THE AUSTRALIAN NATIONAL UNIVERSITY

THE GRADUATE SCHOOL

New Approaches To Using Scientific Data – Statistics, Data Mining And Related Technologies In Research And Research Training

J.H. Maindonald

OCCASIONAL PAPER GS 98/2

Abstract

This paper surveys technological changes that affect the collection, organization, analysis and presentation of data. It considers changes or improvements that ought to influence the research process and direct the use of technology. It explores implications for graduate research training.

The insights of *Evidence-Based Medicine* are widely relevant across many different research areas. Its insights provide a helpful context within which to discuss the use of technological change to improve the research process. Systematic data-based overview has to date received inadequate attention, both in research and in research training. Sharing of research data once results are published would both assist systematic overview and allow further scrutiny where published analyses seem deficient. Deficiencies in data collection and published data analysis are surprisingly common.

Technologies that offer new perspectives on data collection and analysis include data warehousing, data mining, new approaches to data visualization and a variety of computing technologies that are in the tradition of knowledge engineering and machine learning. There is a large overlap of interest with statistics. Statistics is itself changing dramatically as a result of the interplay between theoretical development and the power of new computational tools. I comment briefly on other developing mathematical science application areas – notably molecular biology. The internet offers new possibilities for cooperation across institutional boundaries, for exchange of information between researchers, and for dissemination of research results.

Research training ought to equip students both to use their research skills in areas different from those in which they have been immediately trained, and to respond to the challenge of steadily more demanding standards. There should be an increased emphasis on training to work cooperatively.

"At bottom my critique is pretty simple-minded: Nobody pays much attention to the assumptions, and technology tends to overwhelm common sense."
[Freedman 1987.]

"I personally look forward to the proper balance that will emerge from the mixing of computational algorithm-oriented approaches characterizing the database and computer science communities with the powerful mathematical theories and methods for estimation developed in statistics."
[Fayyad 1998.]

"Statistics has been the most successful information science. Those who ignore statistics are condemned to re-invent it."
[Efron, quoted in Friedman 1997.]

"Some members of the profession are trying hard to make changes, by teaching courses in which substantive questions come first and technique is introduced to find answers. Of course, all too often, technique comes first; data come in as purely decorative illustrations – a practice not confined to statistics departments."
[Freedman 1991.]

Contents

Abstract iii

Preface 1

1. Changing Research Demands 2

Language Issues 2

2. Research Standards 3

Data Collection and Analysis Standards 3

Faults in Published Papers 4

Reasons for Inadequacies 5

Openings for Improvement 5

The Relevance of Information Technologies 6

3. Data-Based Overview 7

The Demand for Data-Based Overview (Systematic Overview) 7

Systematic Overview in Medicine 7

The Neglect of Data Overview 8

Data-Based Overview – Examples and Further Comment 9

Data as a Resource 11

Data Warehousing 12

The Challenge of Large Data Bases 12

The Importance of Data Sharing 12

4. Issues for the Use of Observational Data 13

Historical Experience with the Use of Observational Data 13

Qualitative Research 14

5. Data Mining (Knowledge Discovery in Databases) 14

Commercial Data Mining Packages 15

Mining Business Databases 16

Issues of Size 16

A Computing Perspective on Data Mining 16

Data Mining Tools 17

Building on Statistical Insights 17

False Confessions 18

Re-inventing Statistics 18

The Design of Data Collection 19

Decision Trees, Artificial Neural Networks (Neural Nets), and Related Methods 19

6. Specialised Applications of the Mathematical Sciences 21

Genomics and Bioinformatics 21

Biological Computing – Further Issues 22

Other Mathematical Science Application Areas 22

Cost-Benefit and Cost-Effectiveness Analysis 23

7. New Training Demands for Research Students 23

Research Protocol Design 24

Statistical Consulting Unit Courses 24

Training for Co-operative Research 25

Internet-Based Searches 25

Knowledge Engineering and Machine Learning 26

Bioinformatics (Biological Computing) 26

8. Summary of Main Points 27

9. Appendix 29

Definitions 29

Web Sites 31

10. References 33

Preface

Technological changes offer the research community unprecedented challenges and opportunities. My focus is on changes that affect the collection, organization, analysis and presentation of data. Concluding comments explore implications for graduate research training. Influences from my own contact with research, in horticulture and agriculture, in entomology, in industry, in the humanities and more recently in medicine, are obvious.

I note common present deficiencies in the conduct and reporting of research, but do not attempt to quantify the extent of these deficiencies. The needed data are not available. Published systematic critical scrutiny of all papers which have appeared in a journal or group of journals is unusual. There are a few instances where applied statisticians have conducted overviews of the statistical analysis and presentation in one or another area, but most of these fall well short of systematic regular overview. The scientific community has been slow to apply systematic scientific processes to the scrutiny of its research.

Useful information on the career progress of US PhD students will come from a book, to be published in the next year or two, by Dean Joseph Cerny and Research Director Maresi Nerad from the Graduate Division of the University of California at Berkeley. There is a dearth of quantitative information on Australian PhD graduates and on their later views of their training experience, though this should soon change.

My purpose is both to inform and to stimulate discussion. I will be glad to receive comment and/or pointers to relevant information which I may have missed. I hope, at some later point, to revise this paper in the light of discussion and comment.

I am grateful to Dr John Hooper (Dean, ANU Graduate School) and to Christine Donnelly (Statistical Consulting Unit, ANU Graduate School) for drawing my attention to errors and ambiguities in successive drafts of this paper, and for helpful discussion. I am solely responsible for any remaining errors.

1. Changing Research Demands

An examination of changes of the past two decades gives clues on how research demands may change in the next two decades. We should expect changes of a similar or larger magnitude. Some changes will arise from the attempt to fix problems with our current approaches. Some will be driven by technological advance. Some will be demand driven. Demands to improve the quality and coherence of research are an appropriate point of departure both for examining the place of new information technologies and for considering new training demands.

Features of the present rapidly changing research scene are the demand to gather better data, to gather more comprehensive data, to organize data better, and to make better use of the data we then have. Computing has from its beginnings carried with it the promise to address these concerns. Hence the emergence of computer-oriented departments, societies and journals devoted to *Information Technology*, *Informatics*, *Machine Learning*, *Data Mining*, *Knowledge Discovery in Databases* (KDD), and so on. Hence also the emergence of the data-based approach to the use of research knowledge in medicine that is captured by the name *Evidence-Based Medicine*. Computing is an important tool for evidence-based medicine, but not a driving force.

Information technology is, on the one hand, a general term for technologies for collecting, processing, organizing, interpreting and presenting data. There is also a narrow usage, that has specific reference to computing technologies. The range of usage of the word *informatics*, which is an English rendering of the French *informatique* or the German *Informatik*, is similar.

The methods and insights of Evidence-Based Medicine have potential far-reaching implications for research, both in medicine and in other areas. Unless such insights are taken on board, increased use of computer technology may do little or nothing to enhance research. Its ideas and insights are a good point of departure for discussing how research approaches and research training ought to change. A key issue for Evidence-Based Medicine is, to put it bluntly, to distinguish the data and associated interpretation that merit attention from crap.

Data collection is driven by an understanding of what is worth collecting. In this sense scientific research is driven by theory. It is the genius of science that data may carry within themselves the power to challenge and ultimately destroy the theory which guided their collection. Statistical insights and approaches have a key role in any attempt to extract information from data. Their use is for teasing out pattern, and for distinguishing what may be real from artefacts of the analysis.

Statistics, and related information technologies which make data their focus, have wide application across all research areas. There are in addition methods and approaches that are specific to individual research areas. In the present paper, the interest is in methods and approaches which, although perhaps developed initially to address the requirements of one particular research area, have wide general application. I will focus first on issues that arise from the examination of current published research.

Language Issues

We will encounter a large amount of new terminology. It is hard to tease out what really is new, and what is a new name for what is already known. However annoying the continual addition of new jargon, we cannot ignore it.

Often “Debates are won by those who control the definitions . . .”.[1] They may be won by those whose language best captures a prevailing mood. Apt words will be important in

[1] R. K. Webb, in *Enlightenment and Religion*, ed. Knud Haakonssen, Cambridge 1996, p.310.

convincing a new generation of students that we have thought long and hard about what we ought to teach them. It may help to persuade them that we are active and open-minded participants in the continuing debate about academic priorities.

Fine language must have substance behind it, if we are to improve the quality of research and continue to attract students. Whatever the language used, the content must be driven by what we can justify as genuinely important. We need a sense of deficiencies in current research methodology, a sense of how our research approaches might take advantage of new technologies, a sense of where leading edge research is headed, and a sense of the subject area skills which should underpin that research.

2. Research Standards

Pressures to publish, the vast extent of the refereeing task, and the increasing fragmentation of research between different and largely separate research traditions, have made it difficult to maintain standards over the whole range of technologies that contribute to effective research. The specialist technical demands of individual disciplines – molecular biology, electronic engineering, physiology, biochemistry, etc., continue to advance. These standards are reflected in the high standards which journals demand in the specialist disciplines which they represent.

Standards may not be maintained when authors stray into areas which require expertise that is outside of the specialist discipline or disciplines. Bureaucrats who insist on using the number of pages of published output as a measure of academic performance ought to consider what signals this sends to researchers, to referees, and to journals. Morrison (1998) has an interesting commentary on other deleterious effects of the DEETYA publication rewards system.

Data Collection and Analysis Standards

My reading of published papers, and some published surveys, persuades me that serious problems with the design of data collection and with data analysis are common. A cursory overview of papers in major international journals may be sufficient to reveal examples of serious statistical misinterpretation. I cite examples in a later section. In view of the prevalence of amateur do-it-yourself analyses, there is a case for treating all published analyses as preliminary, pending scrutiny by researchers with relevant statistical science skills! I would expect to find problems with other specialist skills also. To an alarming extent, researchers who have been trained in one area are relying on their own inadequate resources for research which demands skills which they lack.

Concern at the quality of the statistical support for published research has been especially evident in the medical literature (Andersen 1990; Altman 1994). Design, which lays the foundation for everything which follows, is even more important than analysis. This is the point of a hard-hitting article (Chalmers and Grant 1996) which draws lessons from the results of the collaborative eclampsia trial published in 1995. The evidence of this latest trial is that the introduction of diazepam in 1968 is likely to have led to an increase in death rates of women suffering from eclampsia, from 1 in 11 to 1 in 9. Chalmers and Grant comment:

> Today’s report is a triumph for the trialists, but what a scandal that we had to wait 70 years for the answer. … For seventy years, the proponents of various drugs and drug cocktails have hurled disdainful abuse at each other from separate mountain tops, secure in the knowledge that no strong evidence existed that could undermine any one of their multitude of conflicting opinions. …
>
> During a total period of just over five years, far more has been achieved through the collective efforts of 27 centres in 9 developing countries (some that had little or no previous research experience) than has been accomplished during more than half a century of small-scale, poorly controlled, individualistically driven investigative tinkering by others, including many people in the developed world

> who believe they deserve to be regarded as serious investigators. This lack of scientific and professional self-discipline in the developed world, particularly the unwillingness to collaborate in studies of sufficient size, has had substantial human costs.

There is no evidence that other areas are generally better than clinical epidemiology. On the contrary, medical experimentation typically faces greater statistical and other scrutiny than is common elsewhere.

Faults in Published Papers

Recent examples of errors or faults in data collection or analysis in published papers are summarized below. Further examples will be presented later in this paper. Items 1-4 came to my attention in the course of my work, rather than as the result of a systematic search:

1. Much literature in biological anthropology uses secondary data without careful evaluation of sources. In a comparative study of nuclear and mitochondrial genome diversity in humans and chimpanzees that appeared in *Molecular Biology and Evolution* in 1997, the statement is made that "the human material was from various sources." Was it obtained from whoever was available in the laboratory? The source of the chimpanzee material is better documented, though without providing evidence that the animal sources used could be treated as a random sample from an identifiable population. This low tech approach to the design of data collection contrasts starkly with the high tech molecular biology and statistical classification methodology used. For purposes of making a comparative assessment of variability, this Achilles heel can fatally flaw the study.

