Kavli HUMAN Project

Preliminary Study Design

About:

What are the roots of human behavior? How do we make decisions, and what forces shape those decisions? How can we apply what we learn about human health and behavior to inform and improve public policy to make it more effective? The answers to these questions lie in an entirely new way of studying human beings, taking advantage of technological advances and the big data revolution to provide a previously unattainable interdisciplinary data platform to researchers.

Traditional longitudinal studies have been focused on specific domains of inquiry or subsets of the population. In contrast, the Kavli HUMAN Project is the next step in bio-behavioral and longitudinal research by measuring all of the biological and behavioral characteristics that make us human at once, effectively quantifying the human condition.

The Kavli HUMAN Project's data platform will provide researchers with an unprecedented volume and diversity of datatypes to break down the silos between disciplines like neuroscience, genetics, psychology, medicine, and urban informatics in order to unlock new insights into the feedback mechanisms between biology, behavior, and the environment that make up the bio-behavioral complex. Over time, the Kavli HUMAN Project will not only enable groundbreaking science, but provide the context and knowledge to craft evidence-based public policies and improve societal outcomes.

Kavli HUMAN Project

Preliminary Study Design Report

NEW YORK UNIVERSITY
Institute for the Interdisciplinary Study of Decision Making
October 15, 2015

Kavli HUMAN Project
Brooklyn, NY
www.kavlihumanproject.org

ABSTRACT

While the United States alone has a long history of gathering data on its population, going back at least to the census of 1790, longitudinal studies of human populations conducted world-wide have almost uniformly been hand-crafted, small-scale enterprises aimed directly at answering a limited number of predefined questions. The U.S. Health and Retirement Study (HRS), perhaps the broadest and most influential longitudinal survey of all time, gathers surprisingly limited data on financial, social and health matters for older adults once every two years. It is aimed at understanding the impact of aging on social and physical well-being. But despite these limitations, the HRS has revolutionized our understanding of many aspects of aging. A review of longitudinal studies funded by the Kavli Foundation in 2013 concluded that the use of large-scale automated methods for capturing biological and digital data at a synoptic level could revolutionize nearly all aspects of the social and natural sciences, and that such a study was rapidly becoming financially feasible. That review proposed the development of a study methodology that would yield roughly 100x the data density of existing studies. It concluded that an ideally designed platform could achieve this data density at roughly one-fiftieth the cost per data point of existing methodologies.

This report presents this new study methodology, and proposes its use for initiating a highly detailed synoptic longitudinal study of 10,000 individuals in roughly 2,500 households over 20 years in New York City. All aspects of the study methodology are developed in detail from study frame design to data gathering technologies. It is proposed that these detailed data about individuals be complemented with a comprehensive archive of the environmental conditions and events currently being aggregated in New York City. It is then proposed that this combined longitudinal dataset be made accessible to the research community in order to facilitate previously unattainable advancements in medicine, biology, psychology, economics, sociology and anthropology while also fostering evidenced-based public policies. A series of White Papers by leading scholars included in the appendices of this report details the potential benefits of such a study.

Our proposed automated methodology would continuously collect an array of data that includes: genome, microbiome, metabolome, epigenetics, medical records, drug and chemical exposure testing, diet assessment, sleep behavior, physical activity levels and home air and noise quality. Personality traits, IQ, mental health status, long term and working memory, social network structure, communication partners and patterns, continuous location capture, educational history, employment data and detailed financial transaction records are also proposed amongst other data.

Were such a study to be initiated, it would provide for the development of a revolutionary platform for improving the human condition. Its development in New York City, with existing extensive city data sets, would provide the most detailed synoptic discovery data set on humanity ever gathered. This report provides a Preliminary Study Design for the development and deployment of this platform.

