*A New Era in Bioastronomy*
*ASP Conference Series, Vol. 213, 2000*
*G. Lemarchand, K. Meech (eds.)*

# Generalized SETI in a Virtual Observatory

Stanislav G. Djorgovski

*Palomar Observatory, Caltech, Pasadena, CA 91125, USA*

**Abstract.** The advent of new large digital (multi-terabyte) sky surveys and archives, and the data-mining tools for their exploration will provide new opportunities for SETI. A major class of studies using federated, multi-wavelength digital sky archives will involve systematic searches for sources with unusual or anomalous properties, in terms of their spectrum, variability, etc. The concept of SETI can be generalized as a search for such outliers in the parameter spaces of observable source properties. SETI would thus be just one of the many related applications in a national (or global) virtual observatory.

## 1. Introduction: Generalizing the Concept of SETI

The central technical challenge of the traditional (radio-based) SETI is the problem of the "cosmic haystack": discovery of an artificial signal pattern in a vast parameter space, whose axes are usually assumed to be the two celestial coordinates, the frequency, and the modulation (and implicitly the time). SETI has been expanded to other wavelengths (e.g., optical) in searches for manifestations of extraterrestrial civilizations (ETCs), either as intentional signals or as undirected technological signatures, e.g., Dyson spheres in the far-IR, etc.

While the traditional radio-based, and now also optical SETI are certainly worth pursuing, they do bear a clear imprint of the present technological and cultural instant in the history of our own civilization. A more open and less historically or technically biased approach would be desirable: we do not know what kind of manifestations of ET technology or civilizations are most likely to be apparent to us, or easiest to detect or recognize with our existing technology and our limited imagination and intelligence.

It then makes sense to *generalize the concept of SETI as a search for somehow anomalous or unusual patterns or individual sources in any large body of astronomical observations.* In this way, SETI becomes one of the aspects or scientific goals of what is becoming a major new thrust in astronomy, data mining of large digital sky surveys, and searches for rare or new types of astronomical objects and phenomena in them. In a broader sense, we are now exploring a full parameter space defined by a number of observable axes, as discussed, e.g., by Harwit (1975) and Harwit & Hildebrand (1986). In addition to the probable discoveries of new, natural astronomical phenomena or new or peculiar types of astronomical sources, it is possible that some artificial manifestations of ETCs may be found in such searches. This novel approach is being enabled by the vast increase (several orders of magnitude) in the data volumes now generated in astronomy, especially in a number of large, digital sky surveys; and by the corresponding development of computing tools to explore these enormous data sets and parameter spaces generated though their cataloguing and analysis.

## 2. New Astronomy with Multi-Terabyte Digital Sky Surveys and Archives: Towards a Virtual Observatory

We are at the start of a new era of information-rich astronomy. Several ongoing sky surveys over a range of wavelengths are now generating data sets measured in many terabytes, and producing (multi-TB) catalogs of billions of sources, with up to a hundred measured numbers for each object: the increase is both in the *volume* and the *complexity* of the available information. For comparison, the entire human genome is only about 3 GB in size.

Some current examples include the Digital Palomar Observatory Sky Survey (DPOSS), or the Sloan Digital Sky Survey (SDSS) in the visible light, the Two-Micron All-Sky Survey (2MASS) or DENIS in the near-IR, the NRAO VLA Sky Survey (NVSS) and the Faint Images of the Radio Sky at Twenty centimetres (FIRST) in the radio, etc. Many others are now under way or are planned. A useful set of relevant links may be found at `http://digital-sky.org`

Generating multi-terabyte data sets will soon become a norm in the observational astronomy. For the first time we have data sets which are so information rich that they can support scientific studies extending well past their original purposes. Even the surveys or experiments designed with very specific scientific goals in mind (e.g., MACHO searches) can produce a rich body of information which can be used for a multitude of other studies. Large SETI experiments themselves also belong in this class.

An even more powerful data framework can be obtained by federating multiple surveys, e.g., over a wide range of wavelengths, with repeated observations, etc. There is now a growing interest in the development of such a National (and eventually Global) Virtual Observatory (NVO). This will become the fundamental information infrastructure for the astronomy and astrophysics at the turn of the century and beyond. The initial work along these lines has already started by several groups. Some of the relevant discussion can be found in the paper by Szalay & Brunner (1999).

These vast new data sets will enable scientific investigations which were not practically possible with much more limited data sets in the past. Combining surveys done at different wavelengths, from radio to x-rays, would provide a new, panchromatic picture of our universe the objects in it. Of special interest for the generalized SETI may be the surveys which will do repeated observations of the same regions of the sky, looking for variable or fast-moving objects, e.g., the searches for near-Earth asteroids, or searches for transient phenomena such as supernovae, $\gamma$-ray bursts, or their afterglows.

The large size of these data sets will also enable searches for extremely rare types of astronomical objects, and may well lead to the surprizing new discoveries of previously unknown types of objects or astrophysical phenomena. Moreover, such searches can be performed in a fully automated and systematic manner. This is directly relevant to the generalized SETI.

