CSCI 491: Data Visualization

5- From Data to Visualization
Dictionary of Visualizations

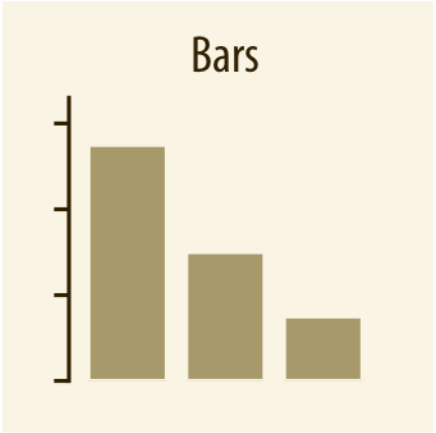
When turn massive data into visual form

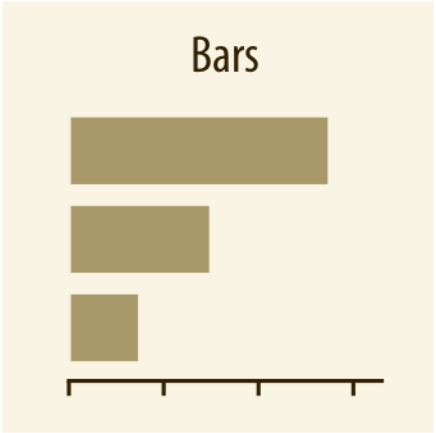
- Creating a visualization requires a number of nuanced judgments. One must determine which questions to ask, identify the appropriate data, and select effective visual encodings to map data values to graphical features such as position, size, shape, and color.

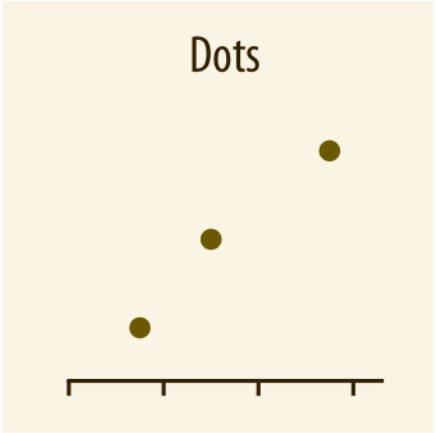
Visualizing amounts

The most common approach to visualizing amounts (i.e., numerical values shown for some set of categories) is using bars, either vertically or horizontally arranged.

However, instead of using bars, we can also place dots at the location where the corresponding bar would end.

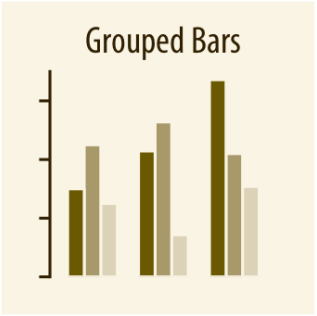


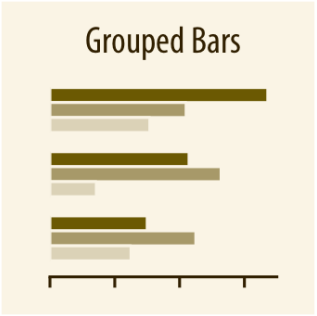


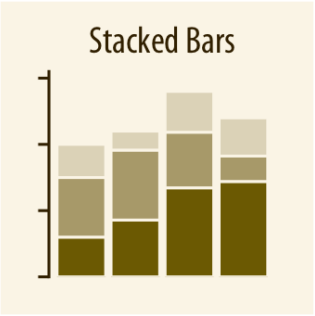


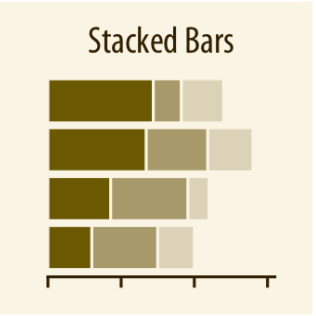
Visualizing amounts (multi-category)

- If there are two or more sets of categories for which we want to show amounts, we can group or stack the bars. We can also map the categories onto the x and y axes and show amounts by color, via a heatmap.



