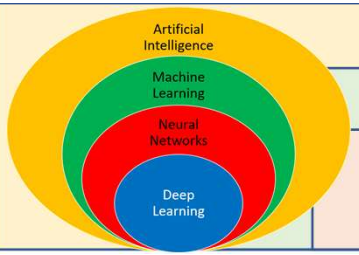


Artificial Intelligence (AI)

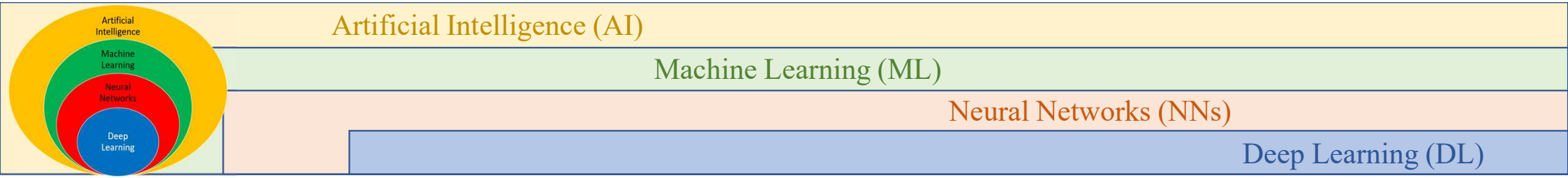
Machine Learning (ML)

Neural Networks (NNs)

Deep Learning (DL)

Advanced Deep Learning

Dr. Rastgoo



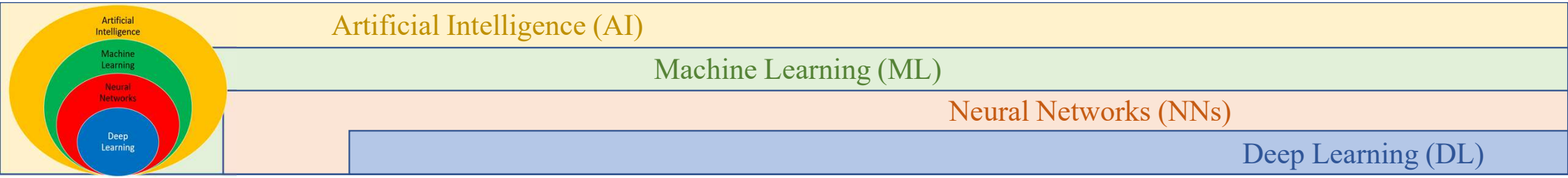
Vision Language Models (VLMs)

- Vision language models (VLMs) are artificial intelligence (AI) models that blend computer vision and natural language processing (NLP) capabilities.

- VLMs learn to map the relationships between text data and visual data such as images or videos, allowing these models to generate text from visual inputs or understand natural language prompts in the context of visual information.

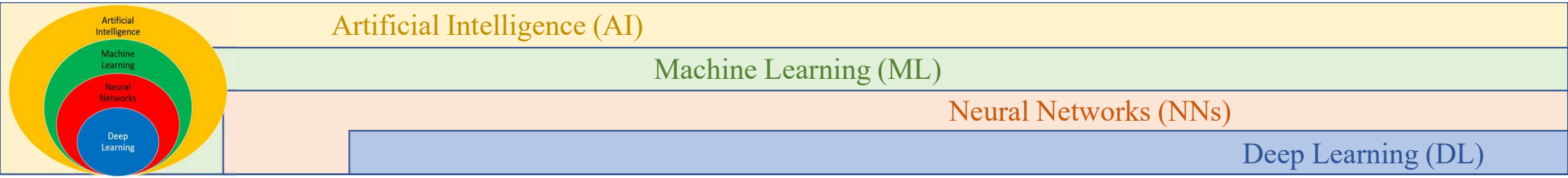
- VLMs, also referred to as visual language models, combine large language models (LLMs) with vision models or visual machine learning (ML) algorithms.

- As multimodal AI systems, VLMs take text and images or videos as input and produce text as output, usually in the form of image or video descriptions, answering questions about an image or identifying parts of an image or objects in a video..



Elements of a vision language model

- Vision language models are typically made up of 2 key components:
- A language encoder
- A vision encoder

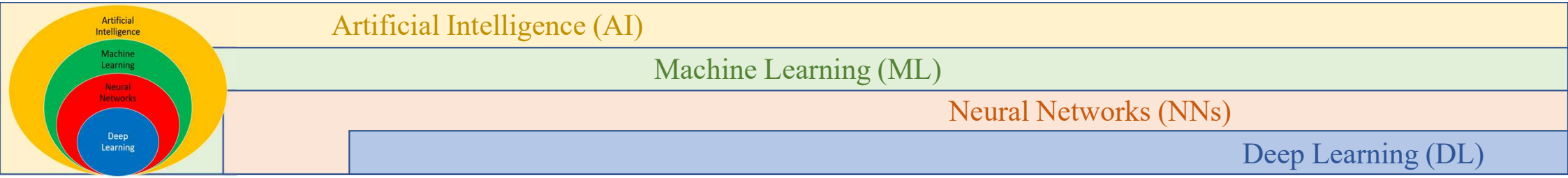


Language encoder

- A language encoder captures the semantic meaning and contextual associations between words and phrases and turns them into text embeddings for AI models to process.

