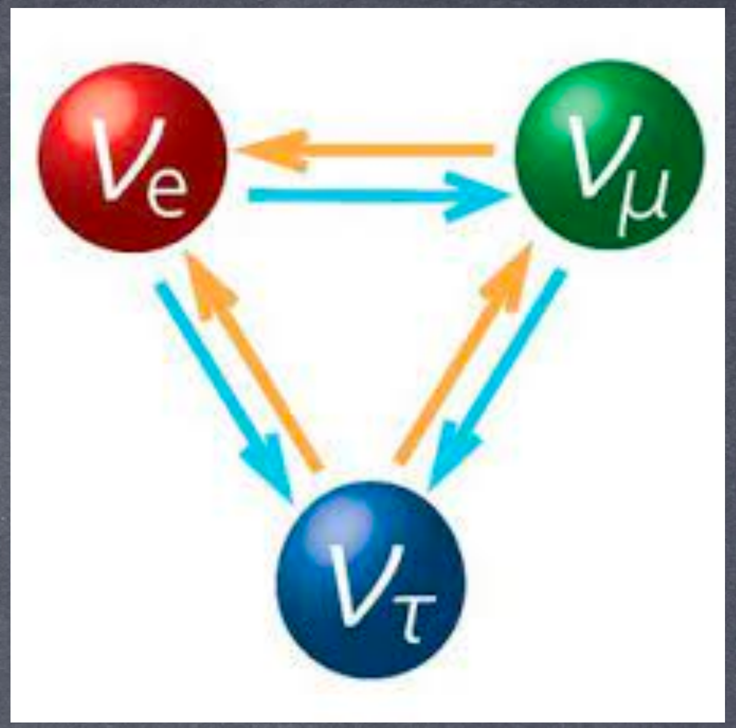
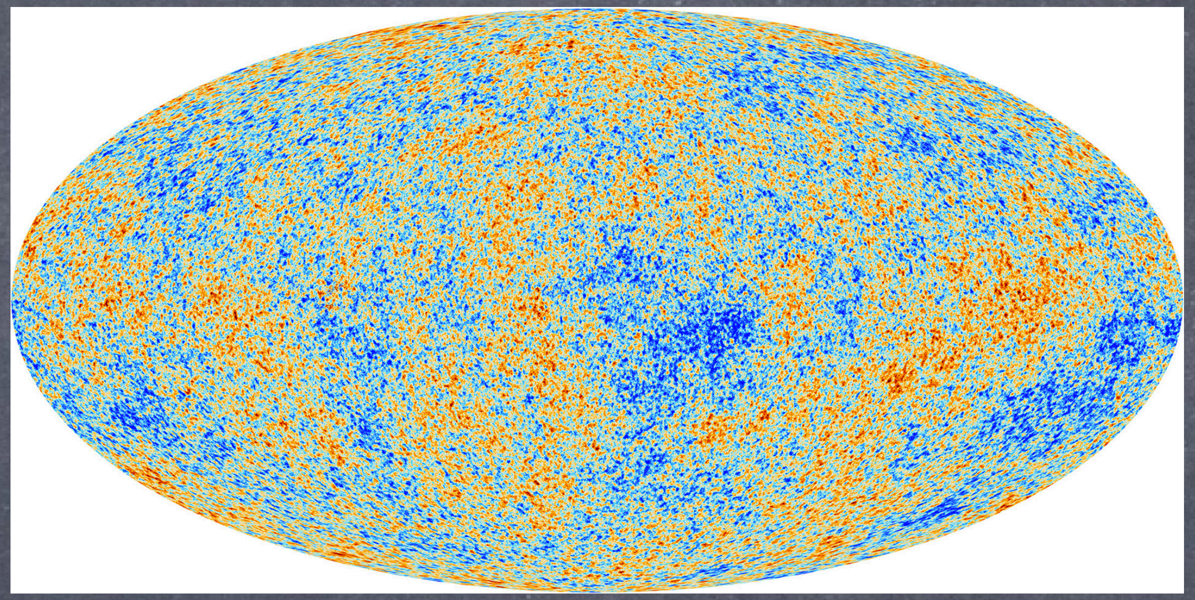


Beyond-Gaussian statistics for cosmological clustering - k-Nearest Neighbor Distributions

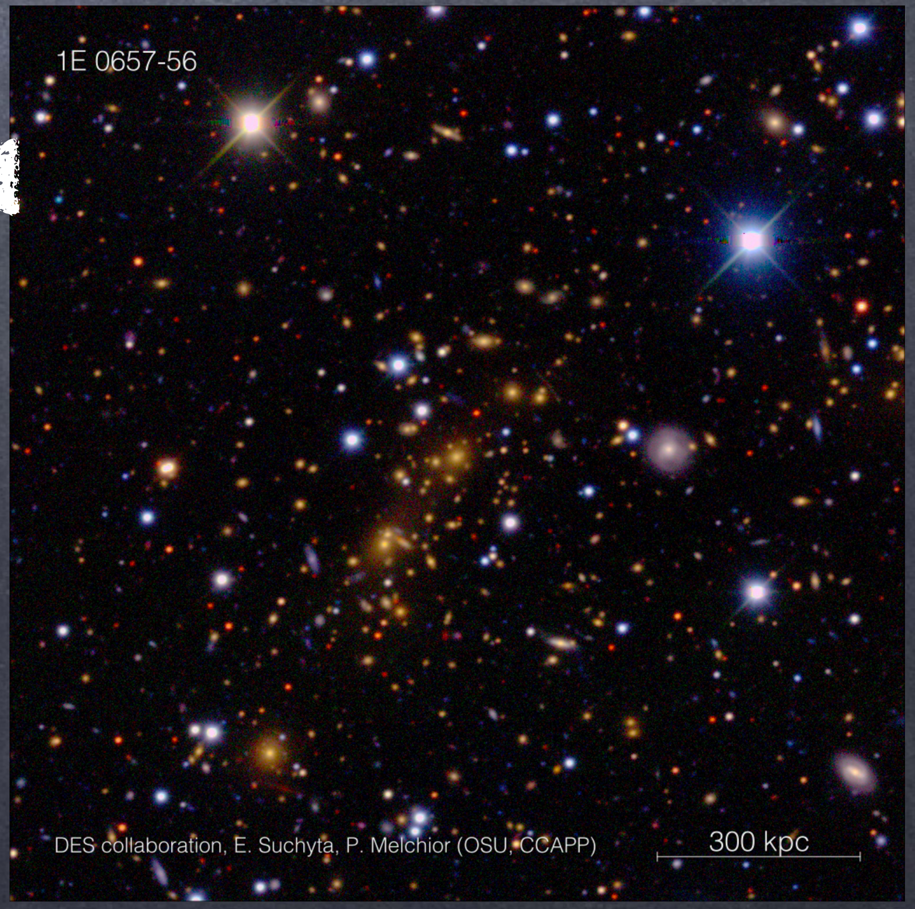
Based on: MNRAS 500(2020) 4, MNRAS 504(2021) 2, MNRAS 511(2022) 2, MNRAS 512 (2022) 3, MNRAS 519 (2023) 4

Arka Banerjee
IISER Pune
@AAPCOS 2023

Background and perturbations

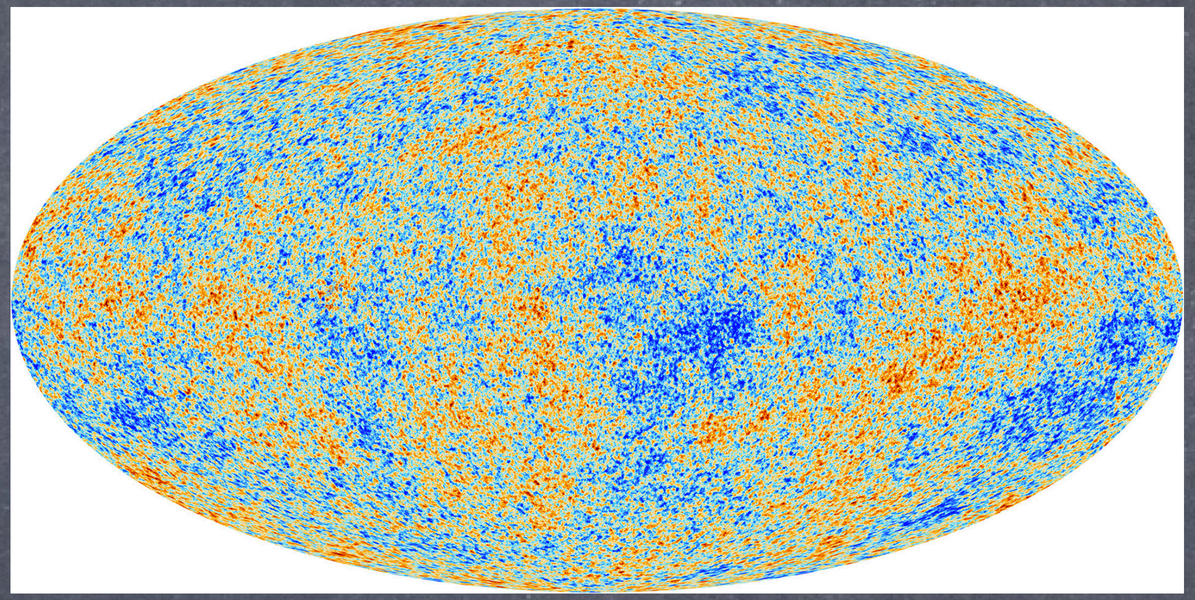


- It is convenient to distinguish between information from the following two phenomena:
  - The expansion rate of the background Universe.
  - The evolution to the perturbations on this background (structure formation).



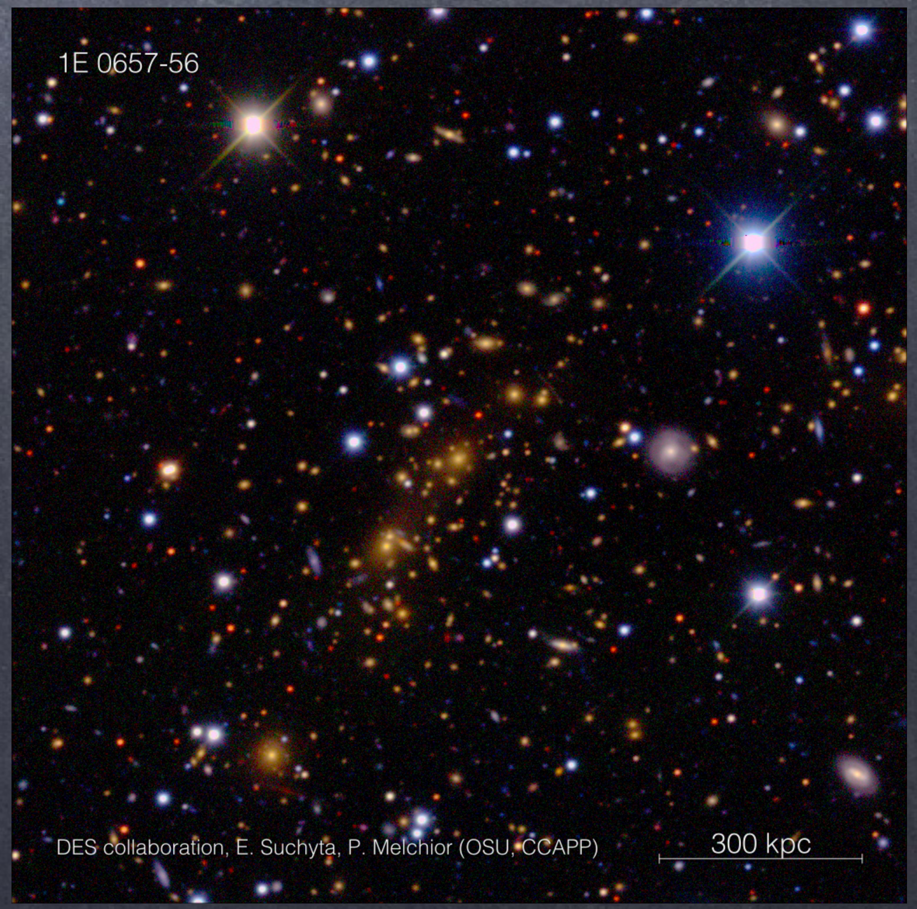
Background and perturbations

- It is convenient to distinguish between information from the following two phenomena:
  - The expansion rate of the background Universe.
  - The evolution to the perturbations on this background (structure formation).
- This evolution is sensitive to the relative abundances of all energy components in the Universe, and their properties.



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

$\delta \sim 10^{4}$



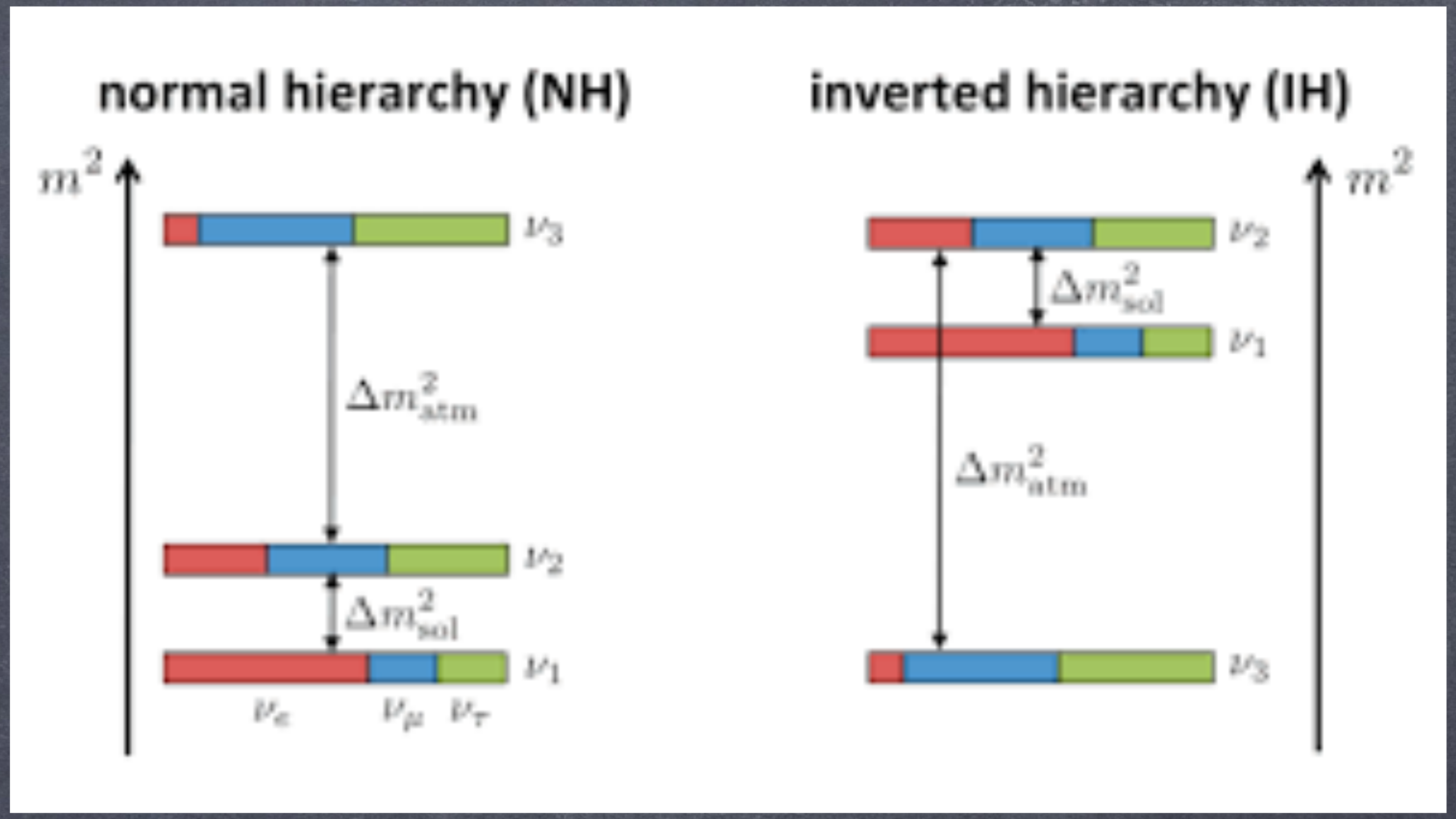
Background and perturbations

- It is convenient to distinguish between information from the following two phenomena:
  - The expansion rate of the background Universe.
  - The evolution to the perturbations on this background (structure formation).
- This evolution is sensitive to the relative abundances of all energy components in the Universe, and their properties.

	Experiment type	Concept	Redshift Range	Primordial FoM	Timescale	Technical Maturity	Comments
DESI	spectro	5000 robotic fiber fed spectrograph on 4m Mayall telescope	$0.1 < z < 2.0$	0.88	now	operating	
Rubin LSST	photo	*ugrizy* wide FoV imaging on a 6.5m effective diameter dedicated telescope	$0 < z < 3$	-	2025-2035	on schedule	Targeting survey for next generation spectroscopic instruments
SPHEREx	narrow-band	Variable Linear Filter imaging on 0.25m aperture from space	$0 < z < 4$	-	2024	on schedule	Focus on primordial non-Gaussianity
MSE+†	spectro	up to 16,000 robotic fiber fed spectrograph on 11.25 m telescope	$1.6 < z < 4$ (ELG+LBG samples)	< 6.1	2029-	high	
MegaMapper	spectro	20,000 robotic fiber fed spectrograph on 6m Magellan clone	$2 < z < 5$	9.4	2029-	high	Builds upon existing hardware and know-how
SpecTel†	spectro	20,000-60,000 robotic fiber fed spectrograph on a dedicated 10m+ class telescope	$1 < z < 6$	< 23	2035-	medium	Potentially very versatile next generation survey instruments
PUMA	21 cm	5000-32000 dish array focused on intensity 21 cm intensity mapping	$0.3 < z < 6$	85 / 26 (32K / 5K optimistic)	2035-	to be demonstrated	Very high effective number density, but $k_\parallel$ modes lost to foregrounds
mm-wave LIM concept	microwave LIM	500-30000 on-chip spectrometers on existing 5-10m telescopes, 80-300 GHz with R$\sim$300-1000	$0 < z < 10$	up to 170	2035 -	to be demonstrated	CMB heritage, can deploy on existing telescopes, signal uncertain, $k_\parallel$ modes lost to foregrounds & resolution

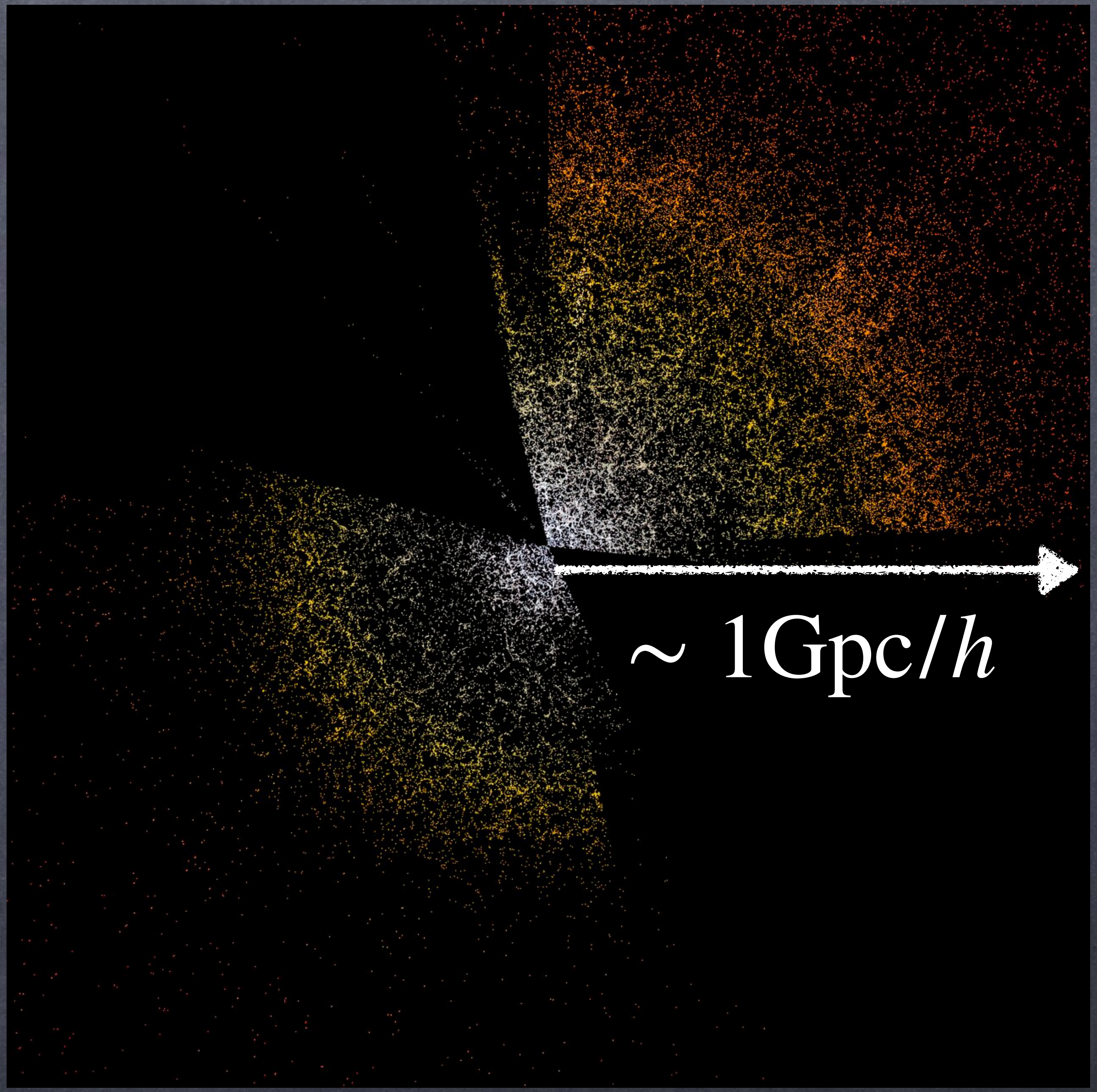
Structure formation: The promise

- What drove inflation? How did it end? Particle spectrum during inflation?
- Is DE consistent with being a cosmological constant at a significantly higher level of accuracy?
- Test the effects of various DM models on structure formation.
- Pin down the total mass of the SM neutrinos. The current bound from cosmology is tantalizingly close to ruling out the inverted hierarchy of neutrino masses.
- Galaxy formation physics, substructure dynamics within halos...



