

FAiND the Best Hardware for your AI Application



Barcelona Supercomputing Center
Centro Nacional de Supercomputación

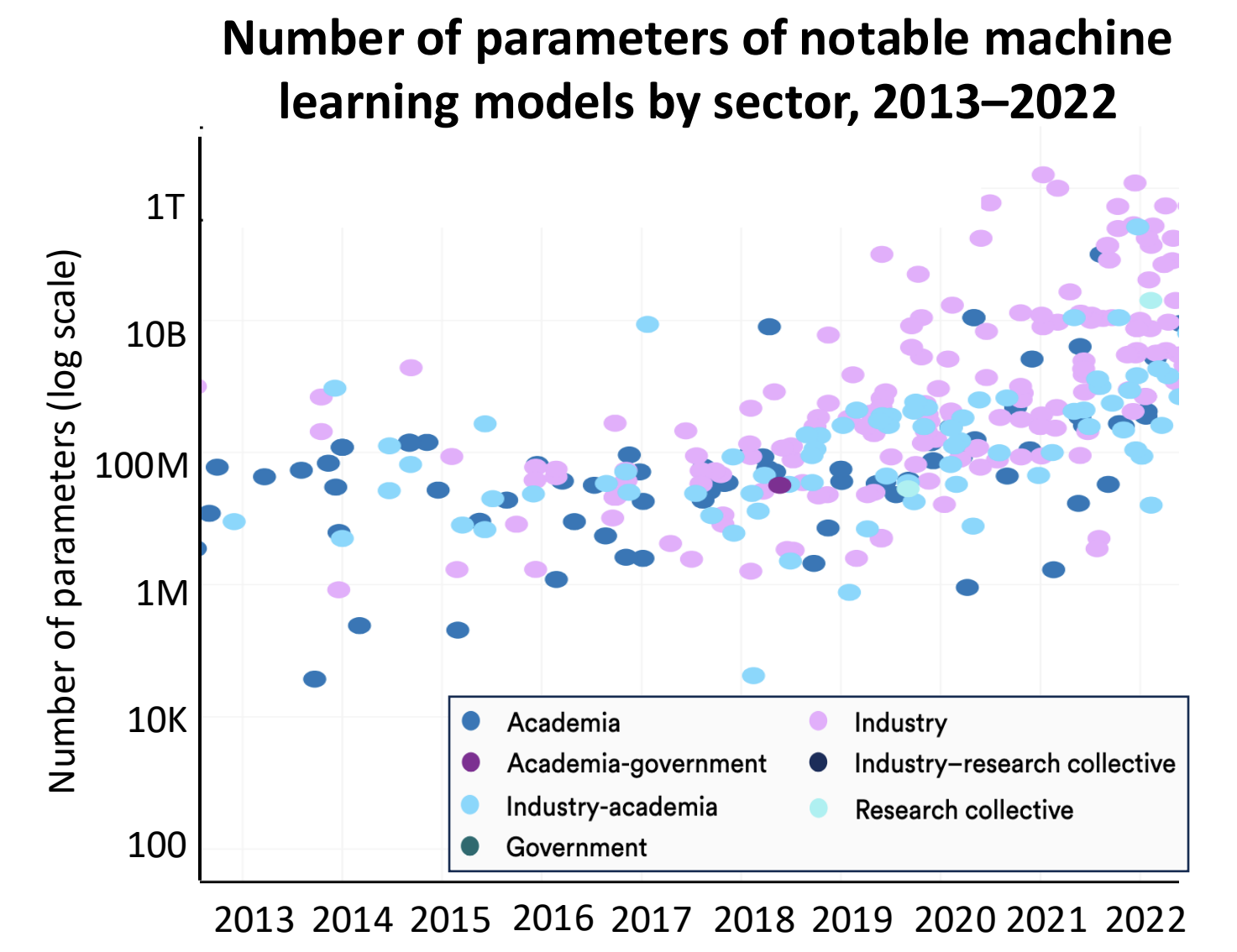
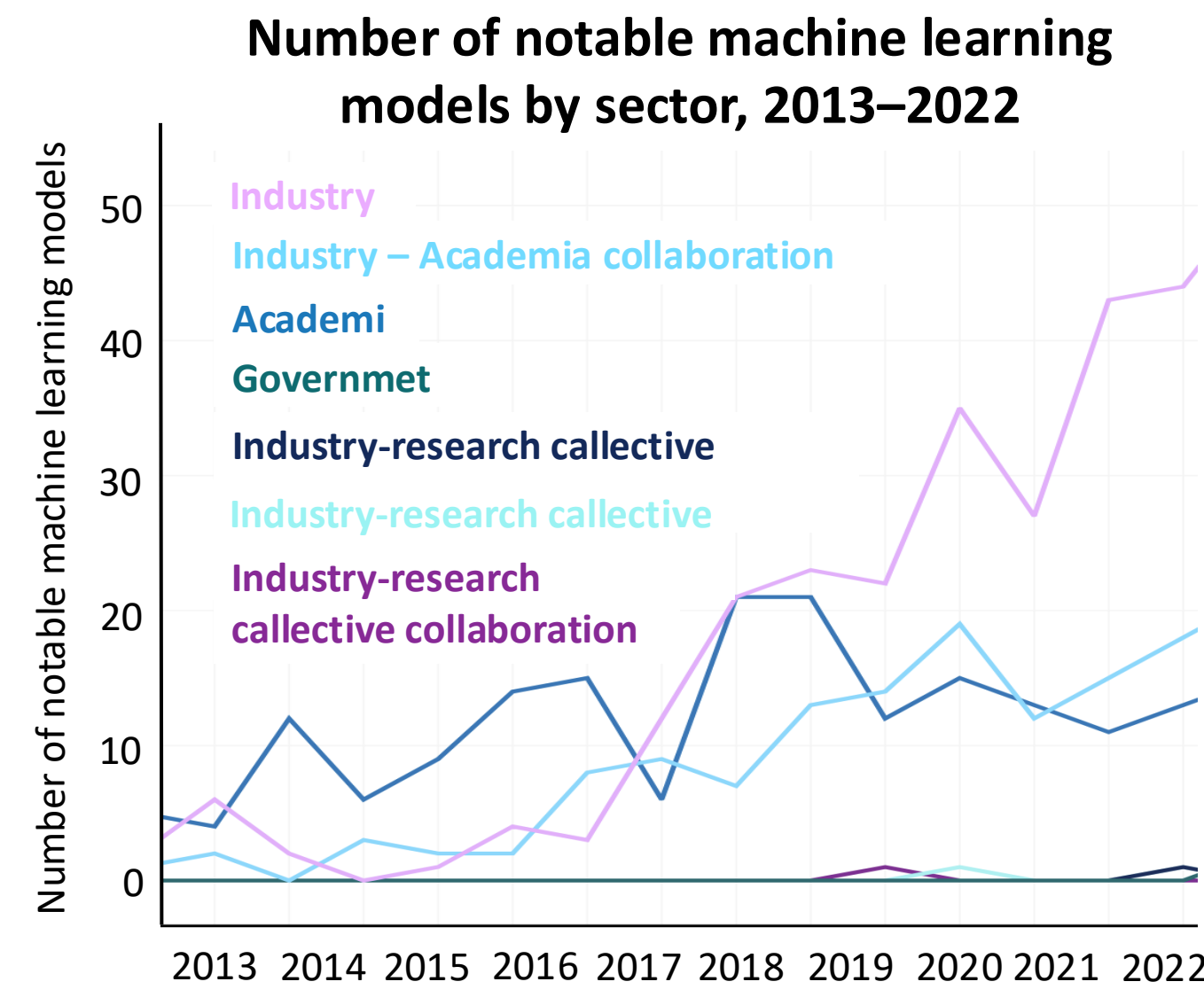
