TÉCNICO LISBOA

Database Systems
Bases de Dados

W07/H3: Data Privacy and Security

Prof. Paulo Carreira

Data Privacy and Protection

Privacy

The right to control personal information and to be free from unwarranted intrusion

- Includes the right to control to create boundaries, and to determine when, how, and to what extent information about oneself is communicated to others.
- Crucial in maintaining individual autonomy, dignity, and freedom
- Societal structures such as companies often have privacy rights related to their proprietary information, trade secrets, and other sensitive business data.

Privacy is a moral concept that applies to humans, to their interactions, and to their property; it does not apply things or to animals

Data Privacy

Data privacy, or Information privacy, refers to the ability of an organization or individual to control what data is collected, used, and disclosed.

In the information age, companies collect large amounts of personal data

- Companies have the legal and ethical duty to protect the data they collect and to guaranteeing that the fundamental right to data privacy of individuals (and other companies) are not violated
- Companies must ensure that data is collected and used in a lawful and transparent manner, respecting individuals' rights, and not using data for purposes that the individual has not consented to

How do they do this?

Data Protection

Data protection is the implementation of measures to safeguard data against unauthorized access, corruption, or loss of data

Companies have a duty to protect data they collect against unauthorized access

Data Protection is a ‘duty'

Technical Measures

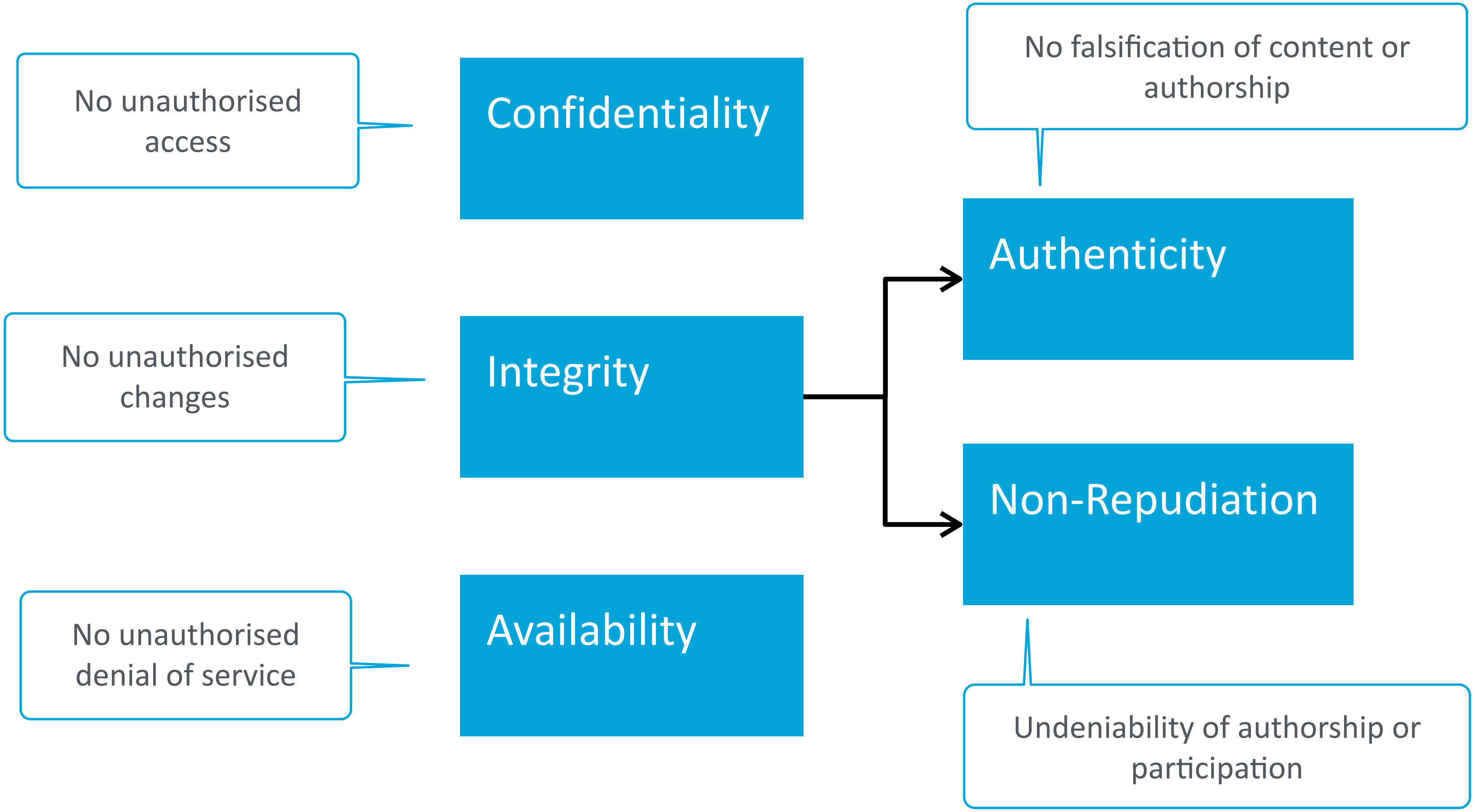
- Firewalls
- Data Loss Prevention
- Encryption
- Backups

Organizational Measures

- Employees Awareness
- Policies and Procedures
- Data minimisation
- Access Control

CIA

CIA Triad



No unauthorised access

Confidentiality

No unauthorised changes

Integrity

No unauthorised denial of service

Availability

No falsification of content or authorship

Authenticity

Non-Repudiation

Undeniability of authorship or participation

CIA *vs* Data Privacy

CIA is enough to ensure Data Protection but insufficient on its own to ensure Data Privacy

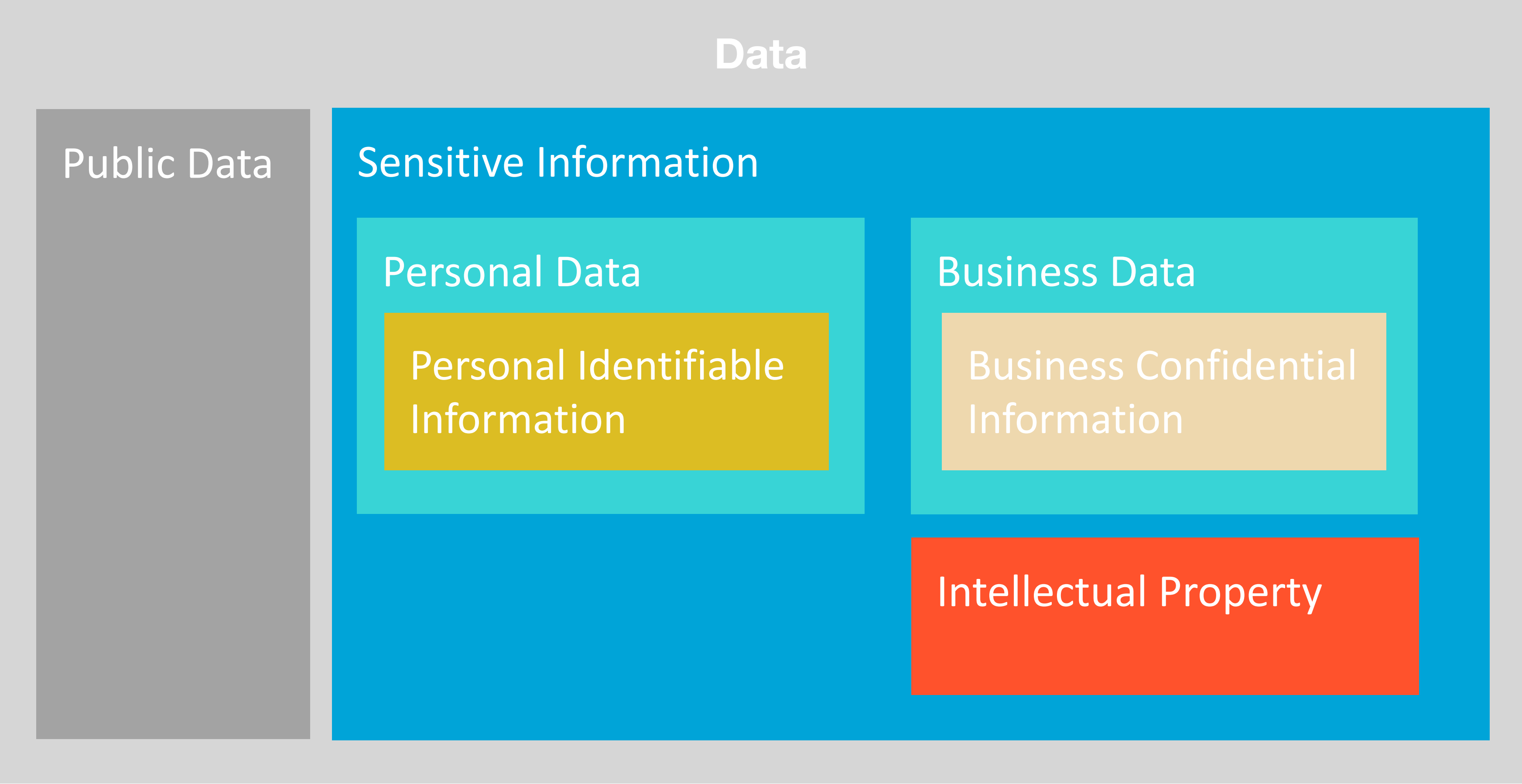
Data privacy also involves notions of consent, transparency, Purpose limitation and Individuals rights

Not guaranteed by C-I-A

Data privacy must also consider how and why data is collected and used, and it must respect individuals' rights regarding their personal data.

