Data Anonymization
A ‘Risky’ affair?

Manjusha Gode

Data Anonymization – A 'Risky' affair?

Agenda

- Basics
- Attacks and risks
- Process of document anonymization
  - Anonymization
  - Risk assessment
- Risk of re-identification
- Challenges
- Conclusion

Basics

- Attack – effort to re-identify an individual from data
- Context – where the data will be shared, intended use
- Risk Threshold – Acceptable value of risk
- Information loss/Data Utility
- Identifiers – Direct and Quasi

Attacks

Four types of plausible attacks:

- Deliberate
- Inadvertent
- Breach
- Demonstration

First three are typically in a restricted environment whereas fourth one is in public domain.

Why to bother?

- Attacks can be real!
- Legal, Financial and reputational impact

Risk Mitigation Mechanisms

- Security and controlled access
- Motive and capacity
- Contracts

Methods of Anonymization

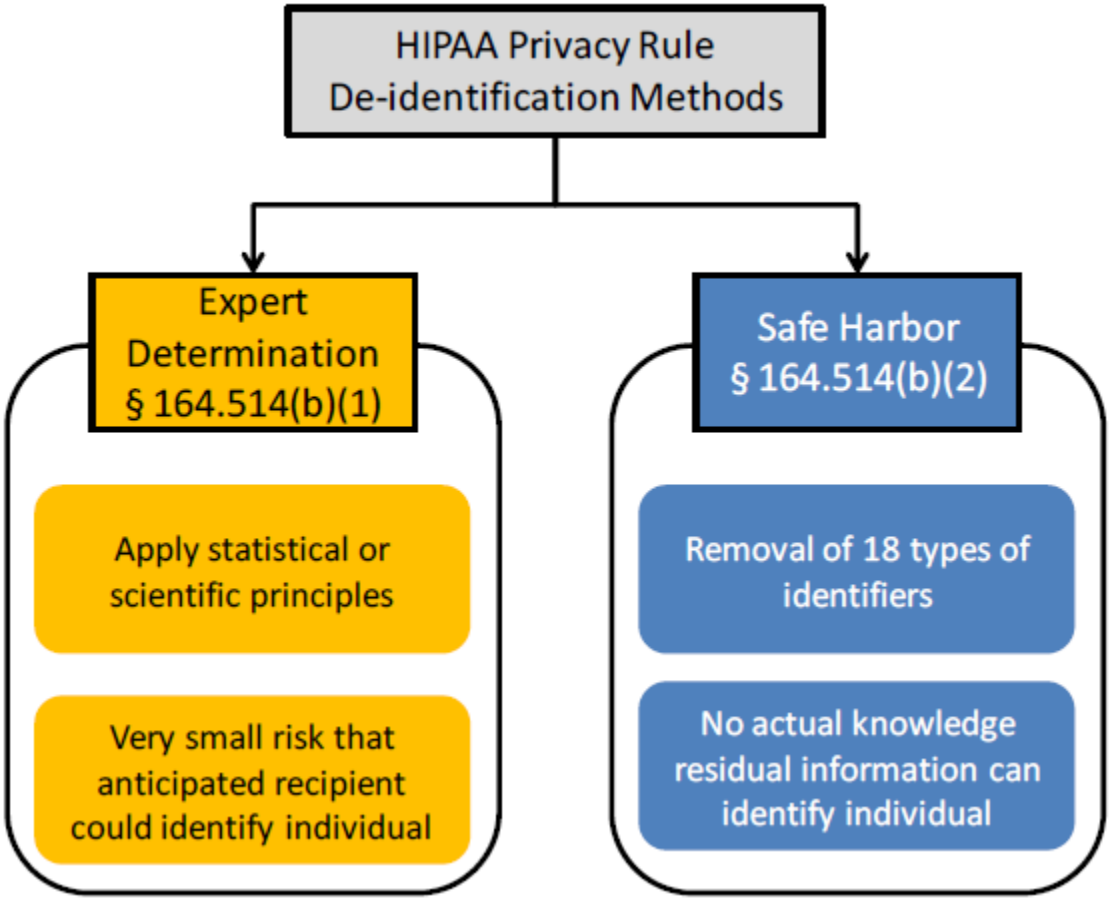
Masking

- Suppression
- Randomization
- Pseudonymization

De-identification

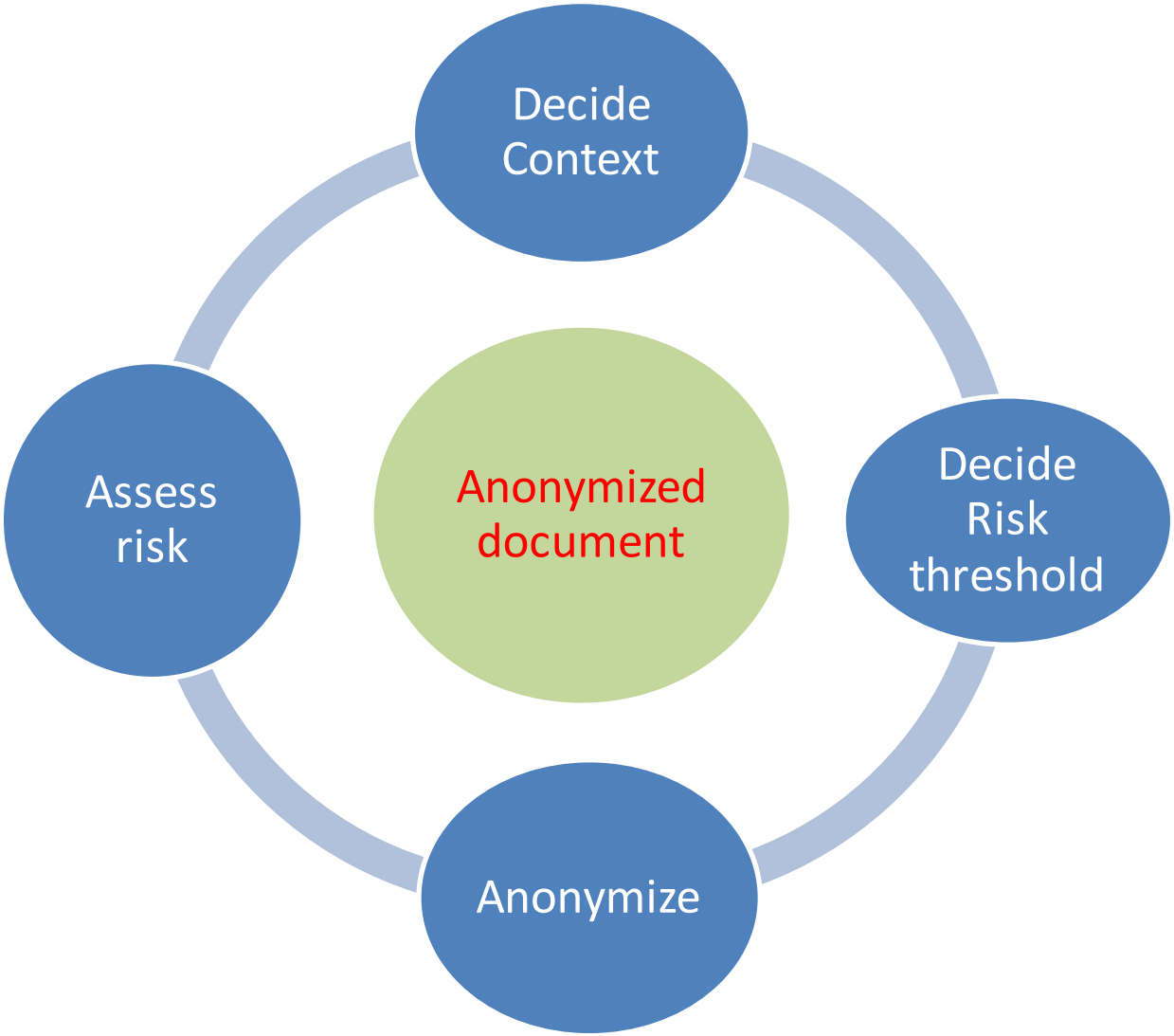
- Generalization
- Sub-sampling
- Suppression

HIPAA – De-identification



1. Names
2. Geographical references
3. Dates (Except Year)
4. Phone numbers
5. Fax numbers
6. Email addresses
7. Social Security numbers
8. Medical record numbers
9. Health plan beneficiary numbers
10. Account numbers
11. Certificate/license numbers
12. Vehicle identifiers
13. Device identifiers and serial numbers
14. Web Universal Resource Locators
15. Internet Protocol (IP) address numbers
16. Biometric identifiers, including finger and voice prints
17. Full face photographic images and any Comparable images;
18. Any other unique identifying number, characteristic

