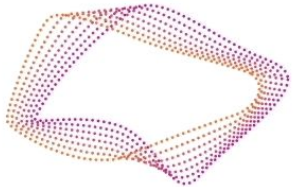


GOVTECH
SINGAPORE

DSAID
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

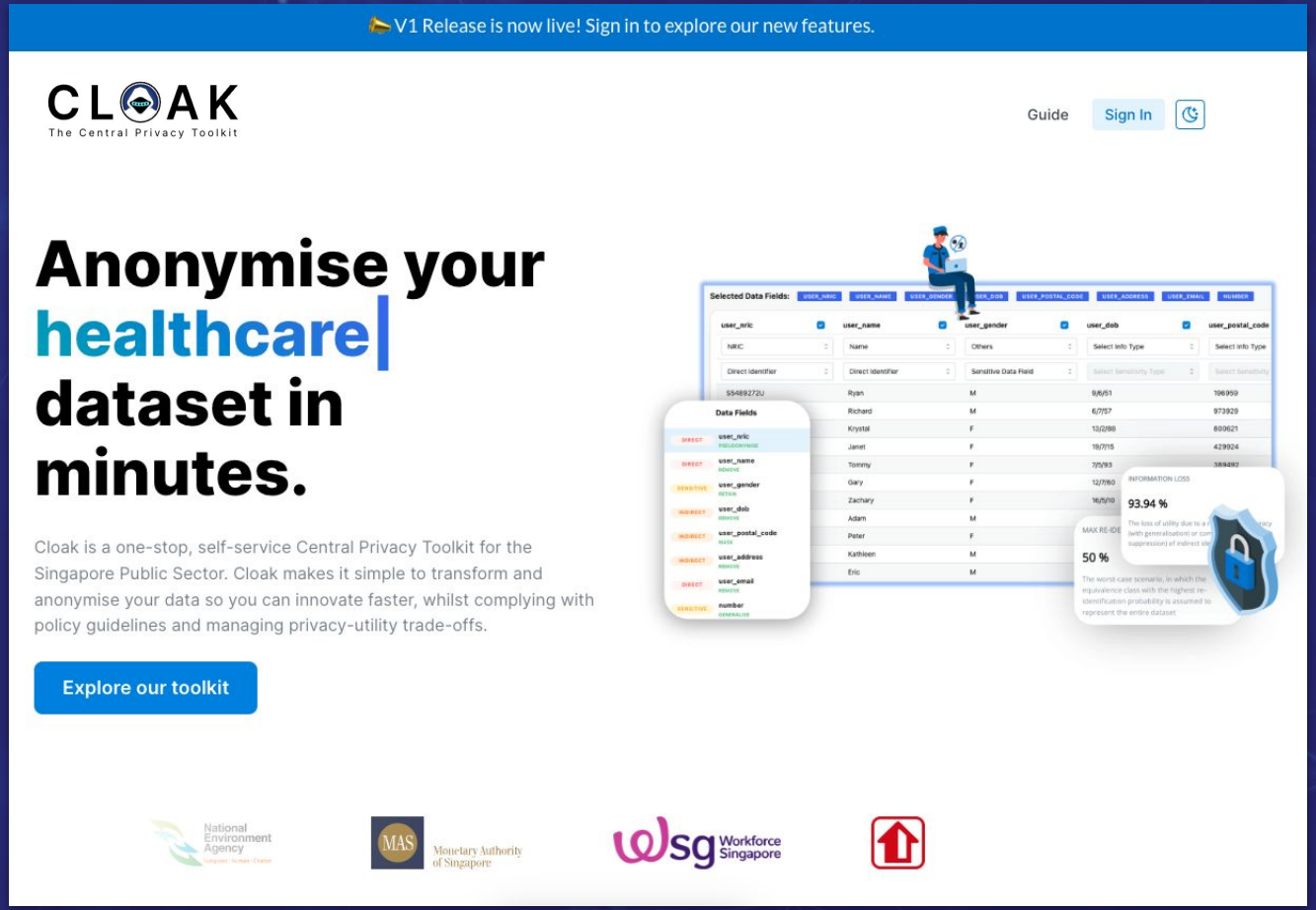
Government Technology Agency, Singapore

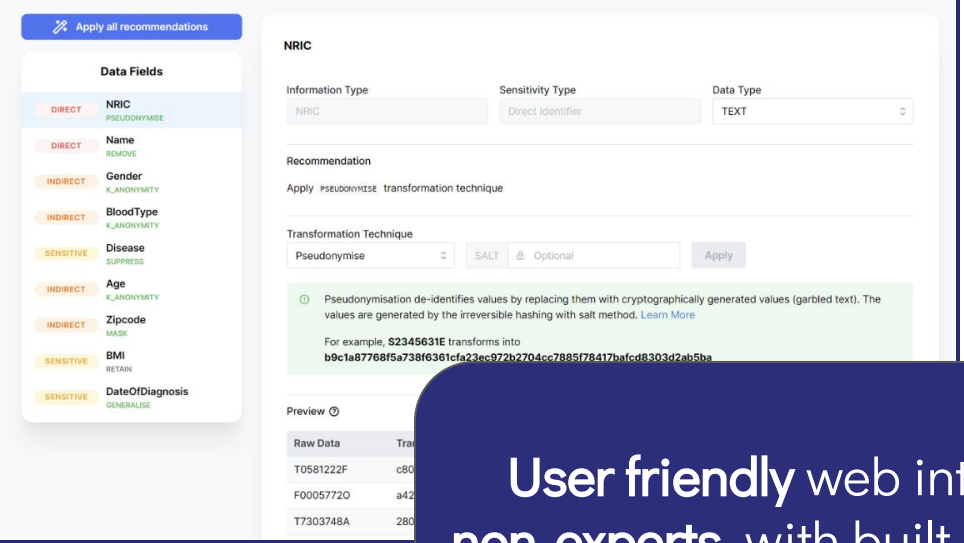
Sep 11, 2023

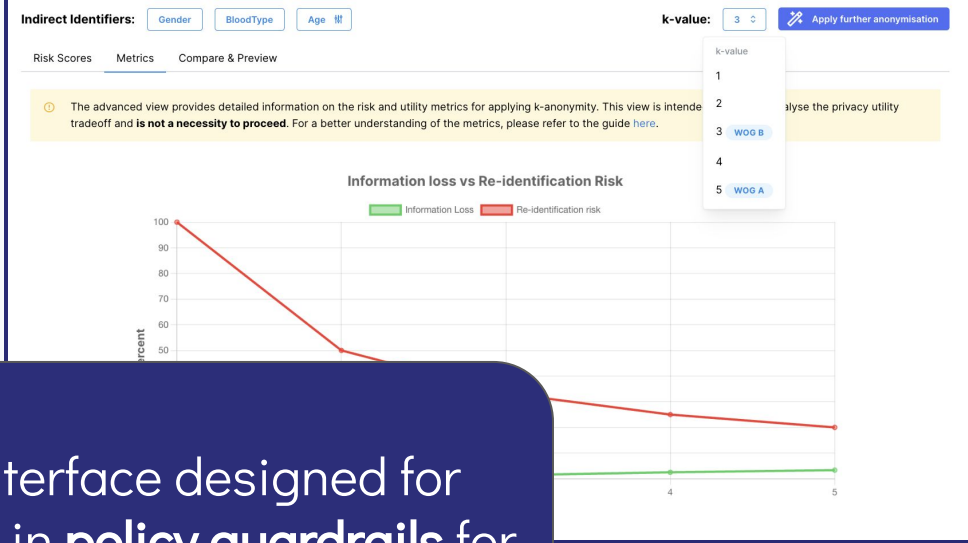
