1976

Statistical Computing - The Vanguard of the Future of Education

John R. Rice
Purdue University, jrr@cs.purdue.edu

Report Number:

76-186

STATISTICAL COMPUTING*
The Vanguard of the Future of Education

by

John R. Rice
Division of Mathematical Sciences
Purdue University

CSD-TR 186
April 1976

*Keynote address: Computer Science and Statistics - 9th symposium on the interface, Harvard Univ., April 1, 1976

STATISTICAL COMPUTING
The Vanguard of the Future of Education

by

John R. Rice
Division of Mathematical Sciences
Purdue University

ABSTRACT

We start by making a case that the future of education involves a shift from "Learning Knowledge" to "Learning How to Use Knowledge." Items presented include the "telephone encyclopedia", the hand held calculators and the built-in Fortran functions. We then examine the current state of this transformation as manifested by special purpose application systems and conclude that statistical computing is furthest along the road to the future. This situation has implications for both education in statistics and software systems for statistics. We single out two specific aspects for more detailed discussion: the liability question and the cost-benefit balance for the development of statistical systems.

1. THE FUTURE OF EDUCATION

I propose that one aspect of the future of education will be a shift from learning knowledge to learning how to use knowledge. By knowledge I mean not only facts (e.g. Columbus discovered America in 1492 or radon is a gas) but many kinds of "how to" information (e.g. how to compute square roots or how to diagram an English sentence). The reason for this shift is that computers and communications systems will place this knowledge at everyone's fingertips and it will be a waste of time and effort to drill all this into everyone's head.

I do not mean to say that we can dispense with all knowledge. Clearly the bulk of the population will still need to know (memorize) basic arithmetic facts. Further, some kinds of knowledge (e.g. English writing) is not likely to be available from a computer system in the foreseeable future. Let me illustrate the direction I foresee with some present or eminent examples.

A. The telephone encyclopedia. In the next decade or two we can expect to be able to subscribe to an encyclopedia service from the telephone company or, perhaps more likely, the TV cable company. If one wants to know who fought in the 30 years war, what a gnu is and how it lives or a summary of the properties of paladium; one merely picks up the phone, dials the encyclopedia and asks. The information returned will not be very sophisticated, but it will extend the breadth of knowledge of even well-educated people by order of magnitude. It will extend the amount of factual information available to the "man-in-the-street" by serveral orders of magnitude. It is interesting to note that the new Encyclopedia Britannica has a strong hierarchical organization, exactly the direction one must go in order to develop a computer based telephone encyclopedia. Note also that inexpensive radio telephones can be more common than transistor radios are today. It is not going to make much sense (especially to kids) for students to memorize the Gettysburg Address or the important battles of the Civil War when they can get that information at any time just by asking.

B. The hand-held calculator. This country is in the process of being flooded by inexpensive hand calculators. I will not repeat the various arguments as to why they will revolutionize the teaching of arithmetic but I just note two points. First, the battle lines are already being drawn in the educational establishment over their use and, second, there are very plausible arguments that their universal adoption can eliminate large chucks of the mathematics curriculum and simultaneously allow Johnny to get the right answeres consistently.

C. The built-in Fortran functions. In the early days of computing most programmers had to know in detail some algorithm for computing the basic mathematical functions like e^x, sin x or log x. These functions arose frequently and had to be calculated by arithmetic within the machine and assembly language programs. Since program libraries were not very common, it was up to the programmer to provide an algorithm. Scientists now compute these functions by simply writing EXP(X), SIN(X) or ALOG(X). They know little about how they might be calculated and nothing about how they are calculated. We have the situation where there is a relatively small number of people who know all the ins and cuts of calculating these functions and a mass of users who rely on their expertice.

D. Symbolic mathematical systems. A standard ingredient in an engineer's education has been learning to do formal differentiation and integration in the calculus courses. Many have complained about spending months learning to calculate things that can be found in the tables that every engineer owns. Educators have replied with some justification that the tables do not have everything and one must learn the various transformations of calculus in order to use the tables. There presently exist computer systems which can differentiate perfectly and which can integrate at a level excelled only by a very few experts. Once these systems become widely and easily available to users, the engineers can legitimately ask why they should use 4% of their total education to learn how to compute things that they can get faster and more reliably by just asking. Indeed, in many applications they do not even need to see these integrals which are immediately used elsewhere in the computations.

The common theme in these examples is the combination of a computer program which embodies an experts knowledge or know how and a communication system that makes this easily accessible to a mass audience. This role as a transfer agent for man's knowledge will eventually be one of the principal applications of computer-communications systems.

2. CURRENT COMPUTERIZED KNOWLEDGE TRANSFER SYSTEMS

Presently the use computer systems for knowledge and know-how transfer is primarily in scientific areas. There are two forms of this; the most wide spread is the program library. Every large scale scientific computer center has such a library and various organizations provide material (e.g. the ACM Collected Algorithms, the BMD statistical programs, the IMSL math-stat library, the IBM SL-MATH library). The second form is the (more or less) integrated system for some area of scientific applications. Here one has a special application area oriented language for communicating with the user. A number of these systems have mathematical underpinnings (e.g. in optimization, ordinary or partial differential equations, statistics) while others contain application area programs almost exclusively. The important point is that some of these are multi-million dollar projects and they provide scientific tools (know-how) to thousands of users. Many, if not most, of these users would be unable to write programs roughly equivalent to those provided by these systems even for the specific problem they have at hand.

