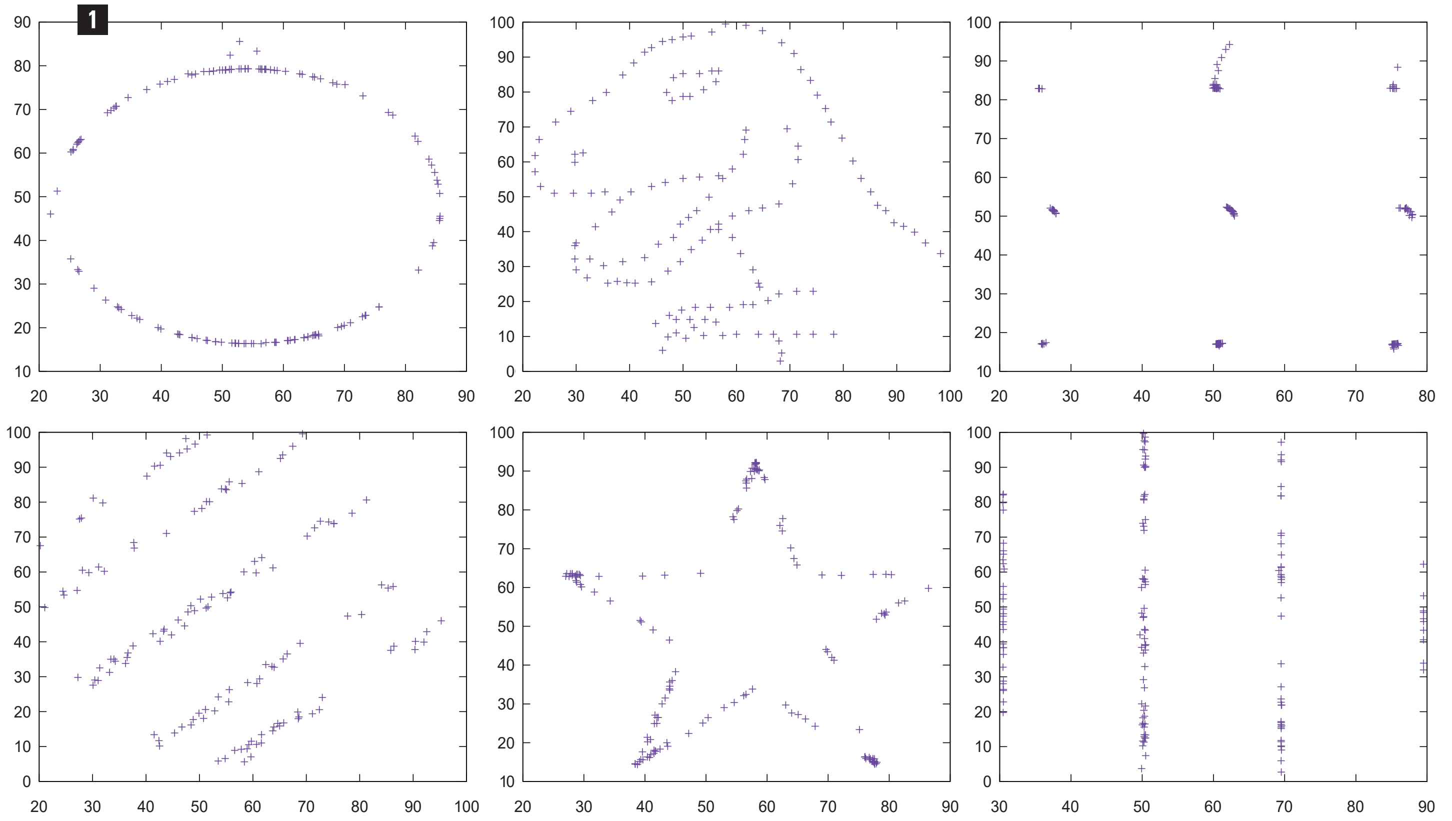
DOI: 10.24425/academiaPAS.2022.144686

1



Visualizing Maps in Practice

Paweł Dłotko, PhD
completed his master's and doctorate in computational mathematics at the Jagiellonian University. He has worked at the University of Pennsylvania in the United States, the Inria Saclay Centre in France, and the University of Swansea in Wales. He leads the Dioscuri Centre in Topological Data Analysis at the PAS Institute of Mathematics.
pawel.dlotko@impan.pl

Visualizations of mathematical functions have myriad applications in our daily lives, from the economy all the way to medicine.

