Social Genome: Putting Big Data to Work for Population Informatics

Hye-Chung Kum, *Texas A&M Health Science Center*
Ashok Krishnamurthy, *University of North Carolina at Chapel Hill*
Ashwin Machanavajjhala, *Duke University*
Stanley C. Ahalt, *University of North Carolina at Chapel Hill*

Data-intensive research using distributed, federated, person-level datasets in near real time has the potential to transform social, behavioral, economic, and health sciences—but issues around privacy, confidentiality, access, and data integration have slowed progress in this area. When technology is properly used to manage both privacy concerns and uncertainty, big data will help move the growing field of population informatics forward.

Nearly all of our activities from birth until death leave digital traces. Health records, wages earned, schools attended—these and countless other data capturing the details of our daily lives serve as our digital social footprint. Collectively, these digital traces—across a group, town, county, state, or nation—form a population's *social genome*, the footprints of our society in general. If properly integrated, analyzed, and interpreted, social genome data could offer crucial insights into how best to serve our greatest societal priorities: healthcare, economics, education, and employment.

Social scientists have long drawn on data collections from governments and elsewhere to track demographic trends, inflation, employment rates, and so on. Now, however, our daily activities leave digital crumbs all over cyberspace, and we have the technology to gather and analyze these crumbs to reveal previously hidden trends. This newfound ability to examine deep analysis-rich questions in near real time using distributed datasets that are large, complex, and diverse has the potential to transform social, behavioral, economic, and health sciences. Population informatics is the burgeoning field at the intersection of social sciences, health sciences, computer science, and statistics that applies quantitative methods and computational tools to answer questions about human populations. Just as bioinformatics has revolutionized biological research, population informatics could catalyze significant advances in our understanding of trends in society, health, and human behavior.

The use of big data has spurred major advances in many areas, from climatology to bioinformatics to business analytics. Unfortunately, social and health sciences are much more complex, relying on person-level information across a population. So far, challenges associated with maintaining privacy and confidentiality, access, data integration, and data management have constrained the use of micro data—person-level data—in these areas of research, leaving rich databases largely untapped.

But improving our capacity to analyze big data collections that involve person-level information is not just interesting science: the results could lead to more informed and effective policy decisions and management of social programs. Social genome data can tell us about how people live, work, respond to change, and make decisions, as well

Published by the IEEE Computer Society

as the collective impact of these individual decisions. Such insights help us understand the root causes of social and public health problems, predict the downstream effects of different policy options, and allocate our collective resources for the greatest impact.

DEFINING SOCIAL GENOME DATA

Information about individual people is critical to understanding society in the same way that the physical genome is critical to understanding an organism. In many ways, the information held in social genome data represent the social being, whereas our physical genome data represent our physical being. A child is born not only with a certain genome sequence but also into a certain social environment—parents, siblings, town, economic status—that influences the life path the child will take. Data on these social environments are just as important in understanding the overall well-being of a person as his or her physical genome. Being able to study the social genome at scale will enable data-driven understanding of important sociological questions such as the long-term effect of the social genome at birth.

To decipher patterns about the ways societies behave and evolve, social scientists must examine how individuals live and interact. The social genome thus represents a core set of data that information scientists can use to explore connections, build theories, and propel breakthroughs in managing a society. But just as with the physical genome, the social genome does not provide the full story; it also contains some useless and erroneous information, so the problem of extracting insights from these data is very challenging.

The field of bioinformatics—now virtually inextricable from the practice of biology as a whole—was catalyzed largely by a single endeavor: the Human Genome Project. Although bioinformatics now includes a plethora of methods and tools beyond genome sequencing, the Human Genome Project provided the focus and structure needed to develop key bioinformatics tools and principles. We need a similar Social Genome Project to catalyze population informatics. The solution we envision and describe here includes a series of region-based social genome projects that could serve as a springboard for developing the tools, analytical methods, and oversight mechanisms needed to transform population informatics to the next level.

CASE STUDIES

Big data is being harnessed for powerful new person-level applications in many areas already. Health informatics analyzes electronic records to improve healthcare delivery and health outcomes for a population. Education informatics relies on school records for education research and delivery. Transit informatics uses real-time GIS data to facilitate public transportation. Business analytics turns operations data into meaningful information for key business functions, such as marketing and client profiling.

