Provenance & Privacy

Modeling Decisions for Artificial Intelligence 2017
MDAI 2017

Oct 18, 2017
Kitakyushu, Japan

Vicenç Torra ➡ *University of Skövde*
Guillermo Navarro-Arribas ➡ *Universitat Autonoma de Barcelona*
David Sanchez-Charles ➡ *CA Technologies*
Victor Muntés-Mulero ➡ *CA Technologies*

Outline

1. Motivation
2. Data Provenance
3. Provenance and Privacy
4. Conclusions

Some historical / motivational notes

2014, Google Spain vs AEPD (Spanish Agency of Data Protection) and Mario Costeja Gonzalez

> “ … an Internet search engine operator is responsible for the processing that it carries out of personal information which appears on web pages published by third parties.”

✔ Confirms the **right to be forgotten**

The right to be forgotten

"... reflects the claim of an individual to have certain data deleted so that third persons can no longer trace them."

Rolf H. Weber, The Right to Be Forgotten More Than a Pandora's Box?, 2 (2011) JIPITEC 120 para 1.

*https://www.jipitec.eu/issues/jipitec-2-2-2011/3084_*

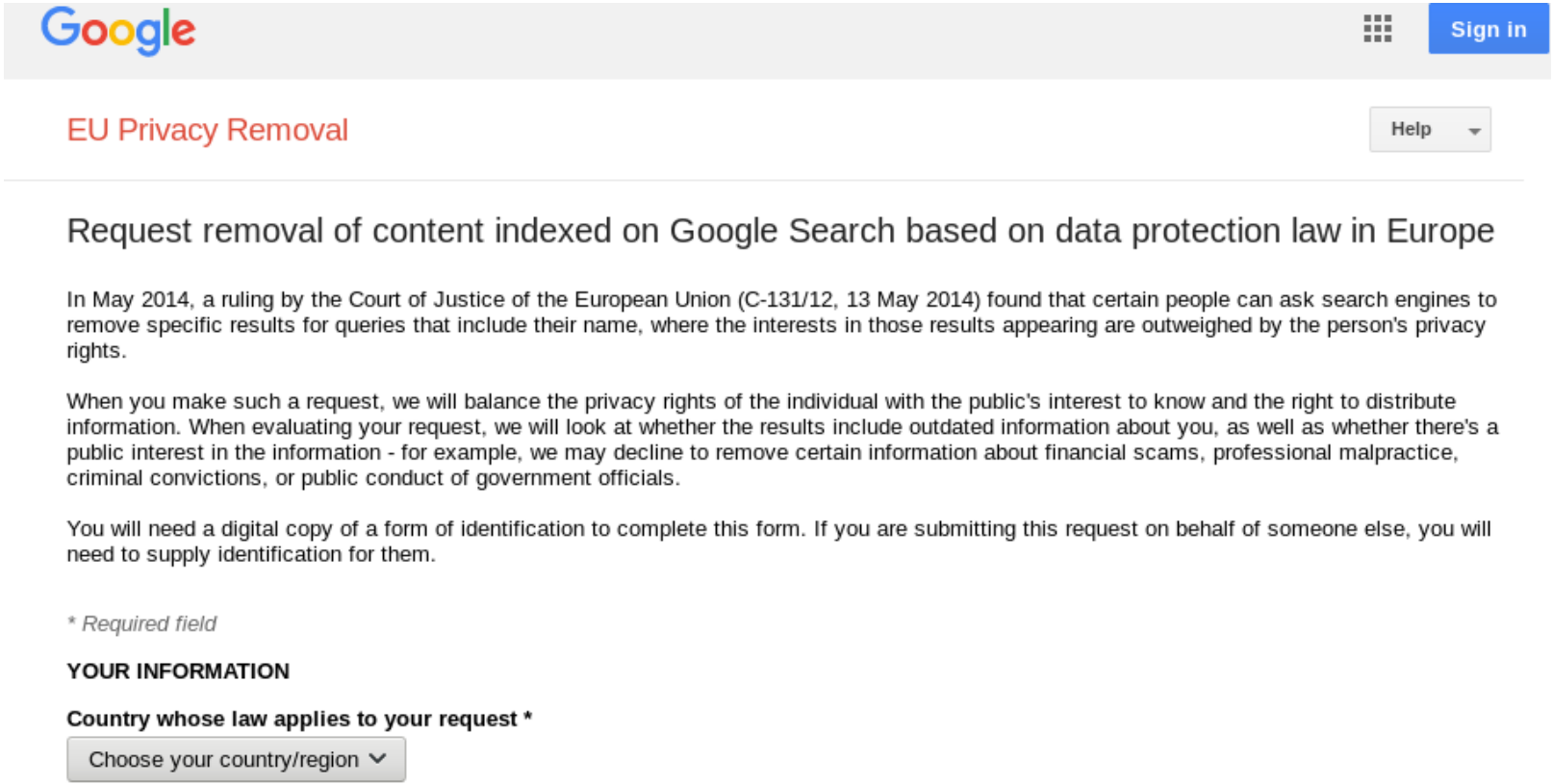
" … the right to silence on past events in life that are no longer occurring."

G. Pino, The Right to Personal Identity in Italian Private Law: Constitutional Interpretation and Judge-Made Rights (2000). THE HARMONIZATION OF PRIVATE LAW IN EUROPE, M. Van Hoecke and F. Ost, eds., Hart Publishing, Oxford, pp. 225-237, 2000. https://ssrn.com/abstract=1737392

Example: Google

Some results might have been eliminated according to the European data protection law.

Right to be forgotten in EU at Google



Google

Sign in

EU Privacy Removal

Help

Request removal of content indexed on Google Search based on data protection law in Europe

In May 2014, a ruling by the Court of Justice of the European Union (C-131/12, 13 May 2014) found that certain people can ask search engines to remove specific results for queries that include their name, where the interests in those results appearing are outweighed by the person's privacy rights.

When you make such a request, we will balance the privacy rights of the individual with the public's interest to know and the right to distribute information. When evaluating your request, we will look at whether the results include outdated information about you, as well as whether there's a public interest in the information - for example, we may decline to remove certain information about financial scams, professional malpractice, criminal convictions, or public conduct of government officials.

You will need a digital copy of a form of identification to complete this form. If you are submitting this request on behalf of someone else, you will need to supply identification for them.

** Required field*

YOUR INFORMATION

Country whose law applies to your request *

Choose your country/region

Right to be forgotten at Google

https://transparencyreport.google.com/eu-privacy/overview

As Oct. 13, 2017:

> 1,900,000 URL removal requests

> 823,000 URLs removed

Sites that are most impacted

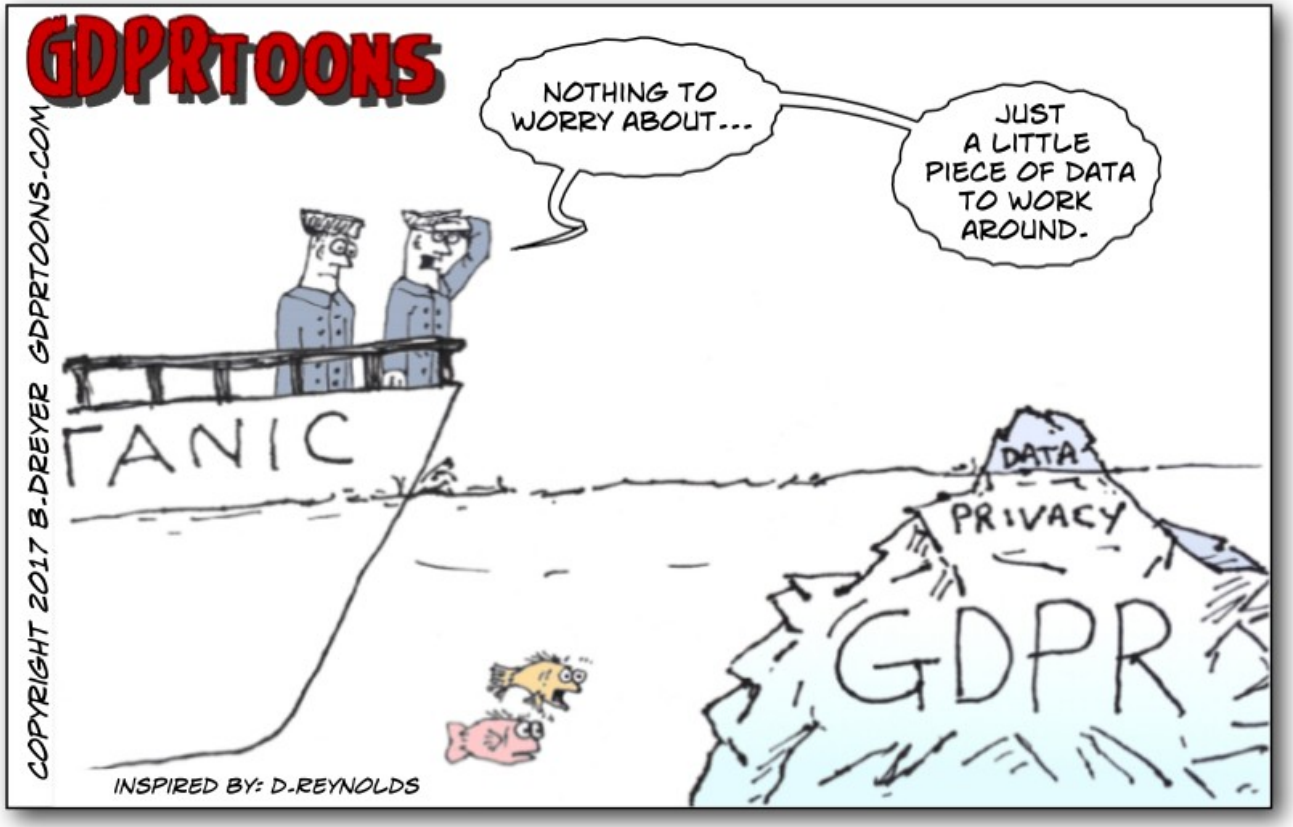
The list below highlights the domains from which we've delisted the most URLs.

Domain	URLs removed	Total URLs requested
www.facebook.com	17,673	41,766
annuaire.118712.fr	11,082	14,864
profileengine.com	11,047	13,185
twitter.com	8,917	21,804
www.youtube.com	8,663	22,855
groups.google.com	8,209	16,455
plus.google.com	7,562	31,041
scontent.cdninstagram.com	5,944	10,639
badoo.com	5,532	10,354
www.wherevent.com	5,439	6,308

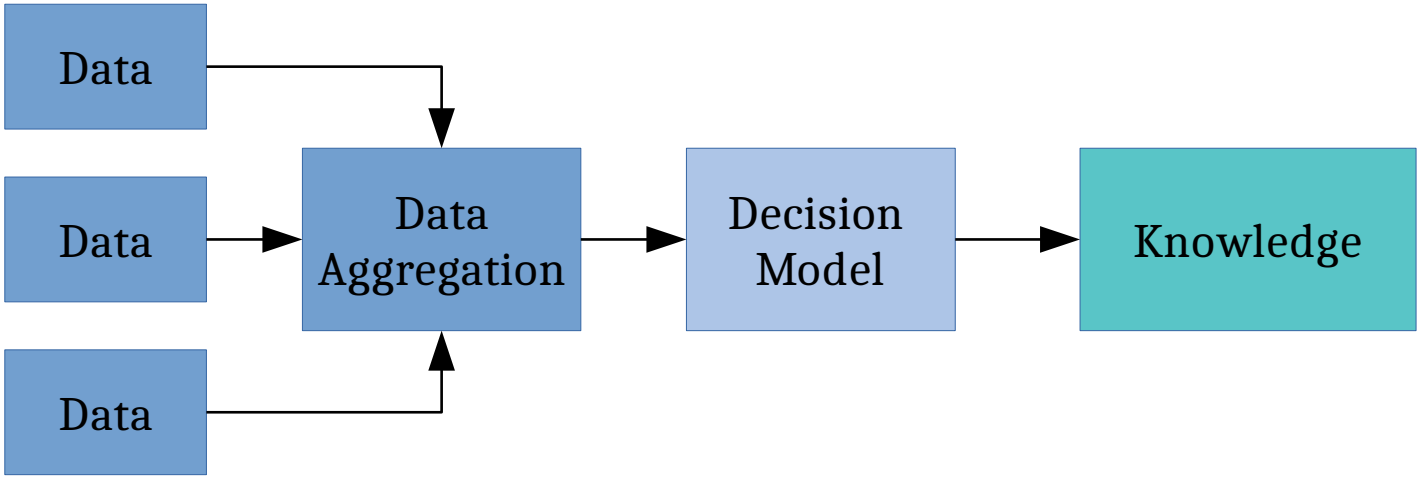
EU General Data Protection Regulation

Adopted in April 27, 2016, and applicable from May 25, 2018.

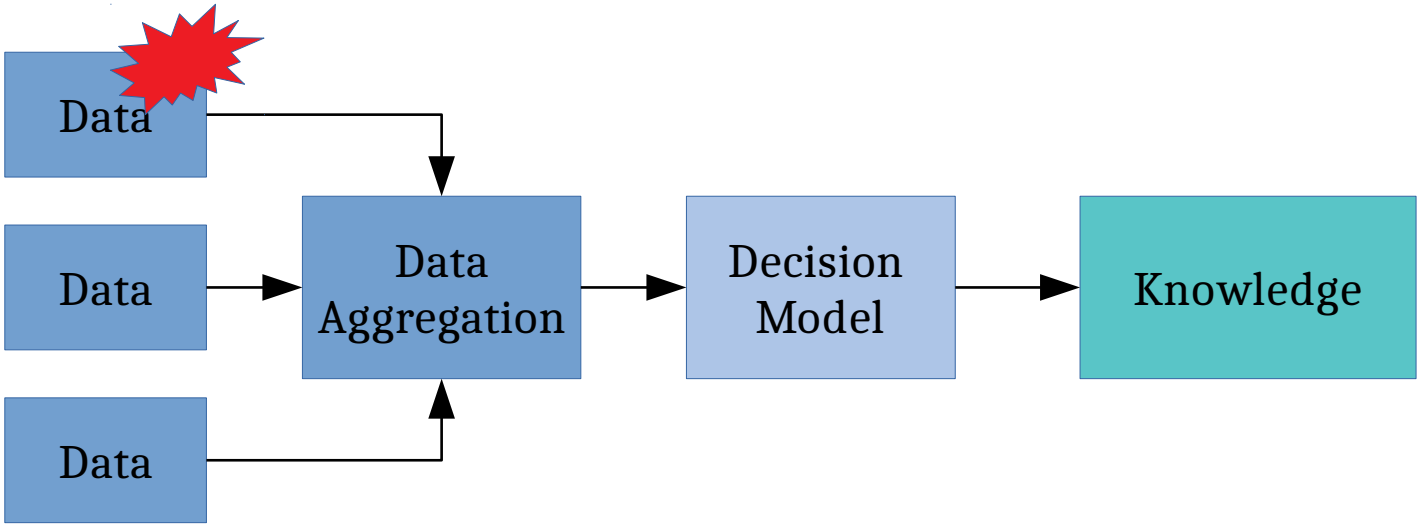
- Right to be forgotten,
- easier access to personal information,
- right to amend your own information.