2. Articles in the *Journal of Economic Entomology* which demonstrate statistical misunderstanding are relatively common. Thomas and Mangan (1997) is something of a record in this respect. They claim that sample size affects the tolerance distribution, i. e. the variation with insecticide dose of the proportion of insects expected to die. It does not. One incorrect formula is correctly derived from another incorrect formula. They suggest, quite wrongly, that a lognormal tolerance distribution makes it appropriate to assume a logit or complementary log-log distribution. They misrepresent results that are presented in authors from whom they quote. They claim to use the Maentel-Haenzel test for testing goodness of fit to an assumed tolerance distribution, an entirely inappropriate use. This is an incomplete list.

3. Papers are common in the Journal of Economic Entomology which fit probit models, even though a plot will make it clear that the data do not at all fit this model. Jessup and Baheer (1990), for example, estimate that their treatment gives a 1 in 300,000 insect survival after 29.4 days, which is blatantly at odds with the proposed 12 day commercial treatment. A plot makes it clear that their statistical model is grossly at odds with the data.

   This widespread use of inappropriate models invalidates inter-region comparisons of estimates of 99% and other lethal mortality points. Published results make it impossible to determine whether the response of the fruit fly *Bactrocera tryoni* to heat treatment is the same in Queensland as in New Caledonia. Where the model is seriously wrong, estimates have a bias that depends on the choice of doses and the allocation of insects between doses. Where two researchers have used the same wrong model, there may be large differences in the bias.

4. Using data from 1910-1992, Nicholls et al. (1996) estimate a straight line relationship between the Southern Oscillation Index (SOI) and all-Australian rainfall. A smooth trend line suggests little or no relationship for values of the SOI less than zero. For values greater than around zero, there does seem to be evidence for higher rainfall at high SOI values. Irrespective of what may be the correct form of response, the statistical analysis which Nicholls et al. present is wrong, and almost certainly

optimistic. It does not allow for sequential dependence in the data. This failure to allow for sequential dependence is a relatively common fault.

5. Further examples of errors and faults may be found in Andersen (1990), Bryan-Jones and Finney (1983), Gartland (1988), Maindonald and Cox (1984), Maindonald (1992) and McCance (1995),

Reasons for Inadequacies

The twin ideas that experimental design is a pushover, and that statistical analysis will soon be reduced to button pressing, do not die easily. Spin doctors who know enough statistics to be dangerous hail each new major improvement to statistical software, if not as the fulfilment of this dream, then as very close to it. Some popular expositors of data mining are once again making the same absurd claims. They encourage amateurs to attempt to turn data analysis into a button-pushing exercise, with results such as I have noted.

Researchers may attempt, without consultation with expert practitioners, design of data collection and statistical analysis tasks for which they are not qualified. They may struggle with older inadequate methods that are now superseded. They may have not updated their knowledge of statistical methods in line with new knowledge, and in line with new heavily computer dependent approaches. It is as though a molecular biologist were to limit him/herself to methods that were in vogue five years ago! The problem is not at all confined to users of statistics – professionals find it difficult to keep their skills up to date. Retraining of statistical specialists is an important issue for the profession.

A contributory factor is that some of the widely used statistical packages have been slow to change. Some areas – psychology and perhaps education – have developed their own statistical traditions, too much cut off from the statistical mainstream and slow to feel the influence of theoretical advances. Gigerenzer et al. (1989) give a fascinating historical account of the origins of these separate traditions.

The mathematics underlying modern methods is often complicated. Users do not, typically, need to trouble themselves about this complication, as the computer takes care of it. The end result may be much closer to the way that users find it natural to think about their problems. Often results are well summarized in a few well-chosen graphs. In addition new computing technologies, to an extent separate from those already identified, are influencing research methodology and opening up new research directions.

Openings for Improvement

Funding bodies must continue to look for ways to get better value from the research dollar. Inevitably, there will be changes. It is salutary to look at changes in those areas (such as clinical epidemiology) where in the past twenty years there have been large advances in approaches to the design of data collection and data analysis, and to consider the implications of these changes for other research. These trends can then be extrapolated a limited distance into the future. This is the motivation for the following list.

1. Requirements to place data and supporting documentation in the public domain, inviting scrutiny and facilitating incorporation, where this is pertinent, into overview studies.
2. Widespread development of mandatory reporting standards comparable to those set out for randomized controlled trials in the CONSORT statement (Begg et al. 1996).
3. Identification of skill gaps which compromise research that otherwise demonstrates high levels of technical skill.
4. In medicine, health and education, the development of mechanisms that will replace multiple small trials by large carefully co-ordinated multi-centre trials.

5. Except where they break radically new ground or where experimental approaches are impossible for ethical or other practical reasons, increased reluctance to fund non-experimental studies.
6. Demands to present cost-benefit or other economic analyses.
7. Insistence, where relevant, that researchers spell out the practical implications of their research. (For example, what are the negatives – false positives and their associated trauma, etc. – that should be set against the ~2 in 1000 reduced deaths from breast cancer screening of women in their 50s and 60s?)
8. Insistence, where appropriate, that researchers use qualitative and quantitative approaches to complement each other.

Items 1 and 2 could be implemented without making any substantial change to refereeing processes. As there are existing models, these are the changes that seem most immediately likely. Moves to collect data into data bases that operate as commercial entities (Transborder 1998) will to a greater or lesser extent work in the other direction, restricting access to data.

Attention to potential skill gaps in ancillary disciplines is more than ever important as we respond to the seductions of new technology - molecular biology, informatics, machine learning, data mining, and so on. The use of a new technology, or of an old technology under a new name, must not be a new opening to dispense with the complementary disciplines needed for an effective study.

The Relevance of Information Technologies

The relevance of information technologies to the above discussion includes the following:

1. There have been huge advances in the methodology for data analysis, taking advantage of the advances in computing hardware and software. There are large differences between statistical packages in the extent to which they have taken up these methodological advances.
2. There has been a large emphasis on methods for *Exploratory Data Analysis*. As we will see, this has a large overlap with the methods and approaches of Data Mining.
3. There have been substantial advances, again taking advantage of the increased power of computing systems, in the methodology for analyzing data from overview studies.
4. Database technology, already a powerful tool for storing and accessing data, has extended to the networking of physically separate databases.
5. There are new data analysis challenges that arise from the sheer size of some databases.
6. There is a new emphasis on the key role of data aggregation for the scientific enterprise. Statisticians, while rightly emphasizing the hazards of making inferences from data that is from disparate sources, may have been slow to take this new emphasis on board.
7. The computer science perspective seems appropriate for attempts at automating the task of making data based inferences. To date, these attempts have had very limited success.

Overall, we see a growth in methodologies which require a pooling of the skills of computer scientists, statisticians and subject area specialists.

3. Data-Based Overview

The Demand for Data-Based Overview (Systematic Overview)

There are, in many areas, serious deficiencies of overview. Again, developments in medicine hint at changes that are needed in other areas. Deficiencies of overview are serious in two areas in which I have worked – disinfestation research, and post-harvest horticultural research. One looks in vain for papers which provide a data-based summary of, for example, major aspects of kiwifruit research, or the use of methyl bromide for codling moth disinfestation.

Data-based overview places the individual studies under critical scrutiny, and places them in context. In a recent review of yield-density studies on green asparagus, Bussell et al. (1998) found large differences within the same locality. Based on commercial experience, it is likely that fertilizer and soil effects, and variety, were the main factors explaining yield differences between trials. Information on relevant factors was so incomplete that it was impossible to draw from the trials themselves any certain inference on factors affecting yield. Two only of the 15 trials gave any information on climate, irrigation and terrain. Four trials gave no information on soil type. The trials give benchmarks against which growers in a local region can compare their own yields. This aside, none of the recent trials have added anything of consequence to what commercial growers already knew – use a modern variety on a sandy or light silt loam soil, plant at the highest density that is practical, and use a fertilizer that is at least as effective as farmyard manure!

Where data-based overview is taken seriously, serious obstacles arise from the ‘file drawer’ problem. Results from a proportion of research studies do not find their way through to publication; they remain in the file drawer. It may then be difficult or impossible to identify all relevant studies. For those studies which are identified, it may be difficult or impossible to get access to raw data. In such areas as clinical medicine, an insistence on some form of international registration of trials at the time of commencement seems desirable. This would ensure that all trials relevant to a particular overview study can be later identified.

Systematic Overview in Medicine

Clinical Epidemiology and related areas of medicine have pioneered approaches to systematic overview and to the summarization of evidence that are useful models for other areas. Systematic Overview is a key methodology for the conduct of studies such as are fostered by the Cochrane Collaboration (Sackett and Oxman 1994), and for Evidence-Based Medicine (Sackett at al. 1997; ScHARR 1998; Moynihan 1998, pp. 213-241). Smith (1996) asks how an ‘evidence-based’ human society would conduct its business. Cochrane type evidence bases are required in many other areas than medicine.

Lessons from experience with medical databases are highly relevant to efforts now under way to collect other types of data, often from disparate sources, in large databases. Draper et al. (1990) describe areas where data-based overview is important. An interesting application is to the estimation of physical constants. Data based overview seems especially important when the literature is extensive, uneven in quality and different biases may be associated with different types of study. Chambers and Altman (1995) should be consulted for an account of systematic overview in medicine.

The advice and insights of evidence-based medicine are in the first instance directed towards researchers. The publisher’s blurb for the journal *Evidence-Based Medicine*, directed to clinicians, argues:

> With 2 million new papers published each year how can you be sure you read all the papers essential for your daily practice, and how can you be sure of the scientific soundness of what you do read?

Researchers have the same interest as clinicians in getting a sense of the conclusions which ought to be drawn from studies to date, as a starting point for their own research. Systematic overview identifies secure knowledge and highlights gaps in research-based knowledge. A particular widespread gap in clinical medicine is in evidence that would assist in tailoring treatments to the special requirements of individual patients. Some papers may have no information on a key covariate, e. g. baseline blood plasma zinc levels in a zinc supplementation trial. Too many papers focus on single end-points where the interest should be in the response profile, i. e. in the pattern of response over time.

There may be several overview studies from which to choose. Just as some papers are so flawed that they merit scant attention, so also for overview studies. Advice and training is needed that will help discriminate the good from the bad. Sackett et al.(1997) and Greenhalgh (1997) emphasize this point, and give advice on the critique of overview studies. If no up to date and clearly authoritative study is available, the researcher's first step must be to attempt his or her own overview.

The demands of data-based overview studies which meet Cochrane Collaboration standards are severe. It is typically easier to do a new study than to undertake a fully adequate overview of existing studies. The technical demands are such that Cochrane Collaboration studies have so far covered only a small proportion of health care. The conduct of overview studies requires special skills that are different from or additional to those of subject area experts. There is evidence that subject area experts do a poorer job than non-experts with experience and skills in the conduct of overview studies (Oxman and Güyatt 1983.)

At present the perspectives of evidence-based medicine, and the importance of Cochrane Collaboration type studies, seem not to be widely recognized outside of medicine. Pressures for change may come from three sources:

1. Researchers in e. g. psychology or education who work on the borderline of clinical medicine may get direct exposure to the ideas and insights of evidence-based medicine.
2. Funding bodies may demand evidence that researchers are following an 'evidence-based' approach.
3. The logic of this general approach to marshaling research evidence is compelling.

Kuhn (1970) and others have argued that research traditions change only when the pressures for change are overwhelming. The inherent logic of the approaches of evidence-based medicine and of the Cochrane Collaboration studies will not, on its own, be enough to bring about widespread adoption of these ideas and insights. Experts whose authority relies on the use of more traditional informal means for assessing the weight of evidence may feel their authority threatened.