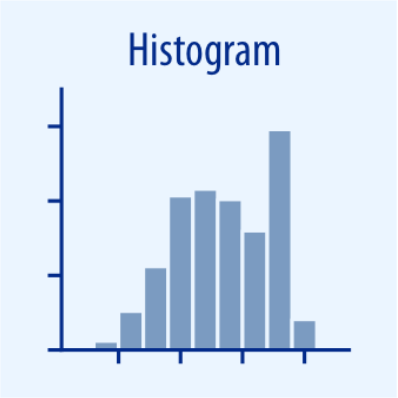


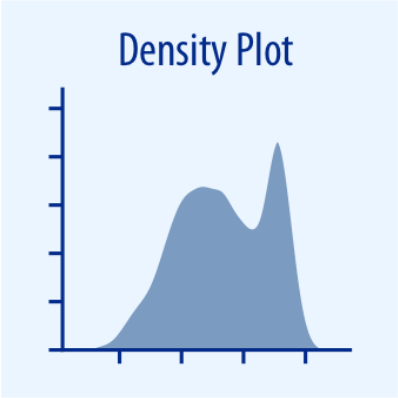


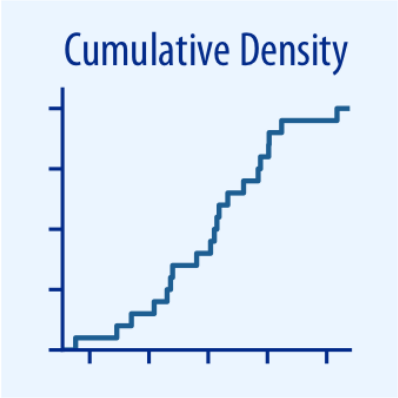


Distributions

- Histograms and density plots provide the most intuitive visualizations of a distribution, but both require arbitrary parameter choices and can be misleading. Cumulative densities and quantile-quantile (q-q) plots always represent the data faithfully but can be more challenging to interpret.

