

# Storytelling with Data

## Module 4: Principles of persuasion and brief proposals

**Scott Spencer**  
Faculty and Lecturer  
Columbia University

# What we've discussed so far

Knaflic's  
*Storytelling with data*

Understand data context  
Choice of appropriate visual display  
Eliminate clutter  
Focus audience attention  
Think like a designer  
Tell a story

Technical audience, employee  
Example 250-word memo  
**Dodgers**, game decisions should optimize expectations  
background > goals > problem > data > method > impact

Adapt to your audience  
Doumont's *Trees, Maps, Theorems*  
Messages, not just information

Identifying events,  
**Citi Bike**, user behaviors  
example case studies  
Measurements of events and behaviors

be concise, every word tell  
Strunk & White's  
*The Elements of style*  
overstatements diminish credibility

TL;DR

Spencer's  
*Scoping a data analytics project*  
decisions > goals and actions >  
methods > data

**step into their shoes!**  
**CAO, CMO, CEO**

beyond the minimum

Columbia University  
*The Writing Center*

What problem is to be solved?  
Is it important?  
Does it have impact?  
Do data play a role in solving the problem?  
Are the right data available?  
Is the organization ready to tackle the problem and take actions from insights?

background > goals > problem > method > impact  
Example **Jakarta** proposal  
*Improving traffic safety through video analysis*  
Technical audience, not employee

complexity last  
Booth's  
*Revising style*

**old before new**

ING's  
*General audience*  
*The Next Rembrandt*

# Getting to storytelling with data



# Agenda

Next deliverable – brief proposal

Today's objectives

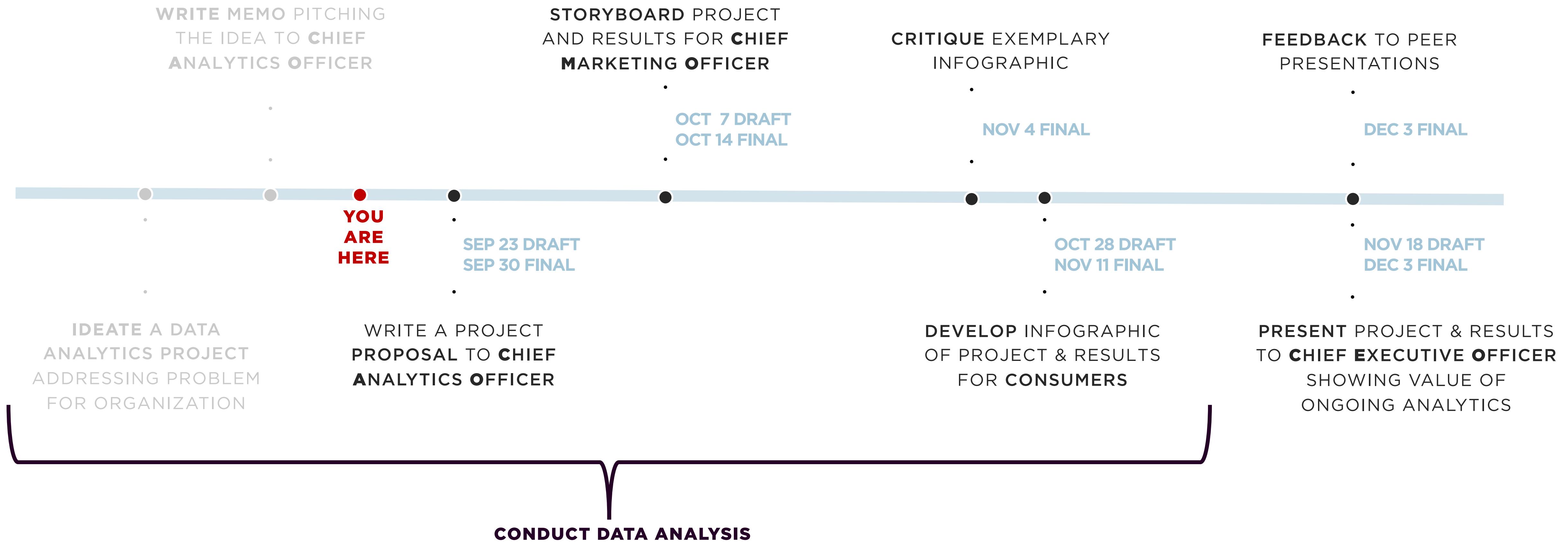
Perspectives on persuasion

Comparison, metaphor, patterns

# Next deliverable

# Upcoming deliverable

**750-word brief proposal** – Write a brief proposal to **CAO** detailing your proposed analytics project. Consider background context, problem, data, solution, and impact. At this point you should have data to start an analysis.



# **Example *draft* brief proposal**

# Example draft proposal. Constraint—750 words or less in main body.

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

To: Scott Powers, Senior Baseball Analyst, Los Angeles Dodgers  
From: Scott Spencer, Faculty and Lecturer, Columbia University

14 February 2019

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

### 1 Our current analyses do not optimize expected wins

Seven terabytes of uncompressed data generated per game overshadow the lack of situational data needed for decision-making that maximizes expected utility. Consider that pitchers, on average, only face 10 percent of major league batters regardless of game state; the reverse is true, too. Or when deciding whether a base runner should attempt to steal against a specific pitcher and catcher in a state of play, say, we are lucky to have any data. Common analyses and heuristics for these situations are inadequate: they not only overfit the data (if any exist), but also offer no manner of estimating changes in probabilities for maximizing *expected utility* (winning the game).

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

### 2 Modeling probabilities for steal success illustrates a broader benefit

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

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

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

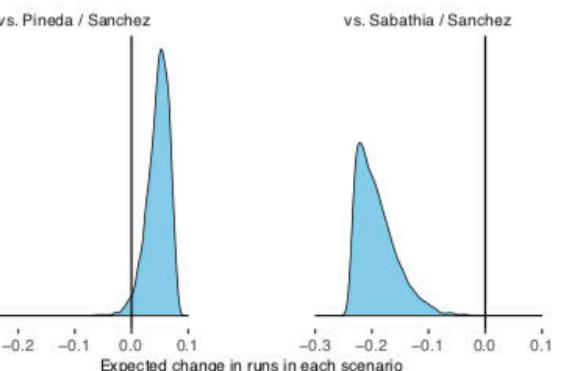


Figure 1. Of the two scenarios, Cain should only attempt to steal against the Sanchez–Pineda duo.

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

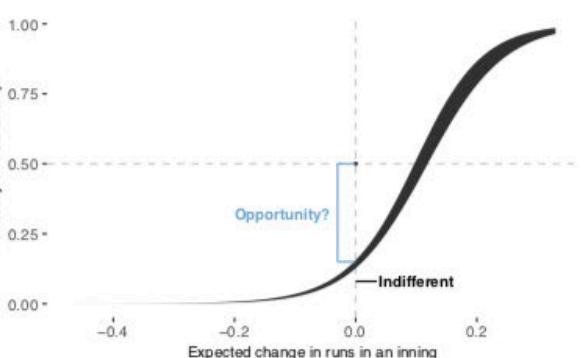


Figure 2. When the change in expected runs is zero, managers should be indifferent to attempted steals, saying go half the time.

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

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

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

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

### 4 For accuracy, compare model results to betting market odds

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

### 5 Conclusion

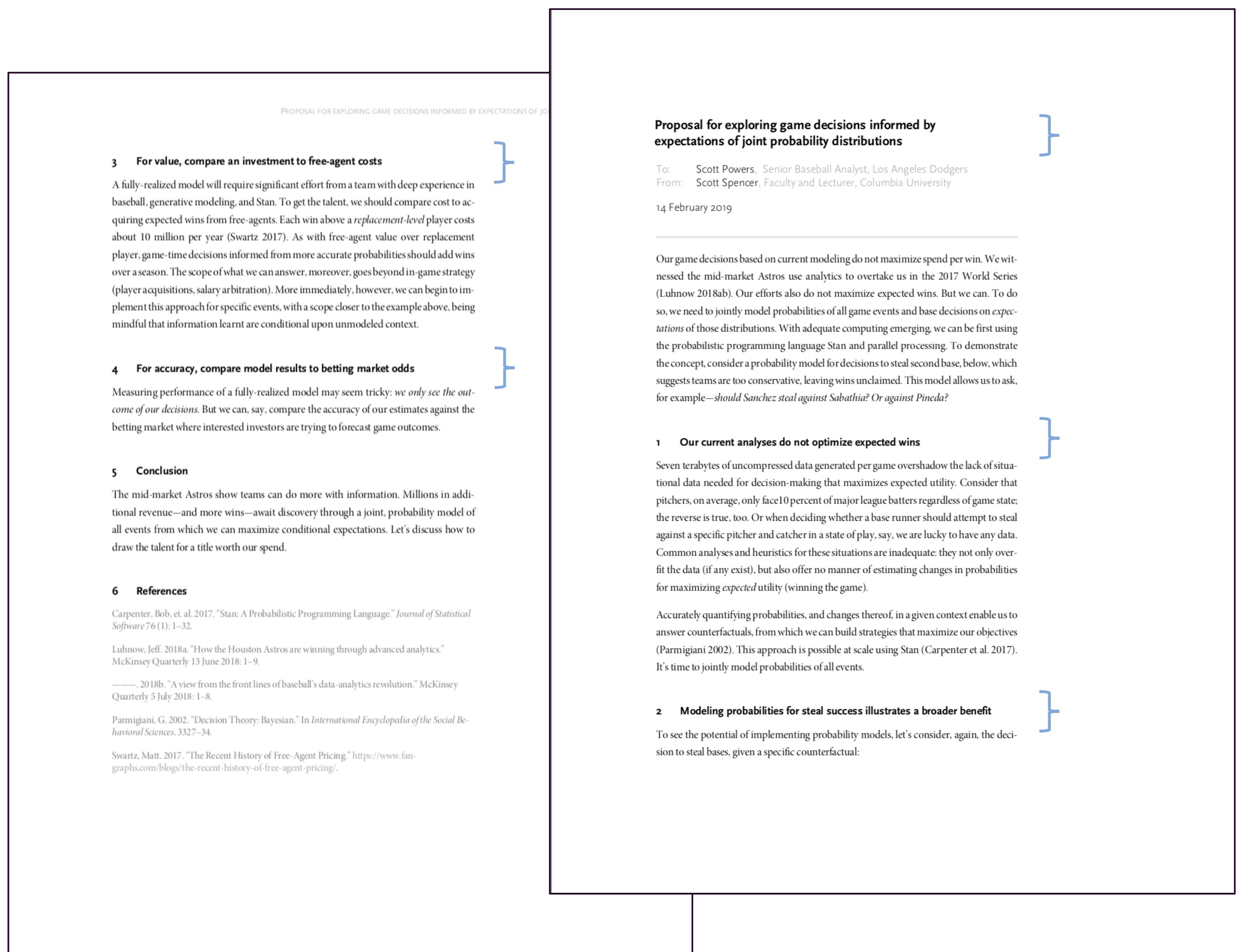
The mid-market Astros show teams can do more with information. Millions in additional revenue—and more wins—await discovery through a joint, probability model of all events from which we can maximize conditional expectations. Let's discuss how to draw the talent for a title worth our spend.

