

GOVTECH
SINGAPORE



Data Science &
Artificial Intelligence
Division

PEPR '23

Privacy in the Public Sector : Lessons Learned and Strategies For Success

Alan Tang, Anshu Singh

Syahri Ikram, Zul Yang, Ivan Wang, Toh Wang Ting,
Ameera Adam, Jason Chew, Raine Tan, Xuanhui Lee

Government Technology Agency, Singapore

Sep 11, 2023

GovTech spearheads digital transformation within Singapore's public sector.



CLOAK (former “enCRYPT”) empowers public to implement **policy-compliant PII detection** and **data anonymization quickly** and **with confidence**.



Guide

Sign In



CLOAK serves as the privacy backbone for the Singapore public sector – supporting both **individual** and **enterprise** use cases.

Anonymise your healthcare dataset in minutes.

Cloak is a one-stop, self-service Central Privacy Toolkit for the Singapore Public Sector. Cloak makes it simple to transform and anonymise your data so you can innovate faster, whilst complying with policy guidelines and managing privacy-utility trade-offs.

Explore our toolkit

user_id	user_name	user_gender	user_dob	user_postal_code
NRIC	Name	Others	Select Info Type	Select Info Type
Direct Identifier	Direct Identifier	Sensitive Data Field	Select Sensitivity Type	Select Sensitivity
55489272J	Ryan	M	9/6/51	106909
	Richard	M	6/7/57	973920
	Krystal	F	13/2/88	800621
	Janet	F	18/7/16	429024
	Tommy	F	7/5/93	388492
	Gary	F	12/7/80	
	Zachary	F	16/5/10	
	Adam	M		
	Peter	F		
	Kathleen	M		
	Eric	M		

INFORMATION LOSS

93.94 %

MAX RE-ID: 50%

The loss of utility due to a (with generalisation) or (with suppression) of indirect use.

The worst case scenario, in which the equivalence class with the highest re-identification probability is assumed to represent the entire dataset.



Apply all recommendations

- Data Fields**
- DIRECT** NRIC (PSEUDONYMISE)
 - DIRECT** Name (REMOVE)
 - INDIRECT** Gender (K_ANONYMITY)
 - INDIRECT** BloodType (K_ANONYMITY)
 - SENSITIVE** Disease (SUPPRESS)
 - INDIRECT** Age (K_ANONYMITY)
 - INDIRECT** Zipcode (MASK)
 - SENSITIVE** BMI (RETAIN)
 - SENSITIVE** DateOfDiagnosis (GENERALISE)

NRIC

Information Type: NRIC
Sensitivity Type: Direct Identifier
Data Type: TEXT

Recommendation: Apply PSEUDONYMISE transformation technique

Transformation Technique: Pseudonymise (SALT Optional) [Apply]

Pseudonymisation de-identifies values by replacing them with cryptographically generated values (garbled text). The values are generated by the irreversible hashing with salt method. [Learn More](#)

For example, **S2345631E** transforms into **b9c1a87768f5a738f6361cfa23ec972b2704cc7885178417bafcd8303d2ab5ba**

Preview

Raw Data	Transformed
T0581222F	c80...
F0005772O	a42...
T7303748A	280...

MOCK DATA GENERATION

Mock Data

Data Fields (3 FIELD ITEMS)

Field Label	Field Type
Profile Group	sg_profile_group (FIELD GROUP) Generate an SG Profile (Name + Race + Gender), 1 column each with customisabl...
Occupation Group	sg_occupation_group (FIELD GROUP) Generate Occupation + Salary + Education Level, 1 column each with customisabl...
Phone Number	sg_numbers Generate a Singaporean phone number with +65 prefix

