OpenDP: An Open-Source Suite of Differential Privacy Tools

Salil Vadhan (PI), Mercè Crosas (co-PI), James Honaker (co-PI),
Gary King (co-PI)

Extract from our proposal to the Sloan Foundation

Project Goal

We propose to lead a community effort to build a system of tools for enabling privacy-protective analysis of sensitive personal data, focused on an open-source library of algorithms for generating differentially private statistical releases. We aim for this platform, OpenDP, to become the standard body of trusted and open-source implementations of differentially private algorithms for statistical analysis and machine learning on sensitive data, and a pathway that rapidly brings the newest algorithmic developments to a wide array of practitioners.

Objectives

We envision OpenDP as an open-source project for the differential privacy community to develop general-purpose, vetted, usable, and scalable tools for differential privacy, which users can simply, robustly and confidently deploy.

Proposed Activities

We will run workshops and provide small research grants to build a community of DP experts committed to an open-source library of DP algorithms and a system to deploy them. Together with this community we will produce a blueprint for library contributions and system deployment, and begin this development.

Expected Products

We will build a small steering group for OpenDP as well as an extended open-source community of experts, and produce a minimum viable system by the end of one year of support. We will establish a solid foundation for long-lasting continued development of the OpenDP system.

Expected Outcomes

We will enable researchers to find, explore and analyze sensitive data, and for government, industry, and other institutions to share such sources. The resulting contributions to knowledge, given the burgeoning new sources of sensitive data, will help shape all fields of knowledge on human behavior.

OpenDP: An Open-Source Suite of Differential Privacy Tools

1. What is the research question and why is it important?

We propose to lead a community effort to build a system of tools for enabling privacy-protective analysis of sensitive personal data, focused on an open-source library of algorithms for generating differentially private statistical releases. We aim for this to become the standard body of trusted and open-source implementations of differentially private algorithms for statistical analysis and machine learning on sensitive data, and a pathway that rapidly brings the newest algorithmic developments to a wide array of practitioners.

When big data intersects with highly sensitive data, both opportunity to society and risks abound. Traditional approaches for sharing sensitive data are known to be ineffective in protecting privacy. Differential Privacy [DMNS06], deriving from roots in cryptography, is a strong mathematical criterion for privacy preservation that also allows for rich statistical analysis of sensitive data. Differentially private algorithms are constructed by carefully introducing "random noise" into statistical analyses so as to obscure the effect of each individual data subject. (See the appendix for more background on differential privacy.)

Despite substantial demand from government agencies, human-subjects research communities, industry, and data archivists, practical adoption of differential privacy remains slow. To address this problem, we propose to encourage and guide the research community in constructing OpenDP, a trustworthy suite of differential privacy tools that will be a public resource for use by any organization wanting to make use of differential privacy.

Some of the use cases we envision for OpenDP are to enable:

- Archival data repositories, such as Dataverse, ICPSR, and Zenodo, to offer academic researchers privacy-preserving access to sensitive data. This would allow both novel secondary reuse and replication access to data that otherwise is commonly locked away in archives, or shared using ineffective "deidentification" methods that fail to protect privacy.
- Government agencies to safely share sensitive data with researchers, data-driven policy makers, and the broader public. For example, although the US Census has invested heavily in differential privacy, they continue to lack a library of algorithms they can promote to other Federal statistical agencies who seek to draw on their expertise.
- Companies to share data on their users and customers with external researchers (as in the Social Science One project funded in part by the Sloan Foundation) or with institutions that bring together several such datasets (as in the proposed "Institute for the Secure Sharing of Online Data" being formulated under a grant from the Sloan Foundation). This enables groundbreaking research on novel informative datasets, often of important social consequence, without the expense of one-off single researcher solutions or the risk of violating user trust (as in the Cambridge Analytica incident).
- Additionally, beyond the context of sensitive data, Differential Privacy has been shown to guard against overfitting and to ensure generalization, and our released tools could also be used for increasing the robustness of empirical findings.

Although we see many more potential use cases for researchers, industry, policymakers and the broader public, we are focused on opening otherwise siloed and sequestered sensitive data to support scientifically oriented research and exploration in the public interest. This presently means we would focus on the "centralized model" of differential privacy, in which there is a trusted "curator" that can store and compute on the entire sensitive dataset to produce statistical releases or synthetic data (which are infused with noise in order to satisfy differential privacy).[1] More details on the

[1] An alternative model for differential privacy is the "local model" where noise infusion is done by the data subjects

first two use cases above can be found in the appendix.

2. What is the state of the research on this question?
As mentioned above, differential privacy has started to have large-scale deployments in industry and government (Google, Apple, Uber, US Census Bureau) and there is also an extensive body of implementation and experimental work in academia. As far as we know, all of this work falls short of our goals for this project as they:

1. are highly tailored to a specific application, with particular data sets and types of analyses to be supported,
2. require more expertise in computer science or differential privacy than our anticipated users would have,
3. and/or have not been vetted by the differential privacy community at large.

In the Privacy Tools Project at Harvard, we have been developing a differential privacy tool, PSI (a Private data Sharing Interface), that aims to address Items 1 & 2 above [GHK+16]. PSI is being designed to integrate into data repositories like the 40 repositories that use the Dataverse software infrastructure, to allow releasing public statistical summaries and providing interactive queries for exploratory data analysis. PSI is planned for beta deployment along with Dataverse 5, targeted for the end of 2019. However, PSI's functionality (in terms of the statistical analyses it supports) is still fairly limited, and the underlying library of implemented differentially private algorithms has not been vetted by the research community at large.

We envision OpenDP as a larger open-source project for the differential privacy community to develop general-purpose, usable, and scalable tools for differential privacy (DP) in which users can have confidence. PSI, along with other implementations of differential privacy in academia and industry, can serve as starting points for components of OpenDP, which we envision will continue to grow and evolve through the community that builds around it. A community-driven open-source library that incorporates contributions from many such projects will have greater capabilities, better trust, and broader adoption, which will lead to faster translation of differential privacy to practice. Moreover, even among current existing government and industry deployments, there has been a stated desire to draw from common, trusted, open codebases, rather than reinvent the underlying primitives continually for each new project, and we believe that OpenDP will provide that codebase and receive enthusiastic code contributions from industry partners.

3. Why is the proposer qualified to address the research question for which funds are being sought?

This project has emerged from the successful collaboration of our team in the Harvard Privacy Tools Project, which has been funded by an NSF Secure and Trustworthy Cyberspace Frontier Project, "Privacy Tools for Sharing Research Data," the Sloan Foundation grant "Applying Theoretical Advances in Privacy to Computational Social Science Practice," and other sources. This multidisciplinary project is a joint effort of Harvard's Institute for Quantitative Social Science

themselves before sending anything to the (now untrusted) curator; this has been used in commercial deployments by companies such as Google, Apple, and Microsoft for collecting data from their customers. An intermediate model between the centralized and local models is the "multiparty model" where several curators hold sensitive datasets about different subjects (e.g., hospitals, each with data about their patients) and the goal is to allow for statistical analysis over the union of the datasets. However, the reduced need for trust in these models has a significant price, either in the accuracy of statistics computed or in computational complexity coming from use of cryptographic techniques like secure multiparty computation. These models would require a significantly different and more complex software architecture than our proposed suite of tools for centralized differential privacy.

(IQSS), the Center for Research on Computation and Society (CRCS), and other centers. One of the key products of this project is the PSI tool mentioned in the previous section, for which the proposers led development.

PI Salil Vadhan is a Harvard College Professor and the Vicky Joseph Professor of Computer Science and Applied Mathematics in the Harvard Paulson School of Engineering and Applied Sciences. He is the lead PI on the Harvard Privacy Tools Project, and was CRCS Director from 2008-11 and 2014-15. Over the past decade, he and his collaborators have obtained numerous results delineating the border between what is possible and impossible with differential privacy. He is a recipient of a Simons Investigator Award, a Gödel Prize, a Guggenheim Fellowship, a Sloan Fellowship, a Phi Beta Kappa Award for Excellence in Teaching, and the ACM Doctoral Dissertation Award.

Co-PI Mercè Crosas is Chief Data Science and Technology Officer at IQSS and Harvard University's Research Data Officer at the Office of Vice Provost for Research. Together with co-PI King, Crosas started the Dataverse software project in 2006 and has since led development of the project. Her role is to provide strategic direction for the architecture, design, and research of the Dataverse project, and lead its community engagement. In particular, Crosas initiated four years ago the successful annual Dataverse Community Meeting, which brings together about 200 Dataverse developers and users from around the world. With this initiative and other outreach efforts, the Dataverse project has now more than 80 contributors, generates 12 releases a year with 550 pull requests, and has 40 installations of the software platform across practically all continents. Crosas also participates in numerous community groups to build standards for research data and software, including data citation and software citation principles and implementation.

Co-PI James Honaker is a Research Associate at CRCS. Previously he has been a Senior Research Scientist at IQSS, and faculty at Penn State and UCLA. He leads development of the PSI: Private data Sharing Interface. His research focuses on statistical software solutions for broad problems in quantitative social science. He is the author of several widely used statistical software packages for quantitative social science, including Amelia (for missing data), Zelig (for statistical inference and interpretation), and TwoRavens (for data exploration and automated machine learning). He won the 2014 Award for Best Statistical Research Software of the Society for Political Methodology.

