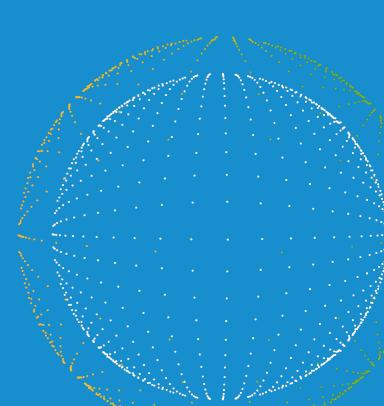
Data Spaces Symposium HPC meets the demands of Al and Data

Jeanette Nilsson RISE

Different perspective on AI, Data and HPC- connecting SME and AI community

- Juan Pelegrin, EC DG CNECT
- Roberta Turra, Cineca
- Andreas Wierse, SICOS
- Kristina Knaving, RISE

Jeanette Nilsson Al/ HPC-Ecosystem expert BDVA, Member of the Board of Directors of RIAG member Sweden's representative in the EU's Al and Business Digitalization Working Group with responsibility for the Partnership for Al, Data and Robotics (Adra)





Al Factories - Amendment to the EuroHPC JU Regulation Council Regulation (EU) 2021/1173

Juan Pelegrin DG Communications Networks, Content and Technology European Commission

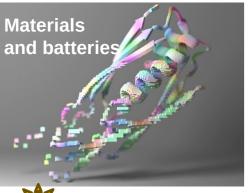
12 March 2024

Large Artificial Intelligence (AI) models

- Large AI models (such as large language models) are a new wave of AI models called generative AI adaptable to various domains and tasks.
- Generative AI models have immense potential to revolutionise multiple sectors.
- Most large AI models (e.g. ChatGPT) are non-European
- Mastery of this technology is of strategic importance for Europe in line with economic security.

Key AI applications







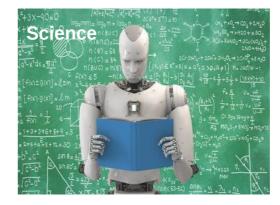


















Protein structure prediction The greatest achievement of AI - and a paradigm shift in science





DeepMind

DeepMind AI cracks 50-year-old problem of protein folding

Program solves scientific problem in 'stunning advance' for understanding machinery of life

DEEPMIND, DE GOOGLE

La IA revela la forma de las proteínas en 3D y abre la puerta a crear nuevos fármacos

V+

 El programa AlphaFold2 resuelve un problema fundamental de la biología de hace 50 años: cómo adquieren su estructura las proteínas



Premio Princesa de Asturias de Investigación Científica y Técnica 2022



Geoffrey Hinton, Yann LeCun, Yoshua Bengio y Demis Hassabis

Premio Fronteras del Conocimiento



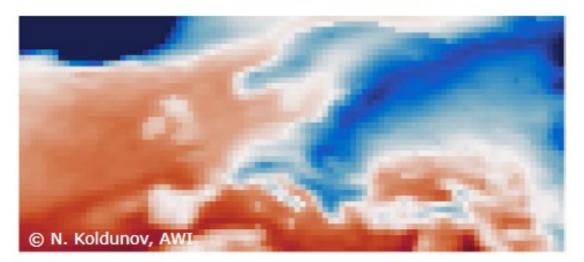
Demis Hassabis, David Baker y John Jumper.