### 6 References

- Carpenter, Bob, et al. 2017. "Stan: A Probabilistic Programming Language." *Journal of Statistical Software* 76 (1): 1–32.  
Luhnow, Jeff. 2018a. "How the Houston Astros are winning through advanced analytics." *McKinsey Quarterly* 13 June 2018: 1–9.  
———. 2018b. "A view from the front lines of baseball's data-analytics revolution." *McKinsey Quarterly* 5 July 2018: 1–8.  
Parmigiani, G. 2002. "Decision Theory: Bayesian." In *International Encyclopedia of the Social Behavioral Sciences*, 3327–34.  
Swartz, Matt. 2017. "The Recent History of Free-Agent Pricing." <https://www.fangraphs.com/blogs/the-recent-history-of-free-agent-pricing/>.

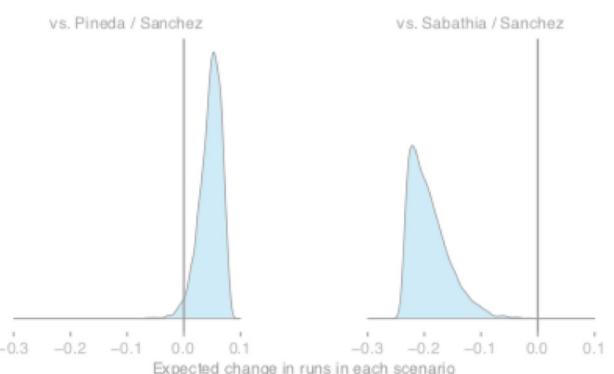
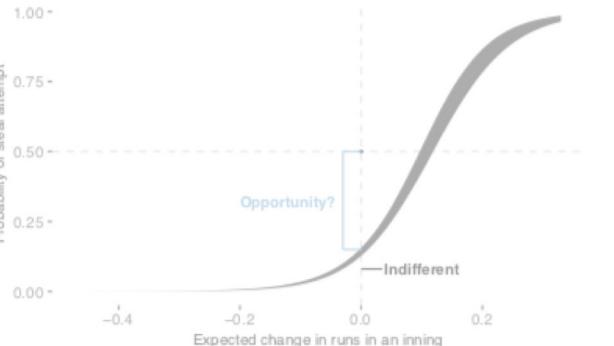
Readability Statistics	
<b>Counts</b>	
Words	720
Characters	3,997
Paragraphs	16
Sentences	35
<b>Averages</b>	
Sentences per Paragraph	4.3
Words per Sentence	18.1
Characters per Word	5.3
<b>Readability</b>	
Flesch Reading Ease	33.2
Flesch-Kincaid Grade Level	13
Passive Sentences	0%

# Example draft proposal. Messaging—Messages first, not just information. Details follow.



Doumont, Jean-Luc. *Trees, Maps, and Theorems*. Principiæ, 2009.

# Example draft proposal. Typography—Grid: two columns with gutter.

<p><b>Proposal for exploring game decisions informed by expectations of joint probability distributions</b></p> <p>To: Scott Powers, Senior Baseball Analyst, Los Angeles Dodgers From: Scott Spencer, Faculty and Lecturer, Columbia University</p> <p>14 February 2019</p> <p>Our game decisions based on current modeling do not maximize spend per win. We witnessed the mid-market Astros use analytics to overtake us in the 2017 World Series (Luhnow 2018ab). Our efforts also do not maximize expected wins. But we can. To do so, we need to jointly model probabilities of all game events and base decisions on <i>expectations</i> of those distributions. With adequate computing emerging, we can be first using the probabilistic programming language Stan and parallel processing. To demonstrate the concept, consider a probability model for decisions to steal second base, below, which suggests teams are too conservative, leaving wins unclaimed. This model allows us to ask, for example—should Sanchez steal against Sabathia? Or against Pineda?</p> <p><b>1 Our current analyses do not optimize expected wins</b></p> <p>Seven terabytes of uncompressed data generated per game overshadow the lack of situational data needed for decision-making that maximizes expected utility. Consider that pitchers, on average, only face 10 percent of major league batters regardless of game state; the reverse is true, too. Or when deciding whether a base runner should attempt to steal against a specific pitcher and catcher in a state of play, say, we are lucky to have any data. Common analyses and heuristics for these situations are inadequate—they not only over-fit the data (if any exist), but also offer no manner of estimating changes in probabilities for maximizing <i>expected</i> utility (winning the game).</p> <p>Accurately quantifying probabilities, and changes thereof, in a given context enable us to answer counterfactuals, from which we can build strategies that maximize our objectives (Parmigiani 2002). This approach is possible at scale using Stan (Carpenter et al. 2017). It's time to jointly model probabilities of all events.</p> <p><b>2 Modeling probabilities for steal success illustrates a broader benefit</b></p> <p>To see the potential of implementing probability models, let's consider, again, the decision to steal bases, given a specific counterfactual:</p>	<p>PROPOSAL FOR EXPLORING GAME DECISIONS INFORMED BY EXPECTATIONS OF JOINT PROBABILITY DISTRIBUTIONS 2</p> <p>In a game against New York Yankees, should Milwaukee Brewers's Lorenzo Cain attempt to steal second base with no one else on base and two outs before the seventh inning, against Gary Sanchez as catcher and Michael Pineda as pitcher? What if against Sanchez and CC Sabathia as pitcher?</p> <p>More specifically, how can we know the <i>expectation</i> that Cain's attempt in each situation increases the probability of expected runs that inning and by how much? Using Stan, I've coded a generative model that along with play outcomes considers various information (runner foot-speed, catcher pop-time) and player characteristics, like pitcher handedness. With the model, we have an answer that also shows the uncertainty. Given 2017 data, this model suggests Cain should steal against Pineda, not Sabathia:</p>  <p>Figure 1. Of the two scenarios, Cain should only attempt to steal against the Sanchez-Pineda duo.</p> <p>Notably, we get these expectations without multiple trials of either scenario. More generally, this model suggests that on average team managers are too conservative, leaving runs unrealized:</p>  <p>Figure 2. When the change in expected runs is zero, managers should be indifferent to attempted steals, saying go half the time. The black band represents the range of variation across managers' decisions. At the intersection of indifference, managers tend to say steal only 10 percent of the time, leaving opportunity.</p> <p>The above is but one example of a more general approach that weighs probabilities of all possible outcomes to maximize expected utility. With broad implementation—jointly modeling the conditional probabilities of all relevant events—we can optimize decisions.</p>	

Müller-Brockmann, Josef. *Grid Systems in Graphic Design*. ARTHUR NIGGLI LTD., 1996. Print.

Tondreau, Beth. *Layout Essentials*. Rockport, 2008. Print.

Samara, Timothy. *Making and Breaking the Grid*. Second. Rockport, 2017. Print.

# Example *draft* proposal. Typography—Layout.

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

**Average line length: 84 characters with spaces  
Butterick recommended 45-90**

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

1 Our current analyses do not optimize expected wins

Seven terabytes of uncompressed data generated per game overshadow the lack of situational data needed for decision-making that maximizes expected utility. Consider that pitchers, on average, only face 10 percent of major league batters regardless of game state; the reverse is true, too. Or when deciding whether a base runner should attempt to steal against a specific pitcher and catcher in a state of play, say, we are lucky to have any data. Common analyses and heuristics for these situations are inadequate: they not only over-fit the data (if any exist), but also offer no manner of estimating changes in probabilities for maximizing *expected* utility (winning the game).

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

2 Modeling probabilities for steal success illustrates a broader benefit

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

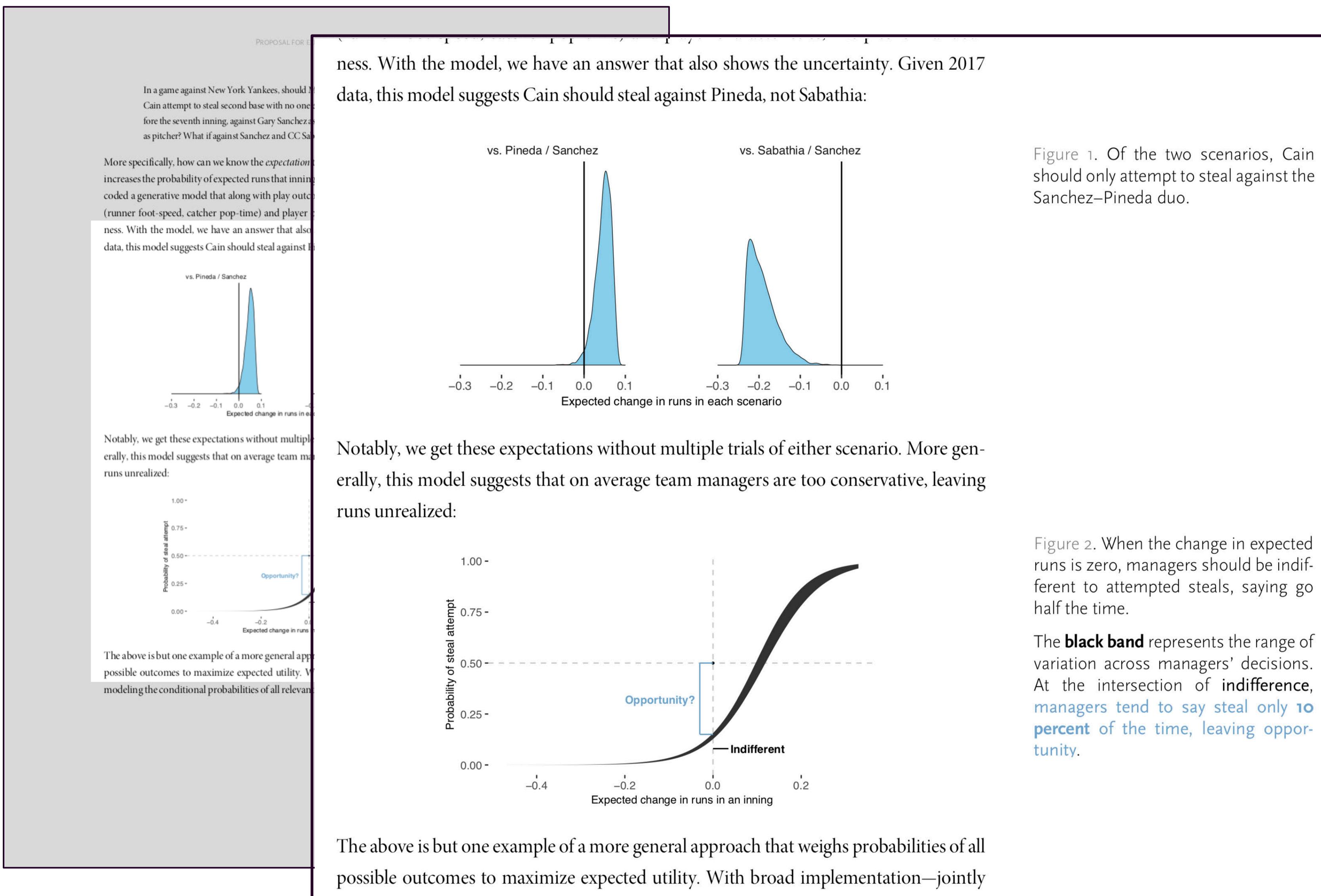
Butterick, Matthew. *Butterick's Practical Typography*. practicaltypography.com. N.p., 2018. Web. 8 Sept. 2018.

