The NIST Differential Privacy Synthetic Data Challenge

NIST
National Institute of Standards and Technology
U.S. Department of Commerce

PSCR OPEN INNOVATION
Solving the Public Safety needs of tomorrow...today

Simons Institute: Foundations to Applications Workshop 3/2019

NIST Challenge

KNEXUS RESEARCH CORPORATION

NIST
National Institute of Standards and Technology
U.S. Department of Commerce

PSCR OPEN INNOVATION
Solving the Public Safety needs of tomorrow...today

Project Details

NIST-Challenge Oversight, PSCR Expertise, Metrology Expertise
Mary Theofanos (PI),
Terese Manley (Prize Manager)

Knexus Research -- Differential Privacy Expertise
Christine Task (NIST Technical Lead)

topcoder -- Phase II Challenge Platform,
Ward Loving (Project Manager),
Sergey Pogodin (Technical Lead)

HeroX -- Phase I Challenge Platform
Kyla Jeffrey (Project Manager)

Research Topic: Differentially Private Synthetic Data
Start Date: June 2018
Application: The First National Challenge in Differential Privacy

As technical lead for the **NIST Differential Privacy Synthetic Data Challenge**, Knexus is providing technical guidance for the first national challenge in differential privacy. Developments coming out of this competition are expected to drive major advances in the practical applications of differential privacy for contexts such as public safety.

Winners from Match #2 will be announced this week
Match #3 begins Next Week, on March 10th 2019. Registration is open now!
https://www.topcoder.com/community/data-science/Differential-Privacy-Synthetic-Data-Challenge

KNEXUS RESEARCH CORPORATION

PSCR Needs Data Analysis:

The **Public Safety Communications Research Division (PSCR)** of the **National Institute of Standards and Technology (NIST)** is sponsoring the Differential Privacy Synthetic Data Challenge to help advance research for public safety communications technologies for America's First Responders

PSCR OPEN INNOVATION

Solving the Public Safety needs of tomorrow...today

As first responders utilize more advanced communications technology, there are opportunities to use data analytics to gain insights from public safety data, inform decision-making and increase safety.

But… we must assure data privacy.

Differentially Private Synthetic Data Generation is a mathematical theory, and set of computational techniques, that provide a method of de-identifying data sets—under the restriction of a quantifiable level of privacy loss. Differentially private synthetic data sets can be safely released to the public, **allowing state and local public safety departments to leverage the power of crowd-sourced analysis to understand and improve their systems**.

NIST Challenge Motivation

Tech Challenges Have Advantages:

Challenges provide researchers with a visible, open and accessible**, shared pathway from theory to practice**.

Challenges grab attention. They **educate the public and potential investors** about new technological possibilities, inviting the audience to follow along with excitement as those possibilities are fulfilled.

Challenges often precede significant acceleration in the development of **commercial products** for new tech.

DARPA URBAN CHALLENGE

Google LUNAR XPRIZE

COLLEGIATE WIND COMPETITION
U.S. DEPARTMENT OF ENERGY

kaggle

UNMANNED AERIAL SYSTEMS
FLIGHT AND PAYLOAD CHALLENGE

PSCR OPEN INNOVATION
Solving the Public Safety needs of tomorrow...today

Overview of the NIST Differential Privacy Synthetic Data Challenge:

- The Challenge began in the summer of 2018 with a concept-building phase where contestants submitted concept papers proposing a mechanism to enable the protection of personally identifiable information while maintaining a dataset's utility for analysis.

- The second phase consists of a sequence of empirical matches throughout fiscal year 2019, where participants with implemented systems compete to produce high quality synthetic data from real data sets.

- This is the **first national challenge** in differential privacy, but ***it's already not the last***.

- The "**data challenge**" format is a well-established archetype, and for public accessibility it makes sense to echo this format… but hosting a ***differentially private*** data challenge requires some **non-trivial adaptations**.

- For the rest of this talk, we'll discuss the complications we encountered and how we chose to address each of them. These are certainly not the only possible solutions, and we invite your feedback and input. In general, we feel these **are important questions for the community to consider.**

NIST Challenge *Challenges*: Contest Procedure

Objective: Create a shared, competitive benchmarking process for Differentially Private Synthetic Data Generators.

Constraints: We… haven't really had one of those before. Contestants will be learning about the behavior of their solutions at the same time we do.

Conclusion: The contest needs to support teams iteratively evaluating, exploring and then refining and resubmitting their solutions.

Solution details: A Sequence of **Three topcoder Marathon Matches**, of increasing difficulty. Each Match has five weeks of **provisional leaderboard** scoring on submitted synthetic data sets, followed by three weeks rigorous **sequestered evaluation** of executable systems and source code, followed by a winners announcement and awarding of prizes.

To support iterative refinement without violating differential privacy, we assume provisional data is public, and keep sequestered data private (using data sets from different years/areas, with different individuals and different distributions).

NIST Challenge *Challenges*: Data and Scoring

Objective: Select the **data sets** to use as the basis for the contest.

Constraints: Data should be relevant to PSCR's applications, of interest to the public audience generally, reflect current synthetic data needs, publicly available (to avoid access restrictions), and it should start out as an achievable objective (not too many variables, not too many values per variable, not too complex correlations)… and then get harder.