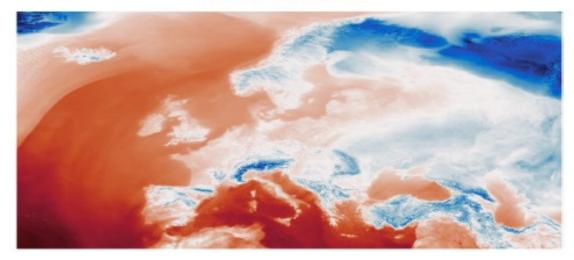
Funded by the European Union

CLIMATE DT: OPERATIONALIZING GLOBAL MULTI-DECADAL CLIMATE **PROJECTIONS AT KM-SCALE – WORKFLOWS SET-UP ON LUMI**



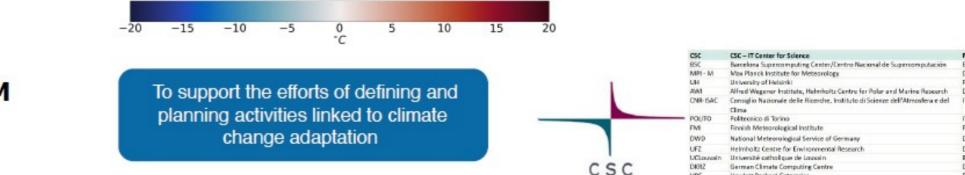
IPCC AR6 (2021), 100km

Digital Twin, 5km



D4/87

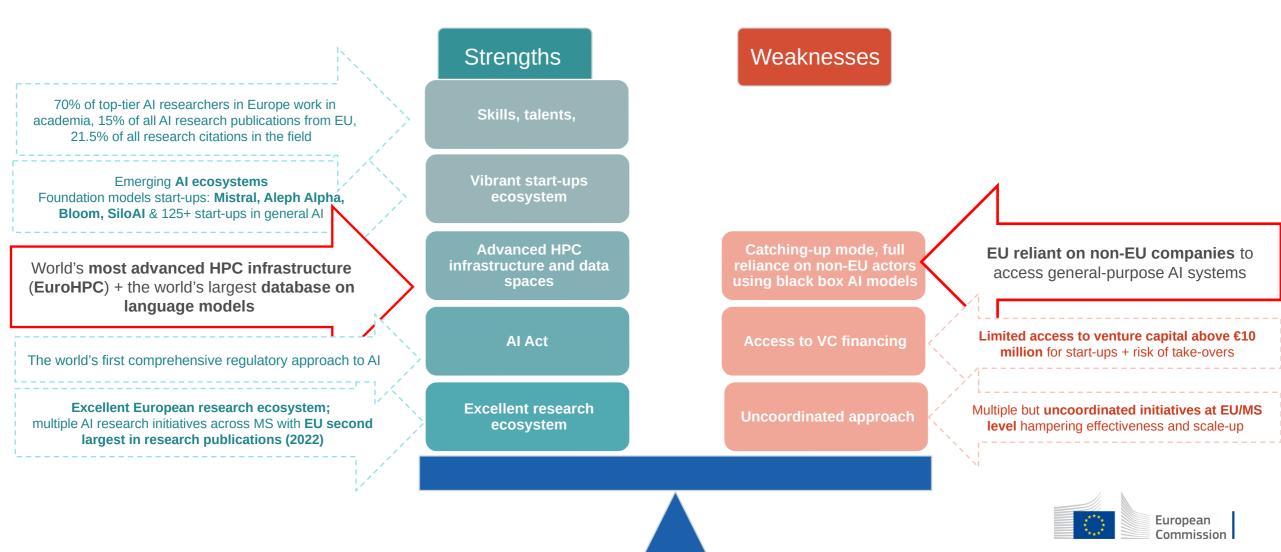
German Climate Computing Centre Hewlett Packard Enterprise



ICON **IFS-NEMO/FESOM**

CECMWF

Towards large AI models: European strengths and weaknesses



Context

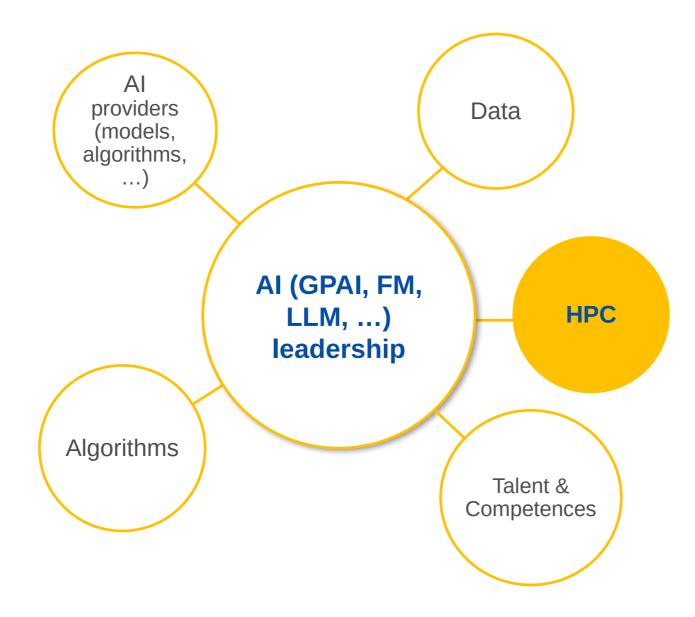
- **Generative AI** and **AI foundation models** are advancing at unprecedented pace and are set to play a pivotal role in shaping the future of technology and society.
- Al is a key policy area of the EU digital strategy. Al in combination with HPC can contribute to a more innovative, efficient, sustainable and competitive economy, while also improving safety, education and healthcare for citizens.



In her 2023 State of the Union address, President von der Leyen announced that **the supercomputing resources of the EuroHPC JU will be made available to European AI startups to train their large-scale models**, contributing to the EU's aim of leading global advances in AI and of achieving responsible and ethical innovation.

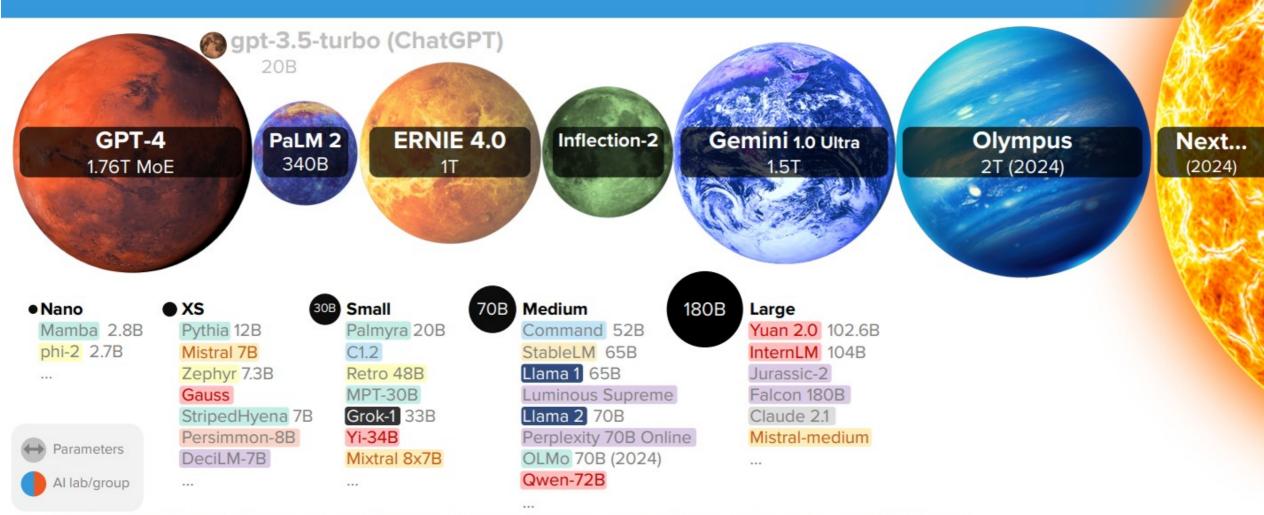


Unlocking the potential of AI





LARGE LANGUAGE MODEL HIGHLIGHTS (DEC/2023)



Sizes linear to scale. Selected highlights only. All models are available. All models are Chinchilla-aligned (20:1 tokens:parameters) https://lifearchitect.ai/chinchilla/ All 200+ models: https://lifearchitect.ai/models-table/ Alan D. Thompson. 2023.

