



PROGRAMME  
DE RECHERCHE  
INNOVATION  
NUMÉRIQUE

# PEPR IA // Embedded AI

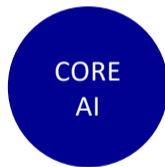
Emergences

Adapting

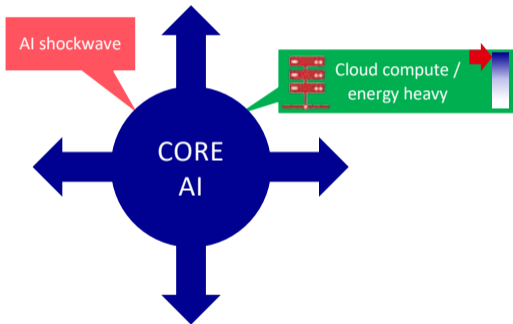
Holigrail

anr<sup>®</sup>  
agence nationale  
de la recherche

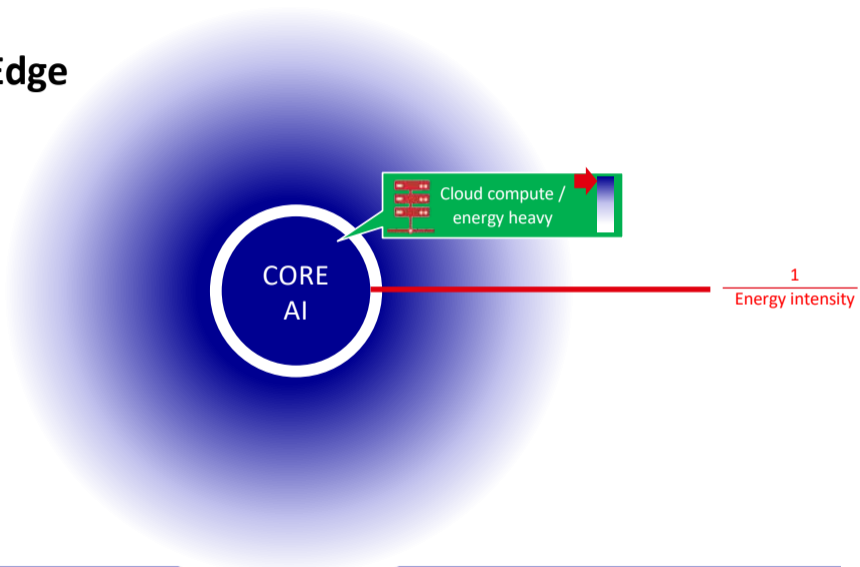
# From Cloud to Edge



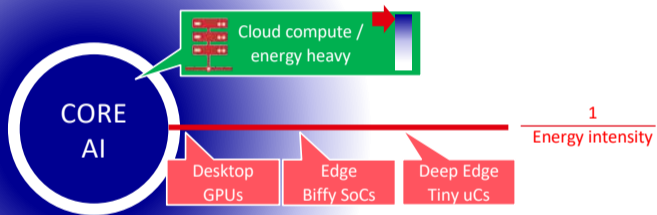
# From Cloud to Edge



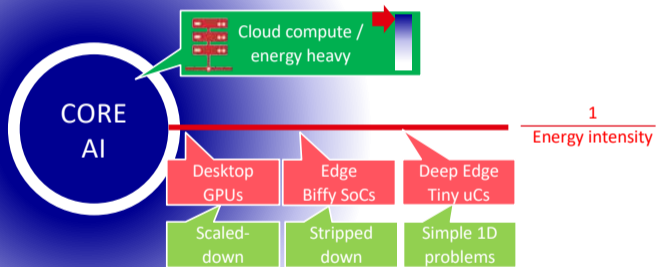
# From Cloud to Edge



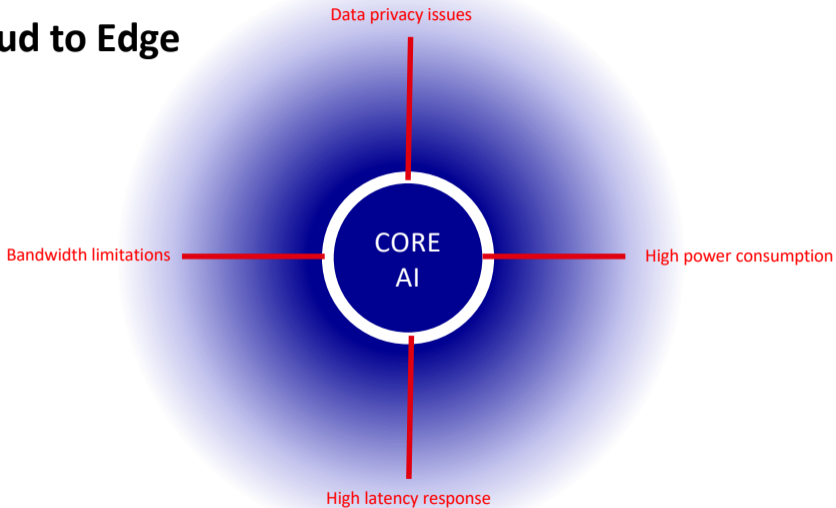
# From Cloud to Edge



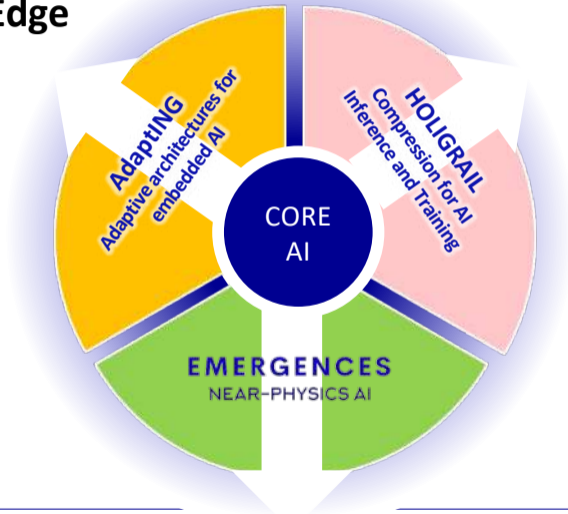
# From Cloud to Edge



# From Cloud to Edge



# From Cloud to Edge







PROGRAMME  
DE RECHERCHE  
INFORMATIQUE  
NUMÉRIQUE

# Emergences

## Near-physics emerging models for embedded AI

Pierre Boulet (CRISTAL)

Julie Grollier (UMPHY)

Fabio Pavanello (INL)

Maxime Pelcat (IETR)

Laurent Perrinet (INT)

Jean-Michel Portal (IN2MP)

Martial Mermillod (LPNC)

Benoit Miramond (LEAT)

**Marina Reyboz (CEA-LIST)**

Sylvain Saighi (IMS)

**Gilles Sassatelli (LIRMM)**

Damien Querlioz (C2N)

Philippe Talatchian (SPINTEC)

Elisa Vianello (CEA-LETI)

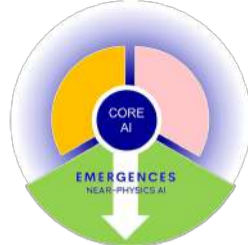


anr<sup>®</sup>  
agence nationale  
de la recherche



# EMERGENCES

## NEAR-PHYSICS AI



## Near-physics emerging models for embedded AI

### Keywords

Energy efficiency  
Embedded AI / Edge AI  
Emerging AI models  
Near-physics AI  
Bio-inspiration

