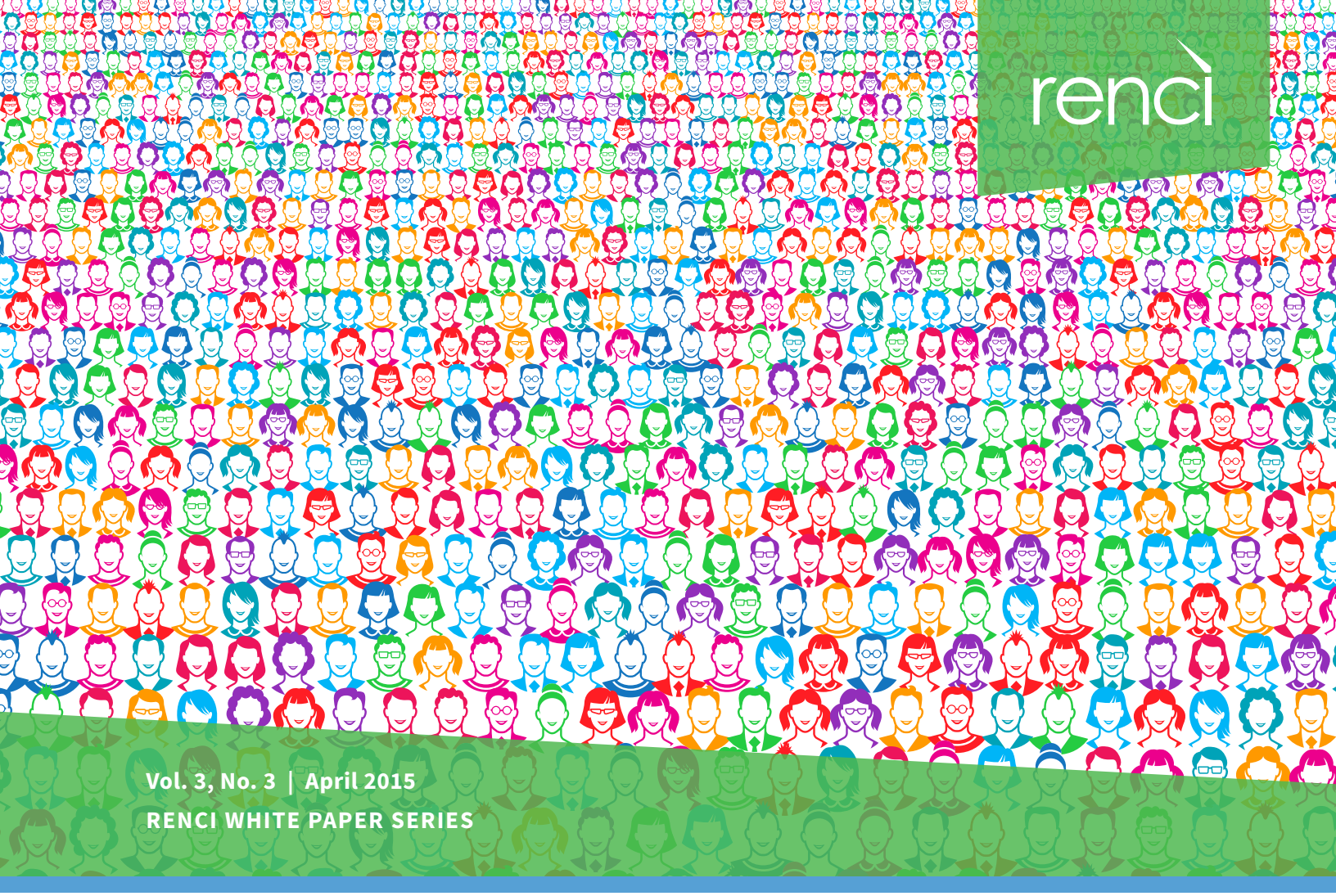


Vol. 3, No. 3 | April 2015
RENCI WHITE PAPER SERIES

The Virtual Institute for Social Research (VISR)

BRINGING BIG DATA TO THE SOCIAL SCIENCES

THOMAS M. CARSEY
Director, Odum Institute for Research in Social Science and Thomas J. Pearsall Distinguished Professor in Political Science, UNC

JONATHAN CRABTREE
Assistant Director for Archives and Information Technology, Odum Institute for Research in Social Science

CHARLES P. SCHMITT
Director of Informatics and Chief Technical Officer, Renaissance Computing Institute (RENCI)

Additional writing support from Anne Frances Johnson and Kathleen Pierce.