Although the data are distinct in each of these fields, the common theme is the application of informatics to process, manage, and analyze individual-level data for group-level insight. Population informatics—accessing existing collections of raw data for secondary purposes—helps drive a deeper understanding of the social genome. However, the key factors that make population informatics difficult are that the data capture many features about a large number of individuals (volume), the data are continuously updated to reflect changes (velocity), the data exist in heterogeneous systems that are redundant yet inconsistent (variety), and the data are incomplete and erroneous (veracity). These four Vs of big data are common to all data-intensive science. Nevertheless, the advantages of being able to integrate, analyze, and interpret massive person-level data collections are clear, as illustrated in the following examples.

Population informatics is the burgeoning field at the intersection of social sciences, health sciences, computer science, and statistics that applies quantitative methods and computational tools to answer questions about human populations.

Economics application

One example of the power of data integration is the Longitudinal Employment Household Dynamics program at the US Census Bureau. LEHD integrates data from censuses, surveys, and administrative records from national and state-based databases across all 50 states to generate information about labor markets. The project enabled economists to use real-world US data to test their models of unemployment dynamics and model-churning behavior, earning some of them a Nobel Prize.

Worker churning, a ubiquitous feature of the US labor market, refers to companies hiring and firing at the same time. Research on worker churning requires the more detailed person-level data available only in the administrative data, and not just the net job loss or creation per company that was traditionally available before this project launched. LEHD vertically integrates data from across multiple geographic areas in one domain—labor. A project that horizontally integrates data across multiple domains—labor, education, and health, for example—would be even more powerful, even if it were confined to one geographic area.

Health application

Another successful integration of different facets of the social genome to gain population insights in near real time is the Google Flu project (www.google.org/flutrends). By combining information about physician visits with individuals' search queries, Google, in collaboration with the US Centers for Disease Control, was able to predict incidences of flu in a more timely and accurate manner than what the CDC could do with just the physician visit information.

A SOCIAL GENOME PROJECT

Researchers are already using person-level data to study population trends. However, many countries lack a national framework for secondary use of such data, leaving each project to develop its own privacy protection and thereby leaving potential vulnerabilities that can degrade public trust in such work. Developing the technology and policies for a virtual "hot cell"—analogous to the shielded rooms used for working with radioactive material—to provide a safe environment for conducting sensitive person-level data research is of critical importance.

The social genome thus represents a core set of data that information scientists can use to explore connections, build theories, and propel breakthroughs in managing a society.

In our vision, each social genome project would establish a regional data gateway—a social genome center—for data relevant to population-level studies in a certain region, such as a state. Generally speaking, these gateways would provide a common portal to multiple databases such as birth, tax, or criminal records where data could be safeguarded while research is conducted. This would be a virtual repository in that data would still be housed physically where most appropriate. Access to data would still be controlled by different data custodians, but the center would facilitate and streamline the process of obtaining access.

The center would not be responsible for integrating or cleaning the heterogeneous data for a particular study. Rather, it would provide the tools that researchers need to clean and integrate the data to meet their requirements by building a hybrid human-machine system that researchers can easily plug into. This would allow each study to optimize the utility of the data to particular research questions. From the users' perspective, these social genome centers would function much like a public library or the federal government's database collection (http://data.gov). But on the back end, the centers would add value to the databases they can access by making the full process—from raw data to the summarized statistics—transparent and available to authorized users in appropriately protected environments as needed. Each center would also be responsible for developing systems, both technical and governance, to protect personal privacy and confidentiality, overseeing processes for providing access to data and developing the software required for such research.

Our envisioned informatics architecture would begin with the data ingestion layer, responsible for getting data into the system securely and creating the loose connections to other data in the system for later use. Loose connections tolerate inconsistencies across datasets that can only be resolved based on the application. The second layer—data access and analysis—would be responsible for giving diverse system users privacy-preserving levels of access and views of the data appropriate for the tasks required to safely turn the social genome data into useful information. Finally, the topmost layer—information delivery—would provide a library of customizable visualization tools that content experts could use to deliver relevant, evidence-based information that could then be included as part of their results and conclusions (for example, real-time reports, graphs, and summary data tables). This layer could easily be integrated with other efforts to make government data more accessible, such as http://data.gov.

