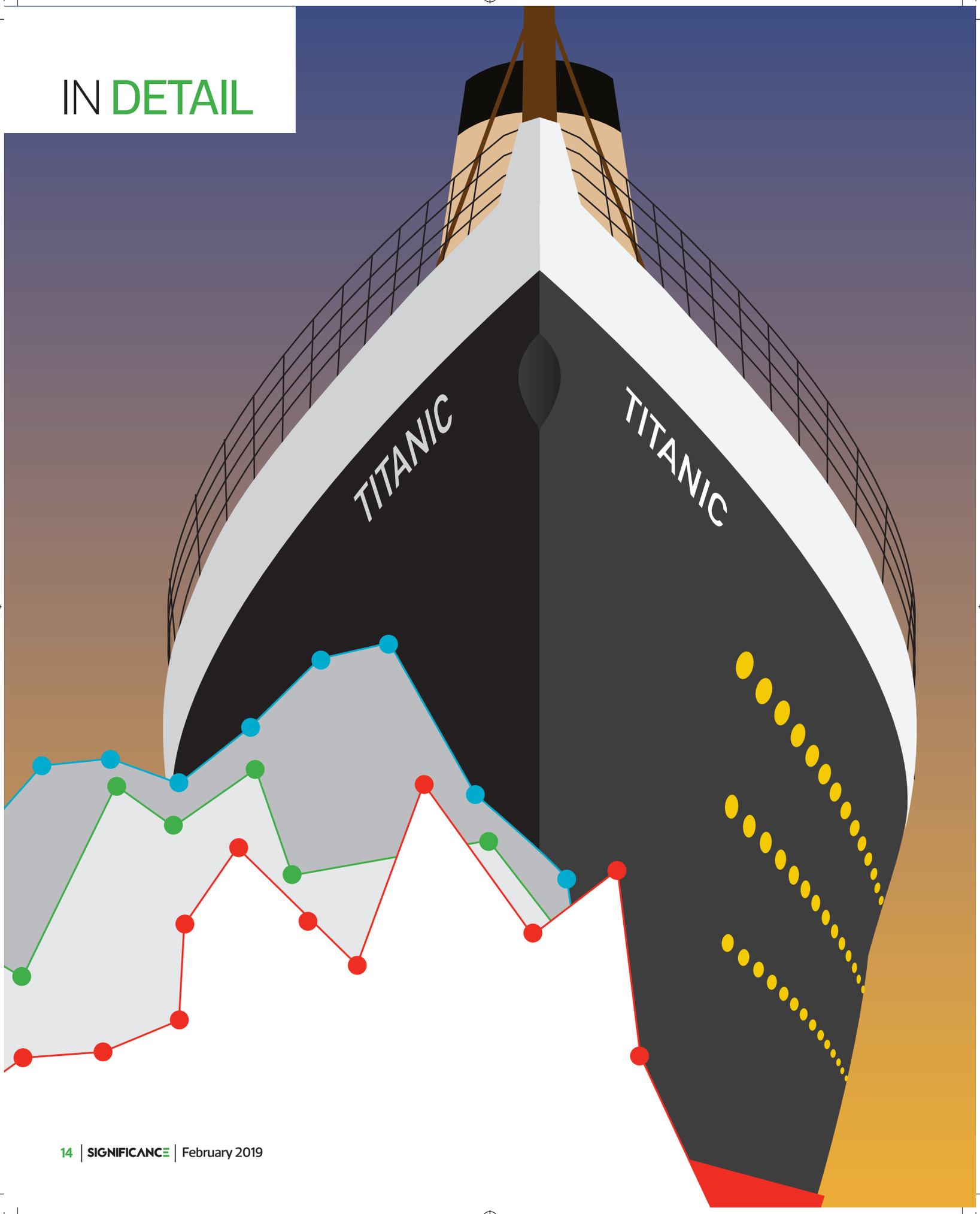
IN DETAIL



Visualising the *Titanic* disaster

The sinking of the Titanic has inspired books, movies and documentaries. But it has also motivated data visualisation designers to find new ways to tell the story of the tragedy. **Michael Friendly**, **Jürgen Symanzik** and **Ortac Onder** review the first graph of the disaster and some recent history

Michael Friendly is a fellow of the American Statistical Association, professor of psychology at York University, and an associate editor of the *Journal of Graphical and Computational Statistics*. His current research work includes the development of graphical methods for data visualisation, and the history of data visualisation.

