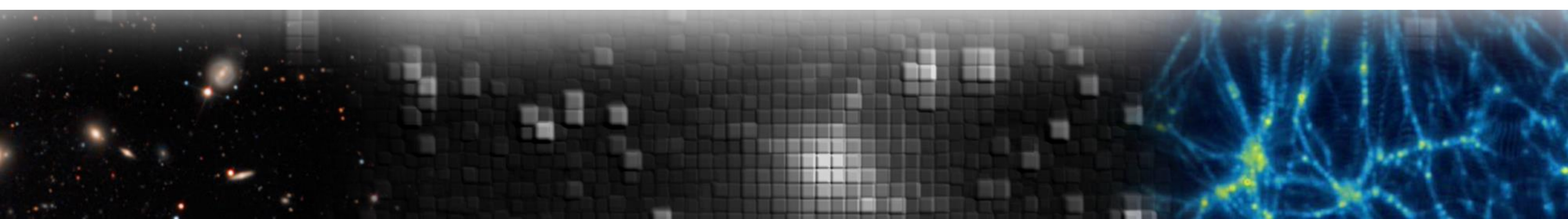
Revealing the Local Cosmic Web from Galaxies by Deep Learning

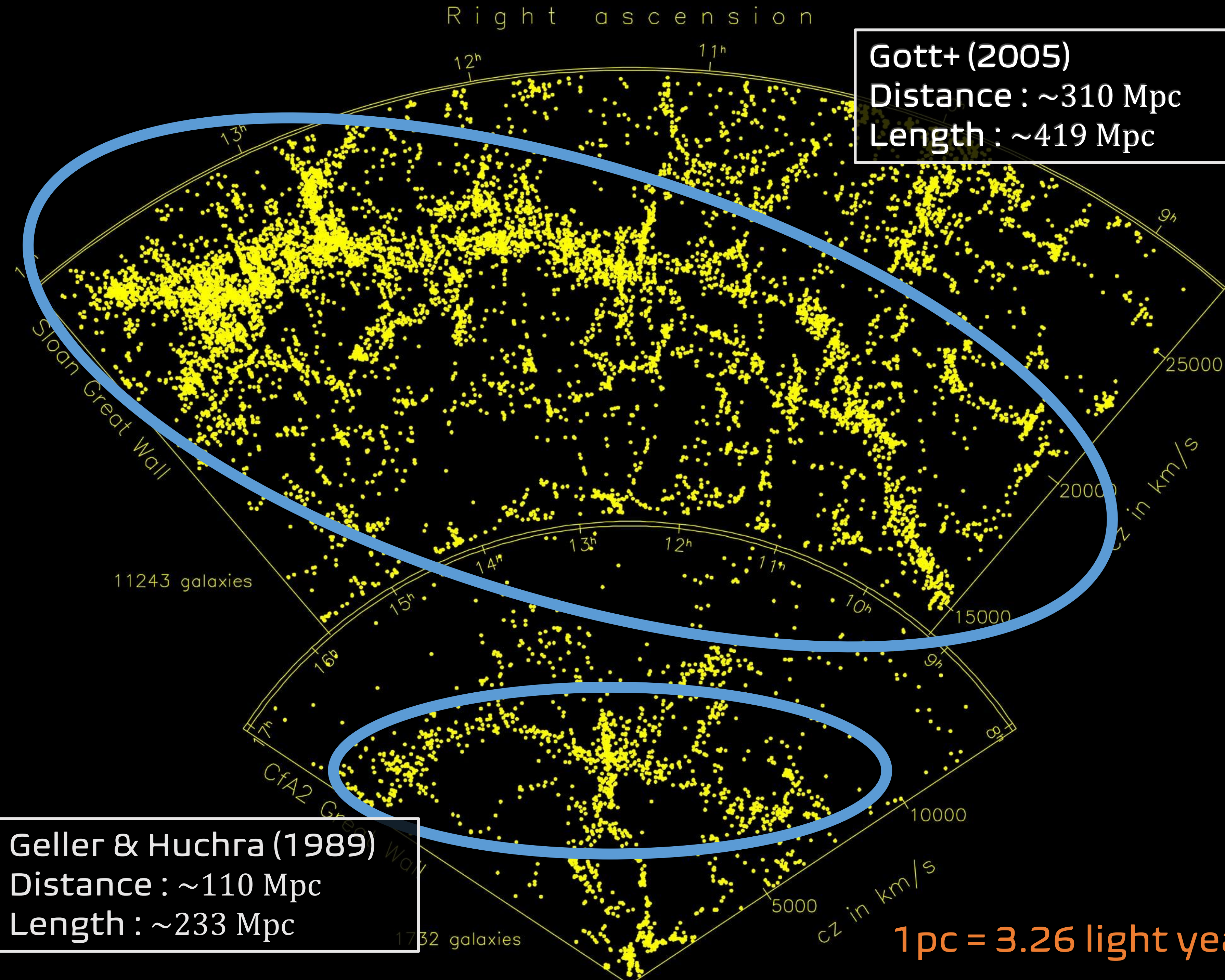
Sungwook E. Hong (KASI)
with Donghui Jeong (PSU), Ho Seong Hwang (SNU) & Juhan Kim (KIAS)

@ Yonsei Lab for Dark Universe Special Lectures

Mar. 30th, 2021



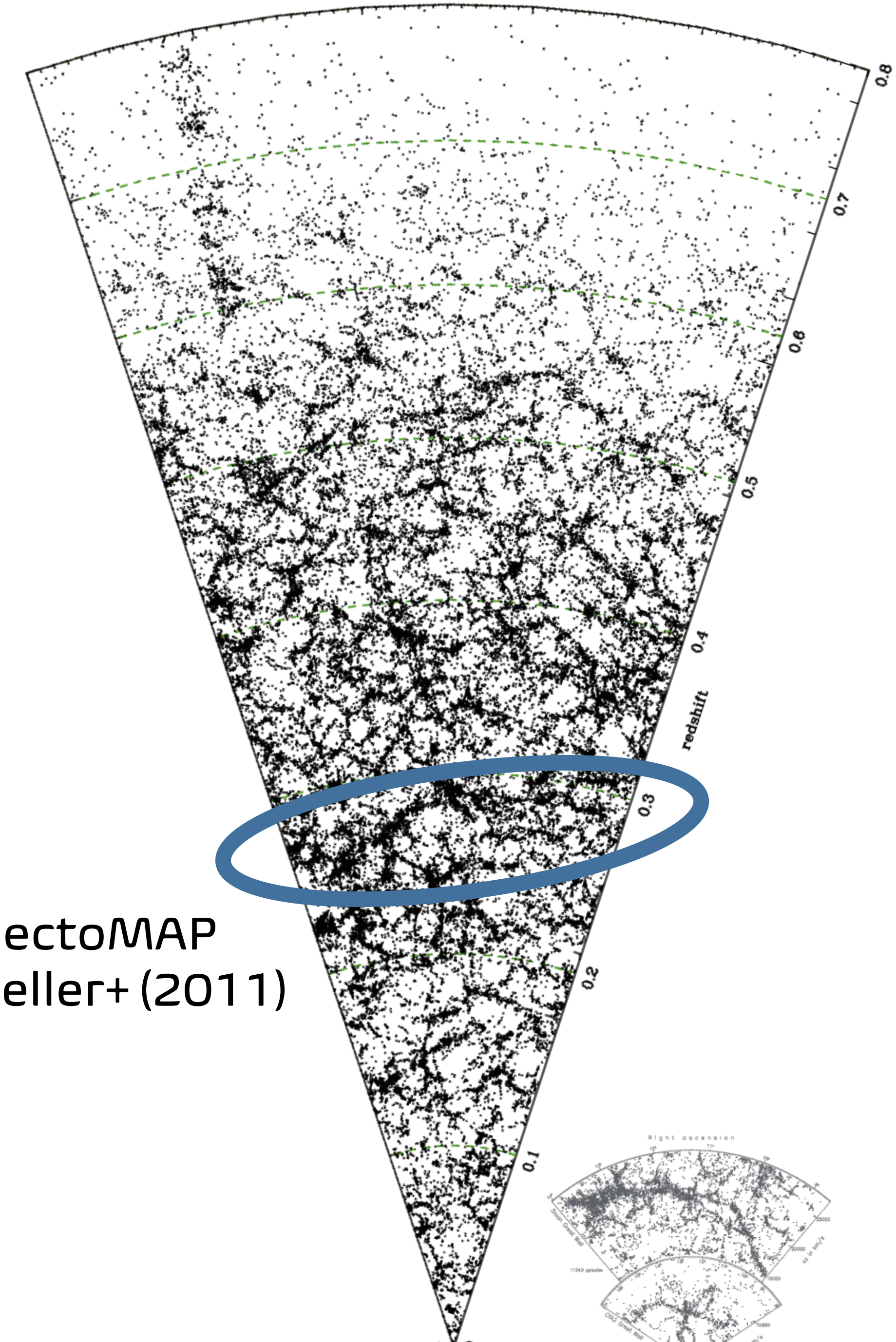
Horizon Run 5 (Lee & Shin+ 2020)



Gott+ (2005)
Distance : ~310 Mpc
Length : ~419 Mpc

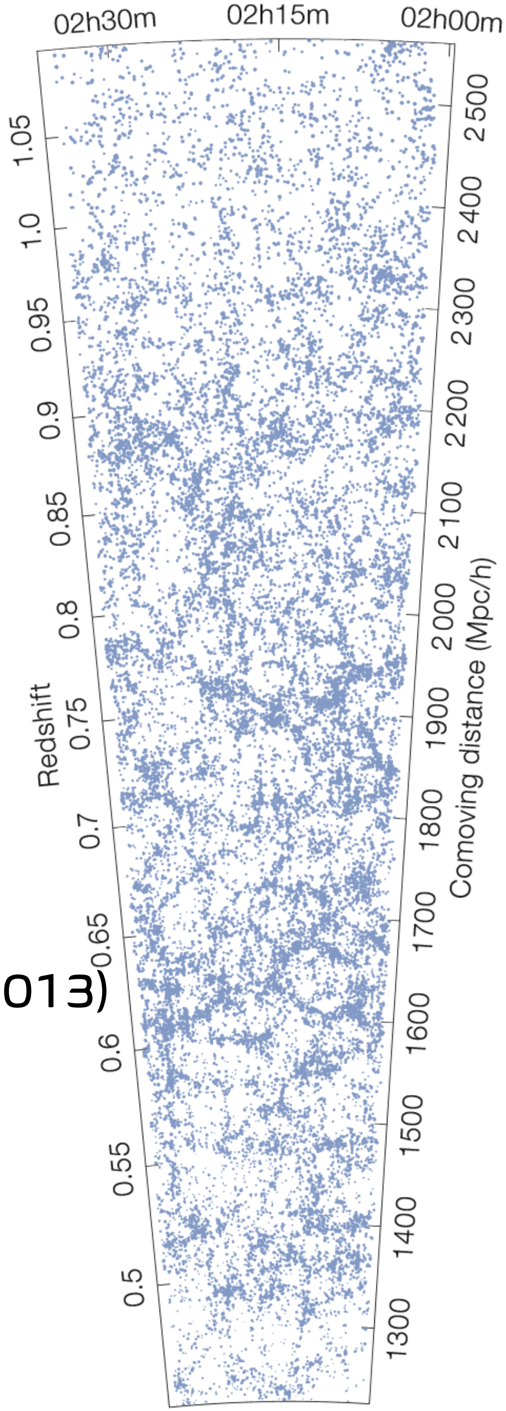
Geller & Huchra (1989)
Distance : ~110 Mpc
Length : ~233 Mpc

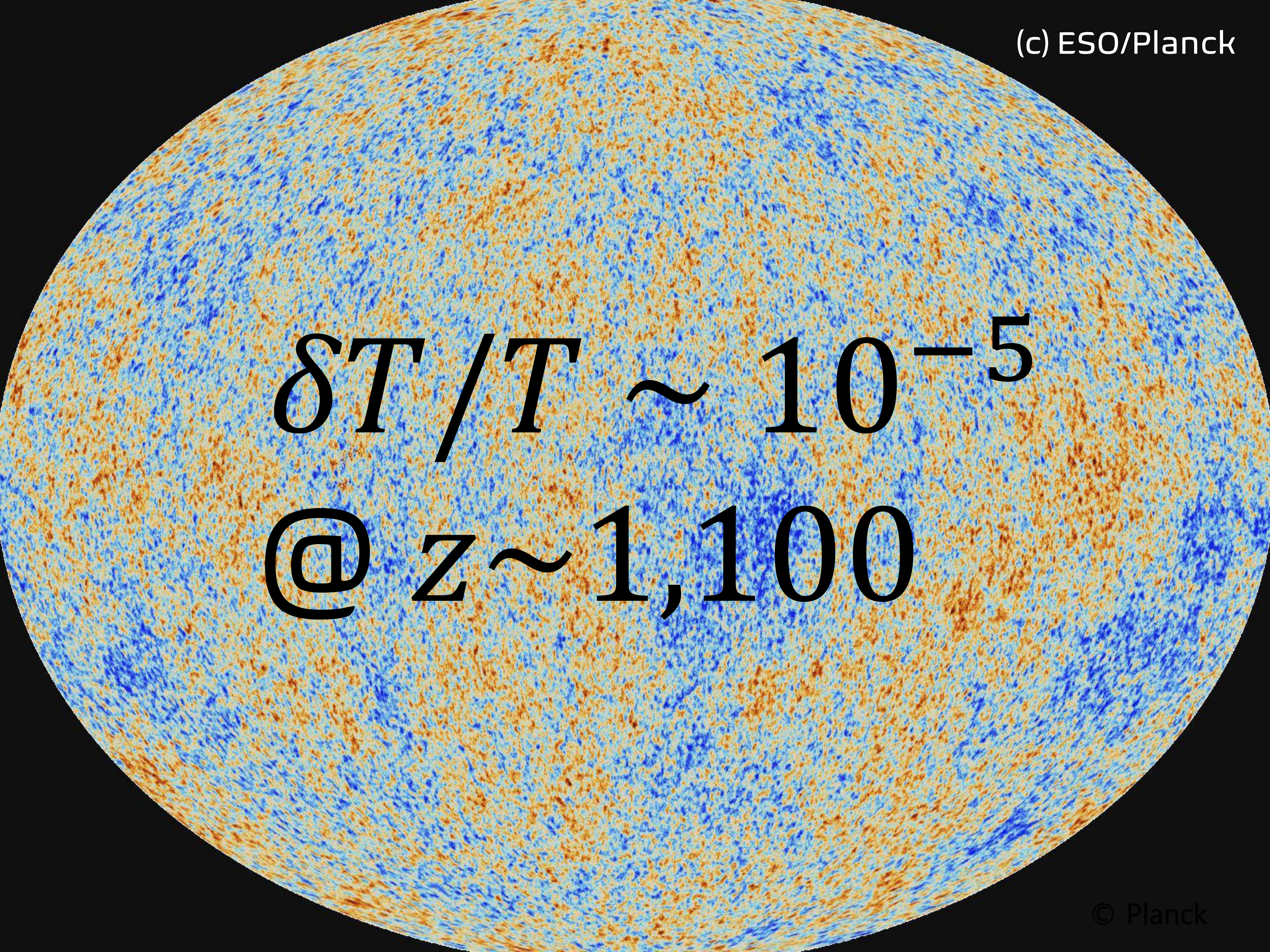
1pc = 3.26 light year



HectoMAP
Geller+ (2011)

VIPERS
Guzzo+ (2013)



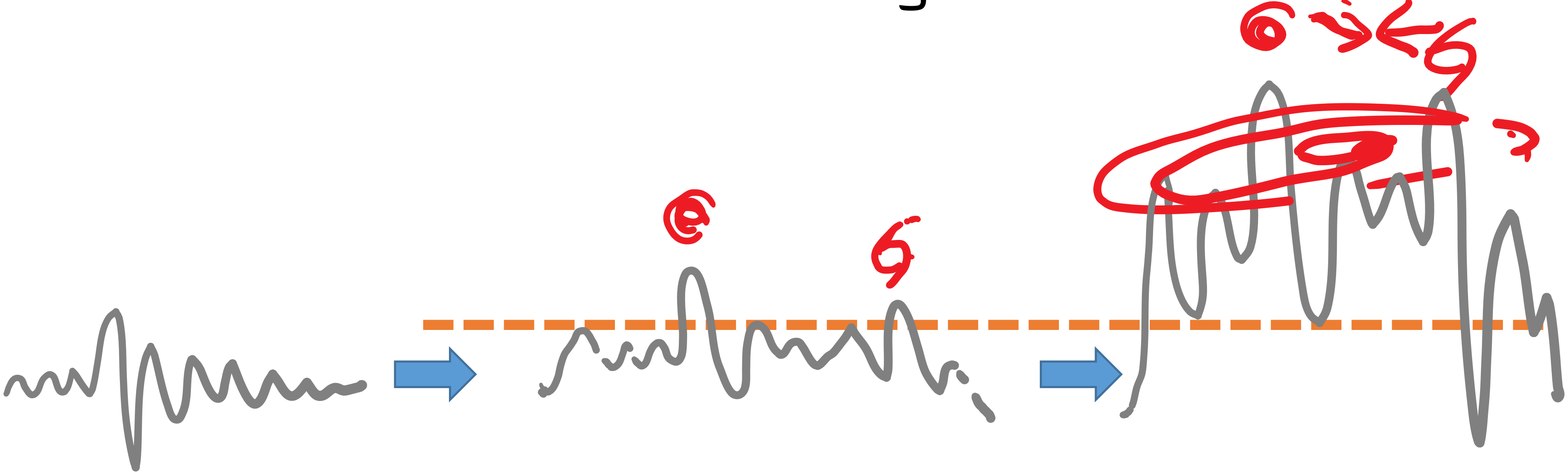


$\delta T / T \sim 10^{-5}$

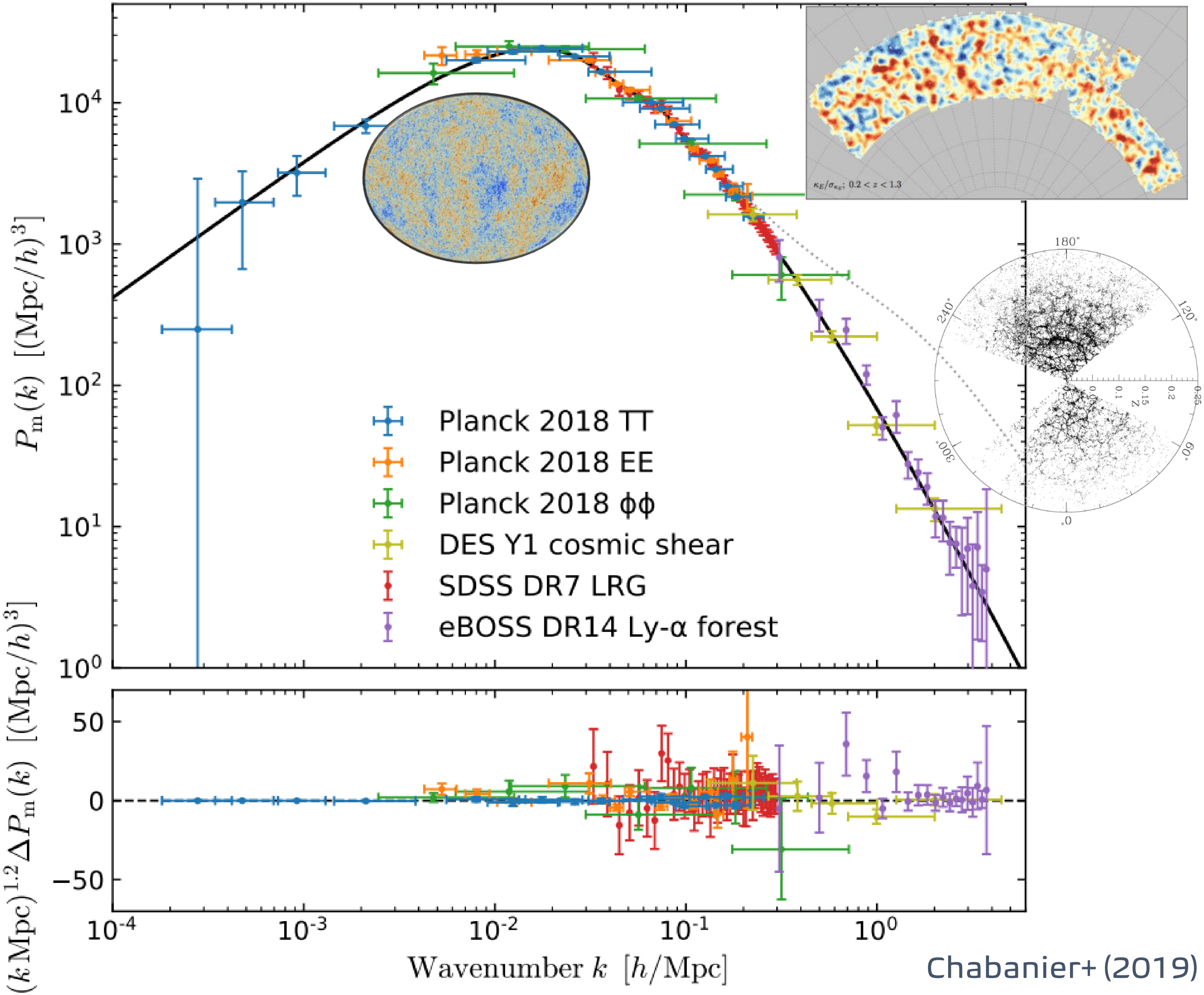
@ $z \sim 1{,}100$

Formation of Large-Scale Structures

- Gravity + Cosmic Expansion
  → Dense becomes denser, sparse becomes more sparse
- Once density is greater than a certain threshold, matter will gravitational collapse irrespective to the cosmic expansion
- Small-scale structures form larger-scale structures

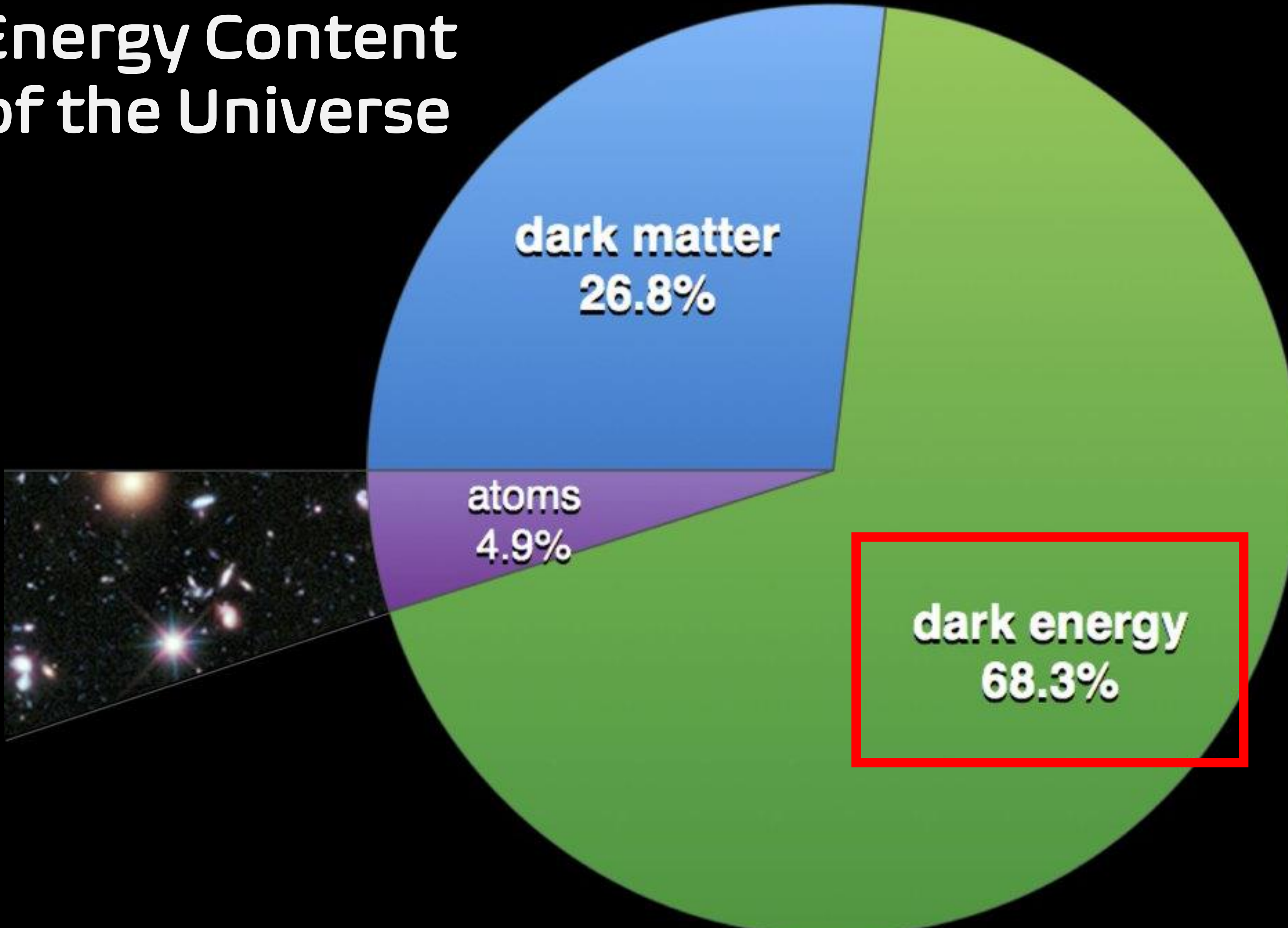


Horizon Run 5 (Lee & Shin+ 2020)

