EPICURE
Unlocking European-level HPC Support

HPC in ARM Architecture Hackathon

4-7 February 2025

Braga, Portugal

AI in ARM Architecture

Alícia Oliveira
alicia.oliveira@inesctec.pt

Co-funded by the European Union

EuroHPC Joint Undertaking

This project has received funding from the High Performance Computing Joint Undertaking under grant agreement No 101139786

Introduction

What is **A**rtificial **I**ntelligence?

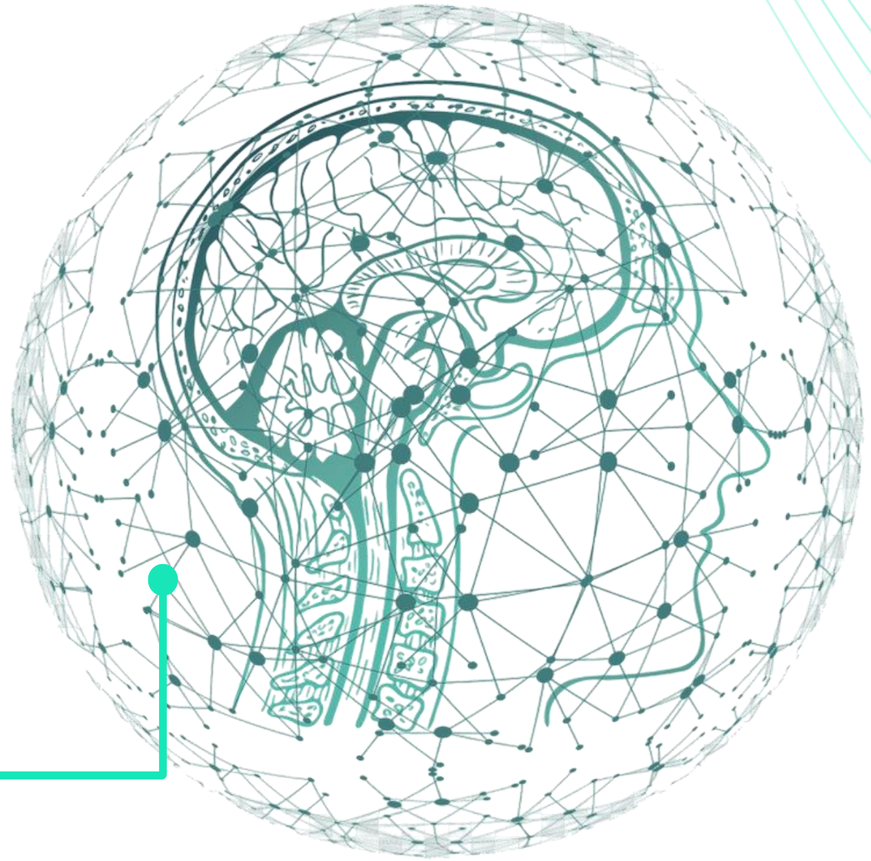
Introduction

What is **A**rtificial **I**ntelligence?

“Artificial Intelligence refers to the development of computer systems for performing tasks that require human intelligence.”