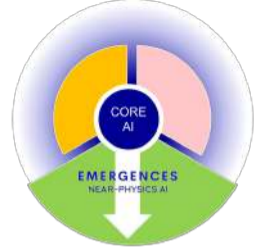
### Key figures

T<sub>O</sub>: September 1<sup>st</sup>, 2023  
Duration: 48 month  
14 Partners  
Nb of PhD: 19  
Nb of Post doc: 13  
TRL: basic research  
Total grant requested: 6.8 M€

# Emergences, near-physics AI

... at the Edge

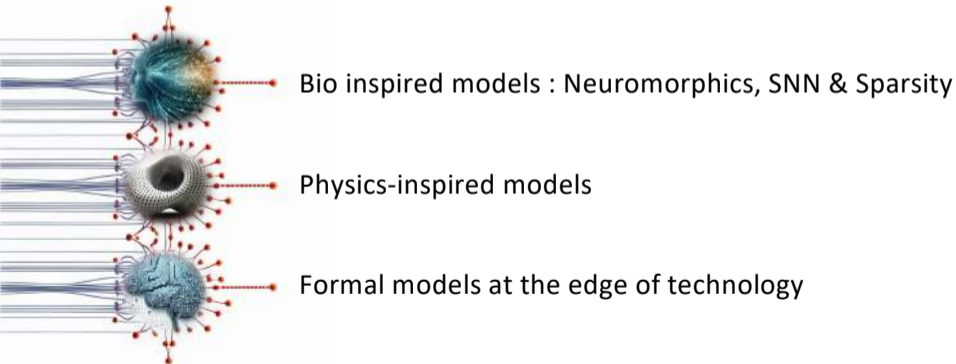
- Exploit the **intrinsic properties of physical devices** for ML
  - Rather than massive linear algebra on energy-hungry digital hardware
- Conventional formal models & training algorithms poorly amenable
  - Emerging models inspired from **physics itself & neurosciences**
  - Alongside **associated training algorithms**
- Shaped & tuned for sustainable AI in sound application domains
  - Environmental monitoring, health



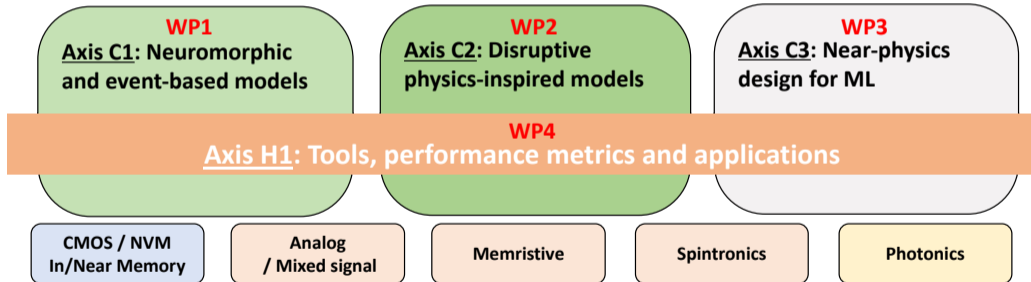
Reduce energy consumption of embedded AI models  
Structure the French landscape of embedded AI

# Which models are we talking about?

3 classes



# Project structure



# WP1

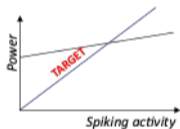
## Event-driven, neuromorphic & sparse models



## WP1 - SNN / Event-driven models

Most sota SNN implementations fail to leverage spiking sparsity energy-wise

→ HW templates that leverage sparsity: NVM & novel techniques

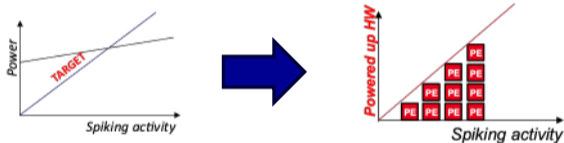


6 PhDs // 4  
co-supervised  
2 post docs

## WP1 - SNN / Event-driven models

Most sota SNN implementations fail to leverage spiking sparsity energy-wise

→ HW templates that leverage sparsity: NVM & novel techniques



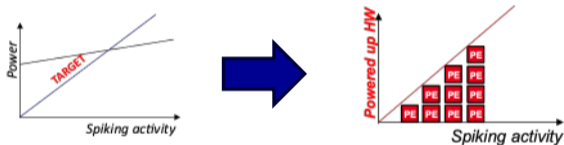
6 PhDs // 4  
co-supervised  
2 post docs