GovTech spearheads digital transformation within Singapore's public sector.

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

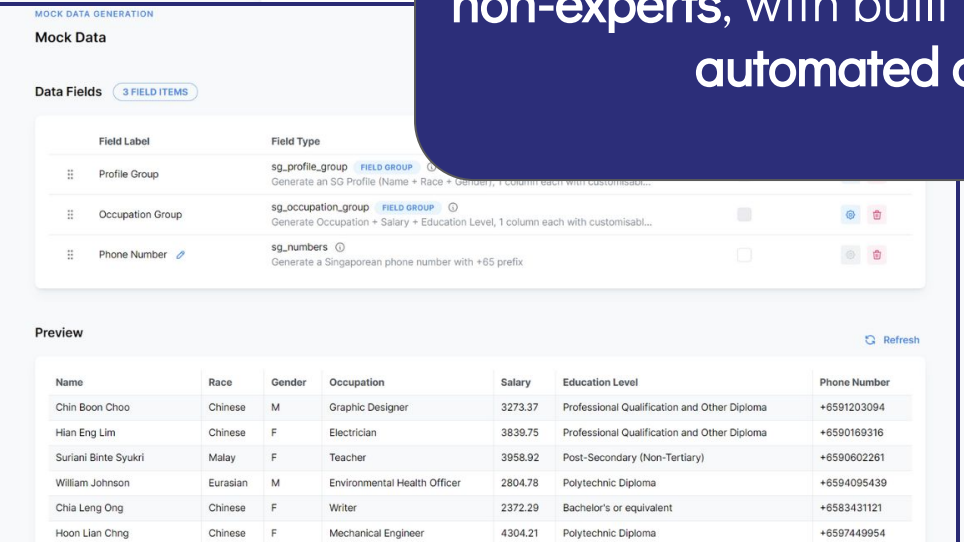
CLOAK serves as the **privacy backbone** for the Singapore public sector – supporting both **individual** and **enterprise** use cases.