Paweł Dłotko

Institute of Mathematics,
Polish Academy of Sciences, Warsaw

Mathematics integrates and generalizes many image-related concepts by associating them with a function, which matches values of defined arguments (the domain) to corresponding output values (the codomain). In mathematics, therefore, the "image" of a function is the set of all output values it produces.

Functions are ubiquitous, present everywhere in our day-to-day lives. Let's take one example: a photo stored in your phone's memory. Its domain is a collection of pixels – tiny squares on the screen – while its codomain is the set of possible colors, typically defined using the additive color model of red, green, and blue (RGB). Each pixel is illuminated with a particular combination of these colors, and they all come together to show your photo. In simplified terms, we

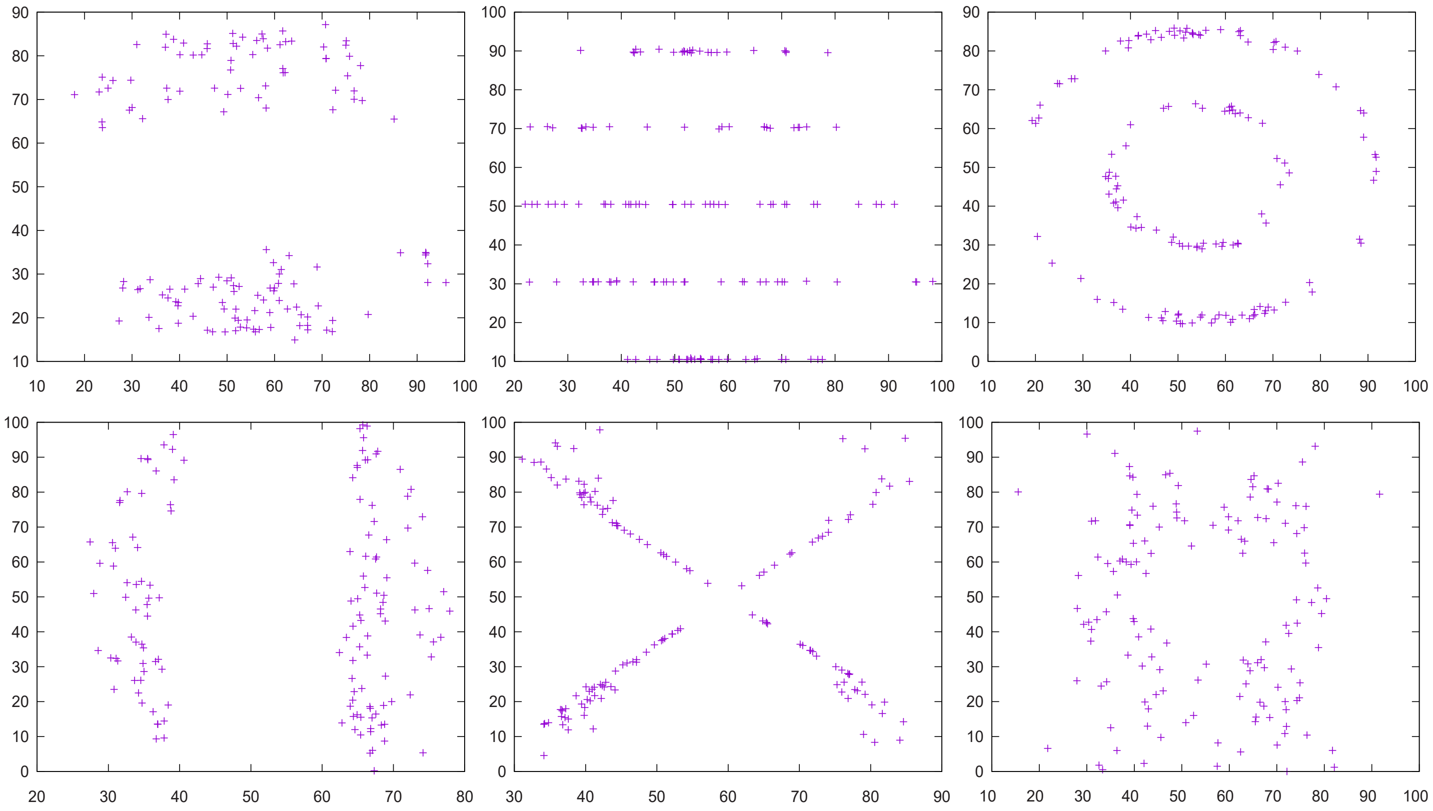


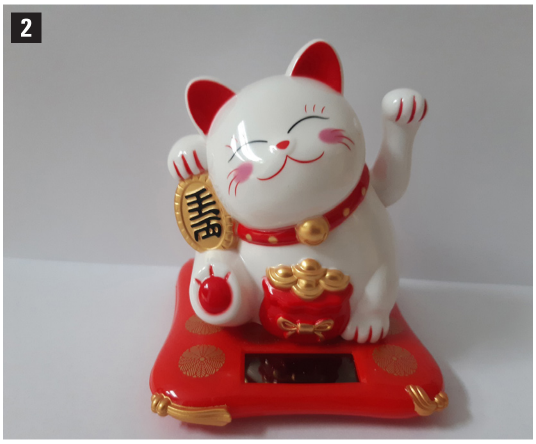
Fig. 1
The Datasaurus Dozen – a set of point clouds that all share very similar descriptive statistics values, but nevertheless exhibit very different shapes

could say that the domain is the plane (phone screen) and the image is a 3D space where numerical values assigned to each dimension describe the intensity of each color. Old black-and-white photographs are easier to interpret. Their domain is, once again, a collection of pixels and the image is a palette of different shades of gray, which can be expressed as a single number representing the saturation of black.

Point clouds