Summary

"Big data" is frequently cited as a potential business opportunity for companies, from banks to pharmaceutical companies to tech startups. However, big data is more than a way to make money or capture online readership. It is first and foremost information about people: their beliefs, experiences, environments, health, and behaviors—online and offline. In academic research, governments, and the non-profit sector, this type of data is studied by social scientists. What social scientists gain from big data has immense benefits for society as a whole.

However, big data is not easy to work with. It requires massive amounts of computing power to store, analyze, and share. Such resources have not yet made the leap into the social sciences. Social science researchers need access to their own vibrant, virtual computing platform to responsibly collect, process, and preserve big data.

The Virtual Institute for Social Research (VISR) provides a platform to address these challenges. This White Paper outlines how VISR accelerates how the social sciences community does science by giving researchers a venue for collaboration, access to supercomputing power, and a full suite of tools to take advantage of big data for society's benefits.

The Challenge

Big data is yielding great gains for businesses. For example, the retailer Target integrates vast amounts of information about its customers' preferences, demographics, and shopping behavior to inform its investments in marketing, store design, and product lines. In one particularly striking illustration of the power of this approach, a father complained to Target about the coupons for baby items the store had been sending to his high school-aged daughter, only to subsequently learn that Target knew his daughter was pregnant well before he did (Duhigg 2012). Such examples offer a glimpse at big data's enormous potential to illuminate previously-hidden trends, but also highlight the profound privacy and ethical issues associated with big data about people.

Big data fuels remarkable new discoveries for researchers and society more broadly. More and more, these revelations and innovations are springing from information about people. As opposed to past big data sets like those in astronomy or climatology, these new data sets include information of a profoundly personal, private nature. For example, medical researchers are no longer looking at just a person's genes, but at how genes are affected by or interact with the person's specific environment or health behaviors to produce health outcomes—just the sort of personal information that has long been the purview of the social scientist. The entry of social science researchers into the big data realm stands to bring tremendous benefits to society.

Social science researchers look for patterns as they collect and analyze data. These patterns are critical to tracking social movements or networks, which can help in understanding political trends, planning military

AT A GLANCE

- The big data revolution holds enormous, and largely unrealized, potential for the social sciences.
- VISR will provide social science researchers access to a powerful, secure cyberinfrastructure where they can store, analyze, and share big data.
- By providing a robust cradle-to-grave platform for social science data, VISR will promote the responsible use of data, enable new collaborations and research insights, and make the most of the nation's research investments.

campaigns or international development efforts, or solving a public health crisis, to name a few examples. Anthropology, education, geography, history, political science, sociology—the researchers in these fields and other social sciences are all separately collecting and operating on thousands of small data sets. Imagine the possibilities if they had access to a powerful collaborative environment that could link those thousands of data sets together, allowing individual scientists to find new data sets ready to be combined or compared for richer insights using modern, computationally intensive tools. However, the average social science researcher today simply does not have access to the massive level of computing power needed to access, process, and understand big data.

Social data is heterogeneous: it is being collected from a huge range of disparate sources such as social networks, online transactions, and field researchers. Social data is about people: their attitudes, purchases, and behaviors. Social data is everywhere: online and offline, global and local. Social data is expensive: learning from data requires investing heavily in computing capabilities. Most importantly, social data is power: big data yields big discoveries. Social scientists need to be able to leverage the power big data holds.

These challenges are too large for one discipline or one university to surmount. Social science researchers need a powerful, secure, streamlined network to unlock the potential of big data and collaborate on solutions to the world's challenges.

Ideas into Action

The Virtual Institute for Social Research (VISR) is the powerful computing network that the social science world has lacked. Initiated by researchers at the Odum Institute for Research in Social Science in partnership with RENCI, two institutes of the University of North Carolina at Chapel Hill, VISR is now in a pilot phase. When fully developed VISR will bring social science investigators together in a virtual collaborative environment with enough computing power to tackle the challenges of big data management, thus putting the full potential of the big data revolution at the fingertips of social scientists around the world.

VISR is not merely a data repository or a high performance computing cluster, although it will provide those services. It is a powerful, integrated technology platform for remotely accessing and working with data that is highly configurable across the data life cycle. It enables users to collect, store, process, share, and collaborate around big data. It meets the unique challenges involved in applying big data for social science research because it is a cyberinfrastructure that is not only powerful and streamlined, but capable of protecting privacy and providing robust data security. It is also cost-efficient and highly adaptable, open to continual improvement to meet the needs of its users.

