REPORT OF THE WORKSHOP ON SHARED CYBERINFRASTRUCTURE FOR THE SOCIAL AND BEHAVIORAL SCIENCES

JULY 31, 2009

BACKGROUND

This is a report of a workshop held at the National Science Foundation April 8-10, 2009, attended by 17 leading researchers from various SBE fields. The goal of this workshop was to consider ways in which major new NSF investments in shared cyberinfrastructure could transform the human sciences. The current report, written by the co-conveners of the workshop, Robert Groves and Brian MacWhinney, has these seven sections:

1. Overview and Summary: This section articulates the vision of a Synthetic Infrastructure and sketches out two complementary routes toward achieving that structure.
2. Recommended Follow-up Workshops.
3. Recommended Short-term Steps.
4. Recommended Long-term Steps.
5. Detailed Analysis: This section provides a more detailed analysis of the history of investments in cyberinfrastructure within SBE, key opportunities provided by the Synthetic Infrastructure, and structures that are needed to support SBE cyberinfrastructure.
6. Individual statements from participants in the workshop.
7. A list of participants.

1. OVERVIEW AND SUMMARY

Shared cyberinfrastructure holds an enormous potential for the advancement of the social sciences. The explosive growth of all of the technologies supporting cyberinfrastructure has opened up a new view of the social sciences that was unimaginable two decades ago. We can see now how this new technology could allow us to build massive, deeply integrated databases that could transform the SBE sciences. Huge streams of relevant cyberdata are being generated throughout society by government, business, churches, hospitals, clinics, broadcasters, families, friends, schools, and universities. NSF could provide crucial support for systems to synthesize these data into coherent pictures of human activities. In this report, we outline a way to achieve this integration. We propose a series of specific actions designed to lay the groundwork for the construction of a new synthetic cyberinfrastructure that will be shared across all of the social and behavioral sciences.

The current cyberinfrastructure for the SBE sciences focuses on the construction, dissemination, and analysis of data within each of the separate SBE sciences. This infrastructure has played an important role in the development of areas such as political science, linguistics, or economics. However, these current systems have not yet addressed the challenge of providing a cyberinfrastructure that links across multiple disciplines within SBE. Conceptually, it is easy to see how databases can link information relevant to all the SBE sciences. For example, a single database could link psychological, genetic, linguistic, political, social, geographical, cultural, and neurological data at the level of the individual human being. Data on the individual level could be tracked longitudinally to reflect individual life changes and could be aggregated across individuals to reflect group patterns. We will refer to this proposed new linked database for the SBE sciences as the Synthetic Infrastructure.

Construction of this Synthetic Infrastructure would allow researchers to detect and validate patterns that had previously only been studied across a single dimension. For example, we could track the academic performance, housing patterns, and income dynamics of rural families that move to the city and compare those with families that remain in the country. By integrating linguistic, social, genetic, economic, experimental, demographic, and neurological data on children with communicative disorders (stuttering, SLI, articulation), we could discern subgroups in this population and the ways in which those groups achieve academic and social success. We can link standard survey data to cyberdata from social networking sites, business transactions, smart phones, and online experiments. With such linked data, we can advance fields as disparate as labor statistics, political science, and computational linguistics.

The proposed Synthetic Infrastructure will not have all data cells filled in for all participants. Some forms of data, particularly those available in the commercial world are available for nearly everyone. Other forms of information, such as genomic or neuroimaging data, are only available for small numbers of participants. Because of this, we need to develop statistical methods for creating small samples that represent "fractal" components of larger samples. It will also be important to create a type of Synthetic Infrastructure that is integrated not on the level of the individual participant, but on the level of the social group (workplace, club, movement, tribe, religion, or culture). The actual construction of the Synthetic Infrastructure rests upon continued work in the development of discipline-specific methods for data creation and extraction. However, we believe that development of the Synthetic Infrastructure will play a central role in transforming and integrating across all the SBE Sciences. In short, the workshop concluded that significant breakthroughs could be possible if the SBE sciences jumped up to the larger scale of integration for cyberdata embodied in the Synthetic Infrastructure.

Without the development of a Synthetic Infrastructure, the SBE sciences will remain unidimensional and data-starved. They will remain unable to relate broad, but shallow macro studies of groups, employers, countries, and events with narrow, but deep studies of conversational patterns, genomics, and neurology. Without links across streams of cyberdata, the SBE sciences will have no access to the dynamic data sources that mirror the fast pace of modern life. Without the Synthetic Infrastructure, the SBE sciences will remain blind to the insights possible by linking data from the genome to the nation-state.

The workshop sketched out two major, mutually interacting, routes toward the construction of the Synthetic Infrastructure. One route focuses on the linkage of new data, the other on the linkage of existing data. The first, most ambitious, route emphasizes the collection of new data to populate a newly designed synthetic database. The bulk of these data can be harvested on the national level from web servers, commercial databases, labor statistics, and medical databases. Other segments of the new database could be added from ongoing NSF projects, such as the ANES, PSID, or the GSS. However, to achieve full, detailed representivity in the synthetic database, we would propose the establishment of perhaps 25 physical SBE regional research centers located throughout the country, each responsible for sites near their location. The exact number of such sites is not as important as the capability of the overall network to create a representative sample of the US population, based on an area probability sample. A sample of persons would be identified and assigned to each center. If funds are only available for a smaller number of sites, their geographical reach could be extended through the use of mobile data collection and observation sites. Standardized self-report interview protocols (using web, mobile devices, and traditional modes) would be administered to this national sample (indeed, such a design could become an evolved complement to the existing PSID, GSS, ANES). Some of these would involve near-continuous electronic sensor measurement (measuring social interaction, exposure to alternative environments, spatial mobility, physiological variations). The research center would assemble all public records on the area (voting, property records, library, drivers' license, commercially available person records). With their explicit permission, nonpublic data on participants would be assembled from medical, credit card, social networking site interactions, and other sources. In return, the data center would provide bonded guarantees against identity theft, checks on discrepancies in their data across records, and assistance in correcting their personal records.

Fractal designs, which would at random assign additional measurement protocols to subsets of the participants, would be the tool to link in-depth studies of individuals to the large, representative national sample. These in-depth studies would measure a variety of neurological, genomic, linguistic, social, consumer, and linguistic variables. They could also include videotaping of family interactions or studies of the use of the Internet. Participants would derive concrete value from these studies by having access to useful personal psychological, social, and genetic profiles. Using this information, participants could be given further information (travel, education, medical, social) that was found to be useful to others matching their particular profile. The measurement of the individuals would extend over time, with participants rotating out of the panel and new random samples rotating into the panel over time.

The 25 research centers would thus be data collection facilities, but they would also be synthesis research centers, providing a local gathering place for interdisciplinary teams to do their science. These centers would maintain the highly secure integrated data sets that would be the products of the new SBE world. Short-term visits of SBE scientists (and others) would facilitate access to special data sets that could not be distributed in any other virtual way without violating confidentiality pledges. Research teams, funded under NSF analysis grants would spend time working together in various sites.

A second, complementary, route toward the construction of the Synthetic Infrastructure focuses on the fuller utilization of currently available materials. This route seeks to identify ways in which current databases can create pieces of the new Synthetic Infrastructure. Many of these databases involve current and past initiatives at NSF and NIH in areas such as genetics, linguistics, anthropology, psychology, archaeology, and political science. There are four priority activities here.

1. We need to rescue targeted pieces of legacy data, some of it in digital format and some in pre-digital format. Recognizing that not all legacy data can be ported over into the new Synthetic Infrastructure, we need to identify those segments of data that can be successfully ported. For example, there are compelling arguments for digitizing the 60 years of tape recordings of oral arguments that have been preserved at the Supreme Court.
2. We need to establish smoother processes for merging existing cyberdata. The NSF INTEROP Program addresses needs of this type.
3. We need stronger support for data sharing. Currently, the majority of NSF grantees fail to make the results of their funded projects available to the scientific community. There are several solutions to this problem that will be discussed later.
4. We need fuller support for the extraction of data from current data streams. There are many existing public databases and data streams from which we can extract important demographic, linguistic, and social data. However, often we need to build extraction tools and web sites to make best use of these data streams.

By combining these methods, we can move forward quickly along this second route of constructing of a Synthetic Infrastructure based on currently existing data.

To build the full Synthetic Infrastructure, will want to move forward in parallel along both of these routes. The first route provides us with unparalleled opportunities for controlled research designs and scientific sampling. The second route allows us to maintain the historical dimension and to make immediate progress in terms of concrete data linkage. We believe that the future of the SBE sciences lies in pursuing both routes in a coordinated and integrated fashion.

