Lifelong Learning and Large Language Models

Çağatay Yıldız
Tübingen AI Center & University of Tübingen

ML in Science Conference
Tübingen, 09.07.2024

Outline

1. **Motivation for lifelong machine learning**

2. **Academic lifelong learning**

3. **Our works**
    - **Disentangled learning**
    - **Lifelong knowledge organisation**
    - **Continual pretraining of LLMs**

4. **Final thoughts**

Motor development in babies

1. Reaching

Video credit: https://www.youtube.com/watch?v=RpssleVZfPA

Motor development in babies

2. Grasping

Video credit: https://www.youtube.com/watch?v=SHrM9X_3OMw

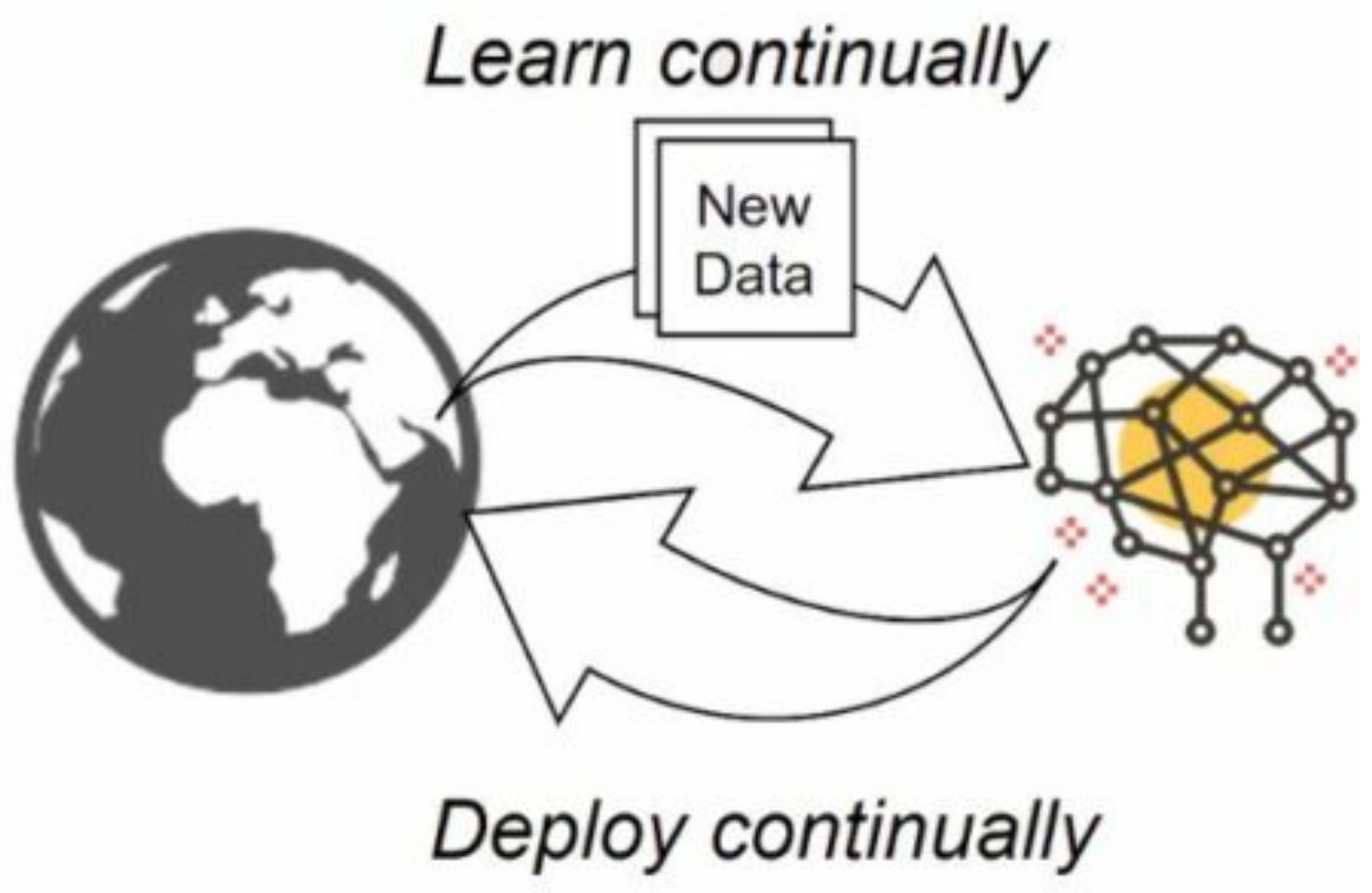
Motor development in babies

3. Pincer grasp

Video credit: https://www.youtube.com/watch?v=pymSN_42DU0

Lifelong learning

Lifelong learning



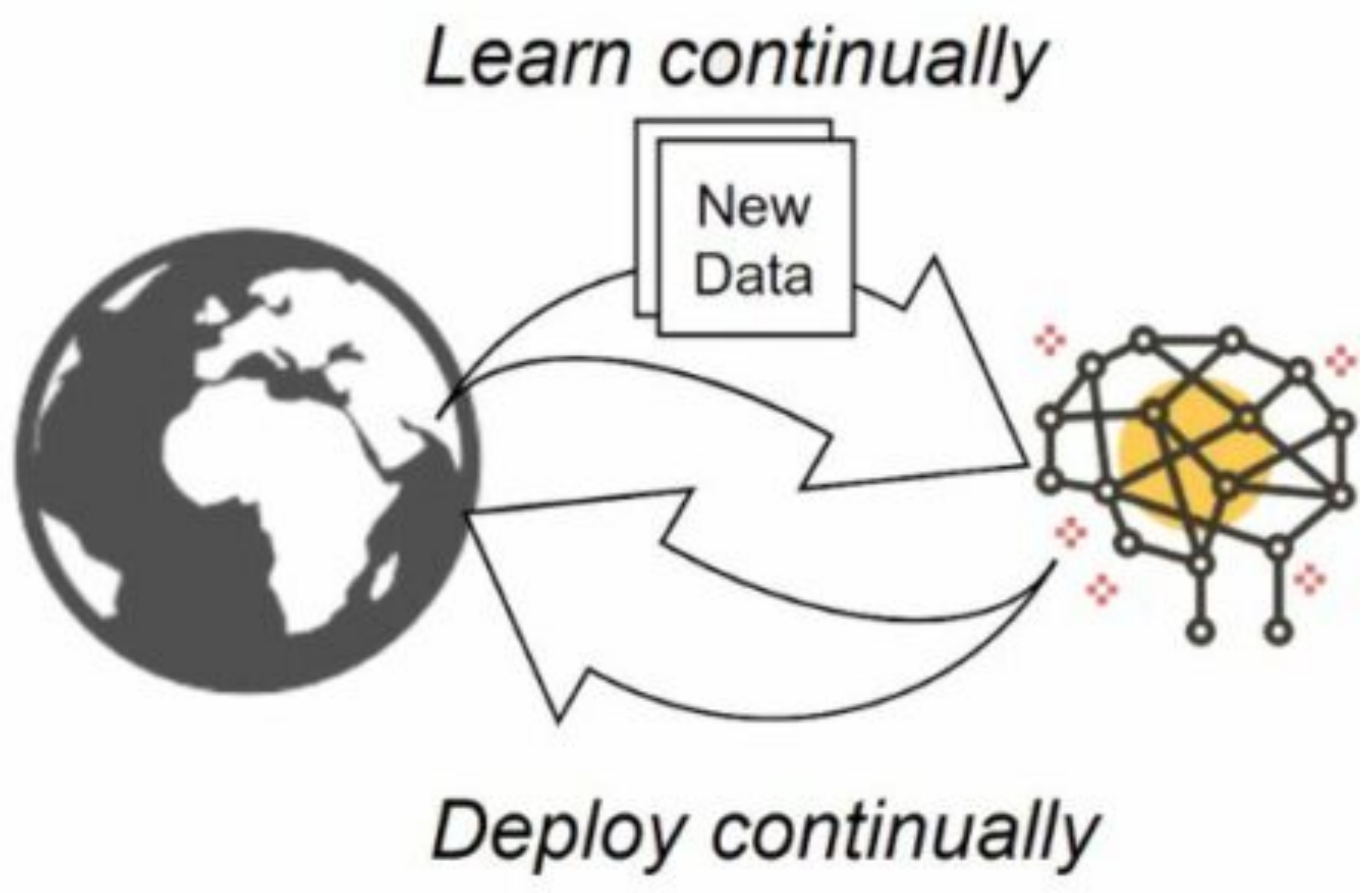
✓ No forgetting

✓ Forward transfer

Image credit: https://imerit.net/blog/a-complete-introduction-to-continual-learning

Lifelong vs static learning

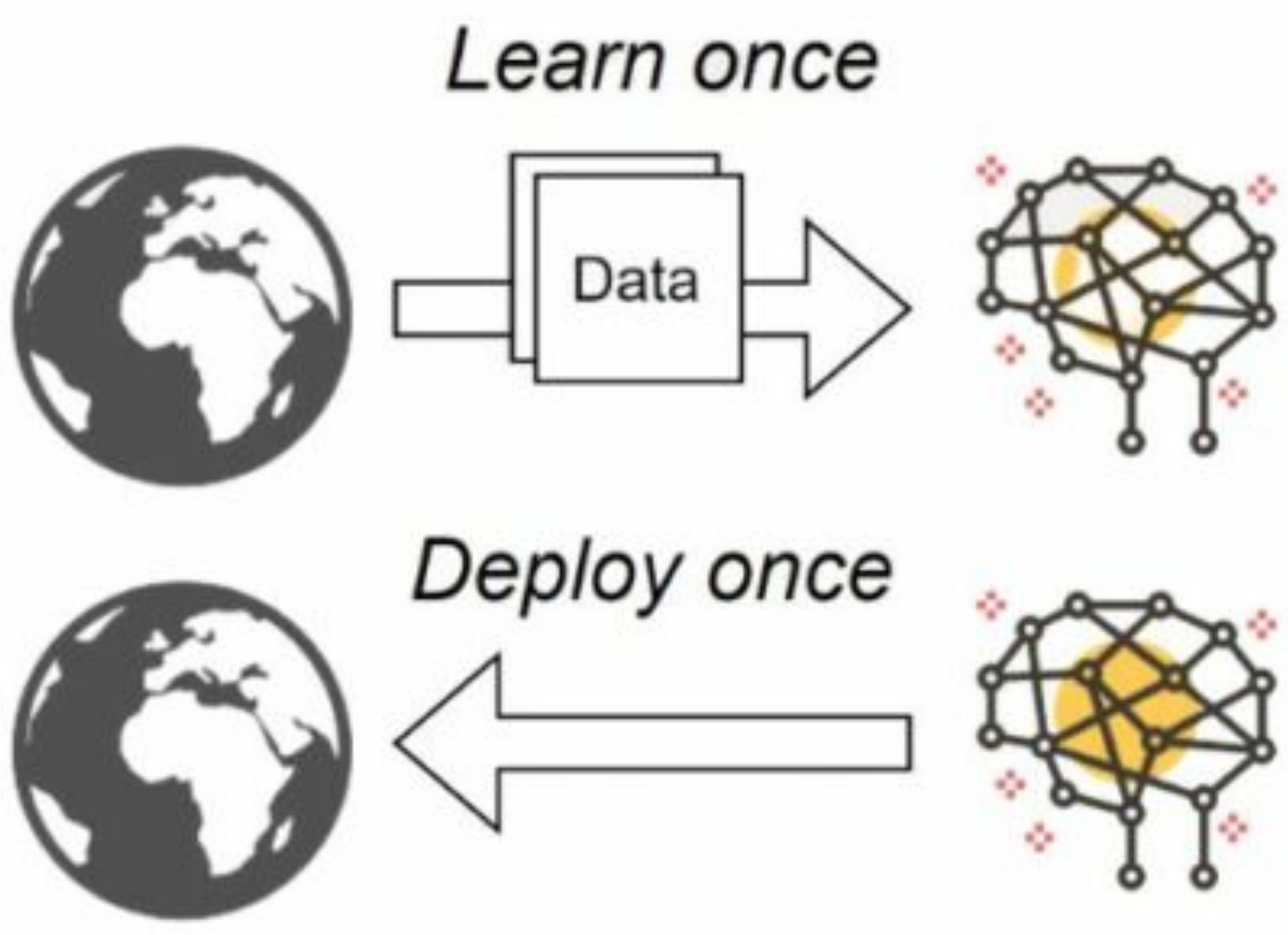
Lifelong ML



✓ No forgetting

✓ Forward transfer

Static ML



✓ Good test performance

Image credit: https://imerit.net/blog/a-complete-introduction-to-continual-learning

Lifelong learning examples

1. Digital assistants
2. Natural language processing
3. Autonomous vehicles
4. Robotics

…

Outline

1. Motivation for lifelong machine learning

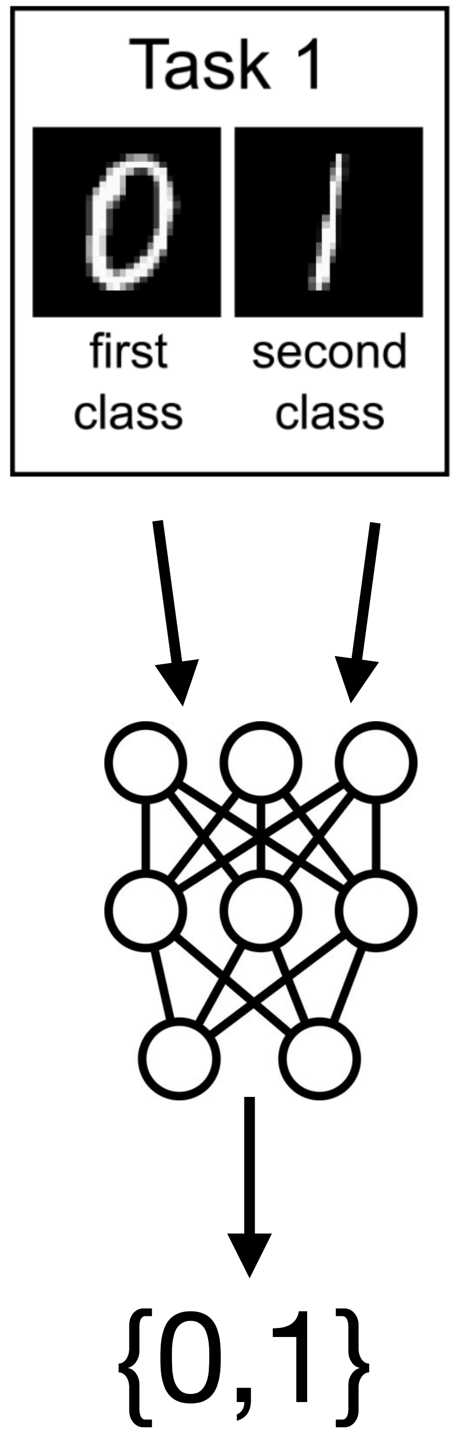
2. **Academic lifelong learning**

3. Our works
    - Disentangled learning
    - Lifelong knowledge organisation
    - Continual pretraining of LLMs

4. Final thoughts

Academic lifelong learning [1]

