DEPARTMENT OF STATISTICS SINGAPORE
Empowering You with Trusted Data

Data Exploration and Research: Use of Synthetic Data in the Singapore Department of Statistics

Dr. Andy Chiang,
Chief Technology Officer, Group Director

At Asia-Pacific Stats Café
"Balancing Access and Confidentiality: The Role of Synthetic Data in Official Statistics"

16 September 2024



DOS's Transformation as First Stop for Data and Data Services

Changing Environment

- Constantly changing economic and social landscape
- Ever-increasing data flows and new technologies, and more emphasis on data-driven decision making
- Greater evidence-based policy formulation with agencies equipping themselves with data capabilities

DOS's Key Role and Key Outcomes

DOS as Capability Centre of Individual and Business Data (CapCen)

(i) Leading National Statistical Office
(ii) Trusted Centre for Individual and Business government administrative data

Data Value Chain

- Collection
- Processing & Integration
- Compilation
- Analyses
- Dissemination & Engagement

DEPARTMENT OF STATISTICS SINGAPORE
Empowering You with Trusted Data

Two legal frameworks provide the mandate for DOS to collect and share data, the first being the Statistics Act

Statistics Act 1973

Primary legislation governing statistical activities in the public sector; provides mandate for data collection and safeguards the confidentiality of information collected.

This Act empowers DOS to share the data collected under the Act as:

i. **aggregated data**; or
ii. **anonymised microdata** to public agencies, researchers commissioned by public agencies and researchers holding full-time appointments in Singapore's Autonomous Universities.

DOS shares aggregated data through



SingStat Website
Informs users on data releases and statistical information, and promotes understanding and exploration with infographics and dashboards



SingStat Table Builder
Facilitates access to over 2,200 data tables from 70 public sector agencies with over 90,000 series for data-driven decision making, research and analyses



SingStat Mobile App
Enables convenient access to commonly referenced data while on-the-go, presenting data in easily understandable charts

DOS shares anonymised microdata through



Anonymised Microdata Access Programme (AMAP)
Allows research using anonymised microdata in a secured environment, and supports complex analyses, e.g., advanced econometrics and regression studies on various economic, social and manpower topics

STATISTICS SINGAPORE
Empowering You with Trusted Data

Second, Public Sector (Governance) Act (PSGA) on administrative data within government for policy making and service delivery