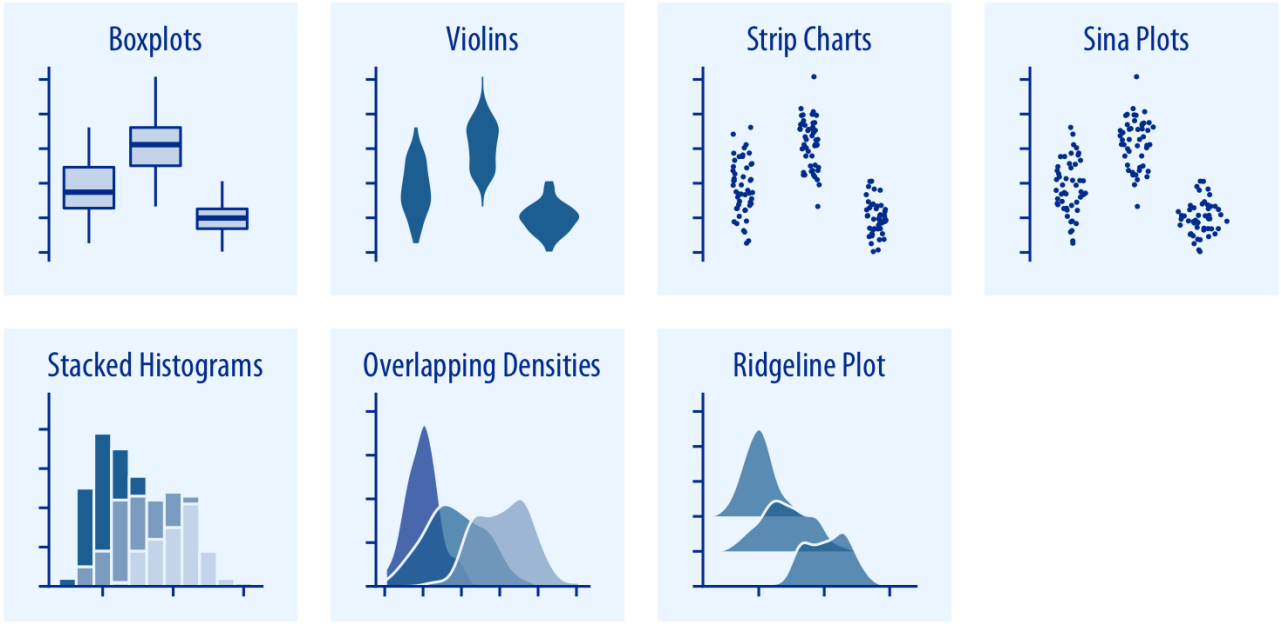






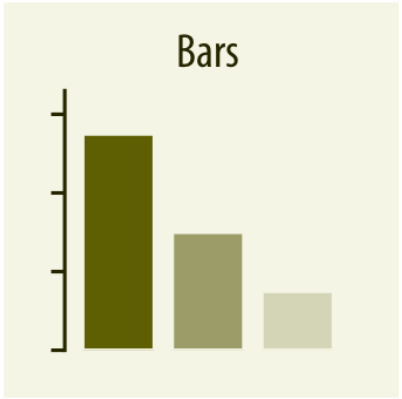
Distributions (many distributions)

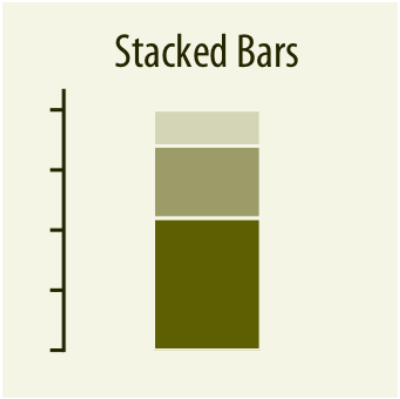
- Boxplots, violin plots, strip charts, and sina plots are useful when we want to visualize many distributions at once and/or if we are primarily interested in overall shifts among the distributions. Stacked histograms and overlapping densities allow a more in-depth comparison of a smaller number of distributions, though stacked histograms can be difficult to interpret and are best avoided. Ridgeline plots can be a useful alternative to violin plots and are often useful when visualizing very large numbers of distributions or changes in distributions over time.



Proportions

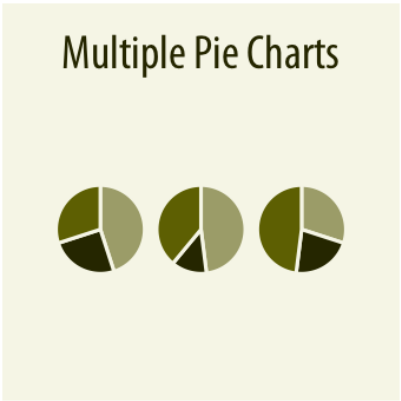
- Proportions can be visualized as pie charts, side-by-side bars, or stacked bars. As for amounts, when we visualize proportions with bars, the bars can be arranged either vertically or horizontally. Pie charts emphasize that the individual parts add up to a whole and highlight simple fractions. However, the individual pieces are more easily compared in side-by-side bars. Stacked bars look awkward for a single set of proportions, but can be useful when comparing multiple sets of proportions.

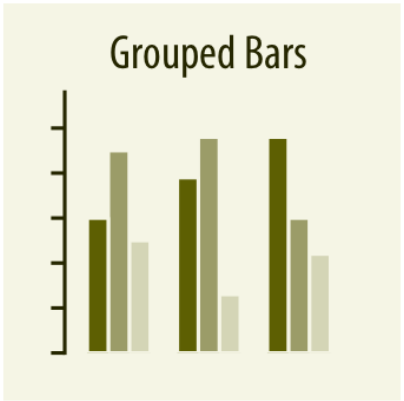


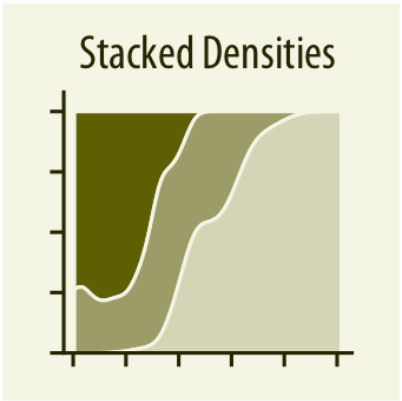


Proportions (multiple sets)

- When visualizing multiple sets of proportions or changes in proportions across conditions, pie charts tend to be space-inefficient and often obscure relationships. Grouped bars work well as long as the number of conditions compared is moderate, and stacked bars can work for large numbers of conditions. Stacked densities are appropriate when the proportions change along a continuous variable.