In data analysis, functions are defined on a multidimensional space. Traditionally, we use two-dimensional space, placing a point in a Cartesian coordinate system with X and Y axes, which allows us to describe it using a pair of numerical values. But if we want to depict a point in 12 dimensions in the same way, we can assign it 12 values which define its individual attributes such as height, width, length, geographical coordinates, and any other seven characteristics we might want to attribute to it. This full set of attributes then describes the position of the point in a 12-dimensional space. Finite sets of points, known as “point clouds,” each have some shape, which frequently carries important information. The Datasaurus Dozen (Fig. 1) provides some excellent examples of different 2D point clouds.

The 12 point clouds shown above exhibit considerably different shapes. However, descriptive statistical values such as mean, standard deviation, and correlation for the X and Y coordinates turn out to be very similar for all 12 of them. The example in Fig. 1 shows, therefore, that data visualization is an extremely important element of data analysis. Here, each point has just two coordinates, so the entire cloud has two dimensions, making generating an image straightforward. But what happens when the cloud under consideration has dozens, hundreds, or even thousands of dimensions? Here let’s use the Maneki-neko cat figurine as an example (Fig. 2). Consider a series of black-and-white photos of the figurine, each measuring 128×128 pixels, all snapped from the same distance at consecutive angles running in a circle all the way around the object.