Sensitive Information

Any information whose **loss**, **corruption** or **misuse** could cause **harm**, **inconvenience**, **embarrassment**, or **loss of reputation** to the **data subject** or **data owners**



Architecture of Information Systems

States of Data

Data at Rest

Data that is stored on physical or virtual disk drives, tape libraries, removable media

- Unauthorised access
- Alteration
- Exfiltration
- Media destruction

Data in Motion

Data being transferred between locations, or programs, over the internet or through a private network

- Interception
- Man-in-the-middle

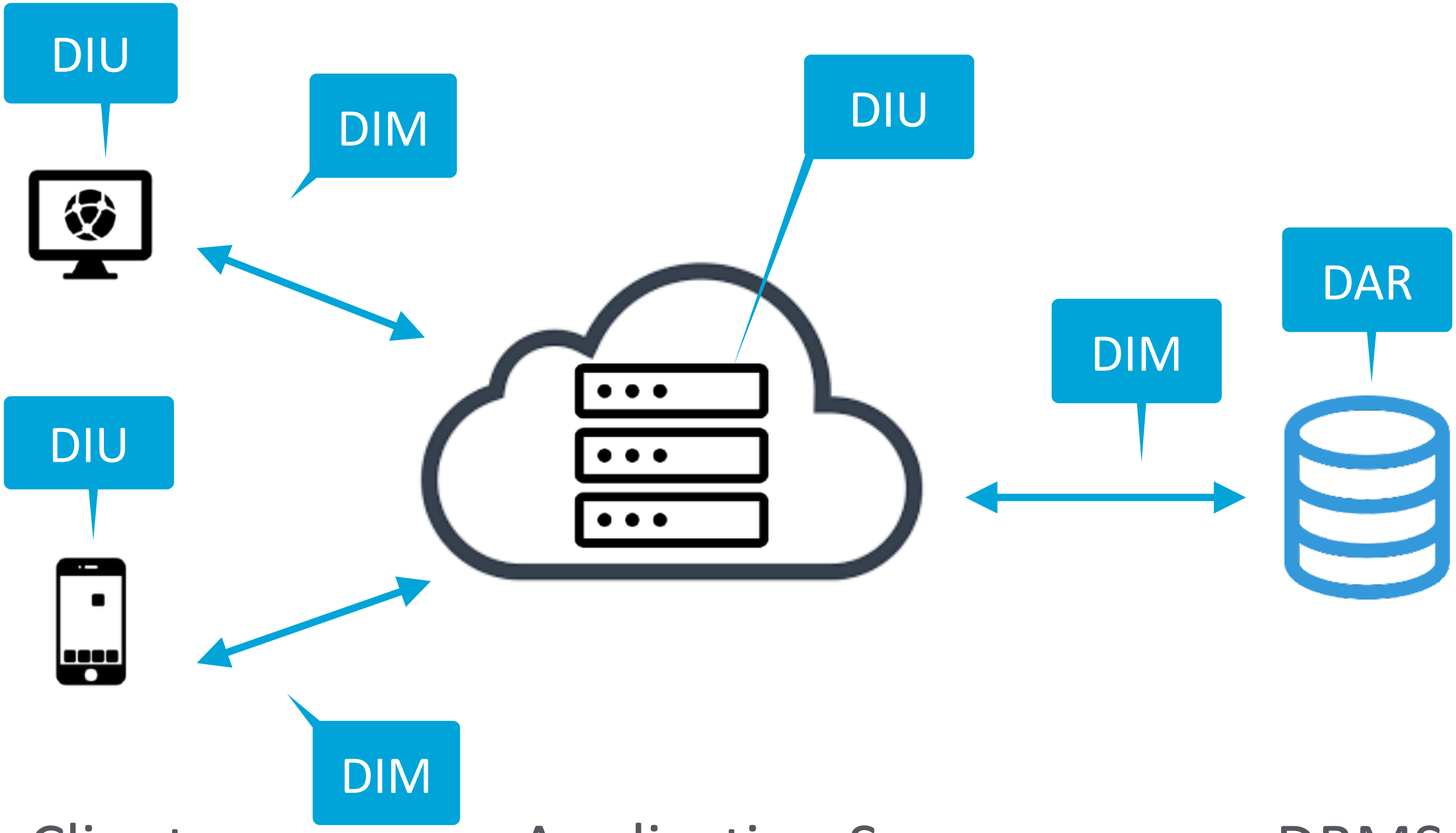
Data in Use

Data currently being processed or used by applications or users

- Trojan
- SQL Injection
- Keylogging
- Memory scrapping

Data is susceptible to distinct threats depending their state

Typical Deployment Architecture

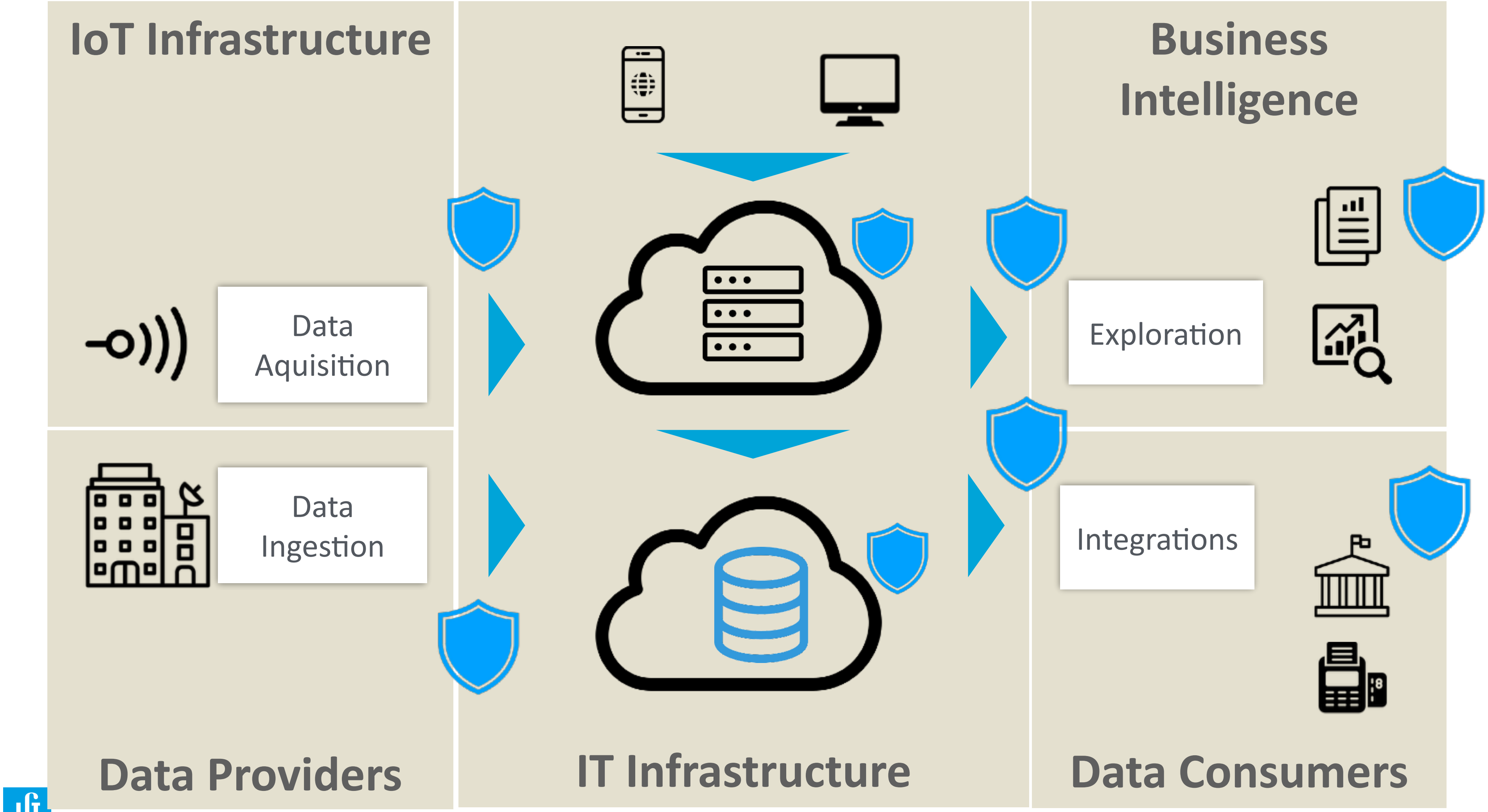


Clients

Application Servers

DBMS

Blueprint of an Information System



IoT Infrastructure

