# Activities

In working directory.

> git init

> git pull git@github.com:brittlundgren/py-astro-stat.git master

Create a directory for each week so far.

Your activity will be located in the week 5 directory.

The data you need for the activity will be in the 'data' directory.

# Python Astro Statistics

Statistics, Data Mining, and Machine Learning in Astronomy
Chapter 4½

June 17th, 2014

Ben Tofflemire & Elijah Bernstein-Cooper

# Today’s Lesson

*Lecture*

1. Bootstrapping
2. Comparing Distributions
3. Selection Effects & The Luminosity Function
4. Histogram Bins and Errors
5. Some helpful Python tips

*Group Activities*

*Group Presentations*

# 1. Bootstrapping Motivation

I want to estimate the uncertainty on a measurement.

Problem:

I have no idea what the underlying error distribution is. I cannot simply use the analytical formulae Karen gave me last week!

Solution:

Who cares what the underlying distribution is, lets bootstrap it!

# Bootstrapping Method

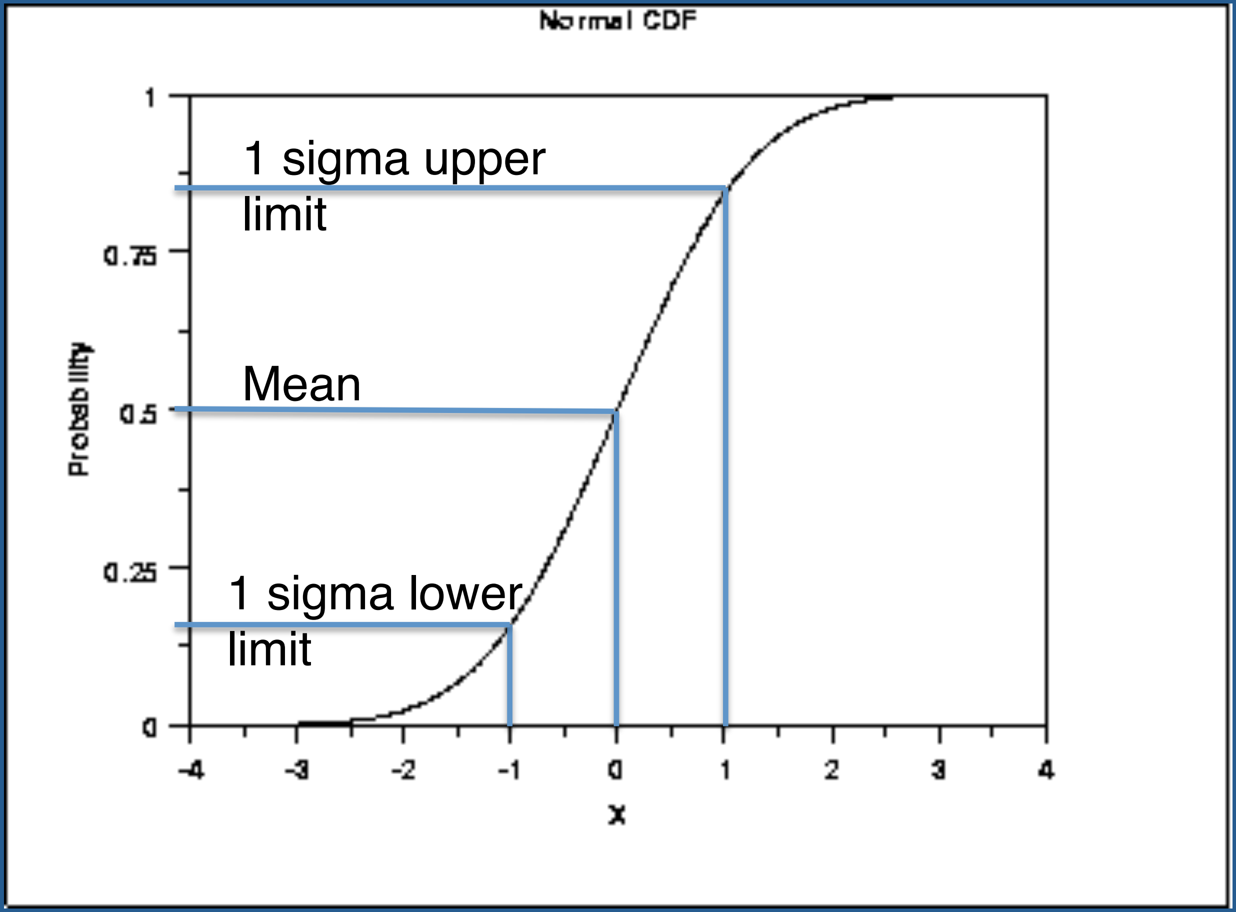
Example: Error in the Mean

1) From your original sample, create N new samples by randomly drawing data from the original with replacement.
   - Make sure it is a <u>UNIFORM</u> random distribution!

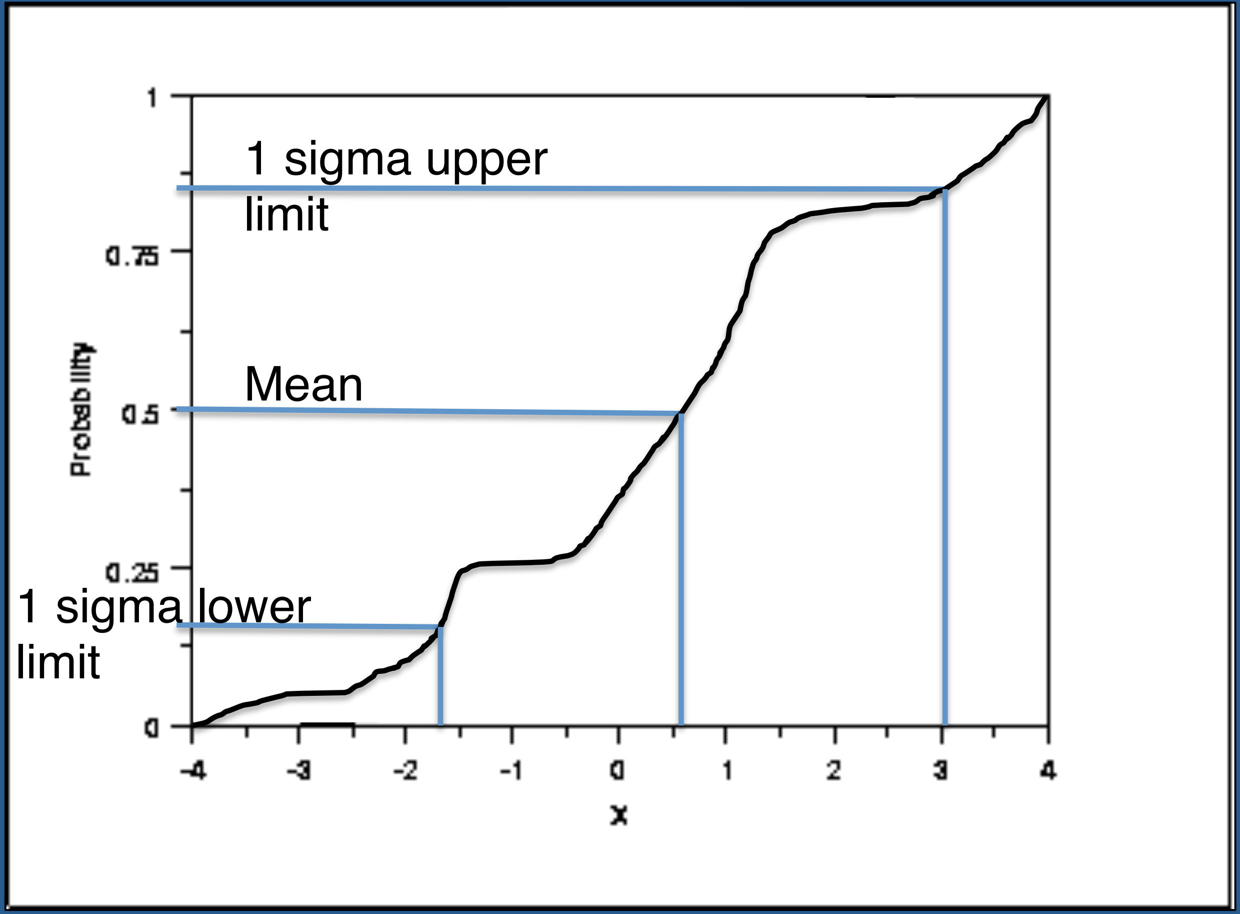
2) Calculate the Mean for each resampling

