Exascale IT Requirements in Astronomy

Joel Primack, University of California, Santa Cruz

Director, University of California High-Performance AstroComputing Center

HUAWEI | HCC 2012

Exascale IT Requirements in Astronomy

Joel Primack, University of California, Santa Cruz

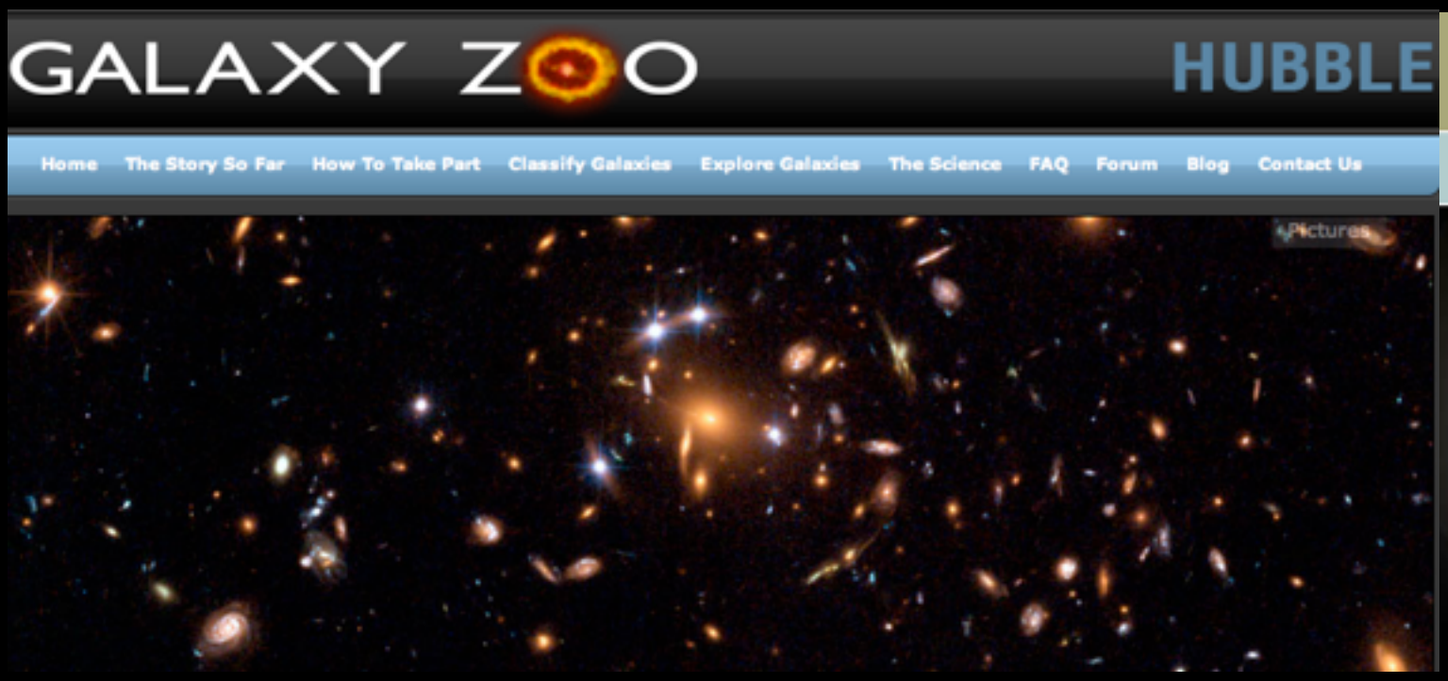
1. What's Special About Astronomy
2. Sloan Digital Sky Survey & HST
3. Large Synoptic Survey Telescope
4. Square Kilometer Array
5. Computational Cosmology

