SCANNING THE HEAVENS

Brian Hayes

Astronomy has become a statistics-bound discipline. In the time of Galileo, finding just four moons orbiting Jupiter was enough to overturn centuries of cosmological thought. As recently as 1929 Edwin P. Hubble detected the expansion of the universe by measuring red shifts in the spectra of just 18 galaxies. By the 1980s, however, when Margaret J. Geller and John P. Huchra were studying the large-scale distribution of matter in the universe, they had to plot the positions of more than 10,000 galaxies to establish the existence of cosmic sheets, filaments and voids. Gleaning further clues to the structure of the universe will require still larger data samples. To that end, a major new survey of the skies is now in preparation. It will catalogue some 50 million galaxies and about 70 million stars; redshift measurements will yield three-dimensional positions for a subset of about a million galaxies and 100,000 quasars.

The new survey, called the Sloan Digital Sky Survey, is being conducted by eight institutions (listed at the end of this article) with major funding from the Alfred P. Sloan Foundation. A new 2.5-meter telescope to be erected at the Apache Point Observatory in New Mexico will be dedicated to the survey. Although the new instrument has some novel and noteworthy features, the telescope is not the key innovation that will make the survey possible. The crucial factor is the technology for digitally recording large numbers of images and spectra and for automating the analysis, recognition and classification of those images and spectra.

A Census of the Skies

The idea of a sky survey is not new. In recent decades the most influential survey has been the National Geographic Society–Palomar Observatory Sky Survey, completed in the early 1950s. It catalogued all of the sky visible from the Northern Hemisphere on 936 pairs of 14-inch-square glass photographic plates. (One plate in each pair was exposed through a red filter and the other through a blue filter.)

Brian Hayes is a former editor of American Scientist. *Address: 211 Dacian Avenue, Durham, NC 27701. Internet: bhayes@mercury.interpath.net.*

The photographic plates of the Palomar survey have been an essential resource for the astronomical community for the past 40 years. Whenever something new turns up in the skies—say a compact radio source, or an x-ray star—one of the first things astronomers do is check the Palomar survey plates to see if the object has a visual counterpart. But photographic plates have grave limitations when it comes to the analysis of large numbers of galactic and stellar images. The galaxies and stars are there by the millions in the silver-halide emulsion, but counting and classifying those images by visual inspection is a formidable task.

One solution is to digitize the photographic plates. This project has been undertaken by Roberta M. Humphreys and her colleagues at the University of Minnesota, using a special-purpose laser scanner. They have scanned all the plates except those close to the plane of the Milky Way, where the clutter of stars and dust makes analysis difficult.

The SDSS, using digital methods from start to finish, will extend to objects 10 to 20 times dimmer than the faintest ones included in the Palomar survey. Instead of recording just two color bands, the survey will measure brightness in five ranges of wavelengths (ultraviolet, green, red and two infrared bands). And, most important, whereas the Palomar survey is strictly two-dimensional—it shows the positions of objects projected on the sky but tells nothing of their distance along the line of sight—the SDSS will provide three-dimensional coordinates for a large subset of galaxies and quasars.

Capturing Photons

Instead of photographic film, the recording medium for the SDSS will be silicon, specifically the detector chips called charge-coupled devices, or CCDs. A CCD is a rectangular array of cells that generate electric charge when they are exposed to light. In the usual mode of operation, the charge is allowed to accumulate in place for some time—this is analogous to the exposure time in a conventional photograph—and then it is read out by passing the packets of charge bucket-brigade style from one cell to the next along each column, and then from column to column to a

sense amplifier, which measures the charge in each packet.

A CCD is an almost ideal astronomical detector. It is highly sensitive, with a quantum efficiency at some wavelengths of 50 percent or more, meaning that one electron is liberated for every two photons intercepted. (The efficiency of photographic emulsions is less than 1 percent.) The silicon detector also has a wide dynamic range: At low temperature each cell emits only a dozen or so electrons in the dark, yet it can hold a few hundred thousand electrons before it saturates in bright light. Unlike photographic film, the response of a CCD is highly linear—the number of electrons emitted is nearly proportional to the number of photons absorbed. The range of wavelengths is also wide, extending from the infrared into the ultraviolet. And spatial resolution is excellent; the CCDs for the Sloan survey have a pixel size of 24 microns, which is roughly equal to the resolution limit of the Palomar survey plates.