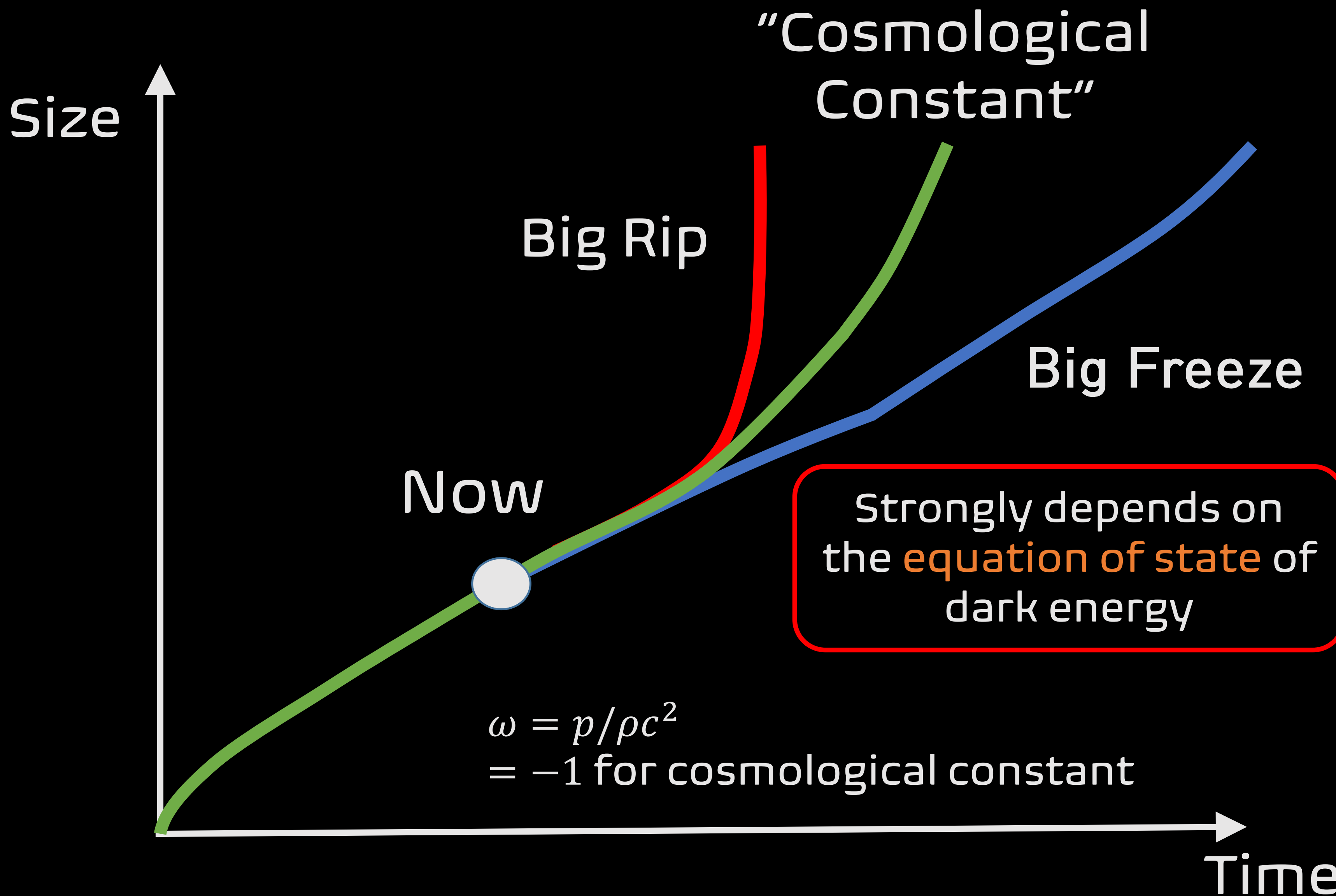


Chabanier+ (2019)

Energy Content of the Universe

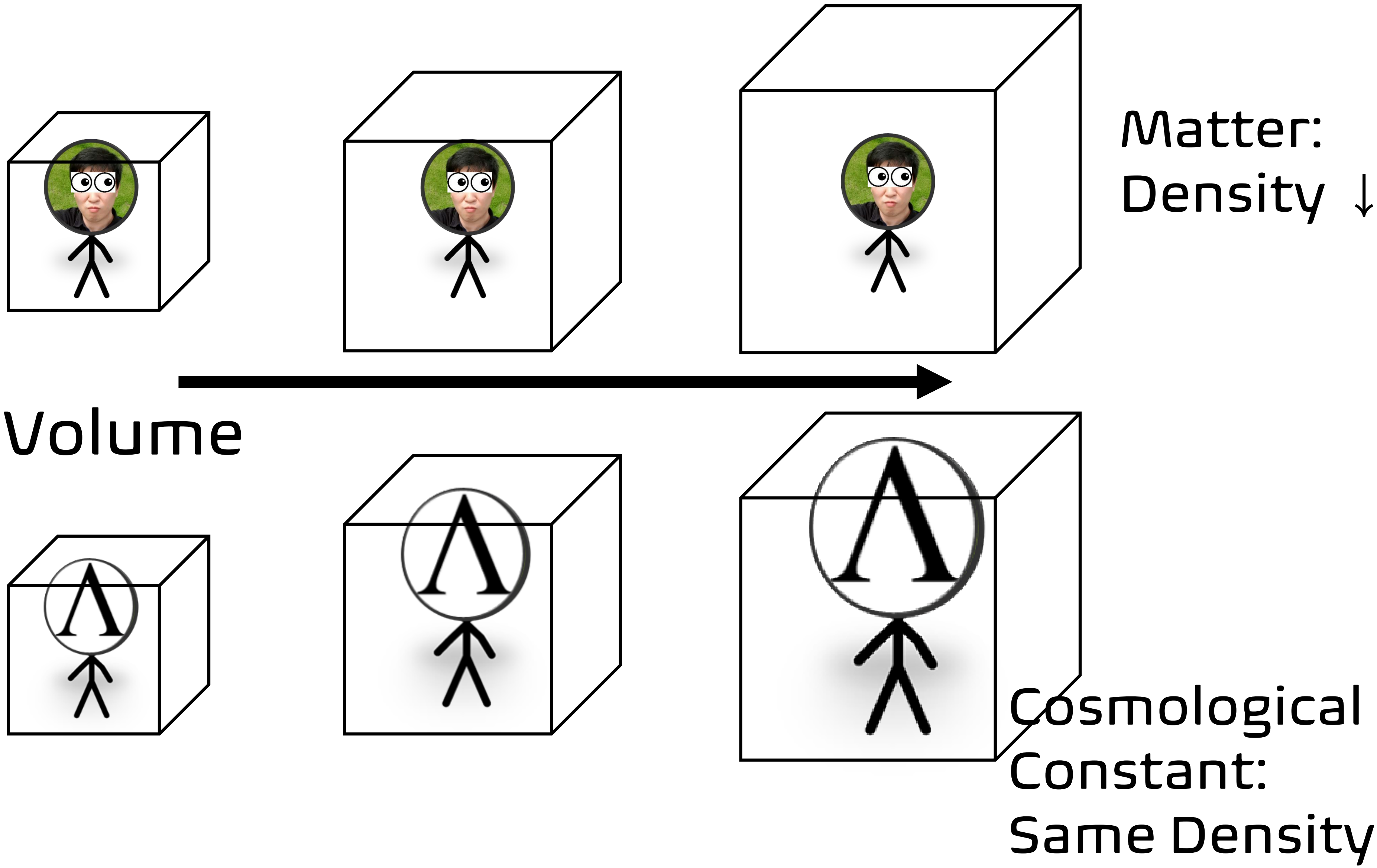


Fate of the Universe

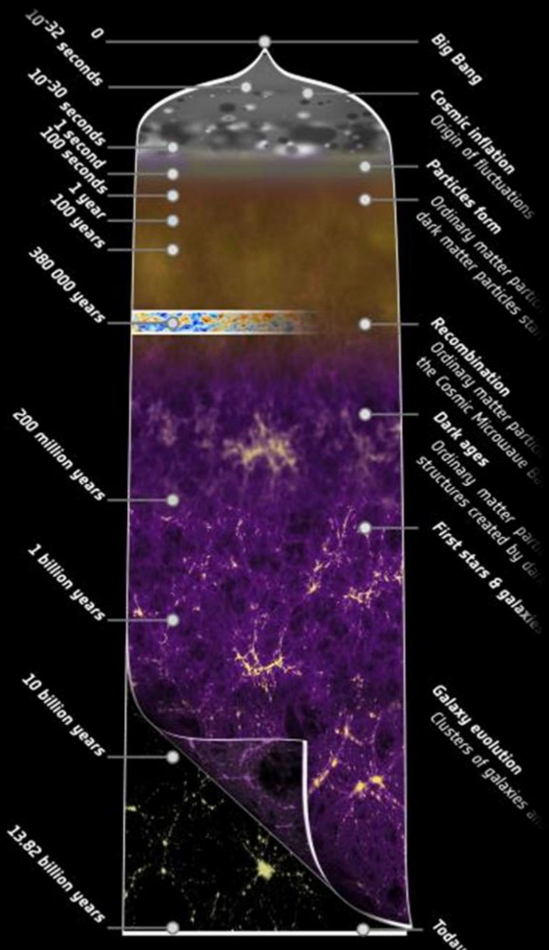


Strongly depends on the equation of state of dark energy

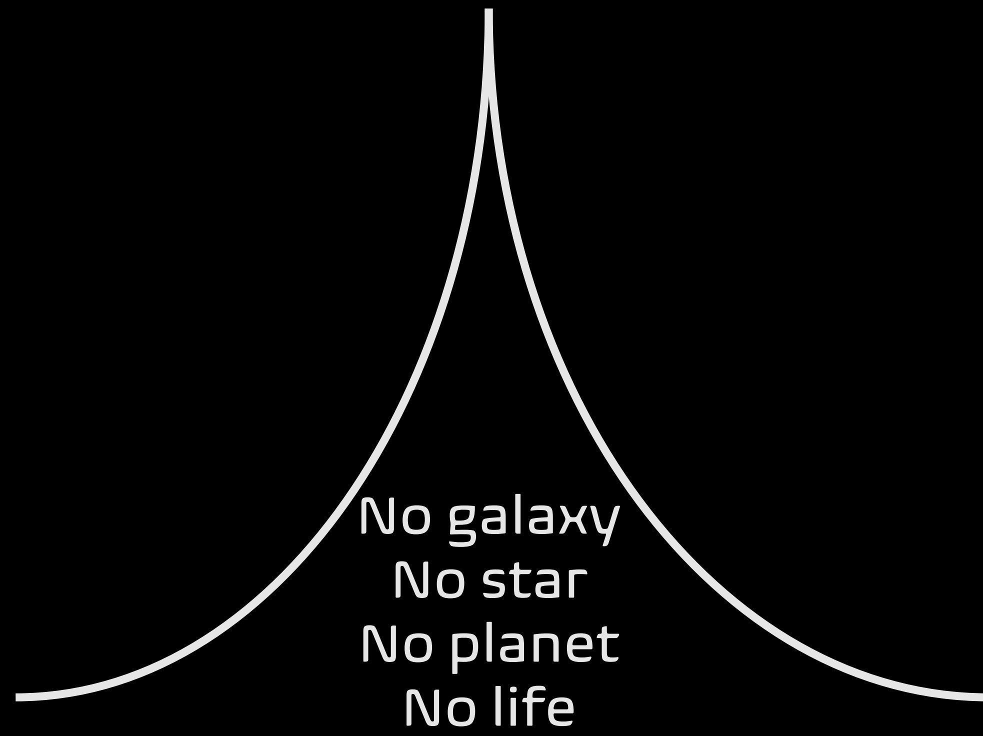
$\omega = p/\rho c^2$
$= -1$ for cosmological constant


Matter:
Density ↓
Volume
Λ
Cosmological
Constant:
Same Density

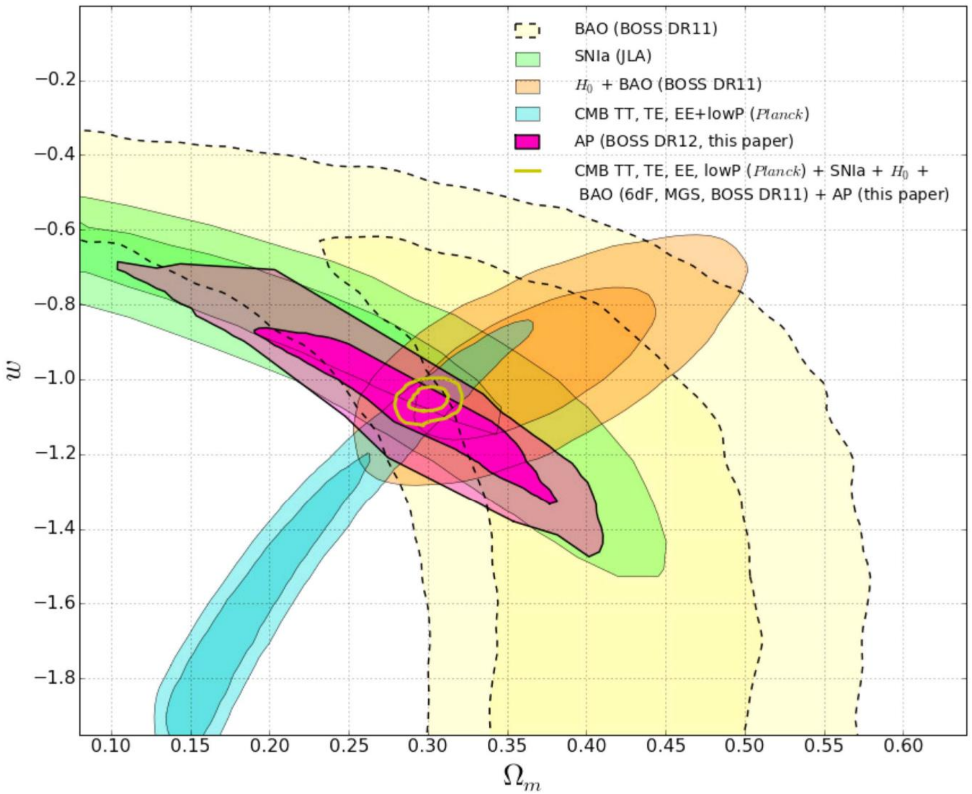
Observed cosmological constant $\sim 10^{-120}$ of theoretical expectation

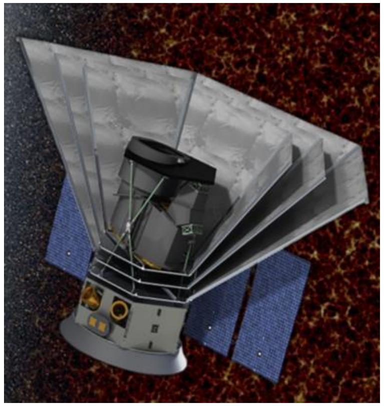


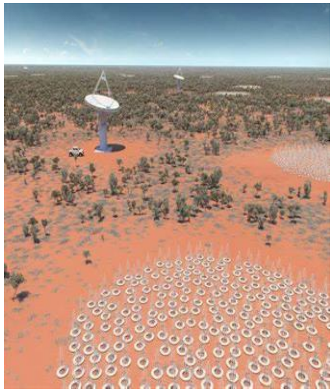
Cosmological constant
with observed value

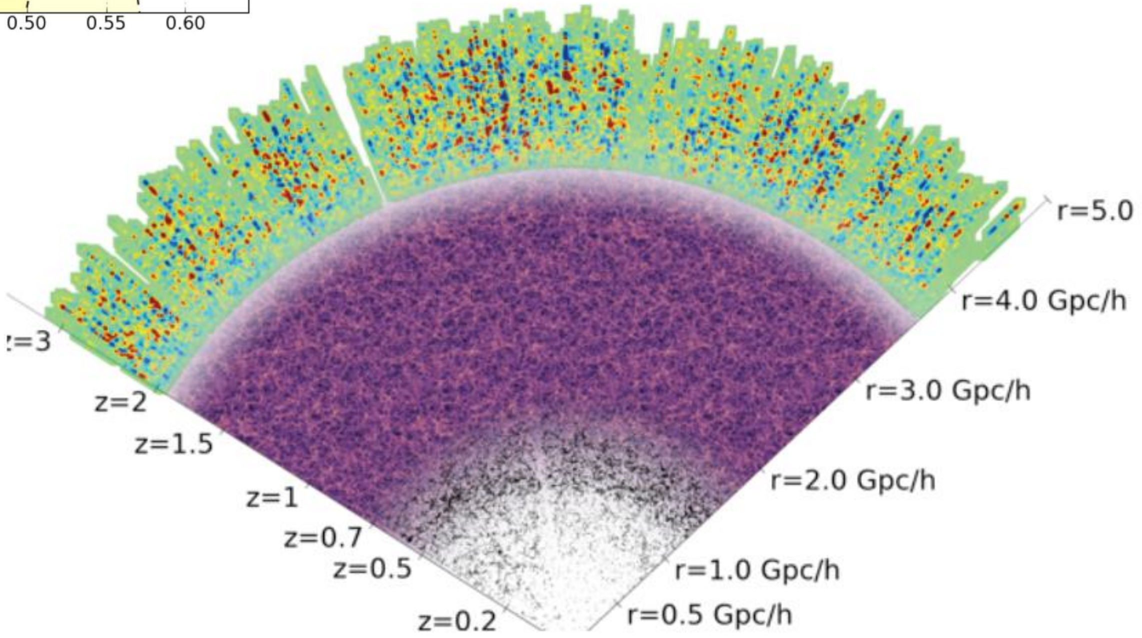


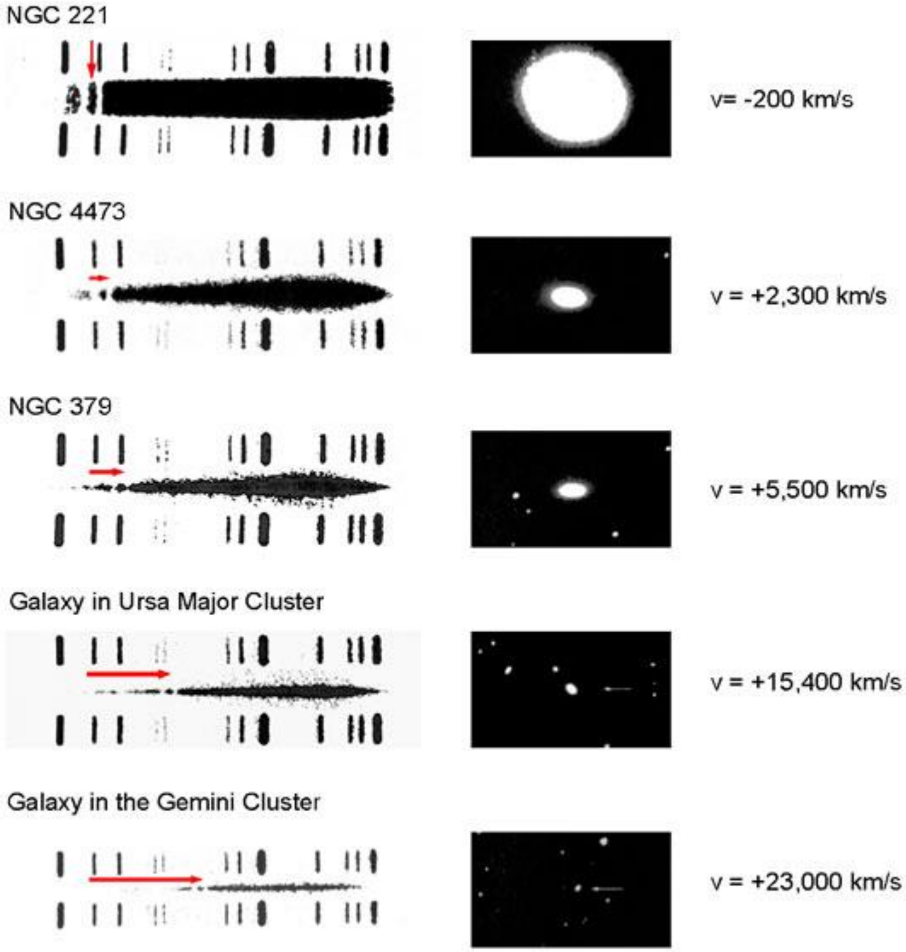
Cosmological constant
with theoretical expectation


BAO (BOSS DR11)
SNIa (JLA)
H_0 + BAO (BOSS DR11)
CMB TT, TE, EE+lowP (*Planck*)
AP (BOSS DR12, this paper)
CMB TT, TE, EE, lowP (*Planck*) + SNIa + H_0 +
BAO (6dF, MGS, BOSS DR11) + AP (this paper)
w
−0.2
−0.4
−0.6
−0.8
−1.0
−1.2
−1.4
−1.6
−1.8
0.10
0.15
0.20
0.25
0.30
0.35
0.40
0.45
0.50
0.55
0.60
Ω_m






r=5.0
r=4.0 Gpc/h
r=3.0 Gpc/h
r=2.0 Gpc/h
r=1.0 Gpc/h
r=0.5 Gpc/h
z=3
z=2
z=1.5
z=1
z=0.7
z=0.5
z=0.2



Hubble's Law?

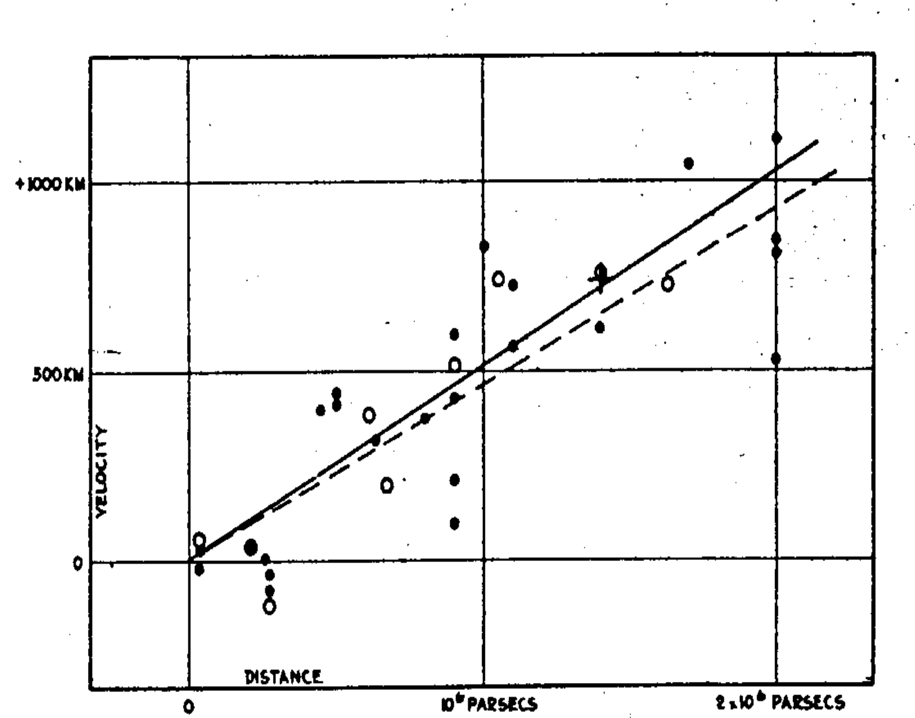
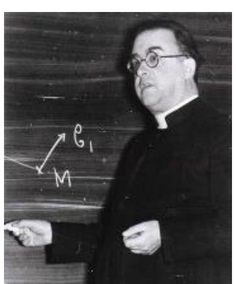


FIGURE 1



Lemaitre (1927)

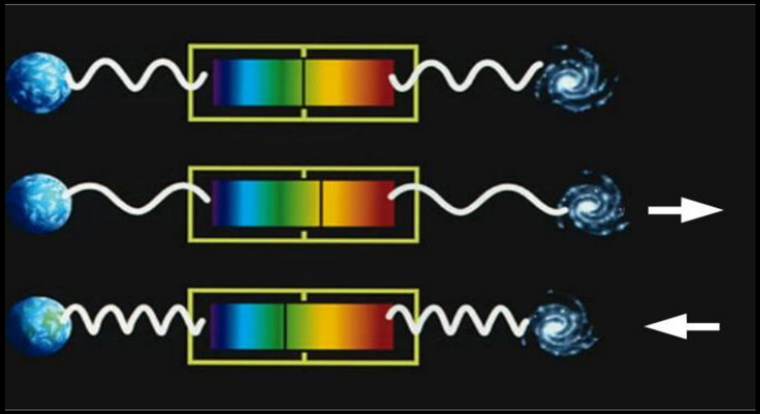
Hubble (1929)

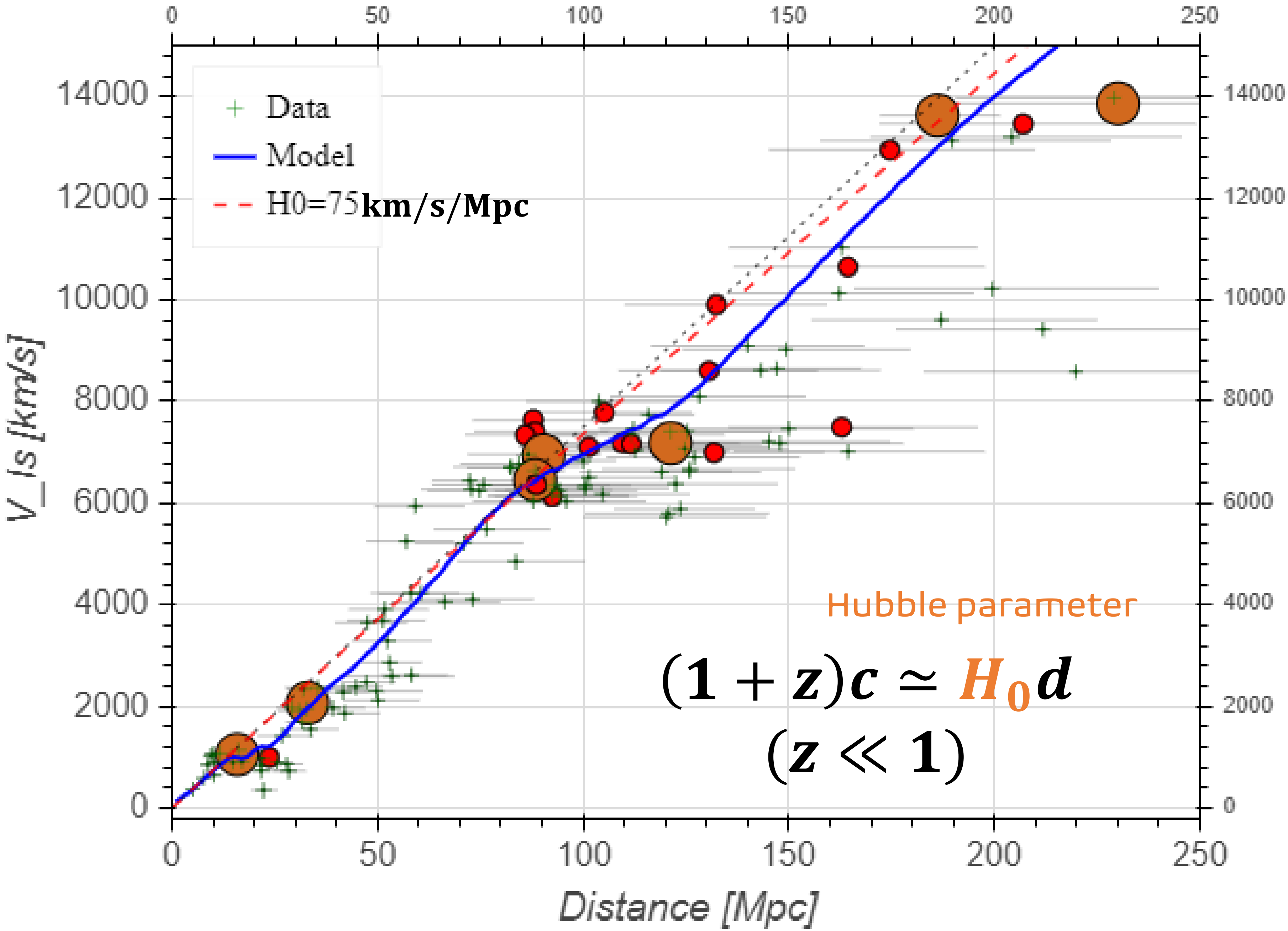
Redshift

- A phenomenon that the wavelength of electromagnetic wave is elongated.

