Distributed Edge Intelligence

Denis Trystram

Univ. Grenoble Alpes, France
Denis.Trystram@Univ-Grenoble-Alpes.fr
MIAI

Eclipse Days
feb. 14, 2020

Agenda of this talk

1. Presentation of MIAI
2. Research program on Edge intelligence
   A case study
3. Personal thoughts about the usage of (smart) IoT

Emergence of IA (in France): a Brief History

- An old topic.
- **March 2018** report of Cédric Villani. cartography of AI in France and some recommendations.
- **early 2019** ANR call (100 millions d'euros) for 4 institutes 3IA, whose objective is:
  Boost research and develop industry
- **mid 2019** Grenoble was selected
  (the other laureat are Paris, Sophia, Toulouse).

Today:

Everybody is doing AI
AI is synonym of Machine Learning (and most specialized Deep learning).

Emergence of IA (in France): a Brief History

- An old topic.
- **March 2018** report of Cédric Villani. cartography of AI in France and some recommendations.
- **early 2019** ANR call (100 millions d'euros) for 4 institutes 3IA, whose objective is:
  Boost research and develop industry
- **mid 2019** Grenoble was selected
  (the other laureat are Paris, Sophia, Toulouse).

Today:

Everybody is doing AI
AI is synonym of Machine Learning (and most specialized Deep learning).

University Grenoble Alpes

Objectives and overall presentation

Institute to support the development of education, research and transfer in ai, at the service of human being and the environment.

- Develop world-class interdisciplinary researches in AI and AI for human beings and the environment.
- Offer attractive courses in AI.
- Sustain innovation in AI and help to develop AI in major companies, SMEs and start-ups.
- Inform and interact with citizens an all aspects of AI.

Public Partners

UNIVERSITÉ Grenoble Alpes

Inria inventeurs du monde numérique

Grenoble INP

cnrs

DE LA RECHERCHE À L'INDUSTRIE cea

GRENOBLE ECOLE DE MANAGEMENT BUSINESS LAB FOR SOCIETY

Inserm La science pour la santé From science to health

Industrial Partners

23 big companies

- Google
- orange
- THALES
- criteo
- Atos
- Hewlett Packard Enterprise
- ArcelorMittal
- Schneider Electric
- msa santé famille retraite services
- SIEMENS
- NAVER LABS
- AFP
- IBM
- TOTAL
- ENEDIS L'ELECTRICITE EN RESEAU
- Constellium
- Medtronic
- facebook
- RENAULT Passion for life
- STMicroelectronics
- Mentor Graphics
- NOKIA
- HUAWEI

21 small and medium companies

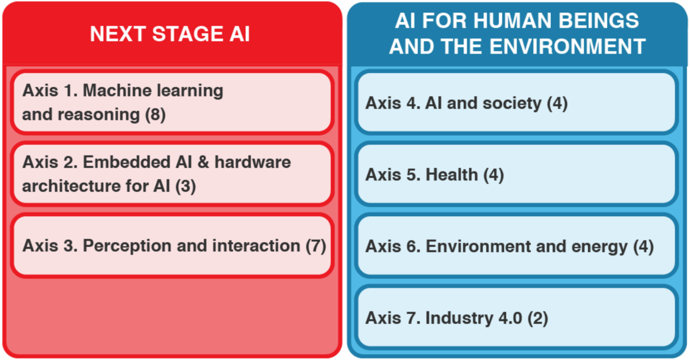
- genOway
- COSMOTECH
- Energy Pool Smart energy management
- ProbaYes Mastering Uncertainty
- neovision
- TexiSense
- PIXYL
- kelkoo
- hurence
- iexec
- ICALPS
- coservit
- ARYBALLE Technologies
- QARNOT COMPUTING
- eBikeLabs
- Quantmetry Building AI with pioneers
- Skopai Deep Tech Insights
- SeqOne
- KOELIS
- Carbon Bee
- INCOM

11 very small companies

- SurgiQual Institute
- THEIA
- COMONGO Aim straight / Viser juste
- ryax technologies
- microlight3D
- P2LINK
- Etisense Jacketed monitoring for life sciences
- Odit-e
- SentinHealth
- neoma LABS

Two main Themes

Organized in Axis and Research Programs.

