Storytelling with Data

Module 02: Scoping a project, finding data, considering an analysis, writing a short memo for data analytics

Scott Spencer
Faculty and Lecturer
Columbia University

Highlights from online discussion

Agenda

Thinking through an analytics project

Citi Bike: group ideas on scoping

Project progression & deliverables

Getting started, finding data

Writing about data analytics

Up next week

Thinking through an analytics project

Scoping a data analytics project

Spencer

Progression

- Decisions
- Goals and actions
- Methods
- Data

} Iterative!

Initial questions

- What problem is to be solved?
- Is the problem important?
- Could an answer 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?

Spencer, Scott. "Scoping a Data Analytics Project." ssp3nc3r.github.io , 3 Jan. 2019.

Citi Bike: group ideas on beginning to scope a project

Citi Bike NYC

Group ideas on beginning to scope a data analytics project

Guerriero, Michael. "Interactive: A Month of Citi Bike." New Yorker., 26 Jul. 2013. Web. 26 Jan. 2019, https://www.newyorker.com/news/news-desk/interactive-a-month-of-citi-bike

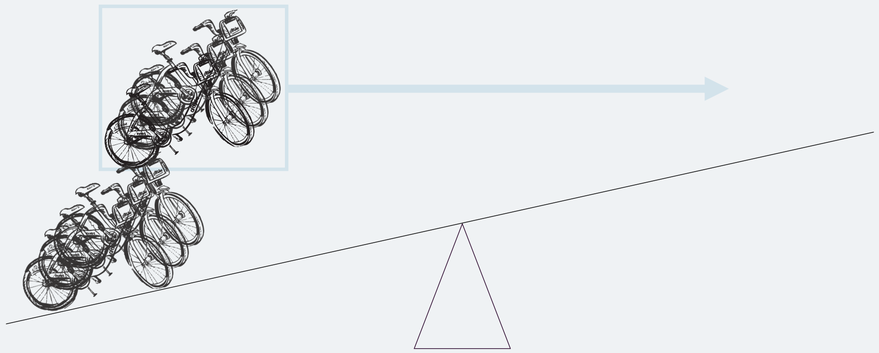
Background

In 2013, the NYC 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.

The Challenge

Making bikes *and* stations available.



> **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 challenge.
>
> **Dani Simons — Citi Bike Spokeswoman**

West Sider. "Citi Bike Racks Continue to Go Empty Just When Upper West Siders Need Them ." westsiderag.com. N.p., 19 Aug. 2017. Web. 2 June 2019.

How can we identify causes, relationships?

Key questions to ask:

Identifying events and user behavior

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**?

Measurements of events and behaviors

How may these events and behaviors have been **measured** and **recorded**?

What **data** are **available**? Where? What form?

May these data be **sufficient to find insights** through analysis, useful for decisions and goals?

Spencer, Scott. “Exploring CitiBike ride data.” ssp3nc3r.github.io , 28 Feb. 2019.

Identifying Citi Bike ride *and related* data

A proposed analysis

Explore 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 **weather**.

Identified data, measurements

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:
http://www.nyc.gov/html/tlc/html/about/trip_record_data.shtml

