PRIVACY AND DATA PROTECTION

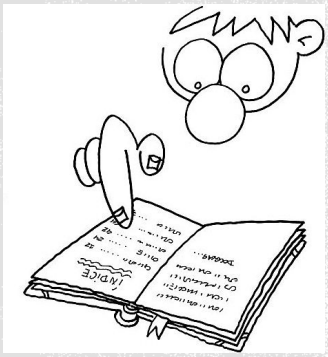
Introduction to Computer and Network Security

Silvio Ranise [silvio.ranise@unitn.it or *ranise@fbk.eu*]

UNIVERSITÀ DI TRENTO

FONDAZIONE BRUNO KESSLER

CONTENTS



2 ON PRIVACY

PRIVACY: DEFINITION 1

- Informational self-determination
- Everyone gets to **control** information about herself/himself
- This means a lot…
  - Who gets to see what (related to access control)
  - Who gets to use what (related to usage control)
  - What they can use it for (related to purpose control)
  - Who they can give it to (related to data transfer control)
  - …

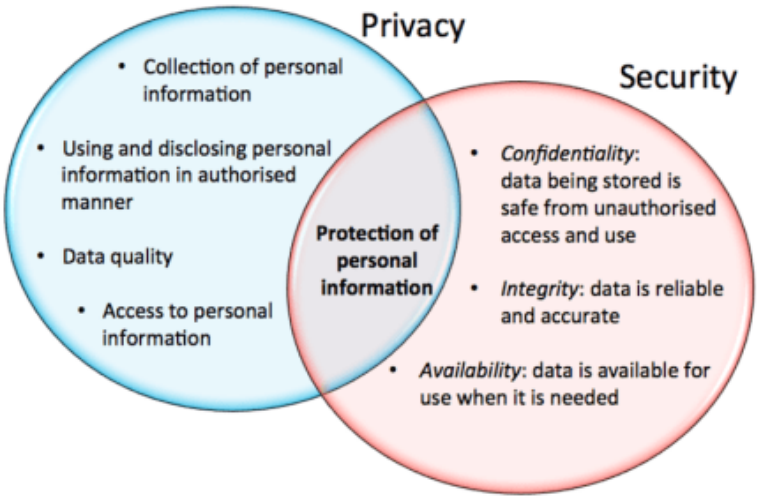
3

PRIVACY: DEFINITION 2

- High-level of difficulty in **correlating** data/actions
  - (in particular, those performed on-line)
- This means that there is some hope to remain anonymous*!*
- This is very difficult to achieve for many reasons…

4

SECURITY VS PRIVACY



5

Type	Property	Description	Threat
Hard privacy	Unlinkability	Hiding the link between two or more actions, identities, and pieces of information.	**L**inkability
	Anonymity	Hiding the link between an identity and an action or a piece of information	**I**dentifiability
	Plausible deniability	Ability to deny having performed an action that other parties can neither confirm nor contradict	**N**on-repudiation
	Undetectability and unobservability	Hiding the user's actvities	**D**etectability
Security	Confidentiality	Hiding the data content or controlled release of data content	**D**isclosure of information
Soft Privacy	Content awareness	User's consciousness regarding his own data	**U**nawareness
	Policy and consent compliance	Data controller to inform the data subject about the system's privacy policy, or allow the data subject to specify consents in compliance with legislation	**N**on compliance

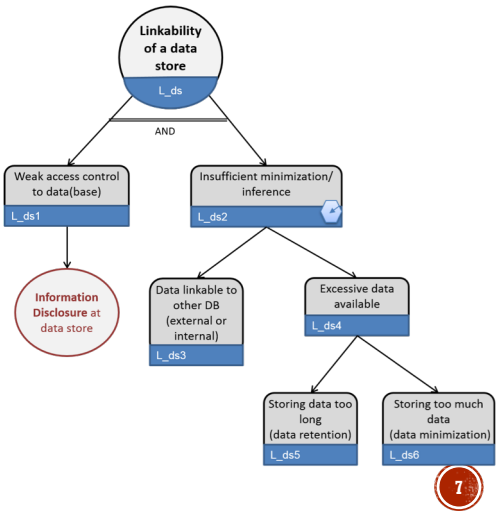
LINDDUN

More info available at https://linddun.org/

6

LINDDUN CHARACTERIZATION OF PRIVACY

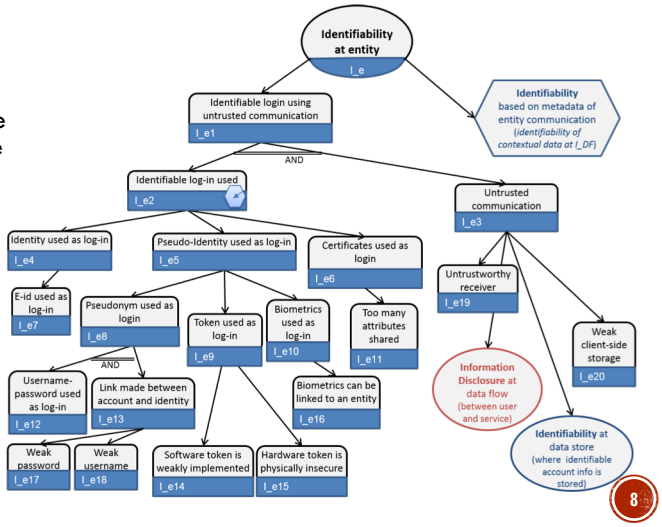
- **Linkability**
  - Being able to sufficiently distinguish whether 2 **items of interest (ioi)** are linked or not, even without knowing the actual identity of the subject of the linkable ioi
  - Examples
    - web page visits by the same user
    - entries in databases related to same person
  - Consequences
    - **Identifiability** when too much linkable information is combined
    - **Inference**: when "group data" is linkable, this can lead to societal harm, like discrimination (e.g. if an insurance company knows that people who live in a certain area get sick more often, they might increase their insurance cost for that target group)



7

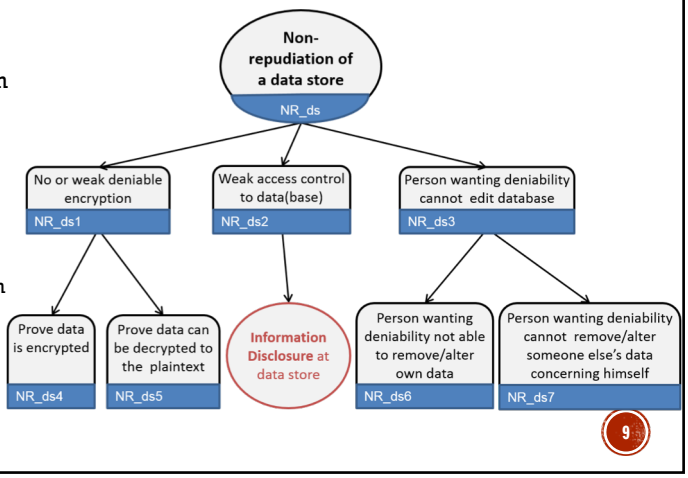
LINDDUN CHARACTERIZATION OF PRIVACY

- **Identifiability**
  - Being able to sufficiently identify the subject within a set of subjects (i.e. the anonymity set). Not being able to hide the link between the identity and the ioi (an action or piece of information).
  - Examples
    - identifying the reader of a web page
    - the person to whom an entry in a database relates
  - Consequences
    - Severe privacy violations (when subject assumes he is anonymous)



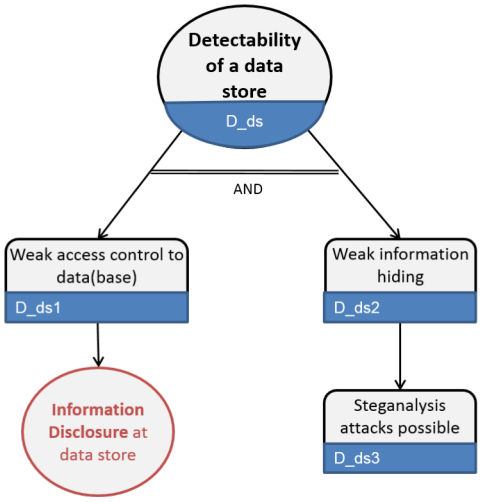
LINDDUN CHARACTERIZATION OF PRIVACY

- **Non-repudiation**
  - Not being able to deny a claim. The attacker can thus prove a user knows, has done or has said something. He can gather evidence to counter the claims of the repudiating party.
  - Examples
    - Problems in on-line voting/whistleblowing
  - Consequences
    - when a person is not able to repudiate an action or piece of information, he can be held accountable (e.g. a whistleblower can be prosecuted)



LINDDUN CHARACTERIZATION OF PRIVACY

- **Detectability**
  - Being able to sufficiently distinguish whether an ioi exists or not. Detectability concerns iois of which the content is not known (to the attacker)
  - Examples
    - Existence of a pregnancy test
  - Consequences
    - By detecting whether an ioi exists, one can deduce certain information, even without actually having access to that information



10

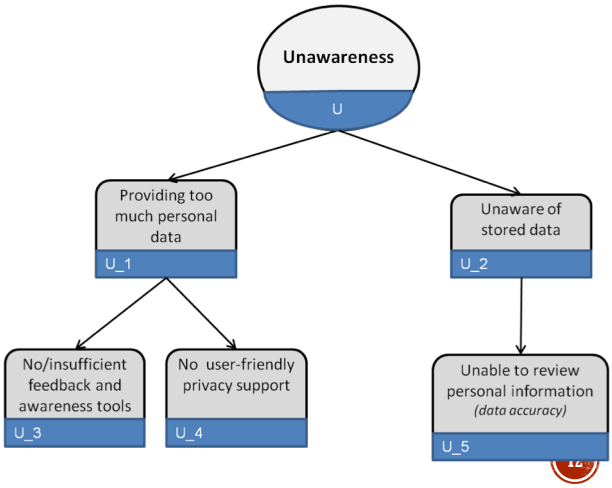
LINDDUN CHARACTERIZATION OF PRIVACY

- **Disclosure of information**
  - Related to (violation of) confidentiality and thus more to security