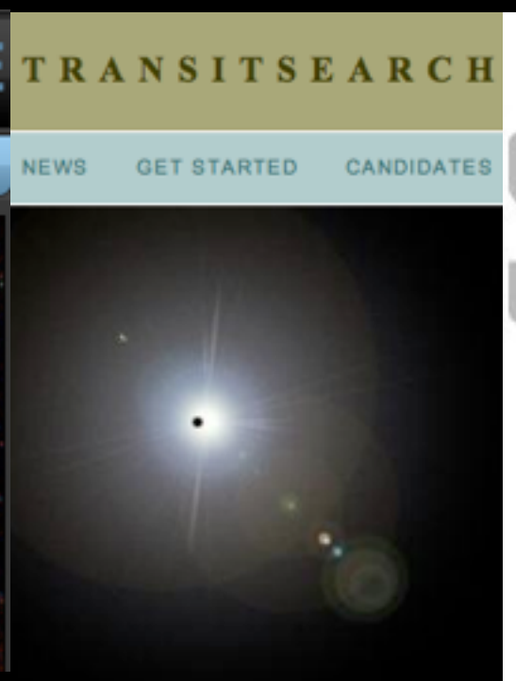
Note: 1000 Gb = Terabyte= Tb = 10^{12} bytes
1000 Tb = Petabyte = Pb = 10^{15} bytes
1000 Pb = Exabyte = Eb = 10^{18} bytes

Bruce Munro's sea of 600,000 CDs ≈ 500 Tb

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 pretty **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) 2012

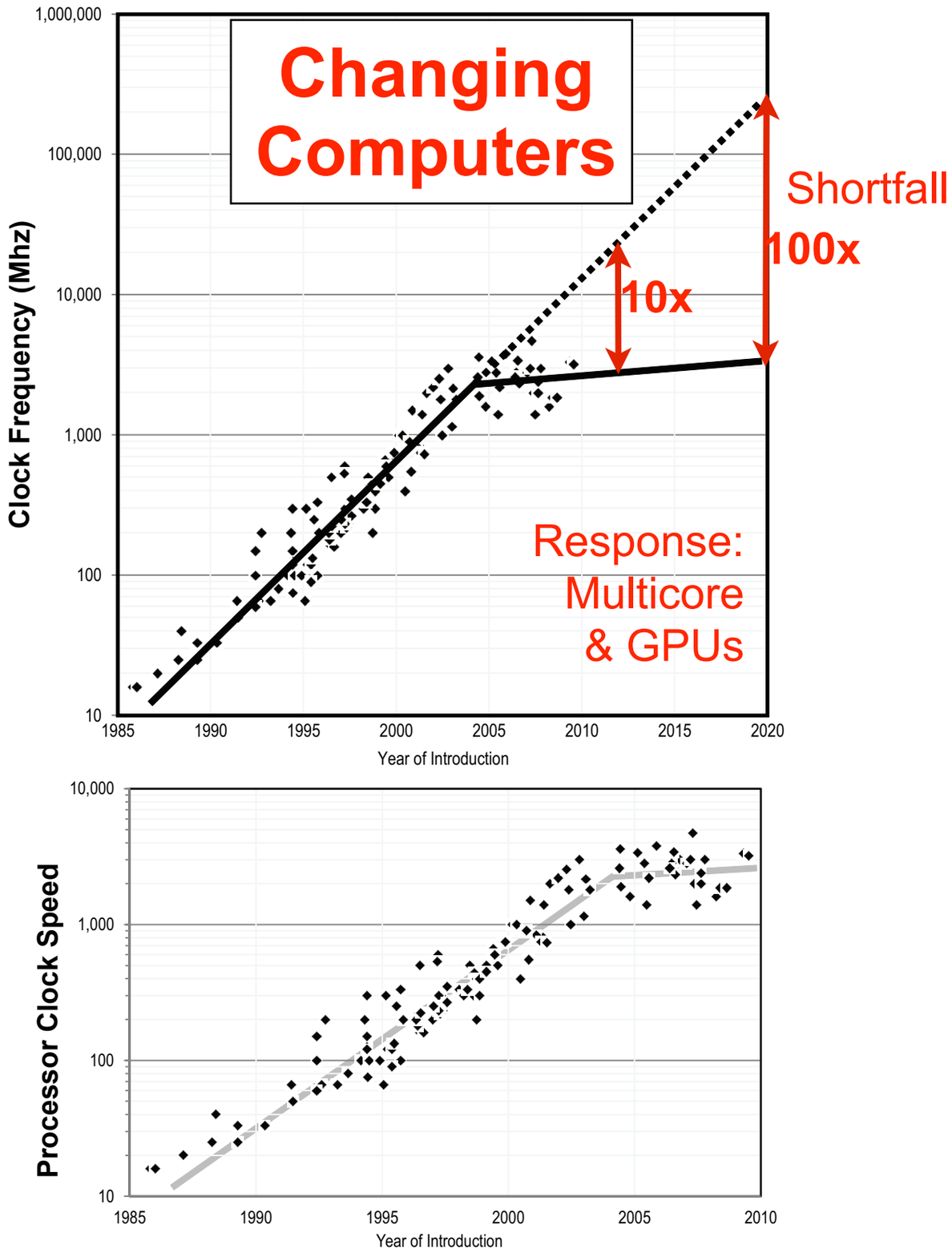
185 Tb of images
25 Tb/year ingest rate
>100 Tb/year retrieval rate