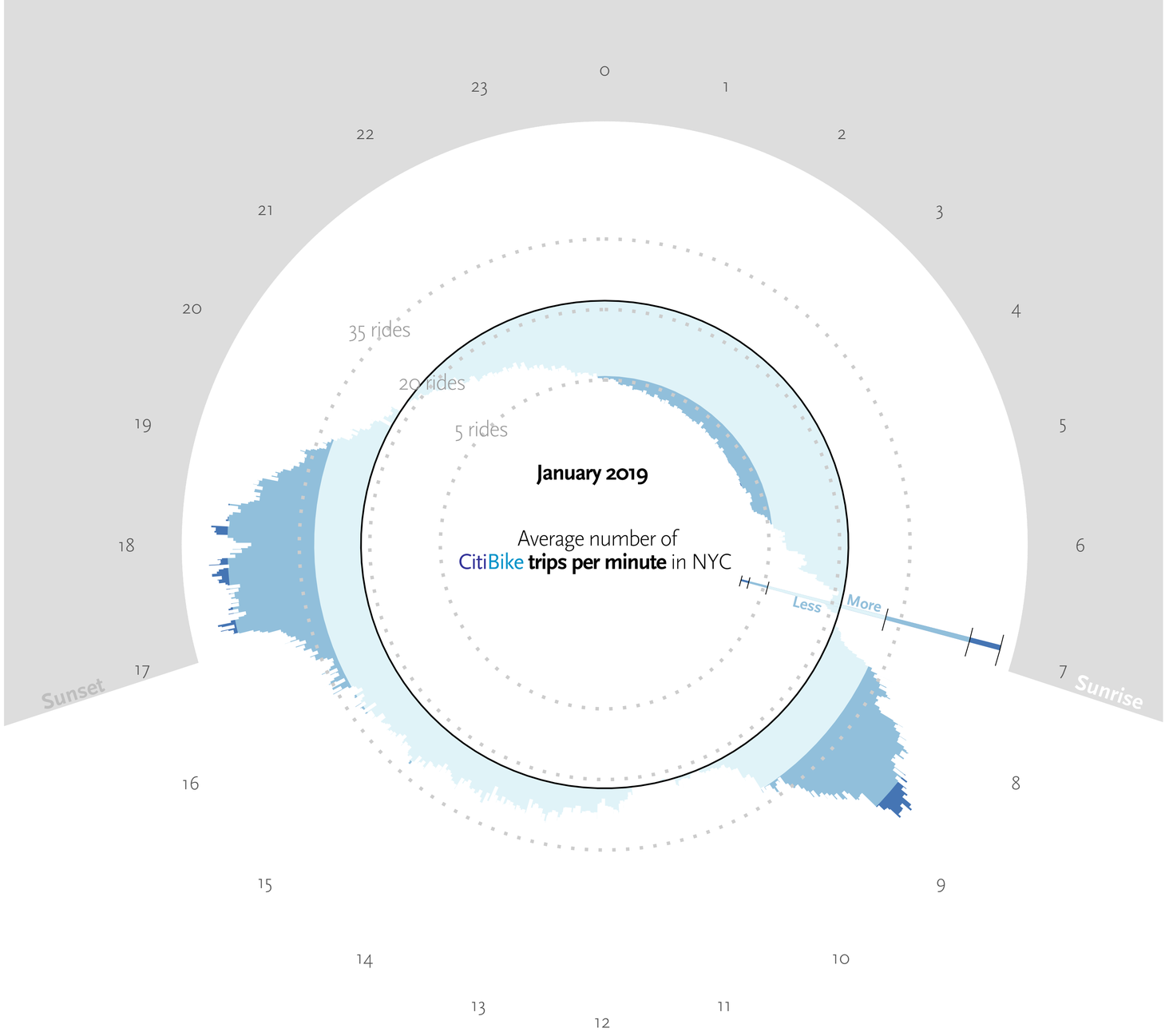
Subway lines entrance/exit locations:
https://data.cityofnewyork.us/Transportation/Subway-Stations/arq3-7z49

Historical **weather**:
https://darksky.net/dev

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

Spencer, Scott. "Exploring CitiBike ride data." ssp3nc3r.github.io , 28 Feb. 2019.