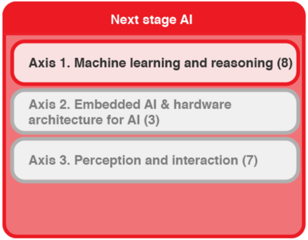


Brief Overview of the research axis

1.1. Machine learning models: invent data efficient and robust models and algorithms, develop lifelong learning and multiple learning solutions, explore new paradigms inspired by cognitive sciences

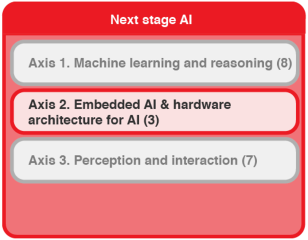
1.2. Statistics and optimization: develop frameworks (convex optimization, statistical physics and random matrix theory) to better understand and optimize large-dimensional models

1.3. Fair and evolvable AI: develop AI systems that explain their decisions, certify fairness and privacy, and understand the mechanisms driving the evolution of knowledge



2.1. Neuro-processing units: develop new hardware/software architectures for energy efficient AI, mainly through spiking neural networks and optimization of existing hardware architectures

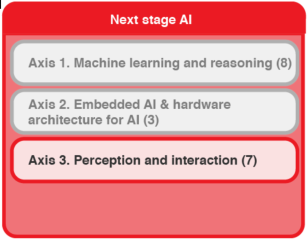
2.2. Distributed intelligence: develop optimised algorithms and adequate orchestration software to process the data produced by distributed AI



3.1. Robotics: develop socially aware robots for enhanced interactions with humans and explore combination of AI techniques with control theory to design efficient and safe robotic systems

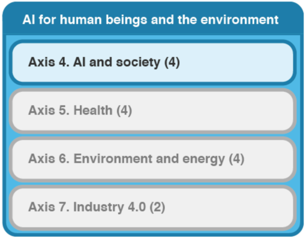
3.2. Natural language and speech processing: develop more versatile speech and language technologies able to learn from fewer examples, adapt to various kinds of perturbations and extend to new languages and contexts

3.3. Computer vision: associate learning algorithms to developments on the representation of objects and their dynamics with the aim to enhance 3D artificial vision and develop new generation of self-supervised visual systems



4.1. AI regulation: develop ethical and legal frameworks for AI, without them being stumbling blocks to the development of AI

4.2. Integration of AI into society: bring together AI approaches and social science methods to explore the integration of AI and the adaptation of algorithms to contexts influencing users



5.1. Real-life 4P medicine*:* develop new AI tools for improved health trajectories and augmented patient empowerment

5.2. Multi-omics: identify new biomarkers from multimodal health data and develop new tools on their basis to compute personalized risk scores

5.3. Computer-assisted medical intelligence: develop new generation of intra-operative AI-based assistants to treat patients more efficiently and less invasively

6.1. AI solutions for natural disasters: identify concentration and location of pollutants, develop better models of subterranean processes, measure the impact of climate change and develop new tools for environmental monitoring and geophysical data assimilation

6.2. Optimizing energy management: develop new generation of AI tools for smart grids, incl. optimization of network topology and prediction of its evolution and the evolution of its components

7.1. Human-centric manufacturing: develop AI techniques to enhance human-machine collaborations (cobotics and augmented reality) and to develop novel decision-making methods for reactive operations and supply chain management

7.2. Predictive quality: develop and use AI techniques to predict quality of new materials, products, production processes, maintenance activities and industrial systems

Educational Aspects

Goal:

train students and professionals (IT experts and application practitioners) at all levels, undergraduate, master, PhD.
Target: 1200 students.

