Storytelling with data

02 | *data* for analytics projects, and elements of *writing*

Scott Spencer | Columbia University


Write **memo** pitching idea to **Chief Analytics Officer**
Present **critique** of exemplary **information graphic**
Build **interactive graphics** for **Chief Marketing Officer** to explore project and insights
Write **critique** of **peer presentation**
YOU ARE HERE
FINAL
FINAL
DRAFT
FINAL
FINAL
DRAFT
FINAL
DRAFT
FINAL
DRAFT
FINAL
Ideate a **data analytics project** addressing problem or opportunity for **your audience**
Write a **project proposal** to **Chief Analytics Officer**
Create **information graphics** of project and insights for an **External Audience**
Present **project** and **insights** to **Chief Executive Officer**
CONDUCT DATA ANALYSIS

***data* for analytics projects**

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).

Structured Data | data that can be stored in a table, and every instance in the table has the same structure (i.e., set of attributes).

Unstructured Data | data where each instance in the data set may have its own internal structure, and this structure is not necessarily the same in every instance.

Nominal types are *names* for categories, classes, or states of things.

Ordinal types are similar to nominal types, except 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.

example data found for class Citi Bike project

Examples of publicly available data sources

Data are recorded of each **bike** unlocked and docked, along with remaining **dock** capacities at the locations, dates, and times of each event: https://www.citibikenyc.com/system-data

Taxi pickup and drop-off locations and times: https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page

Subway lines entrance/exit locations: https://data.cityofnewyork.us/Transportation/Subway-Stations/arq3-7z49

Historical **weather**: https://www.weather.gov/documentation/services-web-api

Traffic data and more: http://www.nyc.gov/html/dot/html/about/datafeeds.shtml#realtime

 example data found for class Citi Bike project

Examples of publicly available data sources

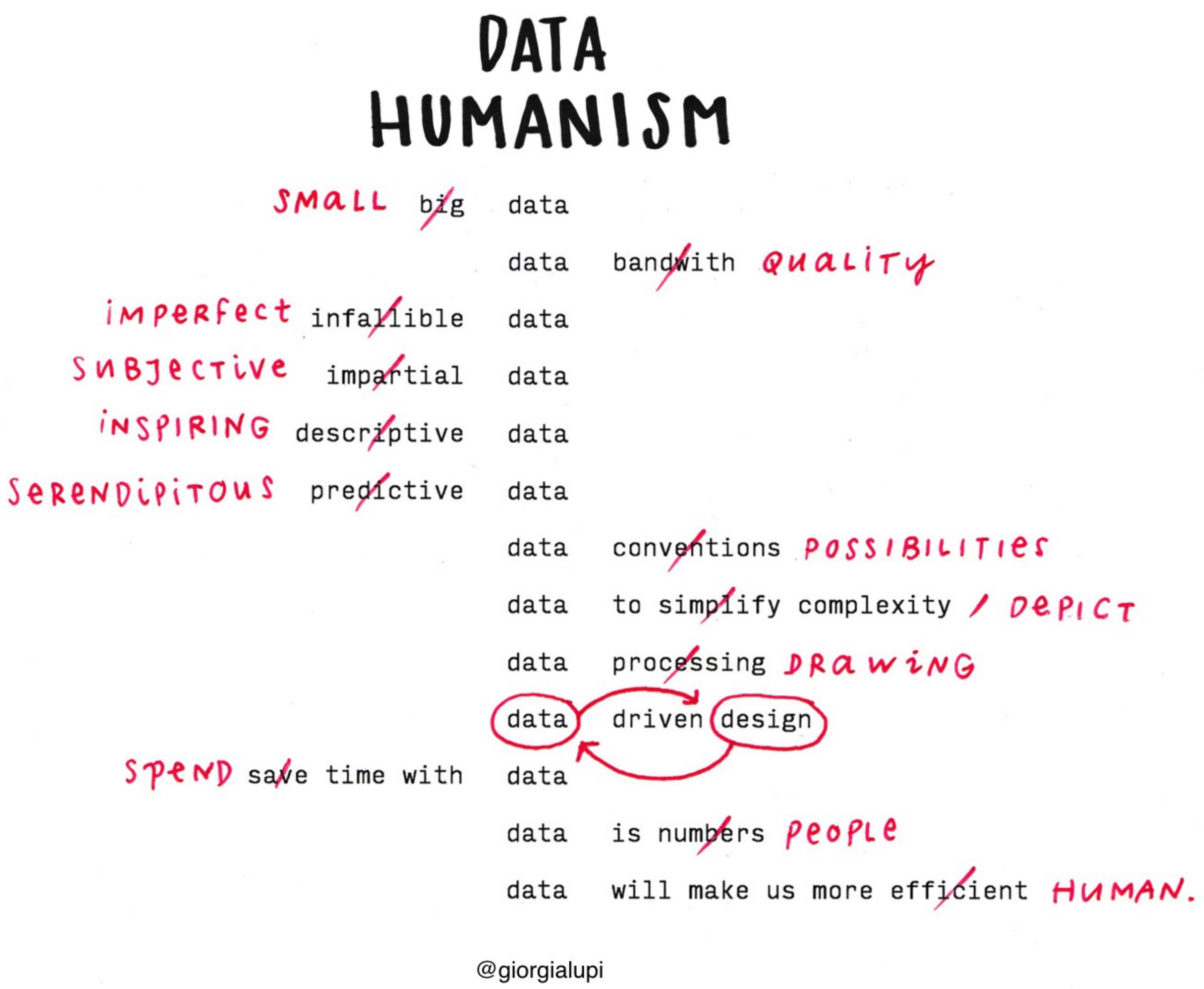
Data are recorded of each **bike** unlocked and docked, along with remaining **dock** capacities at the locations, dates, and times of each event: https://www.citibikenyc.com/system-data