Large Synoptic Survey Telescope (LSST) ~2020

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

Square Kilometer Array (SKA) ~2024

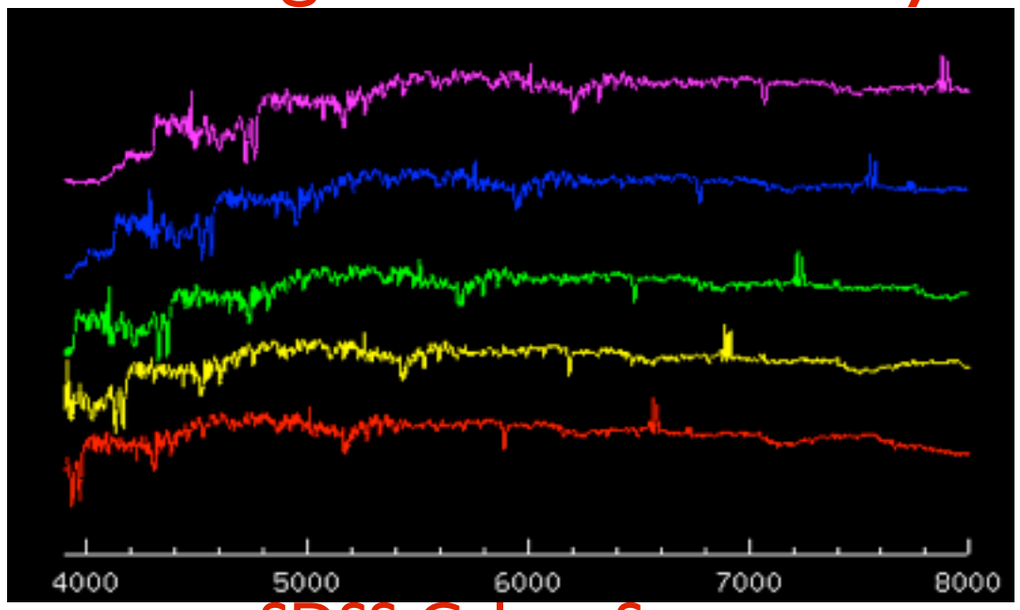
1 Eb per day (> internet traffic today)
100 PFlop/s processing power
~1 Eb processed data/year



Sloan Digital Sky Survey (SDSS)

