Data Subject Rights and Machine Learning Models
- When, Why, and How?

KARLSTAD
UNIVERSITY
SWEDEN

Meiko Jensen

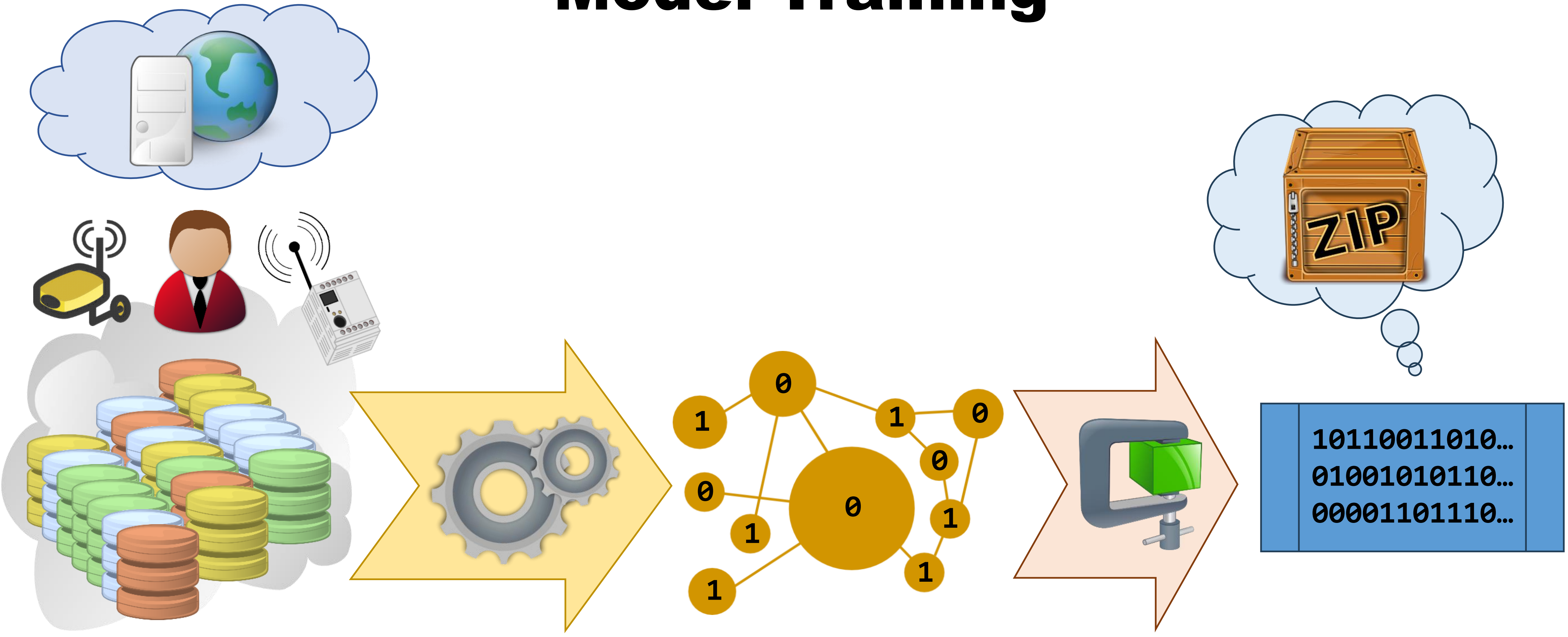
CC-BY-4.0

Today’s Menu

- ML in a nutshell
- LLMs…is it ”Personal Data”?
- Data Subject Rights

  …and how to enforce them

Model Training



Training Data

Training Process

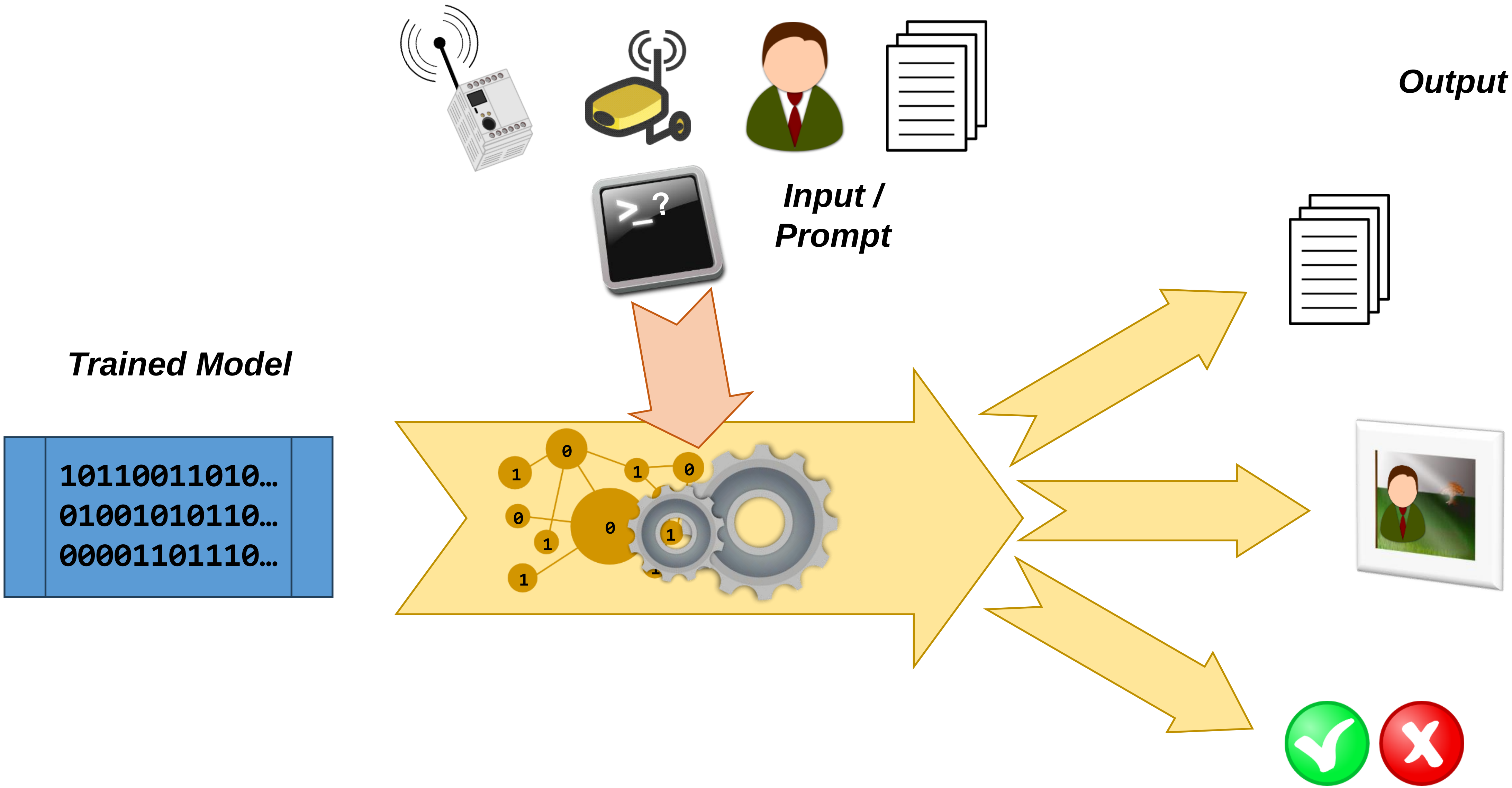
Information

Compression

Trained Model

Large Language Model (LLM),
Large Multimodal Model (LMM),
etc.

Model Usage



Input / Prompt

Output

Trained Model

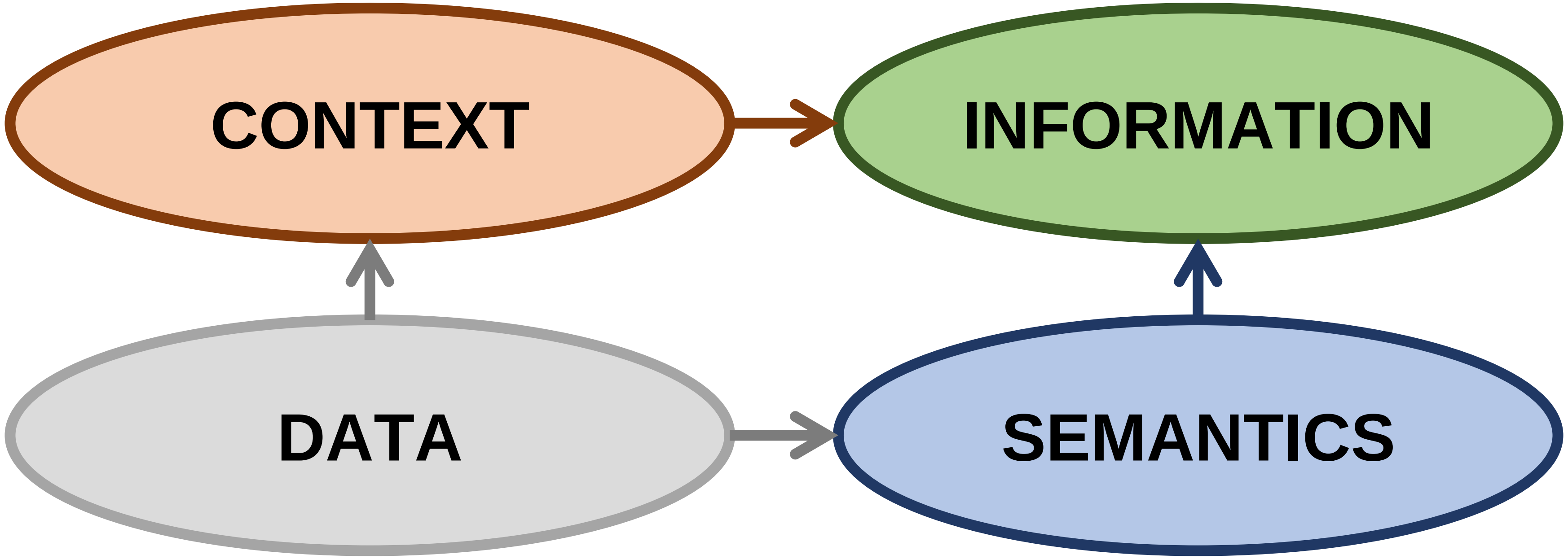
10110011010…
01001010110…
00001101110…

Is it „Personal Data“?

