# Data Privacy Workshop

# Why is Data Privacy Important to Us?

"Data is the **lifeblood of the digital economy and a digital government**. We need to **use and share data as fully as possible** to provide better public services.

…

As the **custodian of a vast amount of data**, the Government takes this responsibility very seriously. We must do out utmost to **minimize the risk of data breaches**.

…"

- Prime Minister Lee Hsien Loong, PSDSRC report November 2019

# Workshop Schedule

| Timeslot | Activity |
|---|---|
| 09:00AM – 09:30AM | Introduction and Terminologies<br>Data Privacy Regulations |
| 09:30AM – 10:00AM | Discussion + Case Study A |
| 10:00AM – 10:30AM | PSDSRC Overview<br>PSDSRC Recommendations 1.1 |
| 10:30AM – 10:50AM | Case Study B<br>**5 min break** |
| 10:50AM – 12:00PM | PSDSRC Recommendations 1.2 + Python hands-on |
| 12:00PM – 01:30PM | **Lunch break** |
| 01:30PM – 02:30PM | PSDSRC Recommendations 1.3 + Python hands-on<br>**5 min break** |
| 02:30PM – 03:30PM | Data Anonymization Concepts and methods |
| 03:30PM – 04:00PM | Conclusion + Q&A |

# Introduction

This presentation was designed to give a quick primer on data privacy concepts and techniques.

In this presentation, we will be covering theory on Data Privacy concepts and techniques as well as practical python examples of how these techniques are done.

Participants are assumed to have some knowledge on basic python programming and have some experience working with data.

# What is Data Privacy?

# What is Data Privacy?

**Data**

information, especially facts or numbers, collected to be examined and considered and used to help decision-making, or information in an electronic form that can be stored and used by a computer

- Cambridge Advanced Learner's Dictionary

**Privacy**

the right that someone has to keep their personal life or personal information secret or known only to a small group of people

- Cambridge Business English Dictionary

# What is Data Privacy?

empowering your users to make their own decisions about who can process their data and for what purpose

- GDPR, EU

freedom from intrusion into the private life or affairs of an individual when that intrusion results from undue or illegal gathering and use of data about that individual

- NISTIR 8053 ISO/IEC 2382

# What is Data Privacy? (In our own words)

- Allow users to make decisions on who can process their **personal data** and for what purpose
- Ensure that usage of users' personal data is limited to the **parties** and **purposes** that they have consented to

# PII vs Personal Data

# Personally identifiable information (PII) vs Personal Data

## PII

**Personally identifiable information (PII)** is any data that could **potentially identify** a specific individual.

**Any information** that can be used to distinguish one person from another and can be used for deanonymizing previously anonymous data can be considered PII.

- Tech Target

Any information about an individual maintained by an agency, including (1) any information that can be **used to distinguish or trace an individual's identity**, such as name, social security number, date and place of birth, mother's maiden name, or biometric records; and (2) any **other information that is linked or linkable to an individual**, such as medical, educational, financial, and employment information.

- NISTIR 8053 under personally identifiable information GAO Report 08-536, NIST SP 800-122

## Personal Data

**'personal data'** means **any information relating to an identified or identifiable natural person** ('data subject'); an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person

- GDPR

any information relating to an identified or identifiable natural person (data subject)

- NISTIR 8053 ISO/TS 25237:2008

# Personally identifiable information (PII) vs Personal Data (Our definition)

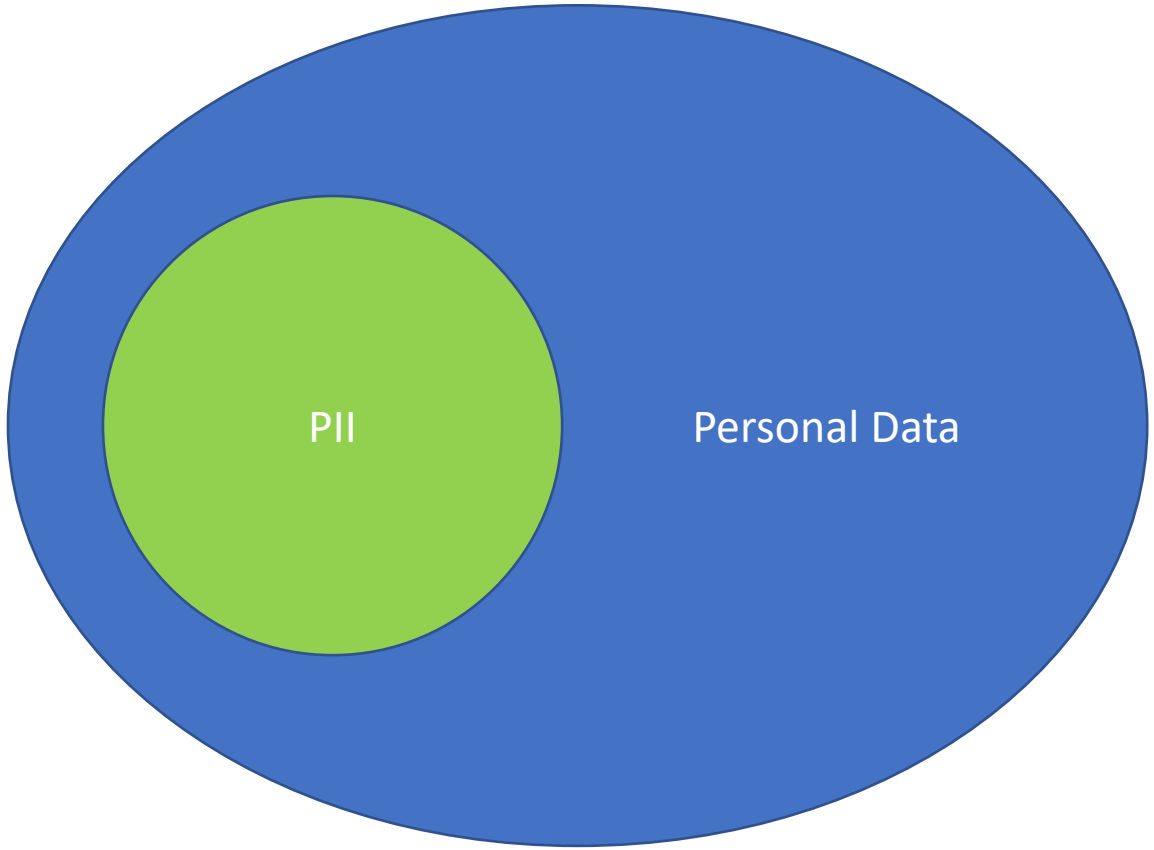
**PII**

- **Directly identifying attributes** of user *(e.g. NRIC, Name, email, mobile number, etc.)*
- **Indirectly identifying attributes** that is directly linked to the user *(e.g. gender, address, religion, birthday)*

**Personal Data**

- **any information relating** to an identified or identifiable natural person *(e.g. any PII, transaction data, web browsing data, geolocation data, etc.)*

# Personally identifiable information (PII) vs Personal Data




Examples of PII:

- NRIC
- Name
- Age / Birthday
- Address
- Religion

Examples of Personal Data that are not PII:

- Geolocation data
- Transaction data
- Web browsing data

# Data Privacy Regulations

# Data Privacy Regulations

- Personal Data Protection Act (PDPA), Singapore
- IM8 – Data Management and Additional Requirements for the Protection of Personal Data, Singapore Government
- *General Data Protection Regulation (GDPR), European Union

* Note: While GDPR has extra-territorial effect (i.e. applies to organizations that handle personal data of EU citizens whether they are EU-based organizations or not), it is rarely applicable to data usage in Singapore public service.

# Data Privacy Regulations – PDPA Obligations

| Obligation | Explanation |
|---|---|
| Consent | • Only collect, use or disclose personal data for **purposes** for which an individual has given his or her **consent**<br>• Allow individuals to **withdraw consent** |
| Purpose Limitation | • May collect, use or disclose personal data about an individual for **reasonable purposes** which the individual has **given consent**<br>• But **may not** require the individual to consent for collection, use or disclosure of his or her personal data **beyond what is reasonable**, as a **condition** of providing a product or service |
| Notification | • **Notify individuals of the purposes** for collection, use or disclosure of their personal data **on or before** such collection, use or disclosure of personal data. |
| Access and Correction Obligation | • Upon request, the **personal data of an individual** and information about the ways in which personal data **has been or may have been used** or disclosed within a year before the request should be provided.<br>• Requirement to **correct any error or omission** in an individual's personal data upon his or her request |

# Data Privacy Regulations – PDPA Obligations

| Obligation | Explanation |
|---|---|
| Accuracy | • Make reasonable effort to ensure that personal data collected is **accurate and complete** |
| Protection | • Make **reasonable security arrangements** to protect the personal data that your organisation possesses or controls |
| Retention Limitation | • **Cease retention** of personal data, or remove the means by which the personal data can be associated with particular individuals, when it is **no longer necessary** for any business or legal purpose. |
| Transfer Limitation | • Transfer personal data to another country **only according to the requirements** prescribed under the regulations, to ensure that the standard of protection provided to the personal data so transferred will be **comparable** to the protection under the PDPA |
| Accountability | • Make information about your **data protection policies, practices and complaints process** available on request.<br>• Designate one or more individuals as a **Data Protection Officer** to ensure that your organisation complies with the PDPA |

# Data Privacy Regulations – IM8 equivalent

| Obligation | IM8 |
|---|---|
| Consent | Additional Requirements for the Protection of Personal Data - Section 1.5 to section 1.14 |
| Purpose Limitation | Additional Requirements for the Protection of Personal Data - Section 1.15 |
| Notification | Additional Requirements for the Protection of Personal Data - Section 1.16 to section 1.19 |
| Access and Correction Obligation | Additional Requirements for the Protection of Personal Data - Section 1.20 to section 1.28 |
| Accuracy | Additional Requirements for the Protection of Personal Data - Section 1.29 |
| Protection | Policy on Data Management – Section 3 |
| Retention Limitation | Additional Requirements for the Protection of Personal Data - Section 1.30 to Section 1.32 |
| Transfer Limitation | Policy on Data Management – Section 4 |
| Accountability | Additional Requirements for the Protection of Personal Data – Section 1.3 |

# Links and resources

- PDPC website: https://www.pdpc.gov.sg/Overview-of-PDPA/The-Legislation/Personal-Data-Protection-Act
- AGC statutes: https://sso.agc.gov.sg/Act/PDPA2012
- IM8 (GSIB Intranet): https://intranet.mof.gov.sg/portal/IM/Themes/IT-Management/Data.aspx

# Discuss: What are the PDPA Obligations before, on and after collection of personal data?

**On or before collection of personal data**

- Consent
- Purpose limitation
- Notification

**After collection of personal data**

- Access and correction obligation
- Accuracy
- Protection
- Retention limitation
- Transfer limitation

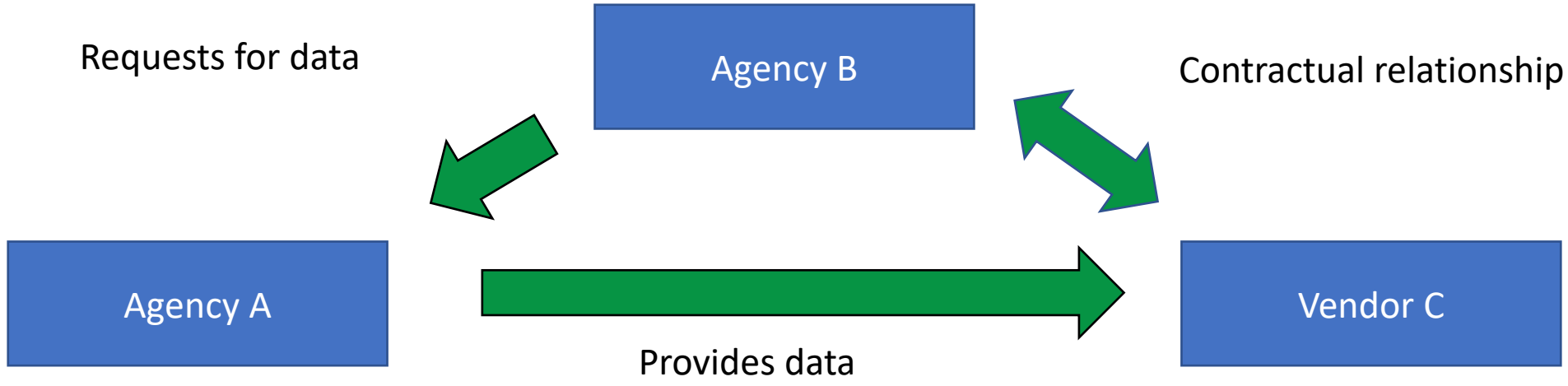
**All time**

- Accountability

# Case Study A (15mins)

Agency A owns a set of corporate contact data (office contact and email). Agency B has a project that requires the corporate contact data from Agency A to be shared to Vendor C for further processing.