THE CHALLENGE

To put big data to work for population informatics, we must overcome some unique challenges. Investing in infrastructures to propel population informatics forward in a coordinated, responsible way can help us unleash the power of big data for the nation's collective benefit. Many more topics, such as secure data ingestion, auditing, and data version controls, are not covered here due to space constraints.

Building a knowledge base platform with uncertainty management

Whereas a company like Amazon owns much of its customer data and can centrally manage and analyze that information, the data sources of greatest value to population informatics research are managed by disparate bodies including hundreds of departments within local, state, and federal governments—birth and death records, Medicare and Medicaid rolls, school enrollment rosters, and criminal records, to name a few. Each agency has its own approach to collecting, labeling, managing, and providing access to data, making it challenging for researchers—even those within government bodies—to integrate data for in-depth analysis. Thus the social genome platform must provide tools to ingest and manage diverse kinds of data, including structured repositories such as medical records, summary statistics such as socioeconomic indicators from the US Census, and real-time unstructured data such as medical notes or tweets. Because social genome centers would primarily deal with individual-level data, the completion of the following two tasks is essential: disambiguating individuals, such as identifying that Bob in education records, Robert in the Social Security dataset, and bob1234 on Twitter, are the same person; and enriching individual information from multiple datasets, such as knowing that Bob volunteers at a retirement home based on his tweets. This process of continuously ingesting, disambiguating, and enriching entities from disparate sources of information is referred to as knowledge base synthesis.[1]

An important difference between knowledge base synthesis and typical ETL (extract-transform-load) tools available for data warehousing is the fact that it must integrate information from multiple domains and maintain multiple versions of the data to satisfy the disparate information needs of multiple users. For instance, a social genome center might host the information from three different state agencies (or domains): education, child welfare, and health. Each agency (or user) is interested in maintaining and curating its own data, as well as in enriching datasets with information from other databases, requiring the disambiguation of people and entities such as hospitals and schools. These agencies might have different uses for the data—the education department might only release statistical summaries, while the health department might want to utilize child-level data for medical interventions (and thus have a lower tolerance for errors in disambiguation and enrichment). Finally, the social genome center has to support multiple versions representing different time points to understand temporal trends.

As the number of domains, users, and versions increases, the complexity of management, disambiguation, and enrichment of individual information only increases. Moreover, users of the social genome data must be able to bring together (disambiguate, consolidate, and analyze) chaotic data into a view that can address particular research questions on the fly. To support multiple data uses, the system cannot create a single, consolidated, and clean view of the data, but rather a framework and tools so that users can manipulate their own views at will and easily. This is fundamentally different from the goals of the typical ETL process, which maintains one clean collection of data. A better approach is to abstract out domain-independent algorithms into a platform layer and to expose a set of plug-ins that different users can customize for knowledge base synthesis.[1] It is important that such plug-ins can support efficient human decision making by quickly pointing out inconsistencies that need to be resolved in the federated data, along with interactive visualization that supports multiple levels of details.

Secure data access

There is a direct relationship between data usability and risk to privacy; greater access to data generally leads to a higher privacy risk but more usability of the data, and more restricted access generally provides better privacy protection at the cost of less usability. The key is to understand data use requirements to design a flexible paradigm that balances the two competing requirements for usability and protection given particular needs. This is sometimes called the privacy-by-design approach to privacy protection. Privacy-by-design looks beyond the narrow view of privacy as anonymity and tailors privacy principles and data protection to the full system, thereby building a safe environment consisting of secure computer systems and policy frameworks, in which data can be analyzed safely. The fundamental design principles for privacy and usability are the minimum-necessary standard—which states that maximum privacy protection is provided when the minimum information needed for the task is accessed at any given time—and the maximum-usability principle—which states that data are most usable when access to the data is least restrictive—in other words, direct remote access is most usable. If we apply these design principles into a secure laboratory for population informatics, the three components of that laboratory must be a well-designed secure computer system, secure software and data to carry out the research in a privacy-preserving manner, and a governance framework.[2]

Privacy-by-design looks beyond the narrow view of privacy as anonymity and tailors privacy principles and data protection to the full system, thereby building a safe environment consisting of secure computer systems and policy frameworks, in which data can be analyzed safely.

