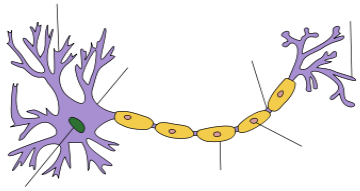
UMONS
Université de Mons



18-19 Mai 2024

Certificat d'université en

Intelligence Artificielle

HackIA24

Atelier d'Intelligence Artificielle (I-ISIA-202)

Edge AI System for smart cities

Sidi Ahmed Mahmoudi

https://hackia.eu/

Contact: sidi.mahmoudi@umons.ac.be

UMONS University of Mons

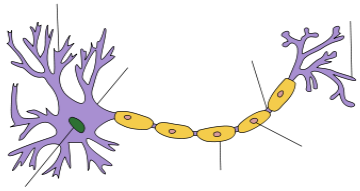
POLYTECH MONS

Faculté des Sciences

numediart
INSTITUT DE RECHERCHE POUR LES TECHNOLOGIES CREATIVES DE L'UMONS

inforTech
UMONS RESEARCH INSTITUTE FOR INFORMATION TECHNOLOGY AND COMPUTER SCIENCE

UMONS
Université de Mons



18-19 Mai 2024

Certificat d'université en

Intelligence Artificielle

Let us Go

Sidi Ahmed Mahmoudi

Contact: sidi.mahmoudi@umons.ac.be

UMONS
University of Mons

POLYTECH MONS

Faculté des Sciences

numediart
INSTITUT DE RECHERCHE POUR LES TECHNOLOGIES CREATIVES DE L'UMONS

inforTech
UMONS RESEARCH INSTITUTE FOR INFORMATION TECHNOLOGY AND COMPUTER SCIENCE

PLAN

I. Introduction & Programme du Workshop

II. Objectif du workshop

III. Phase 1 : développement et entraînement des modèles

IV. Phase 2 : portage de solution sur ressource Edge : Jetson Xavier

V. Phase 3 : optimisation (compression) et explicabilité de modèles

Conclusion

PLAN

I. Introduction & Programme du Workshop

II. Objectif du workshop

III. Phase 1 : développement et entraînement des modèles

IV. Phase 2 : portage de solution sur ressource Edge : Jetson Xavier

V. Phase 3 : optimisation (compression) et explicabilité de modèles

Conclusion

Programme du Workshop

18 Mai 2024

08h00 à 09h00	Accueil des participants
09h00 à 10h00	Présentation du challenge. Pr. Sidi Mahmoudi et pause-café
10h00 à 13h00	Session de travail N° 01
13h00 à 14h00	Lunch
14h00 à 16h45	Session de travail N° 02
16h45 à 17h00	Pause-café
17h00 à 20h00	Session de travail N° 03
20h00 à 21h30	Dîner
21h30 à 23h00	Activité sociale « à définir »

Programme du Workshop

19 Mai 2024

07h00 à 08h30	Petit déjeuner
08h30 à 10h45	Session de travail N° 04
10h45 à 11h00	Pause-café
11h00 à 13h00	Session de travail N° 05
13h00 à 14h00	Lunch Sandwich
14h00 à 17h00	Préparation des démos et pitchs
17h00 à 17h30	Pause-café
17h30 à 18h30	Présentations devant jury
18h30 à 19h30	Remise des prix, cérémonie de clôture et cocktail

Coaching

Coachs Techniques

Pr. Sidi Ahmed Mahmoudi
UMONS

Ir. Jean-Sébastien Lerat
UMONS

Ir. Mohamed Benkedadra
UMONS

Ir. Maxime Gloesener
UMONS

Membres de Jury

Membres du Jury |

Pr. Thierry Dutoit
UMONS

Pr. Pierre Manneback
UMONS

Pr. Souhaib Ben Taieb
UMONS

Pr. Stéphane Dupont
UMONS

Olivier Verscheure
EPFL

Driss Lahem

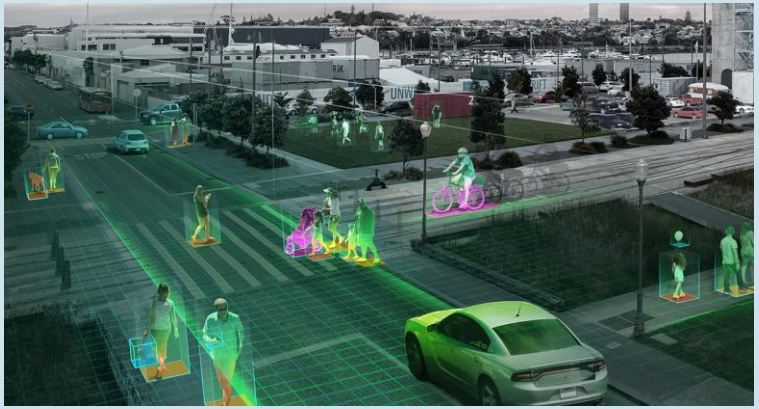
Introduction

- Mise en œuvre et exploitation des connaissances acquises lors des défis IA
- Déploiement de modèles Deep Learning dans des applications concrètes

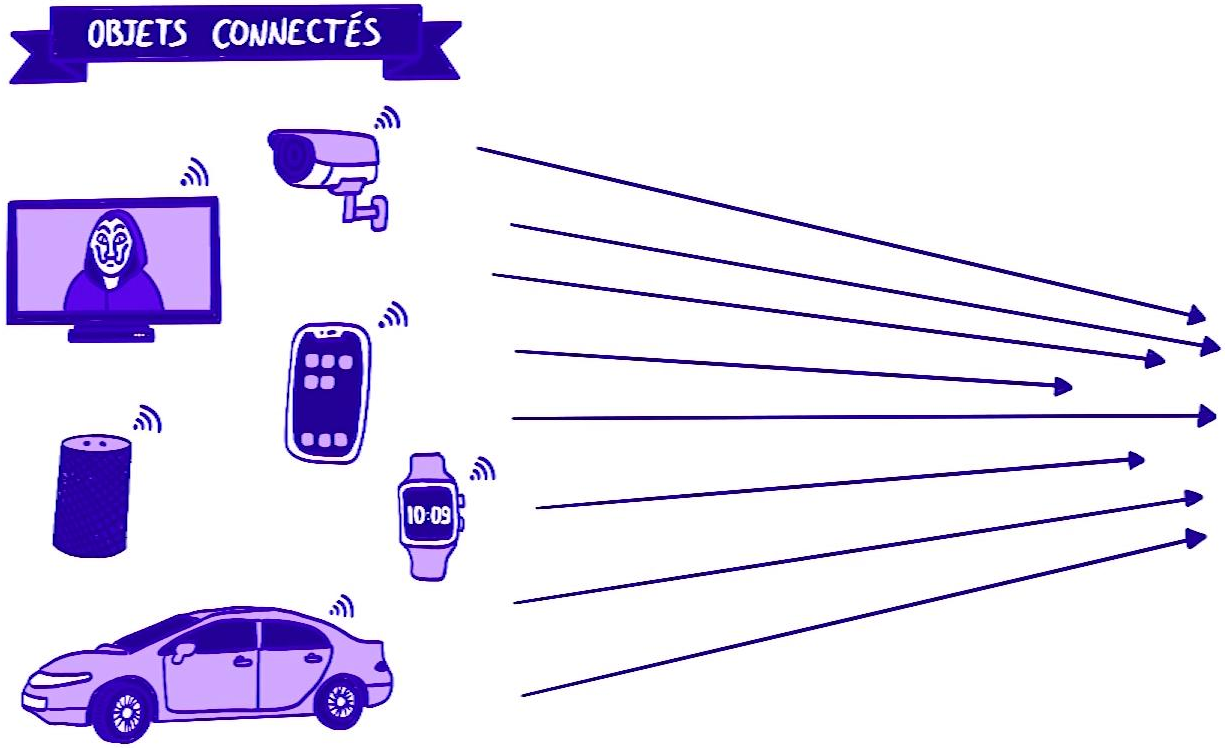
« **modules pour villes intelligentes** »

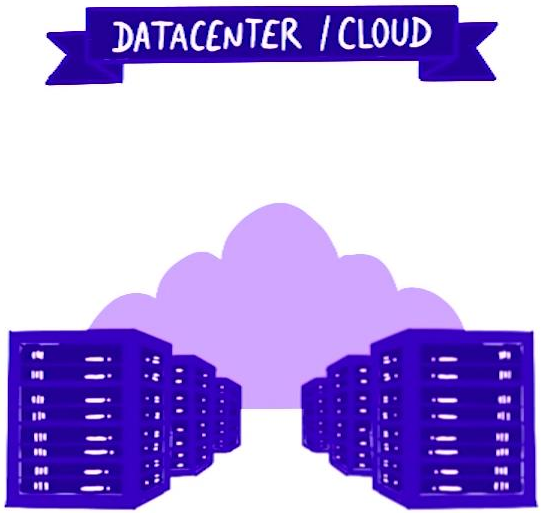
- Application des modèles sur des séquences vidéos au lieu des images