Example, visually exploring the data

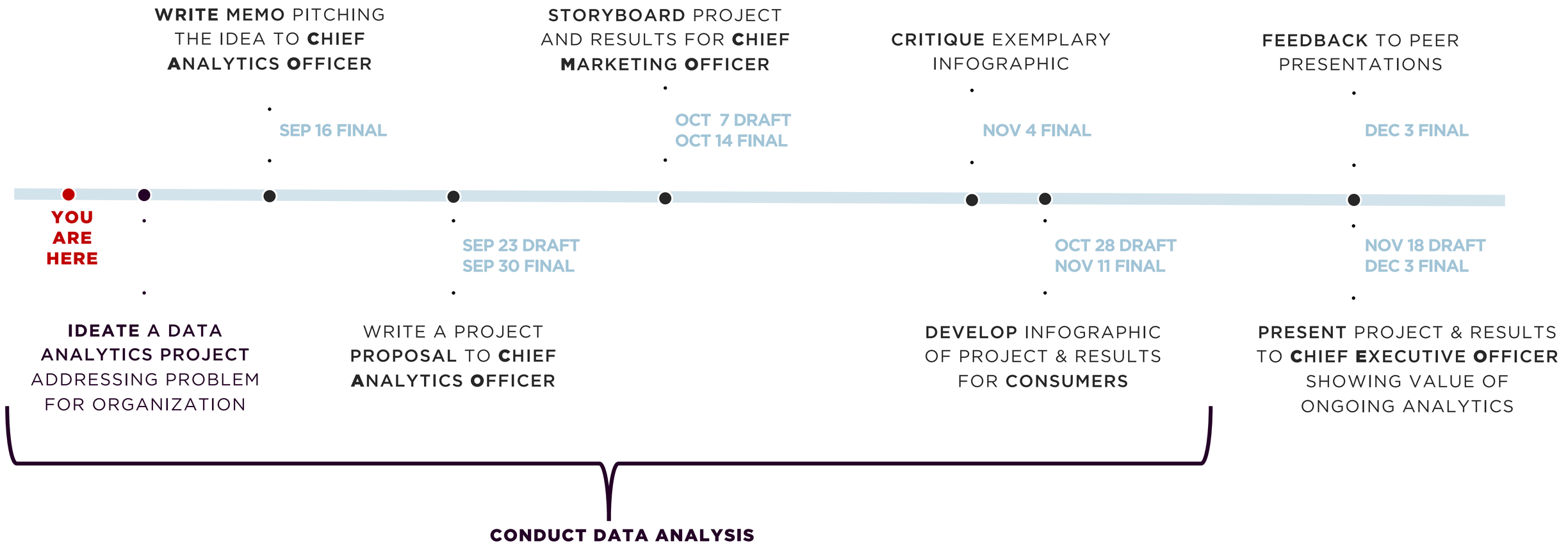


Spencer, Scott. "Exploring CitiBike ride data." ssp3nc3r.github.io , 28 Feb. 2019.

Project progression & deliverables

Project progression

The scenario — You identify a likely business problem in an organization. You develop a data analysis project that provides insights, useful for decision making. You attempt to persuade others to allow the project to move forward, and bring insights that help further organization goals.



Main deliverables

For your chosen company and case study,
as an imagined member of the analytics team ...

250-word memo

Write a memo to **CAO** about an opportunity to leverage analytics. Consider background context, problem, data, solution, and impact.

750-word proposal

On approval of the memo, write a proposal to **CAO**, detailing the anticipated project.

Storyboard

Present project result in storyboard to **CMO**, using narrative forms, and with comparisons, metaphors and other storytelling concepts.

Infographic

Recraft the results, telling the narrative through an infographic for the **public** or **potential consumers**, using data visualization with brand awareness.

Presentation

Construct and deliver a 4-5 minute persuasive presentation with up to 10 slides to the **CEO**, telling the story of the analytics project to convince them of further investment in analytics.

Getting started, finding data

Selecting an entity, starting your project, **now**:

Choose something interesting to you!

Chose an **entity** — for profit or not for profit — and **analytics project** that you may be interested. Do **not** chose one of the example case studies.

Research and imagine a problem the entity may face

The problem may be **similar in complexity**, importance, and interest to that of Citi Bike or to that of the other case studies in our readings.

Research available data, relevant to the problem

You should find **publicly available data** to analyze. This data need not be specific to the entity, just **relevant** to the analysis — for example, weather for Citi Bike.

Consider how you plan to analyze the data

Are the data, or can you **wrangle** it, into a form to perform your analysis within your available time? Assume your analysis will include **aggregating**, **summarizing**, **visualizing**, **regressing**, and most of all, **comparing**.

Example resources

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

Thomson research

Real-time and historical SEC EDGAR filings, scanned images of company annual reports and foreign exchange filings.
https://clio.columbia.edu/databases/5410648

FactSet

Broad financial news on markets and economies or granular news on company earnings, FactSet news and research offers real-time news on the factors that affect markets and research that pulls in data from many trusted sources.
https://library.columbia.edu/find/eresources/databases/factset.html

Clio database search

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.

Web: Munzert, Simon et al. *Automated Data Collection with R*. Wiley, 2015. Print. https://clio.columbia.edu/catalog/12746575

R's **base** installation, and many **R packages** contain built-in datasets. The command data() gives you the base R datasets, and including the installed package name, say, data(package="rethinking") lists the datasets in the package.

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

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

NYC OpenData provides public access to numerous data sets gathered from NYC agencies. https://opendata.cityofnewyork.us/data/

Data.gov is a USA federal collection of datasets. https://www.data.gov

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

Writing about data analytics, introduction

Comparing project details described with varying lengths of an award-winning project:

Improving Traffic Safety Through Video Analysis in Jakarta, Indonesia

Caldeira, Joao et al. "Improving Traffic Safety Through Video Analysis: Pulse Lab Jakarta." dssg.uchicago.edu. N.p., 2018. Web. 2 Feb. 2019. https://github.com/dssg/jakarta_smart_city_traffic_safety_public

Questions for Discussion

For whom is the paper written? Who is the **intended audience**?

