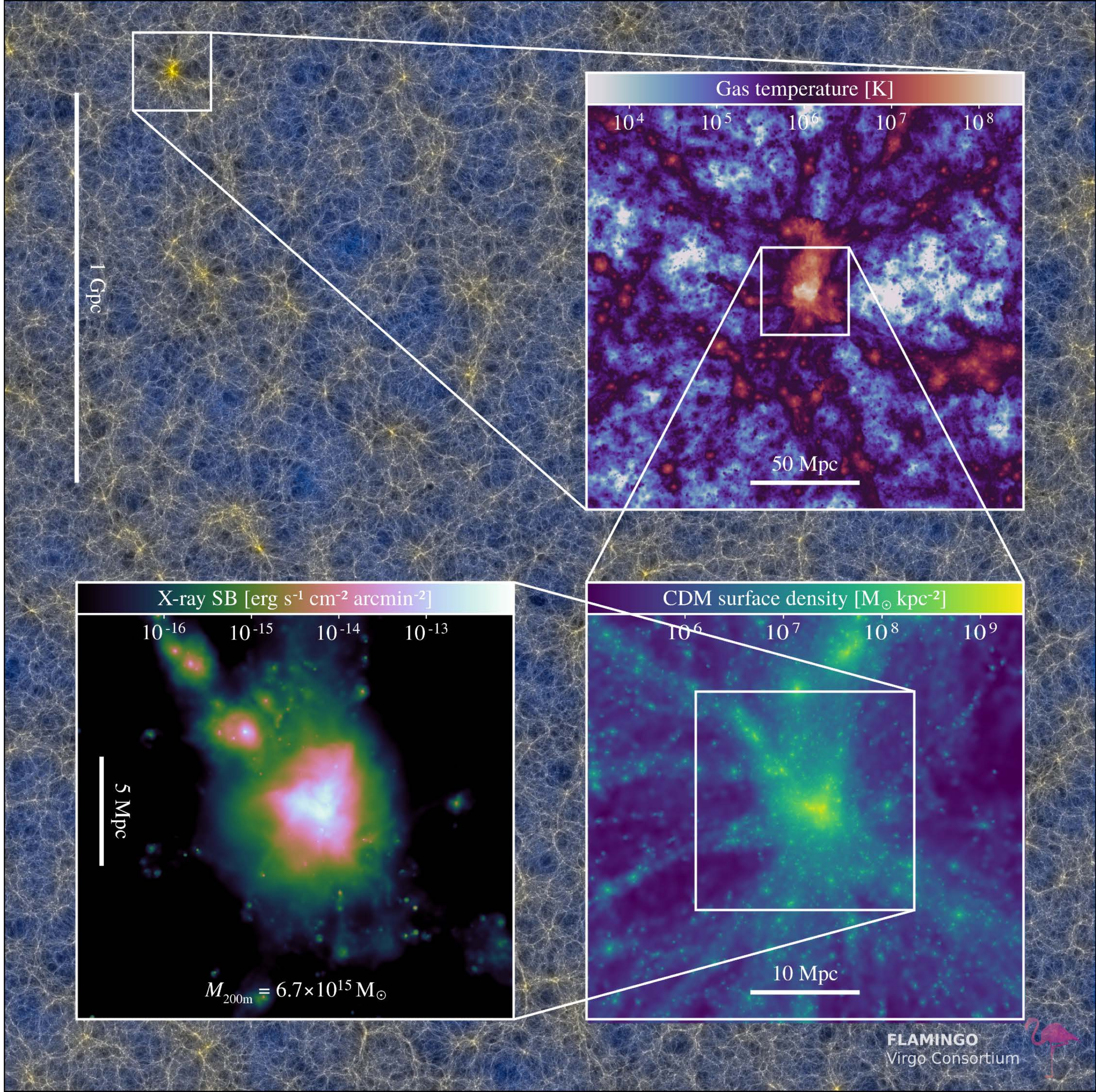
Cosmology via big data in Astrophysics

Emilio Romano-Díaz
Argelander Institut für Astronomie
University of Bonn

05/07/2024

CCS-UKY

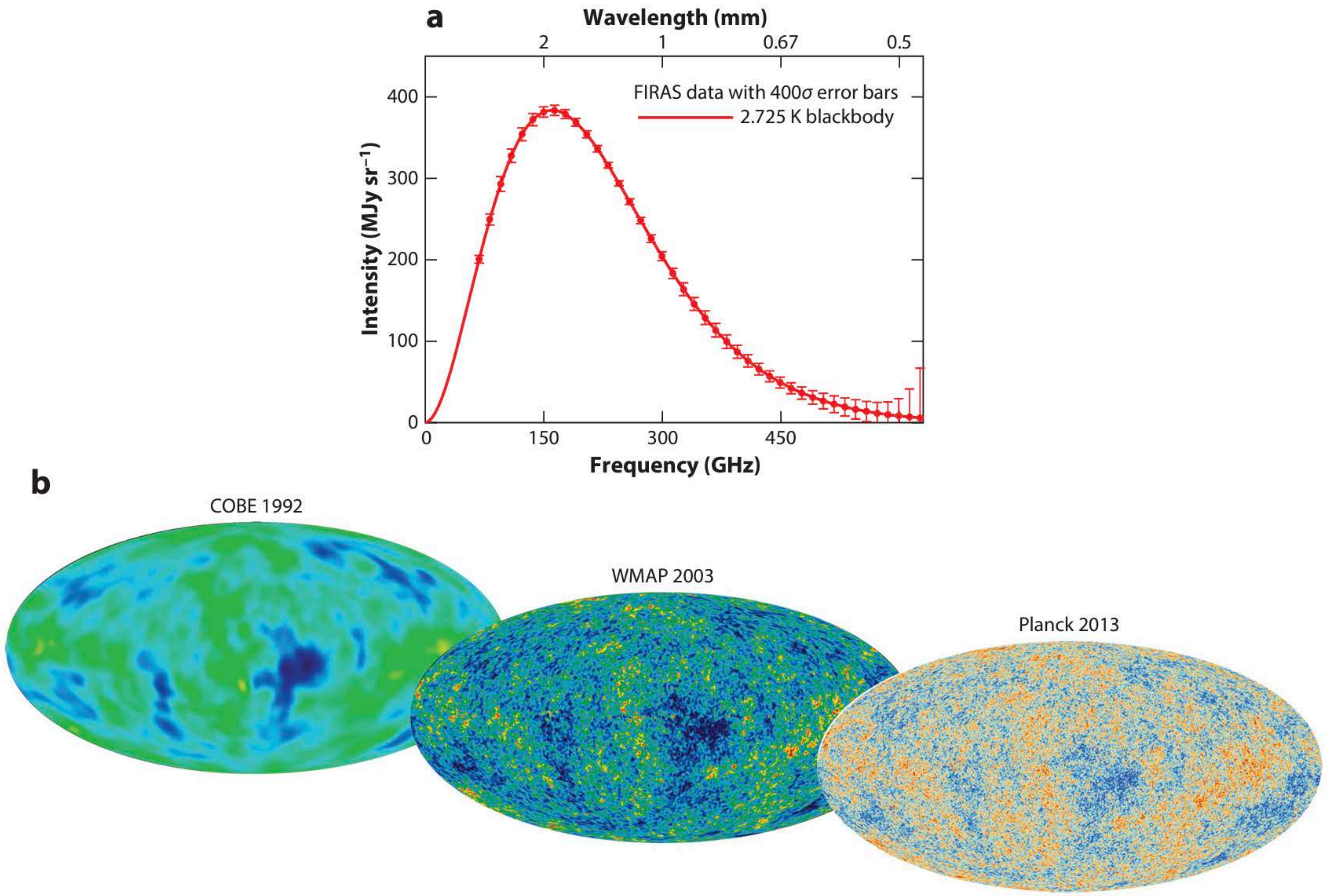


- Cosmology
- Precision Cosmology
- Observational Data: Surveys, LSST, EUCLID
- Simulations: DM-only, Dm+Baryons (Hydrodynamics, MHD, RT)
- Small simulations to feed LSS
- Emulators (future)

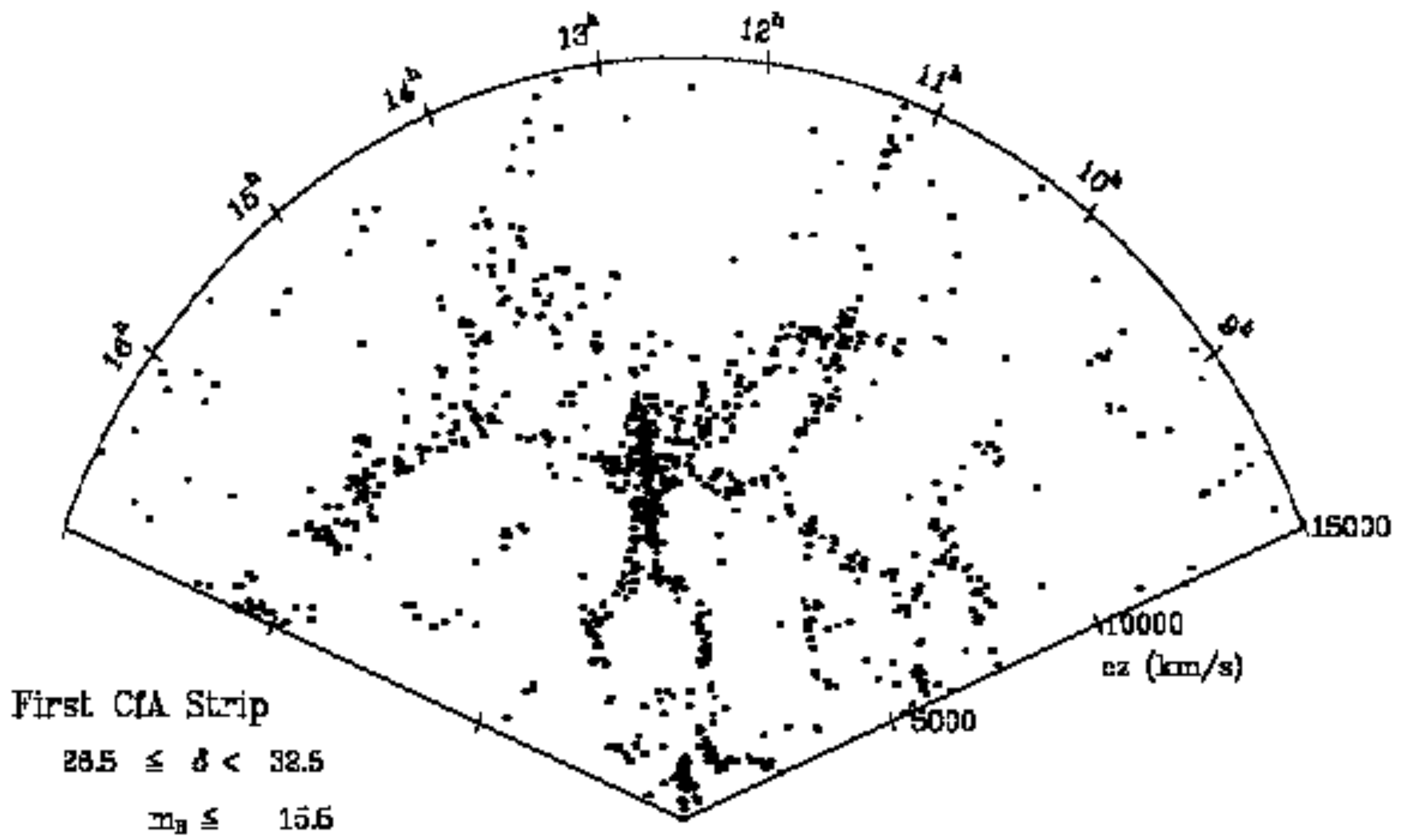
Cosmology

From data starved to big-data science!

- Relatively a "new" field, just some 100 yrs old, with an exponential growth since the 1970's
- LCDM cosmological model described by 6 parameters: n_s, s_8, Ω_m, Ω_L, Ω_b, H_0, t
- Galaxy Surveys

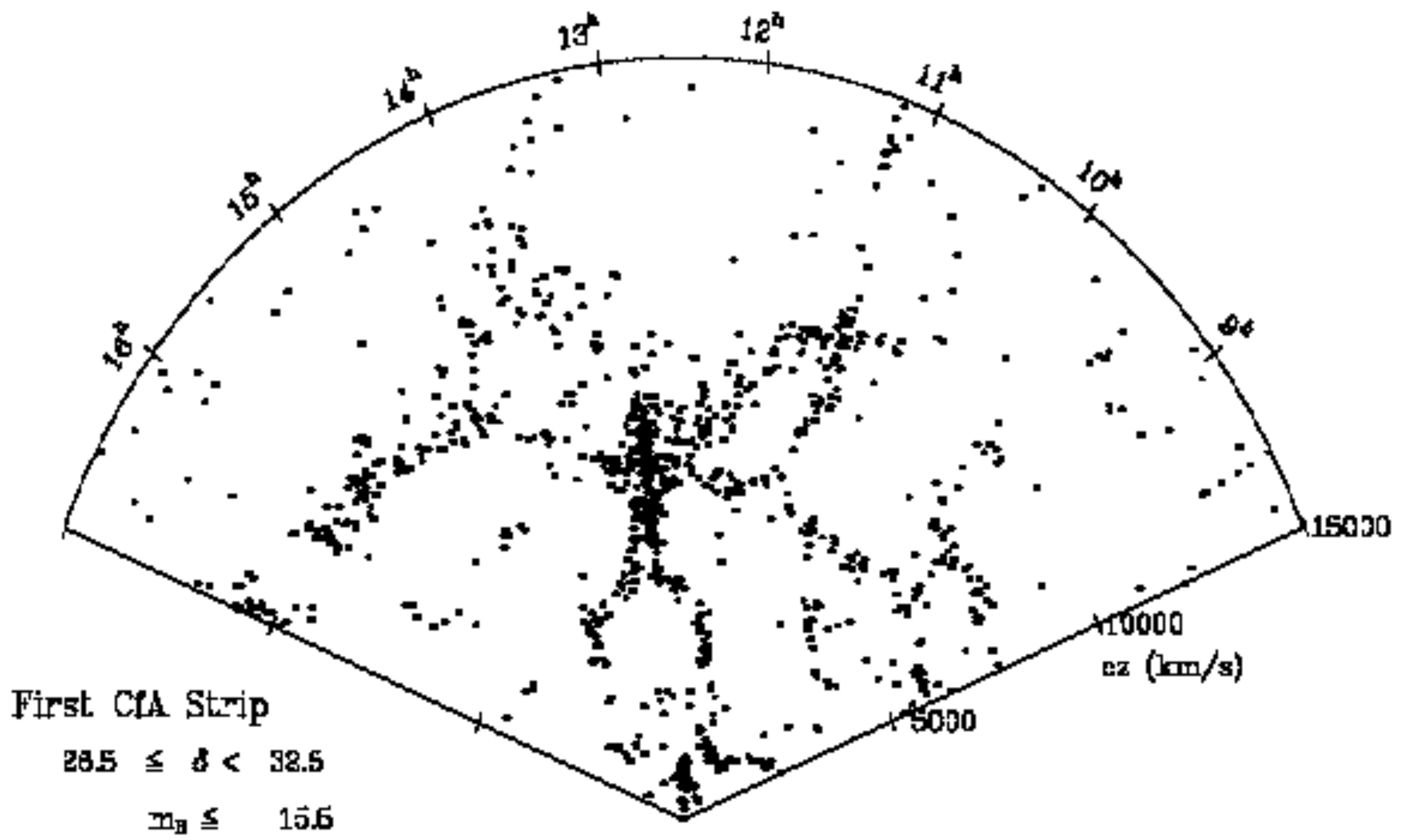


Galaxy Surveys:

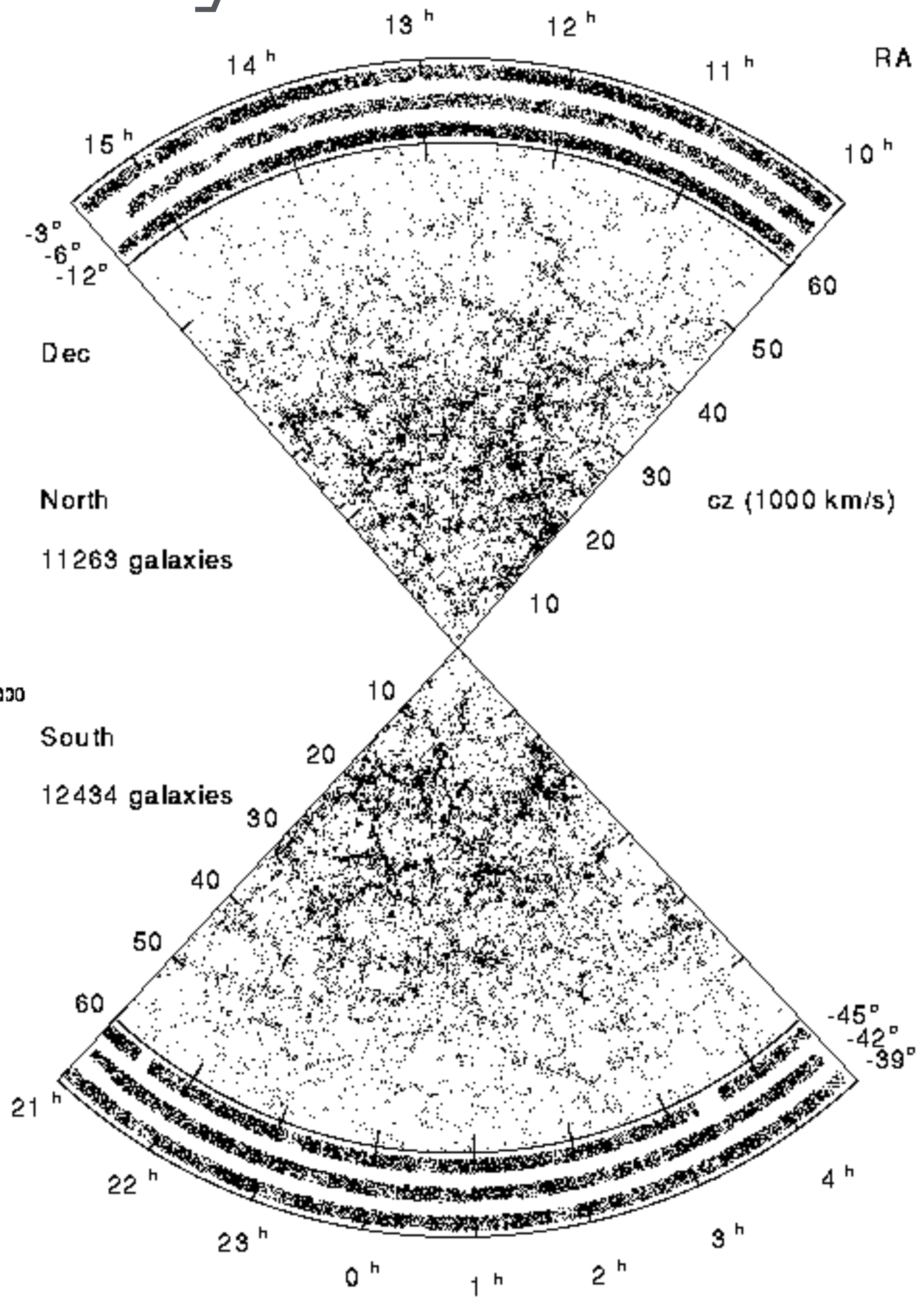


CFA:
1,732 galaxies (1977)
18,000 (1990)

Galaxy Surveys:

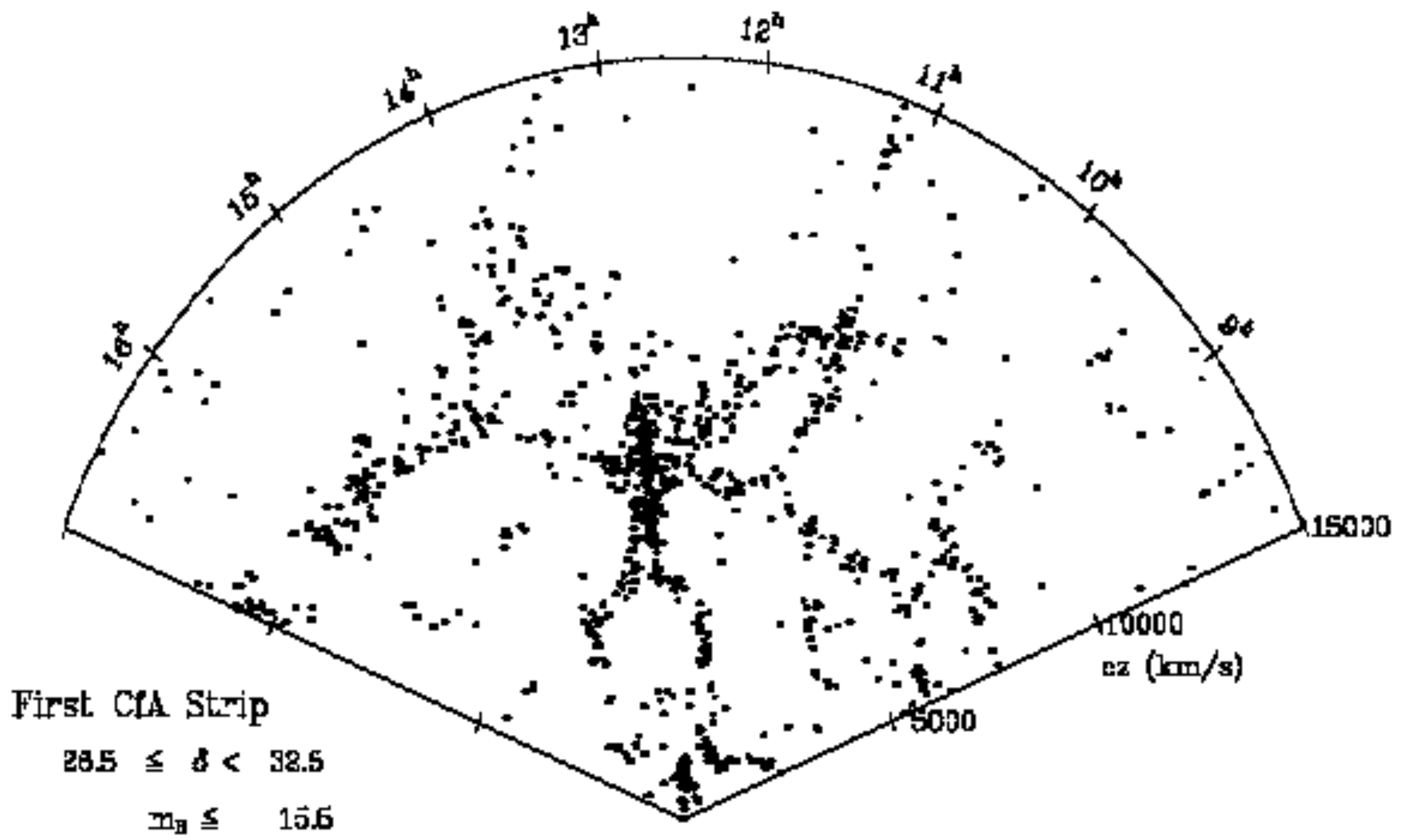


CFA:
1,732 galaxies (1977)
18,000 (1990)

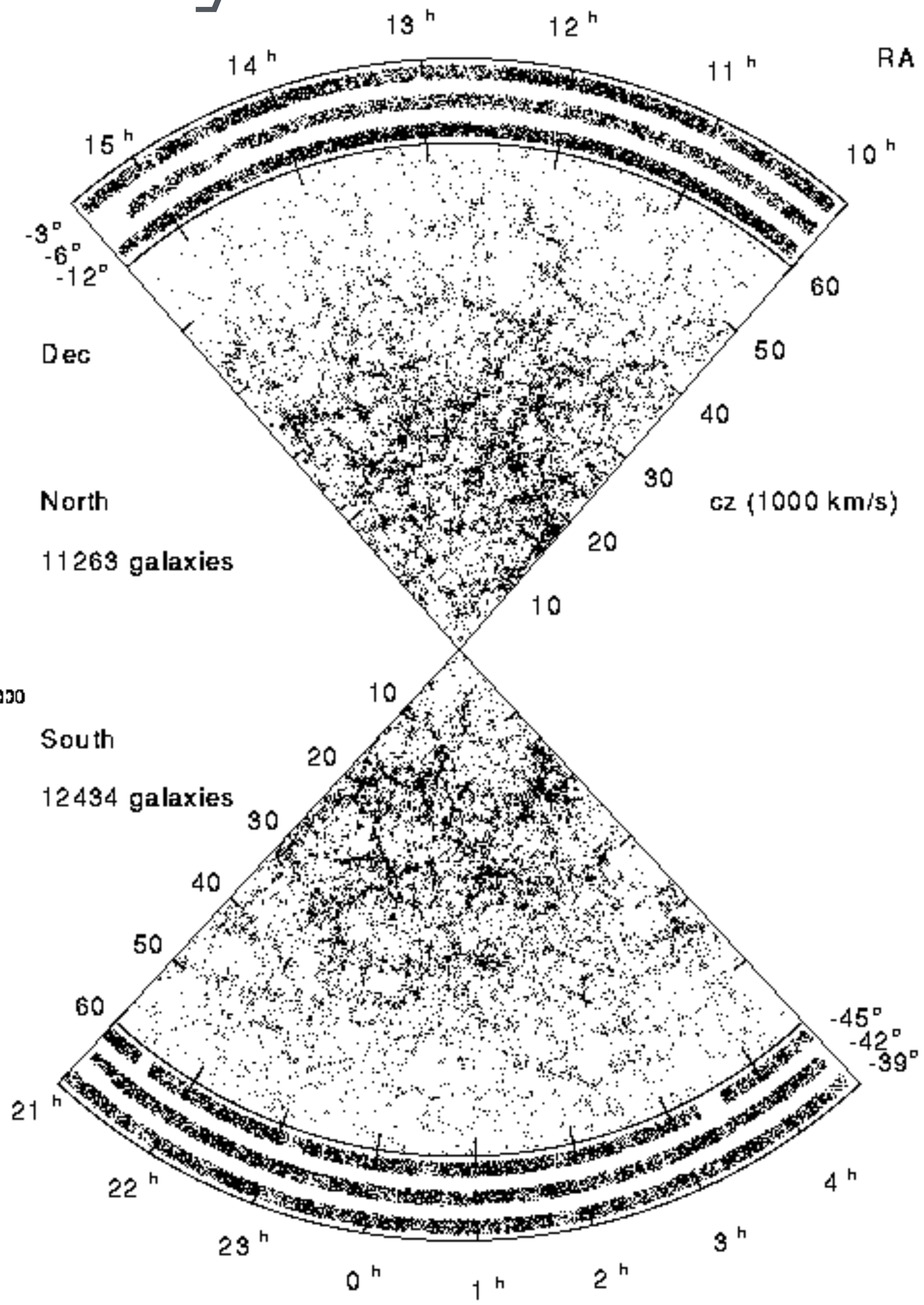


LCRS: 26,418 galaxies (1991)

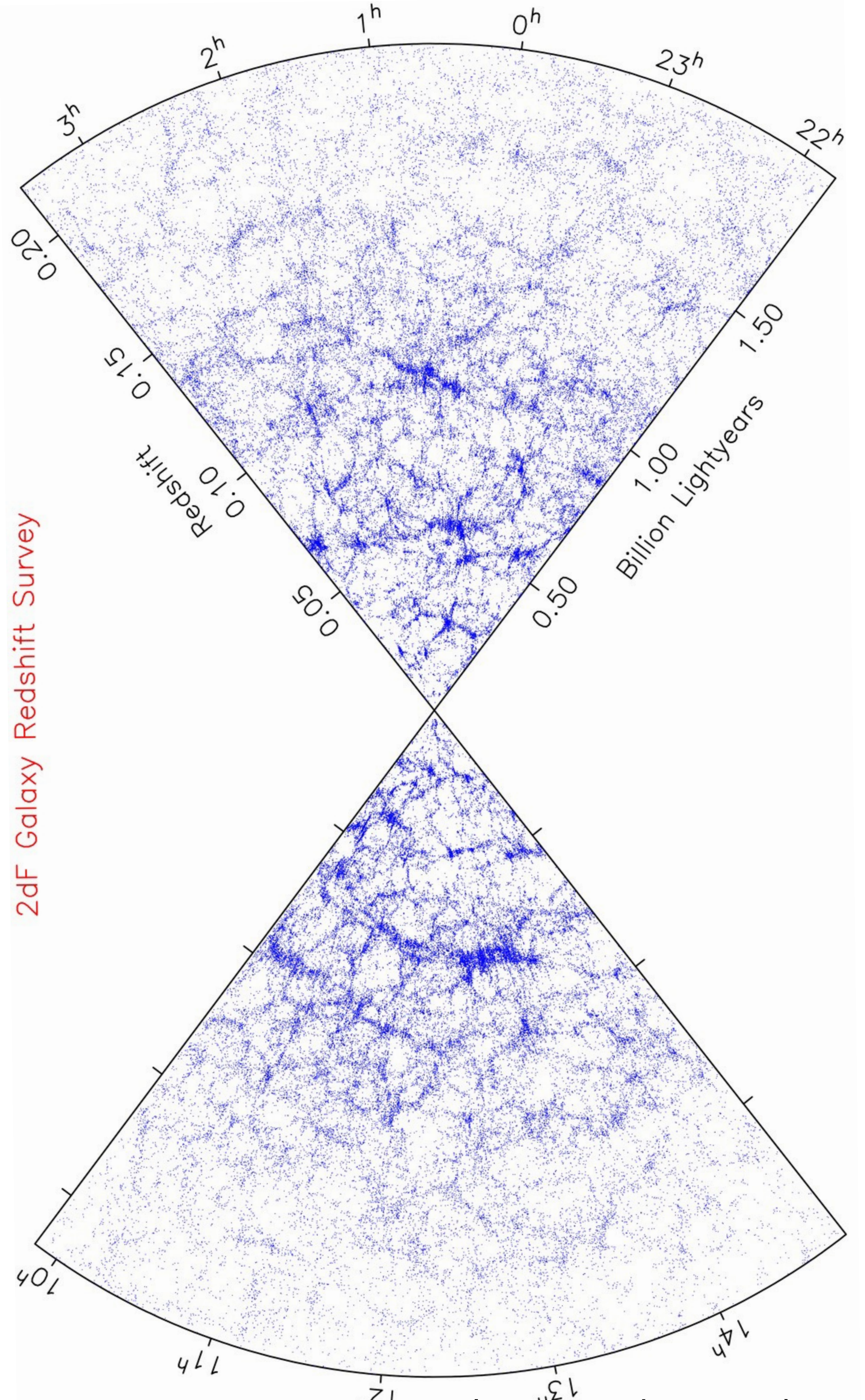
Galaxy Surveys:



CFA:
1,732 galaxies (1977)
18,000 (1990)



LCRS: 26,418 galaxies (1991)



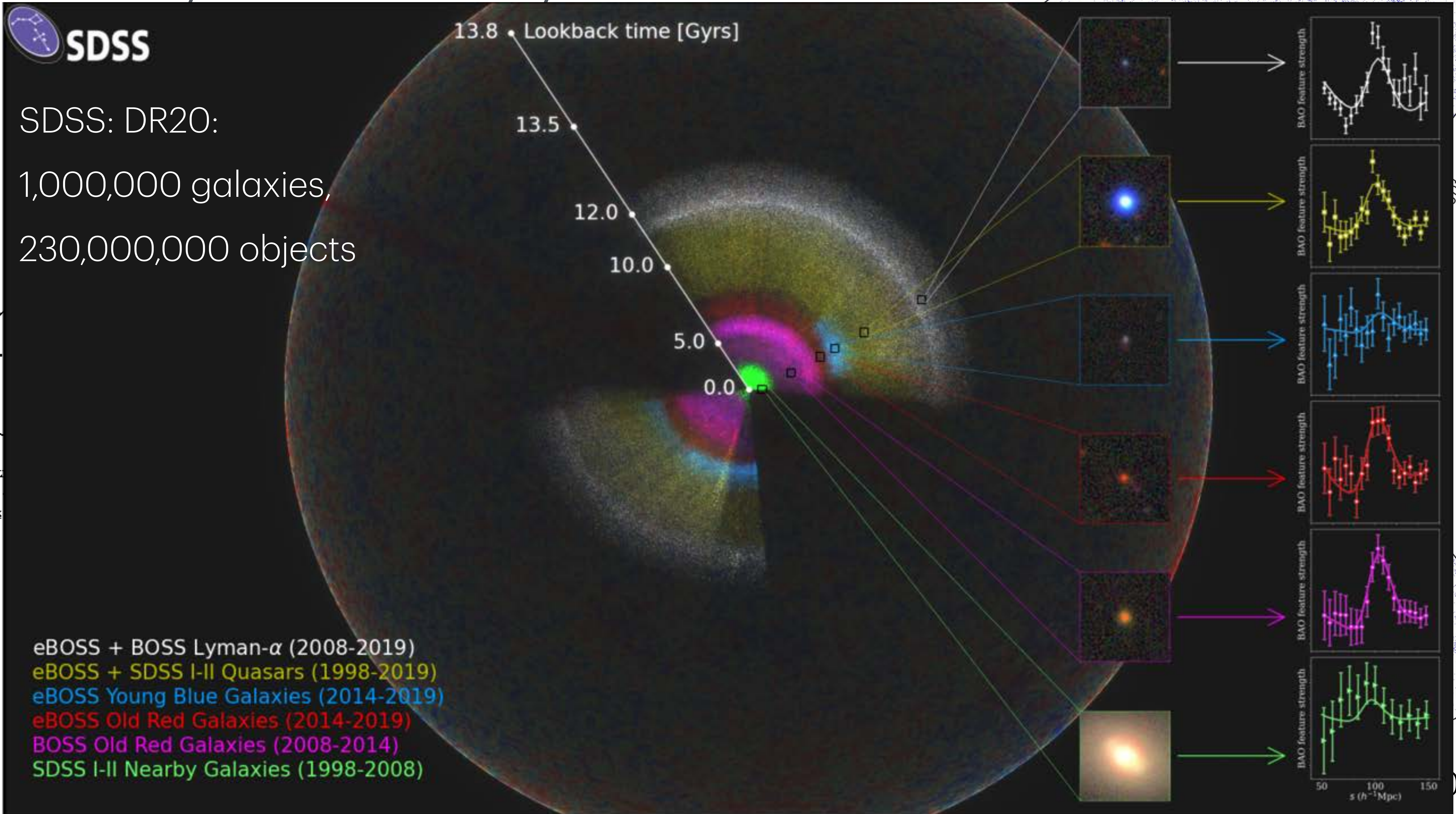
2DF: 221,000 galaxies (2002)

Galaxy Surveys:

SDSS: DR20:

1,000,000 galaxies,

230,000,000 objects

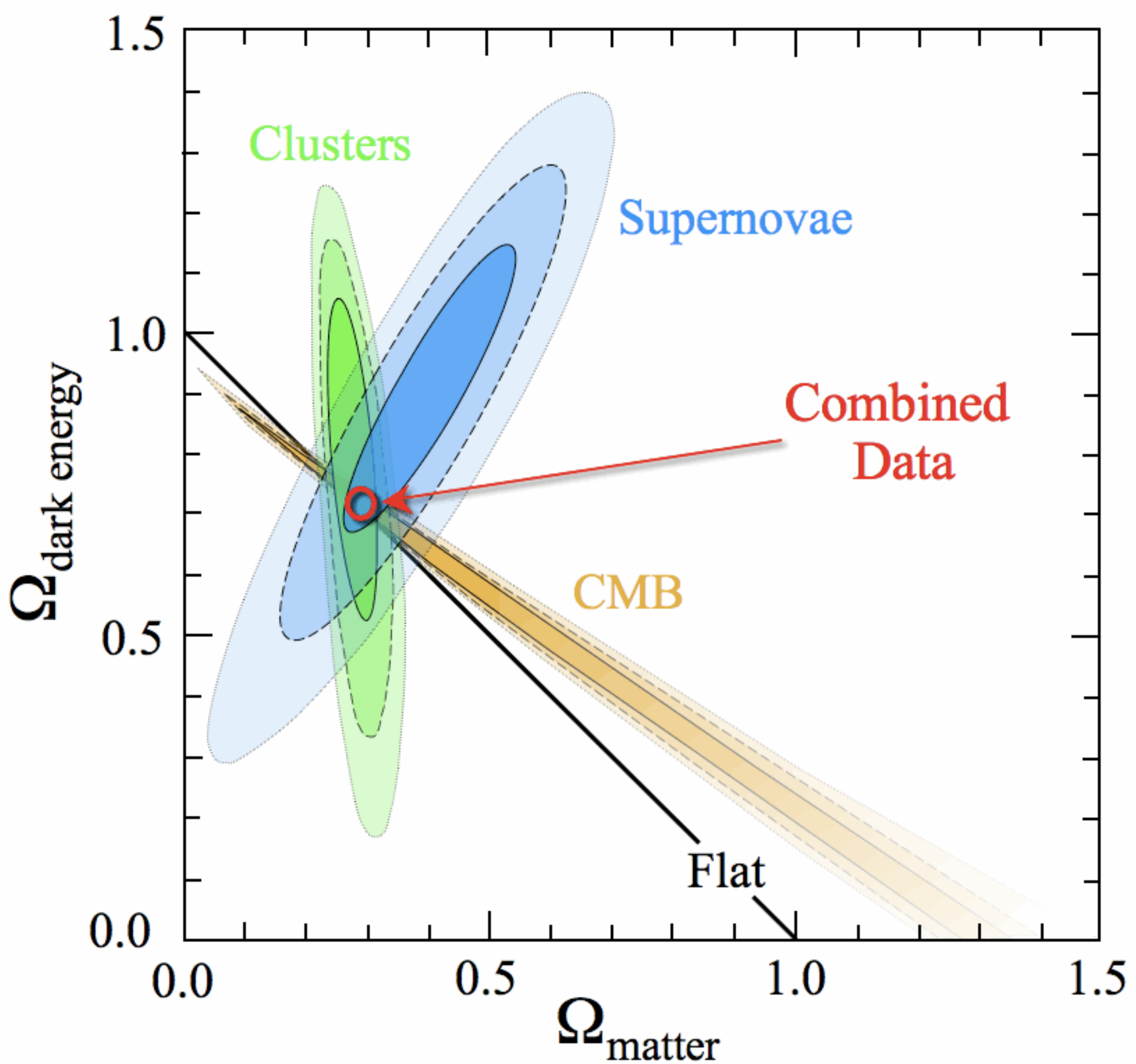


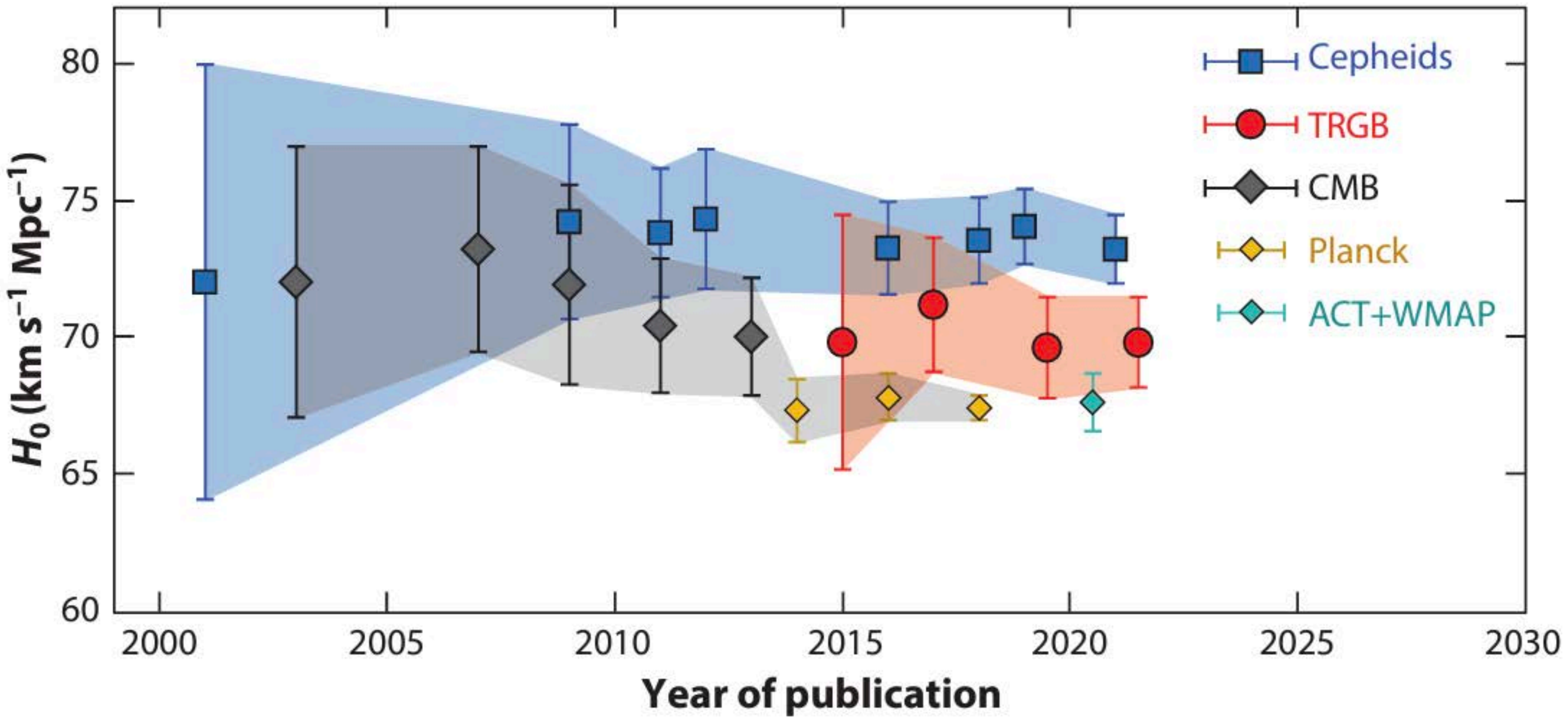
Galaxy Surveys:

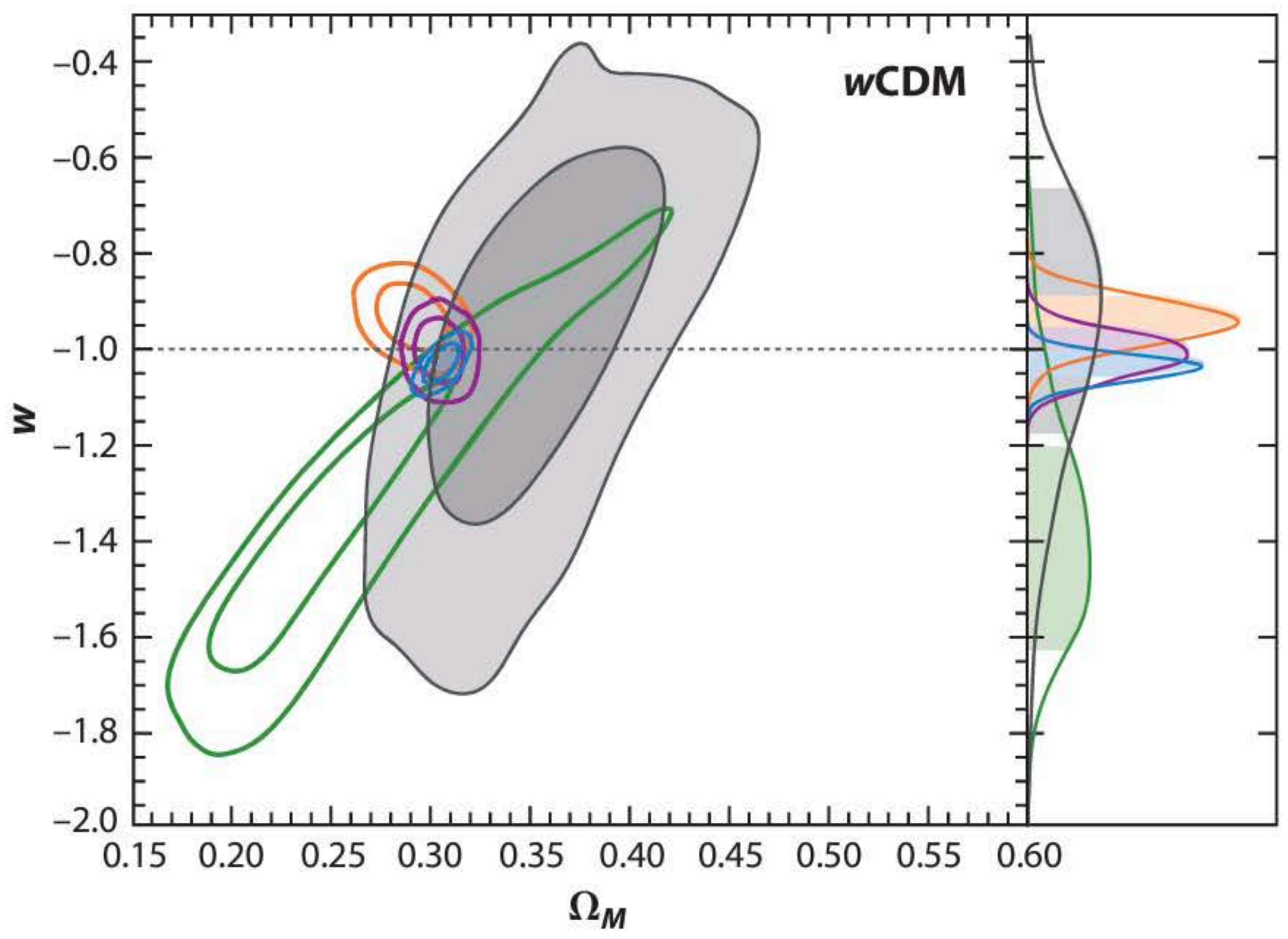
The Vera C. Rubin Observatory (LSST) / EUCLID mission

