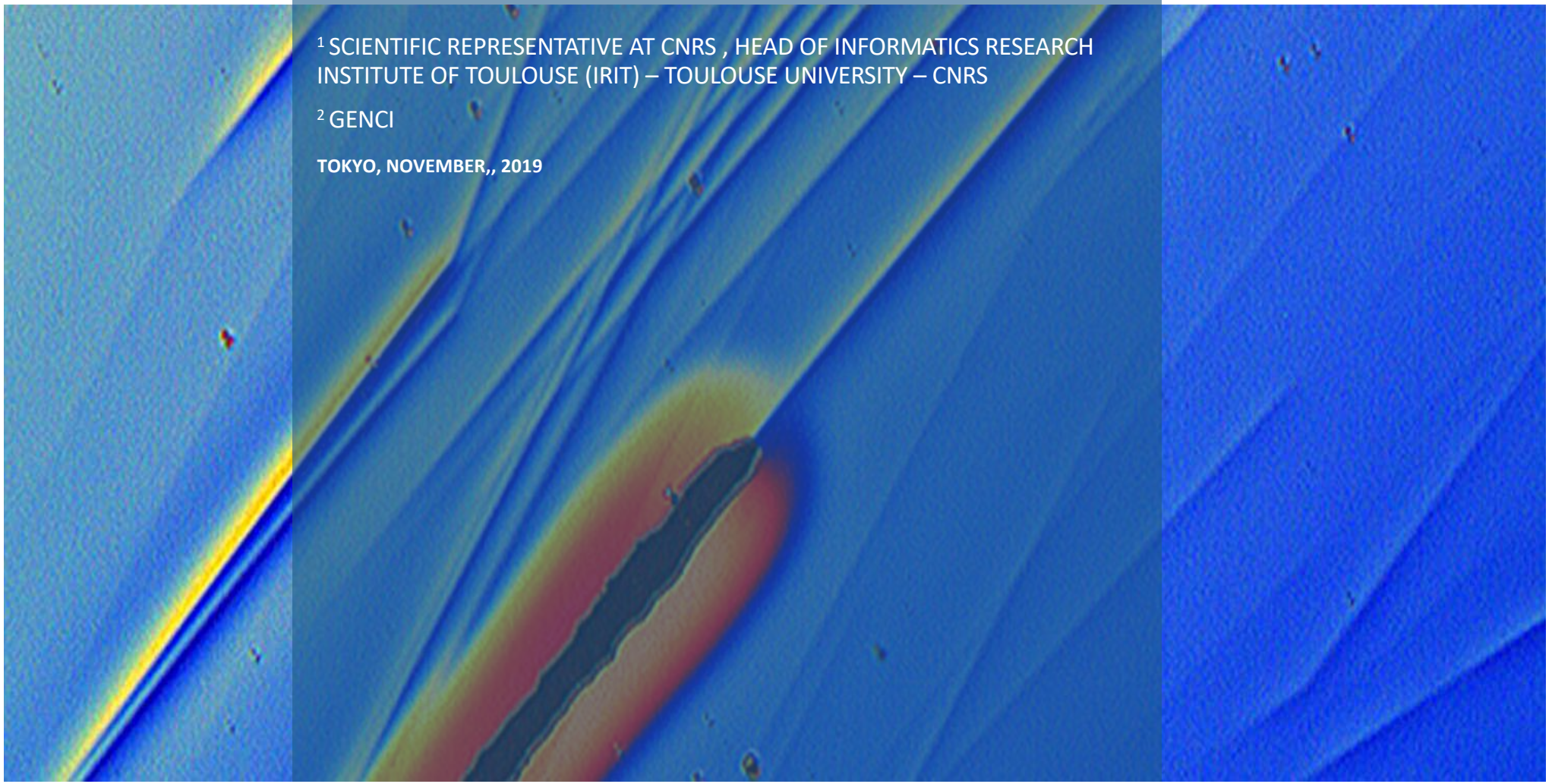
Toward the convergence between HPC, HPDA and IA

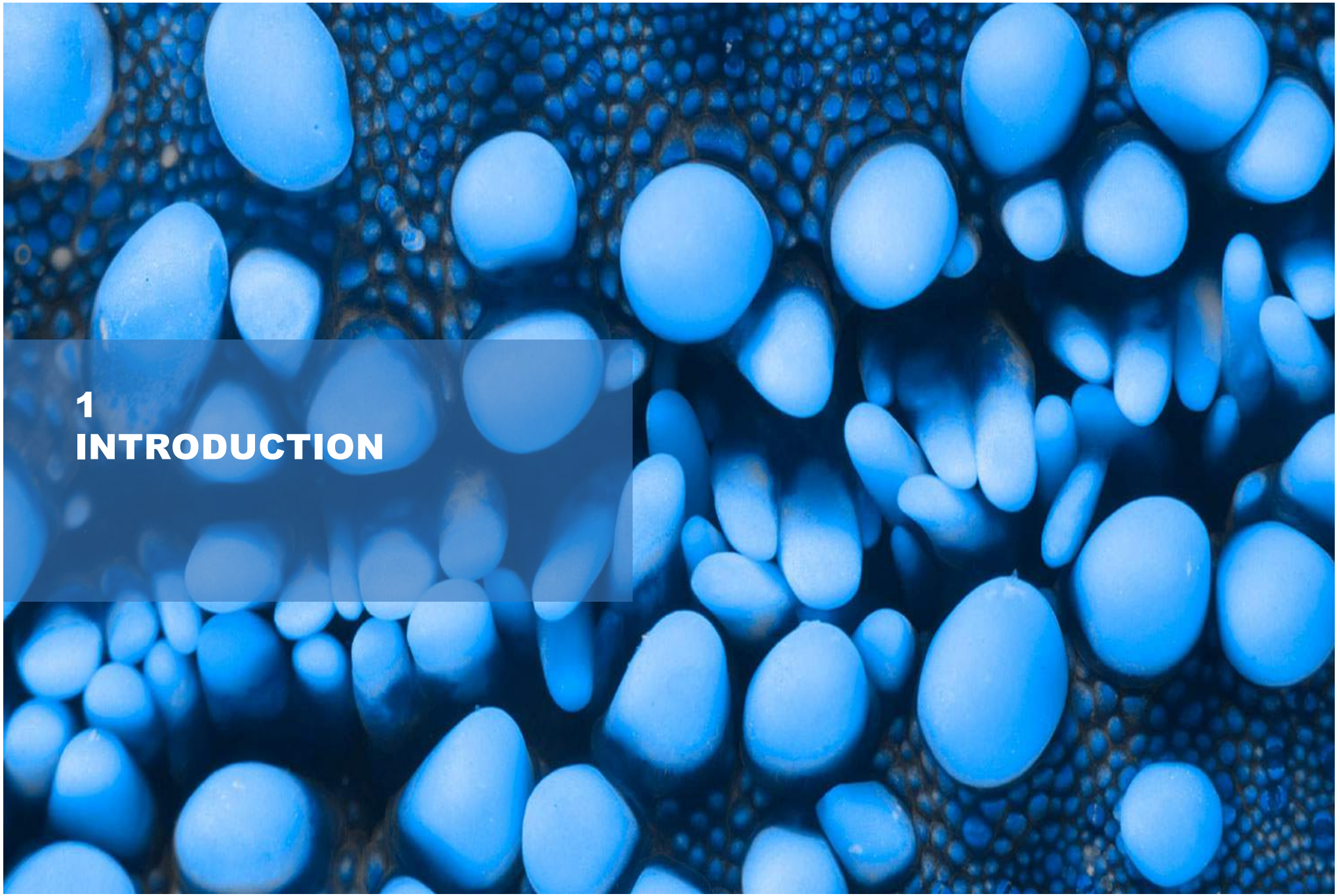
M. DAYDÉ[1] AND S. REQUENA[2]

[1] SCIENTIFIC REPRESENTATIVE AT CNRS , HEAD OF INFORMATICS RESEARCH INSTITUTE OF TOULOUSE (IRIT) – TOULOUSE UNIVERSITY – CNRS

[2] GENCI

TOKYO, NOVEMBER,, 2019





1
INTRODUCTION

1

CNRS IN A FEW NUMBERS

- A 31,500-strong workforce, including 24,600 permanent personnel, 15,128 researchers and 16,509 engineers, technicians and administrative staff carrying out and supporting research
- 2019 budget: 3.4 billion euros
- Over 1,100 research and service units
- More than 52,000 publications in high-caliber international journals each year, 60 % being published jointly with at least one foreign laboratory
- 21 Nobel prize laureates and 12 Fields Medal winners

