A Cautionary Reflection on (Pseudo-)Synthetic Data from Deep Learning on Personal Data

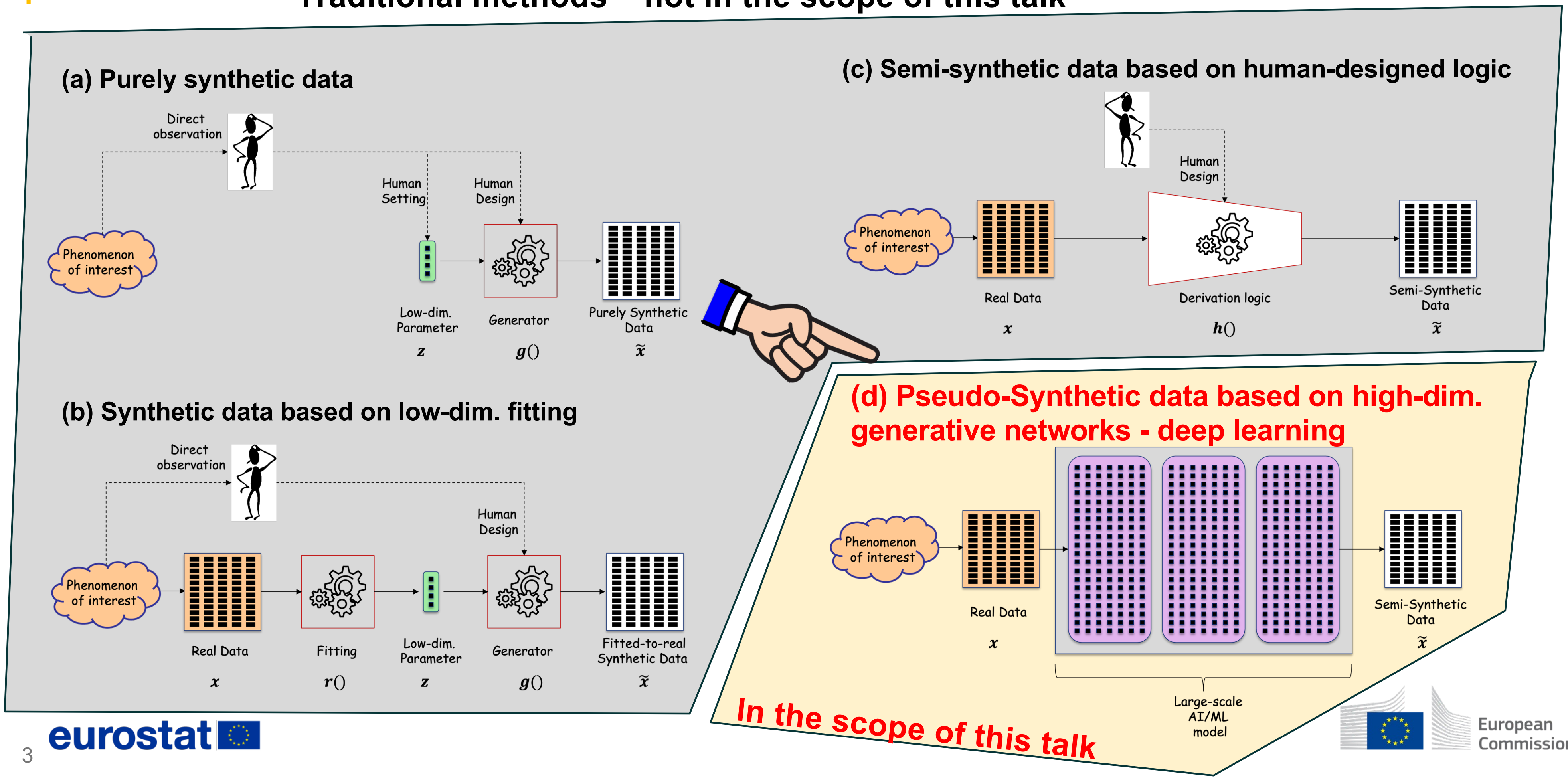
Fabio Ricciato
Eurostat, Unit A.5 Methodology; Innovation in official statistics

Privacy in Statistical Databases conference (PSD 2024)
25-27 September 2024, Antibes, France

Caveat: The information and views set out in this presentation **are those of the author** and do not necessarily reflect the official opinion of the European Union. Neither the European Union institutions and bodies nor any person acting on their behalf may be held responsible for the use which may be made of the information contained therein.

