Project: Evaluation of Noise Infusion for Large-Scale Demographic Sample Survey (SDRN)

Knexus Research LLC
1951 Kidwell Drive, Suite 240,
Vienna, VA 22182
1-855-569-7373

Project Information:

Project No.	ADC-SDRN-23-N02
Contract #:	ADC-2023-605-01
PoP:	10/1/2023 - 9/9/25

Corporate Official:
Mr. Adam Lurie, CEO, (202) 306-0806, adam.lurie@knexus.ai
Principal Investigator (PI):
Dr. Christine Task, christine.task@knexus.ai
Report Prepared By:
Dr. Christine Task, christine.task@knexus.ai

Project Partners:

1. Tumult Labs, 201 W. Main Street, Suite B26 Durham, NC 27701, Tel: 919-444-2973 || *www.tmlt.io* || POC: Dr. Ashwin Machanavajjhala
2. Strategix LLC, 1309 Coffeen Ave, STE 1200, Sheridan, WY 8801, Ph 623 565 9541 || Dr. Jess Stahl

Table of Contents

SDRN Synthetic Data Report 3
Project Overview: Balancing Disclosure Avoidance with Utility 3
Final Candidates Data Deidentification Methods Summary 5
Evaluation of Data Deidentification Approaches 6
Metric Categories 6
PCA (Distribution Shape) Difference on SDR 6
Pairwise Correlation Differences on SDR 7
Linear Regression Differences on SDRN 8
Privacy Metrics on SDR 9
Performance Summary 9
Recommendations 9
Appendix A: Results for Baseline Deidentification Algorithms 10
Reduced Baseline 50-Feature Schema Overview 10
Baseline Deidentification Algorithms 11
Utility and Privacy Metrics 12
Evaluation Results 13

America's DataHub Consortium (ADC), a public-private partnership, implements research opportunities that support the strategic objectives of the National Center for Science and Engineering Statistics (NCSES) within the U.S. National Science Foundation (NSF). These results document research funded through ADC and are being shared to inform interested parties of ongoing activities and to encourage further discussion. Any opinions, findings, conclusions, or recommendations expressed above do not necessarily reflect the views of NCSES or NSF. Please send questions to ncsesweb@nsf.gov. NCSES has reviewed this product for unauthorized disclosure of confidential information and approved its release (NCSES-DRN26-018).

SDRN Synthetic Data Report

Project Overview: Balancing Disclosure Avoidance with Utility

This report summarizes the findings of the Survey of Doctoral Recipients Noise Infusion (SDRN) project, funded by the National Center for Science and Engineering Statistics (NCSES) as part of the National Secure Data Service Demonstration (NSDS-D) project. The SDRN project informs the use of privacy preserving methods in disclosure for the broader application of noise infusion for sample surveys. The report focuses on utility and privacy evaluations for the three top candidate privacy solutions selected by stakeholders and NCSES staff, and the project's final recommendations. The appendix includes evaluation results from a much larger set of baseline algorithms that were initially explored on a smaller data set.

The National Center for Science and Engineering Statistics (NCSES) manages a number of data products, such as the Survey of Doctoral Recipients (SDR). This data is incredibly useful for a variety of purposes, including policy design, economic analysis, and training artificial intelligence models. However, when publishing public microdata, privacy is a significant concern due to the risk of reidentifying individuals.

The variables released in public microdata can be seen as belonging to two groups: **Confidential Features** and **Public Fingerprint Features**. Confidential Features are sensitive pieces of information not typically available to the public, such as salary, job satisfaction, or the reason for a job change. By contrast the Public Fingerprint consists of feature values that are easy to discover from publicly available sources like CVs, social media, or personal websites. This includes information like age, marital status, or employer sector.

When an individual has a unique combination of values in their public fingerprint, their record can be reidentified using public information. Simply reducing the detail of features is often not enough to prevent this, as most individual records remain unique across the many features in a comprehensive dataset. And, while it may not protect privacy, reducing detail does reduce the utility of microdata for analysis.

Modern privacy approaches, however, can allow datasets to contain more detailed information while still maintaining data privacy. Currently, the Survey of Doctorate Recipients (SDR) public-use file and other publicly available SDR information products use multiple disclosure methodologies to protect the confidentiality of SDR respondents, including top-coding and reducing feature detail. This project explored noise infusion to augment other disclosure limitation methodologies currently in use with the SDR public-use microdata file

Synthetic data and noise infusion (differential privacy) were explored as means to increase data accessibility while maintaining strong data privacy. Synthetic data aims to retain the original statistical distributions and structure of the dataset while protecting individual privacy. Differential privacy aims to eliminate the risk of re-identification through data linkage and other privacy attacks. A defining characteristic of differential privacy is that given a sufficiently strong

privacy parameter, individuals cannot be re-identified. Thus, the use of synthetic data and noise infusion offers the potential to strengthen data privacy protection for SDR respondents while enabling the inclusion of more information in the SDR public-use file, expanding access to useful data for researchers and the public.

Top Data Deidentification Methods Summary

This section provides a brief introduction to the three privacy solution candidates that were the focus of the final phase of the project. Appendix A provides results for the initial eight methods evaluated during the baseline phase.