Context

Fields of **AI**

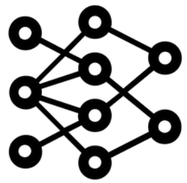


Machine Learning

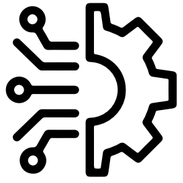
Context

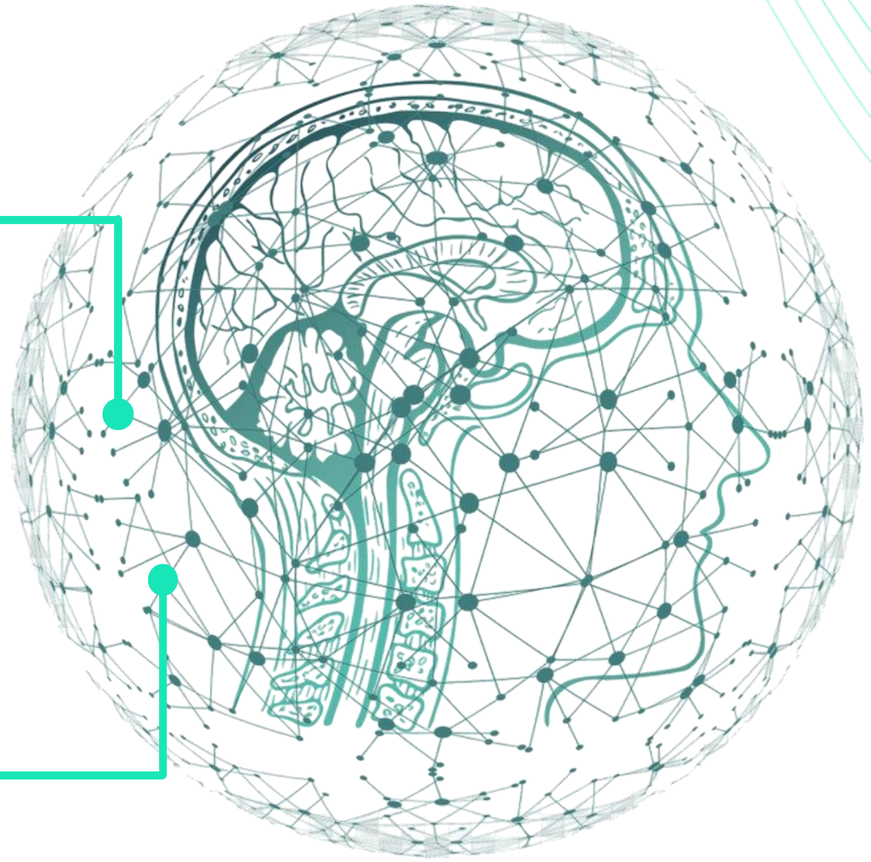
Fields of **AI**

Neural Networks



Machine Learning



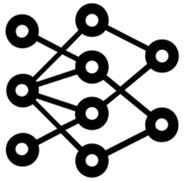


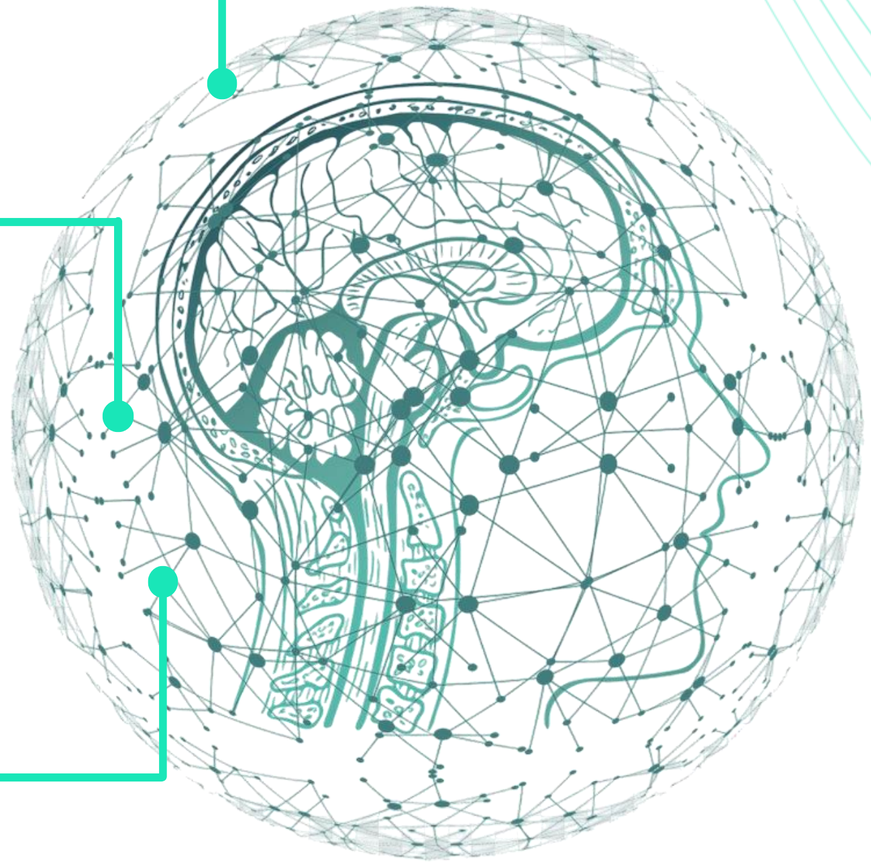
Context

Fields of **AI**

Deep Learning

Neural Networks





Machine Learning

Context

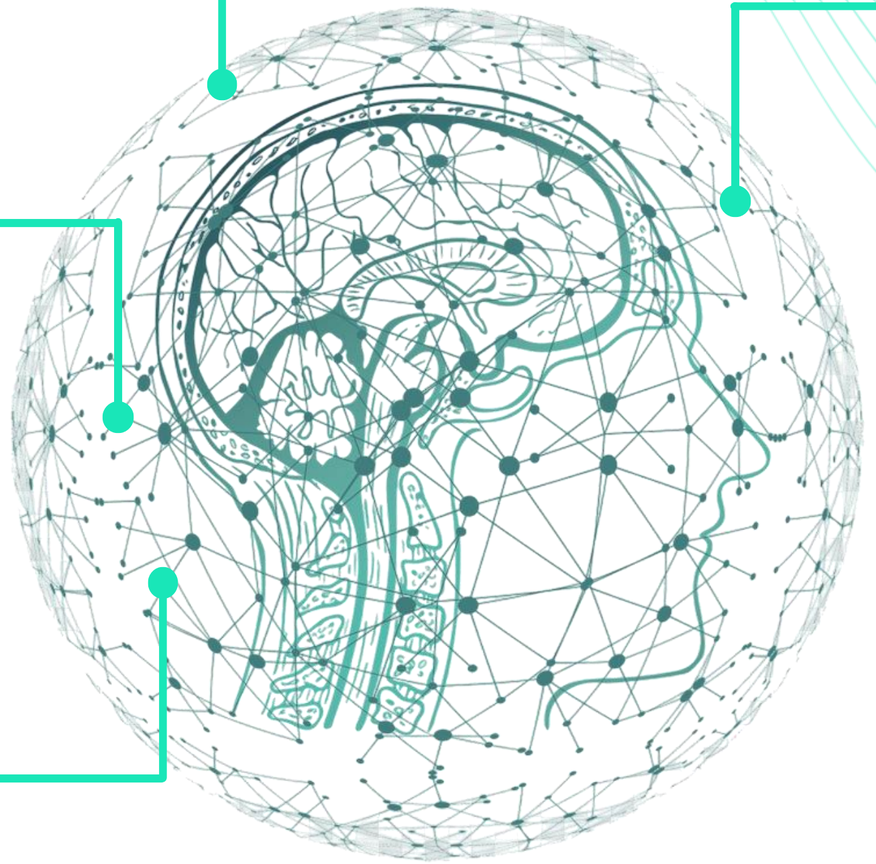
Fields of **AI**

Deep Learning

Natural Language Processing

Neural Networks

Machine Learning



Context

Fields of **AI**

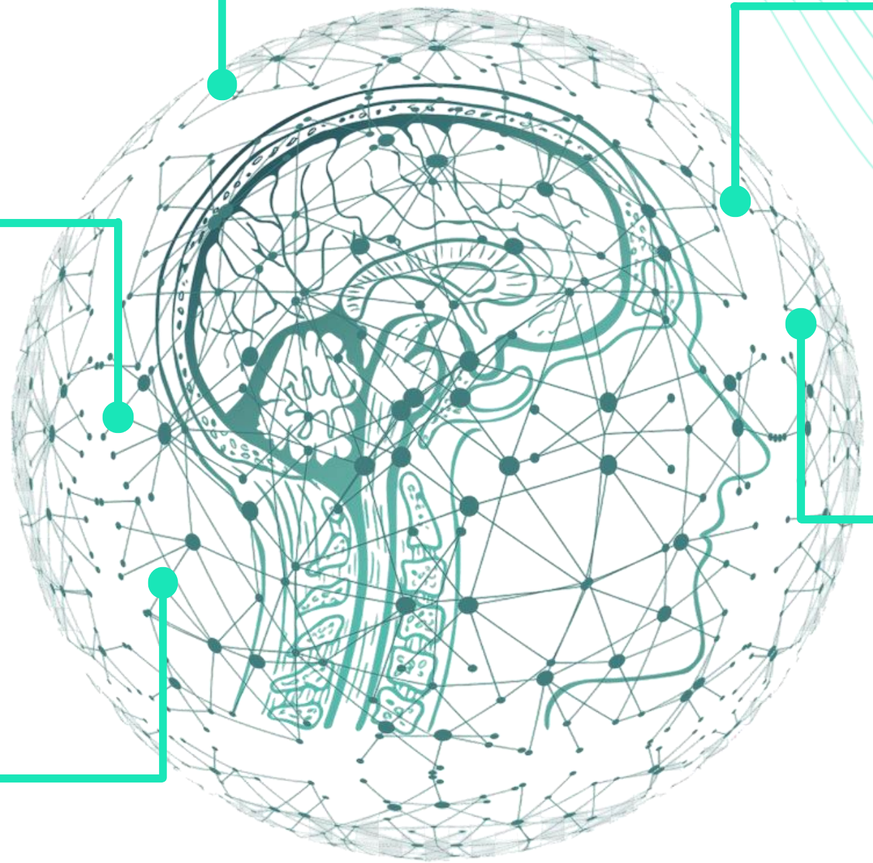
Deep Learning

Natural Language Processing

Neural Networks

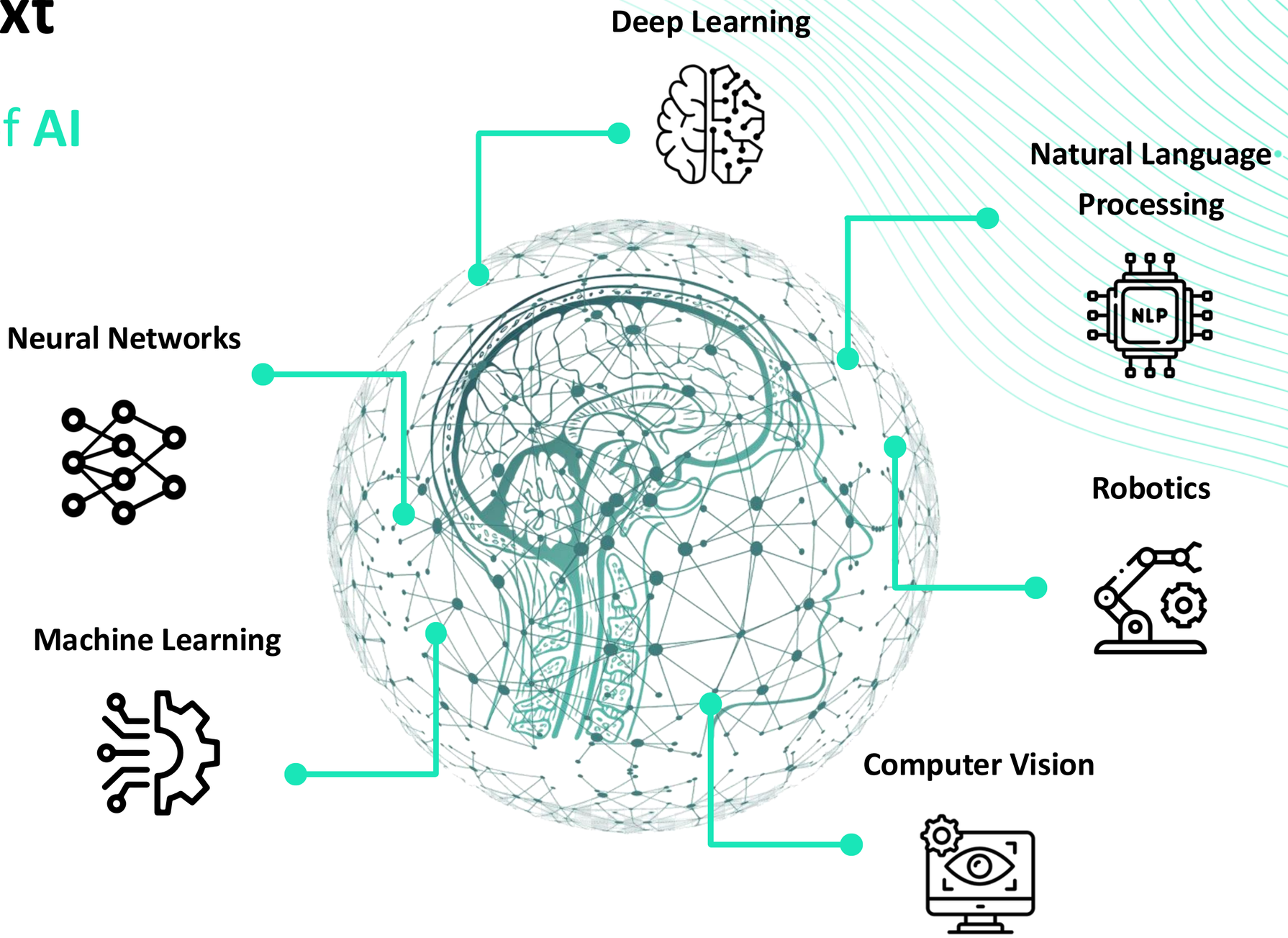
Robotics

Machine Learning



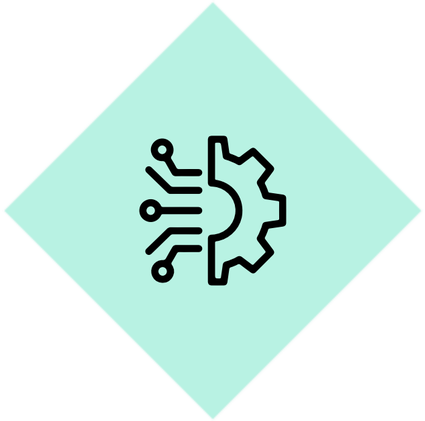
Context

Fields of **AI**



Machine Learning

Fields of **AI**

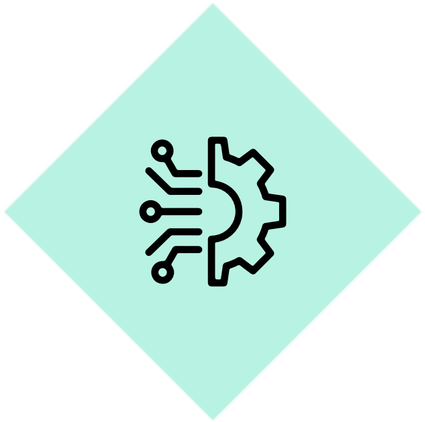


Machine Learning is a field of AI that allows computers to learn patterns from data and make decisions without being explicitly programmed.

ML models improve their performance over time by analyzing more data.

Machine Learning

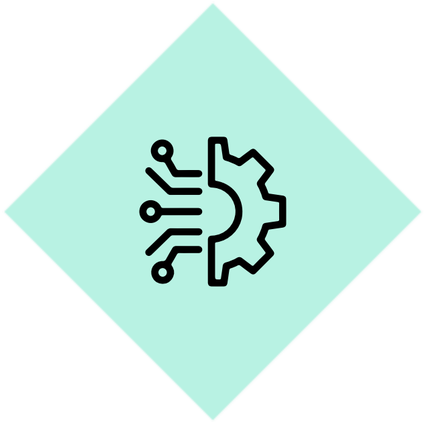
Fields of **AI**



Training **Inference**

Machine Learning

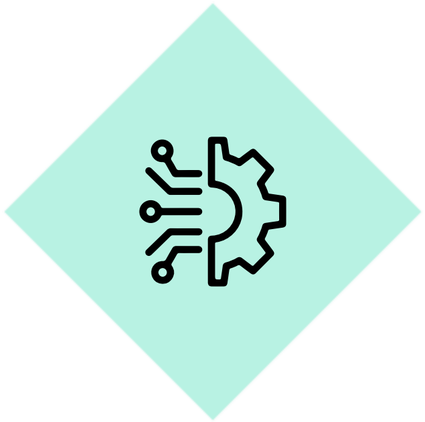
Fields of **AI**