Broadly speaking, the purpose of population informatics is to transform raw administrative data beyond operations into insights that can inform decision making. Table 1 details the computer system, software, and data for four data-access levels—restricted, controlled, monitored, and open—designed around the workflow from raw data to decision based on the four most common activities.[2] These access levels offer optimum privacy protection while still providing maximum usability for the given data and activity, and they help define a comprehensive system for privacy protection for most secondary data analysis in population informatics research.

Table 1. Comparison of risk and usability.*

	Restricted access	Controlled access	Monitored access	Open access
Example systems	RDC (Research Data Center)	Secure medical workspace	Secure Unix servers	Public website
Type of data	Decoupled micro data (high-risk data)	De-identified micro data (medium-risk data)	Aggregate data (low-risk data)	Sanitized data (minimal-risk data)
Privacy-protection methods used	Encryption for decoupling, locked down computer with physical restriction	Locked down virtual machine (VM) to restrict software on the computer and data channels	Information accountability	Disclosure limitation methods
Oversight protocol	Based on the risk and benefit of the research, approval is required	Based on the risk and benefit of the research, approval is required	Must file what and how data are being used, including for what purpose, in advance, but does not require approval; will still support information accountability when breach is suspected	Honor system; no registration or details of use required, but user signs a general agreement with guidelines for appropriate use
Monitor use	On and off the computer	On the computer	On the computer	No monitoring
Usability I U1.1: Software (SW)	Only preinstalled data integration and tabulation software; no query capacity	Requested and approved statistical software only	Any software	Any software
Usability II U1.2: Other data	No outside data allowed	Only preapproved outside data allowed	Any data	Any data
Usability III U2: Access	No remote access	Remote access	Remote access	Remote access
Risk I R1:Cryptographic attack	Highly difficult	Fairly difficult; would have to break into virtual machine	Easy to run sophisticated software with outside data	NA
Risk II R2: Data leakage	Very difficult; memorize data and take out	Physical data leakage (take a picture of monitor)	Electronically take data off the system	NA

*Shaded boxes represent no restriction in a particular dimension; they depict how access levels are fully opened up one at a time from restricted access to open access (allowing for usability 1.1 and 1.2 results in risk 1, and allowing for usability 2 results in risk 2).