Cell Suppression

- **Method:** Redacts data by suppressing identifiable outlier features. It retains records that are not considered identifiable outliers.
- **Privacy:** Addresses quasi-identifiers.
- **Performance:** On SDR data, more than 80% of records see some suppression. It fails to meaningfully capture relationships for smaller groups.

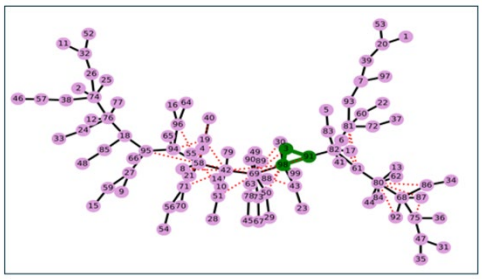
Partially Protected with Primary Suppressions

Table 2	Age					
	0-20	21-40	41-60	61-80	81+	Total
Marital Status						
Married	4	22	28	26	D	82
Separated	D	5	7	D	0	16
Divorced	0	8	9	9	3	29
Widowed	0	D*	5	10	9	25
Never Married	36	7	6	3	D	53
Total	42	43	55	50	15	205

* D is used to indicate suppressed cells

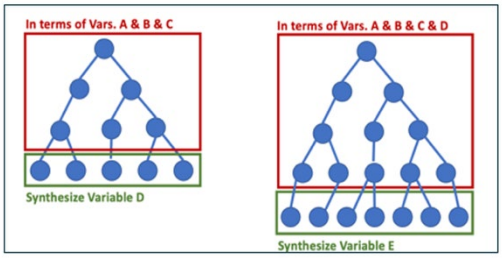
Source: National Center for Education Statistics (NCES)

DP-PGM (Differentially Private Probabilistic Graphical Model)



- **Method:** Learns data by creating a node for each feature and connecting it to other features based on the probability of influence. It generates new records by traveling across these nodes, with values based on feature connections.
- **Privacy:** This method is differentially private, meaning it adds random noise to the probabilities.
- **Performance:** Effective at maintaining the privacy of records, but struggles to capture more complicated feature relationships.

Decision Tree Based Synthetic Data



- **Method:** Learns data by creating a tree for each feature that understands its relationship to all other features. It then generates new data by replacing feature values with new, likely values from the tree.
- **Privacy:** The size of the tree is limited to prevent overfitting.
- **Performance:** Can maintain relationships between features even for large datasets. It provides marginally less privacy protection than DP-PGM but is significantly better than cell suppression.

Evaluation of Data Deidentification Approaches

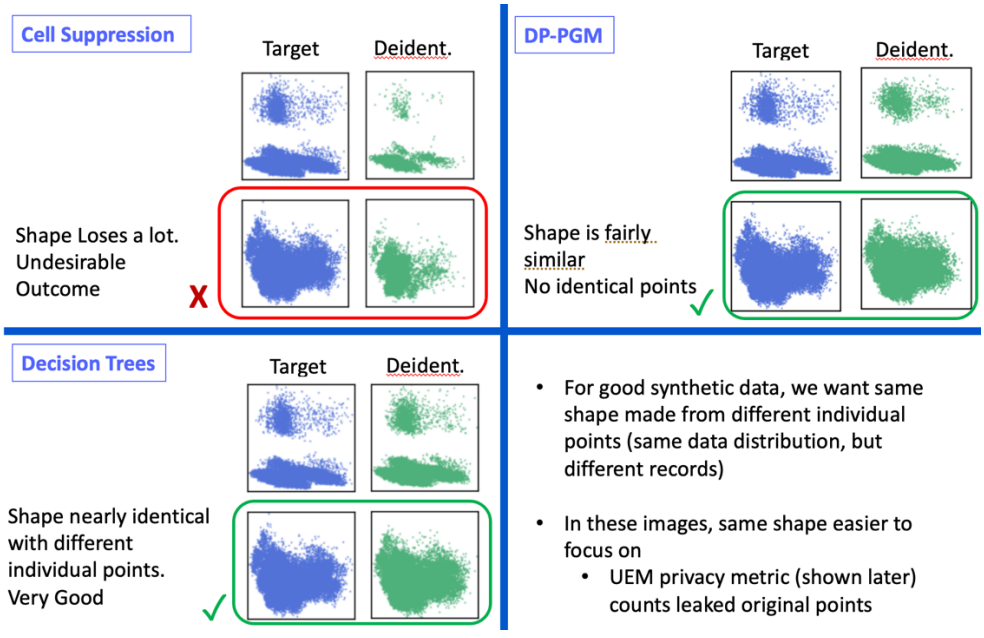
To evaluate the fitness of these privacy approaches for use, a suite of tools and metrics was built to examine both **fidelity** (how similar the synthetic data is to the original) and **privacy performance**. Evaluations were performed using the 250-feature SDR Public Use Microdata.

Metric Categories

- **Overall Fidelity:** How similar does the data distribution look compared to the original data?
- **Regression Utility Metrics:** Compares analyses to each other to see how similar their results are.
- **Privacy Metrics:** How effectively is reidentification prevented?

