Recent progress in the application of high-dimensional statistics to astrophysics and cosmology

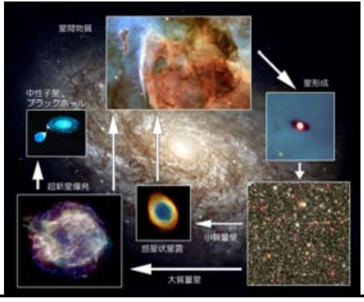
Tsutomu T. TAKEUCHI
1. Division of Particle and Astrophysical Science, Nagoya University, Japan
2. The Research Center for Statistical Machine Learning, the Institute of Statistical Mathematics

New Developments of Theories and Methodologies for Large Complex Data, , Tsukuba, 4-5 Nov., 2022

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.



Collaborators

Suchetha COORAY, Kai T. KONO (河野 海)
Division of Particle and Astrophysical Science, Nagoya University, Japan

Kouichiro NAKANISHI (中西 康一郎)
ALMA Project, National Astronomical Observatory of Japan

Kazuyoshi YATA (矢田 和善), Makoto AOSHIMA(青嶋 誠)
Institute of Mathematics, University of Tsukuba, Japan

Aki ISHII (石井 晶)
Department of Information Sciences, Tokyo University of Science, Japan

Kento EGASHIRA (江頭 健斗)
Graduate School of Science and Technology, University of Tsukuba, Japan

Kohji YOSHIKAWA (吉川 耕司)
Center for Computational Sciences, University of Tsukuba, Japan

Kotaro KOHNO (河野 孝太郎)
Institute of Astronomy, The University of Tokyo, Japan

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.

VERA, NAOJ

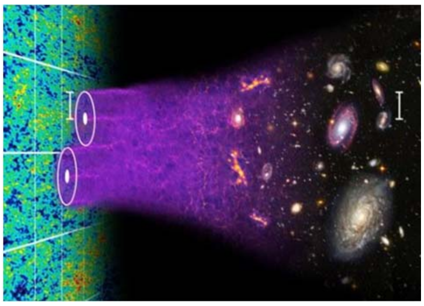
0. Background

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.

1. Introduction

1.1 Structure formation in the Universe

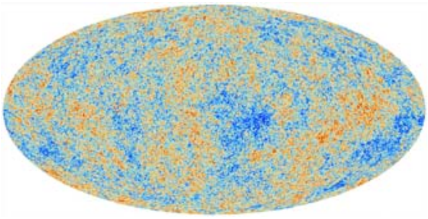


All the structures in the Universe have emerged from a tiny fluctuation at very early epoch (380,000 yr).

1.2 Galaxy formation from the cosmic initial condition

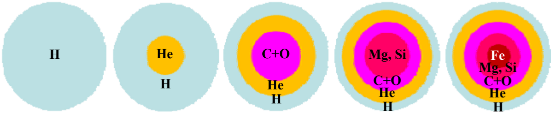
Galaxies are supposed to have formed from a tiny (order of ~ 10^{-5}) fluctuation of matter (mainly dark matter: DM) in the early Universe.

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



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

Life of stars and their nucleosynthesis



Tc 10^7K 10^8K 3×10^8K 10^9K 5×10^9K

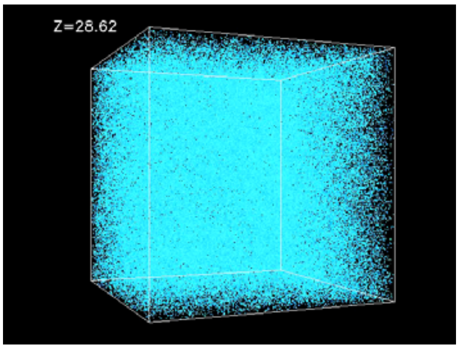
Birth H burning (Main sequence) He burning (Giant) C+O burning Si burning

Stars produce heavy elements by the nuclear reaction, and how far the reaction goes depends on the mass of stars.

Lighter than the Sun

Heavier than the Sun

1.3 Formation and evolution of galaxies



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

Life of stars and their nucleosynthesis II (old scenario)

Where are heavier elements than iron produced?

⇩

Final phase of stars with various masses is related:

1. The death of stars with mass < 8 $M_\odot$

   These stars become unstable and repeat expansion and contraction (thermal pulse instability). Atoms absorb neutrons and turn into heavier elements: s-process

2. The death of stars with mass > 8 $M_\odot$

   Atoms absorb neutrons very quickly during the supernova explosions, and heavy elements are rapidly produced: r-process

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.

Life of stars and their nucleosynthesis II (new scenario)

Where are heavier elements than iron produced?

⇩

Final phase of stars with various masses is related:

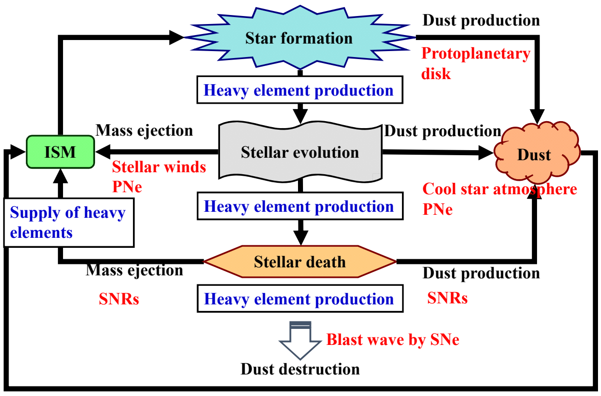
1. The death of stars with mass < 8 $M_\odot$

   These stars become unstable and repeat expansion and contraction (thermal pulse instability). Atoms absorb neutrons and turn into heavier elements: s-process

2. The death of stars with mass > 8 $M_\odot$

   Neutron star binaries explode as kilonovae, and r-process occurs during the explosion.