- Déploiement de solutions Deep Learning sur ressources Edge AI



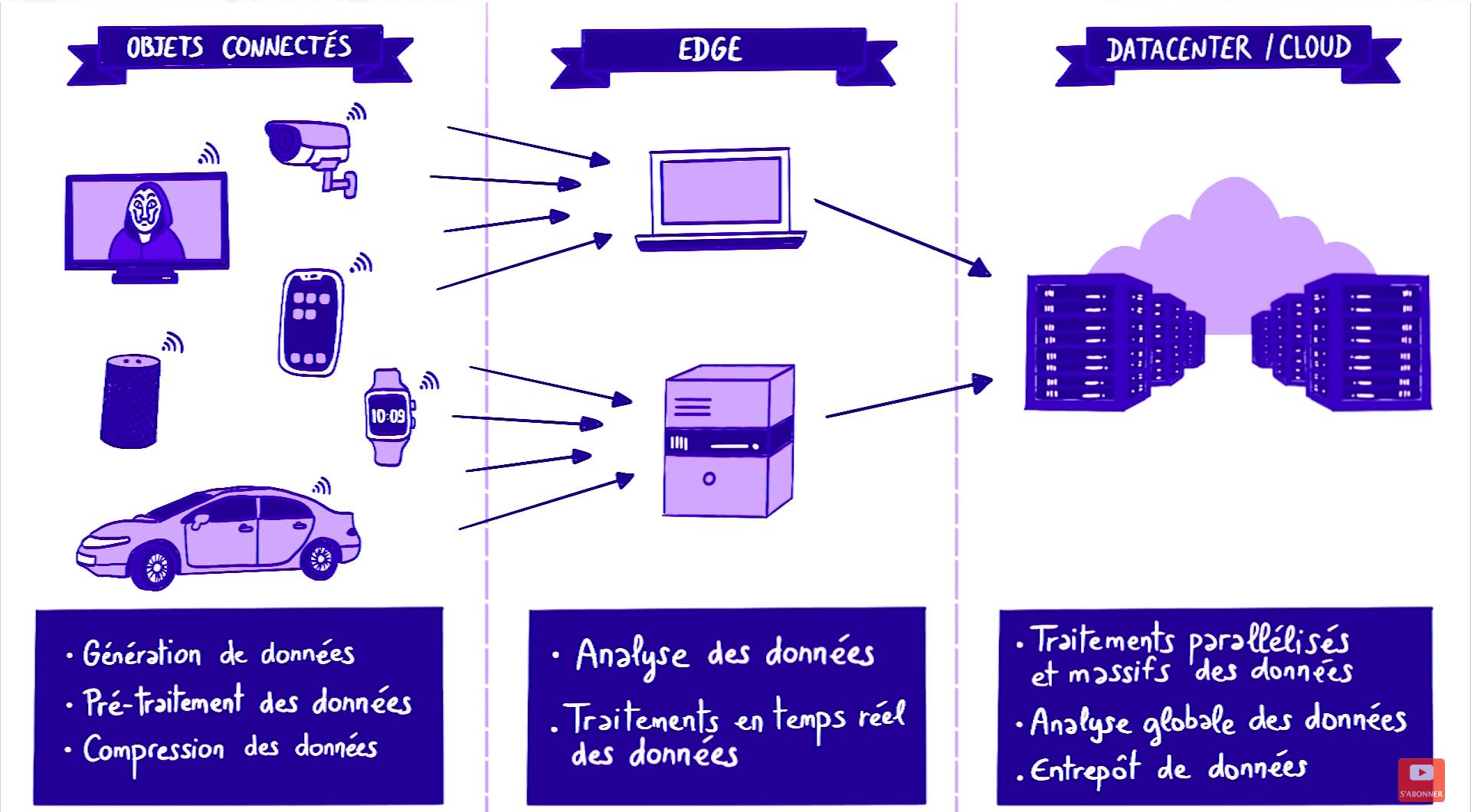
Que signifie Edge Computing ?





- Saturation de la bande passante
- Augmentation de la latence de traitement des données

Que signifie Edge Computing ?



PLAN

I. Introduction & Programme du Workshop

II. Objectif du workshop

III. Phase 1 : développement et entraînement des modèles

IV. Phase 2 : portage de solution sur ressource Edge : Jetson Xavier

V. Phase 3 : optimisation (compression) et explicabilité de modèles

Conclusion

Objectif du Workshop

- Développement de réseaux de neurones profonds pour classifier des images, localiser des objets sur images, reconnaître visages, etc.
- Intégrer les modèles dans un système Edge IA appliqué à des vidéos capturées en temps réel « **module pour villes intelligentes** »
- Portage de la solution sur la ressource Edge IA « **Jetson Xavier** »
- Compromis entre **précision, temps de calcul, espace mémoire et explicabilité**

Objectif du Workshop

- **Projet réalisable en quatre (04) phases :**

1. **Développement et entraînement des modèles Deep Learning ;**
2. **Portage sur ressource Edge : Jetson Xavier ;**
3. **Optimisation/compression de modèles ;**
4. **Explicabilité « XAI » du modèle de classification.**

PLAN

I. Introduction & Programme du Workshop

II. Objectif du workshop

III. Phase 1 : développement et entraînement des modèles

IV. Phase 2 : portage de solution sur ressource Edge : Jetson Xavier

V. Phase 3 : optimisation (compression) et explicabilité de modèles

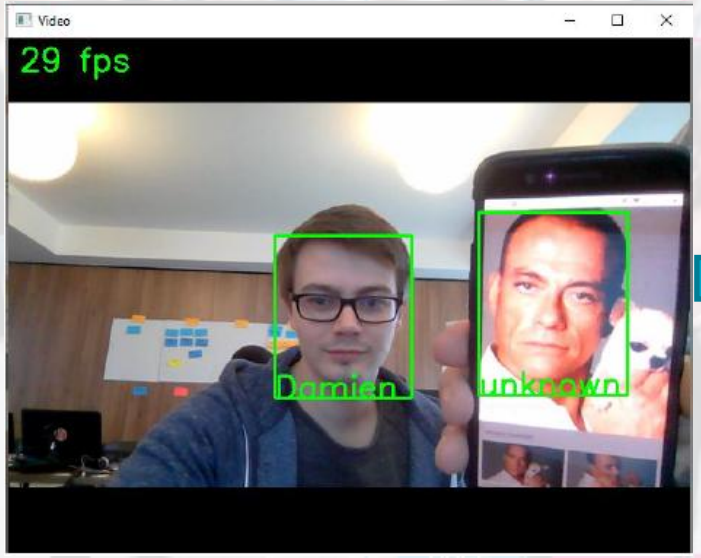
Conclusion

Objectif du Workshop

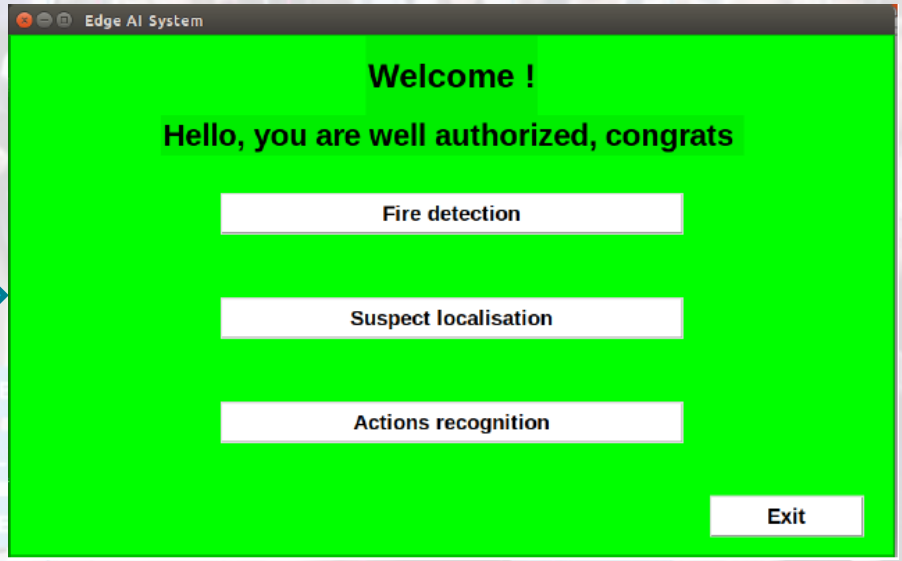
Groupe 1, 2 & 3: Edge AI for Smart Cities

- Développer et entraîner vos modèles sur ressources locales ou cloud
- **Modèle 1 :** classification de feu
- **Modèle 2 :** détection de feu et/ou autres objets (voitures, personnes, etc.)

Face recognition



Edge AI module for smart cities



PLAN

I. Introduction & Programme du Workshop

II. Objectif du workshop

III. Phase 1 : développement et entraînement des modèles

IV. Phase 2 : portage de solution sur ressource Edge : Jetson Xavier

V. Phase 3 : optimisation (compression) et explicabilité de modèles (facultatif)

Conclusion

Portage sur ressource Edge : Jetson Non

Code de démarrage : dossier « HackIA23_Input »

PLAN

I. Introduction & Programme du Workshop

II. Objectif du workshop

III. Phase 1 : développement et entraînement des modèles

IV. Phase 2 : portage de solution sur ressource Edge : Jetson Xavier

V. Phase 3 : optimisation (compression) et explicabilité de modèles

Conclusion

Edge AI for Smart Cities

- **Optimiser et accélérer** l'algorithme : temps d'inférence des modèles
- Privilégier l'utilisation des réseaux de neurones de **plus petites tailles**
- **Compresser** les modèles développées: pruning, quantification, distillation de connaissances, etc.

  ⟹ Vous pouvez combiner les méthodes de compression
- Explicabilité « XAI »: à l'aide du framework **Pytorch**
- **Objectif :** meilleur compromis entre précision, temps de calcul espace mémoire requis.

