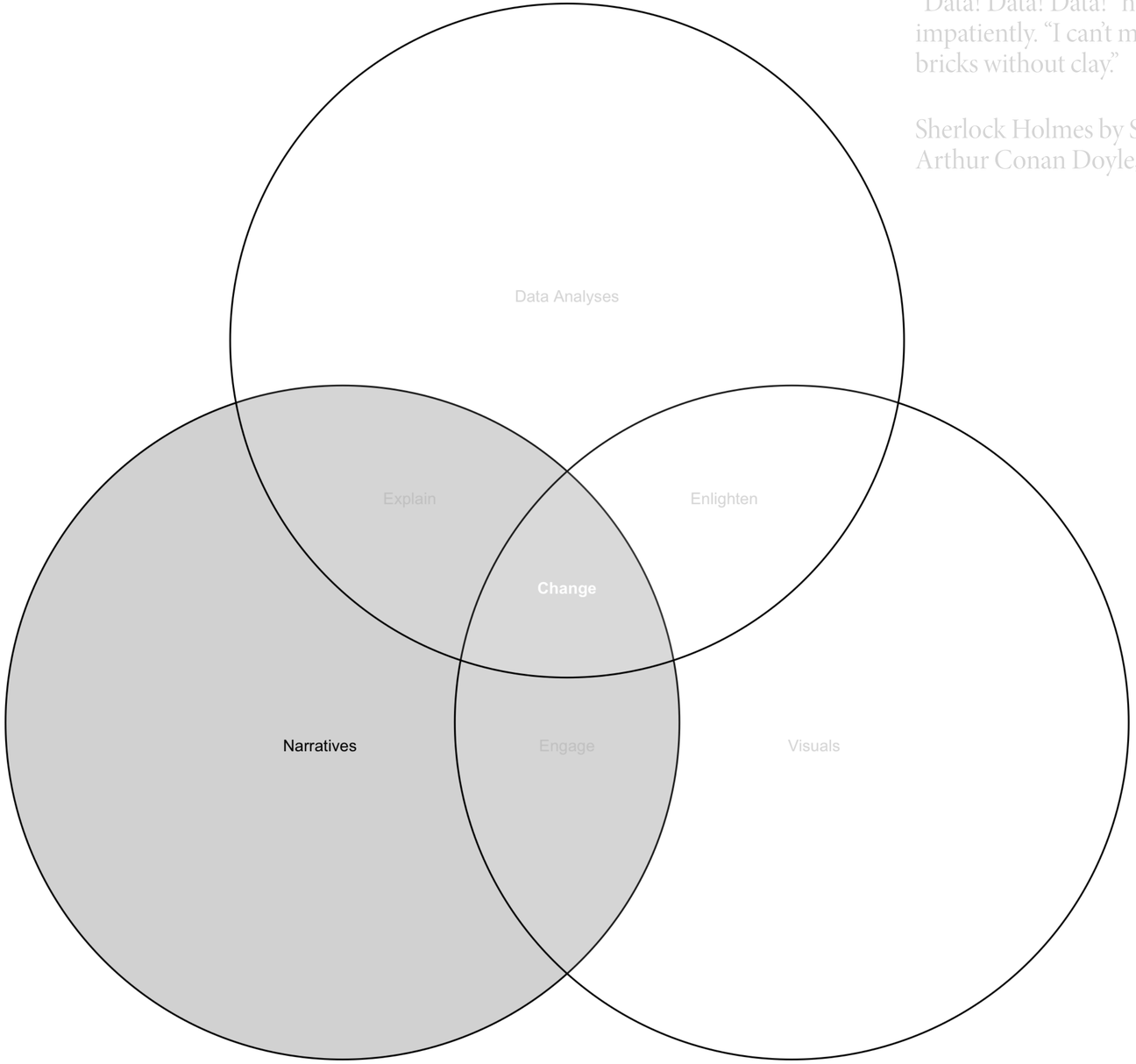
Storytelling with data

05 | review homework two; (re)design for an audience, *continued*; elements of writing

course overview, learn to drive change using data visuals and narrative

"Data! Data! Data!" he cried impatiently. "I can't make bricks without clay."

Sherlock Holmes by Sir Arthur Conan Doyle, *author*



No one ever made a decision because of a number. They need a story.

Daniel Kahneman, *psychologist, behavioral economist, and author*

The greatest value of a picture is when it forces us to notice what we never expected to see.

John W Tukey, *mathematician*

general course deliverable timeline

Individual Work

For learning data visualization
and written narrative techniques

Group work

For building graphics and narrative
into interactive communications

Sept 30	Oct 14	Oct 28	Nov 18	Nov 18	Dec 11	Dec 13
Homework 1 graphics 10%	Homework 2 graphics 10%	Homework 3 writing 10%	Homework 4 graphics 10%	Proposal 15%	Interactive Communication 20%	Multimodal communication 15%
Participation 10%						

homework two review | graphics practice with Citi Bike rebalancing study

homework two review, **questions?**

homework 2.rmd

Knit on Save | Knit | Run

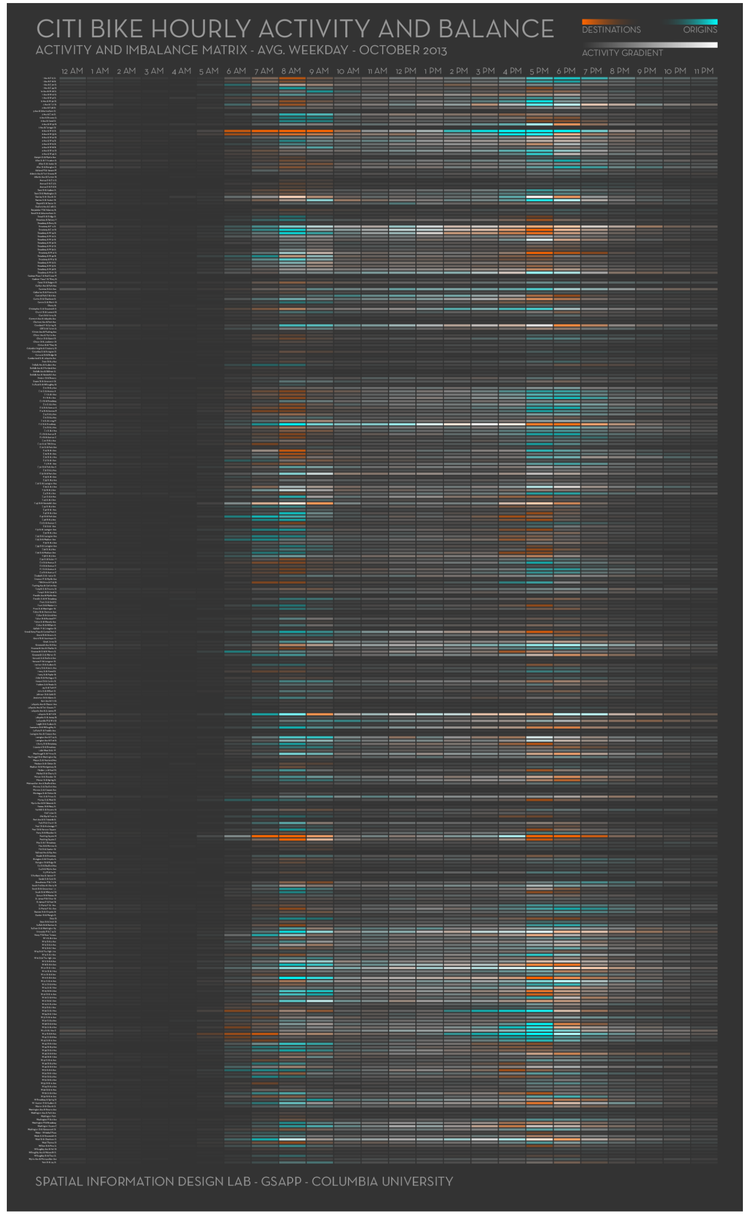
```
1  ---
2  title: 'Homework 2: graphics practice'
3  author: 'Last name, first name'
4  date: ``r format(Sys.Date(), "%Y, %B %d")`'
5  output: distill::distill_article
6  ---
7
8  ```{r setup, include=FALSE}
9  knitr::opts_chunk$set(
10   eval = FALSE,
11   echo = TRUE,
12   message = FALSE,
13   error = FALSE,
14   warning = FALSE)
15 ```
16
17
18
19 # Preliminary
20
21
22
23 For this homework assignment, we'll continue exploring data related to our Citi
   Bike case study as a way to practice the concepts we've been discussing in
   class.
24
25 In our third discussion, we briefly considered an exploratory visualization of
   activity and docking station (im)balance, conducted in 2013 by Columbia
   University's Center for Spatial Research.
   [https://c4sr.columbia.edu/projects/citibike-rebalancing-study](https://c4sr.co
   lumbia.edu/projects/citibike-rebalancing-study).
26
27 As practice in understanding encodings, let's review and reconstruct one of the
   Center's graphics, titled: "CITI BIKE HOURLY ACTIVITY AND BALANCE". You can
   download and zoom in on a high resolution pdf of the graphic here:
   [https://c4sr.columbia.edu/sites/default/files/Activity_Matrix_Composite.pdf](h
   ttps://c4sr.columbia.edu/sites/default/files/Activity_Matrix_Composite.pdf).
28
29
30
31 # Question 1(a) and 1(b) — data types and visual encodings
32
33
34
35 What variables and data types have been encoded?
36
37
38
```