Co-PI Gary King is the Albert J. Weatherhead III University Professor at Harvard University—one of 24 with Harvard's most distinguished faculty title—and Director of the Institute for Quantitative Social Science. King develops and applies empirical methods in many areas of social science research, focusing on innovations that span the range from statistical theory to practical application. King leads the Dataverse project together with co-PI Crosas, and leads Social Science One together with Nathaniel Persily. He will lead our efforts to find social science use cases for OpenDP, as well as our efforts to raise industry support.

All four have been extensively involved in five years of summer programs training and mentoring research fellows (undergraduate, graduate and postdoctoral) in privacy research and tools implementation, as well as spearheading workshops on privacy both for the research community and the broader lay public of quantitative researchers and data archivists. In response to discussions with government and industry about the difficulty finding and training engineers to work in privacy, PI Vadhan and co-PI Honaker have developed and are currently teaching a graduate course examining the pragmatics of software deployments of differential privacy.

4. What is the research methodology?

We orient the construction of our proposed tools around a series of key principles for ensuring

trust, a community structure for enlarging participation, and a series of architectural components for establishing adoption of our tools. We detail these key principles here as the foundation for describing our community building and tool architecture in the next section.

Key Principles: The design and community governance of the system should reflect the following core design principles to allow for the most trusted and capable system:

- *Open Source:* Research and Development advances will be driven by an open-source community, with a core team that will guide the implementation and priorities. It will engage a broad set of DP experts and development contributors worldwide, through appropriate processes and incentives.
- *Security and Privacy:* To guard against security and privacy risks, the additions to the API or to any critical component of the service must be vetted by security experts, additions to the library of differentially private algorithms must be vetted by experts in differential privacy, and the service itself should avoid storing (and in most cases even accessing) raw sensitive data.
- *Scalability:* The service must provide means to scale for datasets of TB to PB size.
- *Extensibility:* The architecture must be modular and allow contributions of new features through a well defined process.

5. What is the work plan?

The key principles above provide the framework from which we build both our community and system architecture, and we now describe these concrete components of our project strategy. Finally we provide our current blueprint for system development, although we expect that this will evolve through community contribution.

Key Phase Zero Community Building As an open-source community effort, and anticipating broad adoption across data partnerships, the community architecture—including governance, appropriate resources, and community engagement—is as important as the technical software architecture. We envision a foundational Phase Zero in which we assemble expert involvement around key tasks, leverage current work by co-PI Crosas on open-source software health,[2] and follow recommendations provided by the NumFocus project[3] to define OpenDP's community architecture and sustainability (see Appendix for more on project sustainability). Key teams include:

- **Principal DP scientist(s):** Oversees the design and implementation of the DP library, budgeting interfaces and other aspects of the privacy preserving architecture.
- **OpenDP architect (CTO or CSO):** Designs and guides implementation of OpenDP architecture. Leads the process to approve the implementation of a new model in the DP service, critical changes in the technical components, and extension of the API.
- **Steering committee:** Assembled from experienced distinguished experts in the curation, analysis, and protection of data as well as open-source software development, this committee periodically reviews the growth priorities and capabilities of the system, as well as the community engagements, and provides guidance and direction to all aspects of the system development.
- **DP oversight committee:** Consists of experts in the development and deployment of differentially private algorithms. They will provide guidance on the architectural choices in OpenDP

[2] Co-PI Crosas leads a joint IMLS and Sloan funded project titled "A Proposed Quantitative Index to Understand and Evaluate the Health of Open Source Projects", which leverages the also Sloan funded CHAOSS project (https://chaoss.community/) and the IMLS funded "Lyrasis: It Takes a Village" (http://lyrasisnow.org/press-release-itav-guidebook/) project to define metrics to evaluate the health of academic open-source software projects

[3] https://numfocus.org/

that implicate differential privacy and the prioritization of methods from the differential privacy research literature to incorporate into OpenDP, as well as serve as an editorial board for the vetting of algorithms and their implementations as they are incorporated into OpenDP.

- **Security oversight committee:** This committee consists of experts in computer security, who will oversee and provide guidance on security aspects of the system architecture, including secure storage, transfer APIs, authentication, information flow, side-channel attacks, and processes for identifying vulnerabilities (e.g. via bug bounties).
- **DP development team:** In charge of technical components requiring DP expertise, such as the library, programming and graphical interfaces, and verification of code.
- **System development team:** In charge of technical development of system architecture, deployment, containers and APIs, and integration with data custodian partners.
- **Open-source coordinator:** In charge of driving the collaborative process, the open-source coordinator collects input from all the stakeholders: the contributors of code to OpenDP, the academic, government, and industry organizations that will deploy and use OpenDP, and individual end-users, so as to help set priorities, grow the community, and foster adoption.

In this hierarchy, teams are primarily made up of staff developers who create robust, polished, enterprise code, while committees are made up of research and practitioner experts who provide expert guidance and vetting, in addition to contributing code coming from their own research. Of these, the three committees—the steering, DP oversight, and security oversight committees—are the key and novel building block of our community plan. These are the points at which we plan to harness the contributions of motivated researchers and experts. The steering committee will consist of experienced distinguished experts representing many of the stakeholders that OpenDP will serve, including the DP research community, privacy policy, social science, medical informatics, data repositories, open-source software development, government, and industry organizations. It will provide guidance on the broad goals of the system and review its growth priorities, accomplishments and milestones. We envision that in steady state, it will meet twice per year, though perhaps more frequently during the ramp-up of OpenDP. The DP oversight committee will be filled with leading experts in the development and deployment of differentially private algorithms. They will recommend the most pressing algorithms to add to the library; oftentimes, additionally committee members would contribute code from their own research to form the first draft of library implementations. The DP committee will oversee the vetting of the mathematical proofs of algorithms before the DP development team begins writing code implementations, and then also oversee the review and verification of the code. Similarly, the security oversight committee would be created from researchers in systems security, and will provide guidance on security aspects of the system architecture. We expect that these oversight committees will meet quarterly in steady state, with some work carried out between meetings (similarly to an editorial board). Our initial ideas for the membership of these committees is listed in the appendix.

While the committees are constructed mostly from researchers, the DP development and system development teams are primarily composed of professional software developers. Some funding from this proposal would go towards some positions in these teams, and our plan would grow these teams with future grants and industry support. To foster contributions from and use by industry, we envision offering six-month residencies, wherein companies could send developers to join the OpenDP team, contribute to OpenDP codebase, and learn and train in differential privacy before returning to their companies, where they may also continue to contribute to the OpenDP codebase as their company makes use of OpenDP's tools. (We have extensive experience in teaching differential privacy through the tutorials we have offered every year in the very successful Privacy Tools summer internship program, and the "Applied Privacy for Data Science" graduate course currently being taught by Honaker and Vadhan.) As we intend for this to be a community-wide

effort, we are keen to cooperate with any commercial and non-commercial partners who agree with the value of such a library.

The team of co-PIs in this proposal, Vadhan, Crosas, Honaker, and King, are committed to the successful flourishing of this project and system and leading it through at least its first year. PI Vadhan will serve in the role of Principal DP scientist for the initial launch of OpenDP, but a newly recruited community member might take on this role in the future. PI Vadhan, and co-PIs King and Crosas will be part of the steering committee, while co-PI Honaker will start off by leading the DP development team and participating on the DP oversight committee. The committees will be filled by community experts in security and privacy from academia and industry recruited in the Phase Zero supported by this proposal.

Key Phase One Deliverables In Phase One we plan to build a deployable privacy preserving solution over the course of a year. The exact architecture and specifications of these components will be driven by community decisions in Phase Zero, but we expect the following components to be crucial to any successful approach.

- *Library of DP Algorithms/Methods:* A library of DP methods will grow as the open-source community contributes new models. A new method will go through vetting managed by the DP oversight committee before it is released and supported through the API.
- *Budgeting Interfaces:* We propose to provide at least two interfaces to the system, for two types of users. The first is a clear intuitive graphical user interface (GUI) that provides a constrained workflow and detailed guidance to users that allows data owners and analysts with no privacy expertise to make informed choices about the balance between accurate answers and accumulated privacy loss. In addition, we propose to provide a more expressive programming interface that allows a sophisticated user to describe more sophisticated and customized statistical releases built up from basic differentially private primitivesMach.
- *API:* A new API needs to be defined for this new service. The API will support submitting a DP request based on one of the supported DP methods in the library. The REST API should be secured by using an authorization protocol such as OAuth2.0. The API should be registered as a SmartAPI to follow best practices from the open-source community,
- *Containers:* The DP service will use a container (Docker, or similar technology) that holds the script of DP algorithms and requests to be applied to the data. The container is pulled from the location of the data source. The analysis takes place in the data enclave where the sensitive data resides, and the raw data never needs to leave the enclave or data source. The results are placed in a location accessible by the DP service.
- *Authentication and Authorization:* OAuth2 protocol (or similar) will be used to authenticate and authorize users to use the API.
- *Large-scale data engine:* For large datasets, it might be necessary to deploy a Spark cluster or similar big data engine to run the DP algorithms in the data source location. In this case, the DP methods will need to be written to support Spark (or in particular SparkR) and take advantage of computing parallelism.