LifeArchitect.ai/models

EuroHPC for large AI models (estimation)

Olympus 2T (2024)	HPC system / Al model	Meluxina - LU (18 petaflops peak performance)	LUMI - FI (550 petaflops peak performance)	Leonardo – IT* (323 petaflops peak performance)	Jupiter-DE Approx. figures
30 BN	Falcon 40 bn parameters	50 days	4 days	3 days	1 day
70 BN	LLaMA (META) 65 bn parameters	80 days	7 days	5 days	1.5 days
200 BN	GPT-4 (OpenAl) Hundreds of bn parameters	6 years	6 months	4 months	6 weeks

Source: Internal analysis



How to capitalise on EuroHPC strengths to develop a highly competitive and innovative AI start-up and research ecosystem in Europe?

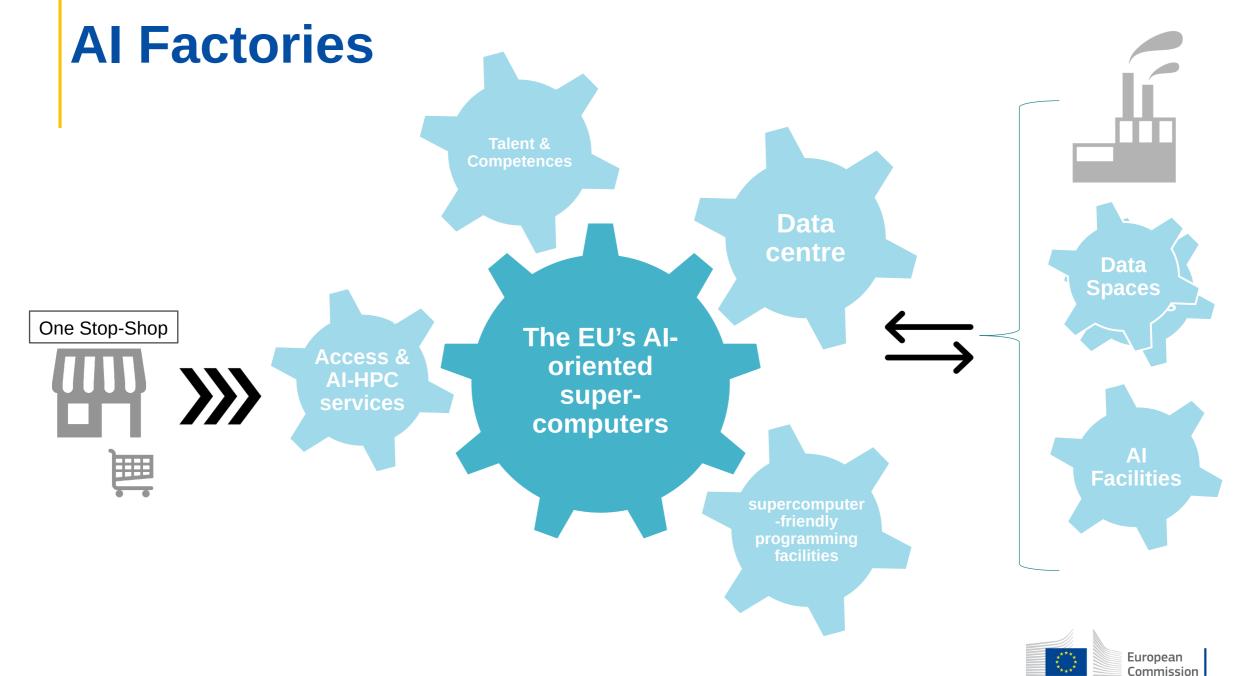
- Procurement of supercomputers.
- Upgrading of existing EuroHPC supercomputers.
- Access to supercomputers.

BUT:

- Current EuroHPC infrastructure and services not optimized for AI.
- Inflexible conditions for upgrading.
- Not considering specific needs of AI user communities.
- No connection with AI ecosystem.

⇒ A targeted amendment to the EuroHPC Regulation is necessary.





Amendments to EuroHPC Regulation (EU) 2021/1173

AI Factory - Activities

- **1.** Acquisition & operation of AI dedicated supercomputers (co-located with data centre)
- **2.** Upgrading with AI existing EuroHPC
- **3.** Provide access to SMEs and start-ups (incl. widening usage)
- **4.** AI supercomputing service centre (algorithms, training- testing- evaluation- validation of AI models, development of large-scale AI applications, ...)
- **5.** Supercomputer-friendly programming facilities (parallelization, usage optimization, ...)
- **6.** Attracting & pooling talent
- 7. Interacting with AI-ecosystem at large & other AI initiatives

Hosting AI supercomputers conditional of establishing AI Factory

European

AI Factory

• Hosting Entity

• AI-Factory related Hosting Entity eligibility criteria / Hosting Entities => One-Stop-Shop

• Hosting Agreement

- Specific hosting conditions for AI-dedicated supercomputers / AI-upgrades
- Acquisition and ownership of Artificial Intelligence-dedicated supercomputers
 - Apply for hosting anytime => no limitations as for high-end/mid-range

• Upgrading of Supercomputers

- Deleting € limit/ Widening scope to request upgrades/ Apply for upgrade anytime => no limitations to apply within 3 years after selection as hosting entity
- EU financial contributions (OPEX/CAPEX) percentage as for original supercomputer to be upgraded

• Use of EuroHPC Supercomputers

• Governing Board defines special access conditions for AI supercomputers / upgrades

• Allocation of Union Access time to EuroHPC supercomputers

• EU share of access time proportional to EU contribution



THANKS



DATAWEEK²⁴

JOIN.LEARN.SHARE.GET VALUE

HPC meets the demands of Al and Data Leonardo: Bridging Al Needs, Access, and HPC

12/03/2024 15:30-16:30 (CET)