PCA (Distribution Shape) Difference on SDR

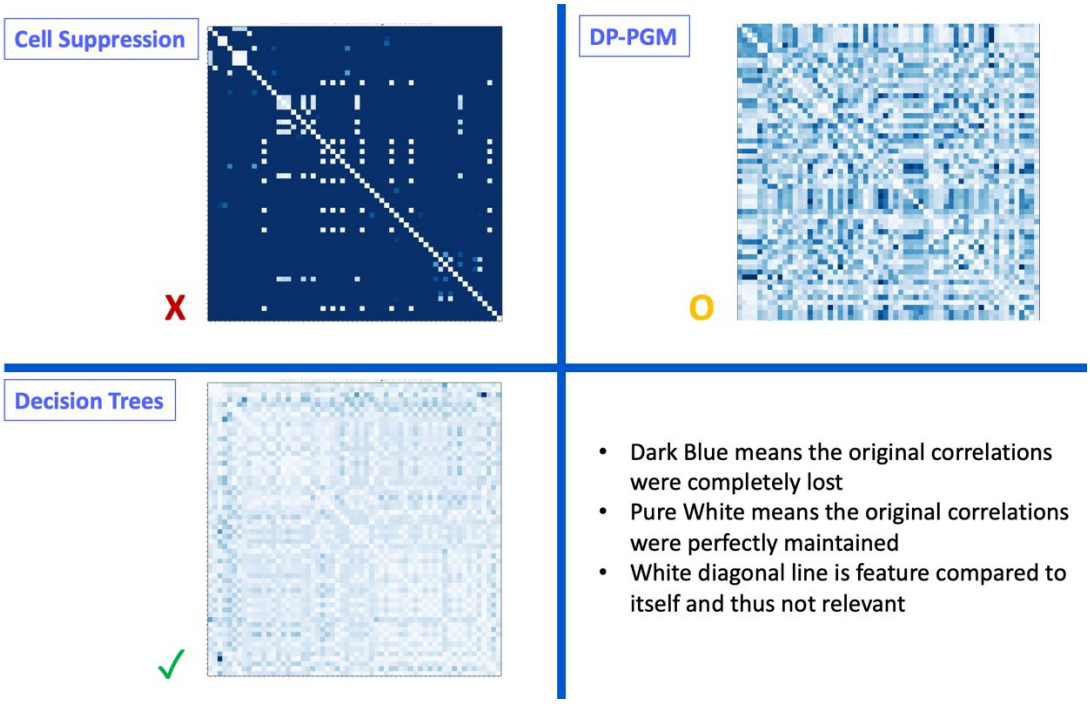
This metric measures data fidelity by using Principal Component Analysis (PCA) to create 2D plots that show the overall shape of the 250-feature Public Use Microdata (PUMS) data. The goal is to compare the shapes of the original and deidentified data. The ideal outcome for good synthetic data is to have the same shape with different individual points, meaning the data distribution is the same but the records are different.



- **Cell Suppression:** The shape loses a lot of information, which is an undesirable outcome.
- **Decision Trees:** The shape is nearly identical, which is a very good outcome.
- **DP-PGM:** The shape is fairly similar, and no identical points are present.

Pairwise Correlation Differences on SDR

This metric measures data fidelity by showing how accurately the privatized data maintains the original relationships between features. In the heatmap below, whiter blocks mean better-maintained correlations, while dark blue means the original correlations were completely lost. The white diagonal line is a feature compared to itself and is not relevant. These evaluations show the results of the 50 worst-performing features for each deidentification approach; the full evaluation reports include pairwise correlations across all 250 features.



- **Cell Suppression:** The image shows a large number of dark blue blocks, indicating that correlations were lost.
- **Decision Trees:** This heatmap is almost entirely white, showing that correlations were very well maintained.
- **DP-PGM:** This heatmap shows some correlation loss but is much better than cell suppression.

Linear Regression Differences on SDRN

This metric measures regression utility by showing the difference in linear regression results between the original and deidentified data. The regression shown here predicts income based on employment category. This indicates how similar the results of a statistical analysis on the two datasets will be. The goal is for the regression lines of the two datasets to match perfectly. The heatmap in this analysis shows the difference in data distribution: purple means too few records were created, while orange means too many. The ideal heatmap would be completely white. In the full evaluation reports multivariate logistic regressions were performed to predict every categorical variable class with sufficient individuals, with similar results.

- **Cell Suppression:** The purple heatmap and divergent lines show that too many people were suppressed, and the regression does not hold closely.
- **Decision Trees:** The matching lines and white heatmap indicate that statistical analysis will yield nearly identical results, even at the extremes.
- **DP-PGM:** The orange heatmap and divergent lines indicate that added noise created too many new people, and the regression does not hold closely.

Privacy Metrics on SDR

These metrics measure resistance to reidentification. It includes a **Membership Attack Accuracy** score, which is the success rate of guessing if a record was in the original dataset using a k-nearest neighbor based attack (an ideal score is 50%, or random guessing), and a **Unique Exact Match Leakage** score, which is the percentage of leaked records that were in the original dataset (an ideal score is 0%).

- **Cell Suppression:**
  - **Membership Attack Accuracy: 73.2%**
  - **Unique Exact Match Leakage: 19.1%**
  - **Result:** An attacker can identify a fair number of records, and about 20% of original records were maintained in the deidentified dataset.
- **Decision Trees:**
  - **Membership Attack Accuracy: 49.5%**
  - **Unique Exact Match Leakage: 0%**
  - **Result:** The attacker is guessing randomly, and no exact records were maintained.
- **DP-PGM:**
  - **Membership Attack Accuracy: 50.2%**
  - **Unique Exact Match Leakage: 0%**
  - **Result:** The attacker is guessing randomly, and no exact records were maintained.