3. THE UNIQUE POSITION OF STATISTICAL COMPUTING.

There are several attributes of statistical computing which make it unique and lead it to being in the vanguard. The nature of the knowledge (as seen by the user) is highly computational. There is a very large, diverse and established user community. The users and others realize (for the most part) that they do not understand statistics or statistical computing very well. One can almost make the analogy with color television where the user turns the knobs to get the pretty pictures without the vaguest understanding or interest in what is behind the knobs. I conjecture that the bulk of the users of statistical computing are in a similar situation and, indeed, there will be a great demand for "one button" programming.

4. SOME PROBLEM AREAS IN THE CREATION OF KNOWLEDGE TRANSFER COMPUTER SYSTEMS.

This projection has profound implications for the educational system, but I will not address myself to them. The requirements for the software in such a system are very stringent and a list of them contains almost all the "motherhood and apple pie" words of computing. The system must be easy to use and communicate with a user in his own terms. It must be robust and reliable in all respects and withstand considerable abuse gracefully. It must be efficient internally; it is not sufficient just for it to "work." Note that both covered wagons and 747's provide coast-to-coast transportation, but there is no market for covered wagons. The systems must be portable, expandable and easily maintained, which imply quality programming and documentation.

I won't elaborate on these requirements because they have been widely discussed in recent years. However, I will caution you that many have heard but not understood the message. Even if you believe that you have got the message, you have to be vigilent to avoid ignoring parts of it from time to time. A less common but potentially crucial question concerning these systems is liability. Who is responsible if a knowledge transfer system gives the wrong answers?

Most people view computer programs and systems like used cars. You pay your money, cross your fingers and take your chances. If the software is given away free, surely there can be no liability associated with it, or can there? Let's examine a few facts associated with this situation:

(a) Not so long ago there were widely circulated statistical packages which would produce wrong answers in situations where one could make a good case that a good program would not.
(b) Most statistical software is produced by professionals in the same sense as doctors, architects or CPA's.
(c) There are numerous instances of doctors losing law suits for errors made in providing free services, even for emergency aid at the scene of accidents.
(d) IBM was sued for millions in the Equity Funding scandal just because they provided the

hardware and utility software for the computer system that was used to perpetrate the fraud.

(e) There are enumerable financial or human disasters that can be caused by wrong answers from statistical systems.

(f) The courts are steadily extending the domain of implied warranties and liability for products and services.

(g) There is a growing surplus of lawyers at a time when work from automobile liability and medical malpractice suits is drying up.

I conclude from this that the creators of statistical systems should be very cautious. They had better test and retest their programs to be sure that they meet "professional" standards. The day might come when we have a bevy of lawyers looking over our shoulders trying to second guess us just as the doctors have now. If this unhappy situation comes to pass then we, just like the doctors, will find ourselves obliged to continually test for eventualities that are extremely unlikely to occur. The result, of course, will be higher costs all around and it will be delicate to decide where the safety features are no longer worth the cost. Note that the costs are borne by the users who are not the ones being protected by the safety testing.

5. THE COST-BENEFIT BALANCE FOR STATISTICAL SYSTEMS.

The statistician who is a casual computer user will probably underestimate the effort for a moderate statistical system which meets typical current standards by one and perhaps two orders of magnitude. Suppose one set out today to develop a new statistical system, one with all new programming but without any research and development of new statistical methods. One should expect to spend the order of 1 or 2 million dollars over a period of 4-5 years in order to produce a stable, documented, portable and quality system. And I mean quality by todays standards, not a system suitable for the mass use that I visualize for the future.

That is a staggering amount of money. The enhancements required for the future will at least double and probably quadruple the cost and this assumes that the communications and language analysis technology is borrowed free of charge. In order to put this effort into perspective, we must consider the costs of other activities in the statistical education sphere:

(a) The cost of preparing a statistics text for production (author's time, editorial time, typesetting and associated overhead) is about \$60,000. If we assume 25 new books per year, then about 1.5 million dollars per year is being invested in preparing educational material in statistics.

(b) The full tuition cost of a semester course in statistics is about \$500. If we assume about 300,000 students take statistics each year, then about 150 million dollars is being invested per year in statistical instruction.

I thus conclude that the cost of preparing a computer system for the transfer of statistical knowledge and know how is very modest compared to the cost of current efforts with the same goal.

What then are the benefits to be gained. It is clear that we cannot eliminate either textbooks or classes in statistics. I believe that if we aim to achieve the same level of ability, then we can substantially reduce the instruction. However, I do not believe this reduction will be significant because we will opt instead for substantially raising the level of ability. I come back to the analogy with the hand-held calculator where almost everyone will get 14850/306 to be 48.529 even 20 years after their last math course. Recall that recent statistics show that 80% of the entering college freshmen cannot do long division correctly. I believe that it is plausible that the universal availability of such a system will at least double and probably quadruple the measured level of ability of statistical students. Furthermore, the level of ability will decay at a much slower rate than at present. Since the current level of ability is worth \$150 million per year (that is what we are paying for it) then the benefit of a statistical knowledge transfer computer system is conservatively valued at \$150 million per year.

We cannot rush out and start building the system on this basis because we do not currently have the communications system required to support it. But we can foresee the required communications and, in the meantime, computer power is currently widely enough available to justify investments in forerunners of such a system.

6. CONCLUSION.

We argue that computer based knowledge transfer systems will revolutionize (evolutionize?) the education process in a variety of subjects. Factual and algorithmic knowledge is most suitable for this evolution. The primary benefit of this evolution will be a very substantial increase (doubling or quadrupling) in the level of ability in these subjects achieved through this approach. The costs may well be very modest compared to these benefits and, indeed, this approach is one that can measurably increase the effectiveness of our education system in a large number of areas.

We suggest that statistics is in a somewhat unique position as the leader of this development and it is likely to remain so.