Data Aquisition

Data Ingestion

Data Providers

IT Infrastructure

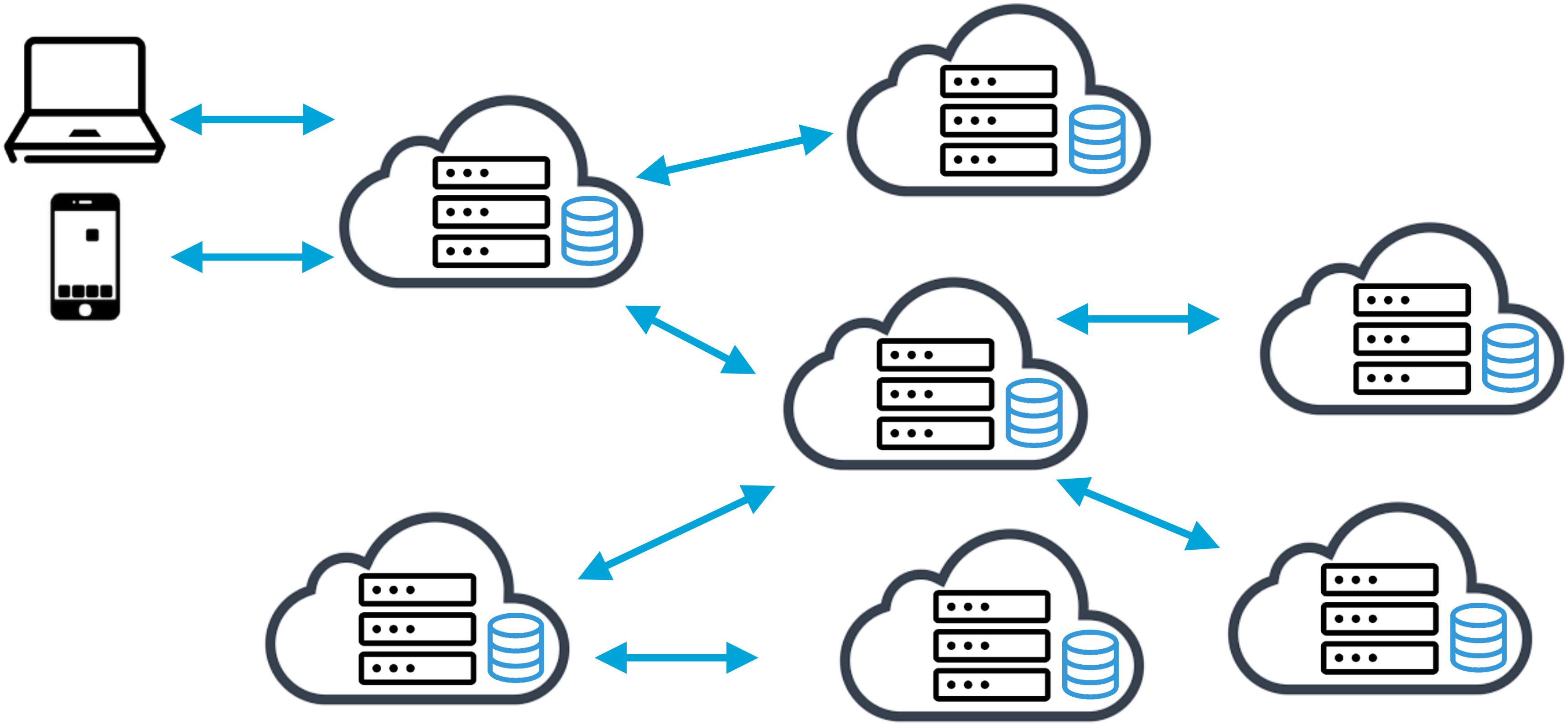
Business Intelligence

Exploration

Integrations

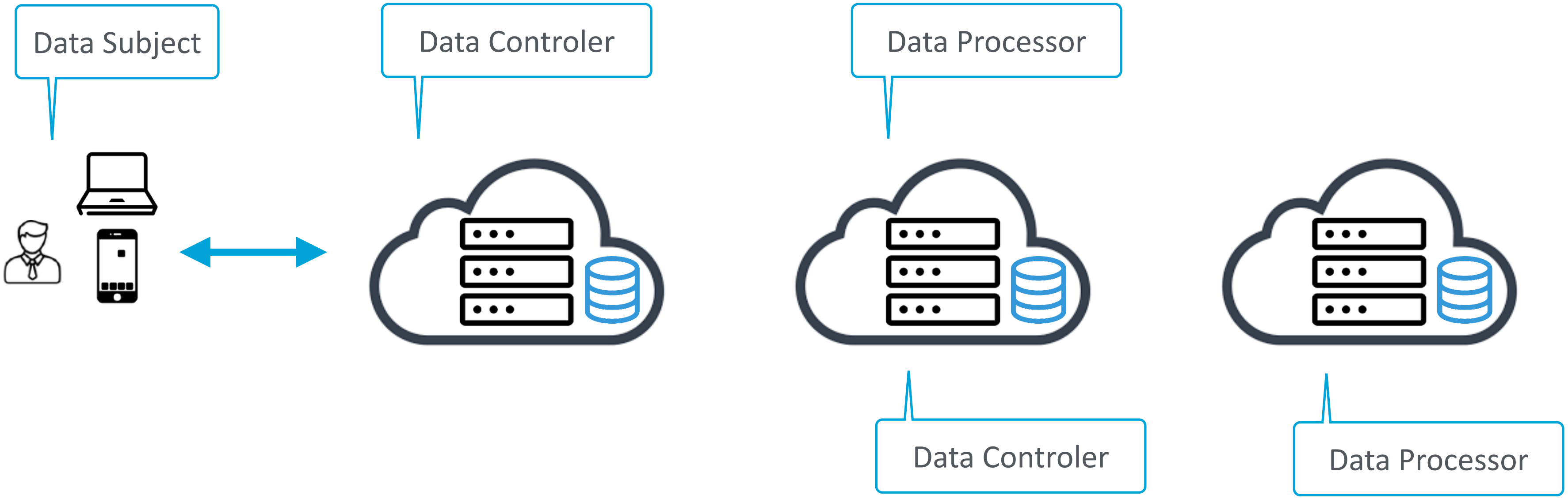
Data Consumers

Typical Deployment Architecture



Actual systems are highly interconnected

Architecture of Responsibility



Architecture of Responsibility

- Data Subjects: Individuals whose personal data is processed and who have rights (to privacy, to access, rectify, erasure, and port their data).

- Data Controllers: They are responsible for ensuring that data processing activities comply with the GDPR, including obtaining consent, implementing security measures, and responding to data subject rights requests.

- Data Processors: Data processors are individuals or organizations that process personal data on behalf of data controllers. They have specific obligations, such as ensuring data security, maintaining records of processing activities, and cooperating with data controllers.

