BIG DATA ASTRONOMY

Joel Primack, UCSC

Director, University of California
High-Performance AstroComputing Center
http://hipacc.ucsc.edu

BIG DATA ASTRONOMY

Joel Primack, University of California, Santa Cruz

1. The Double Dark Universe
2. Sloan Digital Sky Survey & HST
3. Large Synoptic Survey Telescope
4. Square Kilometer Array
5. Computational Cosmology

Note: 1000 Gb = Terabyte= Tb = 10^{12} bytes
1000 Tb = Petabyte = Pb = 10^{15} bytes
1000 Pb = Exabyte = Eb = 10^{18} bytes


Bruce Munro’s sea of 600,000 DVDs ≈ 500 Tb

Composition of the universe:

Atomic matter	4.5%
Cold Dark Matter	25%
Dark Energy	70%

DARK MATTER

+ DARK ENERGY =

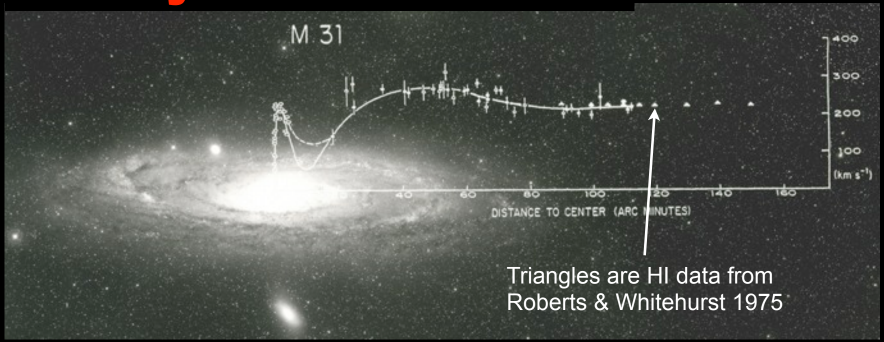
“DOUBLE DARK THEORY”

Technical Name:

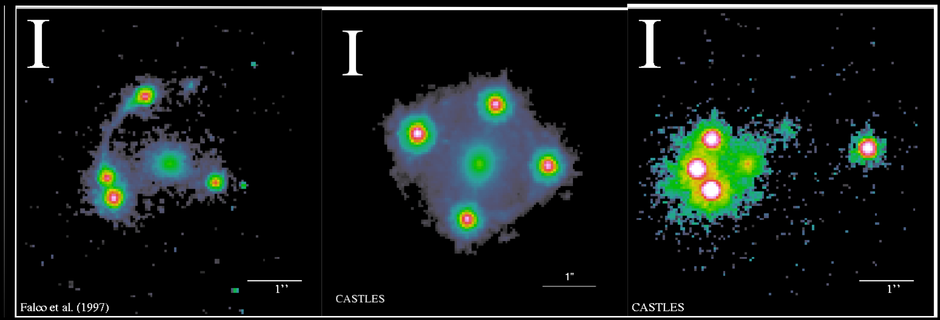
Lambda Cold Dark Matter (ΛCDM)

Examples of the Evidence for DARK MATTER

Galaxy Rotation Curves

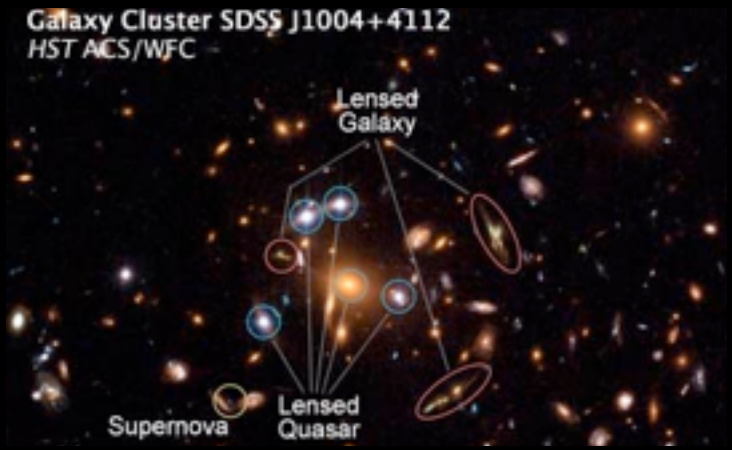


Galaxy Gravitational Lenses



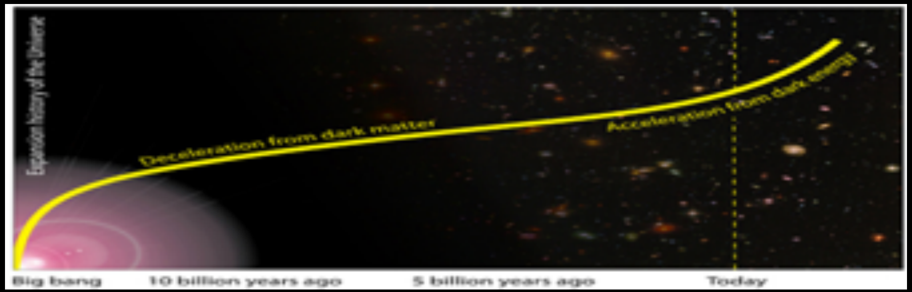
Bullet Cluster

Cluster Gravitational Lenses

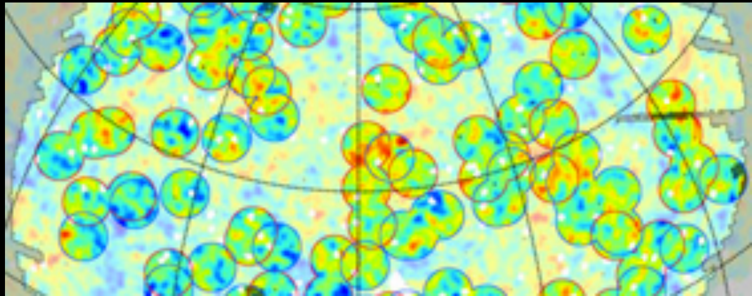


Examples of the Evidence for DARK ENERGY

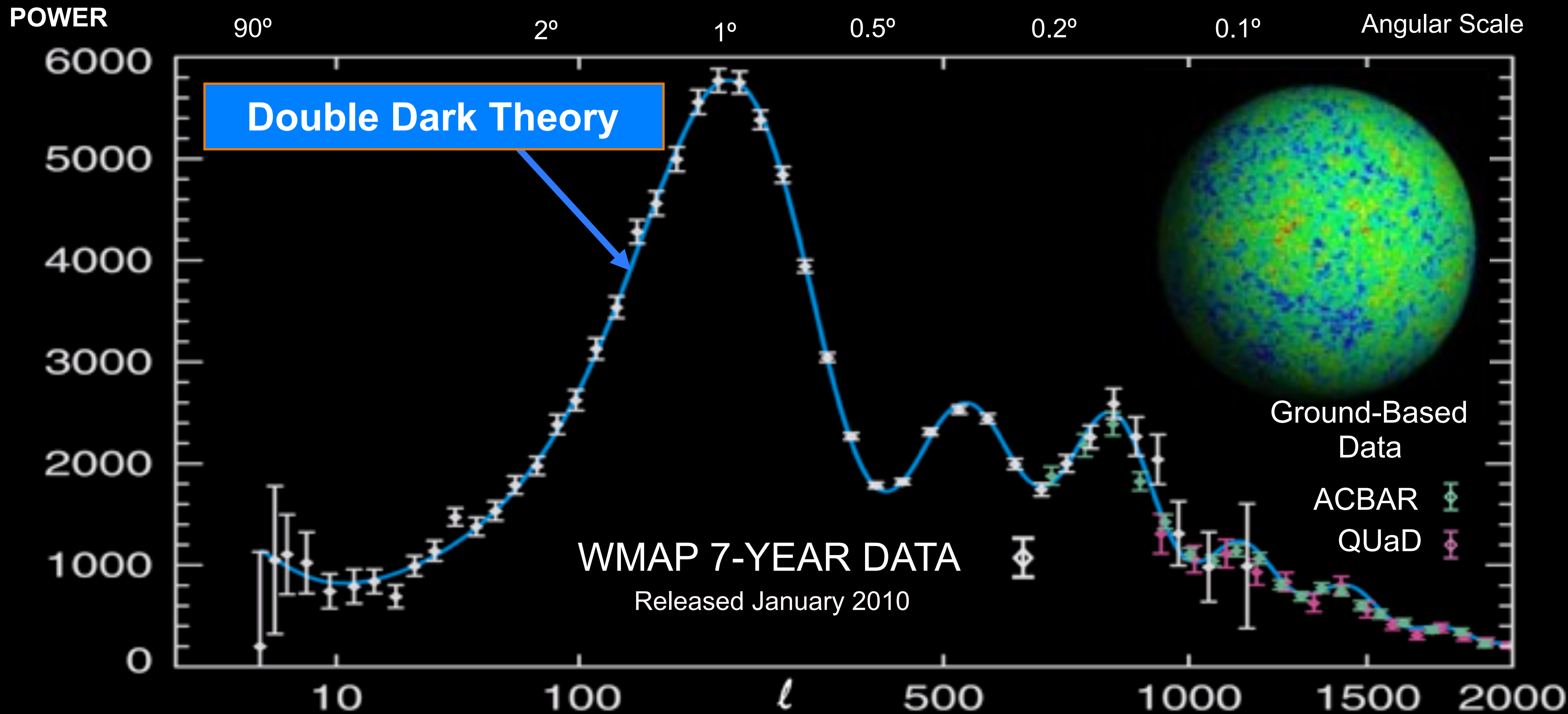
Expansion History of the Universe



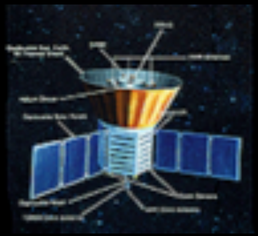
Supercluster and Void ISW

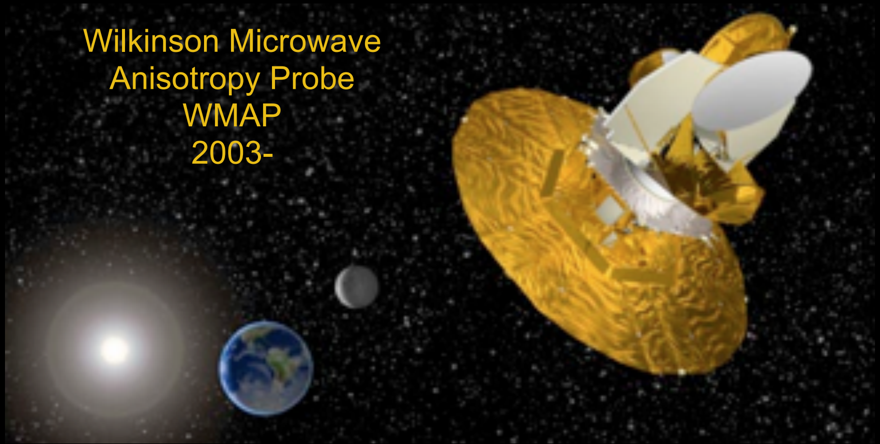


Big Bang Data Agrees with Double Dark Theory!