See also:

Bringhurst, Robert. *The Elements of Typographic Style*. Fourth. Hartley & Marks, 2012. Print.

Rutter, Richard. *Web Typography*. Ampersand Type, 2017. Print.

# Example draft proposal. Graphics as paragraphs; annotating, linking words to data in graphics.



Tufte, Edward R. *The Visual Display of Quantitative Information*. Second. Graphics Press, 2001.

Kay, Matthew. *Figures*. [www.mjskay.com](http://www.mjskay.com). Aug. 2015. Web. 28 Mar. 2019.

Riche, Nathalie Henry et al. Ch. 9, *Communicating Data to an Audience*, in *Data-Driven Storytelling*. CRC Press, 2018.

# Today's Objectives

# Objectives

1

Explain the role of persuasion in getting buy-in for analytics projects.

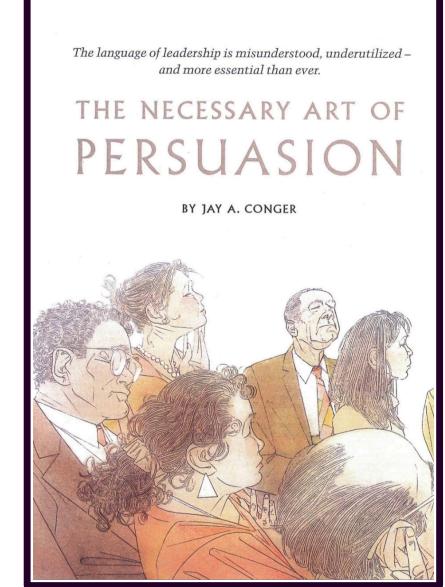
2

Explain the role of persuasion in implementing analytic insights.

3

Employ tools and techniques taught in class to persuade technical and non-technical audiences.

# Perspectives on persuasion



# Necessary art of persuasion

Conger

Conger is an executive educator, coach, and program designer who teaches leadership to companies and individuals.

## Persuading involves four steps

### Establish credibility

First assess your credibility—your knowledge about the strategy, product, or change proposed—by **self reflection** and **asking others**. **Fill in gaps**: gain knowledge; cite outside sources; demonstrate the proposal by starting smaller.

### Find common ground

Study the issues with colleagues; **think through their arguments, evidence, and perspectives**. Address or include them, making your proposal something shared.

### Combine evidence with story, metaphor

Numerical evidence should be **supplemented with** “examples, stories, metaphors, and analogies” to enliven your proposal. This is particularly helpful when presenting **comparable situations** to the one under discussion.

### Connect emotionally

Understand how your audience feels on the issues, and recognize—even share—their. Empathize.

# Narrative Design Patterns for Data-Driven Storytelling

## Riche, co-editors

The editors are researchers and professors with focuses on human-computer interaction and information visualization.



## Classical devices of rhetoric

The classical devices of rhetoric involve **logos** (reason, word), **ethos** (character, ideal), and **pathos** (experience, emotion).

Rhetoric in data-driven stories aim for truth, connect

Though we believe the ultimate **goal of data-driven storytelling is to communicate truth** (most closely to logos), there are traces of both **pathos**, and **ethos in every story, which help connect the narrator with the audience**.

## Patterns for argumentation

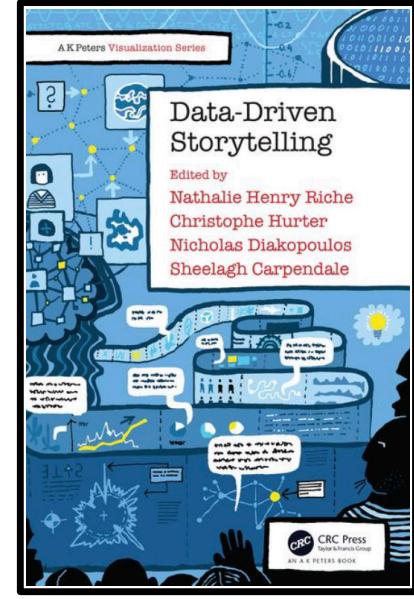
**Argumentation** is the action or process of reasoning systematically in support of an idea, action, or theory.

**Patterns** for argumentation serve the intent of persuading and convincing audiences.

# Narrative Design Patterns for Data-Driven Storytelling

*Riche, co-editors*

The editors are researchers and professors with focuses on human-computer interaction and information visualization.



## Compare

Comparison allows the narrator to make the point about equality of both data sets, to explicitly highlight differences and similarities, or to give reasons for their difference.

**Please approve the hire of 2 FTEs**

to backfill those who quit in the past year

Ticket volume over time

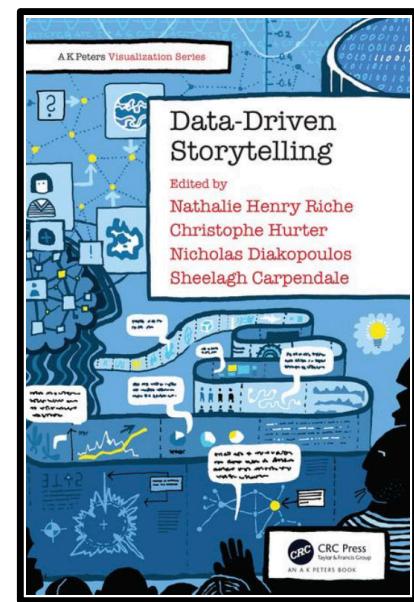


Data source: XYZ Dashboard, as of 12/31/2014 | A detailed analysis on tickets processed per person and time to resolve issues was undertaken to inform this request and can be provided if needed.

# Narrative Design Patterns for Data-Driven Storytelling

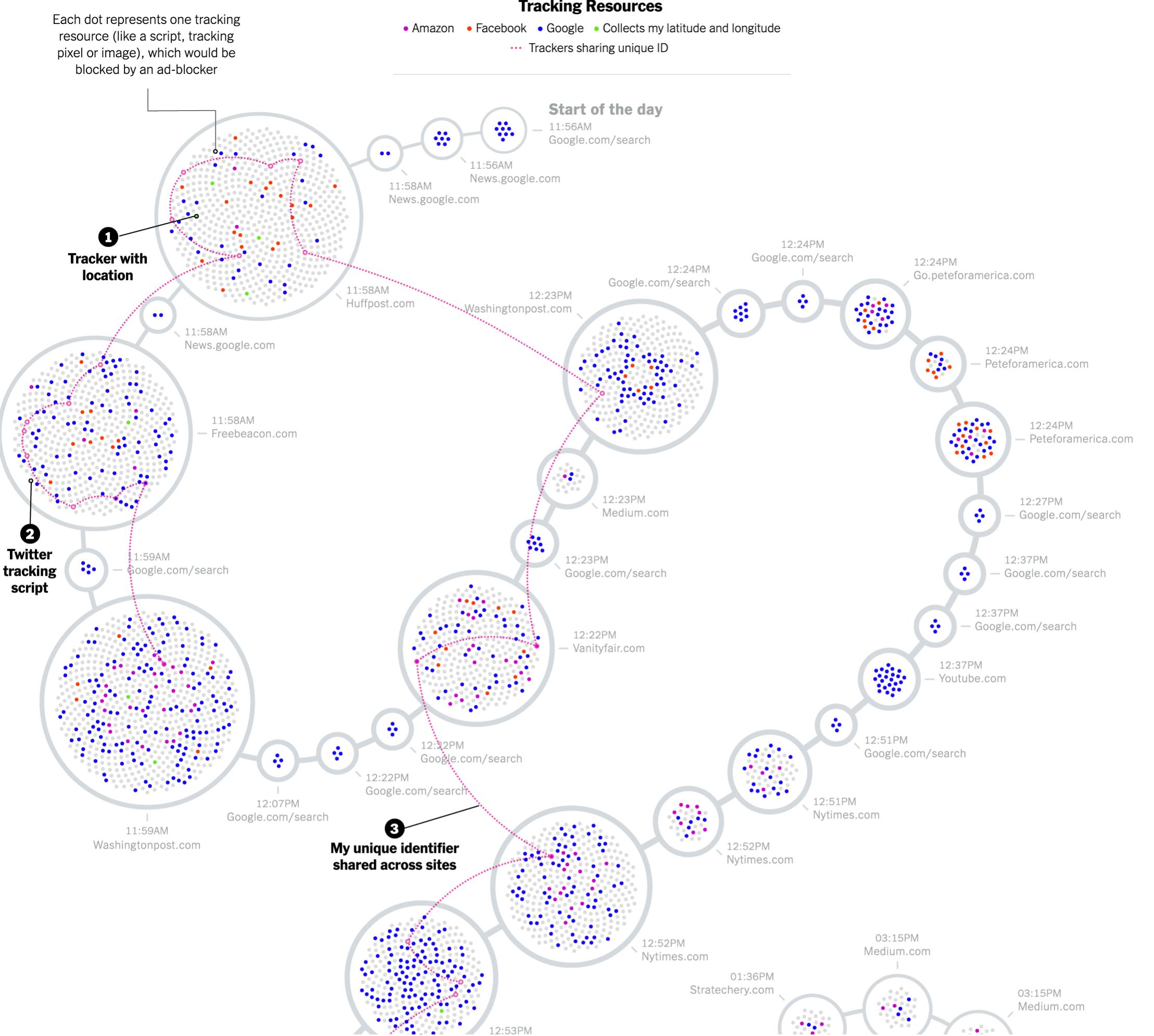
Riche, co-editors

The editors are researchers and professors with focuses on human-computer interaction and information visualization.



## Concretize

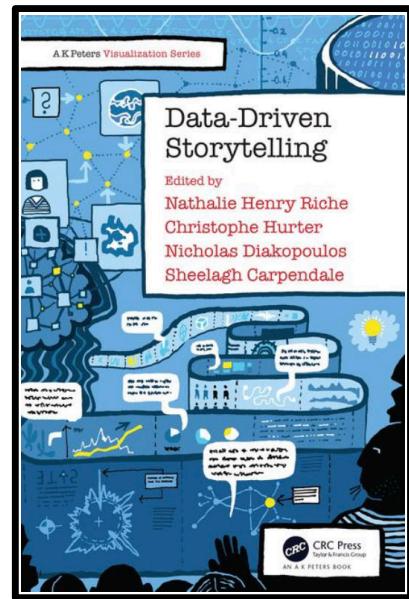
Shows abstract concepts with concrete objects. Concretization usually implies that each data point is represented by an individual visual object (e.g., a point or shape), making them less abstract than aggregated statistics.