3) Create a CDF of the resampled means.
   - This is your error distribution!

# Bootstrapping Method



# Bootstrapping Method



# Jackknife Alternative

# 2. Comparing Distributions

Motivation:

Is my distribution Gaussian?

- Can I use the analytical formulae Karen gave me last week?

Are two samples drawn from the same distribution?

Solution:

Kolmogorov-Smirnov (K-S) Test to the rescue!

# K-S Test Method

Is my distribution Gaussian?

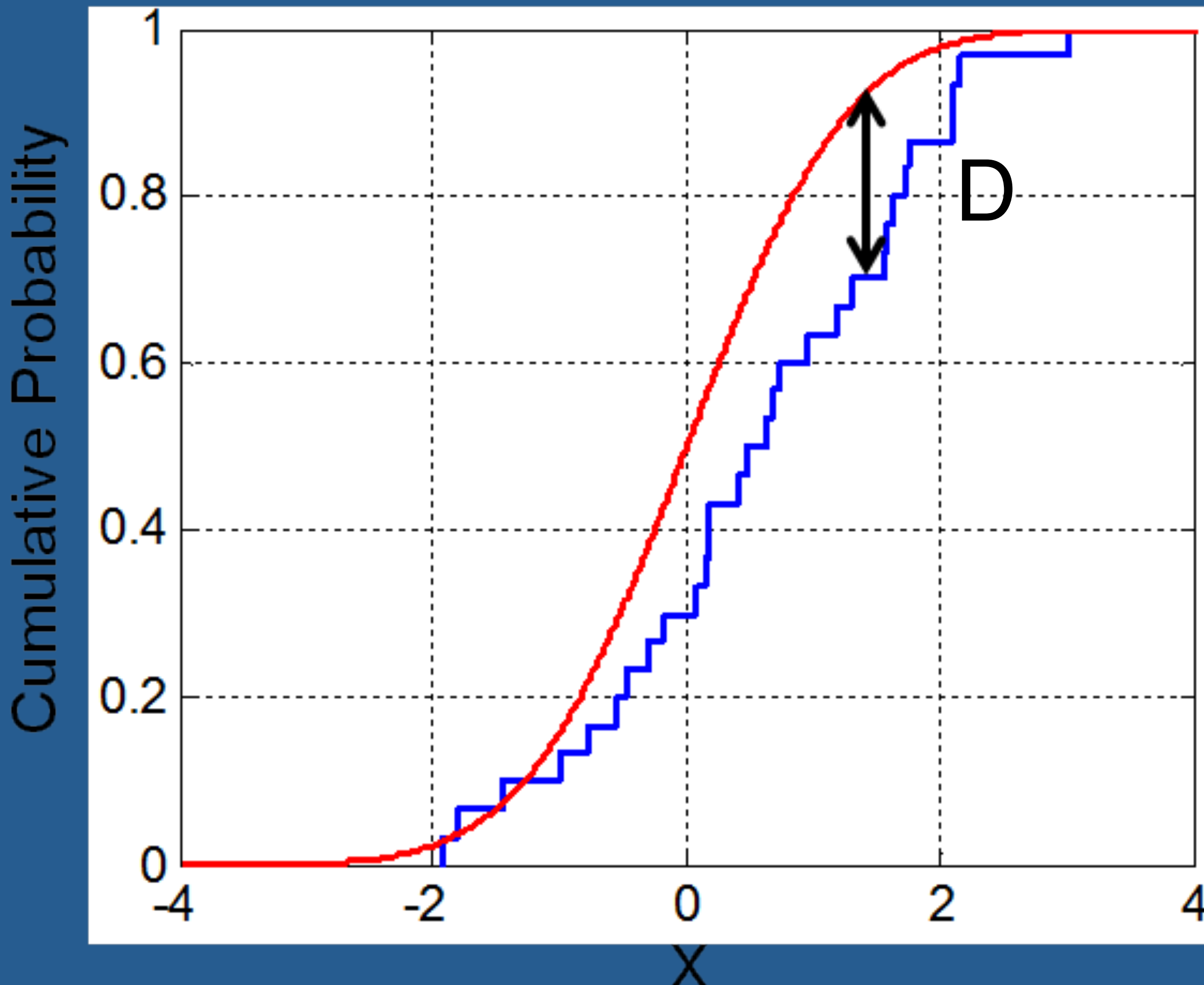
- From your sample create a CDF.

- Find the the maximum vertical difference between your CDF and a Gaussian CDF ('sup' = Supreme).

$$D_n = \sup_x |F_n(x) - F(x)|$$

- If your sample is Gaussian, D -> 0 as n -> infinity.

# K-S Test Method



# K-S Test Method

- The D value corresponds to a probability ($\alpha$) that the two samples were drawn from the same distribution (depends on number of data points in the sample).