Two ways:

1. Strengthen and articulate existing courses.
2. Create new labels (short/dedicated, on-demand courses).

Innovation/Transfer and partnership

Budget:

19 Meuros over 4 years (same amount from the french gov. and for the industrial partners).

MIAI
Grenoble Alpes
Multidisciplinary Institute
In Artificial Intelligence

Contact:

eric.gaussier@univ-grenoble-alpes.fr

For more detail:
https://miai.univ-grenoble-alpes.fr/

Computing needs for AI

sept. 2019 Jean Zay, large computing platform operated by Genci (HPE with more than one thousand Nvidia GPUs).

Such equipments and some target applications (face recognition, justice, ...) feed the fantasm of the *Big Brother* aspect of AI.

Huge variety of computing/digital systems

- HPC and Data centers (clouds)
- Fog – small clusters (hundreds/thousands cores)
- Edge – laptops/smart phones, embedded systems
- IoT – sensors (extreme edge)

Computing needs for AI

sept. 2019 Jean Zay, large computing platform operated by Genci (HPE with more than one thousand Nvidia GPUs).

Such equipments and some target applications (face recognition, justice, ...) feed the fantasm of the *Big Brother* aspect of AI.

Huge variety of computing/digital systems

- HPC and Data centers (clouds)
- Fog – small clusters (hundreds/thousands cores)
- Edge – laptops/smart phones, embedded systems
- IoT – sensors (extreme edge)

Edge Computing

The easy use of black box AI creates a lot of new *needs* everywhere without a (deep) understanding of what happens.

Edge as an alternative.

Compute close to the place where the data are produced.

Many nice features including energy efficiency and privacy.

Edge Computing

The easy use of black box AI creates a lot of new *needs* everywhere without a (deep) understanding of what happens.

Edge as an alternative.

Compute close to the place where the data are produced.

Many nice features including energy efficiency and privacy.

A zoom on Distributed intelligence (topic 2.2)

Structure of the program (2019-2023)

- Academic partners: Frédéric Desprez (Inria, SILECS), Jean-Paul Jamont (LCIS Valence) Noel De Palma (LIG), Denis Trystram (LIG).
- Industrial partners

ryax technologies

Carbon Bee

QARNOT COMPUTING

AtoS

orange™

- 2 academic PhD students, 4 PhD with conventions with companies, 2 other expected PhDs and PostDoc, visiting positions. Platforms and simulators.
- Observers: GFI, Huawei, STmicroelectronics, Total

Topics covered in this program

Edge Intelligence: towards sober and frugal AI.

- Federated (distributed) learning
- on-line learning and learning with small amount of data (streaming)
- Job allocation and service orchestration

Challenges

Manage efficiently such complex infrastructures composed of multiple and heterogeneous digital/computing components.

- Large amount of polymorphic data issued from multiple sources.
- Resources appear and disappear.
- New storage capabilities.
- Highly heterogeneous characteristics of the different components, including hardware, OS kernels, compilers, etc.
- Networks of different types and instability (different latencies).

Edge platform: case study at Qarnot

From Heaters to Computing Resources: **“turn IT waste heat into a viable heating solution for buildings.”**

The actual Qarnot platform (Bordeaux and Paris):

- $\sim$1,000 distributed QRads embedding $\sim$3,000 diskless computing units and several sensors
- $\sim$20 local servers (QBoxes) with memory disks
- 1 global server (QNode) with a centralized storage server

