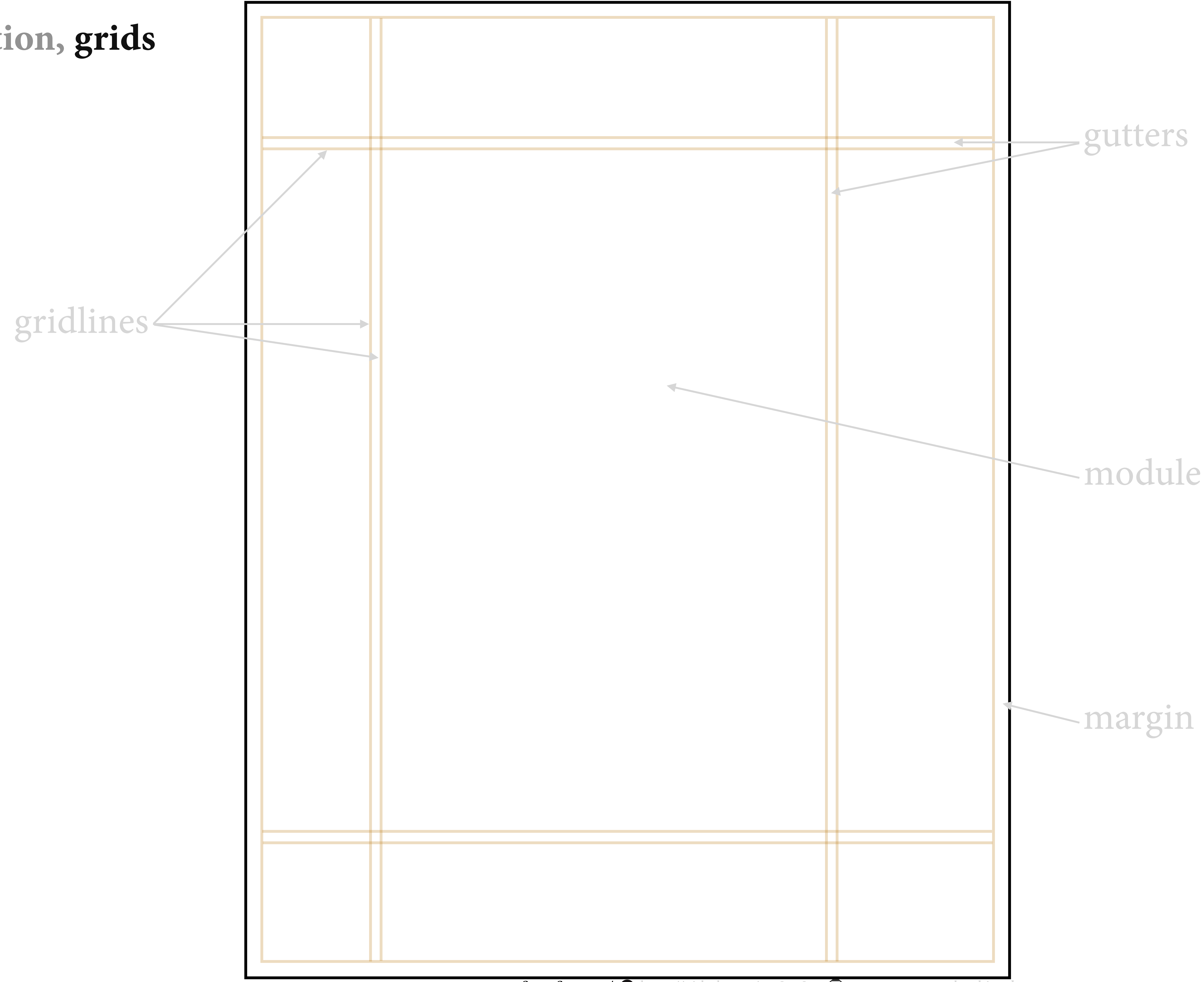
Because the naming power of words was distrusted by Locke, he repeated himself often. Seventeenth-century theories of language, especially Wilkins's scheme for a universal language involving the creation of countless symbols for countless meanings, had centered on this naming power. A new era in the study of language that focused on the ambiguous relationship between sense and reference begins with Locke's distrust.

example 10(b)