Preliminary
Question 1(a) and 1(b) — data types and visual encodings
Question 2 — coordinate systems
Question 3 — comparing encoded data
Question 4 — workflow, tidying and transforming data
Question 5 — transforming data
Question 6 — scaling data
Question 7 — mapping data to visual channels
Question 8 — decoding and interpretation: critical thinking
Bonus — advanced practice
Knit and submit

289:264 | Knit and submit | R Markdown

Console

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



review student research — encoding data with hue, saturation, luminance

(re)design for your audience, *continued*

redesigns, example — original graphic within government publication explaining part of US economy

Bureau of Economic Analysis
U.S. DEPARTMENT OF COMMERCE

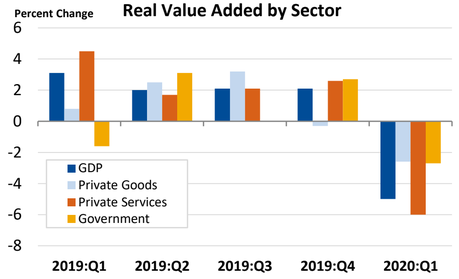
Monday, July 6, 2020
Contact: Jeannine Aversa, (301) 278-9003

Gross Domestic Product by Industry: First Quarter 2020

Accommodation and food services; finance and insurance; and health care and social assistance industries were the leading contributors to the 5.0 percent (annual rate) decrease in gross domestic product (GDP) in the first quarter of 2020.

All sectors of the U.S. economy contributed to the decrease, led by a decline in private services-producing industries. The decline in first-quarter GDP reflected the response to the spread of COVID-19, as governments issued "stay-at-home" orders in March. This led to rapid changes in production, as businesses and schools switched to remote work or canceled operations, and consumers and businesses canceled, restricted, or redirected their spending. For more information, see "Federal Recovery Programs and BEA Statistics: COVID-19 and Recovery" on the BEA website.

Real GDP and Real Value Added by Sector

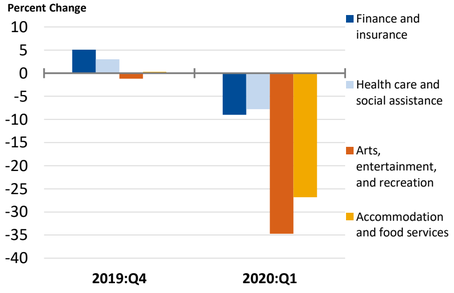


U.S. Bureau of Economic Analysis Seasonally adjusted annual rates

Overall, 17 of 22 industry groups contributed to the first-quarter decline in real GDP. Of the five industry groups that offset the decline in the first-quarter real GDP, agriculture, forestry, fishing, and hunting was the largest contributor, increasing 15.5 percent.

For accommodation and food services, real value added—a measure of an industry's contribution to GDP—decreased 26.8 percent, primarily reflecting a decrease in food services and drinking places.

Real Value Added by Selected Industries



U.S. Bureau of Economic Analysis Seasonally adjusted annual rates

Finance and insurance decreased 9.0 percent, primarily due to a decrease in insurance carriers and related activities.

Health care and social assistance decreased 7.8 percent, primarily reflecting decreases in ambulatory health care services and in hospitals.

Arts, entertainment and recreation decreased 34.7 percent, primarily reflecting a decrease in performing arts, spectator sports, museums, and related activities.

BEA statistics—including GDP, personal income, the balance of payments, foreign direct investment, the input-output accounts, and economic data for states, local areas, and industries—are available at www.bea.gov. E-mail alerts are also available.

Bureau of Economic Analysis. *Gross Domestic Product by Industry: First Quarter 2020.* https://www.bea.gov/sites/default/files/2020-07/gdpind120-fax.pdf.

redesigns, example — what's the point of this graphic? Do encodings intuitively show the point? Let's redesign!

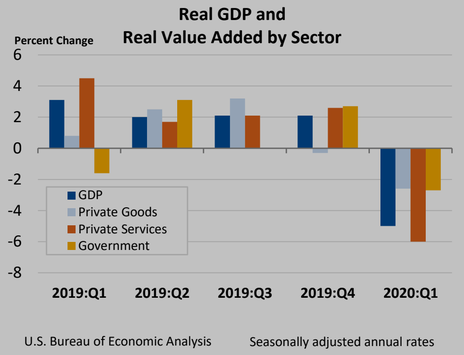
bea
Bureau of Economic Analysis
U.S. DEPARTMENT OF COMMERCE

Monday, July 6, 2020
Contact: Jeannine Aversa, (301) 278-9003

Gross Domestic Product by Industry: First Quarter 2020

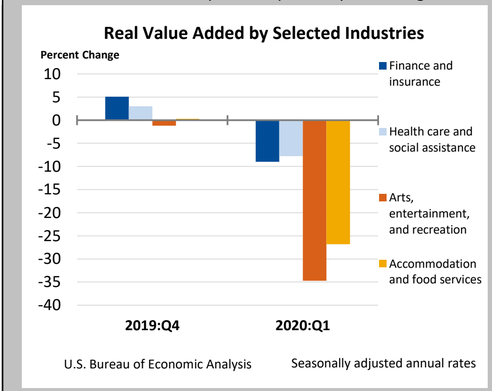
Accommodation and food services; finance and insurance; and health care and social assistance industries were the leading contributors to the 5.0 percent (annual rate) decrease in gross domestic product (GDP) in the first quarter of 2020.

All sectors of the U.S. economy contributed to the decrease, led by a decline in private services-producing industries. The decline in first-quarter GDP reflected the response to the spread of COVID-19, as governments issued "stay-at-home" orders in March. This led to rapid changes in production, as businesses and schools switched to remote work or canceled operations, and consumers and businesses canceled, restricted, or redirected their spending. For more information, see "Federal Recovery Programs and BEA Statistics: COVID-19 and Recovery" on the BEA website.



