Topological Data Analysis for Cosmology and Galaxy Evolution

Tsutomu T. TAKEUCHI

with
Kai KONO, Suchetha COORAY

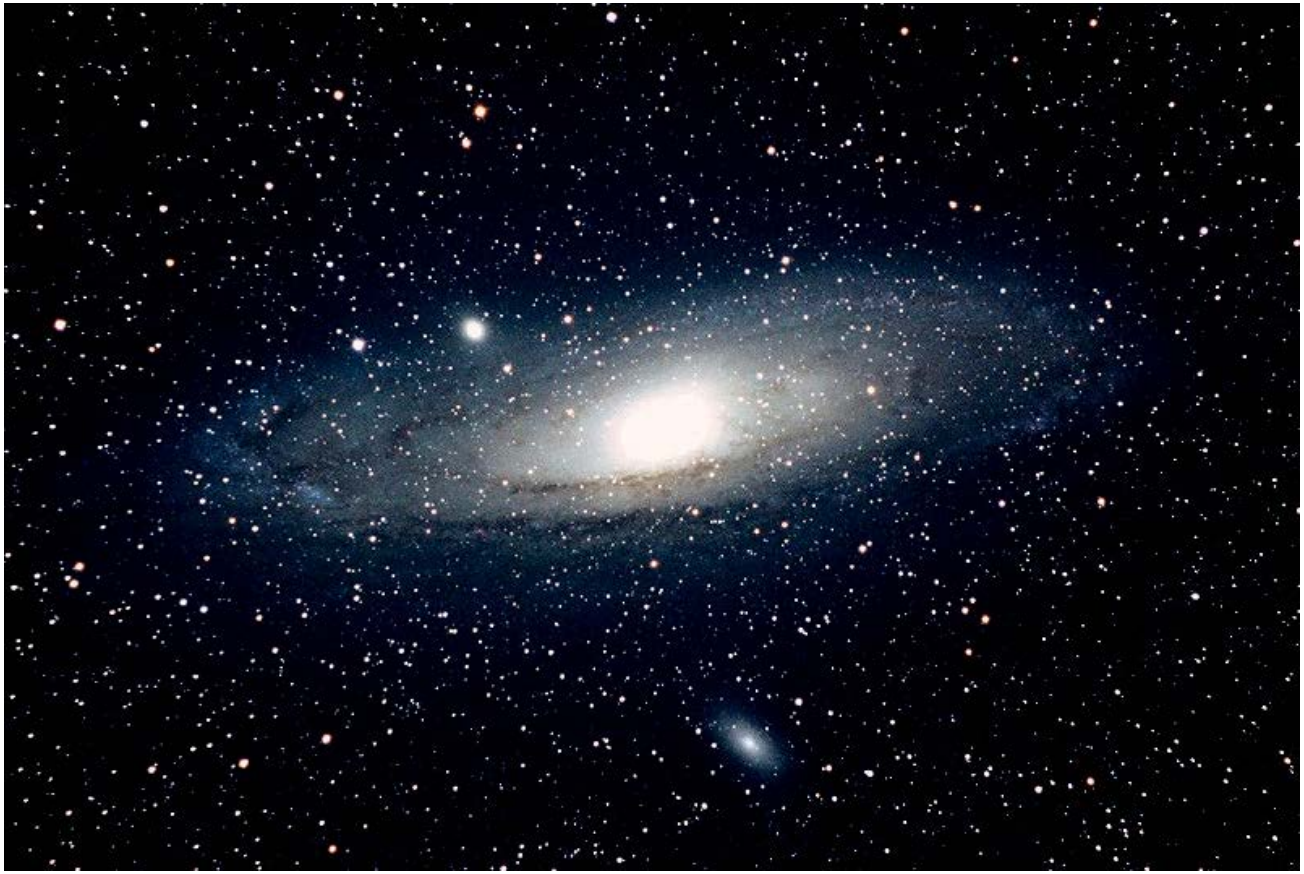
Division of Particle and Astrophysical Science, Nagoya University, Japan

New Development of the Statistical Machine Learning,
ISM, Tachikawa, 30-31 Jan., 2020

0. Background

0.1 What are galaxies?

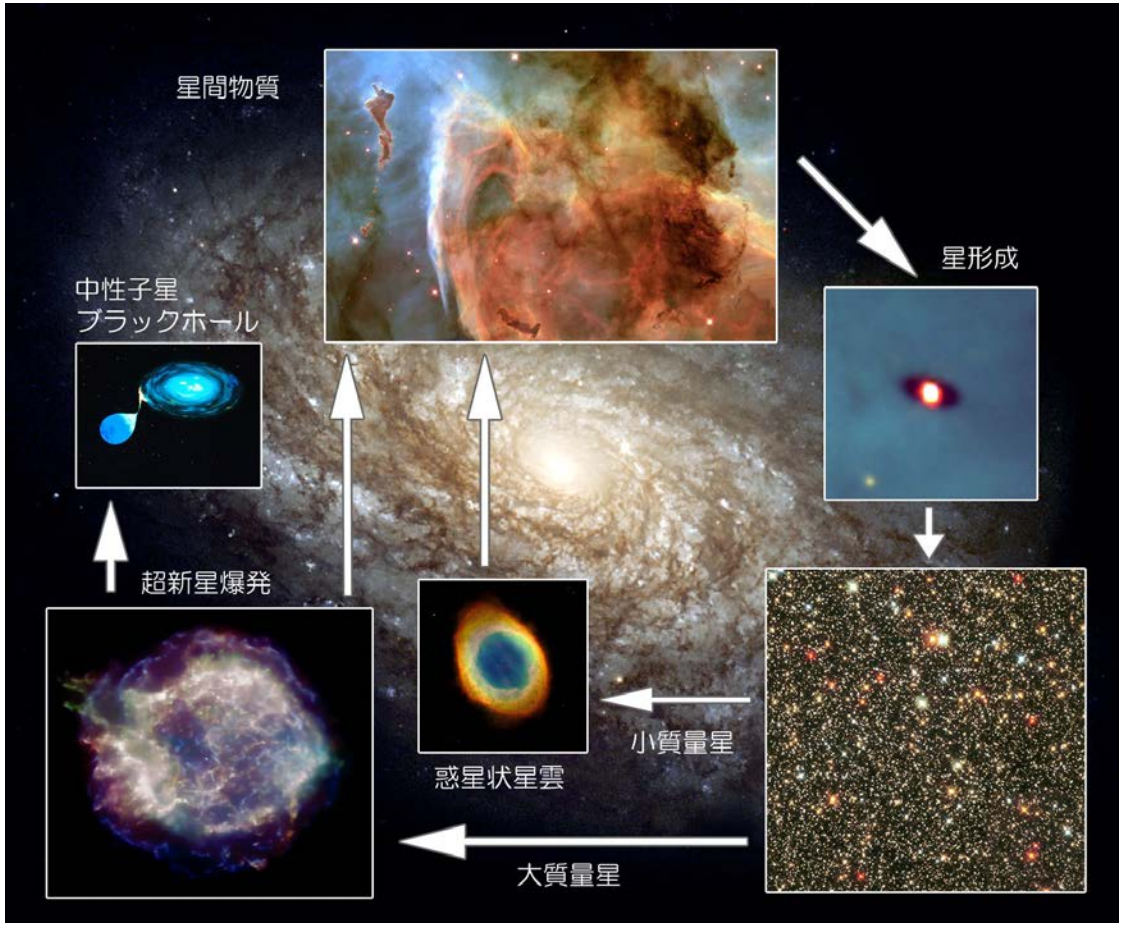
A galaxy is a huge agglomeration of stars, interstellar medium (ISM: gas+dust), and dark matter (DM), a complex system with a complicated interaction between each component.



0. Background

0.1 What are galaxies?

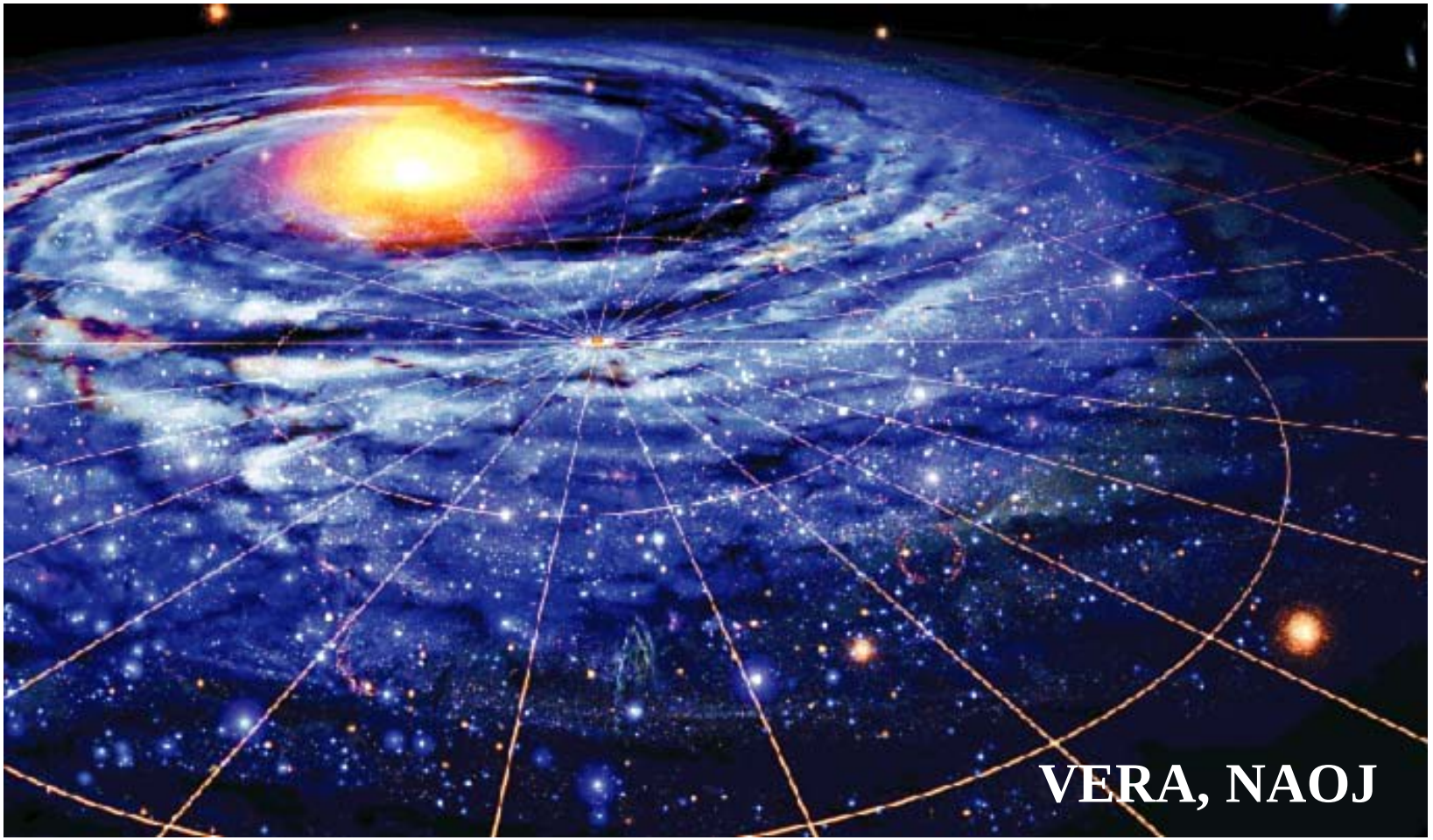
A galaxy is a huge agglomeration of stars, interstellar medium (ISM: gas+dust), and dark matter (DM), a complex system with a complicated interaction between each component.



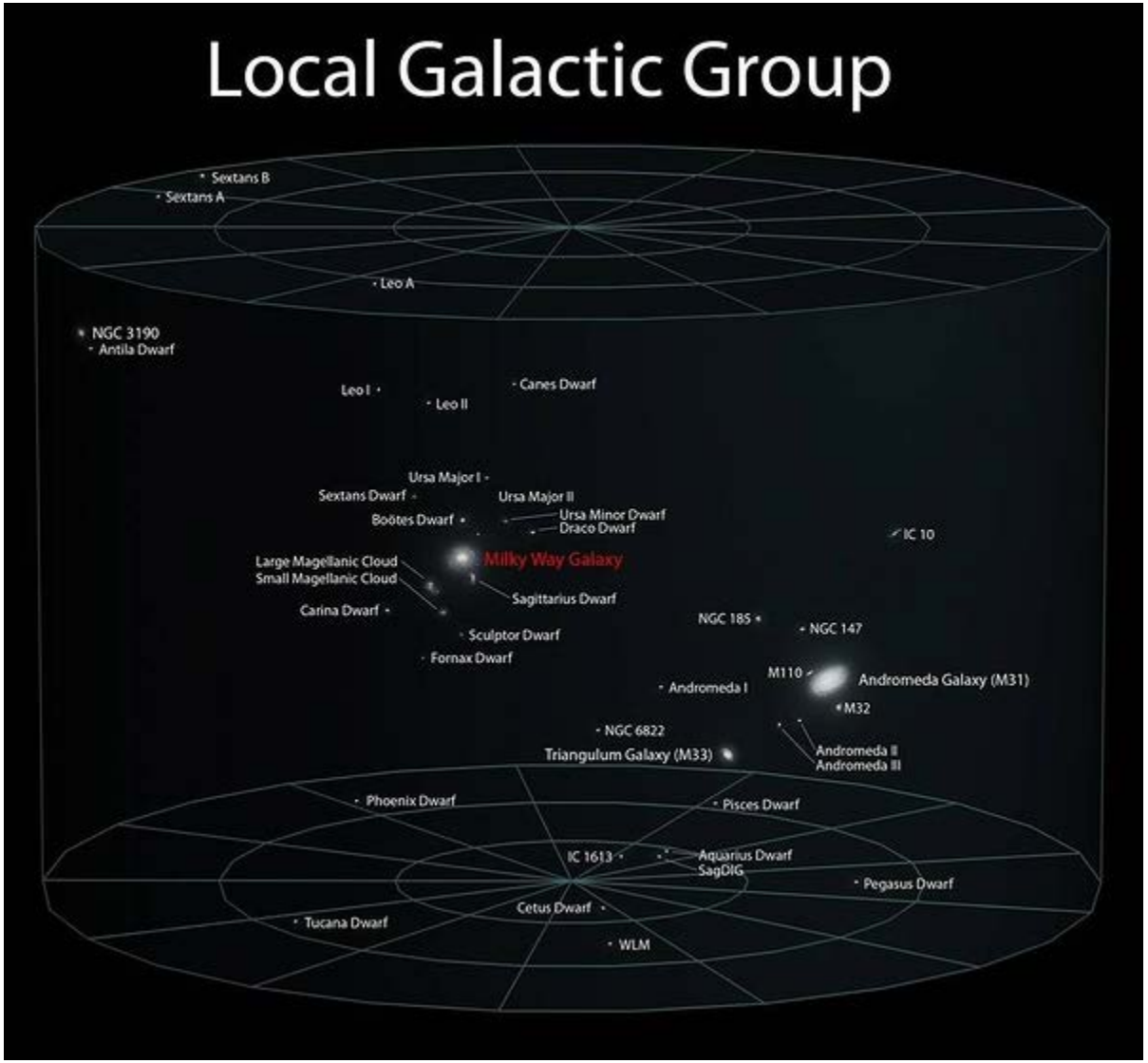
©Shokabo

0.2 Galaxies to the Large-Scale Structures

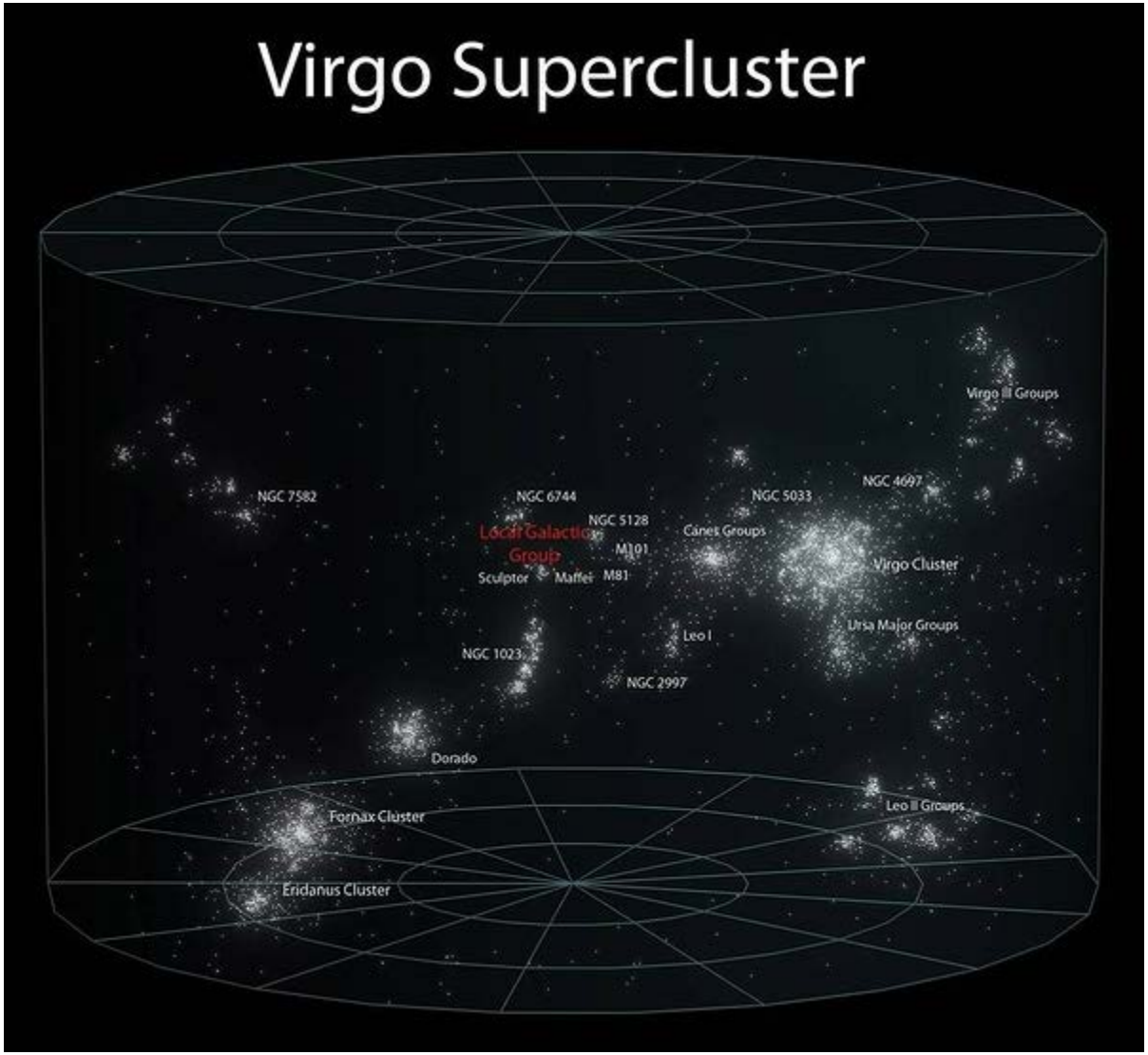
If we observe the Milky Way from outside, it would appear as a disk with spiral structure which consists of gas and stars.



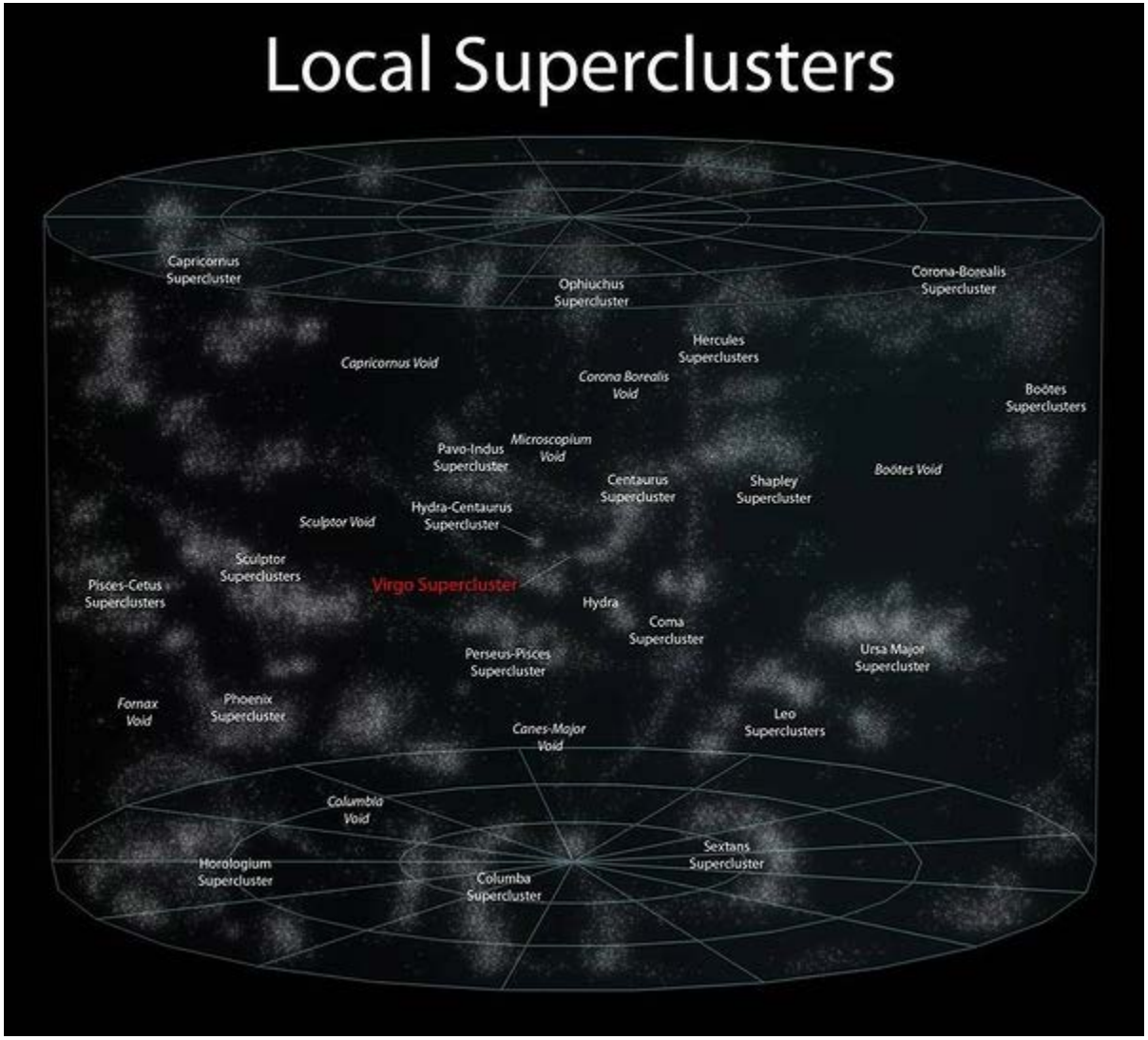
From galaxies to groups and clusters of galaxies



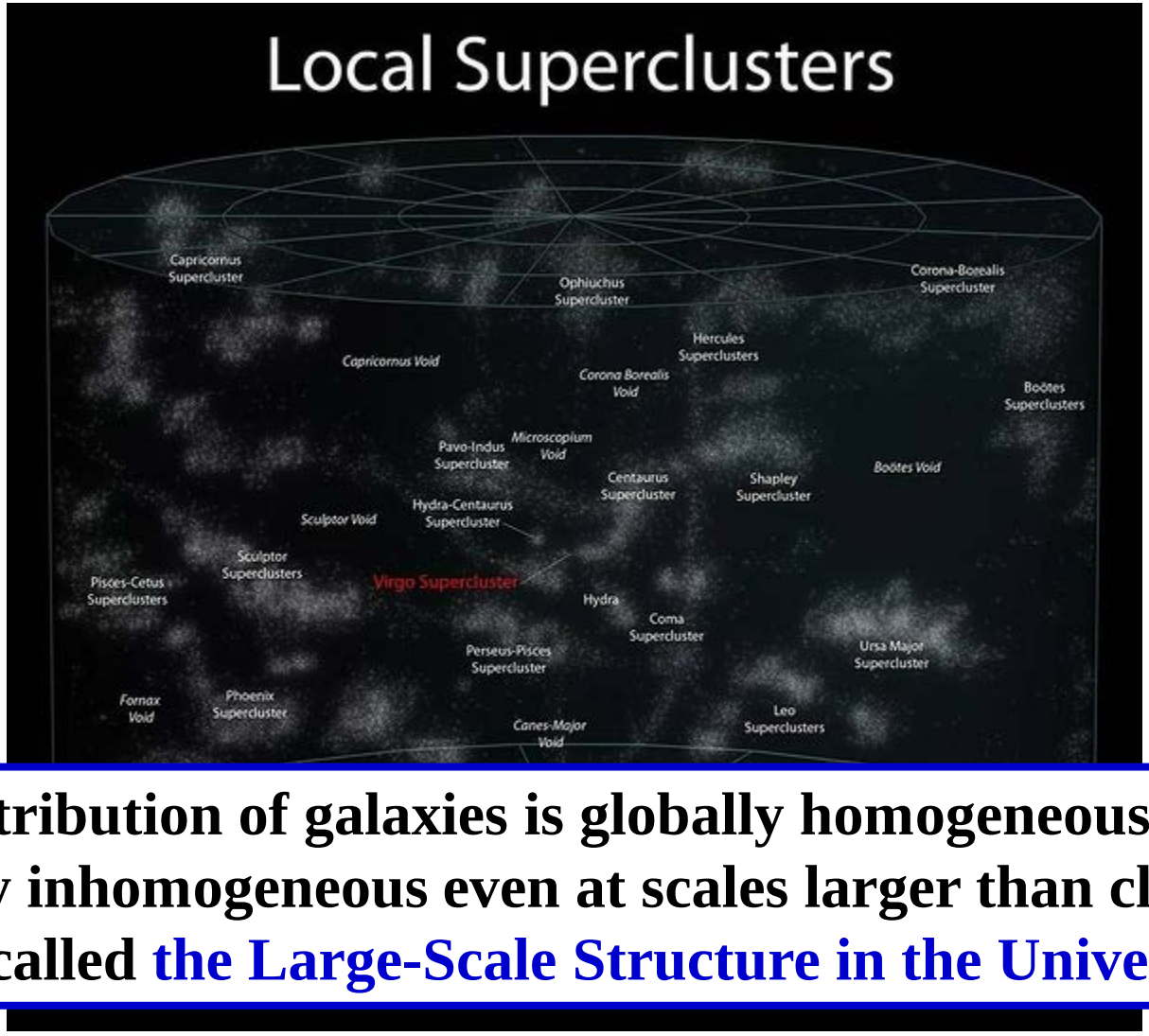
From galaxies to groups and clusters of galaxies



From groups and clusters to the Large-Scale Structure

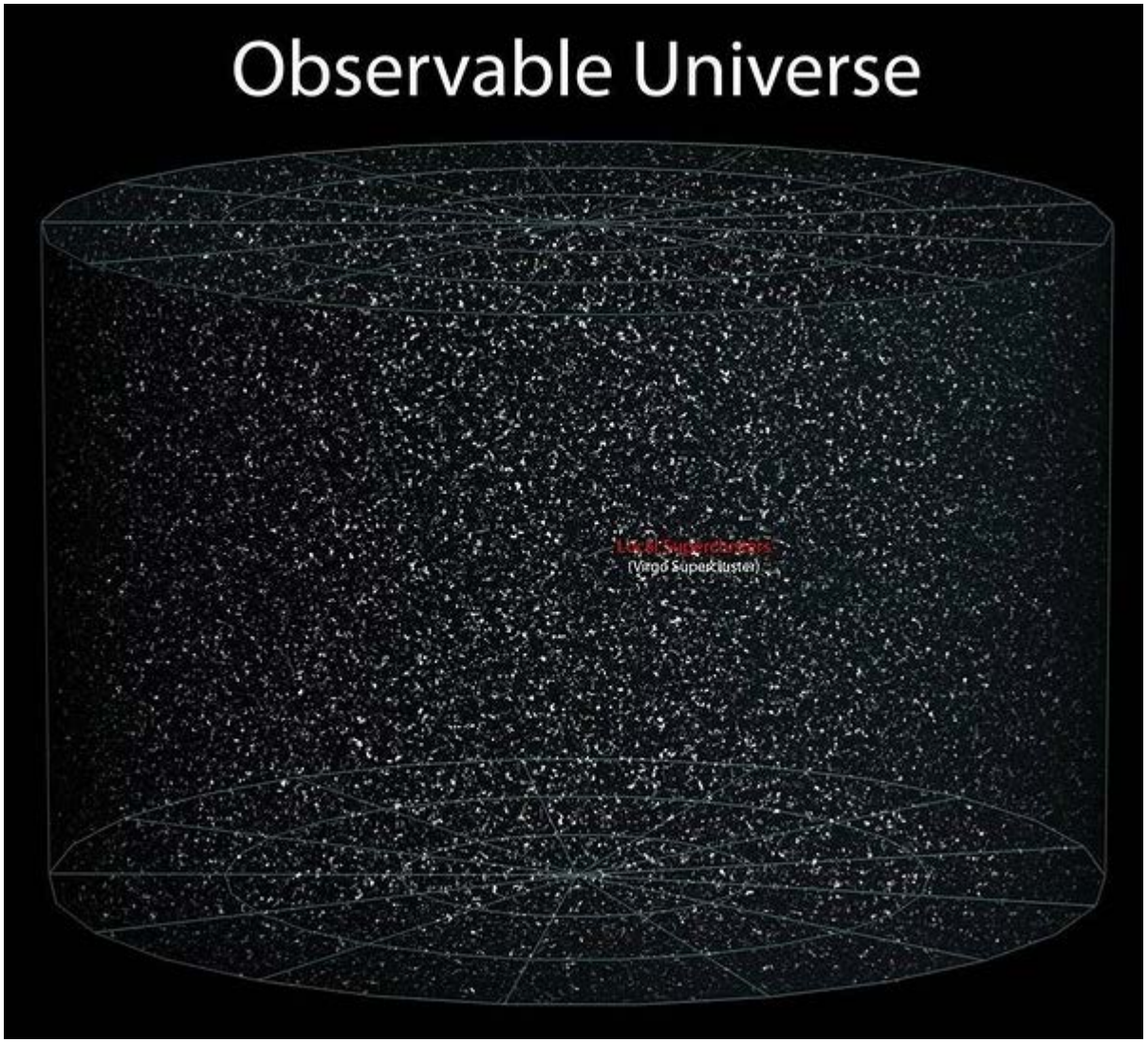


From groups and clusters to the Large-Scale Structure



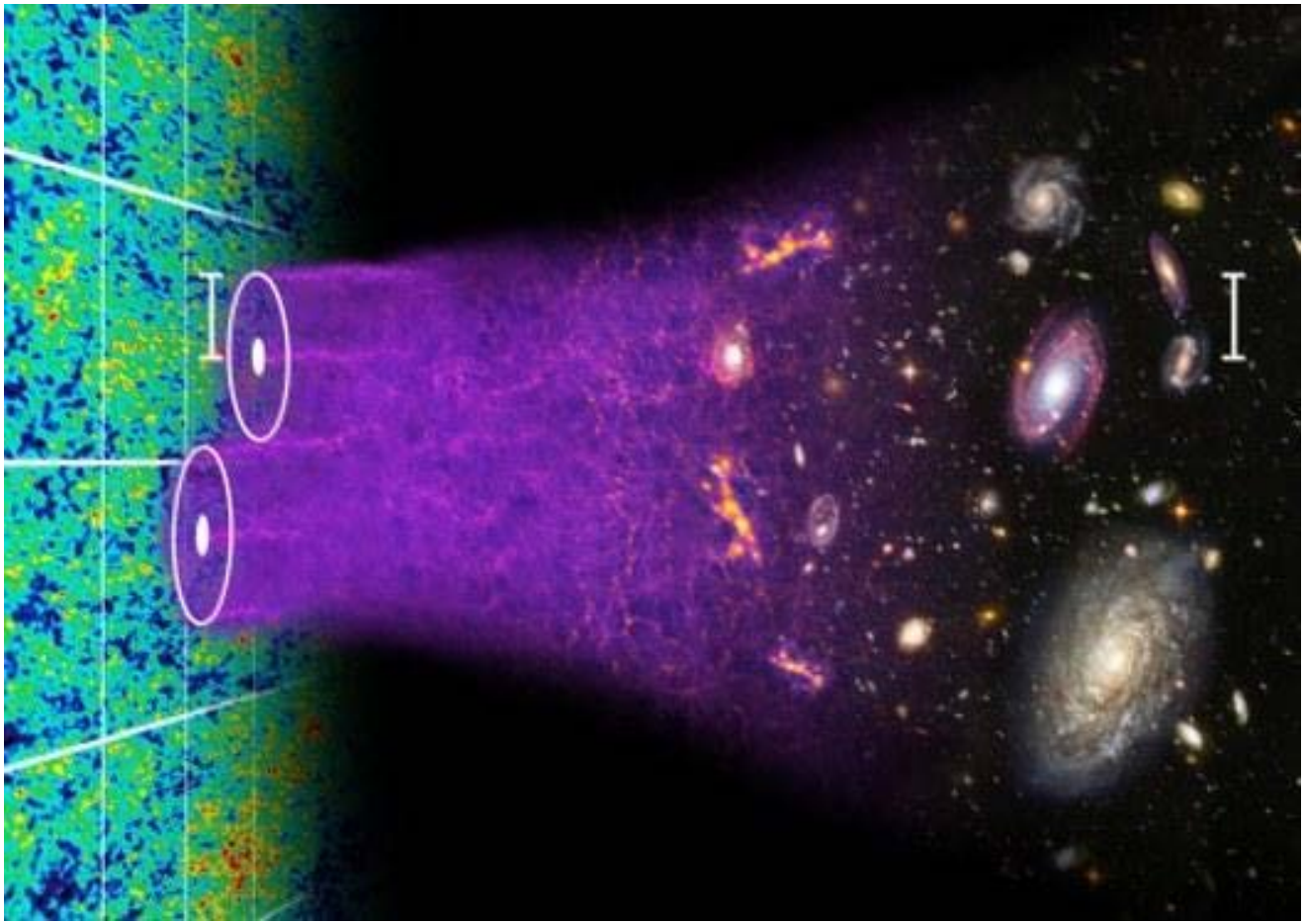
The distribution of galaxies is globally homogeneous, but strongly inhomogeneous even at scales larger than clusters. This is called the Large-Scale Structure in the Universe.