Chemical evolution of galaxies: metal and dust



Star formation

Heavy element production

Stellar evolution

Heavy element production

Stellar death

Heavy element production

ISM

Dust

Dust production

Protoplanetary disk

Mass ejection

Stellar winds

PNe

Dust production

Cool star atmosphere

PNe

Supply of heavy elements

Mass ejection

SNRs

Dust production

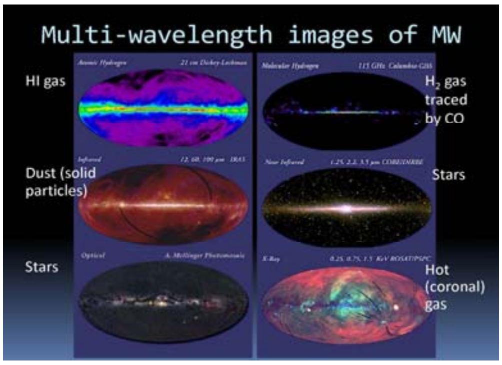
SNRs

Blast wave by SNe

Dust destruction

2. Interstellar Medium (ISM)

2.1 Matter between stars in galaxies



2.2 Properties of the ISM

- Space between the stars within a galaxy is not empty.
- ISM consists of **gas** and **dust** (solid particles; C, Si, O, Fe…)
- Gas-to-dust mass ratio = 100.
- Gas
  - **Hydrogen (H, 92% by number)**
  - **Helium (He, 8%)**
  - **Oxygen, Carbon, etc. (0.1%)**
  - Basic component = HI gas (n = 1 cm^{-3}, T = 100 K, typically)
  - Temperature is determined by the balance between heating and cooling
  - Density is determined by force balance (pressure, gravity)

2.3 ISM phases and star formation

ISM has various phases

1. Plasma (ionized diffuse phase)
2. Neutral gas (mainly neutral hydrogen HI)
3. Molecular gas (mainly molecular hydrogen H_2)

Since gas must become dense enough to form stars, star formation occurs in molecular clouds. Namely,

Atomic gas ⇒ Molecular gas ⇒ Stars

Spatial scales

Spatial scales of galaxies and star formation (SF) are some orders of magnitude different:

Galaxies ~ kpc
Star formation ~ a few pc (for molecular clouds)

However, global properties of galaxies and SF activity are mysteriously correlated in various aspects!

⇒ Meso-scale physics to connect the scales of a galaxy and SF should be explored.

Kennicutt-Schmidt (K-S) law

Stars form in molecular cores.

⇒ It is natural to suppose a relation between the star formation rate (SFR) and gas density. **Schmidt (1959) proposed a relation**

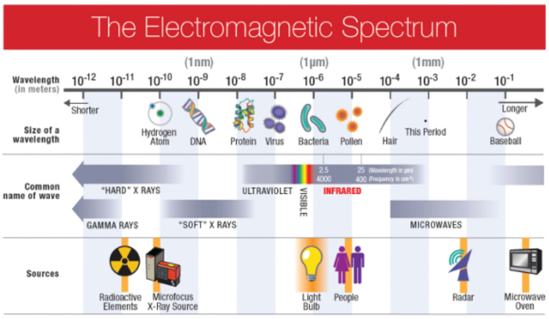
$$\mathrm{SFR} \propto \rho^n.$$

i. n = 1 Density controls star formation.
ii. n = 2 Collision-like process plays a role for star formation

⇒ **The power-law index contains substantial information on what triggers the star formation.**

3. Spectroscopic Observation of the ISM

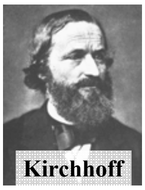
3.1 Electromagnetic waves



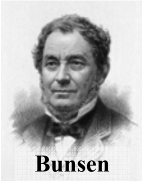
https://www.americanpharmaceuticalreview.com/Featured-Articles/331616-Optical-Spectroscopy-Where-is-it-Going/

3.2 What does spectroscopy tell us?

Kirchhoff and Bunsen showed that the emission lines corresponds to the absorption lines, which can be used for the identification of elements. Kirchhoff immediately pointed out that this leads to a very important application in astronomy.



Kirchhoff

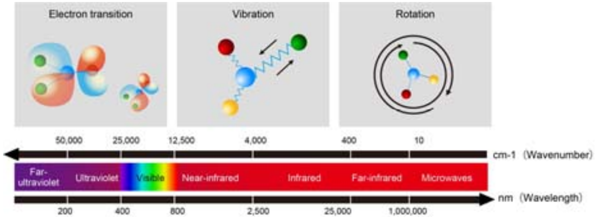


Bunsen

http://www.hao.ucar.edu/public/education/sp/images/kirchhoff.html
https://en.wikipedia.org/wiki/Robert_Bunsen

3.3 Quantum transition to spectral lines

Astronomical spectroscopy brings physical information of the objects in the remote Universe.



https://www.yokogawa.com/about/research-development/inv_center/spectroscopy/

4. Spectral Mapping in Astrophysics

4.1 General situation in astrophysics

Classical statistical analysis

Sample size: n
Data dimension: d

The following condition is implicitly assumed

$$n >> d$$

But this is not the case for many cases in scientific researches. Astronomers and astrophysicists have ever simply given up when they face such type of problem.

4. Spectral Mapping in Astrophysics

4.1 General situation in astrophysics

High-dimensional low-sample size (HDLSS) data analysis

Sample size: n
Data dimension: d

For the HDLSS data, the condition is

$$n << d$$

This condition is often found in e.g., genomic analysis, medical analysis, etc.

In astrophysics, for example, integral field spectroscopy has this property.

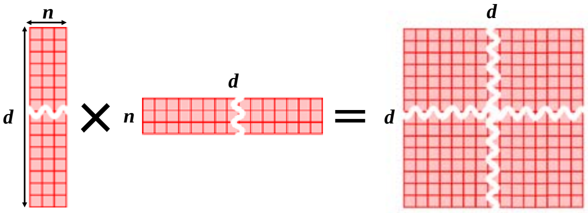
4.2 Geometric Representation

Dual representation of sample covariance matrix

When we draw a set of n samples from the parent population ($d > n$), $\vec{x}_1, \ldots, \vec{x}_n$.

The sample covariance matrix ($d \times d$) is $S = \frac{1}{n} X X^T$,

$X \equiv (x_1, x_2, \ldots, x_n)$



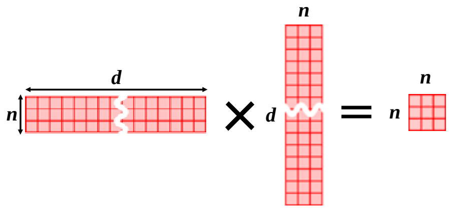
Note that this is a tremendously huge matrix!