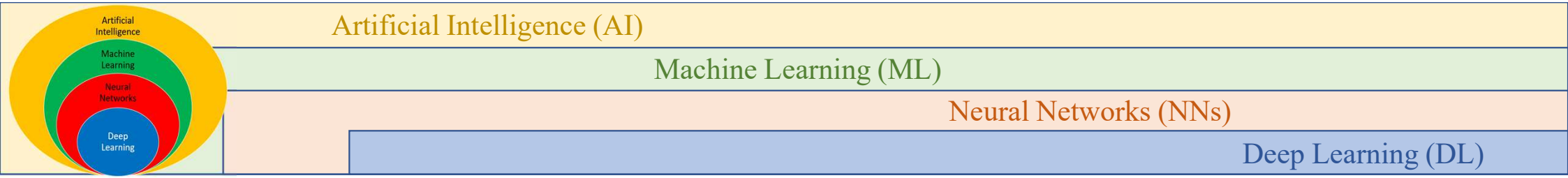
- Most VLMs use a neural network architecture known as the transformer model for their language encoder.

- Examples of transformers include Google's BERT (Bidirectional Encoder Representations from Transformers), one of the first foundation models that underpin many of today's LLMs, and OpenAI's generative pretrained transformer (GPT).



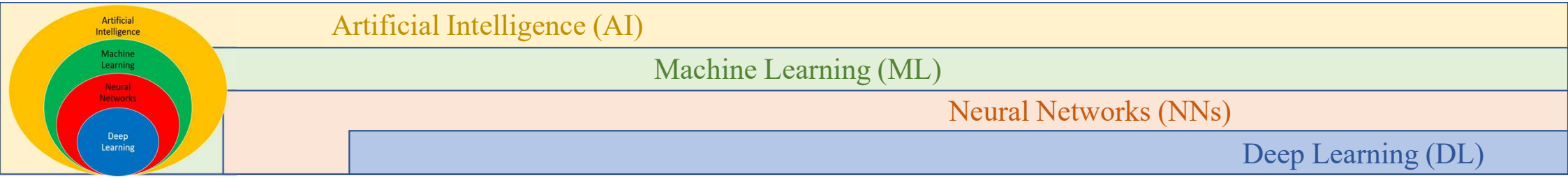
Language encoder

- Here's a brief overview of the transformer architecture:
- Encoders transform input sequences into numerical representations called embeddings that capture the semantics and position of tokens in the input sequence.
- A self-attention mechanism allows transformers to "focus their attention" on the most important tokens in the input sequence, regardless of their position.
- Decoders use this self-attention mechanism and the encoders' embeddings to generate the most statistically probable output sequence.



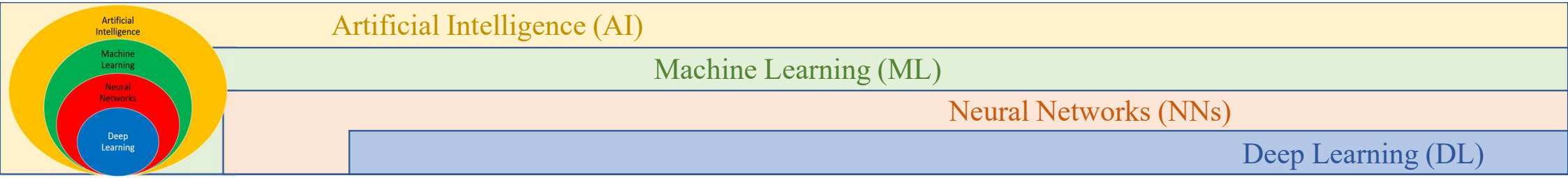
Vision encoder

- A vision encoder extracts vital visual properties such as colors, shapes and textures from an image or video input and converts them into vector embeddings that machine learning models can process.
- Earlier versions of VLMs used deep learning algorithms such as convolutional neural networks for feature extraction.
- More modern vision language models employ a vision transformer (ViT), which applies elements of a transformer-based language model.
- A ViT processes an image into patches and treats them as sequences, akin to tokens in a language transformer.
- The vision transformer then implements self-attention across these patches to create a transformer-based representation of the input image.



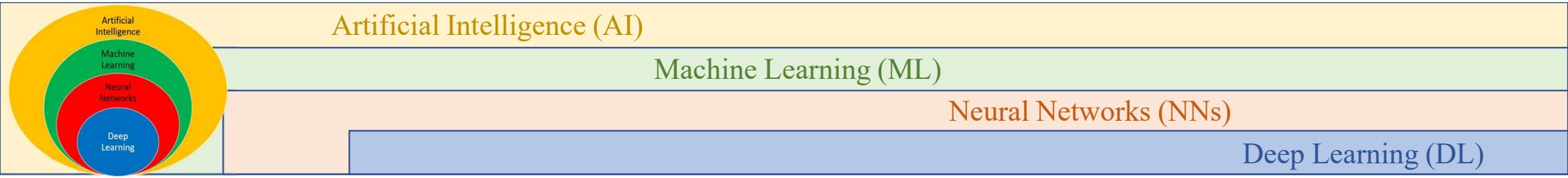
Training vision language models

- Training strategies for vision language models involve aligning and fusing information from both vision and language encoders so the VLM can learn to correlate images with text and make decisions on the 2 modalities together.
- VLM training usually draws upon a mix of approaches:
- Contrastive learning
- Masking
- Generative model training
- Pretrained models.



