Future developments in social science computing

by GEORGE SADOWSKY

The Urban Institute
Washington, D.C.

INTRODUCTION

The set of computer related activities characterized by the term "social science computing" is both diverse and extensive. During the past 15 years, such activities have grown substantially both in scope and in volume and have become increasingly important both for basic research in the social and behavioral sciences and for public policy formation and evaluation.

In this paper I will present an overview of the evolution and present status of social science computing and examine several factors that are likely to have a significant effect upon its development during the next 10 years. Predicting the future is generally hazardous and is complicated in this instance by being dependent upon the development of future computing technology. During the relatively short history of automatic digital computing there have been many predictions—both optimistic and pessimistic—that have returned in short order to plague those who made them. Despite the high risk of error, attempts to predict the future have some value in planning for that future state. Attempts to forecast the future also can often provide a more accurate assessment of the present, a process from which social science computing might benefit considerably.

The term "social science computing" is somewhat misleading in that it implies a homogeneity of computing activities that does not now exist. Activities as diverse as process control in computer controlled experimentation laboratories, dynamic macroeconomic simulation, large survey data collection and analysis, simulation of alternative social policies, dissemination of census results, multivariate statistical techniques, and natural language text processing for content analysis and concordance production are all included. Social science computing activities may be observed as a part of research activity in universities and research institutions, in commercial organizations that produce social science computing products and in those that use them, and in government administrative processes and policy evaluation functions. Statements regarding social science computing generally are true of a subset of these activities only. To the extent that there is a commonality among activities in the field, it is that they are a collection of processes centered upon observed data relating to behavioral phenomena.

Many aspects of social science computing enjoy an independent tradition. The methodologies of commercial data processing, information retrieval, source data capture, statistical inference and digital system simulation all apply to social science computing activity, yet professional and methodological development in these fields generally occurs independently of it. The dependence of social science computing upon a variety of existing methodological traditions reduces its cohesiveness as a discipline and complicates the process of training social scientists for research careers in which a knowledge of quantitative methodologies is becoming increasingly important.

HISTORICAL OVERVIEW

Social science automatic computing activity dates from the introduction of automatic data processing equipment in the 1880s by Herman Hollerith. Hollerith's machinery was initially used in 1887 by the City of Baltimore to tabulate mortality statistics,17 and was later used by other local governments for similar functions. However, its use in sorting and tabulating the results of the 1890 U. S. Decennial Census of Population marks the beginning of large scale social science computing. Hollerith's initial innovations led to the development of a family of "unit record" electromechanical accounting machinery for general data processing applications. A primary use of these machines was to collect, process, and disseminate statistics for public purposes. The introduction of Hollerith's automatic machinery in the census process was followed by a pe-
period of increasing use of automatic accounting machinery by government for processing social and economic statistics. Shortly after the production of electronic digital computers, the first computer produced commercially—a Remington Rand Univac I—was installed at the U. S. Bureau of the Census in 1951 for processing data collected during the 1950 U. S. Decennial Census of Population.²

The automation of statistical computing activities began well before the invention of the automatic digital computer. Many statisticians made intensive use of the commercial unit record equipment available prior to the digital computer. Despite the fact that calculations performed using this equipment were slow, required extensive wiring of plug boards, and required repetitive use of different specialized machines, the resulting procedures were generally more accurate and faster than the corresponding calculations performed by hand. Thus statisticians quickly took advantage of the opportunities offered by automatic digital computation.

During the initial years of automatic digital computer use, social science computing tasks were similar to many other computing tasks. Computers were programmed almost exclusively in primitive assembly languages, the programmer often operated the computing equipment himself, and computing activity was localized around the few computer installations then available. Survey research centers and individual social scientists analyzing empirical data continued to rely heavily upon unit record processing equipment, especially specialized equipment such as the IBM 101 Counter-Sorter. The development of higher level programming languages such as Fortran and Comtran allowed programmers to write programs that were more easily understandable by non-programmers and more easily exportable to colleagues using different computing equipment. During this period the initial “general purpose” statistical and data processing programs were produced; such programs were often referred to as “canned” programs or “program packages.” They were characterized by being machine dependent, having numeric field control cards and fixed field punched cards as input, and producing a limited range of outputs, usually sparsely labelled. Substantial emphasis was placed upon internal execution efficiency.

