Security and Privacy in the Era of Big Data

The SMW, a Technological Solution to the Challenge of Data Leakage

A RENCI/National Consortium for Data Science WHITE PAPER

renci ncds

CONTACT INFORMATION
Charles Schmitt, PhD
Telephone: 919.445.9696
Email: cschmitt@renci.org

Summary

The era of "big data" has ushered in a wealth of opportunities to advance science, improve health care, promote economic growth, reform our educational system, and create new forms of social interaction and entertainment. Yet these opportunities bring with them increasing challenges related to data security and privacy. The challenges include: a lack of effective tools and approaches for securely managing large-scale data and distributed data sets; third party data sharing; vulnerabilities in ever-expanding public databases; and technological advancements that are outpacing policy as it relates to digital security and privacy. Another major challenge is intentional or malicious data leakage. This White Paper discusses the challenge of data leakage and describes an innovative technological solution developed by RENCI staff; namely the Secure Medical Workspace (SMW).

The Challenges

Security and privacy concerns are growing as big data becomes more and more accessible. The collection and aggregation of massive quantities of heterogeneous data are now possible. Large-scale data sharing is becoming routine among scientists, clinicians, businesses, governmental agencies, and citizens. However, the tools and technologies that are being developed to manage these massive data sets are often not designed to incorporate adequate security or privacy measures, in part because we lack sufficient training and a fundamental understanding of how to provide large-scale data security and privacy. We also lack adequate policies to ensure compliance with current approaches to security and privacy. Furthermore, existing technological approaches to security and privacy are increasingly being breached, whether accidentally or intentionally, thus necessitating the continual reassessment and updating of current approaches to prevent data leakage.

Consider recent events related to the surveillance programs of the ***U.S. government***. The surveillance programs of the National Security Administration (NSA) are now common knowledge (Bamford, 2012; 2013; Greenwald, 2013; Greenwald and MacAskill, 2013a,b). The NSA has been at the forefront of efforts to collect and analyze massive amounts of data through its Stellar Wind program, Prism Program, and a variety of other data-intensive programs. The capabilities of the NSA are likely to expand with the opening of the Utah Data Center (Bamford, 2012), which presumably will focus on accessing the "deep web" ("deep net," "invisible web") or the countless records, reports, and other classified information from national and foreign governments. The Center will have massive computing capabilities that will facilitate the storage and analysis of unimaginable amounts of data and enable large-scale code-breaking (or decryption), indexing, and other technological approaches to breach and leak otherwise secure data and information. The Center's computational goal is to achieve computing capabilities on the level of exaflop (10^{18} bytes) by 2018, zettaflop by 2021 (10^{21} bytes), and yottaflop by 2024 (10^{24} bytes). Once these computing capabilities and analytics are developed, history indicates that they will surely become available to the public, either for free or by fee.

At a Glance

- Data leakage is a major challenge in digital security and privacy due to the availability of massive amounts of data, the ability to link disparate data sources, the increase in data sharing, and the absence of policies and procedures that reflect evolving technological capabilities.
- The SMW addresses the challenge of data leakage. It represents a major advance over current technological solutions to the data leakage challenge, including the incorporation of two-factor authentication, DLP technology, virtualization technology, and institution-specific security and privacy policies that can be tailored or updated as needed. The SMW is flexible and scalable and requires very little ongoing IT support.
- Plans are underway at RENCI to develop an enhanced, open source version of the SMW architecture that incorporates hypervisor-based, distributed DLP technology.

The Team

Charles P. Schmitt, PhD, RENCI CTO and Director of Informatics and Data Science; **Michael Shoffner**, RENCI Senior Research Software Architect; **Phillips Owen**, RENCI Senior Research Software Architect; **Xiaosho Wang**, RENCI Senior Biomedical Researcher; **Brent Lamm**, Director of Analytics, UNC Health Care; **Javed Mostafa**, PhD, Professor, School of Information and Library Science, UNC-Chapel Hill; **Mike Barker**, Assistant Vice Chancellor for Research Computing and Learning Technologies, UNC-Chapel Hill; **Ashok Krishnamurthy**, PhD, RENCI Deputy Director and Professor of Computer Science, UNC-Chapel Hill; **Kirk C. Wilhelmsen**, MD, PhD, RENCI Chief Domain Scientist for Genomics and Professor of Genetics and Neurology, UNC-Chapel Hill; **Stanley C. Ahalt**, PhD, RENCI Director and Professor of Computer Science, UNC-Chapel Hill; and **Karamarie Fecho**, PhD, Medical and Scientific Writer for RENCI.

