The Origins of Statistical Computing

During the 1920s and 1930s, computing labs helped establish statistics as a discipline in the United States

by David Alan Grier

The Iowa State University Statistics Lab in 1930

Now that the era of Big Data has arrived and statisticians are searching for ways to handle massive data sets, it is interesting to look back on the early days of statistical computing and just how far it has come. This article originally ran in the September 2006 issue of Amstat News. Enjoy!

Statistical computing became a popular field for study during the 1920s and 1930s, as universities and research labs began to acquire the early IBM mechanical punched card tabulators. They used these machines for tabulating and computing summary statistics and for fitting more complicated statistical models, such as analyses of variance and linear regressions.

These labs proved to be important places for advancing statistical methodology. They helped make Galton's and Pearson's ideas on correlation practical tools that could be used for scientific research. They encouraged researchers to think in terms of large problems with extensive datasets. Without them, modern statistical methodology could have languished as an interesting theory, useful for small problems.

Many of these labs offered their services to physicists, astronomers, biologists, and social scientists. Some created tables of higher mathematical functions; others solved complicated differential equations. A few of these labs, most notably those at Iowa State University and Columbia University, became test beds for early computer scientists, who experimented with new ideas for computing machines and for numerical algorithms.

The larger labs were funded by donations. In the 1920s, there were no instrumentation grants for the mathematical sciences. The scientific infrastructure was developed by Vanevar Bush during and after World War II. The only source of government money for scientific research was the Department of Agriculture. This organization—supportive of empirical research—helped to establish the largest and most sophisticated of the statistical laboratories, the Statistics Lab at Iowa State University.

Some of the names associated with these early labs—James Glover, H. T. Davis, A. E. Brandt, Howard Tolley—are not recognizable by many. They published little and made only marginal contributions to the theory of statistics or the development of computers. Yet these researchers hoped that the combination of computing technology and mathematical statistics would radically change science.

The First Statistical Labs: Before the Tabulator

The earliest statistical laboratories were founded to study economic phenomena, but quickly began to apply their [attention] to problems in the social, biological, and behavioral sciences. Economic applications paid the bills, leased the equipment, and provided the salaries for laboratory workers. One of the first of these laboratories was founded at the University of Michigan by James Glover, a professor of mathematics. Glover was a pioneer actuary and student of financial risk. He began teaching advanced statistics courses in 1904, though they included little of the mathematical statistics that was being developed in England by Karl Pearson and R. A. Fisher.

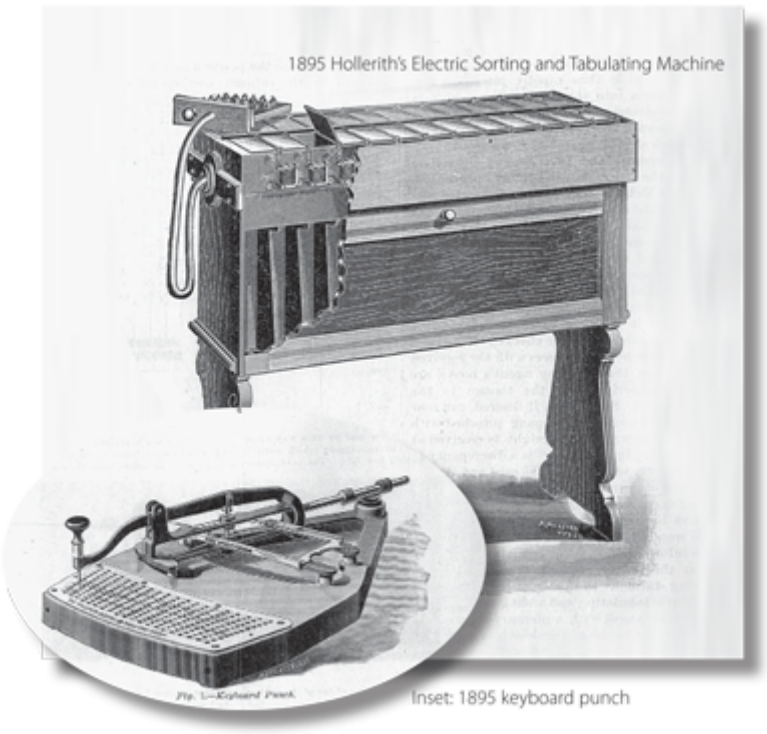
Glover had a small lab operating by about 1910. It was staffed by his students and followed the model of the computing labs found in observatories and astronomy departments. Glover's lab reduced data to summary statistics, created actuarial tables, cross-tabulated data, and made projections from simple statistical models.

