Storytelling with data

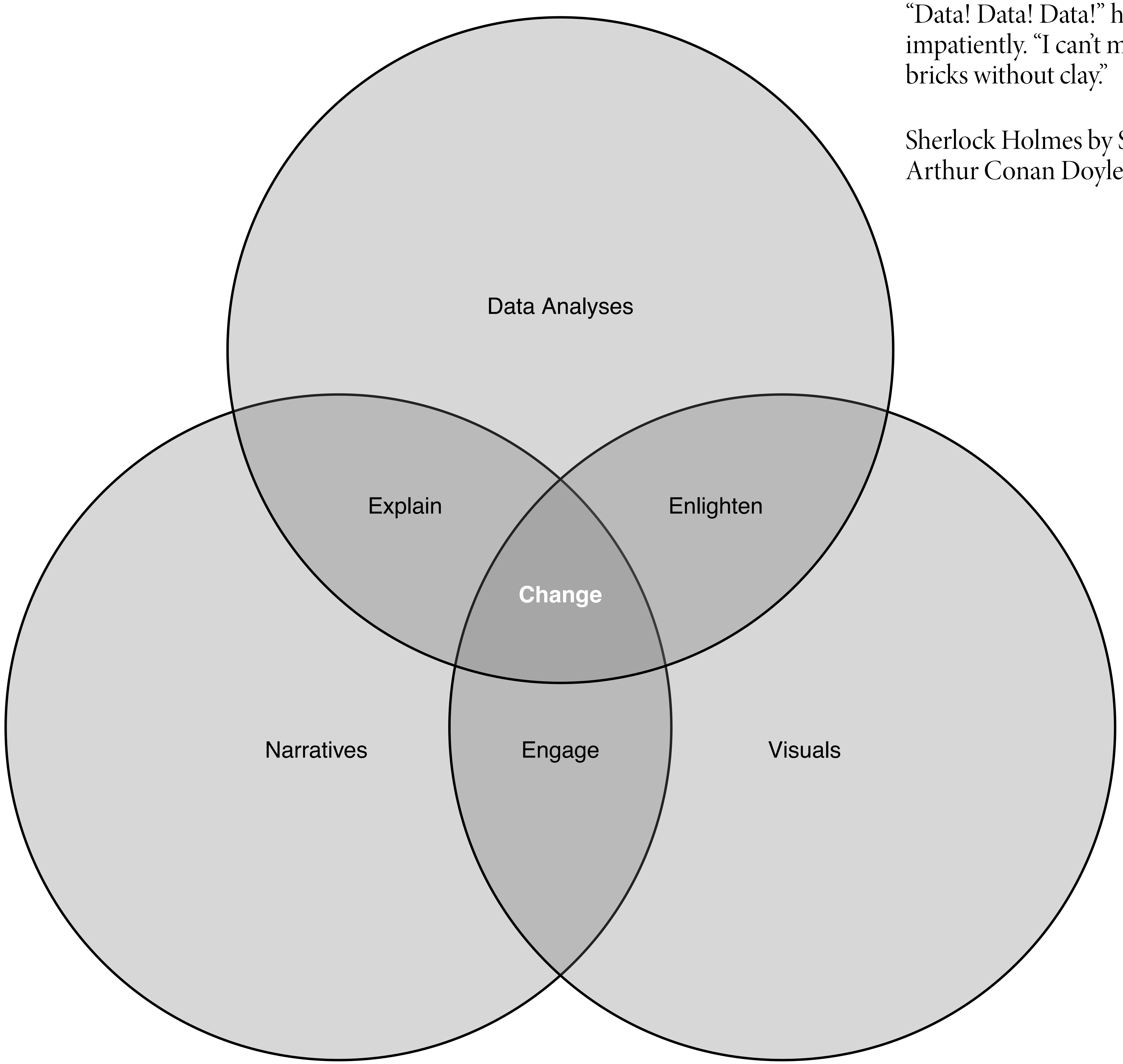
07 | Communicating context, uncertainty, estimates, and forecasts

Scott Spencer | Columbia University

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

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

Sherlock Holmes by Sir Arthur Conan Doyle, *author*



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

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

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

John W Tukey, *mathematician*

general course deliverable timeline

Individual Work

For learning data visualization
and written narrative techniques

Group work

For building graphics and narrative
into interactive communications

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

next deliverables, **individual homework three and group proposal**

Individual Work

For learning data visualization
and written narrative techniques

Group work

For building graphics and narrative
into interactive communications

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

Participation 10%

individual homework four check-in | graphics practice with Citi Bike rebalancing study

individual homework four check-in, **questions?**

homework 4.rmd

Knit on Save · Knit · Run

```
google_analytics: UA-123500360-1
---

```{r setup, include=FALSE}
knitr::opts_chunk$set(
  eval = TRUE,
  echo = TRUE,
  error = FALSE,
  message = FALSE,
  warning = FALSE
)
```

In our previous class demonstrations and homeworks, we practiced exploring CitiBike ride data to gain insights into the bike share's rebalancing efforts. In the process, we gained experience transforming data and mapping data to visual encodings.

First, as a class we practiced using a workflow with CitiBike data to create a new variable, an indicator whether bikes may have been rebalanced. Next, in homework two, we practiced mapping CitiBike ride data onto the three attributes of color: hue, saturation, and luminance. In the process we were able to explore how useage, rebalancing efforts, or both may have changed between 2013 and 2019, and again before and after the pandemic began. This exploration also helped us consider some of the limitations of the particular visualization: it did not consider the effects of rebalancing or bike and docking station availability.

In this assignment, we will try to account for those and other limitations in the visualizations, and in the process gain practice with new data graphics and *explaining* our insights to others.

Preliminary setup

Load libraries to access functions we'll use in this analysis. Of note, if you have not installed these packages, do so outside of this `rmd` file.


```{r}
library(tidyverse) # the usual
library(geojsonio) # for map data
library(broom)     # for map data
library(patchwork) # for organizing multiple graphs
library(ggthemes)  # collection of graph themes
theme_set(theme_tufte(base_family = 'sans'))
```


We'll use the same dataset as in our previous homework. Let's load our data and rename variables (as before),

```



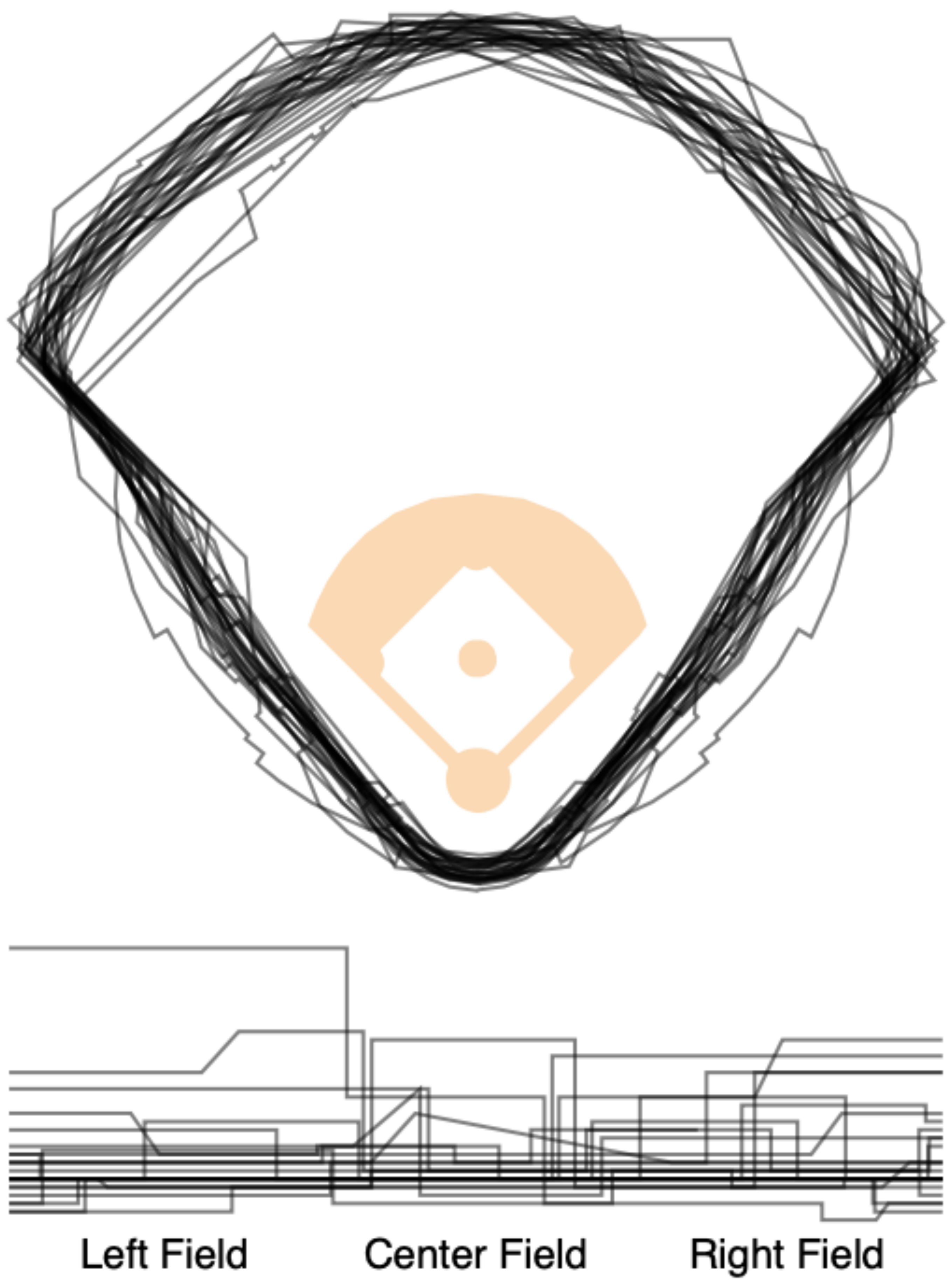
30:1 (Top Level) R Markdown

Console

**group project check-in** | **proposals**

# data in context — the data generating process

# data generating process, meaning of data depends on context — example (*baseball: stadium*, *location*, *weather*, *people*, …)




Spencer, Scott. 1.1.1.2 "Understanding data requires context." In *Data in Wonderland*, 2021.
https://ssp3nc3r.github.io/data_in_wonderland/#understanding-data-requires-context

The focus on collecting "big data" for analyses can miss ***differences in what data represent***.

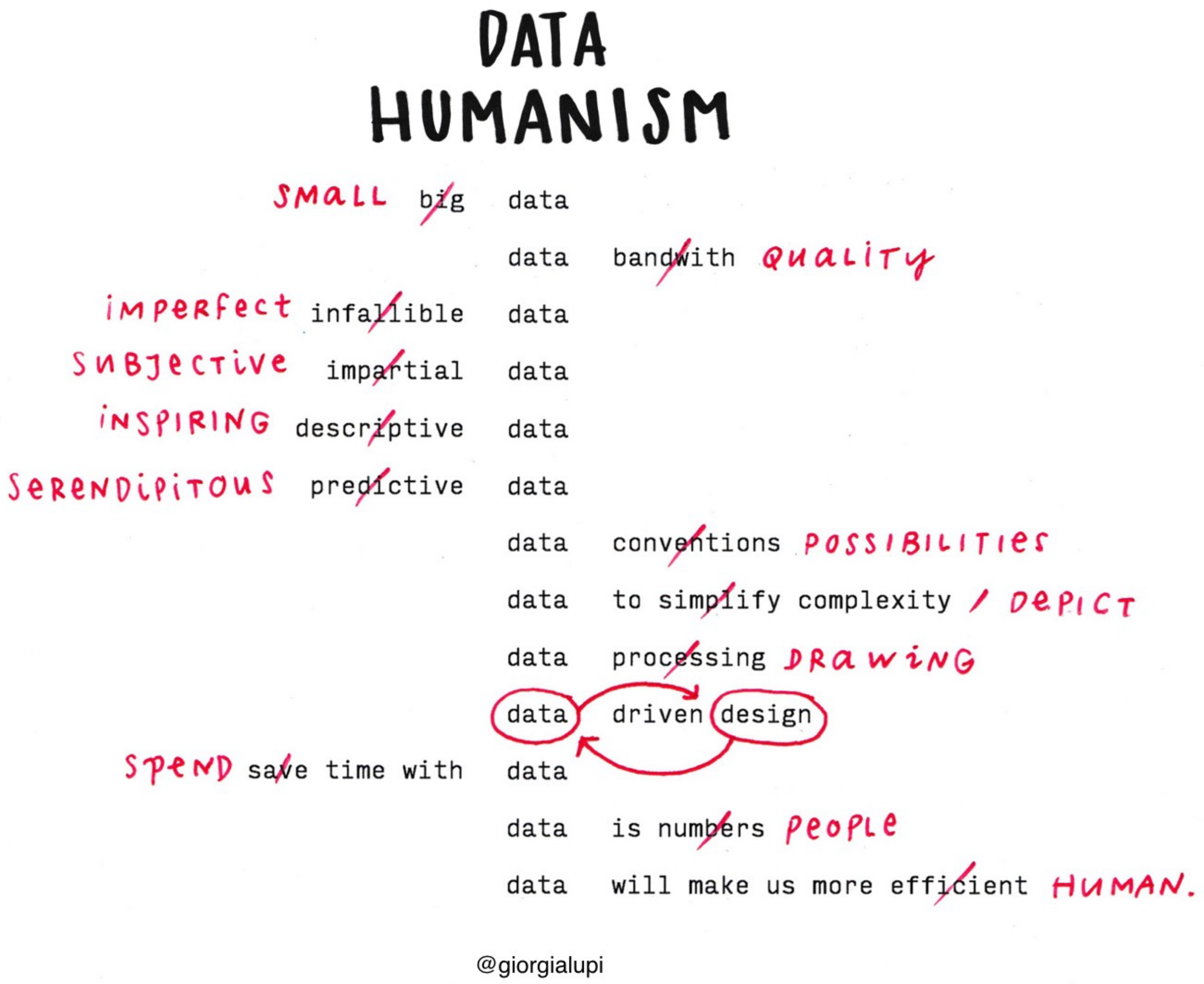
**What** generated each observation? Be specific with context. **How** was each observation measured? **Who** collected each observation? …

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

# data generating process, **data as representation — an understanding of data requires its context!**

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

— Giorgia Lupi, *Information Designer*

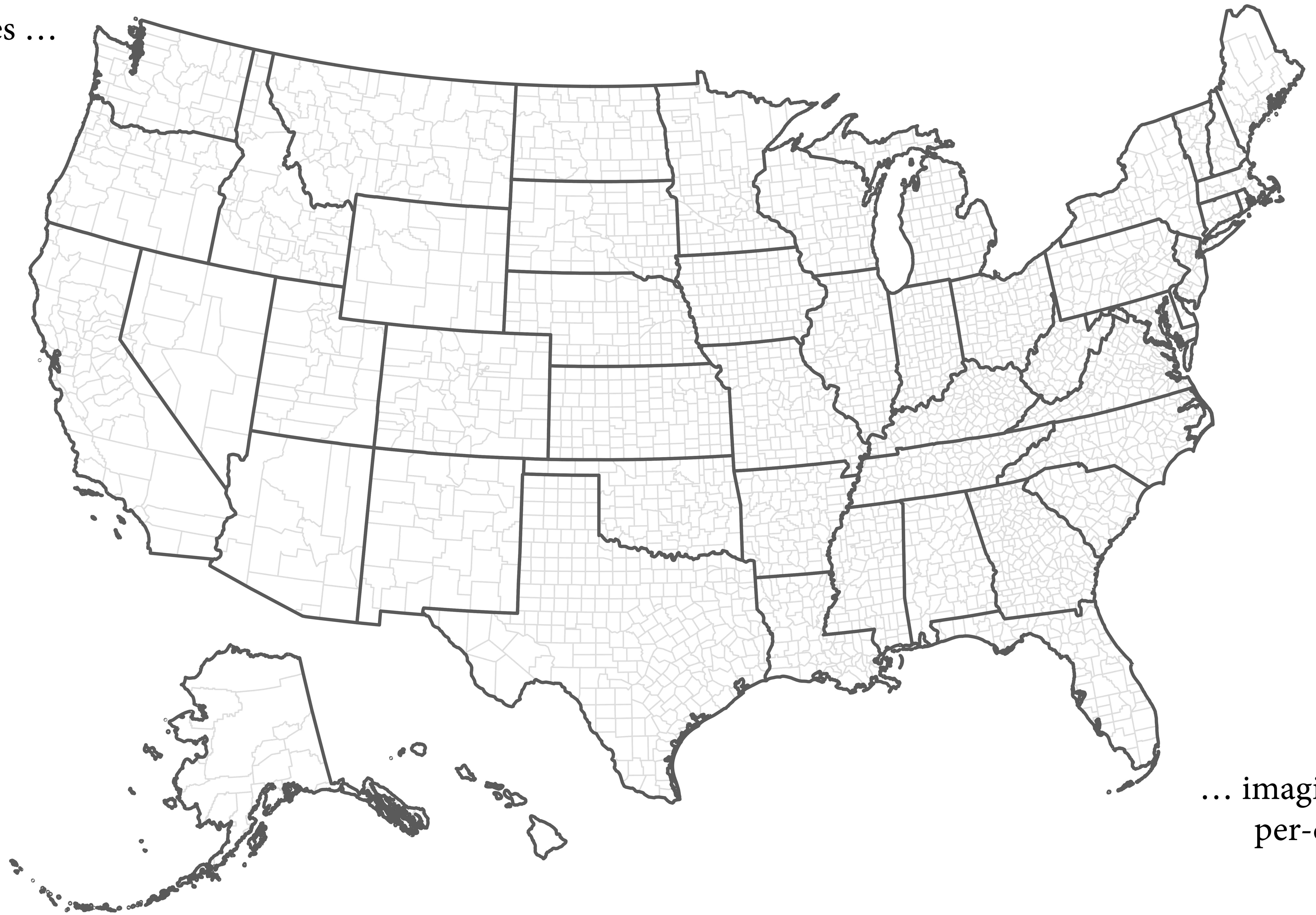



Lupi, Giorgia. "Data Humanism: The Revolutionary Future of Data Visualization." In *PrintMag*, January 30, 2017. https://www.printmag.com/post/data-humanism-future-of-data-visualization.

**understanding elements of variation, a *simulation* study**

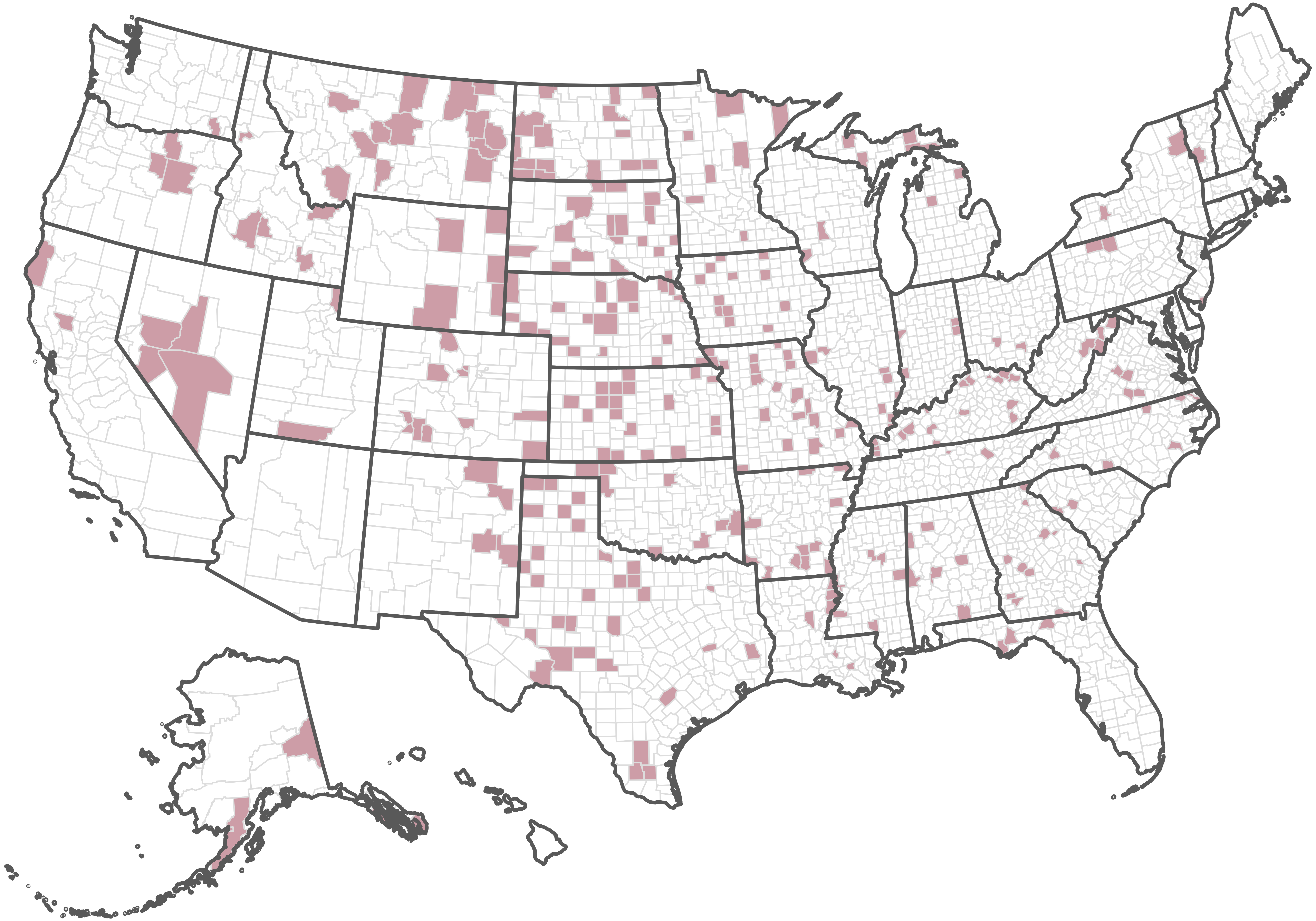
 ***imagine* collecting data on rate of cancer in each county of the United States**