Training



Testing

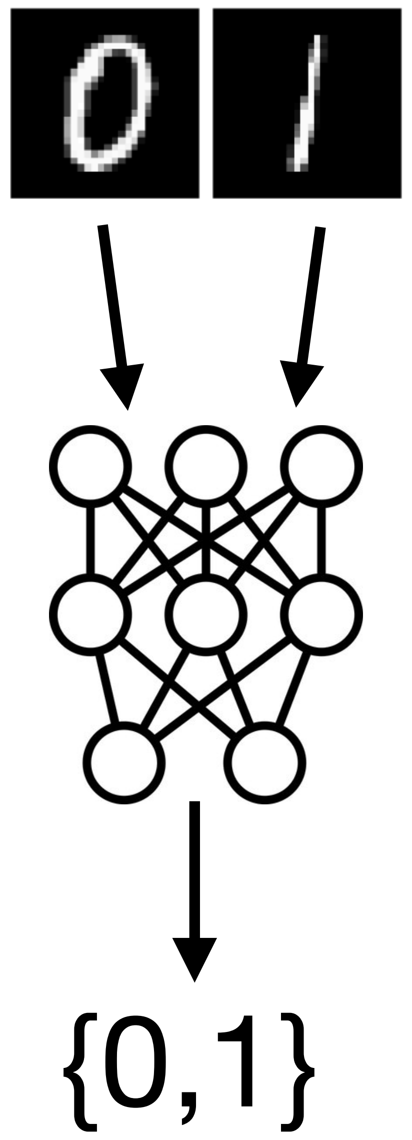
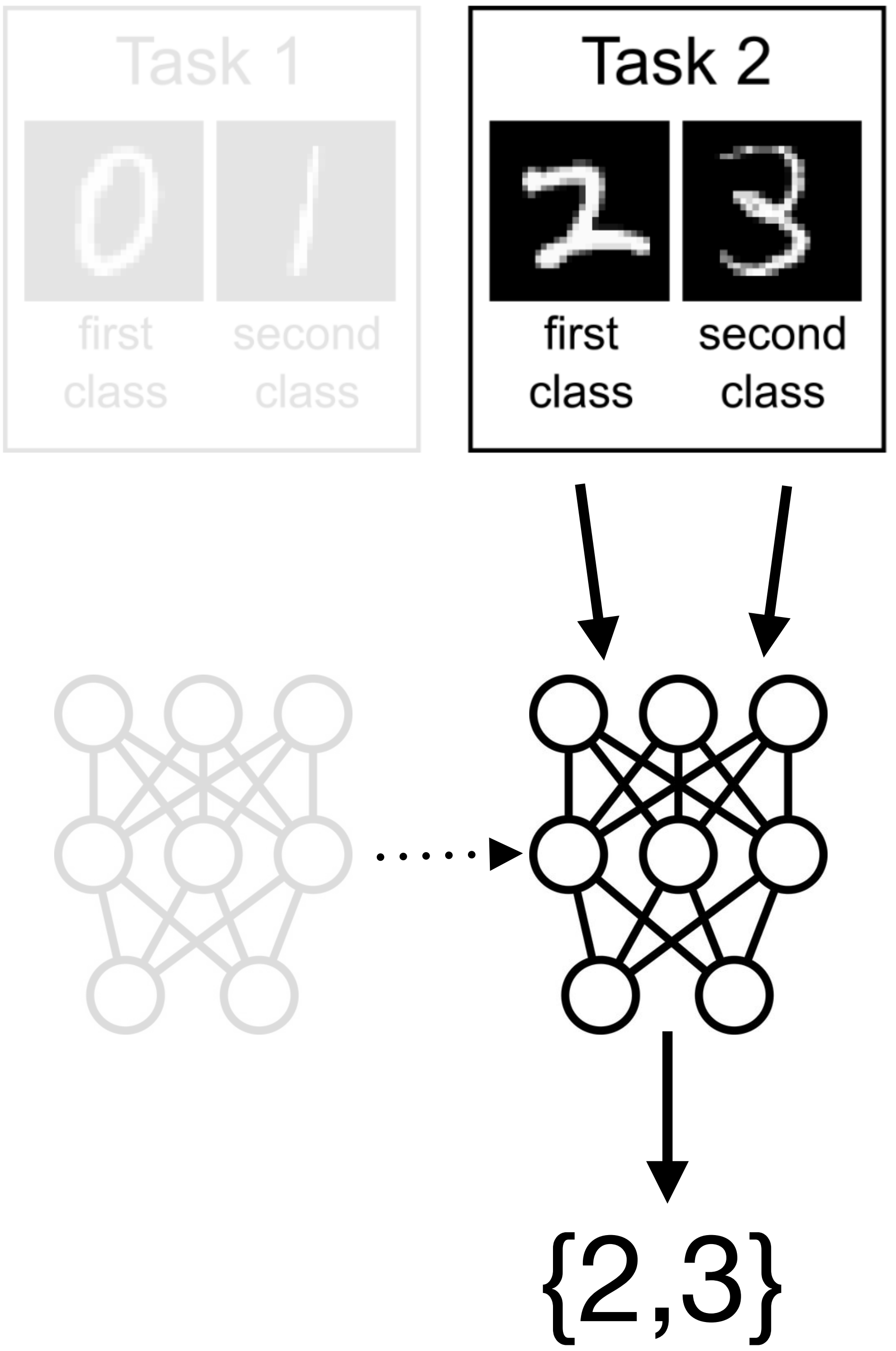


Image credit: [1] van de Ven, G. & Tolias, A. 2019. Three scenarios for continual learning, https://arxiv.org/abs/1904.07734

Academic lifelong learning

Training



Testing

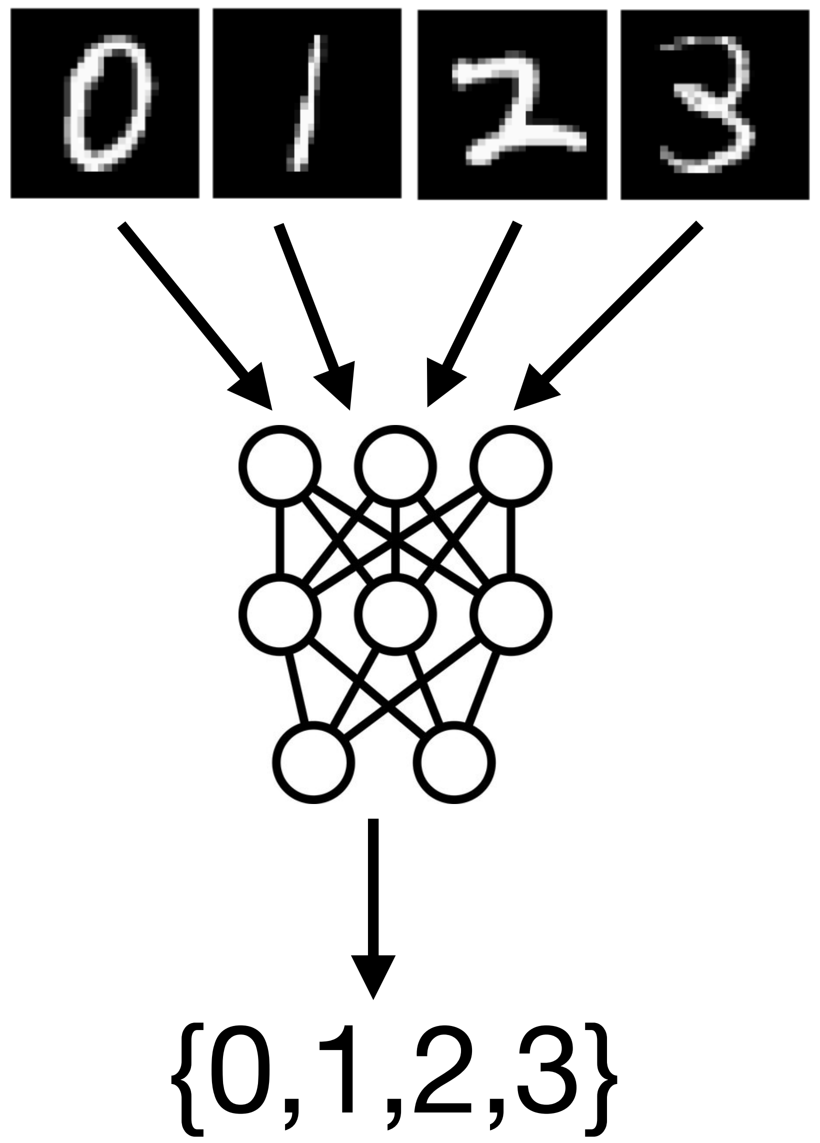
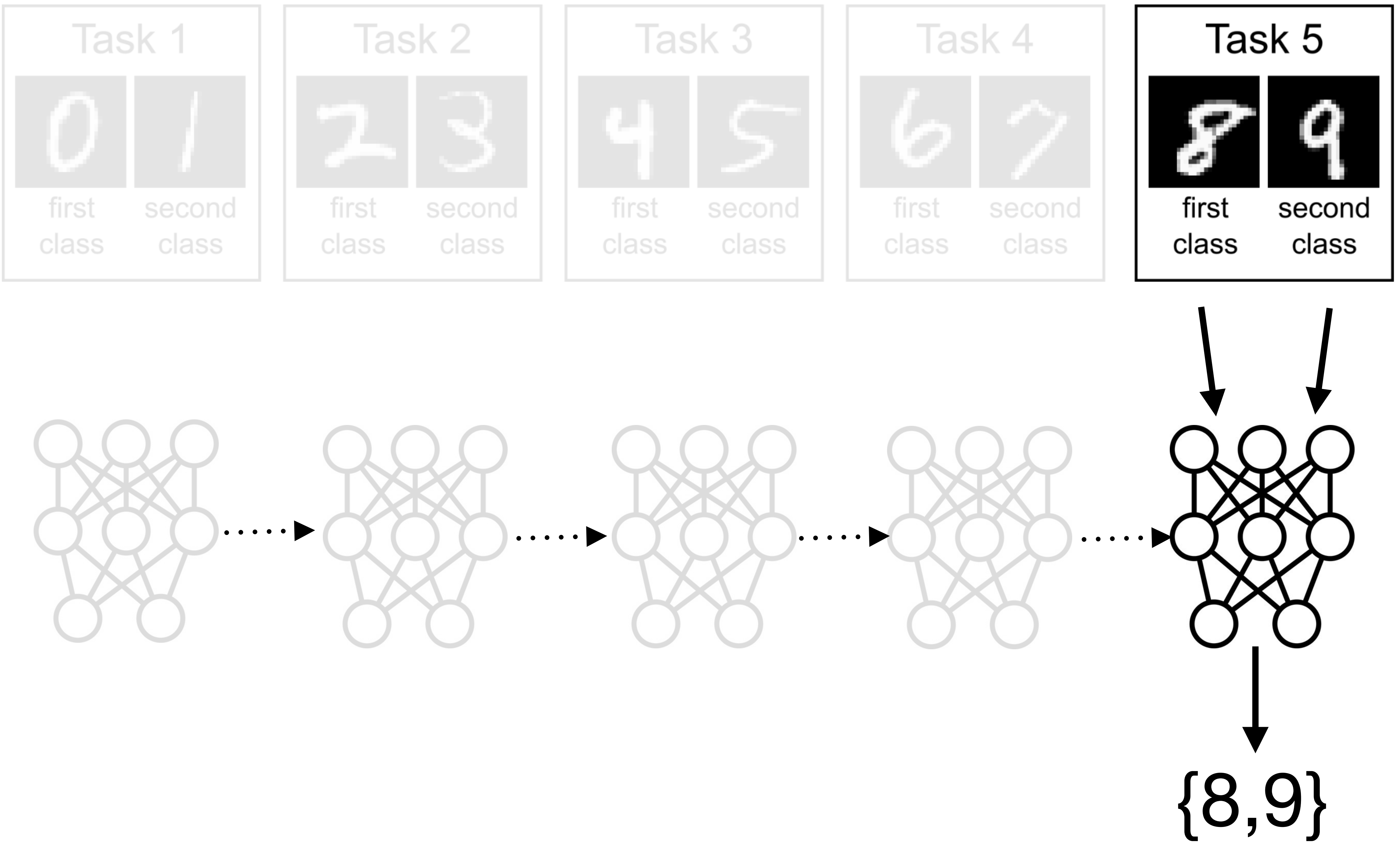


Image credit: [1] van de Ven, G. & Tolias, A. 2019. Three scenarios for continual learning, https://arxiv.org/abs/1904.07734

Academic lifelong learning

Training



Testing

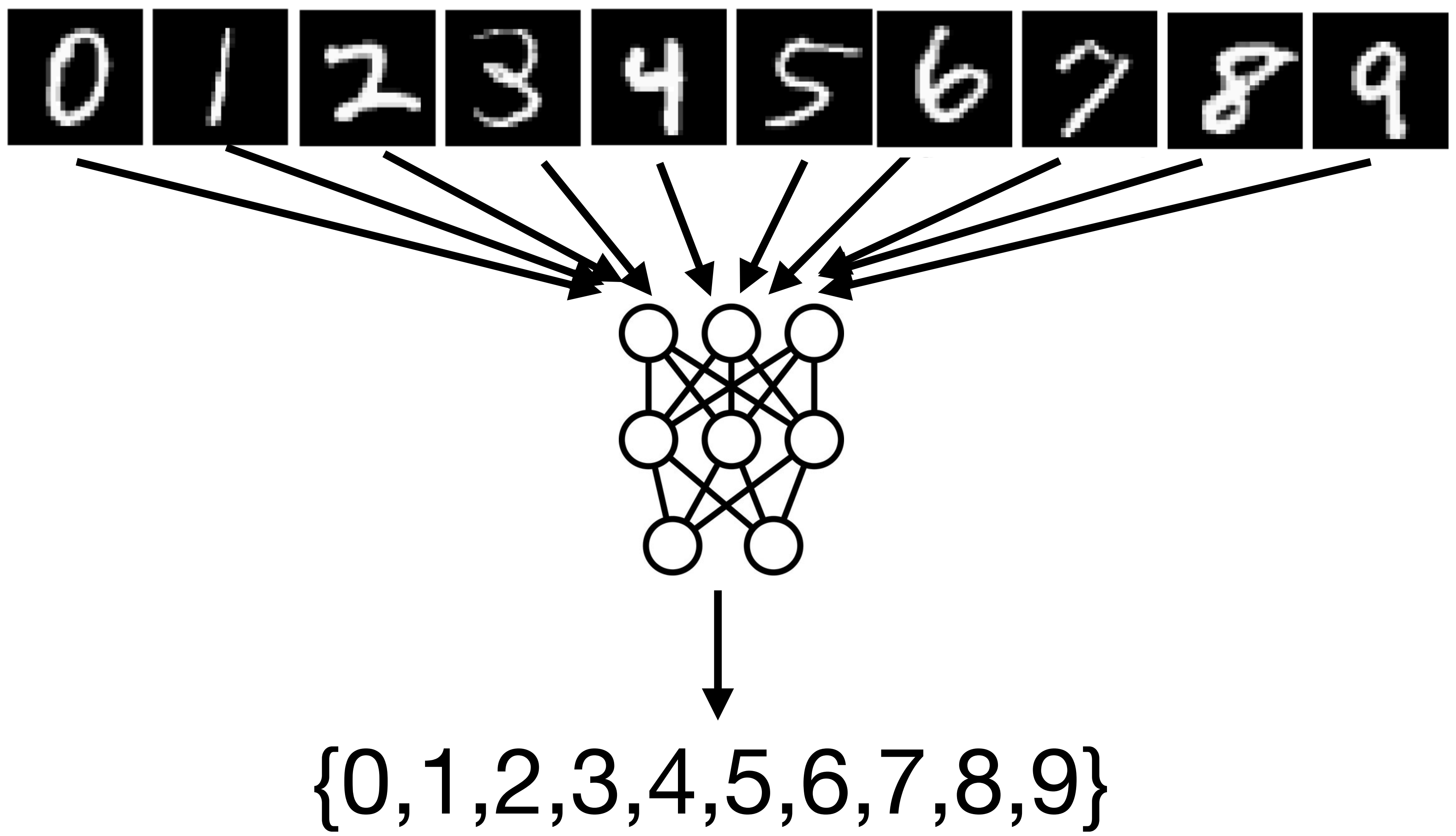


Image credit: [1] van de Ven, G. & Tolias, A. 2019. Three scenarios for continual learning, https://arxiv.org/abs/1904.07734