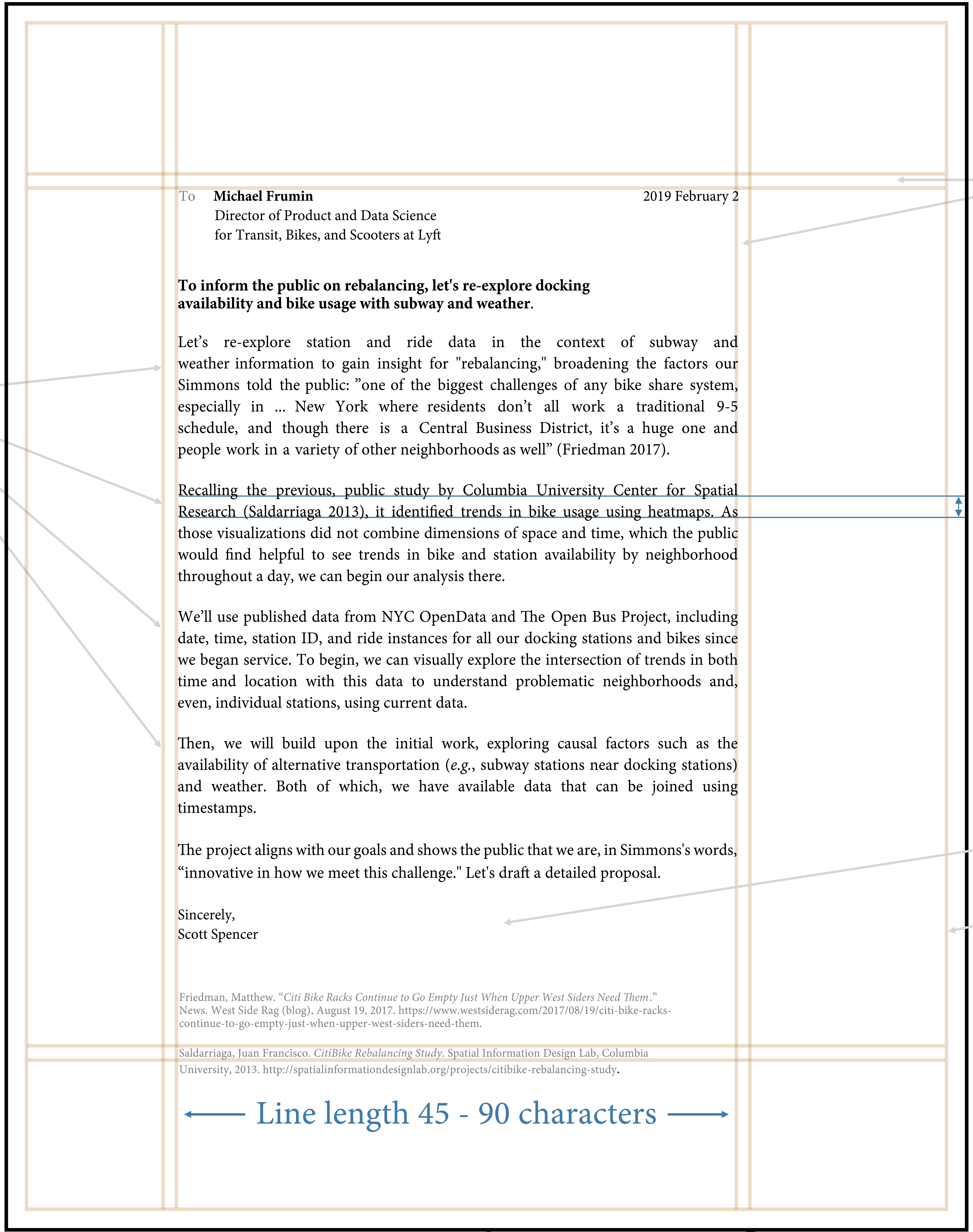
Locke often repeated himself because he distrusted the naming power of words. This naming power had been central to seventeenth-century theories of language, especially Wilkins's scheme for a universal language involving the creation of countless symbols for countless meanings. Locke's distrust begins a new era in the study of language, one that focused on the ambiguous relationship between sense and reference.

visual organization

visual organization, **grids**



visual organization, **grids**


To **Michael Frumin** 2019 February 2
Director of Product and Data Science
for Transit, Bikes, and Scooters at Lyft

To inform the public on rebalancing, let's re-explore docking availability and bike usage with subway and weather.

Let's re-explore station and ride data in the context of subway and weather information to gain insight for "rebalancing," broadening the factors our Simmons told the public: "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" (Friedman 2017).

Recalling the previous, public study by Columbia University Center for Spatial Research (Saldarriaga 2013), it identified trends in bike usage using heatmaps. As those visualizations did not combine dimensions of space and time, which the public would find helpful to see trends in bike and station availability by neighborhood throughout a day, we can begin our analysis there.

We'll use published data from NYC OpenData and The Open Bus Project, including date, time, station ID, and ride instances for all our docking stations and bikes since we began service. To begin, 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 current data.

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 and shows the public that we are, in Simmons's words, "innovative in how we meet this challenge." Let's draft a detailed proposal.

