



Legal Aspects of Anonymization and Pseudonymization

--

Module 3: Pseudo/Anon Terminology

Bud P. Bruegger



Outline Module 3

Pseudo/Anon Terminology

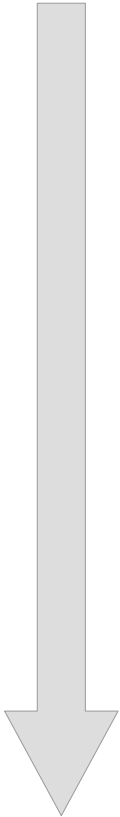
- **Technology Transfer**
- **Conceptualization and Terminologies**
- **Analysis of Legal Texts → Messaging Needs: Tech → Policy**
- **Messaging (impact) Requirements**
- **Resulting Terminologies**
- **Dissemination**

Outline Module 3

Pseudo/Anon Terminology

- **Technology Transfer**
- **Conceptualization and Terminologies**
- **Analysis of Legal Texts → Messaging Needs: Tech → Policy**
- **Messaging (impact) Requirements**
- **Resulting Terminologies**
- **Dissemination**

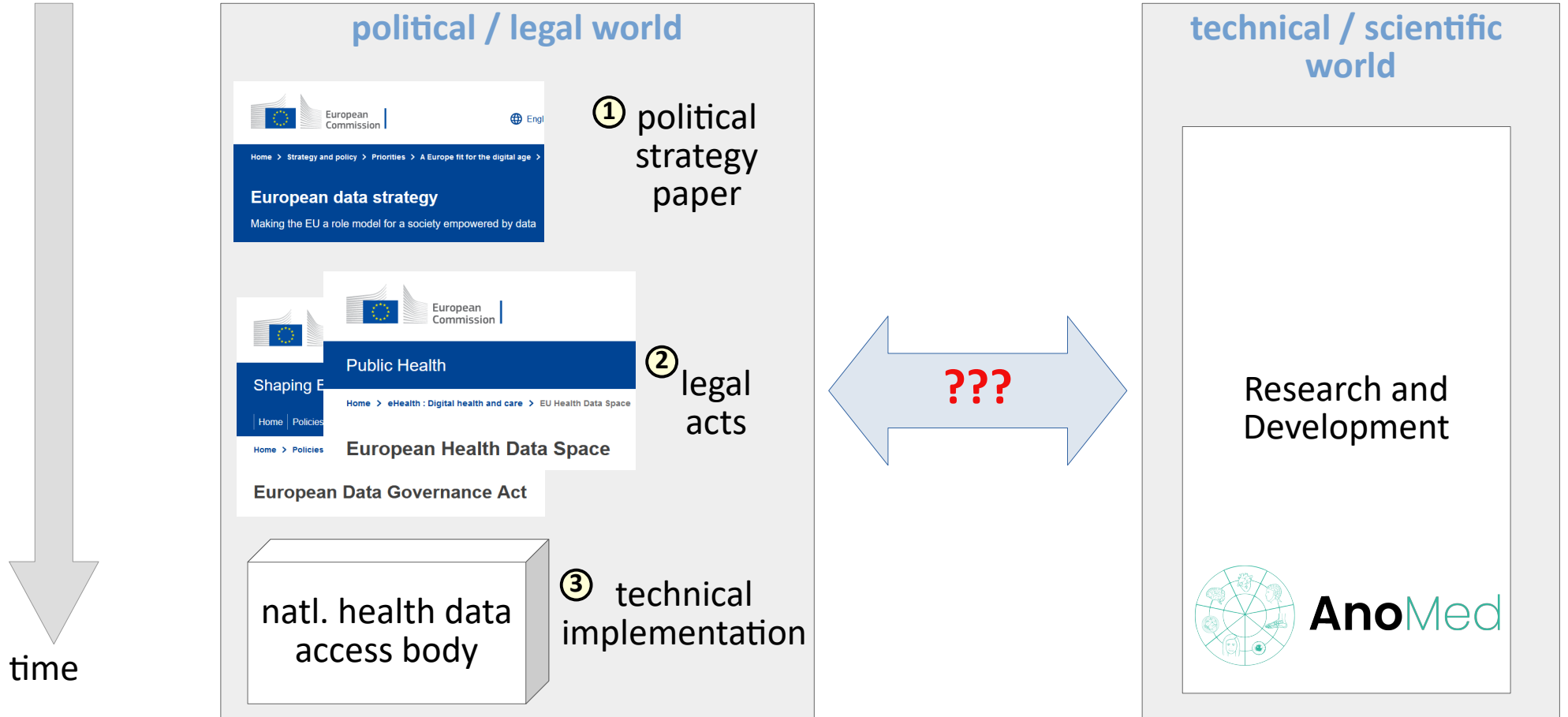
New Technology in Society



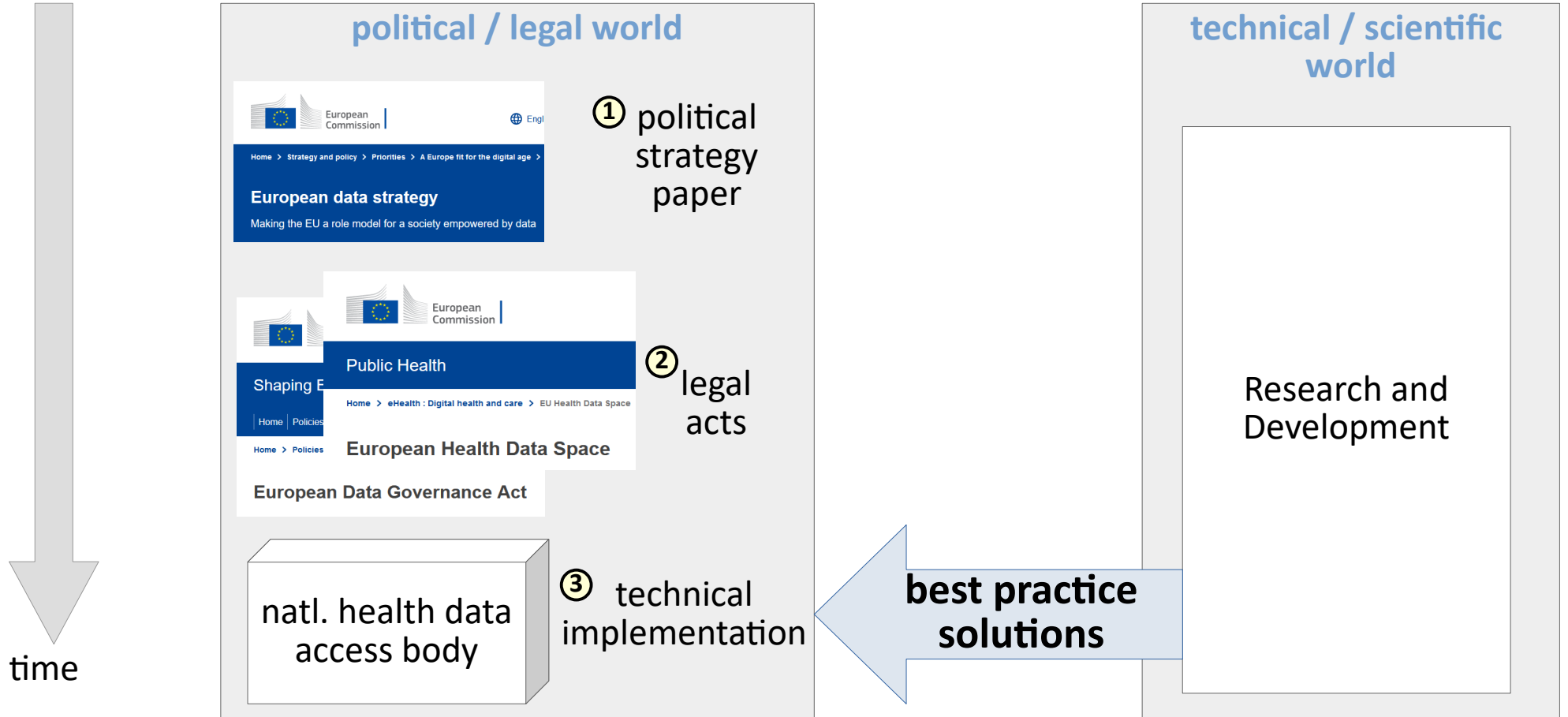
time



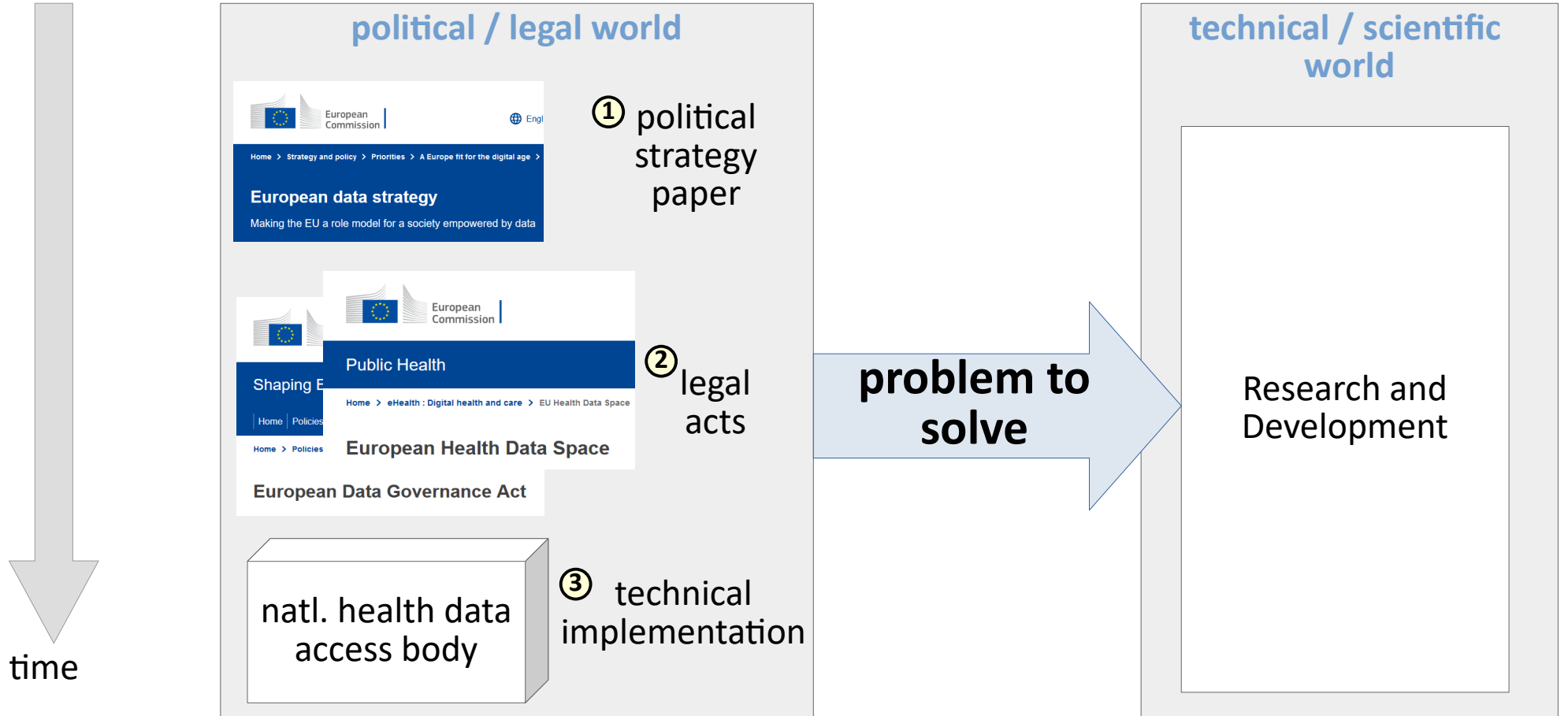
Interaction?



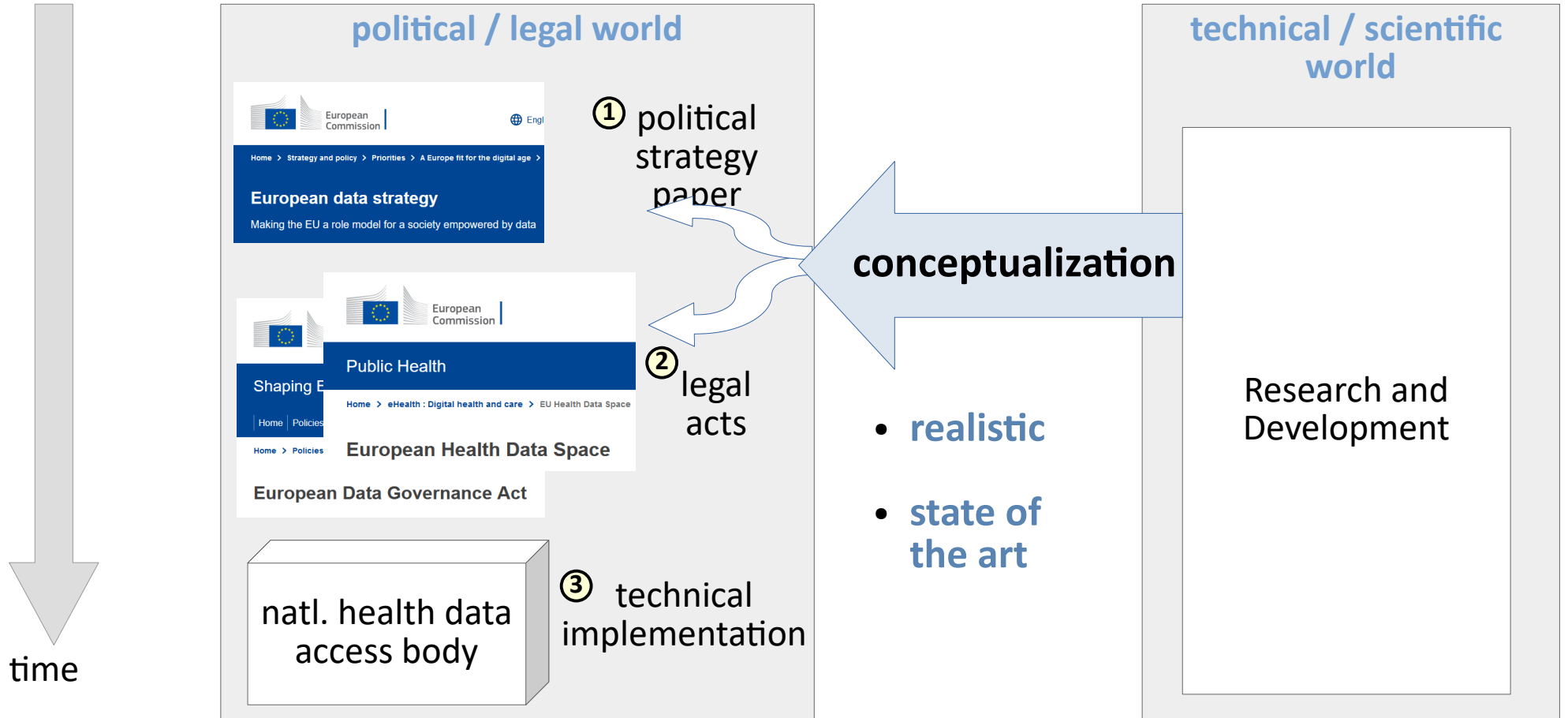
Technology Transfer



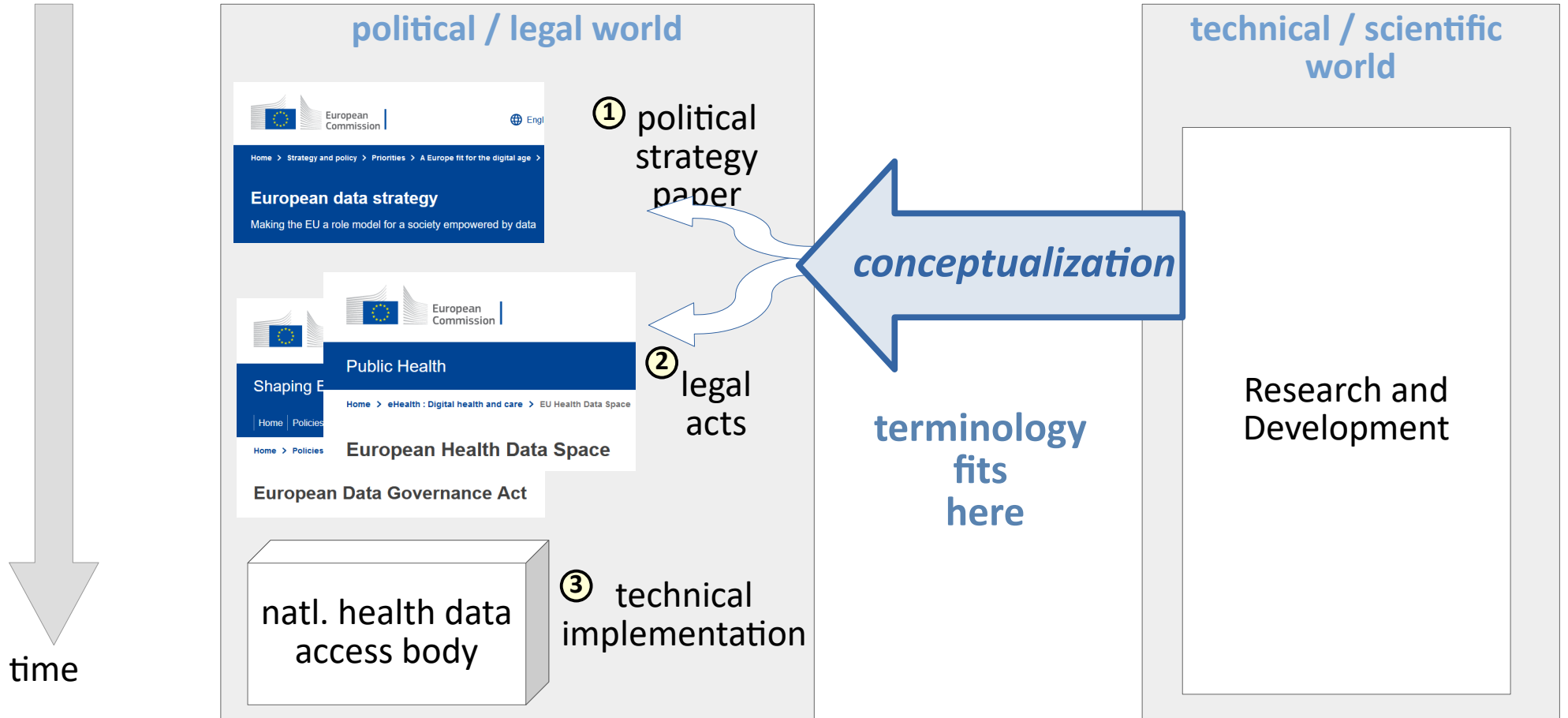
“Legal Transfer”



