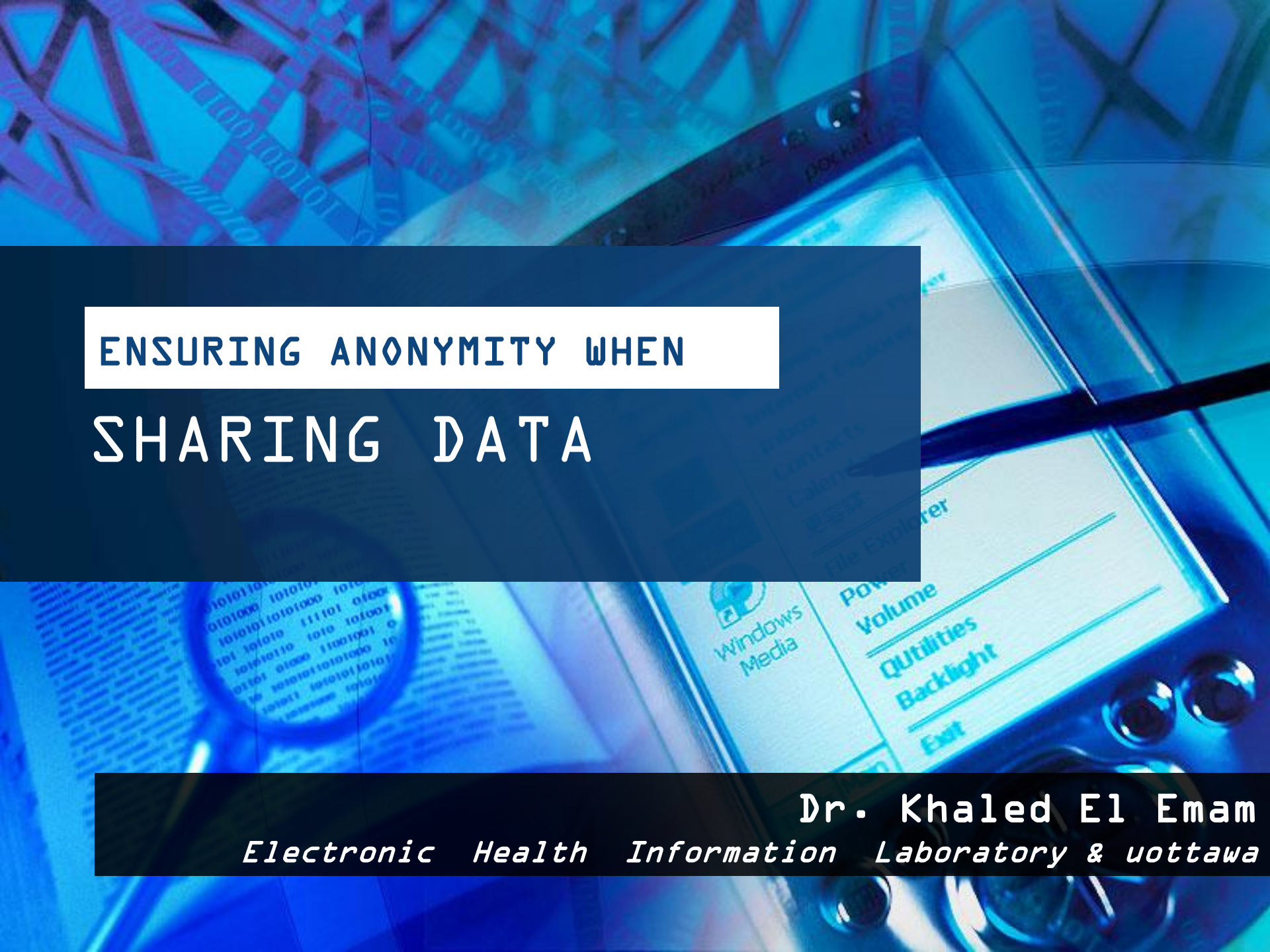


ENSURING ANONYMITY WHEN SHARING DATA

Dr. Khaled El Emam
Electronic Health Information Laboratory & uottawa

ANONYMIZATION

Motivations for Anonymization

- Obtaining patient consent/authorization – not practical for large databases and introduces bias
- Limiting principles / minimal necessary
- Contractual obligations
- Maintain public / consumer / client trust
- Costs of breach notification
- Rising discipline of re-identification attacks

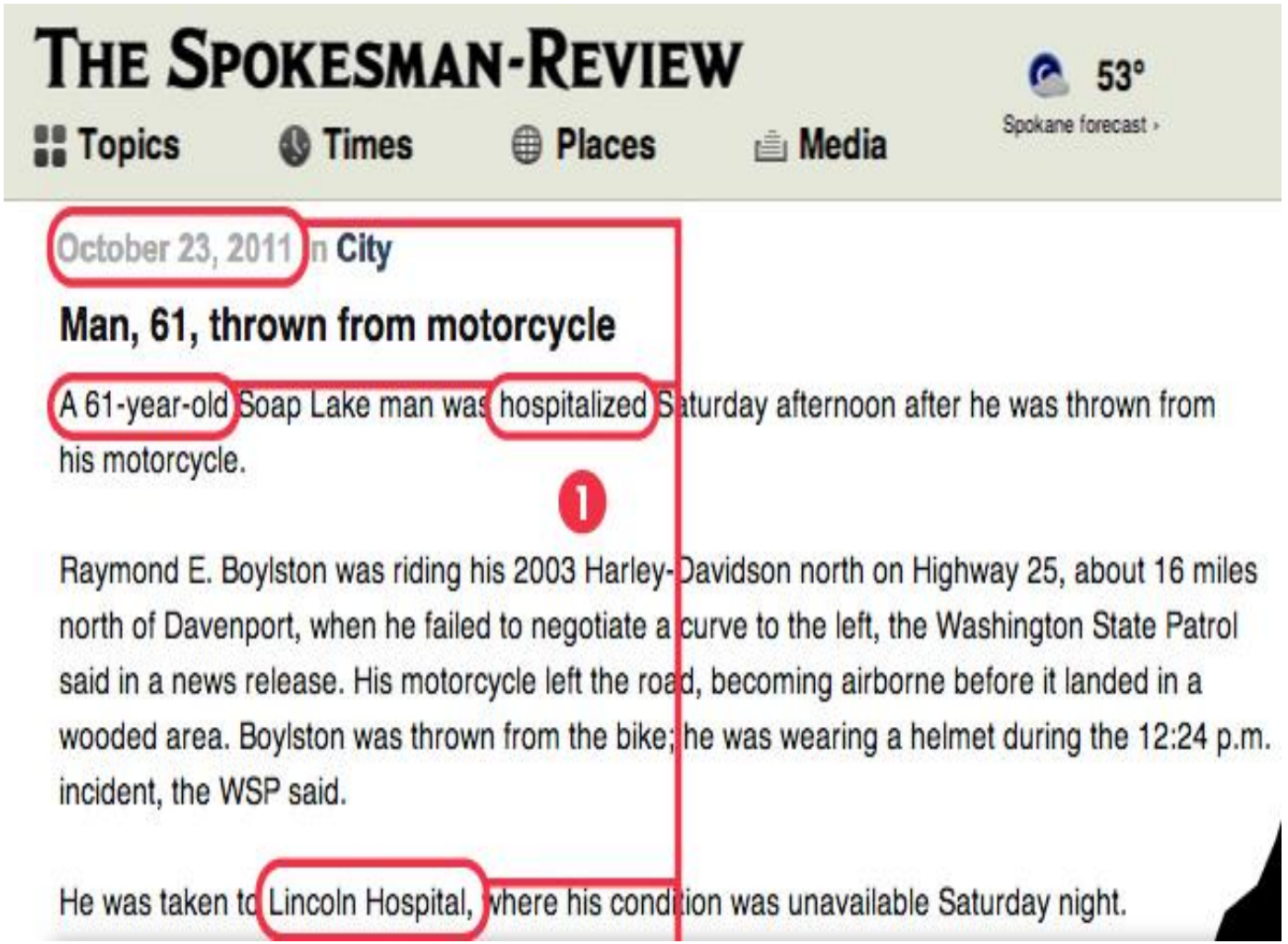
Bill of Health
Examining the intersection of law and health care, biotech & bioethics
A blog by the Petrie-Flom Center and friends

Posted on May 23, 2013 by Meyer Michelle

← Previous Next →

Breaking Good: A Short Ethical Manifesto for the Privacy Researcher

This post is part of *Bill of Health*'s symposium on the Law, Ethics, and Science of Re-Identification Demonstrations. We'll have more contributions throughout the week, and extending at least into early next week. Background on the symposium is here. You can call up all of the symposium contributions here (or by clicking on the "Re-Identification



THE SPOKESMAN-REVIEW

53°
Spokane forecast ›

Topics Times Places Media

October 23, 2011 n City

Man, 61, thrown from motorcycle

A 61-year-old Soap Lake man was hospitalized Saturday afternoon after he was thrown from his motorcycle.

1

Raymond E. Boylston was riding his 2003 Harley-Davidson north on Highway 25, about 16 miles north of Davenport, when he failed to negotiate a curve to the left, the Washington State Patrol said in a news release. His motorcycle left the road, becoming airborne before it landed in a wooded area. Boylston was thrown from the bike; he was wearing a helmet during the 12:24 p.m. incident, the WSP said.

He was taken to Lincoln Hospital, where his condition was unavailable Saturday night.

http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0028071

OPEN ACCESS Freely available online

PLoS one

A Systematic Review of Re-Identification Attacks on Health Data

Khaled El Emam[1,2]*, Elizabeth Jonker[1], Luk Arbuckle[1], Bradley Malin[3,4]

1 Electronic Health Information Laboratory, CHEO Research Institute, Ottawa, Canada, **2** Department of Paediatrics, University of Ottawa, Ottawa, Canada, **3** Department of Biomedical Informatics, Vanderbilt University, Nashville, Tennessee, United States of America, **4** Department of Electrical Engineering and Computer Science, Vanderbilt University, Nashville, Tennessee, United States of America

Abstract

Background: Privacy legislation in most jurisdictions allows the disclosure of health data for secondary purposes without patient consent if it is de-identified. Some recent articles in the medical, legal, and computer science literature have argued that de-identification methods do not provide sufficient protection because they are easy to reverse. Should this be the case, it would have significant and important implications on how health information is disclosed, including: (a) potentially limiting its availability for secondary purposes such as research, and (b) resulting in more identifiable health information being disclosed. Our objectives in this systematic review were to: (a) characterize known re-identification attacks on health data and contrast that to re-identification attacks on other kinds of data, (b) compute the overall proportion of records that have been correctly re-identified in these attacks, and (c) assess whether these demonstrate weaknesses in current de-identification methods.

Anonymization = Risk Management

No Zero Risk

Anonymization Standards

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.