# Narrative Design Patterns for Data-Driven Storytelling

Riche, co-editors

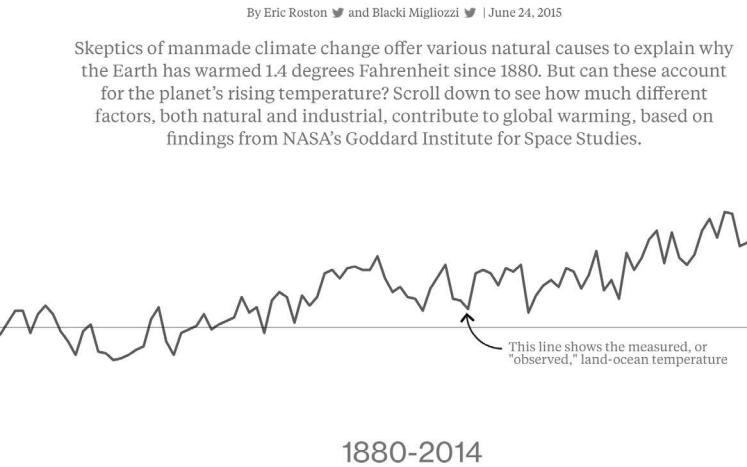
The editors are researchers and professors with focuses on human-computer interaction and information visualization.



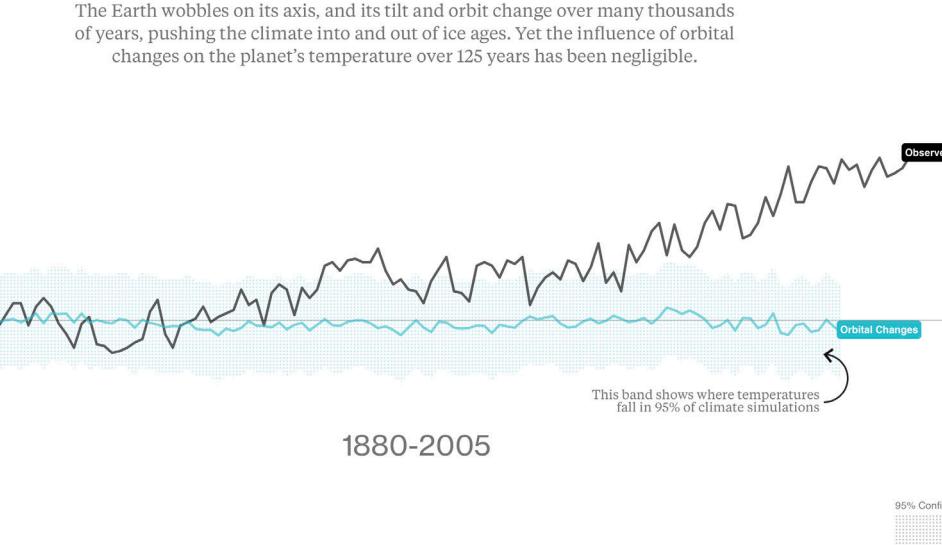
## Repetition

Repetition can increase a message's importance and memorability, and can help tie together different arguments about a given data set. Repetition can be employed as a means to search for an answer in the data.

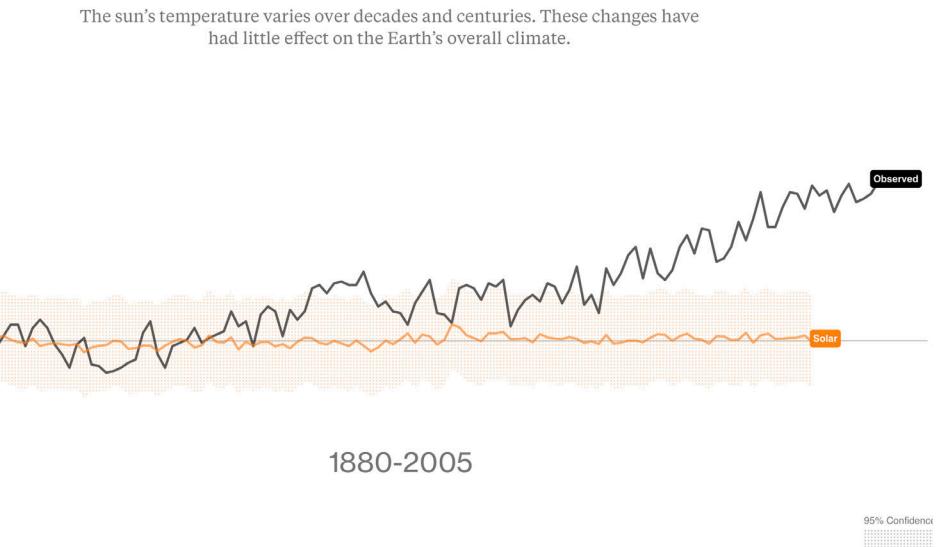
### What's Really Warming the World?



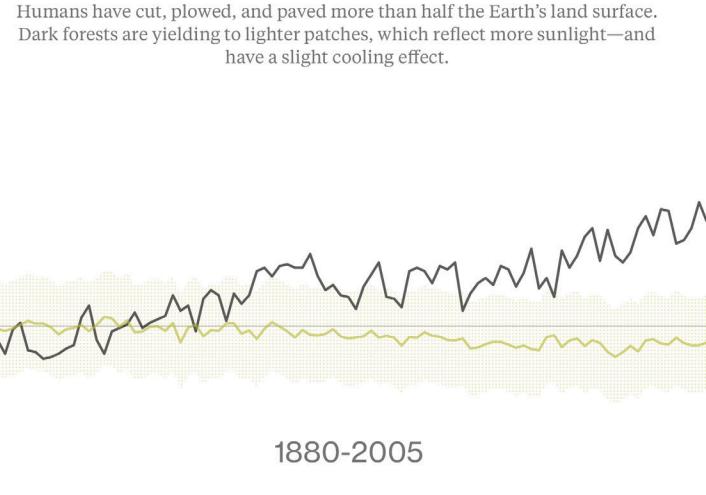
### Is It the Earth's Orbit?



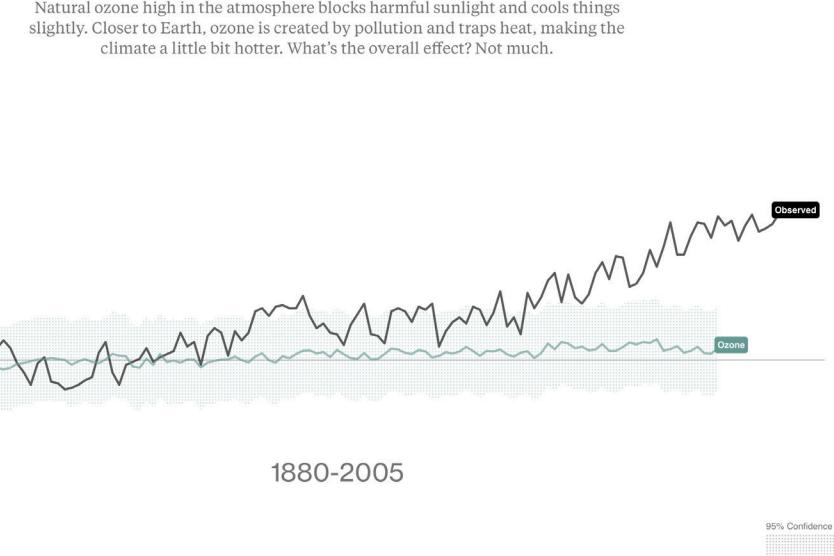
### Is It the Sun?



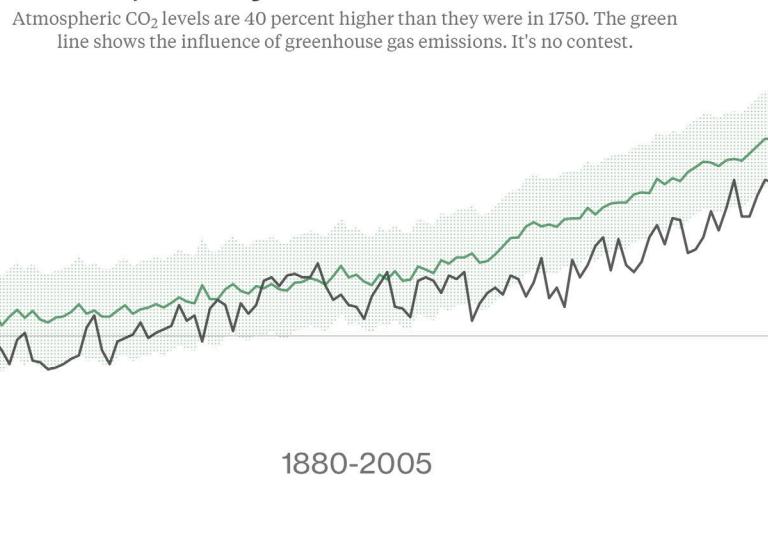
### So If It's Not Nature, Is It Deforestation?



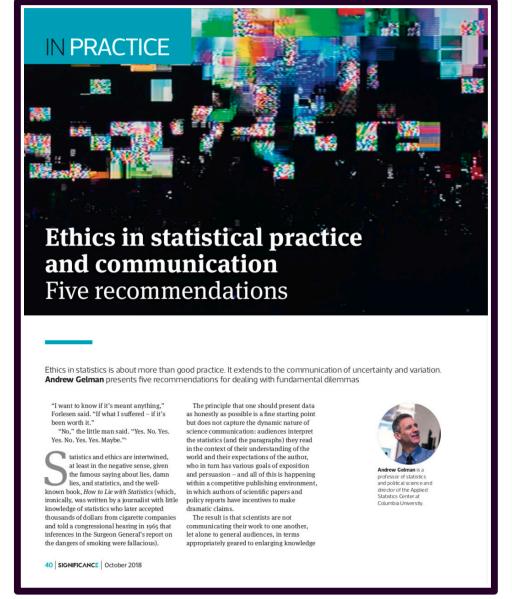
### Or Ozone Pollution?



### No, It Really Is Greenhouse Gases.



# Statistical persuasion



# Ethics in Statistical Practice and Communication

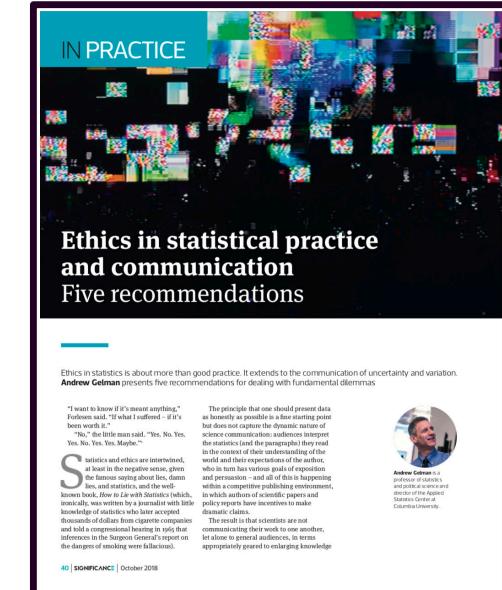
Gelman

Professor of Statistics and Political Science at Columbia University, he is known widely for his work in Bayesian statistics, and has authored several textbooks, including *Teaching Statistics*, and *Bayesian Data Analysis*.