What **information types and categories** does the paper cover? What **details** do you notice?

How is the information **organized**?

How does the information in, and structure of, the paper **compare** with the initial questions I've suggested we consider when scoping an analytics project?

Caldeira, Joao et al. "Improving Traffic Safety Through Video Analysis: Pulse Lab Jakarta." dssg.uchicago.edu. N.p., 2018. Web. 2 Feb. 2019.

Jakarta example — 124-word proposal summary, from analysts' blog

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.

Caldeira, Joao et al. "Improving Traffic Safety Through Video Analysis: Pulse Lab Jakarta." https://dssg.uchicago.edu/project/improving-traffic-safety-through-video-analysis/ N.p., 2018. Web. 2 Feb. 2019.

Jakarta example — details in 124 words

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.

Caldeira, Joao et al. "Improving Traffic Safety Through Video Analysis: Pulse Lab Jakarta." https://dssg.uchicago.edu/project/improving-traffic-safety-through-video-analysis/ N.p., 2018. Web. 2 Feb. 2019.

Jakarta example — structure of blog post

background context

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.

Caldeira, Joao et al. "Improving Traffic Safety Through Video Analysis: Pulse Lab Jakarta." https://dssg.uchicago.edu/project/improving-traffic-safety-through-video-analysis/ N.p., 2018. Web. 2 Feb. 2019.

Jakarta example — structure of blog post

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.

goals, actions, origin of data

Caldeira, Joao et al. "Improving Traffic Safety Through Video Analysis: Pulse Lab Jakarta." https://dssg.uchicago.edu/project/improving-traffic-safety-through-video-analysis/ N.p., 2018. Web. 2 Feb. 2019.

Jakarta example — structure of blog post

problem

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.

Caldeira, Joao et al. "Improving Traffic Safety Through Video Analysis: Pulse Lab Jakarta." https://dssg.uchicago.edu/project/improving-traffic-safety-through-video-analysis/ N.p., 2018. Web. 2 Feb. 2019.

Jakarta example — structure of blog post

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.

method, data

Caldeira, Joao et al. "Improving Traffic Safety Through Video Analysis: Pulse Lab Jakarta." https://dssg.uchicago.edu/project/improving-traffic-safety-through-video-analysis/ N.p., 2018. Web. 2 Feb. 2019.

Jakarta example — structure of blog post

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.

impact, linked to goals, decisions

Caldeira, Joao et al. "Improving Traffic Safety Through Video Analysis: Pulse Lab Jakarta." https://dssg.uchicago.edu/project/improving-traffic-safety-through-video-analysis/ N.p., 2018. Web. 2 Feb. 2019.

Jakarta example — structure of blog post

background context

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.

problem

impact, linked to goals, decisions

goals, actions, origin of data

method, data

Caldeira, Joao et al. "Improving Traffic Safety Through Video Analysis: Pulse Lab Jakarta." https://dssg.uchicago.edu/project/improving-traffic-safety-through-video-analysis/ N.p., 2018. Web. 2 Feb. 2019.

Questions for Discussion

In what ways do the structure and detail of the proposal in the short blog summary **compare** with the final paper?

How do the structure and detail in this case study **compare** with *The Next Rembrandt*, from last lecture?

If you lengthened this summary to 250 words, what additional details do you think a chief analytics officer may want to know before approving the project?

Caldeira, Joao et al. "Improving Traffic Safety Through Video Analysis: Pulse Lab Jakarta." https://dssg.uchicago.edu/project/improving-traffic-safety-through-video-analysis/ N.p., 2018. Web. 2 Feb. 2019.