Conceptualization



Conceptualization

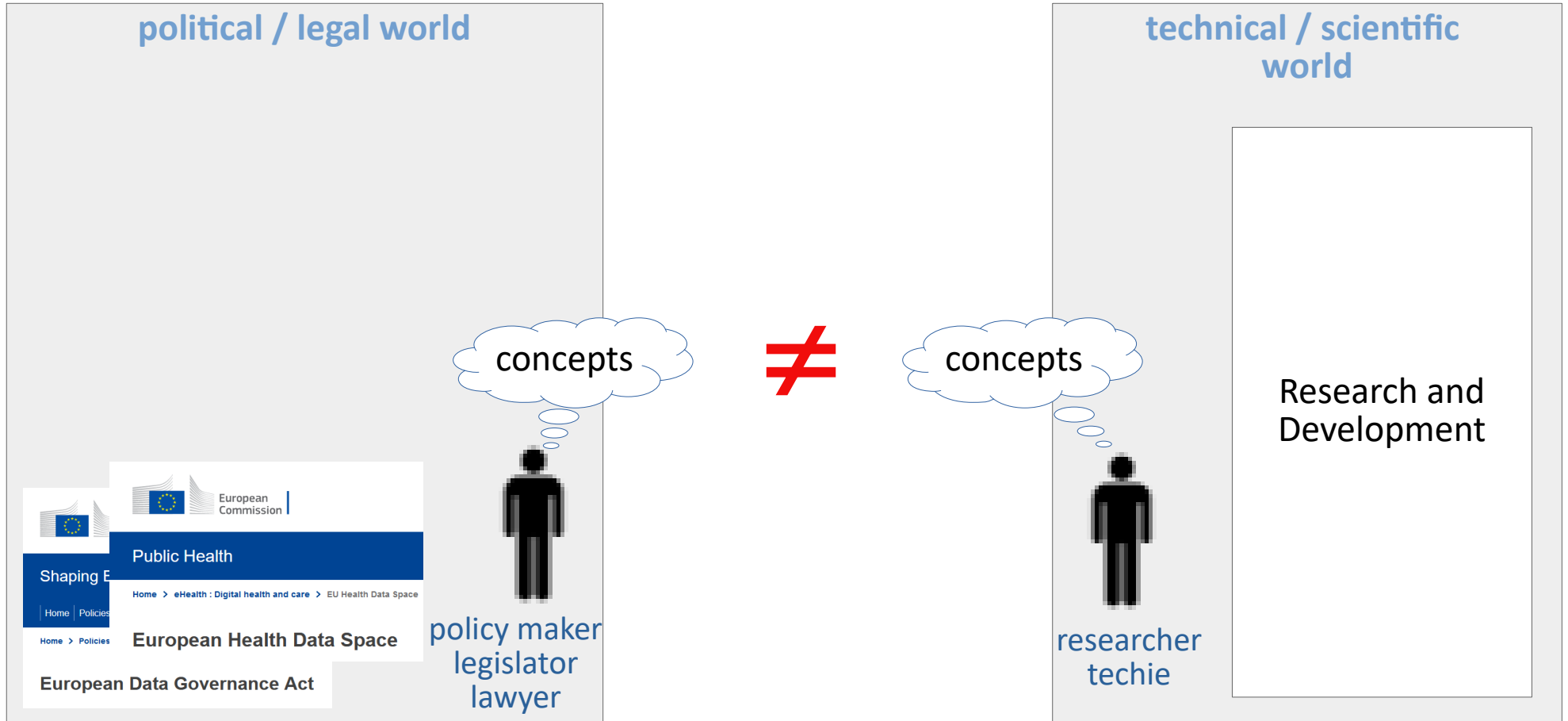


Outline Module 3

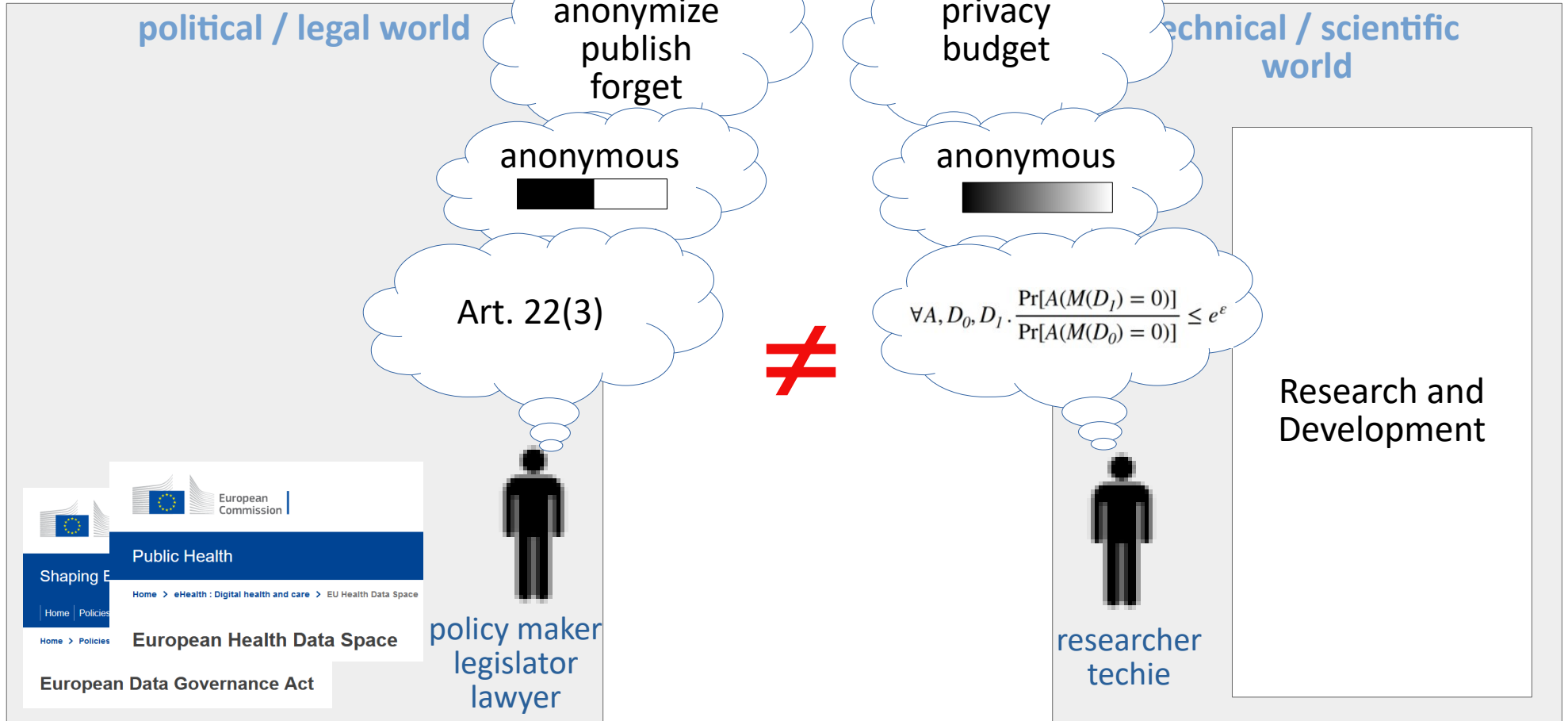
Pseudo/Anon Terminology

- Technology Transfer
- **Conceptualization and Terminologies**
- Analysis of Legal Texts → Messaging Needs: Tech → Policy
- Messaging (impact) Requirements
- Resulting Terminologies
- Dissemination