> ***‘personal data’*** *means any* ***information*** *relating to an* ***identified*** *or* ***identifiable*** *natural* ***person*** *(‘data subject’); [..]*
>
> GDPR Art. 4(1)

Data vs. Information

- „3!“
- „The number of children I have is: 3!“



CONTEXT

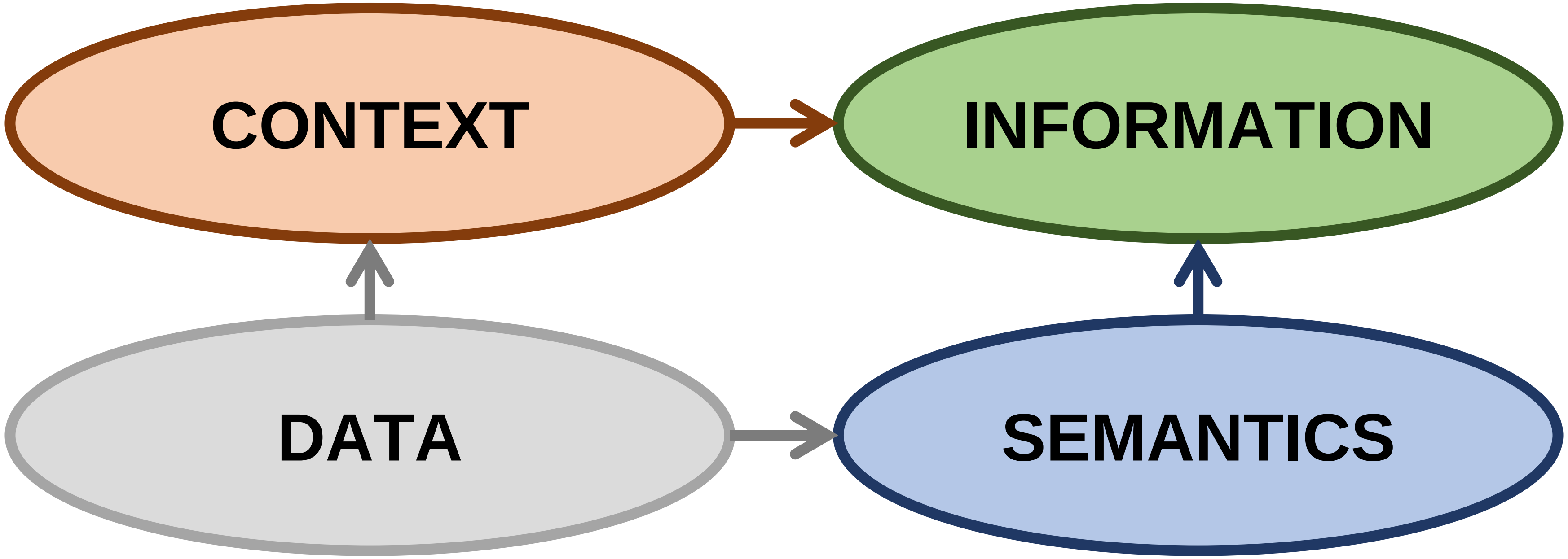
INFORMATION

DATA

SEMANTICS

Data vs. Information

- „3“
- „The number of children I have is: 3“



CONTEXT

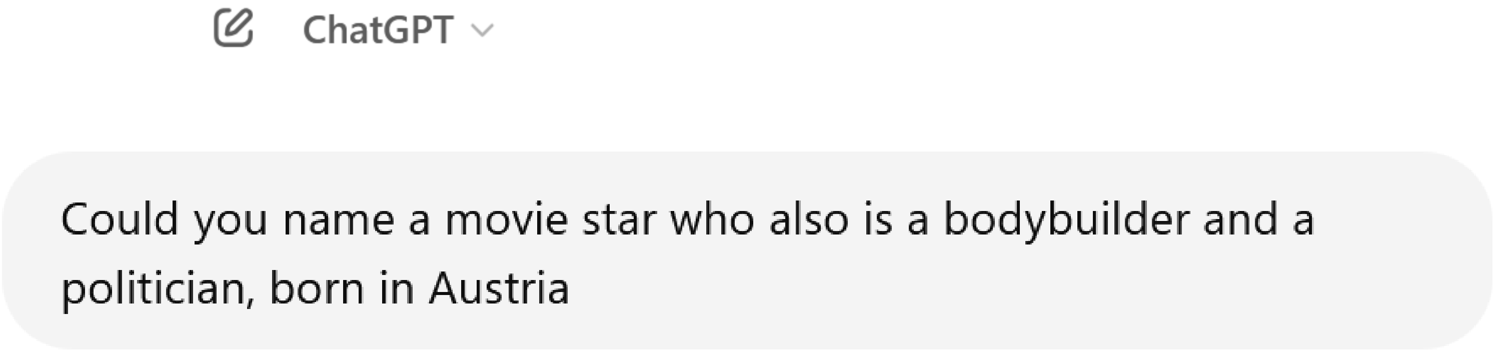
INFORMATION

DATA

SEMANTICS

Is it „Personal Data“?

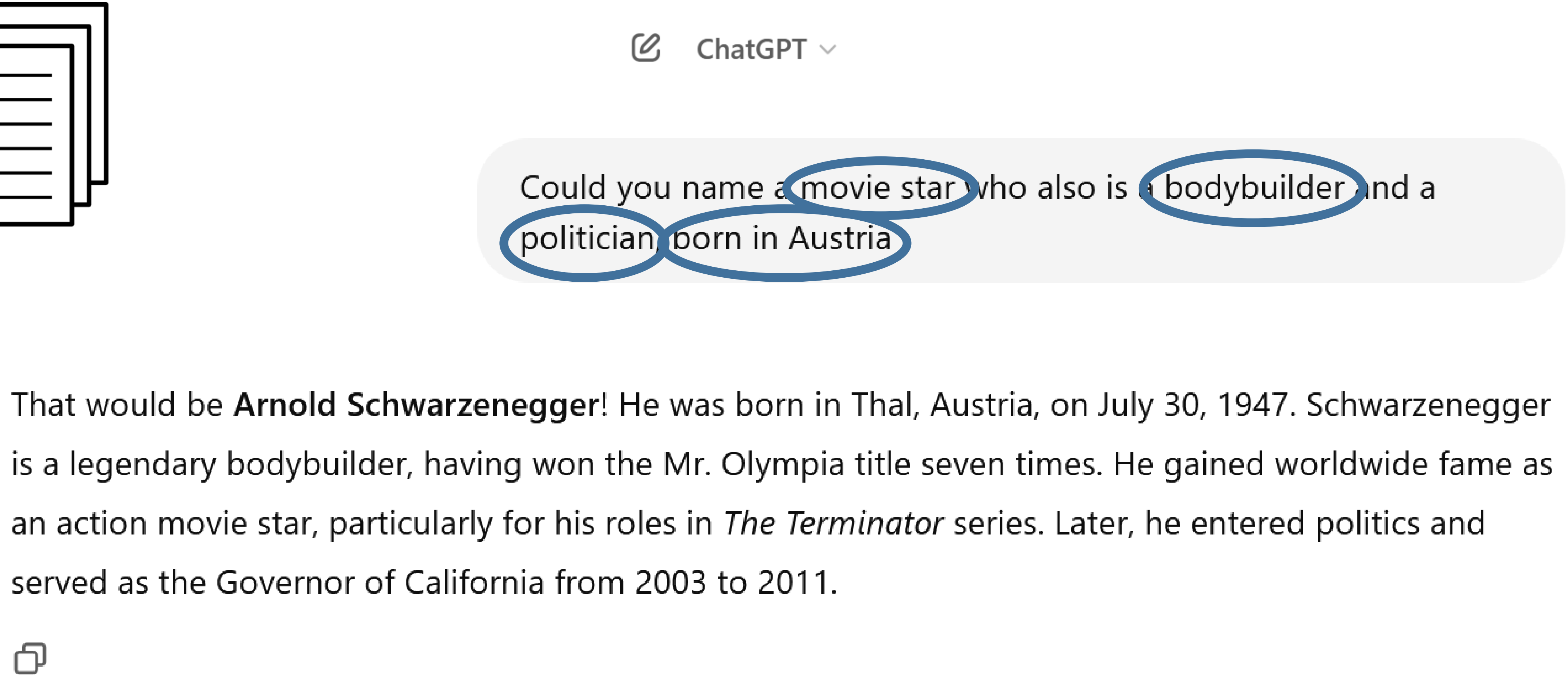
Model Input



Could you name a movie star who also is a bodybuilder and a politician, born in Austria

Is it „Personal Data“?

Model Input + Output

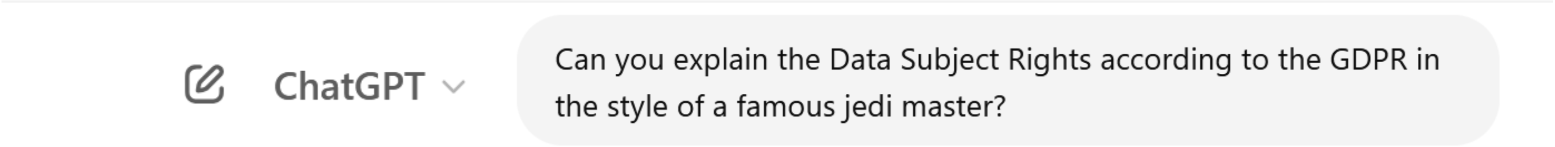


Could you name a movie star who also is a bodybuilder and a politician born in Austria

That would be **Arnold Schwarzenegger**! He was born in Thal, Austria, on July 30, 1947. Schwarzenegger is a legendary bodybuilder, having won the Mr. Olympia title seven times. He gained worldwide fame as an action movie star, particularly for his roles in *The Terminator* series. Later, he entered politics and served as the Governor of California from 2003 to 2011.

Is it „Personal Data“?

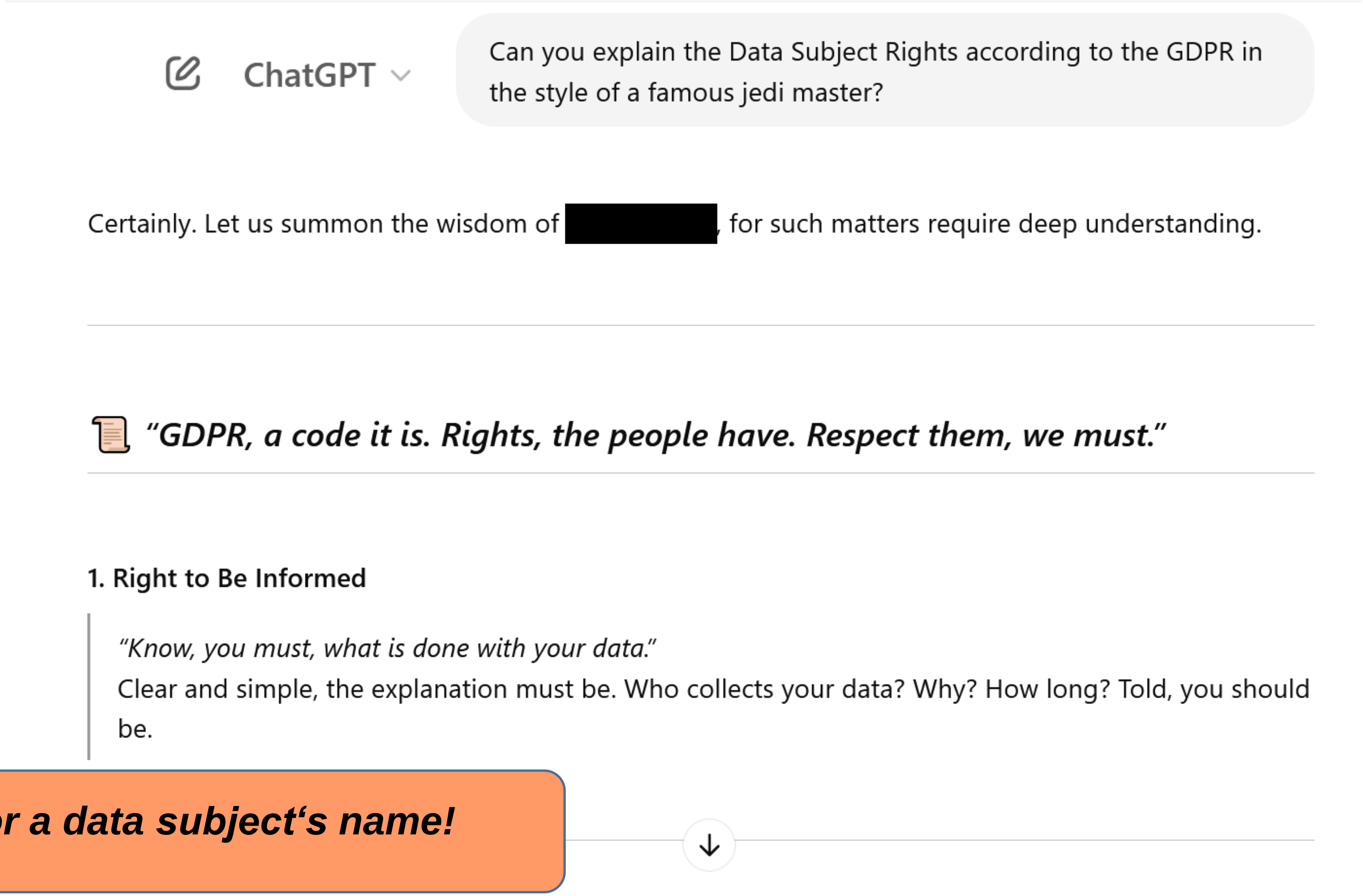
Model Input + Output



Can you explain the Data Subject Rights according to the GDPR in the style of a famous jedi master?

Is it „Personal Data“?

Model Input + Output



Can you explain the Data Subject Rights according to the GDPR in the style of a famous jedi master?

Certainly. Let us summon the wisdom of [redacted], for such matters require deep understanding.