Synthetic vs. pseudo-synthetic

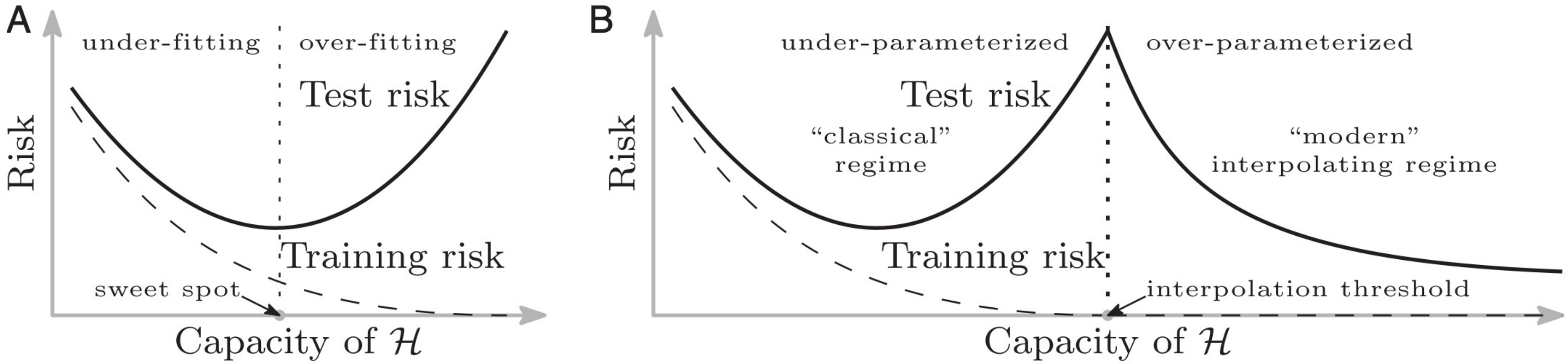
Traditional methods – not in the scope of this talk



Focus on over-parameterised models

Classical ML
Bias-Variance trade-off,
Over-fitting is bad!
Model size << Data size;
Models learn *from* the data
but cannot learn *the data*

Modern ML, Deep Learning
No trade-off between Bias and Variance!
Over-fitting is good!
Model size >> Data size (over-parameterized);
Models learn *from* the data
and also learn *the data*

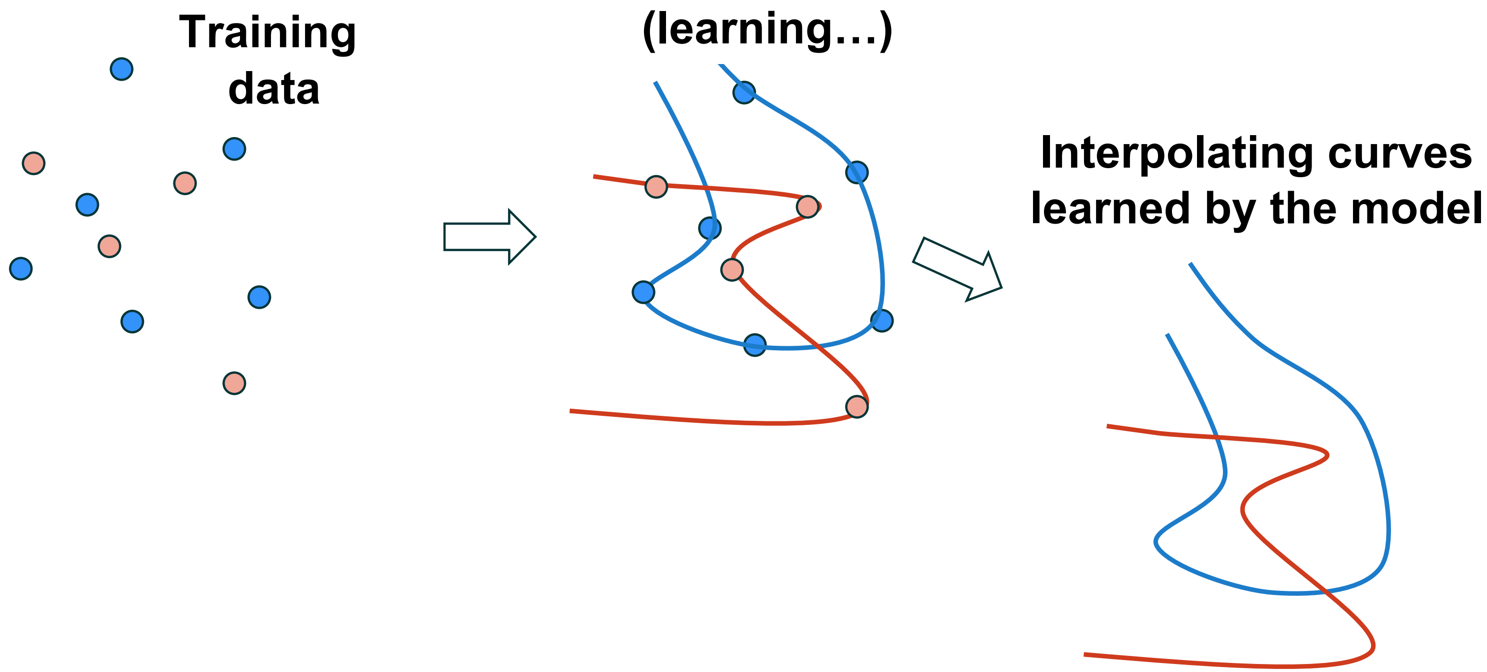


eurostat

Source: M. Belkin, D. Hsu, S. Ma and S. Mandal,
Reconciling modern machine-learning practice and the classical bias–variance trade-off
https://www.pnas.org/doi/full/10.1073/pnas.1903070116