The Neglect of Data Overview

There are many reasons for the past relative neglect of data overview issues. An adequate statistical theory was slow to develop. For a long time there was more than adequate challenge to theoretical skills from developing a theory that would handle data from an individual field site or from an individual clinical trial. There are severe problems in deciding how to weight the separate sources of evidence. Scientists have often been protective of their experiments and their data, which they may believe should stand on their own independently of what other scientists may have been doing. The tradition of analysing separately data from each field experiment or each trial became firmly established. It remains firmly entrenched in horticulture, and in other research areas also. Experimenters who have worked on different sites may each claim the other is 'wrong', where it is unclear whether the difference is a geographical effect, or perhaps due to differences in experimental procedure. In these circumstances researchers who belong to the same research tradition may "hurl disdainful abuse at each other from separate mountain tops."

Data-Based Overview – Examples and Further Comment

1. *Science* recently (Taubes, 1998) carried an article on *The (Political) Science of Salt*. An over-riding issue in this debate has been the relative weight that should be placed on different sources of evidence. I have no doubt that the proper way to make sense of this debate is to place the different studies in a hierarchy:

- Randomised Controlled Trials that followed strict protocols
- Other clinical trials
- Intra-population studies
- Inter-population 'ecologic' studies; migration studies.

It is now widely accepted that the soundest evidence is from randomised controlled trials that follow strict protocols. Such trials, on diet more generally as well as on salt, are now providing insight on the superficially contradictory results that have been obtained from other types of studies. As often, one has to sift out the more directly relevant and reliable sources of information, and use them to interpret less reliable and/or relevant sources of information.

Taubes' article highlights other key points. He discusses the 1988 Intersalt study of the relationship between blood pressure and salt consumption, including both inter- and intra-population components. A key problem for the debate which followed publication of Intersalt was that other researchers were not allowed access to the data.

There are numerous instances where the relative weighting of different sources of evidence and the pooling of evidence were key issues – many of them modern re-runs of the discovery that blood-letting, so far from making you better, is actually dangerous. A recent Cochrane meta-analysis indicates that giving albumin to patients in critical illness increases the risk of death, by around 1 death for every 14 critically ill patients[2] who receive albumin. (Cochrane Injuries Group Albumin Reviewers 1998.)

2. Many of the agricultural fertilizer trials that were conducted in New Zealand over several decades prior to the 1980s were for a long time not analyzed. Not until the 1980s did a series of papers appear in the *New Zealand Journal of Agricultural Research* that provided the first careful overall quantitative evaluation of evident major effects. They highlighted areas which had been over-researched, and identified remaining gaps. There was an inevitable and implicit criticism of individual trials. Nowadays, a reasonable expectation is that such data will feed into a fertilizer database, with data analyses regularly updated to take account of data from new trials.

3. McGuinness (1997) provides evidence of several different competing schools of thought, each convinced it is right, on the teaching of reading. This may be an area where theory has grown like a weed, too little constrained by data from experiments that follow strict protocols such as are now demanded for medical clinical trials. The book is a careful overview of the current evidence, though perhaps overstating the case for her own approach. She rightly criticises the quality of much reading research, to the extent that there has been no direct comparison with competing approaches or that claims have been based on loaded comparisons that have not used appropriate controls. McGuinness's account has many of the elements of the thorough data-based overview that is required.

McGuinness uses research evidence to identify a range of sub-tasks which must all be mastered if children are to learn to read. There is an inexorable logic to the approach which she defends, which includes tests for identifying failure in any sub-task. A key insight is that children should be able to identify the 43 or 44 sounds of spoken English before learning letters or letter combinations which represent these sounds.

[2] The 95% confidence interval was 9 – 32.

> The attempt to work in the other direction, from letter combinations to sounds, introduces too many complications. The theory which she develops seems compelling, because it seems relatively complete and is backed up at key points by research evidence. She presents research evidence which shows that her methods work.
>
> While I find her arguments persuasive, I would like to see an independent critical evaluation. This might lead to one of two conclusions: either present data do indeed back up her claims, or else the jury is still out. If the jury really is still out, the only proper basis for judging between rival claims is a well-conducted randomised controlled trial. Indeed this is the only ethical way to advance educational practice.

Data will enter in other ways into other forms of research synthesis. In his book *Guns, Germs and Steel* Diamond (1997) seeks to explain striking differences between the long-term histories of peoples on different continents and islands in the past 13,000 years. The book is in a sense a sequence of data-based overview studies that are welded together into a brilliant continuous narrative. The data that he quotes are broad brush – numbers of plant and animal species domesticated in different geographical locations, differences in land area and population size, differences in between continents in the diffusion rates of crops and artefacts that seem a result of their different geography, one-sidedness in the transfer of diseases between Europe and the Americas, and a variety of archaeological and phylogenetic data. He limits attention to data which seem to have a clear and relatively unequivocal story to tell. As is inevitable in a book that is intended for a wide audience, the casual reader must largely take Diamond’s facts and figures on faith, accepting that they are adequately accurate for his purpose. Specialist readers will wish to refer back to his sources.

Particularly relevant to my discussion is Diamond’s last chapter, on “The Future of Human History as a Science”. Diamond proposes a research programme that would gather quantitative information intended to test his major claims, and that would provide more accurate quantitative estimates of e. g. the different diffusion rates of crops, artefacts, etc., in the different continents. Diamond’s research synthesis sets the scene for an ongoing research programme. This leads into a wider-ranging discussion of ‘historical science’. There is an overlap of interest with the historical content of astronomy, climatology, earth science and evolutionary biology. A view that sees history as a series of ‘natural experiments’ can be illuminating and insightful.

This discussion might continue, taking me well away from my central themes. Imaginative reconstruction and synthesis readily gets out of hand. Hence the importance of using all available data-based reality checks that we can summon to our aid. Why do I consider that Diamond is broadly right, but reject the elaborate imaginative historical reconstructions of Immanuel Veliskovsky, which Sagan (1979) dissects?

Data as a Resource

A recent study (Transborder 1997) identifies two conflicting trends:

1. The need for scientists to adapt to conducting research with data that come in rapidly increasing quantities, varieties, and modes of dissemination, frequently for purposes far more disciplinary than in the past; and
2. The worldwide trend toward imposition of increasing economic and legal restrictions on access to scientific data gained from publicly funded research.

It seems inevitable that public funding bodies and journals will increasingly insist on putting all data into an archive where it will be available, if necessary under some restricted form of access, to other researchers. This is already a requirement for research funded by the U. K. Economic and Social Research Council (ESRC). It seems inevitable that other funding bodies, seeking increased value from the research dollar, must follow suit. A common heavy reliance on secondary data makes such data archives highly important for social scientists.

Privatization, and pressures on public institutions to recover costs, have in many countries increasingly led to charges for services that were previously free, including the generation and distribution of scientific data. Thus agreements on the free interchange of a certain minimum level of meteorological data have been accompanied by severe restrictions on interchange beyond this level (Transborder 1997).

A major study that had U. S. government support gives the general guideline:

> The value of data lies in their use. Full and open access to scientific data should be adopted as the international norm for the exchange of scientific data derived from publicly funded research. The public-good interests in the full and open access to and use of scientific data need to be balanced against legitimate concerns for the protection of national security, individual privacy, and intellectual property.
> [Transborder 1997.]

There may be far-reaching changes for academic researchers in the United States under a new provision which will require Federal awarding agencies "to ensure that all data produced under an award will be made available to the public through the procedures established under the Freedom of Information act." (Kaiser 1998.) The background to this change was industry concern to get access to university research data on health effects of air pollution. There has been concern that the provision will open up new ways for industry groups to try to discredit research results which they do not like, or even to harass researchers. Presumably, however, environmental groups as well as industry groups will gain the same new opportunities for access to data. There seems wide acceptance that processes for data sharing are in principle desirable, but tempered with concern at the perceived absence of filtering processes. This concern may be unwarranted (Gough 1998), though it will be necessary to address issues of proprietary data and privacy. Clearly there will be some initial discomfort. Against this are the substantial long term benefits which Goldwater (1998) expects from processes for data sharing. Moves to charge for access to data will work in the other direction, creating obstacles to ready access to data.

Appendix C of Transborder (1997) lists a number of successful examples of international exchange and management of data in the natural sciences – nuclear structure, high-energy physics, chemical sciences, genomic sequence and related data, Hubble space telescope, seismic, and meteorological data. The Hubble Space Telescope Archive, operated under a memorandum of understanding between NASA and the European Space Agency, may be a useful model for other areas. All scientific data are archived, kept proprietary to the contributing astronomer for one year, and after that time become available to other astronomers.

Data Warehousing

Data warehousing provides a single point of access, perhaps through networked databases, to data that share a common theme. Data may be from museum collections, DNA studies, surveys, censuses, clinical trials, meteorological records, astronomy, geophysics, laboratory and field experiments, etc. In many areas, it will be important for researchers to learn to access and use such material. There may be virtue in learning no more than it is immediately necessary to know. Next year's system may have a radically improved user interface.

A key part of the data warehousing task is to ensure that data are accurate and consistent, i. e. to develop and implement a single data model for the whole database. Even within a single institution databases from different departments may define data fields differently, or use different units – perhaps pounds instead of kilograms. The use of fields and the structure of the database must be documented as part of the database itself. As far as possible, errors must be identified and removed. It is important to retain contextual information which may later be crucial to data use and interpretation. If these issues are not handled when data are warehoused, they will be an obstacle to later use of the data. Inevitably some errors and gross deficiencies in background information will come to light only when the attempt is made to analyse the data. Data cleaning, and contextual documentation, is a huge challenge to warehousers.

The Challenge of Large Data Bases

Large data bases offer new data retrieval and data analysis challenges (Wegman 1995.) This has become an active and important area for research. Through the Cooperative Research Centre for Advanced Computational Systems, ANU already has a stake in work on the management and analysis of large data sets. The other participants are CSIRO and the Australian arms of three computer companies. There is an On Line Data Archives Program and a Data Mining program. The Advanced Server Technologies Program has as one of its components "techniques for managing large information repositories and thematic information retrieval". The Virtual Environments Program has a data analysis component. There is one further program – Digital Media Libraries.

The Importance of Data Sharing

It has been suggested (Denise Lievesley, pers. comm.) that the archiving of data along with the research design and perhaps a preliminary descriptive analysis might of itself qualify as a publication. In principle, this is a move that I would support. It is clear from the published literature that cases are frequent in which researchers do not have the skills needed to do an adequate job of analysing their data. Handling the analysis at arms length from the data would encourage objective evaluation of the design of data collection and of the quality of the data. Under current practice, authors have an incentive to gloss over evident deficiencies in their data. The matter is not however straightforward, because of the frequent interplay between analysis and a frequent demand for information which was not initially provided to the analyst.

Access to the data makes it possible for others to check analyses which supported the published results. The data become available to researchers who may want to use the information in planning their own study, or who may want to include it in a wider overview study. It will at the same time expose studies, potentially, to continuing data-based critical review. Publication may become a starting point for wide-ranging critical evaluation, not a supposed final imprimatur on the research.

Demands to make data available for archiving are one aspect of the extent to which supervisors must expect that the research environment in which their students work will be more demanding than the environment they have themselves experienced. There are implications for the training of research students. (See section 7). Any attempt to aggregate data that have a common theme seems welcome, in spite of the potential for abuse. The value of individual studies on a topic lies in what they contribute to the total picture.

Careful examination of existing data can be extraordinarily useful, as a preliminary to undertaking one's own study. Where earlier data are available from an archive, this becomes relatively straightforward. Such data can be invaluable for the indications they provide on how the new study should be designed.

4. Issues for the Use of Observational Data

Historical Experience with the Use of Observational Data

In the early 1980s, there was extensive debate between clinical researchers who felt that databases containing largely observational data had an important role in the evaluation of new therapies, and those who felt that the main reliance should be on randomized controlled trials. Jorgensen and Gentleman (1998) give a number of references; see for example Feinstein (1984) and Green and Byar (1984). There is now wide agreement that while observational databases have their uses, e. g. in drawing attention to side effects, they are an unreliable and potentially misleading source of evidence for deciding between alternative therapies.

There will be comparable lessons for the use of databases in other areas of knowledge. After some initial large gains, perhaps equivalent to the early use of observational evidence to establish a strong presumption that smoking causes lung cancer, subsequent gains may be small and hard won. Lessons learned in the context of clinical databases will have to be re-learned in other contexts.