Caveat:

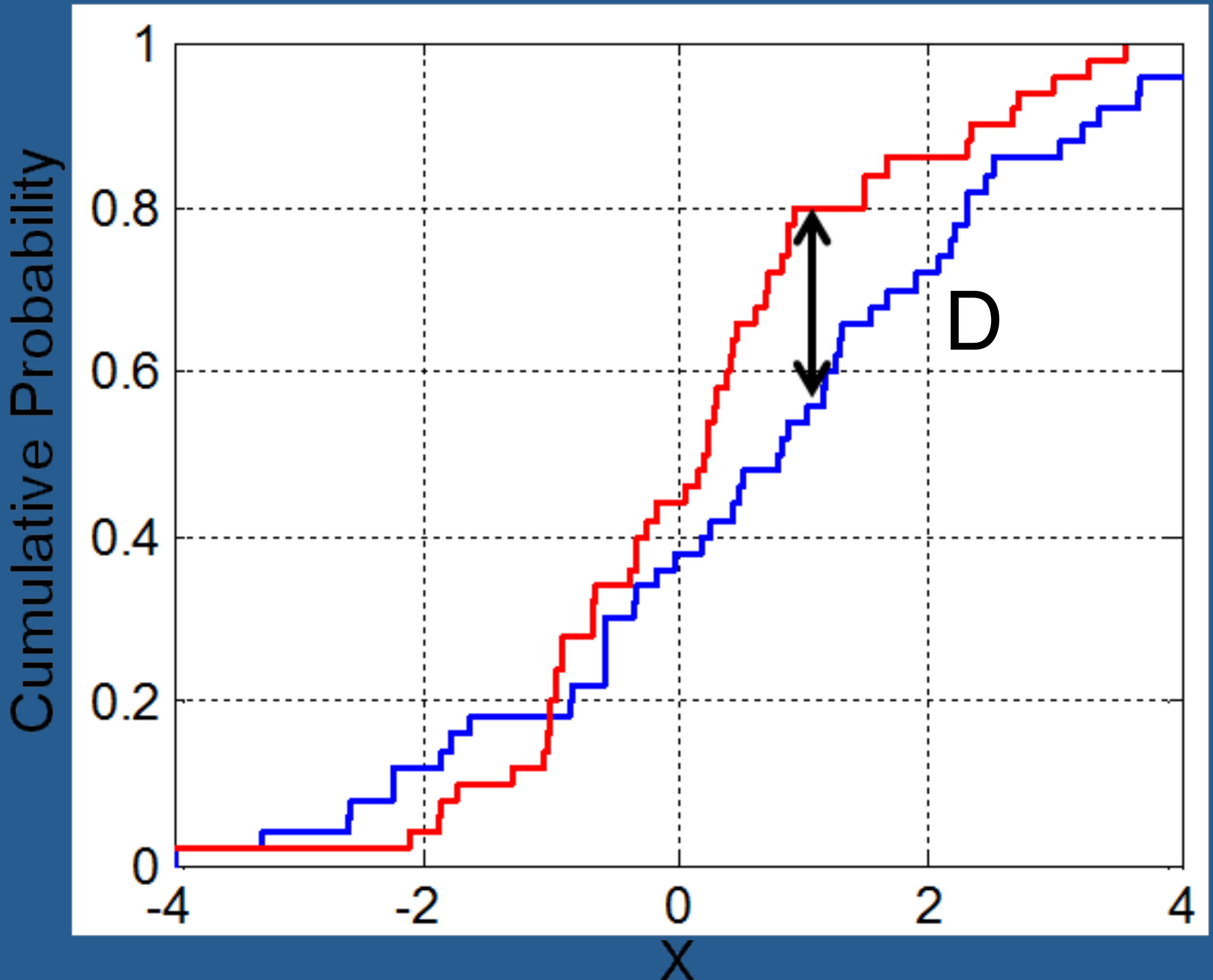
If $\alpha$ is above some limit it does <u>not</u> explicitly tell you your distribution *is* Gaussian, it tells it is *indistinguishable* from a Gaussian distribution.

Upside:

Extremely flexible and simple.

Does not require well sampled data in principle.

# K-S Test Method



# 2. Comparing Distributions

There a plethora of other tools besides the K-S test to compare a sample with a Gaussian distribution.

A few options are:

- Anderson Darling
- Shapiro-Wilk
- $Z_1$
- $Z_2$

All have their own strengths and weaknesses detailed in Chapter 4.7

# 3. Selection Effects

Problem:

Often an observed distribution ($f(x)$) will be biased from the true distribution ($h(x)$) by some selection function ($S(x)$).

In General:

$$h(x) = f(x) / S(x)$$

# 3. Selection Effects

Example: The Luminosity Function

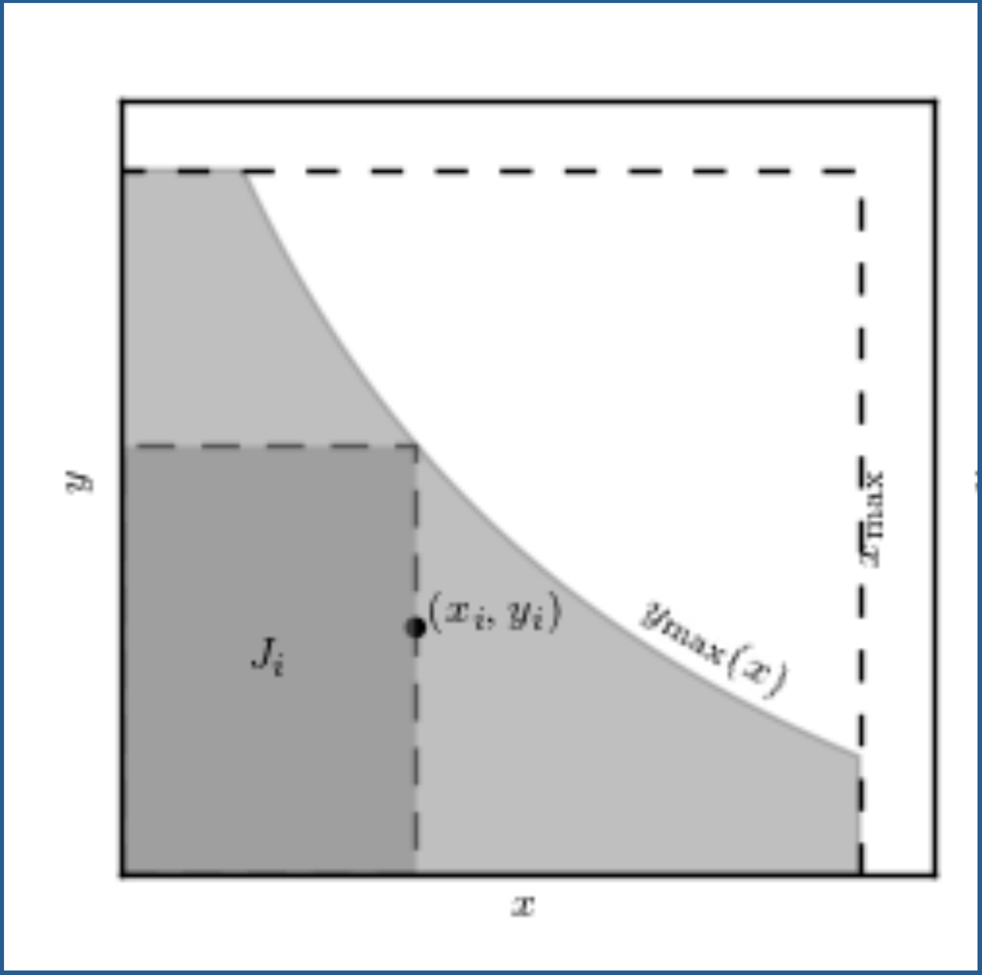
The number of galaxies per unit luminosity (or absolute magnitude M+dM) per unit volume.