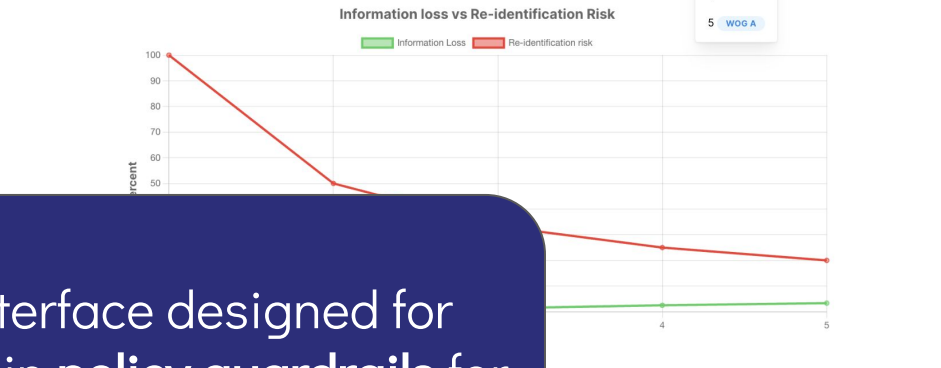
Preview

Name	Race	Gender	Occupation	Salary	Education Level	Phone Number
Chin Boon Choo	Chinese	M	Graphic Designer	3273.37	Professional Qualification and Other Diploma	+6591203094
Hian Eng Lim	Chinese	F	Electrician	3839.75	Professional Qualification and Other Diploma	+6590169316
Suriani Binte Syukri	Malay	F	Teacher	3958.92	Post-Secondary (Non-Tertiary)	+6590602261
William Johnson	Eurasian	M	Environmental Health Officer	2804.78	Polytechnic Diploma	+6594095439
Chia Leng Ong	Chinese	F	Writer	2372.29	Bachelor's or equivalent	+6583431121
Hoon Lian Chng	Chinese	F	Mechanical Engineer	4304.21	Polytechnic Diploma	+6597449954

Indirect Identifiers: Gender BloodType Age

Risk Scores Metrics Compare & Preview

The advanced view provides detailed information on the risk and utility metrics for applying k-anonymity. This view is intended to help you understand the tradeoff and is not a necessity to proceed. For a better understanding of the metrics, please refer to the guide here.



User friendly web interface designed for non-experts, with built in policy guardrails for automated compliance.

Madam,

I am writing to appeal to the Singapore Ministry of Manpower on behalf of my wife, <PERSON> (<SG_NRIC_FIN>), whose work pass is due for renewal. Her date of birth is <DATE_TIME>. Our <LOCATION> address is <SQ_ADDRESS>. My mobile number is <PHONE_NUMBER> and <PERSON>'s number is <PHONE_NUMBER> or you can email us at <EMAIL_ADDRESS>.

We are in urgent need of your assistance in renewing Harin's work pass, as it has been expiring soon. She is employed in the Technology sector and has been an integral part of her company since arriving from Korea.

Despite our best efforts, we have been unable to complete the renewal process on our own, and we are now seeking your help. We have followed all necessary procedures and submitted the required documents, but we have not received any response from the authorities.

I would like to request that the Ministry of Manpower take immediate action to resolve this matter as soon as possible. <PERSON>'s employment is vital to our family's financial stability, and we are in a difficult situation without her income of \$3000 a month. We would not want her bank account to be frozen due to the lack of a work pass. Her bank account is 199-52346-9 at DBS Bank. According to your website, <https://www.mom.gov.sg/contact-us>, it will take about 15 business days, but we hope that it can be shorter by having us link up to expedite the process.

I kindly ask that you provide us with any guidance or assistance necessary to expedite the process.

Used by **over 80** Singapore government agencies.

Land Transport Authority

MSF
MINISTRY OF SOCIAL AND FAMILY DEVELOPMENT



National Environment Agency
Safeguard • Nurture • Cherish

MINISTRY OF MANPOWER



Monetary Authority of Singapore

PUBLIC SERVICE DIVISION
PRIME MINISTER'S OFFICE

INLAND REVENUE AUTHORITY OF SINGAPORE

HTX

SINGAPORE POLICE FORCE
SAFEGUARDING EVERY DAY

GOVTECH SINGAPORE

>1000 datasets anonymized

>1 million API calls

>10 million PII detected and replaced

Est >10,000 man hours saved

... supporting dozens of **LLM** and **data analytics** use cases



#1. Privacy is Hard, But The Risks Are Real

Good privacy explainability is just as important as good privacy engineering.

What We Learned: The average user finds data anonymization **slow** and **technically challenging** - but **necessary**.

Intuitively understands the **need** for **data privacy**

We need to protect our citizens' privacy!
(and I don't want to get in trouble)

Which laws / policies / guidelines are relevant?

Confused by **ambiguous** and **evolving policy** environment



Just wants to get the **job done!**

So... can I share my data?