Privacy-preserving data integration

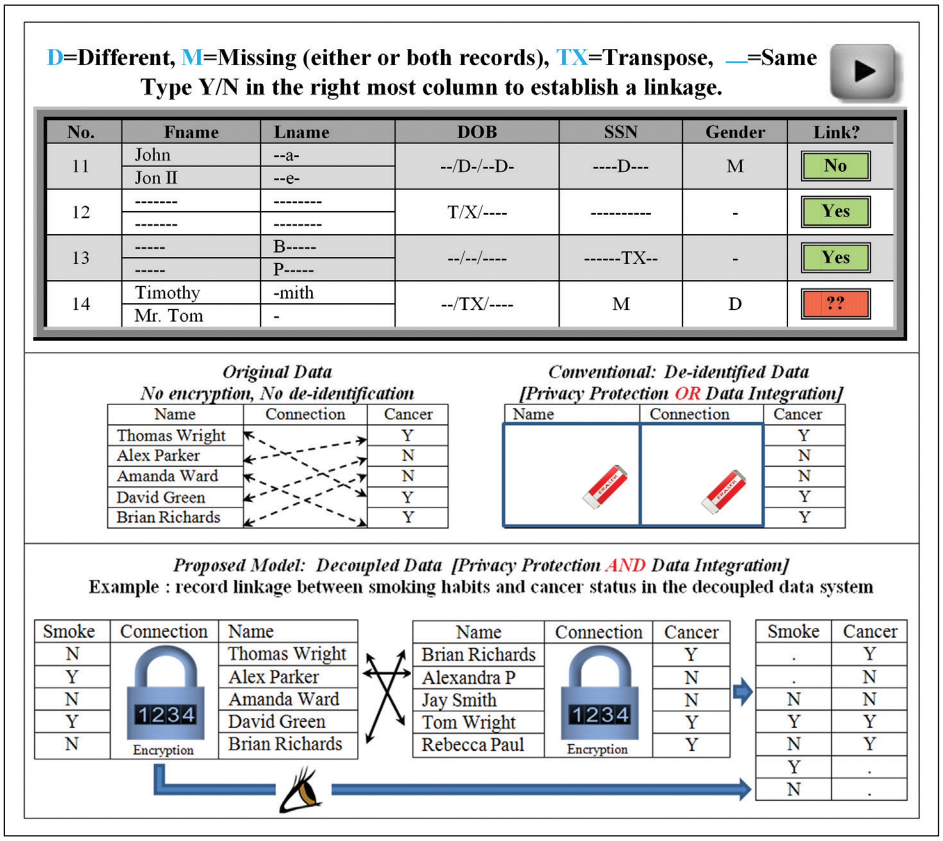
Integrating data from heterogeneous and uncoordinated systems requires record linkage—the critical task of identifying record pairs that belong to the same real-world entity. But considerations of privacy make this a difficult issue for population informatics. Privacy-preserving record linkage is fundamentally different from most privacy-preserving data operations in that the goals of record linkage are precisely to identify the entity represented by the data so that the linkage can be made accurately. For example, it is very important to distinguish between two twins in a dataset so that the two records are not treated as a duplicate record for one person and that the records are not cross-linked. Incorrect identification has the potential to harm the subjects and can also result in serious legal and clinical consequences.

It is critical to understand the distinction between identity disclosure (who is this person?) and sensitive attribute disclosure (does this person have cancer?). Identity disclosure has little potential for harm on its own, but sensitive attribute disclosure is another matter.[3,4] If we define the privacy goal of privacy-preserving record linkage as a guarantee against attribute disclosure, we can develop systems that allow both privacy protection and high-quality record linkage.[4]

Private record linkage computes the set of linked records given a mapping function and outputs the linked records to the two private parties without revealing anything about the nonlinked records. The goal of private record linkage is to securely compute a known mapping function. The first generation of private record linkage methods was made up of hash-based algorithms, which provided strong privacy guarantees but were limited to exact matching. The second generation of methods was built on secure approximate string comparison operations, such as Bloom filters, to support approximate record linkage. Major challenges here are that, in reality, the mapping function is typically not known, and there is a requirement to manually refine the ambiguous links for high-quality data integration.[5]

In practice, trusted third parties with access to all the data perform data linkage and integration. In the US, federal and state health statistics departments and selected research entities are the trusted parties. Several countries operate a data linkage center to support population research. In these centers, the most important protocol for privacy protection is the separation of identifying data and sensitive data to protect against attribute disclosure.[6,7]

High-quality data integration requires human involvement to manage the errors inevitably introduced by imperfect real data. Errors that are not properly managed propagate to subsequent data analyses, leading to incorrect analyses and decisions.[8]



D=Different, M=Missing (either or both records), TX=Transpose, __=Same
Type Y/N in the right most column to establish a linkage.

No.	Fname	Lname	DOB	SSN	Gender	Link?
11	John Jon II	--a- --e-	--/D-/--D-	----D---	M	No
12	------- -------	-------- --------	T/X/----	----------	-	Yes
13	----- -----	B----- P-----	--/--/----	------TX--	-	Yes
14	Timothy Mr. Tom	-mith -	--/TX/----	M	D	??

Original Data
No encryption, No de-identification

Name	Connection	Cancer
Thomas Wright		Y
Alex Parker		N
Amanda Ward		N
David Green		Y
Brian Richards		Y

Conventional: De-identified Data
[Privacy Protection OR Data Integration]

Name	Connection	Cancer
		Y
		N
		N
		Y
		Y

Proposed Model: Decoupled Data [Privacy Protection AND Data Integration]
Example : record linkage between smoking habits and cancer status in the decoupled data system

Smoke	Connection	Name
N		Thomas Wright
Y		Alex Parker
N		Amanda Ward
Y		David Green
N		Brian Richards

Name	Connection	Cancer
Brian Richards		Y
Alexandra P		N
Jay Smith		N
Tom Wright		Y
Rebecca Paul		Y

Smoke	Cancer
.	Y
.	N
N	N
Y	Y
N	Y
Y	.
N	.