4.2 Geometric Representation

Dual representation of sample covariance matrix

When we draw a set of n samples from the parent population ($d > n$), $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$.

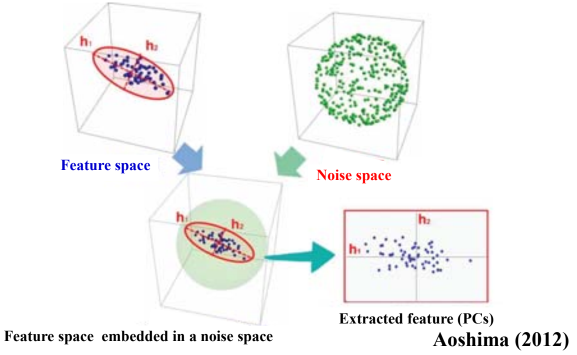
Consider a dual sample covariance matrix ($n \times n$), $\boldsymbol{S}_D = \frac{1}{n}\boldsymbol{X}^T\boldsymbol{X}$



This can be handled much more easily!

High-dimensional PCA

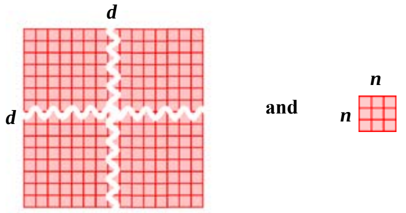
A specially designed PCA, the high-dimensional PCA, can sweep out the noise sphere and extract features of the data.



Aoshima (2012)

Eigenvalues of the dual covariance matrix

When we draw a set of n samples from the parent population ($d > n$), $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$.



share the first n eigenvalues, i.e., the same important statistical information!

4.2 Spectral mapping in astronomy

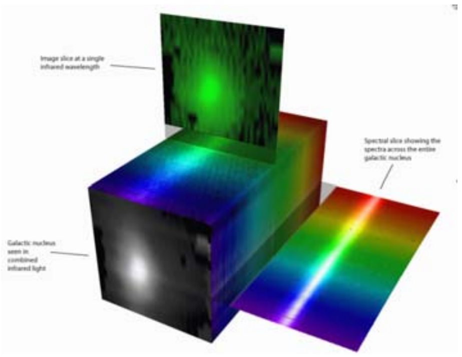
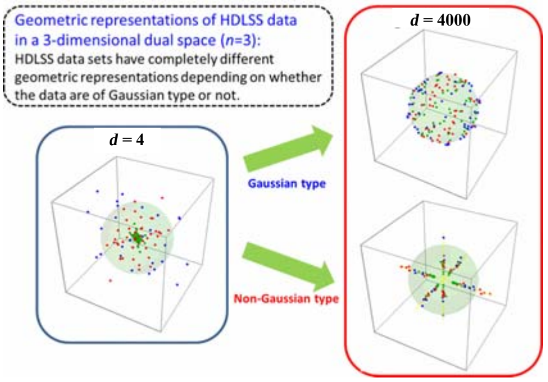


Image slice at a single infrared wavelength
Spectral slice showing the spectra across the entire galactic nucleus
Galactic nucleus seen in combined infrared light

http://ifs.wikidot.com/what-is-ifs

Unusual behavior of high-dimensional data

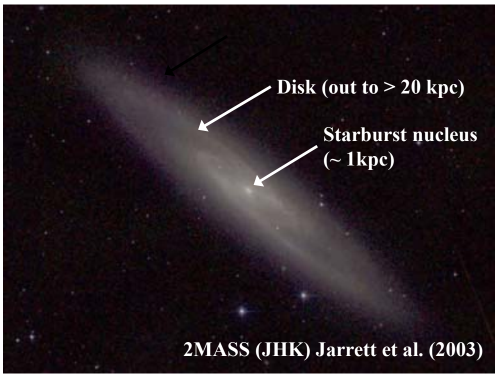


Geometric representations of HDLSS data in a 3-dimensional dual space (n=3):
HDLSS data sets have completely different geometric representations depending on whether the data are of Gaussian type or not.

https://www.math.tsukuba.ac.jp/~aoshima-lab/research.html

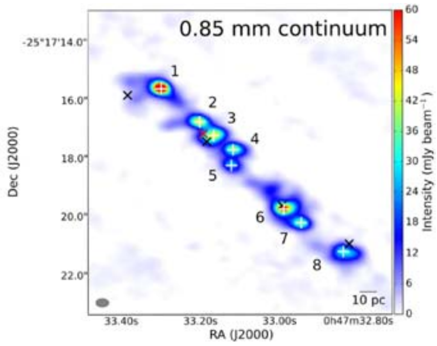
4.3 Actual data: ALMA data cube of NGC253

NGC 253: prototypal starburst



2MASS (JHK) Jarrett et al. (2003)

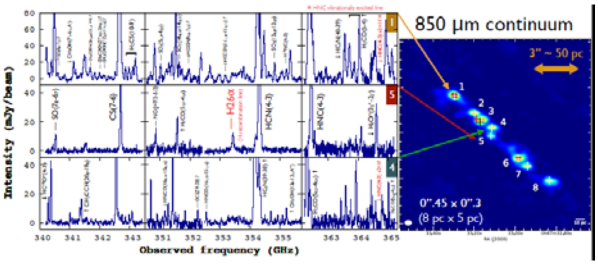
Close up of the starburst



Ando et al. (2017)

Rich in molecular lines

ALMA resolved diverse star-forming activities at ~ 10 pc scale.



ALMA Band7 spectra

Ando et al. (2017)

4.4 Structure of the Data

Data: Ando et al. (2017)

~ spatial dimension 231 × spectral dimension 2248

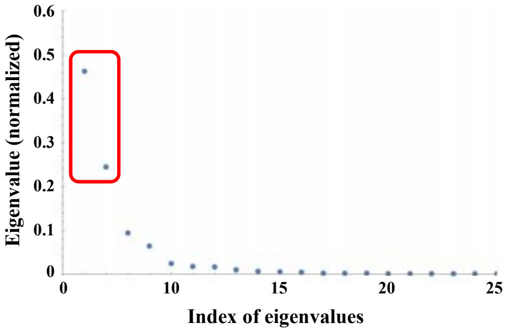
⇒ A case with n = 231 and d = 2248 ($n \ll d$)

Problems from astrophysical side

- Too much information on spectra.
- Too large variety of spectral lines compared to n.