An initial milestone in the development of statistical computing was achieved when W. J. Dixon and his colleagues at UCLA created the original set of biomedical statistical programs in 1962. This set, initially known as the BIMD series, provided the user with an integrated set of programs containing a large number of statistical procedures; the series included such features as a common vocabulary across programs, a comprehensive method of specifying data transformations, adequate user documentation, and readable and intelligible printed and graphic output. The BIMD series surpassed previous products by its completeness, increased user orientation, and emphasis upon exportability to other computer centers. While some aspects of the original BIMD programs now appear primitive, the initial impact of this set of programs upon statistical computing was substantial. The BIMD programs were developed during a time of intense batch program construction activity; some of the batch systems that followed BIMD such as P-STAT³ and DATA-BANK-MASSAGER⁴ were oriented specifically to social science users.

Another milestone in the development of social science computing was the creation and implementation of higher level languages oriented specifically to tasks in the field. Shortly after the concept of higher level languages became generally accepted, a number of social scientists began to develop languages oriented specifically to social science computation. I believe primary credit should go to Arthur Couch and David Armor at Harvard University for their development of the Data-Text language.⁵ Although implemented originally in assembly language, Data-Text provided social scientists with free form statement types linked directly to behavioral science computing procedures. Considerable effort was spent upon organizing printed output to the user’s specifications. The development of Data-Text led to other efforts such as BEAST¹ and SPSS¹ which have been useful as system design experiments and as applications programs for obtaining statistical outputs and tabulations from rectangular data files.

A third milestone in social science computing was the successful exploitation of interactive computing systems for research and instruction in the social sciences. Using computer systems that support economical interactive computing, a number of such programs have been developed. These include IMPRESS¹² for on-line instruction in social science computer based research methods, BASIS² for on-line statistical analysis, TROLL² for on-line time series analysis, equation system estimation, and econometric simulation, and others such as TRACE-III,³ OPS-3,⁷ and ADMINS.¹⁴ These systems are examples of the manner in which interactive environments have been used to support powerful tools for social science computing tasks.

Most social science computing activity has depended critically upon access to behavioral data in machine readable form. Government agencies and survey research organizations pioneered in developing techniques for recording survey data on punch cards and analyzing
these data mechanically. Universities and private survey research centers today have substantial holdings of sample survey data, and much of this material is available for secondary analysis. Several specialized data distribution networks have evolved; for example, the Inter-University Consortium for Political Research at the University of Michigan distributes political data by mail from its archive to over 100 member academic organizations, and the National Bureau of Economic Research makes available its on-line data bank of aggregate economic time series to its subscribers.

Government statistical agencies have historically been important sources of data for quantitative social science research. The publications of these agencies were an important source well before the widespread availability of automatic computing machinery. The U.S. Bureau of the Census provided a wide variety of population, housing, labor force and manufacturing data on an aggregated basis for many years; and after the 1960 census the Bureau initiated a policy of disseminating population microdata on magnetic tape for 1/10000 and 1/1000 random samples of U.S. families. Since 1969, microdata collected by the monthly Current Population Survey have been available in machine readable form to qualified researchers. Availability of public data in machine readable form has increased in recent years as government statistical operations have become more mechanized and as social scientists have learned to obtain and use the data effectively.

CURRENT STATUS

Social scientists have available to them today a wide variety of computer programs for their use. Batch-oriented programs for statistical analysis of rectangular file structures are most common; in addition to the better known and more reliable of these programs, there exist many locally developed programs that perform frequently used statistical procedures. The quality, accuracy and documentation of these programs vary substantially, and recent evidence presented by Longley and Wampler indicates that increased quality control is needed in their production. Computing languages for the social sciences are relatively new; however, there has been sufficient development to indicate the power that such languages provide, and their continued development seems warranted. More recently, interactive programs have been developed that give social scientists flexible control over their procedures and their data. Some very useful interactive tools have already been produced, but much work will be required to exploit current interactive computer systems in a truly productive manner.

An increasingly copious flow of behavioral data is now generated by individual researchers, private firms, public organizations and governmental organizations, and much of it is in machine readable form. The increase in complexity of the processing operations now being used require the input data to be more error free than was required in the past; present consumers of microdata especially must be much more sensitive to noise in their data. There is now increased demand for government data by private researchers, and the conflict between the use of individual data for legitimate public purposes and the confidentiality of such data has yet to be resolved in a satisfactory manner.