Sources of PII

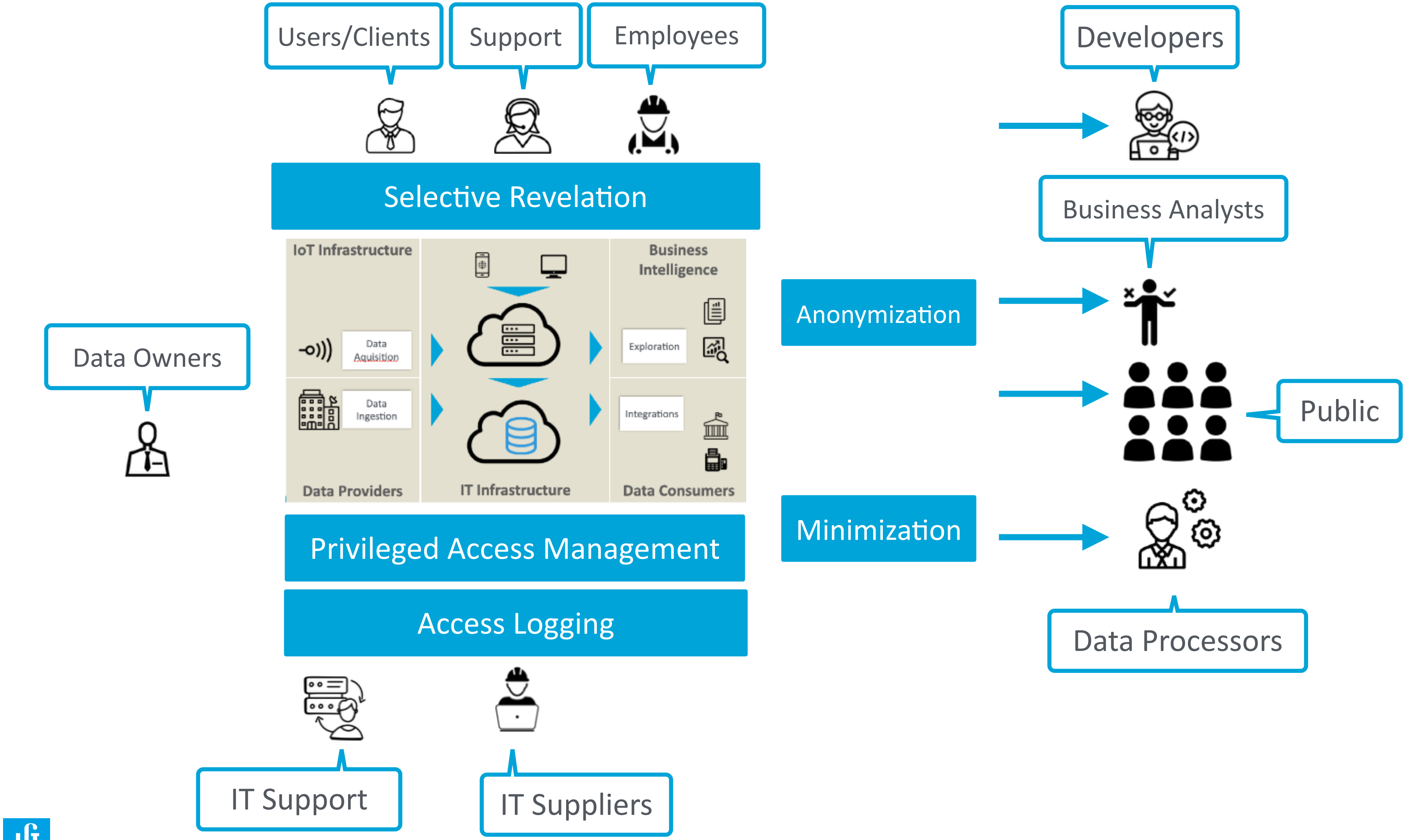
- Government-issued identification: For example, driver’s license, passport, birth certificate, and pension and medical benefits identifiers (e.g., in the United States, Social Security number and Medicare number)
- Contact information: For example, email address, physical address, and telephone numbers
- Online information: For example, Facebook and other social media identifiers, passwords, and

Sources of PII

- Geolocation data: From smartphones, GPS devices, and cameras
- Device address: Such as an IP address of a device connected to the Internet or the media access control (MAC) address of a device connected to a local area networkVerification data: For example, mother's maiden name, pets' and children's names, and high school
- Medical records information: Such as prescriptions, medical records, exams, and medical images
- Biometric and genetic information: Such as fingerprints, retinal scans, and DNA
- Account numbers: Such as bank, insurance, investment, and debit/ credit cards

Data Sharing

Data Stakeholders



Users/Clients

Support

Employees

Selective Revelation

Privileged Access Management

Access Logging

Data Owners

IT Support

IT Suppliers

Developers

Business Analysts

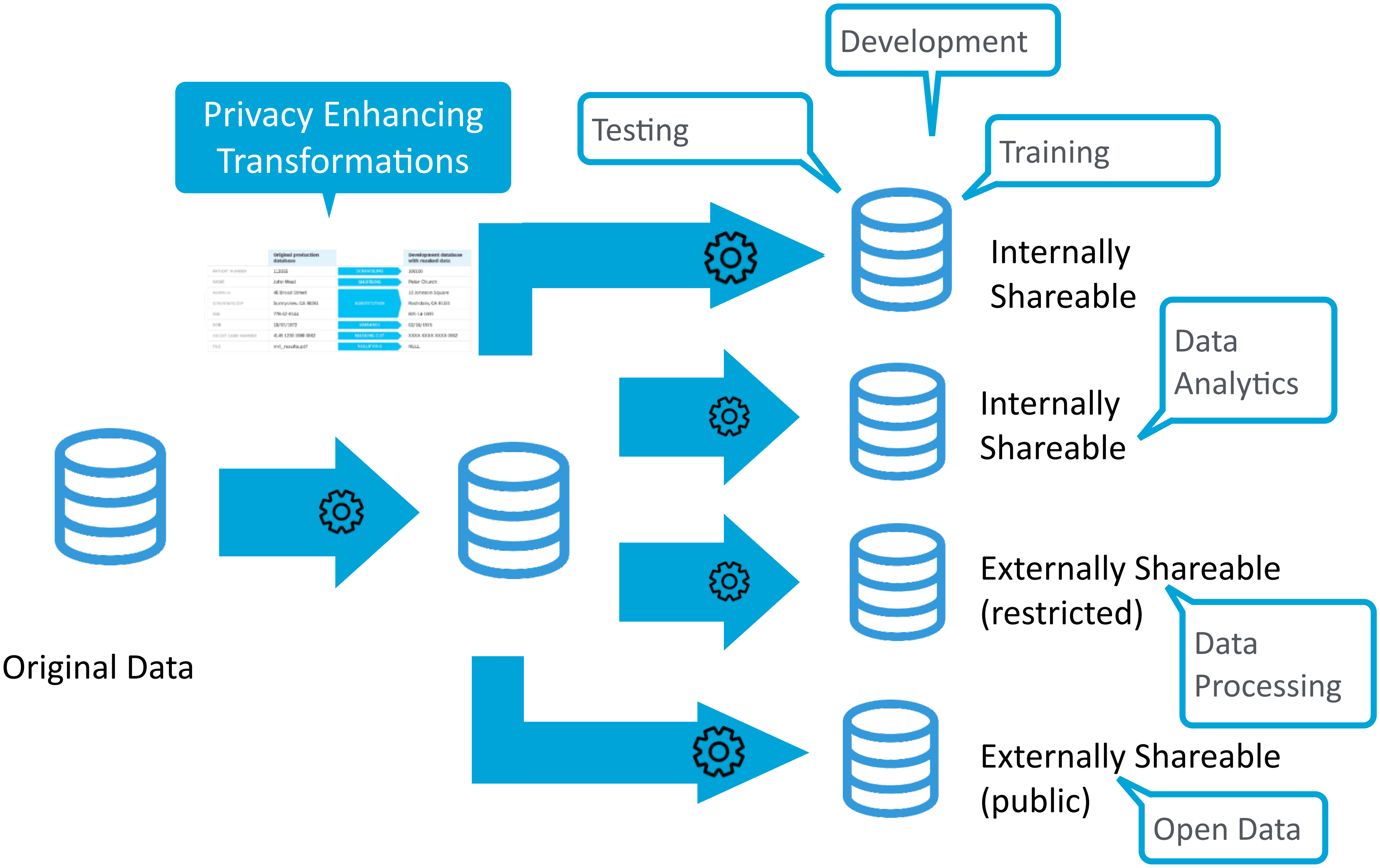
Anonymization

Public

Minimization

Data Processors

Data Privacy Pipeline



Privacy Enhancing Transformations

	Original production database		Development database with masked data
PATIENT NUMBER	113355	SCRAMBLING	300100
NAME	John West	SHUFFLING	Peter Church
ADDRESS	45 Broad Street	SUBSTITUTION	12 Johnson Square
CITY/STATE/ZIP	Sunnyview, CA 90261		Rochdale, CA 81331
SSN	779-62-8144		805-14-1995
DOB	18/07/1972	VARIANCE	02/18/1975
CREDIT CARD NUMBER	4548 1230 3000 0062	MASKING OUT	XXXX XXXX XXXX 0062
FILE	mri_results.pdf	NULLIFYING	NULL

Original Data

Testing

Development

Training

Internally Shareable

Internally Shareable

Data Analytics

Externally Shareable (restricted)

Data Processing

Externally Shareable (public)