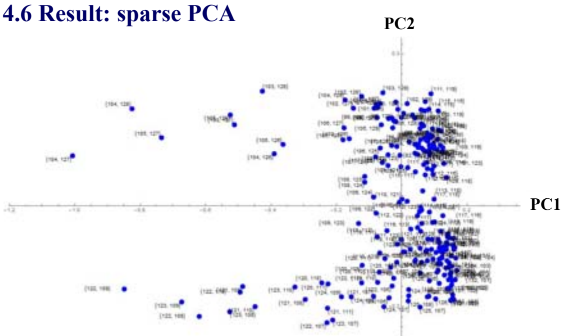
We apply the high-dimensional statistical analysis to the ALMA spectral mapping data of NGC253.

4.5 Result: eigenvalues of the NGC253



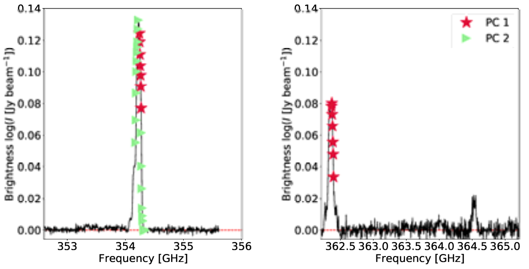
The huge amount of information on the ALMA spectra are basically determined by two largest eigenvalues.

4.6 Result: sparse PCA



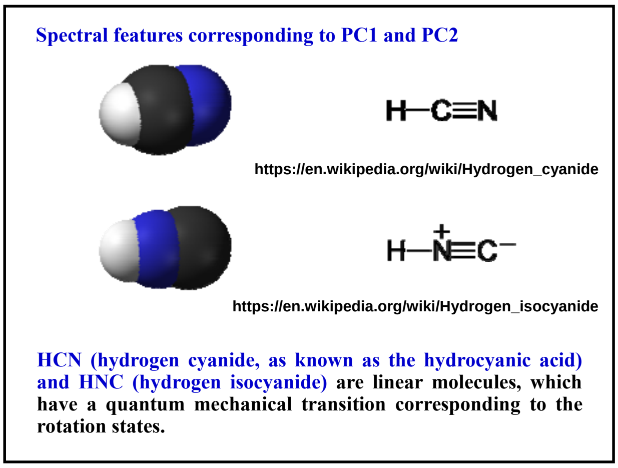
PC1 and 2 consist of ~ 20 elements (spectral features on the resolution units). The key features may be reduced only to a few to several lines!

Responsible spectral features for PC1 and PC2



Takeuchi et al. (2022)

Indeed the PCs correspond to spectral line features. The lines are HCN(4-3) and HNC(4-3). PC1 represents the total intensity, and PC2 the Doppler shift of the lines.

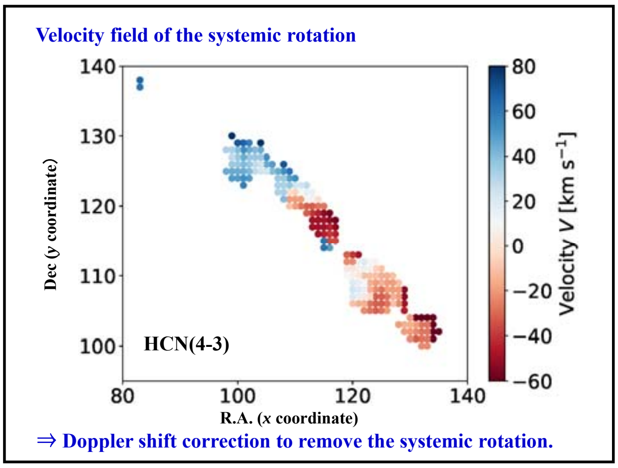


Spectral features corresponding to PC1 and PC2

https://en.wikipedia.org/wiki/Hydrogen_cyanide

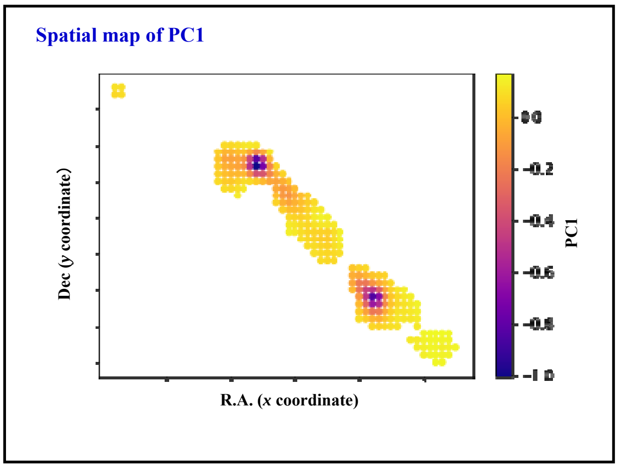
https://en.wikipedia.org/wiki/Hydrogen_isocyanide

HCN (hydrogen cyanide, as known as the hydrocyanic acid) and HNC (hydrogen isocyanide) are linear molecules, which have a quantum mechanical transition corresponding to the rotation states.

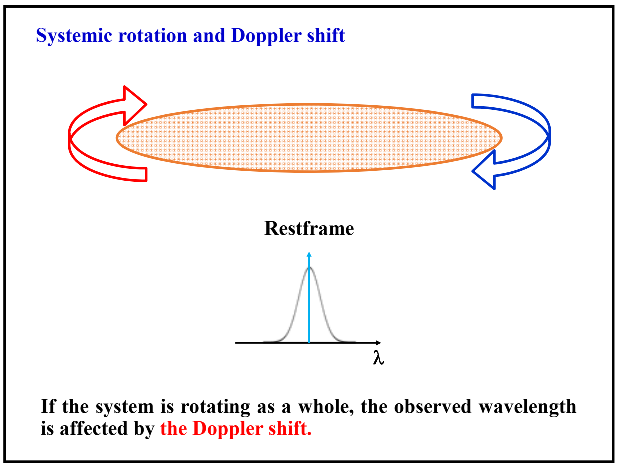


Velocity field of the systemic rotation

⇒ Doppler shift correction to remove the systemic rotation.