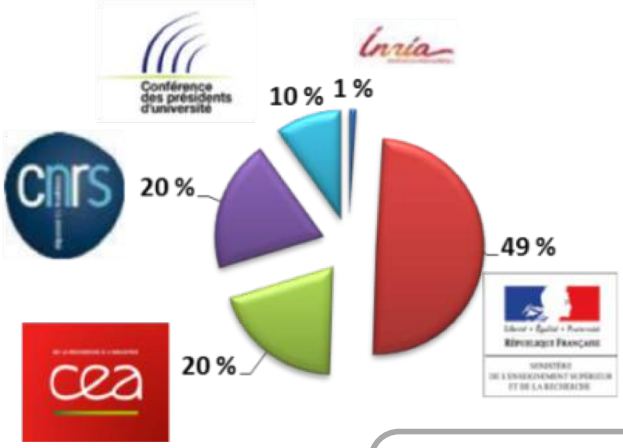
GENCI IN A NUTSHELL (1/2)

National HPC + storage resources management

French strategy for equipping 3 national centres (TGCC, Idris, Cines)
Resources available for researchers from **academia and industry** for open research, 1.9 billion core hours and >600 projects / year

↳ **National dynamics**

Expanded into the regions with



"Civil society"
5 shareholders
2019 budget = €39 M

Participation in the Europe of HPC

GENCI represents France within PRACE
↳ **Joliot Curie as Tier0 (60%)**
↳ **European dynamics**

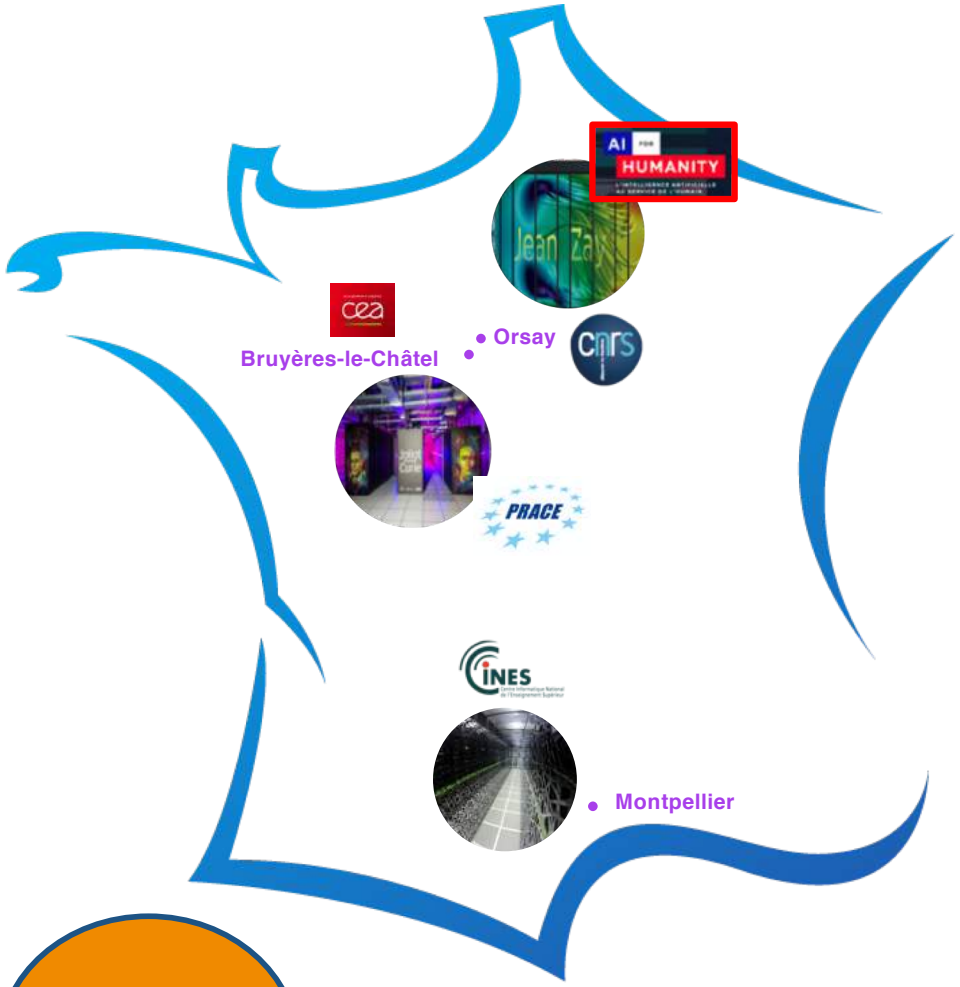
Promotion of HPC

Among academics and industrials
With a specific action towards SMEs

↳ **Democratisation of the use of HPC and AI in the context of CGI's CEI with Teratec**

GENCI IN A NUTSHELL (2/2)

National HPC production systems



>40 PF
= *2/yr

- ❑ 3 national centers (Tiers 1)
  - CINES : Occigen
  - TGCC : Joliot-Curie – (Tier 1 & Tier 0)
  - IDRIS : Jean Zay
- ❑ Multi anual invest. plan(10 yr) - **39 M€/an**
  - 2018 TGCC renewal / 2020 extension
  - 2019 IDRIS renewal / **2021 extension**
  - **2021 CINES renewal**
  - 2022/23 Exascale system (EuroHPC)
- ❑Computing power available
  - TGCC: 23PF (end of 2019)
  - IDRIS: 16 PF (end of 2019)
  - CINES: 3,5PF (2016)
- ❑Complementary architectures and services

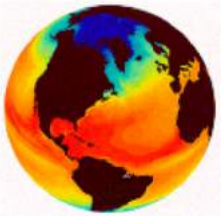
1 CONTEXT

- **« Big data » is the fourth pillar of science**
  - New paradigms in scientific practices
- **Explosion of data volumes**
  - Numerical simulations (high performance computing) and computing power requirements still growing
  - Experiments, observation and monitoring systems (space, land, sea, air)
- **Convergence high performance computing and high throughput data analysis**
  - Complex wide area workflows and their data logistics
  - Software platform of distributed services across a continuum of edge, fog and centralised (Cloud, HPC) infrastructures
- **Different level of maturity across scientific communities**
  - Need to bring communities closer and share beacon of good practices
  - Strengthen new skills and support to users
  - New strategies and architectures… without forgetting energy efficiency !
- **AI is back**