Major problems currently exist within the field of social science computing. Among the most serious appear to be: (1) inadequate standards for data documentation and problems of data transfer; (2) the low level of computational knowledge among social scientists and the lack of adequate training available; (3) the slow rate of diffusion of computing innovations into social science computing, especially in government; (4) problems of program inaccuracy, documentation and transfer; (5) lack of adequate software tools for many types of complex processing operations; and (6) the low level of professionalism in social science computing activities generally.

Some of the problems that exist in social science computing are common to most areas of computing and will be ameliorated by advances in computing generally. Other problems have been aggravated by the lack of resources allocated to the social sciences relative to other disciplines. Still others can be traced to misallocation of those resources made available. Finally, it appears that the current organization of the social science computing industry has itself produced or aggravated a number of present problems.

INDUSTRIAL ORGANIZATION

It seems both appropriate and relevant to employ one of the social sciences, economics, to study the industrial organization of the social science computing industry. For the totality of social science computing activity is an industry; it encompasses a variety of products, their producers and consumers, the markets in which they interact, the mechanisms for distributing goods, the reward structure for producers, the rate of investment in capital, incentives for investment, the rate of technical progress, and other industry characteristics. The institutional structure of social science
computing directly affects the manner in which individuals and organizations allocate resources to produce a wide variety of products. Solutions to current problems in social science computing are as likely to result from organizational change as from technical advances or increased resource supplies.

Most products within the social science computing industry fall into four general groups: (1) computer programs; (2) numeric data files; (3) knowledge regarding numerical methods, statistical procedures and computing techniques; and (4) personal services. Within each of these groups, the products produced are heterogeneous, although there is some evidence of product differentiation that is largely artificial.

Computer programs range in generality from specialized applications programs to general purpose applications systems and translators for high level programming languages. The former are final goods; they represent unique products designed to satisfy specific requirements. General purpose programs are intermediate goods, or producer goods; once created, they are then used repeatedly to produce a variety of final goods specified by non-programming computer users. Data files vary in size and complexity from those having a rectangular structure with few observations and variables to those resulting from the 1970 Decennial Census of Population and Housing.

The market for social science computing products appears to have four important dimensions: (1) locality; (2) computer type; (3) substantive focus; and (4) profit/non-profit. Markets have a strong local component because computing installations serve as centers of knowledge and expertise for their users, and rapid and inexpensive access to assistance is often essential for the success of computer related work. In addition, an organization having a computer center within its administrative jurisdiction frequently raises barriers to use of other centers by their staff. Also, the market is divided according to the manufacturer, model, and configuration of computing equipment available because programs, data and technical information can often be transferred at low cost between quite similar computers but not between dissimilar computers. The market is also stratified by substantive social science discipline or research or policy focus, especially with regard to machine readable data files and computer programs and hardware configurations that perform specialized processes within that discipline. Finally, there is infrequent interchange of programs between commercial firms and academic and research institutions.

There are a large number of producers of social science computing products, many of whom are individual social scientists and programmers. In addition, producers include social science computing groups within research organizations and universities, survey research centers, private firms, and various agencies of federal, state, and local governments. Many of these producers are also consumers; often production is initiated primarily or wholly for internal consumption. Most producers are active primarily within one of the above markets, although private firms can generally respond quickly to changes in demand by acquiring and reallocating resources.

With one outstanding exception, the subsets of the industry corresponding to the above broad product groups exhibit a low degree of industrial concentration, i.e., no small number of producers command a major share of these sub-markets. There currently exist a large number of producers of social science software, and except for large applications systems barriers to entry into the industry are almost non-existent. The number of producers (suppliers) of data is smaller, and entry into this industry generally requires a higher level of investment than entry into program production. Several organizations have made an investment in data banks containing economic time series, and competition to provide these data and associated computer software is now vigorous. Government agencies enjoy a natural monopoly position in collecting and disseminating data when the data must be made available to the government by law and is unlikely to be disclosed otherwise; examples include population census information and individual and corporate tax return data. In addition, government agencies enjoy substantial competitive advantages in collecting such data when it results as a byproduct of on-going administrative procedures or when the resources required to produce the data are larger than most potential producers can mobilize.