Roberta Turra, CINECA

under the umbrella of: Data Spaces Symposium Unite. Innovate. Adopt.



 inded by
 The Data Spaces Support Centre receives funding from the European Union Digital Europe Programme

 e European Union
 under grant agreement n° 101083412



The impact of AI on HPC can be seen from different perspectives

- Tasks / Requirements
- Users
- Access to resources
- Hardware
- Software
- Projects / Apps
- Services





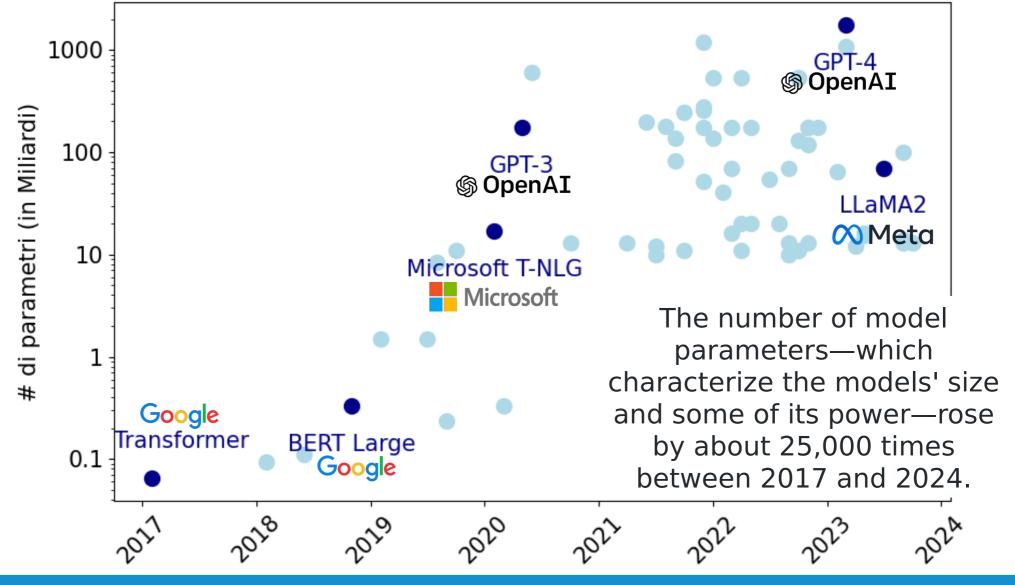
Tasks / requirements
new in HPC (impact
of AI)

- Run physical (deterministic) models that generate data
- Applications: SIMULATIONS (Computational Fluido Dynamics, Molecular Dynamics, ...)

- Data driven ML algorithms (probabilistic) that generate the model
- New workload need optimization of HPC systems for AI tasks
- New applications on texts and images
- With Deep Learning computational power becomes one key element (data – algorithms – compute)



LLM's evolution





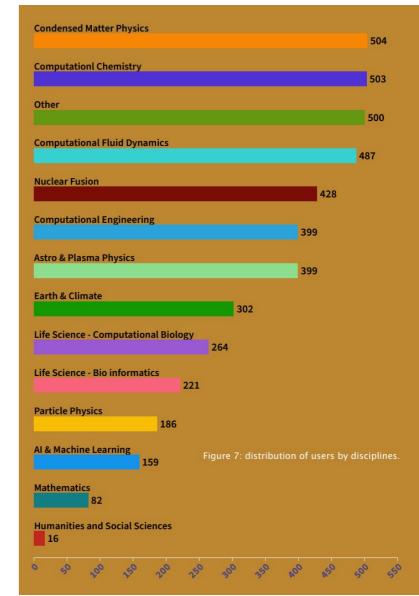
Users

Traditional HPC

- Domain: Natural sciences
- Main typologies:
- Researchers (scientific communities)
- Big companies (e.g. pharmaceutical, Oil and Gas, ...)

What's new in HPC (impact of AI)

- Domain: Social sciences (in the beginning)
- No single community
- New typology: startups



CINECA HPC Annual Report 2022/23 Data Spaces Symposium Unite. Innovate. Adopt.



Access to resources's new in HPC (impact

Traditional HPC

- Scientific merit (with limitations on the number of core/hours)
- Contractual / research agreements
- Public funding (e.g. EU projects)

Specific calls

of AI)

- EuroHPC
- AI-BOOST
- ALT-EDIC
- EuroCC2, FF5, EDIHs, ... (in Italy the National Competence Center for HPC, ICSC)
- Specific agreements (in Cineca: FAIR, FBK, AL, IBM, iGenius)