Jürgen Symanzik, PhD, is a fellow of the American Statistical Association, an elected member of the International Statistical Institute, and professor of statistics at Utah State University. He is interested in everything related to statistical graphics, including software, applications, and their history.

Ortac Onder is a PhD candidate at Schulich School of Business, York University. His current research focuses on quality and performance measurements and improvement in health care.

The sinking of the RMS *Titanic* is one of the most storied shipwrecks in maritime history. Touted as the ultimate in transatlantic travel and said to be "unsinkable", the *Titanic* collided with an iceberg on 14 April 1912 on her maiden voyage and sank shortly thereafter on 15 April, killing 1502 out of 2224 passengers and crew. The sinking of the *Titanic* is not the largest in terms of lives lost. But it is the one that has been documented most thoroughly – in government reports and personal accounts of survivors, and in numerous books and several popular movies.

This is one legacy of the *Titanic* disaster, but it left another: a wealth of data, comprising details of all the passengers and crew, many with names, ages, passenger class and even cabin numbers for those in first and second class.

We recently discovered an early and relatively unknown graph showing survival among the *Titanic* passengers and crew, published less than one month after the disaster. This graph had a surprisingly modern look. It prompted us to review the history of this graph and the variety of uses to which the *Titanic* data have been put in the two decades since the data set became available in machine-readable form.

The first graph

The Sphere was a popular British illustrated weekly newspaper, published by the Illustrated London News Group from January 1900 until June 1964, and dedicated to worldwide reporting on popular issues. On 4 May 1912, only three weeks after the *Titanic* disaster, it published a chart (Figure 1) by the graphic artist G. Bron using data released the week before by the House of Commons.

Bron's graph shows the breakdown of survival among the passengers and crew – by passenger class, gender and age (comparing adults and children) – in what is clearly an early innovation in data display. It combines back-to-back bar charts for those who lived and those who perished, using area of the bars to convey the actual numbers. Within the passenger classes, the numbers and bars are subdivided by gender for adults, while children are shown as a separate group. It also includes two similar summary panels, showing the totals for all passengers and for all passengers and crew.

Today, we might describe this as an early form of a mosaic plot, or as an area-proportional back-to-back array of bar charts. Whatever name we give, it deserves to be admired as an exceptional early example of data visualisation and a tribute to the skills of the illustrator.

Who was G. Bron?

G. Bron was a prolific technical illustrator who worked for *The Sphere*, the *Illustrated London News* and similar publications between about 1910 and 1925. Today, he would be called a data-graphic or info-vis designer, one far ahead of his time. Little about him was previously known, not even his first name. A search in the British Newspaper Archive (bit.ly/2Rzv5dm) turned up over 20 examples of his work, most published in *The Sphere*. In the course of writing this article

Bron's use of back-to-back proportional bar charts to show death versus survival was a stroke of graphic genius