The performance of government in producing data in such a quasi-monopoly position has not been without fault. Under monopoly conditions, economic theory would predict that prices would be higher than if competition existed and the single producer would earn monopoly profits because of it. Government agencies are generally prohibited from pricing their products above average cost at most; but instead the monopoly position appears to encourage inefficient use of resources and technology and a lack of responsiveness to consumer demands. This results in higher prices for the consumer than would have prevailed in a competitive environment, a sluggish rate of innovation and less consumer choice.

The distribution mechanism for social science computing products is in the process of substantial
change. Products have been generally recorded on physical storage media such as punch cards, printed paper, and magnetic tape and these media were physically distributed to consumers. In the past, high costs associated with this distribution technology were partially responsible for product markets being segmented by computer installations and by hardware type. More recently, an increasing number of products are being distributed by transmitting digital data over common carrier telephone networks and specialized communications networks, and it appears that this growth will continue both in relative and absolute terms for some time.

Knowledge about social science computing products, statistical procedures, and related computing techniques reaches potential consumers in a number of ways. Inter-personal communication now forms one of the most important means of transmitting such knowledge. Some dissemination of this information is made in publications of computer societies, social science professional societies, research institutions, and various related academic disciplines. Informal publications, workshops, and conferences also play an important part in collecting and distributing knowledge. However, at the present time there are no information centers or professional societies focused primarily upon social science computing. SIGSOSC, the Special Interest Group for Social and Behavioral Science Computing of the Association for Computing Machinery, is a likely candidate for this position, but its development into a professional society will take at least 5 to 10 years.

Social science computing products are created in accordance with the production processes available to the industry at the time of production. A production process or function is a set of technical rules that specify how goods and services called inputs may be transformed into other goods and services called outputs. The levels of knowledge and technology available dictate what production functions are available for obtaining specified outputs. If a production process requires more than one input, there are usually a variety of combinations of inputs that are capable of producing the same set of outputs, i.e., the production function may allow input substitution. In computing, a typical example of input substitution is provided by the frequent tradeoff in program design between the use of main memory and processor time; using more of one generally provides savings in use of the other. Resource constraints and the relative prices of the inputs enter into the choice made by producers of which production process to use for obtaining output.

For many social science computing products, the cost of producing the first unit of a specific output is high relative to the marginal cost of producing subsequent duplicate units. The cost of duplicating a computer program or a machine readable data file, once the product has been created, is low and requires only the use of well-known duplication procedures. However, the costs paid by social scientists to import programs and data are generally much higher; the principal reason is that the distribution technology for these products is still rather primitive since the products often do not adapt easily to computing environments other than those very similar to the environment in which they were created. The total costs of distribution are therefore often much higher than the costs of duplication.

If the marginal cost of a unit of production is low relative to the initial investment required to produce the first unit, then resources are allocated more efficiently if there exist a few large producers rather than many small producers. Yet with the exception of the production and distribution of large data files and a few other products, it appears that the market is dominated by small producers. There are a number of factors that help to explain this phenomenon. First, many of these products are differentiated slightly but meaningfully for their consumers. Second, computing products may be one of a number of outputs of a joint production process; often another output is increased education for the producer. Also, the very low barriers to entry encourage new producers into these markets at low cost to the new entrants.

The above factors help to explain the existence of small producers of social science computing products, but they do not explain the relative absence of larger producers who are able to make substantial investments in more general and more powerful products. Even for the first unit of production, there appear to be substantial economies of scale. That is, there are some fixed costs in program preparation that must be incurred regardless of program size. Examples of such fixed costs in social science computing systems are code to read the data file and interpret its values, code to provide elementary output displays and code to perform standard data transformations; they are almost always included regardless of how the data are processed. To the extent that fixed costs influence the production process, a more comprehensive and more general program appears to be cheaper to construct than the many smaller programs that duplicate its functions.

A very important factor limiting substantial investment in computing products is that although the return to such investment may be high for the industry as a whole, most of the return cannot be captured by the investor. Another factor is the segmentation of the
market discussed previously; if individual markets are limited in size, then the cost of sales increases and potential returns are more limited.

