DATA SCIENCE

CONFRONTING UNKNOWNS

HOW TO INTERPRET UNCERTAINTY IN COMMON FORMS OF DATA VISUALIZATION

By Jessica Hullman

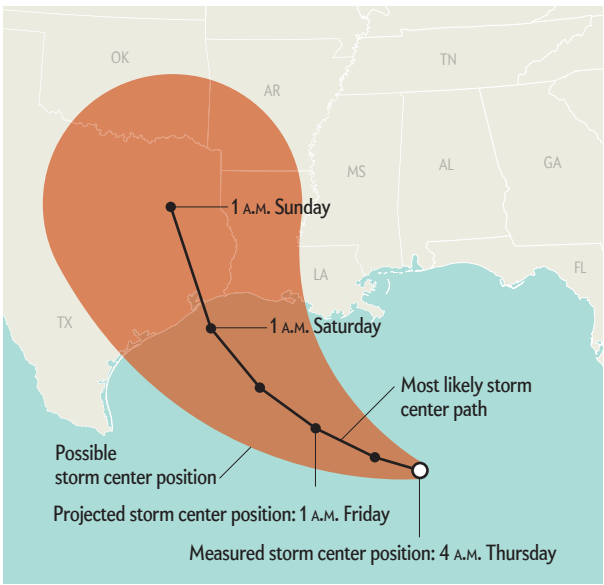
Jessica Hullman is a professor of computer science and journalism at Northwestern University. She and her research group develop and evaluate data-visualization and data-interaction techniques to enhance reasoning about uncertainty.

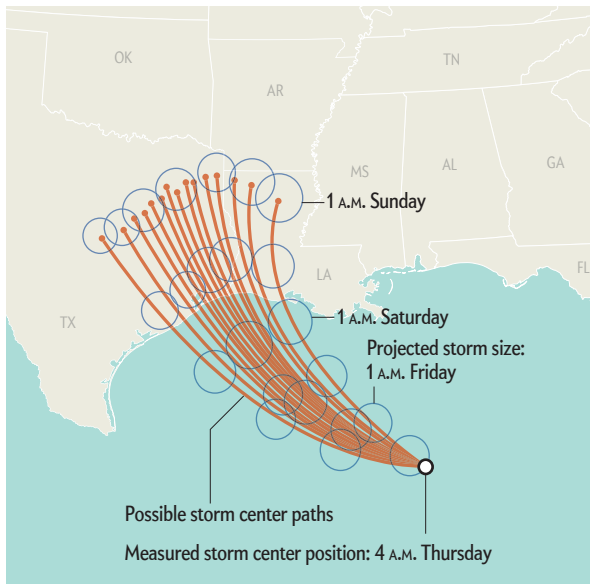
When tracking a hurricane, forecasters often show a map depicting a "cone of uncertainty." It starts as a point—the hurricane's current position—and widens into a swath of territory the storm might cross in the upcoming days. The most likely path is along the centerline of the cone, with the probability falling off toward the edges. The problem: many people misinterpret the cone as the size of the future storm.

Researchers have found that the misunderstanding can be prevented if forecasters instead show a number of possible paths. Yet this approach can also introduce misunderstanding: lots of people think the probability of damage is greater where each path intersects land and less likely between the lines (*maps*).

Uncertainty pervades the data that scientists and all kinds of organizations use to inform decisions. Visual depictions of information can help clarify the uncertainty—or compound confusion. Ideally, visualizations help us make judgments, analytically and emotionally, about the probability of different outcomes. Abundant evidence on human reasoning suggests, however, that when people are asked to make judgments involving probability, they often discount uncertainty. As society increasingly relies on data, graphics designers are grappling with how best to show uncertainty clearly.

What follows is a gallery of visualization techniques for displaying uncertainty, organized roughly from less effective to more effective. Seeing how different approaches are chosen and implemented can help us become more savvy consumers of data and the uncertainty involved.