Academic lifelong learning

Training

Testing

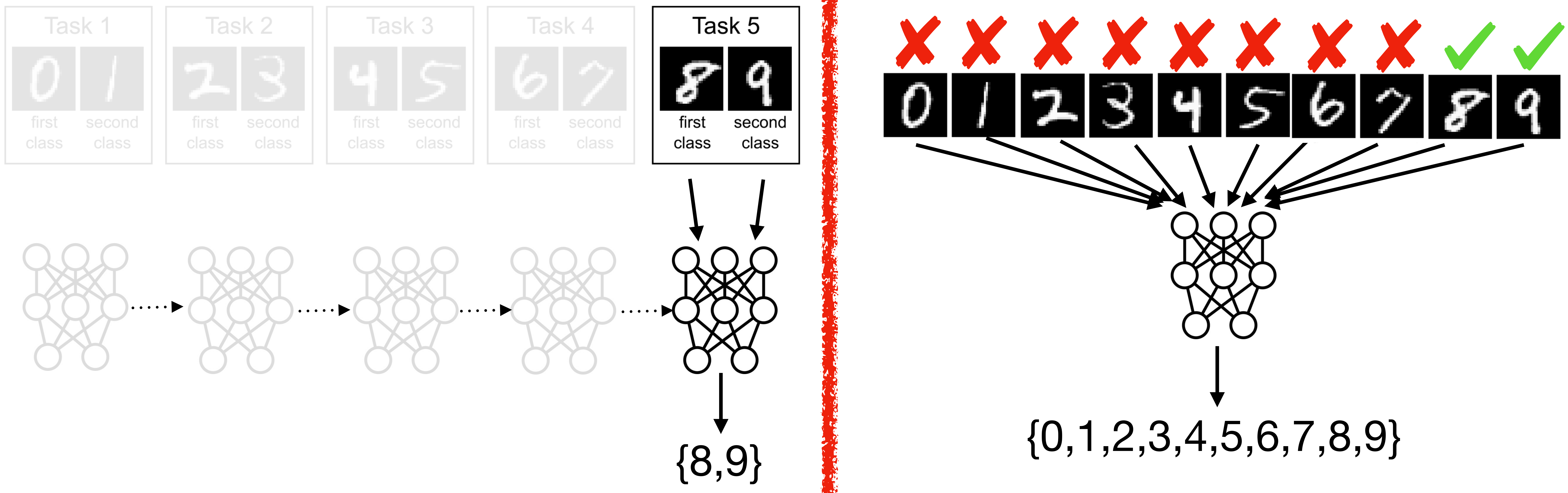
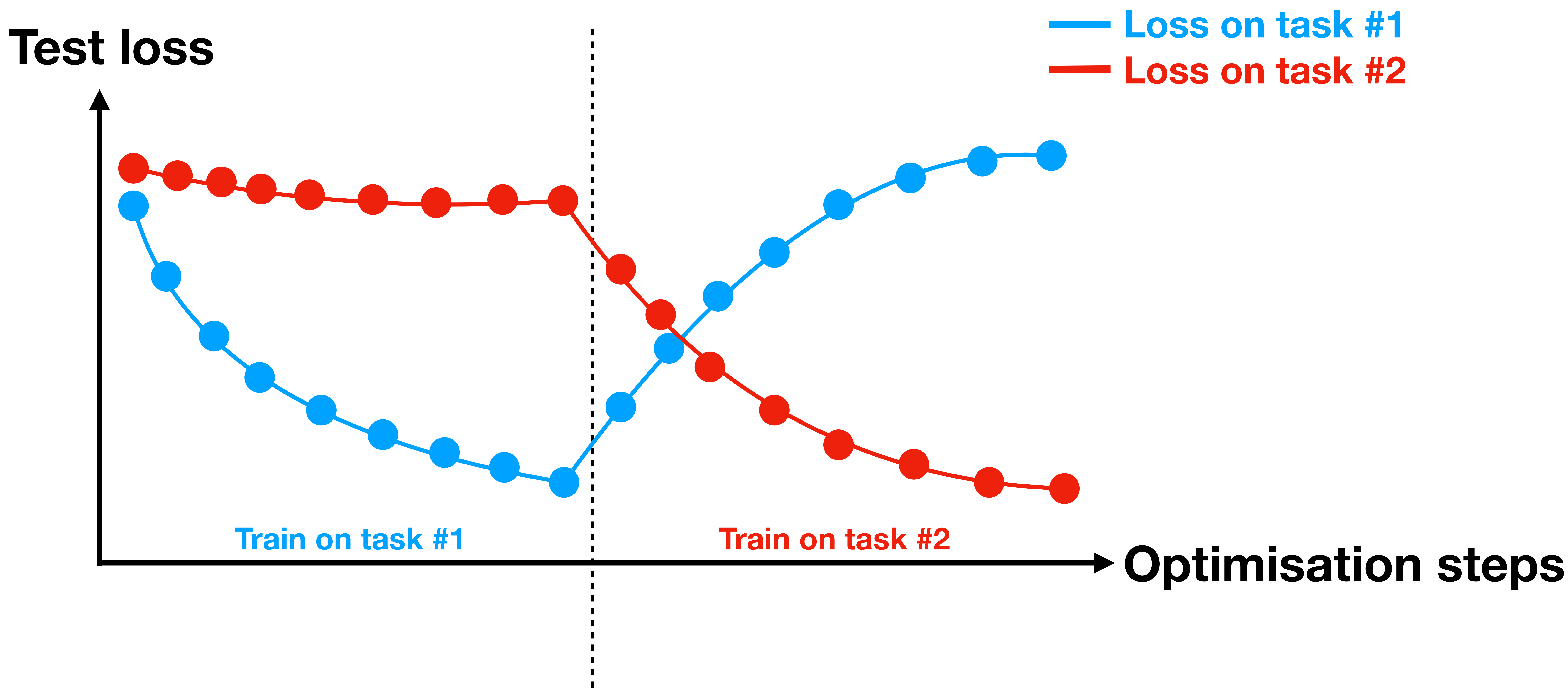


Image credit: [1] van de Ven, G. & Tolias, A. 2019. Three scenarios for continual learning, https://arxiv.org/abs/1904.07734

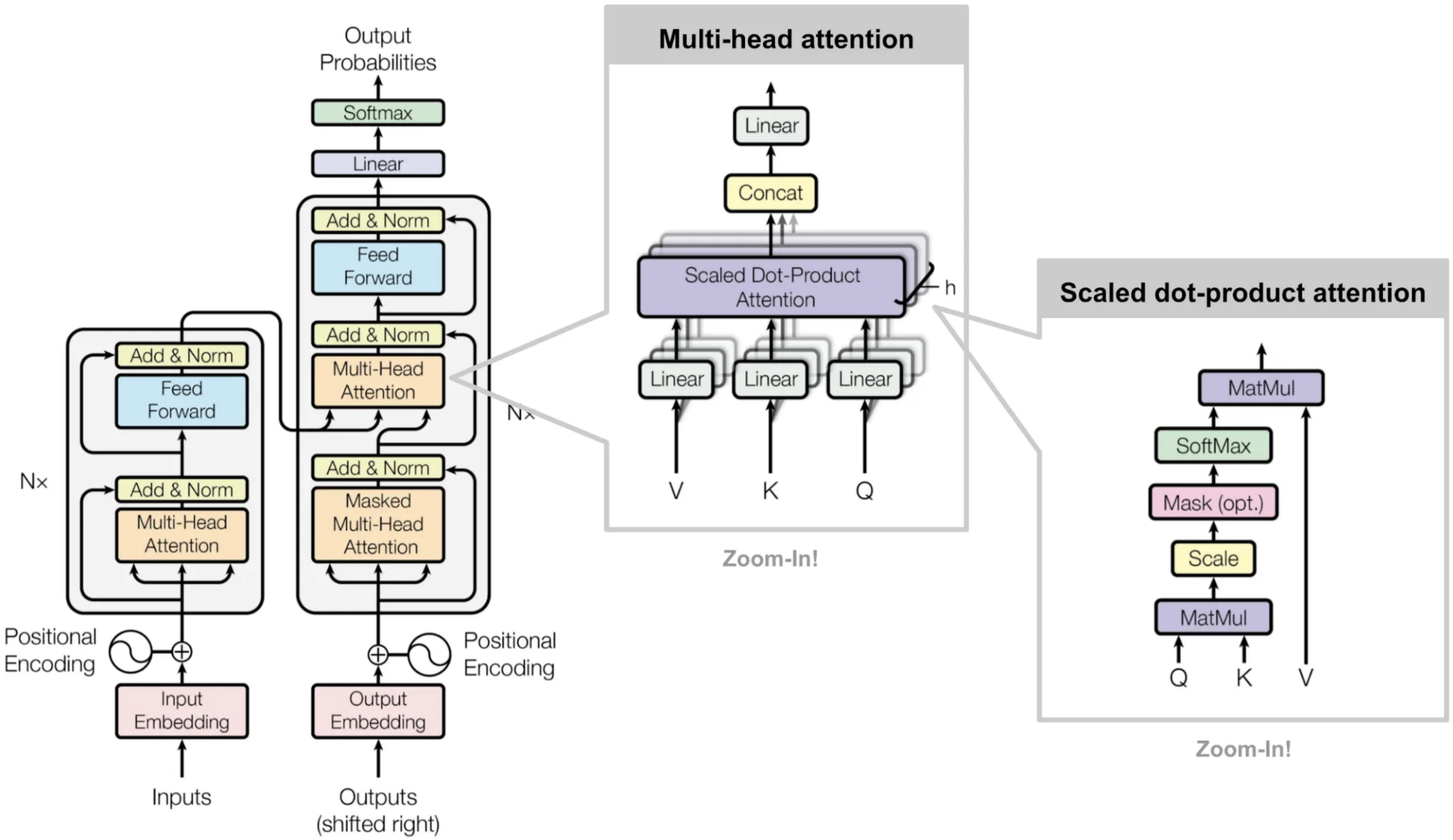
Catastrophic forgetting



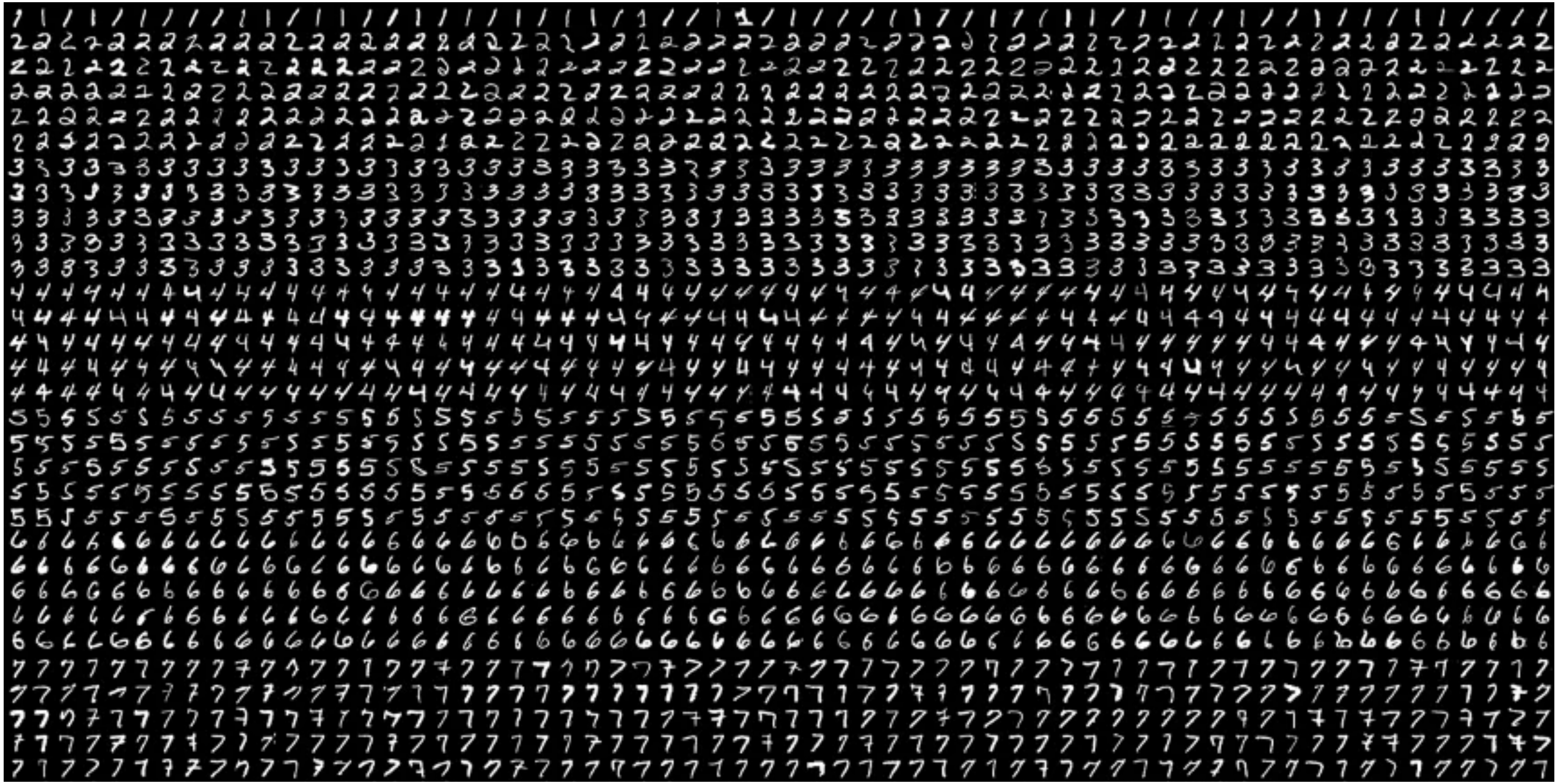
Existing methods?

1. Parameter isolation
2. Regularisation approaches
   I. Functional regularisation
   II. Weight regularisation
3. Replay methods

Issue-1: which weights to adapt?



Issue-2: training complexity?



...