Relevant length scales

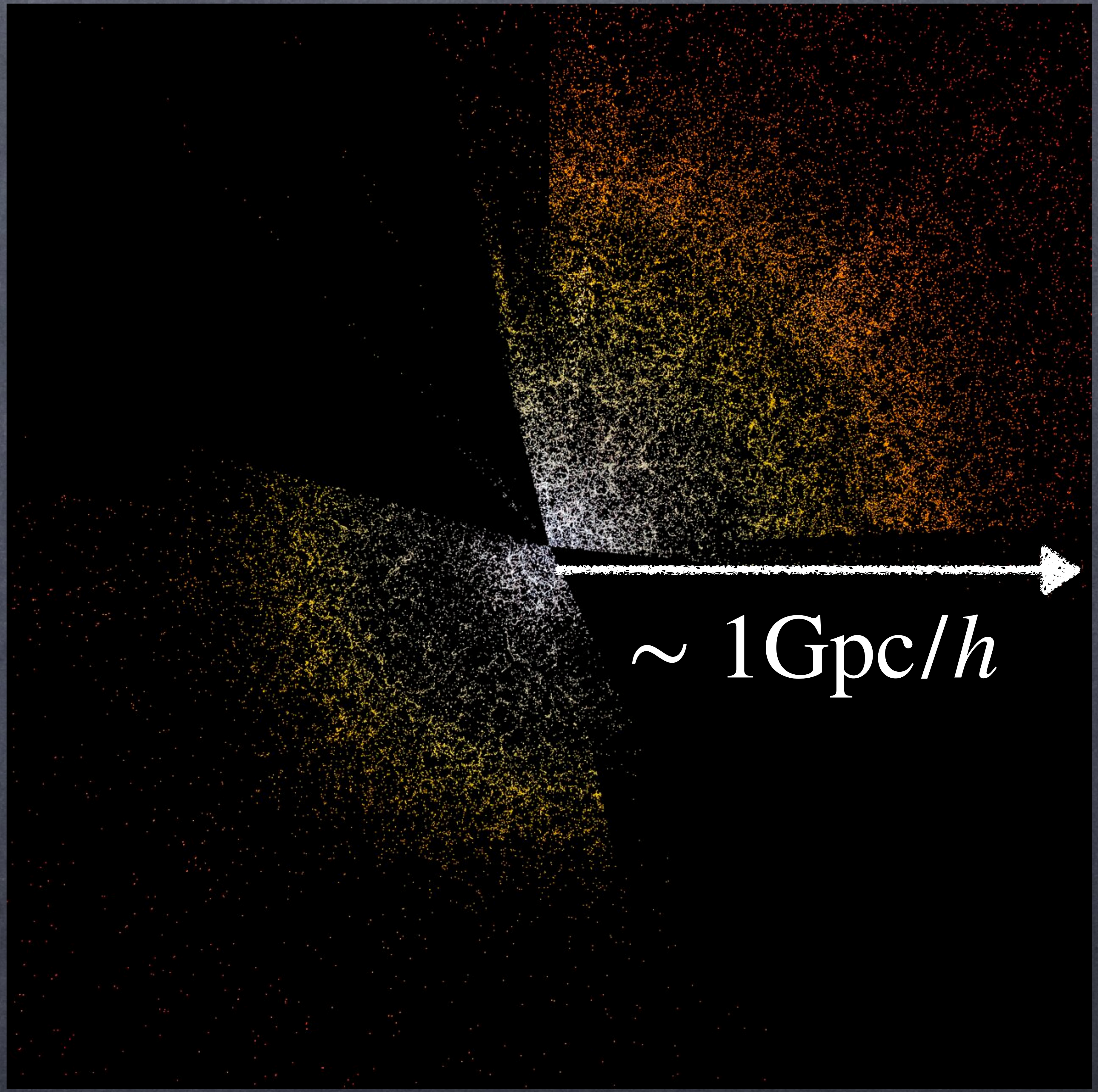
SDSS Collaboration



- The Universe is assumed to be increasingly homogeneous and isotropic on large scales.

Relevant length scales

SDSS Collaboration

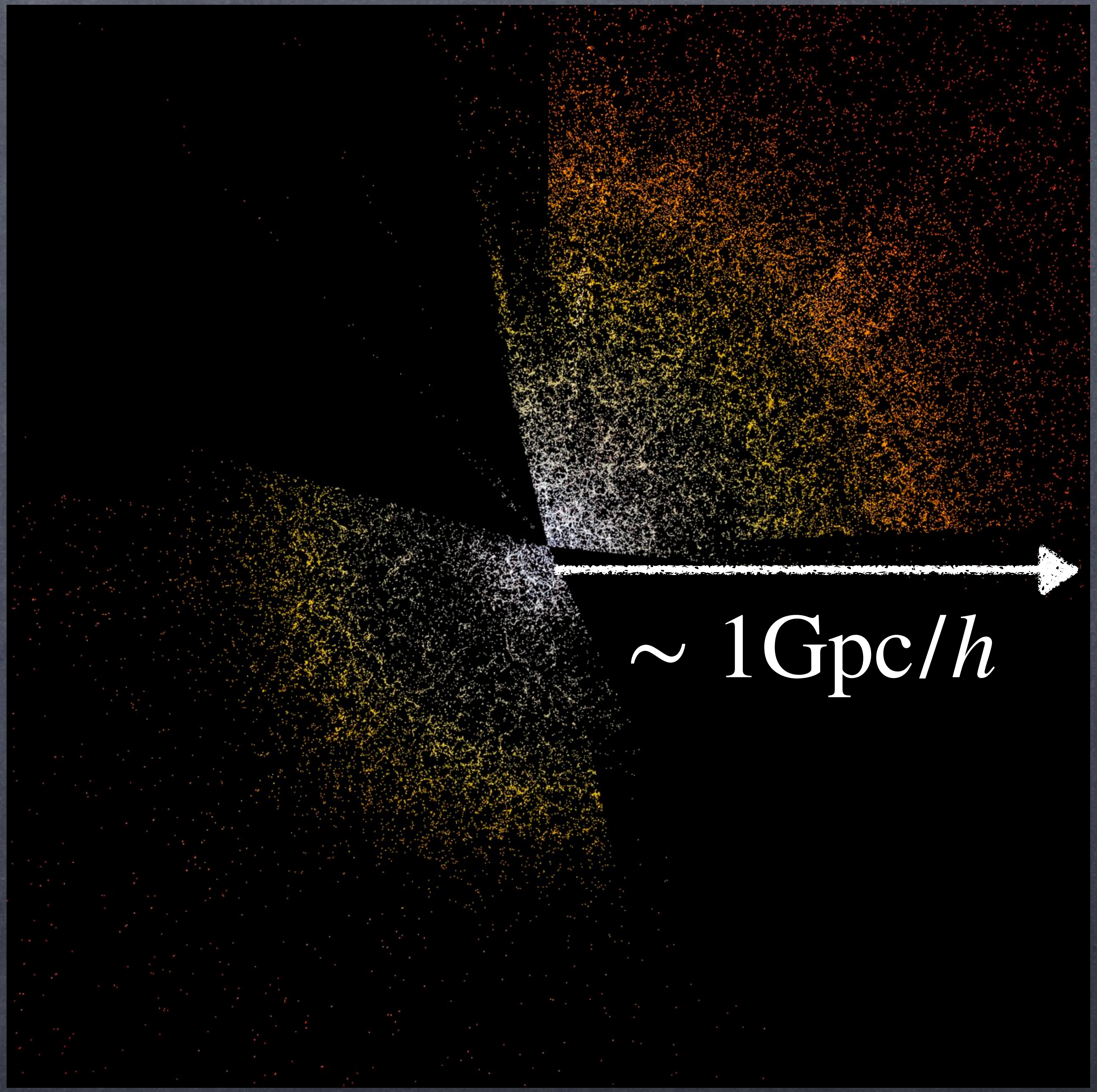


- The Universe is assumed to be increasingly homogeneous and isotropic on large scales.
- Different regions of the Universe will have small fluctuations around the mean, when the volume considered is large.

Relevant length scales

- The Universe is assumed to be increasingly homogeneous and isotropic on large scales.
- Different regions of the Universe will have small fluctuations around the mean, when the volume considered is large.
- This is true even at z=0 (current time).

SDSS Collaboration



Relevant length scales

- Since the density contrast δ is small on large scales, it is possible to use a perturbation theory approach to describe the evolution of δ.

$$\delta(\vec{x}) = \frac{\rho(\vec{x}) - \bar{\rho}}{\bar{\rho}}$$

$$\text{Continuity} : \dot{\delta} = -\frac{1}{a}\overrightarrow{\nabla} \, . \, \vec{v}$$

$$\text{Euler} : \frac{\partial \vec{v}}{\partial t} = -\frac{1}{a}\vec{v} - \frac{1}{a}\overrightarrow{\nabla}\phi$$

$$\text{Poisson} : \nabla^2 \phi = 4\pi G \bar{\rho} a^2 \delta$$

Relevant length scales

- Since the density contrast δ is small on large scales, it is possible to use a perturbation theory approach to describe the evolution of δ.

- Holds down to $\sim 40\mathrm{Mpc}/h$, but needs higher orders in perturbation theory. (For scales, the size of our galaxy is about 20kpc/h).

$$\delta(\vec{x}) = \frac{\rho(\vec{x}) - \bar{\rho}}{\bar{\rho}}$$

$$\text{Continuity} : \dot{\delta} = -\frac{1}{a}\overrightarrow{\nabla} . \vec{v}$$

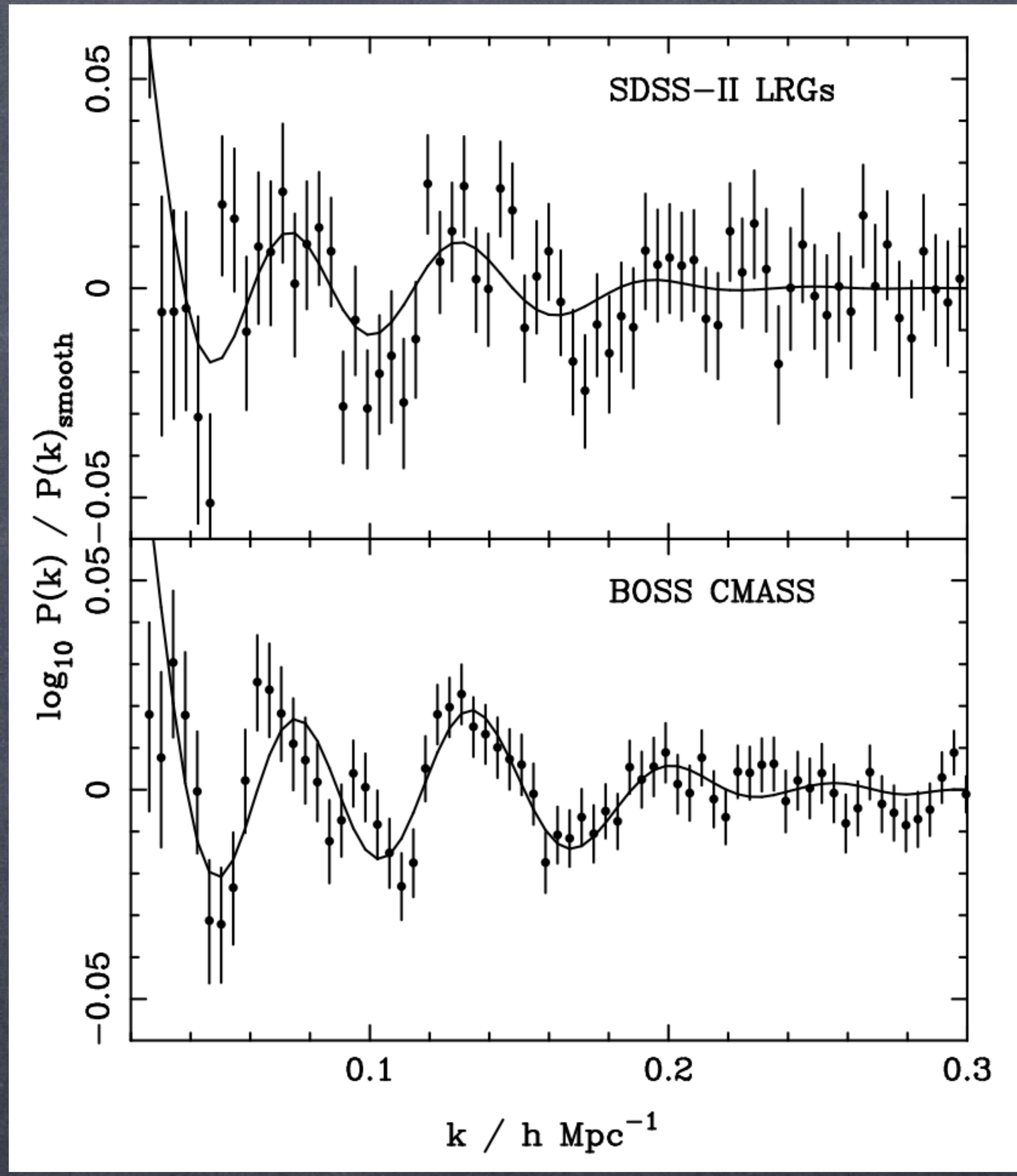
$$\text{Euler} : \frac{\partial \vec{v}}{\partial t} = -\frac{1}{a}\vec{v} - \frac{1}{a}\overrightarrow{\nabla}\phi$$

$$\text{Poisson} : \nabla^2\phi = 4\pi G\bar{\rho}a^2\delta$$

Cosmology from large scales

- Most cosmology analyses in the past and even today focus on information from these large scales.

- We are close (but not quite) to exhausting what we can learn about the Universe from these large scales.



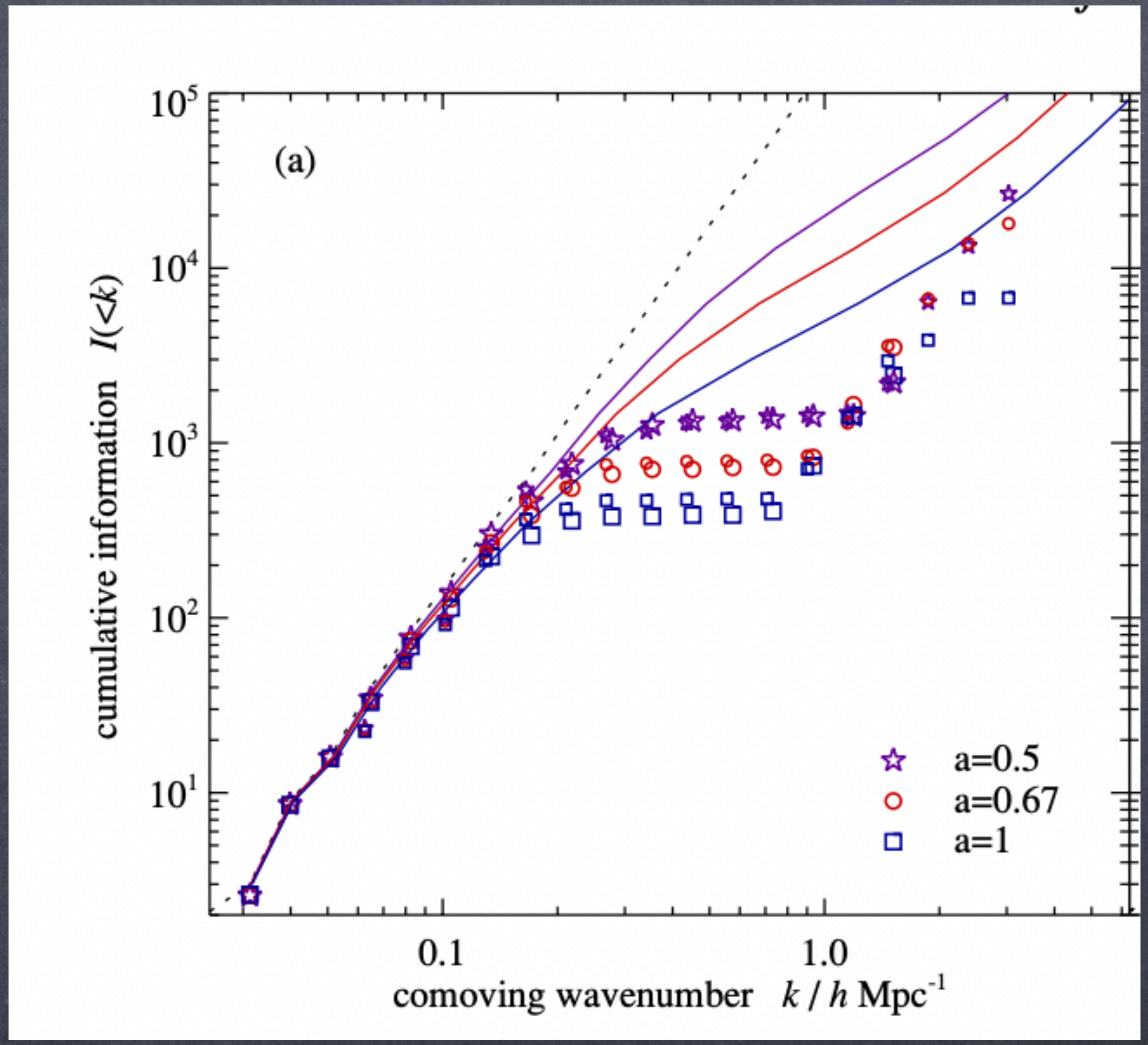
SDSS Collaboration

Why consider smaller scales?

- Many more independent regions within the observable Universe, i.e. greater statistical power.

- The total information naively scales as k_{max}^3. A factor of 2 in scales implies a factor of 8 in the total information.

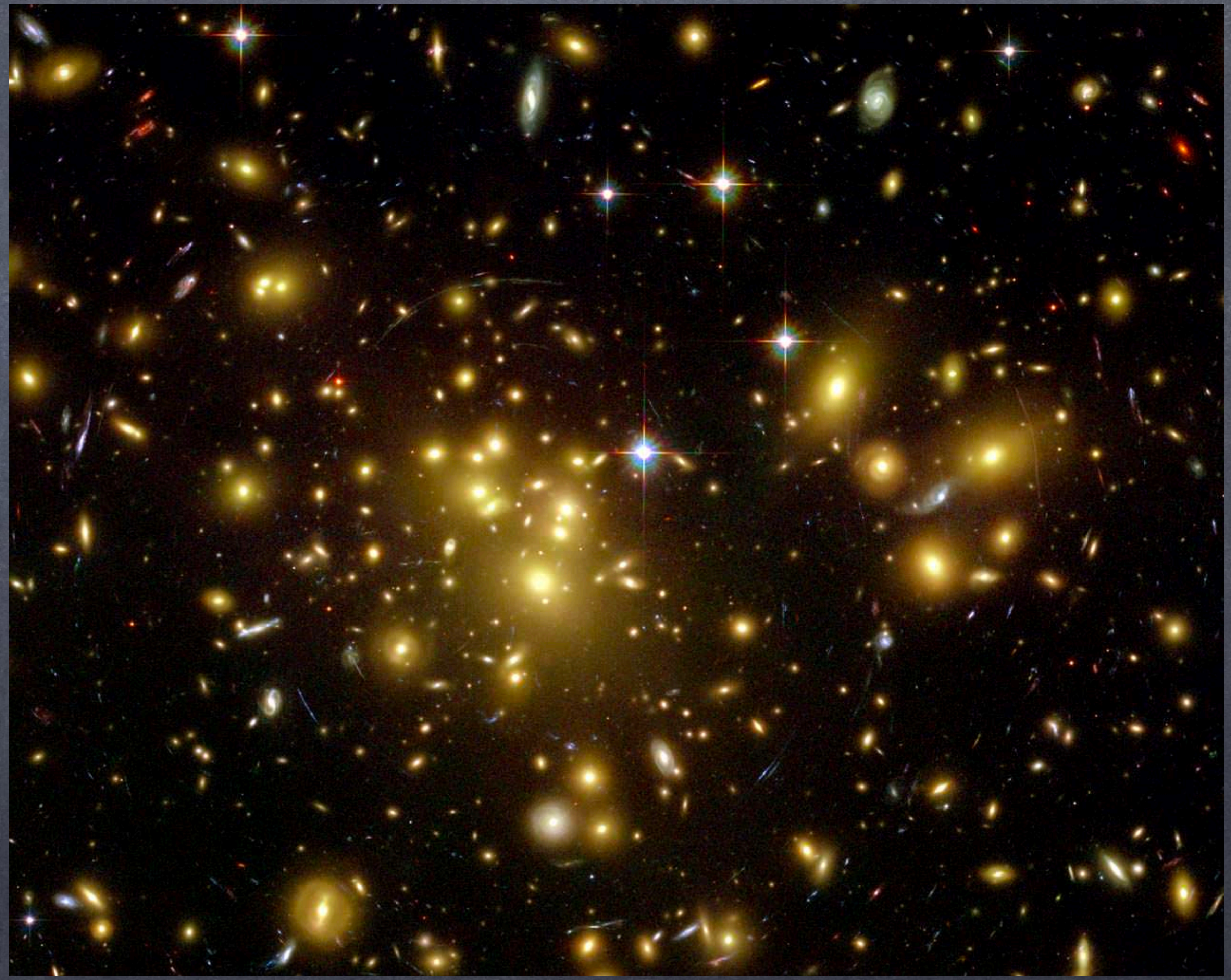
- These scales are already measured in surveys, often at the highest signal-to-noise ratio.



Rimes et al, 2005

Small scales: The challenge

- Density contrast $\delta \gtrsim 1$, so perturbation techniques are not applicable.
- Have to use numerical techniques.