ING et al. "The Next Rembrandt." 5 Apr. 2016. Web. https://www.nextrembrandt.com

Example *draft* 250-word memo

A perfect game

Statistics, probability, computing

Models

Mode, maximum likelihood

Probability distributions

Joint distributions

R language and packages

Inferences

Mean, expectations

Decision theory

Counterfactuals

Simulations

Perfect Game

Los Angeles Dodgers

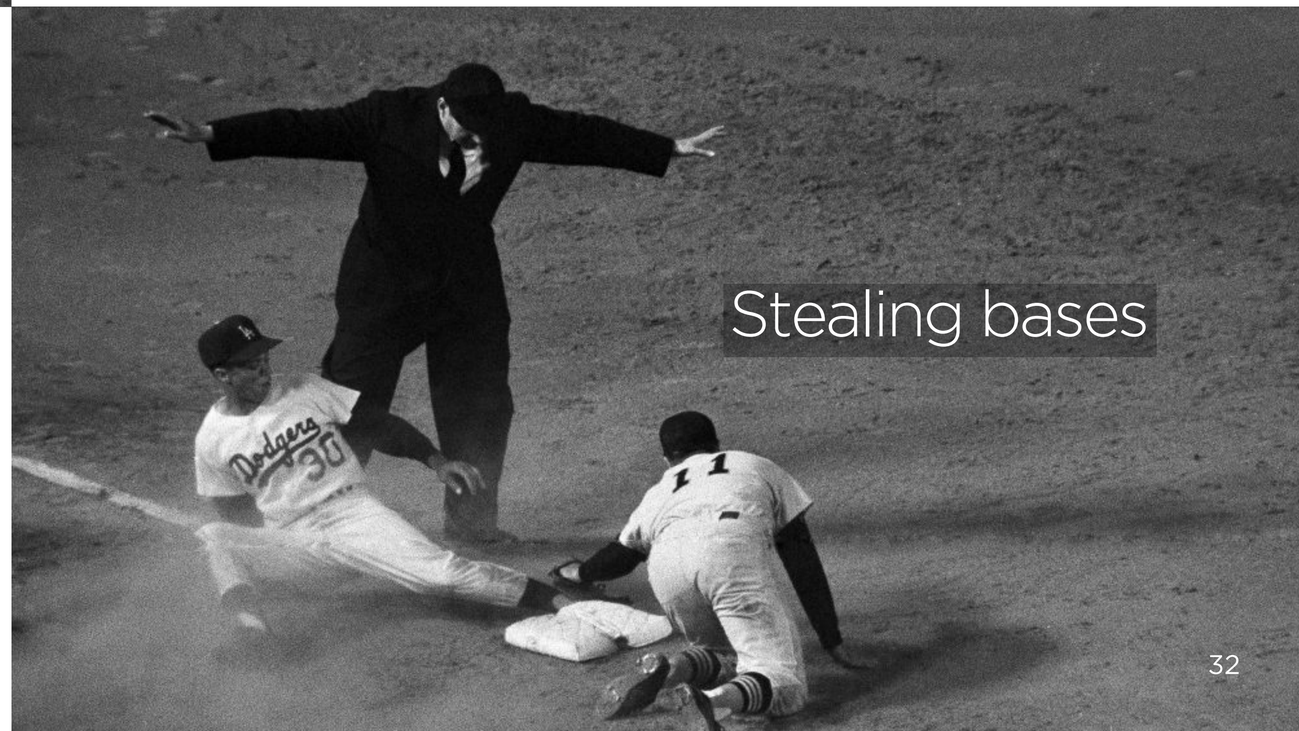
Pitcher Sandy Koufax

Statcast data

(Attempting to) steal a base

Salary cap

Baseball



Stealing bases

To: **Scott Powers**
Director, Quantitative Analytics

2 February 2019

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

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. Decisions from what is most likely, however, 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, compute expectations: 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`. And we can test the concept with decisions to steal.

Sincerely,
Scott Spencer

Questions for Discussion

Who was the **audience**?

Was a **few minutes enough time** to understand?
Did you notice **structure** that guides the audience?

Do the **audiences differ** between this memo,
the *Jakarta* proposal, and *The Next Rembrandt*? If so, how?

Do the **structures and details** of the
communications seem to depend on the audience? If so, how?

Might some **revisions improve** this draft? (We'll revisit this question).

Audience background

Director of Quantitative Analytics
Ph.D. Statistics from Stanford University
Some publications use machine learning
Knows R programming
An employee, knows history of Dodgers

To: **Scott Powers**
Director, Quantitative Analytics

2 February 2019

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

Message first, context

Main body word count, 250

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. Decisions from what is most likely, however, leave wins unclaimed. To claim them, let's base decisions on expectations flowing from decision theory and probability models. A joint [illegible] 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, compute expectations: weight every thing 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`. And we can test the concept with decisions to steal.

