Digital Data Flows Masterclass Series

Class 3: De-Identification, Differential Privacy, and Homomorphic Encryption

30 January 2019

DIGITAL DATA FLOWS MASTERCLASS: EMERGING TECHNOLOGIES

Curriculum	Date*
Session 1: Artificial Intelligence and Machine Learning	25 October 2018 - side event, ICDPPC (Brussels)
Session 2: Location Data: GPS, Wi-Fi, and Spatial Analytics	27 November 2018 - (Brussels)
Session 3: De-Identification: Multi-party Computing, Differential Privacy, and Homomorphic Encryption	30 January 2019 - side event, CPDP (Brussels) (with remote participation)
Session 4: Advertising Technologies: Online Data Flows, Behavioral Targeting, and Cross-Device Tracking	March 2019 - Virtual
Session 5: Mobile Apps: Operating Systems, Software Development Kits (SDKs), and User Controls	April 2019 - Virtual
Session 6: Transportation and Mobility: Video Analytics, Sensors, and Connected Infrastructure	June 2019 - Virtual
Session 7: Biometric Data: Facial Recognition, Voice, and Digital Fingerprints	July 2019 - Virtual
Session 8: Tracking in Physical Spaces: Retail Technologies, Smart Homes, and the "Internet of Things"	Sept. 2019 - Virtual

*dates may change.

visit: https://fpf.org/classes/

Guest Experts:

Khaled El-Imam
Founder and CEO, Privacy Analytics

Prof. Sophie Stalla-Bourdillon
Professor in Information Technology Law and Data Governance, University of Southampton, UK
Senior Privacy Counsel & Legal Engineer, Immuta

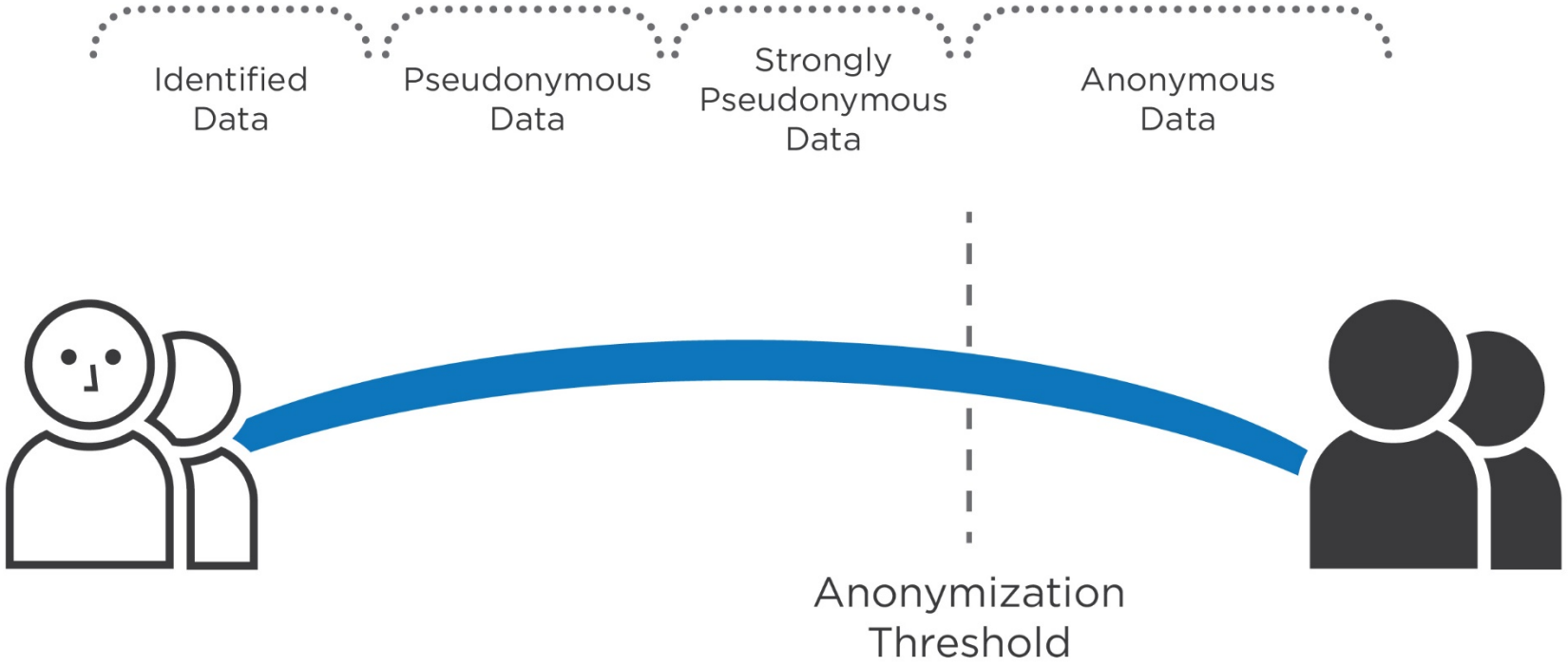
IMS Health & Quintiles are now IQVIA™

Principles of Data De-identification and Pseudonymization

Khaled El Emam

PRIVACY ANALYTICS
an IQVIA company

The Identifiability Spectrum



Types of Identifiers

Examples of direct identifiers: Name, address, telephone number, fax number, MRN, health card number, health plan beneficiary number, VID, license plate number, email address, photograph, biometrics, SSN, SIN, device number, clinical trial record number

Examples of quasi-identifiers: sex, date of birth or age, geographic locations (such as postal codes, census geography, information about proximity to known or unique landmarks), language spoken at home, ethnic origin, total years of schooling, marital status, criminal history, total income, visible minority status, profession, event dates, number of children, high level diagnoses and procedures