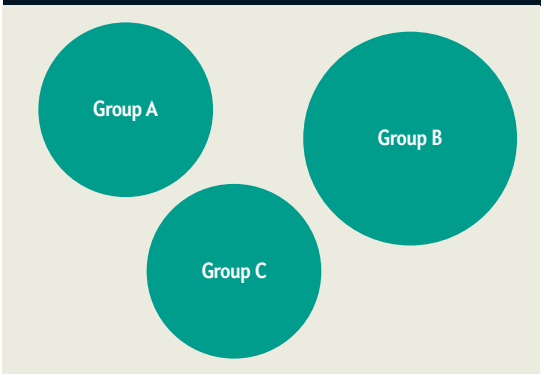


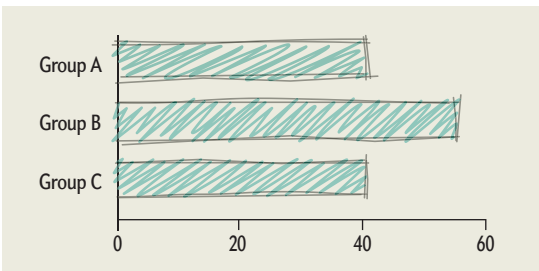


"CONE OF UNCERTAINTY" (*left*) shows where a hurricane may head, according to a group of forecasts. An alternative is to show the specific path predicted by each forecast (*right*). Both approaches have pros and cons in helping people judge the risk they may face, but the one on the right makes it clearer that the path is difficult to predict.

SOURCES: NATIONAL HURRICANE CENTER (*cone of uncertainty*); "VISUALIZING UNCERTAIN TROPICAL CYCLONE PREDICTIONS USING REPRESENTATIVE SAMPLES FROM ENSEMBLES OF FORECAST TRACKS," BY LE LIU ET AL., IN *IEEE TRANSACTIONS ON VISUALIZATION AND COMPUTER GRAPHICS*, VOL. 25; AUGUST 20, 2018 (*multiple storm paths*)

Maps by Tiffany Farrant-Gonzalez





NO QUANTIFICATION

The least effective way to present uncertainty is to not show it at all. Sometimes data designers try to compensate for a lack of specified uncertainty by choosing a technique that implies a level of imprecision but does not quantify it. For example, a designer might map data to a visual variable that is hard for people to define, such as a circle floating in space (*top*) rather than a dot on a graph that has *x* and *y* axes. This approach makes the reader's interpretation more error-prone. Alternatively a designer might use a program that creates a hand-drawn or "sketchy" feel (*bottom*). Both approaches are risky.

PROS

- If readers sense that a visualization is difficult to quantify or is simply impressionistic, they may be more cautious in making inferences or decisions based on it.

CONS

- Readers may not realize that the visualization is intended to convey imprecision and may reach conclusions that have large errors.
- Even if readers recognize that the visualization was chosen to imply imprecision, they have no way of inferring how much uncertainty is involved.

INTERVALS

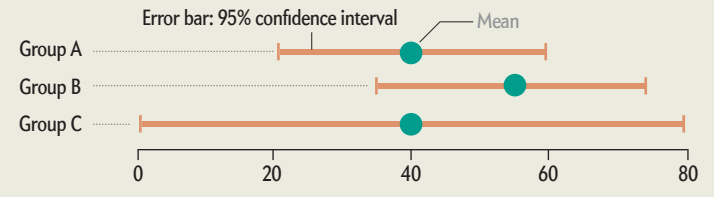
Intervals may be the most common representations of quantified uncertainty. Error bars (*top*) and confidence envelopes (*bottom*) are widely recognized, but even though they seem exact and straightforward, they are notoriously hard to interpret properly. Research shows they are often misunderstood, even by scientists.

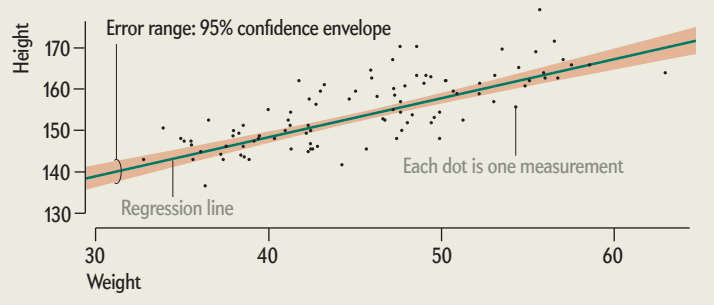
PROS

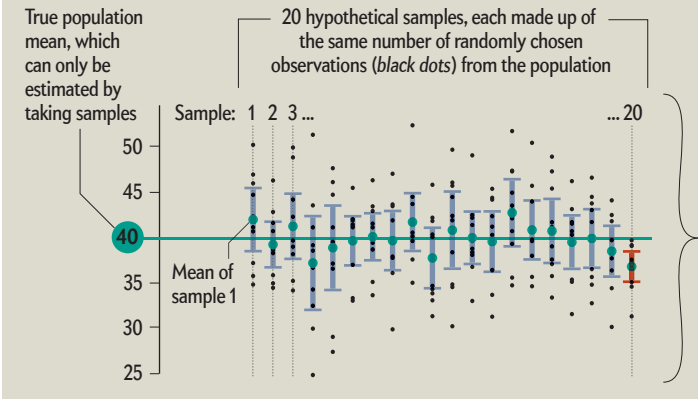
- Widely recognized as a representation of uncertainty.
- Offers a simple format for expressing the possibility of different values.
- The choice of interval can be customized for different types of questions about the same data set. For example, when one is making inferences about the range of values in a population, intervals based on standard deviation are helpful; for inferences about the range of values of a statistic like a mean, intervals based on standard error are appropriate.

