Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA
ICSC Centro Nazionale di Ricerca in HPC, Big Data and Quantum Computing

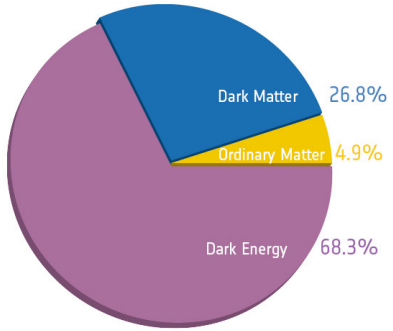
Valorizzazione dell'informazione cosmologica nelle grandi survey di galassie dallo spazio attraverso algoritmi innovativi di Machine Learning

Luigi Guzzo, Università degli Studi di Milano

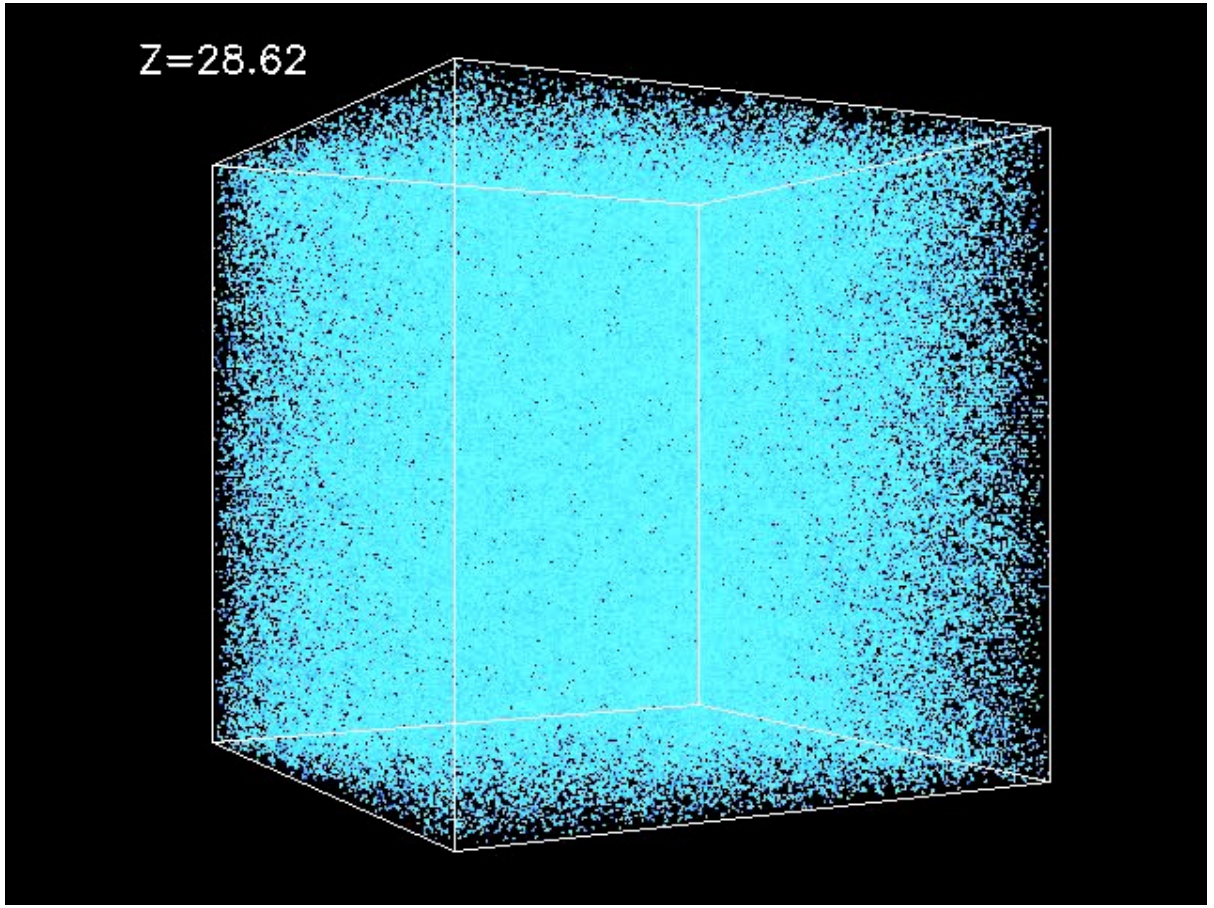
Spoke 3 Progetti Bandi a Cascata, 24/09, 2024