PLAN

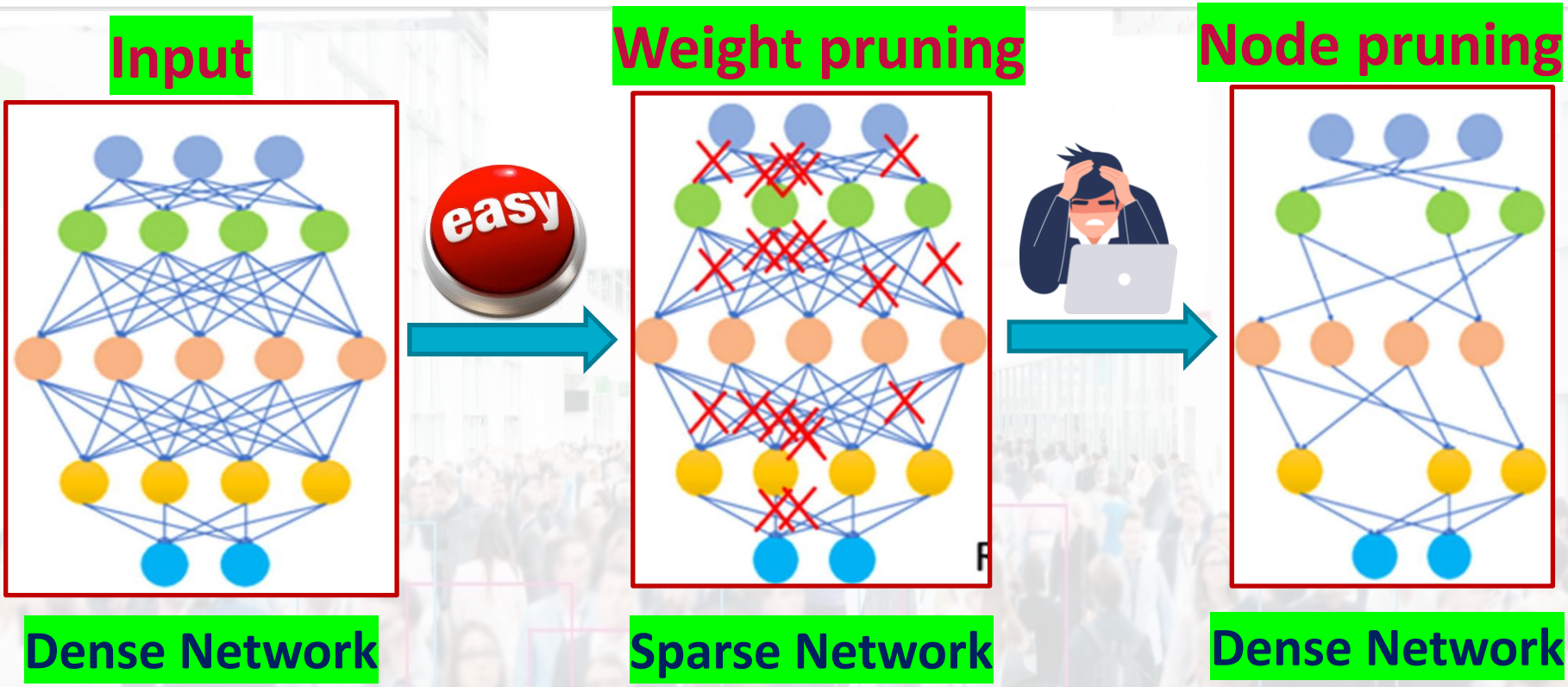
Main approaches of Edge AI in Deep Learning

a. **Pruning**

b. **Quantization**

c. **Knowledge Distillation**

Related work : Pruning

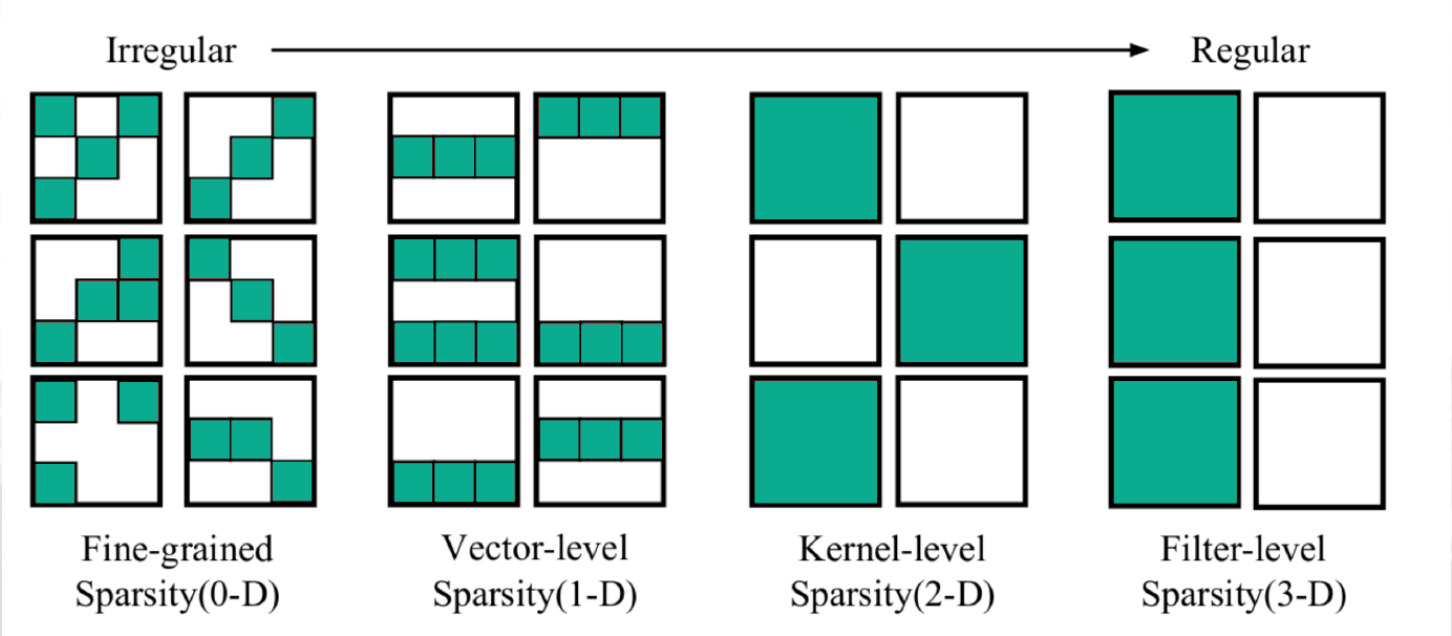


Several elements to analyze before pruning :

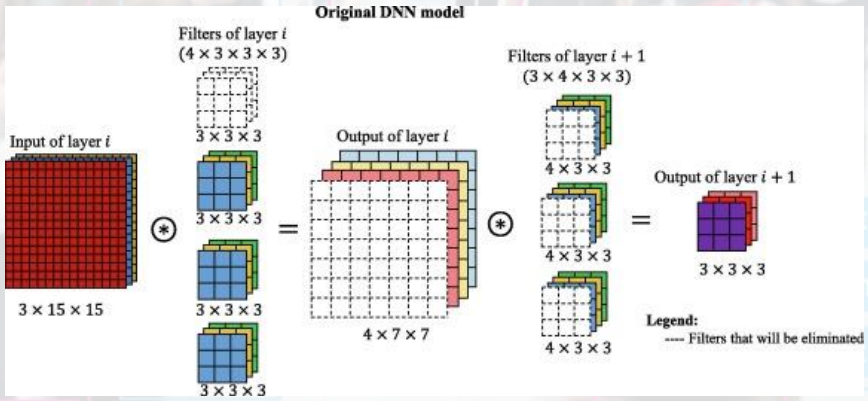
- **Pruning methods**
- **Pruning-zone application**
- **Neron pruning or/and weight pruning**
- **Pruning scheduling**
- **Pruning impact on memory consumption**
- **Pruning impact on model size**
- **Pruning impact on computation time**

Related work : Pruning

- **Pruning methods**

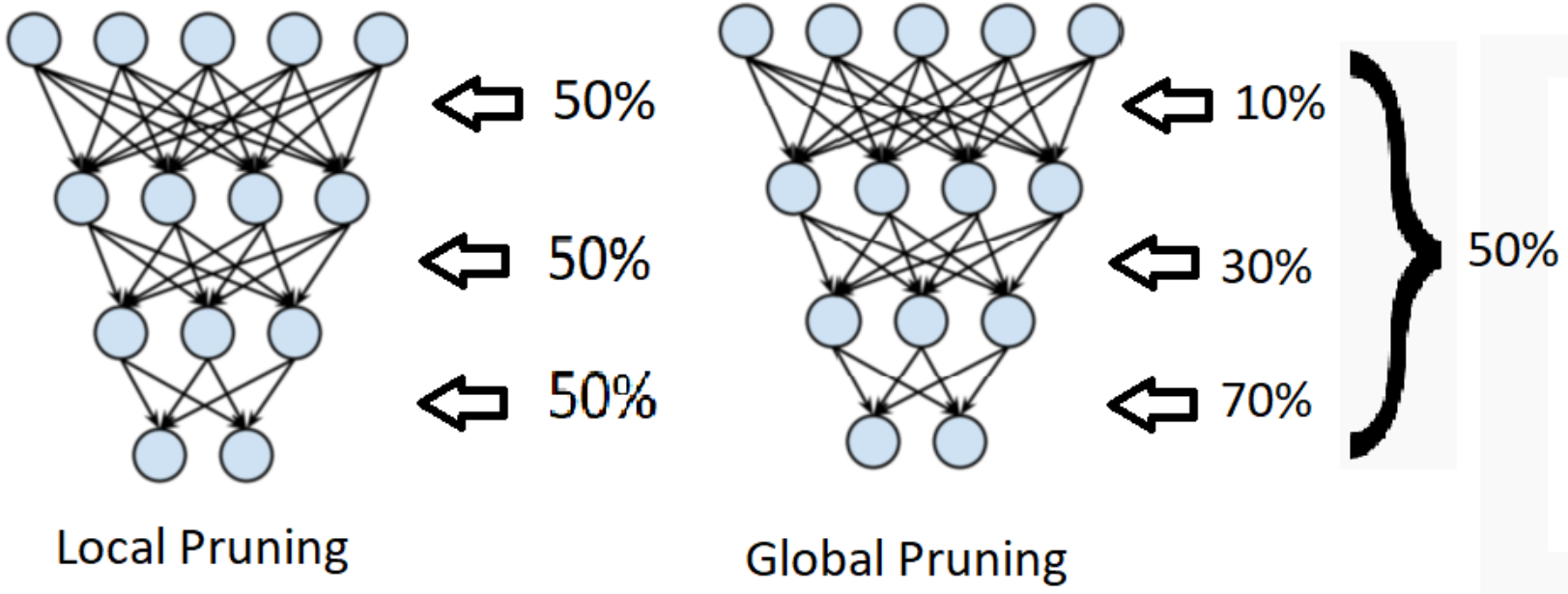


Kernel level sparsity example



Related work : Pruning

- **Pruning zone application**



Related work : Pruning

- **Neurons pruning / weight pruning**

- **Magnitude based pruning**

$$threshold(w_i) = \begin{cases} w_i \text{ if } |w_i| > \lambda \\ 0 \text{ if } |w_i| \leq \lambda \end{cases}$$