Indeed, the NSA has an official technology transfer program (www.nsa.gov/research/tech_transfer/information_processing/index.shtml) and has already funded various business entities to develop analytical tools to enhance their capabilities, including IBM's System S/InfoSphere Streams, which provides high-speed, scalable, complex analytics to manage and interpret text, voice, video, and other heterogeneous data types in real-time (DBMS2, 2009a,b).

Private businesses, hospitals, and biomedical researchers are also making tremendous investments in the collection, storage, and analysis of large-scale data and private information. While the aggregation of such data presents a security concern in itself, another concern is that these rich databases are being shared with other entities, both private and public. Private business have already established partnerships with the government, as highlighted by the Prism Program, which provides the NSA with direct access to the databases of Microsoft, Yahoo, Google, Facebook, PalTalk, YouTube, Skype, AOL, and Apple (Greenwald and MacAskill, 2013a). Hospitals are increasingly adopting Electronic Medical Record (EMR) systems to enable the aggregation of patient data within a hospital and across a hospital system (Charles et al., 2013), and biomedical researchers are tapping into EMR data and new, nontraditional data sources such as social media, sensor-derived data (home, body, environment), and consumer purchasing and mobility patterns. In addition, the National Institutes of Health and other funding agencies are strongly encouraging biomedical researchers to share their research data.

Further, as anyone who has tried to remove themselves from an email or postal advertising list knows, businesses share personal information (for a fee) with other businesses. With more businesses and other parties engaging in third party and "open" market use of personal information, security and privacy breaches are likely to become more frequent. In addition, the tools to aggregate publically available data are now mature enough that they can be accessed and used by persons with little or no formal training in computer science or software programming. While regulations regarding data access and use apply to health care providers and biomedical researchers, few (if any) regulations exist to protect the individual consumer from unauthorized use of his/her data by industry, and few (if any) incentives exist for private industries or other entities to comply with established privacy and security policies and procedures (Nair, 2012). Indeed, hackers (discussed below) routinely capitalize on these inherent weaknesses in current business practices (Honan 2012a,b; Nair, 2012). Perhaps even more alarming: whereas the U.S. government invests massive amounts of money into large-scale, brute-force computing on relatively high-quality data, private businesses typically apply advanced analytics on low-quality data, thus enabling the capabilities of private industry to eventually become more powerful than those of the U.S. government (Bamford, 2012; 2013).

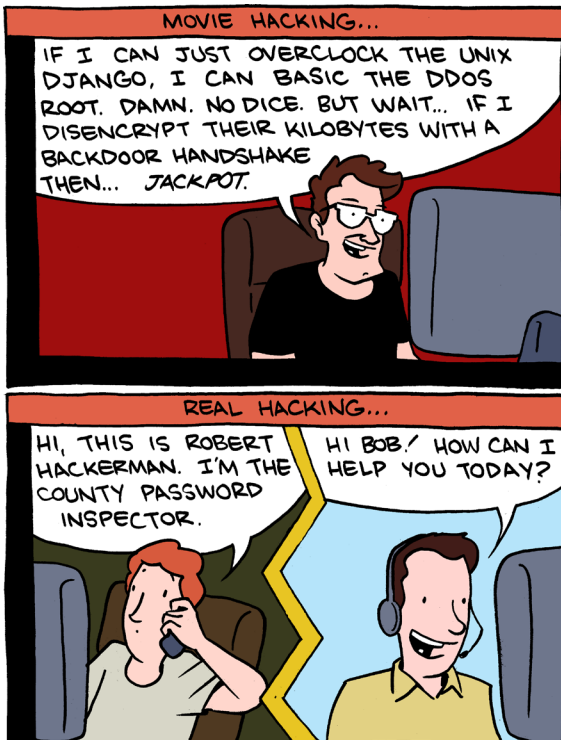
Data hackers have become more damaging in the era of big data due to the availability of large volumes of publically available data, the ability to store massive amounts of data on portable devices such as USB drives and laptops, and the accessibility of simple tools to acquire and integrate disparate data sources. According to the Open Security Foundation's DataLossDB project (http://datalossdb.org), hacking accounts for 28% of all data breach incidents, with theft accounting for an additional 24%, fraud accounting for 12%, and web-related loss accounting for 9% of all data loss incidents. Greater than half (57%) of all data loss incidents involve external parties, but 10% involve malicious actions on the part of internal parties, and an additional 20% involve accidental actions by internal parties.