Introduction: the goals of modern cosmology

- We live in a strange Universe dominated by dark matter and dark energy, which we do not understand



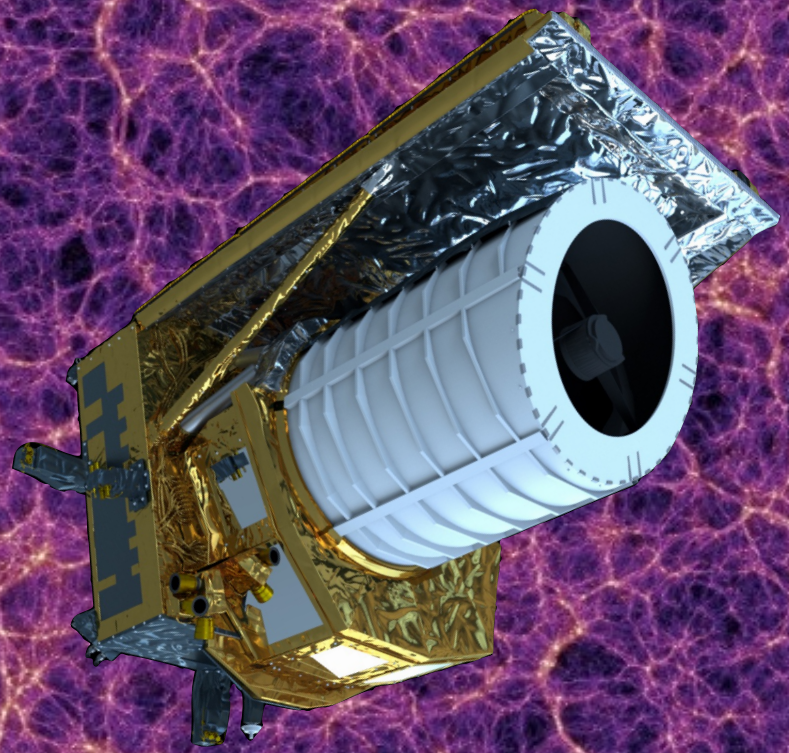
- The ingredients of this cosmic soup shape the formation of the large-scale structure of the Universe



The numbers of the Universe are encoded in its large-scale structure

- To reconstruct the large-scale structure of the Universe and measure cosmological parameters we build larger and larger 3D galaxy maps, through spectroscopic measurements of their distances ("*redshifts*")

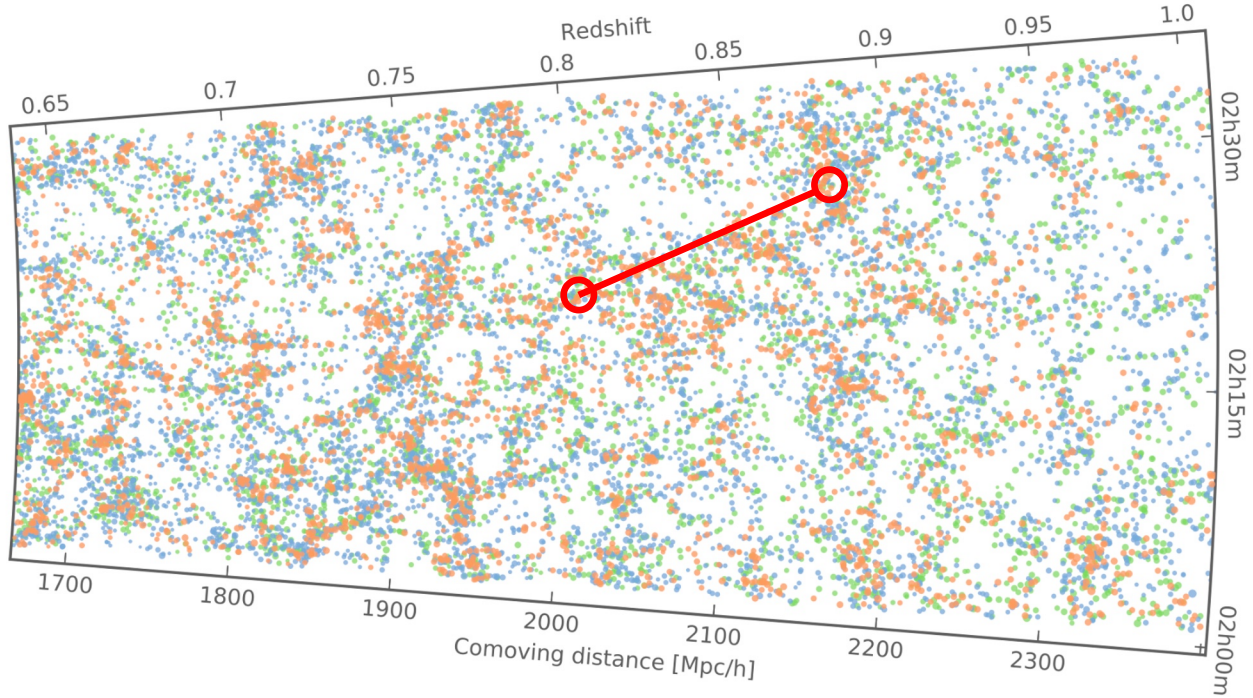
- The ESA Euclid mission is the currently largest such endeavour: in 6 years it will measure redshifts for more than 30 million galaxies to unprecedented distances, over 1/3 of the sky