Overall, 17 of 22 industry groups contributed to the first-quarter decline in real GDP. Of the five industry groups that offset the decline in the first-quarter real GDP, agriculture, forestry, fishing, and hunting was the largest contributor, increasing 15.5 percent.

For accommodation and food services, real value added—a measure of an industry's contribution to GDP—decreased 26.8 percent, primarily reflecting a decrease in food services and drinking places.

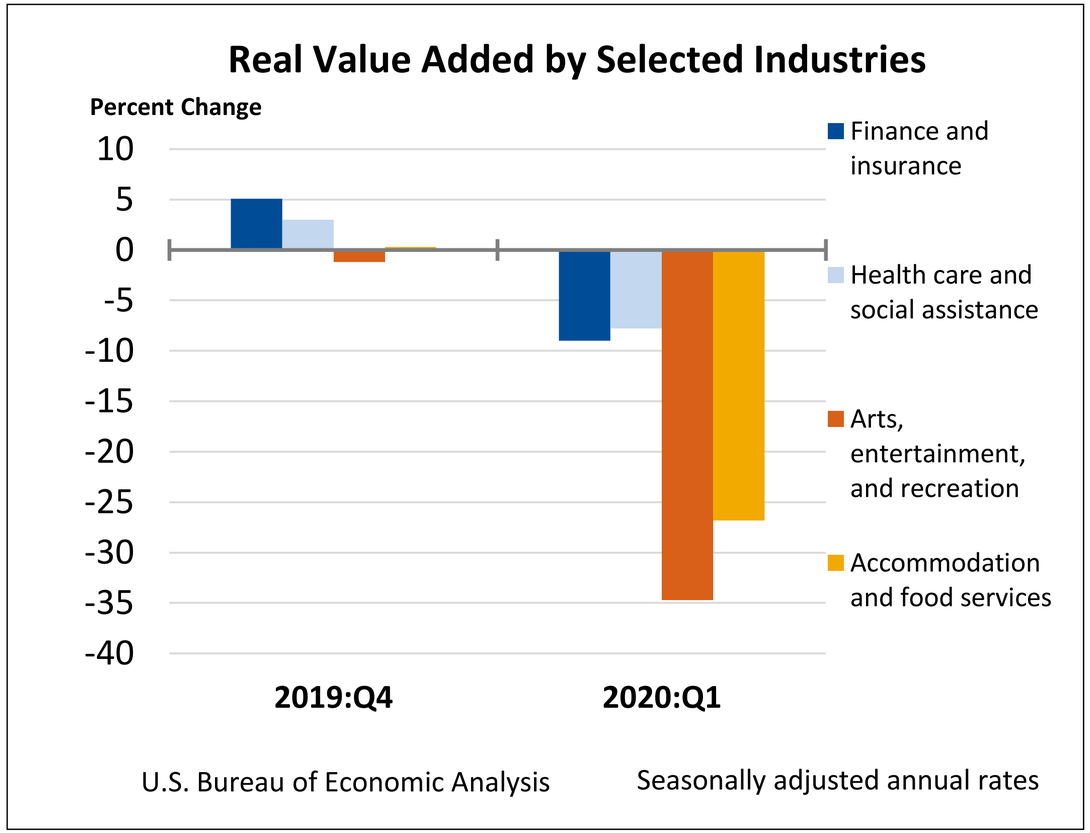


Finance and insurance decreased 9.0 percent, primarily due to a decrease in insurance carriers and related activities.

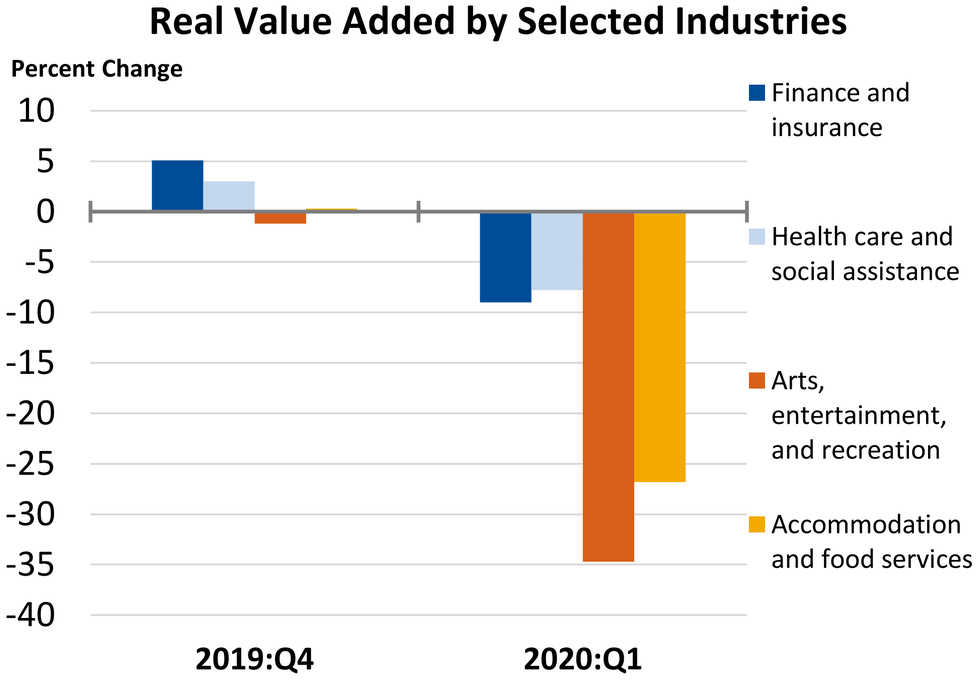
Health care and social assistance decreased 7.8 percent, primarily reflecting decreases in ambulatory health care services and in hospitals.

Arts, entertainment and recreation decreased 34.7 percent, primarily reflecting a decrease in performing arts, spectator sports, museums, and related activities.

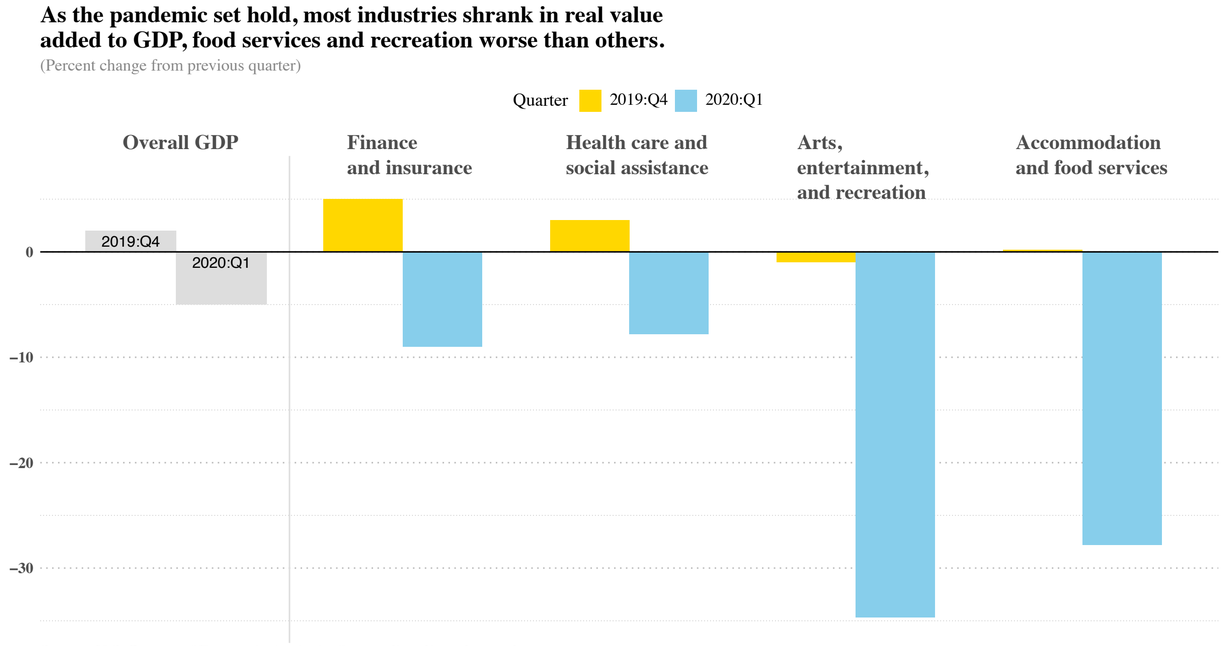
BEA statistics—including GDP, personal income, the balance of payments, foreign direct investment, the input-output accounts, and economic data for states, local areas, and industries—are available at www.bea.gov. E-mail alerts are also available.



