Storytelling With Data

variation and uncertainty

Scott Spencer | Columbia University

Conceptual project timeline

What are variation and uncertainty? Where might each arise?

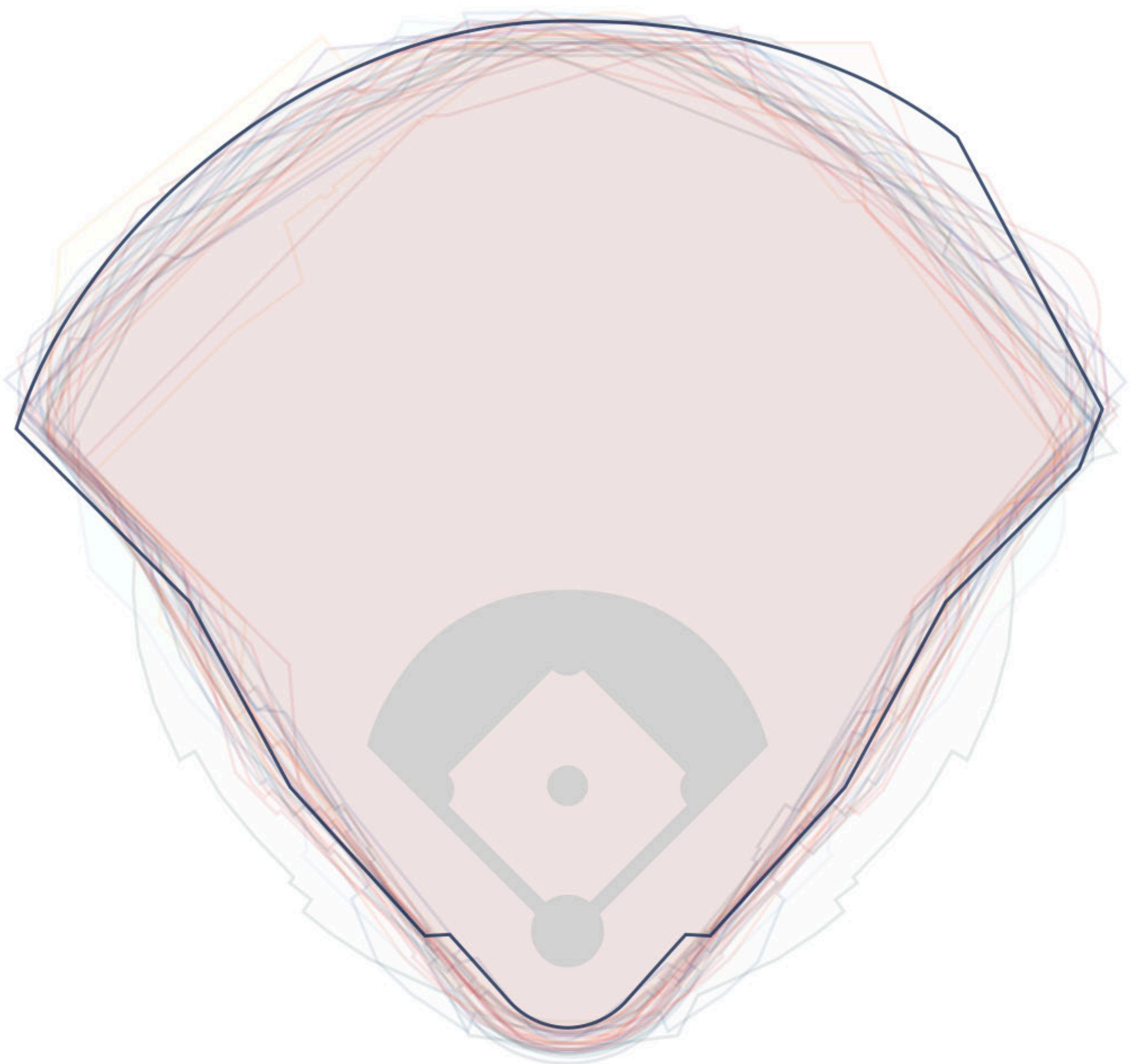
variation in context — *the data generating process*

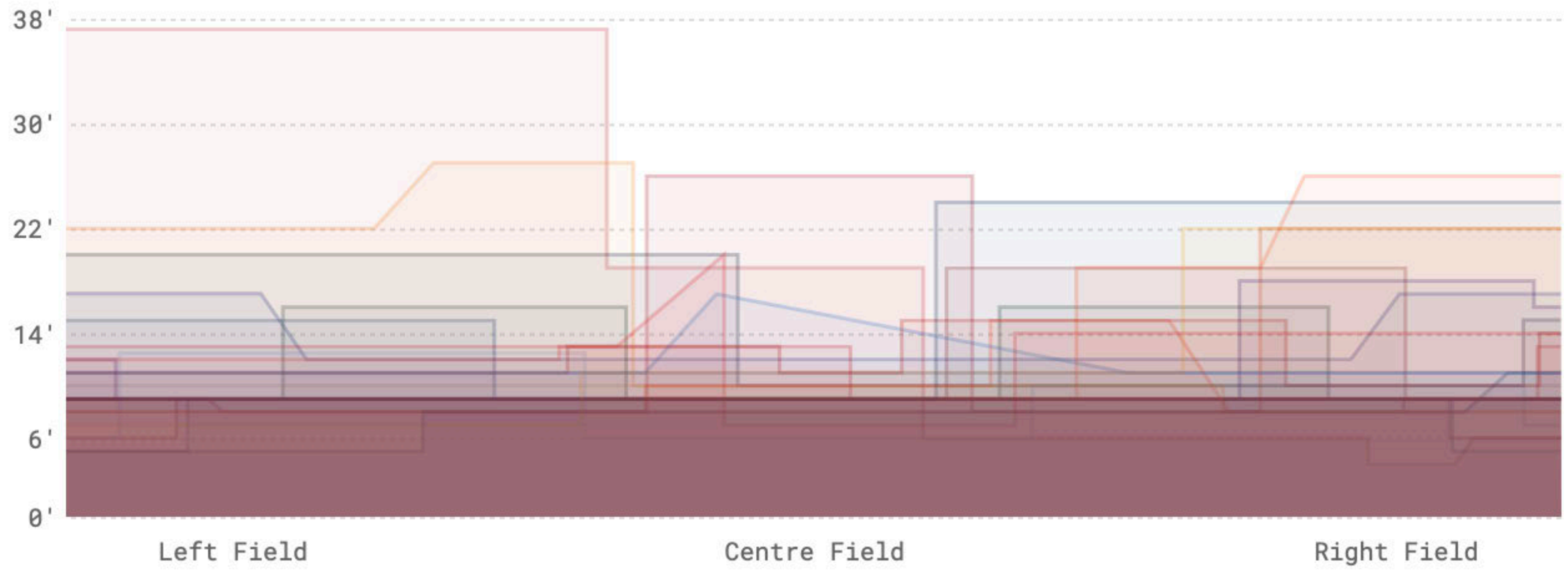
The focus on collecting "big data" for analyses can miss ***differences in what data represent***.

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

Loukissas, Yanni

 | *example — data in baseball depends on stadium, location, weather, people, …*

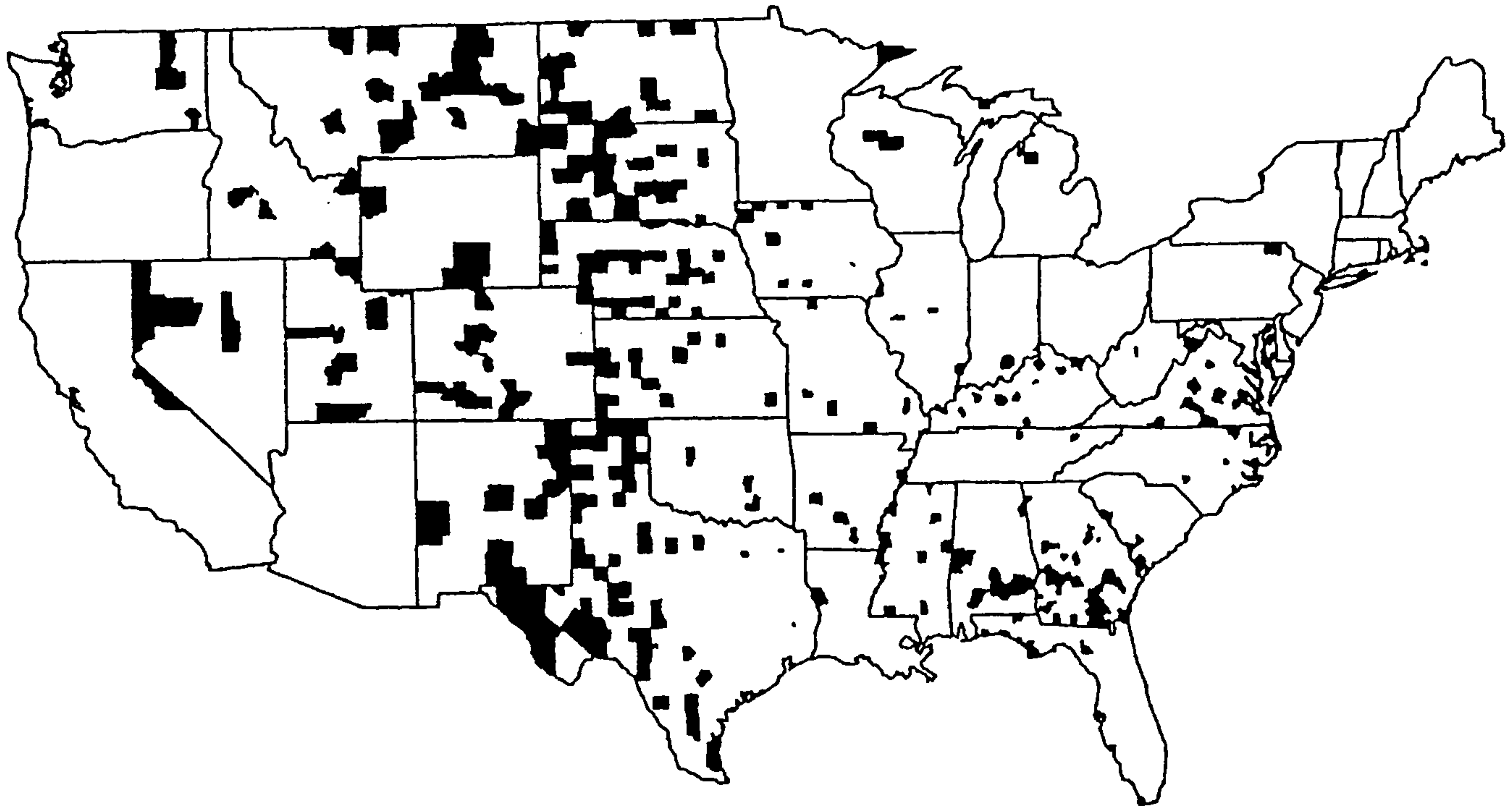


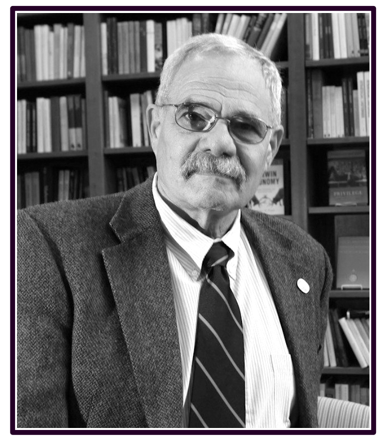


In this map of age-adjusted kidney cancer rates, the counties shaded are those counties that are in the ***lowest* decile of the cancer distribution**.