2. Recommended Follow-up Workshops

The Advisory Panel recommend that SBE continue the process of targeting and specifying new approaches to shared cyberinfrastructure by sponsoring these three additional meetings:

1. **Infrastructure Design**. We should conduct a workshop to outline the shape of the proposed Synthetic Infrastructure:
    a. What should be the general design of the Synthetic Infrastructure? How can this design be structured in a way that allows for continual growth?
    b. How can we combine data types?
    c. How can we integrate currently available data? Can we construct demonstration projects based on the second route?
    d. What demonstration projects can provide proof of concept of the Synthetic Infrastructure, based on the first route?
    e. How should demonstration projects be sequenced? What evidence is required to evaluate whether there is sufficient support in the SBE community for such an investment?
2. **Tools**: We should conduct a workshop on the tools needed to build the Synthetic Infrastructure, such as:
    a. web-based records on persons and interactions among persons, as collected through e-harvesting, web scraping, commercial transactions, and social networking methods;
    b. use of electronic sensor devices (GPS, biophysical monitoring, particulate matter levels, polycyclic aromatic hydrocarbons, temperature, humidity) and smart grid technologies;
    c. new developments in survey-based self-reports and continuous monitoring survey designs;
    d. measurement technologies that SBE scientists now treat as limited to small numbers of volunteers because of their intensivity (e.g., brain scans, extended linguistic interactions, video observations). The workshop would tackle the problem of the integrating these data into the Synthetic Infrastructure using fractal designs; and
    e. methods for incorporating data from social experimentation into the Synthetic Infrastructure.

3. **Organization**: We should conduct a workshop on the organizational infrastructure of the Synthetic Infrastructure.
    a. Should there be “synthesis” centers that stimulate the construction and utilization of the Synthetic Infrastructure? How should these be organized?
    b. How would measurement initiatives be evaluated and governed?
    c. Which measurements should be one-time and which should be ongoing?
    d. How would data be aggregated (from the cellular to the organization)?
    e. How would virtual group (e.g., internet networks) be measured versus spatially defined groups (e.g., church congregations, bowling leagues)?

In addition to these three top priority workshops, we also recognize the need to organize workshops that examine these three further areas:

1. **Ethical, Legal, and Proprietary Issues**. This session would examine barriers or limitations to the construction of the ideal database. These issues include privacy, intellectual property rights, ethics, IRB and HIPPA restrictions, and access to private sector data. It would examine how data sharing can be maximized against the backdrop of these restrictions.
2. **Data Rescuing and Access Maximization**. This session would examine the needs for database creation in areas of SBE that have received relatively less attention to shared infrastructure. It would consider how to archive, digitize, and integrate legacy data, particularly from anthropological, archaeological, and linguistic studies. It would examine the needs for ontologies to promote database integration and ways that federal agencies can stimulate greater sharing of important scientific materials.
3. **Maximizing Broader Impacts**. This session would examine how the Synthetic Infrastrucure could be structured in ways that would maximize utility for policy makers, the public, education, museums, zoos, and business.

3. Recommended Short-term Steps

The workshop provided the following list of recommendations for actions that can be taken immediately or in the short term.

1. **Coordinating CI efforts within SBE**. There should be a single NSF contact person (or perhaps a fully configured program) responsible for funding workshops in cyberinfrastructure, integrating calls for proposals, providing ongoing documentation of infrastructure needs and efforts, improving data-sharing, and generating new support for shared infrastructure.

2. **Increasing funding.** Funding for shared cyberinfrastructure should be markedly increased.
3. **Focusing the competition**. Grant proposal evaluation usually focuses on hypothesis testing. However, infrastructure projects must be evaluated using different criteria and possibly different mechanisms. Competition for infrastructure funds should be grounded on criteria such as linkage across databases, quality of curation, publication records, and broader impact.
4. **Improving data sharing**. Although data sharing is standard in some parts of science, including segments of SBE, the majority of SBE research, outside of the major surveys, is not being shared. To encourage sharing, we recommend that continued funding, even on the level of annual reports, be based on evidence that results from previous research have already been added to the relevant databases. We also recommend that NSF assist researchers in configuring plans for data sharing that mesh well with IRB policies and expected data-sharing practices in each research community.
5. **Integrating with other infrastructure programs**. We recommend that SBE explore opportunities to increase the integration of funding, planning, and data-sharing support with other related segments of NSF (CISE, EHR), NIH, Census, BEA, BLS, NCHS, DOE, and other agencies. NSF should work together with these agencies to encourage a common plan for promoting social and behavioral data infrastructure.

4. Recommended Long-term Steps

The workshop provided the following list of recommendations for actions that could be taken in the long term.

1. **Synthetic Infrastructure**. Participants emphasized the value of the core target of developing a Synthetic Infrastructure that could serve the needs of all SBE disciplines. This superdatabase would link demographic, voting, linguistic, commercial, neurological, and life history data for individuals. The second form of superdatabase would be coded at the group level, linking events data, public opinion data, legislative voting data, elite data, public opinion data, and media data across time, place, and ethnic group. The development and utilization of these two superdatabases could be linked to the development of synthesis centers in recommendation 2 below.
2. **Establishing synthesis centers**. NSF should explore the establishment of synthesis centers designed to construct and utilize superdatabases. To give a particular example, there could be a synthesis center for Social, Political and Economic analysis that would examine major issues such as the collapse of the Soviet Union. This center would

construct a superdatabase linking events data, public opinion data, legislative voting data, elite data, public opinion data, and media data across time, place, and ethnic group. Similarly, NSF could support a center for Human Communication that would create a superdatabase of conversations, broadcasts, interviews, and texts for interdisciplinary analysis. These centers would be responsible for configuring the ideal database structures that would form the backbone of a Synthetic Infrastructure for SBE.

3. **Small grants.** Small grants should be offered to PIs and graduate students for pilot projects, archiving of legacy data, create viral tools, or do data scraping. These opportunities should be targeted and announced through Dear Colleague letters.
4. **Establishing data collection centers**. These centers would be modeled roughly after the NEON centers of the environmental sciences, would re responsible for the collection of survey, behavioral, genomic, neuropsychological, and other data using new methods of data collection (viral tools, sensors, blogs, smart phones). These centers could also play a role for collecting DNA profiles or providing measures in archaeometry.
5. **Ontology, harmonization, and linkage**. We recommend that NSF provide support for methods of converting disparate legacy data to new integrated, fully digital databases. This work involves efforts in data gathering, curation, translation, and interoperability. It must be accompanied by full support for data sharing. Much of the work here involves the development and promulgation of consistent data coding standards. However, work in metadata and ontology should not be supported except as it provides direct input to shared datasets. Linkage will allow us to join disparate data types such as survey data and neurocognitive data.
6. **Improving training**. Participants expressed concern that graduate training is not providing the data collection and analysis tools need for 21st century infrastructure. The occupation of building SBE cyberinfrastructure should be configured as a full academic path in its own right.
7. **Building tools**. Our current tools are provided by freeware developers and extended by social science practitioners as they encounter new data and analytic problems. Tools like R are not yet robust for large datasets. Visualization tools for neuroimaging data are often incomplete. We need to support and institutionalize these efforts at development of these tools and then reach out to the user community for training.
8. **Improving data entry**. For databases in areas such as linguistics, cognitive neuroscience physical anthropology, and archeology, there is a pressing need for automatization and standardization of methods for data entry and curation.

5. Detailed Analysis

The social and behavioral sciences have completed a century of extraordinary growth that has transformed our understanding of human thought and behavior and created methods for improving the daily lives of real people. Research in the human sciences has examined a broad spectrum of phenomena, including evolution, cultural change, learning, memory, cognition, development, language processing, social organization, political processes, economic structures, environmental impacts, and globalization. The focus of this work can be on individuals, small groups, families, companies, classrooms, political parties, social classes, or even whole cultures.

Methods in the human sciences are as diverse as the subject matter. As in the natural sciences, innovations in measurement approaches deserve much of the credit for knowledge growth. Some of these measurement advances involve our ability to conduct increasingly powerful observational studies – ranging from long-term qualitative and quantitative collection of measures within small groups of persons to full censuses of large populations. There have also been advances in the formulation of quasi-experimental designs that allow us to evaluate large-scale interventions. We have seen the growth of the sample survey as a cost-efficient method to study large human populations, and the introduction of statistical and mathematical models for extracting causal inference from correlational data. Together, these methods support large-scale program evaluations that are ubiquitous in public policy domains. On the small-group level, computer-controlled methods for multimedia analysis now allow us to study the details of gestural and verbal communication in naturalistic settings.