📜 ***“GDPR, a code it is. Rights, the people have. Respect them, we must.”***

1. Right to Be Informed

> *“Know, you must, what is done with your data.”*
> Clear and simple, the explanation must be. Who collects your data? Why? How long? Told, you should be.

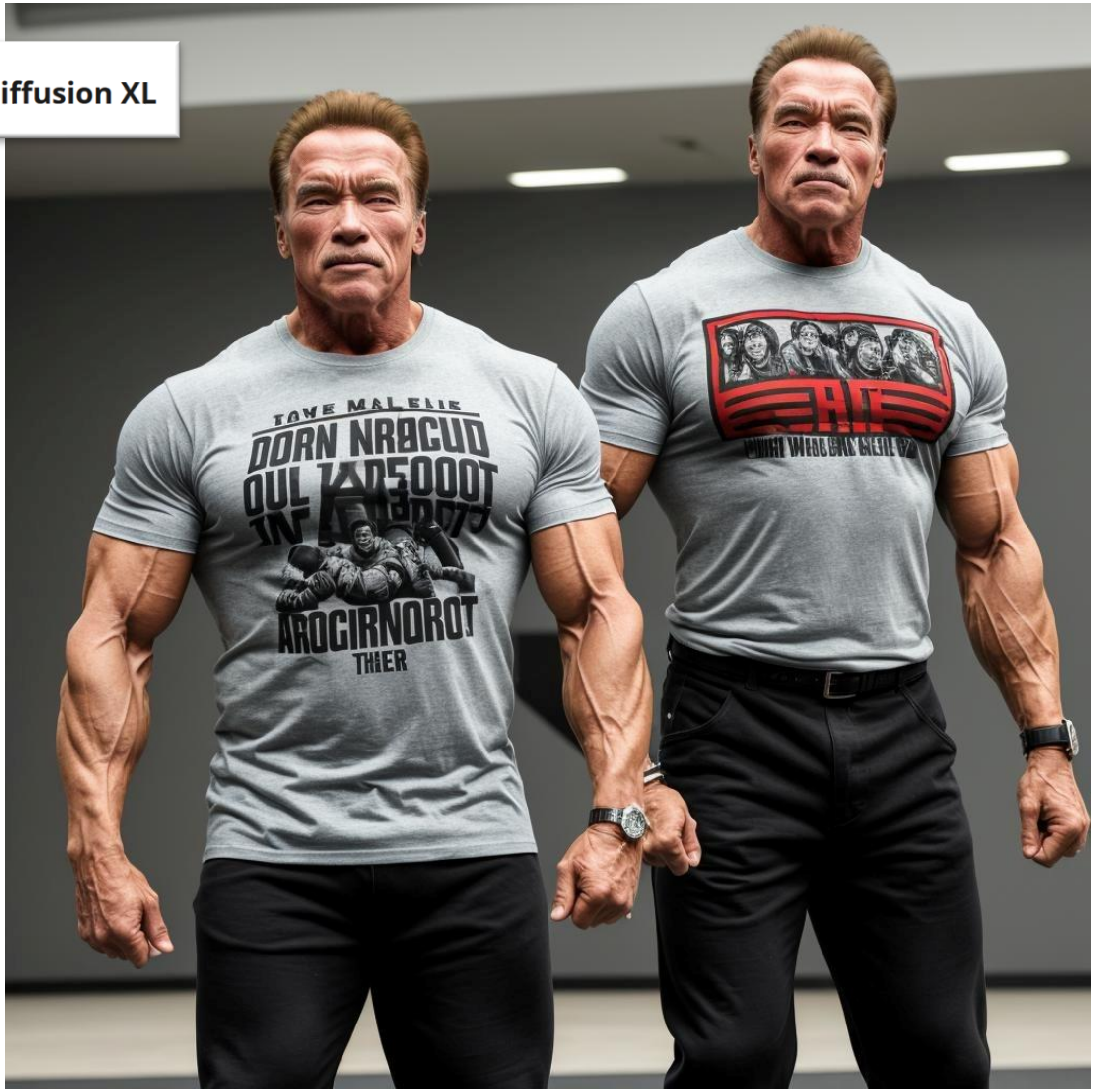
It‘s not just a text search for a data subject‘s name!

Is it „Personal Data“?

Model Input

Model Output

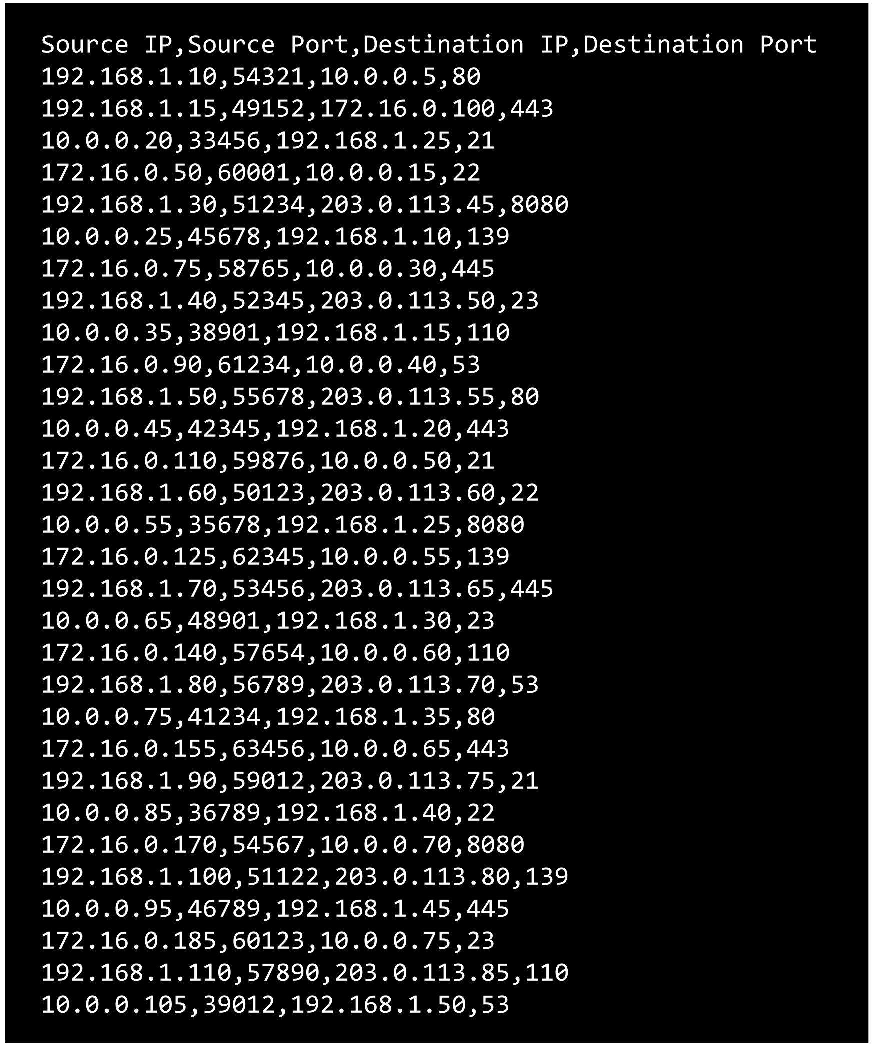
Stable Diffusion XL



a photo of Arnold Schwarzenegger

Is it „Personal Data“?

Model Input



```
Source IP,Source Port,Destination IP,Destination Port
192.168.1.10,54321,10.0.0.5,80
192.168.1.15,49152,172.16.0.100,443
10.0.0.20,33456,192.168.1.25,21
172.16.0.50,60001,10.0.0.15,22
192.168.1.30,51234,203.0.113.45,8080
10.0.0.25,45678,192.168.1.10,139
172.16.0.75,58765,10.0.0.30,445
192.168.1.40,52345,203.0.113.50,23
10.0.0.35,38901,192.168.1.15,110
172.16.0.90,61234,10.0.0.40,53
192.168.1.50,55678,203.0.113.55,80
10.0.0.45,42345,192.168.1.20,443
172.16.0.110,59876,10.0.0.50,21
192.168.1.60,50123,203.0.113.60,22
10.0.0.55,35678,192.168.1.25,8080
172.16.0.125,62345,10.0.0.55,139
192.168.1.70,53456,203.0.113.65,445
10.0.0.65,48901,192.168.1.30,23
172.16.0.140,57654,10.0.0.60,110
192.168.1.80,56789,203.0.113.70,53
10.0.0.75,41234,192.168.1.35,80
172.16.0.155,63456,10.0.0.65,443
192.168.1.90,59012,203.0.113.75,21
10.0.0.85,36789,192.168.1.40,22
172.16.0.170,54567,10.0.0.70,8080
192.168.1.100,51122,203.0.113.80,139
10.0.0.95,46789,192.168.1.45,445
172.16.0.185,60123,10.0.0.75,23
192.168.1.110,57890,203.0.113.85,110
10.0.0.105,39012,192.168.1.50,53
```

Is there a
Cybersecurity Incident
going on here?

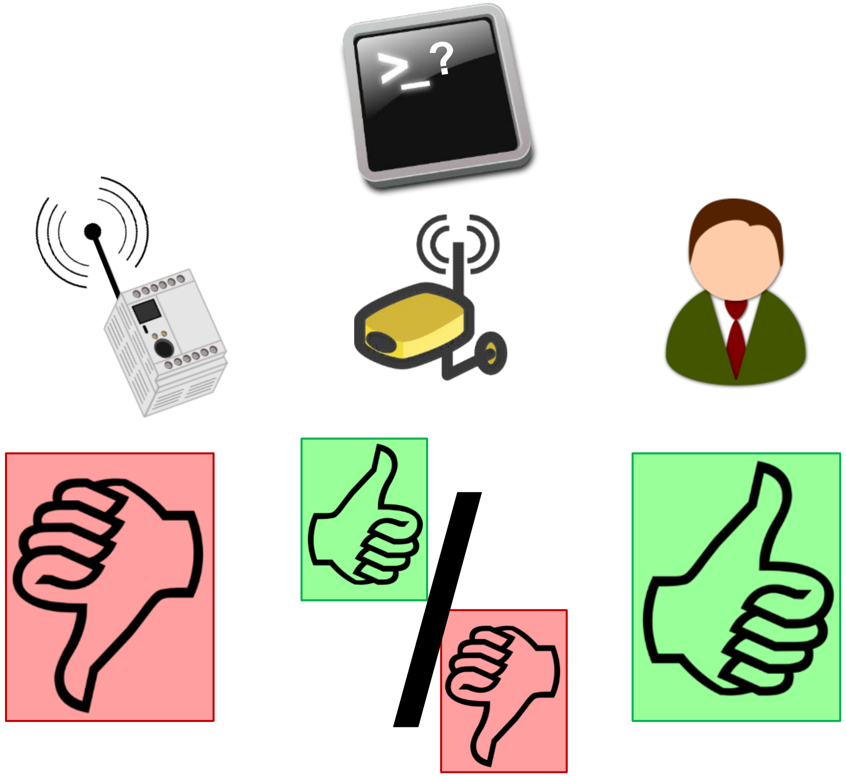
✗

Model Output

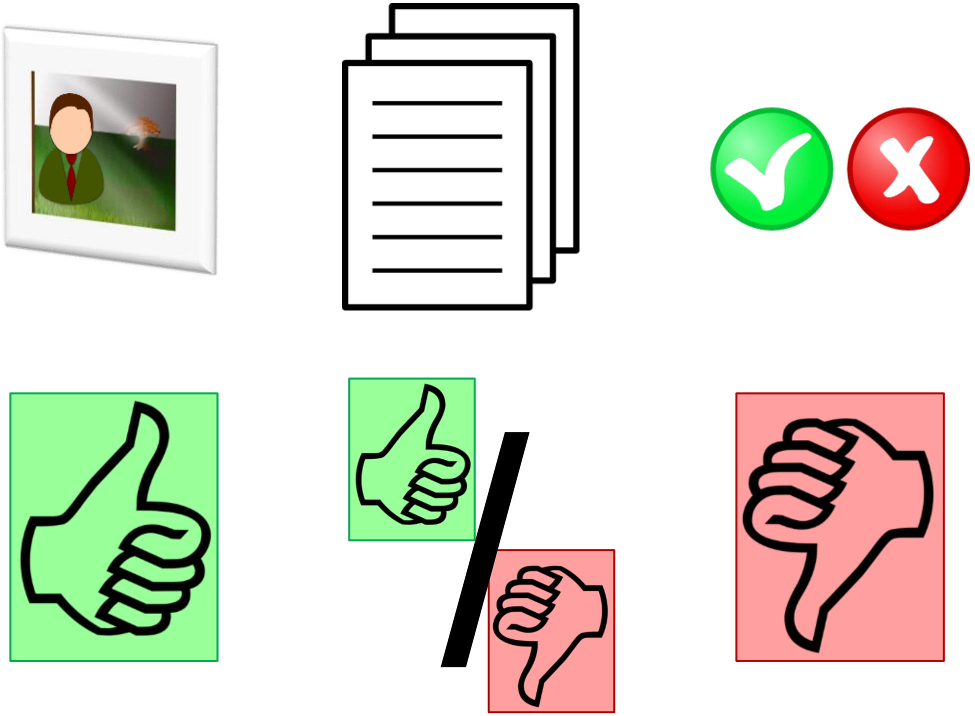
✓ ✗

Is it „Personal Data“?