Glover rarely had more than two computing assistants. Most were his students, and many were women. Even in the early 1900s, the University of Michigan was a coeducational institution, and Glover felt that statistical study was especially appropriate for young women. These women became human computers, the clerical workers who did mathematical calculations before the advent of the electronic computer. Some of these women established prosperous careers in industry, in the U.S. Census Bureau, or in actuarial firms.

Glover's lab inspired Henry Reitz to organize a computing lab at the University of Illinois, where he worked until 1918. Like Glover, Reitz was interested in actuarial work and did projects for insurance companies in Chicago and Springfield.

The First Card Tabulators

The first statistical laboratories used mechanical adding machines such as those made by Monroe, Marchant, or Sunstrand. They began using punched-card tabulators invented by Herman Hollerith for the 1890 U.S. Census. Hollerith formed the Hollerith Tabulating Machine Company to manufacture and market these devices. This company merged with two other firms to form the Computing Tabulating and Recording Company, or C.T.R. In 1924, C.T.R. was renamed International Business Machines, or IBM.



In 1990, the U.S. Census requested agriculture reports, so the Bureau of Agricultural Economics—one of the divisions of the Department of Agriculture—used the punched-card tabulator. By May of 1923, the department consolidated all of its equipment into a single statistical office. The founder of this office was Howard Tolley, not a statistician by training, but a geodesist. Until 1918, he had worked for the U.S. Coastal Survey, where he learned the primitive statistical tools used to create maps.

Tolley and others at the Bureau of Agriculture were interested in the statistical methods of Frederick Winslow Taylor. Taylor was an engineer from Philadelphia, whose writings on scientific management were highly influential in the first decades of this century. He proposed a means of studying the methods of workers and developed some crude statistical techniques for

gathering and analyzing data. These techniques were filled with heuristic and ad hoc methods and were often criticized by Taylor's detractors. They were relatively effective at the time, however, and were studied by many managers who wished to improve production at their plant or in their office.

Tolley was interested in applying Taylor's ideas to fruit markets in New York and cold-storage warehouses along rail lines. Yet, he clearly understood the limits of Taylor's methods and knew that these statistical methods were unable to help him in situations with large amounts of variability, such as estimating crop production and weather damage. His training at the Coastal Survey helped him to understand the relationship between correlation analysis and least squares. During his early years in the Department of Agriculture, he worked to promote least squares analysis. Although researchers were generally interested in this method, they occasionally found it difficult to apply because the punched-card tabulators of the early 1920s were unable to multiply. Tolley apparently found a practical method to compute correlations that required both a punched-card tabulator and a desktop calculator.

The lab in the Department of Agriculture inspired two Iowans, George Snedecor and Henry A. Wallace, to experiment with punched-card statistical computations. Henry Wallace eventually rose to prominence as the vice president of the United States, but during the 1920s, he was the publisher of his family's farm journal, *Wallaces' Practical Farmer*. He was also a self-taught statistician and was interested in the interplay of biology and economics in farm management. During the 1910s, he learned the methods of correlation studies and least squares regression by reading George Udny Yule's book, *An Introduction to the Theory of Statistics* (London: Griffin, 1911). Finding in that book no easy method for solving the normal equations for regression, Wallace devised his own, using an idea that Gauss had applied to an astronomical problem.

In 1923, Henry A. Wallace learned of the new statistics lab at the Department of Agriculture while he was visiting his father, Harry Wallace, who was then the Secretary of Agriculture. Intrigued with the machines, he borrowed a tabulator at a Des Moines insurance firm and taught himself how to use the device to calculate correlations. He would punch data cards and then take them to the offices of the insurance company for tabulating. During the first years of the 1920s, he published more sophisticated statistical studies in the pages of *Wallaces' Farmer*. The last, published in January 1923, was a detailed study of land values in the state.