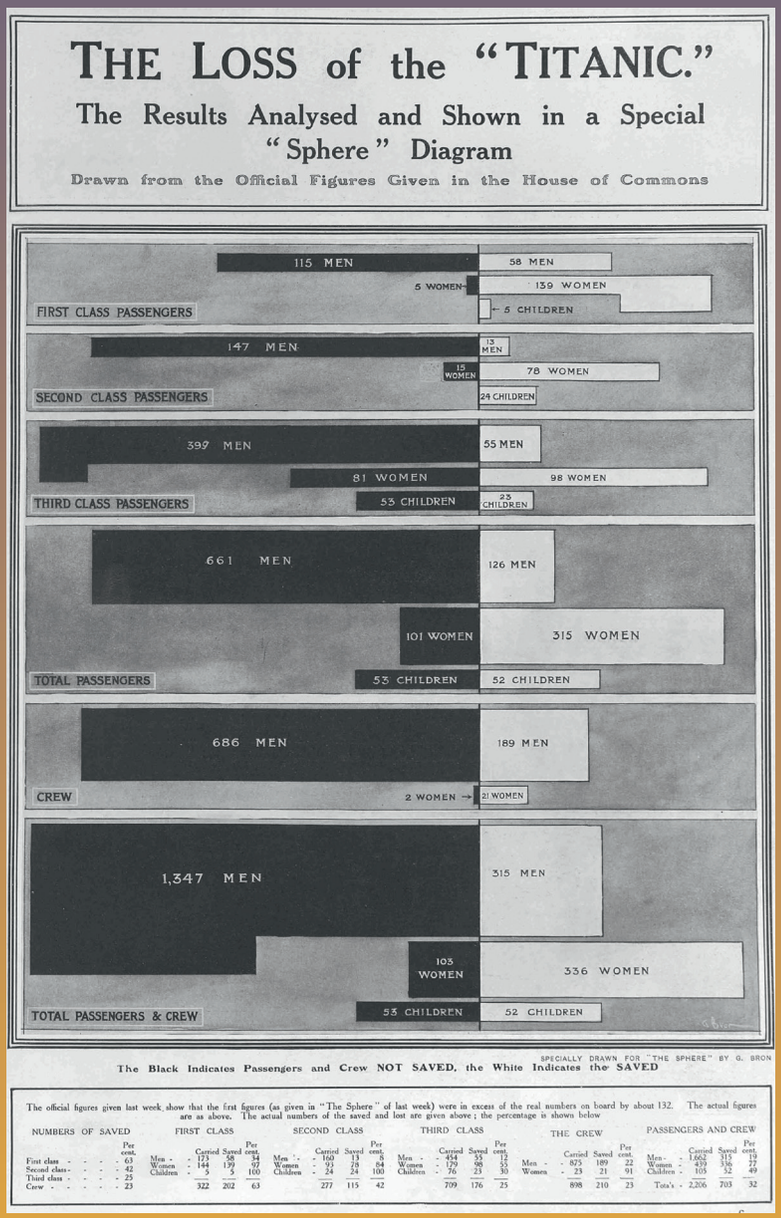


FIGURE 1 G.Bron's chart of "The Loss of the 'Titanic'", from *The Sphere*, 4 May 1912. Each subgroup is shown by a bar whose area is proportional to the numbers of cases. © British Library Board; reproduced with permission of the Mary Evans Library.

we discovered that G. Bron was most likely the pseudonym adopted by William B. Treeby, born in London, but further biographical details are still sketchy (see supplementary material at bit.ly/titanicvis).

By and large, Bron's illustrations were graphic stories, designed to convey an interesting but possibly complex topic visually, in ways in which mere words and numbers could not compete. It is difficult to know what led him to produce his remarkable chart of the *Titanic*. Sometime between the sinking on 15 April and the publication of his *Sphere* graph on 4 May, he became aware of a numerical table classifying passengers and crew that would shortly be published by the House of Commons. We can imagine that he looked at this and asked himself how he could make it comprehensible to his readers. His use of back-to-back proportional bar charts was novel and a stroke of graphic genius. Just a glance showed that, overall, two-thirds of the passengers and crew perished. The separate conditional panels for class showed directly to the eye that survival was greatest in first class and least among the crew. The reader could "drill down" to examine the breakdown by gender and age within each class.

The data

The primary sources of data on the *Titanic* derive from official inquiries launched in Britain and the USA. (Complete documents can be found at titanicinquiry.org.) Shortly after the disaster, the British Parliament authorised the British Board of Trade Inquiry with Lord Mersey as chair. The committee interviewed over 100 witnesses over 36 days of hearings. Their report, issued on 30 July 1912, contained extensive tables of passengers and crew, broken down by age group, gender, class and survival, as well as details on the launching of the 20 lifeboats. In April–May 1912 a similar inquiry was initiated in the US Senate which interviewed 82 witnesses over 18 days. Among other things the report (over 1000 pages) contained lists of the names and addresses of most passengers and crew.

As far as we are aware, the first public data set appeared in 1995 in an article by Robert Dawson, titled "The 'Unusual Episode' Data Revisited", in the *Journal of Statistics Education*.[1] Its classroom use was illustrated by an exercise in statistical thinking, where students were shown tables of deaths and death rates – classified by economic status, age and gender – for an "unusual episode" (without context) and asked to reason about what the causes might have been. The data set contains 2201 observations and the variables Class, Age, Sex and Survived.

In September 1995, Phillip Hind launched encyclopedia-titanica.org, the first publicly available database on all passengers and crew aboard the ship. At the time, it was the only reasonably complete individual list giving details of name, actual age, profession, cabin number, lifeboat number, and so forth. Two surviving canine pets (one named Sun Yat Sen) were also listed.[2] The website now includes photos and biographies on many of the passengers and crew.