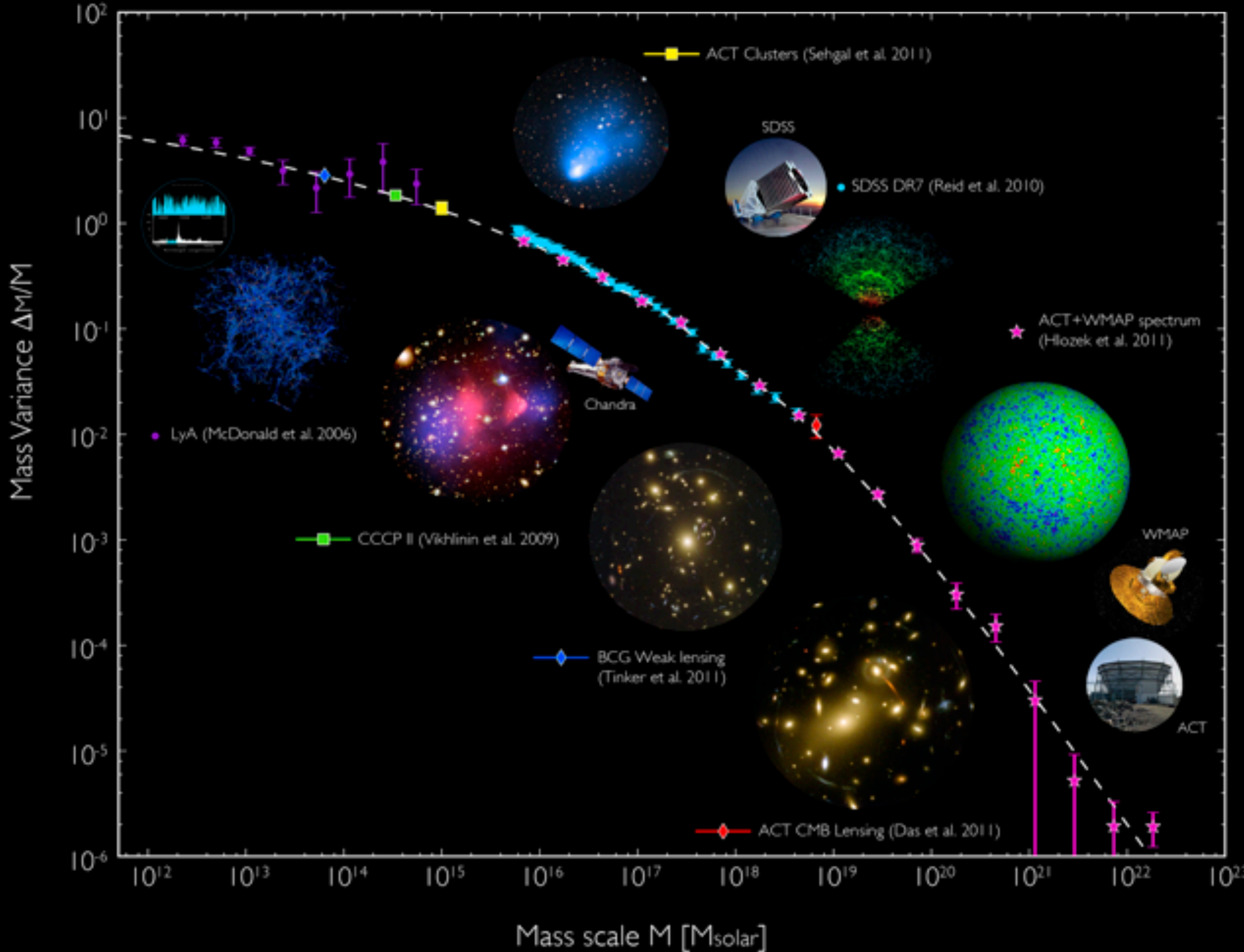


Cosmic Background Explorer COBE 1992

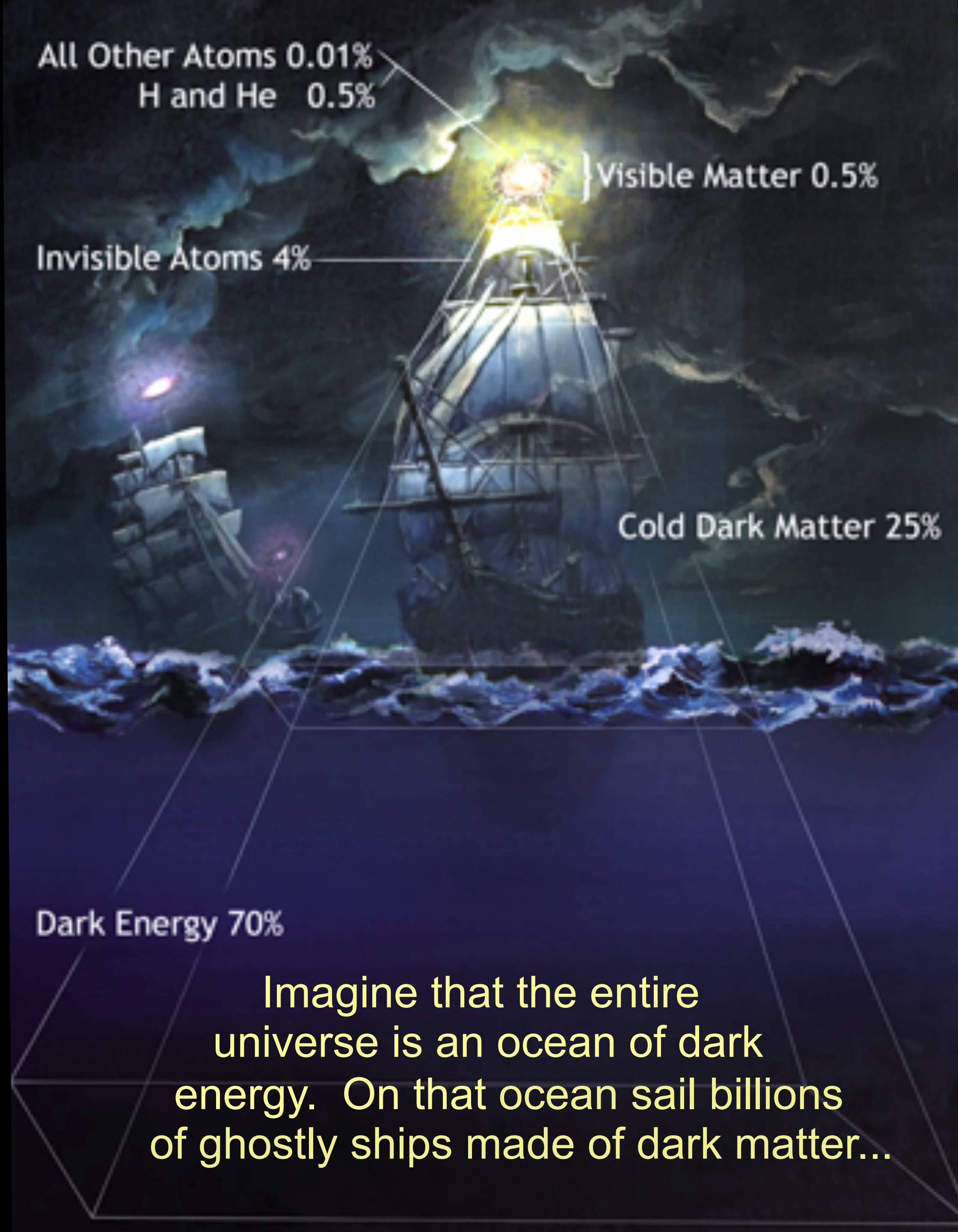




Matter Distribution Agrees with Double Dark Theory!



Matter and Energy Content of the Universe

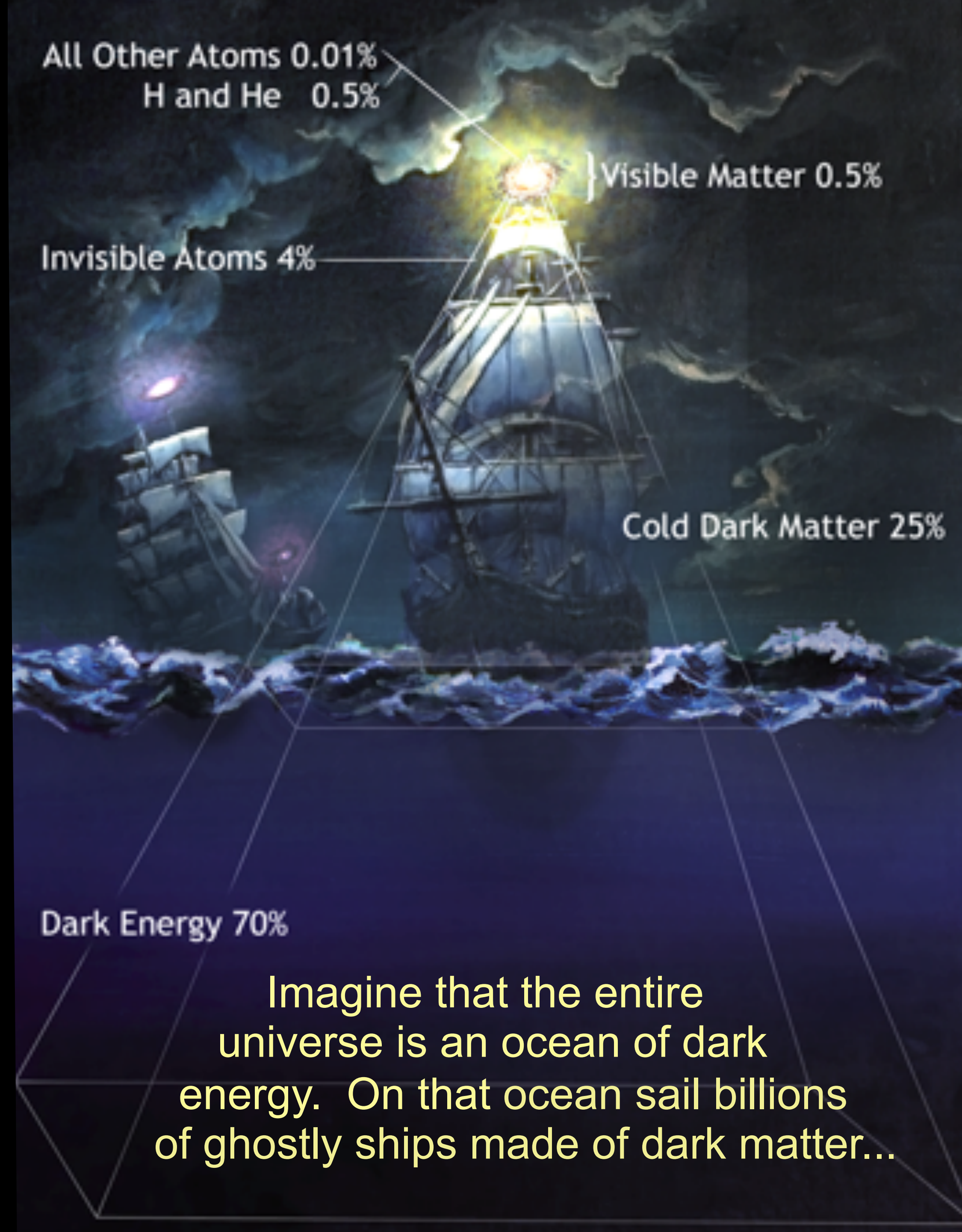


Imagine that the entire universe is an ocean of dark energy. On that ocean sail billions of ghostly ships made of dark matter...