Edge platform: case study at Qarnot

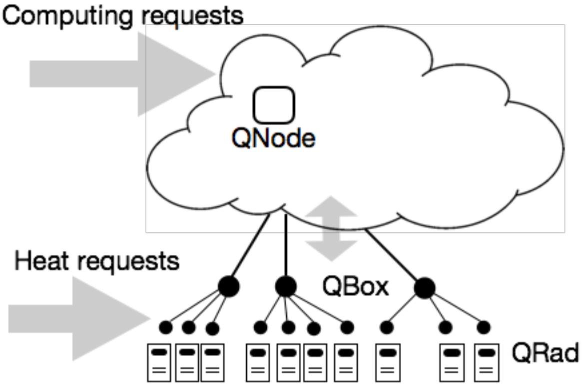
From Heaters to Computing Resources: **"turn IT waste heat into a viable heating solution for buildings."**

The actual Qarnot platform (Bordeaux and Paris):

- ∼1,000 distributed QRads embedding ∼3,000 diskless computing units and several sensors
- ∼20 local servers (QBoxes) with memory disks
- 1 global server (QNode) with a centralized storage server

Credits: `https://www.qarnot.com`

Schematic view of Qarnot Platform



Two types of computing requests

Cloud tasks:

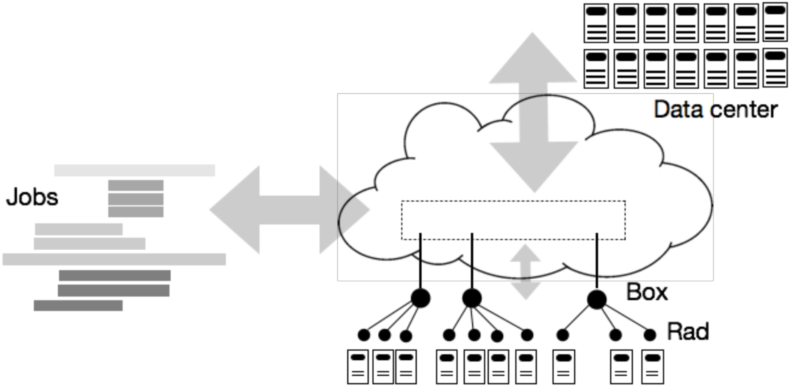
- Submitted to the QNode
- data-set dependencies in the centralized storage
- different priorities (low or high)

local/IoT tasks:

- Submitted to a – local – QBox
- data-set dependencies in the QBox disk
- different priorities (low, high or very high)
- Should be executed **locally**

Abstract view of the Qarnot Platform

Tasks/jobs (groups of **sequential instances**) are submitted on-line.



Platform Dynamicity

Resources appear and disappear over time: **the inhabitants decide according to the weather or their schedule!**

- Available resources when heating is required (QRad is ON)
- Unavailable when ambient air is too warm (QRad is OFF)

Also depends on the task priority

Network uncertainties:

- Link failures
- Congestion

Threats and Opportunities of Edge Computing

Put intelligence at the edge

The number of digital objects connected to Internet is growing exponentially.

IT represents 9% of the energy consumption in our country.
The volume of data storage grows by 35% per year in data centers[1].

The impact of IT is expected to be greater than transports...
AI is a major actor of this growth.

For IoT
the *equipment rate* per person is also growing very fast.

[1] 985 exabytes expected in 2020

We observe two extreme positions:

- Those who consider that IT is bad and dangerous (the *Alarmists*).
- Those who think IT will save the planet.

IoT is individual-centered.

A dream?

Use IoT for changing the practices/usages to preserve the planet.

We observe two extreme positions:

- Those who consider that IT is bad and dangerous (the *Alarmists*).
- Those who think IT will save the planet.

IoT is individual-centered.

A dream?

Use IoT for changing the practices/usages to preserve the planet.

Concluding remarks/Question(s)

What usage for IoT?

- The production of digital objects grows exponentially but most of them lead only to an increase of *comfort*.
- How many small digital components and sensors target an optimization of energy consumption?
- How to avoid Jevons' effect.

Concluding remarks/Question(s)

What usage for IoT?

- The production of digital objects grows exponentially but most of them lead only to an increase of *comfort*.
- How many small digital components and sensors target an optimization of energy consumption?
- How to avoid Jevons' effect.