From the Large-Scale Structure to the Hubble horizon



1. Introduction

1.1 Structure formation in the Universe

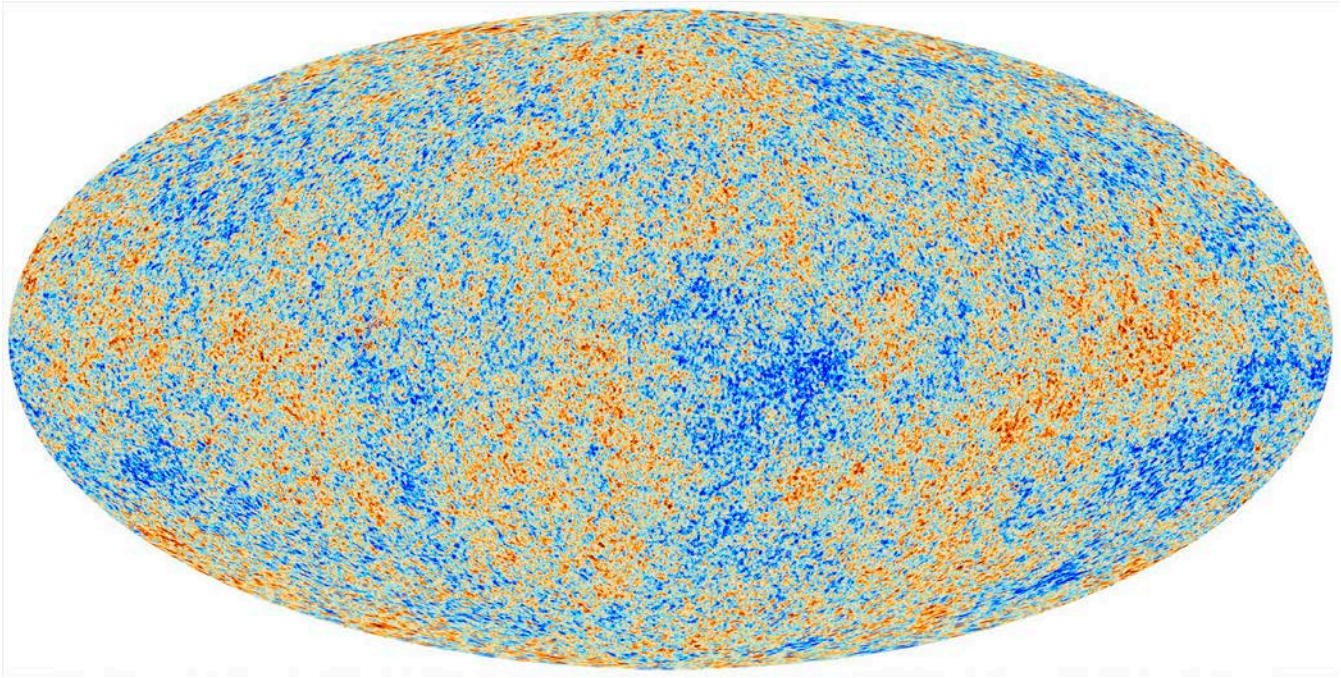


All the structures in the Universe have emerged from a tiny fluctuation at very early epoch (380,000 yr).

1.2 Galaxy formation from the cosmic initial condition

Galaxies are supposed to have formed from a tiny (order of ~ 10^{-5}) fluctuation of matter (mainly dark matter: DM) in the early Universe.

The initial condition is imprinted on the Cosmic Microwave Background (CMB) observed at radio wavelengths.



http://www.rssd.esa.int/index.php?project=Planck

Gaussian random field: initial condition of matter fluctuation

Gaussian random field is a stochastic field whose distribution is described by Gaussian and its Fourier phases have no correlation.

$$f(\delta)\mathrm{d}\delta = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{\delta^2}{2\sigma^2}} \mathrm{d}\delta \qquad \textbf{(1)}$$

$$\langle \phi_{\boldsymbol{k}} \phi_{\boldsymbol{k}'} \rangle \propto \delta_{\mathrm{D}}(\boldsymbol{k} - \boldsymbol{k}') \qquad \textbf{(2)}$$

All the stochastic properties of a field is uniquely characterized by the power spectrum $P(k)$ for Gaussian random fields.

Gaussian random field: initial condition of matter fluctuation

Gaussian random field is a stochastic field whose distribution is described by Gaussian and its Fourier phases have no correlation.

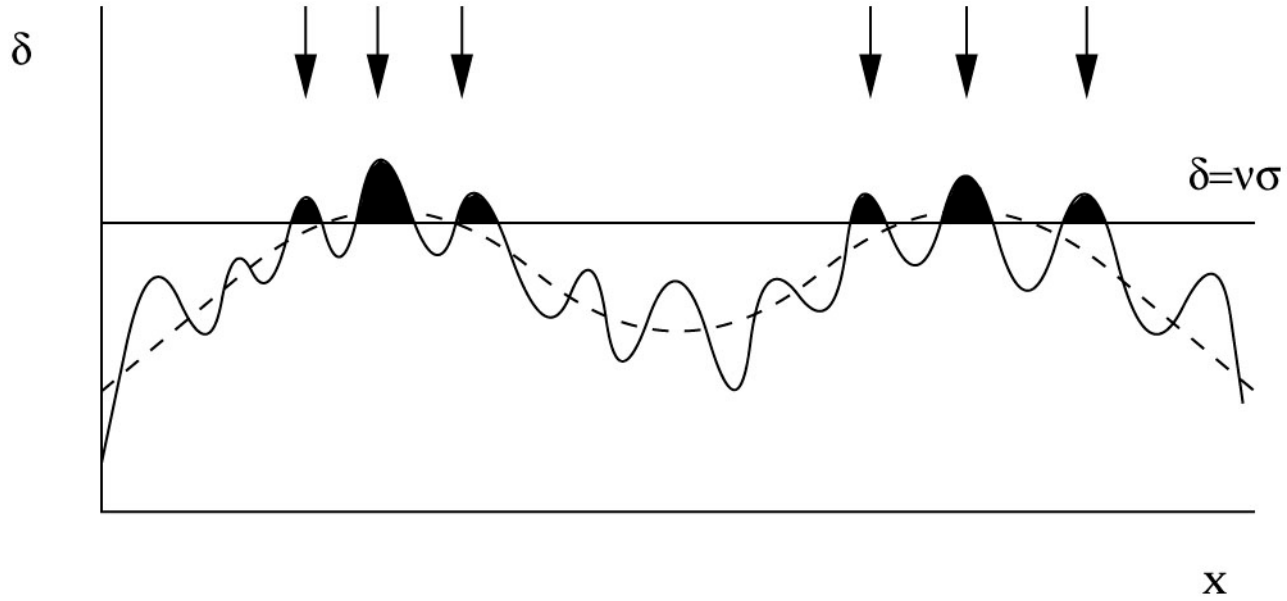
$$f(\delta)\mathrm{d}\delta = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{\delta^2}{2\sigma^2}} \mathrm{d}\delta \qquad (1)$$

$$\langle \phi_{\boldsymbol{k}} \phi_{\boldsymbol{k}'} \rangle \propto \delta_{\mathrm{D}}(\boldsymbol{k} - \boldsymbol{k}') \qquad (2)$$

All the stochastic properties of a field is uniquely characterized by the power spectrum $P(k)$ for Gaussian random fields.

Observationally, the initial density fluctuation in the Universe can be regarded as (almost) Gaussian.

Formation of dark halos: statistical description

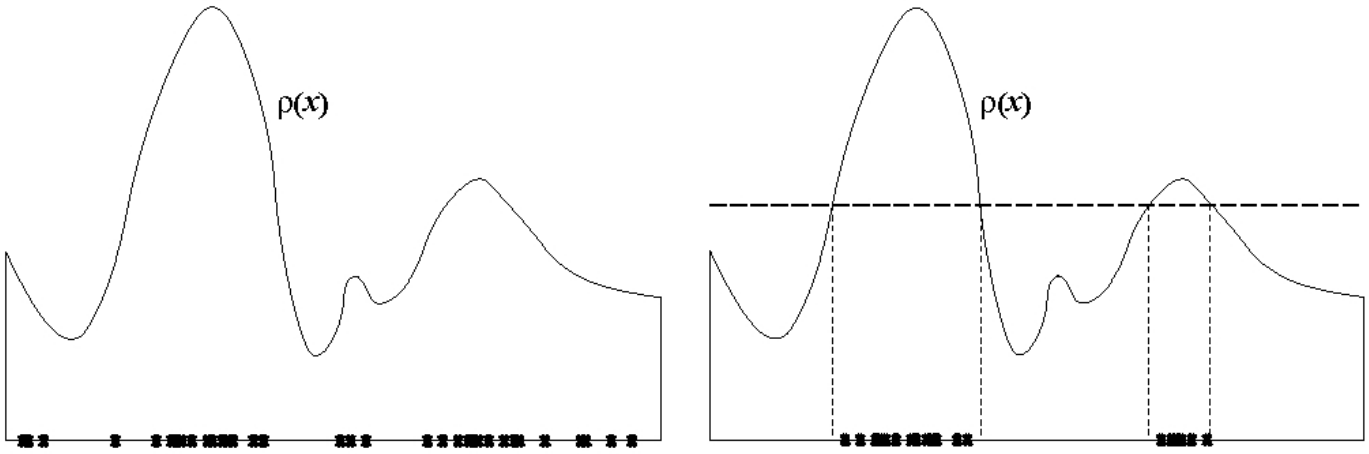


Peacock (1998)

First step

A random Gaussian density field of DM evolves through gravity. When the density of a patch of the field exceeds a certain threshold, the patch starts to be gravitationally bound to form a dark halo.

Formation of dark halos: statistical description

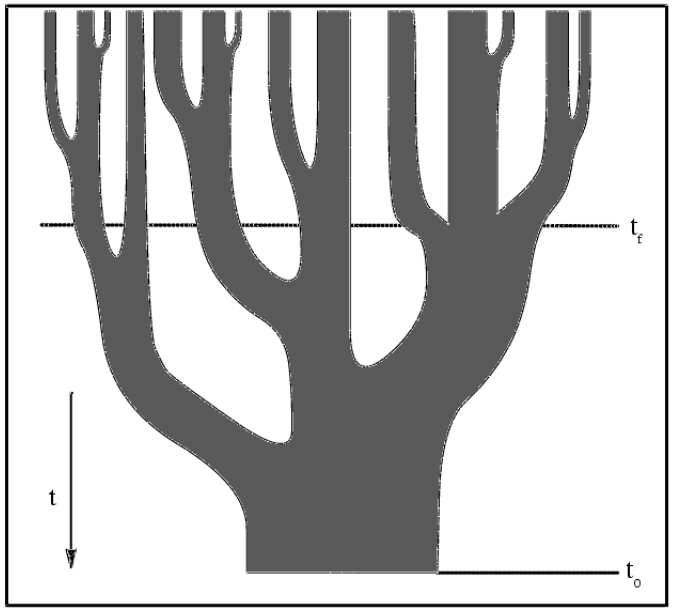


Kerscher (1999)

Second step

In the dark haloes, baryons (normal gas) start to contract to form stars and galaxies. Halos and galaxies are not one-to-one corresponding. From a density field to galaxies, the mathematical treatment changes from a continuous field to a point field.

Formation of dark halos: statistical description



Lacey &Cole (1993)

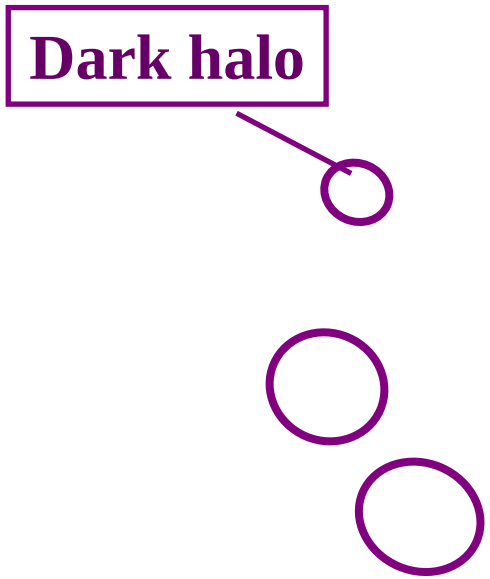
Third step

Dark halos merge with time, and galaxies in them also merge, but not in the same way, because baryons evolve also through electromagnetic interaction. If we want to incorporate galaxy properties with environment etc., a proper treatment is needed.

1.3 Formation and evolution of galaxies

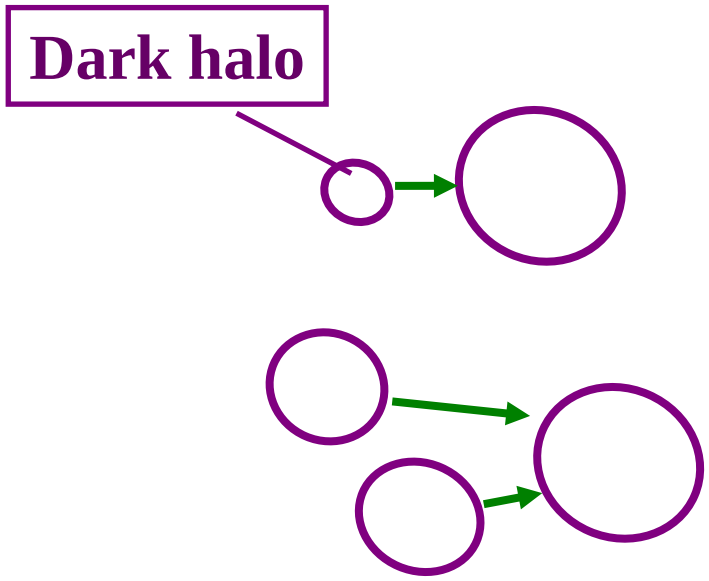
The hierarchical structure formation

The mass in the Universe is known to be dominated by DM. The initial Gaussian fluctuations of DM start to grow by gravitational interactions. Resulting virialized structures are called dark halos.



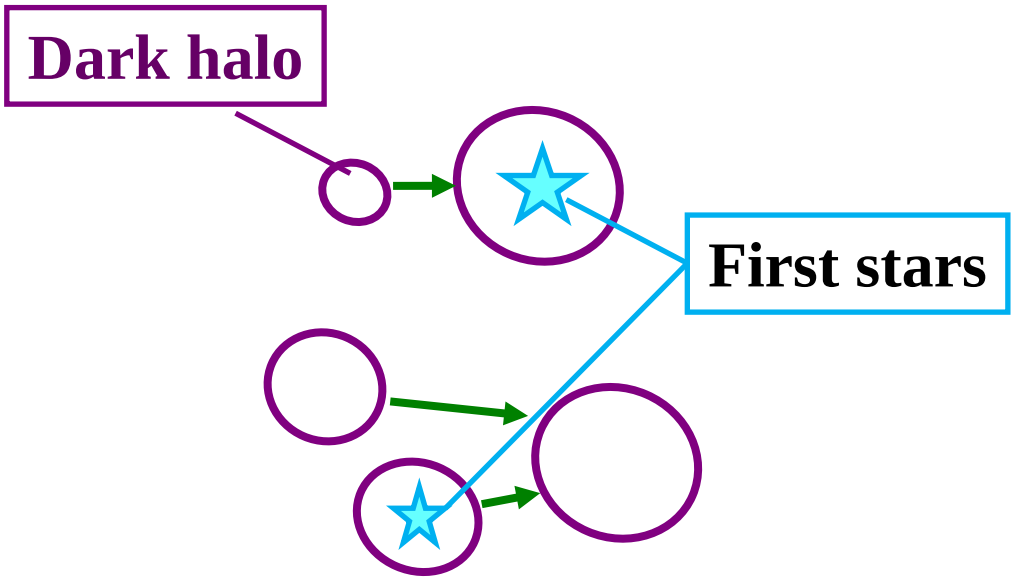
The hierarchical structure formation

The dark halos approach each other and finally merge to form larger halos. The formation proceeds from smaller to larger structures. This is the so-called hierarchical structure formation, currently the most reliable scenario of the structure formation in the Universe.



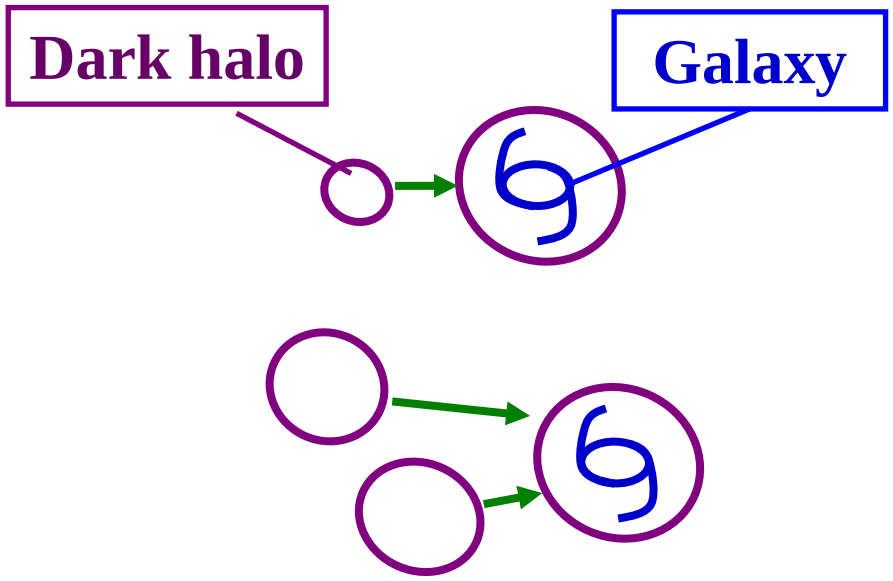
The hierarchical structure formation

During the merging of dark halos, the baryonic gas falls into the gravitational potential wells of DM and is compressed there. First stars are formed in dark halos. When they explode as supernovae, first heavy elements are provided to the Universe.



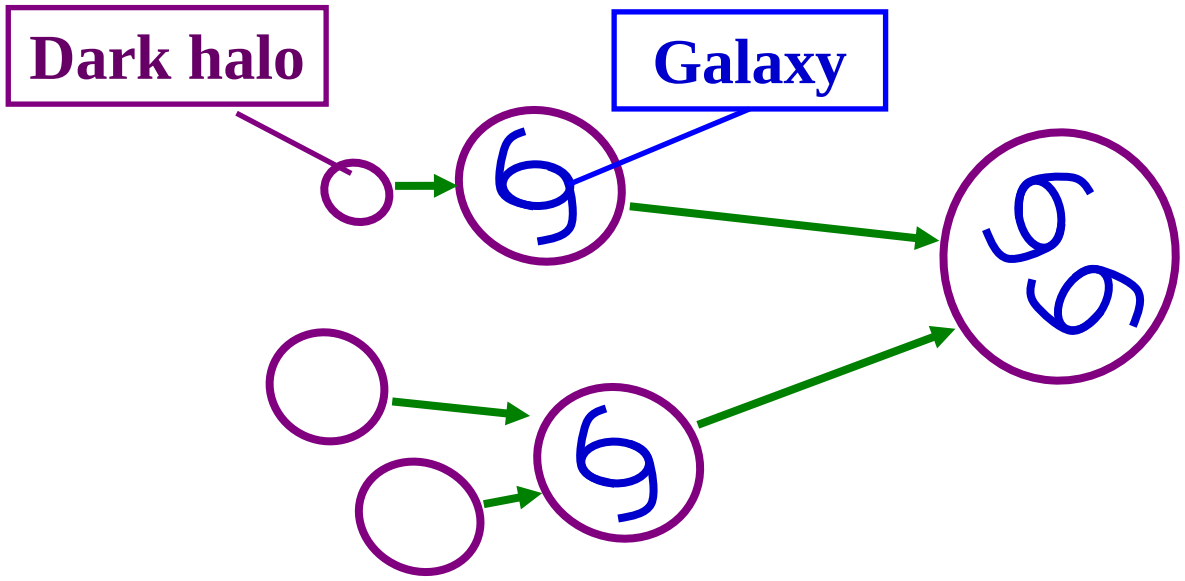
The hierarchical structure formation

The supply of heavy elements makes the condition of star formation much easier. Then, the gas turns into stars collectively, and galaxies form as large agglomerations of stars and remaining gas in dark halos.



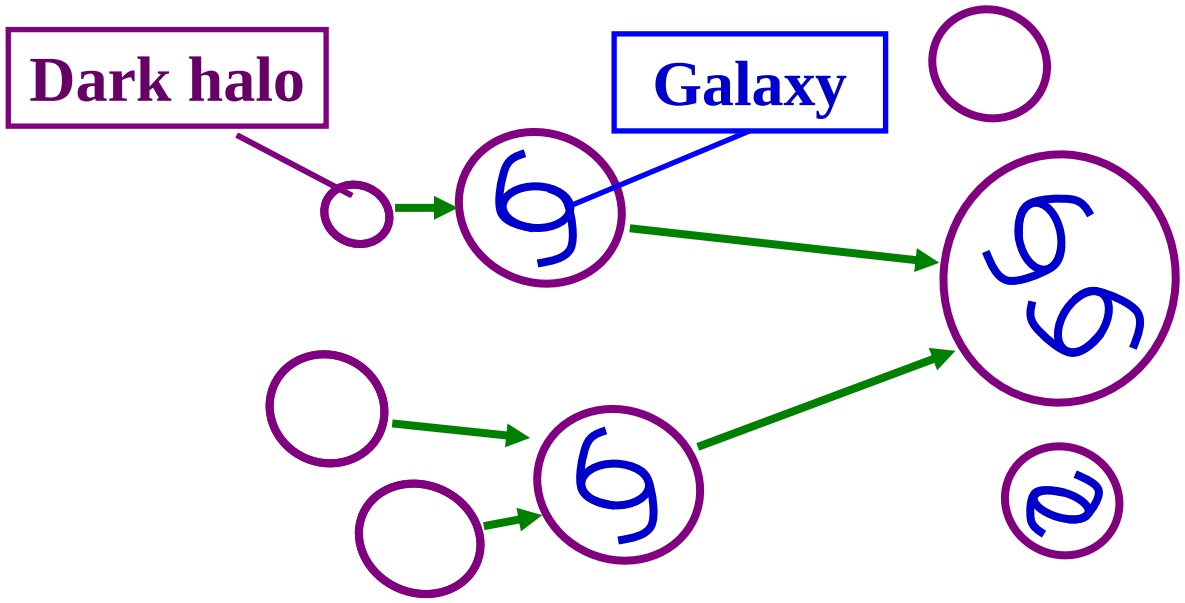
The hierarchical structure formation

Dark halos continue merging and form larger and larger halos. Consequently, galaxies in these halos start to cohabit in the same newly formed halos. Baryonic structures cannot merge as easily as dark halos because of gas pressure.



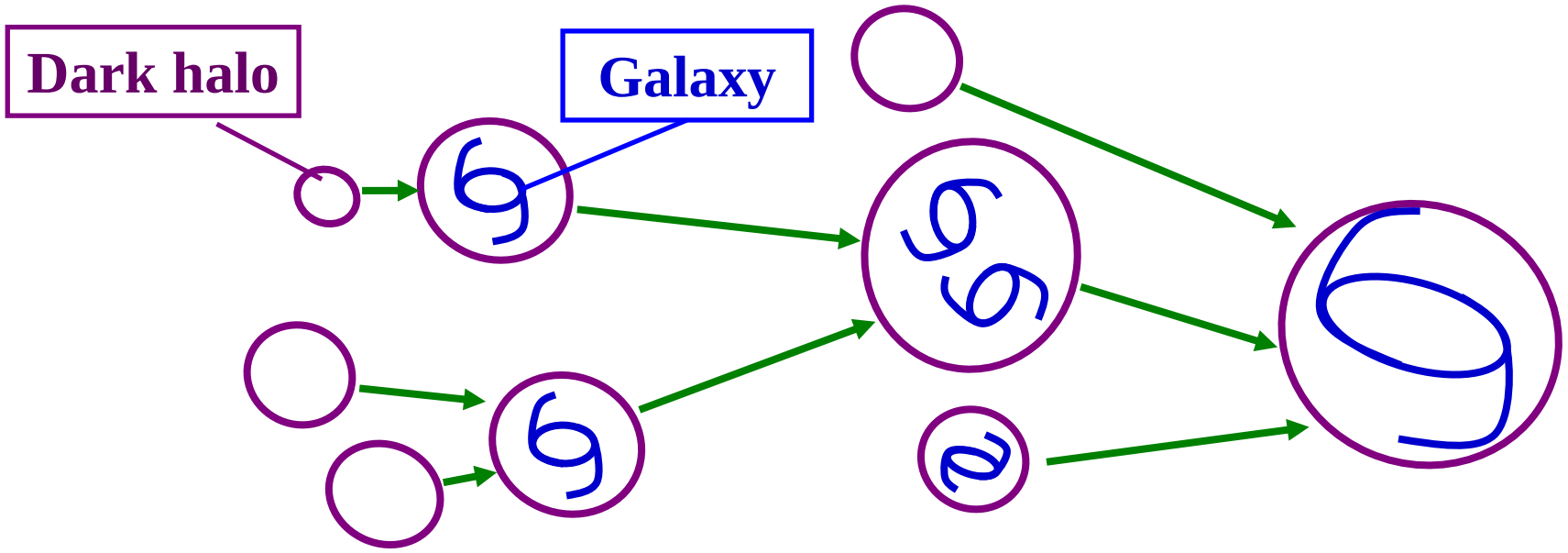
The hierarchical structure formation

Then, sometimes dark halos are occupied by one or more galaxies and sometimes no galaxies. The occupation number is stochastic (but loosely a function of the halo mass). Merging goes on with the cosmic time.