By reading up the relevant clauses of PDPA and IM8, explain what actions should Agency B take before the sharing of data could be done.



# Case Study A (potential solution)

- PDPA **does not apply** in this scenario

*PDPA Section 4(1)(c): Parts III to VI (of the PDPA) shall not impose any obligation on any public agency or an organisation in the course of acting on behalf of a public agency in relation to the collection, use or disclosure of the personal data*

- Additional requirements for the protection of personal data **does not apply** to business contact information

*IM8 Additional requirements for the protection of personal data:*

*The "Additional requirements for the protection of personal data" shall apply to all personal data possessed, controlled, or processed by an Agency except: 1. Personal data of a deceased individual; and 2. Business contact information*

# Case Study A (potential solution - continued)

- Can use **Standardised Terms of Sharing (STS)** and **Standardised Data Sharing Form (SDSF)** for sharing data from Agency A to Agency B with explicit authorization for use by Vendor C for the specified purpose.

*IM8 Data, 4. Data Access and Distribution, Section 1.4:*

*Agencies shall adopt the STS when sharing data with another Agency. The STS defines the respective roles and responsibilities of the data provider, data requestor and the data controller. It clarifies that accountability for the data flows with the movement of the data.*

*IM8 Data, 4. Data Access and Distribution, Section 1.14/G1:*

*Agencies should consider adapting the STS and/or SDSF when sharing data with Non-Government Entities (NGEs) because the governance principles therein are useful starting points for Agencies to establish the Terms of Use.*

- Agency B also has to ensure that there are **controls in place** to govern the handling and processing of data in the form of **contractual clauses** with Vendor C

*IM8 Data, 4. Data Access and Distribution, Section 1.14/G2:*

*NGEs are not subject to the Instruction Manuals (IM). Apart from the Official Secrets Act, the Statutory Bodies and Government Companies (Protection of Secrecy) Act and other sector-specific legislation which may apply, the primary form of control over NGEs would be in the form of contractual clauses governing the handling and processing of such data.*

# PSDSRC Report and its Recommendations

# Public Sector Data Security Review Committee (PSDSRC) Report

- Comprehensive Report on Data Security Practices in across entire Public Service
- Five Key Recommendations

| Desired Outcomes | Key Recommendations |
|---|---|
| **Protects data** and **prevents** data compromises | 1. Enhance technology and processes to effectively protect data against security threats and prevent data compromises. |
| **Detects and responds** to data incidents | 2. Strengthen processes to detect and respond to data incidents swiftly and effectively. |
| **Competent** public officers embodying a **culture of excellence** | 3. Improve culture of excellence around sharing and using data securely, and raise public officers' competencies in safeguarding data |
| **Accountability** for data protection at every level | 4. Enhance frameworks and processes to improve the accountability and transparency of the public sector data security regime |
| **Sustainable and resilient manner** | 5. Introduce and strengthen organisational and governance structures to drive a resilient public sector data security regime that can meet future needs. |

Focus on KR1 as it is directly related to DSAID DE team's job scope in building data solutions

Do note that there are some safeguards that are more focused on general data protection and cybersecurity, which will not be our main focus for data privacy

# Public Sector Data Security Review Committee (PSDSRC) Report

- For Key Recommendation 1, to protect data and prevent data compromises:
  - 13 technical safeguards (prefixed with ‘T’)
  - 10 process safeguards (prefixed with ‘P’)
- PSDSRC goal is to implement relevant measures for 80% of Government systems by end 2021 and for all Government systems by end 2023

**Recommendation 1.1:** Reduce the surface area of attack by minimising data collection, data retention, data access and data downloads.

**Collect and retain data only when necessary**

**P1:** Collect datasets only where necessary

**P2:** Limit retention period of data

**Minimise the proliferation of data to endpoint devices**

**P3**: Isolated Secured Environments for third parties and privileged users

**P4**: Access data by queries instead of data dumps

**P5**: Access sensitive files on secured platforms

**Access and use data for the task at hand**

**T1**: Volume limited and time limited data access

**T2**: Automatic Identity and Access Management Tools

**P6**: Limit and monitor authorised and privileged access

# P1 – Collect datasets only where necessary

What:

Reduces the surface area of attack by **minimising the collection of datasets** that are **unnecessary** or have **no clear identified value** for agencies' operations

Why:

Reduce surface area of attack. Lesser data collected is lesser data to be protected and managed.

How:

- Agencies should not collect data that is already collected by **Single Sources of Truth** as part of the Government Data Architecture. Agencies should obtain such data from the **Trusted Centres** when needed.

# P2 – Limit retention period of data

What:

Reduces the surface area of attack by **minimising the storing of datasets** at agency or at endpoint devices

Why:

Reducing surface area of attack. Lesser data retained at endpoint devices or other agencies will reduce the amount of data to be protected and managed.

How:

- Agencies should **set a retention period** for each dataset collected. The datasets should be **purged** from the officer's laptop after the retention period is over

# P3 – Isolated secure environments for third parties and privileged users

What:

Ensures that **high-risk users** (i.e. users which are entrusted with functions, tasks or data that are highly sensitive) are not able to extract data from Government systems.

Why:

Prevent data leakage by extraction or copying.

How:

- Instead of downloading the dataset onto the vendor's system, the vendor or public officer accesses it through a **Virtual Desktop Infrastructure (VDI)** which prevents data transfer. The vendor or public officer is able to view the data required, but will not be able to copy out the data.

# P4 – Access data by queries instead of data dumps

What:

Reduces the risk that a full database is compromised as **only the necessary fields and records are accessed**.

Why:

Reduce surface area of attack by limiting data retrieval based on need to basis.

How:

- Instead of downloading the whole database with records for all citizens, the officer queries the database and **retrieves only the record of the citizen** that the officer is serving.

# P5 – Access sensitive files on secured platforms

What:

Ensures that access to sensitive files have the **appropriate security safeguards** and are logged and monitored.

Why:

Prevent data leakage by endpoint devices being accessed by unauthorized users and ensure that logs are monitored for malicious activities.

How:

- Instead of downloading data from the database onto his laptop, he should use the collaboration functions of the **Singapore Government Document Collaboration Service (SG-DCS)** to access data on the platform where feasible. This is the equivalent of accessing data on Google Drive without downloading it onto one's laptop.

# T1 – Volume limited and time limited data access

What:

Prevents officers from accessing too much data at one time, and the duration the officer can access it.

Why:

Access should be limited to relevant datasets, or to a subset of the data so that the data users only have access to the data that is needed to carry out their assignment. This reduces the risk of an accidental exfiltration, or outside attacker gaining access to sensitive data through compromising a data user's account. This also reduces the risk of an insider accessing sensitive data when he or she is not authorised to do so.

How:

- **Restrict data access** when the **duration and volume** data access **exceeds predefined limits**. (e.g. VPN time limit)

# T2 – Automatic Identity and Access Management (IAM) tools

What:

Ensures that access to the data is limited only to people authorised to do so.

Why:

Access should be limited to relevant datasets, or to a subset of the data so that the data users only have access to the data that is needed to carry out their assignment. This reduces the risk of an accidental exfiltration, or outside attacker gaining access to sensitive data through compromising a data user's account. This also reduces the risk of an insider accessing sensitive data when he or she is not authorised to do so.

How:

- **IAM tools** automatically manage officers' identity and access rights, ensuring that only authorised persons can access data.
- **Automatic Privileged Identity and Management (PIM)** tools control, monitor, and protect user accounts which have **more access and capabilities than ordinary users** (e.g. administrator accounts) and **enforce more stringent measures** (e.g. 2FA, time-limited access) to protect these accounts.

# P6 – Limit and monitor authorized and privileged access

What:

Reduces the risk that a malicious outsider gains access to an account with access to sensitive data.

Why:

Access should be limited to relevant datasets, or to a subset of the data so that the data users only have access to the data that is needed to carry out their assignment. This reduces the risk of an accidental exfiltration, or outside attacker gaining access to sensitive data through compromising a data user's account. This also reduces the risk of an insider accessing sensitive data when he or she is not authorised to do so.

How:

- Agencies should set strict processes that allow privileged access only where necessary, and ensure access is closely tracked and not shared.
- This is enabled by the technical safeguard, **T2: Automatic Identity and Access Management Tools**.

# Case study B (15mins)

- Agency A needs to conduct a study of social economic trends and has engaged a vendor to help with the analysis of the data. However, Agency A does not currently have the required data to perform the study.
- Make your own assumptions on the attributes of the data (e.g. size, nature, sensitivity, etc.)
- Using the applicable safeguards in PSDSRC recommendation 1.1, list and explain measures that Agency A should take to perform this study.

# Case study B (Potential Solution)

1. (P1) If the data required is already collected by **Single Sources of Truth** as part of the Government Data Architecture, Agency A should obtain such data from the relevant **Trusted Centre(s)**.
2. (P3) Vendor should only be allowed to do analysis on a **Virtual Desktop Infrastructure (VDI)** which prevents data transfer. The vendor or public officer is able to view the data required, but will not be able to copy out the data.
3. (P4) Instead of downloading the whole database for analysis, only the **relevant fields and records** should be extracted for the purpose of the study.
4. (T1 and P2) Once the study has concluded, the vendor **should not be allowed continued access** to the data sets. The data sets provided via the VDI **should be purged** when it is not required for the study anymore.

**Recommendation 1.2:** Enhance the logging and monitoring of data transactions to detect high-risk or suspicious activity.

**Enhance logs and records to more accurately pinpoint high-risk activity and assist in response and remediation**

**P7**: Maintain data lineage

**T3**: Digital watermarking of files

**Detect suspicious activity and alert the user or stop the unauthorised activity automatically**

**T4**: Enhanced logging and active monitoring of data access

**T5**: Email data protection tool

**T6**: Data loss protection tool

# P7 – Maintain data lineage

What:

Identify any unauthorised modification and usage of data flows, and support the remediation of an unauthorised modification to the data.

Why:

Logging and monitoring allows the Government to detect suspicious or high risk activity, and take immediate action to resolve this to prevent data compromises.

How:

- By maintaining a record of **where the data is used**, **how the data is transferred**, **how the data has been changed** and **who has used the data** for what purposes, the agency is able to identify and rectify this unauthorised modification.

# T3 – Digital watermarking of file

What:

Enable investigators to trace from whom the dataset originated from in the event of a data incident.

Why:

Logging and monitoring allows the Government to detect suspicious or high risk activity, and take immediate action to resolve this to prevent data compromises.