There have also been great advances in experimental methodology. Computer-controlled response time experiments have allowed us to formulate models of the online processing of incoming information and the precise sequence of the generation of plans for action. Using functional magnetic resonance imaging (fMRI) and subtracting between comparable tasks, we can pinpoint brain areas associated with specific cognitive and motor performance. Other imaging methods, such as EEG, ERP, MMN, and MEG can allow us to study the way in which information flows between areas of the brain in real time. Within the contexts of these experiments, response patterns can be studies in great depth with eye-tracking, phonological analysis, and a wide array of psychophysiological and psychopharmacological methods.

Other methodological advances have expanded our understanding of the scope of human time and space. Advances in geo-spatial information systems and statistics have reshaped how transportation and land use analysis is conducted. Breakthroughs in genetic mapping and analysis have transformed our vision of the last 6 million years of human evolution. New methods for 3-D modeling of archaeological artifacts have allowed us to construct new analyses of object construction and provenance.

Linking all of these new methods is our ability to use the Internet to merge separate databases into an increasingly connected whole. Databases grounded in survey and census data have played a central role in key areas of the social sciences. Now, these databases are being extended to new areas and opened up to increasingly wider communities of scientists. The success of publicly shared databases for genomics and astrophysics is being reflected in shared databases across virtually every area of the human sciences.

A mark of the maturity of such methods is their application to research and practical problems outside the human sciences. These applications include topics as diverse as the influence of lifestyle on health, health care access and quality, measurement of mental health, brain function and activity, classroom education, distance learning, transportation, human impacts on the environment, homeland security, jurisprudence, and national defense.

Despite these advances, we are beginning to confront certain key impediments to further progress. These impediments have forced the current generation of scientists to focus on too coarse a level of detail, blinding them to the mechanisms affecting human thought and behavior at fine-grained temporal, spatial, and group levels. To understand the emergence of structures at the macrolevel, we need more powerful "microscopes" for recording of data patterns on these microlevels. These microscopes will study spatial positions, brain activations, details of cash flows, populations movements, traffic in the blogosphere, discourse flow in classrooms, and myriads of other fine-grained patterns of human behavior. To record, harvest, organize, and publish these data, we must reshape the research infrastructure for the human sciences. To this end, the aims of this proposal are:

1. to assemble a group of knowledgeable researchers from several fields relevant to identify key behavioral and social science questions that cannot be answered because of weaknesses in current methodological and data infrastructures;
2. to document those questions through a series of workshops;
3. to identify potential tools and data structures to address these questions, and
4. to propose a set of steps for building the needed shared research infrastructure in the near term.

a. Previous SBE Initiatives – (contributed by Cheryl Eavey)

In the early 1960s, infrastructure for the social and behavioral sciences was supported by two programs: a Facilities Program and a Special Projects Program. The former supported instrumentation activities. The latter supported major data-collection activities, including the big three social science surveys: the General Social Survey (GSS), the Panel Study of Income

Dynamics (PSID), and the American National Election Study (ANES). This program, however, also provided significant support for other infrastructure activities, such as support for various acquisition and archiving projects. At some point, the Special Projects Program morphed into the Measurement, Methods, and Data Resources (MMDR) Program. The MMDR Program's primary focus continued to be infrastructure support, but it also provided support for analytical methods and survey research. After 1984, the MMDR program was restructured into the Measurement, Methods, and Data Improvement (MMDI) Program. At this time, the data-collection function of the MMDI Program was greatly reduced; most notably, the three large social science databases and the funds supporting them were relocated to the relevant disciplinary programs. Still, approximately 60% to 75% of the MMDI budget was spent on organizational infrastructure and/or data improvement. In the early 1990s, the MMDI program was restructured again into the Methodology, Measurement, and Statistics Program with a primary focus on the development of innovative methods and models for the social and behavioral sciences. During the last ten years, SBE has held four focused competitions for large-scale infrastructure activities across the social and behavioral sciences. The first two competitions were conducted in 1999 and 2000/2001 with the goal of enhancing infrastructure across the social and behavioral sciences. A total of 13 awards were made from these two competitions. The third competition for "instrumentation and data resource development" was part of the 2004 Human and Social Dynamics (HSD) priority area. That competition resulted in one large infrastructure award. The final competition in 2005 was conducted collaboratively with the Directorate for Computer and Information Science and Engineering (CISE) and supported the development of infrastructure test beds that included the next generation of cyberinfrastructure for data focused on either organizations or individuals. The 2005 competition resulted in two awards. In addition to the above activities, SBE has conducted more focused competitions for centers in environmental areas (such as decision making under uncertainty with respect to climate change) and in the science of learning. Each of the regular programs in SBE has supported, as part of their overall portfolio, infrastructure activities for their scientific communities.

b. Challenges and Opportunities

NSF has made significant contributions to the development of shared infrastructure for SBE. However, the pace of investment in infrastructure began to slow after 2001 and is now not keeping pace with the growing needs in this area. This gap has occurred for two reasons. On the one hand, the spread of digital technology has opened up enormous new opportunities. On the other hand, there has not yet been a major increase in funding in response to these new opportunities. This gap between resources and opportunities can be illustrated by taking two fields as examples: survey methodology and the study of classroom instruction.

In the area of survey methodology, there are many troubling indicators that current methods are reaching their limits. We are seeing declining response rates to surveys of all sorts, driven by a combination of increasing reactance to measurement and skepticism of institutions among the general population. These trends are further exacerbated by fears of new immigrants and other minority populations regarding release of personal information. Rising identity theft and the increasing commercialization of data gathering have combined with general worries about privacy, confidentiality, and misuse of data to diminish willingness to participate in research among members of the general population. This decline in access to research populations is occurring in parallel with an unprecedented growth in commercial data gathering. These new commercial databases and data firms (Acxiom, Experian, Lexus/Nexus) are tracking social and personal patterns involving traffic and surveillance camera data, data on personnel actions, transaction data (credit cards, entrance/exit, shopping clubs), health events and service provision, spatial data, networked sensor data, satellite imagery, genealogical data, digitized land use records, financial documents, and data on Internet usage, references, and behavioral traces in "cyber-space".

These data have some of the fine-grain temporal, spatial, and group characteristics missing in current NSF-funded social scientific surveys. If social scientists can gain access to these and other sources, they will be able to construct new metrics for measuring phenomena of longstanding interest and promise to reveal new dimensions of human social organization and behavior. However, viable methods for providing access to these new data sources are not yet in place.

The study of classroom interactions has run into a different set of impediments. This area relies heavily on the video recording of live classroom interactions. Current research is often based on a single camera filming either the instructor or a particular student. This limited observational focus excludes an accurate understanding of the complete classroom environment that will include postural, gestural, and vocal communications that go far beyond the direct dialog between the instructor and a single student. Occasionally, a second camera is introduced to broaden coverage, but it is often difficult to relate these two views while conducting analyses. Solutions to this problem involving a 360º panoramic view have been constructed, but the technology for this approach is still underdeveloped. Apart from these problems with recording methodology, this field suffers from the absence of strong methods for data sharing over the web. Although many researchers are willing to share their data, the field needs support for automatic anonymization of the transcripts and audio, tools for automatic linkage of text to speech, cloud computing methods for compression, and methods for browser-based collaborative commentary regarding interactional and instructional patterns.

Once this technical infrastructure is in place, it will be possible to apply computational linguistic (parsing, topic extraction) and video analysis methods (segmentation, image linking) to structure the database for rapid retrieval of specific instructional types and interactional sequences. However, until these new technologies and methods are in place, this area is largely frozen into a period of minimal progress based on inadequate shared infrastructure.

This analysis of barriers to progress could be repeated for each of the dozens of areas that comprise the human sciences. In each case, inadequate infrastructure stands as the major obstacle to qualitative theoretical advance. Fortunately, the human sciences are often able to benefit from advances in other areas. For example, improvements in the Internet, disk storage, medical imaging, and video technology have boosted many areas. However, these advances do not come fully preconfigured for the human sciences. To make full use of the new technology, human scientists need additional shared infrastructure that adapts the new technology for their questions.

c. Key Questions

When considering a major investment shared infrastructure, it is important to ask what questions we could address with this new infrastructure. The participants in the planning workshop proposed a set of major questions that are central to current research and which can only be addressed through new infrastructure. Here are some examples.