Many of these components, may be adopted individually by practitioners for use in their own projects. For example, a government agency might use the Library as is to release privacy preserving statistics, an organization with DP expertise might use the Library as a starting point for designing custom DP algorithms (which they then can contribute back to OpenDP), and an industry portal might use the budgeting GUI to facilitate simple queries from non-experts for their own DP algorithms. However, in combination, the set of components will allow the full OpenDP system to be a web service that provides the ability to construct scripts or code to generate differentially

private statistical releases on a sensitive, secure, remote dataset, when supplied with the appropriate metadata to construct the release, and correctly measures the cumulative privacy loss of such releases. OpenDP will expose an Application Programming Interface (API) that allows agents who hold sensitive data, as well as authorized analysts, the ability to describe statistical releases they require from OpenDP's library algorithms, and returns securely to the site that holds the data the code to generate these privacy-preserving releases. These privacy-preserving algorithms will be curated from the expert community of differential privacy researchers, vetted by the DP oversight committee, and published in open-source formats that can be run on common large-scale data science stacks.

Relation to Previous Work by the PIs In our work most directly related to OpenDP, and which was funded in part by a grant from the Sloan Foundation, co-PIs Honaker, King, Vadhan, and collaborators have built the prototype tool, PSI (a Private data Sharing Interface), to demonstrate that the strong privacy guarantees of differential privacy could be brought to practice for use by quantitative social scientists with archival data. While differential privacy is a highly specialized mathematical field completely removed from the topics covered in quantitative statistical training, our goal was to prove that a system could be used by data collectors and analysts without any advanced training, and without any involvement or intervention by computer scientists or other experts in data privacy. In much the same way that most researchers use statistical software without needing to know anything about the underlying computer architecture, we aimed to build a system that allows researchers who want to distribute data to focus on their research topics, and contribute and leverage what they know about their data, while abstracting away the mechanisms that protect privacy. In addition to accessibility, we aimed to build a generalized system that would work across the wide variety of social science datasets that might be deposited in a data repository like those that use the Dataverse platform. In the technical literature, and in the existing industrial deployments, DP algorithms are often highly tailored and tuned to a specific attached use case, while our goal of a system that does not require the intervention of privacy experts means focusing on generalizable techniques that pragmatically and robustly work across the spectrum of data settings. PSI is part of a larger effort to expand the Dataverse repository infrastructure to take on sensitive data, including other important elements such as support for secure storage, authentication, customized terms of use, and other law and policy aspects of sensitive data. This includes integration of Dataverse with the DataTags system [SCB15], a set of standardized levels that define security, access, and contractual (DUA) requirements. Using the DataTags system will help guarantee that these pre-defined requirements assigned to each dataset will be accordingly satisfied. We plan for the release of Dataverse version 5 by end of 2019 to incorporate these features, including configuration to integrate with a beta version of PSI as an external tool, so that the 40 data repositories around the world that use Dataverse can start hosting sensitive data and employing differential privacy as one means for accessing such data.

Our experience in building and deploying PSI will greatly inform our efforts towards OpenDP. We expect that elements of PSI's architecture, and possibly some of the actual software, will be incorporated into the OpenDP suite. However, OpenDP will greatly expand the capabilities and scope of PSI in a number of respects. First, the library of differentially private algorithms currently underlying PSI is quite minimal; OpenDP aims to produce a much more comprehensive library incorporating sophisticated, state-of-the-art methods contributed by researchers throughout the differential privacy community. Second, OpenDP will incorporate architectural insights coming from other efforts in the differential privacy community to design general-purpose differential privacy tools, such as as PinQ [McSherry09], εktelo [ZMK+18], PrivateSQL [KTM+19], Fuzz [HPN11]

and LightDP [ZK17]. Like PSI, most of these tools are currently research prototypes; OpenDP will provide an avenue to turn them into production-ready tools and get them into the hands of many users.

Furthermore, our experience building an active and growing community for the Dataverse project and our current joint IMLS and Sloan-funded project titled "A proposed quantitative index to understand and evaluate the health of open source projects" will inform us how to best build and grow the OpenDP community. On one hand, the experience with Dataverse has showed us the importance of being connected to the community through a diverse set of venues: active mailing lists, bi-weekly community calls, and an annual in-person community meetings. It has also showed us the importance of managing the project using an agile approach, with transparency about the project's roadmap, the issues backlog, and what is being worked on each sprint (2 weeks), as well as a technical and management process that enables the community to easily contribute their code. On the other hand, our current work on open-source software health metrics will help assess quantitatively the growth and sustainability of academic open-source software projects, using expert elicitation and mixed methods approach to define the appropriate (weighted) metrics for each project based on a defined set of project goals. The outcomes of this study can help define and collect the appropriate metrics for OpenDP. The appendix contains more details on our sustainability plan.

OpenDP Architecture

Library: Our fundamental goal is for the OpenDP library to become the standard body of trusted and open-source implementations of differentially private algorithms for statistical analysis and machine learning on sensitive data, and a pathway that rapidly brings the newest algorithmic developments to a wide array of practitioners. To reach this breadth of adoption, we want to support the currently most common definitions of differential privacy (such as pure, approximate, and concentrated differential privacy for tabular data), and be extensible to allow for including other versions in the future.

We have a core library from PSI in R, but presently plan to port and rebuild either in Python, or in C++ with R and Python bindings to allow for contributions in both languages.

Metadata for the algorithms in the library would define which privacy definition(s) the algorithm meets, and allow the system to offer different correct subsets of the library depending on the library user's chosen privacy model. In addition to vetted algorithms to make a library useful in deployment and to practitioners, we have found through our experience with PSI that these algorithms need a flora of surrounding utility functions for various types of use, such as to translate privacy-loss parameters into promises of accuracy in the DP releases, or conversely to determine the required privacy-loss parameter to meet a desired level of accuracy, as well as to construct confidence intervals about releases, and to "post-process" other statistical information that can be released for no privacy loss as functions of previously released information. Our library would also offer composition functions for measuring total privacy-loss across releases across an analysis. We have examples of all of these functions for the present statistics available in our PSI Library in R, including code for the optimal composition theorem for approximate differential privacy [MV16], the theoretical work for which was developed specifically for PSI.

We envision the library having a modular architecture like εktelo [ZMK+18] that isolates and minimizes privacy-critical trusted code that needs to be verified carefully. This privacy-critical code will be verified by a combination of manual vetting by DP experts (the DP oversight committee) and tools for formal verification of DP (which started with Fuzz [HPN11] and has now become much easier to use with systems like LightDP [ZK17]).

Interfaces: A library of differentially private algorithms is not sufficient to enable non-experts to use differential privacy safely; it is also essential to have tools that manage the limited "privacy loss-budget" over many statistical releases by a data holder and/or interactive queries by data analysts. One of the challenges is that it can be difficult to understand the implications of different selections of the privacy loss parameters (namely ϵ and δ), both in terms of privacy and utility, especially when these need to be distributed over many different statistics. We envision that OpenDP will include several budgeting and query interfaces that can be used with the underlying library.

We anticipate that one of these budgeting and query interfaces will be similar to the graphical user interface that we have built for PSI (and indeed this may be PSI's most concrete contribution to OpenDP). The PSI interface exposes the accuracy versus privacy tradeoff to users in easy-to-understand terms and is accompanied by a variety of simple explanations of differential privacy and its parameters that are shown to the user at relevant times, with tailored natural language interpretations. Through this support, analysts who may not have any familiarity with differential privacy can still make decisions about what statistics they would like to release, informed by the level of accuracy that can be expected. We have invested in usability experiments with non-expert data analysts to continue to refine our system, and to demonstrate the potential of this approach. While the PSI use-case required extensive development of a GUI for applied researchers, we believe a programming interface will be an important requirement for many of the more sophisticated users of our library and system. A programming interface would allow users to write scripts using our library and automatically certify that all of the points where the sensitive data is touched are through the vetted differentially private mechanisms in the library, and enforce a privacy-loss budget across an entire analysis or even from cumulative analyses from different researchers whose previous releases have been recorded. PSI does not presently offer a programming interface, and does not have the capability to track the privacy loss budget for analyses that go beyond what can be expressed in its GUI. Instead, we hope to incorporate insights coming from programming frameworks such as PinQ [McSherry09] and εketelo [ZMK+18].

Containerization: We envision OpenDP will not directly touch sensitive data, but instead will generate code to construct differentially private releases, that can be securely sent to the secure locations where the data is presently stored, and the owners (who may verify the certification of the code) can run the code on their own data to generate a differentially private release. This may involve utilizing OpenDP created or sanctioned software components for executing these scripts, such as to guarantee randomness and prevent side-channels in the execution. OpenDP will also guide a schema standard for the storage and sharing of differentially private releases, for interoperability with different users and reusers of OpenDP releases. The service needs to guarantee that only approved and secure requests can be submitted, and the generated code is securely returned. Creating trusted code that can be executed remotely, in a "software as a service" model, will allow our system to provide data owners a way to keep very large or very sensitive data on their own systems, and execute all analyses in their own trust environments. By removing the need to move sensitive data to places where owners lose control, we think we can increase the willingness of data owners to open privacy-preserving access to their data, and also remove the risks of data being exposed in insecure or unintended locations or intercepted in transit.

The appendix contains more details on important system design choices in OpenDP.