- **Neurons Pruning**

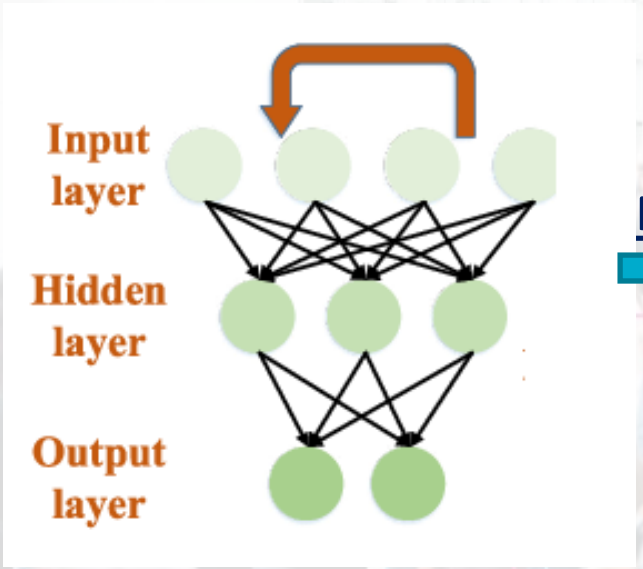
- **Movement based pruning**

$$\sum_t \left(\frac{\partial \mathcal{L}}{\partial W_{i,j}} \right)^{(t)} W_{i,j}^{(t)}$$

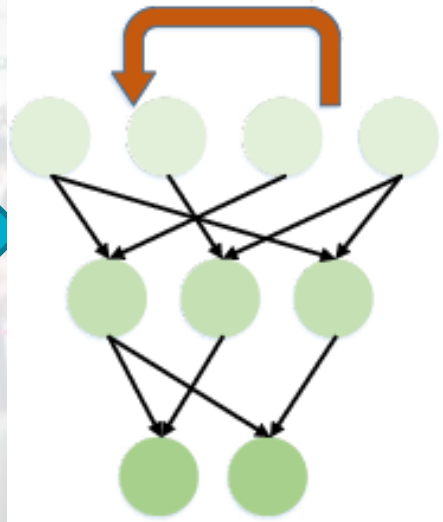
Related work : Pruning

- Pruning scheduling

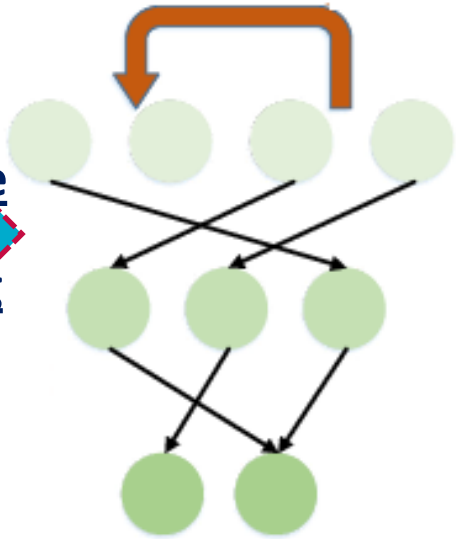
Initial training → **Retraining** → **Retraining**



Pruning



Iterative Pruning



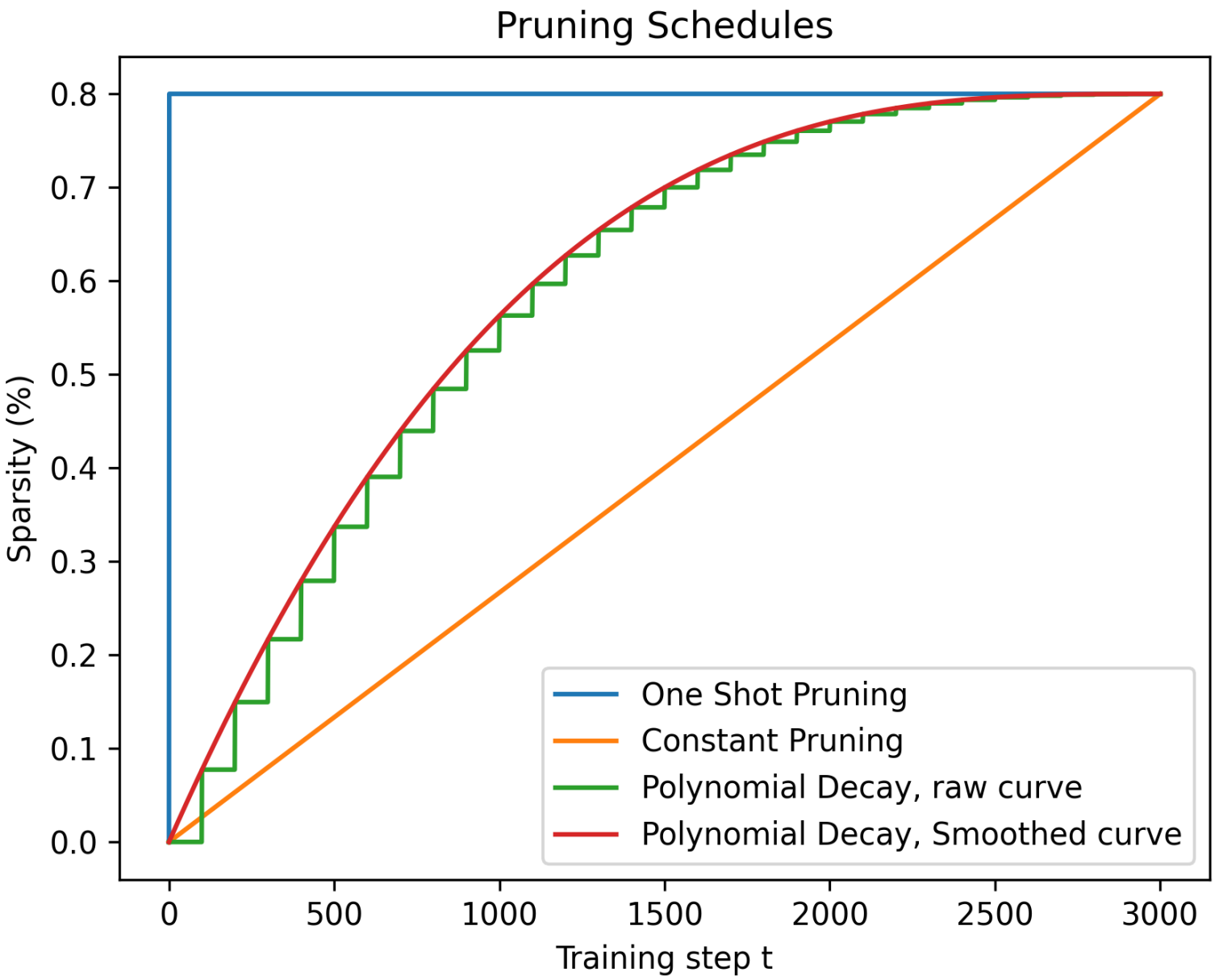
One iteration

Many iterations

One Shot pruning

Iterative pruning

Pruning schedules



PLAN

Main approaches of Edge AI in Deep Learning

a. Pruning

b. Quantization

c. Knowledge Distillation

Related work : Quantization

-0.2	1	0.3
0.1	-0.6	-0.7
1.2	0.4	0

32 bit

quantization

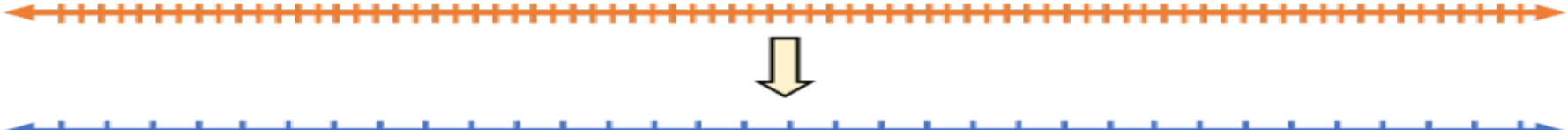
1	3	2
1	0	0
3	2	1

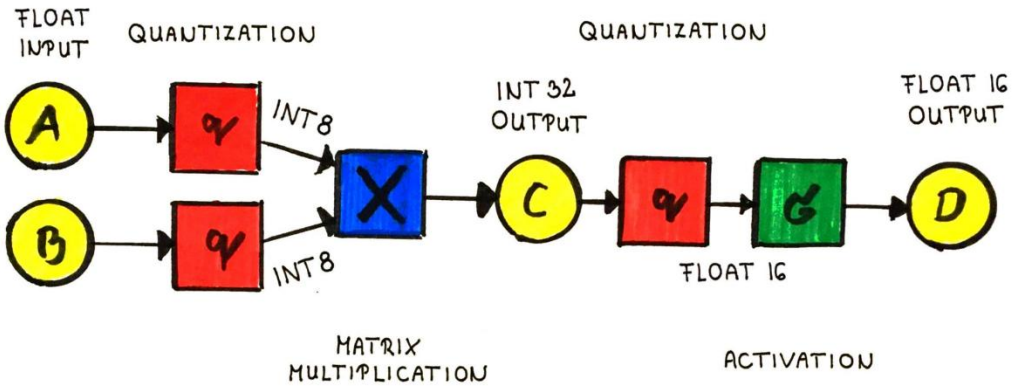
2 bit

+

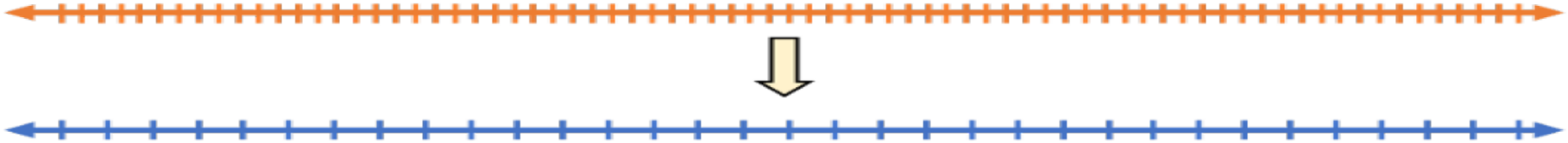
index	[in bits]	value
0	[00]	-0.6
1	[01]	0
2	[10]	0.4
3	[11]	1.1