# WP1 - SNN / Event-driven models

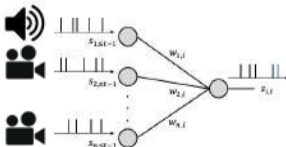
Most sota SNN implementations fail to leverage spiking sparsity energy-wise

→ HW templates that leverage sparsity: NVM & novel techniques



Wide performance gap w.r.t. « formal ML »

→ Exploiting multimodality

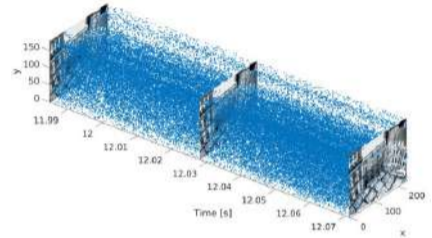


6 PhDs // 4  
 co-supervised  
 2 post docs

# WP1 - SNN / Event-driven models cont'd

Further investigate sparsity..

- Sparsified network topologies
- Event-based sensors for true end-to-end



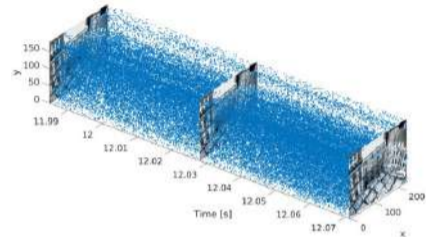
# WP1 - SNN / Event-driven models cont'd

Further investigate sparsity..

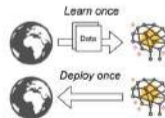
- Sparsified network topologies
- Event-based sensors for true end-to-end

Training is key..

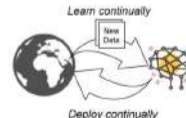
- Novel (online & local) training algorithms
- Incremental learning



Static ML



Adaptive ML



# WP2

## Models inspired from physics

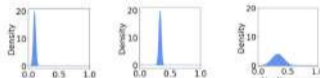


6 PhDs // 5  
co-supervised  
4 post docs

## WP2 – Physics-inspired models

### Stochastic & Bayesian models

→ Exploiting memristive devices properties



6 PhDs // 5  
co-supervised  
4 post docs

## WP2 – Physics-inspired models

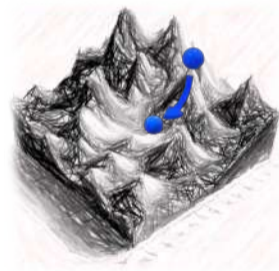
### Stochastic & Bayesian models

→ Exploiting memristive devices properties

### Models having an intrinsic dynamics

→ Mapped onto « tunable » physical memristive hardware (EBM)

$$\frac{ds}{dt} = -\frac{\partial E}{\partial s}$$



6 PhDs // 5  
co-supervised  
4 post docs

# WP2 – Physics-inspired models

Stochastic & Bayesian models

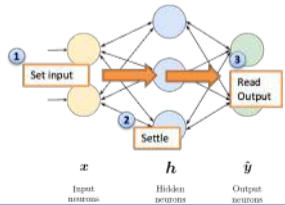
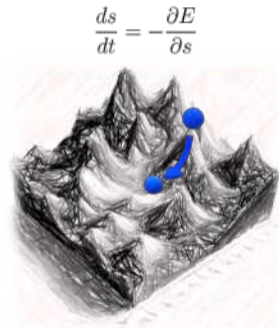
→ Exploiting memristive devices properties

Models having an intrinsic dynamics

→ Mapped onto « tunable » physical memristive hardware (EBM)

Above all, local training algorithms

→ Equilibrium Propagation, NeuralODEs, Forward-Forward...