POWERFUL DATA PROCESSING

VISR's servers offer much-needed computational capability for researchers running complex algorithms on large data sets (or thousands of small data sets). In many cases, its servers' calculations and response times will be far beyond what users currently have available at their home institutions. In addition, VISR will integrate with Jetstream (http://pti.iu.edu/jetstream/), a National Science Foundation-funded project that offers researchers cloud-based, high performance computing power. Jetstream, like VISR, was designed to bring supercomputing to a wider field of researchers, including the social sciences. Its integration with VISR will offer social scientists essentially unlimited computing power.

On its own, VISR can store the massive amounts of data needed to fuel big data social science. It has also been integrated with the Dataverse Project (Dataverse.org), which allows researchers to access secondary data, discover relationships between data sets, and find collaborators. By combining VISR with the Dataverse tools, social scientists will be able to publish, share, reference, extract, and analyze large amounts of data like never before.

ROBUST PRIVACY AND SECURITY PROTECTIONS

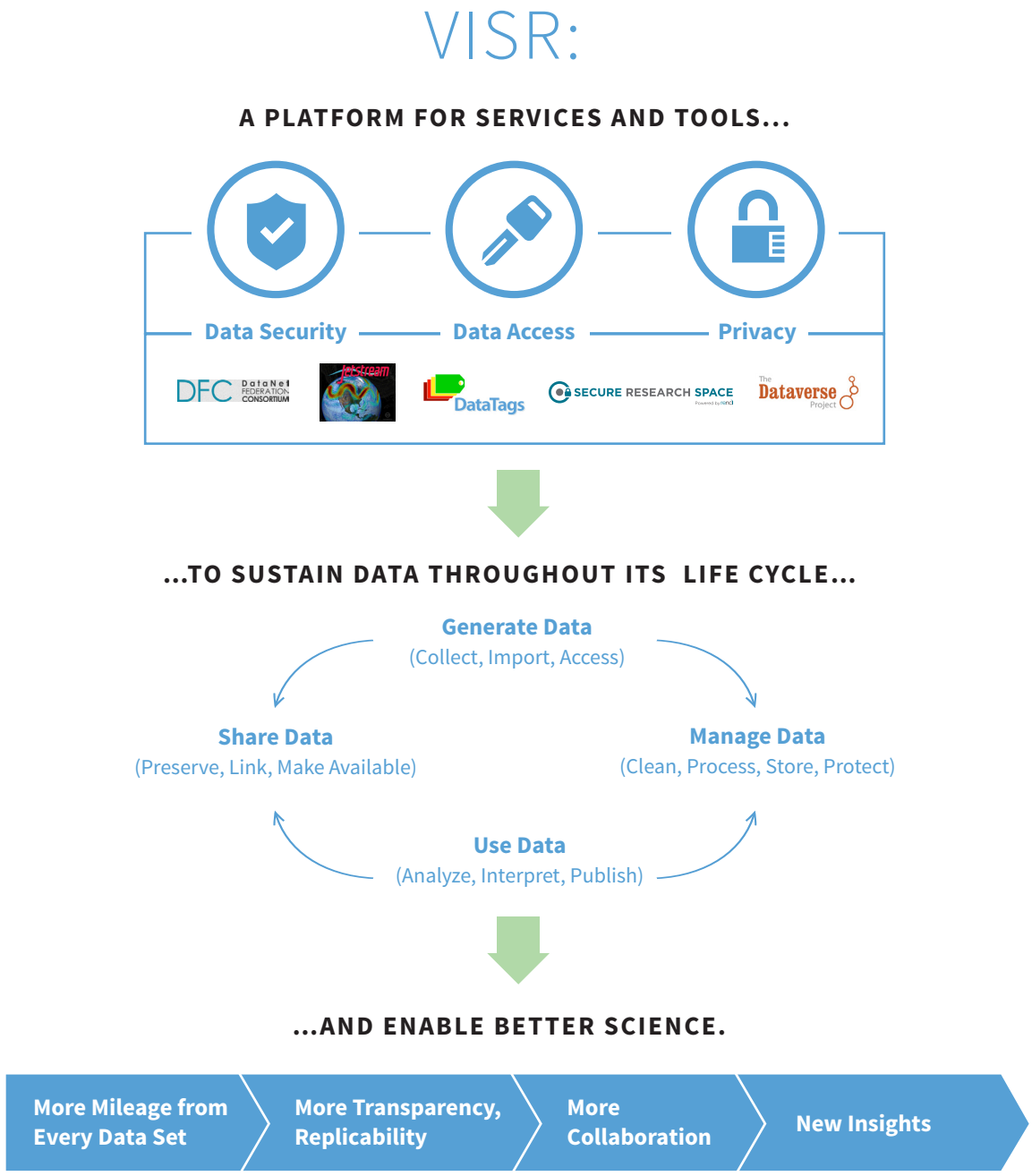
The data sets social scientists work with are often highly sensitive. By integrating several existing tools, VISR is being developed as a secure network with the privacy controls needed to ensure the protection of sensitive data. First, VISR takes advantage of iRODS (irods.org), a proven data management technology whose user-defined rules protect big data while allowing for large-scale computations. This open source software has an active consortium of users around the world, securing its ongoing sustainability (Hammond 2009). In addition, proper data set handling for sensitive data requires knowledge of complex federal and state laws. To address this challenge, DataTags (iq.harvard.edu) helps researchers comply with these restrictions through a simple user-input questionnaire, which is then processed to permanently attach the relevant security and privacy tags on the data. Finally, the Secure Research Space (renci.org), a platform first developed to protect sensitive medical data in a virtual environment, is also available for VISR users to enhance data security.