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

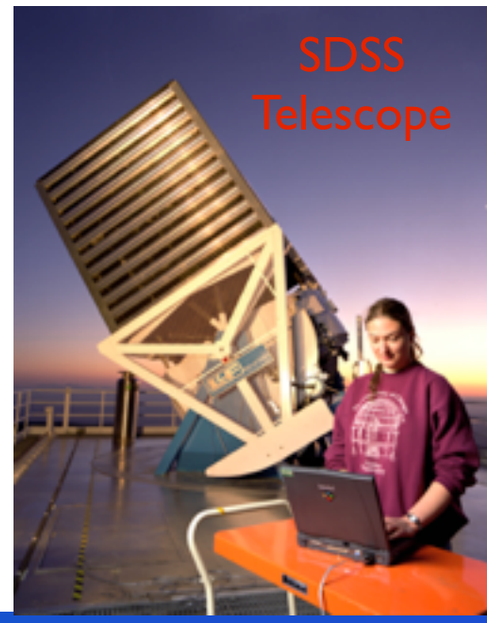
Imaging survey in 5 wavelength bands 5-color images of ¼ of the sky
Spectroscopic redshift survey



SDSS Galaxy Spectra

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

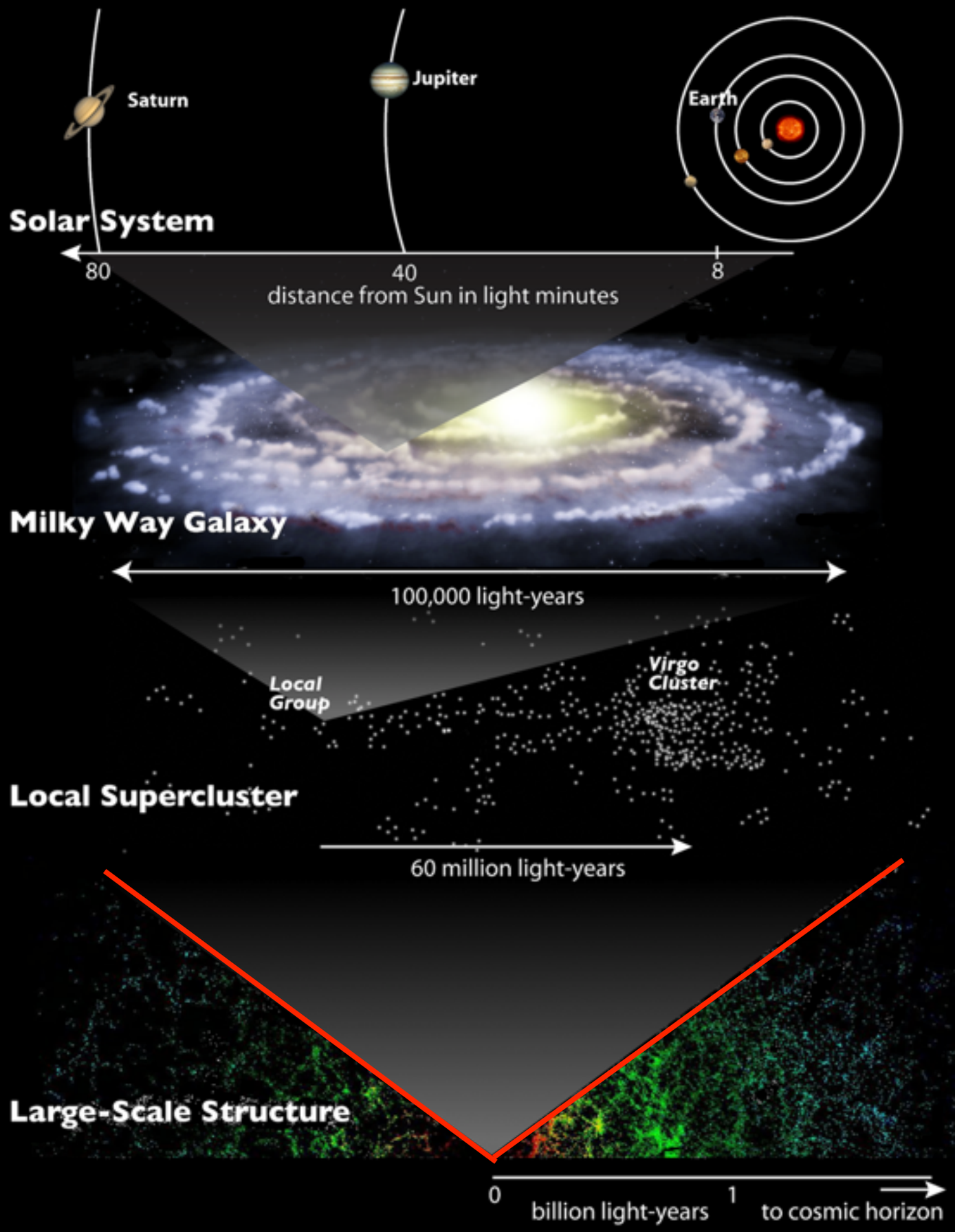
Data is publicly accessed
- 840 million web hits in 9 years, now >1 billion
- 4,000,000 distinct users* vs. 15,000 astronomers
- Basis for ~20,000 scientific papers
- More citations than any telescope including Hubble