Sincerely,
Scott Spencer

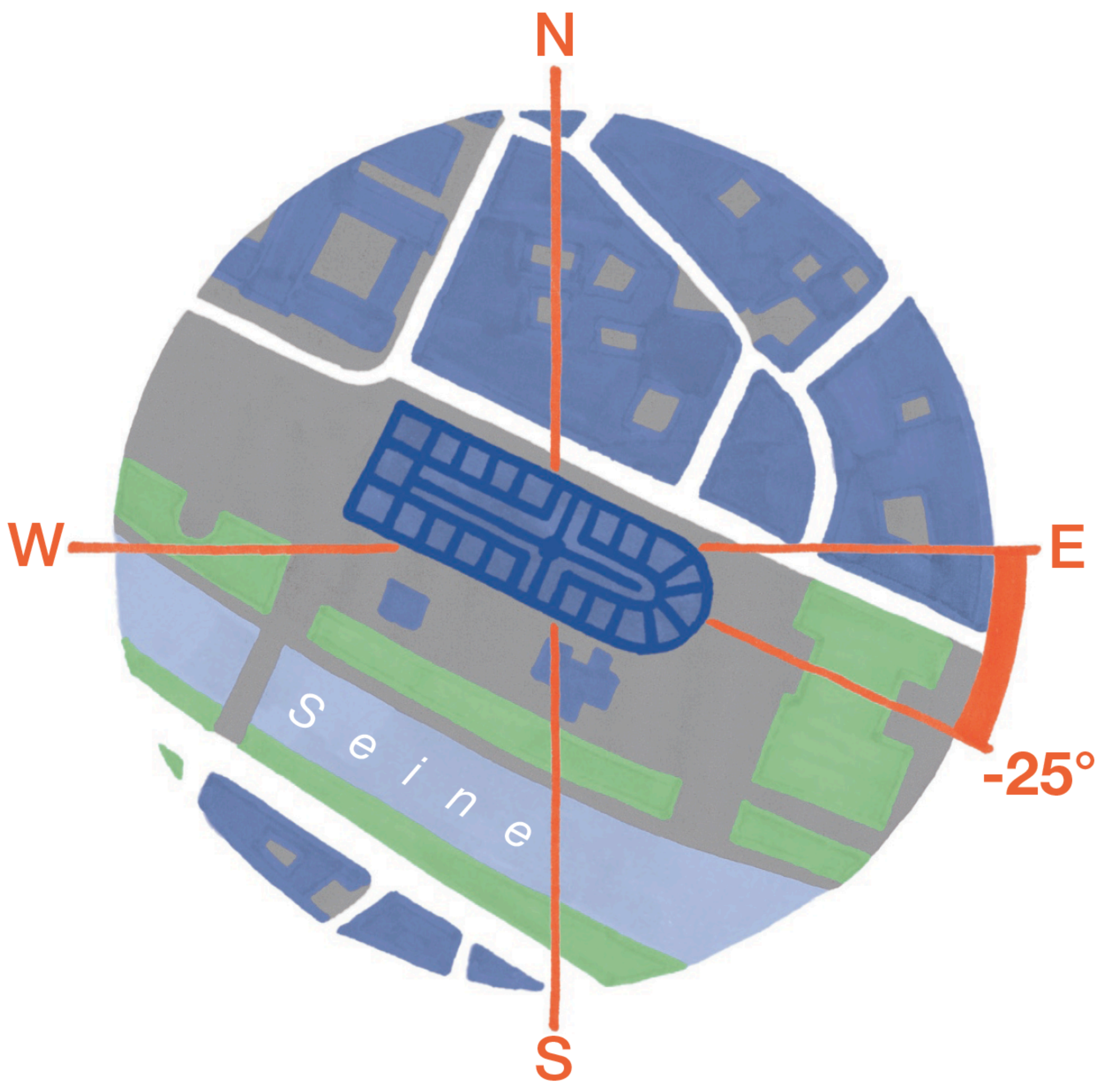
Friedman, Matthew. "*Citi Bike Racks Continue to Go Empty Just When Upper West Siders Need Them*." News. West Side Rag (blog), August 19, 2017. https://www.westsiderag.com/2017/08/19/citi-bike-racks-continue-to-go-empty-just-when-upper-west-siders-need-them.

