**Proposal Submission Template**

**for**

**Privacy Enhancing Technologies (PETs) for Public Health Challenge**

**Background and Overview**

data.org is partnering with a global financial institution, Harvard OpenDP, researchers from the Pontifical Javeriana University, and Sloan Foundation, to runa Privacy-Enhancing Technologies (PETs) for Public Health Challenge to create innovative differential privacy (DP) solutions to unlock privately held commercial sensitive data and enable optimal data-driven decision making in epidemiology.

**Privacy Enhancing Technologies (PETs) for Public Health Challenge**

A critical aspect in the early stages of an epidemic is the rapid epidemiological assessment to guide an effective public health response. The rapid identification of transmission dynamics is paramount for timely intervention and control measures. Although various methods have been developed and used in real-time response, there is still room for improvement of the precision, timeliness, and robustness of the technical solutions. Particularly, financial transaction data has emerged as a new opportunity to improve current epidemiological methods and tools to inform public health response. This challenge focuses on understanding possible correlations between financial transaction activities and public health behavior at different points in the pandemic, while preserving users' privacy. In this regard, we will focus on addressing two-part questions:

1. PART 1: Epidemiological (policy) decision-making – how to demonstrate the usefulness (informational value) of the remaining signals in the data for supporting real-world policy decision support in epidemiology.
2. PART 2: Privacy – how to demonstrate the quality and robustness of privacy-enhancing technology for unlocking privately held, commercially sensitive data.

**Response Format**

1. All the answers must be answered in English in the proposed format (.doc, .docx). Kindly name the submission document in the following manner:

**"<institute's name\_proposaltitle\_PETsChallenge>"**

1. The font style and size should be consistent in the response. (Font style: Arial; Size: 11 or 12pt).
2. **Using this template, all questions must be answered within a maximum of 6 pages (including any references or diagrams). We will not consider any content submitted as appendix or any additional pages beyond the 6-pages.**
3. **We require the following documentation to be included in the proposal submission (on a separate page)**
   * 1. **Annual budget**
     2. **Proposed budget tied to milestones.**
     3. **Entity information (w 8 or w9)**
     4. **Letter of determination (US = c3 status; university or non-profit)**

Thank you for your interest in collaborating with data.org on this important work.

|  |
| --- |
| **PART 1: Epidemiological decision-making** |

**Instructions**: Please answer **question #1** (unconstrained policy scenario) plus at least **1 additional question of your choice** in this section.

1. Policy Scenario 1 (Unconstrained)\*

* **How can privacy-enhanced transactional data be used to improve state-of-the-art epidemiological techniques commonly employed to inform public health response in real time?**
  + Context - In general, participants should focus on how to incorporate privately held financial data into epidemiological analysis, accessible through differential privacy mechanism ensuring users’ privacy, and aiming to design tools that allow for agile use of this information (from the financial data) jointly with open data sources to support real-time decision making.

|  |
| --- |
|  |

1. Policy Scenario 2 (Who infects who)

* **How can privacy-enhanced transactional data be used to inform contact patterns and who infects who matrix?** 
  + Context - In epidemiology, a **who acquires infection from whom (WAIFW) matrix** is a [matrix](https://en.m.wikipedia.org/wiki/Matrix_(mathematics)" \o "Matrix (mathematics)) that describes the rate of transmission of infection between different groups in a population, such as people of different ages, but can be extended to different social activities. Transaction data can have the potential to enhance our understanding of contact patterns by different age groups, profession and commercial activities which may be informative for infectious disease dynamics.

|  |
| --- |
|  |

1. Policy Scenario 3 (Effective reproduction number estimation)

* **How can privacy-enhanced transactional data be used to inform contact patterns, improve real time Rt estimations?**
  + Context - To assess the speed at which an infection spreads in a population is an important task when informing public health response to an epidemic. The instantaneous reproduction number (Rt) describes the average number of secondary cases generated by infectious individuals at a certain time assuming no changes to current conditions and it is commonly employed to characterize spread in real time.

|  |
| --- |
|  |

1. Policy Scenario 4 (Nowcasting)

* ***How* can privacy-enhanced transactional data be used to correct biases in Nowcasting estimations due to population behavioral changes?**
  + Context - Real time public health surveillance is subject to retrospective upward corrections due to the presence of occurred but not yet reported events, which reflects on the epidemic curves as a right truncation bias that should be corrected to enhance situational awareness and accurately inform public health officials and decision making. Statistical nowcasting methods aim to uncover current trends, predicting how strongly the preliminary data will be corrected once reporting catches up. Nowcasting estimations can be negatively affected, for instance, when the hospital system is overwhelmed or because of behavioral changes due to holidays.

