Strength in Diversity: The Advance of Data Analysis

David J. Hand

Department of Mathematics, Imperial College, 180 Queen's Gate,
London SW7 2AZ, UK
d.j.hand@imperial.ac.uk

Abstract. The scientific analysis of data is only around a century old. For most of that century, data analysis was the realm of only one discipline - statistics. As a consequence of the development of the computer, things have changed dramatically and now there are several such disciplines, including machine learning, pattern recognition, and data mining. This paper looks at some of the similarities and some of the differences between these disciplines, noting where they intersect and, perhaps of more interest, where they do not. Particular issues examined include the nature of the data with which they are concerned, the role of mathematics, differences in the objectives, how the different areas of application have led to different aims, and how the different disciplines have led sometimes to the same analytic tools being developed, but also sometimes to different tools being developed. Some conjectures about likely future developments are given.

1 Introduction

This paper gives a personal view of the state of affairs in data analysis. That means that inevitably I will be making general statements, so that most of you will be able to disagree on some details. But I am trying to paint a broad picture, and I hope that you will agree with the overall picture.

We live in very exciting times. In fact, from the perspective of a professional data analyst, I would say we live in the *most* exciting of times. Not so long ago, analysing data was characterised by drudgery, by manual arithmetic, and the need to take great care over numerical trivia. Nowadays, all that has been swept aside, with the burden of tedium having been taken over by the computer. What we are left with are the high-level interpretations and strategic decisions; we look at the summary values derived by the computers and make our statements and draw conclusions and base our actions on these. It is clear from this that the computer has become *the* essential tool for data analysis.

But there is more. The computer has not merely swept aside the tedium. The awesome speed of numerical manipulation has permitted the development of entirely new kinds of data analytic tools, being applied in entirely new ways, to entirely new kinds, and indeed sizes, of data sets. The computer has given us new ways to look at things. The old image, that data analysis was the realm of the boring obsessive, is now so diametrically opposite to the new truth as to be laughable.

J.-F. Boulicaut et al. (Eds.): ECML 2004, LNAI 3201, pp. 18–26, 2004.

This paper describes some of the history, some of the tools, and something of how I see the present status of data analysis. So perhaps I should begin with a definition. *Data analysis is the science of discovery in data, and of processing data to extract evidence so that one can make properly informed decisions*. In brief, data analysis is *applied philosophy of science*: the theory and methods, not of any particular scientific discipline itself, but of *how to find things out*.

2 The Evolution of Data Analytic Disciplines

The origins of data analysis can be traced back as far back as one likes. Think of Kepler and Gauss analysing astronomical data, of Florence Nightingale using plots to demonstrate that soldiers were dying because of poor hygiene rather than military action, of Quetelet's development of 'social mechanics', and the fact that world's oldest statistical society, the Royal Statistical Society, was established in 1834. But these 'origins' really only represent the initial stirrings: it wasn't until the start of the 20th century that a proper scientific discipline of data analysis really began to be formed. That discipline was statistics, and for the first half of the 20th century statistics was the only data analytic game in town. Until around 1950, statistics *was* the science of data analysis. (You will have to permit me some poetic leeway in my choice of dates: 1960 might be more realistic.)

Then, around the middle of the 20th century, the computer arrived and a revolution began. Statistics began to change rapidly in response to the awesome possibilities the computer provided. There is no doubt that, had statistics been born now, at the start of the 21st century, rather than 100 years ago at the start of the 20th, it would be a very different kind of animal. (Would we have the *t*-test?.) Moreover, although statistics was the intellectual owner of data *analysis* up until about 1950, it was never the intellectual owner of *data* per se, and in the following decades other changes occurred which were to challenge the position assumed by statistics. In particular, another discipline grew up, whose primary responsibility was, initially, the storage and manipulation of data. From data manipulation to data analysis was then hardly a large step. Statistics was no longer the only player.

Nowadays, of course, computer science has grown into a vast edifice, and different subdisciplines of it have developed as specialised areas of data analysis, all overlapping with each other and overlapping with their intellectual parent, statistics. These subdisciplines include machine learning, pattern recognition, and data mining, and one could arguably include image processing, neural networks, and perhaps even computational learning theory and other areas also. I cannot avoid remarking that Charles Babbage, typically regarded as one of the fathers of computing with his *analytical engine*, would have been fascinated by these developments: he was also one of the founders of the Royal Statistical Society. Of course, these various data analytic disciplines are not carbon copies of each other. They have subtly different aims and emphases, and often deal with rather different kinds of data (e.g. in terms of data set size, correlations, complexities, etc.). One of my aims in this talk is to examine some of these differences. Moreover, if the computer has been the strongest influence leading to the de-

velopment of new data analytic technologies, application areas have always been and continue to have a similar effect. Thus we have psychometrics, bioinformatics, chemometrics, technometrics, and other areas, all addressing the same sorts of problems, but in different areas. I shall say more about this below.

