Data Science: An Action Plan for Expanding the Technical Areas of the Field of Statistics

William S. Cleveland

Statistics Research, Bell Labs

DOI:10.1002/sam.11239
Published online in Wiley Online Library (wileyonlinelibrary.com).

Abstract: An action plan to expand the technical areas of statistics focuses on the data analyst. The plan sets out six technical areas of work for a university department and advocates a specific allocation of resources devoted to research in each area and to courses in each area. The value of technical work is judged by the extent to which it benefits the data analyst, either directly or indirectly. The plan is also applicable to government research labs and corporate research organizations. © 2014 Wiley Periodicals, Inc. Statistical Analysis and Data Mining 7: 414–417, 2014

1. SUMMARY OF THE PLAN

This article describes a plan to broaden the major areas of technical work of the field of statistics. Because the plan is ambitious and implies substantial change, the altered field will be called 'data science'.

The focus of the plan is practicing the data analyst. A basic premise is that technical areas of data science should be judged by the extent to which they enable the analyst to learn from data. The benefit of an area can be direct or indirect. Tools that are used by the data analyst are of direct benefit. Theories that serve as a basis for developing tools are of indirect benefit. A broad successful theory can have a wide-ranging benefit, affecting data analysis in a fundamental way. For example, the Bayesian theory of statistics affects all methods of estimation and distribution.

The plan sets out six technical areas for a university department and advocates a specific allocation of resources to research and development in each area as a percent of the total resources that are available beyond those needed to teach the courses in the department's curriculum. Furthermore, the allocation applies to the makeup of the curriculum; that is, the allocations are a guideline for the percentage of courses in each of the technical areas. The six areas and their percentages are as follows:

- *Multidisciplinary Investigations (25%)*: data analysis collaborations in a collection of subject matter areas.
- *Models and Methods for Data (20%)*: statistical models; methods of model building; and methods of estimation and distribution based on probabilistic inference.
- *Computing with Data (15%)*: hardware systems; software systems; and computational algorithms.
- *Pedagogy (15%)*: curriculum planning and approaches to teaching for elementary school, secondary school, college, graduate school, continuing education, and corporate training.
- *Tool Evaluation (5%)*: surveys of tools in use in practice, surveys of perceived needs for new tools, and studies of the processes for developing new tools.
- *Theory (20%)*: foundations of data science; general approaches to models and methods, to computing with data, to teaching, and to tool evaluation; mathematical investigations of models and methods, of computing with data, of teaching, and of evaluation.

Universities have been chosen as the setting for implementation because they have been our traditional institutions for innovation, and they can rapidly redirect areas of work by changing what is taught to graduates of data science. But a similar plan would apply to government research labs and corporate research organizations.

Change is needed in the technical areas of data science because critical areas that can be of immense benefit to data analysts need more resources. Computing with data,

Correspondence to: William S. Cleveland (wsc@bell-labs.com)

an area whose importance has recently been recognized by computer scientists needs much more resources. Multidisciplinary investigations are a major activity in some departments but not others. Many students of data science go on to teach, but it is rare to find a course in university curricula on pedagogy. A rigorous evaluation of tools and their development applies to data science what statisticians routinely advocate for process improvement in other disciplines.

The primary agents for change should be university departments themselves. But it is reasonable for departments to look both to university administrators and to funding agencies for resources to assist in bringing about the change.

Please read on for more detail and discussion.

2. MULTIDISCIPLINARY PROJECTS

The single biggest stimulus of new tools and theories of data science is the analysis of data to solve problems posed in terms of the subject matter under investigation. Creative researchers, faced with problems posed by data, will respond with a wealth of new ideas that often apply much more widely than the particular data sets that gave rise to the ideas. If we look back on the history of statistics—for example, R. A. Fisher inventing the design of experiments stimulated by agriculture data, John Tukey inventing numerical spectrum analysis stimulated by physical science and engineering data, and George Box inventing response surface analysis based on chemical process data—we see that the greatest advances have been made by people close to the analysis of data.

Because data are the heat engine for invention in data science, the action plan allocates 25% of resources to data analysis investigations. This does not mean that each and every faculty member needs to analyze data. But data analysis needs to be part of the blood stream of each department and all should be aware of the workings of subject matter investigations and derive stimulus from them.

Students should analyze data. Doing so should be a required, major part of undergraduate and graduate programs in data science.

3. MODELS AND METHODS

The data analyst faces two critical tasks that employ statistical models and methods:(1) specification—the building of a model for the data and (2) estimation and distribution—formal, mathematical-probabilistic inferences, conditional on the model, in which quantities of a model are estimated, and uncertainty is characterized by probability distributions.

A vast array of methods exist for estimation and distribution, but far less effort has gone into methods for model building. True methods have been extensively developed for certain classes of models; one example is classical linear regression models. But many other widely used classes have virtually no methods; one example is random-parameter models (random effects, repeated measures, random coefficients, randomized blocks, etc.).