- A kind of Doppler’s effect when an object radiating electromagnetic wave moves in an opposite direction to the wave propagation direction.

$$1 + z = \frac{\lambda_{\mathrm{obs}}}{\lambda_0} = \frac{v_{\mathbf{los}}}{c}$$





$$(1+z)c \simeq H_0 d$$
$$(z \ll 1)$$

Cosmicflows-3 (http://edd.ifa.hawaii.edu/)

Robertson-Walker spacetime metric

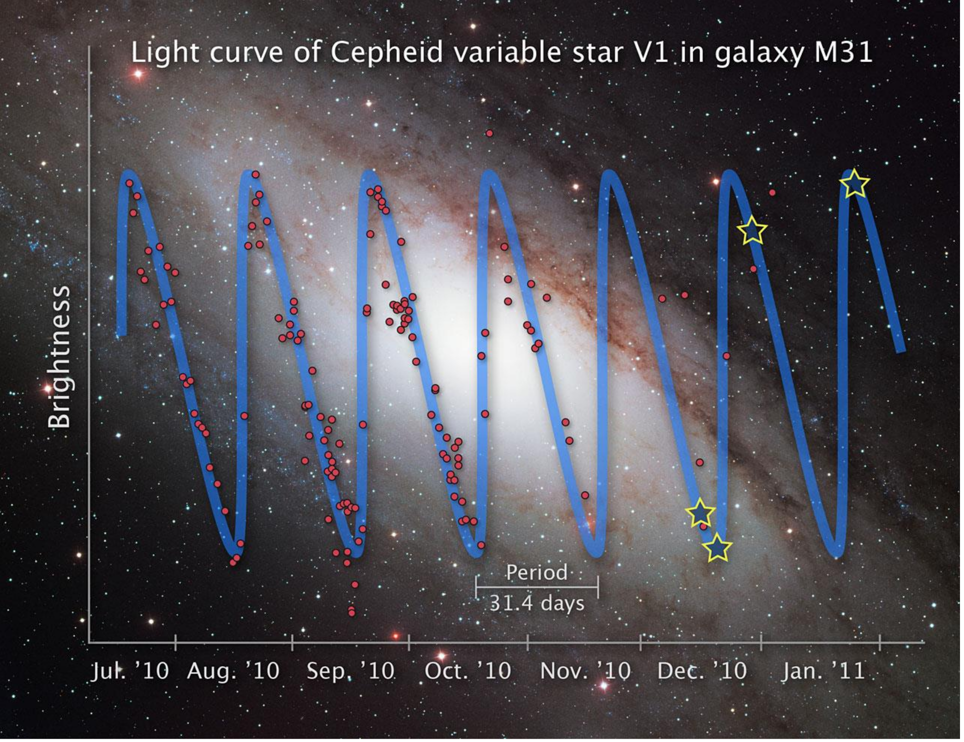
$$ds^2 = c^2 dt^2 - \underset{\text{scale factor}}{\boxed{a^2(t)}}\left[\frac{dr^2}{1 - \underset{\text{curvature}}{\boxed{K}}\, r^2} + r^2\left(d\theta^2 + \sin^2\theta\, d\phi^2\right)\right]$$

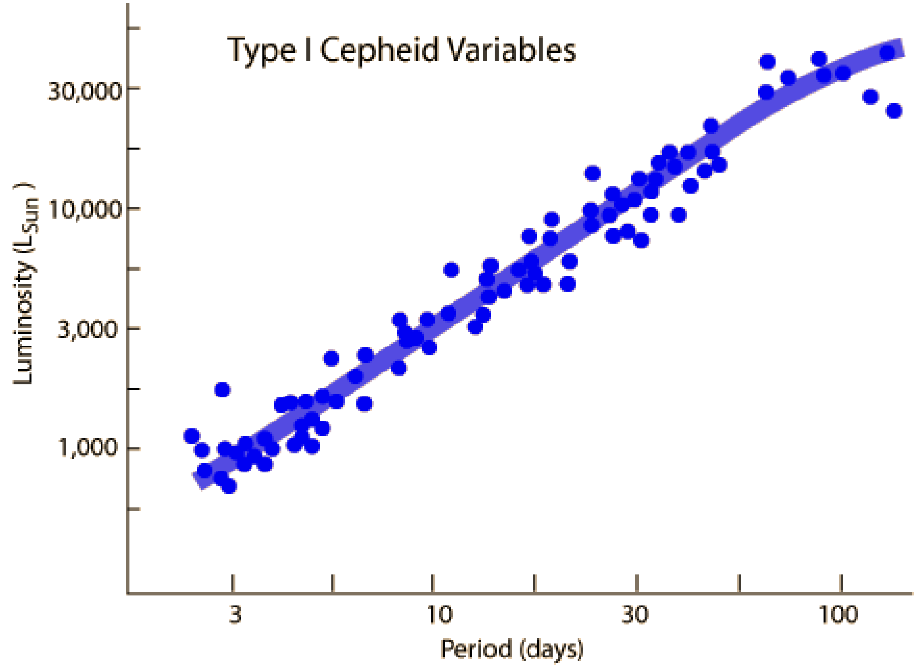
Friedmann equation

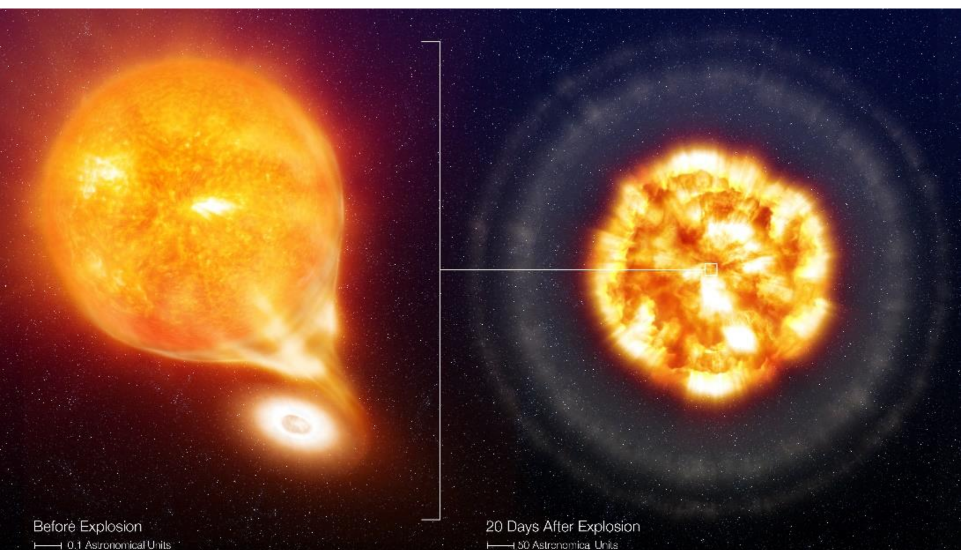
$$\dot{a}^2 + K^2 c^2 = \frac{8\pi G}{3}\rho a^2, \qquad \ddot{a} = -\frac{4\pi G}{c^2}\left(\rho + \frac{p}{c^2}\right) a$$

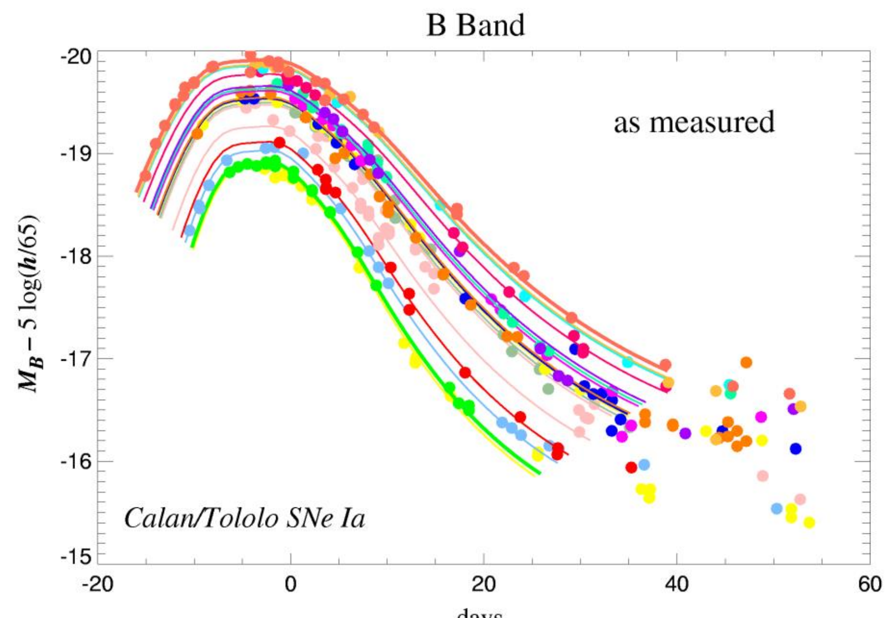
Critical density

$$\rho_c(t) = \frac{3\overset{\text{Hubble parameter}}{\boxed{H(t)^2}}}{8\pi G}$$


Light curve of Cepheid variable star V1 in galaxy M31
Brightness
Period
31.4 days
Jul. '10
Aug. '10
Sep. '10
Oct. '10
Nov. '10
Dec. '10
Jan. '11

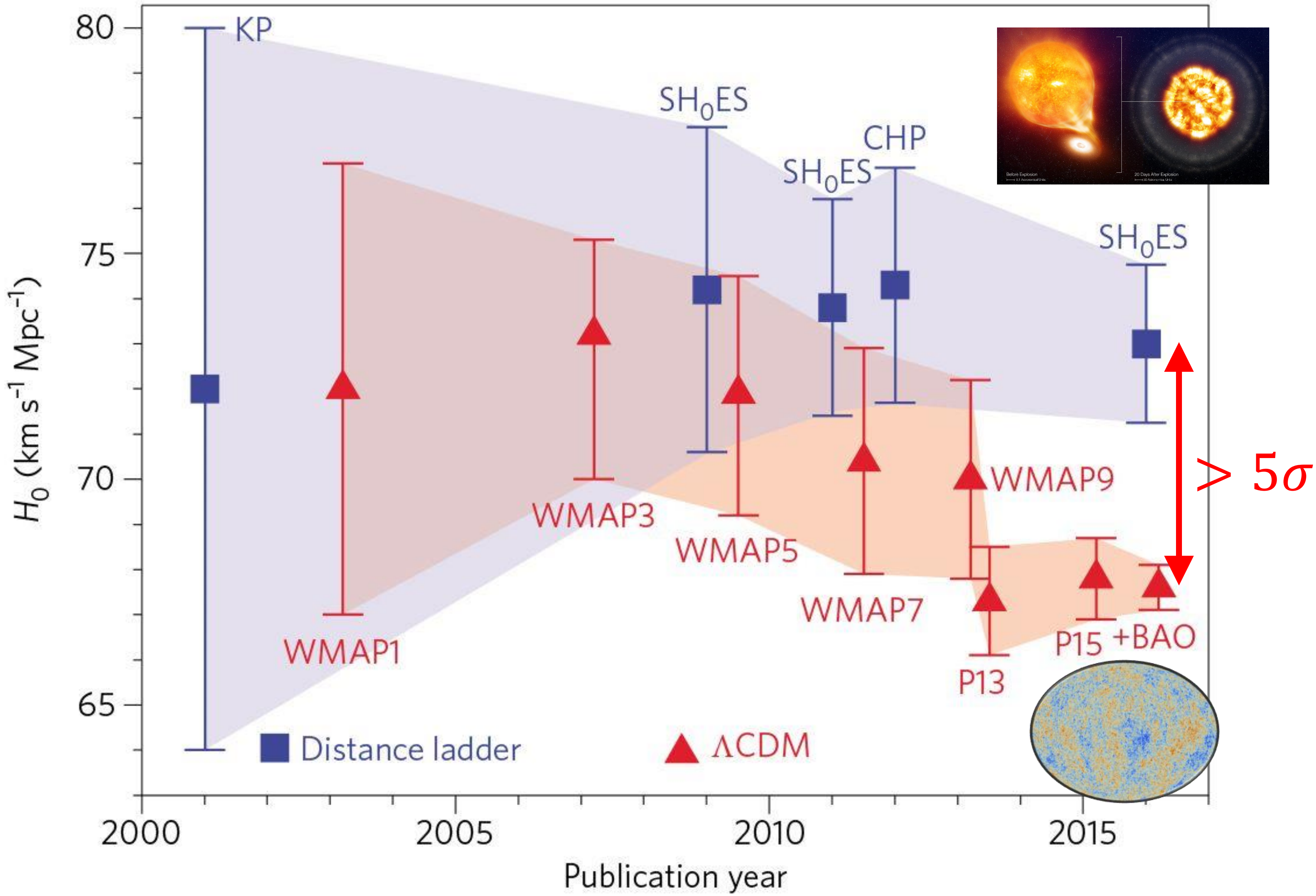

Type I Cepheid Variables
Luminosity (L_{Sun})
30,000
10,000
3,000
1,000
3
10
30
100
Period (days)


Before Explosion
0.1 Astronomical Units
20 Days After Explosion
50 Astronomical Units

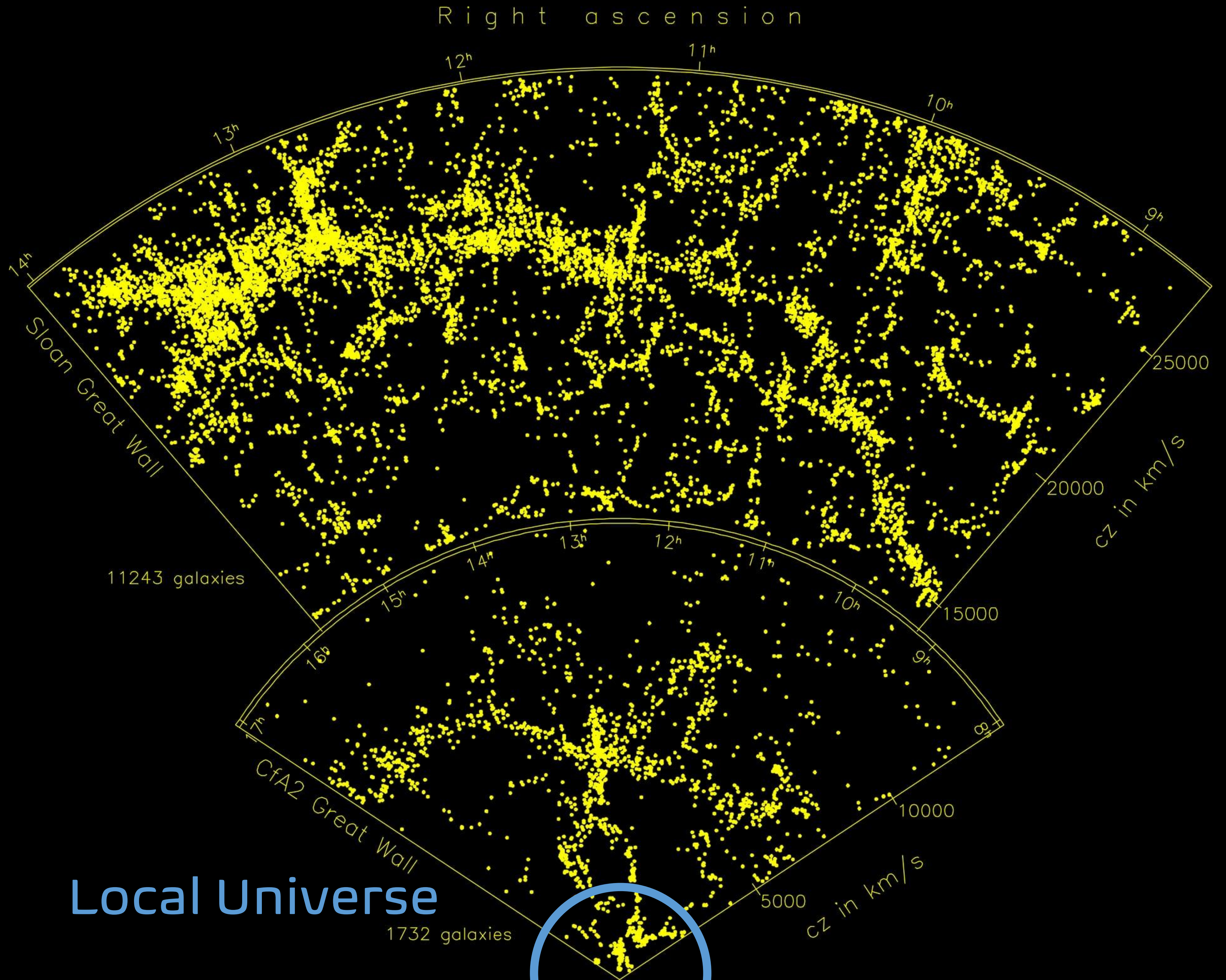

B Band
as measured
$M_B - 5\log(h/65)$
-20
-19
-18
-17
-16
-15
-20
0
20
40
60
days
Calan/Tololo SNe Ia

H_0-Tension?

https://www.nature.com/articles/s41550-017-0121



Local Universe



Why Local Universe is Important?

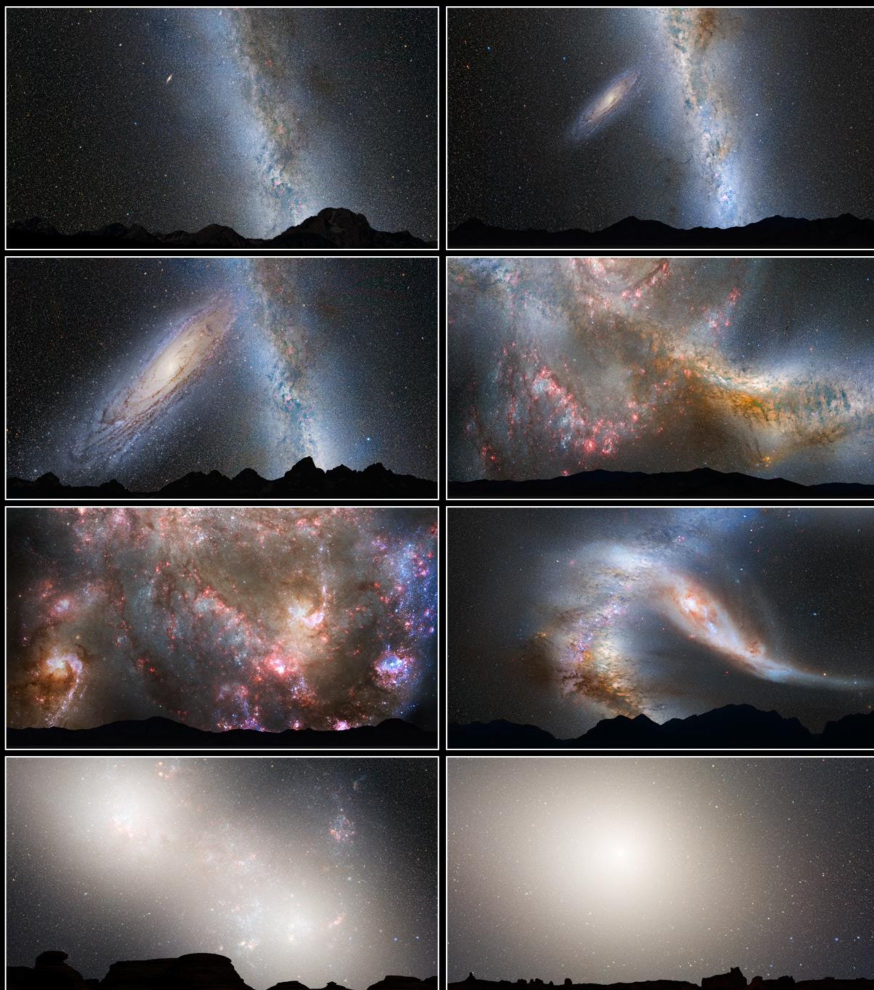
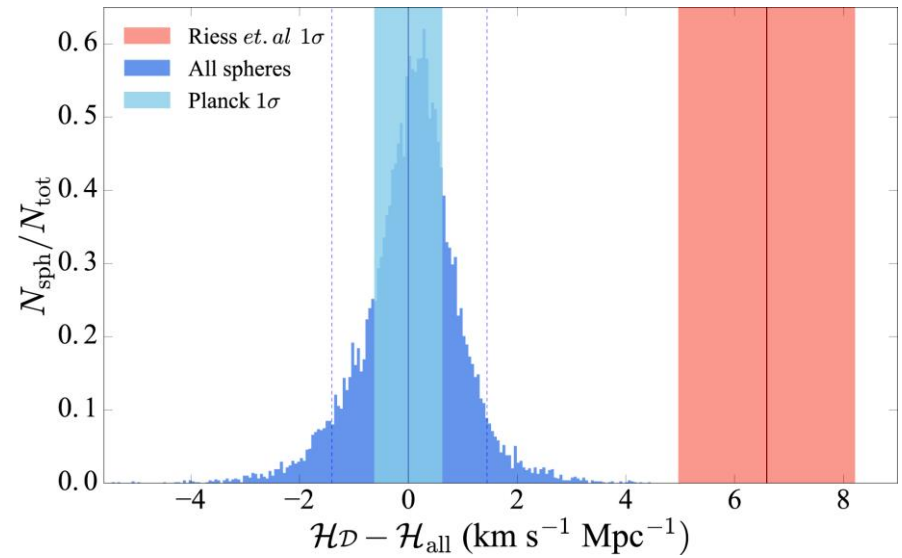