Wallace had become a friend of George Snedecor, who taught the statistics courses at Wallace's alma mater, then named Iowa State College. Impressed with Wallace's knowledge of least squares, Snedecor invited him to teach an advanced course on those methods to college faculty. This class, which met for 10 consecutive Saturdays over the fall and winter of 1924, ended with a demonstration of punched-card calculation. After the class, Snedecor helped Wallace prepare a manuscript on his algorithm for solving normal equations. They jointly published the manuscript in 1925 with the title *Correlation and Machine Calculation*.

The title of Wallace and Snedecor's pamphlet can mislead modern readers. For the most part, the machines the paper refers to are desk calculators, not tabulating machinery. By computing sums of squares and sums of cross-products, a mechanical tabulator could produce quickly a set of normal equations. The same tabulator, however, could not be used easily to solve these equations. It was extremely awkward, if not impossible, to use a 1920s vintage tabulator to solve matrix arithmetic problems. Such problems were solved by human computers who used desk calculators.

Inspired by Wallace, Snedecor devoted much effort to acquiring tabulating machines for his university. He was able to secure them in the fall of 1927 and established a statistical computing lab within the department of mathematics.

The Statistics Lab at Iowa State College

During 1927, Snedecor used the tabulating equipment for every possible application that he could find and presented a detailed report to his chair. He tabulated basic agricultural statistics, tracked the results of agricultural county fairs, and started a punched-card livestock breed book. A colleague used the tabulator to evaluate higher mathematical functions. Another interpolated a function with polynomials.

After a year of operation, Snedecor turned the lab equipment over to the management of one of his students, A. E. Brandt. Brandt had been a professor of farm mechanics at Oregon State University. He enjoyed the subtleties of the tabulators and liked to find new ways of doing calculations. From the economics department, he recruited human computers to help operate the machines and to solve normal equations for regression problems. One of these clerks, Mary Clem, would remain with the statistics lab for 50 years and be identified as the lead human computer of the group.

The computing facility was an important part of a lab that was quickly building statistical expertise. Through the Department of Agriculture, it acquired funds to host summer institutes in statistical theory. The first of these was held in 1927 with British statistician R. A. Fisher. Fisher met with about 50 researchers who were eager to learn his methods. One of these researchers was Henry A. Wallace, who would shortly thereafter leave Iowa and become secretary of agriculture, following in his father's footsteps. By then, Wallace had become fascinated with the problems of weather prediction and had begun a large study in which he attempted correlating heat, humidity, and wind direction with the position of the planets. The work eventually became an embarrassment to Wallace when his political enemies branded it as "weather astrology."

As secretary of agriculture, Wallace did become a champion of statistical studies as a means of planning programs to address social and economic ills. He devised and prepared the Agricultural Adjustment Act, a radical proposal to support farm prices and to alleviate the effect of the Great Depression on American agriculture. This program was the first piece of legislation in Franklin Roosevelt's New Deal, written and implemented within his first 100 days in office. It required the Department of Agriculture to undertake large statistical studies of major farm products, including cotton, corn, tobacco, and pork.

The Agricultural Adjustment Act proved to be a boon for land grant colleges and state experimental farms because they were asked to do much of the local statistical work. It was especially helpful to the Iowa State College Statistical Laboratory. The lab, now independent of the mathematics department, acquired several government contracts in the early days of the New Deal. As the demand for statistical work increased, the lab undertook increasingly larger and more important jobs. By 1936, it was negotiating with the Department of Agriculture to undertake major research projects, including a large master sample of the nation's farms. These projects increased the size of the lab. Over a short period of five years, its budget grew by a factor of 16.

John Atanasoff and an Early Electronic Computer

The rapid expansion of the lab allowed one Iowa State College faculty member to undertake some experiments in computation. That professor, John Atanasoff, held appointments in both the mathematics and physics departments. During the early 1930s, Atanasoff had been studying approximate solutions to differential equations. The last step of his approach required him to solve a large system of linear equations. Knowing that the statistics lab routinely solved such problems when it computed regression models, Atanasoff began to consider how such equations might be solved by using the lab's punched-card equipment.

