Stat Comput (2015) 25:705–711
DOI 10.1007/s11222-015-9565-6

Statistics and computing: the genesis of data science

David J. Hand[1]

Accepted: 8 March 2015 / Published online: 11 June 2015

Abstract The two disciplines of statistics and computing are together the core technologies of data science. The journal *Statistics and Computing* has been instrumental in enhancing the interaction between them over the past quarter century. This has been a period of dramatic change in each of the disciplines, where huge progress has been made, in both fundamental theory and in practice and applications. But it has also been a period of dramatic change in scientific publishing. The evolution of *Statistics and Computing* has reflected both changes, putting it at the cutting edge of progress. But these changes have not reached an end. We can confidently expect even more startling progress in the disciplines and change in the practice of scientific publishing in future years. It is vital that *Statistics and Computing* keeps pace.

Keywords Statistical software · Big data · Data science · Scientific publishing · Open access · Peer review

1 In the beginning

During the second half of the twentieth century, the nature of statistical practice changed. The advent of the computer lifted the burden of arithmetic tedium, gradually transforming statistics from a discipline characterised by the minutiae of calculation to a discipline where one could take the big picture, focusing on understanding and interpretation. Moreover, by 1991, the year in which *Statistics and Computing* was launched, it was very apparent that statistics was just as much a computational discipline as a mathematical one. The journal was launched in recognition of that shift and as an attempt to capitalise and reinforce the synergy which was emerging between statistics and allied domains within computer science, such as machine learning.

The period around 1991 was also the beginning of another revolution, this time in statistical (and, more generally, scientific) publishing, as scientific communication shifted from paper to electronic media. *Statistics and Computing* has spanned this change, so I begin in the next section by describing it. Subsequent sections return to look at the impact of computers on statistics, and where things may be heading.

2 The editorial process

Although only 25 years have passed since the first edition of *Statistics and Computing* appeared, it was launched in the era of hard-copy paper submissions. I am aware that the following is going to sound bizarre to anyone who was awarded their PhD since the turn of the century, but here is a description of the editorial process when the journal began.

Authors would send in three or four paper copies of their submissions. I, as editor, would identify an Editorial Board member to deal with the submission, or I would deal with it myself. We would then post the paper copies to potential referees, along with paper forms for their comments (separate forms to return comments for the authors and confidential comments for the editor). Then, as now, a goodly percentage of potential referees would not reply, or not reply by the deadline, so chasing letters would be sent out. Eventually, the reports would be received, and I would have the task of reaching a decision on the basis of them. As any editor will know, this is often quite hard, as referees' reports

David J. Hand is the Founding Editor of Statistics and Computing.

✉ David J. Hand
d.j.hand@imperial.ac.uk

[1] Statistics and Computing, Imperial College, London, UK

can be diametrically opposed. How does one balance 'a highly original piece of work' against 'derivative and obvious'?

It is a mark of the change that I do not still have the letters that were written back and forth when we were setting up the journal. They were written on paper—and disappeared during the course of various office moves. Had they been email communications, then they may well still exist in an electronic archive.

Physical carriage of papers and reports across the world meant that it was a slow process (although I suspect that this has pros and cons: the 2-week turnaround time for letters to the US meant that one at least had time to think). I would like to pay tribute to my former secretary, the late Liz Ostrowski, who worked for me when we set up the journal and while I was Head of the Statistics Department at the Open University. Without her superb and efficient organisation things would rapidly have disintegrated into chaos.

That description of the editorial process when we set up the journal deliberately focussed on the differences from the present. But it probably obscures something important, and which is going to have an even more radical impact on scientific communication in the future. This is that the striking thing is how *similar* the process described above is to what happens nowadays, even if it was slower and clumsier. Nowadays, the so-called snailmail has been replaced by email, papers are submitted and dealt with electronically, but essentially the process has remained the same: papers are still sent out to referees, who return comments and on which the editor makes a decision. As I explore below, this basic model is changing, and we can expect it to change dramatically, over the next 25 years. It probably does not need stressing how important it is that a journal such as *Statistics and Computing* keeps abreast of those changes.