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

The Modern Scientific Cosmos

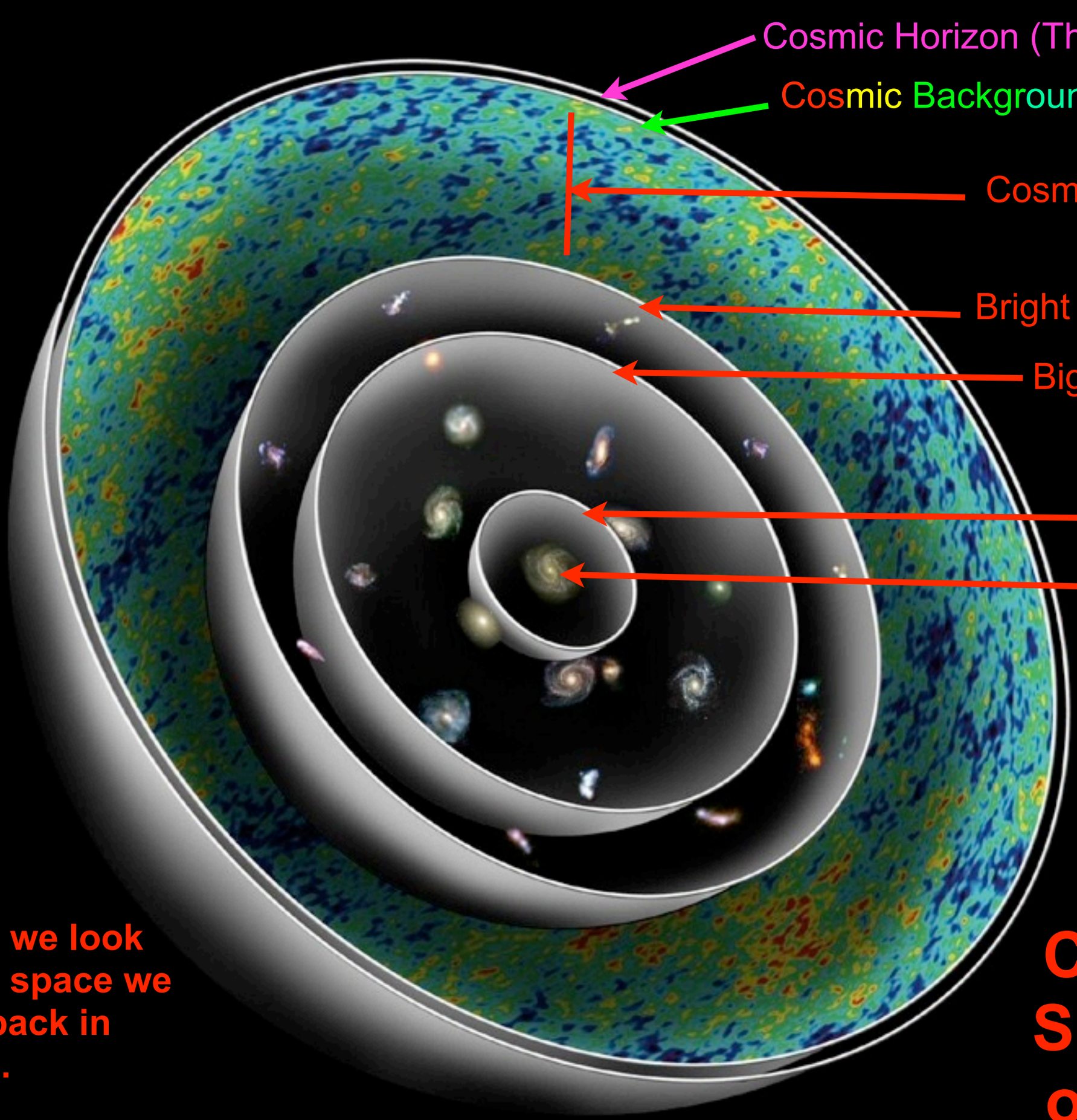
Our Cosmic Address



each dot is a big galaxy

Sloan Digital Sky Survey

GALAXIES MAPPED BY THE SLOAN SURVEY