What does "hashing" or "tokenization" mean?

Not familiar with **data privacy** or **security concepts**

I can't run Python on my work PC!

Lack access to **commercial statistical tools**

What We Did: Built a no-code web interface for non-experts focusing on explainability and simplicity.

1) Select access mode: ⓘ

Refer to the [IM8 Guidelines](#) for more info.

WOG Mode A

For Public Access

- Unrestricted access with little control over end-user.
- Examples: Datathons, light analytics
- Treatment: Heavy anonymisation and mandatory re-identification test

WOG Mode B

For Conditional and Restricted Access

- Controlled access with good control and monitoring over end-user
- Examples: Policy or program evaluation
- Treatment: Lighter anonymisation with optional re-identification test

Custom

For Customised Transformations

- Free-play over all anonymisation techniques
- Examples: Inter-agency sharing, ad-hoc transformations
- Treatment: User-defined anonymisation with optional re-identification test

Data Type

TEXT

Recommendation

Apply PSEUDONYMISE transformation technique

Transformation Technique

Pseudonymise

SALT

Optional

Apply



Pseudonymisation de-identifies values by replacing them with cryptographically generated values (garbled text). The values are generated by the irreversible hashing with salt method. [Learn More](#)

For example, **S2345631E** transforms into

b9c1a87768f5a738f6361cfa23ec972b2704cc7885f78417bafcd8303d2ab5ba

Policy guardrails and recommendations

What We Did: Built a no-code web interface for non-experts focusing on explainability and simplicity.

The screenshot displays the enCRYPT web interface. At the top left, there is a blue button labeled "Apply all recommendations". Below it is a "Data Fields" panel with a list of fields and their status:

- DIRECT** NRIC (PSEUDONYMISE)
- DIRECT** Name (REMOVE)
- INDIRECT** Gender (K_ANONYMITY)
- INDIRECT** BloodType (K_ANONYMITY)
- SENSITIVE** Disease (RETAIN)
- INDIRECT** Age (K_ANONYMITY)
- INDIRECT** Zipcode (RETAIN)
- NON-SENSITIVE** BMI (RETAIN)

The main content area features a help section titled "How does enCRYPT apply anonymisation?" explaining k-anonymity. Below this, "Indirect Identifiers" are selected as Gender, BloodType, and Age. A "k-value" dropdown is set to 1, with a tooltip showing options 1, 2, 3 (WOG B), 5 (WOG A), and 5 (WOG A). A "Suggested k-value for WOG A" tooltip points to the 5 (WOG A) option. A blue button "Apply further anonymisation" is also present.

Navigation tabs include "Risk Scores", "Metrics", and "Compare & Preview". The "Compare & Preview" view shows two panels:

- Transformed Data:** k-value: 1. Status: NOT WOG MODE A COMPLIANT. Metrics: 500 (41.8 % unique rows), 12 (2.4 %) SAFE ROWS (compliant with the k-value of 5), 488 (97.6 %) RISKY ROWS (non-compliant with the k-value of 5).
- Anonymised Data:** k-value: 5. Metrics: 500 (0 % unique rows), 500 (100 %) SAFE ROWS (compliant with the k-value of 5), 0 (0 %) RISKY ROWS (non-compliant with the k-value of 5).

A "Next" button is located at the bottom right of the interface.

“One click” policy compliance

What We Did: Built a no-code web interface for non-experts focusing on explainability and simplicity.

Your data meets the requirements for WOG Mode A.

Data Anonymisation Report

Filename: **final_vehicle_ownership_datas**
enCRYPT Job ID: **20230524-y1N2V6qVX98**
Username: **alan_tang@tech.gov.sg**

Original file		
Date and time	24 May 2023, 12:02 PM (job started)	24 M com
File size	44.98 KB	
Number of data fields	9	

UTILITY

Dataset-level

INFORMATION LOSS

3.33 %

The loss of utility due to a reduction in accuracy (with generalisation) or complete loss (with suppression) of indirect identifiers' values.

NORMALISED AVERAGE EQUIVALENCE CLASS SIZE

1.47

The closeness of equivalence class sizes to the best case of size k. A value of 1 means ideal anonymisation (highest utility), with the equivalence class sizes equal to the given k-value.