Popular interest in the *Titanic* surged with the release of James Cameron's movie in November 1997. Immediately following this, Random House released a boxed set,

Titanic: The Official Story, containing the Mersey report and facsimiles of 18 original documents from London's Public Record Office.[3] These included the *Titanic* deck plans, the final telegram sent from the ship just prior to sinking, newspaper articles excoriating the White Star Line for criminal negligence, lists of deaths recorded in official logs, and so on. It is not explicitly clear what the goal or purpose was, but these materials serve as a model case for courses in history of statistics and archival research.

The *Titanic* data, taken from Dawson, made their first public appearance in a software package in R, version 0.90.1, in December 1999, expressed as a four-way contingency table of counts, classified by Class, Age, Sex and Survived. A variety of other data sets are available in contributed R packages, including `TitanicSurvival` (in the car package), which gives details (name, sex, age, class, survived) on 1309 passengers, and `Lifeboats` (in the vcd package), which gives data on the composition and launch times of the lifeboats. Passenger data from the *Titanic*, split into training and test samples, is also used in a Kaggle prediction competition (bit.ly/2RxcwGU).

Some modern uses

The significance of the disaster and the availability of information regarding the passengers and crew made the *Titanic* data attractive for various uses. The range of disciplines gives a sense of the appeal of these data as a compelling example of popular interest, of a novel graphical method or illustration of some statistical technique. The context makes it easy to tell an interesting story to illustrate a new method or graph.

In statistics, narrowly defined, the data have been used to illustrate graphical methods for categorical data and their use as a visualisation method for log-linear models and related generalised linear models. Recursive partitioning methods (also known as classification and regression trees, or CART models) is another area where the *Titanic* data provide an easily understood concrete example, and this has led to tree-based visualisations.

In a wider scope, encompassing computer science and data science, with an emphasis on predictive modelling and cross-validation, the *Titanic* data provide an important test case; while in the InfoVis community, the data provide a challenge – and opportunity – for graphic designers to try to tell the story of the disaster in a single sheet, containing words, numbers, pictures and data visualisations.

What follows is a selection of a few highlights to illustrate the themes described above. Many more examples can be found on the web page associated with this article (see box and bit.ly/titanicvis).

Mosaic-type plots

Bron's initial graphic idea, to show deaths and survival on the *Titanic*, broken down by passenger class, gender and age group, was brilliant at the time, and perhaps underappreciated in the history of data visualisation. He was on to something important: how to display the proportions of survivors, classified by the other variables he had available. His solution anticipated modern methods.

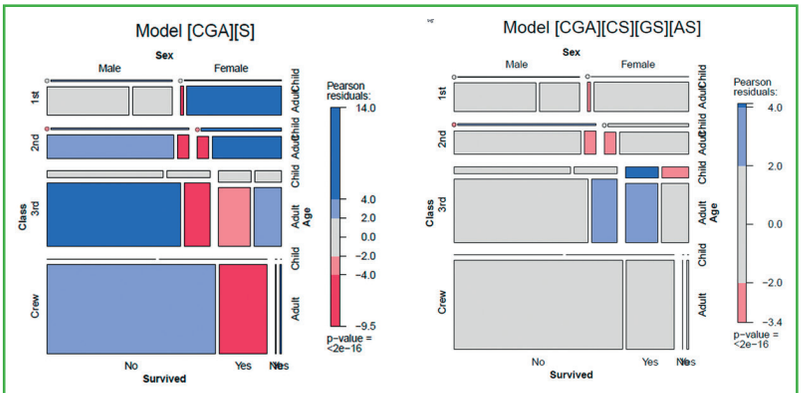


FIGURE 2 Mosaic plots for two log-linear models fitted to the *Titanic* data. Left: joint independence model `[CGA][S]`; right: main effects model `[CGA][CS][GS][AS]`. Source: author graphic, recreated from Friendly (1999).[6]

In the 1990s two new ideas for graphical analysis of categorical data arose for this problem. The *Titanic* data provided great examples because they gave a context and a story to appreciate these new methods.