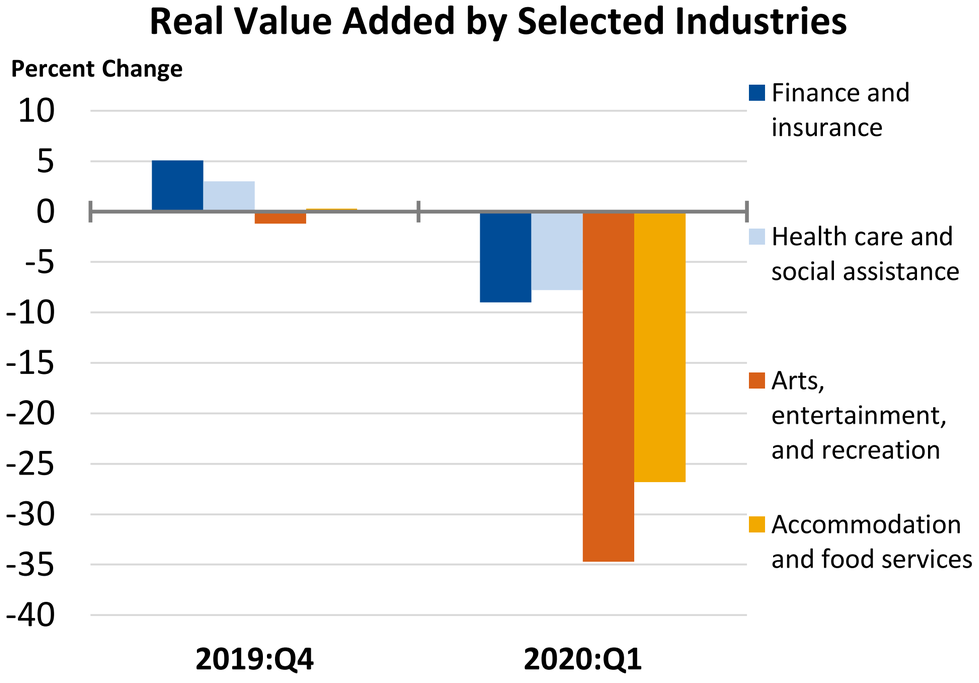
redesigns, example — first possible redesign. Does this redesign more intuitively convey a point?



U.S. Bureau of Economic Analysis Seasonally adjusted annual rates

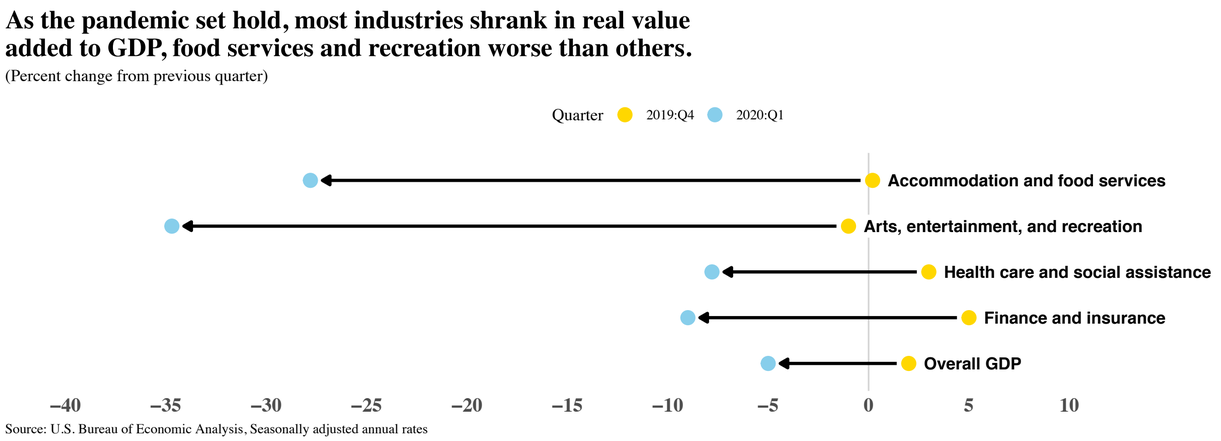


redesigns, example — second possible redesign. Does this redesign more intuitively convey a point?