Figure 1. Secure Decoupled Linkage (SDLink). The SDLink GUI (top) applies data-recoding techniques that display the difference between the attributes that are meaningful for record linkage instead of the actual data. For example, the gender field only indicates same(_), different (D), or missing (M) in one or both fields. Internally, the data are stored in a decoupled data system (bottom), which separates out the identifying attributes from the sensitive attributes and introduces fake data (chaffing). Decoupled data have the same level of privacy protection as deidentified data (mid right), but are much more powerful because researchers can link multiple decoupled datasets safely. Decoupled data, along with chaffing, allow for accurate record linkage with no attribute disclosure.

Recently, researchers have proposed Secure Decoupled Linkage (SDLink) for privacy-preserving interactive record linkage. SDLink is a computerized third-party linkage system that offers safe and high-quality data integration by using a hybrid human-machine system[4] based on three core privacy principles.[4] First, as shown in Figure 1, SDLink decouples the identifying data from the sensitive data via encryption. Second, through chaffing (adding fake data) and universe manipulation (changing the dataset label), SDLink prevents the attribute inference that can occur in group disclosure. For example, if someone you know is on the cancer registry (group disclosure), she must have cancer (attribute disclosure), but this disclosure can be prevented if you know that the list has fake data—that people who do not have cancer are also on the list—or if you did not know this is a cancer registry. Any identity disclosure is additionally minimized by recoding the variables in the GUI (see the top of Figure 1). Only the information that is essential for record linkage is revealed during the linkage process. More research is needed to understand the useful and meaningful differences of the different variable types as well as what people infer from information displayed for linkage. The key is to understand the minimum information required for acceptable linkage and then to design protocols to securely reveal only that information.

Privacy-preserving data analysis

Although several privacy and security challenges arise from unauthorized access or malicious dissemination of data, the results of valid data analyses can also lead to the disclosure of sensitive information about individuals, and thus a confidentiality breach. There is a fine line between an adversary's ability to infer sensitive attributes of an individual and a researcher's ability to learn trends in the population. Hence, mathematically formulating what it means for some data analysis to not breach the privacy of individuals is a challenging task. Understanding these risks well is especially important for data released as open access or as monitored access in the four-level model discussed earlier.

Another challenge in private data analyses is that even if one result does not disclose sensitive information about any individual, a collection of these tasks could potentially lead to a breach. For instance, consider two queries: the number of unemployed males in Durham, North Carolina, and the number of males in Durham other than Bob who are unemployed. While Bob's employment status is not disclosed by either query in isolation, it can be inferred by combining the answers to both queries. Recent work has shown that many supposedly safe methods of releasing data can lead to disclosure of individual information by combining multiple invocations of these algorithms.[9,10]

In fact, a classic result shows that you cannot answer more than an adversarially chosen set of $n(\log n)^2$ queries over a database of n bits such that each query has $o(\sqrt{n})$ errors without the adversary being able to reconstruct the original database. This result poses a fundamental limit on private data analyses and motivates the need to think about private data analysis as a budget-constrained problem. Each query leads to some privacy loss while providing some utility in terms of data analysis. The goal is to achieve the maximum utility under a fixed privacy budget.[9]

As a society, we will have to collectively contemplate the expected norms of ethical personal data use for the greatest benefit.

Differential privacy is a methodology that lets us concretely reason about privacy-budgeted data analysis. An algorithm satisfies differential privacy if, for any two datasets D1 and D2 that differ in one row, the ratio of the likelihood of the algorithm resulting in the same output starting from D1 and D2 is bounded by at most e^{ε}. Thus, if each row in a database corresponds to an individual, then using a differentially private algorithm provably ensures that the output is not sensitive to an arbitrary change in any one individual's input.[10] Differential privacy is powerful because it can be composed—two algorithms that satisfy differential privacy with parameters ε1 and ε2 results in (ε1 + ε2) differential privacy, thus allowing us to apportion a total privacy budget of ε across many subtasks. Differential privacy can allow accurate analyses in certain cases. For instance, one of the LEHD data products boasts of provable differential private protection in the released data (http://onthemap.ces.census.gov).