Risk Assessment Process



Risk Assessment

Confusion matrix

	Identified as PHI	Not identified as PHI
PHI	TP	FN
Not PHI	FP	TN

Deals with the recall

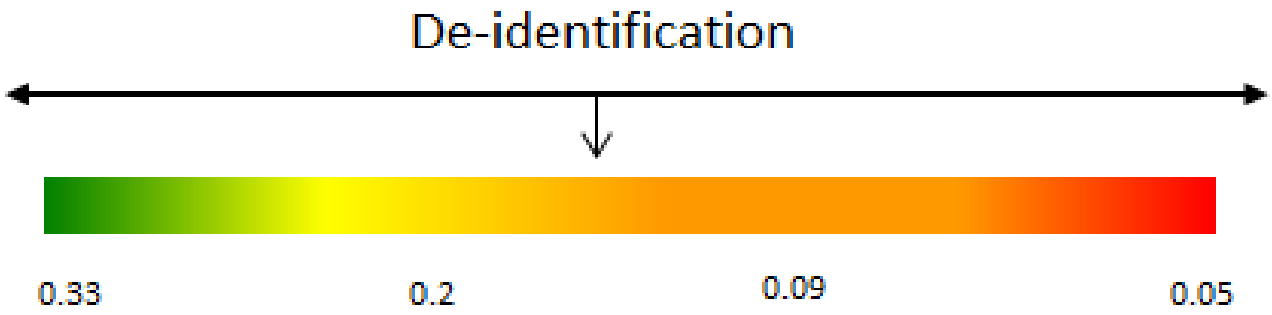
Deals with the Precision

- Catch
- Leak
- Precision
- Recall

Risk Threshold

Non-public trusted recipients

Public Data



- Security controls
- Not sensitive data
- No obvious motive to re-identify the data

- No security controls
- Sensitive data
- Motives to re-identify the data

Risk Calculation

1. **Deliberate:**

$Pr(\text{re-id, attempt}) = Pr(\text{attempt}) \times Pr(\text{re-id} \mid \text{attempt})$

2. **Inadvertent:**

$Pr(\text{re-id, acquaintance}) = Pr(\text{acquaintance}) \times Pr(\text{re-id} \mid \text{acquaintance})$