We note that these **healthy** counties tend to be very rural, midwestern, southern, and western counties.

It is both **easy and tempting to infer** that this outcome is directly due to the clean living of the rural life-style—*no air pollution, no water pollution, access to fresh food without additives, etc*.



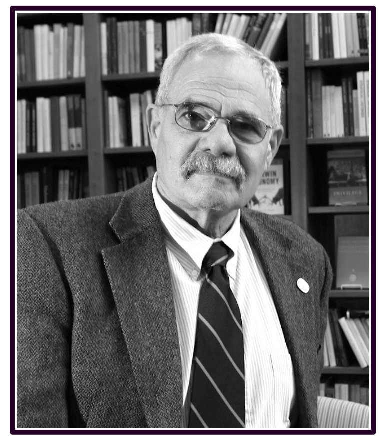


Wainer, Howard

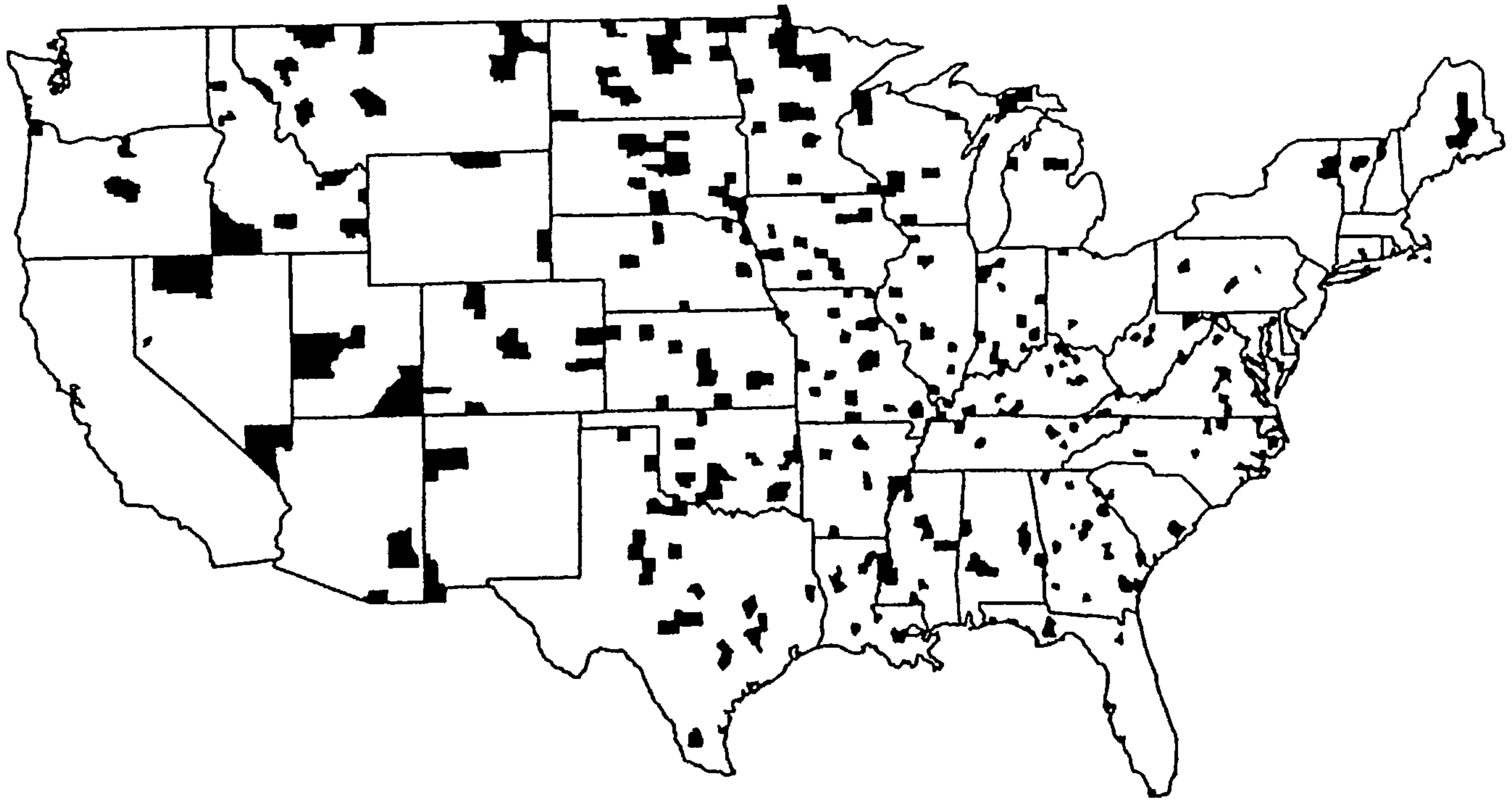
In another map of age-adjusted kidney cancer rates. While it looks very much like figure 1.1, it differs in one important detail—the counties shaded are those counties that are in the ***highest decile of the cancer distribution***.

We note that these **ailing** counties tend to be very rural, midwestern, southern, and western counties.

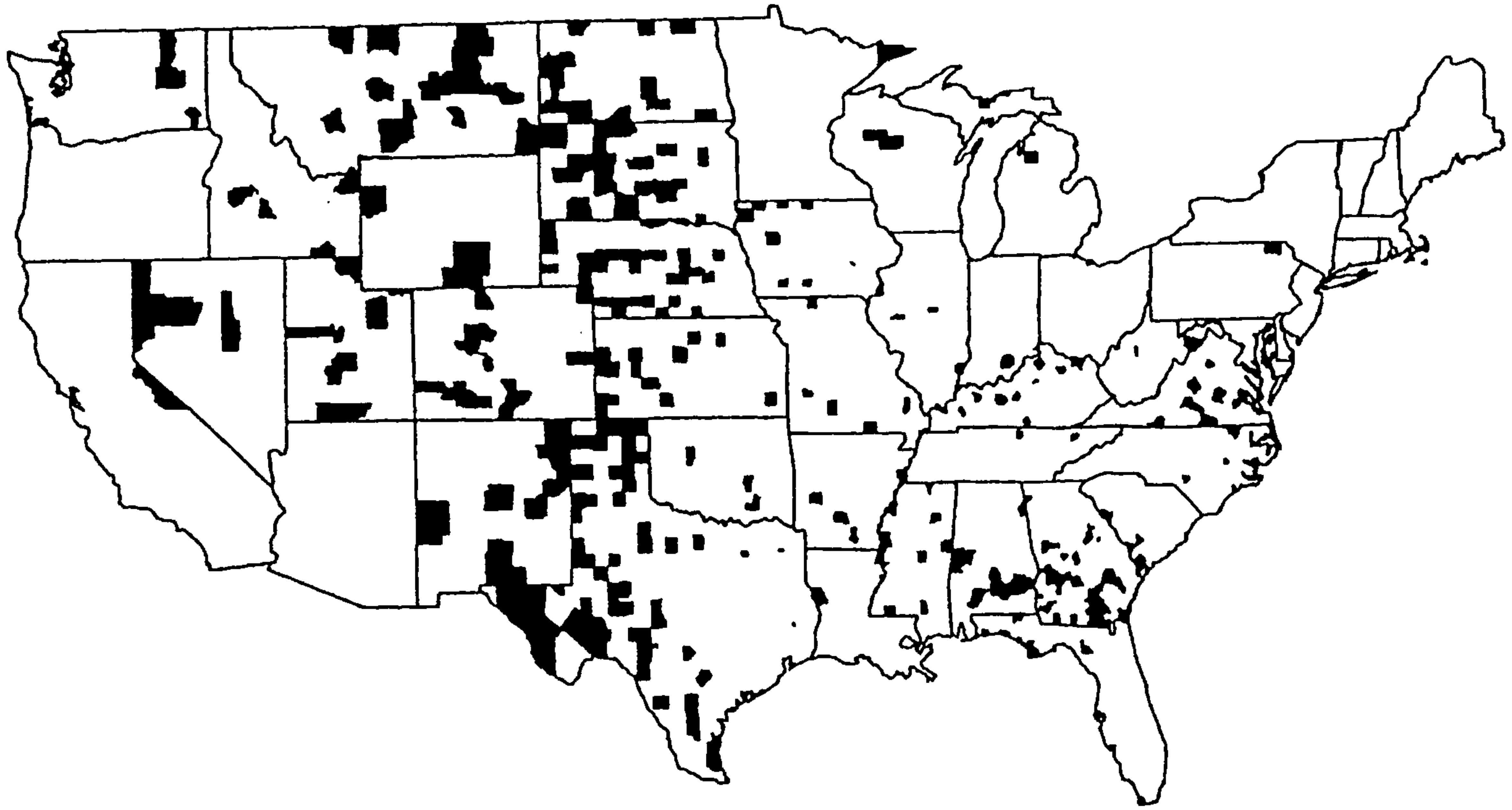
It is both **easy to infer** that this outcome might be directly due to the poverty of the rural lifestyle—*no access to good medical care, a high-fat diet, and too much alcohol, too much tobacco.*