Anonymisation: managing data protection risk code of practice

STATISTICAL POLICY
WORKING PAPER 22 (Second version, 2005)

Report on Statistical Disclosure Limitation Methodology

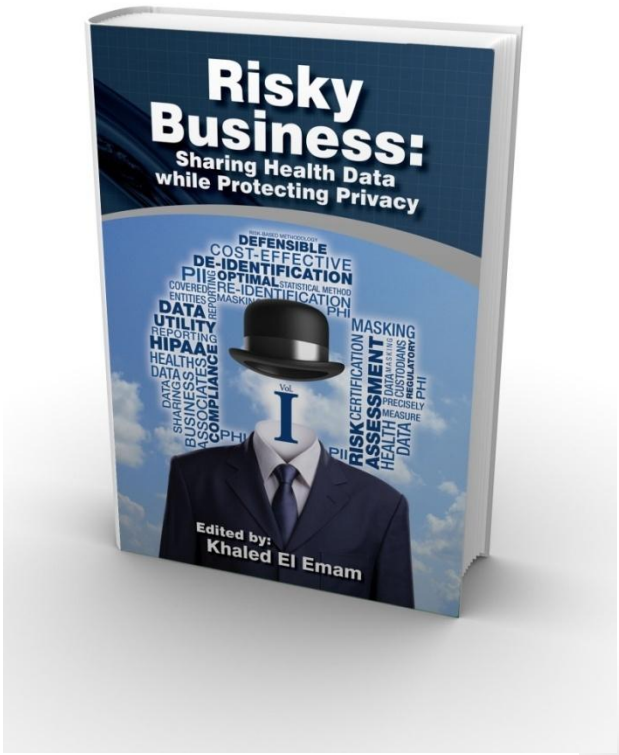
Federal Committee on Statistical Methodology

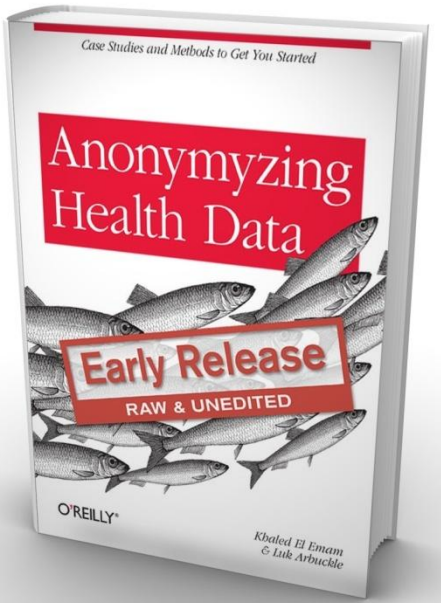
'Best Practice' Guidelines for Managing the Disclosure of De-Identified Health Information

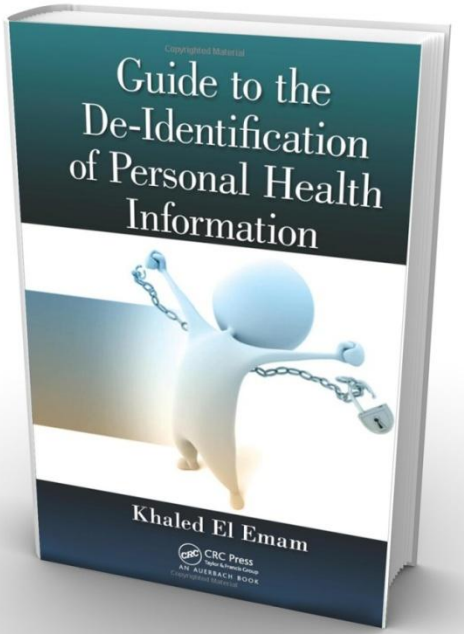
Prepared by the:
Health System Use Technical Advisory Committee
Data De-Identification Working Group