New journals are always springing up. This is not surprising: the scientific enterprise continues to grow and the research frontier to lengthen as more specialised areas develop. In 2009, there were over 25,000 journals in science, technology, and medicine (CasesBlog 2014), with over 1.5 million articles being published per year. New journals in statistics and related data analytic areas regularly appear. At about the same time as *Statistics and Computing* was launched, we also saw other specialised journals appear, such as *Statistics in Medicine* and *Statistical Methods in Medical Research*. And, of course, other journals covering similar domains to *Statistics and Computing*, although perhaps with slightly different emphasis, appeared around the same sort of time. So, for example, *Computational Statistics and Data Analysis* was launched in 1983, *Machine Learning* in 1986, *Computational Statistics Quarterly* in 1986 (changing its name to *Computational Statistics* in 1992), and *The Journal of Machine Learning Research* in 2000.

One of the advantages of launching and editing one's own journal (as opposed to, for example, the journal of a learned society) is that one has considerable freedom to do as one wishes. One can choose the format of the journal, reformat it if you think that would help, introduce novel styles of content, and so on, without being concerned about the weight of history on one's shoulders and the disapproving expressions of oil-painted founders peering down from the walls ('that's not the way the XXX Society does things!').

3 Scientific journals

Since *Statistics and Computing* was founded, the importance of hard-copy printed journals themselves has faded—not the *content*, which is just as important as ever, but the medium. Nowadays, instead of walking through shelves of bound copies of journals, one simply downloads the specific papers one wants—from *anywhere* in the world. Physical storage problems have evaporated. This change has been accompanied by a spate of people offering paper copies of journals for sale (or free, if someone would take them off their hands). My guess is that most ended up being recycled or dumped.

Hard-copy paper publication meant that there was a barrier to entry as far as the creation of new journals was concerned: one had to organise the physical publishing and a distribution process. To a large extent, these barriers have fallen in the electronic era, so that it is easy to set up a new journal, through a website. This has pros and cons, as I shall illustrate below.

Despite these changes, the fact is that most journal websites still show the mark of the pre-electronic era. Papers still appear bunched into Volumes and Issues. There is a sound reason for this if the papers are to appear in hard copy, but no reason if they are merely to appear on a website: they can be electronically published as soon as they are ready and numbered sequentially. Clearly, date of publication remains important—not least to establish priority—and hence the creation of preprint websites like arXiv.

Another change, perhaps not directly related to the advent of the computer, is having a dramatic effect on the journal model. This is the general drive for increased transparency. This manifests itself in movements such as open government and open data. As far as scientific publishing goes, it manifests itself in terms of the so-called *open access*.

As most readers will know, open access refers to the idea that, instead of restricting access to journals to the subscribers (which may be entire organisations, like universities), access should be free to anyone. In the era of web-based scientific publication, this is in principle straightforward: put the papers on the web and let anyone access them. In practice, however, things are not that simple, not least because there is a natural tendency to hang on to the traditional publication model. One

reason for it to be a *natural* tendency is that the traditional model is a *business* model: publishers, learned societies, and other organisations earn substantial revenue from the subscription model of journal publication. If this is going to be replaced, those revenue streams will dry up.