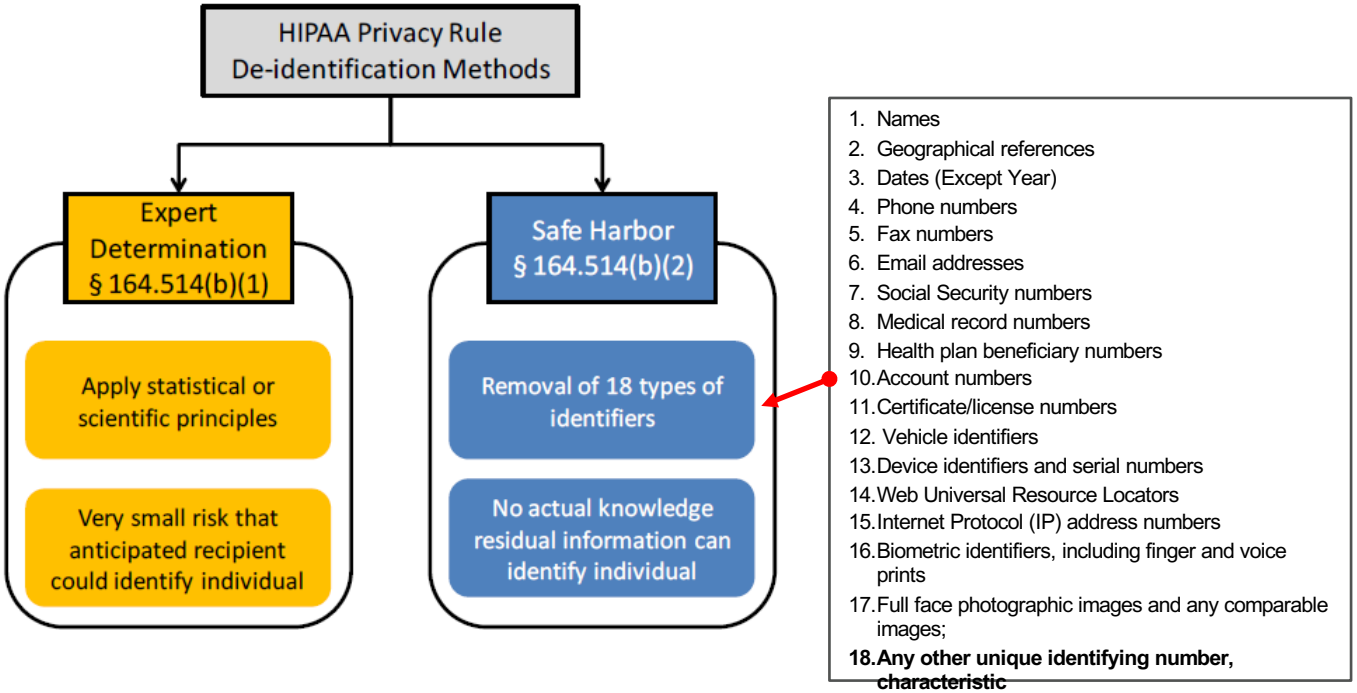
Pseudonymous Data

Examples of direct identifiers: Name, address, telephone number, fax number, MRN, health card number, health plan beneficiary number, VID, license plate number, email address, photograph, biometrics, SSN, SIN, device number, clinical trial record number

X

Examples of quasi-identifiers: sex, date of birth or age, geographic locations (such as postal codes, census geography, information about proximity to known or unique landmarks), language spoken at home, ethnic origin, total years of schooling, marital status, criminal history, total income, visible minority status, profession, event dates, number of children, high level diagnoses and procedures

HIPAA De-identification Standards



A29 / CNIL Anonymization Approaches

Anonymization

- **No clear line between Anonymized and Personnal data**
- **The opinion provides two options to check that a Dataset is anonymized:**

1. **Your dataset has <u>none</u> of the following property:**

   – **Singling out:** possibility to isolate some records of an individual in the dataset;

   – **Linkability:** ability to link, at least, two records concerning the same data subject or a group of data subjects (in the same database or in two different databases);

   – **Inference:** the possibility to deduce, with significant probability, the value of an attribute from the values of a set of other attributes

OR

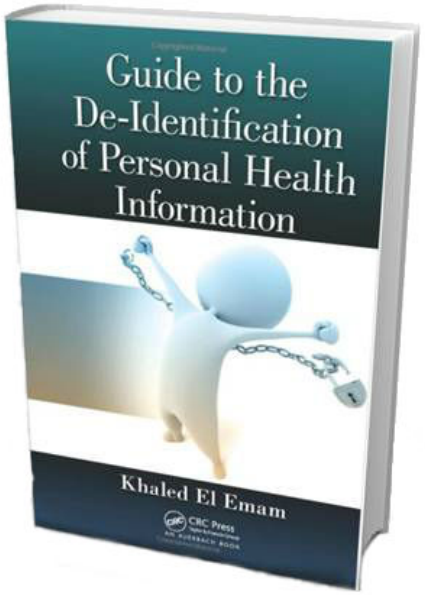
2. **Make analysis of re-identification risk.**

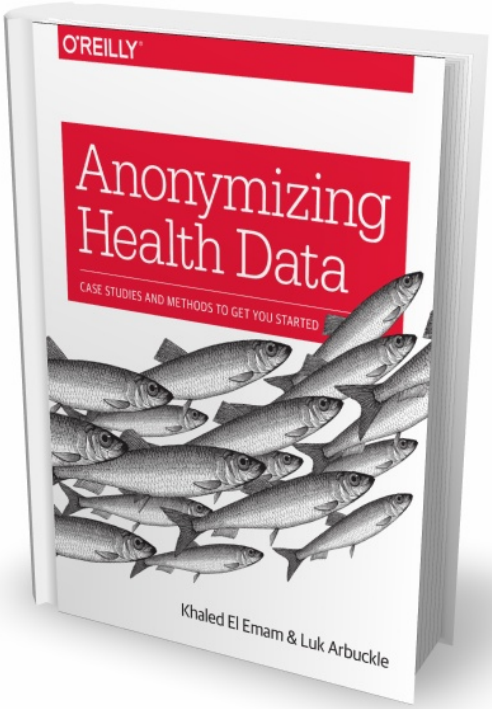
CNIL

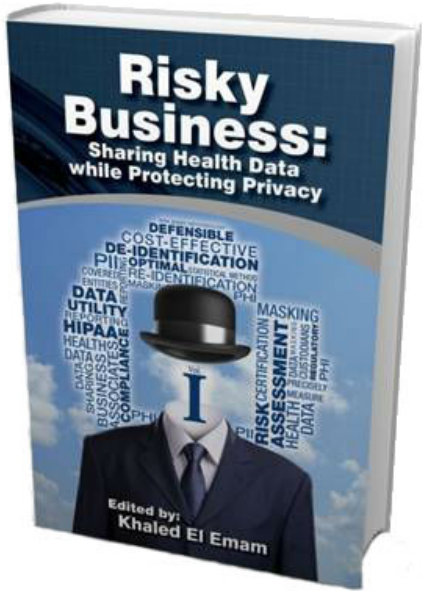
FUTURE OF PRIVACY FORUM

Risk-based Anonymization Methods

Methodology Texts

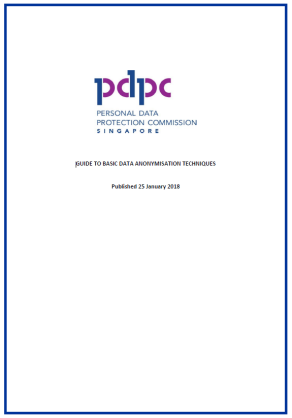


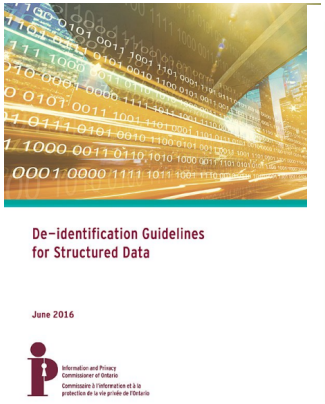


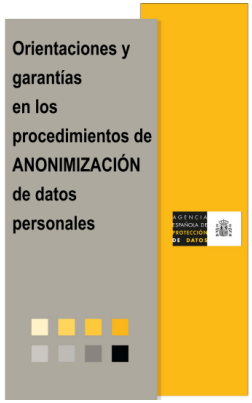


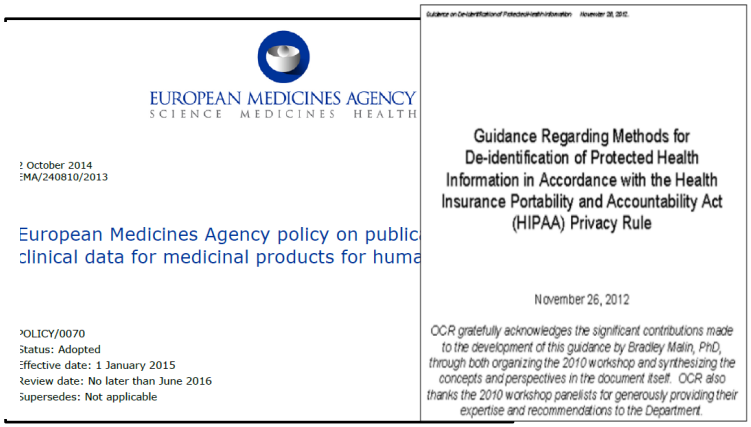
Guidelines


Data protection
Anonymisation: managing data protection risk code of practice
ico.