Hardware

Traditional HPC

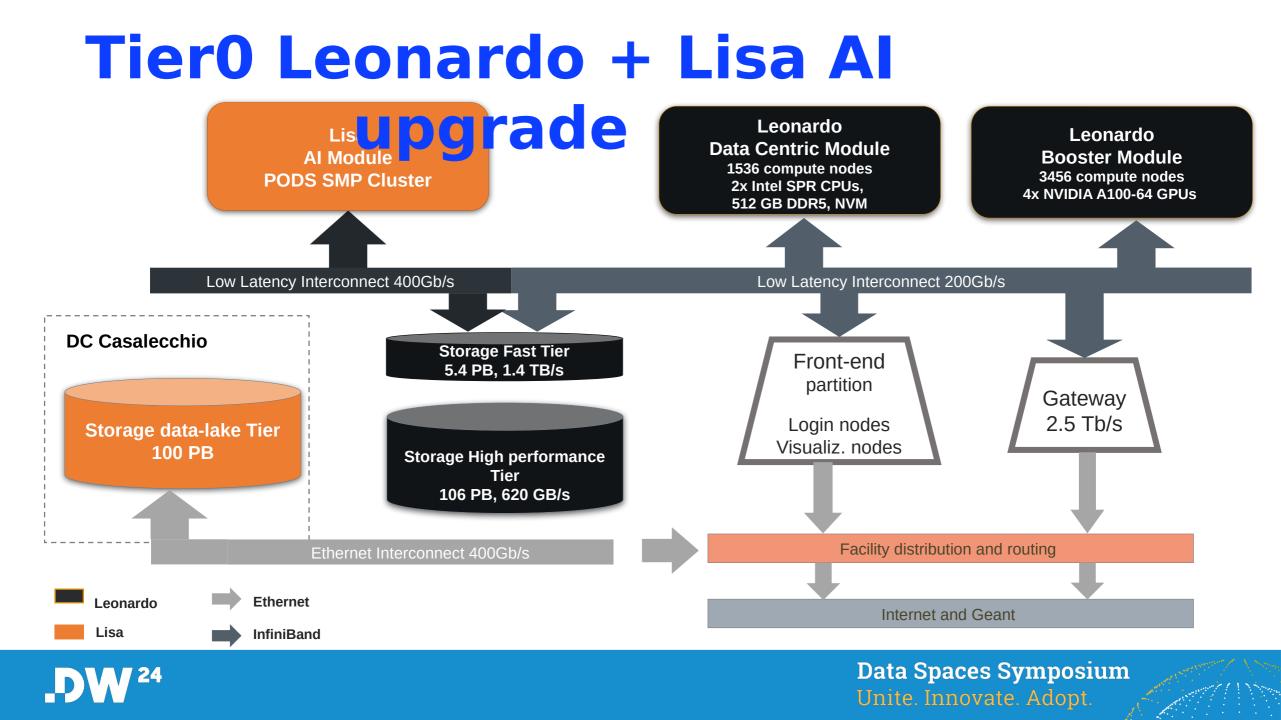
What's new in HPC (impact of AI)

- CPUs
- Double precision

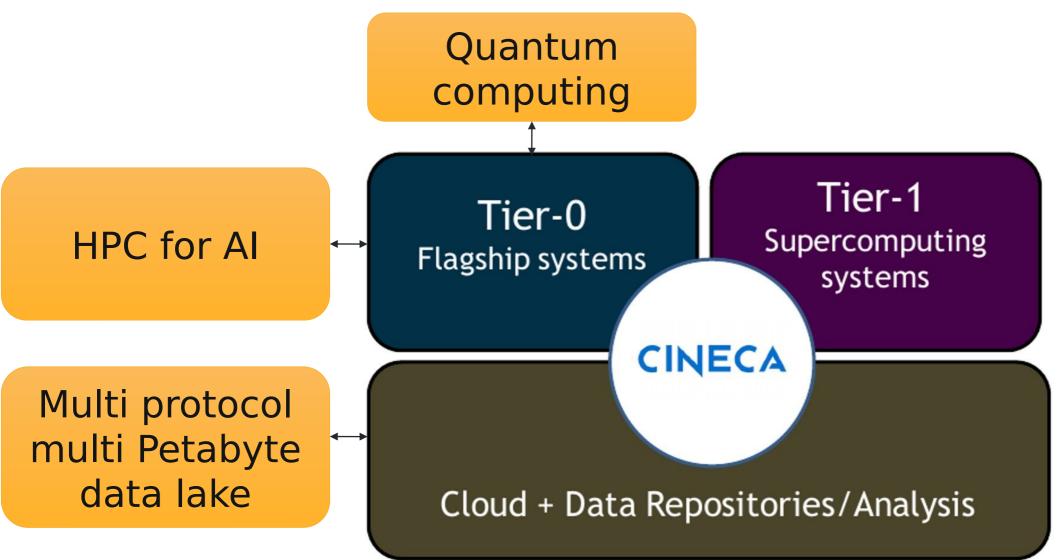
- GPUs
- Bandwidth
- (fast)Storage

Al and machine learning workloads have driven the development and adoption of specialized hardware such as GPUs (Graphics Processing Units), TPUs (Tensor Processing Units), and FPGAs (Field-Programmable Gate Arrays). These **accelerators** are designed to perform parallel operations at high speeds, essential for the massive matrix and vector computations common in Al. This shift has led to a rethinking of HPC infrastructure to accommodate these accelerators, leading to more **heterogeneous computing environments**.



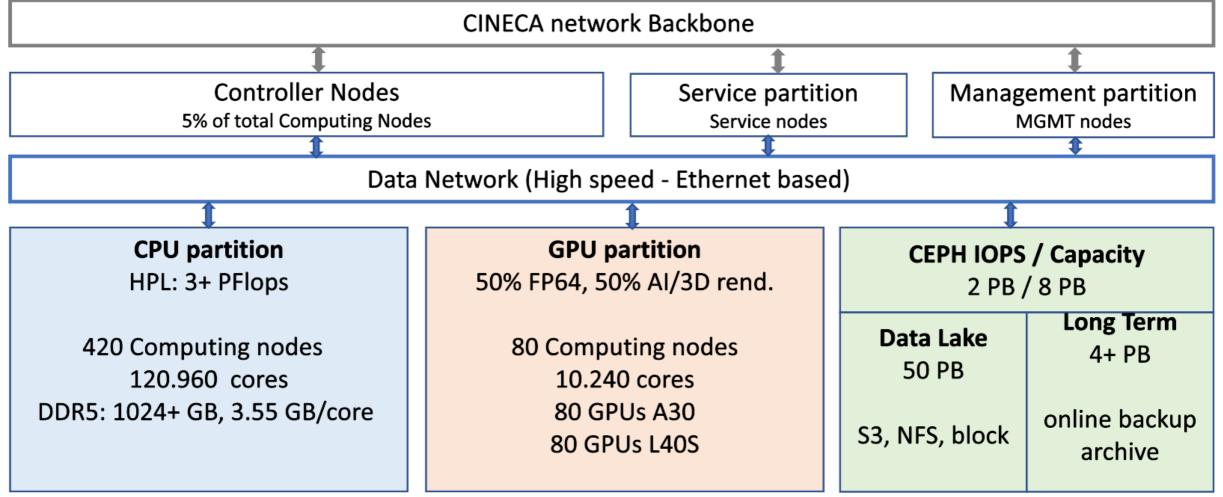


HPC infrastructure





Cloud service evolution





Data Network



Cloud

- Flexible use of resources (dynamic resource allocation, allowing users to scale resources up or down on-demand)
- Interactive Computing Services (enable to provide front-end services for the users, web access, real-time data processing, graphical interfaces for inference ...)
- Environment personalization (operating system and software stack can be selected and installed by the user)
- Security (GDPR compliant)