In one respect silicon still cannot match photographic emulsions: There is no prospect any time soon of making a 14-inch-square CCD. The main CCDs for the SDSS are a little less than two inches across, and at that size they are enormous by the usual standards of the chip-maker's art. Each of the chips has a square array of 2,048 (or 2^{11}) columns and rows, and thus somewhat more than four million pixels. The imaging camera requires 30 of these chips as well as two dozen smaller ones, and the spectrographs will use four more large chips. It is noteworthy that the CCDs will be the most costly component of the survey apparatus, overshadowing even the telescope and its enclosure. The cost would have been higher still if the survey strategy (to be explained below) did not allow the use of chips with certain imperfections.

The SDSS actually consists of three coordinated surveys. The photometric (or brightness-measuring) survey bears the closest resemblance to a conventional photographic survey, in which one of the final products is a mosaic of images of the sky. The astrometric (or star-measuring) survey aims to determine the precise geometric positions of stars and quasars and thus to provide a coordinate frame for the other surveys. The spectroscopic survey will characterize in greater detail the emissions of selected galaxies and quasars, measuring brightness as a function of wavelength over the range from 4,000 to 9,000 angstroms. One important reason for recording spectra is to determine the red shift—the extent to which emission or absorption lines have been shifted toward longer wavelengths, mainly by motions associated with the expansion of the universe. Since the work of Hubble, it has been recognized that red shift is correlated with distance, so that knowing red shifts allows one to construct a three-dimensional map of the universe.

Staring vs. Scanning

The Palomar survey plates, like most astronomical photographs, were made in "staring mode": The camera was pointed at a region of the sky and was held still for the necessary exposure time (which actually entailed moving the telescope to compensate for the earth's rotation). One can imagine another approach, in which the telescope slews across the sky as it is carried by the earth's sidereal motion, and the photographic plate is moved across the focal plane in step with this motion, keeping each image at a fixed point on the emulsion. Such a "scanning mode" is impractical with conventional photography (except in the cameras that record a photo finish at the race track), but it is easy to implement with a CCD. There is no need to physically move the detector chips across the focal plane; instead the telescope is aimed so that images move parallel to the columns of cells on the chip, and the packets of charge are shuttled from cell to cell to keep pace with them. When the image reaches the edge of the chip, the accompanying charge is measured. Scanning mode offers a number of advantages. There is no need to stop at intervals and "change the film"; data can be recorded continuously across a long strip of sky. More important, single-pixel defects in a chip cause negligi-

Figure 1. Strip of sky scanned by a prototype of the SDSS photometric camera shows part of the Coma cluster of galaxies; the brightest galaxy, left of center, is NGC4889. (Image courtesy of Tim McKay, Fermilab.)

ble degradation of the image, since every pixel is averaged over 2,048 CCD cells.

In the camera being built for the photometric and astrometric surveys 30 large CCDs are arranged in six columns and five rows. Images progress down the columns, moving from chip to chip. Each row of chips is equipped with an optical filter that passes a different band of wavelengths. The spacing between columns of chips is a little less than one chip width, which means that the sky will be scanned in narrow stripes, like the interlaced scan lines of a television image, and two passes (with a slight offset) will be needed for full coverage of a strip.

In addition to the 30 large CCDs, the camera will include 24 smaller chips, which have the same 2,048 columns of cells but only 400 rows. Twenty-two of these chips are for the astrometric survey. Fewer rows implies a shorter exposure time in a camera operating in scanning mode, but the briefer exposure is acceptable because the astrometric sensors will examine only relatively bright stars. The remaining two small chips monitor the focus of the optical system.

Recording the spectrum of a galaxy is harder than capturing its image. Light from the galaxy must be focused onto a slit and then dispersed according to wavelength by a prism or diffraction grating. Recording a single spectrum can take from several minutes to several hours, so doing them one at a time is not a practical option if you want to gather a million spectra. For the SDSS an ingenious trick allows the spectra of up to 600 objects to be recorded at once. The trick—already an established technique employed in several smaller surveys—is to capture the light from each galaxy with an optical fiber, which performs the function of the entrance slit in a conventional spectrograph. Bundles of 300 fibers are brought to each of two spectrographs, where the light from each fiber is spread out into a fan-like spectrum that illuminates a few columns of a CCD. How are the fibers positioned to intercept light from the selected galaxies and quasars? It will be done by drilling aluminum "plug plates," with holes at the appropriate positions. The hundreds of fibers coming from the spectrographs are plugged into the holes, and then the plate is installed in the focal plane of the telescope. Some 3,000 of these plates will be needed.

