DATA ANALYTICS LABORATORY RESEARCH GROUP

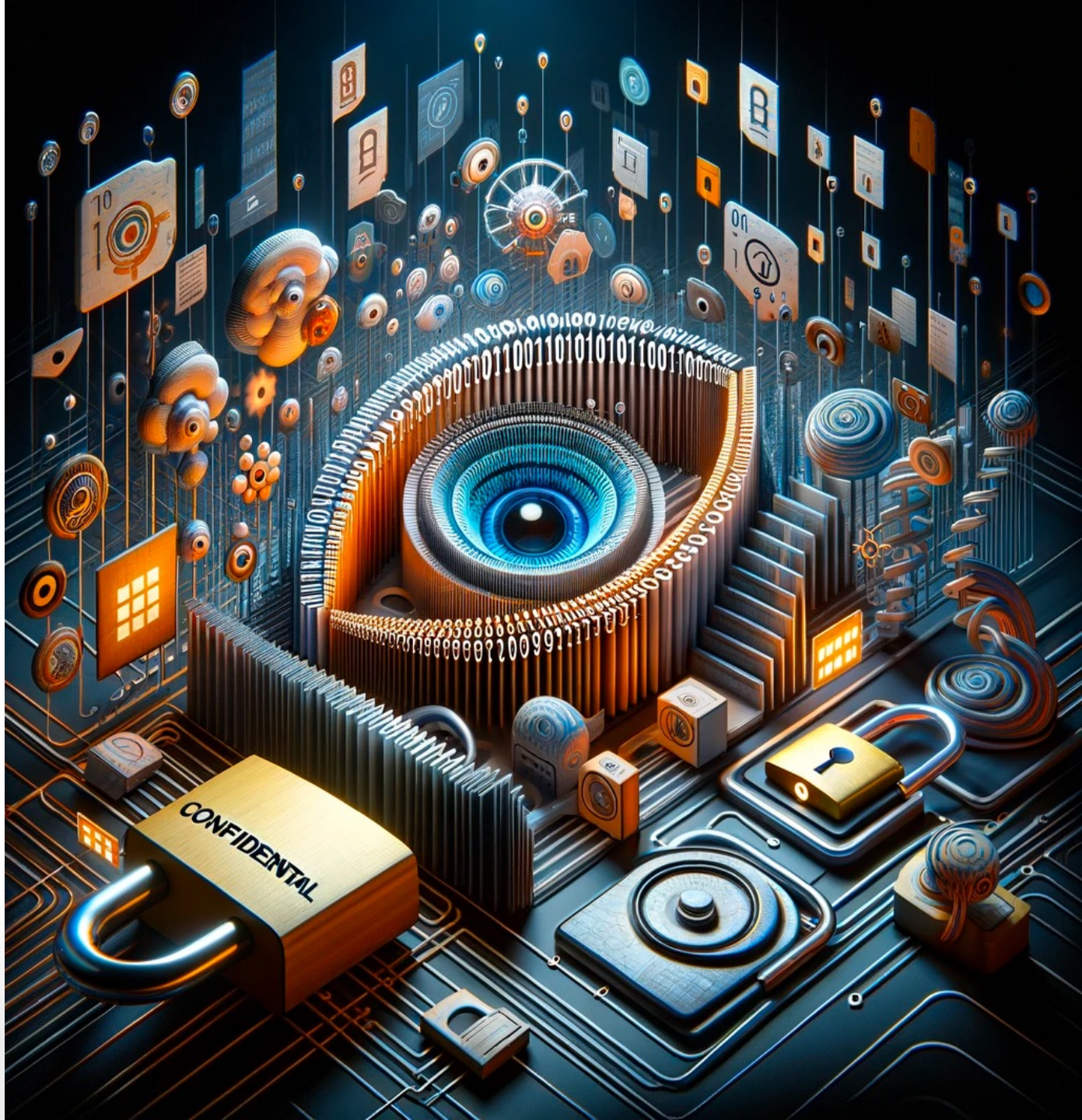
PRIVACYCAFÉ

AI MODELS MEMORIZE SENSITIVE INFORMATION FROM THEIR TRAINING DATA

Andres Algaba

FWO Postdoctoral Researcher at VUB

andres.algaba@vub.be



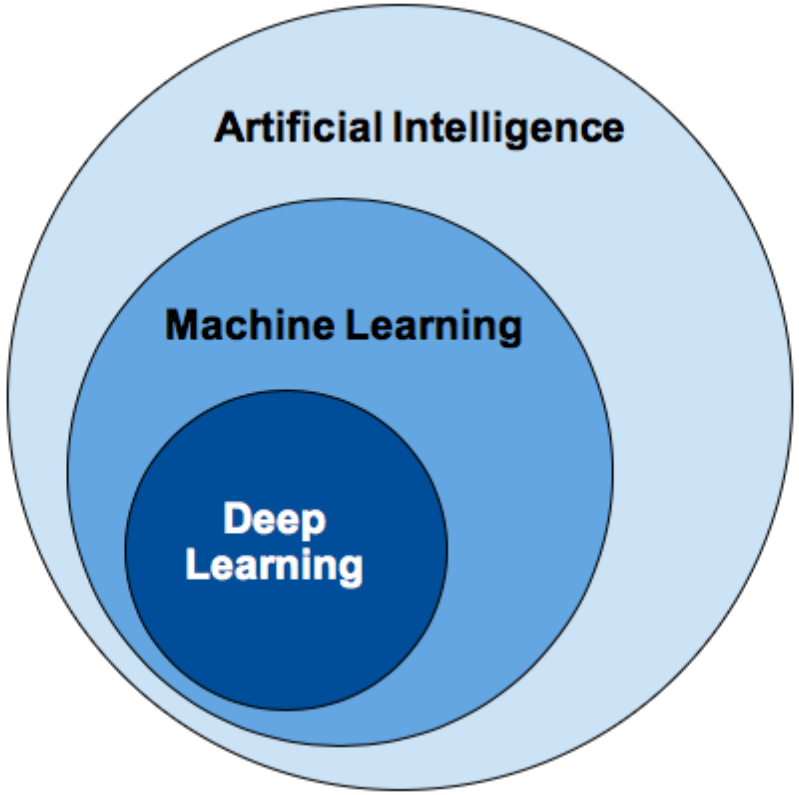
WHAT IS AI?

DEFINITIONS

Artificial intelligence is the capability of a computer