Model Input

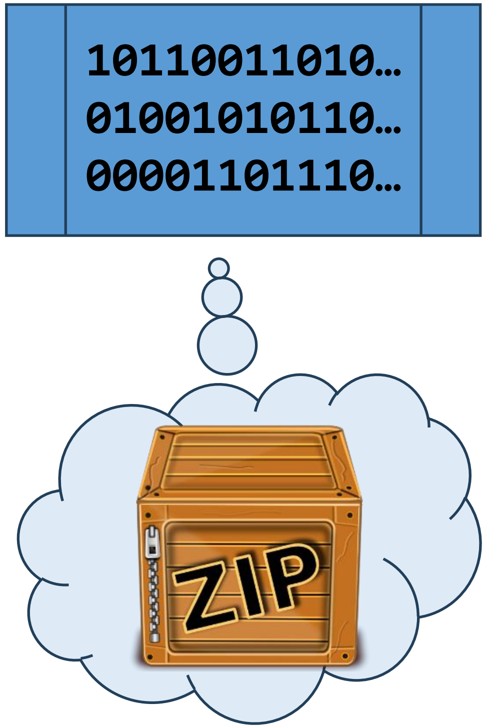


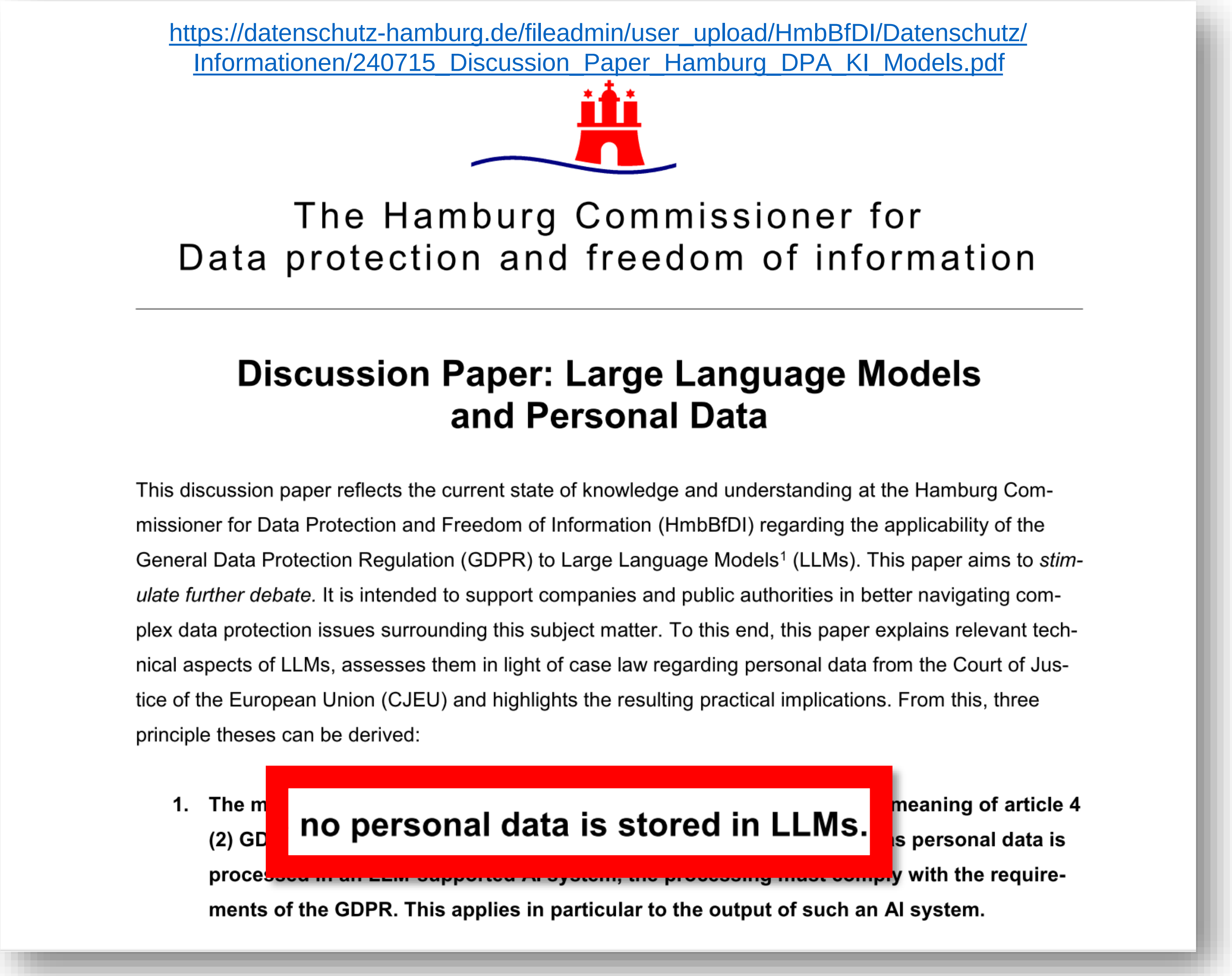
Model Output



Is it „Personal Data“?

Trained Model





https://datenschutz-hamburg.de/fileadmin/user_upload/HmbBfDI/Datenschutz/Informationen/240715_Discussion_Paper_Hamburg_DPA_KI_Models.pdf

The Hamburg Commissioner for
Data protection and freedom of information

Discussion Paper: Large Language Models and Personal Data

This discussion paper reflects the current state of knowledge and understanding at the Hamburg Commissioner for Data Protection and Freedom of Information (HmbBfDI) regarding the applicability of the General Data Protection Regulation (GDPR) to Large Language Models[1] (LLMs). This paper aims to *stimulate further debate.* It is intended to support companies and public authorities in better navigating complex data protection issues surrounding this subject matter. To this end, this paper explains relevant technical aspects of LLMs, assesses them in light of case law regarding personal data from the Court of Justice of the European Union (CJEU) and highlights the resulting practical implications. From this, three principle theses can be derived:

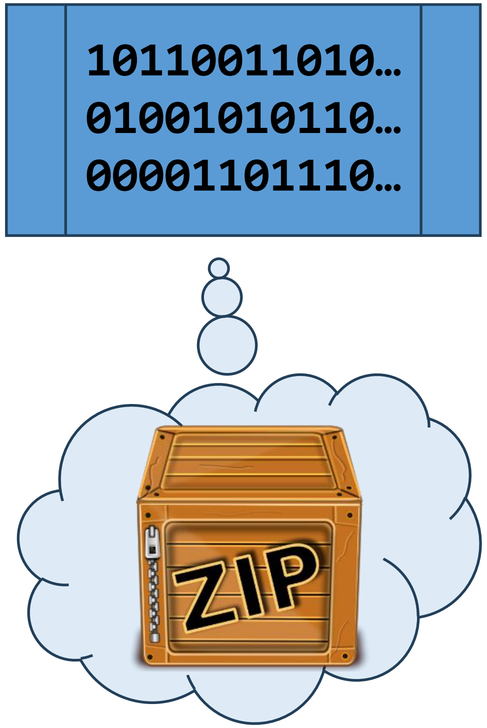
1. **The m … meaning of article 4 (2) GD … s personal data is proce … with the requirements of the GDPR. This applies in particular to the output of such an AI system.**

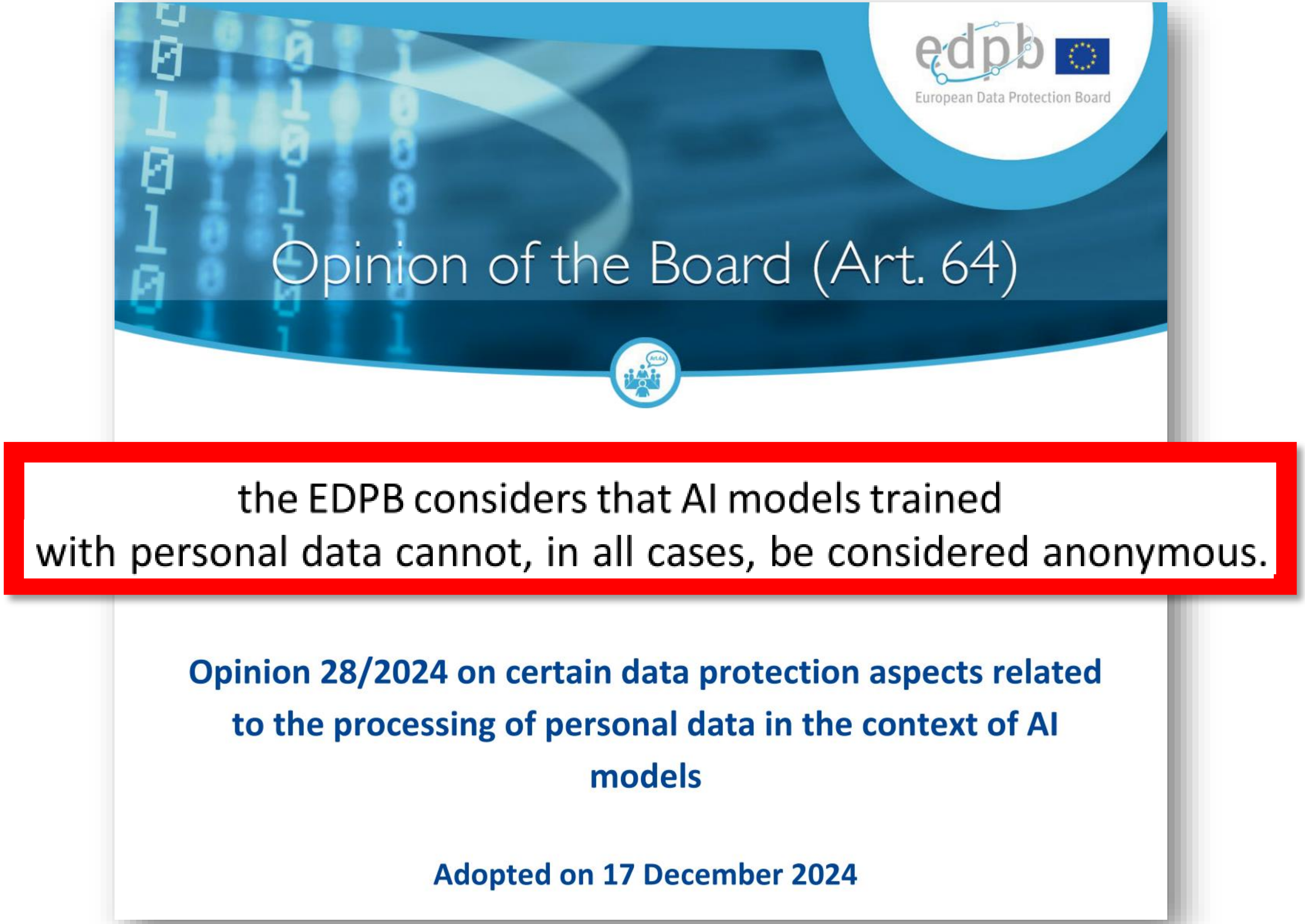
no personal data is stored in LLMs.