Map of United States counties …

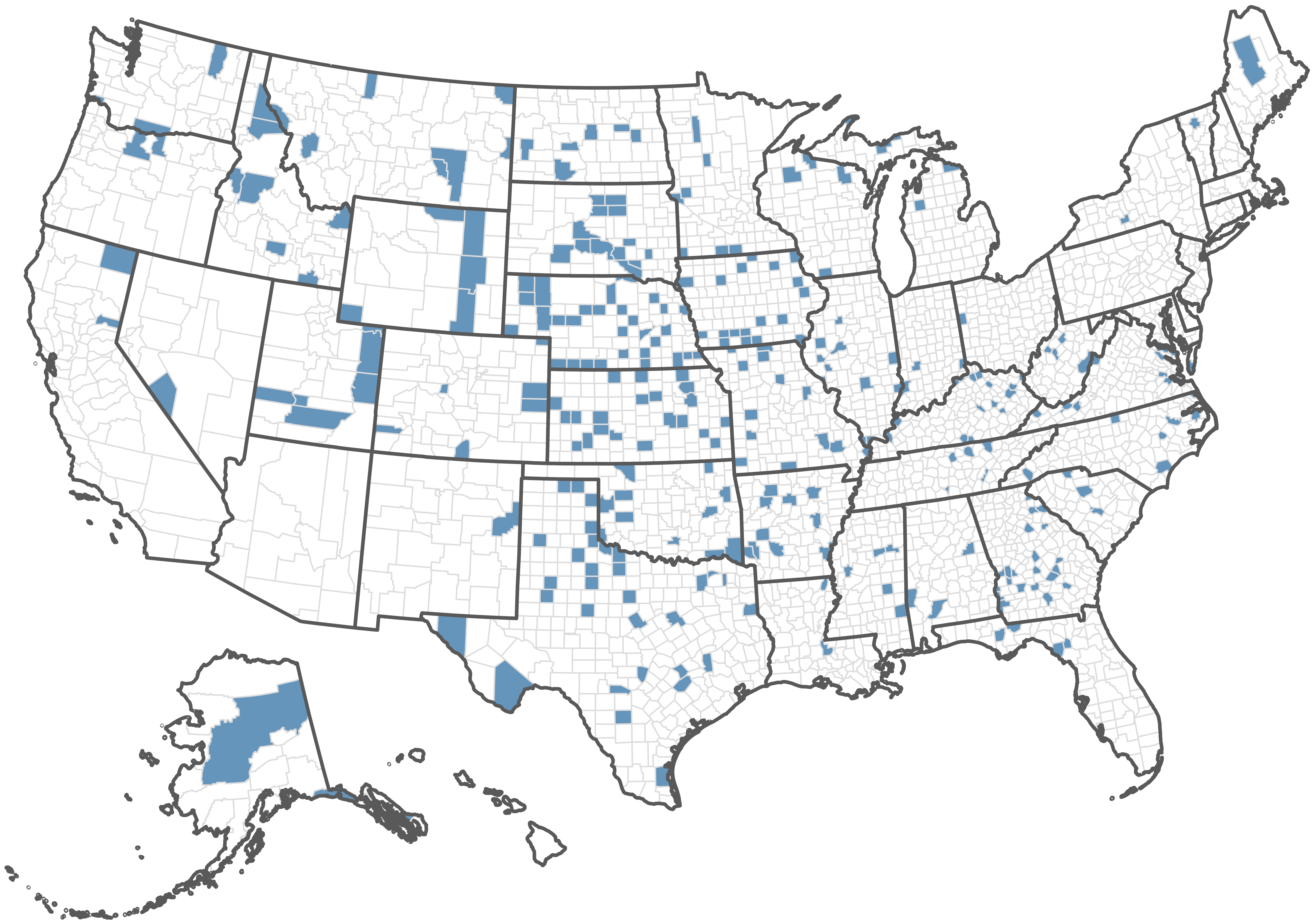


… imagine collecting data on
per-county rate of cancer.

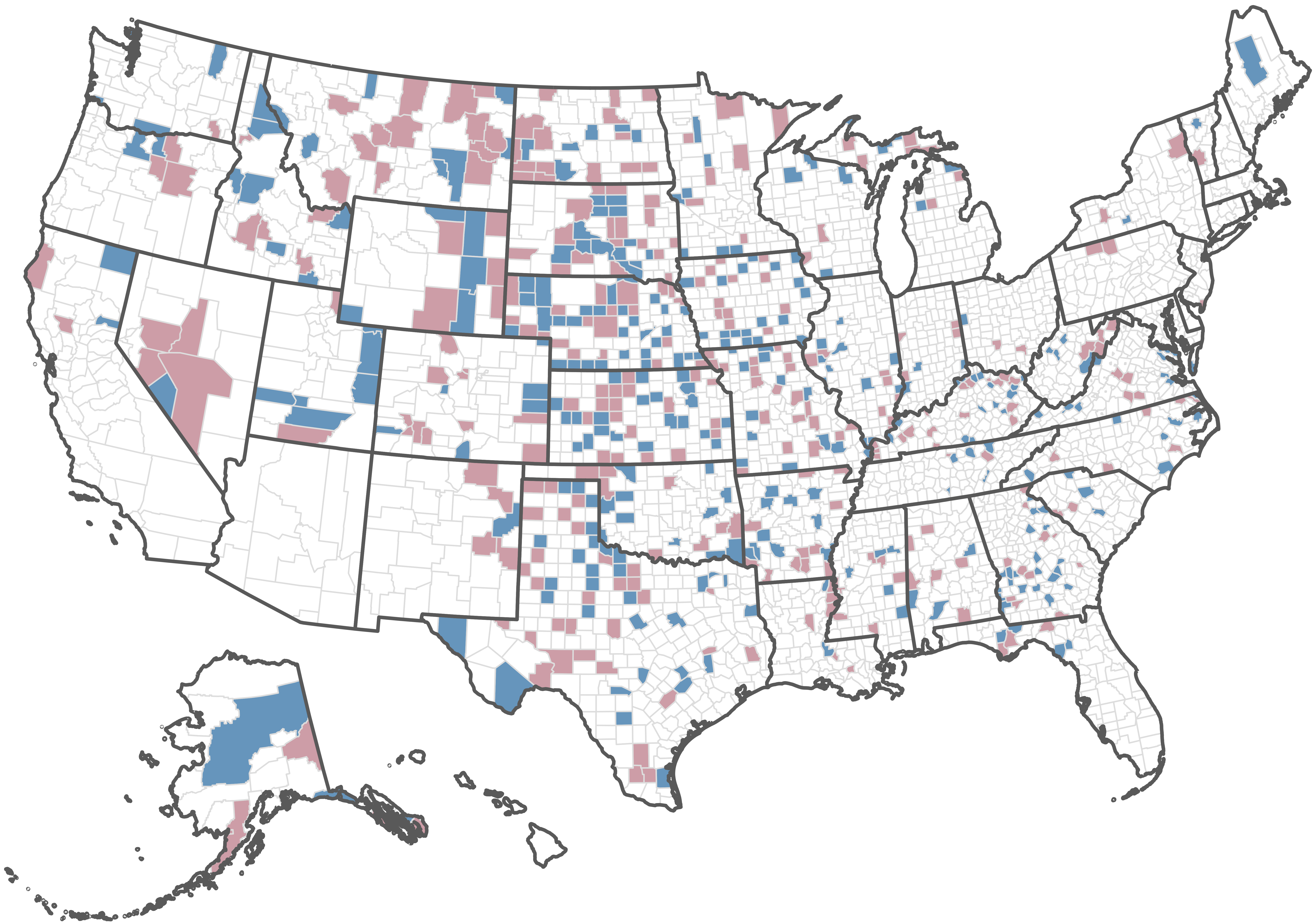
simulation study on variation, **United States counties with *highest* decile of age-adjusted cancer rates**



# simulation study on variation, **United States counties with *lowest* decile of age-adjusted cancer rates**



**simulation study on variation, United States counties with either *lowest* or *highest* decile of age-adjusted cancer rates**



# simulation study on variation, data was *simulated* from a single population rate across all counties — one percent

```
library(tidyverse)
library(ggthemes)
library(tidycensus)

CENSUS_API_KEY <- Sys.getenv("CENSUS_API_KEY")

county_pop <- get_estimates(
  geography = "county",
  product = "population",
  year = 2019,
  key = CENSUS_API_KEY)

set.seed(1)

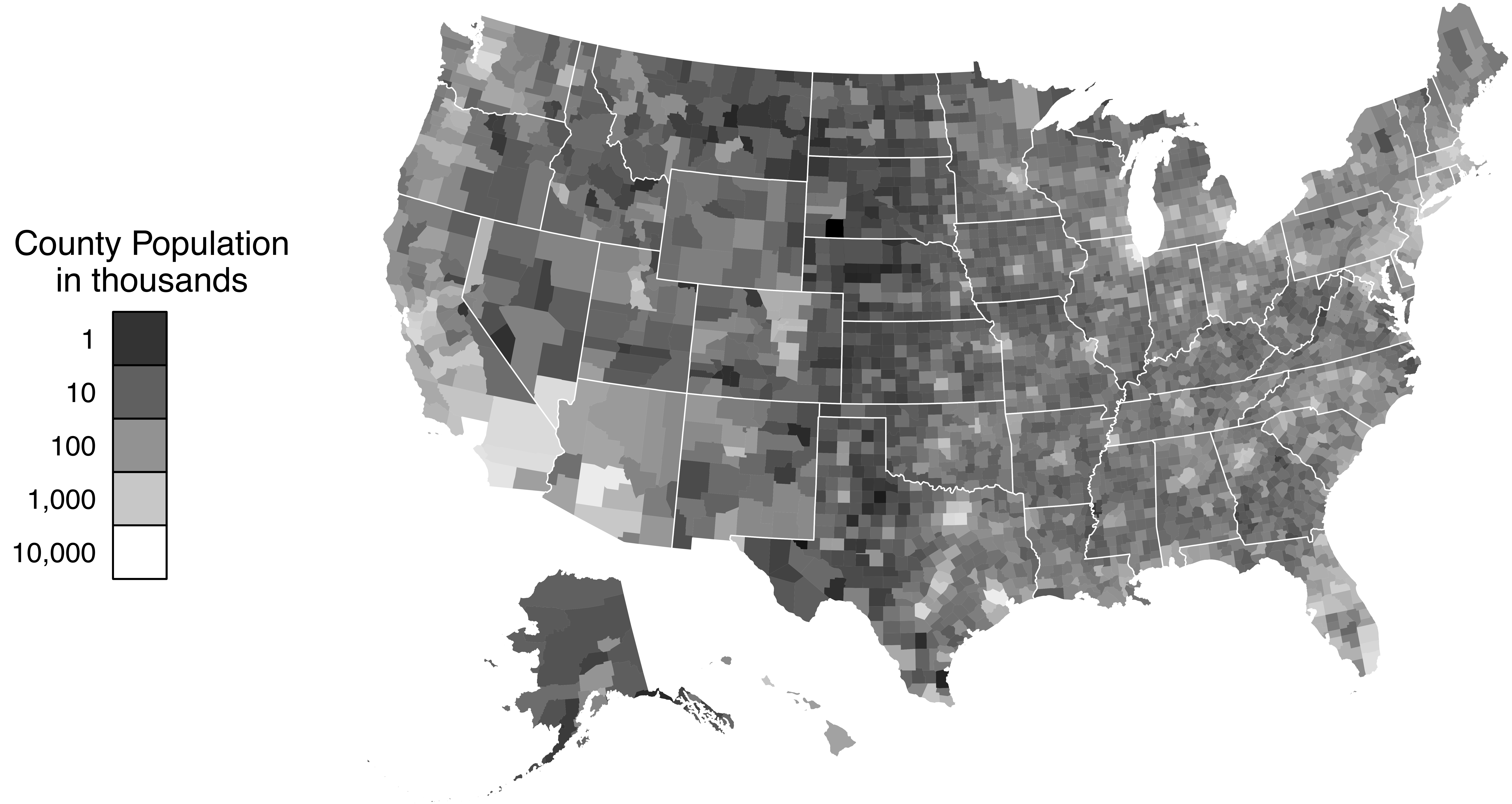
data <-
  county_pop %>%
  pivot_wider(
    names_from = "variable",
    values_from = "value") %>%
  left_join(county_laea) %>%
  rowwise() %>%
  mutate(
    rate_cases = rbinom(n = 1, size = POP, prob = 0.01) / POP * 1000
)
```

```
p <- ggplot() + theme_void()  +
  theme(legend.position = "")


# map encoding lowest decile of rates using luminance of polygon fills
p +
  geom_sf(
    data = data,
    mapping = aes(
      geometry = geometry,
      fill = rate_cases < quantile(rate_cases, probs = 0.1)),
    lwd = 0.2,
    color = '#dddddd'
  ) +
  geom_sf(
    data = state_laea,
    fill = NA
  ) +
  scale_fill_manual(values = c("#ffffff", "#6695BC"))

# map encoding highest decile of rates using luminance of polygon fills
p +
  geom_sf(
    data = data,
    mapping = aes(
      geometry = geometry,
      fill = rate_cases > quantile(rate_cases, probs = 0.9)),
    lwd = 0.2,
    color = '#dddddd'
  ) +
  geom_sf(
    data = state_laea,
    fill = NA
  ) +
  scale_fill_manual(values = c("#ffffff", "#CD9DA7"))
```

simulation study on variation, **United States county populations**



The most dangerous equation

De Moivre's equation:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \quad \therefore \quad \sigma_{\bar{x}} < \sigma$$

σ	the measure of the variability of a population (its standard deviation).
$\sigma_{\bar{x}}$	the variation of averages of subsets of the population.
n	the number of observations in each subset

Why so dangerous?

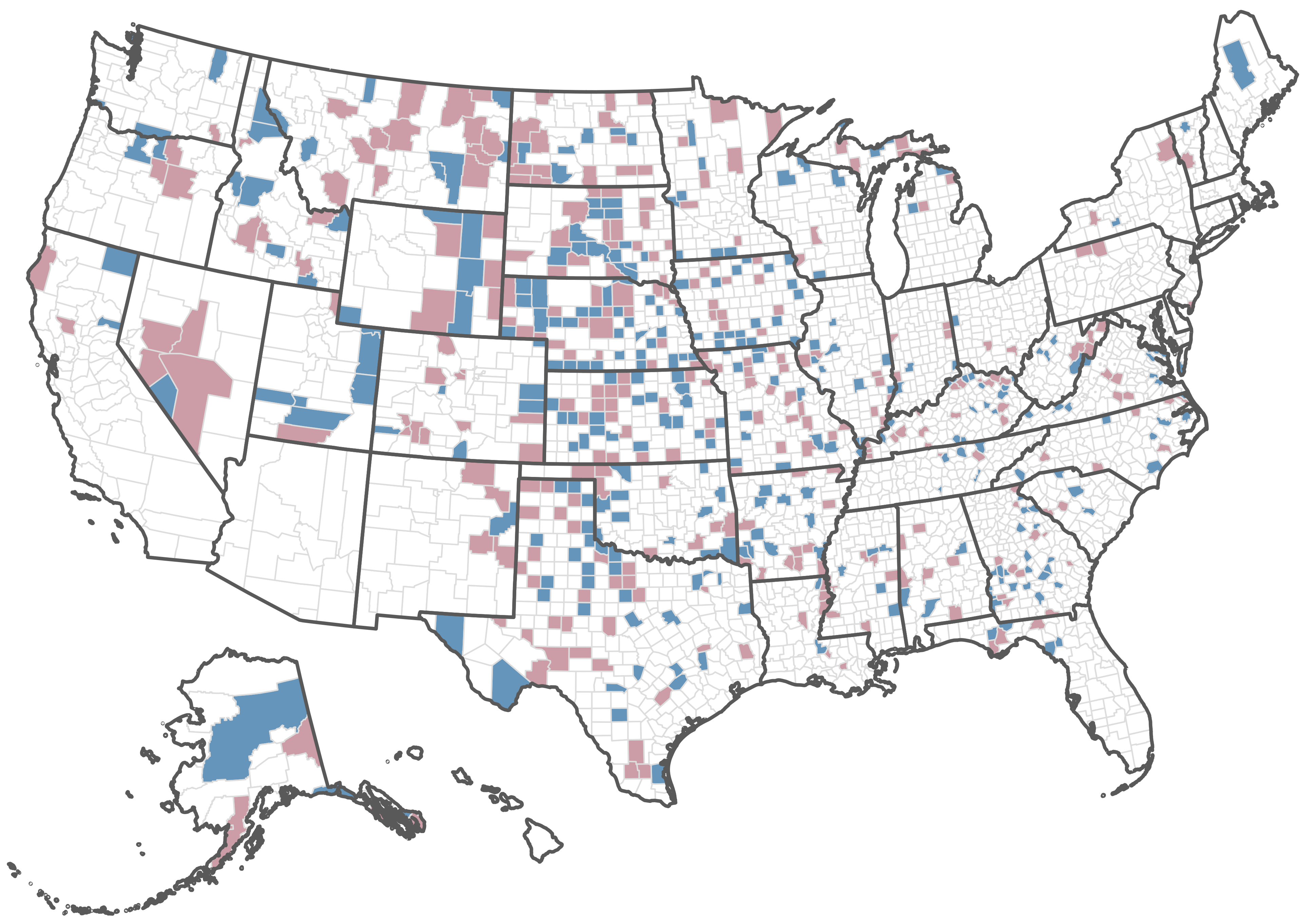
Extreme length of time during which ignorance of it has caused confusion

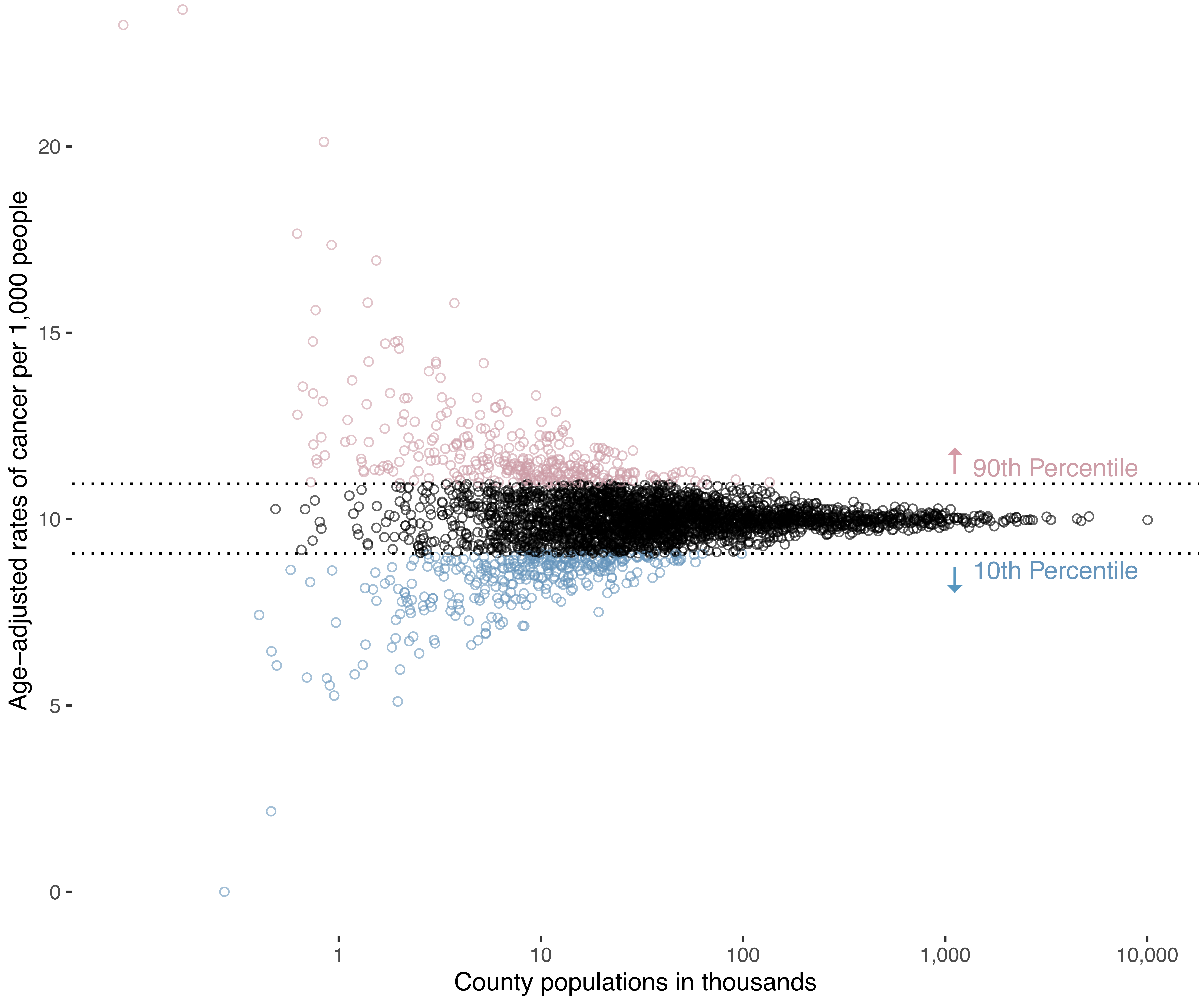
Wide breadth of areas that have been misled

Seriousness of the consequences that ignorance has caused

Wainer, Howard. "The Most Dangerous Equation." In *Picturing the Uncertain World*, 5–20. Princeton University Press, 2009.

simulation study on variation, **simulated, age-adjusted cancer rates ~ county populations | single, true rate of one percent**





communicating uncertainty, overcoming concerns

Concern | people will **misinterpret** quantities of uncertainty, inferring more precision than intended.

Response | Most people *like* getting quantitative information on uncertainty, from it can *get the main message*, and without it are more likely to misinterpret verbal expressions of uncertainty. Posing clear questions guide understanding.

Concern | people **cannot use** probabilities.

Response | laypeople can provide high-quality probability judgments, if they are asked clear questions and given the chance to reflect on them. Communicating uncertainty protects credibility.

Concern | credible intervals may be **used unfairly** in performance evaluations.

Response | probability judgments give us more accuracy about the information; *i.e.*, won't be too confident or lack enough confidence.