32 bit





Related work : Quantization



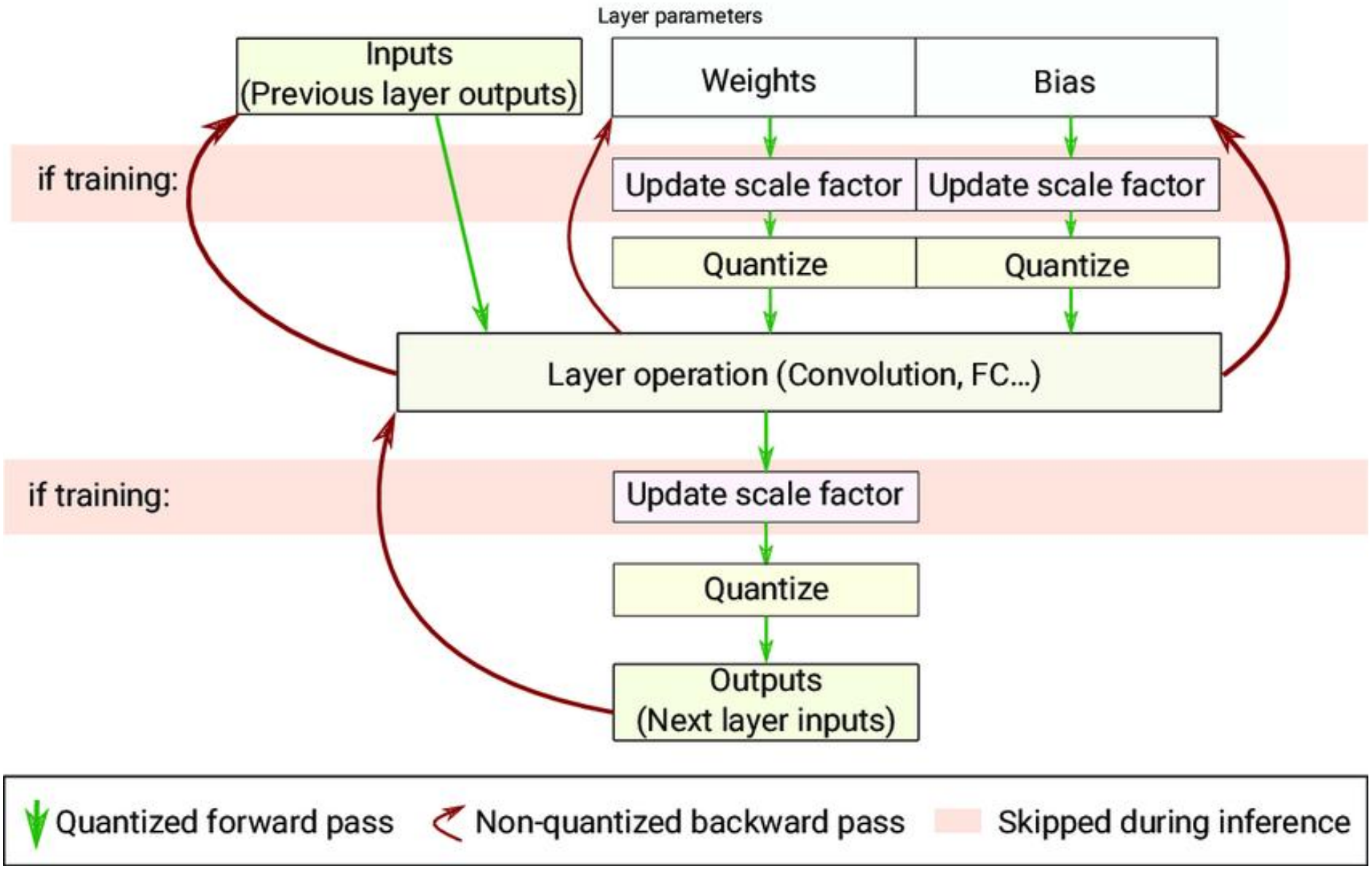
Benefits:

- **Faster arithmetic operations**
- **Reduction in model size**
- **Compatibility with more (and less) devices**

When to apply ?

- **Dynamic Quantization :** quantization of weights only (both fp16 and int8)
- **Static Post training quantization :** quantization of weights/activations (8 bit)
- **Quantization Aware Training**

Quantization Aware Training



PLAN

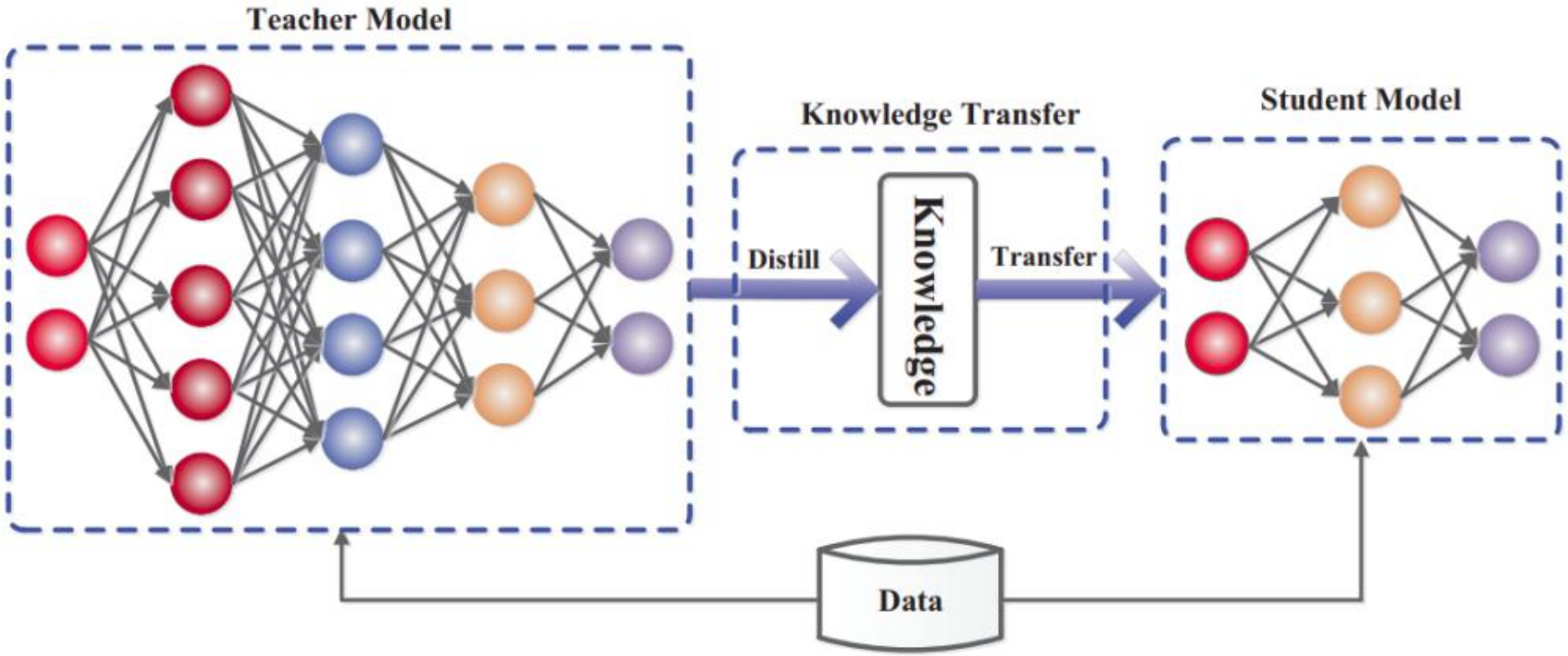
Main approaches of Edge AI in Deep Learning