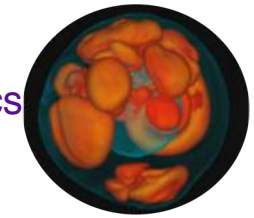
HPC & DATA ANALYTICS

Two strategic tools

For Science

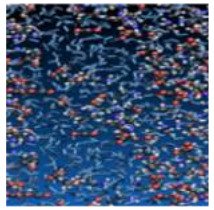


Climate

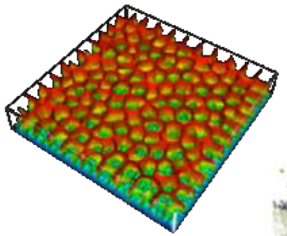


Astrophysics

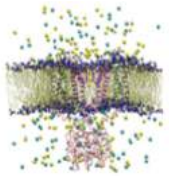
Energy



Chemistry



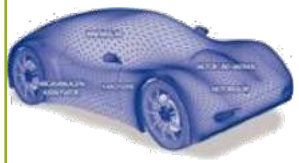
Materials



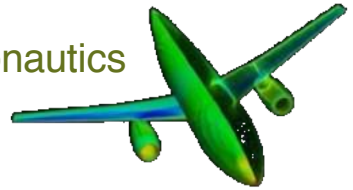
Life Sciences

Humanities

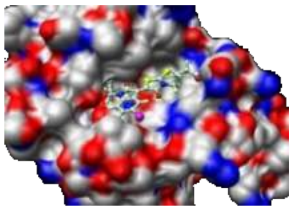
For innovation



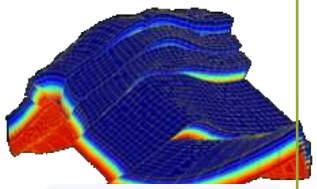
Automotive

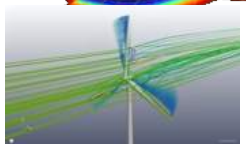


Aeronautics



Pharma

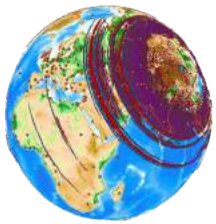




Oil & Gas & Renewables

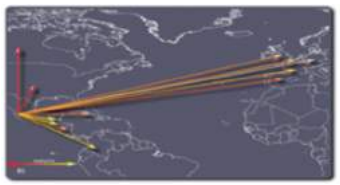
Personlized/ Precision Medecine

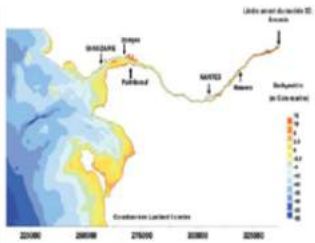
For decision making



Natural risks

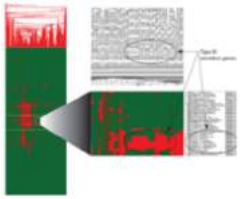
biological and epidemiological risks





Industrial risks

(cyber)Security

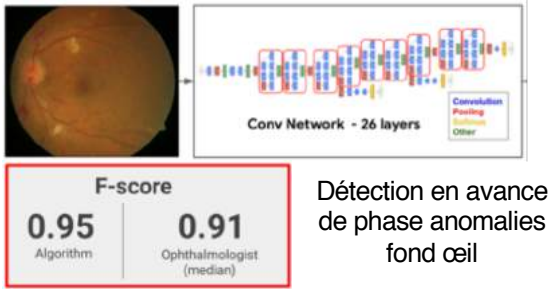


CONVERGENCE HPC – AI

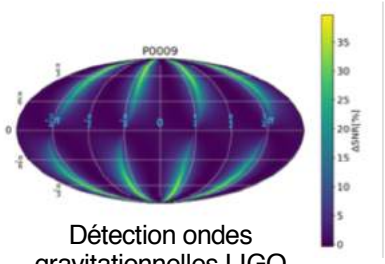
Context

❑Why HPC/HTC need AI ?

- For inferring data flows from large scale scientific instruments
  - Stream access, support of end to end workflows (*edge to tape*)



Détection en avance de phase anomalies fond œil



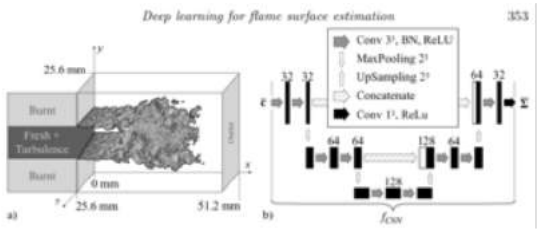
Détection ondes gravitationnelles LIGO

Prévention séismes

- For coupling learnt models and simulation codes (cf Gordon Bell'18) – toward cognitive simulation
  - Interpolation and extrapolation of long trajectory in MD methods, integration of reduced/surrogate models for multiscale , optimisation / UQ (reducing parameter space), forcing of de models (climate), acceleration the convergence of iterative methods, mesh tuning …

Prédication évènements extrêmes climat

Couplage simulation LES combustion et IA



Optimisation crash auto

CONVERGENCE HPC – AI

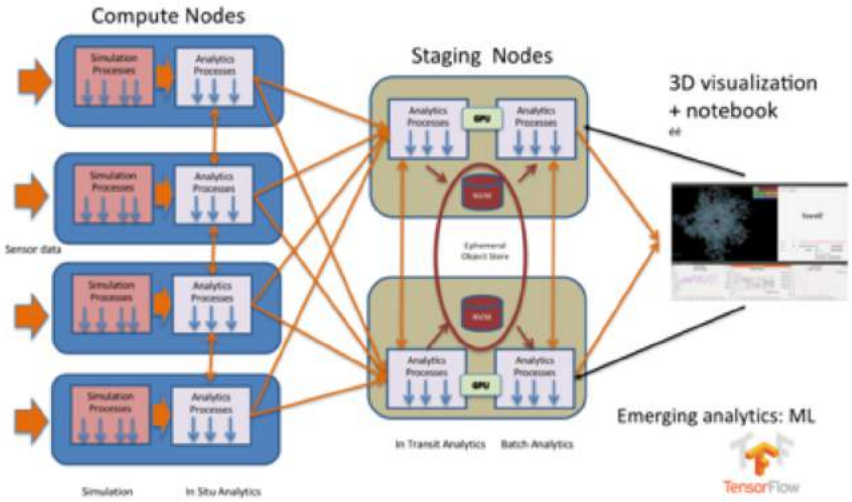
Context

❑Why HPC/HTC need AI ?

- For (in-situ, in-transit) post processing of numerical simulations

simulation → (learning, inference) → storage of pertinent data

smart computational steering



- **Optimise data movement -> minimise energy**

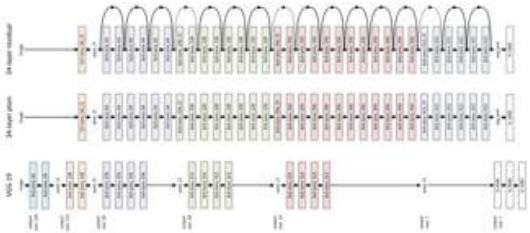
- For better exploiting systems and computing centers
  - AI driven schedulers, improved security, preventive maintenance, optimisation of the infras, …

CONVERGENCE HPC – AI

Context

❑ And AI needs HPC ?

- HPC generates huge amounts of data suitable for AI training
  - Ex : last IPCC campaign in France -> 14 PB of data



- Scale up of the learning phase of neural networks (networks + complex/deep, more data, more classes, …)
  - HPC provide unique 4 levels of parallelism for massive scaling DNN training (© S. Matsuoka)
  1. Hyper parameter search
  2. Data parallelism : different batch data
  3. Domain decomposition : // layer calculations in propagation
  4. Intra chip ILP, vector : // convolutions

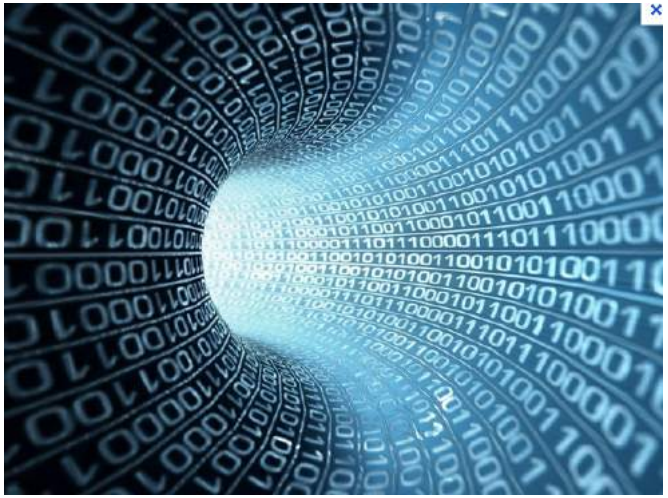
- Auto-tuning of the choice of models (Auto DL/ML, AutoAI), use of federated/transfer learning, …

- Explicability and trust on AI (XAI as a glass box), coupling between formal methods and neural networks, …

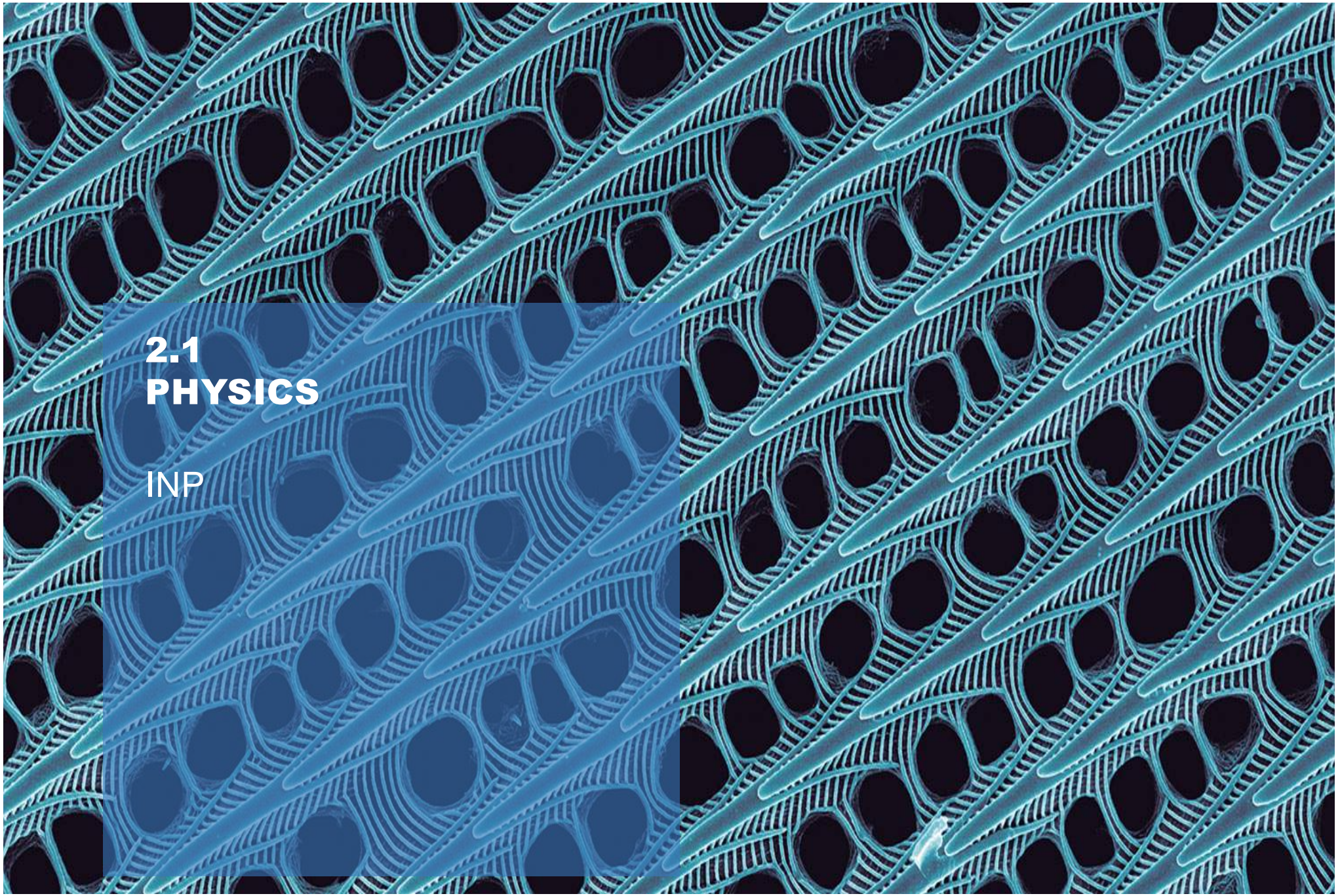
Figure 3 – Les apports de l'IA explicable