**Inference**

6 PhDs // 5  
 co-supervised  
 4 post docs

# WP2 – Physics-inspired models

Stochastic & Bayesian models

→ Exploiting memristive devices properties

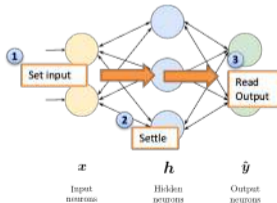
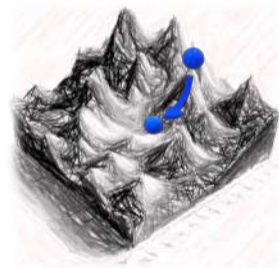
Models having an intrinsic dynamics

→ Mapped onto « tunable » physical memristive hardware (EBM)

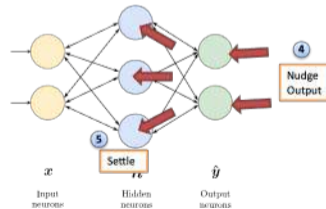
Above all, local training algorithms

→ Equilibrium Propagation, NeuralODEs, Forward-Forward...

$$\frac{ds}{dt} = -\frac{\partial E}{\partial s}$$



**Inference**



**Training**

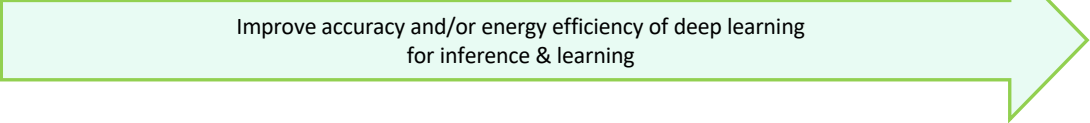


# WP3

## Formal models at the edge of technology



## WP3 – Formal models at the edge of technology

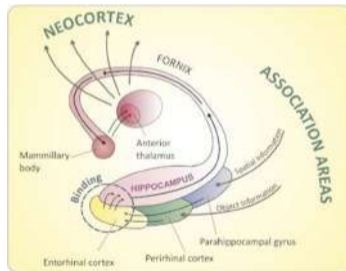


Improve accuracy and/or energy efficiency of deep learning  
for inference & learning

## WP3 – Formal models at the edge of technology

Improve accuracy and/or energy efficiency of deep learning  
for inference & learning

- Embedded multimodal continual learning (link with WP1)
  - Brain inspired model (link with neurosciences)
  - Improve accuracy / robustness
  - Co-design between algorithms & hardware

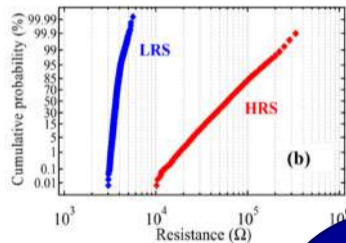


5 PhDs  
All co-supervised  
3 post docs

## WP3 – Formal models at the edge of technology

Improve accuracy and/or energy efficiency of deep learning  
for inference & learning

- Embedded multimodal continual learning (link with WP1)
  - Improve accuracy / robustness
  - Brain inspired model (link with neurosciences)
  - Co-design between algorithms & hardware
- Non Volatile Memory & IMC / NMC technologies
  - Complex linear algebra functions and attentional mechanisms thanks to NVM-based IMC
  - On-chip training with NVM weight storage (variability-aware)

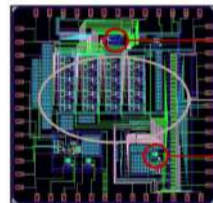


5 PhDs  
All co-supervised  
3 post docs

## WP3 – Formal models at the edge of technology

Improve accuracy and/or energy efficiency of deep learning  
for inference & learning