a. Pruning

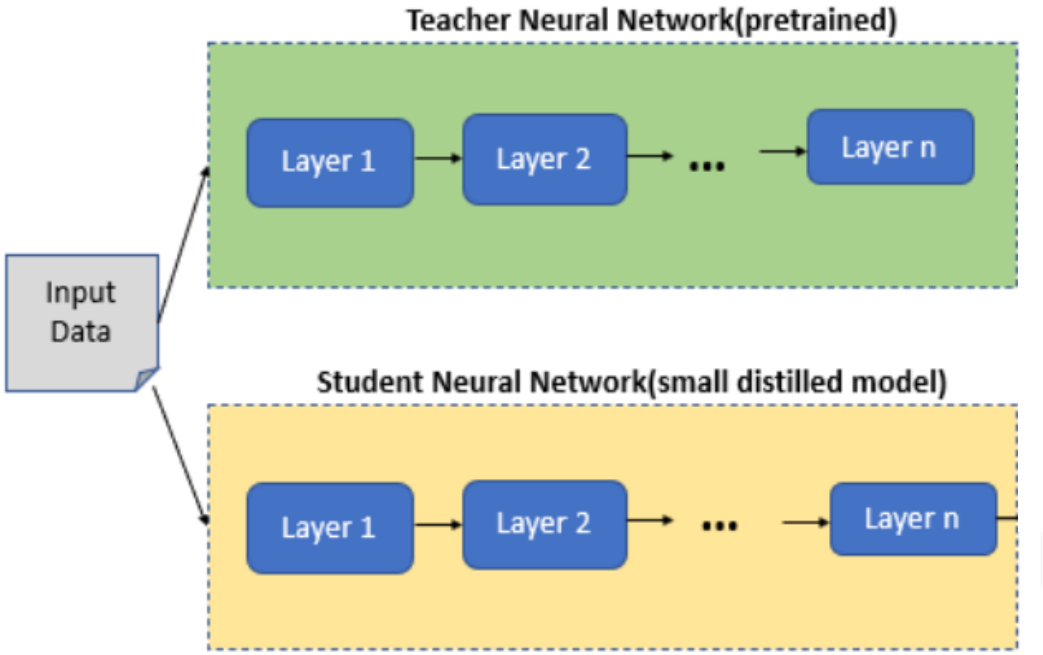
b. Quantization

c. Knowledge Distillation

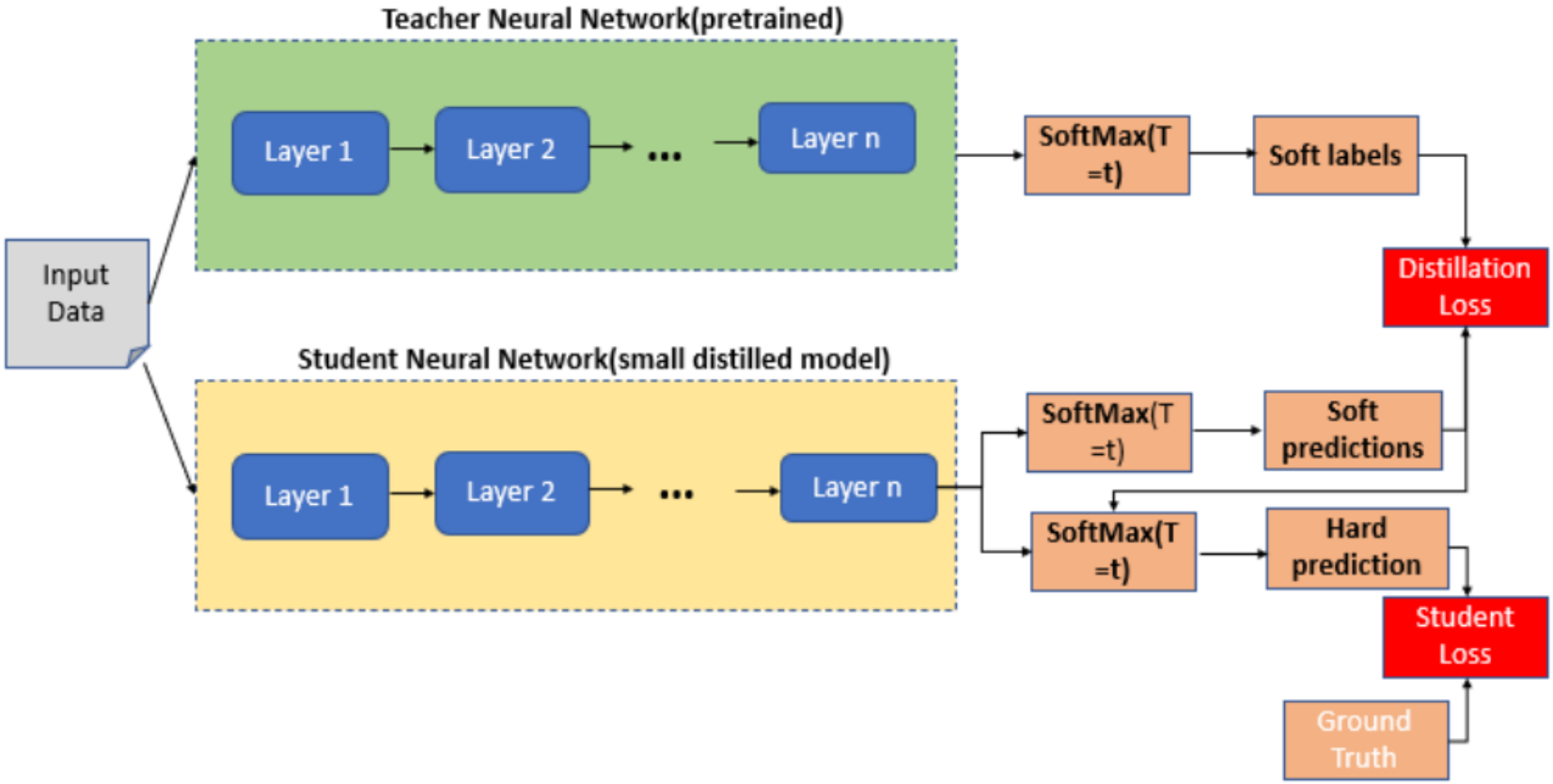
Knowledge Distillation



Knowledge Distillation : Process



Knowledge Distillation : Process



DNN compression : discussion

Pruning

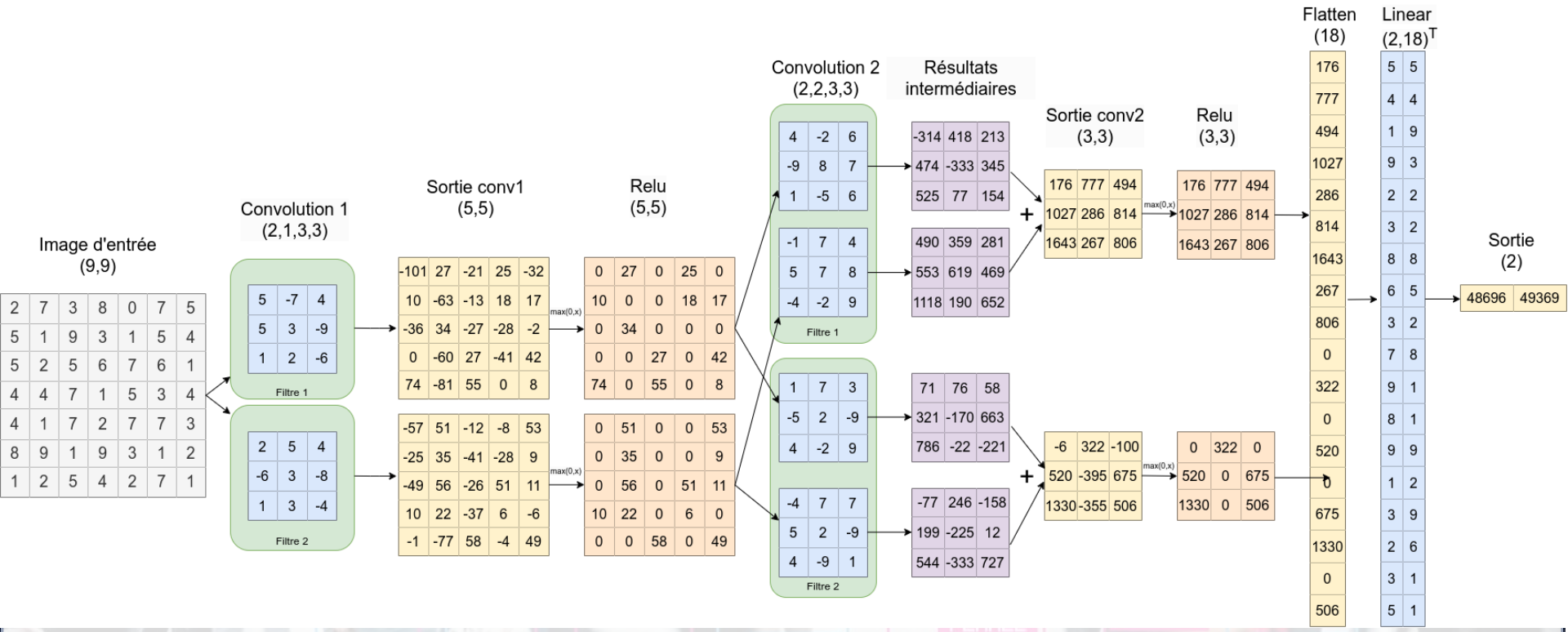
- Major methods generate **sparse** neural networks
- Reduction of Mem size but **no reduction in comp time or RAM consumption**
- Not suitable for Edge AI applications

Block pruning Proposal

- Analyze the dependency of the neural network nodes → blocks
- Calculate the average magnitude of the blocks
- Remove the low magnitude blocks
- **Generate a Dense and pruned** neural network

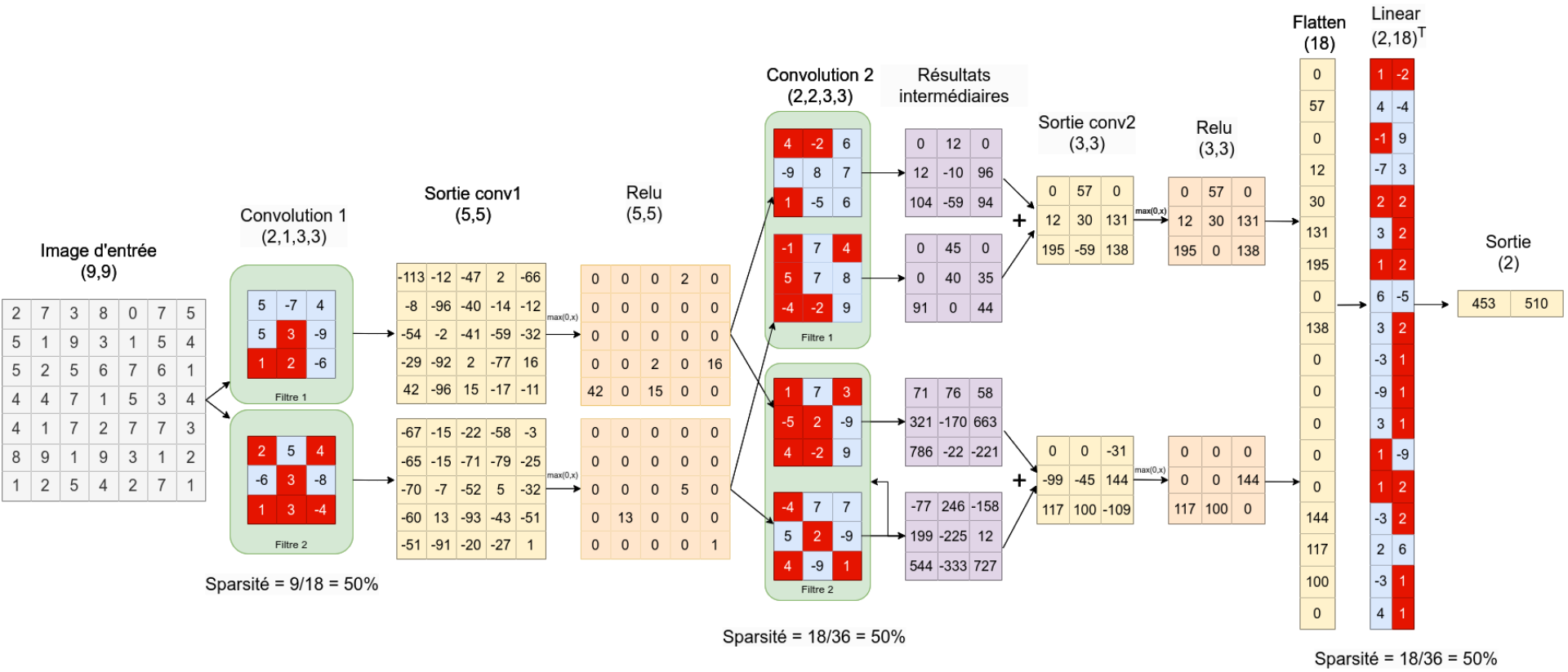
DNN compression : discussion & illustration