Much social science computing activity occurs in universities, research organizations, government agencies and other environments that often do not enter the commercial marketplace to buy and sell these products. Within these organizations, producers are generally individuals, and the rewards sought are primarily nonfinancial; rather they are prestige and recognition for the producer's accomplishment. At the present time, however, the technical activity associated with producing computer products for social scientists does not enjoy high prestige or substantial recognition for the individual involved. "Tool using," or substantive activity, has higher status within social science professions than "tool building," or technical activity. Thus, although changes in the level and allocation of investment in computing tools for social scientists would lead to a more efficient allocation of individual skills and technical resources and would be expected to provide substantial externalities to social science as a whole, the reward structure of academic social science disciplines prevents the individual innovator from collecting his rewards. The lack of a satisfactory reward structure for non-commercial producers of social science computing tools is very unfortunate, since it discourages innovative investment in production for those who are often able to make substantial contributions to the field. The inadequate reward structure contains another unfortunate aspect; it perpetuates a lack of interest and involvement in computer oriented training by persons in the social science professions.

THE FUTURE

During the next ten years, continued substantial growth in social science computing activity is likely to occur. Social scientists currently being trained are taught quantitative methods to a greater degree than any of their predecessors. The amount of data in machine readable form is increasing rapidly. The decreasing cost of computation is shifting the balance in favor of using computers rather than using manual labor or not performing a task at all. Furthermore, as inventories of data and quantitative methods grow and as the number of practitioners in social science computing increases, important issues in social and economic policy become more tractable using the computer. Although growth in social science computing will not be uniform and will be unevenly divided among disciplines and applications, there is little doubt that it will occur.

Advances in computing hardware and software during the past 15 years have been responsible for a dramatic decline in the cost of computing. Current evidence suggests that logic and memory costs will continue to decline substantially during the next decade. These declines in cost will benefit social science computing activity by reducing the cost of those computing inputs.

In addition to declines in computing cost, three specific areas of development within the computer industry are likely to benefit social science computing substantially and should be assisted and encouraged. These areas are: (1) communications terminal technology and interactive system development; (2) digital data networks linking computers; and (3) data storage technology. Developments in these three areas will impact different aspects of the social science computing industry.

The development of reliable and responsive interactive computing systems has already had an impact upon the manner in which social science computing is performed. Previous reliance upon batch computing systems forced social science programs into a mold in which output decisions had to be made simultaneously rather than sequentially, computations producing large amounts of (often unread) output were the rule, and errors in input preparation resulted in increased computer costs and turnaround time delays that were generally large relative to the magnitude of the error. In general, batch systems have had the effect of separating a non-programming investigator from his procedures and his data.

Recent uses of interactive systems provide convincing evidence that it is possible to exploit such hardware in a manner that allows a user considerably more freedom and flexibility with his data and the procedures he employs. The TROLL system provides an example of such a use. Furthermore, in some general interactive systems including the first, CTSS, batch and interactive modes of computation are not antithetical but combine naturally to use common system resources. The appropriate mode of computation may be chosen according to the characteristics of the task and the preferences of the user. Recent systems such as Digital Equipment Corporation's PDP-10 include interactive-batch compatibility as a basic design feature; such a feature is not costly relative to the benefits it provides and should be more common in the future.

The widespread availability of interactive systems plus the development of faster and less expensive user terminals should cause a shift of much social science computing to these systems. Among the reasons are: (1) the ability to specify tasks of analysis incrementally
rather than collectively; (2) the potential of on-line interaction with data bases and the ability to browse through them with flexible computer support functions; (3) the generally lower cost of making a mistake in procedure and the ability to detect it more rapidly; and (4) the resynchronization of the user's processes of hypothesis formulation, testing, and analysis and the computer's use in supporting these functions. Batch processing will not and should not disappear; it will continue to be used for routine tasks in which interaction is not required or is inappropriate.

The computer terminal market has just begun a period of rapid development. In the past the terminal industry was dominated by two large firms, AT&T and IBM, and the market for terminals grew slowly; the rate of innovation in the industry was sluggish. As interactive computing began to increase, the market for terminals grew rapidly and many new firms entered the industry. The results have been increased competition, more product variety for consumers, lower costs, and an increase in the rate of innovation. If the growth in interactive computing continues, the terminal market should continue growing rapidly. Bigger markets should lead to increased specialization, some of which will benefit social science computing directly. For example, a low cost interactive graphics terminal having moderate power is well within the scope of present technology. Only a larger market is required for it to be introduced commercially. Cheaper, faster and more powerful terminals will make widespread use of terminals economically feasible and will allow enhancement of the quality and power of interaction.