3 Data

I toyed briefly with the idea of calling this talk 'analysing tomorrow's data' since one of the striking things about the modern world of data analysis is that the data with which we now have to deal could not have been imagined 100 years ago. Then the data had to be painstakingly collected by hand since there was no alternative, but nowadays much data acquisition is automatic. This has various consequences.

Firstly, astronomically vast data sets are readily acquired. Books on data mining (e.g. [2],[3]), which is that particular data analytic discipline especially concerned with analysing large data sets, illustrate the sorts of sizes which are now being encountered. The word *terabyte* is no longer unusual. When I was taught statistical data analysis, I was taught that first one must familiarise oneself with one's data: plot it this way and that, look for outliers and anomalies, fit simple models and examine diagnostics. With a billion data points (one of the banking data sets I was presented with) this is clearly infeasible. Other problems involve huge numbers of variables, and perhaps relatively few data points, posing complex theoretical as well as practical questions: bioinformatics, genomics, and proteomics are important sources of such problems.

Secondly, one might have thought that automatic data acquisition would mean better data quality, since there would be no subjective human intervention. Unfortunately, this has not turned out to be the case. New ways of collecting data has meant new ways for the data collection process to go wrong. Worse, large data sets can make it more difficult to detect many of the data anomalies.

Data can be of low quality in many ways: individual values may be distorted or absent, entire records may be missing, measurement error may be large, and so on. As discussed below, much of statistics is concerned with *inference* - with making statements about objects or values not seen or measured, on the basis of those which have been. Thus we might want to make statements about other objects from a population, or about the future behaviour of objects. Accurate inferences can only be made if one has accurate information on how the data were collected. Statisticians have therefore predicated their analyses on the assumption that the available observations were drawn in well-specified ways, or that the departures from these ways were understood and could be modelled. Unfortunately, with many of today's data sets, such assumptions often cannot be made. This has sometimes made statisticians (quite properly) wary of analysing such data. But the data still have to be analysed: the questions still need answers. This is one reason why data mining has been so successful, at least at first glance. Data miners have been prepared to examine distorted data, and to attempt to draw conclusions about it. It has to be said, however, that often that willingness has arisen from a position of ignorance, rather than one of awareness of the risks that were being taken. Relatively few reports of the conclusions extracted from a data mining

exercise, for example, qualify those conclusions with a discussion of the possible impact of selectivity bias on the data being analysed. This is interesting because, almost by definition, data mining is secondary data analysis: the analysis of data collected for some other purpose. The data may be of perfect quality for its original purpose (e.g. calculating your grocery bill in the store), but of poor quality for subsequent mining (e.g. because some items were grouped together in the bill).

A third difference between many modern data analysis problems and those of the past is that nowadays they are often dynamic. Electronics permit data to be collected as things happen, and this opens the possibility of of making decisions as the data are collected. An example is in commercial transactions, where a customer can supply information and expects an immediate decision. In such circumstances one does not have the luxury of taking the data back to one's laboratory and analysing it at leisure. Speech recognition is another example. This issue has led to new kinds of analytic tools, with an emphasis on speed and not merely accuracy. No particular area of data analysis seems to have precedence for such problems, but the computer science side, perhaps especially machine learning clearly regards such problems as important.

Although every kind of data analytic discipline must contend with all kinds of data, there is no doubt that different kinds are more familiar in different areas. Computational areas probably place more emphasis on categorical data than on continuous data, and this is reflected in the types of data analytic tools (e.g. methods for extracting association rules) which have been developed.

4 The Role of Mathematics

Modern statistics is often regarded as a branch of mathematics. This is entirely inappropriate. Indeed, the qualitative change induced by the advent of the computer means that statistics could equally be regarded as a branch of computer science.