EQUIVALENCE CLASS SIZE

MAX: **11** MIN: **5**

The groups of records sharing the same values on a set of indirect identifiers.

OVERALL INDIRECT-IDENTIFIERS' TRANSFORMATIONS

LEGEND: Retained Generalised Suppressed

67.47 % **32.53 %**

Column-level

GENDER

Suppressed: **0 %**

Generalised: **0 %**

Retained: **100 %**

Information loss: **0 %**

BLOODTYPE

Suppressed: **0 %**

Generalised: **0 %**

Retained: **100 %**

Information loss: **0 %**

AGE

Suppressed: **0 %**

Generalised: **97.6 %**

Retained: **2.4 %**

Information loss: **9.98 %**

Validation and risk /utility metrics

What We Did: Built a no-code web interface for non-experts focusing on explainability and simplicity.




Glossary of Privacy Terms

- Anonymisation
- De-identification
- Data transformation
- Policy guided access control modes
- Information types
- Sensitivity types

Self-help privacy guides

K-ANONYMITY CEO, 38 → C-Suite, [30-40] CTO, 35 → C-Suite, [30-40]	GENERALISATION 23-12-2021 → [01-12-2021, 30-12-2021] CEO → C-Suite
ENCRYPTION S4686731D → ensdad26526dgfsah	PERTURBATION 23.23 → 23.04
PSEUDONYMISATION S4686731D → had26526sa	SHUFFLING Robert → Emma Emma → Robert
MASKING S4686731D → *****731D emma@tech.com → *****@tech.com	REPLACEMENT Emma → Julie
TRANSPOSITION S4686731D → U608953F	AGGREGATION \$2000 → \$2500 \$3000 → \$2500
RETENTION S4686731D → S4686731D	SUPPRESSION S4686731D → -
	REMOVAL S4686731D →



Name: Julie Watson
Position: C-Suite
NRIC: had26526sa
Email: *****@tech.com
Age: [30-40]
Date of Birth: 1988-Q3

What We Did: Built a no-code web interface for **non-experts** focusing on **explainability** and **simplicity**.



Privacy workshops and clinics

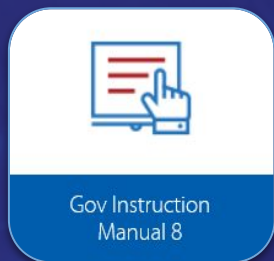


#2.

Strong Policy-Tech Integration Is Key

Data governance regimes often dictate privacy use cases; working with policymakers should be a two-way conversation.

What We Learned: Evolving data privacy regime in Singapore and globally - creates opportunity for a “virtuous cycle” for policy to shape technology solutions, and vice versa.



**PUBLIC SECTOR
(GOVERNANCE) ACT 2018**

Public-Private Data Sharing
Public-Public Data Sharing

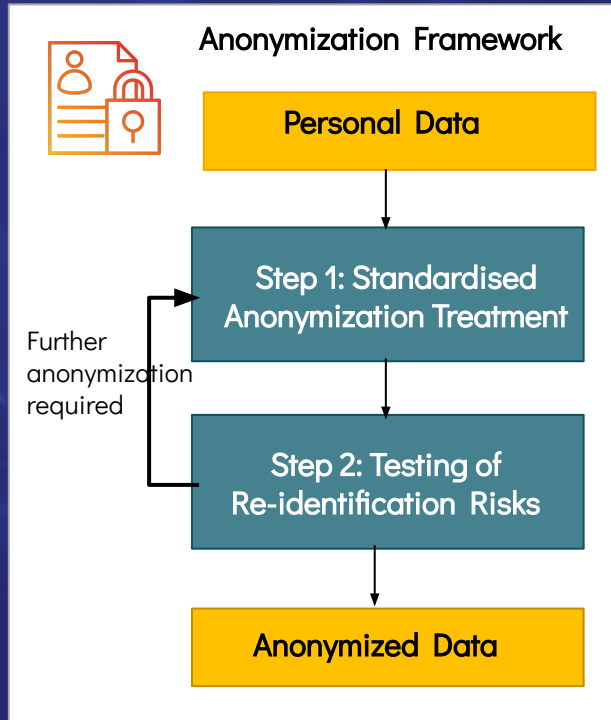


Private-Private Data Sharing
Private-Public Data Sharing



Other Data Privacy, Risk and
Security Recommendations

What We Did: Worked with policymakers to **clarify** and **refine** implementation of **anonymization framework** for **public-private sharing**.