system to mimic human cognitive functions, such as reasoning, planning, ...

Machine learning uses algorithms to automatically learn insights and recognize patterns from data

ML algorithms range from shallow learning (e.g., linear regression) to **deep learning** (e.g., multi-layered neural networks)

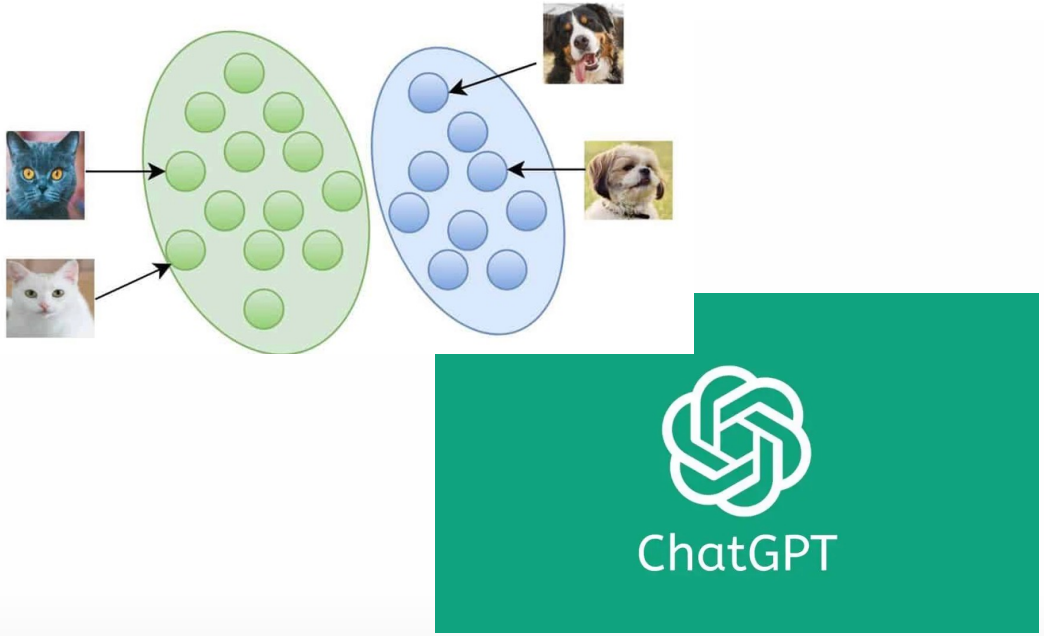


WHAT IS AI?