How:

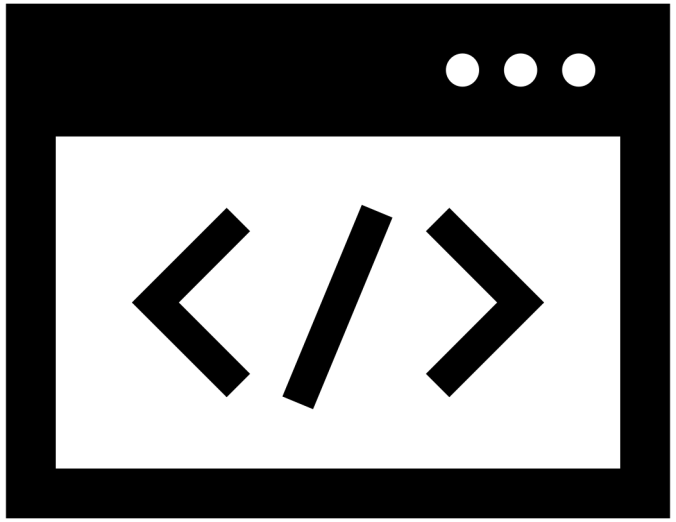
- Adding marking information, such as cryptographic signature. The watermark information can identify the originator of the dataset, prove the authenticity of the file, and is hard to remove from the file.

# Digital watermarking of file

Techniques:

- Steganography: Hiding information within the file which cannot be easily seen
- Cryptographic signatures in file metadata: Embedding cryptographic signatures into file metadata, which can tell us about the integrity of the file and where it originated from

# T3 – Digital watermarking of file



Code Demo

# T4 – Enhanced logging and active monitoring of data access

What:

Keep logs and analyse them to flag out anomalous activity as well as support remediation in the event of a data breach.

Why:

Logging and monitoring allows the Government to detect suspicious or high risk activity, and take immediate action to resolve this to prevent data compromises.

How:

- Logging of data access to sensitive data at greater detail, such as at the individual data query level. The logs should be protected from accidental or deliberate erasure, so that they can reliably show what data has been compromised, how it has been compromised and who was involved.
- Active monitoring of data access to sensitive data by proactive scanning of log files for anomalous data access behaviours, and active checking of data access endpoints' for compliance with data security rules.

# T5 – Email data protection tool

What:

Ensures email senders double-check that they intend to send any email with potentially risky activities (e.g. containing sensitive data, or to suspect addressees) to prevent any accidental or unauthorised disclosure through email.

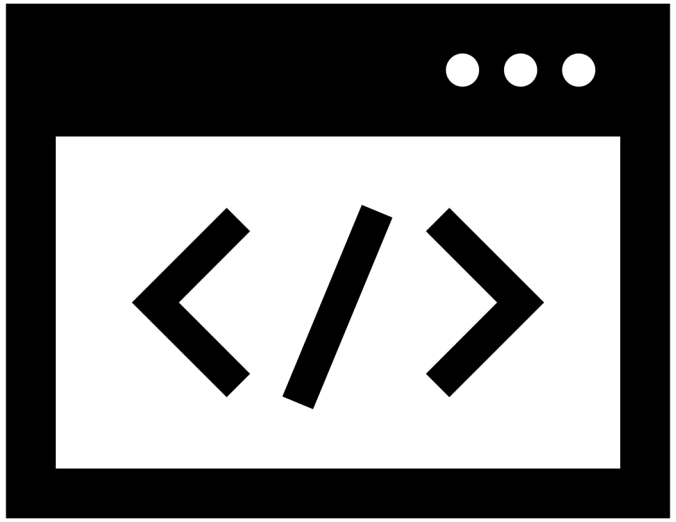
Why:

Prevent any accidental or unauthorised disclosure through email

How:

- Email tools which scan for potentially risky activities (e.g. embedded files in file attachments, large numbers of users in cc lists, email content contains identifiers such as NRICs) and require users to positively affirm that they intend to proceed with the potentially risky activities.

# T5 – Email data protection tool



Code Demo

# T6 – Data loss protection tool

What:

Prevents anomalous activities that are likely to be correlated with malicious activity or data incidents.

Why:

Logging and monitoring allows the Government to detect suspicious or high risk activity, and take immediate action to resolve this to prevent data compromises.

How:

- Monitor computer network and files for anomalous activities (e.g. unexpected downloads of large amounts of data to personal computers) and stop any unauthorised file transfers.

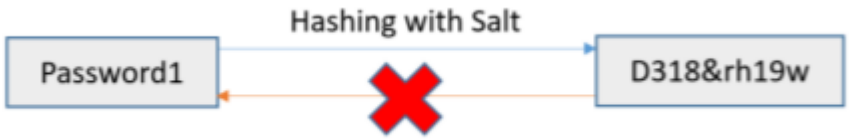
**Recommendation 1.3:** Protect the data directly when it is stored and distributed to render the data unuasable even when extracted or intercepted.


**Render data unusable even if exfiltrated from storage**
**T7:** Hashing with salt
**T8:** Tokenisation
**T9:** Field-level encryption
**P8:** Managing keys to these safeguards

**Partially hide the full data**
**T10:** Obfuscation/ masking/ removal of entity attributes
**T11:** Dataset partitioning

**Protect the data during distribution**
**T12:** Password protecting and encrypting data files
**P9:** Securely distribute password out-of-band
**T13:** Data file integrity verification
**P10:** Distribute files through appropriate secure channels

# T7 – Hashing with salt*




What:

Ensure that sensitive values (e.g. identifiers) cannot be seen or reasonably recovered in the event of a compromise.

Why:

This ensures that even if an attacker were to break into the IT system, the data would be unusable to the attacker.

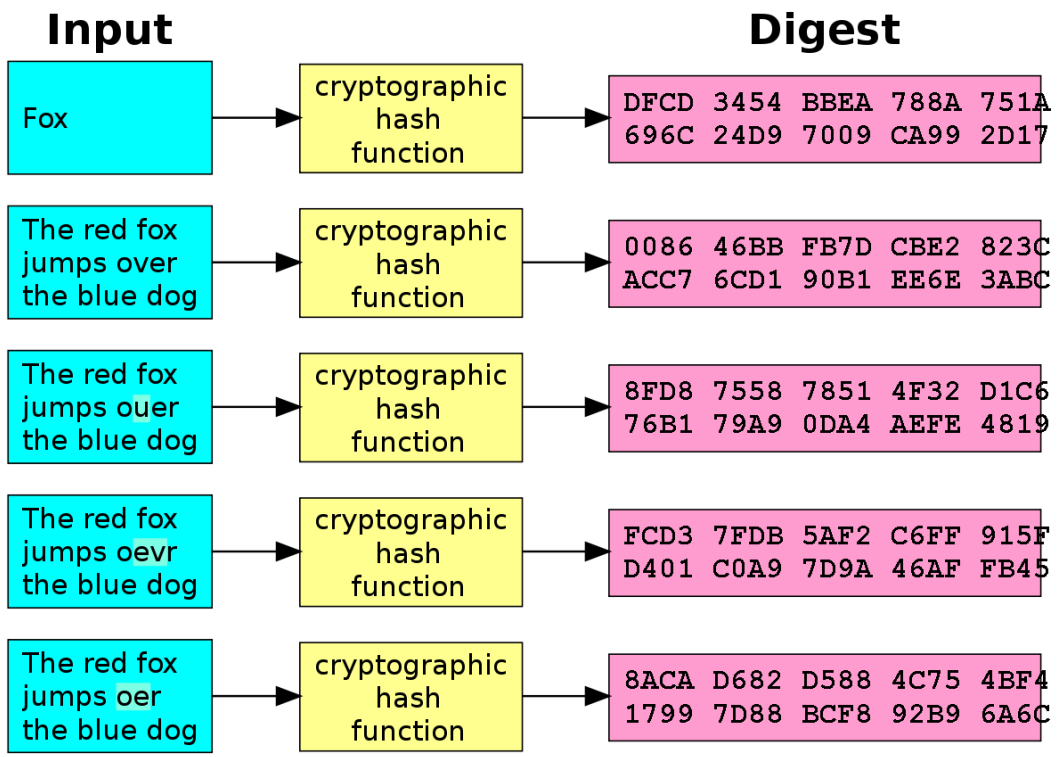
How:

- Replace sensitive values (e.g. identifiers) with an algorithmically derived value that cannot be reversed easily.

*Note: this measure is only appropriate for data fields where the actual values need not be recovered, such as passwords or for aggregated data analytics. Strong hashing functions should be used, such as cryptographic hash function, that cannot be reversed with current computing resources.

# Hashing with salt

- One directional
- One simple non-cryptographic hash function is the modulo (remainder) operator
- Avalanche effect – one bit change can result in a significant change in the resulting hash
- Add salt to protect against rainbow table attacks*
- Available cryptographic hashing algorithms:
  - MD5
  - SHA1
  - SHA256



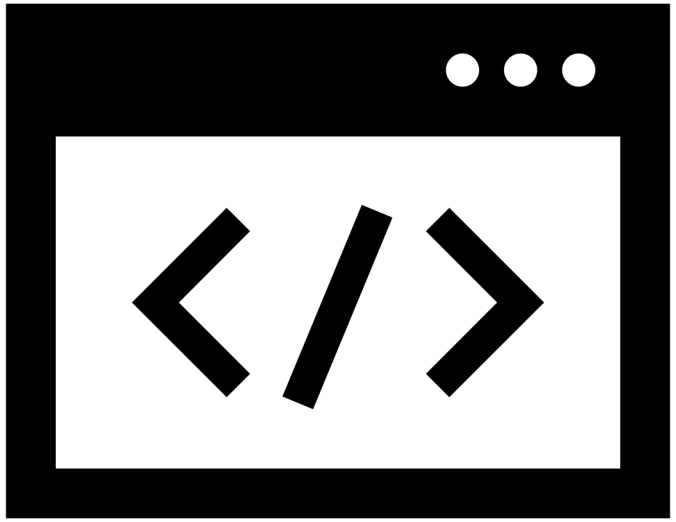

* Rainbow tables: pre-computed table of password hash chains used to crack hashed passwords

# Hashing algorithms at a glance

| Algorithm and variant | | Output size (bits) | Internal state size (bits) | Block size (bits) | Rounds | Operations | Security (in bits) against collision attacks | Capacity against length extension attacks | Performance on Skylake (median cpb)[1] | | First published |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | | long messages | 8 bytes | |
| MD5 (as reference) | | 128 | 128 (4 × 32) | 512 | 64 | And, Xor, Rot, Add (mod $2^{32}$), Or | ≤18 (collisions found)[2] | 0 | 4.99 | 55.00 | 1992 |
| SHA-0 | | 160 | 160 (5 × 32) | 512 | 80 | And, Xor, Rot, Add (mod $2^{32}$), Or | <34 (collisions found) | 0 | ≈ SHA-1 | ≈ SHA-1 | 1993 |
| SHA-1 | | | | | | | <63 (collisions found)[3] | | 3.47 | 52.00 | 1995 |
| SHA-2 | *SHA-224* | 224 | 256 (8 × 32) | 512 | 64 | And, Xor, Rot, Add (mod $2^{32}$), Or, Shr | 112 | 32 | 7.62 | 84.50 | 2004 |
| | *SHA-256* | 256 | | | | | 128 | 0 | 7.63 | 85.25 | 2001 |
| | *SHA-384* | 384 | 512 (8 × 64) | 1024 | 80 | And, Xor, Rot, Add (mod $2^{64}$), Or, Shr | 192 | 128 (≤ 384) | 5.12 | 135.75 | 2001 |
| | *SHA-512* | 512 | | | | | 256 | 0 | 5.06 | 135.50 | |
| | *SHA-512/224* | 224 | | | | | 112 | 288 | ≈ SHA-384 | ≈ SHA-384 | 2012 |
| | *SHA-512/256* | 256 | | | | | 128 | 256 | | | |
| SHA-3 | *SHA3-224* | 224 | 1600 (5 × 5 × 64) | 1152 | 24[4] | And, Xor, Rot, Not | 112 | 448 | 8.12 | 154.25 | 2015 |
| | *SHA3-256* | 256 | | 1088 | | | 128 | 512 | 8.59 | 155.50 | |
| | *SHA3-384* | 384 | | 832 | | | 192 | 768 | 11.06 | 164.00 | |
| | *SHA3-512* | 512 | | 576 | | | 256 | 1024 | 15.88 | 164.00 | |
| | *SHAKE128* | *d* (arbitrary) | | 1344 | | | min(*d*/2, 128) | 256 | 7.08 | 155.25 | |
| | *SHAKE256* | *d* (arbitrary) | | 1088 | | | min(*d*/2, 256) | 512 | 8.59 | 155.50 | |

*n*-bit security means that the attacker would have to perform $2^n$ operations to break it

# T7 – Hashing with salt



Code Demo

# T8 – Tokenization*


What:

Ensure that identifiers cannot be seen in the event of a compromise

Why:

This ensures that even if an attacker were to break into the IT system, the data would be unusable to the attacker.