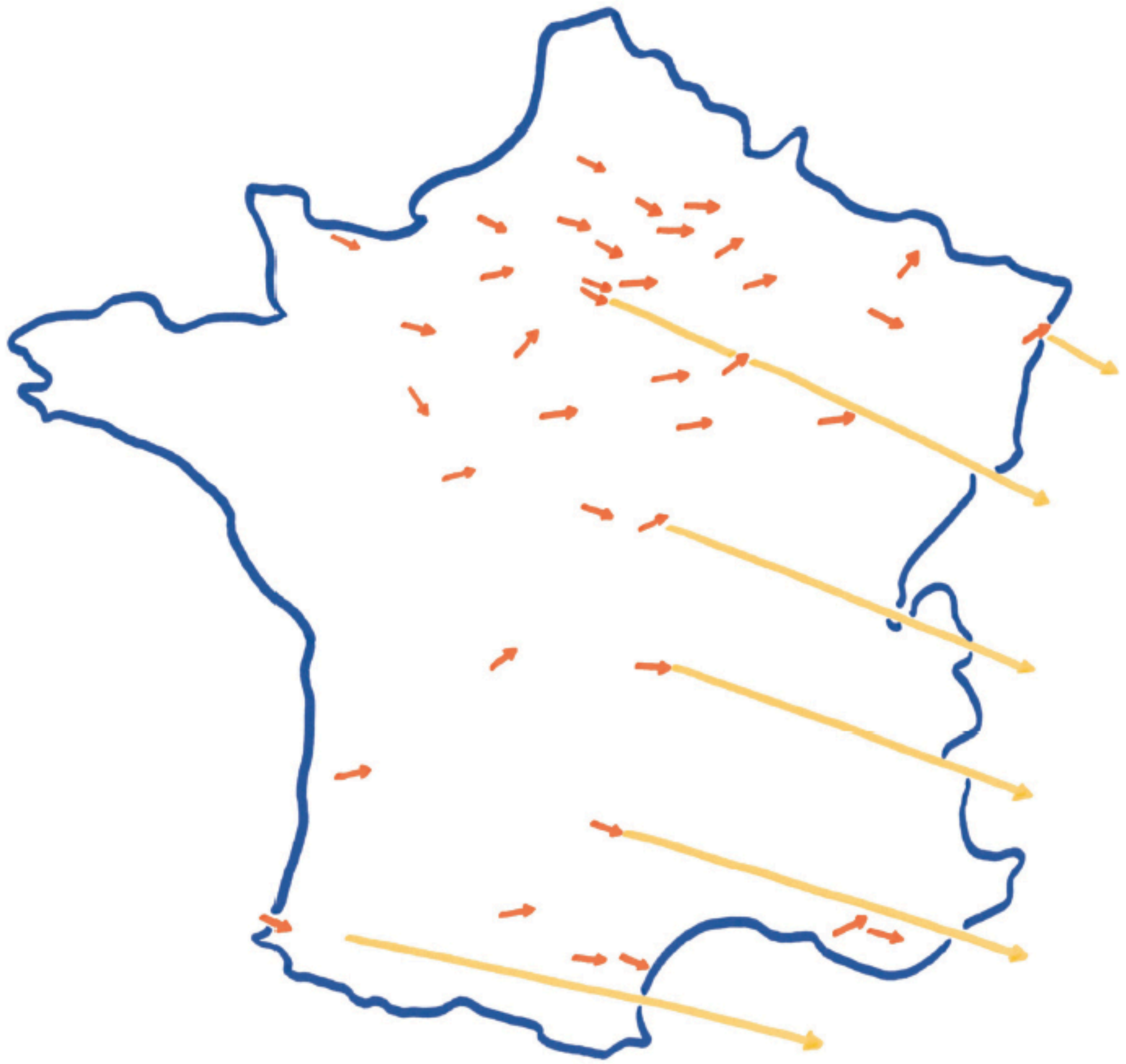
Saldarriaga, Juan Francisco. *CitiBike Rebalancing Study*. Spatial Information Design Lab, Columbia University, 2013. http://spatialinformationdesignlab.org/projects/citibike-rebalancing-study.

> 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.
>
> — Matthew Butterick, *Practical Typography*

for inspiration — another example of *unstructured* data *identified* and *collected* for an analytics project

for inspiration, **data identification and collection**





resources

References

Spencer, Scott."Elements of Writing." In *Data in Wonderland*. 2021. https://ssp3nc3r.github.io/data_in_wonderland.

Andrews, R J. "Focus Attention." In *Info We Trust: How to Inspire the World with Data*. Wiley, 2019.

———. "Imagination to Image." In *Info We Trust: How to Inspire the World with Data*. Wiley, 2019.

———. "Stories of Space, Time, and Data." *YouTube*, April 10, 2020. https://youtu.be/rjI6IfSQZyw.

Booth, Wayne C, Gregory G Columb, Joseph M Williams, Joseph Bizup, and William T Fitzgerald. *The Craft of Research*. Fourth. University of Chicago Press, 2016.

Butterick, Matthew. "*Butterick's Practical Typography*," 2018. https://practicaltypography.com/.

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

Sharot, Tali. *The Influential Mind. What the Brain Reveals about Our Power to Change Others*. Henry Holt and Company, 2017.

Storr, Will. *Science of Storytelling*. New York, NY: Abrams Books, 2020.

Zinsser, William. *On Writing Well*. Sixth. *The Classic Guide to Writing Nonfiction*. Harper Resource, 2001.