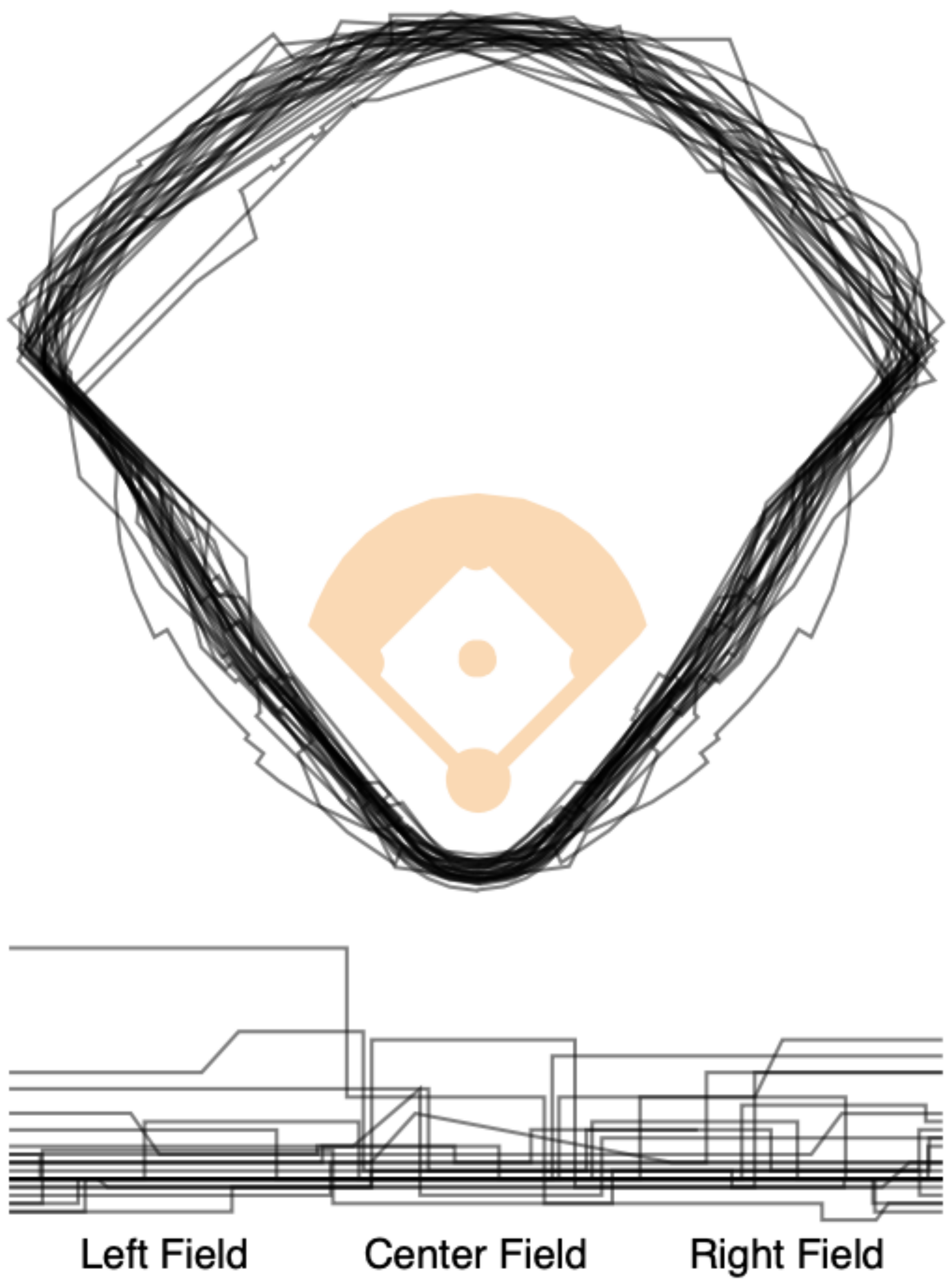
```r
# manually download and unzip file into directory, then load into R
data <- read.csv("JC-202012-citibike-tripdata.csv", header = TRUE)


# or for reproducible analysis and communication, download straight into R
temp <- tempfile()

url <- paste0(
  "https://s3.amazonaws.com/tripdata/JC-",
  "202012",
  "-citibike-tripdata.csv.zip")

download.file(url, temp)

data <- read.csv(unz(temp, "JC-202012-citibike-tripdata.csv"),
                 header = TRUE )
unlink(temp)
```


```r
# setup twitter developer account: https://developer.twitter.com/en/apply-for-access for keys

library(rtweet)

TWITTER_KEY    <- "<enter your key from dev.twitter.com>"
TWITTER_SECRET <- "<enter your key from dev.twitter.com>"
ACCESS_TOKEN   <- "<enter your key from dev.twitter.com>"
ACCESS_SECRET  <- "<enter your key from dev.twitter.com>"

twitter_token <-
  create_token(
    app             = "apan_teaching",
    consumer_key    = TWITTER_KEY,
    consumer_secret = TWITTER_SECRET,
    access_token    = ACCESS_TOKEN,
    access_secret   = ACCESS_SECRET)

cb <- get_timeline('CitiBikeNYC', n = 100, token = twitter_token)
```