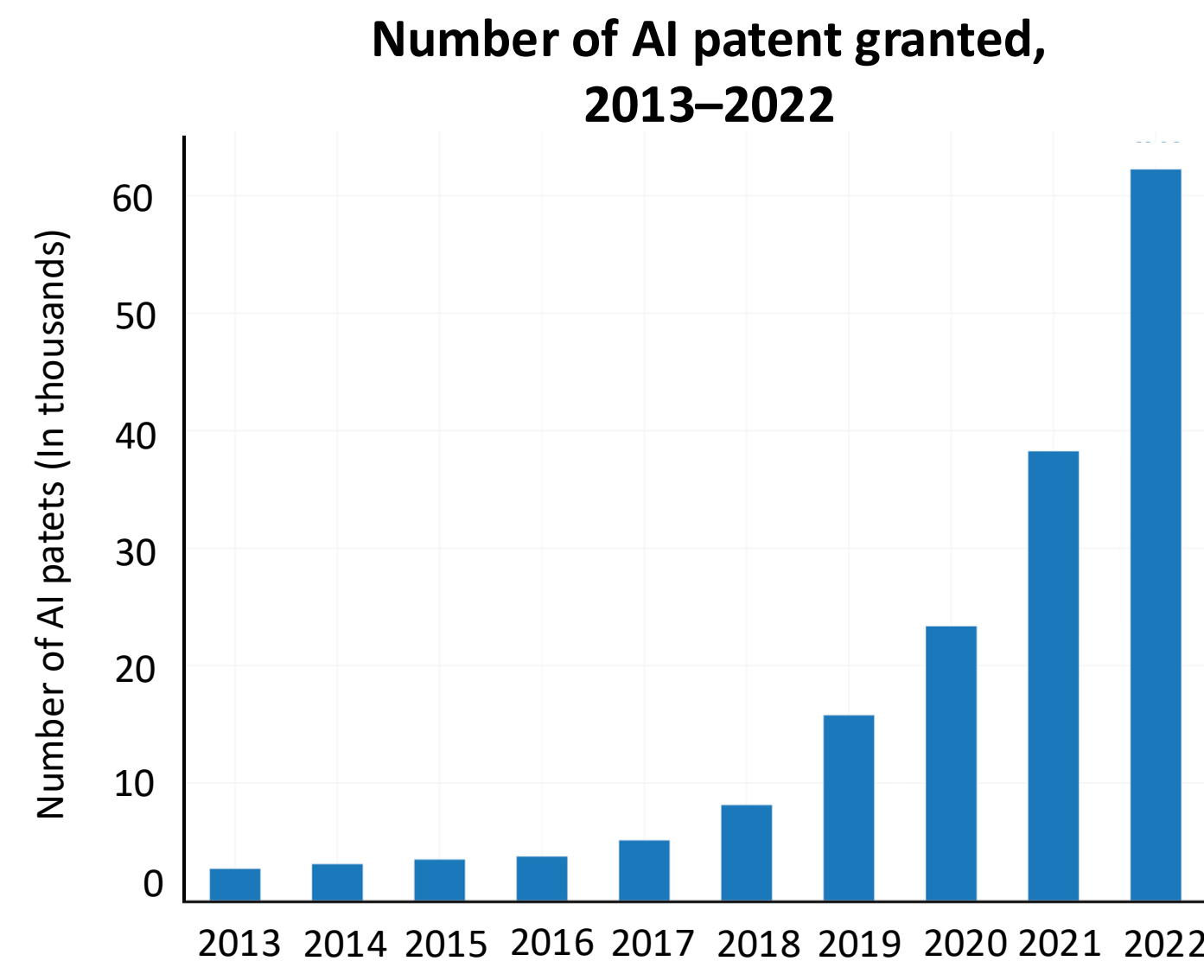
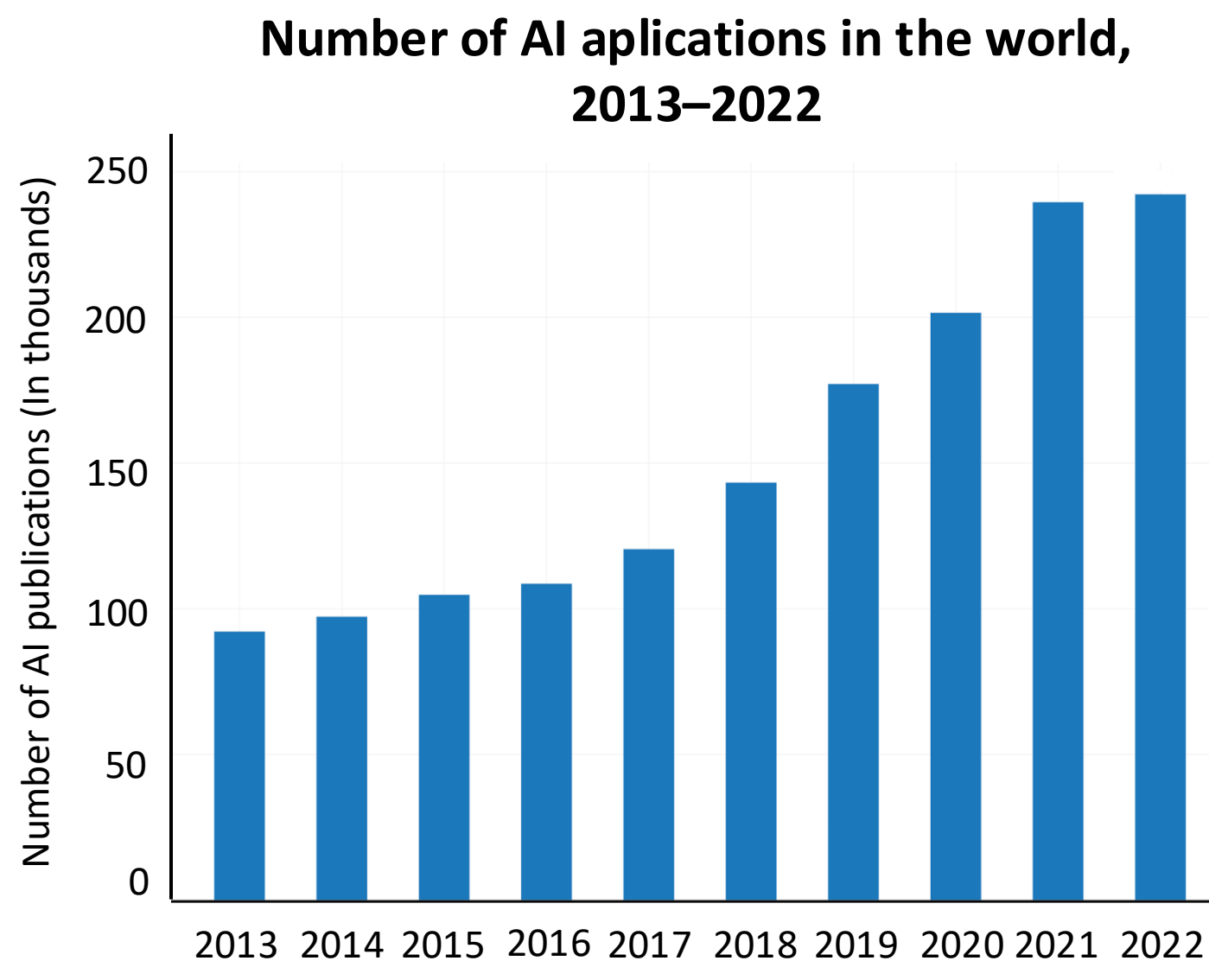
Mariana Carmin¹, Javier Beiro¹, Victor Xirau¹, Pouya Esmaili¹, Petar Radojkovic¹, Eduard Ayguadé¹, Emanuele Confalonieri², Rishabh Dubey², Jason Adlard²