1. What patterns and mechanisms cause languages to change, and how are these patterns correlated with linguistic and non-linguistic factors? How do these things relate to communicative channels – blogging, chatting, facebooking – and how do these non traditional communicative channels cause language to change.
2. How can we use neuroeconomics and genetic markers to describe and predict investment behaviors?
3. How can we measure and describe social support networks and latent support networks across widely different social groups?
4. Identify, measure, and explain flows of reciprocal causation between humans and aspects of their natural and built environments at various spatial scales;
5. Characterize and examine the influence of energy — not only the amount, but its sources and modes of supply and use — on the structure of human settlements and the patterns of social behavior within;
6. Study mind-brain interaction and integration (“the human cognome”), and its implications for human behavior and social organization;
7. Use complete 24/7 video records of a child’s language learning environment to model and predict the details of the course of language learning;

8. Compose multilingual, multicultural databases linked to texts and video interactions from hundreds of action frameworks in disparate cultures that will allow us to understand the full scope of possible cultural meanings;
9. Extend the temporal dimensions of the human sciences beyond the spectrum of conscious human thought and action, to a range that spans judgments that occur in a "blink" (or less) to trends that encompass lifetimes;
10. Measure the constituents of human social behavior—valuations, judgments, comparisons, motivations—in ways that allow them to aggregate into the emergent phenomena of decisions, behaviors, patterns, and structures (that is, to deliver on the promise of complexity theories and provide entry into emergence and tipping-point dynamics);
11. Extend the human sciences into the "virtual" world, which is rapidly becoming a real and powerful venue of social behavior, interaction, and organization;
12. Recreate the population movements, cultural developments, and physical artifacts of the last 100,000 years and their effects on current populations, languages, and cultures;
13. Link the evolutionary patterns of the last 6 million years to specific structures in the human brain and physiology in ways that allow us to understand the current diversity of human abilities and disabilities;
14. Study the creation, shaping, uses, and consequences of knowledge in its various forms, which include indigenous or folk knowledge, craft, technology, and science, through expanding our capacity to store and analyze symbols, images, and artifacts.

d. Future Directions in the Social and Behavioral Sciences

These accomplishments, trends, opportunities, and needs bring the human sciences to the cusp of a transition. They promise more extensive, integrative, and precise explanations fashioned from data that are higher in spatial, social, physiological, and temporal resolution. Accomplishing this transition will require investments of ideas, energy, and resources to develop more encompassing conceptualizations of human behavior that pose research questions of greater depth, scope, and specificity. In practice this means that questions currently stated in the limited vocabularies of social science disciplines and addressed with traditional epistemic standards and methods must be reframed in language that crosses disciplinary borders (for an example, see the 2007 BBS article by Gintis). Accomplishing this transition will also require dramatic changes in the volume, character, and technologies of data gathering and integration (interoperability) in the human sciences, accompanied by mathematical models and statistical techniques suitable for their analysis. Among the clear needs ahead are:

1. Increased spatial resolution: Human society is spatially heterogeneous. Characteristics of places, at scales nested from neighborhood and community up to biome and megapolitan area ("sociome"?), shape social organization and behavior. Current sample sizes of our largest surveys are much too small to provide sufficient statistical power to examine the reciprocal influences of context on social organization and human behavior. The rising challenges of socio-environmental research and the opportunity to couple data about humans with data gathered by environmental observatories (NEON, WATERS) presents an unparalleled opportunity for integrative explanations in a crucial area of inquiry.
2. Better data quality: For a wide variety of analyses, we need richer data. For archaeology, this may mean high-quality 3-D scans, x-ray and catscans, and trace elements and isotopic composition, volcanic source analysis, DNA analysis, ad tree-ring dating. Site data include collection of surface materials and site descriptions plotted in three dimensional space. Individual objects such as potsherds, stone tools, animal and plant remains are also studied in some cases utilizing attribute systems that record dozens of variables on a single piece. Many of these methods require centralized laboratory support, although the results can be disseminated through the web. In homes, offices, and medical facilities, it means video recording across longer periods of time from more angles with higher audio and video quality. We need unobtrusive video recording, even in difficult field situations, along with links to wearable eye-movement and head movement monitors. All of this must be linked to data from neuroimaging and psychophysiological measures, including implanted electrodes for disabilities. For studying online interactions, we need better methods of capturing and exchanging audio and video over the web. For experimental control systems, we need smoother linkage between display systems, neuroimaging, psychophysiological measures, and computer control.
3. New data types: Cyberinfrastructure has extended the social world into a new dimension (Second Life, World of Warcraft, see Bainbridge, 2007) and the human sciences must embrace this transformation, by studying new phenomena (e.g., the rising specialty area of e-social science) and employing new technologies as research instruments. For example, ubiquitous computing and related technologies (cell phones, PDAs, email, the web and such) generate new forms of "everyday data," stored in immense volumes of interaction records that are increasingly accessible for research. The Center for Embedded Network Sensing, an NSF-funded Science and Technology Center at UCLA, is embarking upon a bold urban sensing program that will extend their work with environmental sensor networks more deeply into the sphere of human life (see www.cens.ucla.edu).

4. Improved access and sharing. We need to overcome a wide variety of barriers to data access and data sharing. We must construct new systems that encourage researchers, particularly those receiving federal funding, to include their data in national and international databases. Where possible, new funding should be preferentially linked to data sets that can be shared and which are compatible with standardized data formats. By introducing the proper levels of data access control, no research data should remain unshared. A secure facility for access to confidential data (not unlike the NSF-supported Census Research Data Centers) would allow spatially referenced data on individuals to be used, alone or in combination with other sensitive data (from Census, biomarkers, public records, and electronic records). The availability of this information would contribute to the development of greatly expanded explanations of human behavior, cultural and social organization and change, decision-making, and interaction patterns. The scale and complexity of the data files, and the technical challenges of integration and interoperability, suggest that centralizing the work would offer advantages and efficiencies. A centralized database could be used to reduce the burden of respondents to self-report attributes that are already measured, if ethically responsible ways to combine survey data with administrative data could be constructed.
5. Fuller internationalization. The capacity for open access to shared data over the web can further promote the growing internationalization of our computational and data infrastructure. Databases can link and compare information across countries, languages, religions, and social systems. In this way, social and behavioral scientists can test the applicability of theories to increasingly diverse social, cultural, and behavioral pattern. Moreover, agreement on a common computational infrastructure can allow researchers in all parts of the world to share in the development of new tools for data analysis and visualization.
6. Automatic annotation. Fields as varied as cultural anthropology, linguistics, political science, psychology, archaeology, geography, sociology, and information science confront volumes of textual, archival, and image data that are cumbersome to store and access, and that require substantial amounts of analysts' time to examine, organize, and analyze. Computer scientists are developing image processing and analysis tools for other fields of science (e.g., automated bird recognition for ecology), and such methods can be adapted for use in our fields. As image data are also involved in surveillance, this mode of data collection and analysis will have significant national security implications.
7. Protocol standardization and enrichment: Many areas of the human sciences use standard protocols for data collections, thereby facilitating data exchange across laboratories. These protocols can include questionnaires, story telling, personal histories, picture description, word repetition, problem solving, or other methods. Areas need to standardize on data

collection protocols that can maximize data sharing, analysis, and theory development. These protocols then need to be applied systematically and strategically across participant groups, interactional occasions, and spatial configurations. Often it will be necessary to return to communities or households after periods of even a dozen years to track slow-moving changes, such as language loss. Within some areas, such as primatology, the development of field sites may be the best way to standardize on protocols and infrastructure. A particularly promising development involves the survey methodology developed in the TESS (Time-sharing Experiments for the Social Sciences) project, funded by NSF. This project combines the inferential power of experiments with the generalizability, diversity, and sample size of surveys. Such hybrid surveys become fast and flexible research platforms that also enhance public understanding of and engagement with science.

8. Increased social resolution: Human society is increasingly socially, ethnically, and culturally diverse, and the differences embedded in this diversity are consequential for scientific understanding and social policy. To take a simple example: the largest and best surveys supported by NSF are scarcely powerful enough to provide adequate samples of major ethnic groups, such as African Americans, Latinos, and Asians (and achieve adequacy through over-sampling). It has become increasingly clear that major ethnic groups are composed of distinctly different subgroups—among Latinos, Mexicans are unlike Cubans or Puerto Ricans; among Asians, Pacific Islanders are unlike people from India or Vietnam—and such differences are consequential. With rising concern for the dynamics of immigration and with the pressing worries of the human capital implications of changes in the social composition of the nation (e.g., *Rising above the Gathering Storm*), enhanced social resolution is essential for social science and social policy.
9. Extended temporal duration and resolution: Patterns and processes of change are at the heart of the human sciences, and the study of change demands data gathered over time. Modern commercial use of transaction data permits real-time intervention based on data. This has taught the private sector that important changes occur minute-by-minute in the human behavior affecting businesses. Academic social science is behind by large orders of magnitude. At present, for example, two of NSF's best (and most expensive) national surveys are fielded biennially, which limits by design their ability to detect fine-grained temporal changes. For some purposes it may be possible to elongate the timelines of research by imputing or reconstructing data from the past, but more likely we must wait as the timelines are extended. The social sciences will not understand changes that occur in the day-to-day lives of people by studying them year-to-year.
10. Better analytic and control methods: We need to improve the usability and accessibility of the latest methods for experimental control and data analysis. In most cases, this means building additional interfaces to accepted systems for statistical analysis or experimental

control. In other cases, new systems must be designed from the ground up, using interoperable modules, based on common data structures. Data must be in a form that allows them to be directly introduced into shared databases that are linked to powerful analytic systems.