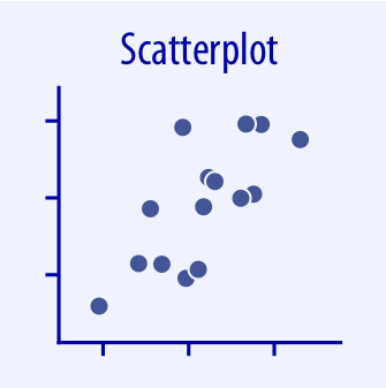


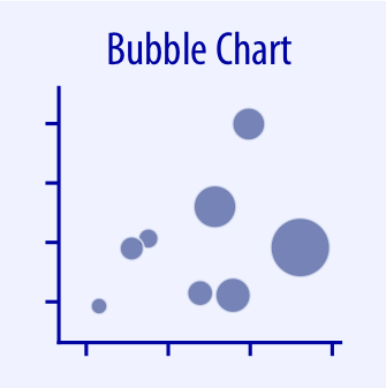
Proportions (multiple grouping)

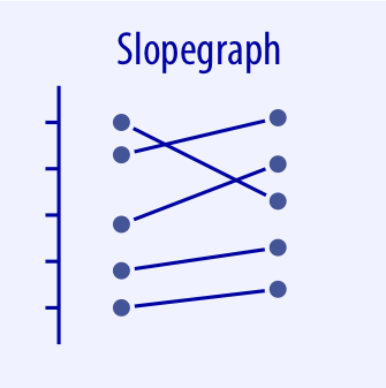
- When proportions are specified according to multiple grouping variables, mosaic plots, treemaps, or parallel sets are useful visualization approaches. Mosaic plots assume that every level of one grouping variable can be combined with every level of another grouping variable, whereas treemaps do not make such an assumption. Treemaps work well even if the subdivisions of one group are entirely distinct from the subdivisions of another. Parallel sets work better than either mosaic plots or treemaps when there are more than two grouping variables.

x–y relationships

- Scatterplots represent the archetypical visualization when we want to show one quantitative variable relative to another. If we have three quantitative variables, we can map one onto the dot size, creating a variant of the scatterplot called a bubble chart. For paired data, where the variables along the x and y axes are measured in the same units, it is generally helpful to add a line indicating x = y. Paired data can also be shown as a slopegraph of paired points connected by straight lines.