The current most popular solution is to require the authors to pay, up-front, to have their papers published, after which anyone can access them. Some of the dangers of this are obvious. Unscrupulous individuals can set up their own journal and charge authors to publish in it, regardless of scientific merit. A nice illustration of this appears on the *That's Mathematics!* website of 14th September, 2012 (Eldredge 2012):

> On August 3, 2012, a certain Professor Marcie Rathke of the University of Southern North Dakota at Hoople submitted a very interesting article to *Advances in Pure Mathematics*, one of the many fine journals put out by Scientific Research Publishing.... This mathematical tour de force was entitled "*Independent, Negative, Canonically Turing Arrows of Equations and Problems in Applied Formal PDE*", and I quote here its intriguing abstract:
>
> > Let $\rho = A$. Is it possible to extend isomorphisms? We show that D' is stochastically orthogonal and trivially affine. In [10], the main result was the construction of p-Cardano, compactly Erdős, Weyl functions. This could shed important light on a conjecture of Conway–d'Alembert.
>
> After a remarkable turnaround time of only 10 days, on August 13, 2012, the editors were pleased to inform Professor Rathke that her submission had been accepted for publication.

As you will have suspected, *Advances in Pure Mathematics* is an open-access journal: it charges authors a fee of US$500 prior to publication.

The obvious question, when an author is confronted with a request to pay for their paper to appear on a website, is what exactly is one paying for (it does have a striking similarity to vanity publishing)? After all, since authors have to prepare their paper in specified formats, there is not a great deal of work for the 'publisher' to do.

Good-quality journals have their refereeing process, of which more below, but since referees are usually unpaid, that can hardly justify the cost. This seems to leave only the 'brand name': one pays more for designer goods, so why not pay to appear in prestigious journals? This is all very well, but it rather distorts the aims and aspirations of science: good science should be published regardless of the wealth of those doing it.

To my mind, the open-access/author-pay model is very much an attempt to hold onto business models of the past. My prediction is that future 'journals' will accept all papers, with no refereeing process, and that papers will accumulate something similar to Facebook's 'likes'. Good papers will be recognised by the community as good. This seems to be the direction in which arXiv is moving. It will also help those concerned with evaluating research impact: the arguments about journal impact factors, citation rates, and so on will be of little relevance. Of course, gaming will be possible (and ways to tackle it will be developed), but gaming is always possible (even for the current refereeing model).

This new model is sometimes described as a 'publish then filter' model, to be contrasted with the current 'filter then publish' model. Since the target readership will have the opportunity to evaluate papers, rather than relying on the views of a small group of relatively arbitrarily selected referees, this may well produce higher quality science.

From the perspective of a journal like *Statistics and Computing*, the question this raises is, is it possible to move from one model to the other? Or is the difference simply too great, so that old-style journals have to be scrapped, to be replaced by a radically new model? It might be possible to make a gradual transition, in which some papers go through a refereeing process, so attracting a stamp of quality, while others do not, and for which readers have to vote in the way described above. That such gradual transitions may be possible is illustrated by journals which mix the traditional subscription approach with the open-access approach.

The classical 'filter then publish' model has other demerits, which the alternative might also be able to overcome, such as

(i) Papers do not have to squeeze through the bottleneck of ill-informed referees. Who amongst us has not had the experience of receiving comments from a referee who (we might claim) clearly did not really grasp the significance of the paper?
(ii) Publication bias is a very well-attested problem, perhaps especially in medical research. The complex filtering process through which papers end up appearing in journals is well known to bias toward statistically significant results—to the extent that it has now become a topic of scientific investigation in its own right. See, for example, Ioannidis (2005).

On the other hand, of course, in contrast to (i), papers would not be improved by well-informed referees' comments. Again, who amongst us has not had the experience of receiving comments from a referee which led to significant improvement of a paper? A system in which papers are amended in response to comments, making them dynamic entities rather than the current crystallised entities, would be a significant advance (provided a history of versions is kept).

4 Statistics then

So much for the changes in scientific, and in particular, statistical publishing over the life of *Statistics and Computing*, but what about statistics itself?

I had intended to begin to answer that question by saying 'statistics was in transition'. However, while that is true, one could argue a case that any *technology* like statistics is always in transition. In the case of statistics, a regular stream of mathematical developments, from the end of the nineteenth century to the present day, often in response to the novel demands of new application domains, has meant that the discipline has kept evolving.