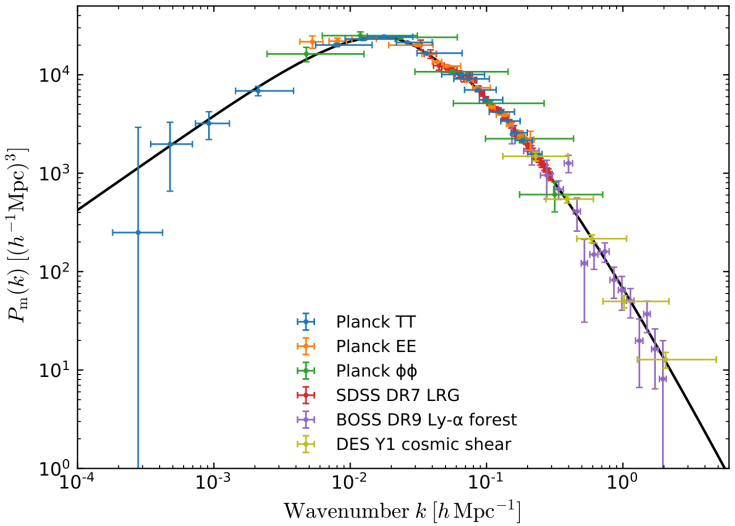


Extracting cosmological parameters

- We extract the "numbers of the Universe" (as the density of dark matter Ω_M, the equation of state of dark energy w, or the growth rate of cosmic structure f), by measuring statistics of galaxy clustering and comparing them to model predictions

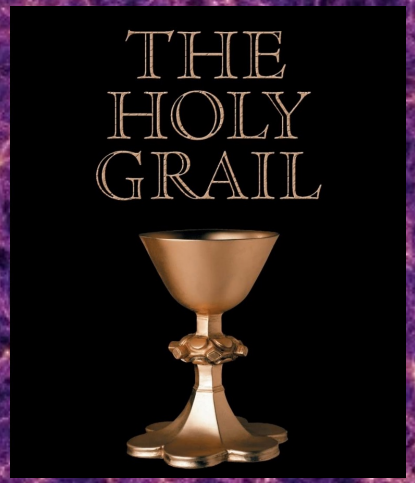
- This entails essentially extracting n-point correlation functions, with n=2, 3, ... in real or Fourier space