Ideate

- ✓ Co-create guidelines
- ✓ Explore and prototype alternative PETs .

Iterate

- ✓ User feedback for **policy review** (e.g. k-anon erodes utility).
- ✓ Discuss **technical approach** and **risk/utility trade-offs**.

Implement

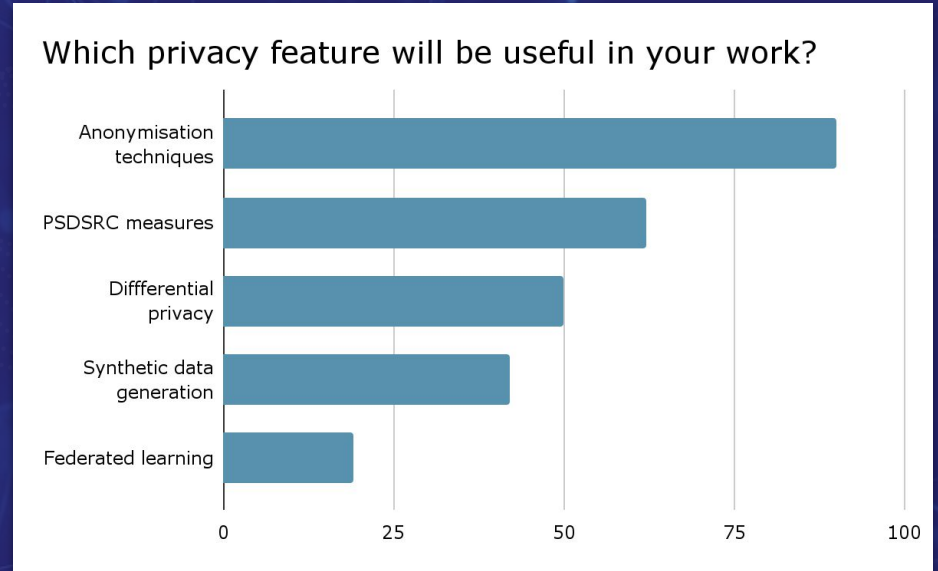
- ✓ **Guided walkthroughs** based on use case and environment.
- ✓ **“One click”** automated compliance.



#3. Optimise For Use Cases, Then Technology

Traditional anonymization can be complemented by cutting-edge PETs which may not yet be ready for mass adoption.

What We Learned: Most users are PET-agnostic; traditional anonymization remains the de facto choice.



What We Did: Adopt a bespoke and measured approach to demonstrate value of PETs to users.

Experiment

Test drive and evaluate industry and open-sourced offerings.



Proof of Concept

Partner agencies on specific use cases to validate risk vs. utility trade-off.



Implement at Scale

Incorporate as central feature in CLOAK Central Privacy Toolkit.

Agency Use Case

Objective: De-risking data prior to sharing with research partners.

Economic Sector Agency

Short term: Explore optimal anonymization approach to be policy compliant, while preserving value of the data.

Medium term: Experiment with alternative PETs (e.g. synthetic data generation) - potentially as part of “policy sandboxes”.



#4.

Privacy Should Be 'Baked In' To The Product Lifecycle.

Privacy-by-design as a rule of thumb for effective adoption and integration.

What We Learned: Privacy tools need to be **tightly integrated** with users' workflows in order to scale.

TRUST

TRUST: Improving health outcomes through trusted data exchange.

AG

Analytics.Gov: Central secured data exploitation platform for government.



The Challenge: Potential **privacy leakages** from usage of **LLM products** in the public sector.