However, by the time *Statistics and Computing* was launched, a more fundamental change was underway - and a change which has continued up to the present day (and which, I must add, has not stopped: the future of the discipline of statistics will be dramatically different from the past).

Whereas, throughout most of the twentieth century, statistics had been regarded as a mathematical discipline (some even, at least for a period, regarded it as a branch of mathematics), it was becoming obvious that, with just as much justification, one could regard it as a computational discipline. The opening sentence of my first editorial, in Volume 1, Number 1, was 'Statistical science is evolving dramatically under the impact of computers.'

Progress in computer technology had led to changes in statistical methodology which can best be described as revolutionary. More than that, progress in computer technology was leading to changes in statistical ideas. The classic example is of course, Bayesian statistics, which has become a practical statistical philosophy, whereas previously computational constraints meant that it could be used on only the simplest of problems. But there are many other areas of statistics where the impact of the computer has been revolutionary. For example, the computer has enabled us

- to do things in a split second which would previously have taken impossible amounts of time, such as inverting large matrices. This facilitated great leaps forward in multivariate analysis and its applications;
- to do things which were simply infeasible before. Examples are inverting even larger matrices (e.g. a million square), iterative methods (leading to tools such as generalised linear models), and simulation (opening all sorts of areas);
- to do things people had not thought of before: various kinds of resampling methods fall into this camp, such as bootstrap methods; and
- to carry out dynamic interactive visualisation of data sets.

These developments were taking place squarely within the realm of statistics, but in parallel other tools were being created in other places—especially in computer science departments. Examples are neural networks, expert systems, and support vector machines. Occasionally, tensions had arisen between the different disciplines, but gradually these tensions were relaxed as each discipline recognised the value of the perspectives the others brought to bear. Thus, for example, when statisticians started to look at neural networks, they were able to embed the estimation procedures (e.g. the so-called 'feed-forward' estimation method) in statistical estimation theory which had been developed over many decades. This made neural network technology more rigorous, leading to better understanding of its properties, and identification of the sorts of applications to which it would be valuable. Another example of the two different perspectives coming together is in belief networks, conditional independence graphs, graphical models, and related areas. These developments, in particular, were brought together in a unique series of conferences on 'AI and Statistics' which started in 1984.

The complementary strengths of different disciplines means that a healthy synergy can emerge when they work together. So, for example, the mathematical heritage of statistics means that it tends to be more cautious, wishing to establish the properties of methods before using them. In contrast, the engineering heritage of computer science has encouraged a more adventurous, try-it-and-see approach. The benefits of working together are obvious. It is perhaps no coincidence that there are hackers in computer science, but not in statistics. What would a statistical hacker be? A crude and oversimplified characterisation might describe computer science as being more aimed at building *tools* (so at my own university, the Department of Computing is in the Faculty of Engineering), while statistics is more aimed at discovering something about nature or making decisions (and, at my own University, the statistics group is in the Faculty of Science).

As well as leading to dramatic changes in the nature of statistics as a discipline, the advent of the computer also led to dramatic changes in statistical practice. This was apparent in the appearance of major statistical packages such as SPSS and SAS, and also in statistical languages such as S+. But the ease with which such tools enabled statistical analyses to be undertaken was not an unalloyed good. Some of us were concerned that the very ease of use, especially by the statistically uninformed, created risks of its own. After all, subject to syntactic constraints, any set of numbers you feed into an analysis will give a result, whether or not it is meaningful. In a real sense, the ease of use of statistical software meant that instead of thinking beforehand about what was the right thing to do (the right question, the right tool), there was the possibility of trying many analyses—and if one failed to produce the result you wanted, another could be tried. Thought was being replaced by computer power, which is not nec-

essarily a good idea [see, for example, Hand (1994, 2014a) for some discussion and examples of this point]. Effective statistical analysis depends critically on understanding the scientific question, so that automatic or rote strategies are high-risk strategies.