SUMMARY

Study Need

While there is a wealth of overlapping longitudinal studies being conducted today in the developed world, no methodology currently exists for gathering a truly synoptic dataset. The result of this limitation is that individual studies are designed to focus narrowly on a small set of measures that are hoped to yield insight into a single domain. In the 1970s, this was also true in many of the physical and genetic sciences. An astronomer hoping to gain insight into black holes manually searched the heavens for individual objects, data about which could be used to build a small database for the study of black holes. A geneticist hoping to explain psychosis might similarly examine single genes in laborious detail in an effort to build a local understanding of the roots of psychopathology. In the 1980s, many disciplines began to abandon this approach, however, in favor of more synoptic and automated methods. In 2000, the Sloan Digital Sky Survey (SDSS), for example, began the automated scanning of deep space in order to build a massive database that today profiles over 500 million celestial objects. The Human Genome Project (HGP) began the development of tools for automated large-scale genetic analysis in 1985 that now have yielded massive databases on the genetic structure of hundreds of organisms. Today, astronomers and geneticists achieve daily progress with these community resources that would have been unimaginable prior to their development.

Unfortunately, the synoptic study of human beings has proven more technologically difficult than either the study of the cosmos or of the genome. Human behavior, human biology, human development, human economic activity and human social interactions have all been studied successfully, but with labor-intensive tools that preclude the measurement scale required for a synoptic project like the SDSS or the HGP. Recently however, that has begun to change. Driven by corporate interests and rapidly maturing digital technologies, detailed measurements of nearly all aspects of human life and behavior are now being made in a piecemeal way by a host of large corporations. Were it possible to aggregate these existing technologies and develop a few key tools for supplementing them, a synoptic study of humanity could be possible today, at a surprisingly low cost. We estimate that the development and deployment of a synoptic study of humanity in an initial urban cohort of 10,000 could be achieved for an initial cost of approximately $10M/yr. Like the HGP, we believe that once the basic platform is developed, costs for future deployments could drop by at least a factor of 2 within 5 years.

Were such a database, or such databases, available to scholars, there is compelling evidence that a host of currently intractable medical, psychological, developmental, economic and sociological problems could be successfully engaged. Such a longitudinal dataset would, for example, definitively identify the many interlocking causes of our obesity epidemic. It would finally allow us to determine the critical factors in successful childhood development. It would define when and why people suffer from Alzheimer's disease by uniting data about environmental exposure to pathogens, diet, behavior, medical care and genetics in a single longitudinal database. It would identify the roots of psychopathology, financial success and educational attainment at unprecedented depth, all at a very low initial cost.

Management Plan

To achieve these goals we propose the development of a study we refer to as the Kavli HUMAN Project (**H**uman **U**nderstanding through **M**easurement and **AN**alysis) (KHP). We propose building the study around a variant of the Department of Energy's "Stage-Gate" process for large-scale study management. In this document we specifically propose a five-stage process: Stage 0: Study Need and Feasibility; Stage 1: Preliminary Study Design; Stage 2/3: Complete Study Design; Stage 4: Construction; Stage 5: Deployment. The Kavli HUMAN group delivered a Study Need and Feasibility analysis to the Kavli Foundation Board in October of 2014. This current document constitutes the Preliminary Study Design required next by our management plan.

Conceptually, the study development and execution process is organized around five principal divisions:

- Measurement and Technology; (What we propose to measure and how)
- Study Frame Design; (How we propose to select, recruit and retain subject-participants)
- Privacy and Security; (Security for physical and digital data, as well as participant consent)
- Education and Public Outreach
- Scientific Agenda; (How data collection meets social and scientific needs)

Each division of the study is currently governed by an *Advisory Council* of 10-15 leading scholars, technologists and social activists with a lead individual who serves as chairperson. Each of the five boards provides support to five complementary in-house staff groups responsible for executing all aspects of that study division. Each internal division is lead by an outstanding practitioner who works with their own staff to accomplish study goals in concert with the relevant Advisory Council.

Overall management of the study is achieved by the Governing Board of the study, through the Study Director who provides day-to-day management of the project. All five Advisory Council Chairs, the Study Director and five to ten additional members sit on the Governing Board of the study. Finally, the Director manages the study in concert with a Chief Scientist and an Administrative Director who oversee scientific and managerial issues, respectively.