Community Engagement The success of OpenDP will rely crucially on buy-in and participation from the wider differential privacy community, to participate in the development and gover-

nance of OpenDP, to contribute code to its libraries, and help in vetting code contributed by others. To this end, PI Vadhan pitched the vision described in the present proposal at the Sloan-sponsored March 2019 workshop "Data Privacy: From Foundations to Applications" at the Simons Institute for the Theory of Computing, where many of the leaders in bringing differential privacy to practice (from academia, industry, and government) were present. The talk received an enthusiastic response and led to an engaged discussion about how to ensure its success. We describe some of the ideas offered below.

Researchers in the differential privacy community have a strong desire to maximize the impact of their work, and OpenDP can provide a channel for such impact. To maximize these incentives, we can provide code contributors with measures of how widely their code is used, and encourage users of the library to cite the researchers who designed the algorithms they used. Along these lines, we have been working to make the PSI library self-citing, so that as an analysis is constructed, all the algorithms that have been used are put into a bibliography that credits the original paper authors, to facilitate the proper citation of works contributed to the library. The cloud deployment of the library in the OpenDP system could also track the usage of algorithms and report metrics to the original authors to show the impact of their work. Another idea is to offer automated benchmarking against a corpus of test datasets (building on DPBench [HMM+16]), so that contributors of algorithms can easily demonstrate the utility of their contributions in papers they produce from their work. A "bug bounty" program can be used to incentivize the community to carefully inspect any new contributions to OpenDP for vulnerabilities.

Our proposal to offer 6-month residencies for software engineers from industry also came from this community discussion, out of a recognition that there is currently a high demand (but low supply) for differential privacy engineers.

6. What will be the output from the research project?

We aim for OpenDP to become a standard body of trusted and open-source implementations of differentially private algorithms for statistical analysis and machine learning on sensitive data, and a pathway that rapidly brings the newest algorithmic developments from the theoretical literatures to a wide array of practitioners. As such we want to build a long-lasting open-source community, with no foreseeable endpoint. To accomplish this we plan a two-phase deployment together with a speedy (Phase Zero) period of community building and engagement, which together form a two-and-a-half-year launch window. We expect a deployable DP solution and a coalesced community of privacy research experts and practitioners by the end of our first year, which is the period for which we are asking for support.

During the **Phase Zero** period we plan to establish the foundational community of researchers and practitioners to build OpenDP. Two major products of this period are linked workshops to establish the open-source community and achieve the buy-in of experts and practitioners. Immediately we plan to host a small workshop, focused on laying a groundwork, and composed primarily of differential privacy experts who have experience on deployed DP systems (such as those on the proposed DP oversight committee, as listed in the Appendix), along with a few additional key experts (e.g. on open-source software development). This workshop will produce the starting whitepaper of both the software architecture and community structure, as well as the method of recruitment to the following workshop. Subsequently, we propose to run a much larger workshop, that is broadly inclusive of the many stakeholders for OpenDP, such as those on the proposed committees (see Appendix). In this workshop, we will introduce the whitepaper from the first workshop as the foundational proposal, and allow the broader community to develop and refine the vision of both the community principles and the software architecture of the project itself. We expect

	Timeline	Key Deliverables	
Phase Zero	0-6 months	Community Building, Organizational Workshops, Architectural Specifications, Board and Leadership Roles Filled.	Funded Period
Phase One	7-12 months	Budgeting Graphical User Interface, Large Scale Data Engine, Continued Library Development, Continued Security Engineering.	Funded Period
	13-18 months	Budgeting Graphical User Interface, Large Scale Data Engine, Continued Library Development, Continued Security Engineering.	
Phase Two	18-30 months	Expanded Use Case Partner Development, Additional Library Contributions, Architecture and Deployment Improvements.	

these workshops to conclude and publish their finalized guidelines within six months. The key deliverable from these workshops should be a formal open-source structure for this project, a core of researchers committed to the required service roles, and an agreed-upon architectural blueprint for the OpenDP library and system development.

During **Phase One** we plan to build a deployable privacy-preserving solution over the course of a year. The first six months of this will be supported by this grant. In the funded period we will deliver a minimal viable product of the OpenDP platform, including a library of differentially private algorithms, an architecture for running DP operations composed of these libraries in remote storage, and a programming interface for using this library that will confirm and verify that an analysis meets the privacy-loss constraints desired by the user.

While both the scale of the system and the timelines we are proposing are ambitious, we believe we can draw on our earlier experience building the more limited differential privacy toolkit, PSI, as well as similar projects of interested collaborators, to kick-start development. This, in combination with a notably strong and broad core of theoretical computer scientists and applied quantitative practitioners in our existing research group, and strong interest from the broader research community and government and industry adopters to join this endeavor, can give this project a rapid launch from which we envision OpenDP will continue to grow and evolve through the community that builds around it.

In addition to the major products of community formation and minimum viable OpenDP release, we expect numerous secondary research developments. In particular, during Phase One, as part of our open-source strategy, we have set aside funds to provide small grants to research contributors to aid in tasks for the system development and the library implementation. We anticipate around 5 small grants to aid researcher and student time, and cement proposed contributions to OpenDP. From this, we expect not only the flourishing of the concrete software and service deliverables already described, but likely new research products, particularly on the pragmatics of DP implementation and experimentation, such as papers and workshop talks, will be additional products of this funding.

Appendices

What is Differential Privacy?

[This section, except for the fifth paragraph, is a verbatim extract from the differential privacy primer [WAB+18], coauthored by co-PI Honaker, PI Vadhan, and collaborators.]

Differential privacy is a strong, mathematical definition of privacy in the context of statistical and machine learning analysis. It is used to enable the collection, analysis, and sharing of a broad range of statistical estimates based on personal data, such as averages, contingency tables, and synthetic data, while protecting the privacy of the individuals in the data.

Differential privacy is not a single tool, but rather a criterion, which many tools for analyzing sensitive personal information have been devised to satisfy. It provides a mathematically provable guarantee of privacy protection against a wide range of privacy attacks, defined as attempts to learn private information specific to individuals from a data release. Privacy attacks include re-identification, record linkage, and differencing attacks, but may also include other attacks currently unknown or unforeseen. These concerns are separate from security attacks, which are characterized by attempts to exploit vulnerabilities in order to gain unauthorized access to a system.

Computer scientists have developed a robust theory for differential privacy over the last fifteen years, and major commercial and government implementations are starting to emerge. Differential privacy mathematically guarantees that anyone viewing the result of a differentially private analysis will essentially make the same inference about any individual's private information, whether or not that individual's private information is included in the input to the analysis.

What can be learned about an individual as a result of her private information being included in a differentially private analysis is limited and quantified by a privacy loss parameter, usually denoted epsilon (ε). Privacy loss can grow as an individual's information is used in multiple analyses, but the increase is bounded as a known function of and the number of analyses performed.

Differentially private algorithms are constructed by carefully introducing random noise into statistical analyses so as to obscure the effect of each individual data subject. Thus, differential privacy reduces the accuracy of statistical analyses, but does so in a quantifiable manner that introduces an explicit privacy-utility tradeoff. As the number n of observations in a dataset grows sufficiently large, the loss in accuracy due to differential privacy generally becomes much smaller than that due to statistical sampling error. However, it can be challenging to maintain high accuracy for studies on modest-sized datasets (or modest-sized subsets of large datasets).

The differential privacy guarantee can be understood in reference to other privacy concepts:

- Differential privacy protects an individual's information essentially as if her information were not used in the analysis at all, in the sense that the outcome of a differentially private algorithm is approximately the same whether the individual's information was used or not.
- Differential privacy ensures that using an individual's data will not reveal essentially any personally identifiable information that is specific to her, or even whether the individual's information was used at all. Here, specific refers to information that cannot be inferred unless the individual's information is used in the analysis.

As these statements suggest, differential privacy is a new way of protecting privacy that is more quantifiable and comprehensive than the concepts of privacy underlying many existing laws, policies, and practices around privacy and data protection. The differential privacy guarantee can be interpreted in reference to these other concepts, and can even accommodate variations in how they are defined across different laws. In many cases, data holders may use differential privacy to demonstrate that they have complied with legal and policy requirements for privacy protection.

Differential privacy is currently in initial stages of implementation and use in various academic,

industry, and government settings, and the number of practical tools providing this guarantee is continually growing. Multiple implementations of differential privacy have been deployed by corporations such as Google, Apple, and Uber, as well as federal agencies like the US Census Bureau. Additional differentially private tools are currently under development across industry and academia.

Some differentially private tools utilize an interactive mechanism, enabling users to submit queries about a dataset and receive corresponding differentially private results, such as custom-generated linear regressions. Other tools are non-interactive, enabling static data or data summaries, such as synthetic data or contingency tables, to be released and used. In addition, some tools rely on a curator model, in which a database administrator has access to and uses private data to generate differentially private data summaries. Others rely on a local model, which does not require individuals to share their private data with a trusted third party, but rather requires individuals to answer questions about their own data in a differentially private manner. In a local model, each of these differentially private answers is not useful on its own, but many of them can be aggregated to perform useful statistical analysis.