For the data analyst, the inequity is unwarranted. Often, the model building phase is the salient part of the analysis, and the estimation and distribution phase is straightforward. Model building is complex because it requires combining information from exploring the data and information from sources external to the data such as subject matter theory and other sets of data. Inevitably, specifications must be chosen by an informal process that balances information from the data, information from sources external to the data, and the desirability of parsimony. Tools that help specification are much needed by data analysts.

4. COMPUTING WITH DATA

Data analysis projects today rely on databases, computer and network hardware, and computer and network software. A collection of models and methods for data analysis will be used only if the collection is implemented in a computing environment that makes the models and methods sufficiently efficient to use. In choosing competing models and methods, analysts will trade effectiveness for efficiency of use.

Historically, the field of data science has concerned itself only with one corner of this large domain—computational algorithms. Here, even though effort has been small compared with that for other areas, the impact has been large. One example is Bayesian methods, where breakthroughs in computational methods took a promising intellectual current and turned it into a highly practical, widely used general approach to statistical inference.

Along with computational methods, computing with data includes database management systems for data analysis, software systems for data analysis, and hardware systems for data analysis. At the moment, most work in systems is in the hands of companies that develop commercial systems for data analysis. What data analysts use in practice is determined by these companies. They work hard to produce best possible systems, but innovation must be tempered by their need to stay profitable. Without the stimulus of research devoted to innovation, progress in computing with data has been slower than it could be, and the creativity that exists within universities has had almost no influence. This is unfortunate because opportunities

for breakthroughs in data analysis are at hand. This has resulted from two technological developments: powerful PCs and high-speed Internet technology, both at very low cost. We have the right to hope today for a low-cost hardware/software environment for data analysis spread out over a high-speed network that allows intensive analysis of very large databases.

Computer scientists, waking up to the value of the information stored, processed, and transmitted by today's computing environments, have attempted to fill the void. One current work is data mining, but its benefit to the data analyst has been limited, because the knowledge among computer scientists about how to think of and approach the analysis of data is limited, just as the knowledge of computing environments by statisticians is limited. A merger of the knowledge bases would produce a powerful force for innovation. This suggests that statisticians should look to computing for knowledge today, just as data science looked to mathematics in the past. This suggests that departments of data science should contain faculty members who devote their careers to advances in computing with data and who form partnerships with computer scientists. This suggests that students in data science should have the opportunity to pursue courses in computer systems and in the mathematics of computing.

John Chambers has demonstrated the ability of statisticians to succeed in research in computing with data and the ability of computer scientists to value highly such work. In 1998, Chambers and the S system for graphics and data analysis won the world's most prestigious software prize, the ACM Software System Award, receiving the citation, 'For The S system, which has forever altered how people analyze, visualize, and manipulate data'. The esteem in which S is held is made clear by the winners of the award in previous years such as UNIX, VisiCalc, TeX, SMALLTALK, PostScript, TCP/IP, the World Wide Web, Mosaic, and Tcl/Tk.

5. PEDAGOGY

A data science department in a university must, of course, concern itself with teaching in its own setting. But it is vital that resources be spent to study pedagogy and to teach pedagogy. It makes sense that such study encompasses more than the university setting; curricula in elementary and secondary schools, company training programs, and continuing education programs are important as well. Education in data science does many things. It trains statisticians. But just as important it trains nonstatisticians, conveying how valuable data science is for learning about the world.

6. TOOL EVALUATION

The outcome of two areas of data science—models and methods, and computing with data—are tools for the data analyst. Work in this area can be made more effective by formal surveys of the practice and needs of data analysts, and formal study of the process of developing tools. In other words, we need to measure and evaluate data science.

Statisticians are the first to step up to assert that process improvement needs process measurement and an analysis of the resulting data. Statisticians should turn this methodology inward to study data science. There should be surveys to determine what methods and models and what computing methods and systems are used by data analysts in practice today. There should be surveys that poll practicing data analysts to determine perceived needs for new tools. There should be studies to determine how the process of tool development can be improved.

7. THEORY

Theories, both mathematical and nonmathematical, are vital to data science. Theoretical work needs to have a clearly delineated outcome for the data analyst, albeit indirect in many cases. Tools of data science—models and methods together with computational methods and computing systems—link data and theory. New data create the need for new tools. New tools need new theory to guide their development.

Mathematics is an important knowledge base for theory. It is far too important to take for granted by requiring the same body of mathematics for all. Students should study mathematics on an as-needed basis. Some will need the finite mathematics often taught in computer science departments. Others might need measure theory and functional analysis.