2
DATA ANALYTICS

2 SCIENTIFIC DATA AT CNRS



- **Well established practices and stewardship**
  - Nuclear, particle, and astro-particule physics (IN2P3),
  - Earth and space sciences (INSU)
  - Biology (INSB)
  - Humanities and social sciences (SHS)
- **On going structuration:** Ecology and environment (INEE)
- **More varied and less mature on data organization:**
  - Physics (INP), Chemistry (INC), Engineering sciences (INSIS)
  - Expected for some large instruments (synchrotrons,…)
- **Data and data processing are research objects**
  - Mathematics (INSMI)
  - Information sciences and technologies (INS2I)

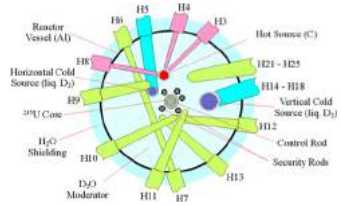


2.1 PHYSICS

INP

2.1 PHYSICS: DATA FROM PHOTON AND NEUTRON FACILITIES

- **Used for protein and large-molecule crystallography, drug discovery, analysing chemicals / geological materiel / …, medical imaging, …**
- **CNRS Teams generate about 500 TB/year on these facilities**
- **European initiatives for data analysis and management: PaNdata, PaNOSC,…**
- **Challenges:**
  - Several scientific communities and devices
  - Increasing volume of data
  - Competency needs for data analysis and management



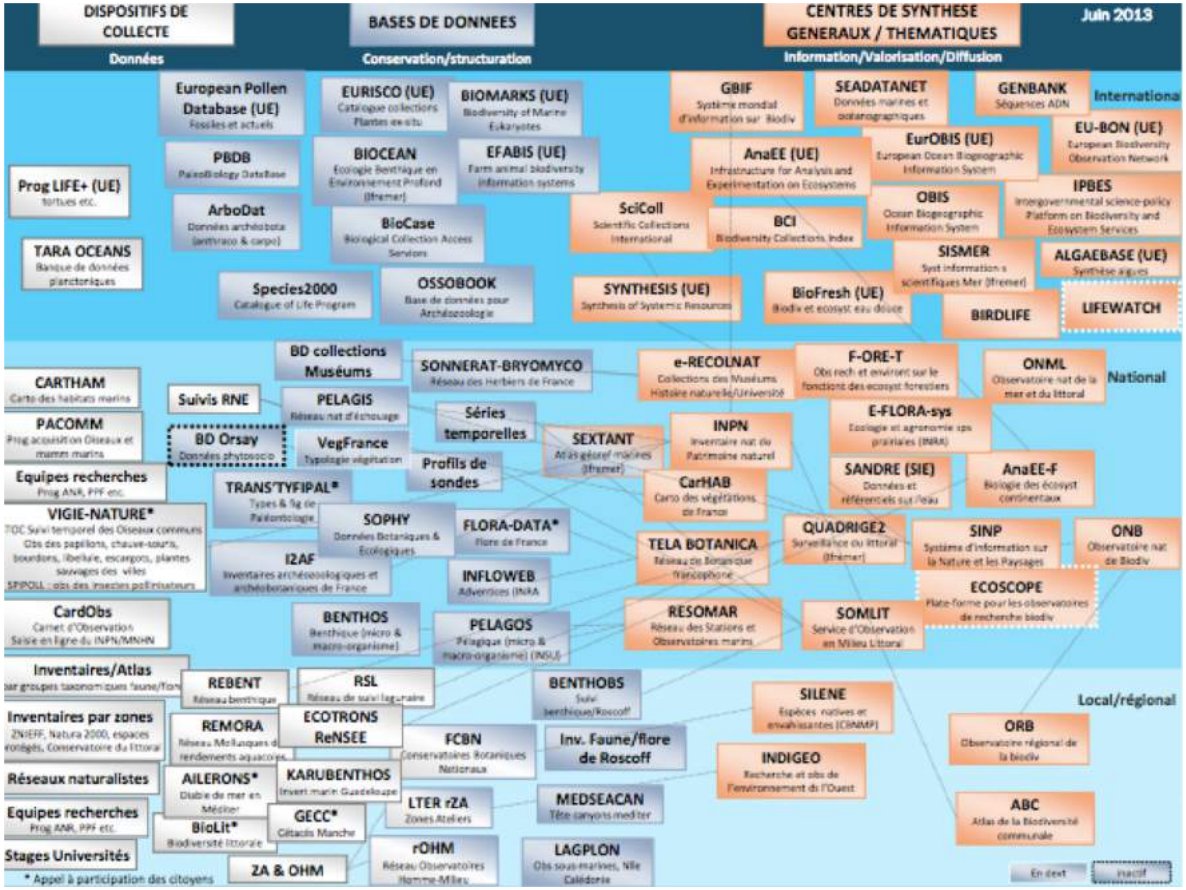
Synchroton	Size	Current production	Future production	Data management
SOLEIL	29 lines, 4600 users, 48 countries	~350 GB/day i.e. ~100 TB / year	**Several PB per year**	Primary, secondary and long term storage with replication (Active Circle)
ESRF	42 lines, 49000 users	2 PB / year	**10-20 PB / year within 5 years**	2 months on disk and 1 year on tape
ILL and Orphée	> 40 instruments, ~1000 users / year	200 TB / year		Data transferred using network to users for analysis. After 2 years data available in ILL catalog.

2.2 ECOLOGY AND ENVIRONMENT

INEE

2.2 ECOLOGY AND ENVIRONMENT – 3 LEVELS:

- **Collecting data**
  - -omics (genomics, …)
  - Collections
  - Observations
  - Experiments
- **Data bases**
  - Organisation
  - Conservation
- **Centers of synthesis**:
  - General / thematic
  - Information / valorisation / diffusion