Initial CNN (Without Pruning)



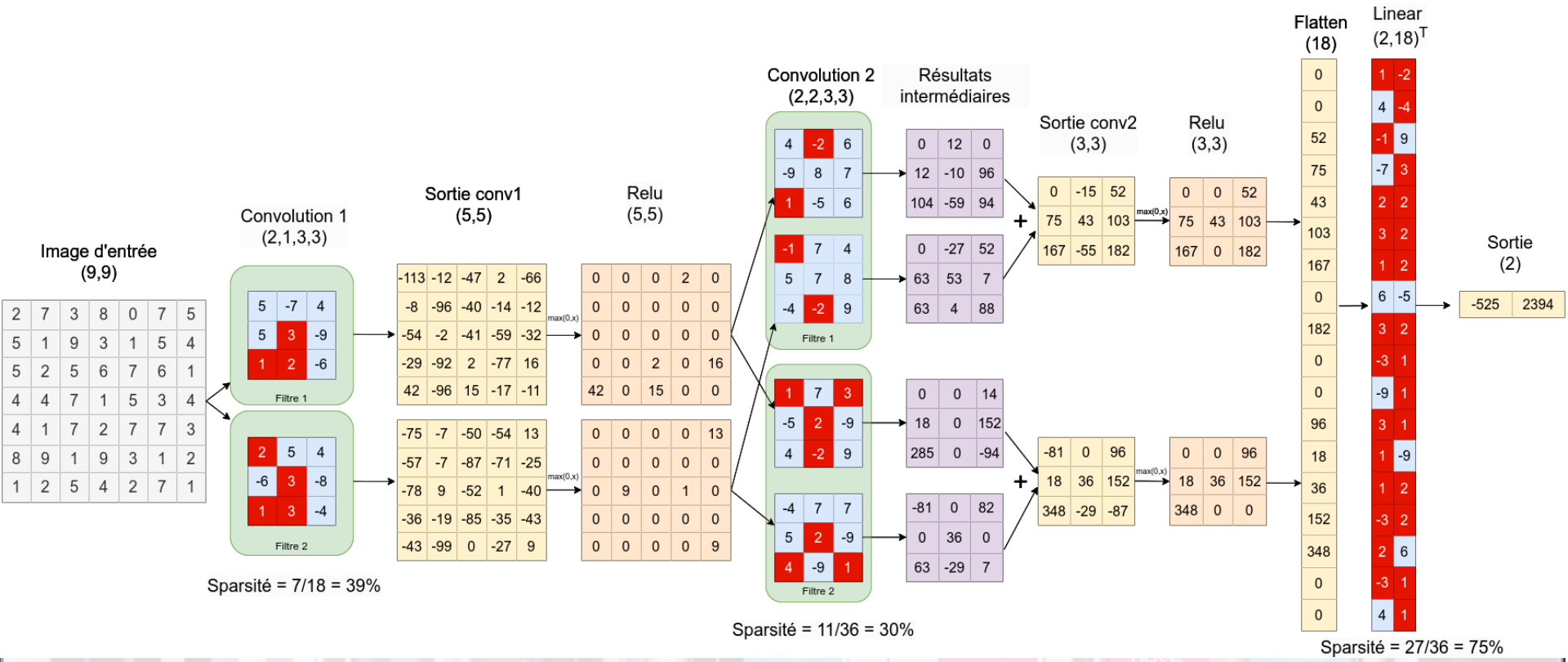
DNN compression : discussion

Unstructured Local Pruning : 50%



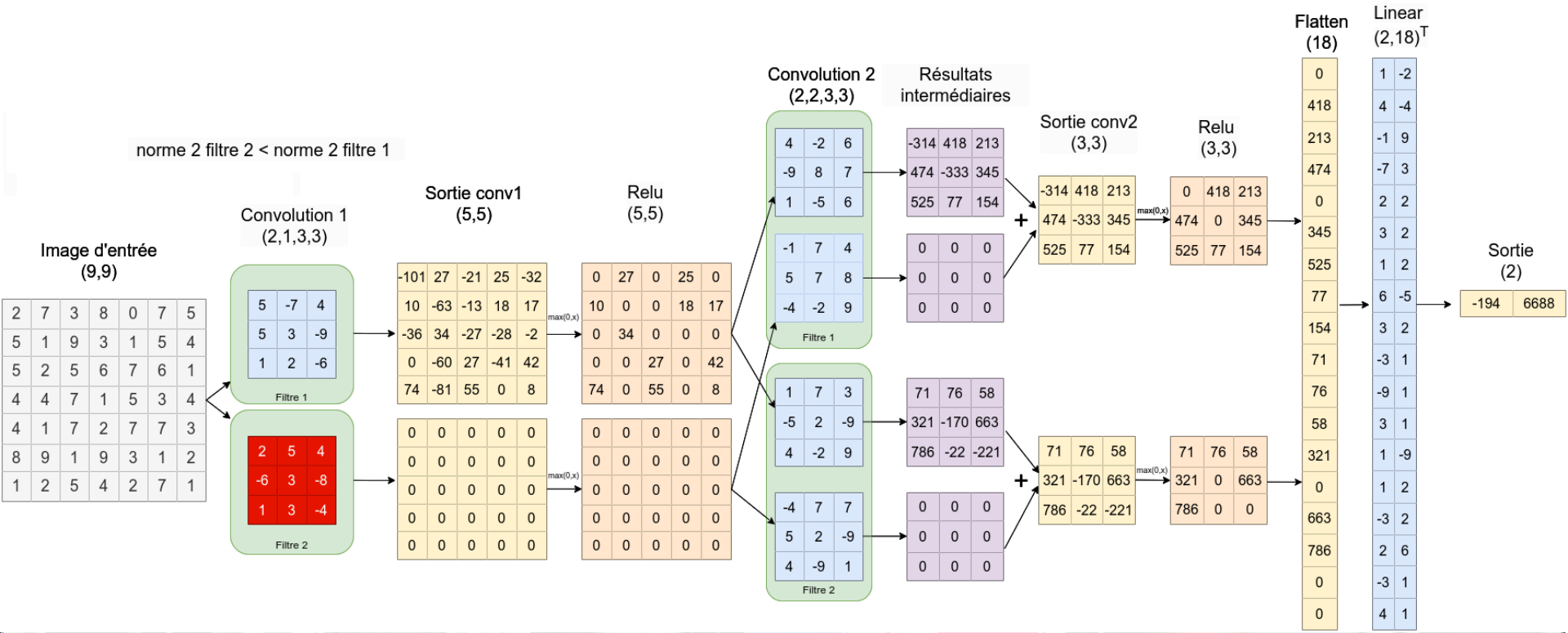
DNN compression : discussion

Unstructured Global Pruning : 50%



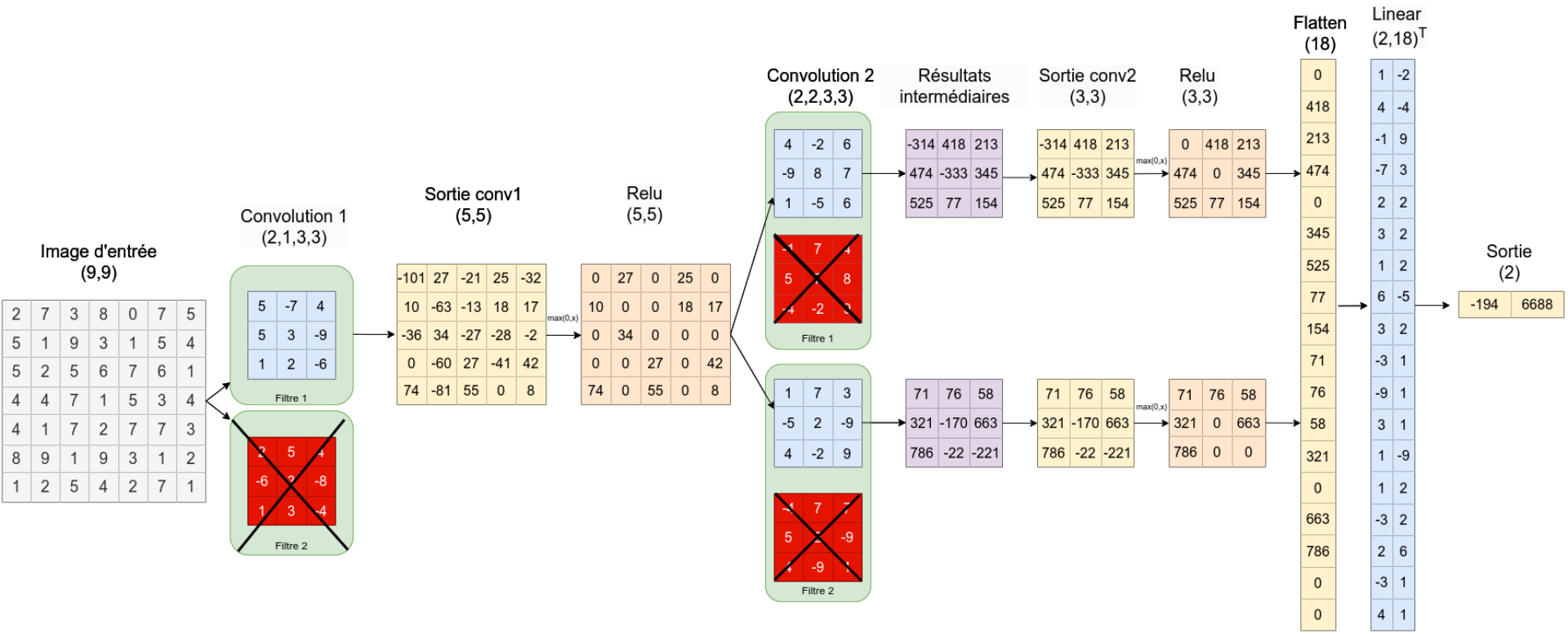
DNN compression : discussion

Structured Filter Pruning



DNN compression : discussion

Structured Block Pruning



Outils de développement

Matériel

Cloud : Google Colab, Google Colab pro, vast ai, etc.