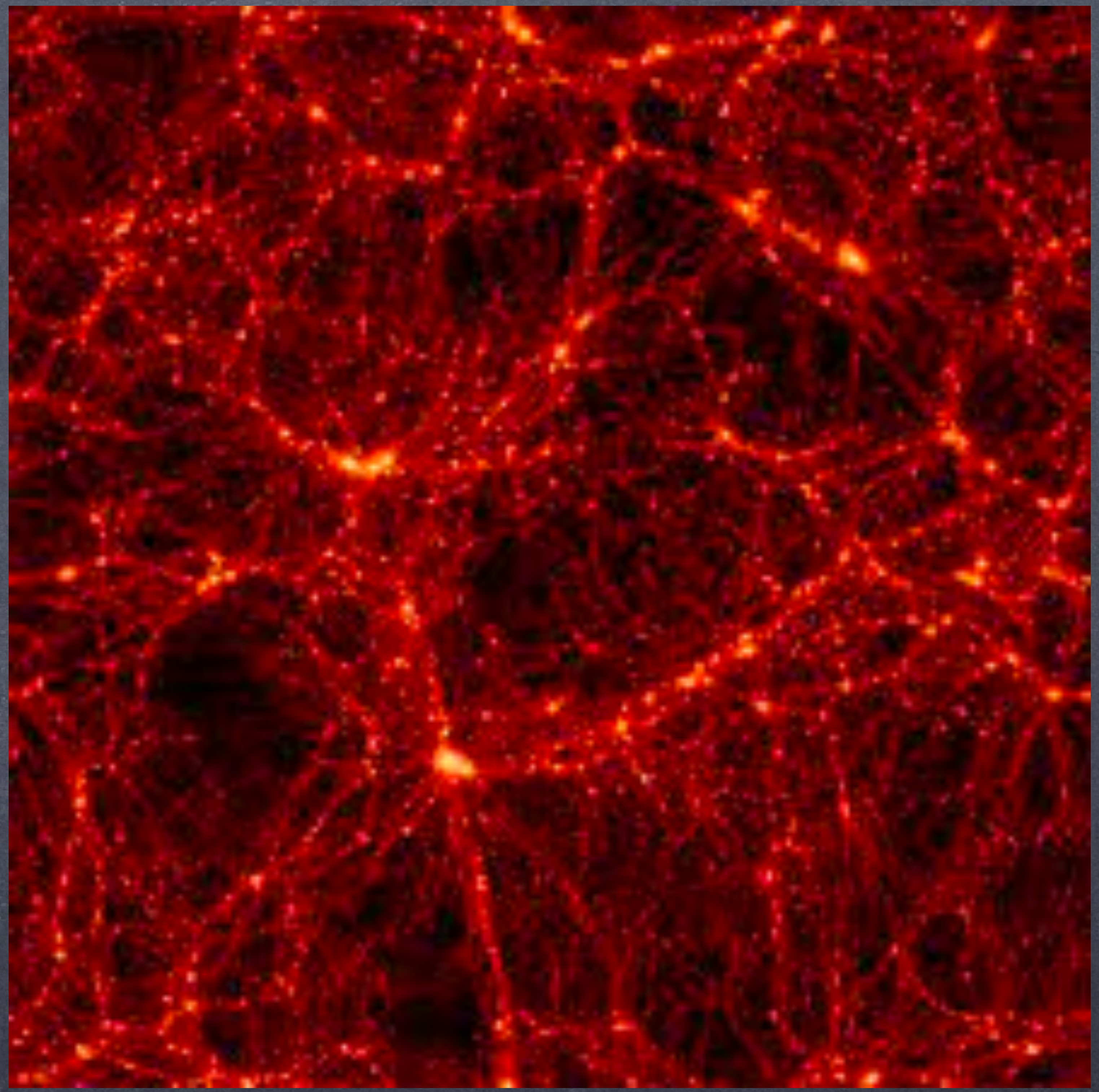


HST

N-body simulations

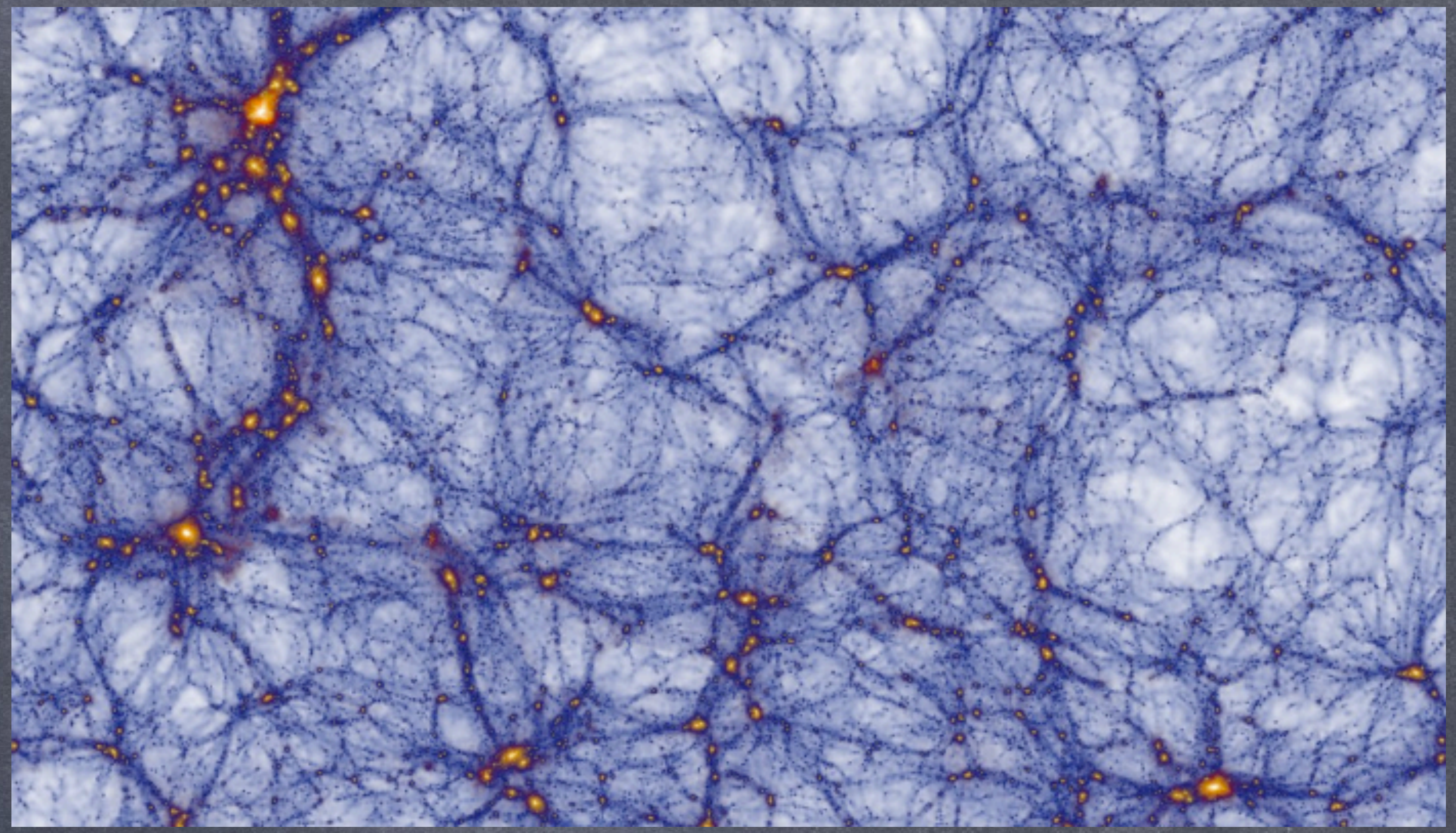
- Generate initial conditions when perturbation theory is still valid, and allow the system to evolve under its own gravity. More particles in a given volume ⇒ higher resolution.

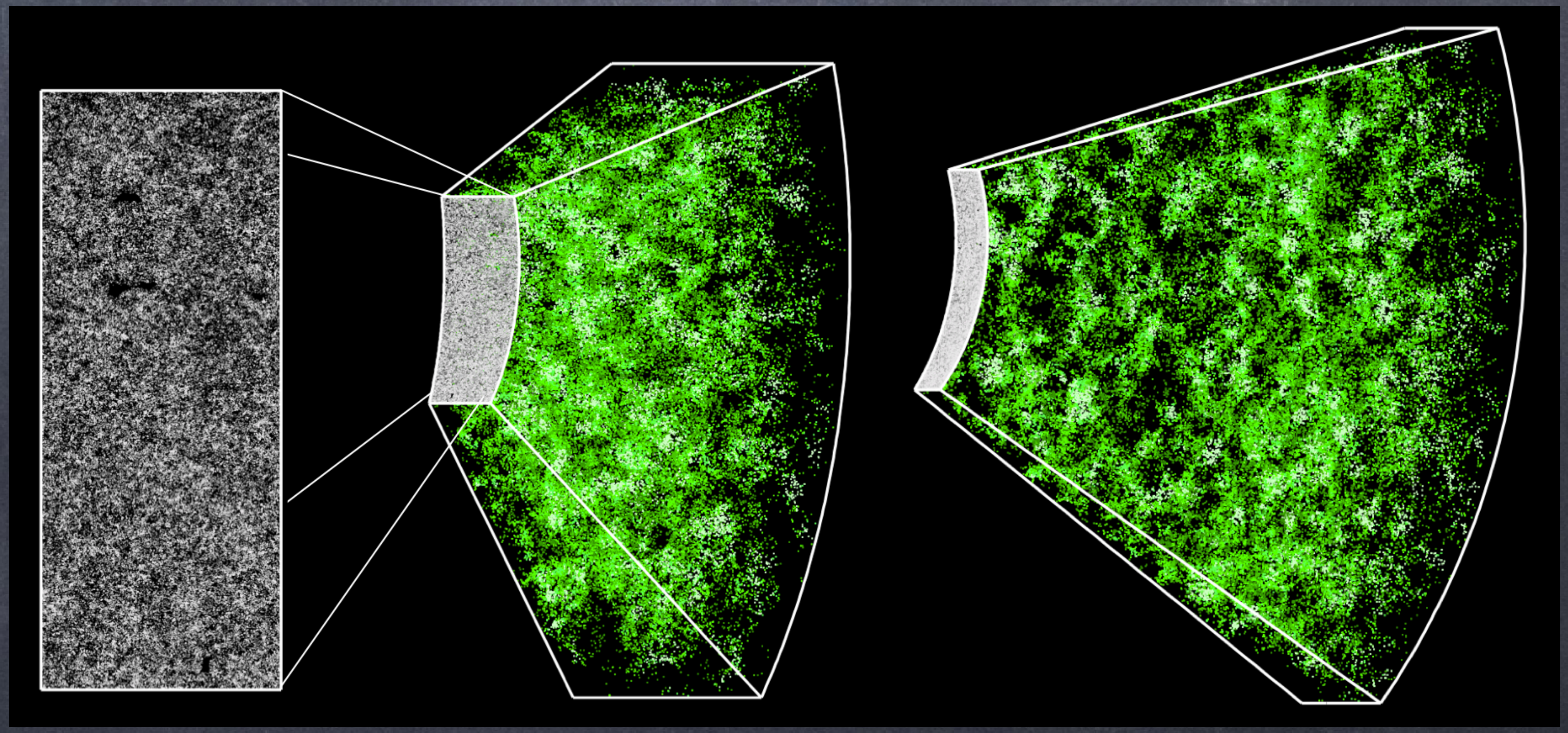
- Naively, such a computation scales as N^2. However, techniques have been developed to allow for a much shallower scaling $\sim N \log N$.



How do we quantify "structure"?

- Need to characterize the spatial distribution of points, say positions of galaxies, statistically. Need the concept of "summary statistics".
- Changing cosmology will change the clustering of data, and therefore the summary.
- More powerful summary statistics will capture more information about the underlying distribution.





Comparing data and theoretical predictions: 2-point functions

$$\delta(\vec{x}) = \frac{\rho(\vec{x}) - \bar{\rho}}{\bar{\rho}}$$

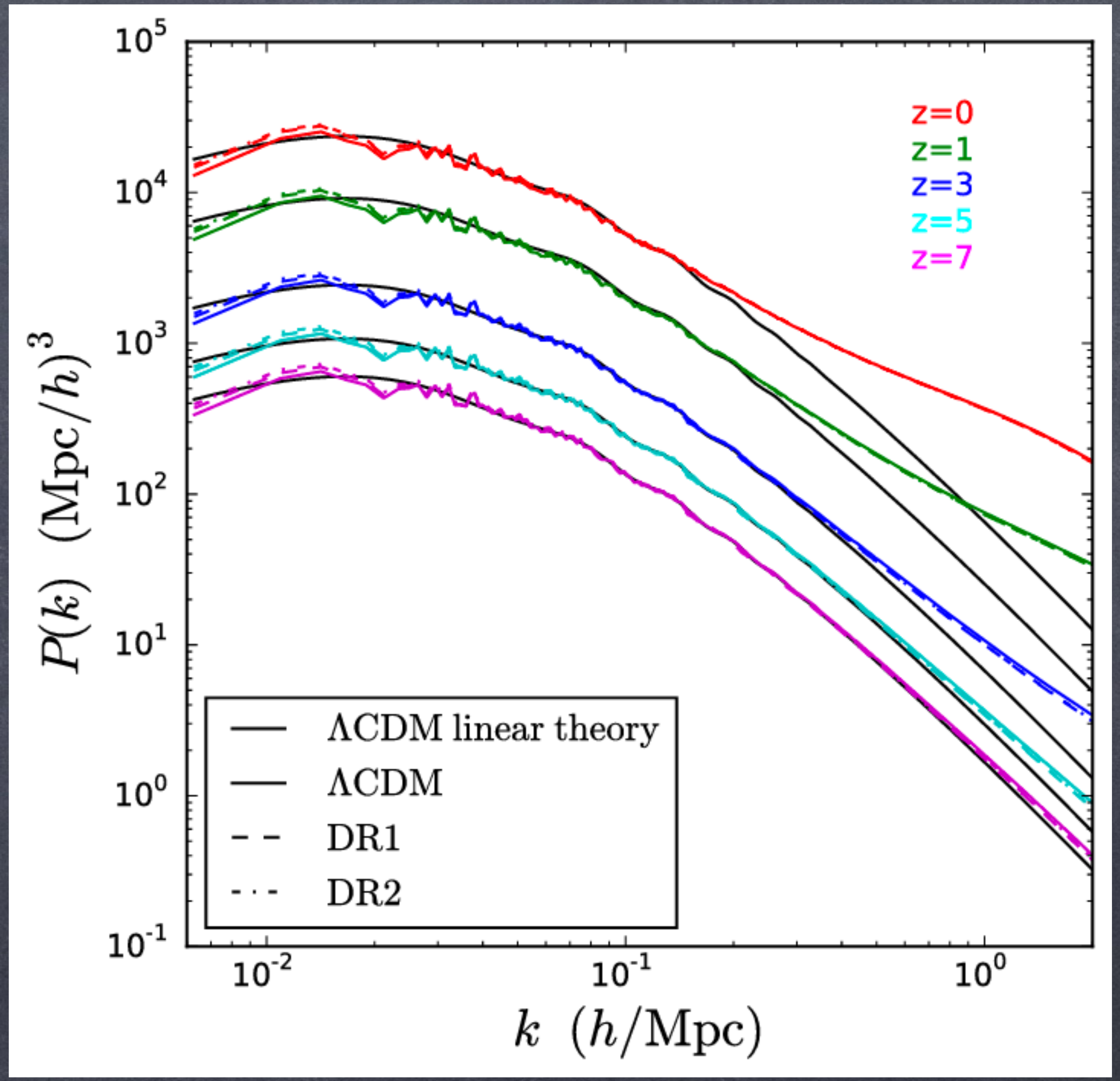
- The most widely used statistical measure in cosmology is the power spectrum $P(k)$, or its Fourier transform $\xi(r)$.

$$\xi(r) = \langle \delta(\vec{x})\delta(\vec{x} + \vec{r}) \rangle_{x,|\vec{r}|=r}$$

$$P(k)\delta^3(\vec{k} - \vec{k}') = \frac{1}{(2\pi)^3}\langle \delta(\vec{k})\delta(\vec{k}') \rangle$$

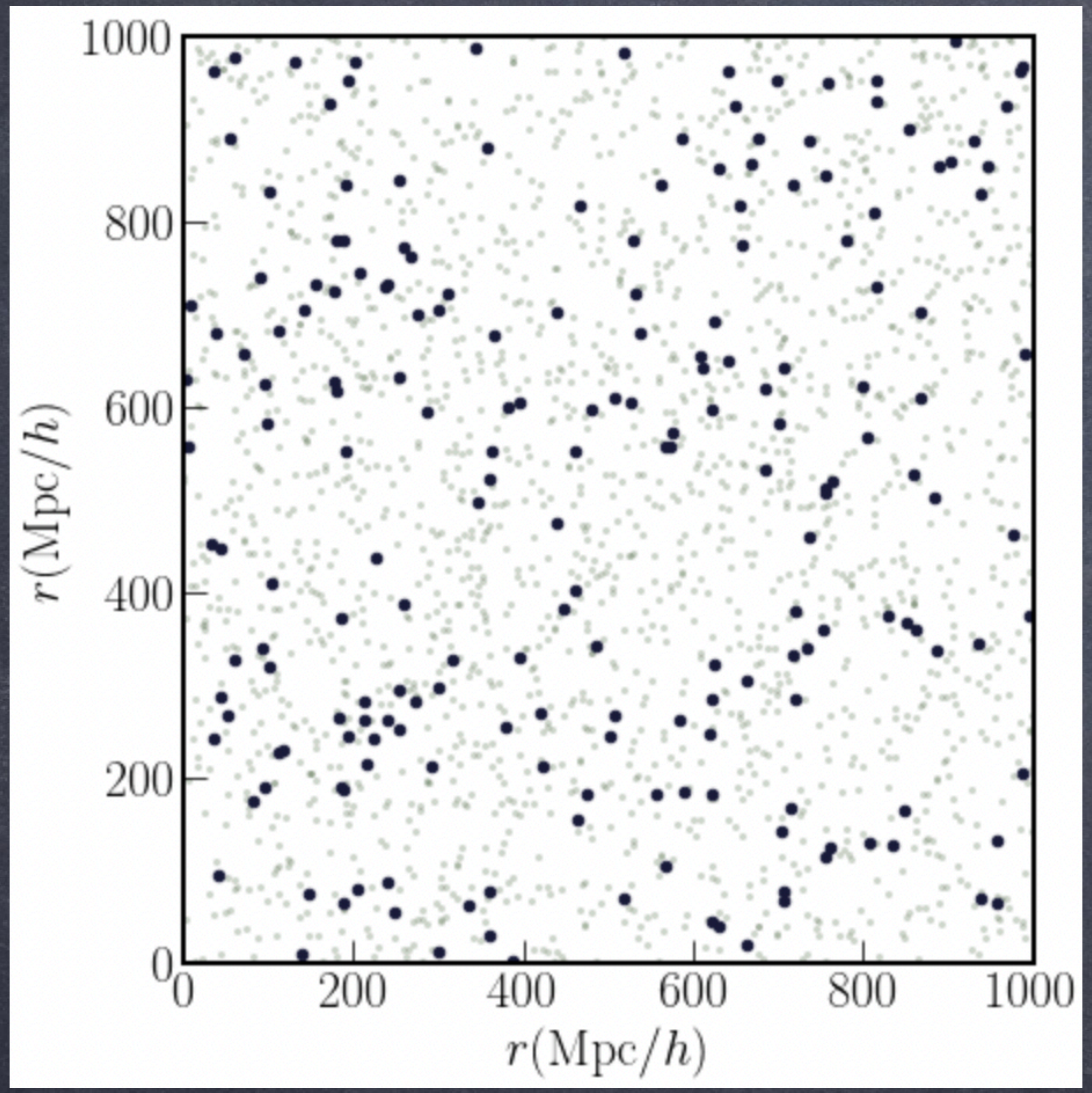
Comparing data and theoretical predictions: 2-point functions

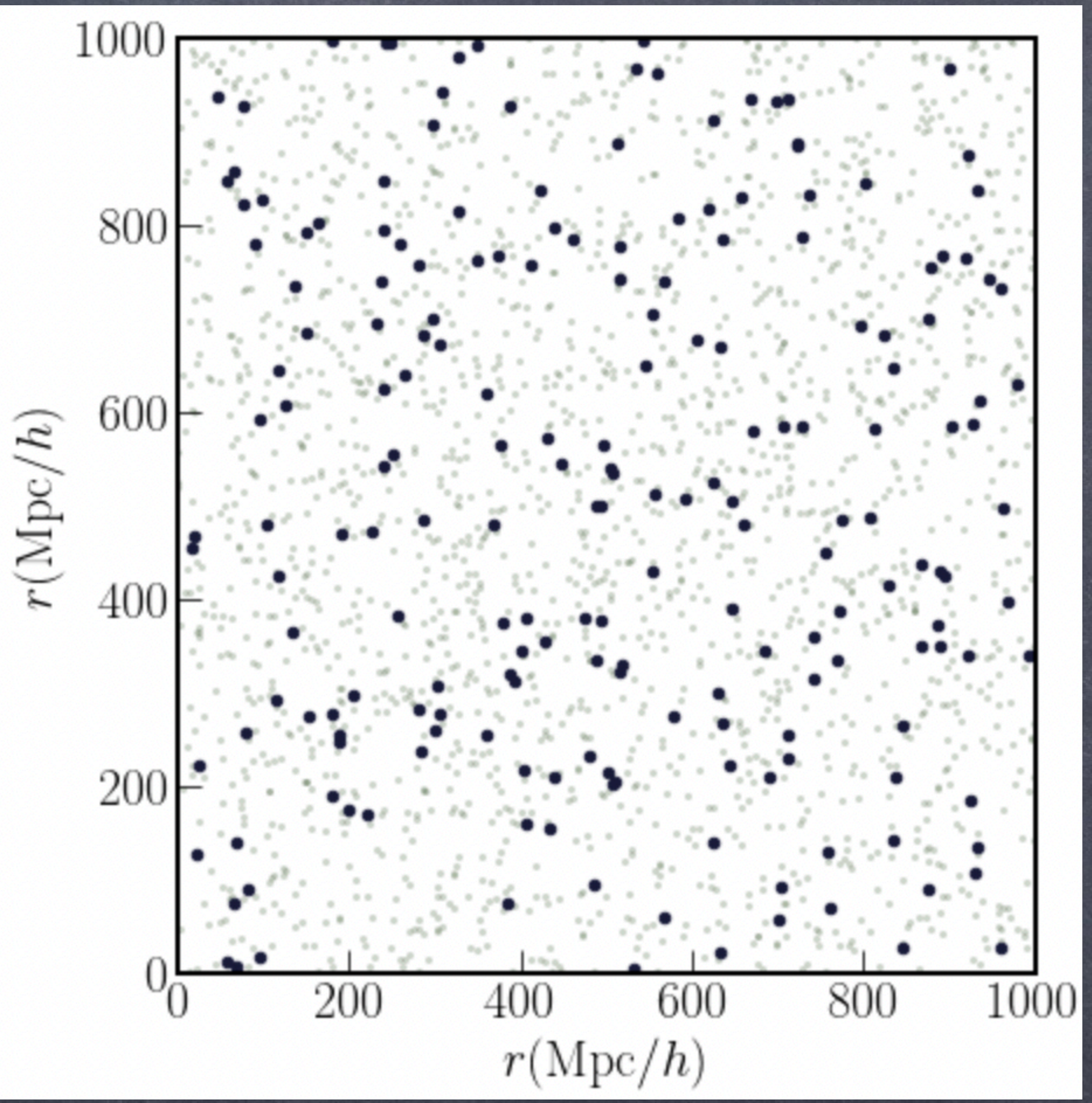
- The most widely used statistical measure in cosmology is the power spectrum $P(k)$, or its Fourier transform $\xi(r)$.



