Appendices

Appendix 1: Curriculum Vitae

CVs including education, employment history, a list of the proposer's most important previous publications, the journals in which they appeared, book reviews, and important honors and other awards have been supplied as individual attachments.

Appendix 2: Conflict of Interest

There are no potential significant conflicts of interest or source of bias related to this proposal on behalf of the primary investigators, key project staff, or the grantee institution.

All activities conducted or hosted by the Program on Information science, which hosts this project, are governed by the Massachusetts Institute of Technology's conflict of interest policy. Activities conducted at Harvard University are governed by substantially similar policies.[1] These policy requires that all MIT and Harvard researchers comply with both Federal and sponsor-specific conflict of interest (COI) disclosure requirements. This policy requires annual disclosure and administrative review of all significant financial interests, and details on how an SFI entity does or does not have a relationship to each sponsored research project.

A significant financial interest explicitly includes but is not limited to:

- Remuneration valued at $5K or more in the last 12 months from any entity other than the university

[1] See http://research.fas.harvard.edu/policies/financial-conflicts-interest-disclosures ; http://law.harvard.edu/faculty/resources/conflict-of-interest-policy.html

- Any equity interest in a non-publicly traded entity; and any equity interest of at least $5K in a publicly traded entity
- Any income related to patents or other intellectual property rights

For information on MIT's full policies on responsible and ethical conduct of interest, including conflict of interest policies and procedures see: http://web.mit.edu/conduct/conflict.html.

Appendix 3: Attention to Diversity

The Massachusetts Institute of Technology, which host this project, is committed to the principle of equal opportunity in education and employment. MIT & Harvard do not discriminate against individuals on the basis of race, color, sex, sexual orientation, gender identity, religion, disability, age, genetic information, veteran status, ancestry, or national or ethnic origin in the administration of its educational policies, admissions policies, employment policies, scholarship and loan programs, and other Institute administered programs and activities.[2] The institute provides a range of resources, events and initiatives support of diversity.[3]

The Program on Information Science, which hosts this project and is part of MIT, respects and values these differences, fosters approaches to problem-solving and decision-making that are multi-dimensional, and strives to creating an atmosphere of civility, collegiality, mutual respect, and inclusion in its staff, in workshops and other events it hosts, and in its educational offerings.

---

[2] *See* http://web.mit.edu/referencepubs/nondiscrimination for the institute's full nondiscrimination statement. See http://www.law.harvard.edu/current/careers/ocs/employers/recruiting-policies-employers/index.html , https://www.seas.harvard.edu/sites/default/files/files/Academic%20Affairs/FAS-Handbook.pdf for relevant Harvard policies.

[3] *See* http://diversity.mit.edu; http://web.mit.edu/facultyworklife/community/diversitymit.html; http://hrweb.mit.edu/diversity.

When hosting workshops, conferences, training and other public events, our approach for non-expert recruitment is to disseminate information using the broadest applicable Institute, professional society, and disciplinary distribution lists and event-calendars. In addition, we actively seek to identify distribution channels, explicitly targeting diverse scholars within the relevant discipline, and to prioritize travel support (where available) to promote diversity and participation of junior scholars.

Appendix 4: Bibliography

- [Adi08] B. Adida. Helios: Web-based open-audit voting. In P. C. van Oorschot, editor, USENIX Security Sympo- sium, pages 335–348. USENIX Association, 2008.
- [Alt12] Micah Altman, "Mitigating Threats to Data Quality Throughout the Curation Lifecycle" (position paper from a workshop, Curating For Quality: Ensuring Data Quality to Enable New Science, Arlington, Virginia, September 10-11, 2012), http://datacuration.web.unc.edu
- [BZ06] M. Barbarao and T. Zeller. A face is exposed for aol searcher 4417749. New York Times, page A1, 9 August 2006.
- [BNS13a] A. Beimel, Nissim, K., and Stemmer, U., "Characterizing the Sample Complexity of Private Learners", in Proceedings of the 4th Conference on Innovations in Theoretical Computer Science, New York, NY, USA, 2013, pp. 97–110.
- [BNS13b] Amos Beimel, Kobbi Nissim, Uri Stemmer: Private Learning and Sanitization: Pure vs. Approximate Differential Privacy. APPROX-RANDOM 2013: 363-378
- [BNS14] Amos Beimel, Kobbi Nissim, Uri Stemmer: Learning Privately with Labeled and Unlabeled Examples. CoRR abs/1407.2662 (2014)
- [BL07] J. Bennett and S. Lanning. The Netflix prize. In *Proceedings of KDD Cup and Workshop*, 2007.
- [BDMN05] A. Blum, C. Dwork, F. McSherry, and K. Nissim. Practical privacy: the SuLQ framework. In C. Li, editor, PODS, pages 128–138. ACM, 2005.
- [BLR08] A. Blum, K. Ligett, and A. Roth. A learning theory approach to non-interactive database privacy. In R. E. Ladner and C. Dwork, editors, STOC, pages 609–618. ACM, 2008.
- [BUV14] M. Bun, Ullman, J., and Vadhan, S., "Fingerprinting Codes and the Price of Approximate Differential Privacy", in Proceedings of the 46th Annual ACM Symposium on Theory of Computing, New York, NY, USA, 2014, pp. 1–10.
- [CCK+13] Y. Chen, Chong, S., Kash, I. A., Moran, T., and Vadhan, S., "Truthful mechanisms for agents that value privacy", in Proceedings of the fourteenth ACM conference on Electronic commerce, Philadelphia, Pennsylvania, USA, 2013, pp. 215-232.

- [Cro11] M. Crosas. The Dataverse Network: An open-source application for sharing, discovering and preserving data. *D-Lib Magazine*, 17(1-2), 2011.
- [Cro13] M. Crosas. A data sharing story. *Journal of eScience Librarianship*, 1(3):173-179, 2013.
- [DN03] I. Dinur and K. Nissim. Revealing information while preserving privacy. In Proceedings of the Twenty- Second ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems, pages 202–210, 2003.
- [DLMV12] Y. Dodis, López-Alt, A., Mironov, I., and Vadhan, S., "Differential Privacy with Imperfect Randomness", in Proceedings of the 32nd International Cryptology Conference (CRYPTO `12), Santa Barbara, CA, 2012, Lecture Notes on Computer Science., vol. 7417, pp. 497–516.
- [DMNS06] C. Dwork, F. McSherry, K. Nissim, and A. Smith. Calibrating noise to sensitivity in private data analysis. In S. Halevi and T. Rabin, editors, *TCC*, volume 3876 of *Lecture Notes in Computer Science*, pages 265–284. Springer, 2006.
- [DNR+09] Cynthia Dwork, Moni Naor, Omer Reingold, Guy N. Rothblum, and Salil Vadhan. 2009. On the complexity of differentially private data release: efficient algorithms and hardness results. In Proceedings of the forty-first annual ACM symposium on Theory of computing (STOC '09). ACM, New York, NY, USA, 381-390. DOI=10.1145/1536414.1536467 http://doi.acm.org/10.1145/1536414.1536467
- [DNV12] Cynthia Dwork, Moni Naor, and Salil Vadhan. 2012. The Privacy of the Analyst and the Power of the State. In Proceedings of the 2012 IEEE 53rd Annual Symposium on Foundations of Computer Science (FOCS '12). IEEE Computer Society, Washington, DC, USA, 400-409. DOI=10.1109/FOCS.2012.87 http://dx.doi.org/10.1109/FOCS.2012.87
- [DN04] C. Dwork and K. Nissim. Privacy-preserving datamining on vertically partitioned databases. In Proceedings of CRYPTO 2004, volume 3152, pages 528–544, 2004.
- [DRV10] Cynthia Dwork, Guy N. Rothblum, and Salil Vadhan. 2010. Boosting and Differential Privacy. In Proceedings of the 2010 IEEE 51st Annual Symposium on Foundations of Computer Science (FOCS '10). IEEE Computer Society, Washington, DC, USA, 51-60. DOI=10.1109/FOCS.2010.12 http://dx.doi.org/10.1109/FOCS.2010.12
- [Fair74] Fair information practice principles. U.S. 1974 Privacy Act, 1974. http://www.ftc.gov/reports/privacy3/fairinfo.shtm.
- [Fel08] J. Felch. DNA databases blocked from the public. Los Angeles Times, page A31, 29 August 2008.
- [FFKN09] Dan Feldman, Amos Fiat, Haim Kaplan, Kobbi Nissim: Private coresets. STOC 2009: 361-370.
- [Fog05] Fogel, K. (2005). Producing open source software: How to run a successful free software project. O'Reilly Media, Inc.
- [FWC+10] Benjamin C. M. Fung, Ke Wang, Rui Chen, and Philip S. Yu. 2010. Privacy-preserving data publishing: A survey of recent developments. ACM Comput. Surv. 42, 4, Article 14 (June 2010), 53 pages. DOI=10.1145/1749603.1749605 http://doi.acm.org/10.1145/1749603.1749605
- [GHH+13] Marco Gaboardi, Andreas Haeberlen, Justin Hsu, Arjun Narayan, Benjamin C. Pierce: Linear dependent types for differential privacy. POPL 2013: 357-370