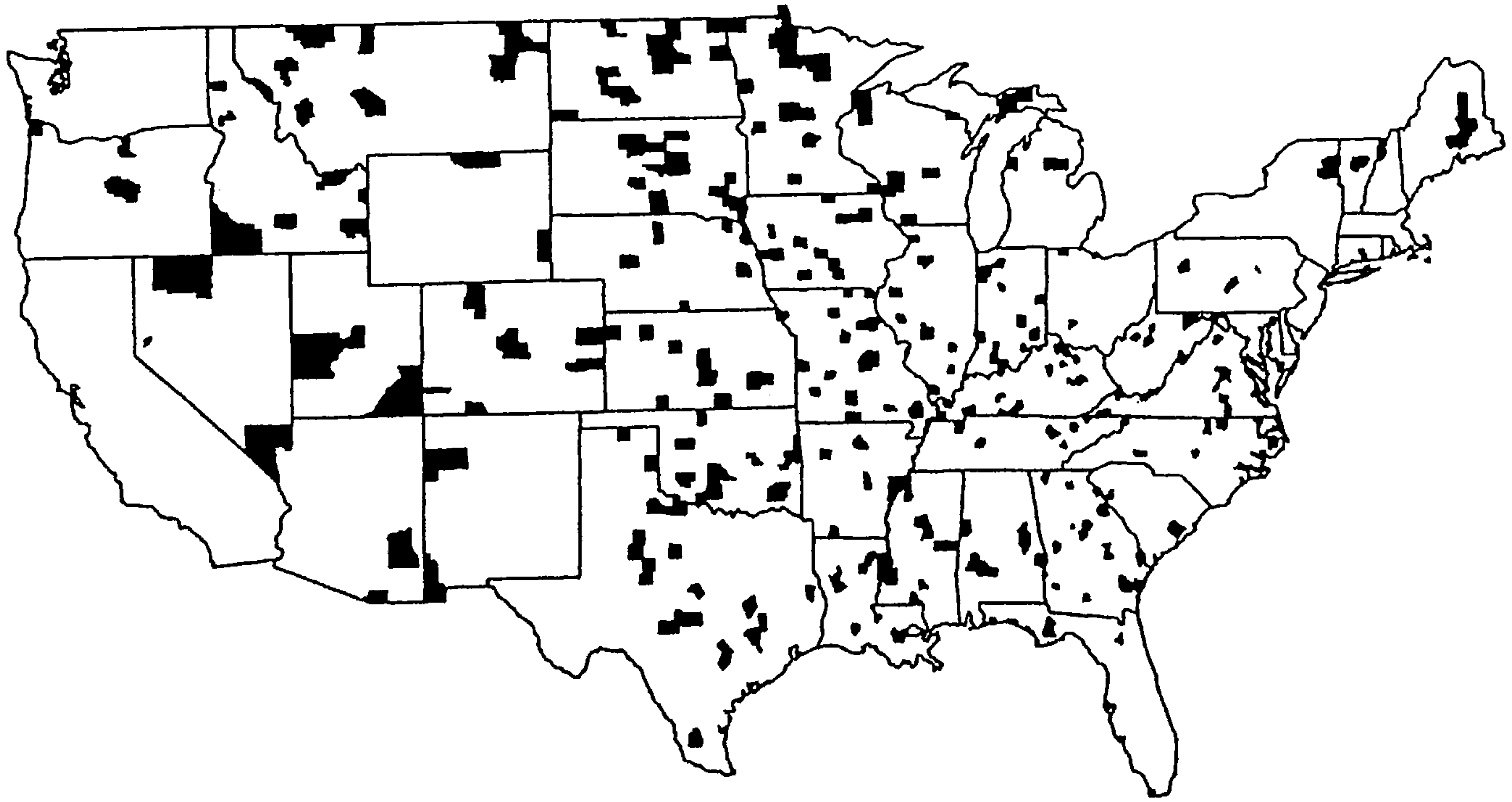


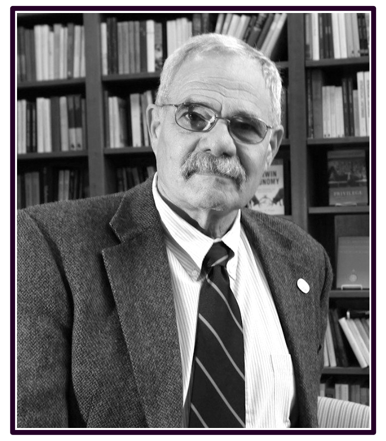
Wainer, Howard



 variation in means are inversely proportional to square root of sample size



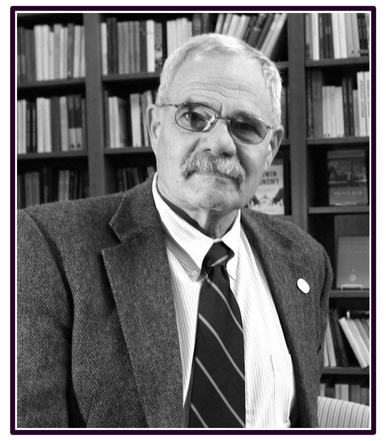




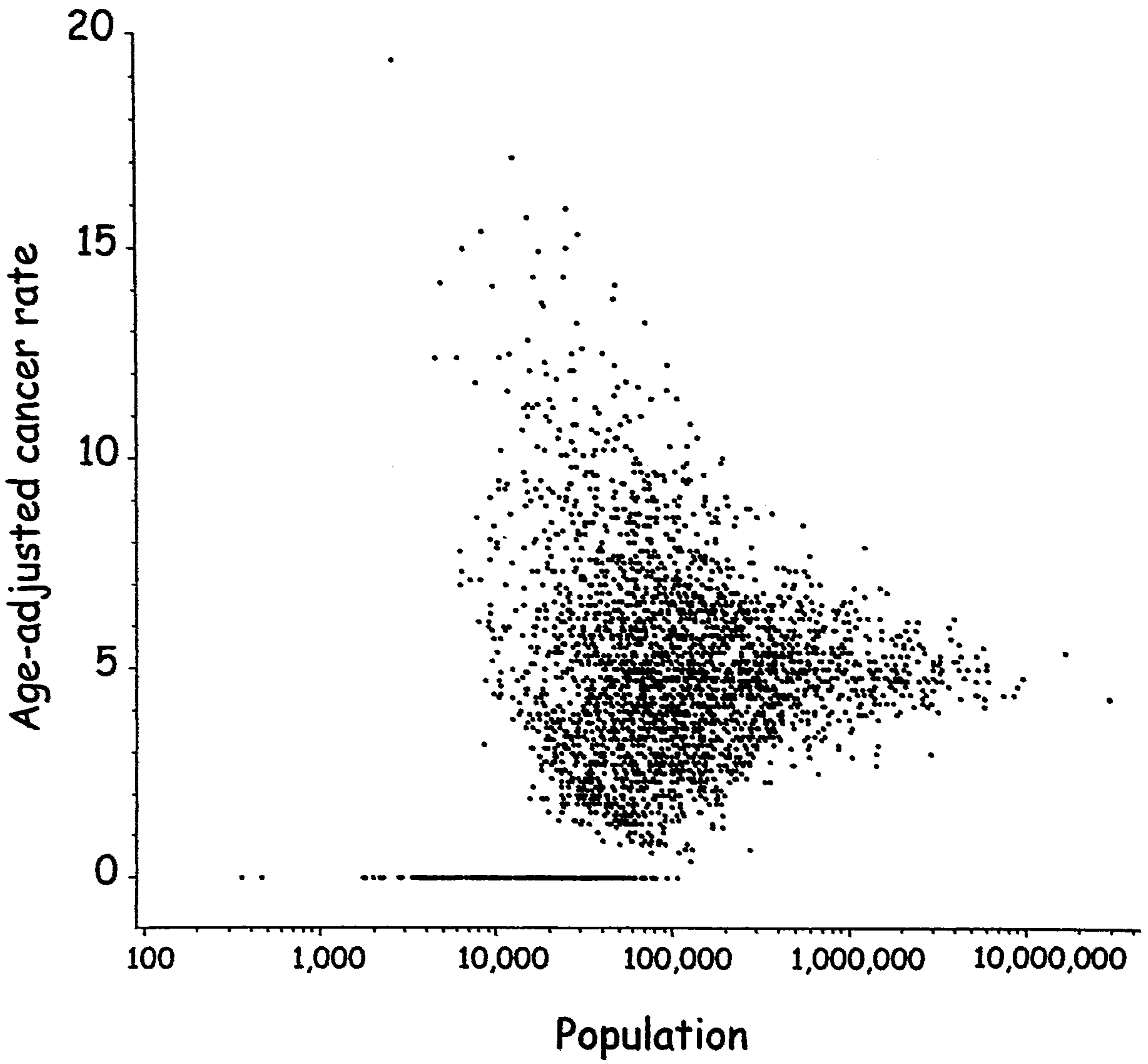
Wainer, Howard

The apparent paradox is explained by variation due to sample size — Moivre's equation in action. The variation in the mean is inversely proportional to the square root of the sample size, and so small counties have much larger variation than large counties.

Our credibility and decisions informed by communication are both improved when we accurately convey variation and uncertainty.



Wainer, Howard

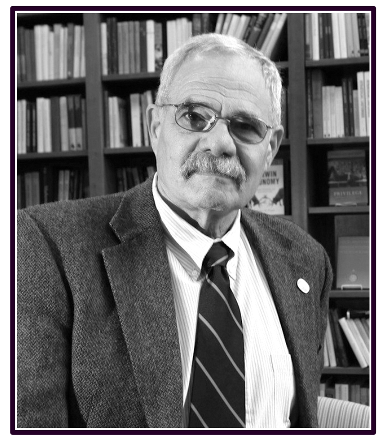


The most dangerous equation

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



Wainer, Howard

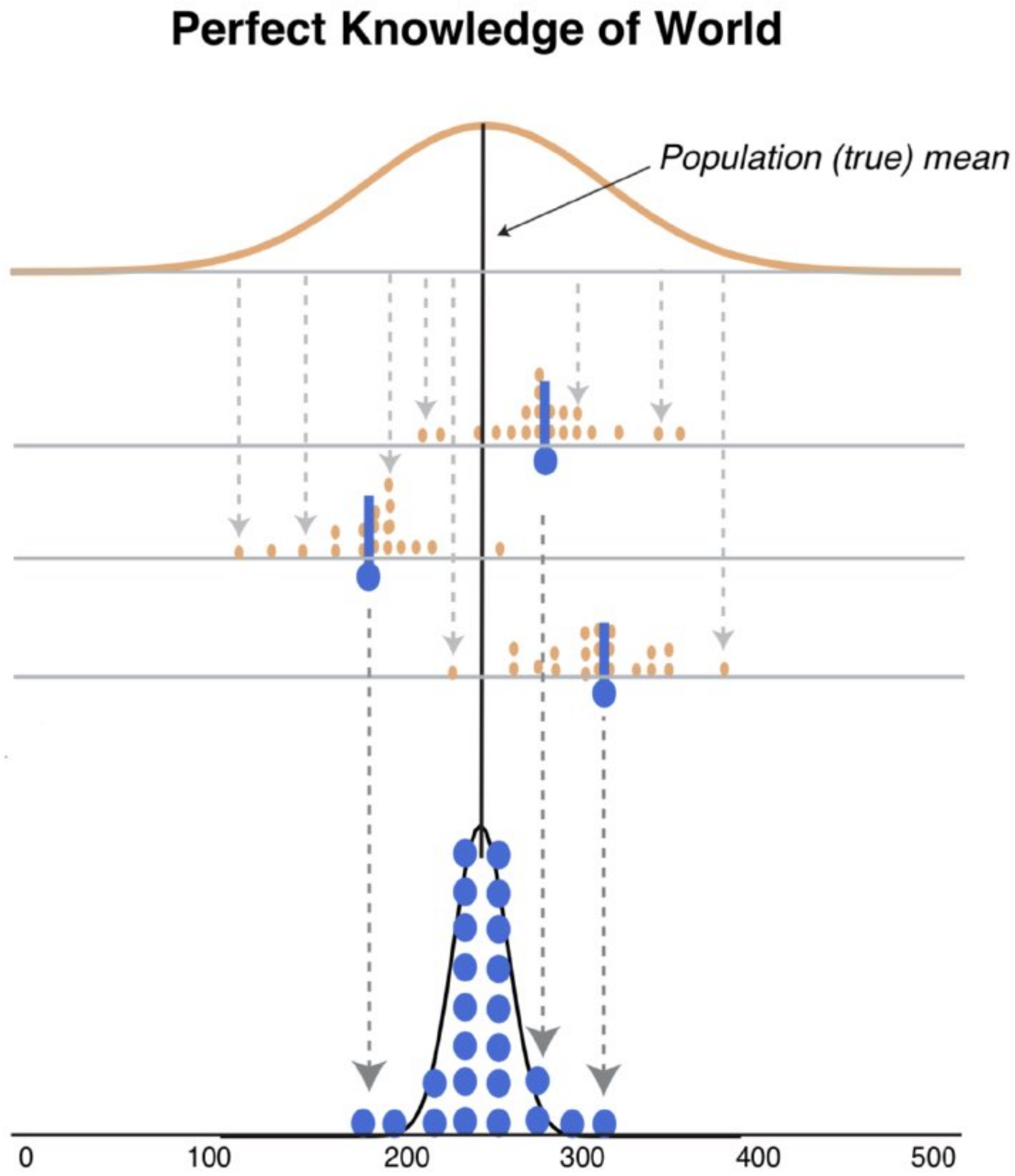
Why so dangerous?