Data leakage can be costly. In March 2012, hackers broke into Utah's Department of Health database and downloaded personal data from 780,000 patients (Social Security Numbers were downloaded for 280,000 patients); Utah's Governor responded by firing the Department's CTO; and the State of Utah was forced to pay-out ~$9 million on security audits, security upgrades, and identity theft protection measures for patients.

—Stewart and The Salt Lake Tribune, 2013

Matt Honan, Wired magazine's technology writer, has written extensively about his own personal experience with hacking (Honan 2012a,b). In less than an hour in August 2012, a hacker named Phobia destroyed his entire digital world (his Google, Apple, Twitter, Gmail, and Amazon accounts, containing much of his work portfolio and photos of his newborn child, were all erased) by acquiring only two pieces of personal information (a billing address and the last four digits of an online-filed credit card) and exploiting security flaws in Apple and Amazon's customer service procedures (automated password resets) and links between Honan's personal online accounts (Twitter link to personal website). In this case, the hacker's motivation was innocuous: he wanted Honan's 3-character Twitter handle. "I honestly

didn't have any heat towards you before this. I just liked your username like I said before," Phobia tweeted, but the damage to Honan was significant.



(Graphic reprinted with permission from Zach Weinersmith, Saturday Morning Breakfast Cereal, http://www.smbc-comics.com/index.php?db=comics&id=2526)

"Anonymous," to provide another example, is a leaderless group of hackers and technology experts who first originated in 2003 and whose motivation appears to be to prevent Internet censorship and control, but this group has breached major data networks and produced huge disruptions in business practices, including multiple "denial of service" attacks again PayPal and MasterCard, presumably in support of Julian Assange's efforts with WikiLeaks (Coleman, 2011; www.wikileaks.org). In July 2013, five hackers (four Russian, one Ukrainian) and an imprisoned U.S. hacker and conspirator were charged with the largest data breach in U.S history, resulting in the leakage of >160 million credit card numbers and associated personal information (names, passwords, PINs), with financial losses that have yet to be fully realized (Voreacos, 2013). The motivation in this case was to sell the data to third party "dump resellers" or illegal organizations and individuals who intend to use the data for their own financial gain.

Clearly, data leakage represents a major problem in today's era of big data. While stopping hackers from getting to data must be a goal, stopping hackers (and authorized users) from removing data from its authorized location is also a critical step to stop data leakage.

Existing Solutions to Protect Against Data Leakage

Ironically, ***oral or written pledges*** remain the most common method to protect against data breach and leakage and ensure compliance with security and privacy policies and procedures. Even the NSA relies on oral pledges to protect against intentional data leakages, but as the recent incident with NSA contractor Edward Snowden shows (Greenwald and MacAskill, 2013; Greenwald et al., 2013), oral pledges are useless if the motivation to leak data is stronger than the motivation to protect it.