Analytical from from Schechter (1979).

$$\phi(M) = \frac{\ln 10}{2.5}\phi^* 10^{0.4(\alpha+1)(M-M^*)} exp\left[-10^{0.4(M-M^*)}\right]$$

# 3. Selection Effects

Example: The Luminosity Function



# 3. Selection Effects

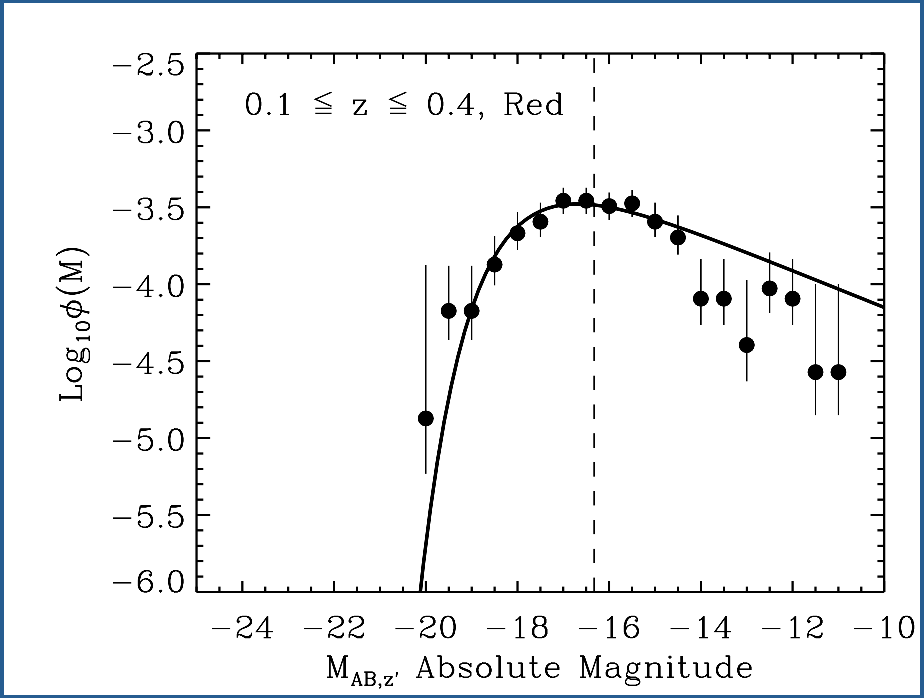
Example: The Luminosity Function

$$\phi(M) = \frac{\ln 10}{2.5}\phi^* 10^{0.4(\alpha+1)(M-M^*)} exp\left[-10^{0.4(M-M^*)}\right]$$

$\phi^*$ : 

$M^*$: 

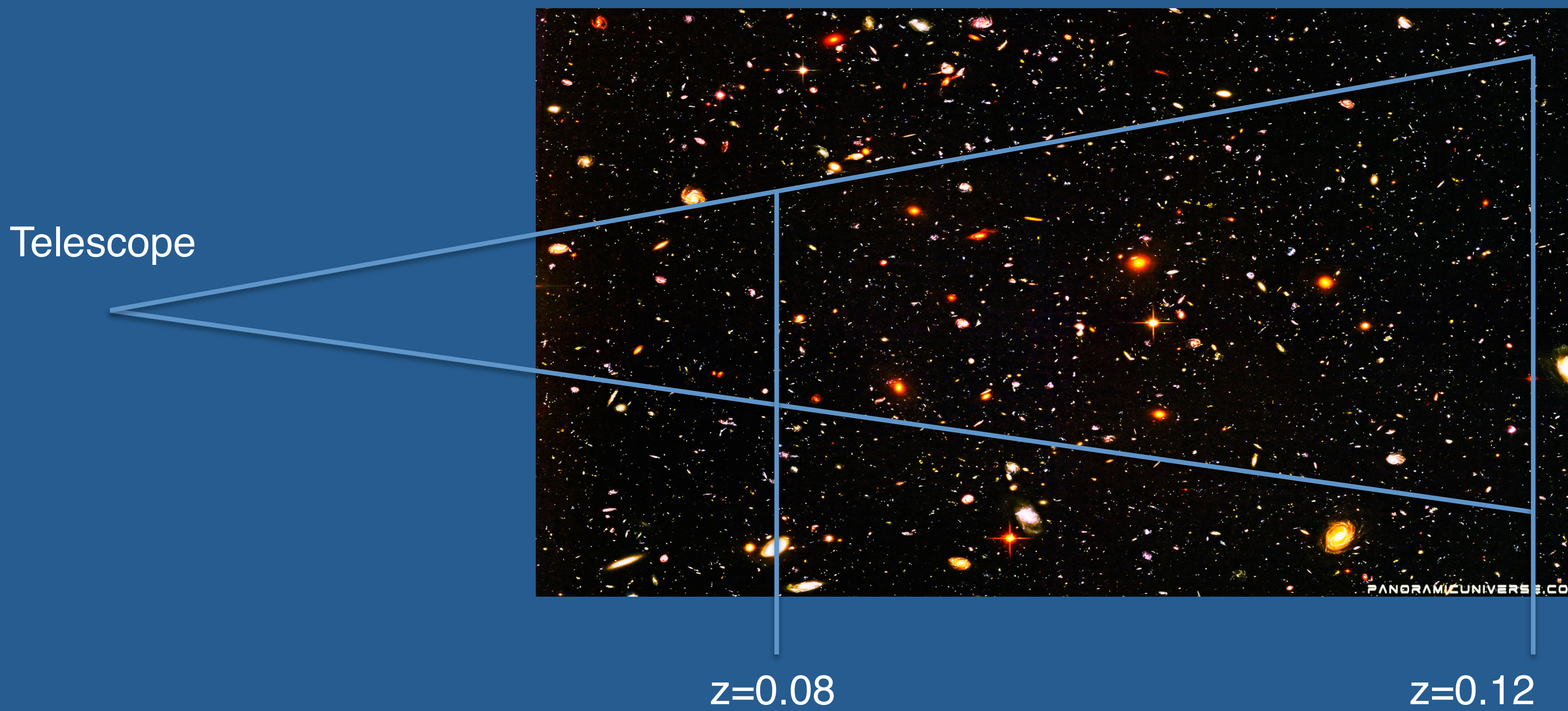
$\alpha$ : 