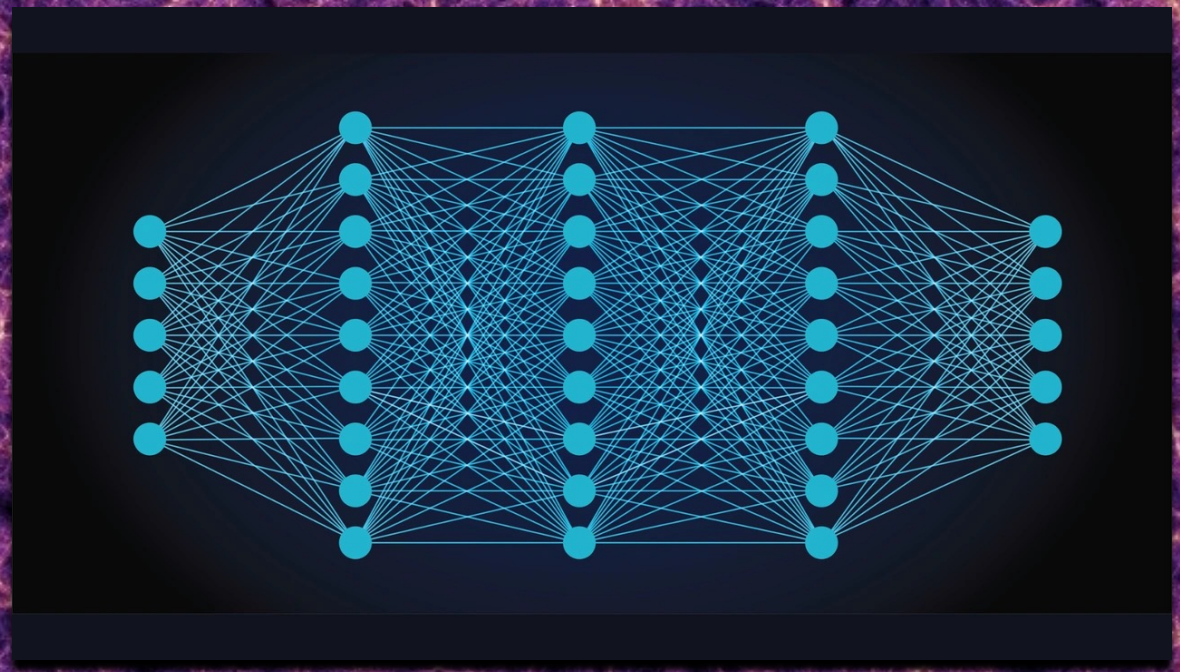


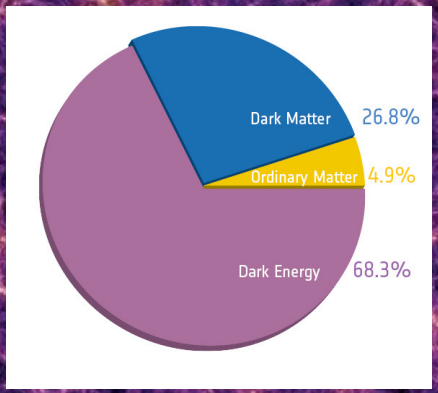


Power spectrum: 2-point statistics in Fourier space

The holy grail: bypass expensive n-point functions (summary statistics) via trained Machine Learning algorithms that "read" the parameters directly from the galaxy field