¹Barcelona Supercomputing Center, Spain ²Micron Technology, Italy



Problem: Overwhelming AI revolution

May you live in interesting times



Source: Stanford's Artificial Intelligence Index Report 2024: Chapter 1: Research and Development.

Part of the solution: FAiNDER.eu

Filter the table

Filter by: Model, Phase, Layer, Operation, Sparsity, Data Types, Memory Footprint, Hardware Platforms

Check all the references

Model's Characteristics

AI models

Model	Phase	Layers	Operation	Sparsity	Data Type	Memory Footprint	Production hardware platform	Features	Industry usage
Convolutional Neural Networks (CNN)		Convolution	Matrix-matrix multiply	Dense matrix, Dense matrix	1-bit 2-bit 8-bit	< 1GB	CPU	Realtime requirements	
		Activations functions	ReLU, Sigmoid or Tanh	Dense matrix, Dense matrix					
Transformers		Concatenation	Vector-vector concat	Dense vector, Dense vector					
		Neural Network	Activation functions	ReLU					
Deep Neural Networks (DNN)	Inference	Embedding	Embedding lookup						
		Look-up aggregation							
Deep Learning Recommendation Systems (DLRS)		Multiple and							
		Activation fu							
Graph Neural Network (GNN)	Inference	Recurrent layers	Weights and bias operations	Vector-matrix multiply Vector-Vector add	Dense vector, Dense matrix Dense vector, Dense vector	< 1GB	GPU TPU	Real-time requirements Sequential execution dependencies	Google: Natural language Translation, Speech rec Driving: Baidu: Speech Speech rec
		Backpropagation through time		Partial derivatives over a vector Vector-vector add	Dense vector Dense vector, Dense vector				
Recurrent Neural Networks (RNN)	Training	Forward pass	All inference operation			< 1GB	GPU TPU	Can exploit model and data parallelism	