## Why statistics?

Consider this paradox: statistics is the science of uncertainty and variation, but data-based claims in the scientific literature tend to be stated deterministically (e.g. “We have discovered ... the effect of X on Y is ... hypothesis H is rejected”).

Is statistical communication about exploration and discovery of the unexpected, or is it about making a persuasive, data-based case to back up an argument?



# Ethics in Statistical Practice and Communication

Gelman

Professor of Statistics and Political Science at Columbia University, he is known widely for his work in Bayesian statistics, and has authored several textbooks, including *Teaching Statistics*, and *Bayesian Data Analysis*.

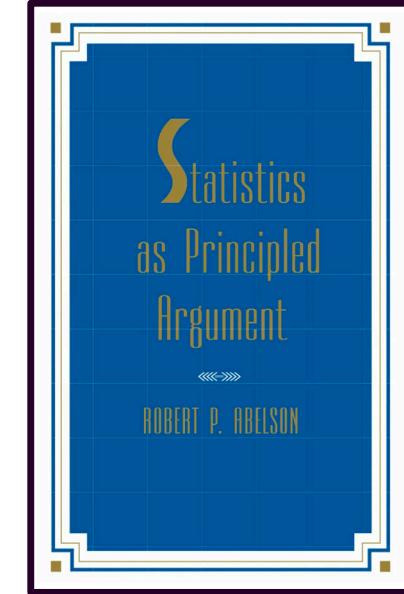
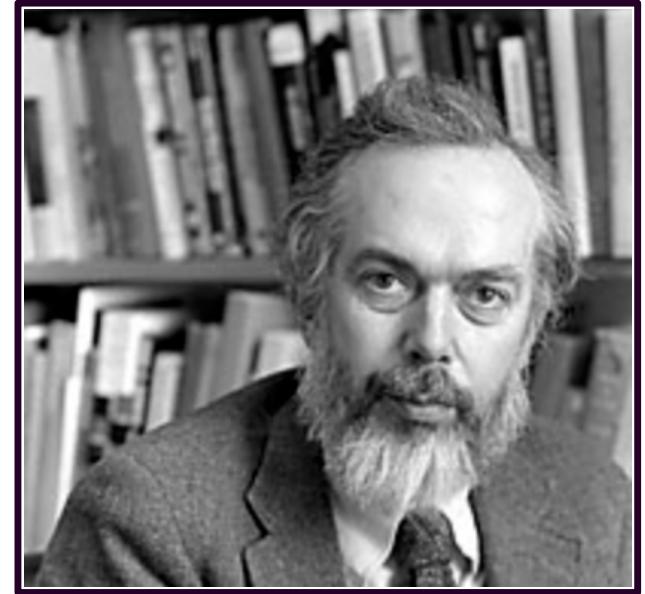
## Exploring and persuading

The answer to this question is necessarily each at different times, and sometimes both at the same time.

Just as you write in part in order to figure out what you are trying to say, so you do statistics not just to learn from data but also to learn what you can learn from data, and to decide how to gather future data to help resolve key uncertainties.

Traditional advice on statistics and ethics focuses on professional integrity, accountability, and responsibility to collaborators and research subjects.

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



# Statistics as principled argument

*Abelson*

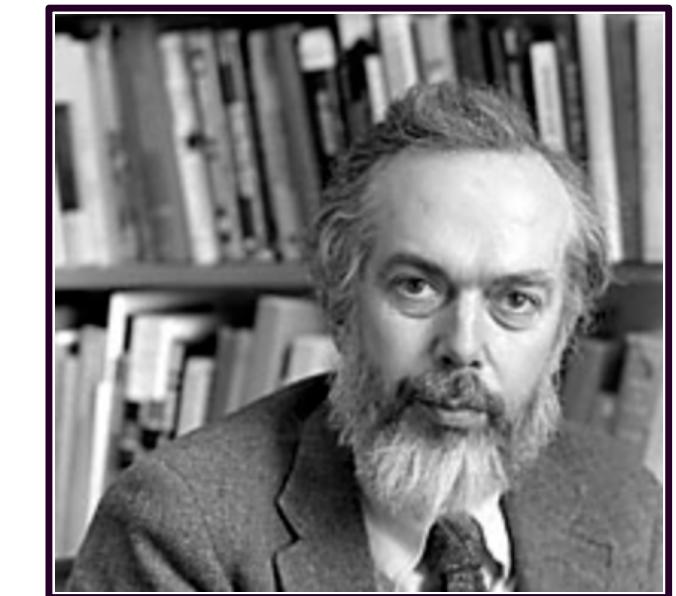
Educated at MIT and Princeton, the late professor of psychology and political science taught at Yale 42 years, consulted for NBC, and was an analyst for three presidential campaigns.

## The purpose of statistics is persuasion

The purpose of statistics is to organize a useful argument from quantitative evidence, using a form of principled rhetoric ... that conveys an interesting and credible point.

To make statistical arguments, it helps to wear different hats

His “image of the ideal statistician, already conceived as a good (but honest!) lawyer and a good storyteller, also includes the virtues of a good detective.”



# Statistics as principled argument

## Abelson

Educated at MIT and Princeton, the late professor of psychology and political science taught at Yale 42 years, consulted for NBC, and was an analyst for three presidential campaigns.

### Comparison gives 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."

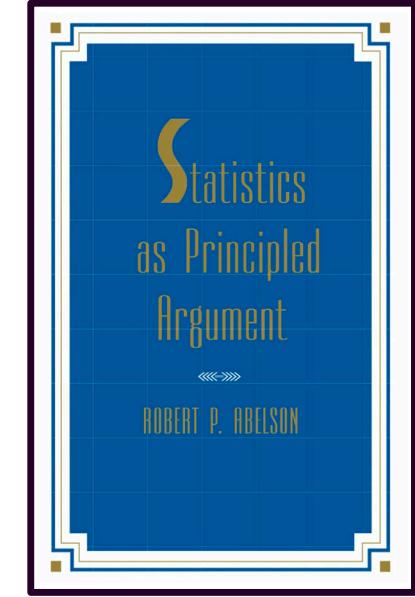
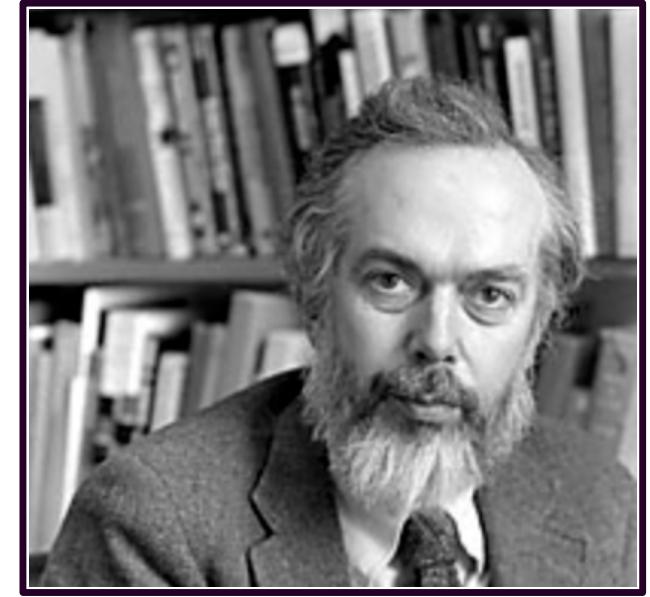
"The average life expectancy of famous orchestral conductors is 73.4 years."

### Why is this important? How unusual is this?

### Consider standards of comparison

Should we compare with orchestra *players*? With *non-famous* conductors, with the *public*? With other *males* in the United States, whose average life expectancy was 68.5 at the time of the study?

With other males who have already reached the age of 32, the average age of appointment to a first conducting post, almost all of whom are male? This group's average life expectancy was 72.0.



# Statistics as principled argument

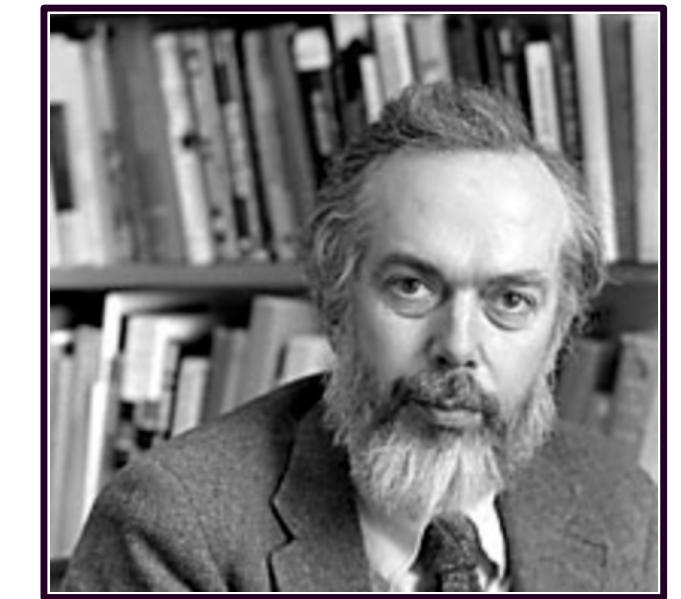
*Abelson*

Educated at MIT and Princeton, the late professor of psychology and political science taught at Yale 42 years, consulted for NBC, and was an analyst for three presidential campaigns.

## Elements of statistical persuasion

Several properties of data, and its analysis and presentation, govern its persuasive force.

- M**agnitude of effects
- A**rticulation of results
- G**enerality of effects
- I**nterestingness of argument
- C**redibility of argument

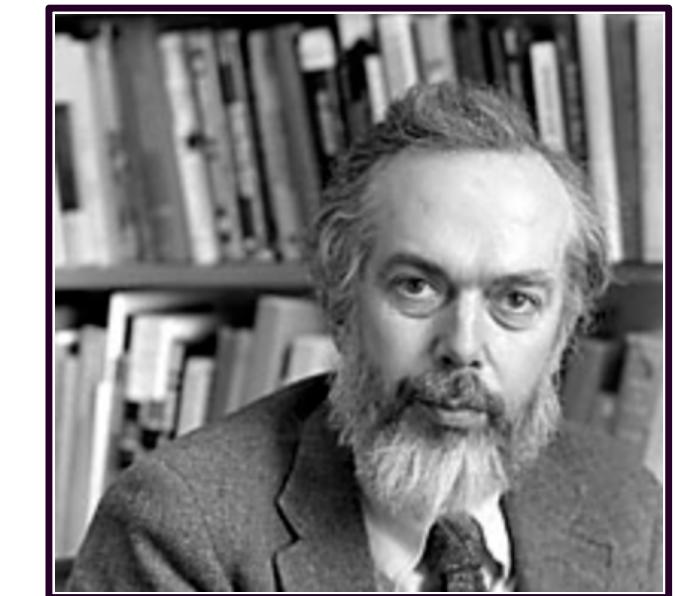


# Statistics as principled argument

## Abelson

Educated at MIT and Princeton, the late professor of psychology and political science taught at Yale 42 years, consulted for NBC, and was an analyst for three presidential campaigns.