October 2010

Anonymization Methods & Examples



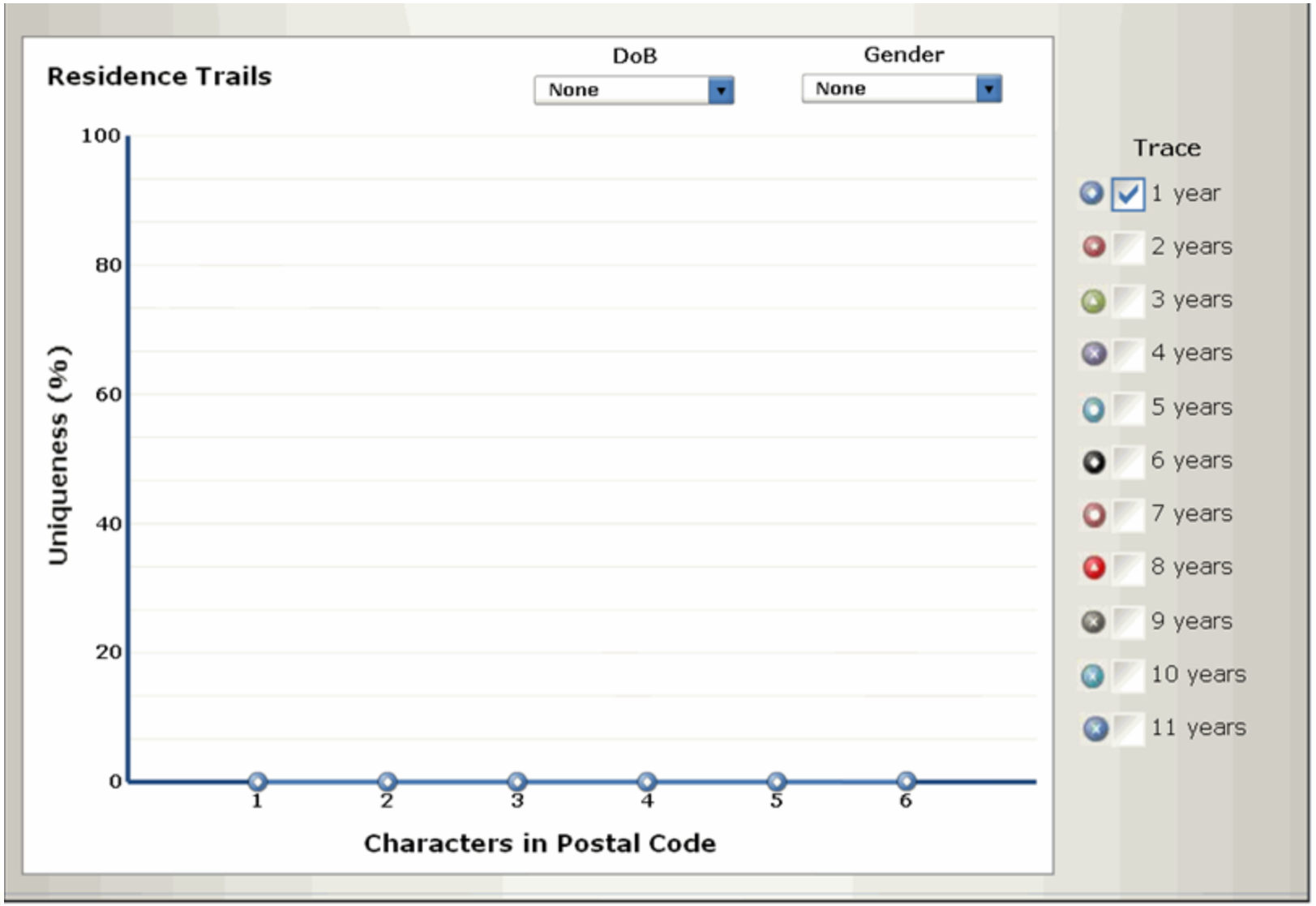


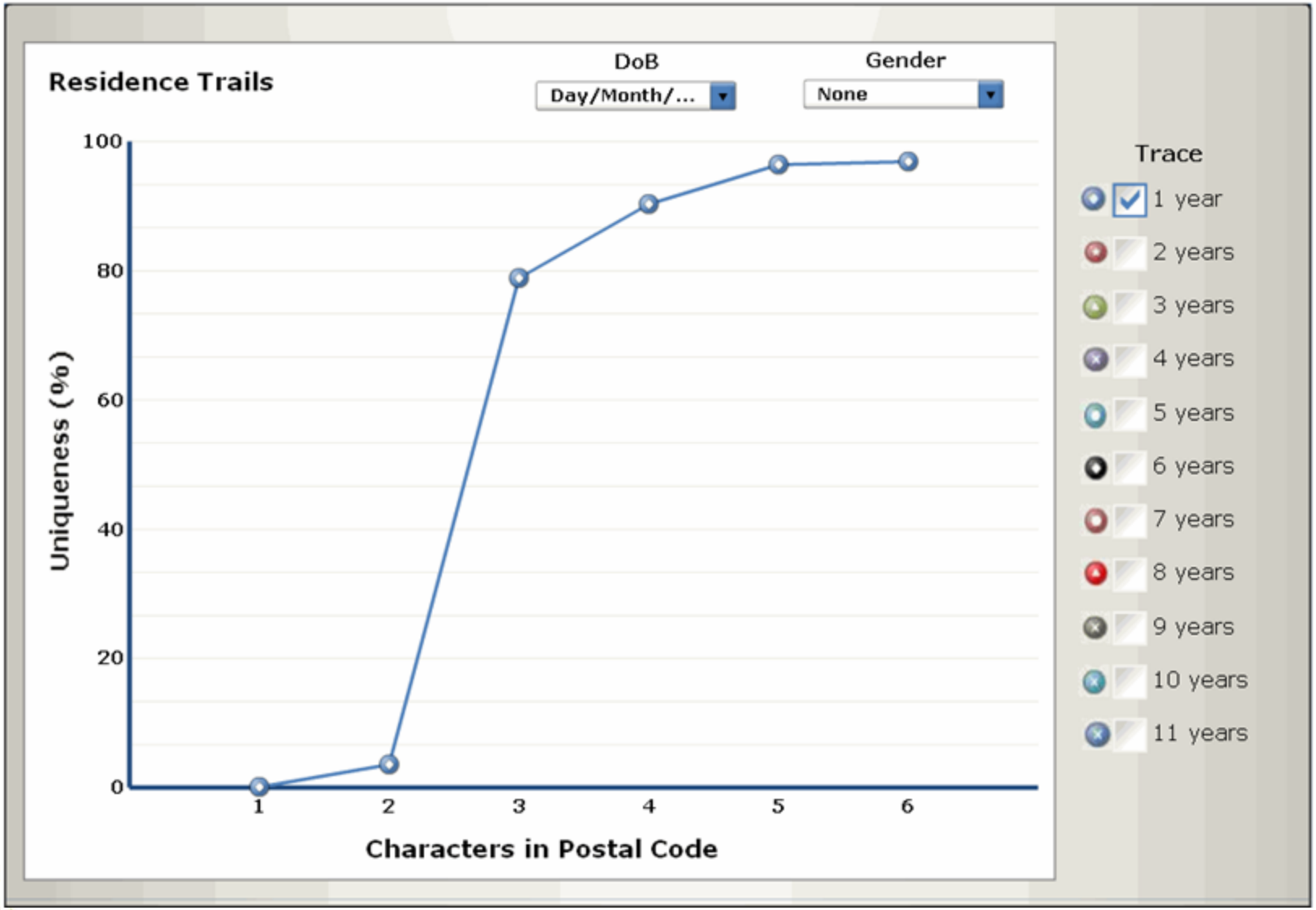


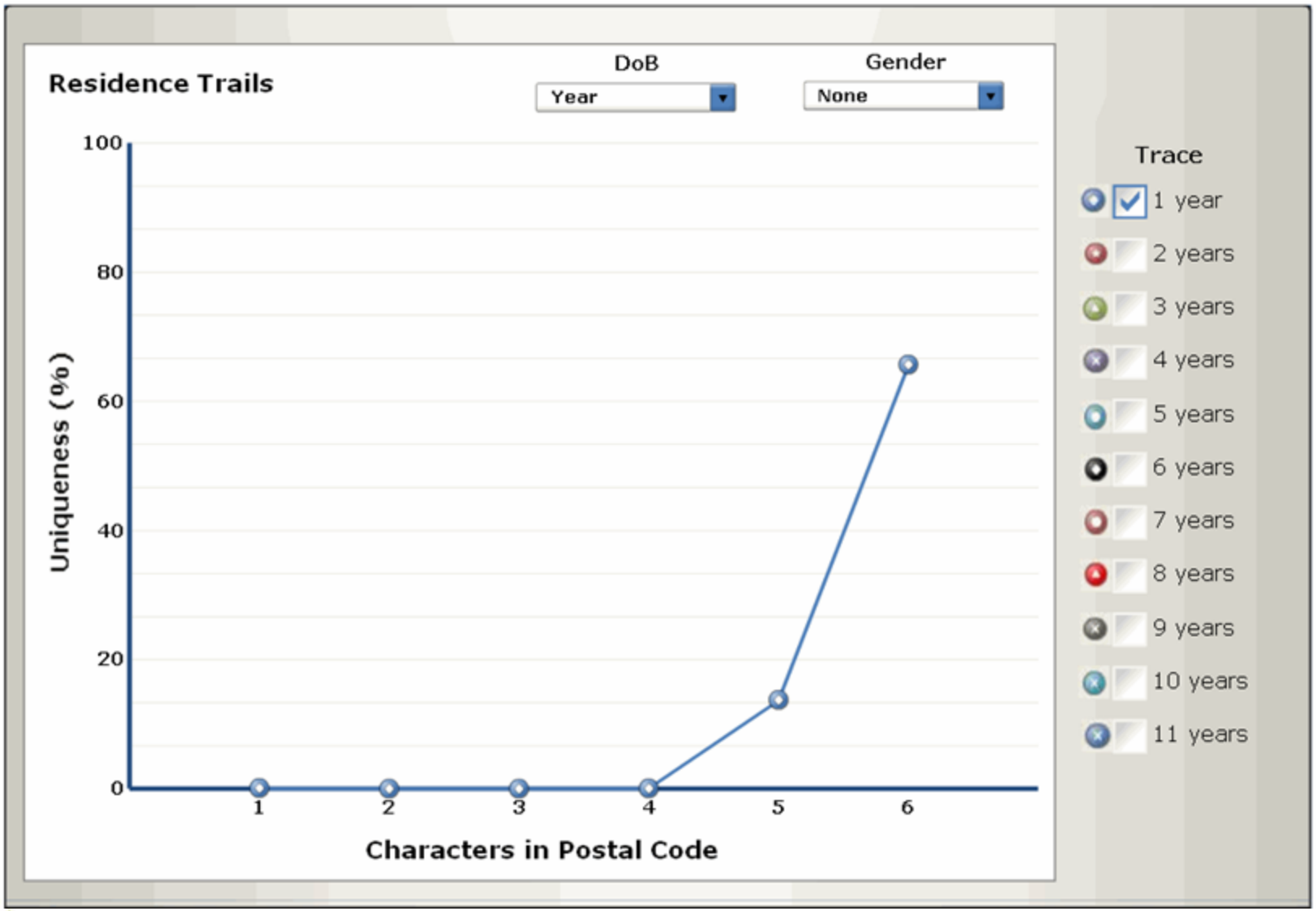
Direct & Indirect Identifiers

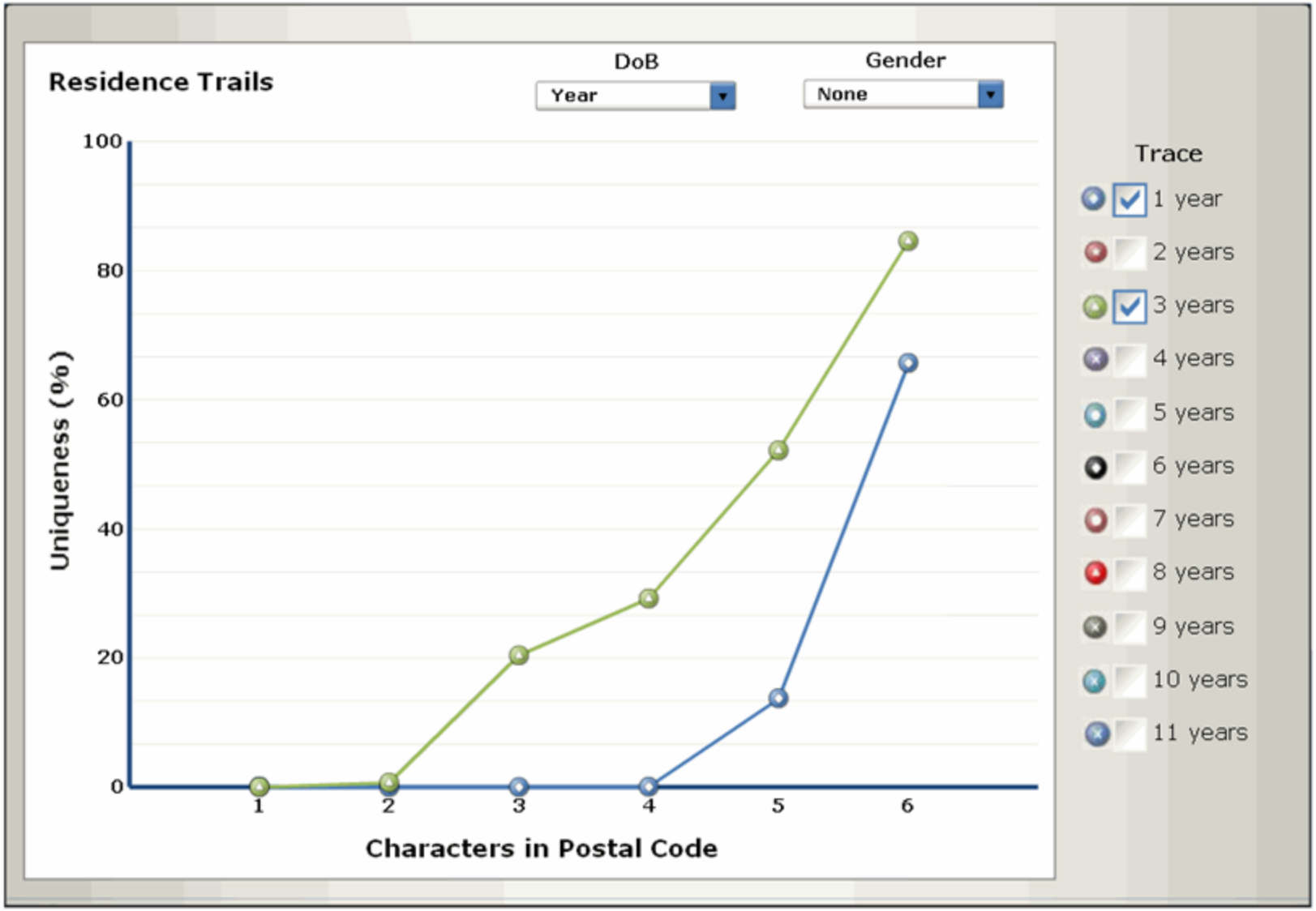
Examples of direct identifiers: Name, address, telephone number, fax number, MRN, health card number, health plan beneficiary number, VID, license plate number, email address, photograph, biometrics, SSN, SIN, device number, clinical trial record number

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

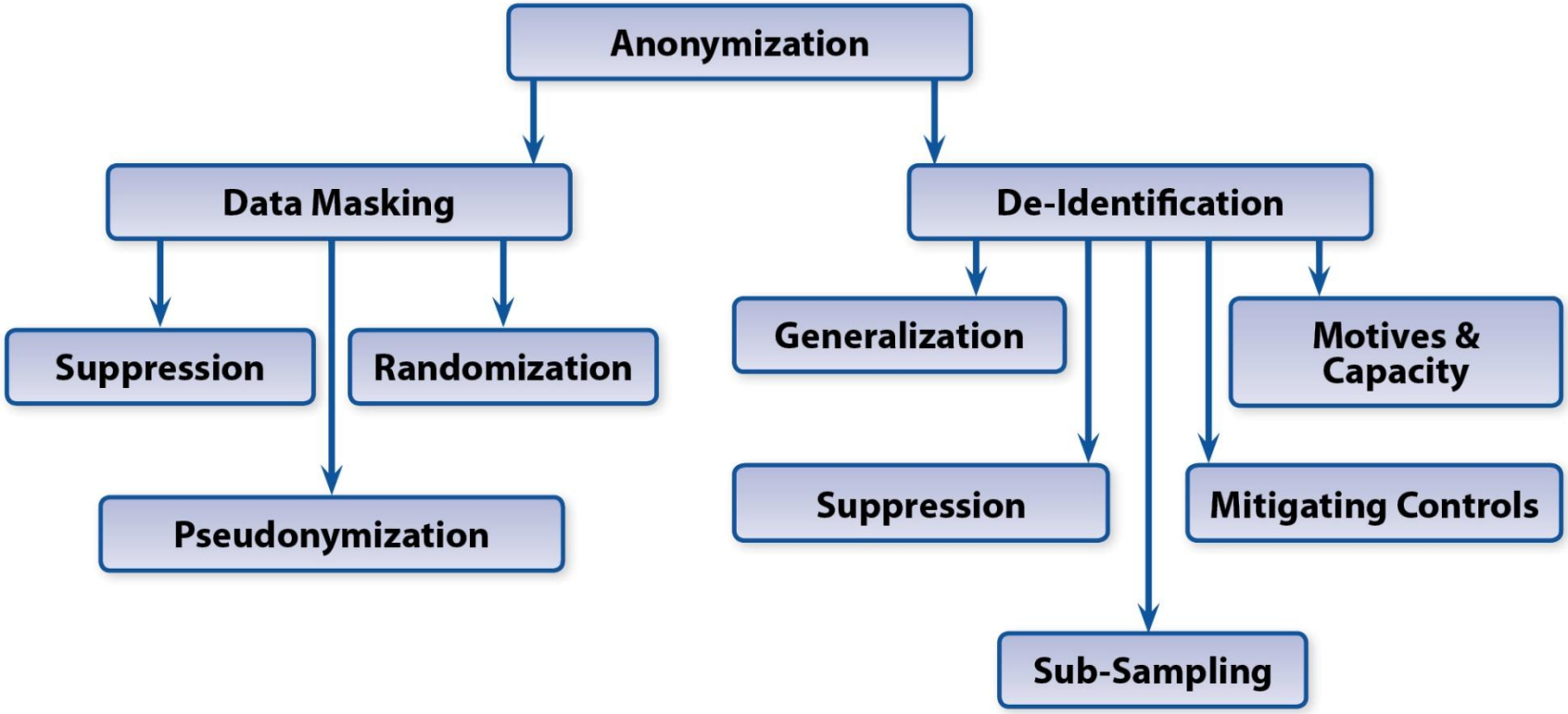

Residence Trails
DoB
None
Gender
None
100
80
60
40
20
0
Uniqueness (%)
1
2
3
4
5
6
Characters in Postal Code
Trace
1 year
2 years
3 years
4 years
5 years
6 years
7 years
8 years
9 years
10 years
11 years


Residence Trails
DoB
Day/Month/...
Gender
None
Uniqueness (%)
100
80
60
40
20
0
1
2
3
4
5
6
Characters in Postal Code
Trace
1 year
2 years
3 years
4 years
5 years
6 years
7 years
8 years
9 years
10 years
11 years

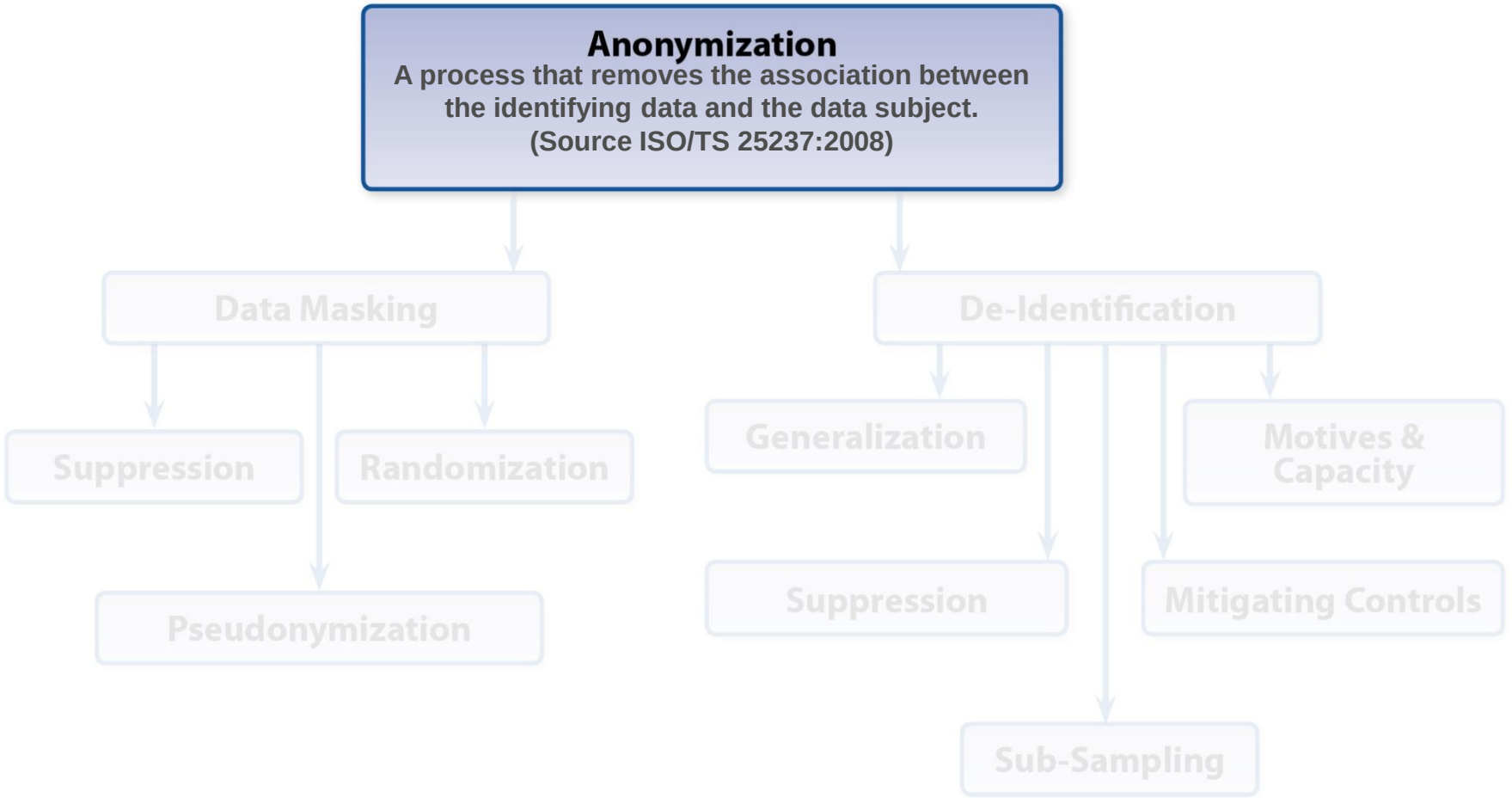

Residence Trails
DoB
Year
Gender
None
Trace
1 year
2 years
3 years
4 years
5 years
6 years
7 years
8 years
9 years
10 years
11 years
Uniqueness (%)
100
80
60
40
20
0
1
2
3
4
5
6
Characters in Postal Code