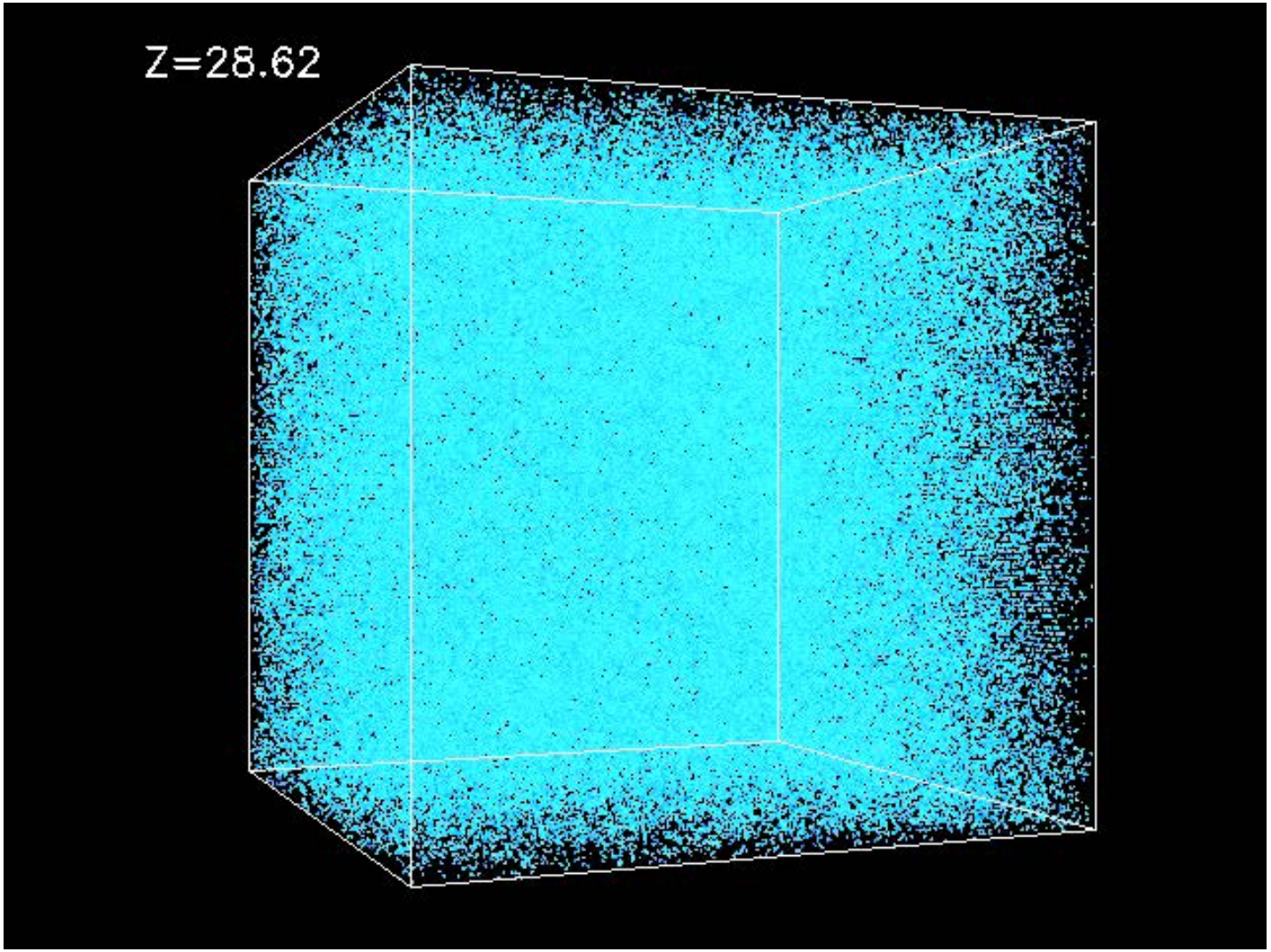


The hierarchical structure formation

Finally, some galaxies merge and form larger galaxies. Present-day large galaxies (up to $M_{\mathrm{baryon}} \sim 10^{12}\ \mathrm{M}_{\odot}$) are thought to have formed in the merger process. Strong merging process is often accompanied by an effective compression of gas, inducing burst of star formation.



Evolution of DM distribution (numerical simulation)

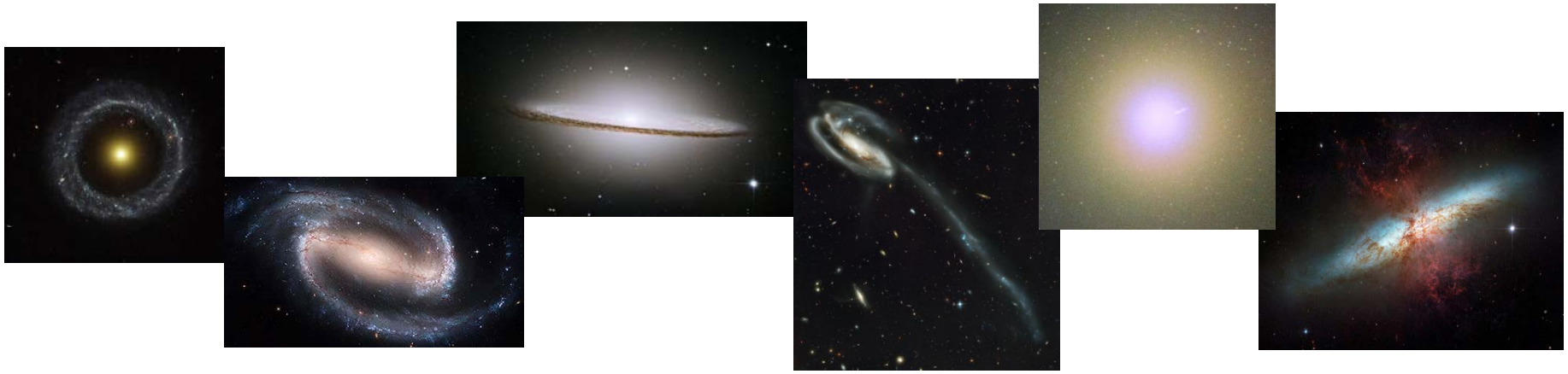


http://cosmicweb.uchicago.edu/filaments.html

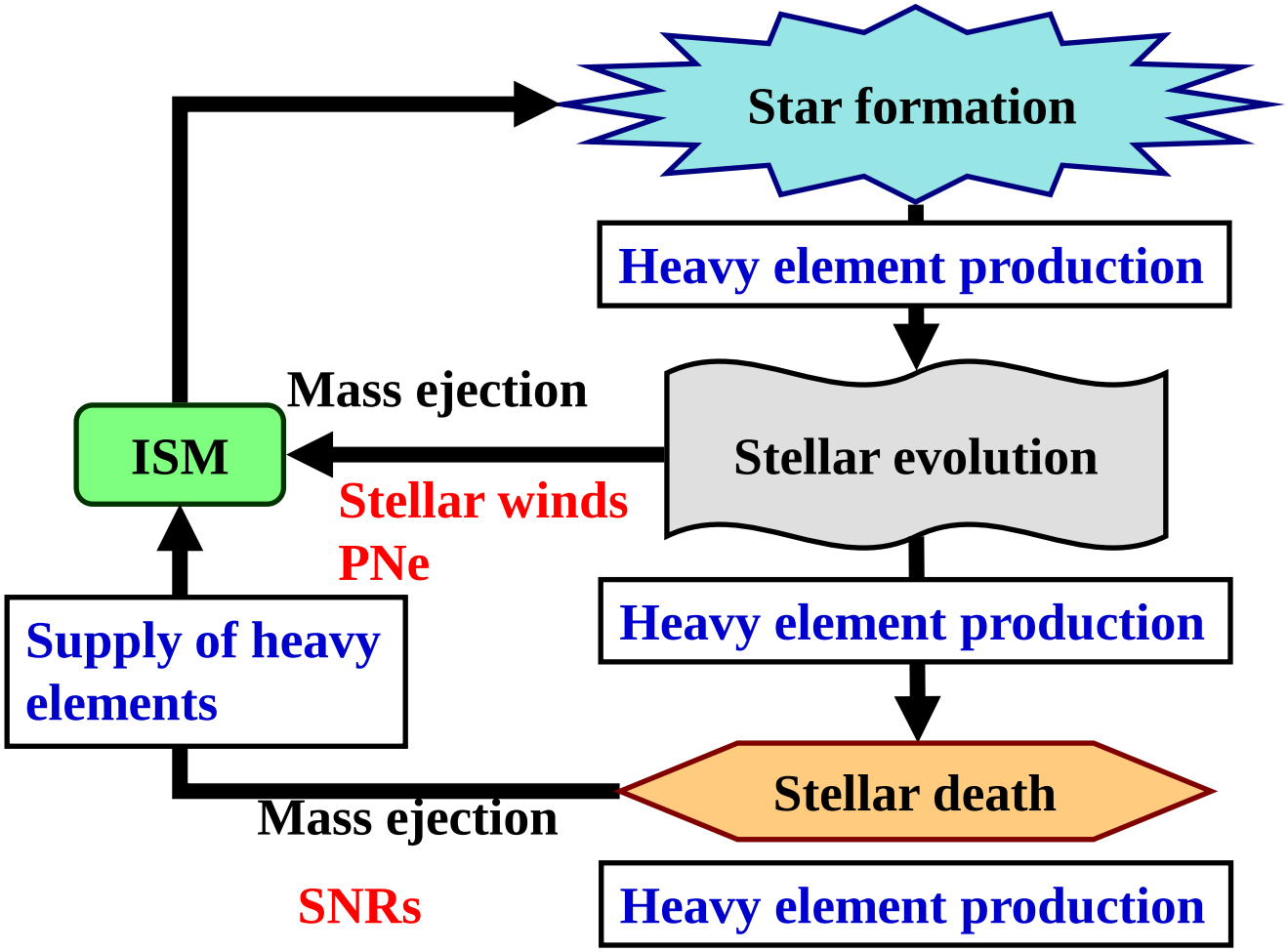
1.4 Internal galaxy evolution

Star formation in galaxies

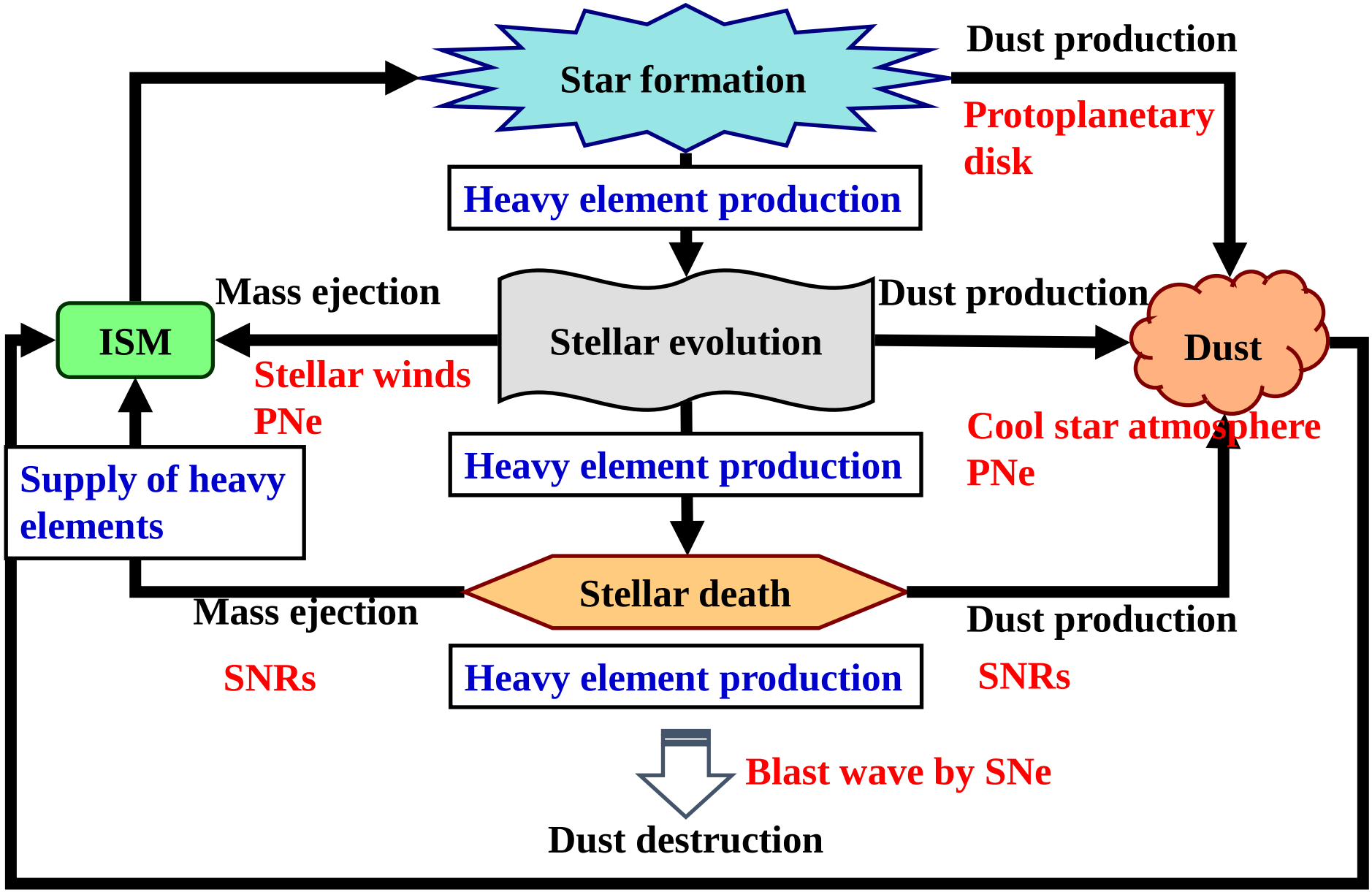
Galaxies have formed at various epochs in the Universe, merged, and grown. In parallel, gas has transformed into stars. Stars die and return back their gas into the ISM, and next generation of star formation proceeds.



Chemical evolution of galaxies: metal and dust



Chemical evolution of galaxies: metal and dust



2. Matter Distribution via Persistent Homology

2.1 Spatial distribution of matter and galaxies

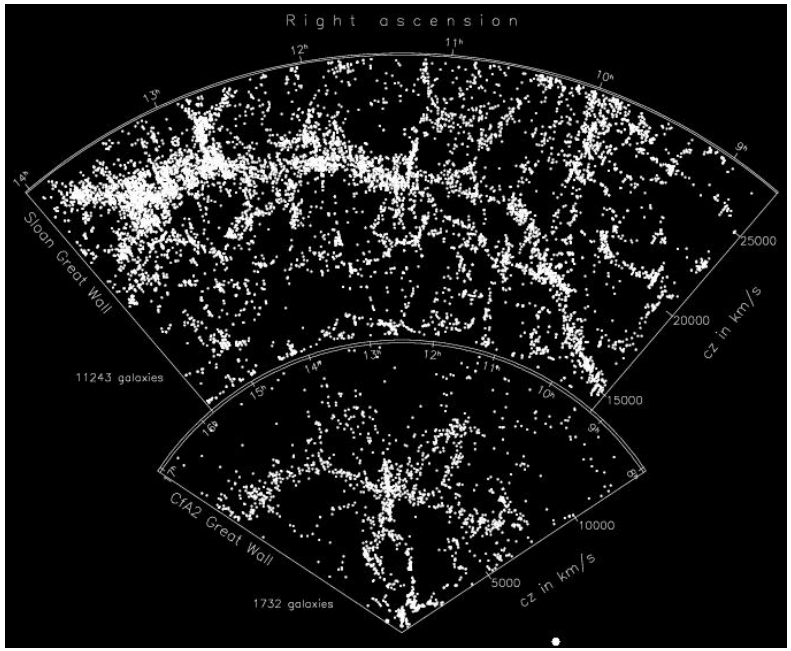
As we have already seen, galaxies have formed from a Gaussian random field of matter (DM and normal matter (baryon)).

2. Matter Distribution via Persistent Homology

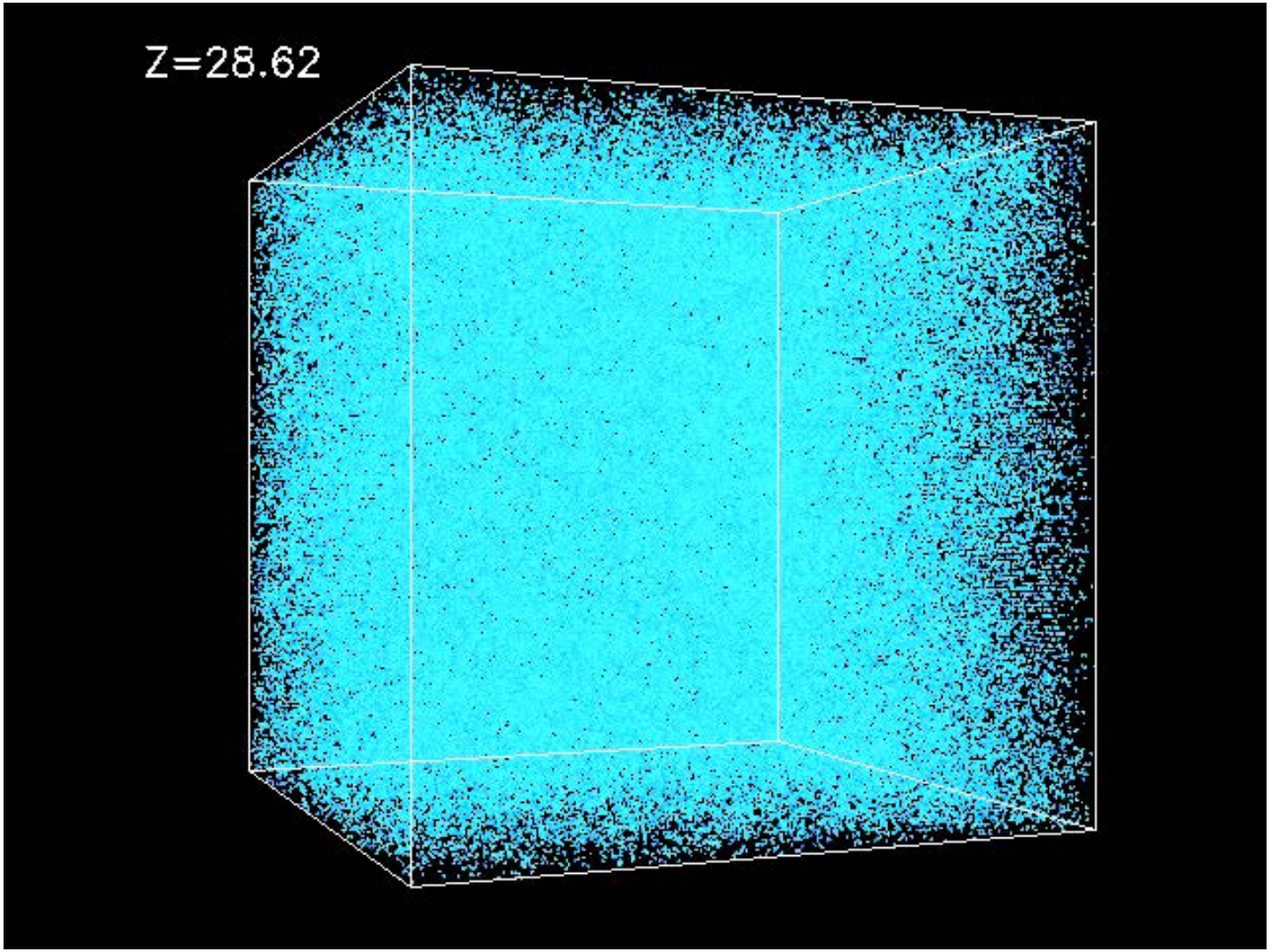
2.1 Spatial distribution of matter and galaxies

As we have already seen, galaxies have formed from a Gaussian random field of matter (DM and normal matter (baryon)).

⇒ Cosmological matter distribution at any epoch should be characterized by a statistical way.



Evolution of DM distribution revisited



http://cosmicweb.uchicago.edu/filaments.html

Traditional characterization of matter distribution

Density fluctuation: $$\delta \equiv \frac{\rho(\vec{x}) - \bar{\rho}}{\bar{\rho}} \quad \textbf{(3)}$$

Dispersion: $$\sigma^2 \equiv \left\langle \delta^2 \right\rangle \quad \textbf{(4)}$$

Fourier component: $$\delta_{\vec{k}} \equiv |\delta_{\vec{k}}| e^{i\phi_{\vec{k}}} = \int \delta(\vec{x}) e^{i\vec{k}\cdot\vec{x}} \mathrm{d}^3 x \quad \textbf{(5)}$$

Power spectrum: $$P(k) \equiv \left\langle |\delta_{\vec{k}}|^2 \right\rangle \quad \textbf{(6)}$$

Higher-order power spectra (bispectrum ⇔ 3-point correlation function; trispectrum ⇔ 4-point correlation function) are also defined.

In general, a set of infinite number of moments (or their Fourier counterparts) are needed to specify the properties of a stochastic field.

Traditional characterization of matter distribution

Density fluctuation: $$\delta \equiv \frac{\rho(\vec{x}) - \bar{\rho}}{\bar{\rho}} \quad \textbf{(3)}$$

Dispersion: $$\sigma^2 \equiv \left\langle \delta^2 \right\rangle \quad \textbf{(4)}$$

Fourier component: $$\delta_{\vec{k}} \equiv |\delta_{\vec{k}}| e^{i\phi_{\vec{k}}} = \int \delta(\vec{x}) e^{i\vec{k}\cdot\vec{x}} \mathrm{d}^3 x \quad \textbf{(5)}$$

Power spectrum: $$P(k) \equiv \left\langle |\delta_{\vec{k}}|^2 \right\rangle \quad \textbf{(6)}$$

Higher-order power spectra (bispectrum ⇔ 3-point correlation function; trispectrum ⇔ 4-point correlation function) are also defined.

Thanks to the near-Gaussian property of the matter distribution, the power spectrum still can play a central role for describing the matter distribution.

2-point correlation function

The correlation function of the density field is defined through the ensemble average

$$\xi(\mathbf{x}_1,\mathbf{x}_2)=\left\langle \delta(\mathbf{x}_1)\delta(\mathbf{x}_2)\right\rangle \qquad (7)$$

This can be expressed through the Fourier transform

$$\xi(\mathbf{x}_1,\mathbf{x}_2)=\frac{1}{(2\pi)^6}\int d^3\mathbf{k}_1 d^3\mathbf{k}_2\; e^{i(\mathbf{k}_1\mathbf{x}_1-\mathbf{k}_2\mathbf{x}_2)}\left\langle \delta(\mathbf{k}_1)\delta^*(\mathbf{k}_2)\right\rangle \qquad (8)$$

Using the translational invariance in Fourier space, we get

$$\xi(\mathbf{x}_1,\mathbf{x}_2)=\frac{1}{(2\pi)^3}\int d^3\mathbf{k}_1 d^3\mathbf{k}_2\; e^{i(\mathbf{k}_1\mathbf{x}_1-\mathbf{k}_2\mathbf{x}_2)}\delta^{(3)}(\mathbf{k}_1-\mathbf{k}_2)P(k_1) \qquad (9)$$

Since the correlation function only depends on the distance,

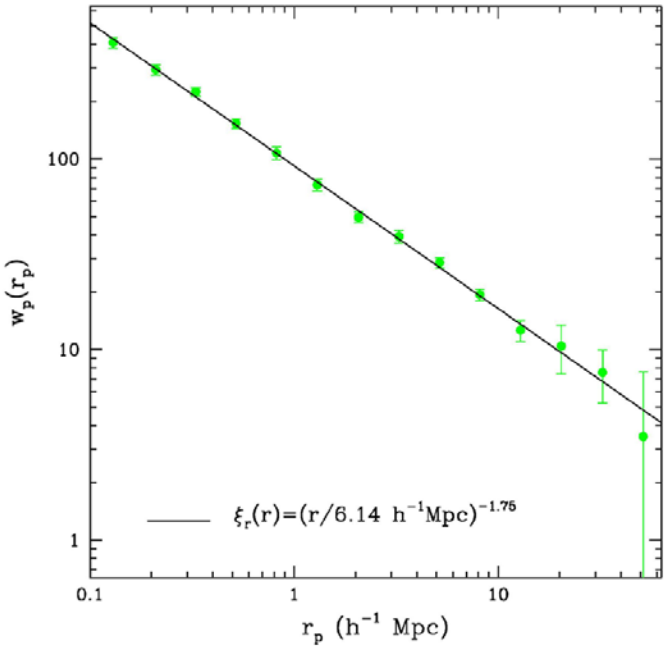
$$\xi(\mathbf{x}_1,\mathbf{x}_2)=\int d^3\mathbf{k}\; e^{i\mathbf{k}(\mathbf{x}_1-\mathbf{x}_2)}P(k)=\xi(\mathbf{x}_1-\mathbf{x}_2) \qquad (10)$$

2-point correlation function

- **First measured by Totsuji & Kihara (1969), then Peebles et al.**
- **A power law is a fairly good approximation (but no strong physical reason)**

$$\xi(r) = \left(\frac{r}{r_0} \right)^{-\gamma} \qquad (11)$$

- **Correlation length r_0= 5.4 h^{-1} Mpc ***
- **Exponent is around γ = 1.8**



*** From the latest measurement, h = 0.7.**

2-point correlation function

- **First measured by Totsuji & Kihara (1969), then Peebles et al.**
- **A power law is a fairly good approximation (but no strong physical reason)**

$$\xi(r) = \left(\frac{r}{r_0} \right)^{-\gamma} \qquad \textbf{(11)}$$

- **Correlation length r_0= 5.4 h^{-1} Mpc**
- **Exponent is around γ = 1.8**

Limitation

To evaluate the statistical properties beyond the Gaussian approximation, we have to introduce a series of infinitely high correlation functions, which is not realistic.

⇒ More flexible method is desired!

2.2 Baryon acoustic oscillation (BAO)

Baryons evolve in a very complicated way via electromagnetic interactions (with radiative heating/cooling, gas pressure, fluid dynamical processes, etc.)

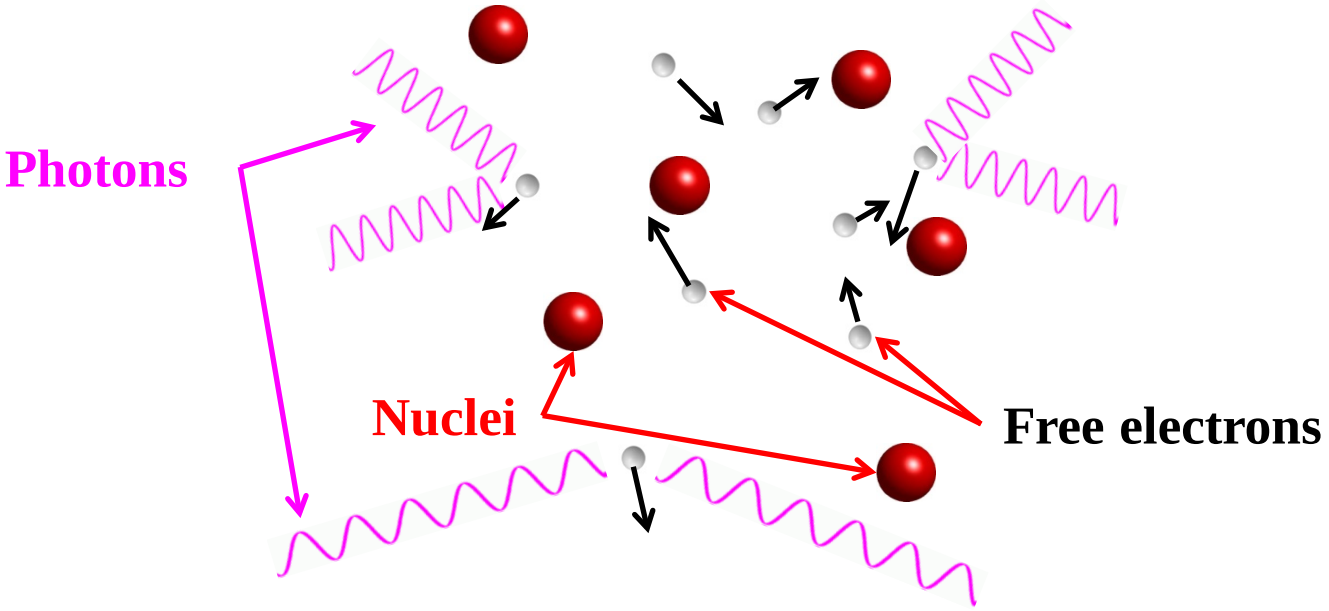
A typical example of such a nontrivial phenomenon is the baryon acoustic oscillation (BAO) generated in the matter-radiation fluid in the early Universe (Peebles & Yu 1970; Sunyaev & Zel'dovich 1970).

State of matter in the early Universe

Matter and radiation were tightly coupled by Thomson scattering

⇒ hot dense matter-photon plasma (two fluid)

$T > 3000$ K

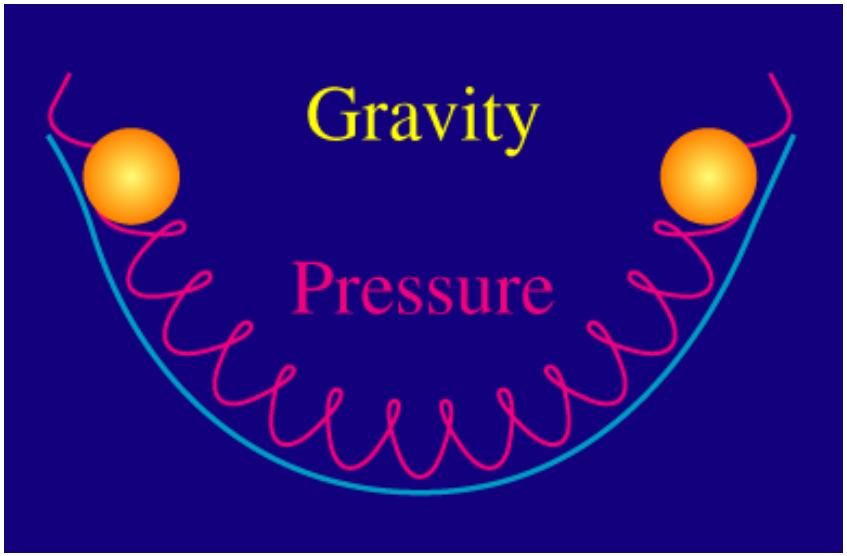


Microscopically, photons are scattered by free electrons and cannot go straight.

Acoustic peak

Consider a point-like initial perturbation in the primordial matter-photon plasma. In the plasma, the matter and photons are locked into a single fluid.

Since the photons are so hot and numerous, the combined fluid has an tremendous pressure with respect to its density.



http://background.uchicago.edu/~whu/physics/physics.html

Acoustic peak

Consider a point-like initial perturbation in the primordial matter-photon plasma. In the plasma, the matter and photons are locked into a single fluid.