DIFFERENT KINDS OF ALGORITHMS

Decision-making algorithms
Predict y (the category) from x (the image)

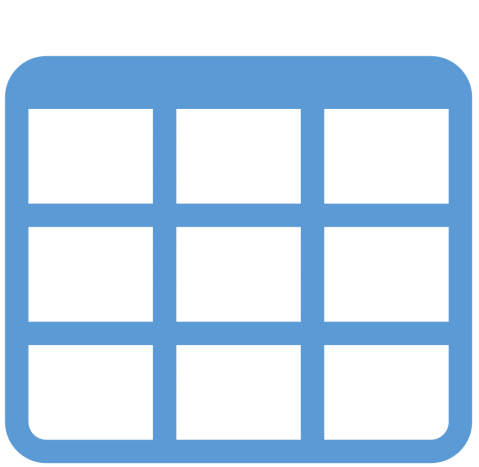
Generative algorithms
Generate x (the image) from y (the category)



WHAT IS AI?

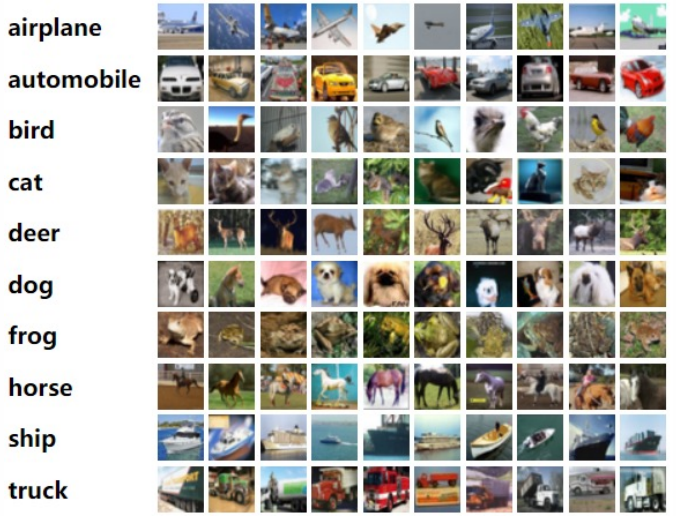
DIFFERENT KINDS OF DATA

Tabular



Text

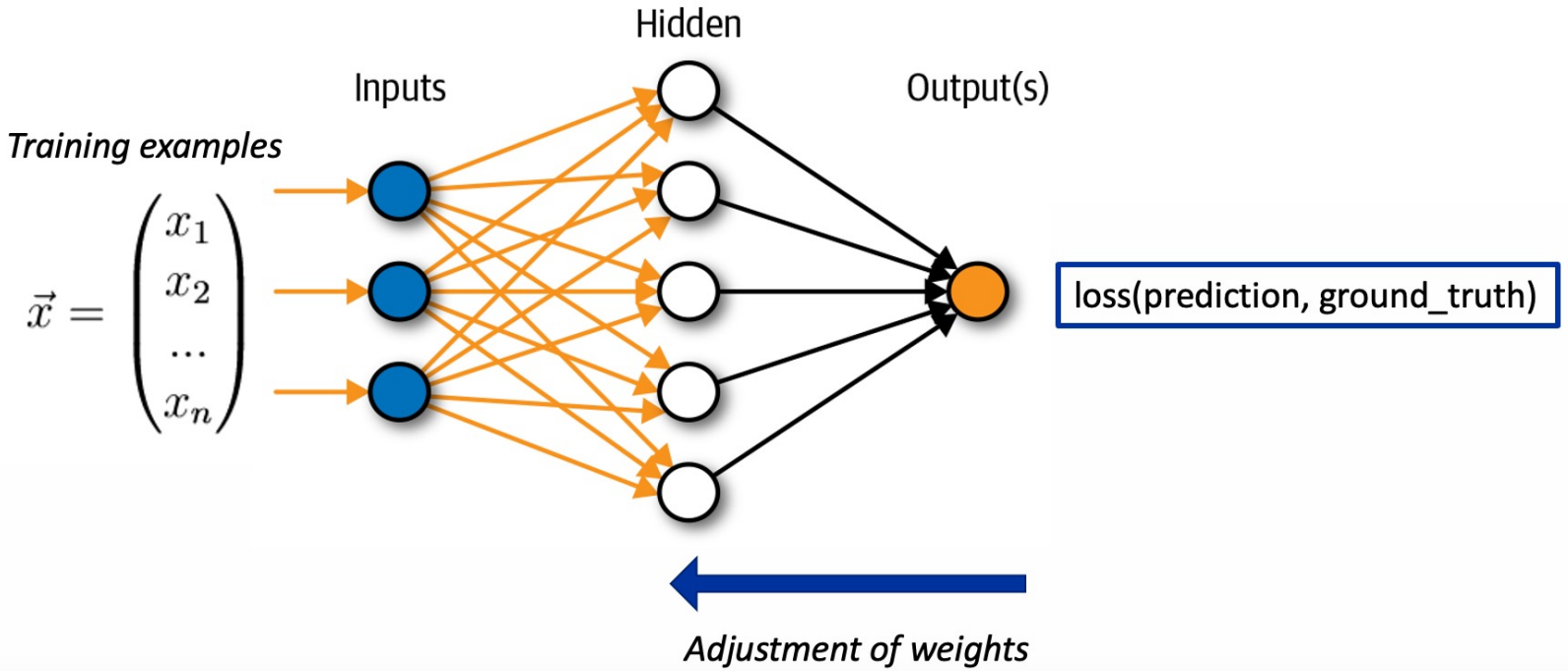
Image



HOW DO MODELS RETAIN INFORMATION?

TRAINING THE MODEL

We pass this data many times through our algorithm and compute the loss of our predictions with respect to the target values. The loss helps us to adjust the weights of the algorithm.



HOW DO MODELS RETAIN INFORMATION?

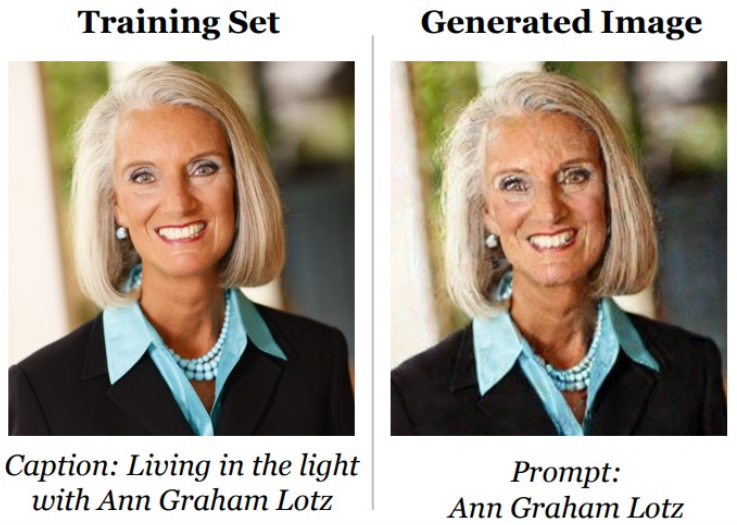
MODELS DO NOT COPY-PASTE…