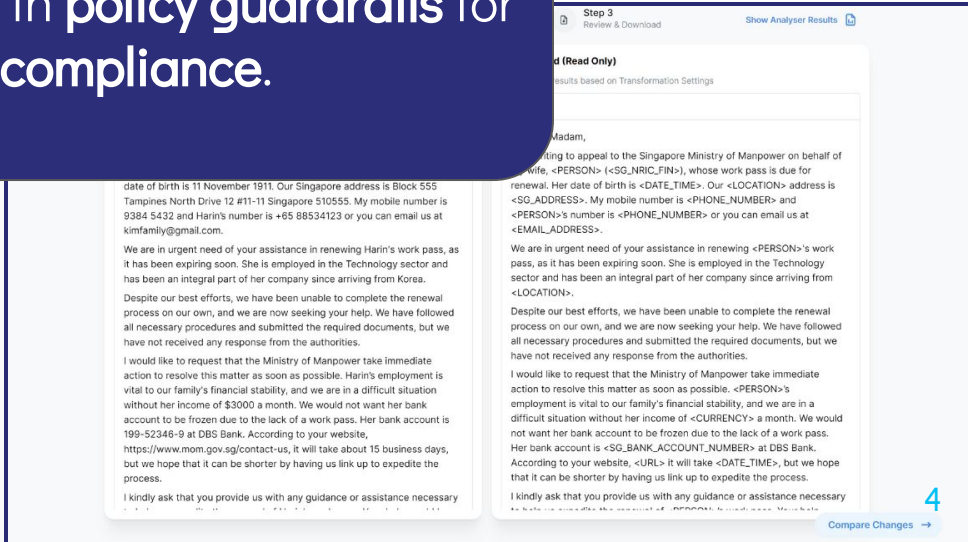






User friendly web interface designed for **non-experts**, with built in **policy guardrails** for **automated compliance**.





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

>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

GOVTECH SINGAPORE

#1.

Privacy is Hard,

But The Risks Are Real

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

GOVTECH SINGAPORE

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

What does "hashing" or "tokenization" mean?