Issue-3: superficial benchmarks

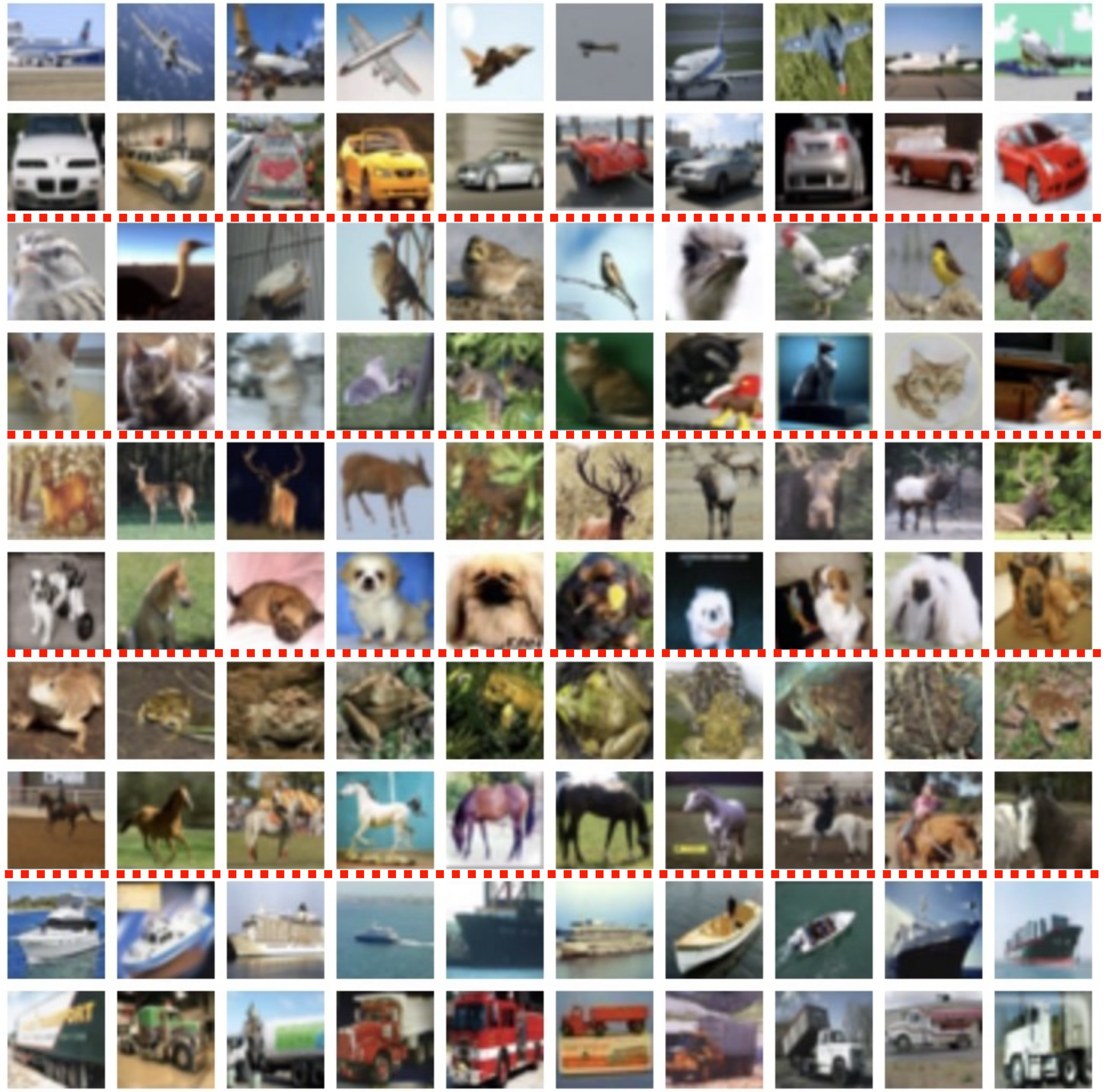
Task #1
airplane
automobile

Task #2
bird
cat

Task #3
deer
dog

Task #4
frog
horse

Task #5
ship
truck



Outline

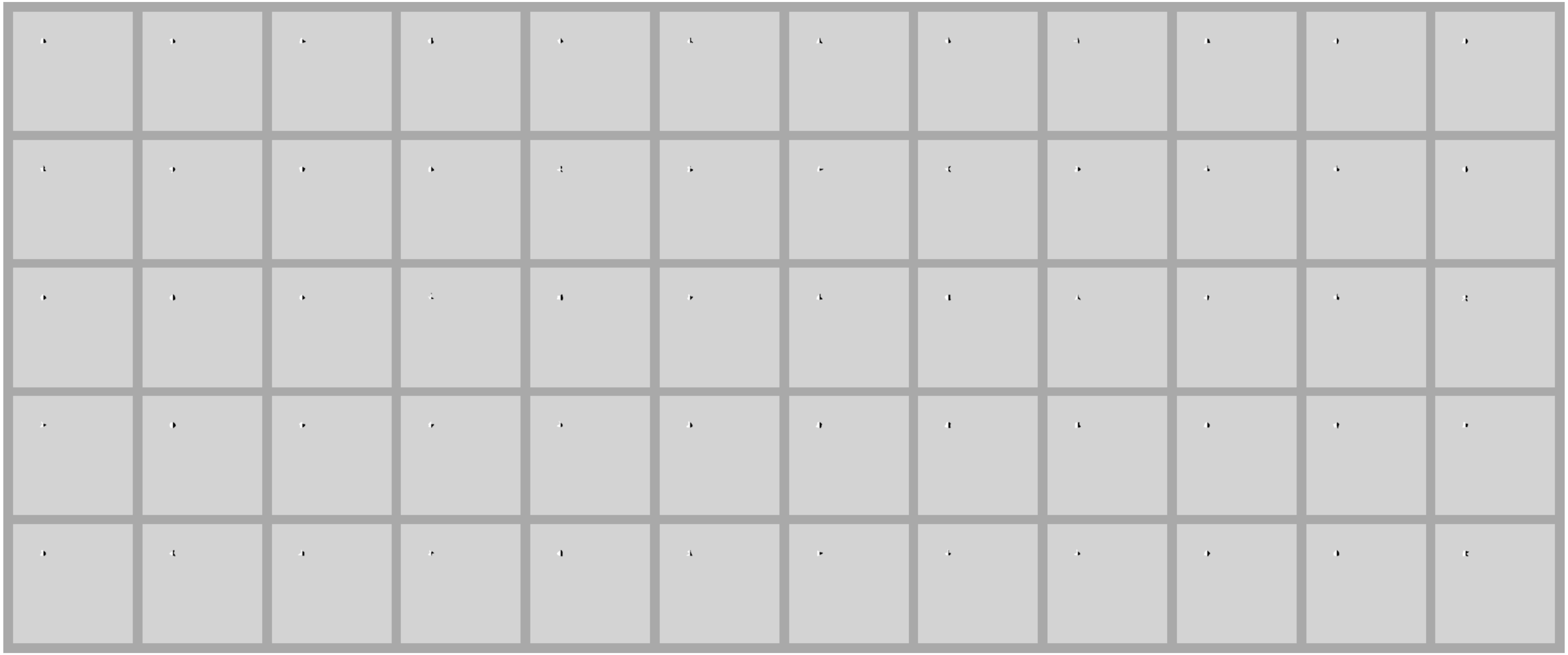
1. Motivation for lifelong machine learning

2. Academic lifelong learning

3. Our works
   - **Disentangled learning**
   - Lifelong knowledge organisation
   - Continual pretraining of LLMs

4. Final thoughts

Our benchmark: infinite dSprites [2]

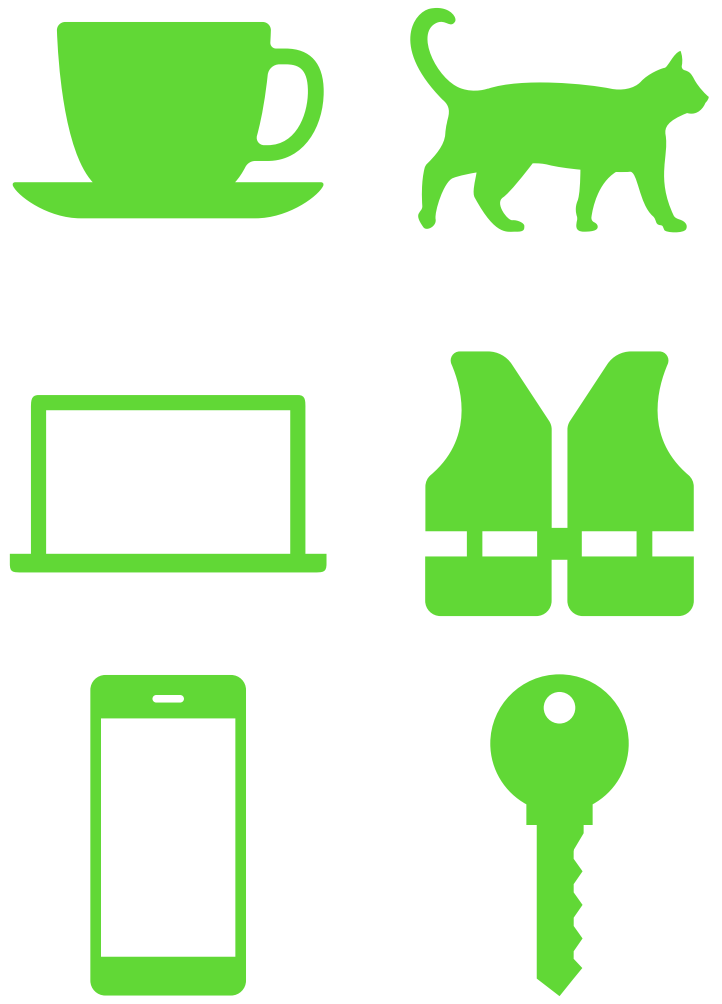


[2] Dziadzio, S., et al. 2024. Infinite dSprites for Disentangled Continual Learning: Separating Memory Edits from Generalization

Disentangled learning [2]

Separate **class/task agnostic transformations** from **what is class/task specific.**

A	cup cat laptop jacket phone key	is still a	cup cat laptop jacket phone key	even if	the image is blurry. the object has a different color. the lighting is low. the camera is angle changes. the image size changes. the image has low resolution.



[2] Dziadzio, S., et al. 2024. Infinite dSprites for Disentangled Continual Learning: Separating Memory Edits from Generalization

Disentangled learning

New class

Prototype buffer

Image credits: https://stackoverflow.com/questions/22844341/feasible-computer-vision-techniques-for-detecting-and-classifying-animal-faces and https://openreview.net/pdf?id=WyEdX2R4er

Disentangled learning

Prototype buffer

Image credits: https://stackoverflow.com/questions/22844341/feasible-computer-vision-techniques-for-detecting-and-classifying-animal-faces and https://openreview.net/pdf?id=WyEdX2R4er

Disentangled learning

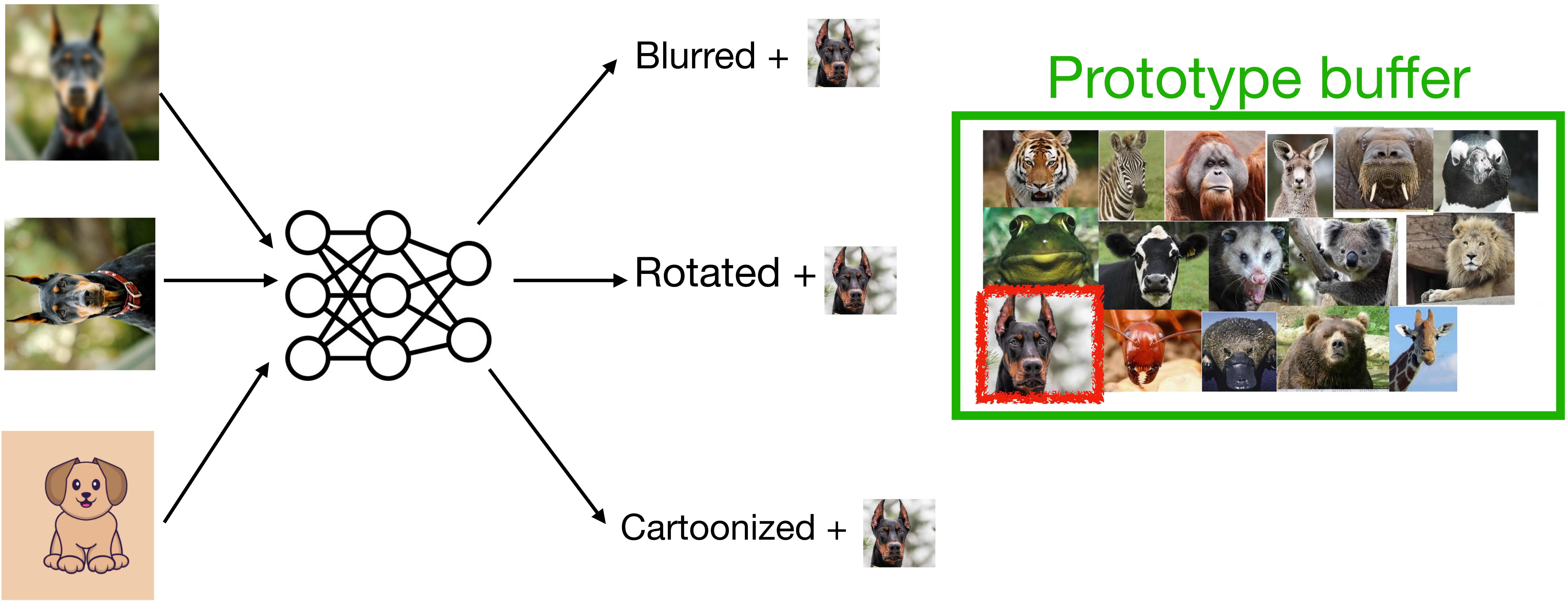


Image credits: https://stackoverflow.com/questions/22844341/feasible-computer-vision-techniques-for-detecting-and-classifying-animal-faces and https://openreview.net/pdf?id=WyEdX2R4er

Disentangled learning

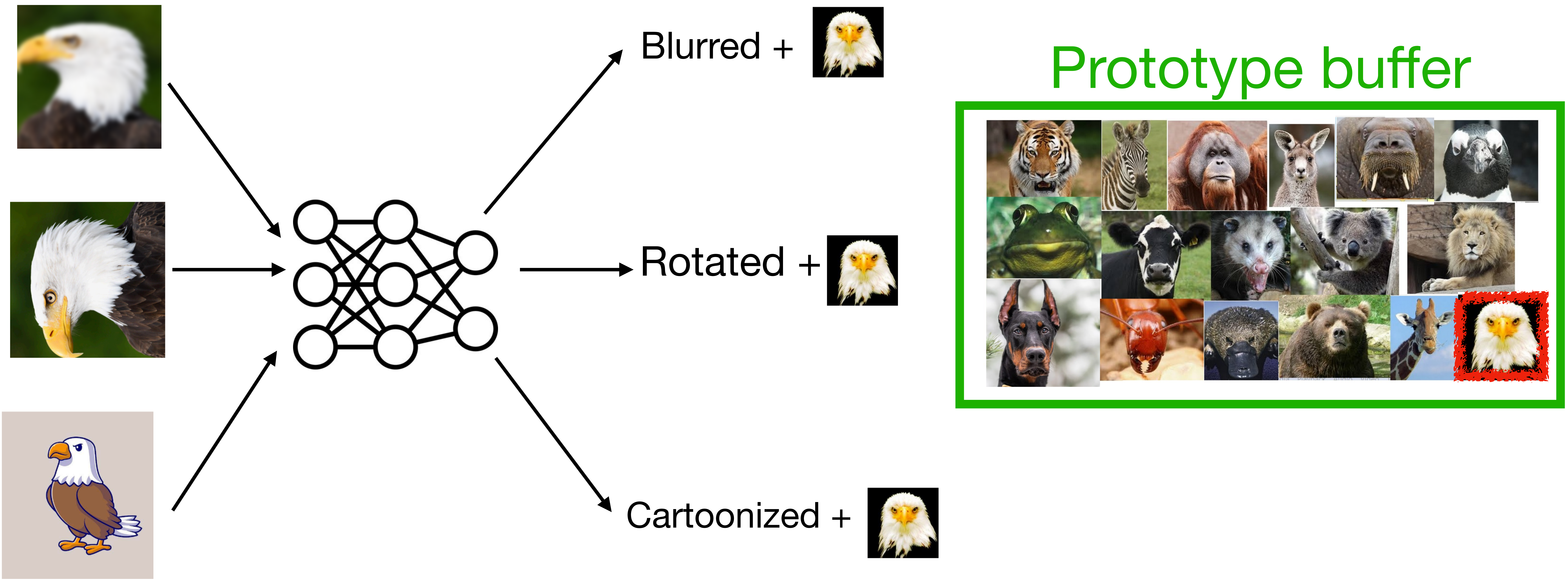
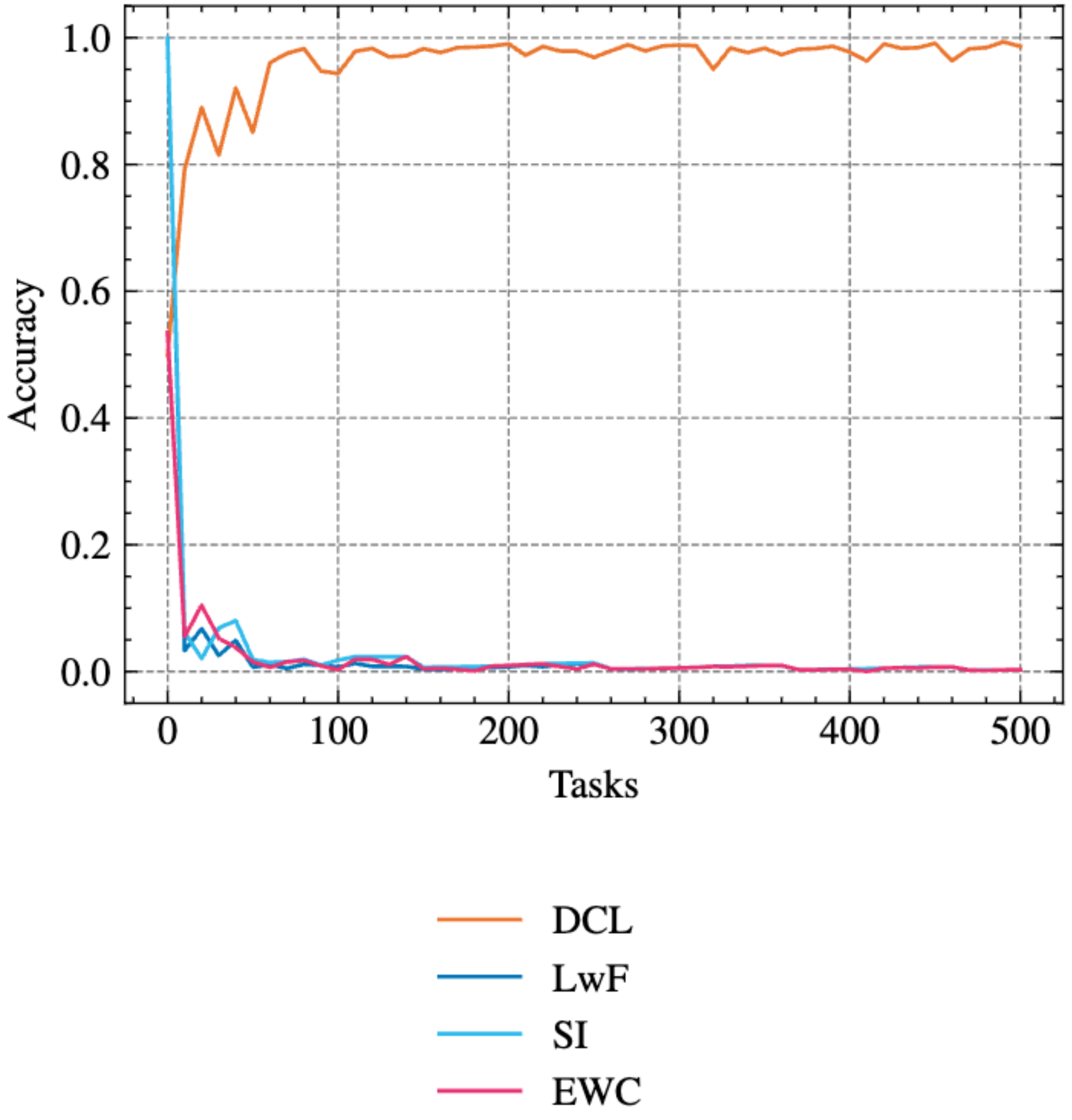