Illustration Sequence of the Milky Way and Andromeda Galaxy Colliding

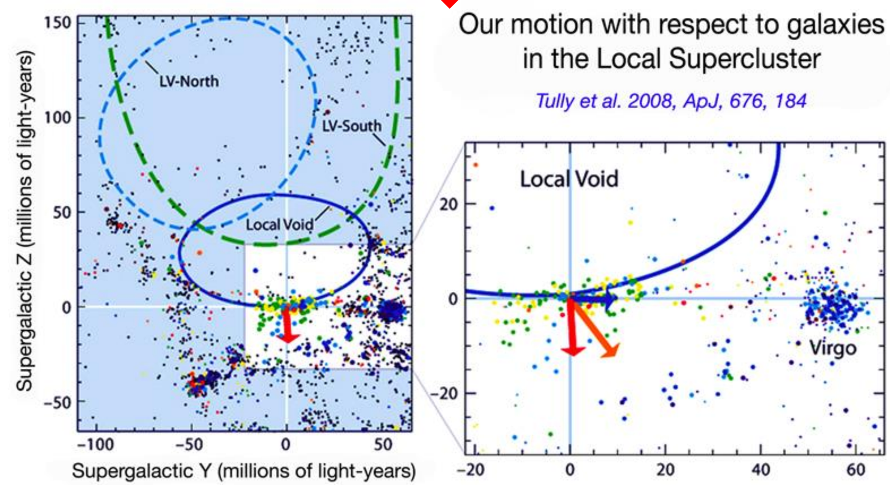
NASA, ESA, Z. Levay and R. van der Marel (STScI), T. Hallas, and A. Mellinger ▪ STScI-PRC12-20b



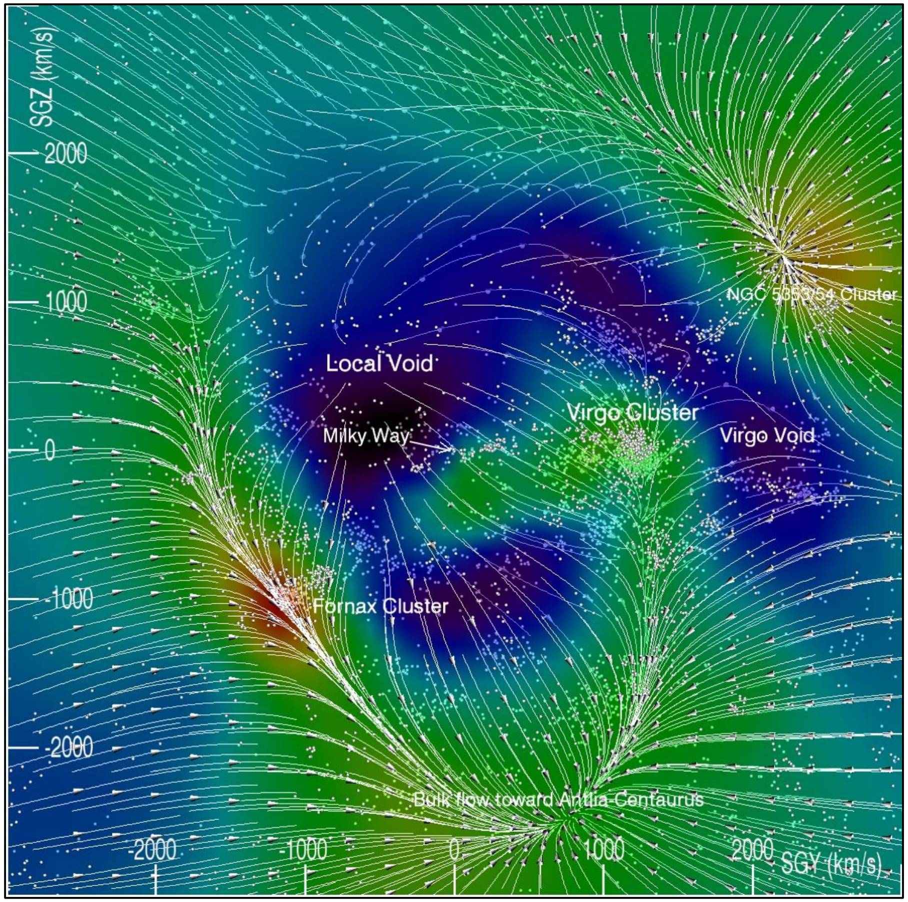
?

Our motion with respect to galaxies in the Local Supercluster

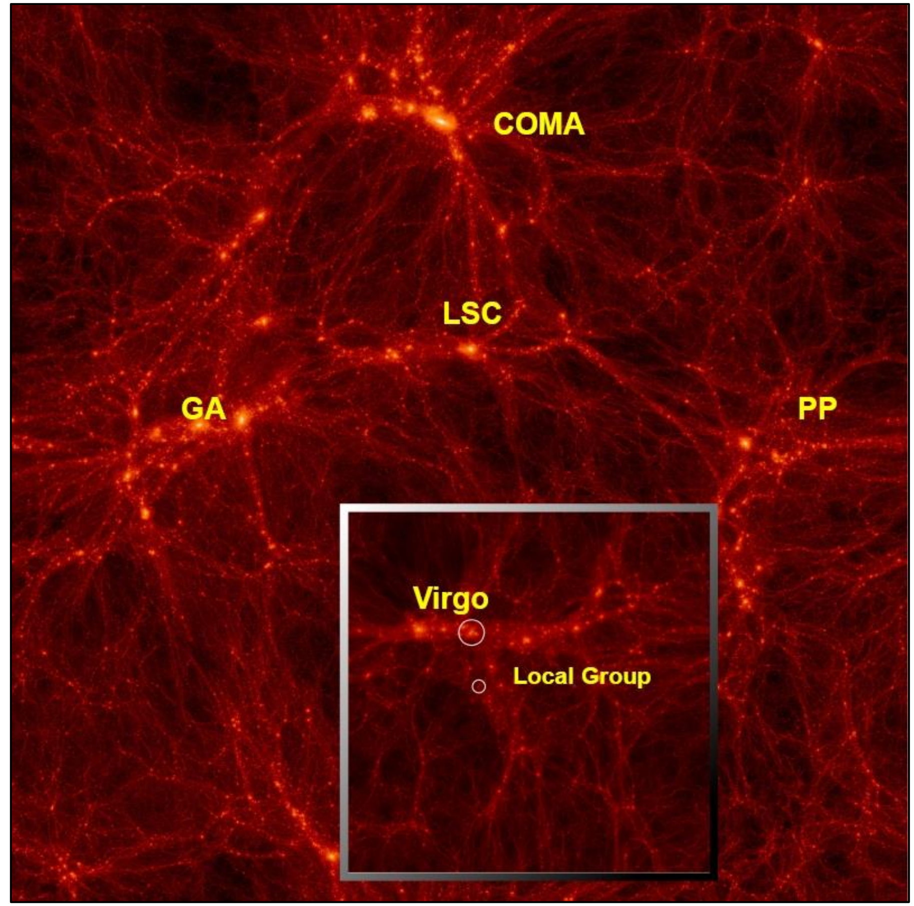
Tully et al. 2008, ApJ, 676, 184



Reconstructing Local Universe?

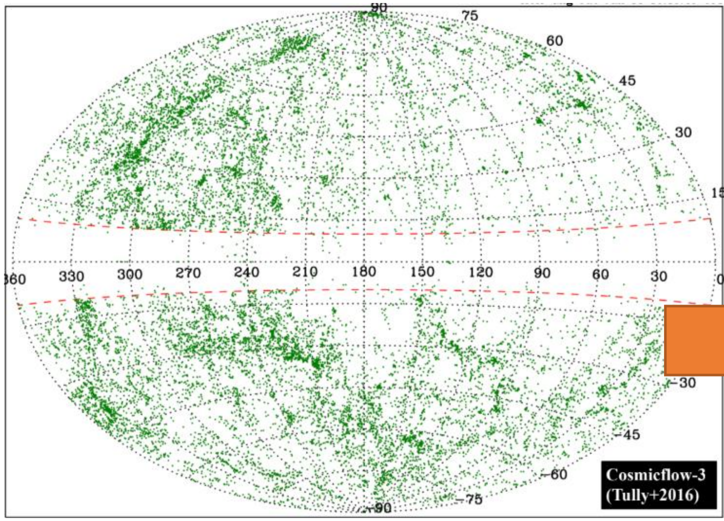


Cosmicflows-1: Courtois+ (2012)

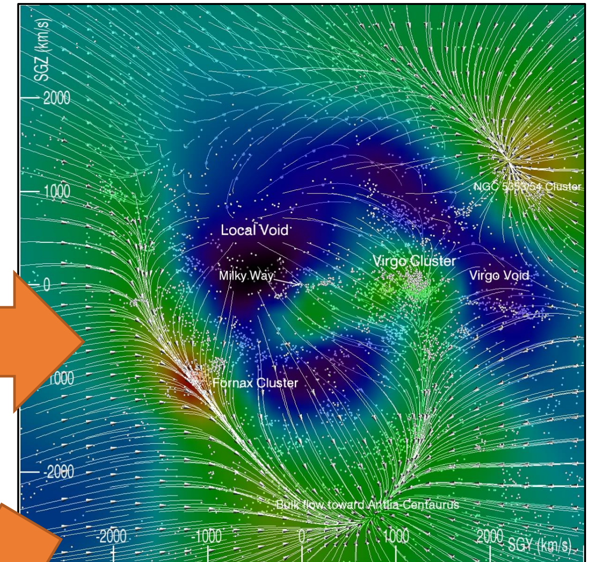


CLUES simulations
(160 Mpc/*h* & 64 Mpc/*h*)
1005.2687

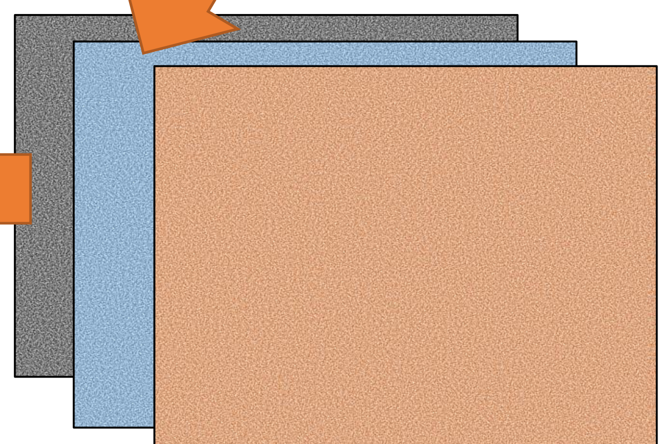
Reconstructing Local Universe?



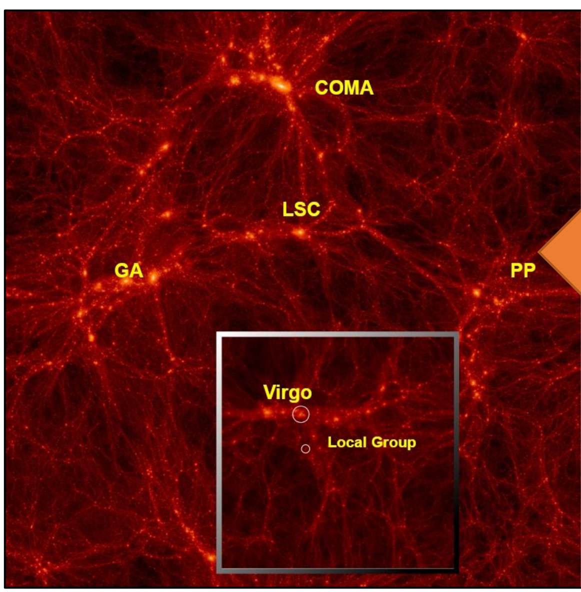
Full-sky observations
(Incomplete; low-latitude mask)



Reconstruct Gaussian field
(> 4Mpc/*h* scale)

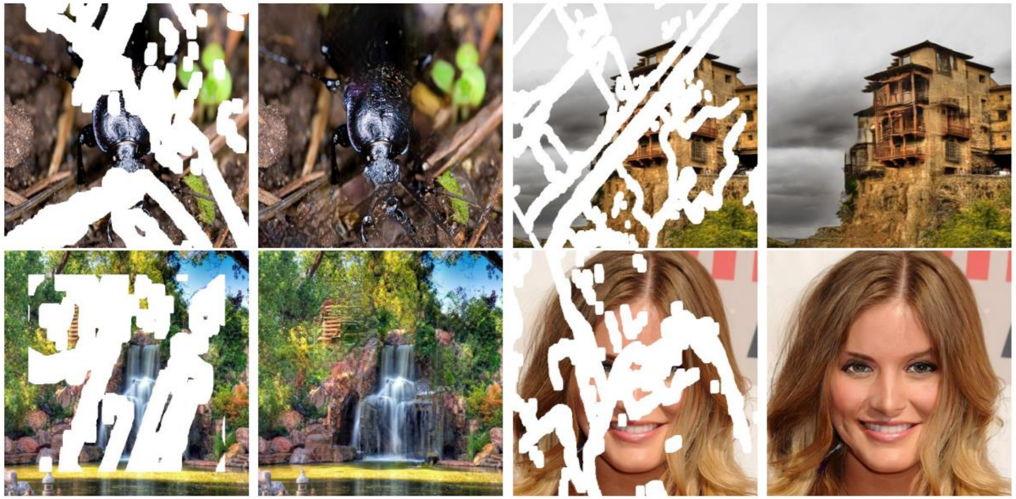


Add different realizations of
small-scales

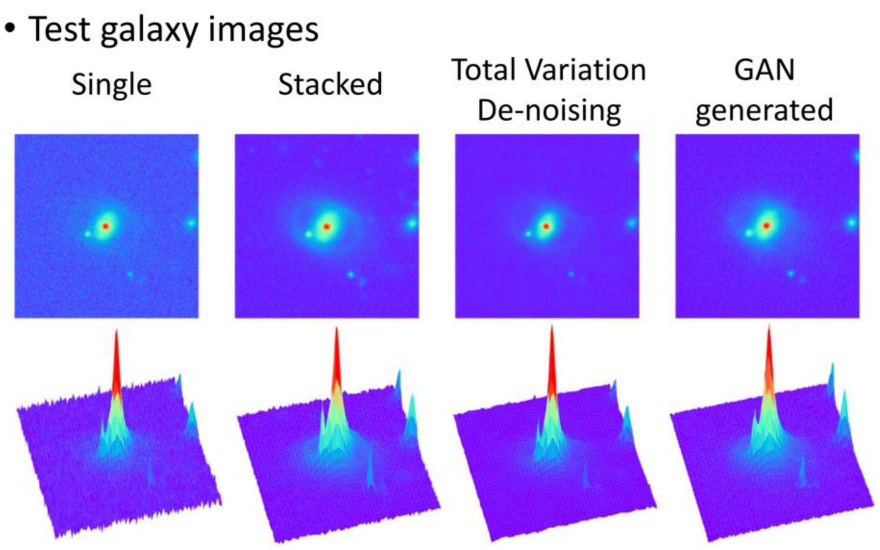


Run simulations
& find **the best?**

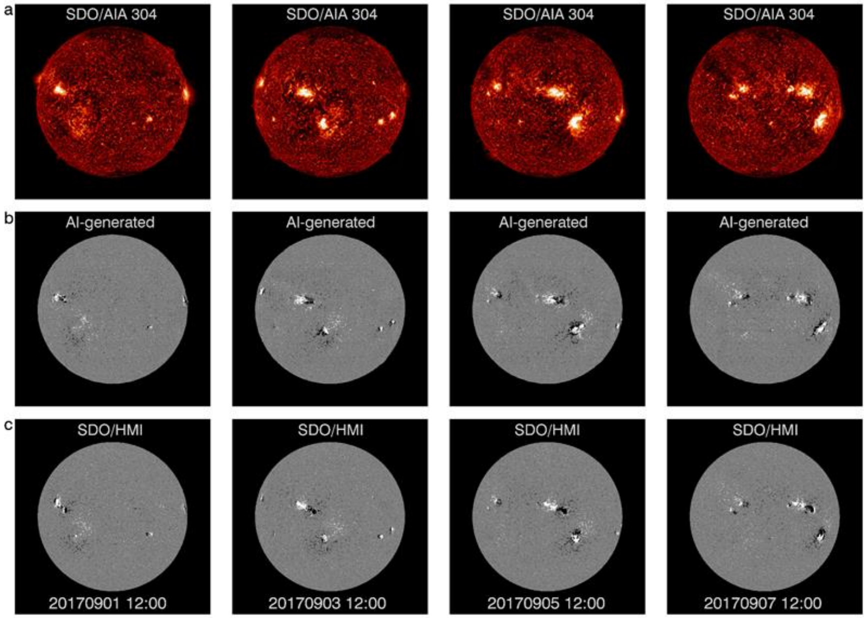
Deep Learning Image Reconstruction



© NVIDIA: 1804.07723

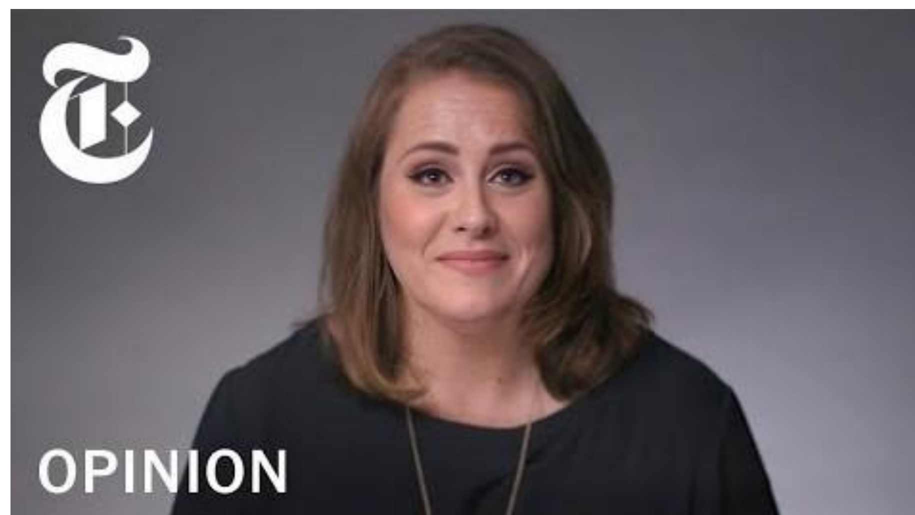


© Youngjun Park



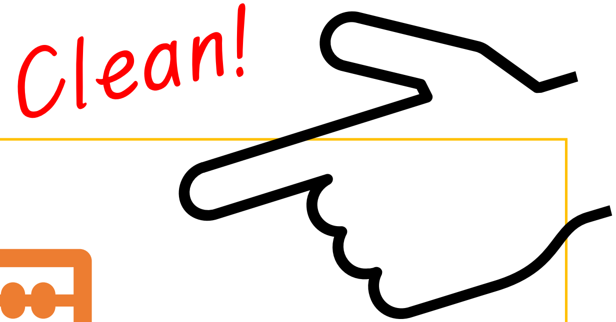
SolarMagGAN: Kim+ (2019)

What is Machine Learning?

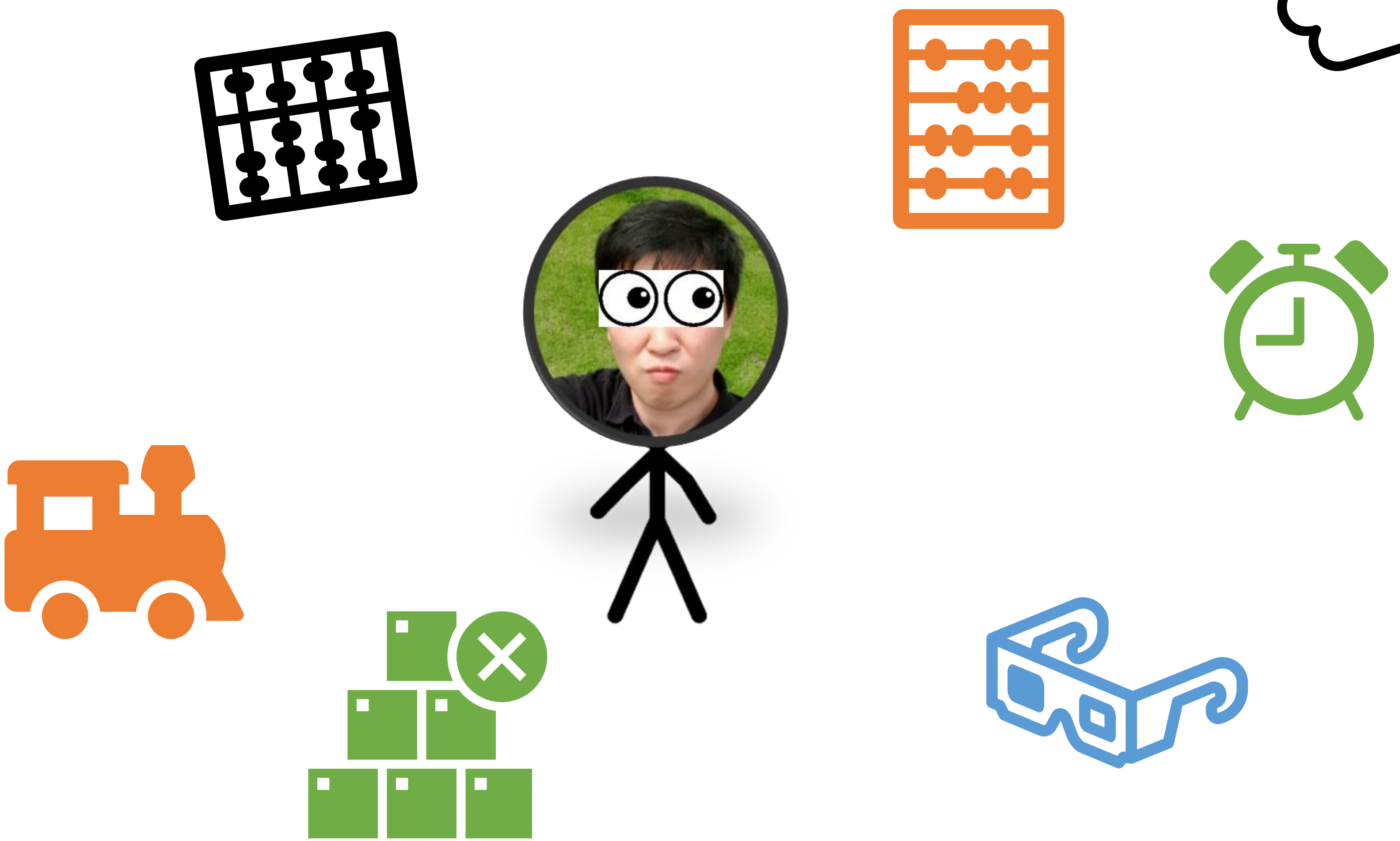


What is ~~Machine~~ Learning?

What is ~~Machine~~ Learning?



Unsupervised



What is ~~Machine~~ Learning?