GDPRTOONS
NOTHING TO WORRY ABOUT...
JUST A LITTLE PIECE OF DATA TO WORK AROUND.
TANIC
DATA
PRIVACY
GDPR
COPYRIGHT 2017 B.DREYER GDPRTOONS.COM
INSPIRED BY: D.REYNOLDS

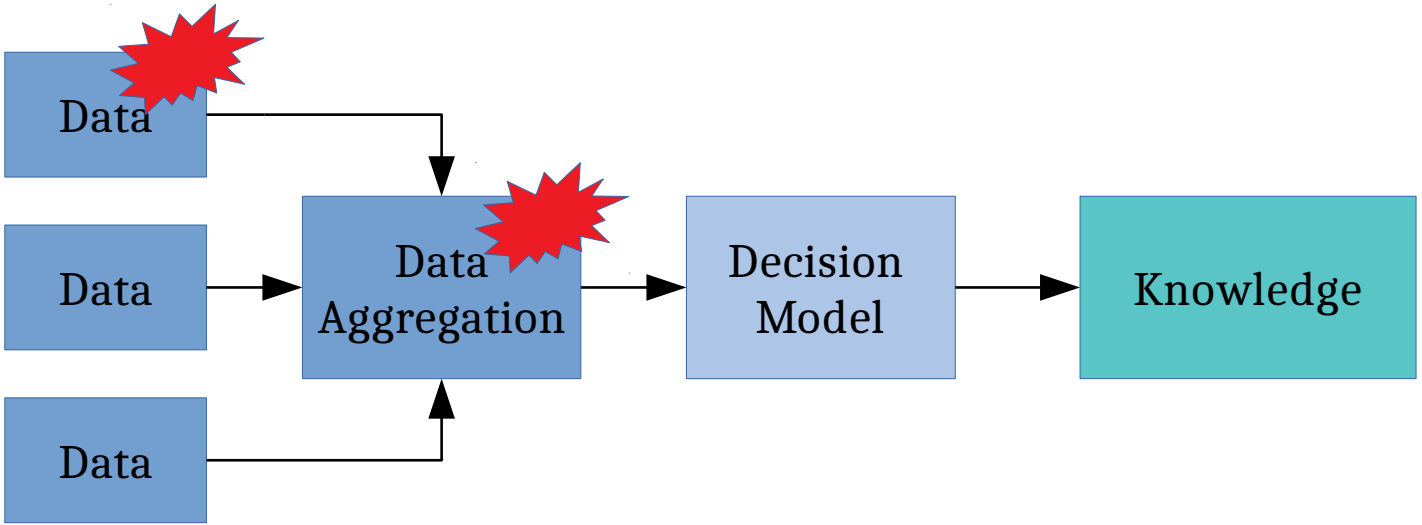
Implications: how do we implement this?



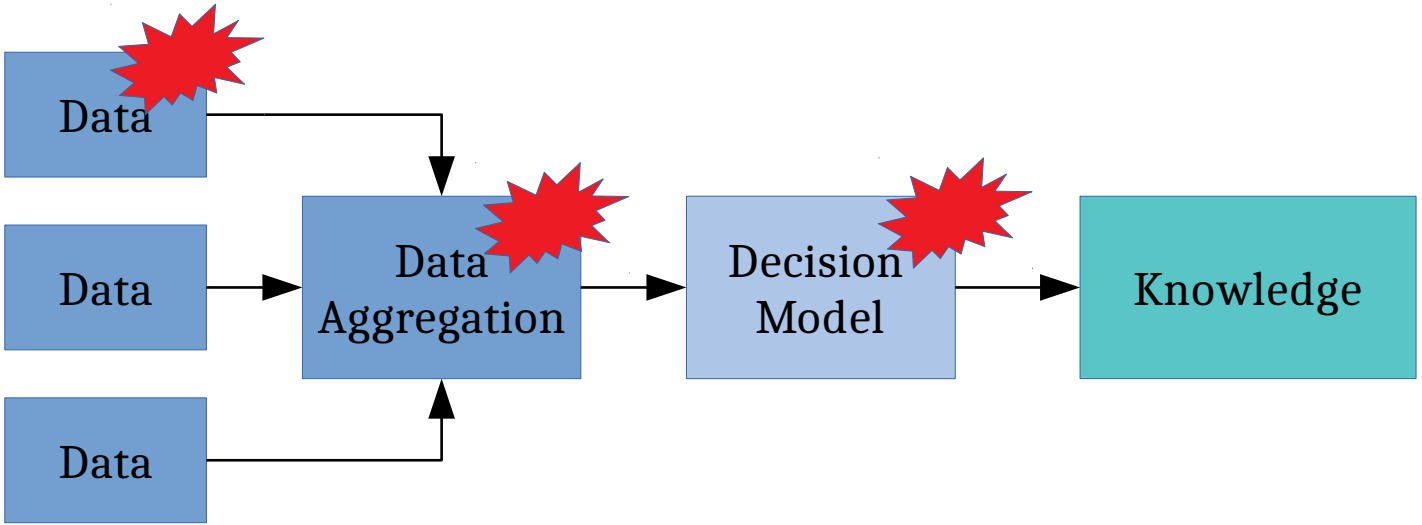
Implications: how do we implement this?



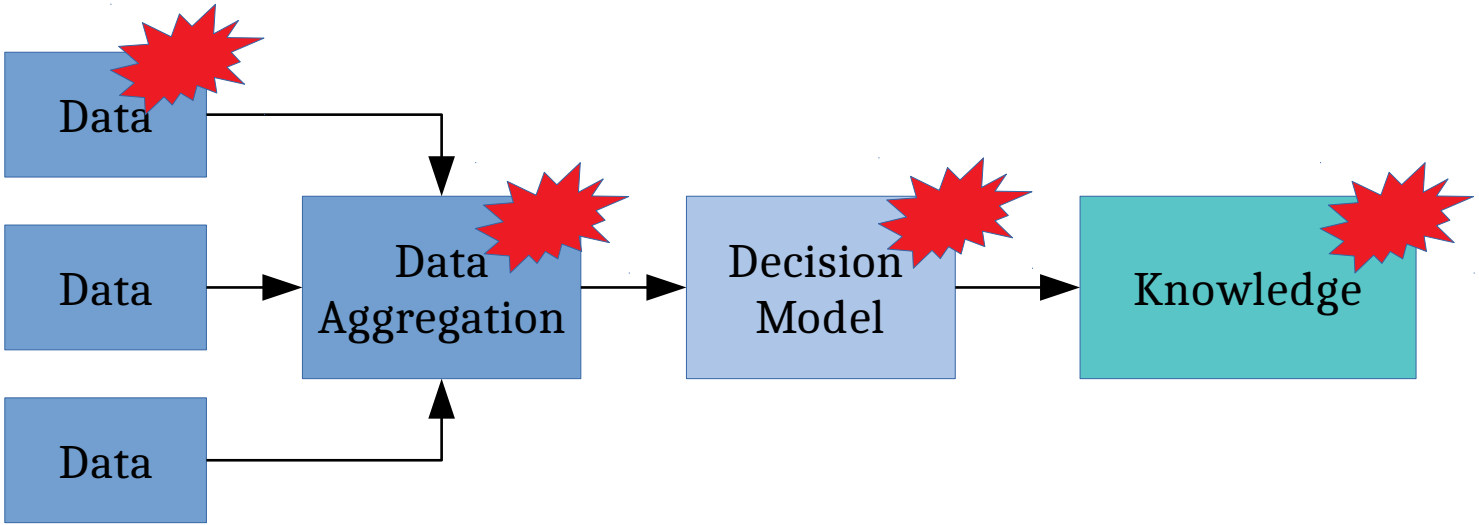
Implications: how do we implement this?