References

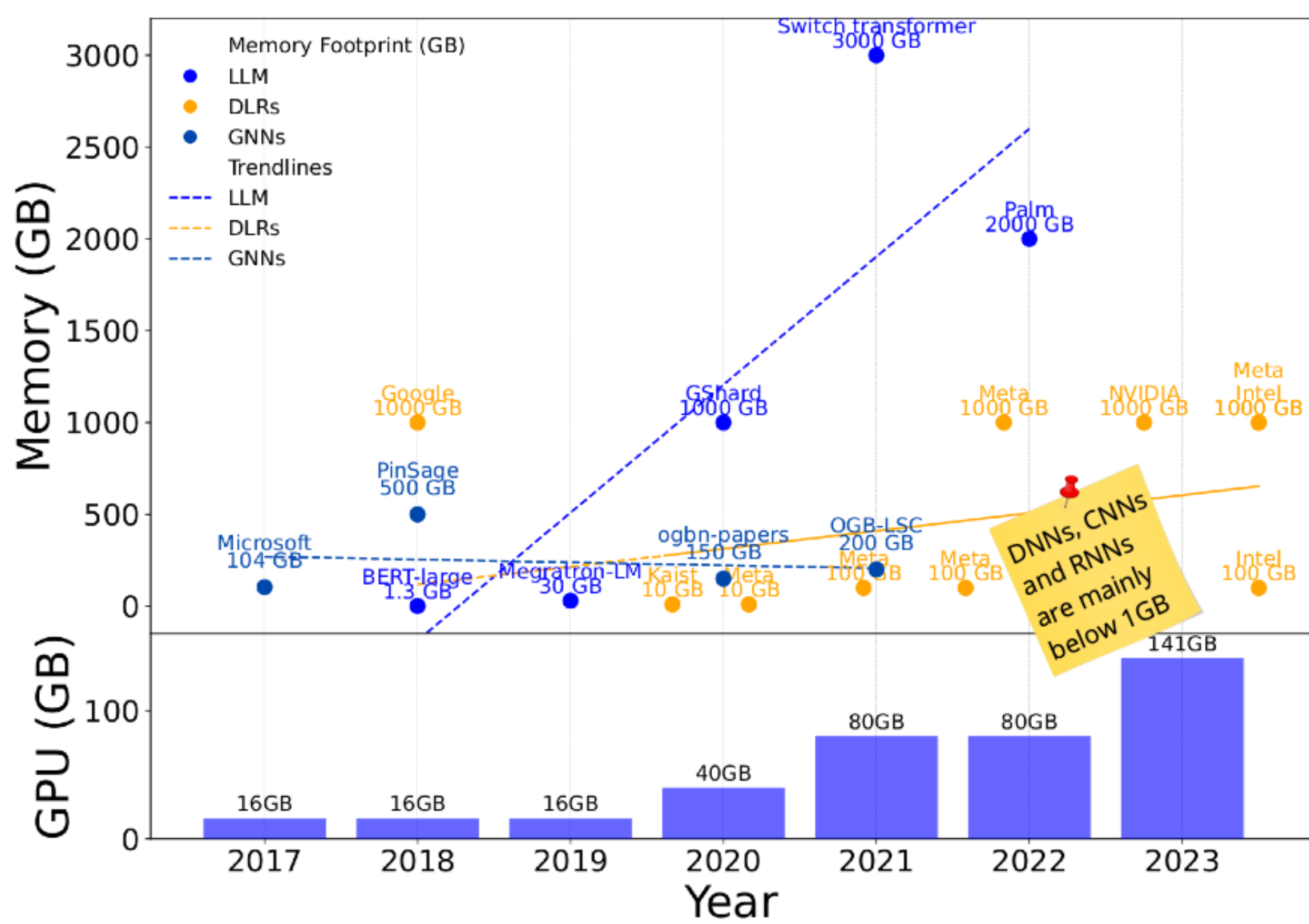
- Robert A Jacobs. Increased rates of convergence through learning rate adaptation. *Neural networks*, 1(4):295-307, 1988. [Copy Text](#) • [View Online](#)
- Justus, D., Brennan, J., Bonner, S. and McGough, A.S., 2018, December. Predicting the computational cost of deep learning models. In 2018 IEEE international conference on big data (Big Data) (pp. 3873-3882). IEEE. [Copy Text](#) • [View Online](#)

Check the statements references

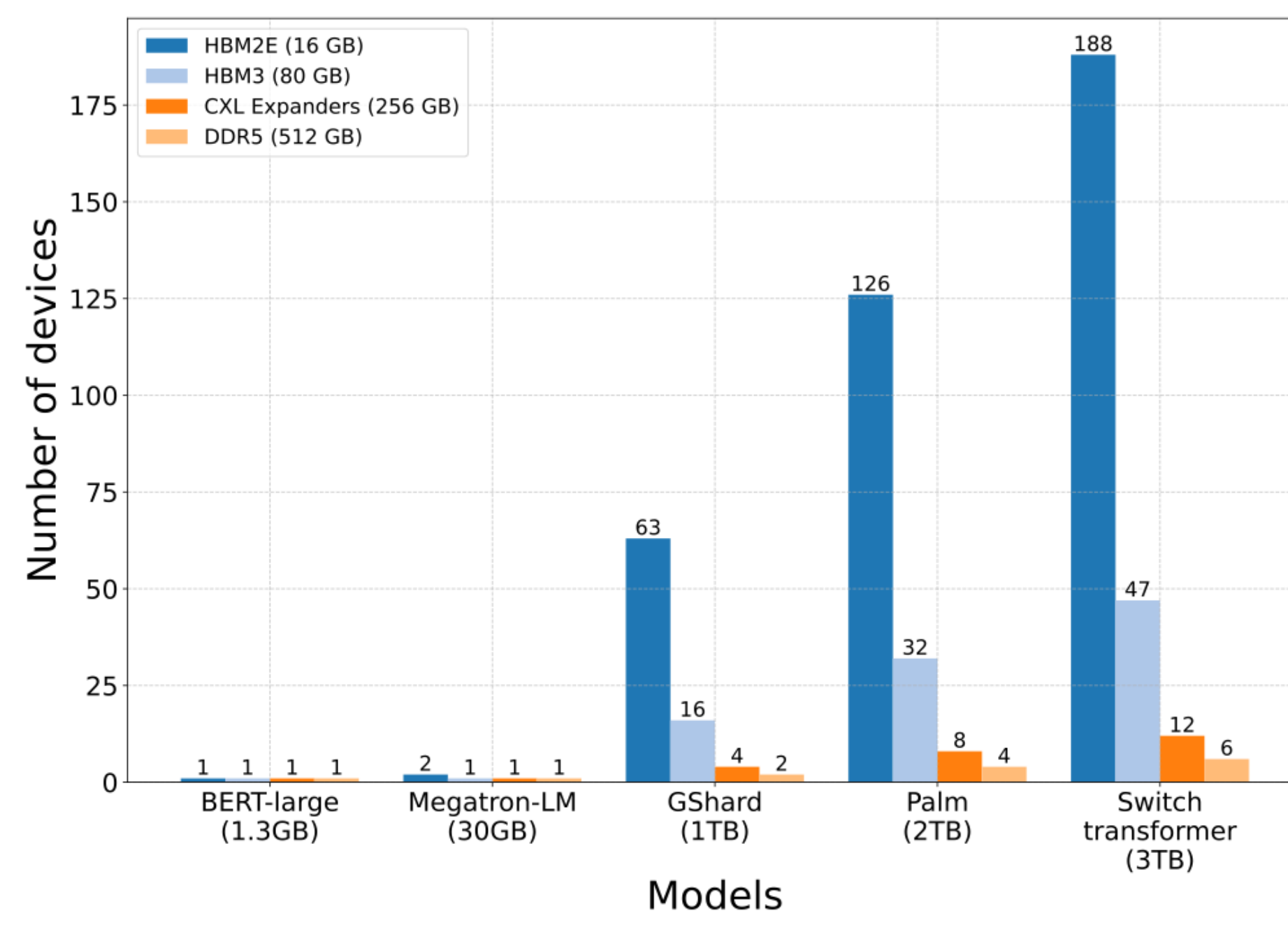
Let's analyse the memory requirements

Memory increase 2017-2023:

- GPU capacity: 9x
- LLM model size: 2038x

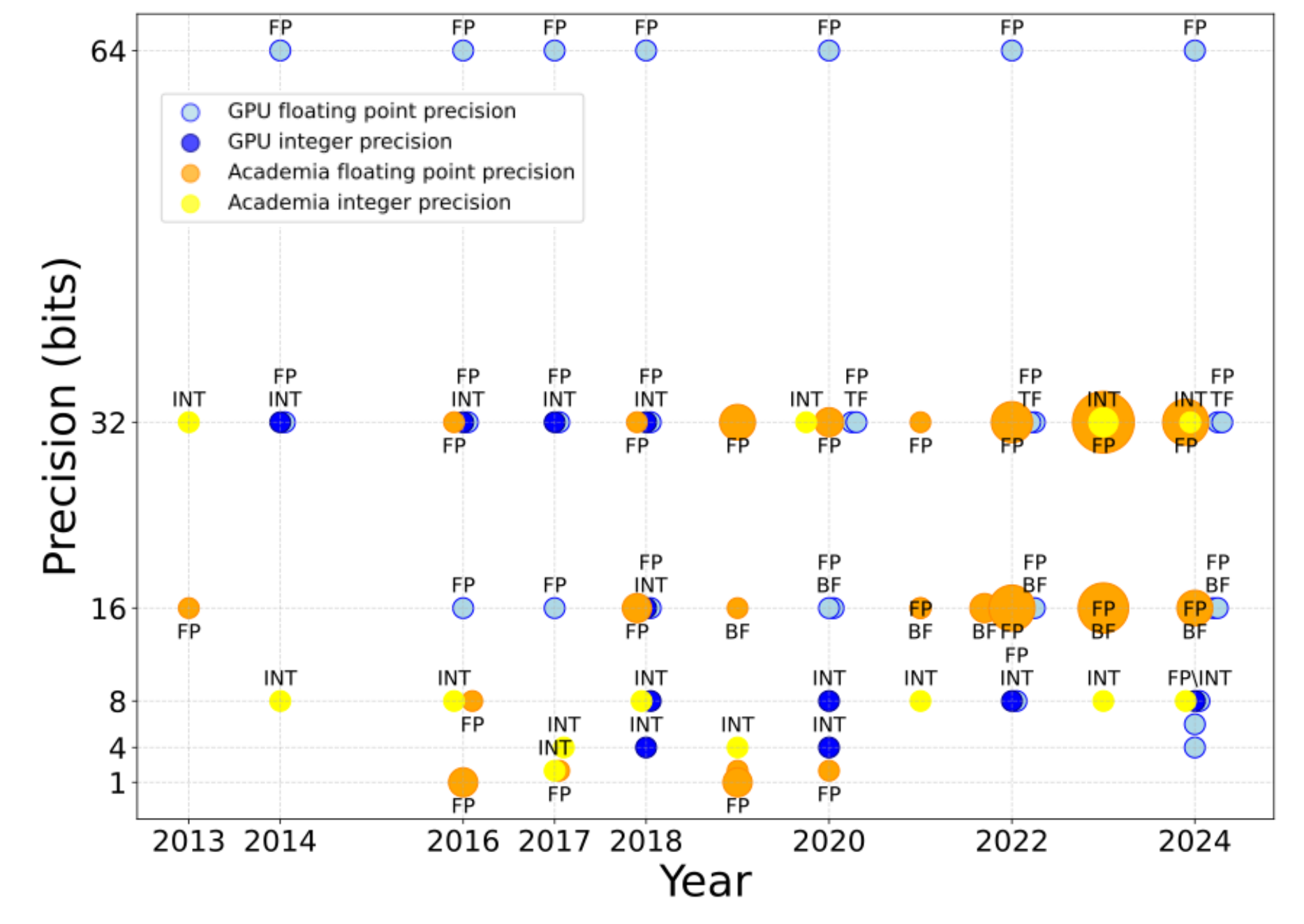


LLMs require tens to hundreds of memory devices



Lower precision datatypes:

- Explored by academia
- Adopted by industry or in commercial products



FAiND our web page!