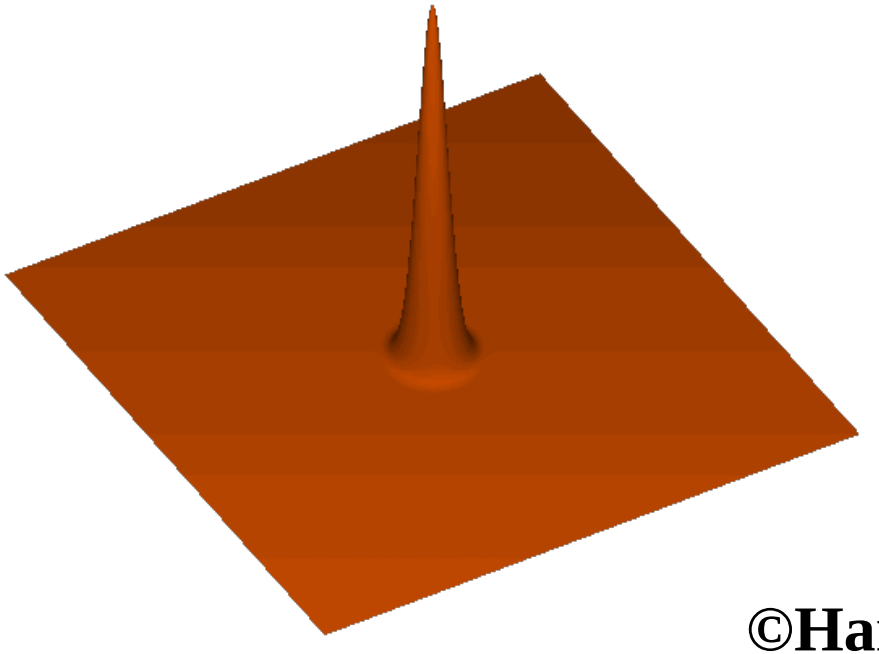
Since the photons are so hot and numerous, the combined fluid has an tremendous pressure with respect to its density.

The pressure tries to equalize itself with the surroundings, which results in an expanding spherical sound wave. The sound speed at this early epoch is evaluated by

$$c_{\mathrm{s}} = \frac{c}{\sqrt{3\left(1 + \frac{3\rho_{\mathrm{bar}}}{4\rho_{\mathrm{rad}}}\right)}} \qquad (12)$$

which is ~ 57 % of the light speed at early epoch.

Acoustic peak



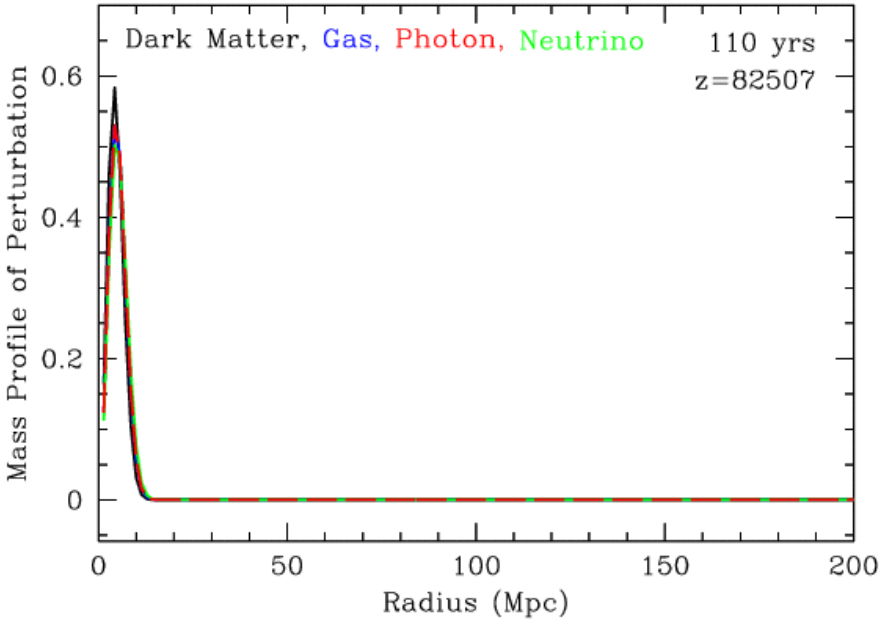
©Harvard University

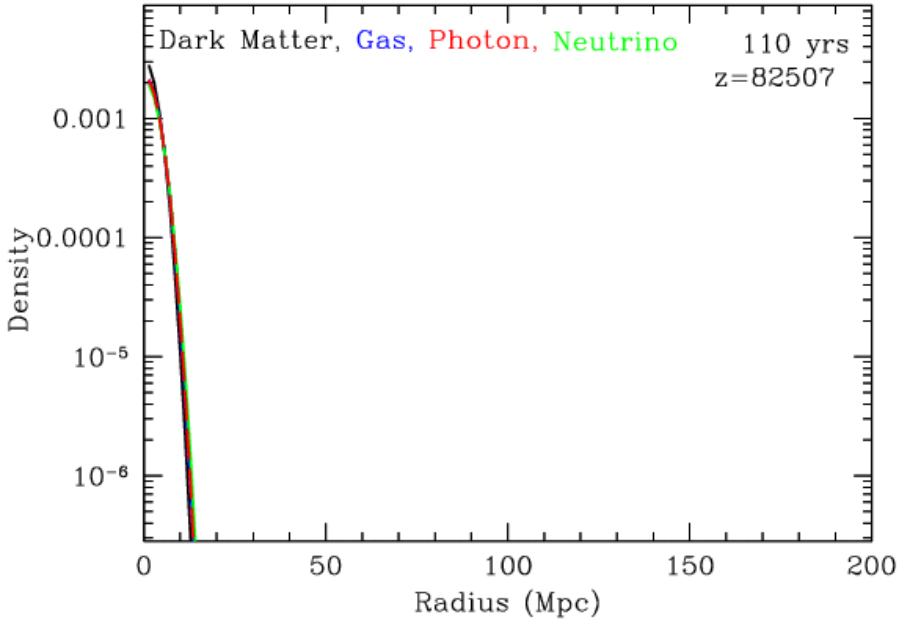
Sound horizon (final radius) r_s is obtained by

$$r_s = \int_0^{t_{dec}} (1+z) c_s \, dt = \int_{z_{dec}}^{\infty} \frac{c_s}{H_0 \sqrt{\Omega_{r,0}(1+z)^4 + \Omega_{m,0}(1+z)^3 + \Omega_{\Lambda 0}}} dz \quad \textbf{(13)}$$

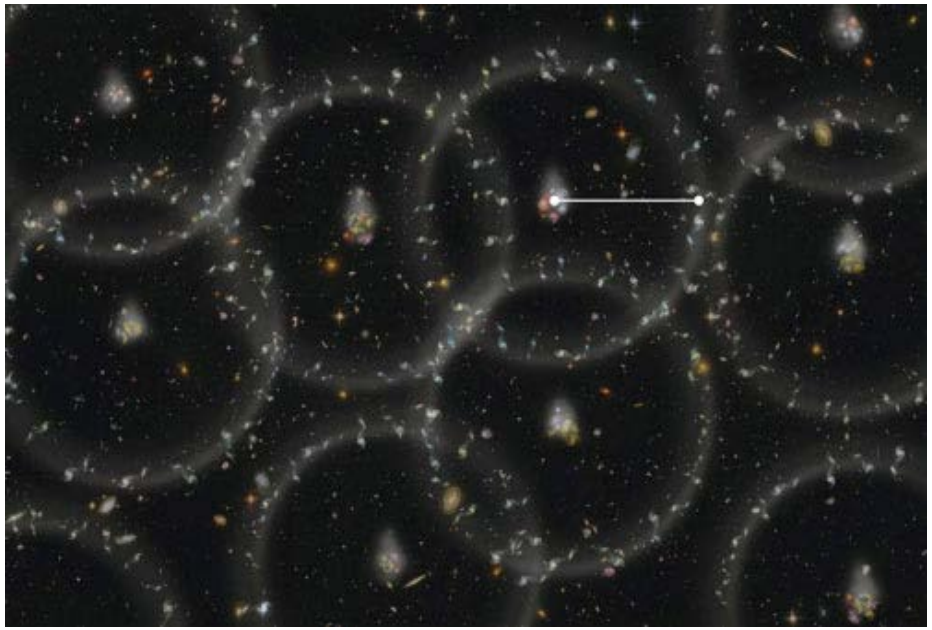
where dec stands for the decoupling.

Acoustic peak

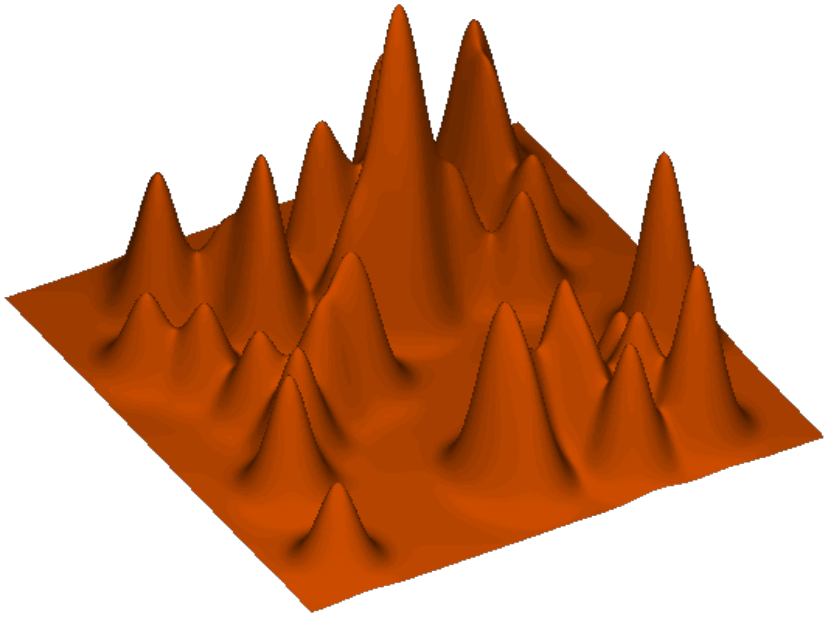




BAO



© UCLA



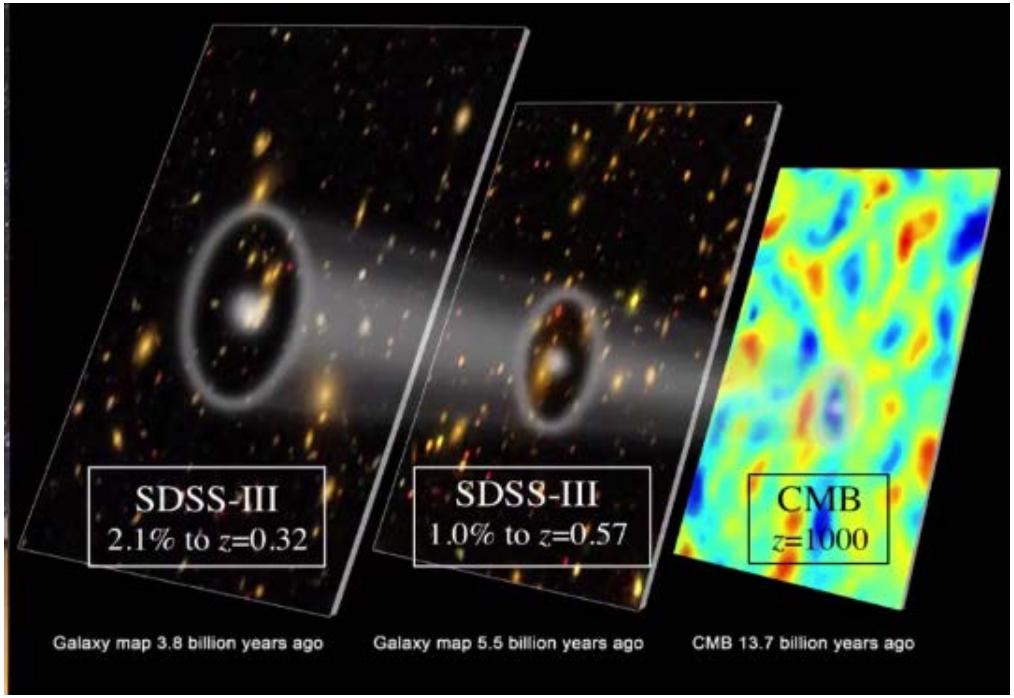
©Harvard University

What we observe is the superposition of many acoustic waves imprinted on the large-scale structure emerged from the primordial fluctuations.

BAO on the correlation function

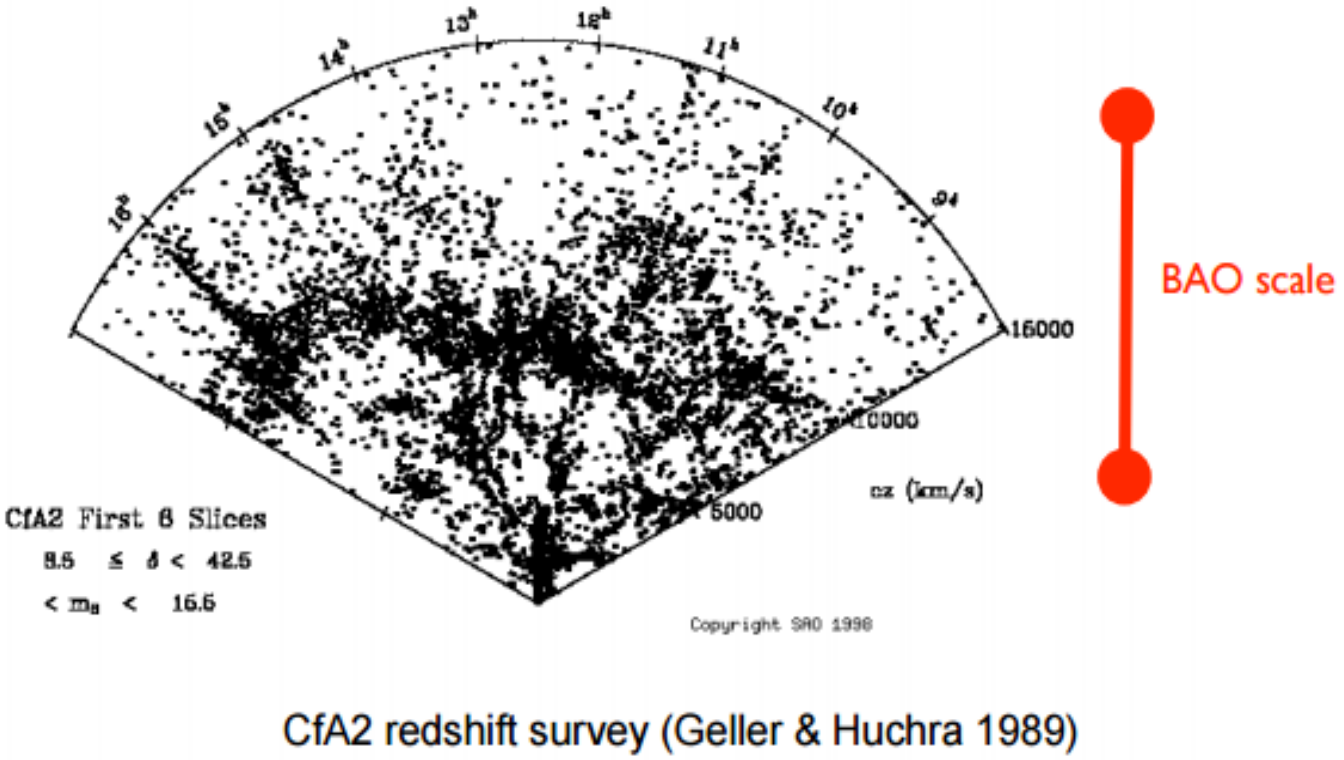
The BAO length scale is constant in comoving coordinates.

⇒ In principle, we can detect the signal on the galaxy 2-point correlation function.



© E.M. Huff, the SDSS-III team, and the South Pole Telescope team

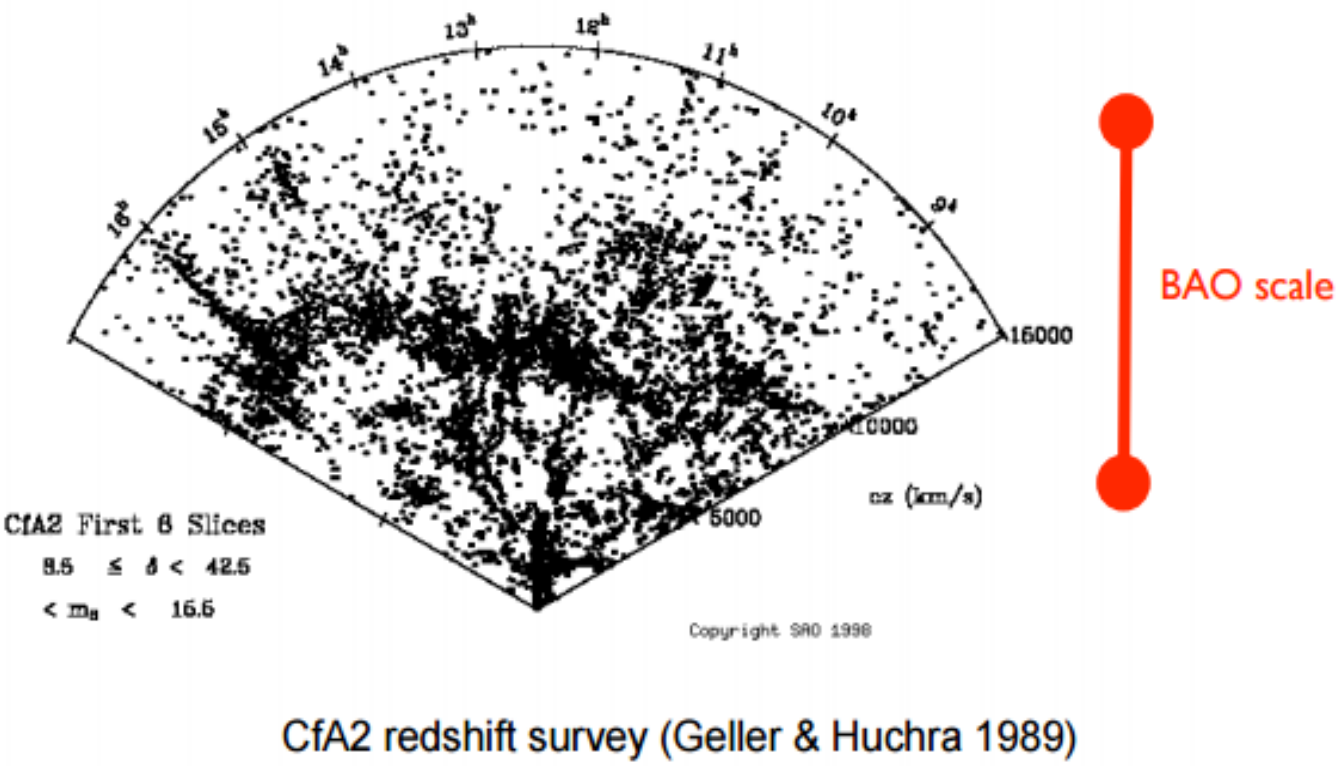
BAO on the correlation function



CfA2 redshift survey (Geller & Huchra 1989)

However… the BAO scale is very large compared with the typical scale of the large-scale structure.

BAO on the correlation function



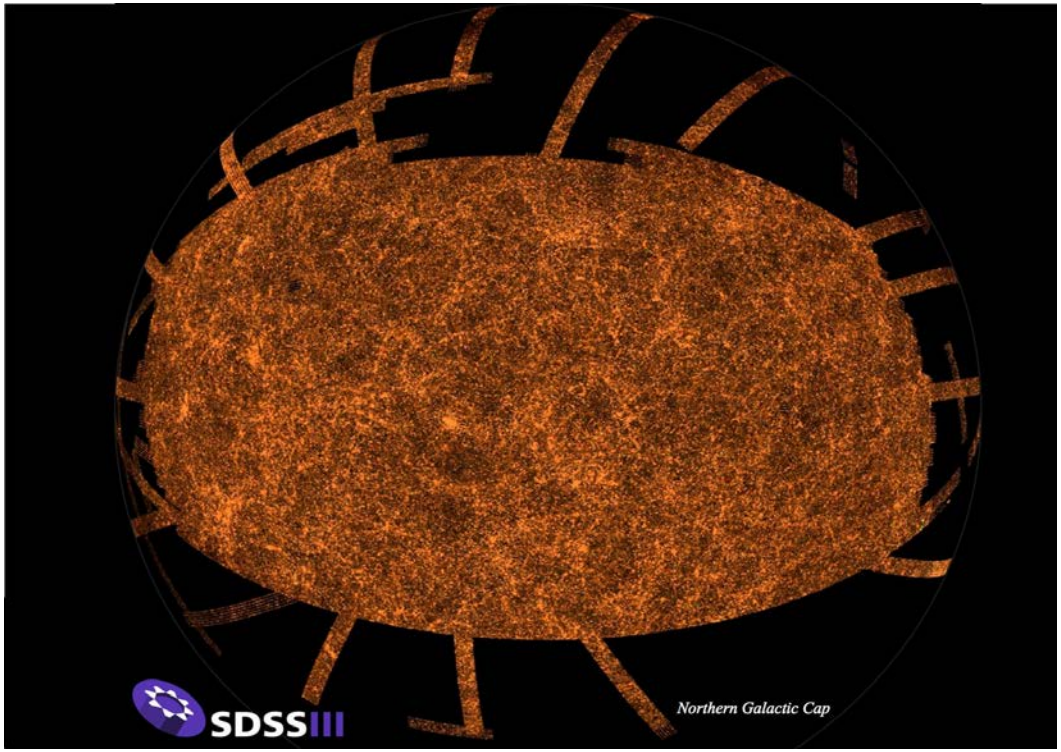
CfA2 redshift survey (Geller & Huchra 1989)

However… the BAO scale is very large compared with the typical scale of the large-scale structure.

⇒ We need a very large galaxy sample with dense sampling, since we must measure the signal at such a large scale.

Sloan Digital Sky Survey (SDSS)

The SDSS is the largest optical photometric and spectroscopic surveys ever existed. It covers one-third of the sky.

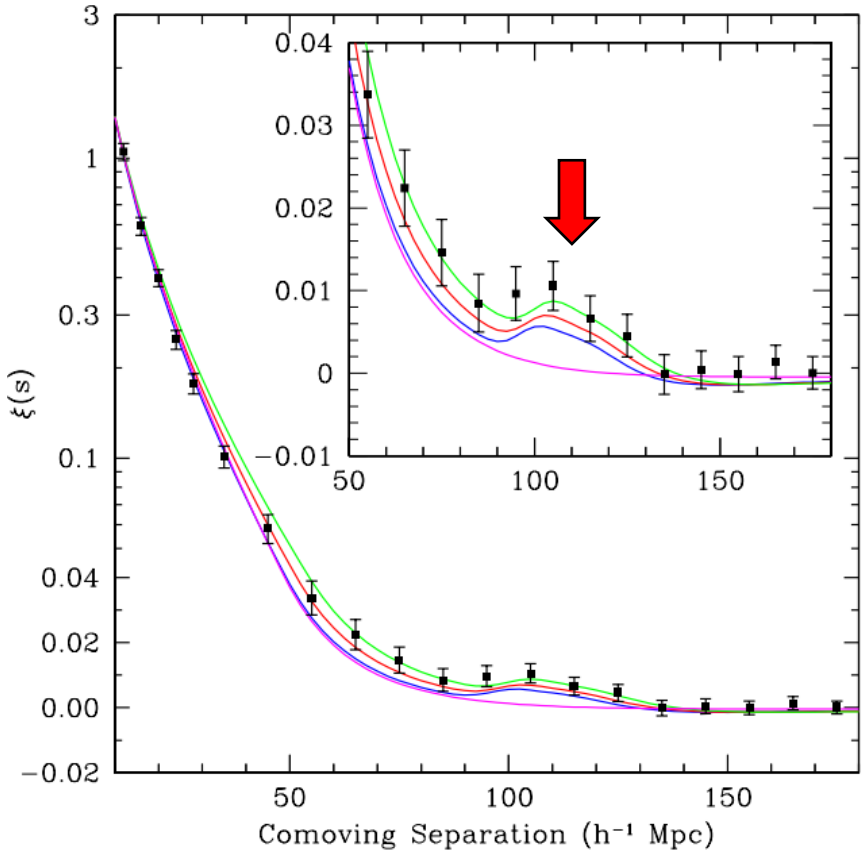


The advent of the SDSS finally made it realistic to detect the BAO signal on the 2-point correlation function.

First detection of the BAO signal on the correlation function

Eisenstein et al. (2005) first detected the signal around 150 Mpc on the 2-point correlation function.

Currently this analysis is only possible with SDSS data.



Eisenstein et al. (2005)

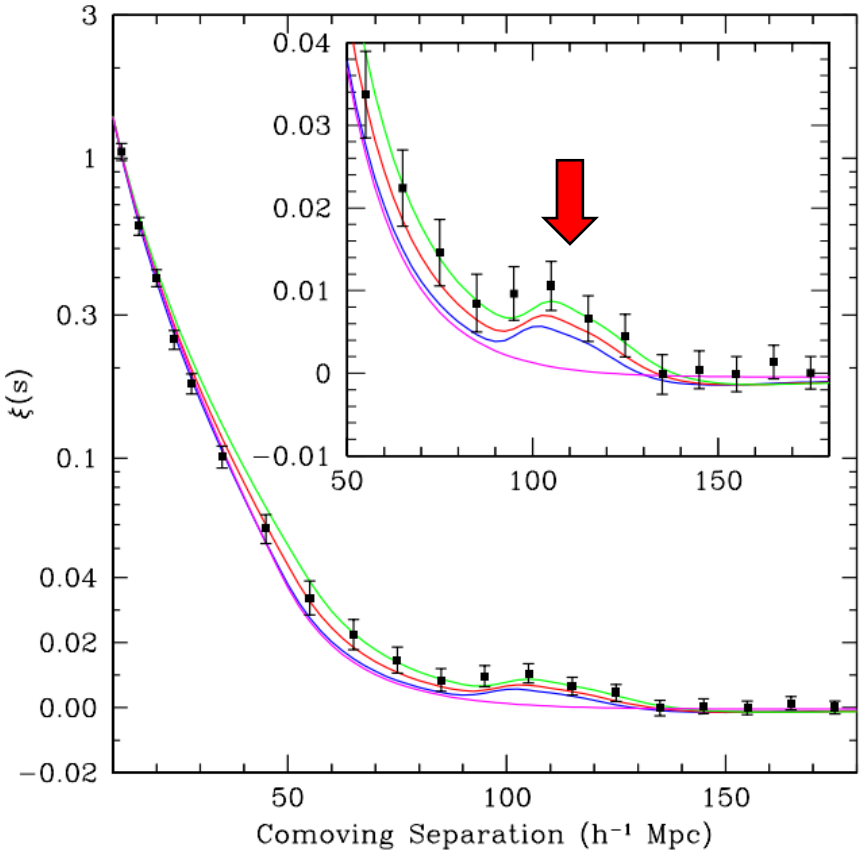
First detection of the BAO signal on the correlation function

Eisenstein et al. (2005) first detected the signal around 150 Mpc on the 2-point correlation function.

Currently this analysis is only possible with SDSS data.

More flexible method is desired!

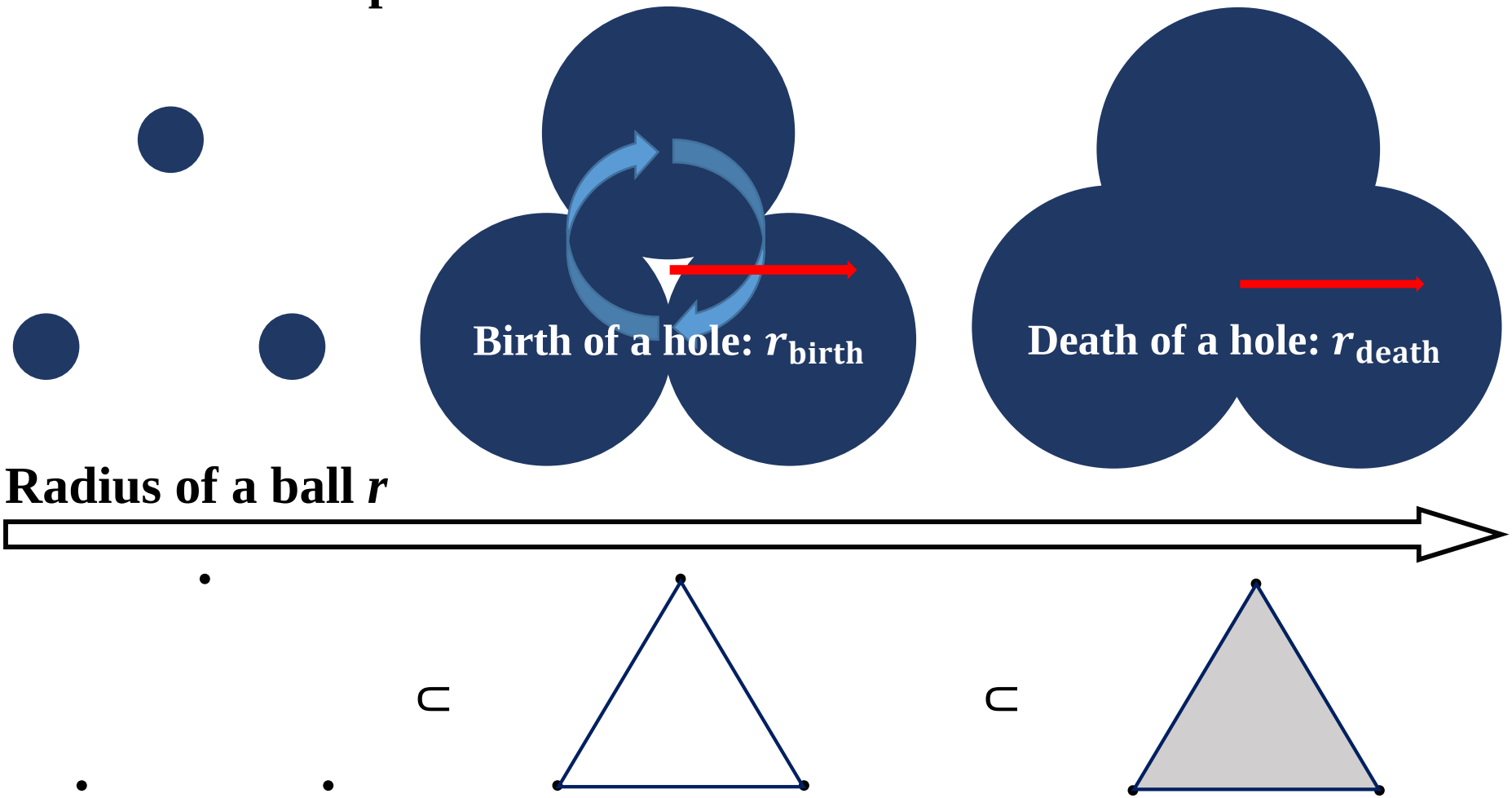
⇒ Persistent homology!