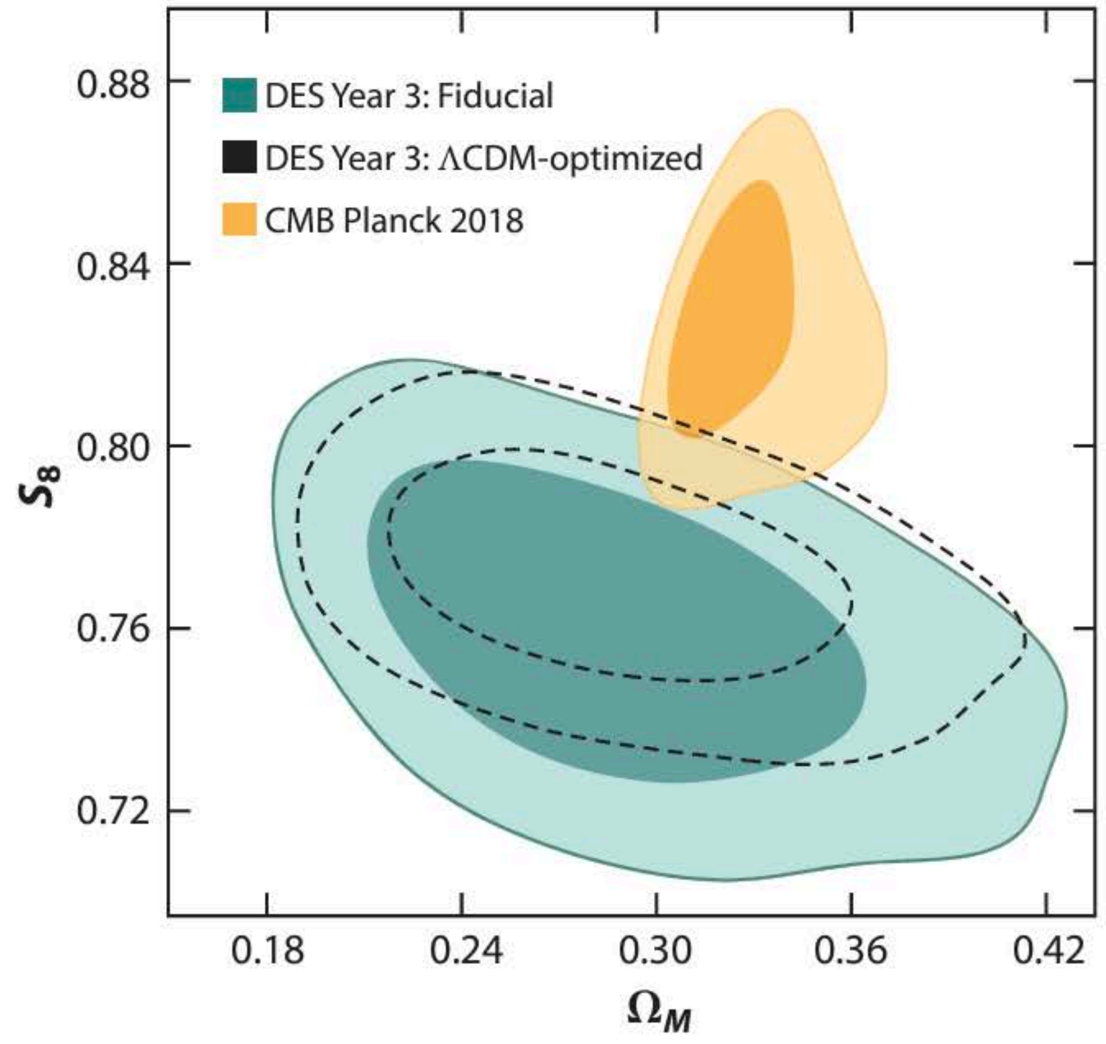
- Telescope with a 3.2Gigapixel CCD camera
- 200,000 pictures/yr => 1.28 PB uncompressed data
- Computer requirements: 250 teraflops & 100PB storage
- Pipeline is being automatised and three different timescales, prompt, daily, and annually.
- Most of the data will be cleaned via ML techniques, while some data will be kept in RAW format
- Sponsored by NCSA & IN2P3 (France)