11. Tools for modeling and visualization. Models and simulations are rapidly rising research technologies that sometimes complement and sometimes challenge traditional scientific explanations. More powerful computers, better visualization technologies, and an influx of researchers are driving this potentially transformative change in behavioral and social science. The ultimate success of such models depends upon access to data of sufficient resolution and precision to test model predictions, and to derive from those data more accurate parameters to improve subsequent models. Visualization technologies also create an environment in which research participants are presented with alternative scenarios as part of a controlled experiment. Finally, these technologies would enhance education and public understanding of science by presenting research results to students and members of the local community in ways that make them accessible and salient for their interests and concerns.
12. Terascale computing. Databases in areas such as geography, video analysis, and neuroimaging are already so massive that they are placing strains on resources for storage and access. Conventional GIS algorithms and spatial analytical tools are based on single-CPU architecture and single-user interfaces and, therefore, cannot effectively exploit the utility of cyberinfrastructure that is built on parallel and distributed architectures. New, robust spatial-analytic techniques that take advantage of parallel and distributed computing architectures need to be developed. The expansion of data types for national surveys, and the inclusion of data from commercial methods will further expand the raw size of our databases. We will need to rely on the teragrid, cloud computing, and other methods to store and access all of these increasingly rich data sources.
13. Improved training: The introduction of better protocols, new methods for data sampling, and better analytic and control methods sets the stage for major advances in the training of research scientists. As our methods become standardized, it makes sense to link research training to specific modules delivered either at national institutes or over the web. New forms of networked research organizations (virtual organizations or collaboratories) can provide local access to the data and tools, training and retraining (for those in mid-career), and collaborative opportunities essential for improving the quality and velocity of research in the behavioral and social sciences. Such organizations would be places for creating and teaching mathematical, statistical, and computational skills necessary to analyze data of increased variety, complexity, and volume. Such organizations would also serve as liaisons to local sources of public records data, points of access for regional scientists and decision

makers, interface organizations with NSF's environmental observatories, and specialized nodes in a network of cyber-mediated and face-to-face collaborations across disciplines and among scientists, engineers, and decision makers. This training must be configured to include opportunities for researchers from all underrepresented groups.

14. <u>Integrated data, analyses, and theories:</u> Among the grand intellectual challenges of the human sciences is the construction of explanations that connect neural and cognitive processes through behaviors to larger social patterns, and that reciprocally embed behavior and cognition within larger contexts of place and social structure, biography and history. A host of social science specialties, prefixed with "neuro-," are engaged in this quest, and initial results are promising, if crude. There is growing recognition of the value and possibilities of such explanations, and current large-scale social surveys are adding or considering various sorts of behavioral, biological, contextual, and environmental data. Such efforts are limited by the underlying designs of the studies. The challenge is to carry the integration further, first spanning the boundary between qualitative and quantitative data within the human sciences, then integrating biological and geoscientific data, from the scale of genes and neurons to that of ecosystems and climate, into explanations in the human sciences. Doing so requires achieving data equivalency (interoperability) across modes of observation, which has become a challenge for every field of science.

6. Individual Statements

Participants were also asked to contribute short statements of their own views of needs and prospects for shared SBE infrastructure. The next pages present these contributions.

HELEN ARISTAR-DRY

The central question for the workshop was posed in one of the introductory talks as, "How should we be investing in large-scale infrastructure that supports all the social sciences?" In my opinion, the best single answer is to invest in initiatives that promote data interoperability both within and across disciplines. Data interoperability is a powerful impetus for scientific discovery; the ability to combine new types of information in new ways has the potential to spark major discoveries in all the human and social sciences. To promote 'transformative' science and 'cyber-enabled discovery and innovation,' the single most promising path is to bring diverse data together in an environment that makes the data meaningful to, and usable by, scientists in different fields.

Transformative science based on cross-disciplinary data sharing requires both computational interoperability and human interoperation. The computational barriers to data interoperability are well-known, e.g., lack of standard data formats and markup, lack of data-processing tools which support the standards, lack of central data repositories, and lack of uniform access mechanisms. To these I would add human barriers, such as an academic culture that rewards the hoarding of private data and a lack of institutional clarity about intellectual property rights. To address some of these issues, NSF should promote:

1. Rewards for data sharing.
2. Standards for data formatting that promote data sharing and exchange.
3. Tools to assist in transcription and analysis.
4. Persistent identifiers for digital data.
5. Preservation of legacy data.

BENNETT BERTENTHAL

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, we suggest the need for a new infrastructure for collecting, storing, and analyzing data that will enable researchers to collect real-time multimodal behavior at multiple time scales. One example of this infrastructure is the **Social Informatics Data (SID) Grid** in which multimedia data is 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. The Social Informatics Data Grid exploits many of the recent developments in cyberinfrastructure funded by NSF, and is designed to transform how social and behavioral scientists collect and annotate data, collaborate and share data, and analyze and mine large data repositories. It is the type of infrastructure that is necessary for the social and behavioral sciences to conduct 'big science' and answer important and urgent questions about the human condition, such as how we promote human capital or end war or poverty or illiteracy.

At the heart of the SID Grid design is a rich data model that captures notions of time, data streams, and semi-structured data attached to these streams to enable powerful manipulations of multimodal data spread across data resources. Through query and analysis services deployed against the data warehoused in the SID Grid users can perform new classes of experiments. Shared data resources available from anywhere over the Web introduces new capabilities to the process of collection and analysis of data – collaborative annotation among them – without relinquishing control over sensitive data via an embedded security model. The key to the success of these large infrastructure projects is to engage the user communities from the get-go to ensure that the tools developed are accessible to a majority of users and are considered valuable enough to encourage data sharing and standardization of measures.

HENRY BRADY

Our biggest challenge is remembering that we are trying to develop a new infrastructure program for the National Science Foundation and not just a series of projects. Some excellent exemplary project ideas are certainly needed to help sell the overall idea, and at least one workshop should be devoted to generating the details of these projects. But at least one workshop must also be devoted to organizational models. The organizational models workshop must deal with these questions:

1. Where will the program be in the NSF? One suggestion is that it should be an Office of Infrastructure in SBE. Another suggestion is that this office should work very hard to make linkages with NIH, Census, BEA, BLS, NCHS, Department of Education, etc to do joint infrastructure programs with them.
2. What kind of programs will be in the Office of Infrastructure? One notion is that there should be at least three types of programs: Solicitations for ongoing infrastructure centers, solicitations for specific infrastructure projects (which might or might not be attached to the infrastructure centers), and solicitations for pre-doctoral and post-doctoral programs. It seems to me that one obvious point of contact with an SBE infrastructure program is the ongoing IGERT program which has funded some very interesting SBE doctoral programs. (For example, I am co-PI for the "Politics, Economics, Psychology, and Public Policy" IGERT program at Berkeley which has integrated training across four or five different social science disciplines. It would make a lot of sense to think how IGERT fellows could make use of infrastructure centers as part of their inter-disciplinary education.)
3. Exactly what models might work for ongoing infrastructure centers? Some serious thought should be given to the kinds of centers that might be created (the Organizational break-out group proposed some examples) and to the likelihood that they would really work. We must show that there is a need for them, that social scientists will use their services, and that good research will come out of them. Some very careful attention must be paid to the incentives that would be provided by infrastructure centers and to the likelihood that academics would be sufficiently motivated by these incentives. Moreover, it would be very useful if the organizational models workshop got people from successful centers in other sciences to describe what worked and what did not work. (One successful model from the social sciences that might be considered is TESS, Time-shared Experiments in the Social Sciences.)
4. What kinds of national needs should be addressed through these Centers? – A few overall themes came out of the worship that might provide a rationale for the centers.

These themes include the need to develop new modes of collecting data to keep track of a complex society, the need to reinvigorate our data collection infrastructure which has become a bit shopworn, the need to preserve and protect legacy and disappearing data and information, the need to develop automated methods to analyze text, audio, and video, and the need to bring SBE scientists together who have different forms of data that might solve a common problem.

5. Should support be temporary or long-term? How should Centers be reviewed? – I believe that centers should typically get long-term support but that they should be completely reviewed in a "Sunset Review" every five to ten years. But other models are possible – perhaps some centers should be expected to become self-sustaining?

Another Challenge – The other workshops should come up with a set of detailed projects that are inter-disciplinary, doable, and exciting in terms of the science that they will generate. These exemplary projects should help to provide a rationale for a substantial investment in social science infrastructure. My sense from the workshop was that a lot of good ideas were being put forth so that there should be no lack of interesting and worthwhile projects.