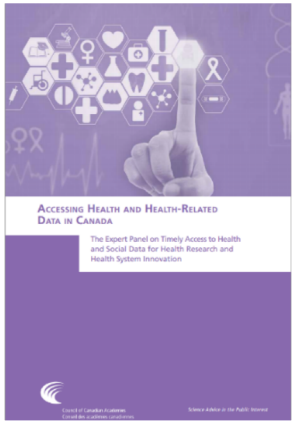

pdpc
PERSONAL DATA PROTECTION COMMISSION SINGAPORE
GUIDE TO BASIC DATA ANONYMISATION TECHNIQUES
Published 25 January 2018


De-identification Guidelines for Structured Data
June 2016
Information and Privacy Commissioner of Ontario
Commissaire à l'information et à la protection de la vie privée de l'Ontario

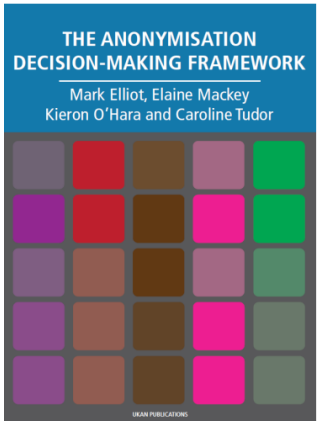

Orientaciones y garantías en los procedimientos de ANONIMIZACIÓN de datos personales

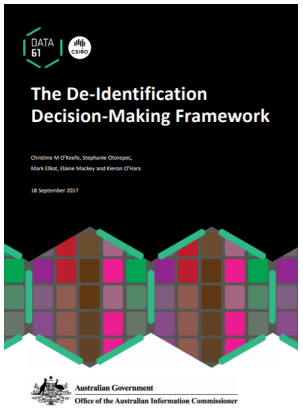

EUROPEAN MEDICINES AGENCY
SCIENCE MEDICINES HEALTH
October 2014
EMA/240810/2013
European Medicines Agency policy on public
clinical data for medicinal products for huma
POLICY/0070
Status: Adopted
Effective date: 1 January 2015
Review date: No later than June 2016
Supersedes: Not applicable
Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule
November 26, 2012
OCR gratefully acknowledges the significant contributions made to the development of this guidance by Bradley Malin, PhD, through both organizing the 2010 workshop and synthesizing the concepts and perspectives in the document itself. OCR also thanks the 2010 workshop panelists for generously providing their expertise and recommendations to the Department.


Sharing Clinical Trial Data
MAXIMIZING BENEFITS, MINIMIZING RISK
INSTITUTE OF MEDICINE


Accessing Health and Health-Related Data in Canada
The Expert Panel on Timely Access to Health and Social Data for Health Research and Health System Innovation


HITRUST
De-Identification Framework
A Consistent, Managed Methodology for the De-Identification of Personal Data and the Sharing of Compliance and Risk Information

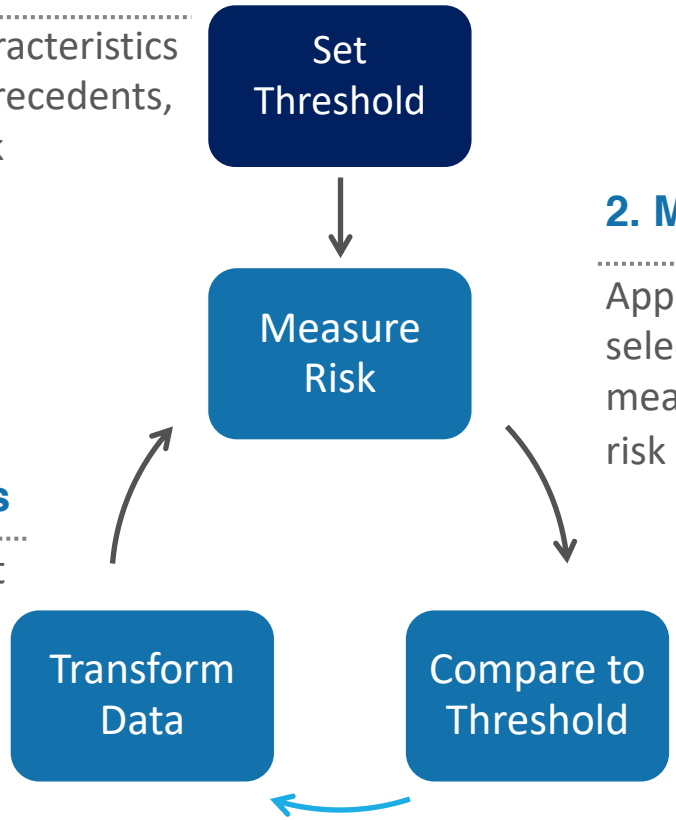

THE ANONYMISATION DECISION-MAKING FRAMEWORK
Mark Elliot, Elaine Mackey
Kieron O'Hara and Caroline Tudor
UKAN PUBLICATIONS


The De-Identification Decision-Making Framework
18 September 2017
Australian Government
Office of the Australian Information Commissioner

Anonymization Cycle

1. Set Risk Threshold

Based on the characteristics of the data and precedents, a quantitative risk threshold is set.



2. Measure Risk

Appropriate metrics are selected and used to measure re-identification risk from the data.

3. Evaluate Risk

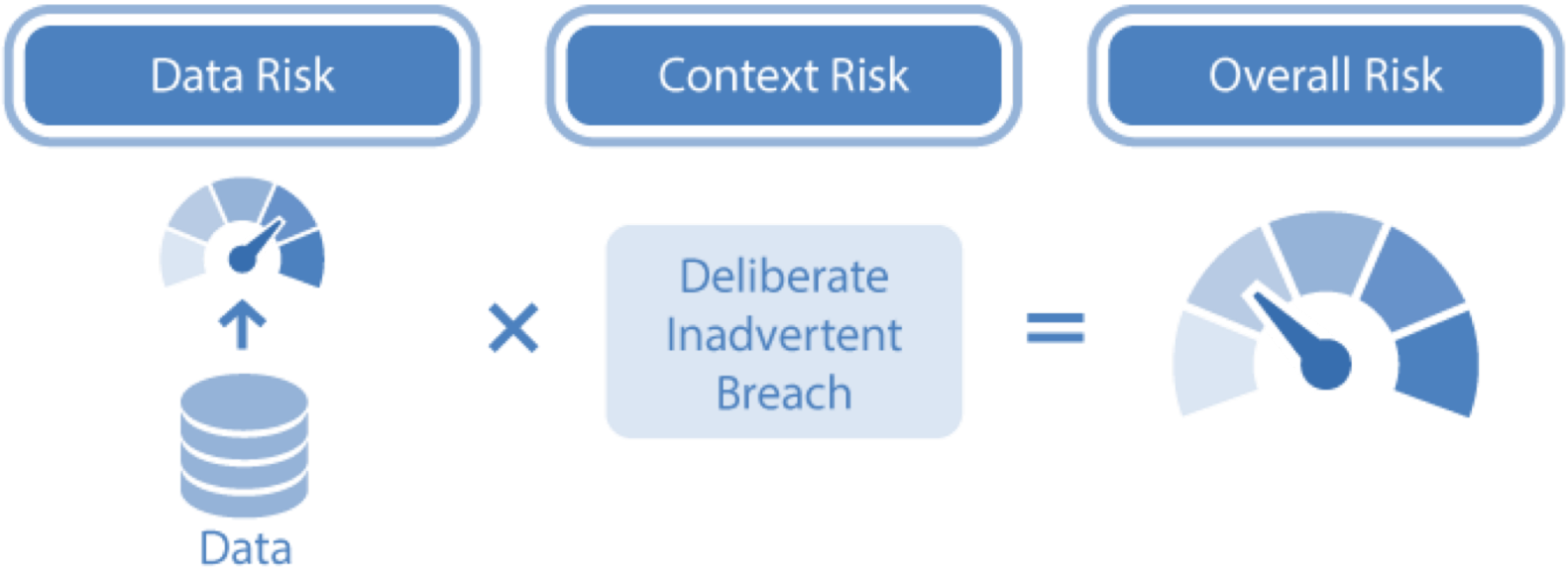
Compare the measured risk against the threshold to determine if it is above or below it.