Dark Matter Ships

on a

Dark Energy Ocean



Matter and Energy Content of the Universe

ΛCDM

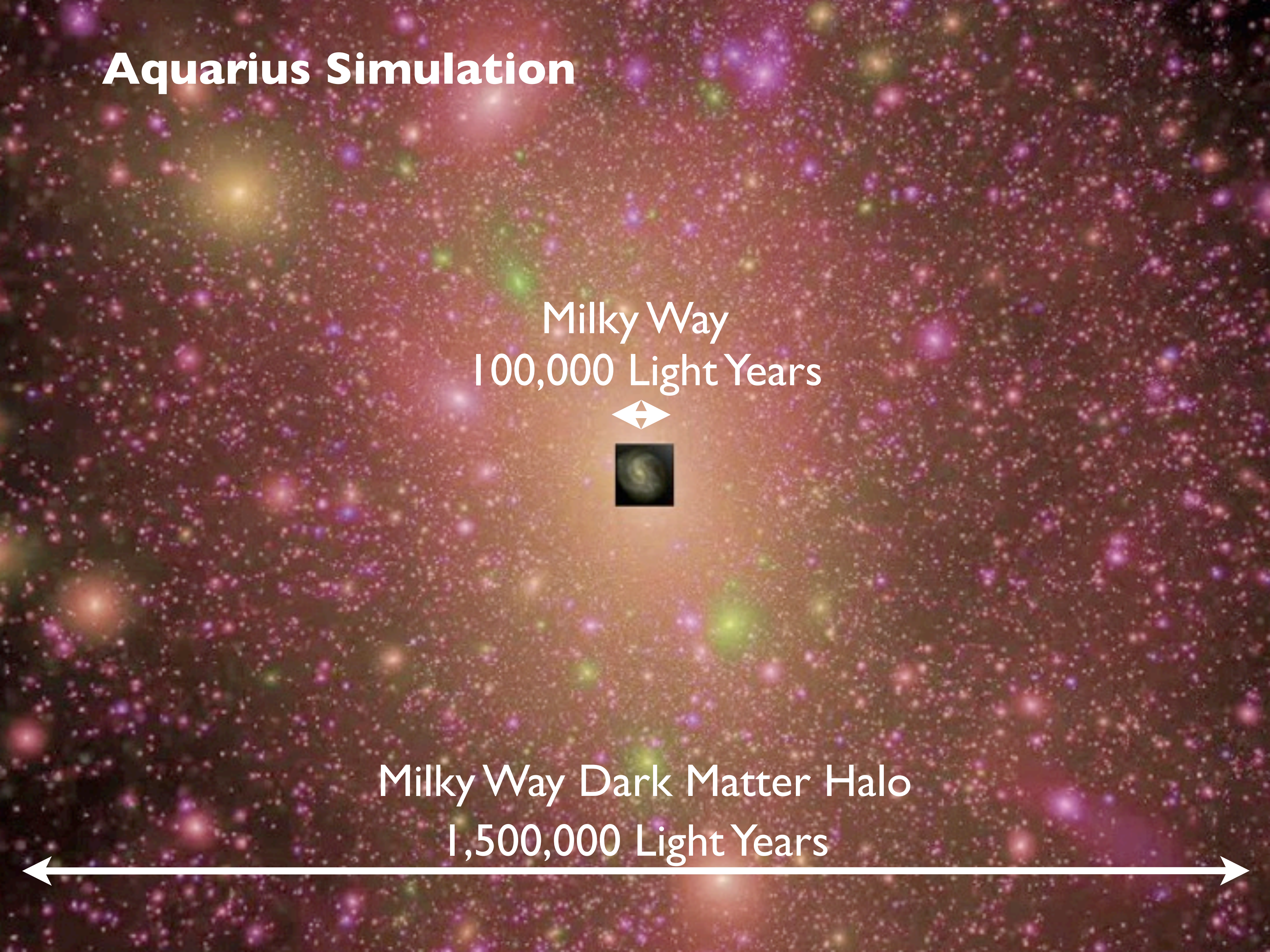
Double Dark Theory

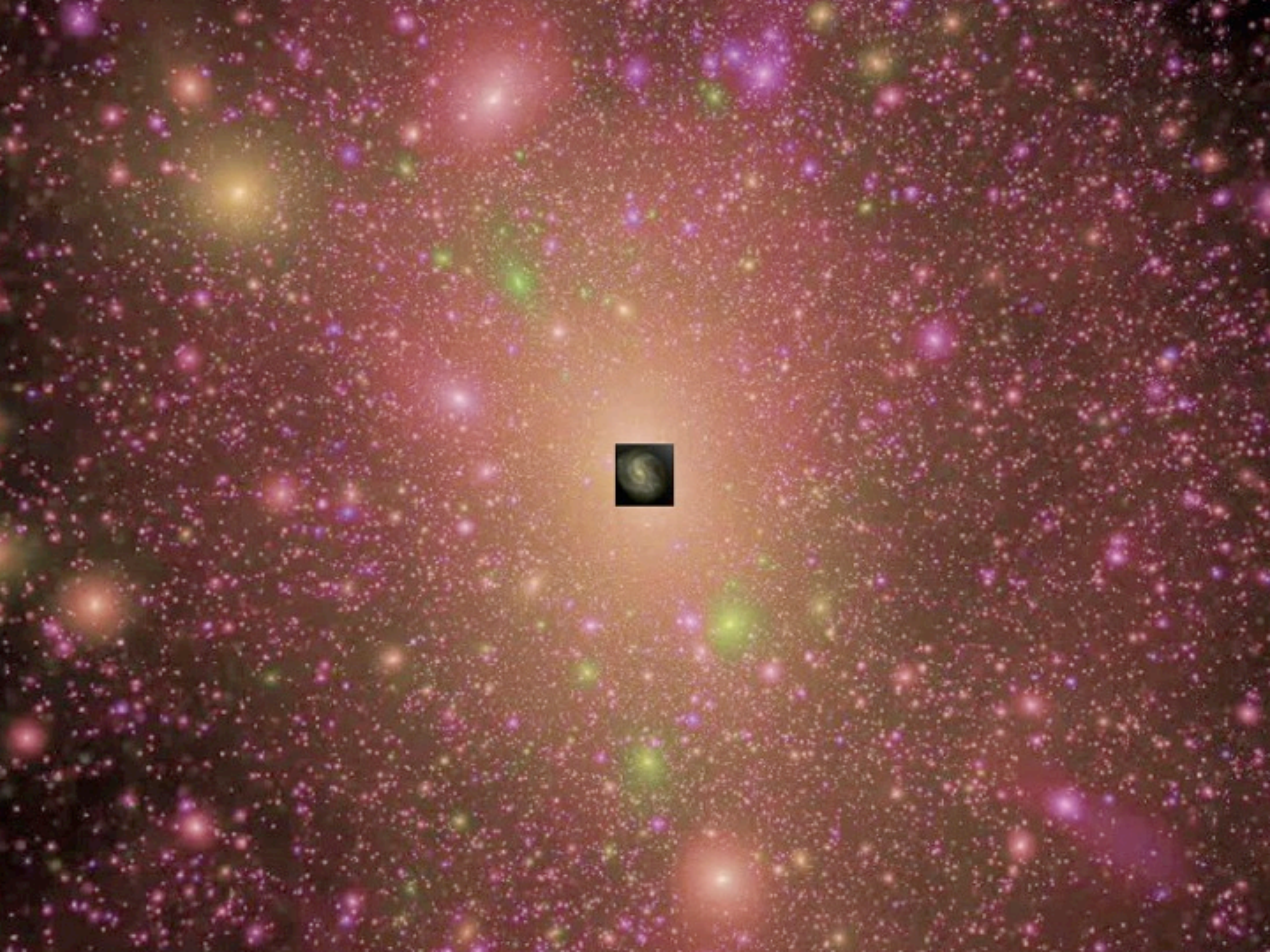
Cosmological Simulations

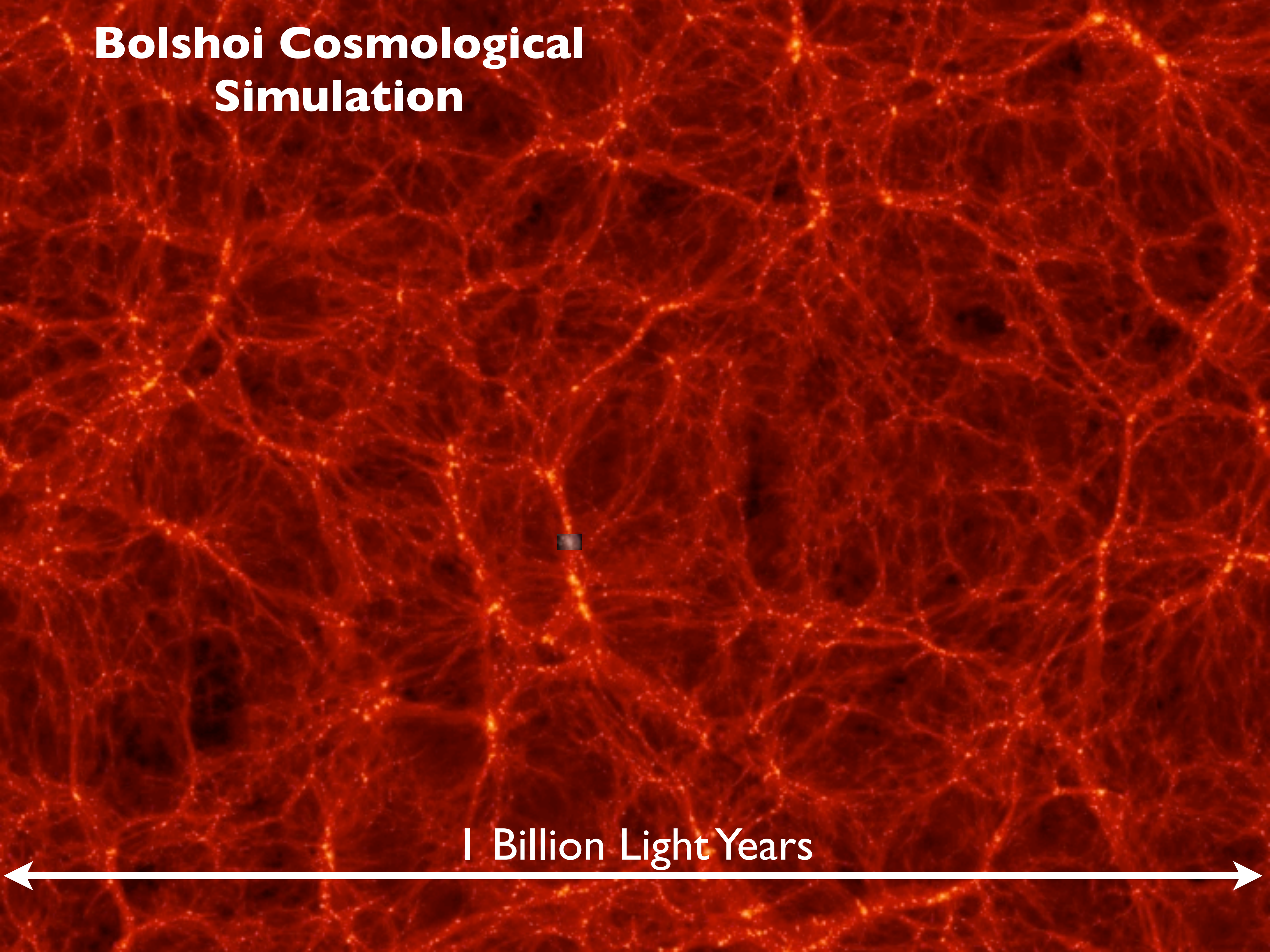
Astronomical observations represent snapshots of moments in time. It is the role of astrophysical theory to produce movies -- both metaphorical and actual -- that link these snapshots together into a coherent physical theory.

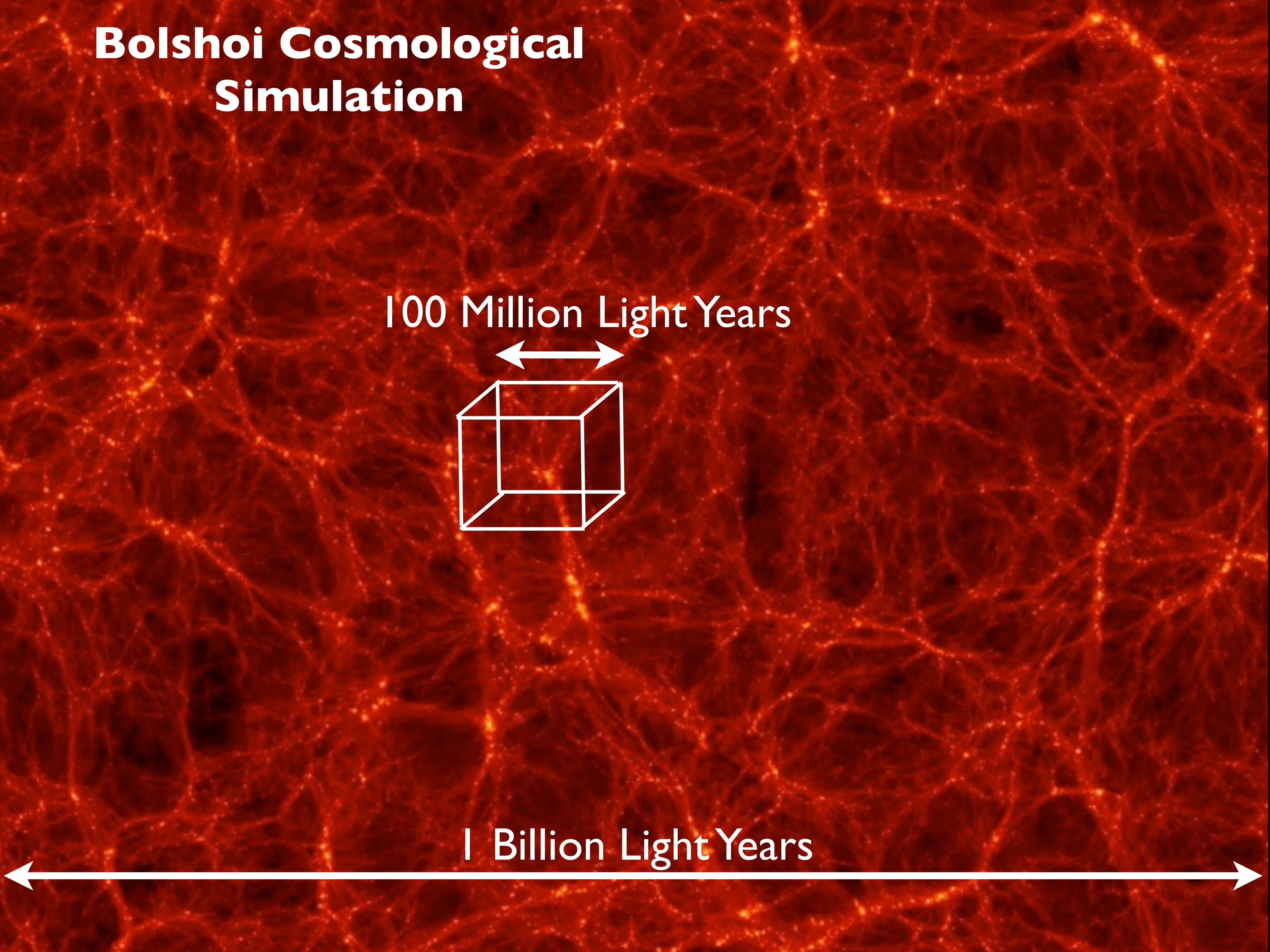
Cosmological dark matter simulations show large scale structure, growth of structure, and dark matter halo properties

Hydrodynamic galaxy formation simulations: evolution of galaxies, formation of galactic spheroids via mergers, galaxy images in all wavebands including stellar evolution and dust


Aquarius Simulation
Milky Way
100,000 Light Years
Milky Way Dark Matter Halo
1,500,000 Light Years




Bolshoi Cosmological Simulation
1 Billion Light Years


Bolshoi Cosmological Simulation
100 Million Light Years
1 Billion Light Years

Bolshoi Cosmological Simulation

100 Million Light Years

Bjork "Dark Matter"

Biophilia


SKY
& TELESCOPE
Dive Deep In
the Lagoon p. 61
JULY 2012
Universe in
a Box
From the Big Bang to Now p. 26

Universe on Fast Forward

STEFAN GOTTLÖBER / LEIBNIZ-INSTITUT FÜR ASTROPHYSIK POTSDAM

Supercomputer modeling is transforming cosmology from a purely observational science into an experimental science.

JOEL R. PRIMACK
& TRUDY E. BELL

https://dl.dropbox.com/u/5495083/Sky%26Telescope%20Bolshoi%20Article.pdf

EVOLVING UNIVERSE

Facing page, left to right: These frames from the Bolshoi simulation depict the universe at redshifts of 10, 3, 1, and 0, which corres-pond to cosmic ages of 490 million years, 2.2 billion years, 6 billion years, and 13.7 billion years (today). The bright areas have high densities of dark matter. As the far left frame shows, Bolshoi starts off with only a modest degree of lumpiness in the distribution of matter. But the subsequent frames demonstrate how gravity, acting over billions of years, gathered matter into long filaments that surround immense voids. Galaxies are concentrated along the filaments, clusters at the nodes.

Bolshoi Merger Tree for the Formation of a Big Cluster Halo

Peter Behroozi

The Bolshoi simulation

ART code

250Mpc/h Box
LCDM

σ_8 = 0.82
h = 0.70

8G particles
1kpc/h force resolution
1e8 Msun/h mass res

dynamical range 262,000
time-steps = 400,000

NASA AMES
supercomputing center
Pleiades computer
13824 cores
12TB RAM
75TB disk storage
6M cpu hrs
18 days wall-clock time

Cosmological parameters are consistent with the latest observations

Force and Mass Resolution are nearly an order of magnitude better than Millennium-I

Force resolution is the same as Millennium-II, in a volume 16x larger

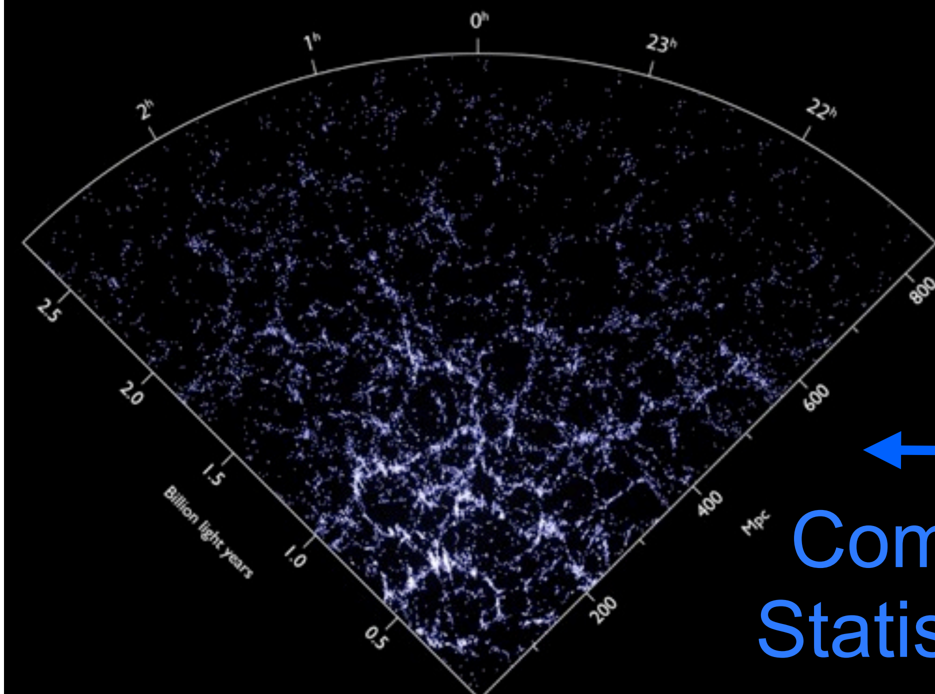
Halo finding is complete to V_{circ} > 50 km/s, using both BDM and ROCKSTAR halo finders

Bolshoi and MultiDark halo catalogs were released in September 2011 at Astro Inst Potsdam; Merger Trees are now available

Observational Data

Sloan Digital Sky Survey

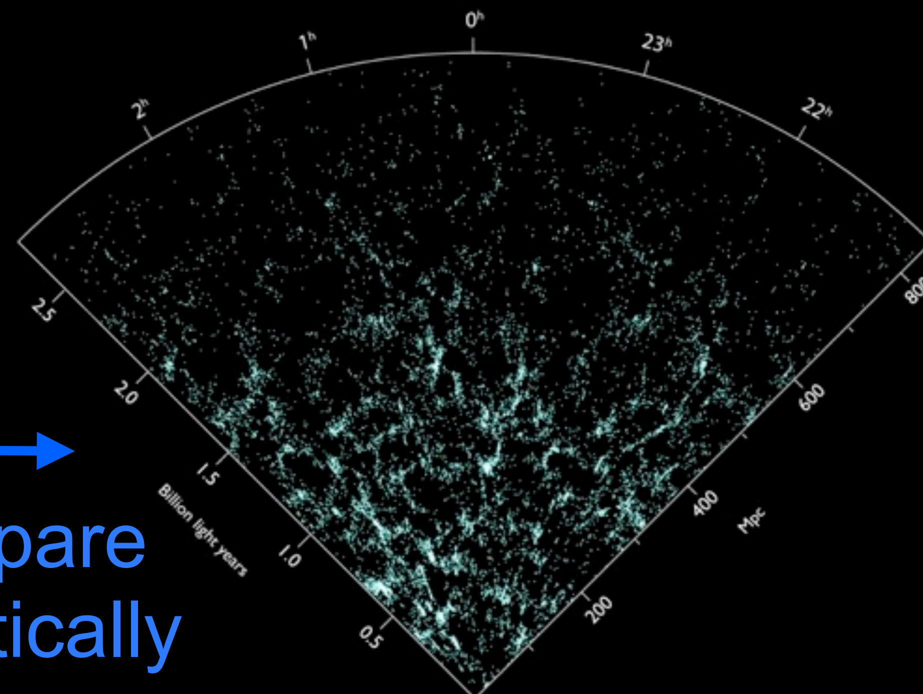
SDSS



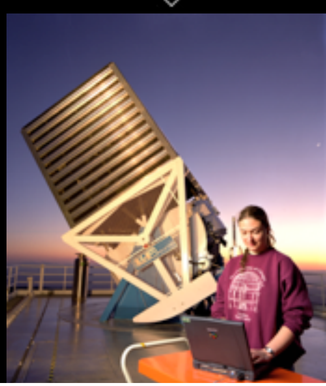
Cosmological Simulation

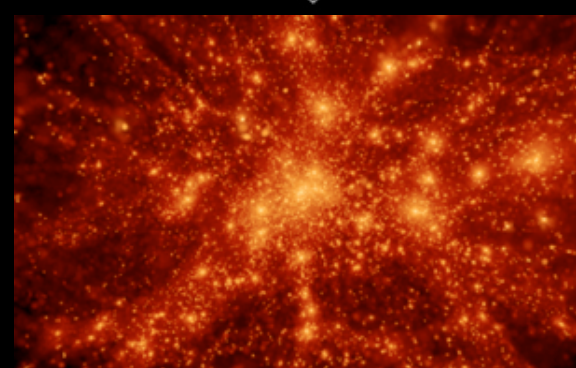
Risa Wechsler, Ralf Kahler, Nina McCurdy