Eisenstein et al. (2005)

2.3 Persistent homology

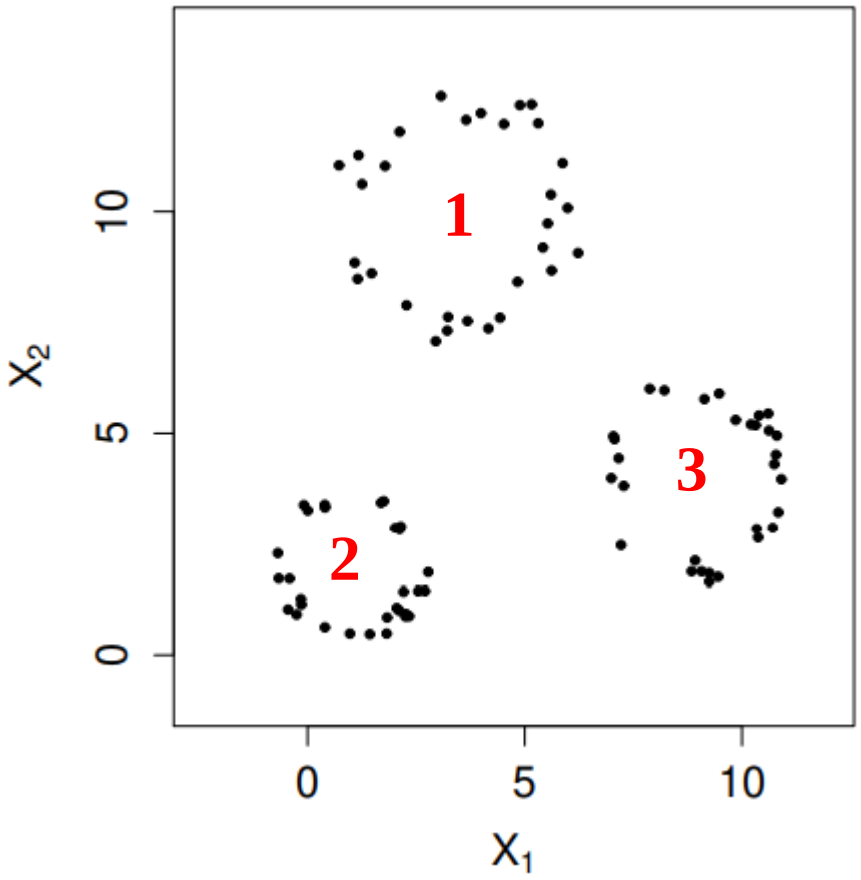
The topological information is characterized with holes constructed from *n*-dimensional sphere with radius *r* from discrete data points.

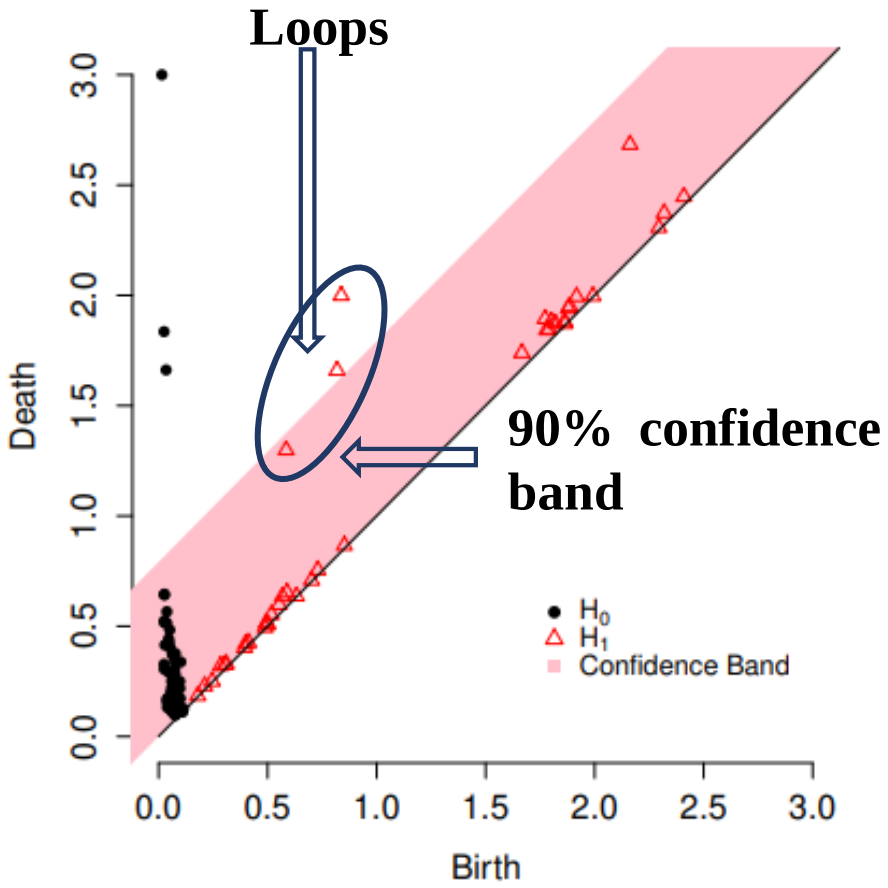


Persistent diagram: finding loops from data

Data with three loops

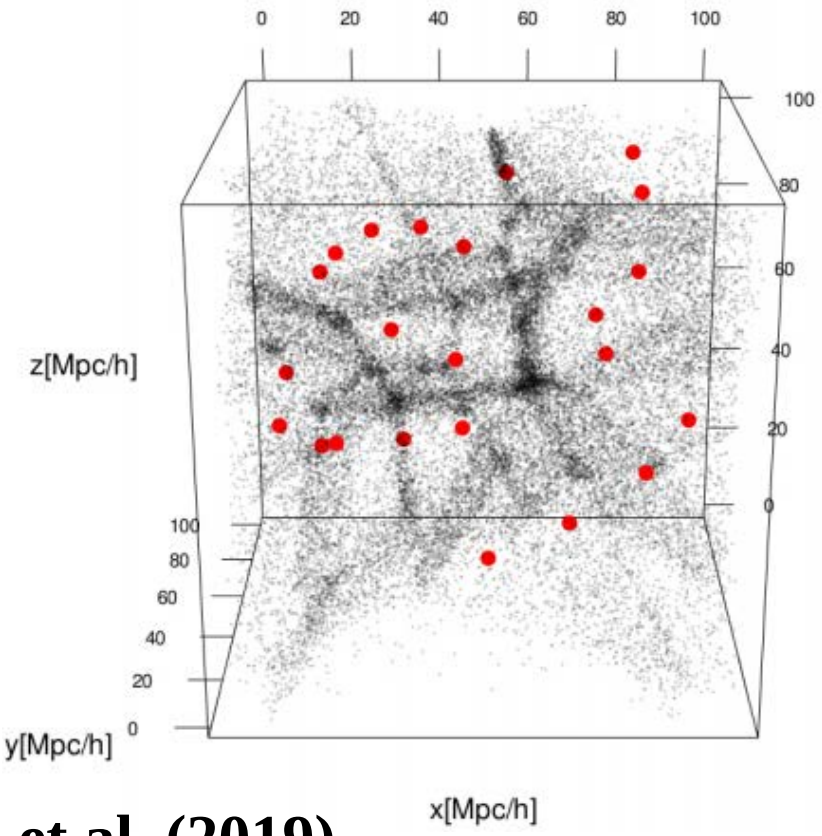
⇒ From persistent diagram, we can find three significant structures.

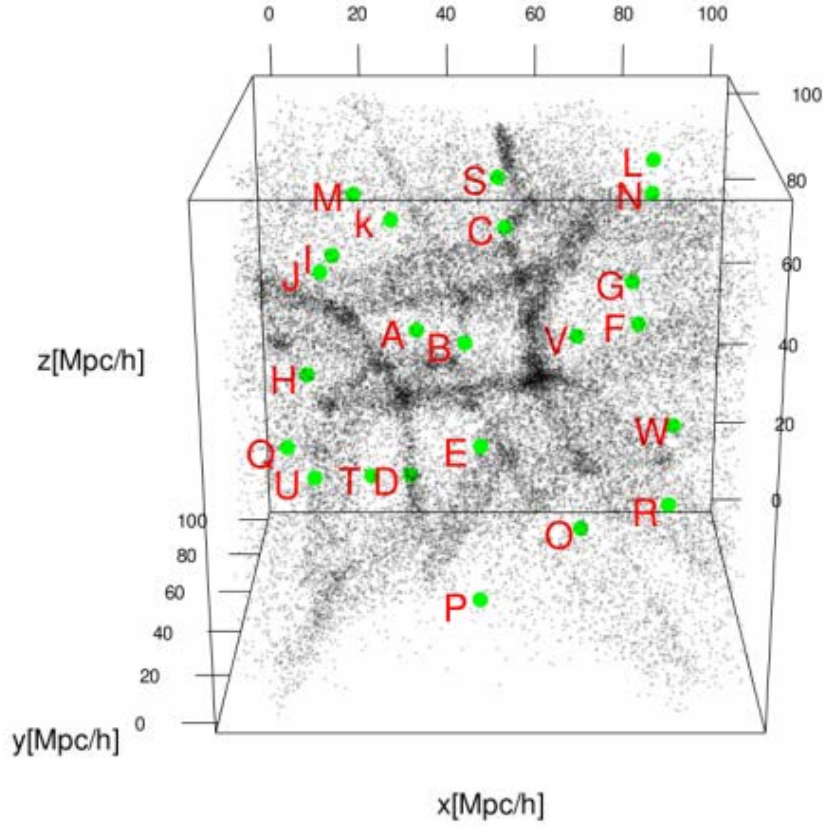




Preceding work: void finding (Xu et al. 2019)

Xu et al. (2019) applied the persistent homology for a large galaxy data set generated from *N*-body simulation (Libeskind et al. 2018) and found 23 voids (A-W) in the simulation data.

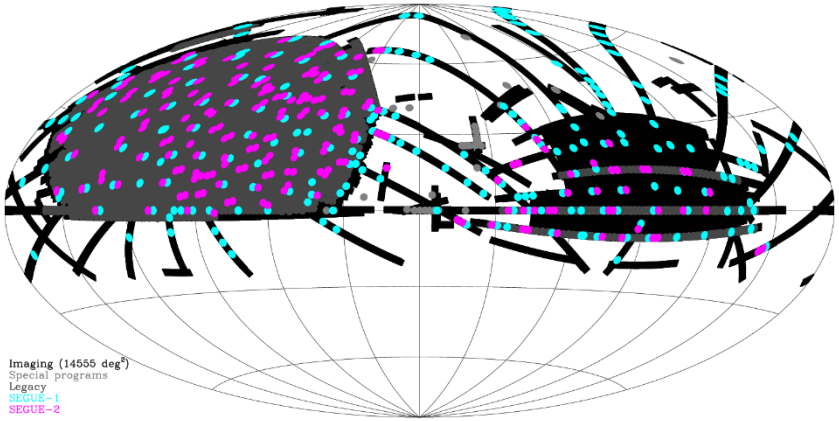




Xu et al. (2019)

Local galaxy data ($z \sim 0$)

SDSS DR12 (Alam et al. 2015) Band: *u, g, r, i, z*
Sky coverage area: 14555 deg^2

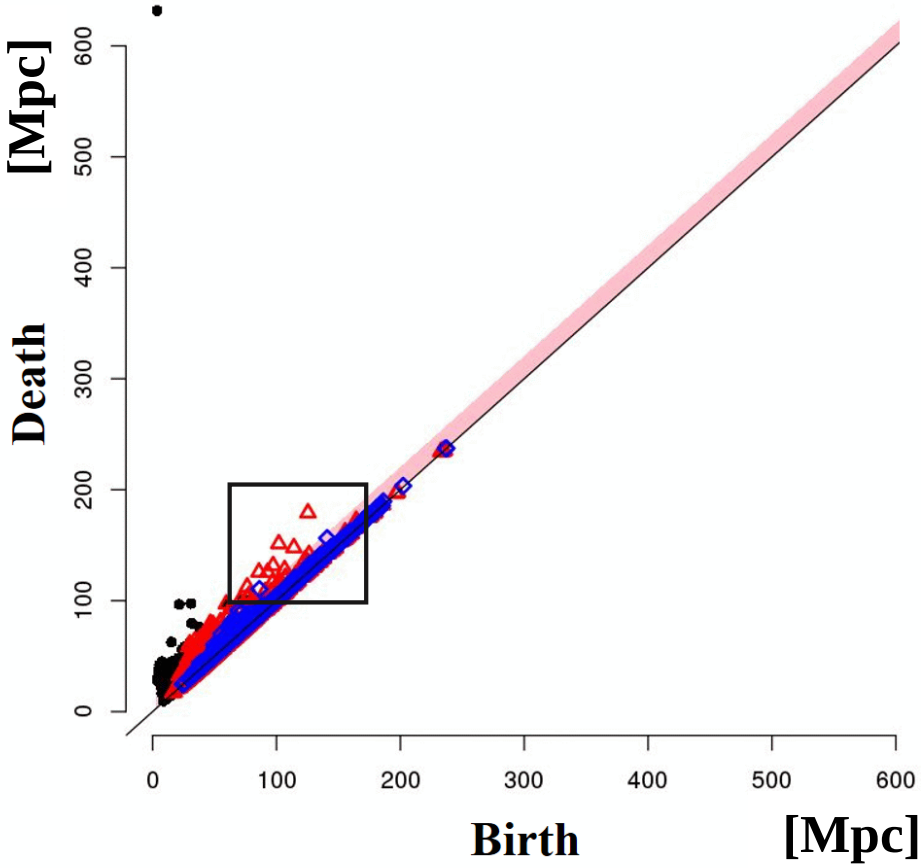


Luminous red galaxies (LRG)

Relatively massive (~ $10^{11\text{-}12}$ $M_\odot$) elliptical galaxies. In this study, we followed selection criteria suggested by Eisenstein et al. (2001). We further randomly sampled 2000 galaxies.

***N.B.* We do not use the full dataset but a sparsely drawn subsample of the SDSS.**

Result at $z \sim 0$



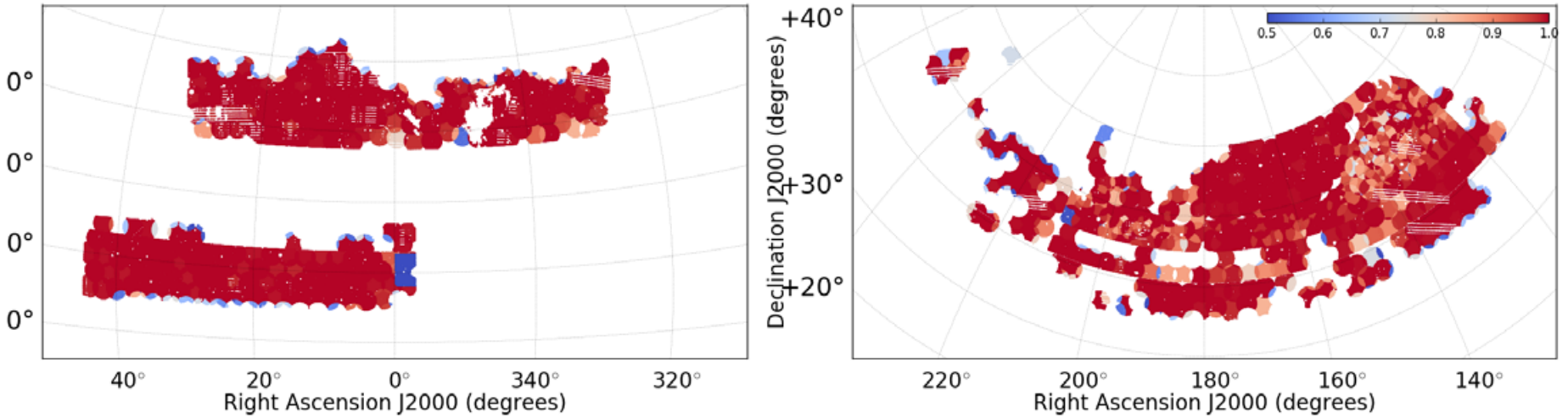
There are 10 loops above the 90% confidence band whose r_{death} is ~ 150 Mpc. This exactly corresponds to the BAO signal.

⇒ Much more efficient than the correlation function method!

Quasar data at z ~ 0-3: explore the evolution

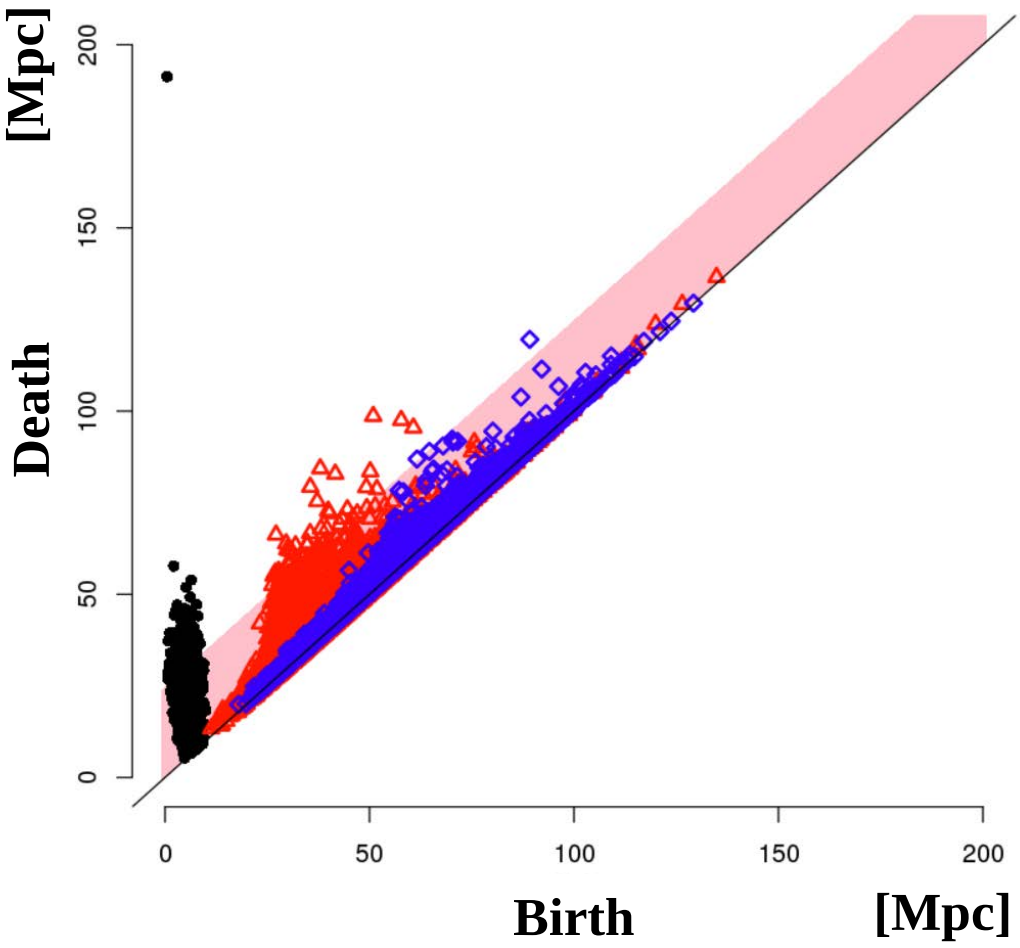
SDSS DR14 quasar catalog (Bautista et al. 2018)

Quasars are though to form in very high density peaks of the fluctuation field. Again we construct a subsample of size 2000 extracted from the parent sample.



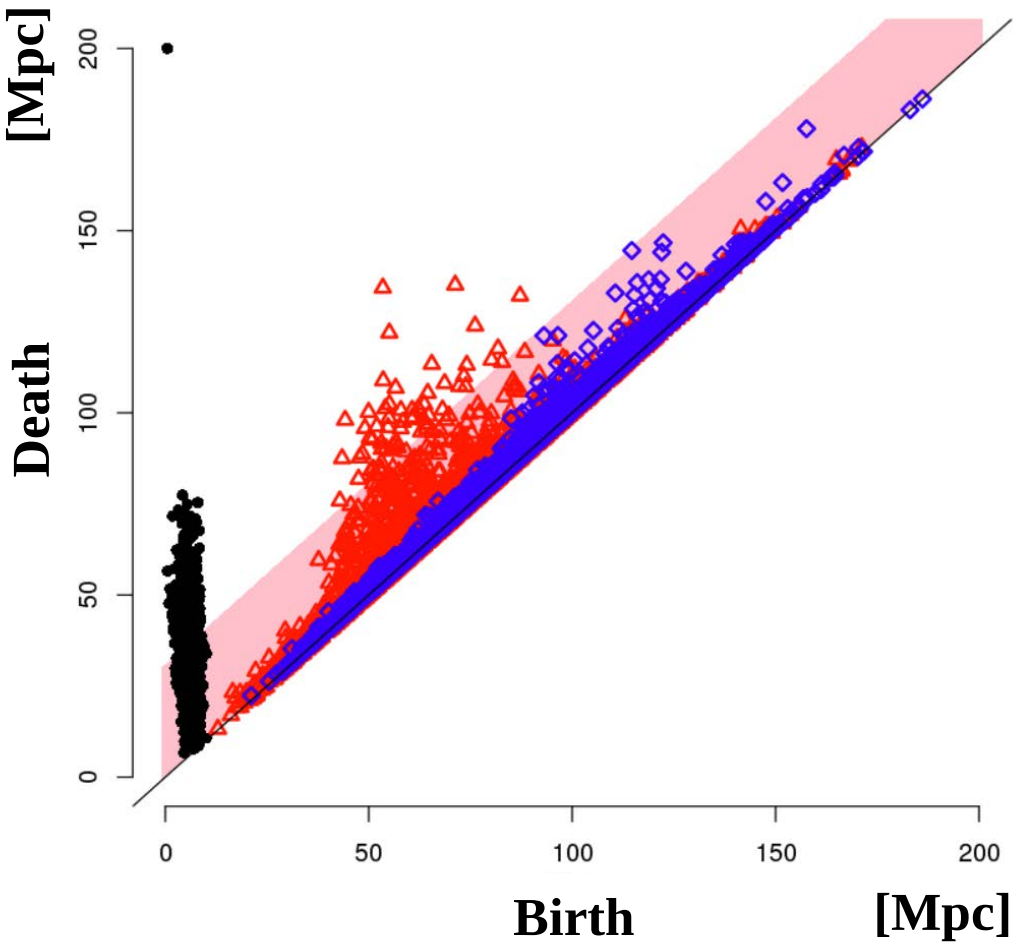
Bautista et al. (2018)

Result at $0 < z < 1$

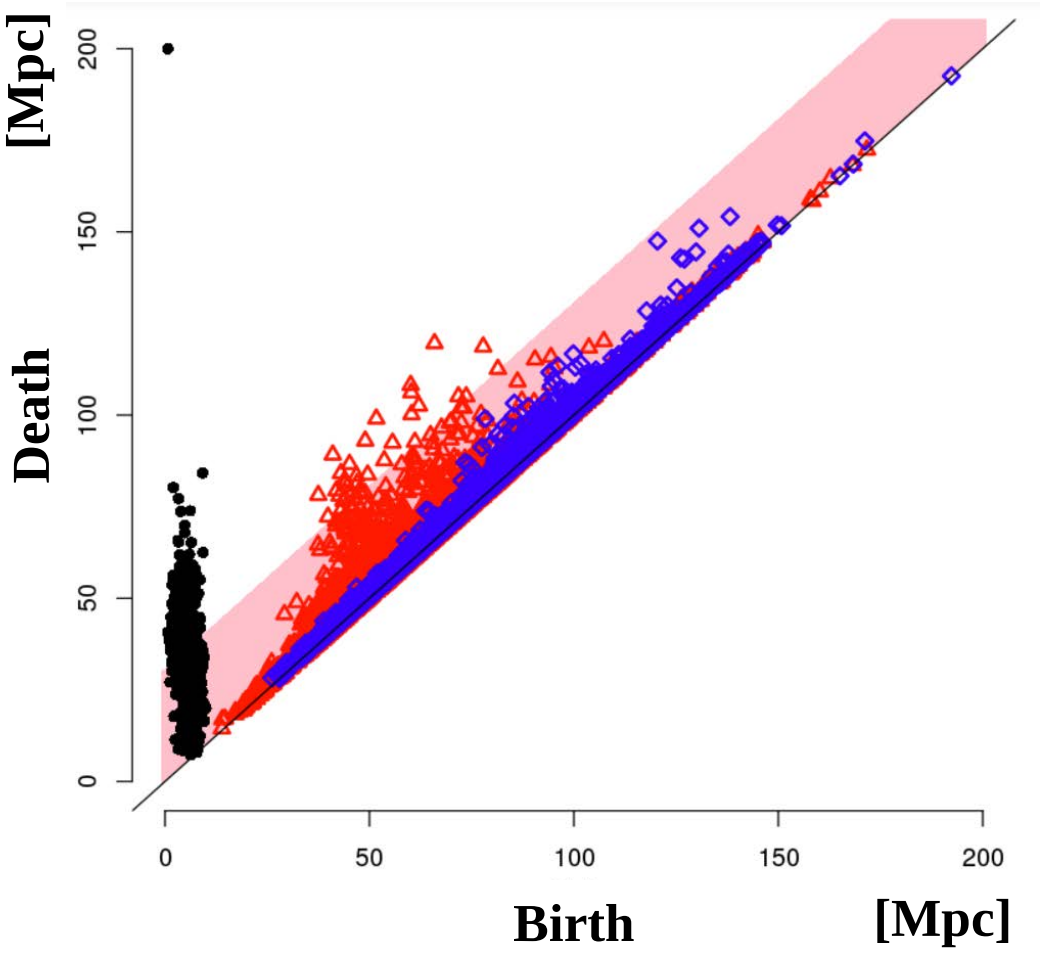


Again we see clear signatures of BAO.

Result at $1 < z < 2$



Result at $2 < z < 3$



However, it is not easy to evaluate the evolution, i.e., compare the difference between different redshifts. We plan to introduce Wasserstein metric to quantify it.

3. Galaxy Evolution through Manifold Learning

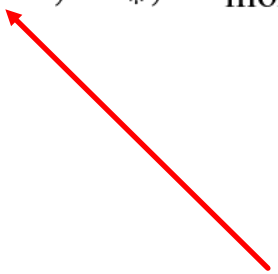
3.1 Galaxy evolution from a modern point of view

Galaxies evolve in various aspects:

3.1 Galaxy evolution from a modern point of view

Galaxies evolve in various aspects:

$$
\begin{aligned}
\mathrm{SFR}(t) &= f_1(\mathrm{SFR}, M_*, M_{\mathrm{mol}}, M_{\mathrm{HI}}, M_{\mathrm{dust}}, M_{\mathrm{halo}}, \delta_{\mathrm{gal}}, \ldots) \\
M_*(t) &= f_2(\mathrm{SFR}, M_*, M_{\mathrm{mol}}, M_{\mathrm{HI}}, M_{\mathrm{dust}}, M_{\mathrm{halo}}, \delta_{\mathrm{gal}}, \ldots) \\
M_{\mathrm{mol}}(t) &= f_3(\mathrm{SFR}, M_*, M_{\mathrm{mol}}, M_{\mathrm{HI}}, M_{\mathrm{dust}}, M_{\mathrm{halo}}, \delta_{\mathrm{gal}}, \ldots) \\
M_{\mathrm{HI}}(t) &= f_4(\mathrm{SFR}, M_*, M_{\mathrm{mol}}, M_{\mathrm{HI}}, M_{\mathrm{dust}}, M_{\mathrm{halo}}, \delta_{\mathrm{gal}}, \ldots) \\
M_{\mathrm{dust}}(t) &= f_5(\mathrm{SFR}, M_*, M_{\mathrm{mol}}, M_{\mathrm{HI}}, M_{\mathrm{dust}}, M_{\mathrm{halo}}, \delta_{\mathrm{gal}}, \ldots) \\
M_{\mathrm{halo}}(t) &= f_6(\mathrm{SFR}, M_*, M_{\mathrm{mol}}, M_{\mathrm{HI}}, M_{\mathrm{dust}}, M_{\mathrm{halo}}, \delta_{\mathrm{gal}}, \ldots) \\
\delta_{\mathrm{gal}}(t) &= f_7(\mathrm{SFR}, M_*, M_{\mathrm{mol}}, M_{\mathrm{HI}}, M_{\mathrm{dust}}, M_{\mathrm{halo}}, \delta_{\mathrm{gal}}, \ldots) \\
&\vdots
\end{aligned}
$$



$$x = x(T|T > t)$$

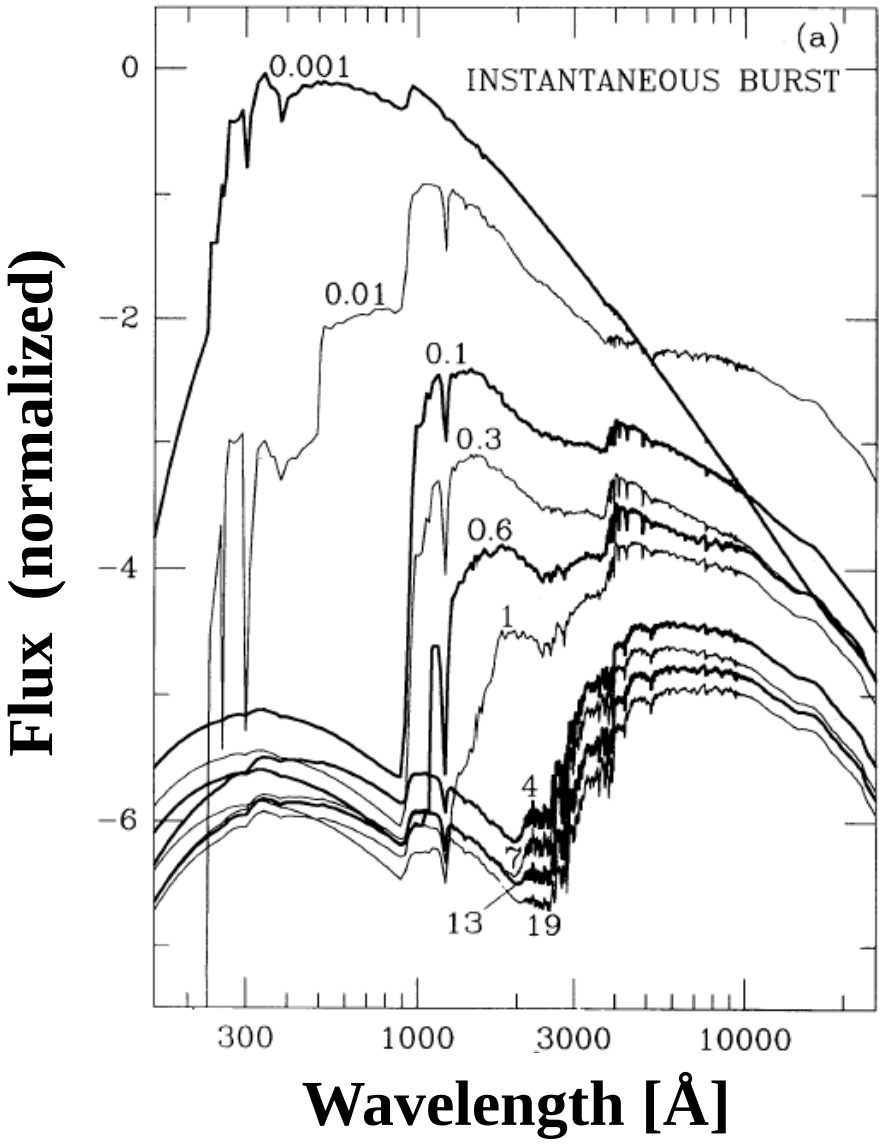
3.1 Galaxy evolution from a modern point of view

Galaxies evolve in various aspects:

$$
\begin{aligned}
\mathrm{SFR}(t) &= f_1(\mathrm{SFR}, M_*, M_{\mathrm{mol}}, M_{\mathrm{HI}}, M_{\mathrm{dust}}, M_{\mathrm{halo}}, \delta_{\mathrm{gal}}, \ldots) \\
M_*(t) &= f_2(\mathrm{SFR}, M_*, M_{\mathrm{mol}}, M_{\mathrm{HI}}, M_{\mathrm{dust}}, M_{\mathrm{halo}}, \delta_{\mathrm{gal}}, \ldots) \\
M_{\mathrm{mol}}(t) &= f_3(\mathrm{SFR}, M_*, M_{\mathrm{mol}}, M_{\mathrm{HI}}, M_{\mathrm{dust}}, M_{\mathrm{halo}}, \delta_{\mathrm{gal}}, \ldots) \\
M_{\mathrm{HI}}(t) &= f_4(\mathrm{SFR}, M_*, M_{\mathrm{mol}}, M_{\mathrm{HI}}, M_{\mathrm{dust}}, M_{\mathrm{halo}}, \delta_{\mathrm{gal}}, \ldots) \\
M_{\mathrm{dust}}(t) &= f_5(\mathrm{SFR}, M_*, M_{\mathrm{mol}}, M_{\mathrm{HI}}, M_{\mathrm{dust}}, M_{\mathrm{halo}}, \delta_{\mathrm{gal}}, \ldots) \\
M_{\mathrm{halo}}(t) &= f_6(\mathrm{SFR}, M_*, M_{\mathrm{mol}}, M_{\mathrm{HI}}, M_{\mathrm{dust}}, M_{\mathrm{halo}}, \delta_{\mathrm{gal}}, \ldots) \\
\delta_{\mathrm{gal}}(t) &= f_7(\mathrm{SFR}, M_*, M_{\mathrm{mol}}, M_{\mathrm{HI}}, M_{\mathrm{dust}}, M_{\mathrm{halo}}, \delta_{\mathrm{gal}}, \ldots) \\
&\vdots
\end{aligned}
$$

This is the formal and ultimate goal of the studies on galaxy evolution, but clearly it is a substantially complicated problem. It is time to define the evolution of galaxies with more objective point of view.