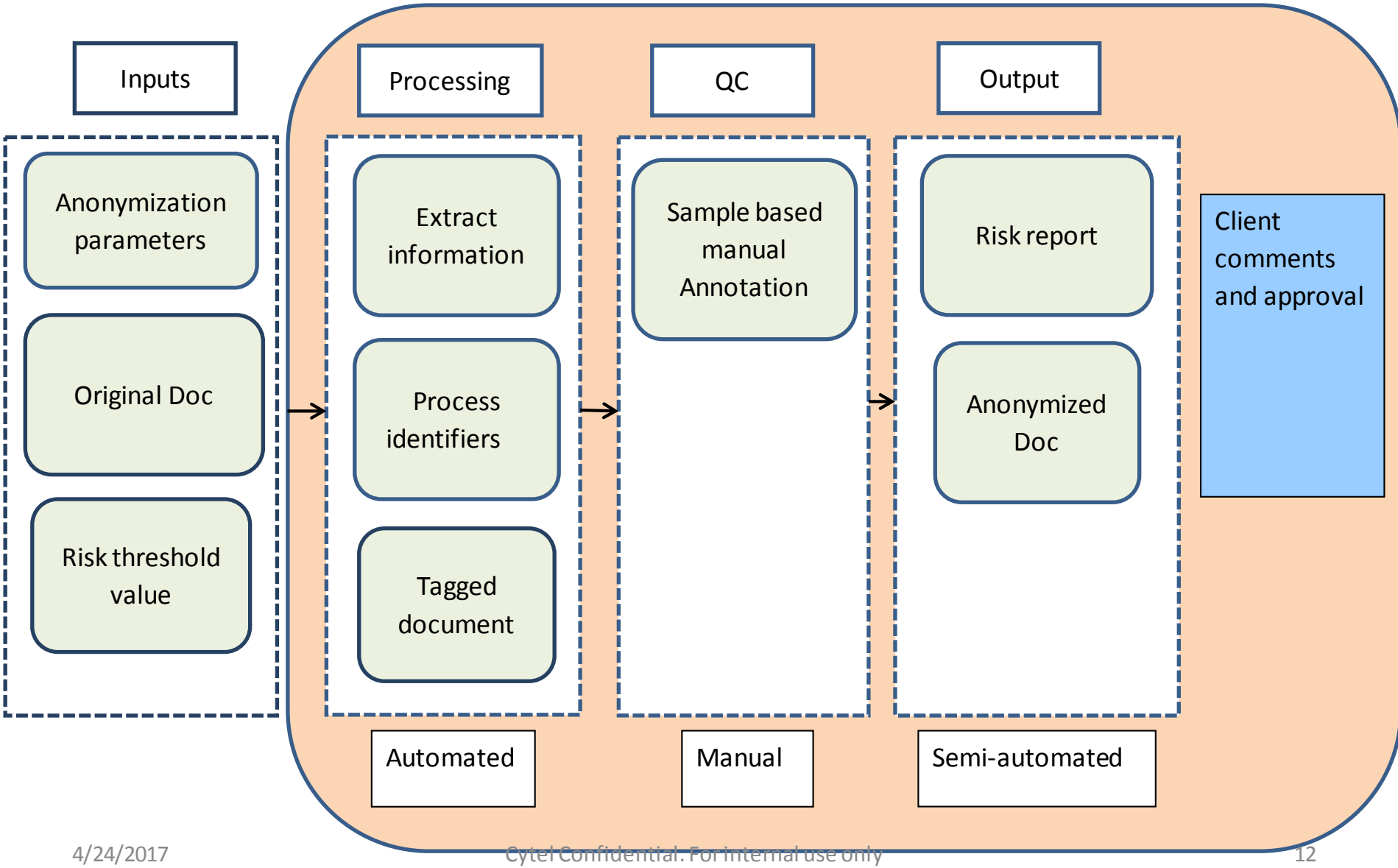
3. **Data breach:**

$Pr(\text{re-id, breach}) = Pr(\text{breach}) \times Pr(\text{re-id} \mid \text{breach})$

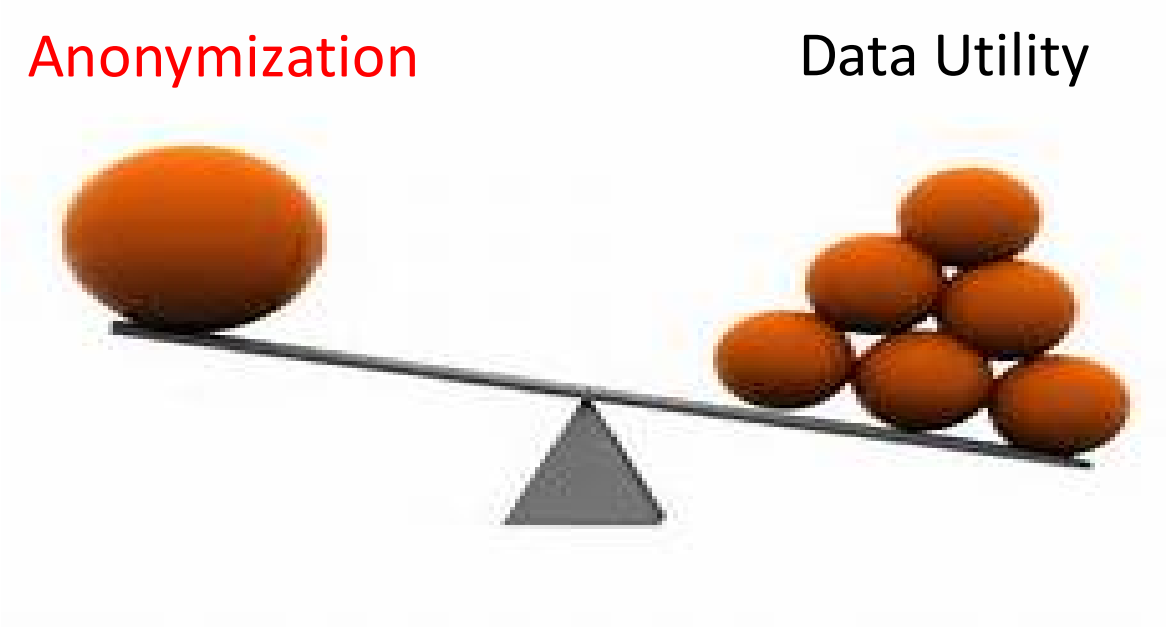
4. **Demonstration:**

Pr(re-id) - based on de-identified data

Our Approach



Risk Vs Data Utility



The goal should be to maximize data utility while meeting the predefined risk thresholds

Challenges

- Context – clarity?
- Availability of information in public domain
- Ensuring exhaustiveness of :
  - Anonymization parameters
  - References of patient data within and across other sources of information
- Instance of an identifier in Graph/image – needs OCR
- Risk in un-redacted part
- Single entries – possible candidates for re-identification

Risk Report Contents

- Risk Determination Statement
- Coverage
- Anonymization process and risk model
- Risk threshold
- Transformations applied
- Risk measurement
- Risk assessment
- Data utility
- References

Summary

- Four types of attacks
- Risk Threshold
- Masking for direct identifiers and de-identification for Quasi identifiers
- Four step process (Iterative)
    - Determine context
    - Determine risk threshold
    - Anonymize
    - Assess risk
- Risk Report – documenting the process of anonymization and risk calculation and assessment

Contact details:

Manjusha Gode

Email: manjusha.gode@cytel.com

manjushagode@yahoo.com

Phone: 98509 71711

References:

1. Guide to the De-identification of personal health information by Khaled El Emam

2. Anonymizing health data by Khaled El Emam

Thanks!