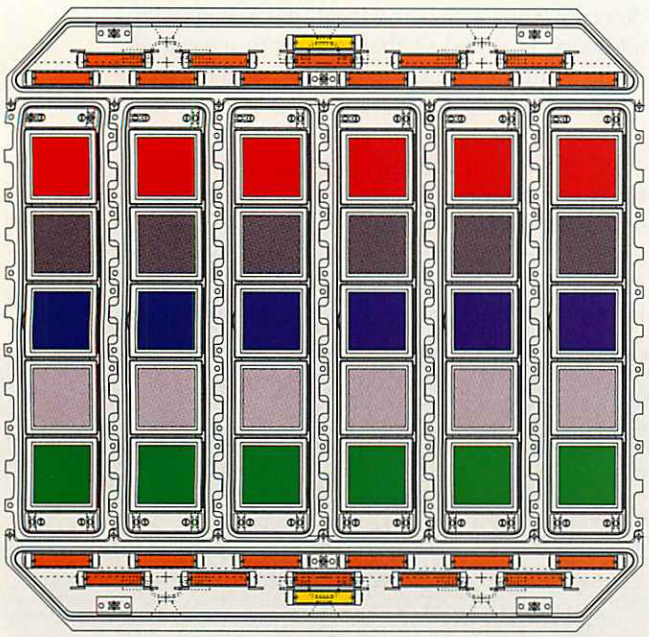

Figure 2. Array of chips serves as the primary detector in the digital sky survey. The large chips for the photometric survey are arranged in six columns and five rows. Each row will be exposed through a different optical filter; proceeding from top to bottom the filters will pass red, infrared, ultraviolet, near infrared and green wavelengths. The 22 smaller chips shown in orange are for the astrometric survey; the two chips shown in yellow monitor focus. (Adapted from a drawing by James E. Gunn, Princeton University.)

Drilling the holes in the plug plates is a straightforward task for a computer-controlled machine tool, but plugging the fibers into the holes is another kind of challenge. Machines tend to be awkward at such work. Human hands can deftly place the fibers, but they cannot be relied on to get 300 fibers in the right holes, without missing any, and then repeat the performance 3,000 times. The solution adopted by the SDSS team is to plug the fibers manually but not worry about getting them in the right holes; workers will assign fibers to holes at random, only making sure that every hole has a fiber. Then a laser will scan across the fibers at the prism end (where they are lined up in a neat row) and a video camera will detect which plate position is associated with each fiber.

The Sloan survey will cover the northern galactic cap—the region of sky north of the Milky Way—which amounts to about a quarter of the celestial sphere. During the autumn months when this part of the sky is inaccessible from New Mexico, the telescope will repeatedly scan a narrow stripe of sky in the southern galactic hemisphere, which will thereby be surveyed deeply. After a year of testing, expected to begin in the spring of 1995, the survey should require five years of observing.

Stars and Galaxies by the Terabyte

Each pixel in the photometric survey will be encoded in 16 bits (or two bytes) of data, allowing for about 65,000 brightness values. The angular size of the pixels is 0.4 seconds of arc squared, and the total survey area encompasses roughly 10,000 degrees squared. A back-of-the-envelope calculation shows that the survey area will be divided into approximately 8×10^{11} pixels, each represented in five colors. At two bytes per color per pixel, the complete data archive must amount to at least 8×10^{12} bytes, or 8 terabytes. In fact there are many areas of overlap in the survey, so that almost 50 percent of the pixels will be surveyed more than once, and as a result the actual data volume for the photometric survey will be roughly 12 terabytes. The spectroscopic data

base will be rather small when measured by this standard: only 3.6×10^{11} bytes, or 360 gigabytes.

During photometric observations the stream of data from the CCDs will be buffered in six large disk drives and then written onto tape cartridges that hold more than five gigabytes. Compression algorithms will reduce the volume of data by a factor of two or more, but even so a full night's observing will produce up to 90 gigabytes of data, or about a dozen tapes. The tapes will be shipped by overnight courier to the Fermi National Accelerator Laboratory near Chicago, which has experience with even larger data flows in high-energy physics experiments.