Particular risks arose when a package offered many different summary statistics, since there was then a temptation to output them all and sift through them to find those which were significant or which supported a hypothesis. With the parallel move to electronic rather than manual data collection, the dangers of multiple testing and overfitting (e.g. in high-dimensional problems) became even greater.

Partly in a response to such risks, there was a flurry of activity developing 'statistical expert systems'—programmes aimed at providing guidance, help, and protection during the course of a statistical analysis [see for example, Nelder (1989) and Hand (1987, 1993)]. This work has led onto graphical user interfaces.

5 The future of statistics

We are clearly living at a time when statistics and computing have come together. Computer science underlies the manipulation and handling of the increasingly large data sets we have to contend with, and statistics underlies the extraction of useful information from these data sets. Indeed, at the time of writing this article, scarcely a day goes by without some mention of 'big data' in the news media.

But big data is not actually all that new. In a very real sense, the phrase 'big data' is merely a media rebranding of the phrase 'data mining' (perhaps via the phrases 'data analytics', 'business analytics', and perhaps to be replaced in turn by the phrase 'data science'). Data mining is defined as 'the discovery of interesting or valuable structures in large data sets.' The first books on data mining began to appear 5 or 10 years after the launch of *Statistics and Computing* [e.g. Fayyad et al. (1996) and Hand et al. (2001)]—and even at that time some very large data sets were around. For example, Walmart carried out some seven billion transactions per year by 1994, and AT&T carried 70 billion long-distance phone calls in 1997.

What was interesting was that while, in many quarters, data mining was met with enthusiasm for the possibilities it opened, in other quarters it was met with suspicion. The enthusiasm reflected the potential medical and scientific discoveries, as well as the commercial opportunities for increased profit and improved products and services consequent on understanding one's customers better. The suspicion seemed to be of two types.

The first arose from an awareness of the dangers of massive unconstrained search. Search long enough, in enough places, for enough kinds of structures, and you are almost guaranteed to come up with something [see Hand (2014b), for some examples of this]. As has been said, 'torture the data long enough and they are bound to confess.' It is curious that, in some quarters (e.g. economics), the phrase 'data mining' became synonymous with such search problems and so became something to be avoided. This was despite the books and journals promoting data mining, and the range of success stories.

The second kind of suspicion arose from a narrow perception of what was meant by data mining, restricting it to refer to human behaviour. A classic example of this was the *Total Information Awareness* programme (later *Terrorist Information Awareness*) in the US, aimed at trying to detect potential terrorists' outrages in advance, from anomalous patterns of consumer behaviour. The perception was that data mining was all about monitoring and snooping on individuals.

As will be clear, these concerns about data mining both arise as a consequence of misinterpreting one small part of the technology or its application as the entire thing. But they did generate some anxiety in the data mining community that the public backlash might constrain methodological research. 'Big data' has, to a large extent, not been so concerned with this—though in some domains (especially personal privacy) the same issues are arising again.

The danger underlying the first kind of suspicion (very extensive search) has, of course, not gone away. However, advances in statistics, allied with the necessary computer power to support these advances (very much in tune with the aims of *Statistics and Computing*), have alleviated them—I am thinking of advances in multiple testing, not least the developments in false discovery rate methods.

Given the similarities between the data mining initiatives and the big data initiatives, it is sensible to ask what we learnt from the data mining experience. One thing, which many big data proponents have not yet recognised, is that discoveries in large data sets are due, in order of decreasing frequency, to

(i) Errors in the data. Data quality is a *critical* issue if big data projects are to succeed. While this is true of all data, it is particularly important for large data sets since the computer is necessarily an intermediary between you and the data: you cannot examine each data point for accuracy.
(ii) Chance. In any large data set, there are almost certain to be unusual structures reflecting mere randomness, rather than anything of genuinely interest (Hand 2014b). Overfitting and excessive search (the physicists' *look elsewhere* effect) can guarantee this. Recall William Kruskal's observation that 'A reasonably perceptive person, with some common sense and a head for figures, can sit down with almost any structured and substantial data

set or statistical compilation and find strange-looking numbers in less than an hour.' (Kruskal 1981, p. 508)

(iii) Real phenomena, but phenomena which are already known or of no interest. The examples I often give are the facts that about half of the married people in the US are female and that in any time series the maxima and minima alternate.