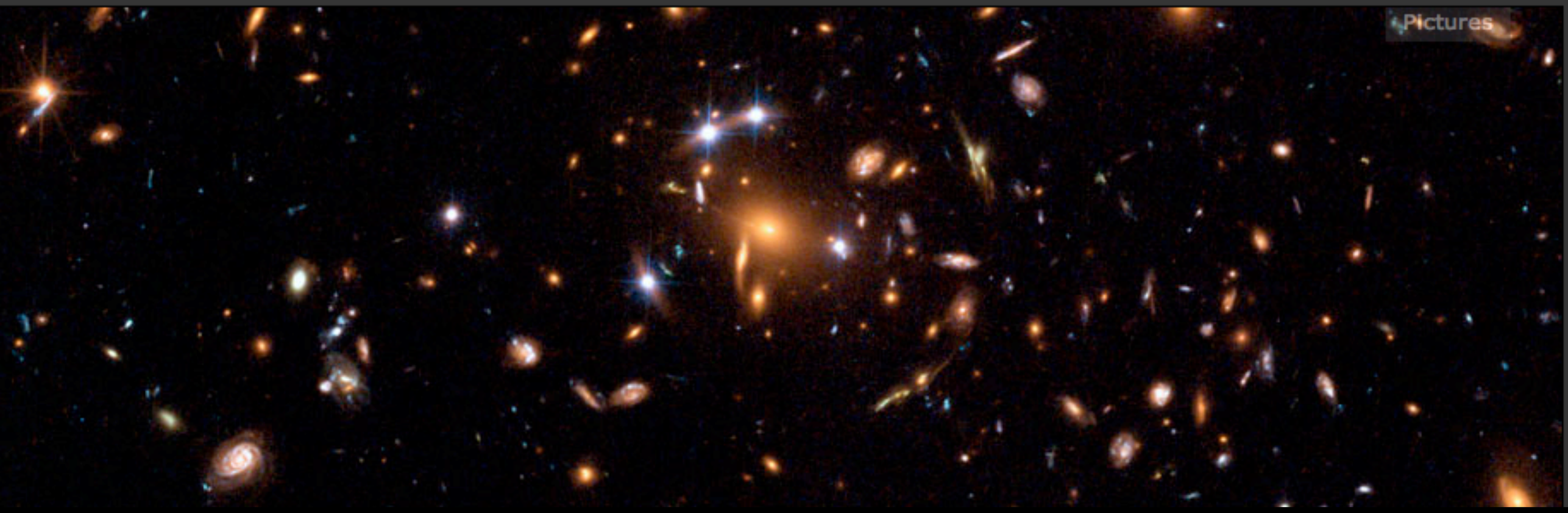


When we look out in space we look back in time…

Cosmic Spheres of Time

GALAXY ZOO HUBBLE

Home | The Story So Far | How To Take Part | Classify Galaxies | Explore Galaxies | The Science | FAQ | Forum | Blog | Contact Us