Differential privacy is supported by a rich and rapidly advancing theory that enables one to reason with mathematical rigor about privacy risk. Adopting this formal approach to privacy yields a number of practical benefits for users:

- Systems that adhere to strong formal definitions like differential privacy provide protection that is robust to a wide range of potential privacy attacks, including attacks that are unknown at the time of deployment. An analyst using differentially private tools need not anticipate particular types of privacy attacks, as the guarantees of differential privacy hold regardless of the attack method that may be used.
- Differential privacy provides provable privacy guarantees with respect to the cumulative risk from successive data releases and is the only existing approach to privacy that provides such a guarantee.
- Differentially private tools also have the benefit of transparency, as it is not necessary to maintain secrecy around a differentially private computation or its parameters. This feature distinguishes differentially private tools from traditional de-identification techniques, which often conceal the extent to which the data have been transformed, thereby leaving data users with uncertainty regarding the accuracy of analyses on the data.
- Differentially private tools can be used to provide broad, public access to data or data summaries while preserving privacy. They can even enable wide access to data that cannot otherwise be shared due to privacy concerns. An important example is the use of differentially private synthetic data generation to produce public-use microdata.

Differentially private tools can, therefore, help enable researchers, policymakers, and businesses to analyze and share sensitive data, while providing strong guarantees of privacy to the individuals in the data.

Exemplar Use Cases

Some brief envisioned use cases are described in sections 1 and 5, and we elaborate some of those here for richer detail and to give a greater vision of the potential impact of this project.

Open Science Data Repositories: Dataverse [King07, Crosas11, Crosas13, King14], developed at Harvard's Institute for Quantitative Social Science (IQSS) since 2006 under co-PIs King and Crosas, enables researchers to share their datasets with the research community through an easy-

to-use, customizable web interface, keeping control of and gaining credit for their data while the underlying infrastructure provides robust support for good data archival and management practices. The Dataverse software has been installed and serves as a research data repository in more than 40 institutions worldwide. The Dataverse repository hosted at Harvard University[4] is open to all researchers, and contains one of the largest collections of small to medium sized research data in the world.

Prior to our work, Dataverse repositories (like most general-purpose data repositories) had almost no support for hosting privacy-sensitive data. Datasets with sensitive information about human subjects were supposed to be "de-identified" before deposit. Unfortunately, research in data privacy has demonstrated convincingly that traditional de-identification does not provide privacy protection. An early and dramatic example of the problems with traditional de-identification was given by Latanya Sweeney in the late 1990s [Sweeney97]. She showed how to re-identify patients in "anonymized" medical insurance claims records that were publicly released by the Massachusetts Group Insurance Commission by matching date of birth, gender, and ZIP code with publicly available voter registration lists. Surprisingly, these three attributes uniquely identify well over 50% of the US population [Sweeney00, Golle06]. In particular, she was able to identify the medical record of William Weld, who was then governor of the state of Massachusetts. There have since been numerous demonstrations of the ease of re-identifying subjects in de-identified datasets, including Netflix movie rentals [NS08], geographical information system data [GS07], genetic databases [MS00, MS01], and internet search engine logs [BZ06]. The risks posed by re-identification are magnified as social science research becomes more data-driven, more external sources of information that can be exploited for re-identification become available, and research datasets are made accessible to the general public based on well-intentioned open data policies.

The current alternative to open data sharing in repositories like Dataverse is that researchers depositing a dataset ("data depositors") could declare their dataset "restricted," in which case the dataset would not be made available for download, and the only way for other researchers to obtain access would be through contacting the data depositor and negotiating terms on an *ad hoc* basis. This approach is also unsatisfactory, as it can require the continued involvement of the data depositor, the negotiations can often take months, and thus it impedes the ability of the research community to verify, replicate, and extend work done by others.

In our work on the Privacy Tools Project, supported by the National Science Foundation, the Sloan Foundation, and the US Census Bureau, we have been building software tools to make privacy-protective and data-sharing more accessible and efficient for researchers with no expertise in data privacy, computer science, or law. These tools enable such researchers to use the methods of differential privacy; to navigate the complex requirements of privacy laws, university data policies, and institutional review board procedures; and to select appropriate data-handling policies and data-sharing agreements. In particular, our tool PSI (described earlier in the proposal), will allow a data depositor to offer differentially private summary statistics and exploratory data analysis to a wide range of Dataverse users, in addition to allowing for approved researchers to apply for access to the raw, sensitive data. As discussed in the proposal, the development of PSI so far has been targeted at integration specifically with Dataverse repositories. Moreover, it currently supports only a small collection of differentially private algorithms and its implementation has not been vetted by the community at large.

OpenDP will allow for differential privacy to be safely deployed in a similar way across many other data repositories in the social and health sciences. It will offer a professional and trusted codebase that integrates easily with other repository platforms, with much more functionality than

[4] http://dataverse.harvard.edu

any single academic research project can offer, and focused on the pragmatic statistical techniques needed by quantitative researchers, including model optimization, inference and hypothesis testing. OpenDP will offer a programming interface so that repositories can accept verified DP scripts to run on the data in their holdings, and a graphical user interface for more simple data exploration.

Government Statistics: In recent years, federal, state, and local agencies have been making increasing amounts of data publicly available pursuant to open data initiatives. Government open data, from segregation and urban mobility to records of bicycle collisions, from geolocated crime incidents to geolocated pothole reports, can support a wide range of social science research about the fine-grained structure of social and political life. Notable open city portals include Boston's BARI[5] and New York's CUSP[6], and the importance of such data is highlighted by the fact that the meta-question of which governments have more open data and why has itself become a social science literature [Ubaldi13, VBS14, ZJ14].

However, most government agencies are ill-equipped to appropriately address the privacy concerns in their data. Consequently many datasets are either withheld or released with inadequate protections. Data published to municipal open data portals often undergo an *ad hoc* de-identification process in which columns deemed identifying are removed or coarsened prior to release in microdata formats, and identifiable and sensitive information can frequently be found in the published data. For example, the open data portal for the City of Boston has published records of 311 requests containing individual-level information, such as addresses of residents requesting assistance from a program offering inspections to households with children suffering from asthma [AWO+16]. Our Privacy Tools team, through ongoing engagement with open government data communities, has developed a framework by which government agencies can match modern privacy tools to the risks and intended uses of their data [AWO+16]. This framework has been advocated by a draft NIST Publication [Garfinkel16] providing guidance to government agencies on applying de-identification techniques. However, despite calling for use of "modern privacy tools" such as differential privacy across government, these tools are not currently available in software that can be used by non-experts, in particular by government open data managers. Our work will start to remedy this situation.

Government agencies will be able to use OpenDP to produce rich statistical summaries of sensitive datasets that can be shared widely without worry that the combination of released statistics will reveal individual-level information (in contrast to data de-identified using traditional means, which have repeatedly been shown to be vulnerable to re-identification). This is similar to how the US Census Bureau plans to use differential privacy to produce public-use microdata samples for the 2020 Decennial Census. (In fact, members of our team are participating in a Census-funded cooperative agreement, "Formal Privacy Models and Title 13," that aims to help the Bureau in this regard. That effort is synergistic with the one we are proposing here.) In addition to tables of statistics that would be of common interest, in principle it is possible to generate differentially private "synthetic data" that reflects many statistical properties of the original dataset and thus can be treated as a safe-to-release proxy for the original dataset. (For example, this can be done by estimating the parameters of a statistical model in a differentially private way, and then generating new data points using the model with estimated parameters.)

Agencies could also provide approved researchers with the OpenDP query interface to run differentially private analyses of interest to them on the data. The reason to limit such access to approved researchers is that every query made increases the privacy loss (ε) measured by differential

[5] https://www.northeastern.edu/csshresearch/bostonarearesearchinitiative/boston-data-portal/
[6] https://cusp.nyu.edu

privacy, and thus there is a finite "privacy budget" of queries that can be made while maintaining a desired level of privacy protection. The privacy budget could be quickly exhausted with a public query interface. On their own, differential privacy tools may not provide sufficient accuracy or support for the statistical methods that a researcher needs to use to obtain publishable results. In such a case, a differentially private query interface can be used for exploratory data analysis, for example for formulating hypotheses. The final analysis could then be carried out in a more controlled manner, for example in an enclave similar to Census Research Data Centers or by having an agency statistician vet and run the analysis for the researcher.

System Details

OpenDP System Properties Building a system that has broad use to a wide community across a spectrum of applications requires an encompassing, modular architecture that can support different materialized requirements. There are lots of choices inherent in any actual deployment of differential privacy to a particular use case—choices in architecture, algorithms, trade-offs between ease of use versus flexibility, even choices of what types of data to support. Our goal is to harness the knowledge and abilities of the community of experts to make decisions reflecting the state of the art, while maintain flexibility when relative trade-offs favor different styles of applications. In general, we want to architect OpenDP so that the system can be adapted in the longer run to have broad adoption and use, while not needlessly slowing immediate application to the currently prioritized use cases. From our experience building PSI, and intensely reviewing other differential privacy deployments, we describe some of the key system decisions needed and how our OpenDP platform will accommodate them. We also show the process by which we plan to grow in our OpenDP product lifecycle in Figure 1.