3.2 Galaxy evolution in multiband luminosity space

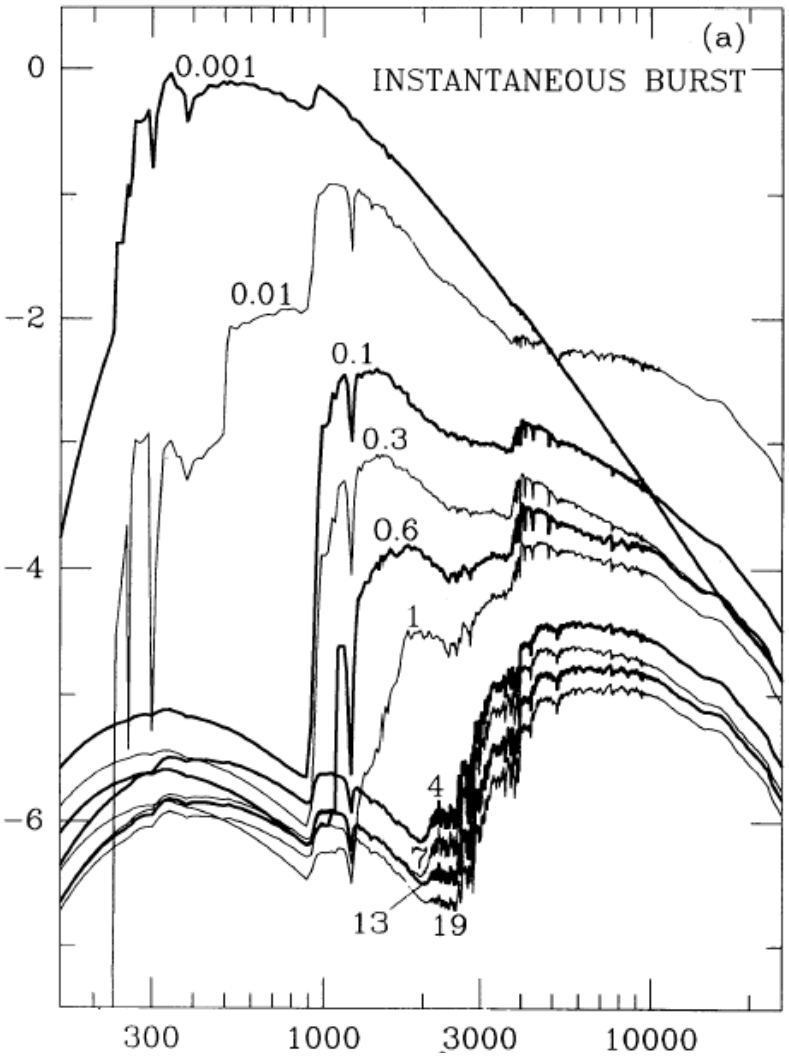


Star formation history (SFH) is one of the key factors of galaxy evolution.

SFH is directly reflected to the spectral luminosity of galaxies.

Bruzual & Charlot (1993)

3.2 Galaxy evolution in multiband luminosity space



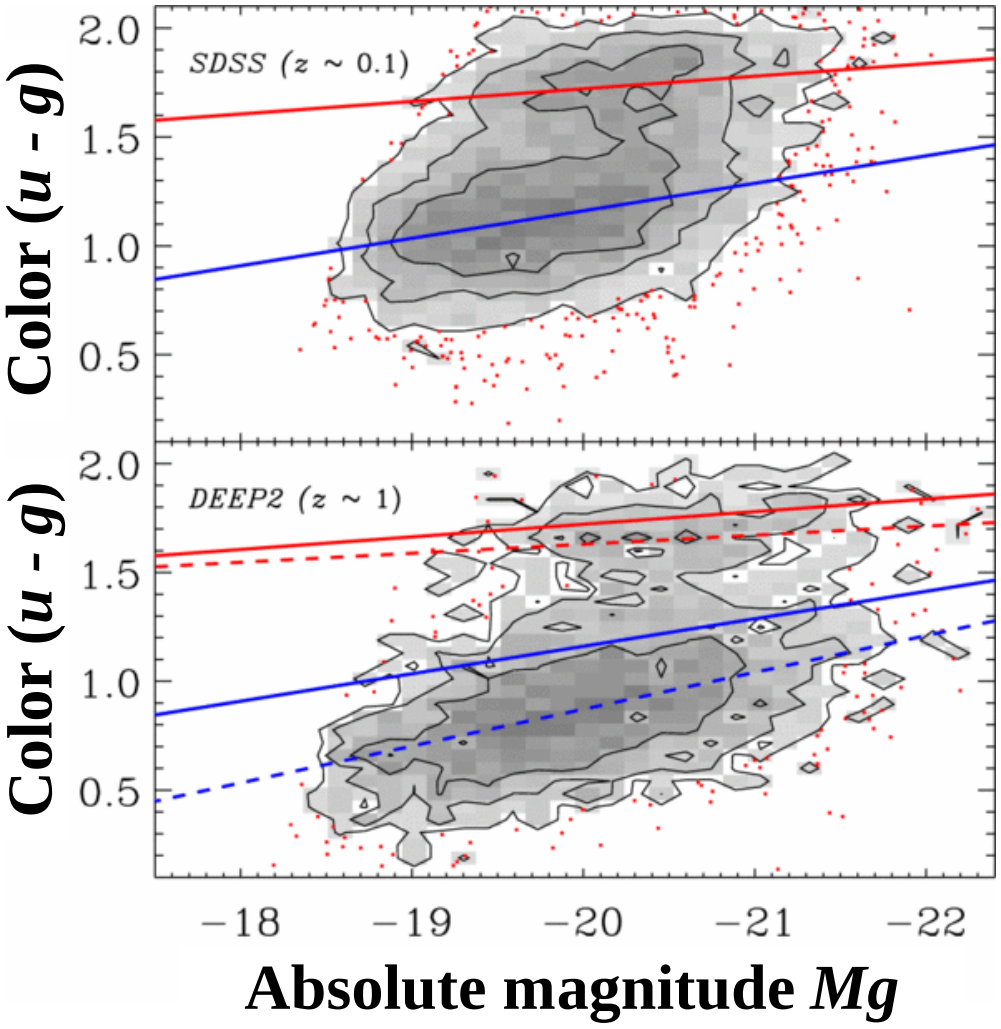
Star formation history (SFH) **is one of the key factors of galaxy evolution.**

SFH is directly reflected to the spectral luminosity of galaxies.

Galaxy evolution related to the SFH will be well represented in the multiwavelength (band) luminosity space.

3.3 Traditional methods and its limitation

Color-magnitude relation



If we plot galaxy luminosity (absolute magnitude) vs. color, a clear dichotomy is found: the color bimodality.

Redder galaxies:

red sequence

Bluer galaxies:

blue cloud

Boundary: green valley

Blanton (2006)

Potential problem in the traditional color-based methods

Colors are basically ratios of two luminosities.

⇒ Selection effect is always too entangled and messy.

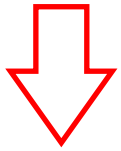
⇒ Completeness test is almost impossible in a simple way.

Potential problem in the traditional color-based methods

Colors are basically ratios of two luminosities.

⇒ Selection effect is always too entangled and messy.

⇒ Completeness test is almost impossible in a simple way.

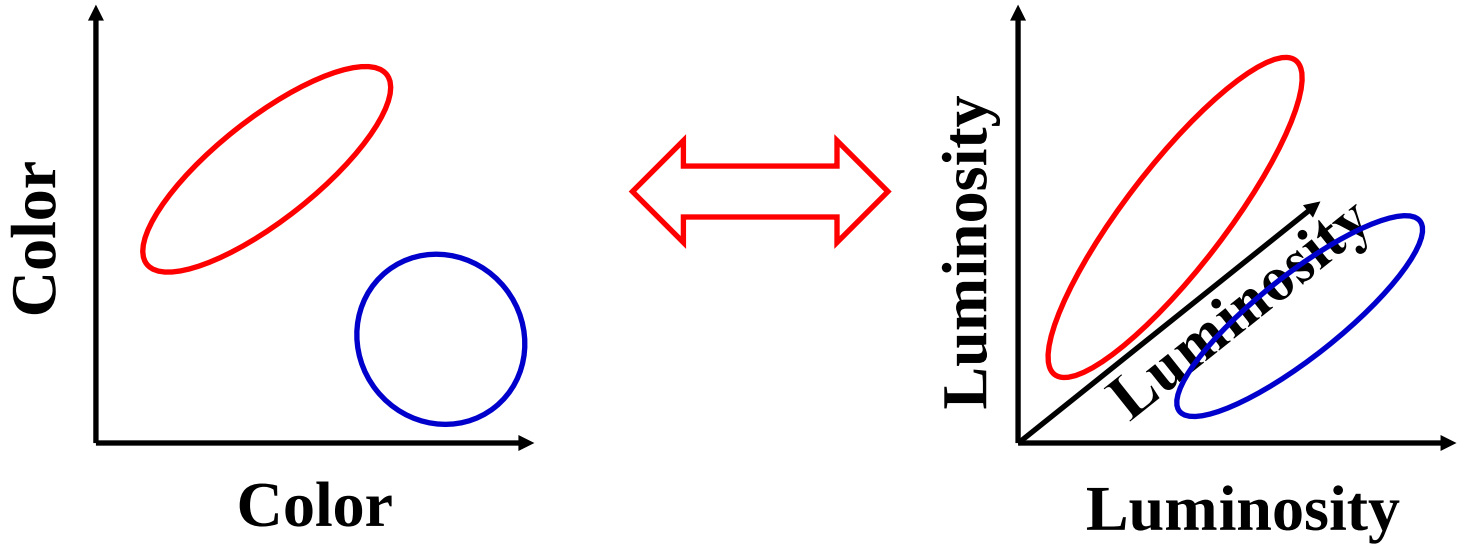


Suggestion: forget about colors!

Instead, we can simply use the distribution of galaxies in a multidimensional luminosity (absolute magnitude) space.

Potential problem in the traditional color-based methods

Since we have a bimodality in color-color space, we must have an equivalent peaks in the multidimensional luminosity space. Color-color plots only show reduced information.



Potential problem in the traditional color-based methods

Since we have a bimodality in color-color space, we must have an equivalent peaks in the multidimensional luminosity space. Color-color plots only show reduced information.

The boundary can be automatically defined by the machine-learning type method.

3.4 Data: RCSED

- **Reference Catalog of galaxy Spectral Energy Distributions (RCSED) (Chilingarian et al. 2016)**
- **Catalog of galaxies produced as join between *GALEX*, SDSS, and UKIDSS catalogs, and processed with state-of-the-art spectral analysis methods**
- **Covers approximately 25% of the sky and contains *k*-corrected ultraviolet-to-near-infrared photometry (11 bands of FUV, NUV, *u*, *g*, *r*, *i*, *z*, *Y*, *J*, *H*, *K*) of some 1 million galaxies, as well as some of their physical properties**

RCSED

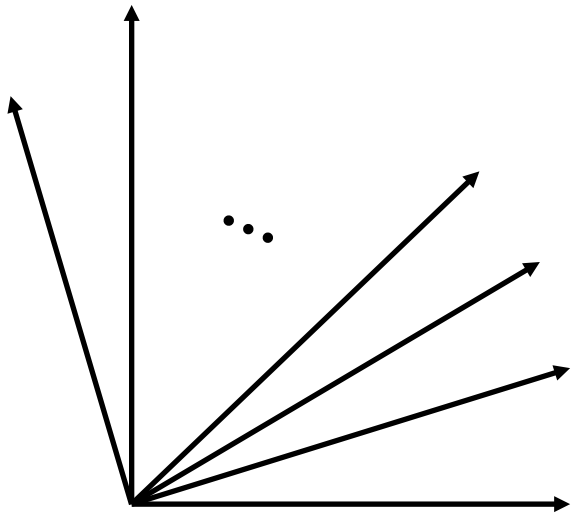
http://rcsed.sai.msu.ru

3.5 Classification in multiwavelength luminosity space

Generate a subsample with all 11 rest-frame magnitudes (FUV, NUV, *u*, *g*, *r*, *i*, *z*, *Y*, *J*, *H*, *K*) ~ 800,000 galaxies

⇒ Construct a volume limited sample that is representative of the whole galaxy sample of ~ 30,000 galaxies.

Perhaps impossible to classify by intuition.



Unsupervised machine learning in luminosity space

Fisher Expectation-Maximization (FEM) algorithm (Bouveryron & Brunet 2012)

1. **Assign initial cluster (class) centers by *k*-means++**
2. **Execute FEM**
   i. **E-step: calculate the complete log-likelihood under the current value of the Gaussian mixture model**
   ii. **F-step: boundary is chosen to maximize the distances between groups, and to minimize the internal scatters**
   iii. **M-step: parameters of Gaussian functions are optimized by maximizing the conditional expectations of the complete log-likelihood**
   iv. **Back to 2.i (E-step) until the result converges.**

Unsupervised machine learning in luminosity space

Fisher Expectation-Maximization (FEM) algorithm (Bouveryron & Brunet 2012)

1. Assign initial cluster (class) centers by *k*-means++
2. **Execute FEM**
   i. **E-step: calculate the complete log-likelihood under the current value of the Gaussian mixture model**
   ii. F-step: boundary is chosen to maximize the distances between groups, and to minimize the internal scatters
   iii. M-step: parameters of Gaussian functions are optimized by maximizing the conditional expectations of the complete log-likelihood
   iv. Back to 2.i (E-step) until the result converges.

Unsupervised machine learning in luminosity space

Fisher Expectation-Maximization (FEM) algorithm (Bouveryron & Brunet 2012)

1. Assign initial cluster (class) centers by *k*-means++
2. **Execute FEM**
   i. E-step: calculate the complete log-likelihood under the current value of the Gaussian mixture model
   ii. **F-step: boundary is chosen to maximize the distances between groups, and to minimize the internal scatters**
   iii. M-step: parameters of Gaussian functions are optimized by maximizing the conditional expectations of the complete log-likelihood
   iv. Back to 2.i (E-step) until the result converges.

Unsupervised machine learning in luminosity space

Fisher Expectation-Maximization (FEM) algorithm (Bouveryron & Brunet 2012)

1. Assign initial cluster (class) centers by *k*-means++
2. **Execute FEM**
   i. E-step: calculate the complete log-likelihood under the current value of the Gaussian mixture model
   ii. F-step: boundary is chosen to maximize the distances between groups, and to minimize the internal scatters
   iii. **M-step: parameters of Gaussian functions are optimized by maximizing the conditional expectations of the complete log-likelihood**
   iv. Back to 2.i (E-step) until the result converges.

Unsupervised machine learning in luminosity space

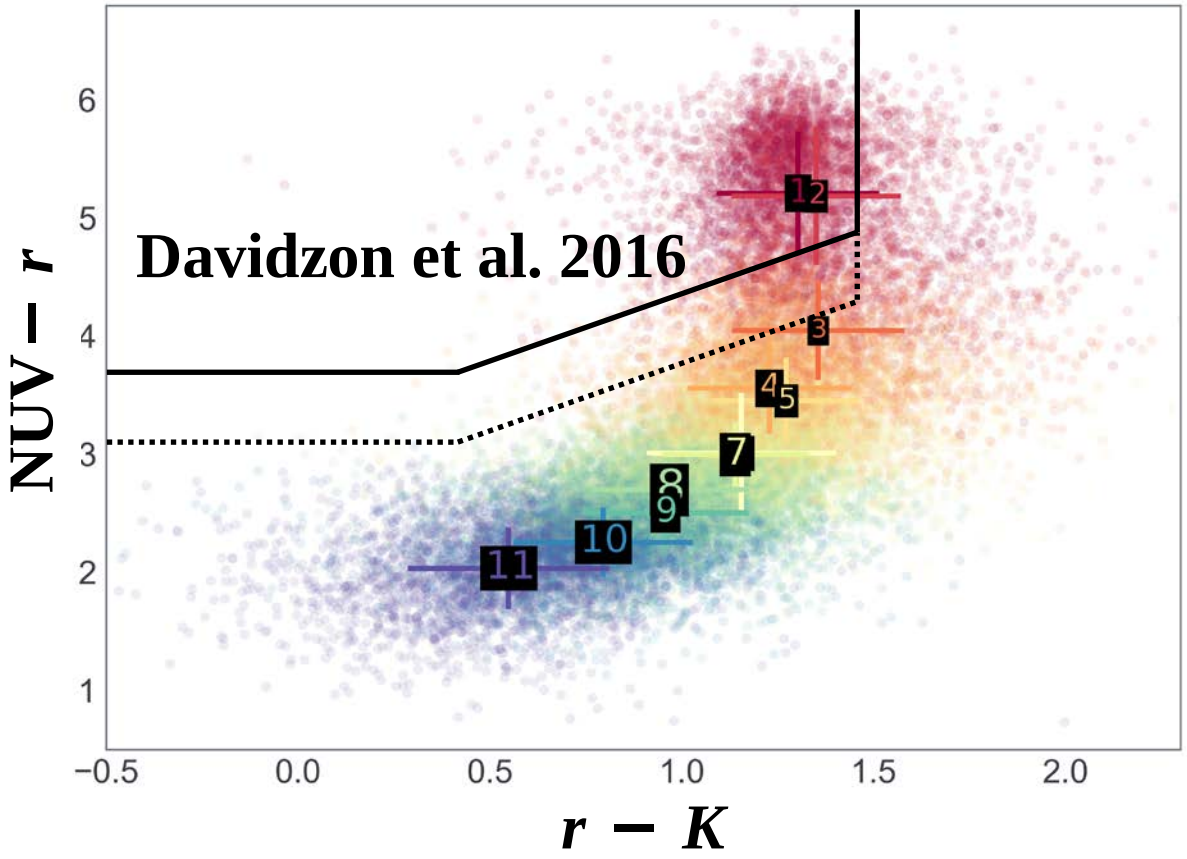
Fisher Expectation-Maximization (FEM) algorithm (Bouveryron & Brunet 2012)

1. Assign initial cluster (class) centers by *k*-means++
2. Execute FEM
   i. E-step: calculate the complete log-likelihood under the current value of the Gaussian mixture model
   ii. F-step: boundary is chosen to maximize the distances between groups, and to minimize the internal scatters
   iii. M-step: parameters of Gaussian functions are optimized by maximizing the conditional expectations of the complete log-likelihood
   iv. Back to 2.i (E-step) until the result converges.

3.6 Classification result

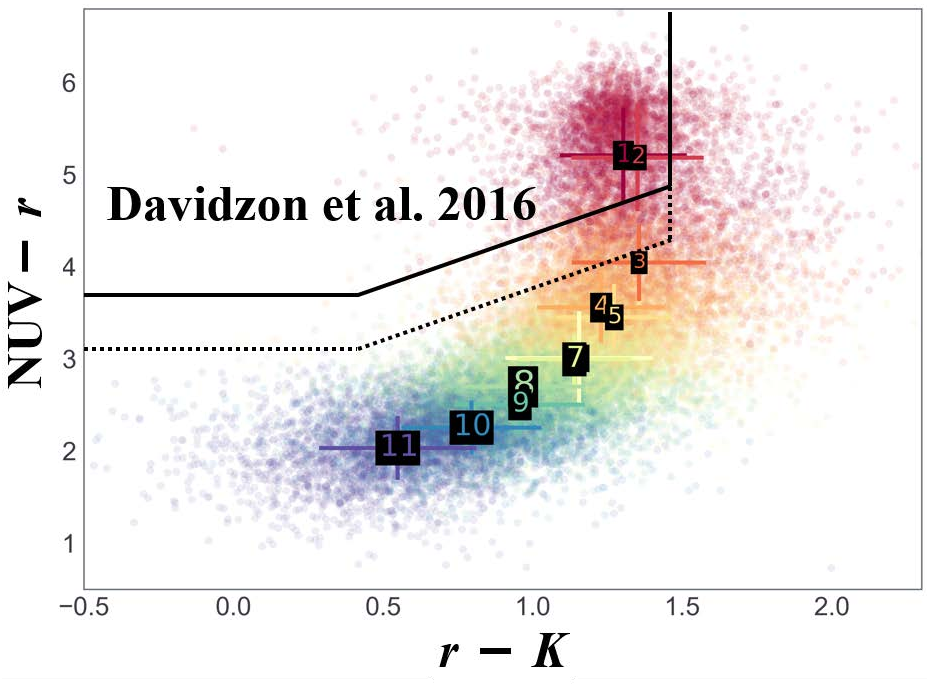
Classification from machine learning

We classified the whole galaxy sample into 11 classes (based on Siudek et al. 2018).



Classification to manifold learning

We classified the whole galaxy sample into 11 classes (based on Siudek et al. 2018).

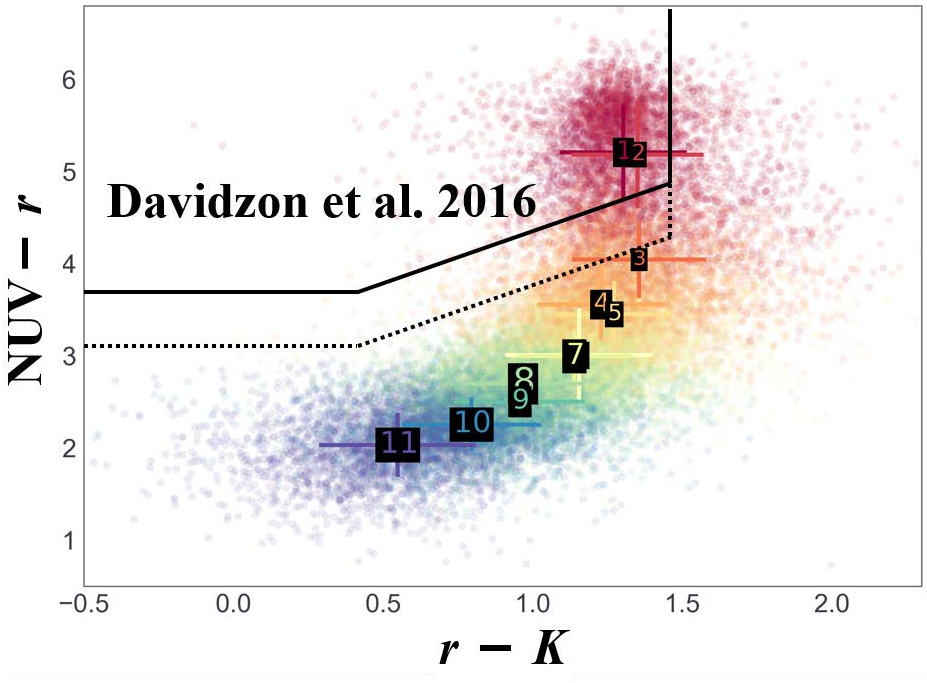


However, the groups do not seem to be sharply separated on this projection.

Intrinsic smooth structure?

Classification to manifold learning

We classified the whole galaxy sample into 11 classes (based on Siudek et al. 2018).



However, the groups do not seem to be sharply separated on this projection.

⇩

Intrinsic smooth structure?

⇩

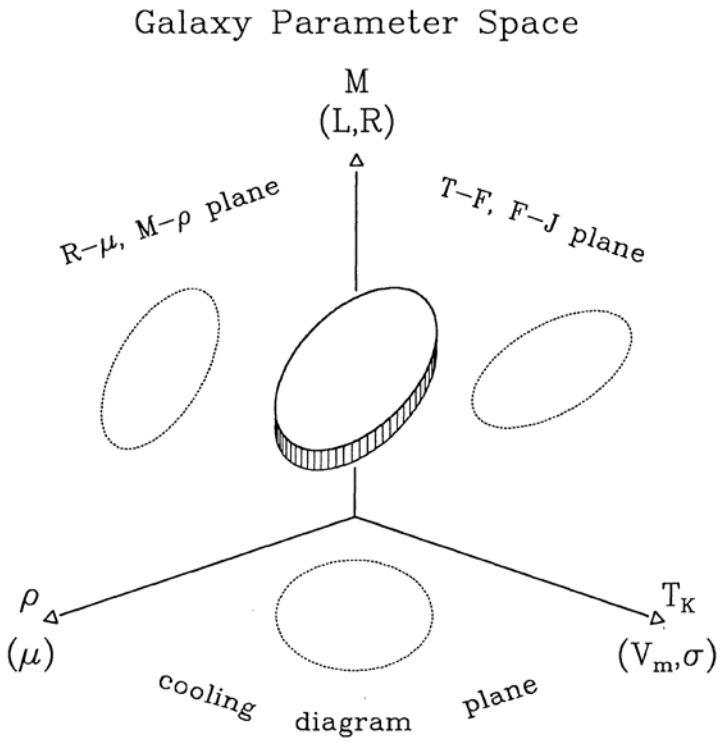
A manifold!

3.7 Galaxy evolution from a galaxy manifold

Galaxy manifold

Historically, in 80's, astronomers introduced a method of classical multivariate analysis such as PCA to find and unify various scaling relations (e.g., Djorgovski 1992).

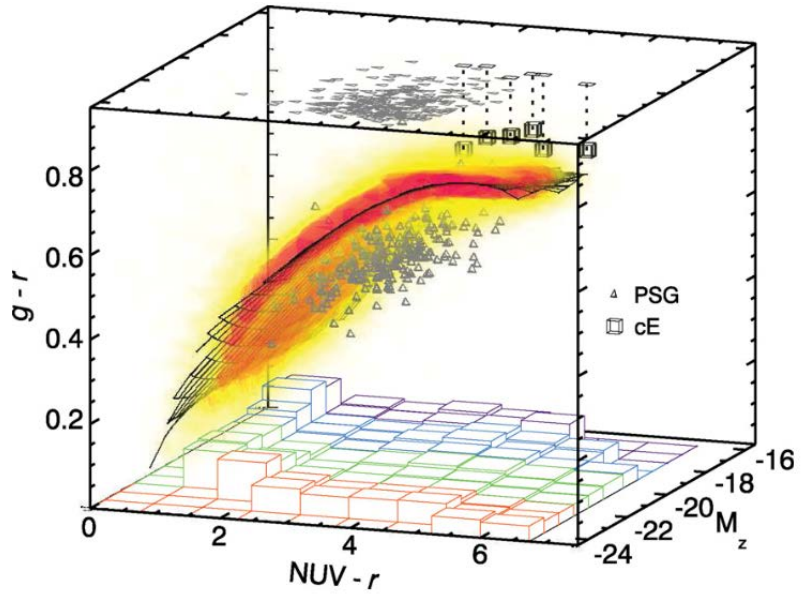
However, since classical PCA-type analysis could only find linear structure in the feature space, the idea worked only to a limited problems, and have been once forgotten.



