The size and complexity of modern datasets has far outstripped the capabilities of common existing methods for visualization. Current tools work well for small numbers of data points, where each point is represented by many pixels and is thus individually perceivable by the human visual system. However, when there are many more data points than pixels, it is crucial to accurately convey the *distribution* of points, i.e., the overall structure and pattern of the data, which typically emerges only indirectly via patterns of overplotting and summing of pixel values. Achieving a faithful visualization with scatterplots or heatmaps usually requires knowledge of the underlying distribution of datapoints, either from a priori expertise or through exploration, which presents serious conceptual and practical problems for understanding new, unknown large datasets.

In this talk, I will describe a new approach, called Datashading, that is a novel adaptation of the visualization pipeline which provides a principled way to visualize large and complex datasets. Datashading is based on the simple idea of using binning techniques to retain the original data values as far into the visualization pipeline as possible, even to the pixel level. As implemented in the new Python datashader library (https://github.com/bokeh/datashader), this approach allows algorithmic computations to replace trial-and-error approaches at each stage of processing.

I will demonstrate how these techniques can be used for flexible, interactive, and practical visualization of even extremely large datasets with billions of points.