European Commission

Over-parameterized models fit the data

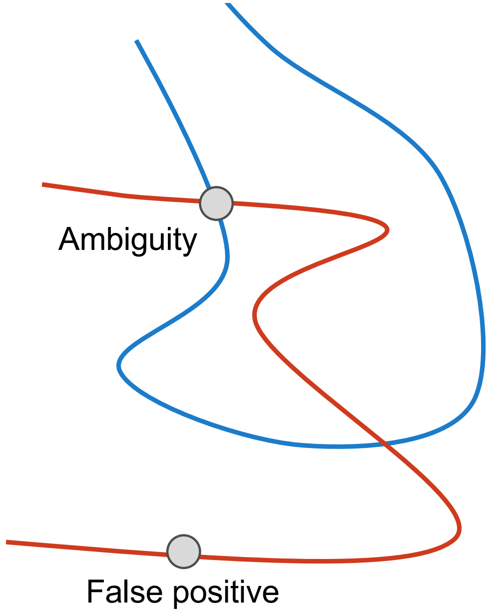


Over-parameterized models fit the data

If you know the fitting curves
(= have access to trained model)
you can easily perform
- Attribute Discovery (AD) (with low ambiguities)
- Membership Inference (MI) (with low false positives)

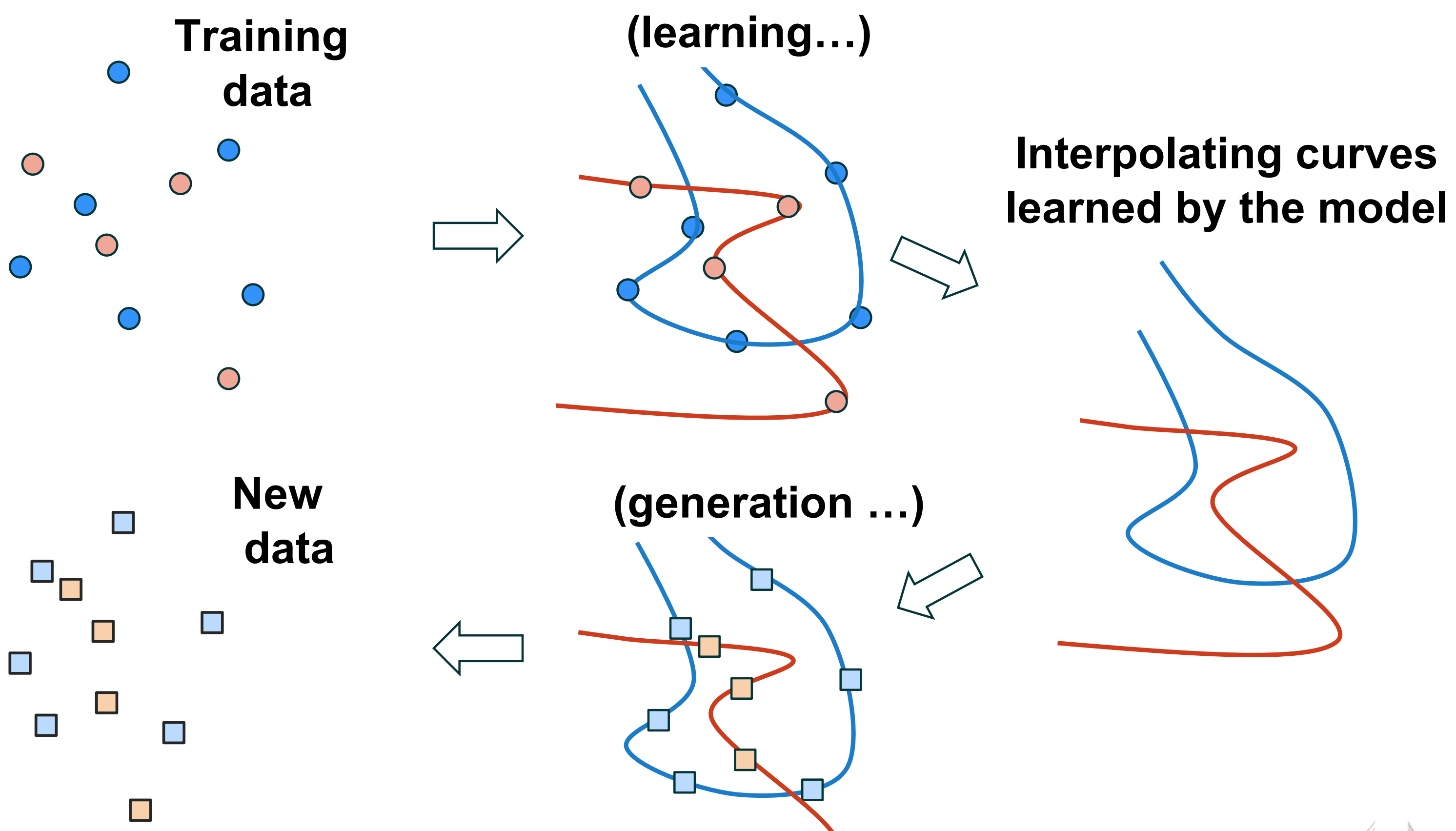
Shouldn't that be sufficient to qualify the fitting curve – hence the trained model – as personal data?