How:

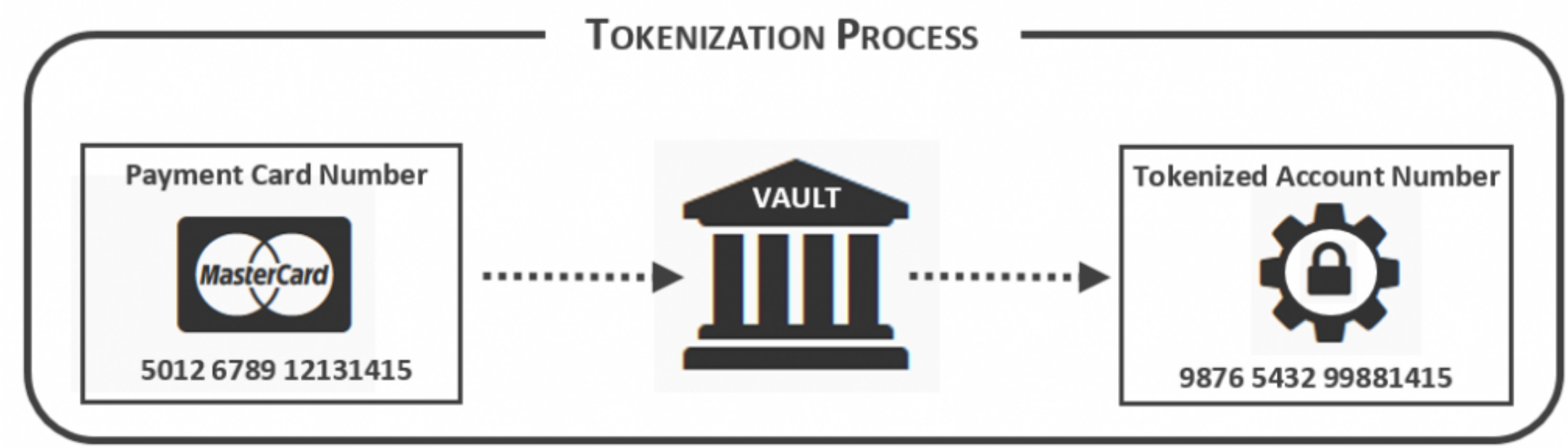
- Replace identifiers and attributes with a different value known only to the agency. The different values are randomly generated and saved in a lookup table/token vault

*Note: the underlying technique of field-level encryption achieves the same function as “T9. Field Level Encryption”.

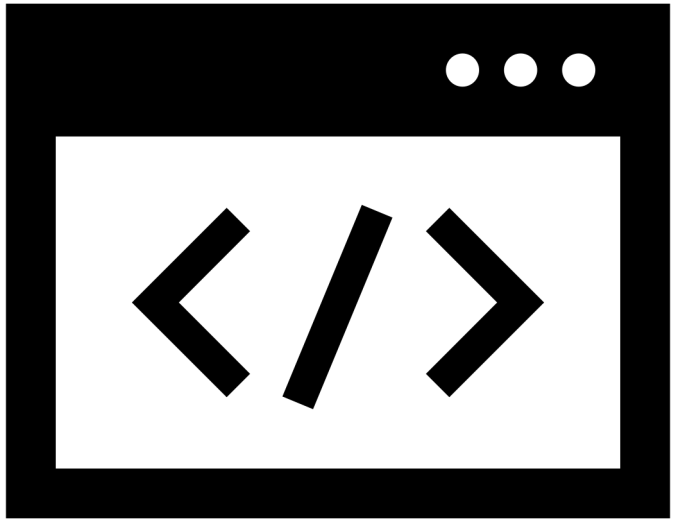
This technique is appropriate for data fields where the actual values need to be recovered, such as identifiers required for service delivery

# Tokenization

- Randomly generates a token value for each original value and store the mapping in a token vault



# T8 – Tokenization



Code Demo

# T9 – Field-level encryption*


What:

Ensure that sensitive values cannot be seen in the event of a compromise

Why:

This ensures that even if an attacker were to break into the IT system, the data would be unusable to the attacker.

How:

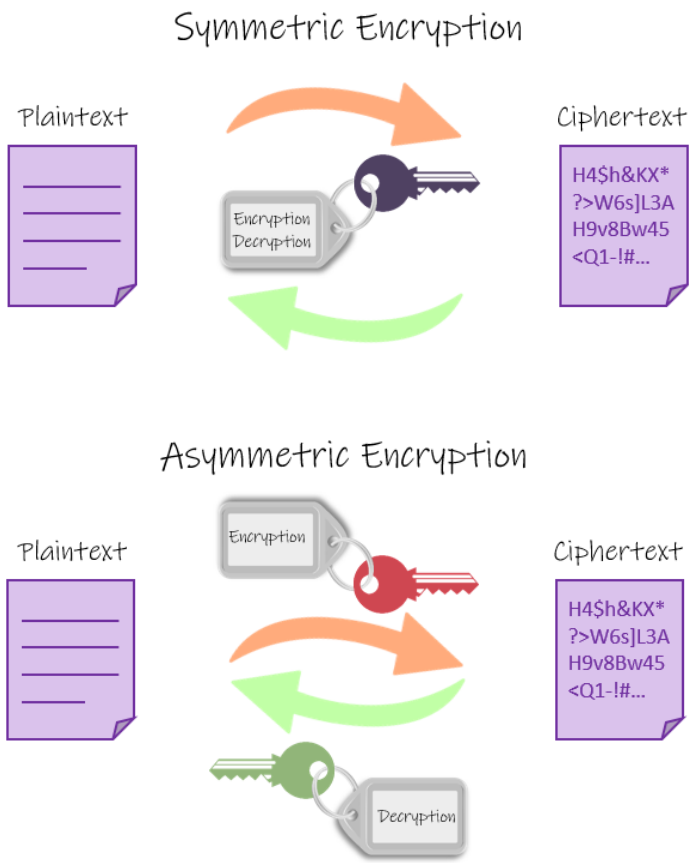
- Encrypting specific data fields to hide the true value. A different secret encryption key is to be used for each field.

*Note: the underlying technique of field-level encryption achieves the same function as “T8. Tokenisation”.

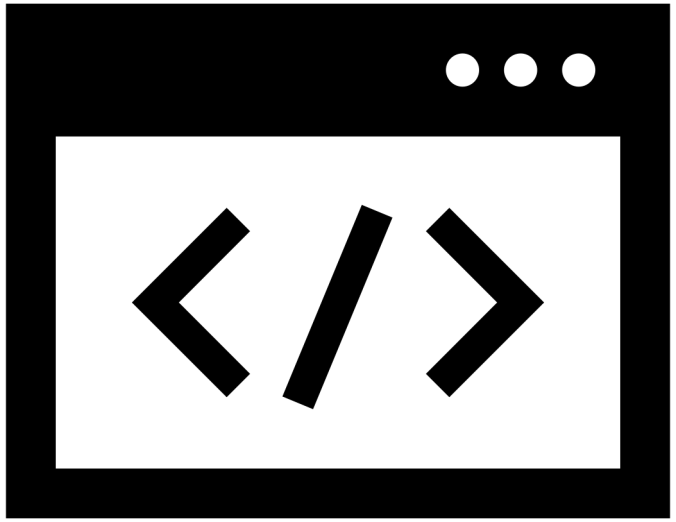
The technical implementation of field-level encryption uses a mathematical encryption function instead of a lookup table for tokenisation. Field-level encryption is more appropriate where the actual values need to be frequently recovered. (e.g. deidentification of data set is required by other agencies as part of their BAU usage, de-tokenization will be too slow for such a purpose, instead the field is encrypted at rest but decrypted when accessing it)

# Encryption basics

- Two way functions (i.e. can encrypt and decrypt)
- Symmetric
  - Only 1 key which does both encryption and decryption
  - E.g. AES
- Asymmetric
  - 2 keys (a.k.a keypairs), when one is used for encryption only the other key can be used to decrypt the data
  - E.g. RSA algorithm



# T9 – Field-level encryption (symmetric key)



Code Demo

# P8 – Manage keys to data protection technical safeguards

What:

Ensure the effectiveness of the technical safeguards of tokenisation and field-level encryption by keeping the “key” safe

Why:

This ensures that even if an attacker were to break into the IT system, the data would be unusable to the attacker.

How:

- Implement processes to ensure that the officer holding the key is not the same person as the officer using the data.

# T10 – Obfuscation/ masking/ removal of entity attributes*


What:

Ensure that the exact sensitive values cannot be seen or ever recovered in the event of a compromise, although approximate or noisy values might still be seen.

Why:

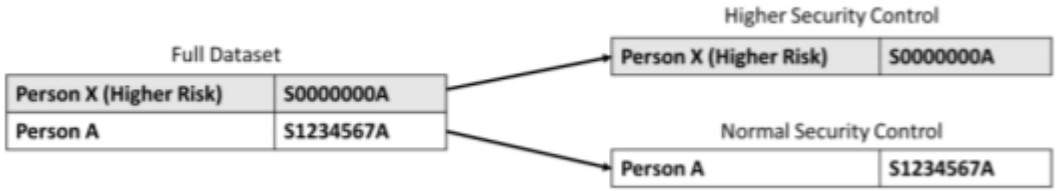
This ensures that even if an attacker were to break into the IT system, the damage would be limited as he/she would have no access to the full data.

How:

- Hide the true value of the attributes by adding noise, banding the data, or masking out portions of the value. Attributes not relevant for data usage should be removed.

*Note: This measure is appropriate where the exact values are sensitive, but noisy values (that are less sensitive) are sufficient for usage and exploitation.

# T11 – Dataset partitioning




What:

Ensure that information on selected entities or attributes will not be compromised even if the larger database has been compromised. This could include individuals in vulnerable positions or sensitive attributes.

Why:

This ensures that even if an attacker were to break into the IT system, the damage would be limited as he/she would have no access to the full data.

How:

- Break a dataset into smaller datasets by segmenting out selected entities or attributes and apply different access controls to each of the partitions.
- The partitioning of the dataset can be achieved by either: (a) physically partitioning of the dataset in different storage locations; or (b) virtually partitioning of the datasets in different virtually isolated partitions. Each dataset partition, physical or virtual, would have different access controls.

# T12 – Password protecting and encrypting files

What:

Ensure that only the receiver with the password can access the file.

Why:

Another common source of data incidents is when a public officer mistakenly distributes the data file to unauthorised parties, for example, via email. An unauthorised party could also intercept the data during transit. Data must therefore be protected during the distribution phase as well.

How:

- Secure a file using encryption and password such that only authorised users can access and change the content.

# P9 – Securely distribute passwords out-of-band

What:

Ensure the effectiveness of the technical safeguards of password protecting and encrypting files by sending the passwords to secured files via a separate message on a different channel.

Why:

Another common source of data incidents is when a public officer mistakenly distributes the data file to unauthorised parties, for example, via email. An unauthorised party could also intercept the data during transit. Data must therefore be protected during the distribution phase as well.

If the password was sent together with the email, the unauthorised party would have access to the underlying dataset.

How:

- Password to encrypted file is transmitted through a different channel, such as through text or call or trusted Instant Messaging services.