11

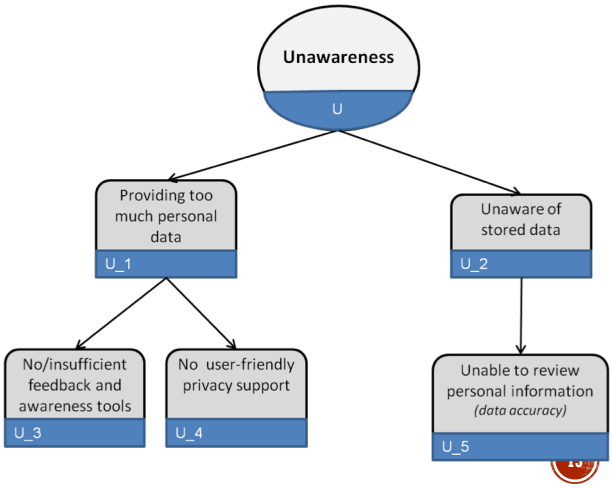
LINDDUN CHARACTERIZATION OF PRIVACY

- **Unawareness**
  - Being unaware of the consequences of sharing information. Often users are not aware of the impact of sharing data. This can be data shared to friends on facebook, but also personal information shared with other services (i.e. loyalty cards, online services, …)
  - Ideally, all users (data providers) should be clearly informed and educated of the consequences of sharing data using (online) services.
  - Consequences
    - The more information is available, the easier it can be linked (and identified)

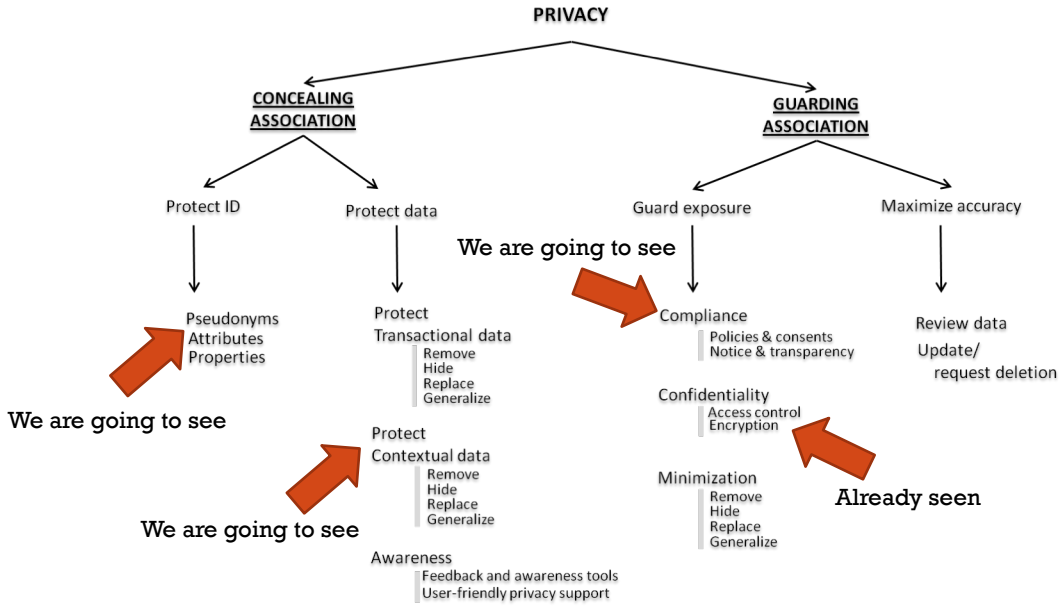


LINDDUN CHARACTERIZATION OF PRIVACY

- **Non-compliance**
  - Not being compliant with legislation, regulations, and corporate policies
  - Consequences
    - Fines (when violating legislation, or not adhering to the communicated corporate policies)
    - Loss of image, credibility, etc.



LINDDUN: MITIGATIONS



15

ON ANONYMIZATION

DATA ANONYMIZATION

- Remove **Personally Identifying Information** (PII)
  - Examples: Name, Social Security number, phone number, email, address…
- **Problem**: PII has no technical meaning
- It is defined in disclosure notification laws
  - If certain information is lost, consumer must be notified
- In privacy breaches, any information can be personally identifying
  - Example: Netflix Prize dataset…

NETFLIX DATA LEAK

Problem
- data can be de-anonymized by considering other sources of knowledge!!!

- Netflix launched the Netflix prize
  - Improve movie recommendations algorithm of at least 10% of the used algorithm and get a $1 million prize (if best solution)
- For the first edition in 2006, Netflix released 100 million supposedly anonymized movie ratings. Each included
  - a unique subscriber ID, the movie title, year of release and the date on which the subscriber rated the movie
- Just 16 days later, two University of Texas researchers announced that they had identified some of the Netflix users in the data set
- How? In some cases, **researchers were able to identify targets by matching their Netflix reviews with data from other sites like IMDb**
- Netflix continued the contest and named a $1 million winner. But when Netflix tried to launch another contest in 2009 -- with subscriber data including gender, zip code, and age -- a woman (a lesbian mother who is not open about her sexual orientation) filed a suit as Jane Doe saying that
  - *"To some, renting a movie such as Brokeback Mountain or even The Passion of the Christ can be a personal issue that they would not want published to the world"*
- After months of back-and-forth, Netflix called off the second contest and settled the lawsuit

SOME DETAILS ABOUT LINKAGE ATTACKS (1)

Massachusetts hospital discharge dataset

Medical Data Released as Anonymous

SSN	Name	...city	Date Of Birth	Sex	ZIP	Marital Status	Problem
			09/27/64	female	02139	divorced	hypertension
			09/30/64	female	02139	divorced	obesity
		asian	04/18/64	male	02139	married	chest pain
		asian	04/15/64	male	02139	married	obesity
		black	03/13/63	male	02138	married	hypertension
		black	03/18/63	male	02138	married	shortness of breath
		black	09/13/64	female	02141	married	shortness of breath
		black	09/07/64	female	02141	married	obesity
		white	05/14/61	male	02138	single	chest pain
		white	05/08/61	male	02138	single	obesity
		white	09/15/61	female	02142	widow	shortness of breath

Voter List

Name	Address	City	ZIP	DOB	Sex	Party	.....
.....	.....	.....	.....	.....	.....	.....	
.....	.....	.....	.....	.....	.....	.....	
Sue J. Carlson	1459 Main St.	Cambridge	*02142*	*9/15/61*	*female*	democrat	.....
.....	.....	.....	.....	.....	.....	.....	

Figure 1: ...e-identifying anonymous data by linking to external data

Public voter dataset

SOME DETAILS ABOUT LINKAGE ATTACKS (2)

- Attacker learns sensitive data by joining two datasets on common attributes
  - Anonymized dataset with sensitive attributes – Example: age, race, symptoms
  - Non-sensitive dataset with individual identifiers – Example: name, address, age, race
- Demographic attributes (age, ZIP code, race, etc.) are very common in datasets with information about individuals
- **Quasi-identifiers** are crucial: (birthdate, ZIP code, gender) uniquely identifies 63% of US population nowadays
  - **Quasi-identifiers** are pieces of information that are not of themselves unique identifiers, but are sufficiently well correlated with an entity that they can be combined with other quasi-identifiers to create a unique identifier
- Publishing a record with a quasi-identifier is as bad as publishing it with an explicit identity
- Eliminating quasi-identifiers is not desirable
  - Example: users of the dataset want to study distribution of diseases by age & ZIP code

SOME DETAILS ABOUT LINKAGE ATTACKS (3)

- Identifiers and sensitive attributes

Key Attribute		Quasi-identifier		Sensitive attribute
Name	DOB	Gender	Zipcode	Disease
Andre	1/21/76	Male	53715	Heart Disease
Beth	4/13/86	Female	53715	Hepatitis
Carol	2/28/76	Male	53703	Brochitis
Dan	1/21/76	Male	53703	Broken Arm
Ellen	4/13/86	Female	53706	Flu
Eric	2/28/76	Female	53706	Hang Nail

MITIGATING LINKAGE ATTACK

- **k-anonymity**
  - The information for each person contained in the released table cannot be distinguished from at least k-1 individuals whose information also appears in the release
  - Example:
    - you try to identify a man in the released table, but the only information you have is his birth date and gender
    - there are k men in the table with the same birth date and gender.
- Any quasi-identifier present in the released table must appear in at least k records
  - Goal: each record is indistinguishable from at least k-1 other records

HOW TO ACHIEVE K-ANONYMITY

- Generalization
  - Replace quasi-identifiers with less specific but semantically consistent values until get k identical
  - Example: partition ordered-value domains into intervals
    - Age: 29, 22, 21 → 2*
    - Gender: Male, Female → *
- Suppression
  - When generalization causes too much information loss

K=2
Quasi-identifiers: Race, Birth, Gender, ZIP

	Race	Birth	Gender	ZIP	Problem
t1	Black	1965	m	0214*	short breath
t2	Black	1965	m	0214*	chest pain
t3	Black	1965	f	0213*	hypertension
t4	Black	1965	f	0213*	hypertension
t5	Black	1964	f	0213*	obesity
t6	Black	1964	f	0213*	chest pain
t7	White	1964	m	0213*	chest pain
t8	White	1964	m	0213*	obesity
t9	White	1964	m	0213*	short breath
t10	White	1967	m	0213*	chest pain
t11	White	1967	m	0213*	chest pain