4. Apply Transformations

If the measured risk does not meet the threshold, specific transformations are applied to reduce the risk.

VUB BRUSSELS PRIVACY HUB FUTURE OF PRIVACY FORUM

Measuring Overall Risk



Measuring Data Risk

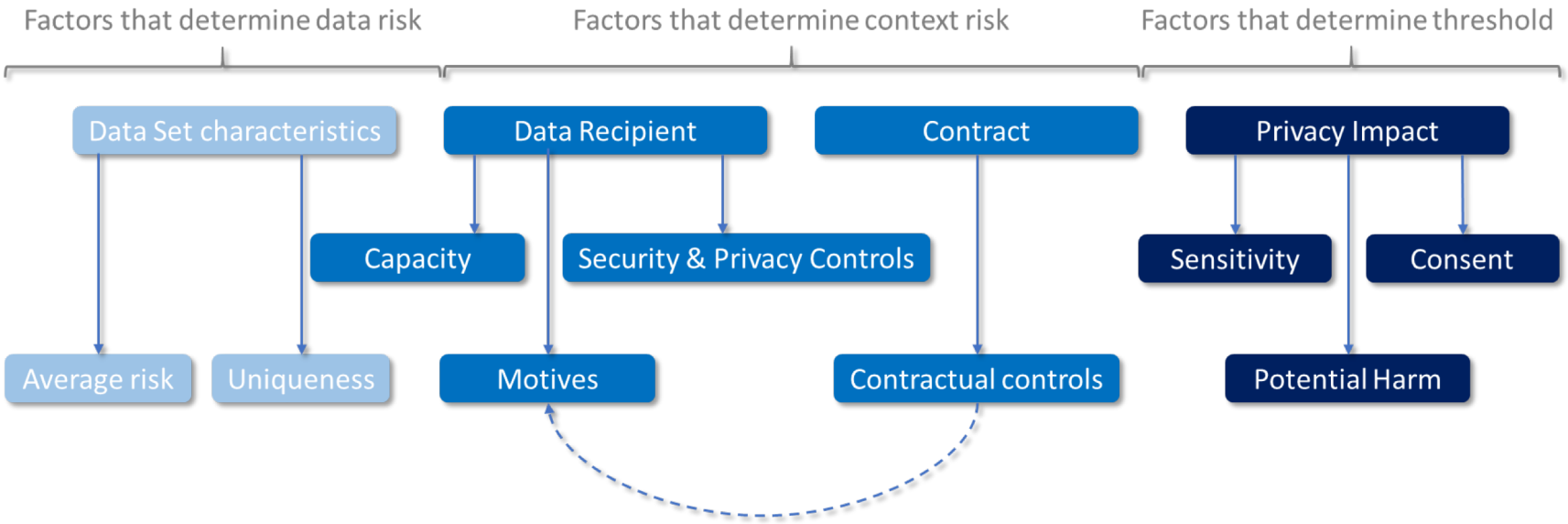
DIRECT IDENTIFIERS			QUASI-IDENTIFIERS		OTHER VARIABLES		
ID	Name	Telephone No.	Sex	Year of Birth	Lab Test	Lab Result	Pay Delay
1	John Smith	(412) 668-5468	M	1959	Albumin, Serum	4.8	37
2	Alan Smith	(413) 822-5074	M	1969	Creatine Kinase	86	36
3	Alice Brown	(416) 886-5314	F	1955	Alkaline Phospha		52
4	Hercules Green	(613)763-5254	M	1959	Bilirubin		36
5	Alicia Freds	(613) 586-6222	F	1942	BUN/Creatinine		82
6	Gill Stringer	(954) 699-5423	F	1975	Calcium, Seru		34
7	Marie Kirkpatrick	(416) 786-6212	F	1966	Free Thyroxine I		23
8	Leslie Hall	(905) 668-6581	F	1987	Globulin, Total	3.5	9
9	Douglas Henry	(416) 423-5965	M	1959	B-type Natriuretic peptide	134	38
10	Fred Thompson	(416) 421-7719	M	1967	Creatine Kinase	80	21

3

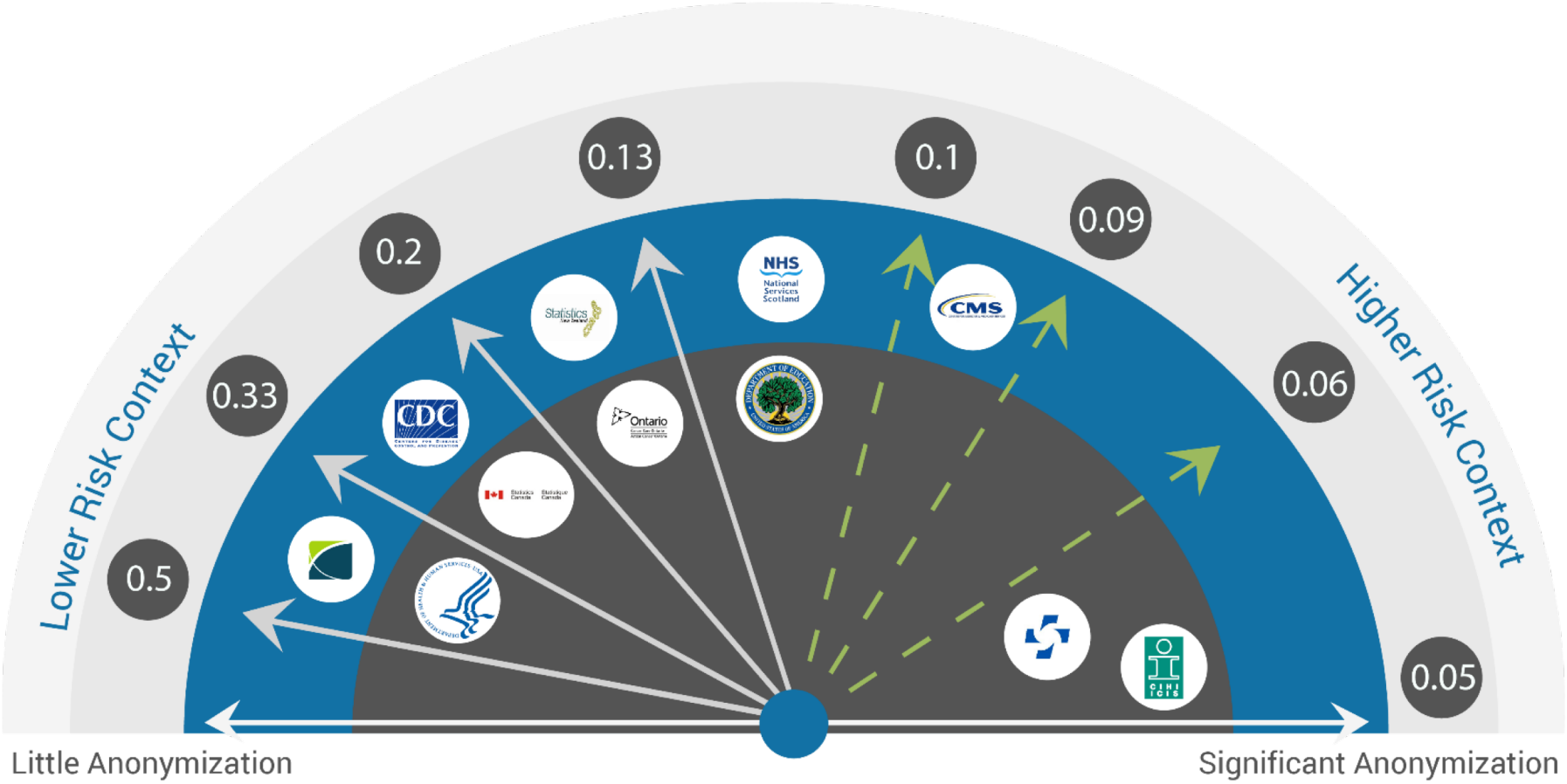
Two quasi-identifiers matching in three cells within a data set