# T13 – Data file integrity verification

What:

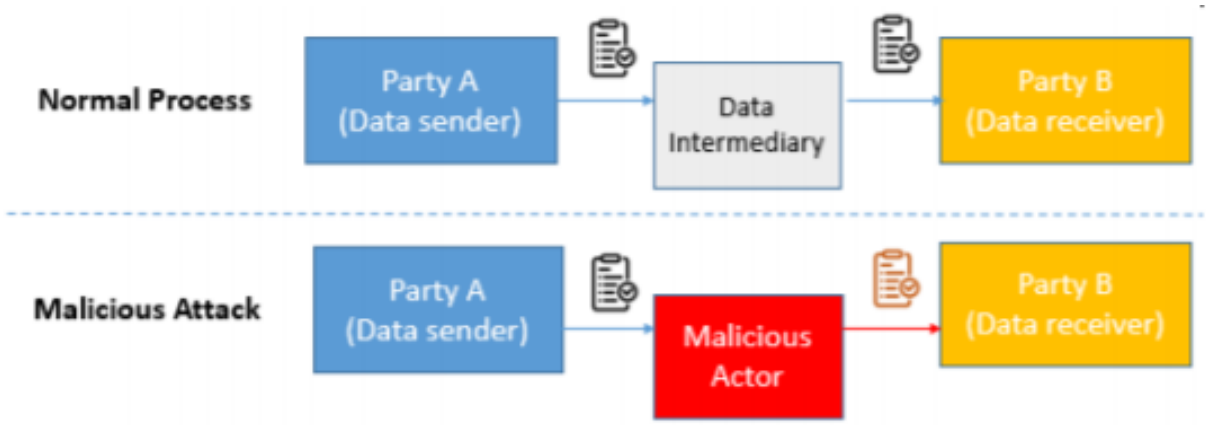
Ensure that the receiver gets the same file that the original sender intended.

Why:

This measure allows the recipient to verify that the data is the same as what the sender has sent, and protects such data against malicious attacks.

How:

- Original data sender provides a checksum or digital signature that confirms the integrity of a data file.

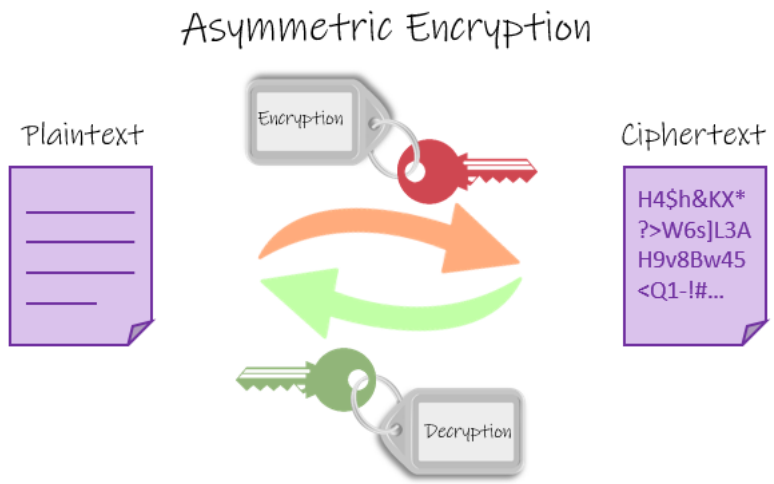


# File integrity verification with Asymmetric Key Encryption (Public Key Cryptography)

## What is public key cryptography?

A cryptographic system that uses asymmetric key encryption algorithm to:

1. Ensure that **only the intended recipient** can decrypt the file/message
2. Ensure that the file/message is **sent by the correct sende**r
3. Ensure that the file/message is **not corrupted or tampered with** (i.e. File integrity verification)



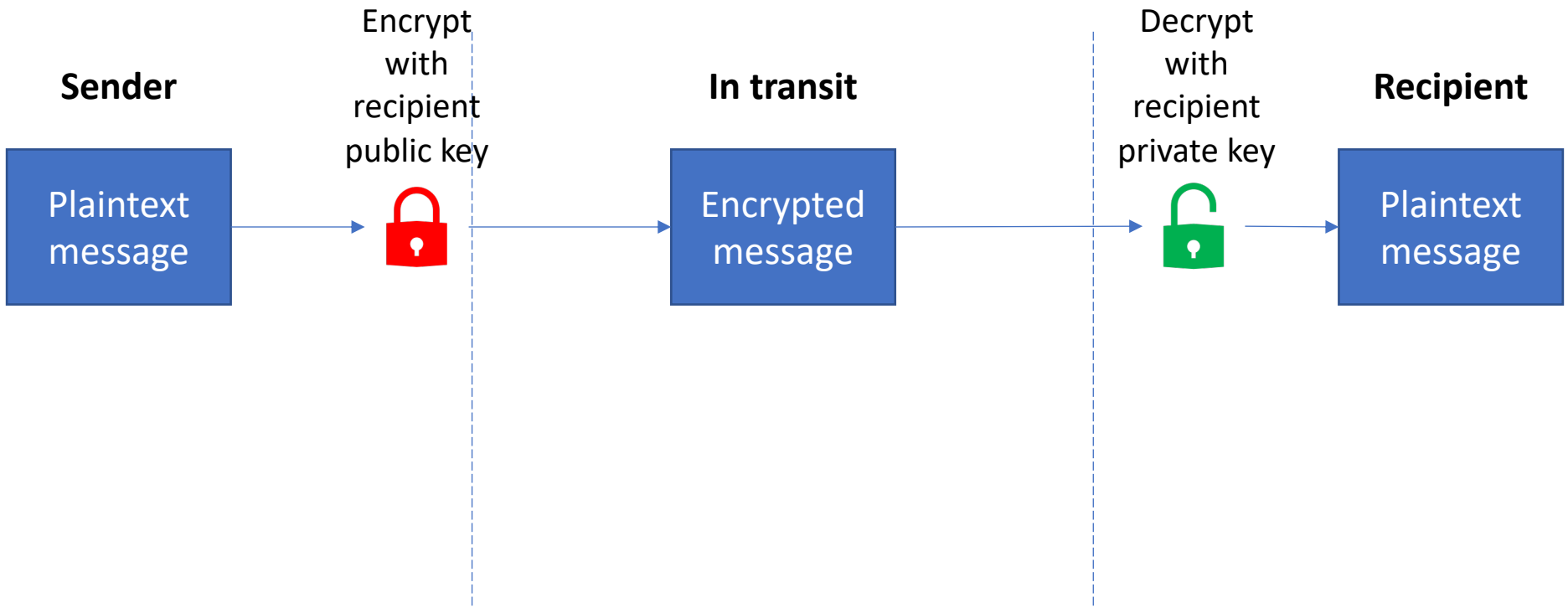

## Recap on what is asymmetric key encryption:

- Encryption keys come in **pairs**
- When one key is used to encrypt, **only the other key** can be used to decrypt (and vice versa)
- In public key cryptography, one key would be a **private key** (known only to the owner) and the other will be a **public key** (known to everyone)

# File integrity verification with Asymmetric Key Encryption (Public Key Cryptography)

Feature 1:

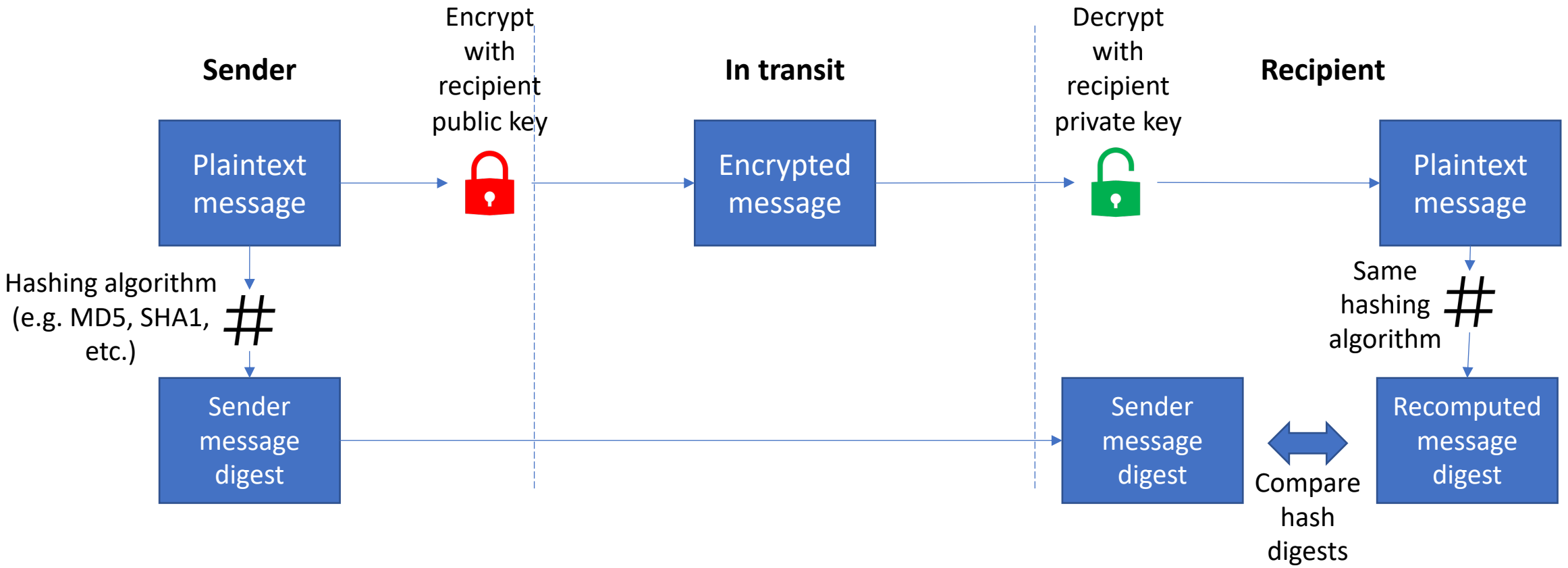
Ensure that **only the intended recipient** can decrypt the file/message



# File integrity verification with Asymmetric Key Encryption (Public Key Cryptography)

Feature 2:

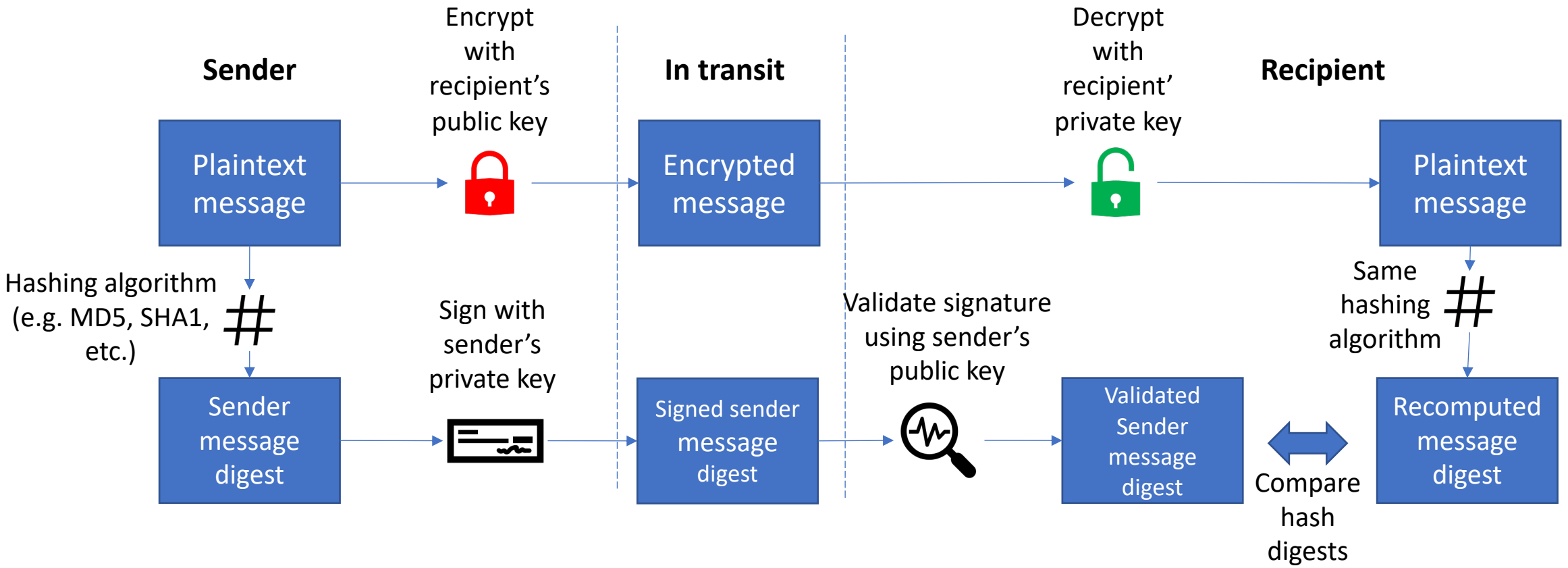
Ensure that message is **not corrupted or tampered with**