Baugh et al 2015

2-point functions for discrete tracers

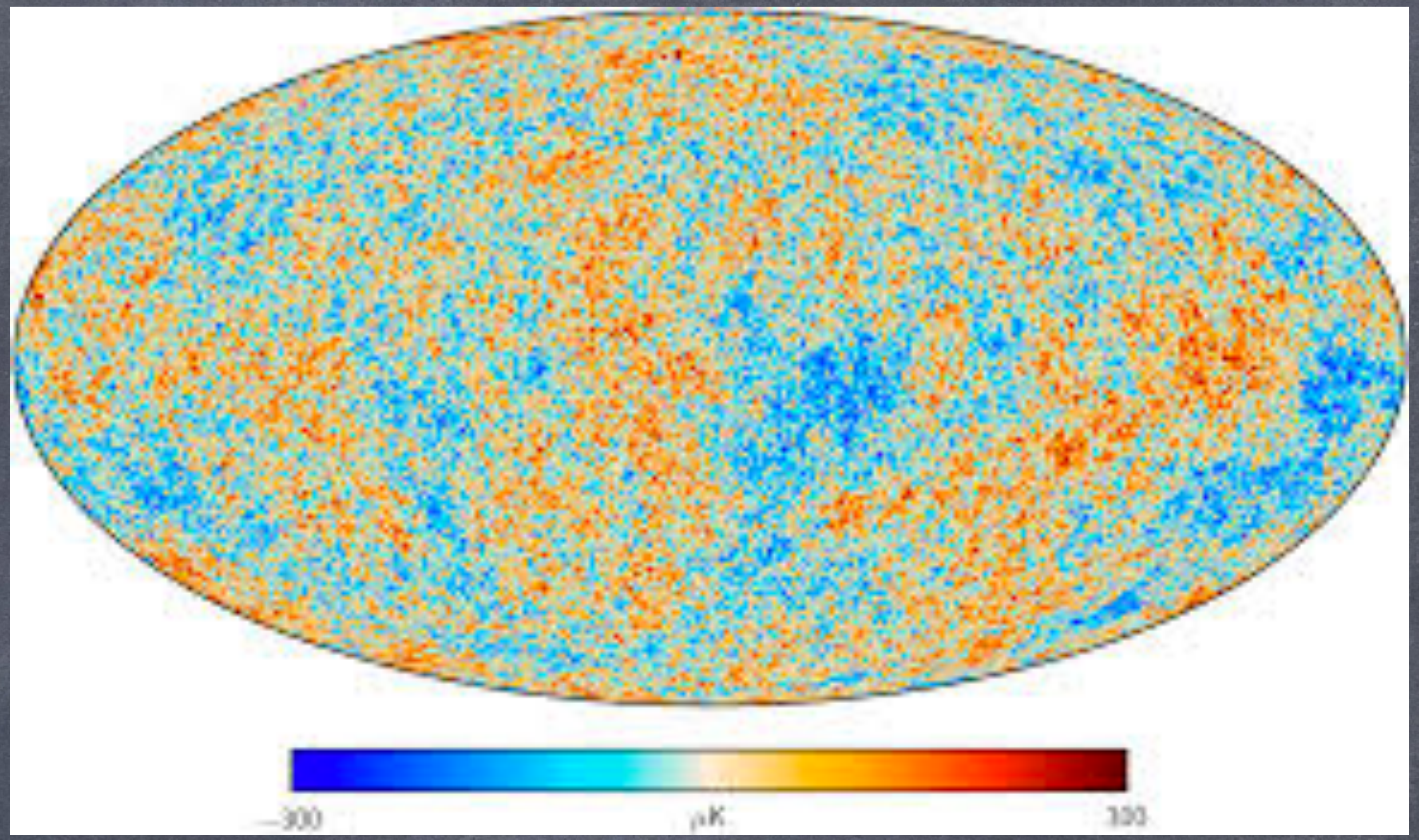




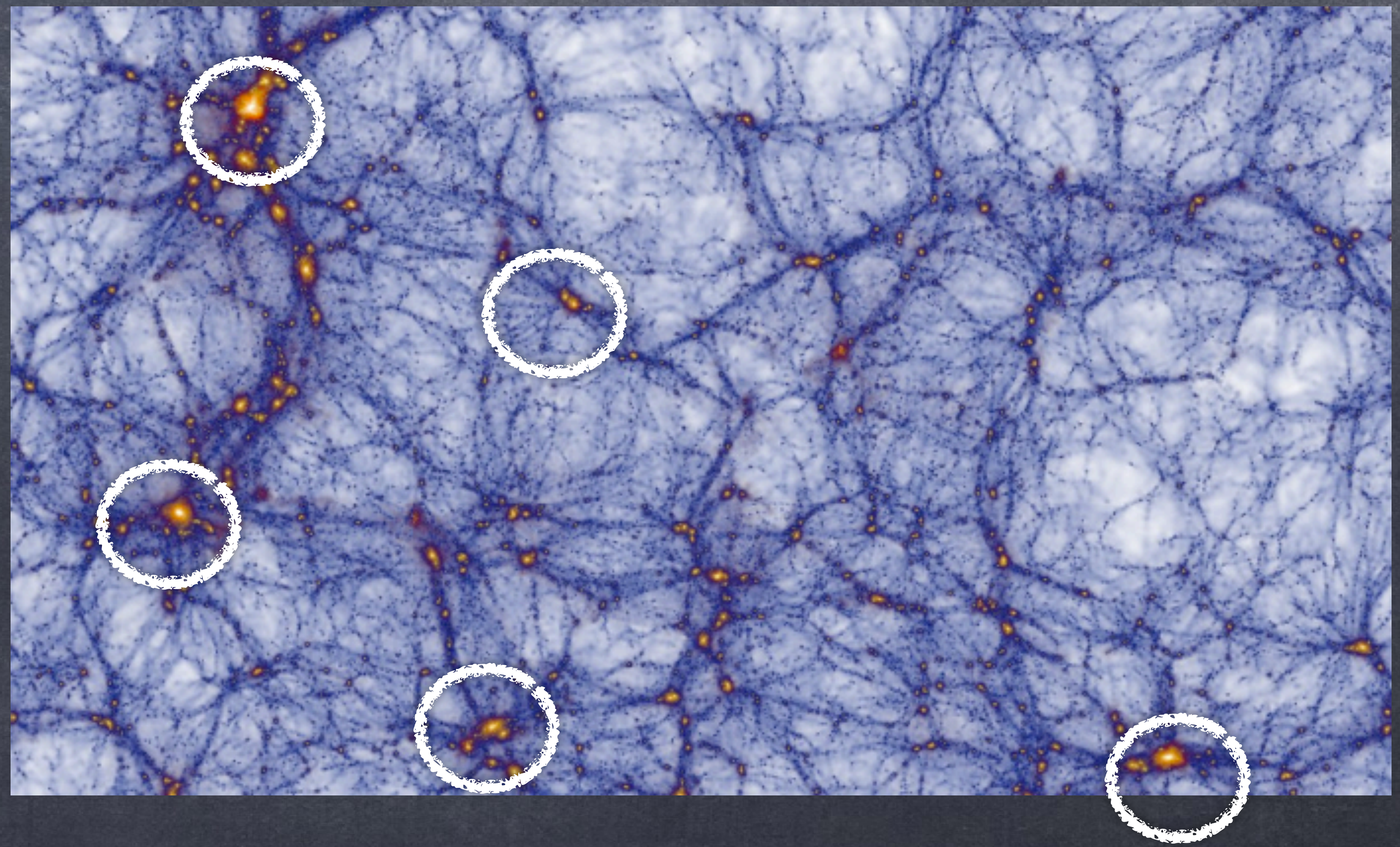
$$\xi(r) = \frac{\langle DD(r) \rangle}{\langle RR(r) \rangle} - 1$$

2-point correlations

- The power spectrum, or the 2pt correlation function is the complete summary statistic of a gaussian random field.
- Does not capture all the information when the density field becomes non-Gaussian.
- To make full use of information on small scales, we need to explore statistics beyond the 2-pt functions.

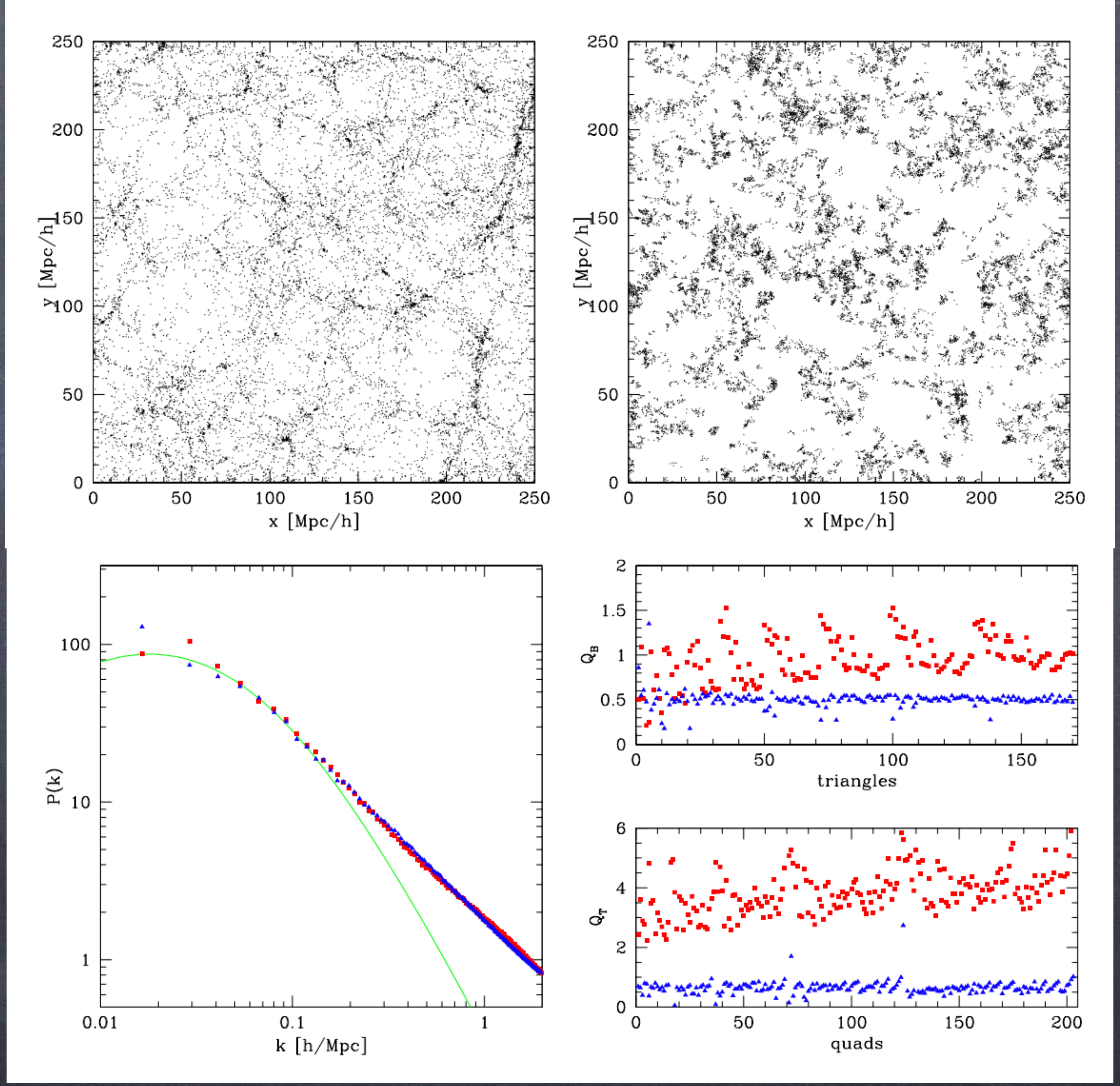


Planck, 2018



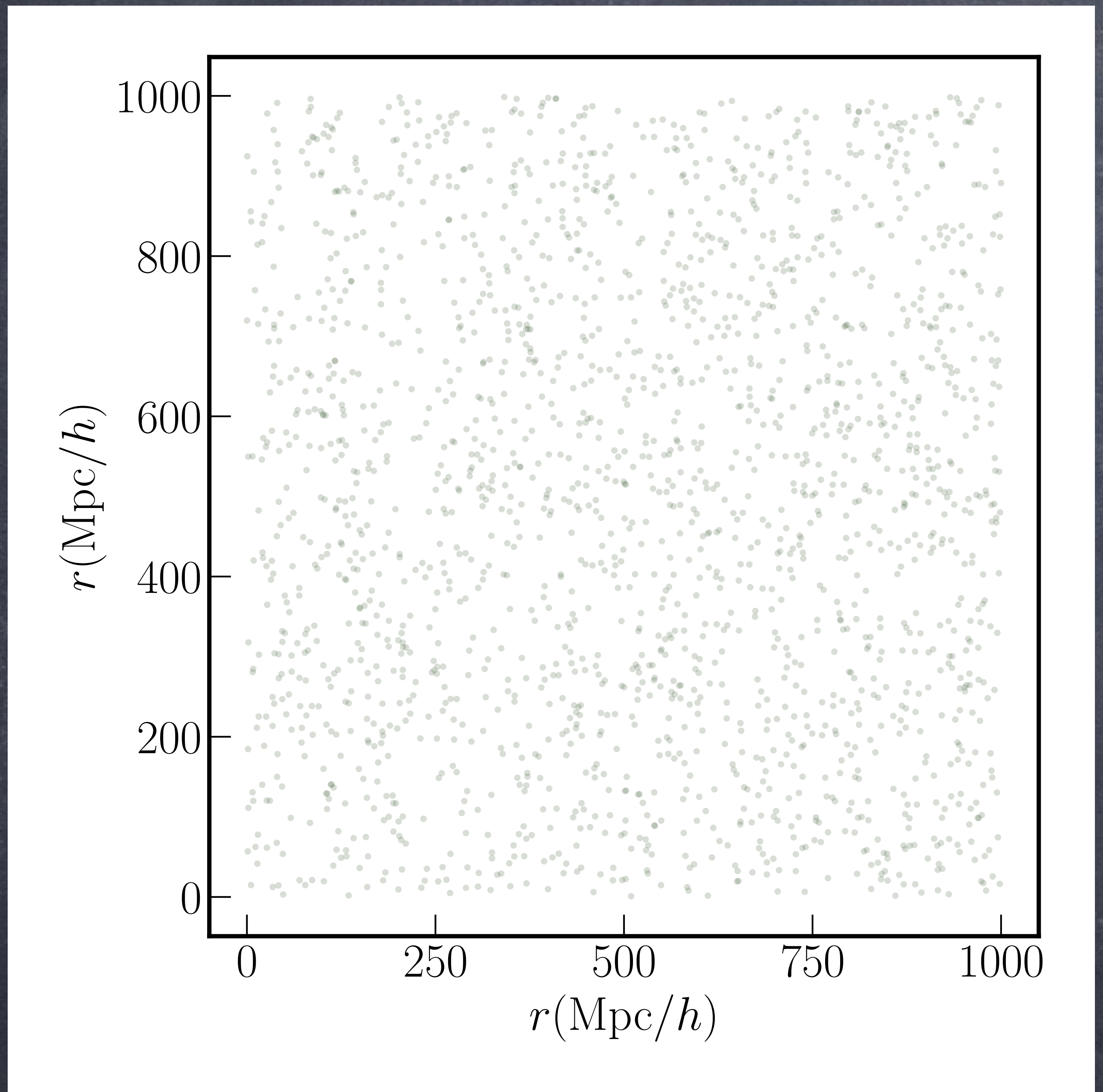
Beyond the 2PCF: Higher order N-point correlations

- Consider higher N-point correlation functions. The 3PCF (bispectrum) already has a lot of extra information, but computationally expensive to compute.
- Becomes computationally prohibitive as we generalize to higher N-PCF.



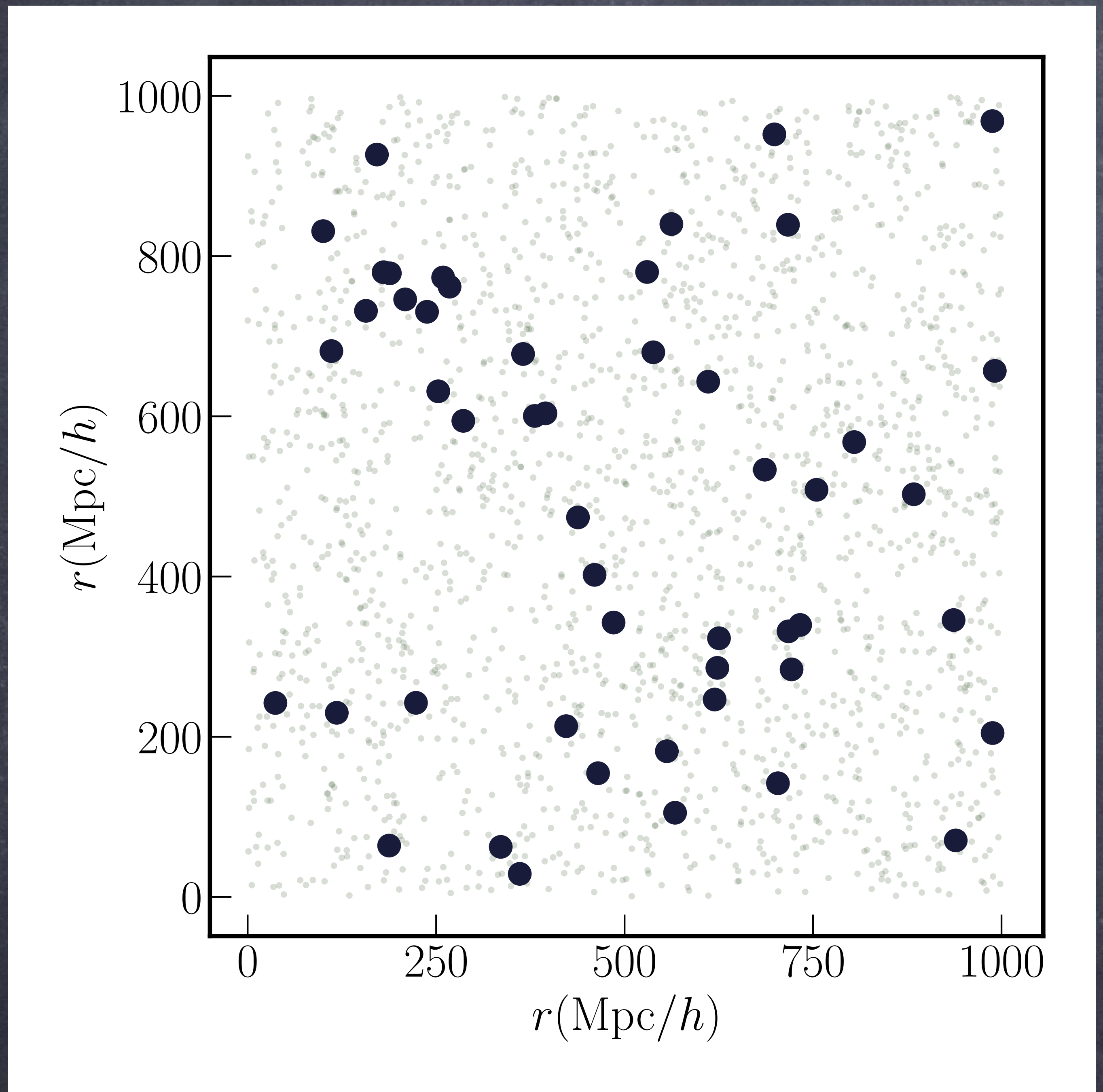
Sefusatti and Scoccimarro, 2004

Using a new statistical measure for discrete data: k Nearest Neighbor distributions



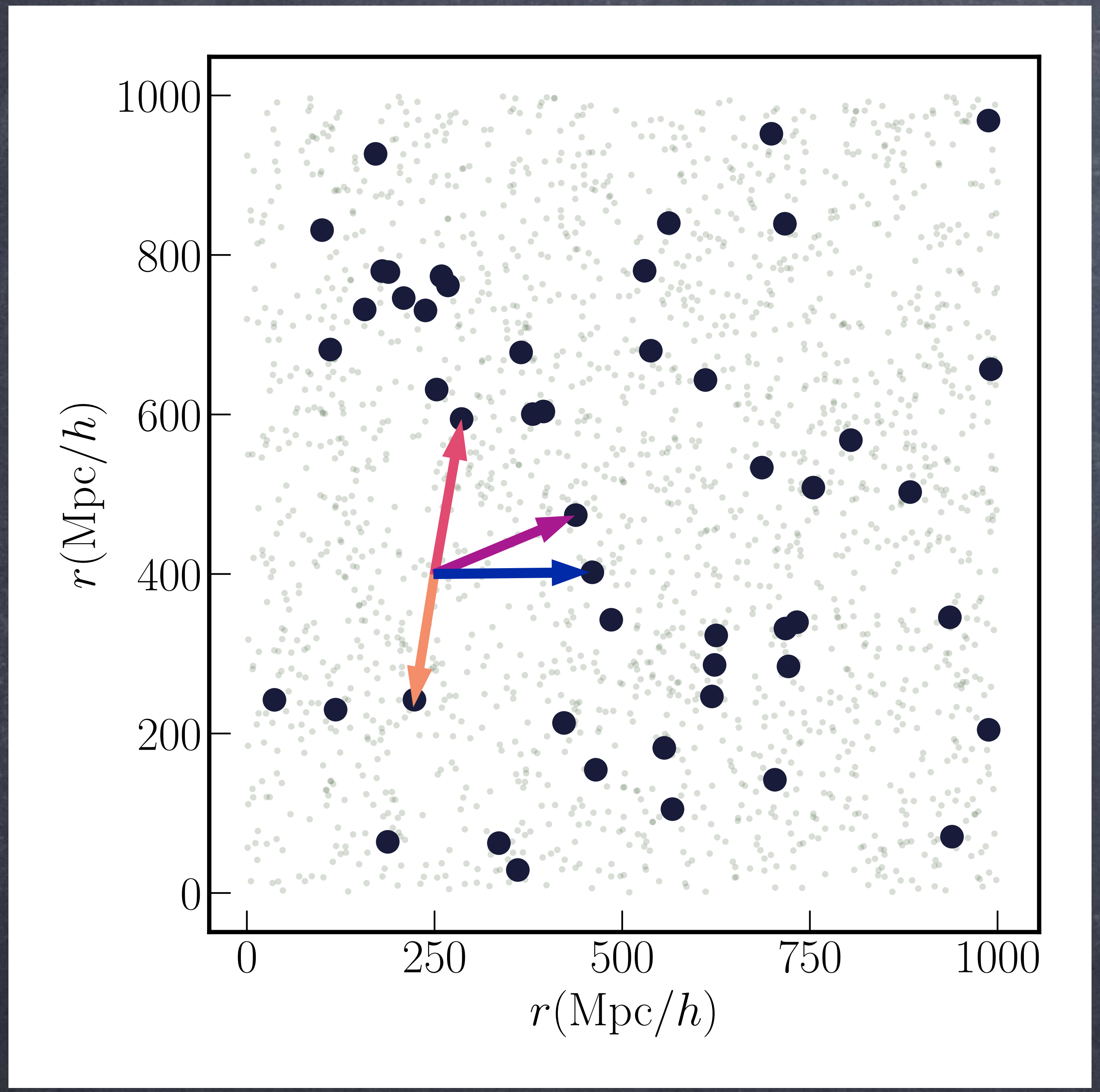
- Sample the volume densely with a set of query points.