Conclusion: **Event Data and Survey Data** (not time series data, transaction data, or image/audio data this time)

Solution Details:

Match #1 and Match #2: San Francisco Fire Event Data. Over a decade of data with features such as priority, response time, location, unit type, etc.

Match #3: [redacted]

NIST Challenge *Challenges*: Data and Scoring

Objective: A functional definition of whether one synthetic data set is 'better' than another.

Constraints: It has to run very efficiently, be fair, reasonably data independent, and we don't need only one… we need at least three *non-redundant* metrics, in order to increase the rigor of the scoring across each of the three matches. It should also capture the needs and preferences of the data user community.

Conclusion: Randomized Heuristics! Each match **adds a new scoring metric to the existing set.**

Solution Details:

Provisional Leaderboard Scoring is done on three submitted synthetic data-sets, generated at three specified values of epsilon, with the resulting three scores averaged together.

Sequestered Final Scoring is done with repeated trials, additional values of epsilon as needed, and final score is computed a privacy/accuracy AUC (Area Under Curve)

The same scoring metric is used in both the provisional and sequestered phase of each match:

Match #1: Randomized, normalized 3-Marginal based distance metric

Match #2: 3-Marginal, and randomized Row-pool similarity based metric

Match #3: 3-Marginal, Row-pool, and [redacted]

[Match #1 & #2 scoring metrics developed by Sergey Pogodin]

NIST Challenge *Challenges*: Differential Privacy Validation

Objective: Prevent ‘cheating’ from (often unintentional) violations of Differential Privacy, which will generally result in high accuracy scores. DP validation must keep the leaderboard reasonably reliable during the provisional phase, and make every effort during the sequestered phase to ensure prizes are only awarded to valid solutions.

Constraints: Even though some probabilistic black box DP verification systems exist, we didn’t have the time or budget to implement and adapt them to our needs on this particular project.

Conclusion: SME Review Panel. Relying on human resources instead, we needed to be as efficient as possible, and considerate of volunteers’ time.

Solution details:

Provisional Phase**:** To earn a **1000x score boost**, contestants must submit clear, complete privacy proofs to pass a **Differential Privacy Prescreen**, occurring as a weekly SME review telcon. The prescreen is a quick check to ensure the contestant is making a good faith effort to satisfy differential privacy and there are no obvious errors.

Sequestered Phase: Invited contestants submit source code, code guide/documentation, updated algorithm specification and privacy proof for a thorough **Final Differential Privacy Validation** by the SME review team. Solutions failing validation are eliminated from prize eligibility.

NIST Challenge *Challenges*: Algorithms That Exist As Software

Objective: NIST is part of the US Department of Commerce. NIST provides guidance that helps US corporations address technical needs. As a vital outcome of this contest, we would like to have stable, usable, well-engineered (ideally open sourced) software solutions that can be further evaluated by NIST experts, contributing towards NIST's efforts to issue official guidance on DP Synthetic Data.

Constraints: Teams may begin with academic prototypes that have only been used inside their research groups, to generate results for specific research papers.

Conclusion: Make initial participation in the match accessible for research prototypes, and increase code requirements over the course of the match.

Solution Details: Each Match begins with minimal software requirements (**simply submit correctly formatted synthetic data sets to earn a provisional leaderboard score**), and these are increased throughout the match: invitees to the sequestered round must have standardized delta/epsilon input, no hardcoded data schemas (schema given as input), and thorough code documentation aligned with algorithm documentation. Their solutions then undergo source code review by multiple SME, and their docker containers are run by the TC tech lead—If either encounter problems, they are informed and may be able to fix and resubmit. Prize-winners leave the match with money, but also with an **externally evaluated code base** that will be more easily shared, tested, and used by other researchers, potentially forming a stable basis for future production-level solutions. ***Participate in the contest and we'll provide a free two-month bootcamp for your DP Synthetic Data solution.***

NIST Challenge Match #3

Match #3 Begins Next Week, 3/10/19, Registration Is Open Now:

Data:

Provisional Phase Data—1940 Census Persons Level Data for Colorado

Sequestered Phase Data—1940 Census Persons Level Data for [Not Colorado]

Scoring:

- To Catch **Long Tails**: Income Inequality based metrics
- To Catch Degradation of Accuracy over **Differences of Differences**: Pay-gap based metrics

Final Outcomes:

- $62K in prizes for Match #3!
- $4K bonus for top 5 challenge winners who provide their full solution in an open source repository for use by all interested parties
- Further NIST research, as a metrology lab, to identify and establish metrics and methods for evaluating synthetic data
- Goal of disseminating the lessons learned and approaches taken, through journal special issue, conferences, workshops, talks. Challenge participants will be invited to contribute

Register Here: https://www.topcoder.com/community/data-science/Differential-Privacy-Synthetic-Data-Challenge

(…or just google "NIST Differential Privacy Challenge")

NIST Challenge Motivation

PSCR Research Priorities:

PSCR OPEN INNOVATION

Solving the Public Safety needs of tomorrow...today

NIST's **Public Safety Communications Research** division has strong commitments to both **public safety research** and the preservation of security and privacy, including the use of de-identification.

It is well known that privacy in data release is an **important area for the Federal Government (which has an Open Data Policy), state governments, the public safety sector and many commercial non-governmental organizations**. Developments coming out of this competition would hopefully drive major advances in the practical applications of differential privacy for these organizations.

PSCR is sponsoring this exciting data science competition to help advance research for public safety communications technologies for **America's First Responders**