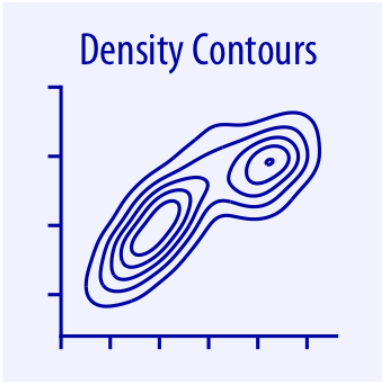


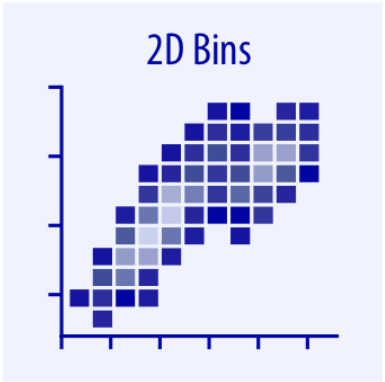


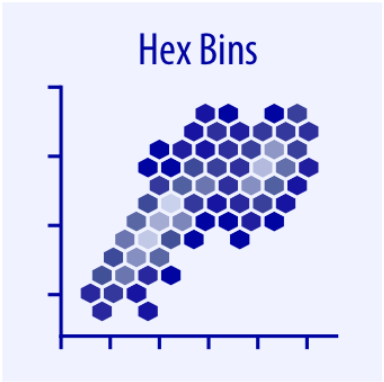


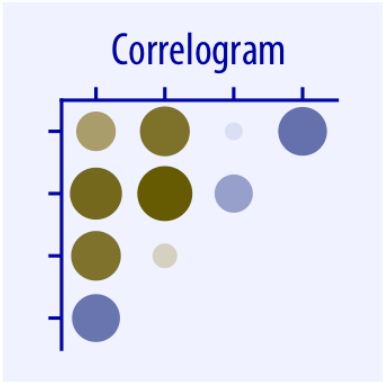
x–y relationships (large number of points)

- For large numbers of points, regular scatterplots can become uninformative due to overplotting. In this case, contour lines, 2D bins, or hex bins may provide an alternative. When we want to visualize more than two quantities, on the other hand, we may choose to plot correlation coefficients in the form of a correlogram instead of the underlying raw data.



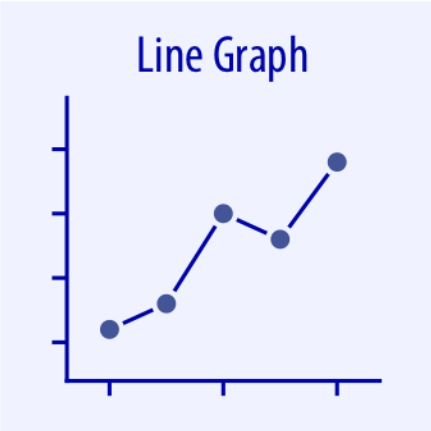


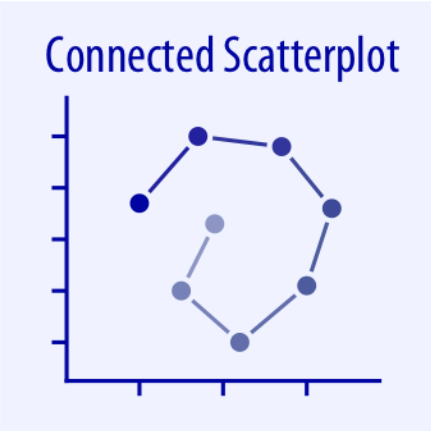


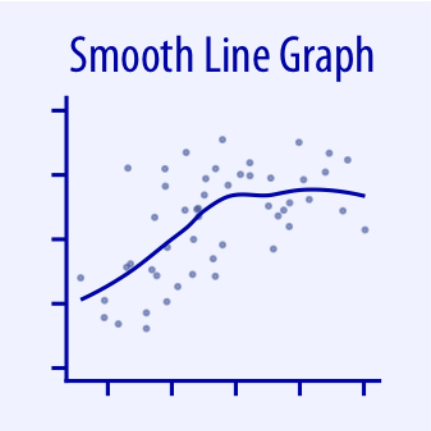


x–y relationships (trends)

- When the x axis represents time or a strictly increasing quantity such as a treatment dose, we commonly draw line graphs. If we have a temporal sequence of two response variables we can draw a connected scatterplot, where we first plot the two response variables in a scatterplot and then connect dots corresponding to adjacent time points. We can use smooth lines to represent trends in a larger dataset.