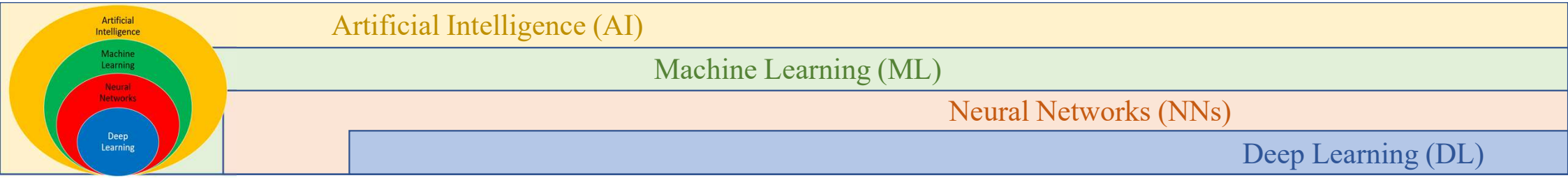
Contrastive learning

- Contrastive learning maps the image and text embeddings from both encoders into a joint or shared embedding space.
- The VLM is trained on datasets of image-text pairs and learns to minimize the distance between embeddings of matching pairs and maximize it for nonmatching pairs.
- A common contrastive learning algorithm is CLIP (Contrastive Language-Image Pretraining).
- CLIP was trained on 400 million image-caption pairs taken from the internet and demonstrated high zero-shot classification accuracy.



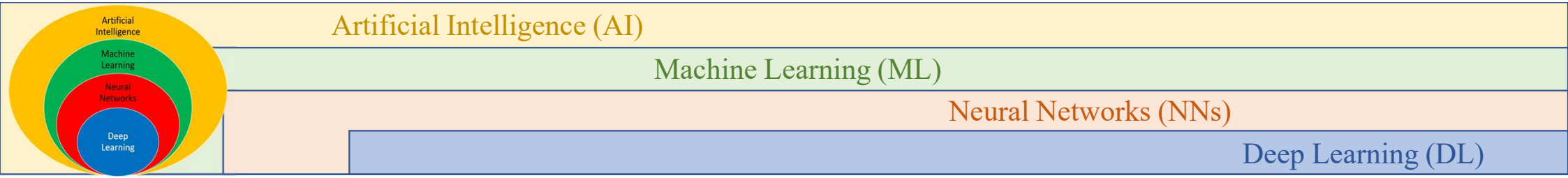
Masking

- Masking is another training technique where visual language models learn to predict randomly obscured parts of an input text or image.
- In masked language modeling, VLMs learn to fill in the missing words in a text caption given an unmasked image.
- Meanwhile, in masked image modeling, VLMs learn to reconstruct the hidden pixels in an image given an unmasked caption.
- An example of a model that uses masking is FLAVA (Foundational Language And Vision Alignment). FLAVA employs a vision transformer as an image encoder and a transformer architecture for both its language encoder and multimodal encoder.
- The multimodal encoder applies a cross-attention mechanism to integrate textual and visual information. FLAVA's training encompasses masked modeling along with contrastive learning.



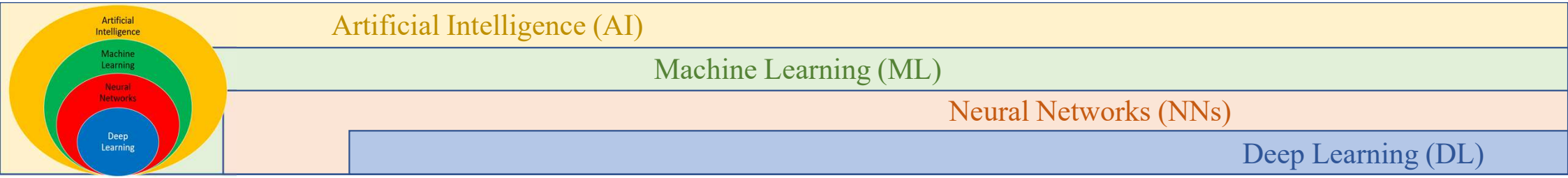
Generative model training

- Generative model training for VLMs entails learning to generate new data.
- Text-to-image generation produces images from the input text, while image-to-text generation produces text, such as captions, image descriptions or summaries, from an input image.
- Examples of popular text-to-image models include diffusion models, such as Google's Imagen, Midjourney, OpenAI's DALL-E (beginning with DALL-E 2) and Stability AI's Stable Diffusion.



Pretrained models

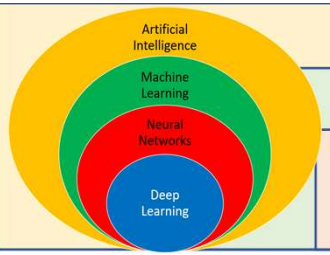
- Training vision language models from scratch can be resource-intensive and expensive, so VLMs can instead be built from pretrained models.
- A pretrained LLM and a pretrained vision encoder can be used, with an added mapping network layer that aligns or projects the visual representation of an image to the LLM's input space.
- LLaVA (Large Language and Vision Assistant) is an example of a VLM developed from pretrained models. This multimodal model uses the Vicuna LLM and the CLIP ViT as a vision encoder, with their outputs merged into a shared dimensional space using a linear projector.
- Vicuna is a chat assistant trained by fine-tuning Llama 2 on user-shared conversations collected from ShareGPT.



Pretrained models

- Gathering high-quality training data for VLMs can be tedious, but there are existing datasets that can be used for pretraining, optimization and fine-tuning for more specific downstream tasks.

- For instance, ImageNet contains millions of annotated images, while COCO has thousands of labeled images for large-scale captioning, object detection and segmentation. Similarly, the LAION dataset consists of billions of multilingual image-text pairs.