Residence Trails
DoB
Year
Gender
None
Uniqueness (%)
100
80
60
40
20
0
1
2
3
4
5
6
Characters in Postal Code
Trace
1 year
2 years
3 years
4 years
5 years
6 years
7 years
8 years
9 years
10 years
11 years

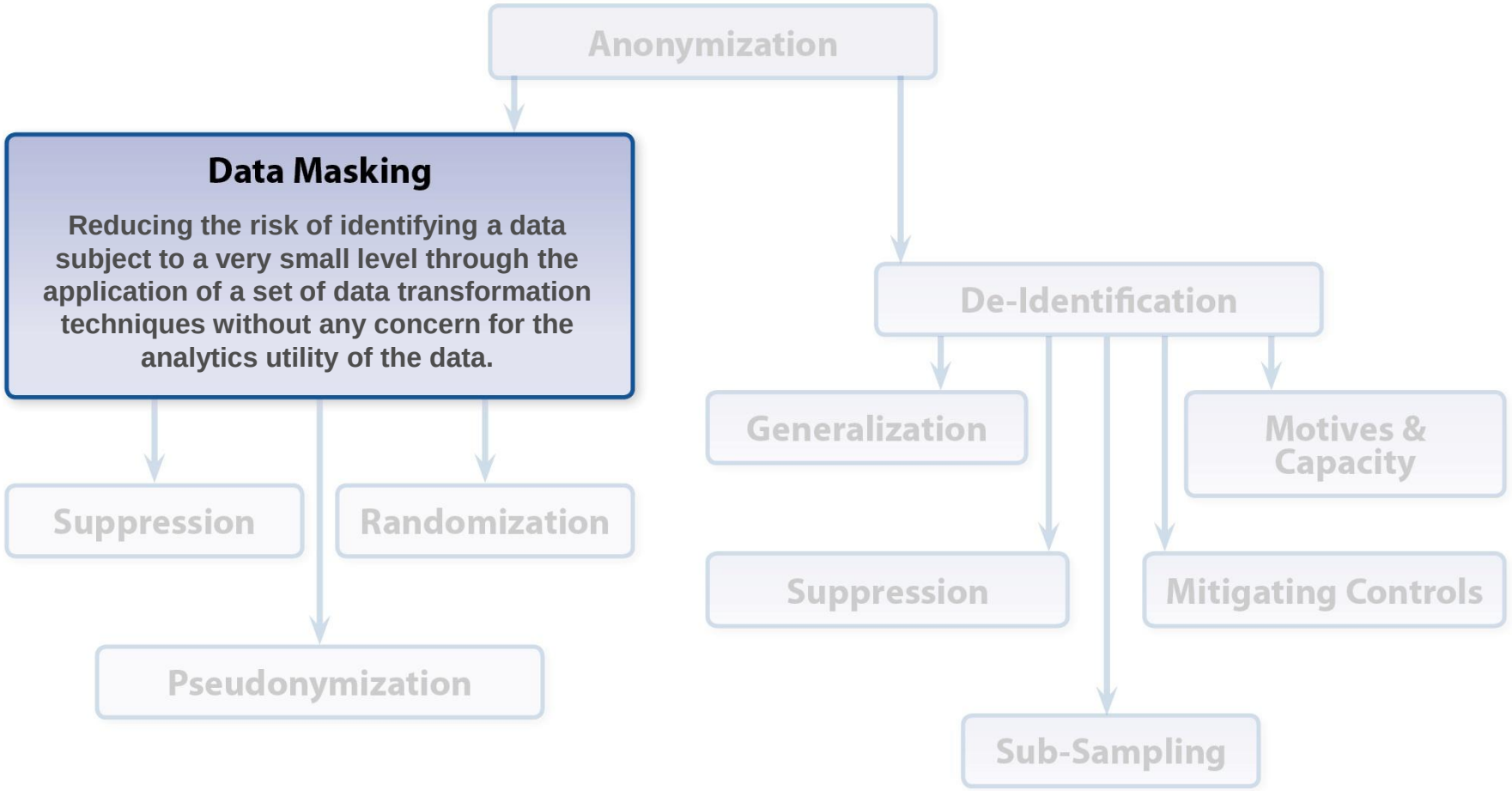
Landscape


Anonymization
Data Masking
De-Identification
Suppression
Randomization
Pseudonymization
Generalization
Motives & Capacity
Suppression
Mitigating Controls
Sub-Sampling

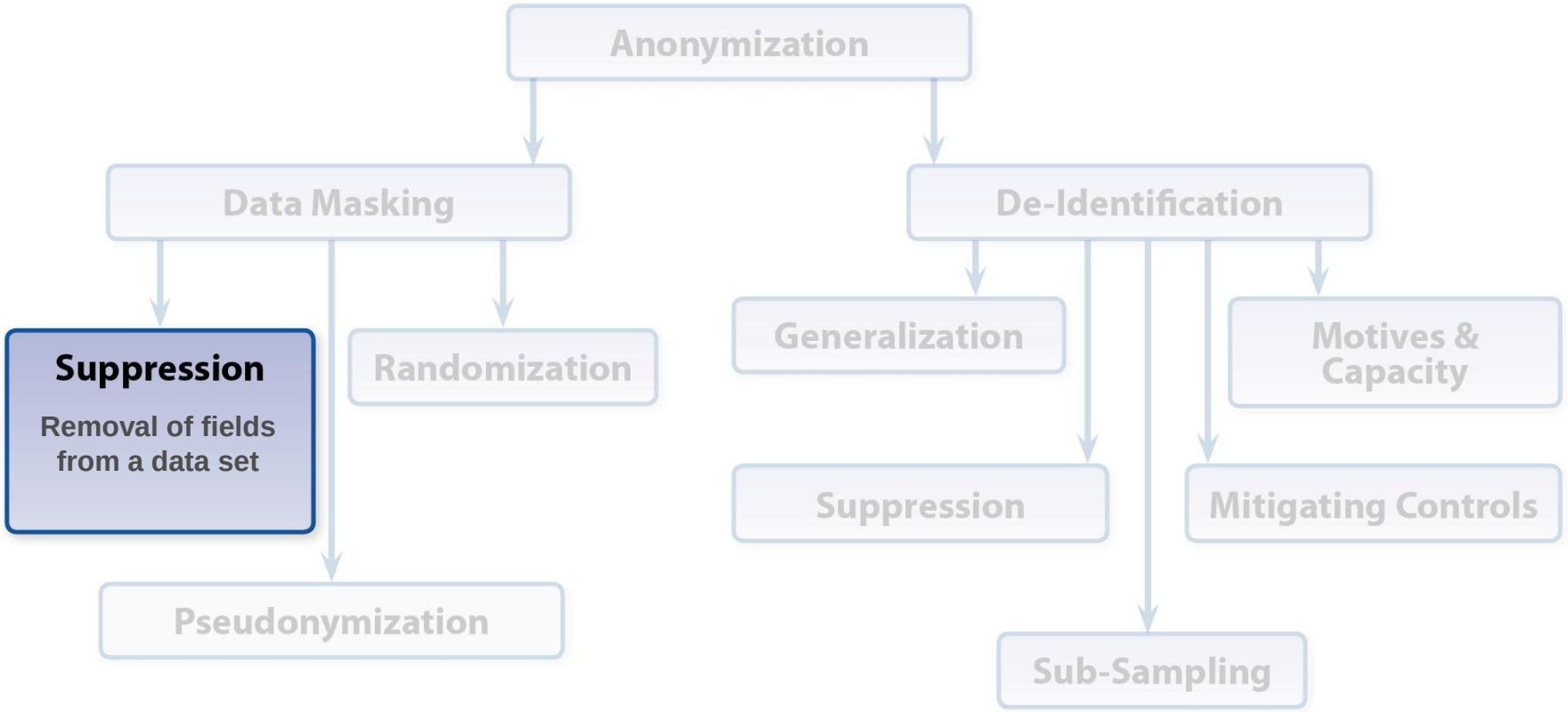
Landscape


Anonymization
A process that removes the association between the identifying data and the data subject.
(Source ISO/TS 25237:2008)
Data Masking
De-Identification
Suppression
Randomization
Pseudonymization
Generalization
Motives & Capacity
Suppression
Mitigating Controls
Sub-Sampling

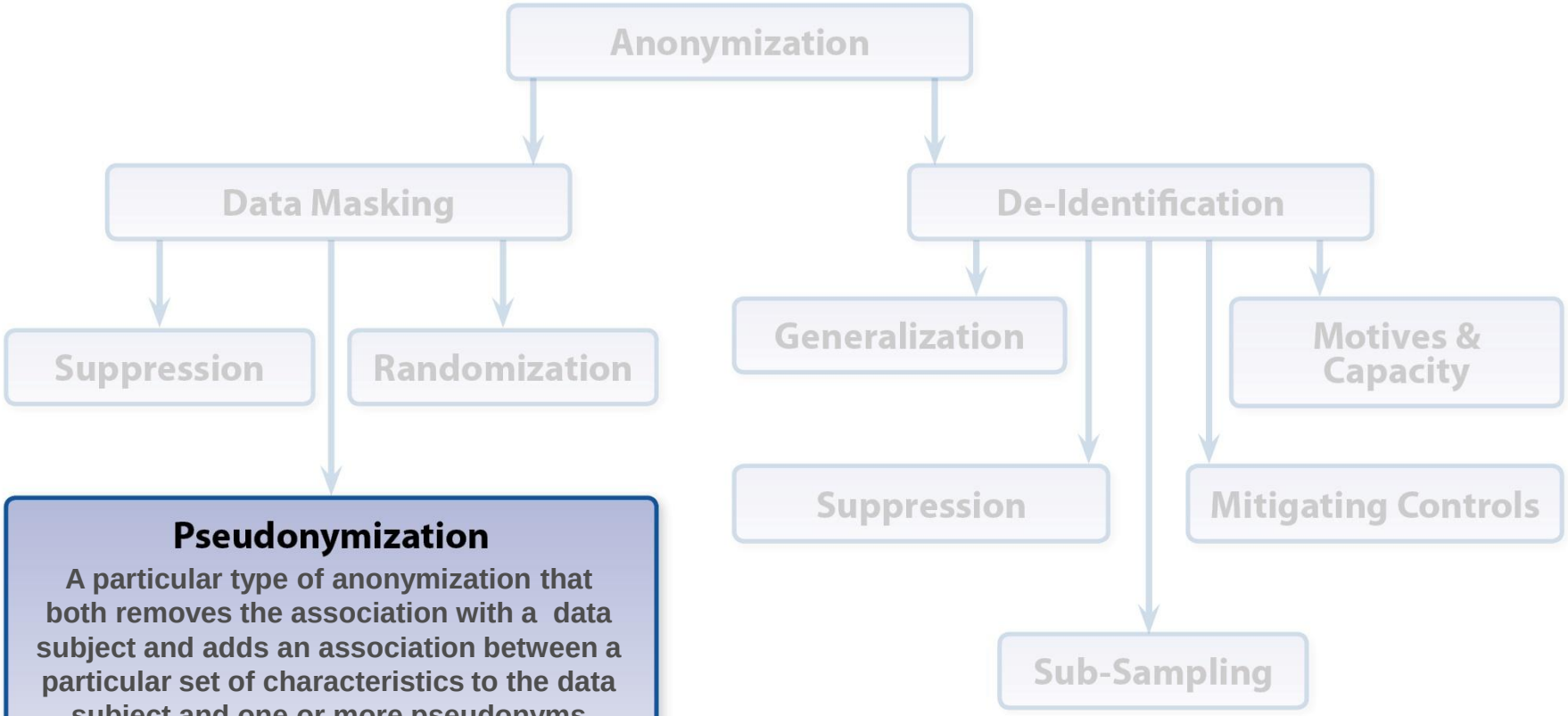
Landscape


Anonymization
Data Masking
Reducing the risk of identifying a data subject to a very small level through the application of a set of data transformation techniques without any concern for the analytics utility of the data.
Suppression
Randomization
Pseudonymization
De-Identification
Generalization
Suppression
Sub-Sampling
Motives & Capacity
Mitigating Controls