Bolshoi

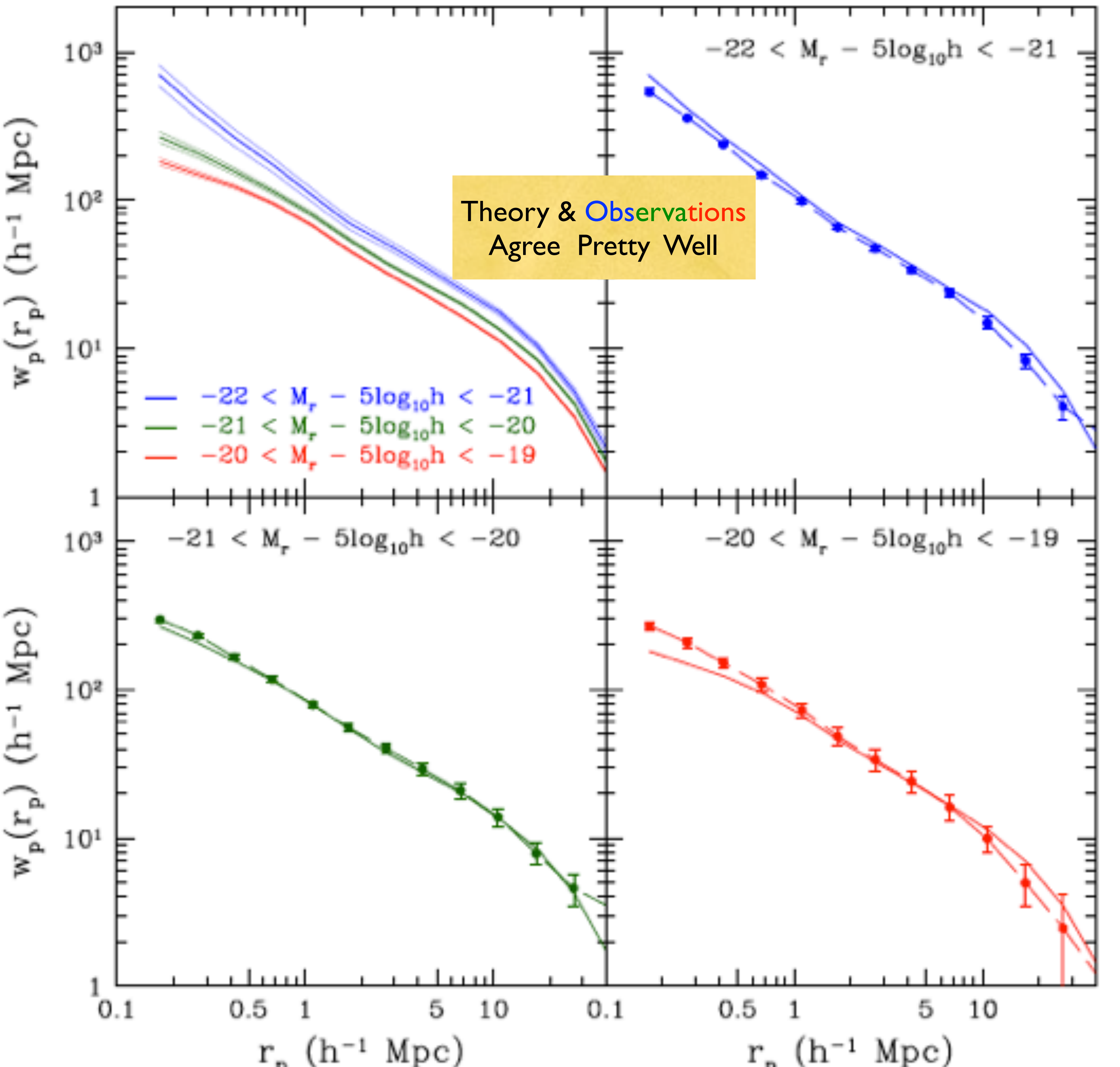


Compare Statistically





Bolshoi Projected Galaxy Correlation Functions

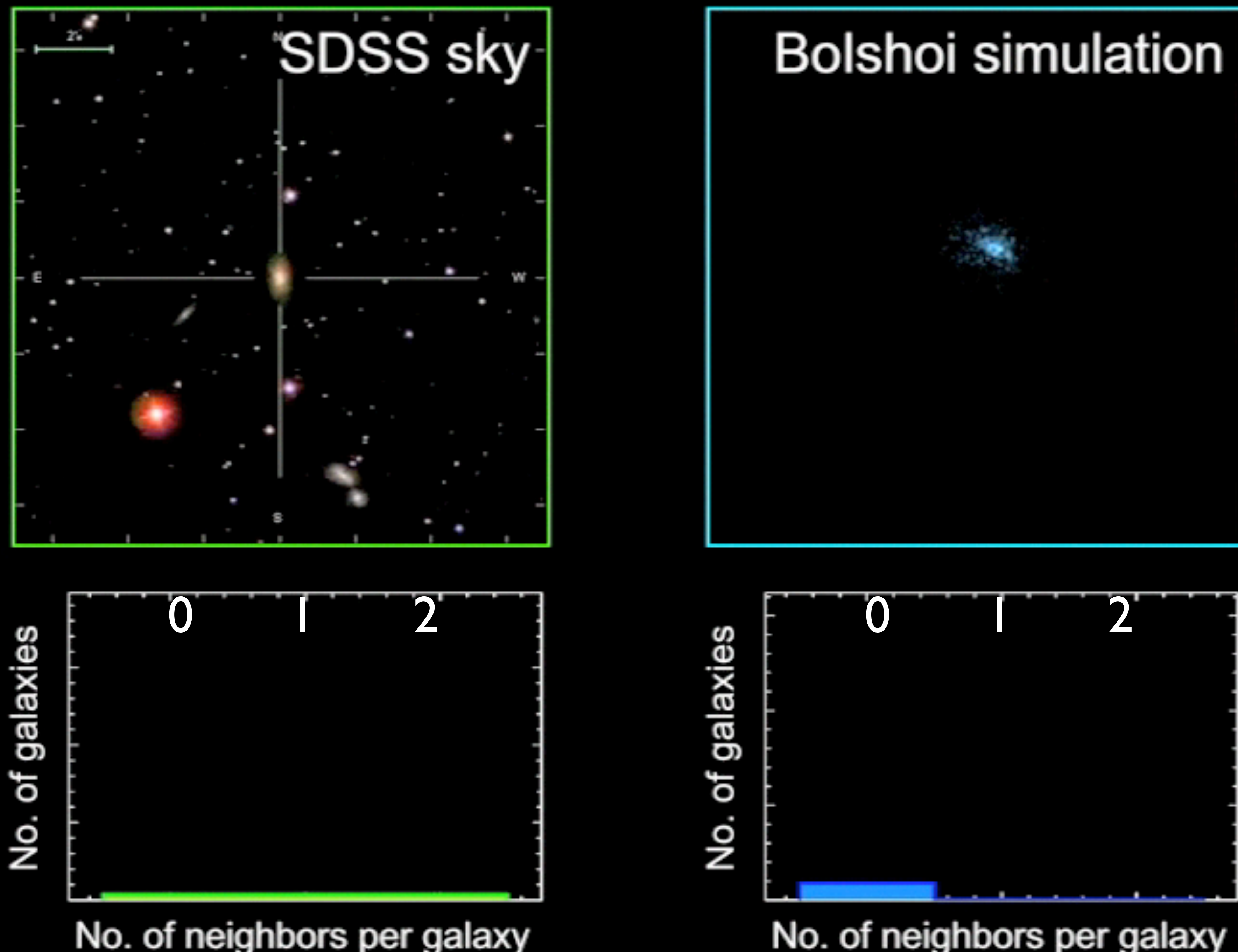


The correlation function of SDSS galaxies vs. Bolshoi galaxies using halo ~~abundance matching,~~ with scatter using our stochastic abundance matching method. This results in a better than 20% agreement with SDSS. *Top left:* correlation functinon in three magnitude bins, showing Poisson uncertainties as thin lines. *Remaining panels:* correlation function in each luminosity bin compared with SDSS galaxies (points with error bars: Zehavi et al. 2010).

Trujillo-Gomez, Klypin, Primack, & Romanowsky 2011 ApJ

The Milky Way has two large satellite galaxies, the small and large Magellanic Clouds

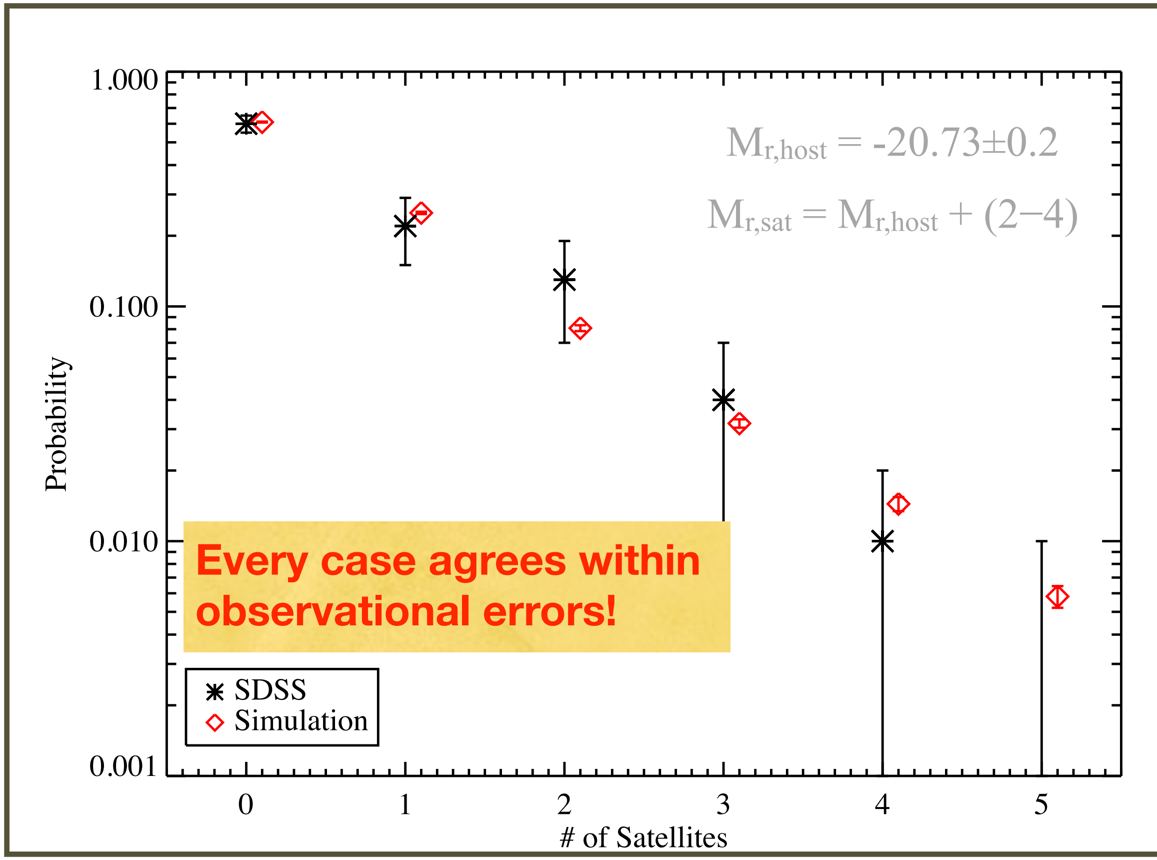
The Bolshoi simulation + halo abundance matching predicts the likelihood of this


SDSS sky
Bolshoi simulation
0 1 2
No. of galaxies
No. of neighbors per galaxy
0 1 2
No. of galaxies
No. of neighbors per galaxy

- Apply the same absolute magnitude and isolation cuts to Bolshoi+SHAM galaxies as to SDSS:
  - Identify all objects with absolute ${}^{0.1}M_r$ = -20.73±0.2 and observed m_r < 17.6
  - Probe out to z = 0.15, a volume of roughly 500 $(Mpc/h)^3$
  - leaves us with 3,200 objects.
- Comparison of Bolshoi with SDSS observations is in close agreement, well within observed statistical error bars.

# of Subs	Prob (obs)	Prob (sim)
0	60%	61%
1	22%	25%
2	13%	8.1%
3	4%	3.2%
4	1%	1.4%
5	0%	0.58%

Statistics of MW bright satellites: SDSS data vs. Bolshoi simulation



Busha et al. 2011 ApJ
Liu et al. 2011 ApJ

Risa Wechsler

Similarly good agreement with SDSS for brighter satellites with spectroscopic redshifts compared with Millennium-II using abundance matching -- Tollerud, Boylan-Kolchin, et al. 2011 ApJ

Astronomical data has several advantages:

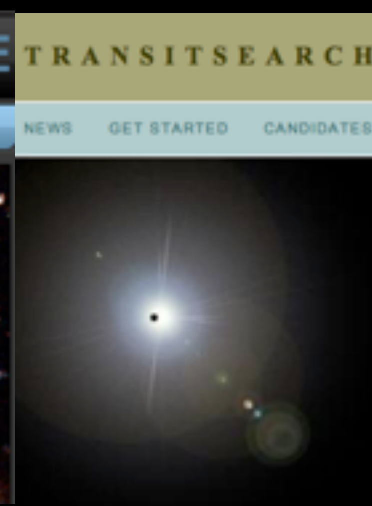
The data tends to be pretty **clean**

The data is (mostly) **non-proprietary**

The research is (mostly) **funded**

The data is **big** and **sexy**

and there's a lot of **public involvement:**



Big Challenges of AstroComputing

Big Data

Sloan Digital Sky Survey (SDSS) 2008

2.5 Terapixels of images
40 Tb raw data ➠120 Tb processed
35 Tb catalogs

Mikulski Archive for Space Telescopes (MAST)

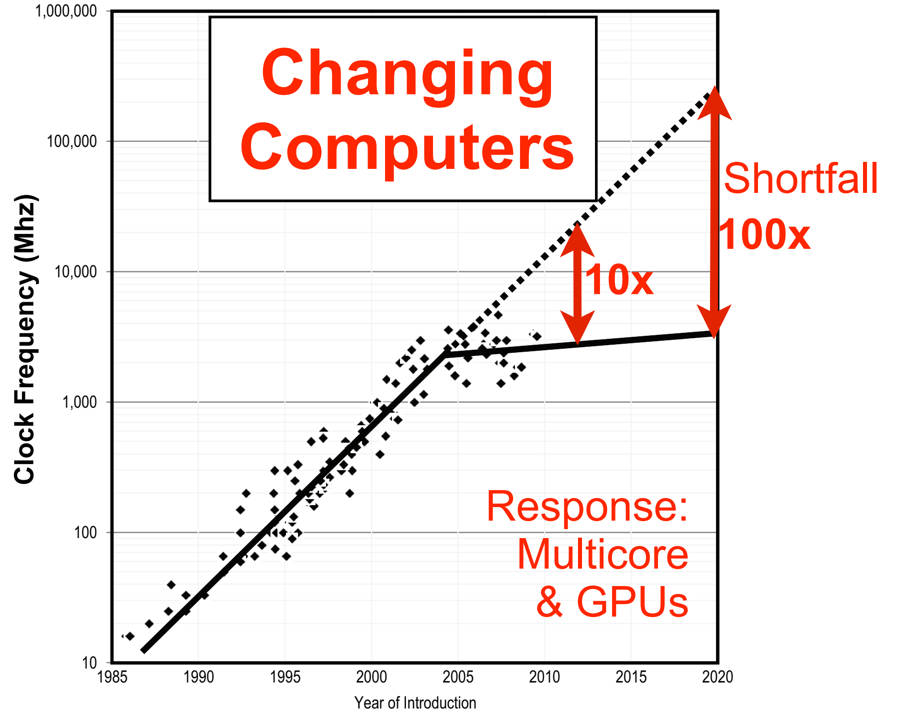
185 Tb of images
25 Tb/year ingest rate
>100 Tb/year retrieval rate

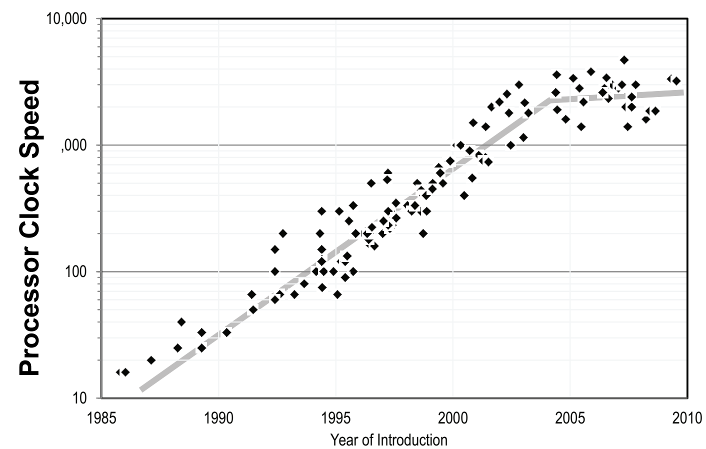
Large Synoptic Survey Telescope (LSST) 2014