Measurement and Technology

The utility of the entire undertaking proposed here rests on the notion that a fairly complete synoptic profile of the same 10,000 individuals can be achieved over 20 years at a reasonably low cost. A critical feature of this plan is that it must be accomplished in a sufficiently non-invasive manner that subject-participants both remain in the study and remain largely uninfluenced in their daily behavior by the study.

Several key assets make this possible. First, the development of digital data technologies by the private sector means that nearly all of the digital data we would require can now be gathered non-invasively and at scale. Second, the rapidly falling cost and wide availability of technologies for biosample analysis make a huge range of physical sampling technologies feasible at scale. Together, contemporary digital and physical sampling technologies place a detailed characterization of individuals within reach, given minimal technological and platform development. Third, the recent development of large-scale Geographic Information Systems (GIS) makes it possible to combine the proposed datasets with very detailed characterizations of our subject-participant's environment at a micro-level. GIS-based municipal and academic data about everything from local weather to hyperspectral image-based assessments of environmental toxins at the resolution of a city block can be combined with the participant-level data to complete the proposed dataset.

GIS databases of this kind are being developed throughout the world today. Although the most complete such database is now being built in New York City, comparable databases are under construction in many areas and the number of such

databases is predicted to grow radically in the next 5 years. While this makes New York City a logical site for platform development, we expect that within three years a number of additional sites will likely be available. Finally, we note that the 10,000 participants are embedded in roughly 2,500 households. Because families and households are studied together, many of our measurements (for example genetic measurements) have much higher statistical power than would be expected in a sample of 10,000 independent individuals. This sampling structure also assures that developing children are studied within the matrix of their families.

Demographics and Home Environment

The proposed measurement systems begin with standard demographic data about each participant of the kind gathered in a traditional study. This is supplemented by easily achieved measurements of the home environment, both derived from publically available databases and from a simple survey made by study personnel at the time of participant recruitment. NYC public datasets and GIS databases provide additional neighborhood-level data at a spatial precision of a single building lot and a higher than daily temporal precision.

Biological/Medical Samples

At the time of enrollment, and at roughly 3-year intervals thereafter, biological and physical data is gathered on each participant. We note that this is the most invasive (and expensive) part of the proposed study, and these measurements take place over the course of half a day. During this period we obtain a basic medical exam; a blood sample for genetics and blood chemistry; a urine sample for toxicology; a saliva sample for oral microbiome, genetics (in children) and salivary hormone analysis; a hair sample for toxicology and a stool sample for gut microbiome.

These biological samples are supplemented with medical data from the participant's electronic medical record, to which the study gains continuous electronic access. We also gain ongoing electronic access to all medical-diagnostic codes filed with the state and state prescription records.

Educational-Occupational

In addition to gathering traditional educational data, we propose to gain ongoing electronic access to Department of Education data on all young subjects and on the schools attended by our subjects. Similar data on college performance, when applicable, will be gathered from commercial databases. Traditional occupational data will also be gathered in an ongoing manner.

Financial

A detailed financial profile will be gathered at enrollment but more importantly, ongoing financial data will be gathered automatically at a daily level. Automatic web-based systems will allow data at the transaction level to be acquired for the vast majority of our participant's economic activity.

Smartphone

Perhaps the single most important data collection tool employed by the study will be each participant's smartphone. Smartphones allow us to gather a host of data about the participants including: continuous location data, activity data, and social network data from email, telephone and texting. The smartphones, and an accompanying Bluetooth beacon technology we are developing, also provide information on how and when family members – including children – interact. The smartphones also allow us to invite participants to complete very brief psychological and social questionnaires offered on a weekly basis. Aggregating data over many years these questionnaires provide an unparalleled data stream – the details of which are described in the body of this report.

Summary

The sections above capture many of the data streams that the study can gather. The detailed structure of our measurement process, how acceptability of these measurements to participants is assessed and maintained, and a host of additional measures, are discussed in the body of this report.