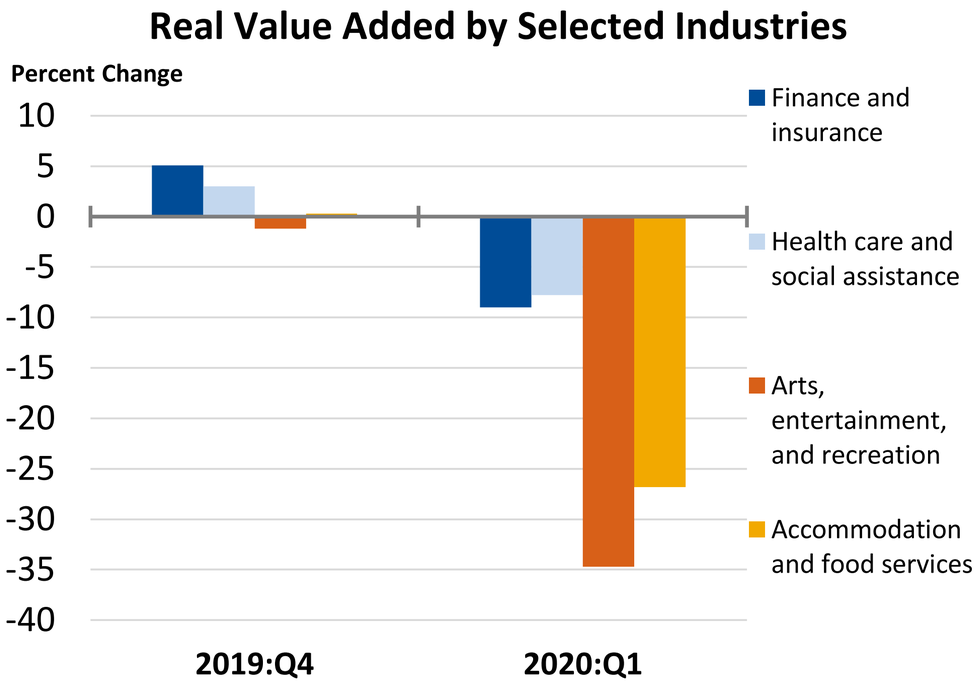


U.S. Bureau of Economic Analysis

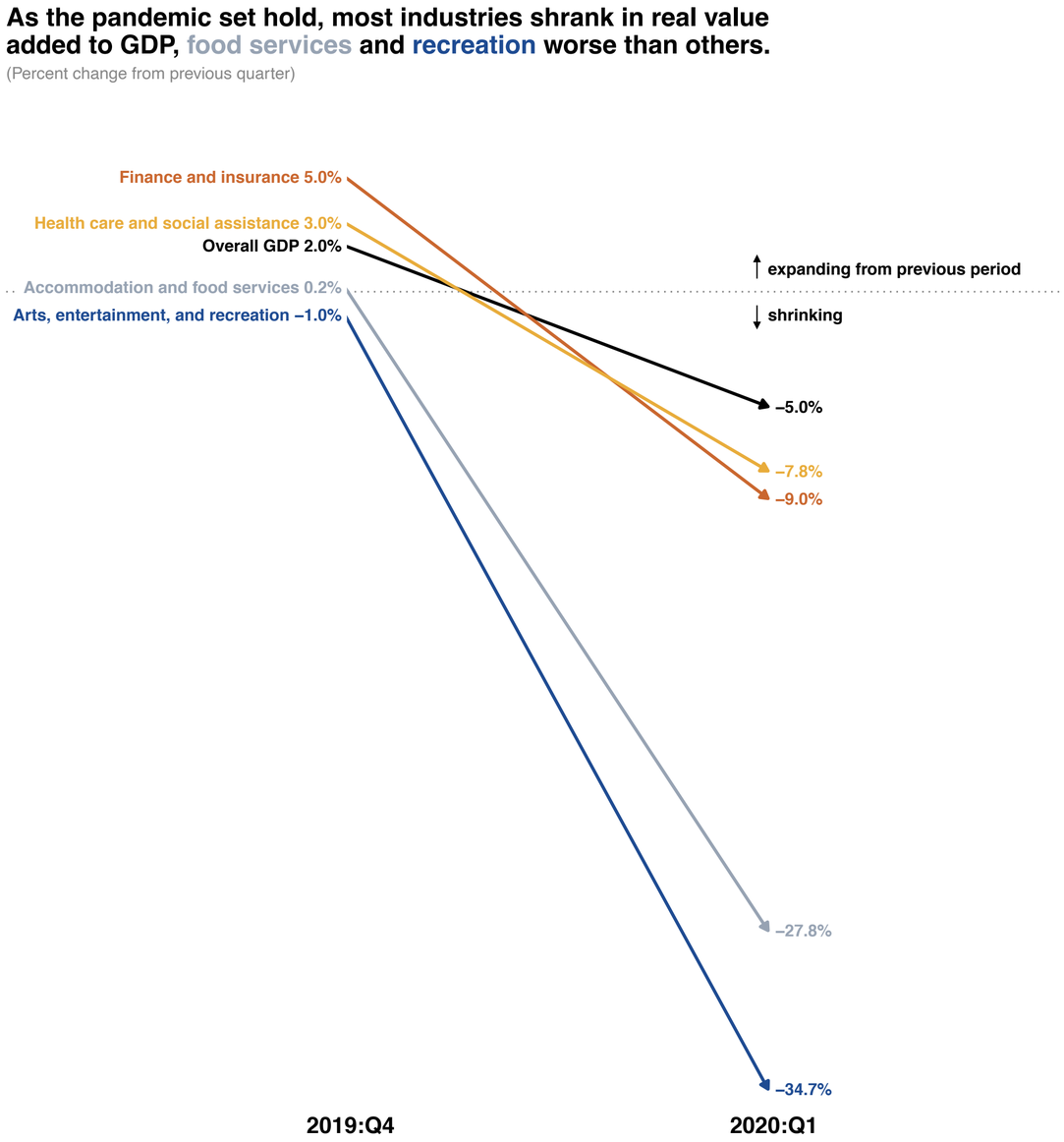
Seasonally adjusted annual rates



redesigns, example — third possible redesign. Does this redesign more intuitively convey a point?



U.S. Bureau of Economic Analysis Seasonally adjusted annual rates



Source: U.S. Bureau of Economic Analysis, Seasonally adjusted annual rates

elements of writing, *fundamentals*

I write entirely to find out what I'm thinking, what I'm looking at, what I see, and what it means.

— Didion, Joan, *writer*

fundamentals, why we communicate — with *others*

Get our audience(s) to

pay attention to,
understand,
(be able to) **act upon**

a maximum of **messages**,
given **constraints**.

fundamentals, use messages, ***not*** just information

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

A concentration in urban areas (175 µg/m^3) is unacceptably high.

"A *message* differs from raw *information* in that it presents 'intelligent added value,' that is, something [new for your audience] to understand about the information."

— Doumont, *Trees, maps, and theorems*.

fundamentals, three laws of communication applied to narrative

Adapt to your audience

Maximize the signal-to-noise ratio

Use effective redundancy

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.

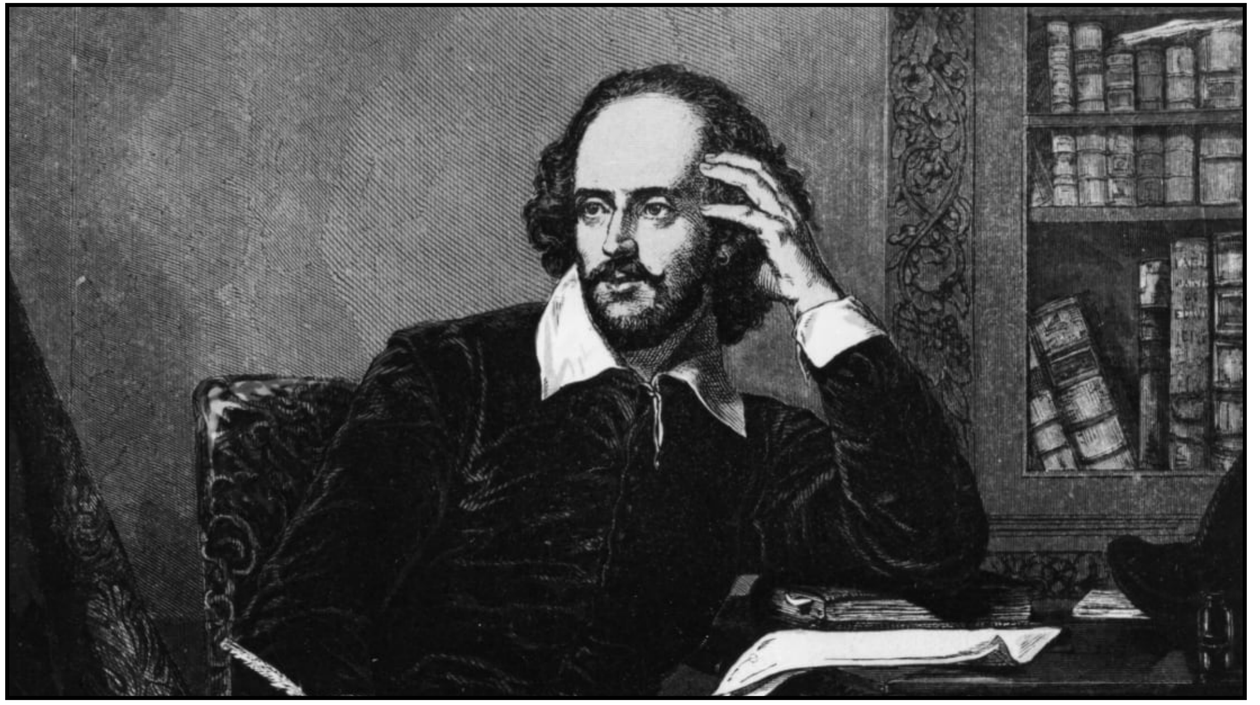
Focusing on what you and your audience have in common, rather than what you disagree about, enables change.