- Embedded multimodal continual learning (link with WP1)
  - Improve accuracy / robustness
  - Brain inspired model (link with neurosciences)
  - Co-design between algorithms & hardware
- Non Volatile Memory & IMC / NMC technologies
  - Complex linear algebra functions and attentional mechanisms thanks to NVM-based IMC
  - On-chip training with NVM weight storage (variability-aware)
- Hybrid photonic/electronic schemes
  - For ultra-large ANNs
  - Possibly incorporating NVM devices



Photodiodes  
where 3-D  
waveguide optical spatial filters  
will be printed on top



Control circuitry to  
test all kind of things



HfOx memristors

5 PhDs  
All co-supervised  
3 post docs

# WP4

## Tools, performance metrics & applications



## WP4 – Tools, performance metrics & applications

Design a **strategy** to : Evaluate the applicability of proposed contributions to other models  
Perform comparative analysis using representative criteria

- Benchmarking
  - Datasets
  - KPIs: perf, power, tolerance, sustainability etc.
  - Benchmarking protocols and domain-specific recommendations
- Tooling
  - Common tools at most (simulators, porting of training algorithms, bridges where possible)
  - Tools for DSE / AutoML / scalability analysis
- Applications
  - Health
  - Monitoring (environment)
  - Wearables

**3-years Platform engineer**  
2 PhDs including 1  
co-supervision  
4 post docs

# Specific project organisation

## Workshop every 6 months

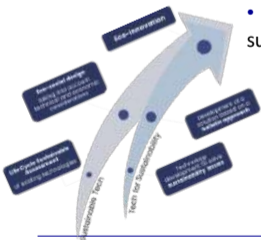
- To assure collaborative & cooperative strategies
- To assure a global multi-disciplinary approach
  - Strong link with neurosciences
  - Interest for societal impact: invitation of philosophers/sociologists
  - Interest for the new legislation : invitation of jurists



# Specific project organisation

## Workshop every 6 months

- To assure collaborative & cooperative strategies
- To assure a global multi-disciplinary approach
  - Strong link with neurosciences
  - Interest for societal impact: invitation of philosophers/sociologists
  - Interest for the new legislation : invitation of jurists
  - Development of sustainable approaches :  
sustainable tech (dev specific KPIs) + tech for sustainability (specific use cases)



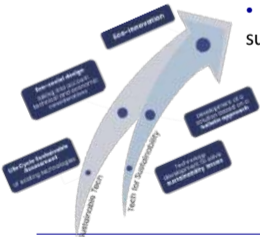
# Specific project organisation

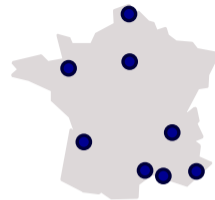
## Workshop every 6 months

- To assure collaborative & cooperative strategies
- To assure a global multi-disciplinary approach
  - Strong link with neurosciences
  - Interest for societal impact: invitation of philosophers/sociologists
  - Interest for the new legislation : invitation of jurists
  - Development of sustainable approaches :  
sustainable tech (dev specific KPIs) + tech for sustainability (specific use cases)

### Advisory board

- Michel Paindavoine
- Christian Gamrat
  - David Bol
  - Ian O'Connor





# Roles and achievements of partners

## LEAT

SNN  
 Multimodality  
 & HW imp.  
 WP1

## IMS

SNN  
 NVMs  
 WP1

## CRISTAL

SNN  
 Tools  
 WP1, WP4

## LIRMM

Digital & mixed signal  
 AI4CAD  
 WP1, WP2, Management

B. Miramond



S. Saighi

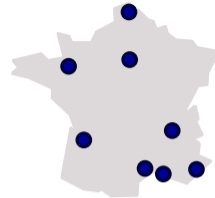


P. Boulet



G. Sassatelli





# Roles and achievements of partners

## SPINTEC

Stochastic & prob.  
MRAM

WP2

## C2N

Stochastic & EBM models

WP2

## UMPHY

Emerg. models  
Bio-inspired train.