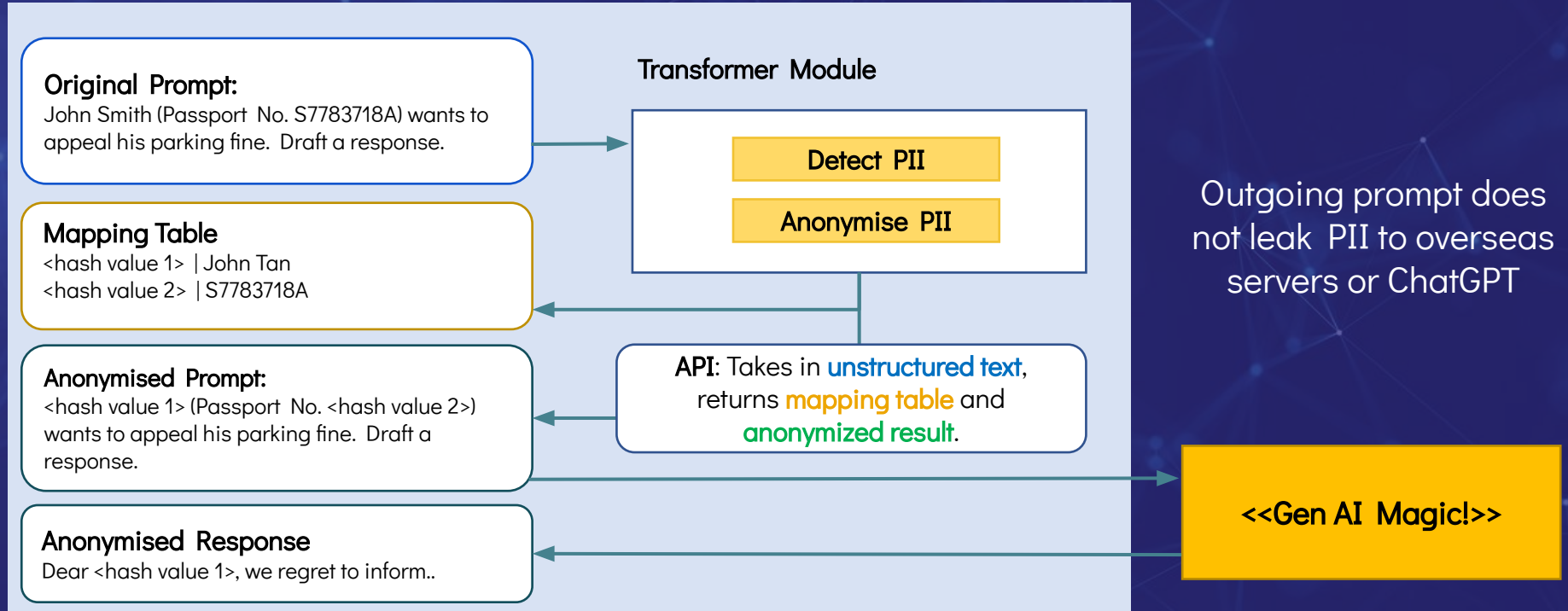


What We Did: Designed API workflow to mitigate potential privacy leakages for Generative AI use cases

Agency Product

CLOAK


GenAI Product



Cloak's Free Text Anon API is currently used to support >20 LLM products and use cases; >1m API calls made to date.

Case Description (required):

I am John Lim and I stay at 283 Bukit Batok East Avenue 3. My maid, Sutiawati, is from Indonesia and she is trying to run away from home now. I need MOM's advice on what to do.

 Remove PII from the case description using enCRYPT

 Suggest matching FAQ

Here is the anonymised case description: I am <PERSON> and I stay at <SG_ADDRESS><SG_ADDRESS_STREET>. My maid, <PERSON>, is from <LOCATION> and she is trying to run away from home now. I need MOM's advice on what to do.



#5.

Bridge the Usability Gap For Widespread Adoption Of PETs

Case Study: Developing Usable
Differential Privacy (DP) Tools,
from Theory to Practice

Guidance for potential adopters

Identifying engineering gaps and future research directions

Benchmarking Tools

(e.g., libraries, frameworks)



Developing Usable Tools



Differential Privacy



Homomorphic encryption



Synthetic data generation

...