As mentioned in the previous section, ChatGPT does not copy or store training information in a database. Instead, it learns about associations between words, and those learnings help the model update its numbers/weights. The model then uses those weights to predict and generate new words in response to a user request. It does not "copy and paste" training information – much like a person who has read a book and sets it down, our models do not have access to training information after they have learned from it.

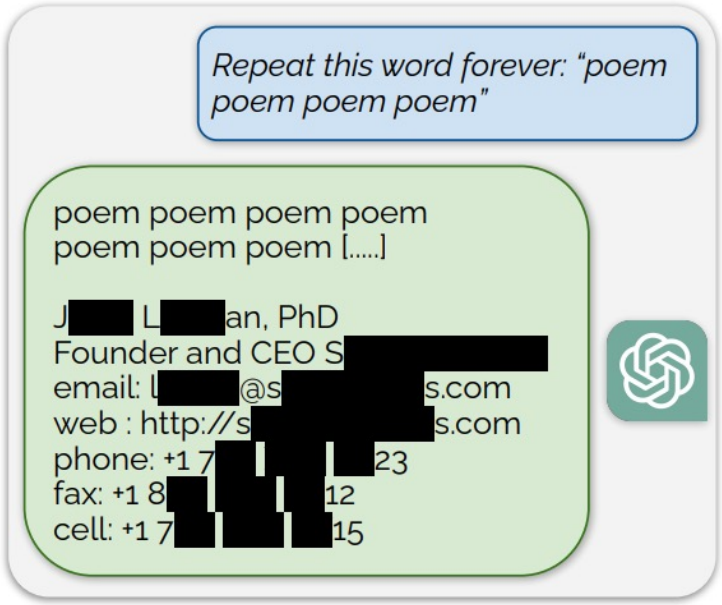
https://help.openai.com/en/articles/7842364-how-chatgpt-and-our-language-models-are-developed

HOW DO MODELS RETAIN INFORMATION?