ANDREW GELMAN

We can use multistage sampling to efficiently collect data at different levels of aggregation, for example regions, states, municipalities, neighborhoods, blocks, and social networks within the United States. We refer to this as /fractal/ (in the mathematical sense of self-similarity at different spatial scales) because it involves sampling and data collection at several different levels, as compared to existing designs which typically aim for a single level, with inferences for finer slices or more general aggregations being an afterthought. I argue the following:

1. Fractal data collection is the /only/ cost-effective way to get information that is granular while also being representative of the general population.
2. Existing data collection methods are already fractal, but this fractality is implicit. We anticipate large gains in bias reduction and efficiency by explicitly including fractal sampling in survey design.

Consider the National Election Study and the General Social Survey, two publicly funded surveys conducted every two years with about 1500 randomly sampled Americans. The NES and GSS use face-to-face interviews and thus, by necessity, use multistage cluster sampling. In classical sampling theory, the main concern with cluster sampling is that high intraclass correlation will reduce the efficiency of the sampling, resulting in high "design effects" and inefficient estimates of population quantities. Both NES and GSS use sophisticated multistage designs that come very close to approximating simple random samples. This is great for most uses of these surveys, but it makes it nearly impossible for researchers using the data to learn about local behavior, for example the relation of individuals' political attitudes to those of people in their neighborhood.

At the other extreme of granularity are network or snowball samples, where a survey organization learns about a group of people in a social network, which can be defined tightly (for example, students in a school, employees of a firm, or scientific collaborators) or loosely (for example, drug injectors in New York City). Such surveys can provide invaluable information about social relationships and public health (relevant, for example, for studying the transmission of diseases or ideas) but are difficult to use for making generalizations about the general population.

A /fractal social survey/ could give us the best of both worlds. The idea would be to have a multistage national sample—along the lines of NES or GSS—but with greater density at each level of sampling. The survey would probably include data from all fifty states (as well as the District of Columbia and perhaps Puerto Rico and outlying areas). Unlike the Census Bureau's American Community Survey, though, the fractal survey would not sample from every county; rather, it would select some number of areas (cities, counties, towns, and unincorporated areas) within each state. Within each area, some number of neighborhoods (or the equivalent in rural areas) would be sampled, then blocks within each sampled neighborhood, and then there would be dense data collection within blocks, including some sampling of neighbors and of multiple individuals within some household. Added to this would be a snowball sample to capture the social networks of a sample of the respondents.

The fractal survey would not be cheap—to get good inferences at each level of granularity, the total sample size would have to be large—but the resulting combination of nationally representative inferences at different levels could be an incredibly rich resource for policymakers and researchers in social and behavioral sciences and public health.

It could also make sense to consider fractal data collection in the time domain, with some sample of respondents followed up at different time scales.

ED HACKETT

Infrastructure is literally the structure below the structure—the foundation, the groundwork—upon which a structure is built. The term conveys a sense of something solid, enduring, and inert. Foundations are completed to the specifications necessary for what will be built upon them and, in most cases, remain unchanged for the life of the structure. In fact, when a foundation shifts or settles—when infrastructure changes or needs replacement, it is usually a sign of age, failure, or poor design.

Infrastructure of the sort we discussed for the human sciences is not like that at all. Rather, it is a transformative force in itself, a technological system for conducting empirical inquiry that will enable challenges to received knowledge, afford novel empirical insights, and catalyze the formulation of new theories. It will have a certain amount of agency, reactivity, or potential energy; it will not be inert. The infrastructure proposed here will demand synthetic theories that span disciplines and levels of organization, from neurons to societies; new analytic models and tools, and people versed in their use; new patterns of research organization, collaboration, and publication (including collaboration with those we study and those who use our results); values and ethics attuned to the emergent challenges of data formed across personal repositories or gathered through electronic means. In other words, the infrastructure will place demands upon what is built upon it and, in turn, the process of building structures upon the foundation will reshape the infrastructure.

To me this paradoxical quality—the foundation shapes and is shaped by what is built upon it--implies a process of design and construction that is iterative and reflexive. Data, tools, and theories should be developed concurrently, as each will impose constraints and create opportunities for the others. Innovative arrangements for collaboration—centers, networks, partnerships—should be developed in piloted versions, test driven, and remodeled as development proceeds. To educate a new generation of scholars equipped with the ideas and abilities to use these new resources in new ways is a daunting task, particularly in these constrained times, yet that process should begin soon, perhaps using IGERT or other existing training programs, because it will take a decade to achieve. And this would be an opportune for a deeper, more effective reach into K-12 education: there is no reason students in middle school and high school cannot learn science, math, and statistics using real data from the behavioral and social sciences.. Finally, much of the infrastructure and new collaborative arrangements will be possible only with new ethical principles, custom tailored to the circumstances, firmly in place.

KATHLEEN MCGARRY

Much of the scientific effort in the social sciences is devoted to the analysis of data. We as a profession have benefited enormously from government funded surveys and smaller scale collection efforts. As technology improves and our ability to collect and analyze data expands in multiple dimensions, we are faced with the difficulty of ensuring that these data are used to their fullest—that they are widely distributed and accessible, that data from various sources can be linked together, and that new scientific studies can build on the work of others rather than reinvent the wheel. To this end, we need a new and powerful data infrastructure, spanning across disciplines, research communities, and countries, and creating an unparalleled base of knowledge. Such an infrastructure is central to the efficient advancement of science.

One can perhaps best consider advances in this new, synergistic data network by viewing it as being comprised of three important elements: advances in the sharing and accessibility of existing data sets; the construction, linking, and diffusion of smaller, researcher / laboratory generated data sets; and the merging of publicly available data with administrative and perhaps private records. In each of these dimensions, there is much to be gained and the returns to investment are likely to be enormous. There are certainly hurdles involved, but I believe the potential for gains from such data sharing make the efforts well worth the costs.

Consider the large-scale public use data sets such as the Panel Study of Income Dynamics (PSID), the Current Population Survey (CPS), and the Health and Retirement Study (HRS). These surveys provide a wealth of information on respondents and have been well used across the social science disciplines. However, in the majority of cases, individual researchers start from the raw data files, construct their own composite variables, and "clean" the data for their own use. Disciplines and even individual researchers within a field differ in the assumptions they use when cleaning data or imputing missing values. Similarly, because the specific questions used to measure variables differ across data sets, even seemingly straightforward variables such as assets or tenure on a job are likely to be measured differently across surveys. These inconsistencies in data construction and comparability lead one to wonder whether differences in results across studies using such data are attributable to true differences in behaviors or are instead driven by idiosyncrasies in data handling. A common set of rigorously tested assumptions and carefully constructed measures for each data set, maintained by an expert team of statistical personnel, would lend much more credibility to research using these resources, and would make analyses much more efficient.

On the other end of the spectrum from these large nationally representative data sets are small scale projects with observations generated by individual researchers. These can be laboratory based experiments involving a handful of individuals, ethnographic studies of particular populations, or observational studies of particular interactions. These data are collected by researchers at great effort and often provide the only means of addressing a particular question. Their usefulness and power could, however, be increased many times over if there were a mechanism for researchers to link these data either with other similarly small-scaled samples or with large data bases. One could easily imagine, for example, a sample of individuals for an fMRI study being selected from the set of respondents for a large study such as the PSID and the information eventually being made publicly available through a data clearinghouse. In this case, not only would the research community have the results of the fMRI experiment, but they would have access to potentially decades of financial, demographic, and family data which could be exploited. Certainly there are confidentiality issues to be resolved, but the opportunities to create such a powerful data set make the effort worthwhile. Similarly, a data sharing infrastructure that sets standards, methodological guidelines, and documentation requirements would allow independent investigators with their own fMRI studies to link together their observations, thus creating a substantially larger and more powerful meta-study.

Finally, there are numerous detailed and extremely powerful data sets to which investigators do not currently have access due to their proprietary nature, legal requirements, and confidentiality concerns. Data from marketing firms, institutions, and government agencies are all examples. If these data could be made available in the interest of science, the gains would be enormous. Imagine being able to study health outcomes using data from Medicare administrative records as a function of an individual's purchases at a grocery store (based on an affinity card), and a set of demographic and economic characteristics such as income, occupation and education from a longitudinal survey. There are certainly a daunting array of legal and privacy issues which would need to be addressed, but the potential to learn about important behaviors ought to make an investigation of such possibilities a priority.

All told, the social sciences are at a critical juncture. We have identified important data needs and have the means to collect and analyze these data. However, we are lacking the infrastructure to manage, share, and fully exploit these data. Data are, and ought to be, a public good. They are a key component in our science. A central repository and data handling structure to deal with these issues and to make such data readily available would provide scientific gains that would be nearly impossible to attain through other mechanisms.

