

IMS Health & Quintiles are now

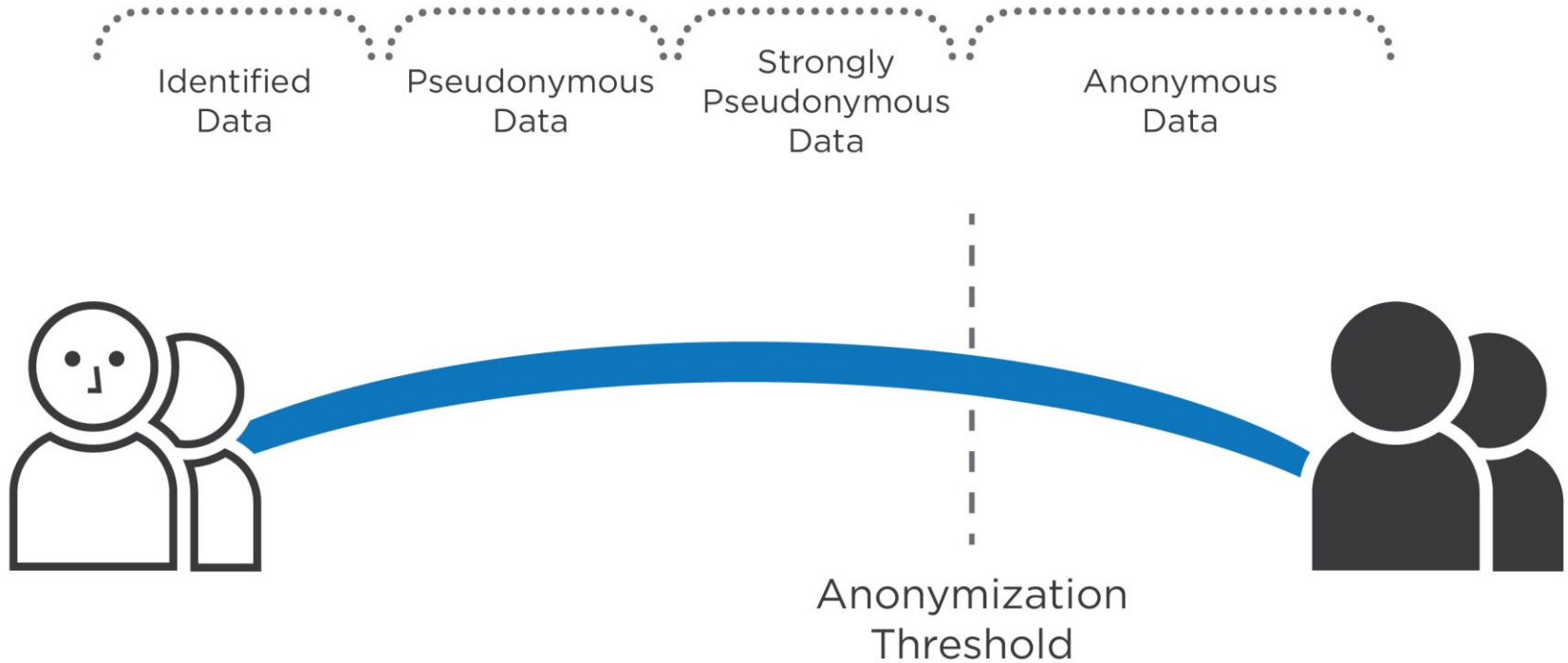


# Principles of De-identification

Khaled El Emam



# 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

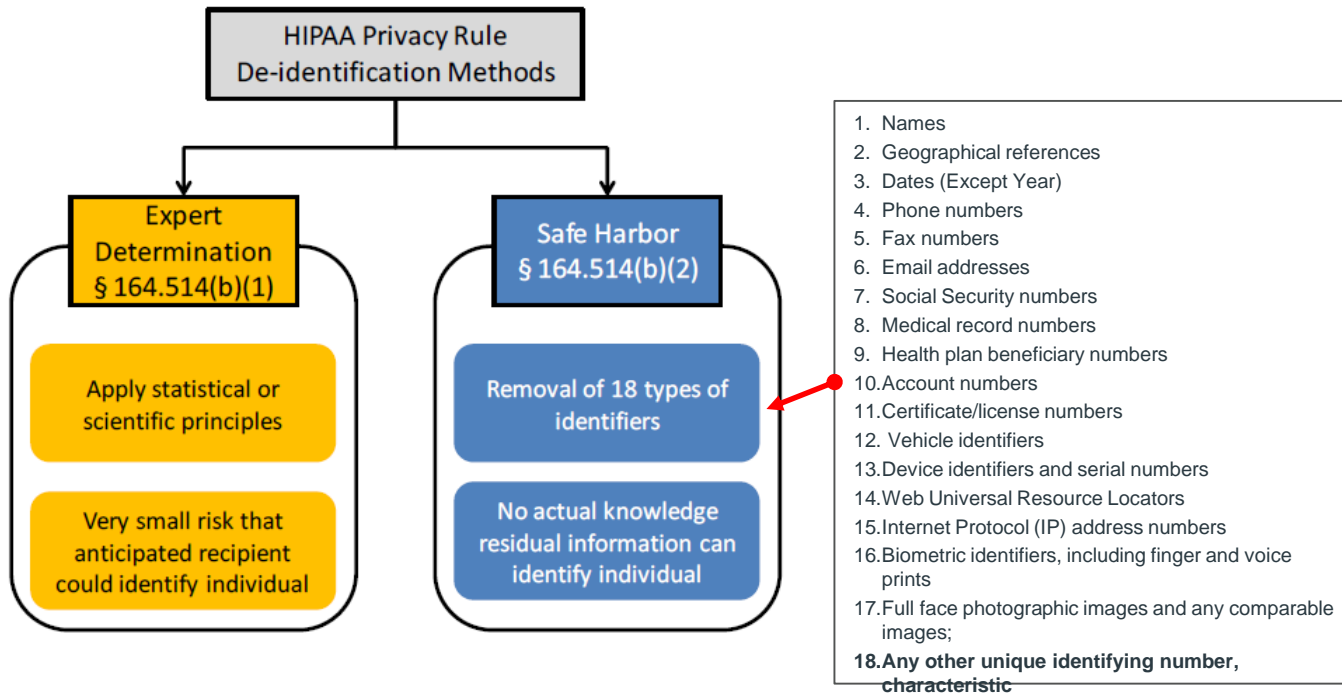
**Examples of quasi-identifiers:** sex, date of birth or age, geographic locations (such as postal codes, census geography, information about proximity to known or unique landmarks), language spoken at home, ethnic origin, total years of schooling, marital status, criminal history, total income, visible 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

**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 Personal data
- The opinion provides two options to check that a Dataset is anonymized:
  1. Your dataset has none 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.




# Guidelines



Data protection


## Anonymisation: managing data protection risk code of practice



**pdpc**  
PERSONAL DATA PROTECTION COMMISSION SINGAPORE


GUIDE TO BASIC DATA ANONYMISATION TECHNIQUES

Published 23 January 2018




### De-identification Guidelines for Structured Data

June 2016



### Orientaciones y garantías en los procedimientos de ANONIMIZACIÓN de datos personales



**EUROPEAN MEDICINES AGENCY**  
SCIENCE MEDICINES HEALTH

1 October 2014  
EMA/240810/2013

### European Medicines Agency policy on public clinical data for medicinal products for human use

POLICY/0070  
Status: Adopted  
Effective date: 1 January 2015  
Review date: No later than June 2016  
Supersedes: Not applicable

Guidance on De-identification of Protected Health Information November 26, 2012

### 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 OF THE NATIONAL ACADEMIES




### 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

Office of the Information Commissioner

**HITRUST**




### De-identification Framework

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

March 2015

### THE ANONYMISATION DECISION-MAKING FRAMEWORK

Mark Elliot, Elaine Mackey  
Kieron O'Hara and Caroline Tudor




HEALTH RESEARCH INSTITUTE

**DATA 61**

### The De-identification Decision-Making Framework

Christine M O'Hara, Stephanie O'Hara, Mark Elliot, Elaine Mackey and Kieron O'Hara

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.

