Cyberinfrastructure for the Social and Behavioral Sciences

Social and behavioral science lies at the intersection of computer vision, database design, psycholinguistics, cognitive and social neuroscience, psychology, linguistics, education, anthropology, sociology, and high-speed computing and networking.

By Bennett I. Berterenthal, PhD

Social and behavioral science requires the ability to compare, measure, and search for patterns in semi-structured and heterogeneous data. The challenge is to integrate information over time, place, and types of data in order to scale up the opportunities for comparisons. Once these diverse datasets are integrated, tools are necessary for annotation and analysis of the different data types, including voice, video, images, text, and integer and real numbers.

Currently, investigators studying the neural, cognitive, and social behaviors of humans lack the tools to assess multiple measures at multiple levels simultaneously and to store and analyze these measures in a common database. Significant conceptual, technical, and analytic advances are necessary for understanding multimodal human behaviors at different time scales. This new field lies at the intersection of computer vision, database design, psycholinguistics, cognitive and social neuroscience, psychology, linguistics, education, anthropology, sociology, and high-speed computing and networking. Successful collaboration among these diverse disciplines requires a "data interface" (e.g., shared datasets and databases), a "service interface" (e.g., shared tools for analysis), and an intellectual interface (e.g., shared problems and theories) to support multidisciplinary research.

In response to this need, my colleagues and I are developing a Social Informatics Data (SID) Grid that will enable researchers to collect real-time multimodal behavior at multiple time scales. Multimedia data will be stored in a distributed data warehouse that employs Web- and Grid services to support data collection, storage, access, exploration, annotation, integration, analysis, and mining of individual and combined data sets. Although a number of groups are currently developing large data archives for research purposes, our project is unique because it is focused on streaming data, such as videos and physiological measures, which change over time. The ability to investigate multiple measures of human behavior simultaneously and at different time scales is essential to understanding how behavior is dynamic, multi-level, and multi-causal.

Digital Data Collection and Testing Dynamic Models

A primary component of this project is the development of a core facility, referred to as a SuperLab, equipped for monitoring the physiological and behavioral responses of participants during experimental and observational studies. Data to be collected in this type of lab will include: frame synchronized multi-camera video, multi-channel audio, motion capture, eye movements, electrophysiological measures, such as electromyograph (EMG), electromyograph (EMG), heart rate (ECG), and respiration, body movements, and bioassays, such as cortisol or oxytocin. Each of these data streams will be sampled at different rates ranging from .01 Hz to 44.1 kHz, and will be collected within the same experimental session. Currently, no laboratory in the social or behavioral sciences is equipped to concurrently measure more than two or three responses simultaneously. The SuperLab will enable digital data collection at a scale that far exceeds current practices.

As a specific example of how data collection in a SuperLab will affect the modeling and analysis of time-sampled data, consider how the underlying dynamics of an observable social behavior can be investigated with the acquisition of multiple measures at multiple time scales. In any dynamic system, there are several "state" as well as observable variables that are changing with time, but their relations are rarely measured directly. In this new paradigm, one set of measures will record the temporal features of observable behaviors (i.e., coding of video and audio, eye movements) in a social context, whereas a different set of measures (e.g., heart rate, facial EMG, EEG, salivary cortisol) will investigate the "state" variables that underlie these behaviors. By temporally aligning the observed variables with the state variables, new experiments conducted in the SuperLab can move beyond the investigation of observable variables to incorporate underlying "state" variables and their dynamic interactions with the observable variables.

In theory, this data collection scenario could be implemented today, but the sheer volume of data and number of inputs that need to be controlled and synchronized represents a huge challenge to even the most technologically sophisticated labs in the social and behavioral sciences. Our goal is to essentially establish the SuperLab as a public facility that could be shared in the same way as telescopes or particle accelerators are shared by researchers in the physical sciences. As a complement to this lab, we plan to provide an ensemble of network and grid services that will significantly reduce the burden for individual users who want to collect multimodal data or annotate and analyze previously collected corpuses of data stored in centralized or distributed data archives. These software tools and data will be bundled together in a publicly accessible data grid.

Overview of the Social Informatics Data Grid

Data collected in the SuperLab or through any other facility (such as a Brain Research Imaging Center) will be stored in a distributed database with tools provided by our Social Informatics Data (SID) Grid. Consistent with the recommendations of the Atkins Report, the SID Grid is designed as a multi-tiered architecture for long-term, distributed, and stable data and metadata repositories that institutionalize community data holdings. As can be seen in Figure 1, the types of facilities and services to be provided at the most general level (shaded layers) will generalize across different disciplines and project-specific applications. At the next layer, cyberinfrastructure support will be customized for specific disciplines and projects. Thus, the SID Grid will include both domain general as well as domain specific applications and services.