Open Data

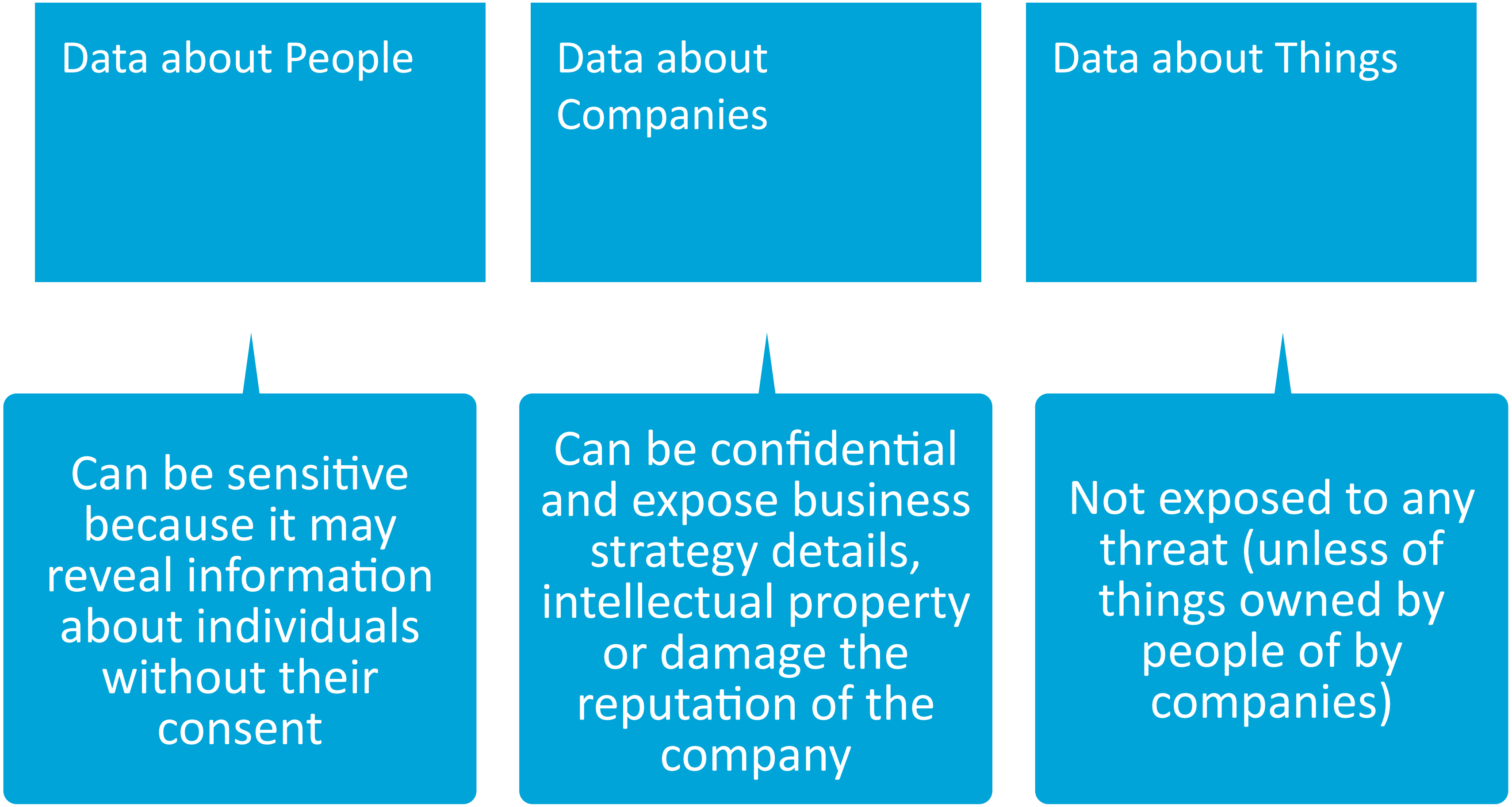
Data Sharing

Data must be share to realize its value or to be processed by a third party

Since data sharing is irreversible

How do we guarantee that only the right data is shared

Categories of Data



Data about People

Can be sensitive because it may reveal information about individuals without their consent

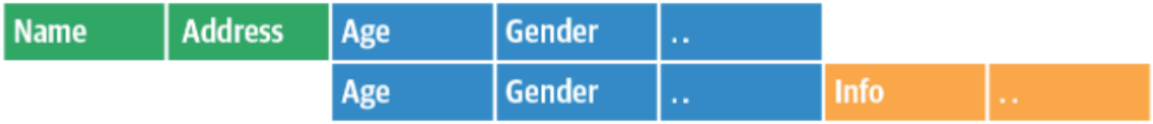
Data about Companies

Can be confidential and expose business strategy details, intellectual property or damage the reputation of the company

Data about Things

Not exposed to any threat (unless of things owned by people of by companies)

Personal Data



Name	Address	Age	Gender	..		
		Age	Gender	..	Info	..

Personal data is information about an identifiable individual and consist of attributes related to individuals

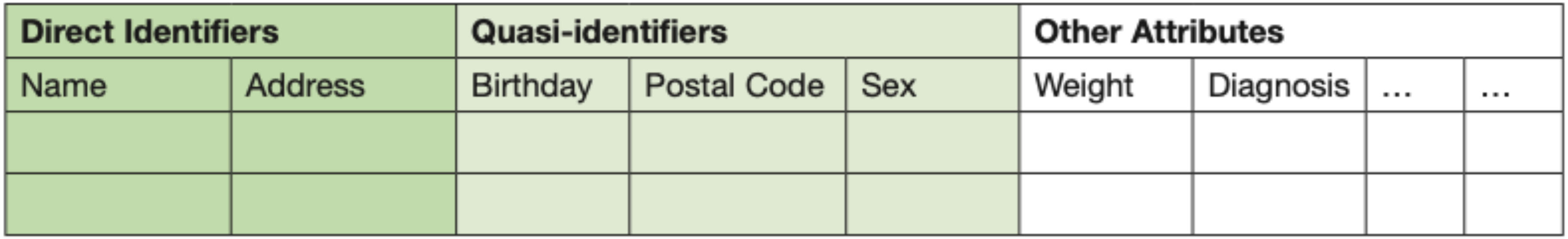
With respect to privacy of personal data, attributes can be classified as Direct Identifiers, Indirect Identifiers, and Other

Classification of Attributes

- Direct Identifiers: Direct identifiers are attribute values that allow an individual to be identified and also allow other data to be linked to that individual. Examples of identifying variables include name, email address, home address, telephone number, health insurance number, and Social Security number.

- Indirect Identifiers (or quasi-identifiers): Indirect (or quasi-) identifiers are attributes data that by themselves doe not identify a specific individual, but that can be can be used in combination to identify an individual, or can be linked with other dataset present of future to identify an individual.

Direct vs. Indirect Identifiers

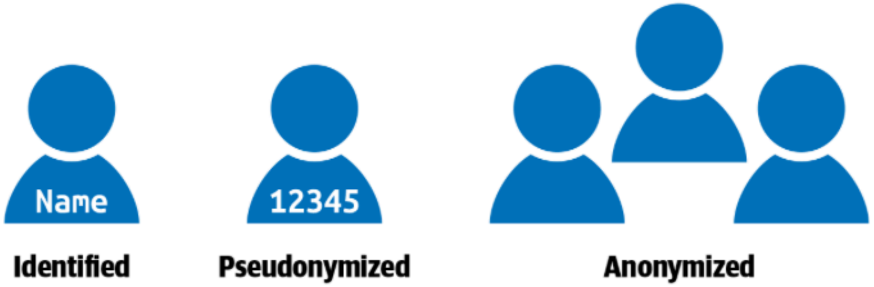


Direct Identifiers		Quasi-identifiers			Other Attributes			
Name	Address	Birthday	Postal Code	Sex	Weight	Diagnosis	…	…

The distinction matters, because you may conceal direct identifiers but share quasi identifiers and have the illusion that your data is anonymous. This was the mistake of AOL.

Identifiability

Identifiability Spectrum



- Identifiable data is data that can be directly associated to an individual.
- Pseudonymized data …
- Anonymized data or (de-identified data) is data that cannot be reasonably linked to an individual.

Disclosure Risks

Disclosure Risks

The possibility that private or confidential information can be revealed is technically known as disclosure risk.

- (Re-) Identification
- Addressing
- Attribution
- Linking
- Inference

Identifiability

Re-identification refers to the ability to correctly assign a record to an identifiable (with a high probability)

For example, an adversary might determine that the record with the key 123ABC belongs to Mary Jones; this reveals that all the information in that record is associated with Mary Jones.

Anonymization standards that exist today would typically only address this specific issue of protecting against identity disclosure

Addressability

Addressability refers to a the situation where a pseudonym that can be used to target (or “address”) a specific individual (not necessarily an identifiable individual).

For example, an advertiser could send the pseudonym and the advertisement to an ISP that then links the pseudonym to a specific device ID and sends that advertisement to that device. The ISP already knows the identity of the consumer, and the advertiser never gets to know the identity of the consumer. In that case the pseudonym is addressable but not identifiable to the advertiser.