Using a new statistical measure for discrete data: k Nearest Neighbor distributions



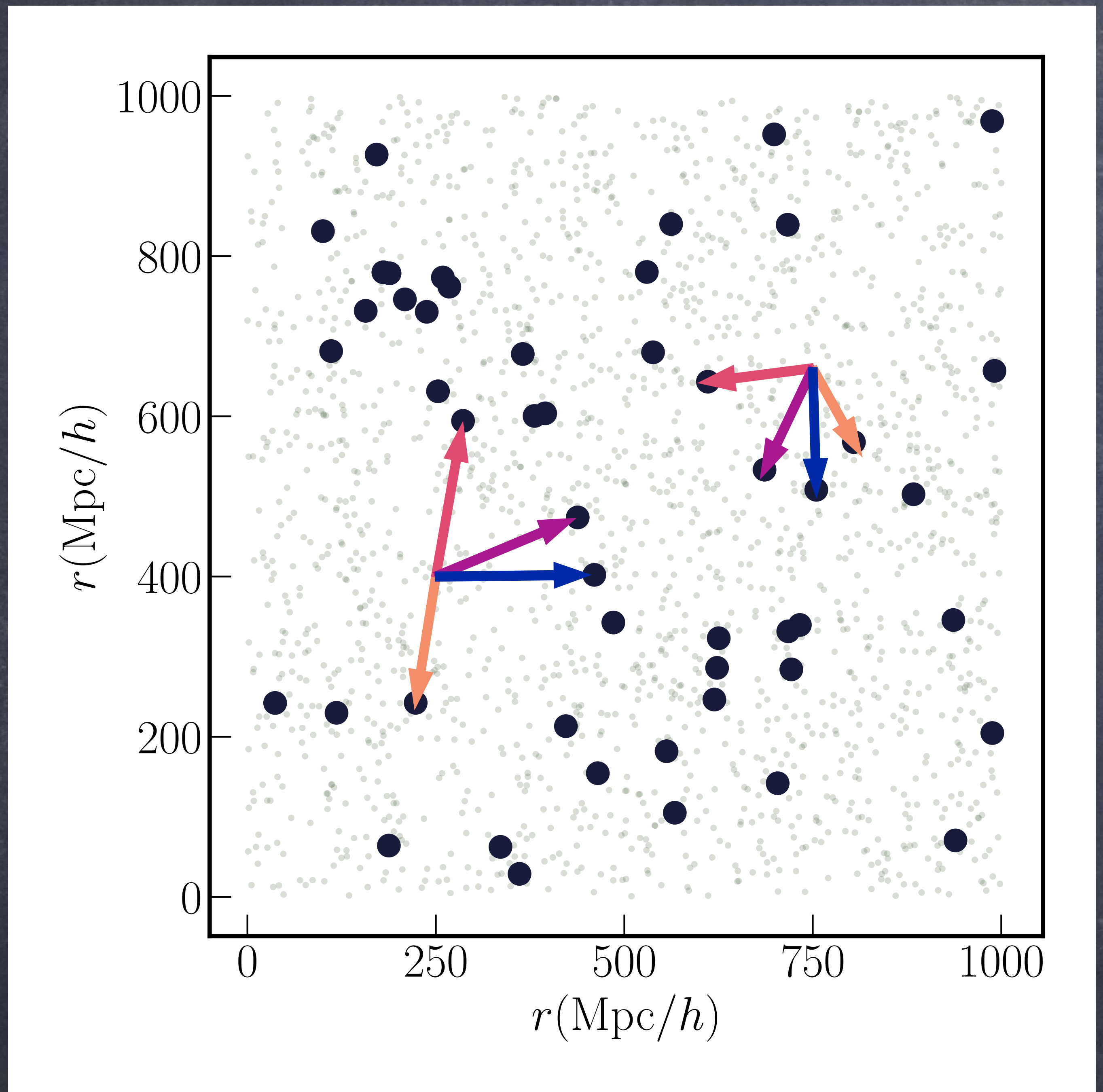
- Sample the volume densely with a set of query points.

Using a new statistical measure for discrete data: k Nearest Neighbor distributions



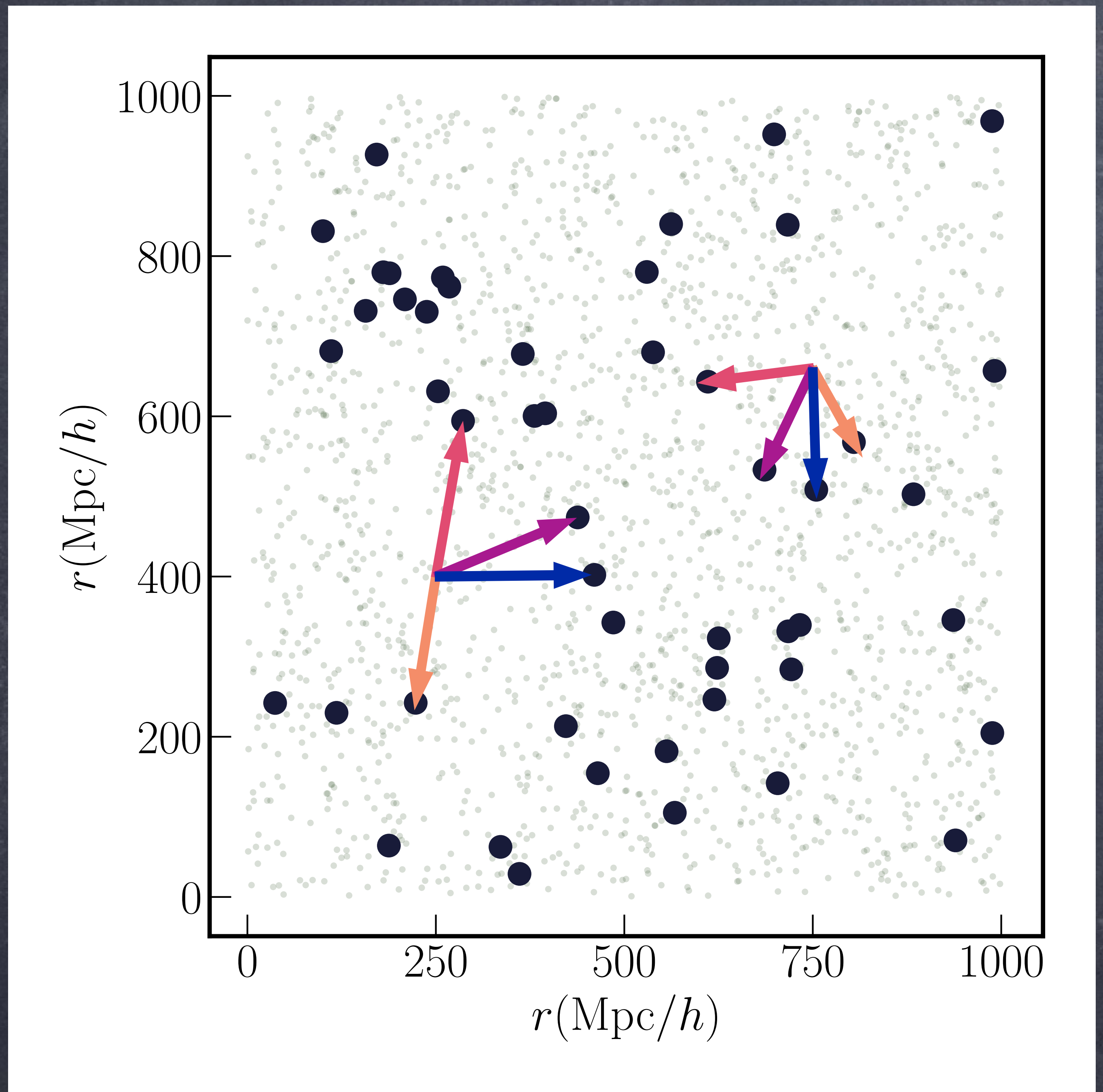
- Sample the volume densely with a set of query points.
- For each of the query points, use a tree structure to efficiently find the distance to the 1st, 2nd, … k-th nearest neighbor data points.

Using a new statistical measure for discrete data: k Nearest Neighbor distributions

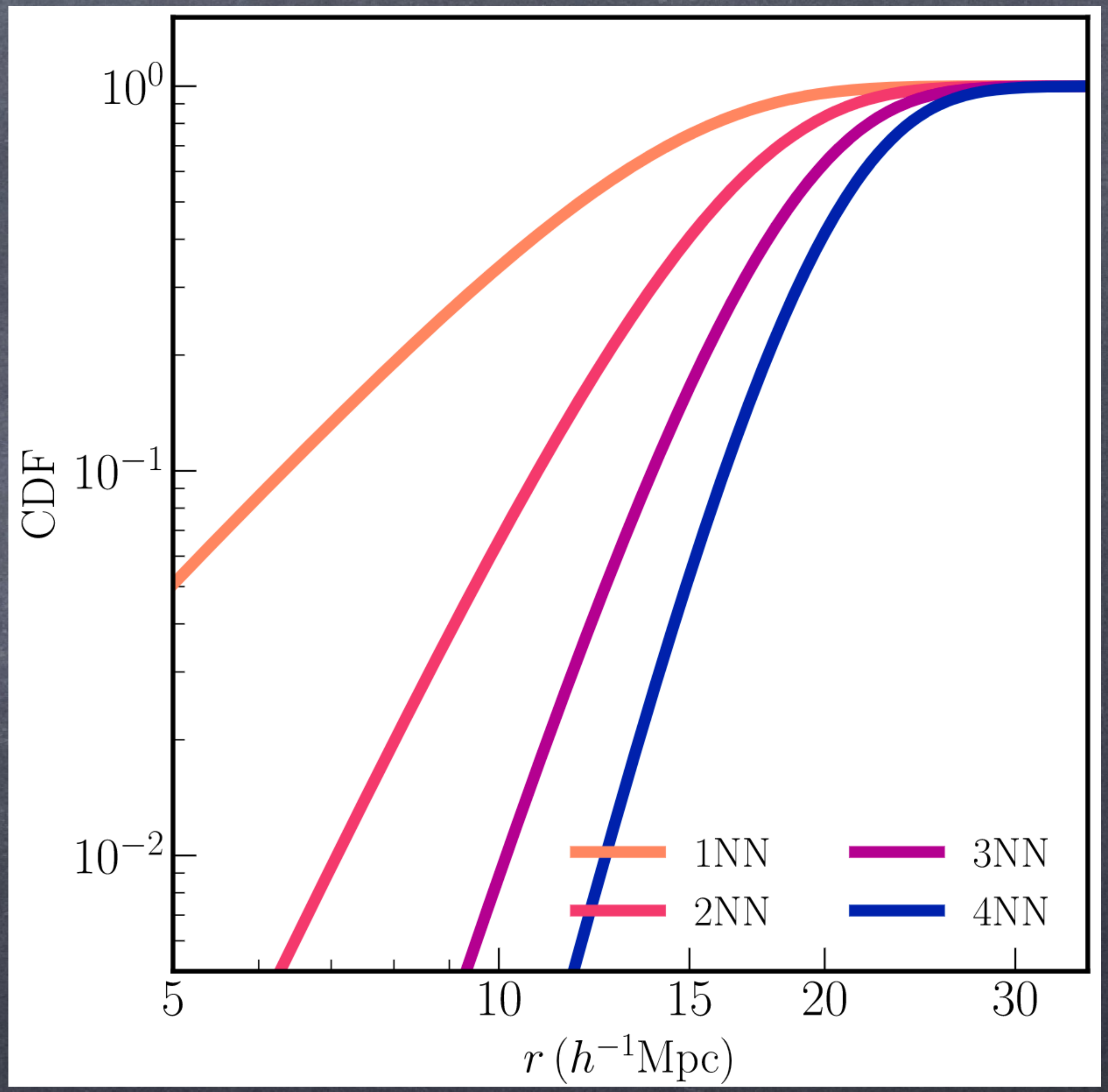


- Sample the volume densely with a set of query points.
- For each of the query points, use a tree structure to efficiently find the distance to the 1st, 2nd, ... k-th nearest neighbor data points.

Using a new statistical measure for discrete data: k Nearest Neighbor distributions



- Sample the volume densely with a set of query points.
- For each of the query points, use a tree structure to efficiently find the distance to the 1st, 2nd, … k-th nearest neighbor data points.
- For a given k, sort the distances to get the Empirical CDF of the distances.
- Takeaway: a) A single measurement procedure is sufficient for a range of scales. b) Not computationally expensive to measure higher k distributions. (~20 seconds on a single core)



Small scales ← r (Mpc)

What do the kNN distributions measure?

- The measurement can be connected to cumulative counts: $\mathrm{CDF}_{1\mathrm{NN}}(R) = \mathscr{P}_{>0}(V)\,|_{V=4/3\pi R^3}$
- The generating function for the distributions can be written in terms of integrals over all (connected) N-point correlations in the data:

$$P(z, V) = \frac{1 - \exp\left[\sum_{k=1}^{\infty} \frac{\bar{n}^k (z-1)^k}{k!} \xi^{(k)}(V)\right]}{1 - z} \qquad \xi^{(k)}(V) = \int_V \ldots \int_V d^3\mathbf{r}_1 \ldots \mathbf{r}_k \xi^C(\mathbf{r}_1, \ldots, \mathbf{r}_k)$$

- Each kNN-CDF measures a different combination of the N-point correlation functions:
  - $$\mathrm{CDF}_{1\mathrm{NN}}(V) = 1 - \exp\left[\sum_{k=1}^{\infty} \frac{(-\bar{n})^k}{k!} \xi^{(k)}(V)\right]$$
  - $$\mathrm{CDF}_{2\mathrm{NN}}(V) = 1 - \exp\left[\sum_{k=1}^{\infty} \frac{(-\bar{n})^k}{k!} \xi^{(k)}(V)\right] - \left(\frac{(-\bar{n})^{(k-1)}}{(k-1)!} \xi^{(k)}(V)\right) \exp\left[\sum_{k=1}^{\infty} \frac{(-\bar{n})^k}{k!} \xi^{(k)}(V)\right]$$

What do the kNN distributions measure?

- The measurement can be connected to cumulative counts: $\mathrm{CDF}_{1\mathrm{NN}}(R) = \mathscr{P}_{>0}(V)\,|_{V=4/3\pi R^3}$

- The generating function for the distributions can be written in terms of integrals over all (connected) N-point correlations in the data:

$$P(z,V) = \frac{1 - \exp\left[\sum_{k=1}^{\infty} \frac{\bar{n}^k (z-1)^k}{k!} \xi^{(k)}(V)\right]}{1-z} \qquad \xi^{(k)}(V) = \int_V \ldots \int_V d^3\mathbf{r}_1 \ldots \mathbf{r}_k \xi^C(\mathbf{r}_1, \ldots, \mathbf{r}_k)$$

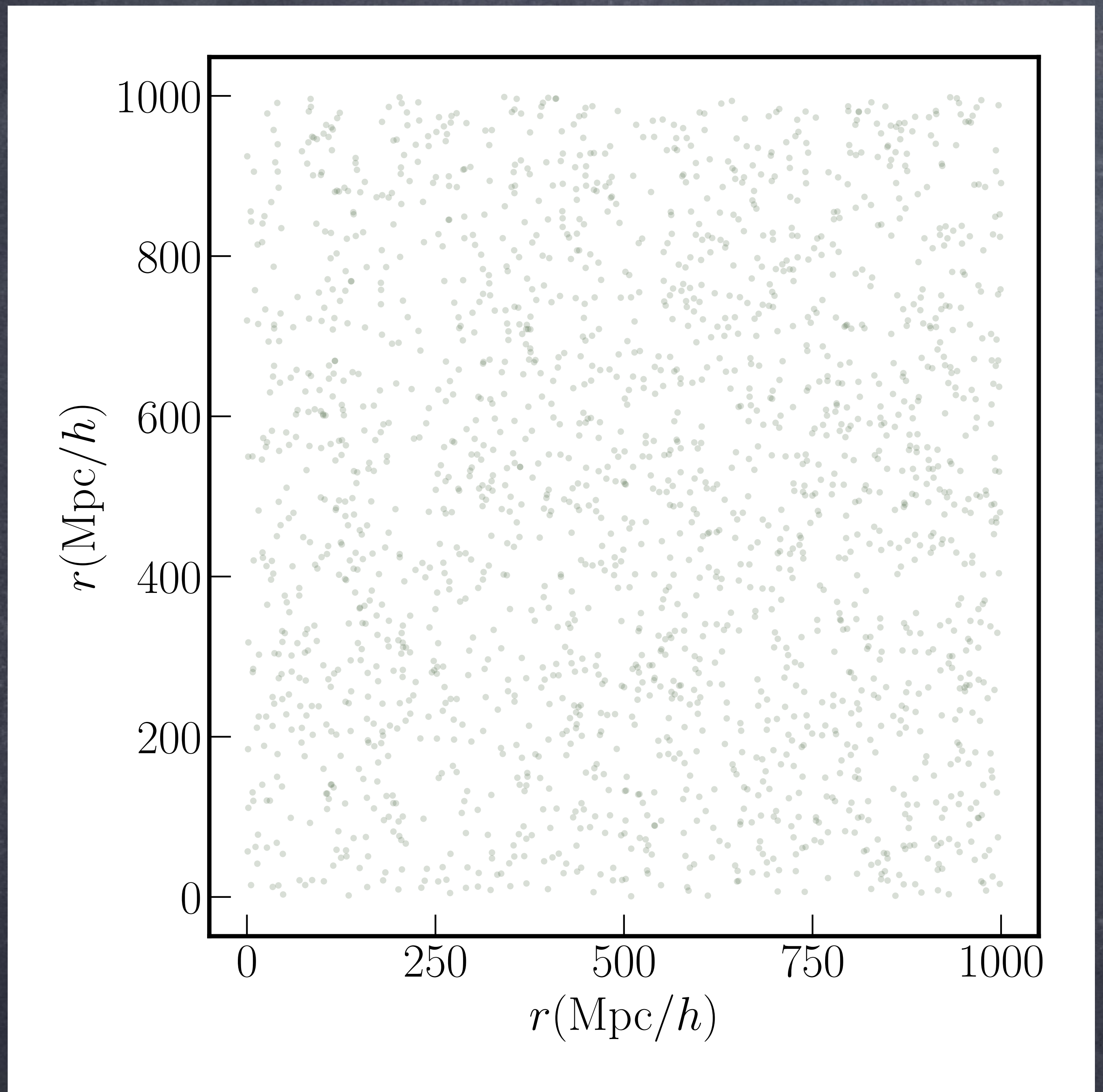
- Each kNN-CDF measures a different 1 point averages of the underlying continuous field smoothed on scale R:

  - $\mathrm{CDF}_{1\mathrm{NN}}(V) = 1 - \left\langle \exp\left[-\bar{n}V\left(1+\delta_R\right)\right]\right\rangle$

  - $\mathrm{CDF}_{2\mathrm{NN}}(V) = 1 - \left\langle \exp\left[-\bar{n}V\left(1+\delta_R\right)\right]\right\rangle - \left\langle \left(\bar{n}V\left(1+\delta_R\right)\right)\exp\left[-\bar{n}V\left(1+\delta_R\right)\right]\right\rangle$

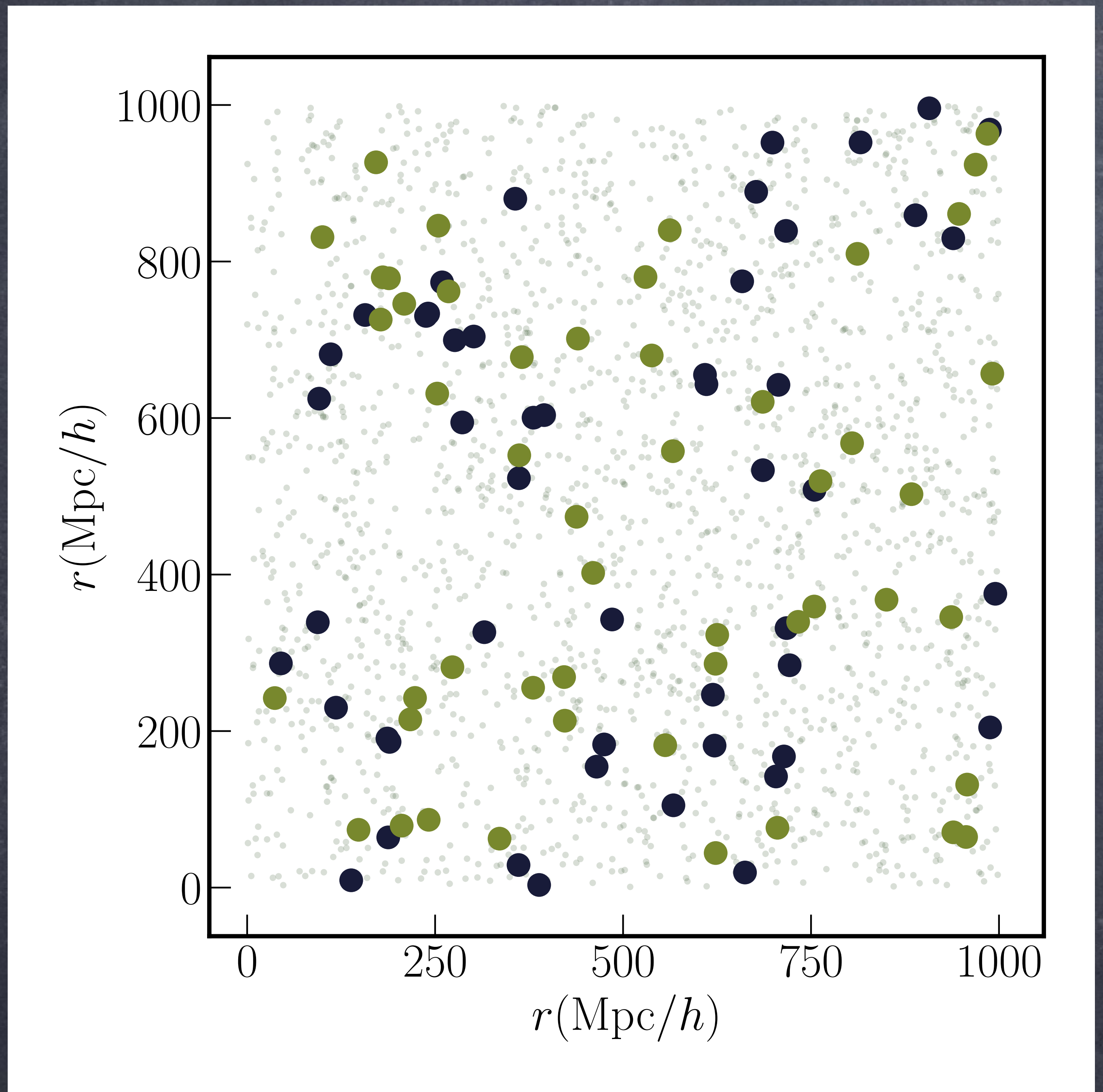
Cross-correlations with nearest neighbors

- Sample the volume densely with a set of random points.