Factors Affecting Risk

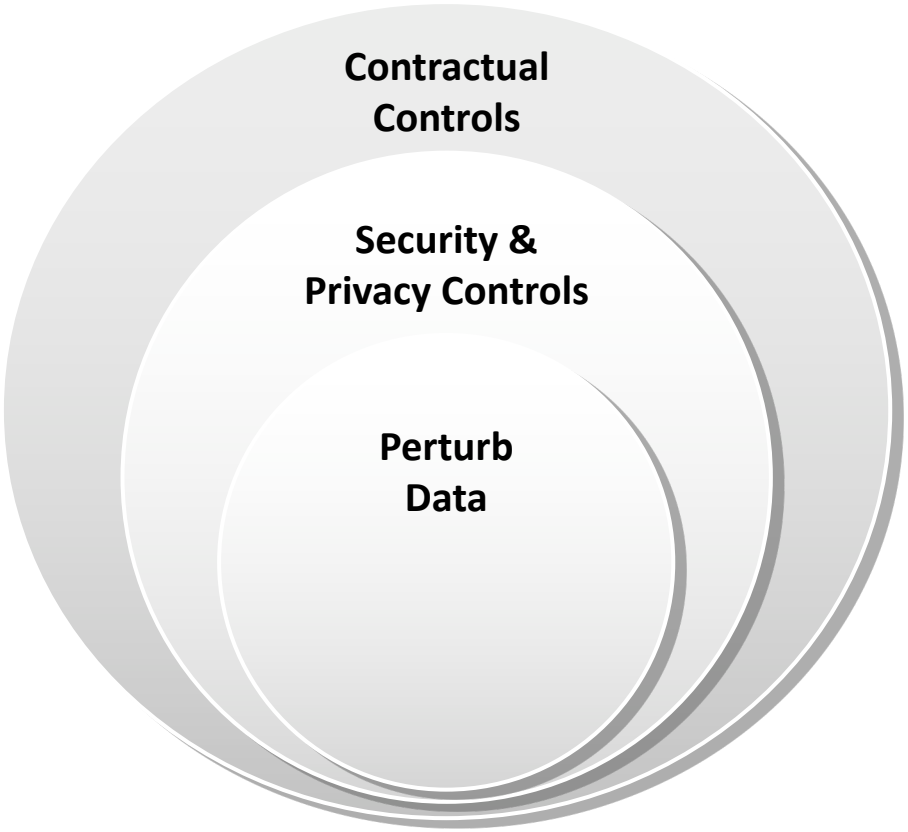
Multiple factors below are taken into account to properly de-identify (anonymize) data. The data risk and context risk are measured and then compared against the risk threshold.



Spectrum of Identifiability



Layers of Protection



IMS Health & Quintiles are now IQVIA™

Contact

kelemam@privacy-analytics.com

@kelemam

www.privacy-analytics.com

PRIVACY ANALYTICS
an IQVIA company

Introduction to differential privacy

Prof. Sophie Stalla-Bourdillon
Professor in Information Technology Law and Data Governance, University of Southampton, UK
Senior Privacy Counsel & Legal Engineer, Immuta

VUB BRUSSELS PRIVACY HUB FUTURE OF PRIVACY FORUM

Overview

3 things for today

1. What is differential privacy?
2. What is the promise of differential privacy?
3. What are the use cases for differential privacy?

3 things for today

1. What is differential privacy?
2. What is the promise of differential privacy?
3. What are the use cases for differential privacy?

Souvenir, souvenir…

Digressing

Have you heard about masking, pseudonymizing, and generalization techniques (e.g. k-anonymity)?

Differential privacy is ….

Differential privacy is …. different in that it is based on randomization

Let’s assume I want to de-identify

i.e. make sure it is not possible to attribute the data to individuals

I could mask obvious direct identifiers (unique identifiers) or quasi direct identifiers (names, addresses) or sensitive data (sexual orientation)…

But then I could have a long list of attributes which when combined together could make the individual easily identifiable (gender, profession, age, location)

I could round attributes

Ex:

- Classify profession into higher classes
- Use age range

But then I am not necessarily preventing linking (with external databases)

What is differential privacy?

Why does differential privacy stand out?

Differential privacy is “privacy by process”

Cynthia Dwork

It’s query based… rather than data based

Injection of randomized noise

Play time

- Pick a number between 1 and 4 (and keep it secret)
- Question: have you ever cheated on your tax return?
- Raise your hand if either