Implications: how do we implement this?



Implications: how do we implement this?



Implications: how do we implement this?

- Personal data is typically aggregated and used in decision models.
- Deletion and amendment may require reconsideration of inferences made from the models, and reconsideration of the knowledge extracted from the data.
- Need to know who contributed to data models and decisions
- Update models and decisions when they are affected by data deletions and updates.

Data Provenance

Data Provenance (or Data Lineage): information about data origin and its processing. An historical record of the inputs, and processes that influenced data.

- Like metadata or annotations.
- Improves data quality and accountability
- **Enables the implementation of the right to be forgotten and the right to amend.**

Data Provenance Types

Common to consider 2 types of provenance data:

- **Where** provenance: origin of the data
  - where it comes from?

- **Why** provenance: process that generated the data
  - why or how it got to the database?

Provenance representation

- Provenance record:

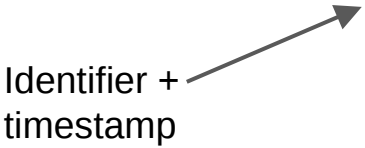
$$(seqID, p, \{(A_1, v_1), \ldots,(A_n, v_n)\}, (A, v))$$

- **Chains**: time-ordered sequence of provenance records (actor, process applied to the data)
- **Graphs**: partially ordered provenance records (as directed acyclic graphs)

Provenance representation

- Provenance record:

$$(seqID, p, \{(A_1, v_1), \ldots,(A_n, v_n)\}, (A, v))$$

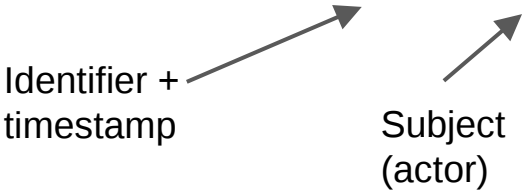


- **Chains**: time-ordered sequence of provenance records (actor, process applied to the data)
- **Graphs**: partially ordered provenance records (as directed acyclic graphs)

Provenance representation

- Provenance record:

$$(seqID, p, \{(A_1, v_1), \ldots,(A_n, v_n)\}, (A, v))$$

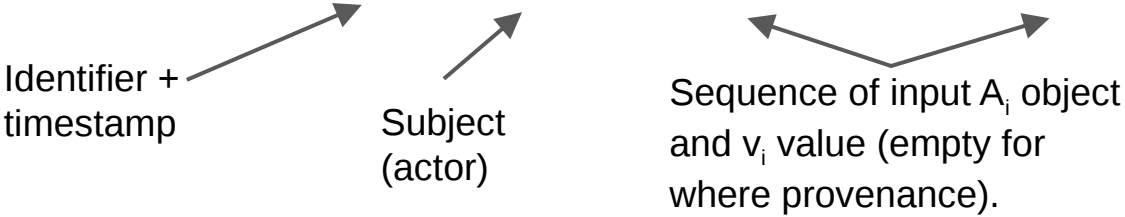


- **Chains**: time-ordered sequence of provenance records (actor, process applied to the data)
- **Graphs**: partially ordered provenance records (as directed acyclic graphs)

Provenance representation

- Provenance record:

$$(seqID, p, \{(A_1, v_1), \ldots,(A_n, v_n)\}, (A, v))$$

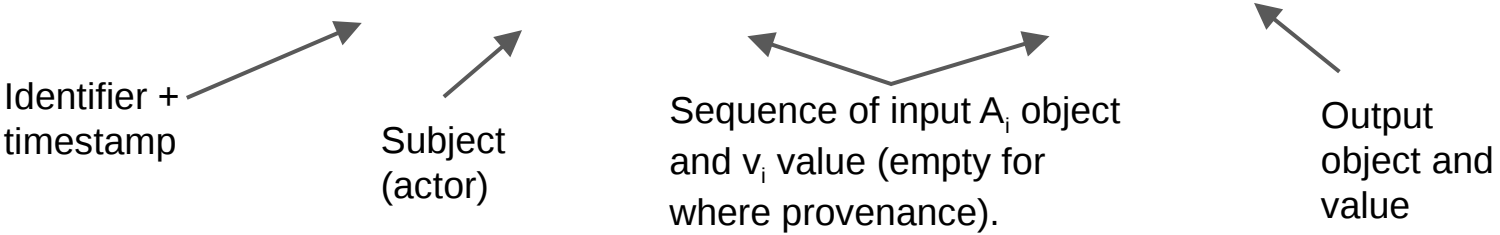


- **Chains**: time-ordered sequence of provenance records (actor, process applied to the data)
- **Graphs**: partially ordered provenance records (as directed acyclic graphs)

Provenance representation

- Provenance record:

$$(seqID, p, \{(A_1, v_1), \ldots,(A_n, v_n)\}, (A, v))$$



- **Chains**: time-ordered sequence of provenance records (actor, process applied to the data)
- **Graphs**: partially ordered provenance records (as directed acyclic graphs)

W3C’s PROV

“ … provenance is defined as a record that describes the people, institutions, entities, and activities involved in producing, influencing, or delivering a piece of data or a thing.”

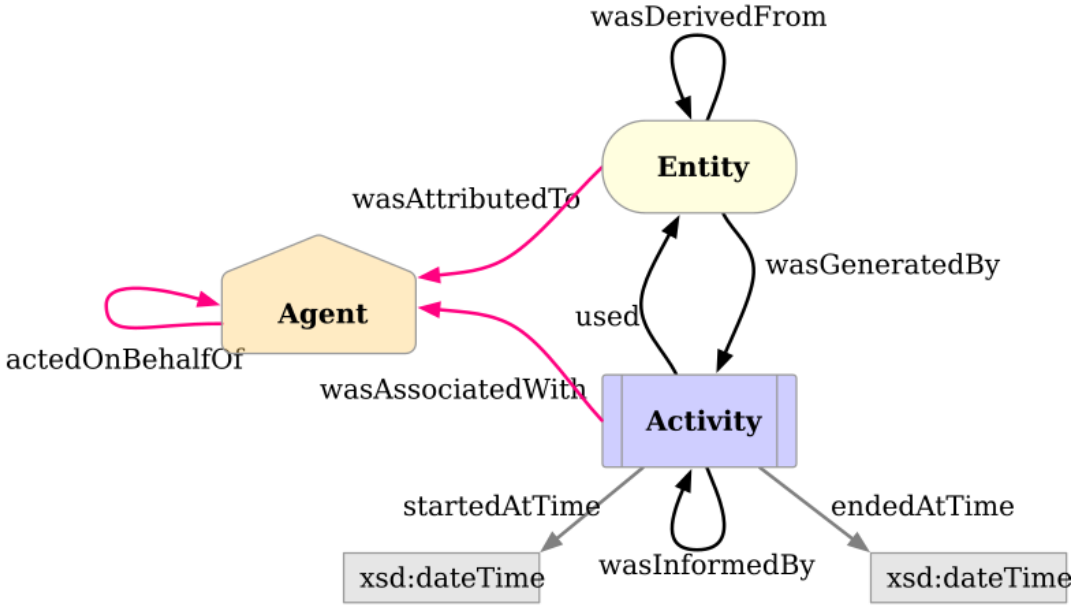
- (Extensive) set of specifications by W3C for the inter-operable interchange of provenance information in heterogeneous environments such as the Web.
- Most accepted standard for provenance information, but
- Actually not widely used!
- Not (yet?) adopted by provenance aware DBs.