Supervised

Find a function $F(\ldots)$

$$F(\overrightarrow{x_i}) \approx \overrightarrow{y_i}$$

Output we estimate

for given data $\{\overrightarrow{x_i}\}$ and $\{\overrightarrow{y_i}\}$

Input

Unsupervised

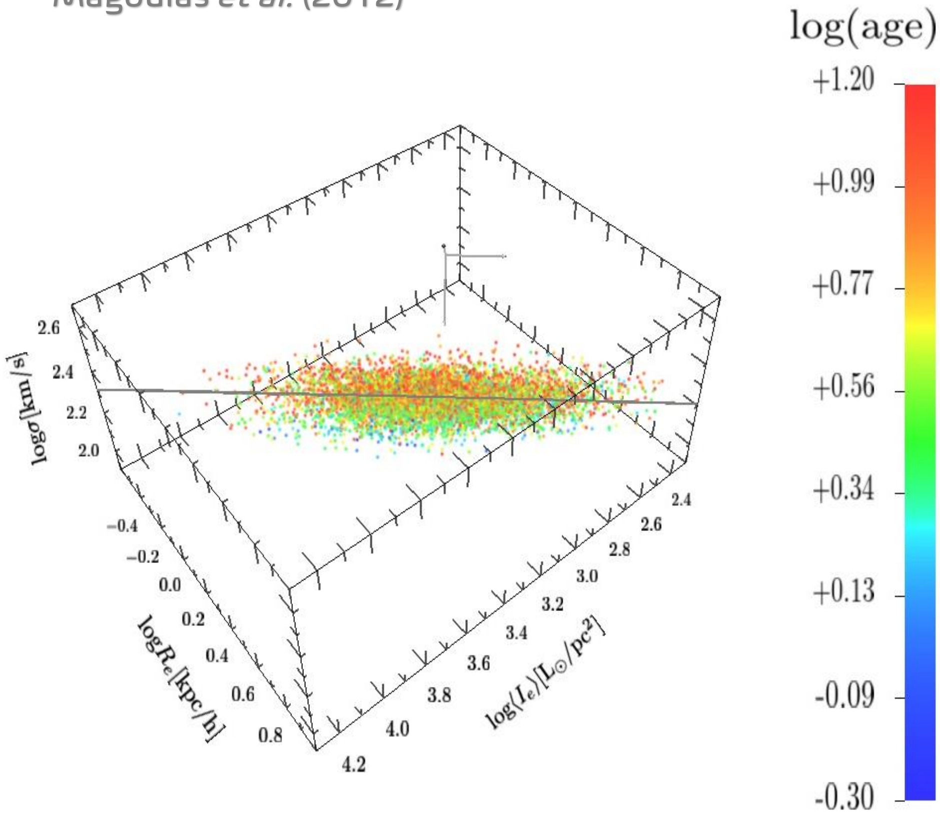
Find a function $F(\ldots)$

$$F(\overrightarrow{x_i}) = \overrightarrow{y_i}$$

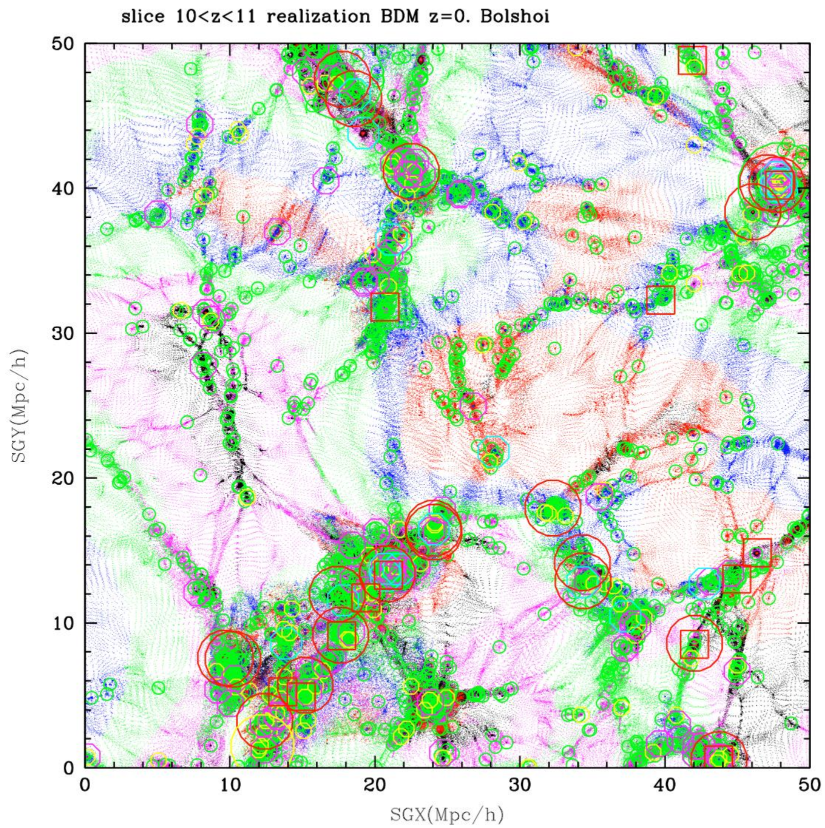
Output we can define

where $\{\overrightarrow{y_i}\}$ is something useful for data $\{\overrightarrow{x_i}\}$

Magoulas *et al.* (2012)



© Bolshoi Cosmological Simulation



Least-squared fitting w/ polynomial/exponential form

→ Supervised learning

$\{\vec{x_i}\}$: $\{(\mathrm{R_e}, \sigma)_i\}$
$\{\vec{y_i}\}$: $\{\langle I_e\rangle_i\}$

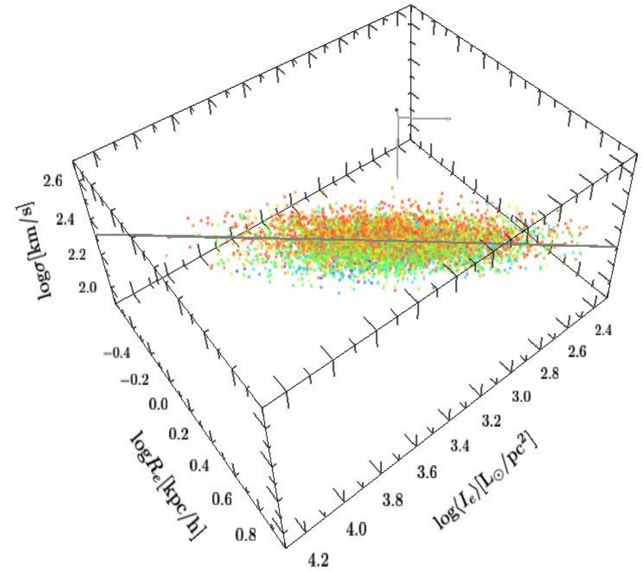
Friend-of-friends halo finder

→ Unsupervised learning

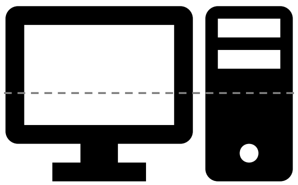
$\{\vec{x_i}\}$: particle position
$\{\vec{y_i}\}$: group label

Traditional Learning (or Fitting)

R_e	σ	$\langle I_e \rangle$
X.XXXXXX	X.XXXXXX	X.XXXXXX
X.XXXXXX	X.XXXXXX	X.XXXXXX
X.XXXXXX	X.XXXXXX	X.XXXXXX
X.XXXXXX	X.XXXXXX	X.XXXXXX
X.XXXXXX	X.XXXXXX	X.XXXXXX
…	…	…



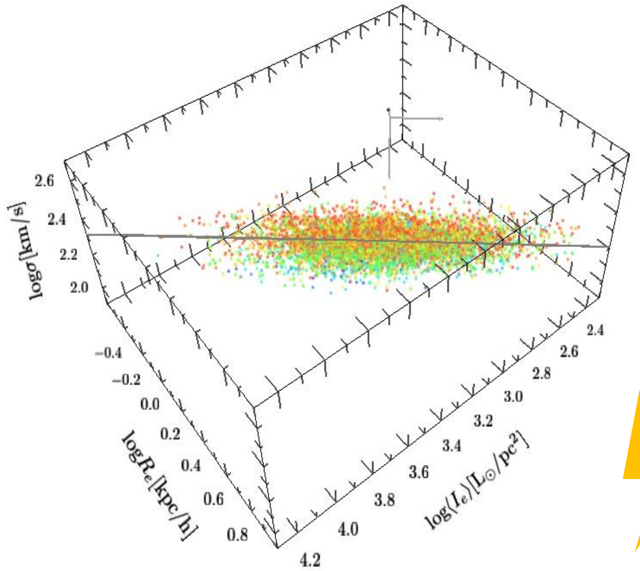
$$\log\langle I_e \rangle$$
$$= \beta_1 \log R_e + \beta_2 \log \sigma + \beta_3$$

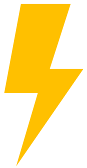


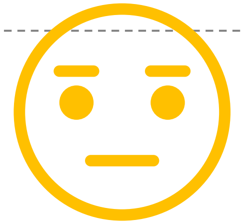
$$(\beta_1, \beta_2, \beta_3) = (\dots, \dots, \dots)$$

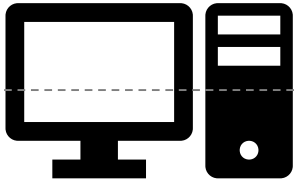
Machine Learning

R_e	σ	$\langle I_e \rangle$
X.XXXXXX	X.XXXXXX	X.XXXXXX
X.XXXXXX	X.XXXXXX	X.XXXXXX
X.XXXXXX	X.XXXXXX	X.XXXXXX
X.XXXXXX	X.XXXXXX	X.XXXXXX
X.XXXXXX	X.XXXXXX	X.XXXXXX
...	...	...









Learning Strategy

$$\langle I_e \rangle (R_e, \sigma) \simeq \cdots$$

Machine Learning: Pros

- ML can be used without specific description.

→ Good when we need to find something beyond the known relation.

- ML can find very complex $F(\ldots)$.

→ Good when the relation is hard to formulate.

- Once found, ML can usually calculate $F(\ldots)$ very fast.

→ Good when the true relation is computationally expensive.

Machine Learning: Cons

- $F(\dots)$ derived from ML is usually very complex.

→ Harder to get clear physical understanding.

- Controlling *learning strategy* is somewhat different from controlling *physical* parameters.

→ May take extra efforts to get used to the methods.

Machine Learning

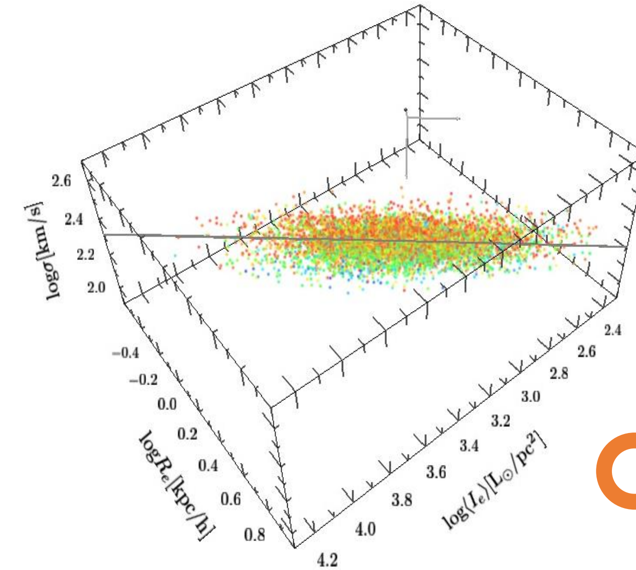
- All results should be lead by physics
  - Software is no replacement for physical understanding
- ML is a useful technique in three cases:
  - No known or understandable physical mechanism exists:
    - Dark energy
  - Physics known, but computing the required quantities is challenging:
    - Photometric redshifts
    - Large numbers of simulations - non-linear emulators
  - Paradigm testing: looking for things not predicted by any physics or models
    - Unknown unknown
    - Model-independent tests

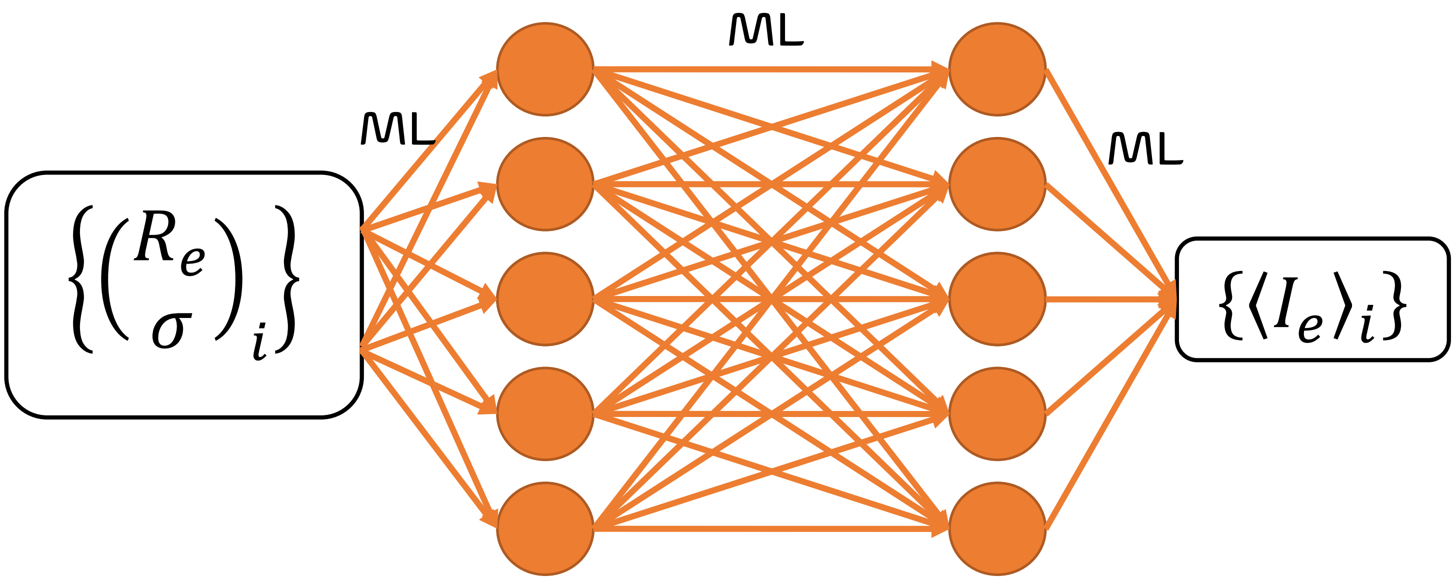
Techniques (examples):

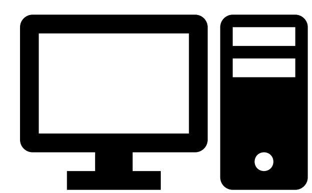
1. Classification/catagoriszation
   - Support Vector Machines
   - Naive Bayes
   - Perceptron/CNN
   - Decision Trees
2. Clustering
   - Hierarchical
   - k-means
   - DBSCAN
   - Density
3. Interpolation/Regression
   - Gaussian Process Regression
   - Kernel Ridge Regression
   - Multivariate adaptive regression splines

Deep Learning

R_e	σ	$\langle I_e \rangle$
X.XXXXXX	X.XXXXXX	X.XXXXXX
X.XXXXXX	X.XXXXXX	X.XXXXXX
X.XXXXXX	X.XXXXXX	X.XXXXXX
X.XXXXXX	X.XXXXXX	X.XXXXXX
X.XXXXXX	X.XXXXXX	X.XXXXXX
…	…	…

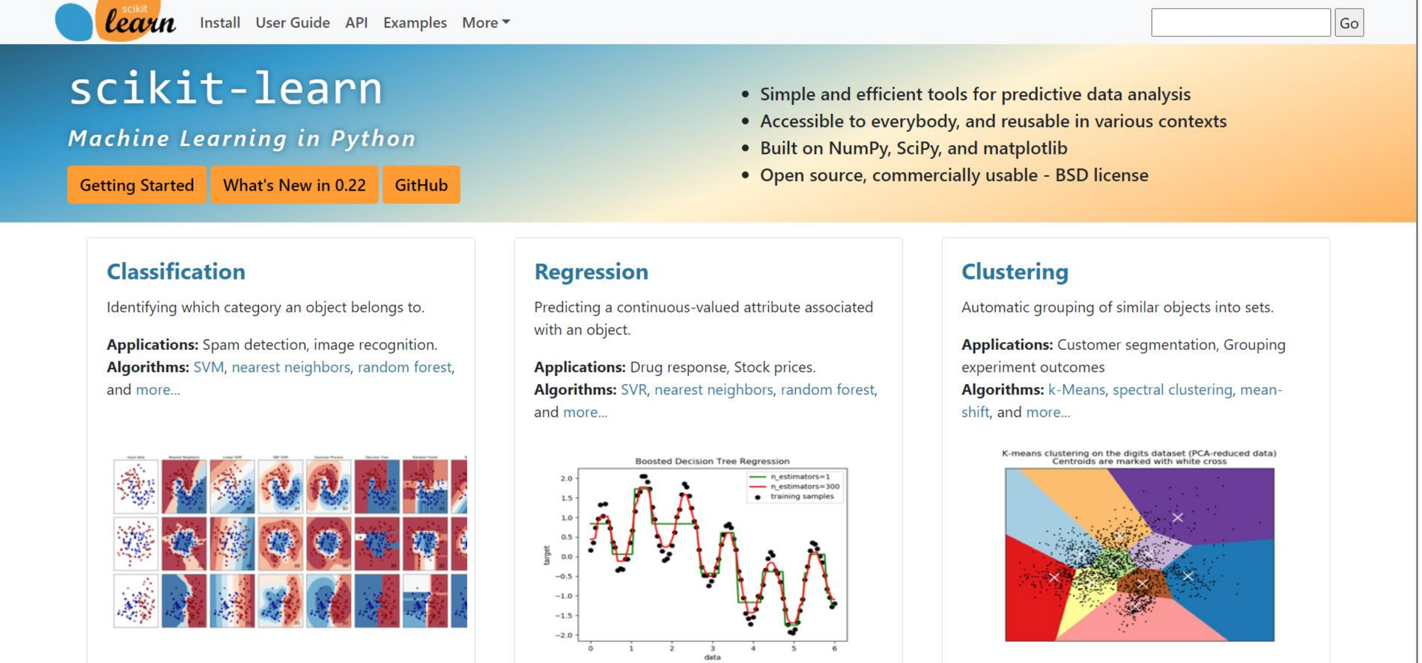






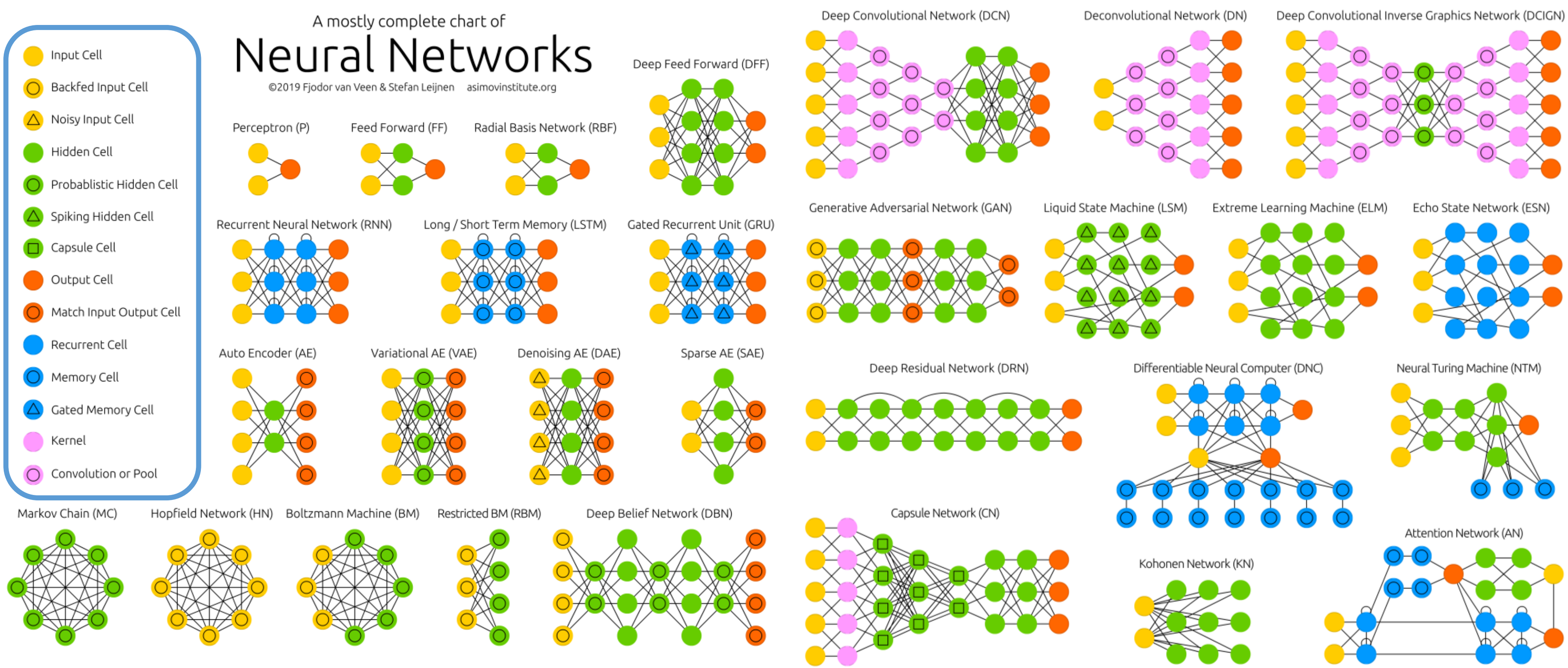
Deep Learning: Pros & Cons

- The physical understanding of results becomes even harder than usual ML
- Building DL architectures allow too many degrees of freedom
- Successful DL requires many training samples & computational cost (usually on GPUs)
- If successful, DL can provide very high estimating power in complex problems



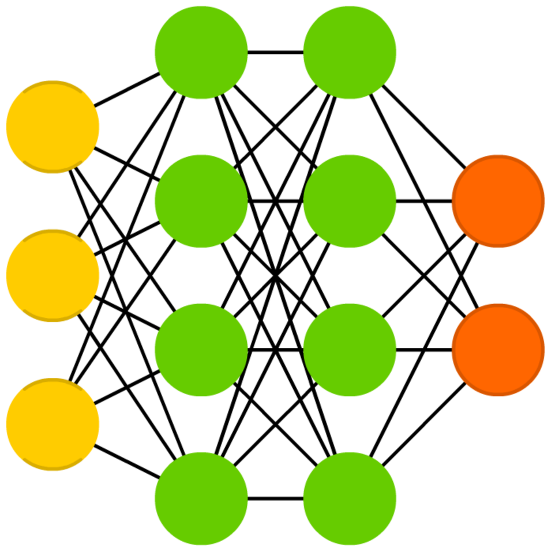
ML: scikit-learn (http://scikit-learn.org/)

DL: tensorflow (http://www.tensorflow.org/)

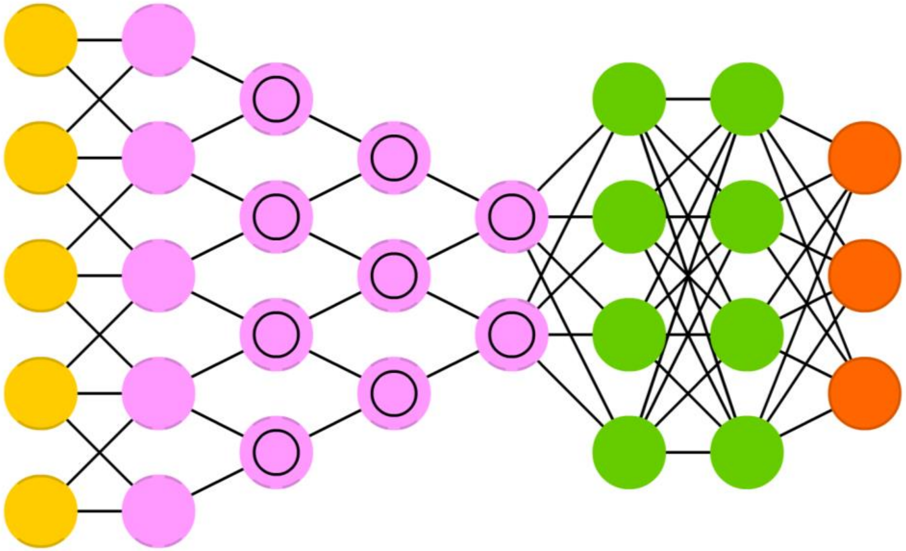


https://www.asimovinstitute.org/neural-network-zoo/

Deep Feed Forward (DFF)



Deep Convolutional Network (DCN)



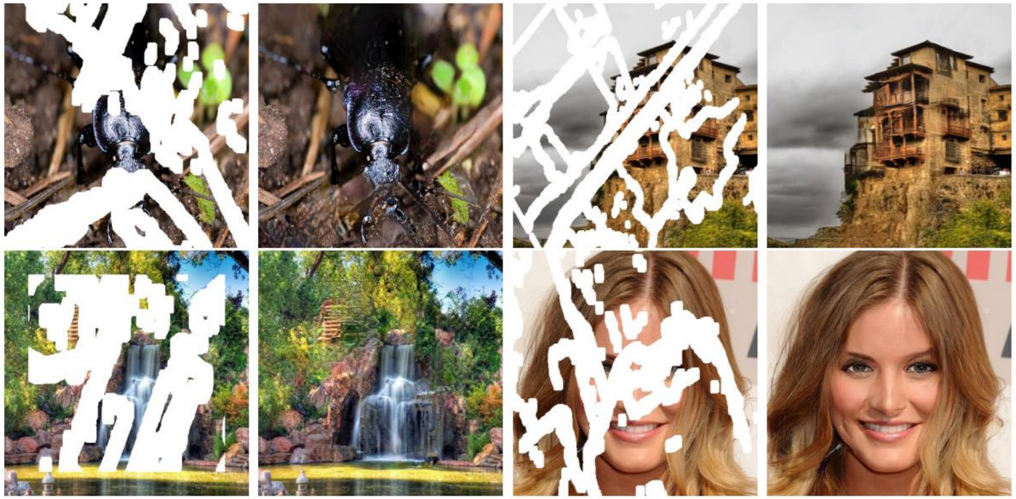
Fully connected (FC) layer

- Produce connections for all elements between layers
- Useful if each element is rather independent (e.g., different physical parameters)

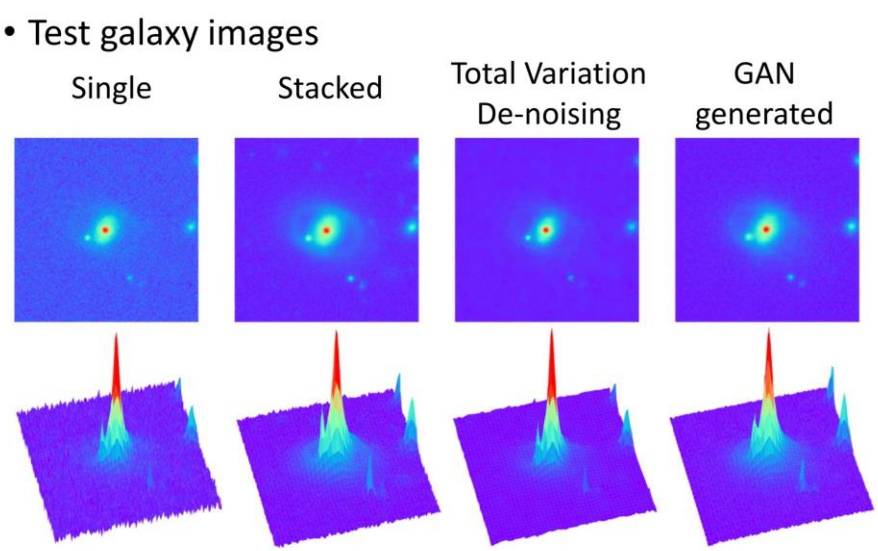
Convolutional (Conv) layer

- Apply convolution kernels with fixed sizes
- Useful if ordering of elements is important (e.g., spectrum, image)