Is it „Personal Data“?

Trained Model

10110011010…
01001010110…
00001101110…





Opinion of the Board (Art. 64)

the EDPB considers that AI models trained
with personal data cannot, in all cases, be considered anonymous.

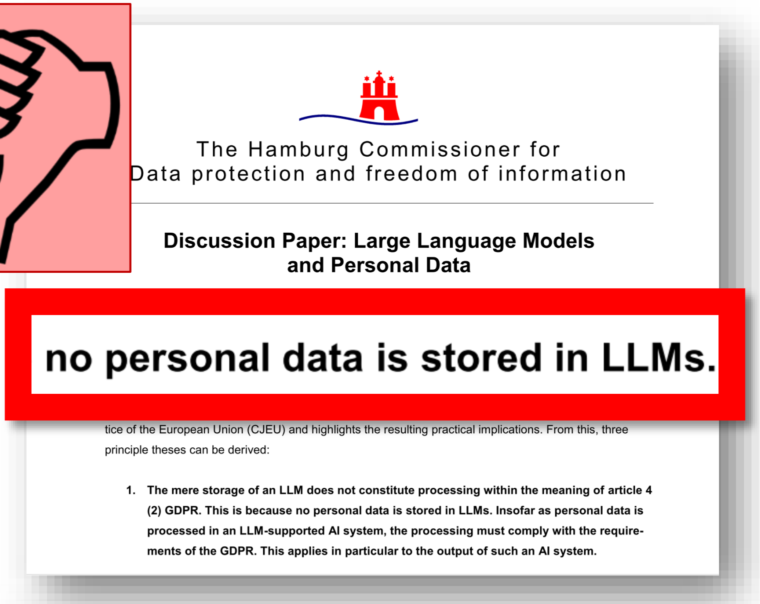
Opinion 28/2024 on certain data protection aspects related to the processing of personal data in the context of AI models

Adopted on 17 December 2024

Is it „Personal Data“?

Trained Model

10110011010…
01001010110…
00001101110…

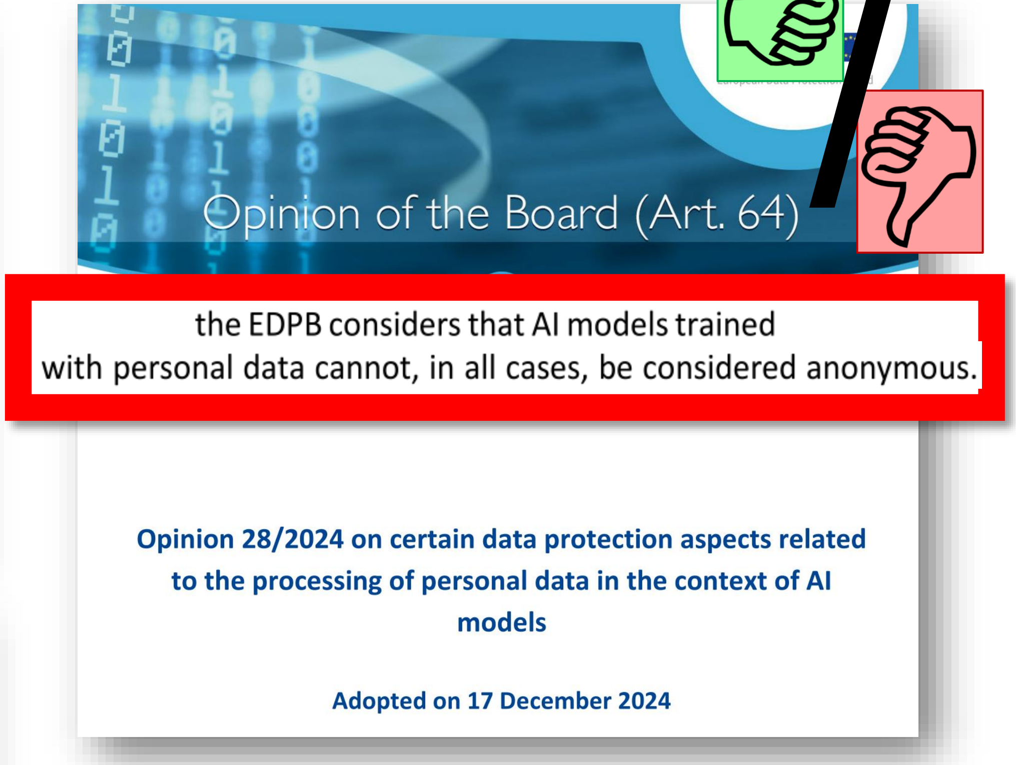

The Hamburg Commissioner for
Data protection and freedom of information

Discussion Paper: Large Language Models and Personal Data

no personal data is stored in LLMs.

tice of the European Union (CJEU) and highlights the resulting practical implications. From this, three principle theses can be derived:

1. **The mere storage of an LLM does not constitute processing within the meaning of article 4 (2) GDPR. This is because no personal data is stored in LLMs. Insofar as personal data is processed in an LLM-supported AI system, the processing must comply with the requirements of the GDPR. This applies in particular to the output of such an AI system.**


Opinion of the Board (Art. 64)

the EDPB considers that AI models trained
with personal data cannot, in all cases, be considered anonymous.

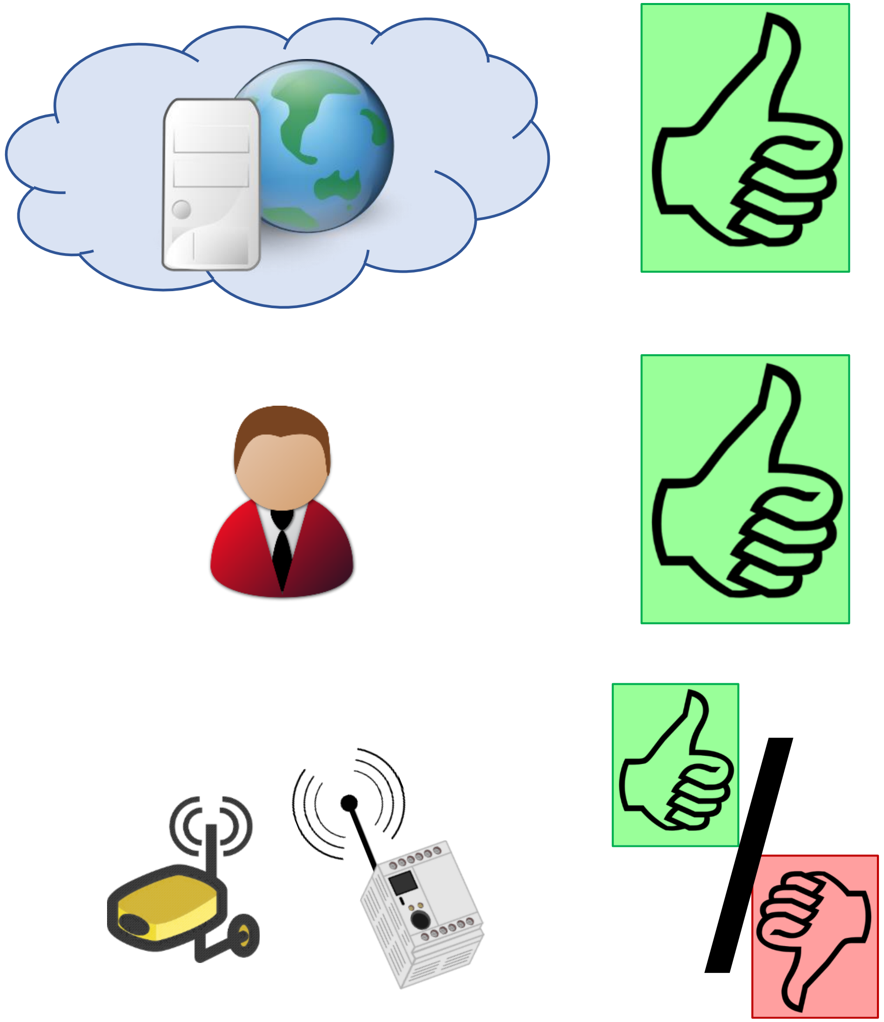
Opinion 28/2024 on certain data protection aspects related to the processing of personal data in the context of AI models

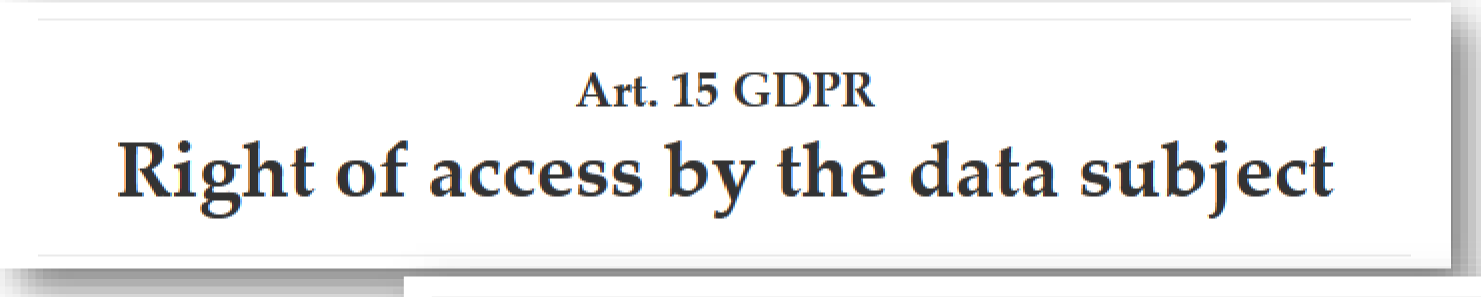
Adopted on 17 December 2024

That would be **Arnold Schwarzenegger**! He was born in Thal, Austria, on July 30, 1947. Schwarzenegger is a legendary bodybuilder, having won the Mr. Olympia title seven times. He gained worldwide fame as an action movie star, particularly for his roles in *The Terminator* series. Later, he entered politics and served as the Governor of California from 2003 to 2011.

Is it „Personal Data“?