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

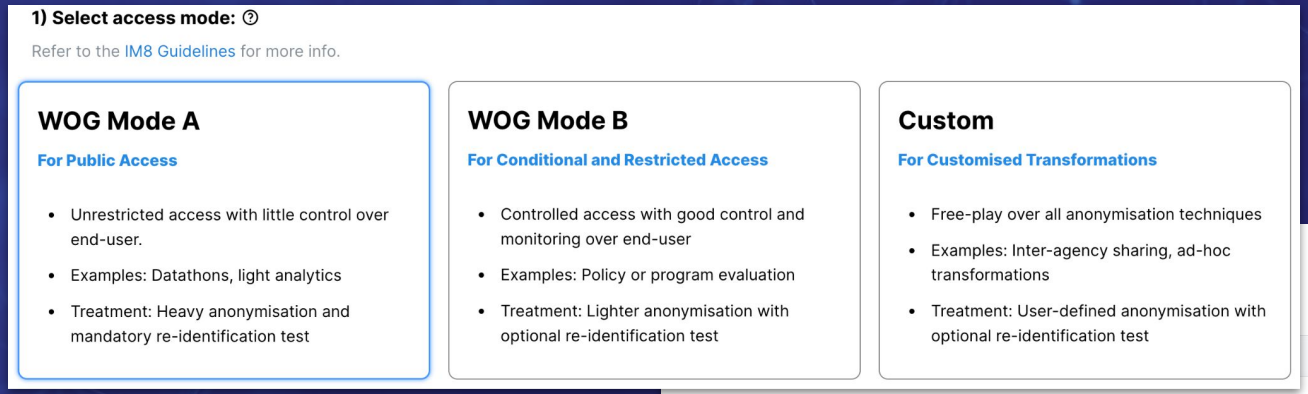
Just wants to get the **job done!**

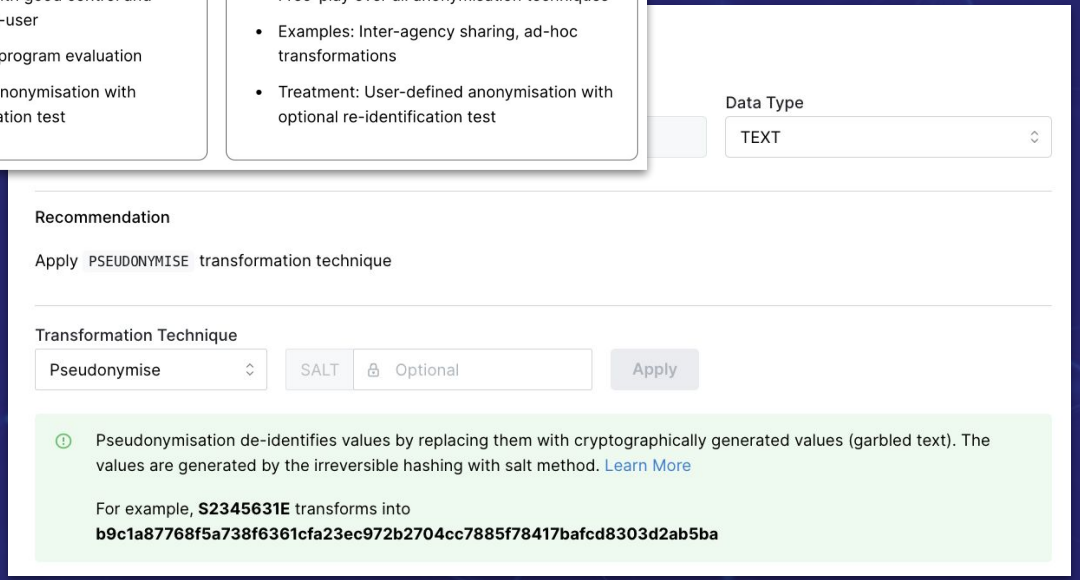
So... can I share my data?

I can't run Python on my work PC!

Lack access to **commercial statistical tools**

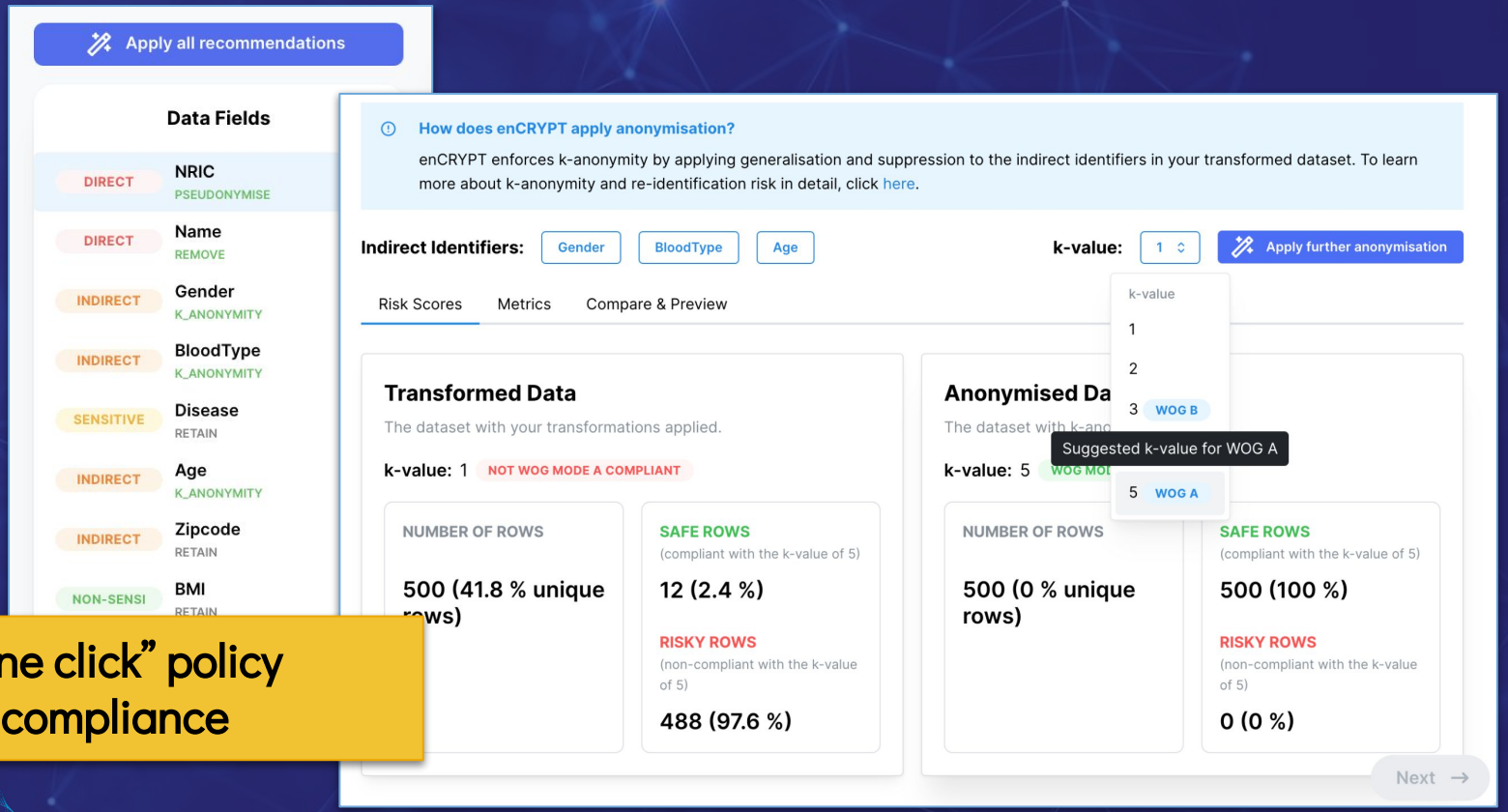
What We Did: Built a **no-code** web interface for **non-experts** focusing on **explainability** and **simplicity**.