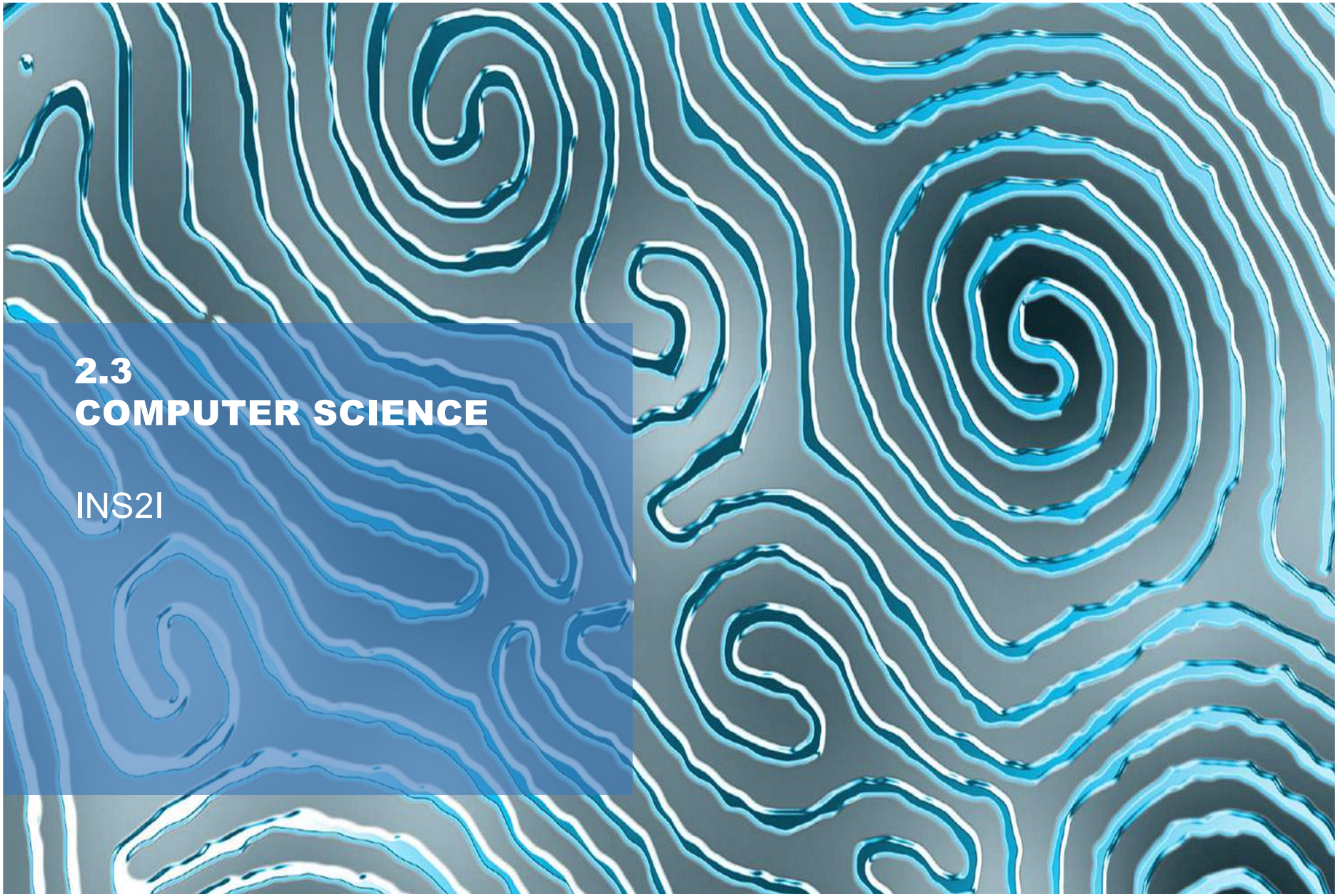


2.2

PATRINAT : « Natural Heritage » (UMS2006)
AFB-CNRS-MNHN

- Organize the interoperability of nature data systems to support their sharing.
- Manage national databanks of species and their conservation status, of habitats, of geology and of protected areas.
- Organize the validation, the diffusion and the valorization of all these data allowing to identify the needs of naturalistic knowledge on the French territories.
- Coordinate the establishment of the National Biodiversity Data Center (PNDB)

Founded in January 2017, PatriNat provides expertise and knowledge management missions through a center of expertise on biodiversity and a data center on nature via the INPN platform which gathers to date, more than 42 million data



2.3 COMPUTER SCIENCE

INS2I

2.3 DATA IN INS2I

- Data Science and AI at the heart of research in INS2I
  - Large scale infrastructures for data processing and AI: High Performance Data Analytics (HPDA)
  - Innovative data processing and analysis methods: collect, indexation, storage, management, exploitation, valorisation, decision-making, accessibility and visualization
  - AI techniques
  - Expected evolutions in: automatization of information extraction processes, better architectures for Big Data and AI, improved security, exploitation of Open Data
- Digital factory, security and user acceptability are some of the important topics

Conclusions of CNRS report on practises related to data (june 2018)

www.cnrs.fr

- **Organization, management, scientific exploitation / valorization of data produced ithin CNRS are major challenges: amounts of data explode**
- **Big Data concerns all CNRS research institutes at various levels**:
  - Well established practices and *stewardship* of data (IN2P3, INSU, INSB, INSHS).
  - Structuration going on to answer to the needs (INEE)
  - Still embryonic consideration of the data except for some TGIR (INC, INP, INSIS)
  - Data and data analytics are research topics (INSMI, INS2I)
- **Logistics and Stewardship of data are two of the main challenges in addition to recruit or train more data scientist**
- **Important issues: interdisciplinarity, platforms, user support, scalable data analysis workflows, multi-sources data, convergence HPC / HDA** (High-End Data Analysis) **in complex workflows**
- **Define new strategies / architectures**
- **Energy efficiency !**

A new paradigm for the research

www.cnrs.fr

- Scientific activity undergoes an epistemological upheaval = > new forms of production of knowledge and emergence of several sub-disciplines.
- **New fields of investigation born at the interfaces of the scientific disciplines:**
  - **bioinformatics**, **computational neurosciences**, **cyber-security**, **digital humanities**, **geoinformatics**, **e-health...**
  - Example : **astroinformatics that incorporates astronomy, astrophysics, computer science and signal processing**
  - …

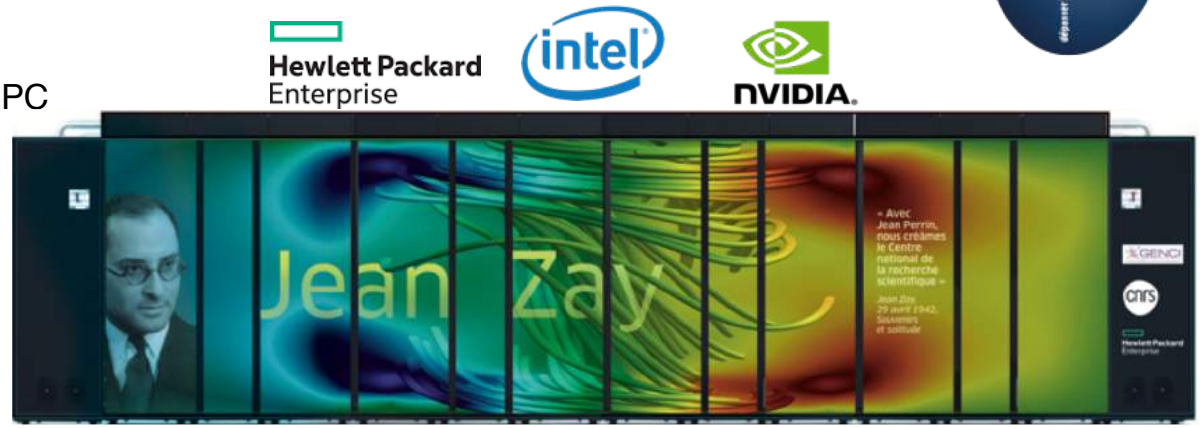
3.
FRENCH NATIONAL PROGRAM FOR AI

3 FRENCH NATIONAL PROGRAM FOR AI

- **National AI initiative announced by President Macron following recommendations of Villani's reports:**
  - 4 Interdisciplinary Artificial Intelligence Institutes selected including the ANITI project in Toulouse
  - 15 Pflops GENCI supercomputer for AI installed at CNRS IDRIS
  - Support to user communities and training experts are crucial
- **Strategy presented by the President focuses on 4 major challenges:**
  - Reinforcing the AI ecosystem in order to attract the very best talents
  - Developing an open data policy, above all in sectors where France already has the potential for excellence, such as healthcare
  - Creating a regulatory and financial framework favouring emergence of "AI champions", through provision of special support to AI research projects and startups
  - Giving thought to AI regulation and ethics, to ensure its development in line with the very best standards of acceptability for citizens

INTRODUCING JEAN ZAY @ IDRIS

One of the biggest converged system in Europe



❑ Objectives

- **Support** with **sovereign** and **leading edge** HPC facilities the French AI research community
- **Foster** synergies between AI and HPC communities
- To be **integrated** into the French AI plan

❑ **Converged** system ?