22

K-ANONYMITY IN ACTION

- By linking these two tables, you still don't learn Andre's problem

Released table

	Race	Birth	Gender	ZIP	Problem
t1	Black	1965	m	0214*	short breath
t2	Black	1965	m	0214*	chest pain
t3	Black	1965	f	0213*	hypertension
t4	Black	1965	f	0213*	hypertension
t5	Black	1964	f	0213*	obesity
t6	Black	1964	f	0213*	chest pain
t7	White	1964	m	0213*	chest pain
t8	White	1964	m	0213*	obesity
t9	White	1964	m	0213*	short breath
t10	White	1967	m	0213*	chest pain
t11	White	1967	m	0213*	chest pain

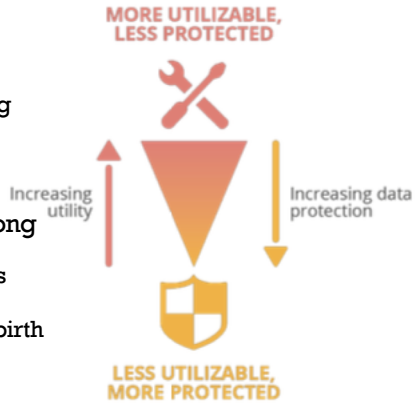
External data source

Name	Birth	Gender	ZIP	Race
Andre	1964	m	02135	White
Beth	1964	f	55410	Black
Carol	1964	f	90210	White
Dan	1967	m	02174	White
Ellen	1968	f	02237	White

23

BEWARE OF THE FOLLOWING TRADE-OFF

- A pseudonym may contain some information on the original identifier
  - Therefore, every such type of pseudonym carries the risk of being subject to a re-identification attack
  - For example, the linkage attack seen above
- In many cases, the additional information on the original identifier contained in the pseudonym is kept for linkage among pseudonyms themselves
  - Example: a pseudonym may keep the year of birth from a person's birth date as part of the pseudonym
  - It is so feasible to categorise pseudonyms based on their year of birth
- This may be an intentional feature of the pseudonymisation function allowing for some utility of the pseudonyms created, taking into account the potential loss of protection caused by this pseudonymisation approach



24

DEFINITION OF PSEUDONYMIZATION FUNCTIONS

- A pseudonymisation function maps identifiers to pseudonyms
- Requirement
  - Let $Id1$ and $Id2$ be two distinct identifiers and $pseudo1$ and $pseudo2$ be their corresponding pseudonyms
  - A pseudonymisation function must verify that $pseudo1$ is different than $pseudo2$
    - Otherwise, the recovery of the identifier could be ambiguous
- However, a single identifier Id can be associated to multiple pseudonyms ($pseudo1$, $pseudo2$...) as long as it is possible for the pseudonymisation entity to invert this operation
- In all cases, there exists some additional information that allows the association of the pseudonyms with the original identifiers (called **pseudonymisation secret**)
  - The simplest case of pseudonymisation secret is the **pseudonymisation mapping table**

25

IMPLEMENTATION OF PSEUDONYMIZATION FUNCTIONS (1)

- **Counter**
  - identifiers are substituted by a number chosen by a monotonic counter
  - A seed s is set to 0 (for instance) and then it is incremented
  - Values produced by the counter never repeat to prevent any ambiguity
  - *Pros*
    - Simplicity
  - *Cons*
    - sequential character of the counter can provide information on the order of the data
    - scalability issue for large datasets (full pseudonymisation mapping table needs to be stored)
- **Pseudo Random Number Generator**
  - Similar to counter although more complex solution, difficult to implement (avoid repetitions), but no sequential issue

26

IMPLEMENTATION OF PSEUDONYMIZATION FUNCTIONS (2)

- **Cryptographic Hash Function**
  - The digest of the identifier is the pseudonym
  - It is considered inadequate for pseudonymisation as it is prone to brute force and dictionary attacks
- **Encryption**
  - Usually, block ciphers like the AES are used for pseudonymization
  - The block cipher is used to encrypt an identifier using a secret key, which is both the pseudonymisation and recovery secret
  - Padding is used as the size of identifiers is usually less than the block size

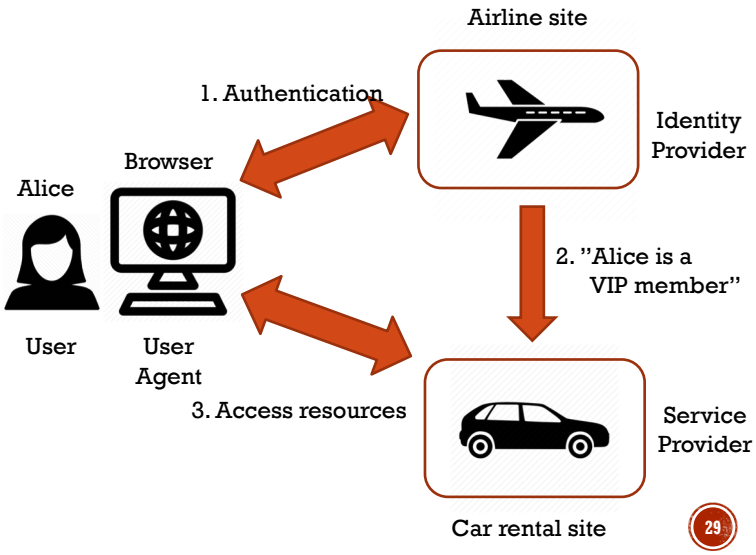
27

28

PSEUDONYMS IN SAML

RECALL THE FOLLOWING SCENARIO

- Consider Alice visits an airline website for making her trip
- For booking her flight, she provides her credentials to airline website
- After booking, she found a link to car rental (from airline website)
- She visits car rental website
- Alice rents a car without signing in again…
- … because the car rental site trusts the airline site when transmitting authentication authentication assertions such as 2

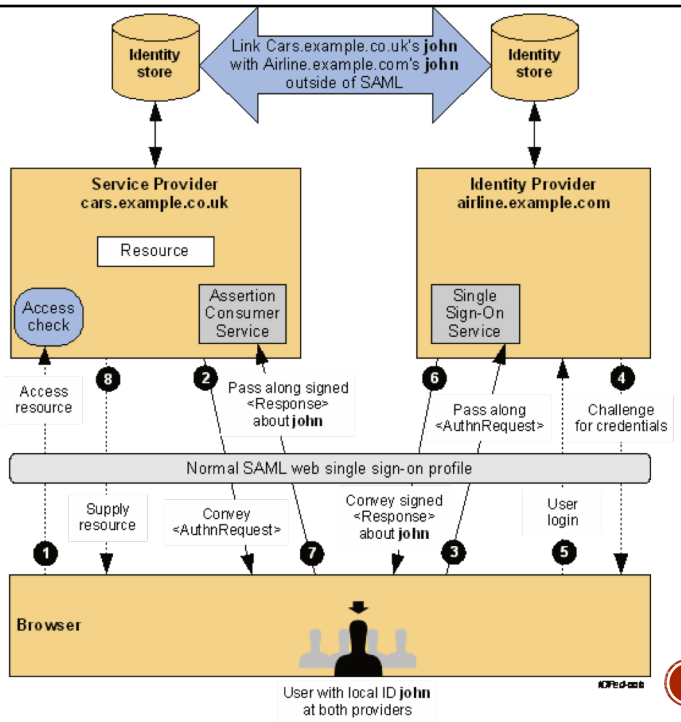


RECALL ALSO THAT…

- … SAML supports privacy (in particular, correlation of actions) by using pseudonyms of two kinds
  - **persistent pseudonyms** established between an identity and a service provider
  - ***one-time or* transient identifiers** ensure that every time a certain user accesses a given service provider through a SSO operation from an identity provider, that **service provider will be unable to recognize them as the same individual as might have previously visited**
- We now consider on the airline/car-rental scenario how the two types of pseudonyms are implemented
- This boils downs to understand how the message exchanged among users, identity providers, and service providers are tied to **individual local and federated user identities shared between participants**

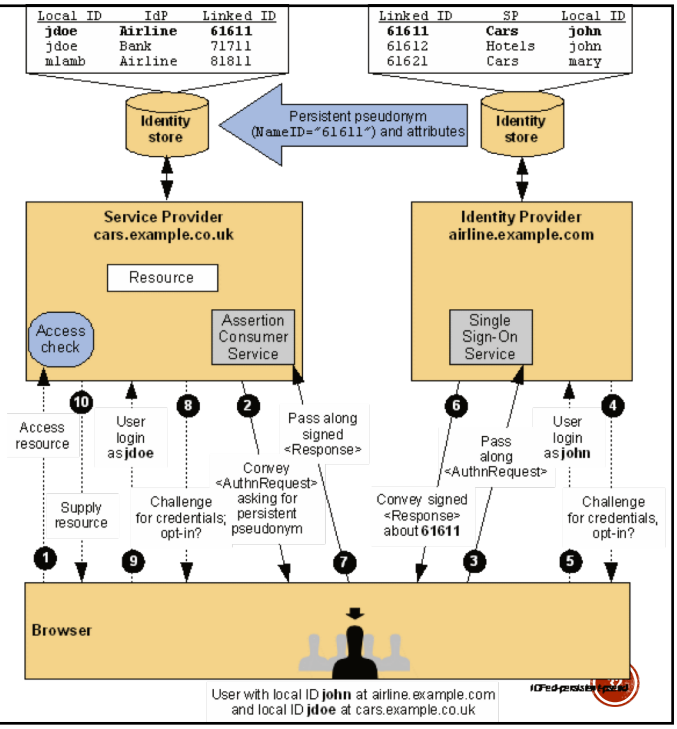
SAML IDENTITY FEDERATION (1)

- **Federation via Out-of-Band Account Linking**
- The establishment of federated identities for users and the association of those identities to local user identities can be performed without the use of SAML protocols and assertions
- No privacy is provided, standard way of performing identity federation



SAML IDENTITY FEDERATION (2)

- **Federation via Persistent Pseudonym Identifiers**
- The car rental site take advantage of SAML 2.0's ability to dynamically establish a federated identity for a user as part of the web SSO message exchange
- SAML 2.0 provides the NameIDPolicy element on the AuthnRequest to allow the SP to constrain such dynamic behaviour
- The user jdoe on cars.example.co.uk wishes to federate this account with his john account on the IdP, airline.example.com