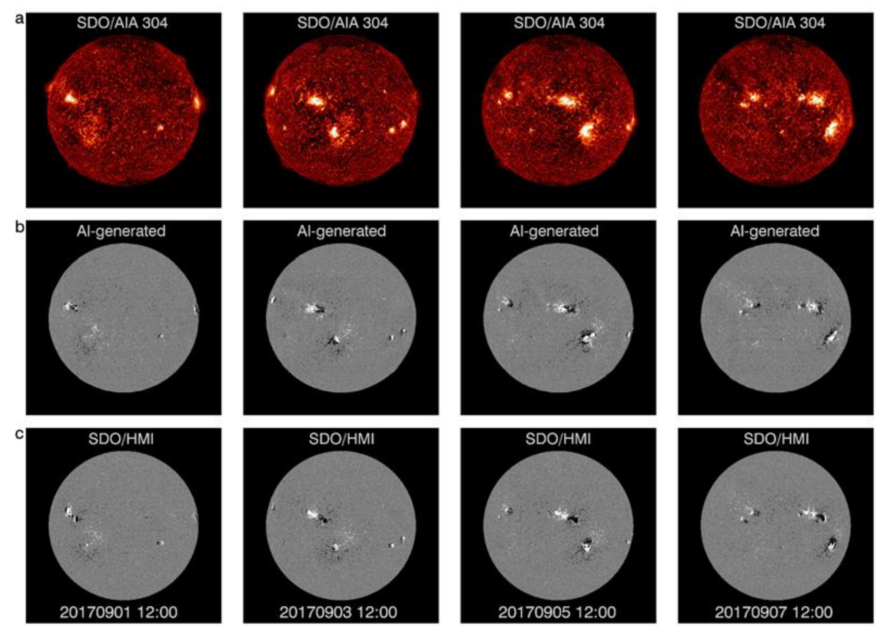
Deep Learning Image Reconstruction



© NVIDIA: arXiv:1804.07723

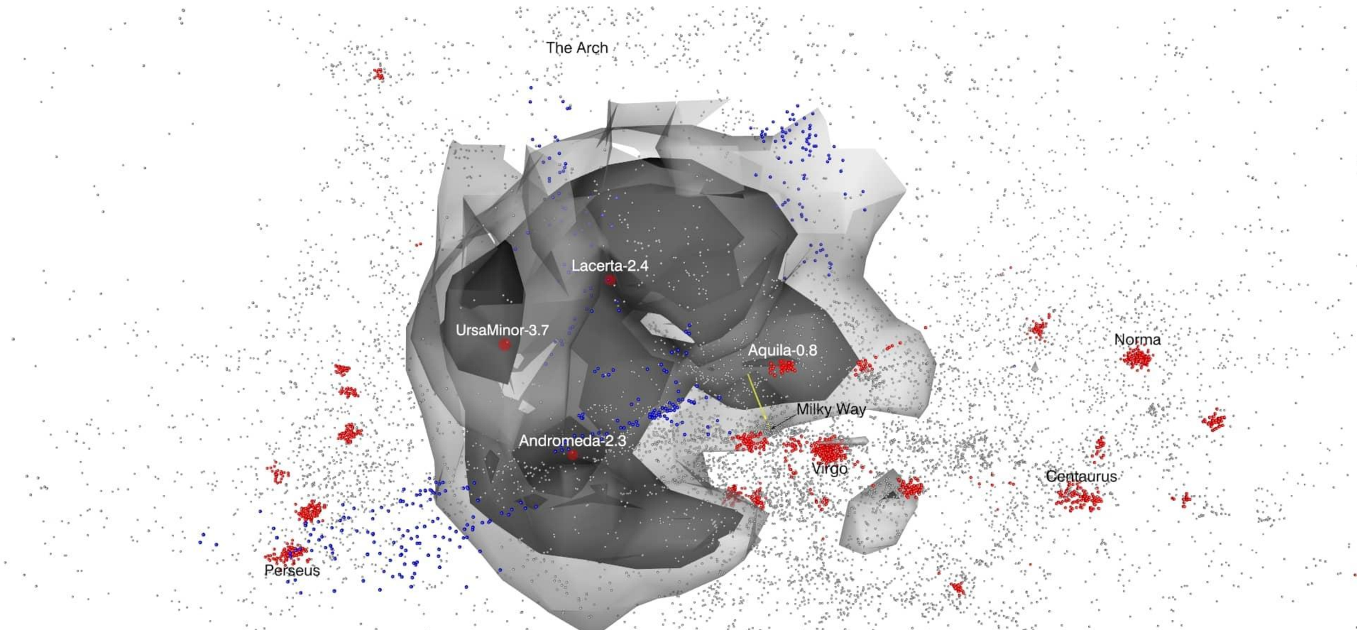


© Youngjun Park



SolarMagGAN: Kim+ (2019)

Observational Data: Cosmicflows-3



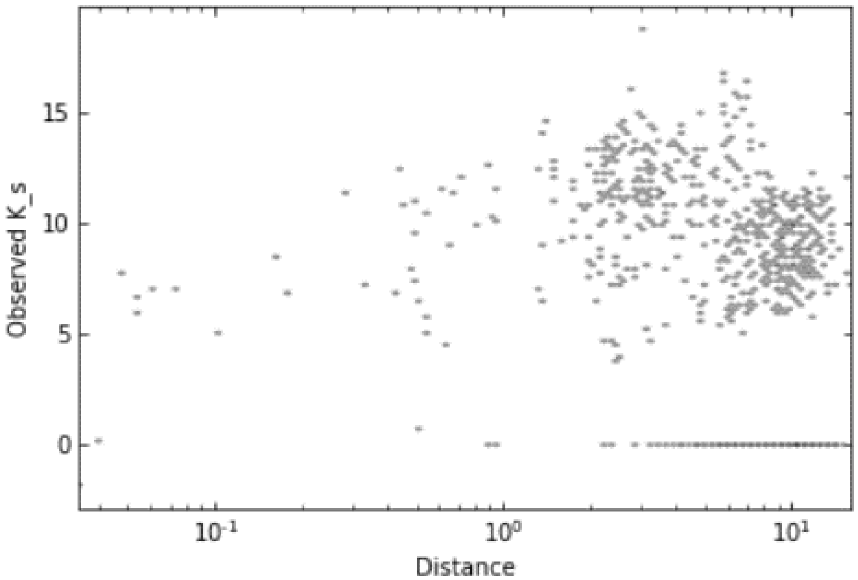
Cosmicflows-3: COSMOGRAPHY OF THE LOCAL VOID

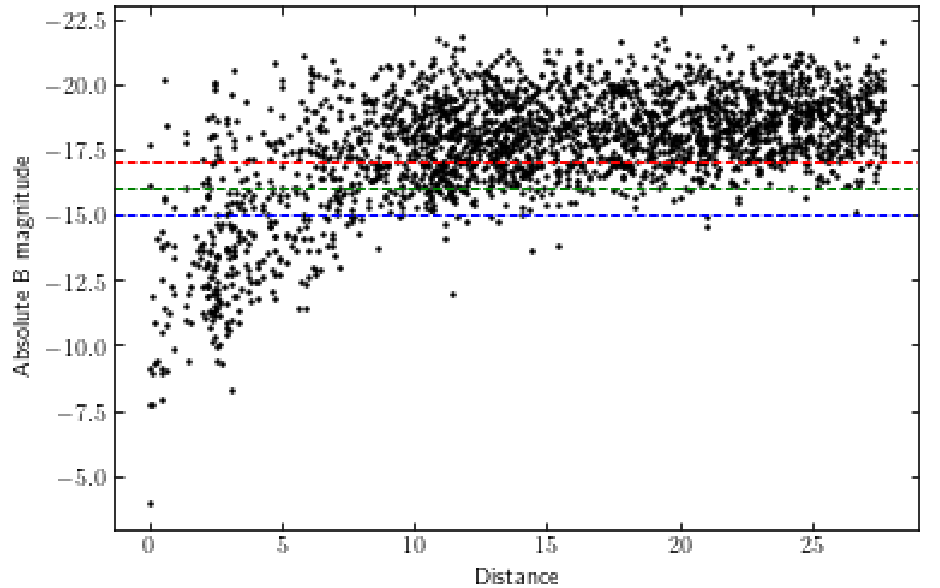
by

R. Brent Tully, Daniel Pomarède, Romain Graziani, Yehuda Hoffman, Hélène M. Courtois and Edward J. Shaya

$H_0 = 100h$ km/s/Mpc

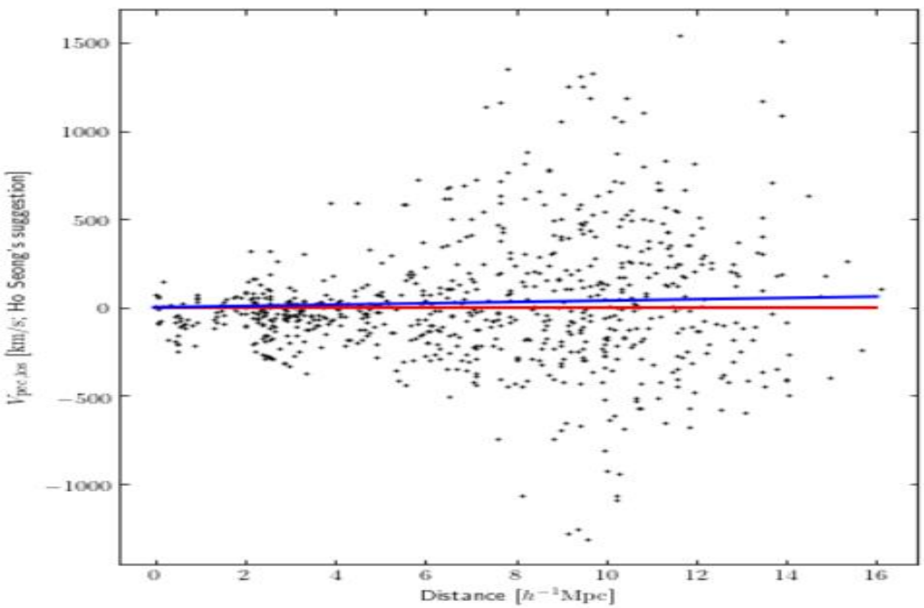
- Provides both distance and LOS velocity
- Use 40Mpc/h cube volume, with 0.3125Mpc/h resolution
- Exclude region close to the Galactic plane ($|b| < 10$ deg)
- Apply absolute B-band magnitude cut ($M_B < -16$ mag)
- Peculiar velocity = LOS velocity (Galactic standard of rest) – Hubble flow





~30% of galaxies do not contain proper K_s-band magnitude. Possibly due to the limitation of 2MASS+ observation.

M_B < -16 mag is sufficient for making volume-limited sample up to $R \sim 30$Mpc/h.



Tested both $H_0 = 67$km/s/Mpc (Planck cosmology) and 75km/s/Mpc (Cosmicflows-3 best-fit) for Hubble flow.
Found no big difference on the prediction.

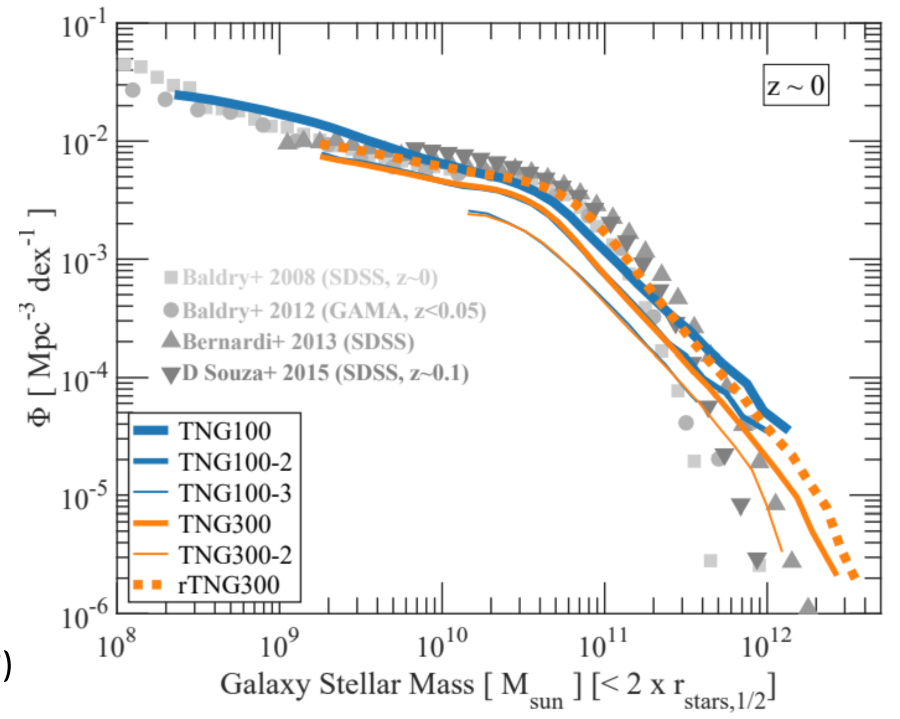
Training & Validation: Illustris-TNG

Training & Validation: Illustris-TNG

- Key Selection Criteria: Similar to Cosmicflows-3
  - Origins : *MW*-like galaxies with stellar mass $4 \times 10^{10} \sim 10^{11}\ M_{sun}$.
  - Use galaxies with $M_B < -16$ mag & $|b| > 10$ deg

- Fiducial Simulation: TNG300-1
  - 205Mpc/*h* boxsize
  - $2{,}500^3$ DM & gas particles

  → Resolution correction with galaxy number density cut rather than face values of M_B and M_*

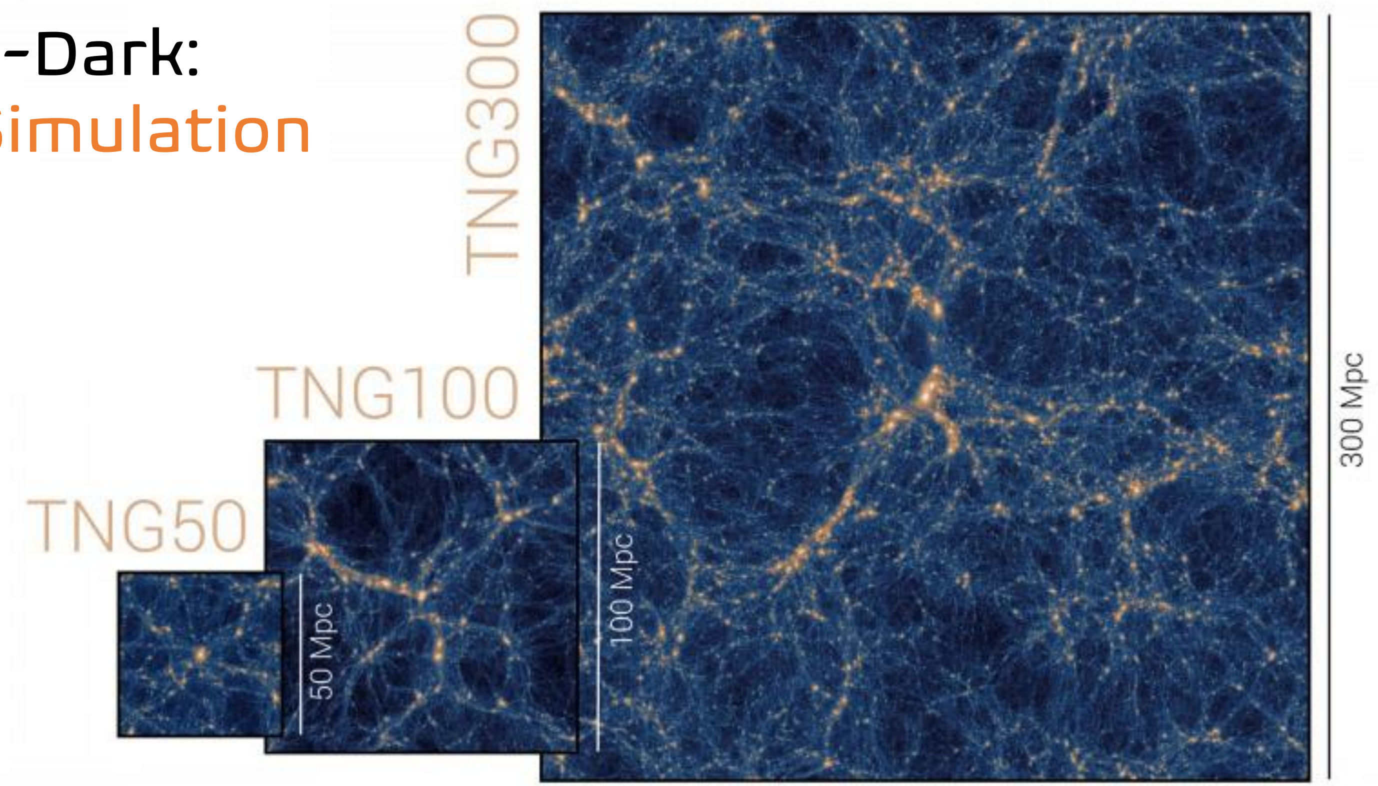


Pillepich+ (2017)

Additional Tests for Different Setups

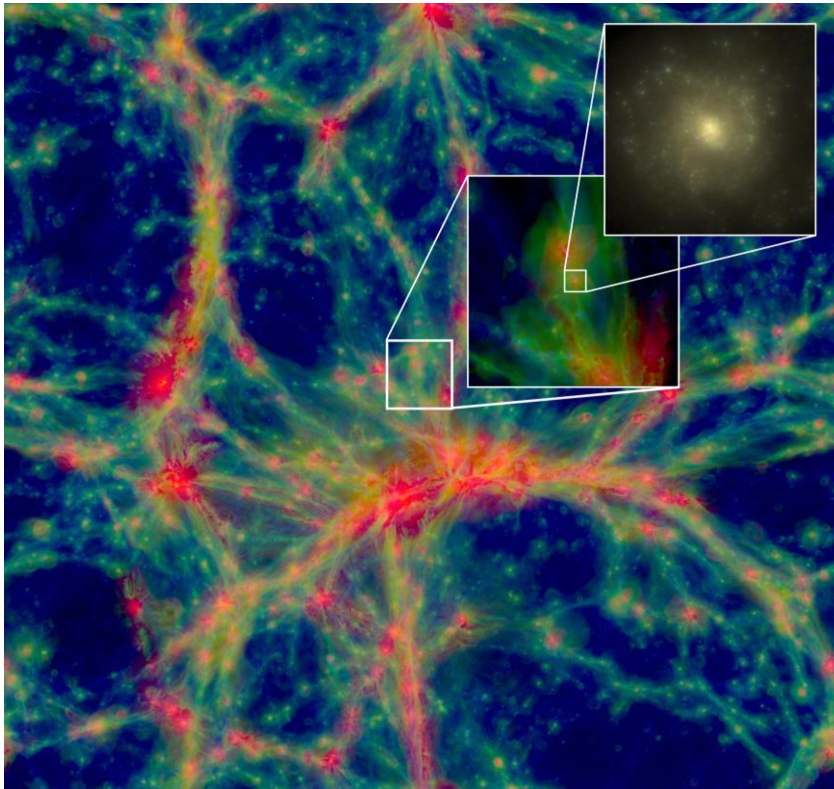
- TNG100-1: Higher Resolution & Lower Boxsize
  - 75Mpc/h boxsize
  - $1{,}820^3$ DM & gas particles
  - Directly use M_B and M_*
  - Use 20Mpc/h subsamples with 64^3 grids

- TNG300-1-Dark: DM-only Simulation

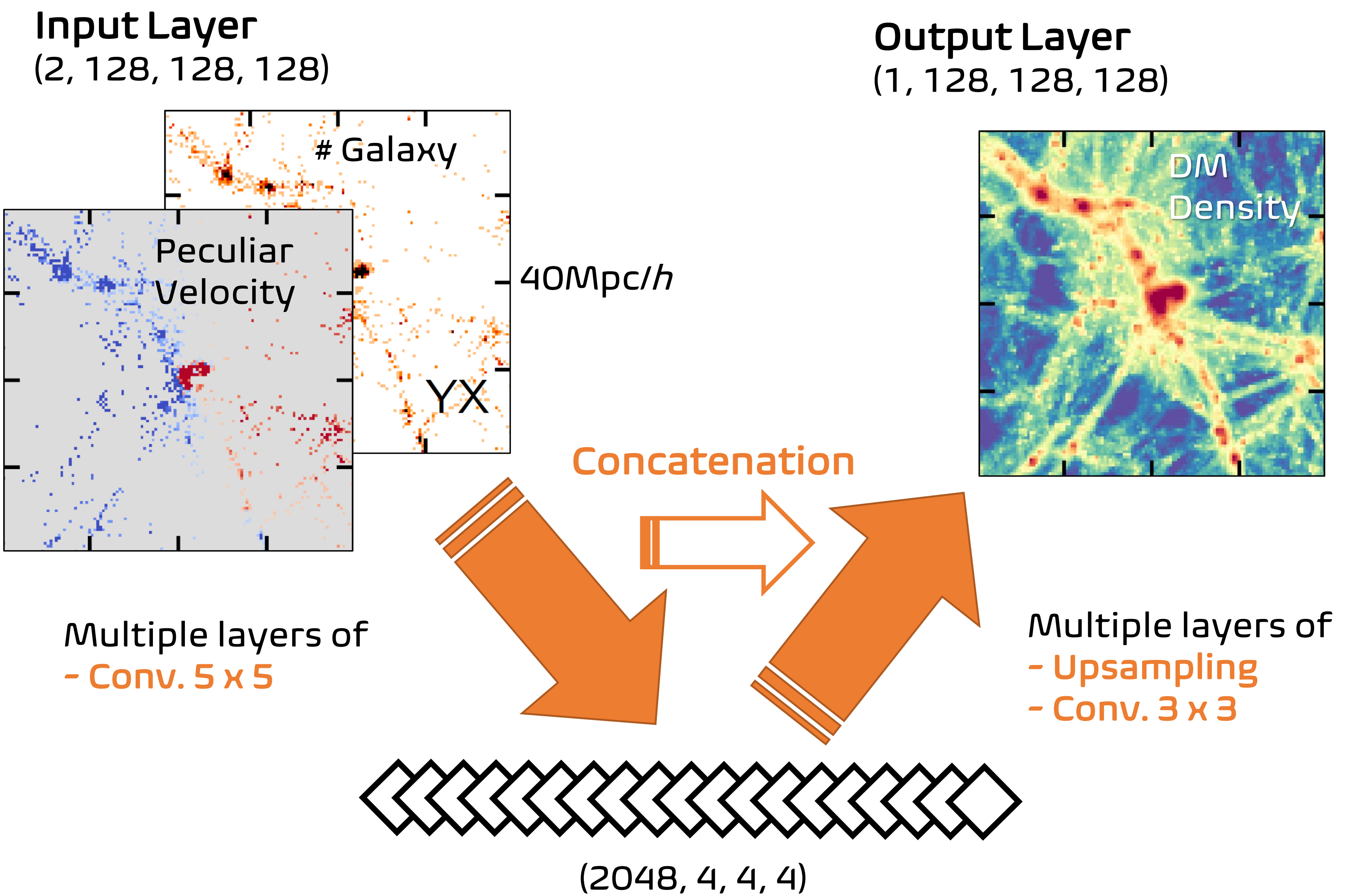


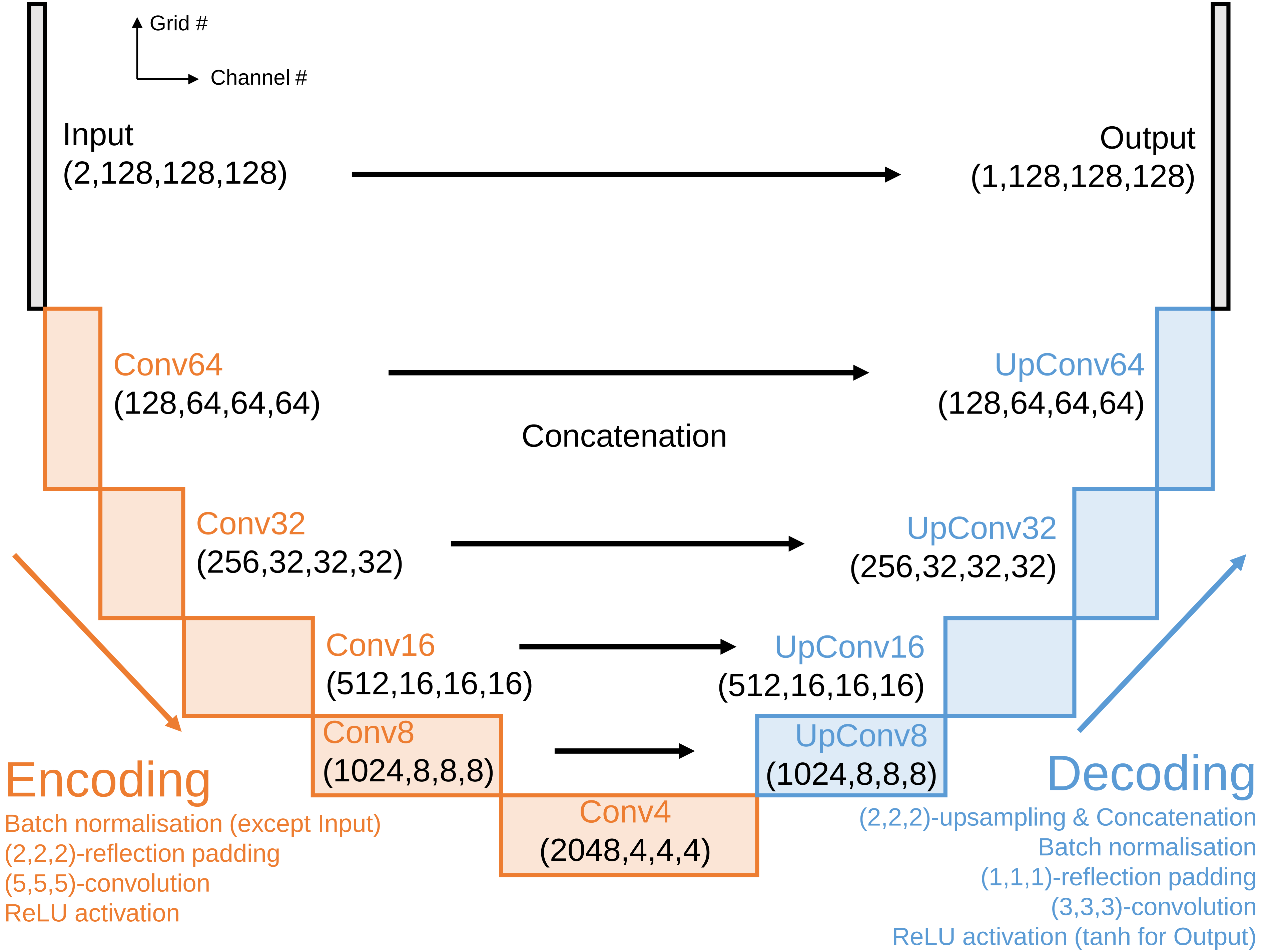
Additional Tests for Different Setups

- EAGLE RefL0100N1504: Different Hydrodynamics
  - 67.77Mpc/h boxsize
  - 1504^3 DM & gas particles
    → Similar resolution to TNG-100-1
  - Magnitude information is available only for massive galaxies
    ($M_* > 10^{8.5}\, M_{sun}$)
    → Galaxy number density cut instead of using M_B