→ you have actually already cheated on your tax return or

→ Picked the number 3

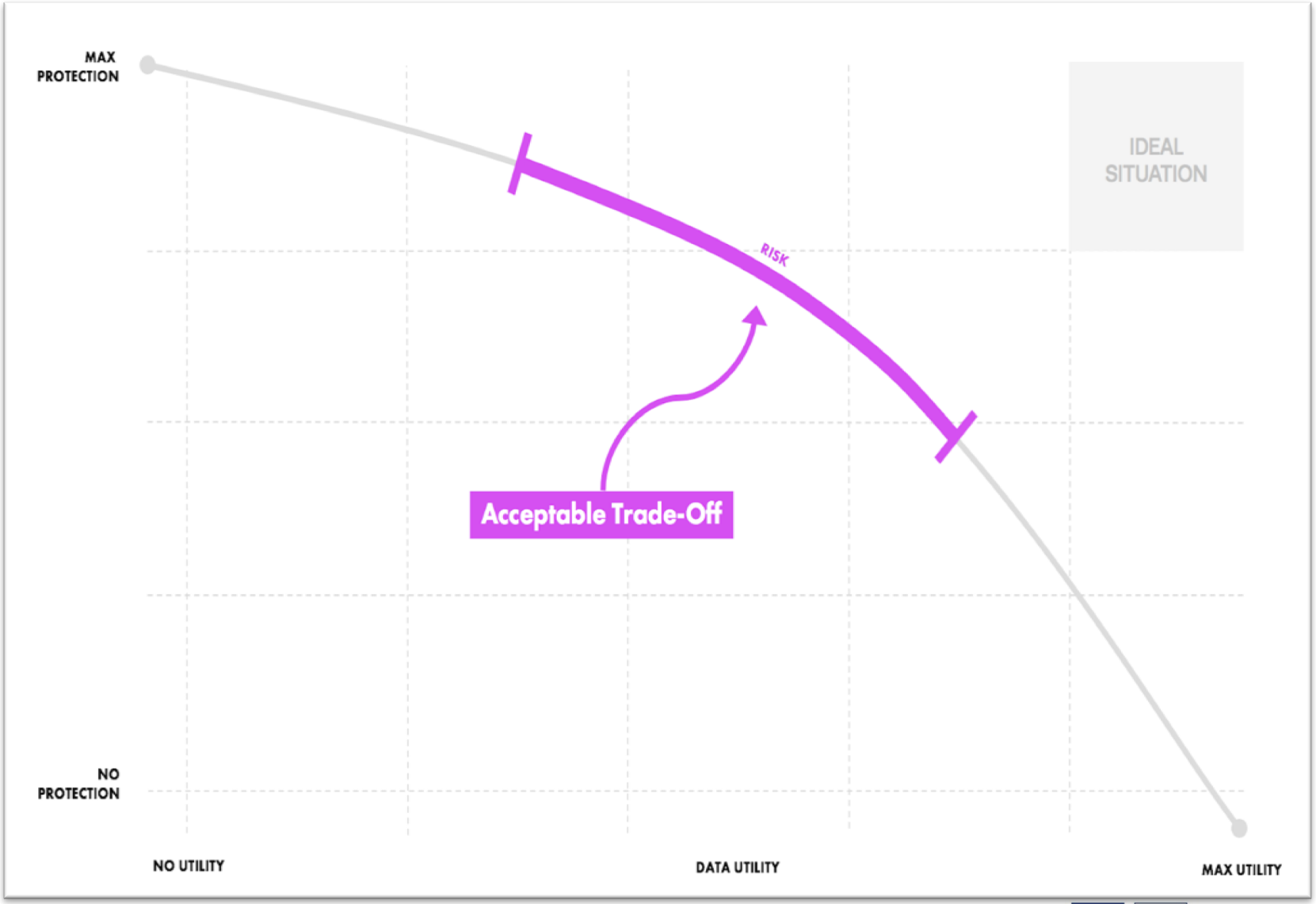
When is the injection done?

Can be done at the time the query is made

Wow!

Wow!

You can tailor the protection to the query!


MAX
PROTECTION
NO
PROTECTION
IDEAL
SITUATION
RISK
Acceptable Trade-Off
NO UTILITY
DATA UTILITY
MAX UTILITY

GREAT!

In particular if you are interested in maximizing both utility and data protection

Randomization is essential in a world where new data sets are created daily, ie. to mitigate linking

Differential privacy at core

Definition

A technique that ensures a data analyst always receives the equivalent query result from a data set, ***regardless of whether an individual's data is included in that data or whether it is excluded from it.***

This is one mathematical description for the protection of privacy. You should never be able to learn anything specific about an individual from a data set if you can't even tell her data is contained in the data.

How does it work?

The technique is flexible and has been implemented in a variety of ways, but at core it is comprised of **3 functions**:

- Query limitations
- Statistical noise
- “Privacy loss”

1. Query limitations

Only ***aggregate*** queries can be made against a data set. Ex: min, max, average, count and sum.

2. Noise

A **calibrated amount of noise** is injected into query results to protect specific data points while allowing meaningful information to be drawn from aggregate trends/patterns.

3. Privacy Loss

Theoretically, given infinite questions an analyst could still learn something about individuals even if restricted by the above techniques. So differential privacy mandates a “budget” that limits the number and specificity of questions that can be asked. If you exhaust that budget (or **your privacy loss is above a certain threshold**) it will be possible to derive inferences. There are multiple ways to implement this, but this basically disallows infinite questions.

3 things for today

1. What is differential privacy?
2. What is the promise of differential privacy?
3. What are the use cases for differential privacy?

The promise

“Differential privacy promises that the probability of harm is not significantly increased by the choice [of the individual] to participate [to the differentially private data set].”

Cynthia Dwork

The individual can thus deny his participation into the dataset!

As a result

Injection of randomizws noise brings immunization

Immunization to what?

“Differential privacy is immune to post-processing. “

Cynthia Dwork

With DP you can mitigate 3 types of risks

- linkability: possibility to link records of the same individual together
- singling out: possibility to isolate some records
- inference : possibility to derive, with significant probability, the value of an attribute from the values of other attributes

3 things for today

1. What is differential privacy?
2. What is the promise of differential privacy?
3. What are the use cases for differential privacy?

Let’s pick 4 use cases

To analyse data in order to improve a manufacturing process (use case 1)

To analyse data in order to improve products/services (use case 2)

To create customer profiles to ensure maintenance of products/services (use case 3)

To derive insights about individuals in order to offer new goods or services, target advertising (use case 4)

In which use cases can I use differential privacy?

Answer time

All of them as long as

1. I am not applying the profile upon individuals yet
2. I am creating profiles based on aggregates

“Tell me the exact IP address for a user who purchased item X”

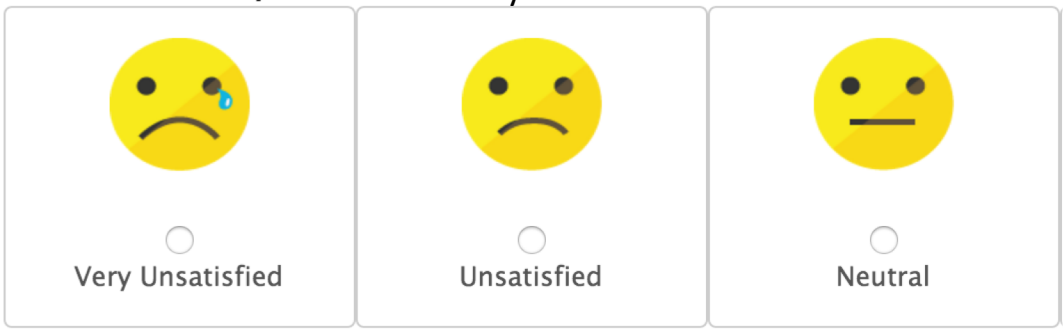

“Tell me the exact IP address
for a user who purchased item
X”

Satisfaction survey

How satisfied are you with DP today?

Very Unsatisfied	Unsatisfied	Neutral	Satisfied	Very Satisfied
				
○	○	○	○	○

How satisfied are you with DP today?



Wait!

What if we talked about ‘attribute-based’ or local differential privacy?

This works if the individual does not mind his participation to the dataset be known.

Noise is used to alter the values of each record. The idea is that any potentially **embarrassing/sensitive information** which appears for a user could appear there by chance.

Upshot: The data subject gains privacy through the ability to deny the contents of their record. The output is a noised record.

Play time bis

- Pick a number between 1 and 4 (and keep it secret)
- Question: have you ever cheated on your tax return?
- Raise your hand if either

→ you have actually already cheated on your tax return or

→ Picked the number 3

This was attribute-based differential privacy in reality…

When does the difference between data set-based DP and attribute-based DP matter?

Let’s assume I am in charge of creating a data set of people investigated for tax fraud and opening the data set to queries.

Interesting feature of attribute-based DP

The data subject could be choosing how much randomness to apply to the submission, and the value could be different for different data subjects.

If the data subject is planning multiple submissions they might choose to increase the amount of randomization per submission, so that after they have made all submissions the net privacy loss does not exceed a users tolerance.