Artificial Intelligence (AI)

Machine Learning (ML)

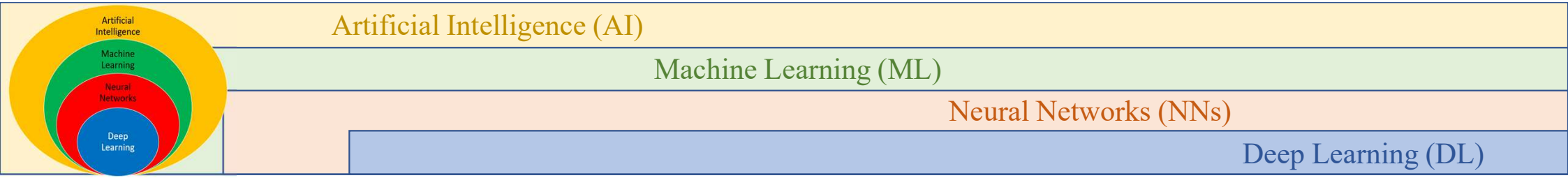
Neural Networks (NNs)

Deep Learning (DL)

Vision language model use cases

- VLMs can bridge the gap between visual and linguistic information. What previously required 2 separate AI models for each modality can now be combined into 1 model.

- VLMs can be used for a range of vision language tasks:
- Captioning and summarization
- Image generation
- Image search and retrieval
- Image segmentation
- Object detection
- Visual question answering (VQA).



Captioning and summarization

- Vision language models can generate detailed image captions or descriptions.
- They can also summarize videos and visual information in documents, such as medical images for healthcare settings or equipment repair charts in manufacturing facilities.

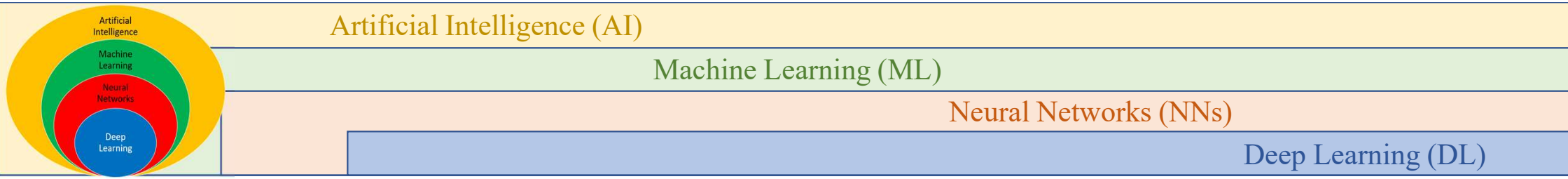


Image generation

- Text-to-image generators such as DALL-E, Imagen, Midjourney and Stable Diffusion can aid in creating art or images to accompany written content.

- Businesses can also use these tools during the design and prototyping phases, helping visualize product ideas.

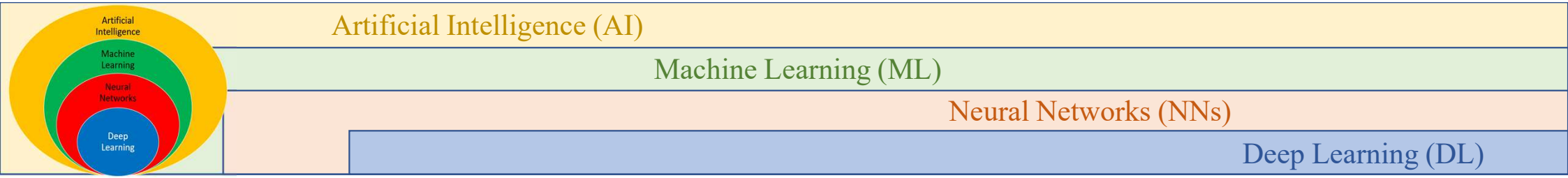


Image search and retrieval

- VLMs can search through large image galleries or video databases and retrieve relevant photos or videos based on a natural language query.

- This can improve the user experience for shoppers on e-commerce websites, for instance, assisting them with finding a particular item or navigating a vast catalog.

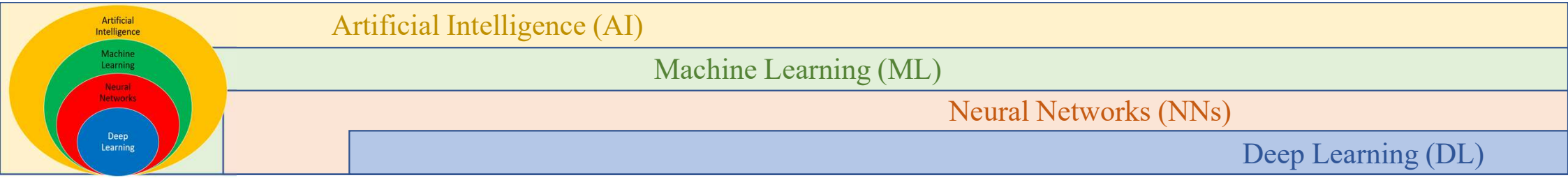
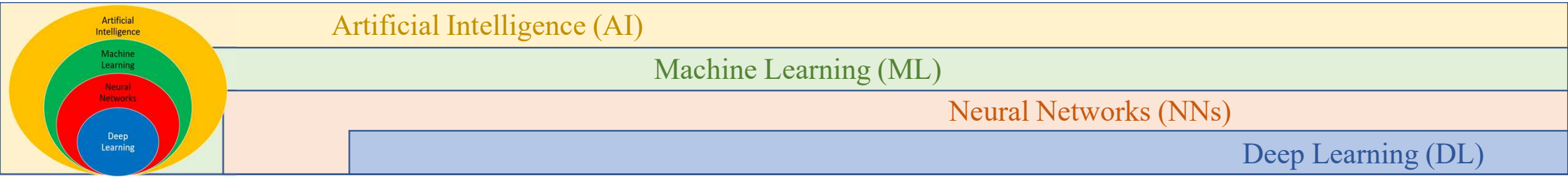