Landscape


Anonymization
Data Masking
De-Identification
Suppression
Removal of fields from a data set
Randomization
Pseudonymization
Generalization
Suppression
Sub-Sampling
Motives & Capacity
Mitigating Controls

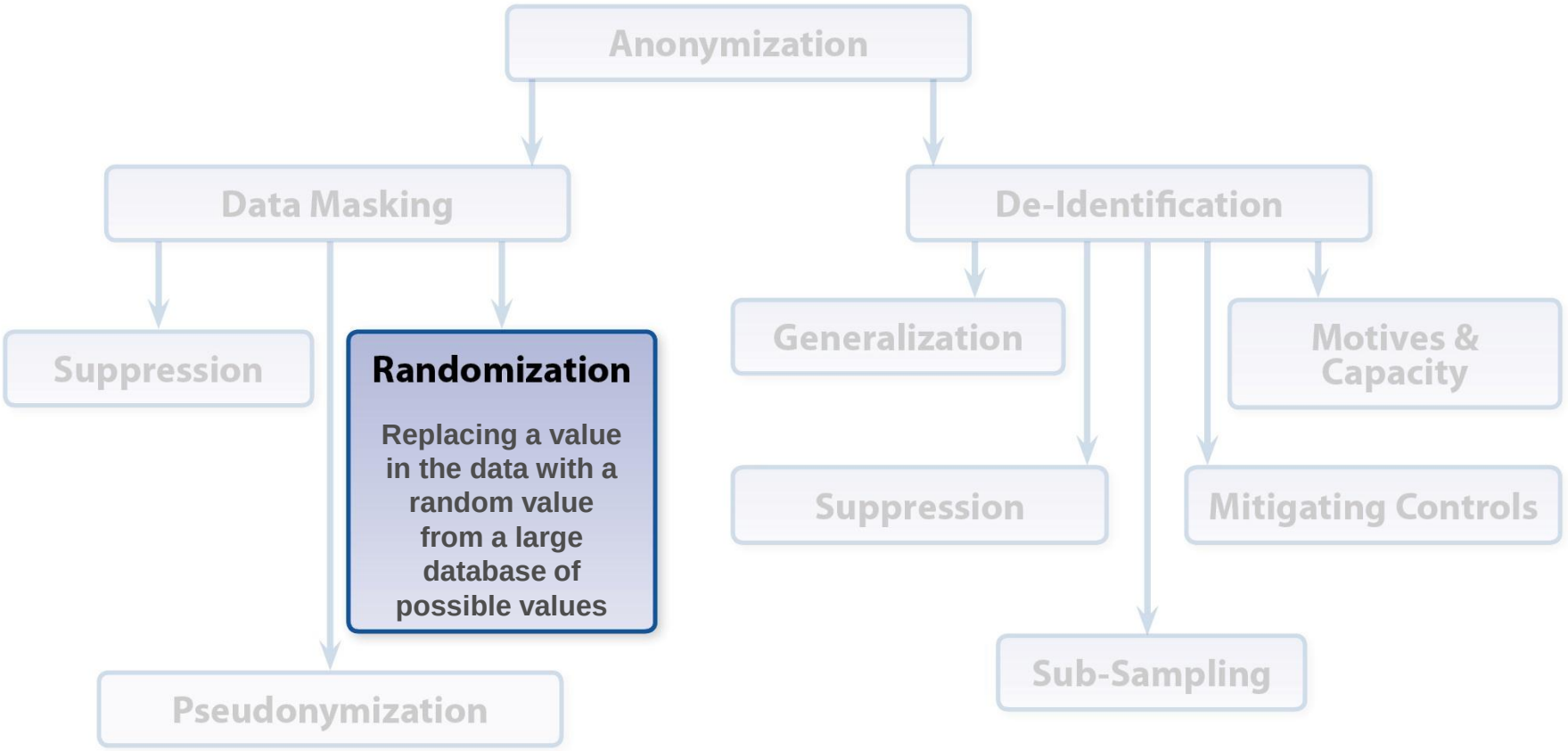
Landscape



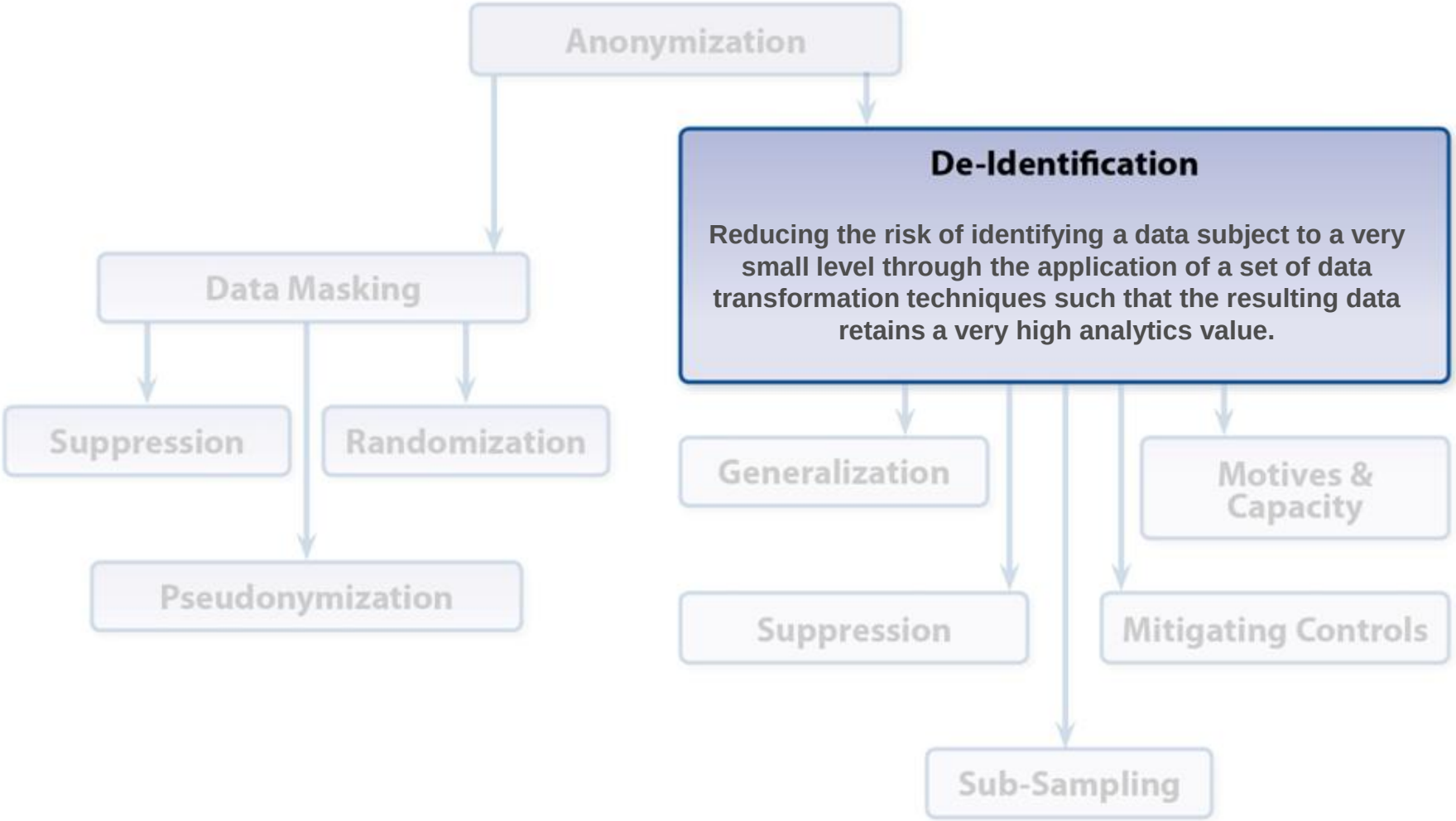
Pseudonymization

A particular type of anonymization that both removes the association with a data subject and adds an association between a particular set of characteristics to the data subject and one or more pseudonyms (Source: ISO/TS 25237:2008)

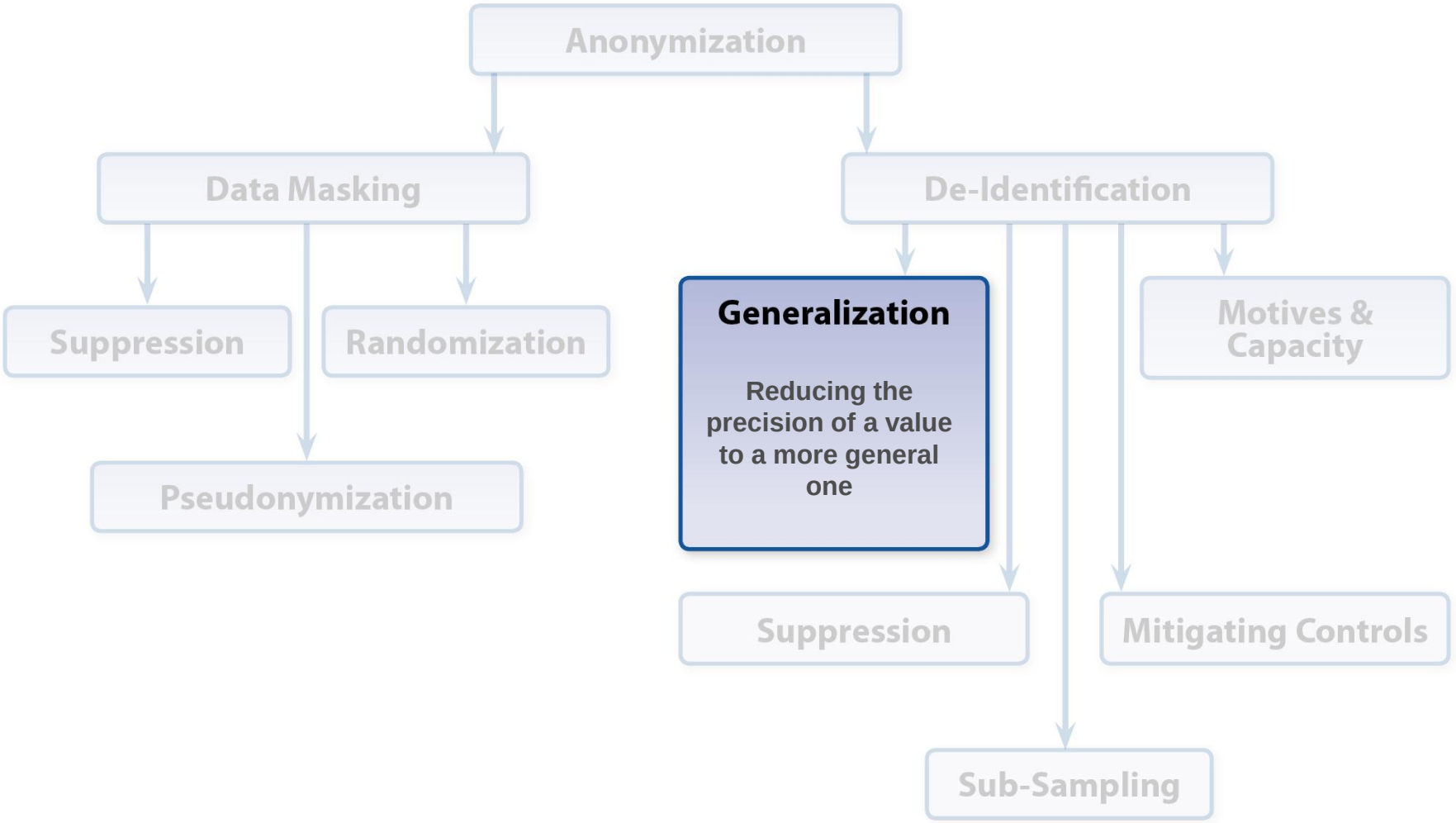
Landscape


Anonymization
Data Masking
De-Identification
Suppression
Randomization
Replacing a value in the data with a random value from a large database of possible values
Pseudonymization
Generalization
Motives & Capacity
Suppression
Mitigating Controls
Sub-Sampling