15 Tb per night for 10 years
100 Pb image archive
20 Pb final database catalog

Square Kilometer Array (SKA) ~2024

1 Eb per day (> internet traffic today)
100 PFlop/s processing power
~1 Eb processed data/year



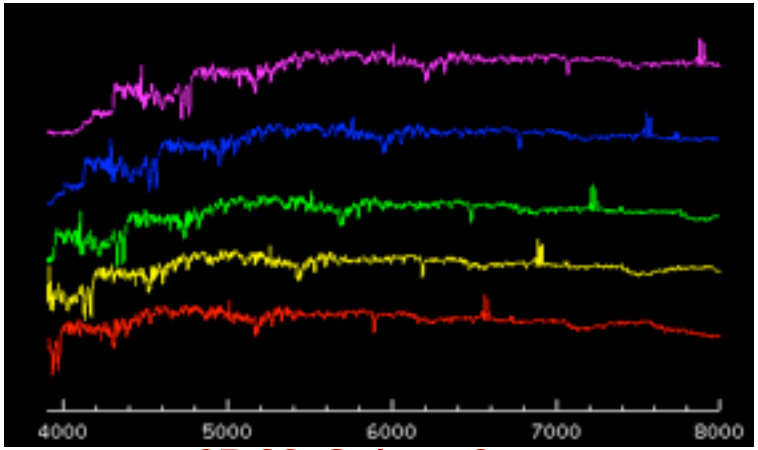


Sloan Digital Sky Survey (SDSS)

Sloan Digital Sky Survey 1992-2008
"The Cosmic Genome Project"

Imaging survey in 5 wavelength bands: 5-color images of ¼ of the sky

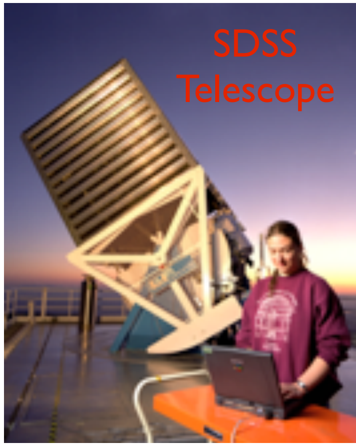
Spectroscopic redshift survey



SDSS Galaxy Spectra

Massive Data
- 2.5 Terapixels of images
- 40 Tb raw data ➠ 120 Tb processed
- 35 Tb catalogs

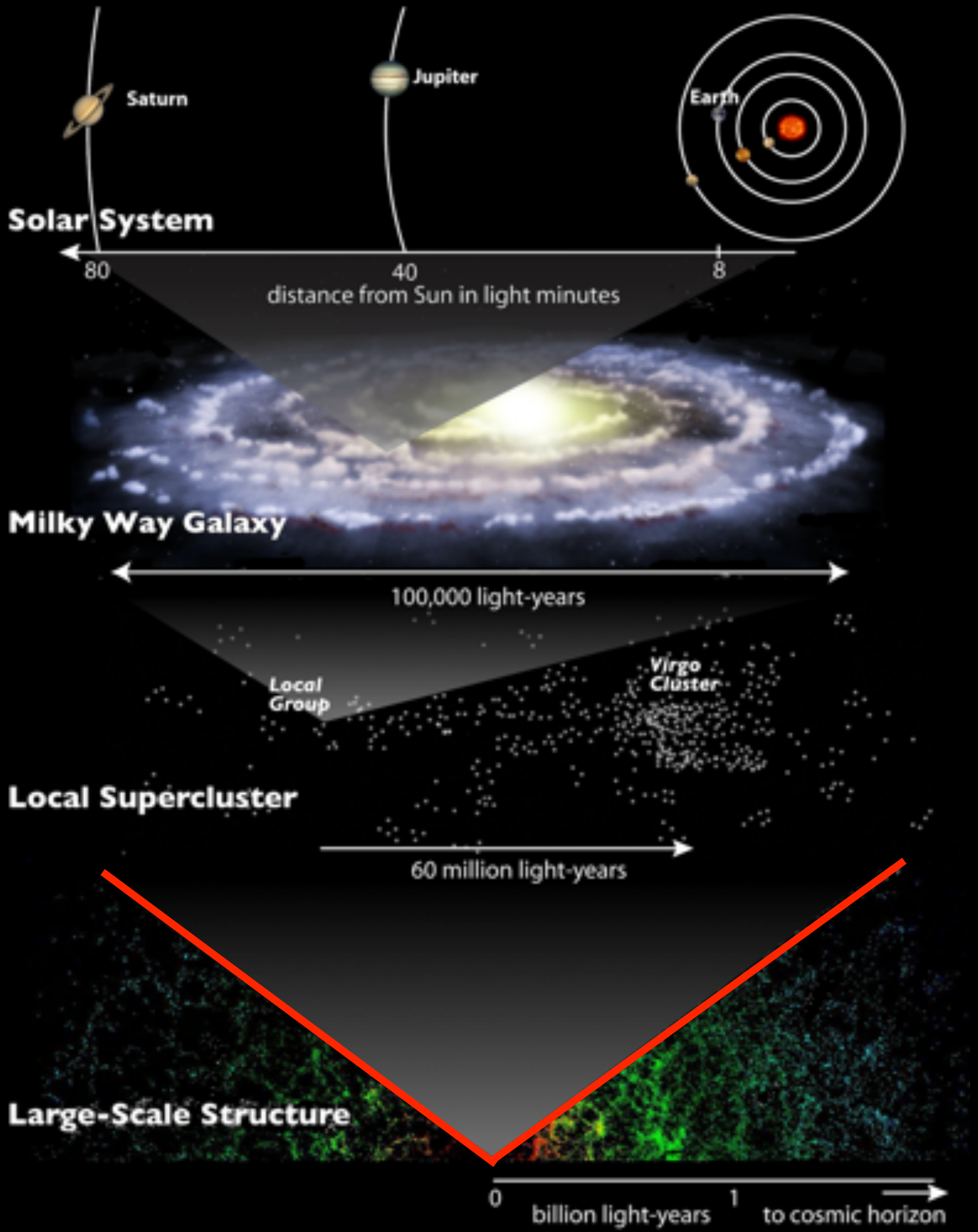
Data is publicly accessed
- 840 million web hits in 9 years, now >1 billion
- 4,000,000 distinct users* vs. 15,000 astronomers
- Basis for ~20,000 scientific papers
- More citations than any telescope including Hubble



* Having fun looking at data no one had ever seen before!

The Modern Scientific Cosmos

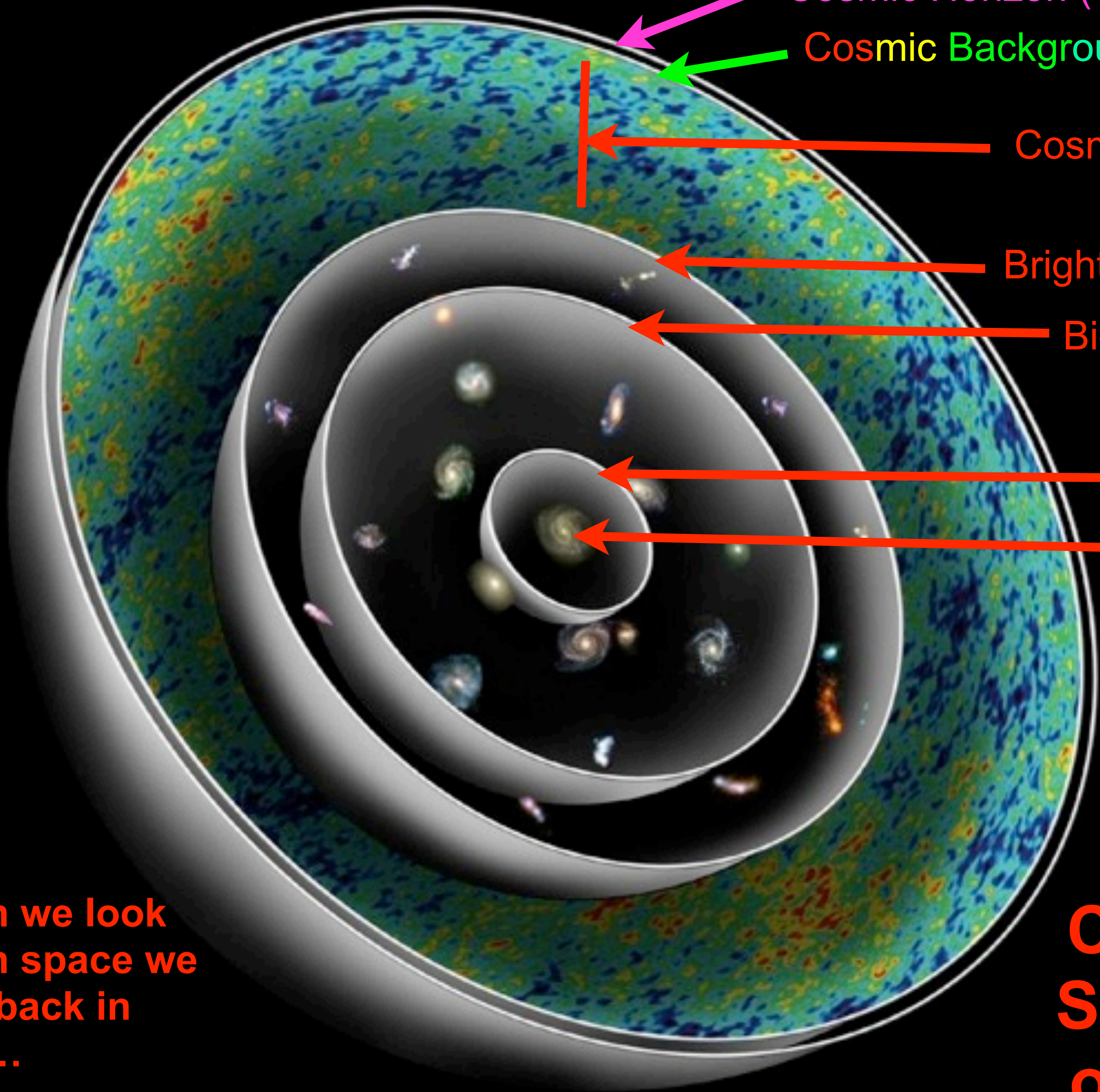
Our Cosmic Address



each dot is a big galaxy

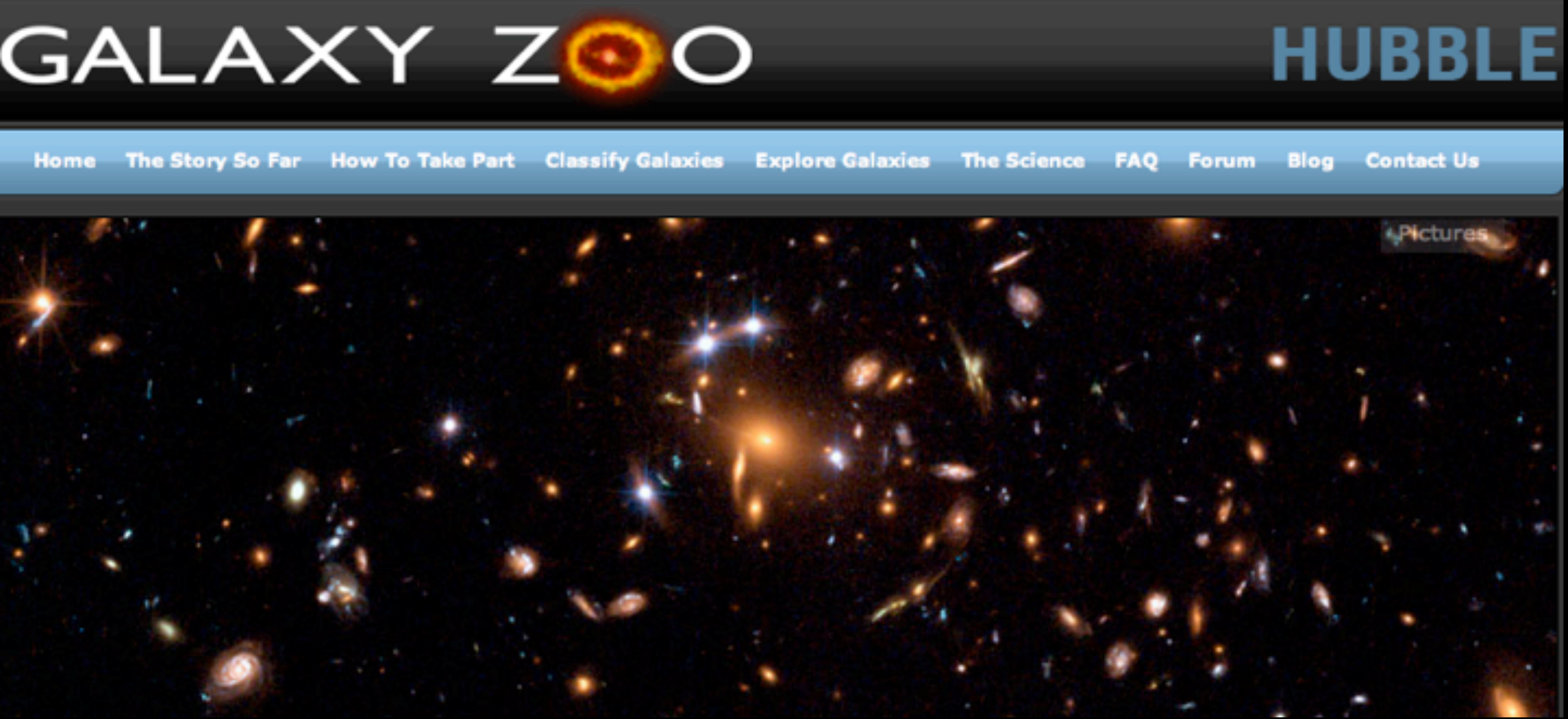
Sloan Digital Sky Survey

GALAXIES MAPPED BY THE SLOAN SURVEY



When we look out in space we look back in time…

Cosmic Spheres of Time



Galaxy Zoo started as an offshoot of the Sloan Digital Sky Survey

40 million visual classifications by the public

>250,000 people participating (blogs, poems, ...)

Amazing original discovery by a schoolteacher (Voorwerp)

Excellent coverage by CNN, BBC, NY Times, Washington Post

http://www.galaxyzoo.org

Mikulski Archive for Space Telescopes
(MAST)

What is the STScI archive?