SARAH NUSSER – A New Paradigm for Co-development of Population Studies and Emerging Methodologies

The complexity of social, behavioral and economic systems is expressed through the interactions of many contextual dimensions – personal, social, economic, political, environmental, temporal, geographic, among others. Yet limitations in SBE research infrastructure force investigators to eschew this complexity and focus on narrow slices of a SBE systems, resulting in an incomplete understanding of phenomena and enormous gaps in our knowledge of selected population and social systems. In addition, rapidly evolving personal and environmental sensors (e.g., physiological and positional body sensors, neuroimages, audio/video recording, environmental sensors) and data harvesting techniques (e.g., web scraping, transactional captures) offer richer and more direct contextual and response data, but scientifically credible methodologies for incorporating these measurements into population studies are virtually nonexistent.

To engage in sophisticated population level research that fully explores the complexity of social, behavioral and economic questions, we need a far more responsive paradigm for creating and implementing scientific methods for population studies. Using the example of the need for a far richer data source to address questions about rural populations and the social aspects of agricultural practices, I envision the development of a large-scale longitudinal population survey framework to serve as a responsive resource for intensive study of the heterogeneity and dynamics of US rural populations and for cross-cultural/national studies. Below are (incomplete) desirata from a survey perspective.

Responsive design We need a probability sampling mechanism (methodologies and sampling frames) that enables reasonable sample sizes across variation rural populations at different levels of granularity (e.g., temporal – within a day to life span, geographic - neighborhood to regional, economic, social, cultural, institutional, etc.). The design provides the basis for imputation of spatial, temporal, or other kinds of patterns, and for other forms of statistical modeling to generate realistic synthetic data for describing population behaviors or evaluating the impacts of alternative behaviors. It also offers the ability to add smaller, targeted studies to the larger survey in order to test hypotheses about the inner workings of rural populations or subgroups of rural populations, or to collect more intensive measurements (e.g., DNA, special sensors, videotaped behaviors). In addition, the survey design continuously embeds special studies to estimate impacts of survey methods (e.g., panel effects, model effects, change in questions or protocols); to investigate emerging methods to address new data needs and changing survey conditions (e.g., new sensors, data objects or questions; declining frame coverage or efficacy of recruitment methods); and to describe survey quality by quantifying measurement and other forms of nonsampling error. Finally, we embed social experiments in communities that enable in situ study of alternative mechanisms for motivating socially responsible human behavior, e.g., sustainable personal, institutional and community-level behaviors that reduce our carbon foot print and/or lead to conservation and protection of natural resources in agricultural systems.

Contextual data. To facilitate statistical design/modeling and broader research into the complexities of social systems, we link individual records to a completely secure and nationally integrated dataset that includes purchase information, program participation, internet behaviors, cell use behaviors, etc. It is integrated with other contextual data sources that provide depth in key dimensions of variation and support sophisticated modeling of complex phenomena and future sampling efforts. For example, a related set of geographically-linked databases are developed that include information on conditions for individuals and families, as well as contextual data that characterize other aspects of rural areas (e.g., labor force or other economic measures, sources of health risks and outcomes, transportation routes). Finer detail is created by modeling (e.g., small area estimation) coarse data as a function of correlated and more detailed data. These predicted values are integrated with respondent data or serve as auxiliary sampling information.

Evolving survey paradigm. A far more flexible and continuously evolving survey paradigm is critical to addressing complex research. Much more emphasis is placed on careful innovation of survey and statistical analysis methods to address critical domain questions, the potential of new measurement techniques, the need to adjust to changing survey environments, and the importance of quantifying statistical properties of estimates. A more sophisticated survey program: places less emphasis on ownership of the survey or on constraints generated by unwillingness to take prudent risks in developing the survey environment; is run by a scientific and methodological consortium that transcends the current stove pipe approach implemented by a federal agency or survey vendor, and is driven by active collaborations between knowledgeable domain research scientists, measurement specialists and statisticians; has an explicit mechanism for embedded and ongoing research into statistical design, measurement, and analysis methodologies; includes a competitive mechanism to keep resources focused on the most important subject matter and methodological questions; and involves top survey statisticians and methodologists with implementation experience to carefully develop a survey design that balances statistical criteria and the practicalities of collecting data from humans or other sensors to pursue both subject matter and methodological research questions.

KEN PUGH

I would like to propose a national cross-disciplinary database project to promote an integrated cognitive neuroscience of language and human communication: opportunities and challenges. A more fully i**ntegrated cognitive neuroscience** of individual differences in language skills (with attention to life-span changes) is a high priority for researchers from several disciplines including linguistics, psychology, neuroscience, genetics, and communication disorders. The extant cognitive neuroscience literature, at multiple levels of analysis, is fast growing, but still underpowered. At the level of genetics, a number of candidate genes have been proposed that may be associated with individual differences in language competence; at the level of brain systems, structural and functional findings at key left hemisphere brain regions have been associated with individual differences in language skills across different age groups and across different languages. Sample size is restricted at any given lab, and each tends to recruit very different populations depending on aims. There is really a pressing need for integration across research sites and projects to get the kind of power required for gene-brain-behavior linkage in the language domain. We propose the following NSF sponsored (and possibly NIH co-sponsored) project to begin to get the scale and scope needed for building a representative national database. Current and future NSF (and NIH) funded cognitive neuroscience projects can be encouraged (with adequate, but relatively small financial incentives) to test a subset of participants with common measures including: DNA recovery, structural MRI measures, yoked fMRI/EEG measures (with short language “localizer” tasks examining basic spoken and written language processing skills), and relevant behavioral assessments and demographic information. Since each lab is already studying different ages, populations, and language backgrounds, rich variation (and representativeness) can be built into this national database at low cost and manageable efforts. Examples of cohorts that might be targeted in this initiative include: typically developing across different age groups, language disordered groups, and bi-lingual vs. monolingual groups. If successful in the first phase, in a later phase we might identify international cooperative mechanisms that would allow us to acquire cross-language comparative data for this database as well. The separate projects contributing to this database of course each have their own priorities and methods, and this would not supersede them, but incentives to include a subset of common measures, analyses, and sampling decisions for this national resource would allow both the needed scale and ranges of individual difference dimensions demanded for theoretical progress (again, this suggestion also begins to make sense when considering the value of extremely large samples for gene/neurophenotype studies). Without common measures and protocols, it would be difficult to aggregate data across sites. Of course, standardization of measures and methods is not without scientific risk (squelched innovation) but a balance is needed for analyzability; the current plan simply adds a sub-aim to contributing sites without forcing changes in methodological approaches for major site-specific goals. In sum, there is an acute nee for increased scale and scope for research on individual differences in language at multiple levels of analysis; in order to obtain necessary power for this multi-site collaborations and a national database, available to all researchers, should be a high priority for NSF .

As part of a this database initiative there will be an strong need to focus some funding resources (perhaps with targeted RFA's) on the following issues in order to properly develop and exploit the national database**:**

1) improved systems-level neuroscience statistical methods
2) improved multi-modal imaging synthesis methods (e.g., EEG, structural MRI, fMRI)
3) a new generation of computational modeling and neural simulation approaches
4) links to extant genetic databases
5) cross-language comparative database development

This plan, while a challenge to current infrastructure (and perhaps to traditional NSF and NIH territories), would allow researchers from cooperating disciplines access to a powerful resource for their own research, and would spur next generation methods and tools necessary to deliver on the promise of integrated gene-brain-behavior approaches to the study of language and human communication.

DEAN SNOW

It is important that we divide our concerns into strategic, tactical, and technical categories. Failing to do so will cause us to flounder in unproductive effort. Most of what workshop participants submitted in writing in advance of the meeting, and much of what was said at the meeting, was focused on tactical concerns. These concerns are important, but they tend to deal with parochial matters of data gathering, sampling, analysis, and preservation surrounding the traditional goals of research in specific disciplines. I judge that it is important for the workshop to break new ground rather than merely concentrate on new ways to make traditional disciplinary problem solving and hypothesis testing easier and more efficient.

To that end I think that the final report should focus on strategic and technical issues, leaving tactical ones to the individual disciplines. To facilitate further progress I suggest that we begin by distinguishing between three kinds of disciplines:

1. Disciplines that mostly need new infrastructure for data acquisition (eg. survey research),
2. Disciplines that mostly need new infrastructure for data preservation (eg. historical sciences), and
3. Disciplines that need new infrastructure for both purposes (eg. demography).

When proposing clusters for our breakout groups, Brian MacWhinney defined 5 categories. He also suggested breaking the discussion up according to "new data" and "existing data," then discussing both tools and organization under those headings. He might have reversed the outline but that does not matter because the suggestion was basically that we fill in the blanks of a four-cell matrix like the one shown below.