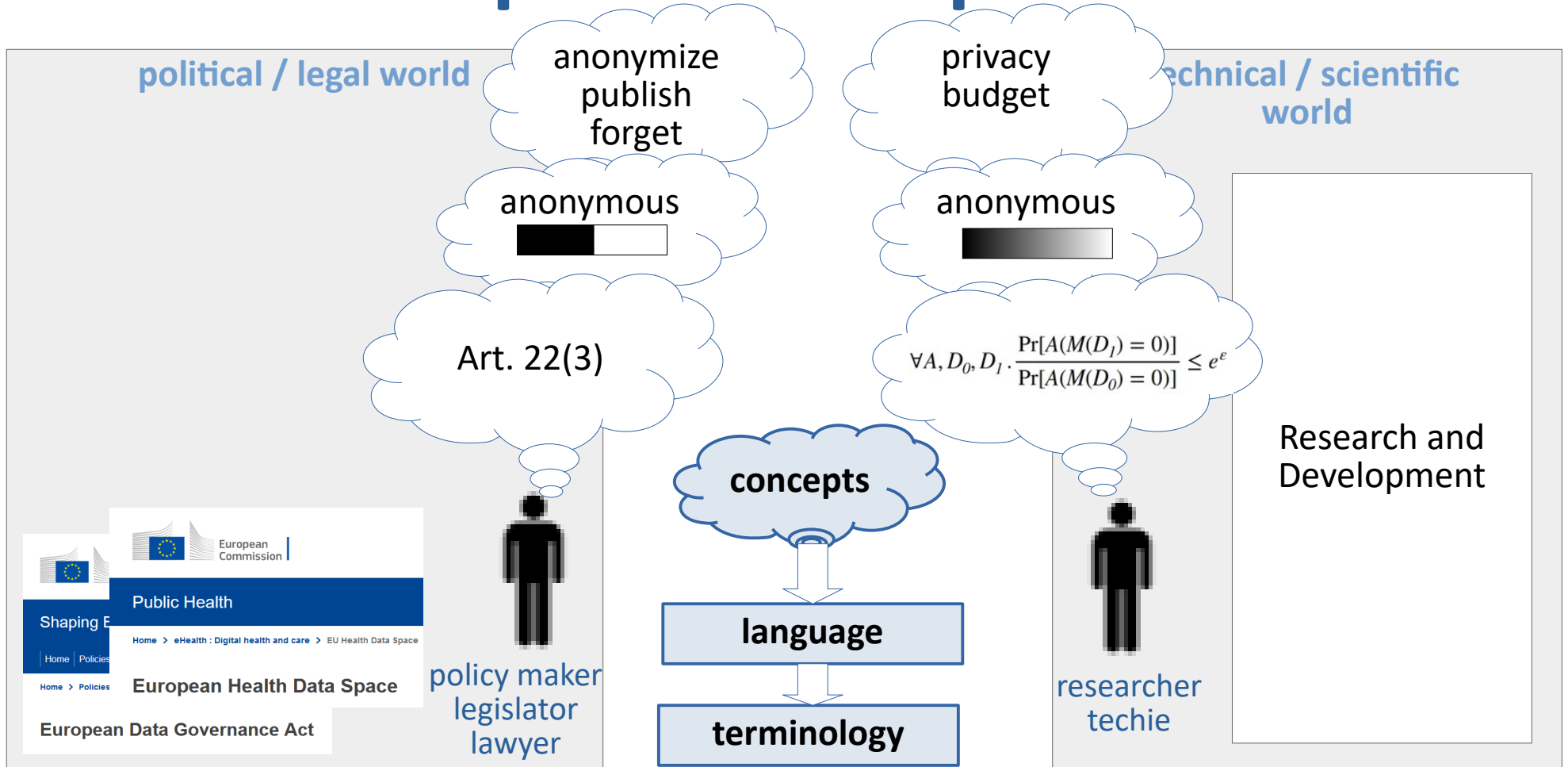
The Conceptual Gap



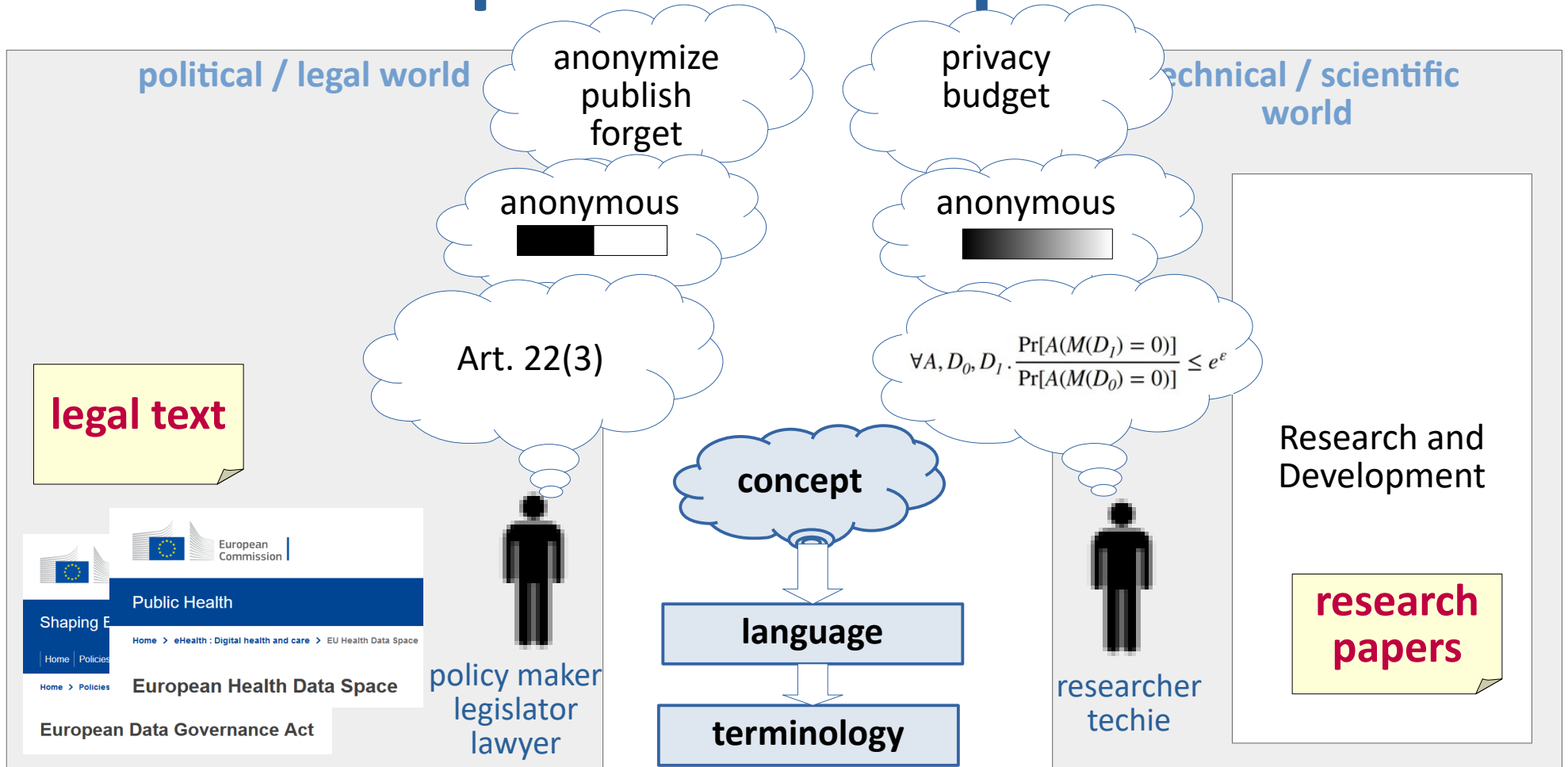
Example Differences



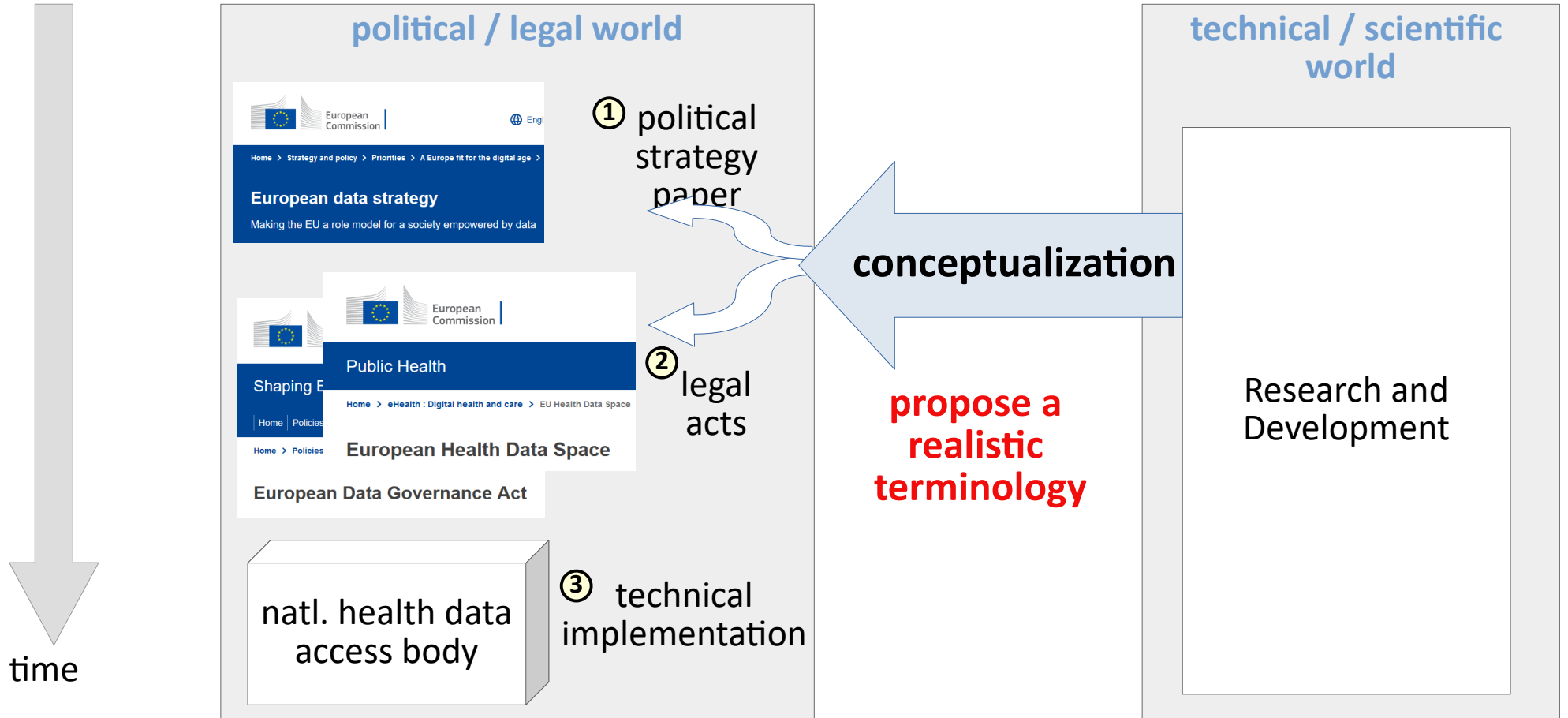
how to capture conceptualization?



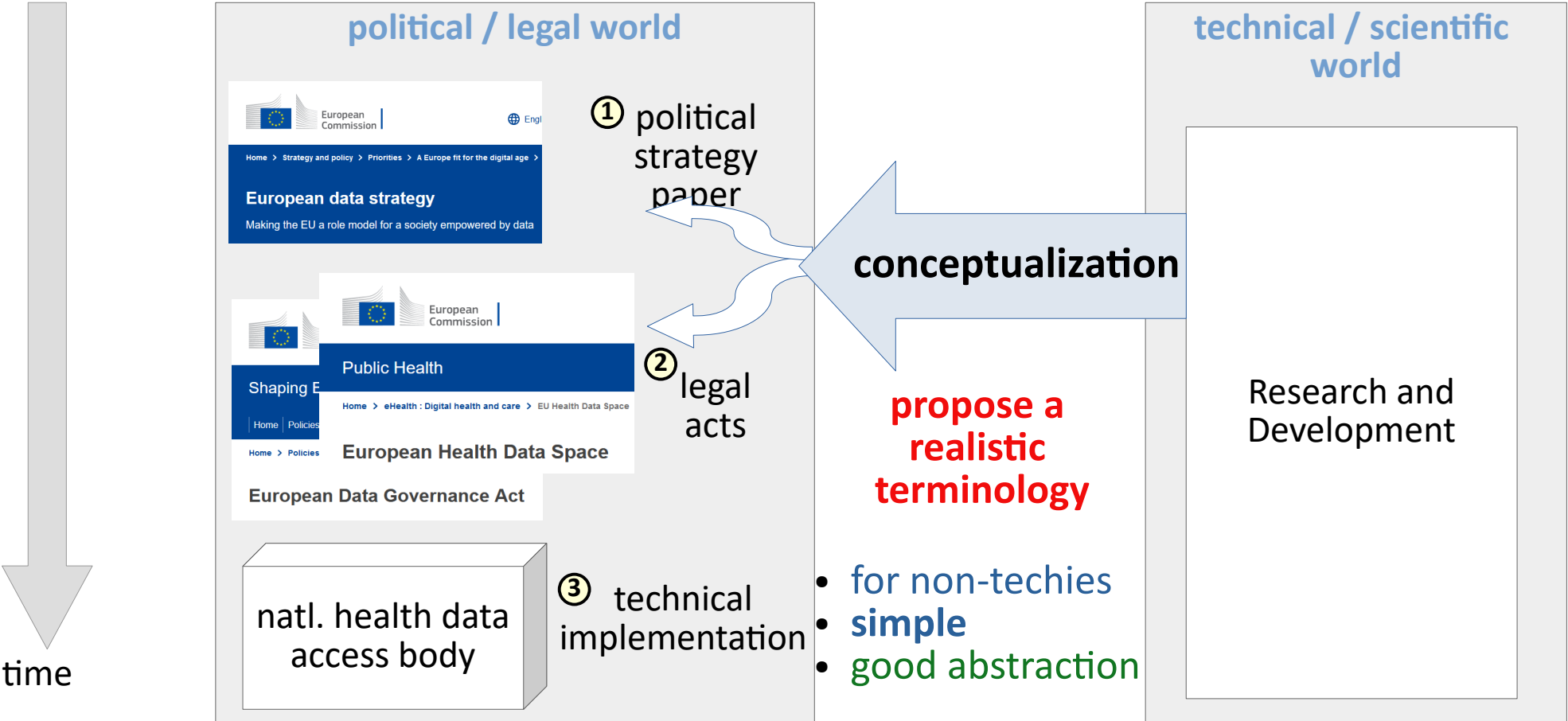
how to capture conceptualization?



how to affect conceptualization



how to affect conceptualization?



ULD's unique position


political / legal world


technical / scientific world



ULD 
Unabhängiges Landeszentrum für
Datenschutz Schleswig-Holstein
Research Department


one foot
in the
legal world


get toes wet
in technical
research

 **AnoMed**

Outline Module 3

Pseudo/Anon Terminology

- Technology Transfer
- Conceptualization and Terminologies
- **Analysis of Legal Texts** → Messaging Needs: Tech → Policy
- Messaging (impact) Requirements
- Resulting Terminologies
- Dissemination

Objective of Analysis

– Conceptualization


- reverse-engineering of text
- how do policy-makers think about technology

– Shortcomings

- mismatches between technical and policy concepts?
 - what is unrealistic?
- what “messaging” does new terminology need to convey?
 - bridge gap between technical and policy concepts

– Compatibility