- [GS06] S. L. Garfinkel and M. D. Smith, editors. Data Surveillance, IEEE Security & Privacy, volume 4. IEEE, 2006. Special Issue.
- [GR11] Arpita Ghosh and Aaron Roth. 2011. Selling privacy at auction. In Proceedings of the 12th ACM conference on Electronic commerce (EC '11). ACM, New York, NY, USA, 199-208. DOI=10.1145/1993574.1993605 http://doi.acm.org/10.1145/1993574.1993605
- [GS07] M. Gutmann and P. Stern. Putting People on the Map: Protecting Confidentiality with Linked Social- Spatial Data. National Academy Press, Washington, DC, 2007.
- [Hig08] Sarah Higgins, "The DCC Curation Lifecycle Model," International Journal of Digital Curation 3, no. 1 (2008): 134-140, http://www.dcc.ac.uk/resources/curation-lifecycle-model
- [HRN+14] Ho, Andrew Dean and Reich, Justin and Nesterko, Sergiy O and Seaton, Daniel Thomas and Mullaney, Tommy and Waldo, Jim and Chuang, Isaac, HarvardX and MITx: The First Year of Open Online Courses, Fall 2012-Summer 2013 (January 21, 2014). Ho, A. D., Reich, J., Nesterko, S., Seaton, D. T., Mullaney, T., Waldo, J., & Chuang, I. (2014). HarvardX and MITx: The first year of open online courses (HarvardX and MITx Working Paper No. 1).. Available at SSRN: http://ssrn.com/abstract=2381263 or http://dx.doi.org/10.2139/ssrn.2381263
- [HD14] J. Honaker and V. D'Orazio. Statistical Modeling by Gesture: A graphical, browser-based statistical interface for data repositories. Extended Proceedings of ACM Hypertext 2014.
- [HRZ11] J. J. Horton, D. G. Rand, and R. J. Zeckhauser. The online laboratory: Conducting experiments in a real labor market. Experimental Economics, 14:399–425, 2011. Available at SSRN: http://ssrn.com/abstract=1591202.
- [HDF+10] A. Hundepool, J. Domingo-Ferrer, L. Franconi, S. Giessing, R. Lenz, J. Naylor, E. Schulte Nordholt, G. Seri, P.-P. De Wolf. A Network of Excellence in the European Statistical System in the field of Statistical Disclosure Control (ESSNET), Handbook on Statistical Disclosure Control. January 2010. http://neon.vb.cbs.nl/casc/SDC_Handbook.pdf
- [IKL08] K. Imai, G. King, and O. Lau. Toward a common framework for statistical analysis and development. Journal of Computational and Graphical Statistics, 17(4):892–913, 2008.
- [IQSS06] IQSS Dataverse Network (Created 2006). The Institute for Quantitative Social Science at Harvard University.
- [KLN+11] Shiva Prasad Kasiviswanathan, Homin K. Lee, Kobbi Nissim, Sofya Raskhodnikova, Adam Smith: What Can We Learn Privately? SIAM J. Comput. 40(3): 793-826 (2011).
- [Kin07] G. King. An introduction to the Dataverse Network as an infrastructure for data sharing. Sociological Methods and Research, 36:173-199, 2007.
- [Kin09] G. King. The changing evidence base of social science research. In K. Schlozman and N. Nie, editors, *The Future of Political Science: 100 Perspectives*. Routledge Press, 2009.
- [Kin14] G. King. Restructuring the social sciences: Reflection from Harvard's Institute for Quantitative Social Science. *PS: Political Science and Politics*, 47(1):165-172, 2014.

- [KTW00] G. King, M. Tomz, and J. Wittenberg. Making the most of statistical analyses: Improving interpretation and presentation. American Journal of Political Science, 44(2):347-361, 2000.
- [KNRS13] S. P. Kasiviswanathan, Nissim, K., Raskhodnikova, S., and Smith, A., "Analyzing Graphs with Node Differential Privacy", in Proceedings of the 10th Theory of Cryptography Conference on Theory of Cryptography, Berlin, Heidelberg, 2013, pp. 457–476.
- [LPA+09] D. Lazer, A. Pentland, L. Adamic, S. Aral, A.-L. Barabasi, D. Brewer, N. Christakis, J. F. N. Contractor, M. Gutmann, T. Jebara, G. King, M. Macy, D. Roy, , and M. V. Alstyne. Computational social science. Science, 2009.
- [LKG+08] K. Lewis, J. Kaufman, M. Gonzalez, A. Wimmer, and N. Christakis. Tastes, ties and time: a new social network dataset using facebook. In *Social Networks*, volume 30, 2008.
- [LKG+09] K. Lewis, J. Kaufman, M. Gonzalez, A. Wimmer, and N. Christakis. Tastes, ties and time dataverse. Technical report, Harvard IQSS, Cambridge, MA, 2009.
- [MKA+08] Ashwin Machanavajjhala, Daniel Kifer, John M. Abowd, Johannes Gehrke, Lars Vilhuber: Privacy: Theory meets Practice on the Map. ICDE 2008: 277-286
- [MS00] B. Malin and L. Sweeney. Determining the identifiability of DNA database entries. In *Proceedings, Journal of the Medical Informatics Association*, Washington, DC, 2000. Hanley Belfus.
- [MS01] B.Malin and L. Sweeney. Reidentification of DNA through an automated linkage process. In *Proceedings, Journal of the Medical Informatics Association*, Washington, DC, 2001. Hanley Belfus.
- [MS02] B.Malin and L. Sweeney. Pacific symposium on biocomputing 2002. In *Inferring Genotype from Clinical Phenotype through a Knowledge Based Algorithm*, Singapore, 2002. World Scientific.
- [MMP+10] A. McGregor, Mironov, I., Pitassi, T., Reingold, O., Talwar, K., and Vadhan, S., "The Limits of Two-Party Differential Privacy", in Proceedings of the 51st Annual {IEEE} Symposium on Foundations of Computer Science (FOCS `10), Las Vegas, NV, 2010, pp. 81–90.
- [McS09] F. McSherry. Privacy integrated queries: an extensible platform for privacy-preserving data analysis. SIGMOD Conference 2009: 19-30
- [MPRV09] I. Mironov, Pandey, O., Reingold, O., and Vadhan, S., "Computational Differential Privacy", in Advances in Cryptology–-CRYPTO `09, Santa Barbara, CA, 2009, vol. 5677, pp. 126–142.
- [MTS+12] Prashanth Mohan, Abhradeep Thakurta, Elaine Shi, Dawn Song, and David Culler. 2012. GUPT: privacy preserving data analysis made easy. In Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data (SIGMOD '12). ACM, New York, NY, USA, 349-360. DOI=10.1145/2213836.2213876 http://doi.acm.org/10.1145/2213836.2213876
- [NS08] A. Narayanan and V. Shmatikov. Robust de-anonymization of large sparse datasets. In IEEE Symposium on Research in Security and Privacy, Oakland, CA, 2008. IEEE.
- [NOS12] Kobbi Nissim, Claudio Orlandi, and Rann Smorodinsky. 2012. Privacy-aware mechanism design. In Proceedings of the 13th ACM Conference on Electronic

Commerce (EC '12). ACM, New York, NY, USA, 774-789. DOI=10.1145/2229012.2229073 http://doi.acm.org/10.1145/2229012.2229073

- [NRS07] Kobbi Nissim, Sofya Raskhodnikova, Adam Smith: Smooth sensitivity and sampling in private data analysis. STOC 2007: 75-84
- [NST12] Kobbi Nissim, Rann Smorodinsky, and Moshe Tennenholtz. 2012. Approximately optimal mechanism design via differential privacy. In Proceedings of the 3rd Innovations in Theoretical Computer Science Conference (ITCS '12). ACM, New York, NY, USA, 203-213. DOI=10.1145/2090236.2090254 http://doi.acm.org/10.1145/2090236.2090254
- [NVX14] Kobbi Nissim, Salil Vadhan, and David Xiao. 2014. Redrawing the boundaries on purchasing data from privacy-sensitive individuals. In Proceedings of the 5th conference on Innovations in theoretical computer science (ITCS '14). ACM, New York, NY, USA, 411-422. DOI=10.1145/2554797.2554835 http://doi.acm.org/10.1145/2554797.2554835
- [OBB+12] L. Ohno-Machado, V. Bafna, A.A. Boxwala, B.E. Chapman, W.W. Chapman, K. Chaudhuri, M.E. Day, C. Farcas, N.D. Heintzman, X. Jiang, H. Kim, J. Kim, M.E. Matheny, F.S. Resnic, S.A. Vinterbo, and the iDASH team. iDASH: integrating data for analysis, anonymization, and sharing. J Am Med Inform Assoc 2012;19:196-201 doi:10.1136/amiajnl-2011-000538. http://jamia.bmjjournals.com/content/19/2/196.full
- [RAW+10] Jason Reed, Adam J. Aviv, Daniel Wagner, Andreas Haeberlen, Benjamin C. Pierce, Jonathan M. Smith: Differential privacy for collaborative security. EUROSEC 2010:1-7
- [RP10] Jason Reed, Benjamin C. Pierce: Distance makes the types grow stronger: a calculus for differential privacy. ICFP 2010: 157-168
- [RRK+13] Reidenberg, Joel; Russell, N. Cameron; Kovnot, Jordan; Norton, Thomas B.; Cloutier, Ryan; and Alvarado, Daniela, "Privacy and Cloud Computing in Public Schools" (2013). Center on Law and Information Policy. Book 2. http://ir.lawnet.fordham.edu/clip/2
- [Ros04] Rosen, L. (2004). Open source licensing. Prentice Hall PTR.; Laurent, A. M. S. (2008). Understanding open source and free software licensing. O'Reilly.
- [RSK+10] Indrajit Roy, Srinath T. V. Setty, Ann Kilzer, Vitaly Shmatikov, Emmett Witchel. Airavat: Security and Privacy for MapReduce. NSDI 2010: 297-312
- [Sin10] R. Singel. NetFlix cancels recommendation contest after privacy lawsuit. Wired.com ThreatLevel Blog, 12 March 2010. http://www.wired.com/threatlevel/2010/03/netflix-cancels-contest/.
- [Smi09] Adam Smith. 2009. Asymptotically Optimal and Private Statistical Estimation. In Proceedings of the 8th International Conference on Cryptology and Network Security (CANS '09), Juan A. Garay, Atsuko Miyaji, and Akira Otsuka (Eds.). Springer-Verlag, Berlin, Heidelberg, 53-57. DOI=10.1007/978-3-642-10433-6_4 http://dx.doi.org/10.1007/978-3-642-10433-6_4
- [Swe96] L. Sweeney. Replacing personally-identifying information in medical records: the scrub system. In Proceedings, Journal of the Medical Informatics Association, Washington, DC, 1996. Hanley Belfus.
- [Swe97a] L. Sweeney. Iterative profiler. Technical report, MIT, Cambridge, MA, 1997. Sealed by order of court in Southern Illinoisan v Department of Health.

