POINTS OF VIEW

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.

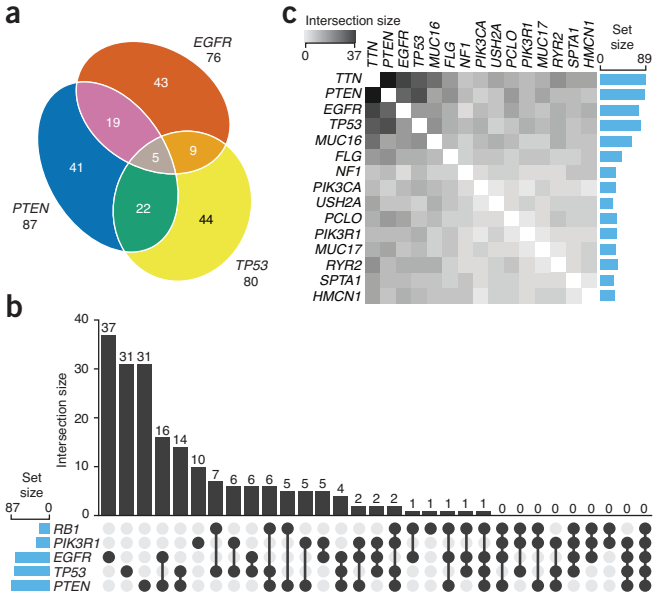

Figure 1 | Set visualization techniques. (**a**) Euler diagram displaying the intersections of three genes. Sets are genes mutated in tumors of patients with glioblastoma multiforme[3], and set intersections indicate genes that are co-mutated. The number of patients shown in **a**, **b** and **c** varies 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.