First, mosaic plots, proposed by Hartigan and Kleiner, provided a new graphic method for visualising multivariate frequency tables, in a single view rather than a collection of one-way or two-way diagrams.[4] The essential idea was a mosaic of rectangles ("tiles"), with the area of each made proportional to the cell frequency. Friendly connected these with log-linear models by shading the cells in relation to residuals in a given model.[5,6]

For example, Figure 2 shows the result of fitting two models to the *Titanic* data. The left-hand panel shows the fit of the model with the symbolic formula `[CGA][S]`, which asserts that Survival (`[S]`) is independent of Class, Gender and Age jointly. This is the baseline, null model. The pattern of shading (blue for positive residuals, red for negative) shows that important associations remain unaccounted for: that is, one or more of Class, Gender and Age affects Survival. The right-hand panel shows the fit of the model `[CGA][CS][GS][AS]`, which allows "main effect" associations of each of Class, Gender and Age with Survival. This model fits much better, but still shows significant lack of fit. The pattern of residuals here suggests some interactions are present: adding the term `[GAS]` would allow an interaction of Gender and Age – "women and children first"; adding the association `[CGS]` would allow Survival to depend on the combinations of Class and Gender.

Second, interactive software for visualising and manipulating multivariate contingency tables was developed. MANET[7] and MONDRIAN[8] from the Augsburg lab are notable here. Hofmann illustrated how selecting a category of one variable (Survived) highlighted those cases in all other views.[9] Valero-Mora *et al.* used the *Titanic* data to illustrate *ViSta*, a

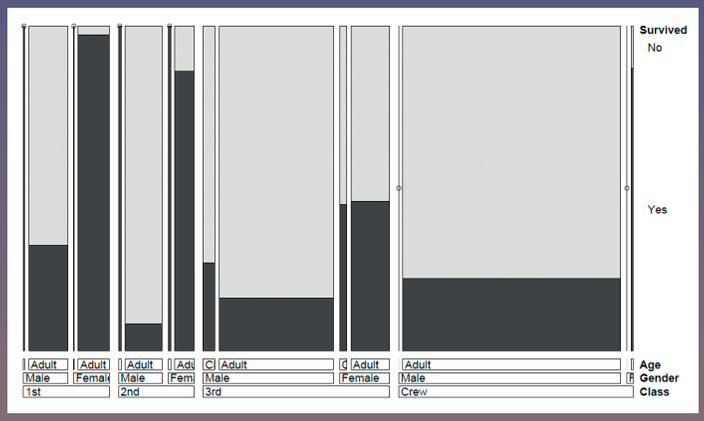


FIGURE 3 Double-decker plot of the *Titanic* data. Each bar has an area proportional to the frequency in the table. The proportion that survived is shaded black. Source: Author graphic, recreated from Meyer et al.[12]

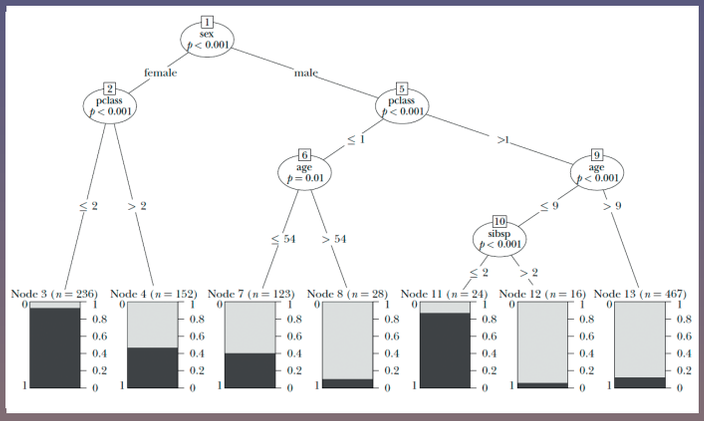


FIGURE 4 Graphic representation of a conditional inference tree, predicting survival from sex, passenger class, age and family size.[14] © American Economic Association; reproduced with permission of the *Journal of Economic Perspectives*.

system combining multiple interacting windows, using both log-linear models and multiple correspondence analysis.[10]

Figure 3 shows a double-decker plot, a variation of a mosaic plot in which the tiles for all predictors (Class, Gender and Age) are split horizontally and the response variable (Survived) is split vertically.[11] In this type of plot, the widths of the bars are proportional to the joint frequencies of C, G, and A. When each bar is split vertically by Survived, the heights of the black bars are proportional to the conditional probabilities of S given C, G, and A. If Survival were independent of all predictors (the model `[CGA][S]` shown in Figure 2 (left)), the black bars would all have the same height. Note that showing the bars for survivors and those who perished back-to-back would give something similar to Bron's chart.