Training	Inference
Develops intelligence by learning from data	Applies learned intelligence to new data

Machine Learning

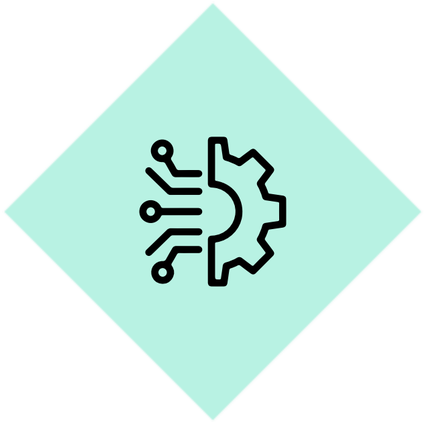
Fields of **AI**



Training	Inference
Develops intelligence by learning from data	Applies learned intelligence to new data
Model is trained using labeled data	Model analyzes and classifies unseen data

Machine Learning

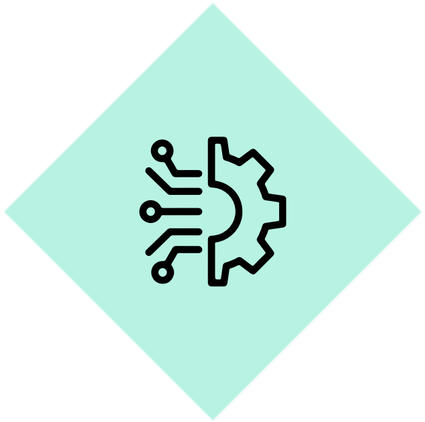
Fields of **AI**



Training	Inference
Develops intelligence by learning from data	Applies learned intelligence to new data
Model is trained using labeled data	Model analyzes and classifies unseen data
Requires large datasets for learning patterns	Uses trained knowledge to make predictions

Machine Learning

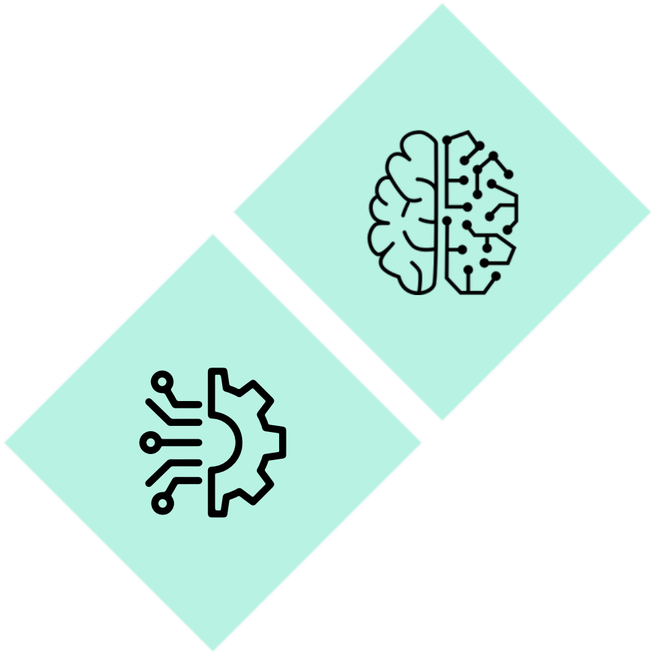
Fields of **AI**



Training	Inference
Develops intelligence by learning from data	Applies learned intelligence to new data
Model is trained using labeled data	Model analyzes and classifies unseen data
Requires large datasets for learning patterns	Uses trained knowledge to make predictions
Requires high-power hardware	Optimized for efficiency and low latency

Deep Learning

Fields of **AI**



Deep Learning is a subset of ML that uses artificial neural networks with multiple layers to process and analyze complex data.