Passwords and controlled access via permissions remain the most common technological approaches to protect against unauthorized access to data. Passwords have been used throughout history to authenticate personal identity. In the digital world, they consist of a string of typographical characters used for authentication to approve access to a computer system or other type of digital resource. While passwords can improve data security, they are not without limitations. Those limitations include the fact that passwords can be easily transferred from one person to another without permission of the owner of the data. People are also notoriously bad at password management and continue to rely on passwords that can be easily hacked (e.g., password rotation involving simple changes to an underlying password, the use of family names or birthdates) (Nair, 2012). Traditional approaches to password resets and the use of temporary passwords introduce vulnerabilities. Indeed, customer service staff members at Apple and Amazon inadvertently assisted in the hacking of Matt Honan's digital files through password security vulnerabilities (i.e., answers to password reset security questions that an outsider could easily discover or find and the default password reset solution of sending a temporary password to a person's email) (Honan, 2012a,b; Nair, 2012). More secure solutions to reset passwords (e.g., blocked access, postal mail requests to reset passwords) are considered too cumbersome on the user and thus unlikely to be adopted.

Two-factor (or multi-factor) authentication represents an improvement over the simple password. While not restricted to the digital world, two-factor

authentication historically requires that a user submits two of three authentication factors before gaining access to data or another resource. Those factors generally include: (1) something a user knows (a password); (2) something a user has (a physical bank card); and (3) something a user is (a biometric characteristic such as a fingerprint) (Federal Financial Institutions Examination Council, http://www.ffiec.gov/pdf/authentication_guidance.pdf). The most common example of this security approach is the ATM machine, which requires a bank card and a Personal Identification Number (typically four digits). Digital application typically requires the use of two passwords. While better than password-only (one-factor) authentication, two-factor authentication is vulnerable to the same limitations as passwords.

Data Leakage Prevention (DLP) Technology is a relatively new approach to data security that was introduced in the mid-2000s and designed to protect against the leakage of sensitive or confidential data, particularly by insiders with malicious intent (Weinberg, 2008; Ouellet for Gartner, 2013). DLP technology involves the inspection of data packets by location and file classification and the restriction of data movement within and outgoing from an internal network through the enforcement of policies that are based on data location and file classification. Limitations of DLP technology include the fact that it can be too stringent for both end users and Information Technology (IT) staff if the policies are rigid, thus discouraging use and compliance. DLP technology also does not protect against accidental or intentional data leakage through the removal of data via a computer that is removed from the environment protected by the DLP technology. Additionally, as with all policy-based solutions, DLP technology is difficult to implement in a distributed environment; a centralized environment is generally more conducive to this technology.

Ideas Into Action: The Secure Medical Workspace

The Secure Medical Workspace, or ***SMW***, was developed by RENCI in collaboration with the University of North Carolina (UNC) and UNC Health Care as a novel, comprehensive solution to the challenge of data leakage. Specifically, the SMW was developed to prevent the intentional or inadvertent leakage of patient data from EMR systems through printing, downloading, emailing, physical removal of disks, etc.—a major vulnerability of EMR systems (Weinberg, 2008; Oullet for Garner, 2013; Shoffner et al., 2013).

In the past, health care professionals and researchers at UNC who requested sensitive data from systems within UNC Health Care were provided, after obtaining proper authorization, data extracts sent via secure transfer methods (e.g., secure email) by UNC's Medical Information Management Department. While this model served to ensure that sensitive data were provided only under sanctioned circumstances, and while it triggered the proper disclosure process, there was no way to control the security of the data after it was provided to the requester, and the data extracts could be printed, copied to external drives, or emailed.

The SMW provides approved requesters with easy access to sensitive patient data for patient care or for approved clinical research on a secure virtual workspace, coupled with the ability to prevent (or allow with a challenge and auditing by IT staff) the physical removal of data from a central, secure, storage environment (Mostafa 2010; Kamerick and Mostafa 2011; Owen et al., 2011; Mostafa and Shoffner 2012; Shoffner, 2012; Shoffner et al., 2013).

Key technological features include: (1) two-factor authentication to gain access to the SMW; (2) virtualization technology to provide access to sensitive data on a secure, centralized, network-based file server equipped with typical Microsoft Windows file techniques and other software resources to open, view, manipulate, and analyze sensitive data within the secure virtual workspace but without the ability to physically remove sensitive data; (3) preconfigured virtual machine images that conform to and automatically implement UNC's security policies (e.g., password changes every 30 days, automated anti-virus software updates, disallowing new software installation); (4) use of encryption for data in motion and data at rest; and (5) IT management capabilities to provide for easy administration of the entire solution and equip IT administrators with an audit trail of data removed from the secured environment.