The goal in developing the SID Grid is to design a resource structure that generalizes to all behavioral and social science research domains including tools to transform and analyze heterogeneous data types. For example, the National Center for Data Mining (Robert Grossman, Director), which is a partner in this project, has already contributed to the development of XML standards for several common data preparation operations, and will continue to develop additional standards during the project period with a special focus on data preparation services for transforming time series data and streaming data.

Although we eventually plan to develop a data grid that will generalize across all social science domains, it is necessary to begin by focusing on specific domains, because some of the resources required by the SID Grid are specific to the needs of a particular research community. During the initial funding period, we will focus on resources for three specific domains: Multimodal communication, cognitive and social neuroscience, and neurobiology of social behavior in humans and animals. Previously collected corpora and data archives in raw or partially analyzed forms will be integrated into the SID Grid via a translation layer. For example, we have already negotiated with Brian MacWhinney to make TalkBank and the Child Language Data Exchange System (CHILDES) accessible through the SID Grid. We also plan an interface that will integrate software tools developed by domain experts to provide additional services for annotation and analysis. Annotation tools, such as Elan and MacVisSTA (see Figure 2), will soon be accessible via the SID Grid. Although these tools were created for annotation of multimodal communication, we are planning to expand their functionality so that they will be available for storing, displaying and coding additional types of data, including eye movements, physiological measures, and motion tracking in a time synchronous fashion. Our goal is not to "reinvent the wheel" but to incorporate existing databases and tools that are currently available and make them accessible through a common website. By doing so, we can provide interfaces for using data and tools interchangeably, which will expand applications beyond those that were intended by the original developers.

Collaboration and Standardization

Another motivation for creating the SID Grid is to provide new and more expansive opportunities for collaboration. We are developing a multi-platform, multi-task tool for working with and manipulating time-synchronized data. This interface tool, built to use SID Grid services, will provide the means for collaborative annotation and analysis of SID Grid data sets. Results, hypotheses, derived data streams, annotations, and metadata will be available over the Grid to widely distributed communities of collaborators.

The SID Grid will be complemented by support for a virtual collaboration environment based on the Access Grid (AG) version 2.0 software developed by the Stevens' lab at Argonne National Laboratory and University of Chicago. In essence, the AG is an ensemble of network and computing resources that supports group-to-group human interaction in real-time across the grid. It consists of large-format multimedia displays, presentation and interactive software environments, interfaces to AG middleware, and interfaces to remote visualization environments. The Access Grid is already deployed at over 200 research and development sites worldwide. It is being used on a daily basis to conduct distributed meetings, seminars, and virtual conferences. Since the AG is based on open source software and uses the Internet for its streaming data transport, it is not uncommon for groups to leave their AG nodes operational 24 hours a day thereby creating a persistent shared working environment between multiple sites.
The creation of these cyber tools will provide some of the needed infrastructure for supporting collaborative research in the social and behavioral sciences. This infrastructure will encourage data sharing and accelerate the development of standards for collecting and coding physiological and behavioral data. For purposes of outreach and dissemination, we are creating a website with tutorials for using the SID Grid, organizing workshops on use of the infrastructure, and soliciting researchers to conduct multimodal research in the SuperLab. The availability of these databases and software tools could change how we educate the next generation of scientists.

Opportunities and Challenges

The resources and services available through grid computing are already transforming fields, such as particle physics and bioinformatics, but they involve dedicated partnerships between domain experts and computer scientists. If the tools developed for the social and behavioral sciences are to have a comparable impact, then it is necessary that they are designed to facilitate research without prescribing how research should be conducted. This goal represents a delicate balance between meeting the needs of the research communities and developing new tools that are compatible with a grid infrastructure.

Our approach to achieving this goal is to solicit the advice and feedback of the research communities by recruiting working groups of domain experts who are willing to meet once or twice a year with the developers of the SID Grid to prioritize needs and offer feedback on the tools already developed. We are currently in the process of forming three working groups to assist in the development of specialized tools for multimodal communication, cognitive and social neuroscience, and neurobiology of social behavior, but plan to recruit additional working groups as we expand our efforts in the future. The long-term impact of this project should be a better understanding of how social scientists and computer scientists can work together to develop the field of “social informatics”.

About the Author

Bennett I. Bertenthal is a Professor of Psychology and Computational Neuroscience at the University of Chicago. He is also a Senior Fellow of the Computation Institute at the University of Chicago. Prior to this appointment, he was the Assistant Director of the Social, Behavioral, and Economic Sciences (SBE) Directorate of the National Science Foundation (NSF) from October 1, 1996 to December 31, 1999. Dr. Bertenthal is the author of more than 100 publications on perceptual, motor and cognitive development, developmental cognitive neuroscience, visual processing of motion information, perception and production of biological motions, nonlinear modeling of posture and gait, and science policy. He is a fellow of the American Association for the Advancement of Science, the American Psychological Society, and the American Psychological Association. Dr. Bertenthal was the recipient of a Career Development Award (1985-90) from the National Institutes of Health, and received the American Psychological Association’s Boyd R. McCandless Young Scientist Award for distinguished research in 1985.