A 5-step guide to data visualization

elsevier.com/connect/a-5-step-guide-to-data-visualization

Data has been described as the new raw material for business and the "oil of the 21st century."

The volume of data used in business, research and technological development is massive and continues to grow. For instance at Elsevier, there are about 700 million articles per year downloaded from ScienceDirect, 80,000 institution profiles on Scopus, 13 million researcher profiles on Scopus and 3 million researcher profiles on Mendeley. It becomes harder and harder for a user to grab a key message from this universe of data.

That's where data visualization comes in: summarizing and presenting large data in simple and easy-to-understand visualizations to give readers insightful information.

There are many advanced visualizations (e.g., networks, 3D-models and map overlays) used for specialized purposes such as 3D medical imaging, urban transportation simulation, and disaster relief monitoring. But regardless of the complexity of a visualization, its purpose is to help readers see a pattern or trend in the data being analyzed, rather than having them read tedious descriptions such as: "A's profit was more than B by 2.9% in 2000, and despite a profit growth of 25% in 2001, A's profit became less than B by 3.5% in 2001." A good visualization summarizes information and organizes in a way that enables the reader to focus on the points that are relevant to the key message being conveyed.

For our projects in Elsevier's Analytical Services, we are constantly looking for ways to advance data analysis and visualization. For example, in our analysis of research performance, research collaboration is an area with vast amounts of data available. In our report *Comparative Benchmarking of European and US Research Collaboration and Researcher Mobility* for Science Europe, cross-state and international collaborative data were available but could not be interpreted easily via the usual tables and x-y plots. To figure out the story behind the data, we built a chart that shows networks to identify collaboration links between countries and to see the impact of each collaboration. See the article "Telling stories with big data" for how our team works with governments, funding bodies, universities, and researchers to provide data-based evidence to inform strategic decisions in research.

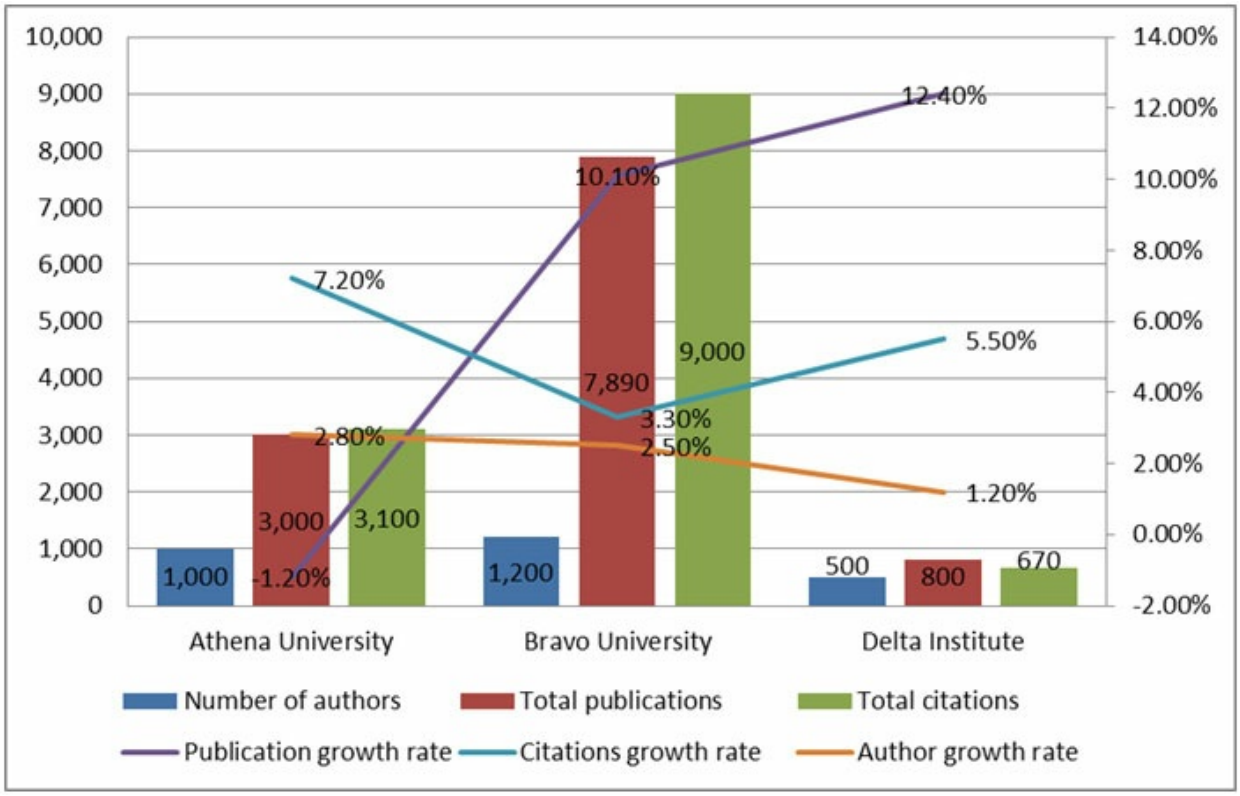
Here, we offer a brief guide consisting of 5 steps for anyone who wants to communicate an observation or explain an analysis clearly with tables, graphs, charts and diagrams, keeping in mind that creating a good visualization is an iterative process.