Analyzed Legal Texts

-  **Guidelines on Anonymization:** draft
 - official legal interpretation → concepts
 - compatibility!
- **Data Governance Act (*DGA*):** regulation, in force
- **European Health Data Space (*EHDS*):** regulation, proposal

Legal Interpretation of the GDPR

Q: “what is *anonymous*?”

Authorities:

initial interpretation:
(often)



highest authority:



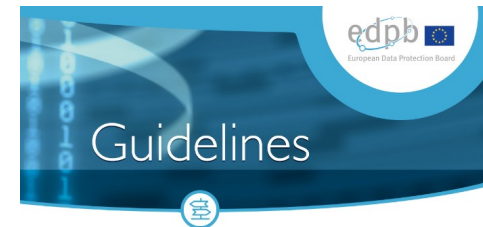
Courts

optional:
for concrete cases

issues



European
Court of Justice
(en: ECJ; de: EuGH)

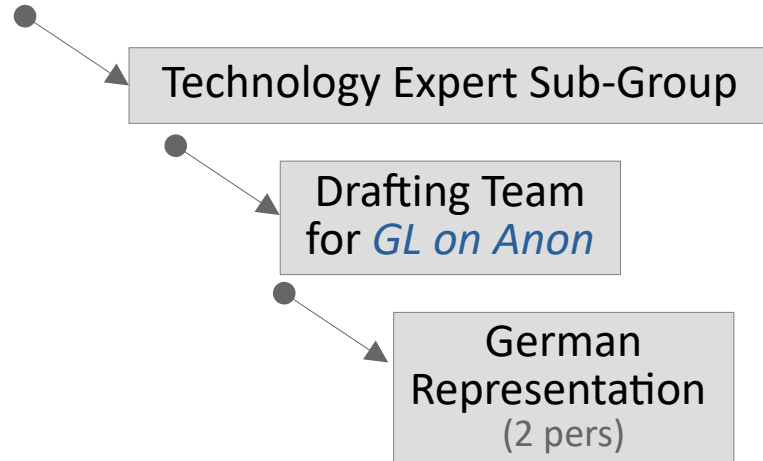


now: working on **Anonymisation**

Process: EDPB GL on Anonymization



27 natl. Supervisory Authorities + EDPS



Process: German Participation



27 natl. Supervisory Authorities + EDPS



16 Länder + Bund

DSK
DatenSchutzKonferenz

Technology Expert Sub-Group

Drafting Team
for *GL on Anon*

German
Representation
(2 pers)

Arbeits Kreis Technik

Feedback Team

VidConfs: feedback
on **advanced** drafts

Process: ULD Participation



27 natl. Supervisory Authorities + EDPS



16 Länder + Bund

DSK
DatenSchutzKonferenz

Technology Expert Sub-Group

Drafting Team
for *GL on Anon*

German
Representation
(2 pers)