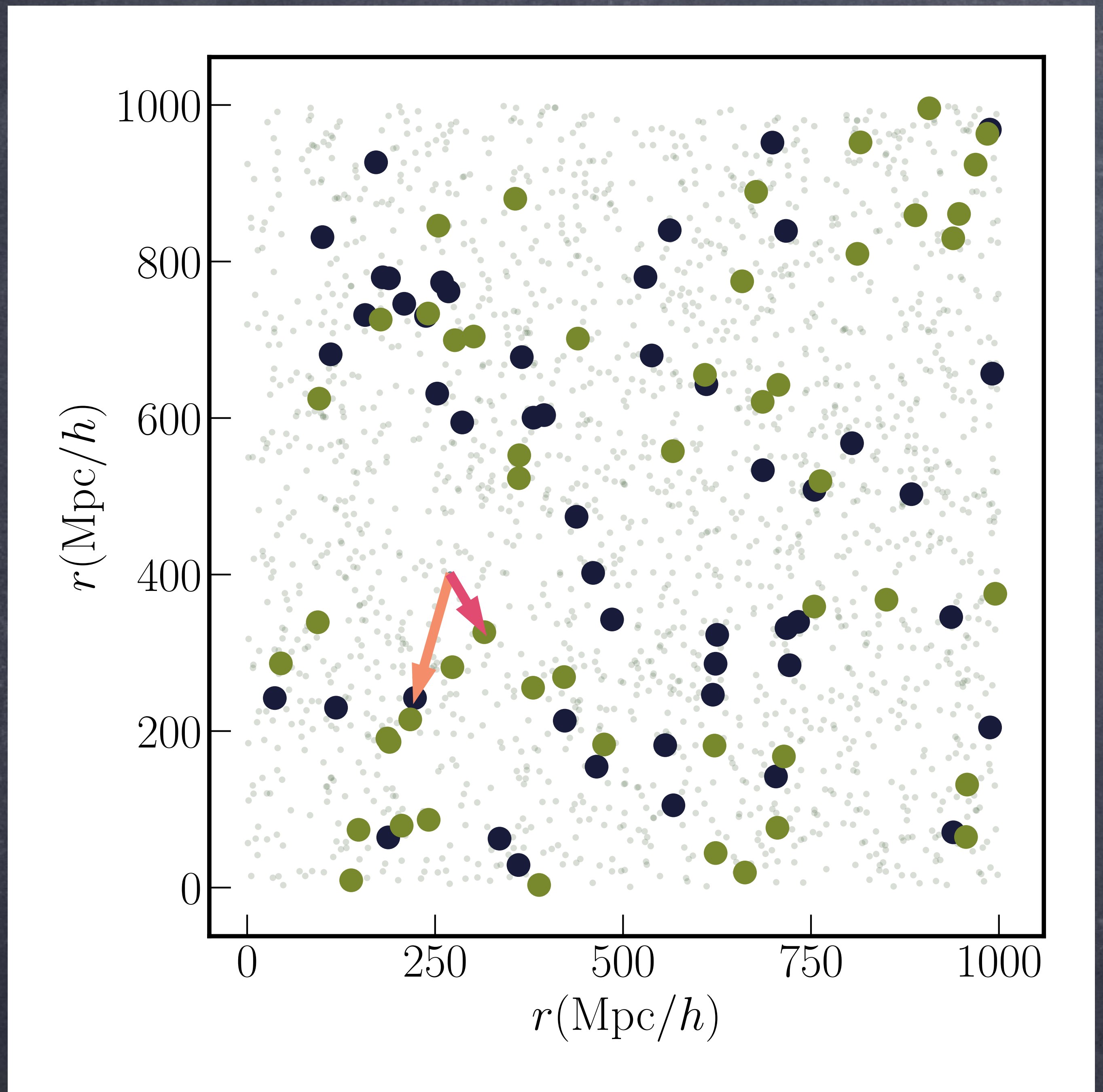


Cross-correlations with nearest neighbors

- Sample the volume densely with a set of query points.

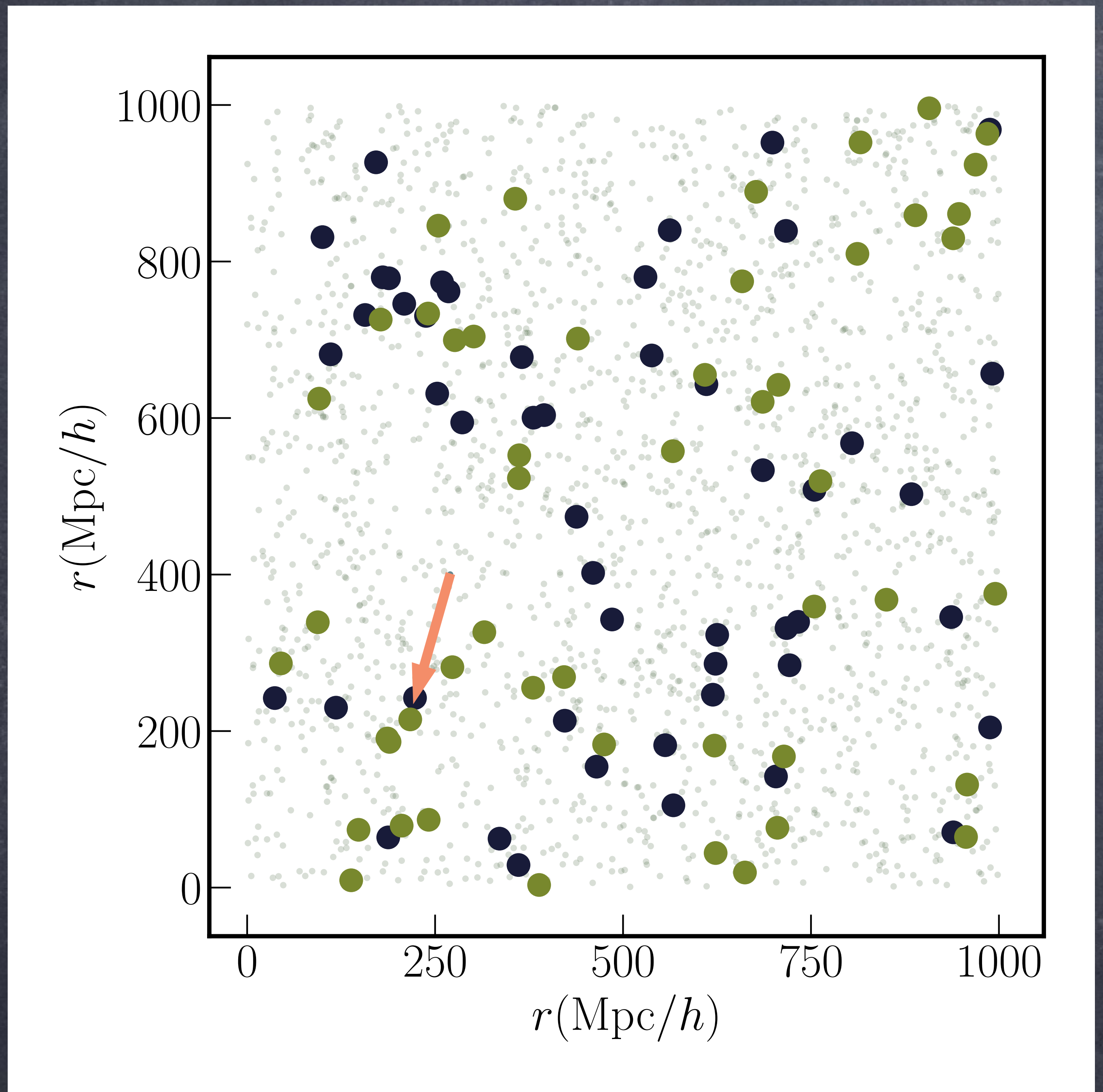


Cross-correlations with nearest neighbors



- Sample the volume densely with a set of query points.
- For each query point, find the distance to the nearest data point of each dataset.
- For each query point, pick the larger distance.

Cross-correlations with nearest neighbors

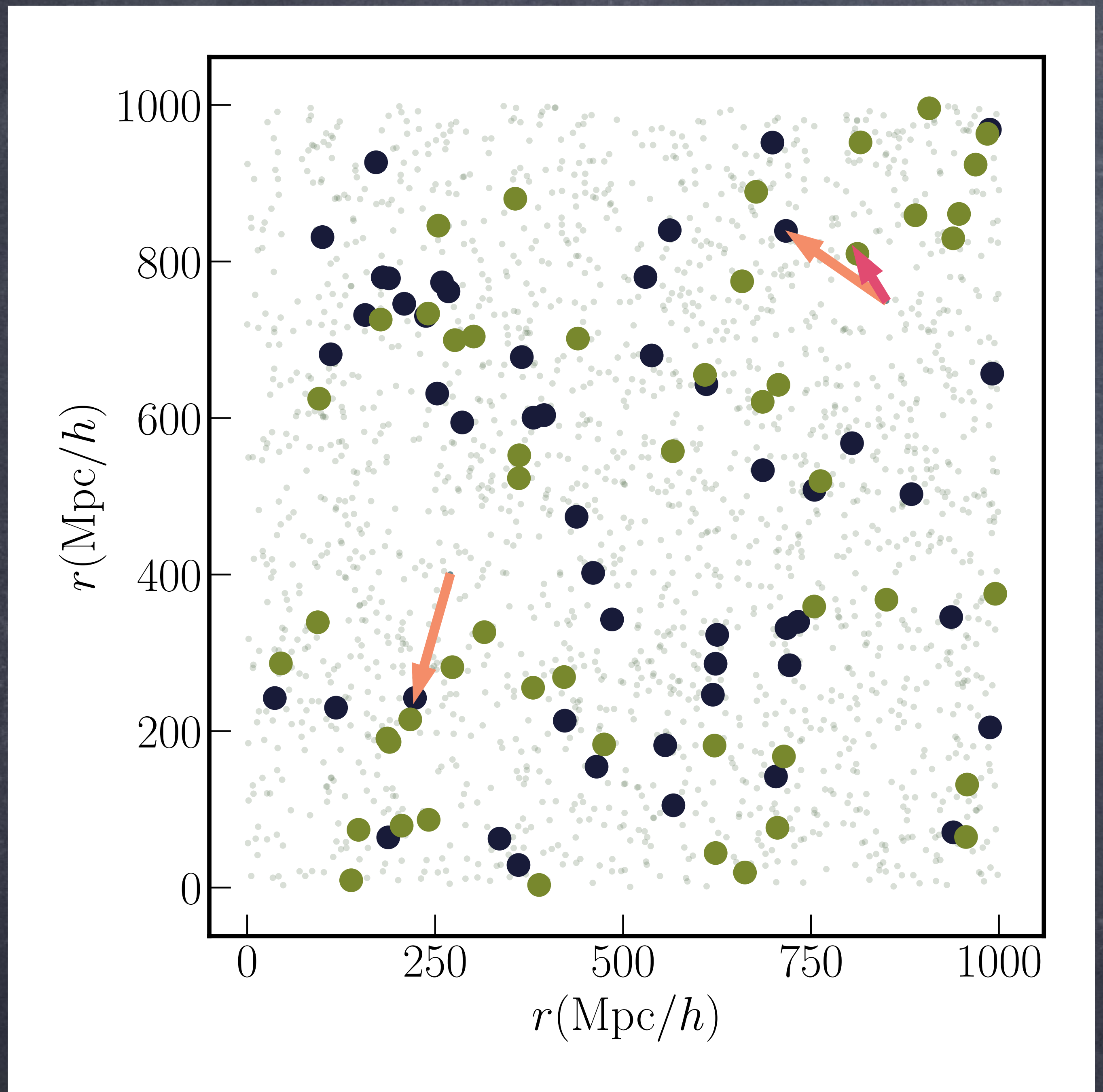


- Sample the volume densely with a set of query points.

- For each query point, find the distance to the nearest data point of each dataset.

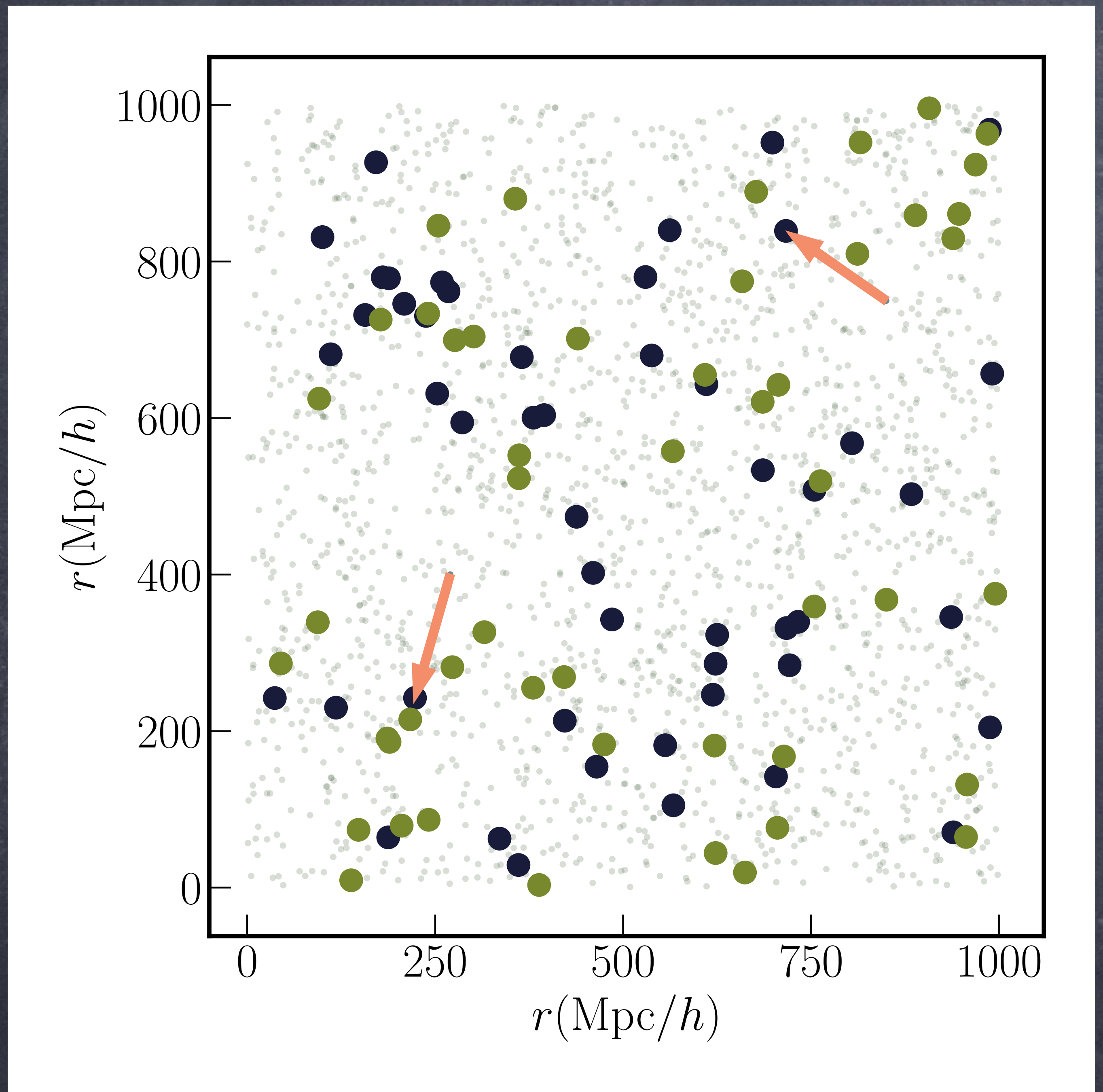
- For each query point, pick the larger distance.

Cross-correlations with nearest neighbors



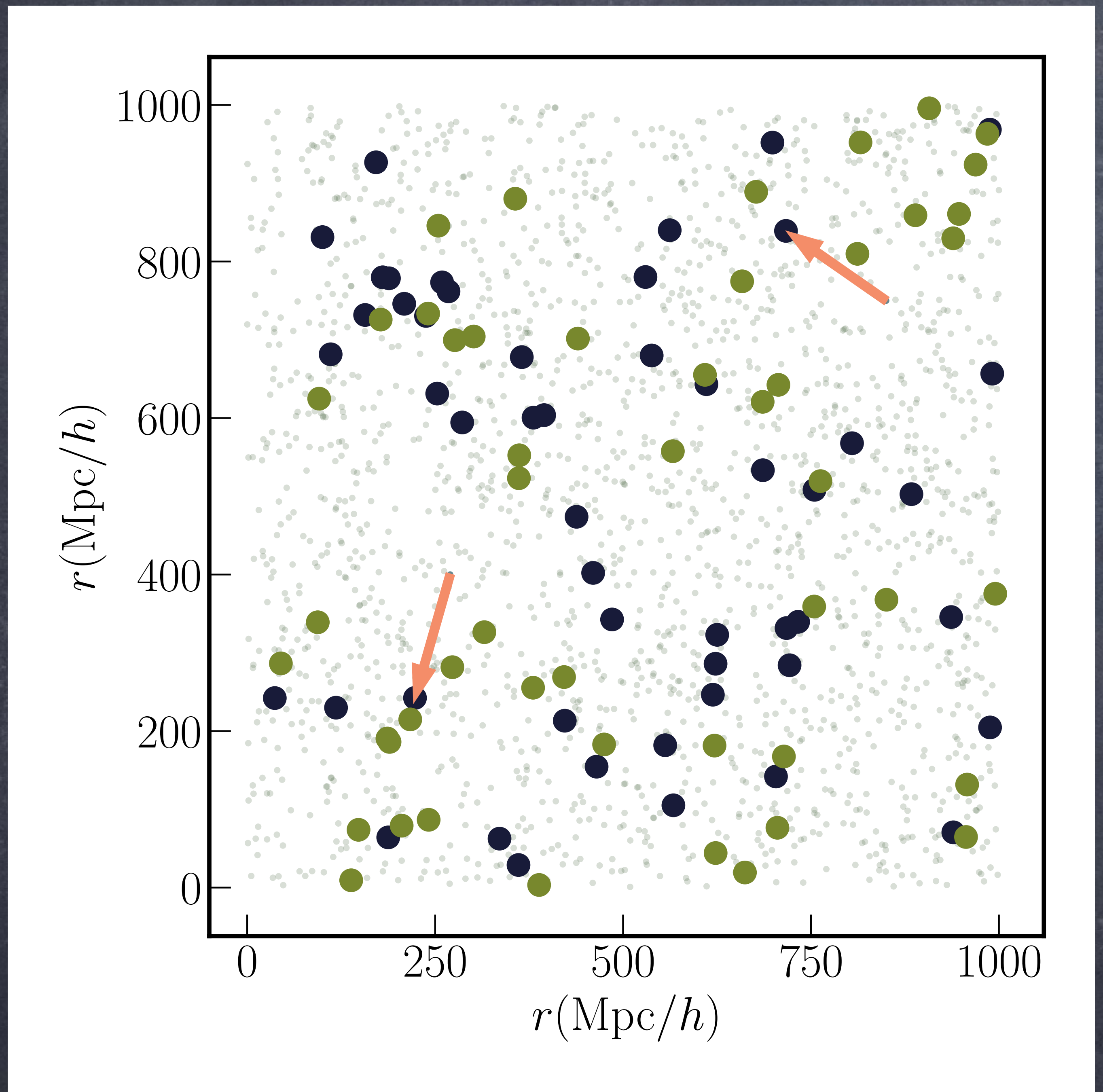
- Sample the volume densely with a set of query points.
- For each query point, find the distance to the nearest data point of each dataset.
- For each query point, pick the larger distance.