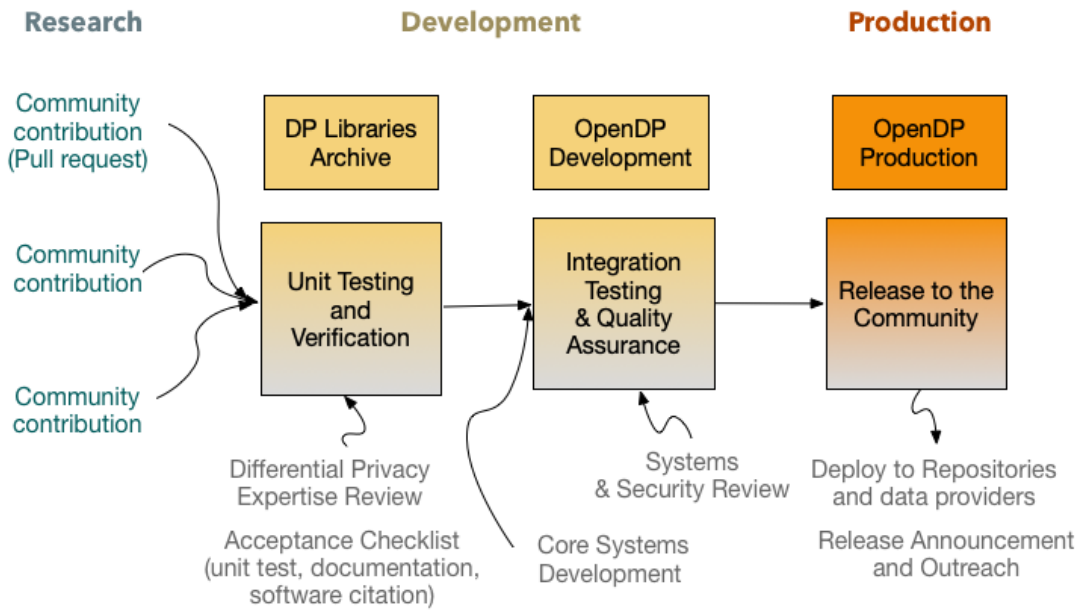

Figure 1: *OpenDP product lifecycle.*

Definitional choices: The most fundamental choices involve the very definition of differential privacy. At an abstract level, for an algorithm to meet the definition of differential privacy requires

that distribution of outputs be close from two neighboring datasets. The exact definition of *close* (which provides the quantitative measure of privacy loss), as well as the definition of *neighboring* (which captures the unit of privacy protection, e.g. individuals, households, relationships, etc.) can vary. We want to support the currently most common definitions of differential privacy, and then allow the OpenDP user to choose among these for their own use case.

- *Privacy-loss definitions:* Currently we envision supporting pure, approximate, Rényi and concentrated differential privacy, and be extensible to allow for including other versions in the future.
- *Neighbor relations:* We aim for the OpenDP library to support both the "delete one observation" and "change one observation" definitions of neighboring, especially as certain types of operations and types of algorithms are easier to work with under rival neighborhood definitions.
- *Composition:* The exact differential privacy definition employed defines how such releases compose. We would offer composition functions appropriate to the definitions that are covered in the library. We would also propose an object structure for batches of differentially private releases from the OpenDP library that automatically measures the composed privacy loss of the contained releases.

Algorithm and architecture choices: Another central choice is what differentially private algorithms to include and prioritize. The possible literature is vast to cover, but some choices and emphases we envision are:

- *Modularity:* Given that very many DP algorithms are built up from a core set of differentially private primitives, we would architect our library to isolate and facilitate the easy recombination of the underlying common primitive DP constituent methods. Drawing on the insights in systems such as PinQ [McSherry09], εktelo [ZMK+18], and our own PSI [GHK+16], we would separate transformations on the underlying sensitive data from methods that inject noise and produce safe differentially private releases from "post-processing" on those releases. This approach isolates trusted, privacy-critical code and simplifies the task of verifying that differential privacy is satisfied.
- *Simple summary statistics:* Exploratory data analysis (EDA) [Tukey77] is crucial to both releasing predetermined query workloads and allowing adaptive interactive learning for data discovery. Univariate summary statistics, and simple low-dimensional measures of relationships form the backbone of EDA, and would form a necessary emphasis in the OpenDP library.
- *Optimization:* For more complex statistical and machine learning models, generalized differentially private approaches to optimization often allow a model of interest to be constructed in a conventional manner, then optimized and released with privacy-preservation, and such optimizers would be another key focus [ACG+16, CMS11].
- *Uncertainty:* The quantification of uncertainty (e.g. through confidence intervals and p-values), and privacy-preserving hypothesis testing are necessary capabilities any DP system oriented toward open science research will have to support. We will emphasize including these abilities into the OpenDP library [GLRV18, KV18].

Data structure choices: Our present goal is to create a minimum viable product within one year that will provide secure, trusted releases with community vetted code, with a focus on supporting scientifically oriented research and exploration in the public interest. Towards this fast-paced goal, we initially plan to support datasets that are simple flat tables where the unit of observation is an individual data subject. However, many more complex forms of data require their own tailored algorithms (and even privacy definitions). We would plan to extend the system to the following types of data, prioritized based on the interests of the community of researchers and the needs of

the use cases, practitioners, and industry partners who join our community.

- *Relational data* is common in industry and administrative settings. The sensitivity of joins between datasets poses a difficult problem that recent work such as the Flex library [JNS18] and PrivSQL [KTM+19] have begun to address in practical systems.
- *Graph and network data* is key to studying many sensitive social phenomena such as social networks and communication, and presents unique challenges especially with regard preserving privacy at the level of the node (which typically corresponds to an individual data subject) [KNRS13, BBDS13].
- *Location data*, even aggregated to large regions, can quickly uniquely fingerprint an individual [DHVB13]. Pioneering work such as OnTheMap [AAG+09] was one of the first large scale deployments of differential privacy, and this continues to be an important application field [MIC+13].
- *Time-series and online data*, contains repeated observations over time of the same observations, causing unique challenges for composition of differential privacy. Algorithms for continuously updating data [DNPR10] and serially dependent observations [SCR+11] are key to any such use cases.

System exploit concerns: Another important aspect of building this library is verifying that the implementations actually satisfy the guarantees of differential privacy, rather than potentially leaking sensitive information due to bugs in the code or gaps between a programming language and its textbook abstraction. For example, issues that OpenDP will address include:

- *Side channels/Covert channels:* sensitive information can be inadvertently leaked through the the timing of executions or writing to system globals [HPN11].
- *Random numbers:* Many standard differentially private algorithms call for generating random numbers from continuous probability distributions (such as the Laplace distribution [DMNS06]). This can be a source of vulnerabilities because floating-point arithmetic is only an approximation to real-number arithmetic [Mironov12] and default random number generators are predictable [Krawczyk92]. Thus OpenDP will require the use of cryptographically strong random number generators, and make careful use of discretization (e.g. fixed-point arithmetic) that is explicitly analyzed for privacy, like in the recent work of PI Vadhan [BV18].
- *Verification:* in general, proving that an implementation of an algorithm in code matches the algorithm proved in theory can require great human expertise, but significant progress has been made in automated verification of implementations of differential privacy [BGHP17, ZK18, AH17], and we plan to use formal verification techniques as a resource and assistance to aid in the vetting of contributions to the OpenDP library.

Attention to Diversity

We are committed to gender and racial diversity in all the teams associated with this project: the security oversight committee, the DP oversight committee, the steering committee, the DP development team, the system development team. One of two principal goals is to create a vibrant community of researchers and practitioners, and such a community is successful only if it is welcoming, tolerant, fosters understanding and respect, and actively brings in a full diversity of experiences and viewpoints.

This approach has been a building block of constructing the team on the Privacy Tools Project. As a key example, we have used summer fellowships (undergraduate, graduate, postdoc) to form the collaboration communities for most of our projects, such as PSI. Over the six years of this program, we have been very active in recruiting fellows to bring diversity to our group, with a broad

distribution of our call for applications including specific outreach at/to the Tapia Conference on Diversity in Computing, Harvard's Women in Computer Science (WICS) organization, WECode (Women Engineers Code Conference), Women in Theory (WIT), and Lawrence Livermore National Laboratories Office of Strategic Diversity and Inclusion Programs. Year-by-year, about thirty percent of our project participants have been women and nine percent underrepresented minorities.

Project Participants for Privacy Tools
In years 1 & 2, of 55 students and postdocs, 15 were women and 6 underrepresented minorities.
In year 3, of 63 project participants, 18 were women and at least 8 were underrepresented minorities.
In year 4, of 69 project participants, 19 were women and at least 5 were underrepresented minorities.
In year 5, of 51 project participants, 16 were women and at least 3 were underrepresented minorities.
In our sixth year, of 19 project participants, 7 are women and at least 1 is from an underrepresented minority.

We will use a similar approach to outreach as we recruit for participation in OpenDP. We have given thought to diversity in the potential membership of our committees (listed in an earlier appendix). Of the 35 suggested members, 13 are women and 3 are underrepresented minorities.

As a declaration of values and a signal of who we aim to be as a community, we will have a code of conduct for our workshops that solidify our aspirations and commitments. The codes of conduct of the Society for Political Methodology [7] and their statement of diversity[8] partially modelled on the resources of the Geek Feminism Wiki[9] are exemplars of how we would construct our own code of conduct. Our partner in open-source scientific tools and community building, NumFOCUS, has successfully created a code of conduct [10] and reporting mechanisms we will discuss with them.

Beyond policies and towards action, attention to diversity in our community will be a primary objective of one of the funded positions in this proposal, the Open-source coordinator, charged with building a vibrant, healthy community of contributors and users.