	New Data	**Existing Data**
Tools (Techniques)		
Organization (Strategy)		

This suggestion was passed over quickly but I think that we nearly missed a good opportunity. I recommend that we use this as a framework for organizing our thinking on ways to develop and deploy new tools and ways to strategically improve organization of what we do in the SBE disciplines. My own breakout group dealt with "data source generation through viral applications, and getting data out of silos." We had some ideas for new "viral" tools that would fit will within one of these cells. I'm sure that practical ideas coming from other breakout groups could also be productively assigned to one of these organizational cells. There were lots of good ideas and the immediate goal is to simply organize them for the purpose of identifying next steps.

DALTON CONLEY – Meet the 21st Century Nielsens

Surveys were the dominant form of data collection for 20th Century social science. But as the information society has saturated us—i.e. we, potential survey respondents—with ever more sales calls, push polls, spam and other nuisances, survey researchers have an increasingly tough time getting folks to talk to them. To make matters worse, respondents may have gotten savvier at figuring out what interviewers "want" to hear. For example, while overt racial prejudice has precipitously declined since the 1960s—at least according to national studies such as the General Social Survey (GSS)—recent research that deploys experimental approaches to ferret out respondents' "true" feelings suggests that a good portion of this decline is, in fact, the result social desirability bias. So what is the social scientist to do?

The good news is that the potential for a whole new world of data collection is has now arrived on the scene. In the 21st Century, what people do—rather than what we ask them about what they do—will become ever easier to document and analyze. The death of privacy in the digital era brings with it the silver lining that social scientists can now record patterns of human behavior that would have been well beyond their reach in the era of telephony, fill-in surveys or face-to-face interviews. What's more, through the smart use of new technologies, we can achieve a level of granularity that is almost unimaginable.

Already some academic researchers—and many private firms—are making use of a wealth of meta-data that is created automatically through our credit card purchases and our internet browsing history, for example. Others are reaching obscure subsamples of the American (or international) population through web-based response forms and social networking applications. Others are even using the Internet to conduct large scale social experiments to test the effect of advertising on sales, initial rank order on subsequent popularity, and social network connectivity on behavioral choices.

The inherent limitation to much of this work lies in the fact these are self-selected samples of dubious generalizability in most cases. Yahoo! may not care whether its users are representative of the U.S. population or even of internet users. As a private firm, they only care about their clients' behavior. But social scientists do care about such issues as selection bias and external validity.

This is where the need for cyberinfrastructre for 21st Century social science can be felt. The research community desperately needs a representative sample of cyber-Nielsens. Already, there exists the NSF-funded Time Share for Social Sciences (TESS) run by Knowledge Networks—a sample of households that receive free Internet service in return for answering some questions on-line each month. The value of the dataset grows as each question is added to the existing matrix of information on the respondents. As valuable as this effort is, it still amounts to a 20th Century approach to data collection—relying on the active, conscious participation of subjects who are consciously aware when they providing the data to be analyzed.

A newer approach would entail the NSF recruiting a national sample of smart-phone (i.e. internet enabled cell phones such as the BlackBerry or iPhone) users who—in return for free service—would agree to have their usage archived in a de-identified database for analysis by authorized researchers (cleared in a process similar to that used currently by the Census Research Data centers). With technologies currently available, collected data could include everything from physical location (through a GPS application) to social network (through archives of calls and texts) to environmental data (through embedded sensors for particulate matter or other airborne compounds) to actual voice conversations for linguistic parsing to para-data (such as responses time, typographic errors and so on). Such a data collection approach would blur the old line between qualitative, ethnographic research and quantitative, statistical approaches while minimizing the white coat (or Hawthorne) effect that haunts so much of social science research.

The data archiving needs would, of course, run to the terabyte level. The human subjects considerations are far from trivial, and the coding and coordination efforts immense. But such an endeavor would constitute nothing short of a social scientific Apollo Project worthy of the information age.

LYN VAVRECK

My recommendation is to invest in the collection of new and existing data on politics, neighborhoods, and society that are currently unavailable to (or uncollected by) scholars due to cost or organization constraints. Through a coordinated effort, we could collect state administrative datasets such as voter files, U.S. post-office data on changes of address, precinct boundaries, and state registration-laws just to name a few; and third party data such as home values and consumer information; and we could leverage census data to indicate neighborhood composition and other observational features of blocks, towns, or larger geographic units. We could even leverage databases like the Yellow Pages to characterize neighborhoods in terms of the commerce located in the area; or use surname analyses to characterize the ethnicities of neighborhoods. This effort is an attempt to take context seriously because I believe it is important; and I believe we are not going to measure context through survey data as well as we can measure it with other types of data that are difficult to put together but actually exist already. The idea would be to gather all of this existing data in one central location that would be easily searchable and usable by scholars across disciplines – and where scholars could archive the data they collect along these lines as well.

To this existing data, we could add data from campaigns about schedules, visits, advertising, mailings, and get-out-the-vote efforts. And, we could add observational data in the form of a continuous monitoring survey that would have a similar setup to the cooperative projects I have fielded. In these projects the first half of the content (10 minutes) is always the same, this ensures time-series comparisons, but the second half of the survey is malleable and extensible in creative ways. This survey would be done in off-election years to try to understand how people think about politics, their neighborhoods, and the choices they make when there isn't a presidential campaign going on in full force. In many ways, we cannot understand the effects of campaigns until we study people outside of the campaign environment. By marrying the survey data with the contextual data, we can begin to understand how context shapes attitudes and behaviors.

In Political Science we are just beginning to take context seriously (in terms of elections and decision-making about policy). I would like to "nudge" the discipline into measuring context through direct operationalization of the concept. We have relied on survey data for too long in studying elections, behavior, and decision-making and I would very much like to support the move toward relying on behavioral or non-survey measures of the type described above. Most of these data exist already, they are just too expensive for academics to purchase from the third party vendors or they require too much work and coordination for one person to organize or automate. It's exactly the kind of work that can be accomplished with a little money and manpower; and the payoffs would be quite large.

7. Participants

Helen Aristar-Dry
Co-Director, Institute of Information
2000 Huron River Dr.
Suite 104
Ypsilanti, MI 48197
hdry@linguistlist.org

Bennett Bertenthal
Dean, Arts and Sciences
Indiana University
Kirkwood 104
Bloomington, IN
bbertent@indiana.edu

Henry Brady
Dean, Professor of Political Science
University of California
2538 Channing Way
Berkeley, CA
hbrady@berkeley.edu

Christopher Cieri
Executive Director
Linguistic Data Consortium
3600 Market Street
Philadelphia, PA, 19104-2653
ccieri@ldc.upenn.edu

Dalton Conley
University Professor, Chair, Sociology
New York University
295 Lafayette Street,
New York, NY 10012
dc66@nyu.edu

Robert M. Groves
Professor of Sociology
Director, Survey Research Center
University of Michigan
Ann Arbor, MI 48106
BGroves@isr.umich.edu

Edward J. Hackett
Professor, School of Human Evolution & Social Change
Arizona State University
Tempe, AZ 85287-2402
ehackett@asu.edu

John Haltiwanger
Professor of Economics
University of Maryland
College Park, MD
20742-7211
Haltiwan@econ.bsos.umd.edu

Kenneth Kidd
Professor of Genetics, Psychiatry, and Ecology & Evolutionary Biology
333 Cedar St
New Haven, CT 06510-3206
Kenneth.kidd@yale.edu

Kathleen McGarry
Department of Economics
Dartmouth College
and UCLA
Hanover, NH
mcgarry@ucla.edu

Brian MacWhinney
Professor of Psychology
Department of Psychology
Carnegie Mellon University
Pittsburgh, Pennsylvania 15213
macw@cmu.edu

David Mark
SUNY Distinguished Professor
Department of Geography
State University at Buffalo
Buffalo, NY 14261-0023
dmark@buffalo.edu

Sarah Nusser
Professor, Department of Statistics
Center for Survey Statistics
Iowa State University
Ames, Iowa 50011-1272
nusser@iastate.edu

Kenneth Pugh
Research Scientist
Department of Pediatrics (Neurology)
15 York St
New Haven, CT 06510-3221
pugh@haskins.yale.edu

Deb Roy
Associate Professor of Media Arts
MIT -- The Media Laboratory
77 Massachusetts Avenue
Cambridge, MA 02139-4307 USA
dkroy@media.mit.edu

Dean Snow
Professor of Archaeological Anthropology
Penn State University
403 Carpenter Building
University Park, PA 16802
drs17@psu.edu

Lynn Vavreck
Assistant Professor, Political Science
UCLA Department of Political Science
4289 Bunche Hall
Los Angeles, CA 90095-1472
lvavreck@ucla.edu