Performance Summary

- **Cell Suppression:** This method loses information on every feature correlation and results in a messy data distribution that is easy to reidentify. With 80% of records losing some information, it has poor overall performance.
- **DP-PGM:** This method is effective at preventing reidentification, but it struggles with preserving correlations and has more difficulty with smaller subpopulations.
- **Decision Trees:** This method is the best performer. It maintains correlations well and handles odd feature values effectively. It also prevents reidentification better than cell suppression while maintaining a higher number of records.

Conclusions and Recommendations

After review of these results and consultation with Stakeholders and the NCSES Disclosure Review Board, it was decided to recommend Partial Synthesis with Decision Trees as the final privacy solution for the Survey for Doctoral Recipients. Partial Synthesis uses a synthetic data generator to produce new replacement values for the public fingerprint features that are most likely to be leveraged in a reidentification attack. Like Cell Suppression, it does not change the values of confidential features, so the majority of the data retains its exact ground truth values for analysis. But unlike Cell Suppression, vulnerable features are replaced rather than suppressed, so smaller subpopulations remain available for analysis. Partial Synthesis via Decision Trees is currently in production use on the American Community Survey. Partial synthesis provides weaker privacy protection than differential privacy but reduces practical reidentification risk.

Appendix A: Results for Baseline Deidentification Algorithms

Here we summarize the initial evaluation results of baseline deidentification algorithms for the Survey of Doctoral Recipients (SDRN).

Reduced Baseline 50-Feature Schema Overview

All deidentification algorithms were evaluated first on a reduced-size baseline dataset consisting of the following SDR features (including merged checkbox features A16-A36). These were chosen as having the strongest mutual correlations and the most utility for analysts:

- **A16:** academic positions
- **A22:** job-required technical expertise at the bachelor's level or higher in natural sciences, social sciences, or other fields
- **A25:** primary and secondary reasons for taking a postdoc
- **A29:** the primary and secondary reasons for working outside the field of the highest degree
- **A36:** the impact of COVID on salary and how it is reflected
- **AGEGRP:** Age Group (5-year intervals) (recoded for public use)
- **BTHUS:** Place of birth (U.S./Non-U.S.)
- **CTZN:** Citizenship or visa status
- **DIFAGEGR:** Physical abilities: earliest age experienced difficulties (5-year intervals)
- **EARNP:** Total earned income before deductions in previous year (recoded for public use)
- **EMSECDT/EMSECSM:** Employer sector (detailed/summary codes)
- **GENDER:** Gender
- **MARSTA:** Marital status
- **MINRTY:** Underrepresented minority (URM) indicator. Excludes Non-Hispanic White and Asian
- **OBSNUM:** Observation number
- **RACETHMP:** Race/ethnicity (recoded for public use)
- **SALARYP:** Salary on principal job (annualized and recoded for public use)
- **WKSLYR/WKSWK:** Number of weeks worked last year/per week
- **CHSCH:** Reason for changing employer or job: school-related reasons
- **CHFAMCOV:** Reason for changing employer or job: family-related reasons due to COVID
- **CHLAYCOV:** Reason for changing employer or job: laid off or job terminated due to COVID
- **CHRET:** Reason for changing employer or job: retired
- **CHOT:** Reason for changing employer or job: other
- **CHFAM:** Reason for changing employer or job: family-related reasons
- **CHLOC:** Reason for changing employer or job: job location
- **CHCON:** Reason for changing employer or job: working conditions
- **CHPAY:** Reason for changing employer or job: pay, promotion
- **CHCHG:** Reason for changing employer or job: change in career/professional interests
- **CHLAY:** Reason for changing employer or job: laid off/job terminated
- **WRKG:** Working for pay or profit during reference week
- **LOOKWK:** Not working, looking for work
- **EMUS:** Employer location (U.S./Non-U.S.)
- **CHUN12:** Children living in household indicator: under age 12
- **APROD:** Work activity on principal job (10% indicator): production, operations, maintenance
- **WAMGMT:** Work activity on principal job (10% indicator): managing or supervising people/project
- **WAAPRSH:** Work activity on principal job (10% indicator): applied research
- **EMED:** Indicator for educational institution employer
- **WASALE:** Work activity on principal job (10% indicator): sales, purchasing, marketing

- **WABRSH:** Work activity on principal job (10% indicator): basic research
- **WAOT:** Work activity on principal job (10% indicator): other
- **WAQM:** Work activity on principal job (10% indicator): quality or productivity management
- **WATEA:** Work activity on principal job (10% indicator): teaching
- **WACOM:** Work activity on principal job (10% indicator): computer applications
- **WAACC:** Work activity on principal job (10% indicator): accounting, finance, contracts
- **WASVC:** Work activity on principal job (10% indicator): professional services
- **WADSN:** Work activity on principal job (10% indicator): design
- **WAEMRL:** Work activity on principal job (10% indicator): human resources
- **WADEV:** Work activity on principal job (10% indicator): development

Baseline Deidentification Algorithms

The baseline SDRN evaluations explored a wider variety of potential deidentification solutions. Approaches that performed poorly were not continued to the larger 250 feature data set.