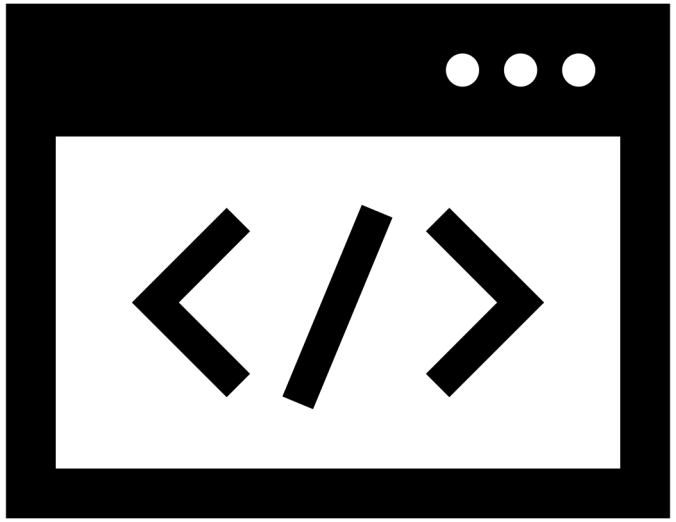
# File integrity verification with Asymmetric Key Encryption (Public Key Cryptography)

Feature 3:

Ensure that file is received from **correct sender**



# T13 – Data file integrity verification



Code Demo

# P10 – Distribute files through appropriate secure channels

What:

Ensure that the distribution channels for sensitive files have the appropriate security safeguards.

Why:

To protect data during the distribution phase from being intercepted.

How:

- The distribution channel that the officer uses should encrypt the file during transmission so that any attacker who maliciously intercepts the file would not be able to retrieve the original contents. (e.g. using HTTPS, SFTP, etc.)

# Quiz time

# Data Anonymization Concepts and Methods

# Data Anonymization Concepts and Methods

- Techniques specified in PSDSRC 2019
  - Hashing with salt
  - Tokenization
  - Encryption (Symmetric/Asymmetric)
  - Obfuscation/Data Masking
- Concepts
  - K-Anonymity
  - Differential Privacy
- Techniques not explicitly specified in PSDSRC 2019
  - Format Preserving Encryption
  - Synthetic Data
  - Homomorphic encryption
  - Data Perturbation
  - Generalization
  - Data Swapping
  - Secure multiparty compute

# Concepts

# K-Anonymity

Definition:

k-anonymity is a ***property*** possessed by certain anonymized data where for k-anonymity to be achieved, there need to be **at least k individuals in the dataset who share the set of attributes** that might become identifying for each individual.

Ways of achieving k-anonymity:

- Generalization:

  replace quasi-identifiers* with less specific, but semantically consistent values until get k identical values and partition ordered-value domains into intervals.

- Suppression:

  when generalization causes too much information loss then the quasi-identifier is omitted, not released at all. This is common with outliers.

*quasi-identifiers: indirect identifiers like gender, religion, postal code, birthday

# K-Anonymity

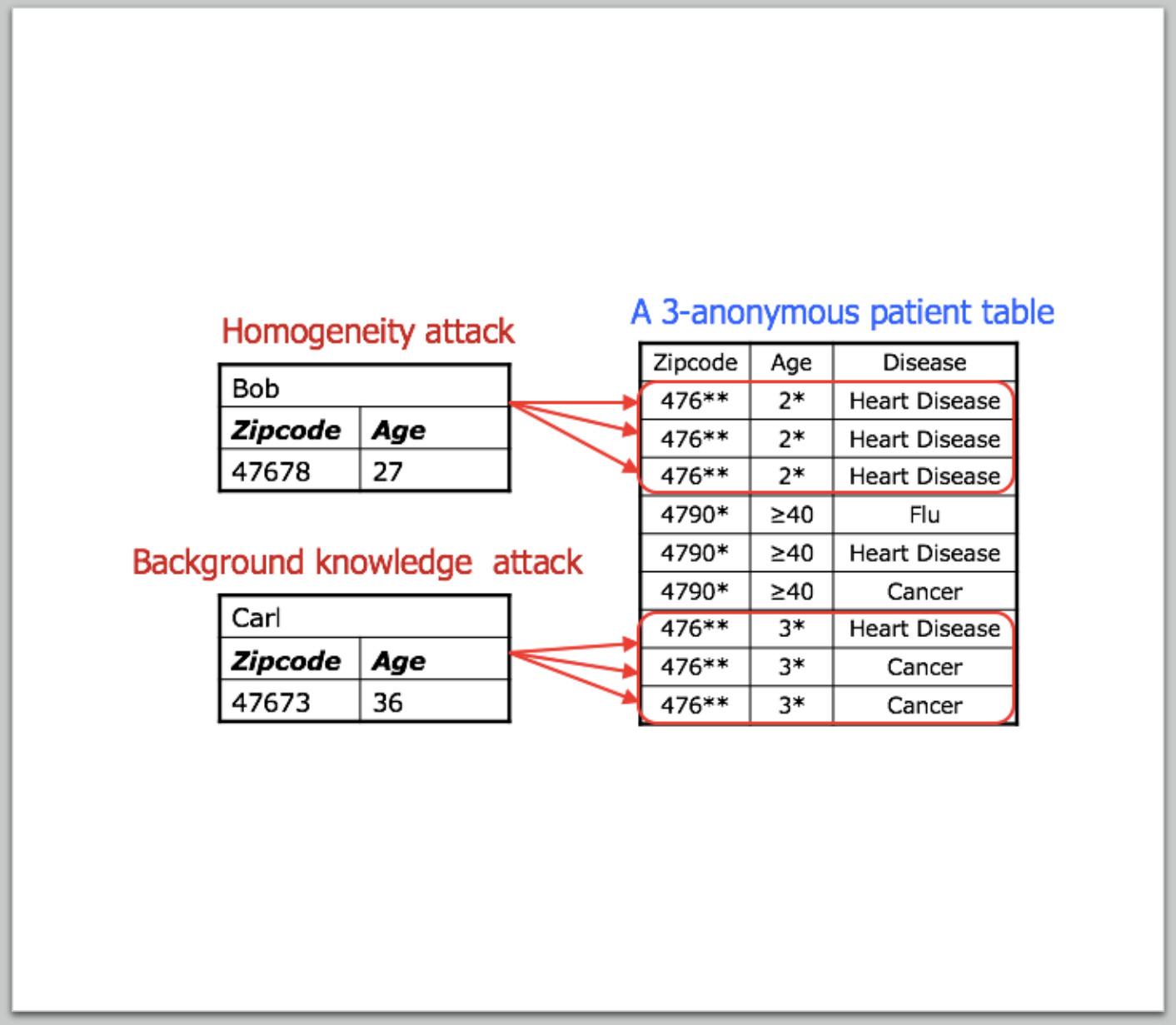
Attacks on K-Anonymity:

- Homogeneity Attack:

  This attack leverages the case where all the values for a sensitive value within a set of *k* records are identical. In such cases, even though the data has been *k*-anonymized, the sensitive value for the set of *k* records may be exactly predicted.

- Background Knowledge Attack:

  This attack leverages an association between one or more quasi-identifier attributes with the sensitive attribute to reduce the set of possible values for the sensitive attribute.



# Differential privacy (ϵ)

Definition:

Differential privacy in itself is not a technology. It's a ***property*** that describes some systems — a mathematical guarantee that your privacy won't be violated if your data are used for analysis. A system that is ***differentially private*** allows analysis while protecting sensitive data behind a veil of uncertainty.

Measurement

- ϵ is a measure of how private, and how noisy, a data release is. Higher values of ϵ indicate more accurate, less private answers; low-ϵ systems give highly random answers that don't let would-be attackers learn much at all.
- There's not much consensus about what values of ϵ are actually "private enough." Most experts agree that:
  - values between 0 and 1 are very good,
  - values above 10 are not,
  - and values between 1 and 10 are various degrees of "better than nothing."
- The parameter ϵ is exponential: by one measure, a system with ϵ = 1 is almost three times more private than ϵ = 2, and over 8,000 times more private than ϵ = 10.

# Differential privacy

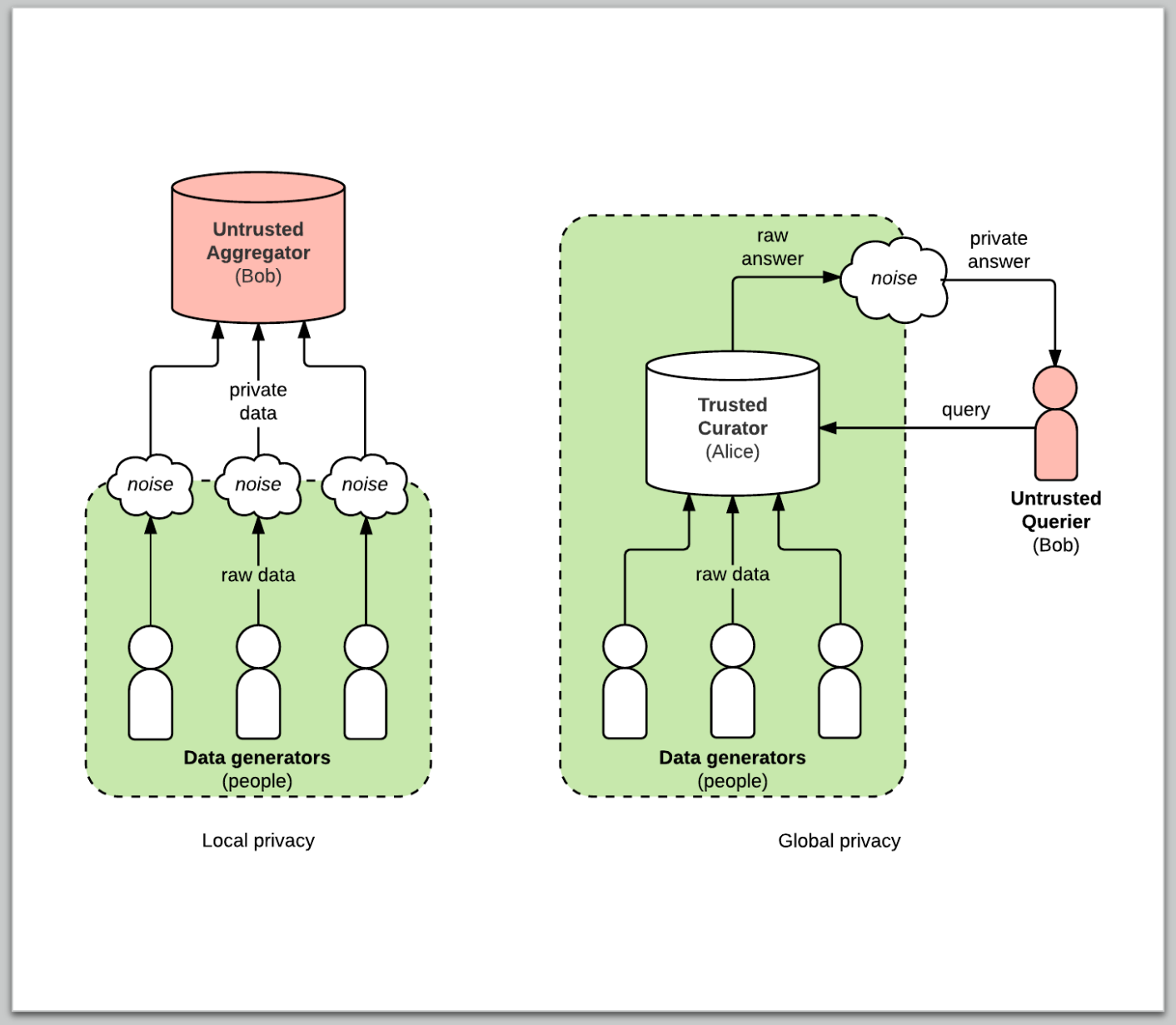
Ways of achieving ϵ differential privacy:

- Global:

  In a globally private system, one trusted party (or curator) has access to raw, private data from lots of different people. The curator does analysis on the raw data and adds noise to answers after the fact.