Interpolating curves learned by the model



European Commission

Generation of new points …



Dissimilarity ≠ Privacy (1/2)

The new data points , may be all well separated
from the original data points , (no matchings, minimum distance) but …

… under certain conditions the new data points allow reconstructing the learned fitting curves:

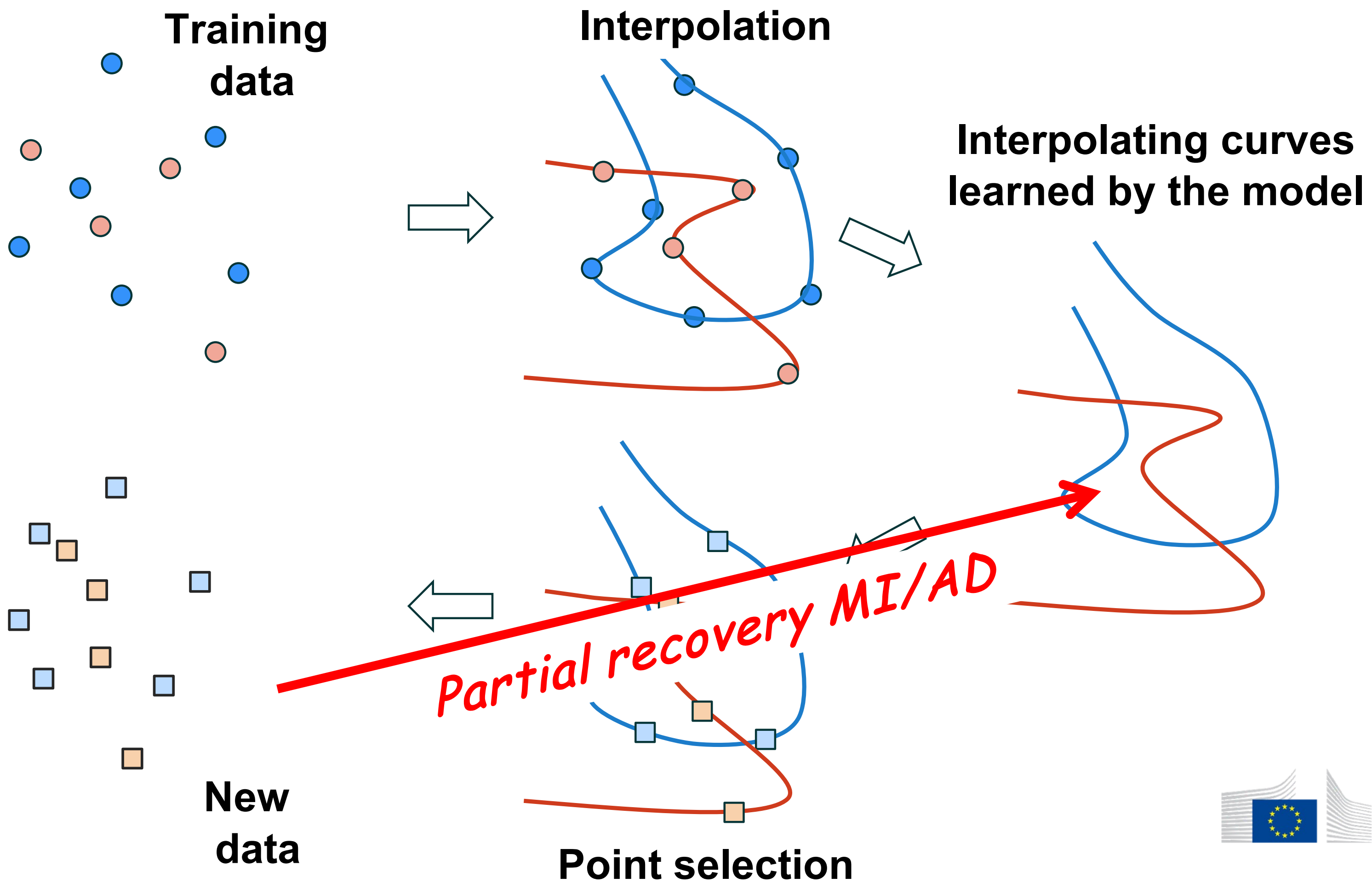
- curve belonging to parametric family with *n* degrees of freedom + number of new data points is at least *n*
- the parametric family is known (or can be guessed) by the attacker