Sincerely,
Scott Spencer

Readability Statistics

Counts	
Words	250
Characters	1,357
Paragraphs	4
Sentences	14
Averages	
Sentences per Paragraph	3.5
Words per Sentence	17.8
Characters per Word	5.2
Readability	
Flesch Reading Ease	43.5
Flesch-Kincaid Grade Level	11.6
Passive Sentences	0%

To: **Scott Powers**
Director, Quantitative Analytics

2 February 2019

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

Context, orient the audience

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. Decisions from what is most likely, however, 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, compute expectations: 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`. And we can test the concept with decisions to steal.

Sincerely,
Scott Spencer

To: **Scott Powers**
Director, Quantitative Analytics

2 February 2019

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

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. Decisions from what is most likely, however, 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.

Goal, action problem

After defining our objective (e.g. optimize expected runs) we will, from Statcast data, compute expectations: 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`. And we can test the concept with decisions to steal.

Sincerely,
Scott Spencer

To: **Scott Powers**
Director, Quantitative Analytics

2 February 2019

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

Proposed solution

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. Decisions from what is most likely, however, 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, compute expectations: 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`. And we can test the concept with decisions to steal.

Sincerely,
Scott Spencer

To: **Scott Powers**
Director, Quantitative Analytics

2 February 2019

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

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. Decisions from what is most likely, however, 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, compute expectations: 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.

Objective, data, methods

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`. And we can test the concept with decisions to steal.

Sincerely,
Scott Spencer

To: **Scott Powers**
Director, Quantitative Analytics

2 February 2019

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

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. Decisions from what is most likely, however, 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, compute expectations: 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.

Eventual benefit

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`. And we can test the concept with decisions to steal.

Sincerely,
Scott Spencer

To: **Scott Powers**
Director, Quantitative Analytics

2 February 2019

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

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. Decisions from what is most likely, however, 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, compute expectations: 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.

Benefit, comparison, financial

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`. And we can test the concept with decisions to steal.

Sincerely,
Scott Spencer

To: **Scott Powers**
Director, Quantitative Analytics

2 February 2019

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

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. Decisions from what is most likely, however, 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, compute expectations: 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`. And we can test the concept with decisions to steal.

Limitations, short-term solutions

Sincerely,
Scott Spencer

To: **Scott Powers**
Director, Quantitative Analytics

2 February 2019

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

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. Decisions from what is most likely, however, 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, compute expectations: 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`. And we can test the concept with decisions to steal.

Reminder of proposal

Sincerely,
Scott Spencer

Author of memo

To: **Scott Powers**
Director, Quantitative Analytics

Audience background

Senior Analyst
Ph.D. Statistics from Stanford University
Some publications use machine learning
Knows R programming
An employee, knows history of Dodgers

2 February 2019

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

Message first, context

Context, orient the audience

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. Decisions from what is most likely, however, 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.

Goal, action problem

Proposed solution

Objective, data, methods

After defining our objective (e.g. optimize expected runs) we will, from Statcast data, compute expectations: 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.

Eventual benefit

Benefit, comparison, financial

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.

Limitations, short-term solutions

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`. And we can test the concept with decisions to steal.

Reminder of proposal

Sincerely,
Scott Spencer

Author of memo

Group help on case studies

Group help on case studies

Work with your peers to **ideate**, and **give feedback** on,
your ideas for possible choices of **entity** and **business problem**,
your **data analytics project**, and **potential sources of data**.

Then, let's **share** some of these ideas together.

On deck through next week

For Next Week, Module 3:

Agenda next week

Decide on your data analytics project; **locate** and begin **collecting** available data for analysis.

Turn in a **final**, concise **memo** to CAO on your pitch to use data analytics to gain insight into the business problem you have identified.

The minimum

Booth, Wayne C et al. *The Craft of Research*. Fourth. University of Chicago Press, 2016, Chapter 17 *Revising Style: Telling Your Story Clearly*

Consider whether the Jakarta blog post and the Dodger's draft memo implemented any of these principles, and if so, how.

Doumont, Jean-Luc. *Trees, Maps, and Theorems: Effective communication for rational minds*. Principiæ, 2009. selected pages.

Understand how to structure materials and content from specific audiences' points of view.

Butterick, Matthew. *Butterick's Practical Typography*. practicaltypography.com. N.p., 2018. Web. 8 Sept. 2018.