Information Products Appendix

We plan to release a deployable differential privacy solution within the 12 months funded by this project, and to continue to grow this system over a futher 18 months. All the component code for this system will be publicallly available on GitHub and licensed under Apache Licence 2.0, as detailed in the sustainability plan. The core components released under this system are:

- *Library of DP Algorithms/Methods:* A library of DP methods will grow as the open-source community contributes new models. A new method will go through vetting managed by the DP oversight committee before it is released and supported through the API.
- *Budgeting Interfaces:* We propose to provide at least two interfaces to the system, for two types of users. The first is a clear intuitive graphical user interface (GUI) that provides a constrained workflow and detailed guidance to users that allows data owners and analysts with no privacy expertise to make informed choices about the balance between accurate answers and accumulated privacy loss. In addition, we propose to provide a more expressive programming interface that allows a sophisticated user to describe more sophisticated and customized statistical releases built up from basic differentially private primitives.
- *API:* A new API needs to be defined for this new service. The API will support submitting a DP request based on one of the supported DP methods in the library. The REST API should be secured by using an authorization protocol such as OAuth2.0. The API should be registered as a SmartAPI to follow best practices from the open-source community,
- *Containers:* The DP service will use a container (Docker, or similar technology) that holds the script of DP algorithms and requests to be applied to the data. The container is pulled from the location of the

[7] https://www.cambridge.org/core/membership/spm/about-us/diversity-and-inclusion/code-of-conduct-at-spm-events

[8] https://www.cambridge.org/core/membership/spm/about-us/diversity-and-inclusion

[9] https://geekfeminism.wikia.org/wiki/Conference_anti-harassment

[10] https://numfocus.org/code-of-conduct

data source. The analysis takes place in the data enclave where the sensitive data resides, and the raw data never needs to leave the enclave or data source. The results are placed in a location accessible by the DP service.

- *Authentication and Authorization:* OAuth2 protocol (or similar) will be used to authenticate and authorize users to use the API.
- *Large-scale data engine:* For large datasets, it might be necessary to deploy a Spark cluster or similar big data engine to run the DP algorithms in the data source location. In this case, the DP methods will need to be written to support Spark (or in particular SparkR) and take advantage of computing parallelism.

Sustainability Plan: This project adheres to community standards and National Digital Stewardship Alliance sustainability factors, applied to both software and data.[11] The partners in the project will work together to support diversified roles that can live beyond the span of the project. These are the roles or categories identified in successful data repositories and research software projects [Lee12].

- Service provider: *"Development, maintenance and support of a centralized preservation environment where other parties can transfer resources"*
- Enabler: *"Development, maintenance and support of software tools and systems that other institutions can install and run in their own environments"*
- Facilitator: *"Convening of forums for discussion and interaction among interested professionals, support for development of communities of practice, local testing of technical approaches to share experiences with others, development and dissemination of guidance documents"*

We have used these ideas and roles with success to promote the sustainability of the Dataverse project. In OpenDP we plan to expand them to support the sustainability of our privacy-sensitive tools and privacy tools integrated with Dataverse. This will be delivered in the following way:

- The Harvard Dataverse repository will initially serve as the *service provider* for any privacy-sensitive datasets described in the use cases, as well as additional privacy-sensitive datasets deposited by the research community. The IQSS Dataverse team, in collaboration with the Harvard Library and the Harvard University Information Technology, will provide the ongoing maintenance and support needed to sustain this centralized archival repository. As the system grows, more service providers can be added.
- All code for OpenDP (including libraries, APIs, GUI and programming interfaces and code for deployment) will be available through the GitHub repository, and distributed under Apache License 2.0. (Dataverse software also is distributed under Apache License 2.0.) Documentation with instructions on how to install the software will be provided with each release. Additionally, the OpenDP project aims to grow the underlying library into a shared community resource for privacy researchers and practitioners and will create extensive modes of developer and user support to stimulate community contribution including detailed developer/user websites, mailing lists and issue queues for opening the architecture and development goals, and a robust system of verification and testing for contributed code. Through all this, the DP development team and System development team will fulfill the *enabler* role in this project.
- The Open-source coordinator, located at IQSS, and working with the Steering committee and Principal DP scientist, and with the involvement of the broader OpenDP community, will fulfill the *facilitator* role in this project. Currently at IQSS the Dataverse team sustains and helps grow a community of users and developers. This includes 80 GitHub contributors, 75 community calls sp far, a community list (forums) with 534 members, and an annual community meeting with about 200 attendees. As relayed in the book [Millington12], when a community becomes too large and diversified, it is recommended to create sub-community groups. In this case, we propose to create sub-communities on "privacy tools for sharing data to support scientific research", following the same tactics used by the Dataverse community: an annual meeting co-hosted with (but separate from) the Dataverse Community meeting, a forum for discussion on privacy tools (associated with the current Dataverse community group), and leveraging the on-going Dataverse community calls to engage several times a year a group of users and stakeholders interested in privacy tools and sharing sensitive data.

We will also use the resources and advice provided by NumFocus to develop further our sustainability

[11] The Software Sustainability Institute. https://www.software.ac.uk/. At The University of Edinburgh. Also the National Digital Stewardship Alliance (NDSA). http://www.digitalpreservation.gov/ndsa/NDSAtoDLF.html. Library of Congress.

plan. We aim for OpenDP to become a NumFocus affiliated project and thus have access to its Sustainability Program, which, as defined in the NumFocus website, focuses on four main goals: 1) connect the NumFOCUS projects to each other to jointly develop and share information on sustainability strategies, 2) connect project leads with people with relevant expertise and networks, 3) provide training on skills related to open source sustainability, including business and financial planning, marketing strategies, community engagement, governance, etc.; and 4) support infrastructure that would help the projects more effectively manage finances, necessary technical resources, and client and business relationships.

This proposal funds core elements for the first year of this project. We have outlined a 2.5 year timeline to rapidly grow the system and reach expanded use cases with partners. We are investigating support from interested industry partners. Some of these dialogues are advanced, but are not yet at the stage of a funding agreement. We believe that an initial commitment by the Sloan Foundation to foster the community of experts will in turn attract industry support for other tasks and future years, and if successful this system could to continue to rapidly expand beyond our initial 2.5 year timeline. However, if we reach our deployment goals and then look primarily for maintenance and sustainability we project the estimated incremental cost for maintaining the OpenDP system is to fund the Open-source coordinator at one-half time, and two staff developers (one for library maintenance and development, and one for backend and deployment development).

Bibliography

[ACG+16] Abadi, M., Chu, A., Goodfellow, I., McMahan, H.B., Mironov, I., Talwar, K. and Zhang, L., 2016. "Deep learning with differential privacy." In Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security (pp. 308–318). ACM.

[AH17] Albarghouthi, A. and Hsu, J., 2017. "Synthesizing coupling proofs of differential privacy." Proceedings of the ACM on Programming Languages, 2(POPL), p.58.

[AWO+16] Altman, M., Wood, A., O'Brien, D., Vadhan, S., and Gasser, U., 2016, "Towards a Modern Approach to Privacy-Aware Government Data Releases" Berkeley Technology Law Journal, 3, 30. 2016.

[AAG+09] Andersson, F., Abowd, J. M., Graham, M., Wu, J., and Vilhuber, L. 2009. "Formal Privacy Guarantees and Analytical Validity of OnTheMap Public-use Data."

[BV18] Victor Balcer, V., and Vadhan, S., 2018. "Differential Privacy on Finite Computers." In Anna R. Karlin, editor, 9th Innovations in Theoretical Computer Science Conference (ITCS 2018), volume 94 of Leibniz Interna- tional Proceedings in Informatics (LIPIcs), (pp. 43:1–43:21), Dagstuhl, Germany, 2018. Schloss Dagstuhl Leibniz-Zentrum fuer Informatik. ISBN 978-3-95977-060-6. doi:10.4230/LIPIcs.ITCS.2018.43. URL http://drops.dagstuhl.de/opus/volltexte/2018/8353.

[BZ06] Barbarao, M., and Zeller, T., 2006. "A face is exposed for aol searcher 4417749." New York Times, page A1, 9 August 2006.

[BGHP16] Barthe, G., Gaboardi, M., Hsu, J. and Pierce, B., 2016. "Programming language techniques for differential privacy." ACM SIGLOG News, 3(1), pp.34-53.

[BBDS13] Blocki, J., Blum, A., Datta, A., and Sheffet, O., 2013. "Differentially private data analysis of social networks via restricted sensitivity." In Proceedings of the 4th conference on Innovations in Theoretical Computer Science (pp. 87-96). ACM.

[CMS11] Chaudhuri, K., Monteleoni, C. and Sarwate, A.D., 2011. Differentially private empirical risk minimization. Journal of Machine Learning Research, 12(Mar), (pp. 1069–1109).

[Crosas11] Crosas, M., 2011. "The Dataverse network: An open-source application for sharing, discovering and preserving data." D-Lib Magazine 17 (12). doi:1045/january2011-crosas.

[Crosas13] Crosas, M., 2013. "A data sharing story." Journal of eScience Librarianship 1 (3): 17379.

[MKHS15] Crosas, M., King, G., Honaker, J., and Sweeney, L., 2015. "Automating Open Science for Big Data." The ANNALS of the American Academy of Political and Social Science, 659, 1, Pp. 260-273.