Data represents real life. It is a snapshot of the world in the same way that a picture catches a small moment in time. Numbers are always placeholders for something else, a way to capture a point of view—but sometimes this can get lost.

— Giorgia Lupi, *Information Designer*




Left Field
Center Field
Right Field

importance of *comparison* and *change*

comparison | necessary for meaning

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, Robert, *Statistician, Professor*

The fundamental analytical act in statistical reasoning is to answer the question 'Compared with what?'

— Tufte, Edward, *Statistician, Professor, Data Visualization Expert*

 example — importance of statement?

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

 example — choice of comparison depends on point of message or goal

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

average life expectancy of males in United States, 68.5 years

life expectancy of males at least 32 years old,
average appointment age of a first conducting post, 72.0 years

Questions so far on your analytics projects and data? — your soundboard

(business) communication, *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*

Get our audience(s) to

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

↓

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

A concentration of 175 µg per m³ has been observed in urban areas.

A concentration in urban areas (175 µg/m³) 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*.

Adapt to your audience

Maximize the signal-to-noise ratio

Use effective redundancy

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 examples for discussion and group exercise

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

Motivation

Message

Details

Appendix

?

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.

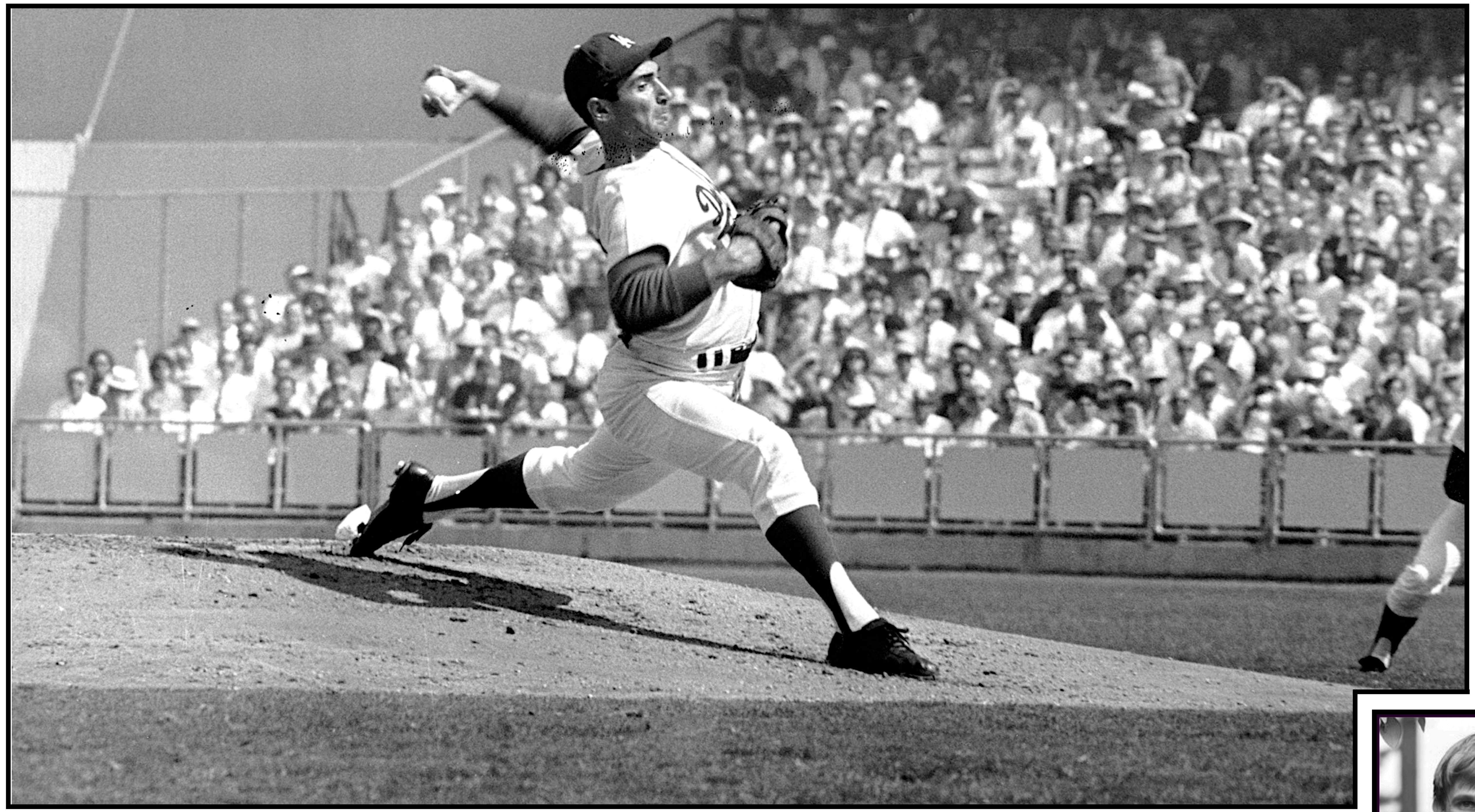
Then, we will build upon the initial work, exploring causal factors such as the availability of alternative transportation (*e.g.*, subway stations near docking stations) and weather. Both of which, we have available data that can be joined using timestamps.

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.