However, all students need an intensive grounding in mathematical probability, but it must be probability in the sense of random variation and random variables, as opposed to probability in the sense of measure theory and measurable functions, which is needed by only a few students. The reason goes back to the data. Data vary, and that variation typically needs to be thought of probabilistically, and a superb intuition for probabilistic variation becomes the basis for expert modeling of the variation.

Not all theory is mathematical. In fact, the most fundamental theories of data science are distinctly nonmathematical. For example, the fundamentals of the Bayesian theory of inductive inference involve nonmathematical ideas about combining information from the data and information external to the data. Basic ideas are conveniently expressed by simple mathematical expressions, but mathematics is surely

not at issue. Mathematical theory is a means of investigation and can shed light on all areas of data science including the fundamental theories.

8. OUTCOMES

One outcome of the plan is that computer science joins mathematics as an area of competency for the field of data science. This enlarges the intellectual foundations. It implies partnerships with computer scientists just as there are now partnerships with mathematicians. It implies that students develop skills in computing as well as mathematics. It implies faculty expertise in computing as well as mathematics. The reason goes back to the data analyst. Today, exciting new frontiers of data science that hold great promise for analysts involve computing with data.

Another outcome results from extensive involvement in data analysis projects. It carries statistical thinking to subject matter disciplines. This is vital. A very limited view of data science is that it is practiced by statisticians. The wide view is that data science is practiced by statisticians and subject matter analysts alike, blurring exactly who is and who in not a statistician. The wide view is the realistic one because all the statisticians in the world would not have time to analyze a tiny fraction of the databases in the world. The wide view has far greater promise of a widespread influence of the intellectual content of the field of data science.

Two other outcomes result from making explicit the link between theory and data. First, it guides theory in important ways. Second, it can substantially increase the domain of support for theory. Departments can delineate the link explicitly in requests for funding for multidisciplinary investigations. This brings the possibility of funding for the theory of data science from sources that support other subject matter disciplines.

Finally, a successful implementation of the plan will bring new, exciting areas of research and development to data science.

9. REFERENCES AND ACKNOWLEDGMENTS

No single advocacy here is new. Each of the six technical areas has been discussed in the literature; the references at the end have colored the plan, but are a small sample of the available writings. Furthermore, there are many instances of university departments having addressed the growth of one or more of the technical areas set out here.

Still, an important aspect of the specifics of this plan is that the pieces fit together, reinforcing one another, and all are important for the data analyst. Much credit for the specifics goes to my colleagues at Bell Labs, both current and past.

I am grateful to Nick Fisher, whose idea was used to address the topic, and to Diane Lambert for very important suggestions for clarifying some of the points made here.

REFERENCES

[1] G. E. P. Box, Science and statistics, J Am Stat Assoc 71 (1976), 791–799.
[2] J. Chambers, Computing with data: concepts and challenges, Am Stat 53 (1999), 73–84.
[3] W. S. Cleveland, Visualizing Data, Summit, NJ, Hobart Press, 1993.
[4] G. W. Cobb and D. S. Moore, Mathematics, statistics, and teaching, Am Math Mon 104 (1997), 801–823.
[5] J. W. Cooley and J. W. Tukey, An algorithm for the machine calculation of complex Fourier series, Math Comput 19 (1965), 297–301.
[6] R. A. Fisher, The foundations of theoretical statistics, Philos Trans R Soc Lond A 222 (1922), 309–368.
[7] J. H. Friedman, The Role of Statistics in the Data Revolution, Statistics Department, Stanford University, Technical Report, 2000.
[8] A. E. Gelfand and A. F. M. Smith, Sampling-based approaches to calculating marginal densities, J Am Stat Assoc 85 (1990), 398–409.
[9] D. W. Marquardt, Statistical consulting in industry, Am Stat 33 (1979), 102–107.
[10] D. S. Moore, G. W. Cobb, J. Garfield, and W. Q. Meeker, Statistics education fin de siècle, Am Stat 49 (1995), 250–260.
[11] F. Mosteller, Broadening the scope of statistics and statistical education, Am Stat 42 (1988), 93–99.
[12] D. Nichols, Future Directions for the Teaching and Learning of Statistics at the Tertiary Level, Department of Statistics and Econometrics, Australian National University, Technical Report, 2000.
[13] D. Nolan and T. Speed, Teaching statistics theory through applications, Am Stat 53 (1999), 370–375.
[14] A. F. M. Smith, Public Policy Issues as a Route to Statistical Awareness, Department of Mathematics, Imperial College, Technical Report, 2000.
[15] J. W. Tukey and M. B. Wilk, Data analysis and statistics: an expository overview, In The Collected Works of John W. Tukey, L. V. Jones, ed. New York, Chapman & Hall, 1986, 549–578.
[16] E. J. Wegman, Visions: The Evolution of Statistics, Center for Computational Statistics, George Mason University, Technical Report, 2000.