It is particularly effective for handling large datasets and requires more computing power and data than traditional ML methods.

Deep Learning Frameworks

What is **PyTorch**?

Deep Learning Frameworks

What is **PyTorch**?

PyTorch is an open-source DL framework.

Deep Learning Frameworks

What is **PyTorch**?

PyTorch is an open-source DL framework.

PyTorch is widely used for ML and AI applications, especially in research and production.

Deep Learning Frameworks

What is **PyTorch**?

PyTorch is an open-source DL framework.

PyTorch is widely used for ML and AI applications, especially in research and production.

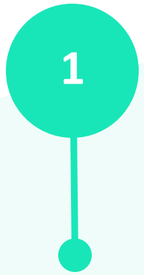
PyTorch provides flexibility, ease of use, and dynamic computation graphs.

Deep Learning Frameworks

How **PyTorch** works?

Deep Learning Frameworks

How **PyTorch** works?



Prepare the data

Deep Learning Frameworks

How **PyTorch** works?

1

Prepare the data

2

Convert data into tensors

Deep Learning Frameworks

How **PyTorch** works?

1. Prepare the data
2. Convert data into tensors
3. Load the data

Deep Learning Frameworks

How **PyTorch** works?

1. Prepare the data
2. Convert data into tensors
3. Load the data
4. Define the model

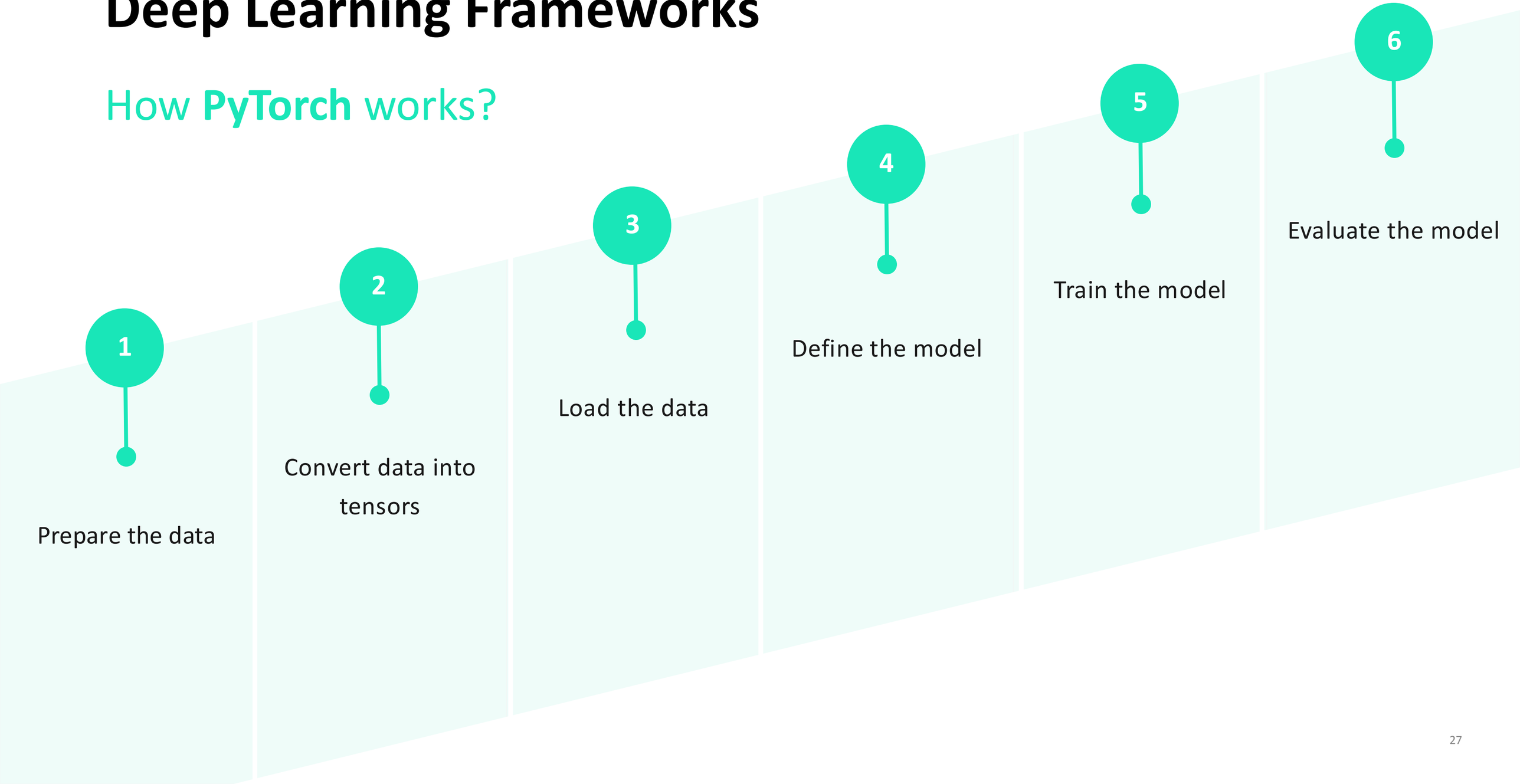
Deep Learning Frameworks

How **PyTorch** works?

1. Prepare the data
2. Convert data into tensors
3. Load the data
4. Define the model
5. Train the model

Deep Learning Frameworks

How **PyTorch** works?



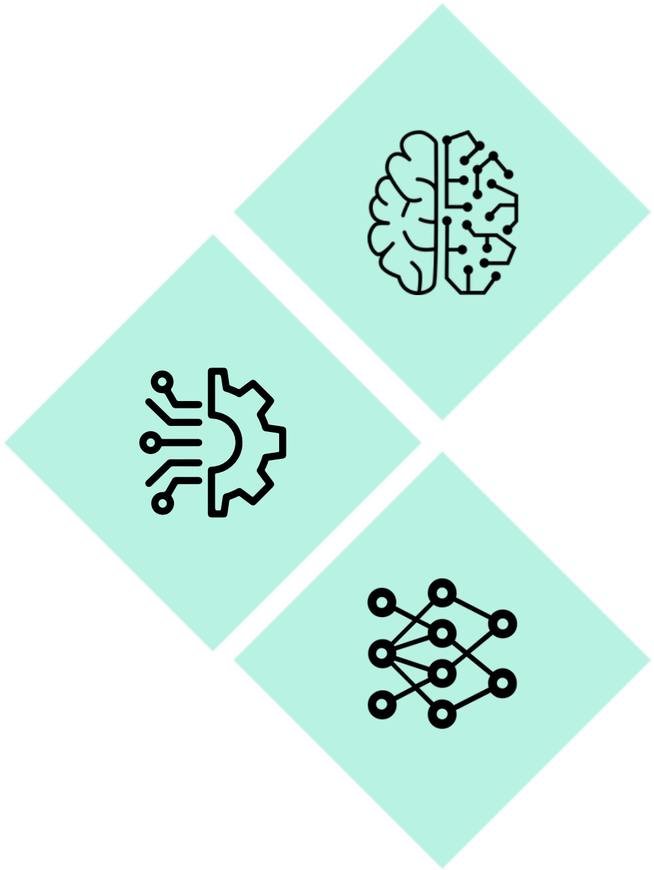
Deep Learning Frameworks

How **PyTorch** works?