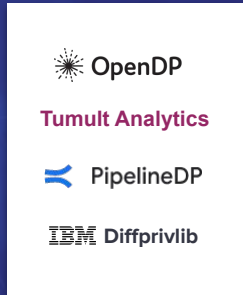
Taking Stakeholders' Perspectives

(e.g., data handlers, product managers, policymakers)



Understand the practical and organisational challenges, opportunities and perceptions

Key Desiderata For Benchmarking Differential Privacy Libraries and Frameworks



ANALYTICS	SECURITY & RELIABILITY	USABILITY	DIFFERENTIAL PRIVACY FEATURES	PERFORMANCE
DIVERSE SUMMARY STATISTICS	CRYPTOGRAPHICALLY SECURE RANDOM NUMBER GENERATION	EASE-OF-USE (FAMILIAR PYTHON/SQL APIS and ECOSYSTEM)	PRIVACY BUDGET ACCOUNTING	SCALABILITY ON ARBITRARY DATASET SIZES (TIME AND MEMORY CONSUMPTION)
GROUP BY AND FILTERS	PREVENTS TIMING ATTACKS	ACCURACY ADJUSTMENTS	DIVERSE MECHANISMS	ACCURACY ON DATASETS OF DIVERSE CHARACTERISTICS
JOIN ON PUBLIC/PRIVATE DATASETS	PREVENTS FLOATING POINT VULNERABILITY	UNCERTAINTY ESTIMATION	STRICT AND RELAXED PRIVACY DEFINITIONS	
ADAPTIVE/INTERACTIVE QUERYING	PREVENTS OTHER KNOWN/UNDERLYING LIBRARIES VULNERABILITIES	RISK ESTIMATION	BASIC AND ADVANCED COMPOSITIONS	
USER-DEFINED FUNCTIONS	ROBUSTNESS TO PRIVACY GUARANTEES	DISTRIBUTED COMPUTING FOR LARGE DATASETS	CONFIGURE PROTECTED CHANGE	
DETERMINISTIC FUNCTIONS (TRANSFORMATIONS)	TRANSPARENCY ON UNADDRESSED VULNERABILITIES	PROTECTED OUTPUT CONSISTENCY	CONFIGURE CONTRIBUTION BOUNDING	
	MATURITY	AUTOMATED BOUNDS COMPUTATIONS	PRIVACY DEFINITION CASTING	
		AUTOMATED PRIVATE PARTITIONS SELECTION FOR GROUP BY	POPULATION AMPLIFICATION	
		HANDLING NULLS/NaNs/INFINITE VALUES		
		PRE/POST PROCESSING FUNCTIONS		

We curated DP Python libraries and frameworks* developed by prominent researchers and institutions that have the potential for wider adoption.

*Based-on the latest version available at the time of development

Built on top of work by Garrido, Gonzalo Munilla, et al. "Lessons learned: Surveying the practicality of differential privacy in the industry." (2022).

Significant changes are expected as they continue to evolve with new features, improvements, and research in differential privacy

- **Spark interest** in practitioners, investors, library creators and research community
- Differences among the tools can be bridged with **engineering efforts**
- **Implementation** can impact the **computational and utility performance** of the outputs
- Abstractions can be catalyst for adoption but **building blocks (core libraries)** can be needed for further optimizations



<https://medium.com/dsaid-govtech>

DSAlD Data Privacy Protection Capability Centre (DPPCC)

from



Sharing Data with Differential Privacy
and A Data Practitioner's Guide to
Benchmarking Differential Privacy
Python Tools

Developing A Usable General-Purpose Web Interface To Generate Private Statistics

Risk, accuracy, uncertainty estimation and visualisation

Privacy budget splitting

Privacy parameter recommendations

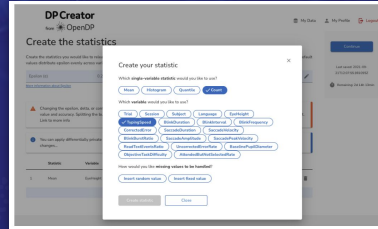
Can provide post-processing functionality to get consistent output?

Can automate or provide recommendations on hyperparameters computations?

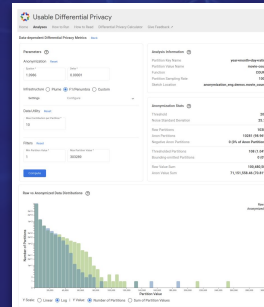
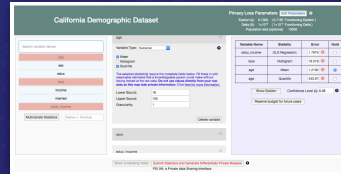
Can minimise user inputs?