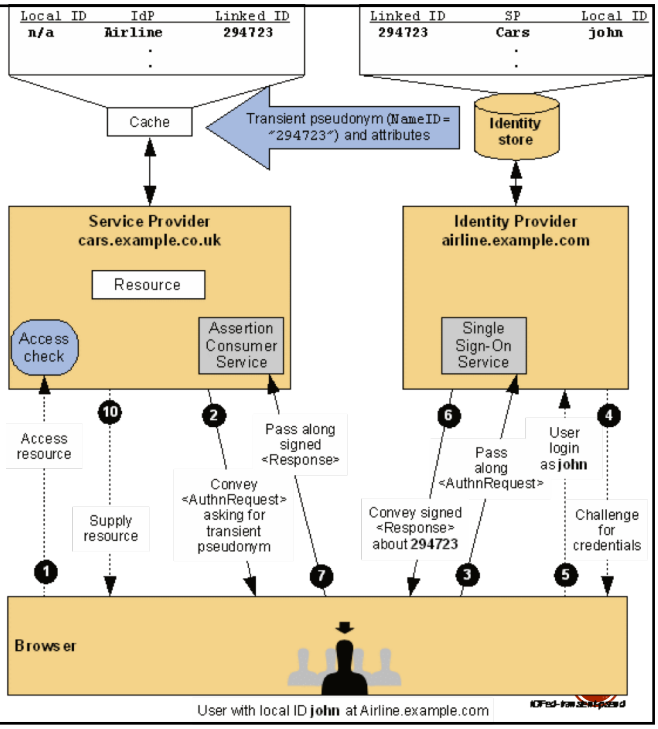


SAML IDENTITY FEDERATION (2): SOME DETAILS

- Service provider initiated flow
- User is challenged to enter credentials in the airline.example.com identity provider for username john (local to the identity provider)
- If authentication is successful, the identity provider creates a persistent identifier (61611) to be used for the session at the service provider: the authentication assertion does not include the local name john
- Upon reception and validation of the authentication assertion, the service provider checks if there is already a federation
  - If so, the username jdoe (local to the service provider) is retrieved from the association table and the SP checks if the user has the right to access the requested resource (this means checking if she is a VIP member)
  - If not, the service provider challenges the user to enter valid credentials for the username jdoe (local to the service provider), an association between the two usernames is entered in the table and the flow continues in a similar way as in the previous case

SAML IDENTITY FEDERATION (3)

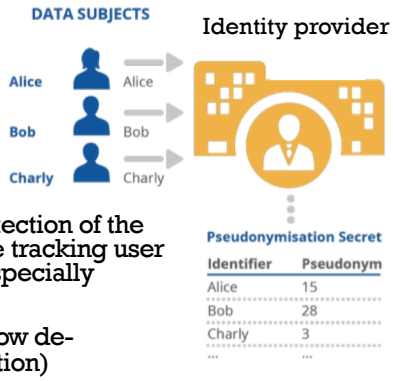
- **Federation Using Transient Pseudonym Identifiers**
-
- The car rental site take advantage of SAML 2.0's ability to dynamically establish a federated identity for a user as part of the web SSO message exchange
- SAML 2.0 provides the NameIDPolicy element on the AuthnRequest to allow the SP to constrain such dynamic behaviour
- The user jdoe on cars.example.co.uk wishes to federate this account with his john account on the IdP, airline.example.com



SAML IDENTITY FEDERATION (3): SOME DETAILS

- Service provider initiated flow
- User is challenged to enter credentials in the airline.example.com identity provider for username john (local to the identity provider)
- If authentication is successful, the identity provider creates a temporary identifier (294723) to be used for the session at the service provider: the authentication assertion does not include the name john but it contains contains a statement about the fact that the authenticated user is a VIP member
- Upon reception and validation of the authentication assertion,
  - the supplied transient name identifier is used to dynamically create a session for the user at the SP
  - the membership level attribute might be used to perform an access check on the requested resource

A REMARK



- The techniques that we have seen above allow for a better protection of the anonymity of users with respect to the services providers since tracking user activities is made more complex by the use of pseudonyms (especially transient ones)
- Still, identity providers can track user activities as they know how de-anonymize the pseudonyms (since they implement anonymization)
  - Abstractly the situation is depicted in the right corner
- However, notice that
  - If the identity providers use the pseudonyms also internally, instead of the user identifiers, this already offers a better protection to the privacy of users in case of data breaches if the pseudonymization secret is adequately protected by suitable security mechanisms (e.g., cryptography and access control)
  - If the identity providers delete from time to time the pseudonymization secret (there may be legal constraints that regulate the time interval between deletion, then even identity providers may have difficulty in reconstructing the entire tracking histories of users

36

37

GDPR

DATA PROTECTION: LEGAL PROVISIONS (1)

- Data privacy laws and regulations vary from country to country and even from state to state
- The European Union's General Data Protection Regulation (GDPR) went into effect in May, 25 2018
- **Compliance** with any one set of rules is complicated and challenging
- The basics of data protection and privacy apply across the board and include:
  - safeguarding data
  - getting consent from the person whose data is being collected
  - identifying the regulations that apply to the organization in question and the data it collects
  - ensuring employees are fully trained in the nuances of data privacy and security

38

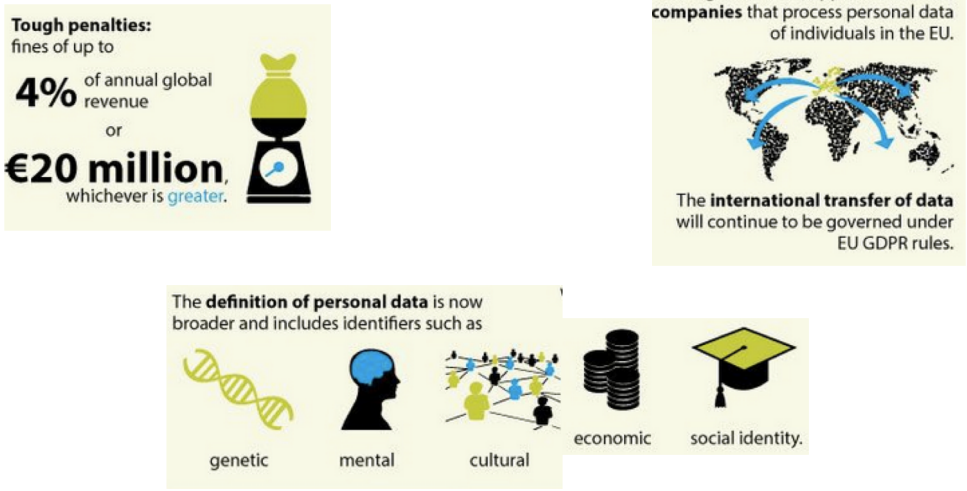
DATA PROTECTION: LEGAL PROVISIONS (2)

- The GDPR covers all EU citizens' data regardless of where the organization collecting the data is located
- GDPR requirements include:
  - Barring businesses from storing or using an individual's personally identifiable information without that person's express consent
  - Requiring companies to notify all affected people and the supervising authority within 72 hours of a data breach
  - For businesses that process or monitor data on a large scale, having a data protection officer who's responsible for data governance and ensuring the company complies with GDPR
  - Fines for not complying can be as much as €20 million or 4% of the previous fiscal year's worldwide turnover, depending on which is larger
  - Performing the Data Protection Impact Assessment (DPIA) for every data processing activity

39

GENERAL DATA PROTECTION REGULATION (GDPR)

OVERVIEW

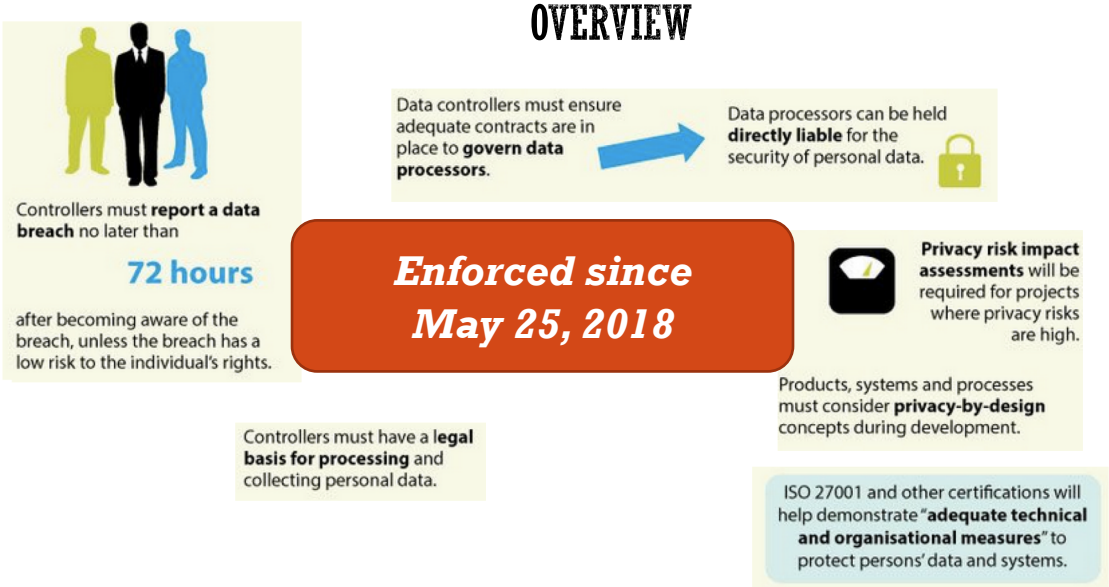


40

Pictures taken from inforgraphics at https://www.pinterest.fr/pin/396739048415981377/

GENERAL DATA PROTECTION REGULATION (GDPR)

OVERVIEW



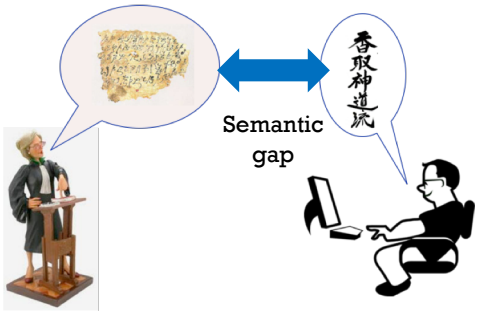
41

Pictures taken from inforgraphics at https://www.pinterest.fr/pin/396739048415981377/

GENERAL DATA PROTECTION REGULATION (GDPR)

FOCUS

Controllers must have a **legal basis for processing** and collecting personal data.