Policy guardrails and recommendations

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



“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 %

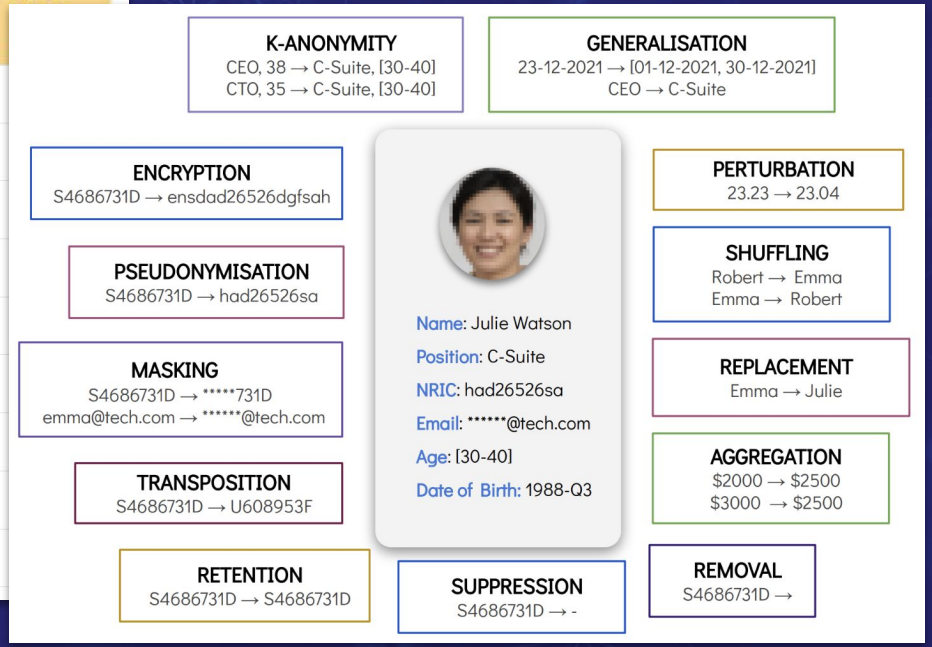
AGE

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**.

Self-help privacy guides



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

Privacy workshops and clinics

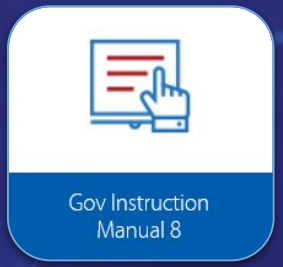
#2.

Strong Policy-Tech Integration Is Key

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

GOVTECH SINGAPORE

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**.