Much of the data analysis can be done off-line, without any severe time constraints, but a few important tasks must be synchronized with the survey schedule. In particular, the results of the photometric survey will be used to select galaxies and quasar candidates for the spectroscopic survey, and so the computer programs that distinguish among these various kinds of objects will have to keep up with the flow of data from the telescope. The aim is to extract the list of spectroscopic targets from a strip of sky within a week.

Distinguishing a galaxy from a star is quite easy in the case of nearby galaxies, which can extend over an area of sky as large as the full moon. But most of the 50 million galaxies detected by the survey will be so distant that they will cover only a few pixels on the CCD detectors. Making matters worse, stellar images are not perfectly pointlike, but are spread out to some extent by atmospheric effects and the telescope optics. Nevertheless, algorithms for discriminating between stars and galaxies work with high reliability; they are based on statistical measures of both fuzziness and asymmetry in the image. Distinguishing quasars from stars calls for a different technique; by definition quasars (or quasi-stellar objects) yield a starlike image, and so geometric criteria are useless, but the color of a quasar (as recorded in the five bands) is highly distinctive.

Accurate classification of objects and selection of targets for spectroscopy is crucial to the success of the survey. The aim is to collect a large sample of distant objects with as few systematic biases as possible; in particular, all galaxies brighter than a certain limiting magnitude are to be included in the spectroscopic sample. One possible bias in the selection process comes from the dimming and reddening of objects by interstellar dust, which is more abundant near the plane of the Milky Way; adjustments for this effect must be built into the selection algorithm. Other biases could creep into the data through peculiarities of the survey itself. For example, spectra cannot be recorded simultaneously from two galaxies that are closer than three millimeters in the focal plane of the telescope, because that is the thickness of the sheathed optical fibers inserted into the plug plates. If no steps were taken to overcome this problem, the survey sample would be deficient in galaxies that have close neighbors. The solution is to make the spectroscopic fields overlap enough that most members of close pairs can be recorded in separate passes over the sky.

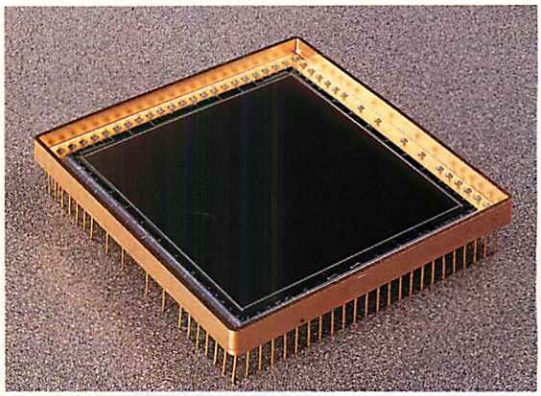

Figure 3. Charge-coupled device is an array of 2,048 × 2,048 pixels. The active area of the chip is a square 4.5 centimeters on a side. (Photograph courtesy of Scientific Imaging Technologies, manufacturer of the chip.)

Subsequent analysis of the data will greatly refine the distinctions made during on-line processing. Algorithms will attempt to classify galaxies according to their morphology (spirals, ellipticals, etc.) and color. The spectra will also be classified, and red shifts will be calculated. The stars too can be categorized from the relative distribution of radiant energy in the five color bands, distinguishing red giants, for example, from main-sequence stars and white dwarfs. The product of this analysis will be a data base of more than 100 million objects, which astronomers will be able to explore in ways that would have been unthinkable with photographic materials. The fact that the data base has been compiled almost entirely by algorithmic methods rather than by human judgment offers a peculiar collateral benefit. Even if the selection and classification algorithms are not perfectly correct in

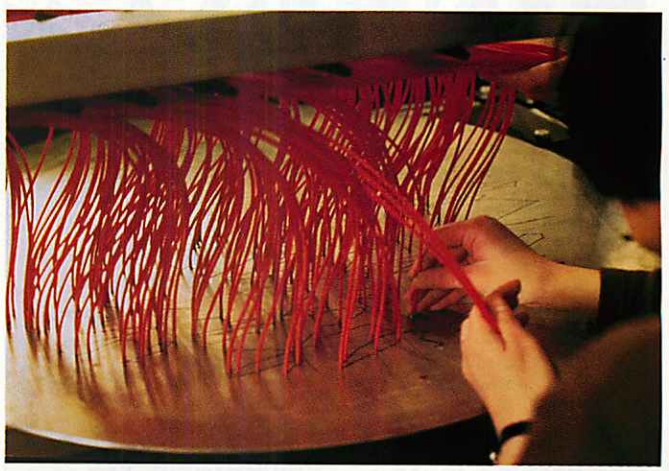

Figure 4. Optical fibers plugged into drilled aluminum plates will carry light from the focal plane of the telescope to two spectrographs. Here an early trial of the plate-plugging operation is shown. (Photograph courtesy of Astrophysical Research Consortium.)

their decisions, they are perfectly consistent, and so the criteria behind those decisions can be stated precisely. It will always be possible to find out exactly why an object has been included in or excluded from a given class.