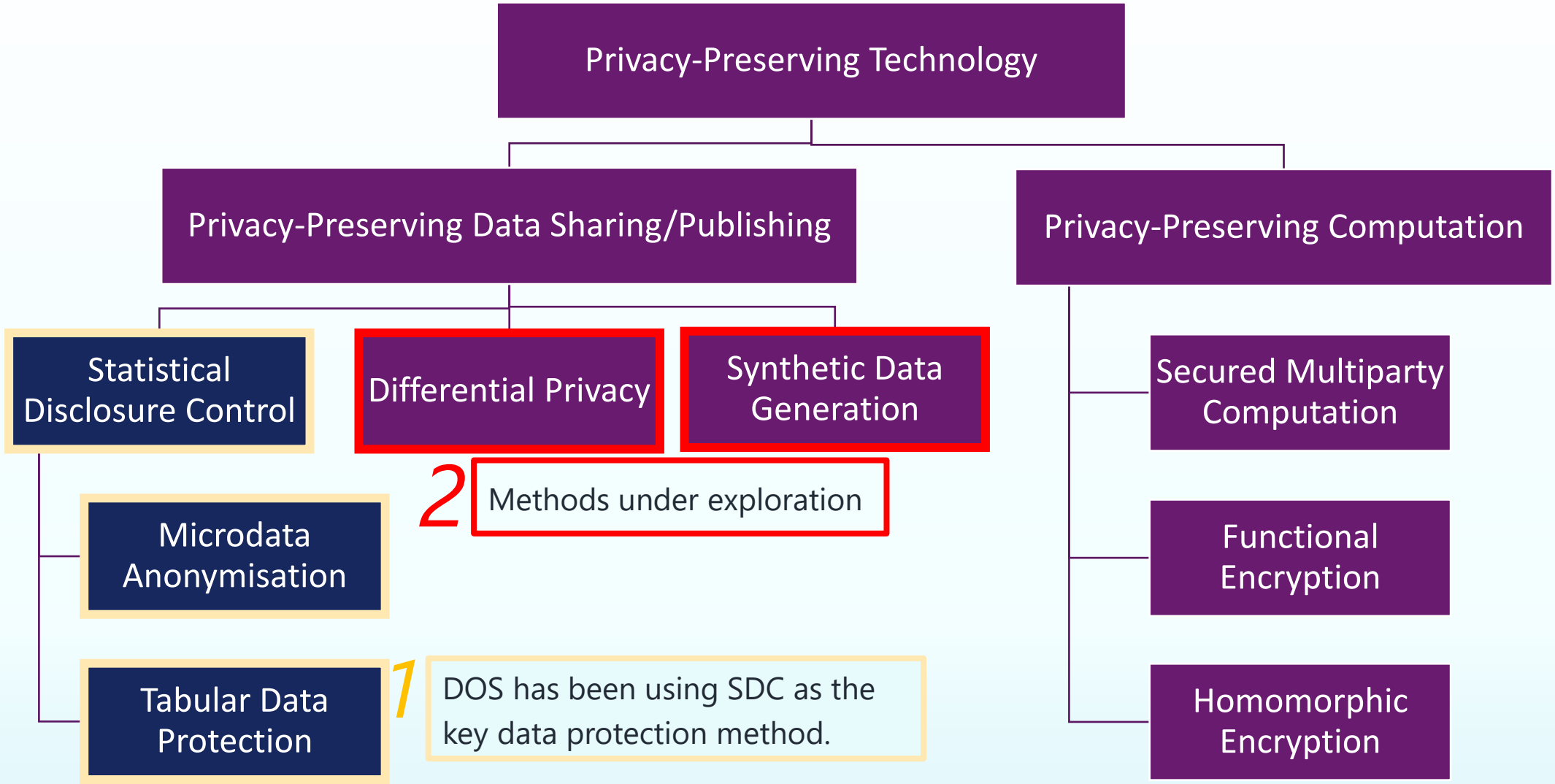
Public Sector (Governance) Act 2018

Enables data sharing within government for policy making and service delivery, while imposing strong safeguards such as criminal penalties on public officers who commit acts of data misuse or disclosure without authorisation, including the unauthorised re-identification of anonymised data.

DOS Trusted Centre for Individual and Business Data empowers government agencies to self-discover data insights for policy planning, operations and service delivery by providing public officers with access to a broad range of PSGA core administrative data.

DOS TC does not hold or share data collected under the Statistics Act.

DOS's Privacy-Preserving Technology Framework



Differential Privacy and Synthetic Data

- DOS has developed some capability on *differential privacy (DP) data and synthetic data*
- As differentially private synthetic data and synthetic data are not attributable to the records in the original data, they hold promise to improve data sharing while still protecting against the disclosure of confidential information
- Synthetic data are currently shared with government agencies for data exploration and facilitating longitudinal data research projects

Differential Privacy

- 2 approaches for implementing DP:
  1. DP aggregate statistics
  2. DP synthetic data
- Synthetic data approach is generally easier:
  - No need to split the privacy-loss budget over the statistics to be generated.
  - Maintains consistency between all statistics generated.
- We worked with the National University of Singapore N-Cript Research Centre to test their DP synthetic data generation algorithm.
- DOS has presented DP research at ISI World Congress 2023 on "*Privacy on a Subset of Dataset Variables*"

Synthetic data samples for longitudinal data research projects



DOS Longitudinal Data Analytics use case:
To facilitate data exploration and advanced modelling scripts from researchers

- DOS supports government agencies on longitudinal data research projects in multiple ways:

  (1) Self-service using DOS Remote Query (RQ) application to obtain aggregated statistical tables/regressions without seeing the underlying records

  (2) Self-service using anonymised microdata in the data lab

  (3) Service request where DOS officers perform the analysis and provide aggregated results

- For (3) service request, the government agencies researchers can request for synthetic data sample in advance so that they can develop more advanced econometric, longitudinal modelling scripts for DOS officers to run on the actual records and provide the aggregated results
- This gives the researchers a good idea how the actual dataset is structured, (instead of only referencing the metadata), reducing potential errors in their scripts before sending to DOS.
- DOS officers can run the codes more smoothly on the actual dataset, and this reduces project turnaround time.

Synthetic data samples for Trusted Centre Data Exploration

DOS Trusted Centre use case: To facilitate data exploration

Small samples of synthetic data are produced for more highly-demanded core administrative datasets.

The method of generating synthetic data (Regression Trees in Synthpop in R) is used to ensure that statistical aggregates are representative of the original dataset at the broad level and not at the detailed level.

Agencies use the synthetic data samples for initial data exploration and then request for the actual data if they decide to pursue the study.

Users are cautioned that the small samples are not for generating statistics due to reduced data utility.

How is the synthetic data generated?

1 **Pre-processing**
Categorical variables with many categories and low frequency count in some categories are grouped up to reduce generation run time.

2 **Generation**
"Synthpop" package in R and classification and regression tree models to generate variables. The visit sequence of the synthesis is defined using the "varrank" package in R.

3 **Post-generation processing**
- *Derived fields.* Derive the synthesised value of the derived fields from the synthesised raw fields to preserve cross-field relationships
- *Consistency checks and processing*

4 **Evaluation**
- *Utility and trend checks. V*erified through data utility assessments, summary statistics

DEPARTMENT OF STATISTICS SINGAPORE
Empowering You with Trusted Data

DEPARTMENT OF STATISTICS SINGAPORE
Empowering You with Trusted Data



Thank you

Our Vision

National Statistical Service of Quality, Integrity and Expertise

Our Mission

We Deliver Insightful Statistics and Trusted Statistical Services that Empower Decision Making

Connect with us:

www.singstat.gov.sg

info@singstat.gov.sg

@SingStat

@singstat_dos

@SingStatvideo

@sg-department-of-statistics