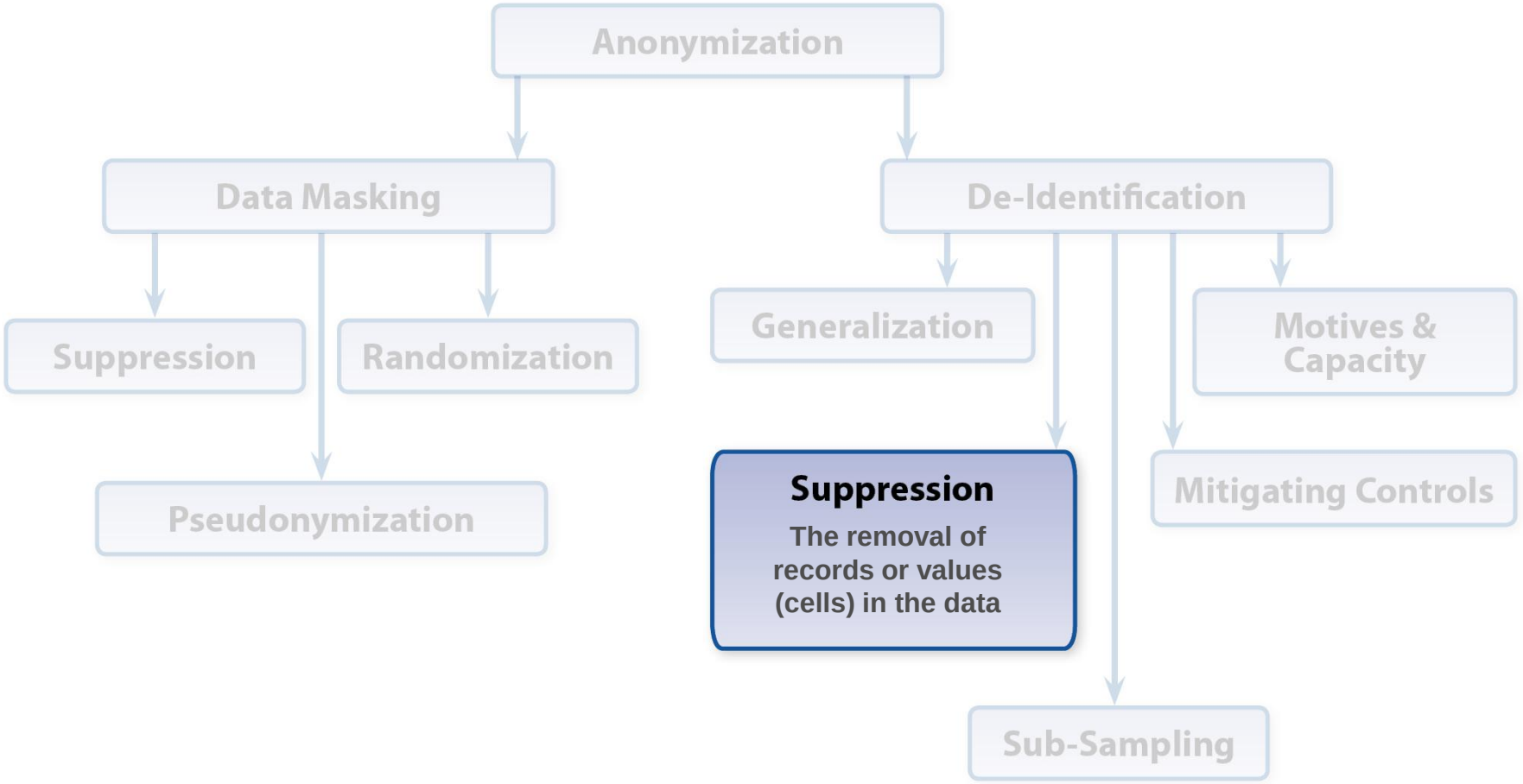
Landscape


Anonymization
De-Identification
Reducing the risk of identifying a data subject to a very small level through the application of a set of data transformation techniques such that the resulting data retains a very high analytics value.
Data Masking
Suppression
Randomization
Pseudonymization
Generalization
Motives & Capacity
Suppression
Mitigating Controls
Sub-Sampling

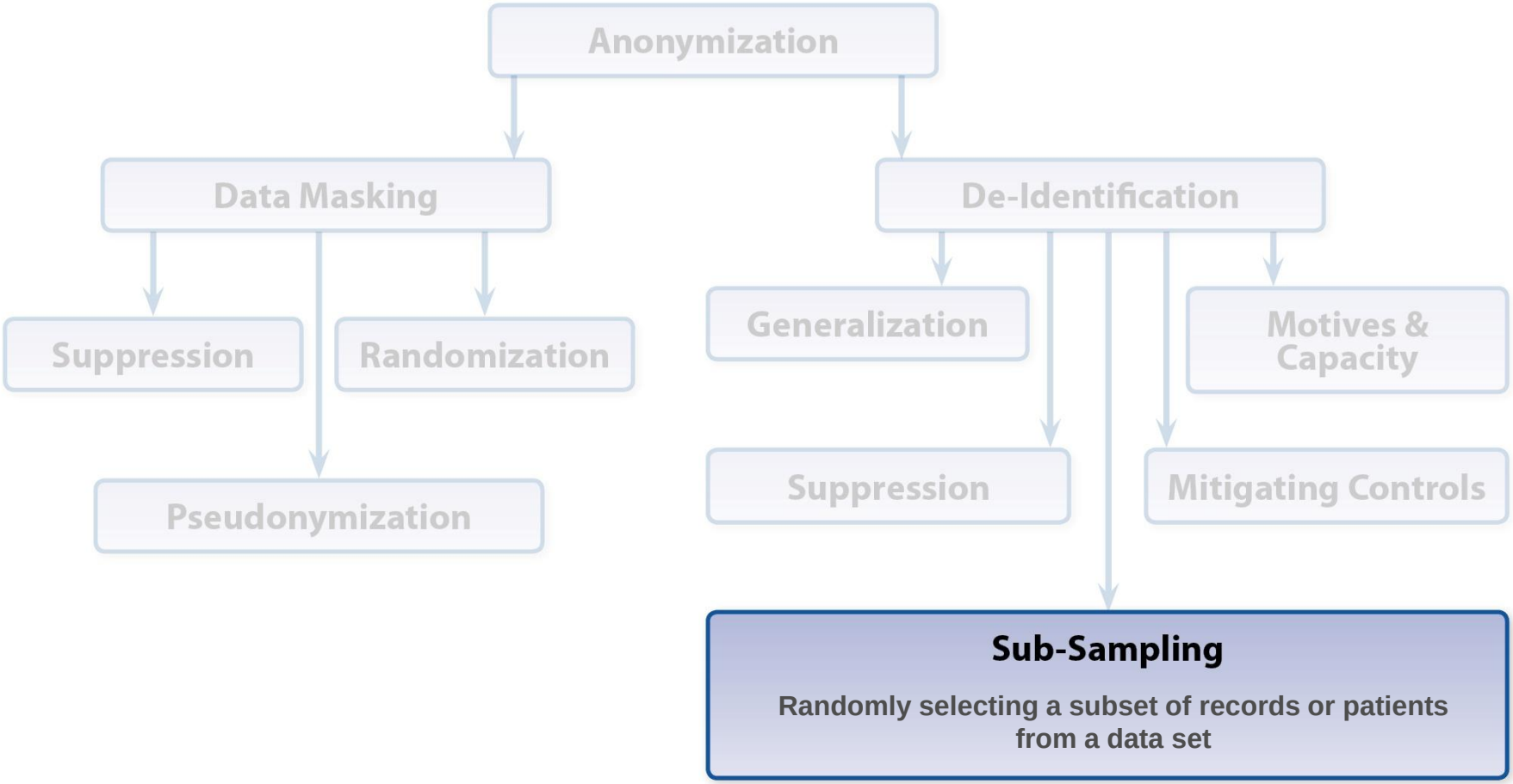
Landscape


Anonymization
Data Masking
De-Identification
Suppression
Randomization
Pseudonymization
Generalization
Reducing the precision of a value to a more general one
Motives & Capacity
Suppression
Mitigating Controls
Sub-Sampling

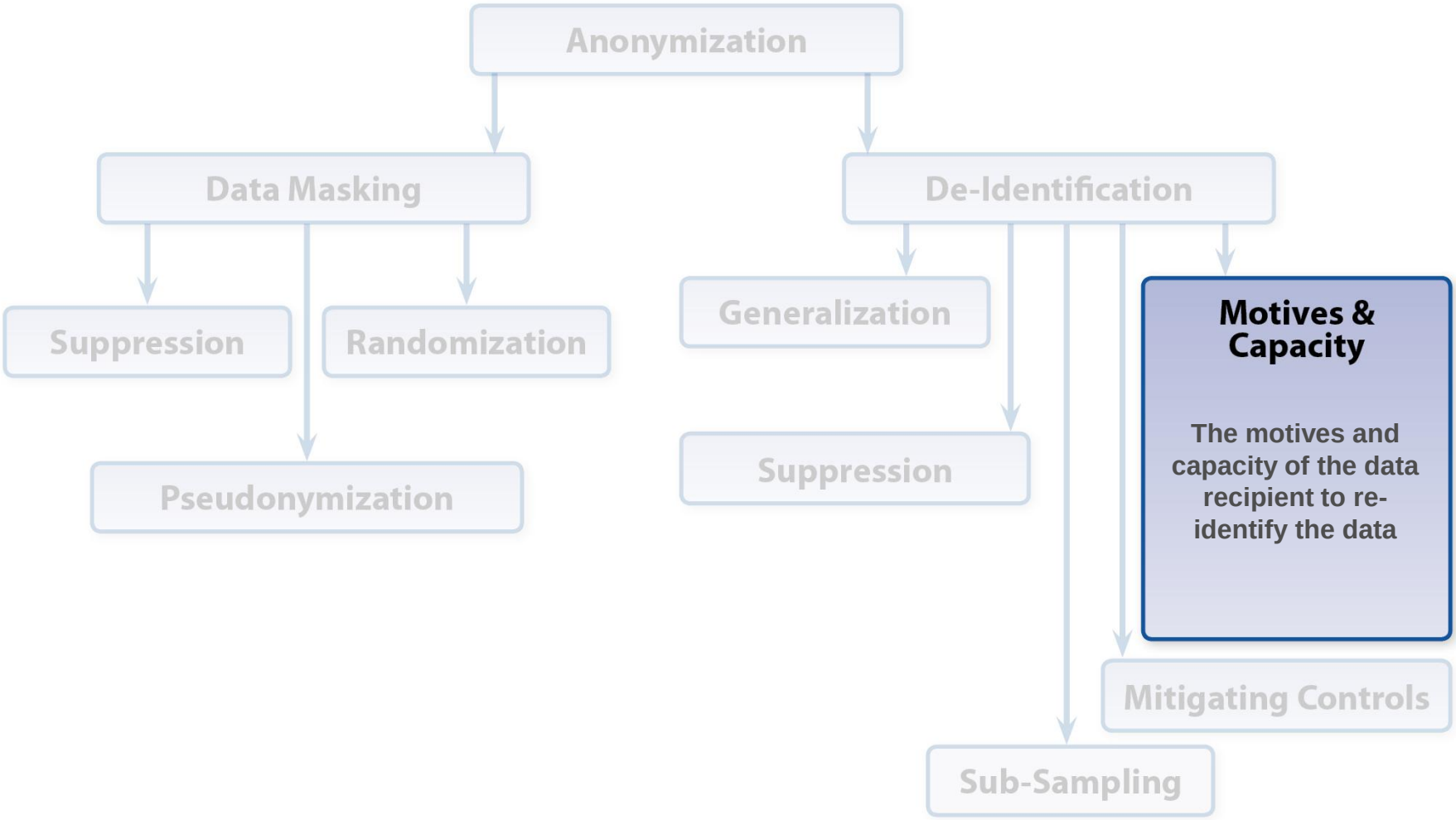
Landscape


Anonymization
Data Masking
De-Identification
Suppression
Randomization
Generalization
Motives & Capacity
Pseudonymization
Suppression
The removal of records or values (cells) in the data
Mitigating Controls
Sub-Sampling