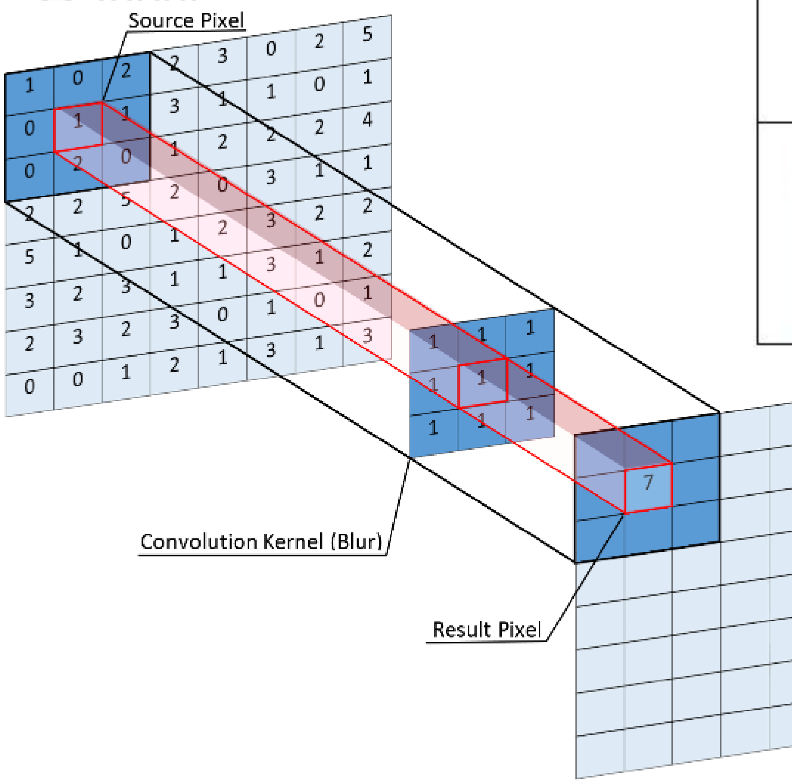


Method: UNet-like CNN





Grid #
Channel #
Input
(2,128,128,128)
Output
(1,128,128,128)
Conv64
(128,64,64,64)
UpConv64
(128,64,64,64)
Concatenation
Conv32
(256,32,32,32)
UpConv32
(256,32,32,32)
Conv16
(512,16,16,16)
UpConv16
(512,16,16,16)
Conv8
(1024,8,8,8)
UpConv8
(1024,8,8,8)
Conv4
(2048,4,4,4)
Encoding
Batch normalisation (except Input)
(2,2,2)-reflection padding
(5,5,5)-convolution
ReLU activation
Decoding
(2,2,2)-upsampling & Concatenation
Batch normalisation
(1,1,1)-reflection padding
(3,3,3)-convolution
ReLU activation (tanh for Output)



Original	Gaussian Blur	Sharpen	Edge Detection
$\begin{bmatrix}0&0&0\\0&1&0\\0&0&0\end{bmatrix}$	$\frac{1}{16}\begin{bmatrix}1&2&1\\2&4&2\\1&2&1\end{bmatrix}$	$\begin{bmatrix}0&-1&0\\-1&5&-1\\0&-1&0\end{bmatrix}$	$\begin{bmatrix}-1&-1&-1\\-1&8&-1\\-1&-1&-1\end{bmatrix}$
			

Extract Features

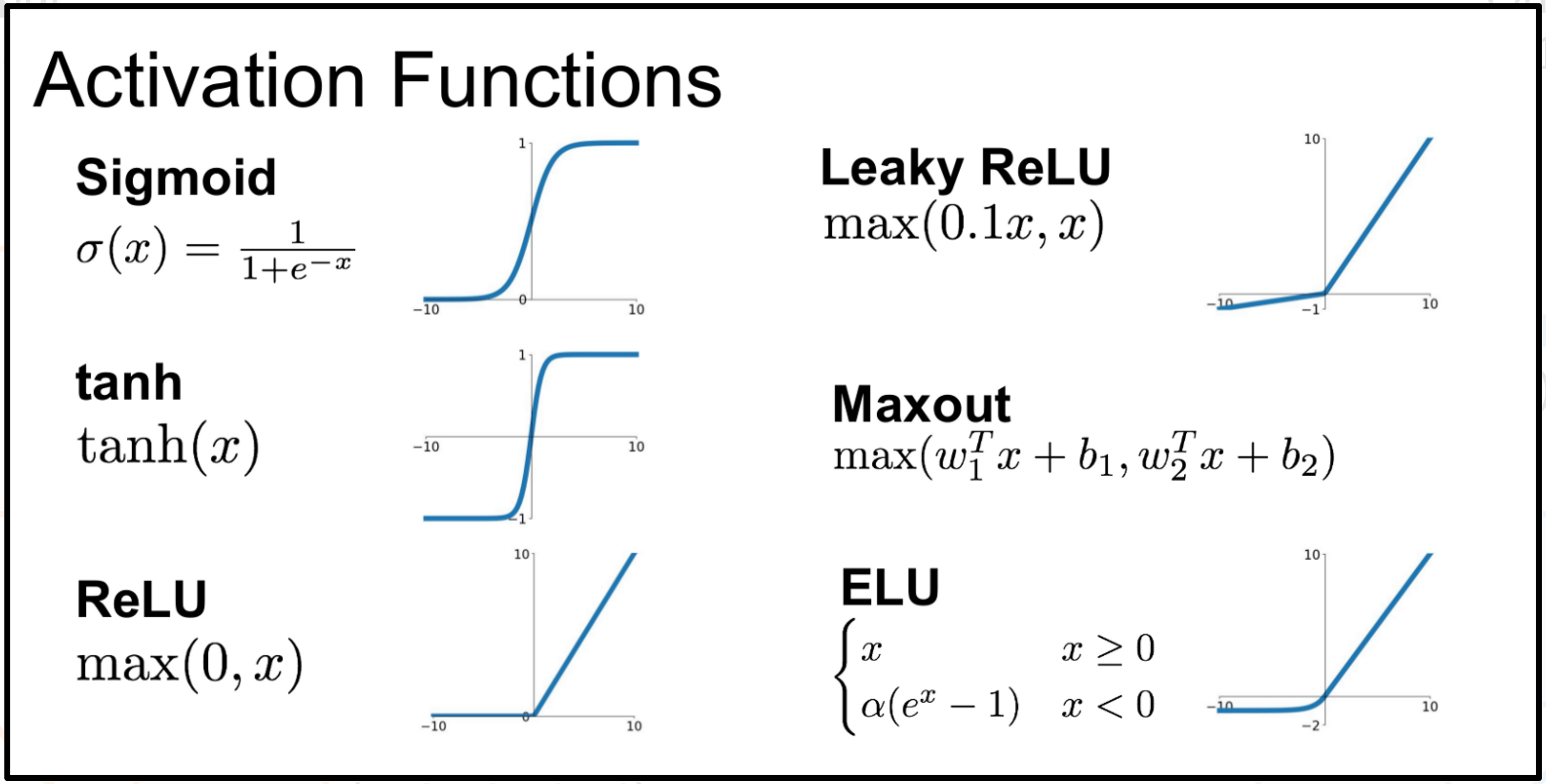
Encoding

Batch normalisation (except Input)
(2,2,2)-reflection padding
(5,5,5)-convolution
ReLU activation

Decoding

(2,2,2)-upsampling & Concatenation
Batch normalisation
(1,1,1)-reflection padding
(3,3,3)-convolution
ReLU activation (tanh for Output)

Activation Functions



Add Nonlinearity

Encoding

Batch normalisation (except Input)
(2,2,2)-reflection padding
(5,5,5)-convolution
ReLU activation

Decoding

(2,2,2)-upsampling & Concatenation
Batch normalisation
(1,1,1)-reflection padding
(3,3,3)-convolution
ReLU activation (tanh for Output)

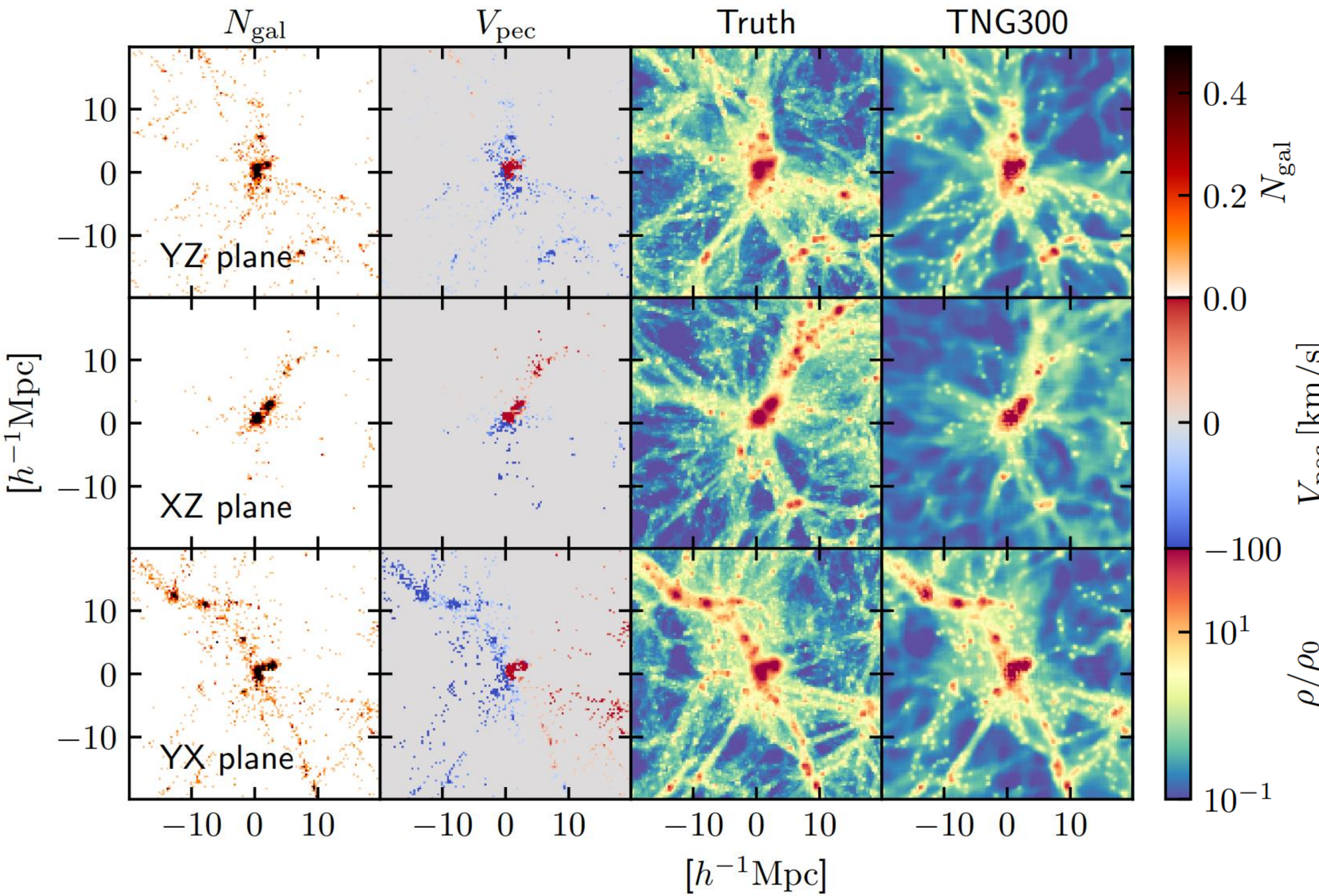
Method: UNet-like CNN

- Number of independent centers
  - Training: 10,629
  - Validation: 1,256
- Loss function

$$\mathcal{L} \propto \sum \left(\log_{10} \rho_{\mathrm{truth}} - \log_{10} \rho_{\mathrm{pred}}\right)^2$$

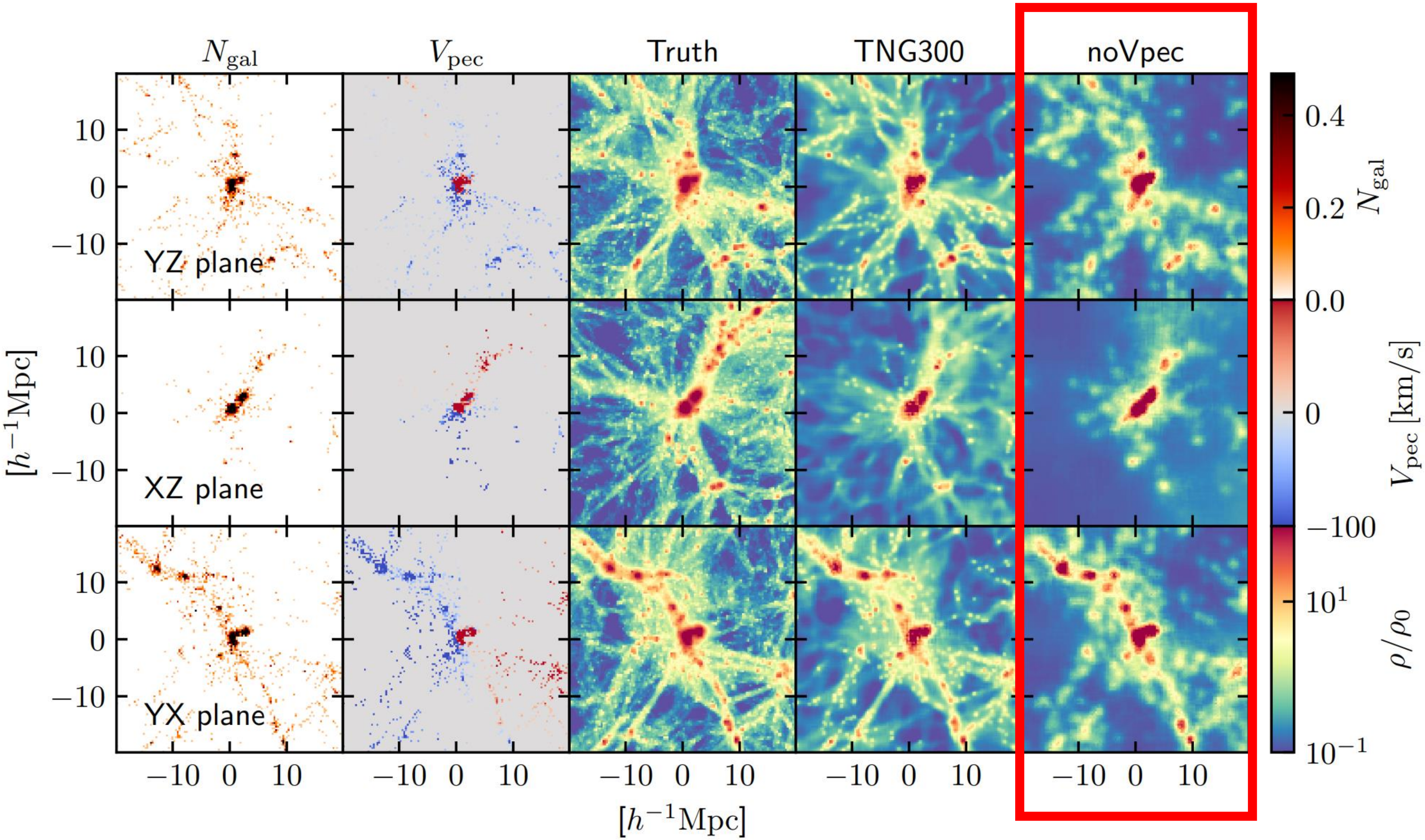
- Tool : Tensorflow 2 / Keras
- 200~400 epochs; 3~4 days w/ NVIDIA Tesla V100(s)

Performance Test



Hong, Jeong, Hwang & Kim; 2008.1738

Cf) Training without Peculiar Velocity

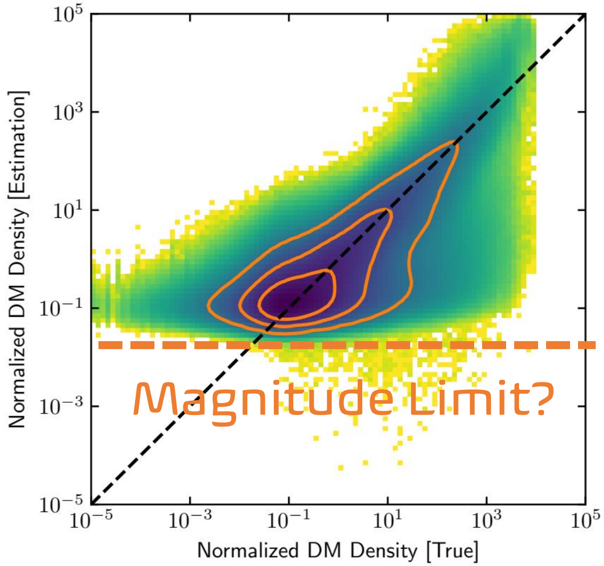


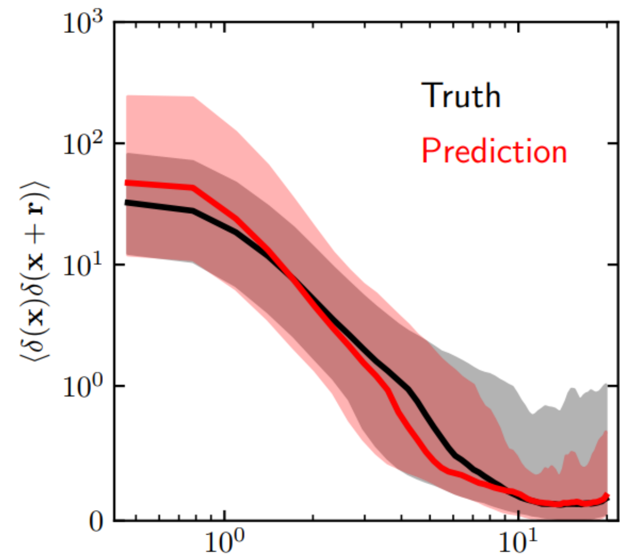
Performance Test

Pixel-to-Pixel

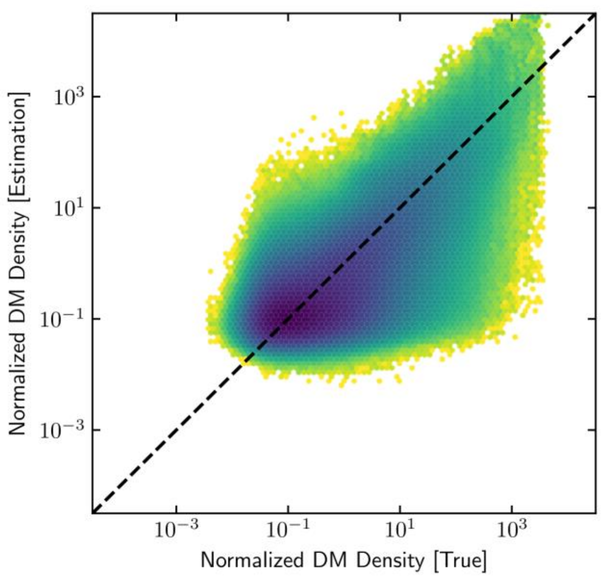
Two-point Correlation Function

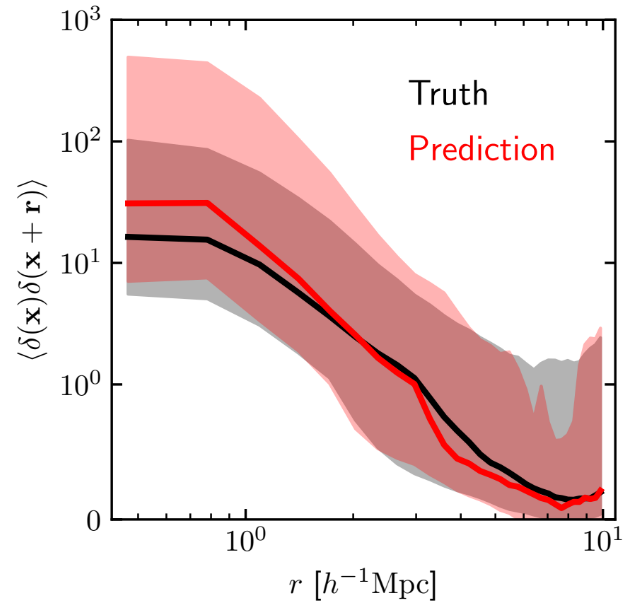
TNG300-1



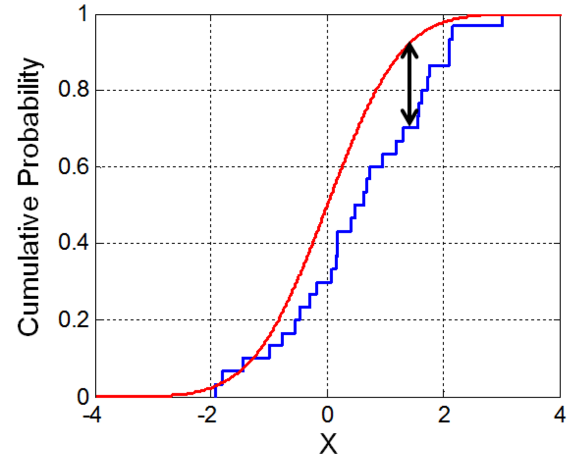


Cf) EAGLE
(Different Hydrodynamics)