Image segmentation

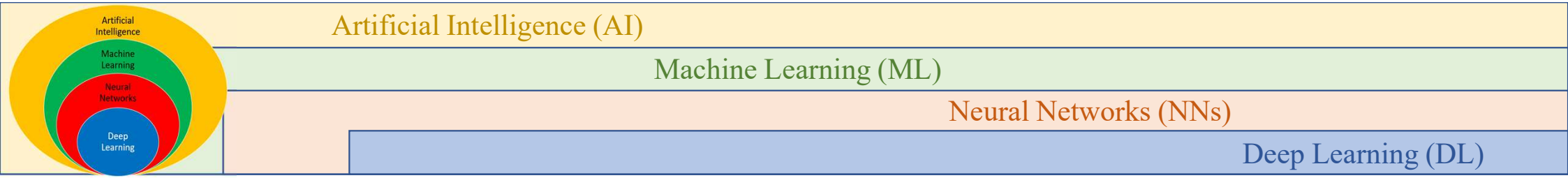
- A visual language model can partition an image into segments based on the spatial features that it has learned about and extracted from the image.
- The VLM can then supply text descriptions of those segments.
- It can also generate bounding boxes to localize objects or provide other forms of annotation such as labels or colored highlighting to specify sections of an image relating to a query.
- This can be valuable for predictive maintenance, for example, helping analyze images or videos of factory floors to detect potential equipment defects in real time.



Object detection

- Vision language models can recognize and classify objects within an image and provide contextual descriptions such as an object's position relative to other visual elements.

- Object detection can be used in robotics, for instance, allowing robots to better understand their environment and comprehend visual instructions.

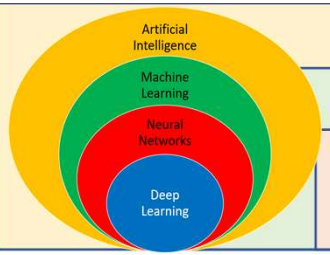


Visual question answering (VQA)

- VLMs can answer questions about images or videos, demonstrating their visual reasoning skills. This can help with image or video analysis and can even be extended to agentic AI applications.

- In the transportation sector, for example, AI agents can be tasked with analyzing road inspection videos and identifying hazards such as damaged road signs, faulty traffic lights and potholes.

- Then, they can be prompted to produce a maintenance report outlining the location and description of those hazards.



Artificial Intelligence (AI)

Machine Learning (ML)

Neural Networks (NNs)

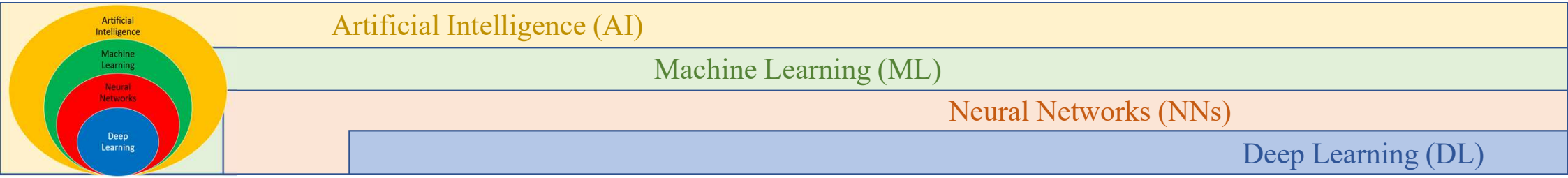
Deep Learning (DL)

Examples of VLMs

- Vision language models are advancing rapidly, with the potential to be as widespread as current advanced LLMs.

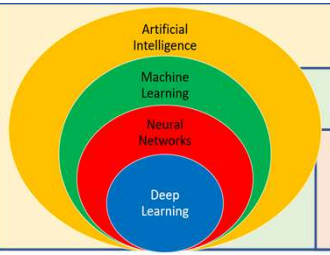
- Here are some examples of popular VLMs:

- DeepSeek-VL2
- Gemini 2.0 Flash
- GPT-4o
- Llama 3.2
- NVLM
- Qwen 2.5-VL.



DeepSeek-VL2

- DeepSeek-VL2 is an open source vision language model with 4.5 billion parameters from the Chinese AI startup DeepSeek.
- It's made up of a vision encoder, a vision language adapter and the DeepSeekMoE LLM, which takes on a Mixture of Experts (MoE) architecture.
- DeepSeek-VL2 has a tiny variant with 1 billion parameters and a small variant with 2.8 billion parameters.