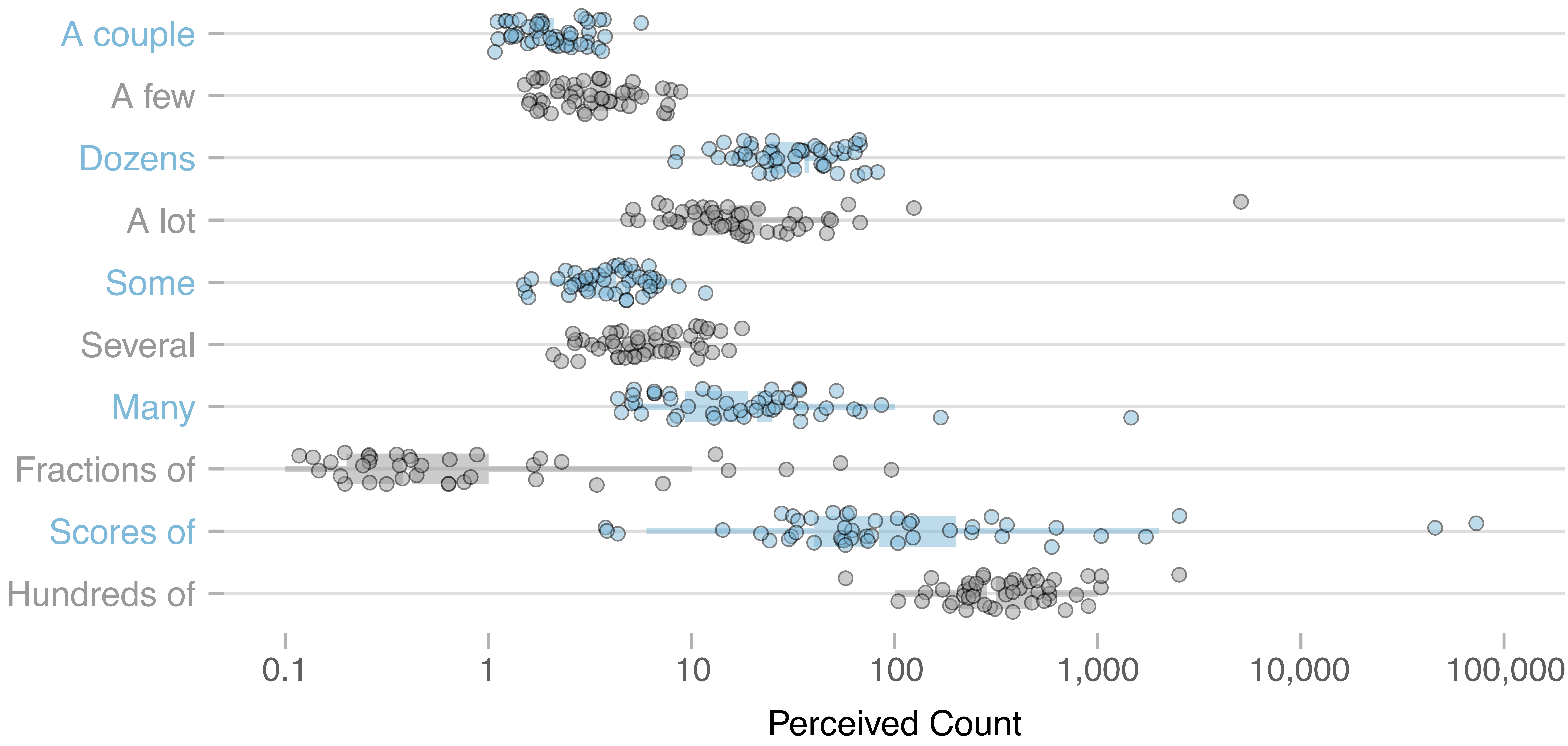
Fischhoff, Baruch. *Communicating Uncertainty: Fulfilling the Duty to Inform*. Issues in Science and Technology 28, no. 4 (August 2012): 63–70.

expressing uncertainty and variation — *le mot juste*?

What [*probability / number*] would you assign to the phrase "[*phrase*]"?

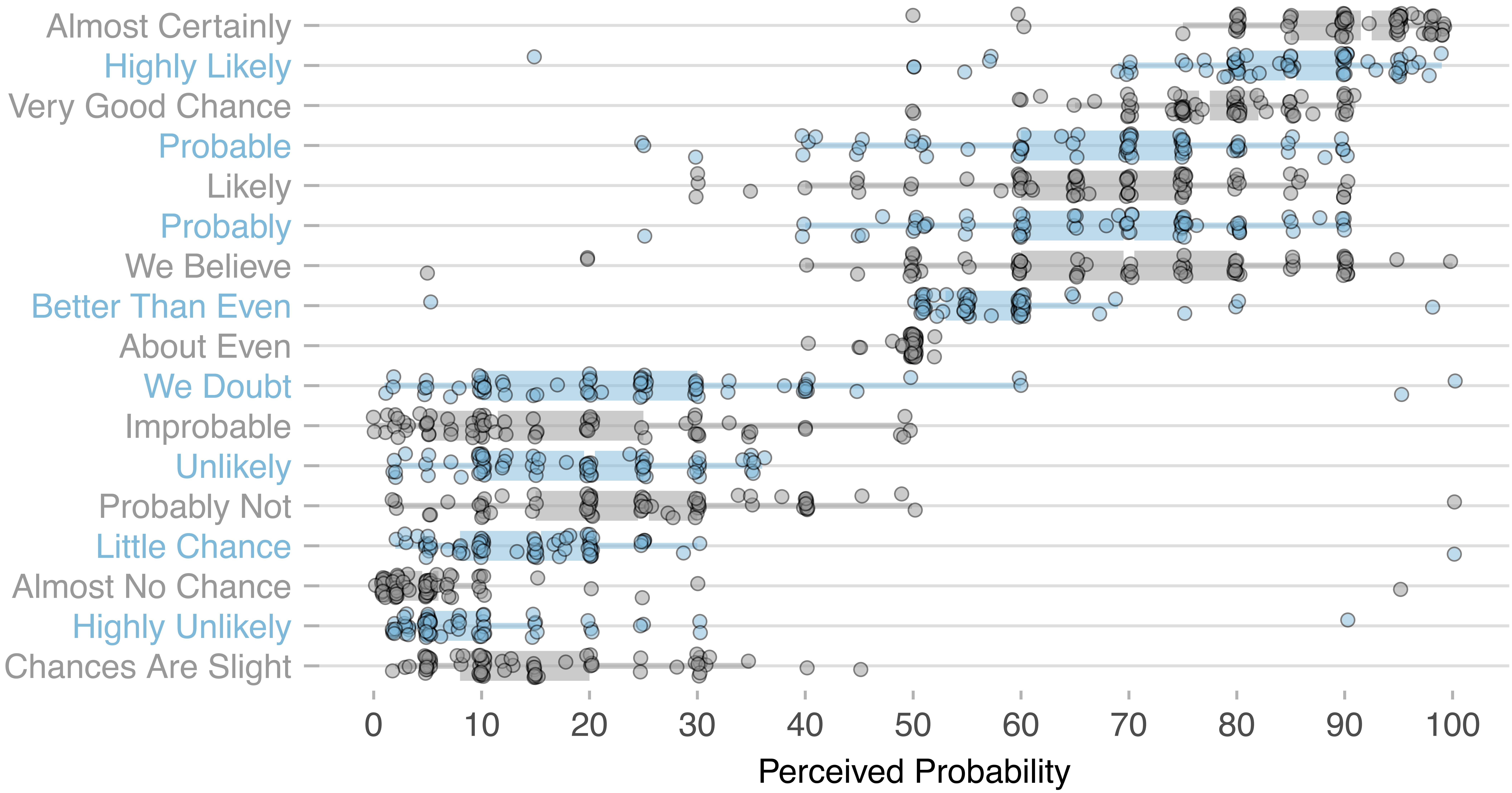
zonination. "Perceptions of Probability and Numbers," August 2015. https://github.com/zonination/perceptions.

uncertainty in language, survey question — What *number* would you assign to the phrase "[phrase]"?



zonination. "Perceptions of Probability and Numbers," August 2015. https://github.com/zonination/perceptions.

uncertainty in language, **survey question — What *probability* would you assign to the phrase "[phrase]"?**



zonination. "Perceptions of Probability and Numbers," August 2015. https://github.com/zonination/perceptions.

types of *uncertainty* in analyses

model specifications and selections

Do the models (parameters, data, functions) represent the underlying process intended for inference and account for data collection?

estimations in model parameters

parameters represent variation in observations, measurement error, etc

whether computations work as intended

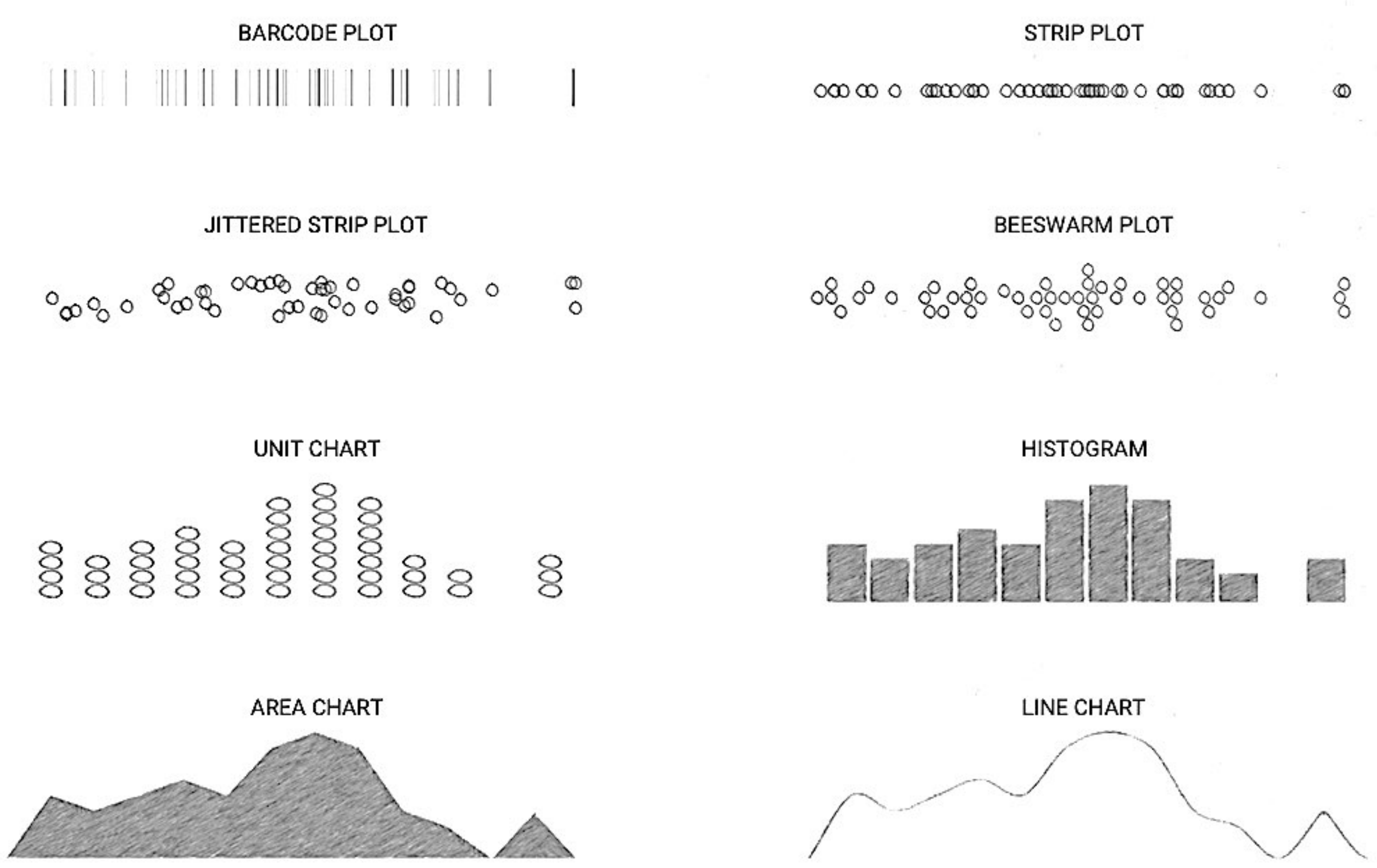
e.g., calculation overflows, underflows, coding mistakes

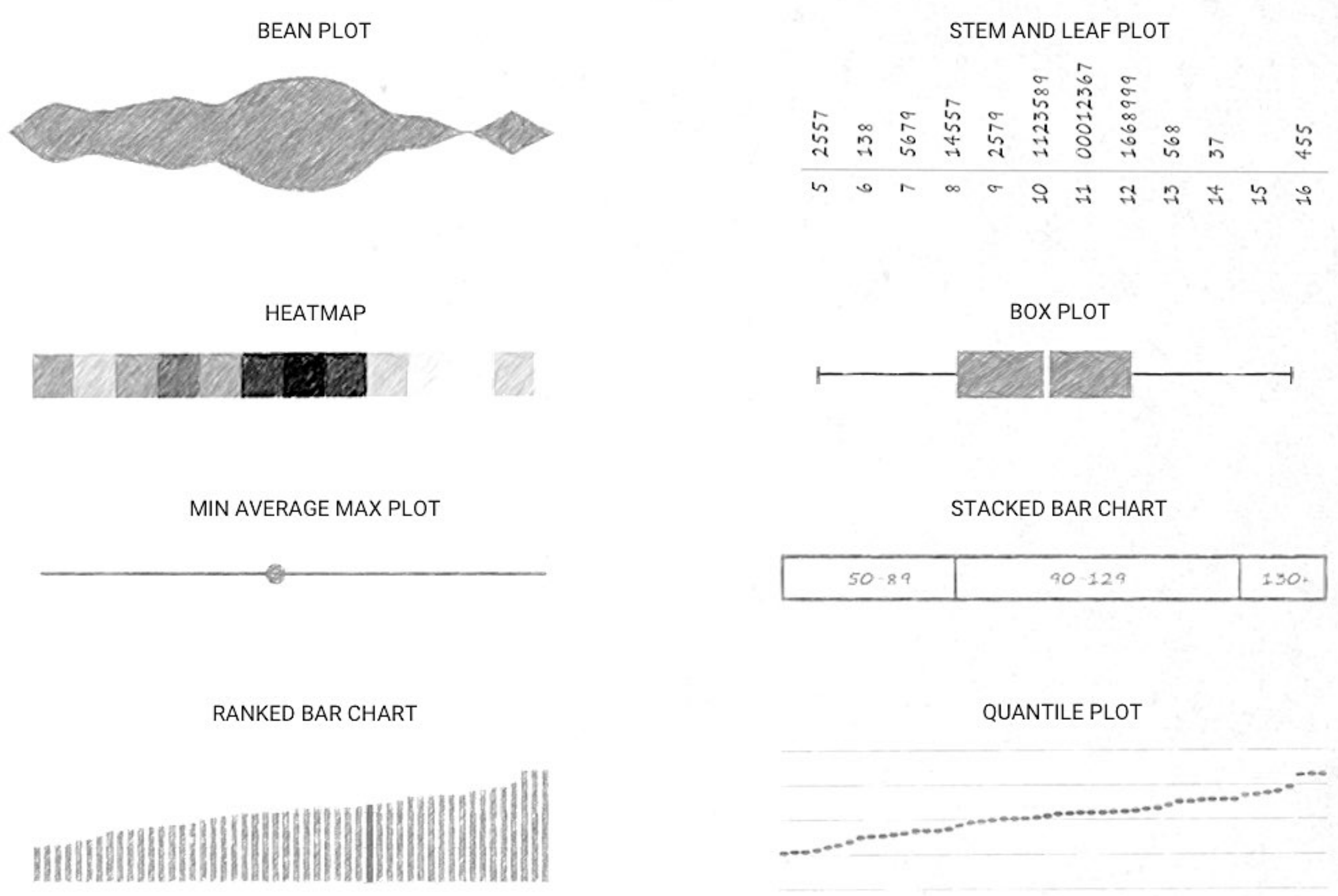
decisions from model outputs

look to decision theory, utility functions

expressing uncertainty, estimates, forecasts — distinguishing them

Example taxonomies of common visual encodings for *variation* in *measures*, and for *estimates* …

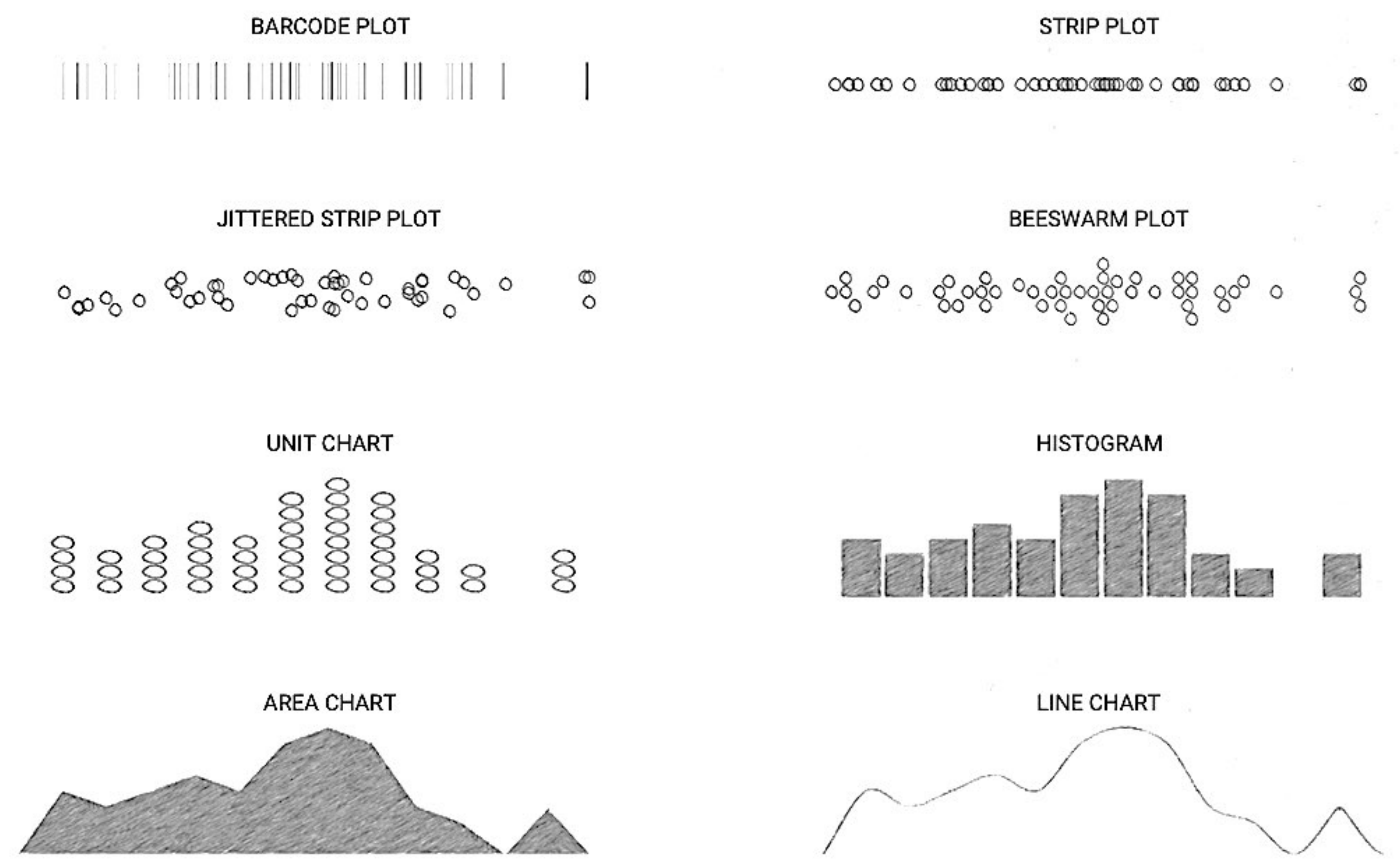


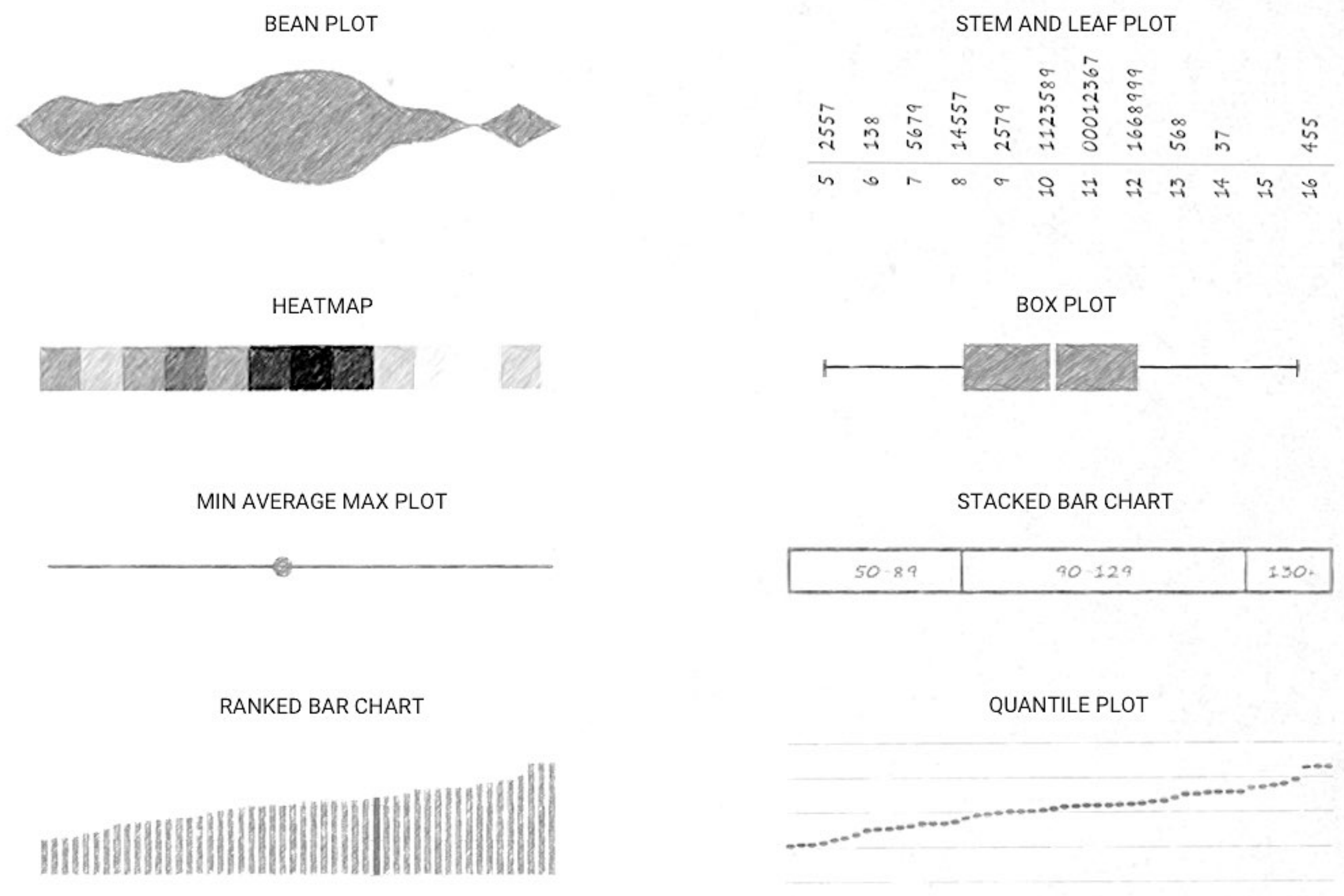


Cherdarchuk, Joey. "*Visualizing Distributions*." Business. Dark Horse Analytics (blog), November 8, 2016. https://www.darkhorseanalytics.com/blog/visualizing-distributions-3.