Djorgovski (1992)

Galaxy manifold

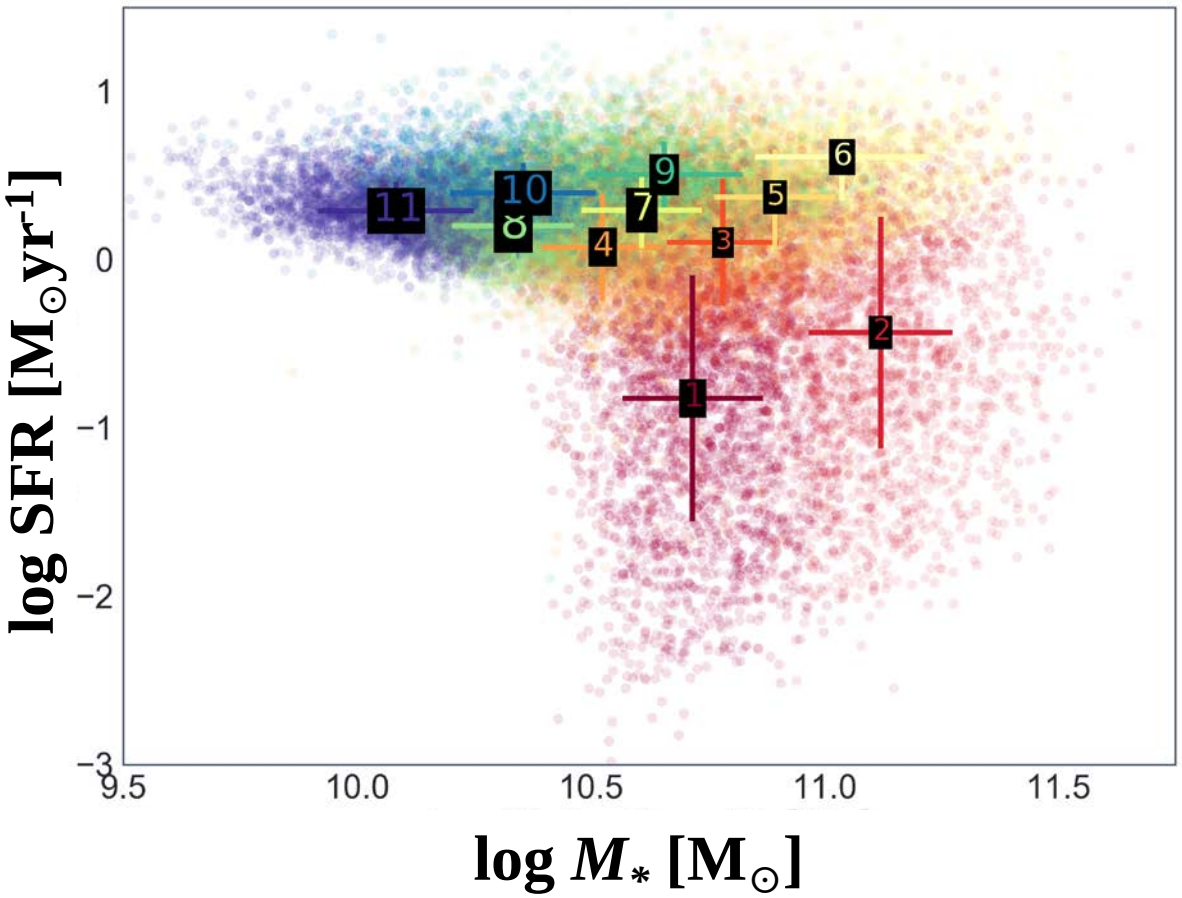
Some preceding studies have suggested the existence of a smooth relation of galaxies in the 3D color–color–magnitude space smoothly continuing from the blue cloud to the red sequence (e.g. Chilingarian et al. 2012).



Chilingarian et al. (2012)

⇒ general idea of a low dimensional submanifold existing in a higher dimensional feature space: revival of the galaxy manifold!

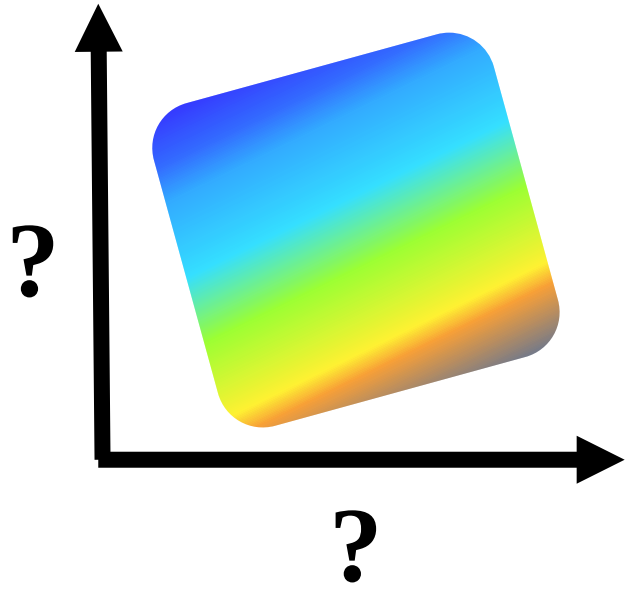
Galaxy manifold



A strongly nonlinear structure was discovered on the SFR-M_* plane. This is also regarded as one of the projections of the galaxy manifold.

Quest for the optimal representation of the manifold

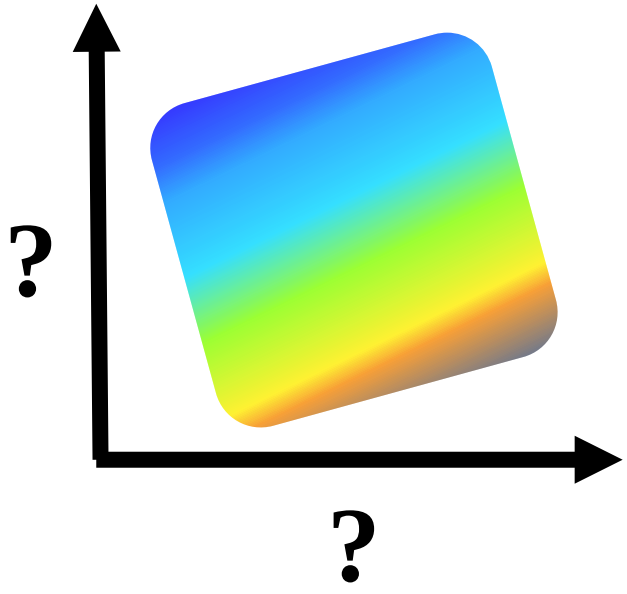
Can we use a machine learning technique to identify what parameters represent the manifold the best?



Quest for the optimal representation of the manifold

Can we use a machine learning technique to identify what parameters represent the manifold the best?

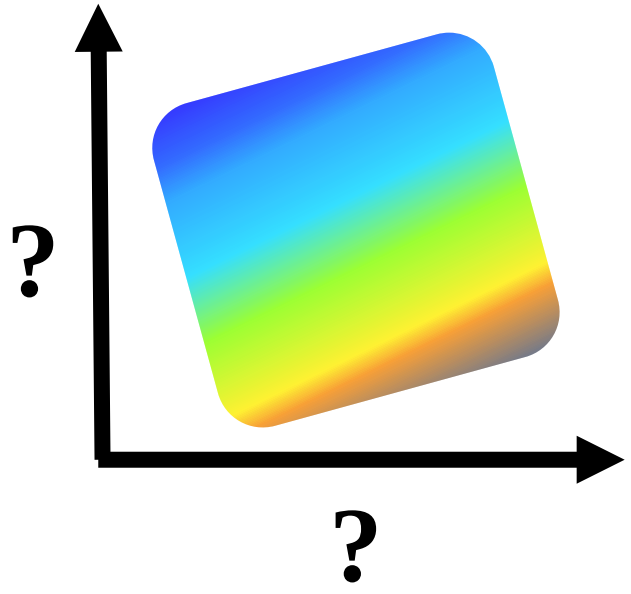
⇒ Yes, if we ask a right question to the machine learning.



Quest for the optimal representation of the manifold

Can we use a machine learning technique to identify what parameters represent the manifold the best?

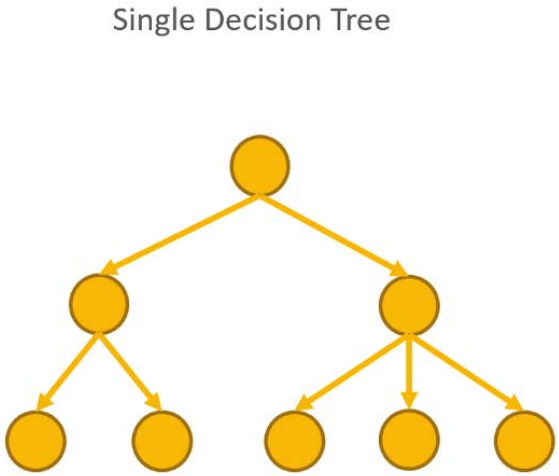
⇒ Yes, if we ask a right question to the machine learning.



Since most of the astronomers are interested in luminosities (colors), we focus on the search for the best-representing photometric bands in this work.

Random forests for feature selection

Random Forests or random decision forests are clustering algorithms that generates random decision trees based on the given features.

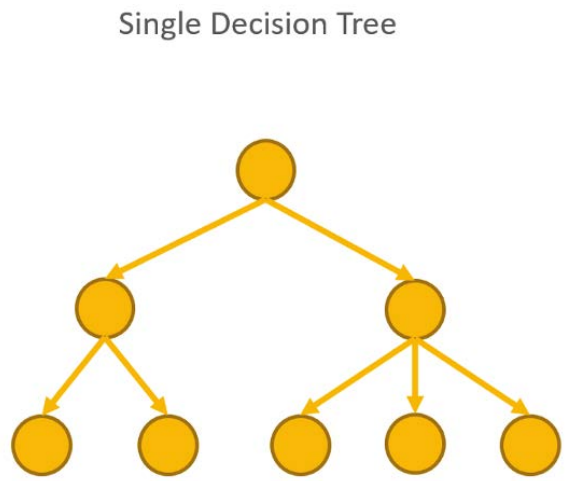


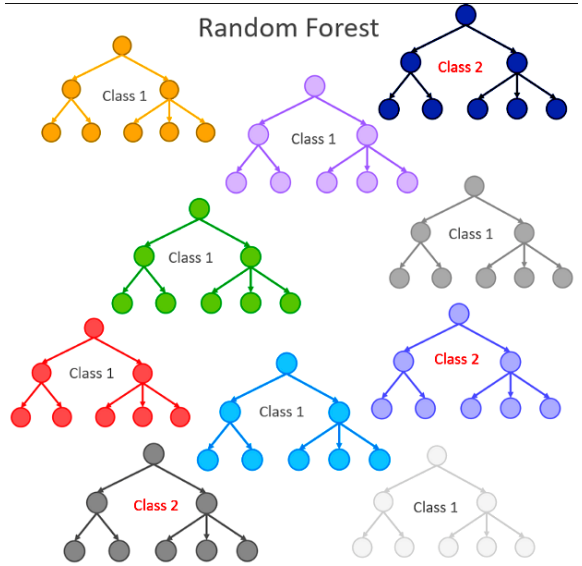
https://towardsdatascience.com/from-a-single-decision-tree-to-a-random-forest-b9523be65147

Random forests for feature selection

Random Forests or random decision forests are clustering algorithms that generates random decision trees based on the given features.

Many of such trees produce a random forest.





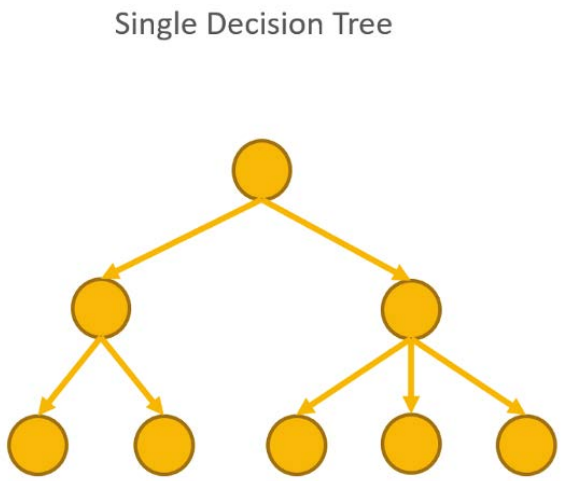
https://towardsdatascience.com/from-a-single-decision-tree-to-a-random-forest-b9523be65147

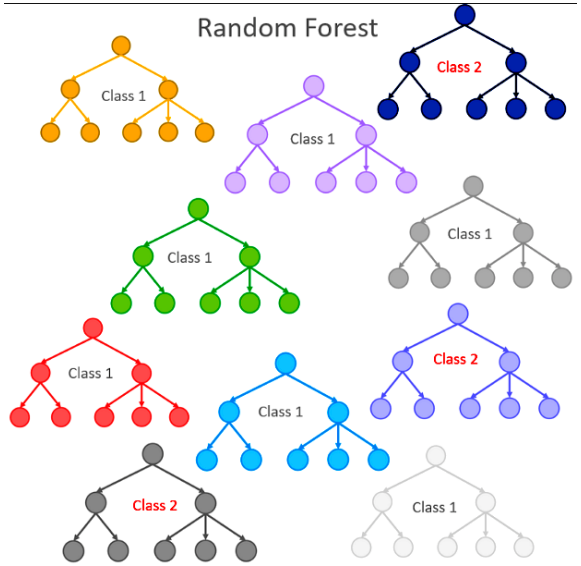
Random forests for feature selection

Random Forests or random decision forests are clustering algorithms that generates random decision trees based on the given features.

Many of such trees produce a random forest.

By providing the answer (class membership), the random forest will learn the important features.





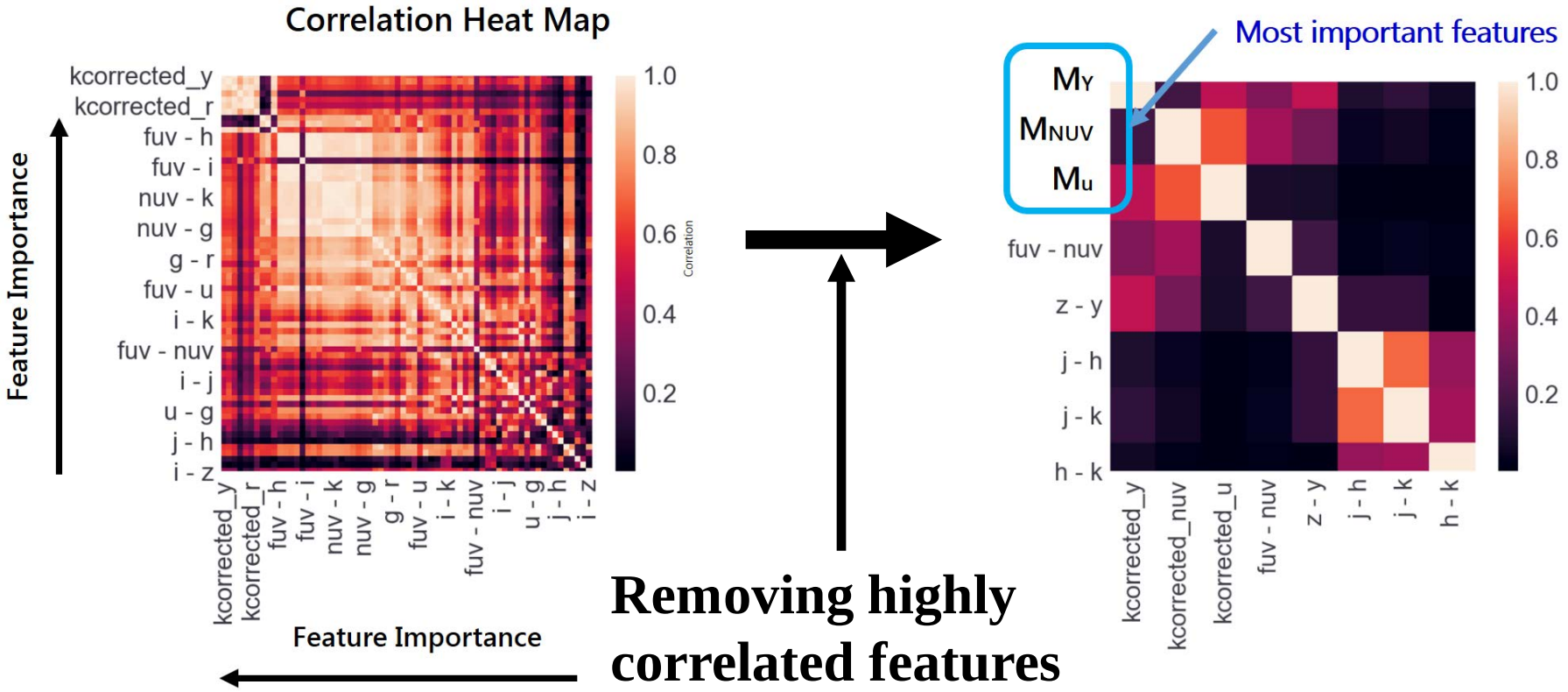
https://towardsdatascience.com/from-a-single-decision-tree-to-a-random-forest-b9523be65147

Random forests for feature selection

Input all the magnitudes and colors with class information.

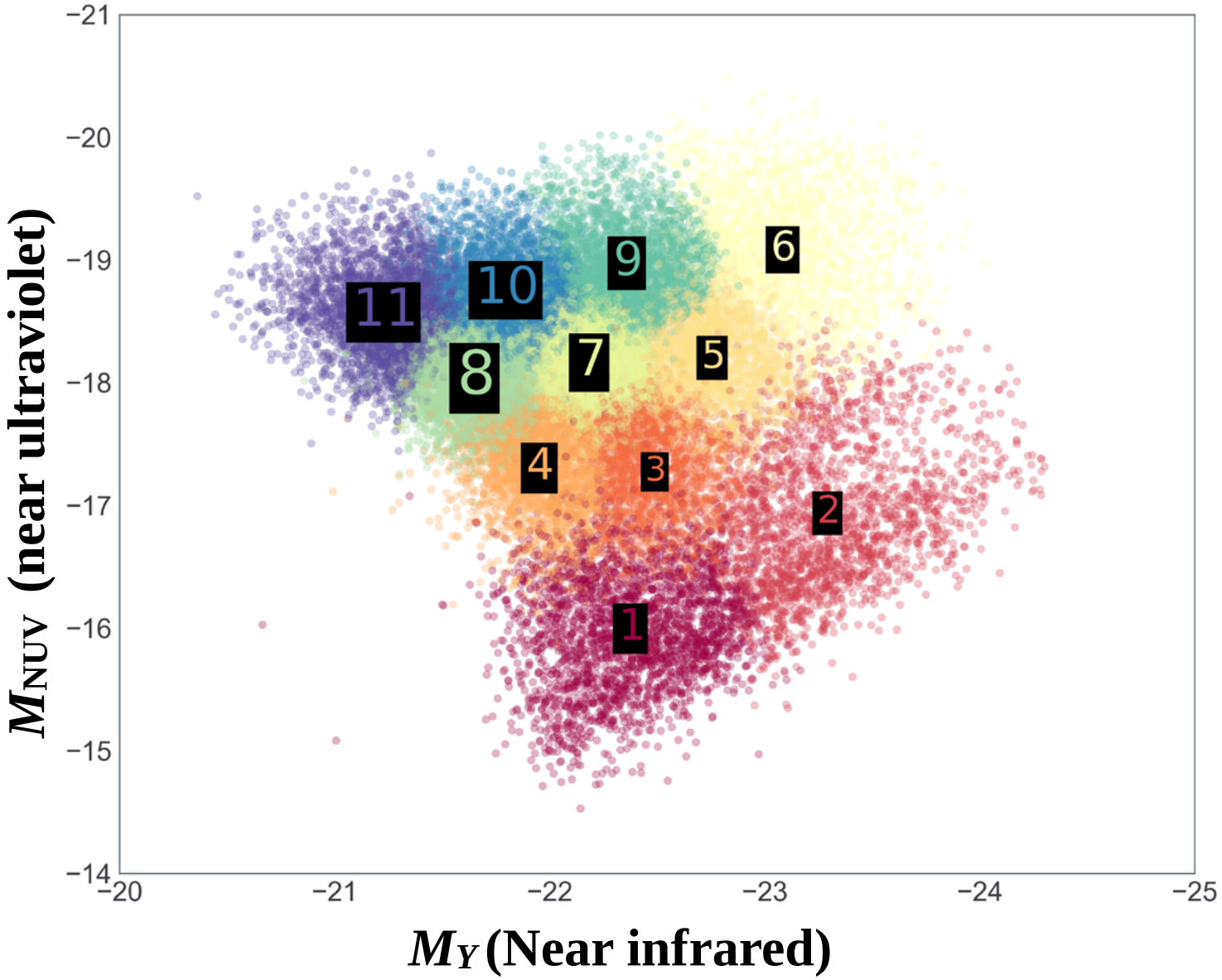
→ **Random forest** →

Features ranked based on the importance

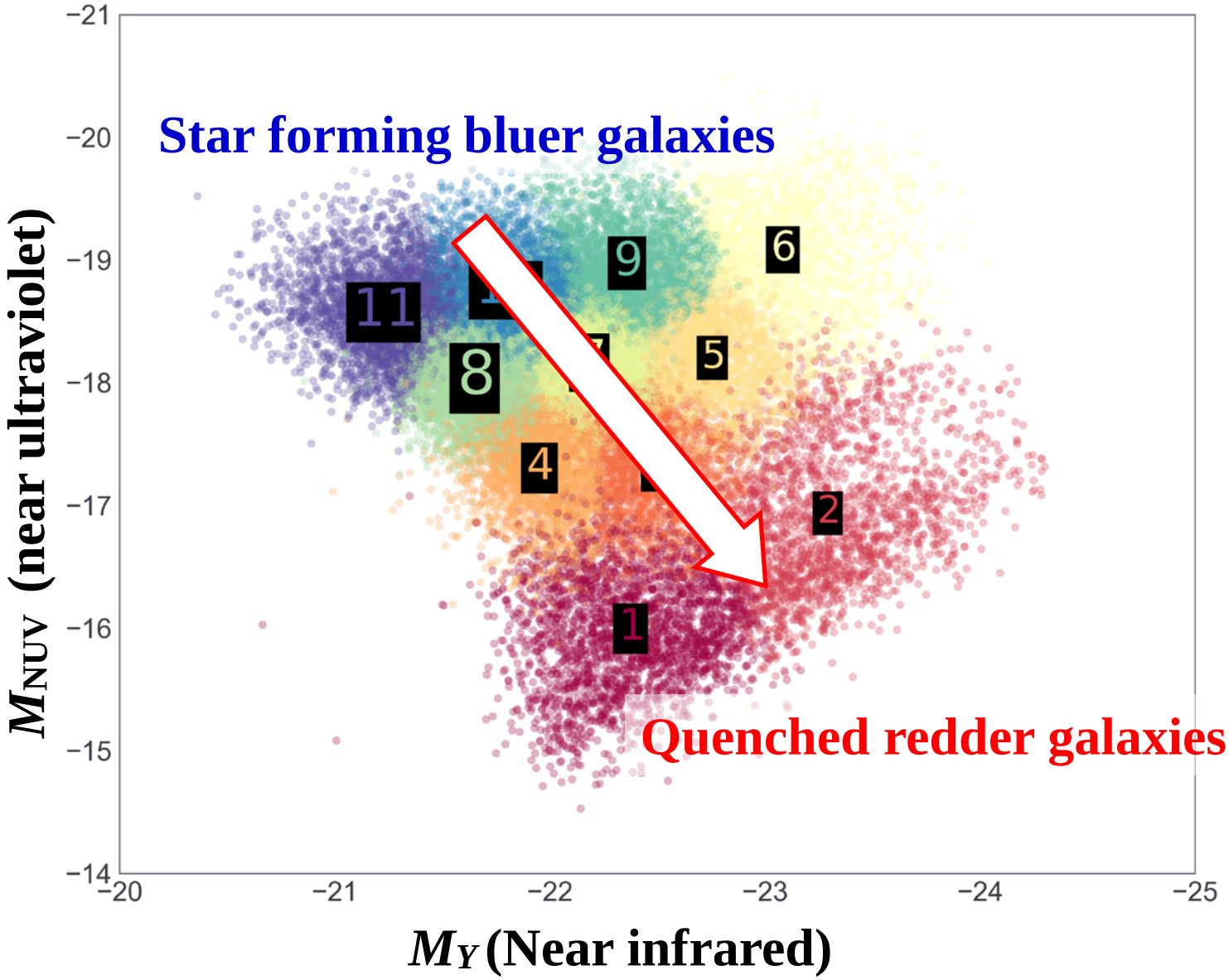


Removing highly correlated features

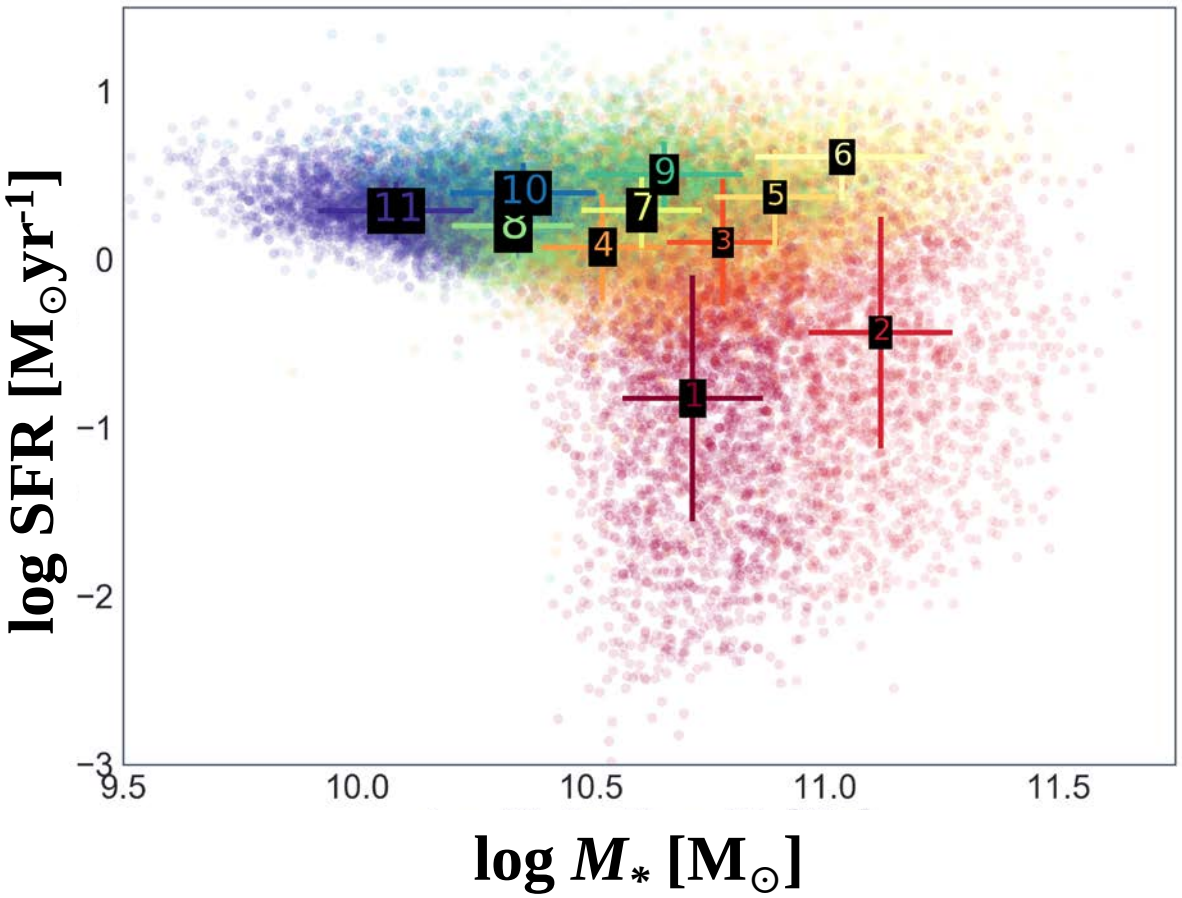
Two-dimensional galaxy manifold



Two-dimensional galaxy manifold

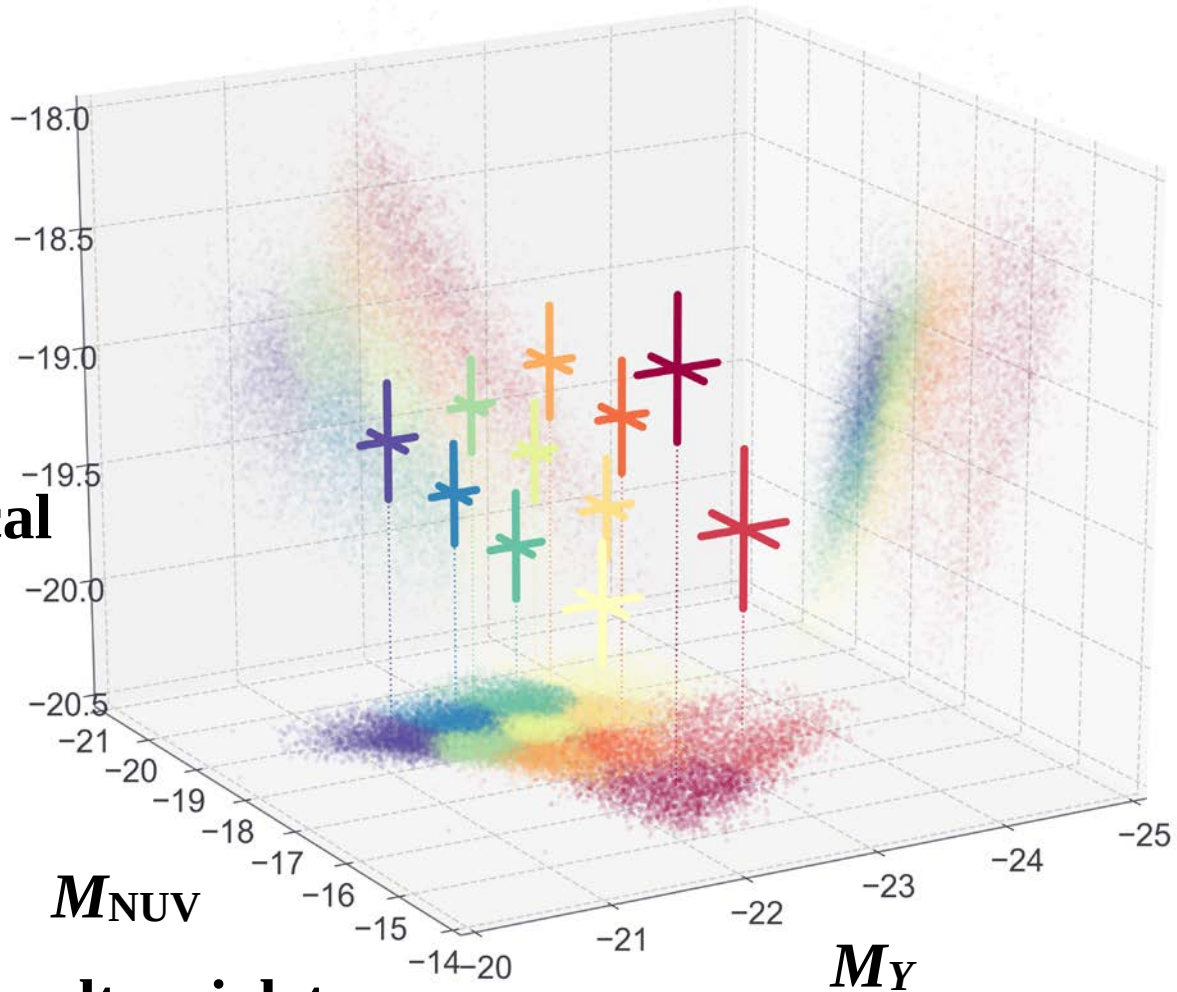


Two dimensional galaxy manifold: M_*-SFR plane revisited



Now it is clear that the M_*-SFR relation is a projection of the 2-dim manifold. The blue galaxy populations are strongly degenerated (seen from edge-on): galaxy main sequence.