# 3. Selection Effects

Example: The Luminosity Function

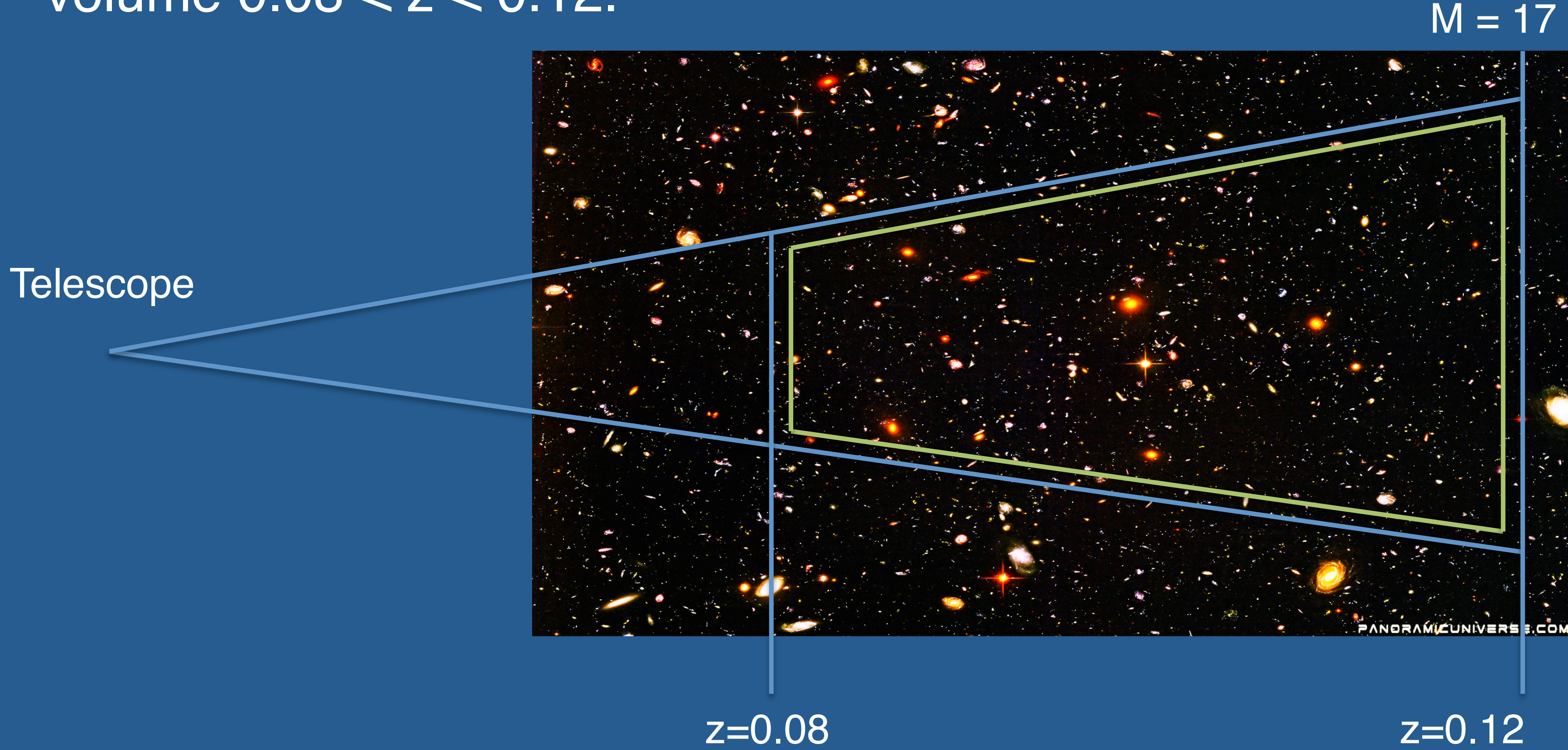
Say we care to probe galaxies with $-11 < M < -21$ in the volume $0.08 < z < 0.12$.



# 3. Selection Effects

Example: The Luminosity Function

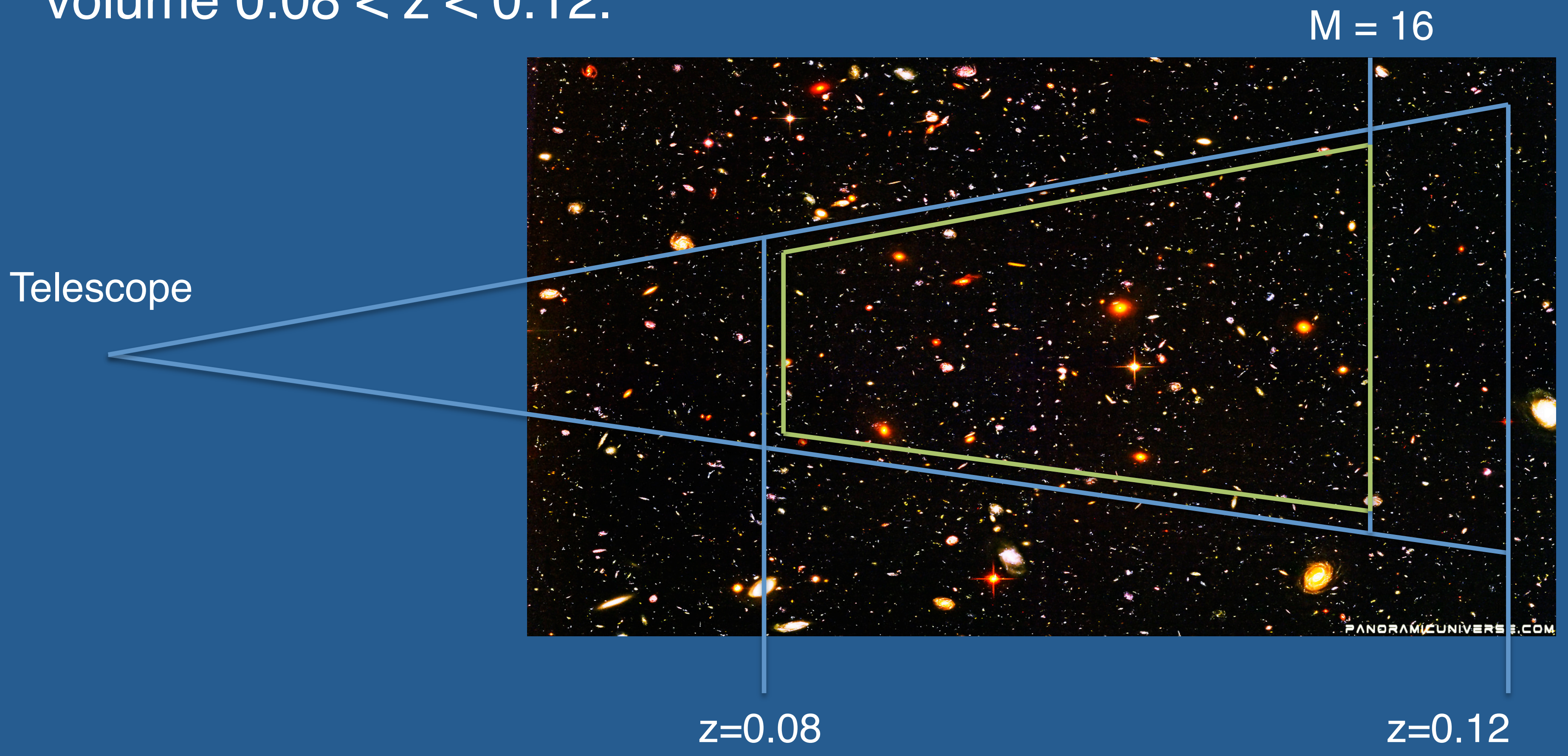
Say we care to probe galaxies with -11 < M < -21 in the volume 0.08 < z < 0.12.