In a sense statistics, and data analysis more generally, is the opposite of mathematics. Mathematics begins with assumptions about the structure of the universe of discourse (the axioms) and seeks to deduce the consequences. Data analysis, on the other hand, begins with observations of the consequences (the data) and seeks to infer something about the structure of the universe. One consequence of this is that one can be a good mathematician without understanding anything about any area to which the mathematics will be applied – one primarily needs facility with mathematical symbol manipulation – but one cannot be a good statistician without being able to relate the analysis to the world from which the data arose. This is why one hears of mathematics prodigies, but never statistics prodigies. Analysis requires understanding.

There are other differences as well. Nowadays a computer is an essential and indispensable tool for statistics, but one can still do much mathematics without a computer. This is brought home to our undergraduate students, taking mathematics degrees, with substantial statistical components, when they come to use software: statistical software packages such as Splus, R, SAS, SPSS, Stata, etc., are very different from mathematical packages such as Maple and Mathematica. Carrying out even fairly basic statistical analyses using the latter can be a non-trivial exercise.

David Finney has commented that it is no more true to describe statistics as a branch of mathematics than it would be to describe engineering as a branch of mathematics, and John Nelder has said '*The main danger, I believe, in allowing the ethos of mathematics to gain too much influence in statistics is that statisticians will be tempted into types of abstraction that they believe will be thought respectable by mathematicians rather than pursuing ideas of value to statistics*.'

There is no doubt that the misconception of statistics as mathematics has been detrimental in the past, especially in commercial and business applications. Data mining, in particular took advantage of this - its very name spells glamour and excitement, the possibility of gaining a market edge for free. But there are also other examples where the image of statistics slowed its uptake. For example, experimental design (that branch of statistics concerned with efficient and cost effective ways to collect data) was used in only relatively few sectors (mostly manufacturing). Reformulations of experimental design ideas under names such as the Taguchi method and Six Sigma, however, have had a big impact. If anything ought to convince my academic colleagues of the power of packaging and presentation, then it should be these examples.

5 Several Cultures Separated by a Common Language

The writer George Bernard Shaw once described England and America as '*two cultures divided by a common language*', and I sometimes feel that the same applies to the various data analytic disciplines. Over the years, I have seen several intense debates between proponents of the different disciplines. Part of the reason for this lies in the different philosophical approaches to investigation. Statistics, perhaps because of its mathematical links, places a premium on proof and mathematical demonstration of the properties of data analytic tools. For example, demonstrating mathematically that an algorithm will always converge. Machine learning, on the other hand, places more emphasis on empirical testing. Of course there is overlap. Most methodological statistics papers include at least one example of the methods applied to real problems, and most machine learning papers describe the ideas in mathematical terms, but there is a clear difference in what is regarded as of central importance.

Another reason for the debates has been that many of the ideas were developed in parallel, by researchers naturally keen to establish their priority and reputation. This led to claims to the effect that 'we developed it first' or 'we demonstrated that property years ago.' This was certainly evident in the debates on recursive partitioning tree classifiers, which were developed in parallel by the machine learning and statistics communities.

Misunderstandings can also arise because different schools place emphasis on different things. Early computer science perspectives on data mining stressed the finding of patterns in databases. This is perfectly natural: it is something often required (e.g. what percentage of my employees earn more than €x p.a.?). However, this is of limited interest to a statistician, who will normally want to make an inference to a wider population or to the future (e.g. what percentage of my employees are likely to earn more than €x p.a. next year?). Much work on association analysis has ignored this inferential aspect. Moreover, much work has also made a false causal assumption:

while it is *interesting* to know that ten times as many people who bought A also bought B, it is *valuable* to know that if people can be induced to buy A they will also buy B, and the two are not the same.

While there have been tensions between the different areas when they develop similar models, each from their own perspective, there is no doubt that these tensions can be immensely beneficial from a scientific perspective. A nice example of this is the work on feedforward neural networks. These originally came from the computer (or, one might argue, the cybernetics, electrical engineering, or even biological) side of things. The perspective of a set of fairly simple interacting processors dominated. Later, however, statisticians became involved and translated the ideas into mathematical terms: such models can be written as nested sequences of nonlinear transformations of linear combinations of variables. Once written in fairly standard terms, one can apply the statistical results of a century of theoretical work. In particular, one could explain that the early neural network claims of very substantial improvement in predictive power were likely to be in large part due to overfitting the design data, and to present ideas and tools for avoiding this problem. Of course, nowadays all these are well understood by the neural network community, but this was certainly not the case in the early days (I can remember papers presenting absurdly overoptimistic claims), even though statisticians had known about the issues for decades.