Software

Traditional HPC

- Specialized software libraries, compilers, and tools optimized for performance on specific hardware architectures (e.g. Quantum espresso)
- Repositories

What's new in HPC (impact of AI)

- Pytorch, tensorflow
- New libraries
- New modules (e.g. CinecaAI)

• Data Lake



Projects / Apps Traditional HPC

- Particle Physics
- Condensed Matter Physics
- •Astrophysics and Plasma Physics
- •Computational Chemistry
- Computational Engineering
- Computational Fluid Dynamics
- •Earth and Climate Science
- •Life Science Computational biology
- •Life Science Bioinformatics

What's new in HPC (impact of AI)

• A new scientific area in project classification: AI & Machine learning

 Scientific and innovation (commercial) projects that either develop new AI algorithms or use AI in their application domain



Projects / App

National research projects (ISCRA C) running on Cineca infrastructure in

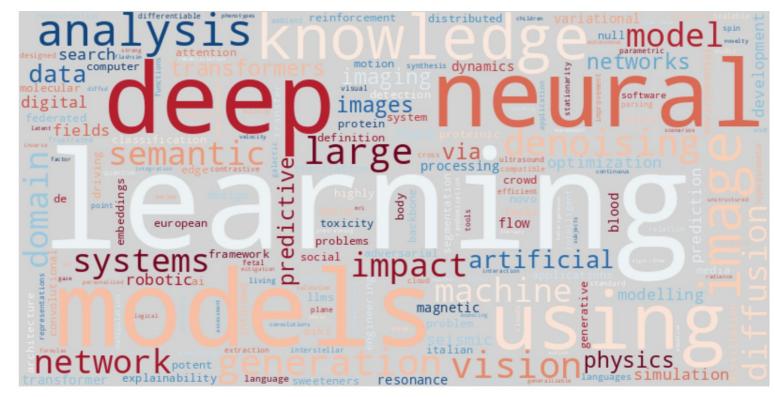
Project focus	# of abstracts	
AI & ML	101	
Big Data/Bioinformatics	39	
COVID-19	7	
HPC	581	
Quantum Computing	18	
Total	746	

of abstracts for each cluster 9protein, molecule 8 code, software 7 -6 -Cluster anguage, Ilm 5 noise, seismic 3 -2 -1 -0 10 15 19 1 5 # of abstracts



Projects / Apps

Word cloud in the ISCRA C project titles





Climate and Weather

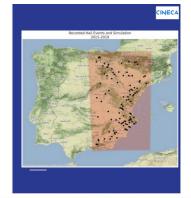
- Day-ahead hail prediction to prevent losses in agriculture
- DAMAS (https://euhubs4data.eu/experiments/damas/) Data-driven Model for the Analysis of Sea-state
- Climate Byte (https://euhubs4data.eu/experiments/climaterisk-assessment-for-agriculture-insurance-using-big-dataand-artificial-intelligence/)

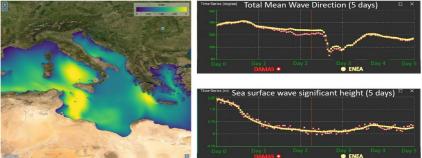
Climate risk assessment for agriculture insurance

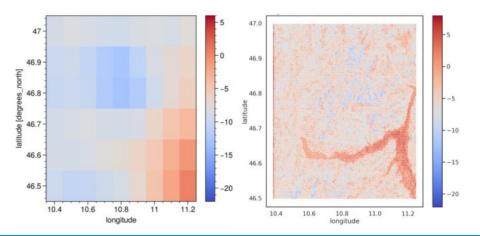
ILLUMIA

(https://www.ifabfoundation.org/ifab-activities/projects/aigeneral-circulation-model/)

- AI General Circulation Model
- OptimESM (https://optimesm-he.eu/) Optimal High Resolution Earth System Models for Exploring Future Climate Change
- Weather4Energy (https://www.supercomputing-icsc.it/en/spoke-4-earthclimate-en/)



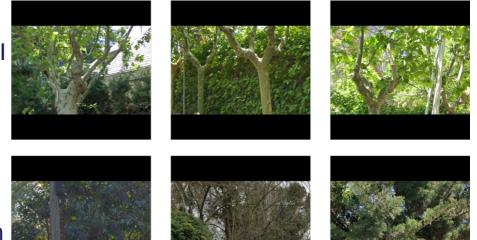




Earth Observation

Arboria https://euhobs4data.eu/experiments/arboria/) An open-data based urban forest solution improved by AI optimization

- Digital Twin of Bologna city
- Mobility Square (https://euhubs4data.eu/experiments/sustainablemobility-recommendation-engine-for-tourism/) Sustainable mobility recommendation engine for tourism







Other relevant ongoing projects

Cataloguing and meta-dating of artistic heritage

Support the Ministry of Culture in the cataloguing and metadating of artistic heritage through the use of image analysis techniques (object detection, classification, captioning) Legislative support using artificial intelligence and LLMs

Design and implementation of a semi-automatic laws impact evaluation system that can generate new high quality text laws and improve impact