Artificial Intelligence (AI)

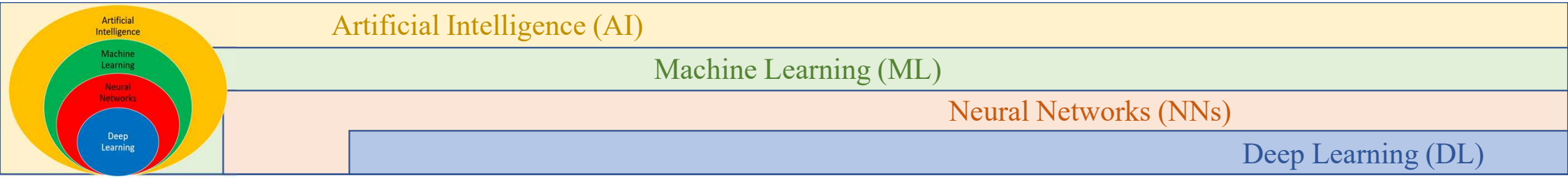
Machine Learning (ML)

Neural Networks (NNs)

Deep Learning (DL)

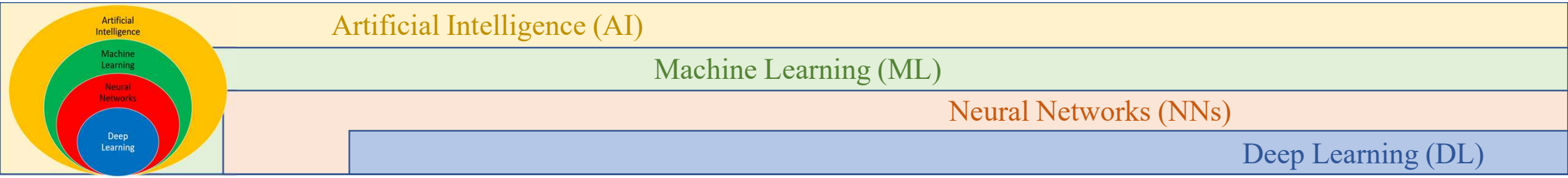
DeepSeek-VL2

- DeepSeek-VL2 is an open source vision language model with 4.5 billion parameters from the Chinese AI startup DeepSeek.
- It's made up of a vision encoder, a vision language adapter and the DeepSeekMoE LLM, which takes on a Mixture of Experts (MoE) architecture.
- DeepSeek-VL2 has a tiny variant with 1 billion parameters and a small variant with 2.8 billion parameters.



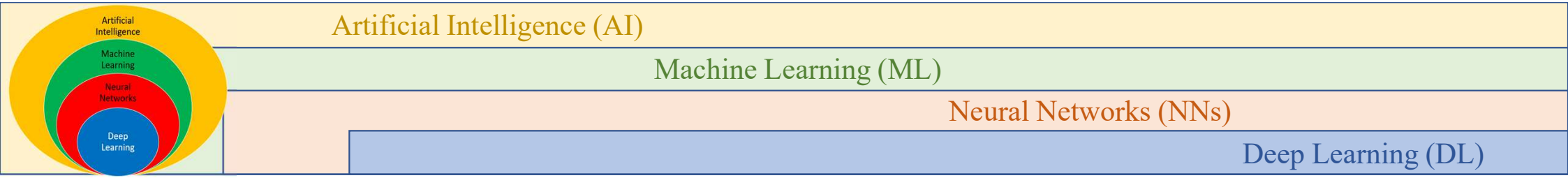
Gemini 2.0 Flash

- Gemini 2.0 Flash is part of the Google Gemini suite of models.
- Input modalities include audio, image, text and video, with a text-only output.
- An image generation feature is on its way.



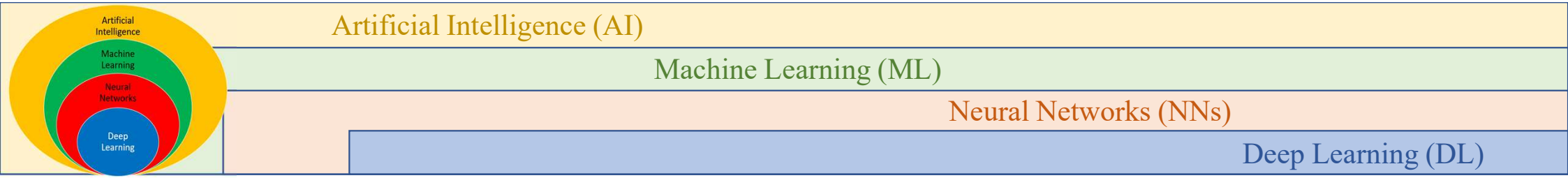
GPT-4o

- OpenAI's GPT-4o is a single model trained end-to-end across audio, vision and text data.
- It can accept a mixture of audio, image, text and video inputs and produce any combination of audio, image and text outputs, with the same neural network processing all inputs and outputs.
- Its smaller counterpart, GPT-4o mini, supports both image and text inputs and generates text outputs.



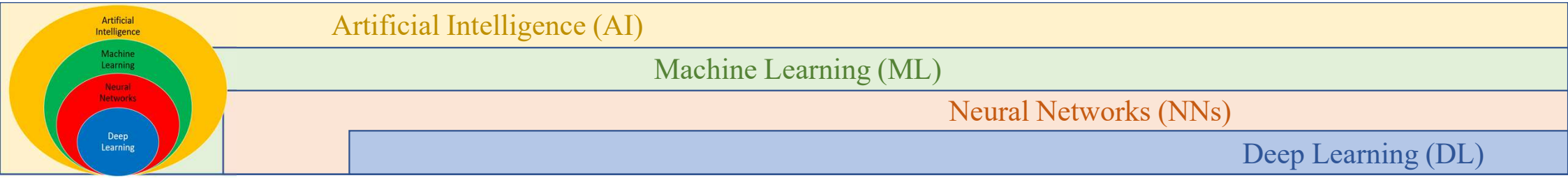
Llama 3.2

- The Llama 3.2 open source models include 2 VLMs in 11 and 90 billion parameter sizes.
- Inputs can be a combination of text and images, with a text-only output.
- According to Meta, the VLM architecture consists of a ViT image encoder, a video adapter and an image adapter.
- The separately trained image adapter has a series of cross-attention layers that feed image encoder representations into the pretrained Llama 3.1 LLM.