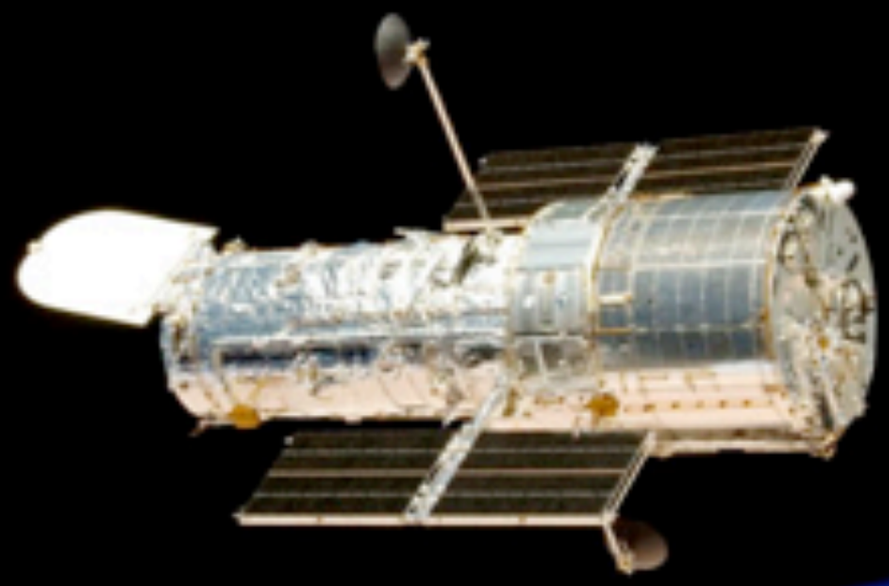
Galaxy Zoo started as an offshoot of the Sloan Digital Sky Survey

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

GALAXY ZOO HUBBLE

Home The Story So Far How To Take Part Classify Galaxies Explore Galaxies The Science FAQ Forum Blog Contact Us

Pictures



http://www.galaxyzoo.org

Welcome to Galaxy Zoo, where you can help astronomers explore the Universe

Galaxy Zoo: Hubble uses gorgeous imagery of hundreds of thousands of galaxies drawn from NASA's Hubble Space Telescope archive. To understand how these galaxies, and our own, formed we need your help to classify them according to their shapes — a task at which your brain is better than even the most advanced computer. If you're quick, you may even be the first person in history to see each of the galaxies you're asked to classify.

Classifier Log In

Click here to log in

- Register
- Forgotten Password?

Explore galaxies

Mikulski Archive for Space Telescopes (MAST)

What is the STScI archive?

Mikulski Archive for Space Telescopes: **MAST**

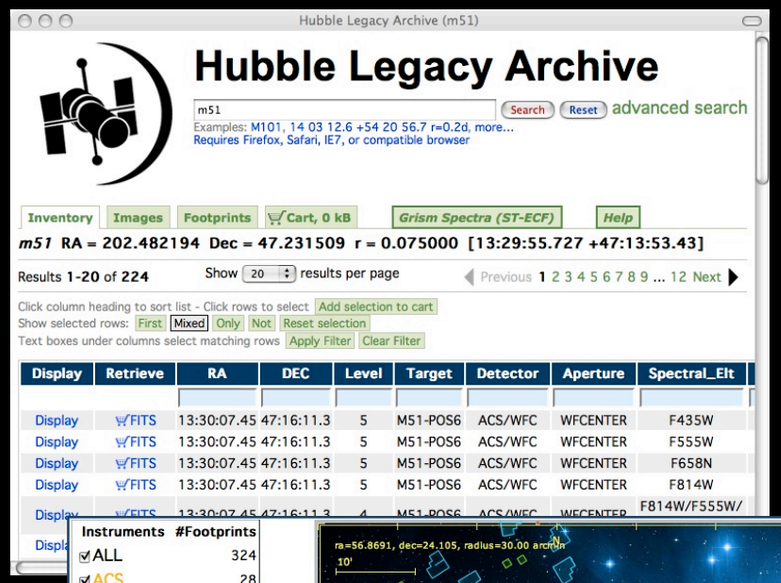
- Data

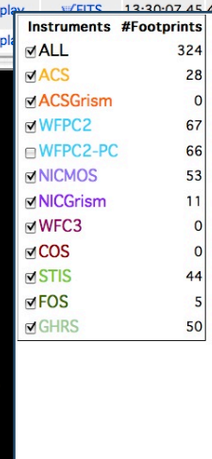
  ~185 TB of images, spectra, catalogs, time series