If the computer is leading to a unification of the data analytic schools, so also are some theoretical developments. The prime examples here, of course, are Bayesian ideas. Bayes's theorem tells us how we should update our knowledge in the light of new information. This is the very essence of learning, so it is not surprising that machine learning uses these ideas. With the advent of practical computational tools for evaluating high dimensional integrals, such as MCMC, statistics has also undergone a dramatic Bayesian revolution, not only in terms of dynamic updating models but also in terms of model averaging. Indeed, model averaging, like the understanding of overfitting (indeed, closely connected to it), has led to deep theoretical advances. Tools such as boosting and bagging are based on these sorts of principles. Boosting, in particular, is interesting from our perspective because it illustrates the potential synergy which can arise from the disparate emphases of the different disciplines. Originally developed by the machine learning community, who proposed it on fairly intuitive grounds and showed that it worked in practical applications, it was then explored theoretically by statisticians, who showed its strong links to generalised additive models, a well-understood class of statistical tools. The most recent tool to experience this initial development, followed by an exposure to the ideas and viewpoints of other data analytic disciplines, is that of support vector machines.

In fact, perceptrons (the progenitor of support vector machines) and logistic discrimination provide a very nice illustration of the difference in emphasis between, in this case, statistical and machine learning models for classification. Logistic discrimination fits a model to the probability that an object with given features $\mathbf{x}$ will belong to class 0 rather than class 1. Typically, the model is fitted by finding the parameters which maximise the design set log likelihood:

$$\log L \propto \sum_{i=1}^{n} \log \hat{p}\left(0 \mid \mathbf{x}_i\right). \tag{1}$$

Classification is then effected by comparing an estimated probability with a threshold. It is immediately clear from (1) that all design set data points contribute - it is really an average of contributions. This is fine if one's model $\hat{p}\left(0 \mid \mathbf{x}\right)$ has the form of the 'true' function $p\left(0 \mid \mathbf{x}\right)$. But this is a brave assumption. It is likely that the model is not perfect. If so, one must question the wisdom of letting data points with estimated probability far from the classification threshold contribute the same amount to the fit criterion (1) as do those near to it (see [4]). In contrast, perceptron models focus attention on whether or not the design set points are correctly classified: quality of fit of a model far from the decision surface, which is broadly irrelevant to classification performance, does not come into it.

An example of another area which has been developed in rather different ways by different disciplines is the area I call *pattern discovery*. This is the search for, identification of, and description of anomalously high local densities of data points. The computer science literature has focused on algorithms for finding such configurations. In particular, a great deal of work has occurred when the data are character strings, in, especially text search (e.g. web search engines) and nucleotide sequences. In contrast, the statistical work has concentrated on the inference problem, developing scan statistics for deciding whether a local anomaly represents a real underlying structure is just random variation of a background model. Ideas of this kind have been developed in many application areas, including bioinformatics, technical stock chart analysis, astronomy, market basket analysis, and others, but the realisation that they are all tackling very similar problems appears to be only recent.

Implicit in the last two paragraphs is one of the fundamental differences in emphasis between computational and statistical approaches to data analysis - again an understandable difference in view of their origins. This is the emphasis of the computational approaches on algorithms (e.g. the perceptron error-correcting algorithm) and the emphasis of the statistical approaches on models (e.g. the logistic discrimination model). Both algorithms and models are, of course, important when tackling real problems.

It is my own personal view that one can also characterise the difference between the two perspectives, at least to some extent, in terms of degree of risk. The computational schools seem often prepared to try something without the assurance that it will work, or that it will always work, but in the hope (or knowledge from previous analyses) that it will sometimes work. The statistical schools seem more risk averse, requiring more assurance before carrying out an analysis. Perhaps this is illustrated by the approaches to pattern discovery mentioned above: the data mining community develops algorithms with which to detect possible patterns, while the statistical community develops tools to tell whether they are real or merely chance. Once again, both perspectives are valuable, especially in tandem: adventurous risk-taking offers the possibility of major breakthroughs, while careful analysis shows one that the method gives reliable results.

6 Future Tools and Application Areas?

Of course, the various data analytic disciplines are constantly evolving. We live in very exciting times because of the tools which have been developed over the past few decades, but that development has not stopped. If anything, it has accelerated and will continue to do so as the computational infrastructure continues to develop. This means faster and larger (in terms of all dimensions of datasets). Judging from the past, this will translate into analytic tools about which one previously could only have dreamt, and, further, into tools one could not even have imagined.

If the computer is one force driving the development of new data analytic tools, I can see at least two others.