DP Creator From Harvard's OpenDP Privacy Team (2022)



PSI: a Private Data Sharing Interface (2016)



Usable Differential Privacy (2020)

ViP: Visualizing Privacy-Utility Trade-Offs in Differentially Private Data Releases (2022)

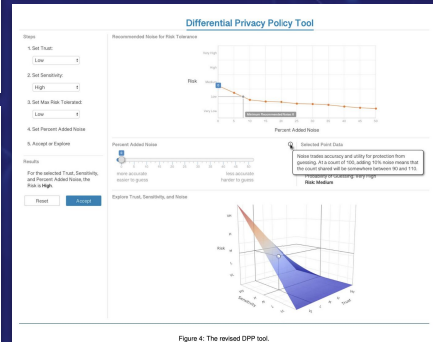
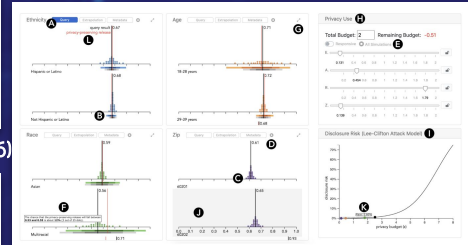


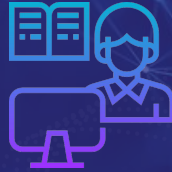
Figure 4: The revised DPP tool.

DPP: Decision Support for Sharing Data Using Differential Privacy (2021)

User Study To Gather Perceptions From Various Government Agencies

#PARTICIPANTS: 18

- Stakeholders from various government agencies and teams
- Unfamiliar with differential privacy



STUDY DURATION: 1h

- Semi-structured interview
- Educating, followed by conducting think-aloud experiments on the step-by-step preliminary design



Our preliminary web interface design used for the study

The screenshot displays a web interface with a four-step process flow:

- Step 1:** Data depositor mode (indicated by a green checkmark)
- Step 2:** Validate Your Use Case and Upload Dataset (indicated by a green checkmark)
- Step 3:** Configure Dataset Protection Details (indicated by a green checkmark)
- Step 4:** Generate Private Statistics and Manage Privacy Budget (indicated by a blue button)
- Step 5:** Download Private Statistics and Report (indicated by a blue button)

Step 3 - Configure privacy budgets and generate private statistics

Based on the sensitivity of your dataset, we have allocated a privacy budget that can be distributed among the various statistics you wish to generate using differential privacy. We have also provided a guide to assist you in configuring the values of the privacy parameters and understanding their implications.

Quick links and guide to settings the values

Clamping bounds
The values are clamped to be within lower and upper bounds to limit the influence of any individual.

What is good way to set the values?
It is often a good idea to choose clamping bounds that aren't absolute limits over the data range, but are such that most values would fall within these bounds. [Learn more.](#)

How does it impact the output?
If these bounds are too tight, the release may be biased, because values outside these bounds are replaced with the nearest bound. On the other hand, if these bounds are too wide, the respective release will have greater variance.

Key Findings/Recommendations From The User Study on Government Agencies

Need for :

- **Awareness and engagements** to try out the technology
- **Communicating clear expectations** around the use cases (e.g., requires shift in data science practices)
- **Designing for different user modes:** data depositor, data analyst, and negotiation to address data sharing and requesting requirements
- **Designing to build confidence in decision-making** (e.g., optimisation and guidance for error minimisation, uncertainty estimation and providing metadata)
- **Guidelines for privacy budgeting** based on risk, data sensitivity and trust

Data Depositor



Data Analysts



Negotiation.



Next Steps

- The Singapore government is **advancing towards adopting differential** privacy by
 - **expanding user education** to raise awareness
 - **surfacing use cases** from various agencies
 - **creating policy on suitable privacy parameters**
 - **developing tools** to get data holder/analysts started with differential privacy
- **Exploring other advanced PETs** like Synthetic Data Generation and Homomorphic Encryption to overcome our data sharing challenges.



Thank You

Meet our data privacy team!



Alan Tang
Lead Product Manager
alantang@dsaid.gov.sg



Anshu Singh
Research Engineer
anshu@dsaid.gov.sg



Syhari Ikram
Data Engineer
syhari@dsaid.gov.sg



Zul Yang
Software Engineer
zul@dsaid.gov.sg



Ivan Wang
Data Engineer
ivan@dsaid.gov.sg



Wang Ting
Data Engineer
wangting@dsaid.gov.sg



Ameera Adam
Data Engineer
ameera@dsaid.gov.sg