Mikulski Archive for Space Telescopes: **MAST**

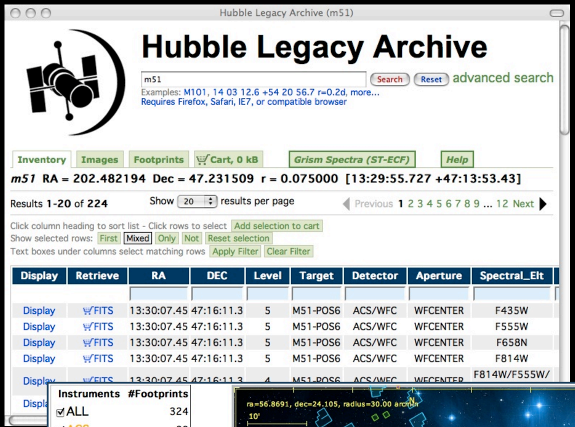
- Data

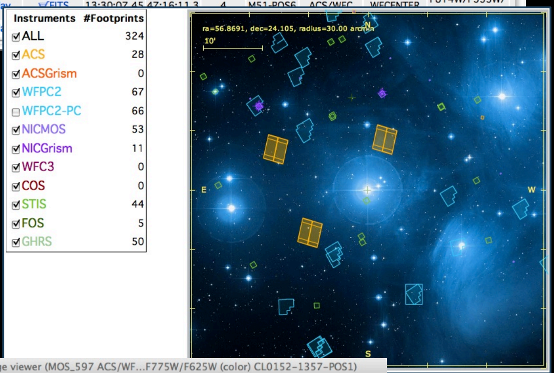
  ~185 TB of images, spectra, catalogs, time series

- Metadata

  ~10^6 HST observations (plus other missions)

  Documentation, publication links, …





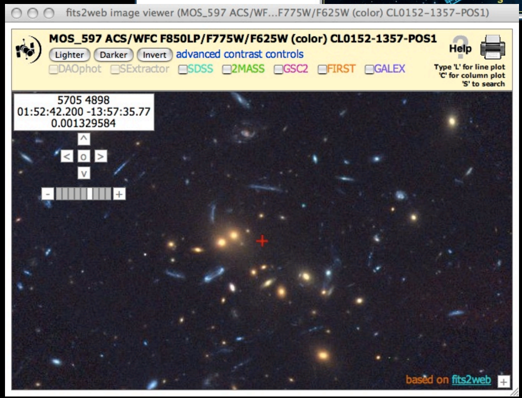
- User interfaces

  Search, browse, plot, explore

  Browser-based interfaces

  Help desk/User support


```
<VOTABLE>
<DESCRIPTION>STScI Hubble Legacy Archive SIAP</DESCRIPTION>
  <INFO name="QUERY_STATUS" value="OK"></INFO>
  <RESOURCE type="results">
    <PARAM datatype="char" name="INPUT:POS" value="210.802458,54
    <PARAM datatype="double" name="INPUT:SIZE" value="0.240000"
    <PARAM datatype="char" name="INPUT:FORMAT" value="FITS" arra
    <PARAM datatype="char" name="INPUT:imagetype" value="best" a
    <PARAM datatype="char" name="INPUT:inst" value="acs,wfpc2,ni
    <PARAM datatype="int" name="INPUT:hrcmatch" value="0"></PARA
    <PARAM datatype="double" name="INPUT:zoom" value="1.000000">
    <PARAM datatype="double" name="INPUT:autoscale" value="99.50
    <PARAM datatype="int" name="INPUT:asinh" value="1"></PARAM>
    <PARAM datatype="char" arraysize="*" name="refframe" ucd="VO
    <PARAM datatype="char" arraysize="*" name="projection" ucd="
    <TABLE>
```

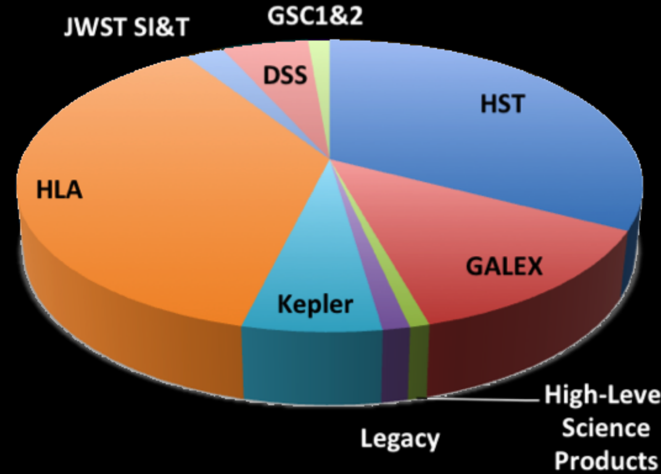
- Services

  VO services, data retrieval, image cutouts, …

  (UIs are built around VO services)

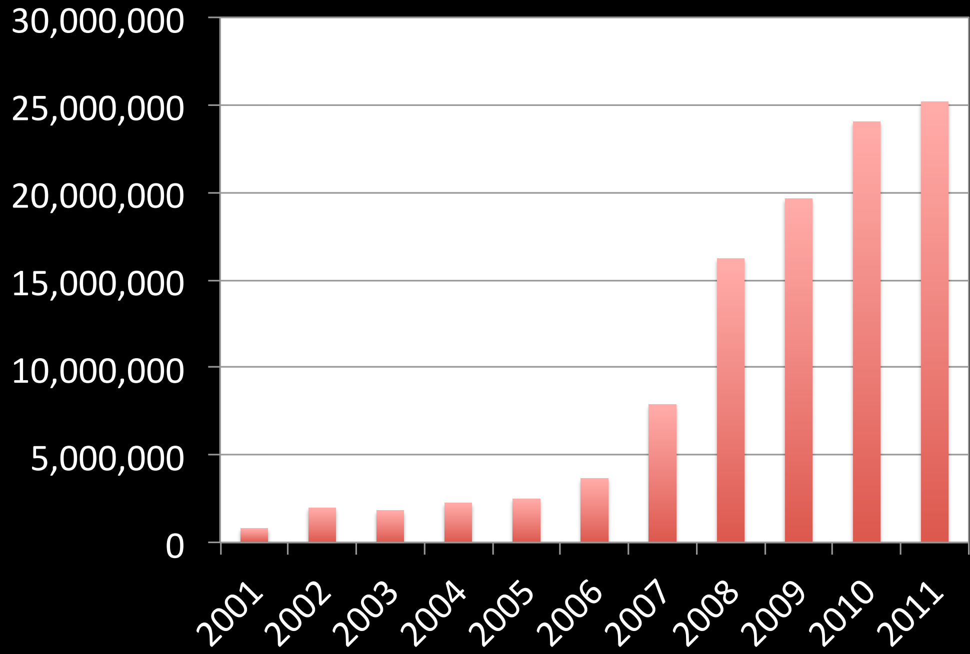
The MAST Archive: 2 minute summary

- ~ 185 TBytes (62 TB HST, 79 TB HLA)
- Ingest rate: > 25 TB/yr
- Retrievals: > 100 TB/yr
  - Distributed volume ~4x ingest

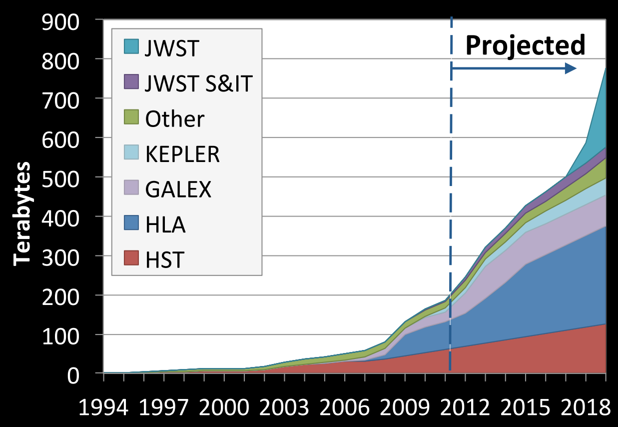


Data holdings

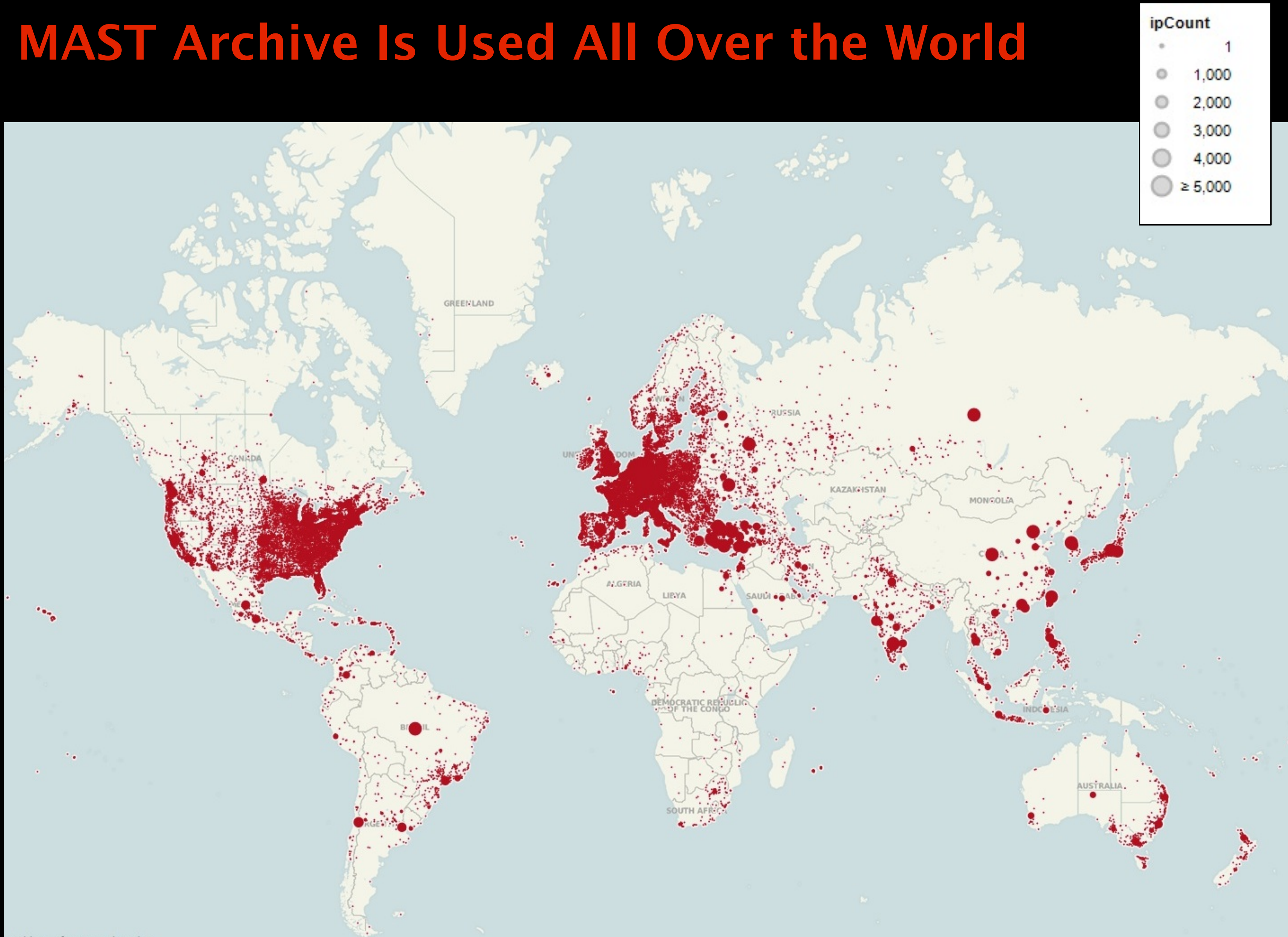
Number of searches per year



Past & projected volume



MAST Archive Is Used All Over the World

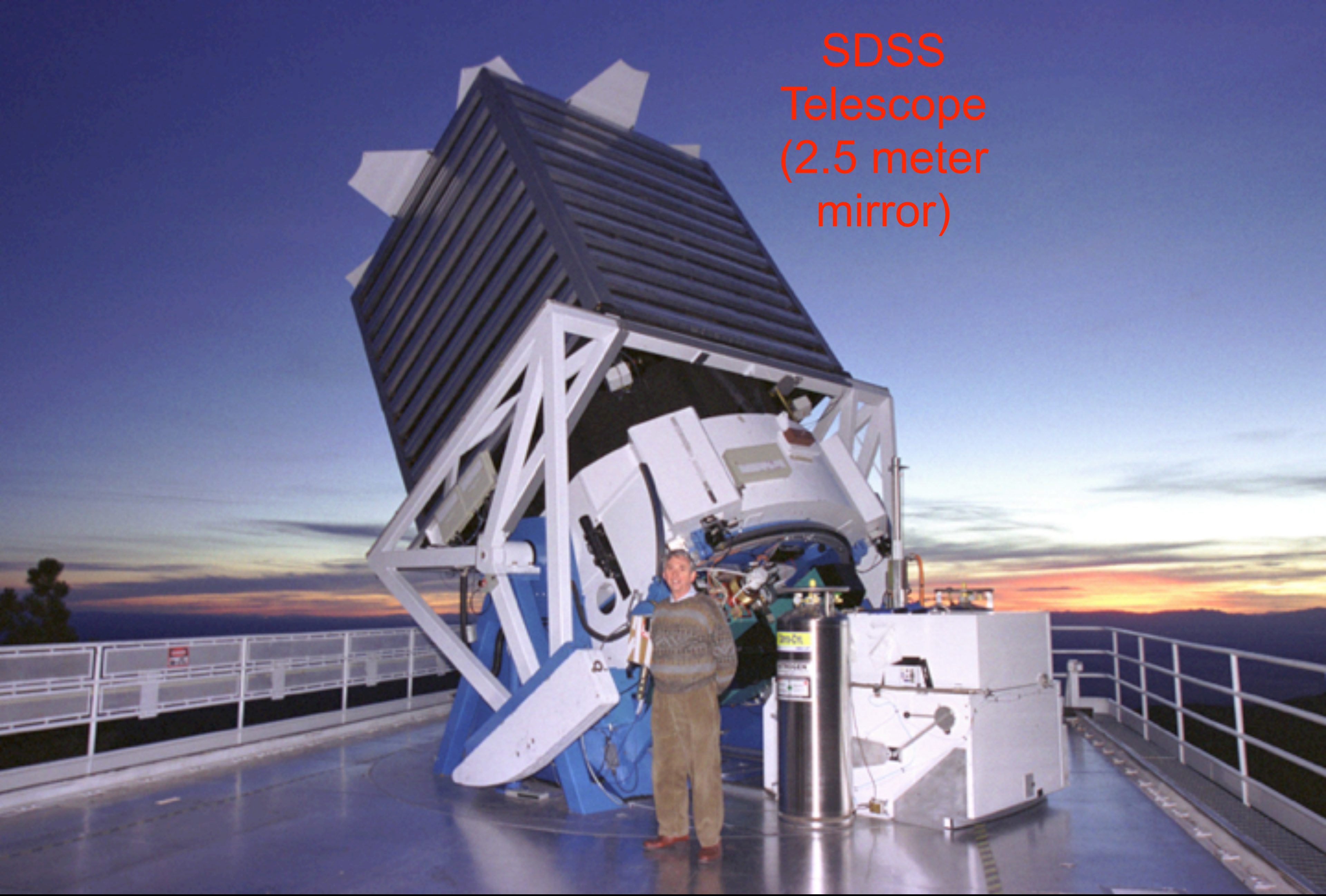


Virtual Observatory

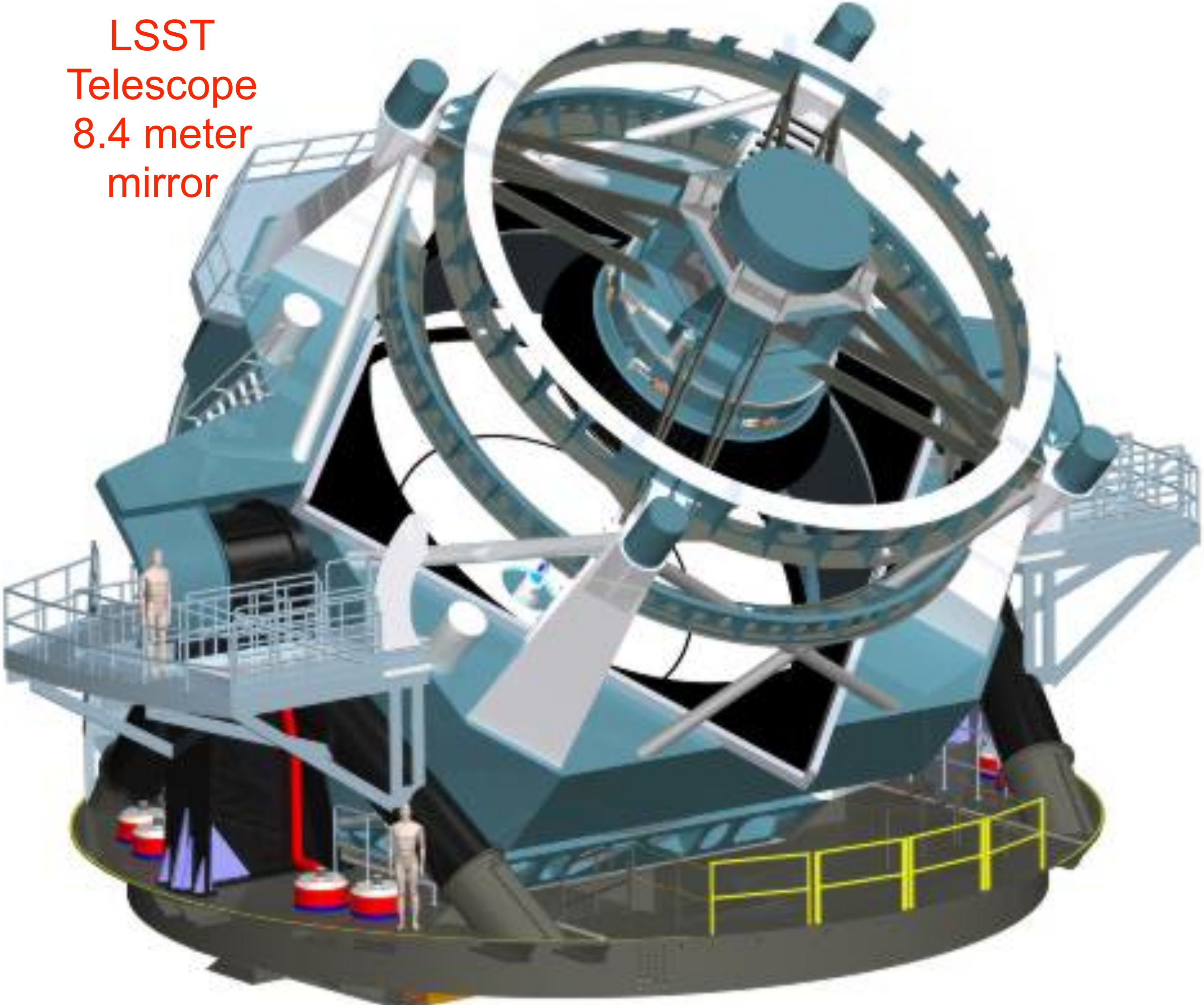
- Started with NSF ITR project, "Building the Framework for the National Virtual Observatory", collaboration of 20 groups
  - *Astronomy data centers*
  - *National observatories*
  - *Supercomputer centers*
  - *University departments*
  - *Computer science/information technology specialists*
- Similar projects now in 15 countries world-wide