- [Swe97b] L. Sweeney. Weaving technology and policy together to maintain confidentiality. In *Journal of Law*. Medicine and Ethics, volume 25, 1997.
- [Swe00] L. Sweeney. Uniqueness of Simple Demographics in the US Population. Technical report, Carnegie Mellon University, Data Privacy Lab, Pittsburgh, PA, 2000.
- [Swe02a] L. Sweeney. Achieving k-anonymity privacy protection using generalization and suppression. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 10(5):571–588, 2002.
- [Swe02b] L. Sweeney. k-anonymity: a model for protecting privacy. In International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, volume 10, 2002.
- [Swe03] L. Sweeney. Identifiability of de-identified pharmacy data. Technical report, Carnegie Mellon University, Data Privacy Lab, Pittsburgh, PA, 2003.
- [Swe09] L. Sweeney. Identifiability of de-identified clinical trial data. Technical report, Carnegie Mellon University, Data Privacy Lab, Pittsburgh, PA, 2009.
- [TUV12] J. Thaler, Ullman, J., and Vadhan, S. P., "Faster Algorithms for Privately Releasing Marginals", in Automata, Languages, and Programming - 39th International Colloquium, ICALP 2012, Warwick, UK, 2012, Lecture Notes in Computer Science., vol. 7391.
- [UV11] J. Ullman and Vadhan, S., "PCPs and the Hardness of Generating Synthetic Data", in Proceedings of the 8th IACR Theory of Cryptography Conference (TCC `11), Providence, RI, 2011, Lecture Notes on Computer Science., vol. 5978, pp. 572–587.
- [Uni04] United States Census Bureau, Census Confidentiality and Privacy, 1790-2002 (2004).
- [VAA+11] S. Vadhan, D. Abrams, M. Altman, C. Dwork, P. Kominers, S. D. Kominers, H. R. Lewis, T. Moran, G. Rothblum, and S. Vadhan. Comments on advance notice of proposed rulemaking: Human subjects research protections: Enhancing protections for research subjects and reducing burden, delay, and ambiguity for investigators, Docket ID number HHS-OPHS20110005. http://www.regulations.gov/#!documentDetail;D=HHS-OPHS-2011-0005-1101, October 2011.
- [WW01] Leon Willenborg and Ton de Waal, Elements of Statistical Disclosure Control, vol. 155, Lecture Notes in Statistics (New York: Springer Verlag, 2001).
- [Zim10] M. Zimmer. But the data is already public: on the ethics of research in facebook. Ethics and Information Technology, 12:313–325, 2010. 10.1007/s10676-010-9227-5.

Appendix 5: Information Products

A central goal of the project and its host institutions is to produce information products that are freely available to the research and informatics community. To facilitate this goal:

- All research articles will be made available under open access licenses
- All software produced by the project will be made available under an Open Software Initiative approved open software license

- All data used in publications will be formally cited, following best practices. (See https://www.force11.org/datacitation)
- All data collected by the project and necessary for replication of any publication will be made available through a public archive.

Publication Availability

The Massachusetts Institute of Technology and Harvard University, which hosts the project, is committed to disseminating the fruits of its research and scholarship as widely as possible and have adopted a broad open-access policies.[4] In keeping with that commitment, all scholarly articles produced by the project will be made available to the public under selected an open-access license.

Software Availability

Any license used will permit use, dissemination, and modification of the software for commercial and non-commercial purposes [Ros04]. We have selected the Apache v.2 software license as a default. However, compatibility with previous licenses used for core components, may require the use of a GPL v.3 license in some cases.

The Apache license is one of the most well-established licenses in use for open software development. It is a permissive license that is designed to be easily integrated with code under other licenses, including other open source licenses (such as MIT and GPL v. 3) as well as commercial products. Furthermore, unlike most other licenses, it contains explicit grant of patent rights where that is needed to operate, modify and distribute the software, thus eliminating a potential substantial barrier to commercial reuse.

[4] See https://libraries.mit.edu/scholarly/mit-open-access/open-access-at-mit/mit-open-access-policy, https://osc.hul.harvard.edu/policies

The software developed will be designed to be transferable, so that development can continue independently after the completion of the proposed project. Transferability is legally enabled through license of software itself and any required patents, using the Apache license above. However, effective transferability requires more than licensing -- in practice, for software to live beyond its original creators requires that it be developed using a community development process [Fog05]. The community development process has several key elements, which we will follow throughout the proposed work:

- Community development requires that the source code must be managed transparently in a well-known repository. This enables potential contributors to the project to examine the software, track development, and contribute directly. We plan to use the GitHub repository for all active development. While maintaining a local backup for disaster recovery purposes
- Community development requires that knowledge about the software codebase and development practices needs to be captured as documentation; and that documentation needs to be maintained and shared through a well-known public repository. We will capture all design decisions, specification, use cases, schemas, requirements and other software documentation, and make it available through the GitHub repository, to accompany the source tree itself.
- Community development requires that clear processes be established for reporting bugs, requesting features, tracking development milestones. We will use the standard GitHub repository features for these activities, to promote transparency and community engagement.

- Finally, community development requires that the development process itself be open and transparent. A basic part of this is conducting development in the open – such that discussions about development processes and about particular design and development decisions occur through public project email lists (or other similar public mechanisms) rather than through private channels. Effective community software development also requires establishing and communicating a "social contract" that clarifies expectations with respect to how to make contributions to the code; how contributions are reviewed, tested, documented and accepted; how to report issues and requests; how to engage in development discussions; how contributions will be attributed and credited; and how members of the community can join the project and earn decision making authority over areas of development.

Data Management

1. *Data to be collected.* Raw data collected directly by the entire project will comprise bibliographies of legal resources, and text corpuses of data management plans and related agreements
2. *Audience.* The primary audience for this data are researchers in two groups – disciplinary scientists aiming to verify or extend published findings, and methodologists and computer scientists aiming to evaluate new methods on real domain data.
3. *Access and Sharing.* All data refined data, replication data for publications, and raw data (with the exception of the raw calibration video) will be deposited in the Program for Informatics Data Archive. This archive is a part of the Harvard Dataverse Network (DVN). The IQSS DVN is a public repository, hosted at by the Harvard Libraries and, and backed up in perpetuity by the endowment of the Henry A. Murray Archive,

established in 1976. The Harvard DVN facilitates data access by providing descriptive and variable/question-level search; topical browsing; data extraction and reformatting; and on-line analysis. The Harvard DVN is open to all researchers for access and deposit, and MIT Libraries contract with Harvard for additional ongoing technical support and data access enhancements. All data will be deposited at least 90 days prior to the expiration of the award. Such data may be embargoed until the publication of research based on the data or until one year after the expiration of the award, whichever is sooner. Users will be required to agree to click-through terms that prohibit unlawful uses and intentional violations of privacy, and require attribution. Use of the data will be otherwise unrestricted and free of charge, under a well-recognized open data license (CC-BY-SA version 4).

4. *Formats*. Immediately after collection, any quantitative data produced will be converted to R and CSV formats. These formats are fully supported by the DVN, which will perform archival format migration; metadata extraction; and validity checks. Deposit in these formats will also enable on-line analysis; variable-level search; data extraction and reformatting; and other enhanced access capabilities. Documentation will be deposited in PDF/a, or plain-text formats, to ensure long-term accessibility.
5. *Documentation, Metadata and Bibliographic Information.* The project will create documentation detailing the sources, coding, and editing of all data, in sufficient detail to enable another researcher to replicate them from original sources; and descriptive metadata for each study including a title, author, abstract, descriptive keywords, and file descriptions. The project will include bibliographic information for any publication by the project based on that data. The Dataverse Network system's "templating" feature will

be used for consistency of information across studies. The DVN system automatically generates persistent identifiers, citations, and Universal Numeric Fingerprints for studies; extracts and indexes variable descriptions, missing-value codes and labels; creates variable-level summary statistics; and facilitates open distribution of meta-data with a variety of standard formats (DDI v 2.0, Dublin Core, and USMARC) and protocols (OAI-PMH and Z39.50).