Tree diagrams

Cross-classified data can also be displayed as tree diagrams of various types, with branches corresponding to splits of the categories for variables in some order. Tree-maps are a simple example, similar to mosaic plots in that they also display a measure of size by areas of rectangles.[13]

Other graphic methods

The *Titanic* data served to illustrate, or even motivate, a wide variety of graphical and analytic methods. A few are mentioned in this article, but more can be found online at bit.ly/titanicvis, including:

- **Venn diagrams,** used to show the overlapping sets of the categories.
- **Trilinear plots,** used to display the composition of the survivors in the lifeboats (men, women and children, and crew).
- **Nomograms,** interactive graphics used to show the predicted probability of survival for various settings of the predictors.
- **Parallel sets,** an extension of parallel coordinate plots for categorical variables.
- **Dot plots and nonparametric smoothed curves,** showing survival probability.

A more powerful use arises in connection with classification trees as models for an outcome variable such as survival. For a binary response, these are similar to a series of logistic regression models, where predictors are chosen to maximise predictive accuracy at each step. Pruning methods and cross-validation are used to control model complexity and minimise out-of-sample classification error. Varian was among the first to use the *Titanic* data for this purpose.[14]

Figure 4 gives the result of fitting a conditional inference tree ("ctree") predicting survival from sex, class, age and a measure of family size (`sibsp` = number of siblings plus spouse aboard). The first node splits the data by sex. The second divides by class. The third node (in the right branch) splits males by age, and those aged 9 and under are further split by `sibsp`. The bars at the bottom show the survival rate in each terminal node. As opposed to log-linear models and generalised linear models, classification trees are somewhat more intuitive when shown visually, and have the additional advantage that what might be complex interaction terms in linear models can be easily fitted by successive splits on the branches to improve prediction.

Information visualisation

Following in the footsteps of Bron, modern graphic designers continue to be inspired by the tragedy of the *Titanic* and challenge themselves to tell the story of the disaster in ways that are both visually appealing and provide sufficient details. Unlike statistical graphs which usually focus on just one aspect, an information graphic often attempts to tell the entire story all on one sheet, as in a poster presentation.

The best example we have found of this genre is the graphic produced by Andrew Barr and Richard Johnson for the *National Post* (Figure 5, and bit.ly/2RyjdbL). This illustration strikes us as a tour de force of visual story telling: numbers, words and