CRADLE-TO-GRAVE DATA MANAGEMENT

There is a growing demand for data transparency across its whole life cycle, from collection to publication. When fully deployed, VISR's full suite of tools will streamline the cumbersome process of working with data sets of any size. Researchers can use VISR from the start of their investigations, or import existing data. Using VISR's data collection tools will solve many problems at once: data storage, data cleaning, and access for others. Data collected in VISR from the start will be easy to access, work with, share, and eventually publish.

VISR also solves the problem of data preservation by acting as a stable storage site that provides permanent citations, making research forever harvestable and open to replication.

A natural partner for VISR in the management of data across its life cycle is the DataNet Federation Consortium (DFC, datafed.org). This consortium's NSF-funded mission is to implement a complete data management infrastructure whose parameters will keep massive data collections stable and accessible, despite ever-evolving technology. Because the Odum Institute is a partner in both VISR and the DataNet Federation Consortium, VISR will be able to take advantage of DFC's extensive, essential infrastructure immediately.



COLLABORATION AND EDUCATION

One great advantage that underlies all of VISR's capabilities—power, security, ease-of-use—is its tremendous potential to foster collaboration. Most social scientists today work independently or in small groups, constrained by small data sets and limited resources. VISR will enable researchers to work together around huge sets of data, or tens of thousands of small sets that span a variety of disciplines. Sharing data and making new connections will have enormous benefits, leading to new discoveries within traditional fields of study while building new connections across them (Gutmann & Friedlander 2011). With VISR, researchers can work together to actively explore available data, generate new data, and collaborate throughout the data life cycle, all while accurately preserving the record of science for future investigators.

VISR will also include a valuable training and education component beyond simply teaching its users how to take advantage of VISR tools. Training on the use of big data will help investigators test new computational and analytical approaches and uncover new insights. Training focused on data collaboration will help investigators find and cultivate new connections with other researchers or data sets. Finally, VISR's educational tools will tackle broader data problems such as the critical ethical implications of big data collection and use.

EFFICIENCY AND ADAPTABILITY

In addition to being a powerful, secure, and streamlined site for collaboration, VISR is being built as a cost-efficient, sustainable system that can continue to serve its users far into the future without overly burdensome ongoing expenses. Many of the components of VISR already exist and are integrated into VISR rather than being

BLAZING THE TRAIL

VISR was inspired by a number of successful initiatives in other fields that have gained dedicated, enthusiastic users. These thriving, dynamic networks have successfully addressed researchers' need to easily and securely collaborate around big data in a variety of contexts:

- **The iPlant Collaborative** (iplantcollaborative.org) gives researchers in the life sciences access to high-powered computing, easy data sharing, and seamless collaborations. Its global user base is continually growing, and, at the recommendation of the NSF, iPlant has branched out into iAnimal and iMicrobe.
- **The National Database for Autism Research (NDAR)** (ndar.nih.gov) is a secure, powerful infrastructure serving a robust community of researchers studying autism. Using cloud computing resources, specialized harmonization tools, and thorough, guided tutorials, users collaborate around data sets, methodologies, and queries from anywhere in the world.
- **The INCF Dataspace** (incf.org) is an international network where neuroscience researchers can share, analyze, and collaborate around big data in order to accelerate innovation in brain research.

built from scratch. Piggy-backing on these existing resources makes VISR efficient, effective, and adaptable. Whatever its users want to do with data, VISR and its many integrated tools can do it.