Study Frame Design

In order for the study to yield a meaningful characterization of the society in which it is conducted, it is essential that it be based on a randomized sampling strategy. Like the U.S. Health and Retirement Survey and other successful large-scale longitudinal studies, the proposed study must yield a quantitatively tractable cross-section of the study population. For the purposes of our initial study we propose here a cross-section of the city of New York, including all five boroughs. Restricting the initial study to New York City provides several advantages: it allows a sample of only 10,000 individuals to provide significant statistical power because it places them within a limited set of environments, it makes participant recruitment and tracking geographically tractable at low cost, and it allows us to leverage the data infrastructure of New York which is more advanced than any other urban environment at this time.

In order to maximize the power of the measurements we make on these participants we propose to study entire households. Our proposal is to begin by randomly selecting, using a statistical model of the city, roughly 2,500 "seed" individuals. The randomly selected seeds identify roughly 2,500 households – each of which becomes a participant family. Seeds are selected until a total study population of 10,000 is reached. Our analysis indicates that successful recruitment of 10,000 participants can be accomplished in 30-36 months at the proposed cost-level.

The sampling strategy is discussed in detail in this document. It rests on a flat sample of households anchored to a geographic study frame at the single dwelling level. A statistical oversample of three groups, young children, pre-teens and elders, is specified. Details of our sampling frame and model design are also discussed in this document.

Like many areas in the United States, New York is a multi-lingual city. We propose conducting the study in both English and Spanish. This allows us to reach better than 88% of New York City households at the high school language level. New York City also includes a significant group of Chinese and Russian speakers (an additional 6% of the population) as well as a host of other languages. We propose to develop a "Simple English" version of the study to increase the probability of reaching members of these difficult to reach populations.

Our proposed recruitment procedure is based on a modification of the standard Dillman method. This involves a recruiter making multiple incentivized contacts with a candidate household prior to recruitment. Our proposed method is a lengthened and more incentivized version of the technique employed by the Health and Retirement Survey that achieves roughly 85% recruitment rates in their population. The recruitment teams will be fielded in-house rather than being subcontracted to maximize our ability to develop and control the study platform. Our goal is to achieve a 25% recruitment rate in our population with our in-house teams.

We propose to employ three types of incentives: money and items that have a real monetary value, items and activities that build community good will and giving participants personal and population-level feedback and data. The literature shows that providing even a small monetary incentive (rather than none at all) creates an increase in response rates. Results from other longitudinal studies suggest that incentives other than money engender goodwill and feelings of trust with respondents, as people who feel part of something are more likely to continue participation, and over time the financial cost of these items is often less than would be spent for monetary incentives. Existing studies suggest that using these incentives should enable us to maintain a roughly 95% annual subject retention

rate after initial subject attritions. Treatment effects engendered by our incentive structure are discussed in the complete document.

Privacy and Security

The goal of the Kavli HUMAN Project is to create a resource for scholars that can be used to address a wide range of basic research and policy questions. However, this comprehensive data set will contain a wide range of sensitive material, including personally identifiable information, so access must be balanced against the need to protect the security of the data and the privacy of the participants. Successful management of the rich store of electronic data and biological samples is critical to the success of the project and thus, we have invested substantial resources in the design and implementation of this aspect of the study.

Our intention is to provide the highest possible degree of privacy and security to our subjects. To do this we propose to implement the NIST SP 800-53 security framework. This is a superset of the regulations required by the HIPAA and FERPA acts. It also details security protocols, data handling and data accounting procedures. In essence, our security plan rests upon a dual firewall system. Raw data resides inside a super-secure inner firewall accessible only to system operators. Data inside the inner firewall is partitioned into roughly six secure independent subdomains. Physical and digital access to data inside the inner firewall is only available after two-factor biometric identification of a sysop. Access to the data can, by hardware design, only be accomplished inside the physically secure data facility after two-factor identification of operators at the physical level. All system-wide operations require the consent of two system operators who have been physically and digitally two-factor identified.