There has been much theoretical examination of differential privacy, but how to apply this framework to practical individual data is an area of active research, including understanding optimal methods to apportion privacy budgets to sets of overlapping data analyses, minimizing the noise introduced by differentially private methods in sparse data, and customizing and relaxing differential privacy in applications involving correlations, sparse data, and time-varying data.

THE BROADER PROBLEM: PRIVACY, CONFIDENTIALITY, AND ETHICS

In the computer science literature, privacy refers broadly to collection, maintenance, disclosure, and control of, and access to, information about individuals.[11] It is helpful to note that in many other fields privacy refers more narrowly to safe data collection (data input), whereas confidentiality refers to safe information disclosure (data output).[3] Kenneth Prewitt, former director of the US Census Bureau, states that, privacy is akin to "don't ask" and confidentiality is akin to "don't tell." Some security technologies are applicable to both, and others are specific to only one purpose.

Accidental or purposeful misuse of social genome data has the potential to cause harm to individuals. In addition, privacy and confidentiality breaches can lead to legal consequences, especially in government and research settings. Thus, privacy and confidentiality protection is critical to the success of population informatics research. Protecting privacy and confidentiality in secondary data analysis is complex and requires a holistic approach involving technology, statistics, governance, and a shift in culture of information accountability through transparency rather than secrecy. Information accountability focuses on monitoring use of sensitive data to hold users of that data accountable for any misuse.[12] For example, protection of financial credit history data is mainly based on information accountability, where all parties know who used what information for what purposes with strict laws to hold them all accountable.

Governance models also play an important role in maximizing protection. Helen Nissenbaum provides a practical legal framework for privacy protection of personal information referred to as contextual integrity—that is, privacy protection depends on the context and the expected norms of protection given a particular situation.[13] From a technical standpoint, these privacy standards result in policy requirements on digital data about who has access to which data, for what purpose, and how the data should be maintained. The most relevant question for population informatics research is, "What are the expected norms of ethical conduct for doing research with person-level data in a given society?" Each country must start a discourse on the ethics of data analysis that draws on personal data.

Given proper oversight mechanisms, people might willingly donate data if it means more appropriate allocation of tax dollars and a greater impact for government programs—just as they willingly share blood samples for research that has the potential to save lives. The challenge lies in establishing the proper oversight mechanisms. As a society, we will have to collectively contemplate the expected norms of ethical personal data use for the greatest benefit. Then, we can apply various privacy, confidentiality, and security technologies to hold researchers accountable and ensure that all research using social genome data is conducted within the legal and ethical boundaries.

Big data holds tremendous, yet untapped, potential for informing evidence-based decision making at many levels. To unleash this potential, significant investments in infrastructure are worthwhile to give researchers the ability to develop the necessary tools for integrating, managing, and using social and health data with proper oversight. A series of regional social genome projects would provide unprecedented access to integrated, high-quality, robust datasets and offer the strategic focus and development space needed for population informatics to mature in a responsible, coordinated manner. To provide initial investments and ensure long-term maintenance, we envision regional social genome initiatives as regional-national-academic consortia in which each participant contributes data, funding, and resources and reaps downstream benefits. Social genome initiatives provide tremendous opportunity for both research and public programs, and investments in them will provide benefits for decades to come.

Acknowledgments

We thank Mary Whitton, Michael K. Reiter, Ronald Rindfuss, and Dean F. Duncan for their valuable comments in developing our ideas and article.