Arbeits Kreis Technik

 Schleswig-Holstein

Feedback Team

ULD 

VidConfs: feedback
on *advanced drafts*

Process: EDPB GL on Anonymization



27 natl. Supervisory Authorities + EDPS

Technology Expert Sub-Group

Drafting Team for *GL on Anon*

German Representation (2 pers)

VidConfs: feedback on *advanced drafts*



16 Länder + Bund

DSK DatenSchutzKonferenz

Arbeits Kreis Technik

Schleswig-Holstein

Feedback Team

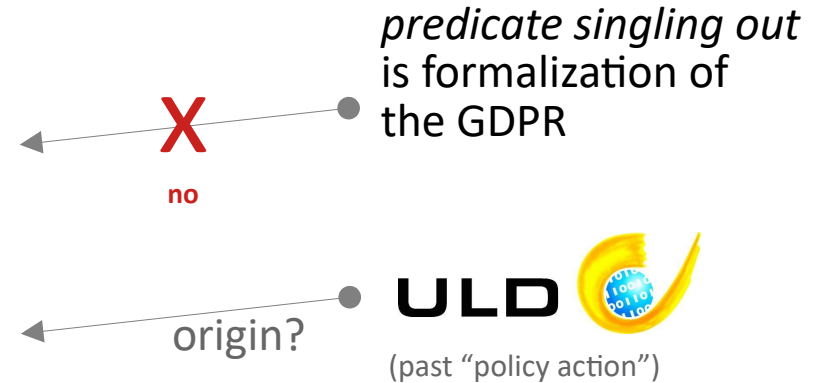


$1 / ((27 + 1) / (16 + 1)) = 0.2\% \text{ influence}$
 M.S. Länder

Analysis: EDPB GL on Anonymization

ULD: Access to Confidential Draft:

- **K-Anonymity is here to stay**
 - singling out possible but still anonymous
- **“supposedly anonymous”**
 - unexpected re-identification = data breach
 - risk made explicit
- **legal (not technical) interpretation**
 - when exactly is something anonymous?
 - how exactly to assess the risk of re-identification?
- **~ 70 pages, complex**



technical interpretation
of
legal interpretation?

Analysis: EDPB GL on Anonymization

ULD: Access to Confidential Draft:

- **K-Anonymity is here to stay**
 - singling out possible but still anonymous
- **legal (not technical) interpretation**
 - when exactly is something anonymous?
 - how exactly to assess the risk of re-identification?
- **~ 70 pages**

predicate singling out is formalization of the GDPR

X
no

technical interpretation
of
legal interpretation?

Analysis: DGA & EHDS

- **Binary Notion of Anonymity:**
 - [yes / no]
 - max 3 states of data: [personal | pseudonymous | anonymous]
- **Risk often “abstracted away”**
 - risk of re-identification acknowledged, **BUT:**
 - anonymize → publish → forget
 - unclear: who validates anonymization? How?
- **No awareness of Privacy Budget**
 - large-scale mandatory publishing of “anonymized data”
 - unclear what happens if re-identified at large scale
 - **ticking bomb?**



Inuit have many words for snow!

Needed Messages (subset)

- “*anonymization*” does not always result in *anonymous*
 - “anonymization” → “identity-reduction”
- Not all “anonymization” techniques are equal
 - render graphically visible via *scope*
- The outcome of “anonymization” may not be decidable
 - success state
 - result states to express uncertainty of success
- Analysis needs to consider *multiple disclosures*
 - render graphically visible
- Pseudonyms should not be shared globally
 - concept: “*pseudonym domain*”
 - concept: “*2nd-level pseudonymization*”

Outline Module 3

Pseudo/Anon Terminology

- Technology Transfer
- Conceptualization and Terminologies
- Analysis of Legal Texts → Messaging Needs: Tech → Policy
- **Messaging (impact) Requirements**
- Resulting Terminologies
- Dissemination

Messaging Requirements

– **Audience:** busy non-technical people

- short, TL;DR
- attractive
- simple

how would you explain
it to a 5 year old?

– **Compatible w/ unavoidable Terminologies**

- laws
- EDPB GL

Outline Module 3

Pseudo/Anon Terminology

- Technology Transfer
- Conceptualization and Terminologies
- Analysis of Legal Texts → Messaging Needs: Tech → Policy
- Messaging (impact) Requirements
- **Resulting Terminologies**
- Dissemination

Result Overview

- **2 Terminologies**
 - **identity-reduction**
 - 3 concept pages
 - 1 glossary page (term → definition)
 - 2 sheet handout: front back
 - **pseudonymization**
 - glossary (term → definition)
 - 3 illustrating figures
 - short (1 sheet) and long version (2 pages)

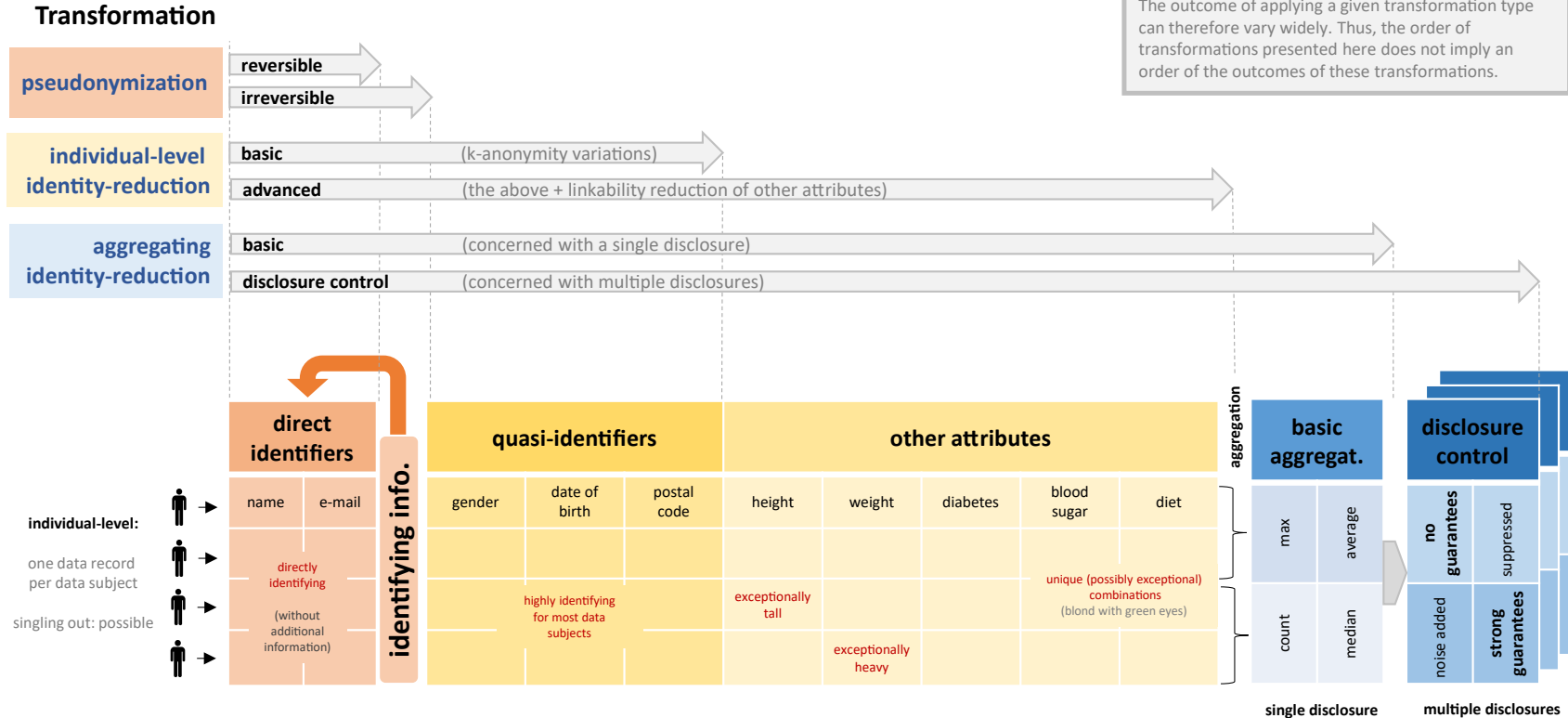
1 The Scope of Identity-Reduction Transformations

Disclaimer:

This taxonomy cannot attempt to answer the question of when data can be considered to be anonymous.

This depends on the data, on the parameters of the transformations, on the available additional information, the state of the art of re-identification, the motivation and resources of possible attackers, ...

The outcome of applying a given transformation type can therefore vary widely. Thus, the order of transformations presented here does not imply an order of the outcomes of these transformations.