6. *Storage, backup, replication, and versioning.* The IQSS DVN provides automatic version (revision) control over all deposited materials and no versions of deposited material are destroyed except where such destruction is legally required. All systems providing on-line storage for the DVN are contained in a physically secured facility that is continually monitored. System backups are made on a daily basis. Replicas of data are held by independent archives as part of the Data-PASS archival partnership, regularly updated, and regularly validated, using the LOCKSS system.
7. *Security.* The IQSS DVN complies with Harvard University requirements for good computer use practices. The University has developed extensive technical and administrative procedures to ensure consistent and systematic information security. "Good practice" requirements include system security requirements (e.g., idle session timeouts; disabling of generic accounts; inhibiting password guessing) operational requirements (e.g., breach reporting; patching; password complexity; logging); and regular auditing and review. The full University security policy can be found at http://security.harvard.edu.
8. *Budget.* The cost of preparing data and documentation will be borne by the project, and is already reflected in the personnel costs included in the current budget. The incremental

cost of permanent archiving activities will be borne by IQSS, supported through the Henry A. Murray Archive endowment.

9. *Privacy, Intellectual Property, Other Legal Requirements.* Information collected will not be encumbered with personal privacy restrictions. Information may be redacted to remove organizational identity. Data released by the project will not be encumbered with intellectual property rights (including copyright, database rights, license restrictions, trade secret, patent or trademark) by any party (including the investigators, investigators' institutions, and data providers.); nor is subject to any additional legal requirements. Depositing data with the IQSS DVN does not require a transfer of copyright, but instead grant permission for IQSS to re-disseminate the data and to transform the data as necessary for preservation and access.
10. *Archiving, Preservation, Long-term Access*. The IQSS DVN commits to good archival practice, including independent geo-spatially distributed replication, a succession plan for holdings, and regular content migration. Should the archiving entity be unable to perform, transfer agreements with the Data-PASS partnership ensure the continued preservation of the data by partner institutions. All data under this study will also be made available for replication by any party under the CC-attribution license, using the LOCKSS protocols – which is fully supported by the DVN system.
11. *Adherence.* Adherence to this plan will be checked at least ninety-days prior to the expiration of the award by the MIT P.I. Adherence checks will include review of the DVN content, number of studies released, availability for each study of subsettable/preservation friendly data formats (possibly embargoed, but listed); availability of documentation (public); and correctness of data citation, including UNF

integrity check.

Appendix 7: Institutional Background

This project is part of a collaboration between the Program for Information Science at MIT; the Center for Research in Computation and Society (CRCS) in the Harvard School of Engineering and Applied Sciences (SEAS); the Institute for Quantitative Social Science (IQSS), a university-wide center that is based in the Harvard Faculty of Arts & Sciences; and the Berkman Center for Internet & Society, a university-wide center that was founded at Harvard Law School.

Program for Information Science

The Program for Information Science (http://informatics.mit.edu), led by co-PI Micah Altman, seeks to solve emerging problems in information management that are essential to support new and innovative services, and to amplify the impact that MIT as a whole has on the development of information science, information policy, and scholarly communication through participation the development of standards, policy, and methods related to information science and information management.

Since 2012, under Dr. Altman's direction, the Program has published over fifteen scholarly publications; publicly archived several research data collections; released two major open source software packages; hosted three research interns; and garnered four research grants and prizes. This research program, under prior leadership lead to the development of the DSpace digital repository system, which is now used by hundreds of systems, and to the popular Simile semantic visualization tools.

The Program is a part of the MIT Libraries, which provides MIT students and researchers with access to over 3 million printed volumes and 55,000 databases and electronic journals. The

library system also has responsibility for the Institute Archives, MIT theses, and the DSpace@MIT institutional repository.The MIT libraries maintains membership for MIT in many leading professional Association of Research Libraries, Coalition for Networked information, Council on Library and Information Resources, Digital Preservation Network, DDI Alliance, Educause, HathiTrust, International Federation of Library Associations, National Digital Stewardship Alliance, NISO, OCLC Research Partners, ORCID, and Portico. Dr. Altman, as director of Research, serves as representative to the Digital Preservation Network, National Digital Stewardship Alliance, NISO, OCLC Research, and ORCID. Senior staff in the library play leadership roles in many of these organizations;

Dr. Altman also serves as chair of the National Digital Stewardship Alliance; serves on the boards of directors for ORCID and iSolon; on the executive board for the American Political Science Association's section on Information technology and politics; the steering committee for the Data-Preservation Alliance for Social Science; on the technical advisory boards of Force11 and The Qualitative Data Archive; and on the editorial boards of The American Journal of Political Science, Social Science Computer Review, The Journal of Information Technology and Politics and Statistical Associates Publishers.

Center for Research on Computation & Society (CRCS)

CRCS was founded in 2004 to drive innovative computer science research and technology towards problems of importance to society. It has done this by bringing Harvard's computer science faculty and students together with postdoctoral fellows and visiting scholars that have expertise in relevant areas, and hosting an interdisciplinary seminar series to enable interaction with social scientists, legal scholars, policy makers, and other computer scientists. In addition to numerous publications by faculty and fellows in first-tier computer science venues,

past contributions of CRCS include the Helios web-based open-audit voting system by CRCS fellow Ben Adida [3], a workshop on Data Surveillance led by CRCS fellow Simson Garfinkel that led to a special issue of IEEE Security & Privacy [30], and the development of courses on privacy, usable security, and cryptography co-taught by CRCS fellows. The current activities of CRCS are structured around three focus areas: Privacy & Security, Economics & Computer Science, and Health Care Informatics. PI Salil Vadhan served as the faculty director of CRCS from 2008-2011 and 2014, and has extensive experience in the foundations of cryptography and differential privacy. The computer science research efforts in our project are based in CRCS; in particular, Kobbi Nissim is a CRCS visiting scholar, Or Sheffet is a CRCS postdoctoral fellow, and Vishesh Karwa will be joining as a CRCS postdoctoral fellow in Spring 2014. Nissim is one of the founders of differential privacy, and his 2003 paper "Revealing Information while Preserving Privacy," which started that line of work, was recently awarded the Alberto O. Mendelzon Test of Time Award from the ACM Principles of Database Systems (PODS) conference. Karwa and Sheffet both have expertise and strong publication records in differential privacy, with Karwa coming from a statistics background and Sheffet from a computer science background.

Institute for Quantitative Social Science (IQSS)

IQSS was founded in 2005 by co-PI Gary King as a university-wide institute with a dual scientific mission [King 2014]. First, IQSS catalyzes research to understand and solve major problems that affect society and the well-being of human populations, by bringing together diverse researchers and approaches from multiple disciplines. Second, IQSS develops analytical tools for social and health sciences, focusing on open collaborative tools for computational social science, statistical analysis, and data sharing and preservation. One of IQSS's initiatives is the

Dataverse project – which is open source, web 2.0 software for data sharing, preservation, citation, and analysis now hosted at universities around the world [King 2007, Crosas 2014]. Each Dataverse distributes virtual archives (called "dataverses") to hundreds of researchers and institutions; each dataverse provides all the services of a professional archive on your web site, and with your branding, but without having to install anything locally. Harvard hosts an instance of the Dataverse software which now comprises the largest catalogue of social science research data in the world. Dataverse software also integrates with Zelig, a framework that allows a large body of different statistical models in the R statistical language to be used from a unified call structure [Imai King Lau 2008]. It is also a modeling architecture that interprets these statistical models in a substantively meaningful fashion [King Tomz Wittenberg 2000].

Co-PI Gary King is the Albert J. Weatherhead III University Professor at Harvard University, and founder and Director of the Institute for Quantitative Social Science. He is an original author of the Zelig statistical package, and the PI of the Dataverse platform. His research spans quantitative methodology and computational methods across the social sciences in more than 150 journal articles, 8 books, and more than 20 research software packages.

Co-PI Mercè Crosas is Director of Data Science at IQSS. She has led the design, architecture and implementation of the Dataverse since the project started 6 years ago, and has given multiple talks and training sessions on data sharing, analysis, management and preservation. She has contributed to defining and implementing policies for the management and dissemination of public and confidential research data for IQSS and the Dataverse [Crosas 2013]. In 2011, Latanya Sweeney's Data Privacy Lab moved to IQSS from Carnegie Mellon University, reflecting IQSS' strong interest in taking on privacy-sensitive datasets.

Senior personnel James Honaker is Senior Research Scientist in the Data Science Group at the Institute for Quantitative Social Science (IQSS). He works on quantitative methods for problems in social science data, and received his PhD from Harvard in 2004. He is an author of the TwoRavens statistical interface and the most recent version 5 of Zelig, as well as other statistical research software such as the Amelia software package for missing data, and won the 2014 Statistical Software Award of the Society for Political Methodology for statistical software that makes a significant research contribution. Prior to IQSS, James taught on the faculties of UCLA and Penn State.