Linkability

Linkability refers to the ability to link records that belong to the same individual together (not necessarily an identifiable individual)

Imagine linking pseudanonymized dataset with incidences of medical conditions that includes sex, birthdate and postal code, and with a pseudanonymized dataset of income with the same attributes.

Attribution

Attribution refer to ability obtains one or more attributes for a specific individual by associating it with group information.

For example, if a hospital releases information showing that all current female patients aged 56 to 60 have cancer, and if Alice Smith is a 56-year-old female who is known to be an inpatient at the hospital, then Alice Smith's diagnosis is revealed, even though her individual de-identified medical records cannot be distinguished from the others.

Attribution is specific form Linkability

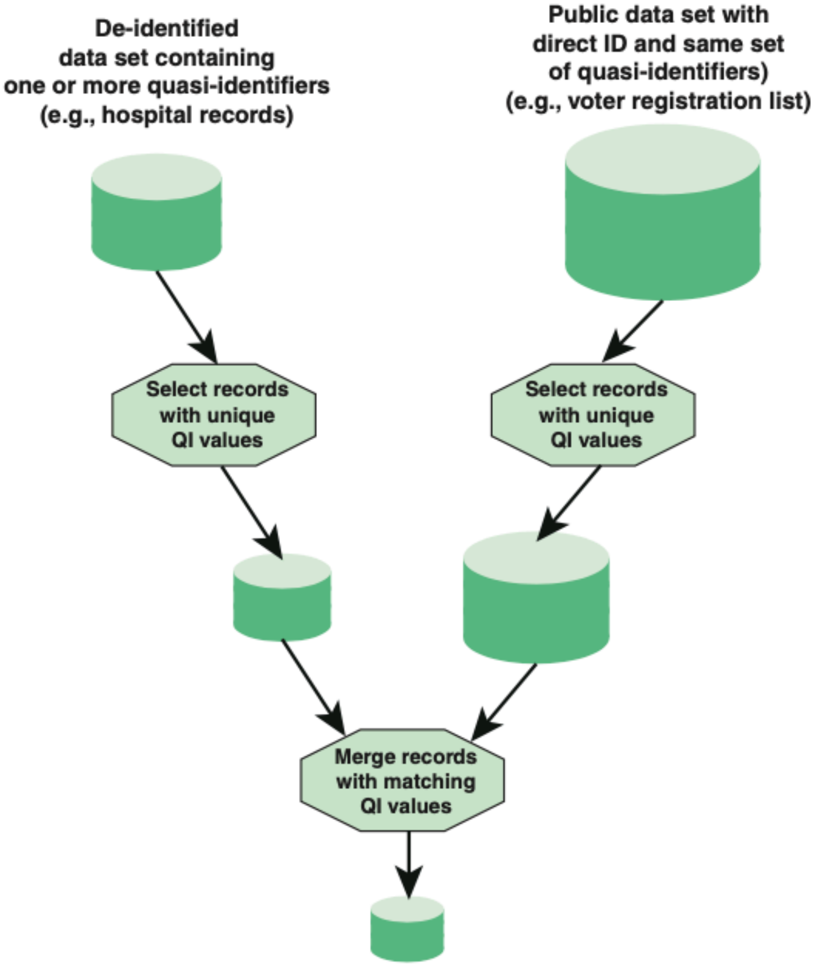
Inference

Inference refers to the possibility of learn something new about an individual or group in the data more accurately than would have otherwise been possible using a series of reasoning steps

As a particular instance, the data may show a high correlation between income and purchase price of a home. Because the purchase price of a home is typically public information, a third party might use this information to infer the income of a data subject.

Privacy Attacks

Linkage Attacks



Privacy Enahncing Technique

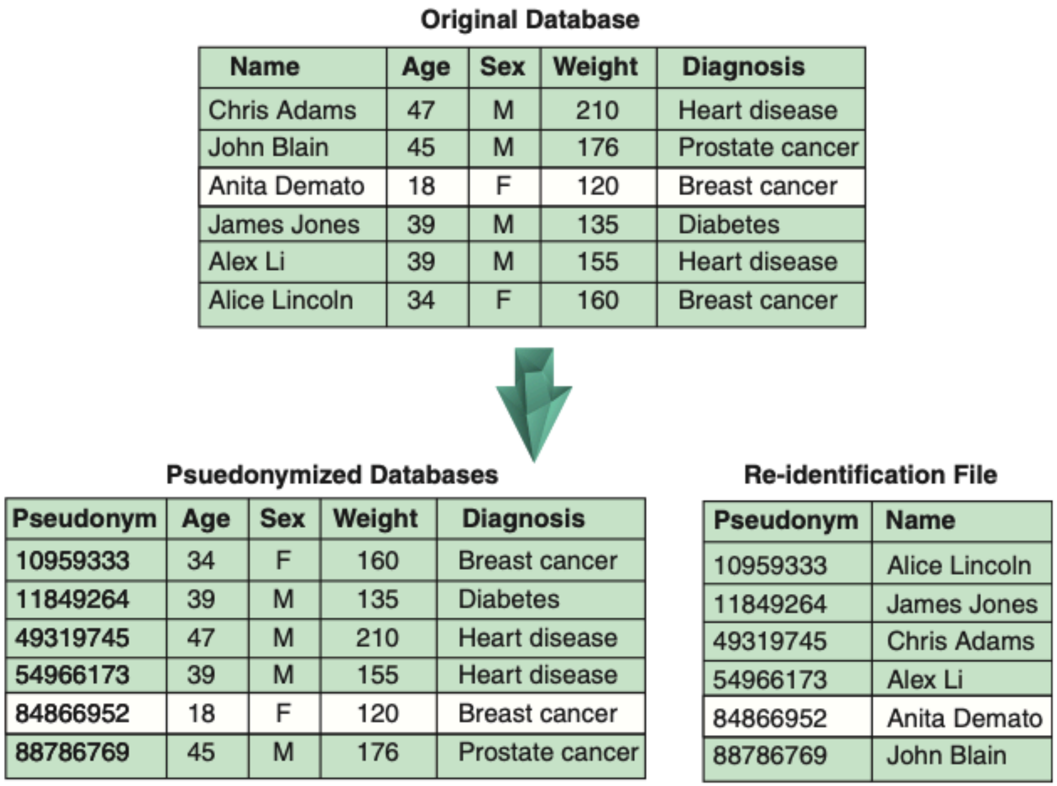
Privacy Enhancing Techniques

Privacy enhancing techniques apply transformation to data so that removing enough direct identifiers and quasi-identifiers makes the identification of individuals harder.

- Pseudonymization
- Data Masking
- Generalization
- Differential Privacy
- Synthetic Data

Pseudonymization

Pseudonymization de-identifies data values by substituting private identifiers with fake identifiers or pseudonyms.



Original Database

Name	Age	Sex	Weight	Diagnosis
Chris Adams	47	M	210	Heart disease
John Blain	45	M	176	Prostate cancer
Anita Demato	18	F	120	Breast cancer
James Jones	39	M	135	Diabetes
Alex Li	39	M	155	Heart disease
Alice Lincoln	34	F	160	Breast cancer

Psuedonymized Databases

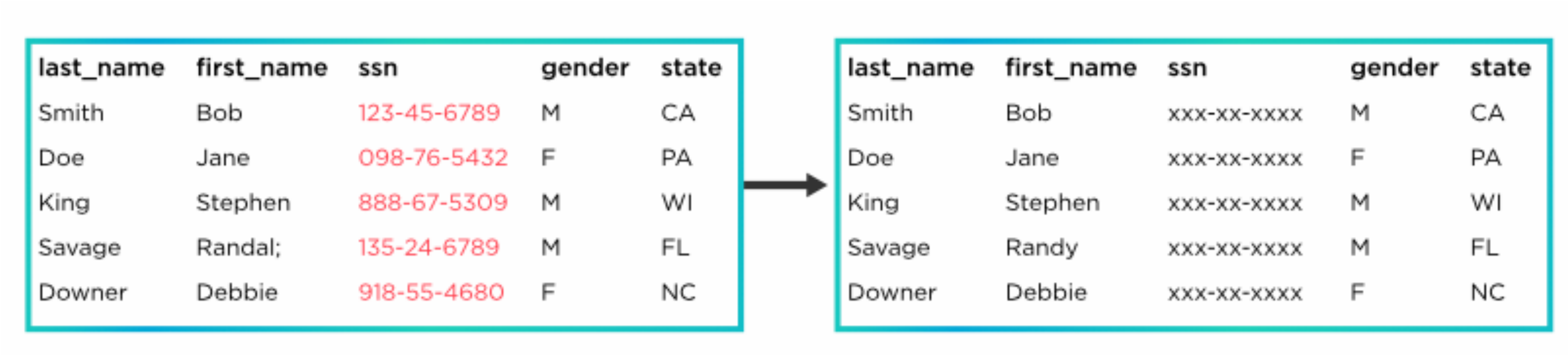
Pseudonym	Age	Sex	Weight	Diagnosis
10959333	34	F	160	Breast cancer
11849264	39	M	135	Diabetes
49319745	47	M	210	Heart disease
54966173	39	M	155	Heart disease
84866952	18	F	120	Breast cancer
88786769	45	M	176	Prostate cancer