encoding uncertainty, estimates, forecasts, distinguishing measurements from estimates

Example taxonomies of common visual encodings for *variation* in *measures*, and for *estimates* …





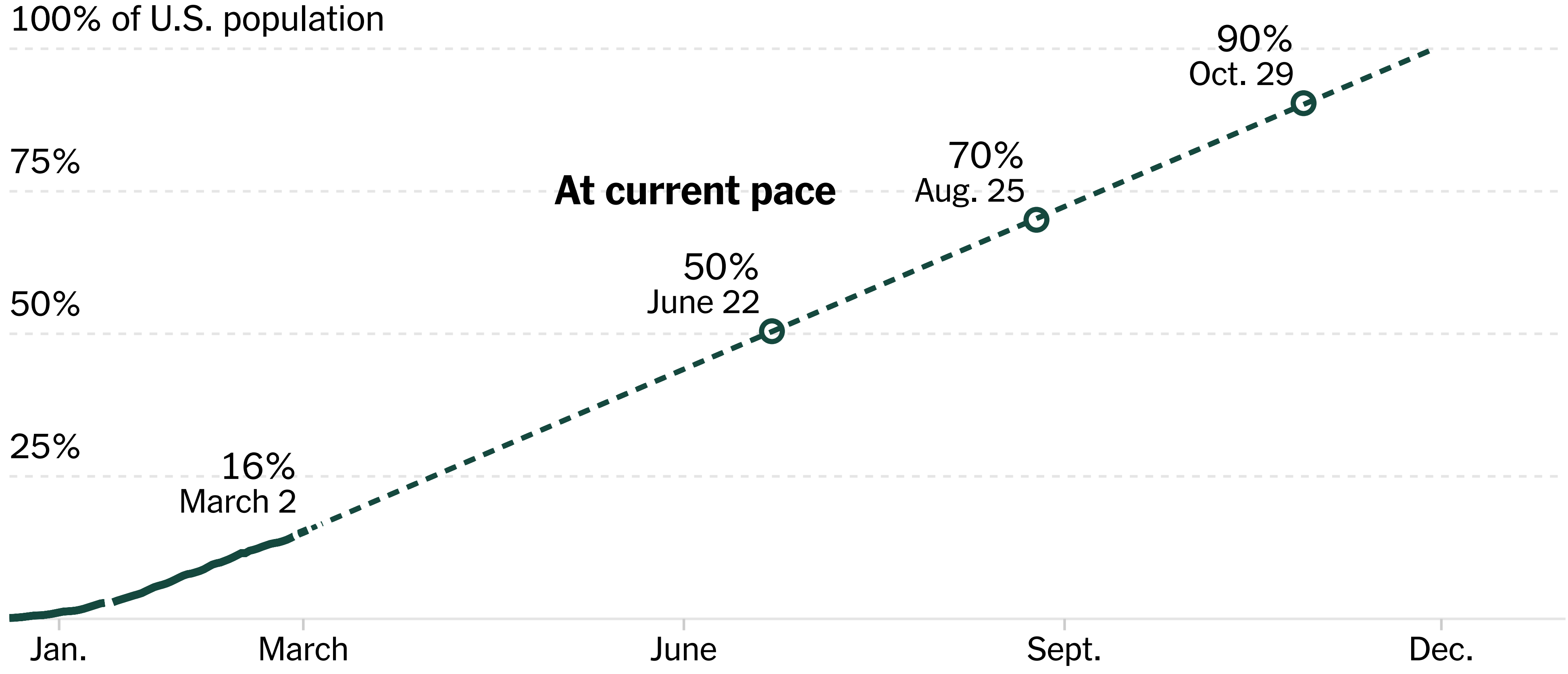
Measurements are observed…

… but estimates are *not* observed measures — *they are modeled from measures* — be clear about distinguishing them with words, encodings, and annotations.

The projection below only shows the share of the total population with at least one shot based on the current rate of vaccination, but it provides a rough indication of when the virus's spread could begin to stall.

When a given share of the U.S. population might be at least partially vaccinated
The current vaccination rate is based on average daily increase in first doses administered over the past week.

Average daily first doses in last 7 days: 1,030,068



Source: Centers for Disease Control and Prevention | Note: Data from Dec. 20 to Jan. 12 are for all doses administered. Data for Jan. 13 is unavailable. Projections could change if additional vaccines are authorized.

If the country maintains its current pace of administering first doses, about half of the total population would be at least partially vaccinated around late June, and nearly all around late October, assuming supply pledges are met and vaccines are eventually available to children.

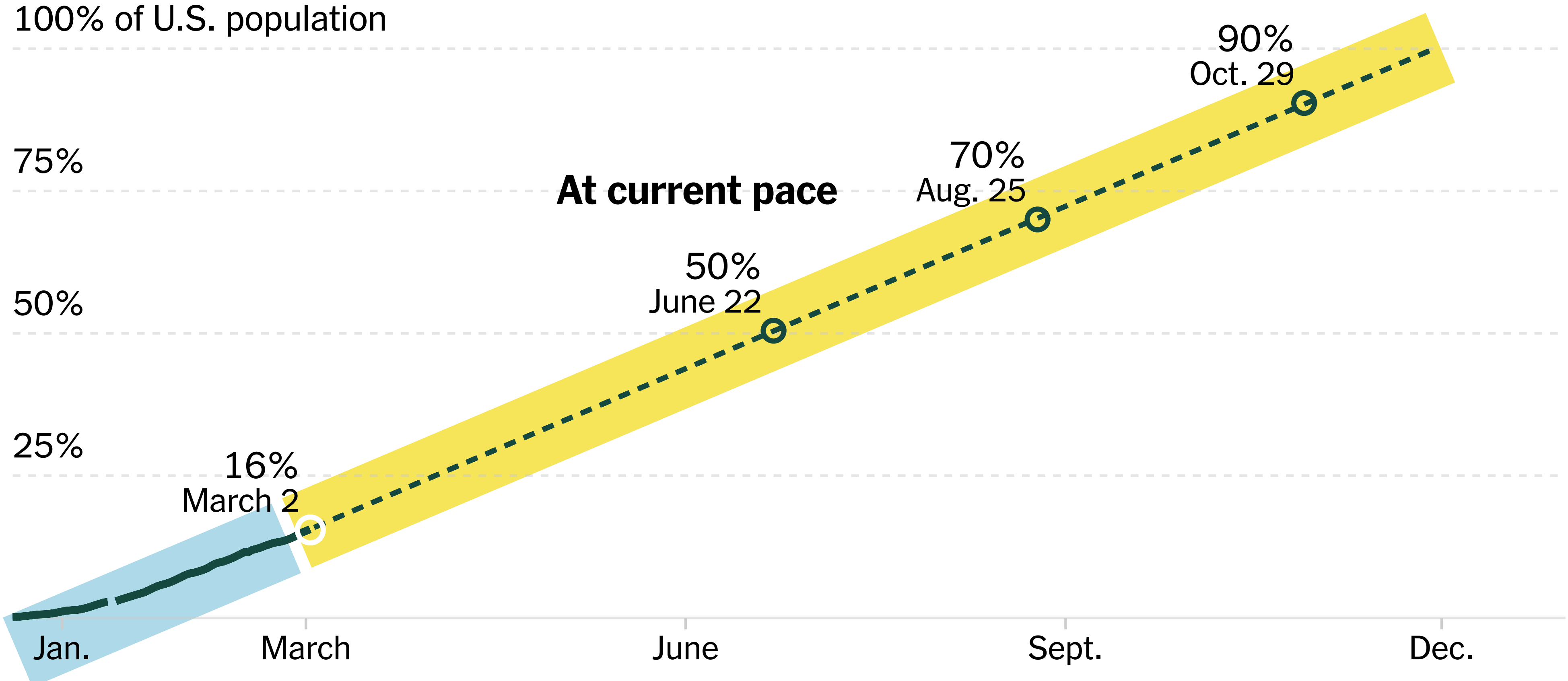
New York Times. "*Covid-19 Vaccine Rollout: State by State*." News, March 1, 2021. https://www.nytimes.com/interactive/2020/us/covid-19-vaccine-doses.html.

The projection below only shows the share of the total population with at least one shot based on the current rate of vaccination, but it provides a rough indication of when the virus's spread could begin to stall.

When a given share of the U.S. population might be at least partially vaccinated

The current vaccination rate is based on average daily increase in first doses administered over the past week.

Average daily first doses in last 7 days: 1,030,068



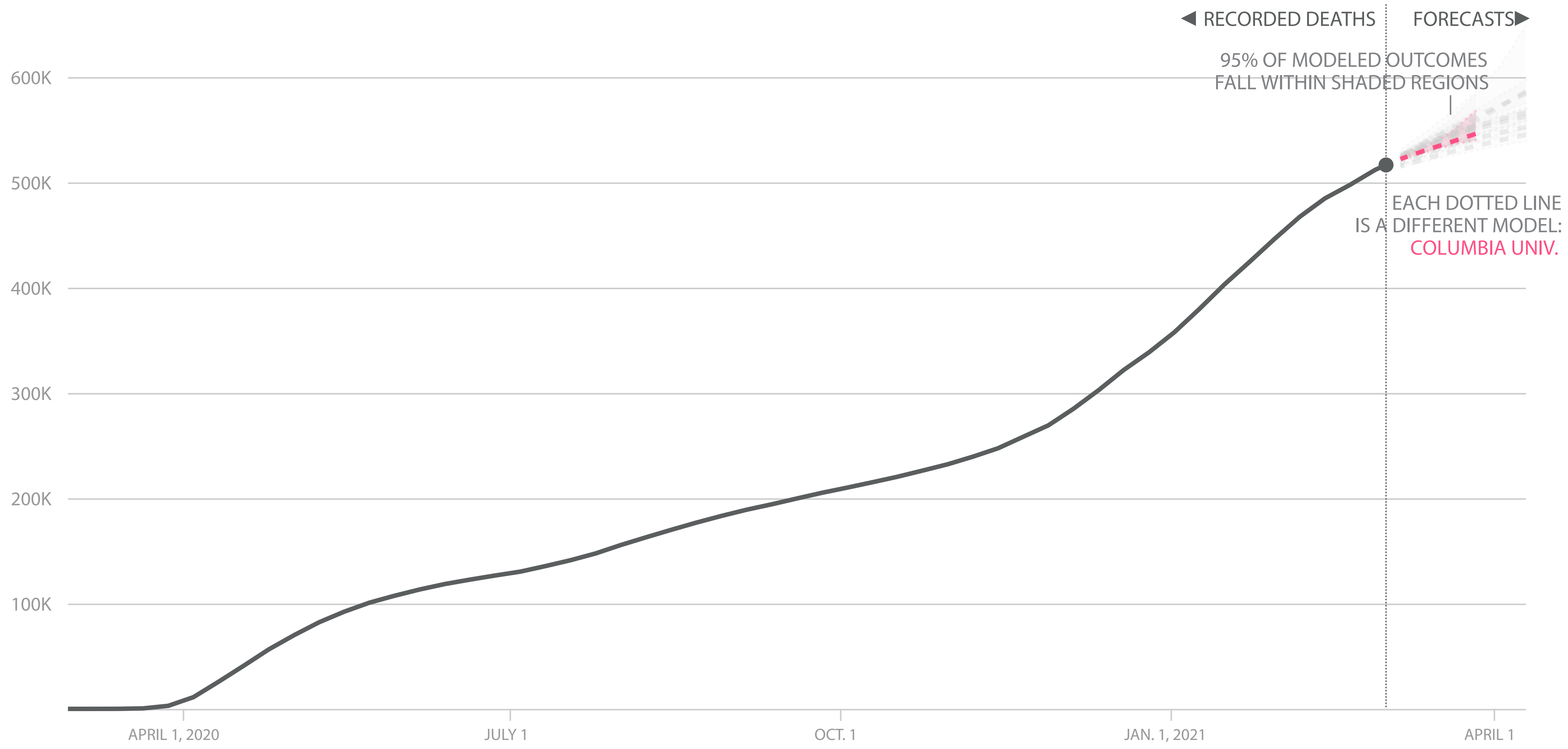
Source: Centers for Disease Control and Prevention | Note: Data from Dec. 20 to Jan. 12 are for all doses administered. Data for Jan. 13 is unavailable. Projections could change if additional vaccines are authorized.

If the country maintains its current pace of administering first doses, about half of the total population would be at least partially vaccinated around late June, and nearly all around late October, assuming supply pledges are met and vaccines are eventually available to children.

New York Times. "*Covid-19 Vaccine Rollout: State by State*." News, March 1, 2021. https://www.nytimes.com/interactive/2020/us/covid-19-vaccine-doses.html.

Models predicting the potential spread of the COVID-19 pandemic have become a fixture of American life. Yet each model tells a different story about the loss of life to come, making it hard to know which one is "right." But COVID-19 models aren't made to be unquestioned oracles. They're not trying to tell us one precise future, but rather the range of possibilities given the facts on the ground.

One of their more sober tasks is predicting the number of Americans who will die due to COVID-19. FiveThirtyEight — with the help of data compiled by the COVID-19 Forecast Hub — has assembled 11 models published by scientists to illustrate possible trajectories of the pandemic's death toll. In doing so, we hope to make them more accessible, as well as highlight how the assumptions underlying the models can lead to vastly different estimates. Here are the models' U.S. fatality projections for the coming weeks.



Forecasts like these are useful because they help us understand the most likely outcomes as well as best- and worst-case possibilities — and they can help policymakers make decisions that can lead us closer to those best-case outcomes.

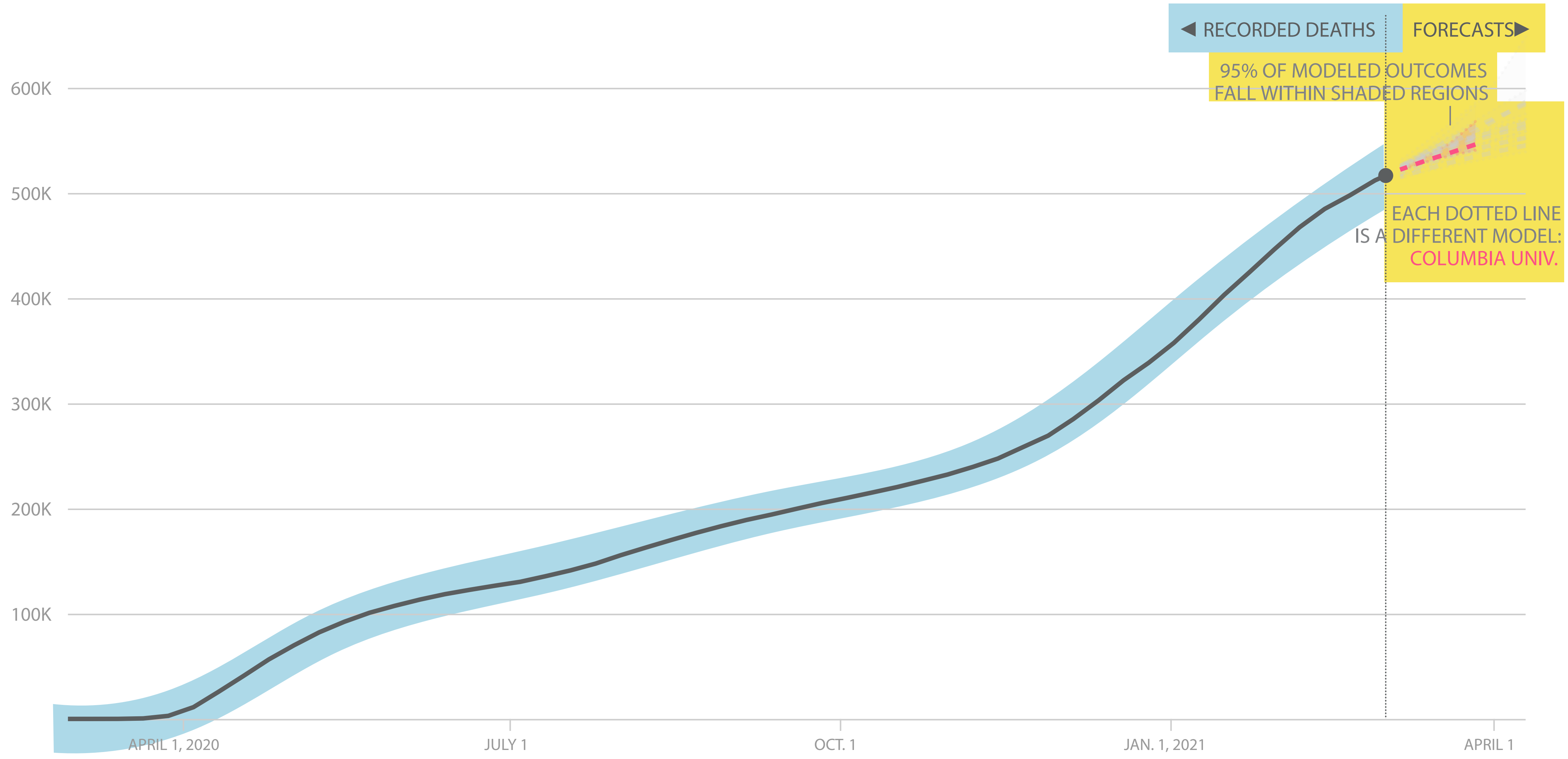
And looking at multiple models is better than looking at just one because it's difficult to know which model will match reality the closest. Even when models disagree, understanding why they are dierent can give us valuable insight.

Best, Ryan, and Jay Boice. "*Where The Latest COVID-19 Models Think We're Headed — And Why They Disagree.*" News. FiveThirtyEight, March 2, 2021. https://projects.fivethirtyeight.com/covid-forecasts/.

encoding uncertainty, estimates, forecasts, distinguishing measurements from estimates — examples

Models predicting the potential spread of the COVID-19 pandemic have become a fixture of American life. Yet each model tells a different story about the loss of life to come, making it hard to know which one is "right." But COVID-19 models aren't made to be unquestioned oracles. They're not trying to tell us one precise future, but rather the range of possibilities given the facts on the ground.

One of their more sober tasks is predicting the number of Americans who will die due to COVID-19. FiveThirtyEight — with the help of data compiled by the COVID-19 Forecast Hub — has assembled 11 models published by scientists to illustrate possible trajectories of the pandemic's death toll. In doing so, we hope to make them more accessible, as well as highlight how the assumptions underlying the models can lead to vastly different estimates. Here are the models' U.S. fatality projections for the coming weeks.



Forecasts like these are useful because they help us understand the most likely outcomes as well as best- and worst-case possibilities — and they can help policymakers make decisions that can lead us closer to those best-case outcomes.

And looking at multiple models is better than looking at just one because it's difficult to know which model will match reality the closest. Even when models disagree, understanding why they are dierent can give us valuable insight.

Best, Ryan, and Jay Boice. "*Where The Latest COVID-19 Models Think We're Headed — And Why They Disagree.*" News. FiveThirtyEight, March 2, 2021. https://projects.fivethirtyeight.com/covid-forecasts/.

> 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?

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:

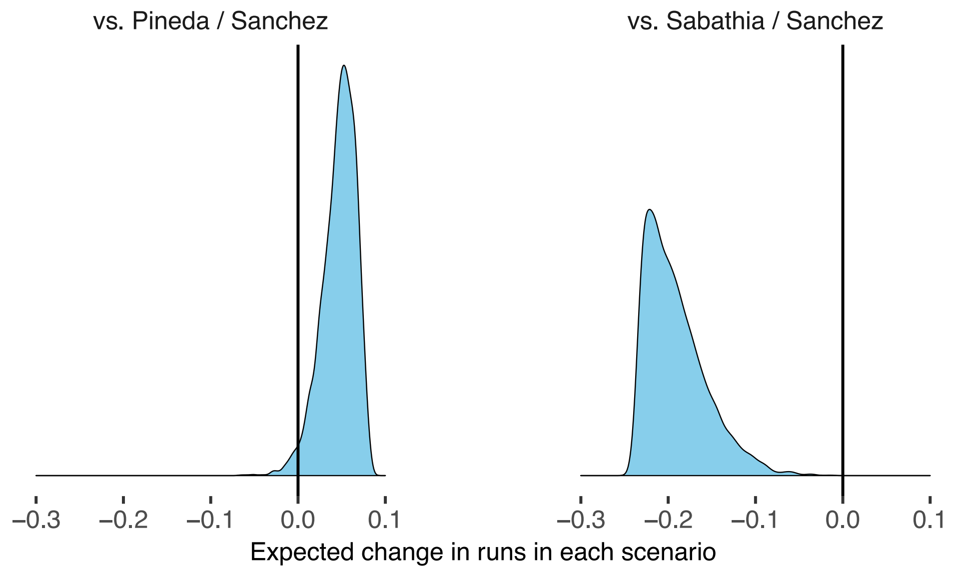


Figure 1. Of the two scenarios, Cain should only attempt to steal against the Sanchez–Pineda duo.

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:

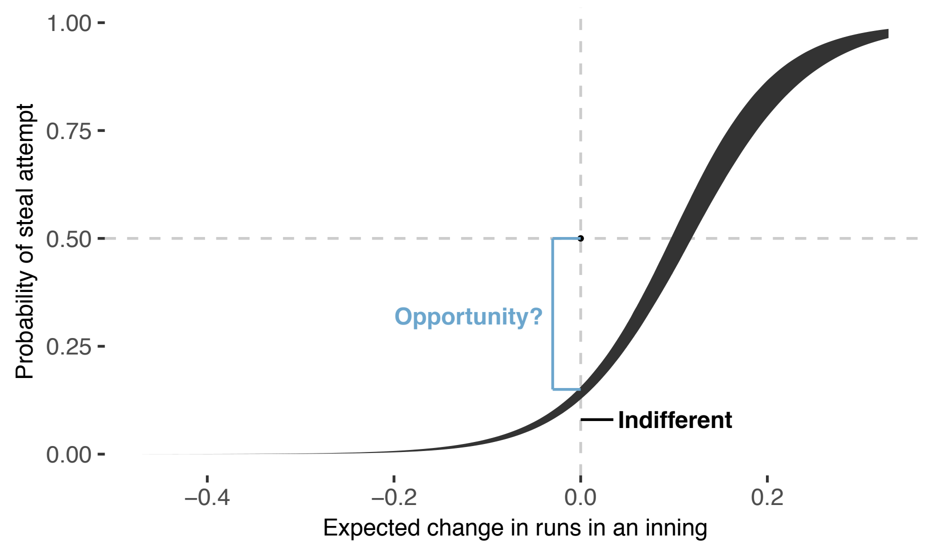


Figure 2. When the change in expected runs is zero, managers should be indifferent to attempted steals, saying go half the time.