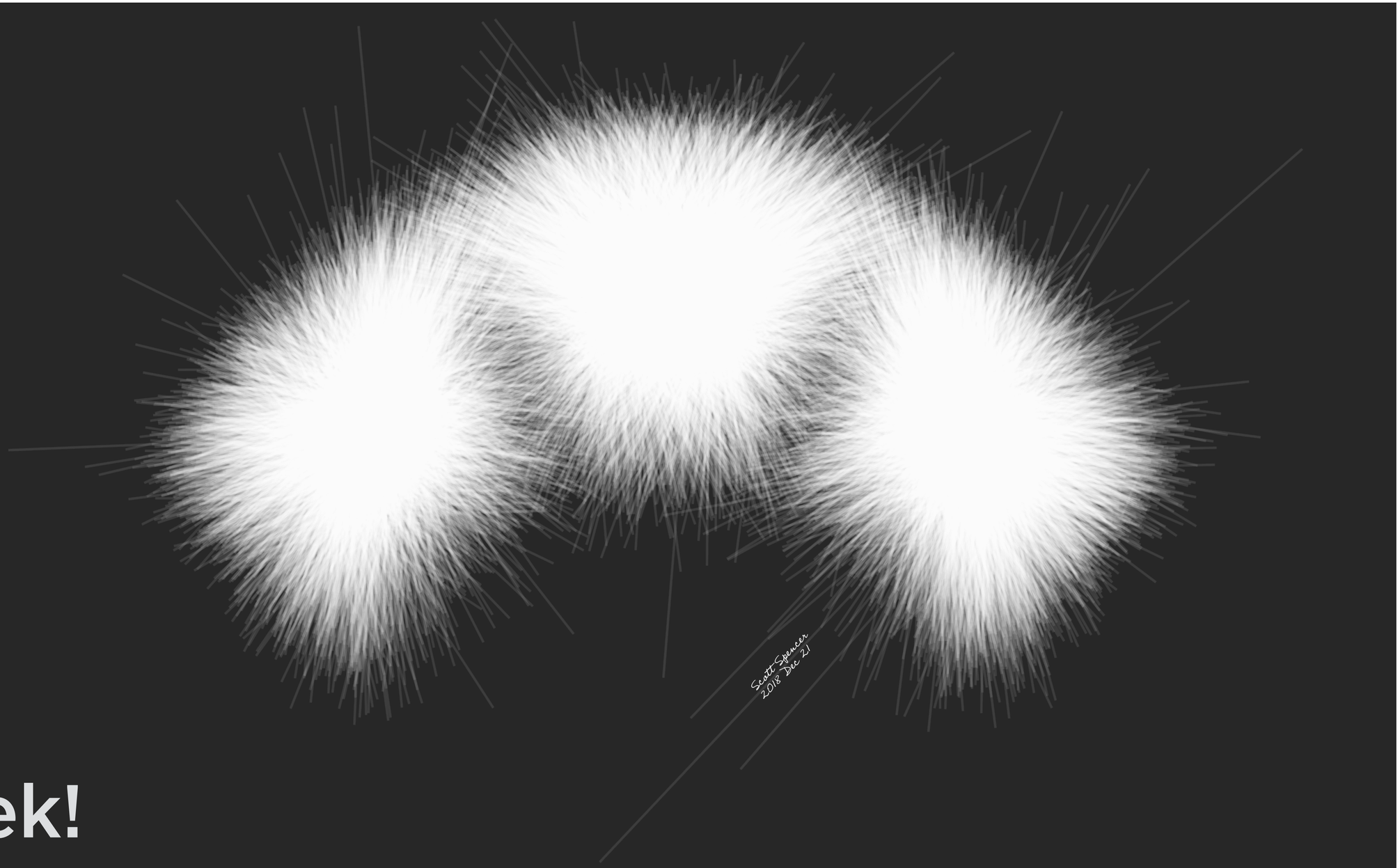
Consider how some of these visual components of writing assist your audience.

Kay, Matthew. "Figures." www.mjskay.com. N.p., Aug. 2015. Web. 28 Mar. 2019.

Consider how information is visually connected.

Continuing feedback guides your course:

Got it!	What information from this lecture did you find particularly interesting or helpful?
Say what?	What may be an additional idea useful for scoping or describing data analytics projects that may complement those we discussed in class?
Let's get on with it.	In what domains beyond your current job or studies are you interested in applying analytics?



See you next week!

Spencer, S. *Baseball Abstract 01*. 21 December 2018. Web.
https://ssp3nc3r.github.io/post/baseball-abstract-01/