Traditional Statistical Disclosure Control:

- **Cell Suppression:** Outlier records are surpressed according to quasi-identifier features. Library: SDCMicro

Non Differentially Private Synthetic Data:

- **Decision Trees:** One tree model is trained for each feature in the data, then records are synthesized one feature at a time.
  Library: R Synthpop (the more computationally efficient XSyn library was used for the 250 feature solution)
- **Gaussian Copula:** The entire original distribution is used to fit a high-dimensional Gaussian Copula ball, which is then used to generate new data. Library: Synthetic Data Vault, FastML
- **TVAE:** An auto-encoder translates the data to a distribution-dependent encoding space and then selects new records from that space, "new ChatGTP like AI". Library: Synthetic Data Vault, Tabular Variational Auto-Encoder (TVAE)

Differentially Private Synthetic Data:

- **MST:** Uses a simple co-occurrence probability network to capture feature relationships, with added noise to protect privacy. Library: SmartNoise (OpenDP), Maximum Spanning Tree (MST)
- **AIM:** A workload-adaptive algorithm that iteratively selects the most useful measurements to approximate the input data. Library: SmartNoise (OpenDP) Adaptive Iterative Mechanism (AIM)
- **Tumult Baseline DP-PGM & with Logical Skips:** An "**infant version**" of a custom Tumult solution that accurately captures univariate distributions and skip logic, and will be tailored for SDRN data in the future.

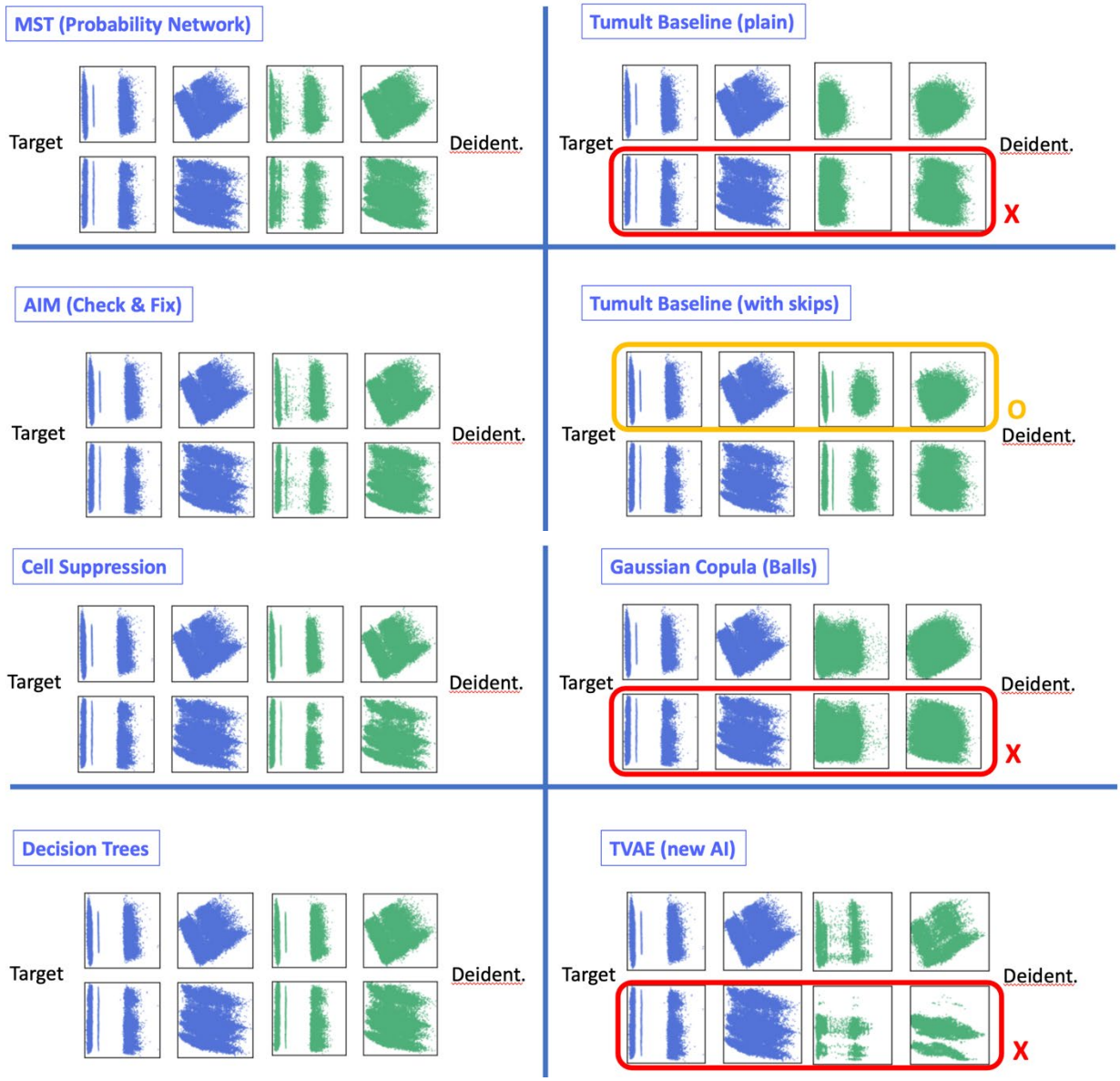
Utility and Privacy Metrics

These metrics are the same as those defined above in the main report, with the addition of Logistic Regression Accuracy Comparison which performs a very comprehensive evaluation of data deidentification's impact on regression analysis (which was not computationally tractable on the larger feature set).