1 The Scope of Identity-Reduction Transformations

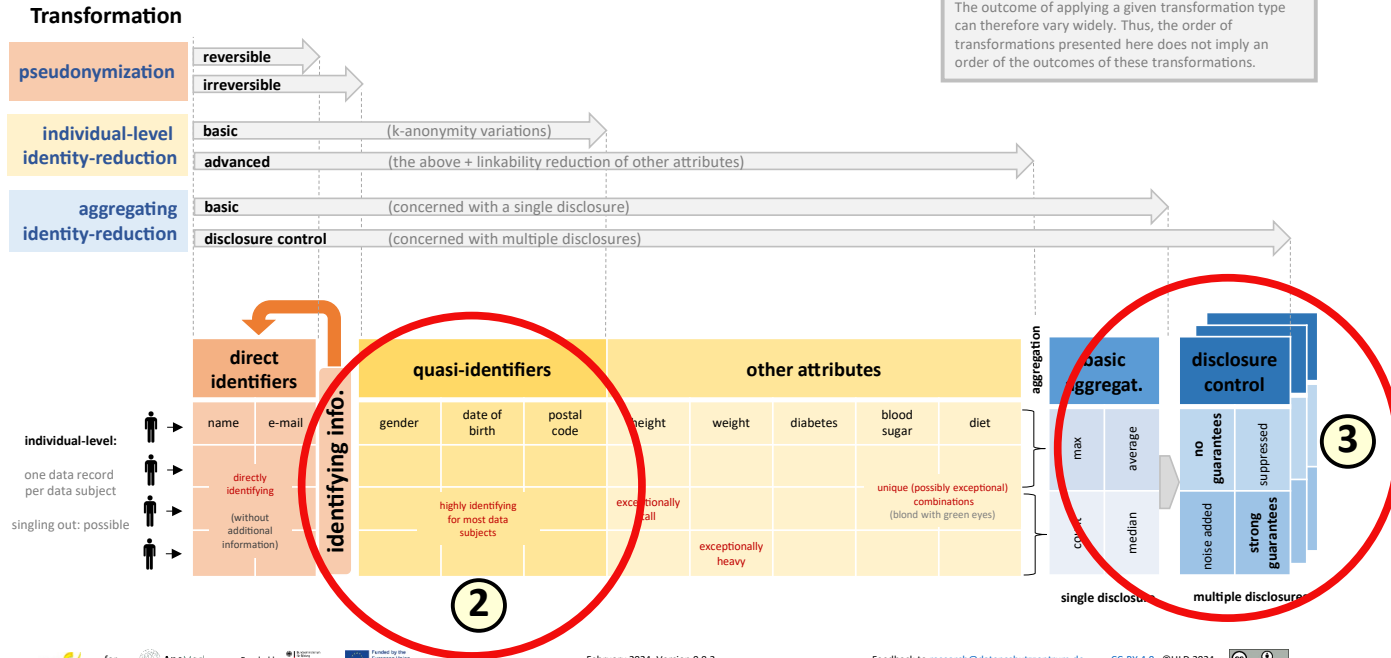
1

Disclaimer:
 This taxonomy cannot attempt to answer the question of when data can be considered to be anonymous.
 This depends on the data, on the parameters of the transformations, on the available additional information, the state of the art of re-identification, the motivation and resources of possible attackers, ...
 The outcome of applying a given transformation type can therefore vary widely. Thus, the order of transformations presented here does not imply an order of the outcomes of these transformations.

1) “anonymization” result not always anonymous

2) “scope” can be a subset of data
 → risk becomes evident

3) scope can be **multiple disclosures:**
 (privacy budget)



2

3

2 A Taxonomy of Identity-Reduction Transformations

Identity Reduction Type		Transformation of Data Elements	Re-Identification Attacks	Possible Outcomes
data pseudonymization	reversible	Direct identifiers are eliminated or transformed (but identifying information is kept)	<ul style="list-style-type: none"> Spontaneous recognition Linkage on: <ul style="list-style-type: none"> Inversion secret quasi-identifiers Unique combinations of other attributes (individ.-level: singling out is trivial)	<i>Pseudonymous Data</i>
	irreversible	In addition: Identifying information is eliminated	Same as above, minus: linkage on inversion secret (individ.-level: singling out is trivial)	<i>Pseudonymous Data</i>
individual-level identity-reduction (aka. <i>record-level, micro data</i>)	basic	In addition: Quasi-identifiers are transformed such that for each possible tuple of quasi-identifiers, there are at least K-1 tuples with undistinguishable values <ul style="list-style-type: none"> Distinction is based on equality or similarity (depending on variance of the quasi-identifiers) Transformations include generalization and suppression 	Same as above, minus: Linkage on quasi-identifiers (individ.-level: singling out is trivial)	<i>Advanced Pseudonymous Data</i> <i>Supposedly Anonymous Data</i>
	advanced	In addition: Other attributes are transformed to protect against linkage <ul style="list-style-type: none"> Transformations include generalization, suppression, top- and bottom-coding, slicing, -data swapping, and noise injection 	Same as above, but: Spontaneous Recognition and linkage on other attributes is rendered more difficult or impossible (individ.-level: singling out is trivial)	<i>Advanced Pseudonymous Data</i> <i>Supposedly Anonymous Data</i> <i>Successfully Anonymous Data</i>
aggregating identity-reduction	basic	For a single disclosure, all individual-level data is transformed such that the resulting values relate to groups of at least C persons	Singling out (followed by linking) possible by inference over multiple disclosures. (reconstruction attacks [1])	<i>Advanced Pseudonymous Data</i> <i>Supposedly Anonymous Data</i> <i>Successfully Anonymous Data</i>
	disclosure control see Art. 2(4) Commission Regulation 557/2013	In addition: The aggregate values are further protected against known or even arbitrary singling out attacks across multiple disclosures.	Singling out over multiple disclosures is rendered difficult or impossible.	<i>Supposedly Anonymous Data</i> <i>Successfully Anonymous Data</i>

2 A Taxonomy of Identity-Reduction Transformations

colors:
link to scope page

taxonomy

terms to distinguish
different
transformations

Identity Reduction Type	Transformation of Data Elements	Re-identification Attacks	Possible Outcomes
data pseudonymization	reversible Direct identifiers are eliminated or transformed (but identifying information is kept)	Spontaneous recognition Linkage on: • Inversion secret • quasi-identifiers • Unique combinations of other attributes (individ.-level: singling out is trivial)	Pseudonymous Data
	irreversible In addition: Identifying information is eliminated	Same as above, minus: linkage on inversion secret (individ.-level: singling out is trivial)	Pseudonymous Data
individual-level identity-reduction (aka. record-level, micro data)	basic In addition: Quasi-identifiers are transformed such that for each possible tuple of quasi-identifiers, there are at least K-1 tuples with undistinguishable values • Distinction is based on equality or similarity (depending on variance of the quasi-identifiers) • Transformations include generalization and suppression	Same as above, minus: Linkage on quasi-identifiers (individ.-level: singling out is trivial)	Pseudonymous Data Advanced Pseudonymous Data Supposedly Anonymous Data
	advanced In addition: Other attributes are transformed to protect against linkage • Transformations include generalization, suppression, top-and-bottom-coding, slicing, data swapping, and noise injection	Same as above, but: Spontaneous Recognition and linkage on other attributes is rendered more difficult or impossible (individ.-level: singling out is trivial)	Advanced Pseudonymous Data Supposedly Anonymous Data Successfully Anonymous Data
aggregating identity-reduction	basic For a single disclosure, all individual-level data is transformed such that the resulting values relate to group of at least C persons	Singling out (followed by linking possible by inference over multiple disclosures. (reconstruction attacks [2]))	Advanced Pseudonymous Data Supposedly Anonymous Data Successfully Anonymous Data
	disclosure control In addition: The aggregate values are further protected against known or even arbitrary singling out attacks across multiple disclosures.	Singling out over multiple disclosures is rendered difficult or impossible.	Supposedly Anonymous Data Successfully Anonymous Data

link to outcomes

“anonymization”
doesn't necessarily
result in anonymous
data

minimal
explanation

re-identification risk
rendered visible

3 Categories of Data

Possible Outcomes of Identity-Reduction Transformations

Disclaimer:

The data category cannot be determined from the data alone.

While there are indicators for data being personal, no technical test exists that guarantees anonymity. Data categories are therefore the result of a risk assessment which takes factors beyond just the data into account.

Data Category

Possibilities of (Re-)Identification

Fully Identified Personal Data

- **direct identification** is possible (since data is unchanged)

(Basic) Pseudonymous Data

personal data (Recital 26 GDPR)

- direct identification is no longer possible
- **only indirect identification** using **additional information** is possible

Advanced Pseudonymous Data

likely still personal data

- direct identification is no longer possible
- **even indirect identification** is rendered **difficult** or **prevented** (but with unknown success)

Supposedly Anonymous Data

likely anonymous

but future re-identification cannot be excluded

- **all relevant known re-identification attacks are excluded**
- **thorough assessment of re-identification risk** results in low risk

Successfully Anonymous Data

certainly anonymous

future practical re-identification can be excluded

- **re-identification can be practically^[1] excluded**
- strong guarantees or thorough assessment of re-identification risk

[1] *practically* here means considering any party who can reasonably likely gain access to the data, its reasonably likely means, and taking into account technological developments.



More words for snow!