PAWEŁ DŁOTKO

Fig. 2
The Maneki-neko cat figurine. A set of photographs of the figurine has its own shape, determined by the various angles from which each photo is taken, and we can understand this shape from its graph

3

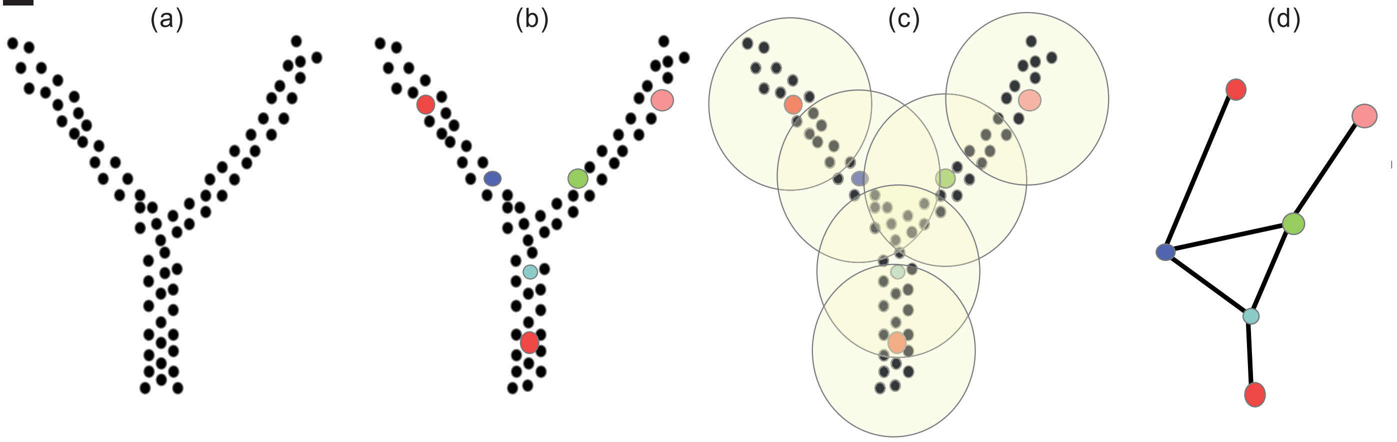


Fig. 3
The concept of the "ball mapper" (BM) algorithm, illustrated with the example of a 2D point cloud

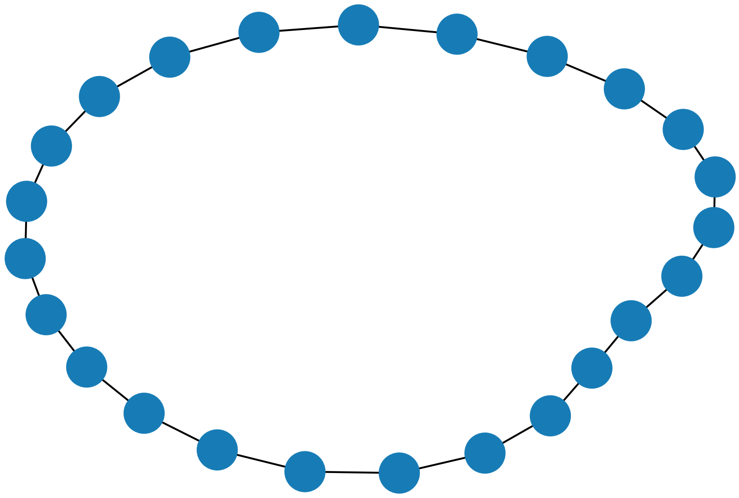
Fig. 4
Graph of vertices corresponding to photos of the Maneki-neko figurine in a 16,384-dimensional space

The pixels in each photo, numbered from 1 to 16,384 (128×128), can together be taken as describing a single point in a 16,384-dimensional space – each photo, then, represents a single "point." The subsequent coordinates describing this point correspond to the level of brightness (on a gray scale) of each individual pixel in the given photo. By collecting all the photos and assigning them corresponding points, we obtain a point cloud. The shape of the resulting cloud can then be visualized using a mapper algorithm, so as to "illustrate" the nature of the complex, multidimensional dataset as an abstract graph (consisting of vertices and edges).

How does this work? Fig. 3 shows the concept that underlies one such algorithm, called a "ball mapper" (BM). It is illustrated with a Y-shaped 2D point cloud (Fig. 3a).