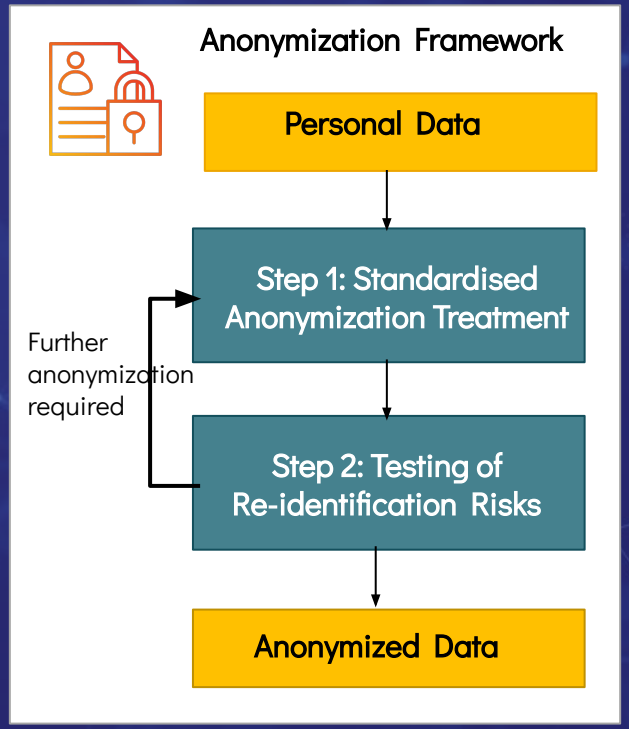


Public-Private Data Sharing
Public-Public Data Sharing

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 .

Implement

- ✔ **Guided walkthroughs** based on use case and environment.
- ✔ **"One click"** automated compliance.

Iterate

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

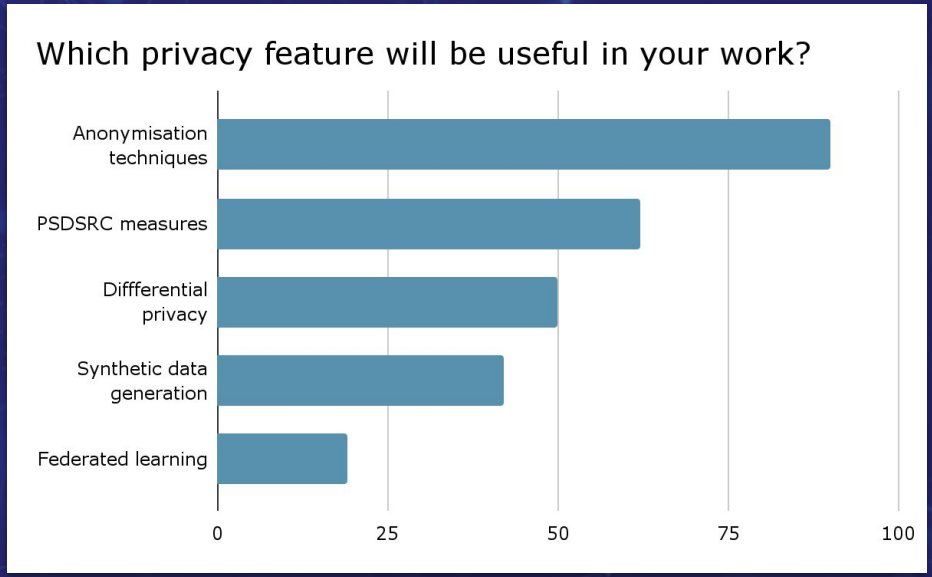
#3.

Optimise For Use Cases, Then Technology

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

GOVTECH SINGAPORE

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.

GOVTECH SINGAPORE

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

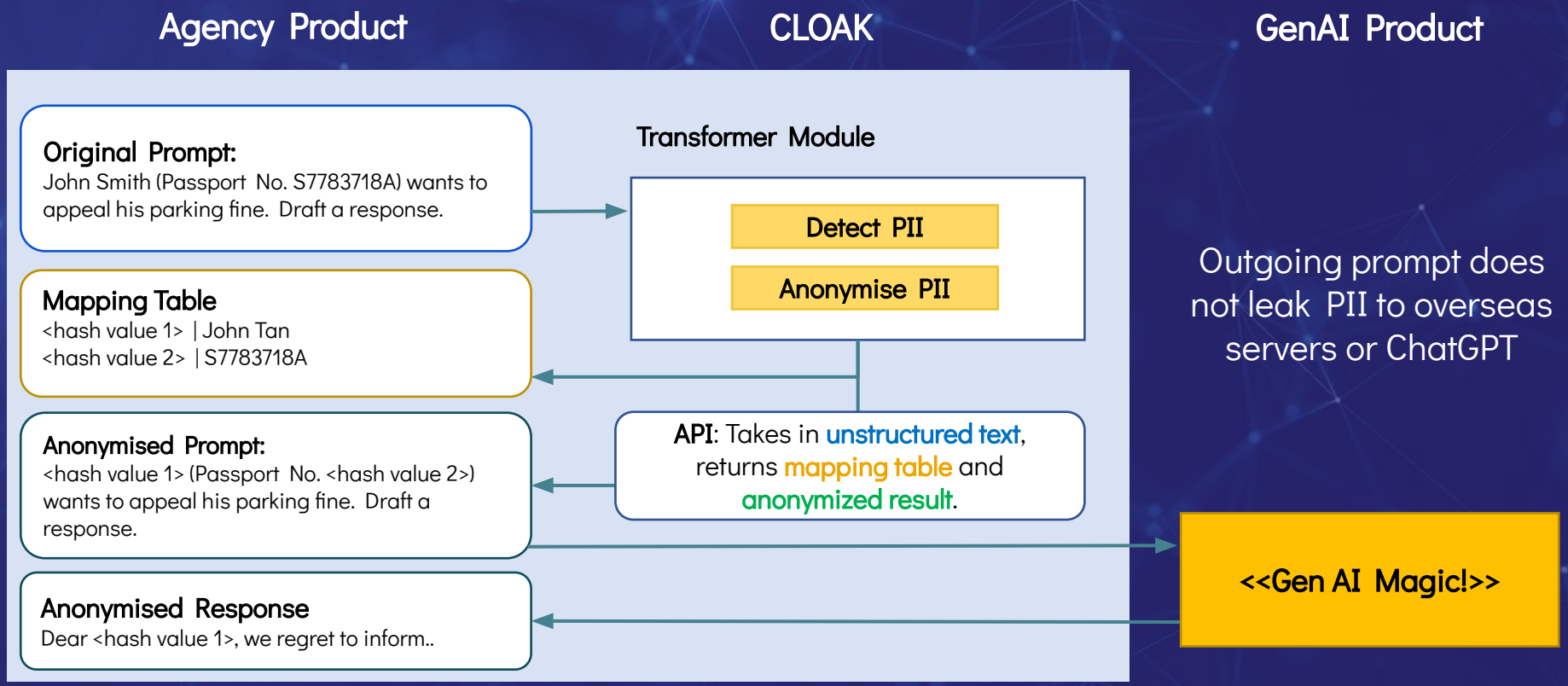
TRUST: Improving health outcomes through trusted data exchange.