Step 1 — Be clear on the question

When creating a visualization, the first step is to be clear on the question to be answered – or alternatively answering the question: "How will the visualization help the reader?"

Having a clear question to answer helps avoid a common problem in data visualization: Comparing "apples" to "oranges." Consider a hypothetical dataset (see Table 1) in which we have information on an institution's total number of authors, publications, citations and their respective growth rate for a given year. A bad example of visualization is shown in Figure 2, where all the variables are included in the same chart. Plotting variables of different types in the same chart is seldom a good idea because a distracted reader may be misled to compare variables that are not comparable. For example, it does not make sense to observe that all the institutions have fewer authors than the total number of publications, or that the publication growth rate rises from Athena University, to Bravo University to Delta Institution. Busier graphs are just harder to read and process, and this is the case when you have multiple y-axes; it is not always clear which variable corresponds to which axis. Simply put, a bad visualization confuses rather than clarifies.

Name	Number of authors	Total publications	Total citations	Author growth rate	Publication growth rate	Citations growth rate
Athena University	1,000	3,000	3,100	2.8%	5.2%	7.2%
Bravo University	1,200	7,890	9,000	2.5%	10.1%	3.3%
Delta Institute	500	800	670	1.2%	12.4%	5.5%



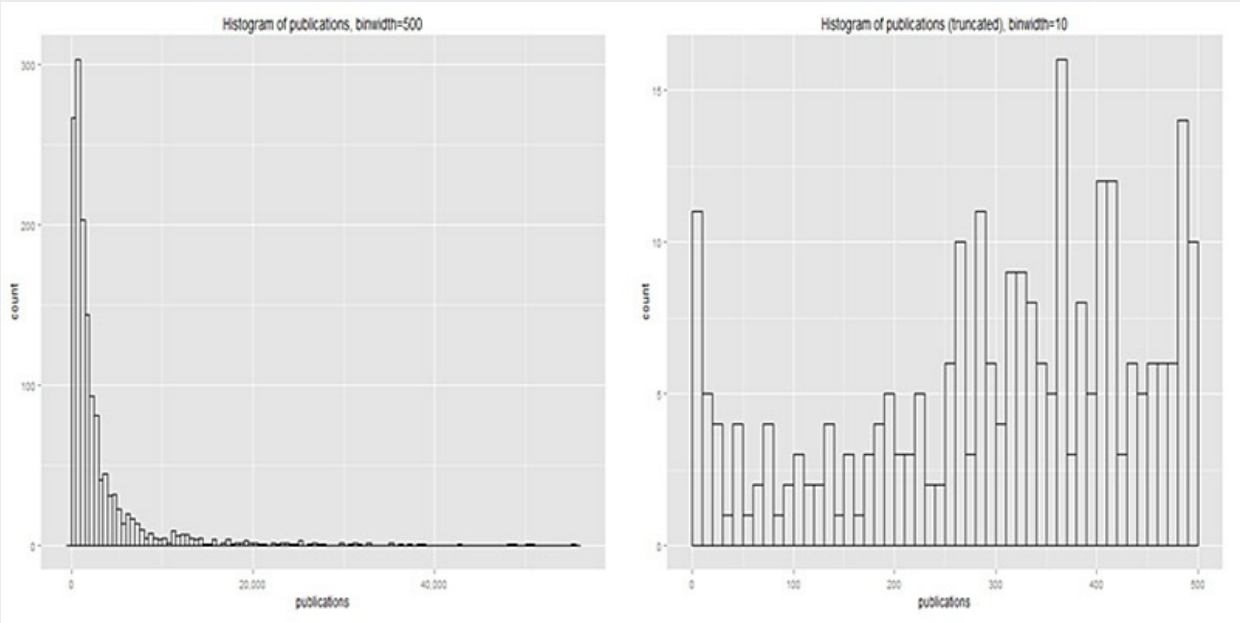
Step 2 — Know your data and start with basic visualizations