- M **Magnitude of effects.** The strength of a statistical argument is enhanced in accord with the quantitative magnitude of support for its qualitative claim. Consider describing effect sizes like the difference between means, not dichotomous tests.
- A **Articulation of results.** The degree of comprehensible detail in which conclusions are phrased. This is a form of specificity. We want to honestly describe and frame our results to maximize clarity (minimizing exceptions or limitations to the result) and parsimony (focusing on consistent, connected claims).
- G **Generality of effects.** This is the breadth of applicability of the conclusions. Over what context can the results be replicated?
- I **Interestingness of argument.** For a statistical story to be theoretically interesting, it must have the potential, through empirical analysis, to change what people believe about an important issue.
- C **Credibility of argument.** Refers to the believability of a research claim, and requires both methodological soundness and theoretical coherence.



# Statistics as principled argument

## Abelson

Educated at MIT and Princeton, the late professor of psychology and political science taught at Yale 42 years, consulted for NBC, and was an analyst for three presidential campaigns.

MAGIC

p-values  
say little,  
can mislead

```
> y <- rnorm(n = 100000, mean = 0, sd = 1)
> x <- rnorm(n = 100000, mean = 0, sd = 1)
> model_fit <- lm(y ~ x)
> summary(model_fit)
```

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

Residuals:

Min	1Q	Median	3Q	Max
-4.6381	-0.6755	0.0064	0.6705	4.0234

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	0.001911	0.003157	0.606	0.5448
x	-0.008707	0.003149	-2.765	0.0057 **
---				

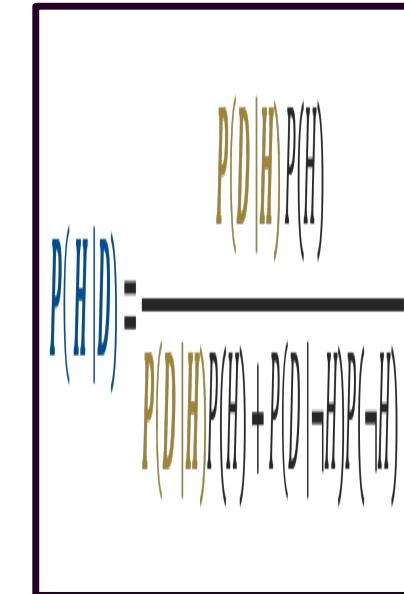
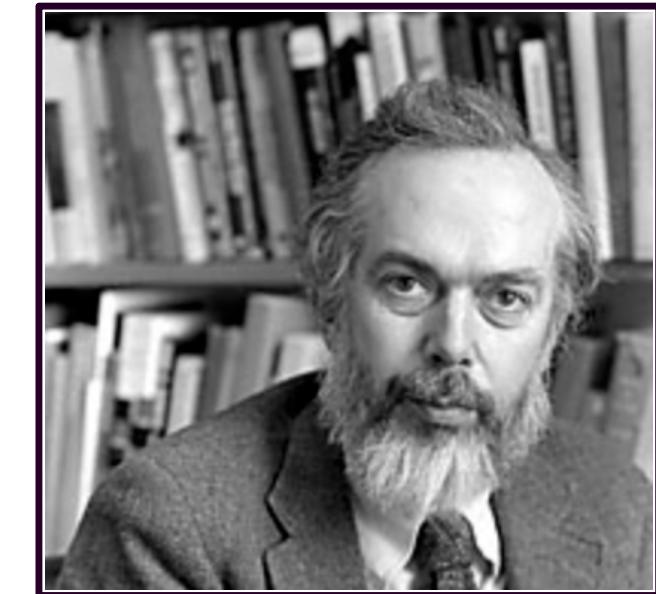
**Signif. codes:** 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9982 on 99998 degrees of freedom  
Multiple R-squared: 7.643e-05, Adjusted R-squared: 6.643e-05  
F-statistic: 7.644 on 1 and 99998 DF, p-value: 0.005699

**Magnitude of effects.** The strength of a statistical argument is enhanced in accord with the quantitative magnitude of support for its qualitative claim. **Consider describing effect sizes like the difference between means, not dichotomous tests.**

The information yield from null hypothesis tests is ordinarily quite modest, because **all one carries away is a possibly misleading accept-reject decision.**

**p-value < 0.01**  
**Stars, significant !?**



# Statistics as principled argument

Abelson

Educated at MIT and Princeton, the late professor of psychology and political science taught at Yale 42 years, consulted for NBC, and was an analyst for three presidential campaigns.

A p-value less than 0.01 is **not**:

Instead, it means:  
 $P(D | H)$

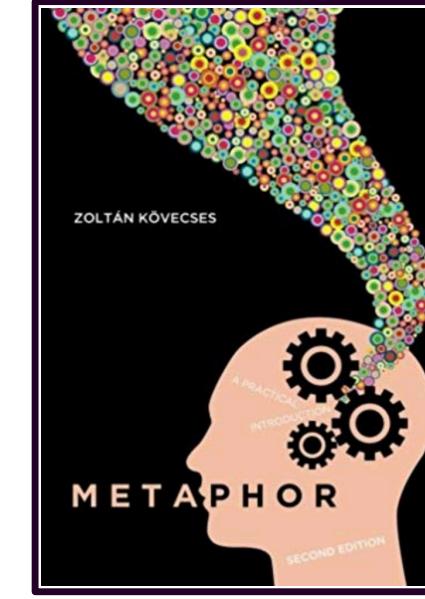
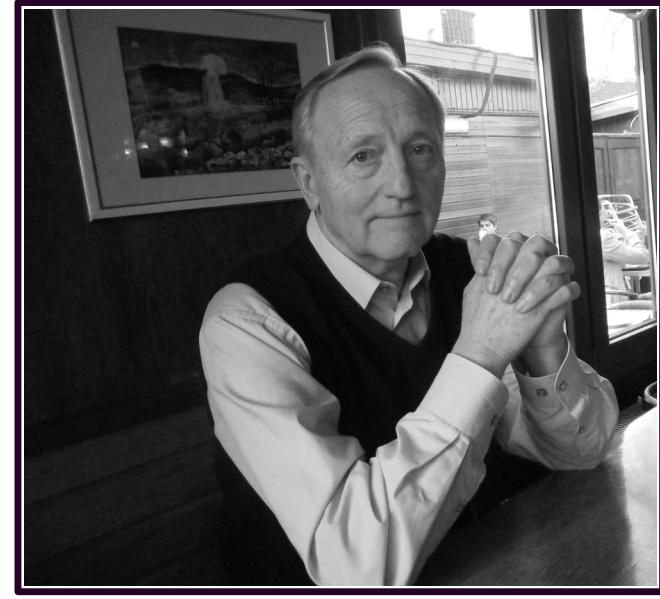
Having observed the data, the probability that the null hypothesis is true is less than one in a hundred.

If it were true that there were no systematic difference between the means in the populations from which the samples came, then the probability that the observed means would have been as different as they were, or more different, is less than one in a hundred.

This being strong grounds for doubting the viability of the null hypothesis, the null hypothesis is rejected.

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

# Comparing abstract to familiar

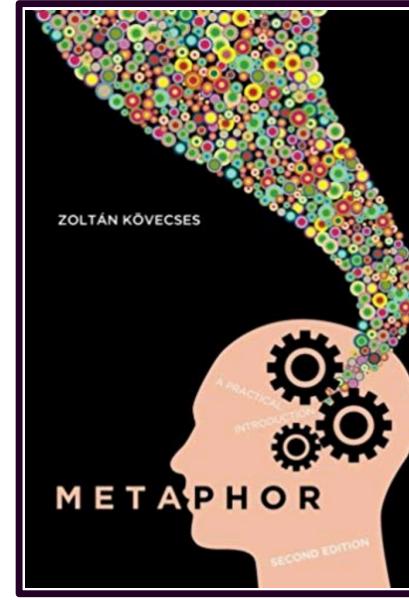
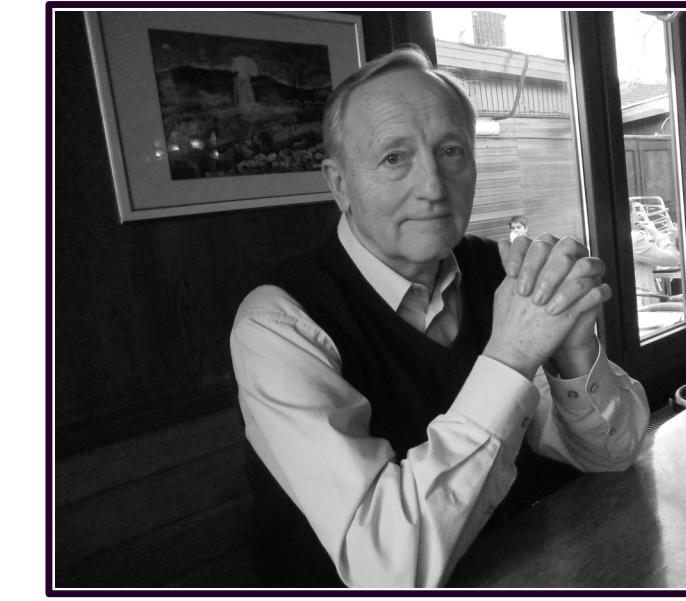


# Metaphor: a practical introduction

## Kővecses

He is professor of linguistics at Eötvös Loránd University, Budapest. He researches language and conceptualization of emotions, cross-cultural variation in metaphor, and the issue of the relationship between language, mind, and culture.

Metaphor adds to persuasiveness by **reforming abstract concepts into something more familiar to our senses**, signaling particular aspects of importance, memorializing the concept, or providing coherence throughout a writing.



# Metaphor: a practical introduction

## Kővecses

He is professor of linguistics at Eötvös Loránd University, Budapest. He researches language and conceptualization of emotions, cross-cultural variation in metaphor, and the issue of the relationship between language, mind, and culture.

### Mapping

Source Domain > Target Domain

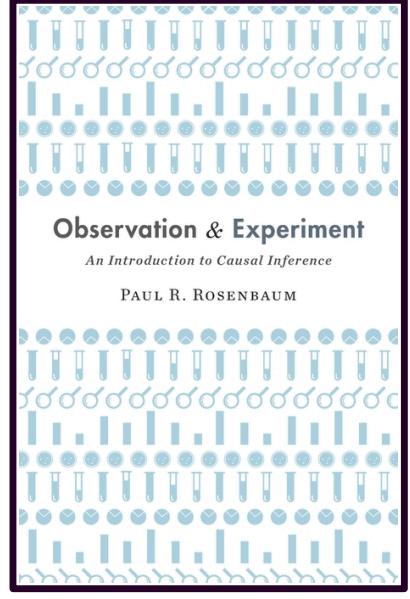
### Target domains

The abstract concepts we need help explaining

### Common source domains

- Human body
- Animals
- Plants
- Buildings and constructions
- Machines and tools
- Games and Sport
- Money
- Cooking and food
- Heat and cold
- Light and darkness
- Movement and direction

**Example:** uses poetry about travel (source domain) to explain the distinction between covariate and outcome (target domain):



# Observation & Experiment

## Rosenbaum

He is Professor of Statistics at the Wharton School and a Senior Fellow of the Leonard Davis Institute of Health Economics, University of Pennsylvania. His book epitomizes the idea that “the most important ideas in statistics can be clearly explained in plain English, with little or no math.”