WP2

## INL

Photonic dev.

WP2

P. Talatchian



D. Querlioz

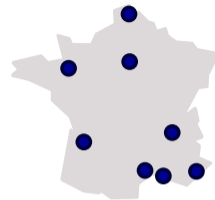


J. Grollier



F. Pavanello





# Roles and achievements of partners

## CEA-LETI

Stochastic & prob. models  
NVMs

WP2, WP3

## CEA-LIST

Emerging models & NVMs  
HW implementations (IMC)

WP2, WP3, WP4, Management

## IM2NP

NVMs & hardware imp.

WP3

## Neurosciences partners

## INT & LPNC

E. Vianello



M. Reyboz



A. Molnos



J.M. Portal



L. Perrinet



M. Mermillod



## Project outcomes



- Very tiny ML
- Very low power
- Autonomous systems  
(energy & training)

## Project outcomes



- Very tiny ML
- Very low power
- Autonomous systems  
(energy & training)



Example : smart  
autonomous sensors  
for environment  
monitoring

## Project outcomes



- Very tiny ML
- Very low power
- Autonomous systems (energy & training)



Example :  
autonomous  
for environ-  
monitoring

Could help lay down the foundations for AI HW as an alternative to current mainstream GPUs/ASICs for some workloads

- Without forgetting scalability

Provide guidance towards  
a choice of model, a training algorithm & a given hardware solution on a per use-case basis





# THANKS!



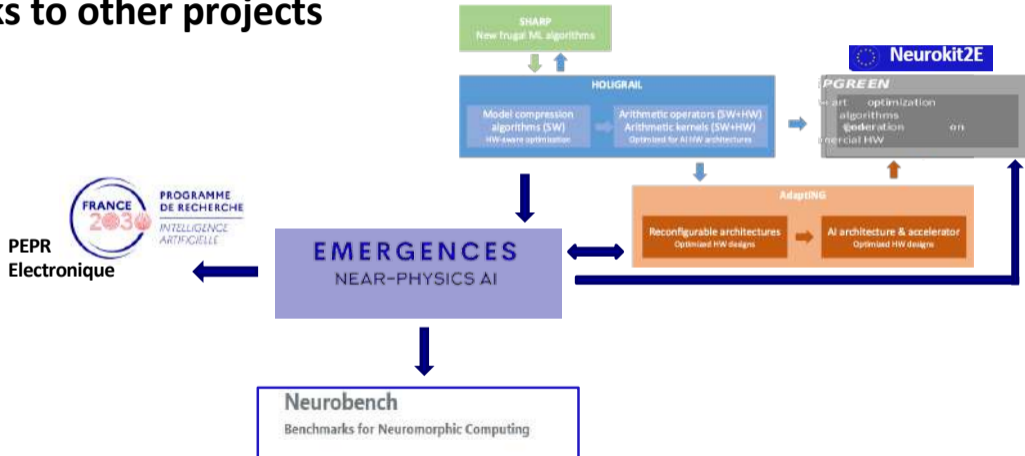
# IA : NOTRE AMBITION POUR LA FRANCE

MARS 2024

## Recommandation n° 5

*Faire de la France un pionnier de l'IA pour la planète en renforçant la transparence environnementale, la recherche dans des modèles à faible impact, et l'utilisation de l'IA au service des transitions énergétique et environnementales.*

# Links to other projects



# Dissemination & Exploitation

## Dissemination activities :

- Publication activities in ML & embedded venues
- Proactive dissemination in networks (Hipeac, GDRs) & workshops

## Rather « basic research project » still :

- Dissemination to the industry too
- Leveraging existing partners' industrial collaborations through tools & applications
- Link with higher TRL France2030 & EU projects having SMEs in the loop : DeepGreen, Neurokit

## Links to other projects