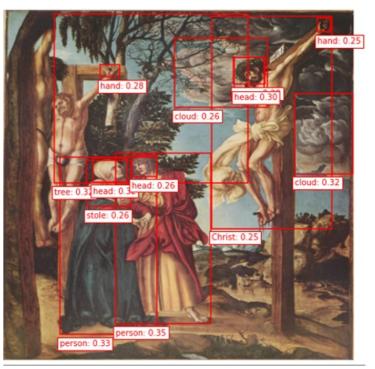
Retrieval Augmented Generation for Systematic Literature Review

Natural Language Processing for healthy life expectancy



Data Centre Digital Twin for Sustainable Exascale Computing Design massive Data Center graph models capturing the spatiotemporal dependencies

Design massive Data Center graph models capturing the spatiotemporal dependencies between computation, nodes, and cooling equipment and conduct analytics to predict the impact of the spatial power distribution on cooling efficiency and cost (https://graph-massivizer.eu/project/data-center-digital-twin/)



Services

Traditional HPC

- User support
- Specialistic support
- Training
- PoC

What's new in HPC (impact of AI)

- Data catalogue
- Free inference for widely adopted open-source models
- Fine-tuning

• ...



Challenges

• Ethics

Sustainability

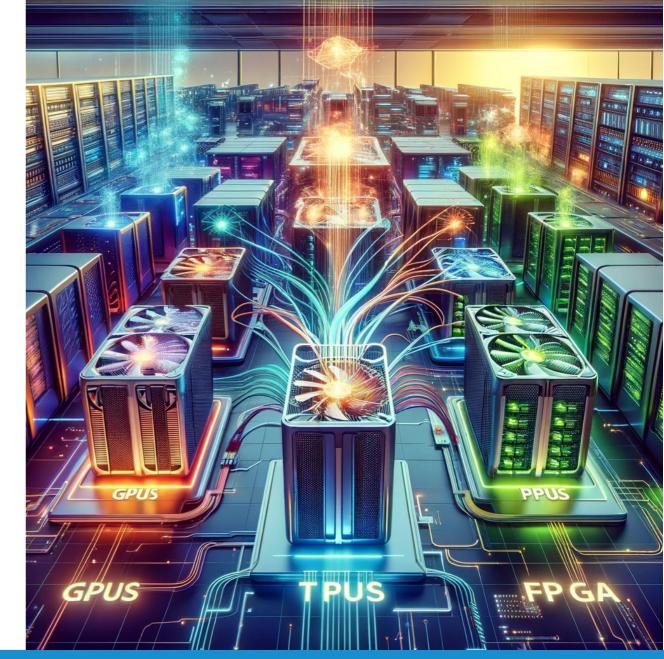


Data Spaces Symposium Unite. Innovate. Adopt.



What's next

- The transformation is going on
- We learn by doing, by actively participating to AI projects and by providing user support in the most computationally challenging tasks (e.g. LLM)





Data Spaces Symposium Unite. Innovate. Adopt.



Thank you!

65

DATA SPACES

SUPPORT CENTRE

Funded by The Data Spaces Support Centre receives funding from the European Union Digital Europe Programme the European Union under grant agreement n° 101083412



Data Spaces Symposium Unite. Innovate. Adopt. 🦯

Darmstadtium | Frankfurt region



Wir vermitteln Höchstleistung.

Bridging HPC Access for SMEs EuroCC Initiatives and the German Ecosystem Dr. Andreas Wierse Managing Director SICOS BW GmbH

1985

Lothar Späth buys a Cray 2 ...





... and creates 10 new positions for SME consultants

1995

Foundation of hww and HLRS



T-Systems

2011

Foundation SICOS BW GmbH





Focus: SMEs





The SICOS BW Tasks

Facilitate SME-Access to HPC Systems and Large Scale Data Facilities

> Support SMEs in finding Competence Partners

Support the Centers in the Optimisation of their Offering towards Industry







SME

Wir vermitteln Höchstleistung.



media

solution center baden

württemberg

OSC(S

Automotive Solution Center for Simulation





From Concepts to Implementation

- EuroCC kicked-off as Research & Innovation Action in **September 2020** (End of Phase 1: 31.12.2022)
 - Phase 2 started 1st of January 2023 (duration 3 years)

32 nations

28 Beneficiaries, 60 Affiliated Entities, 11 Associated Partners

- Germany, Bulgaria, Austria, Croatia, Cyprus, Czech Republic, Denmark, Estonia, Finland, Greece, Hungary, Ireland, Italy, Lithuania, Latvia, Poland, Portugal, Romania, Serbia, Slovenia, Spain, Sweden, France, Netherlands, Belgium, Luxembourg, Slovakia, Norway, Turkey, Republic of North Macedonia, Iceland, Montenegro
- Funding: 50% EuroHPC JU, ~50% from the respective states
- https://www.eurocc-access.eu/
- LinkedIn: EuroCC, Twitter: @EuroCC project



SICOS

What is the Background of EuroCC?



European High-Performance Computing Joint Undertaking ('granting authority'), under the powers delegated by the European Commission ('European Commission') The Topics:

- **HPC** High-Performance Computing
- **HPDA** High-Performance Data Analytics
- Al Artificial Intelligence



The mission of EuroCC 2 is to continue the **establishment of a network of National Centres of Competence (NCC)** in the most efficient way, while continuing to address the differences in the maturity of HPC deployment in Europe, for which improvement has already been noted.

What are National Competence Centers?



National Competence Centers (NCCs) are organizations or institutions that are **designated as experts** in a particular field or technology by a national government. The purpose of NCCs is to **coordinate and enhance** the development of **national expertise and capabilities** in a specific area.

They typically play a key role in **promoting research**, **development**, and innovation, and may also be involved in **providing training**, **education**, and other services to support the growth of a particular industry or field. NCCs may be established in a **variety of fields**, such as advanced manufacturing, biotechnology, renewable energy, or information technology.