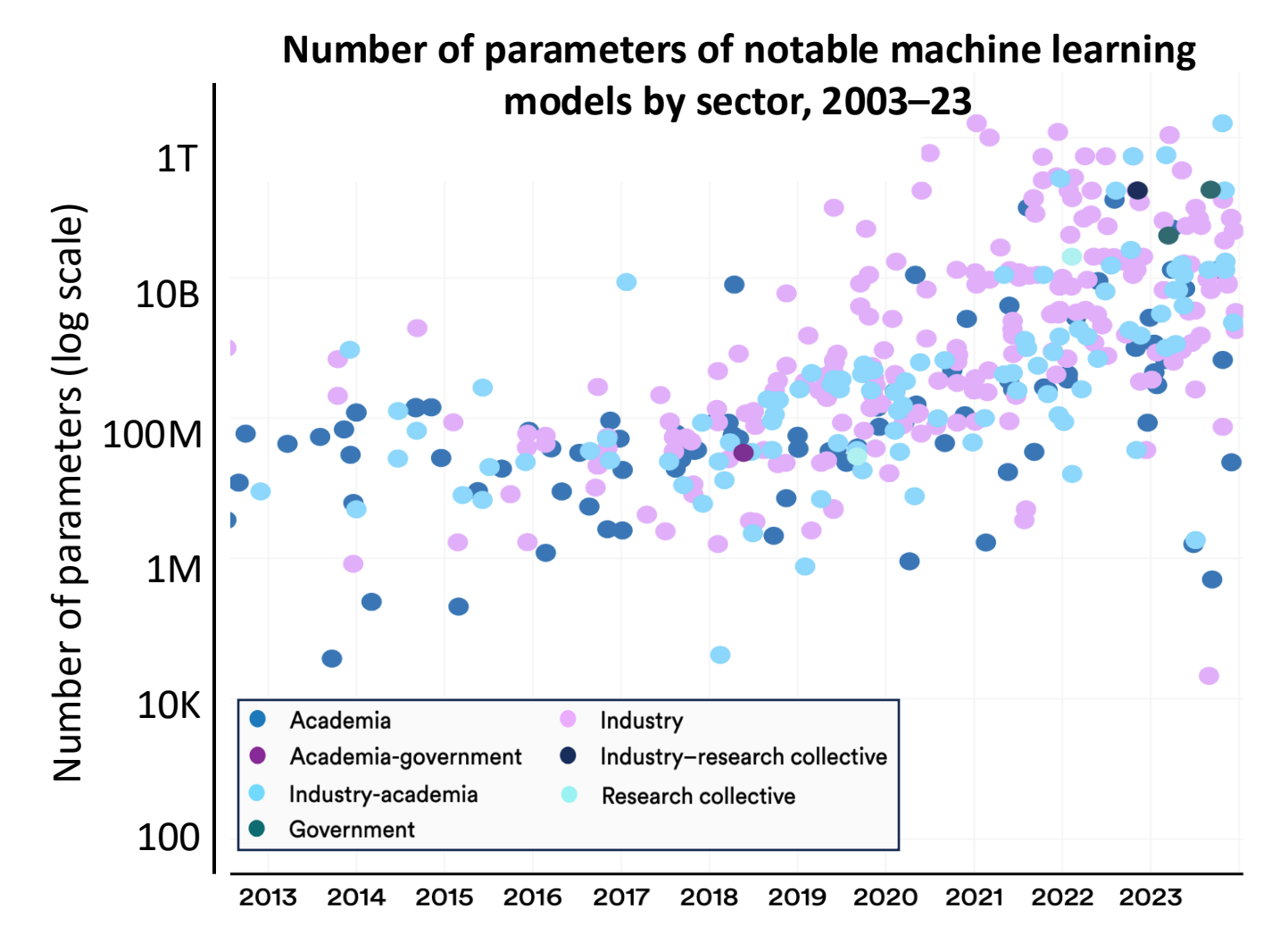
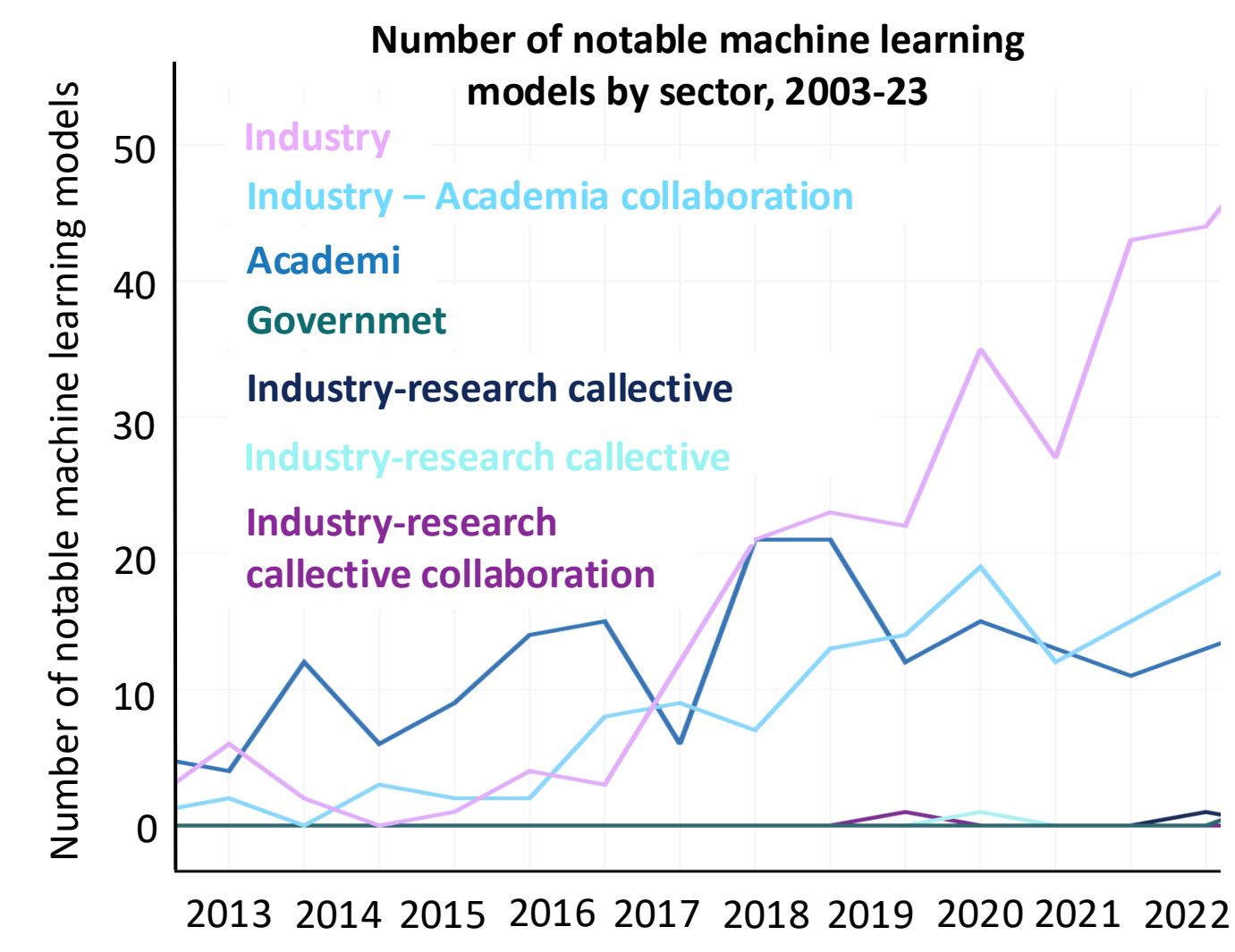