Even if observational data have been collected with extraordinary care, and the data sources are fully documented, attempts to use the significance of regression coefficients to argue for causation are, without other supporting evidence, inevitably flawed. An exception is where it is possible to argue that the data are 'quasi-experimental', i. e. they have the characteristics of experimental data. Snow's (1855) study of the causes of cholera outbreaks is a famous example.

The attempt to derive causation from statistical evidence of association commits the 'ecologic' fallacy. There are many reasons for this failure, the most fundamental of which is that causation may go in either direction or may arise from association with a third variable that may not even have been measured. There may be temporal or spatial correlation, or clustering effects. There are data quality issues, especially likely to be serious when the details of data collection are not known. Finally, many authors make technical mistakes in their use of the methodology, such as not allowing for variable selection bias. All these problems are apparent in a Kanarek et al. (1984) paper which argued, wrongly if the evidence in this paper is the only basis for the claim, that asbestos fibres in the drinking water cause lung cancer. Freedman (1991) uses this paper in drawing attention to the gross over-optimism with which many sociologists have used regression techniques. As Freedman (1987) says:

> At bottom my critique is pretty simple-minded: Nobody pays much attention to the assumptions, and technology tends to overwhelm common sense.

The problem is fundamental, and no technological statistical or computing fix is possible. Causal inference is a difficult and challenging area in which to work. It is more difficult where experiments are impossible. See also Fraker and Maynard (1987), who made a comparison between the use of non-experimental data to compare manpower programs, and results from experiments that compared the same programs. Path analysis and Linear Structural Relations have even more serious problems than regression. Bartholomew (1995), who is sympathetic to the aims of social science users of these models, has a brief judicious summary.

Freedman argues that any case for causative effects must build on many sources of evidence, in which a careful use of regression modelling may have a small part. This makes the research task demanding. Rosenberg (1968) is a good basic text that treats many of the important issues, in a way that stays close to the data.

For just these same reasons, information from existing data bases will rarely be as useful as evidence that is directly collected in order to answer a question that is of interest. This restricts the sorts of questions which data mining, to which we will turn in the next section, can address. It may be able to tell us that men who buy beer on a Friday night commonly also buy disposable nappies. It cannot tell us, without other supporting knowledge, how to increase the sales of either product.

Qualitative Research

The current enthusiasm for qualitative research may in part be a response to the perceived futility of inadequate quantitative research. Quantitative and qualitative research should be complementary, working in tandem to build a convincing case, a point which has had scant attention in the literature.

Quantitative research must start from qualitative judgments on what research questions are to be pursued, and may rely on qualitative insights for the interpretation of results. It may be even easier to do poor qualitative research, perhaps largely reflecting the researcher’s own prejudices, than to do poor quantitative research. Good qualitative research is every bit as demanding as good quantitative research, and should not be seen as an easy alternative which avoids the heavy demands of quantitative research. Greenhalgh (1997, chapter 11) has a useful brief discussion, and gives references which interested readers can pursue further. While acknowledging the importance of qualitative research, she is critical of much of what has appeared under this name.

Note that researchers outside of social science may use other names for approaches that have strong connections with qualitative research. Thus Scholtes (1988), who does mix quantitative and ‘qualitative’ approaches, discusses at length methods for generating and honing ideas.

5. Data Mining (Knowledge Discovery in Databases)

Data are a valuable resource. As such, perhaps one can *mine* the resource for its nuggets of gold. In part the interest in *data mining* has been driven by individuals and organizations who find themselves with large data holdings, which they feel ought to be sources of valuable information. They may have little idea what to do with them. The interest has been fanned by hardware and software computer vendors who are looking for new market niches.

There is no firm distinction between data mining and statistics. Much commercial data mining activity uses relatively conventional statistical methods. A difference is that data miners may be working with quite huge data sets. Hence Friedman’s (1998) definition of data mining as the “computer automated exploratory data analysis of (usually) large complex data sets.” A data set with values of twenty or thirty variables for each of several hundred thousand records is, in the context of commercial data mining, small.

A simple example of ‘exploratory’ data mining is the use of medical practice variations as a starting point for questions about operating or prescribing practices. McPherson (1990) quotes standardised rates for hysterectomy that were six times as high in the United States as in Norway. Such a huge difference calls for investigation and comment. In a classical statistical sense, the data miner is looking for outliers. Detection of fraud, in large clinical trials, or in business records, provides another example. What sorts of unusual patterns might make closer scrutiny desirable?

This more exploratory form of data mining applies a search process to a data set, often a very large data set, and looks for interesting associations. While the data may initially have been collected to answer some primary question or questions, the expectation is that there will be other interesting and potentially useful information in the data. Most experienced statisticians have at some time encountered unexpected and interesting results when, as a prelude to the main analysis, they have set out to do a careful exploratory analysis. Is it possible to set up automatic processes that may bring such results to attention? Jorgensen and Gentleman (1998) cite examples of data sets where there is bound to be unmined interesting information – fisheries data collected by Australian and New Zealand agencies over a number of years, secondary information in databases on clinical trials, and databases of routinely collected business information.

Much of the focus of data mining research has been on ways to find views of data that highlight 'interesting' or unusual features – a search for what statisticians would call 'outliers'. Friedman (1997) lists a number of approaches. *Exploratory Data Analysis*, a name invented by John Tukey, is in the spirit of Fayyad's description of data mining. Research on data visualization is in this same tradition.

Some data mining approaches are fairly specific to individual research areas, such as astrophysics at one extreme or business data processing at the other. Students in those areas ought perhaps to gain a general familiarity. It is not clear to me that this work has yet produced results which ought to be widely taught, rather than to particular specialized groups of students.

Commercial Data Mining Packages

Modesty has not sat well with vendors of commercial data mining packages. Friedman (1997) suggests that as in gold rushes of the past, their interest is in "mining the miners". Data mining may be sold as an investment that is small *relative* to the cost of a large database system, one that will add value. Additionally a powerful sales argument is that without these new powerful tools businesses will suffer disadvantage by comparison with adventurous competitors, who it is suggested are already reaching out eagerly to seize the new data mining opportunities. Even if purchasers later feel cheated, they are unlikely to admit that they made a bad business decision. From the beginnings of electronic computing, ambitious salespeople have oversold new types of computer systems. It is well to recall the exaggerated hype that accompanied early work on artificial intelligence.

> Alas for AI [Artificial Intelligence], the funding came screaming in with lots of strings attached and unrealistic expectations, and the results were pitifully few. Most of the applications didn't work – for good reasons: they were hard problems and still are. It was essentially, in much of the AI community, hubris – arrogance about one's capabilities and potentials, which just failed. The systems did not do what they claimed. But remember, often it wasn't the scientists who were doing the claiming.
>
> ….
>
> One impediment is the perception by much of the computer and management culture that making something work is primarily a matter of getting the right specifications and interpreting them, so to speak – making a program that satisfies those specifications. In a very fundamental way that is just plain wrong.
> [Selfridge 1996.]

It is in this context that skeptical comments from the statistician Peter Huber seem relevant:

> I do not think that I am doing injustice to the present situation by contending that data mining is still a nearly empty hull, held in place by hot air, and serving as a place-holder for more substantive contents to come.
> [Huber, quoted in Fayyad 1998.]

Elder and Abbott (1998) compare the features of "leading data mining tools". Several tools that are primarily statistical systems, including S-PLUS (MathSoft 1997), appear in the comparison.

Mining Business Databases

There is a growing literature on experience with business data mining . Most of this reflects the experience of very large organizations. They have the most to gain from effective use of their large databases, and they can employ the specialists needed to handle analysis and interpretation. Commercial confidentiality prevents the publication of independent objective evaluations in refereed journals. From the commentary which does appear in the published literature, the following points emerge

- In building a data mining model a large proportion of the time – one ballpark estimate (Kelly, quoted in Fabris 1998) is 75% – goes on data validation.
- Efforts at data mining may draw attention to the poor quality of the data, perhaps to the extent that analysis is pointless. The Gartner Group estimates that only about 10% of collected data are analyzed. (Simoudis 1996.)
- Data mining teams need to complement business expertise with statistical expertise. (Fabris 1998.)

Attempts to mine collections of scientific data from disparate sources will encounter these same obstacles.

Issues of Size

Note that many physically large data sets have very limited information at the level of information that matters for the intended use of the data. A huge data base of information on the effects of hospital management practices may turn out to have information on a rather small number of hospitals. Each hospital is a different 'case'. There is an important distinction between those data sets that, using a realistic definition of case, really do have information on vast number of 'cases', and those which have information on only a small number of cases. Where the number of cases is modest the size of the data base is an issue only to the extent that it creates problems for summarizing the data for input to a conventional form of analysis.

The use of appropriate averaging to reduce the number of cases may be essential for the use of classical statistical methods with large data sets. There may be a dependence structure which classical approaches cannot model adequately. If Hampel (1998) is right, long term dependencies in space or time, of a kind that classical methods do not model well, are a pervasive feature of data sets which have very large numbers of observations.

A Computing Perspective on Data Mining

Fayyad argues that there may be a misunderstanding of the aims of data mining.

> Data mining is not about automating data analysis. Data mining is about making analysis more convenient, scaling analysis algorithms to large databases, and providing data owners with easy-to-use tools to help them navigate, visualize, summarize, and model data. …
>
> I personally look forward to the proper balance that will emerge from the mixing of computational algorithm-oriented approaches characterizing the database and computer science communities with the powerful mathematical theories and methods for estimation developed in statistics.
> [Fayyad 1998.]

These aims are modest, attainable, and far removed from the hype that has often surrounded data mining. The systems that Fayyad describes could be extraordinarily useful partners to statistical experts. They take on, at most, very limited aspects of the statistical expert's task. Friedman (1997) quotes other definitions from the data mining literature. For example:

> Data mining is a set of methods used in the knowledge discovery process to distinguish previously unknown relationships and patterns within data. [Ferruza].

Elsewhere Fayyad (1996) himself seems to enlarge the scope of data mining to include statistical analysis. He defines the primary tasks of data mining as classification, regression, clustering summarisation, dependency modelling, and change and deviation detection. Perhaps Fayyad has in mind a heavily automated use of these tools, as a preliminary to careful statistical examination of anything that seems to warrant more careful examination.

Data Mining Tools

Tree-based regression and *neural nets* have been widely promoted as data mining tools. Both these methods are beginning to attract interest from the statistical community. They are most commonly applied to discrimination problems, e. g. a bank may want to distinguish good from bad lending risks. Fayyad (1996) distinguishes *Knowledge Discovery in Data Bases* (KDD) from *Data Mining*. KDD, it is said, refers to the overall process of discovering useful knowledge from data, while *data mining* refers to the initial step of extracting patterns, without the additional steps designed to check whether these patterns are meaningful.

Friedman (1998) describes a Stanford University course which presents a broadly statistical perspective on data mining. The course is aimed at statistical and computing specialists, and provides a broad coverage of techniques that come both from the statistical and from the computing literature. Methodology is classified under the headings: decision tree induction (tree-based regression, etc.), rule induction, association rules (market basket analysis), clustering, and hot spot analysis. Hot spot analysis looks for subgroups of cases which show particularly strong patterns, e. g. banks will wish to identify mortgage holders who are at especially high risk of defaulting on their payments.

Building on Statistical Insights

There is now wide acceptance that progress in data mining will demand a merging of the insights of computing specialists with those of theoretical and applied statisticians. There are general points, relevant to all use of databases, which ought to be widely known and understood:

- Inferences can only be as good as the data allow.
- There is a potential for serious loss of information in the process of moving data from one medium to another, i. e. now, from paper to electronic storage.
- Data must be ‘good’ for their intended use.
- If you mix bad coin with good, the bad coin will ‘drive out’ the good. Mixing unreliable data in with highly accurate data, without discrimination, reduces all data to the level of their least reliable components.
- Analyses must have regard to data structure. As noted above, many physically large data sets are, from a statistical perspective, small!
- There are no royal routes to getting data to yield their insights, not yet anyway, and I expect not anytime soon.