Main difficulties of compliance

- Several legal provisions: national and EU
  - Need for harmonization
- Generic provisions, e.g.
  - all data is accurate and, where necessary, kept up to date
  - data be kept for no longer than necessary
  - data is kept secure
  - data is transferred only to countries that offer adequate data protection

How to translate these into technical requirements?

42

Pictures taken from inforgraphics at https://www.pinterest.fr/pin/396739048415981377/

GENERAL DATA PROTECTION REGULATION (GDPR)

FOCUS

Privacy risk impact assessments will be required for projects where privacy risks are high.

Products, systems and processes must consider **privacy-by-design** concepts during development.

ISO 27001 and other certifications will help demonstrate "**adequate technical and organisational measures**" to protect persons' data and systems.

It's not a matter of *if* a cybersecurity breach will happen, but *when*

Random cybersecurity expert

RISK-BASED APPROACH TO CYBERSECURITY

43

44 GDPR: SELECTED ARTICLES

Full text available at
https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32016R0679

ART. 4: DEFINITIONS

- For the purposes of this regulation:
  - '**personal data**' means any information relating to an identified or identifiable natural person ('data subject'); an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person;
  - '**processing**' means any operation or set of operations which is performed on personal data or on sets of personal data, whether or not by automated means, such as collection, recording, organisation, structuring, storage, adaptation or alteration, retrieval, consultation, use, disclosure by transmission, dissemination or otherwise making available, alignment or combination, restriction, erasure or destruction;
  - '**pseudonymisation**' means the processing of personal data in such a manner that the personal data can no longer be attributed to a specific data subject without the use of additional information, provided that such additional information is kept separately and is subject to technical and organisational measures to ensure that the personal data are not attributed to an identified or identifiable natural person;

45

ART. 4: DEFINITIONS (CONT'D)

- '**controller**' means the natural or legal person, public authority, agency or other body which, alone or jointly with others, determines the purposes and means of the processing of personal data; where the purposes and means of such processing are determined by Union or Member State law, the controller or the specific criteria for its nomination may be provided for by Union or Member State law;
- '**processor**' means a natural or legal person, public authority, agency or other body which processes personal data on behalf of the controller;
- '**consent**' of the data subject means any freely given, specific, informed and unambiguous indication of the data subject's wishes by which he or she, by a statement or by a clear affirmative action, signifies agreement to the processing of personal data relating to him or her;
- '**personal data breach**' means a breach of security leading to the accidental or unlawful destruction, loss, alteration, unauthorised disclosure of, or access to, personal data transmitted, stored or otherwise processed;

ART. 7: CONDITIONS FOR CONSENT

1. Where processing is based on consent, the controller shall be able to demonstrate that the data subject has consented to processing of his or her personal data.

2. If the data subject's consent is given in the context of a written declaration which also concerns other matters, the request for consent shall be presented in a manner which is clearly distinguishable from the other matters, in an intelligible and easily accessible form, using clear and plain language. Any part of such a declaration which constitutes an infringement of this Regulation shall not be binding.

3. The data subject shall have the right to withdraw his or her consent at any time. The withdrawal of consent shall not affect the lawfulness of processing based on consent before its withdrawal. Prior to giving consent, the data subject shall be informed thereof. It shall be as easy to withdraw as to give consent.

ART. 32: SECURITY OF PROCESSING

1. Taking into account the state of the art, the costs of implementation and the nature, scope, context and purposes of processing as well as the risk of varying likelihood and severity for the rights and freedoms of natural persons, the controller and the processor shall implement appropriate technical and organisational measures to ensure a level of security appropriate to the risk, including inter alia as appropriate:

a) the pseudonymisation and encryption of personal data;
b) the ability to ensure the ongoing confidentiality, integrity, availability and resilience of processing systems and services;
c) […]

2. In assessing the appropriate level of security account shall be taken in particular of the risks that are presented by processing, in particular from accidental or unlawful destruction, loss, alteration, unauthorised disclosure of, or access to personal data transmitted, stored or otherwise processed.

48

ART. 33: NOTIFICATION OF A PERSONAL DATA BREACH TO THE SUPERVISORY AUTHORITY

1. In the case of a personal data breach, the controller shall without undue delay and, where feasible, not later than 72 hours after having become aware of it, notify the personal data breach to the supervisory authority competent […] unless the personal data breach is unlikely to result in a risk to the rights and freedoms of natural persons […]

2. The processor shall notify the controller without undue delay after becoming aware of a personal data breach.

3. The notification referred to in paragraph 1 shall at least:

a) describe the nature of the personal data breach including where possible, the categories and approximate number of data subjects concerned and the categories and approximate number of personal data records concerned;
b) communicate the name and contact details of the data protection officer or other contact point where more information can be obtained;
c) describe the likely consequences of the personal data breach;
d) describe the measures taken or proposed to be taken by the controller to address the personal data breach, including, where appropriate, measures to mitigate its possible adverse effects

49

ART. 34: COMMUNICATION OF A PERSONAL DATA BREACH TO THE DATA SUBJECT

1. When the personal data breach is likely to result in a high risk to the rights and freedoms of natural persons, the controller shall communicate the personal data breach to the data subject without undue delay.

2. The communication to the data subject referred to in paragraph 1 of this Article shall describe in clear and plain language the nature of the personal data breach and contain at least the information and measures referred to in points (b), (c) and (d) of Article 33(3).

3. The communication to the data subject referred to in paragraph 1 shall not be required if any of the following conditions are met:

a) the controller has implemented appropriate technical and organisational protection measures, and those measures were applied to the personal data affected by the personal data breach, in particular those that render the personal data unintelligible to any person who is not authorised to access it, such as encryption;

b) the controller has taken subsequent measures which ensure that the high risk to the rights and freedoms of data subjects referred to in paragraph 1 is no longer likely to materialise;

c) it would involve disproportionate effort. In such a case, there shall instead be a public communication or similar measure whereby the data subjects are informed in an equally effective manner.

50

ART. 35: DATA PROTECTION IMPACT ASSESSMENT

1. Where a type of processing in particular using new technologies, and taking into account the nature, scope, context and purposes of the processing, is likely to result in a high risk to the rights and freedoms of natural persons, the controller shall, prior to the processing, carry out an assessment of the impact of the envisaged processing operations on the protection of personal data […]

3. A data protection impact assessment referred to in paragraph 1 shall in particular be required in the case of:

a) a systematic and extensive evaluation of personal aspects relating to natural persons which is based on automated processing, including profiling, and on which decisions are based that produce legal effects concerning the natural person or similarly significantly affect the natural person;

b) processing on a large scale of special categories of data referred to in Article 9(1), or of personal data relating to criminal convictions and offences referred to in Article 10; or

c) a systematic monitoring of a publicly accessible area on a large scale.

51

ART. 35: DATA PROTECTION IMPACT ASSESSMENT (CONT'D)

7. The assessment shall contain at least:

a) a systematic description of the envisaged processing operations and the purposes of the processing, including, where applicable, the legitimate interest pursued by the controller;

b) an assessment of the necessity and proportionality of the processing operations in relation to the purposes;

c) an **assessment of the risks to the rights and freedoms of data subjects** referred to in paragraph 1; and

d) the measures envisaged to address the risks, including safeguards, security measures and mechanisms to ensure the protection of personal data and to demonstrate compliance with this Regulation taking into account the rights and legitimate interests of data subjects and other persons concerned.

52

53

RISK EVALUATION

BASIC CONCEPTS

- MEASURE
  - Definition
    - **Concrete & objective attribute**
  - Examples
    - percentage of fully patched systems within an organization
    - length of time between the release of a patch and its installation on a system
    - level of access to a system that a vulnerability could provide
- METRIC
  - Definition
    - **Abstract & somewhat subjective attribute**
  - Examples
    - how well an organization's systems are secured against external threats
    - how effective the organization's incident response team is

Approximate the value of a metric by collecting & analyzing groups of measures

1. Select metrics
2. Determine what measures can support selected metrics

54

MULTIPLE LEVEL METRICS

- Organizations should have multiple levels of metrics, each geared toward a particular type of audience
  - Examples
    - TECHNICAL SECURITY STAFF interested in lower-level metrics related to the effectiveness of particular types of security controls (e.g., malicious code detection capabilities)
    - SECURITY MANAGEMENT interested in higher-level metrics regarding the organization's security posture (e.g., overall effectiveness of incident prevention & handling capabilities)
  - Lower-level metrics for tactical decisions
  - Higher-level metrics for strategic decisions
  - Lower-level metrics used as input to higher-level metrics

Measures shall be collected via automated means to be more
- accurate and
- frequently collectable

Collected measures shall be
- analyzed and
- visualized in dashboards

55

ACCURACY OF METRICS

- Dependent on the accuracy of measures
- Problems
  - Imprecise definition
    - Example: % of fully patched systems
      Include only OS patches? Also include service/application patches?
      Patches have been installed or also activated by, e.g., rebooting/ configuration changes?
  - Ambiguous terminology
    - Example: #of port scans If an attacker scans ports on 100 hosts, is it one port scan or 100 port scans?
      If the attacker performed the same scan but only scanned one host each day, is it one port scan or 100 port scans?
  - Inconsistent measurement methods
    - Example: patch status
      An OS might report only on OS patches, while another OS might also include some application patches