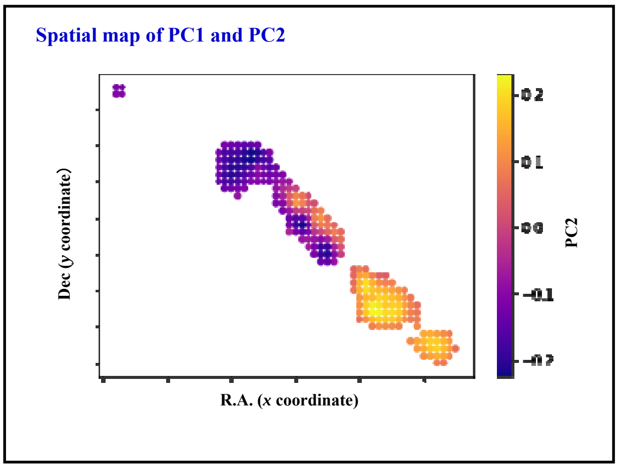


Spatial map of PC1

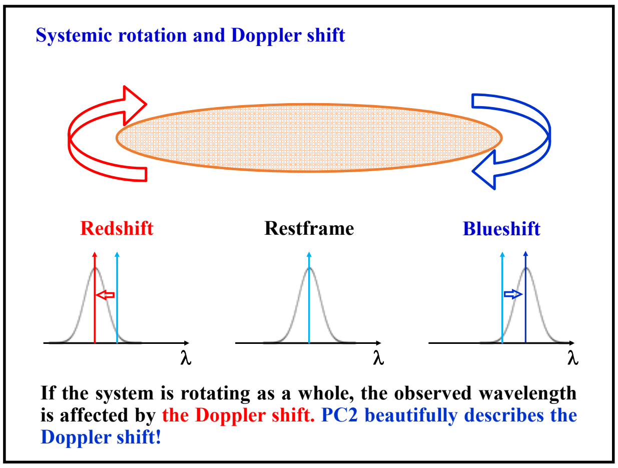


Systemic rotation and Doppler shift

If the system is rotating as a whole, the observed wavelength is affected by the Doppler shift.



Spatial map of PC1 and PC2

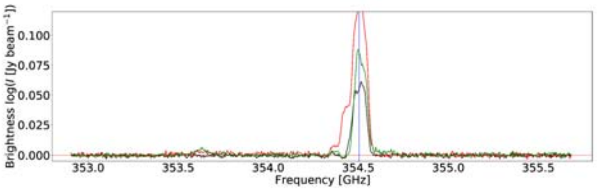


Systemic rotation and Doppler shift

If the system is rotating as a whole, the observed wavelength is affected by the Doppler shift. PC2 beautifully describes the Doppler shift!

4.7 Main analysis

Doppler shift correction

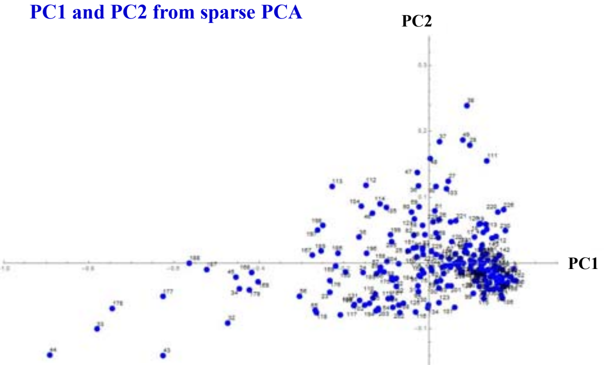


Takeuchi et al. (2021)

We estimated the peculiar velocity field (mainly due to the systemic rotation of the central region of NGC253) by averaging the results from HCN(4-3), HNC(4-3) and CS(7-6) lines, and corrected the Doppler shift.

Due to this correction, the final data dimension is $d = 1971$.

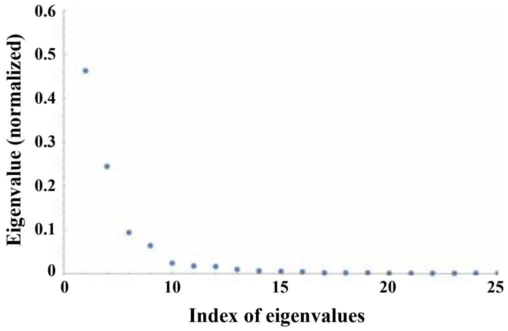
PC1 and PC2 from sparse PCA



Takeuchi et al. (2021)

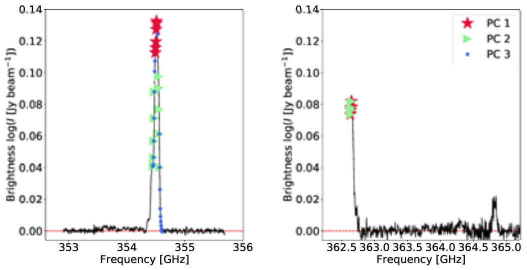
Butterfly-like pattern completely disappeared.

Eigenvalues of the NGC253 before Doppler correction



Takeuchi et al. (2021)

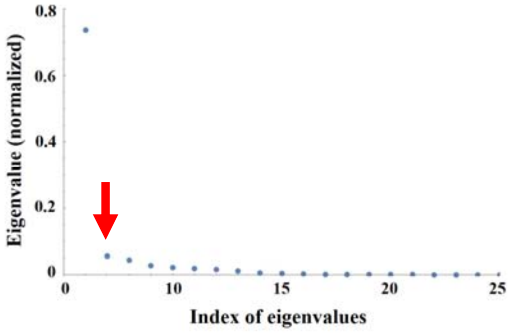
Responsible spectral features for PC1, PC2 and PC3



Takeuchi et al. (2022)

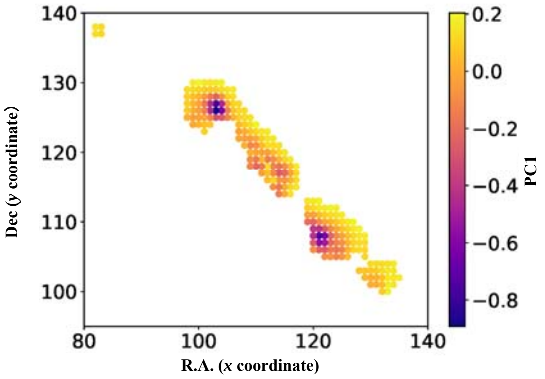
Now PC1 more clearly represents the total intensity, and PC2 and 3 represent smaller-scale velocity structures.