— Tali Sharot, *The Influential Mind*

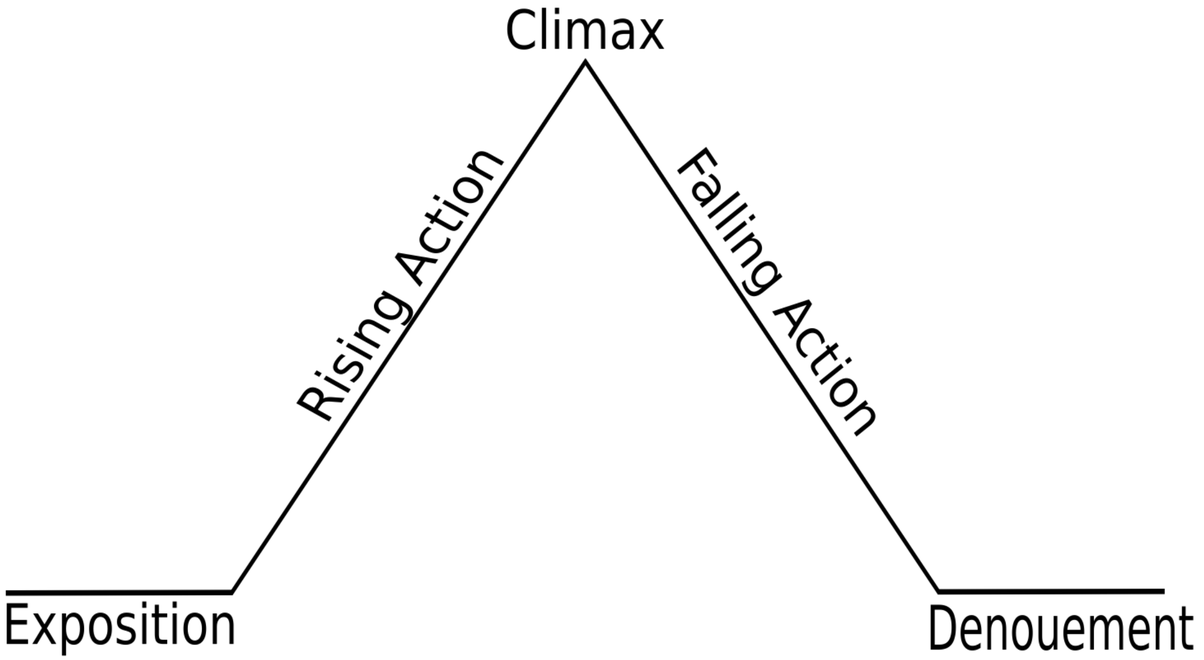
communication structure, first, motivation and message

Motivation	Make the audience receptive to the topic of the communication
Message	Once you have their attention, tell them your main message
Details	Next, support this message: tell them how you got there
Appendix	Last of all, present separately what fewer will want to know

communication structure, **story or narrative**

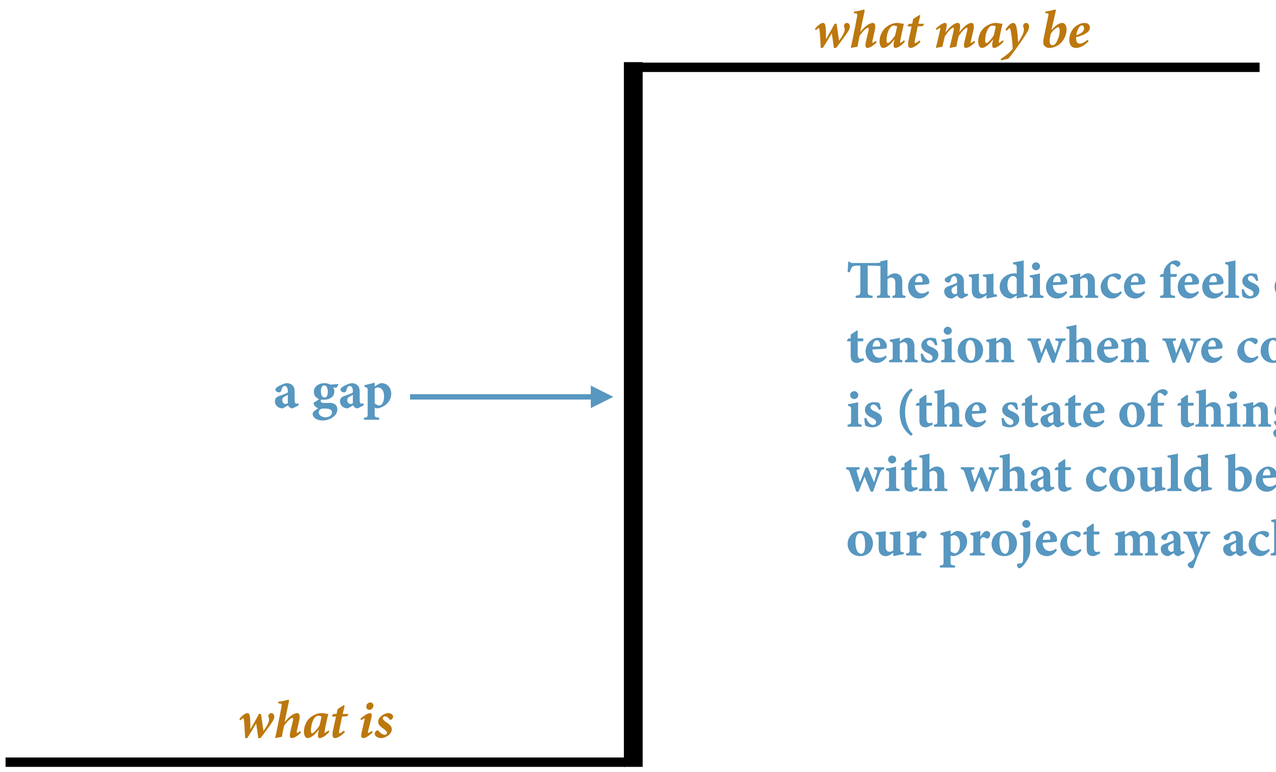


communication structure, story — from Shakespeare to data science?!