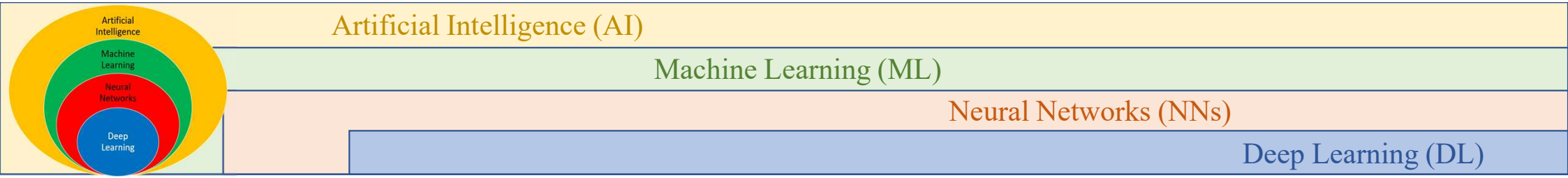
NVLM

- NVLM is a family of multimodal models from NVIDIA.
- NVLM-D is a decoder-only model that feeds image tokens directly into the LLM decoder.
- NVLM-X employs cross-attention to process image tokens and is more efficient for handling high-resolution images.
- NVLM-H takes on a hybrid architecture that combines the decoder-only and cross-attention approaches, improving computational efficiency and reasoning capabilities.



Qwen 2.5-VL

- Qwen 2.5-VL is the flagship vision language model of the Chinese cloud computing company Alibaba Cloud.
- It comes in 3, 7 and 72 billion parameter sizes.
- The model uses a ViT vision encoder and the Qwen 2.5 LLM.
- It can understand videos over an hour long and can navigate desktop and smartphone interfaces.



Vision language model benchmarks

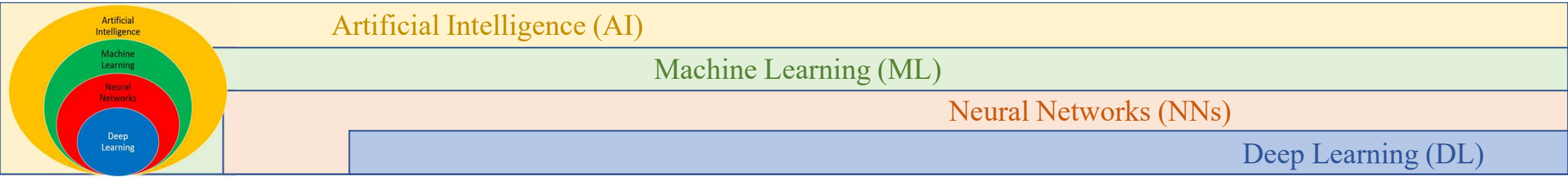
- Like LLMs, VLMs also have their own benchmarks. Each benchmark might have its own leaderboard, but there are also independent leaderboards such as the OpenVLM Leaderboard hosted on Hugging Face that rank open source vision language models based on various metrics.

- Here are some common benchmarks for visual language models:

- MathVista is a benchmark for visual mathematical reasoning.

- MMBench has a collection of multiple-choice questions covering several evaluation dimensions, including object localization, optical character recognition (OCR) and more.

- MMMU (Massive Multidiscipline Multimodal Understanding) contains multimodal multiple-choice challenges across various subjects to measure knowledge, perception and reasoning skills.



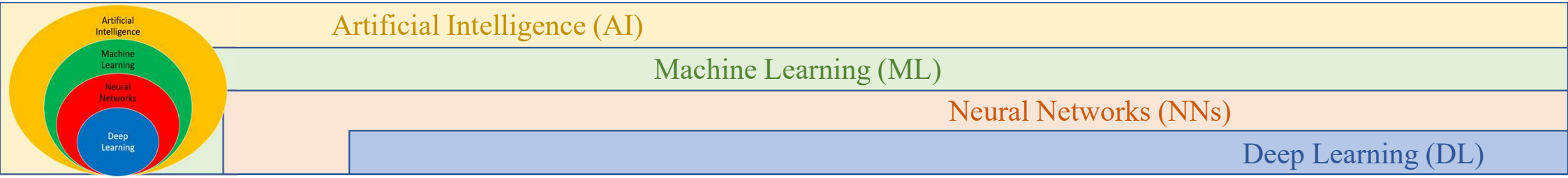
Vision language model benchmarks

- MM-Vet assesses the integration of different VLM capabilities, such as language generation, spatial awareness and more.

- OCRBench focuses on the OCR abilities of VLMs. It consists of 5 components: document-oriented VQA, handwritten mathematical expression recognition, key information extraction, text recognition and scene text-centric VQA.

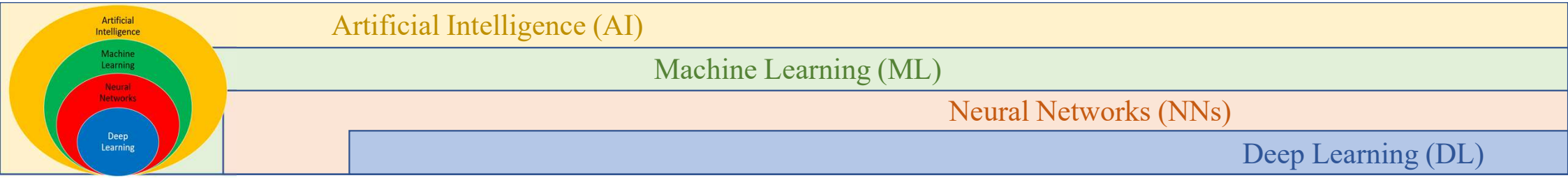
- VQA is one of the earliest VLM benchmarks. The dataset encompasses open-ended questions about images. Other VQA derivatives include GQA (question answering on image scene graphs), OK-VQA (requires outside knowledge for visual question answering), ScienceQA (science question answering) and TextVQA (visual reasoning based on text in images).