→ **the new data points are as exposed to MI/AD as the trained model: shouldn't they too qualify as personal data?**

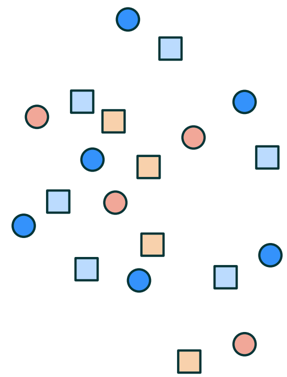
Condition on the generation process

Condition on attacker's knowledge

Partial recovery (MI/AD risk)



Dissimilarity ≠ Privacy (2/2)



The new data points , may be all well separated
from the original data points , (no matchings, minimum distance) but …

… under certain conditions the new data points allow reconstructing the learned fitting curves:

- curve belonging to parametric family with *n* degrees of freedom & number of new data points is at least *n*
- the parametric family is known (or can be guessed) by the attacker

→ **the new data points are as exposed to MI/AD as the trained model: shouldn't they too qualify as personal data?**

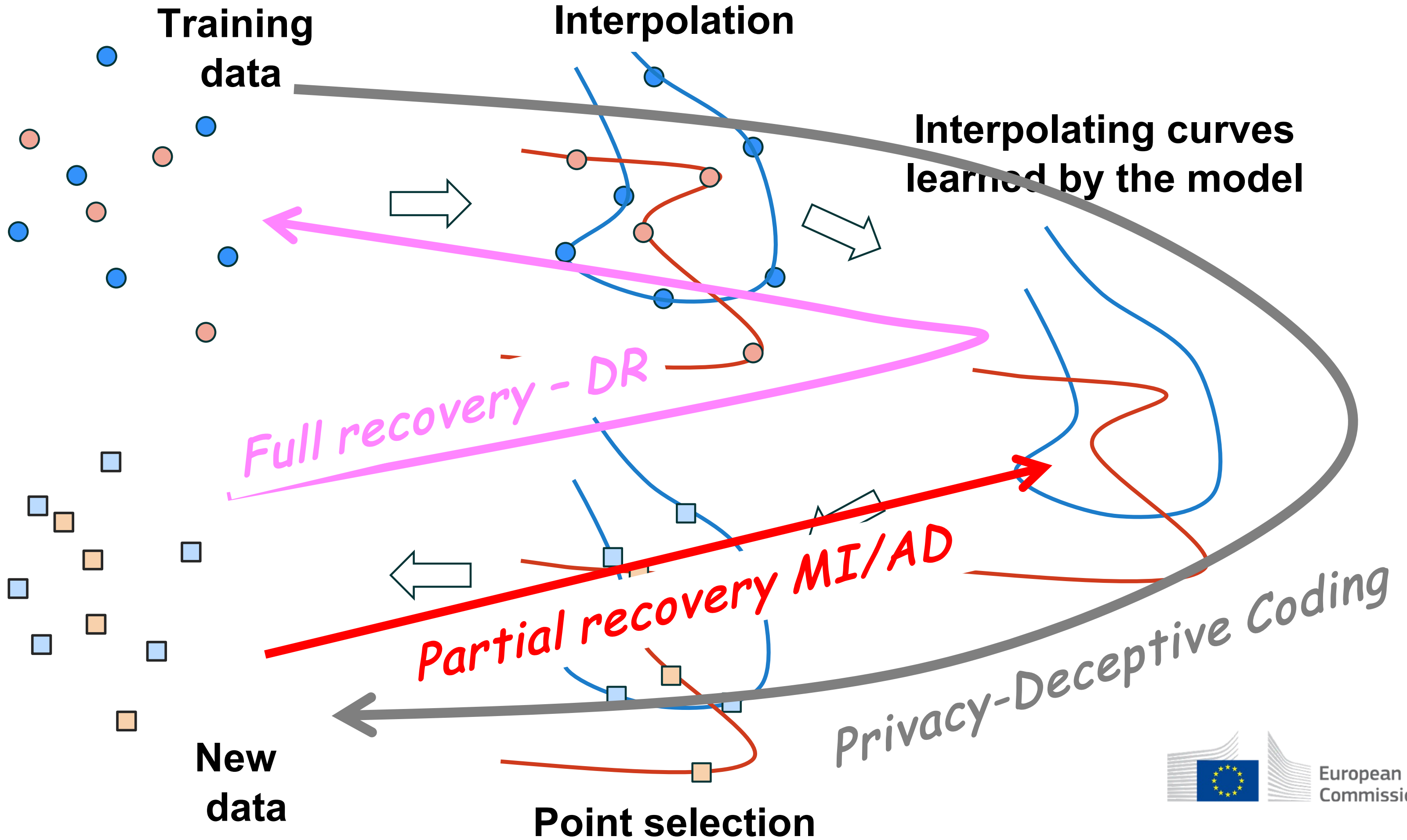
… furthermore, under some additional conditions the new data points would allow recovering exactly the original data points (Database Reconstruction, DR)