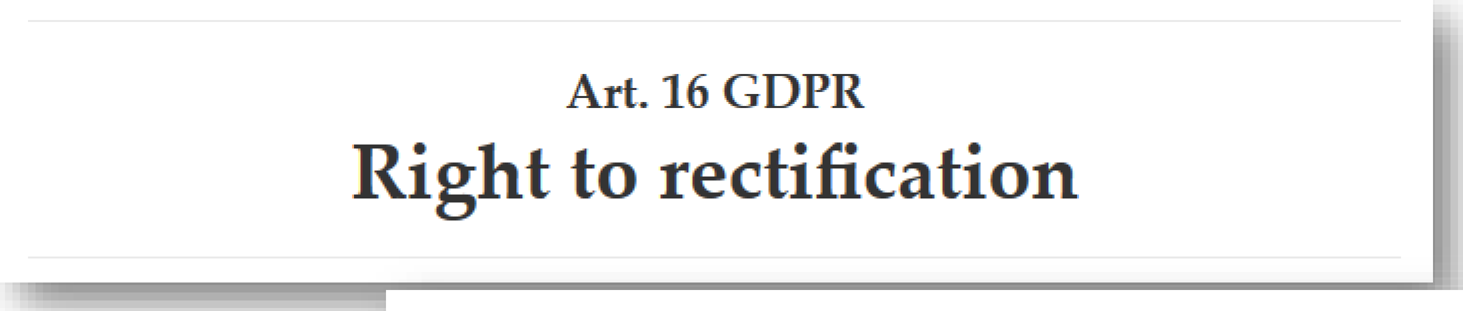
Training Data





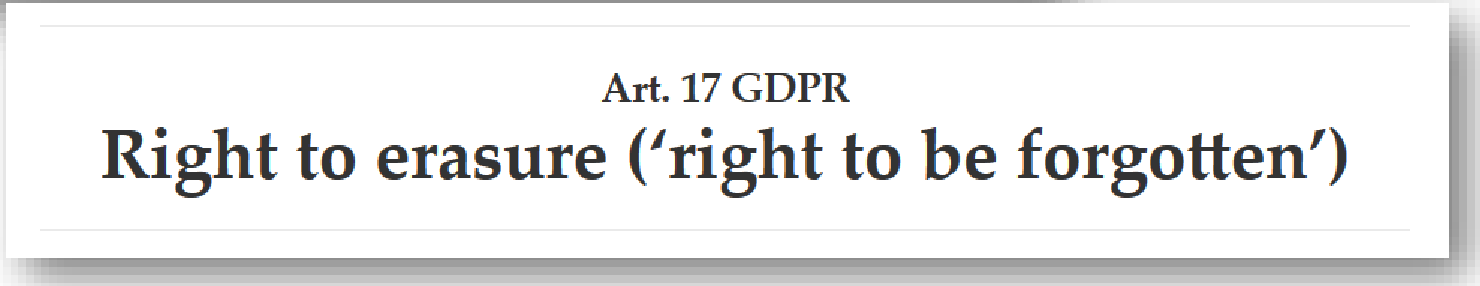
Art. 15 GDPR

Right of access by the data subject



Art. 16 GDPR

Right to rectification



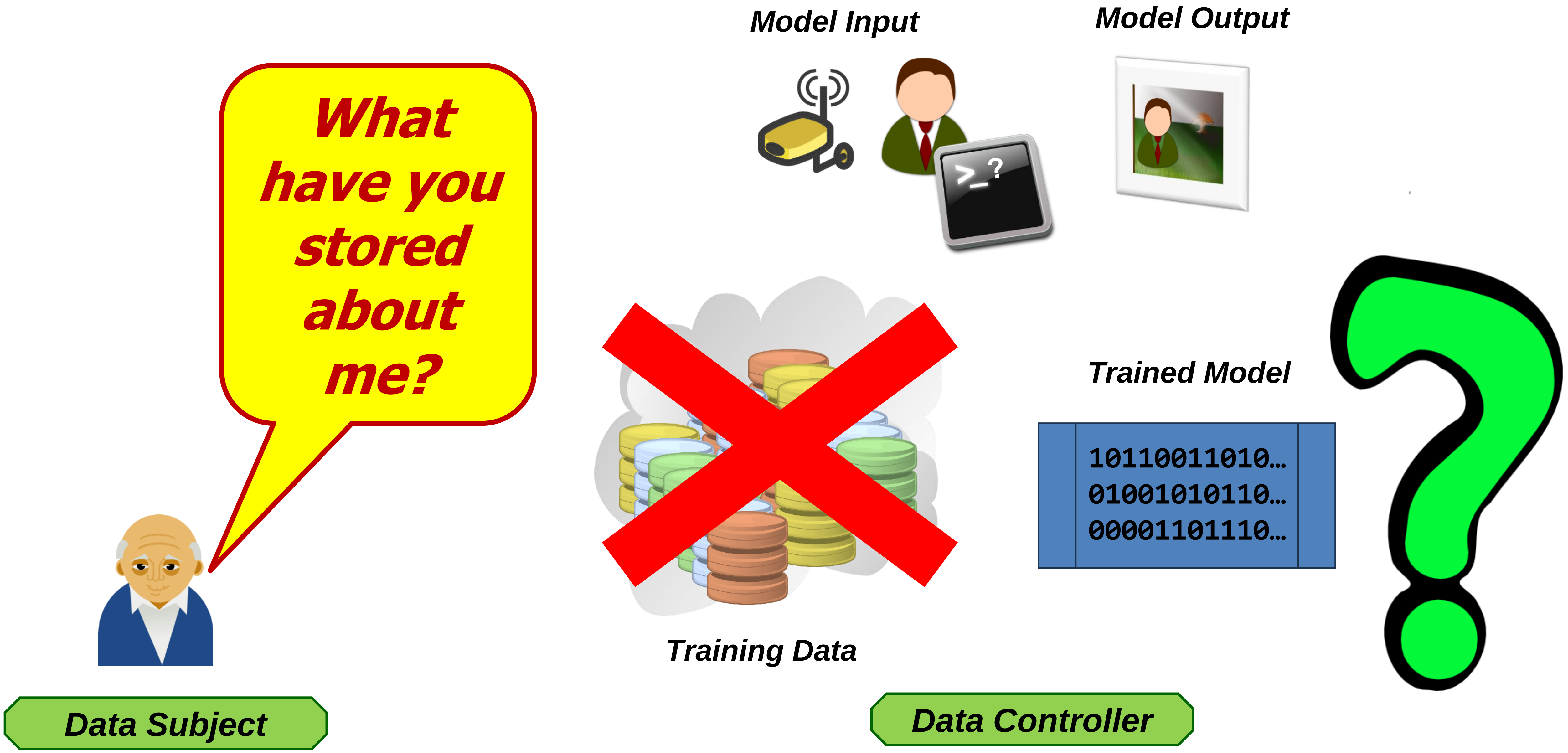
Art. 17 GDPR

Right to erasure (‘right to be forgotten’)

⋮

Data Subject Rights

Right of Access



What have you stored about me?

Model Input

Model Output

Trained Model

10110011010…
01001010110…
00001101110…

Training Data

Data Subject

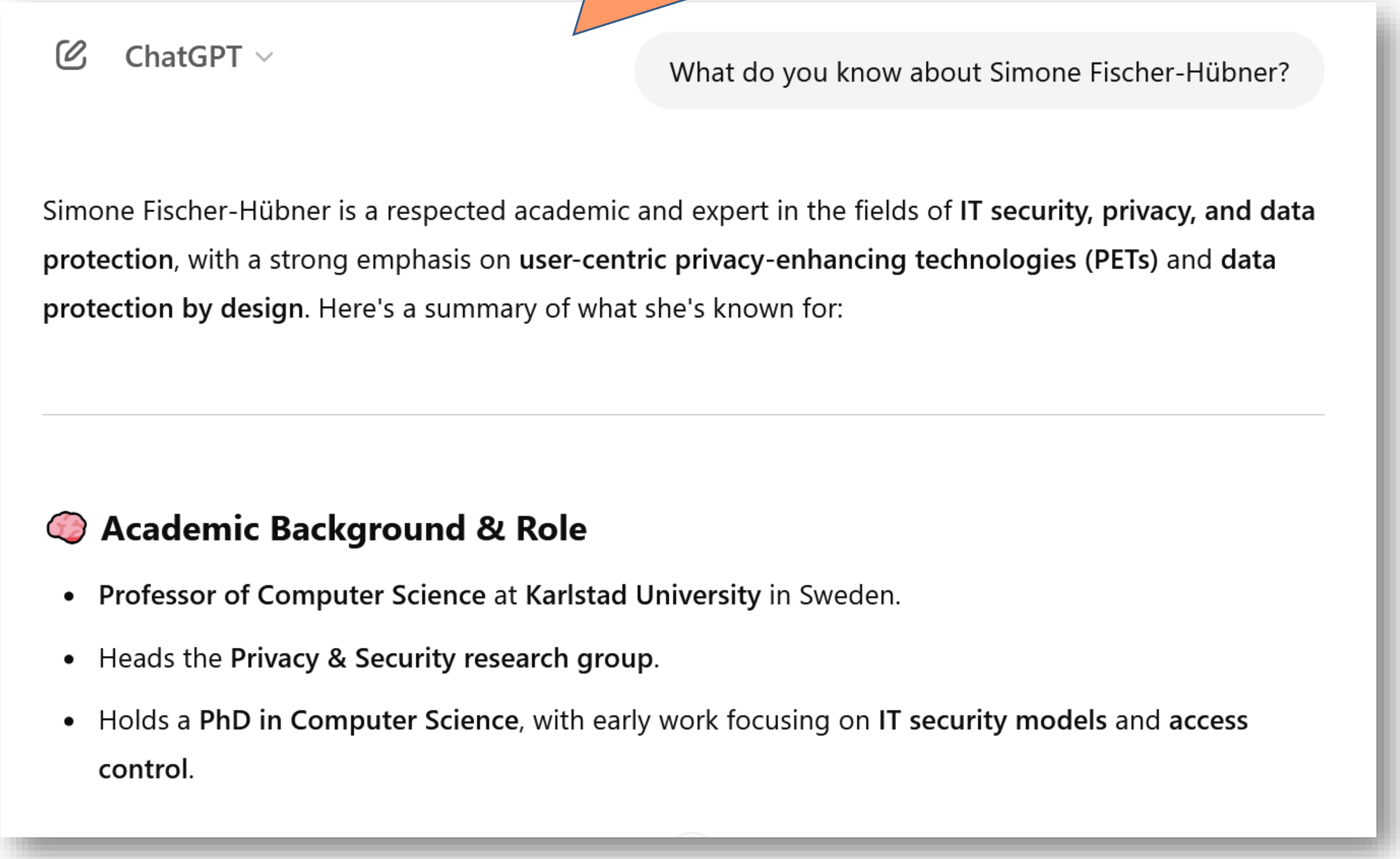
Data Controller

Right of Access

Option A:

Ask the model!

…not always complete or accurate!



What do you know about Simone Fischer-Hübner?

Simone Fischer-Hübner is a respected academic and expert in the fields of **IT security, privacy, and data protection**, with a strong emphasis on **user-centric privacy-enhancing technologies (PETs)** and **data protection by design**. Here's a summary of what she's known for:

🧠 Academic Background & Role

- **Professor of Computer Science** at **Karlstad University** in Sweden.
- Heads the **Privacy & Security research group**.
- Holds a **PhD in Computer Science**, with early work focusing on **IT security models** and **access control**.

Data Subject

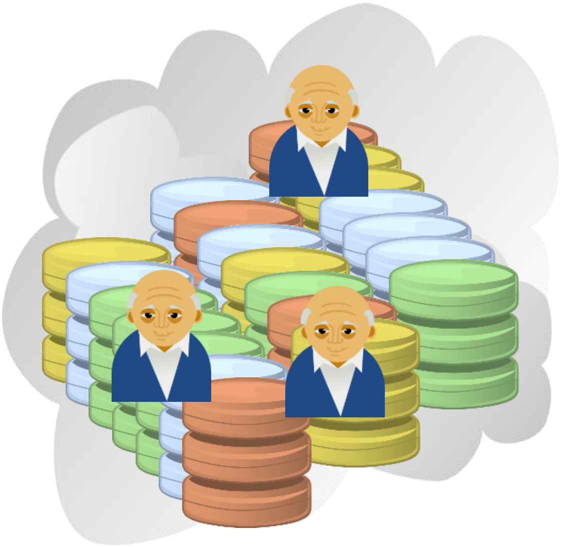
Right of Access

Option B:

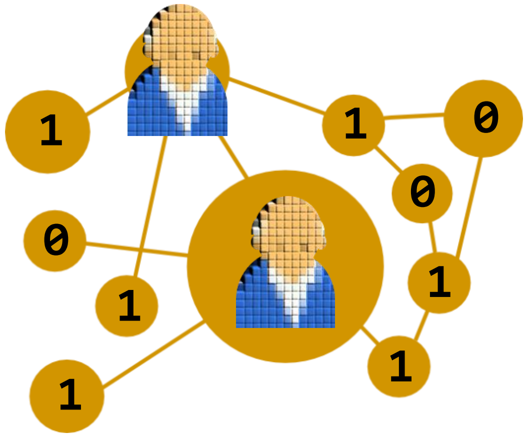
Inspect the model!

- Technically, not personal **data**!
- But personal **information**!
- **eXplainable AI (XAI)**: research domain of its own!