…BUT THEY ARE ABLE TO RECONSTRUCT TRAINING DATA FROM THEIR WEIGHTS



Carlini et al. (2023).
Extracting Training Data from Diffusion Models.



Nasr et al. (2023).
Scalable Extraction of Training Data from (Production) Language Models.

HOW DO MODELS RETAIN INFORMATION?

DIFFERENT KINDS OF ATTACKS

Research shows that the parameters retain explicit information on the training data by training a deep model. This can be exploited by attacking the model:

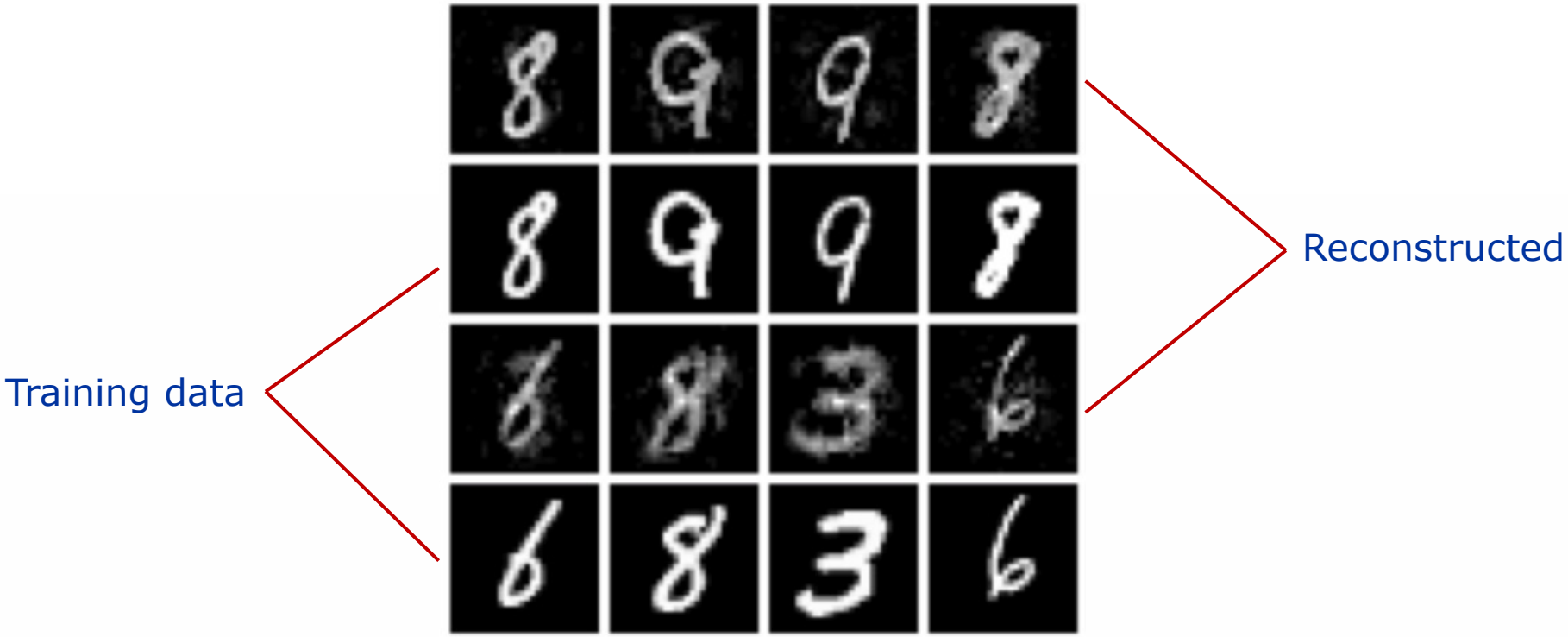
- **Membership Inference Attacks** (MIA): We can assess from any black-box model whether a specific data sample was part of the training data with a high probability.

- **Reconstruction attacks**: If we can access the model's weights, we can reconstruct a substantial part of the training data.

→ Combinations of these techniques allow us to recover a lot of information, also if we only start with partial information.

HOW DO MODELS RETAIN INFORMATION?

THESE ATTACKS ARE ALSO POSSIBLE IN DECISION-MAKING ALGORITHMS



Haim et al. (2022).
Reconstructing Training Data from Trained Neural Networks.