References

1. K. Bellare et al., "WOO: A Scalable and Multi-Tenant Platform for Continuous Knowledge Base Synthesis," *VLDB Endowment*, vol. 6, no. 11, 2013, pp. 1114–1125.
2. H.-C. Kum and S. Ahalt, "Privacy by Design: Understanding Data Access Models for Secondary Data," *AMIA Summits on Translational Science Proc.*, vol. 2013, 2013, pp. 126-130.
3. S. Fienberg, "Confidentiality, Privacy and Disclosure Limitation," *Encyclopedia of Social Measurement*, Academic Press, 2005, pp. 463–469.
4. H.-C. Kum et al., "Privacy Preserving Interactive Record Linkage," to appear in *J. Am. Informatics Assoc.*, 2014, doi: 10.1136/amiajn/-2013-00216/.
5. D. Vatsalan, P. Christen, and V.S. Verykios, "A Taxonomy of Privacy-Preserving Record Linkage Techniques," *Information Systems*, vol. 38, no. 6, 2013, pp. 946–969.
6. C.J. Bradley et al., "Health Services Research and Data Linkages: Issues, Methods, and Directions for the Future," *Health Services Research*, 16 Oct. 2010, pp. 1468–1488.
7. D. Holman et al., "A Decade of Data Linkage in Western Australia: Strategic Design, Applications and Benefits of the WA Data Linkage System," *Australian Health Rev.*, vol. 32, 2008, pp. 766–777.
8. P. Lahiri and M.D. Larsen, "Regression Analysis with Linked Data," *J. Am. Statistical Assoc.*, vol. 100, no. 469, 2005, pp. 222–230.
9. I. Dinur and K. Nissim, "Revealing Information while Preserving Privacy," *Proc. 22nd ACM SIGMOD-SIGACT-SIGART Symp. Principles of Database Systems* (PODS 03), ACM, 2003, pp. 202–210.
10. C. Dwork, "Differential Privacy: A Survey of Results," *Proc. 5th Int'l Conf. Theory and Applications of Models of Computation* (TAMC 08), M. Agrawal et al., eds., Springer, 2008, pp. 1–19.
11. K. Prewitt, "Why It Matters to Distinguish between Privacy and Confidentiality," *J. Privacy and Confidentiality*, vol. 3, no. 2, 2011, article 3.
12. D.J. Weitzner et al., "Information Accountability," *Comm. ACM*, vol. 51, no. 6, 2008, pp. 82–87.
13. H. Nissenbaum, "Privacy as Contextual Integrity," *Washington Law Rev.*, vol. 79, no. 1, 2004, pp. 19–158.

Hye-Chung Kum *is an associate professor at the Texas A&M Health Science Center School of Public Health, Department of Health Policy and Management, with an adjunct appointment in the School of Medicine Baylor Scott & White, Department of Pediatrics. She is also an adjunct associate professor at the University of North Carolina at Chapel Hill's Department of Computer Science, where she leads the Population Informatics Research Group. Her research interests include privacy-preserving entity resolution, secure data infrastructure for safe analysis of sensitive data, population informatics, data science, health services research, and health informatics. Kum received a PhD in computer science from the University of North Carolina at Chapel Hill. Contact her at kum@srph.tamhsc.edu.*

Ashok Krishnamurthy *is deputy director at the Renaissance Computing Institute (RENCI) and adjunct professor in the Department of Computer Science at the University of North Carolina at Chapel Hill. His research interests are in computational science and engineering, and signal and image processing. Ashok received a PhD in electrical and computer engineering from the University of Florida. He is a member of IEEE and ACM. Contact him at ashok@renci.org.*

Ashwin Machanavajjhala *is an assistant professor in the Department of Computer Science at Duke University. His primary research interests include data privacy, systems for massive data analytics, and statistical methods for information extraction and entity resolution. Machanavajjhala received a PhD in computer science from Cornell University. He is a member of ACM. Contact him at ashwin@cs.duke.edu.*

Stanley C. Ahalt *is director of RENCI and a professor in the Department of Computer Science at the University of North Carolina at Chapel Hill. His research interests include signal image and video processing, high-performance scientific and industrial computing, and secure data integration. Ahalt received a PhD in electrical and computer engineering from Clemson University. He is a member of the IEEE Computer Society, the Coalition for Academic and Scientific Computing (CASC), the Research Data Alliance (RDA), and the National Consortium for Data Science (NCDS). Contact him at ahalt@renci.org.*