# 3. Selection Effects

Example: The Luminosity Function

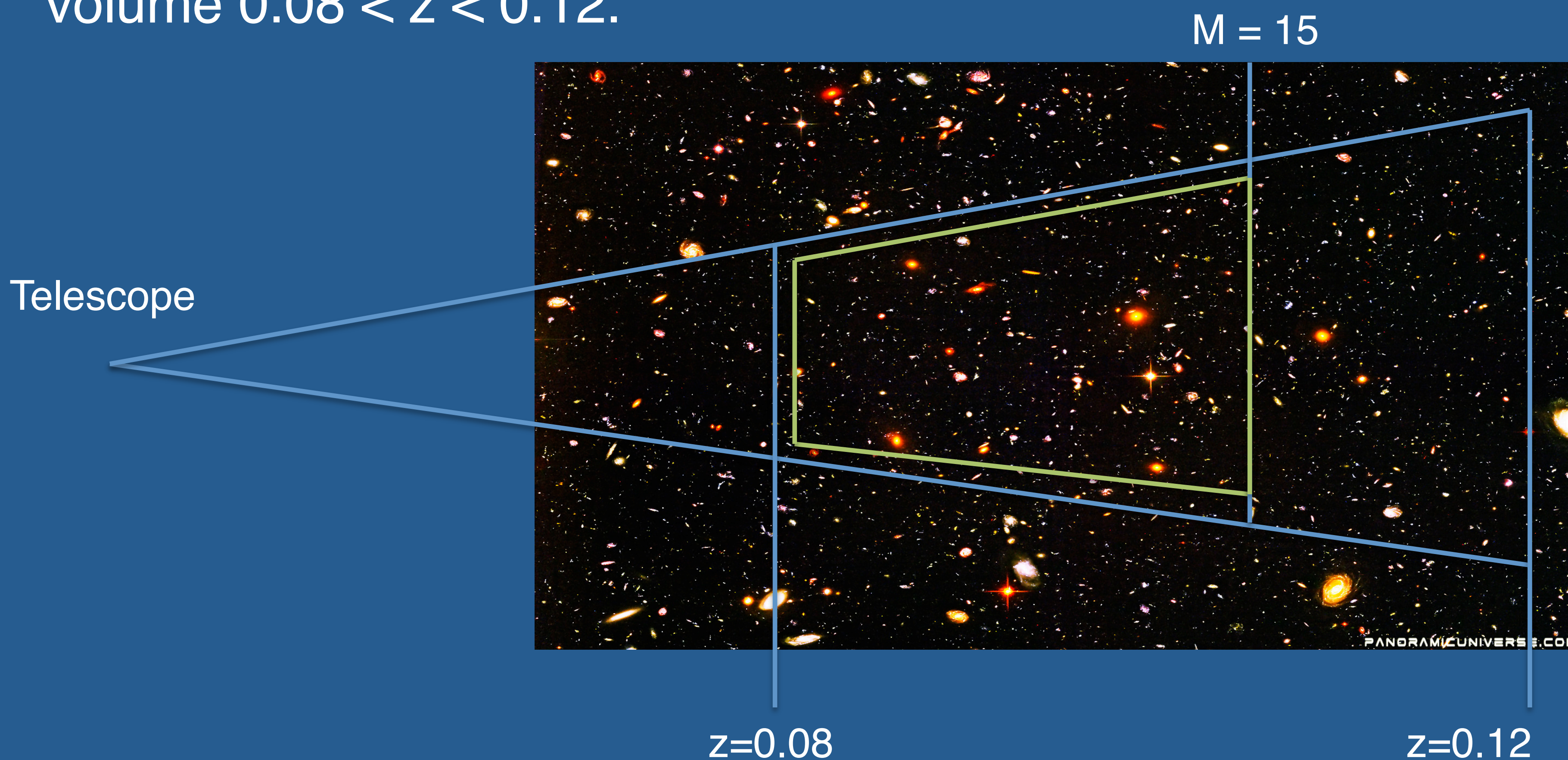
Say we care to probe galaxies with -11 < M < -21 in the volume 0.08 < z < 0.12.