Berkman Center for Internet & Society

The Berkman Center for Internet & Society at Harvard University is a research program founded to explore cyberspace, share in its study, and help pioneer its development. Founded in 1997, through a generous gift from Jack N. and Lillian R. Berkman, the Center is home to an ever-growing community of faculty, fellows, staff, and affiliates working on projects that span the broad range of intersections between cyberspace, technology, and society. Led by a diverse group of faculty directors from many Harvard schools, including Executive Director and Harvard Law School Professor of Practice, co-PI Dr. Urs Gasser, the Berkman Center has an established track record of conducting research on a variety of legal topics related to technology and the law, including privacy, governance, intellectual property, cybersecurity, antitrust, content control, and electronic commerce as well as other dimensions of emerging technological trends and related societal shifts. The Berkman Center's Cyberlaw Clinic, a program that provides innovative, hands-on training and course credit to Harvard Law students who, under careful supervision, offer legal and policy research, guidance, and representation to a variety of real-world clients, including research projects and institutional entities. The Clinic regularly engages with issues

related to information privacy and policy. Beyond the Clinic, the Berkman Center has a number of privacy-related initiatives, including the Berkman Center's Youth and Media project and Student Privacy Initiative, which have empirically and qualitatively explored the dimensions of privacy in the context of youth online and in new classroom technologies. Related to these efforts, the Berkman Center has published a number of papers, reports, best practice guidelines, and other publications that explicate for practitioners the nuances of policy and law as well as the real-world practices that affect privacy. The Berkman Center has also hosted a series of workshops, events, and lectures on privacy related issues. In addition to the Berkman Center's research contributions to the field, privacy is also a focus in the classroom. For example, co-PI Gasser teaches Comparative Online Privacy, which is offered annually at Harvard Law School.

In Fall 2009, under the leadership of co-PIs Gasser and Vadhan, Berkman and CRCS formally joined their fellowship programs to enable the greater interdisciplinary interaction sought by both sides. Berkman and CRCS faculty and fellows have shared a weekly "Fellows Hour," a weekly discussion seminar, and a biweekly technical seminar series, and interdisciplinary research group meetings on a number of topics (especially privacy). Our Privacy Tools project emerged from these collaborative activities, along with the shared experience of the three centers (CRCS, Berkman, IQSS) in dealing with the privacy issues surrounding a particular Facebook dataset [52].

Co-PI Dr. Urs Gasser is the Executive Director of the Berkman Center and a Professor of Practice at Harvard Law School. He is a visiting professor at the University of St. Gallen (Switzerland) and at KEIO University (Japan), and he teaches at Fudan University School of Management (China). Urs Gasser serves as a trustee on the board of the NEXA Center for Internet & Society at the University of Torino and on the board of the Research Center for

Information Law at the University of St. Gallen, and is a member of the International Advisory Board of the Alexander von Humboldt Institute for Internet and Society in Berlin. He is a Fellow at the Gruter Institute for Law and Behavioral Research.

Professor Gasser's research and teaching activities – which focus on technology and information law, policy, and societal issues, and place particular emphasis on privacy law and policy – and his substantive expertise make him well-suited for the role of co-PI on this project. Throughout his research, Professor Gasser has closely examined privacy both in the US and internationally in a number of contexts that are relevant to the focus of the Privacy Tools project, including, among others, children and use of technology, cybersecurity, consumer protection, cloud computing, and electronic health records. At various levels, he has consulted with and provided guidance to government officials, policymakers, and private industry, and written numerous articles and books, and frequently comments in the media on these and related topics.

Professor Gasser has an established history of collaboratively working with leading organizations and computer scientists, technologists, and sociologists – including members of the current Privacy Tools project team, Salil Vadhan and Latanya Sweeney – and routinely participates interdisciplinary research and problem solving efforts. As an experienced a lawyer, he also has a deep understanding of the practical requirements needed to develop the specialized legal instruments that are needed to support and complement the efforts of the computer scientists, statisticians, and social scientists working on the project.

David O'Brien is a senior manager at the Berkman Center for Internet & Society. He has contributed legal research to and led the Berkman Center's efforts across a variety of projects, publications, and initiatives, spanning the topics of privacy, intellectual property, cloud computing, cybersecurity, digital publishing, and internet governance. Under the direction of

Berkman Executive Director, Urs Gasser, David currently leads the Berkman Center's research contributions to the Privacy Tools for Sharing Research Data project. David holds a J.D. from Northeastern University's School of Law, and was admitted to the Massachusetts bar in 2009.

_____Alexandra Wood is a research fellow at the Berkman Center for Internet & Society and a member of the Privacy Tools for Sharing Research Data legal team. Her research explores new and existing legal, regulatory, and contractual approaches to data privacy and contributes to the legal research, analysis, documentation, and conceptual framework supporting the development of the DataTags questionnaire, tagging architecture, and modular license generator tools. Before joining the Berkman Center, she served as a legal fellow with U.S. Senator Barbara Boxer and as a law clerk with the Center for Democracy & Technology and the Electronic Privacy Information Center.

Appendix 8: Other Synergistic Activities

- We plan to continue our extensive collaboration with Cynthia Dwork (Distinguished Scientist, Microsoft Research). Dwork is one of the founders of differential privacy and has collaborated with co-PI Vadhan extensively during the past six years [DNR+09, DNV12, DRV10, VAA+11], and is co-PI on a Sloan project "Towards Practicing Privacy." The continuation of this collaboration will be facilitated by our incoming postdoctoral fellow Vishesh Karwa, who is currently doing an internship with Dwork, working on her Sloan project on education data (one of our selected use cases).
- The National Science Foundation has recently awarded Co-PI Altman a collaborative grant with EdGE at TERC, Landmark College, and MIT, *Revealing the Invisible (RtI),* which brings together expertise in learning sciences, cognitive psychology, and data sciences to advance core knowledge about how big data, enhanced with biometric

information, can aid in the study of learning. The goals of this exploratory research are to understand how the data exhaust from online educational interventions, using digital games as a model, can be used to customize optimal learning experiences, and, more broadly, to evaluate how data exhaust can reveal basic cognitive activities that are prerequisites for learning. As part of this research Dr. Altman will be designing embeddable open-source instruments to collect attentional data from existing embedded cameras in laptops, phones and tablets, and to manage the privacy protection of that data stream in combination with the existing big-data exhaust stream. *RtI* will serve as a concrete example in the use-case analysis within the proposed project, and the results of the proposed project will inform development of the *RtI* instrumentation.

- The National Digital Stewardship Alliance (NDSA) is a consortium of institutions that are committed to the long-term preservation of digital information. The NDSA comprises over 160 participating institutional members, in collaboration with the Library of Congress, and collectively manages over a hundred petabytes of content. Dr. Altman is currently the Chair of the alliance, and a lead author of the *National Agenda for Digital Stewardship*, which summarizes priorities and opportunities for the alliance and for long-term data access. Managing privacy issues related to data has emerged as a priority the alliance, and is explicitly highlighted in the *Agenda*. Selected *NDSA* content will serve as a concrete example in the use-case analysis within the proposed project, and the results of the proposed project will inform recommendations for "best practices" developed and disseminated through NDSA.
- Dr. Altman has established ongoing voluntary research collaborations around data management and information privacy with other MIT units, including IMES, CSAIL, and

the MIT Media Lab. These collaborations have contributed to the development of the White House/OSTP workshop on privacy and big data, to review of the policy commentaries mentioned above, and to the development of a range of potential use-cases for applications of privacy tools. The proposed project will provide opportunities for joint collaboration with postdocs at CSAIL and the Media Lab to develop lifecycle and policy analyses related to use cases emerging from these units.

- The Program for Information Science has established operational collaborations to develop services and policies with the Office of Institutional Research (which is responsible for dissemination of MITx data), the Office of General Council (which is responsible for developing privacy policies across MIT), and the Office of Digital Learning (which is responsible for MITx operations). The proposed project will provide opportunities to collaborate with these offices to test and refine the application of the informatics, legal, and computational tools, in practical setting.
- The Berkman Center has a number of ongoing initiatives and projects related to connected learning, education technology, and privacy that we hope to leverage opportunities to share knowledge and collaborate with the Privacy Tools project between these projects, particularly along the lines of expanding our education technology and privacy research. For example, Co-PI Urs Gasser serves on the Harvard Vice Provost's Committee on Advances in Learning, chaired by Peter Bol, that contributes directly to the HarvardX Initiative, which is Harvard's collaboration with EdX. Other researchers in residence at the Berkman Center also work closely with these efforts. For instance, Justin Reich, a fellow at the Berkman Center who works on our CopyrightX program, is also the Richard L. Menschel Research Fellow at HarvardX. In this role, he leads and

conducts research on massive open online courses (MOOCs) and other HarvardX ventures, many of which are deployed and generate data on the EdX platform. Additionally, CopyrightX, a networked course taught by Berkman Center faculty director Terry Fisher and supported by the Berkman team is affiliated with HarvardX. Other ongoing Berkman Center projects are related more generally to education technology, connected learning, and privacy. This includes the Berkman Center's Student Privacy Initiative, which explores the practical and legal implications of using new technologies in the K-12 classroom, as well as a recently announced NSF-funded Cyber Learning project.

- The IQSS founder and director (and our collaborator) Gary King is on the HarvardX faculty committee, and CRCS faculty member Jim Waldo (and Harvard University CTO) is on the HarvardX support committee, providing us with additional access to the goals and challenges of working with and sharing HarvardX data (as well as to the data itself, which is being stored at IQSS).

Appendix 9: Supplementary Details on Our Approach and Prior Work (Section 2)

A9.1 Differential privacy: mathematical theory and practical tools. Recall that a differentially private interface carefully injects a small amount of "random noise" into each query evaluation (in sophisticated ways) so as to allow global statistical information to be computed accurately while ensuring that information specific to individuals cannot leak. Formally, differential privacy requires that what an adversary sees when interacting the system on a given dataset is essentially the same -- up to a (1+epsilon) multiplicative factor in the probability distribution - as it would be if we removed any individual's data from the dataset. This means that, regardless of what auxiliary information an adversary has (e.g., publicly

available datasets, detailed knowledge about the individual being targeted), it still cannot extract information specific to that individual from the differentially private system.