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

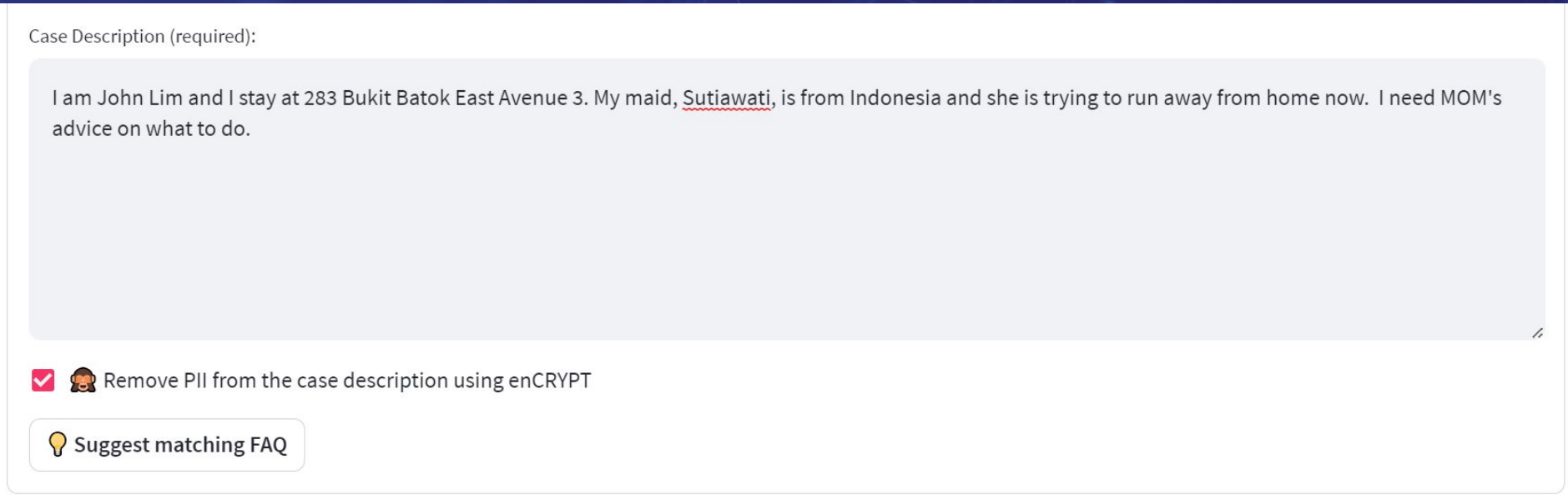
The Challenge: Potential **privacy leakages** from usage of **LLM products** in the public sector.

Original Prompt:
John Smith (Passport No. S7783718A) wants to appeal his parking fine. Draft a response.

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



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



#5.