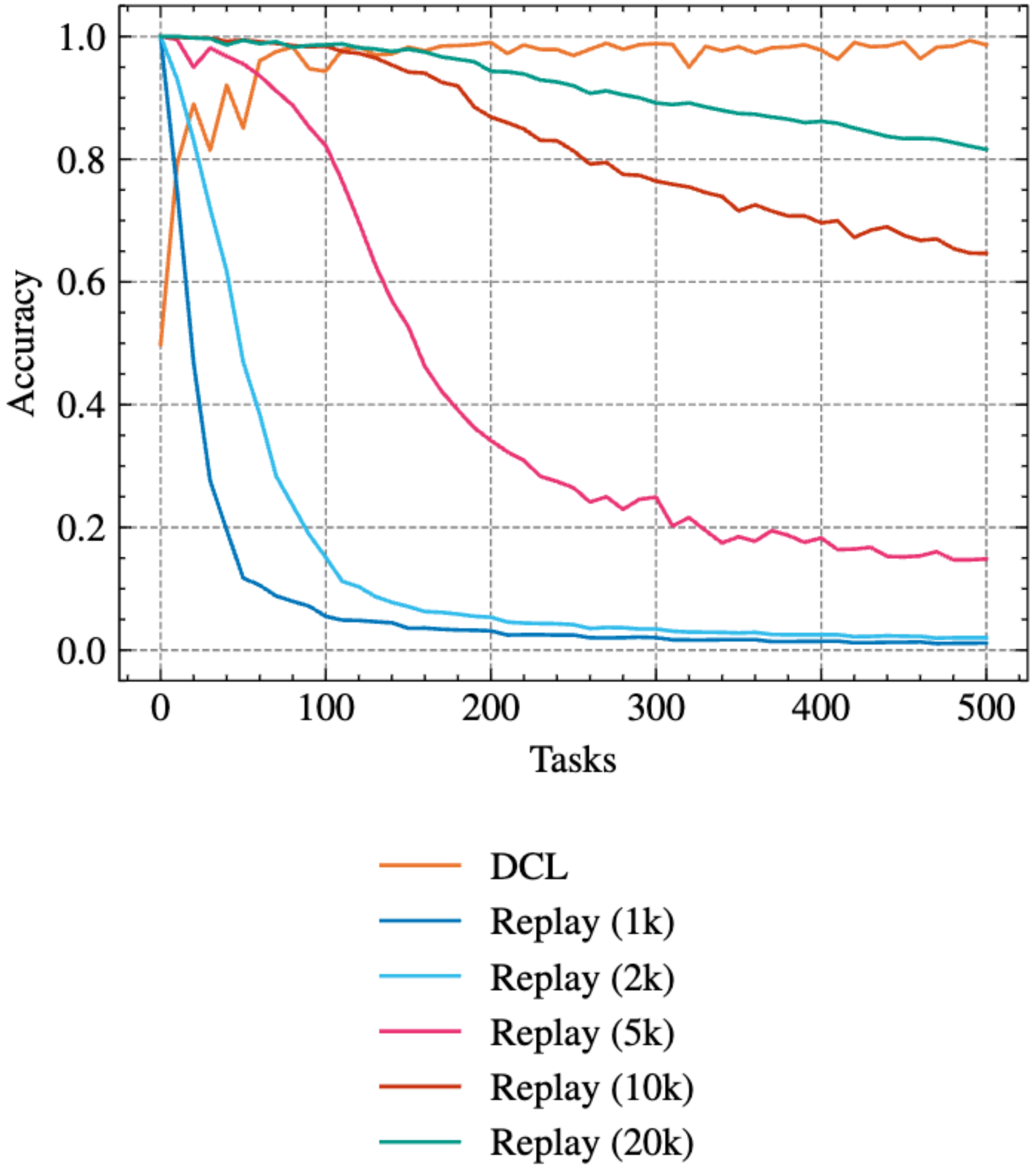
Image credits: https://stackoverflow.com/questions/22844341/feasible-computer-vision-techniques-for-detecting-and-classifying-animal-faces and https://openreview.net/pdf?id=WyEdX2R4er

Disentangled learning

Our method vs standard baselines



Our method vs golden baseline



Latent disentangled learning*

- Toys-200 dataset
- Latent prototypes






* Strüber, Joschka. in preparation for MSc thesis.

Outline

1. Motivation for lifelong machine learning

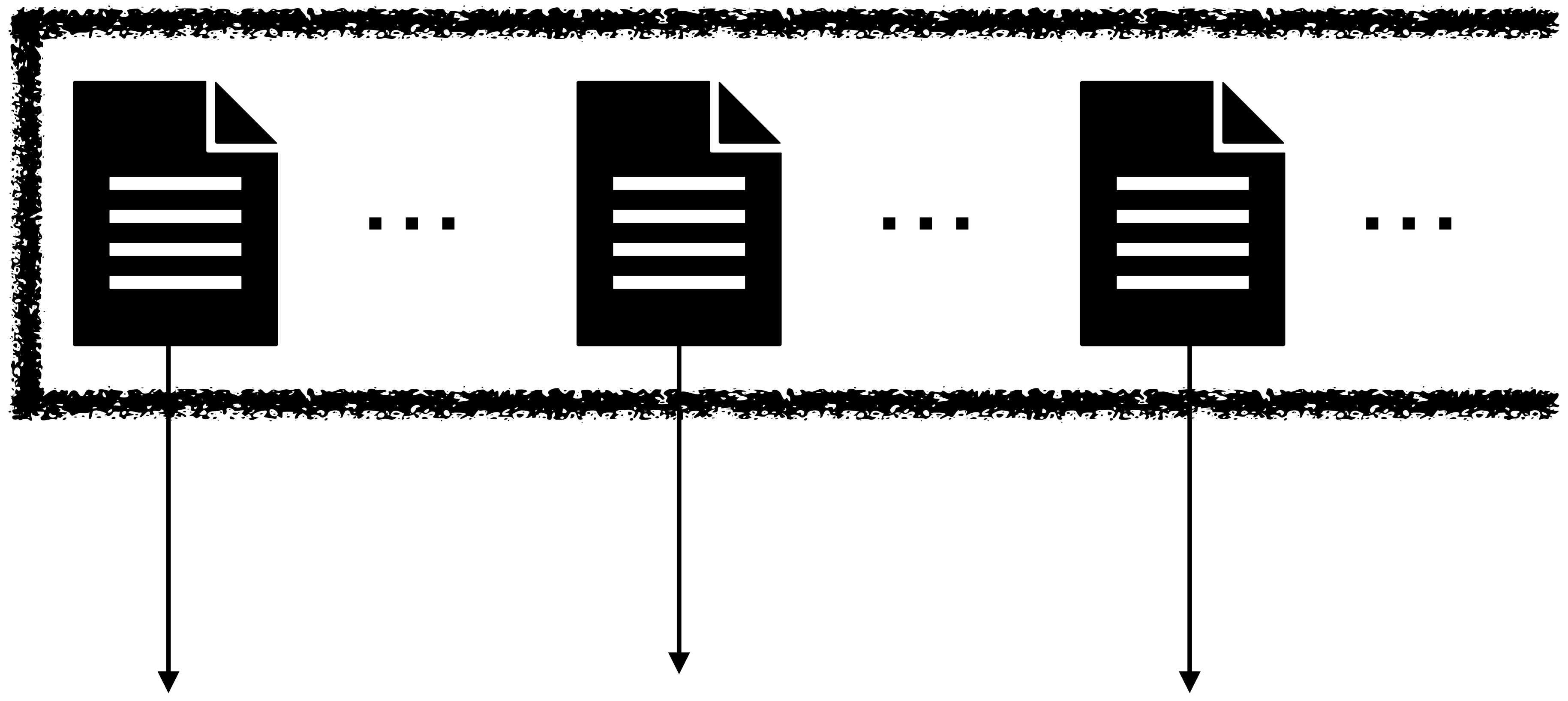
2. Academic lifelong learning

3. **Our works**
    - Disentangled learning
    - **Lifelong knowledge organisation**
    - Continual pretraining of LLMs

4. Final thoughts

Lifelong knowledge organization

Incoming text documents



… In a press meeting, US President Obama said …

… recently elected President Mr. Trump appointed …

… Joe Biden took over the office from Donald …

Question bank

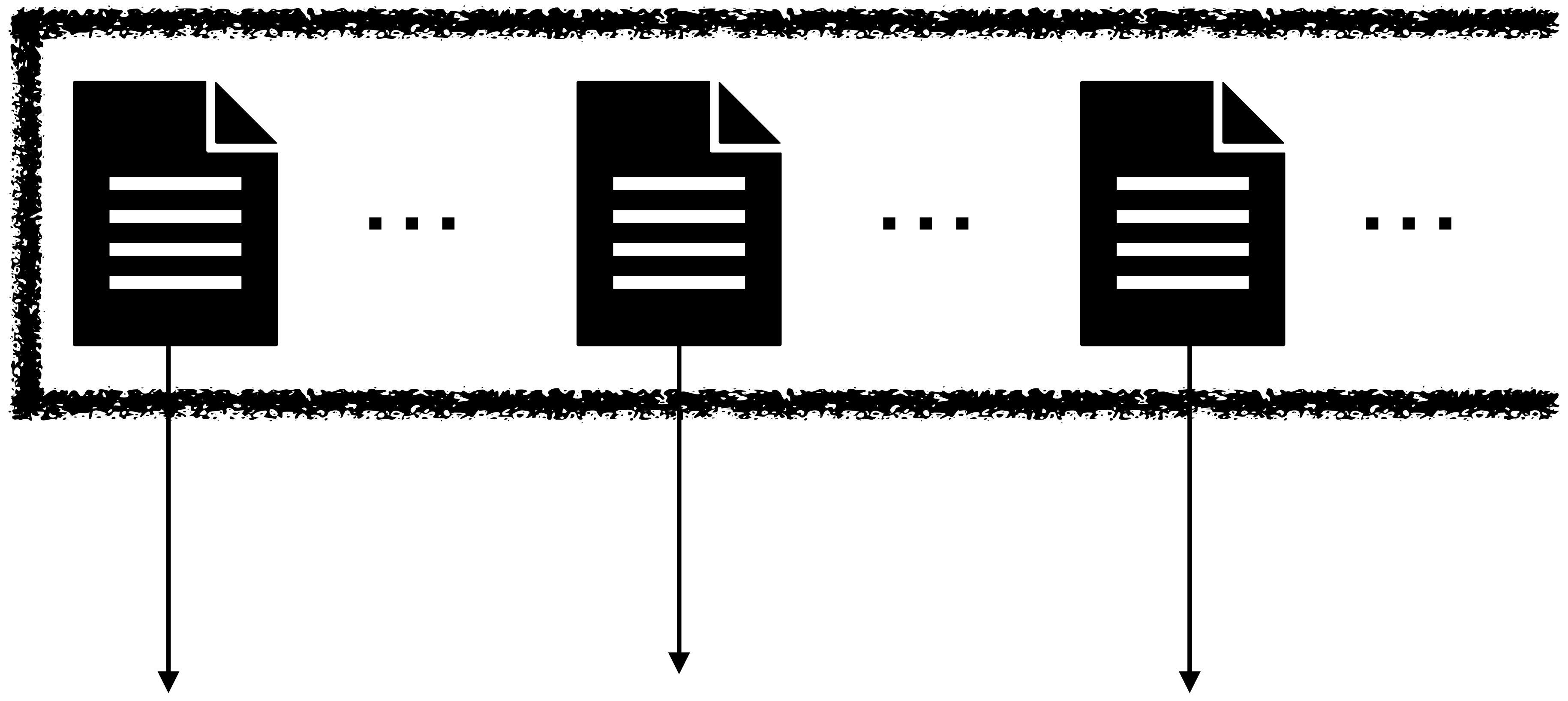
Who is the President of the USA?