- HPC + HPDA + AI

❑ New dynamic access modes

- Elastic pool of resources
- Support of containers, notebooks, …

❑Big Milestones

- March 2018 : French "**AI for Humanity**"
- Jan 2019 : Contract between HPE and GENCI announced
- May 2019 : Installation
- July 2019 : Grands challenges
- October : 2019 Full production
- *Start 2020 : 2nd upgrade following ongoing contract of progress*
- *End 2020 : Major upgrade planned*

16PF

❑A balanced architecture

- HPE SGI 8600
- Scalar partition (HPC): 1528 nodes, 3056 CPU CSL 6248, **61 120 cores**, OPA
- Converged partitions – **1292 GPU**
  - 261 thin nodes, **1044 GPU V100 32GB**, 4xOPA
  - 31 fat nodes, **248 GPU V100 32 GB**, 4xOPA

❑Storage

- **1.3 PB @ 360GB/s** full flash (L1)
- **30 PB @ 150GB/s** HDD (L2)
- SpectrumScale parallel filesystem

INTRODUCING JEAN ZAY @ IDRIS

Some recent updates

❑Grand challenges projects

- Partial opening to few teams during final acceptance (3 months)
- More than 300% oversubscription
- 32 projects finally accepted
  - 18 HPC **and 14 AI**
  - From academia and industry
  - HPC on GPU : astrophysics, combustion, material, electromagnetism, medicine

❑1st look on the AI Challenges

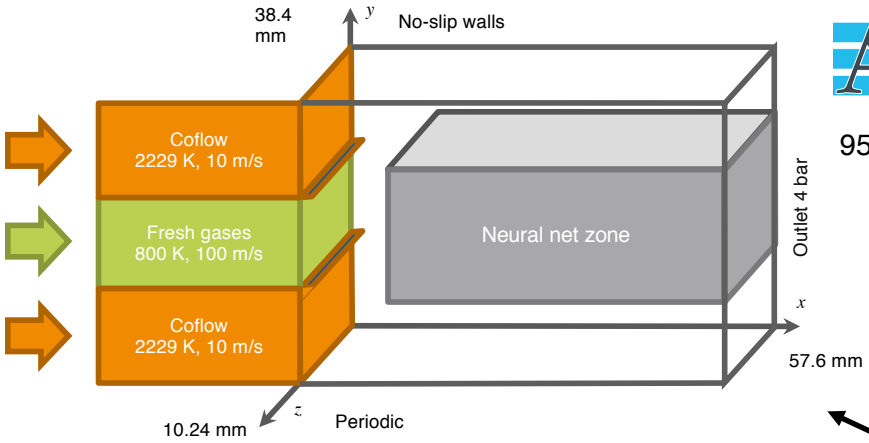
- From 4 to 1044 GPUs
- medicine, vision, detection activity & weak signals, autonomous driving and crash modelling, in-situ post processing + simulation, image, finance, cryptography and security, adversarial attacks, game theory, agents, video, NLP, autoML and new AI models, energy, geology, plant recognition, …

❑ Full opening of the system : November 2019

- Already more than 800% oversubscription of the GPU partition for HPC and AI workloads !

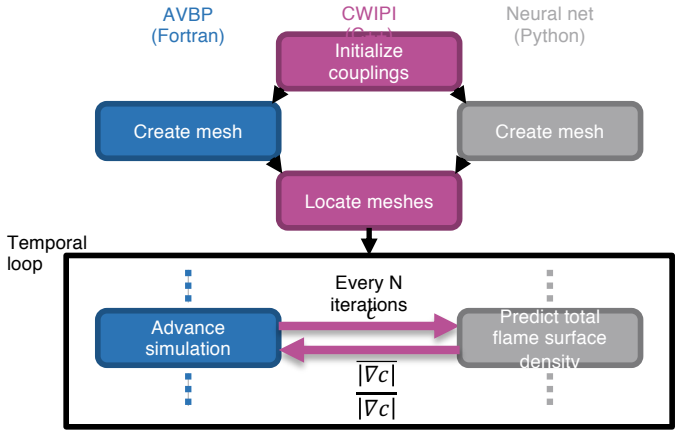
SOME EARLY RESULTS

HPC and AI for high fidelity combustion

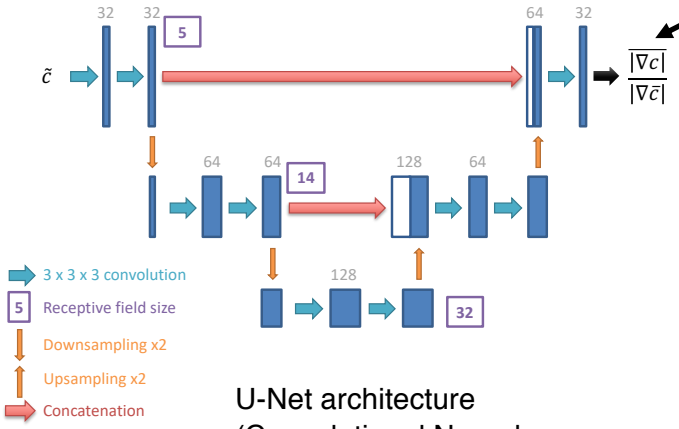


AVBP

95M cells

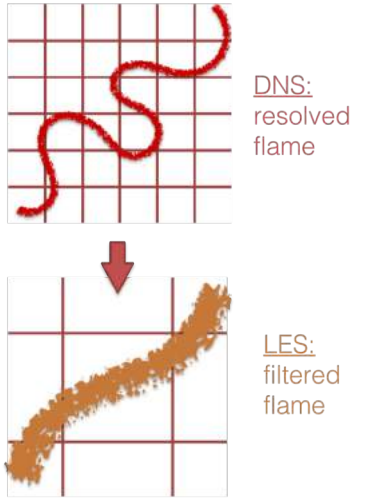


Large-scale LES of a slot jet flame **coupled** with a neural network-based wrinkling model



U-Net architecture (Convolutional Neural Network)

Computing the total surface of the flame front is very important in combustion simulations. In LES this requires a model for the unresolved subgrid-scale wrinkling of the flame front



DNS: resolved flame

LES: filtered flame

First results

- Successful co-simulations **up to 2816 CPUS AND 256 GPUs.**
- First large-scale *a posteriori* test of the network
- Analysis and recommendations for further studies are on-going

SOME EARLY RESULTS

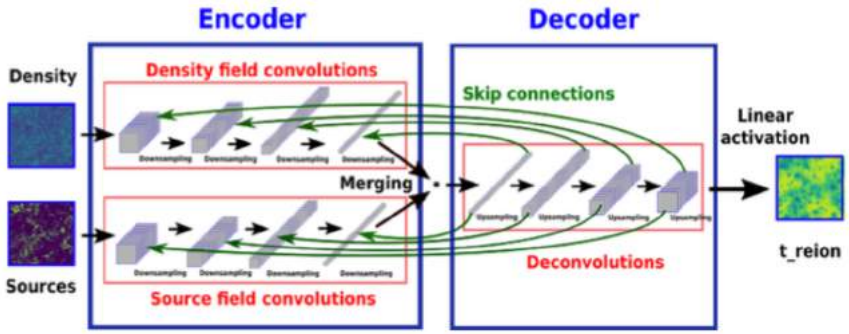
Coupling learn models and simulation models in cosmology

❑Principal Investigator : D. Aubert from Obervatoire de Strasbourg

Observatoire astronomique de Strasbourg

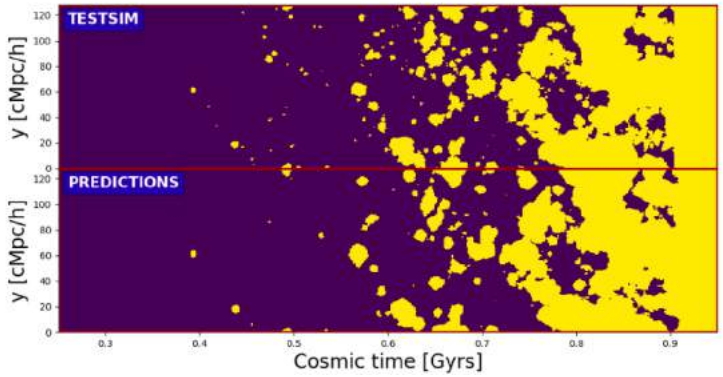
❑Study of the **reonization** of the Universe = 1 Gyrs after Big Bang

- Will be observed soon by instruments like EELT, JWST or SKA
- For the moment only based on massive and costly simulations coupling gravitation, hydrodynamics and radiative transport
- Idea : couple gravitation/hydrodynamics numerical models AND learnt radiative transport models



- Use of auto-encoders based on TensorFlow and Keras

- Methodology already validated for small cubes of 128 Mpc/256^3
- Target = 128 Mpc on meshes of 1024^3



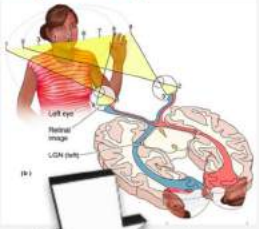
SOME EARLY RESULTS

Using simulation and AI for understanding brain functions

Learning the diffusion MRI signature of brain cytoarchitecture using AI & large scale simulations to decipher the anatomy of brain functions - C. Poupon & K. Ginsburger

human brain

Functional Networks (using fMRI)