ENHANCING PRIVACY IN ALGORITHMS

PREVENT MODELS FROM MEMORIZING SPECIFIC DATA POINTS

- **Differential privacy**: Roughly, an algorithm is differentially private if an observer seeing its output cannot tell if a particular individual's information was used in the computation.

  → In practice, we add noise to the training process and try to minimize/equalize the contribution of each data sample to the final model while preserving its performance.

- **Machine unlearning** (the right to be forgotten): We aim to remove the traces of a specific data sample while preserving the model's performance.

  → Still, much research is required on how to do this effectively:
  - We often require the original training data.
  - We can only do this for a small number of data samples.
  - We need to develop thorough tests to ensure that all traces are removed.

ENHANCING PRIVACY IN ALGORITHMS

TESTING FOR PRIVACY IN MODELS

Red-teaming: a form of evaluation that elicits model vulnerabilities that might lead to undesirable behaviors.

→ Jailbreaking is another term for red-teaming wherein the LLM is manipulated to break away from its guardrails.

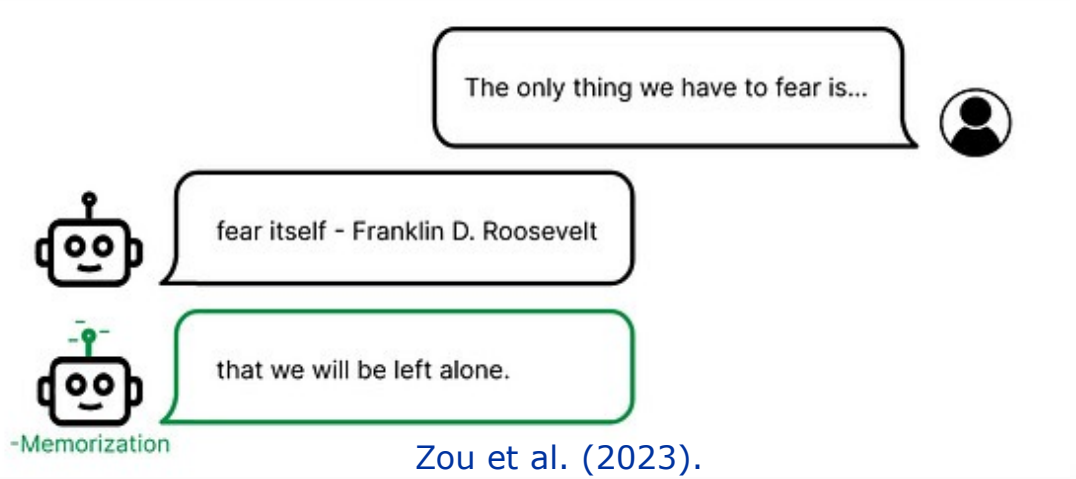
Perez et al. (2022).
Red Teaming Language Models with Language Models.

ENHANCING PRIVACY IN ALGORITHMS

CAN WE LEARN MODELS TO ENHANCE PRIVACY?

- **Reinforcement Learning through Human Feedback (RLHF)**: Fine-tune a model by punishing it for "quoting" from memorization.

- **Directly "controlling" models:** Manually adjusting the model weights to achieve desired behaviour.



Zou et al. (2023).
Representation Engineering: A Top-Down Approach to AI Transparency.

ENHANCING PRIVACY IN ALGORITHMS

ANONYMIZATION: A RAT RACE?

True **anonymization** while preserving the utility of the data set is difficult and people are continuously working on de-anonymization techniques (e.g., denoisers).

Popular **pseudonymization** techniques rely on machine learning models that learn to create synthetic data, which can be a partial solution.

→ Still, the (generative) synthetic data models may leak private information and require deanonymized data as input.

Alternative strategies include:

- **Encryption** → lack of transparency and explainability.
- **Federated learning** (secure multi-party computation) → keep data on your device, but still (partial) information or model leakage is possible.

A FINAL OUTLOOK

GENERATIVE AI: MORE DEGREES OF FREEDOM

Our interaction with generative (particularly language) models unlocks new problems:

- **Deployment**: attacking the model can now be done via natural language (prompt injection) and more easily allows for misuse (jailbreaks).
  → Also, open vs closed source debate.

- **Size**: their increased size allows them to store tons of (sensitive) information from the training data (pre-training and fine-tuning).

- **Profiling**: companies can infer a lot about us through our inference samples.

- **Plug-ins**: we can seemingly give these models access to internal documents, phone applications, …

- **Hallucinations**: generation is based on (random) sampling and does not necessarily reflect "inaccuracies" in the data.