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.

Spencer, Scott. Proposal to Scott Powers. "*Proposal for Exploring Game Decisions Informed by Expectations of Joint Probability Distributions*." February 14, 2019.

> 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?

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:

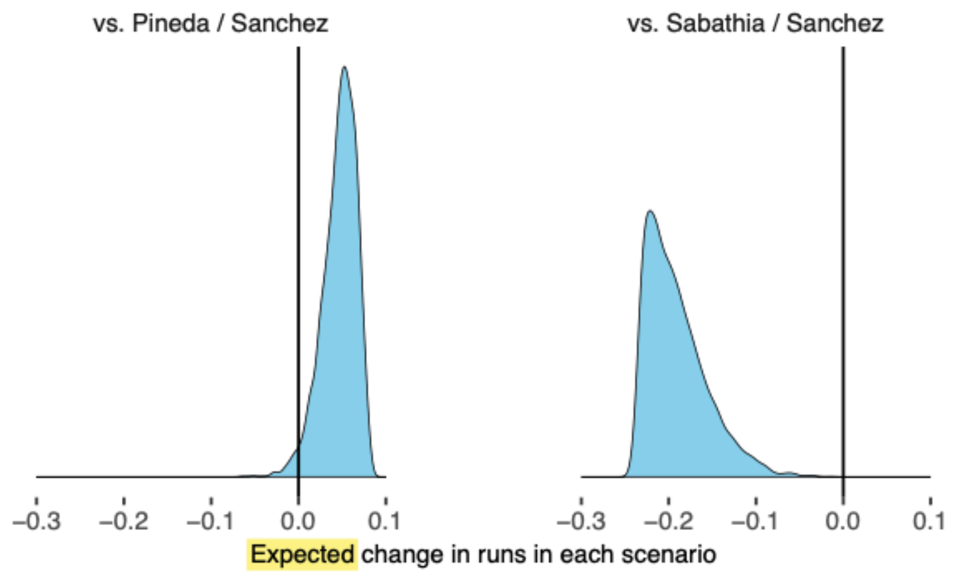


Figure 1. Of the two scenarios, Cain should only attempt to steal against the Sanchez–Pineda duo.

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:

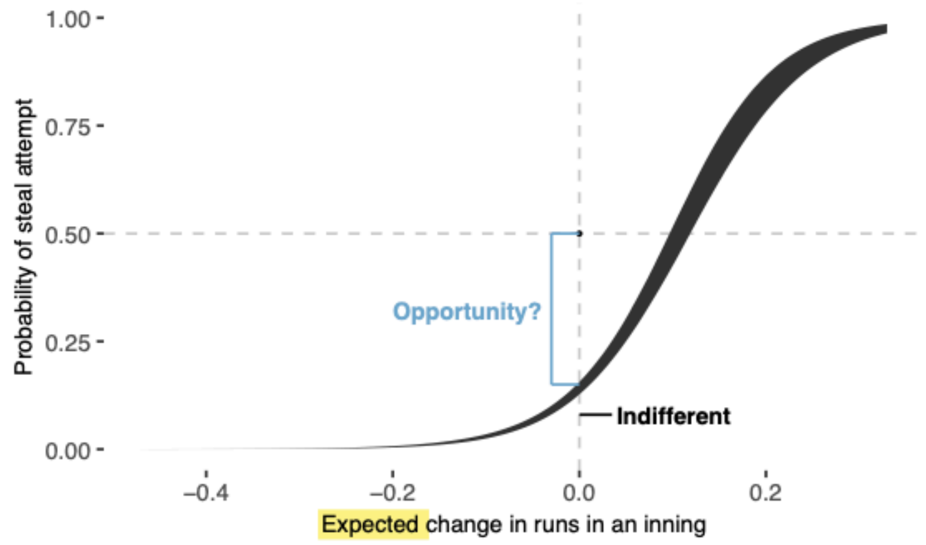


Figure 2. When the change in expected runs is zero, managers should be indifferent to attempted steals, saying go half the time.

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.

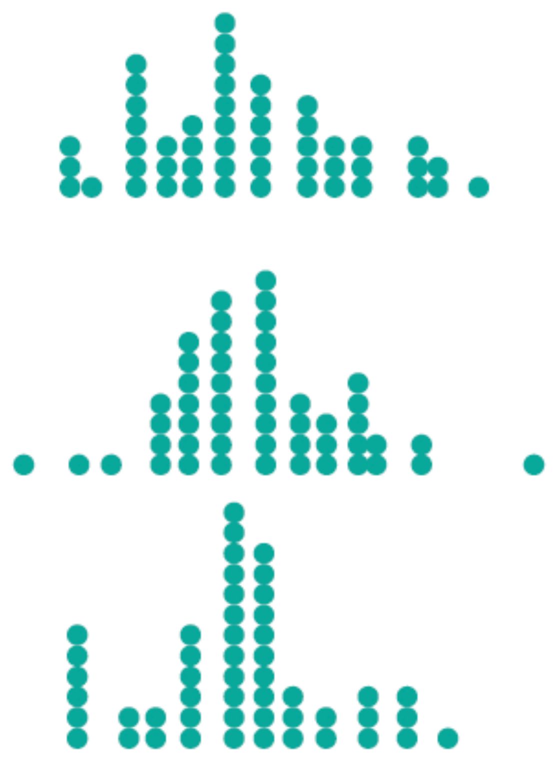
Spencer, Scott. Proposal to Scott Powers. "*Proposal for Exploring Game Decisions Informed by Expectations of Joint Probability Distributions*." February 14, 2019.

expressing uncertainty, estimates, forecasts — *recent* ideas for encoding

 discretizing distributions to improve decisions — quantile dot plots

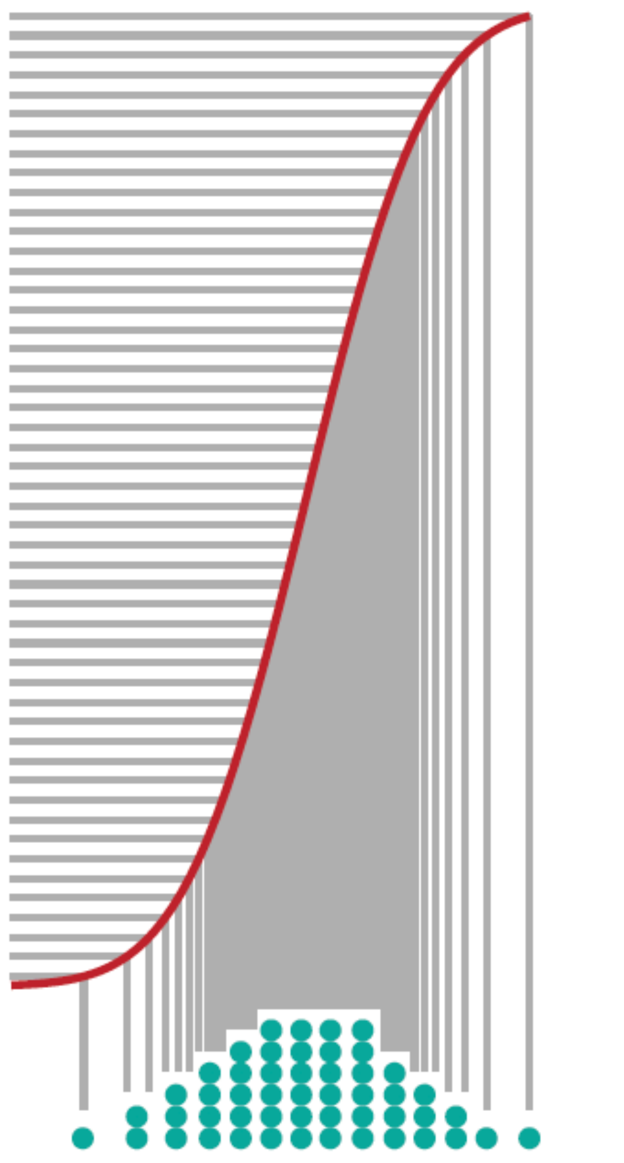
Probability density of Normal distribution

To generate a discrete plot of this distribution, we could try taking **random draws** from it. However, **this approach is noisy**: it may be very different from one instance to the next.



Probability density of Normal distribution

Instead, we use the **quantile function (inverse CDF)** of the distribution to generate "draws" from evenly-spaced quantiles.

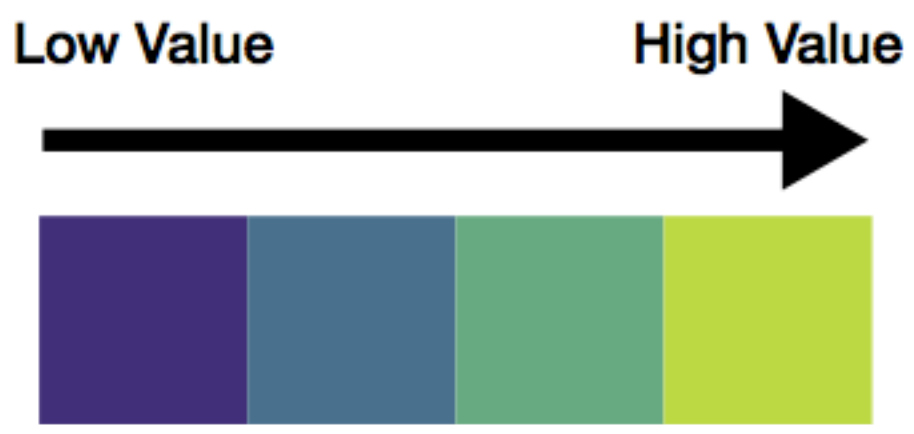


We plot the quantile "draws" using a Wilkinsonian dotplot, yielding what we call a **quantile dotplot**: a consistent discrete representation of a probability distribution.

By using quantiles we facilitate interval estimation from frequencies: e.g., knowing there are 50 dots here, if we are willing to miss our bus **3/50** times, we can count **3 dots** from the left to get a one-sided **94% (1 – 3/50) prediction interval** corresponding to that risk tolerance.

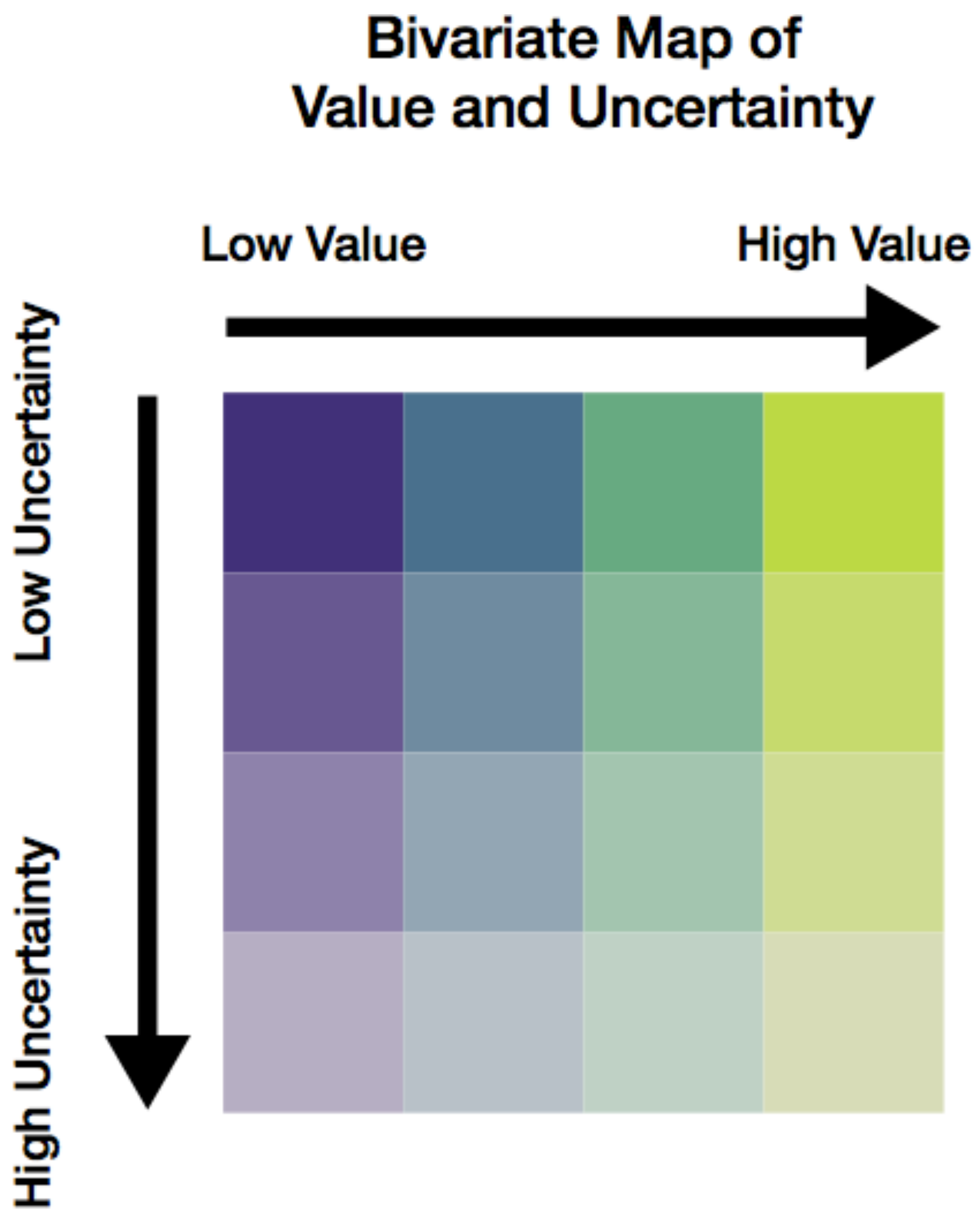
Fernandes, M., Walls, L., Munson, S., Hullman, J., & Kay, M. (2018). *Uncertainty Displays Using Quantile Dotplots or CDFs Improve Transit Decision-Making.*Â Conference on Human Factors in Computing Systems - CHI '18. doi:Â 10.1145/3173574.3173718

 using color to encode uncertainty — value suppressing uncertainty palettes



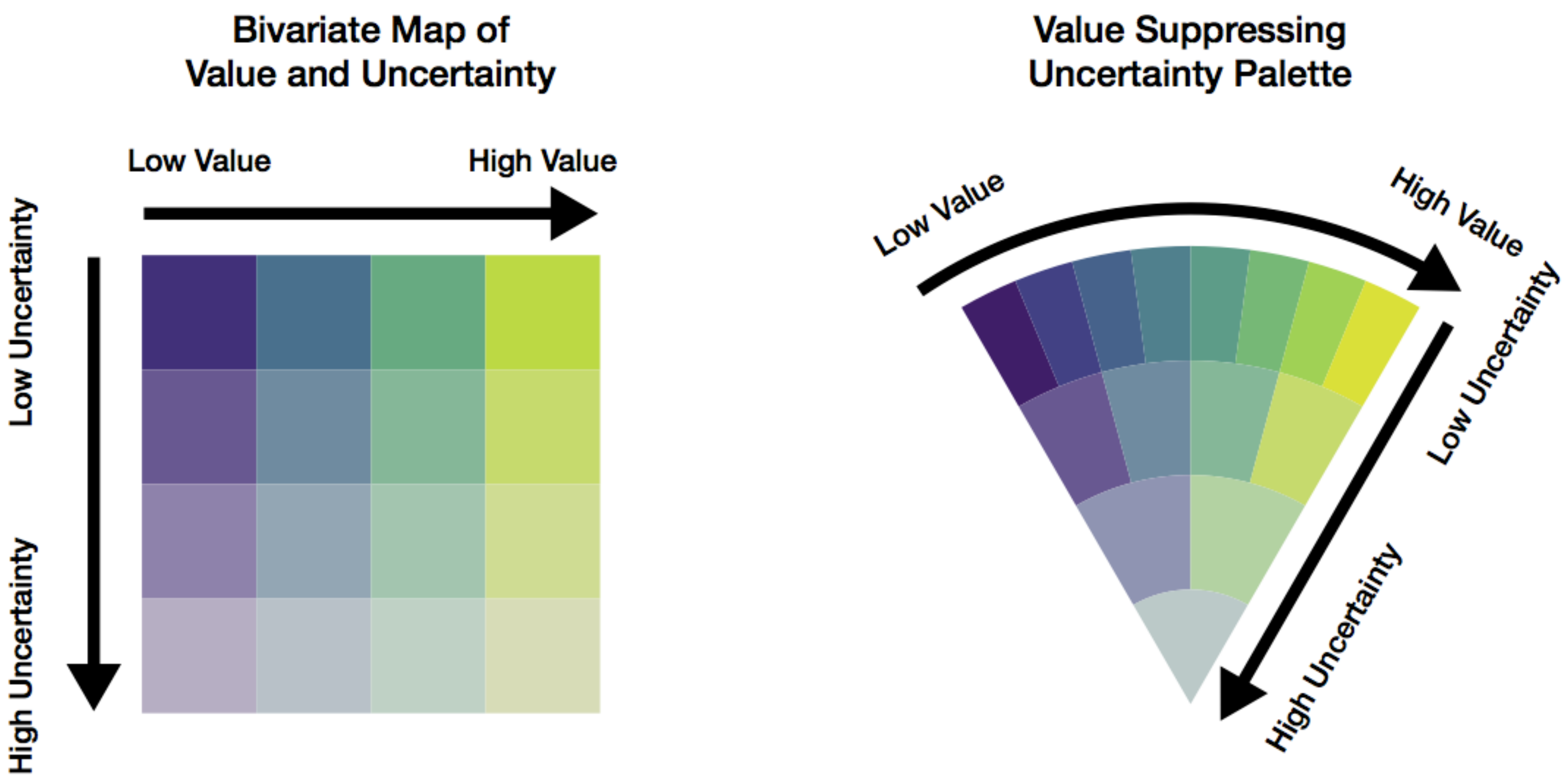
Correll, Michael, Dominik Moritz, and Jeffrey Heer. "Value-Suppressing Uncertainty Palettes." In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems - CHI '18*, 1–11. Montreal QC, Canada: ACM Press, 2018.

recent ideas for encoding, **using color to encode uncertainty — value suppressing uncertainty palettes**



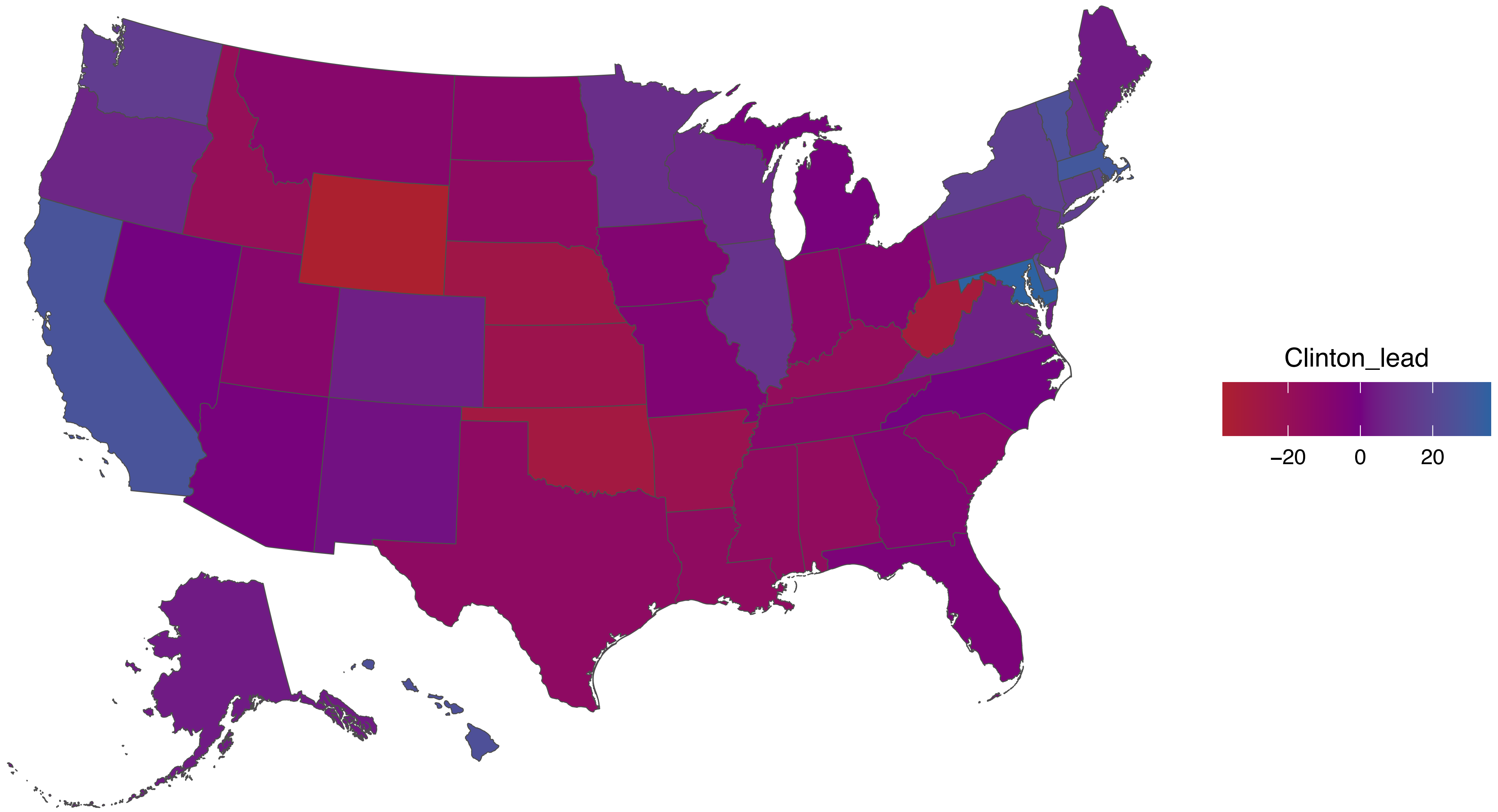
Correll, Michael, Dominik Moritz, and Jeffrey Heer. "Value-Suppressing Uncertainty Palettes." In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems - CHI '18*, 1–11. Montreal QC, Canada: ACM Press, 2018.