|  |
| --- |
|  |

1. Policy Scenario 5 (Forecasting)

* ***How can* privacy-enhanced transactional data be incorporated as a predictor in forecasting epidemic curves such as cases, deaths or Rt while improving accuracy?**
  + Context - Forecasting is the use of current and past knowledge to predict future values or patterns in data within a prediction interval. During the COVID-19 pandemic, epidemic forecasting models were used to obtain predictions to inform timely decisions about healthcare systems needs or the implementation of non-pharmaceutical interventions to reduce transmission [3]. However, due to uncertainties about the underlying epidemic process, unpredictable human behavior or even future interventions, only short-term reliable forecasting can be made in real time. A customary practice to assess these limitations is to design scenarios based on well-defined sets of conditions to assist stakeholders and decision makers in long-term planning.

|  |
| --- |
|  |

|  |
| --- |
| **PART 2: Privacy** |

**Instructions**: Please answer all the following questions. However, if you think there are further considerations which should be included, please include them and they will be taken into account during the scoring.

1. Privacy-Utility tradeoff:  Provide the best evidence you can that your proposed method can achieve a good privacy-utility tradeoff, namely provide a statistical release that is rich and accurate enough to be useful for the policy decision support you describe below in (2), while providing strong differential privacy protections to the merchants represented in the data.  In making your case, describe and motivate an appropriate measure of utility for your method and its intended use (e.g. RMSE, classification accuracy, precision, and recall, etc.) and try to estimate it, for example via Monte Carlo experiments or back-of-the-envelope analysis.

Context –

* Differential privacy protections: we expect merchant-level differential privacy, where adjacent datasets differ in the addition or removal of records that are all associated with a single merchant. The dataset is broken up by merchants. We consider that a merchant is attached to a single location. It means that if a brand has multiple stores at different locations, each store is treated as a different merchant.
* Privacy measure: the choice of the privacy measure (zCDP, approx DP, etc.) is open. When making this choice, participants should have in mind that the privacy measure along with the privacy-loss parameter will have to be validated in later phases by the data curator.
* Implementation: solutions that have all of the privacy reasoning done by the OpenDP library are preferred. If the OpenDP library does not provide some privacy reasoning/mechanisms that are needed, contestants should implement them as new OpenDP transformations or measurements (and try to minimize the number of those so that we have fewer components that need manual review). OpenDP is currently available in Python but if needed, participants can assume that OpenDP will also be available in R to execute their project proposals in later phases of the challenge.

|  |
| --- |
|  |

1. Scalability: Provide the best evidence you can that your proposed method will be feasible to execute on datasets of the size and dimensionality as the Challenge dataset to be made available in Phase 2 and ideally beyond to larger datasets of a similar type.  Your evidence can include discussions of computing time, memory usage, parallelizability, etc.

|  |
| --- |
|  |

1. Alternatives: Compare your proposed method to other possible differentially private methods that could be used for the same problem, and explain why yours is the preferred choice.  Highlight any particularly novel design choices in your solution.

|  |
| --- |
|  |

1. Implementability in OpenDP: Describe the ways in which the OpenDP Library already supports some of the functionality you need, and the ways in which it would need to be extended to implement your solution.

|  |
| --- |
|  |

1. Statistical suitability: Evaluate the statistical properties of your method and the impact they may have on the intended application.  For example, does it provide biased estimates, and what might the impact of that bias be?  Might important subpopulations be drowned out by the level of aggregation and noise, and is this an inevitable consequence of privacy or something that might be avoided with other methods?

|  |
| --- |
|  |

1. Choice of data release:  With a limited privacy-loss budget, it is only possible to do a small number of statistical releases on the dataset.  Make the case that your proposal is important and valuable enough to be one of them.

|  |
| --- |
|  |

1. Responsible Technology: Please provide any additional considerations such as ethics, human-in-the-loop, replicability, auditability, and biases in the method you have chosen to solve the given policy problem.

Context –

* Responsible Technology: Generative AI is still an experimental technology and has known issues and limitations, such as tendency to "hallucinate", limited diversity in underpinning training data, lack of transparency and auditability in how it generates its responses, among others. Under these conditions we need to go further than mere adherence to high level principles and identify localized pertinent criteria for responsible applications of the technology.

|  |
| --- |
|  |