- Local:

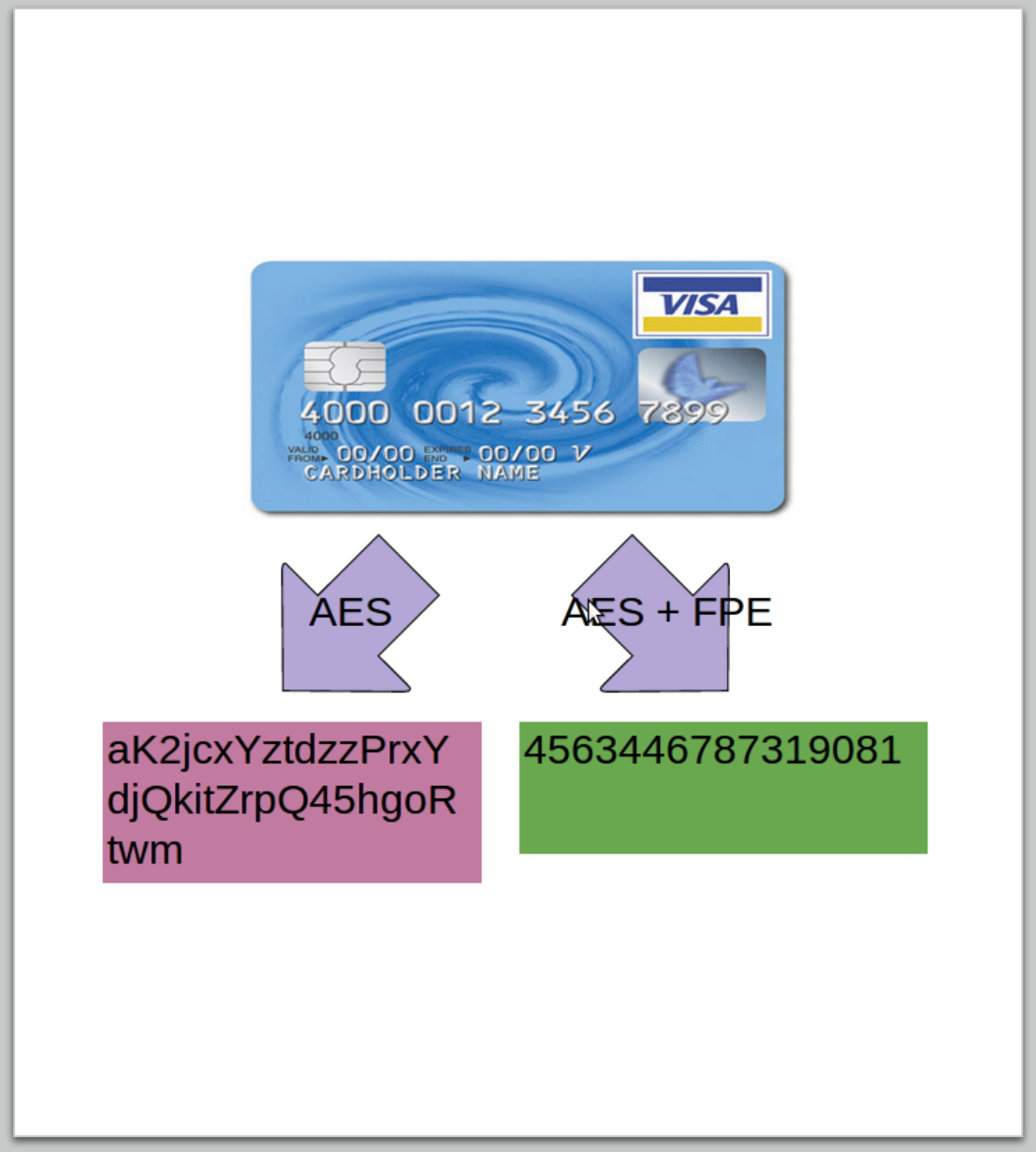
  With local privacy, there is no trusted party; each person is responsible for adding noise to their own data before they share it.



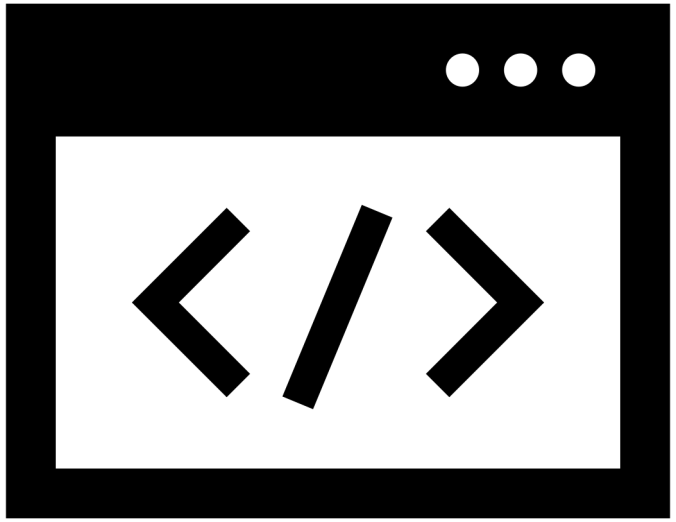
# Other techniques not explicitly specified in PSDSRC 2019

# Format Preserving Encryption (FPE)

- Some systems are not able to accept the format of encrypted ciphertext and can only store data in the same format as the plain text
- FPE is a technique which convert the ciphertext into the same format as the plain text



# Format Preserving Encryption (FPE)



Code Demo

# Synthetic Data Generation

## Definition:

Synthetic dataset is a repository of data that is generated programmatically that attempts to model the original data but at the same time preserve privacy. This allows data mining to be done but with reduced risk of sacrificing privacy.

## Categories:

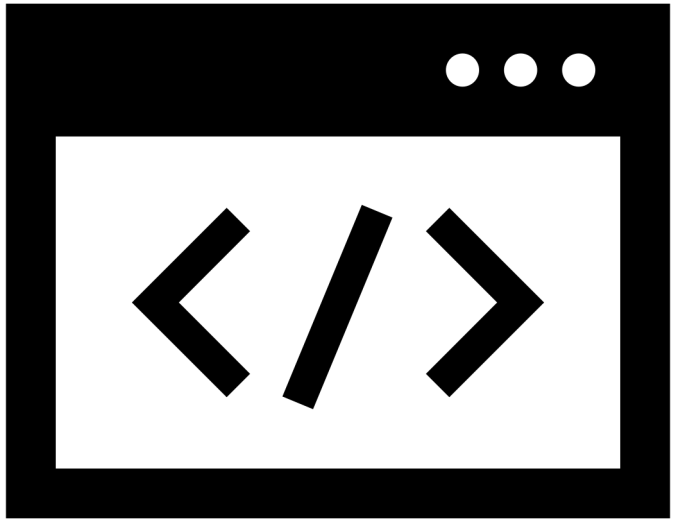
- Fully synthetic:

  Fully synthetic data generators identify the density function of attributes in the original data and estimate the parameters of these density functions. Then for each attribute, privacy protected series are generated by randomly picking up the values from the estimated density functions.

- Partially synthetic:

  Partially synthetic data generators replaces only values of the selected sensitive attribute with synthetic values. The original values are replaces only if it posses high risk of disclosure. Masking the original values with synthetic values prevents re-identification thus preserving privacy in the published data.

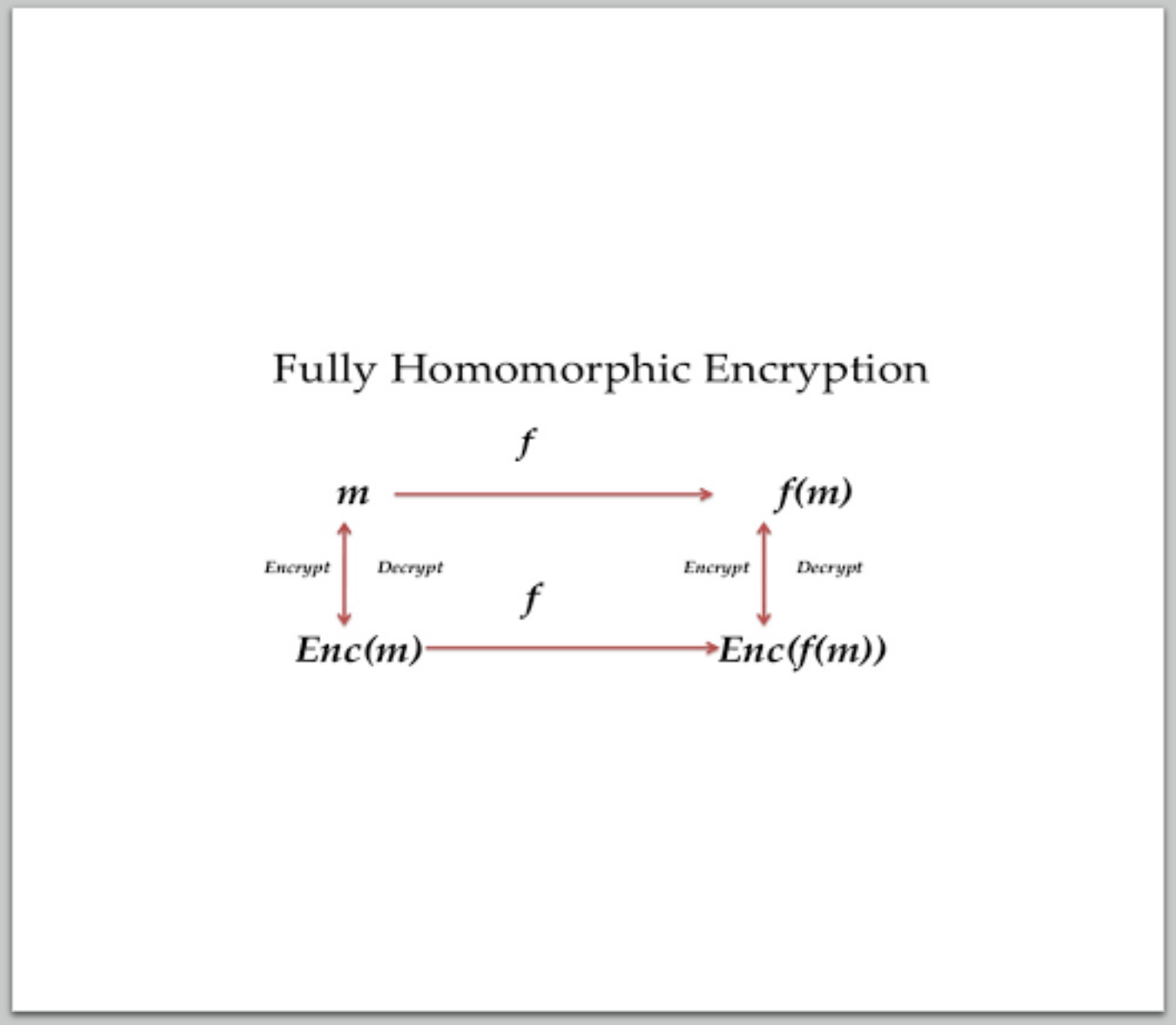
# Synthetic Data Generation



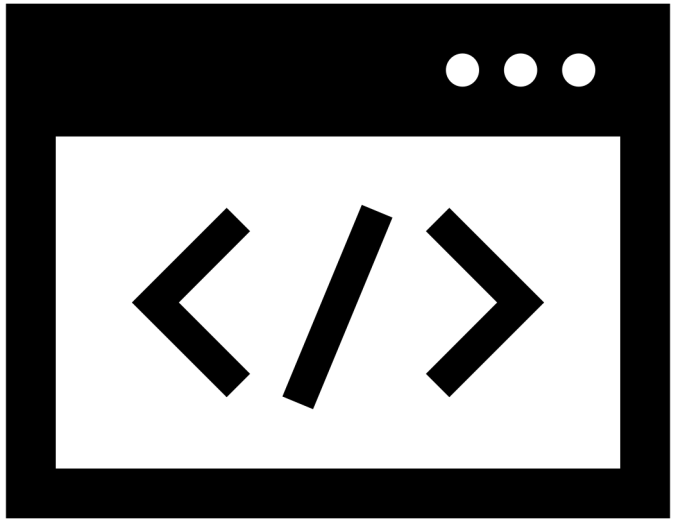
Code Demo

# Homomorphic encryption

- The purpose of homomorphic encryption is to allow computation on encrypted data.
- Thus data can remain confidential while it is processed, enabling useful tasks to be accomplished with data residing in untrusted environments.
- Example of Partially Homomorphic Encryption is the RSA algorithm where addition and multiplication could be done
- One of the recent Full Homomorphic Encryption is CKKS which supports rounding operations and is intended to be used for encrypted machine learning



# Homomorphic encryption



Code Demo

# Data Perturbation

Definition:

Data perturbation is a data security technique that adds ‘noise’ to databases allowing individual record confidentiality.