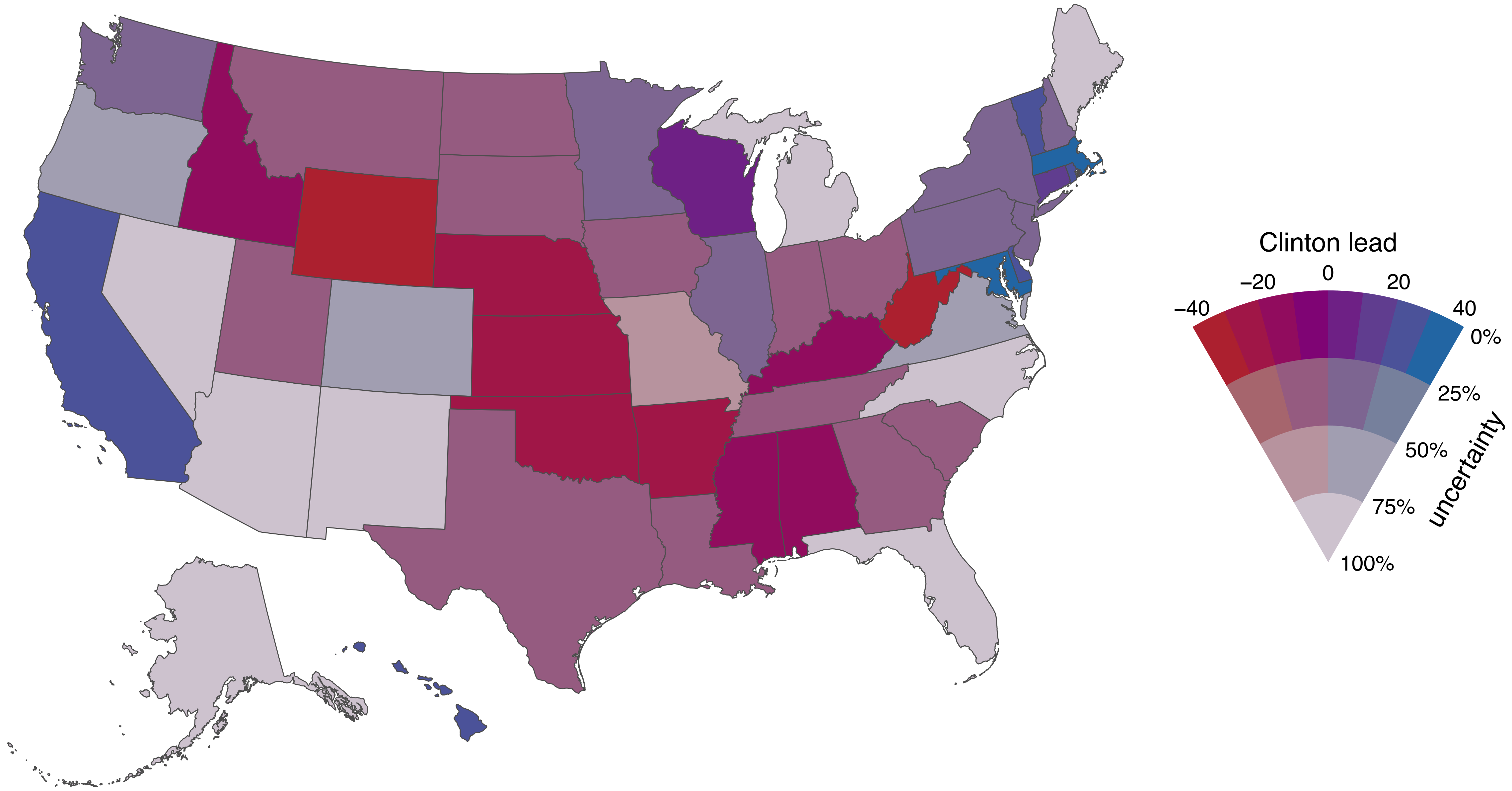
 using color to encode uncertainty — value suppressing uncertainty palettes



Correll, Michael, Dominik Moritz, and Jeffrey Heer. "Value-Suppressing Uncertainty Palettes." In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems - CHI '18*, 1–11. Montreal QC, Canada: ACM Press, 2018.

recent ideas for encoding, using color to encode uncertainty — value suppressing uncertainty palettes

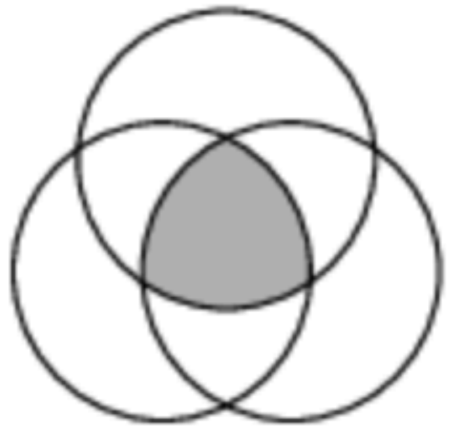
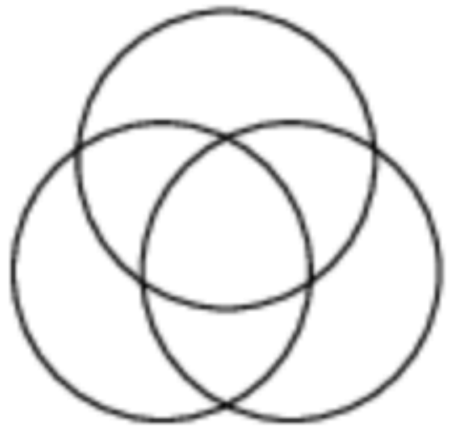
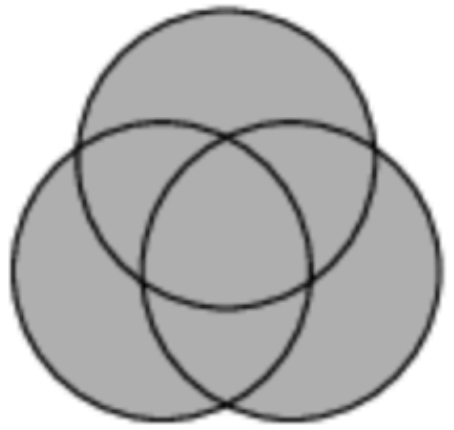
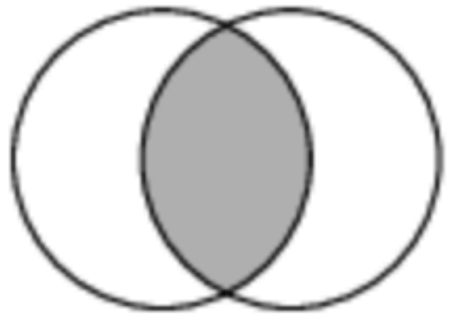
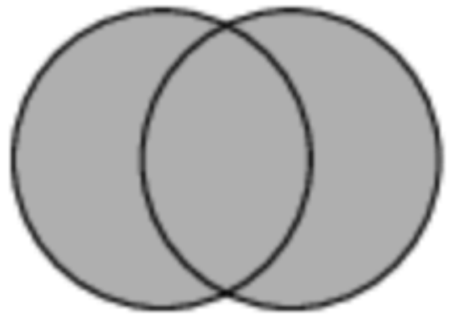




Claus O. Wilke (2021). *multiscales: Multivariate scales for 'ggplot2'*. R package version 0.1.0. https://github.com/clauswilke/multiscales

expressing uncertainty by joining
***grammars* of graphics *and* probability**

graphics + probability grammars, **a grammar for expressing probability, examples in notation**

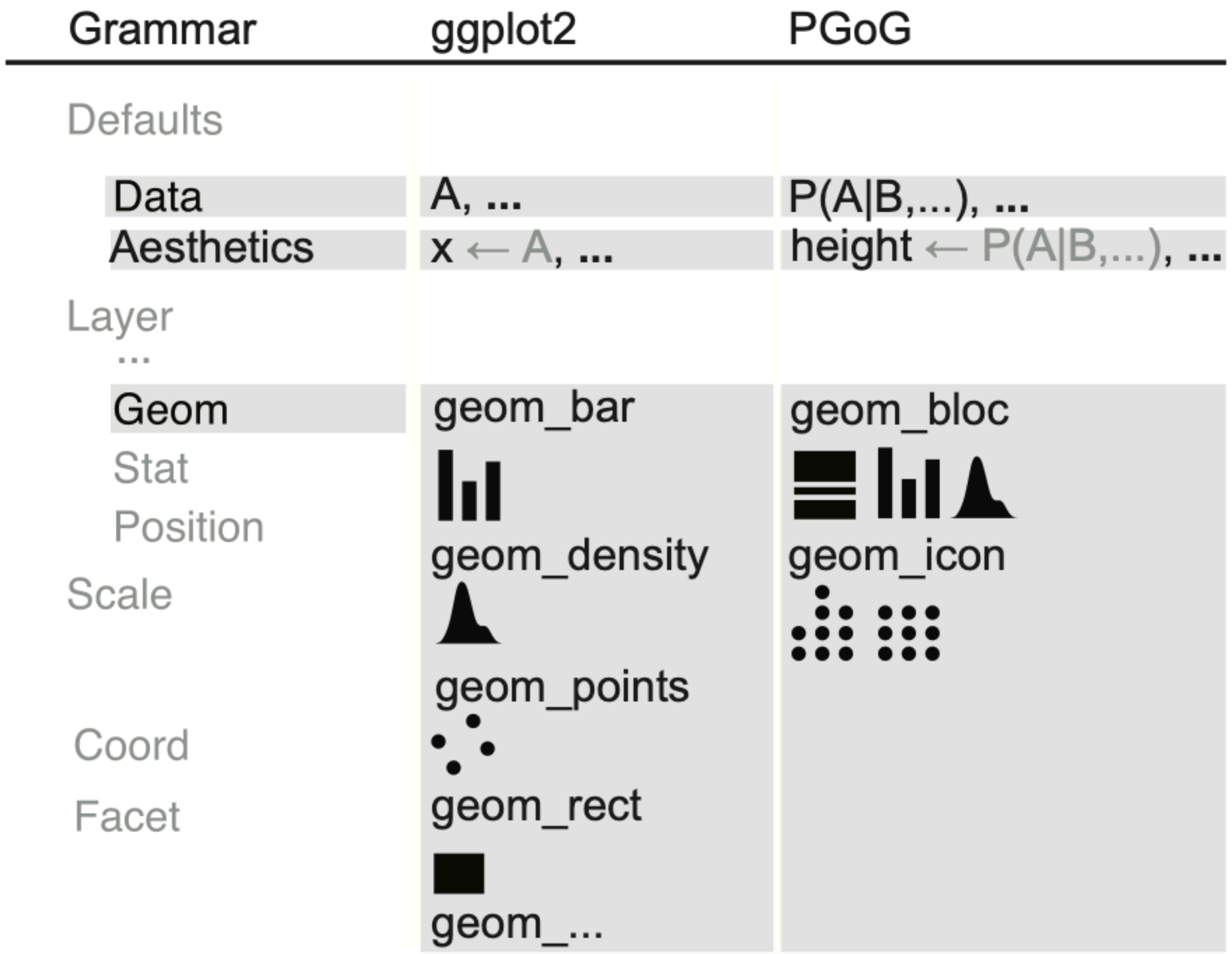
Words	Symbols	Venn diagram
“all”	A and B and C / $A \cap B \cap C$	
“none”		
“at least one”	A or B or C / $A \cup B \cup C$	
“both A and B”	A and B / $A \cap B$	
“A or B”	A or B / $A \cup B$	

$P(A)$ **marginal** ***probability*** **of event A**

$P(A \mid B, C)$ **conditional** ***probability*** **of A *given* B and C**

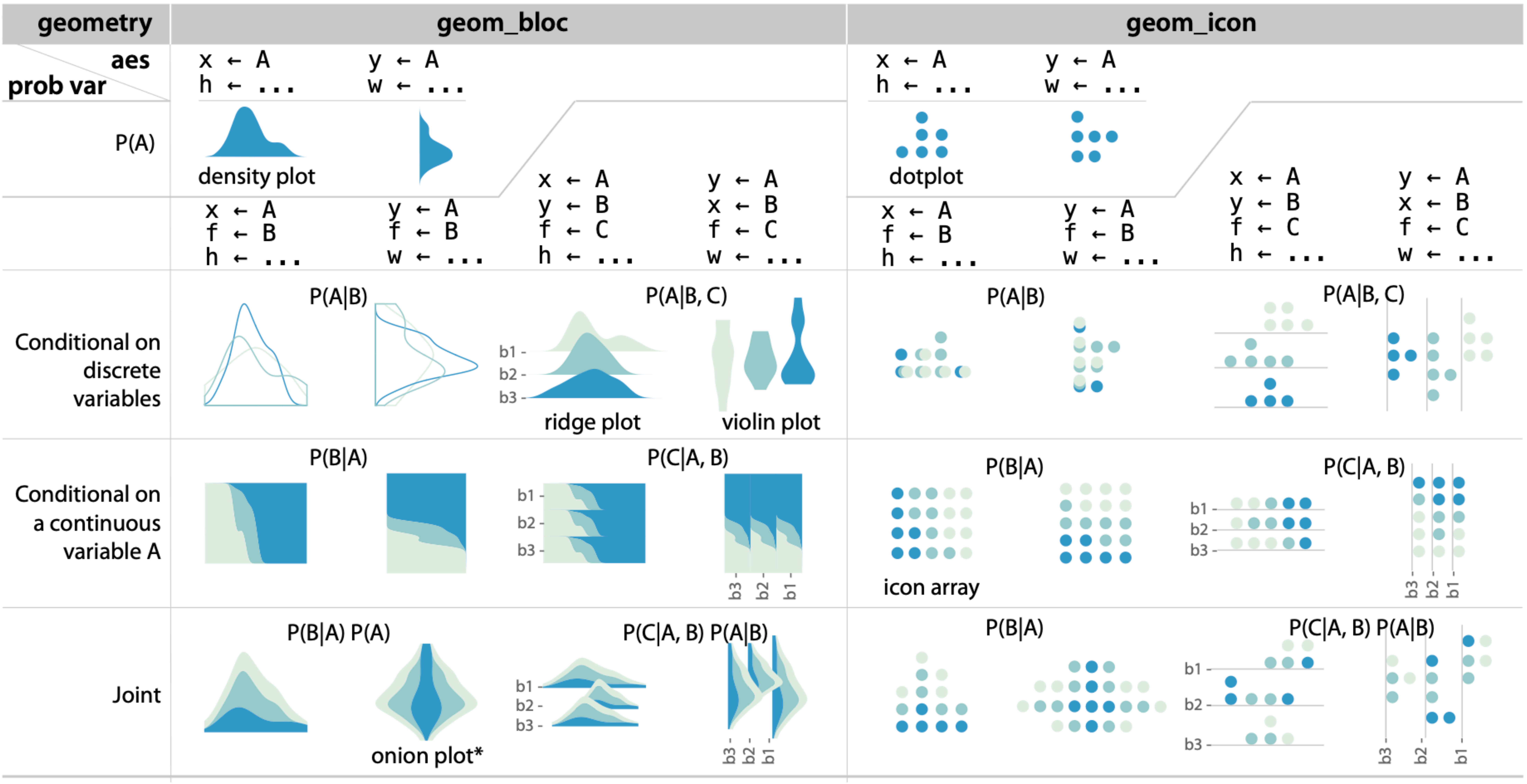
$P(A \mid B)P(B)$ **joint** ***probability*** **of A *and* B**

 joining the layered grammar of graphics with expressions for probability



Kay, Matthew. 2020. *ggdist: Visualizations of Distributions and Uncertainty*. R package version 2.4.0, https://mjskay.github.io/ggdist/.

graphics + probability grammars, example (partial) implementation using `ggdist`



Kay, Matthew. 2020. *ggdist: Visualizations of Distributions and Uncertainty*. R package version 2.4.0, https://mjskay.github.io/ggdist/.

 applied to class example — CitiBike rebalancing study

resources

References

Spencer, Scott. Sec. 1.1.1.2 and 2.3 In *Data in Wonderland*. 2021. https://ssp3nc3r.github.io/data_in_wonderland.

2PI360. "*Scientific Visualization: Principles of Posterior Visualization*," February 2015. https://ctg2pi.wordpress.com/2015/02/24/principles-of-posterior-visualization/.

———. "*Scientific Visualization: Visualizing Uncertainty in Dynamic Variables*," June 2015. https://ctg2pi.wordpress.com/2015/06/23/visualizing-uncertainty-in-dynamic-variables/#more-119.

Barclay, Scott, Rex V Brown, Clinton W Kelly III, Cameron R Peterson, Lawrence D Phillips, and Judith Selvidge. "Handbook for Decision Analysis." Decisions and Designs, Inc., 1977.

Correll, Michael, Dominik Moritz, and Jeffrey Heer. 2018. "Value-Suppressing Uncertainty Palettes." In Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems - CHI '18, 1–11. Montreal QC, Canada: ACM Press. https://dl-acm-org.ezproxy.cul.columbia.edu/doi/10.1145/3173574.3174216.

Fischhoff, Baruch. *Communicating Uncertainty: Fulfilling the Duty to Inform*. Issues in Science and Technology 28, no. 4 (August 2012): 63–70.

Hullman, Jessica. *Confronting Unknowns: How to Interpret Uncertainty in Common Forms of Visualization*. Scientific American, September 2019.

———. "*Why Authors Don't Visualize Uncertainty*." IEEE Transactions on Visualization and Computer Graphics 26, no. 1 (January 2020): 130–39.

Kampourakis, Kostas, and Kevin McCain. *Uncertainty: How It Makes Science Advance*. New York: Oxford University Press, 2020.

Kay, Matthew. 2020. *ggdist: Visualizations of Distributions and Uncertainty*. R package version 2.4.0, https://mjskay.github.io/ggdist/.

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

Lupi, Giorgia. "Data Humanism: The Revolutionary Future of Data Visualization." In *PrintMag*, January 30, 2017. https://www.printmag.com/post/data-humanism-future-of-data-visualization..

Schwabish, Jonathan A. 2021. "Distribution." In *Better Data Visualizations: A Guide for Scholars, Researchers, and Wonks*. New York: Columbia University Press.

Song, Hayeong, and Danielle Albers Szafir. "*Where's My Data? Evaluating Visualizations with Missing Data*." IEEE Transactions on Visualization and Computer Graphics 25, no. 1 (September 2018): 914–24.

Spiegelhalter, David. "Risk and Uncertainty Communication." *Annual Review of Statistics and Its Application* 4, no. 1 (March 7, 2017): 31–60. https://doi.org/10.1146/annurev-statistics-010814-020148.

Wainer, Howard. "The Most Dangerous Equation." In *Picturing the Uncertain World*, 5–20. Princeton University Press, 2009.

Wilke, C. Visualizing Uncertainty. In *Fundamentals of Data Visualization: A Primer on Making Informative and Compelling Figures*. First edition. Sebastopol, CA: O'Reilly Media, 2019.

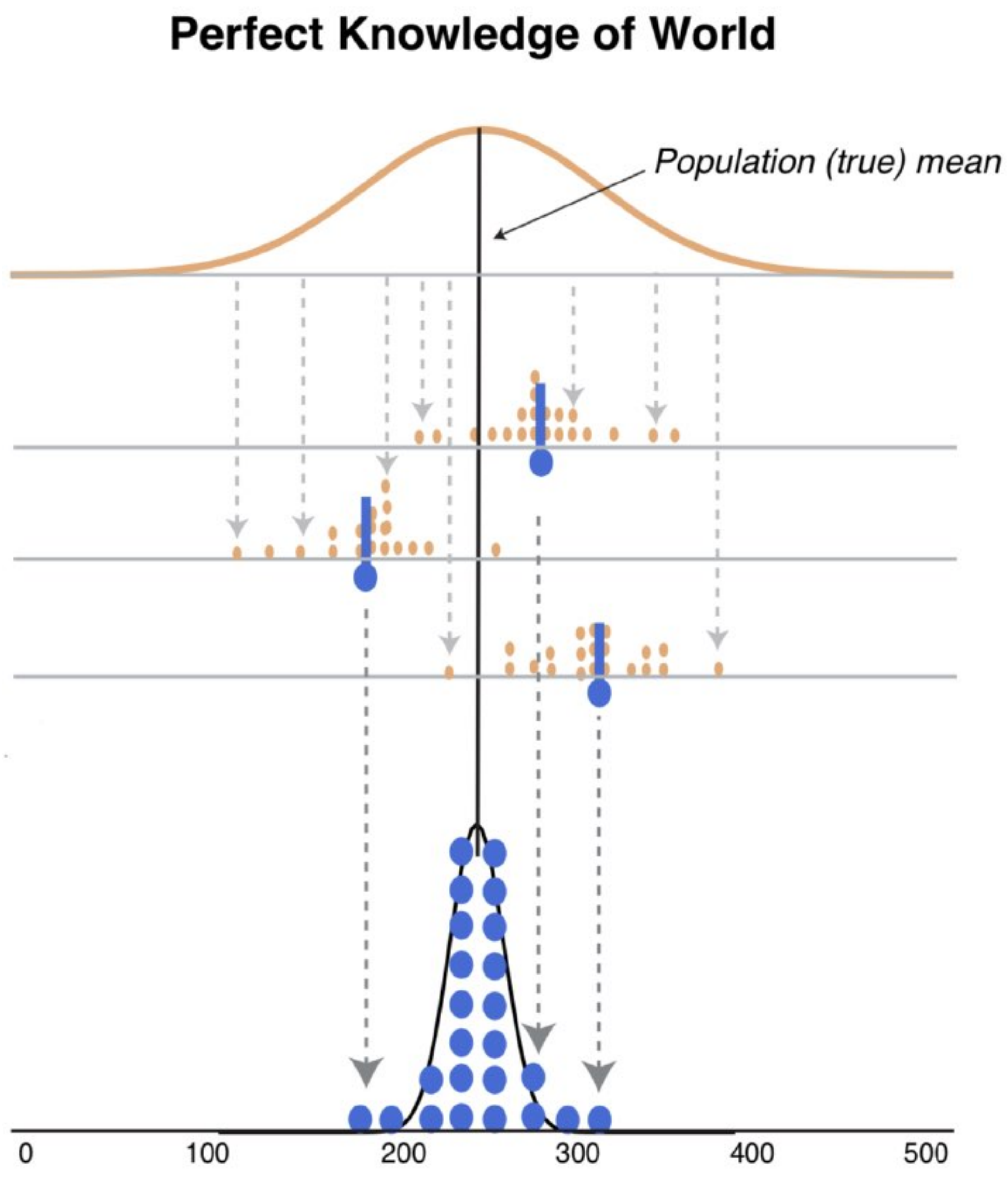
Wilkinson, Leland. "Ch. 15 Uncertainty." In *The Grammar of Graphics*, Second. Springer, 2005.