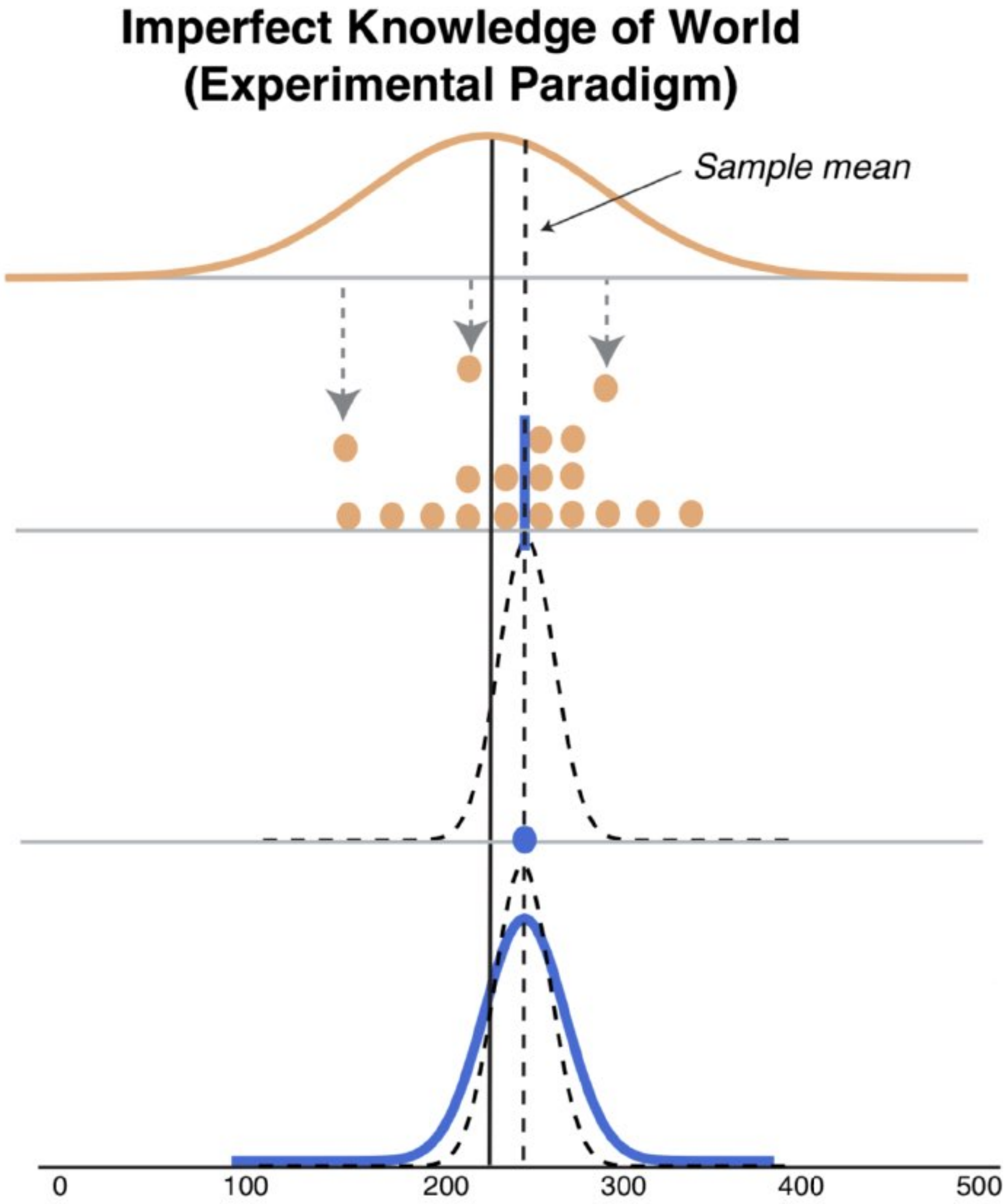
Extreme length of time during which ignorance of it has caused confusion

Wide breadth of areas that have been misled

Seriousness of the consequences that ignorance has caused

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





Hullman, Jessica

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

communicating variation and uncertainty

What obstacles have you found in communicating uncertainty?

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

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

Fischhoff, Baruch

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.

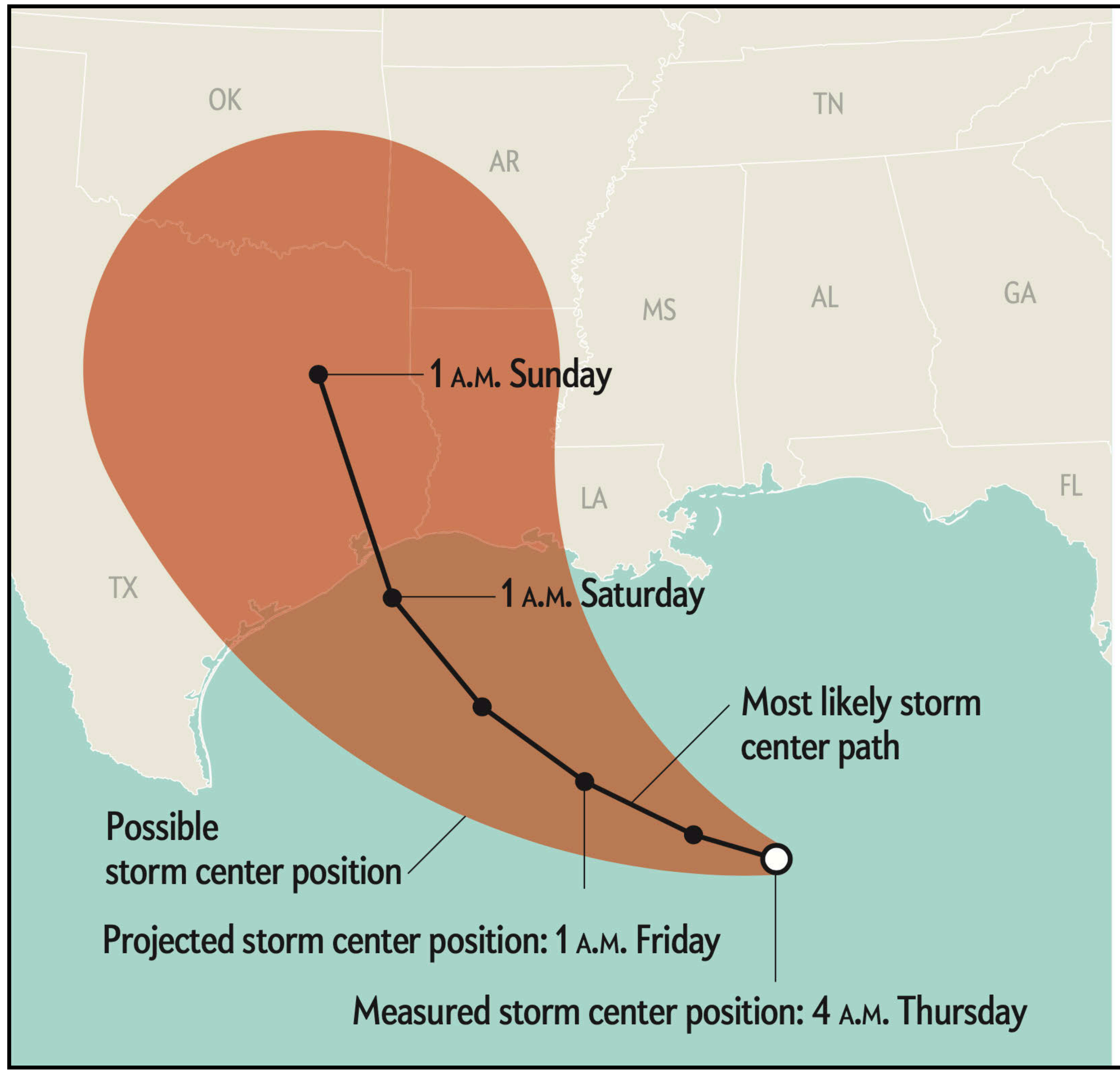
Concern | credible intervals may be **used unfairly** in performance evaluations.