Once it is fully deployed, VISR can be the first and last place for data collection, storage, management, analysis, publication, preservation, and sharing in the social sciences. Its capabilities will enable more researchers to do science the way that science is meant to be done—preserving a data record and encouraging replication. Simply put, VISR will help social scientists to do science the right way.

The Upshot

VISR is currently in the prototype stage. Because its underlying infrastructure and tools already exist, a full deployment is feasible within the next two years and there is a high level of interest in VISR from other research institutions. Outreach among researchers, industry, and government agencies will also increase VISR's visibility and use in the global social science community. The realm of social science is expanding as researchers integrate big data into their studies; collaborations with experts in fields such as computer science, statistics, and health science are increasingly needed to foster innovation and make discoveries about individuals and society.

The training components planned for the future mean that VISR can also play a leading role in the education that working with big data requires. The complexities of big data are still new, which makes training for all users incredibly important (Ahalt & Kelly 2013). Committed, engaged researchers who come to VISR will find a powerful resource that allows them to share, manipulate, cite, extract, publish, and ultimately do amazing things with big data.

The Big Picture

The big data about people being captured today is all about extremes. It is extremely sensitive, it is extremely private, and it is extremely large. Only a powerful, secure, and easy-to-use platform can accommodate the needs of social science researchers seeking to unlock the potential of the big data revolution. VISR is that platform.

VISR gives researchers the tools to do science the right way, right from the beginning. Its computational capabilities are far beyond what most social science researchers currently have access to. The ability to crowdsource data sets from around the world, and from all the disciplines of social science, will lead to discoveries and innovations that would not otherwise have seen the light of day.

The big data revolution is revolutionary precisely because it is now based around data about people, as opposed to natural phenomena. Social scientists must be able to effectively collect, calculate, and collaborate around big data. What they learn will become what we, as a society, learn.

REFERENCES

Ahalt, S. and Kelly, K. (2013). The Big Data Talent Gap. UNC Executive Development. Retrieved March 10, 2015, from http://renci.org/wp-content/uploads/2013/08/The-Big-Data-Talent-Gap-White-Paper1.pdf

DataNet Federation Consortium. Retrieved March 18, 2015, from http://datafed.org/

DataTags. The Institute for Quantitative Social Science, Harvard University. Retrieved March 10, 2015, from http://datascience.iq.harvard.edu/about-datatags

Dataverse.org. Retrieved March 10, 2015, from http://dataverse.org/

Duhigg, C. (2012). The power of habit: Why we do what we do in life and business. New York: Random House.

Gutmann, M. and Friedlander, A. (2011). Rebuilding the mosaic: fostering research in the social, behavioral, and economic sciences at the National Science Foundation in the next decade. Arlington, VA: National Science Foundation. Retrieved March 10, 2015, from http://www.nsf.gov/pubs/2011/nsf11086/nsf11086.pdf

Hammond, J.S. (2009). Best practices: improve development effectiveness through strategic adoption of open source. Cambridge, MA: Forrester Research, Inc. Retrieved March 10, 2015, from http://download.microsoft.com/download/7/1/B/71BEC711-FBB4-400A-984C-0DCCC36248E0/BP-OpenSource.pdf

INCF Dataspace. INCF. Retrieved March 10, 2015, from http://www.incf.org/resources/data-space

iPlant Collaborative. Retrieved March 10, 2015, from http://www.iplantcollaborative.org/

iRODS. Retrieved March 10, 2015, from http://irods.org/

Jetstream. Retrieved March 18, 2015, from http://pti.iu.edu/jetstream/

National Database for Autism Research. National Institutes of Health. Retrieved March 10, 2015, from https://ndar.nih.gov/

Secure Medical Workspace. RENCI. Retrieved March 10, 2015, from http://renci.org/research/secure-medical-workspace/

ABOUT RENCI:

RENCI, an institute of the University of North Carolina at Chapel Hill, develops and deploys advanced technologies to enable research, innovation, and economic development. For more information, see www.renci.org

HOW TO REFERENCE THIS PAPER:

Carsey, T., Crabtree, J., & Schmitt, C. (2015): The Virtual Institute for Social Research: Bringing Big Data to the Social Sciences. RENCI, University of North Carolina at Chapel Hill. Text. http://dx.doi.org/10.7921/G0W66HP5

renci

THE UNIVERSITY of NORTH CAROLINA at CHAPEL HILL

UNC THE ODUM INSTITUTE