Precision Cosmology

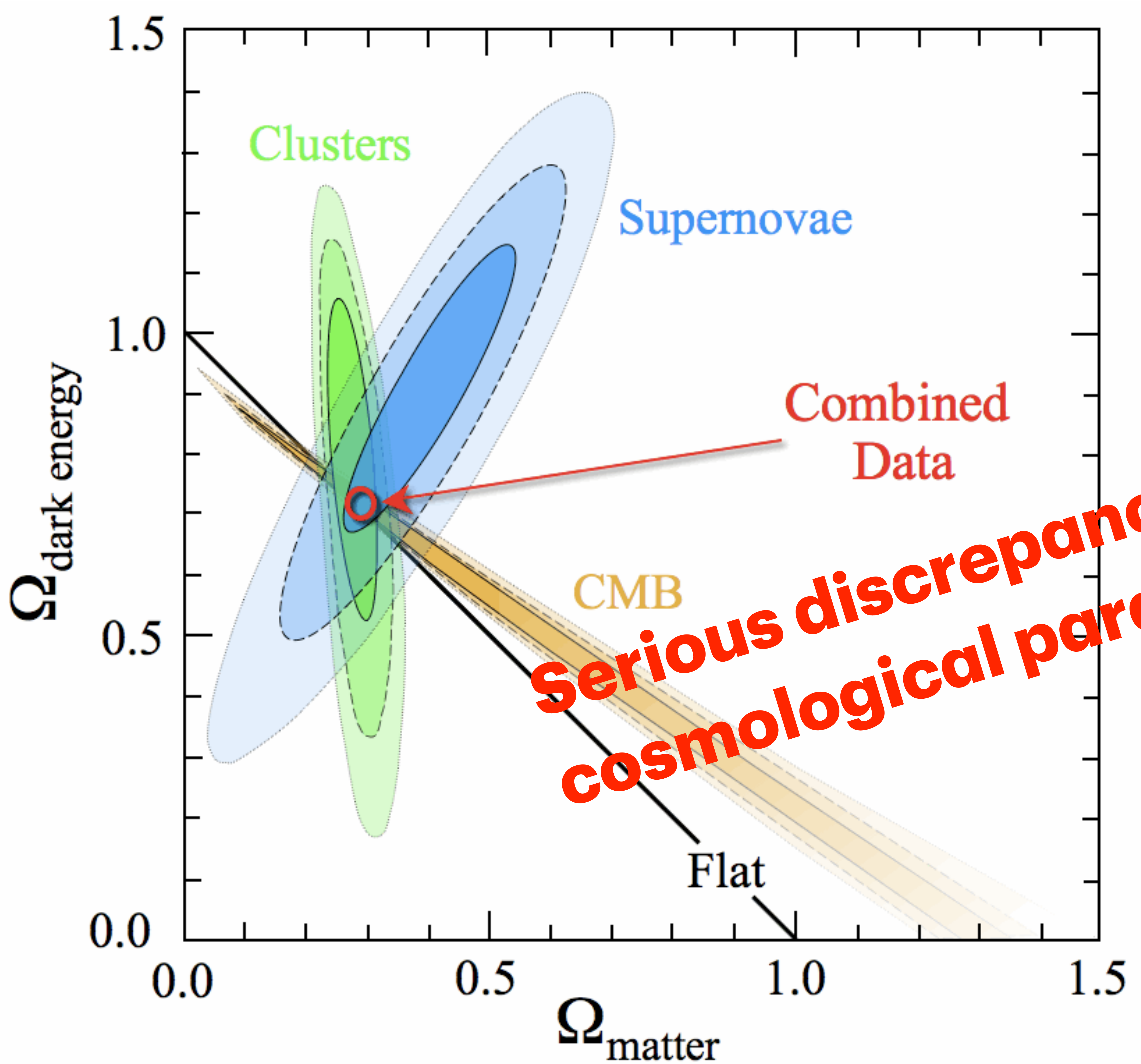


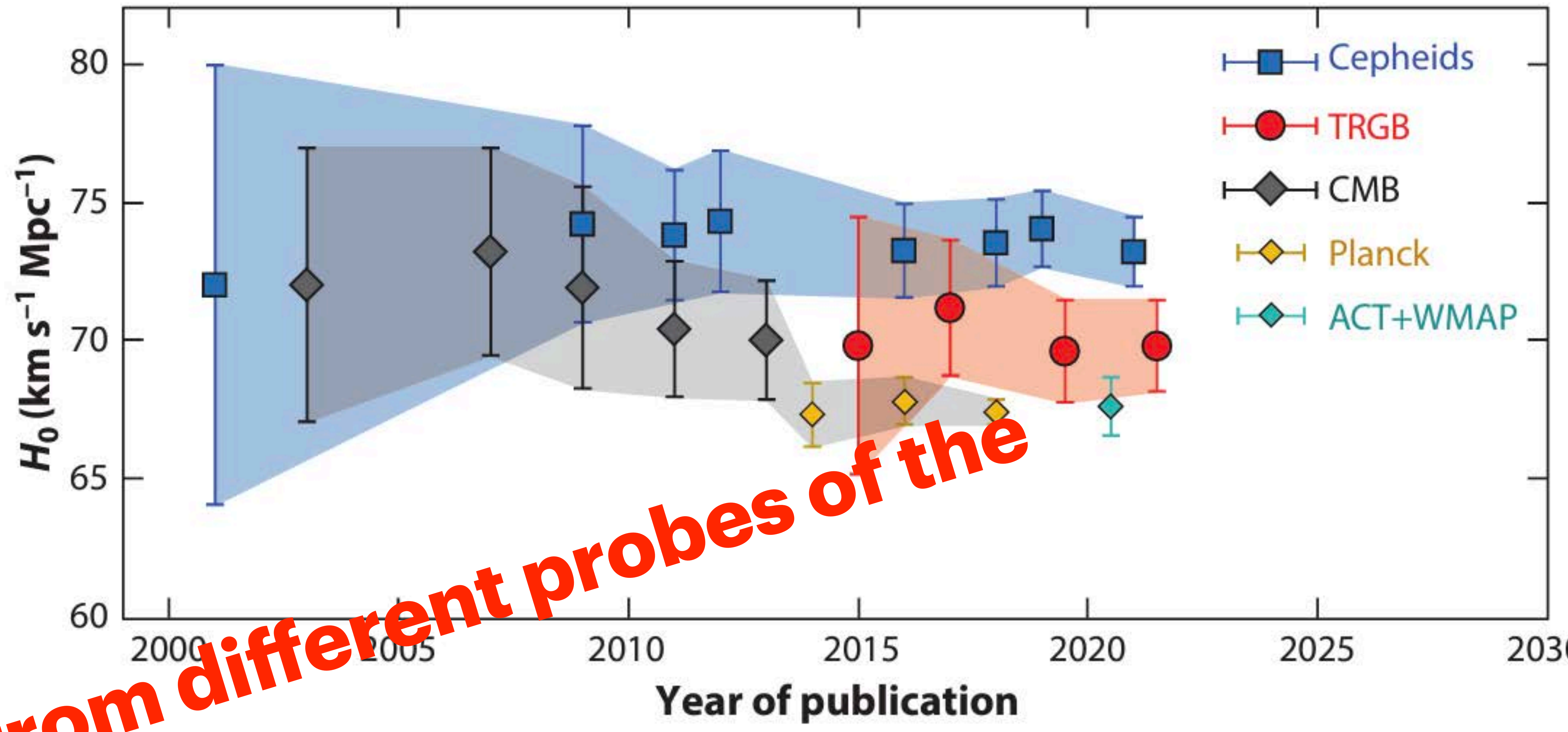




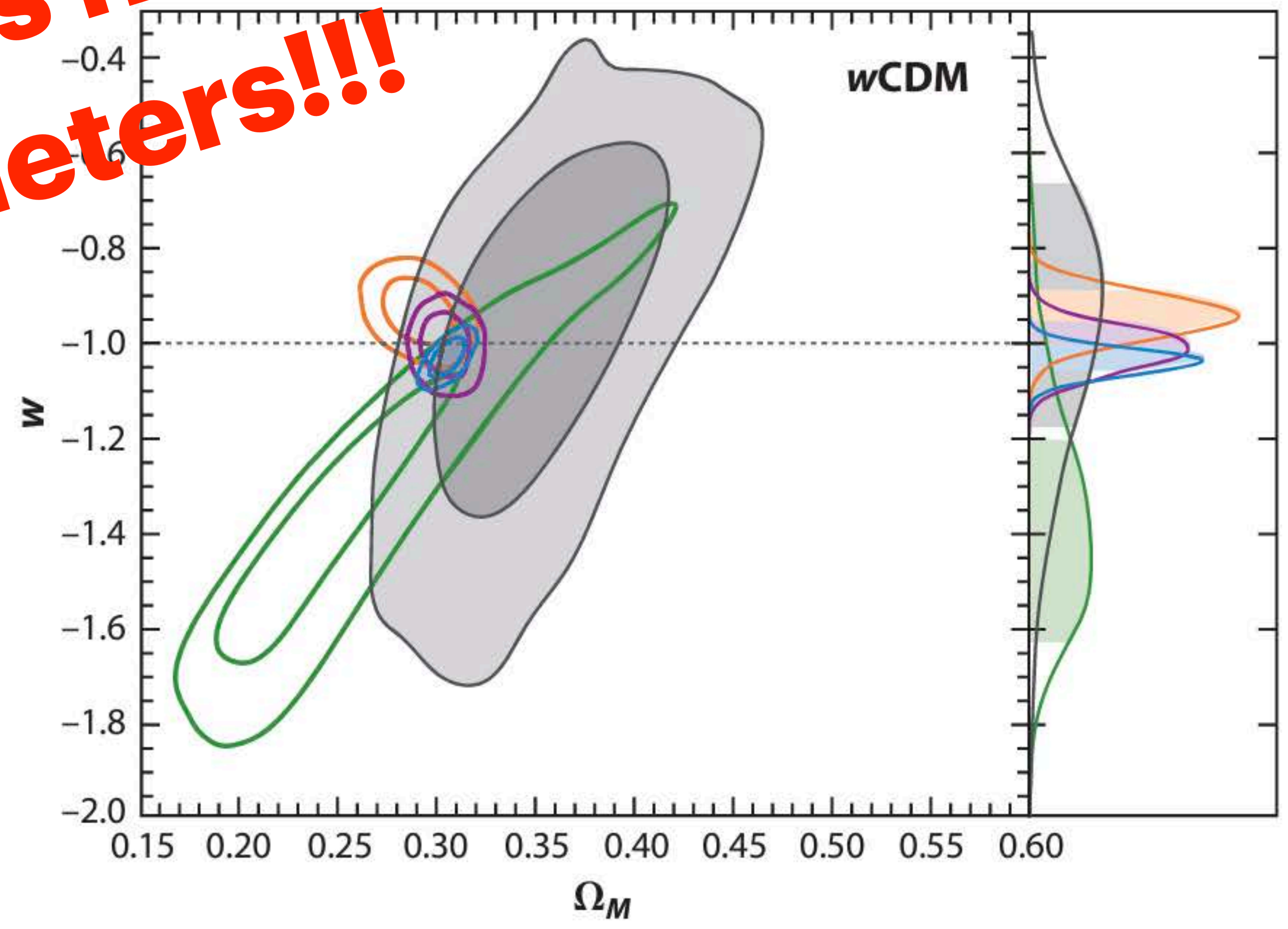


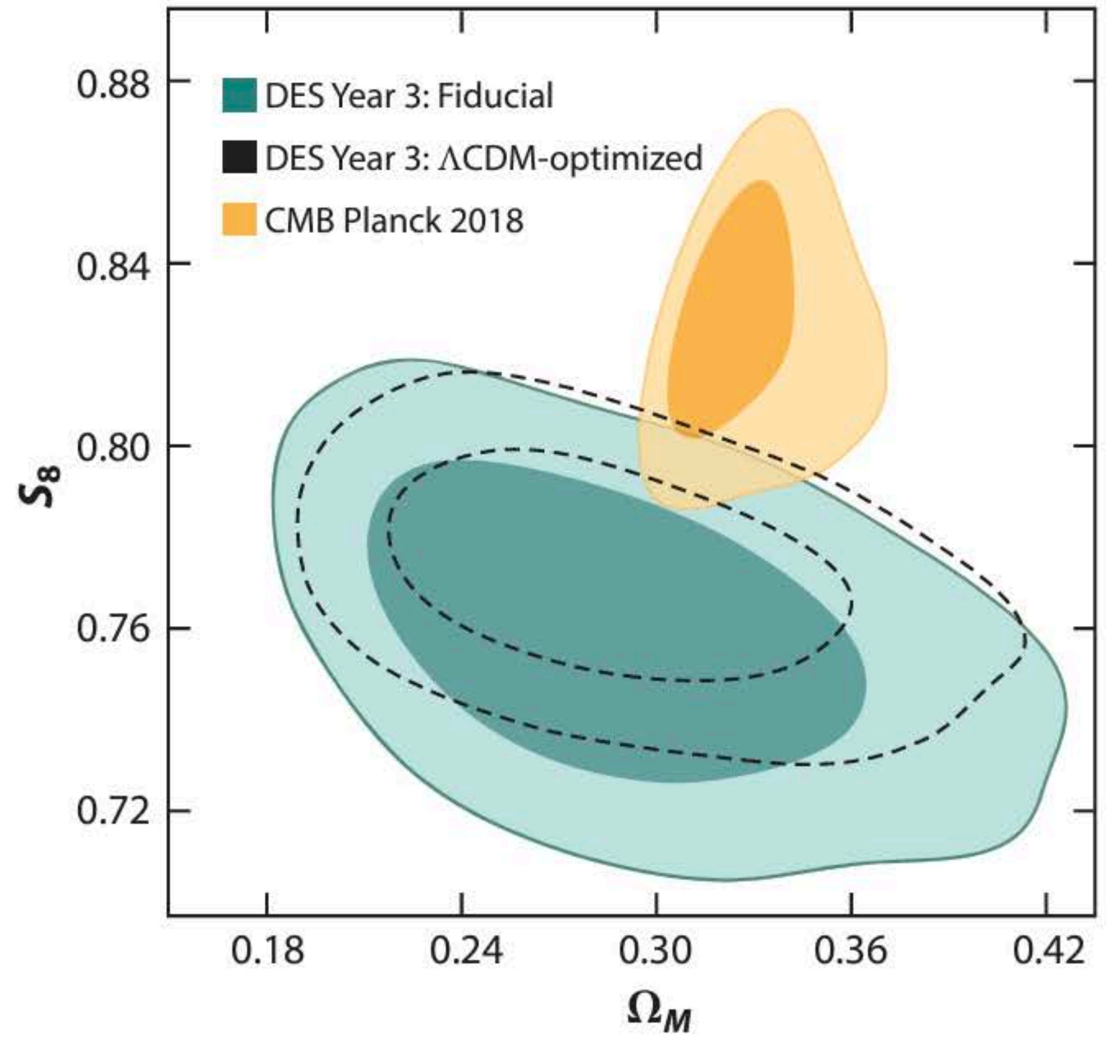
Precision Cosmology



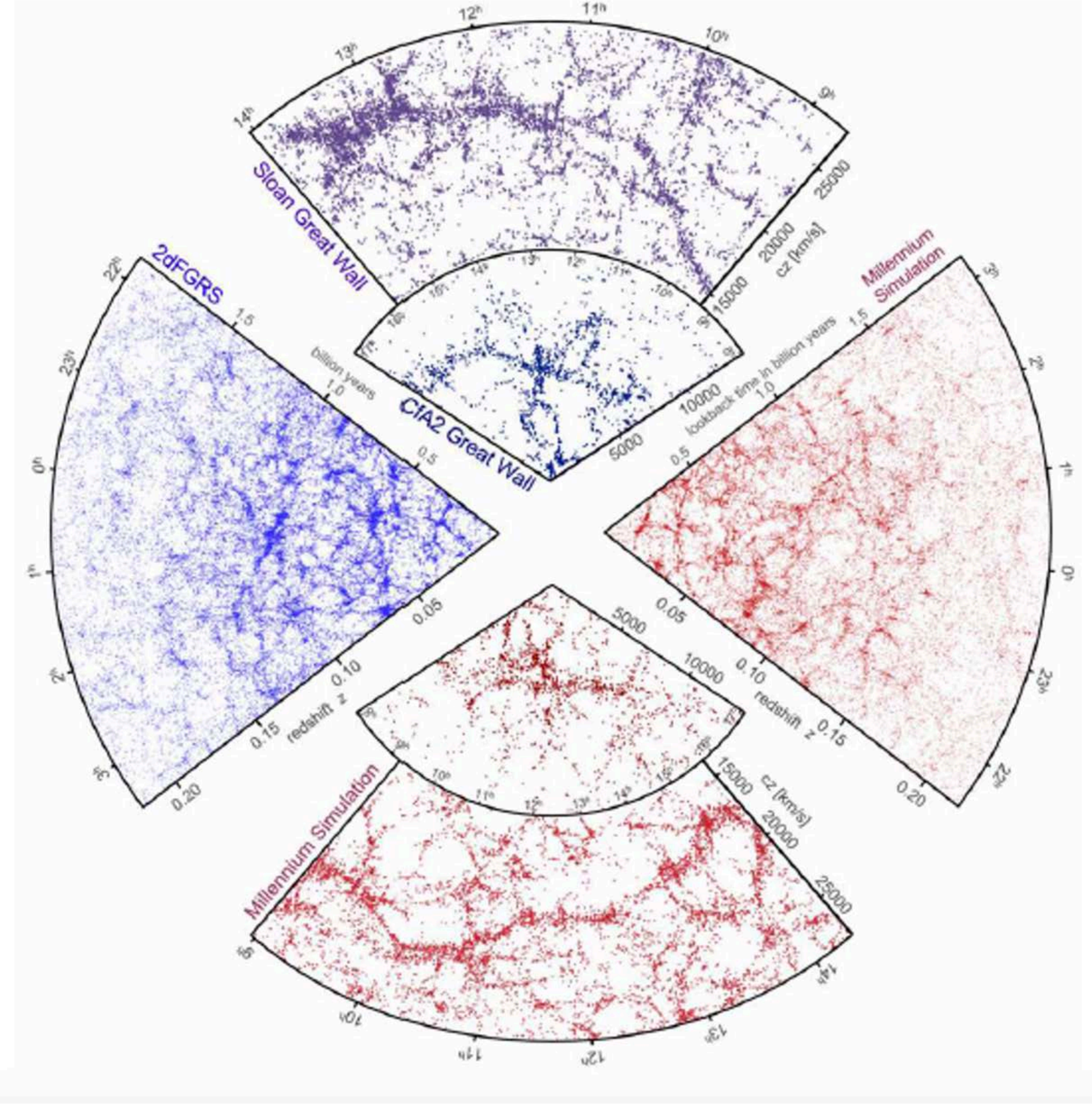


Serious discrepancies from different probes of the cosmological parameters!!!



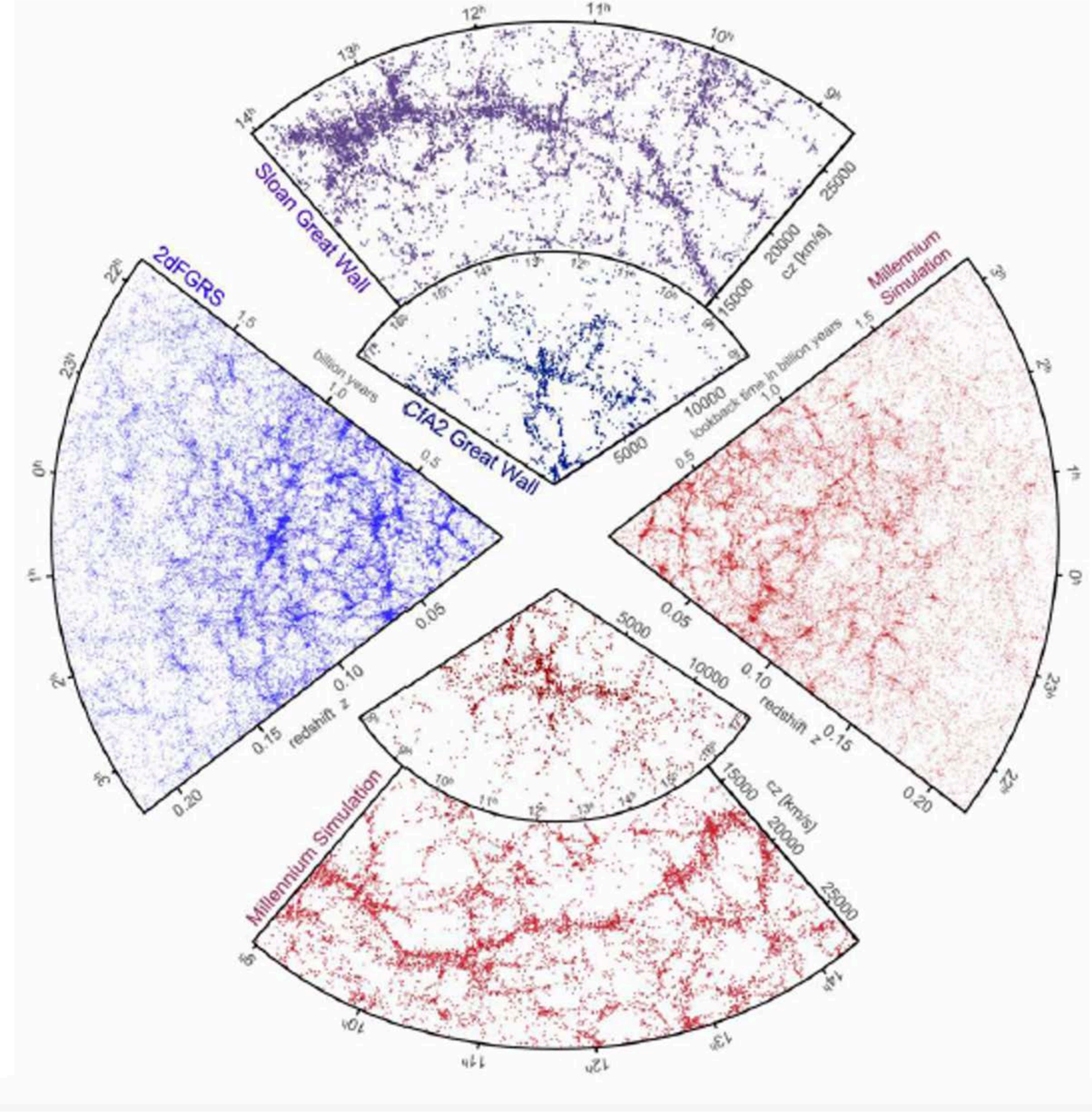


Need for numerical modelling!



Need for numerical modelling!

1. Large volumes
2. DM & baryons
3. Large number of galaxies
4. Different types of galaxies



Simulations

The need for SCC

Typical galaxy stellar masses (10^{10-11} Mo)

DM haloes ~10^{11-12}Mo

Large volumes: > 1Gpc^3

Simulations

The need for SCC

Typical galaxy stellar masses (10^{10-11} Mo)

DM haloes ~10^{11-12}Mo

Large volumes: > 1Gpc^3

Large number of particles:
2X3600^3 (DM + baryons)
>100TB snapshots + additional data:
~10s PB

Simulations

The need for SCC

Typical galaxy stellar masses (10^{10-11} Mo)

