POINTS OF VIEW

Data exploration

Enhancement of pattern discovery through graphical representation of data.

Data visualization can serve two distinct purposes: to communicate research findings and to guide the data-exploration process as the scientific story is unfolding. Each goal entails a different approach to data representation, but sound graphic design principles are important in both. This column is the first in a series that will focus on data-visualization techniques intended to support data exploration.

Exploring data to understand the underlying structure is fundamentally different from presenting known characteristics of the data. In a presentation, a researcher has already identified an interesting structure in the data and is trying to highlight it. In exploration, the researcher suspects that regularities are present but does not know exactly what they are. Instead of emphasizing any one aspect, graphical representation is used to provide overviews in which meaningful patterns may be found.

Patterns are the essence of data exploration, and the eye's ability to discern form makes visual display integral to the process. The visual display of quantitative information can help us see connections in the data. Unlike tables of numbers in which there is little visual connection between the elements, graphs allow us to easily detect data objects with similar physical properties and assemble them into a formation. Data exploration is an iterative process in which expectations and hypotheses guide a graphical organization of the data, and patterns observed in the data germinate new or refined hypotheses.

It is essential to look at data in a graphical form and not rely solely on computational metrics. Anscombe's quartet[1] is a compelling example of this (**Fig. 1**). The four sets of numbers in the quartet have many identical summary statistics (for example, mean of *x* values, mean of *y* values, variances, correlations and regression lines) but vary wildly when graphed. In this example there are only two variables in four groups. In realistic scenarios, however, where datasets are typically much larger, the question of how to display the data visually is substantially more complex.

a

I		II		III		IV	
x	*y*	*x*	*y*	*x*	*y*	*x*	*y*
10	8.04	10	9.14	10	7.46	8	6.58
8	6.95	8	8.14	8	6.77	8	5.76
13	7.58	13	8.74	13	12.74	8	7.71
9	8.81	9	8.77	9	7.11	8	8.84
11	8.33	11	9.26	11	7.81	8	8.47
14	9.96	14	8.10	14	8.84	8	7.04
6	7.24	6	6.13	6	6.08	8	5.25
4	4.26	4	3.10	4	5.39	19	12.5
12	10.84	12	9.13	12	8.15	8	5.56
7	4.82	7	7.26	7	6.42	8	7.91
5	5.68	5	4.74	5	5.73	8	6.89

b

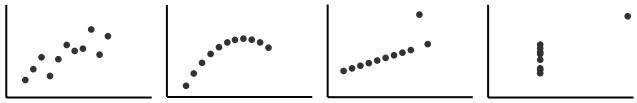

Figure 1 | Anscombe's quartet. (**a**) The four sets of numbers that form Anscombe's quartet. (**b**) The highly distinctive graphs that result from plotting the data in **a**.

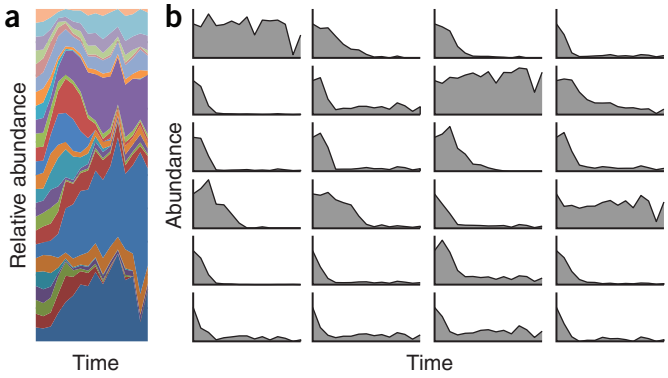

Figure 2 | Small multiples. (**a**) A stack graph showing the relative proportions of 24 cell lines over time. (**b**) Individual growth curves for the data graphed in **a**.

With a high-dimensional dataset, a common exploratory goal is to find 'classes of behavior' among multiple components (for example, genes, populations, samples and so on). A useful strategy is to create simple representations of low-dimensional 'slices' of the data. Ideally we want to restrict the complexity to one plot for each component. To make the visual task of finding commonality between the plots simpler, ensure consistency between the elements being inspected. For example, using a common scale allows the plots to be directly compared.

In the example depicted in **Figure 2**, 24 types of cells had been cultured together in an attempt to study the cells' growth characteristics in a mixture. Representing the relative abundance of all the cell types as a stack graph (**Fig. 2a**) makes it clear that different populations fare differently in this community over time. However, because of the interdependencies between all curves in a stack graph, it is difficult to see additional trends in this overview. By plotting the abundance of each population as a function of time (**Fig. 2b**) several common behaviors can be observed. As the research objective translates to categorizing shapes of curves, we support this visual task by filling the area under the curves, which accentuates their form.

Displaying too much data simultaneously often presents a visual burden that should be avoided. To address this, some data must be left out. In **Figure 2**, for example, we limited our observations to one of four replicates. In instances where the number of components is high (for example, if we had 1,000 instead of 24 cell populations), sampling a subset is a sensible option. As we begin to understand the structure that underlies the data, we can point to features of the data that are of less interest and can therefore be removed from our plots. Focusing on a small number of remaining features allows us to bring additional components into the graphs and gradually attain a more global view of the data.

Over the next several months, we will investigate visualization techniques for extracting meaning from data.

COMPETING FINANCIAL INTERESTS
The authors declare no competing financial interests.

Noam Shoresh & Bang Wong

1. Anscombe, F. J. *Am. Statistician* **27**, 17–21 (1973).

Noam Shoresh is a senior computational biologist at the Broad Institute. Bang Wong is the creative director of the Broad Institute and an adjunct assistant professor in the Department of Art as Applied to Medicine at The Johns Hopkins University School of Medicine.