CONS

- Ambiguity in what is shown: intervals may represent standard deviation, standard error or something else. Each has a unique interpretation.
- Readers can make "deterministic construal errors"—interpreting the ends of the error bar as the high and low values in observed measurements rather than estimates denoting uncertainty.
- Error bars can lead to "within-the-bar bias," common in bar charts. Below, readers may see the bar values to the right of the dots as more probable than the bar values to the left.
- Easy to ignore the uncertainty regions in favor of the central tendency, which may lead to incorrect decisions.







Even when calculated perfectly, on average, 1 in 20 of the 95% confidence intervals will not contain the population mean.

WHAT DOES A CONFIDENCE INTERVAL MEAN?

A natural interpretation of an error bar or confidence envelope that denotes 95 percent confidence is that the interval has a 95 percent chance of containing the true value. Yet it actually refers to the percentage of confidence intervals that would include the true value if an infinite number of random samples of the same size were pulled from the data and each time a 95 percent confidence interval was constructed. Although in practice this pervasive misinterpretation may not drastically change decisions, the fact that even scientists make such mistakes shows how challenging it can be to interpret uncertainty depictions correctly.

Graphics by Jessica Hullman and Jen Christiansen

PROBABILITY DENSITY MAPS

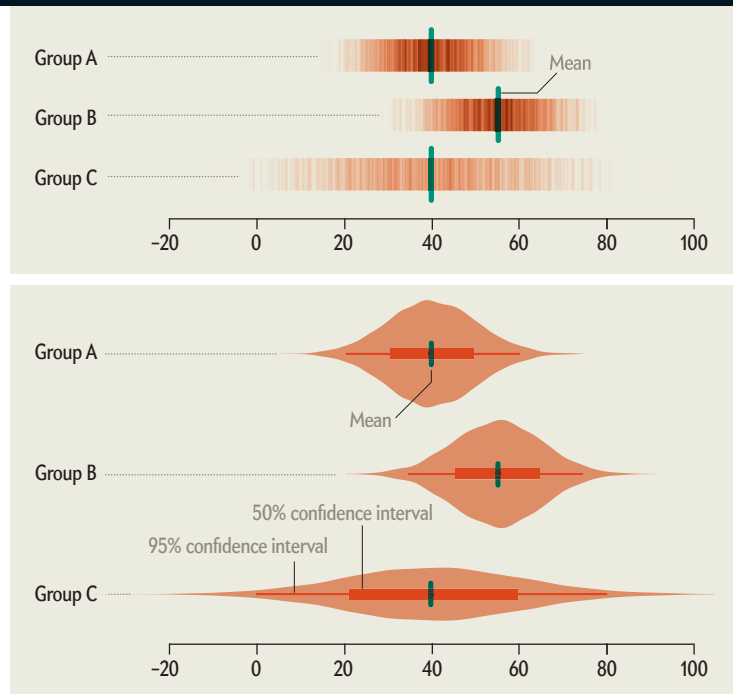
Designers can map uncertainty directly to a visual property of the visualization. For example, a gradient plot (*top*) can shift from dark color (high probability) at the center to lighter color (low probability) at the edges. In a violin plot (*bottom*), wider points mean greater probability. Mapping probability density to a visual variable displays uncertainty in greater detail than interval methods (error bars and confidence envelopes), but its effectiveness depends on how well readers can perceive differences in shading, height or other visual properties.