The definition of measure and its measurament method are driven by what the organization is trying to accomplish

Avoid as much as possible qualitative measures; at least those without a predefined scale

56

SELECTION OF MEASURES

- Only measures supporting selected metrics are needed

Many suggestions in the security community for what measures organizations should collect

- Clearly and accurately specify dependencies among measures in the metrics using them

Little work has been done to determine the value of measures in real-world operational environments, including which measures are most supportive of particular metrics

57

USE OF MEASURES

- Selected measures support the determination of the chosen metrics
- Determining how to combine the values of the measures into a metric
- Some measures may be more important than others in the scope of a metric
  - It is difficult to quantify what weight each measure should be given

Although high-level metrics may stay the same, low-level metrics need to change over time as the security posture of the organization changes

Continuous use of measures and adaptive metrics to reflect evolving CyberSecurity posture needs

Empirical research may provide organizations with a factual basis for weighting measures instead of either guessing or weighting each measure equally

58

RISK: A DEFINITION

- [Likelihood of adverse event] × [Impact of the adverse event]
- Likelihood evaluated over a specific time period
  - Example: 1 year *annual loss expectancy* is evaluated
- Definition above considers risk due to a single specific cause
- When statistically independent multiple causes are considered the individual risks need to be added to obtain overall risk

Likelihood =
[Pr. exploitable weakness in the sys]
×
[Pr. weakness is exploited]

Prop. of system

External factor (e.g., hacker motivation)

- Risk is often measured conditionally, by assuming that some of the factors are equal to unity and thus can be dropped
- Example: replace the impact factor in the definition of risk by 1, so that conditional risk = probability of the adverse event
  - This is standard in reliability theory

59

HOW TO USE RISK IN CYBERSECURITY

- Adverse event = Vulnerability = Weakness allowing an attacker to reduce system's information assurance
  - Example: SW vulnerability is a SW defect/bug that can be exploited by an attacker to cause loss or harm
- Stochastic models can be used to evaluate [likelihood of adverse event]
  - Example: stochastic model of SW vulnerability lifecycle in
    - Joh & Malaiya. *A Framework for Software Security Risk Evaluation using the Vulnerability Lifecycle and CVSS Metrics*. Proc. Int. Ws. on Risk and Trust in Extended Enterprises, 2010.
- Metrics from available vulnerability databases can be used to evaluate [impact of adverse event]
  - Example: for SW vulnerabilities can be used the Common Vulnerability Scoring Systems (CVSS)
    - Mell, Scarfone & Romanosky. *CVSS: A complete Guide to the Common Vulnerability Scoring System Version 2.0*. Forum of Incident Response and Security Teams (FIRST), 2007.

60

HOW TO USE RISK IN CYBERSECURITY (CONT'D)

- Several vulnerabilities: i=1, …, n
- System risk = SUM over i=1, …, n of (Li × Ii)
  - Li = likelihood of exploitation of vulnerability i
  - Ii = impact of exploitation of vulnerability i
- Risk matrix: impact & likelihood are divided into a set of discrete intervals
  - Example: 5 impact levels from [*Very Low-Very High]* & 5 likelihood levels [*Very Unlikely-Frequent]* Each level is assigned a *rating*
- Scales used for likelihood & impact can be linear or non-linear
  - Logarithmic scale for both has some advantages; e.g., risk becomes "additive"

	CONSEQUENCE				
LIKELIHOOD	Insignificant 1	Minor 2	Moderate 3	Major 4	Catastrophic 5
Almost Certain 5	5	10	15	20	25
Likely 4	4	8	12	16	20
Possible 3	3	6	9	12	15
Unlikely 2	2	4	6	8	10
Rare 1	1	2	3	4	5

61

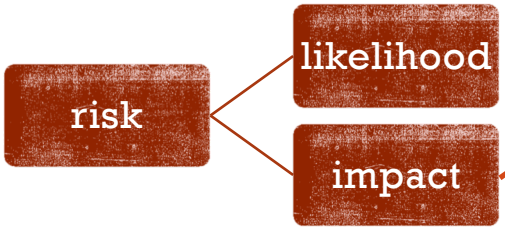
GIVING VALUES TO LIKELIHOOD & IMPACT

- CVSS: industrial standard for assessing security vulnerabilities; attempts to evaluate the degree of risks posed by vulnerabilities
  - Vendor independent framework for communicating likelihood & impact of known vulnerabilities
  - Metrics(*scores)* are computed using proxies of vulnerabilities suggested by security experts
- CVSS scores for known vulnerabilities are available on major public vulnerability dbs (e.g., National Vulnerability Database, NVD)
- For a score, few qualitative levels defined and numerical value associated with them
- Base score: intrinsic characteristics of vulnerability, include 2 sub-scores:
  - Exploitability: attempt to measure how easy it is to exploit the vulnerability
  - Impact: measures how a vulnerability will impact an IT asset in terms of losses in Confidentiality, Integrity & Availability
- Empirical formula to compute Base score combining the 2 sub-scores
  - Goal: ranking of vulnerabilities based on the risk posed by them

62

RISK PERSPECTIVES

risk = likelihood x impact



Likelihood	Negligible (1)	Marginal (2)	Serious (3)	Critical (4)	Catastrophic (5)
	Severity of Impact				
Highly Probable (5)	1x5= 5	2x5= 10	3x5= 15	4x5= 20	5x5= 25
Probable (4)	1x4= 4	2x4= 8	3x4= 12	4x4= 16	5x4= 20
Occasional (3)	1x3= 3	2x3= 6	3x3= 9	4x3= 12	5x3= 15
Remote (2)	1x2= 2	2x2= 4	3x2= 6	4x2= 8	5x2= 10
Improbable (1)	1x1= 1	2x1= 2	3x1= 3	4x1= 4	5x1= 5

- A standard like ISO 27001:2013 is a guideline for risk evaluation of organization/companies (i.e. **impact on business goals**)
- The GDPR is similar to the ISO 27001:2013 with the exception that the **impact is evaluated wrt the data subject**

63

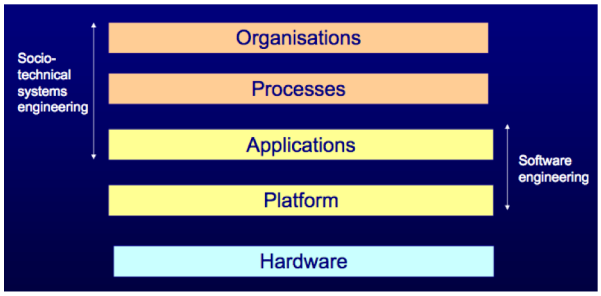
64

DATA PROTECTION: EXAMPLES

SOCIO-TECHNICAL SYSTEM

Term coined in 1960 by Trist and Emery. Have a look at http://www.moderntimesworkplace.com/archives/archives.html

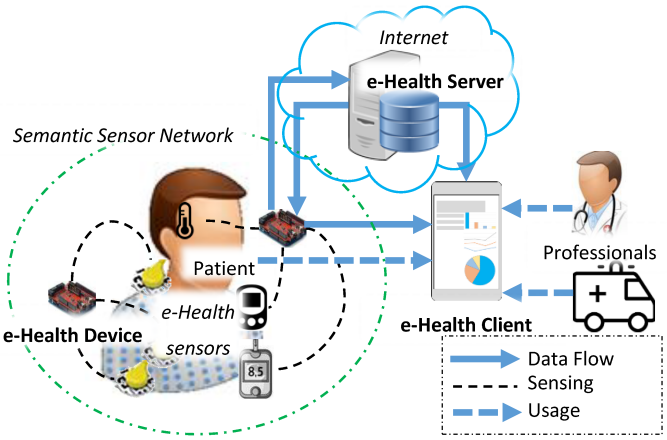
- **Socio-Technical System (STS)** = any system that involves interaction between humans, machines, and environmental aspects of the work system
- **Socio-Technical System (STS)** = any system that represents social human interaction through technology
- Many areas of study from business to health look for ways to bridge social interaction with technology, a STS can frame these studies into easy to read models that will help businessmen and health experts find a **middle ground between technology and social interactions** to develop new technologies that will benefit them and the people

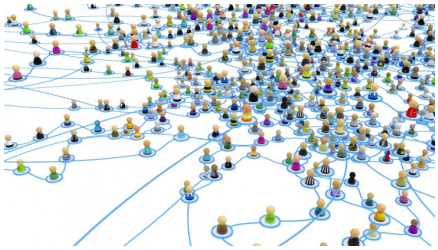


Picture from: **Ian Sommerville, 2004. Software Engineering, 7th edition. Chapter 2 Slide 1 Systems engineering 1.** Available at http://slideplayer.com/slide/5797346/