The next step after identifying the visualization's objective is building a basic diagram –this can be a bar chart, line chart, flow chart, scatterplot, surface plot, map, networks, and more – depending on what data are available. In the course of identifying the key message or messages the chart should convey, we must be clear about several things:

- What variables are we trying to plot?
- What do the x-axis and y-axis refer to?

- Does the size of data points mean anything?
- Does the color in the chart mean anything?
- Are we trying to identify trends over time or correlation between variables?

While some people use different chart types interchangeably, it is not a best practice; different charts are best used to show different types of information. Line charts, for example, are most useful for showing trends over time, or potential correlation between two variables. When there are many data points in your dataset, it may be easier to visualize the data using a scatterplot instead. Histograms, on the other hand, show the distribution of the data, and the shape of the histogram can change depending on the size of the *bin width*, as can be seen in Figure 1. (When making a histogram, you are essentially making a bar chart that shows how many data points fall into a certain range. That range is called the bin width.)



Also, as you will see, you may decide to refine or change the chart type once you complete the next step.

Step 3 — Identify messages of the visualization, and generate the most informative indicator

Consider a different hypothetical dataset of publication details of a particular institution shown in Table 2. The most important step in visualization is to know the dataset well and what each variable represents. The simple sorted table reads that for Subject A, the institution published 633 articles, which accounts for 39% of all articles in the institution; for the same period this subject represents 44% of all articles published globally where 27,738 articles were published. Note, however, in this case the percentages in column (B) add up to more than 100% as some articles are tagged in multiple subject areas.

In this example, we want to know how much the institution publishes in each subject area. While the number of publications is a useful indicator, it becomes more informative when positioned in the context of:

1. The total research output of the institution, represented in column B, and

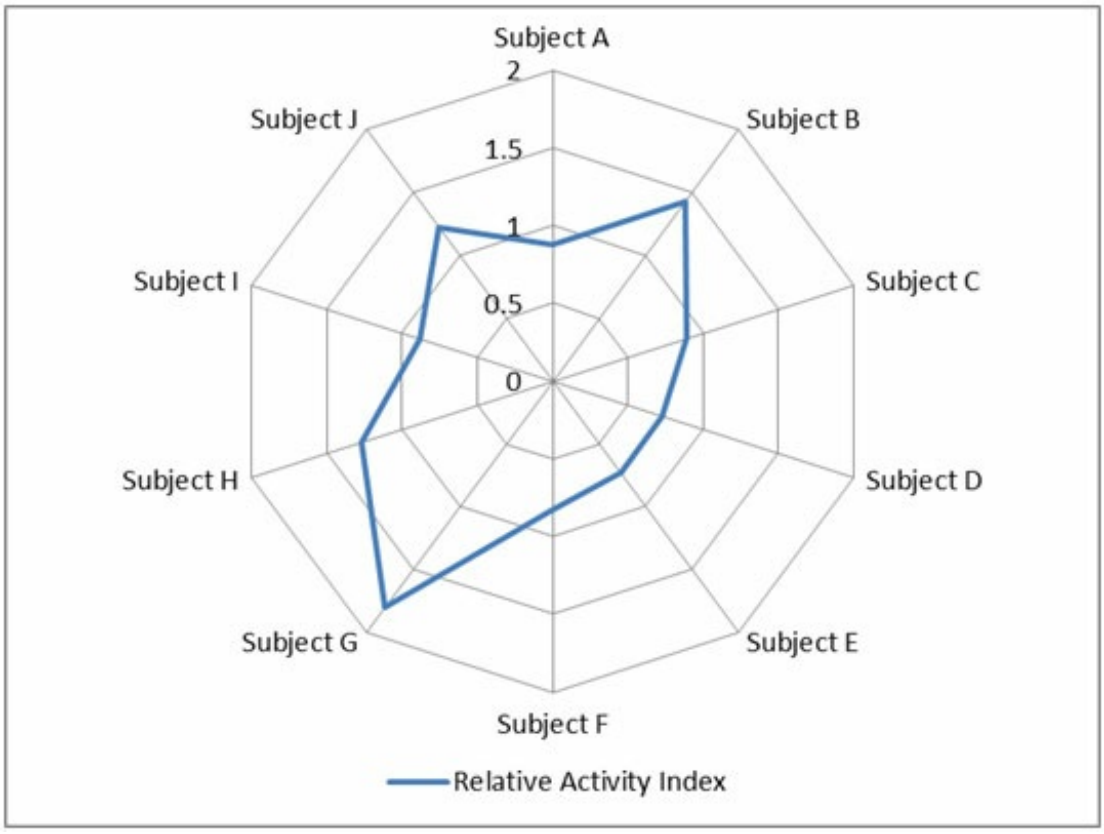
2. The global activity of that subject.

Hence, we can generate a relative activity index (RAI) where a value of 1.0 indicates that an institution's research activity in a field corresponds exactly with global activity in that field; an RAI higher than 1.0 implies a greater emphasis while lower than 1.0 represents a lesser focus as compared to the world. To generate this indicator, we divide the value in column (B) by column (D) in Table 2.

Subject	(A) Publications	(B) Publication (%)	(C) World	(D) World (%)	(E) Relative Activity Index
Subject A	633	39%	27,738	44%	0.88
Subject B	579	35%	15,718	25%	1.43
Subject C	247	15%	10,759	17%	0.89
Subject D	227	14%	12,012	19%	0.73
Subject E	149	9%	7,907	13%	0.73
Subject F	76	5%	3,563	6%	0.83
Subject G	67	4%	1,439	2%	1.8
Subject H	39	2%	1,191	2%	1.27
Subject I	38	2%	1,672	3%	0.88
Subject J	33	2%	1,051	2%	1.22