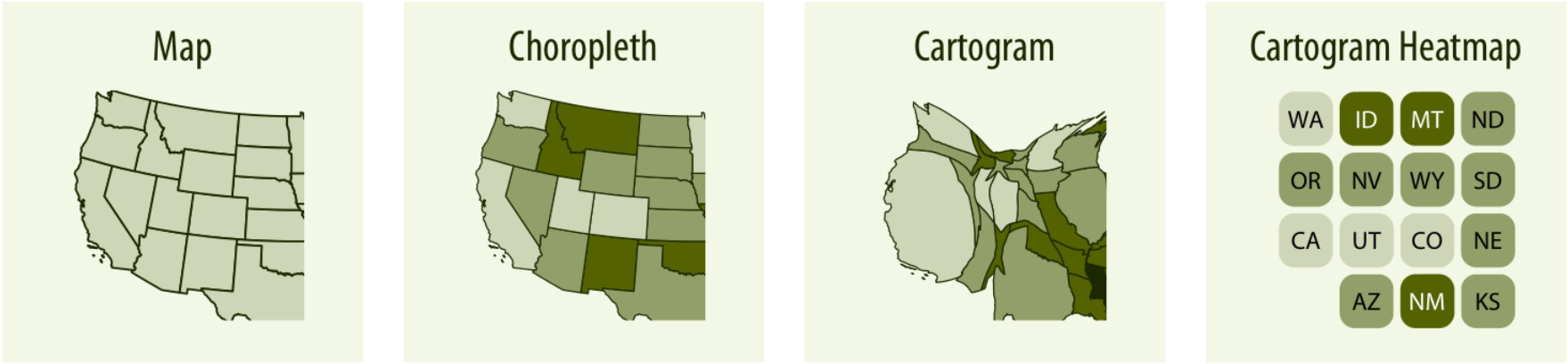






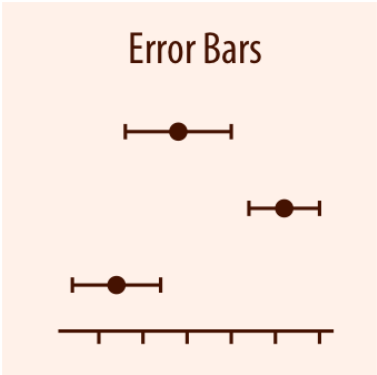
Geospatial Data

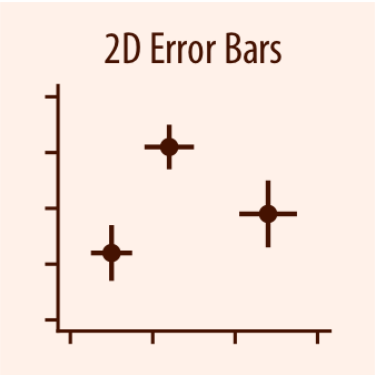
- The primary mode of showing geospatial data is in the form of a map. A map takes coordinates on the globe and projects them onto a flat surface, such that shapes and distances on the globe are approximately represented by shapes and distances in the 2D representation. In addition, we can show data values in different regions by coloring those regions in the map according to the data. Such a map is called a choropleth. In some cases, it may be helpful to distort the different regions according to some other quantity or simplify each region into a square. Such visualizations are called cartograms (see "Cartograms").

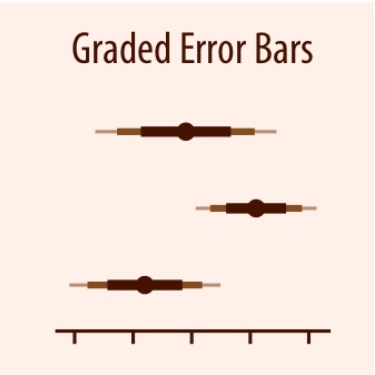


Uncertainty

- Error bars are meant to indicate the range of likely values for some estimate or measurement. They extend horizontally and/or vertically from some reference point representing the estimate or measurement.
- Reference points can be shown in various ways, such as by dots or by bars. Graded error bars show multiple ranges at the same time, where each range corresponds to a different degree of confidence. They are in effect multiple error bars with different line thicknesses plotted on top of each other.