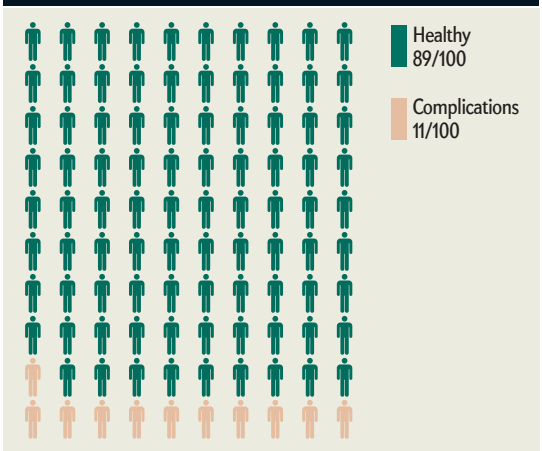
PROS

- Often well aligned with intuition: dark shading or hard boundaries are certain; light shading or fuzzy boundaries are uncertain.
- Avoids common biases such as those raised by intervals.

CONS

- Readers may not recognize that density reflects probability.
- Readers often equate the part of the visualization that is easiest to read (darkest, widest) with the data values themselves and misinterpret the parts that are harder to read (lightest, most narrow) as the uncertainty.
- Estimates can be biased to the darkest or highest points.
- Can be difficult to infer specific probability values.





ARRAYS OF ICONS

Reframing a probability such as 30 percent as a frequency—three out of 10—can make it easier for people to understand uncertainty and consequently use such information appropriately. People may better understand discrete probabilities because they run into them in everyday experiences.

PROS

- More self-explanatory than some other techniques because readers can readily see that probability is analogous to the number of times a symbol appears.
- Readers can make quick estimates if a small number of symbols is used because our visual system recognizes small quantities immediately without counting.

CONS

- Designed to present only a single probability.

MULTIPLE SAMPLES IN SPACE

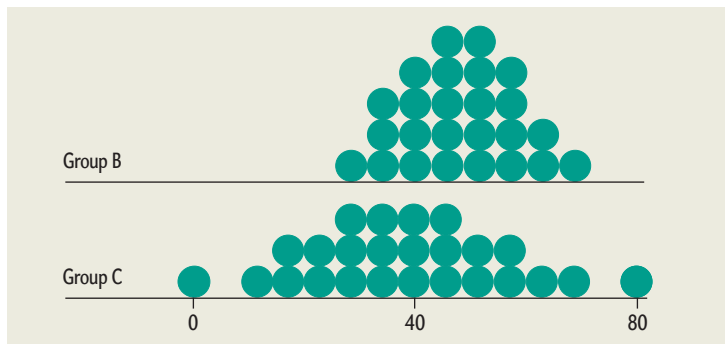
Plotting of multiple samples in space can be used to show probability in a discrete format for one or more variable quantities. One example of this approach is a quantile dot plot. It shows a number of distinct cases from the quantiles of the data distribution, so that the number of dots (such as two dots high or five dots high, in the example below) conveys probability. When there is uncertainty about parameter values from which estimates are drawn, such as initial conditions, samples can be generated that vary these parameters. and can be shown in a single visualization.

PROS

- A designer can choose how many data samples to present, aiming to show enough to convey the distribution but not so many that it becomes difficult for a reader to make out the individual samples.

CONS

- Plotting many data samples can result in occlusion, making probability estimates more error-prone.
- Sampling introduces imprecision, especially if the underlying distribution is heavily skewed by outliers.

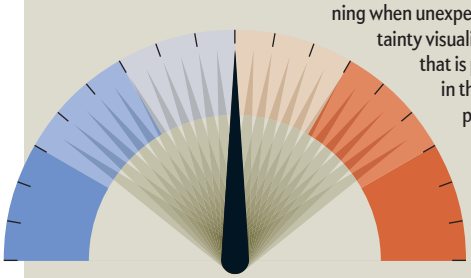


Graphics by Jessica Hullman and Jen Christiansen

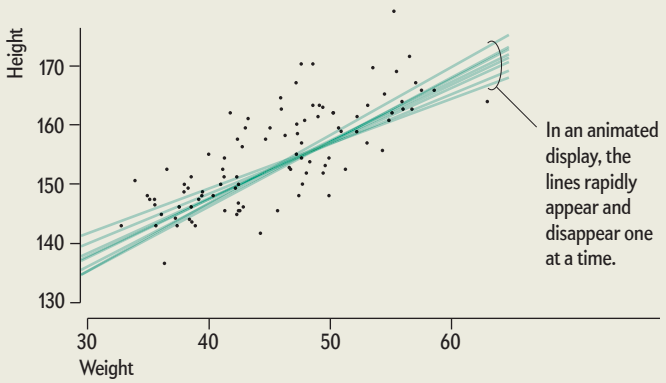
A JITTERY ELECTION NIGHT NEEDLE

Sometimes an uncertainty visualization is controversial. On the night of the 2016 presidential election in the U.S., the *New York Times* introduced an animated gauge on its Web site to display predictions about the outcome. A continuum of colored areas made up the background, from a landslide Clinton win (*left*) to a landslide Trump win (*right*). The data model behind the gauge updated several times a minute as new local results came in. An animated needle jiggled back and forth rapidly, even more frequently than the model was updated.