Getting a relevant graphical view or views is usually essential to getting data to yield their insights (Cleveland 1993.)

In a chapter entitled “Reservations to Automatic Modelling in Statistics” Elder and Pregibon (1996) comment on the data mining pre-occupation with problems which have not yielded well to conventional statistical approaches. They warn that data miners must take on board statistical insights regarding the potential for spurious associations, selection bias, and issues of substance versus statistical significance. I find it surprising that they do not discuss at any length issues of data quality and structure.

Elder and Pregibon give a rough chronology of significant contributions to statistics, relevant to the KDD community, since the 1960s. As Elder and Pregibon explain:

> . . . this time period coincides with the significant increases in computing power and memory, powerful and expressive programming languages, and general accessibility to computing that has propelled us into the Information Age. In effect this started a slow but deliberate shift in the statistical community, whereby important influences and enablers were to come from computing rather than from mathematics.

Unfortunately SPSS, SAS and some other widely used statistical packages continue to reflect, in the style of statistical analysis which they promote, older approaches to statistical analysis which have not adequately taken on board the insights that have been stimulated by the computer revolution and by allied advances in mathematical theory. Actually SAS and to some extent SPSS mix new methodology with older approaches, in a manner that may thoroughly confuse the novice. A new SAS product – JMP – does reflect modern approaches much better. It has a strong linkage between statistical analysis and graphical presentation. Packages which lack this linkage should not be taken seriously.

Kolsky (1998), in a useful brief note on data mining, comments that "the same statistical issues that have plagued statisticians alike in their analysis efforts have not, in any way, been resolved by the use of Data Mining software." Edelstein (1998), quoted in a newspaper report, draws attention to data mining myths:

> . . . that data mining tools need no guidance; that data mining models explain behaviour; that data mining requires no data analysis skill; that it eliminates the need to understand your business and your data; and that data mining tools are different from traditional statistics.

Note however that data mining researchers, from both a computing and a statistics background, are actively developing new data mining tools. Some of these new tools will in due course become part of the stock-in-trade of professional statisticians.

False Confessions

Data mining, like all forms of exploratory analysis, is open to abuse. Under torture, the data may yield false confessions. Data mining readily becomes data dredging, a practice that well deserves its bad reputation. Classical inferential procedures may require substantial adaptation, or have little relevance, for data mining applications with large data sets.

Re-inventing Statistics

Efron (quoted in Friedman 1997) has argued that:

> Statistics has been the most successful information science.
> Those who ignore statistics are condemned to re-invent it.

Efron is perhaps saying that there are no good alternatives to perspectives which statistics offers on the design of data collection, on data analysis, and on modelling. Friedman argues that, because of the intensity and exclusiveness of its love affair with mathematics, statistics risks losing its pre-eminence as a theoretically based information science. Academic statistics must enlarge its purview to take in all aspects of data collection, data manipulation and data interpretation. Unless this happens other more strongly computer based disciplines will increasingly offer that broader perspective, taking what they need from statistics, and consigning academic statistics to a supporting role.

Even more serious may be the tendency, found both among statisticians and among computing oriented data miners, to foster a technique-driven approach to data analysis. Freedman (1991, p. 357), responding to Mason’s discussion of his paper, comments:

> Mason goes on to say, “Statistics has definitely evolved into a field in which people can do their work without actually seeing and doing applications.” If anything, he is being tactful. Some members of the profession are trying hard to make changes, by teaching courses in which substantive questions come first and technique is introduced to find answers. Of course, all too often, technique comes first; data come in as purely decorative illustrations – a practice not confined to statistics departments.

Currently a problem-driven exposition of modern statistical approaches seems a priority for training researchers in application areas. Particular data mining pattern recognition approaches are relevant to some specialists. Otherwise the promise or hope of widely applicable extensive and powerful abilities for automated data analysis remains a dream for the future. One reason is that the statistical computing tools that are needed to support automated data analysis are not yet mature. Even the best implementations of some well-understood and important methodologies have serious deficiencies. For example, Generalized Additive Models offer a methodology for handling general nonlinear responses, and seem a huge advance on classical linear methods. There are serious deficiencies in all of the current implementations, which would seriously compromise any automated system that attempted to use them as a basis for an automated system. Tools for handling Bayesian statistical analysis are even less mature.

A Holy Grail for statistical computing research had been and perhaps remains the building of a statistical expert system. There is at least one expert system for industrial experimental design that has been relatively successful. A more ambitious project called REX (Gale and Pregibon 1984), aiming to build a system for regression analysis, has now for the time being been abandoned. Subsequently Nelder and Pregibon worked on an expert system called GLIMPSE. This too seems to have foundered. The wilder dreams of some data miners may be seen as an attempt to revive the quest for this Holy Grail.

The Design of Data Collection

Data mining searches for information different from what the data were initially collected to provide. Where data mining or other forms of exploratory data analysis provide evidence of valuable ancillary information, this may suggest some redesign of the data collection. Ongoing data mining exercises are far more likely to yield useful information if subsequent data collection has regard to the information that data miners may hope to get from it.

Moves to link museum taxonomic collections in huge databases provide an example. Because of the scarcity of data on species distribution and abundance, there will be attempts to use this taxonomic data for distribution and abundance assessment purposes for which it was not designed. There is the same potential for misleading inferences as from human inter-population and migration studies. To what extent is it reasonable to expect taxonomists to redirect their field collection so that future data are better able to address issues of distribution and abundance?

Decision Trees, Artificial Neural Networks (Neural Nets), and Related Methods

We discuss these in further detail because they seem, at present, the data mining tools which are favoured by software developers. More sophisticated practitioners combine these with more classical techniques, notably logistic regression. Largely, these tools are used for discrimination, where they compete with more classical statistical methods. Hybrid methods, which combine two different types of method with the aim of getting the best of each, are also available. Thus CART[3], which started out as a decision tree system, has now been extended to incorporate logistic regression, either as an adjunct or as an alternative to decision trees.

A tree-based classification gives a decision tree that is similar to a botanical classification key, except that all splits choose between two alternatives. In the simplest version of the methodology, splits are formed one at a time in sequence.

An (artificial) neural network consists of a large number of processing elements (neurones) and a complex pattern of linkages (synapses). Each linkage has a connection weight. The learning process adjusts the connection weight. A neural net should be thought of as a mathematical model for a learning process, rather than as a model for a brain! There are various different specialised types of neural nets. They seem good pattern recognition devices, and good at solving problems where algorithmic solutions suffer from undue computational complexity.

There are close connections between certain specialised types of neural nets and approaches developed by statisticians. Neural nets have attracted wide interest from engineers, computer scientists, biologists, neurophysiologists, psychologists and statisticians. Highly parallel systems may be required for efficient implementation. Neural nets that operate as black boxes, providing an answer but without clues on why a particular classification was chosen, are unsatisfactory when models are used in order to gain scientific insight. Depending on how they are used, tree-based methods may be open to the same objection.

Lim et al. (1997) compare 33 different methods, including twenty-two decision tree methods and two different types of neural net, for handling 32 classification problems. This is a careful and thorough piece of work. Over the 32 problems, the neural net programs were among the poorer performers. For purposes of generalizing to a wider population from which these data sets were drawn, the differences between methods were not however significant at the usual 5% level.

The neural net ‘Frequently Asked Questions’ information at http://cvor.pe.wvu.edu/faq/nnfaq.htm gives a good overview. There is a list of books that identifies the best and the worst. Comments on two of the worst are interesting:

> Both Blum and Welstead contribute to the dangerous myth that any idiot can use a neural net by dumping in whatever data are handy and letting it train for a few days. They both have little or no discussion of generalization, validation, and overfitting.

For statistical problems, neural nets provide a range of models which extend currently available models. Contrary to claims sometimes made in popular literature, neural nets require exactly the same kinds of assumptions as more conventional statistical models. Statistical inference requires the usual types of assumptions about the error structure.

Decision trees and neural nets seem most effective with very large data sets, with at least some tens of thousands of records. For smaller data sets, parametric methods which build in more structure may be preferable. In the trade-off between sampling variance and model bias, sampling variance may be more serious in data sets with some hundreds of records, while model bias may be more important in data sets with tens of thousands of records. Biases that are inherent in the data themselves are unaffected by sample size.

For statistical perspectives on neural nets, see Ripley (1996), Elder and Pregibon (1996), and Cheng and Titterington (1994). The theory and software have not yet developed to a point where neural nets are everyday tools for practicing statisticians.

Friedman and Fisher (1998) discuss an approach, in the same realm of ideas as decision tree methods, which for some regression applications seems to offer advantages over a decision tree approach. Their method uses all variables to define a small part of the space

[3] Details of CART may be found at the site http://www.salford-systems.com

spanned by explanatory variables that is ‘peeled off’ at each split, leading to an increase in the value of the target variable in the part of the subspace that remains.

6. Specialised Applications of the Mathematical Sciences

I have focused on methodologies that have wide relevance over many different application areas. Here I will comment briefly on developments that have more specific relevance.

There are large differences between research areas in the extent and nature of their demand for mathematical skills. Within a research area, there may be large changes as the area develops and matures and finds commercial applications. Geophysics has for several decades made severe mathematical demands. There has for some time been a demand for highly trained mathematical specialists in commercial finance. More recently, there has been evident demand for mathematical and computing specialists in molecular biology, generated in part by interest in exploiting the commercial opportunities of the new technology. Names that are commonly used for a range of disciplines that have resulted from this merging of mathematical science tools with tools from molecular biology are *bioinformatics, biological computing and computational genomics*.

Genomics and Bioinformatics

Genomics is a name for the study of the genome, i. e. of the DNA code which living organisms carry. There is tremendous commercial interest (Service 1998), and a large new demand for the skills of mathematicians, statisticians and computer scientists (Eliot 1996; Skoufos 1998). Biochemical methods for determining the proteins for which genes code are slow. Often, experts can make substantial progress towards determining protein structure by direct examination of the sequence of bases. *Bioinformatics* is often used as a general term for applications which involve skills in one or more of these mathematical sciences. *Biological computing* has a similar usage.

Genomics may be the basis for a third technological revolution, after the industrial revolution and the computing revolution. Multi-national companies, and some governments, are making huge investments in genomics-based industries. There are potential huge implications for health, medicine, pharmaceuticals, food production, and agriculture. For example, newly gained knowledge of the human genome, combined with the relatively automated rational drug design approaches, offer radically new possibilities for the development of pharmaceuticals. There are exciting new possibilities for the development of new industrial biochemical processes. Much of Australia’s production is strongly biology-based, making it important for Australia to be a major player in these emerging technologies and associated industries.

Biologists who wish to be active players in the new technologies will need new skills. As already noted, there is a heavy demand for a new type of specialist, with skills that have a strongly mathematical and computing orientation. We can expect to emerge, within bioinformatics, new genomics disciplines that require strong mathematics and computing skills. Those who have skills in these new disciplines seem likely to dominate the field.

Mishra (1998) notes

> The science of computational genomics and bio-informatics have been created out of this massive sea of sequence data and the need to establish functionality of genes largely based on similarities discerned at the level of the DNA code; bypassing the need for extensive biochemical characterization.

Training opportunities for bioinformatics specialists are currently limited. Skoufos (1998) comments that there are currently only five North American programs that offer PhDs in bioinformatics or computational biology. Any university which can move quickly to respond to the new demand is likely to tap a ready market. Industrial jobs are going, predominantly, to applicants with a knowledge of programming languages, and relevant mathematics and statistics. These may get, currently, higher priority than biological knowledge.

Important Australian bioinformatics sites are those for AGIC (Australian Genomic Information Centre), ANGIS (Australian National Genomic Information Service), and CMIS Bioinformatics.

Biological Computing – Further Issues

Topics set down for discussion at the Seventh International Conference on Intelligent Systems for Molecular Biology, to be held in Heidelberg in August 1999, indicate the range of mathematical and computing applications in molecular biology. These include molecular structure, genomics, molecular sequence analysis, evolution and phylogenetics, molecular interactions, metabolic pathways, regulatory networks, developmental control, and molecular biology generally. "Emphasis is placed on the validation of methods using real data sets, on practical applications in the biological sciences, and on development of novel computational techniques."