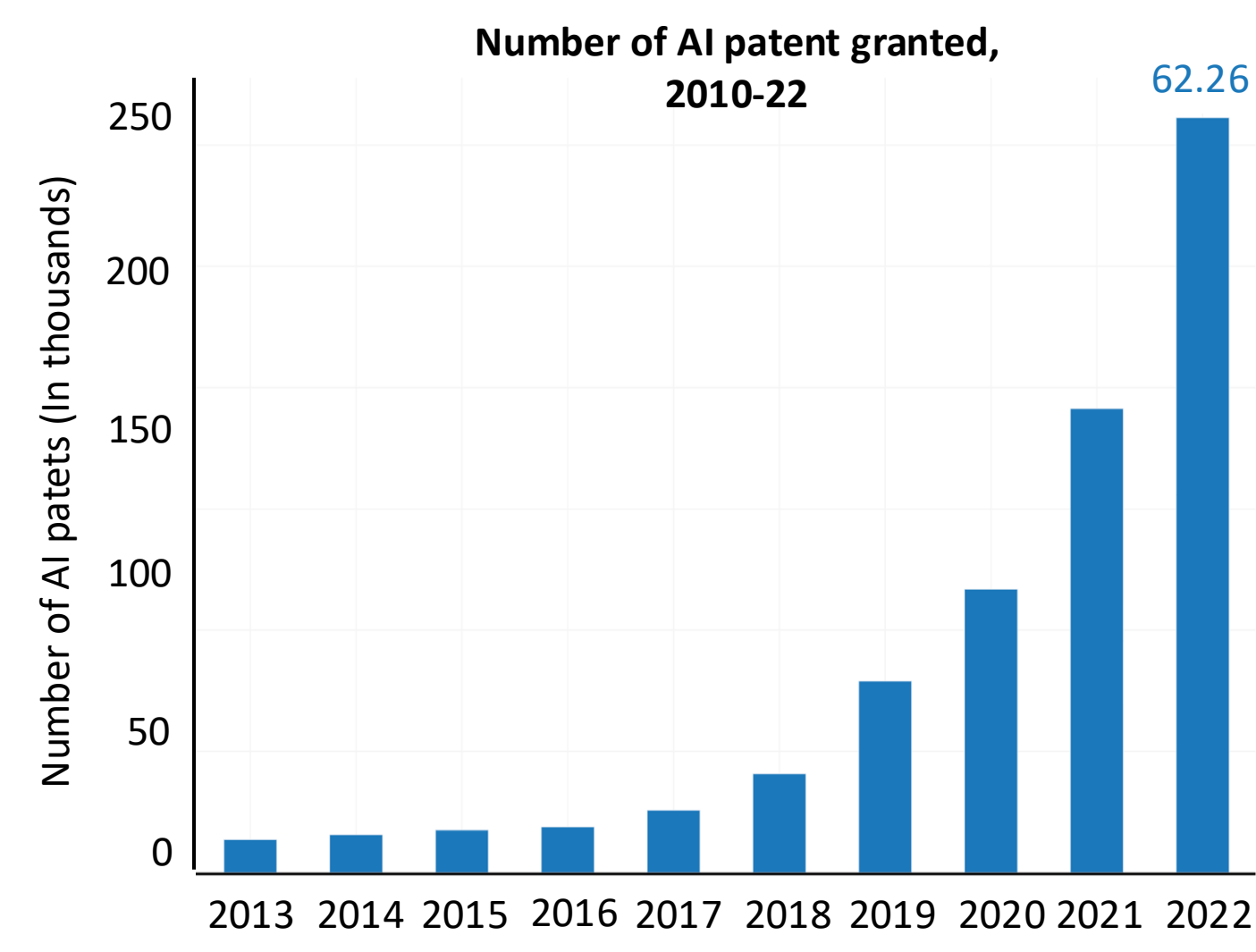
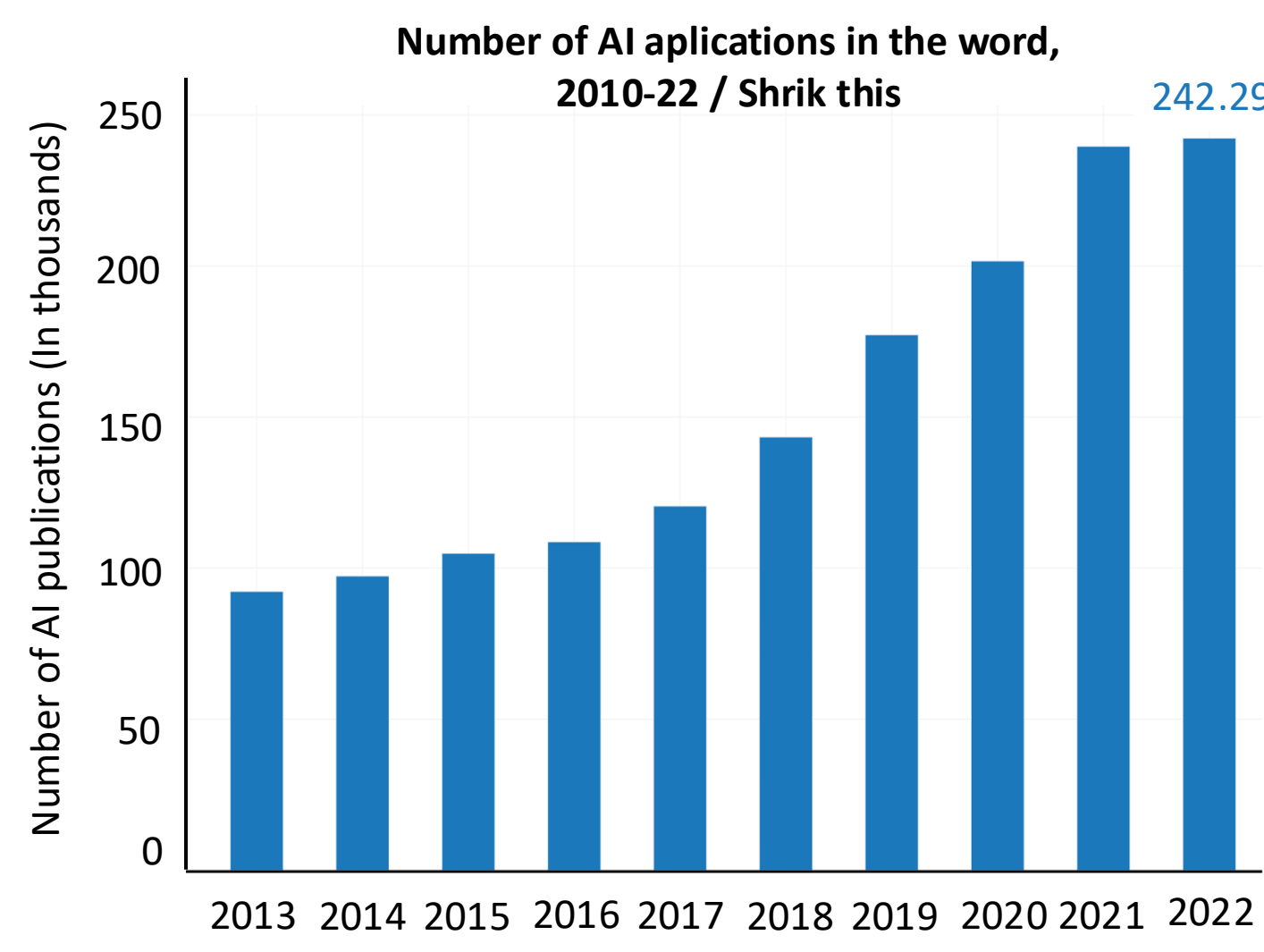