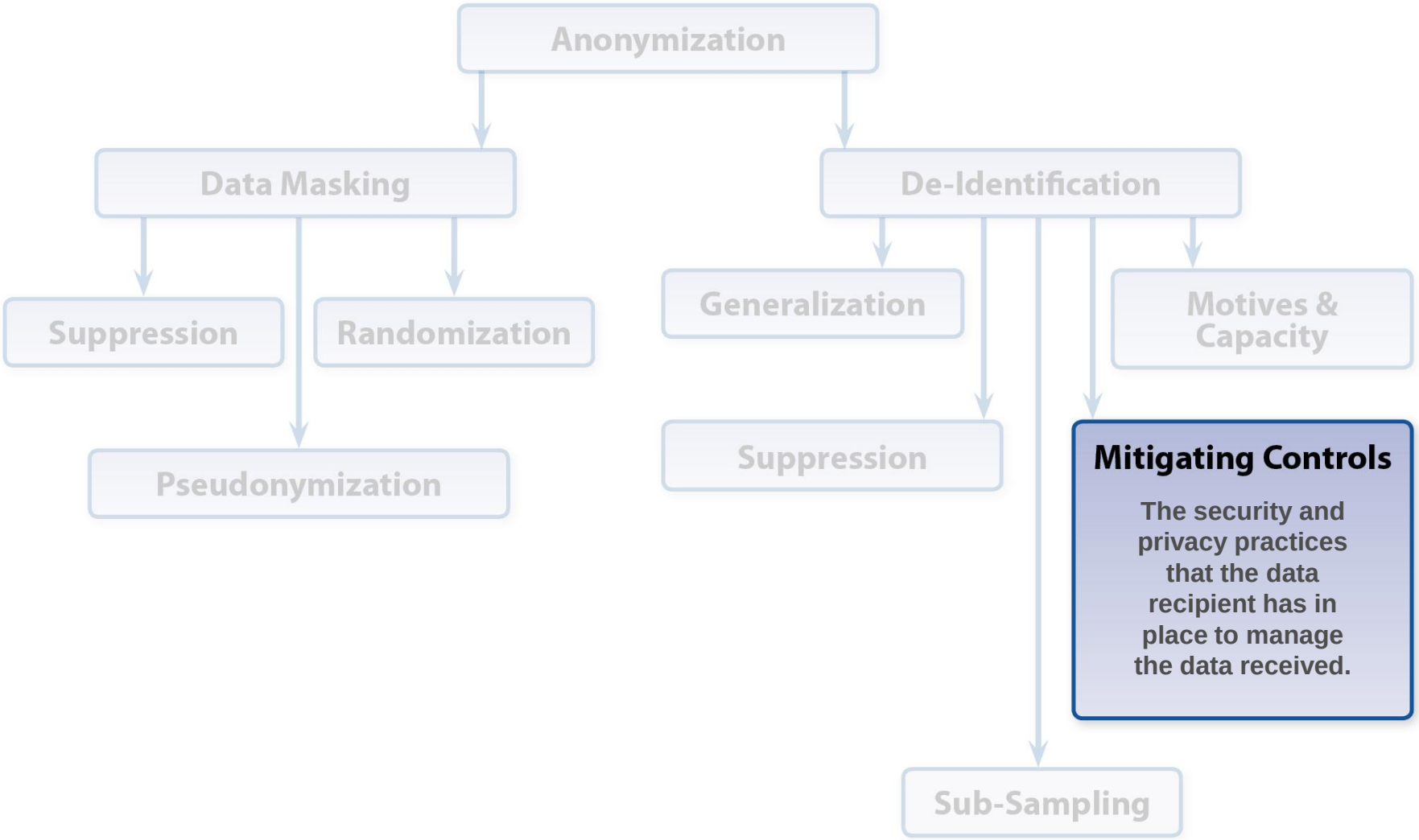
Landscape


Anonymization
Data Masking
De-Identification
Suppression
Randomization
Generalization
Motives & Capacity
Pseudonymization
Suppression
Mitigating Controls
Sub-Sampling
Randomly selecting a subset of records or patients from a data set

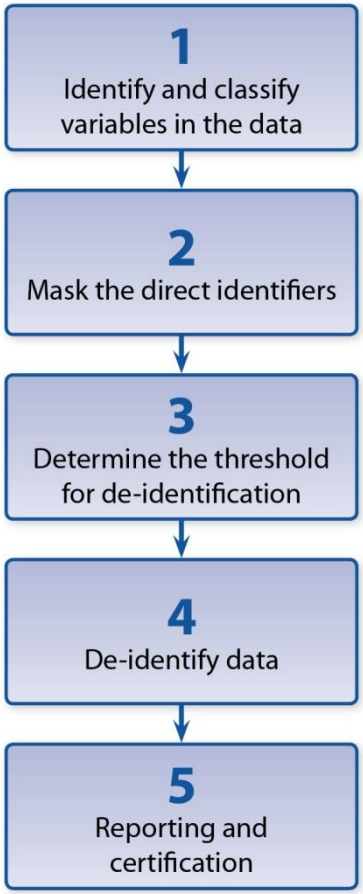
Landscape


Anonymization
Data Masking
De-Identification
Suppression
Randomization
Generalization
Suppression
Pseudonymization
Motives & Capacity
The motives and capacity of the data recipient to re-identify the data
Mitigating Controls
Sub-Sampling

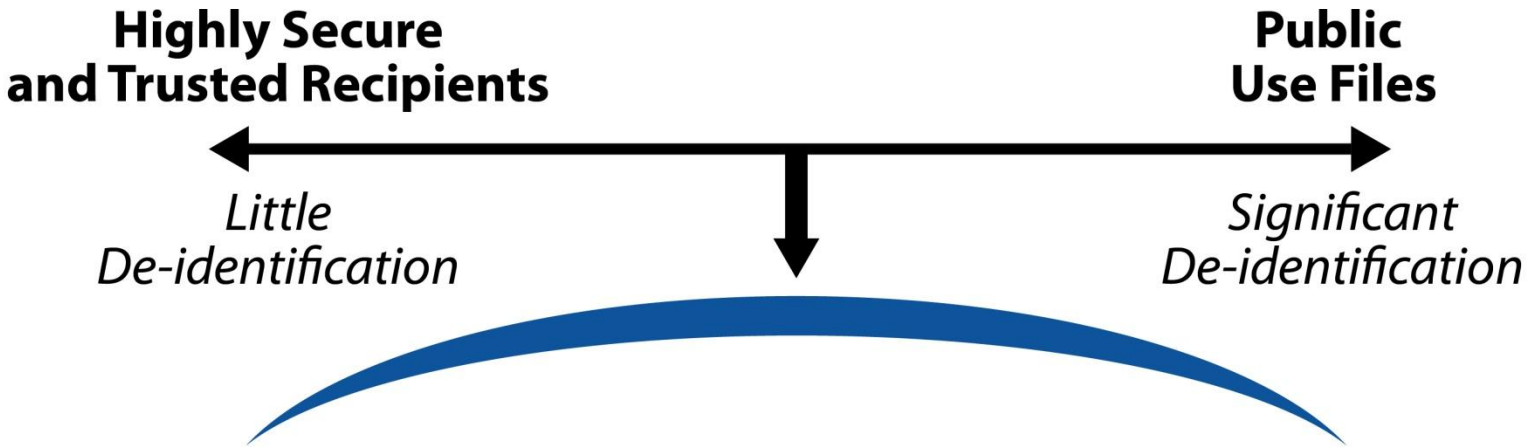
Landscape


Anonymization
Data Masking
De-Identification
Suppression
Randomization
Generalization
Motives & Capacity
Pseudonymization
Suppression
Mitigating Controls
The security and privacy practices that the data recipient has in place to manage the data received.
Sub-Sampling

Anonymization Process



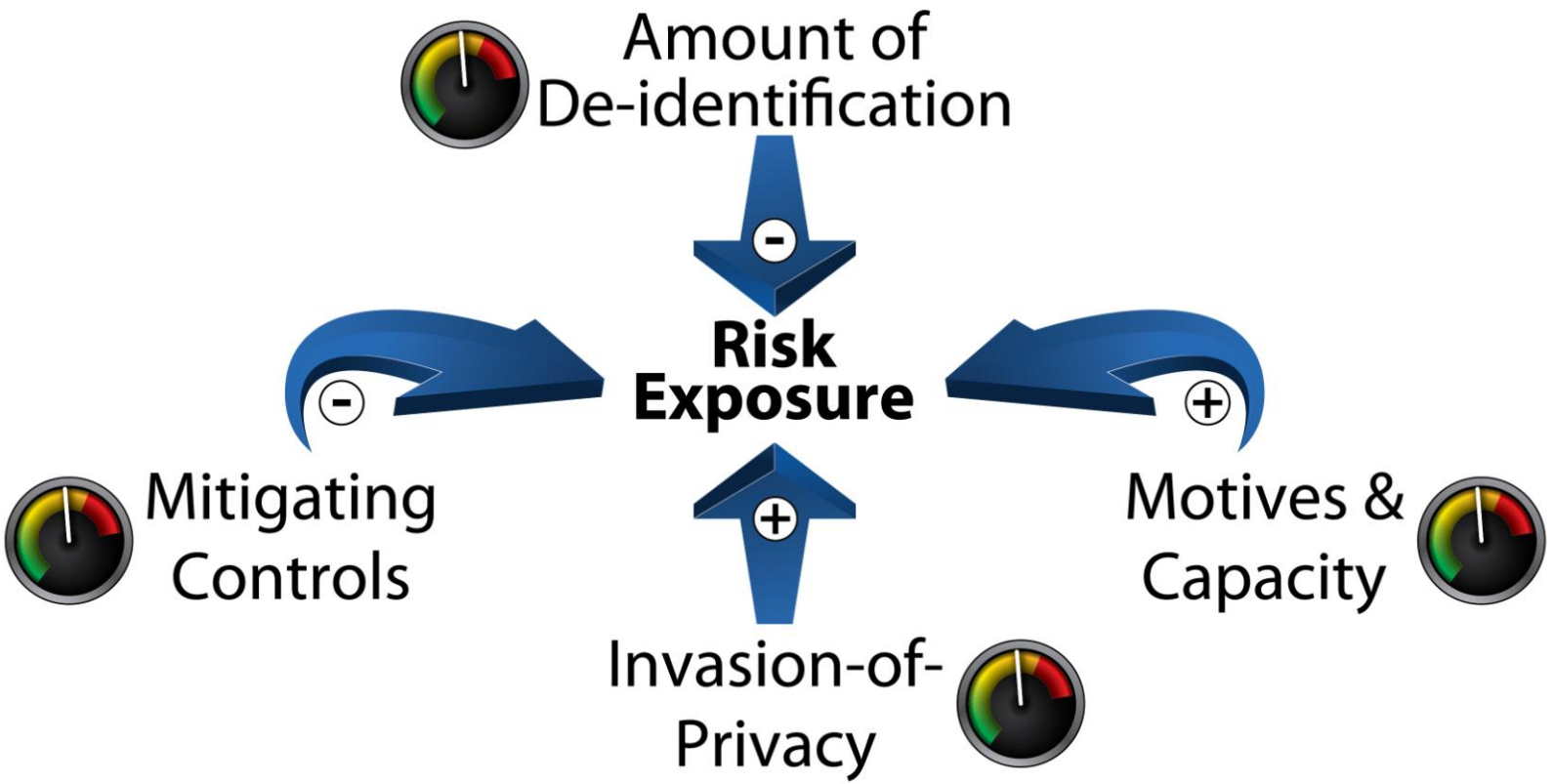
Choosing A Risk Threshold



- *Strong security & privacy practices*
- *Not sensitive data*
- *Consent sought or authority*
- *No motives to re-identify*