Set  
Threshold



Measure  
Risk

## 2. Measure Risk

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

## 4. Apply Transformations

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

Transform  
Data

Compare to  
Threshold



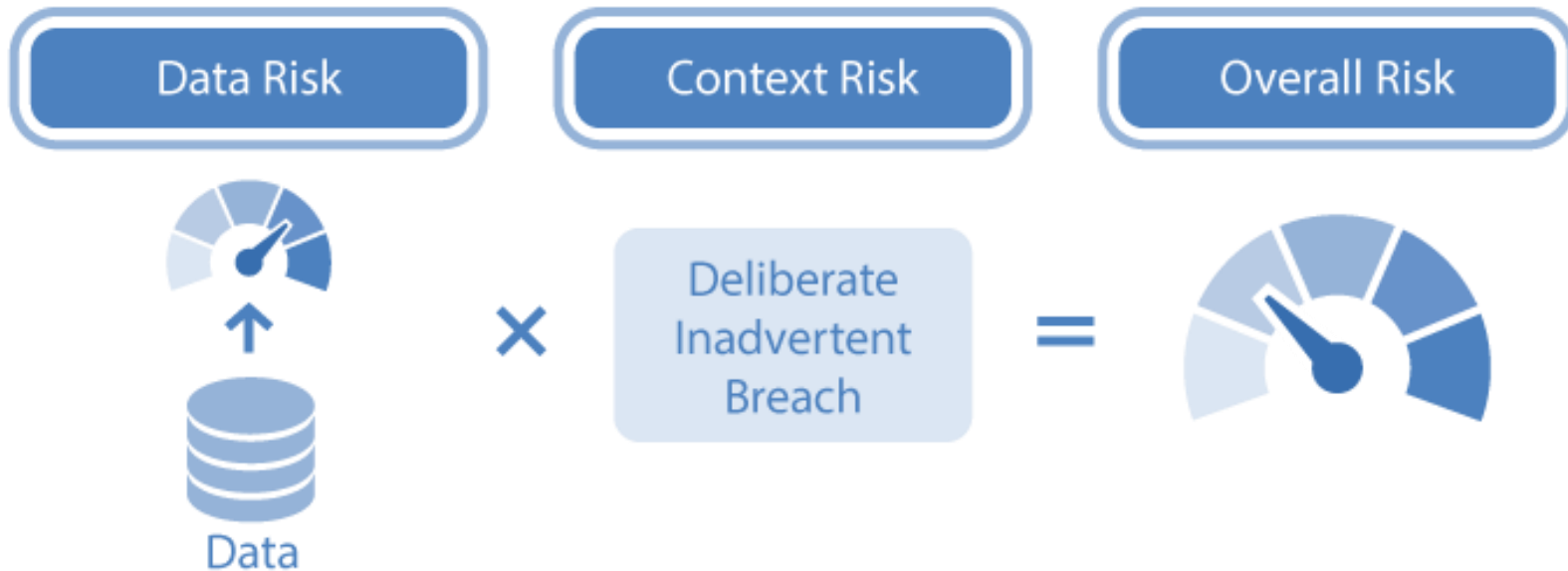
## 3. Evaluate Risk

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





# Measuring Overall Risk



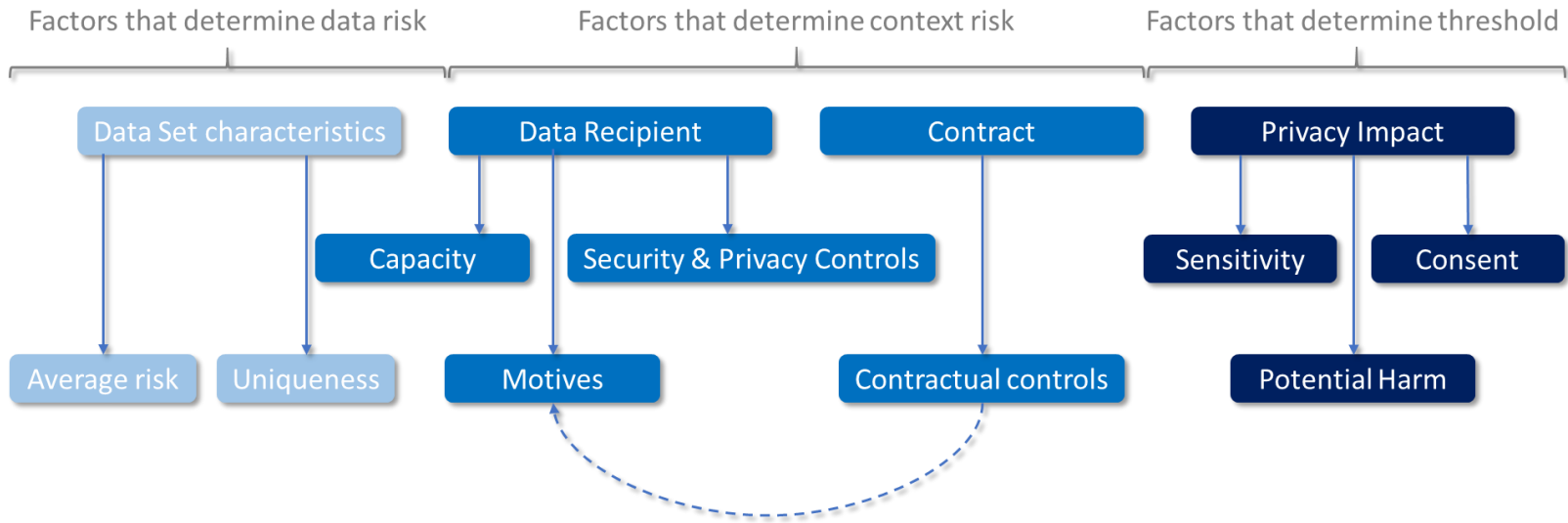