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

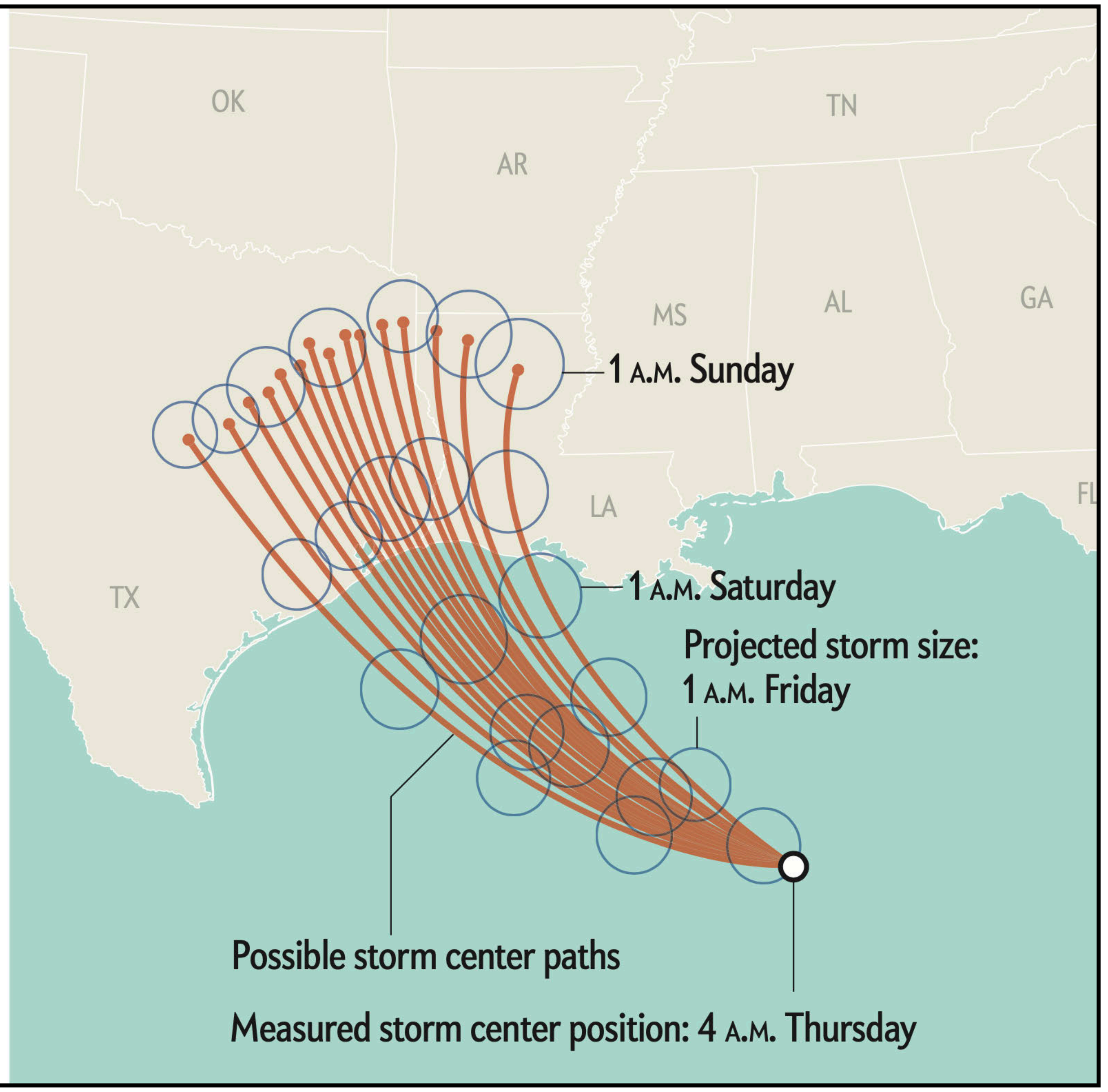
visually encoding uncertainty

 | ***consider alternative encodings and how perception may differ***

Uncertainty in storm path misperceived as growth in size

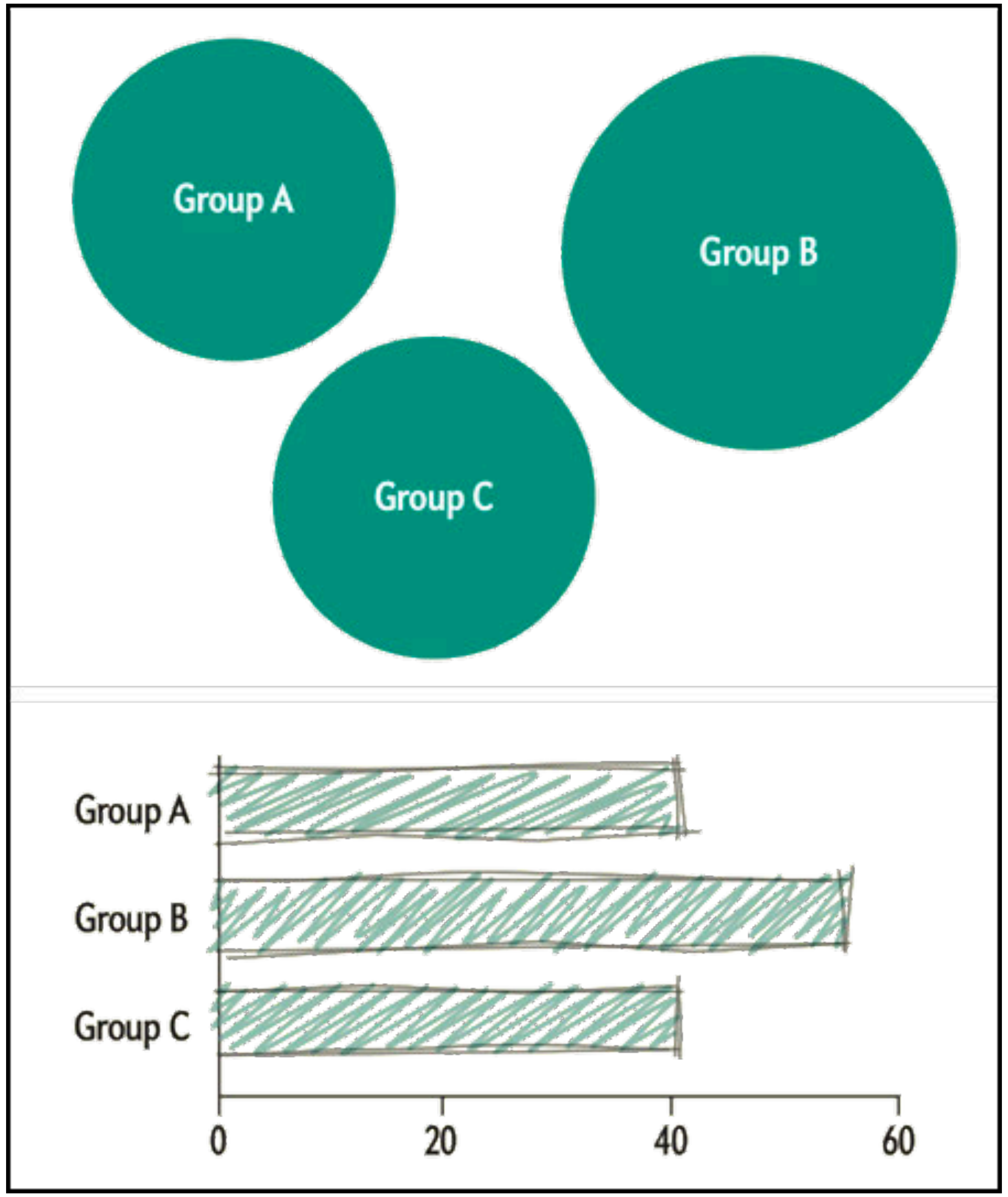


Alternative way to express uncertainty of storm path



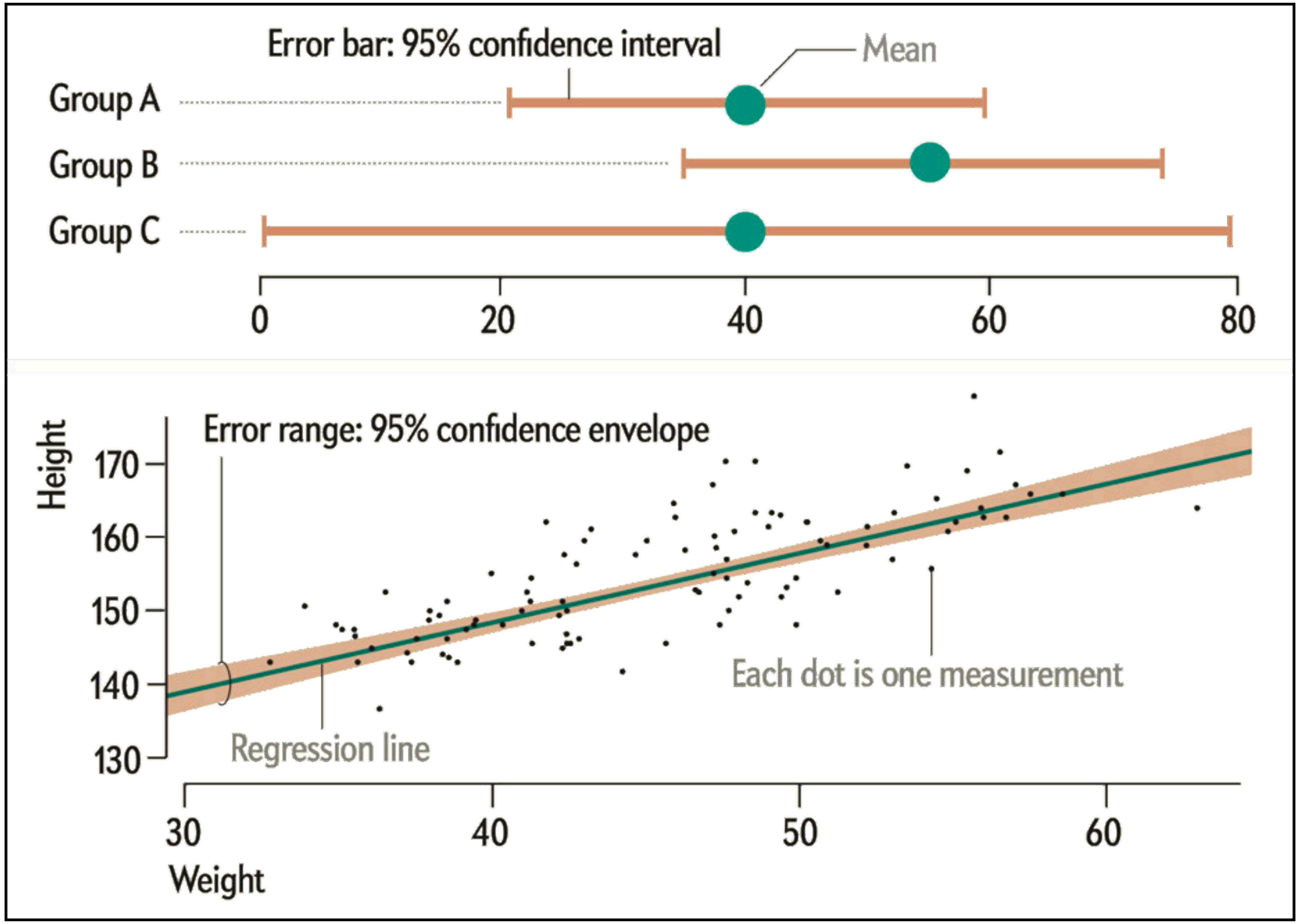
Hullman, Jessica

 | ***no quantification occurs most — provides least information for decisions***



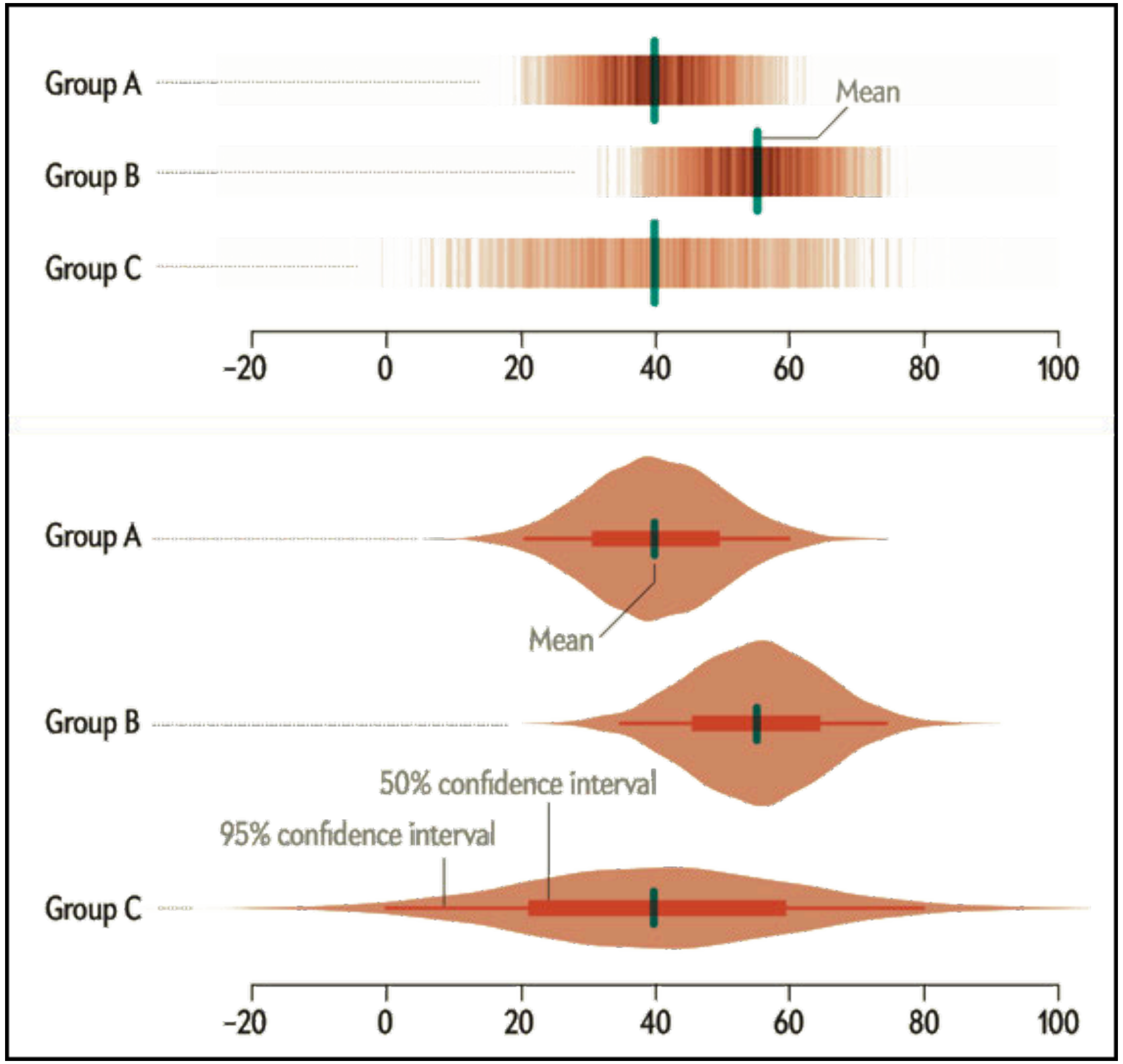
Hullman, Jessica

 intervals are perhaps the most common encodings for uncertainty



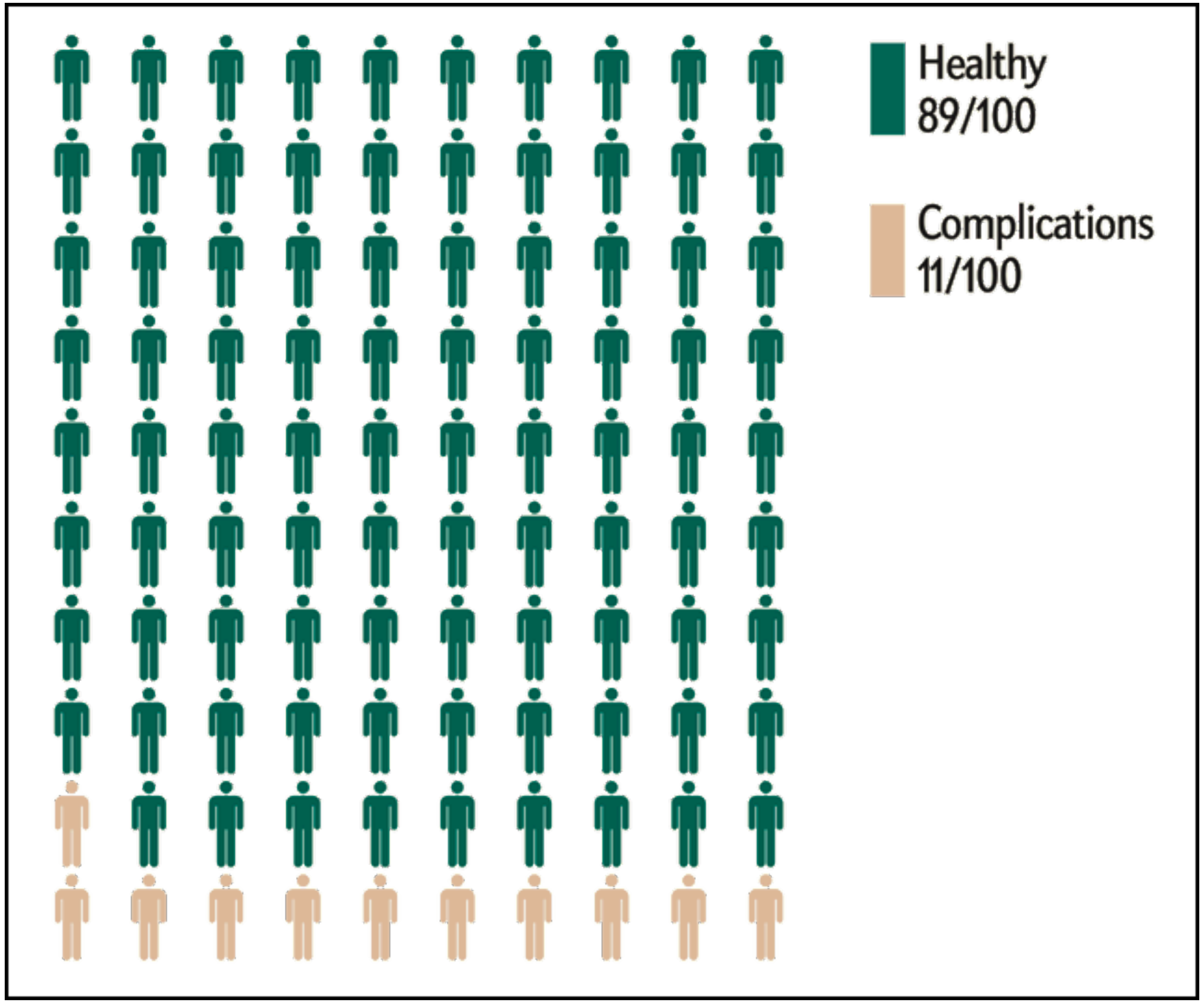
Hullman, Jessica

 probability densities tend to provide the most information about data