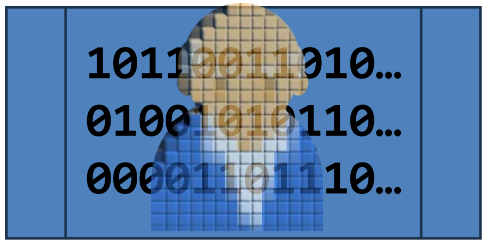
Data Subject



Training Data

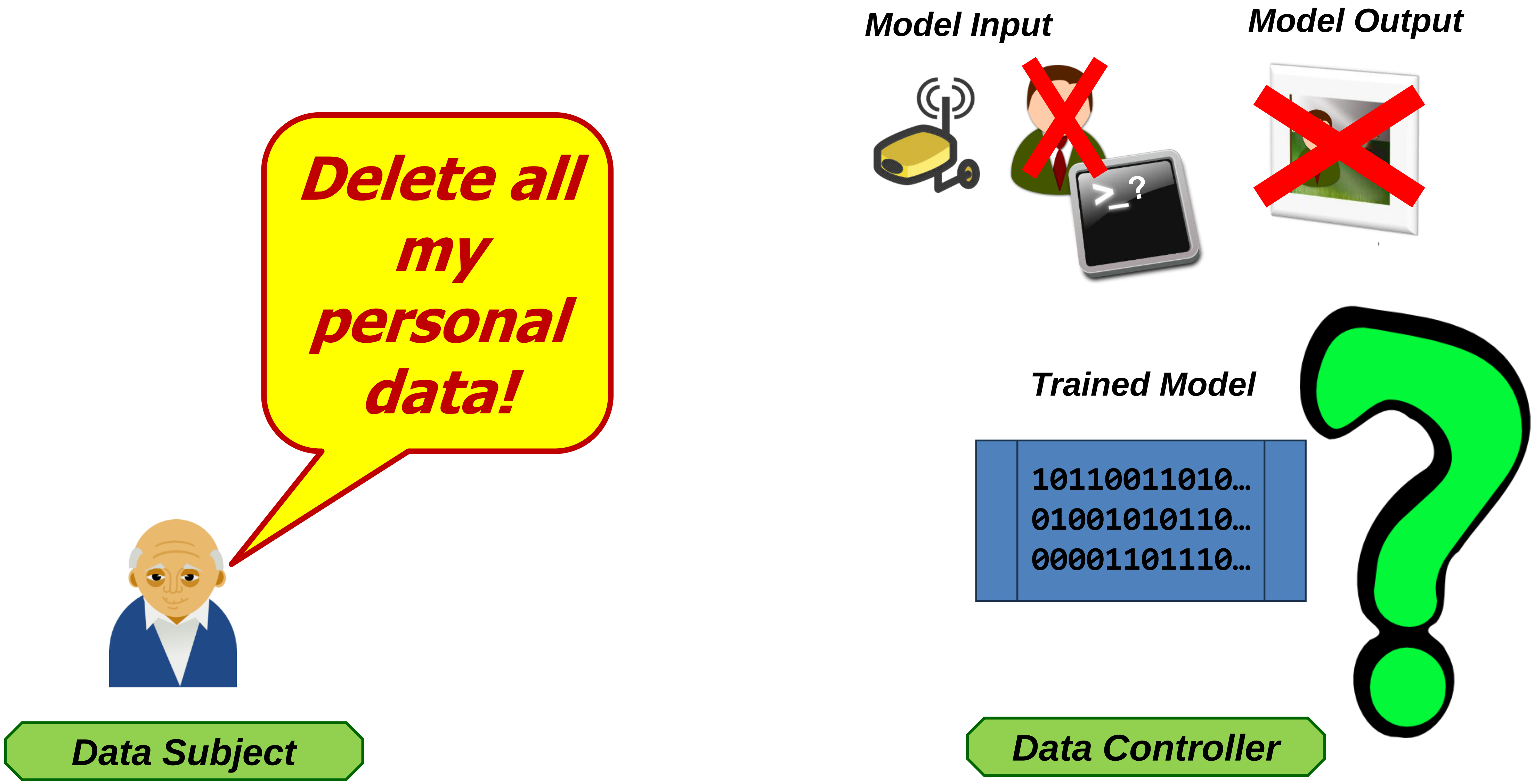


Information



Trained Model

Right to Erasure ("Right to be forgotten!")



Model Input

Model Output

Trained Model

```
10110011010…
01001010110…
00001101110…
```

Delete all my personal data!

Data Subject

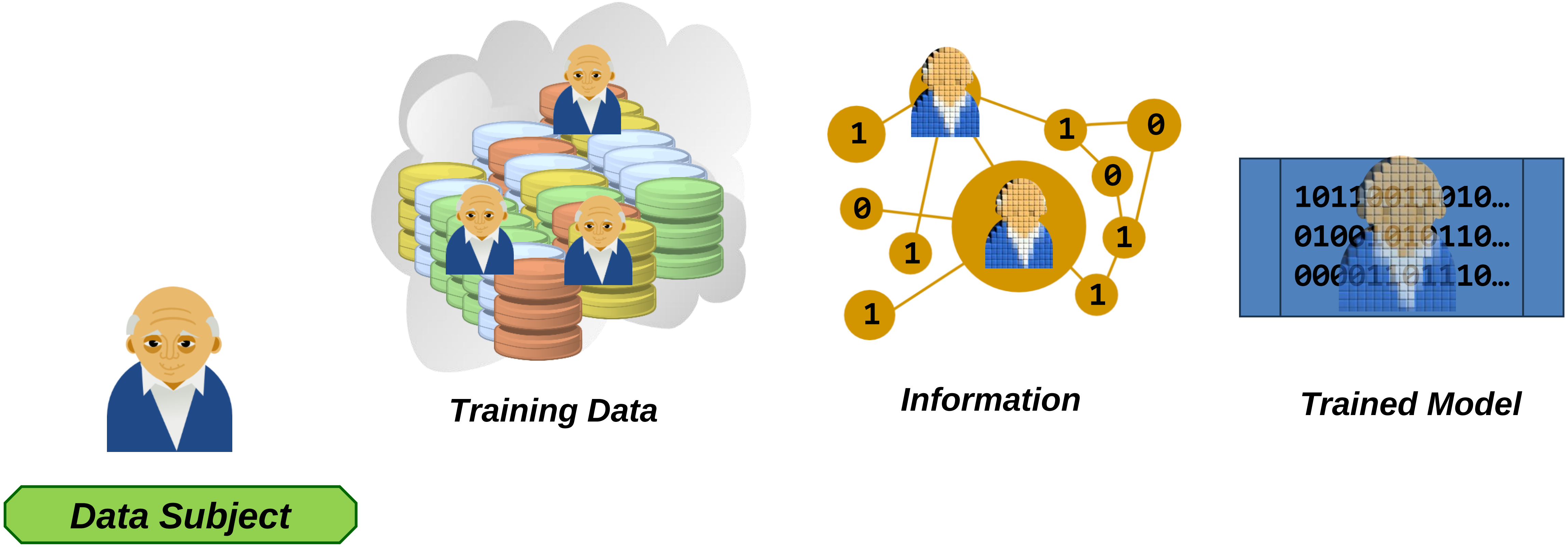
Data Controller

Right to Erasure ("Right to be forgotten!")

Option A:

Remove data subject from training data,

then start from scratch!



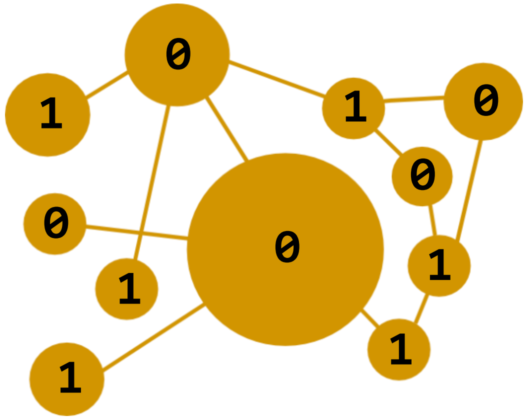
Right to Erasure ("Right to be forgotten!")

Option A:

Remove data subject from training data,

then start from scratch!

Redacted Training Data



New Information

Data Subject

Right to Erasure ("Right to be forgotten!")

Option A:

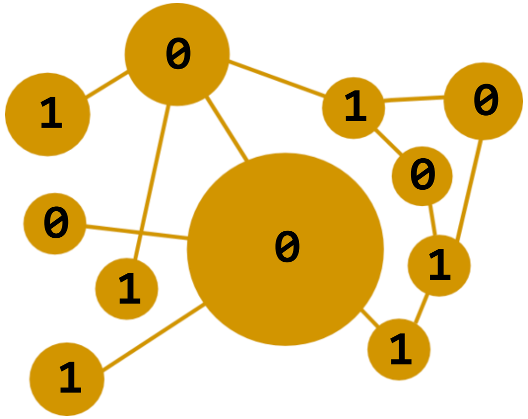
Remove data subject from training data,

then start from scratch!

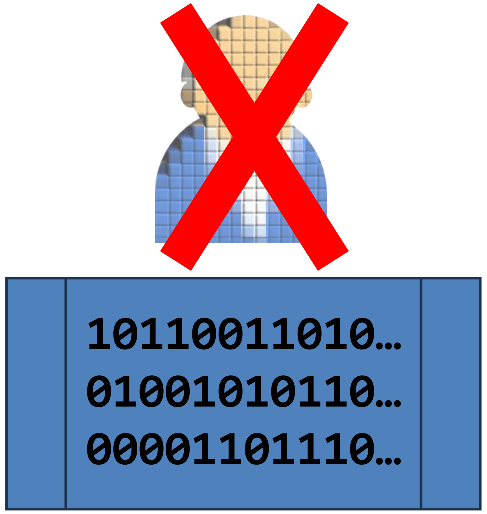
…not likely to happen for LLMs!

Data Subject

Redacted Training Data



New Information



New Trained Model

Right to Erasure ("Right to be forgotten!")

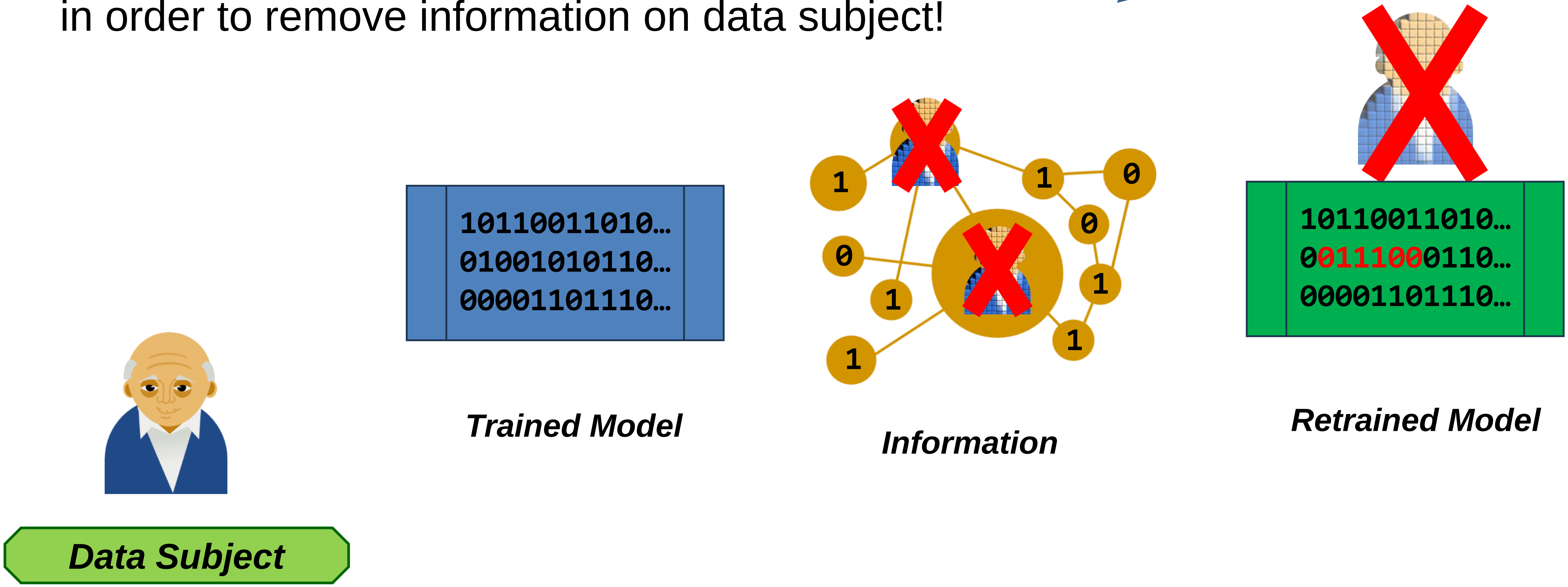
Option B:

Retrain the model

in order to remove information on data subject!

…shown to be ineffective sometimes!

…how to distribute the retrained model?