There is now a rich body of theoretical work on differential privacy, with some striking results showing that very general classes of data analysis tasks are compatible with the strong privacy protections of differential privacy, and, in many cases, the differentially private algorithms have the same asymptotic performance as the best non-private algorithms as n, the number of individuals in the dataset, tends to infinity.

Our tools will make differentially private statistics available through TwoRavens, which is a graphical user interface for quantitative analysis, (illustrated in *Figure C*) constructed by our team members, that allows users at all levels of statistical expertise to explore their data, and appropriately construct and interpret statistical models [HD14]. The interface is a browser-based, thin client, with the data remaining in an online repository, and the statistical modeling occurring on a remote server. This architecture, where both the data and the analysis on the data remains remote from the user interface, can be an enforceable way to grant access to differentially private statistics, without permitting access to confidential raw data.

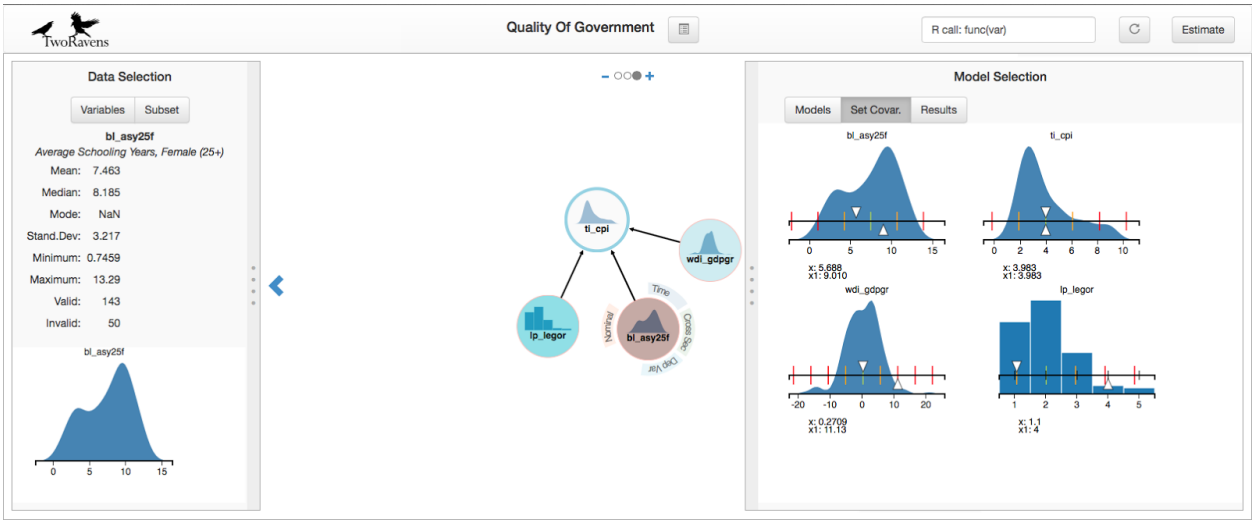

Figure C: Example of TwoRavens interface with data on Dataverse, showing exploratory statistics (left), construction of a statistical model (center), and interpretation of estimated results (right).

There are a number of other ongoing efforts to bring differential privacy to practice, including lots of experimental work and optimization of the performance of differentially private algorithms for specific applications (e.g. [MKA+08,OBB+12]), the design of general-purpose programming tools for differential privacy (e.g. [McS09,RP10,GHH+13]), and the incorporation of differential privacy into larger computer systems for privacy and security (e.g. [RSK+10,RAW10,MTS+12]). Recall that what distinguishes our effort is the focus on incorporating differential privacy into the infrastructure of a data repository (Dataverse) in way that fits directly into the workflow of researchers sharing data, and does so in a way that is optimized for the wide variety of datasets, users, and analyses that a general data repository needs to handle. While we cannot hope to match the performance of algorithms tuned to a very specific type of data and specific type of analysis, our goal is to enable preliminary analysis, whereby a user can determine whether it is worth applying for access to the raw data (a more involved process that will involve a data use agreement and possibly IRB approval). In addition, it is important to note that we are using differential privacy to provide access where it is currently unavailable, rather than using it to constrain access that researchers currently have. We believe that this positioning makes our tools more likely to be adopted, and provides a low-risk way to start assessing the practical performance of differential privacy.

A9.2 Legal research and analysis. Over the past year, the legal team has been focusing on a literature review analyzing US federal and state statutory and regulatory provisions, legal scholarship, and common contractual approaches relevant to the access, use, and disclosure of medical, education, and government records. To this end, the Berkman team has been

systematically identifying and cataloging federal and state laws and regulations relevant to the sharing of research data and drafting legal memoranda analyzing the provisions that affect researchers' and data repositories' access, use, and redisclosure of personal information protected by each of these laws. This research has informed a series of reports, including forthcoming publications comparing various statutory and regulatory definitions of and approaches to privacy. The team has also begun work to package these privacy law resources to become part of a publicly-available toolkit.

A primary research focus has been on contractual approaches to research data sharing. The Berkman team obtained from Harvard offices and data repositories that handle research data, as well as from data repositories hosted at other institutions, more than one hundred data use agreements, memoranda of understanding, and policies used by data repositories, data enclaves, research studies, academic institutions, federal and state government agencies, nonprofit organizations, and businesses. The team then analyzed each contract by mapping every contractual term within it to general categories, such as provisions describing the contents and sensitivity of the data; the restrictions on access, use, and disclosure; the data provider's rights and responsibilities; the data confidentiality, security, and retention procedures to be followed; the assignment of liability between the parties; and enforcement and penalties. Within each of these general categories of provisions, the team identified alternative approaches that researchers, repositories, and others have adopted to address potential issues. This work serves as a foundation for research on both traditional and new integrated approaches to sharing research data.

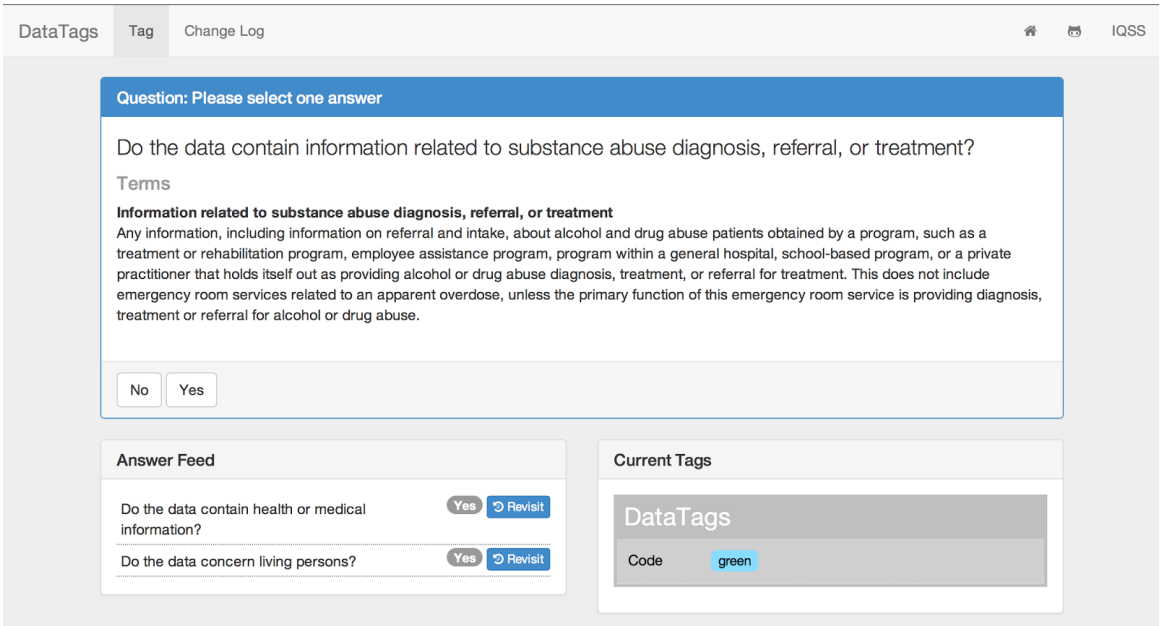

Figure D: Example of the user interface for the DataTags tools, towards the beginning of a user interview concerning medical subject research data.

The prototype version of the DataTags system is illustrated in *Figure D*. It handles the deposit of datasets containing medical, education, and government records protected under many of the key federal data privacy laws and regulations in these areas. The supported laws and regulations include the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule, the Family Educational Rights and Privacy Act of 1974 (FERPA), the Protection of Pupil Rights Amendment (PPRA), the Privacy Act of 1974, and the Confidential Information Protection and Statistical Efficiency Act of 2002 (CIPSEA). The team is beginning initial usability testing of the prototype DataTags tools by selected groups of stakeholders, including institutional review board members and data repository administrators. Efforts are underway to add support for other areas of law and to build a modular license agreement generator that will create custom license agreements to govern subsequent use and sharing of the dataset consistent with the relevant privacy interests and restrictions.