In the first step of the BM algorithm, a uniformly-distributed subset, marked as colored dots, is extracted from the point cloud (Fig. 3b). These colored dots are then taken as the center of "balls" of a certain radius, which together contain all the points in the cloud (Fig. 3c). From this collection of balls, we then build up a graph (Fig. 3d). Importantly, the vertices in the resulting graph should not be seen as situated in space – rather, they are abstract objects. In our example, each vertex represents a ball shown in Fig. 3c, and the edges connect only those balls that share some overlap in the original Y-shaped point cloud. The procedure is easy enough to imagine in 2D, but in fact it can be applied to point clouds of any dimension. After applying the procedure to the photos of the Maneki-neko figurine (taken as points in a 16,384-dimensional space), we obtained the image below, showing a closed-cycle graph – a path we can travel around by hopping from one vertex to another (Fig. 4).

4



The dimensions and shape of the chosen graph represent the nature of our data. Each photo represents a point in 16,384-dimensional space. Each pair of subsequent photos are very similar, therefore the corresponding points are close together. However, as we start from the first photo, each subsequent image becomes more distinct, and after almost a full revolution they start becoming similar again. The same happens to the points representing each photo: they start off by becoming more distant from the starting point only to return to it from the other direction. This is how we return to the starting point after completing the full circle in a multidimensional space.

Visualizing functions

Let's consider a situation when every point in a cloud is assigned an additional value, returned by a certain function on the point cloud. The mean value of such a function for the points covered by a given "ball" can be visualized using a color scale on the vertices of the graph. This can be illustrated, for instance, by a set of banknotes, each of which is a data point described by four different characteristics. The function then assigns each banknote a value: it can be genuine (function value 0) or fake (1). Converting this dataset to an image using the BM algorithm produces a Y-shaped graph (Fig. 5).

The color of the vertices in the resulting graph conveys information about the nature of the banknotes it represents (the mean value of the points falling within the corresponding ball), with purple indicating genuine and yellow fake banknotes. The two yellow arms evident in the graph suggest that we are dealing with two distinctly different kinds of fake banknotes: there are two different sets of yellow vertices not directly connected and situated a certain distance from one another.

Multidimensional data is actually very common, and its visualization is crucial in many fields. As a last example, let us consider a dataset collected by the Netherlands Cancer Institute, describing the activity of various genes in a group of breast cancer patients. The multidimensional points in the dataset represent the activity of a thousand different genes in each patient. Our aim is, based on this data, to provide a prognosis and propose effective targeted therapy for each patient. The classic mapper algorithm (developed at Stanford in 2007) produces a graph that represents an image of the activity of a thousand genes, with the function being patient survival rates.

Like in the case of the banknotes, the resulting domain turns out to be Y-shaped. Of particular interest are the two arms on the left side of the graph (Fig. 6). The arm pointing downwards represents "triple-negative" cancers with a poor prognosis and low survival rates. The arm pointing upwards is more positive: the end shows (in red) a previously unknown subgroup of patients with high survival rates. Further analysis reveals that this subgroup is characterized by cancers that are Estrogen Receptor-positive (ER+), with high levels of c-MYB and low levels of innate inflammatory genes. The combination of these three parameters is currently being used to devise therapies. This subgroup of patients was identified by imaging multidimensional data.