…

…

Lifelong knowledge organization

Incoming text documents

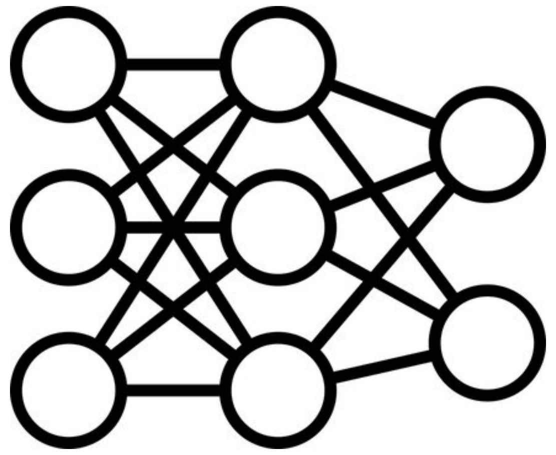


… In a press meeting, US President Obama said …

… recently elected President Mr. Trump appointed …

… Joe Biden took over the office from Donald …

Language model



Question bank

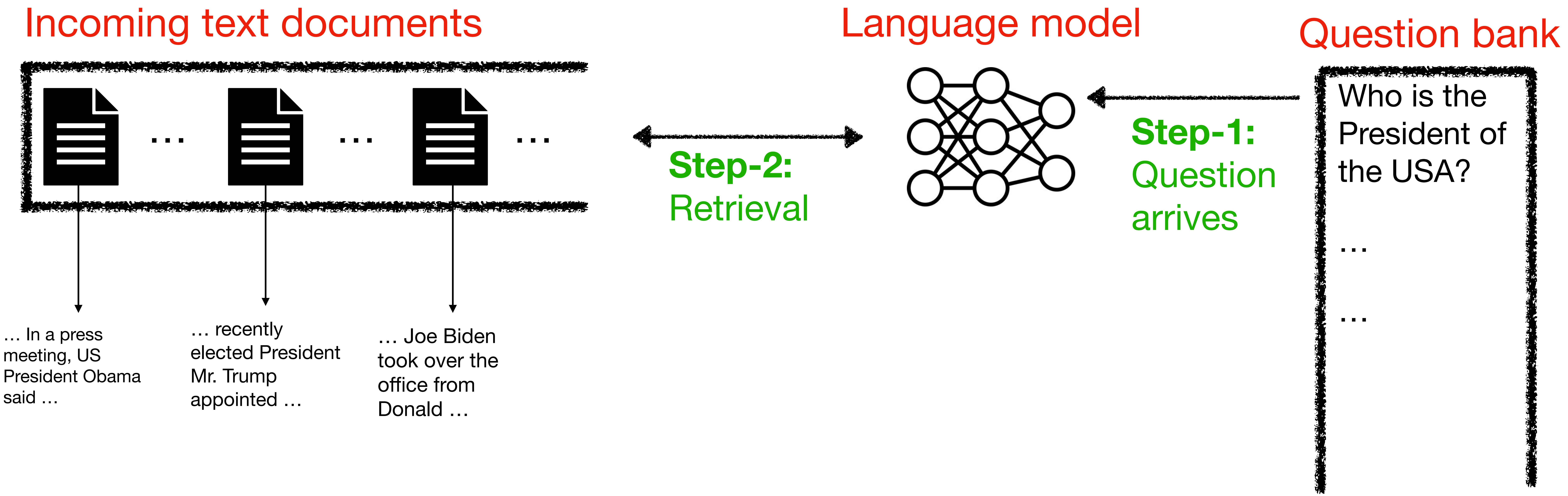
Who is the President of the USA?

…

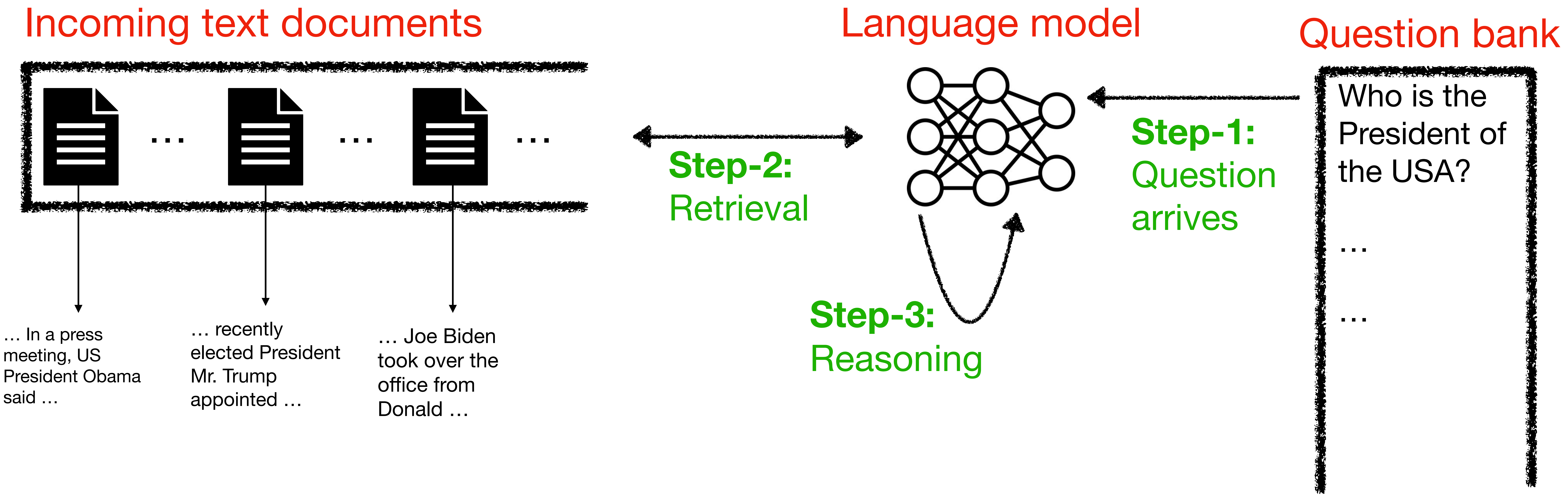
…

Lifelong knowledge organization

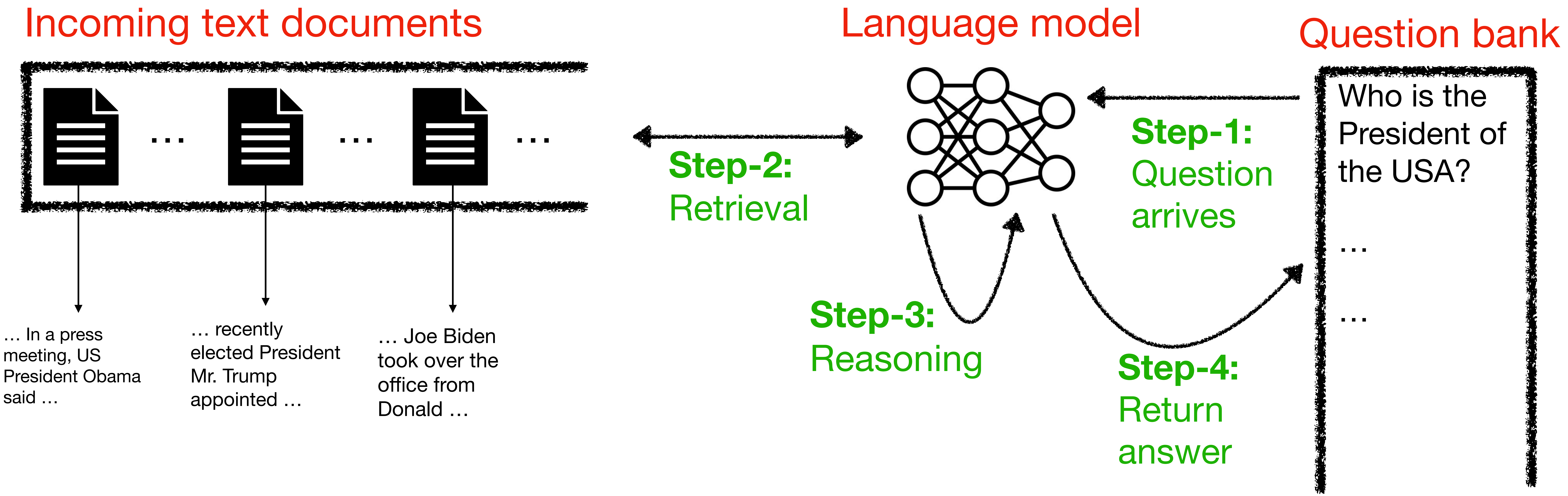
Lifelong knowledge organization



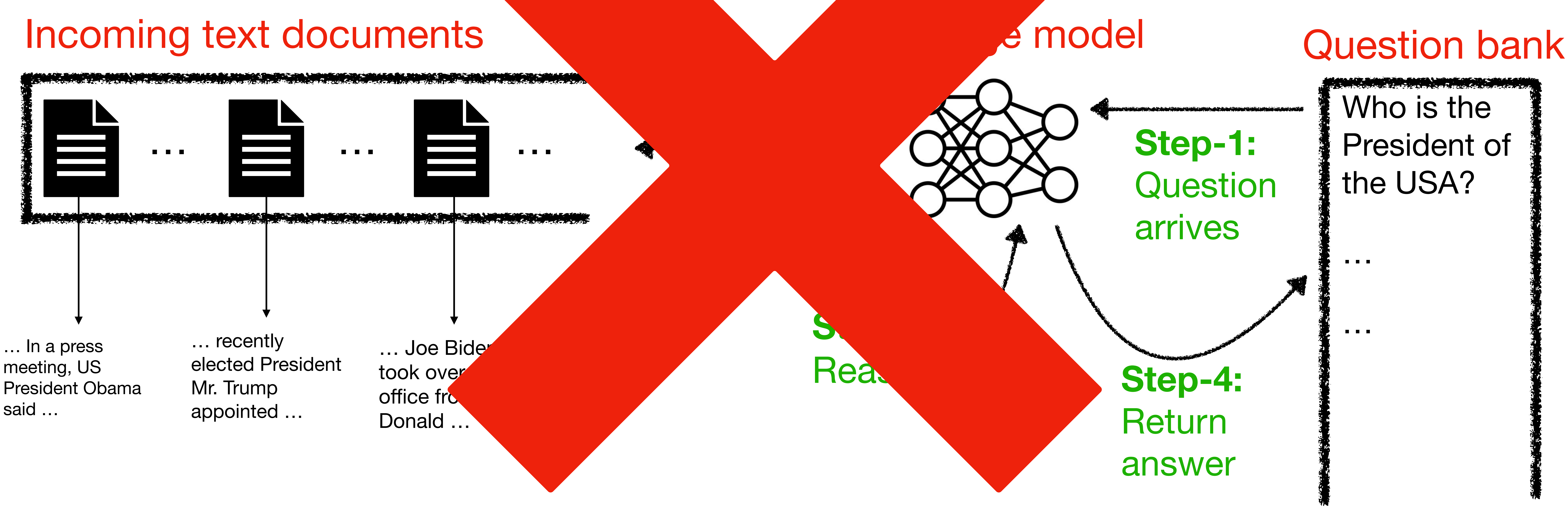
Lifelong knowledge organization



Lifelong knowledge organization

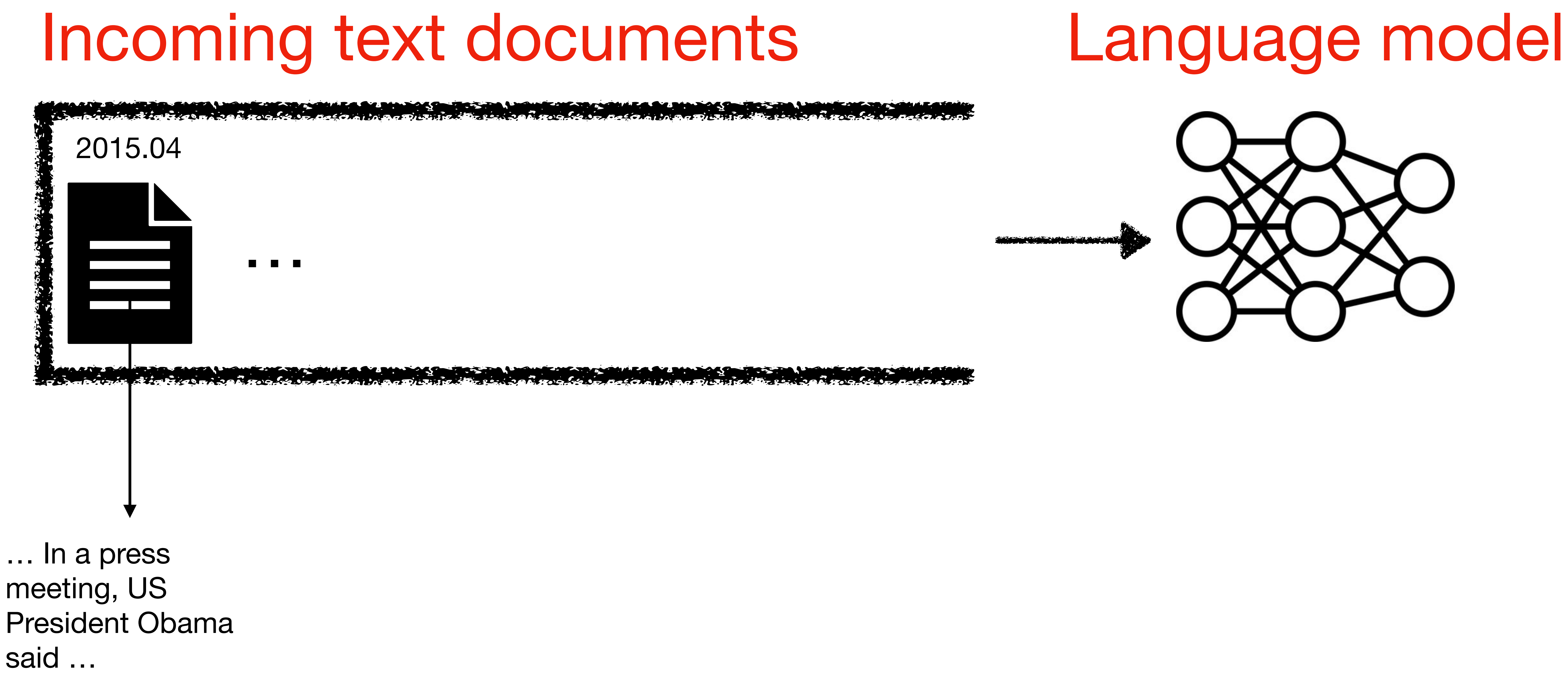


Lifelong k wledge ganization



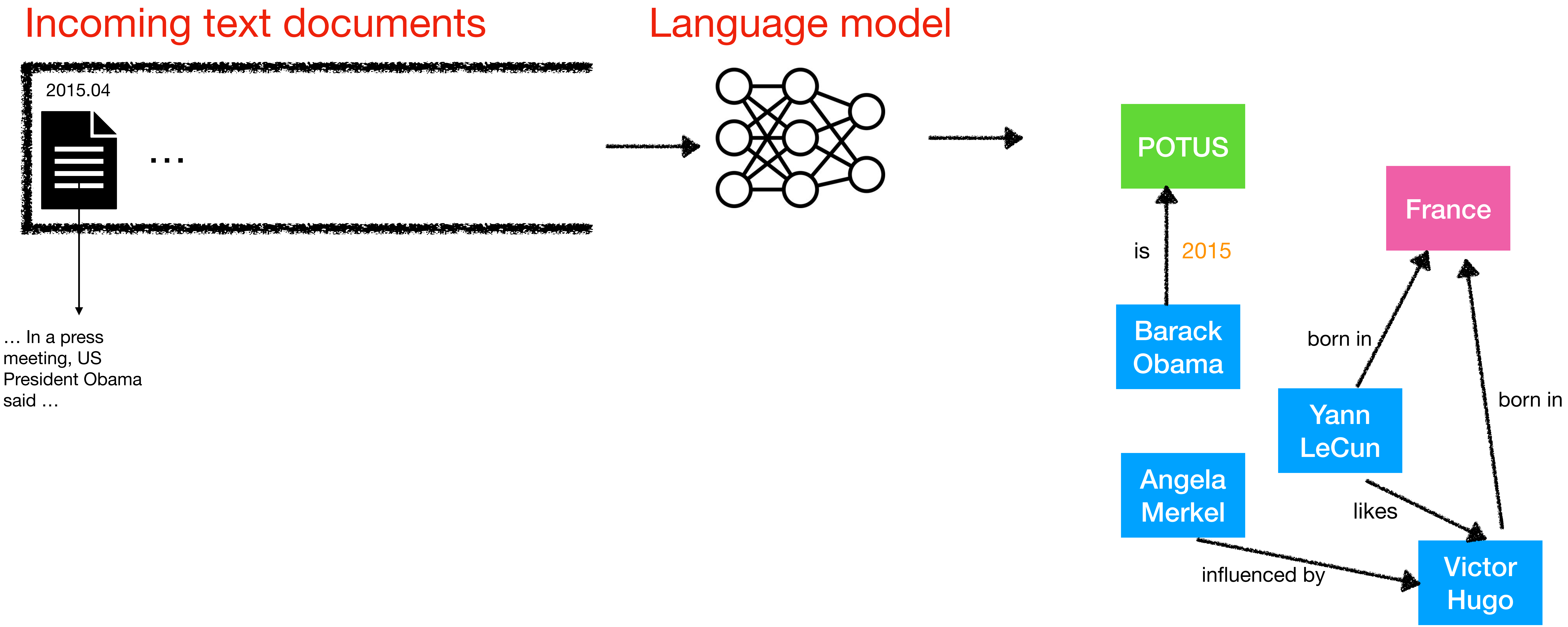
1. If the evidence texts conflict
2. If the evidence conflicts with parametric memory