⇒ International Virtual Observatory Alliance

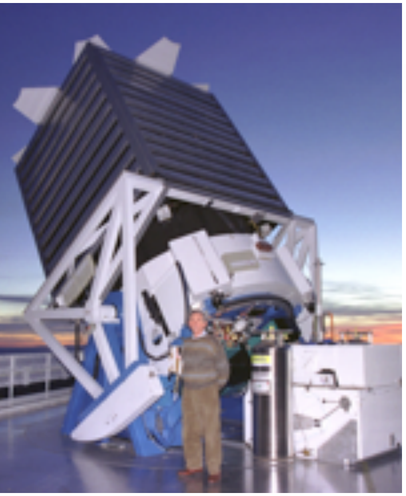
NSF+NASA=>


SDSS
Telescope
(2.5 meter
mirror)

LSST
Telescope
8.4 meter
mirror

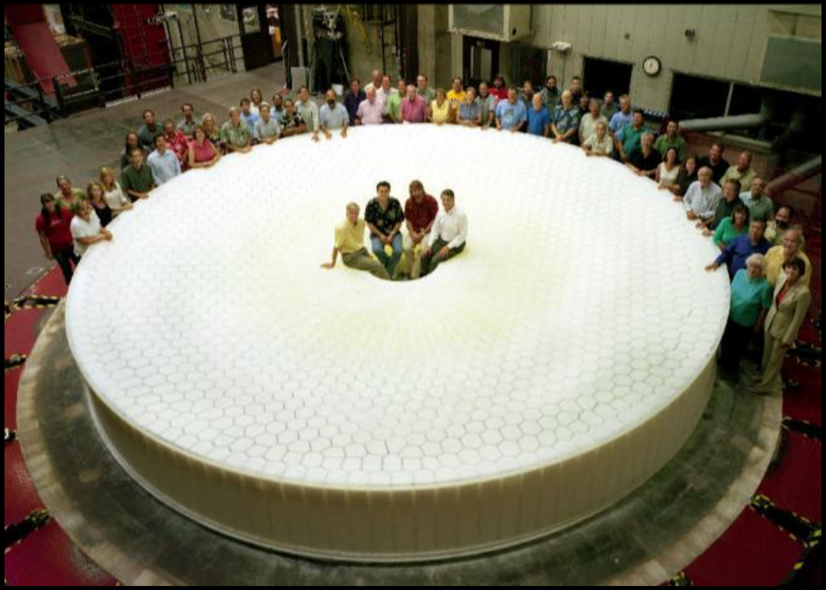


SDSS
Telescope
2.5 meter
mirror



Large Synoptic Survey Telescope (LSST)

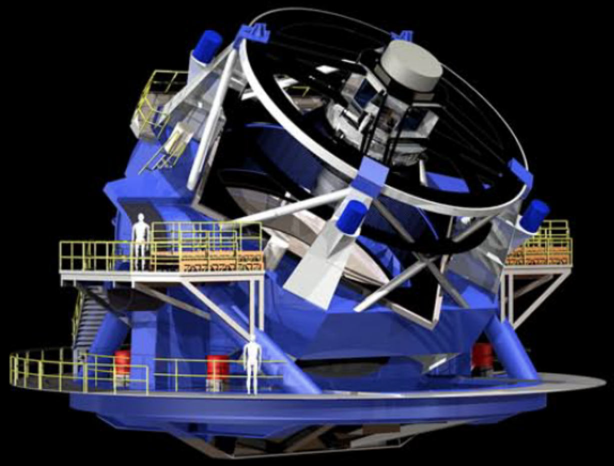
Primary/Tertiary cast from a single borosilicate blank.

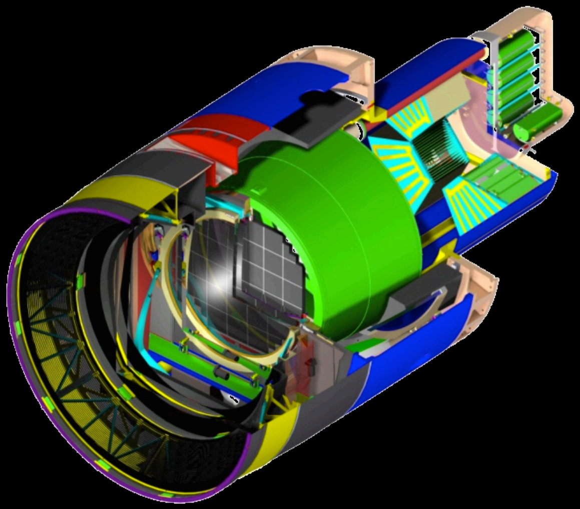


- Primary-Tertiary was cast in the spring of 2008.
- Secondary fabricated by Corning in 2009.

Large Synoptic Survey Telescope **2014**

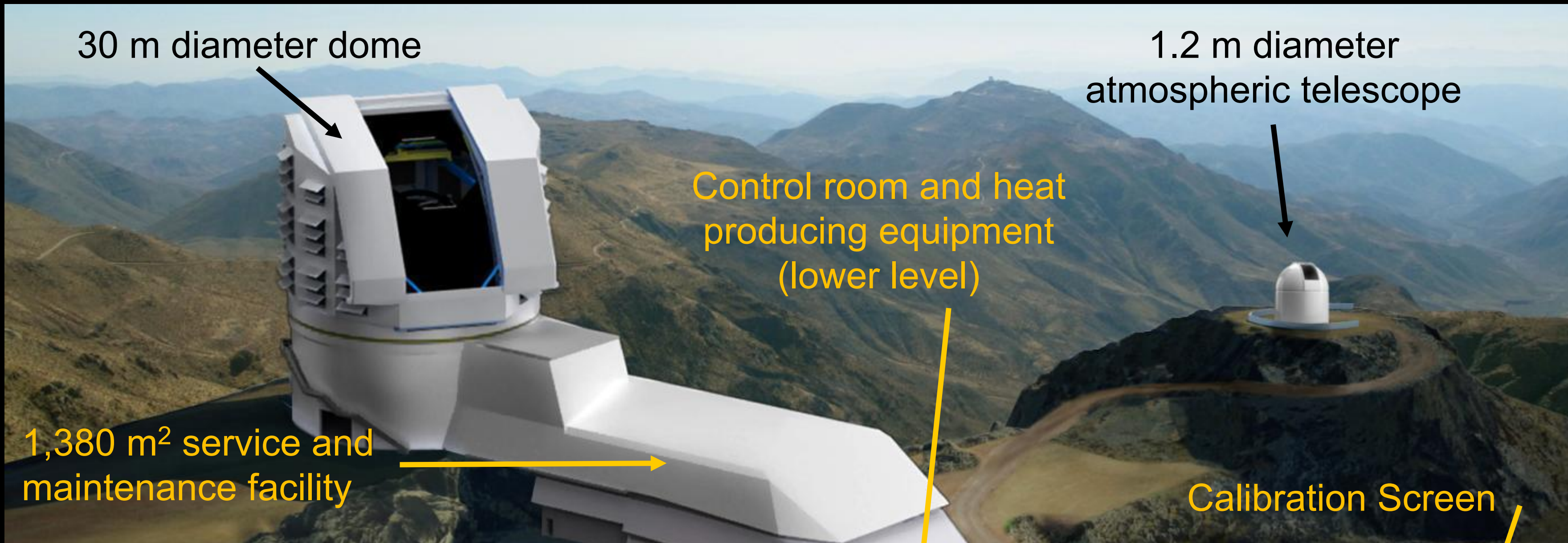
- **Wide field and deep**
  - 27000 sq deg (wide)
  - 100 - 200 sq deg (deep)
  - 10 years
- **Broad range of science**
  - Dark energy
  - Galactic structure
  - Census of the Solar system
  - Transient universe
- **3.2 Gpixel camera**
  - 9.6 sq degree FOV
  - ugrizy filters

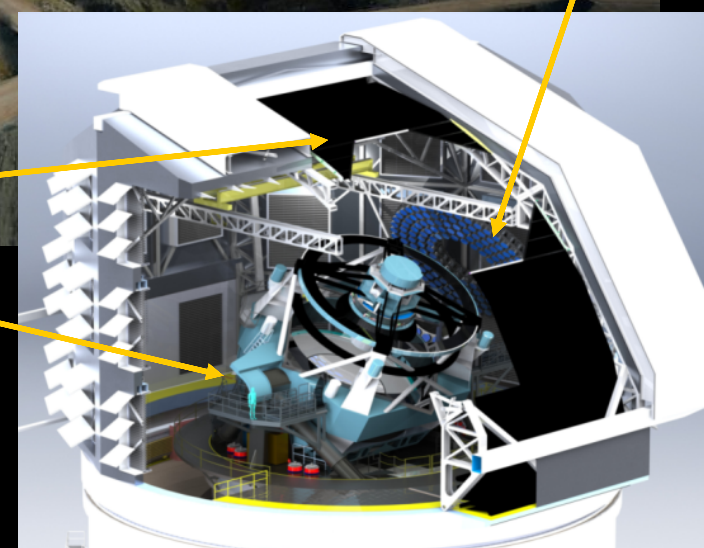




The LSST Site and Base Facilities in Chile

8.4m survey telescope and 1.2m atmospheric telescope





LSST's most remarkable data product will be a 10-year "movie" of the entire sky. This time-lapse coverage of the night sky will open up time-domain astronomy.

Processing the data flow from the LSST

- **Each "Visit" comprises a pair of back-to-back exposures**
  - 2x15 sec exposure; duration = 34 seconds with readout
- **The data volume associated with this cadence is unprecedented**
  - **one 6-gigabyte image every 17 seconds**
  - **15 terabytes of raw scientific image data / night**
  - **100-petabyte final image data archive**
  - **20-petabyte final database catalog**
  - **2 million real time events per night every night for 10 years**
  - **1000 new supernovas discovered every night!**

Precision Cosmology: Constraints on Dark Energy

- LSST will probe the nature of Dark Energy via a distinct set of complementary probes:
  - SNe Ia's as "standard candles"
  - Baryon acoustic oscillations as a "standard rulers"
  - Studies of growth of structure via weak gravitational lensing
  - Studies of growth of structure via clusters of galaxies

- In conjunction with one another, this rich spectrum of tests is crucial for reduction of systematics and dependence on nuisance parameters.

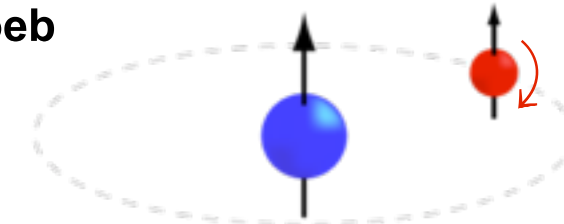
- These tests also provide interesting constraints on other topics in fundamental physics: the nature of inflation, modifications to GR, the masses of neutrinos.

Square Kilometer Array (SKA)

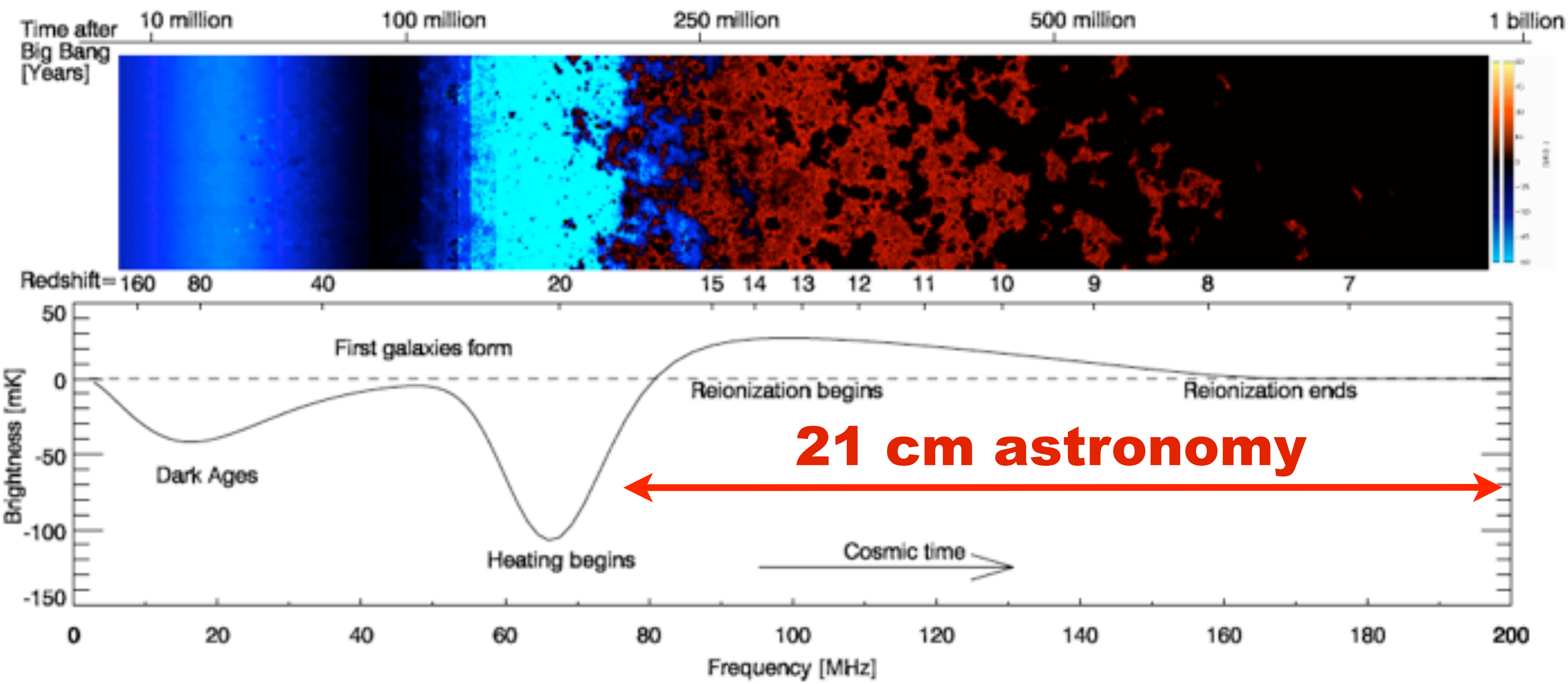
21 cm Cosmology in the 21st Century

Jonathan R. Pritchard & Abraham Loeb

Rep. Prog. Phys. 75, 086901 (2012)



The First Billion Years



The Square Kilometre Array

Exploring the Universe with the world's largest radio telescope

The project timeline

2024	Full science operations with phase two
2020-24	Phase two construction
2020	Full science operations with phase one
2016-20	Phase one construction
2013-15	Detailed design and pre-construction phase
2012	Site selection South Africa & Western Australia
2011	Establish SKA organisation as a legal entity
2008-12	Telescope conceptual design
2006	Short listing of suitable sites
1991	Concept

Facts and figures

The SKA will contain thousands of antennas with a combined collecting area of about one square kilometre (that's 1 000 000 square metres!).