[DHVB13] De Montjoye, Y. A., Hidalgo, C. A., Verleysen, M., and Blondel, V. D. 2013. Unique in the crowd: The privacy bounds of human mobility. Scientific reports, 3, 1376.

[DMNS06] Dwork, C., McSherry, F., Nissim, K., and Smith, A., 2006, March. Calibrating noise to sensitivity in private data analysis. In Theory of cryptography conference (pp. 265–284). Springer, Berlin, Heidelberg.

[DNPR10] Dwork, C., Naor, M., Pitassi, T. and Rothblum, G.N., 2010. "Differential privacy under continual observation." In Proceedings of the forty-second ACM symposium on Theory of computing (pp. 715–724). ACM.

[DSSU17] Dwork, C., Smith, A., Steinke, T. and Ullman, J., 2017. Exposed! A Survey of Attacks on Private Data. Annual Review of Statistics and Its Application.

[GHK+16] Gaboardi, M., Honaker, J., King, G., Murtagh, J., Nissim, K., Ullman, J. and Vadhan, S., 2016. Psi (Ψ): a Private data Sharing Interface. arXiv preprint arXiv:1609.04340. Also deployed prototype at http://psiprivacy.org/about.

[GLRV18] Gaboardi, M., Lim, H., Rogers, R., and Vadhan, S., 2016. "Differentially Private Chi-Squared Hypothesis Testing: Goodness of Fit and Independence Testing". Proceedings of the 35th International Conference on Machine Learning, Stockholm, Sweden, PMLR 80, 2018.

[Garfinkel16] Garfinkel, S., 2016. "De-Identifying Government Datasets (2nd Draft)" (No. NIST Special Publication (SP) 800-188 (Draft)). National Institute of Standards and Technology. http://csrc.nist.gov/publications/drafts/800-188/sp800_188_draft2.pdf.

[GAP18] Garfinkel, S., Abowd, J., Powazek, S., 2018. "Issues Encountered Deploying Differential Privacy." In Proceedings of the 2018 Workshop on Privacy in the Electronic Society (pp. 133–137). ACM.
[Golle06] Golle, P., 2006. "Revisiting the uniqueness of simple demographics in the US population." In Proceedings of the 5th ACM workshop on Privacy in electronic society (pp. 77–80). ACM.

[GS07] Gutmann, M., and Stern, P., 2007. "Putting People on the Map: Protecting Confidentiality with Linked Social-Spatial Data." National Academy Press, Washington, DC.

[HPN11] Haeberlen, A., Pierce, B.C. and Narayan, A., 2011, August. Differential Privacy Under Fire. In USENIX Security Symposium.

[HMM+16] Hay, M., Machanavajjhala, A., Miklau, G., Chen, Y. and Zhang, D., 2016, June. Principled evaluation of differentially private algorithms using dpbench. In Proceedings of the 2016 International Conference on Management of Data (pp. 139-154). ACM.

[JNS18] Johnson, N., Near, J. P., and Song, D., 2018. "Towards practical differential privacy for SQL queries." Proceedings of the VLDB Endowment, 11(5), (pp. 526–539).

[KV18] Karwa, V. and Vadhan, S., 2018. "Finite Sample Differentially Private Confidence Intervals." 9th Innovations in Theoretical Computer Science Conference (ITCS 2018).

[KNJ17] Kasiviswanathan, S., Nissim, K., and Jin, H., 2017. Private Incremental Regression. in the ACM SIGMOD/PODS Conference (PODS 2017).

[KNRS13] Kasiviswanathan, S. P., Nissim, K., Raskhodnikova, S., and Smith, A. 2013 . "Analyzing graphs with node differential privacy." In Theory of Cryptography Conference (pp. 457–476). Springer, Berlin, Heidelberg.

[KKNO16] Kellaris, G., Kollios, G., Nissim, K., and O'Neill, A., 2016. Generic Attacks on Secure Outsourced Databases. 23rd ACM Conference on Computer and Communications Security.

[King07] King, G., 2007."An introduction to the Dataverse network as an infrastructure for data sharing." Sociological Methods and Research 36:173–99.

[King14] King, G., 2014. "Restructuring the social sciences: reflections from Harvards Institute for Quantitative Social Science." PS: Political Science and Politics 47 (1): 16572. 2014.

[KTM+19] Kotsogiannis, I., Tao, Y., Machanavajjhala, A., Miklau, G. and Hay, M., 2019, January. Architecting a Differentially Private SQL Engine. 9th Biennial Conference on Innovative Data Systems Research

(CIDR 19), Asilomar, California, USA.

[Krawczyk92] Krawczyk, H., 1992. "How to predict congruential generators." Journal of Algorithms, 13(4), 527–545.

[Lee12] Lee, C., 2012. "States of Sustainability: A Review of State Projects funded by the National Digital Information Infrastructure and Preservation Program (NDIIPP)." http://www.digitalpreservation.gov/multimedia/documents/ndiipp-states-report032612_final.pdf

[MS00] Malin, B. and Sweeney, L., 2000. "Determining the identifiability of DNA database entries." In Proceedings of the AMIA Symposium (p. 537). American Medical Informatics Association.

[MS01] Malin, B. and Sweeney, L., 2001. "Re-identification of DNA through an automated linkage process." In Proceedings of the AMIA Symposium (p. 423). American Medical Informatics Association.

[McSherry09] McSherry, F.D., 2009, June. Privacy integrated queries: an extensible platform for privacy-preserving data analysis. In Proceedings of the 2009 ACM SIGMOD International Conference on Management of data (pp. 19-30). ACM.

[Millington12] Millington, R., 2012. "Buzzing Communities: How to Build Bigger, Better, and More Active Online Communities." 2012.

[MIC+13] Mir, D. J., Isaacman, S., Cceres, R., Martonosi, M., and Wright, R. N., 2013 . "Dp-where: Differentially private modeling of human mobility." In 2013 IEEE international conference on big data (pp. 580–588). IEEE.

[Mironov12] Mironov, I., 2012. "On significance of the least significant bits for differential privacy." In Proceedings of the 2012 ACM conference on Computer and communications security (pp. 650–661). ACM.

[MV16] Murtagh, J. and Vadhan, S., 2016, January. The complexity of computing the optimal composition of differential privacy. In Theory of Cryptography Conference (pp. 157-175). Springer, Berlin, Heidelberg.

[NS08] Narayanan, A., and Shmatikov., V. 2008. "Robust de-anonymization of large sparse datasets." In IEEE Symposium on Research in Security and Privacy, Oakland, CA, IEEE.
[NBW+18] Nissim, K., Bembenek, A., Wood, A., Bun, M., Gaboardi, M., Gasser, U., O'Brien, D., Steinke, T., and Vadhan, S., 2018. "Bridging the Gap between Computer Science and Legal Approaches to Privacy"In , 2nd ed., 31: Pp. 687-780. Harvard Journal of Law Technology.

[NSV16] Nissim, K., Stemmer, U., and Vadhan, S., 2016. "Locating a Small Cluster Privately." In PODS 2016. ACM SIGMOD/PODS Conference, San Francisco, USA, 2016.

[SCR+11] Shi, E., Chan, H.T.H., Rieffel, E., Chow, R. and Song, D., 2011. "Privacy-preserving aggregation of time-series data." In Annual Network Distributed System Security Symposium (NDSS). Internet Society.

[Sweeney97] Sweeney, L., 1997. "Weaving technology and policy together to maintain confidentiality." The Journal of Law, Medicine Ethics, 25(2-3), (pp. 98–110).

[Sweeney00]Sweeney, L., 2000. Uniqueness of simple demographics in the US population. LIDAP-WP4, 2000.

[SCB15] Sweeney, L., Crosas, M., Bar-Sinai, M. (2015). "Sharing sensitive data with confidence: The datatags system." Technology Science.

[Tukey77] Tukey, J., 1977. "Exploratory Data Analysis." Addison-Wesley.

[Ubaldi13] Ubaldi, B., 2013. "Open government data: Towards empirical analysis of open government data initiatives." OECD Working Papers on Public Governance, (22). https://doi.org/10.1787/5k46bj4f03s7-en

[VBS14] Veljković, N., Bogdanović-Dinić, S., and L. Stoimenov, L., 2014. "Benchmarking open government: An open data perspective." Government Information Quarterly, 31(2), (pp. 278–290).

[WAB+18] Wood, A., Altman, M., Bembenek, A., Bun, M., Gaboardi, M., Honaker, J., Nissim, K., O'Brien, D., Steinke, T., and Vadhan, S., 2018. "Differential Privacy: A Primer for a Non-technical Audience" Vanderbilt Journal of Entertainment Technology Law 21(1):209-275.

[ZK17] Zhang, D. and Kifer, D., 2017, January. "LightDP: towards automating differential privacy proofs." In ACM SIGPLAN Notices (Vol. 52, No. 1, pp. 888–901). ACM.

[ZMK+18] Zhang, D., McKenna, R., Kotsogiannis, I., Hay, M., Machanavajjhala, A. and Miklau, G., 2018, May. Ektelo: A framework for defining differentially-private computations. In Proceedings of the 2018 International Conference on Management of Data (pp. 115–130). ACM.

[ZJ14] Zuiderwijk, A., and Janssen, M., 2014. "Open data policies, their implementation and impact: A framework for comparison." Government Information Quarterly, 31(1), (pp. 17–29).