Lifelong knowledge organization*

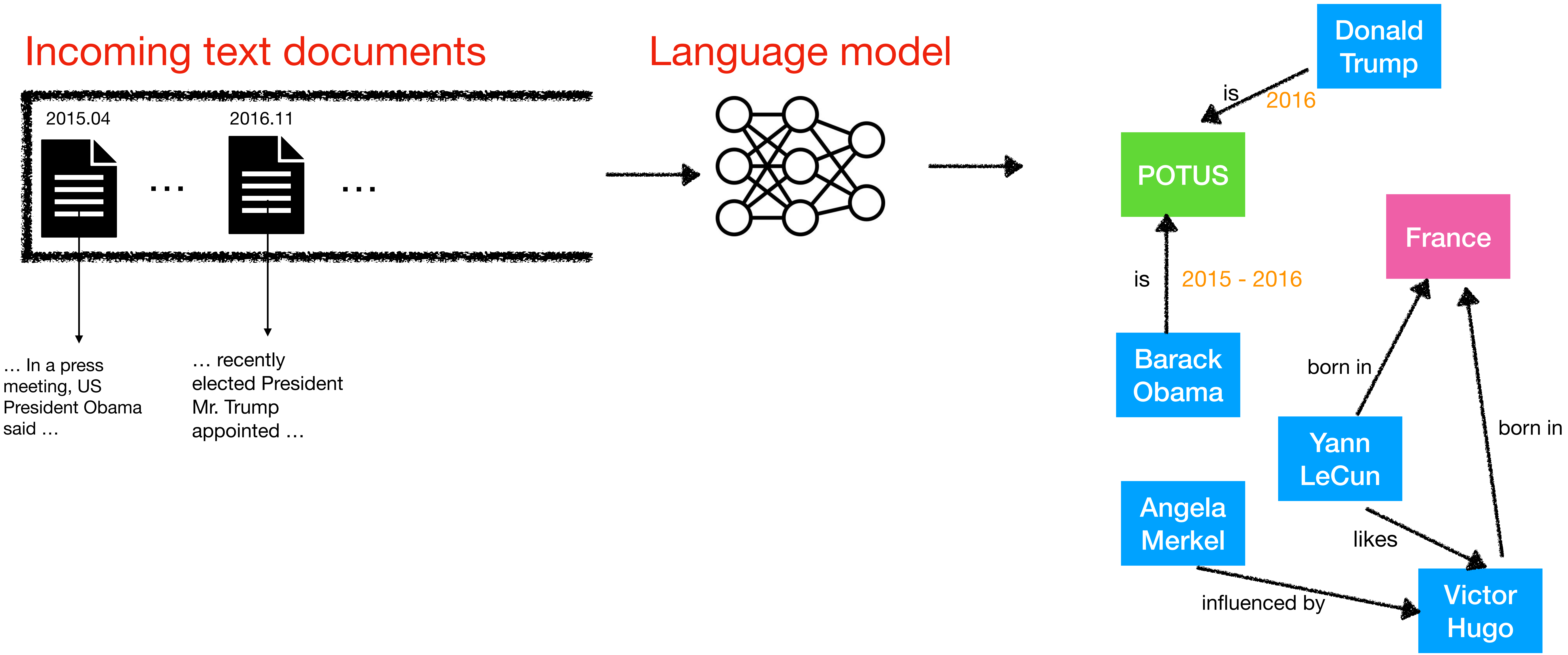


* Özer, Atahan. In preparation for MSc thesis.

Lifelong knowledge organization



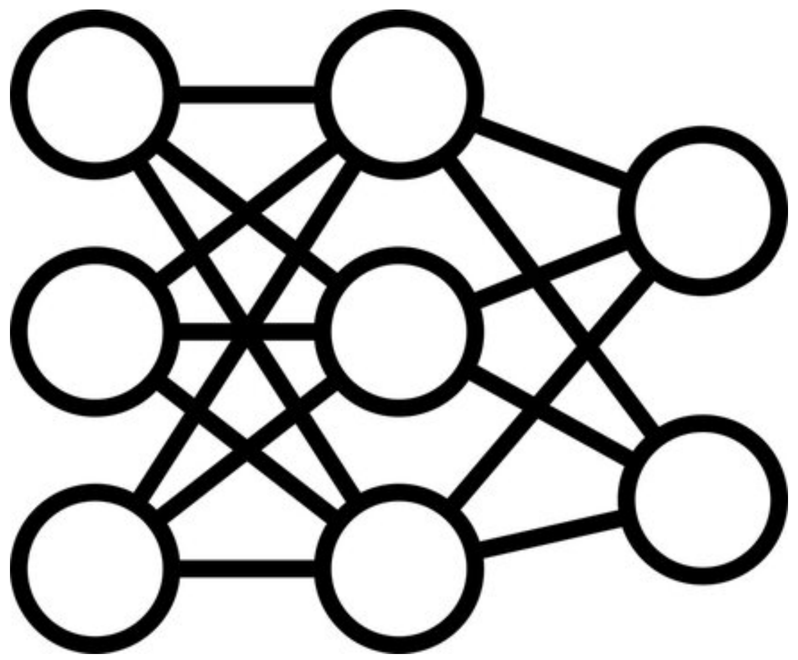
Lifelong knowledge organization



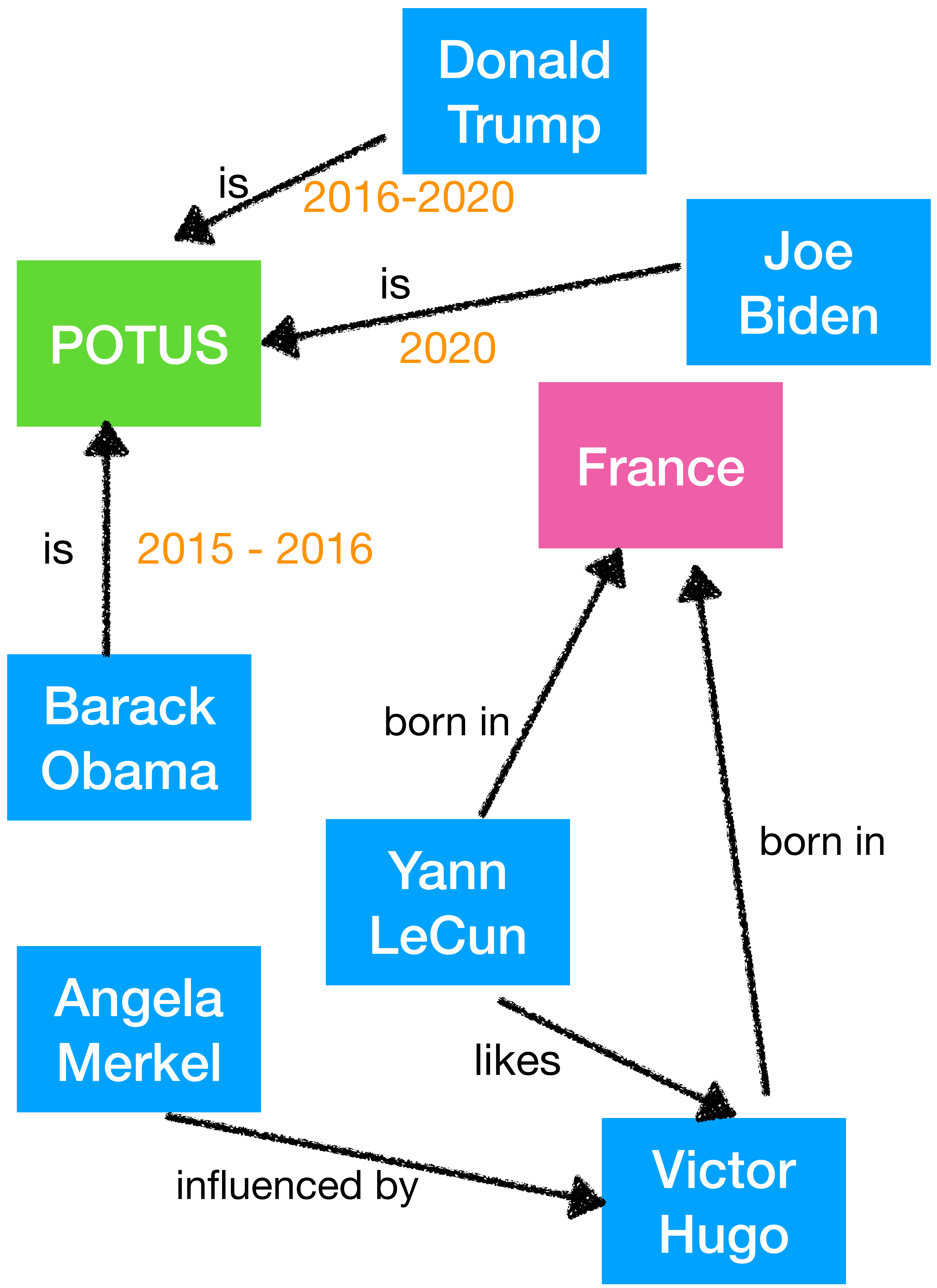
Lifelong knowledge organization

Disentangled lifelong knowledge organization

Reasoning engine



Knowledge base



Back to Prolog (but improved)?

Reasoning engine

```
mother(X,Y):-
   parent(X,Y),female(X).

father(X,Y):-
   parent(X,Y),male(X).

haschild(X):-
   parent(X,_).
```

Knowledge base

```
female(pammi).
female(lizza).
female(patty).
female(anny).
male(jimmy).
male(bobby).
male(tomy).
male(pitter).
parent(pammi,bobby).
parent(tomy,bobby).
parent(tomy,lizza).
parent(bobby,anny).
...
```

Example from: https://www.educba.com/prolog-family-tree/

Outline

1. Motivation for lifelong machine learning

2. Academic lifelong learning

3. Our works
    - Disentangled learning
    - Lifelong knowledge organisation
    - **Continual pretraining of LLMs**

4. Final thoughts

Pretraining & fine-tuning

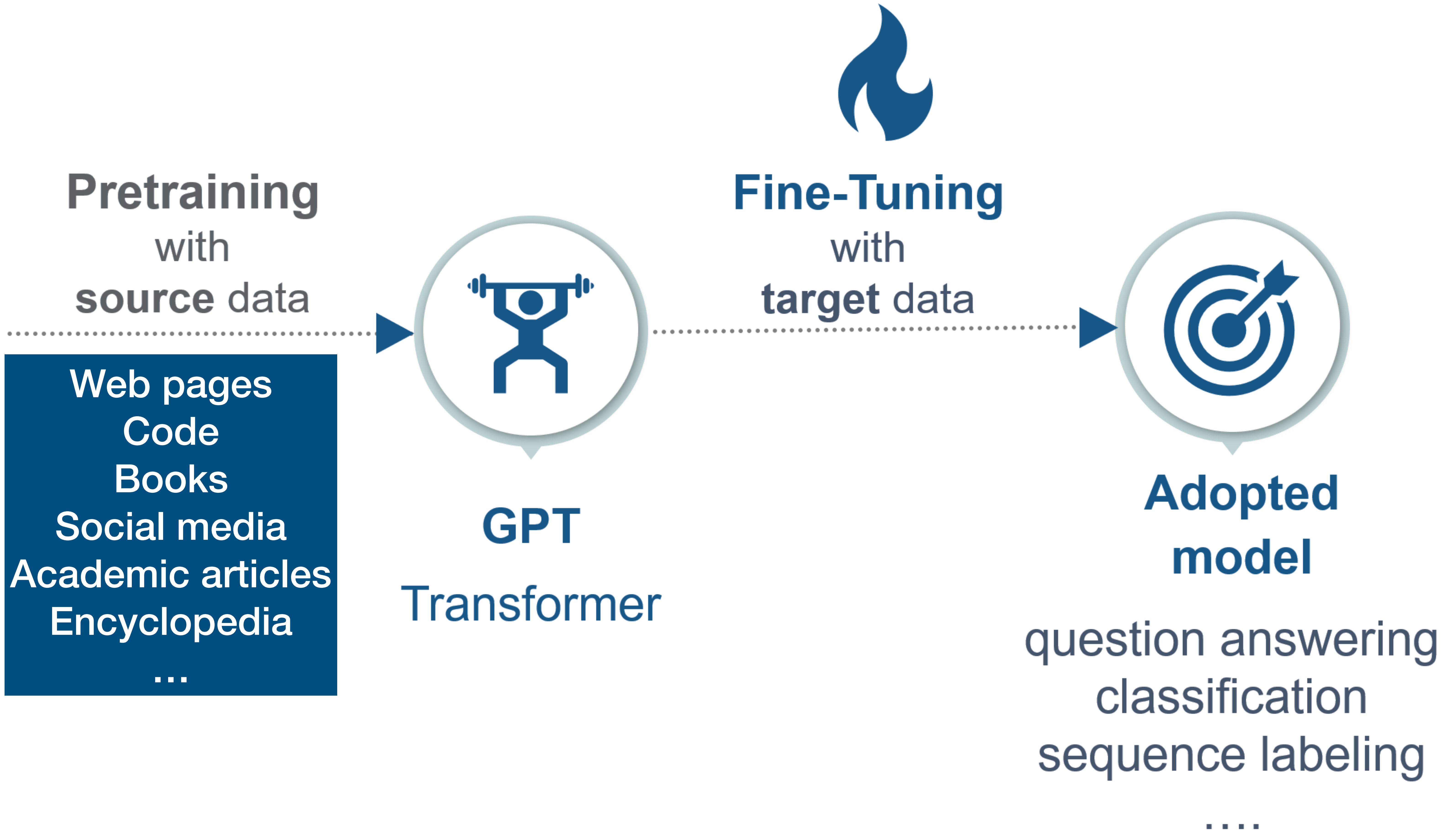
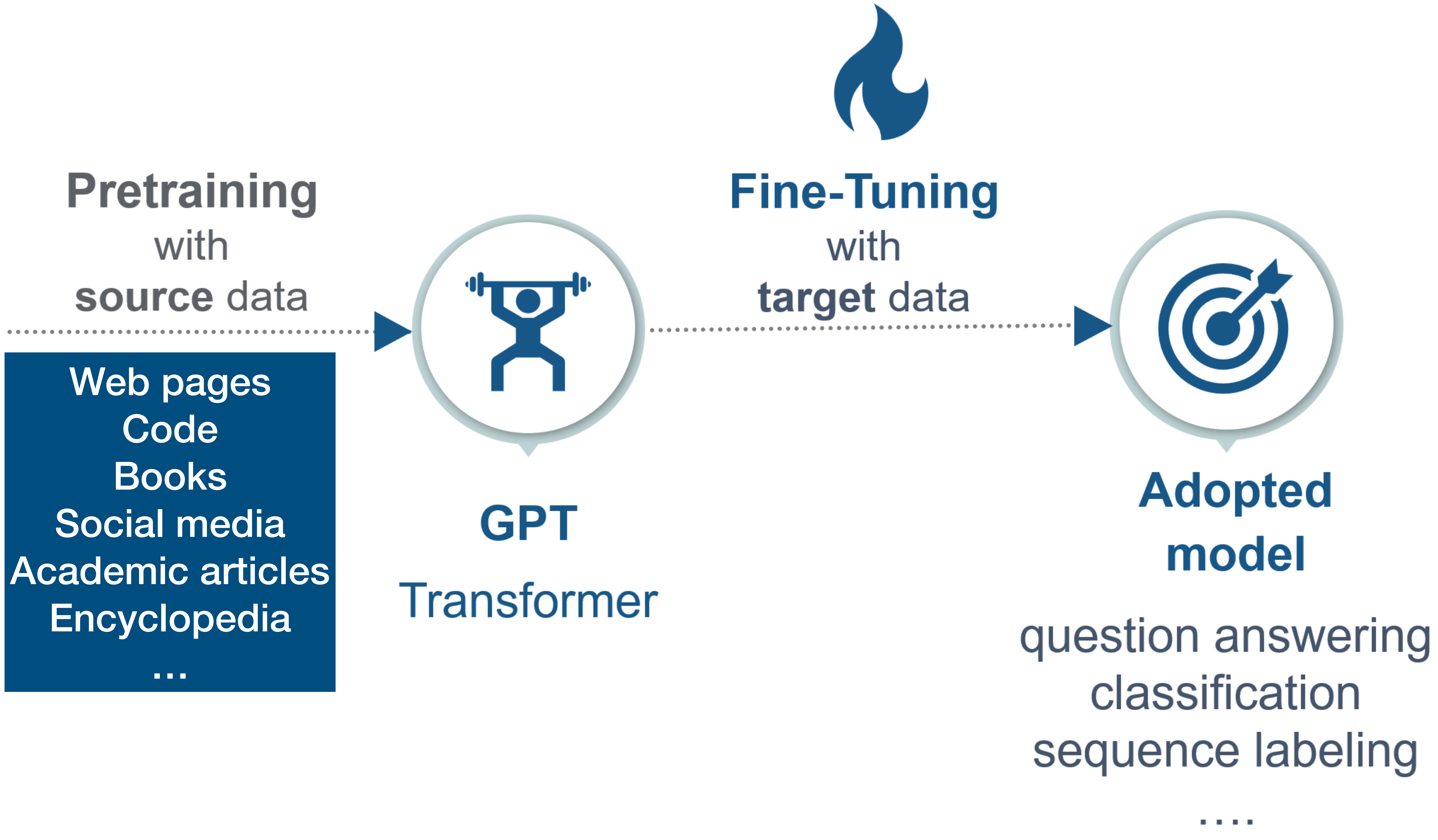


Image credit: https://slds-lmu.github.io/seminar_nlp_ss20/transfer-learning-for-nlp-i.html

Pretraining & fine-tuning



New data available on the Internet everyday!