Performance Test



Kolmogorov-Smirnov statistics

Different magnitude cut

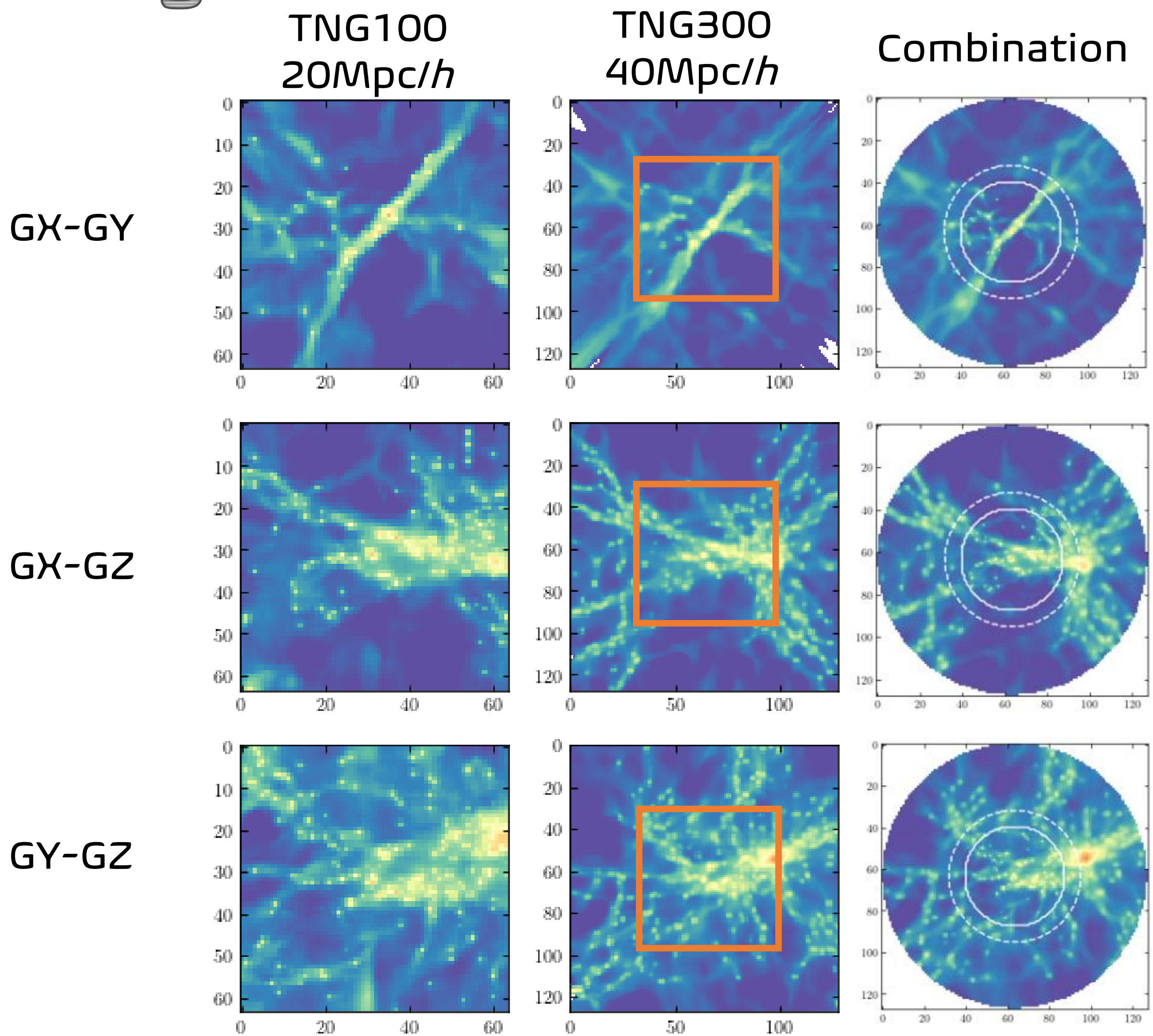
Model	$\log_{10}(\rho_{\rm pred}/\rho_{\rm truth})$	$\mathrm{KS}(\xi_{\rm pred}, \xi_{\rm truth})$		
		$0-1\,\mathrm{Mpc}/h$	$1-3\,\mathrm{Mpc}/h$	$3-10\,\mathrm{Mpc}/h$
TNG100	-0.014 ± 0.543	0.263 ± 0.035	0.175 ± 0.087	0.130 ± 0.042
EAGLE-TNG100	$+0.129 \pm 0.491$	0.171 ± 0.055	0.152 ± 0.047	0.149 ± 0.040
TNG300	-0.020 ± 0.451	0.153 ± 0.035	0.134 ± 0.040	0.163 ± 0.017
16mag	-0.008 ± 0.468	0.109 ± 0.010	0.161 ± 0.033	0.254 ± 0.016
17mag	$+0.017 \pm 0.481$	0.143 ± 0.037	0.168 ± 0.018	0.251 ± 0.019
noVpec	$+0.016 \pm 0.481$	0.367 ± 0.115	0.407 ± 0.061	0.170 ± 0.036
stellarMass	-0.050 ± 0.471	0.186 ± 0.056	0.218 ± 0.016	0.269 ± 0.021
DMhalo	$+0.002 \pm 0.481$	0.264 ± 0.029	0.243 ± 0.030	0.263 ± 0.034

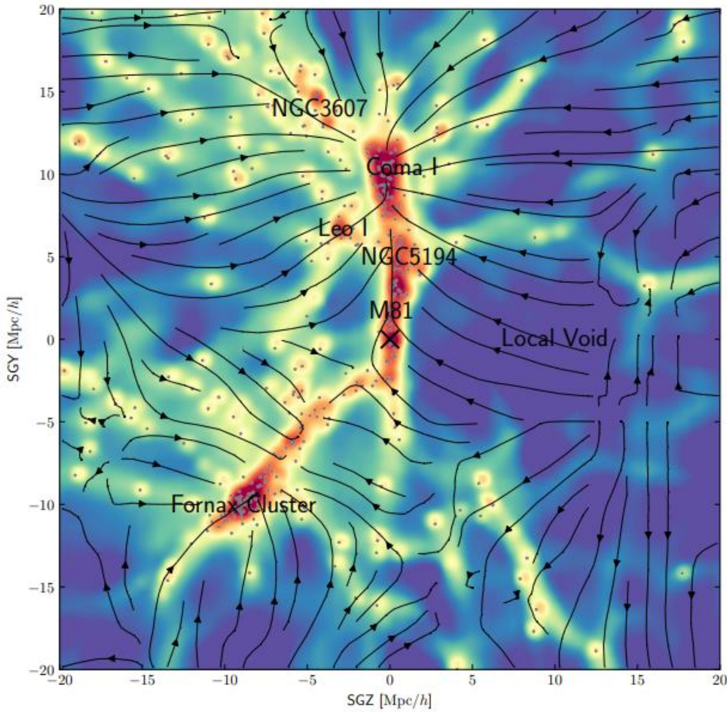
Without peculiar velocity

Stellar mass, instead of galaxy number, as input

Dark matter-only simulation

Convergence Test: Local Universe

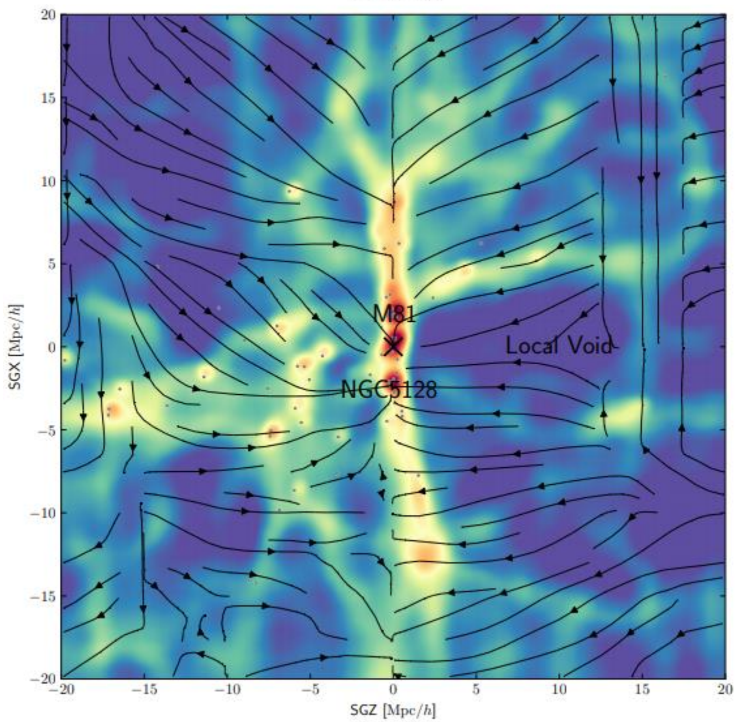


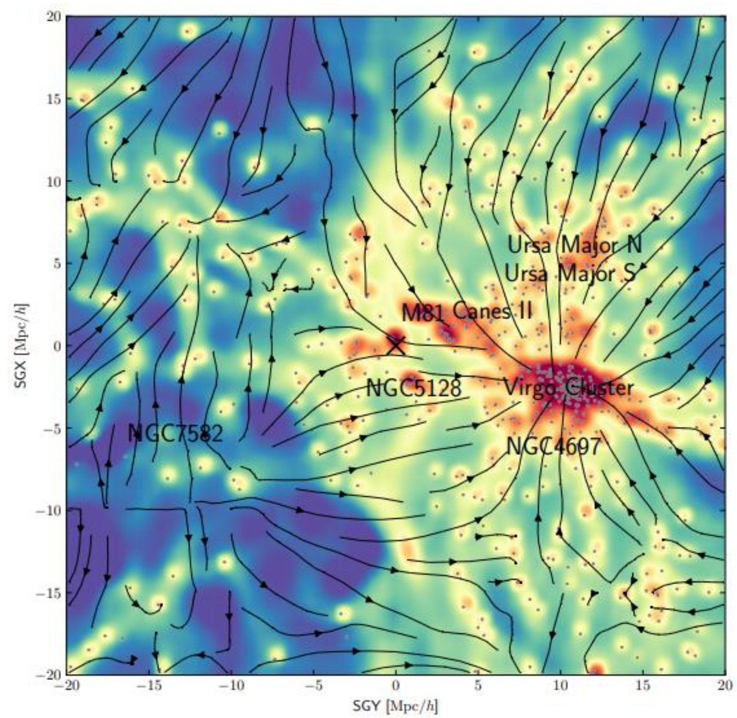


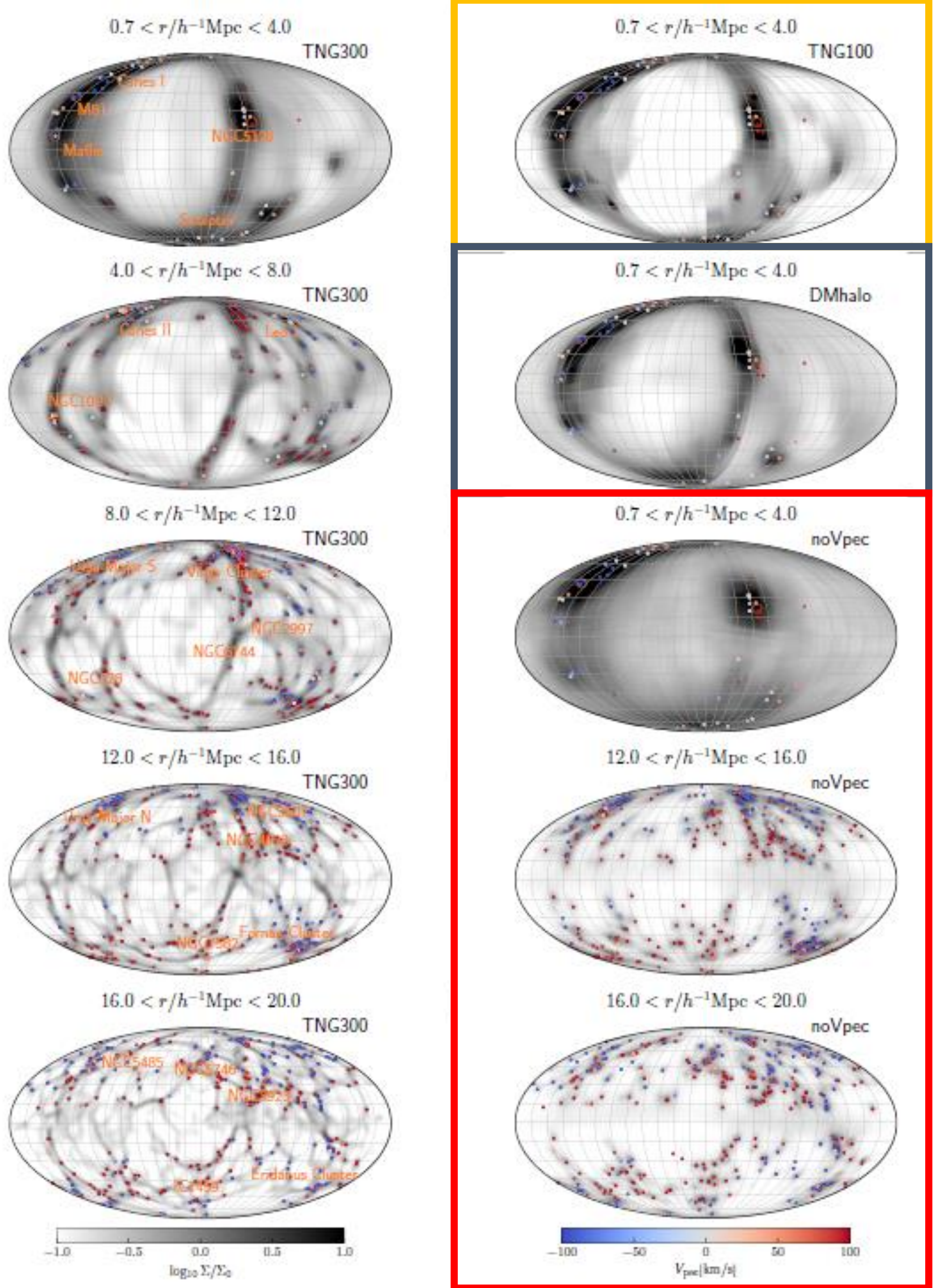
Color: DM Density

Arrows: Gradient of Grav. Potential

Thickness: 4 Mpc/*h*







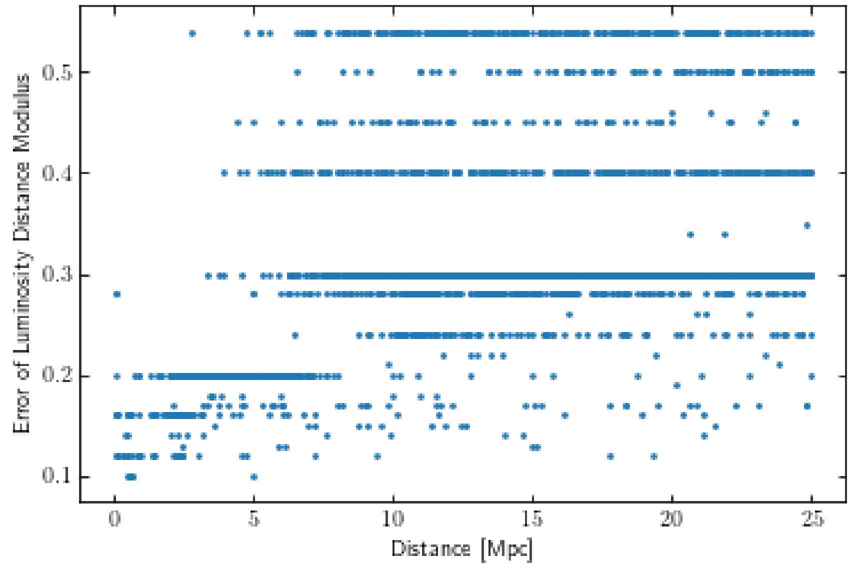
TNG100-1
High-resolution
& Low-volume

TNG300-1-Dark
DM-only
Simulation

Reconstruction
without using
Peculiar
Velocity

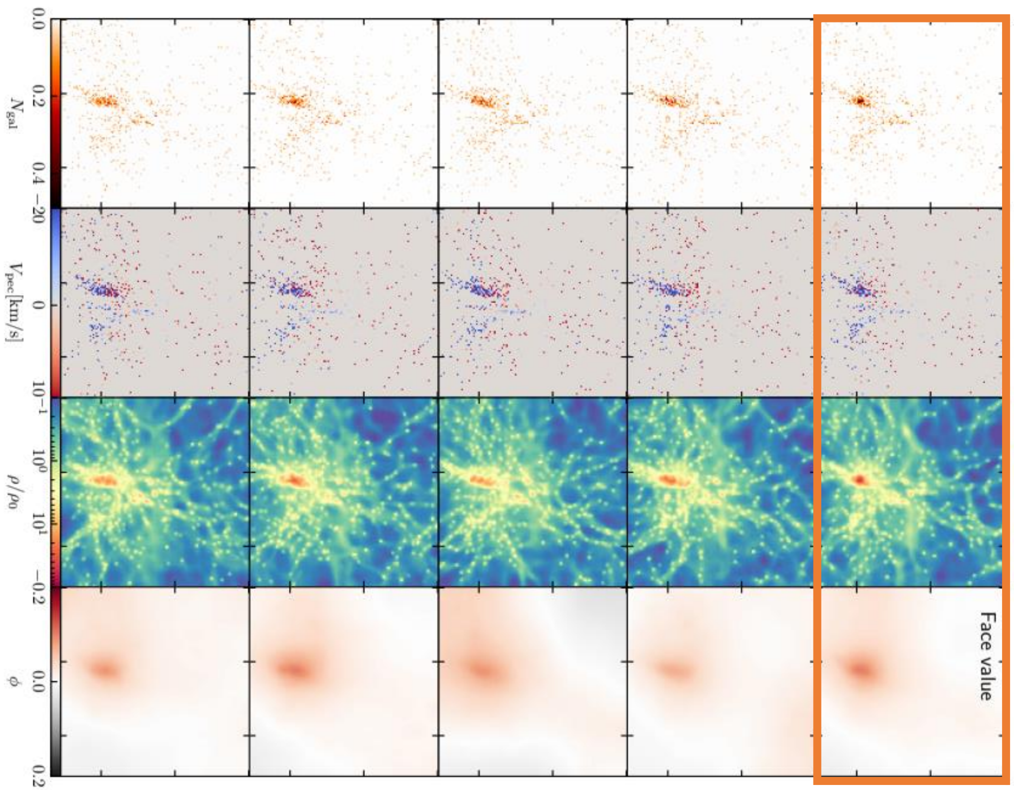
Stress Test: Adding Distance Error

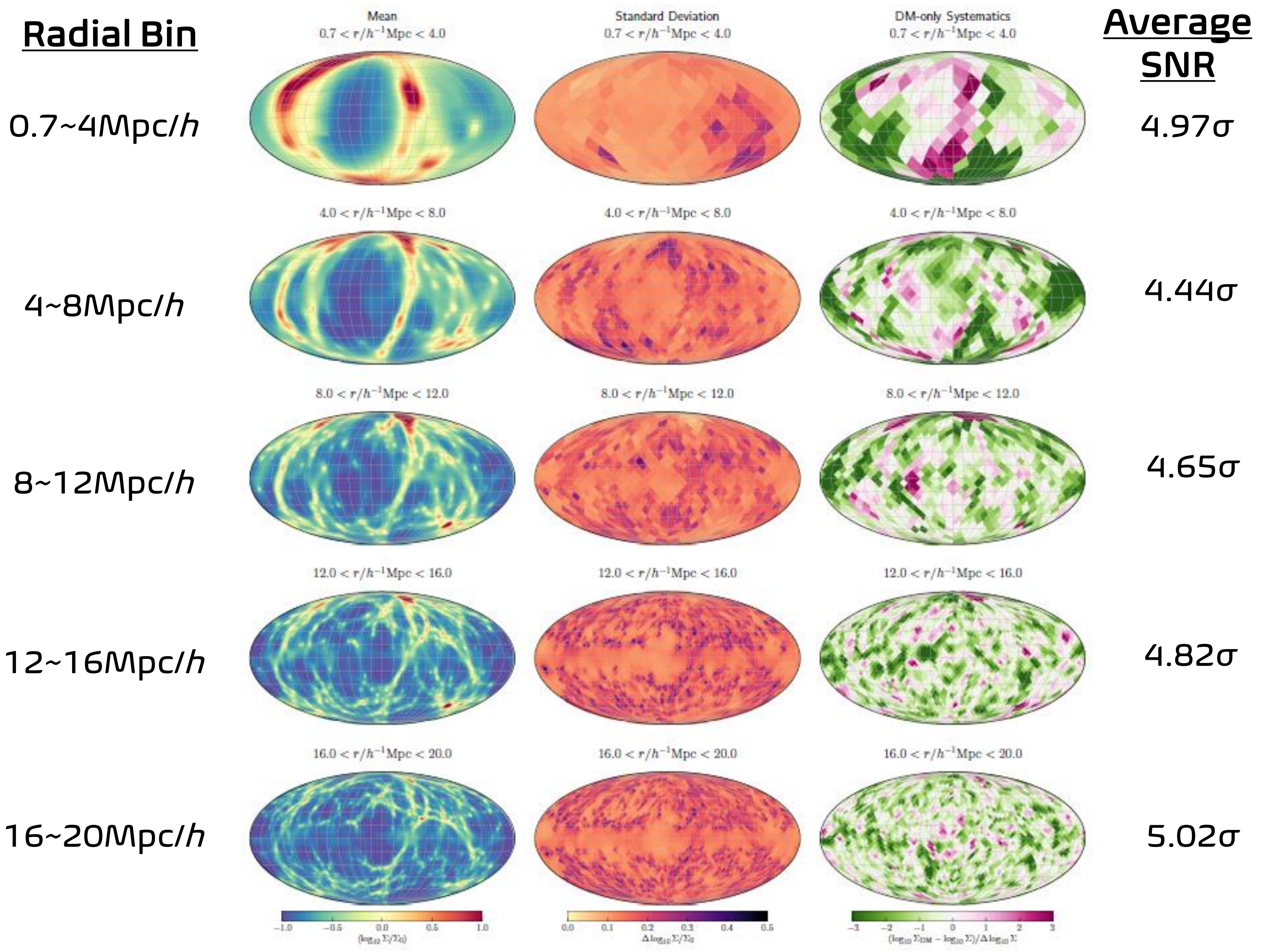
Cosmicflows-3 has
0.1~0.5 of 1σ error of
luminosity distance modulus
→ 5~30% of distance error



Create 1,000 sets of
Cosmicflows-3 catalogues by
applying normal random
distribution of luminosity
distance modulus.

Recalculate peculiar velocity
with new distance.





Radial Bin
Average SNR
Mean
Standard Deviation
DM-only Systematics
0.7 < r/h^-1 Mpc < 4.0
4.0 < r/h^-1 Mpc < 8.0
8.0 < r/h^-1 Mpc < 12.0
12.0 < r/h^-1 Mpc < 16.0
16.0 < r/h^-1 Mpc < 20.0
0.7~4Mpc/h
4.97σ
4~8Mpc/h
4.44σ
8~12Mpc/h
4.65σ
12~16Mpc/h
4.82σ
16~20Mpc/h
5.02σ

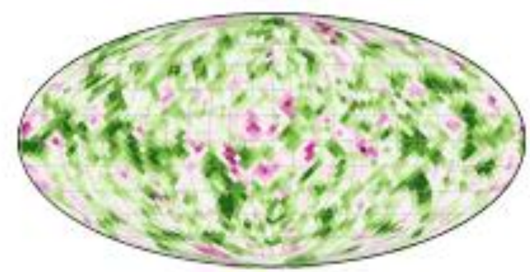
On-sky Average Systematics

$$\Delta_{\rm sys} \equiv \frac{|\log_{10}\Sigma - \log_{10}\Sigma_{\rm TNG300}|}{\Delta\log_{10}\Sigma_{\rm TNG300}}$$

Different assumption on the Hubble parameter: $H_0 = 75\ \mathrm{km/s/Mpc}$

Comparison Model	$0.7-4\,\mathrm{Mpc}/h$	$4-8\,\mathrm{Mpc}/h$	$8-12\,\mathrm{Mpc}/h$	$12-16\,\mathrm{Mpc}/h$	$16-20\,\mathrm{Mpc}/h$
TNG100	$2.281\ (1.837^{+1.993}_{-1.104})$	$1.474\ (1.196^{+1.414}_{-0.842})$	-	-	-
$\boxed{\text{diffH0}}$	$0.212\ (0.171^{+0.223}_{-0.115})$	$0.162\ (0.133^{+0.148}_{-0.092})$	$0.154\ (0.116^{+0.161}_{-0.083})$	$0.152\ (0.117^{+0.153}_{-0.082})$	$0.160\ (0.128^{+0.151}_{-0.092})$
16mag	$1.032\ (0.949^{+0.748}_{-0.647})$	$1.093\ (0.868^{+1.089}_{-0.611})$	$0.862\ (0.716^{+0.729}_{-0.508})$	$0.785\ (0.641^{+0.751}_{-0.455})$	$0.804\ (0.631^{+0.790}_{-0.443})$
17mag	$1.178\ (0.901^{+1.081}_{-0.572})$	$1.105\ (0.889^{+1.026}_{-0.621})$	$1.001\ (0.815^{+0.947}_{-0.575})$	$0.887\ (0.726^{+0.862}_{-0.502})$	$0.898\ (0.734^{+0.833}_{-0.506})$
noVpec	$1.935\ (1.715^{+1.919}_{-1.359})$	$1.105\ (0.834^{+1.120}_{-0.631})$	$0.943\ (0.701^{+0.890}_{-0.524})$	$0.828\ (0.672^{+0.751}_{-0.470})$	$0.750\ (0.626^{+0.742}_{-0.440})$
stellarMass	$1.544\ (1.256^{+1.435}_{-0.843})$	$1.175\ (0.946^{+1.156}_{-0.684})$	$0.925\ (0.734^{+0.909}_{-0.521})$	$0.877\ (0.692^{+0.837}_{-0.485})$	$0.907\ (0.713^{+0.899}_{-0.490})$
DMhalo	$1.737\ (1.154^{+2.253}_{-0.863})$	$1.445\ (1.127^{+1.414}_{-0.816})$	$1.176\ (0.913^{+1.097}_{-0.610})$	$1.057\ (0.846^{+1.029}_{-0.595})$	$0.957\ (0.796^{+0.889}_{-0.574})$





Summary

- Deep Learning can reconstruct DM density & potential of the Local Universe from galaxy distribution.
  → Possibility of future detailed constrained simulation

- Our result is consistent with different boxsizes & simulation setups.

- Even with distance measurement error,
  the 2D projection shows reasonable mapping with average signal-to-ratio 4.8σ.

- Using peculiar velocity is crucial for reconstructing a few Mpc-scale Cosmic Web.

- Deep & complete galaxy survey of Local Universe is mandatory for better prediction of underdense region.