The SKA central computer will have processing power of about 100 Petaflops/s.

The SKA will use enough optical fibre to wrap twice around the earth.

The dishes of the SKA will produce 10 times the 2012 global internet traffic.

The SKA will have 50 times the sensitivity and 10,000 times the survey speed of the best current-day radio telescopes.

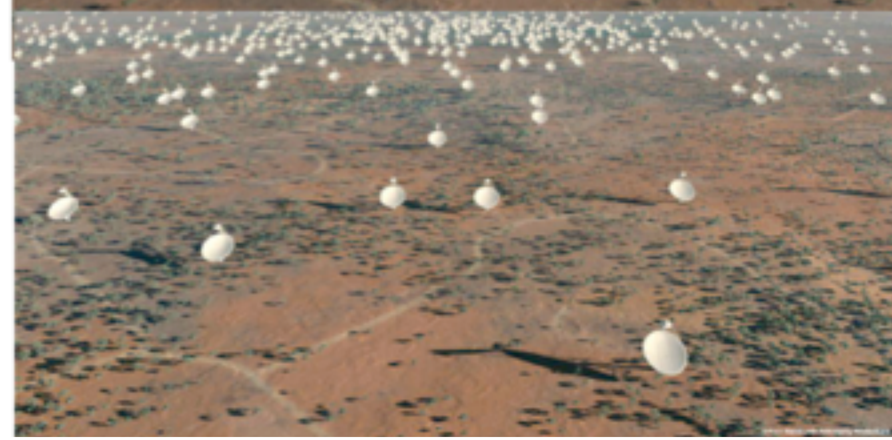
Square Kilometer Array Locations

Square Kilometer Array Antenna Types

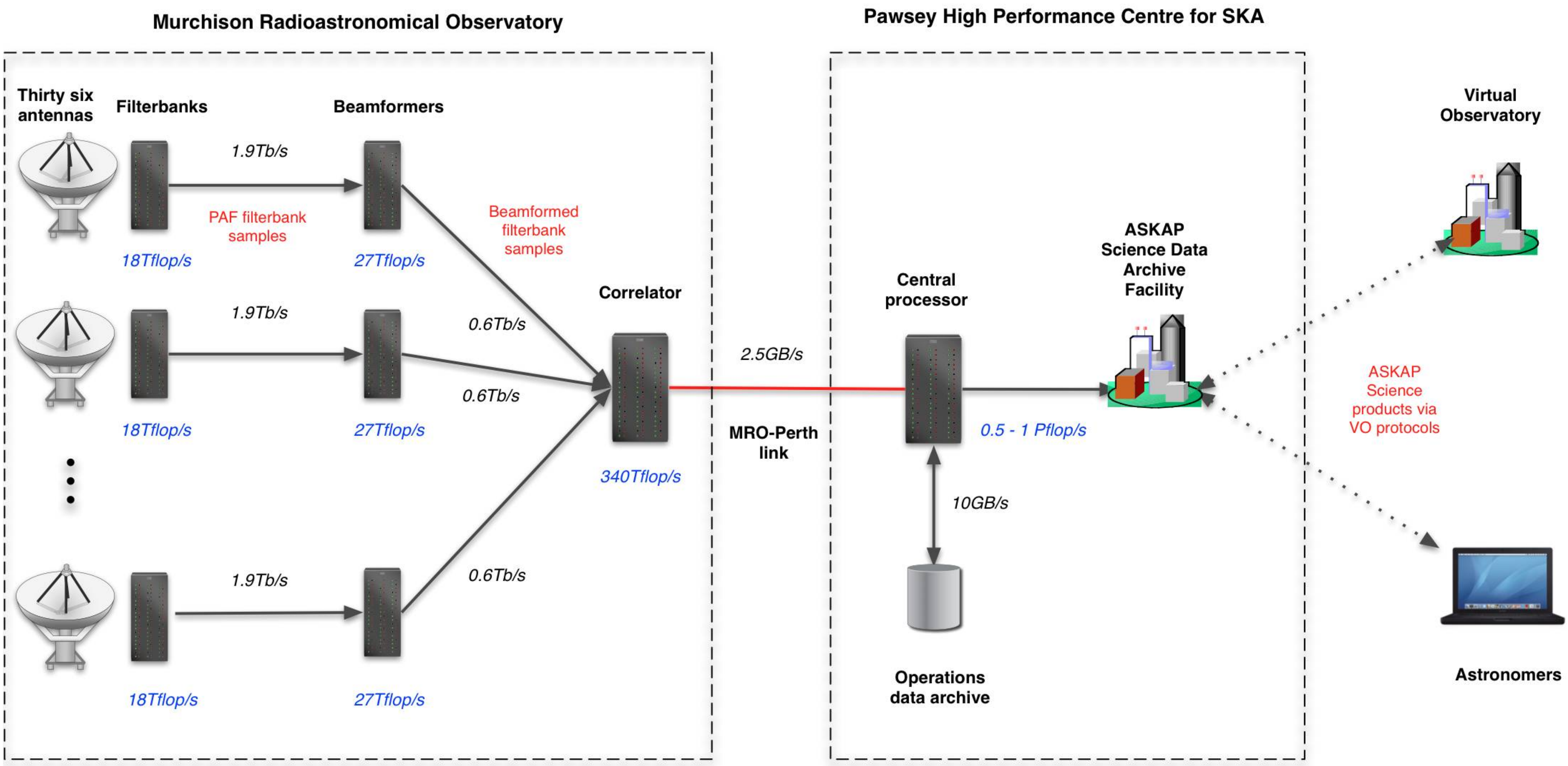
Sparse Aperture Arrays

Dense Aperture Arrays

Radio Dishes



Australian SKA Pathfinder - ASKAP



Total output data rate per antenna = 0.6Tbps.

Big Challenges of AstroComputing

Big Data

Sloan Digital Sky Survey (SDSS) 2008

2.5 Terapixels of images
40 Tb raw data ➠120 Tb processed
35 Tb catalogs

Mikulski Archive for Space Telescopes (MAST)

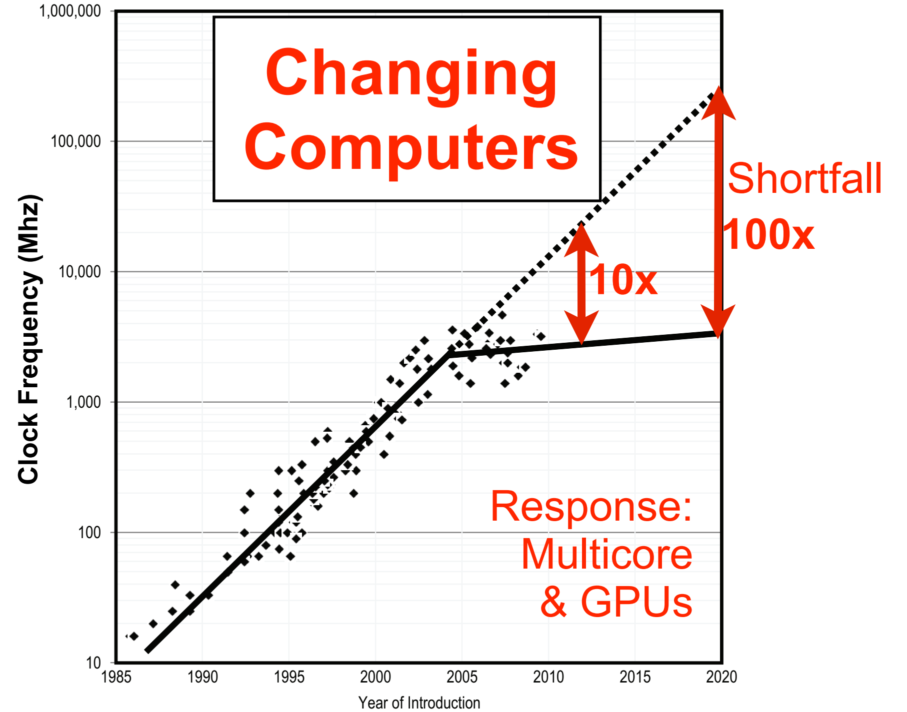
185 Tb of images
25 Tb/year ingest rate
>100 Tb/year retrieval rate

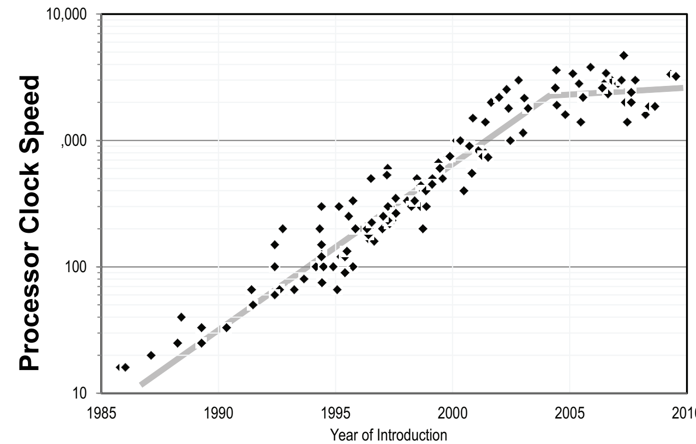
Large Synoptic Survey Telescope (LSST) 2014

15 Tb per night for 10 years
100 Pb image archive
20 Pb final database catalog

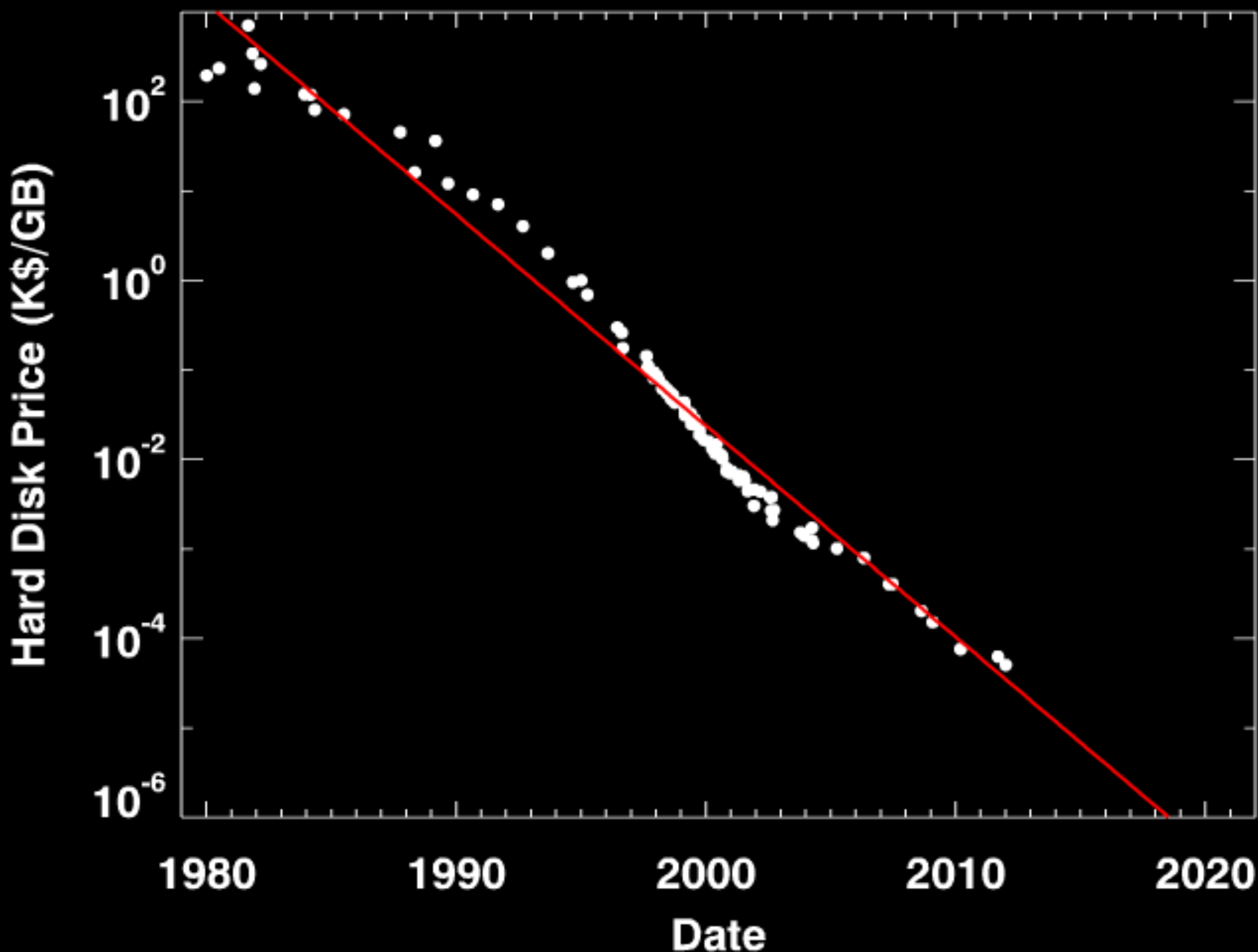
Square Kilometer Array (SKA) ~2024

1 Eb per day (> internet traffic today)
100 PFlop/s processing power
~1 Eb processed data/year





Disk Cost per Gigabyte



The Big Data Future in Astronomy

Exponential growth in computing power and detectors and falling cost of data storage has enabled vast increases in

- Ambitious surveys, with massive storage for archives

- Simulation realism - virtual experiments on the universe

Astronomy is becoming dominated by surveys and simulations

How can we understand such huge amounts of data?
We need data microscopes and telescopes!
We have to analyze outputs as the supercomputers run
Users will send questions (algorithms) to where the data is stored and get back answers (not raw data)

High Performance Scientific Computing Needs

The challenges facing us are

- "Big data" -- too large to move -- from more powerful observations, larger computer outputs, and falling storage costs
- Changing high-performance computer architecture -- from networked single processors to multicore and GPUs

These challenges demand new collaborations between natural scientists and computer scientists to develop

- Tools and scientific programmers to convert legacy code and write new codes efficient on multicore/GPU architectures, including fault tolerance and automatic load balancing
- New ways to visualize and analyze big data remotely
- Train new generations of scientific computer users
- Improve education and outreach


Triumphs and tribulations of ΛCDM, the double dark theory

Joel R. Primack

Review

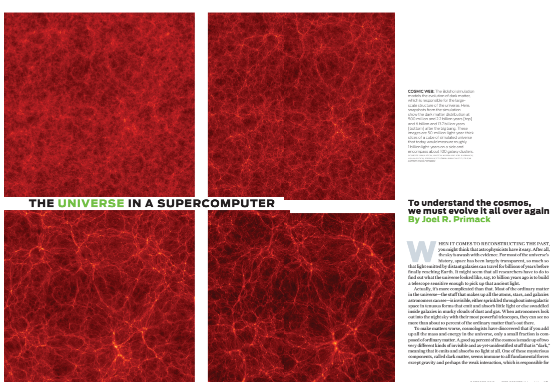
Joel R. Primack, Triumphs and Tribulations of ΛCDM, the Double Dark Theory, Annalen der Physik **524**, 535 (2012)

goo.gl/TFDYf

Popular Articles

Joel R. Primack and Trudy Bell, Universe in a Box, Sky & Telescope, July 2012

bit.ly/QV7D58



Joel R. Primack, The Universe in a Supercomputer, IEEE Spectrum, October 2012

bit.ly/QqWhUQ