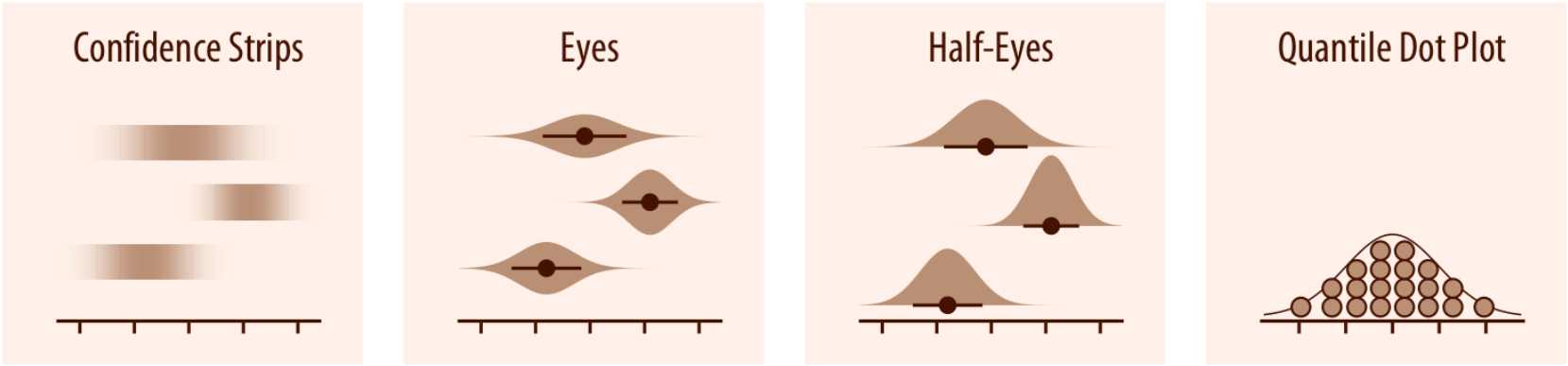






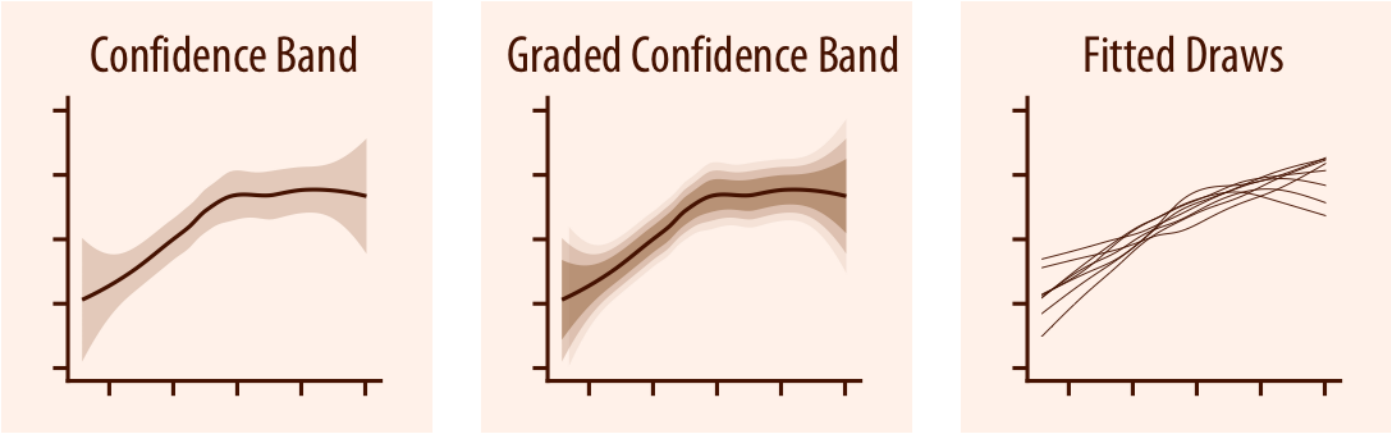
Uncertainty (confidence range)

- To achieve a more detailed visualization, we can visualize the actual confidence or posterior distributions.
- Confidence strips provide a visual sense of uncertainty but are difficult to read accurately. Eyes and half-eyes combine error bars with approaches to visualize distributions (violins and ridgelines, respectively), and thus show both precise ranges for some confidence levels and the overall uncertainty distribution. A quantile dot plot can serve as an alternative visualization of an uncertainty distribution. Because it shows the distribution in discrete units, the quantile dot plot is not as precise but can be easier to read than the continuous distribution shown by a violin or ridgeline plot.



Uncertainty (confidence band)

- For smooth line graphs, the equivalent of an error bar is a confidence band. It shows a range of values the line might pass through at a given confidence level. Like with error bars, we can draw graded confidence bands that show multiple confidence levels at once. We can also show individual fitted draws in lieu of or in addition to the confidence bands.



More reading

- Storytelling with data, chapter 2, "choosing an effective visual"
- A Tour through the Visualization Zoo