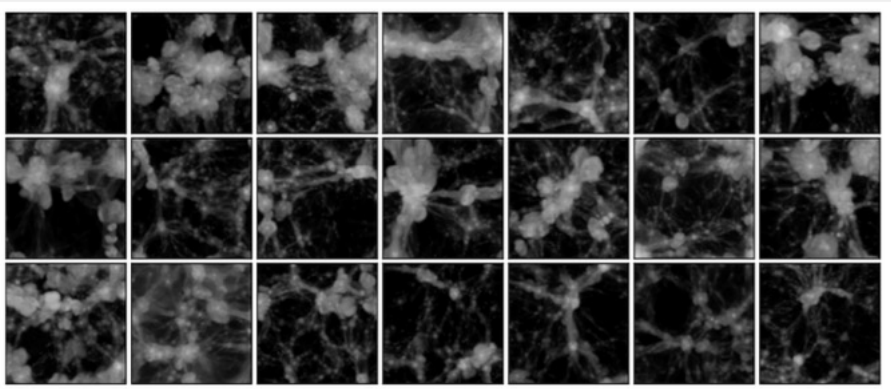




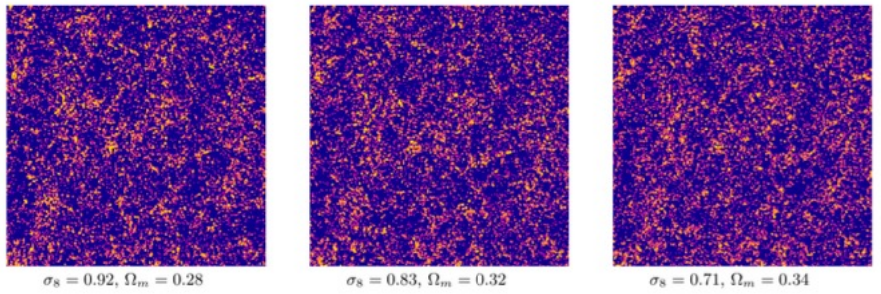


1) Early attempts on cosmological simulations, using CNNs

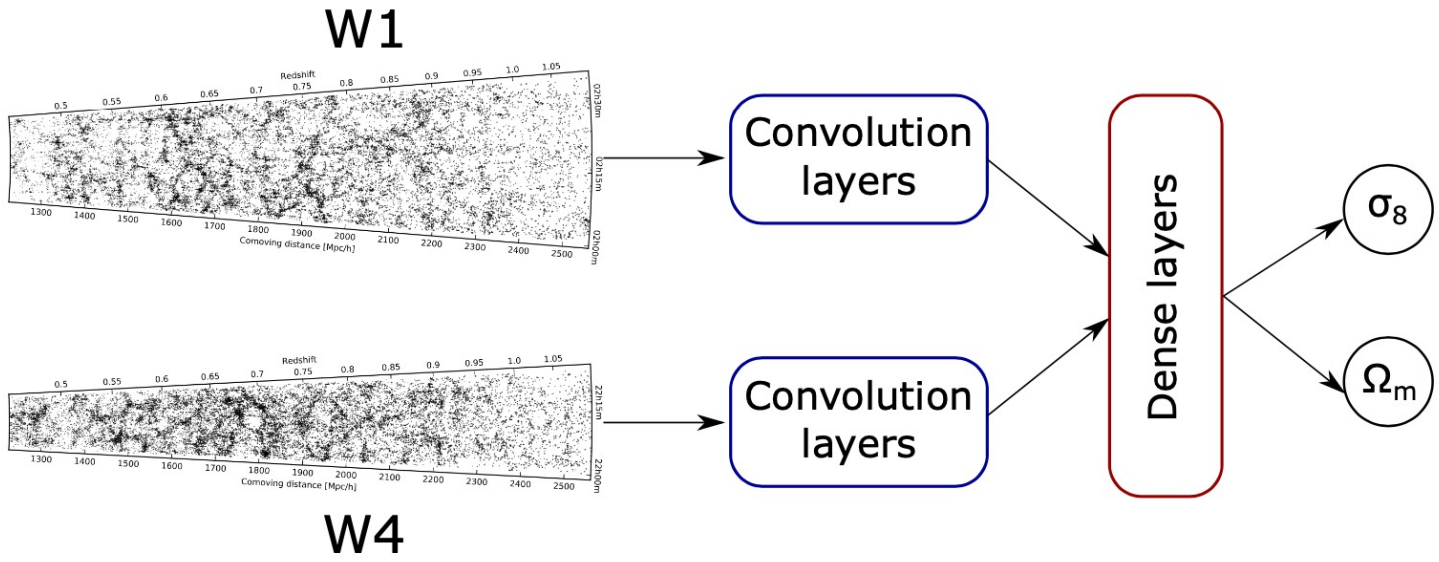
In Villaescusa-Navarro et al. (2021):



In Ntampaka et al. (2020):



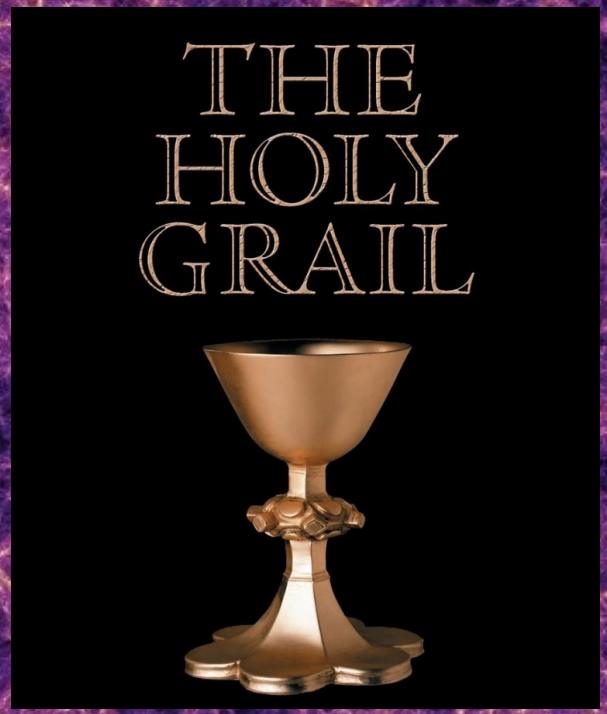
2) Early attempts on real data (2D slices, using CNNs)



- The ambitious plan is a field-level analysis of VIPERS with a convolutional neural network.
- We use only observational information: the angular position and the redshift of the objects.