1. Prepare the data
2. Convert data into tensors
3. Load the data
4. Define the model
5. Train the model
6. Evaluate the model
7. Inference

Neural Networks

Fields of **AI**

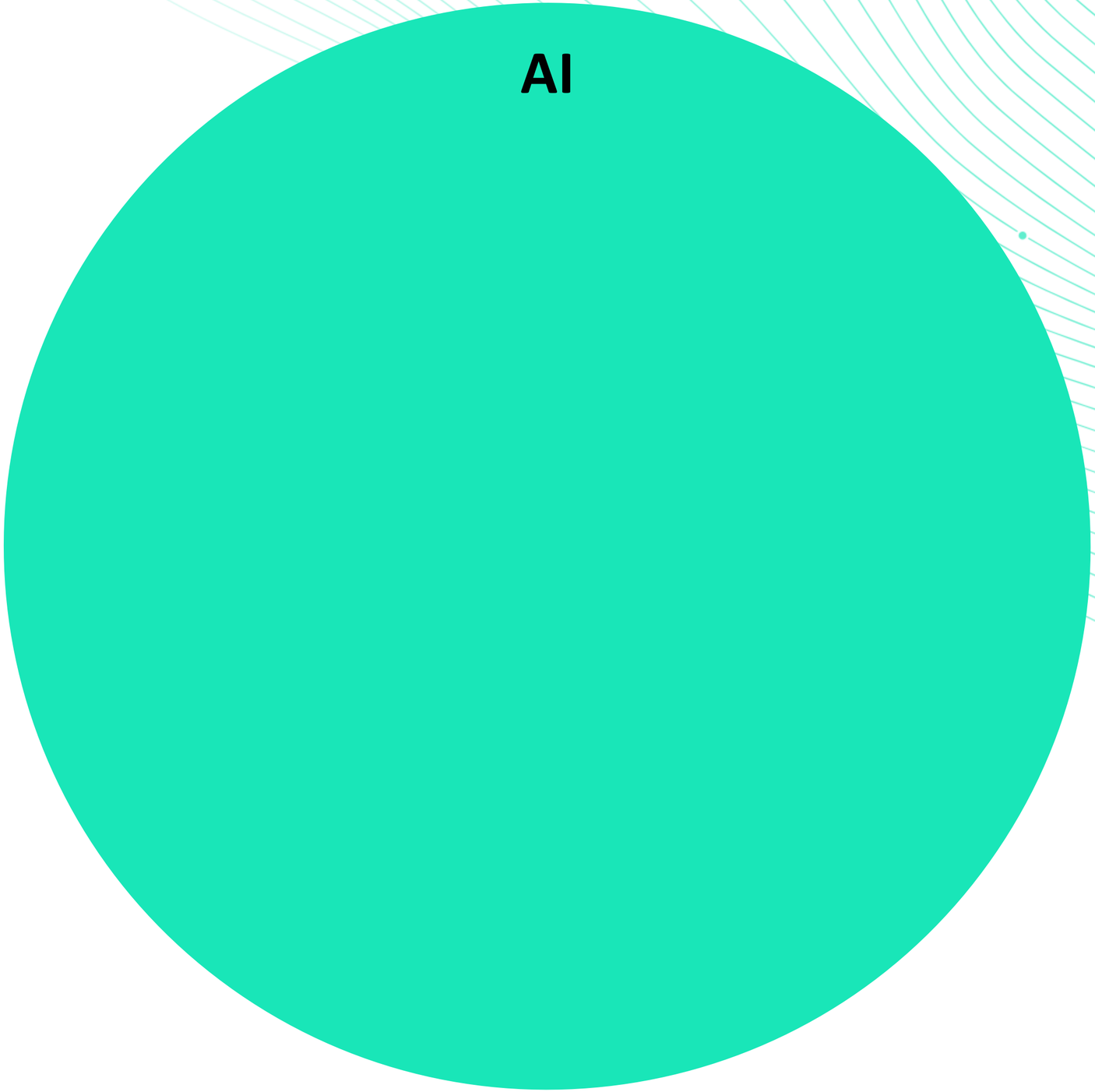


Neural Networks are a type of ML model inspired by the human brain. They consist of interconnected layers of artificial neurons that process and learn from input data.

Simple neural networks are used in ML, while deep neural networks are the foundation of DL.