Edge: Environnement « Deeplearning » préconfiguré

Matériel embarqué

Jetson Xavier

Jetson Nano

Movidius (avec Raspberry PI)

Google Coral

Développement

Git : outils de gestion de versions, de partage de code, etc.

Collaboration

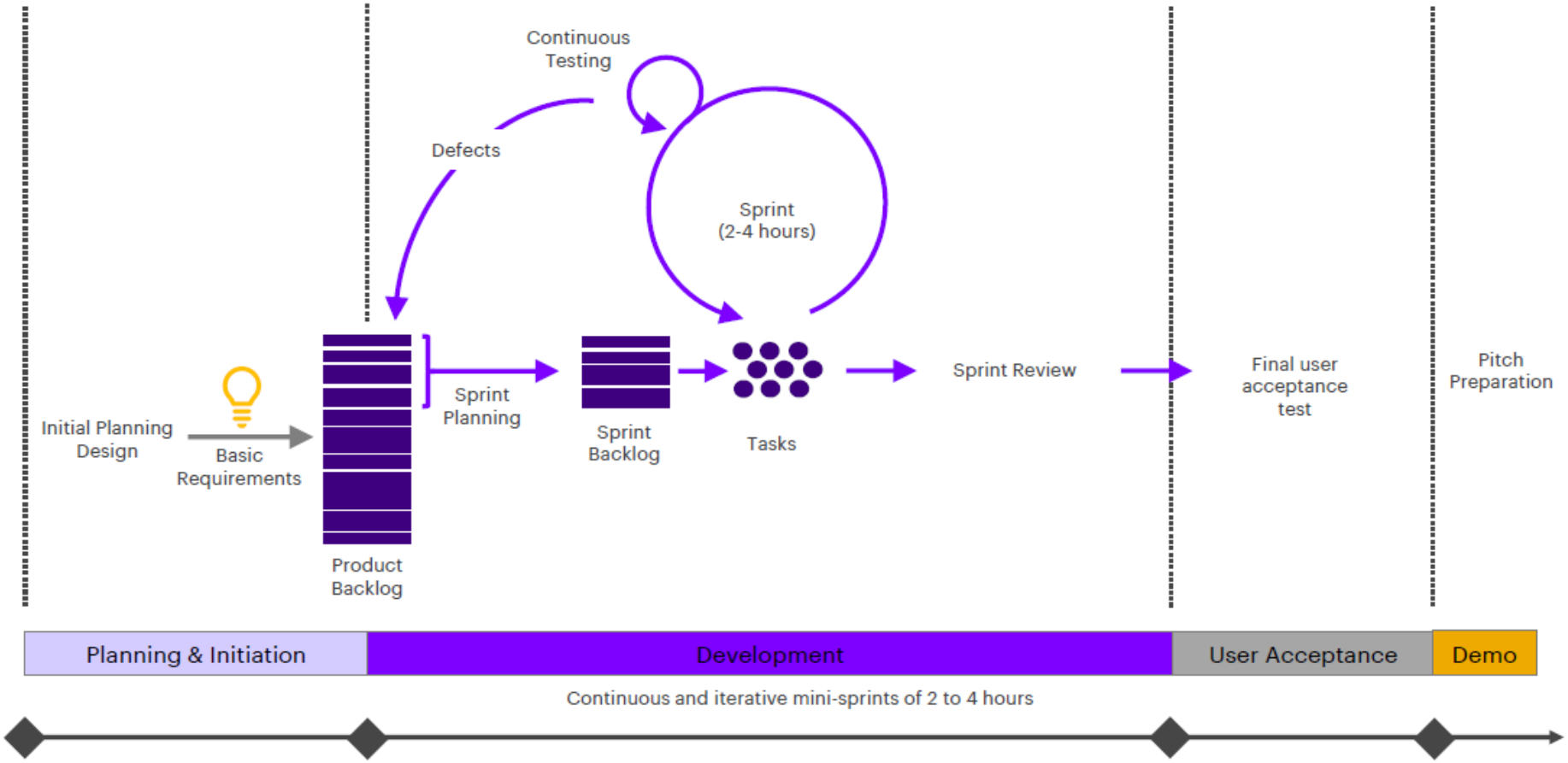
Rejoindre le groupe Slack

Partage de bases de données annotées et autres

Jetson Xavier

GPU	512-core Volta GPU with Tensor Cores
CPU	8-core ARM v8.2 64-bit CPU, 8MB L2 + 4MB L3
Memory	32GB 256-Bit LPDDR4x | 137GB/s
Storage	32GB eMMC 5.1
DL Accelerator	(2x) NVDLA Engines
Vision Accelerator	7-way VLIW Vision Processor
Encoder/Decoder	(2x) 4Kp60 | HEVC/(2x) 4Kp60 | 12-Bit Support
Size	105 mm x 105 mm x 65 mm
Deployment	Module (Jetson AGX Xavier)

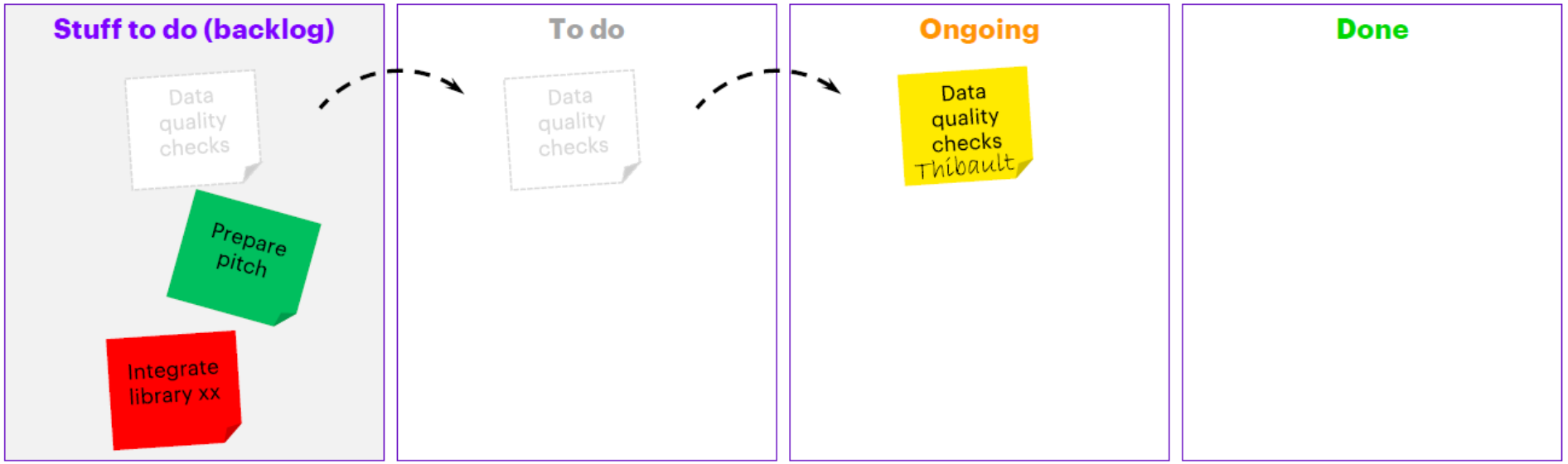
Organisation du travail en équipe



Organisation du travail en équipe

HACKATHON & WORKSHOP AI

Organisation du travail en équipe



Organisation du travail en équipe

HACKATHON & WORKSHOP AI

Stuff to do (backlog)	To do	Ongoing	Done
Integrate library xx		Data quality checks Thibault	Prepare pitch Marc

PLAN

I. Introduction & Programme du Workshop

II. Objectif du workshop

III. Phase 1 : développement et entraînement des modèles

IV. Phase 2 : portage de solution sur ressource Edge : Jetson Xavier

V. Phase 3 : optimisation (compression) et explicabilité de modèles

Conclusion

Edge AI for Smart Cities

- Module Edge AI pour villes intelligentes
- Intégration de modèles offrant une haute précision
- Traitement temps réel à partir de vidéos de la Webcam
- Portage sur matériel Edge (optimisation, réduction de mémoire, etc.)
- D'autres pistes :
  - personnalisation de l'interface graphique
  - Envoie de notification et alertes (sms, emails, etc.)
  - Innovations personnelles

Edge AI for Smart Cities

- Remise du travail et résultats sur Moodle (pour le **25/05/2024**)
- Présentation des résultats en présentiel
- Présentation des résultats devant le Jury le **Dimanche à partir de 17h30**
- Présentation : **pitch + démo**
- Durée de chaque présentation (groupe) : **15 minutes**
- Remise des prix à partir de **18h30** (30/04)
  - **Performance Award :** offert par Accenture et le certificat « Hands on AI »
  - **Innovation Award:** offert par l'institut Numédiart (UMONS)
  - **Pitch Award :** offert par l'institut InforTech (UMONS)

Conclusion

More details on : https://hackia.eu/

Let us start