Image credit: https://slds-lmu.github.io/seminar_nlp_ss20/transfer-learning-for-nlp-i.html

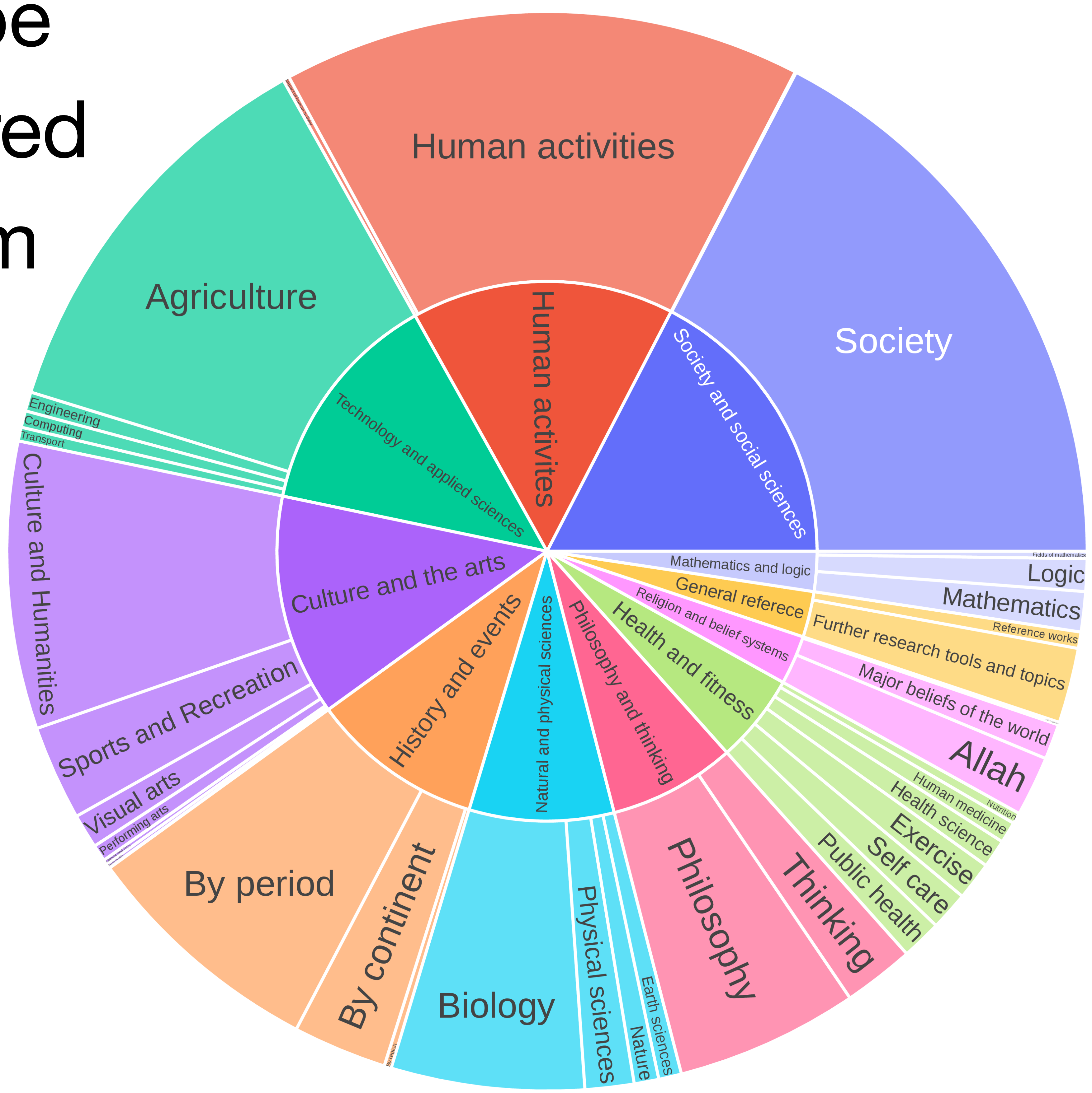
Continual pretraining of LLMs [3]

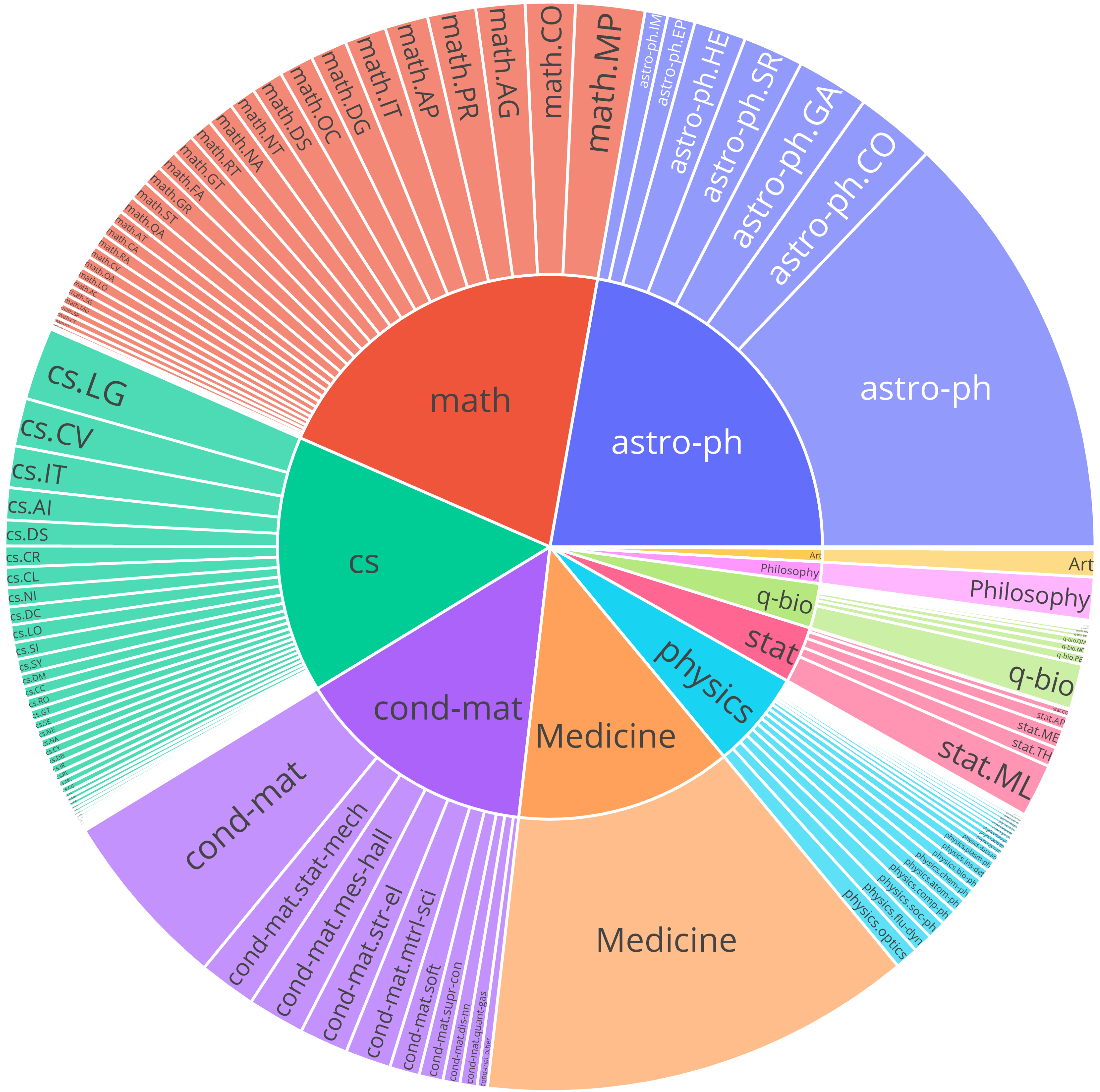
Keep training on new data!

Train domains could be

(i) Semantically ordered

(ii) Completely random





[3] [submitted] Yildiz, Çağatay, et al. 2024. Investigating Continual Pretraining in Large Language Models: Insights and Implications. Image credit: https://arxiv.org/pdf/2210.07370

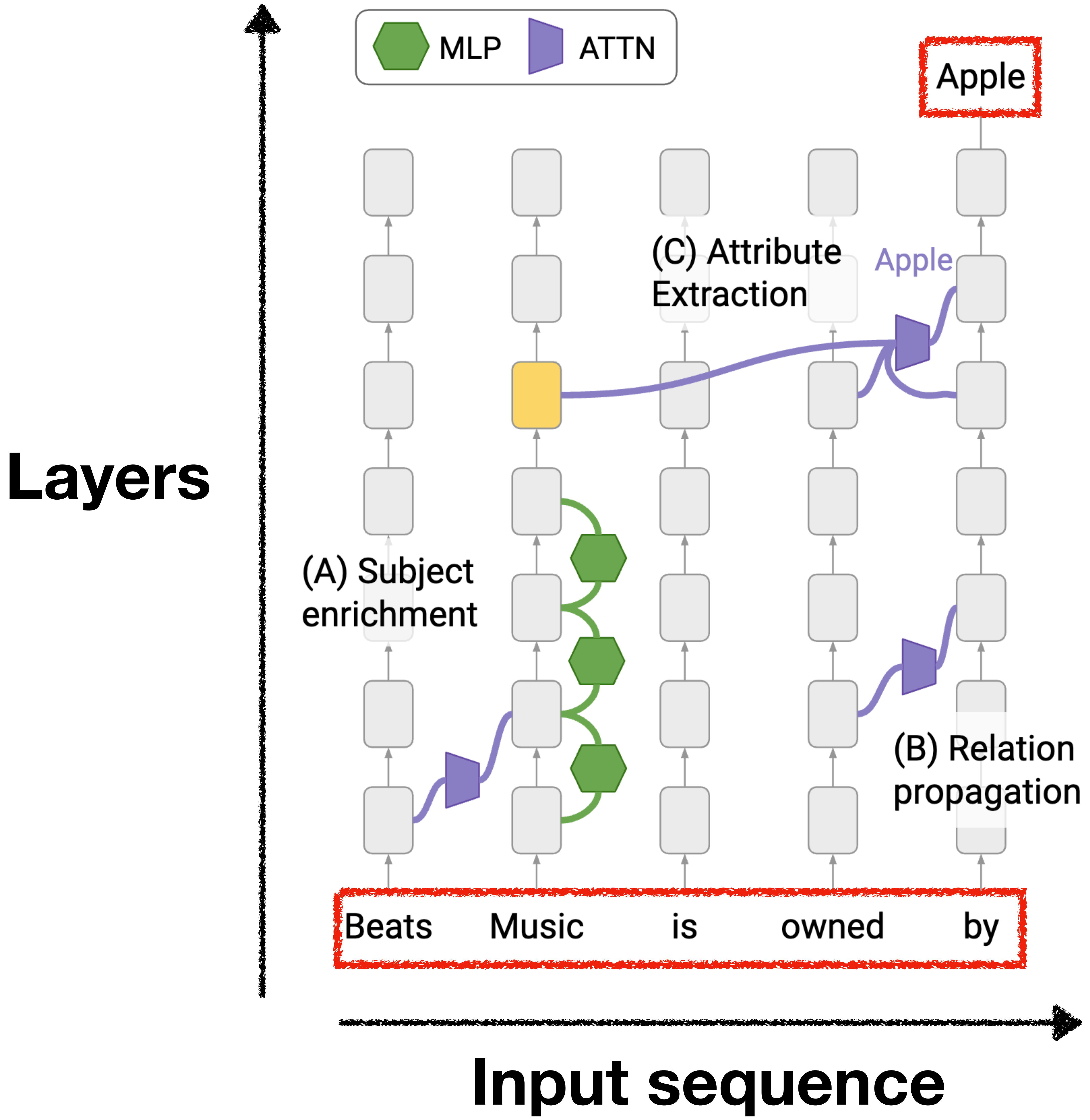
Image credit: [4] Reid et al. 2022. M2D2: A massively multi-domain language modeling dataset

Continual pretraining of LLMs

- Continual pretraining
  - degrades large models, e.g., Llama-7B, if not enough data.
  - improves small models, e.g., GPT2-L.
- Semantic similarity enhances domain specialization.
- Randomizing training domain order improves knowledge accumulation.

Mechanistic understanding of forgetting*

Which stage of knowledge propagation is influenced the most by continual pretraining?



* Sharma, Nitin. In preparation for MSc thesis.
Image credit: https://arxiv.org/pdf/2304.14767

Outline

1. Motivation for lifelong machine learning

2. Academic lifelong learning

3. Our works
    - Disentangled learning
    - Lifelong knowledge organisation
    - Continual pretraining of LLMs

4. **Final thoughts**

Final thoughts

- Lifelong learning is an underappreciated ML paradigm that
  - has applications to smart devices, LLMs, robotics, etc.
  - becomes more relevant with foundation models.
  - has not been studied thoroughly (yet).
- Modular architectures is the way to prevent catastrophic forgetting.
- LLMs can organize knowledge.
- Maybe rethinking the scope of lifelong learning with foundation models?

Thanks!






References

[1] van de Ven, G. & Tolias, A. 2019. Three scenarios for continual learning, https://arxiv.org/abs/1904.07734

[2] Dziadzio, S., et al. 2024. Infinite dSprites for Disentangled Continual Learning: Separating Memory Edits from Generalization

[3] [submitted] Yildiz, Çağatay, et al. 2024. Investigating Continual Pretraining in Large Language Models: Insights and Implications.

[4] Reid et al. 2022. M2D2: A massively multi-domain language modeling dataset