PROV-DM

PROV-DM: **The PROV Data Model**

- **Types**: Entity, Activity, Agent

- **Relations**: Generation, Usage, Communications, Derivation, Attribution, Association, Delegation.



Source: https://www.w3.org/TR/prov-o/, PROV-O: The PROV Ontology, W3C Recommendation 30 April 2013

Provenance properties

Classical properties:

- **Completeness**: all actions relevant to computation should be detected and represented.
- **Efficiency**: provenance introduces space and time overhead. Also provenance need to be efficiently computable.

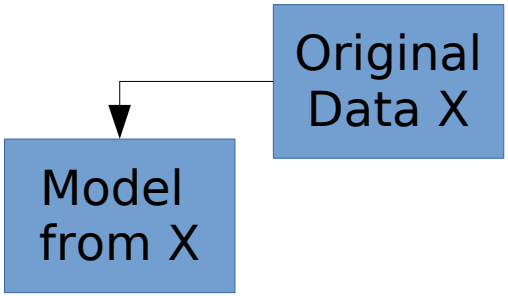
Secure provenance:

Ensure **security and privacy of provenance data** (which is naturally sensitive)

Secure provenance properties

- **Distributed**: data can flow through untrusted environments.
- **Integrity**: prevent forging of data provenance, unauthorized modifications, repudiation, ...
- **Availability**: ensure secure, fast, and reliable access to provenance data.
- **Privacy and confidentiality**: prevent unauthorized disclosure.