# 3. Selection Effects

Example: The Luminosity Function

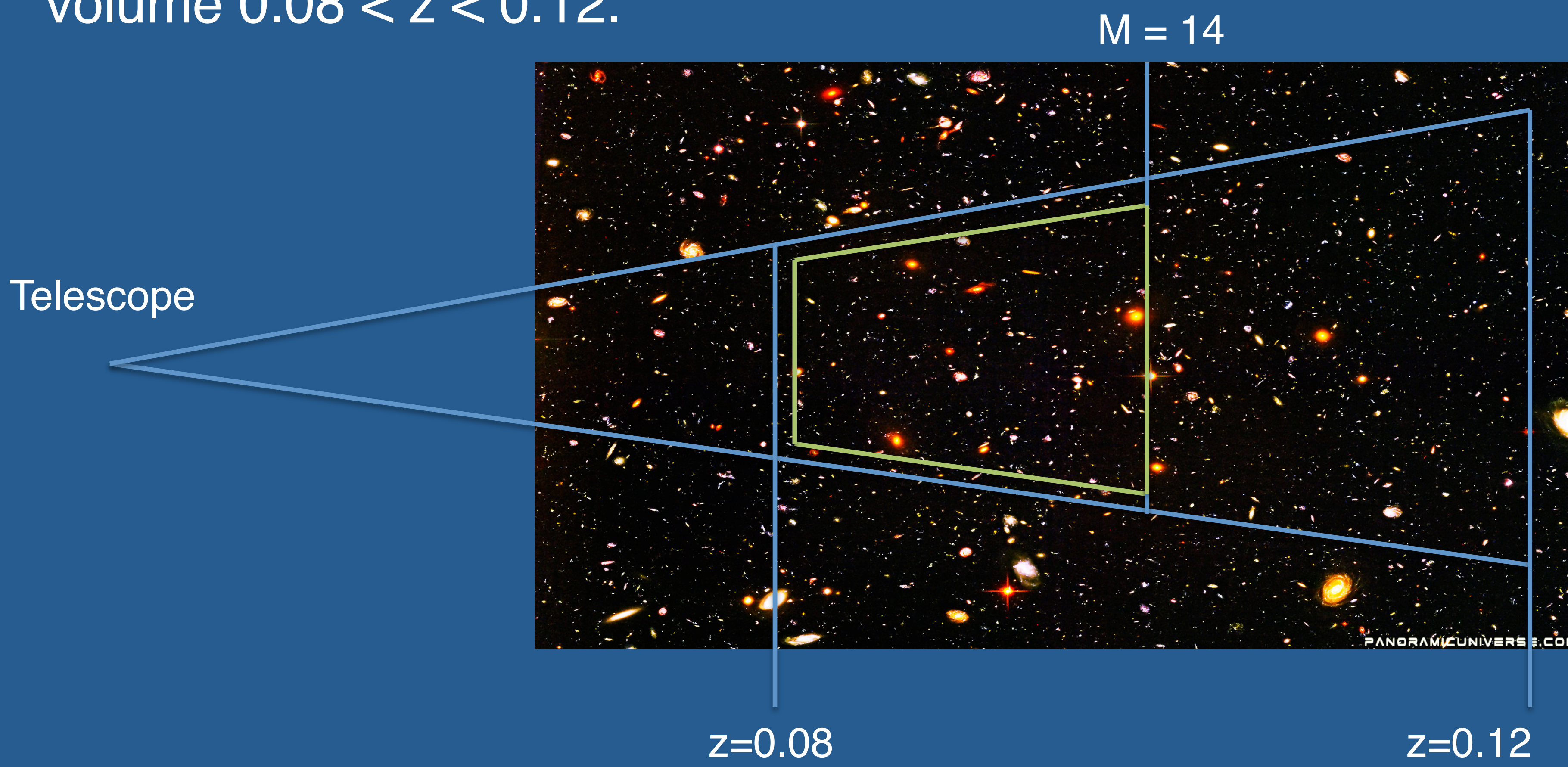
Say we care to probe galaxies with $-11 < M < -21$ in the volume $0.08 < z < 0.12$.



# 3. Selection Effects

Example: The Luminosity Function

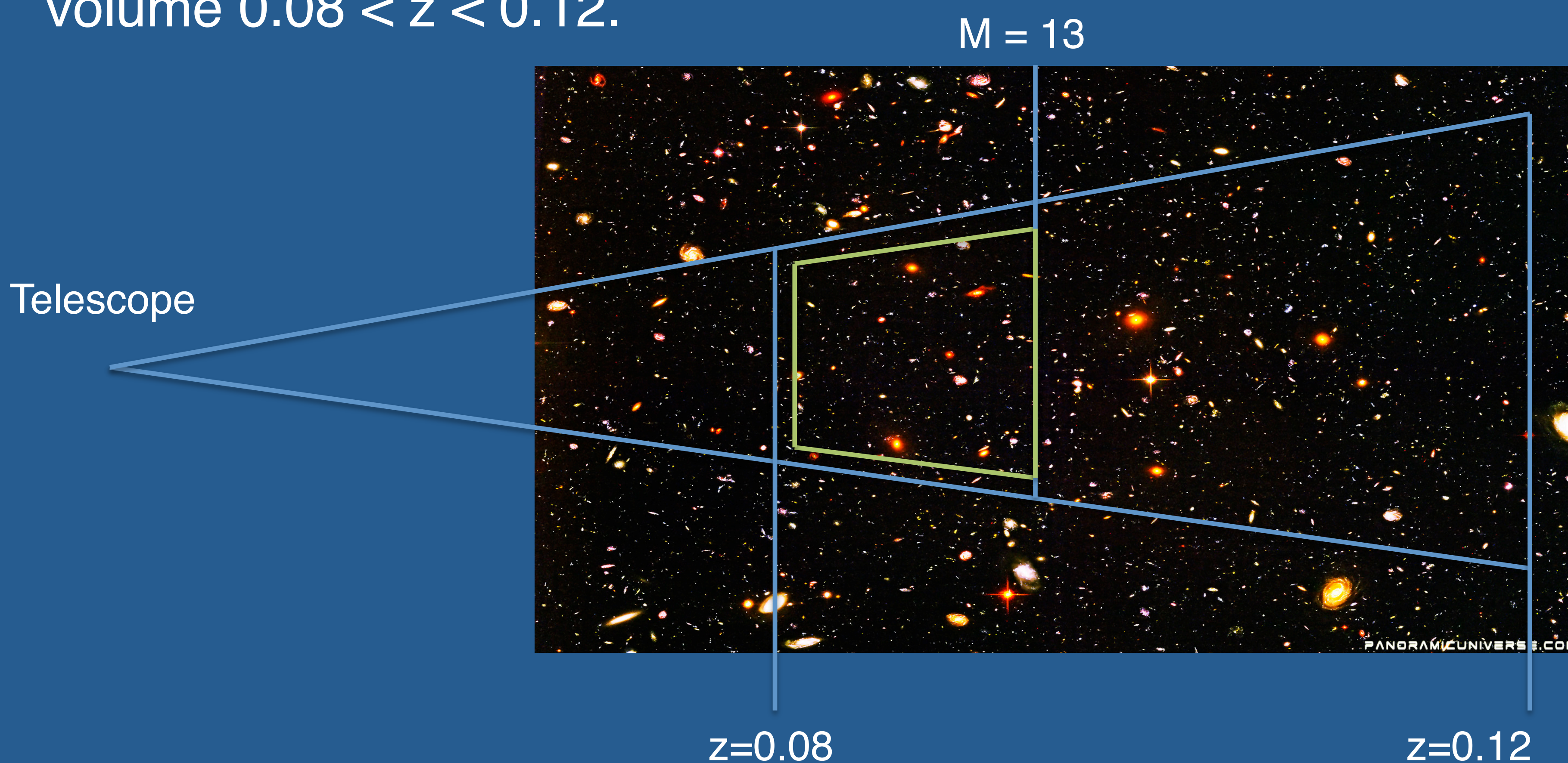
Say we care to probe galaxies with -11 < M < -21 in the volume 0.08 < z < 0.12.