Hullman, Jessica

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

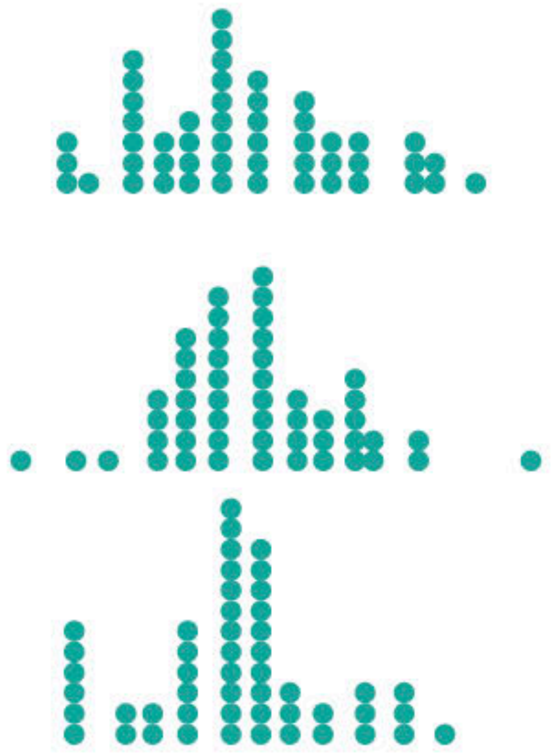


Hullman, Jessica

 | *quantile dot plots create countable distributions — improves decoding accuracy*

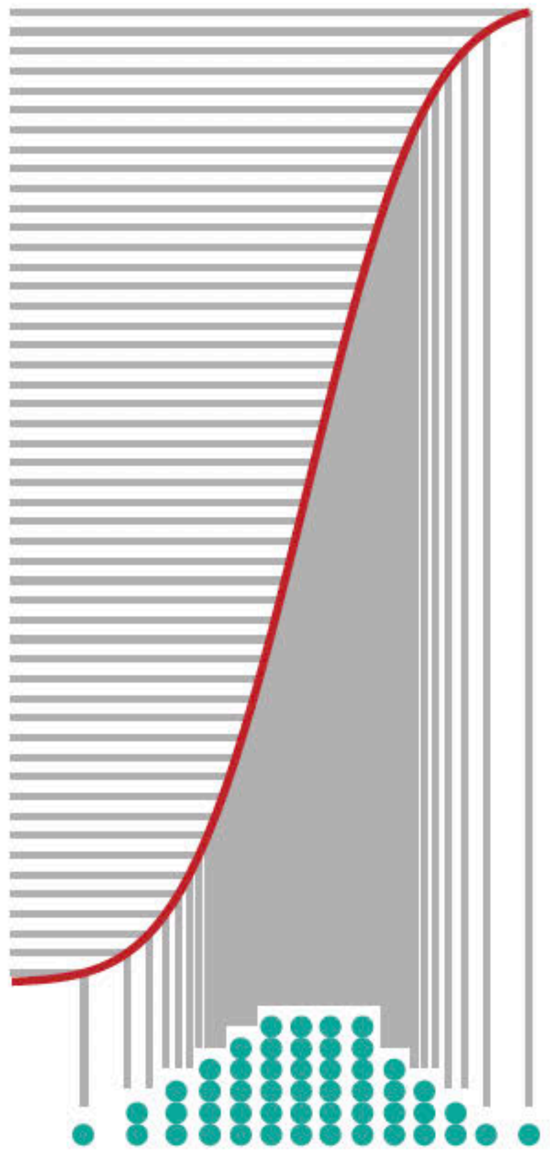
Probability density of Normal distribution

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.



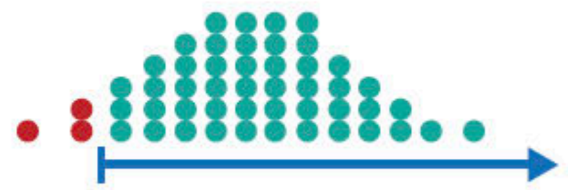
Probability density of Normal distribution

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



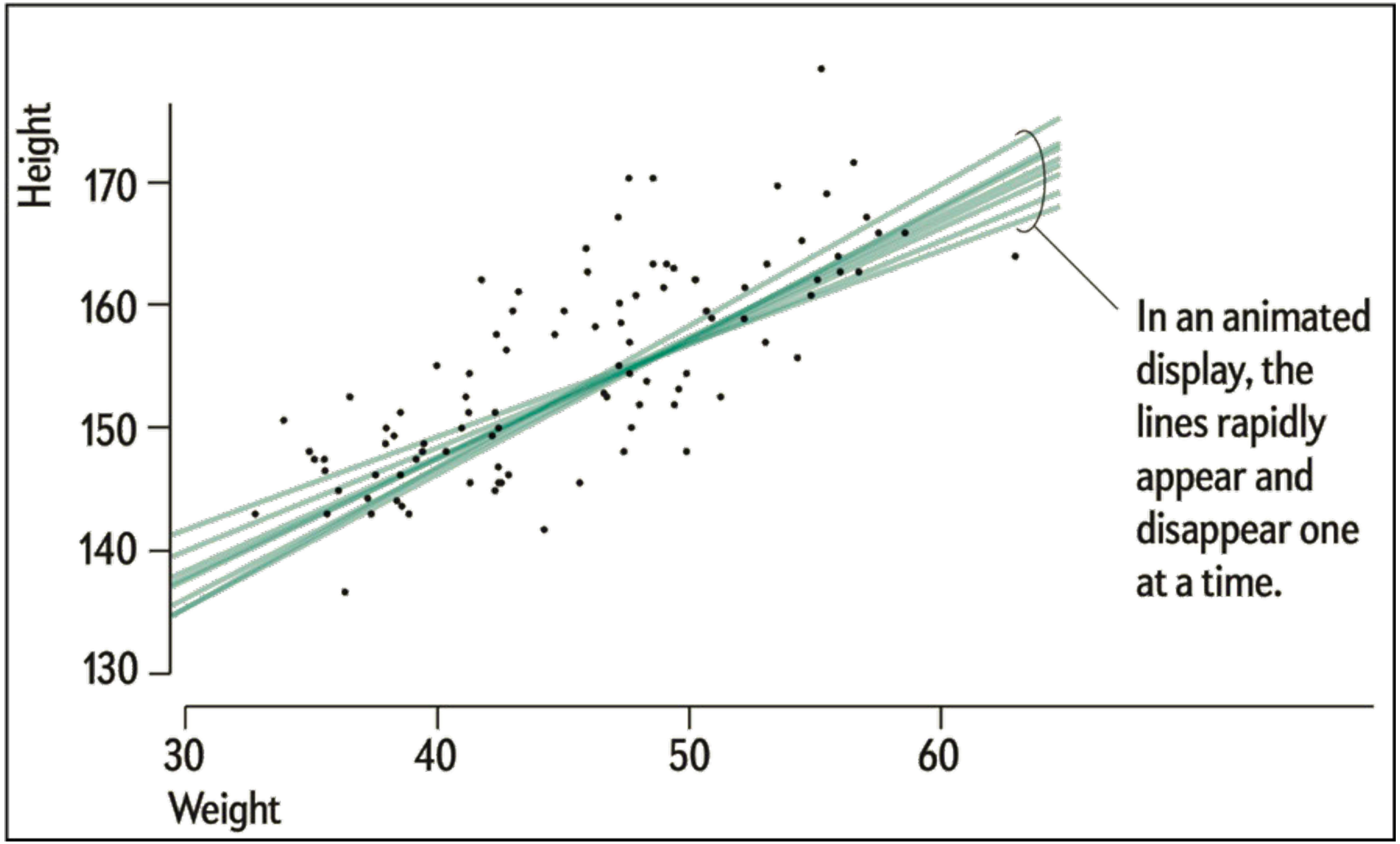
We plot the quantile "draws" using a Wilkinsonian dotplot, yielding what we call a **quantile dotplot**: a consistent discrete representation of a probability distribution.