DM haloes ~10^{11-12}Mo

Large volumes: > 1Gpc^3

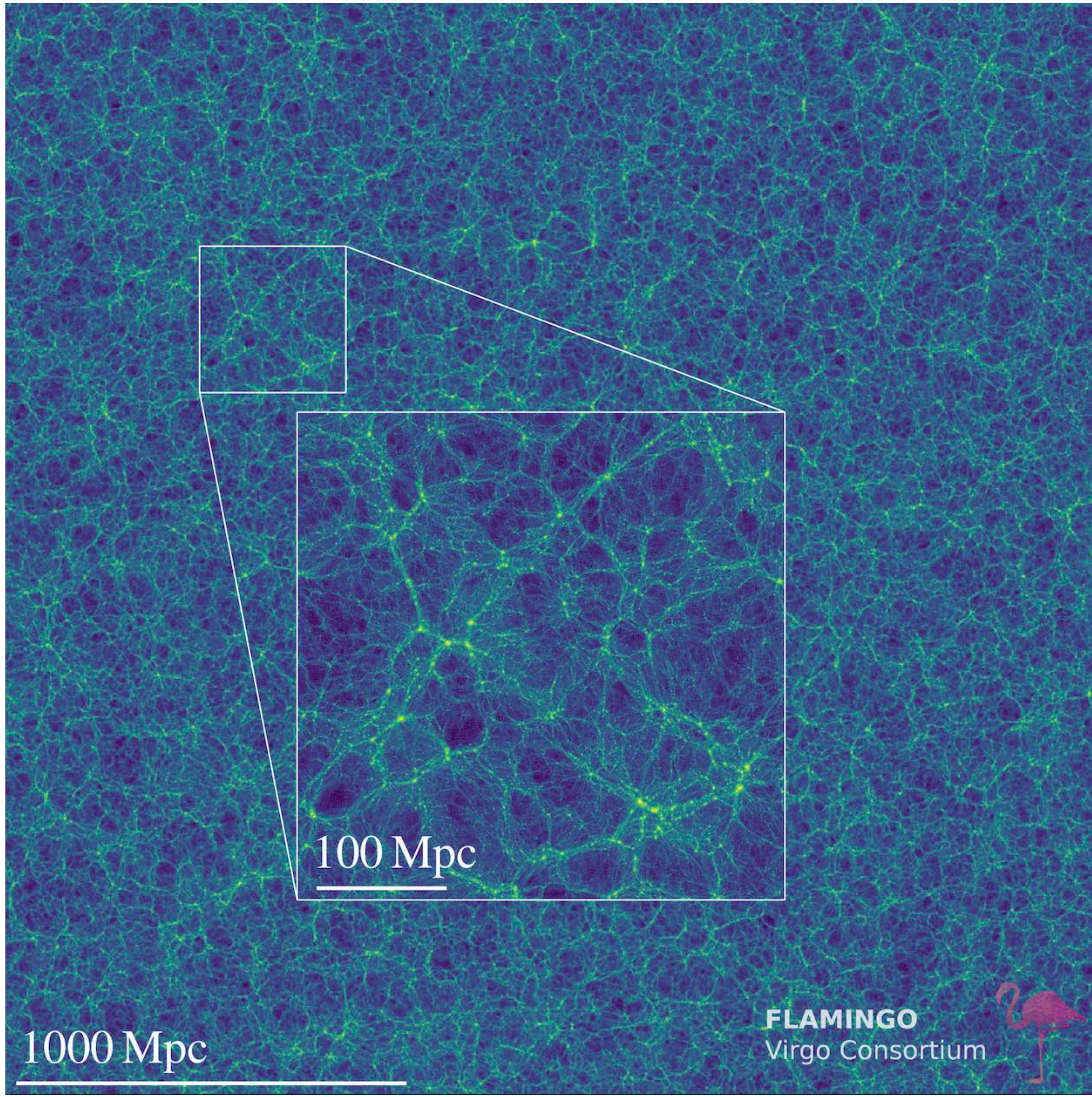
Large number of particles:
2X3600^3 (DM + baryons)
>100TB snapshots + additional data:
~10s PB

- Computationally expensive
- Difficult to handle
- Analysis
- Not all info is relevant
- Not sustainable…

Simulations (The FLAMINGO suite case)

Full-hydro Large-scale structure simulations with All-sky Mapping for the Interpretation of Next Generation Observations (Schaller et al 24)

- Flagship simulation:
- Boxsize 2.8 Gpc
- Particles: 2X5040^3 (DM + Baryons)
- +200TB per output

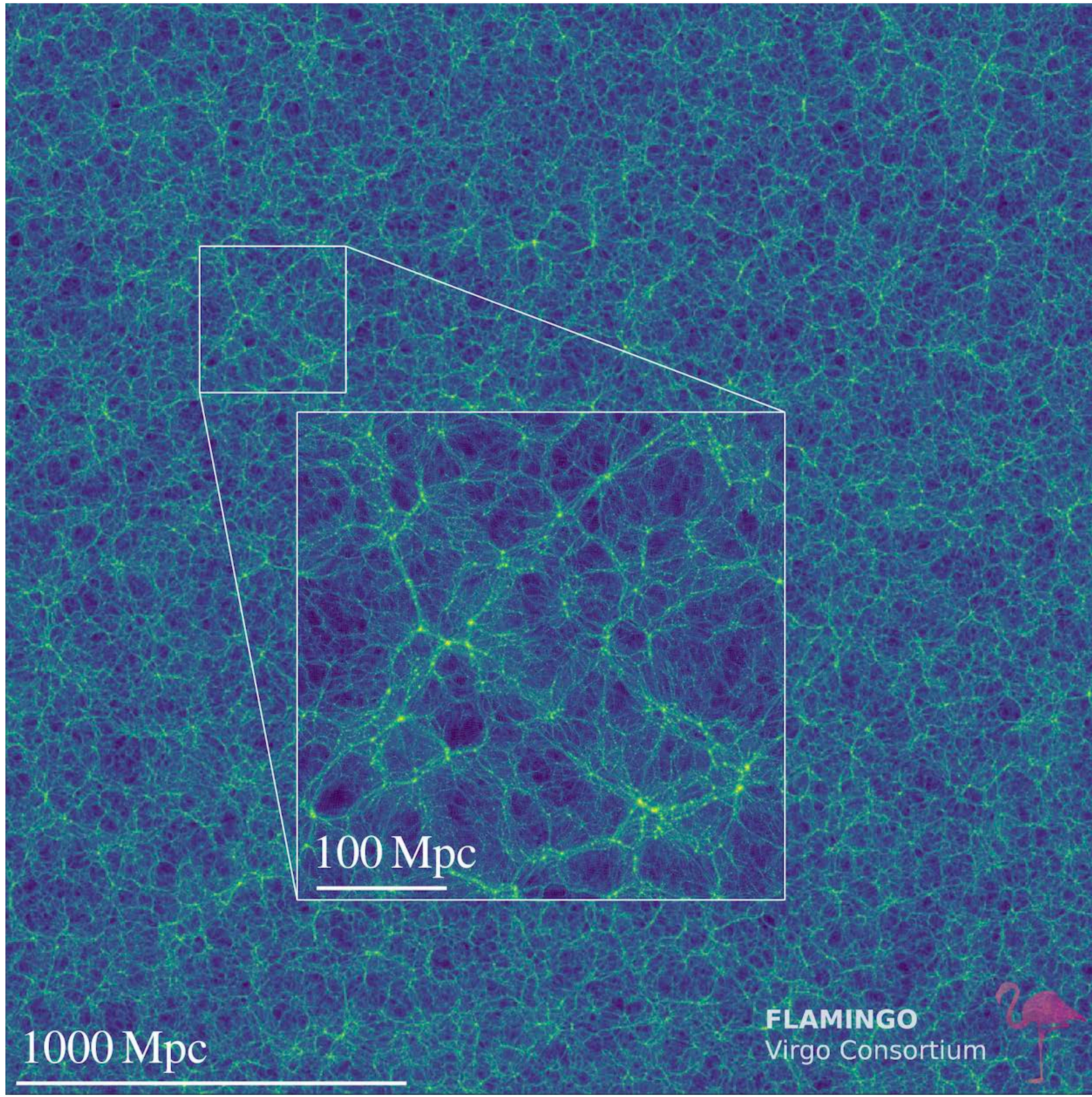


Simulations (The FLAMINGO suite case)

Optimisation: (also "typical" for other codes)

- SWIFT: new code developed from scratch, with borrowed ideas from the community
- Hybrid shared & distributed memory
- Task based parallelisation (not data based)
- aimed at exploiting modern HP cluster architectures
- Optimal parallelisation (almost linear at $>10^4$ cores)

Gravity
Cell based (Oct trees, FMM) + PM (FFT)

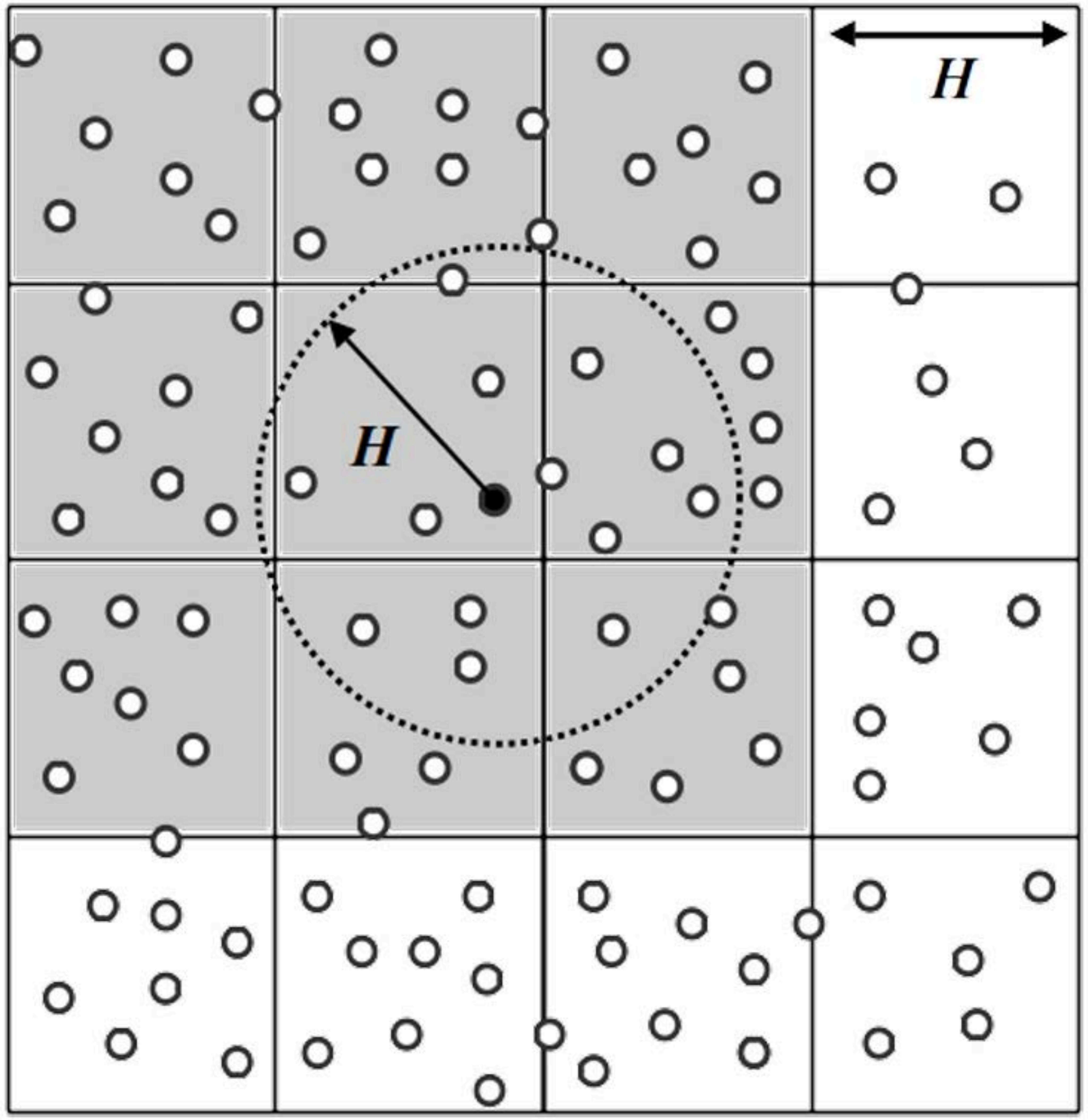


Simulations (The FLAMINGO suite case)

Optimisation: (also “typical” for other codes)

- SWIFT: new code developed from scratch, with borrowed ideas from the community
- Hybrid shared & distributed memory
- Task based parallelism (not data based)
- aimed at exploiting modern HP cluster architectures
- Optimal parallelisation (almost linear at >10^4 cores)

  Gravity
  Cell based (Oct trees, FMM) + PM (FFT)
  all operations vectorised over nodes



Simulations (The FLAMINGO suite case)

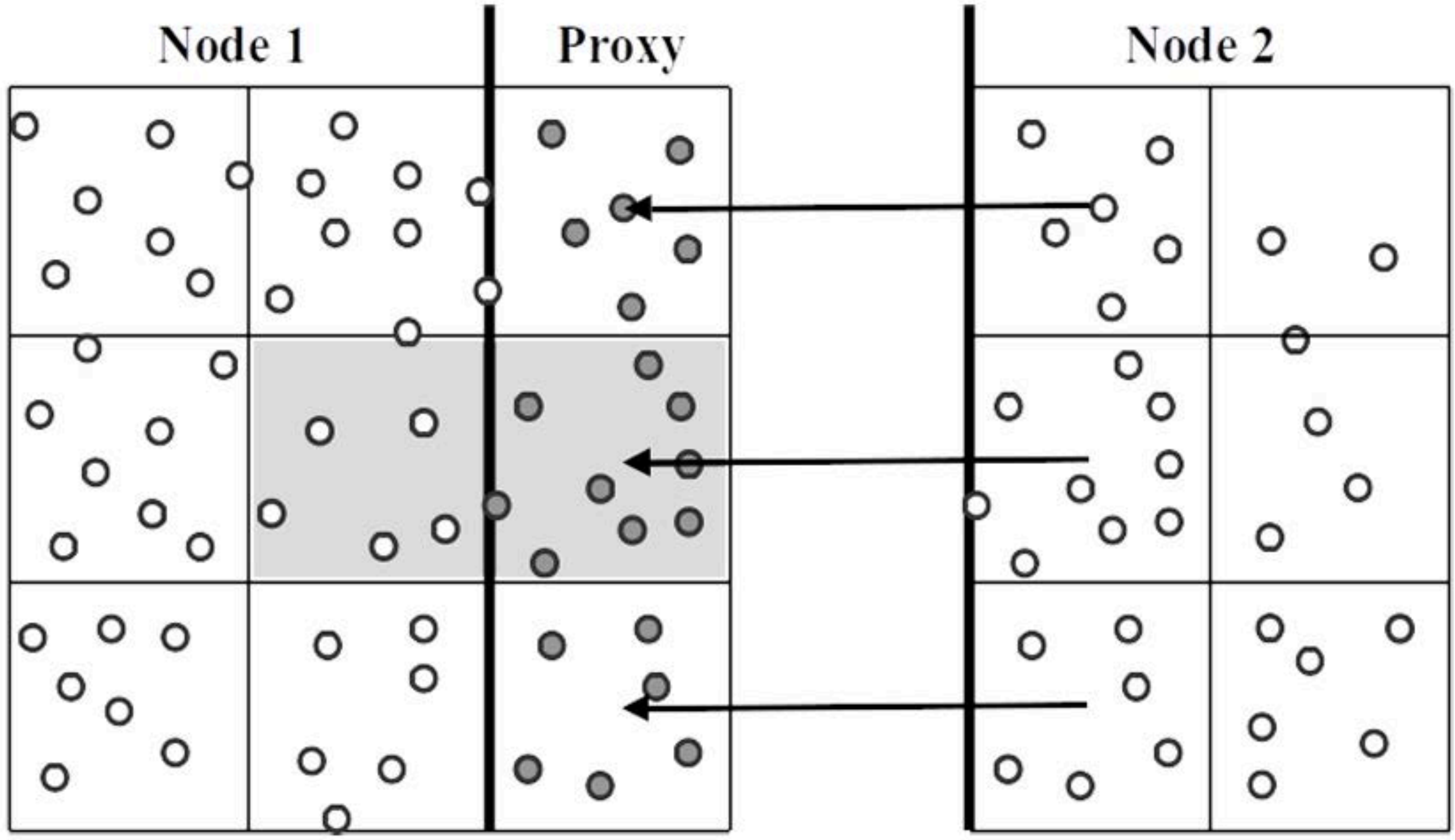
Optimisation: (also "typical" for other codes)

- SWIFT: new code developed from scratch, with borrowed ideas from the community
- Hybrid shared & distributed memory
- Task based parallelism (not data based)
- aimed at exploiting modern HP cluster architectures
- Optimal parallelisation (almost linear at >10^4 cores)

  Gravity

  Cell based (Oct trees, FMM) + PM (FFT)

  minimal communication / nodes



Simulations

Data storage

- Large number of tasks => compressed HDF5 format (Gadget-like type) (The HDFGroup 2022): **x**, **v**, mass, ID….
- Metadata: units, data decomposition
- Particles are stored in the snapshots in order of the domain cells they belong to (easily retrieved specific regions: haloes)
- Output:
  - Single file: all nodes writing in parallel
  - Multiple files: every node write their own
  - Check-points, each task writes
- Example: without compressing: 960 files, 64 TB ==> 260s @ 250GB/s!
  compressing: 11 TB ==> 1260s @ 9GB/s

Simulations

Data storage

- Large number of tasks => compressed HDF5 format (Gadget-like type) (The HDFGroup 2022): **x**, **v**, mass, ID….
- Metadata: units, data decomposition
- Particles are stored in the snapshots in order of the domain cells they belong to (easily retrieved specific regions: haloes)
- Output:
  - Single file: all nodes writing in parallel
  - Multiple files: every node write their own
  - Checkpoints, each task writes

Awesome!