Bridge the Usability Gap For Widespread Adoption Of PETs

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

GOVTECH SINGAPORE

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

OpenDP

Tumult Analytics

PipelineDP

IBM Diffprivlib

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

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		

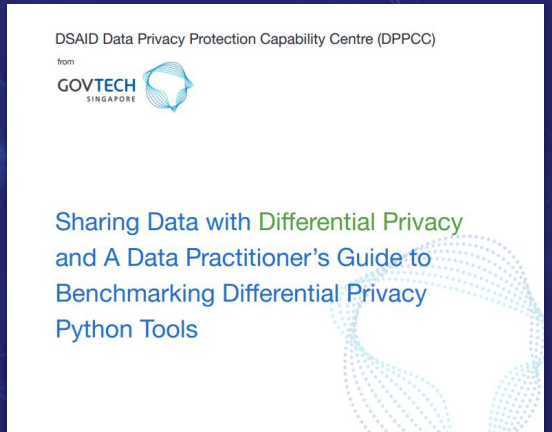
*Based-on the latest version available at the time of development

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



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

Risk, accuracy, uncertainty estimation and visualisation

Privacy budget splitting

Privacy parameter recommendations

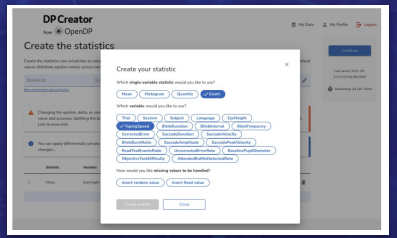
Can provide post-processing functionality to get consistent output?

Can automate or provide recommendations on hyperparameters computations?

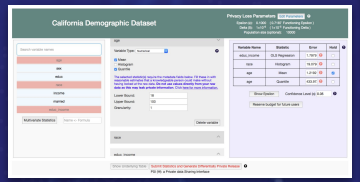
Can minimise user inputs?

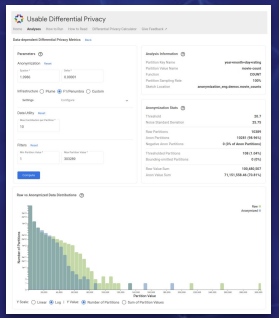
●●●

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



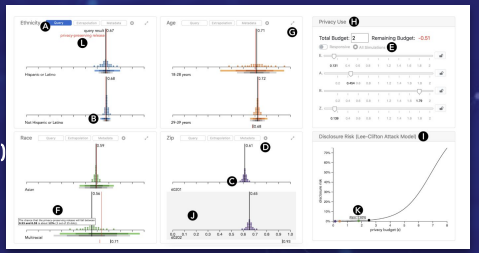
PSI*: a Private Data Sharing Interface (2016)

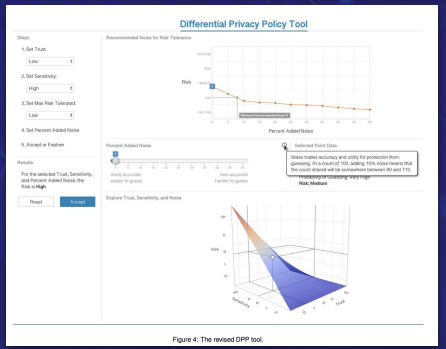




Usable Differential Privacy (2020)

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





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

*DP Creator evolved from PSI

User Study To Gather Perceptions From Various Government Agencies

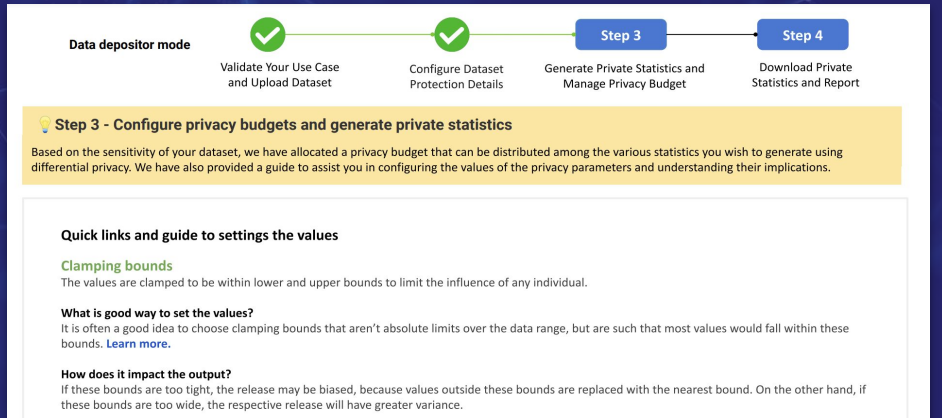
#PARTICIPANTS: 18

- 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



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

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
syahri@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