- new data are picked along the curve according to some criterion designed purposedly to be reversible (e.g., fixed distance from original data points)
- such criterion is known (or can be guessed) by the attacker

Condition on the generation process *G*

Condition on attacker's knowledge *K*

Full recovery (DR risk)



What humans can design, machines can learn

- We introduce the notion of **Privacy-Deceptive Coding** scheme = data generation that allows full (DR) or partial recovery (AD,MI) of the training data. It can be …
- **designed** manually by a rogue human (e.g., polynomial interpolation);

- **learned intentionally** by a rogue AI/ML network designed deliberately to learn a reversible Privacy-Deceptive Coding

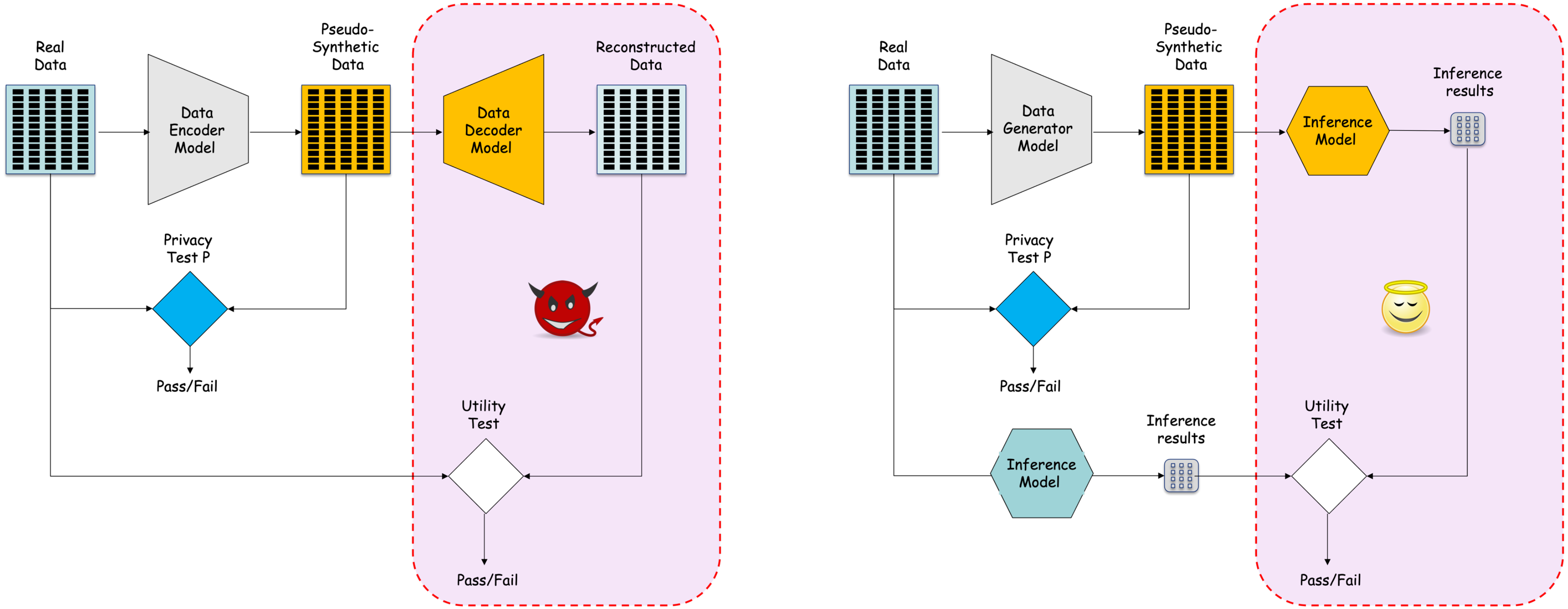
- **learned unintentionally by a non-rogue** AI/ML network designed with the declare purpose of "maximizing utility" (?)
  - Can you prove that your pseudo-synthetic data generation network has NOT ended up learning some kind of privacy-deceptive coding *G*? And do you even understand what it has learned?
  - Can you make sure that potential attackers won't acquire (or guess) the auxiliary knowledge *K?*