Eigenvalues of the NGC253 after Doppler correction

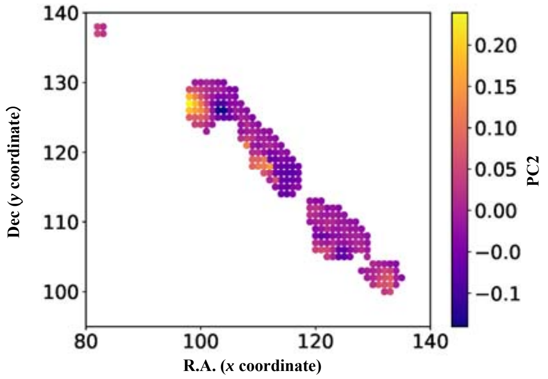


Takeuchi et al. (2021)

Spatial map of PC1 after Doppler correction



Spatial map of PC2 after Doppler correction



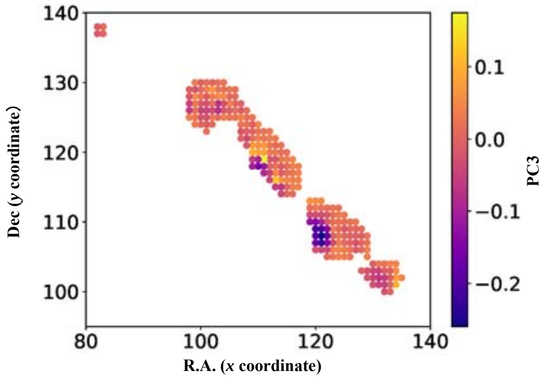
What do we see from the Doppler-corrected map?

NGC253

- Pure starburst: SFR in the central molecular zone is 2 $M_\odot$ yr^{-1} (Rieke et al. 1980; Keto et al. 1999)
- Intense outflow (Matsubayashi et al. 2009; Bolatto et al. 2013)

Indeed the outflow phenomenon is mainly delineated by PC3.

Spatial map of PC3 after Doppler correction



5. Analysis of the HI Forest

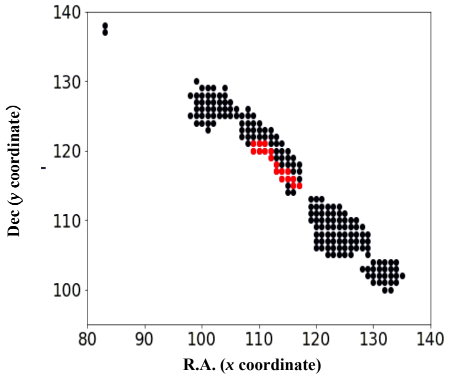
5.1 What is the HI forest?

Quasars

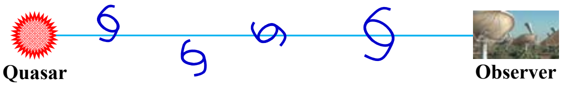
There are very luminous energetic objects at the center of galaxies. They are powered by central black holes, called quasars.

They are luminous at all the electromagnetic wavelength ranges.

Anomaly regions in the velocity field



HI forest



HI stands for atomic hydrogen.

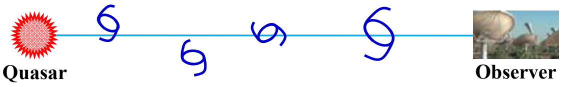
First galaxies are believed to have plenty of HI gas, and they make a shadow on the spectrum of quasars as absorption lines.

These absorption lines look like a forest on the quasar spectrum.

Especially, the absorption line systems caused by the atomic hydrogen is referred to as the HI forest.

5.2 Prospect and difficulty in the analysis of HI forest

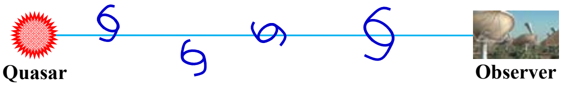
Prospect



- The HI forest carry the information on the spatial distribution of primordial galaxies. This provides a very important clue to the formation of first galaxies.
- The HI forest absorption line systems have evolved into galaxies at later epochs of the Universe. Their evolution might be reflected to the absorption lines.

5.2 Prospect and difficulty in the analysis of HI forest

Difficulty



The background quasars are very rare.

Even by the next-generation radio observational facility Square Kilometre Array (SKA), only a few tens of quasars are expected., while the absorption lines are numerous.

⇒ HDLSS data!

We constructed a new analysis method based on the high-dimensional statistical analysis.

5. Summary

1. Galaxies ubiquitously exist in the present-day Universe, but they have been formed from a tiny fluctuation of matter in the early Universe.
2. Evolution of galaxies is mainly driven by the star formation, a transition from ISM to stars.
3. Various phases of the ISM are related, and the evolution of the ISM is a key to complete the understanding of the galaxy evolution.
4. Spectroscopic observations are of vital importance to extract and interpret the information of matter in galaxies.

5. Summary

5. Spectroscopic mapping and similar methods are fundamentally important to reveal the ISM physics, but the data are high-dimensional low sample size.
6. We applied the high-dimensional PCA on the NGC253 spectral map. ALMA mapping data are typically HDLSS in general, and in this case $n = 231$ and $d = 2228$.
7. Very large variety in the molecular line spectra of NGC253 map can be described only by two PCs! Each PC consists of ~ 20 elements, much fewer than d. Because these elements may be a part of same features, the key features may be reduced to several.

5. Summary

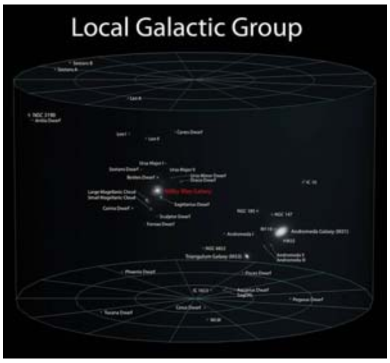
8. The high-dimensional sparse PCA successfully chose two PCs that reproduce the general properties of the ALMA spectroscopic map of NGC253.
9. The controlling feature was HCN(4-3) rotational lines. PC1 describes the total intensity of the lines, and PC2 represents the Doppler shift caused by the systemic rotation.
10. After correcting the Doppler shift due to the systemic rotation, we could obtain information on the smaller-scale velocity field described by PC2 (new) and PC3. These may be caused by outflow phenomena of starburst regions.