Appendix 10: Supplementary Details on Research Directions (Section 3)

A10.1 Incentives and differential privacy. Ghosh and Roth [GR11] considered a model where a data analyst elicits 'data subjects' valuations for privacy, pays some of the subjects for their data, and then carries out a statistical computation on the result using a differentially private algorithm A. In their model, the subjects each experience a privacy loss equal to the product of their individual valuation for privacy and the level of differential privacy provided to them by the algorithm A. They seek mechanisms that are truthful (subjects are incentivized to reveal their true valuations for privacy), individually rational (subjects don't experience a net loss in utility), make finite payments, and provide accurate results.

Under this model, they provided a negative result for the case where agents' valuations for privacy (not only their "data") are considered sensitive (e.g., because caring a lot about privacy might be correlated with having something to hide), and positive results for the case where agents' valuations for privacy are not considered sensitive.

In recent work, visiting scholar Nissim and co-PI Vadhan have extended the Ghosh-Roth work in several ways. One critique of the Ghosh-Roth model, raised in [NST12], is the use of the differential privacy parameter epsilon as an *exact* characterization of the privacy loss experienced by subjects. This is unrealistic, because the actual privacy loss experienced can depend on other things, such as the adversary's background knowledge, the inputs of other data subjects to the mechanism, and the particular output produced by the mechanism. Thus, in [NOS12, CCK+13], we provide alternative models, where the differential privacy parameter is only an upper bound on the privacy loss, and we allow the actual privacy loss to depend on other variables. We also show that one can obtain interesting positive results using this model (mostly for problems of interest in economics, such as social choice problems). In a new work [NVX14], we substantially strengthen the Ghosh-Roth negative result to hold for a much wider class of

privacy valuation functions (not just differential privacy), and show that the negative result can be bypassed if we assume the relation between one's valuation for privacy and one's data value is monotone.

A10.2 Incentives and a Lifecycle Analysis

Historically, control over the privacy of data in social science has been based on one of two models. Where the data publisher has few resources or little expertise, ad-hoc deidentification of data is often used, typically by applying suppression of extreme values and generalization of measurements. In cases where the data publisher has relatively substantial resources and expertise, the official statistics model—in which the data are disseminated primarily as pre-determined summary statistics—is typical. The primary disclosure threats considered in this model are identification of an individual or inappropriate integration of data across multiple government organizations; and traditional statistical disclosure limitations are employed for protection [See for example, HDF10; WW01; Uni04].

In contract, in the privacy use cases being considered in this proposal, disclosure threats, methods, products and many other characteristics vary widely. Building on our own research and that of others we have developed an expanded taxonomy for privacy use cases.[5] A prototype chart of use case features, presented below in Table 1, illustrates a portion of the features that will be used to characterize privacy use cases. This typology of use case characteristics is intended to summarize their key features of research data management that are potentially relevant to stakeholders, data management, and the applications of privacy methods. *Table 1. Example Use Case Features*

[5] *See, e.g.,* [WW01]; [HDF+10]; [FWC+10]; [Alt12].

Families of Features	*Specific Characteristics*
Data characteristics	● Logical Structure (e.g., single relation, multiple relational, network/graph, semi-structured, geospatial, aggregate table) ● Source ● Unit of observation ● Attribute measurement type (e.g., continuous/discrete; ratio/interval/ordinal/nominal scale; associated schema/ontology) ● Performance characteristics (e.g., dimensionality/number of measures, number of observation/volume, sparseness, heterogeneity/variety, frequency of updates/velocity) ● Quality characteristics (e.g., measurement error, metadata, completeness, total error)
Disclosure scenarios	● Source of threat (e.g., natural, unintentional, intentional) ● Areas of vulnerability (e.g., data, software, logistical, physical, social engineering) ● Attacker objectives, background knowledge, and capability (e.g., "nosy neighbor," "business competitor," "muckraking journalist," "panopticon," "intrusive employer/insurer") ● Breach criteria/disclosure concept
Legal/institutional context of data collection	● Consent (e.g., open consent, active but limited consent, passive/implicit consent, awareness of data collection, unawareness of data collection, surreptitious data collection) ● Jurisdiction where collection takes place ● Special legal relationship with subject (e.g., student relationship under FERPA, patient relationship under HIPAA) ● Status of individual/institution responsible for data collection (e.g., a HIPAA regulated entity, an entity subject to the Common Rule, 45 C.F.R. part 46)
Information lifecycle stages	● Lifecycle stages managed/in scope (e.g., data creation/experimental intervention, data collection/initial

	transmission, data storage/ingest/entry into research environment, processing, internal sharing and collaboration, analysis dissemination/publication, verification/scientometric/educational/ scientific reuse, long-term access) ● Information management policies
Analytic results	● Form of output (e.g., summary scalars, summary table, model parameters, data extract, static data publication, static visualization, dynamic visualization, statistical/model diagnostics) ● Analysis methodology (e.g., contingency tables/counting queries, summary statistics/function estimation, regression models/GLM, general model-based statistical estimation/MLE/MCMC, bootstraps/randomization/data partitioning, data mining/heuristics/custom algorithms) ● Analysis goal (e.g., rule-based, theory formation, existence proof, verification, descriptive inference, forecasting, causal inference, mechanistic inference) ● Utility/loss/quality measure (e.g., entropy, mean squared error, realism, validity of descriptive/predictive/causal statistical inference)
Characteristics of data related to informational harm	● Attribute identifiability characteristics (e.g., direct identifiers, quasi-identifiers/publicly observable fixed characteristics of individuals) ● Attribute sensitivity (e.g., descriptions of criminal conduct, health status, income, political affiliations) ● Statistical identifiability risks (e.g., reidentification/record-linkage risks, information/learning risks) ● Expected types of harms from reidentification (e.g., loss of insurability, loss of employability, market discrimination, criminal liability, psychological harm, loss of reputation, emotional harm, and loss of dignity (dignitary harm); social harms to a vulnerable group (e.g., stereotyping), price discrimination against vulnerable

	groups, market failures; chilling of speech and action; potential for political discrimination; potential blackmail and other abuses) ● Expected magnitude of harm, if identification occurs (e.g., minimal, moderate, severe)
Lifecycle stakeholders	● Stakeholder types (e.g., consumer, producer, funder, host institution, researcher, regulator, subject, citizen, journal) ● Stakeholder capacities/resources (e.g., technical expertise, infrastructural capacity, budget, staffing resources) ● Trust relationships
Current approaches	● Regulations/policy ● Legal controls ● Statistical/computational disclosure control methodology ● Information security controls

A full lifecycle description would trace multiple scenarios within each broad use case. And a full model would examine these features at each lifecycle stage, along with the specific relevant characteristics of actors at that stage, actions, and information objects.

A10.3 Massive data. Secure storage of very-large-scale data, including continuously updated or streaming data, is beyond the current architecture of Dataverse. Similarly, while R has exploded in usage among statisticians and quantitative researchers, due to its open contribution structure, interpreted language implementation, and simplified memory management, that same decentralization and ease of development often leads to computational approaches that scales poorly for large problems. Moreover, our current tools for incorporating differential privacy into Dataverse has focused on regression-styled statistical modeling of causal effects, while big data exploration often heavily relies on machine learning methods for (unsupervised) clustering and (supervised) prediction or forecasting.

We plan to take the architecture we have developed for Zelig in R, and blueprint how that same architecture could be mirrored in Scala and Java for increased computational ability in distributed settings. A growing number of statistics and analytics tools have been written for analysis of big data on distributed systems in these languages (such as Apache Mahout, Weka, MALLET) including clustering algorithms such as *k*-means and latent class analysis, and supervised learning such as naive bayes and tree-based algorithms. Clustering and supervised/semi-supervised learning under differential privacy was studied theoretically by visiting scholar Kobbi Nissim [BDMN05, NRS07, KLN+11, FFKN09, BNS13a, BNS13b, BNS14], and we plan to continue this investigation of the possibility and limitations of differentially private adaptations of clustering algorithms and aggregation techniques, as well as the ramifications of ensembles of differentially private models for forecasting. The k-means algorithm from [BDMN05] was implemented in PINQ [McS09] and AIRAVAT [RSK+10]. This algorithm uses sum queries answered with Gaussian Noise. We will examine the potential of incorporating in practical implementations more advanced ideas and techniques that emerged in the theoretical studies since [BDMN05]. This will facilitate easier access to big data tools and better interpretation to large-scale data models, as well as access, transparency, reproducibility and validation of scientific results in private big data settings.

A10.4 Use Case: Online education data. MOOCs and other new educational technologies offer great possibilities for education research. For instance, relationships between student patterns of engagement with course materials, performance on assessments, and pedagogical interventions can be tracked and studied over time, often with a very large sample size. The multi-institutional *Asilomar Convention*[6] affirmed that "maximizing the benefits of

6 *See* http://asilomar-highered.info.

learning research requires the sharing of data, discovery, and technology." However, the sharing of educational data for research raises significant privacy concerns, and is highly constrained by statutes and regulations, such as the Family Educational Rights and Privacy Act (FERPA) and related student privacy laws at the state level. Attempting to balance data access and privacy concerns, MIT and Harvard recently released a de-identified version of the data from their MOOC platforms MITx and HarvardX, through the Harvard Dataverse.[7] The dataset went through an elaborate, labor-intensive deidentification process, yet it was still admitted that "there is always a risk that the data will be re-identified."