The development and availability of general purpose computer networks is important because they will remove significant limitations on market size for social science computing products. By a general purpose network I refer to a network that can transmit digital data between any two of its members without knowing the content of the transmission. Such a network might exist on a permanent basis, such as the ARPA network with its dedicated interface message processors and reserved communications circuits, or it might consist only of interface protocols between each computer in the network and the network circuit itself. The existence of general purpose networks should alter this resource allocation pattern. Suppliers of programs and data bases will have access to a much larger market for their products; the prospect of larger returns should serve to increase both the degree of specialization of those producers and the amount of investment they are willing to make in their products.

For consumers of social science computing products, such networks will increase substantially the number and variety of products available to them. As an example, consider the problem of accessing data files residing at a remote computer center. In the past, data have generally been exported by moving the data file from its initial computing environment to the data requestor's computing environment; usually the environments are technically and administratively dissimilar. The technical problems involved in such transfers are solvable, but they are generally time consuming and tedious. Computer networks offer viable technical alternatives to this procedure. Using a network, it would be possible to transmit a request for processing and output to the computer environment in which the data are stored and transmit the results back to the requestor. If the cost of transmission is low and there are few administrative barriers to access, every user can regard any program or data file on the network as being within his own facility and available to him at only moderately higher cost. Alternatively, the data could be transferred directly to the user's computer, since the existence of the network implies functional interfaces between each computer in the network and the network circuit itself.

The availability of computer networks should substantially reduce the inconvenience and costs of accessing remote computers, provided that the institutional and administrative problems of interorganizational computer access can be solved. This will create larger markets for social science computing tools and will lead producers to raise their level of investment in the products they offer. In addition, there should take place a trend toward specialization in the construction of tools for use, and collection and maintenance of specialized data bases at various research-oriented computer centers. An example of this process is provided by TROLL, an on-line system for macroeconomic estimation and simulation. TROLL is available to its customers on-line using the AT&T telephone network. Partially as an outgrowth of the development of TROLL, a Computer Center for Economics and Management Sciences has been established by the National Bureau of Economic Research to continue and extend the production of quantitative methodology and computing tools. It is anticipated that access to programs and data will be made on-line,
using the telephone system to support network communication.

Initially, the existence of network communication links will be far more important than how the network is implemented. The existence of some form of interprocessor communication will allow substantial changes in how and where social science computing is accomplished. The configuration and the nature of implementation of the network is a matter of economics; it depends upon such factors as anticipated pattern and volume of traffic, number and location of subscribers, and required response characteristics. Until a general purpose network becomes available, more primitive network communication can be supported using the existing telephone network. It seems unwise to build a computing network dedicated to social science computing traffic either now or in the near future. Such an act would be comparable to allocating some of the industry's limited resources to design special computer architecture that would be dedicated to social science computing.

Of all components of computer hardware, the current status and use of data storage technology represents a major constraint on the fashioning of better social science computing programs. As social science programs grow in scope, they grow in memory required also. It is now common for large social science applications systems to use elaborate manual segmentation systems to accommodate themselves to a small amount of address space. Either the memory resources for ameliorating these restrictions do not exist, or the techniques for accessing them are not known or available.

In the absence of a fundamental breakthrough in providing substantially lower cost immediate access memory, computers with virtual memory organization\(^4\) appear to offer a solution to this problem. Virtual memory machines provide both increased program address space and data address space; the latter would provide the ability to process more complex data structures without making a substantial investment in file manipulation software. Virtual memory environments for building and using social science computing tools should obviate some of the ungraceful degradation characteristics which many social science computing programs exhibit upon reaching a capacity constraint. Perhaps most important, the availability of virtual memory would remove most memory management concerns from programmers and allow them to concentrate upon producing more useful computing tools. The opportunity cost of not using automatic memory management techniques is measured in terms of programmer time used in its place and investigator time waiting longer for results. This cost is already substantial in some cases, and it will increase as the cost of hardware declines relative to the cost of labor.

If the technical developments described above materialize, then their impact upon the social science computing industry should have very beneficial effects upon its performance. However, several problems exist within the industry that detract from its performance and that are not likely to be altered by changes in technology.