(iv) Real phenomena which are newly discovered and are of value or genuine interest.

Large data sets do present novel challenges. On the one hand, there are data manipulation challenges, such as how to sort, merge, or select from massive collections. And on the other, there are deep inferential challenges, such as tackling multiplicity, detecting anomalies, modelling tail distributions, and so on. If the development of statistics has been driven by its use in diverse application domains, along with (more recently) the power provided by the computer, then we can expect the new challenges from massive data sets to lead to a new generation of statistical ideas. What is clear about all this is that it is the combination of statistics and computing which represents the future of data science.

If 'big data' represents a novel data analytic domain, there are other, closely related domains, which also pose new challenges for statisticians. One is the so-called 'administrative data'. These are data collected as a side-effect of some other exercise, rather than to answer the questions now being put to them. One important challenge of administrative data is measuring and communicating uncertainty. Classically, statistics has coped with stochastic uncertainty arising from sampling or randomisation considerations, or with belief uncertainty arising from Bayesian interpretations of probability. Administrative data present novel kinds of data quality issues and their consequent uncertainties. The fact that there are many ways in which values can be distorted and many reasons for which they may be missing means that the mathematical elegance of confidence or credibility intervals will be difficult to replicate. The problem is that the notion that one has 'all' the data is rarely true, and reality can depart from 'all' in a wide variety of ways.

The second closely related domain is that of 'open data'. On the one hand, this has been promoted as a necessity for good science: the fact that other researchers should have access to the data, to check the assumptions, test their own models, and so on. And on the other hand, open data has been promoted as a necessity for transparency in government: enabling the public to see that policy is based on a sound evidence base, and to evaluate our legislators, public services, and others to see that they are doing a good job. But open data bumps up against the fact that there have to be limits. Individuals may not want their medical records, financial affairs, and personal relationships accessible for everyone to see. Once again, there are intersecting statistical and computational challenges: how can data be released without compromising privacy? what about the possibility of linking data sets—huge promise for good (e.g. linking dietary, lifestyle, and medical data sets), but also great potential for intrusion. Such questions are the focus of much current research, at the intersection of statistics and computing.

Big data, and its allied concepts of administrative data and open data, represents just one class of areas through which data science is going to have massive impact on our lives. There are many others. There have always been many different kinds of users of statistics, ranging from expert professional statisticians, with PhDs in the subject, through specialists in other disciplines, such as psychologists, astronomers, and medical doctors, who needed to have some understanding of and facility with at least certain classes of statistical ideas and tools, all the way to the interested layman, who wished to make sense of figures reported in the newspapers. It seems fairly clear that a grasp of statistical concepts will become even more important as time passes, to a wider cross-section of people. As far back as 1938, H.G. Wells was able to write '*a certain elementary training in statistical method is becoming as necessary for anyone living in this world of today as reading and writing*' (Wells 1938). This is even more true nowadays and will be yet more true in the future.

Another development which has begun and which will accelerate as time passes is the need to extract useful understanding from unstructured or complex data: text, images, networks, and so on. Simulations are widely used—through Markov Chain Monte Carlo, certainly, but also in simulating large physics experiments, galaxy formation, economies, and biological phenomena. This is likely to become even more important in the future—and again represents a perfect illustration of the awesome synergy arising from statistics and computing.

Streaming data is now an area of research in its own right. Examples include the financial markets, intensive care monitoring, fault detection in engines, cybersecurity, and many other domains. In one sense, this is an area of research which has just begun, but it is an area which will be critically important and find very widespread use within the next few years—not least because of the advent of the 'Internet of Things'.