- Benchmarking VLMs can be time-consuming, but a few tools can help simplify the process. VLMEvalKit is an open source assessment toolkit that allows for one-command evaluation of VLMs. Another assessment suite is LMMs-Eval, which also provides a command-line interface for evaluation..



Challenges of VLMs

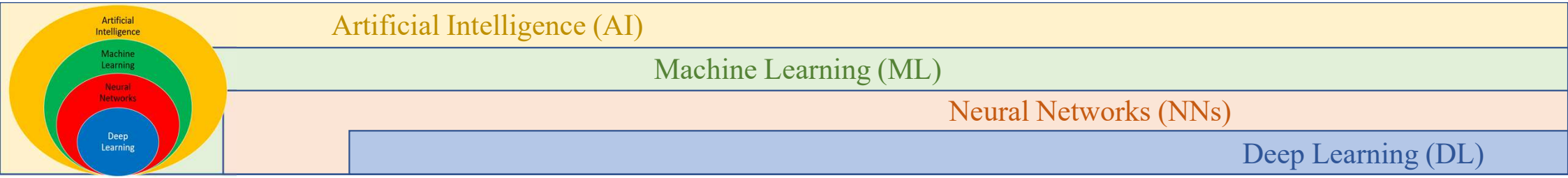
- As with any AI system, VLMs still need to contend with the risks of AI. Enterprises must keep this in mind as they consider integrating vision language models into their internal workflows or implementing them for commercial applications.
- Here are some challenges associated with VLMs:
- Bias
- Cost and complexity
- Generalization
- Hallucinations



Bias

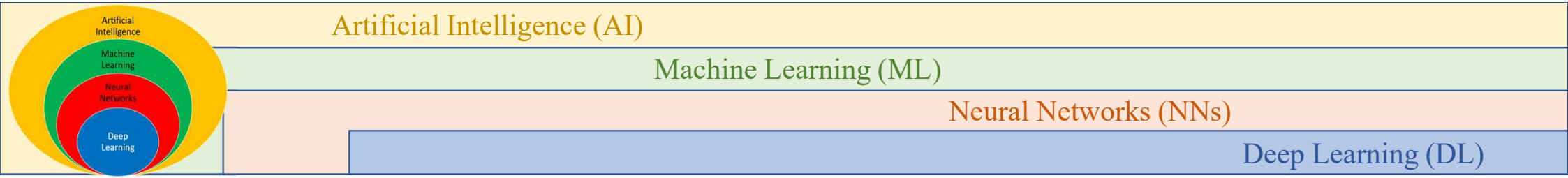
- Visual language models can learn from the biases that might be present in the real-world data they're trained on or from the pretrained models they're built upon.

- Using diverse data sources and incorporating human oversight throughout the process can help mitigate bias.



Cost and complexity

- Vision models and language models are already complex on their own, so merging them can further increase their complexity.
- This complexity leads to the need for more computing resources, making it difficult to deploy VLMs on a large scale.
- Companies must be prepared to invest in the required resources for developing, training and deploying these models.

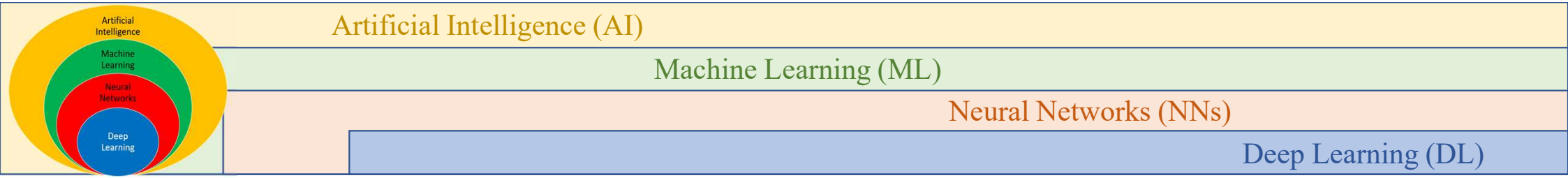


Generalization

- VLMs might falter when it comes to generalization, which is a model's ability to adapt to and make accurate predictions on new, never-before-seen data.

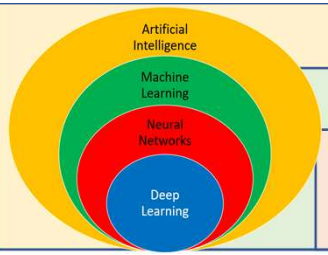
- A balanced dataset that includes outliers or edge cases and employs zero-shot learning can allow VLMs to adapt to novel concepts or atypical image-text combinations.

- IBM's LiveXiv benchmark for visual document understanding tasks can also help. LiveXiv is a dynamic benchmark that's automatically updated monthly, assessing VLMs on questions and images they have likely never seen before.



Hallucinations

- Vision language models can be prone to AI hallucinations.
- Validating the results of these models is a crucial step to make sure they're factually accurate.



Artificial Intelligence (AI)

Machine Learning (ML)

Neural Networks (NNs)

Deep Learning (DL)