Subscribe to our newsletter!



Contact us!

- ✉ mariana.carmin@bsc.es
- ✉ javier.beiro@bsc.es



Model's Characteristics

AI models

Filter the table

Filter by: Model, Phase, Layer, Operation, Sparsity, Data Types, Memory Footprint, Hardware Platforms

Clear all filters [References](#)

Model	Phase	Layers	Operation	Sparsity	Data Type	Memory Footprint	Production hardware platform	Features	Industry usage
Convolutional Neural Networks (CNN)		Convolution	Matrix-matrix multiply	Dense matrix, Dense matrix	1-bit, 2-bit, 8-bit	< 1GB	CPU	Realtime requirements	
Transformers		Concatenation	Vector-vector concat	Dense vector, Dense vector					
Deep Neural Networks (DNN)	Inference	Neural Network	Activation functions	ReLU	Vector				
Deep Learning Recommendation Systems (DLRS)		Embedding	Embedding lookup	Lookup	agregation				
Graph Neural Network (GNN)	Inference		Multiple and	Activation fu					
Recurrent Neural Networks (RNN)	Inference	Recurrent layers	Weights and bias operations	Vector-matrix multiply	Dense vector, Dense matrix		INT8, FP16, BF16	< 1GB	GPU, TPU, Real-time requirements, Sequential execution dependencies
	Training	Backpropagation through time	Partial derivatives over a vector	Vector-vector add	Dense vector		FP32, FP16	< 1GB	GPU, TPU, Can exploit model and data parallelism
		Forward pass	All inference operation						Google: Natural language Translation, Speech recognition; Baidu: Speech recognition

References

- Robert A Jacobs. Increased rates of convergence through learning rate adaptation. *Neural networks*, 1(4):295-307, 1988. [Copy Text](#) • [View Online](#)
- Justus, D., Brennan, J., Bonner, S. and McGough, A.S., 2018, December. Predicting the computational cost of deep learning models. In *2018 IEEE International conference on big data (Big Data)* (pp. 3873-3882). IEEE. [Copy Text](#) • [View Online](#)

Check all the references

Check the statements references

Filter the table

Filter by

Clear all filters [References](#)

- Model
- Phase
- Layer
- Operation
- Sparsity
- Data Types
- Memory Footprint
- Hardware Platforms

Model	Phase	Layers	Operation	Sparsity	Data Type	Memory Footprint	Production hardware platforms	Features	Industry Usage
Convolutional Neural Networks (CNN)		Convolution	Matrix-matrix multiply	Dense matrix, Dense matrix	1-bit 2-bit 8-bit	< 1GB	CPU	Realtime requirements	
		Activations functions	ReLU, Sigmoid or Tanh	Dense matrix, Dense matrix					
Transformers		Concatenation	Vector-vector concat	Dense vector, Dense vector					
Deep Neural Networks (DNN)	Inference	Neural Network	Activation functions	ReLU					
Deep Learning Recommendation Systems (DLRS)		Embedding	Embedding lookup						
			Lookup agregation						
Graph Neural Network (GNN)	Inference		Multiple and						
Recurrent Neural Networks (RNN)	Inference	Recurrent layers	Weights and bias operations	Vector-matrix multiply	Dense vector, Dense matrix		INT8 FP16 BF16		
				Vector-Vector add	Dense vector, Dense vector				
			Backpropagation through time	Partial derivatives over a vector	Dense vector				
Training		Forward pass	Vector-vector add	Dense vector, Dense vector		FP32 FP16		< 1GB	
			All inference operation	...					

AI models

[References](#) X

- Robert A Jacobs. Increased rates of convergence through learning rate adaptation. *Neural networks*, 1(4):295–307, 1988.
[Copy Text](#) • [View Online](#)
- Justus, D., Brennan, J., Bonner, S. and McGough, A.S., 2018, December. Predicting the computational cost of deep learning models. In *2018 IEEE international conference on big data (Big Data)* (pp. 3873-3882). IEEE.
[Copy Text](#) • [View Online](#)

Check individual references

What can you FAiND?

FAiNDER has an explore page where you can FAiND a table that contains:

1. The state-of-the-art AI models.
2. The different characteristics of the models including layers, operations, data sparsity, data types, memory footprint, hardware platforms, general requirements and industry usage.
3. On each cell the specific data available for that model.
4. The list of references including the papers from which the data was taken.

The screenshot displays the FAiNDER explore page interface. A table lists various AI models, with columns for Model, Platform, Layers, Operation, Sparsity, Data Type, Performance metrics, Features, and Industry usage. Callouts highlight specific elements:

- 1:** A list of model types: CNN, DLRS, DNN, GNN, RNN, and Transformers.
- 2:** A header bar for the table columns: Layers, Operation, Sparsity, and Data Type.
- 3:** A callout box labeled 'GPU' pointing to a cell in the Hardware platform column.
- 4:** A 'References' pop-up window showing a list of papers, including:
 - Robert A. Jacobs. Increased rates of convergence through learning rate adaptation. *Neural networks*, 1(4):295-307, 1988. [Copy Text](#) • [View Online](#)
 - Justus, D., Brennan, J., Bonner, S. and McGough, A.S., 2018, December. Predicting the computational cost of deep learning models. In 2018 IEEE International conference on big data (Big Data) (pp. 3873-3882). IEEE. [Copy Text](#) • [View Online](#)