In order to provide scholars with easy access to the dataset, we propose to provide a mechanism for scholars to request an anonymized subset of the data we refer to as a *datamart*. Datamarts are the object with which scholars interact, and they reside inside the outer firewall of the KHP. Datamarts, by design, do not contain personally identifying data. Scholars perform their analyses on datamarts to extract causal relationships, test hypotheses or to engage in data discovery operations. The product of these analyses can be exported outside the facility firewalls for publication or further analysis. All datamarts are destroyed after the study for which they were requested is complete. In addition to these secure data-types, fully anonymous statistical profiles of the database will be made available on the web.

In order to insure that none of the data gathered and stored can be subpoenaed, the data will be covered by an NIH Certificate of Confidentiality (NIH CoC). The NIH CoC insures that all of the data resident in the data storage facility is private and confidential, even in the face of a government or other third party request.

To ensure data survivability and recoverability, the entire database will be routinely encrypted using the AES 256-bit technology. The database will then be fragmented to non-recoverability and stored off-site in a 10-partition/10-location fragmentary backup via point-to-point transfers to remote backup servers. Note that no element of the database ever leaves the facility without these two stages of encryption: 256 bit AES encryption followed by encrypted fragmentation.

Finally, an external "red team," tasked with breaking down our security protocols, will continuously test the security of our system. This external group, discussed in the full document, is incentivized to identify weaknesses in our security system by placing it under continuous attack.

Education and Public Outreach

Education and Public Outreach (EPO) is a critical component of the proposed study, especially given the novel nature of the proposed study at this time. But for the KHP to reach its potential, open lines of communication must be fostered with all of the project's stakeholders – scholars, potential

participants and potential funders in both the foundation and government communities.

Probably the greatest challenge to the study is to ensure that potential participants understand the importance of the study and feel confident about the study's ability to protect and respect them. Many stakeholders initially express great concern over the highly detailed nature of the data we propose to gather and many scholars express concern that it will be difficult to identify participants willing to undergo this level of scrutiny. One of the central missions of the EPO team is to make sure that subjects understand that nearly all of the data aggregated by the study is pre-existing. One example of such a data type currently collected by existing technology is geo-location. Many potential subjects do not realize that corporate actors continuously track all cell phones with an accuracy of meters. One goal of the EPO team is to make sure that potential subjects view the study as aggregators of existing data rather than as creators of novel and worrisome datasets.

Perhaps even more important is that the study be seen as a public actor working for the good of the community. Potential subjects often express concern and discomfort with corporate data gathering. To eliminate this concern the EPO team will have to engage community leaders, and socially visible figures to insure that the goals of the project are well understood by potential participants.

The EPO team also must engage a number of other communities ranging from scholars to policy makers, and each will require an individualized message. Specifics of the task faced by the educational and public outreach unit are provided in this document in significant detail. Media plans for untoward events are also addressed.

Scientific Agenda

The proposed dataset is fundamentally an enterprise of *discovery* science. The goal of building the database is to provide a general-purpose tool, like the SDSS map of the cosmos, which can be used to understand almost any aspect of human behavior and biology. But it is equally important to demonstrate both *how* a database of this kind could be used and *what* a database of this kind could contribute. By developing 'use cases,' the builders of the database can develop a better understanding of what exact form the database should take. Perhaps just as importantly, powerful use case examples can reassure funding agencies that databases of this kind will quickly yield powerful translational insights.

For that reason the Scientific Agenda group is tasked with identifying leading scholars and asking then to consider how a database of this kind could be used to revolutionize their fields of study. Scholars report their conclusions in prospective analyses we refer to as "White Papers." This document includes information on over a dozen such White Papers from a range of disciplines. For example, a leading team of clinical scientists from Harvard University explains how such a database could provide the critical tool for developing a new kind of "Real Time Medicine" that could be deployed in any population of patients. Another team from the University of Michigan and Harvard University explain how such a dataset could finally explain how and why cognitive decline proceeds at such different paces in different individuals. A team from Washington University and New York University describe how this dataset could resolve fundamental questions about who is at risk for different forms of substance abuse. A team from the University of Washington and Harvard University explain how this dataset could be used to get to the heart of the problem of obesity. A team from the University of Michigan and Northwestern University explain how these data could resolve important questions about what kinds of educational options are best suited for different kinds of students. Other papers explore the importance of "field studies" of neuroscience, the need for longitudinal studies of cognitive function, the prevalence of child abuse and the design of cities and the nature of psychological development in adolescents.