# 3. Selection Effects

Example: The Luminosity Function

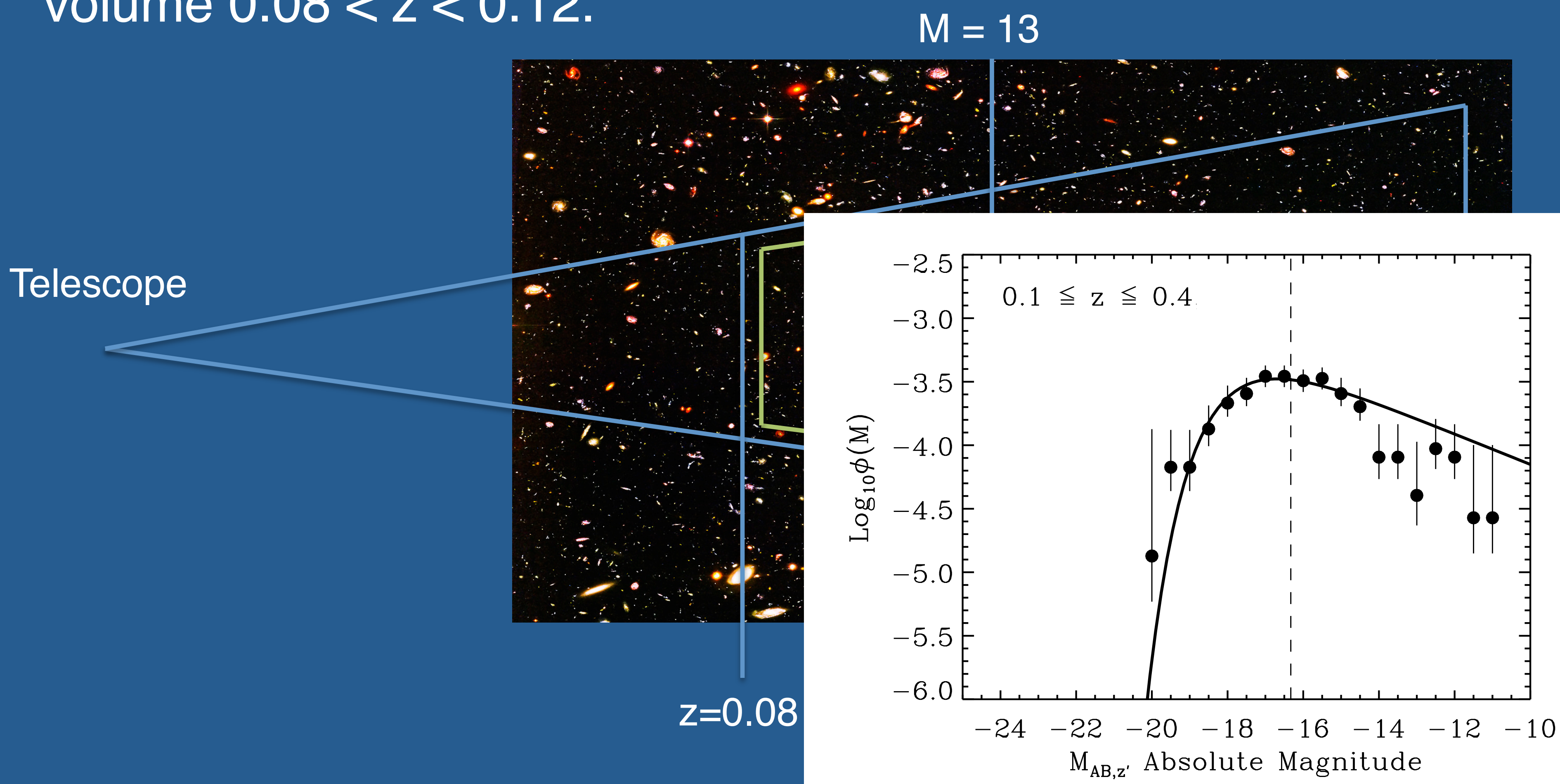
Say we care to probe galaxies with $-11 < M < -21$ in the volume $0.08 < z < 0.12$.



# 3. Selection Effects

Example: The Luminosity Function

Say we care to probe galaxies with -11 < M < -21 in the volume 0.08 < z < 0.12.



# 3. Selection Effects

Example: The Luminosity Function

$1/V_{max}$ Method:

Advantage:

Conceptually simple and easy to implement.

Disadvantage:

Assumes the Luminosity function is constant with increasing distance.

Alternative:

Lynden-Bell C- method

- Does not assume LF is constant
- Does not make sense to me, but hey, it’s in chapter 4.9

# 4. Histogram Bins and Errors

When making a histogram:

- How big do you make your bins?

- How do you calculate the error on that bin?

# 4. Histogram Bins and Errors

When making a histogram:

- How big do you make your bins?
  - Scott: $\Delta_b = 3.5\sigma / N^{1/3}$
  - Freedman-Diaconis: $\Delta_b = 2\ (q_{75}\text{-}q_{25}) / N1/3 = 2.7\sigma_G / N^{1/3}$

- How do you calculate the error on that bin?
  - If the number of data point is a bin is large,
    - $\sigma = \sqrt{N}$ (Gaussian)
  - If the number is data points per bin is below ~15
    - Have to use a Poisson formalism.

# 5. Helpful Python Tips

Docstrings are descriptive text for a function.

To read descriptions use the help() function.
For Ipython: '?' after any variable, function, class, or module

Example:
>>> import numpy as np
>>> help(np.argmax) ---------------------→

Written by ''' text ''' after a function definition

```
Help on function argmax in module numpy.core.fromnumeric:

argmax(a, axis=None)
    Indices of the maximum values along an axis.

    Parameters
    ----------
    a : array_like
        Input array.
    axis : int, optional
        By default, the index is into the flattened array, otherwise
        along the specified axis.

    Returns
    -------
    index_array : ndarray of ints
        Array of indices into the array. It has the same shape as `a.shape`
        with the dimension along `axis` removed.

    See Also
    --------
    ndarray.argmax, argmin
    amax : The maximum value along a given axis.
    unravel_index : Convert a flat index into an index tuple.

    Notes
    -----
    In case of multiple occurrences of the maximum values, the indices
    corresponding to the first occurrence are returned.

    Examples
    --------
    >>> a = np.arange(6).reshape(2,3)
    >>> a
    array([[0, 1, 2],
           [3, 4, 5]])
    >>> np.argmax(a)
    5
    >>> np.argmax(a, axis=0)
    array([1, 1, 1])
    >>> np.argmax(a, axis=1)
    array([2, 2])

    >>> b = np.arange(6)
    >>> b[1] = 5
    >>> b
    array([0, 5, 2, 3, 4, 5])
    >>> np.argmax(b) # Only the first occurrence is returned.
    1
```

# Activities

In working directory.

> git init

> git pull git@github.com:brittlundgren/py-astro-stat.git master

Create a directory for each week so far.

Your activity will be located in the week 5 directory.

The data you need for the activity will be in the ‘data’ directory.