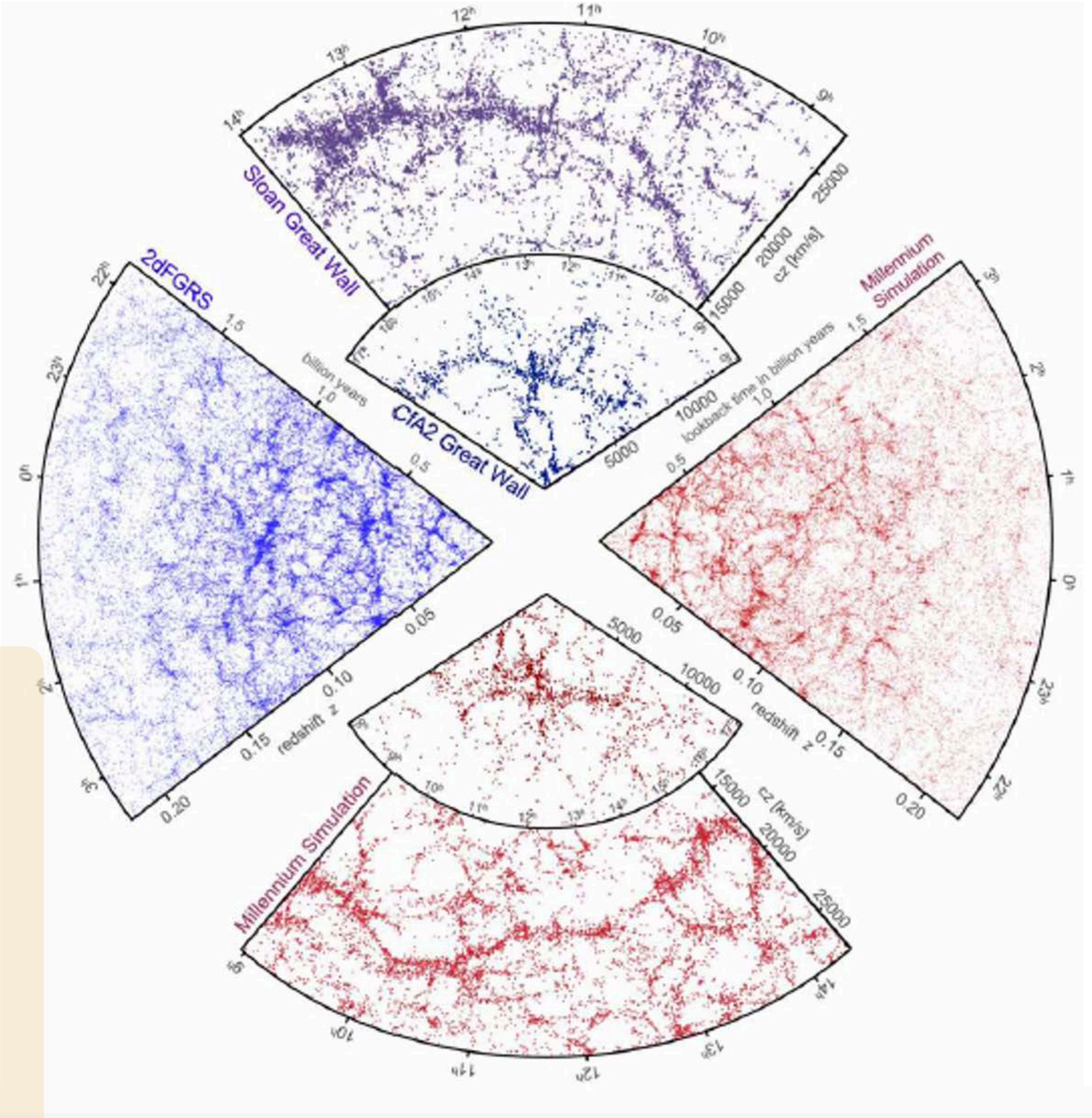
But not good enough!!!

- Example: without compressing: 960 files, 64 TB ==> 260s @ 250GB/s!
  compressing: 11 TB ==> 1260s @ 9GB/s

Simulations

Data storage: Clever ways :)

- Light cones: mimicking real observations



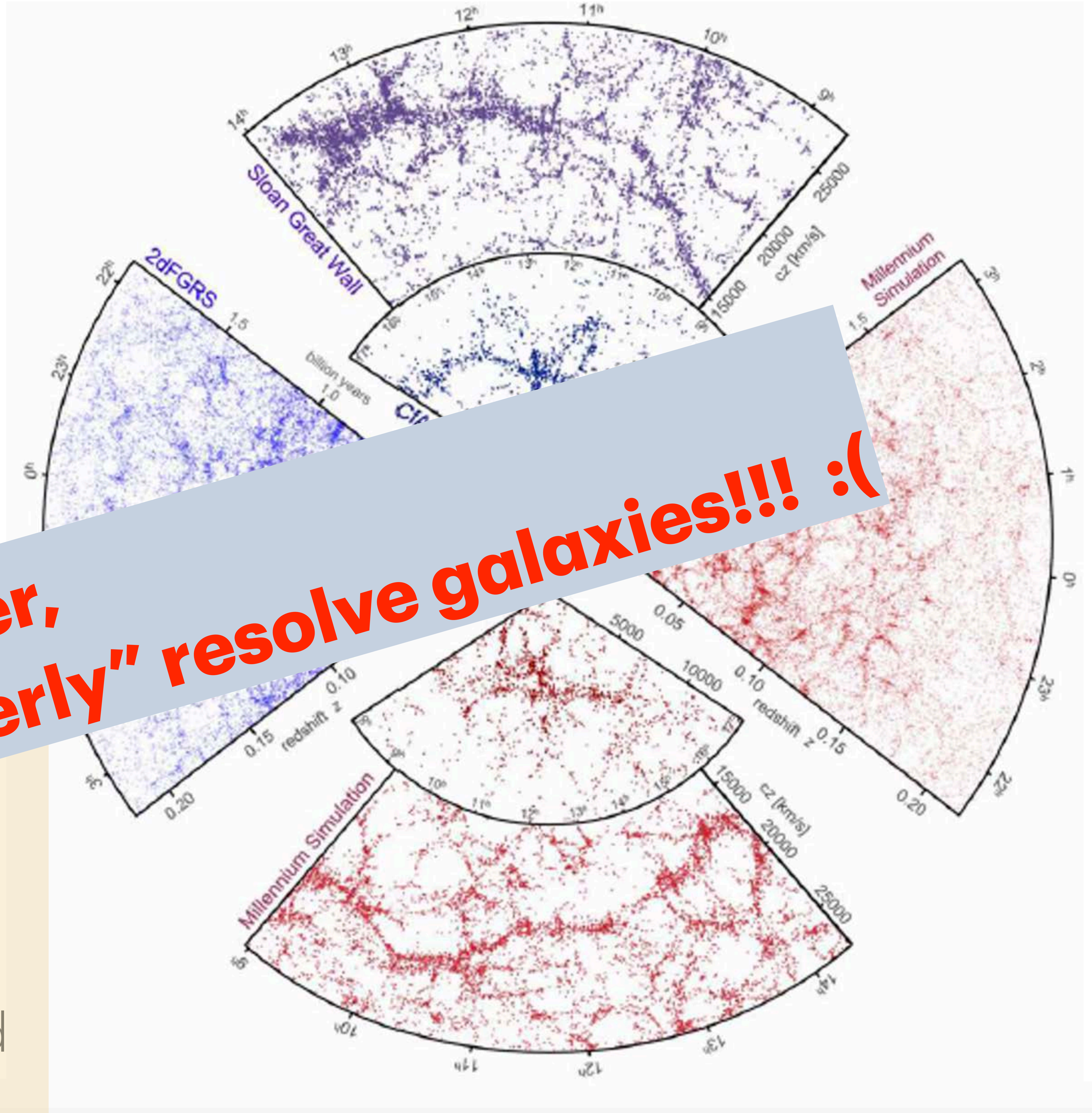
- Save different data type
  - Structure on the fly
    Halo finders
    LSS
  - Properties
    Density field on a grid
    Power spectrum

- Continuous Simulation Data Stream (CSDS): create a database of updates only for the particles needed Hausammann et al. (2022)

Simulations

Data storage: Clever ways :)

- Light cones: mimicking real observations



- Save diff…

  - S…
  - Properties
    Density field on a grid
    Power spectrum

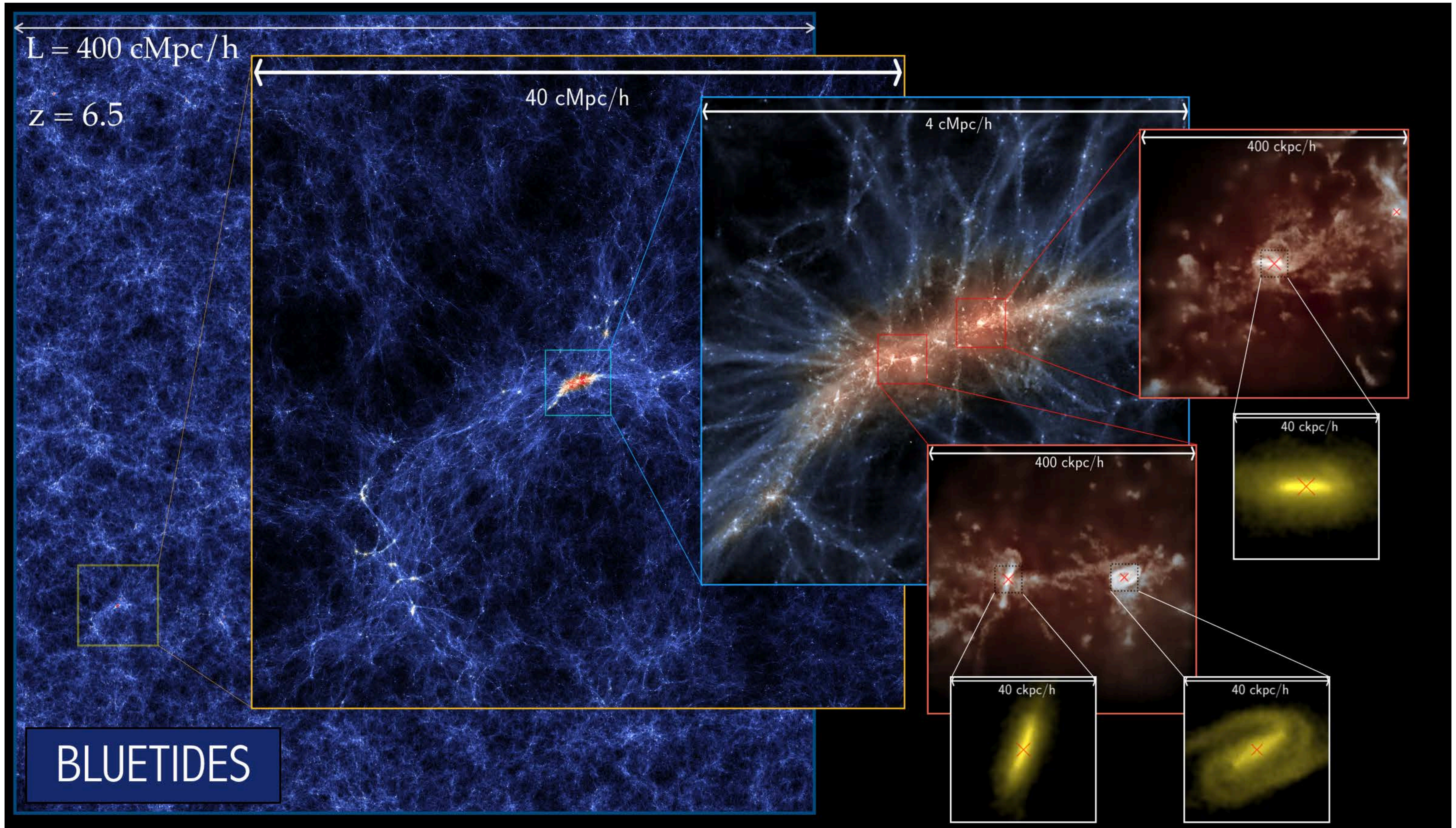
However,
It is still NOT enough to “properly” resolve galaxies!!! :(

- Continuous Simulation Data Stream (CSDS): create a database of updates only for the particles needed Hausammann et al. (2022)

Simulations: Astrophysics…

The need for "more accurate" modelling: small scales

- Zoom-in simulations of particular regions of interest,



High resolution: ~10^7particles

Accurate follow up of gastrophysics

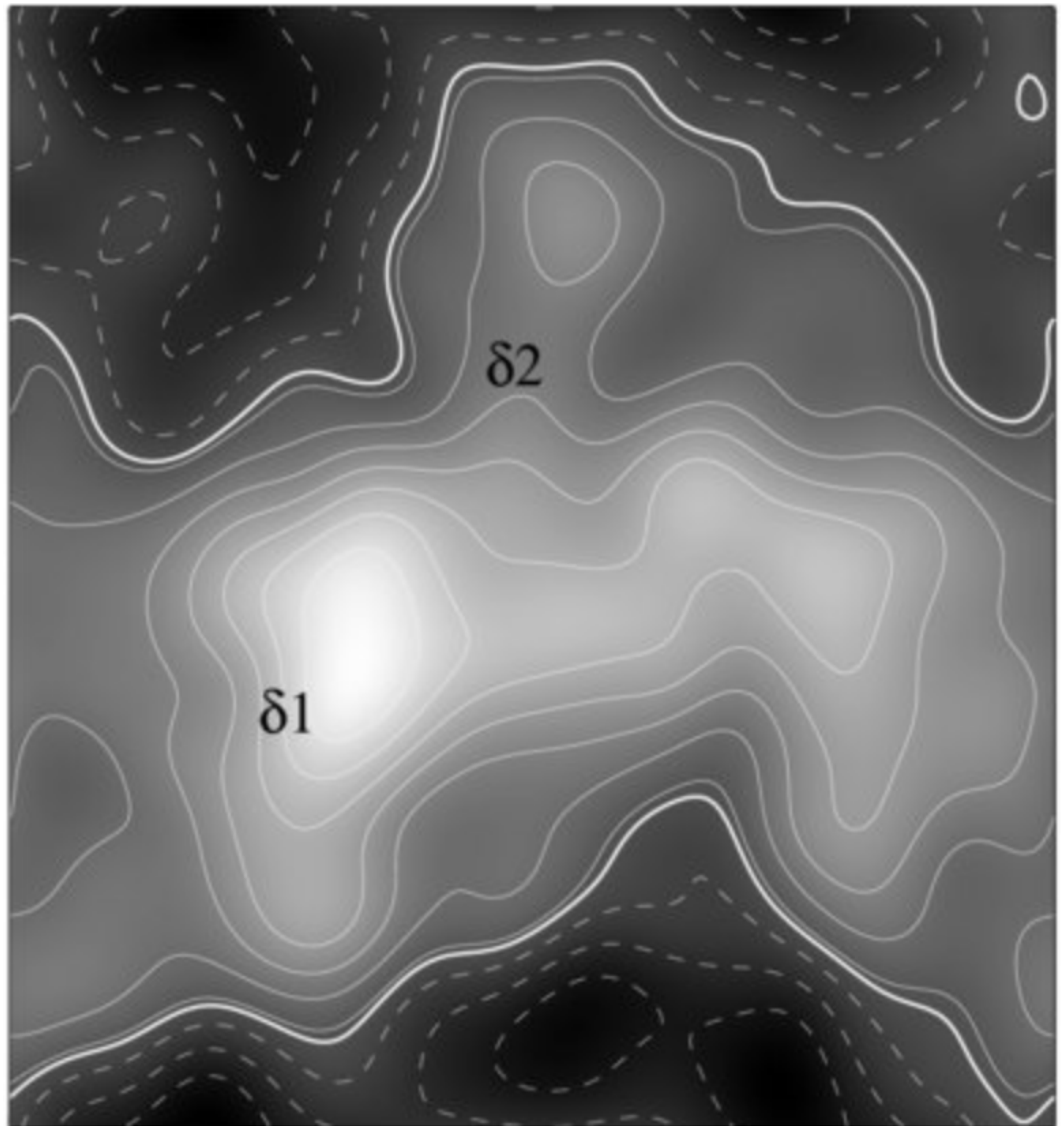
Bi, Shlosman, Romano-Diaz (2022)

Simulations: Astrophysics...