Ongoing work by the Scientific Agenda group seeks to expand the domains of inquiry we examine so as to test and retest the utility of the current study

design. Details of these high impact uses of the proposed database can be found in the Scientific Agenda section of this document. Complete copies of many of these White Papers can be found in Appendix K.

Appendices

The complete document also includes a detailed risk analysis and risk management plans for the proposed study. This risk analysis examines and evaluates all possible risks to the study and its participants. Risk mediation plans are proposed for all high risk items.

A detailed acquisition plan is also provided. This takes the form of a preliminary year-by-year budget for the proposed study. It identifies all costs and relates these costs to proposed funding sources. The acquisition plan precedes a more detailed financial model that will be developed for the final study design.

Specific reports from each of the Advisory Councils are attached. These documents reflect preliminary meetings by the Councils and served as the basis for developing the detailed documents in each of the five study domains.

Some of the measurements proposed in the Measurement section rely on novel technologies developed in house and a detailed assessment of existing wearable technologies. A report on the novel Bluetooth technologies, which are of particular importance for the study of younger children is the first of these sections. The second provides a detailed study-specific assessment of the utility of existing wearable technologies for passive data gathering.

A report by the leading privacy and security law firm Hogan-Lovells, commissioned by the study, is included to provide an external assessment of the privacy and security risks faced by the study.

A proposal from the Synack Corporation for the development of "red-team" IT capabilities to ensure ongoing testing of the database is appended next.

An overall communication plan commissioned from the leading market research firm Berlin-Rosen is appended to provide an external view of the risks and challenges faced by the study in its external communications.

A summary of the acceptability of intensive data gathering of this kind to potential subjects is presented. This independently commissioned study was funded by the NIH in order to assess the willingness of subjects to participate in the NIH Precision Medicine Initiative. It was conducted by the market research firm GfK and provides an excellent overview of subject acceptability issues for English and Spanish speaking US populations.

Biographies for the members of the KHP governance team project are also included to provide a view of the prestigious scholars who advise on decisions about study design and implementation.

Finally, the Appendices conclude with a set of representative Kavli HUMAN Project White Papers developed by the Scientific Agenda Group.

Conclusion

Based on this detailed analysis the study group has drawn the conclusion that it is now possible to develop a platform for large-scale data gathering in a representative human population at an acceptably low cost. Like other large-scale discovery projects conducted over the last three decades, this platform could revolutionize our understanding of its study target: human beings behaving in natural environments. Our analysis suggests that although this could be one of the highest impact discovery dataset projects ever undertaken, it would also be one of the cheapest. Indeed, were the platform to be developed, it would likely reduce the cost of data gathering about humans by a factor of 20-50 times. The existence of such a resource would also eliminate the need for the many partially

overlapping studies conducted today, suggesting very high overall cost savings.

More importantly, the study would advance a completely novel depth of understanding, of the kind accomplished by the Human Genome Project, in the domain of human brain, behavior, biology and society. While it seems clear today that this deeper understanding would yield many societal and scholarly benefits of the kind documented in our White Papers, it is impossible to say with any certainty where the revolution engendered by such a study would stop. Just as it would have been difficult in 1990 to imagine the genetic revolution of today brought on by the Human Genome Project, it is very difficult to say where the HUMAN Project will lead us. As Prof. Steven Koonin, a former Provost of CalTech and former Under Secretary of Energy for science recently put it:

> *"Great science always comes from new instrumentation. When Galileo first turned the telescope on the heavens it opened up a great vista for understanding our place in the universe. When van Leeuwenhoek first looked at a cell through the microscope it opened up a whole vista of understanding biological systems. I think that the Kavli HUMAN Project, with the technologies, the data analysis, and the understanding we have of humans now, has the potential to do the same sort of thing for individuals and the society that is made up of them."*