## 3. Data Mining the Sky: Searching for Outliers in Parameter Spaces

This great scientific opportunity comes with a commensurate technological challenge: how to manage, combine, analyze and explore these vast amounts of information, and to do it quickly and efficiently? The technical issues include optimal storage and access to such complex data, combining sky surveys done at different wavelengths, visualization, searches through them, etc.. Such problems are common to all data-intensive fields, and require a development of a new generation of computing and data management and exploration tools.

There is a great deal of interest and ongoing research in such software technologies in the computer science community, known as Knowledge Discovery in Databases (KDD). See for example, the volume edited by Fayyad *et al.* (1996), or the links available at `http://www.acm.org/sigkdd/links.html`

Examples of the techniques of interest include various clustering analysis methods (whereby the data are associated with a finite number of clouds in the parameter space of measured properties), unsupervised classification methods (algorithms which determine how many statistically distinct classes of objects are there in the parameter space, and divide the objects accordingly), automated pattern recognition, Bayesian networks, intelligent software agents, a variety of advanced visualization techniques, etc.

A general procedure may be to use some unsupervised classification or clustering technique to partition a large parameter space of measured source properties into a number of clusters (multi-dimensional data clouds) which contain the bulk of all detected sources, e.g., normal stars or galaxies of different spectroscopic types and morphologies, etc. Statistically significant deviant sources outside the major clusters may be then selected for further studies and analysis.

Simple but very successful initial applications of such techniques in DPOSS, 2MASS, and SDSS include discoveries of numerous high-redshift quasars, brown dwarfs, and other objects with peculiar spectra, all of them found as outliers in the parameter space of colors. This usually involves classification of all detected objects in a given survey as stars (unresolved) or galaxies (resolved), on the basis of a morphological information present in the images, using some objective, automated, supervised classifier method such as Decision Trees or Neural Nets. One then forms colors (flux ratios) from the survey bands. Among the unresolved sources, normal stars form a well-defined locus of points in the color space, which shifts and changes shape as a function of the magnitude. Objects with extreme colors, e.g., brown dwarfs, or objects significantly displaced from the stellar locus in any direction (usually quasars) can then be selected for follow-up spectroscopy. Adding the data from as many wavelengths as possible clearly increases the discriminatory power and scientific scope of such searches. An example in the SETI context are searches for Dyson spheres as FIR excess objects.

An even more interesting approach involves exploration of data sets with some time-domain information. This can be used to find objects which vary by a large amount over some time scales, including one-time transients, and to search for objects with unusual proper motions (e.g., on peculiar orbits within the Solar system; cf. Valdes & Freitas 1983).

A detailed, multi-wavelength mapping of the sky and the properties of "ordinary" sources would define objectively where peculiar (including possibly artificial) sources may be found in the parameter space of observables. This in turn can be used to design new surveys to explore specifically heretofore poorly covered, or apparently empty regions of the parameter space.

## 4. Concluding Remarks: SETI in a Virtual Observatory

Astronomy is becoming immensely data-rich. A variety of ongoing or forthcoming digital sky surveys over a broad range of frequencies and time scales will provide a systematic and repeated coverage of the sky unlike anything we had in the past. The tools to explore these vast new data sets quickly and efficiently are being actively developed. This will enable a major new thrust of observational astronomy, based on the data mining of federated digital sky surveys and archives, e.g., in the framework of the NVO.

A major class of experiments with these new data sets and archives will involve searches for rare or unusual types of sources, separated from the bulk of the "ordinary" astronomical sources in some subset of the measured parameter space. This may include parameters such as the broad-band (or sometimes a high-resolution) spectra and colors, variability on a range of time scales, proper motion (if any), in some cases polarization, etc. Some of these outliers may turn out to represent intentional or unintentional manifestations of ETCs.

This will be a new way of doing observational astronomy: with a computer and a data archive, rather than a telescope. Any scientist or a student (or even a serious amateur) anywhere, with a network connection and a clever idea would be able to do a first-rate science, perhaps also including some novel SETI experiments. A generalized SETI would thus become a part of the mainstream data-mining astronomy. Moreover, the free access to such major data sets and exploration tools over the web would enable a broader academic and amateur community to participate in such research, testing a wide range of ideas and approaches – some of which may even be successful.

**Acknowledgments.** This work was supported in part by the Bressler Foundation. The author would like to thank R. Brunner, T. Prince, A. Szalay, R. Gal, and many other NVO enthusiasts for stimulating discussions.

## References

Fayyad, U., Piatetsky-Shapiro, G., Smyth, P., & Uthurusamy, R. (eds.) 1996, *Advances in Knowledge Discovery and Data Mining*, Cambridge, MA: AIII Press/The MIT Press

Harwit, M. 1975, *QJRAS*, **16**, 378

Harwit, M., & Hildebrand, R. 1986, *Nature*, **320**, 724

Szalay, A., & Brunner, R. 1999, in *Future Generation og Computer Systems*, Elsevier, North Holland, **16**, p. 63 http://www.arXiv.org/abs/astro-ph/9812335

Valdes, F., & Freitas, R. 1983, *Icarus*, **53**, 453