The first problem concerns access to data. While there has been much progress in recent years by private firms making data available in machine readable form, most behavioral and economic data generated by government agencies and private researchers is unknown to other potential users or underutilized by them. The principal reasons appear to be that: (1) within government agencies, quasi-monopoly positions have discouraged efficiency in both production and distribution; and (2) in academic and research institutions, the prevailing reward structure offers few professional rewards for data production alone. As a result, it is common to observe much duplication of data collection activity, underutilization of existing data sources, and lack of availability of existing timely and detailed data for research and policy analysis. Another problem is closely related. Social science disciplines currently assign lower status to most "technical" work and higher status to "substantive" work. To the extent that this status difference is perceived, new entrants into social science professions are more likely to elect to do substantive work. Those entrants with strong preferences for the computational aspects of the social sciences are more likely to gravitate toward academic computer science rather than toward involvement in a service role in which room for professional growth appears limited.

Within the present industry structure, academically oriented producers of both programs and data find it difficult to obtain academic rewards for their output. The production of a useful, accurate and well-documented social science computer program or machine readable data file may generate substantial externalities for social science as a whole, yet there is no adequate mechanism that allows its producer to capture these externalities. The long run effect of this situation must be to discourage production by those able to produce successfully in other fields.

The problem of developing a reward structure for the social science computing industry is not a new one. Its solution could result in substantial benefits to social scientists in terms of substantially more efficient use of the computing and data resources available to them.
CONCLUSION

During the next ten years there will be a substantial increase in demand for social science computing products from government, universities, research institutions and private businesses. A continued decline in the cost of computing will benefit all these consumers. In addition, the development of reliable general purpose interactive systems and inexpensive terminals will enhance the quality and power of social science computing activities; general purpose networks will enlarge traditional markets and lead to increased competition and better consumer choice; and the development of quite large addressable memories should remove an important existing constraint on the development of large applications systems and complex file processing tools. These developments should be supported by the social science computing industry. Problems concerning access to data and achieving a reward structure for social science computing exist and are important, but remain to be solved in a satisfactory manner.

ACKNOWLEDGMENTS

I would like to acknowledge helpful discussions with and criticisms from Mr. Ben A. Okner and Mr. Gregory A. Marks.

REFERENCES

1 J W BEAN S W KIDD G SADOWSKY B D SHARP
   The BEAST: A user oriented procedural language for social science research
   Processed; Washington The Brookings Institution June 13 1968
2 R BUHLER
   P-STAT: An evolving user-oriented language for statistical analysis of social science data
   (Describes P-STAT tape version 42.) Processed; Princeton Computer Center Princeton University 1966
3 A S COOPERBAND W H MOORE JR R J MEEKER G H SHURE
   TRACE-III: An implicit programming system for inductive data analysis
4 P A CRISMAN ed
   The compatible time-sharing system: A programmer's guide
   2nd ed Cambridge The MIT Press 1965
5 P J DENNING
   Virtual memory
6 W J DIXON ed
   BMD biomedical computer programs
   Berkeley and Los Angeles University of California Press 1967
7 M GREENBERGER M M JONES J H MORRIS JR D N NESS
   On-line computation and simulation: The OPS-3 system
   Cambridge The MIT Press 1965
8 The data-text system: A computer language for social science research designed for numerical analysis of data and content analysis of text
   Harvard University Department of Social Relations (Preliminary Manual) Processed; Cambridge Harvard University 1967
9 J M HENDERSON R E QUANDT
   Microeconomic theory: A mathematical approach
10 C C HOLT
   A system of information centers for research and decision making
   American Economic Review Vol 60 May 1970 pp 149-165
11 J W LONGLEY
   An appraisal of least squares programs for the electronic computer from the point of view of the user
12 W MALING
   Preliminary version of the TROLL/1 reference manual
   Processed; Cambridge Massachusetts Institute of Technology 1971
13 M C McCracken
   A computer system for econometric research
   Processed; Ottawa Economic Council of Canada 1966
14 S D Mcintosh D M Griffel
   The ADMINS primer—November 1968
   Processed; Cambridge Massachusetts Institute of Technology (Center for International Studies) 1968
15 E D MEYERS JR
   Project IMPRESS: Time-sharing in the social sciences
16 N NIE D BENT C H HULL
   SPSS: Statistical package for the social sciences
17 P J REX JR
   Herman Hollerith, the first statistical engineer
18 L G ROBERTS B D WESSLER
   Computer network development to achieve resource sharing
19 S ROSEN
   Electronic computers: A historical survey
   Computing Surveys Vol 1 March 1969 pp 7-36
20 D L ROSS
   Burroughs Advanced Statistical Inquiry System (BASIS)
   Processed; Detroit Burroughs Corporation 1969
21 R H WAMPLER
   An evaluation of linear least squares computer programs