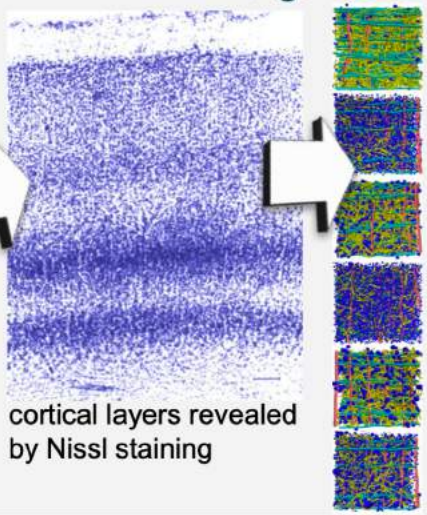


Activation of visual functional areas in the occipital lobe

cerebral cortex populated by neurons and glial cells

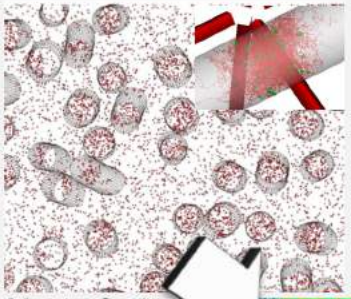
link

Large scale simulation of realistic cortical layers

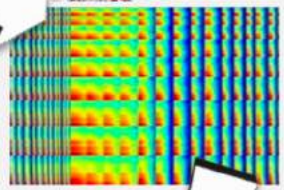


cortical layers revealed by Nissl staining

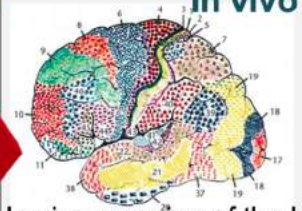
Large scale simulations of virtual cortical layers



Monte-Carlo simulations of Diffusion MRI Signatures for each virtual cortex sample

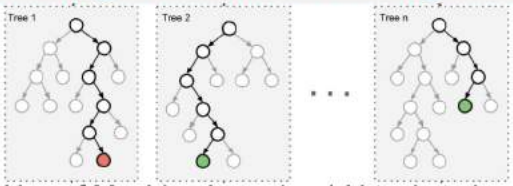


In vivo cytoarchitectonics mapping



In vivo mapping of the human cortex cytoarchitecture using the decoder from MRI acquisitions

Training a Machine Learning decoder



Use of Machine Learning / AI tools to learn the link between the Diffusion MRI signature and the cortex cytoarchitectural features

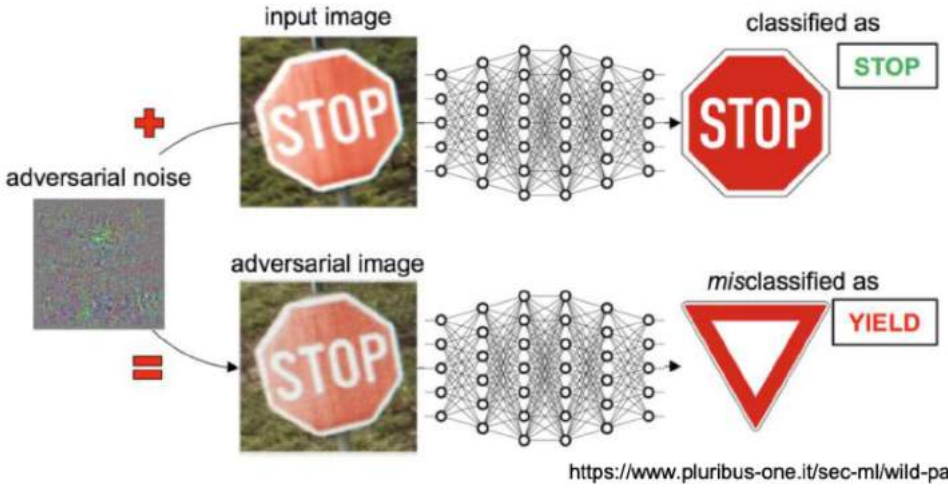
cea Institut des sciences du vivant Frédéric Joliot

université PARIS-SACLAY

GENCI

SOME EARLY RESULTS

HPC for addressing adversarial attacks of neural networks



- Context
  - Neural Networks can be fooled using inputs crafted by malicious users
  - Various attacks exist: $l1$, $l2$, $l\infty$
  - Adversarial Training is a technique to train networks that are robust against one specific attack

- **Goal: Train neural networks that are robust against all attacks**
- Use of Jean Zay
  - Large scale training using CIFAR10 and ImageNet on **up to 100 GPUs**
- Preliminary results from LAMSADE team
  - Randomized Adversarial Training (RAT) is robust against l_∞ AND l_2 type attacks *(Araujo et al. Robust Neural Networks using Randomized Adversarial Training)*

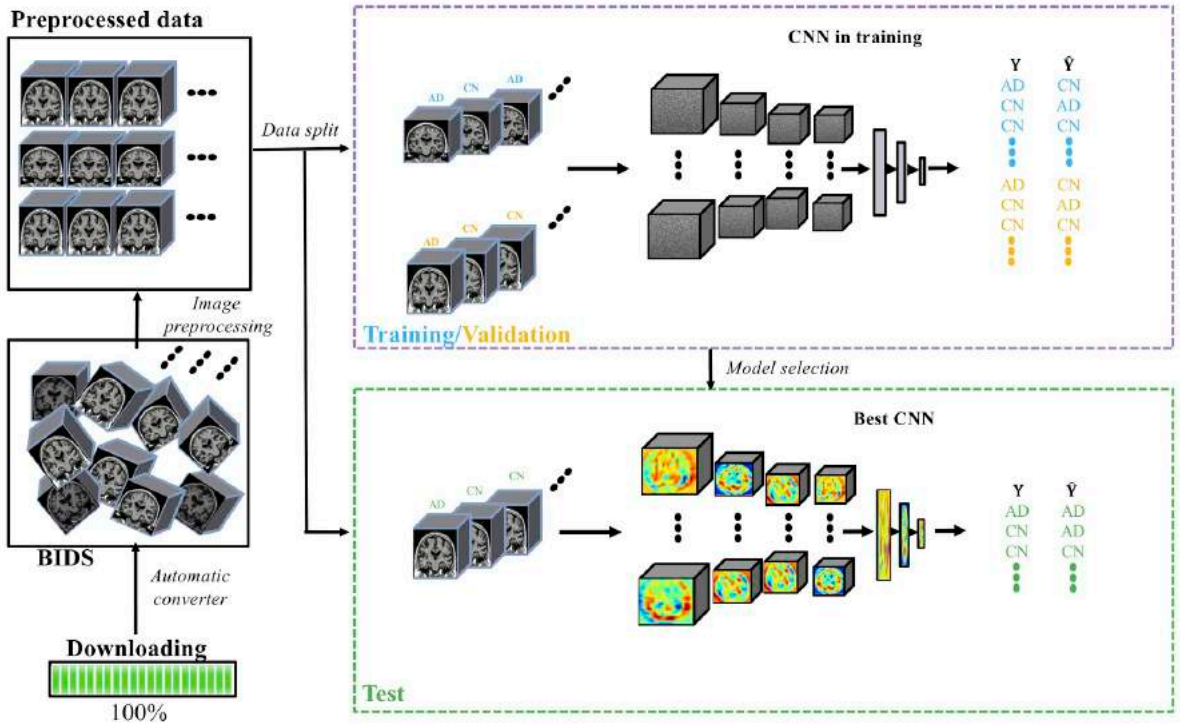
SOME EARLY RESULTS

ADAGE: computer-Aided Diagnosis of AGE-related brain diseases

Principal investigator: Olivier Colliot, ARAMIS Lab (CNRS, Inria, Inserm, Sorbonne, ICM)

Develop and validate deep learning tools for diagnosis from very large scale medical imaging data

Over 20,000 patients
Up to 128 GPUs used



Results:

- prediction of Alzheimer's with high accuracy

Ongoing:

- model of ageing
- diagnosis of other pathologies

ARAMIS LAB
BRAIN DATA SCIENCE

NATIONAL CALL FOR INTERDISCIPLINARY ARTIFICIAL INTELLIGENCE INSTITUTES

www.cnrs.fr

- Following call for proposals 12 applications issued including the one from Toulouse
- 4 Interdisciplinary Artificial Intelligence Institutes selected:
  - Prairie in Paris
  - Grenoble
  - Nice
  - Toulouse
  - Guidelines: 1/3 funding from industry, 1/3 from governement, 1/3 from valorization of academic research participation
  - Budget 100 M€ from MESRI (20-25M€ per project)