(M. Cagliari, PhD thesis, 2024)

The fundamental problem for applications to real galaxy catalogues: WE HAVE NO TRAINING SAMPLES

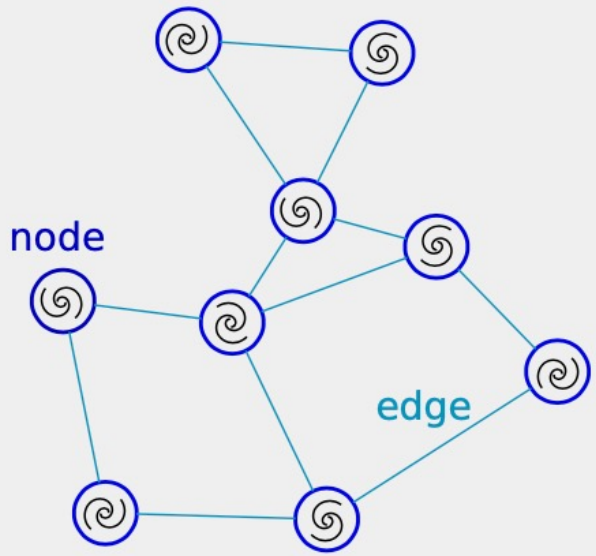


MLS Technical Objectives, Methodologies and Solutions

1. **Identify and optimise ML algorithms for cosmological inference:**

   a. **CNN applications to 2D samples? (drawing from M. Cagliari PhD thesis results)**
   b. **Graph Neural Networks (GNN) as descriptors of galaxy relationships (Tosone+ 2023)**
   c. **Physically Informed NNs in the cosmological context (no application so far)**

- CNN are natural for simulation data
- GNN are natural for real galaxy data



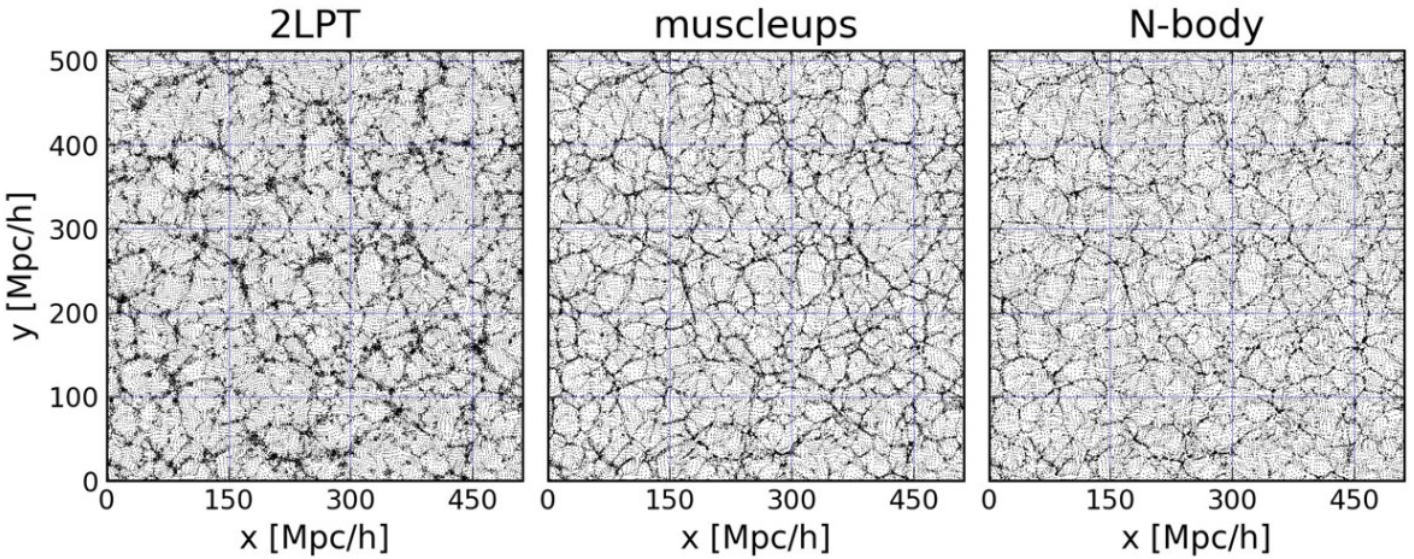
credit Marina Cagliari

MLS Technical Objectives, Methodologies and Solutions

2. Build realistic training samples from numerical simulations:

a. Build fast dark matter skeletons from perturbation theory surrogated simulations (Pinocchio, Monaco+ 2013; MUSCLE-UPS, Tosone+ 2022) vs full n-body

b. "Illuminate" dark-matter haloes through realistic recipes (X-ray clusters, easier, and galaxies, more complicated), via physically informed NNs, to build next-generation training sets