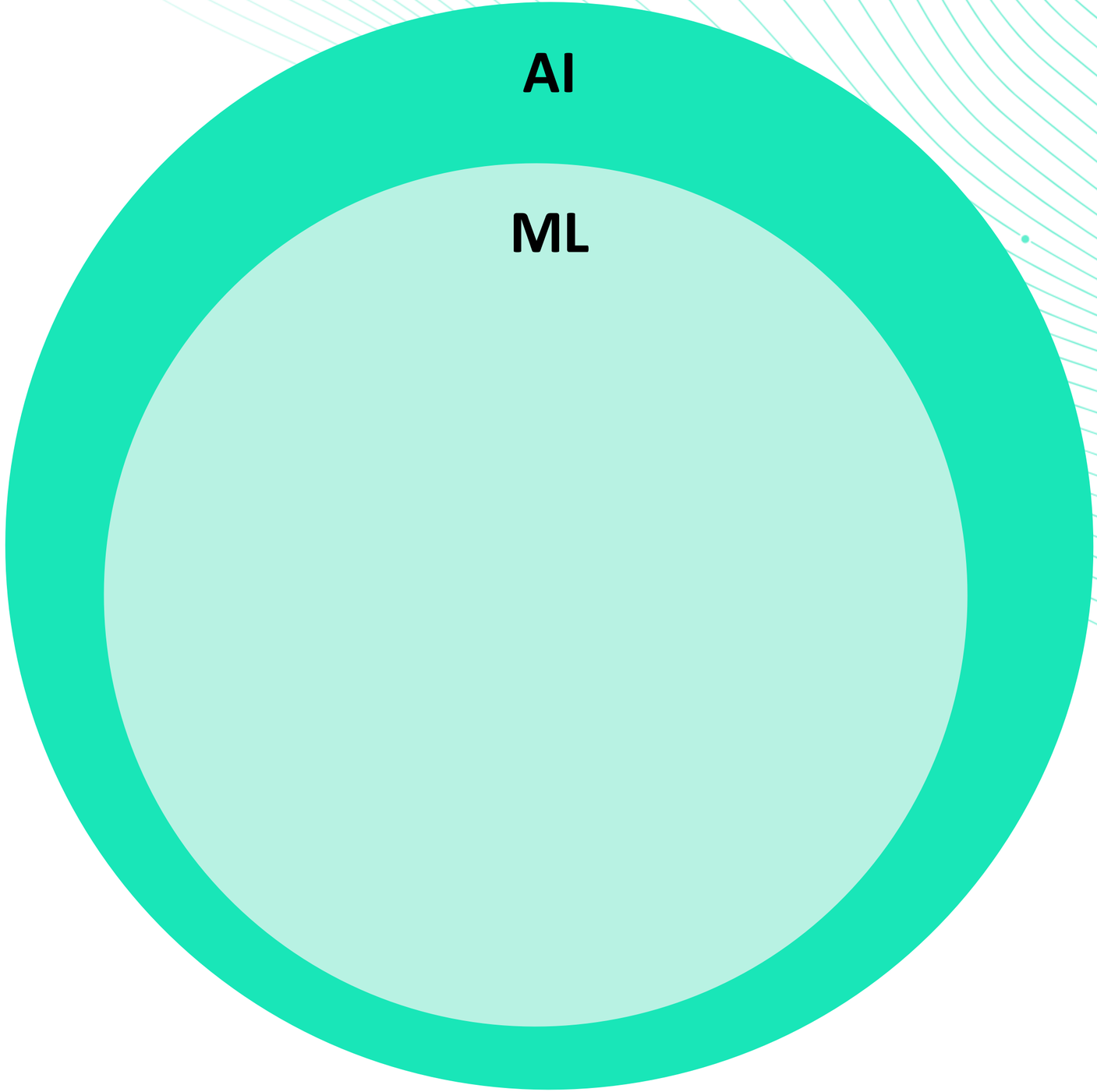
Overview of AI

Fields of **AI**



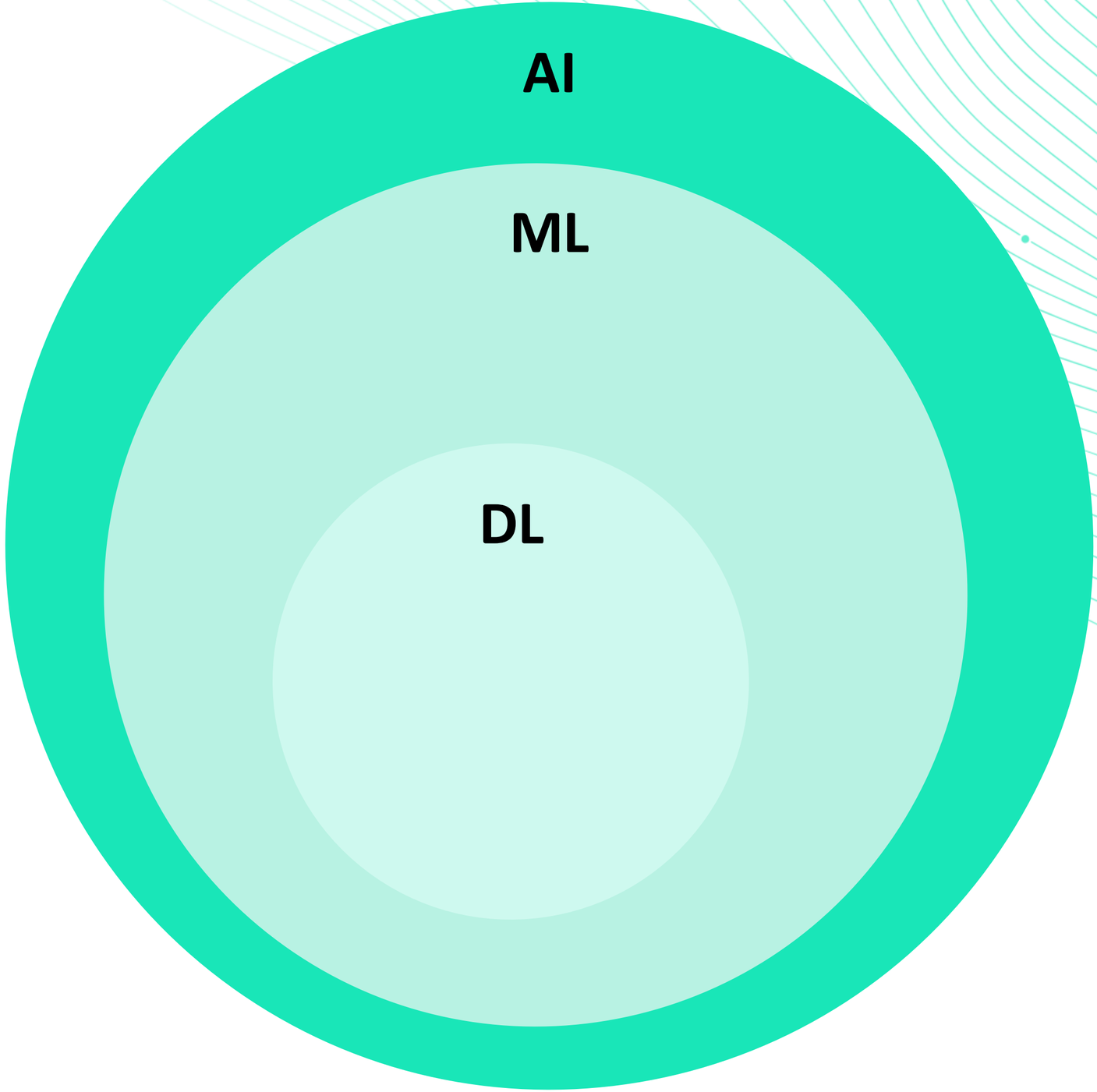
Overview of AI

Fields of AI



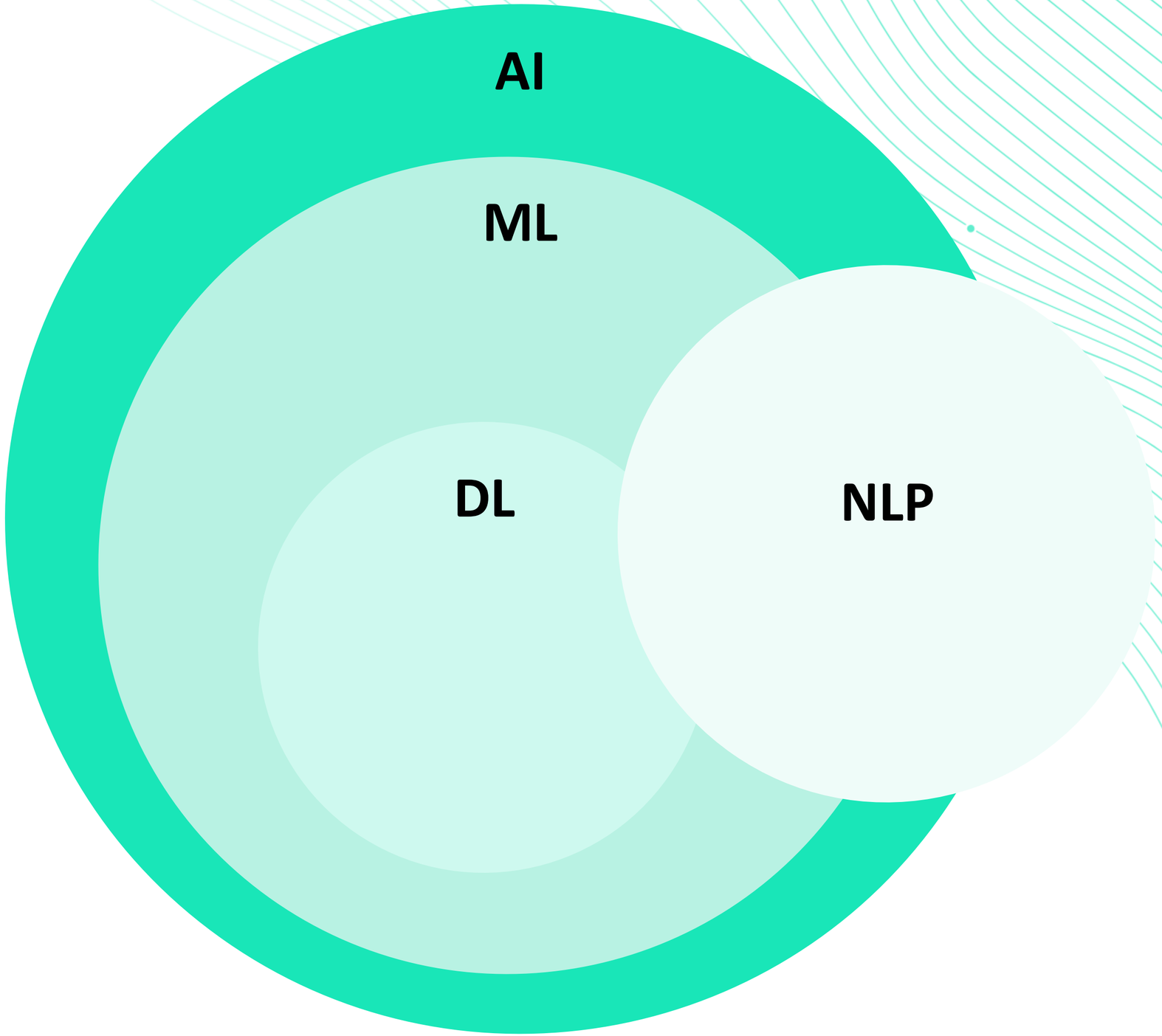
Overview of AI

Fields of AI



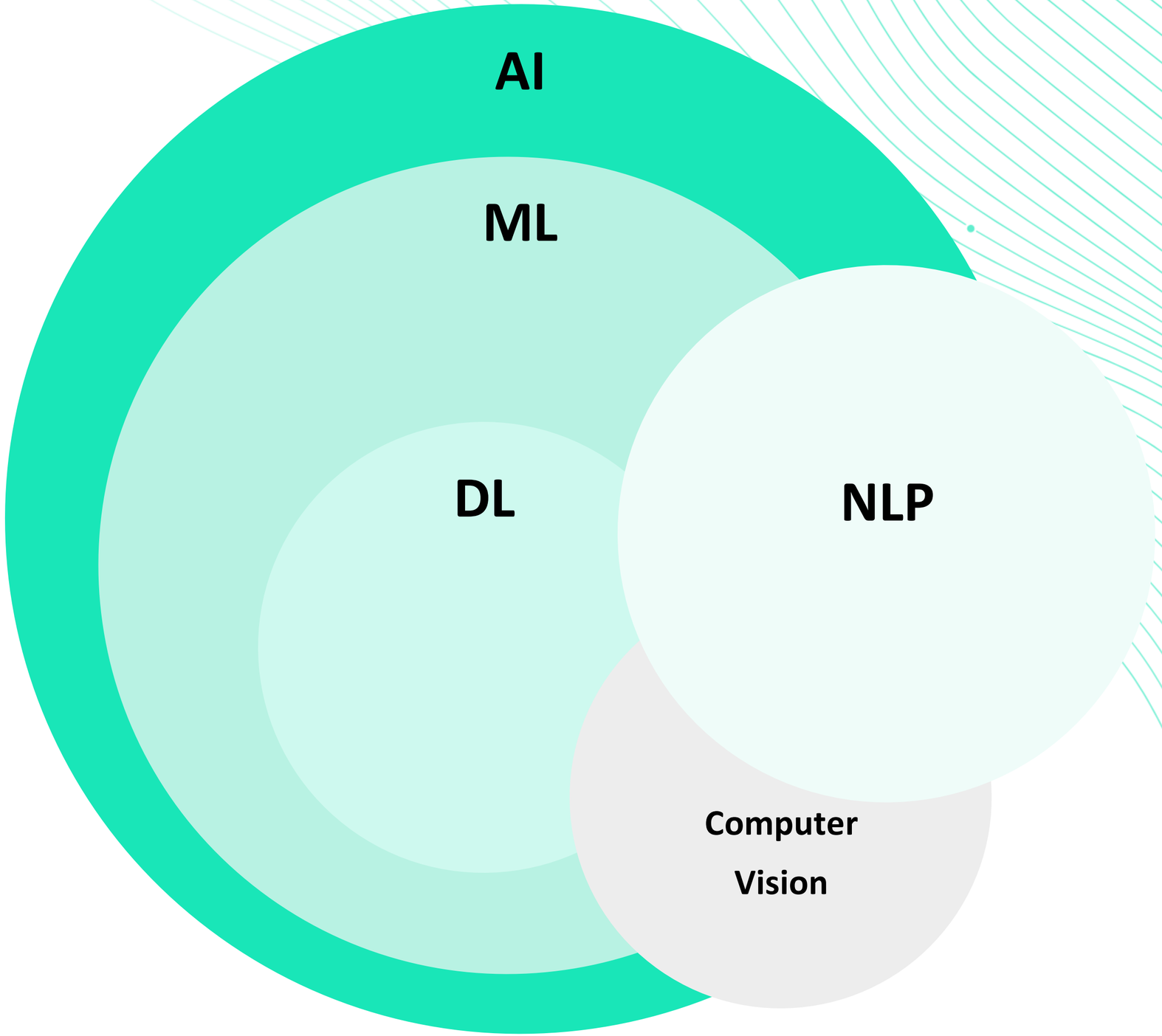
Overview of AI

Fields of AI



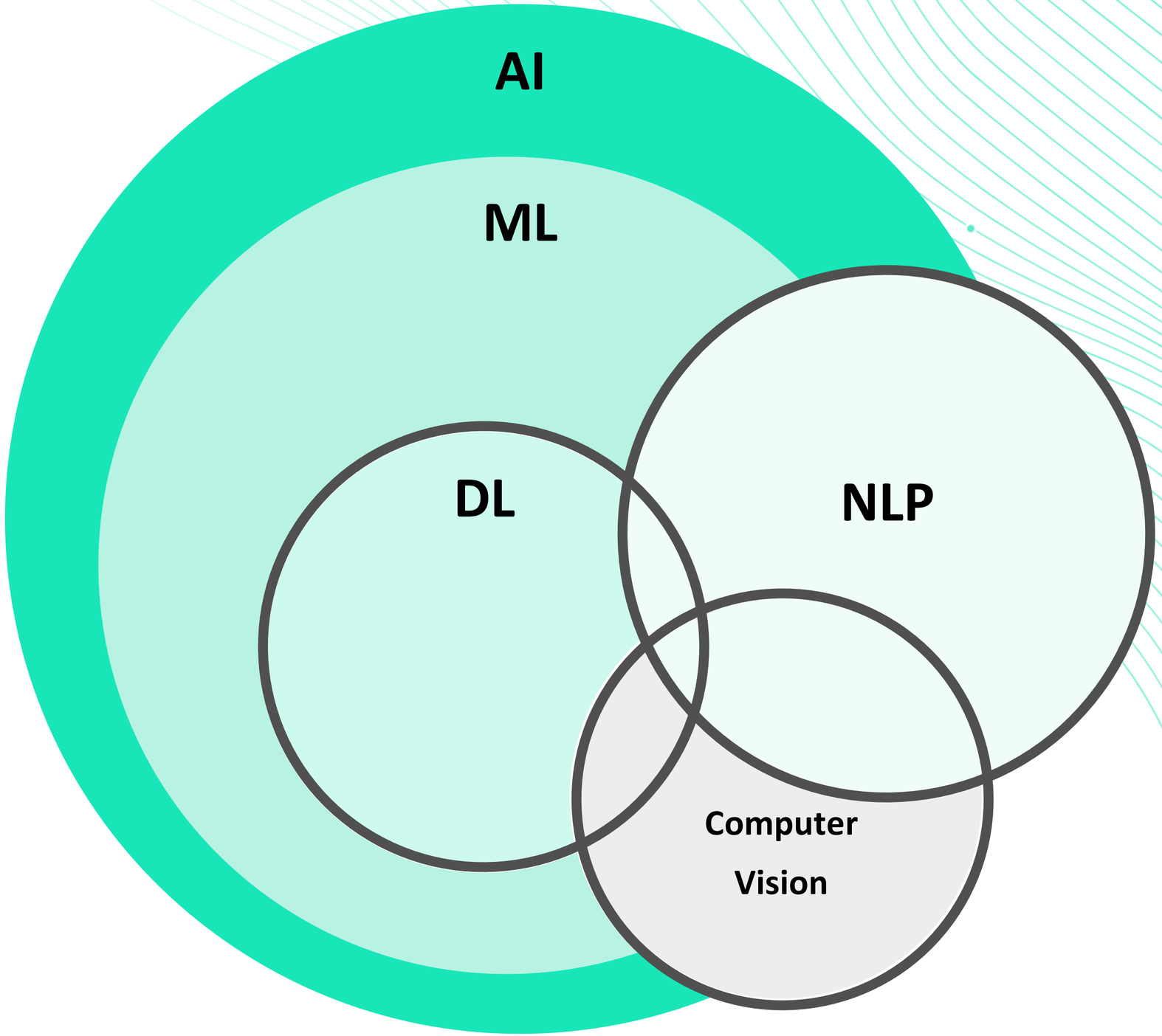
Overview of AI

Fields of AI



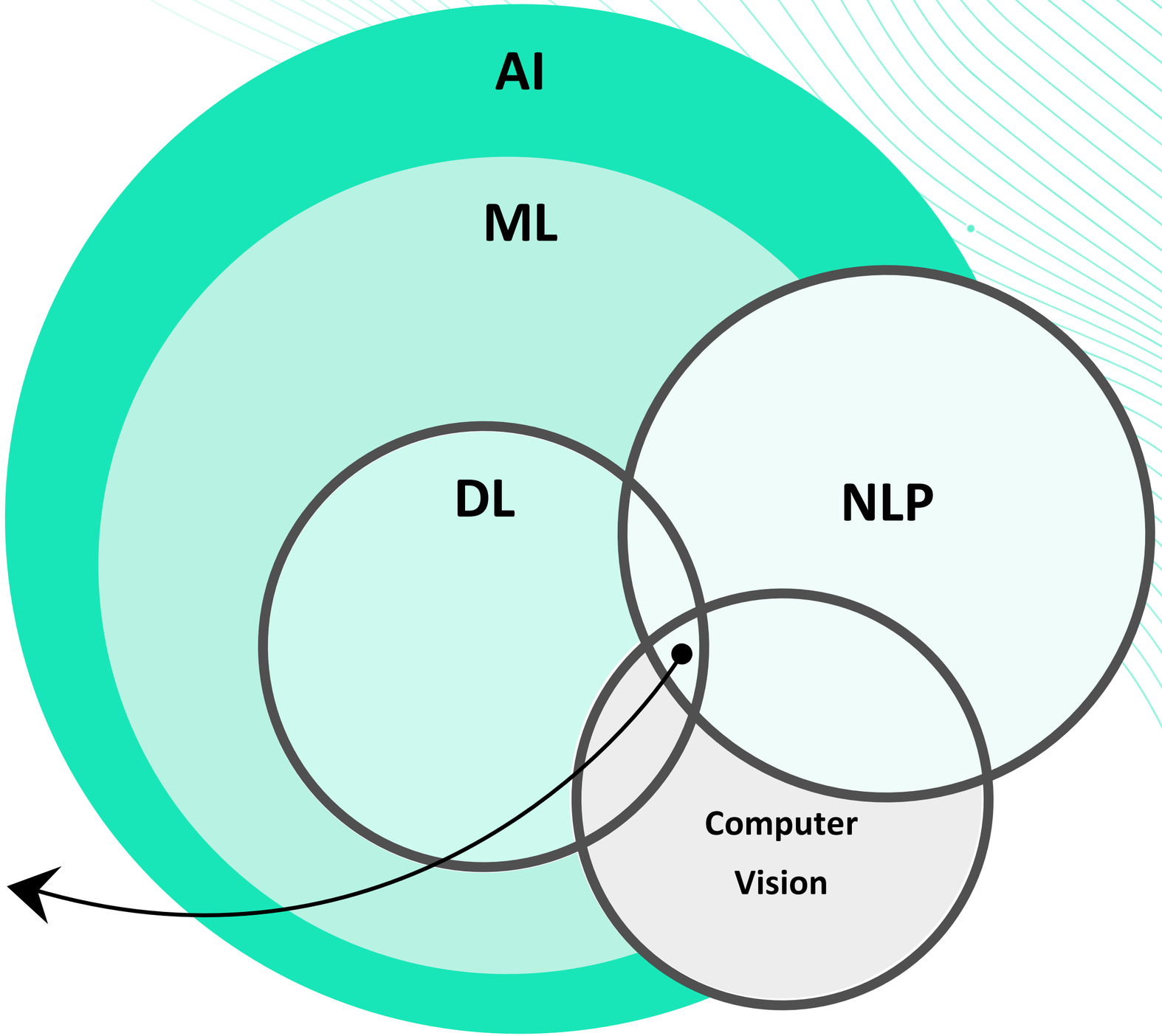
Overview of AI

Fields of AI



Overview of AI

Fields of AI



Generative AI
&
Large Language Models

Large Language Models

What are **L**arge **L**anguage **M**odels?

Large Language Models

What are **L**arge **L**anguage **M**odels?

“Large Language Models are a subset of AI that has been trained on vast quantities of text data to produce human-like responses to dialogue or other natural language inputs.”

Large Language Models

What are **L**arge **L**anguage **M**odels?

“Large Language Models are a subset of AI that has been trained on vast quantities of text data to produce human-like responses to dialogue or other natural language inputs.”

To produce these natural language responses, LLMs make use of DL models, which use multi-layered neural networks to process, analyze, and make predictions with complex data.

Large Language Models

Applications

Large Language Models

Applications

High-Performance Computing

What is **H**igh-**P**erformance **C**omputing?

High-Performance Computing