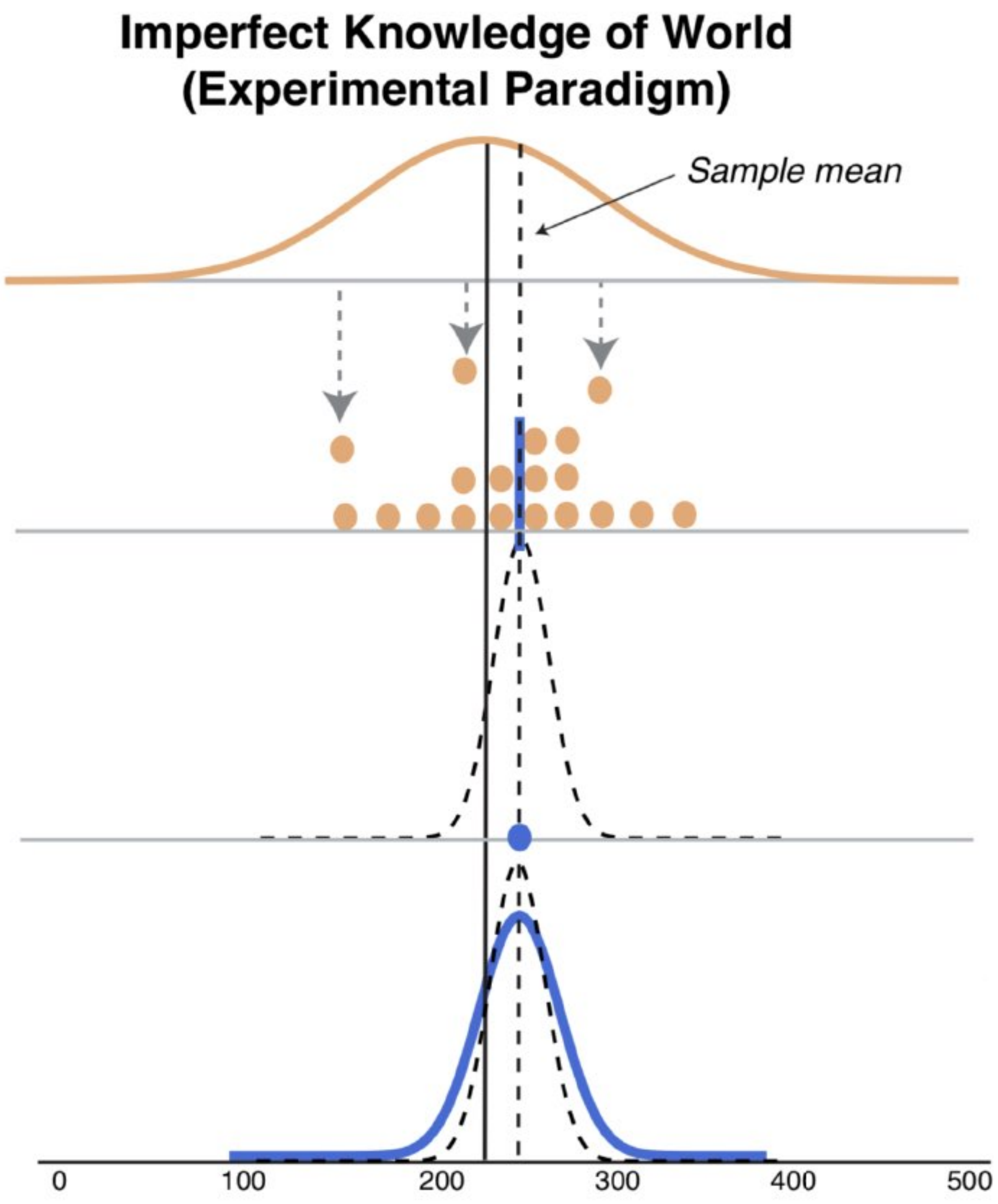
zonination. "Perceptions of Probability and Numbers," August 2015. https://github.com/zonination/perceptions.

supplemental

uncertainty

 sample distributions differ from the population for which we want to infer something

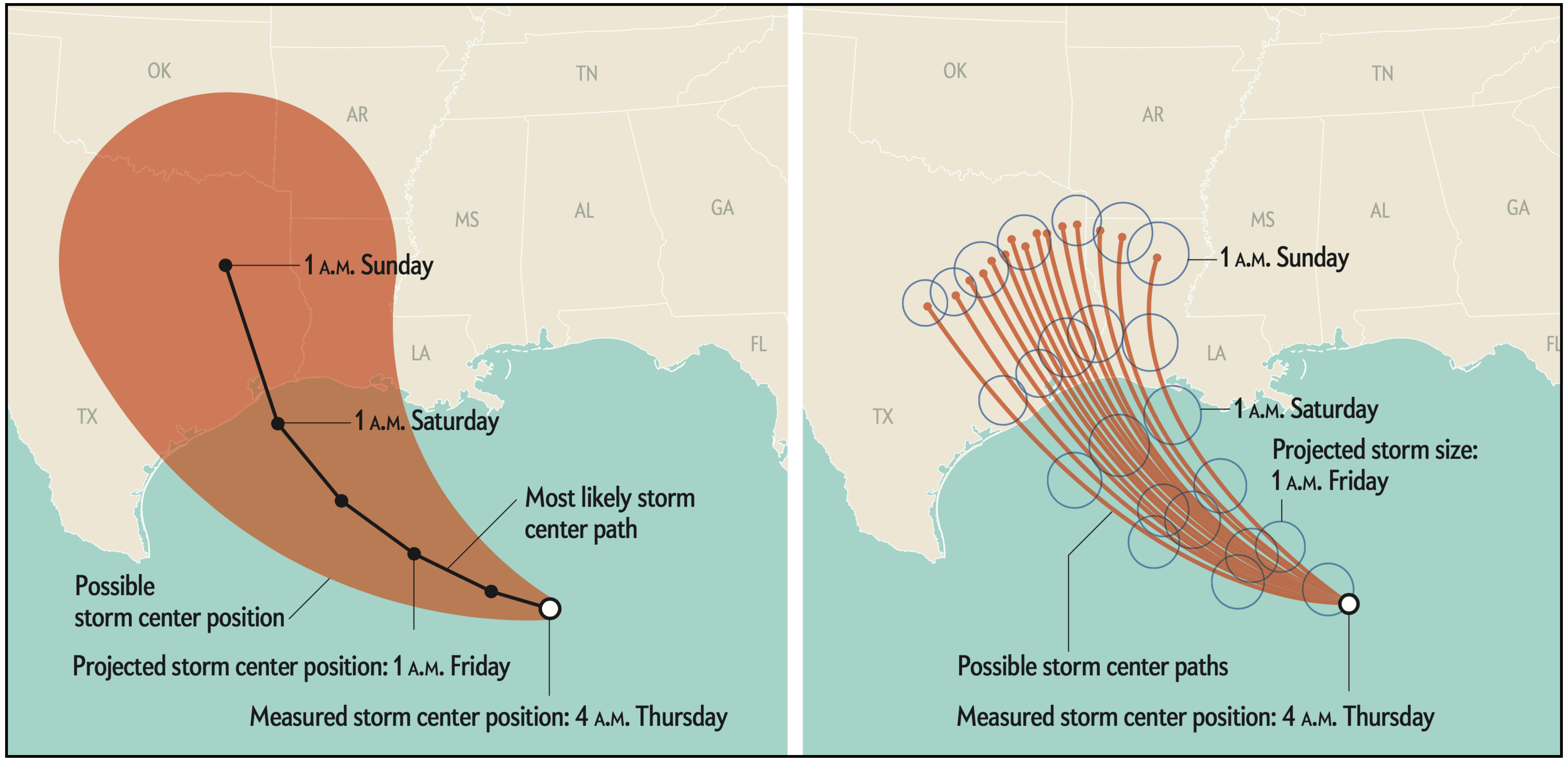




 | *consider alternative encodings and how perception may differ*

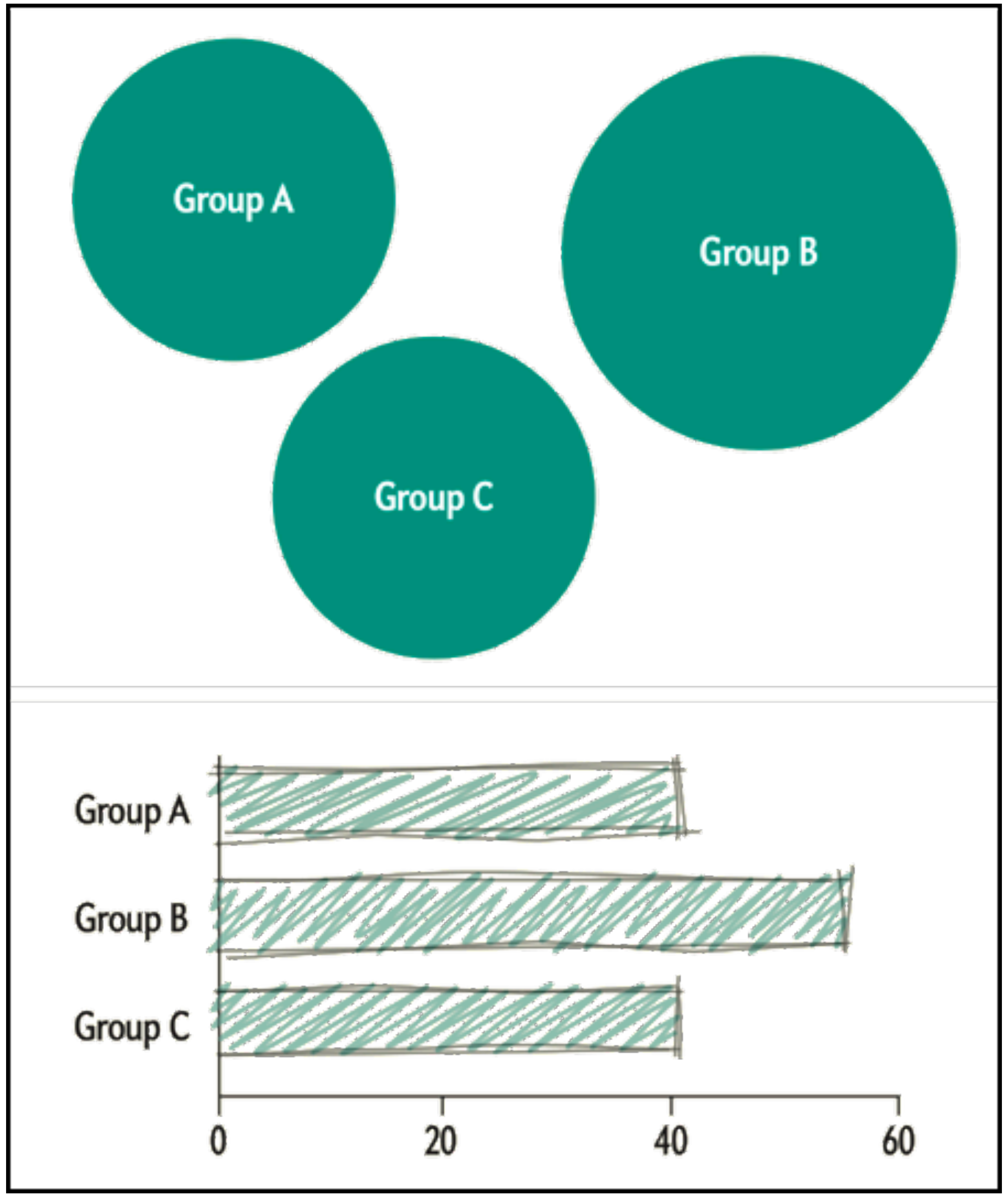
Uncertainty in storm path misperceived as growth in size

Alternative way to express uncertainty of storm path



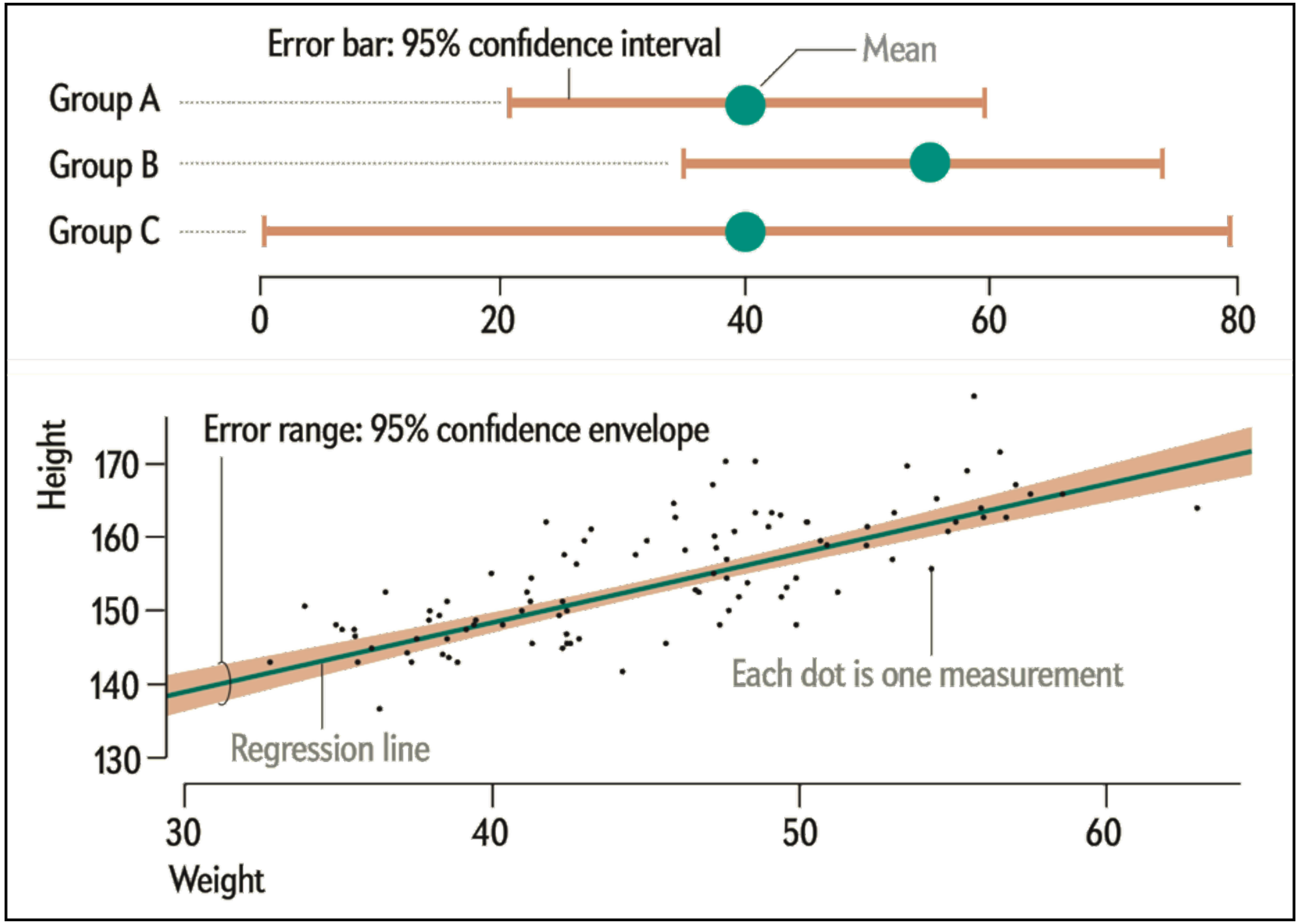
Hullman, Jessica. *Confronting Unknowns: How to Interpret Uncertainty in Common Forms of Visualization*. Scientific American, September 2019.

 no quantification occurs most — provides least information for decisions



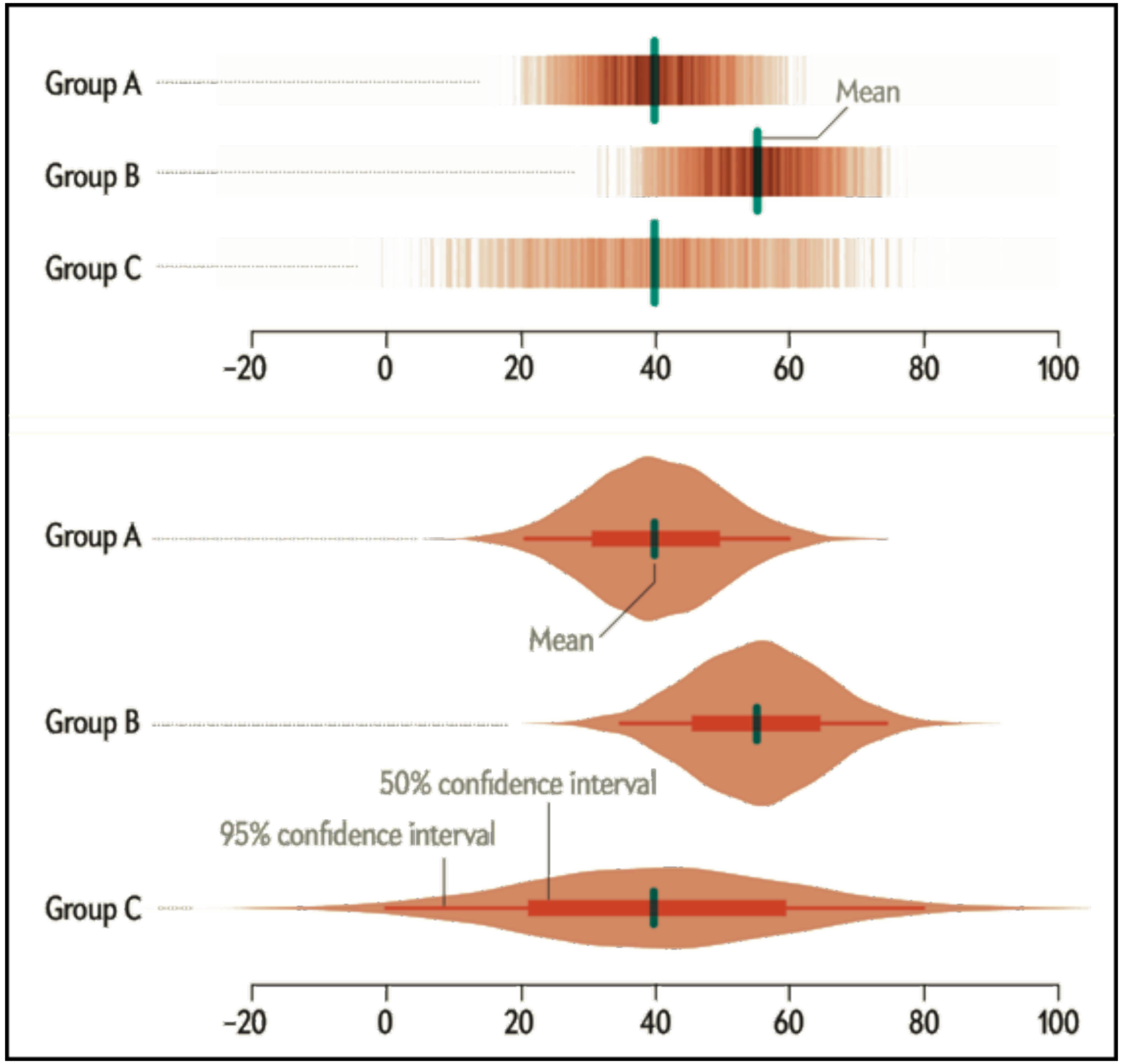
Hullman, Jessica. *Confronting Unknowns: How to Interpret Uncertainty in Common Forms of Visualization*. Scientific American, September 2019.

 | *intervals are perhaps the most common encodings for uncertainty*



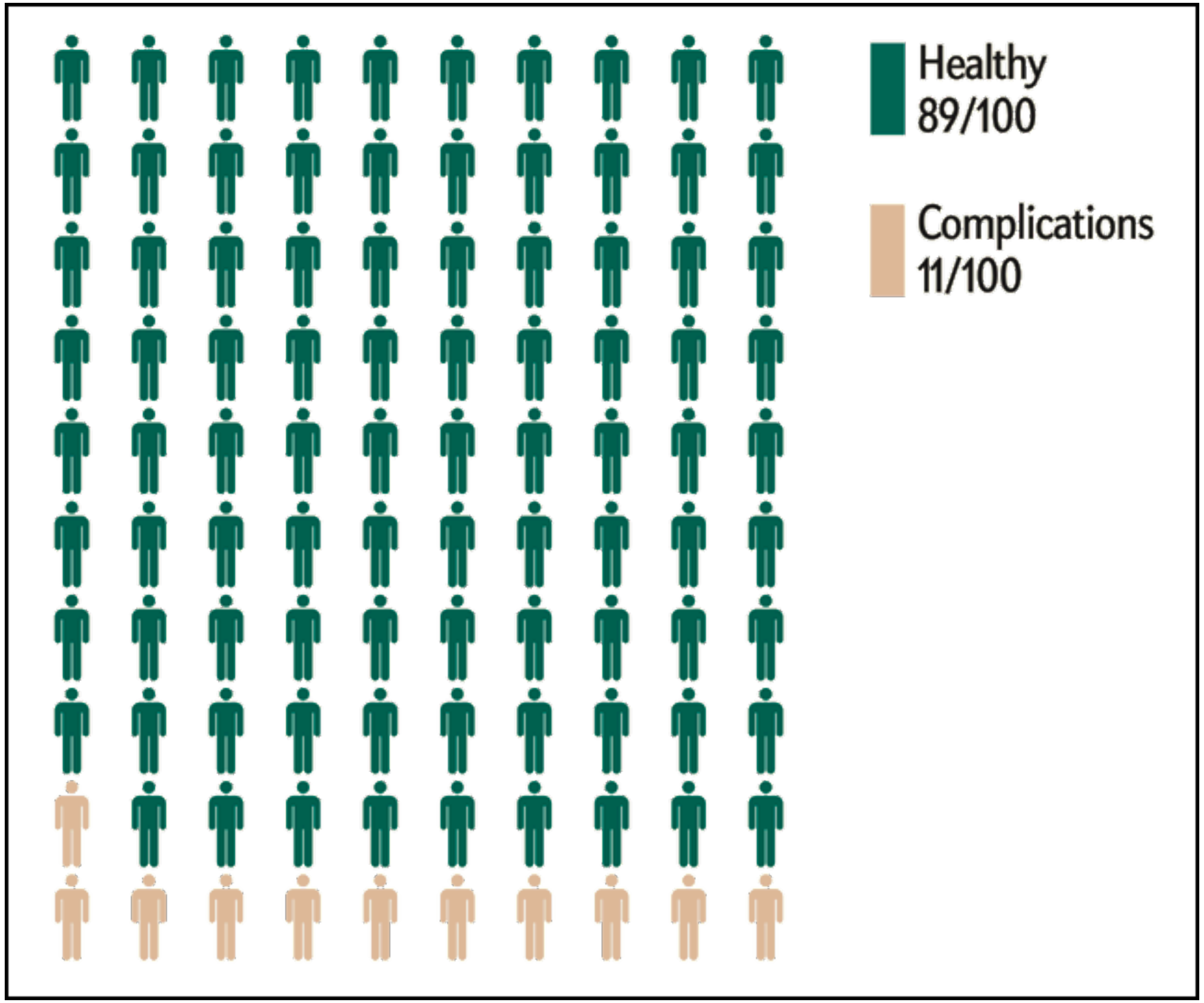
Hullman, Jessica. *Confronting Unknowns: How to Interpret Uncertainty in Common Forms of Visualization*. Scientific American, September 2019.