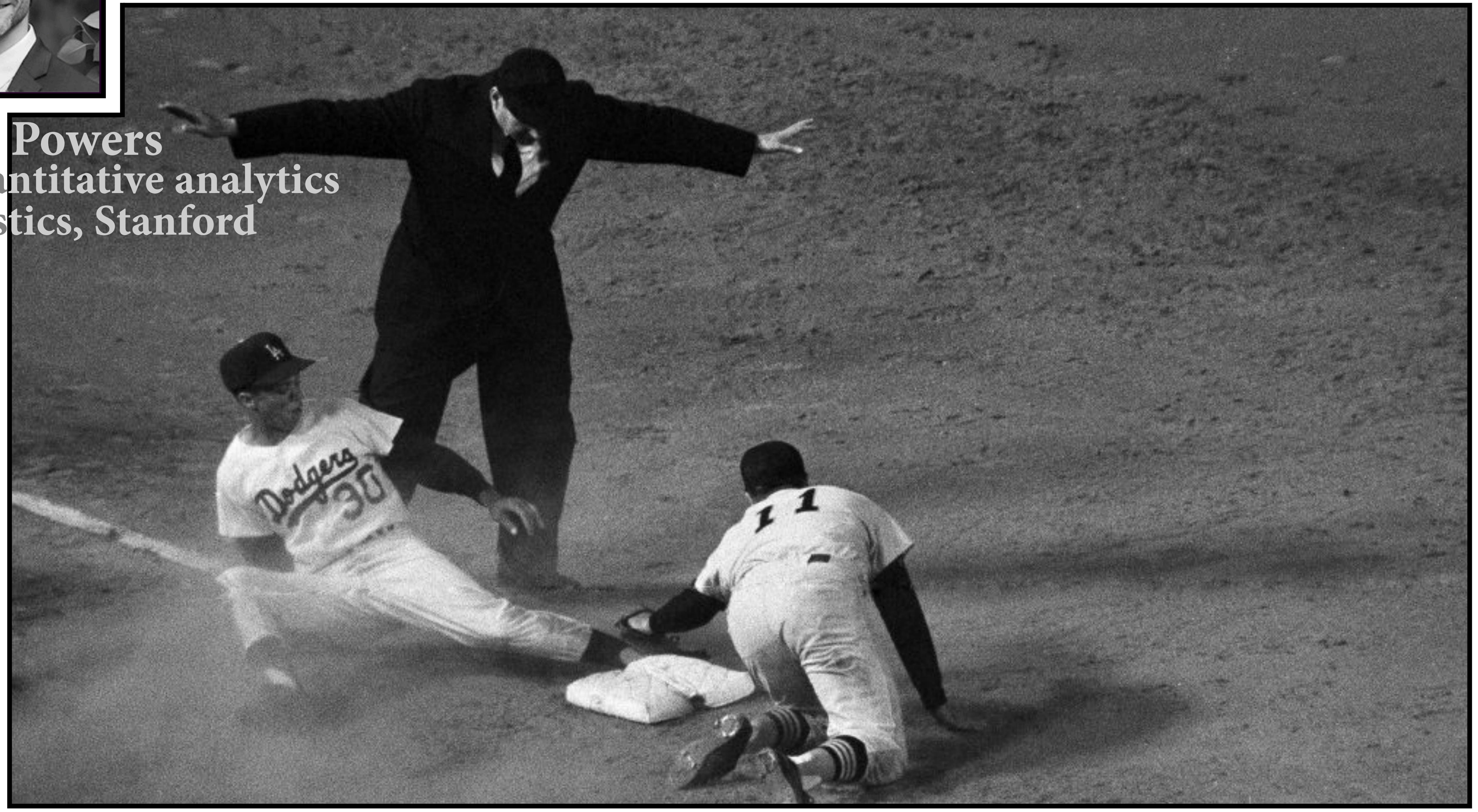
Saldarriaga, Juan Francisco. "*CitiBike Rebalancing Study*." Spatial Information Design Lab, Columbia University, 2013. https://c4sr.columbia.edu/projects/citibike-rebalancing-study.



statistics, probability, computing

Scott Powers
Director of quantitative analytics
PhD Statistics, Stanford

baseball



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

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

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"

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.

resources

References

Spencer, Scott. "Data Measures in Analytics Projects", "Understanding Data Requires Context", and "Elements of Writing." In *Data in Wonderland*. 2021. https://ssp3nc3r.github.io/data_in_wonderland.

Caldeira, Joao, Alex Fout, Aniket Kesari, Raesetje Sefala, Joe Walsh, and Katy Dupre. *Improving Traffic Safety Through Video Analysis: Pulse Lab Jakarta*. Data Science for Social Good. 2018. https://dssg.uchicago.edu/project/improving-traffic-safety-through-video-analysis/ .

Caldeira, Joao, Alex Fout, Aniket Kesari, Raesetje Sefala, Joseph Walsh, Katy Dupre, Muhammad Rizal Khaefi, et al. *Improving Traffic Safety Through Video Analysis in Jakarta, Indonesia*. In Conference on Neural Information Processing Systems NeurIPS, 1–5, 2018.

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

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.

Kelleher, John D, and Brendan Tierney. *Data Science*. MIT Press, 2018.

———. "What Are Data, and What Is a Data Set?" In *Data Science*. MIT Press, 2018.

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

Lupi, Giorgia. "*DATA HUMANISM: The Revolution Will Be Visualized*." Print 70, no. 3 (2016): 76–85.

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

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.

Vickars, Sam. "*The Irregular Outfields of Baseball*." Business. The Data Face (blog), April 2019. http://thedataface.com/2019/04/sports/baseballs-irregular-outfields .

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