Conclusions

Why don’t we think about differential privacy more often?

BRUSSELS PRIVACY HUB

FUTURE OF PRIVACY FORUM

DP is a useful and promising PET!

DP should be used more often!

DP’s Beauty!

It can be combined with other PETS!

Think about masking for example (direct identifiers)

It can be tailored to the query and done on the fly!

A flying PET!

Questions?

Prof. Sophie Stalla-Bourdillon
Professor in Information Technology Law and Data Governance, University of Southampton, UK
Senior Privacy Counsel & Legal Engineer, Immuta

VUB BRUSSELS PRIVACY HUB FUTURE OF PRIVACY FORUM

Overview of Practical Secure Computation Techniques

Homomorphic encryption and secret sharing

Khaled El Emam

Homomorphic Encryption Schemes

Mutli-party computation

Trust models

Trusted 3rd party

Passive adversary (honest-but-curious)

Fully malicious

secure Multi-party

computation

Cons:

speed, complex, bespoke

Pros:

Security guarantees

APPLICATIONS

Medical research
Trustworthy electronic voting
Law enforcement/national security
Financial transactions
Advertising/marketing
Social media
Mobile communications
cloud computing
Genomic privacy

Public key

Encryption

Public key

Encryption

$$c = \mathsf{Enc}_{pk}(m)$$

$$m = \mathsf{Dec}_{sk}(c)$$

Randomized Public key

Encryption

$$c_1 = \mathsf{Enc}_{pk}(m, r_1)$$

$$c_2 = \mathsf{Enc}_{pk}(m, r_2)$$

Randomized Public key

Encryption

$$c_1 = \mathsf{Enc}_{pk}(m, r_1)$$
$$c_2 = \mathsf{Enc}_{pk}(m, r_2)$$

$$If\ r_1 \neq r_2 \text{ then } c_1 \neq c_2, but$$
$$Dec_{sk}(c_1) = Dec_{sk}(c_2) = m$$

Randomized Public key

Encryption

Notation denoting an encrypted plaintext

$$[a] = \mathsf{Enc}_{pk}(a, r)$$

$$[b] = \mathsf{Enc}_{pk}(b, r')$$

Additively Homomorphic Encryption

Additively Homomorphic

Encryption

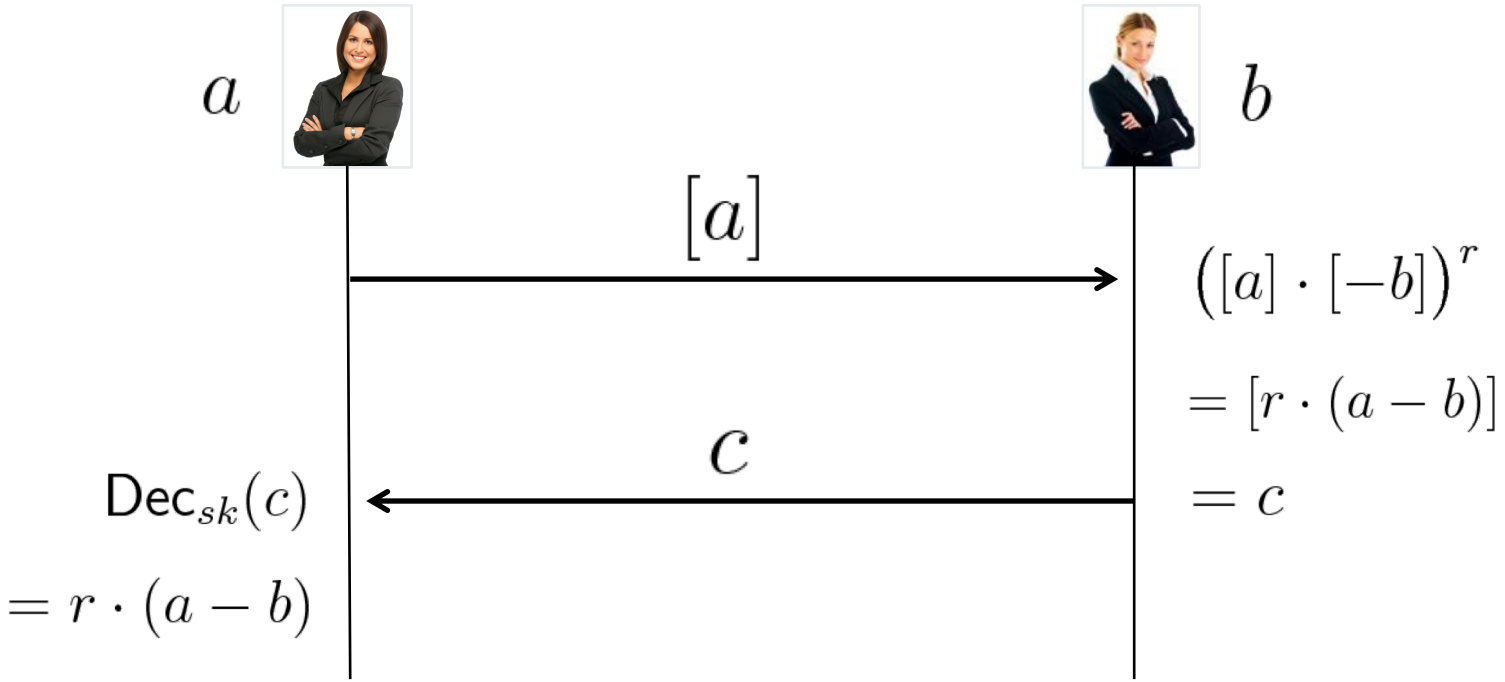
$$[a] \cdot [b] = [a + b]$$

Additively Homomorphic

Encryption

$$[a]^b = [ab]$$

2-party SECURE comparison Protocol



Additively

Homomorphic schemes

Paillier (1999)

$$[a] \cdot [b] = [a + b]$$

Exponential elgamal (1984,1997)