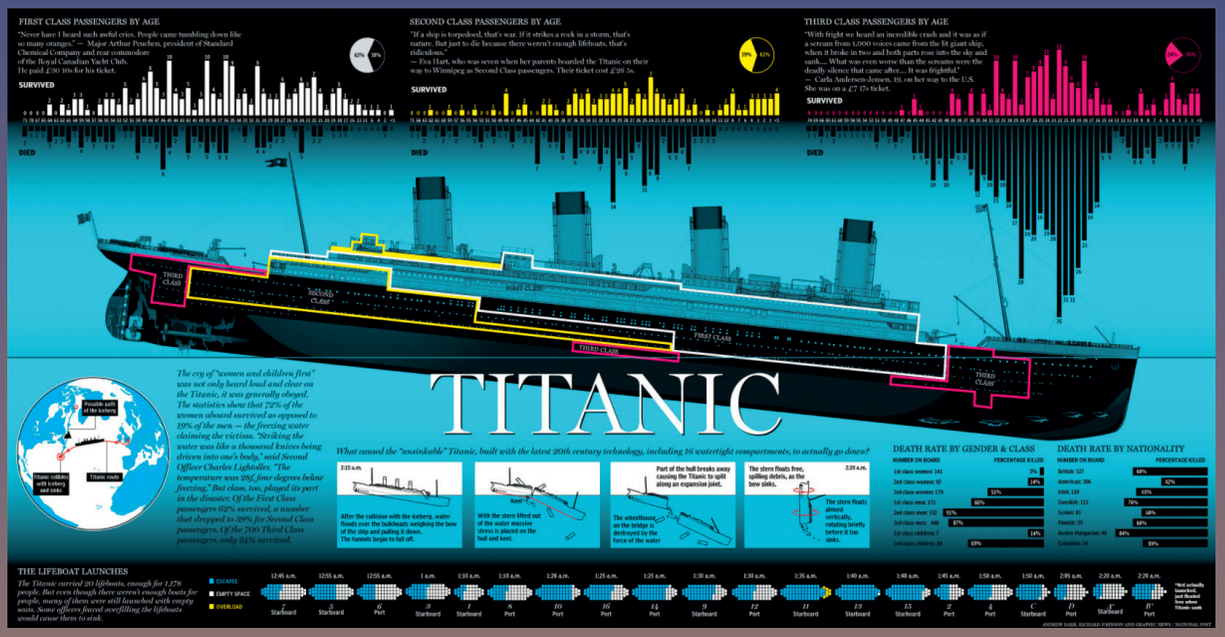


FIGURE 5 Infographic by Barr and Johnson (bit.ly/2RyjdbL) telling a graphic story of survival on the *Titanic*. Notable is the integration of rich numerical information shown in graphs with images providing context and visual explanation. Material republished with the express permission of: National Post, a division of Postmedia Network Inc.

pictures (both images and graphs) are woven seamlessly into a narrative.

The top portion contains back-to-back bar plots of the passengers by age and class, showing the age distributions of those who survived and those who died, with pie charts summarising survival by class. It uses colours keyed to the locations of cabins for the classes in the dominant graphic of the ship.

The bottom portion shows the loading of the lifeboats in the order they were launched, shaded to show the proportion of seats that were filled. It is clear that those launched early and those launched just before the ship sank were only partially filled. Other charts at the lower right give the death rates by gender and class and by nationality of the passenger. A text box gives an interpretation of survival, including the ideas of "women and children first" and the declining survival according to class.

Past and future

The sinking of the *Titanic* was surely a tragedy, but, unlike other historical events resulting in great loss of life, it left behind detailed information on the individuals involved – both victims and survivors – whose stories attracted wide interest. Bron's 1912 chart should be appreciated as an attempt at visual explanation far ahead of its time: the idea that survival could be understood through graphic displays.

We started this project with the discovery of Bron's chart, and the thought that it would be useful to collect and catalogue the various ways in which the *Titanic* data had been depicted in graphs over the past century. We were pleasantly surprised by the wide range of graphical methods and other applications we found. This attests to the compelling nature of the *Titanic* disaster and to the desires of modern graphical developers and designers to illustrate their methods and skills by continuing to tell the *Titanic* story. We believe that the *Titanic* data still have much to offer to graphic designers and visual story tellers. ■

References

1. Dawson, R. J. M. G. (1995) The "unusual episode" data revisited. *Journal of Statistics Education*, **3**(3).

2. Friendly, M. (2000) *Visualizing Categorical Data*. Cary, NC: SAS Institute.

3. Anonymous (1997) *Titanic: The Official Story: April 14-15, 1912*. New York: Random House.

4. Hartigan, J. A. and Kleiner, B. (1981) Mosaics for contingency tables. In W. F. Eddy (ed.), *Computer Science and Statistics: Proceedings of the 13th Symposium on the Interface* (pp. 268–273). New York: Springer-Verlag.

5. Friendly, M. (1994) Mosaic displays for multi-way contingency tables. *Journal of the American Statistical Association*, **89**, 190–200.

6. Friendly, M. (1999) Extending mosaic displays: Marginal, conditional, and partial views of categorical data. *Journal of Computational and Graphical Statistics*, **8**, 373–395.

7. Hofmann, H., Unwin, A. and Theus, M. (1997) MANET (software application). http://www.rosuda.org/MANET/

8. Theus, M. (2002) Interactive data visualization using Mondrian. *Journal of Statistical Software*, **7**(2). http://www.jstatsoft.org/v07/i11/

9. Hofmann, H. (1998) Simpson on board the *Titanic*? Interactive methods for dealing with multivariate categorical data. *Statistical Computing & Statistical Graphics Newsletter*, **9**, 16–19.

10. Valero-Mora, P. M., Young, F. W. and Friendly, M. (2003) Visualizing categorical data in *ViSta*. *Computational Statistics & Data Analysis*, **43**, 495–508.

11. Hofmann, H. (2001). Generalized odds ratios for visual modeling. *Journal of Computational and Graphical Statistics*, **10**, 628–640.

12. Meyer, D., Zeileis, A. and Hornik, K. (2006) The strucplot framework: Visualizing multi-way contingency tables with vcd. *Journal of Statistical Software*, **17**, 1–48

13. Shneiderman, B. (1992) Tree visualization with tree-maps: A 2-D space-filling approach. *ACM Transactions on Graphics*, **11**(1), 92–99.

14. Varian, H. R. (2014) Big data: New tricks for econometrics. *Journal of Economic Perspectives*, **28**(2), 3–28.