If we accurately measure an outcome, we see one of its two potential values: the value that occurs under the treatment the patient actually received. **We can never see the outcome a patient would have exhibited under the treatment the patient did not receive.** . . . Perhaps the distinction between covariate and outcome is most vivid, most palpable, in Robert Frost’s poem “The Road Not Taken” (1916):

**Two roads diverged in a yellow wood**  
And sorry I **could not travel both**  
And be one traveler, long I stood  
And looked down one as far as I could  
To where it bent in the undergrowth

Frost creates the mood attending a decision, one whose full consequences we cannot see or anticipate: “Knowing how way leads on to way,” we will not see the road not taken. As it was for Frost in a **yellow wood**, so it is for a patient at risk of death in the ProCESS Trial, and so it will be in every causal question.



# Ride against the flow

Spencer

For the past six springs, New Yorkers pedaled past colorful blossoms on their way to work, home, or just cruising.

Yet, some cruisers wearied whilst scouting a docking station [•]. And some on foot languished curbside without saddle to straddle.

**Empty and full docking stations sprout like dandelions** under the sun and moon, shown in 10 minute increments. Availability waxes and wanes by time and place.

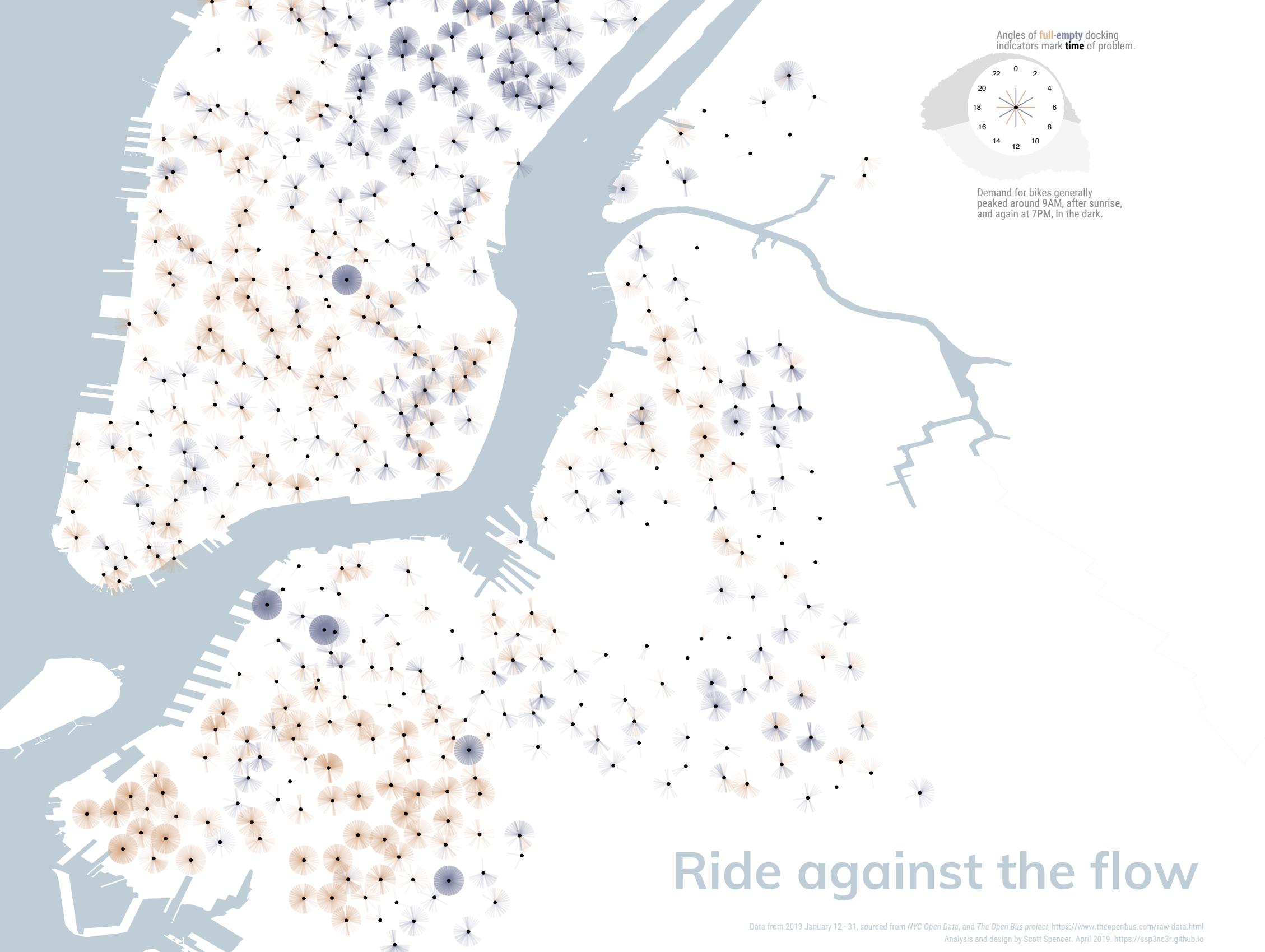
A ride against the flow is a joy ride for us all.

For the past six springs, New Yorkers pedaled past colorful blossoms on their way to work, home, or just cruising.

Yet, some cruisers wearied whilst scouting a docking station [•]. And some on foot languished curbside without saddle to straddle.

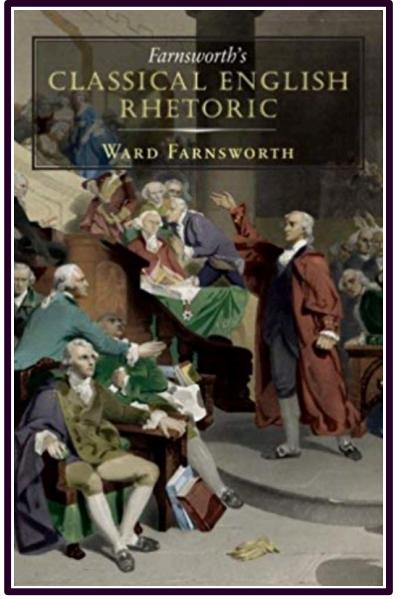
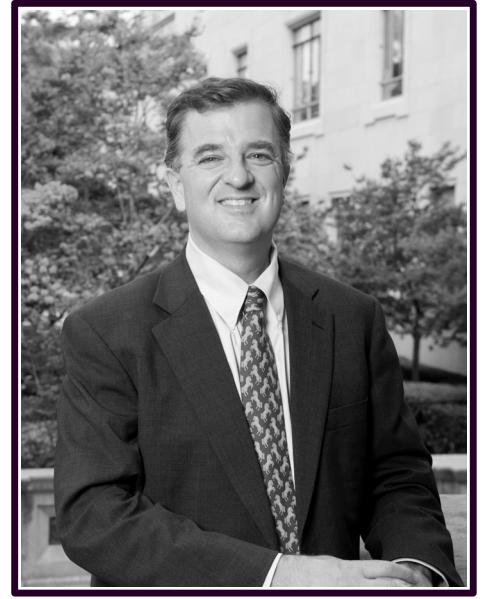
**Empty and full** docking stations sprout like dandelions under the sun and moon, shown in 10 minute increments. Availability waxes and wanes by time and place.

A ride against the flow is a joy ride for us all.





# **Patterns that compare, organize, grab attention**



# Classical English Rhetoric

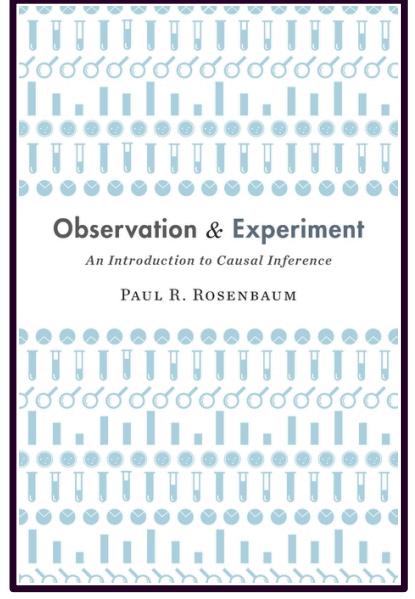
## Farnsworth

He is dean and professor of the University of Texas School of Law. Before teaching, he graduated from University of Chicago Law School, clerked for Supreme Court Justice Kennedy, and served as advisor to an international tribunal in the Hague.

**Use patterns  
to compare,  
grab attention,  
add emphasis**

We can use patterns to “make the words they arrange more emphatic or memorable or otherwise effective.” These patterns can be the most effective and efficient ways to show comparisons and contrasts.

**Example:** Reversal of structure, repetition at the end



# Observation & Experiment

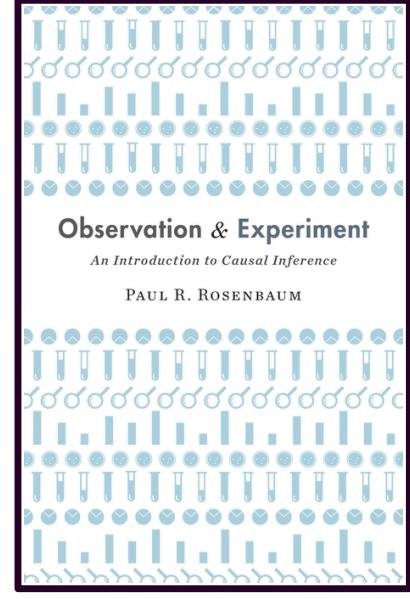
## Rosenbaum

He is Professor of Statistics at the Wharton School and a Senior Fellow of the Leonard Davis Institute of Health Economics, University of Pennsylvania. His book epitomizes the idea that “the most important ideas in statistics can be clearly explained in plain English, with little or no math.”

“

A **covariate** is a quantity determined prior to treatment assignment. In the Pro-CESS Trial, the age of the patient at the time of admission to the emergency room **was a covariate**. The gender of the patient **was a covariate**. Whether the patient was admitted from a nursing home **was a covariate**.