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.



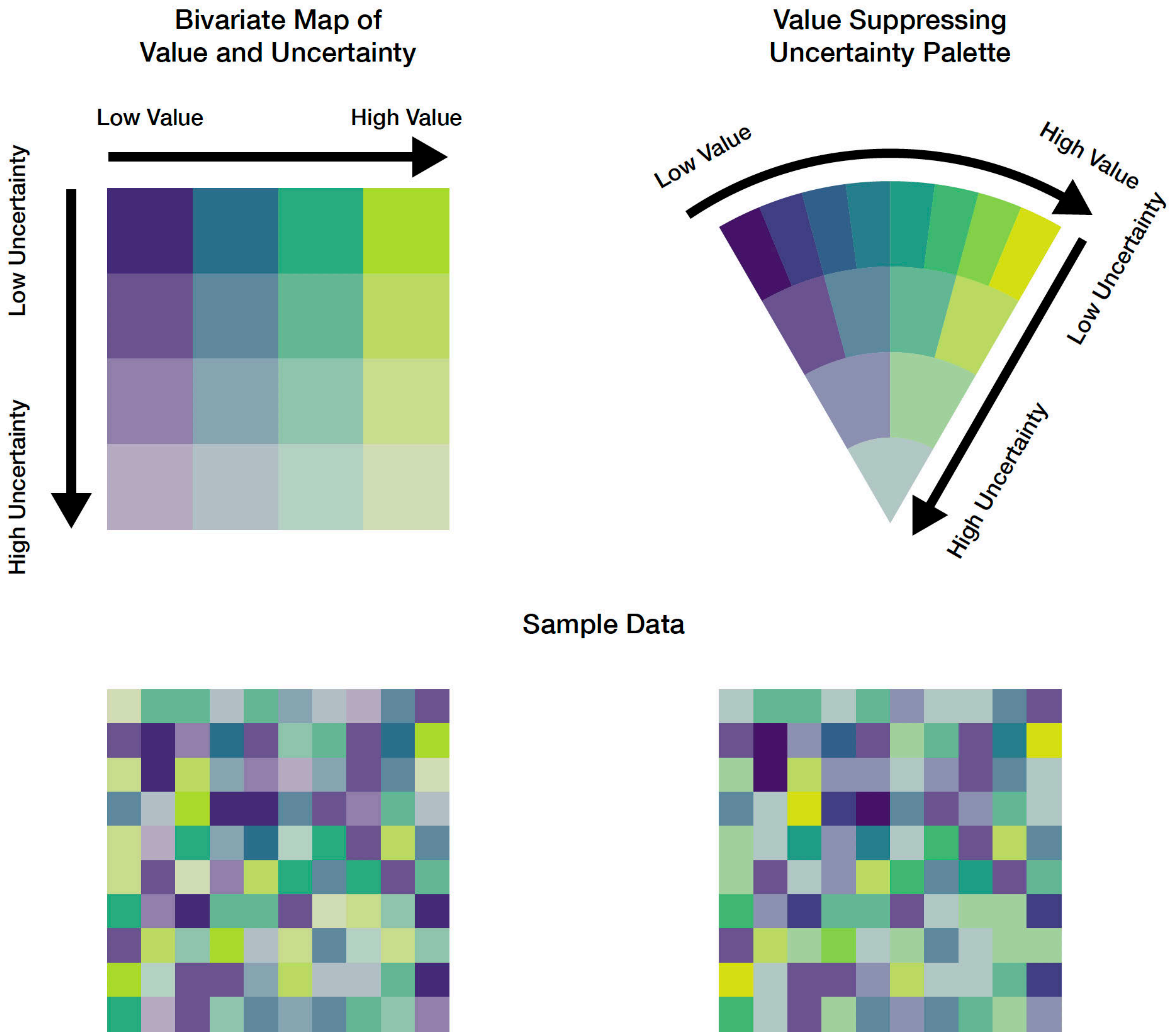
Kay, Matthew & co-authors

 hypothetical outcome plots (showing samples serially) may help people feel uncertainty