$$[g^a] \cdot [g^b] = [g^{a+b}]$$

fully

Homomorphic schemes

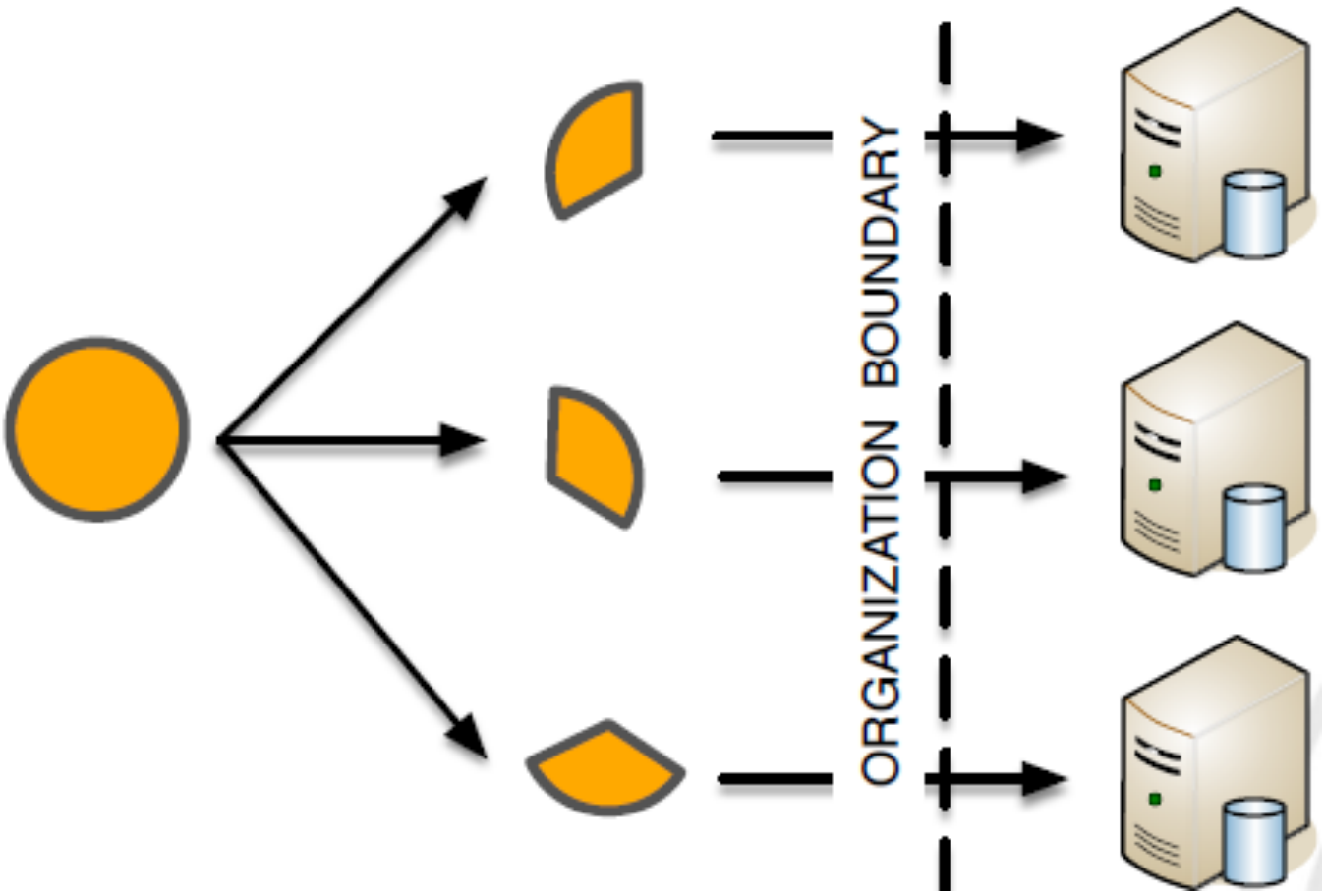
gentry (2009)

$$[a] \otimes [b] = [ab]$$

$$[a] \oplus [b] = [a + b]$$

Secret Sharing

Secret Sharing Protocols



User A
Age: $a = 25$

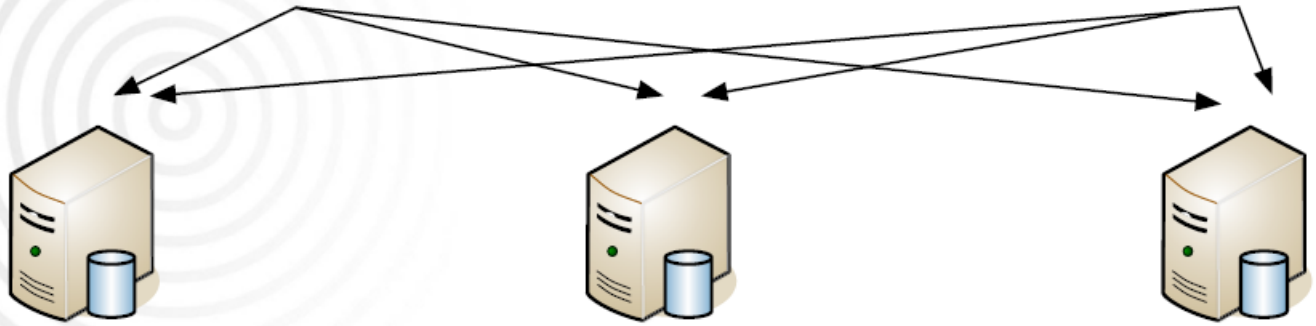
User B
Age: $b = 33$

User A
Age: $a = 25$

1. Choose randomly $a_1 = 57$
2. Choose randomly $a_2 = 13$
3. Compute $a_3 = 25 - 57 - 13 = -45 \equiv 55 \bmod 100$
4. Distribute a_k between the three servers

User B
Age: $b = 33$

1. Choose randomly $b_1 = 44$
2. Choose randomly $b_2 = 57$
3. Compute $b_3 = 33 - 44 - 57 = -68 \equiv 32 \bmod 100$
4. Distribute b_k between the three servers



User A
Age: $a = 25$

1. Choose randomly $a_1 = 57$
2. Choose randomly $a_2 = 13$
3. Compute $a_3 = 25 - 57 - 13 = -45 \equiv 55 \bmod 100$
4. Distribute a_k between the three servers

User B
Age: $b = 33$

1. Choose randomly $b_1 = 44$
2. Choose randomly $b_2 = 57$
3. Compute $b_3 = 33 - 44 - 57 = -68 \equiv 32 \bmod 100$
4. Distribute b_k between the three servers

$a_1 = 57$
$b_1 = 44$
$c_1 = a_1 + b_1 = 101 \equiv 1 \bmod 100$

Server 1

$a_2 = 13$
$b_2 = 57$
$c_2 = a_2 + b_2 = 70 \equiv 70 \bmod 100$

Server 2

$a_3 = 55$
$b_3 = 32$
$c_3 = a_3 + b_3 = 87 \equiv 87 \bmod 100$

Server 3

User C

User A
Age: $a = 25$

1. Choose randomly $a_1 = 57$
2. Choose randomly $a_2 = 13$
3. Compute $a_3 = 25 - 57 - 13 = -45 \equiv 55 \bmod 100$
4. Distribute a_k between the three servers

User B
Age: $b = 33$

1. Choose randomly $b_1 = 44$
2. Choose randomly $b_2 = 57$
3. Compute $b_3 = 33 - 44 - 57 = -68 \equiv 32 \bmod 100$
4. Distribute b_k between the three servers

Server 1

$a_1 = 57$
$b_1 = 44$
$c_1 = a_1 + b_1 = 101 \equiv 1 \bmod 100$

Server 2

$a_2 = 13$
$b_2 = 57$
$c_2 = a_2 + b_2 = 70 \equiv 70 \bmod 100$

Server 3

$a_3 = 55$
$b_3 = 32$
$c_3 = a_3 + b_3 = 87 \equiv 87 \bmod 100$

User C

C computes $c = 1 + 70 + 87 = 158 \equiv 58 \bmod 100$
C learns that the sum of ages of A and B is 58.
Neither C nor the servers learn the individual ages of A and B.

In Practice

Practical Considerations

General purpose analysis tools still need to work on anonymized data or strongly pseudonymized data in a secure environment – otherwise the analysis results can leak information about the individual level data

Otherwise, very specific protocols have to be implemented and their security properties analyzed.

You want secure computation methods that are easily understood (i.e., you are not limited to the two people in the world who can understand and extend them).

Performance on large data is an important criterion to consider – it is not always obvious that all techniques and protocols will scale for the types of computations that you need to run.

We are seeing commercial tools on the market now that are focused on solving very specific problems – good progress there.

Acknowledgements

Some of the slides came from a presentation I gave a few years ago with Professor Aleksander Essex.

Some of the slides came from a presentation of the Sharemind platform describing secret sharing protocols.

Contact

kelemam@privacy-analytics.com

@kelemam

www.privacy-analytics.com

Questions?