The first of these are application areas, mentioned above. Certainly, the growth of statistics over the 20th century was strongly directed by the applications. Thus agricultural requirements led to the early development of experimental design, psychology motivated the development of factor analysis and other latent variable models, medicine led to survival analysis, and so on. In other areas, speech recognition stimulated work on hidden Markov models, robotics stimulated work on reinforcement learning, etc. Of course, once developments have been started, and the power of the tools being developed has been recognised, other application areas rapidly adopt the tools.

As with the impact of developing computational infrastructure, I see no reason to expect this influence of application areas to stop. We are currently witnessing the particular requirements of genomic, proteomic, and related data leading to the development of new analytic tools; for example, methods for handling *fat data* - data involving many (perhaps tens of thousands of) variables, but few (perhaps a few tens of) data points. Mathematical finance is likewise an area which is shifting its centre of gravity towards analysis. Until recently characterised by mathematical areas such as stochastic calculus, it is increasingly recognised that data analysis is also needed - the values of the model parameters must come from somewhere. More generally, the area of personal finance is beginning to provide a rich source of novel problems, requiring novel solutions. The world wide web, of course, is another source of new types of data, and new problems. This area, in particular, is a source of data which is characterised by its dynamic properties, and I expect the analysis of dynamic data to play an even more crucial role in future developments. Decisions in telecoms systems, even in day-to-day purchasing transactions, are needed *now*, not after a leisurely three months' analysis of a customer's track record and characteristics. Delay loses business.

The second additional driving force I can see is also not really a new one. It has always been with us, but it will lead to the development of new kinds of tools, in response to new demands and also enabled by the advancing computational infrastructure. This is the need to model finer and finer aspects of the presenting problems. A recent example of this is in the analysis of repeated measures data. The last two decades have witnessed a very exciting efflorescence of ideas for tackling such data. The essential problem is to recognise and take account of the fact that repeated measurements data are likely to be correlated (with the (multiple) series being too short to use time series ideas). Classical assumptions of independence are all very well, but more accurate models and predictions result when the dependence is modelled. Another

example of such 'finer aspects' of the presenting problem, which has typically been ignored up until now, is the fact that predictive models are likely to be applied to data drawn from distributions different from that from which the design data were drawn (perhaps a case for dynamic models). There are many other examples.

There is, however, a cautionary comment to be made in connection with this driving force. It is easy to go too far. There is little point is developing a method to cope with some aspect of the data if the inaccuracies induced by that aspect are trivial in comparison with those arising from other causes. Data analysis is not a merely mathematical exercise of data manipulation.

If we data analysts live in exciting times, I think it is clear that the future will be even more exciting. Looking back on the past it is obvious that the tensions between the different data analytic disciplines have, in the end, been beneficial: we can learn from the perspectives and emphases of the other approaches. In particular, we should learn that the other disciplines can almost certainly shed light on and help each of us gain greater understanding of what we are trying to do. We should look for the *synergies*, not the *antagonisms*.

I'd like to conclude with two quotations. The first is from John Chambers, the computational statistician who developed Splus and who won the 1998 ACM Software System Award for that work. He wrote: *'Greater statistics can be defined simply, if loosely, as everything related to learning from data, from the first planning or collection to the last presentation or report. Lesser statistics is the body of specifically statistical methodology that has evolved within the profession - roughly, statistics as defined by texts, journals, and doctoral dissertations. Greater statistics tends to be inclusive, eclectic with respect to methodology, closely associated with other disciplines, and practiced by many outside of academia and often outside of professional statistics. Lesser statistics tends to be exclusive, oriented to mathematical techniques, less frequently collaborative with other disciplines, and primarily practiced by members of university departments of statistics.'* [1]

John has called the discipline of data analysis 'greater statistics', but I am sure we can all recognise what we do in his description. What we call it is not important. As Juliet puts it in Act II, Scene ii of Shakespeare's *Romeo and Juliet*:

> *'What's in a name? that which we call a rose*
> *By any other name would smell as sweet.'*

References

1. Chambers J.M. Greater or lesser statistics: a choice for future research. *Statistics and Computing*, **3**, (1993) 182-184.
2. Giudici P. *Applied Data Mining*. Chichester: Wiley. (2003)
3. Hand D.J., Mannila H., and Smyth P. *Principles of Data Mining*, Cambridge, Massachusetts: MIT Press. (2001)
4. Hand D.J. and Vinciotti V. Local versus global models for classification problems: fitting models where it matters. *The American Statistician*. **57**, (2003) 124-131.