The 12 terabytes of raw survey data will be preserved at Fermilab, but various condensations of it will surely prove more useful for most purposes. An atlas of images will include a full catalogue of all the objects detected by the survey along with a "postage stamp" image of each object and its immediate surroundings. In compressed form such an atlas might occupy 80 gigabytes, or about 130 CD-ROMs. The catalogue without the images might fit on 25 CD-ROMs. A reduced-resolution pixel map of the entire survey area along with an archive of all the recorded spectra would fit on about 250 CD-ROMs. Compared with the Palomar photographic plates, the cost of duplicating and distributing all of this information is negligible.

The Urge to Catalogue

The premier scientific problem addressed by the Sloan survey is understanding the large-scale structure of the universe. The frothlike voids and sheets discovered by Geller and Huchra are on a scale of a few hundred million light-years, and there are tantalizing hints of even vaster structures, perhaps an order of magnitude larger. At the largest scale, however, the universe is expected to be homogeneous and isotropic—essentially featureless. Thus the SDSS may reveal the largest objects in existence. The results could have a direct bearing on the most vexed and vexing question in modern cosmology: What is the unseen "dark matter" that seems to guide the motions of galaxies?

The survey will doubtless be put to many other purposes as well. Quasar studies—notably the search for patterns in their distribution and for evidence bearing on their early evolution—could be transformed by a well-defined sample of 100,000 objects (100 times more than most earlier samples). Astronomers interested in the shape, the evolution and the interactions of galaxies will have a similar bonanza of new data, and the survey will even illuminate the distribution of matter in the halo of our own galaxy.

These contributions to knowledge are more than enough to justify the effort being expended on the SDSS, but the survey can also be seen in a broader context. The present moment is one when the encyclopedic cataloguing of resources appears to be the fashion in many areas of science. Molecular biologists have their human genome project. The National Biological Survey is taking inventory of all the flora and fauna of the U.S. In the earth sciences, the satellites of the Earth Observing System will stare down at the planet with the same implacable and all-comprehending gaze that the SDSS telescope will direct upward. This is not to suggest that astronomers are merely following a bandwagon; nor is it to suggest that scientists everywhere have been struck by some sudden impulse to count, collect and classify everything in sight. More likely the impulse has always been there, and what has come rather suddenly is the ability to act on it.

Bibliography

Note: The Sloan Digital Sky Survey has eight sponsoring institutions: the University of Chicago, the Fermi National Accelerator Laboratory, the Institute for Advanced Study, the Japan Promotion Group, the Johns Hopkins University, Princeton University, the U.S. Naval Observatory and the University of Washington. The survey will be conducted under the auspices of the Astrophysical Research Consortium (ARC), which operates the Apache Point Observatory. The SDSS subcommittee of the ARC is chaired by Jeremiah P. Ostriker of Princeton; the project director is Donald G. York of Chicago; the project scientist is James E. Gunn of Princeton; the chair of the Science Advisory Committee is Neta A. Bahcall of Princeton; and the survey director is Richard G. Kron of Chicago and Fermilab.

A Digital Sky Survey of the Northern Galactic Cap. Volume I, Science. 1994. A proposal to the National Science Foundation. University of Chicago, Department of Astronomy.

Kron, R. G. 1992. A wide-field telescope for high-latitude surveys—CCD multiband photometry and multifiber spectroscopy. In *European Southern Observatory Conference on Progress in Telescope and Instrumentation Technologies*, ed. M.-H. Ulrich. ESO Conference and Workshop Proceedings No. 42, p. 635.

Gunn, J. E., and G. R. Knapp. 1993. The Sloan Digital Sky Survey. In *Sky Surveys: Protostars to Protogalaxies*, B. T. Soifer (ed.). Astronomical Society of the Pacific Conference Series Vol. 43, p. 267.