 | *probability densities tend to provide the most information about data*



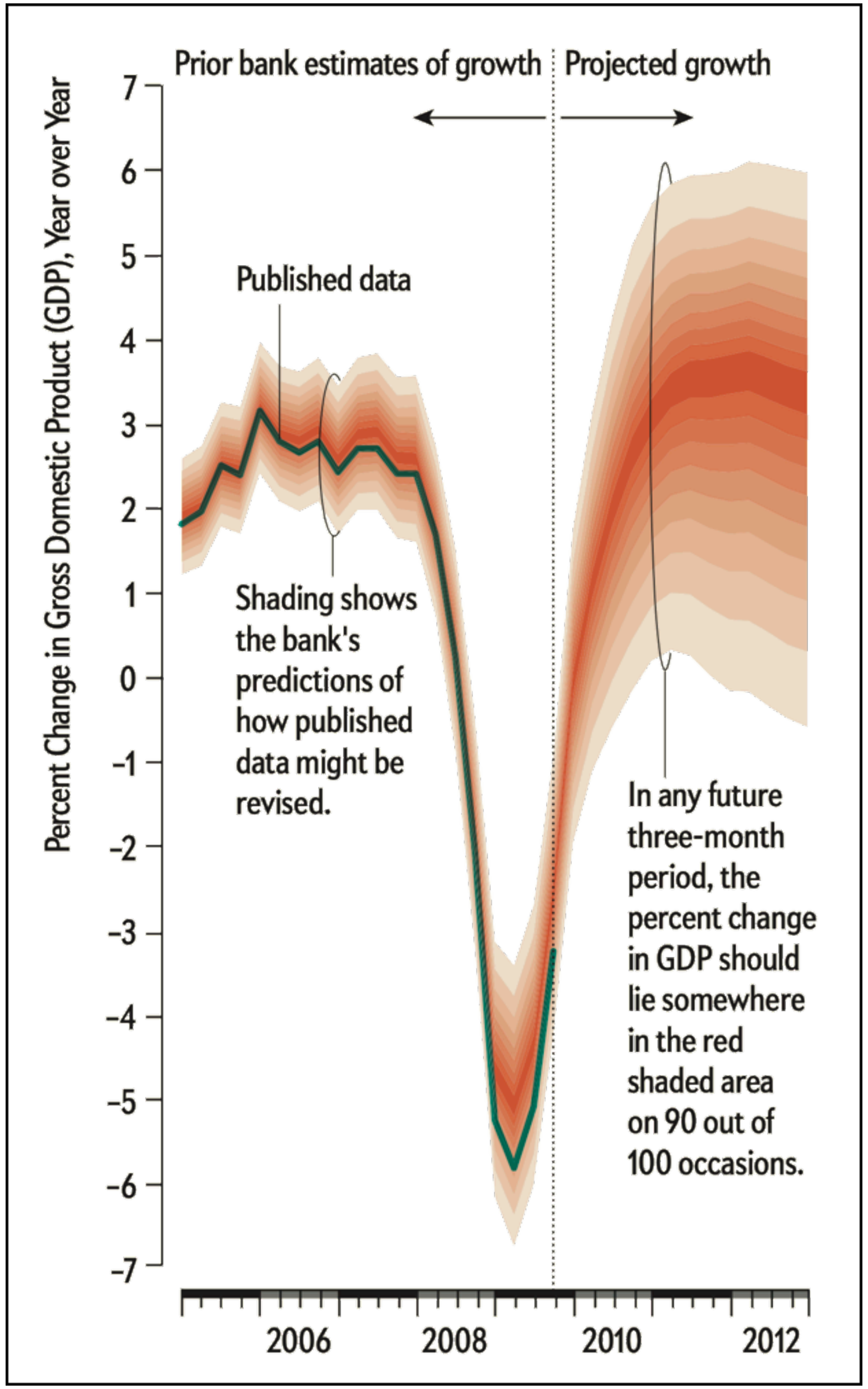
Hullman, Jessica. *Confronting Unknowns: How to Interpret Uncertainty in Common Forms of Visualization.* Scientific American, September 2019.

 | ***arrays of icons — people tend to think discretely, relate to familiar objects***



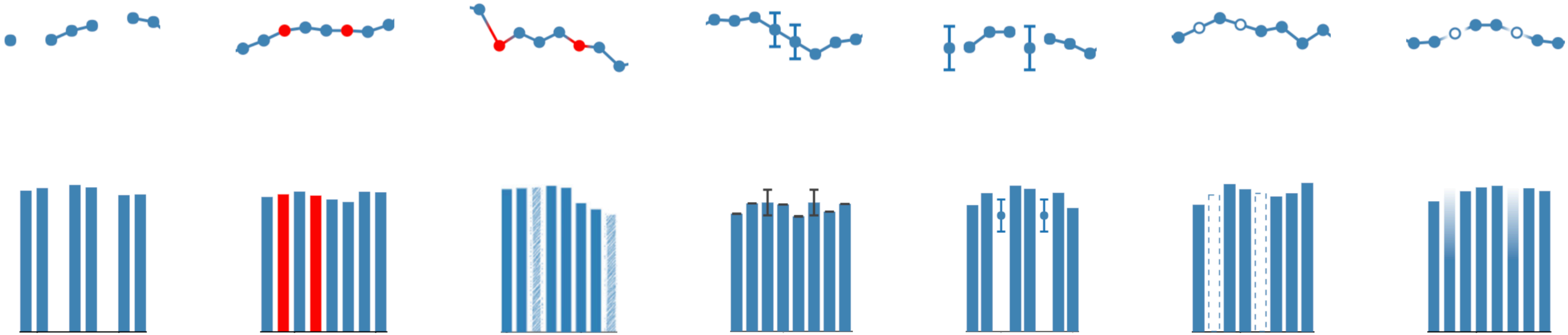
Hullman, Jessica. *Confronting Unknowns: How to Interpret Uncertainty in Common Forms of Visualization.* Scientific American, September 2019.

encoding uncertainty | *typical communication solutions may combine approaches*



Hullman, Jessica. *Confronting Unknowns: How to Interpret Uncertainty in Common Forms of Visualization*. Scientific American, September 2019.

 | *example ways we can show missing data, whether omitted or imputed*



Song, Hayeong, and Danielle Albers Szafir. "Where's My Data? Evaluating Visualizations with Missing Data." IEEE Transactions on Visualization and Computer Graphics 25, no. 1 (September 2018): 914–24.

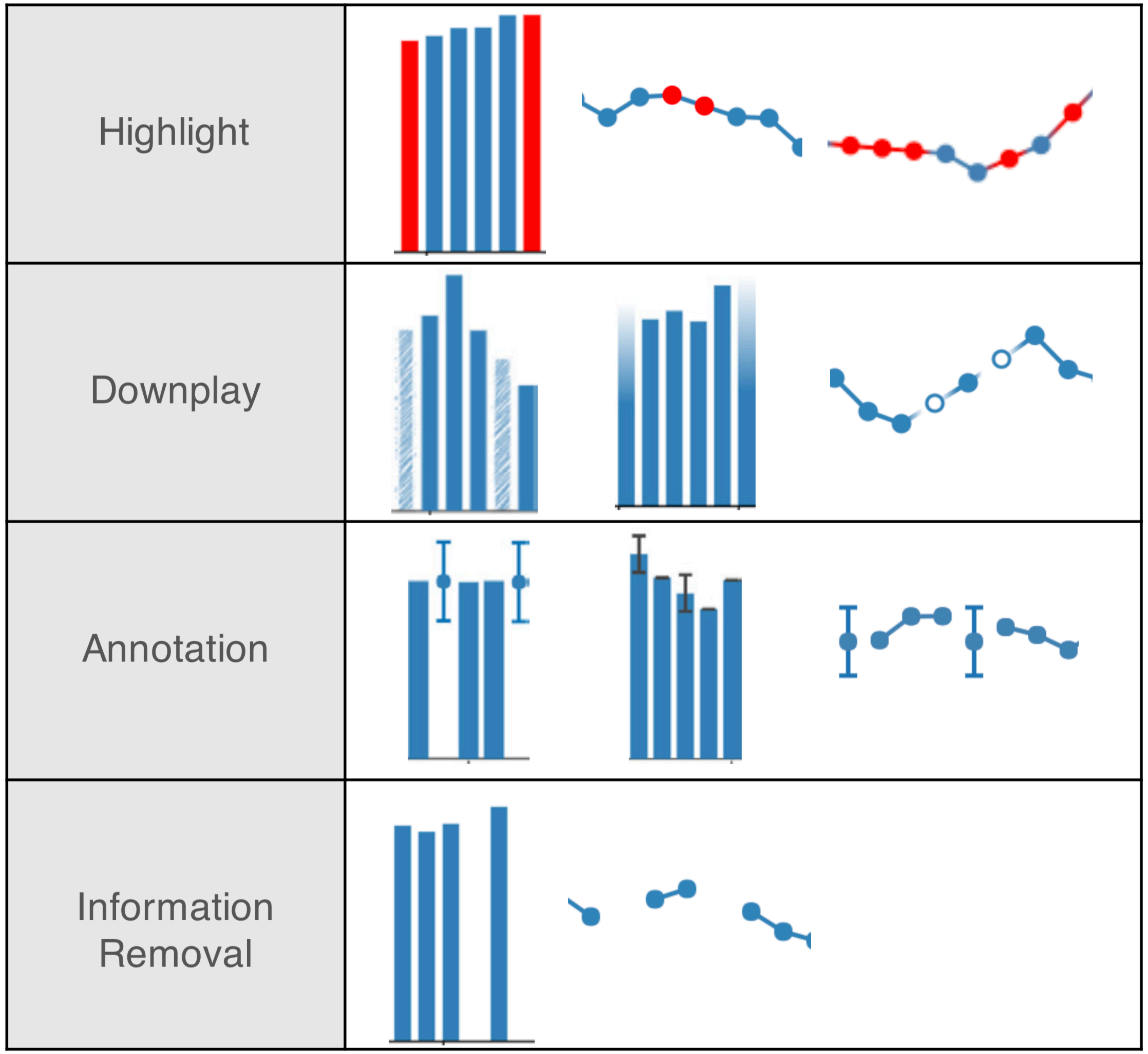
 | ***perception and confidence of data depend on form of communicating about missing values***

Perceived data quality and **confidence** generally degrade as the amount of missing data increases.

Data visualized by **highlighting** missing values tends to be seen as ***higher quality than*** **downplay or information** removal.

Information removal can significantly degrade perceptions of data quality, and confidence. These methods even lead to incorrect responses if missing values break the visual continuity of a visualization.

Modeling missing values (imputation) leads to higher perceptions of quality and confidence ***in analysis***.



Song, Hayeong, and Danielle Albers Szafir. "Where's My Data? Evaluating Visualizations with Missing Data." IEEE Transactions on Visualization and Computer Graphics 25, no. 1 (September 2018): 914–24.

bringing teachings together — *draft* proposal as example

data in narrative, **proposal as a multi-level narrative — title, headings, body, captions**

Spencer, Scott. (Draft) Proposal to Scott Powers. "*Proposal for Exploring Game Decisions Informed by Expectations of Joint Probability Distributions*." February 14, 2019.

"Orderliness **adds credibility** to the information and **induces confidence**. Information presented with clear and logically set out titles, subtitles, texts, illustrations and captions will not only be read more quickly and easily but the information will also be **better understood**."

— Müller-Brockmann, *Grid systems in graphic design*

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 face10 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 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?

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:

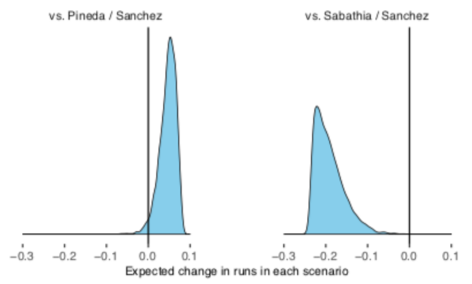


Figure 1. Of the two scenarios, Cain should only attempt to steal against the Sanchez–Pineda duo.

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

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

data in narrative, messages first, details follow

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 face10 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:

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.

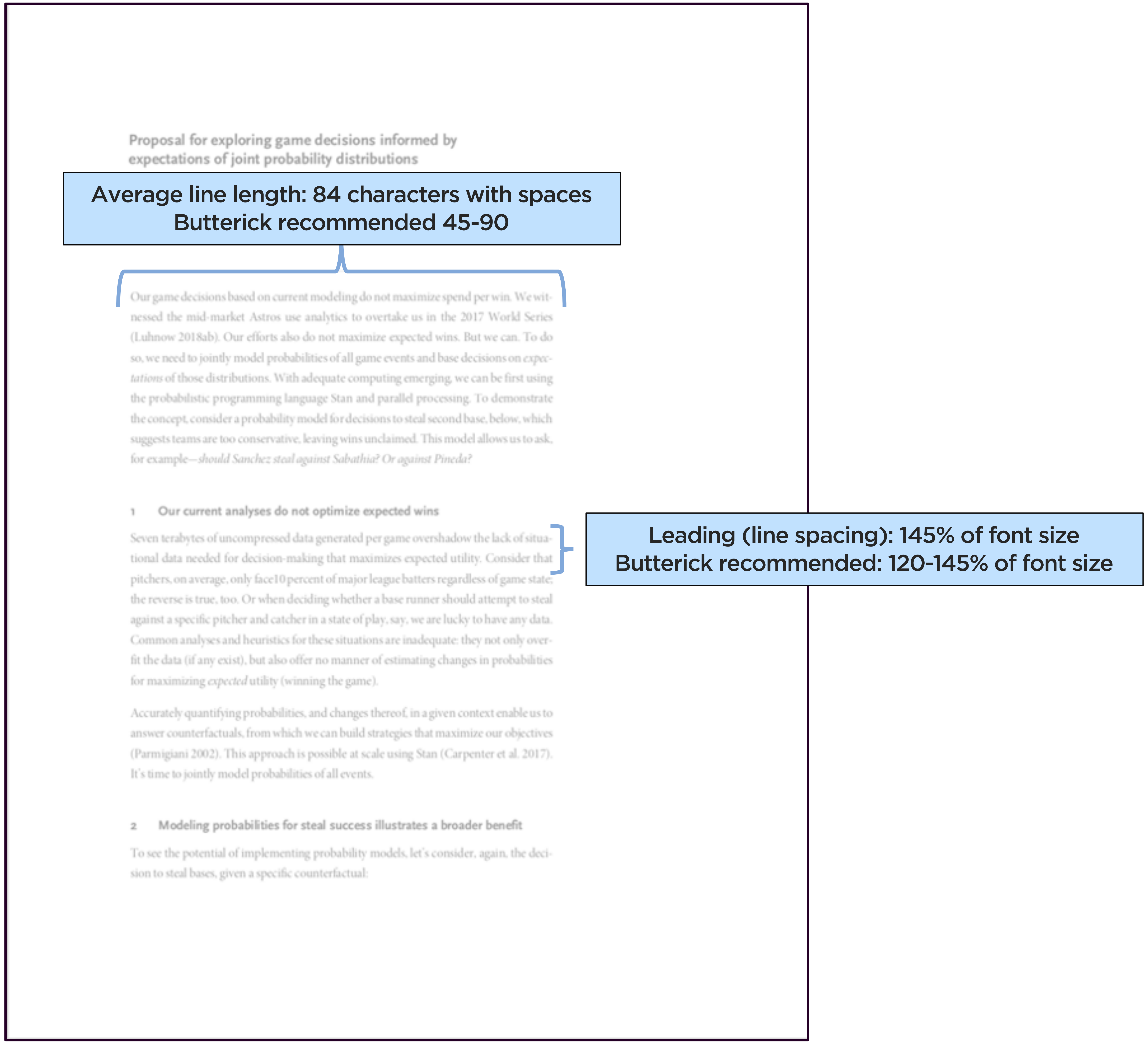
Parmigiani, G. 2002. "Decision Theory: Bayesian." In *International Encyclopedia of the Social Behavioral Sciences*, 3327–34.

Swartz, Matt. 2017. "The Recent History of Free-Agent Pricing." https://www.fangraphs.com/blogs/the-recent-history-of-free-agent-pricing/.

Get our audience(s) to **pay attention** to,
understand,
(be able to) **act upon**

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

— Doumont, *Trees, Maps, Theorems*



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

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 face10 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:

"Most readers are looking for reasons to stop reading. . . . Readers have other demands on their time. . . . The goal of most professional writing is persuasion, and attention is a prerequisite for persuasion. Good typography can help your reader devote less attention to the mechanics of reading and more attention to your message."

— Butterick, Matthew, *Practical Typography*

data in narrative, data graphics as paragraphs about data — linking narrative and data

“Words, graphics, and tables are different mechanisms with but a *single purpose*—the presentation of information. Why should the flow of information be broken up into different places on the page…?”

— Edward Tufte, *The Visual Display of Quantitative Information*

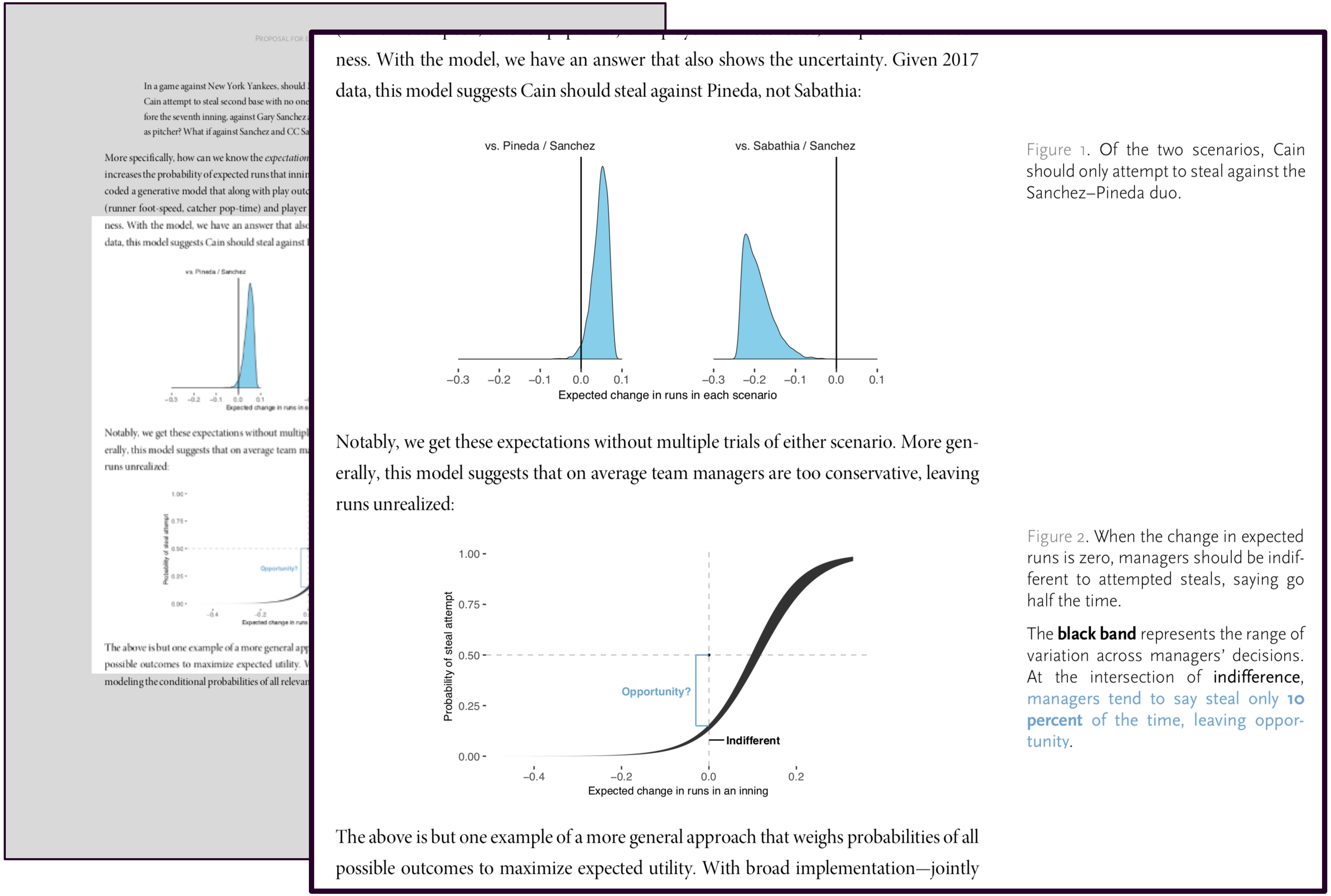

ness. With the model, we have an answer that also shows the uncertainty. Given 2017 data, this model suggests Cain should steal against Pineda, not Sabathia:

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