The SMW architecture incorporates commercial technologies (VMWare for virtualization, WebSense for endpoint DLP) that were assessed against needed user features and a comprehensive matrix required by institutional (in this case, UNC) security policies (Barker and Reed, 2013).

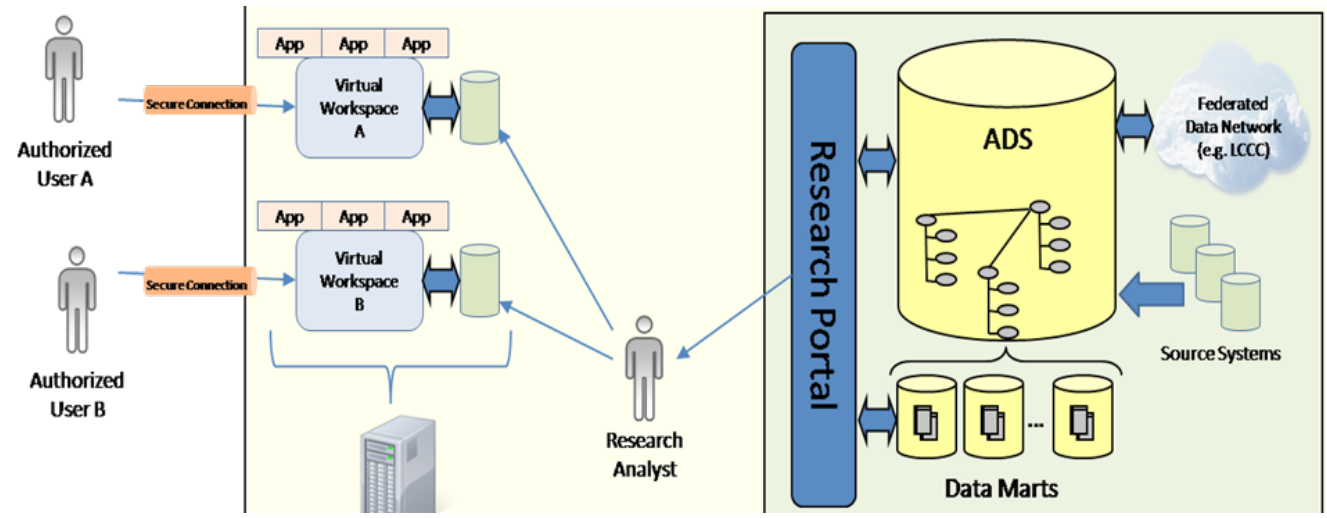


Figure 1. The Secure Medical Workspace. From Shoffner et al., *Clin Trans Sci.* 6:222-225. 2013. Reprinted with permission from John Wiley & Sone Publishers.

The SMW is flexible in that it: (1) allows researchers to install and use (within the secured workspace) analytical tools and other applications that are necessary for specific research projects; (2) enables policies to be updated and/or revised as needed; and (3) is capable of supporting both small research projects and very large, data-intensive research projects involving multiple institutions (i.e., data capacity is defined only by the institution's underlying infrastructure). The SMW is also easy to install and deploy and requires very little IT support after installation. There is no limit to the number of instances of the SMW that can be deployed. Of note, the SMW architecture is available as open source technology.

A patent for the Secure Medical Workspace has been filed to enable future commercialization (Schmitt, Chase, Baldine, Coposky, Shoffner, Lamm, Mostafa, U.S. Patent Cooperation Treaty, Serial No. 61/675,780 for SECURE RESEARCH SPACE), but an open source version of the architecture will remain accessible.

UNC has deployed the SMW for data management in three large, data-intensive research projects: NCGENES (North Carolina Clinical Genomic Evaluation by NextGen Exome Sequencing) in the Department of Genetics; SAS Healthcare Analytics and the UNC Survivorship Cohort Central Tracking System, a collaboration between UNC's Lineberger Comprehensive Cancer Center and SAS; and the Integrated Cancer Information and Surveillance System at the Lineberger Comprehensive Cancer Center.