What is **H**igh-**P**erformance **C**omputing?

“HPC is a technology that uses clusters of powerful processors that work in parallel to process multidimensional datasets and solve complex problems at extremely high speeds.”

AI in ARM Architecture

Why **ARM** for **AI**?

Accelerating AI Everywhere—From Cloud to Edge

AI in ARM Architecture

Why to use **ARM** for **AI** on Deucalion?

AI in ARM Architecture

Why to use **ARM** for **AI** on Deucalion?

Scalable Distributed Training

- Deucalion has multiple ARM-based nodes (1632) that enable parallel DL training, distributing workloads efficiently across these nodes.

AI in ARM Architecture

Why to use **ARM** for **AI** on Deucalion?

Scalable Distributed Training

- Deucalion has multiple ARM-based nodes (1632) that enable parallel DL training, distributing workloads efficiently across these nodes.
- Enables efficient training of large datasets across multiple nodes.

AI in ARM Architecture

Why to use **ARM** for **AI** on Deucalion?

Scalable Distributed Training

- Deucalion has multiple ARM-based nodes (1632) that enable parallel DL training, distributing workloads efficiently across these nodes.
- Enables efficient training of large datasets across multiple nodes.

Less Power Consumption

- ARM architecture is energy-efficient, reducing overall power consumption.

AI in ARM Architecture

Why to use **ARM** for **AI** on Deucalion?

Scalable Distributed Training