There are large and important statistical issues in genetic sequencing, in the sourcing of material, and in the supporting information that is provided when this information is incorporated into databases. Enthusiasm for data mining types of operations should not lead to a neglect of applications of the mathematical sciences to fundamental science.

Genetic data are rarely taken from a random sample of the population which they are thought to represent. For some uses of the data this may not matter. For others it clearly does matter. Ascertainment is a name for purposive sample selection. Comuzzie et al. (1999) consider the ascertainment of a sample based both on medical and genotypic alcohol dependence criteria, in order to increase the power to detect linkage. Different plausible approaches to correcting for the effect of the ascertainment bias on likelihood ratio estimates give substantially different answers. Bias problems may arise whether the sample selection is purposive or merely haphazard, creating large problems for inferences from material gathered using current procedures and stored in genetic databases. This is a serious issue for the use of genetic databases.

Problems may arise because different kinds of mathematical and computing specialists do not always communicate well with each other, or understand their own need of help in areas in which they are not competent. Thus individuals with a traditional mathematical training may not be well equipped to tackle statistical issues. Statisticians may wrestle with issues which require assistance from an expert in combinatorics.

Other Mathematical Science Application Areas

Detailed discussion would take me too far from the central themes of the present paper. Two recommendations in a recent U. S. report (Senior Assessment Panel 1998) relate to interactions with other disciplines, and warrant particular mention:

- Broaden graduate and undergraduate education in the mathematical sciences.
- Encourage and foster interactions between university-based mathematical scientists and users of mathematics in industry, government, and other disciplines in universities.

The report draws attention to missed opportunities that may arise because related ideas may develop in two different fields of mathematical application, but with use of notation and jargon which is so different that there is little or no cross-fertilization. An example familiar in my own area of knowledge is the relatively independent development of *hierarchical analysis of variance* and *repeated measures modelling* in statistics, *generalizability theory* in psychology and *multi-level modelling* in educational theory.

Finding 2 of the report relates to "Interactions with Users of Mathematics". This considers both the interactions of academic mathematics with industry and interaction with other disciplines. Often, the report argues, such interactions are often "obscured by the inward focus of mathematics and science departments". There is more to say:

> **The structure of universities mitigates against interdisciplinary research.** While the above finding criticizes mathematical scientists for not collaborating more actively with other scientists and engineers, part of the fault lies with the organization and culture of universities, here and abroad, which restrains collaboration across scientific boundaries. The academic award system does not encourage collaboration; in fact, individuals who straddle fields reduce their chances of tenure. …

Cost-Benefit and Cost-Effectiveness Analysis

Researchers face increasing pressure to supplement their findings with a cost-benefit or cost-effectiveness analysis. In a comparison of two or more courses of action, cost-benefit analysis attempts to account for financial costs and financial benefits that accrue to all parties. Spillover effects which do not incur a direct financial cost or benefit – damage to the environment or increased noise as a result of a new motorway, are not included. *Cost-effectiveness analysis* is a more limited form of comparison, used when it is appropriate to express the outcome in terms of a single variable, e. g. number of pregnant women saved from contracting rubella per $10,000 expenditure.

It is necessary to understand the traps and limitations of these studies. If they are to be useful, they must get right both the subject area evidence and the economic assessment. The analysis is only as compelling as its weakest component.

For example, DeBaun and Sox (1991) investigate the optimum screening decision threshold for treatment of lead poisoning. They assume no poisoning cost and no treatment benefit for patients below a stated blood lead threshold, and a fixed poisoning cost and a fixed treatment benefit for patients above such a threshold, assumptions which are too crude to give useful results. The assumption that these costs and benefits change linearly with blood lead level, which may still be too crude, might change the conclusions.

Chapter 10 of Greenhalgh (1997), together with the references which she gives, are a helpful starting point for literature on the use of economic analyses in medicine and health.

7. New Training Demands for Research Students

Above, I have discussed research quality issues that bring with them new training demands. There are further reasons for improved training. Even before we examine the impact of new technology, we have a long list of issues which require attention. These issues are mostly, in a broad sense, statistical issues.

Our traditional systems of PhD training have relied heavily on a research apprenticeship model where the novice learns from a expert or ‘master’. They appear to work well for training in laboratory skills. They do not work so well where a major part of the task is the synthesizing of existing knowledge and data, as a background to new research. They do not always work well for research that requires skills in which the experts are themselves inadequately equipped.

Graduates who have a strong training in research methodology are not only likely to do a better job in their specialist area; they emerge better equipped to apply their research skills to areas different from those in which they were trained. They are more marketable.

It is now common for researchers to work for several years on a post-doctoral fellowship before gaining a permanent academic appointment. Specialized subject area training thus continues well beyond the PhD. It then makes sense for the PhD to have a strong focus on wide-ranging methodological skills as a grounding for all later research.

Research Protocol Design

A good model for more directed training may be the course in Research Protocol Design (Course CCEB 661) with which I assisted while at the Centre for Clinical Epidemiology and Biostatistics at the University of Newcastle. In this course, medical graduate students spend a full year designing a research study. As the design develops, it receives critical review from a team which includes, at a minimum, a biostatistician and an epidemiologist. I found it fascinating and rewarding to work with highly intelligent and strongly motivated medical graduates, some of them well-established as consultants, as they struggled to form a clear view of current knowledge, to construct a clear statement of research objectives, and to set down a clear plan of research. Researchers who have not had the benefit of such training are at a disadvantage. Attempting a major research project in clinical epidemiology or health or the social sciences or applied biology without such preparation may be a recipe for a seriously flawed study.

Statistical Consulting Unit Courses

The Graduate School courses in statistics address wide-ranging technical statistical issues[4]. They have been well attended. My course on Research Protocol Design this past year has addressed wider research planning issues. These courses complement the consultations which statistical consulting unit staff offer to individual students, where specific statistical design and analysis issues are discussed.

Our particular focus has been experimental and sampling design, and data analysis. Some expansion into general research planning issues, and data overview, seems desirable. Providing this training is backed up with access to specialist statistical skills where these are needed, there will be a better research outcome, and graduates will emerge better trained.

Statistical presentation issues are important for the perspective they give on the design of data collection, and on analysis. Problems with justifying the choice and use of data gain new force when the researcher must justify his/her procedure to a wider public (Maindonald 1992.) The choice of a relatively optimal form of presentation is, often, not trivial.

We have continued to introduce new leading edge approaches, including resampling methods, tree-based regression, hierarchical modelling (known in some quarters as generalizability theory), repeated measures modelling, and generalized additive modeling. In addition we have offered courses on modern statistical computing environments. In 1999 we may offer survival analysis. Our approach has been practical and example-based. Some refocussing of our courses is desirable, to draw attention to Data Mining and Systematic Overview perspectives.

We need to emphasize changes in statistical approaches which have emerged in the past ten or fifteen years. This pace of change will continue, though likely directions for change may be clearer than they were ten or fifteen years ago. As was pointed out earlier, the net effect of many of these changes is to give results in a form that is easier to describe in subject area terms than were the older analyses. Often results can be summed up in a few well-chosen graphs. Those who are steeped in the older approaches may have unlearning to do!

Courses on database technology and on web-based literature searches would be a useful complement to our courses. Note also the courses on social science research methods that are offered each year in late January and early February, under the auspices of the Australian Consortium for Social and Political Research Incorporated (ACSPRI)[5].

[4] A listing of 1998 courses, and course summaries, are at the web address http://www.anu.edu.au/graduate/scu/course98_s2.html

[5] Details are at the web address http://ssda.anu.edu.au/ACSPRI/COURSES/SUMMER/

Training for Co-operative Research

In industry and Government, researchers must often work as part of a team. The Total Quality Management movement has a strong emphasis on training in teamwork skills, perhaps following Scholtes (1988). See also Peters (1989). A strength of the Scholtes book is its emphasis on methods for generating and honing ideas. As noted earlier, there are links with qualitative research approaches which have become popular in social science.

Current PhD training, with its almost exclusive focus on individual effort, may not be ideal training for working in or with the teams which Scholtes and Peters describe. Teamwork skills may, typically, be less important for pure science projects than for highly applied studies. Thus a study on bruising when apples are transported by road (Maindonald and Finch 1986) required skills in horticulture, engineering and statistical experimental design. The trial might have benefited greatly from the insights of orchardists.

Often, there will be problems on which a group of students who are drawn from different institutions can work co-operatively, communicating by electronic mail. Where several apprentice researchers can be found, internationally, who are working on relatively similar topics, Cochrane Collaboration studies provide a model for a teamwork approach for reviewing the current state of knowledge. Such studies require a high degree of collaboration between participating specialists – statisticians as well as medical specialists. Documentation of the division of responsibility for the total task should satisfy any demands from examiners to identify the contributions of particular team members.

Internet-Based Searches

The Internet provides a facility, not otherwise available, for rapid dissemination of the latest research information. In many specialist areas, it has become a crucial tool for gaining access to the latest research knowledge. There are now some journals which subscribers can download from the internet. Professional electronic mail groups provide previously unheard of opportunities to listen in to expert discussions. It may be just as easy to get an answer from a colleague on the other side of the world as from a colleague in the same corridor. The internet is a source of new software libraries as they become available. The current release of the highly successful R project, in which an international team of collaborators is steadily enhancing a modern public domain Statistical Language, can be downloaded from the internet. The R project relies heavily on the internet for exchange of ideas, information and code.

The proponents of evidence-based medicine have taken up the new opportunities of the internet with enthusiasm. Detailed advice on web-based searches, with examples, is a major feature of the Sackett et al. manual (1997, chapter 2). They comment (p.55):

> The Internet is a veritable bouillabaisse for finding information, with a huge and outrageously expanding pot, and you never know when you stick your fork in what tasty morsel or bit of fishy debris you will stab …

Sackett et al. (p. 72) are scathing about information from textbooks:

> Are textbooks obsolete? Their bloated girth and rapid dating of the action parts of textbooks on diagnosis, prognosis and therapy make it difficult to believe they will survive the electronic age any better than dinosaurs did the Ice Age. Unfortunately, textbooks don't smell as their contents rot, so readers will need to develop alternative crap detectors to avoid poisoning their minds and robbing their patients of current best care. … (Fortunately the principles of science do not age so quickly!)

There is a section (pp. 72-76) on "Teaching skills on how to search". Sackett et al. comment

> Clinicians learn how to search for the best external evidence in different ways. Some are entirely self-taught …, some by watching colleagues search, some from reading books like this one and progressively more from organized seminars, workshops and short courses run by expert searchers. …
>
> Given the opportunity, people can learn to do searches as competently as librarians (for sensitivity, if not specificity, whether or not they have had formal training from librarians …

A small amount of training can help avoid time-wasting, and reduce the risk that key information will be missed. There is detailed technical discussion in Chambers and Altman (1995) of issues that arise from the use of internet-based searches in systematic reviews in medicine. It is important to know what might have been missed!

Librarians are likely to be useful partners, both in constructing short courses on internet based searches and in handling initial training. Library staff provided several effective short courses for graduate students during the time that I worked at the Centre for Clinical Epidemiology and Biostatistics at the University of Newcastle. Chapter 2 of Sackett at al. (1997) would be a good place to start in looking for ideas on the construction of training aimed at other specialist areas.

Knowledge Engineering and Machine Learning

A nice addition to Statistical Consulting Unit courses would be a visionary course on future fallout from current research in Knowledge Engineering and Machine Learning. It should be tempered with an account of past failures to deliver on often visionary promises! Our task is to keep learning from our students and from the literature, and to be sensitive to demands to make new information technologies accessible.

Bioinformatics (Biological Computing)

The *Scientific Opportunities Evaluation Workshop* (SCOPE) workshop, organized by Professor Adrian Gibbs and held on December 9 1998 at the ANU Research School of Biological Sciences, identified broad areas of research activity and resources available in Canberra. The range of expertise on the ANU campus and in CSIRO seems a good basis for mounting initiatives in genomics and other areas of bioinformatics. Genomics is of such importance to the nation and to ANU’s future as a leader in biological research training that it is vital to find ways around current funding roadblocks.