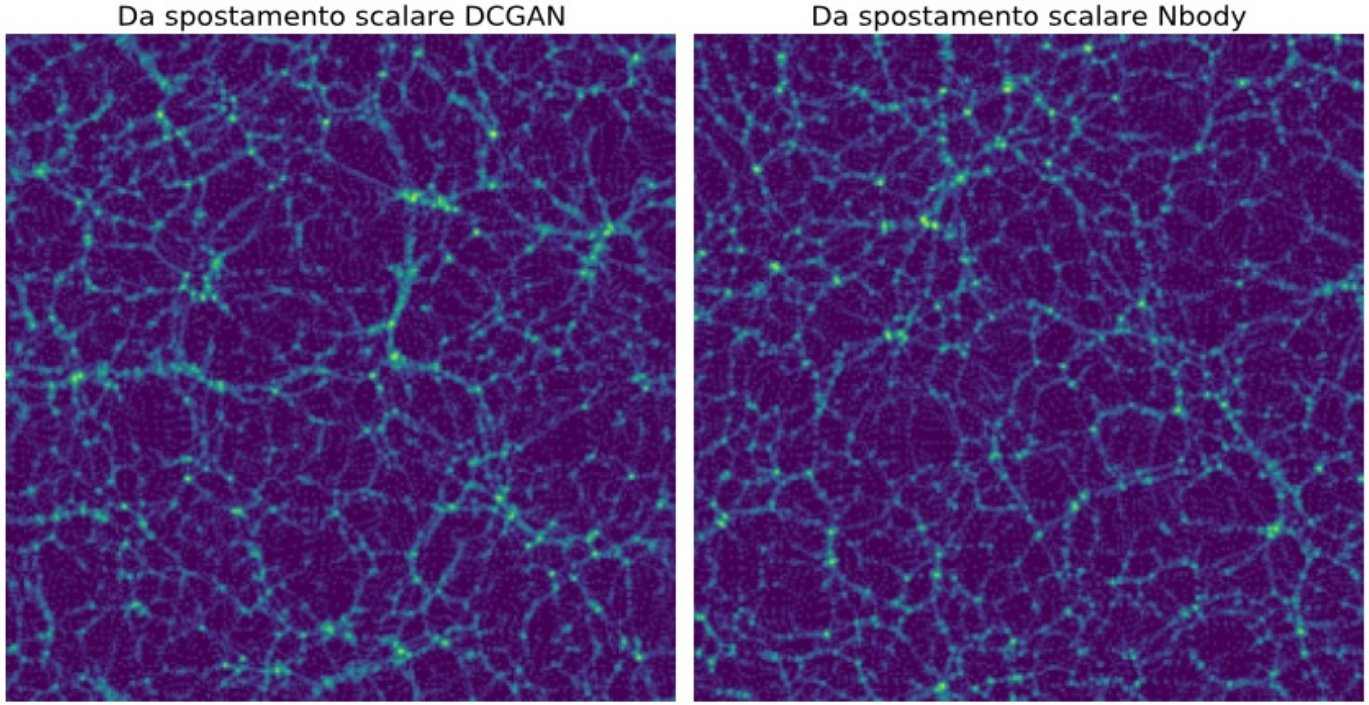


(Tosone+, 2021)

MLS Technical Objectives, Methodologies and Solutions

2. Build realistic training samples from numerical simulations:

- Another option to build fast "dark matter skeletons" : replicate n-body outputs via ML
- "ML helps ML"