- Deucalion has multiple ARM-based nodes (1632) that enable parallel DL training, distributing workloads efficiently across these nodes.
- Enables efficient training of large datasets across multiple nodes.

Less Power Consumption

- ARM architecture is energy-efficient, reducing overall power consumption.
- Deucalion can run large-scale AI workloads while being more sustainable.

AI Workloads on ARM

Demo 1: Training a Neural Network on ARM

AI Workloads on ARM

Demo 2: Inference on ARM

AI Workloads on ARM

Demo 3: ARM vs. x86 - Comparison using Deepseek

Performance Comparison: Inference

ARM vs x86

Question:

“What is the difference between ARM and x86 architectures?”

Performance Comparison: Inference

ARM vs x86

Question:

“What is the difference between ARM and x86 architectures?”

	ARM	x86
Meta-Llama-3-8B	1m37s	1h18m54s
DeepSeek-R1-Distill-Llama-8B	4m44s	2h50m52s

EPICURE

Unlocking European-level HPC Support

Thank you!

pmo-epicure@postit.csc.fi

Follow us

Co-funded by the European Union

EuroHPC Joint Undertaking

This project has received funding from the European High Performance Computing Joint Undertaking under grant agreementNo.101139786. Views and opinions expressed are, however, those of the author(s) only and do not necessarily reflect those of the European Union orEuroHPC Joint Undertaking. Neither the European Union nor the granting authority can be held responsible for them.