The methods and knowledge of molecular biology, including recent mathematical methodologies, have become widely important to all researchers whose focus is in some sense biology. This creates a demand for training that is accessible to students, both at undergraduate and graduate level, who are not mathematical science specialists. This includes anthropology, education and psychology, as well as biology itself.

There is a demand from industry for specialists with a high level of skill in relevant mathematical and computing tools. New course initiatives seem needed, both at the undergraduate and at the graduate level. Offering a few mathematical science courses to molecular biology students will not be an adequate response. The same mathematical science depth is required, though with different content, as in specialist mathematics and computing courses.

8. Summary of Main Points

The best clues on how research demands may change in the next two decades come from examining the most innovative changes of the past two decades.

Some changes will be driven by attempts to fix problems with our present approaches, some will be driven by technology and some will be driven by demands from funding bodies or from users of results.

The approaches of evidence-based medicine have large implications for medical research as well as for medical practice. These carry across to many or most other research areas.

It may seem a truism that new research should build on a basis of careful data-based overview of research to date. It is necessary to make it true!

Archiving of data when results are published may become standard practice in all areas. Access to earlier data would often be a huge help in improving the design of later studies. Such access may be essential for the conduct of high quality data-based overview studies. High quality data should be seen as a valuable resource.

For much published research, publication should be seen as a first step in exposing results to critical evaluation, not as a final imprimatur.

Web-based searches of relevant databases have become essential for literature review.

Large and often networked databases are making huge changes to the science information base. However there are serious potential problems with quality, relevance, and access to key background information. In addition, the emergence of charging systems which treat data as a commercial commodity may place serious restrictions on access.

Data mining, although oversold, emphasizes the new problems and opportunities that arise from data warehousing, and from the creation of new, often large, databases. It may be seen as an attempt to automate the processes by which statistical analysts often encounter unexpected information that is aside from the main purpose of the analysis. There are strong connections with Exploratory Data Analysis (EDA).

The history of the use of evidence from databases in clinical medicine gives insight into the sorts of problems that can be expected in other areas when such databases are ‘mined’ for their information. Insights gained by practical statisticians remain highly important. Data must be ‘good’ for its intended use.

In the past ten years there have been large changes in the methodology for data analysis, taking advantage of advances in computer software and hardware. Effective data mining must build on these methods and insights.

The biological sciences, and allied research areas, have been dramatically affected by the huge advances of molecular biology. There are demands for a new type of molecular biologist, with strong mathematical and computing science skills. The challenge is to find ways to respond quickly to the new training demands.

Reward systems are required that will encourage academic researchers to co-operate across disciplinary boundaries.

Graduates will benefit from a broader training in research methodology, both because it will improve training in their main area of research and because it will better prepare them to tackle other types of research problem.

Graduate students should have, as a component of their training, experience of working co-operatively. Often, there will be problems on which a group of students who are drawn from different institutions can work co-operatively, communicating by electronic mail. It may often be appropriate to extend the literature into a co-operative data-based overview exercise.

There are several ‘information technology’ areas where short courses, additional to current Statistical Consulting Unit courses, may be desirable. These include specialized internet searching and database use. There is room for more visionary courses that take up such themes as machine learning and knowledge engineering.

9. Appendix

Definitions

Biological Informatics:

> "Biological informatics, then, is concerned with developing and using computer, statistical, and other tools in the collection, organization, dissemination, and use of information to solve problems in the life sciences."
>
> [From the Centre for Biological Informatics home page: http://biology.usgs.gov/cbi/aboutcbi/informatics.html]

This seems little different from the way I would define statistics.

Bioinformatics:

This name seems largely restricted to the discourse of molecular biology. It is used as a general term for applications of the mathematical and computing sciences in molecular biology.

Cost-Benefit Analysis:

In a comparison of two or more courses of action, cost-benefit analysis attempts to account for financial costs and financial benefits that accrue to all parties. Spillover effects which do not incur a direct financial cost or benefit – damage to the environment or increased noise as a result of a new motorway, are typically not included. *Cost-effectiveness analysis* is a more limited form of comparison, used when it is appropriate to express the outcome in terms of a single variable, e. g. number of cures per $10,000 expenditure.

Data mining:

> "The computer automated exploratory data analysis of (usually) large complex data sets."
> [Friedman 1998]
>
> "Data mining is a set of methods used in the knowledge discovery process to distinguish previously unknown relationships and patterns within data."
> [Ferruza, quoted in Friedman 1998].

Definitions vary widely.

Evidence-Based Medicine:

> "Evidence-based medicine is the conscientious, explicit and judicious use of current best evidence in making decisions about the care of individual patients."
> [Sackett et al. 1998.]

Informatics:

> "Informatics: Research on, development of, and use of technological, sociological, and organizational tools and approaches for the dynamic acquisition, indexing, dissemination, storage, querying, retrieval, visualization, integration, analysis, synthesis, sharing (which includes electronic means of collaboration), and publication of data such that economic and other benefits may be derived from the information by users of all sections of society."
>
> [From the US President's Committee of Advisors on Science and Technology: http://www.whitehouse.gov/WH/EOP/OSTP/NSTC/PCAST/pcast.html]

Knowledge Discovery In Databases:

> This is sometimes identified with data mining. Or KDD may be described as the overall process of discovering useful knowledge from data, while data mining may be identified as the initial step of extracting patterns.

Machine Learning:

> “Machine Learning is the study of computer algorithms that improve automatically through experience. Applications range from datamining programs that discover general rules in large data sets, to information filtering systems that automatically learn users' interests.”
> [Mitchell 1997.]

Fayyad (1996) finds an overlap between machine learning and data mining, in the interest in the study of theories and algorithms for extracting patterns and models from data. From this point of view, KDD is concerned to know which of these patterns are useful or interesting. However, it seems hard to claim that there has been meaningful learning, unless the machine is somehow able to detect which patterns are useful. Lecture slides to accompany Mitchell (1997) summarize issues for machine learning thus:

- What algorithms can approximate functions well (and when)?
- How does number of training examples influence accuracy?
- How does complexity of hypothesis representation impact it?
- How does noisy data influence accuracy?
- What are the theoretical limits of learnability?
- How can prior knowledge of learner help?
- What clues can we get from biological learning systems?
- How can systems alter their own representations?

The first six seem statistical issues.

Statistics:

> The science of collecting, organizing, analyzing and presenting data.

This definition, which closely represents the point of view of practitioners, is broader than the view of statistics that is presented in much academic teaching of statistics. Academic statistics is often limited to the mathematical theory and computing tools that address the design of experiments, sampling design, and data analysis.

Web Sites

Bioinformatics and Genomics

AGIC (Australian Genomic Information Centre):
http://www.angis.org.au/AboutANGIS/AGIC/

ANGIS (Australian National Genomic Information Service): http://www.angis.org.au/

Bioinformatics Servers: http://www.ii.uib.no/~inge/list.html

Biotechnology Newswatch (published by the McGraw-Hill companies):
http://www.mhenergy.com/demos/biotech/

Course summary for a course on mathematical biology:
http://cs.nyu.edu/cs/faculty/mishra/COURSES/COBIO/cobio.html
[This has interesting discussion on central issues in computational genomics.]

Pharmaceutical Research and Manufacturers of America web page on genomics:
http://www.phrma.org/genomics/

Biological and Medical Informatics

Australian National University Bioinformatics Group:
http://life.anu.edu.au:80/index.html

Biological Informatics (USGS Biological Resources):
http://biology.usgs.gov/aboutcbi/informatics.html

CMIS Bioinformatics: http://www.cmis.csiro.au/sis/bio.htm

Delta – Descriptive Language for Taxonomy: http://osprey.erin.gov.au/delta/delta.html

European Centre for Medical Informatics, Statistics, Statistics and Epidemiology of Charles University and Academy of Sciences of Czech Republic:
http://test.euromise.cz/english

International Journal of Medical Informatics:
http://emlinux.uivt.cas.cz/english/material/elsivier/index.html

Medical Informatics Links: http://medschl-www.mc.duke/dukemi/misc/links.html

MSc/Diploma in Medical Informatics (City University, London):
http://www.city.ac.uk/mim/mscmi.htm (This has a link to the City University ‘School of Informatics”)

Health Informatics Sites and Companies: http://info.ex.ac.uk/cimh/ssites.html

There are a large number of sites devoted to medical informatics.

Data Mining, Decision Trees, and Neural Nets

CART and Salford Systems Consulting Group: http://www.salford-systems.com

Data Mining and Knowledge Discovery (web page for journal):
http://www.research.microsoft.com/research/datamine/jdmkd-cfp2.htm

Data Mining Frequently Asked Questions: http://www.kdnuggets.com/references.html

Decision Trees: http://www.stat.wisc.edu/p/stat/ftp/pub/loh/treeprogs/quest1.7/techrep.zip

Neural Net Frequently Asked Questions: http://cvor.pe.wvu.edu/faq/nnfaq.htm

Data Archives and Issues

(ANU) Social Science Data Archives (SSDA):
http://www.ssda.anu.edu.au/SSDA/about-the-ssda.html

CODATA interdisciplinary committee of the International Council of Scientific Unions:

Global Change Related Data Sets; Federal Committee on Environment and Natural Resources (CENR)’s Subcommittee on Global Change Research:
http://www.gcdis.usgcrp.gov/lsm.html

International Survey Centre (data archive): http://www.international-survey.org/index.html

International Social Survey Programme (data archive): http://www.isp.org

The Data Archive (UK): http://dawww.essex.ac.uk

Evidence-Based Medicine and the Cochrane Collaboration

These activities provide interesting and suggestive models for making the results of research more accessible, and for improving the quality of research, in other areas.

Bandolier (online journal of evidence-based medicine; can be downloaded without charge): http://www.jr2.ox.ac.uk:80/Bandolier/

Cochrane Collaboration (Systematic Overview in Medicine & Health): http://www.cochrane.de/

Evidence-Based Medicine: http://cebm.jr2.ox.ac.uk/index.extras

Evidence-Based Health; various links: http://cebm.jr2.ox.ac.uk/docs/otherebmgen.html

Evidence-Based Practice on the Internet: http://www.shef.ac.uk/~scharr/ir/netting.html

Financial Mathematics

Online resources, commentary, and advice: http://www.contingencyanalysis.com/

Machine Learning

Online machine learning resources, and pointers to machine learning sites: http://www.ai.univie.ac.at/oefai/ml/ml-resources.html

Public Domain Statistical Language – R

R (Public Domain Statistical Language): http://www.ci.tuwien.ac.at/~hornik/R/R-FAQ.html

Research on the Careers of Graduate Students

FASEB (Federation of American Societies for Experimental Biology) Consensus Conference on Graduate Education http://www.faseb.org/opar/educrpt.html

Study of US graduate students conducted at the University of California at Berkeley http://amber.berkeley.edu:5900/publications/NEWS/Fall97/F7STUDY.htm

Workshop on employment outcomes of doctorates in science and engineering http://www.cpst.org/PAGES/cpst/site/pr012.htm

Statistics Courses Available to ANU Researchers

Australian Consortium for Social and Political Research Incorporated (ACSPRI): http://ssda.anu.edu.au/ACSPRI/COURSES/index.html

Statistical Consulting Unit of the Graduate School (1998): http://www.anu.edu.au/graduate/scu/course98_s2.html

10. References

Altman, D. G. 1994. The scandal of poor medical research. Lancet 308: 283-284.

Andersen, Bjorn 1990. Methodological Errors in Medical Research: an incomplete catalogue. Blackwell Scientific Publications, Oxford.

Bartholomew, D. J. 1995. What is statistics? Journal of the Royal Statistical Society A 158: 1-20.

Begg, C., Cho, M., Eastwood, S., Horton, R., Moher, D., Olkin, I., Pitkin, R.,Rennie, D., Schulz, K. F., Simel, D., and Stroup, D. F. 1996. Improving the Quality of Reporting of Randomised Controlled Trials: the CONSORT Statement. Journal of the American Medical Association 276: 637 - 639.