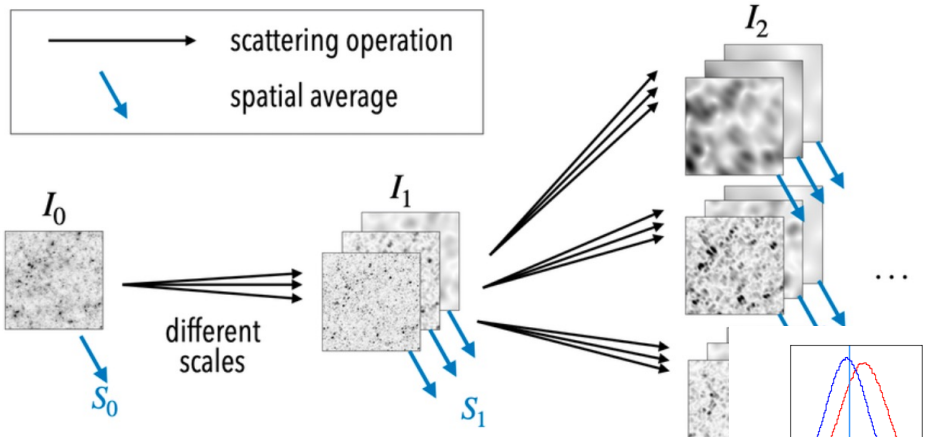


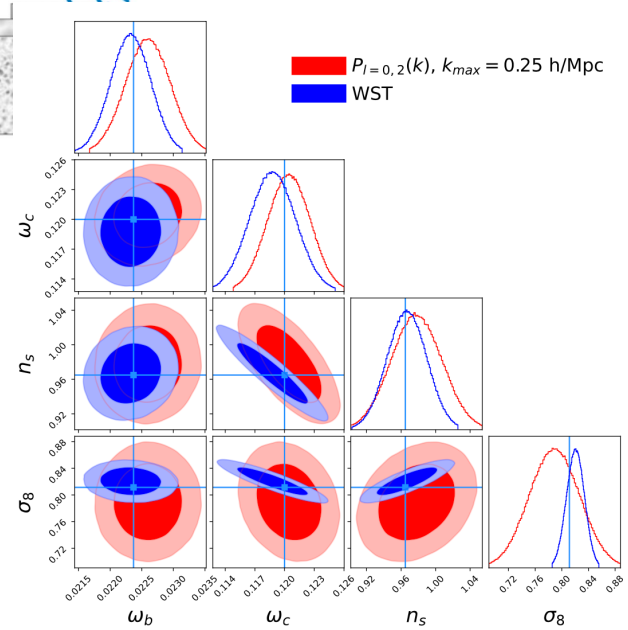
Emulate Lagrangian displacement field against N-body output
Use CNN-UNET (Sofia Chiarenza BSc. thesis) or Gen. Adversarial Networks (GAN, Marco Chiarenza Bs.c thesis)

MLS Technical Objectives, Methodologies and Solutions

3. **Compare with non-ML field-level parameter estimations techniques**

   a. **E.g., Wavelet Scattering Transform and other techniques to capture higher-order information via filtering/weighting of the galaxy field**





(S. Chiarenza, MSc thesis)

MLS: involved Staff and new recruitments

STAFF:

- **Luigi Guzzo (PO), Università di Milano** (expertise: cosmological observations and modelling, Euclid)
- **Davide Maino (PA), Università di Milano** (expertise: cosmological observations and theory, Euclid)

RECRUITED THROUGH MLS:

- **1-year RTD contract starting October 1st** (Davide Bianchi,– expertise: large galaxy surveys, statistical analysis of clustering)

EXTERNAL CONSULTANT SUPPORTED BY MLS:

- **Fondazione Clement Fillietroz - ONLUS, Osservatorio Astronomico della Regione Autonoma Valle d'Aosta** (expertise: cosmological simulations, Machine Learning applications in astrophysics and genomics)

MLS : timescale, Milestones, SAL

		SAL 1			SAL 2			SAL 3			SAL FINALE		
		T1			T2			T3			T4		
WP	Titolo	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	M11	M12
1	Metodi ML per applicazioni cosmologiche												
2	Costruzione di mock training samples sintetici realistici												
3	Inferenza Cosmologica												
	MILESTONES			MS1			MS2			MS3			MS4

MS1: Identificazione dati e algoritmi. Definizione della tipologia di dati (galassie, ammassi) e delle simulazioni (N-Body / LPT) ; confronto e identificazione algoritmi

MS2: Rilascio algoritmo specifico. Completamento architettura e definizione dell'algoritmo «physically aware». ottimizzato sui dati selezionati alla MS1

MS3: Rilascio set cataloghi sintetici di «galassie» per training costruito dalle simulazioni. Dataset completo su cui addestrare il modello di ML/DL generato a MS2.

MS4: Applicazione e validazione dei risultati cosmologici. Applicazione a «test case» dell'algoritmo e confronto parametri cosmologici con output metodi tradizionali e altri metodi field-level (es. WST).