UNC is also evaluating the cost effectiveness of the SMW for campus-wide adoption as a central component of its Information Security Plan following an assessment of the technology by UNC's Research Computing Department (Barker and Reed, 2013). Three other academic medical schools (Duke University, The Ohio State University, and the University of Pittsburgh) have either adopted the SMW or have committed to do so by 2014.

RENCI, in collaboration with investigators in the Department of Computer Science at UNC, is developing new approaches to improve data leakage protection technology under a grant from the Department of Homeland Security (Fabian Monrose, PhD, Principal Investigator). This new work is based on existing security technologies developed by Monrose and colleagues for application in forensic science to enable the tagging and tracking of data access and use from within a virtualization hypervisor (Krishnan and Monrose, 2009; Krishnan et al., 2010).

The new approach will enable the tagging, tracking, and blocking of access to sensitive data, as well as data sets derived from those data, while the data resides in a virtual workspace. This capability will offer adopters the ability to apply restrictive security policies only to those data sets that contain sensitive data, thus reducing unnecessary restrictions on the flow of data into and out of the secure virtual environment and allowing IT staff to audit sensitive data as it is modified, saved, and copied by different user applications. Importantly, the technology resides partially within the hypervisor, thereby reducing the risk of malicious attacks intended to compromise the DLP technology.

The Upshot

The SMW technology addresses many of the current security and privacy challenges in the protection against data leakage in the era of big data:

1. Intentional breaches related to broken verbal or written agreements or malicious attacks are prevented due to the inability to remove data from the SMW.
2. Third-party use by non-authorized users of the data is likewise prohibited.
3. Two-factor authentication is required to access the technology and use of encryption is allowed for data at rest and in motion.
4. DLP technology is embedded within the technology and improvements to existing DLP approaches are under development.
5. Security and privacy policies can be adapted to a given institution's policies and automated and updated as needed.
6. The architecture is being developed as open source software and thus available for use by any group, regardless of budgetary concerns.
7. The technology is easy to implement and maintain, with very little ongoing effort on the part of IT staff.
8. The technology enables the incorporation of access to user-specific software within the virtual workspace environment.
9. The technology is scalable and can be implemented in a distributed environment.
10. The technology can be adapted for numerous applications and varied environments.

The Big Picture

Existing technologies will continue to evolve as needs in data security and privacy are recognized and additional vulnerabilities are realized. The SMW is a novel technological solution to the challenge of data leakage. We predict that this technology will be refined and advanced as new vulnerabilities are realized. As discussed, we already are developing hypervisor-based, distributed DLP technology to incorporate into the system, as well as investigating approaches to embed the SMW technology with data management systems such as iRODS (www.irods.org).

The SMW technology represents a unique advancement in protection against data leakage. While developed initially to secure EMR data, we envision the application of the SMW technology in virtually any scenario in which the security and privacy of big data are of monumental importance. Examples include financial institutions, academic institutions, and private industries that wish to brand themselves as being free from unacknowledged third-party use of their customers' data.

Acknowledgments

This project was supported in part by the National Center for Advancing Translational Sciences, National Institutes of Health (UL1TR000083). Karen Green provided editorial and design support for the preparation of this white paper, and Tasha Wilhelmsen provided proofreading and assistance with the references. RENCI provided funding for this support.

References

Bamford, J. (2013). Five myths about the National Security Agency. *The Washington Post*. http://articles.washingtonpost.com/2013-06-21/opinions/40114085_1_national-security-agency-foreign-intelligence-surveillance-court-guardian. [Accessed June 25, 2013]

Bamford, J. (2012). The NSA is building the country's biggest spy center (watch what you say). *WIRED*. http://www.wired.com/threatlevel/2012/03/ff_nsadatacenter. [Accessed June 25, 2013]