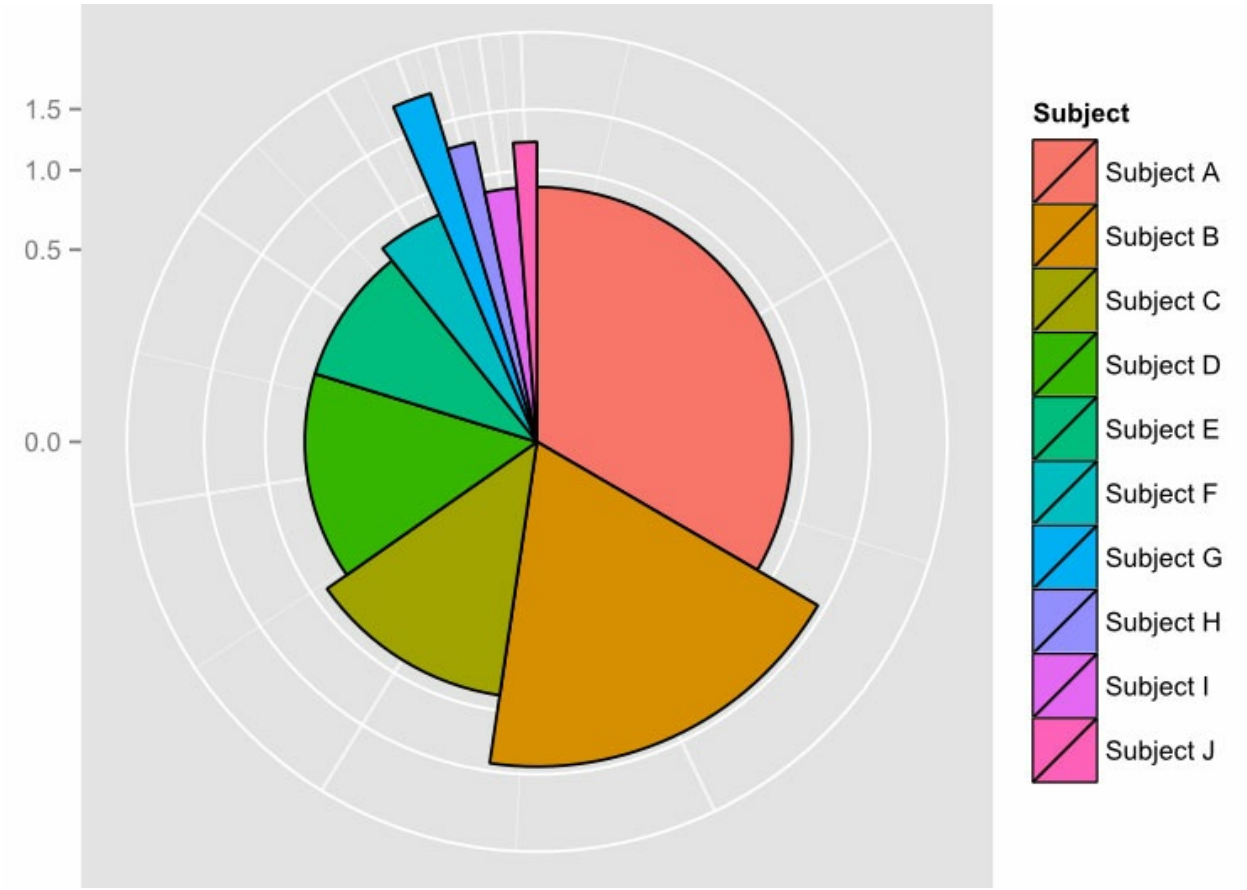
Step 4 — Choose the right chart type

Now, we can plot the relative activity index in a radar plot for comparison, with the attention focused on the subjects that have the highest (or lowest) relative activity index. For instance, subject G has the highest activity index (at 1.8); however, the total world publication in this subject is much smaller as compared to the other subjects. This cannot be seen in the radar plot (see Figure 3). Another limitation of the radar plot is that it suggests a connection between the axes when in this case there is none (subjects are not connected to each other in any way).