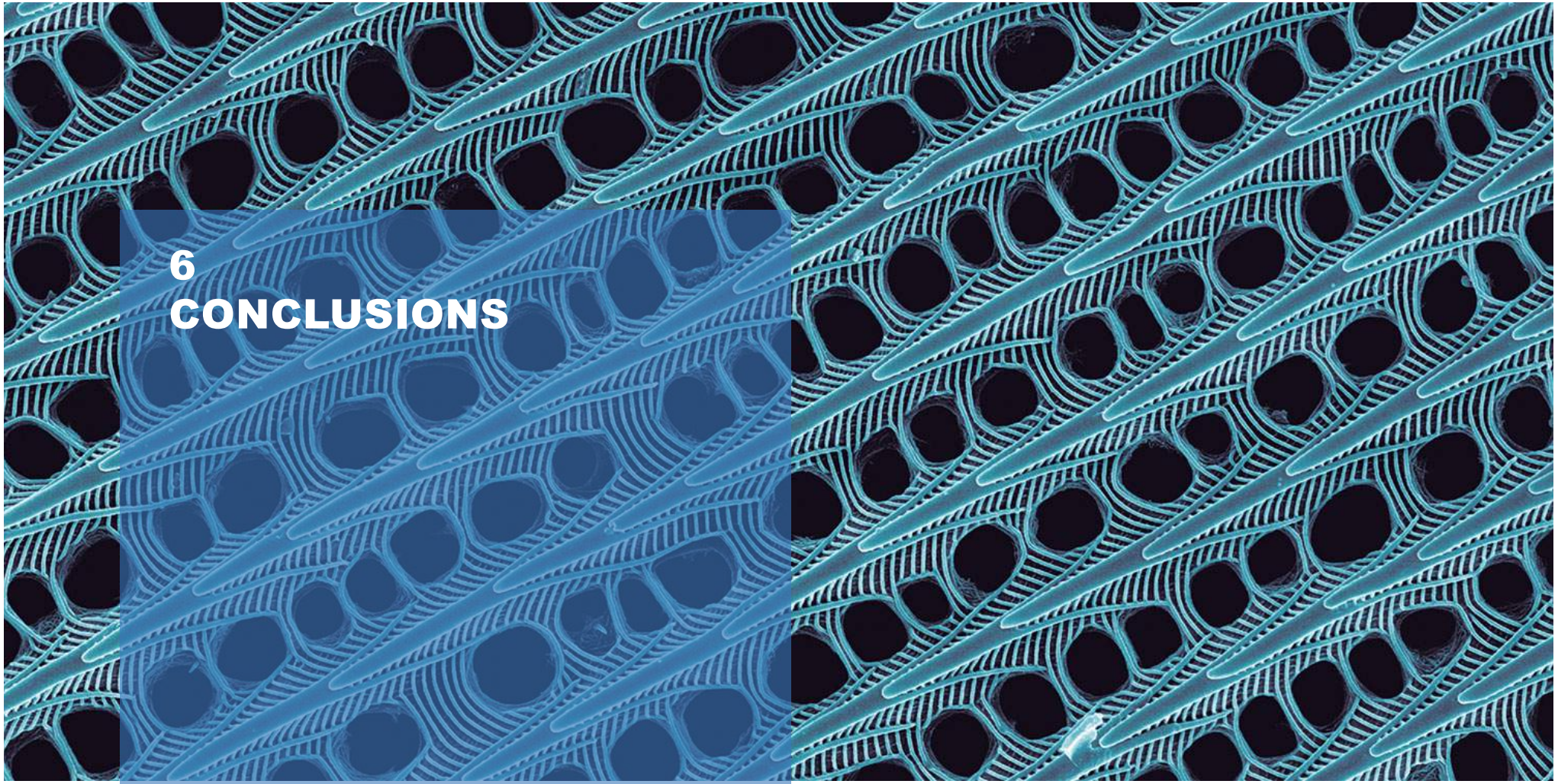
ANITI INTERDISCIPLINARY ARTIFICIAL INTELLIGENCE INSTITUTE IN TOULOUSE

- Hybrid AI with main application in transports

ANITI: SOME NUMBERS

- **2 strategical application areas: mobility and transportation**
- **+200 researchers**
- **3 integrative programs:**
  - Acceptability, fair representative data for AI
  - Certifiable AI toward autonomous critical systems
  - Assistants for design, decision and optimized industry processes
- **+50 partners**
- **100 M€ budget**

www.cnrs.fr



6
CONCLUSIONS

AND NEXT ?

Paving the path toward converged Exascale

- **Mission:** Establish an integrated world-class supercomputing and data infrastructure and support a highly competitive and innovative HPC and Big Data ecosystem
- **Objectives**
  1. ***An integrated world-class supercomputing and data infrastructure***
     - 2 pre-exascale + 2-3 petascale by 2020; 2 exascale by 2022/2023 (1 EU tech); post-exascale infrastructure by 2027
     - federation of HPC infrastructures at European level
     - hybrid HPC/Quantum infrastructure
  2. ***Research and innovation for a HPC and Big Data ecosystem***
     - an integrated European HPC R&I agenda
     - independent HPC technology supply
     - excellence in HPC applications and use
     - HPC Competence Centres, training/skills, outreach

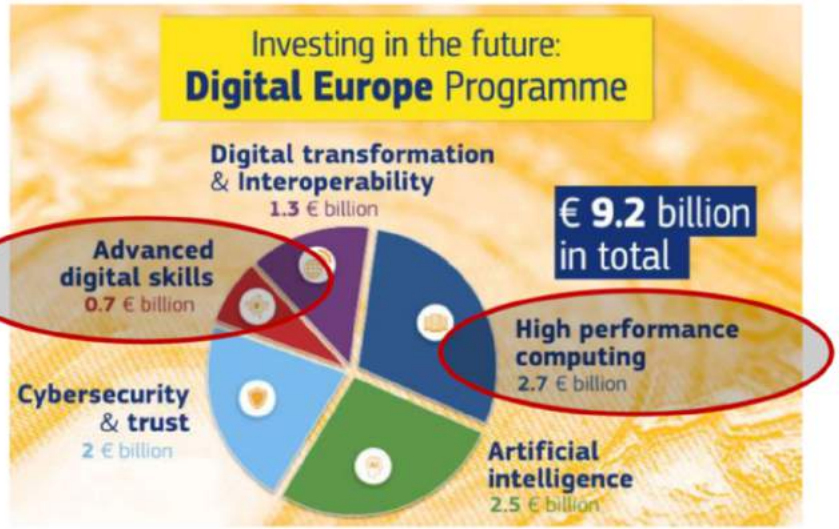
Infrastructure & Operations

R&I: Tech,Apps &Skills

HPC Ecosystem

- >1 Md€ allocated for 1st phase (Pre Exascale)
- Target of 3 Md€ in FP9 for 2nd phase (Exascale)

as Hosting Entity ?



2019/20
Jean Zay
14 to >30 PF

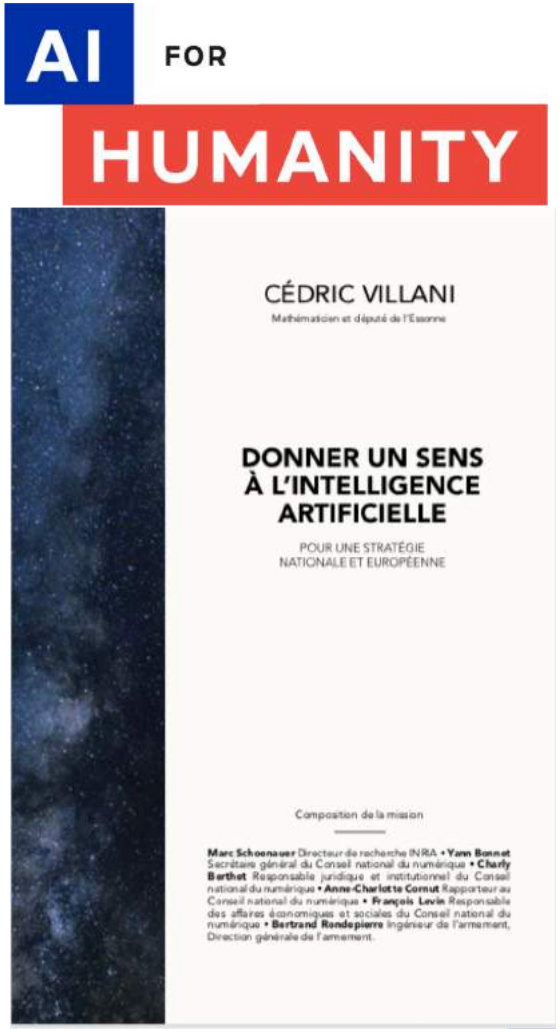
Assessment / deployment of new architectures

Training / user support, new services

2022/23
Converged Exascale
>x00 PF

BUILDING A WHOLE ECOSYSTEM IS MANDATORY

- **HPC + Big Data + AI not only a ressource management problem but a *change of paradigm in scientific research.* Need for a new holistic approach built around scientific challenges:**
  - Inter/pluridisciplinarity (computer science, mathematics and other disciplines),
  - New methods / algorithms : software challenges
  - Compute / data infrastructures in synergy with large-scale instruments, experimental platforms and observation systems
- **Support to user communities and training experts are crucial**



Thank for your attention