Cross-correlations with nearest neighbors

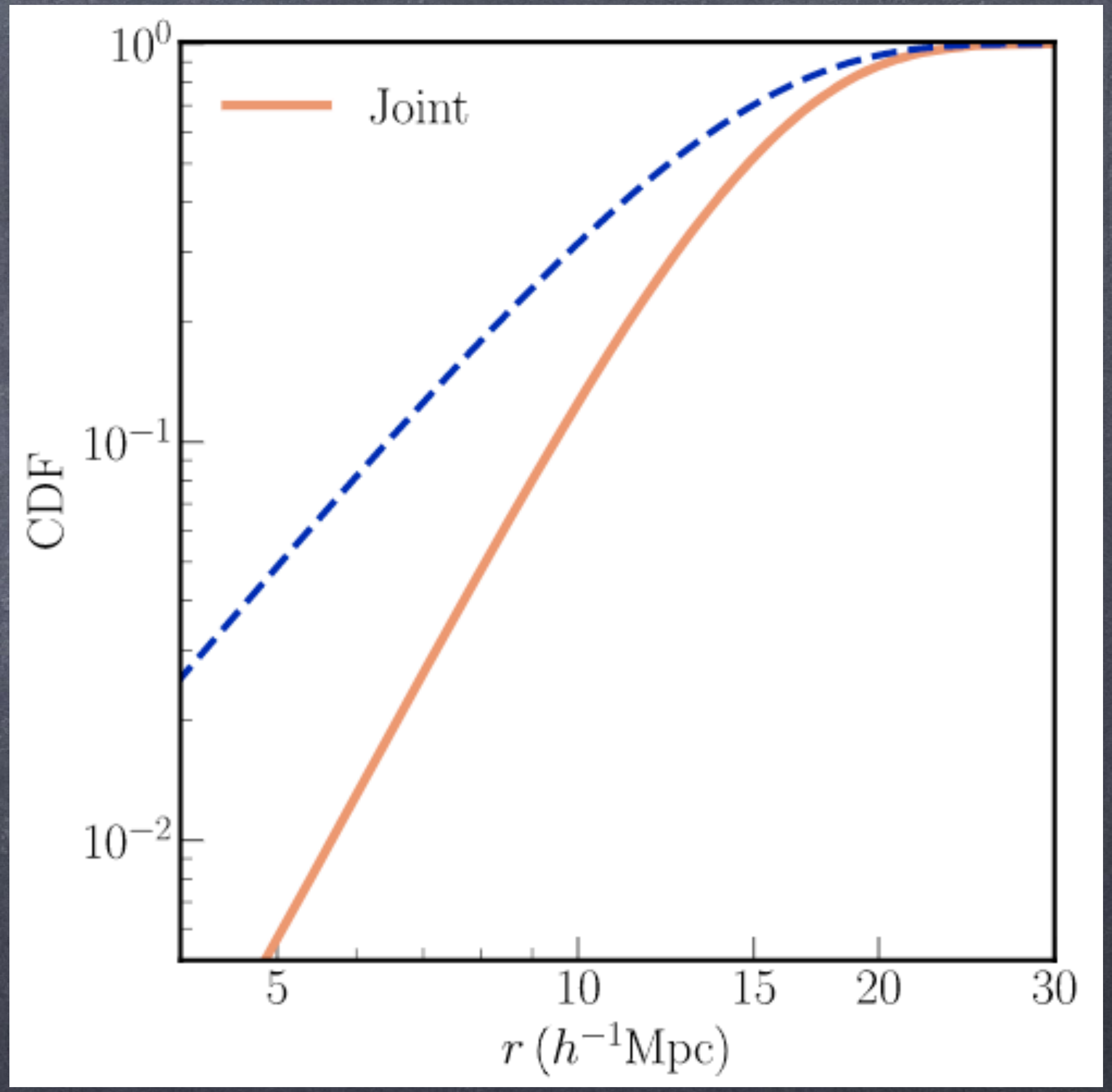


- Sample the volume densely with a set of query points.
- For each query point, find the distance to the nearest data point of each dataset.
- For each query point, pick the larger distance.

Cross-correlations with nearest neighbors



- Sample the volume densely with a set of query points.
- For each query point, find the distance to the nearest data point of each dataset.
- For each query point, pick the larger distance.
- Sort distances, get the empirical (joint) CDF.
- Generalize to the (k_1,k_2) nearest neighbor distributions.



Cross-correlations with nearest neighbors

- For a single set of particles, $\mathrm{CDF}_k(r) = \mathscr{P}_{>k-1}(V)$. Similarly, $\mathrm{CDF}_{k_1,k_2}(r) = \mathscr{P}_{>k_1-1,>k_2-1}(V)$.
- The generating function for $\mathscr{P}_{k_1,k_2}(V)$ is given by

$$P(z_1, z_2|V) = \exp\left[\sum_{k_1=0}^{\infty}\sum_{k_2=0}^{\infty}\frac{\bar{n}_1^{k_1}(z_1-1)^{k_1}}{k_1!}\frac{\bar{n}_2^{k_2}(z_2-1)^{k_2}}{k_2!} \times \int_V d^3\boldsymbol{r}_1 \ldots d^3\boldsymbol{r}_{k_1} d^2\boldsymbol{r}'_1 \ldots d^3\boldsymbol{r}'_{k_2}\,\xi^{(k_1,k_2)}\right]$$

- The generating function for $\mathscr{P}_{>k_1,>k_2}$ is

$$C(z_1, z_2|V) = \frac{1 - P_1(z_1|V) - P_2(z_2|V) + P(z_1, z_2|V)}{(1-z_1)(1-z_2)}$$

Cross-correlations with nearest neighbors

- It is quite easy to isolate the parts of these measurements which depends only on the cross-correlations.

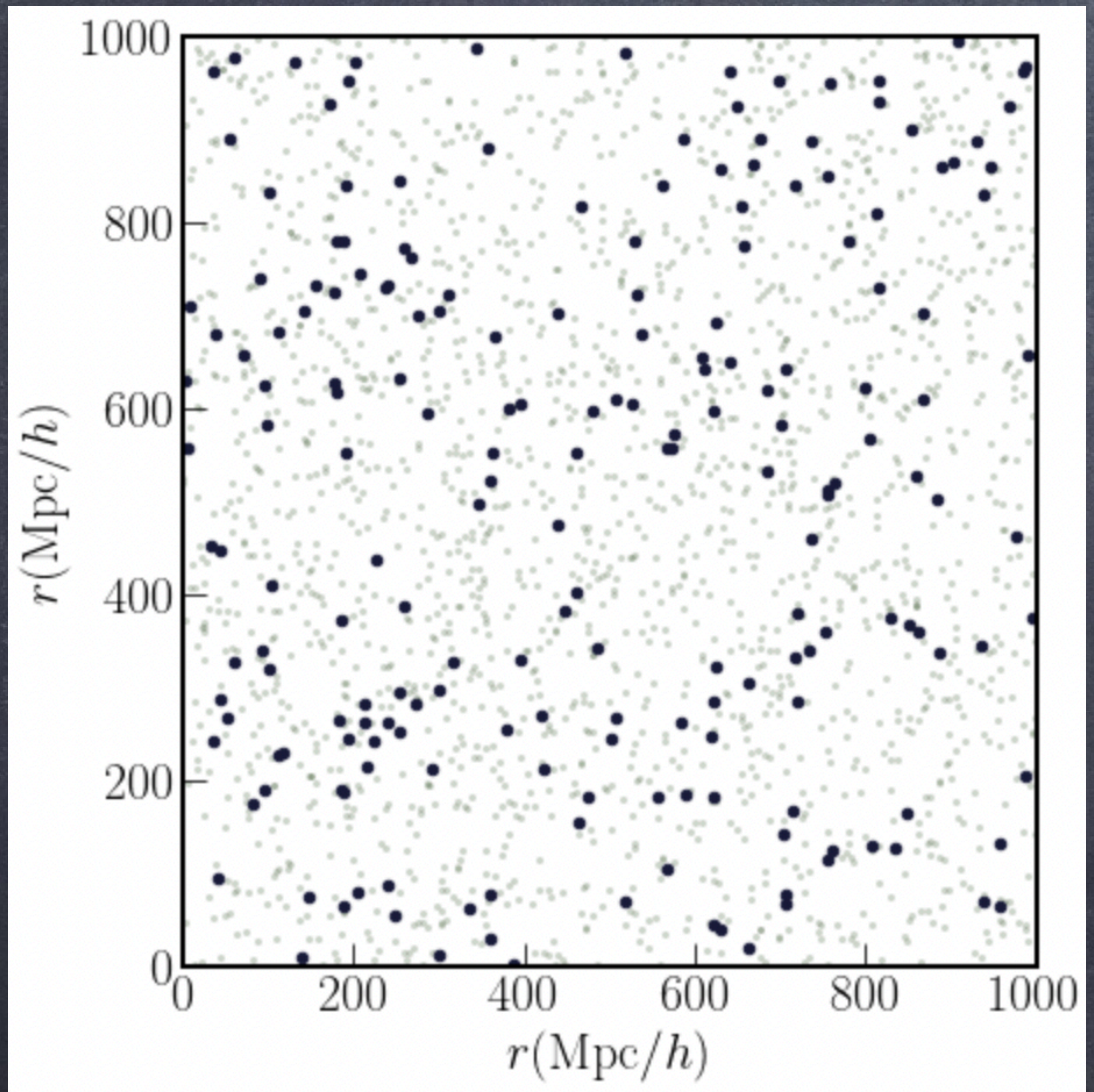
- For completely uncorrelated datasets, $\mathscr{P}_{>k_1,>k_2}(V) = \mathscr{P}_{>k_1}(V) \times \mathscr{P}_{>k_2}(V)$.

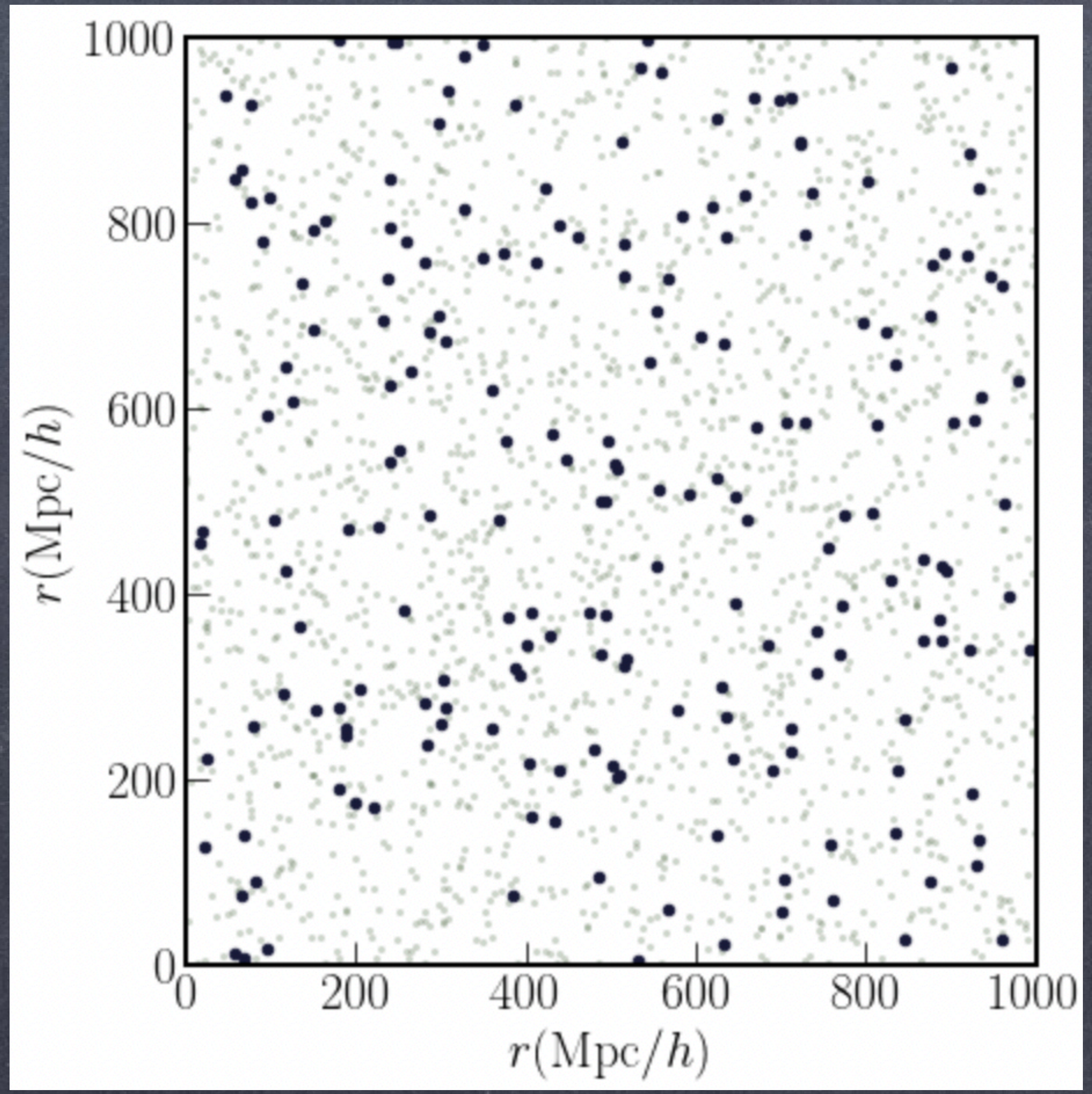
$$\xi'(r) = \mathrm{CDF}_{k_1,k_2}(r) - \mathrm{CDF}_{k_1}^{(1)}(r)\mathrm{CDF}_{k_2}^{(2)}(r)$$

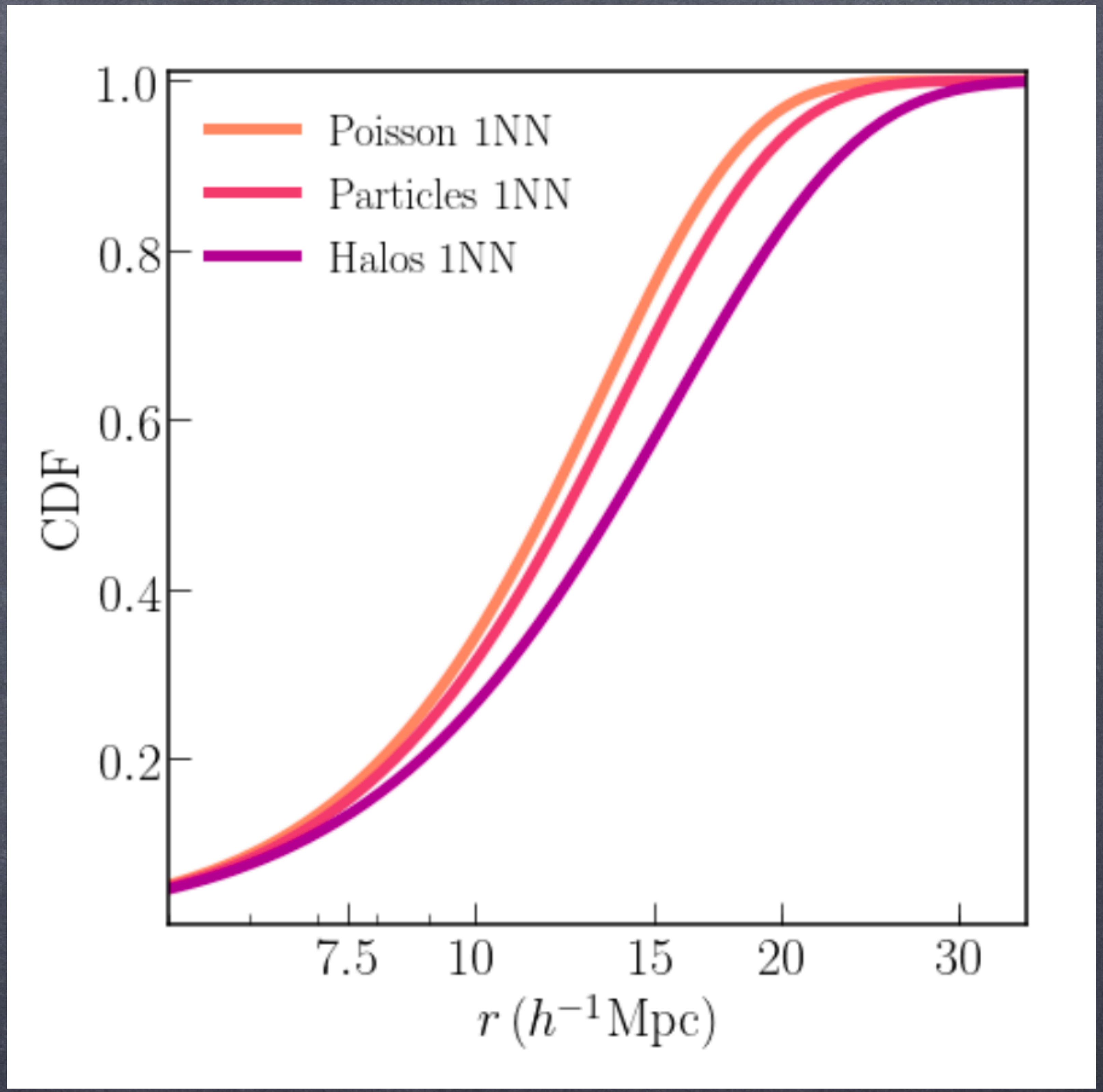
- When this is 0, the two sets are uncorrelated.

Physical intuition for the kNNs

- More clustered the data points (at fixed number density), the more prominent the voids, i.e. the CDF extends out to larger distances.



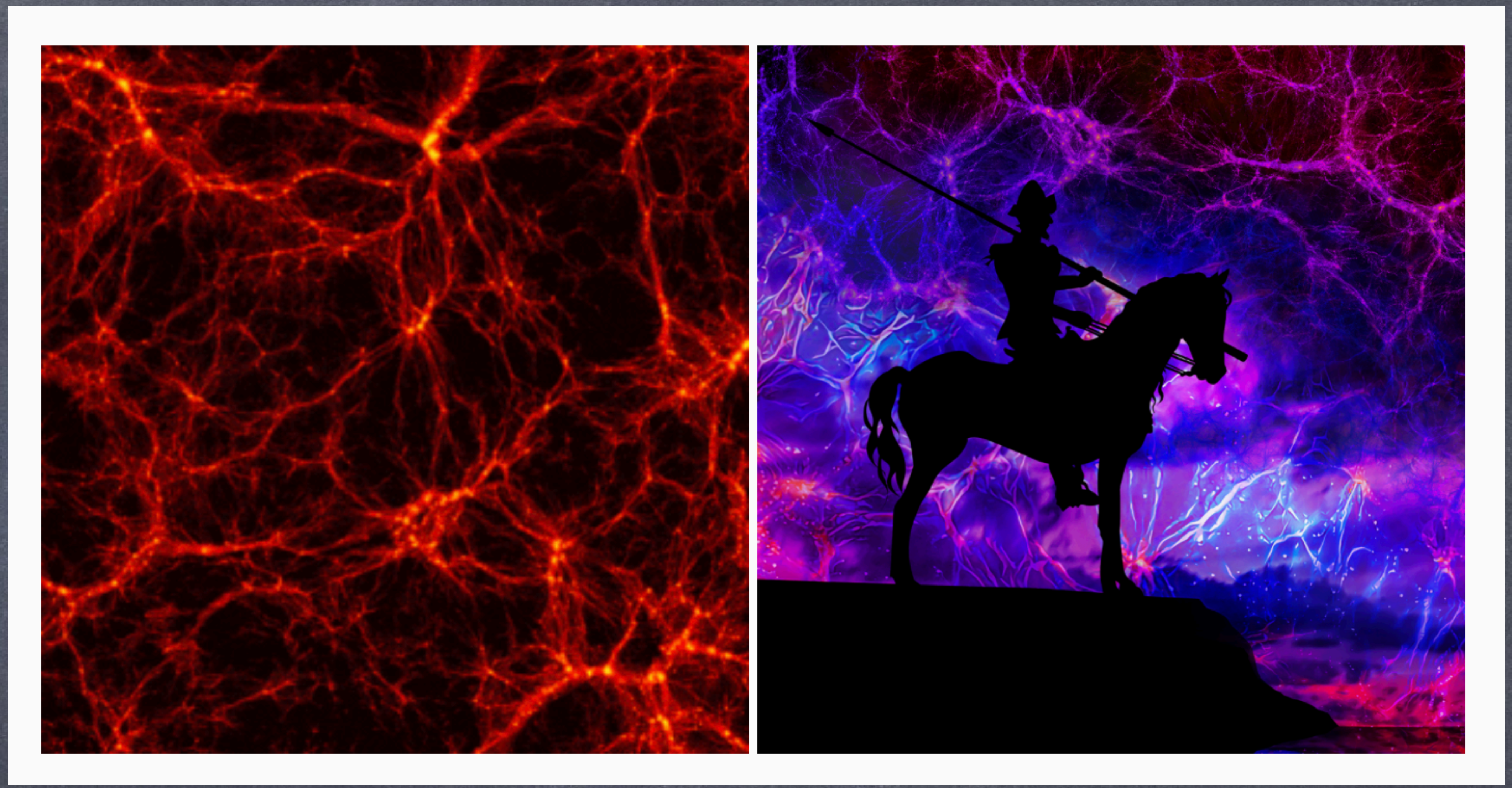




Banerjee & Abel, 2020

Cosmological information with kNN distributions

- We use a Fisher matrix analysis to test how sensitive kNN statistics are to various cosmological parameters, compared to the 2-point function.
- We use the same underlying simulation data to compute the two sets of statistics, and compare their change as a function of change in the values of the cosmological parameters.