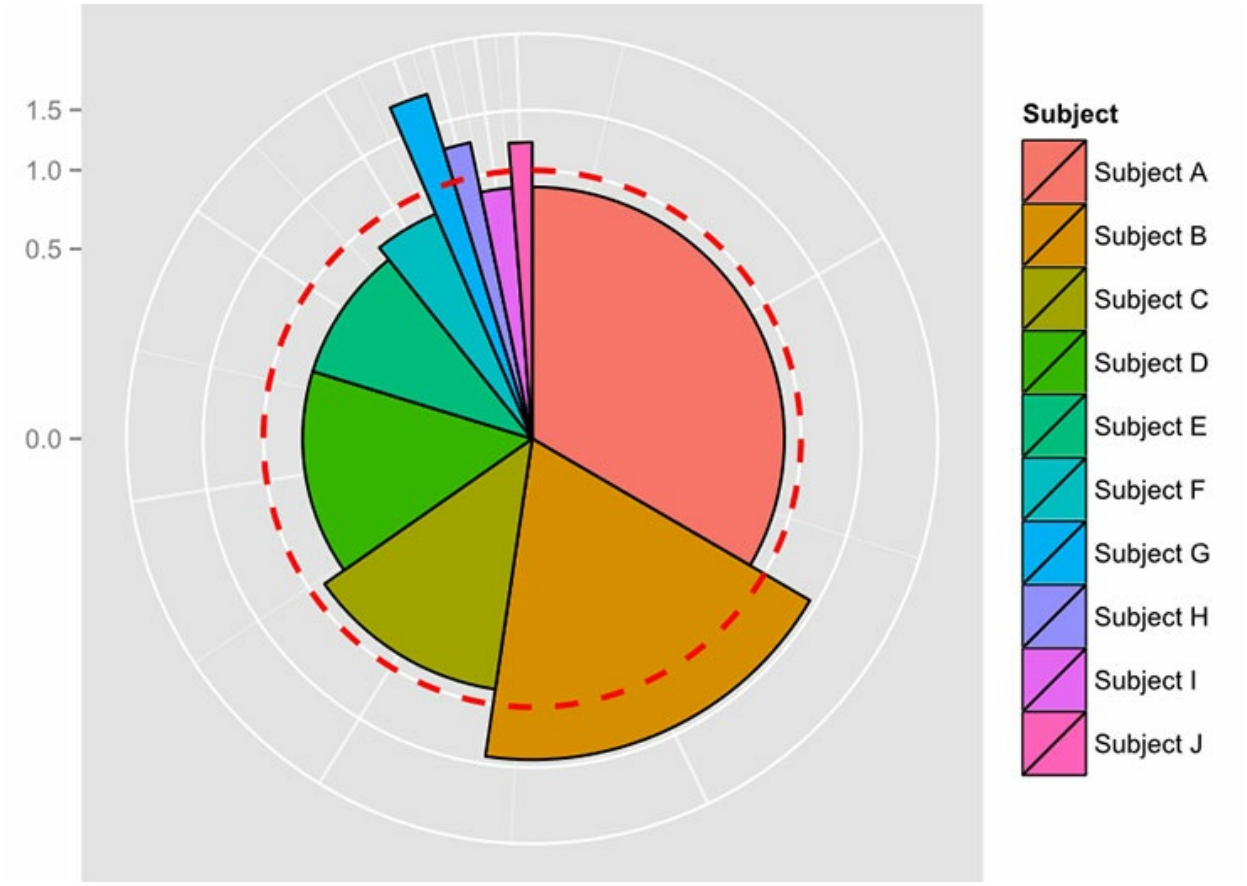
Normalization (in this case RAI) is a very common and useful approach to data interpretation, but needs to be used in context in order for its viewers to make the right conclusion. In our example, identifying high focus (Subject G) on really small subject is not very meaningful.

To get a sense of where the most activity is taking place, we can combine volume and relative activity index in a single chart. Using a spie chart as shown in Figure 4, we represent the volume of articles published by the area of pie slices and the relative activity index by the radius length. Note that in this case the radial axis is of quadratic scale (as compared typical linear scales used in Figure 3). Subject B now stands out with high volume (represented by the area of the pie slice) and high relative activity index (represented by the long radius length).