The need for "more accurate" modelling: small scales

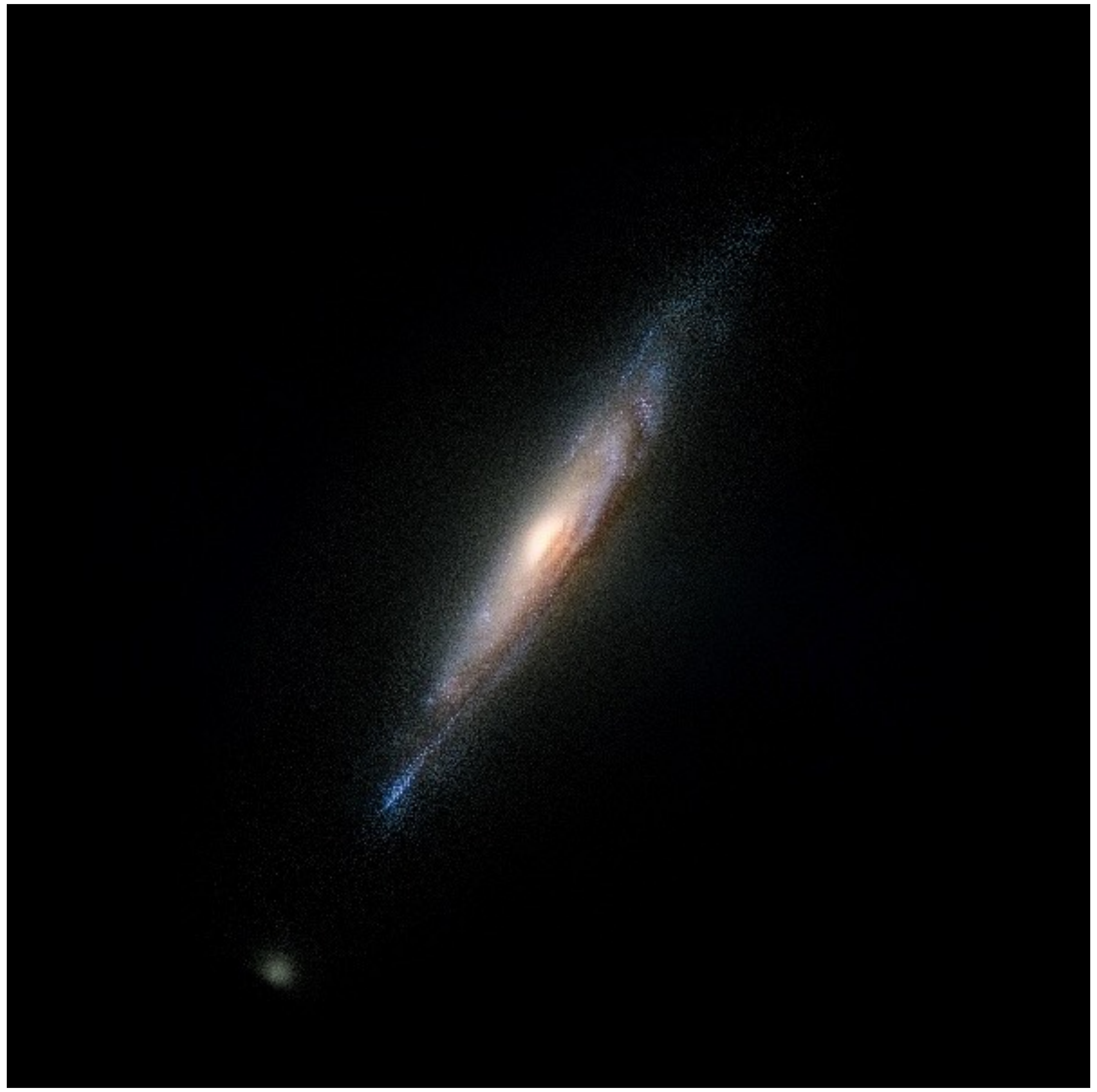
- Constrained realizations: design your own object!
  Zoom-in simulations of particular regions of interest,

High resolution: ~10^7particles
Accurate follow up of gastrophysics



One single halo
very efficient use of resources
Detailed time evolution
Total data: few TBs

Romano-Diaz, Shlosman et al (2006, 11, 14)



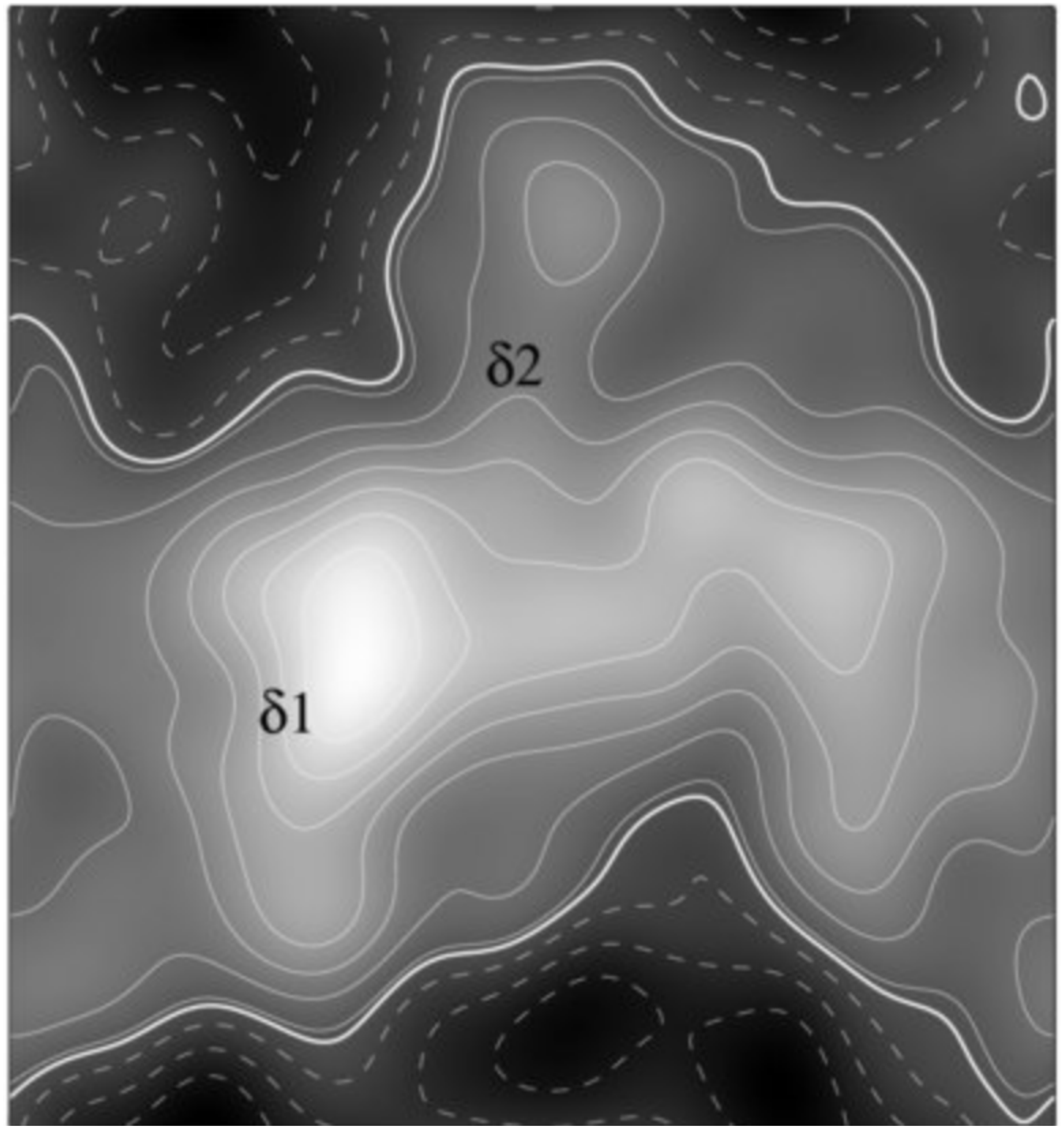
Romano-Diaz et al (2017)

Simulations: Astrophysics...

The need for "more accurate" modelling: small scales

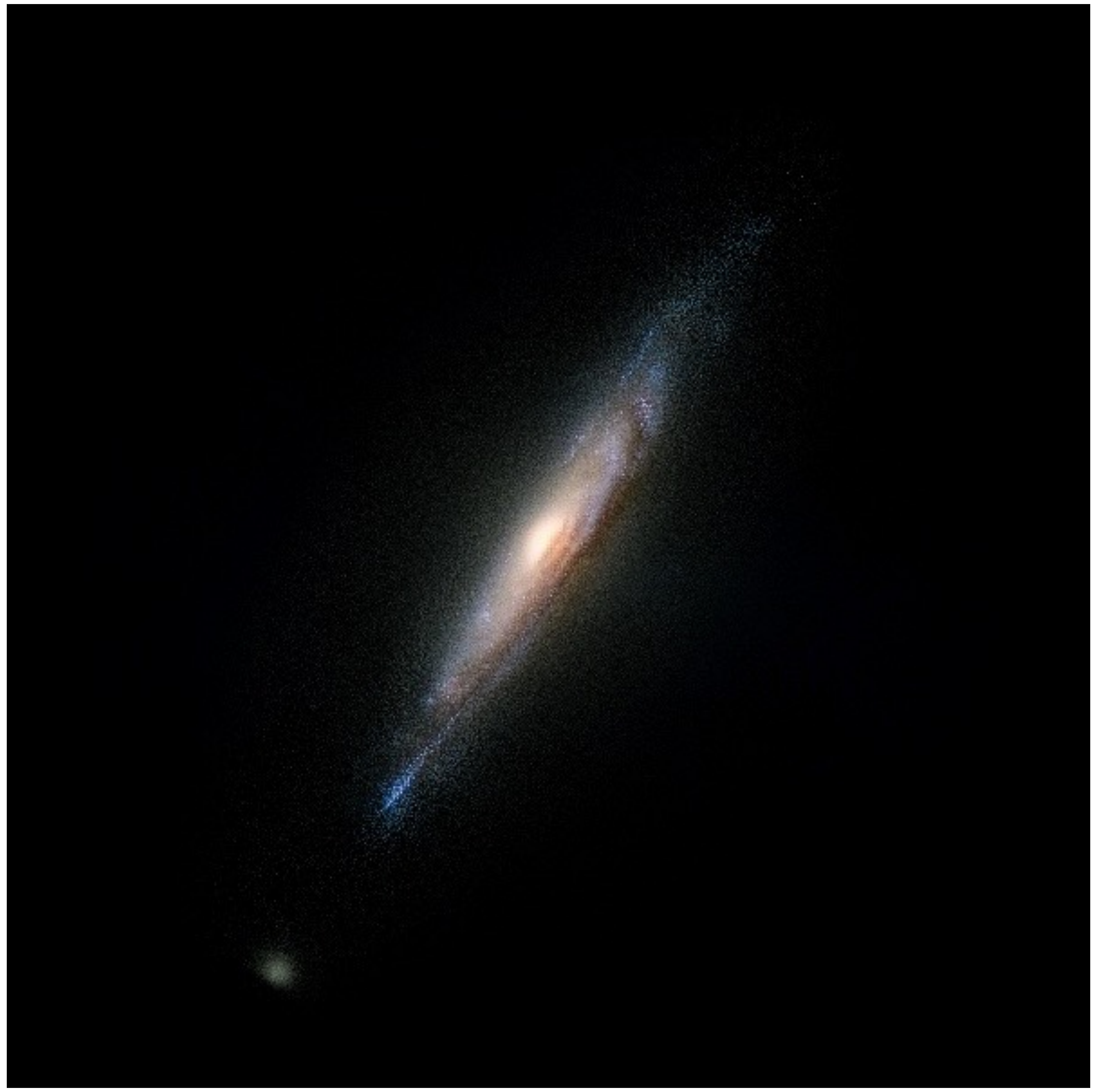
- Constrained realizations: design your own object!
  Zoom-in simulations of particular regions of interest,

High resolution: ~10^7particles
Accurate follow up of gastrophysics



One single halo
very efficient use of resources
Detailed time evolution
Total data: few TBs
Drawback: only few objects

Romano-Diaz, Shlosman et al (2006, 11, 14)



Romano-Diaz et al (2017)

Simulations

The need for "emulators": calibration of the physics...

Use of small scale simulation and ML (Gaussian process emulators) (Kugel et al. 2023)

Fitting of sub-grid parameters to the calibration data

+

Data augmented from small scale simulations

Gaussian emulators trained on N-node latin hypercube simulations

A different emulator is build for each variable

→

Emulator predictions are fit to the data via MCMC of the parameter space accounting for errors in data and the emulator

Putting everything together

- Precision Cosmology is driving the field at the theoretical, numerical and observational levels.
- The need to enter into the Exascale era drive the theoretical & numerical fields to be more demanding, computationally more powerful and friendlier.
- In need of “new ways” of coding (need for expert programmers) capable of exploiting the present and future SC capabilities.
- In combination with Bayesian statistics one can design a more complete, meaningful physical machinery…
- But we are still far from being there….