Condition on the generation process *G*

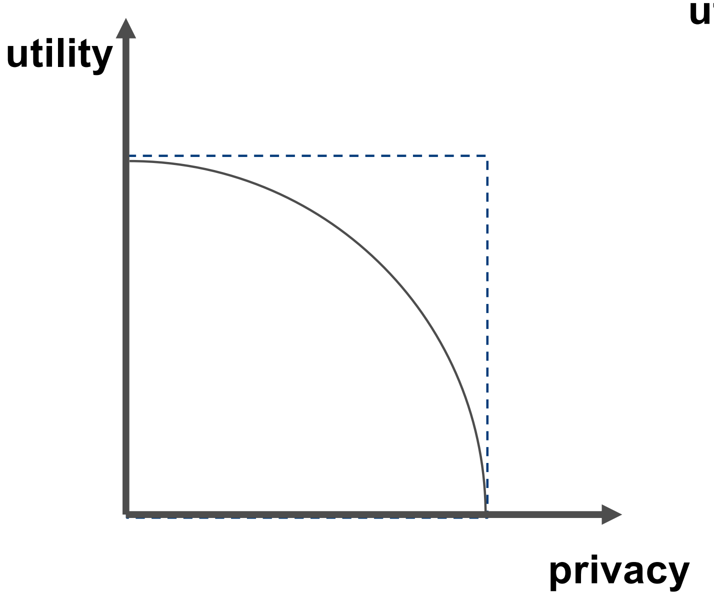
Condition on attacker's knowledge *K*

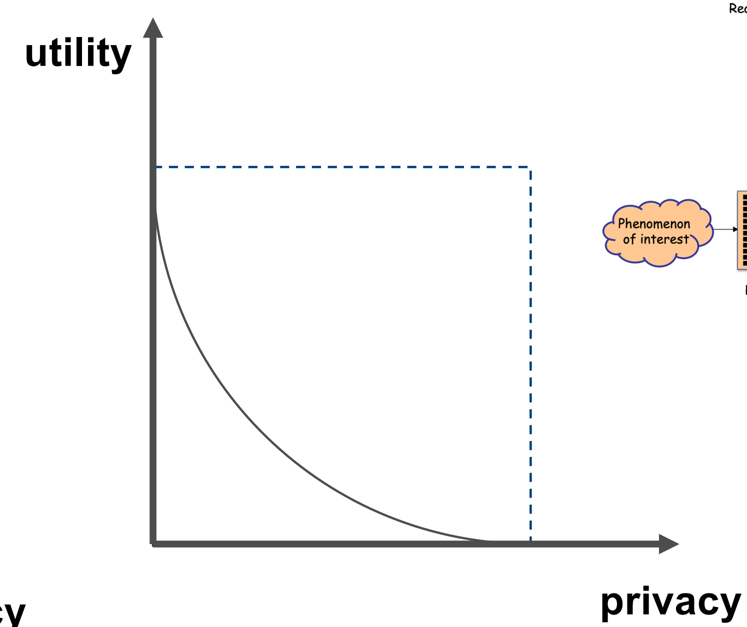
European Commission

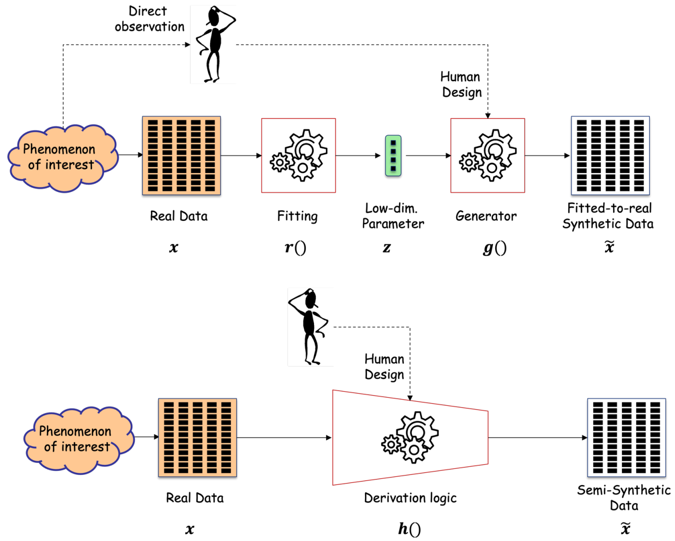
Intentional vs Unintentional learning



Utility-Privacy frontier



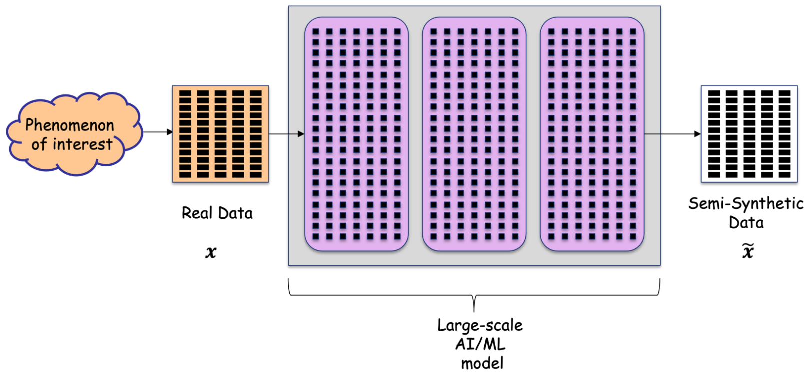


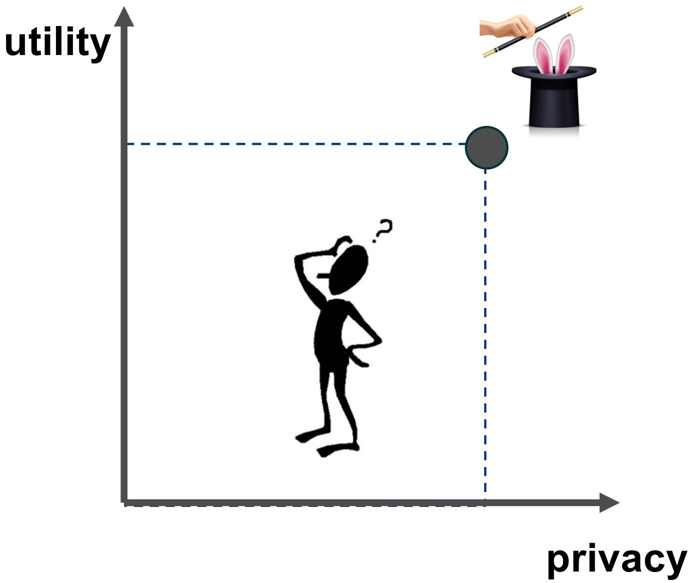


In the traditional schemes, **interpretability** (manual design) and **dimensionality bottleneck** (parsimony) allow to assess where we stand on the Utility-Privacy frontier

European Commission

Utility-Privacy frontier





With data generators based on over-parameterized models, both **interpretability** and the **dimensionality bottleneck** are gone**.**

We may still assess utility, but how to assess privacy?

NB: confusing "privacy" with "dissimilarity" between the new and original data lead to the illusion that we can "jump over" the utility-privacy frontier

Take-home messages

- Read the paper by Belkin https://www.pnas.org/doi/full/10.1073/pnas.1903070116
- Dissimilarity ≠ Privacy
  - **Dissimilarity metrics** are widely used (and may be meaningful) in contexts when (1) the data **generation process is known**, as in traditional human-design methods, or (2) when a **dimensionality bottleneck** along the generation process rules out the possibility of learning a Privacy-Deceptive Coding scheme
  - **Dissimilarity metrics** alone cannot be used to assess privacy when neither (1) or (2) are there, as is the case with large scale deep learning networks
- Privacy assessment needs knoweledge *and interpretability* of the data generation process → no interpretability, no privacy!
- Pseudo-synthetic data generated by deep learning on personal data should be considered, precautionarily, as personal data