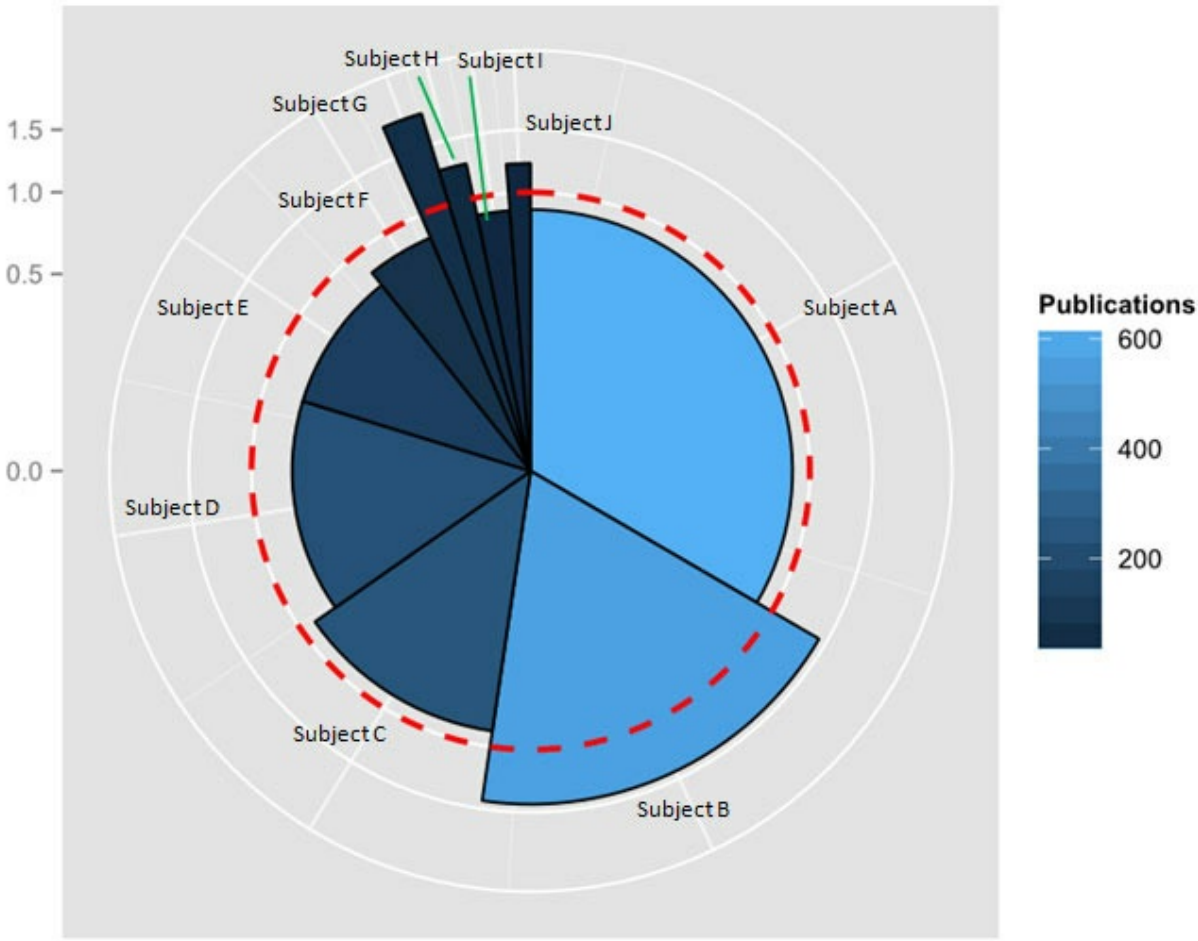


Step 5 — Use color, size, scale, shapes and labels to direct attention to the key messages

Measuring the length of the radius may not be easily done by eye, and in this example, since the relative activity index of 1.0 indicates that the institution's research activity in a field corresponds exactly with global activity in that field, we can guide our readers by including the reference value of 1.0 (represented by red dotted circle) in Figure 5. Subjects whose radius exceeds the reference line can then be identified easily.



We can further help readers identify the subjects with the most publications using the color of the pie slices to vary based on the publication volume, as shown in the legend. The corresponding subjects can then be labeled clearly for easy identification (see Figure 5).



Conclusion

There are many ways to visualize data, new tools and chart types appear constantly, and each strives to create more attractive and informative charts than before. We suggest focusing on the principle that a visualization should clarify and summarize the key message rather than confusing and overloading the reader with superfluous information.

Elsevier Connect Contributors

Georgin Lau, Content & Analytics Product Manager for Elsevier APAC, is based in Singapore and focuses on the APAC Analytics market. She holds a master's degree in statistics from the National University of Singapore and is also a visualization expert who seeks to present data in the most understandable way. One of her current projects is the research assessment study on the discipline of brain research. Prior to joining Elsevier, she worked on many Singapore and overseas government projects as a consultant.

Dr. Lei Pan, Content & Analytics Product Manager for Elsevier EMEA, is based in Amsterdam. She holds a master in Economics from Erasmus University Rotterdam and a PhD in Economics from the VU University Amsterdam. In her current role, she serves clients in government, universities and funding bodies, and specializes in analyses using bibliometric and economic data.

Other contributors were Jeroen Baas, Head of Data Science, and Dr. Judith Kamalski, Head of Analytical Services, in Elsevier's Research Management.