5. Summary

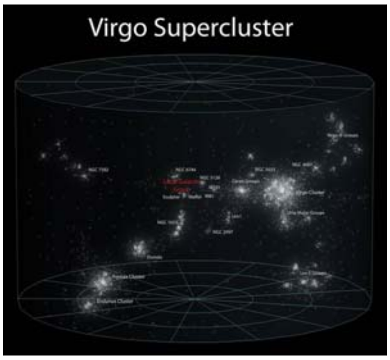
11. The spatial distribution and evolution of the hydrogen absorption line systems (HI forest) can be efficiently explored by the high-dimensional statistical analysis.

Appendix

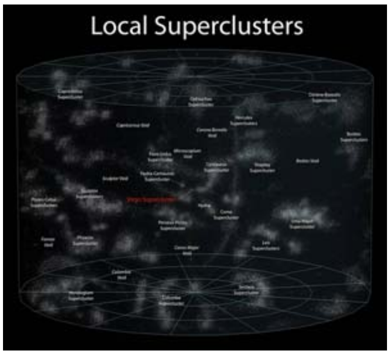
From galaxies to groups and clusters of galaxies



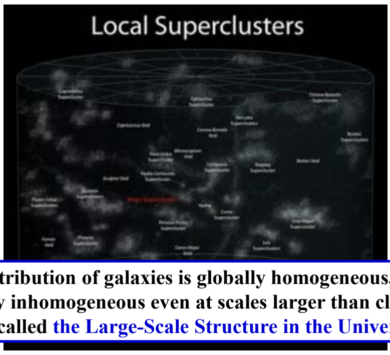
From galaxies to groups and clusters of galaxies



From groups and clusters to the Large-Scale Structure

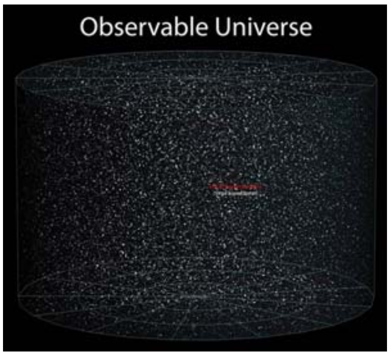


From groups and clusters to the Large-Scale Structure



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.

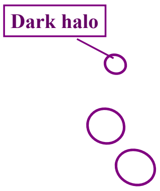
From the Large-Scale Structure to the Hubble horizon



1.3 Formation and evolution of galaxies

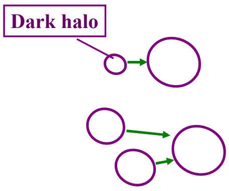
The hierarchical structure formation

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.



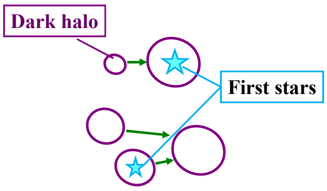
The hierarchical structure formation

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.



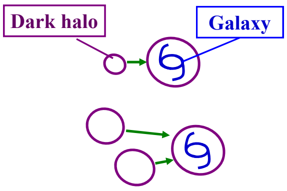
The hierarchical structure formation

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.



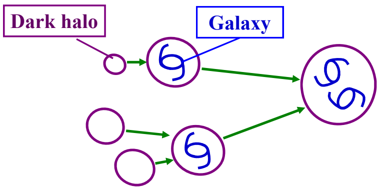
The hierarchical structure formation

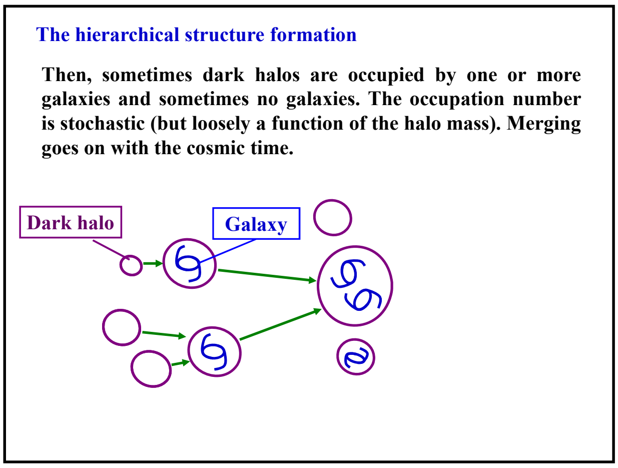
The supply of heavy elements makes the condition of star formation much easier. Then, the gas turns into stars collectively, and galaxies form as large agglomerations of stars and remaining gas in dark halos.



The hierarchical structure formation

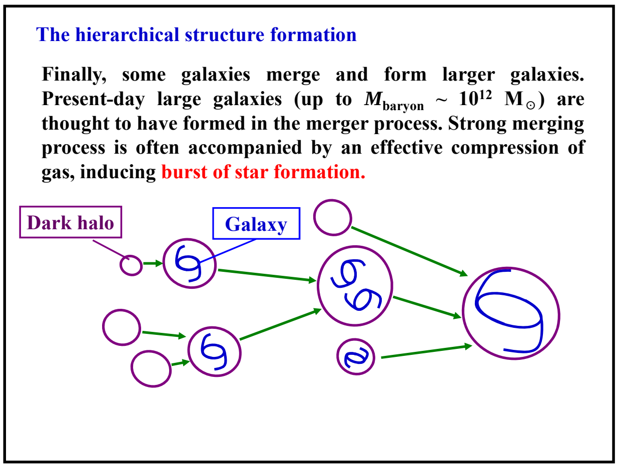
Dark halos continue merging and form larger and larger halos. Consequently, galaxies in these halos start to cohabit in the same newly formed halos. Baryonic structures cannot merge as easily as dark halos because of gas pressure.