Seeing a constantly moving visualization made many viewers anxious on an evening when unexpected events transpired. Uncertainty visualizations *should* provoke anxiety that is proportional to the uncertainty in the data. But after decades of people seeing only static projections for election outcomes that allowed viewers to overlook uncertainty, suddenly shifting to a visualization that provoked a visceral reaction to uncertainty was unsettling.



Chance of winning election



MULTIPLE SAMPLES IN TIME

Plotting multiple possible outcomes as frames in an animation makes uncertainty visceral and much harder to ignore. This technique, called hypothetical outcome plots, can be used for simple and complex visualizations. Perceptual studies indicate that people are surprisingly adept at inferring the distribution of data from the frequency of occurrences: we do not necessarily need to count the number of times an event occurs to estimate its probability. One important factor is the speed of events, which must be fast enough so that people can see a sufficient number of samples yet slow enough for them to consciously register what they saw.

PROS

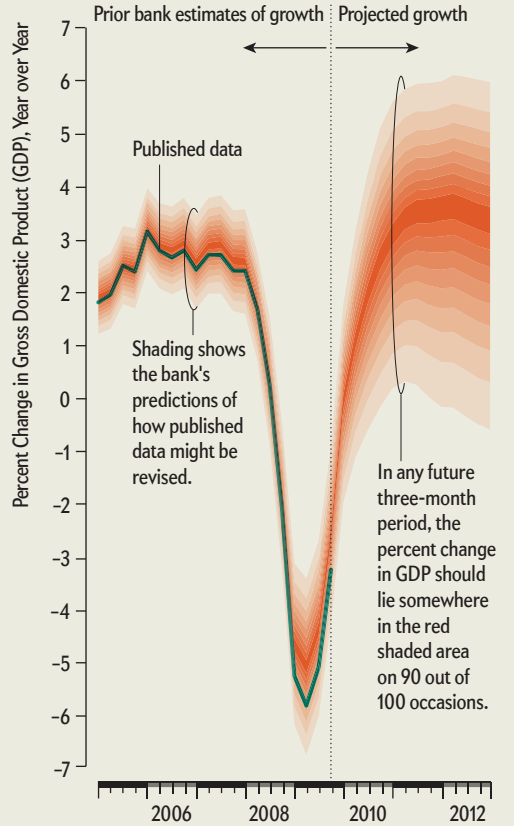
- The human visual system can estimate probability fairly accurately without having to deliberately count the items presented.
- Can be applied widely across different data types and visualization styles.
- Animation makes it possible to estimate probabilities involving multiple variables, which is difficult with static plots.

CONS

- Sampling introduces imprecision, especially if the distribution is heavily skewed by outliers.
- No guarantees on how many individual samples a user will pay attention to.
- Requires creating a dynamic or animated visualization, which some formats such as scientific papers may not yet easily support.

HYBRID APPROACHES

Designers can create effective uncertainty visualizations by combining different techniques rather than choosing a standard chart "type." One example is a fan chart, made famous by the Bank of England (*shown*). It depicts data up to the present (*left side of dotted line*), then projections into the future (*right side*); uncertainty about the past is an important component in assessing uncertainty about the future. The fan chart presents probability from higher chance (*dark shading*) to lesser chance (*light shading*) in multiple bands that represent different levels of confidence, which the reader can choose from. Readers can perceive the information through the position of the edges of the bands, as well as lightness versus darkness. Some modern software packages for statistical graphics and modeling make it easy to combine uncertainty visualization approaches.



SOURCE: *INFLATION REPORT*, BANK OF ENGLAND, FEBRUARY 2010 (*GDP chart*)

MORE TO EXPLORE

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

Visualizing Uncertainty. Claus O. Wilke in *Fundamentals of Data Visualization*. O'Reilly Media, 2019.

Uncertainty + Visualization, Explained. Blog series by Jessica Hullman and Matthew Kay. https://medium.com/multiple-views-visualization-research-explained

FROM OUR ARCHIVES

Saving Big Data from Itself. Alex "Sandy" Pentland; August 2014.

scientificamerican.com/magazine/sa

Illustration by Tiffany Farrant-Gonzalez (election needle)