Categories:

- Probability distribution perturbation:

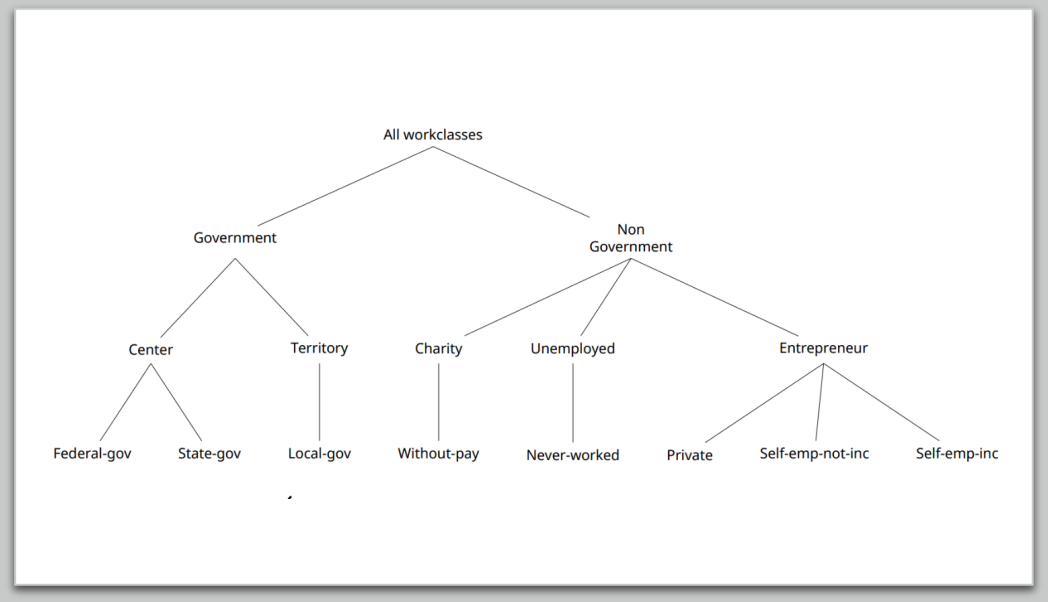
  considers the selected data and replaces it from the same distribution sample or from the distribution itself

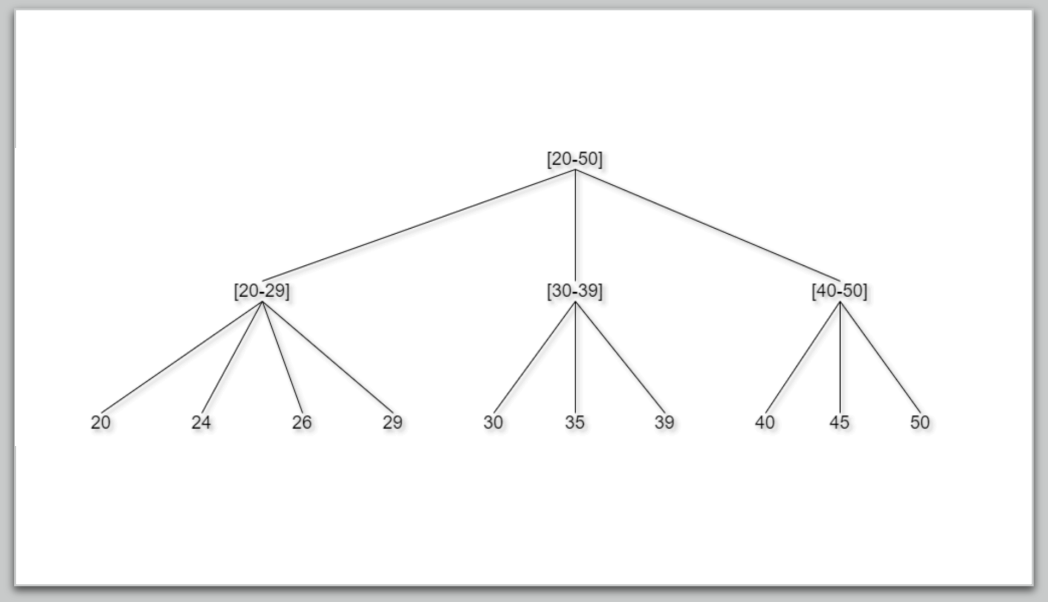
- Value distortion perturbation :

  perturbs data by multiplicative or additive noise, or other randomized processes

# Generalization

- replace a data value with a less precise one via binning, reformatting, rounding or truncating, which preserves data utility and protects against linkage attacks*




*linkage attacks: adversaries collect auxiliary information about a certain individual from multiple data sources and then combine that data to form a whole picture about their target, which is often an individual's personally identifiable information.

# Data swapping

- interchanging values of individual records so that removes the relationship between the record and the respondent while still preserving some of the statistical frequency counts

**Original Responses**

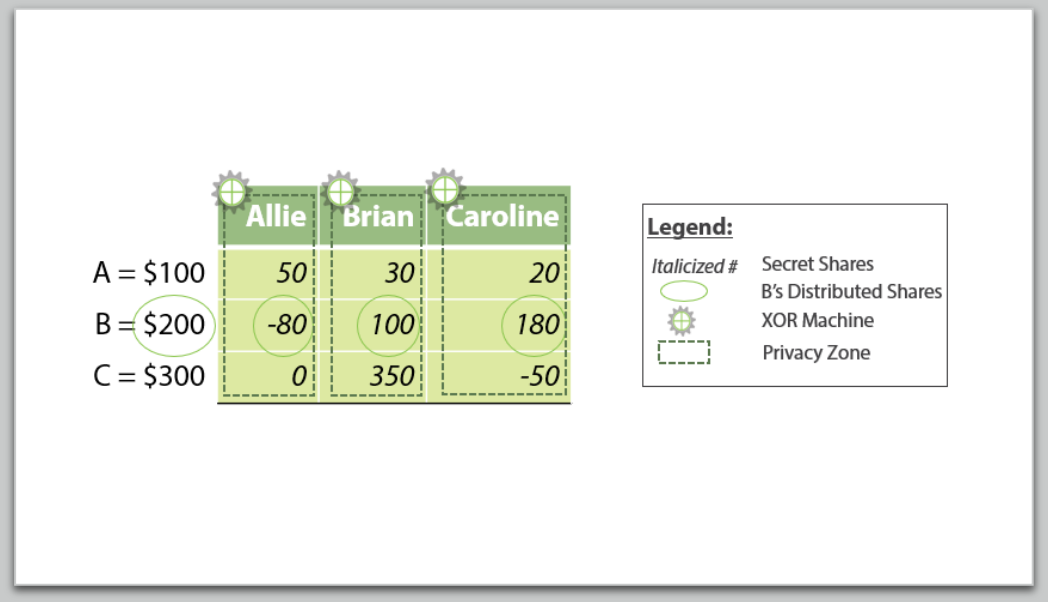
| # | Age | Income |
|---|---|---|
| 1 | 21 | 20,000 |
| 2 | 24 | 30,000 |
| 3 | 35 | 30,000 |
| 4 | 36 | 25,000 |
| 5 | 45 | 55,000 |
| 6 | 50 | 15,000 |

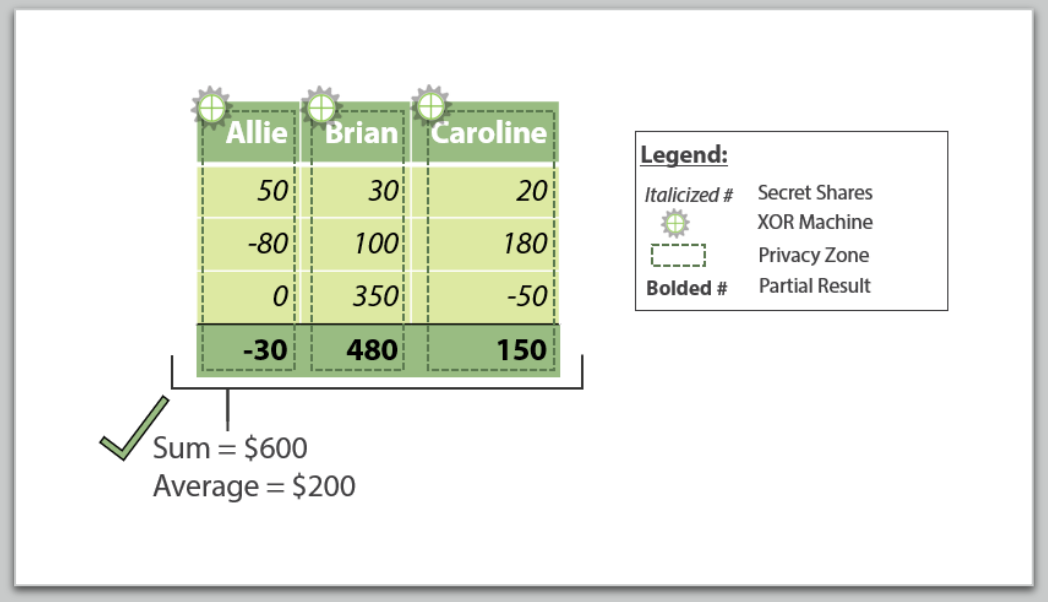
**Responses After Swap #1**

| # | Age | Income |
|---|---|---|
| 1 | 21 | 15,000 |
| 2 | 24 | 30,000 |
| 3 | 35 | 30,000 |
| 4 | 36 | 55,000 |
| 5 | 45 | 25,000 |
| 6 | 50 | 20,000 |

# Secure Multiparty Compute

- Secure multiparty computation (MPC / SMPC) is a cryptographic protocol that distributes a computation across multiple parties where no individual party can see the other parties' data.
- In this simplified example (right), Allie, Brian and Caroline's salaries are $100, $200 and $300 respectively. These are then broken down and distributed to 3 separate secret shares held by Allie, Brian and Caroline respectively.
- When held separately, the numbers are not useful, but when added up, it becomes useful.



# Summary

- Data is the **lifeblood of the digital economy and a digital government**.
- Data privacy is an enabler that allows us to **use and share data as fully as possible** to provide better public services.
- When in doubt, do check the relevant **policies in the IM8** and **obligations in the PDPA**.

Questions?