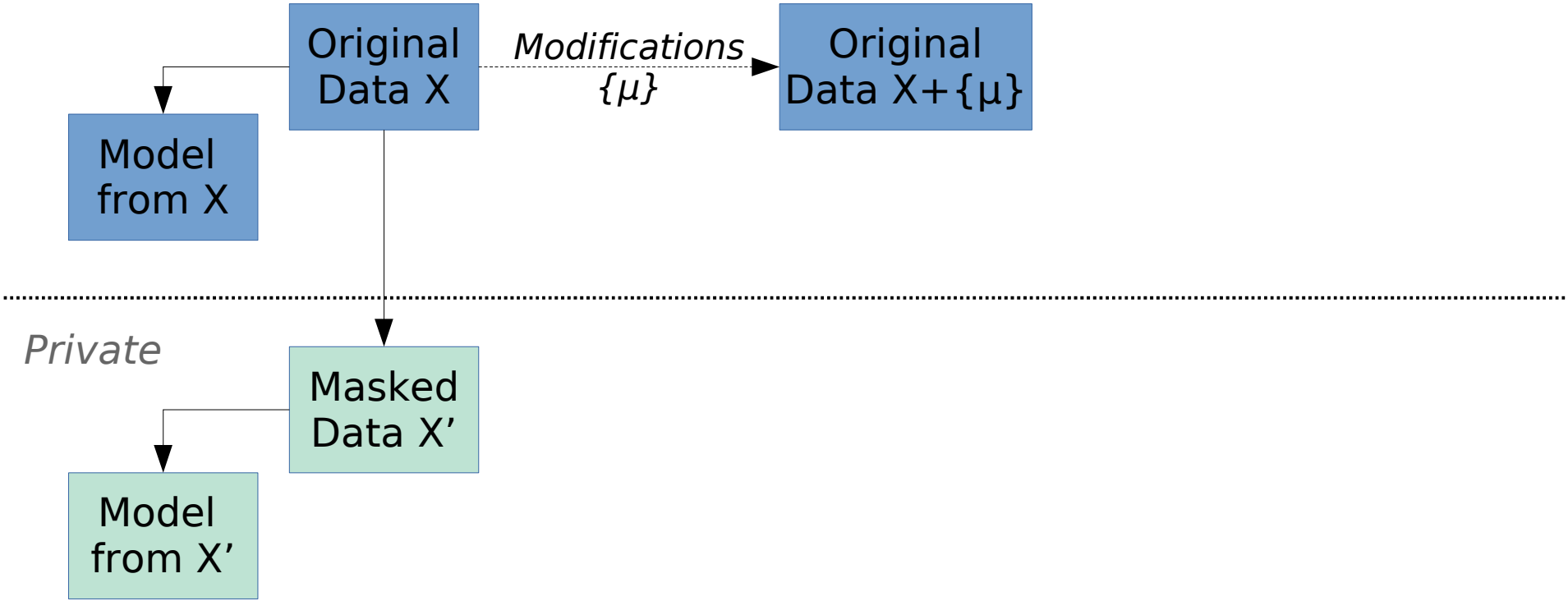
Complex scenario



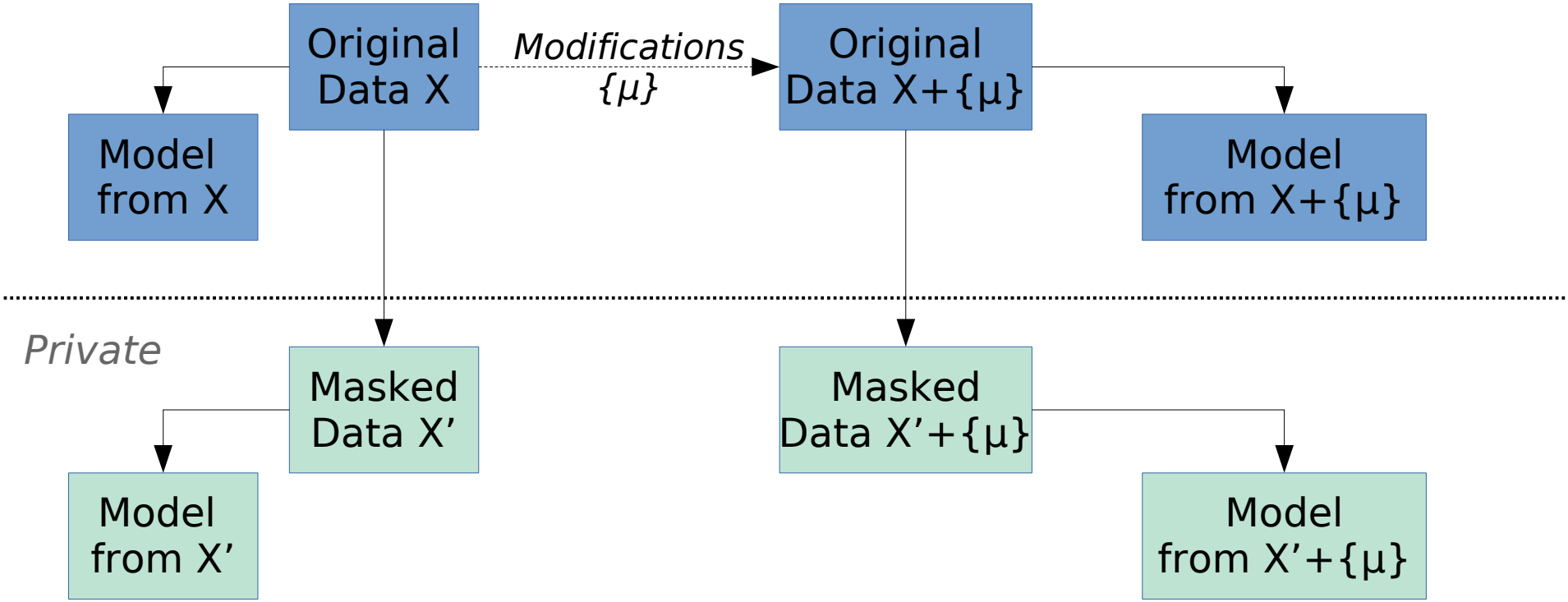
Complex scenario

Complex scenario

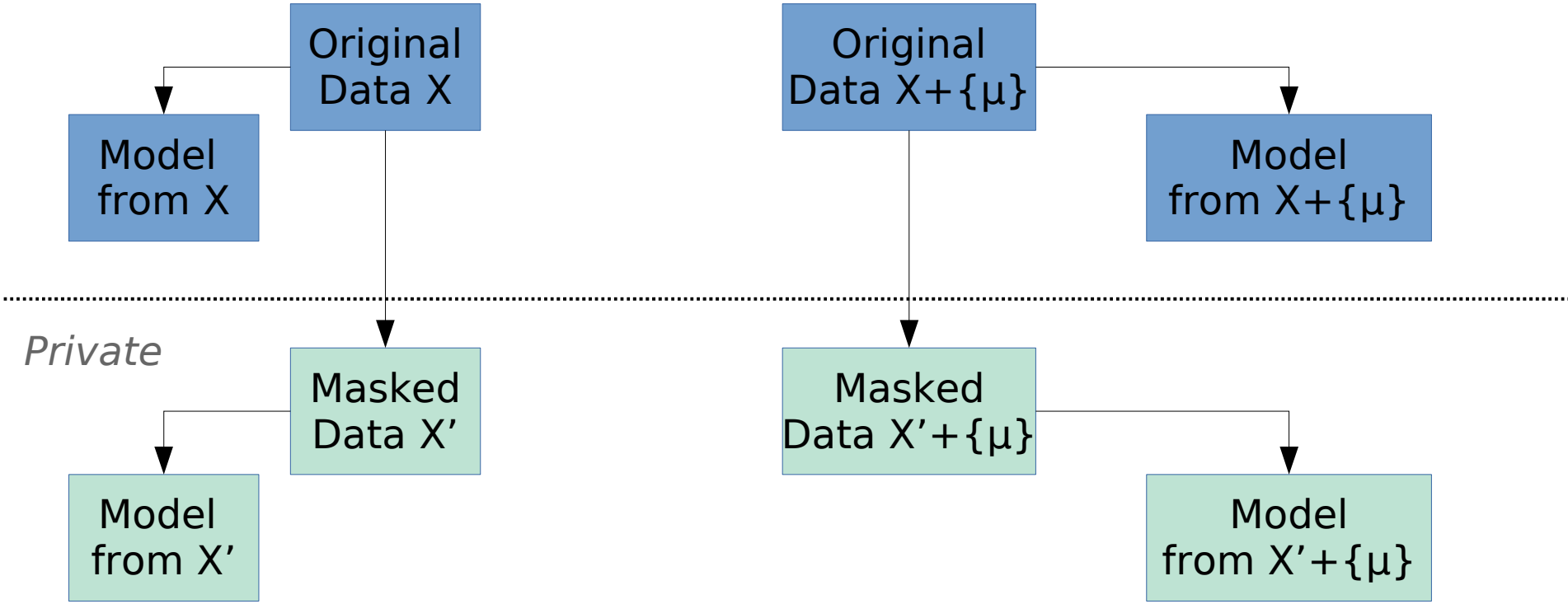
Complex scenario



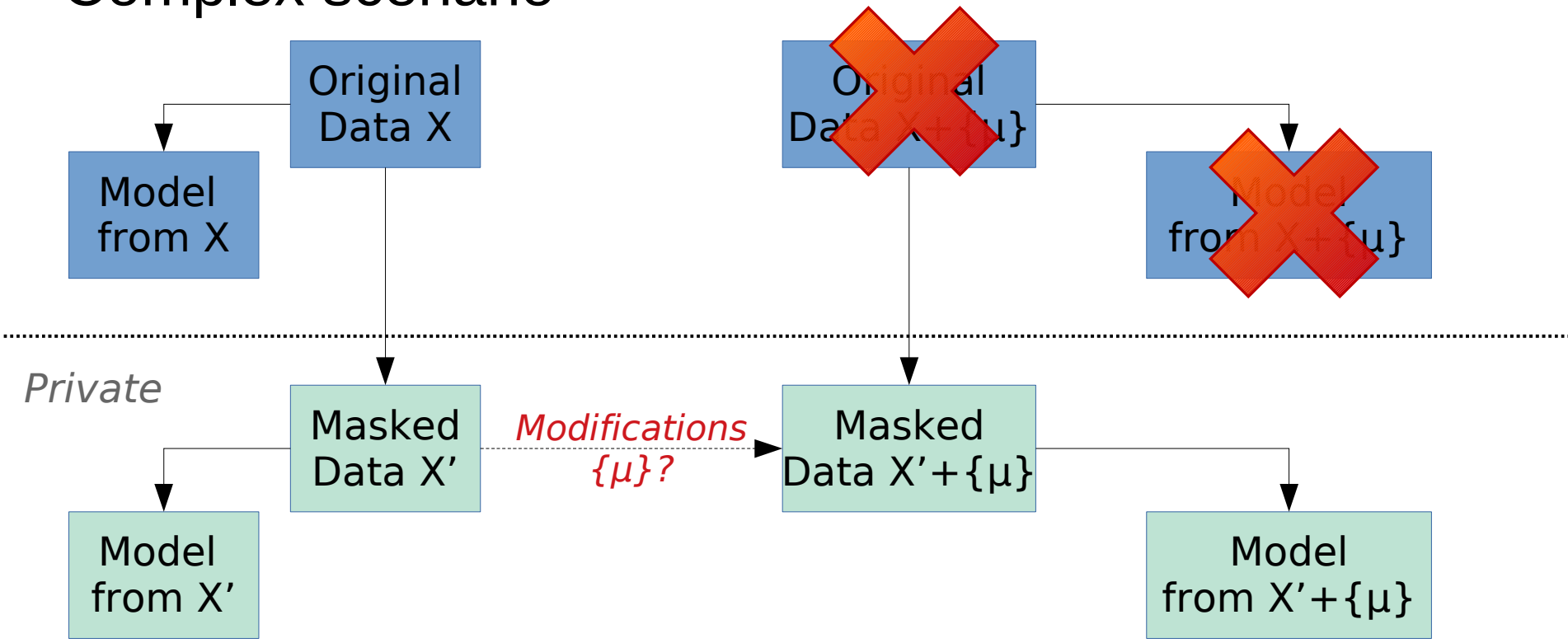
Complex scenario



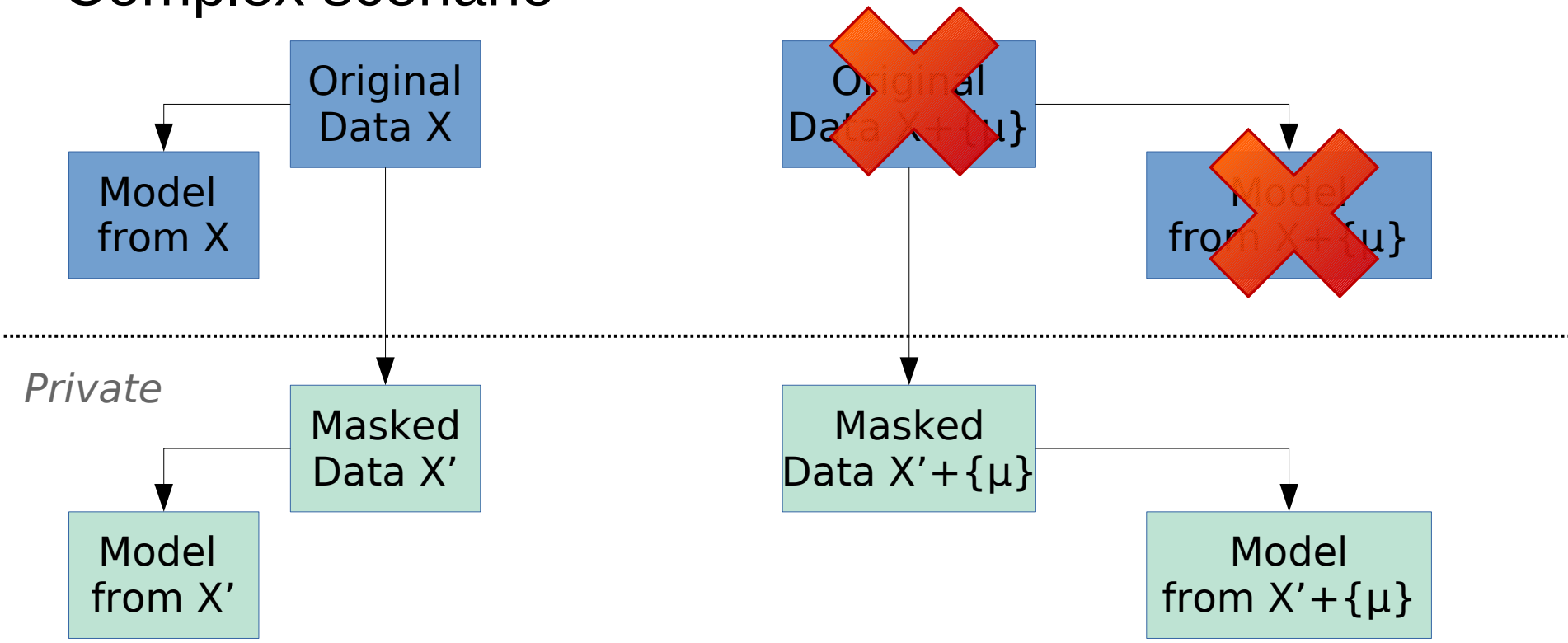
Complex scenario



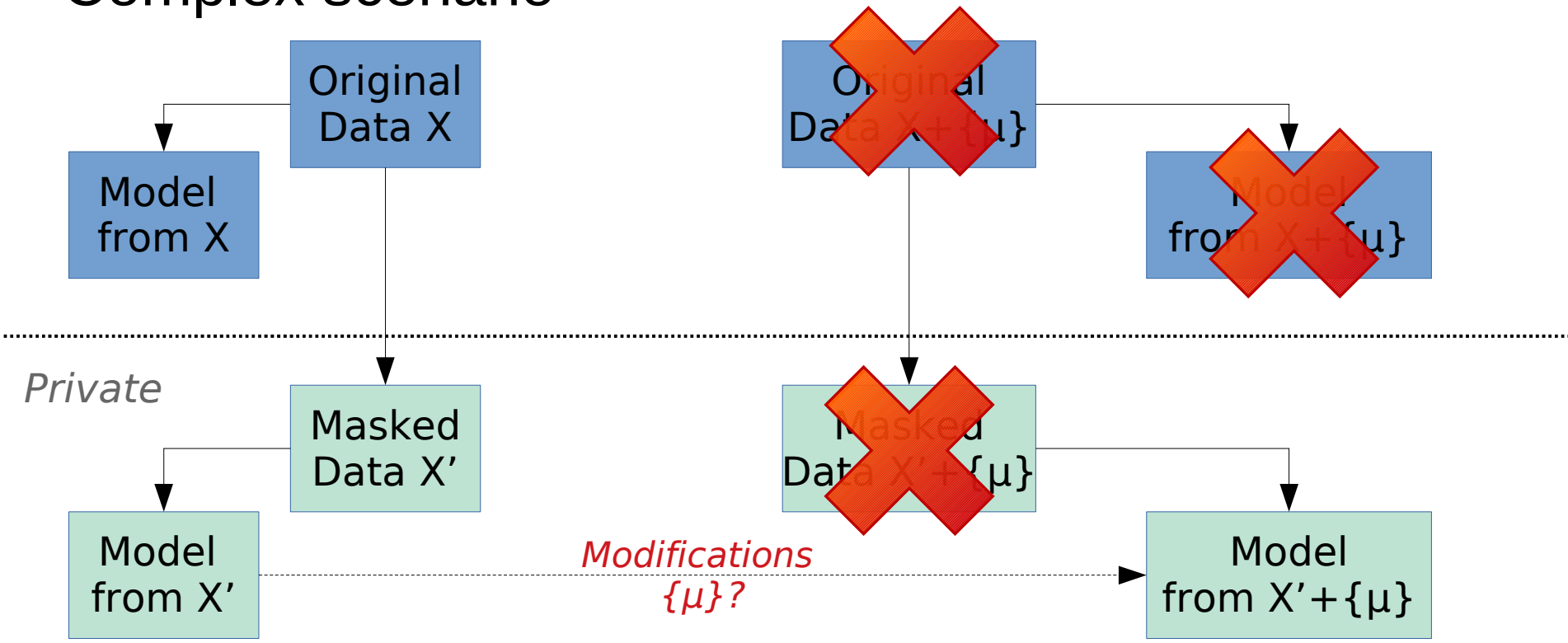
Complex scenario



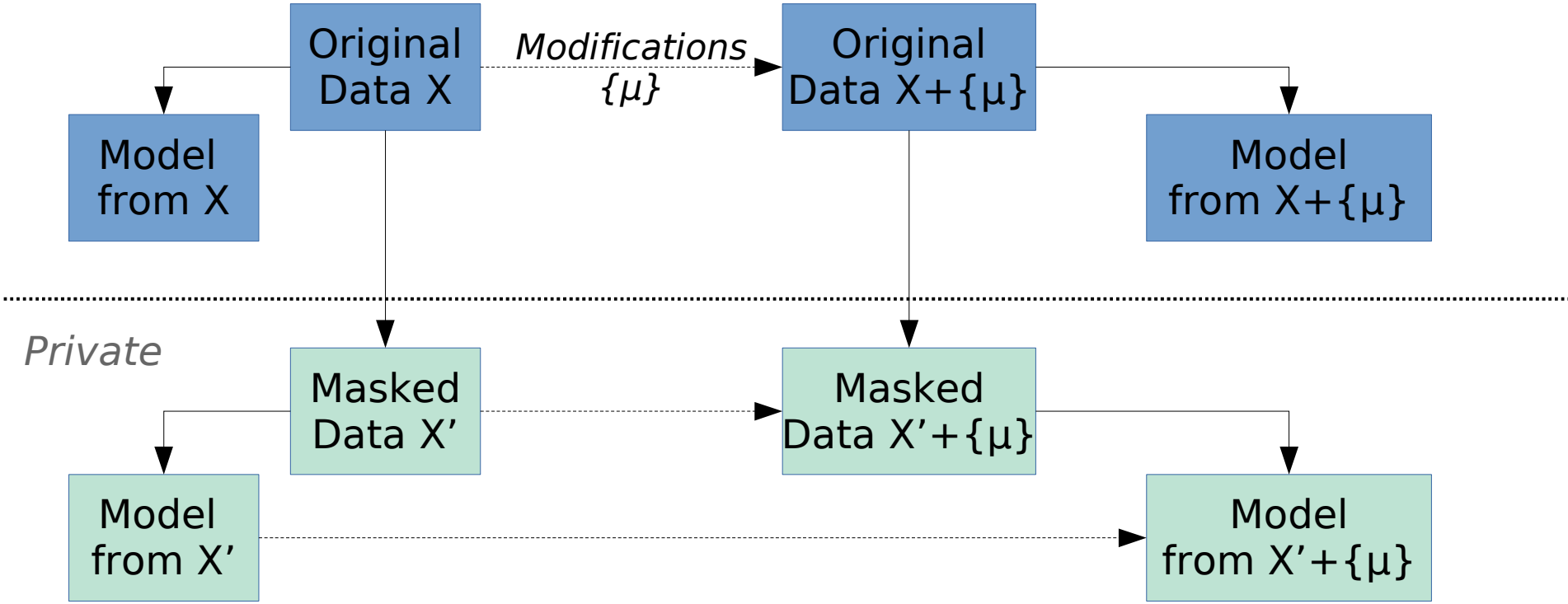
Complex scenario



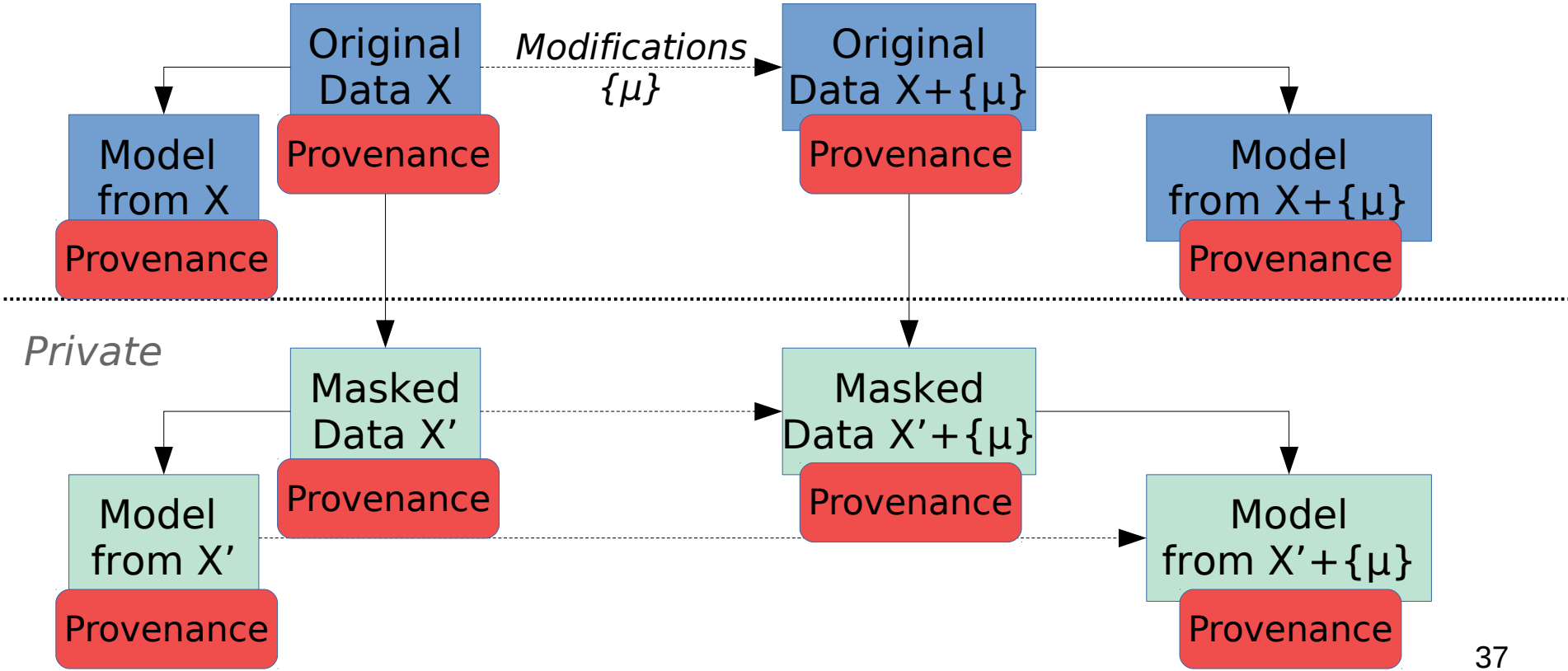
Complex scenario



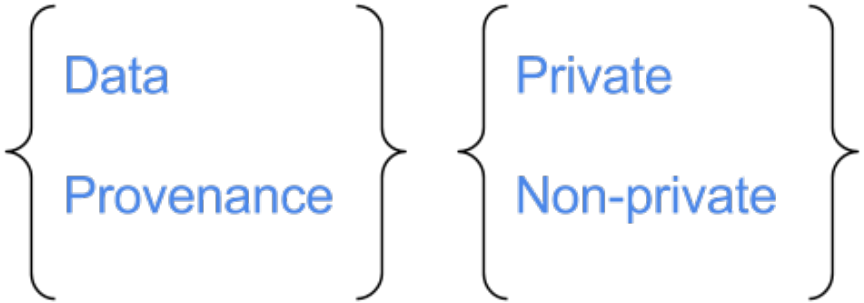
Complex scenario with Provenance



Complex scenario with Provenance



Private Data vs Private Provenance

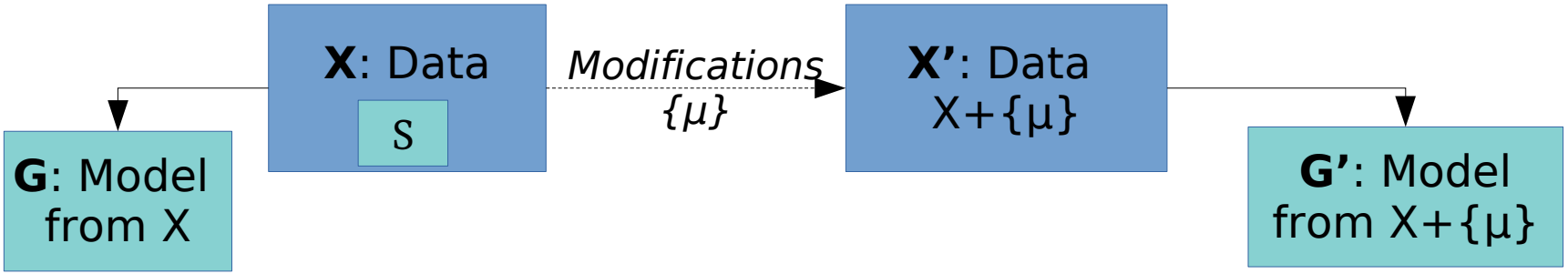


Private Data	Private Provenance
Non-private Data	Private Provenance
Private Data	Non-private Provenance