- **Overall Utility Metrics:**
  - **PCA:** Uses dimension reduction to directly plot the data distribution in 2 dimensions.
  - **Pairwise Correlations**: Pearson's correlation difference between the original data and the deidentified data.
- **Regression Utility Metrics:**
  - **Income vs Employer Sector Linear Regression and Heatmap**: Selected linear regression accuracy, as compared to the ground truth accuracy.
  - **Logistic regression accuracy comparison**: Multivariate regression accuracy across all categorical features and all classes with sufficient records (>100), as compared to the ground truth data accuracy.
- **Privacy Metrics:**
  - **Unique Exact Match metric:** Checks for exact unique record reproduction.
  - **Membership Attack metric:** Measures membership inference rates for the most vulnerable individuals, using a KNN based attack.

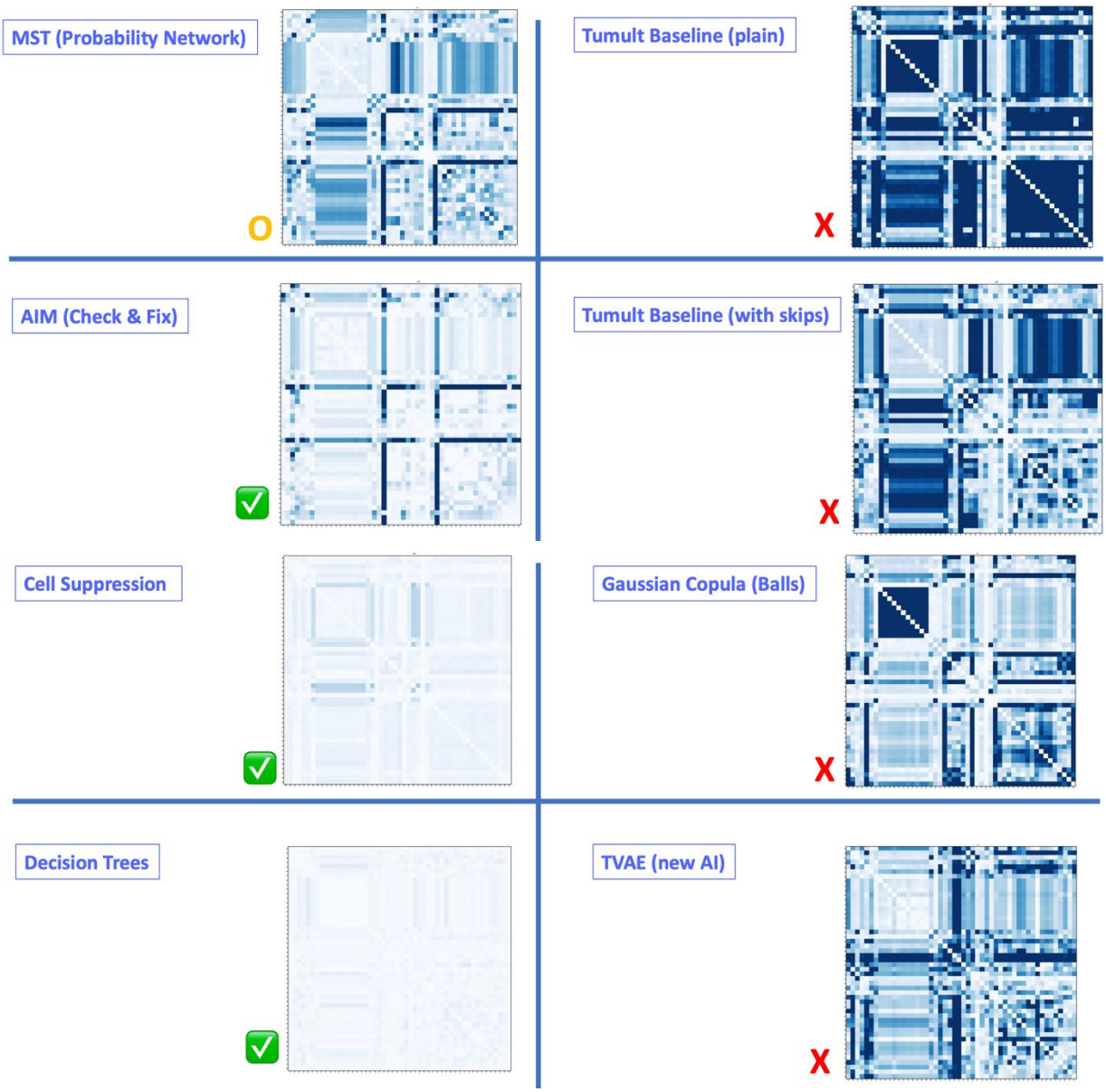
Evaluation Results

PCA



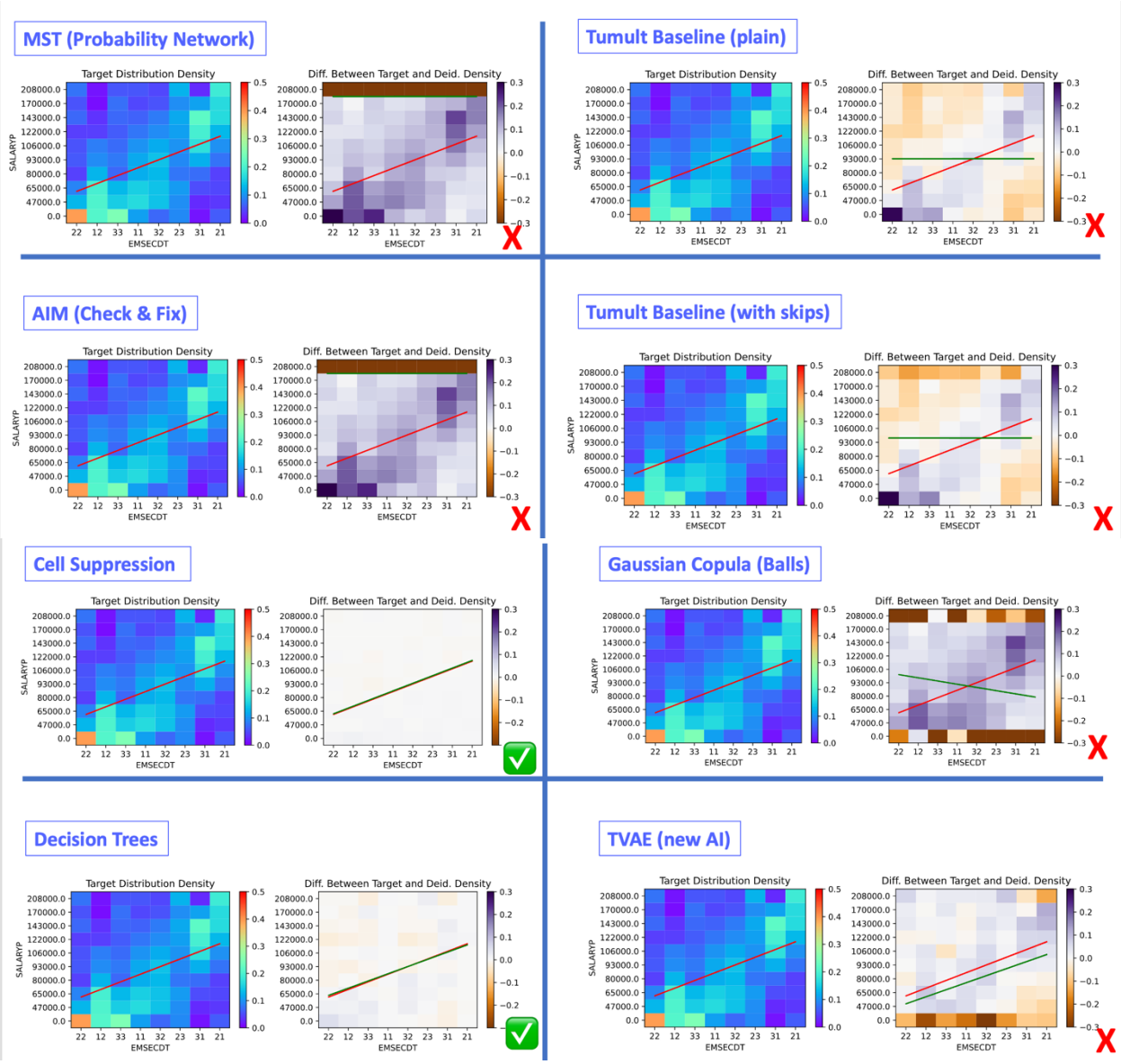
- The non-differentially private synthetic data techniques differed significantly from each other in terms of their ability to retain the original distribution shape. The Gaussian Copula (Synthetic Data Vault FastML) approach removed most of the original structure, and the TVAE approach introduced both error and new structure that wasn't in the original data. Meanwhile, the Decision trees retained the original shapes identically.
- This earliest version of the Tumult algorithm (plain) also had difficulty maintaining the structure. The second version (skips) improved on this by encoding all logical zeroes from the original data.