5

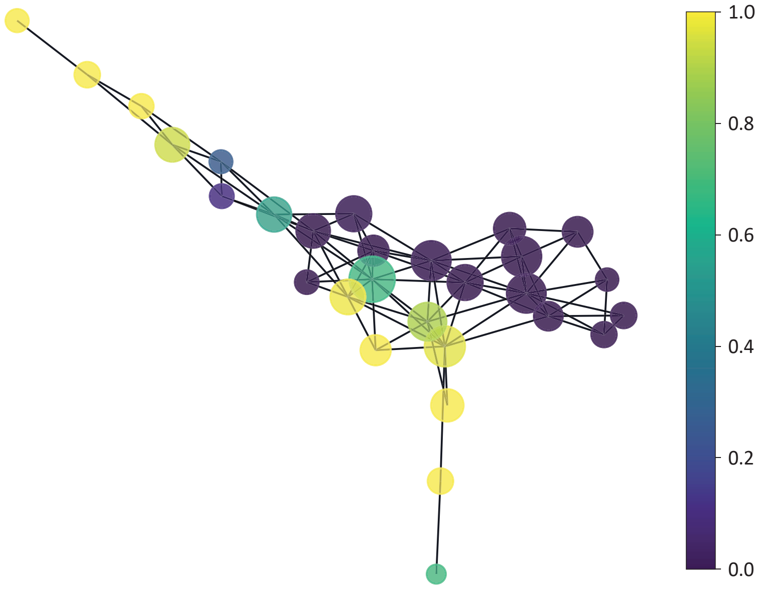


Fig. 5
Four-dimensional space showing the characteristics of genuine (purple) and fake (yellow) banknotes

There are many other examples of multidimensional data and its analysis. Although state-of-the-art machine learning and AI methods make it possible to analyze data blindly, the ability to effectively visualize data and directly analyze the resulting images can provide an additional level of understanding of a given phenomenon (as in the above examples), and it can help scientists posit rational, objective explanations. Mathematical methods need to be supplemented with effective and user-friendly implementations, accessible to anyone. These and other methods are often a source of new discoveries in many fields, stretching from the humanities, through chemistry, physics, and medicine, all the way to pure mathematics. Its myriad applications go to show that contemporary computational mathematics truly is the "queen of the sciences." ■

6

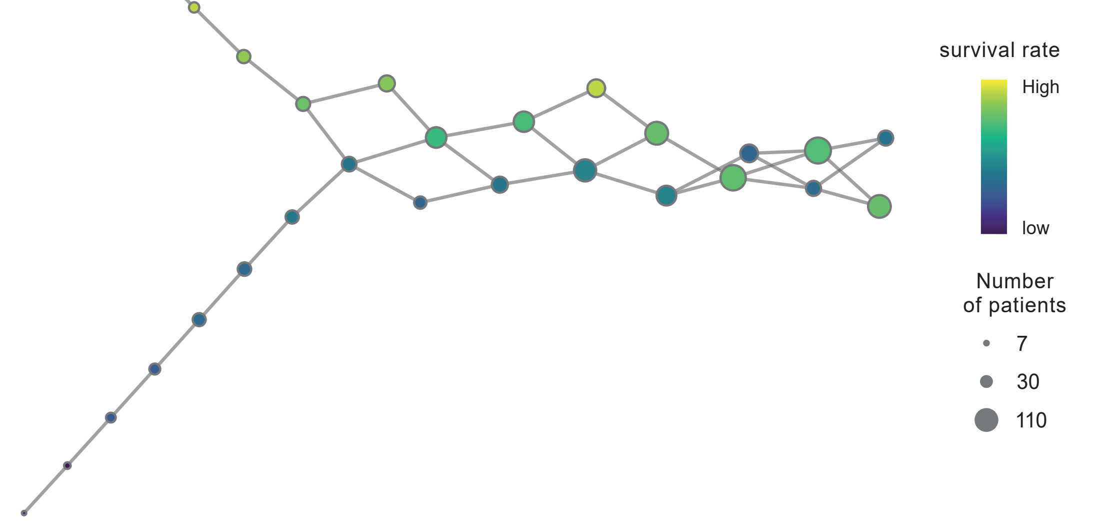


Fig. 6
Image of the activity of a thousand genes for breast cancer patients. The highlighted red branch in the upper left corner represents a previously unknown subgroup with a high survival rate

Further reading:

Explanation of the Ball Mapper Algorithm, https://www.youtube.com/watch?v=M9Dm1nl_zSQ&ab_channel=Pawe%C5%82D%C5%82otko

Ghrist, R. Barcodes: The Persistent Topology of Data, *Bulletin of the American Mathematical Society* 45(1) 2008. https://www.ams.org/journals/bull/2008-45-01/S0273-0979-07-01191-3/S0273-0979-07-01191-3.pdf

Feng, M. et al. Connecting the Dots: Discovering the "Shape" of Data. *Frontiers for Young Minds*. 2021. https://kids.frontiersin.org/articles/10.3389/frym.2021.551557