Visualisation, and dynamic interaction with graphical displays, perhaps by means of virtual reality, keeps promising exciting breakthroughs. Again it is a perfect example of what can be done when the disciplines of statistics and computing work in harness. I am sure that we will see extraordinary developments within the next decade.

I also predict—I'm going out a little on a limb here—that we will see much more development of ensemble systems and parallel computing models for tackling inferential problems.

6 Conclusion

The title 'conclusion', for this section, is intended, on the one hand, to be traditional, for that's how many papers end, and on the other hand to be ironic. It's ironic because, far from being at the conclusion of this two-legged journey of statistics and computing, I would suggest we are still near the start: the computer has changed statistics beyond recognition, but the extent of that change has just begun. Computer technology will continue to advance—look at all the work on quantum computation—and we can expect statistical technology to advance on the back of it.

I would like, however, to sound a cautionary note: the computer cannot replace statistical thinking. In recent months, I have seen mistaken conclusions drawn in a variety of studies, for a variety of reasons: because the researchers did not understand the difference between fixed and random effects, because they did not understand the properties of different measures of correlation, because they did not understand the significance of the distributional assumptions underlying a model, because they failed to grasp the difference between conditional and unconditional relationships, and for other reasons. The computer may allow us to explore data in ways we could not previously have imagined, but that does not mean we can ignore the fundamentals. In pulling the two disciplines together, *Statistics and Computing* has a particularly important role to play.

I knew, in 1991, when I launched the journal, that it was tapping into the future. The success of the journal, and its continued growing prestige, has shown that I was right. But it could only have achieved that through the efforts of my successors as editor. I would like to pay tribute to Wayne Oldford, Gilles Celeux, and now Mark Girolami. Without them, their hard work, and their insights, the journal would not be what it is today.

The last paragraph of my opening editorial from Volume 1, Number 1, read 'The two disciplines of statistics and computer science have more and more to offer each other. The purpose of this journal is to support and strengthen that relationship.' The first sentence remains true—to the extent that in many areas the two disciplines are so closely interwoven that they are now indistinguishable. And it is manifestly clear from the papers published in recent issues that the aim expressed in the second sentence is also being achieved.

References

CasesBlog. (http://casesblog.blogspot.co.uk/2011/03/there-are-25400-scientific-journals-and.html (2014). Accessed 18 Nov 2014

Eldredge, N.: Mathgen paper accepted. http://thatsmathematics.com/blog/archives/102 (2012). Accessed 23 Nov 2014

Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, P., Uthurusamy, R.: Advances in Knowledge Discovery and Data Mining. AAAI Press / The MIT Press, Menlo Park (1996)

Hand, D.J.: A statistical knowledge enhancement system. J. R. Stat. Soc. Ser. A **150**, 335–345 (1987)

Hand, D.J.: Artificial Intelligence Frontiers in Statistics. Chapman and Hall, London (1993)

Hand, D.J.: Deconstructing statistical questions (with discussion). J. R. Stat. Soc. Ser. A **157**, 317–356 (1994)

Hand, D.J.: Solving the right problem. Bull. Inst. Math. Stat. **43**(3), 6 (2014a)

Hand, D.J.: The Improbability Principle: Why Coincidences, Miracles, and Rare Events Happen Every Day. Scientific American/FSG, New York (2014b)

Hand, D.J., Mannila, H., Smyth, P.: Principles of Data Mining. The MIT Press, Cambridge (2001)

Ioannidis, J.P.A.: Why most published research findings are false. PloS Med. **2**, 696–701 (2005)

Kruskal, W.: Statistics in society: problems unsolved and unformulated. J. Am. Stat. Assoc. **76**, 505–515 (1981)

Nelder, J.A.: A statistical expert system: some experiences in constructing a knowledge-based front-end for GLIM. Am. J. Math. Manag. Sci. **9**, 371–388 (1989)

Wells, H.G.: World Brain. Methuen and Co, London (1938)