- *No security & privacy practices*
- *Sensitive data*
- *No authority to disclose data*
- *No consent sought*
- *Strong motives to re-identify*

Managing Re-identification Risk



Gordon Case

- Federal court case around an access request for information from the Canadian adverse drug event database
- Case between Health Canada and the CBC
- Judge ruled that the province of the affected individuals should not be released
- One of the main drivers was the ability to re-identify individuals from information in the adverse event record if it includes province
- Sets a precedent on identifiability of province

BMC Medical Informatics and Decision Making

BioMed Central
The Open Access Publisher

Evaluating the risk of patient re-identification from adverse drug event reports

BMC Medical Informatics and Decision Making 2013, **13**:114 doi:10.1186/1472-6947-13-114

Khaled El Emam (kelemam@uottawa.ca)
Fida K Dankar (fdankar@ehealthinformation.ca)
Angelica Neisa (aneisa@ehealthinformation.ca)
Elizabeth Jonker (ejonker@ehealthinformation.ca)

ISSN 1472-6947

Article type Research article

Submission date 29 April 2013

Acceptance date 25 September 2013

Publication date 5 October 2013

Article URL http://www.biomedcentral.com/1472-6947/13/114

Our Analysis

Table 4 The percentage of ADE deaths that are at a high risk of re-identification by matching to an obituary for different combinations of quasi-identifiers

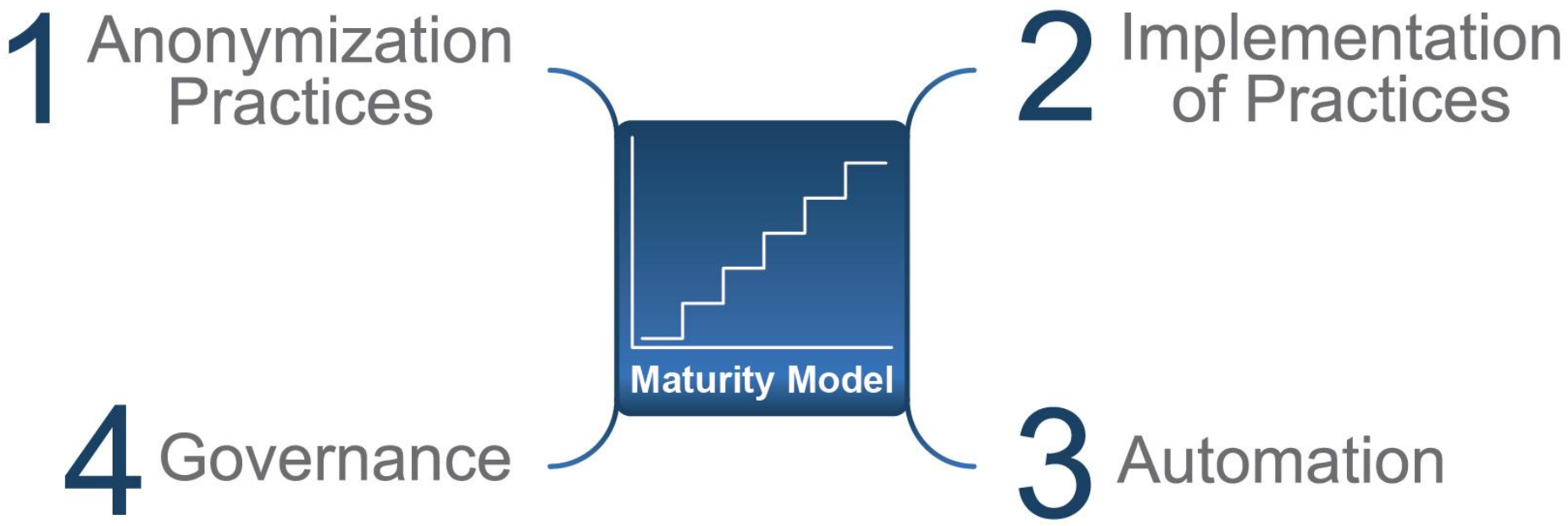
Quasi-identifiers						ADE records at risk (%)		
Province	Age at death	Gender	Day of report	Month of report	Year of report	p = 0.5	p = 0.7	p = 0.9
	X	X	X	X	X	1.95	3.46	5.05
	X	X		X	X	0	0	0
	X	X			X	0	0	0
	2 yr	X			X	0	0	0
	5 yr	X			X	0	0	0
	10 yr	X			X	0	0	0
X	X	X	X	X	X	18.44	25.17	30.78
X	X	X		X	X	0.21	0.4	0.63
X	X	X			X	0.12	0.24	0.39
X	2 yr	X			X	0.04	0.07	0.13
X	5 yr	X			X	0.02	0.03	0.05
X	10 yr	X			X	0	0	0.01

We considered age converted into a 2 year interval, 5 year interval, and a 10 year interval.

Our Analysis

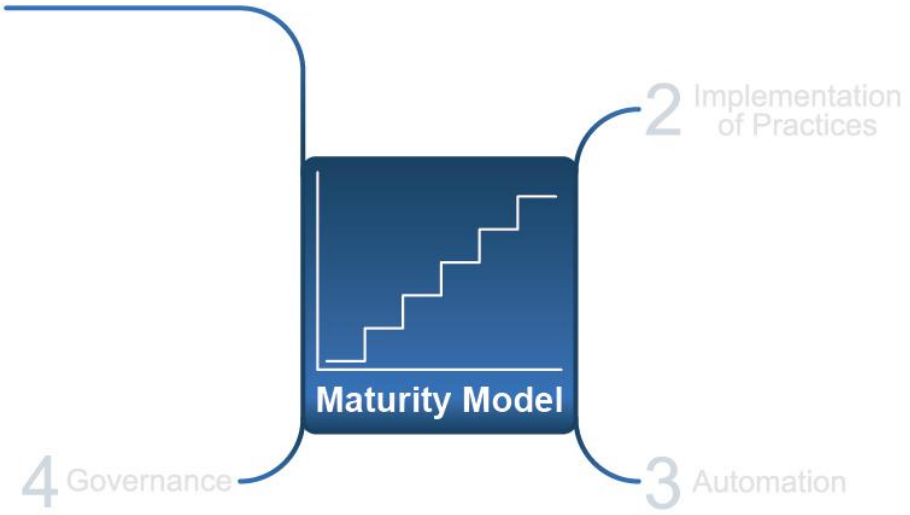
- It would have been possible to release province information if the date of death was reduced in precision
- A detailed analysis using the above risk-based methodology allows us to evaluate these tradeoffs, and find a solution that protects individual privacy but also facilitates the sharing of data

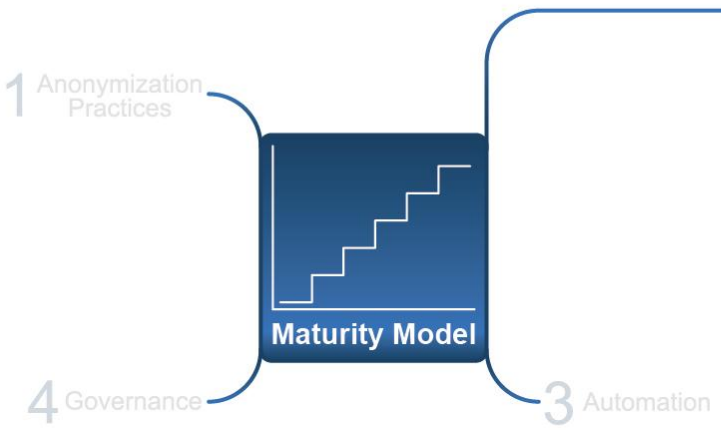
Implementation Framework



1 Anonymization Practices

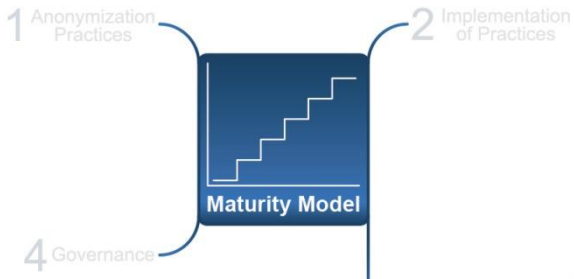
- Techniques
- Standards
- Controls
- Case Studies



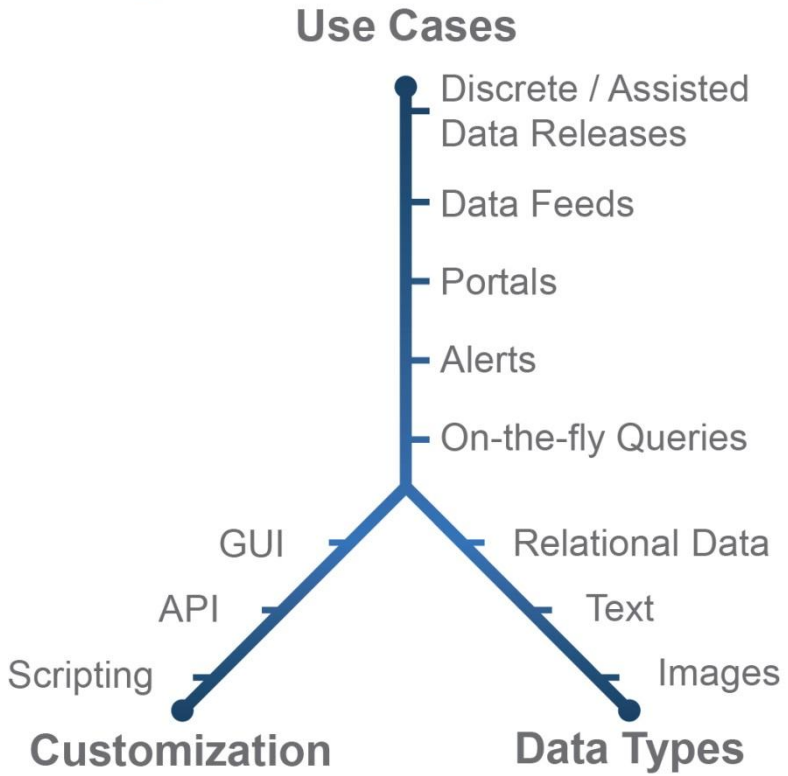


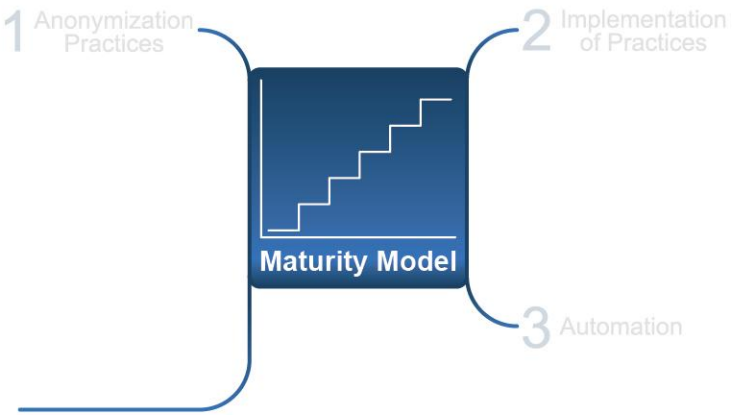
2 Implementation of Practices

- Process
- Roles
- Training
- Organizational Structure
- Performance Measurement



3 Automation





4 Governance

- Legal Oversight
- Audits
- Committees
- Tracking

Contact

kelemam@ehealthinformation.ca

@kelemam

www.ehealthinformation.ca

PRIVACYANALYTICS
Data Anonymization Solutions