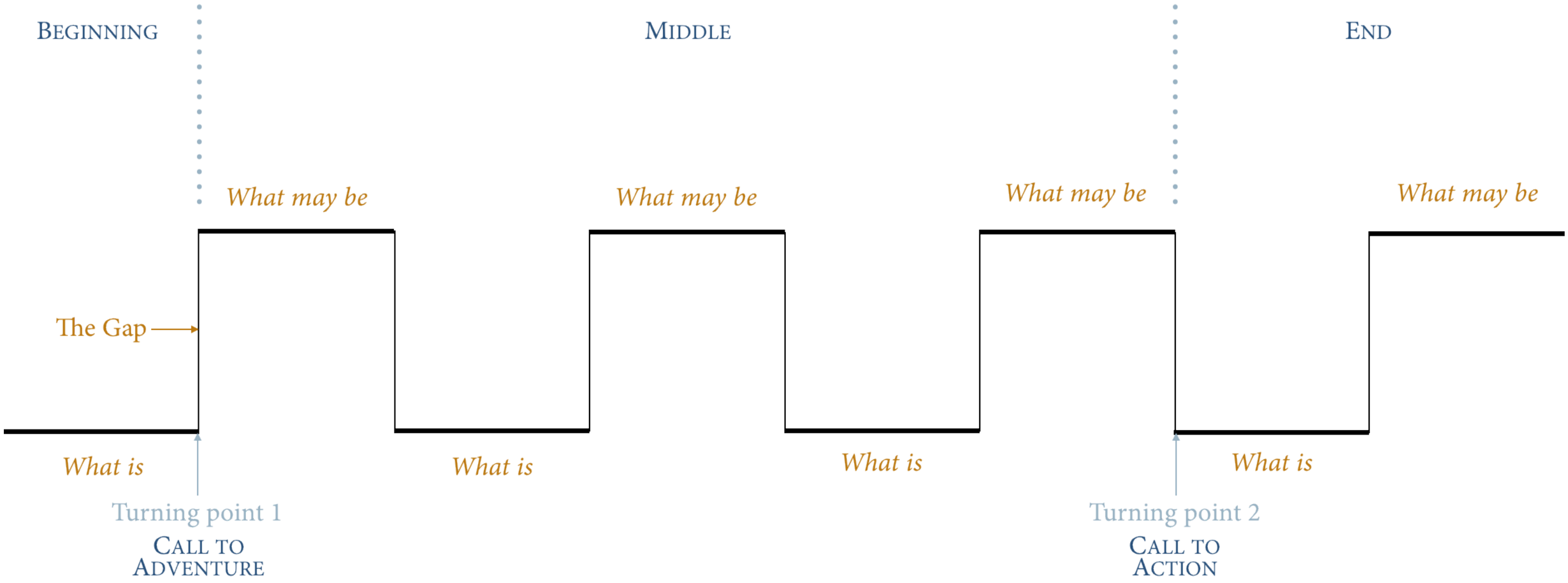
unexpected change

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, old before new

old

new

sentence structure, old before new

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.

communication examples for discussion

CHIEF ANALYTICS OFFICER | heads up a company's data analytics operations, transforming data into business value, and drives data-related business change.

examples for discussion, (more) examples of analytics executives

Kelly Jin
Chief Analytics Officer
City of New York

B.A. Economics, Univ. Penn.
Post-Grad. Ed. in Data Science
Previous analytics appointments

Michael Frumin
Director of Product and Data Science
for Transit, Bikes, and Scooters at Lyft

B.S. Computer Science, Stanford
M.S. Operations Research, MIT
20 years experience with data

Scott Powers
Director of Quantitative Analysis
Los Angeles Dodgers

Ph.D. Statistics, Stanford Univ.
Fluent in R, Publications in
Machine Learning

Blair Borgia
Director of Data Intelligence
ERGO, a startup tech marketing firm

B.A. Math, Eastern. Mich. Univ.
Certifications in Python & SQL
20 years experience with data

examples for discussion, first example *draft* memo

Motivation

Message

Details

Appendix

?

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

2019 February 2

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 a 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.

examples for discussion, first example *draft* memo

Starting with common ground?

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

2019 February 2

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 a 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.

examples for discussion, first example *draft* memo

Unexpected change, information gap?

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

2019 February 2

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 a 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.

examples for discussion, first example *draft* memo

Does the ending echo the lead?

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

2019 February 2

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 a 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.

examples for discussion, first example *draft* memo

Old before new?

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

2019 February 2

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 a 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.

Details?

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

2019 February 2

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 a 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.

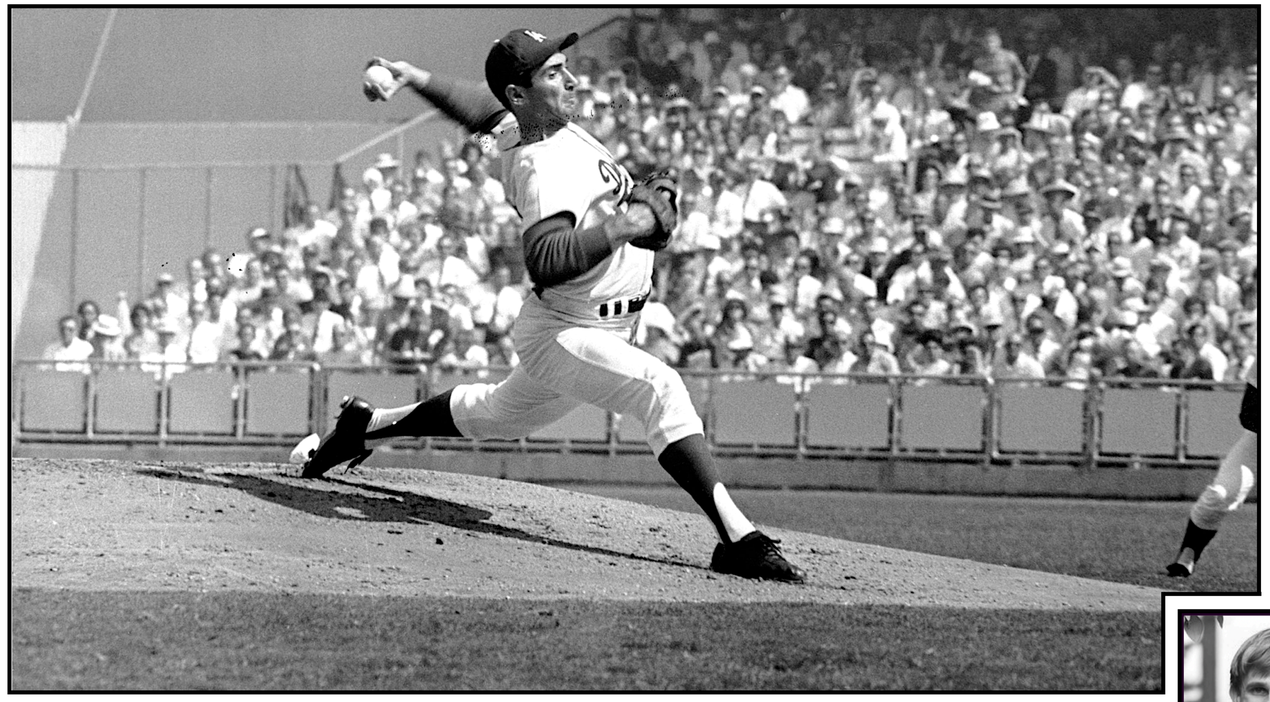
Then, we will build upon the initial work, exploring causal factors such as the availability of alternative transportation (*e.g.*, subway stations near docking stations) and weather. Both of which, we have available data that can be joined using timestamps.