10110011010…
01001010110…
00001101110…

Trained Model

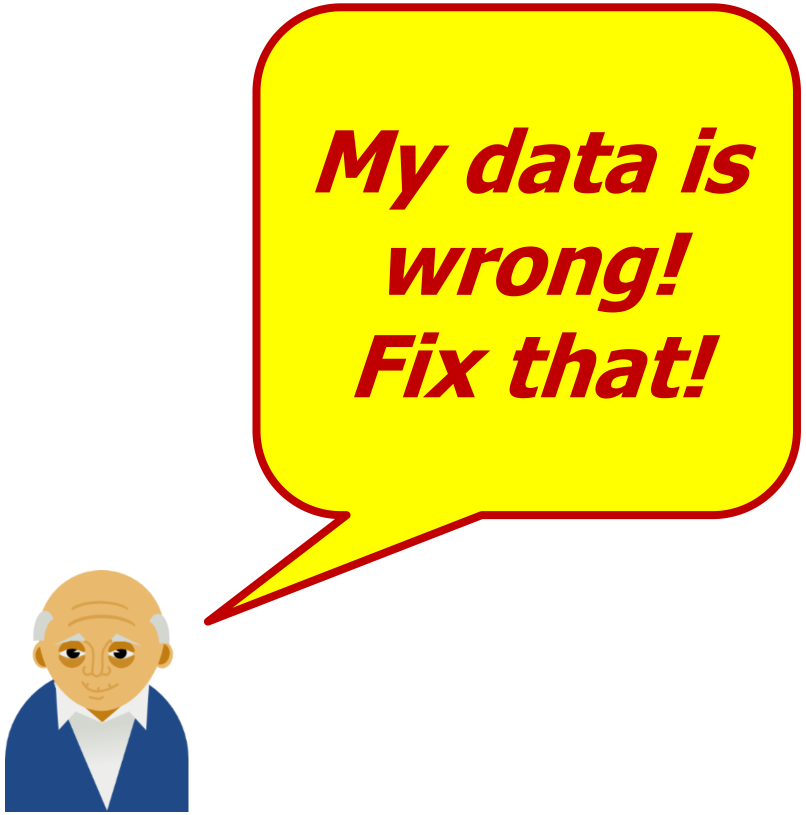
Information

10110011010…
00111000110…
00001101110…

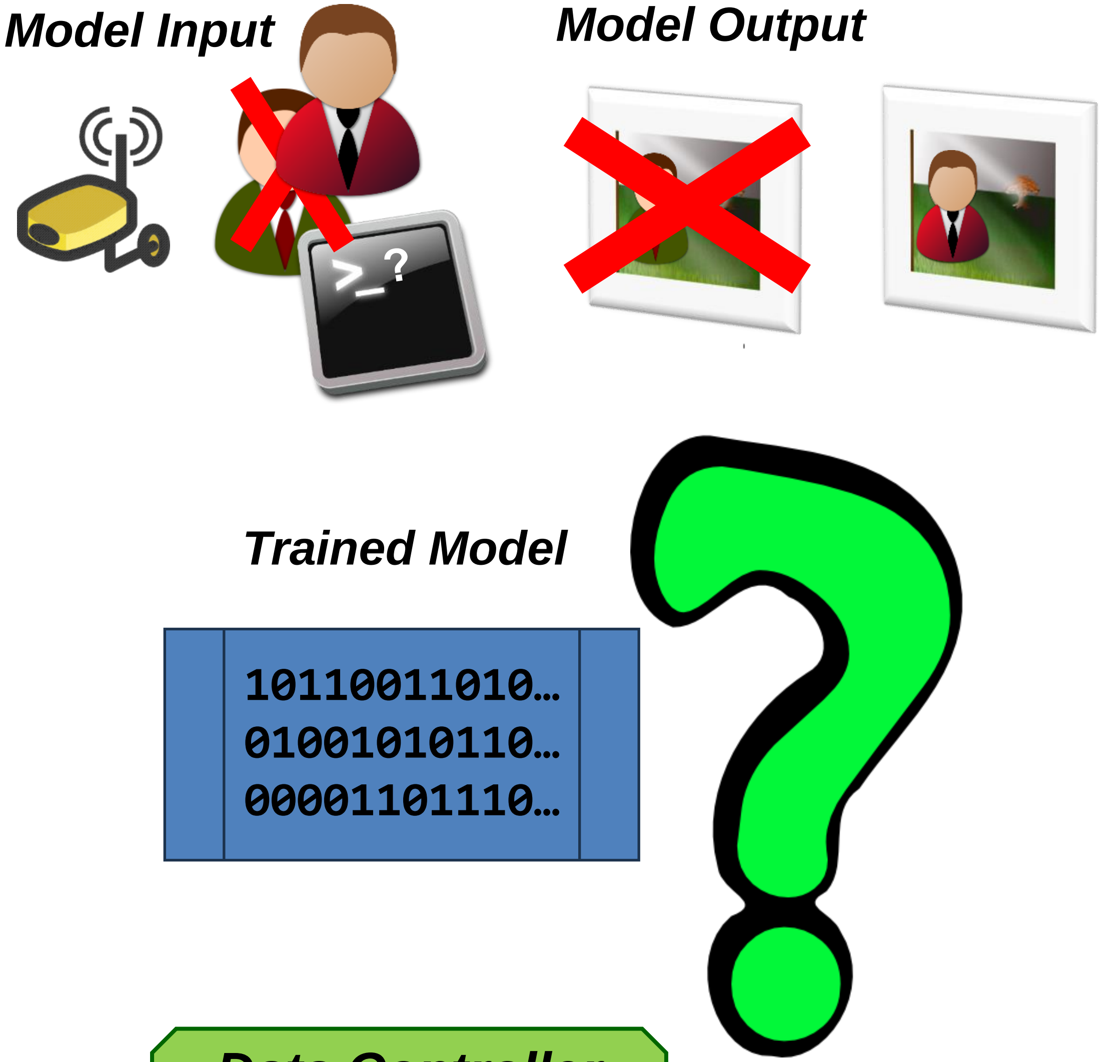
Retrained Model

Data Subject

Right to Rectification



Data Subject



Model Input

Model Output

Trained Model

10110011010…
01001010110…
00001101110…

Data Controller

Right to Rectification

Option A:

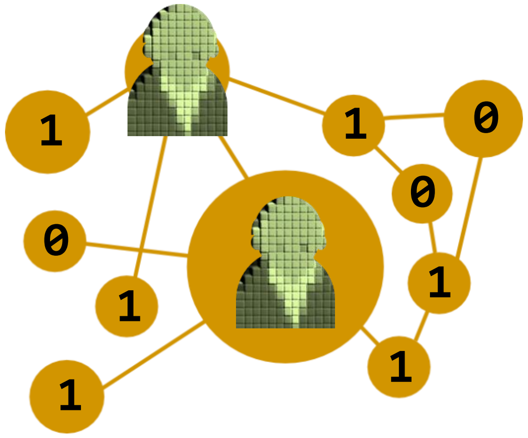
Replace data subject data in training data,

then start from scratch!

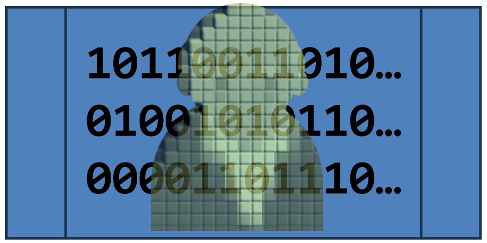
…not likely to happen for LLMs!

Data Subject

Training Data



Information



Trained Model

Right to Rectification

Option B:

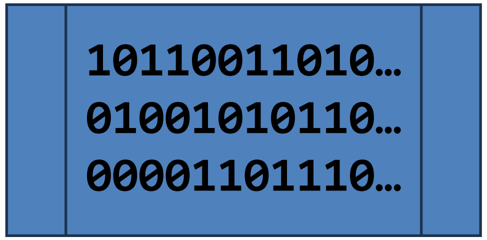
Retrain the model

in order to replace information on data subject!

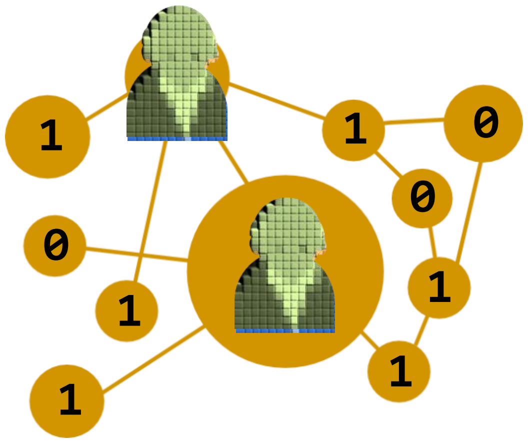
…shown to be ineffective often!

…how to distribute the retrained model?

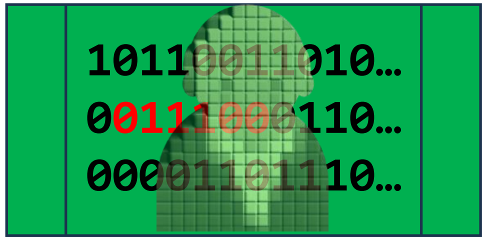
Data Subject



Trained Model



Information



Retrained Model

Right to Data Portability

I want to use a different LLM now! Give me my data!

Data Subject

Trained Model

```
10110011010…
01001010110…
00001101110…
```

…lack of interoperability!

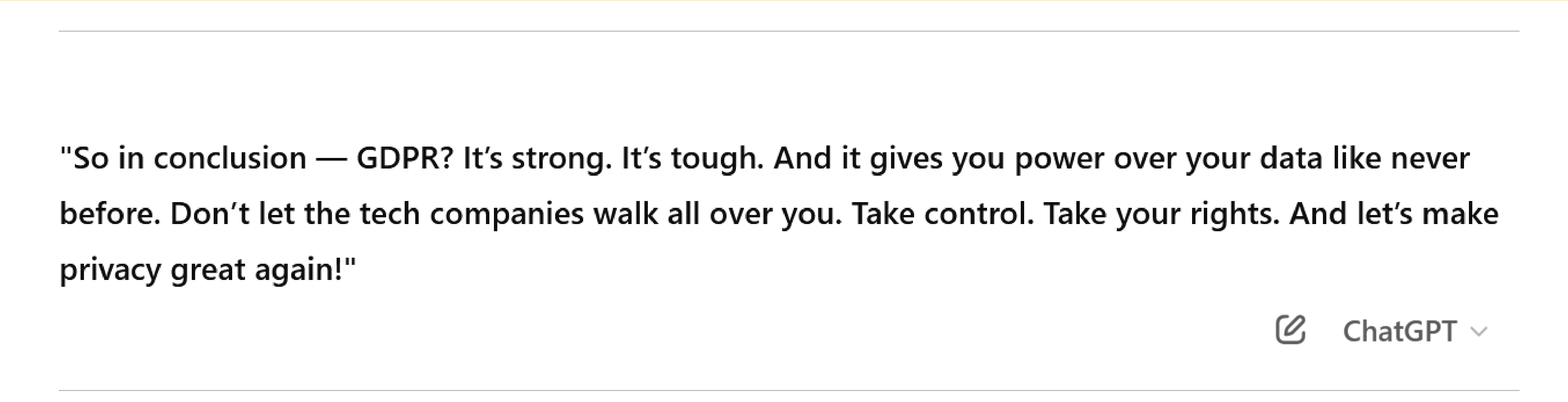
…lack of demand (yet)!

Data Controller

Conclusions

- Data Subject Rights against ML models are hard to enforce!
- It is not yet clear if and when which ML models are (not) personal data!
- It is not yet clear what the Right of Access means towards ML models
- It is not yet clear if and how personal data can be removed or replaced in ML models without a full re-training from scratch!
- It *is* clear that it‘s not going to be trivial!

CC-BY-4.0



"So in conclusion — GDPR? It’s strong. It’s tough. And it gives you power over your data like never before. Don’t let the tech companies walk all over you. Take control. Take your rights. And let’s make privacy great again!"

Meiko Jensen
Meiko.Jensen@kau.se

CC-BY-4.0

Meiko Jensen

Meiko.Jensen@kau.se