**Example:** Repetition at the start, parallel structure



# Observation & Experiment

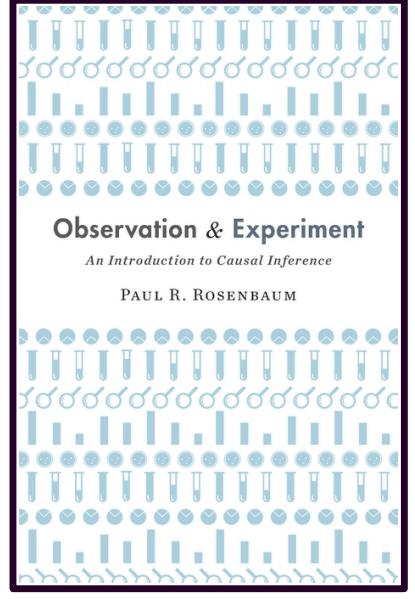
Rosenbaum

He is Professor of Statistics at the Wharton School and a Senior Fellow of the Leonard Davis Institute of Health Economics, University of Pennsylvania. His book epitomizes the idea that “the most important ideas in statistics can be clearly explained in plain English, with little or no math.”



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

**Example:** Asking questions and answering them



# Observation & Experiment

## Rosenbaum

He is Professor of Statistics at the Wharton School and a Senior Fellow of the Leonard Davis Institute of Health Economics, University of Pennsylvania. His book epitomizes the idea that “the most important ideas in statistics can be clearly explained in plain English, with little or no math.”

“

Where did Fisher’s null distribution come from?  
From the coin in Fisher’s hand.



# Statistical Modeling, Causal Inference, and Social Science

Gelman

Professor of Statistics and Political Science at Columbia University, he is known widely for his work in Bayesian statistics, and has authored several textbooks, including Teaching Statistics, and Bayesian Data Analysis.

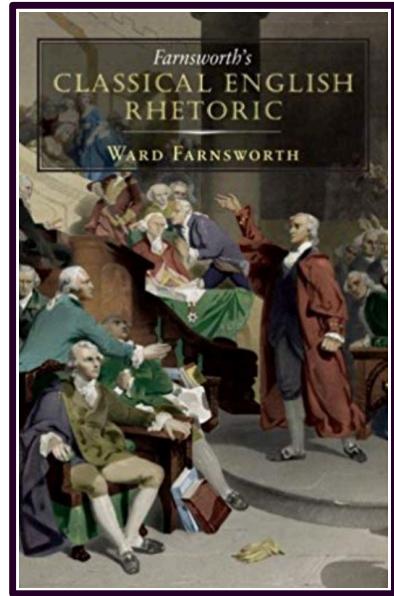
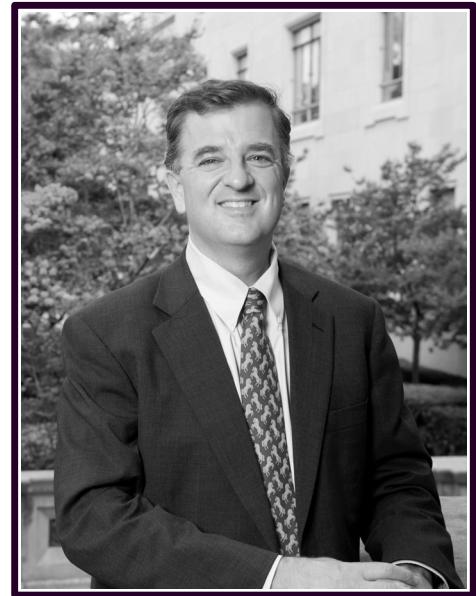
The screenshot shows a blog post titled "The most important aspect of a statistical analysis is not what you do with the data, it's what data you use" (survey adjustment edition). The post discusses the importance of choosing the right variables in statistical analysis, mentioning that complex methods like MRP can sometimes lead to better results than simpler ones like partial pooling. The author quotes David Littlejohn's comment: "The most important aspect of a statistical analysis is not what you do with the data, it's what data you use".

## Example: Inversion of words

“

The most important aspect of a statistical analysis is not what **you** do with the **data**, it's what **data you** use.

## Repetition of words & phrases



# Classical English Rhetoric

## Farnsworth

He is dean and professor of the University of Texas School of Law. Before teaching, he graduated from University of Chicago Law School, clerked for Supreme Court Justice Kennedy, and served as advisor to an international tribunal in the Hague.

## Structural matters

simple repetition (*epizeuxis, epimone*)  
repetition at the start (*anaphora*)  
repetition at the end (*epistrophe*)  
repetition at the start and end (*symploce*)  
repeating the ending at the beginning (*anadiplosis*)  
repetition of the root (*polyptoton*)

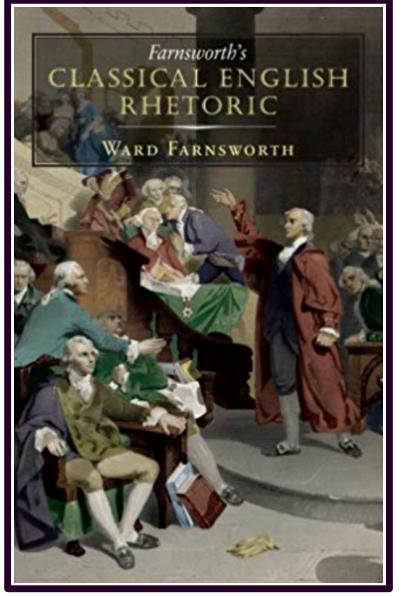
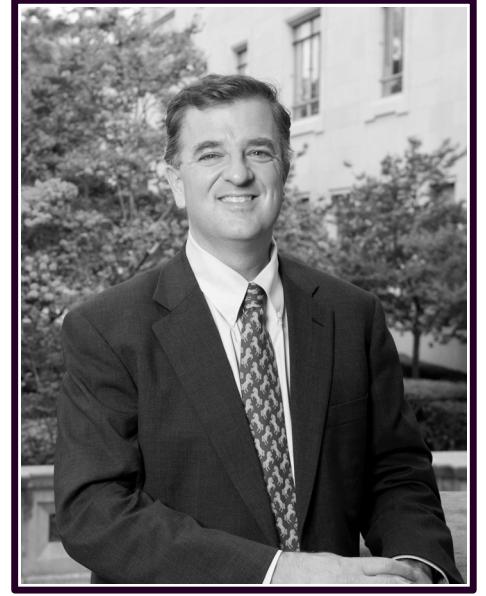
## Dramatic devices

parallel structure (*isocolon*)  
reversal of structure (*chiasmus*)  
inversion of words (*anastrophe*)  
leaving out words (*ellipsis*)  
  
saying things by not saying them (*præteritio*)  
correcting oneself (*metanoia*)  
rhetorical uses of the negative (*litotes*)  
rhetorical questions (*erotema*)  
asking questions and answering them (*hypophora*)  
anticipating objections and meeting them (*prolepsis*)

# How unexpected patterns work

Unexpected word placement calls attention to them, creates emphasis by coming earlier than expected or violating the reader's expectations. Note that, to violate expectations necessarily means reserving a technique like inversion for just the point to be made, lest the reader come to expect it —

**more is less, less is more.** Secondly, it can create an attractive rhythm. Thirdly, when the words that bring full meaning come later, it can add suspense, and finish more climactic.



# Classical English Rhetoric

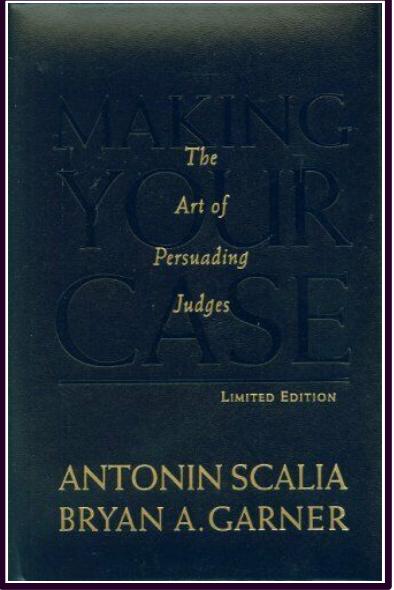
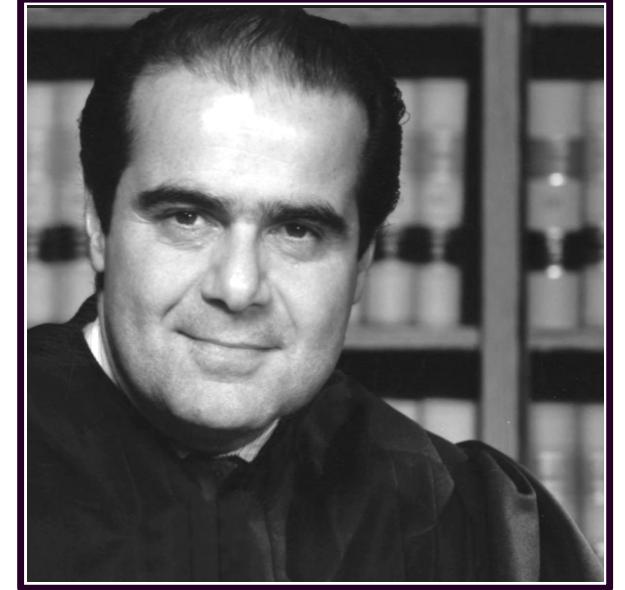
## Farnsworth

He is dean and professor of the University of Texas School of Law. Before teaching, he graduated from University of Chicago Law School, clerked for Supreme Court Justice Kennedy, and served as advisor to an international tribunal in the Hague.

**Immersion  
precedes  
implementation**

Seeing just a few examples invites direct imitation of them, which tends to be clumsy. Immersion in many examples allows them to do their work by way of a subtler process of influence, with a gentler and happier effect on the resulting style.

# Point made



# Making your Case

*Scalia & Garner*

The authors—a former Supreme Court Justice and renowned legal writer—have taught lawyers of the highest acclaim how to persuade judges and juries alike. The peculiarities of legal persuasion are slight compared with the principles it shares with persuasion more generally.

**Close powerfully,  
state explicitly what  
the audience should do**



Persuasive argument neither comes to an abrupt halt nor trails off in a grab-bag of minor points. The art of rhetoric features what is known as the peroration—the conclusion of argument, which is **meant to move the listener to act on** what the preceding argument has logically described.

# Wrapping up

# For Next Week, Module 5:

## Agenda next week

Next deliverable, draft 750-word (or less) proposal  
Audience analysis

### The minimum

Kahneman, Daniel, Dan Lovallo, and Olivier Sibony.  
*Before You Make That Big Decision ...* Harvard  
Business Review 89.6 (2011): 50–60. Print.

Read to understand common limitations and  
approaches to reasoning and making decisions  
amid uncertainty.

Dragicevic, Pierre. “Fair Statistical Communication  
in HCI.” *Modern Statistical Methods for HCI*.  
Springer International Publishing, 2016. 1–40.

Read to consider what may be important in  
communicating statistical analysis. Also,  
consider the graphical displays integrated  
into the writing.

Healy, Kieran. *Data Visualization*. Princeton  
University Press, 2019. Web. <https://socviz.co>

This is a great resource if you need help  
implementing visual displays in R.

# Craft this course for you,

## Turtles and hares?

Of what we covered so far, what material or concept would you like further review? Or are you ready as a rabbit to get on with it?

## Practice is important

Outside of class assignments, how, and how often, do you practice writing? I recommend keeping a data science journal, writing something, anything on your mind about data science each week.

**See you  
next week!**