Barker, M., & Reed, M. S. C. (2013). A research environment for high risk data. Presented at the Research Data Management Implementations Workshop. Chicago, IL, USA: The University of Chicago. http://rdmi.uchicago.edu/sites/rdmi.uchicago.edu/files/uploads/Barker,%20M%20and%20Reed,%20M_A%20Research%20Environment%20for%20High%20Risk%20Data.pdf. [Accessed September 13, 2013]

Coleman, E. G. (2011). Anonymous: From the lulz to collective action. *MediaCommons*. http://mediacommons.futureofthebook.org/tne/pieces/anonymous-lulz-collective-action. [Accessed July 8, 2013]

Charles, D., King, J., Patel, V., & Furukawa, M. F. (2013). Adoption of Electronic Health Record systems among U.S. non-federal acute care hospitals: 2008-2012. *ONC Data Brief No. 9.* Washington, DC, USA: Office of the National Coordinator for Health Information Technology. http://www.healthit.gov/sites/default/files/oncdata-brief9final.pdf. [Accessed July 15, 2013]

DBMS2. (2009a). Followup on IBM System S/InfoSphere Streams. DBMS2. http://www.dbms2.com/2009/05/18/followup-on-ibm-system-sinfosphere-streams. [Accessed August 2, 2013]

DBMS2. (2009b). IBM System S Streams, aka InfoSphere Streams, aka stream processing, aka "please don't call it CEP. DBMS2. http://www.dbms2.com/2009/05/13/ibm-system-s-infosphere-streams-processing. [Accessed August 2, 2013]

Greenwald, G. (2013). XKeyscore: NSA tool collects 'nearly everything a user does on the internet'. *The Guardian*. http://www.theguardian.com/world/2013/jul/31/nsa-top-secret-program-online-data. [Accessed July 31, 2013]

Greenwald, G., & MacAskill, E. (2013a). NSA Prism program taps in to user data of Apple, Google and others. *The Guardian.* http://www.theguardian.com/world/2013/jun/06/us-tech-giants-nsa-data. [Accessed July 3, 2013]

Greenwald, G., MacAskill, E., & Poitras, L. (2013b). Edward Snowden: The whistleblower behind the NSA surveillance revelations. The 29-year-old source behind the biggest intelligence leak in the NSA's history explains his motives, his uncertain future and why he never intended on hiding in the shadows. *The Guardian*. http://www.theguardian.com/world/2013/jun/09/edward-snowden-nsa-whistleblower-surveillance. [Accessed August 7, 2013]

Honan, M. (2012a). How Apple and Amazon security flaws led to my epic hacking. *WIRED*. http://www.wired.com/gadgetlab/2012/08/apple-amazon-mat-honan-hacking. [Accessed July 8, 2013]

Honan, M. (2012b). Mat Honan: How I resurrected my digital life after an epic hacking. *WIRED*. http://www.wired.com/gadgetlab/2012/08/mat-honan-data-recovery. [Accessed July 8, 2012]

Kamerick, M., Mostafa, J., & Ervin, D. (2011). Challenges and emerging solutions for securing medical data and applications. Presented at the 2011 annual meeting of the American Medical Informatics Association, San Francisco, CA, USA.

Krishnan, S., Snow, K. Z., & Monrose, F. (2010). Trail of bytes: Efficient support for forensic analysis. Proceedings of the 17th ACM Conference on Computer and Communications Security - CCS '10, pp. 50-60. New York, New York, USA: ACM Press. http://cs.unc.edu/~fabian/papers/trail10.pdf. [Accessed July 31, 2013]

Krishnan, S., & Monrose, F. (2009). TimeCapsule: Secure recording of accesses to a protected datastore. Proceedings of the 1st ACM Workshop on Virtual Machine Security - VMSec '09, pp. 23-32. New York, New York, USA: ACM Press.