Re-identification File

Pseudonym	Name
10959333	Alice Lincoln
11849264	James Jones
49319745	Chris Adams
54966173	Alex Li
84866952	Anita Demato
88786769	John Blain

Data Masking

It is the process of hiding values in a data set so that the data is still accessible, but the original values cannot be reversed



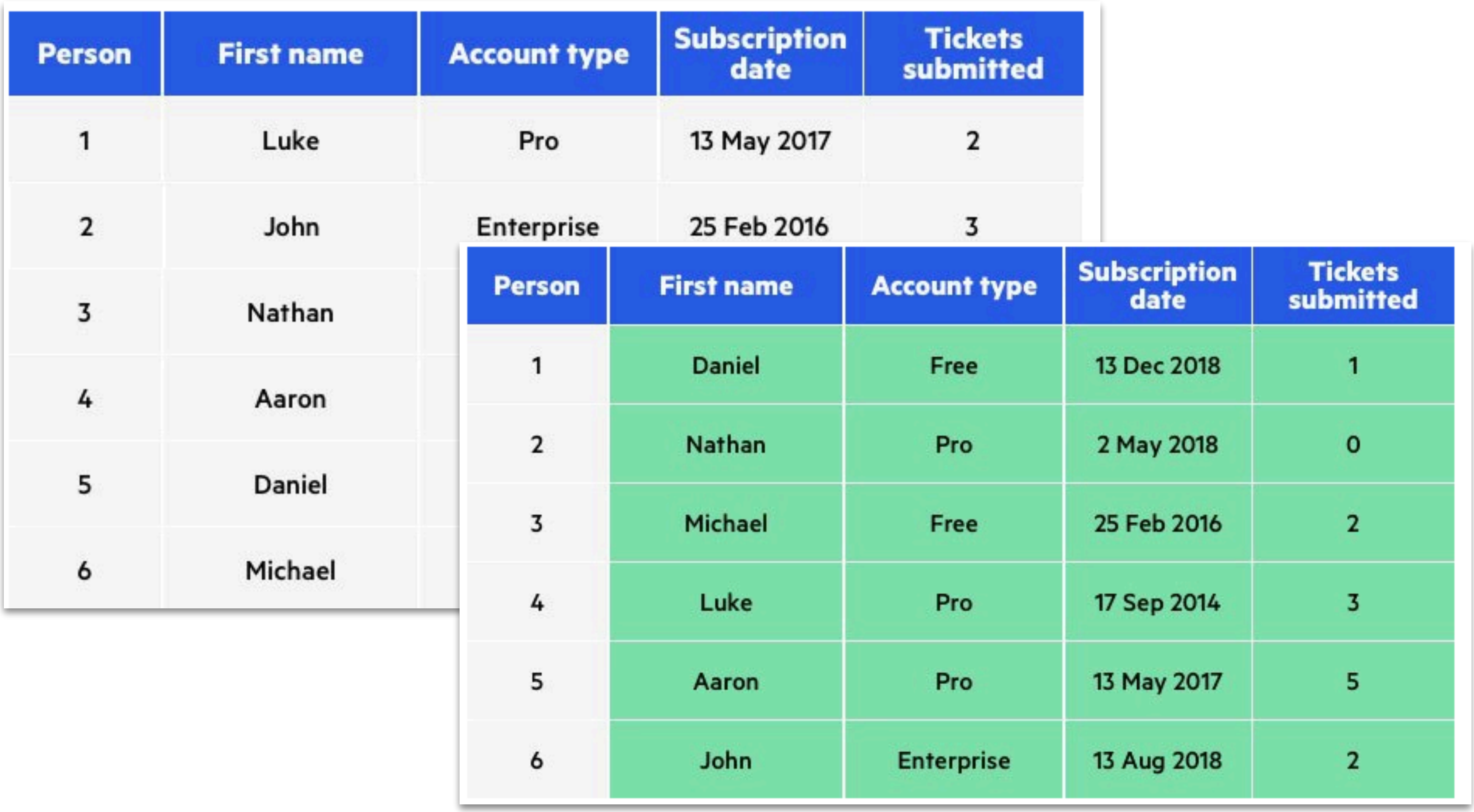
last_name	first_name	ssn	gender	state
Smith	Bob	123-45-6789	M	CA
Doe	Jane	098-76-5432	F	PA
King	Stephen	888-67-5309	M	WI
Savage	Randal;	135-24-6789	M	FL
Downer	Debbie	918-55-4680	F	NC

last_name	first_name	ssn	gender	state
Smith	Bob	xxx-xx-xxxx	M	CA
Doe	Jane	xxx-xx-xxxx	F	PA
King	Stephen	xxx-xx-xxxx	M	WI
Savage	Randy	xxx-xx-xxxx	M	FL
Downer	Debbie	xxx-xx-xxxx	F	NC

Data Masking Techniques

- Substitution
- Scrambling
- Suppression or Redaction
- Nulling
- Encryption
- Hashing
- Perturbation (or randomisation)
- Shuffling (or swapping)

Shuffling

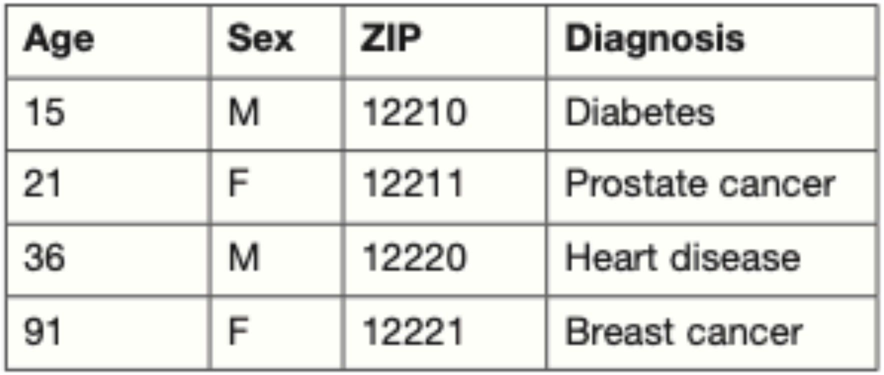


Person	First name	Account type	Subscription date	Tickets submitted
1	Luke	Pro	13 May 2017	2
2	John	Enterprise	25 Feb 2016	3
3	Nathan			
4	Aaron			
5	Daniel			
6	Michael			

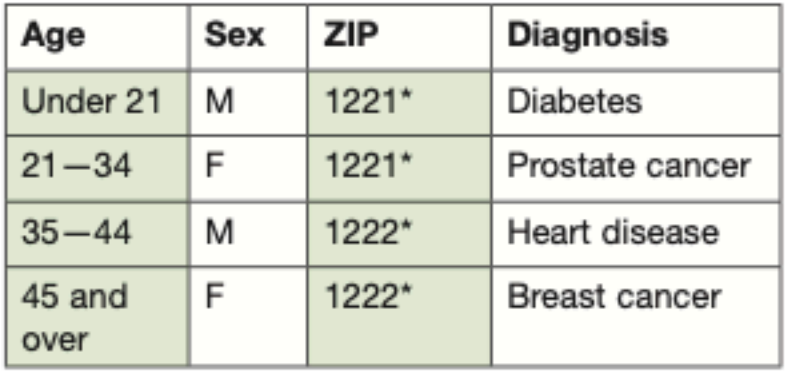
Person	First name	Account type	Subscription date	Tickets submitted
1	Daniel	Free	13 Dec 2018	1
2	Nathan	Pro	2 May 2018	0
3	Michael	Free	25 Feb 2016	2
4	Luke	Pro	17 Sep 2014	3
5	Aaron	Pro	13 May 2017	5
6	John	Enterprise	13 Aug 2018	2

Data Generalization

The process of deliberately decreasing the precision of a dataset to make it less identifiable



Age	Sex	ZIP	Diagnosis
15	M	12210	Diabetes
21	F	12211	Prostate cancer
36	M	12220	Heart disease
91	F	12221	Breast cancer