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

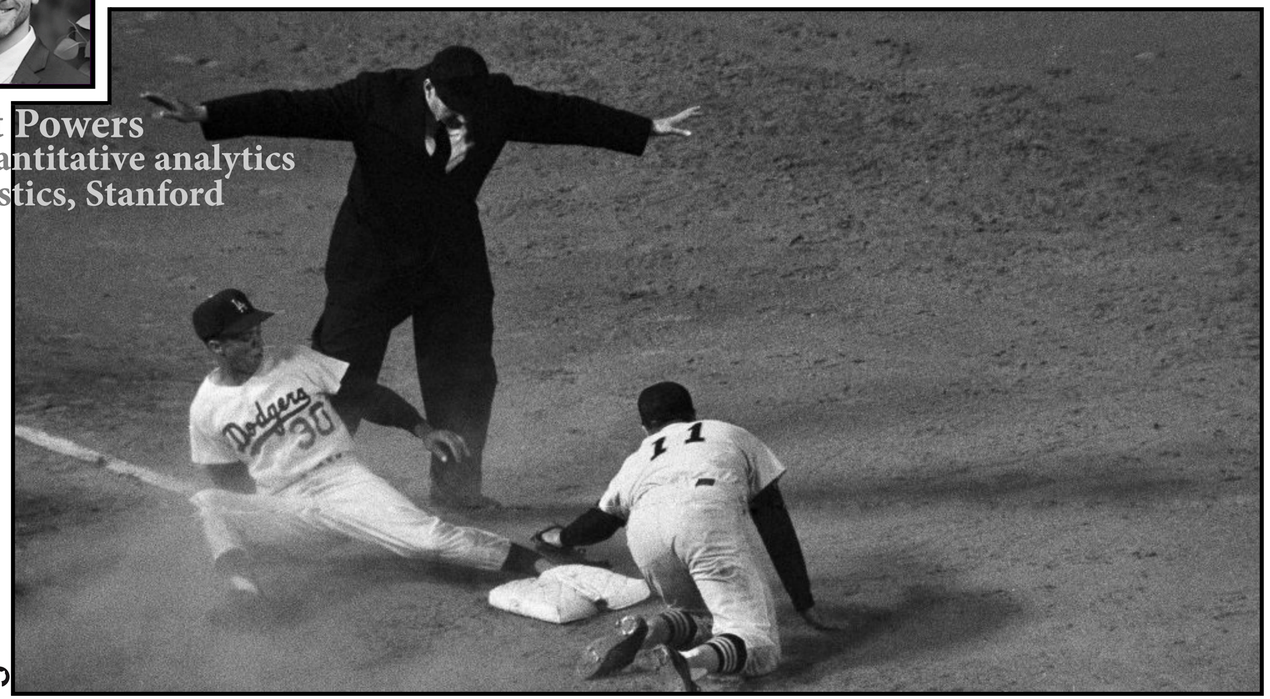
Friedman, Matthew. "*Citi Bike Racks Continue to Go Empty Just When Upper West Siders Need Them*." News. West Side Rag (blog), August 19, 2017. https://www.westsiderag.com/2017/08/19/citi-bike-racks-continue-to-go-empty-just-when-upper-west-siders-need-them.

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



statistics, probability, computing

Scott Powers
Director of quantitative analytics
PhD Statistics, Stanford



baseball

examples for discussion, second example *draft* memo

Motivation

Message

Details

Appendix

?

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

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

Unexpected change, information gap?

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

Does the ending echo the lead?

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

Old before new?

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

Details?

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

individual memo, group projects — data resources

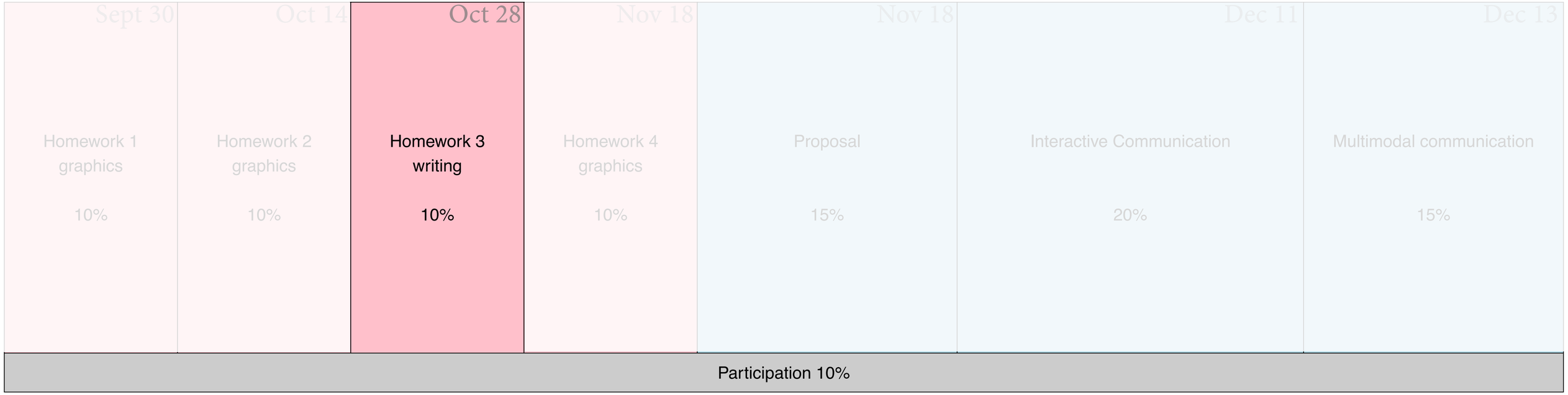
next deliverable, homework three

Individual Work

For learning data visualization
and written narrative techniques

Group work

For building graphics and narrative
into interactive communications



analytics projects, a few (of many, many) starting points for finding — and get help finding — data

Columbia University Library Research Data Services

Research Data Services is jointly supported by the Libraries and CUIT, providing support and consulting for research data needs at Columbia University. Our **expert staff are available to help** with many aspects of the research data lifecycle including **research**, **data management**, **finding data**, recommendations for **cleaning** and **understanding** data, **mapping** and **visualizing** your data.

https://library.columbia.edu/services/research-data-services.html

Columbia Library Clio database search

Real-time and historical SEC EDGAR filings, scanned images of company annual reports and foreign exchange filings.

https://clio.columbia.edu/databases?q=research+reports

Social media: Ravindran, Sharan Kumar, and Vikram Garg. *Mastering Social Media Mining with R*. Packt Publishing, 2015. Print. Clio: https://clio.columbia.edu/catalog/14225862

Web: Munzert, Simon et al. *Automated Data Collection with R*. Wiley, 2015. Print. Clio: https://clio.columbia.edu/catalog/11269563

R's base installation, and many R **packages** contain built-in datasets. The command `data(package = .packages(all.available = TRUE))` lists all data available in all your *installed* packages.

The **General Social Survey** includes more than 40 years of personal-interview survey questions on social characteristics and attitudes in the United States. http://gss.norc.org

Global cities' OpenData provides public access to numerous global cities' data sets gathered from their agencies: *e.g.*, New York City https://opendata.cityofnewyork.us/data/; London https://data.london.gov.uk; Hong Kong https://data.gov.hk/en/

Data.gov is a USA federal collection of datasets. https://www.data.gov **Of note, other countries offer this too.**

Kaggle is an online community of data scientists owned by Google who publish data sets, over 14,000 now, for public use. https://www.kaggle.com/datasets

Google Dataset Search is just like a regular Google search, but focused on datasets. https://toolbox.google.com/datasetsearch

Google Trends is provides data on the relative interest of any keyword searches over time.: https://trends.google.com/trends/

resources

References

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

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.

Doumont, Jean-Luc. "Fundamentals." In *Trees, Maps, and Theorems. Effective Communication for Rational Minds*. Principiæ, 2009.

Spencer, Scott. Memo to Scott Powers, L.A. Dodgers. "*Our Game Decisions Should Optimize Expectations; Let's Test the Concept by Modeling Decisions to Steal*." February 2, 2019.

———. Memo to Michael Frumin, Citi Bike. "*To Inform Rebalancing, Let's Explore Bike and Docking Availability in the Context of Subway and Weather Information*." February 2, 2019.

Sharot, Tali. "(Priors) Does Evidence Change Beliefs?" In *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.

Zetlin, Minda. "*What Is a Chief Analytics Officer? The Exec Who Turns Data into Decisions*." CIO, November 2, 2017.

Zinsser, William. "The Lead and the Ending." In *On Writing Well*, Sixth. The Classic Guide to Writing Nonfiction. Harper Resource, 2001.

extra

group exercise

group exercise, revise analytics write-up for new audience

Improving traffic safety through video analysis in Jakarta

group exercise, revise write-up for new audience

"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

group exercise, revise write-up for new audience — head of data & analytics, Jakarta

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"

group exercise, revise to motivate *Jakarta's analytics executive* to request project proposal — 250 word limit

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.