Non-private Data	Non-private Provenance

Privacy Models

- Privacy applied to data and/or provenance
  - *k*-anonymity
  - Re-identification
  - Differential privacy
  - ...
- **Be careful!**: Need to take the whole scenario into account:
  - Avoid inferences from multiple publications of masked data or models
  - Avoid inferences from changes in the same dataset or model
  - ...

Integral Privacy: problem description



Privacy problem (attacker):

- Given, **S** subset of **X**, **G**, and **G'**:
  - find the set **M={μ}** of possible modifications **μ** consistent with data **S** and knowledge **G**, **G'**.
  - Find elements in **X \ S**.

Integral Privacy: definitions

- *i*-integral privacy when **M** is large and such that the intersection of all subsets is empty.
- *i*-integral privacy à la *k*-anonymity, when the set **M** contains at least *k* alternatives.
- *k*-anonymous integral privacy when the set **M** has at least *k* minimal elements. (Modifications define a lattice).

Conclusions

- Data **Provenance** is becoming important for industry (at least in the EU)
- **Privacy** makes the whole thing even more complex
- There is a lack of tools and knowledge about provenance and privacy
- It is indubitably difficult:
  - Big data ➡ Huge Provenance
  - Privacy is always tricky
  - Need to convince lawyers and the like

Thank you for your attention

- Vicenç Torra: vtorra@his.se
- Guillermo Navarro-Arribas: guillermo.navarro@uab.cat
- David Sanchez-Charles: david.sanchez@ca.com
- Victor Muntés-Mulero: victor.muntes@ca.com