5 states of data (up from 3)

uncertainty of “anonymization” success

compatibility with EDPB GL
 ”supposedly anonymous”

Identity-Reduction: The Technical Perspective

3 Categories of Data

Possible Outcomes of Identity-Reduction Transformations

Disclaimer:

The data category cannot be determined from the data alone.

While there are indicators for data being personal, no technical test exists that guarantees anonymity. Data categories are therefore the result of a risk assessment which takes factors beyond just the data into account.

Data Category

Possibilities of (Re-)Identification

Fully Identified Personal Data	<ul style="list-style-type: none"> • direct identification is possible (since data is unchanged)
<i>(Basic) Pseudonymous Data</i> personal data (Recital 26 GDPR)	<ul style="list-style-type: none"> • direct identification is no longer possible • only indirect identification using additional information is possible
Advanced Pseudonymous Data likely still personal data	<ul style="list-style-type: none"> • direct identification is no longer possible • even indirect identification is rendered difficult or prevented (but with unknown success)
Supposedly Anonymous Data likely anonymous but future re-identification cannot be excluded	<ul style="list-style-type: none"> • all relevant known re-identification attacks are excluded • thorough assessment of re-identification risk results in low risk
Successfully Anonymous Data certainly anonymous future practical re-identification can be excluded	<ul style="list-style-type: none"> • re-identification can be practically^[1] excluded • strong guarantees or thorough assessment of re-identification risk

[1] practically here means considering any party who can reasonably likely gain access to the data, its reasonably likely means, and taking into account technological developments.

terms that are used on other slides

Direct Identifier: A direct identifier is a value or value combination that is commonly known to be related to a given natural person or where a known procedure of limited effort can be used to establish such a relation. Direct Identifiers are often unique in a given context. Examples include a person's name, address, phone number, coordinates of residence, etc.

Relation to a natural person: A value is related to a natural person if, with a significant likelihood, the person has (positive relation) or has not (negative relation) a certain property described by that value.

Quasi-Identifier: A quasi-identifier is a value that is expected to be known about a natural person or easy to find out. Combinations of quasi-identifiers are often unique for a majority of persons. Examples include age, gender, and place of birth.

Singling Out: Singling out is a processing step executed on a data set that, for at least one data subject, results in some data value that is related to a (possibly unknown) person. Such processing can be a trivial lookup in the data set or require sophisticated inference that possibly uses additional information. Singling out through inference can also require the combination of multiple data sets as for example used in reconstruction attacks of statistical data [↔].

Inference: Inference is the process of deriving information from a data set that is not evident. Inference typically applies knowledge of functional dependencies between values, known correlations, known probability distributions, or other dependencies of values that can be expressed with models (including machine learning models). Types of inference include *attribute inference* where the result of the inference are new values that are related to the same data subject, and *membership inference* where, based on some known values of a person, it can be established that this person is indeed a data subject.

Linkage: Linkage is the process of establishing a relation between a singled-out value and an actual natural person. Simple forms of linkage *match* combinations of values of the data set with an external data set that contains direct identifiers. More sophisticated forms of linkage match on values derived by inference or use inference without matching. Linkage is only possible if at least one value relating to the data subject can be singled out.

Matching: Matching is a kind of Linkage based on comparison. The comparison can be based on equality of invariant values or the similarity or closeness of values that change.

Spontaneous Recognition: Spontaneous recognition is a kind of Linkage in which a human observer of a data set matches a singled out combination of values to the known values of a familiar person (relative, colleague, acquaintance, etc.). It uses additional information about the data subject that is knowledge much rather than materialized as data.

Aggregation: Aggregation is a mapping from values relating to multiple persons to a value that relates to a group of persons. Examples include statistics, machine learning models, and decision trees.

Generalization: Generalization maps values to a coarser scale of measurement such that the number of possible values is reduced. Examples include re-classification of nominal values and the definition of intervals of ordinal, ratio or interval values. Generalization can involve multiple values as in mapping weight and height into a body mass index or mapping possible coordinates to districts or zones.

Suppression: Suppression eliminates values from the data set. This can be a single (for example exceptional) value, all values (i.e., a record) of a given data subject, or an attribute for all data subjects.

Top- and Bottom-Coding: Top- and Bottom-Coding is a transformation in which all values above or below a certain threshold are mapped to the same output value that represents (e.g., "above 220 cm")

Noise Injection: Noise injection is a transformation that adds random noise to data values.

Slicing: Slicing is a transformation that splits a high-dimensional data set into multiple lower-dimensional ones.

Data swapping: Data swapping is a transformation in which values belonging to different data subjects (typically belonging to some group) are swapped.

Pseudonymization Terminology for Policy Makers

Version 0.19

pseudonymization: (defined in Art. 4(5) GDPR)
A manner of processing in which directly identifying data elements (additional information) are kept separate and protected against unauthorized use in order to prevent the identification of data subjects during the processing of pseudonymous data.

data pseudonymization:
Data pseudonymization is a transformation of fully identified personal data that separates pseudonymous data and identifying information

pseudonymous data:
Pseudonymous data is personal data in which data subjects cannot be identified without the use of additional information; identifying data that results from data pseudonymization is one kind of additional information.

additional information: (defined in Art. 4(5) GDPR)
Additional information is any information suited to be combined (typically by linked) with pseudonymized data in order to identify (at least some) data subjects. One kind of additional information is the identifying information that results from data pseudonymization; other kinds of suitable additional information can exist and be held, either by the controller or by external parties.

identifying information:
Identifying information is a kind of additional information that is the result of data pseudonymization and is kept separately and protected during pseudonymization. It permits to establish a one- or bi-directional relation between fully identifying data elements and the pseudonyms used in pseudonymous data.

pseudonymization reversal information: (used in Art. 44(3) EHDS)
Pseudonymization reversal information is bi-directional identifying information that permits to map from pseudonyms to fully identifying data elements.

pseudonymization reversal: (used in Art. 44(3) EHDS)
Pseudonymization reversal is the inverse of data pseudonymization that maps pseudonymous data plus pseudonymization reversal information to fully identified personal data.

reversible pseudonymization:
Reversible pseudonymization is pseudonymization in which the pseudonymization reversal information is kept available to enable a full or partial pseudonymization reversal.

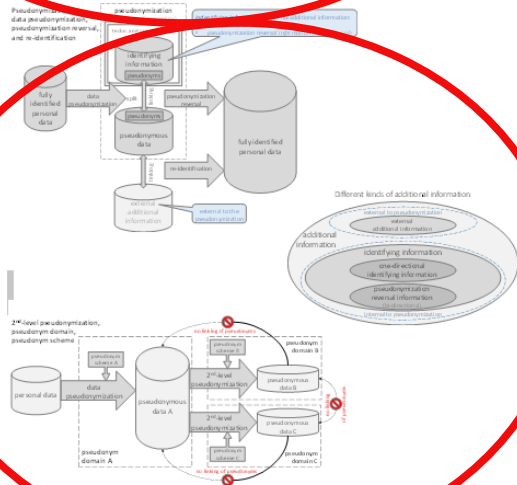
irreversible pseudonymization:
Irreversible pseudonymization is pseudonymization in which the pseudonymization reversal information is not or no longer kept such that the controller is unable to perform a full or partial pseudonymization reversal.

pseudonym:
A pseudonym is a handle for data subjects used on both the pseudonymous data and the identifying information.

pseudonym scheme:
A pseudonym scheme is the manner in which pseudonyms are created during data pseudonymization.

pseudonym domain: The context in which a single pseudonym scheme is applied and consequently each data subject is identified by a unique pseudonym that allows linking of data elements belonging to the same data subject.

2nd-level pseudonymization:
2nd-level pseudonymization is a transformation that replaces the pseudonyms in pseudonymous data with newly created ones. It uses a separate pseudonym scheme to create a distinct ("unlinkable") pseudonym domain.



“pseudonymization domain”

“2nd level pseudonymization”

EHDS: same pseudonym for all users?

Visualize concepts

compatibility with:

- EDPB GL on Pseudonymization (draft)
- EHDS

Outline Module 3

Pseudo/Anon Terminology

- Technology Transfer
- Conceptualization and Terminologies
- Analysis of Legal Texts → Messaging Needs: Tech → Policy
- Messaging (impact) Requirements
- Resulting Terminologies
- **Dissemination**

Dissemination Strategy

- **Target Audiences:**

- **EDPB Drafting Team for Anon Guidelines**
- **EC / policy makers for Data Spaces**



- **ULD official Approval**

- **Engage a few champions**

- **feedback on draft** → they take ownership
- **indirect distribution**

- **Unique Selling Point:**

- simple, small, attractive
- competes with EDPB GL: very legal complex, long

Identity-Reduction:

status:

- draft sent to 7 Champions
- feedback from 3 (2 natl. DPAs)
- feedback from 1 DPA actuatable
- fixed new version
- awaits ULD approval

Conclusions Module 3

- **Technology transfer towards policy world is necessary to make our work relevant**
- **Technology transfer starts with concepts**
- **A terminology can transfer concepts**
- **A suited terminology has been developed in *AnoMed***
- **Dissemination through selected champions has started**
- **time will show..**