Mostafa, J. (2010). Supporting security in applications and services for secondary data use: UNC TraCS Data Warehouse experience. Presented at the 2010 CTSA Annual Informatics Meeting, Bethesda, MD, USA p. 6. https://www.signup4.net/UPLOAD/BOOZ12A/CTSA28E/Abstracts_20101008.pdf. [Accessed July 31, 2013]

Mostafa, J., & Shoffner, M. (2012). Distributed and cross-institutional secure medical access for research and quality improvement. Presented at the 2012 CTSA Annual Informatics Meeting, Bethesda, MD, USA, p. 7. https://www.ctsacentral.org/sites/default/files/documents/Abstract_Book_2012_0.pdf. [Accessed July 31, 2013]

Nair, L. V. J. (2012) How you are helping hackers steal your data. hongkiat.com. http://www.hongkiat.com/blog/keeping-online-data-safe. [Accessed July 8, 2013]

Ouellet, E. (2013). Magic quadrant for content-aware data loss prevention. Publication ID G00224160, Stamford, CT, USA: Gartner. http://www.computerlinks.de/FMS/22876.magic_quadrant_for_content_aware_data_loss_prevent.pdf. [Accessed August 6, 2013]

Owen, P., Shoffner, M., Wang, X., Schmitt, C., Lamm, B., & Mostafa, J. (2011). Secure Medical Research Workspace. RENCI Technical Report No. TR-11-01. Chapel Hill, NC, USA: RENCI. http://www.renci.org/publications/technical-reports. [Accessed July 31, 2013]

Shoffner, M. (2012). Handling "hot" health data without getting burned. Presented at the Strata Rx Conference, San Francisco, CA, USA. http://strataconf.com/rx2012/public/schedule/detail/26231. [Accessed August 8, 2013]

Shoffner, M., Owen, P., Mostafa, J., Lamm, B., Wang, X., Schmitt, C. P., & Ahalt, S. C. (2013). The Secure Medical Research Workspace: An IT infrastructure to enable secure research on clinical data. *Clin Trans Sci.* 6 (3), 222-225.

Stewart, K. (2013). Report: Utah's health data breach was a costly mistake. The Salt Lake Tribune. http://www.sltrib.com/sltrib/news/56210404-78/security-breach-health-data.html.csp. [Accessed July 19, 2013]

Voreacos, D. (2013). 5 hackers charged in largest data-breach scheme in U.S. *Bloomberg*. http://www.bloomberg.com/news/2013-07-25/5-hackers-charged-in-largest-data-breach-scheme-in-u-s-.html. [Accessed July 26, 2013]

Weinberg, N. (2008). Data leakage prevention: Hot technology for 2008. *NETWORKWORLD*. http://www.networkworld.com/research/2008/011408-8-techs-data-leakage.html. [Accessed July 19, 2013]

Websites

Integrated Rules-Oriented Data System (iRODS), www.irods.org.

National Consortium for Data Science, www.data2discovery.org.

National Security Agency, Central Security Service, www.nsa.gov/research/tech_transfer/information_processing/index.shtml.

Open Security Foundation, DataLossDB project, datalossdb.org.

WikiLeaks, www.wikileaks.org.

About RENCI

RENCI, an institute of UNC Chapel Hill, develops and deploys advanced technologies to enable research discoveries and practical innovations. RENCI partners with scientists, policy makers, and industry to engage and solve the problems that affect North Carolina, the U.S., and the world. RENCI is a collaboration involving UNC Chapel Hill, Duke University and North Carolina State University. For more information, see www.renci.org.

About NCDS

The National Consortium for Data Science (NCDS) is a public-private partnership that offers a foundation for advancing data science research, educating the next generation of data scientists, and translating data innovations into economic opportunity. The NCDS formed in 2012 in the Research Triangle area of North Carolina. For more information, see www.data2discovery.org

How to reference this paper:
Schmitt, C., Shoffner, M., Owen P., Wang, X., Lamm, B., Mostafa, J., Barker, M., Krishnamurthy, A., Wilhelmsen, K., Ahalt, S., & Fecho, K. (2013): Security and Privacy in the Era of Big Data: The SMW, a Technological Solution to the Challenge of Data Leakage. RENCI, University of North Carolina at Chapel Hill. Text. http://dx.doi.org/10.7921/G0WD3XHT

Vol. 1, No. 2 in the RENCI White Paper Series, November 2013. Created in collabortion with the **National Consortium for Data Science** (www.data2discovery.org).