# Measuring Data Risk

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 Phosph		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		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

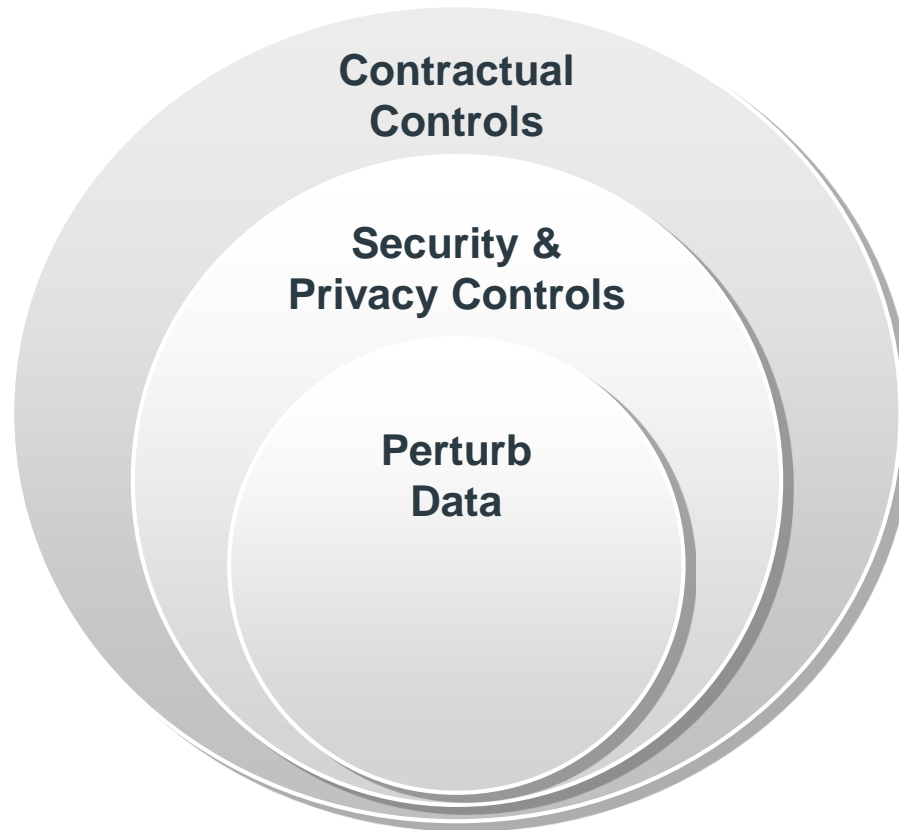
# Factors Affecting Risk

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.





# Layers of Protection



# Contact



[kelemam@privacy-analytics.com](mailto:kelemam@privacy-analytics.com)



[@kelemam](https://twitter.com/kelemam)



[www.privacy-analytics.com](http://www.privacy-analytics.com)