Bryan-Jones, J. and Finney, D. J. 1983. On an error in “Instructions to Authors”. HortScience 18: 279-282.

Bussell, W. T., Maindonald, J. H. and Morton, J. R. 1997. What is a correct plant density for transplanted green asparagus? New Zealand Journal of Crop & Horticultural Science 25: 359-368.

Chalmers, I. and Altman, D. G., eds. 1995. Systematic Reviews. BMJ Publishing Group, London.

Chalmers, I. and Grant, A. 1996. Salutory lessons from the collaborative eclampsia trial. Evidence-Based Medicine 1: 39. Available at http://www.acponline.org/journals/ebm/janfeb96/notebook.htm

Cheng, B. and Titterington, D. M. 1994. Neural networks: A review from a statistical perspective. Statistical Science 9: 2-54.

Cleveland, W. S. 1993. Visualizing Data. Hobart, Summit, New Jersey.

Cochrane Injuries Group Albumin Reviewers 1998. Human albumin administration in critically ill patients: systematic review of randomised controlled trials. British Medical Journal 317: 235-240.

Comuzzie, A. G., Williams, J. T., and Blangero, J. 1999. The effect of ascertainment correction on linkage results in the COGA data set: A cautionary note. Genetic Epidemiology, to appear.

DeBaun, M. R. and Sox, H. C. 1991. Setting the optimal erythrocyte protoporphyrin screening decision threshold for lead poisoning: A decision analytic approach. Pediatrics 88: 121-131.

Diamond, J. M. 1997. Guns, germs, and steel : the fates of human societies. Random House, London.

Draper, D; Gaver, D P; Goel, P K; Greenhouse, J B; Hedges L V; Morris, C N; Tucker, J R; Waternaux, C M 1992. Combining Information. Statistical Issues and Opportunities for Research. National Academy Press, Washington D.C.

Edelstein. H. 1998. Quoted in Wilson, R. Beware of fool’s gold in the data mine. Canberra Times, Tuesday Nov. 10 1998, p. 55.

Elder, J. F. & Abbott, D. W. 1998. A comparison of leading data mining tools. Fourth International Conference on Knowledge Discovery & Data Mining. [Available from http://www.dataminingtlab.com]

Elder, J. F. and Pregibon, D. 1996. A statistical perspective on knowledge discovery in databases. In Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P. and Uthurusamy, R.: Advances in Knowledge Discovery and Data Mining, pp. 83-113. AAAI Press/MIT Press, Cambridge, Massachusetts.

Fabris, P. 1998. Advanced Navigation. CIO Magazine, May 15. [Available from http://www.cio.com/archive/051598_mining.html]

Fayyad, U. 1998. Editorial. Data Mining and Knowledge Discovery 2: 5-7.

Fayyad, U. M., Piatetsky-Shapiro, G. and Smyth, P. 1996. From data mining to knowledge discovery: An overview. In Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P. and Uthurusamy, R.: Advances in Knowledge Discovery and Data Mining, pp. 1-34. AAAI Press/MIT Press, Cambridge, Massachusetts.

Feinstein, A. 1984. The role of observational databases in the evaluation of therapy. Statistics in Medicine 3: 341-345.

Fraker, T. and Maynard, R. 1987. The adequacy of comparison group designs for evaluations of employment-related programs. Journal of Human Resources 22: 194-227.

Freedman, D. A. 1987. As others see us: A case study in path analysis, with discussion by K. Hope, C. A. Achen, P. M. Bentler, N. Cliff, J. Fox, S. Karlin, B. O. Muthen, Rogosa, D., T. J. Rothenberg, E. Seneta, and H. Wold. Journal of Educational Statistics 12: 101-223.

Freedman, D. A. 1991. Statistical models and shoe leather, with discussion by R. Berk, H. M. Blalock and W. Mason. In Marsden, P., ed., Sociological Methodology 1991, pp.291-358.

Friedman, J. H. 1997. Data Mining and Statistics. What’s the Connection? Proc. of the 29th Symposium on the Interface: Computing Science and Statistics, May 1997, Houston, Texas.

Friedman, J. H. 1998. Statistics 315B: Statistical Aspects of Data Mining (Winter 1998). Available from http://www.stanford.edu/~jhf/Stat315B.html

Friedman, J. H. and Fisher, N. I. 1998. Bump hunting in high dimensional data. Available from http://www-stat.stanford.edu/reports/friedman/SuperGEM/prim.ps.Z

Gale, W. A. and Pregibon, D. 1984. REX: an Expert System for Regression Analysis, Proc. Compstat 84, Prague, pp. 242-248.

Gartland, J. 1988. Orthopaedic clinical research. Deficiencies in experimental design and in determinations of outcome. Journal of Bone and Joint Surgery 70: 1357-1364.

Gigerenzer et al. 1989. The Empire of Chance. Cambridge University Press, Cambridge UK.

Goldwater, W. H. 1998. Freedom of information requests (letter to the Editor). Science 282: 1823 (Dec 4 1998).

Gough, M. 1998. Freedom of information requests (letter to the Editor). Science 282: 1823 (Dec 4 1998).

Green, S. B. and Byar, D. P. 1984. Using observational data from registries to compare treatments: the fallacy of omnimetrics. Statistics in Medicine 3: 361-370.

Greenhalgh, T. 1997. How to read a paper. The basics of evidence based medicine. BMJ Publishing Group, London.

Hampel, F. 1998. Is statistics too difficult? Canadian Journal of Statistics 26: 497-513.

House Committee on Science 1998. Unlocking Our Future. Toward a New National Science Policy. Available from http://www.house.gov/science/science_policy_report.htm

Jessup, A. J. and Baheer, A. 1990. Low-temperature storage as a quarantine treatment for kiwifruit infested with Dacus tryoni (Diptera: Tephritidae). Journal of Economic Entomology 83: 2317-2319.

Jorgensen, M. and Gentleman, R. 1998. Data mining. Chance 11: 34-39 & 42.

Kaiser, J. 1998. New law could open up old lab books. Science 282: 1023.

Kanarek, M. S., Conforti, P. M., Jackson, L. A., Cooper, R. C. and Murchio, J. C. 1980. Asbestos in drinking water and cancer incidence in the San Francisco Bay area. American Journal of Epidemiology 112: 54-72.

Kolsky, J. 1998. Statistics and data mining in the analysis of large data sets. Available from http://www.infosense.com/news/article/article1.html

Kuhn, T., 2nd edn, 1970. The Structure of Scientific Revolutions. University of Chicago Press, Chicago.

Lim, T.-S., Loh, W.-Y. and Shih, Y.-S. 1997, revised 1998. An empirical comparison of decision trees and other classification methods. Technical report 979, Department of Statistics, University of Wisconsin. [Available from http://www.stat.wisc.edu/p/stat/ftp/pub/loh/treeprogs/quest1.7/techrep.zip]

Maindonald J. H. 1992. Statistical design, analysis and presentation issues. New Zealand Journal of Agricultural Research 35: 121-141.

Maindonald J. H. and Cox, N. R. 1984. Use of statistical evidence in some recent issues of DSIR agricultural journals. New Zealand Journal of Agricultural Research 27: 597-610.

Maindonald, J. H. and Finch, G. R. 1986. Apple transport in wooden bins. New Zealand Journal of Technology 2: 171-177.

Marshall, Eliot. 1996. Hot property: Biologists who compute. Science 272: 1730-1732.

MathSoft 1997. S-PLUS 4 Guide to Statistics. Data Analysis Products Division, MathSoft, Seattle.

McCance, I. 1995. Assessment of statistical procedures used in papers in the Australian Veterinary Journal. Australian Veterinary Journal 72: 322-330.

McGuinness, D. 1997. Why our Children Can’t Read. The Free Press, New York.

McPherson, K. 1990. Why do variations occur? In Anderson, T. F. and Mooney, G., eds.: The Challenges of Medical Practice Variations, pp.16-35. Macmillan Press, London.

Mishra, B. 1998. Special Topics in Math Biology. Computational Genomics: G63.2856.002/G22.3033.006 [Course Summary]. Available from http://cs.nyu.edu/cs/faculty/mishra/COURSES/COBIO/cobio.html

Mitchell, T. 1997. Machine Learning. McGraw-Hill, New York.

Morrison, R. 1998. Communicators count the costs. (An interview with Robyn Williams.) Available from http://www.abc.net.au/rn/science/ockham/stories/s17313.htm

Moynihan, R. 1998. Too Much Medicine. Australian Broadcasting Corporation.

Nicholls, N., Lavery, B., Frederiksen, C. and Drosdowsky, W. 1996. Recent apparent changes in relationships between the El Nino – southern oscillation and Australian rainfall and temperature. Geophysical Research Letters 23: 3357-3360.

Oxman, A. D. and Güyatt, G. H. 1983. The science of reviewing research. Annals of the New York Academy of Sciences 703: 125-131.

Peters, T. 1989. Thriving on Chaos. Pan Books, London.

Ripley, B. D. 1996. Pattern Recognition and Neural Networks. Cambridge University Press, Cambridge U. K.

Rosenberg, M. 1968. The Logic of Survey Analysis. Basic Books, New York.

Sackett, D. L. and Oxman, A. D., eds. 1994. The Cochrane Collaboration Handbook. Cochrane Collaboration, Oxford.

Sackett, D. L., Richardson, W. S., Rosenberg, W. M. C. and Haynes, R. B. 1997. Evidence-Based Medicine. Churchill Livingstone, New York.

Sackett, D. L., Rosenberg, W. M. C., Gray, J. A. M., Haynes, R. B. and Richardson, W. S. 1998. Evidence-Based Medicine: What it is and what it isn't. Available from http://cebm.jr2.ox.ac.uk/ebmisisnt.html

Sagan, C. 1979. Broca's Brain. Random House, New York.

ScHARR (School of Health and Related Research, University of Sheffield). 1998. Netting the Evidence. A ScHARR Introduction to Evidence Based Practice on the Internet. Available at http://www.shef.ac.uk/~scharr/ir/netting.html

Scholtes, P. 1988. The Team Handbook. Joiner Associates, Madison, Wisconsin.

Selfridge, P. 1996. In from the start. IEEE Expert 11: 15-17 & 84-86.

Senior Assessment Panel 1998. Report on the senior assessment panel of the international assessment of the U. S. mathematical sciences. Available from http://www.nsf.gov/pubs/1998/nsf9895/start.htm

Service, R. F. 1998. Chemical industry rushes towards greener pastures. Science 282: 608-610 (Oct 23 1998).

Simoudis, E. 1996. Reality check for data mining. IEEE Expert 11:26-33.

Skoufos, E. 1998. Bioinformatics. Scientific discipline or support field. In HMS Beagle: The BioMedNet Magazine (http://hmsbeagle.com/hmsbeagle/1997/01/resnews/meeting.htm), Issue 43 (Nov. 27).

Smith, A. F. M. 1996. Mad cows and ecstasy: chance and choice in an evidence-based society. Journal of the Royal Statistical Society A 159: 367-383.

Snow, John. (1855) 1965. On the mode of communication of cholera. Reprint ed., Hafner, New York.

Taubes, G. 1998. The (Political) Science of Salt. Science 281: 898-907 (14 August).

The Data Archive 1996. The Data Archive. Sharing and preserving research data. University of Essex.

Thomas, D. B. and Mangan, R. L. 1997. Modelling thermal death in the Mexican fruit fly (Diptera: Tephritidae). Journal of Economic Entomology 90: 527-534.

Transborder 1997. Bits of Power. Issues in Global Access to Scientific Data/Committee on Issues in the Transborder Flow of Scientific Data, U.S. National Committee for CODATA, Commission on Physical Sciences, Mathematics, and Applications, National Research Council. National Academy Press, Washington D. C. Available from http://www.nap.edu/readingroom/books/BitsOfPower/index.html/

Wegman, E. J. 1995. Huge data sets and the frontiers of computational feasibility. Journal of Computational and Graphical Statistics 4: 281-295.