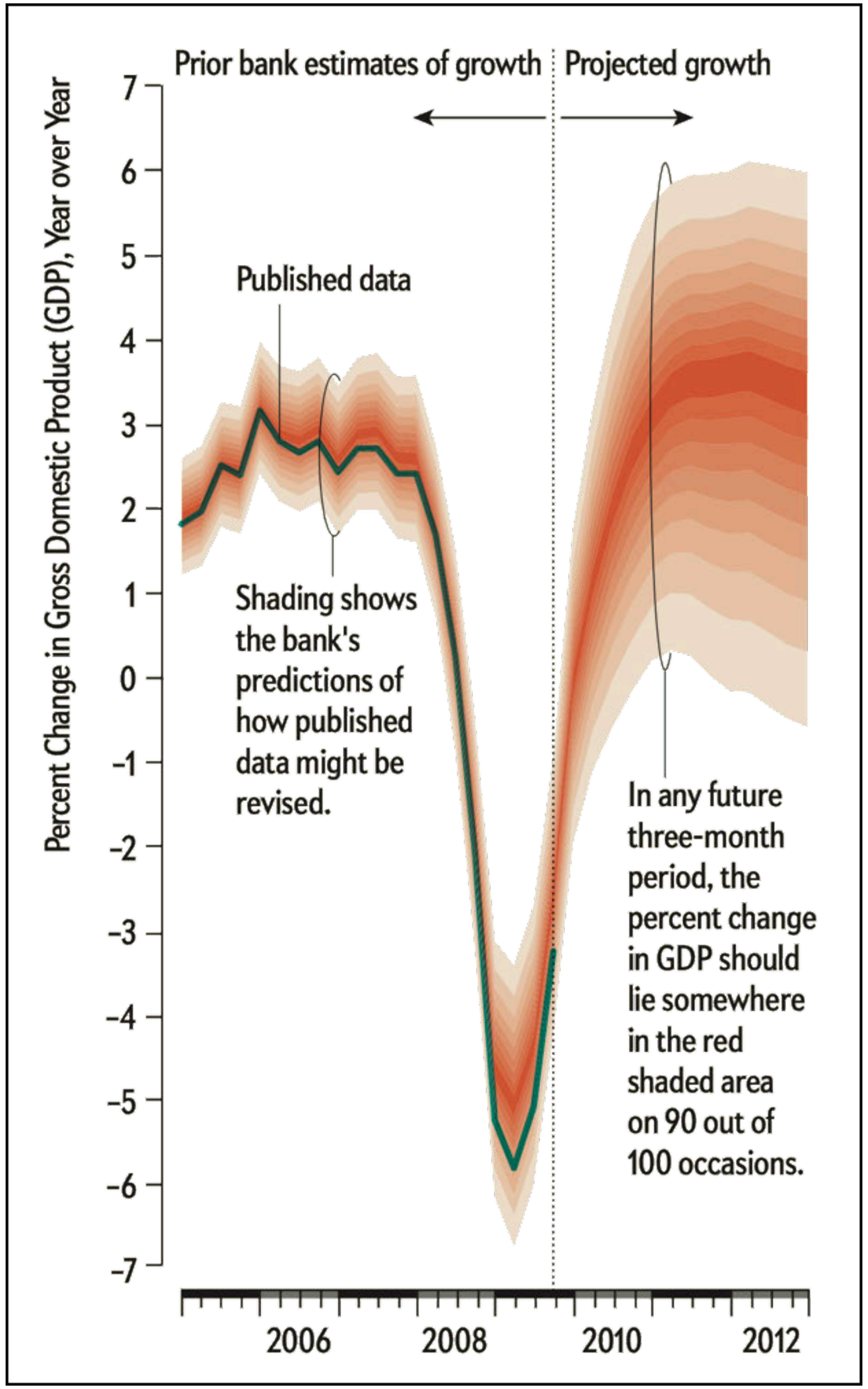
Hullman, Jessica

 | ***mapping uncertainty to color channel (hue, saturation, luminance)***



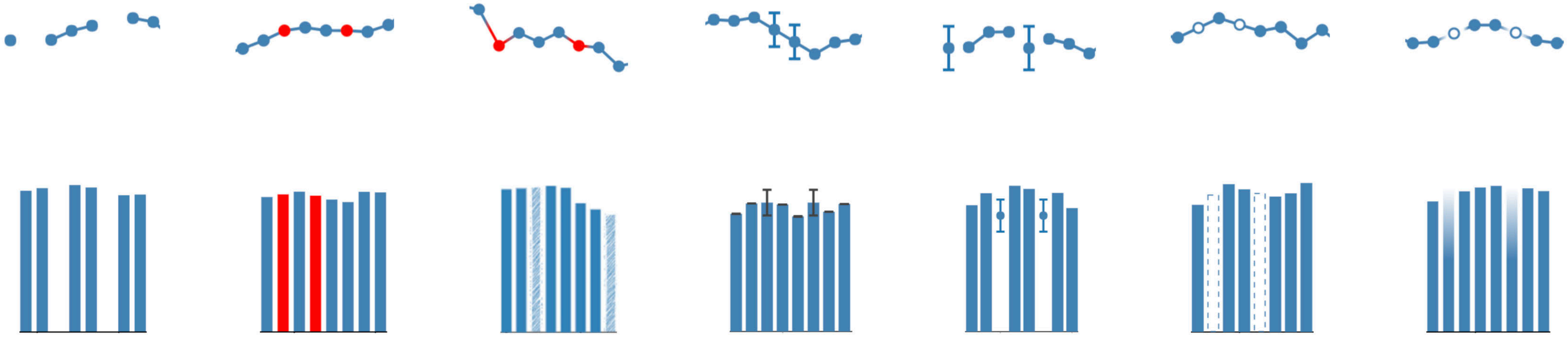
Correll, Michael & co-authors

 typical communication solutions may combine approaches



Hullman, Jessica

encoding uncertainty about missing data



Song & Szafir

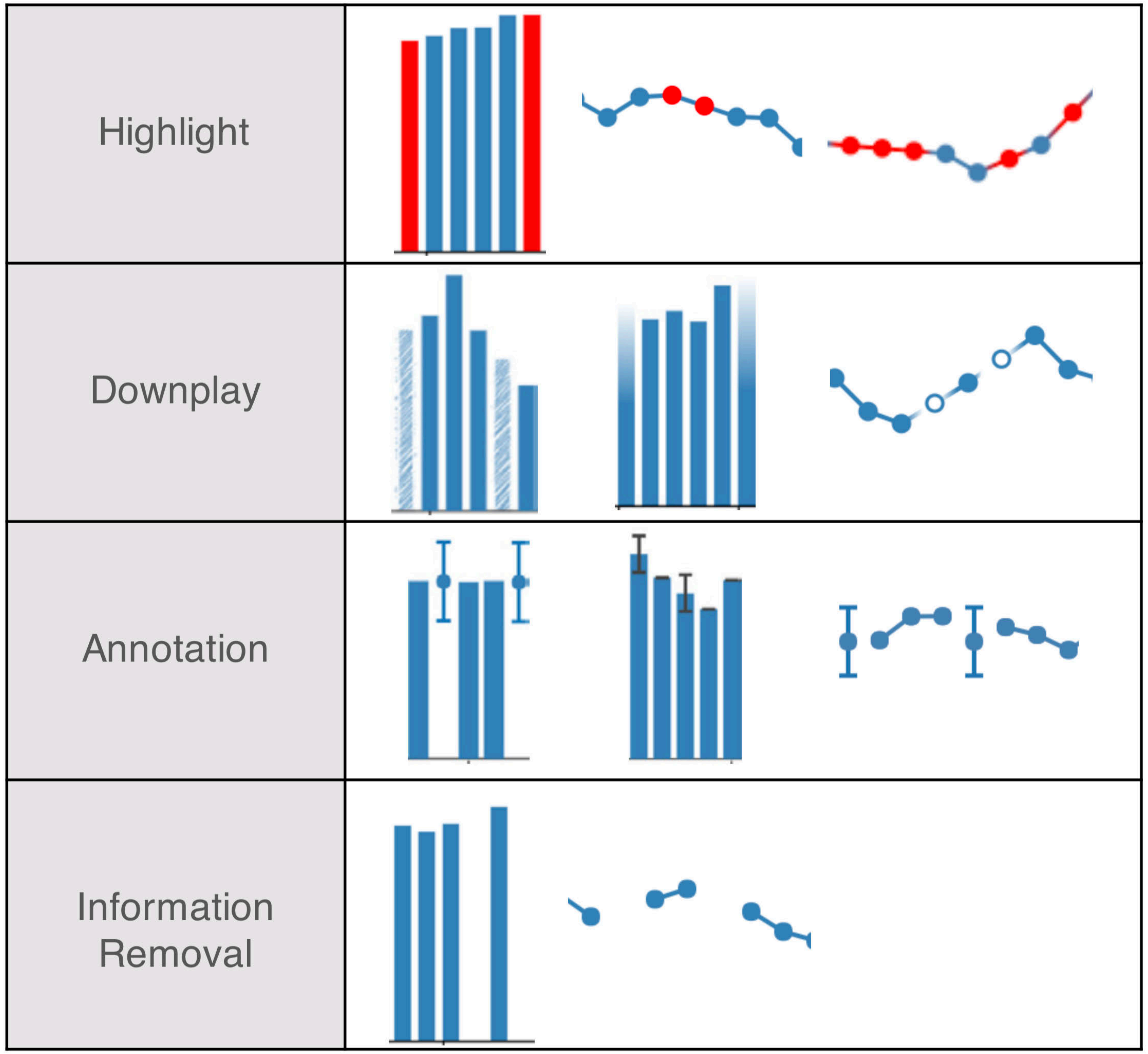
 | ***perception and confidence of data depend on form of communicating about missing values***

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.

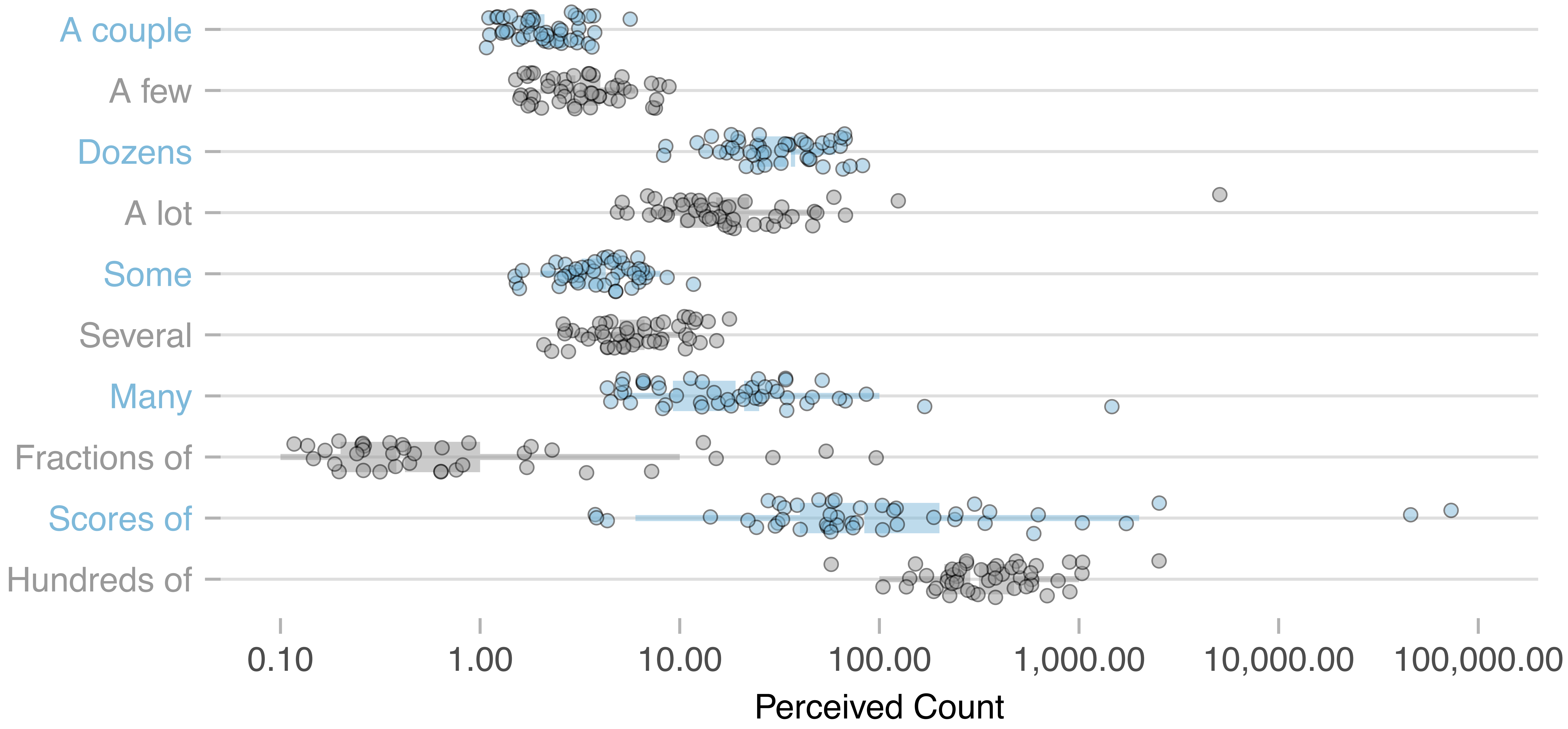
Modeling missing values (imputation) leads to higher perceptions of quality and confidence ***in analysis***.



Song & Szafir

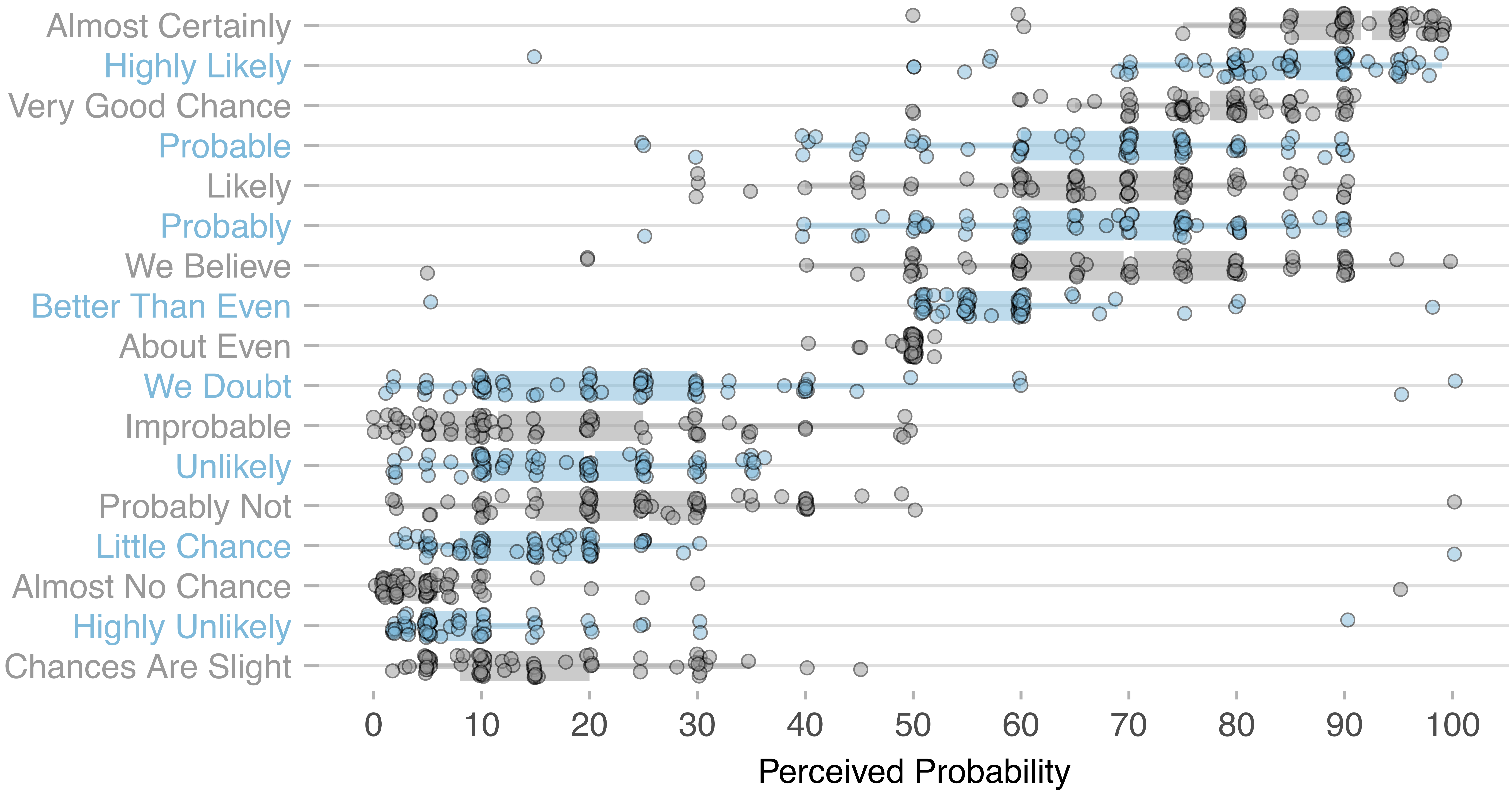
words expressing uncertainty matter too

uncertainty | *people vary in their interpretation of words communicating quantity*



Barclay and zonination

uncertainty | ***people vary in their interpretation of words communicating probability***



Barclay and zonination

Summer suggestion: Bayesian analysis and decision theory

References

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. "*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.

Fernandes, Michael, Logan Walls, Sean Munson, Jessica Hullman, and Matthew Kay. "*Uncertainty Displays Using Quantile Dotplots or CDFs Improve Transit Decision-Making*." In The 2018 CHI Conference, 1–12. New York, New York, USA: ACM Press, 2018.

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, Tara Kola, Jessica R Hullman, and Sean A Munson. "*When (ish) Is My Bus? User-Centered Visualizations of Uncertainty in Everyday, Mobile Predictive Systems*." In The 2016 CHI Conference, 5092–5103. New York, New York, USA: ACM Press, 2016.

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

McElreath, Richard. *Statistical Rethinking: A Bayesian Course with Examples in R and Stan*. 2nd ed. CRC Texts in Statistical Science. Boca Raton: Taylor and Francis, CRC Press, 2020.

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.

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

Wainer, Howard. *Picturing the Uncertain World. How to Understand, Communicate, and Control Uncertainty through Graphical Display*. Princeton University Press, 2009.

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