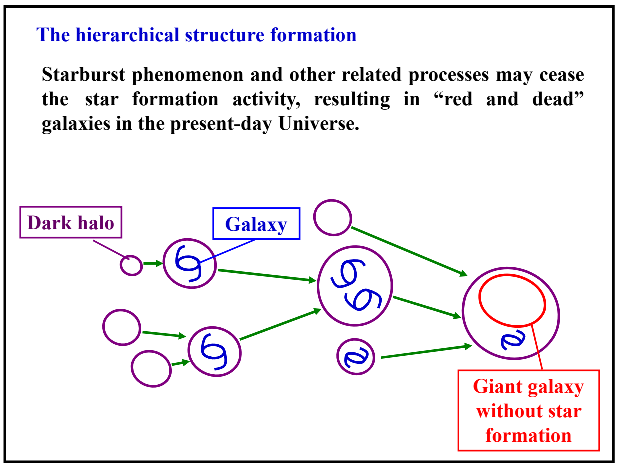
The hierarchical structure formation

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.



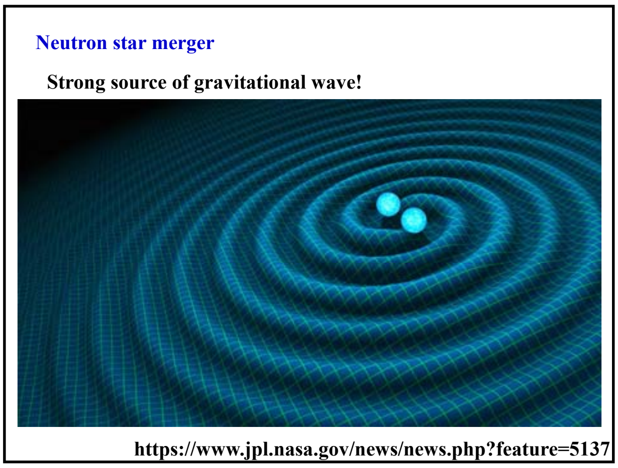
The hierarchical structure formation

Finally, some galaxies merge and form larger galaxies. Present-day large galaxies (up to $M_{\rm baryon}$ ~ 10^{12} $M_\odot$) 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.



The hierarchical structure formation

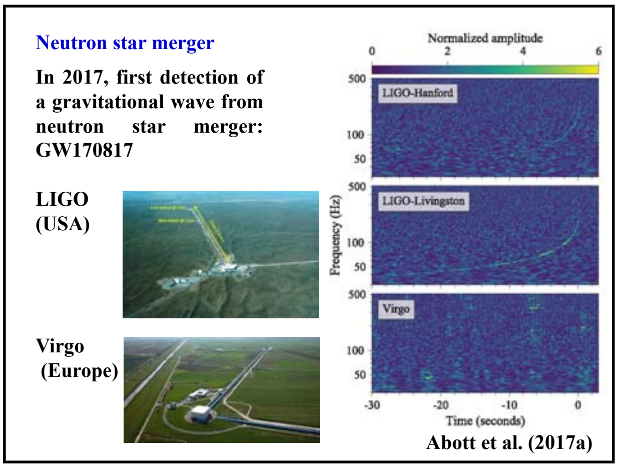
Starburst phenomenon and other related processes may cease the star formation activity, resulting in “red and dead” galaxies in the present-day Universe.



Neutron star merger

Strong source of gravitational wave!

https://www.jpl.nasa.gov/news/news.php?feature=5137



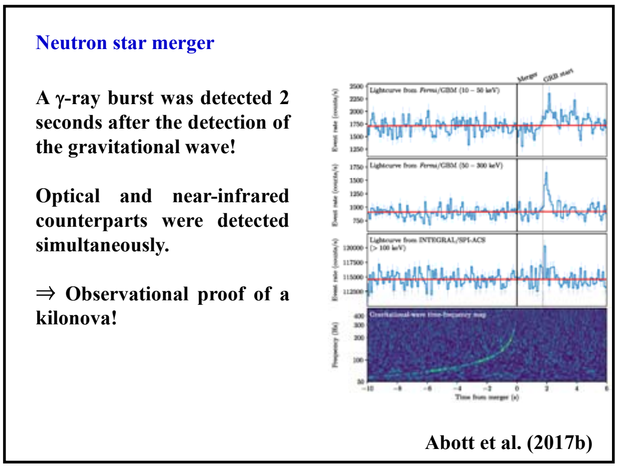
Neutron star merger

In 2017, first detection of a gravitational wave from neutron star merger: GW170817

LIGO (USA)

Virgo (Europe)

Abott et al. (2017a)



Neutron star merger

A γ-ray burst was detected 2 seconds after the detection of the gravitational wave!

Optical and near-infrared counterparts were detected simultaneously.

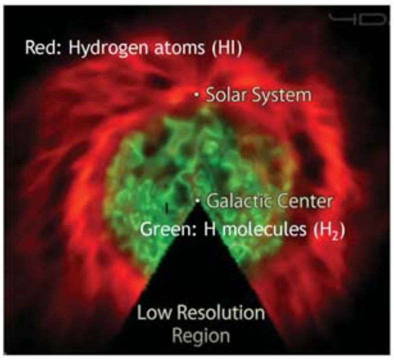
⇒ Observational proof of a kilonova!

Abott et al. (2017b)

Phase of the ISM

	State of H & C	Temperature	Densities (H/cm^3)	Volume fraction
HII Regions & Planetary Nebulae	H, C Ionized	5000 K	0.5	< 1%
Diffuse ISM	H, C Ionized	1,000,000 K	0.01	50%
Diffuse Atomic	$H_2 < 0.1$ C Ionized	30-100 K	10-100	30%
Diffuse Molecular	$0.1 < H_2 < 50\%$ $C^+ > 50\%$	30-100 K	100-500	10%
Translucent Molecular	$H_2 \sim 1$ $C^+ < 0.5$, $CO < 0.9$	15-50 K	500-5000?	Small
Dense Molecular	$H_2 \sim 1$ $CO > 0.9$	10-50 K	$> 10^4$	10%

Spatial distribution of HI and H_2 in the Milky Way



Unusual behavior of high-dimensional data

For high-dimensional data, classical limit theorems do not work. If we wrongly assume them, we would be lead to a wrong conclusion.

Simplest example: for the sample mean

$$\bar{\vec{x}} = \frac{1}{n}\sum_{i=1}^{n} \vec{x}_i$$

1. as *d* /*n*→ 0

$$\| \bar{\vec{x}} - \vec{\mu} \| \xrightarrow{P} \vec{0}$$

2. as *d* /*n*→ ∞

$$\| \bar{\vec{x}} - \vec{\mu} \| \xrightarrow{P} \infty$$

This striking property is referred to as the strong inconsistency.