65

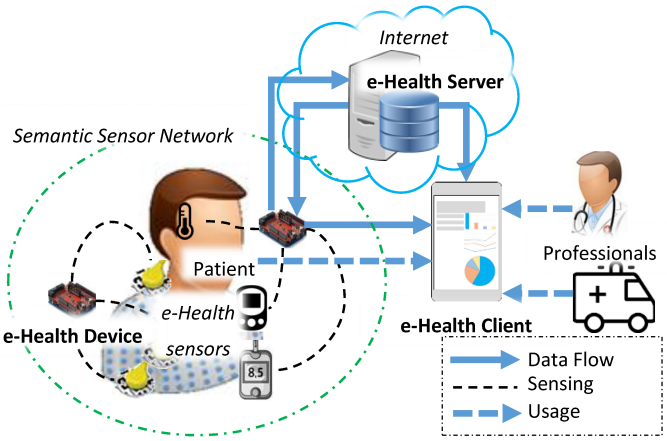
SOCIO-TECHNICAL SYSTEM: SOME EXAMPLES

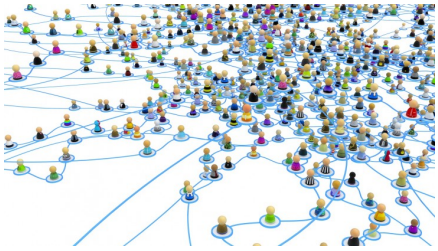




66

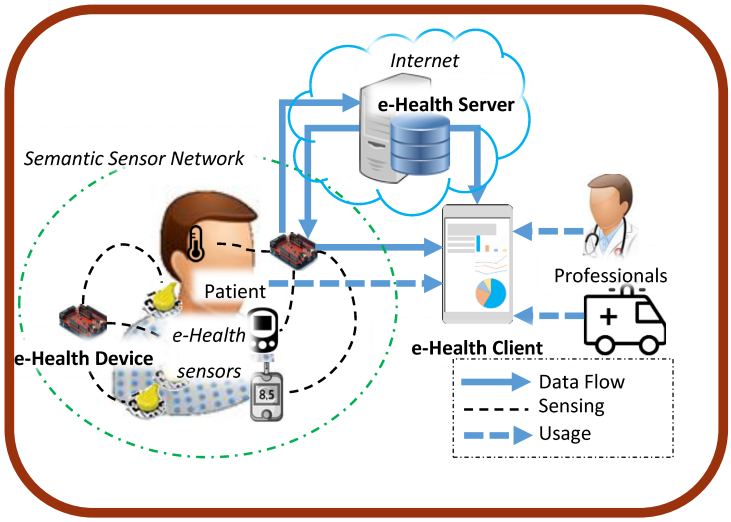
SOCIO-TECHNICAL SYSTEM: SOME EXAMPLES

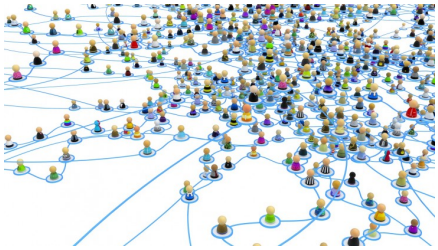




67

SOCIO-TECHNICAL SYSTEM: SOME EXAMPLES





68

EXAMPLE: ANTHEM

- Biggest healthcare breach to date
- January 29, 2015: **78.8 million patient records** stolen
- Leakage of highly sensitive data, including
  - names, Social Security numbers, home addresses, and dates of birth

- **IMPACT**: persons whose data was stolen could have resulting problems about **identity theft for the rest of their lives!!!**

Some details of the attack

1. user within one of Anthem's subsidiaries opened a **phishing email** containing malicious content
2. download of malicious files to the user's computer and allowed **hackers to gain remote access to** that computer and dozens of other systems within the Anthem enterprise, including **Anthem's data warehouse**
3. attacker was able to move laterally across Anthem systems and **escalate privileges**, gaining increasingly greater ability to access information and make changes in Anthem's environment
4. The attacker utilized at least 50 accounts and compromised at least 90 systems within the Anthem enterprise environment including, eventually, the company's enterprise data warehouse
5. Queries to that data warehouse resulted in access to an **exfiltration of approximately 78.8 million unique user records**

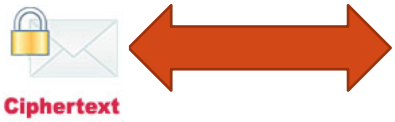
69

ANTHEM: MAIN PROBLEM

The Health Insurance Portability and Accountability Act of 1996 (**HIPAA**) is a federal law that required the creation of national standards to protect sensitive patient health information from being disclosed without the patient's consent or knowledge

- Data in transit: encrypted
- Data at rest: un-encrypted
  - This is **ok** with HIPAA!

Anthem Data warehouse



70

ANTHEM: MAIN PROBLEM... REALLY?

- Data in transit: encrypted
- Data at rest: un-encrypted
  - This is **ok** with HIPAA!

Anthem Data warehouse

Would encrypting the data (at rest) have prevented the data from being stolen?

Probably not after the admin's credentials were compromised as cryptographic keys become accessible…

71

KEY OBSERVATIONS

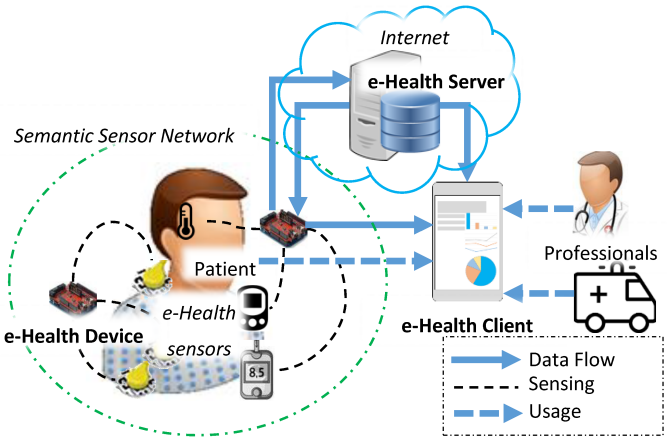
- Cryptographic keys are sensitive data stored in a computer system:
  - **Other security mechanisms** (e.g., access control) **have to protect these keys**
- Cryptography (alone) is rarely ever the solution to a security problem
  - False sense of security
- Cryptography is a translation mechanism
  - *Converting a security problem into a key management/protection problem*

- Cryptography is no silver bullet
- Other security mechanisms are also important, such as
  - **Authentication**
  - **Access control**

More than 80% of data breaches due to authentication and/or access control problems!