European Commission

Thank you

© European Union 2024

Unless otherwise noted the reuse of this presentation is authorised under the CC BY 4.0 license. For any use or reproduction of elements that are not owned by the EU, permission may need to be sought directly from the respective right holders.

Backup slides
(in case of questions)

eurostat

European Commission

Role of auxiliary knowledge *K*

- Question. For the attack to succeed, the attacker must have some auxiliary knowledge *K.* So if we keep the data generation model secret, we can release the pseudo-synthetic data safely – can't we?

- Answer. Think of the pseudo-synthetic data generation process as being **analogous to an encryption scheme**, with auxiliary knowledge *K* being the analogous of the ciphering key. And be reminded that there are «weak» encryption schemes that could be cracked by cryptoanalysis (e.g., earlier versions of GSM encryption)

  - Is your pseudo-synthetic data generation akin to «robust encryption» or rather «weak encryption»? How difficult is to crack it? Can *K* be guessed or anyway recovered from cryptoanalysis, when the attacker knows something about the original data?

  - And would you trust using a black-box encryption scheme that is provided to you by the same company that sells cracking software to the adversaries?

European Commission

https://en.wikipedia.org/wiki/A5/1

Research directions

- To enforce a “dimensionality bottlenck” (e.g., limiting the number of nodes in some network layer) we need to ensure that:

Model size (capacity) << Data size

- Question: What is the intrinsic “size” of the data? Can we compute it?
- Answer: I don’t know and I think it’s a very interesting open research question (I would look in the direction of Kolmogorov complexity…)

European Commission