In our project, we will examine the applicability of the tools we are developing to the online education use case. As discussed above, our work on implementations of differential privacy tools is not optimized for data from specific domains such as education data, but rather is aimed at providing general-purpose tools that can handle the wide variety of datasets that are deposited in repositories such as Dataverse. At the same time, it can be very informative to explore how much utility these general tools offer in the context of specific real-world use cases, and online education data is a special case that raises a number of legal and technical issues related to privacy while also attracting a significant amount of interest from researchers.

We will engage in a variety of activities to study the application of privacy-preserving tools to online education datasets. For instance, personnel from the Berkman Center will explore the laws, policies, and agreements relevant to the collection, use, and disclosure of datasets from online education programs—such as FERPA, legislative developments at the state level, Department of Education guidance materials, and institutional policies and agreements—and engage with the relevant stakeholders working in this space. We will also conduct usability

[7] Harvard Dataverse, "HarvardX-MITx Person-Course Academic Year 2013 De-Identified dataset, version 2.0," http://thedata.harvard.edu/dvn/dv/mxhx.

testing of our tools and outreach, such as hosting workshops, working meetings, or focus groups, with appropriate stakeholders. Potential participants and collaborators include researchers collecting and analyzing online education datasets as well as those affiliated with HarvardX, MITx, and related projects mentioned above in **Section 1.3** and in **Appendix 7**. Such an investigation will demonstrate concrete applications for the use of integrated privacy-preserving tools, provide a baseline for developing systems optimized for the online education domain, and point the way to future extensions of our tools.

A10.5 Use Case: Common Rule data. Most social science research is governed by the Common Rule, 45 C.F.R. part 46, which is aimed at protecting human subjects in federally funded research. The Common Rule includes requirements for obtaining the informed consent of participants, and for the operation of institutional review boards (IRBs), which are tasked with determining whether research studies at a given institution are ethical, and in particular assessing the level of risk posed to the subjects. With the shift towards data-driven, computational social science research, IRBs are increasingly being forced to evaluate informational risks, or those that arise from inappropriate use or disclosure of information (i.e.g., failure of data privacy or security). Positing that most current IRBs are ill-equipped to evaluate informational risks, the Department of Health and Human Services proposed in 2011 to adopt the HIPAA de-identification standard in the Common Rule (i.e., the removal of certain fields would render a dataset presumptively de-identified and thus free of informational risks). In addition to providing a simple rule that is easy for researchers and IRBs to apply, the HHS proposal has the potential advantage of harmonizing the Common Rule and HIPAA, both of which cover research done with data that are gathered in a clinical setting.

Personnel from the Berkman Center are studying how the integrated privacy-preserving tools being developed apply to the collection, use, and sharing of data regulated by the human subjects research protection laws. The Berkman team has conducted legal research to identify and analyze the relevant laws, policies, and agreements relevant, and will expand their current analysis of human subjects research data by conducting a review of institutional policies and consent agreements. Personnel will obtain collections of institutional review board policies and consent forms for human subjects research studies from university sources, such as the data repositories at Harvard and MIT. The team will then conduct a type analysis to cluster these provisions from these policies and agreements in order to identify common typologies in approaches to collecting, using, and sharing these categories of data. These efforts will enhance the development of appropriate legal instruments and practices for sharing data in a wide range of contexts.

A10.6 Use Case: Economic data. The proprietary nature of many new sources of economic data poses a significant barrier to advances in computational economic research. For this reason, we propose to explore a third data privacy use case involving economic data disclosed by businesses and protected by nondisclosure agreements (NDAs). This use case contemplates datasets such as online auction data from eBay and pricing data from Amazon that provide a rich source of information for research in the fields of microeconomics, game theory, and auction theory. We will begin a review of the common approaches to obtaining and analyzing commercial economic data, and the incentives and barriers to the disclosure of such data for research purposes. To conduct this review, we will conduct outreach, such as by holding workshops, working meetings, or focus groups, with industry stakeholders and, where possible,

obtain and analyze collections of NDAs for economic data from university sources, such as the Offices of Sponsored Programs at Harvard and MIT.

For this analysis, we aim to compare the attributes of the stakeholder approaches, including NDAs, that emerge during this review to our existing taxonomies and tag frameworks, and to existing literature on current and proposed practices in economics exemplified by the OKF Open Economics Principles[8] (the development of which was contributed to by PIs Altman and Crosas), the Open Science Framework badges definitions,[9] and emerging replication policies in economics journals.[10] These efforts will enhance the development of appropriate technical and legal approaches for sharing and replicating data from a range of commercial sources, and inform the development of metadata schemas, legal agreements, "good practices," and draft regulatory language in the project.

Appendix 11: Budget Narrative – Detailed Description by Institution

The effort funded by the budget is allocated to the following activities (in roughly equivalent share):

- Explore how new computational and legal privacy tools we are developing can be applied or extended to handle massive data used in computational social science and education research.
- Develop institutional and stakeholder analyses for managing research data privacy and systemic policy consequences of applying new computational and legal privacy concepts and tools.

[8] http://openeconomics.net/principles/
[9] http://centerforopenscience.org/journals/
[10] http://openeconomics.net/resources/data-policies-of-economic-journals/

- Test the Dataverse System in support of selected privacy use cases, and design an architecture for secure privacy protecting large-scale archival data.
- Expand research collaborations to engage with data privacy research at MIT, other experts on differential privacy, the edX platform for MOOCs, and several related Sloan projects.
- Extend new legal and computational tools, and create model frameworks to manage private data restricted by nondisclosure agreements

  Detailed institutional budget narratives are listed below.

Berkman Center for Internet & Society

Co-PI Gasser will develop and oversee specific research objectives and activities for the legal research team, and contribute to dissemination of results; this effort is not part of the budget request. The budget requested will be used to support key personnel, student research assistance, small events and workshops, and personnel travel. The budget will support project manager O'Brien, who will devote 25% to ongoing management of the project and contribute to legal research and analysis, and dissemination of results. The requested budget will also support 100% research fellow Wood's time on the project in years 2 and 3, who will contribute to legal research and analysis, and dissemination of results. The budget supports research assistance from law school students during the academic year semesters and summer. This enables the Berkman Center to utilize 1-2 research assistants (10 hours per week) each fall and spring academic semester, and a total of 4 interns during the summer periods of the project (35 hours per week). In addition to salaries, the requested budget includes support for 1-2 small events of approximately 20 participants, hosted locally, and anticipating 1-3 individuals who require support for travel to the event. If necessary, these events would be used to convene critical

stakeholders and representatives from the private sector or other institutions, such as institutional review boards, to share expertise and consult on the legal and technical practices associated with certain types of data (e.g., economic data and NDAs), institutional policies (e.g., IRB policies at other institutions), and the like. Finally, the requested budget would enable nominal travel for 1-2 key personnel per year to travel to conduct project outreach and presentations.

Institute for Quantitative Social Science

Co-PI Gary King, Director of the Institute for Quantitative Social Science (IQSS), will provide advice and guidance on the direction of the project, and feedback during the development phase and once milestones are achieved. Professor King is not requesting salary support from this grant application.

Mercè Crosas is Director of Data Science at IQSS. In this role she oversees all aspects of development and architecture of the Dataverse Network application, as well as directing Data acquisition, curation services, and development of software tools. She will contribute one half-month per year overseeing development of Dataverse, and providing guidance and management for expansion of Dataverse to secure storage of large-scale data. James Honaker is Senior Research Scientist and lead of the Statistics and Analytics Group in the Data Science Program of IQSS. He is a lead on both the Zelig and TwoRavens statistical software projects. He will contribute one half-month per year, researching differentially private adaptations of machine learning algorithms, blueprinting the expansion of the Zelig architecture to large-scale data in the Java and Scala domains, and adapting TwoRavens to the data use cases for differentially private statistical exploration. Three months of time, per year, is requested to partially fund time of an academic Post-Doctoral scholar, who will research algorithms for differentially private machine learning, and modifying the TwoRavens statistical interface to best work in the context of the

explored data use cases. Similarly, three months of time, per year, is requested to partially fund time of a staff software developer, who will develop the architecture for secure large-scale data storage.

Program for Information Science

PI Altman will provide scientific guidance, supervise the project postdoc, and contribute to dissemination of results; this effort is not part of the budget request. The requested budget will support one postdoctoral fellow, to be hired, for two years, aided by 10 hours/week undergraduate research assistance. Nominal assistance for postdoc conference travel, and computing support is also requested.

Center for Research on Computation & Society

Co-PI Vadhan will provide leadership and take responsibility for our computer science research efforts. The year 1 budget is devoted to the support of visiting scholar Kobbi Nissim (of Ben-Gurion University), who is one of the founders of differential privacy and has been a part of the senior leadership of our Privacy Tools project for the past two years. When supported by the proposed grant in 2015, he will supervise graduate and undergraduate students on research related to the project (e.g. coming out of a graduate course on differential privacy that he will teach with postdoctoral fellow Or Sheffet during Fall 2014), and continue to collaborate with PI Vadhan on shaping all aspects of our work in differential privacy, including economic analyses of differential privacy. The year 2 budget includes a small amount of additional visitor support. In addition to Nissim, our Privacy Tools project has greatly benefited from numerous other collaborations and visitors (including with two other founders of differential privacy, Cynthia Dwork and Adam Smith, as well as with Sofya Raskhodnikova and David Xiao), and the budget will help us continue fruitful interactions of this type. The budget for years 2 and 3 also includes

staff support for our project coordinator, after the corresponding funding from our NSF grant expires.