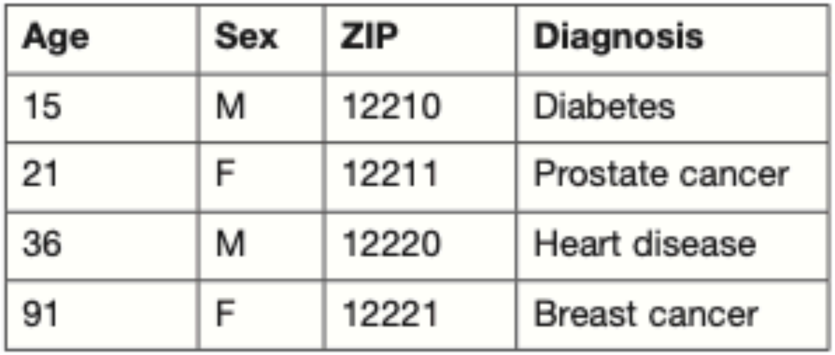


Age	Sex	ZIP	Diagnosis
Under 21	M	1221*	Diabetes
21—34	F	1221*	Prostate cancer
35—44	M	1222*	Heart disease
45 and over	F	1222*	Breast cancer

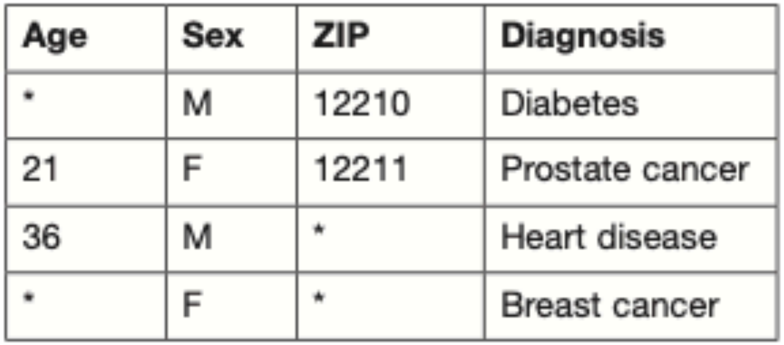
Data Generalization Techniques

- Blurring
- Averaging
- Tokenization
- Bucketing
- Sub-sampling

Example of Suppression

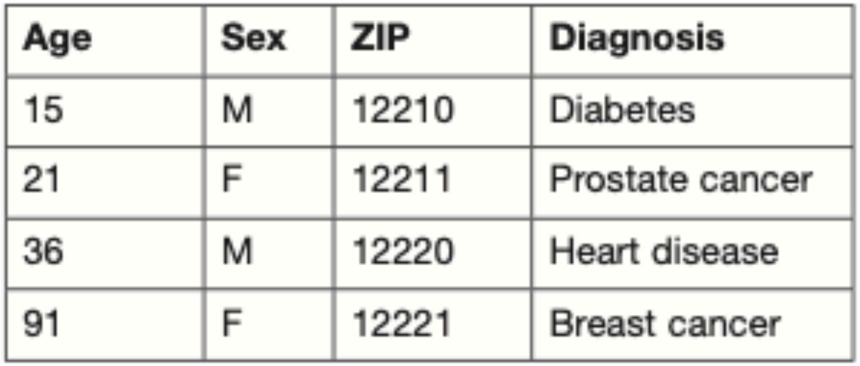


Age	Sex	ZIP	Diagnosis
15	M	12210	Diabetes
21	F	12211	Prostate cancer
36	M	12220	Heart disease
91	F	12221	Breast cancer

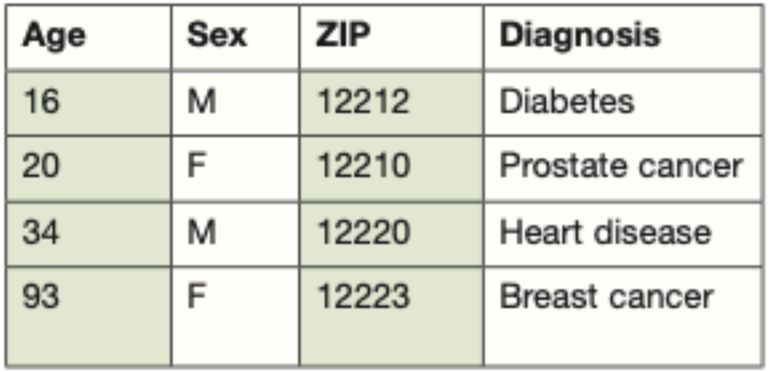


Age	Sex	ZIP	Diagnosis
*	M	12210	Diabetes
21	F	12211	Prostate cancer
36	M	*	Heart disease
*	F	*	Breast cancer

Example of Perturbation

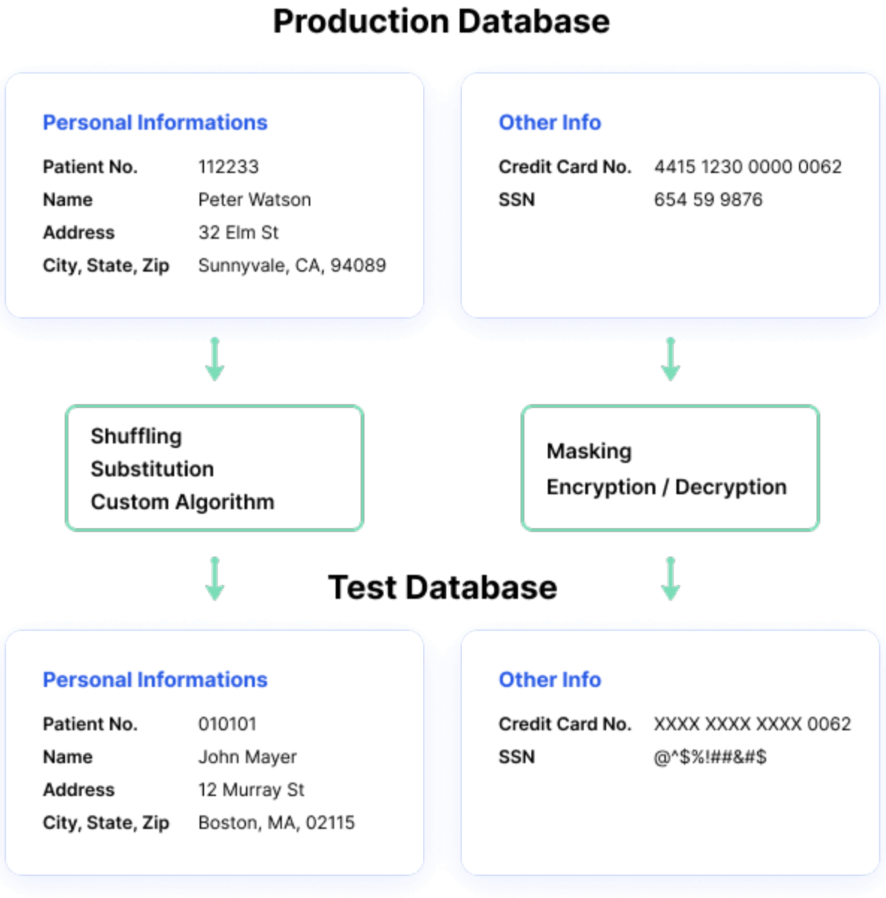


Age	Sex	ZIP	Diagnosis
15	M	12210	Diabetes
21	F	12211	Prostate cancer
36	M	12220	Heart disease
91	F	12221	Breast cancer



Age	Sex	ZIP	Diagnosis
16	M	12212	Diabetes
20	F	12210	Prostate cancer
34	M	12220	Heart disease
93	F	12223	Breast cancer

Example of Distinct Techniques



Production Database

Personal Informations

Patient No.	112233
Name	Peter Watson
Address	32 Elm St
City, State, Zip	Sunnyvale, CA, 94089

Other Info

Credit Card No.	4415 1230 0000 0062
SSN	654 59 9876

Shuffling
Substitution
Custom Algorithm

Masking
Encryption / Decryption

Test Database

Personal Informations

Patient No.	010101
Name	John Mayer
Address	12 Murray St
City, State, Zip	Boston, MA, 02115

Other Info

Credit Card No.	XXXX XXXX XXXX 0062
SSN	@^$%!##&#$

Data Residency, Transfer, and Sovereignty

Data Sovereignty

Data sovereignty refers to the concept that information or data is subject to the laws and governance structures of the country in which it is collected or processed

If an organization collects data in country A but processes or stores it in country B, it needs to comply with the data protection and privacy laws of both countries.

- Anonymized Data: Can be transferred
- Pseudonymized and Encrypted Data: cannot be transferred because it can be reversed

EU companies cannot transfer data of European citizens to datacenters in jurisdictions that do not offer the same levels of protection

Data Residency

Data residency refers to the geographical location where an organization's data is stored on premises or on the cloud

Since countries have different laws and regulations about data privacy, protection, and how data can be accessed or transferred. Data residency is a significant concern since these laws can influence the selection of cloud service providers and the location of their data centers.

EU companies cannot transfer data of European citizens to datacenters in jurisdictions that do not offer the same levels of protection

TÉCNICO LISBOA