Three-dimensional galaxy manifold



M_u

Short optical (3500 Å)

M_{NUV}

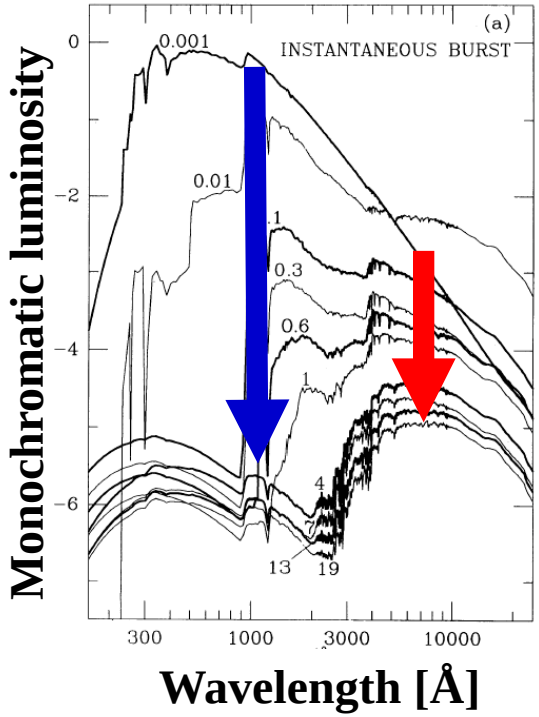
Near ultraviolet (1700 ~ 2800 Å)

M_Y

Near Infrared (~ 10000 Å)

Important implications

The distribution in M_{NUV}-M_Y can be explained intuitively by the evolution of a single-population stellar spectrum.

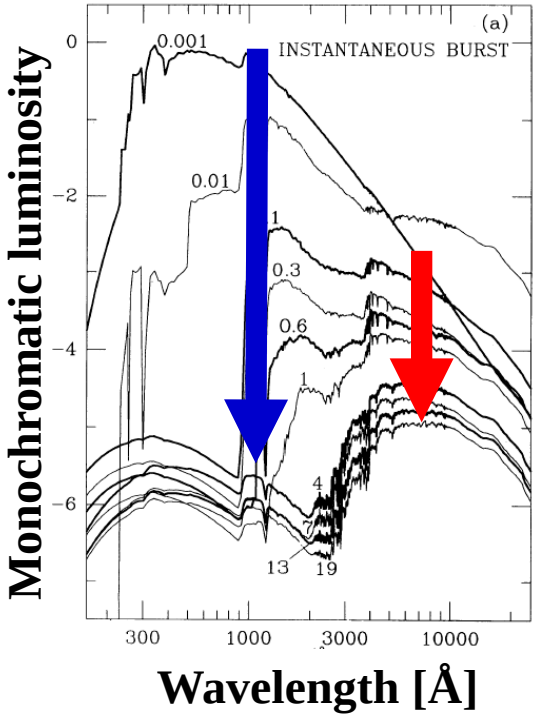


Bruzual & Charlot (1993)

Important implications

The distribution in M_{NUV}-M_Y can be explained intuitively by the evolution of a single-population stellar spectrum.

- **Surprisingly the galaxy spectra are discriminated only by a few broadband luminosities, and NOT complicated combinations of quantities.**
- **Multimodality, and dispersions in classical diagrams are merely a consequence of unsuitable projection of the galaxy manifold.**

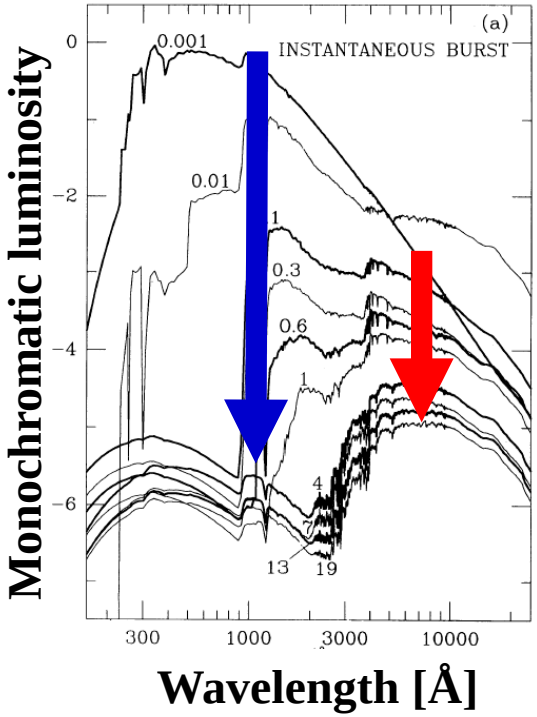


Bruzual & Charlot (1993)

Important implications

The distribution in M_{NUV}-M_Y can be explained intuitively by the evolution of a single-population stellar spectrum.

- **Surprisingly the galaxy spectra are discriminated only by a few broadband luminosities, and NOT complicated combinations of quantities.**
- **Multimodality, and dispersions in classical diagrams are merely a consequence of unsuitable projection of the galaxy manifold.**



Bruzual & Charlot (1993)

It is a good time to rethink how to parametrize galaxies. We should characterize the evolution by a continuous parameter.

4. Conclusions

We applied two different methods of the TDA, the persistent homology and manifold learning to the galaxy survey data.

1. **Persistent homology in cosmology**
   - **In the local Universe, we found ~ 10 loops above the 90% confidence band, whose average radius is *r* ~ 150 Mpc from the 2000 galaxy subsample extracted from SDSS DR12 data.**
   - **To explore the effect of evolution, we applied the persistent homology to SDSS DR14 QSO data. We discovered the BAO signal at all the cosmic age. We plan to introduce Wasserstein metric to quantify it.**
   - **This demonstrates the power of the TDA method for the studies on cosmology and galaxy evolution.**

4. Conclusions

2. **Manifold learning for galaxy evolution**
   - **We demonstrated how machine learning methods can aid our understanding of galaxy evolution.**
   - **We found highly nonlinear continuous structure in the multidimensional luminosity space: the galaxy manifold. This is a revived, improved version of the concept discussed in 80's.**
   - **All the known empirical relations are projection of the manifold.**
   - **The galaxy manifold represents the evolutionary sequence of galaxies. Possibly we can parametrize the galaxy evolution by a few parameters.**

4. Conclusions

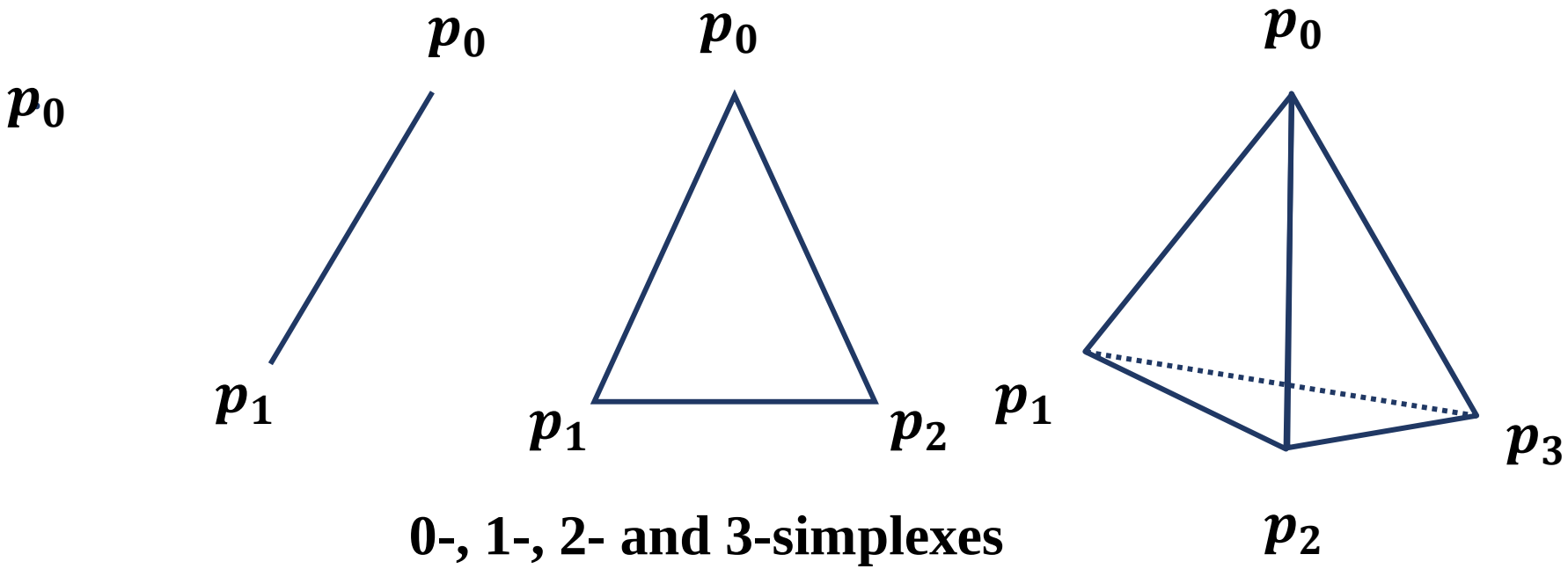
Take-home message

Machine learning can serve as a powerful and revolutionary method to develop fundamental part of astrophysics!

Appendix

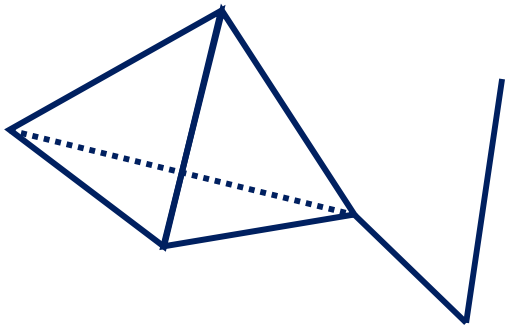
Simplicial complex

A simplex is a generalization of the notion of a triangle or tetrahedron to arbitrary dimensions.



0-, 1-, 2- and 3-simplexes

A simplicial complex is a set composed of points, line segments, triangles and their n -dimensional counterparts.



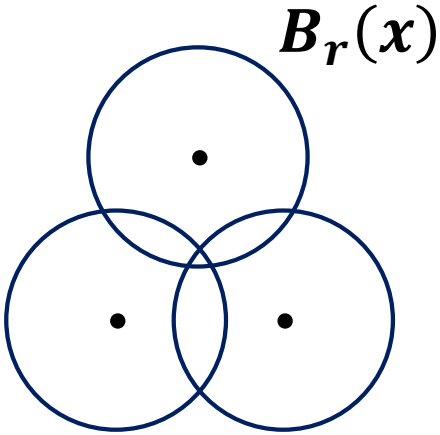
Čech complex

For a set P, the center of spheres,

$$P = \{x_i \in \mathbb{R}^N | i = 1, \dots, n\}$$

$$B_r(x_i) = \{y \in \mathbb{R}^N | |y - x_i| \leq r\}$$ **Sphere with radius r.**

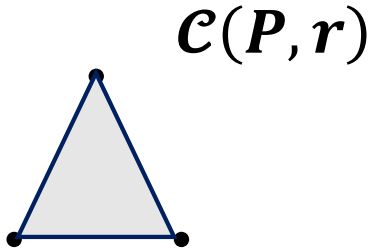
$$\mathcal{C}(P, r) = \left\{ |x_{i_0} \cdots x_{i_k}| \bigcap_{j=0}^{m} B_r\left(x_{i_j}\right) \neq \emptyset \right\}$$



The Čech complex $\mathcal{C}(P, r)$ is the nerve of

$$\Phi = \{B_r(x_i) | \, x_i \in P\}$$

$$\mathcal{C}(P, r) = \mathcal{N}(\Phi)$$



Voronoi diagrams and Delaunay complex

For a set P of the center of spheres,

$$P = \{x_i \in \mathbb{R}^N \,|i = 1, \dots, n\}$$

a Voronoi diagram $\mathbb{R}^N$ can be expressed with Voronoi cells V_i for points x_i.

$$V_i = \left\{x \in \mathbb{R}^N \,||x - x_i| \leq \left|x - x_j\right|, j \neq i\right\}$$ **Voronoi cell**

$$\mathbb{R}^N = \bigcup_{i=1}^{m} V_i$$ **Voronoi decomposition**

Then, Delauney complex $\mathcal{D}(P)$ can be given as the nerve $\mathcal{N}(\Phi)$ of the Voronoi diagram where Φ is a convex closed set.

$$\Phi = \{V_i | i = i, \dots, m\}$$

$$\mathcal{D}(P) = \mathcal{N}(\Phi)$$

Voronoi diagrams and Delaunay complex

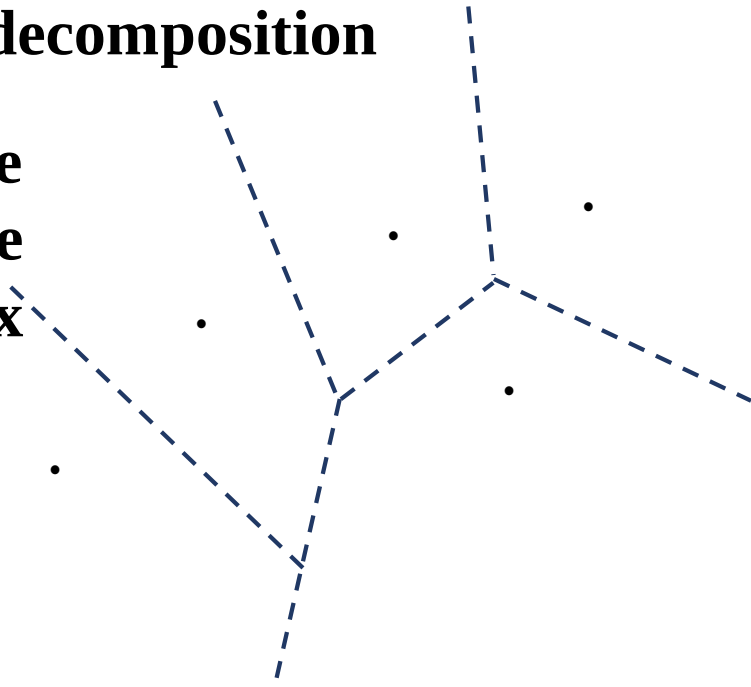
For a set P of the center of spheres,

$$P = \{x_i \in \mathbb{R}^N \,|i = 1, \dots, n\}$$

a Voronoi diagram $\mathbb{R}^N$ can be expressed with Voronoi cells V_i for points x_i.

$$V_i = \left\{x \in \mathbb{R}^N \,||x - x_i| \leq \left|x - x_j\right|, j \neq i\right\}$$ **Voronoi cell**

$$\mathbb{R}^N = \bigcup_{i=1}^{m} V_i$$ **Voronoi decomposition**



Then, Delauney complex $\mathcal{D}(P)$ can be given as the nerve $\mathcal{N}(\Phi)$ of the Voronoi diagram where Φ is a convex closed set.

$$\Phi = \{V_i | i = i, \dots, m\}$$

$$\mathcal{D}(P) = \mathcal{N}(\Phi)$$

Voronoi diagrams and Delaunay complex

For a set P of the center of spheres,

$$P = \{x_i \in \mathbb{R}^N \,|i = 1, \dots, n\}$$

a Voronoi diagram $\mathbb{R}^N$ can be expressed with Voronoi cells V_i for points x_i.

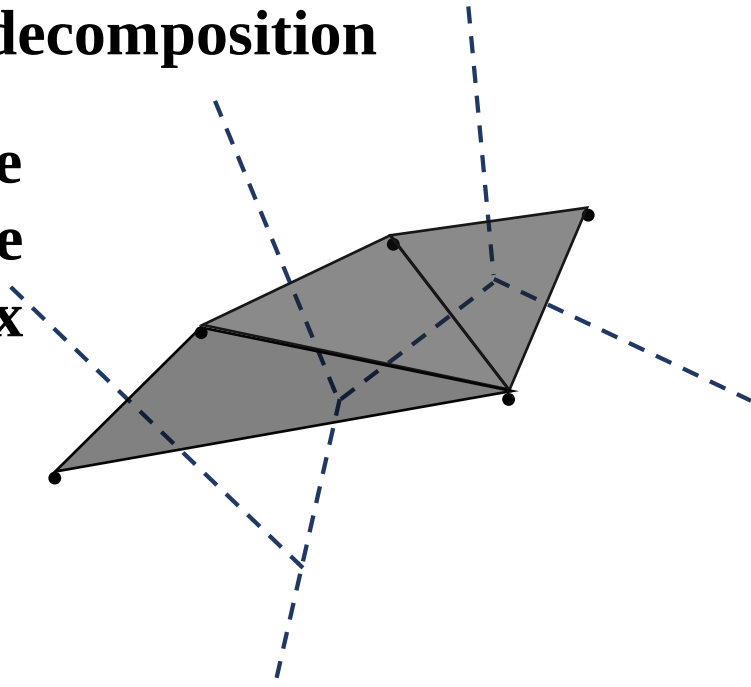
$$V_i = \left\{x \in \mathbb{R}^N \,||x - x_i| \leq \left|x - x_j\right|, j \neq i\right\}$$ **Voronoi cell**

$$\mathbb{R}^N = \bigcup_{i=1}^{m} V_i$$ **Voronoi decomposition**

Then, Delauney complex $\mathcal{D}(P)$ can be given as the nerve $\mathcal{N}(\Phi)$ of the Voronoi diagram where Φ is a convex closed set.

$$\Phi = \{V_i | i = i, \dots, m\}$$

$$\mathcal{D}(P) = \mathcal{N}(\Phi)$$



Alpha complex

We further define a intersection of $B_r(x_i)$ and V_i as W_i.

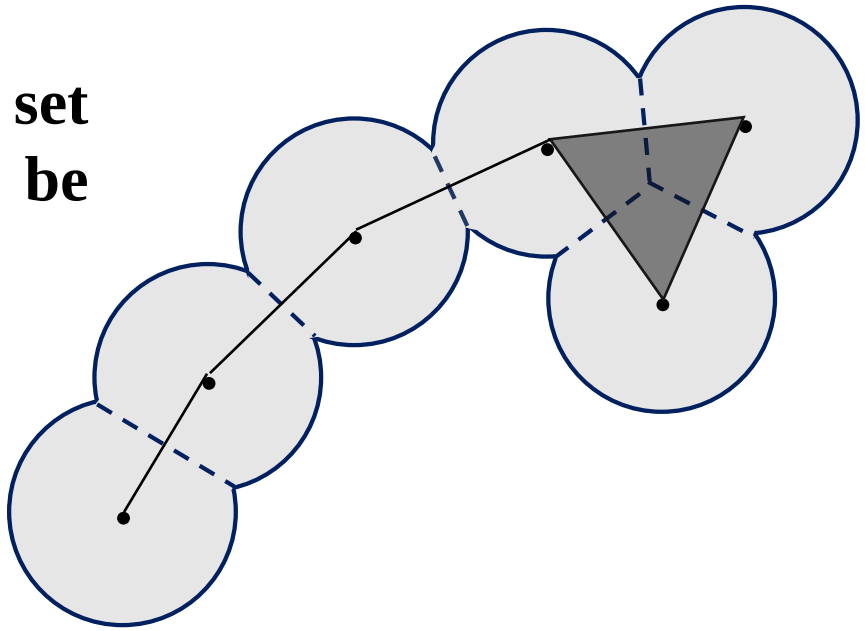
$$W_i = B_r(x_i) \cap V_i$$

$$X_r = \bigcup_{i=1}^{m} W_i$$

An alpha complex $\alpha(P, r)$ for a set of the center of spheres can be defined as below.

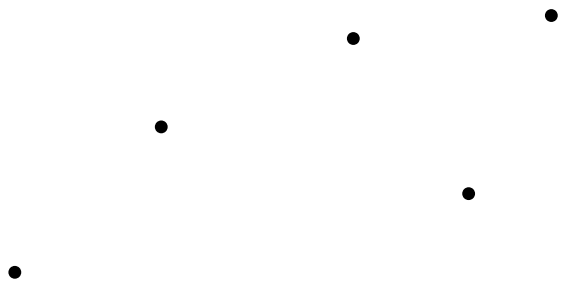
$$\Psi = \{W_i | i = 1, \ldots, m\}$$

$$\alpha(P, r) = \mathcal{N}(\Psi)$$

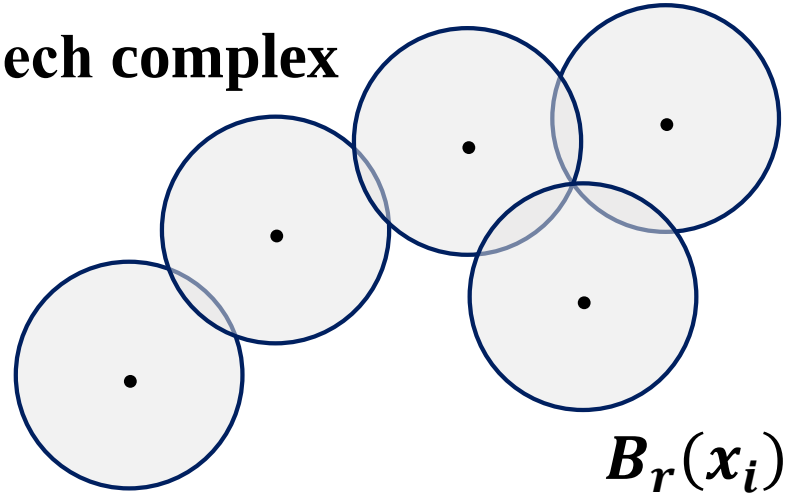


Summary of introduced quantities

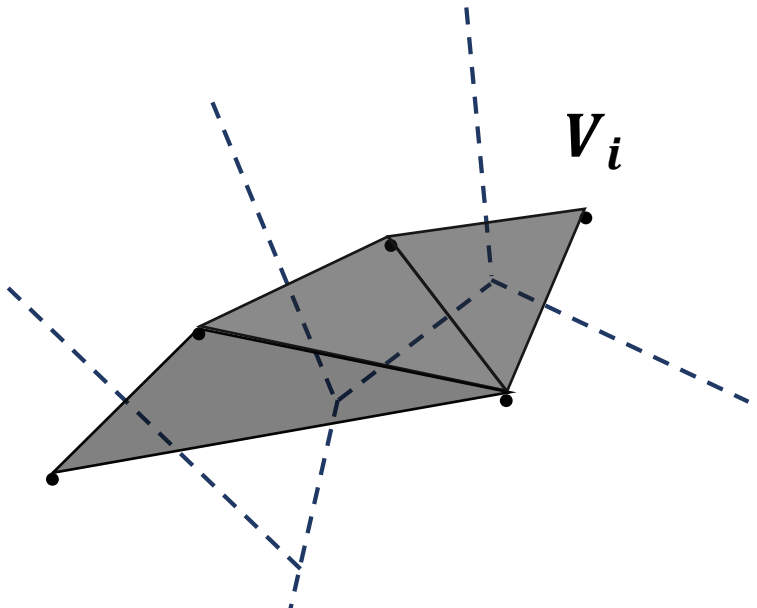
Data points



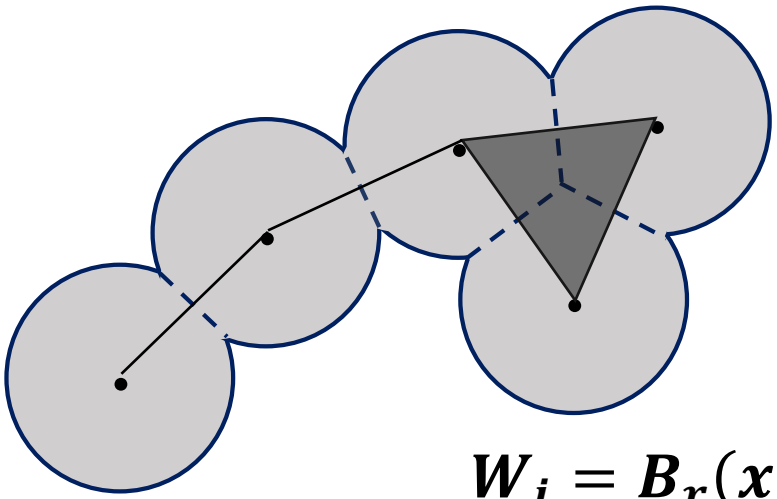
Čech complex



Delaunay complex

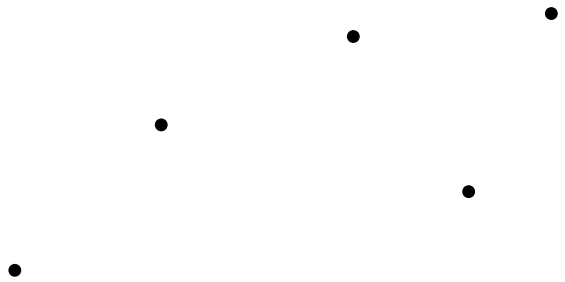


Alpha complex

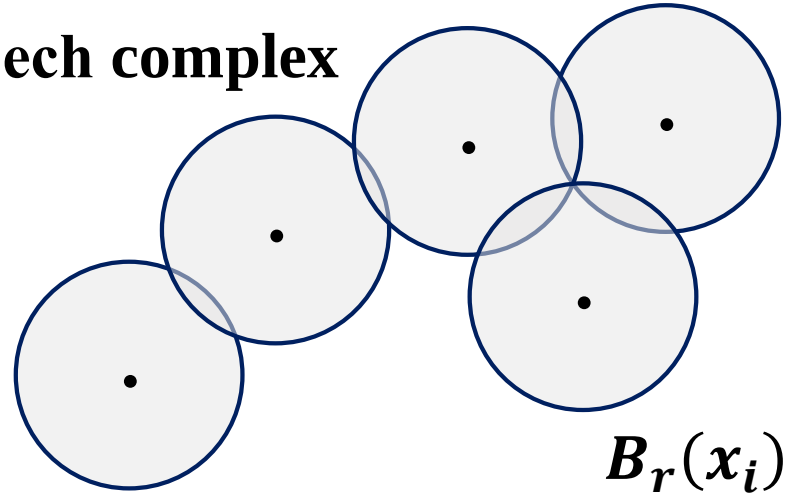


Summary of introduced quantities

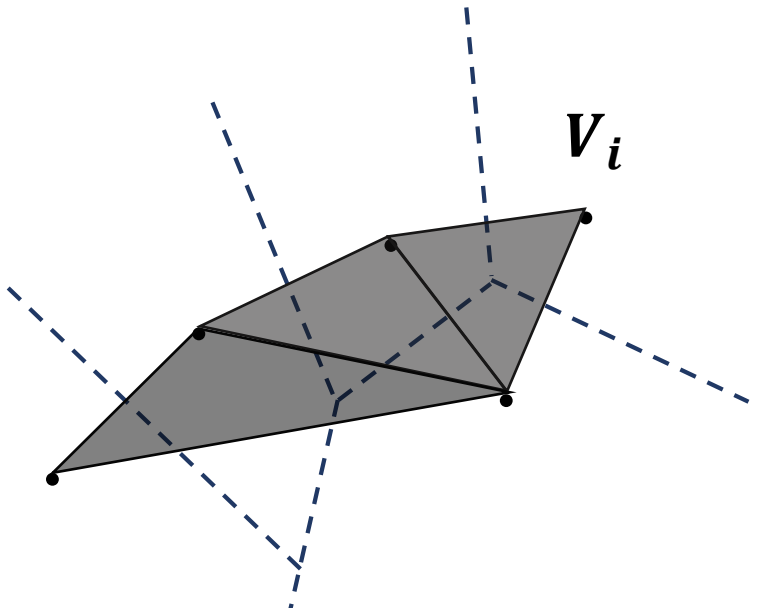
Data points



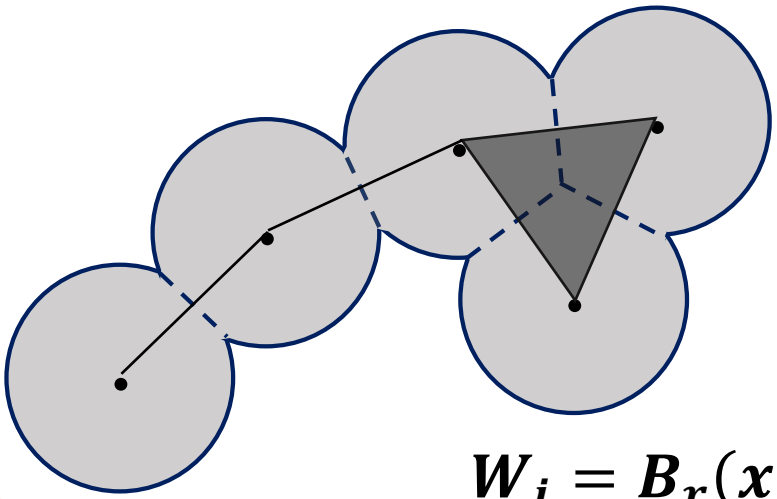
Čech complex



Delaunay complex



Alpha complex



The good points of alpha complex

Alpha complex is a subcomplex of Čech and Delaunay complex because $W_i \subset B_r(x_i), V_i$.

Since we can reduce the number of simplexes, alpha complex is preferred.

∵ Although the Čech complex and the alpha complex is homotopically equivalent (HE), the dimension of simplicial complex which is HE to the Čech complex can be higher than N.