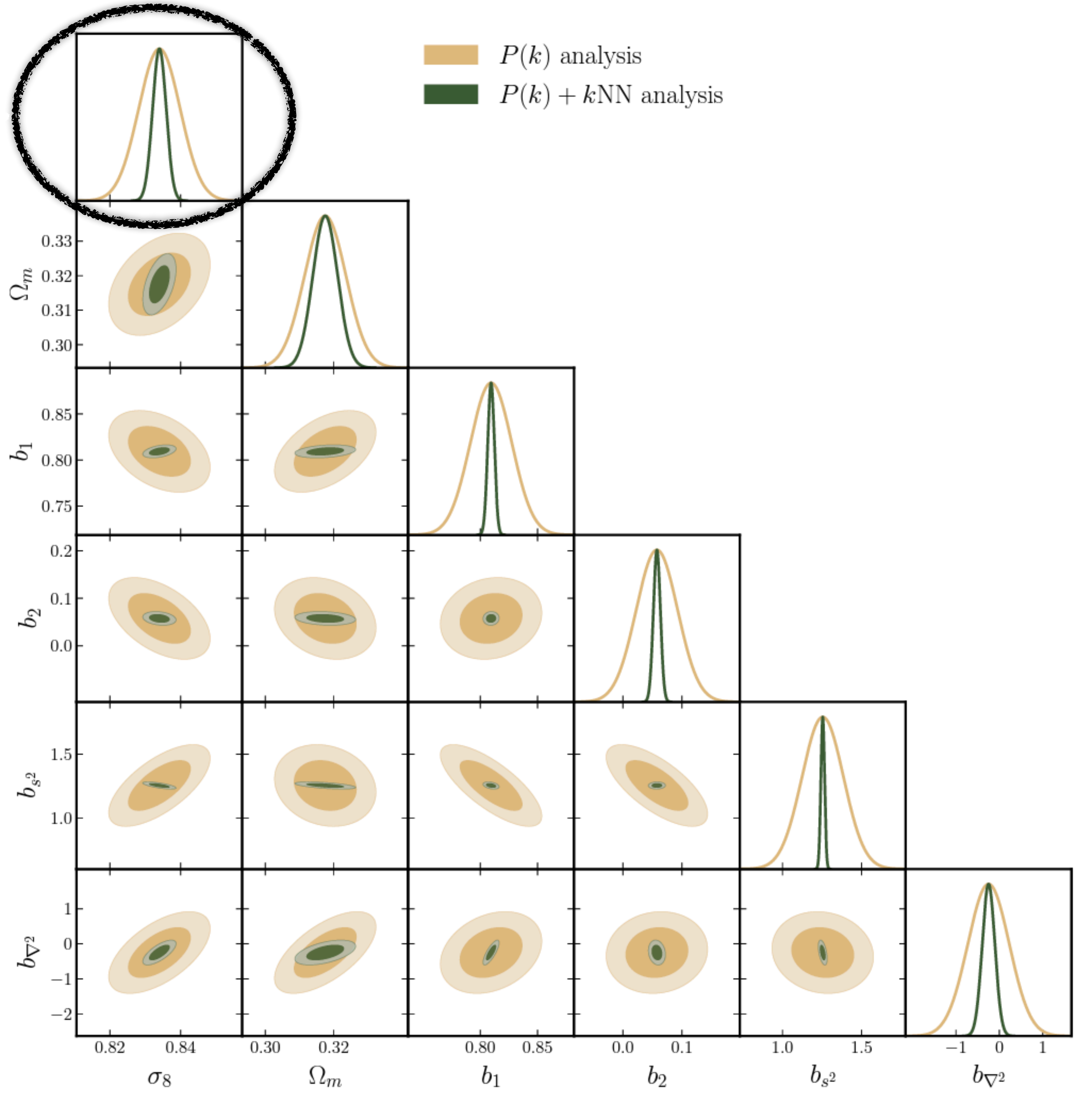


THE QUIJOTE SIMULATIONS

Francisco Villaescusa-Navarro[1,2,†], ChangHoon Hahn[3,4], Elena Massara[1,5], Arka Banerjee[6,7,8], Ana Maria Delgado[9,1], Doogesh Kodi Ramanah[10,11], Tom Charnock[10], Elena Giusarma[1,12], Yin Li[1,3,4,13,31], Erwan Allys[14], Antoine Brochard[15,16], Cora Uhlemann[17,18], Chi-Ting Chiang[19], Siyu He[1], Alice Pisani[2], Andrej Obuljen[5], Yu Feng[3,4], Emanuele Castorina[3,4], Gabriella Contardo[1], Christina D. Kreisch[2], Andrina Nicola[2], Justin Alsing[20,1], Roman Scoccimarro[21], Licia Verde[22,23], Matteo Viel[24,25,26,27], Shirley Ho[1,2,28], Stephane Mallat[29,30], Benjamin Wandelt[10,11,1], David N. Spergel[2,1]

Improvements in parameter constraints

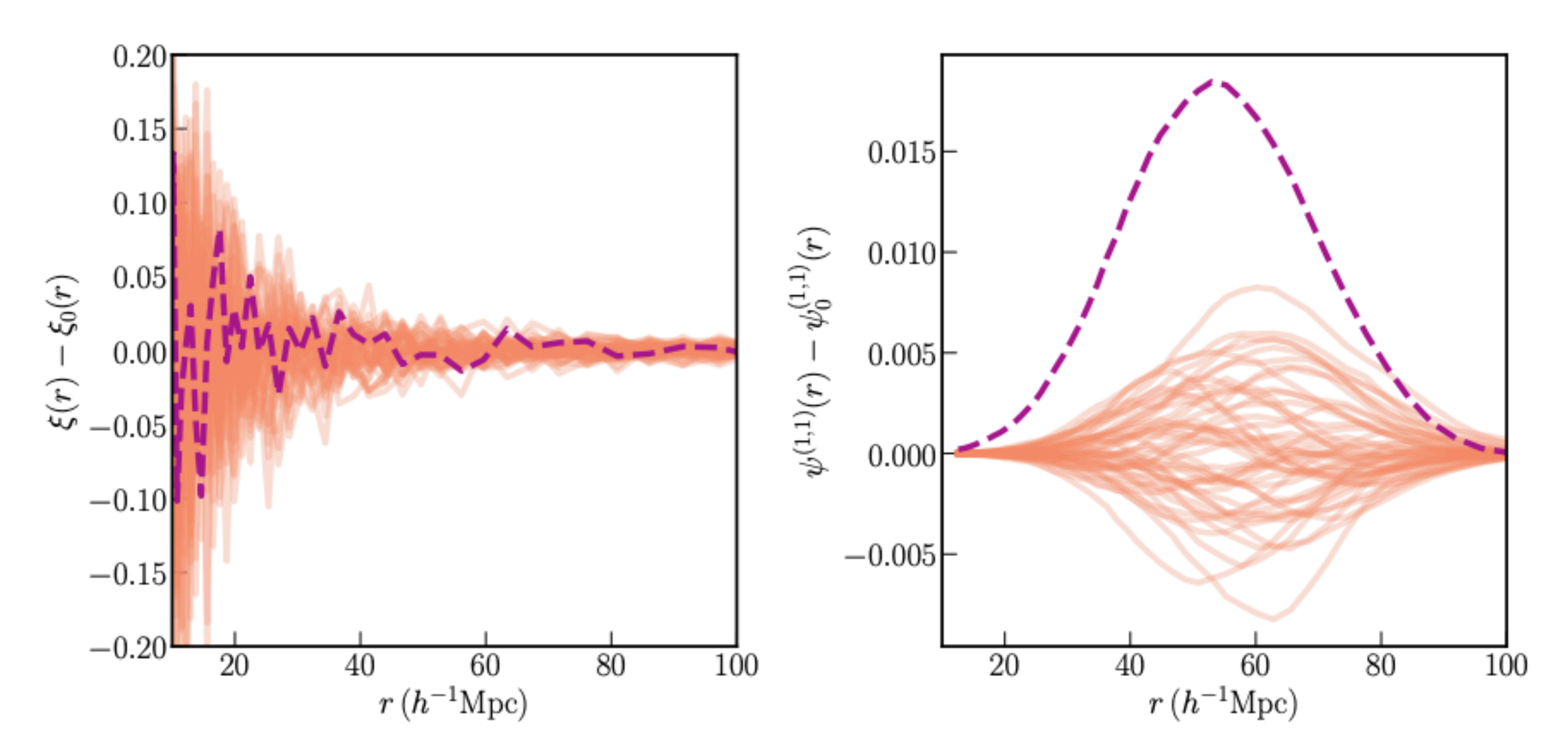
Factor of 3 improvement



Banerjee et al, 2022

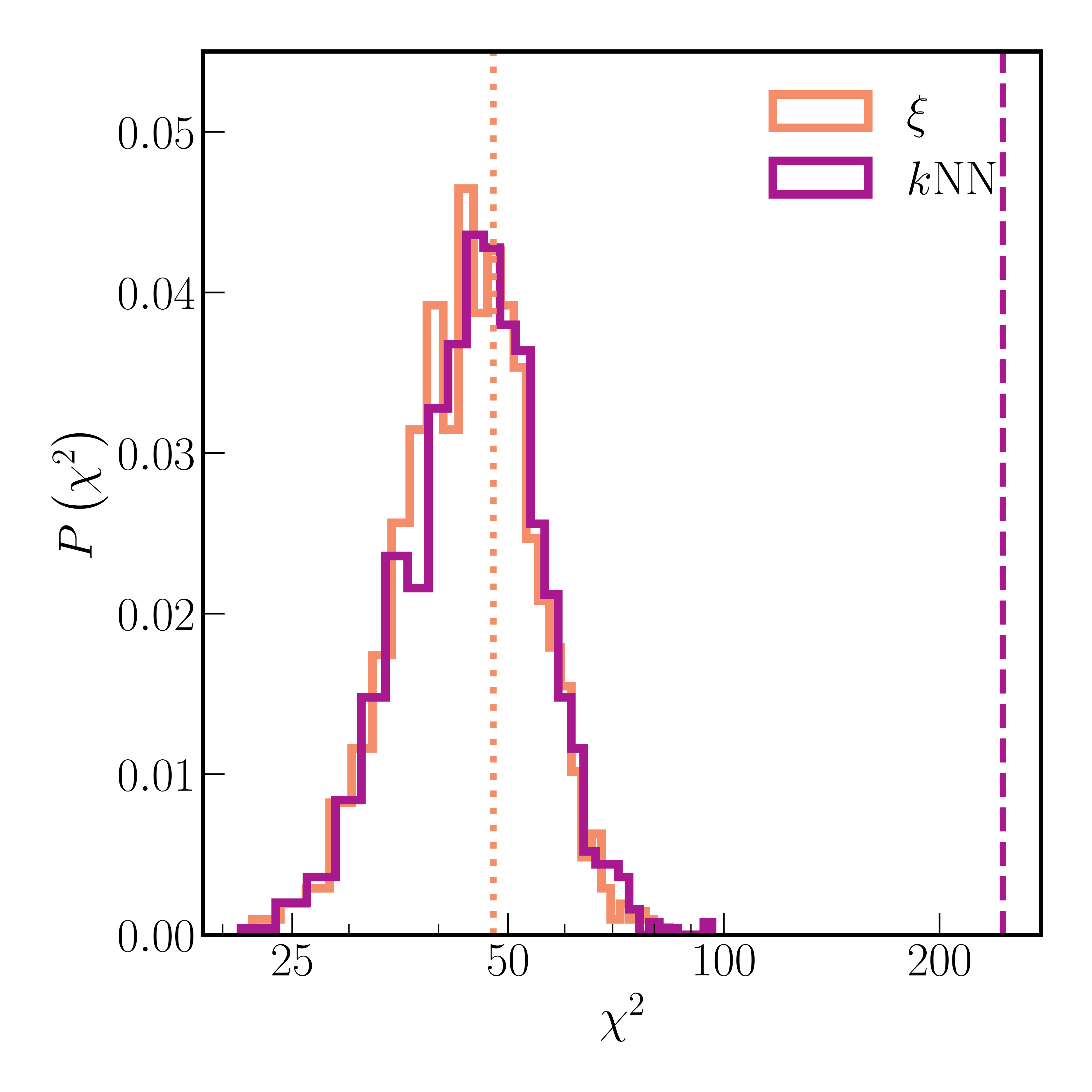
Detection of cross-correlations for sparse samples

$$\bar{n} = 10^{-6} h^3 \mathrm{Mpc}^{-3}$$



Banerjee & Abel, 2021

Detection of cross-correlations for sparse samples

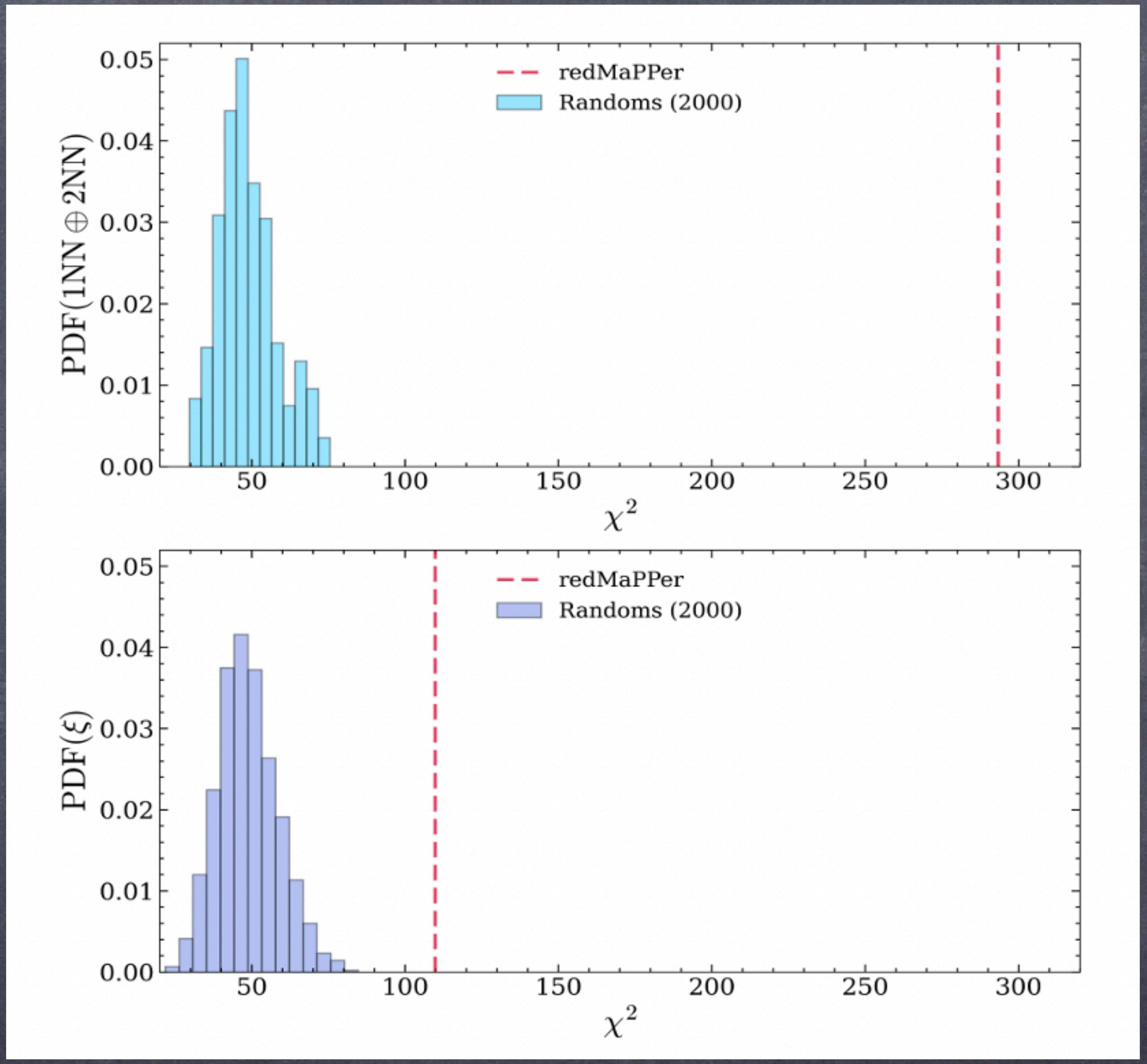


Banerjee & Abel, 2021

First application to data

Detection of spatial clustering in the 1000 richest SDSS DR8 redMaPPer clusters with Nearest Neighbor distributions

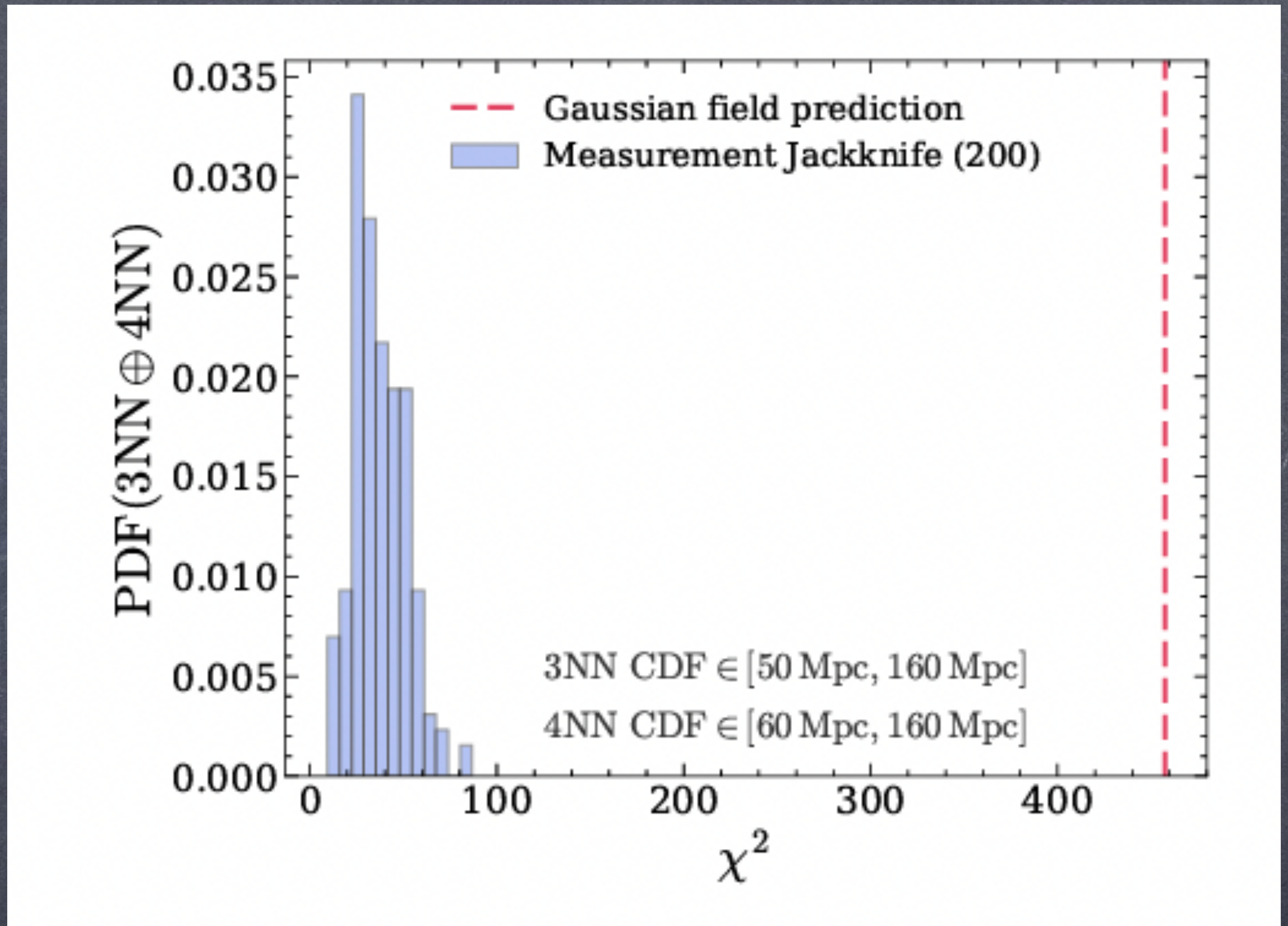
Yunchong Wang,[1,2]⋆ Arka Banerjee [4] and Tom Abel [1,2,3]



First application to data

Detection of spatial clustering in the 1000 richest SDSS DR8 redMaPPer clusters with Nearest Neighbor distributions

Yunchong Wang,[1,2*] Arka Banerjee [4] and Tom Abel [1,2,3]



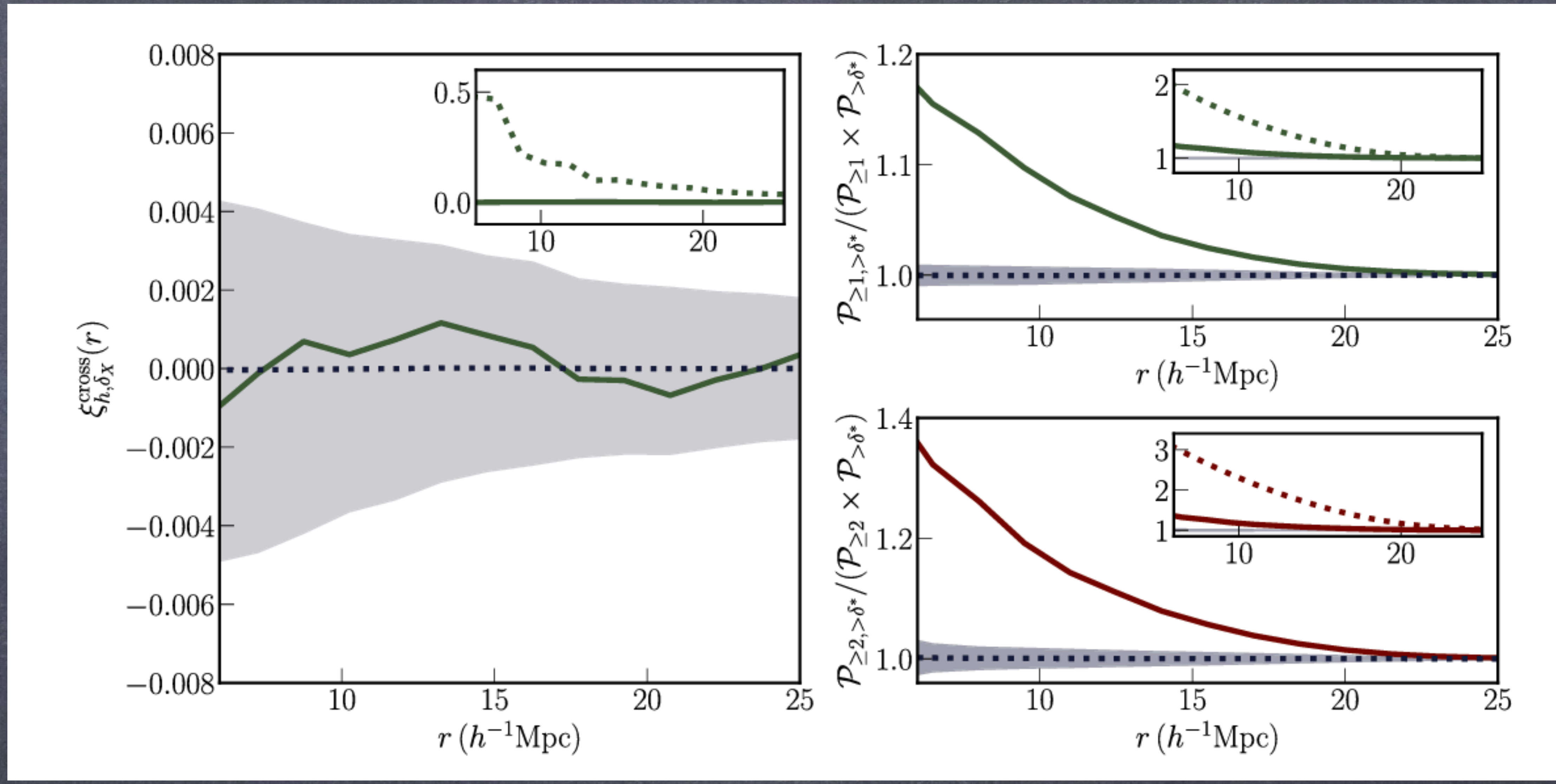
Point-field cross correlations

Monthly Notices of the ROYAL ASTRONOMICAL SOCIETY

MNRAS **519,** 4856–4868 (2023) https://doi.org/10.1093/mnras/stac3813
Advance Access publication 2022 December 30

Tracer-field cross-correlations with k-nearest neighbour distributions

Arka Banerjee[1★] and Tom Abel[2,3,4]



Summary

- Understanding structure formation in the Universe can help answer some of the most fundamental questions in physics (inflation, DM, DE, massive neutrinos, additional light species, ...)

- Large amounts of untapped information on small, nonlinear scales.

- Need to go beyond 2 point statistics. kNN distributions offer a computationally cheap and interpretable path to higher-order statistics. Shows much greater statistical constraining power.

- Many potential applications in cosmology (also GW clustering) and beyond. Deep connections to geometrical and topological measures of clustering such as Minkowski functionals and Betti numbers.

Thank you!