They may work with universities, research institutes, and private sector companies to support the development of new technologies, products, and services. NCCs are often funded by the national government or through partnerships with industry and academia, and may have a variety of goals, including improving the competitiveness of national industries, promoting economic















EURO

Smart Data Innovation Lab



EXCELLERAT



Contact

77



SICOS BW GmbH

Nobelstraße 19 70569 Stuttgart

- www.sicos-bw.de
- 🔀 wierse@sicos-bw.de

**** 0711-342033-0



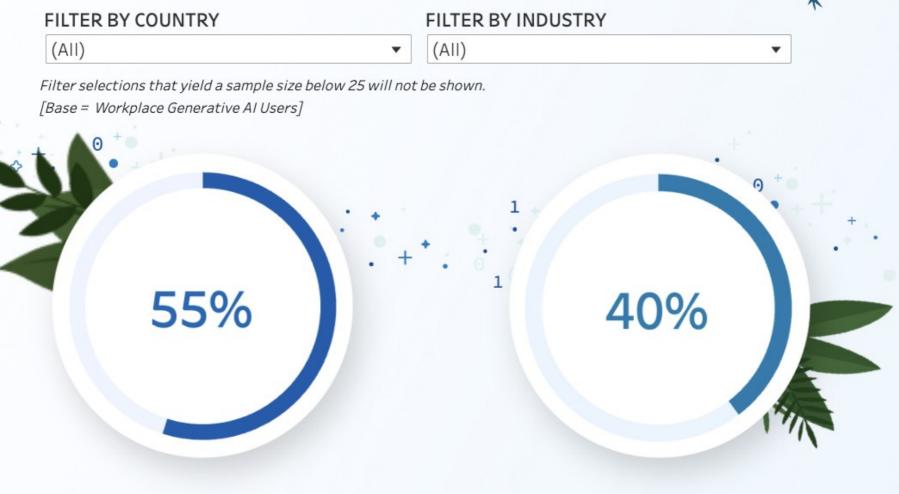


Safeguarding the Information Ecosystems

Kristina Knaving

Senior Researcher, RISE

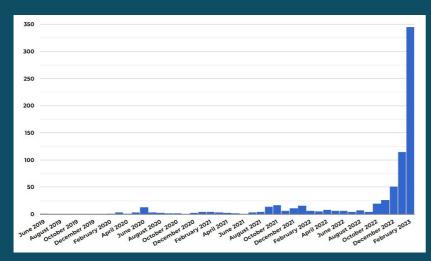
Usage of Unapproved and Banned Generative AI Tools at Work



have used unapproved generative AI tools at work of workplace generative AI users have used banned tools at work

Misinformation and desinformation are poisoning our information ecosystems

- Individuals as well as malicious actors
- Internet, intranets, research
- Human information gathering at risk, also LLM summarising of content
- Model collapse and feedback loops



Submissions to Clarkesworld

Fake Al images are showing up in Google search – and it's a problem					
1000	e Larsen er 28, 2023 4:41PM	SHARE			
Listen to a	rticle 2 minutes	alli			
Google	Israel Kamakawiwo'ole Images Videos News Pronunciation W	X I 🛃 😨 Q			
	<section-header></section-header>	rview Sorge Video Liter Video Liter Video Liter Video Liter Video Liter Video Liter Video Liter Video Liter Video Liter			

Qu	ora			Оре	n in App	Q	
្រា ហ			ث	Q	0		
Did Stannis kill his daughter in the books?							
Answ		ollow	→Q Request	(j Detail		nore	
All rela	ated (38)	~		Reco	ommende	ed 🗸	
	Assistant General kr		Poe Al bot that	excels at	answe	× • Nov 30	

In the "A Song of Ice and Fire" book series by George R.R. Martin, Stannis Baratheon does indeed sacrifice his daughter, Shireen, with the assistance of Melisandre, in a desperate attempt to gain favor from the Lord of Light and improve his chances in the war for the Iron Throne. This event occurs in the fifth book, "A Dance with Dragons."

te

Sponsored by Colt Technology Services

🗘 Upvo

Connect more. Do more

More digital infrastructure solutions to grow your business. An extraordinary customer experience.

	Learn More	ď
C Upvote · 330	₽	
C Upvote · 128	♥ D 17 € 2	•••
Snow. · Upda	ated 3y	

Short answer: no.

Long answer: N000000000000.

Ok, so in short, Shireen is alive and well right now in the books. A lot of people think she will be burned, but that hasn't happened yet.

However, the idea that Stannis will be involved is unlikely. He's at Winterfell like in the books, while Shireen is at the Wall... with her mother and Melisandre. With Jon recently dead and Stannis presumed dead (probably falsely, but unless certain theories are true, everyone at the Wall believes the Pink Letter), I think the idea that







Bias

Bias exists in all data and all humans, and cannot be entirely removed....

...but model bias likely to continue to be less diverse than human bias.

We often focus on easily perceived biases "Microbias" can be a problem over many interactions



Generalisation and convergence – "The AI SAMENESS"

LLMs turn statistical majorities into absolutes



Here are two images depicting a professional-looking individual in a business suit, embodying the essence of a successful CEO in a modern corporate office setting.









Semi-synthetic data

- Al-powered tools makes "human-made" content into semi-synthetic
 - Often without users' knowledge
- Number of tools with AI support grow rapidly- AI filters, adjusts, suggests, generates







Safeguarding the information ecosystems

Suggested solutions

- Al detectors
 - Risk for false positives and negatives
- Mandatory labelling , watermarking and shibboleths
 - Only works for regulated models and model use
- Known sender
 - Privacy issues
- Human attention is needed, even though expensive
- Trusted repositories, verified data sets, curated content
- Educating users to facilitate human-in-the-loop
- Technical soluutions, such as expand modalities



Trusted data repositories

- Curated content / verified sender
- For information gathering, but also for training models

- Who decides what is trustworthy and less biased?
- Who owns the data? Will usable data be locked into organisations?
- Who will be able to train on especially valuable data, such as research articles?



Panel and Summary

HPC meets the demands of AI and Data Room: Vanadium