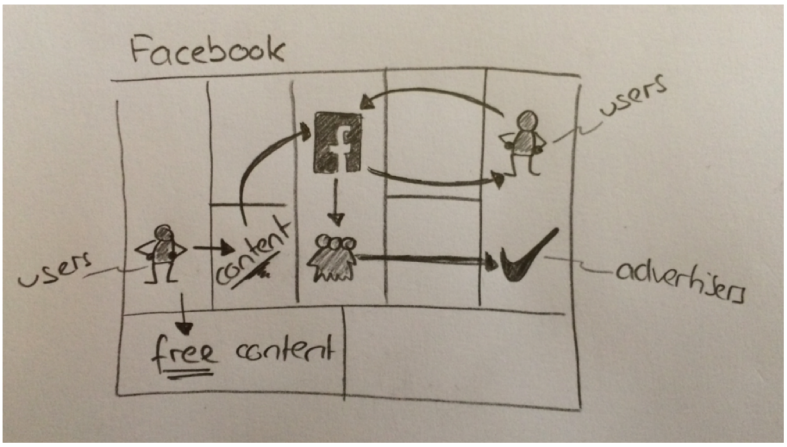
72

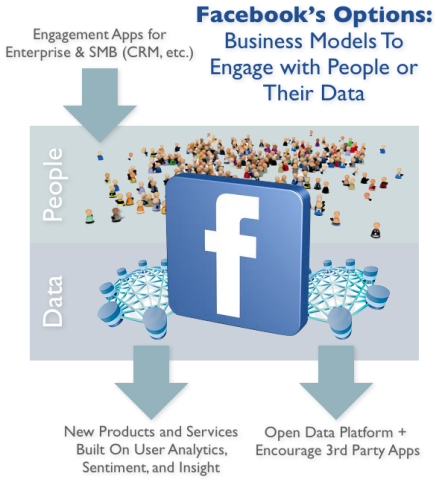
SOCIO-TECHNICAL SYSTEM: SOME EXAMPLES



73

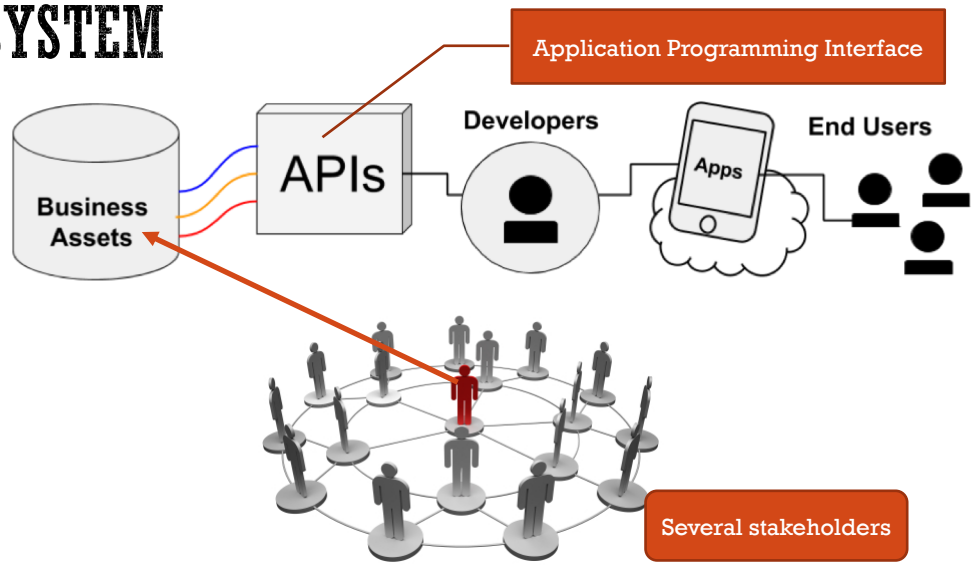
FACEBOOK BUSINESS MODEL





74

FACEBOOK AS OPEN SOCIO-TECHNICAL SYSTEM



75

WHY THE WORD "OPEN" MATTERS

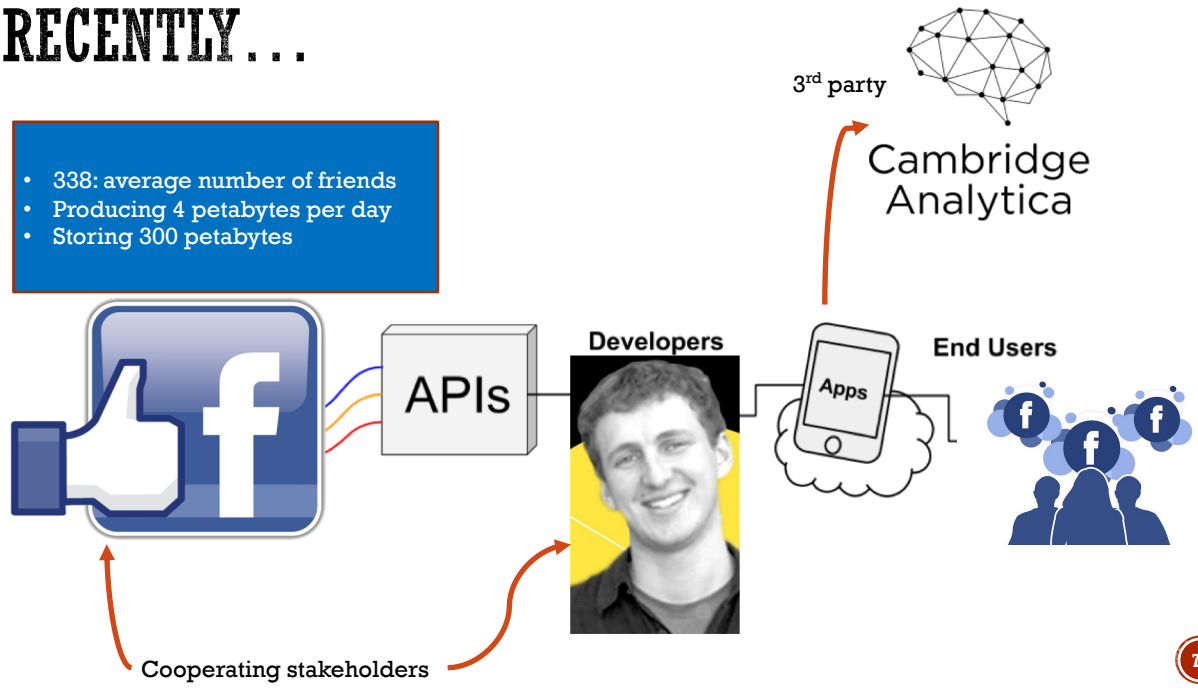
- When someone uses Facebook Login to connect with apps and services, they grant those apps access to a range of information from their Facebook profile
- In **2015, Facebook also allowed apps to access some information from the friend networks of people who used Facebook Login, even though those friends may not have agreed to share their data**
- Although Kogan gained access to this vast data set in a legitimate way and through the proper channels that governed developers on Facebook at that time, **he shared it with Cambridge Analytica**, violating Facebook's policies
- Plus, the **app was presented as a personality quiz, but the data collected was used for political marketing**

- Problem: how can Facebook (or other social networks) prevent bad actors from using platforms to harvest personal data?
- Solution(?): offer better visibility into how user information is handled by third-party apps

76

RECENTLY . . .

- 338: average number of friends
- Producing 4 petabytes per day
- Storing 300 petabytes



77

RECENTLY . . .

- 338: average number of friends
- Producing 4 petabytes per day
- Storing 300 petabytes

Is this a data breach?

3rd party

Cambridge Analytica

APIs

Developers

Apps

End Users

Cooperating stakeholders

78

RECENTLY . . .

- 338: average number of friends
- Producing 4 petabytes per day
- Storing 300 petabytes

Is this a data breach?

3rd party

Cambridge Analytica

GDPR-Art. 4: Definitions
[…]
12) '**personal data breach**' means a breach of security leading to the accidental or unlawful destruction, loss, alteration, unauthorised disclosure of, or access to, personal data transmitted, stored or otherwise processed;
[…]

Cooperating stakeholders

79

WHY IS THIS A BAD STORY?

- **Scale**: 87 millions involved starting from as few as around 200,000 users who installed the app by Kogan
- **Impact**: steering USA presidential election by targeted fake news (?) Damaging democracy…

80

81 TAKEAWAYS

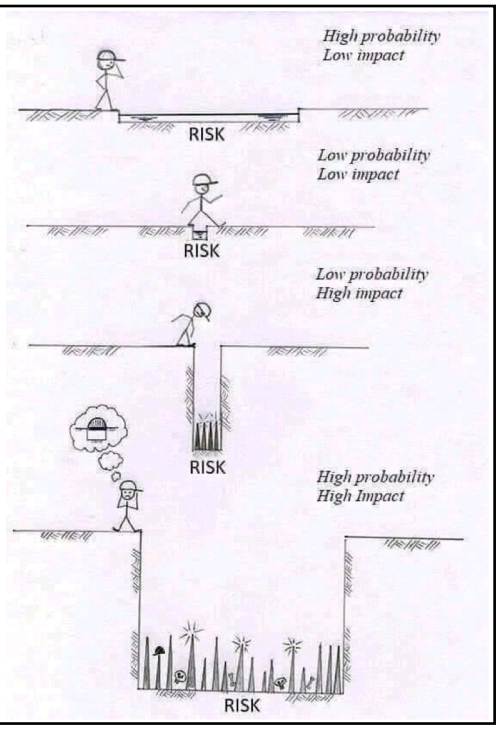
Key to combine
- Technolgical
- Legal

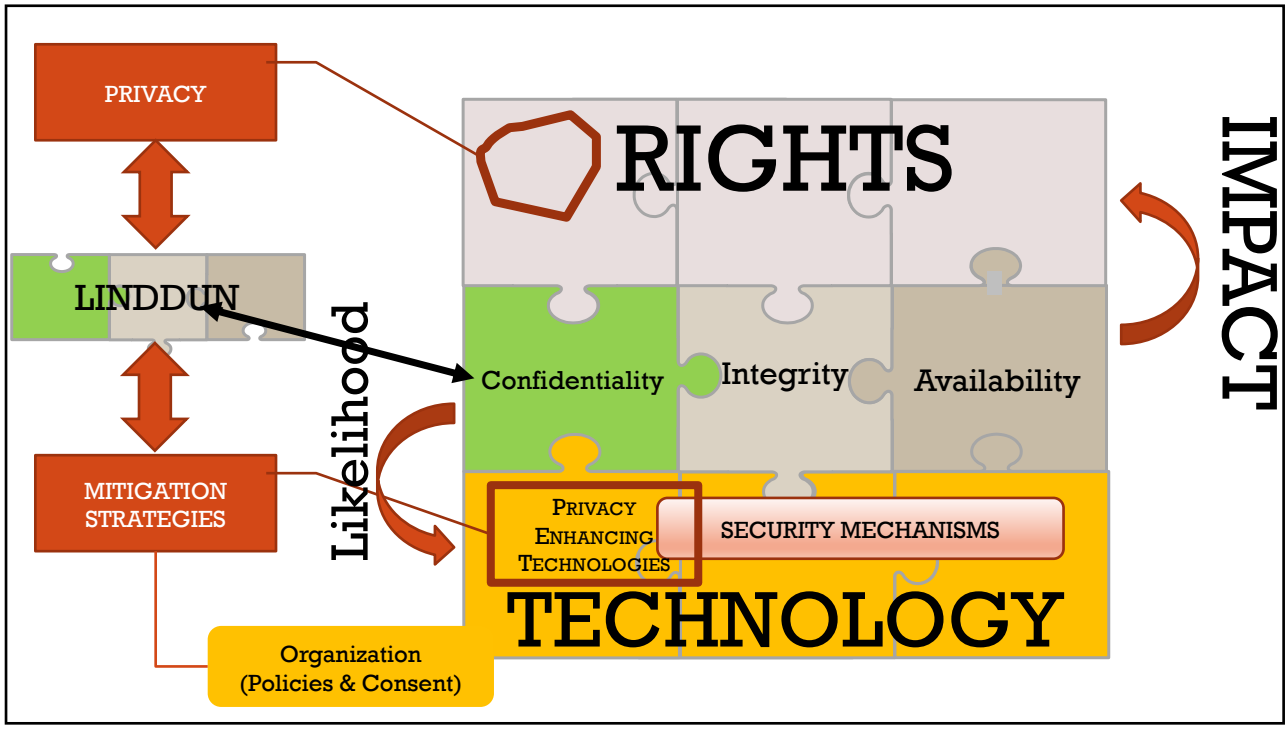
aspects of security, privacy, and fundamental rights

82

RISK MATRIX

	1	2	3	4	5
	Severity of Impact				
Likelihood	Negligible	Marginal	Serious	Critical	Catastrophic
5 Highly Probable	1x5= 5	2x5= 10	3x5= 15	4x5= 20	5x5= 25
4 Probable	1x4= 4	2x4= 8	3x4= 12	4x4= 16	5x4= 20
3 Occasional	1x3= 3	2x3= 6	3x3= 9	4x3= 12	5x3= 15
2 Remote	1x2= 2	2x2= 4	3x2= 6	4x2= 8	5x2= 10
1 Improbable	1x1= 1	2x1= 2	3x1= 3	4x1= 4	5x1= 5





TWO EXAMPLES

In other words, data protection (as concieved by the GDPR) is an even broader and more subtle notion than privacy!

- Violation of integrity
  - Consider a database containing the data of citizens (e.g., salary, owner of house, …) enabling them to ask for tax reductions
  - If an attacker modify such data, it is possible that a citizen will be denied the tax reduction even though he/she is will entitled to do so
- Violation of availability
  - Consider an on-line system to register kids to school; the system accepts requests only during a limited period of time
  - If an attacker mount a denial of service attack during the time window during which it is possible to submit registration requests, then the right to education can be reduced

RECAP QUESTIONS

- What is the likelihood of an event? What is the impact of an event? What is the risk of an event? What is the risk matrix?
- Define the notion of privacy. What is k-anonimity?
- What is data protection? What is LINDDUN?
- What is the scope of application of the GDPR? Who is the data controller? Who is the data processort? What is the Data Protection Impact Assessment? Why does the GDPR propose a risk-based approach to data protection?
- What is the difference between the risk evaluated for an organization and the risk of a data processing activity in the GDPR?