Between 1934 and 1937, he and Brandt experimented with lab equipment. Brandt and Atanasoff modified an IBM tabulator to analyze an atomic spectrum. To do this, they constructed a special circuit that allowed a tabulator to compute all possible differences from a list of numbers. Once they had completed this experiment, they began to work directly on the problem of solving linear systems. Atanasoff sketched a design for the necessary circuits but never completed the task. Before

they made much progress on the project, Brandt left the lab to join the Bureau of Soil Conservation. With Brandt gone, Atanasoff undertook no further experiments. He ultimately decided that his design was too difficult to complete and abandoned it.

Even though Atanasoff lost interest in the punched-card machines, he did not forget about the problem of solving systems of linear equations. In what is a well-known story among computer historians, Atanasoff set off on a long drive across Iowa to think about this problem sometime during the winter of 1937-1938. Several hundred miles later, at a roadside bar in Illinois, he conceived the basic elements for a machine to solve systems of linear equations. The proposed machine had a lot of similarities with modern computers. It was electronic and had a memory unit, a central processor, and binary arithmetic. He built a small prototype of this machine in 1939 and prepared a proposal in 1940 for a full working model, a proposal he used to solicit funds for the machine.

Atanasoff saw his machine within the context of a computing lab and that it would solve linear systems "at low cost and for technical and research purposes." Atanasoff then listed nine possible applications for his machine.

Atanasoff built his machine between 1940 and 1942. With the start of World War II, he abandoned his creation when he left the college to join the Naval Ordnance Laboratory in Washington, DC. Atanasoff's machine would have remained in obscurity were it not for John Mauchly, one of the inventors of the ENIAC computer. Mauchly met Atanasoff at a conference in May 1941 and visited Atanasoff 's lab in Ames, where he studied Atanasoff 's machine. During this visit, Mauchly learned a great deal about electronics and about computing machines. Some of these ideas found their way into the ENIAC design and remained a point of contention between the two computer pioneers for the rest of their careers.

The Statistics Lab of the Cowles Commission

Although the statistical lab at Iowa State University became one of the major centers of statistical research, a similar lab at Indiana University grew to be a center of econometric research. This lab was founded by H. T. Davis and was championed by Colorado financier Alfred Cowles. Davis was a new member of the mathematics faculty when the dean asked him to form a statistical lab in 1927. A small local foundation had agreed to finance such a group so that they might better understand the fluctuations in the economy.

Like Snedecor, Davis was quick to use the lab to explore a diverse range of computational problems. One of the first was an optical interference computation for a physicist colleague. When he realized that his computers had used the wrong values in making the computation, Davis convinced his colleague to perform the experiment a second time, using the values that his computers had mistakenly employed. Although he did statistical calculations for business and economics professors, Davis took little interest in statistical research. Instead, he studied numerical analysis and used the lab to create tables of higher mathematical functions.

Cowles helped revive Davis's interest in statistics. An amateur statistician, Cowles was interested in large regression studies of the economy. He envisioned studies that would collect thousands of observations and fit regression models to 20 or 30 independent variables. In 1931, he approached Davis and asked him for help in undertaking such computations. Davis immediately realized that his small computing lab, staffed by human computers, would be unable to do the necessary computing. He urged Cowles to lease punched-card equipment and helped him establish a statistical lab near Cowles's offices in Boulder, Colorado. Davis spent the next several summers with Cowles, helping him develop the necessary mathematical techniques. This organization became known as the Cowles Commission.

The Last Days of Statistical Labs

World War II marked the glory days of the statistical lab and the start of its slow decline. The Office of Scientific Research and Development financed dozens of statistical projects and organized computing labs to find concrete numerical answers. At Columbia University, Abraham Wald operated a lab of 20 human computers to develop a theory of sequential testing. University of California statistician Jerzy Neyman worked with a New York computing group to help the Air Corps clear the Normandy beaches of mines in preparation for the D-Day landings. Neyman used a geometric model to estimate the number of mines that would survive being bombed by the Air Corps. The New York group, which had begun operation as a W.P.A. project, calculated the actual estimates.

As universities and corporations built centralized computing services, the statistical labs faded away. Their old calculators were shelved and the punched-card equipment returned to IBM as statisticians purchased computer time from a computing center. This trend was reversed in the early 1970s, when inexpensive minicomputers first appeared on the market. The popularity of the personal computer and the widespread availability of statistical software ensured that not only every department would have a computing facility, but that every statistician could do more computing in an hour than the old Department of Agriculture statistical lab could have done during the entire year of 1924.