Pairwise Correlations



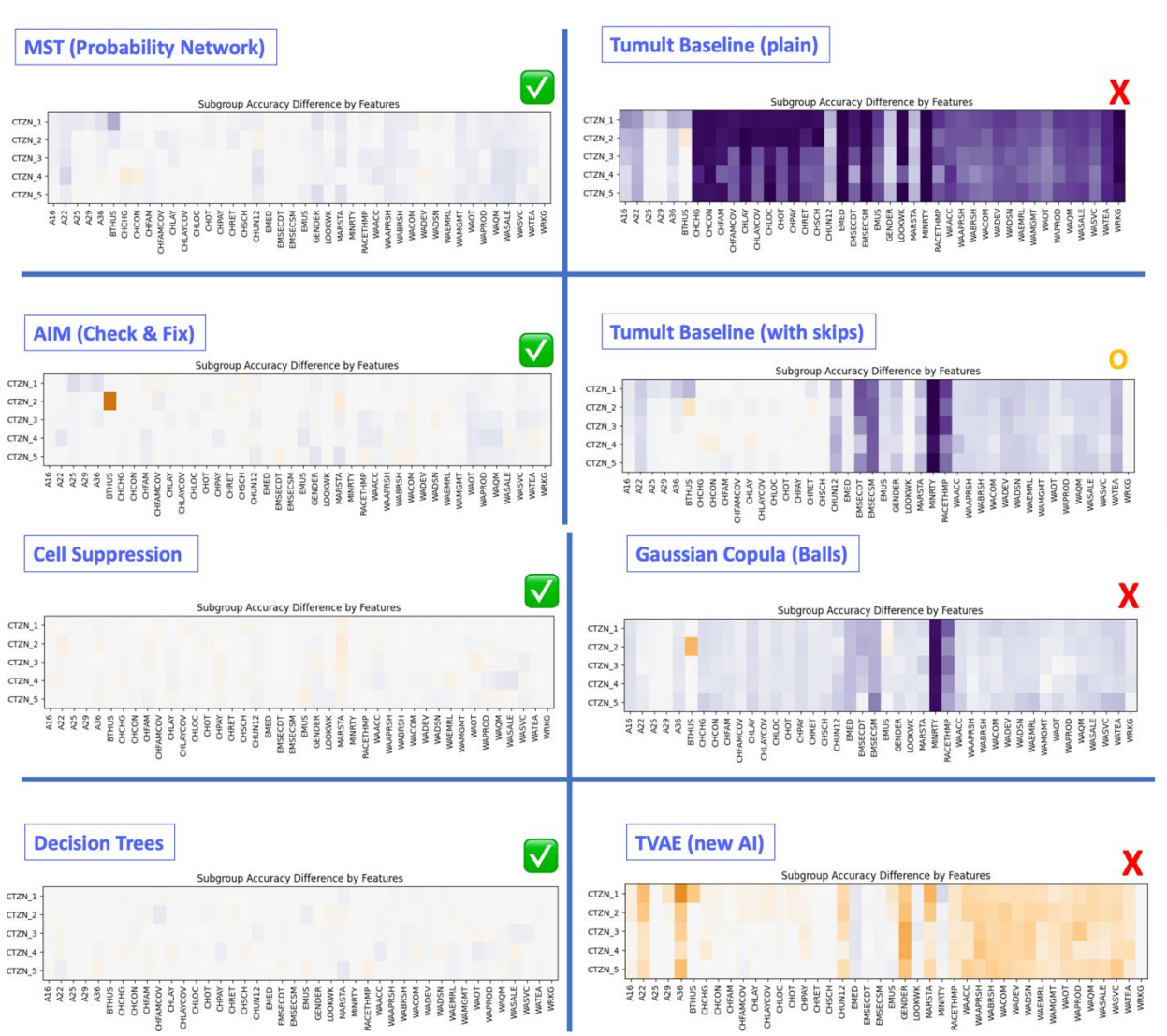
- The TVAE and Gaussian Copula non-DP synthetic data methods had difficulty maintaining correlations; as seen in the PCA results they weren't able to retain the distribution of the original data.
- Decision trees outperformed cell suppression very slightly because they do not suppress data and thus don't lose correlations for subpopulations.
- The mature differential privacy methods (AIM, MST) outperformed the baseline Tumult solutions, but all differential privacy methods had difficult retaining correlations with numerical features.

Linear Regression Heatmap



- The heatmaps show the relationship between salary and employer sector. The regression lines for the deidentified data often deviated from the target data regression line.
- Decision trees performed well, and cell suppression matched exactly. Other methods had varying levels of difficulty.

Regression Accuracy



- The heatmaps for regression accuracy showed differences for various features and citizenship statuses. Purple indicates lower accuracy in the synthetic data as compared to the ground truth data (due to lost feature relationships), while orange indicates higher accuracy. Multivariate regression accuracy is computed across all categorical features and all classes with sufficient records (>100).
- Decision trees, cell suppression and two mature differentially private methods (AIM, MST) performed well.
- TVAE, a non-DP synthetic data method, simplified the relationships in the data until logistic regression was actually much *more* likely to be successful on the synthetic data than on the real data. Results on the synthetic data here would produce overconfidence about relationships in the real data.

Privacy

Exact Unique Record Reproduction:

- **Non-DP Methods:** The unique exact record leakage rate was 93% for Cell Suppression, 4% for Decision Trees, and 0% for both Gaussian Copula and TVAE.
- **DP Methods (epsilon = 10):** The unique exact record leakage rate was 0% all four differentially private techniques.

Most Vulnerable Membership Attack (KNN):

- **Non-DP Methods:** The attack accuracy for a pool size of 100 was 83% for Cell Suppression, 75% for Decision Trees, and 54% for both Gaussian Copula and TVAE. An attack accuracy of 50% means the attack failed.
- **DP Methods (epsilon = 10):** The attack accuracy was 51% for MST, 43% for AIM, 59% for Tumult Baseline (plain), and 46% for Tumult Baseline (with skips) for a pool size of 100.

Baseline Evaluation Outcomes

Stakeholders selected Cell Suppression, DP-PGM (the mature Tumult algorithm using a similar approach to MST but with additional measures for preserving important correlations) and Decision Tree Based Synthesis as the three methods to move forwards for evaluation over the full 250 feature data.

When moving from 50 baseline features to the final 250 features presented in the main report, Decision Trees significantly improved their performance on privacy, due to additional trees introducing additional randomness to the algorithm. The final tuned Tumult algorithm, DP-PGM, also improved significantly in utility from its original baseline version. Cell suppression performed poorly on utility and privacy with the larger feature set. These results are presented in the main report.