- Metadata

  ~10^6 HST observations (plus other missions)

  Documentation, publication links, …



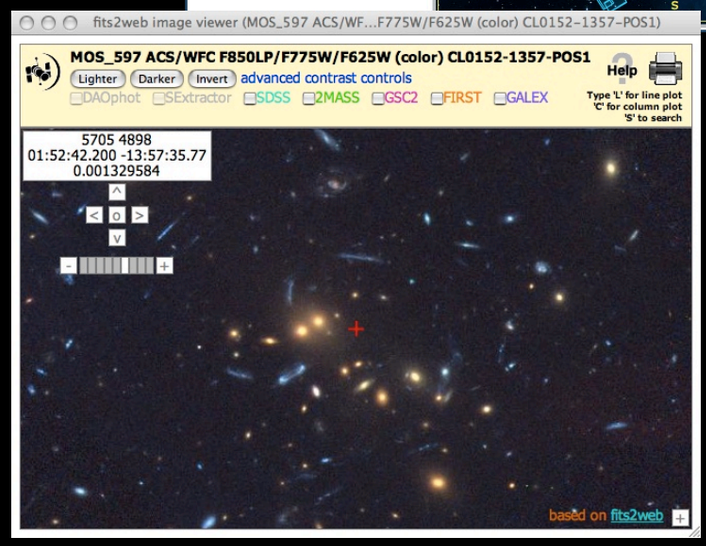


- User interfaces

  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>
```

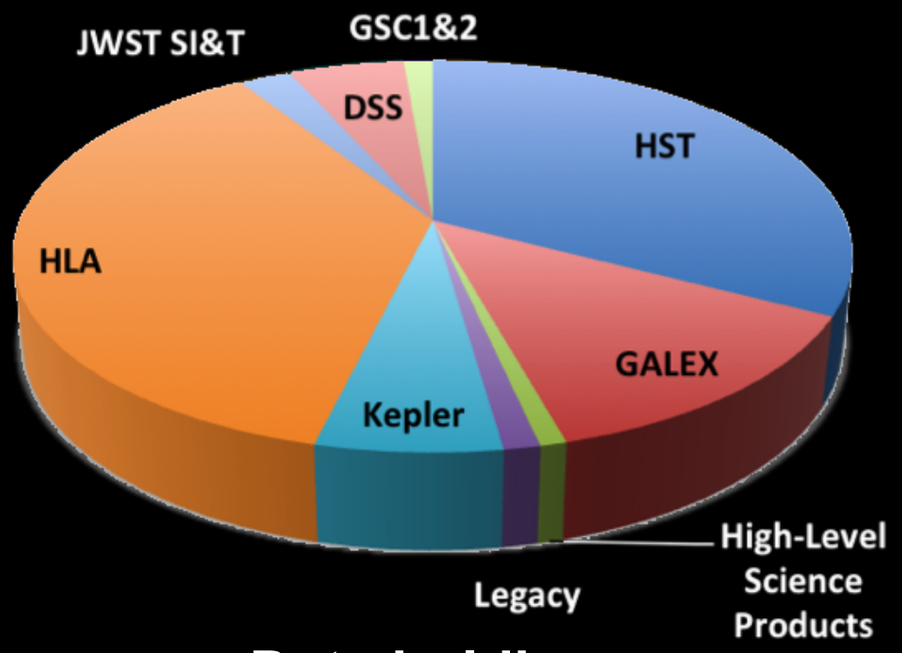
- Services

  VO services, data retrieval, image cutouts, …

  (UIs are built around VO services)

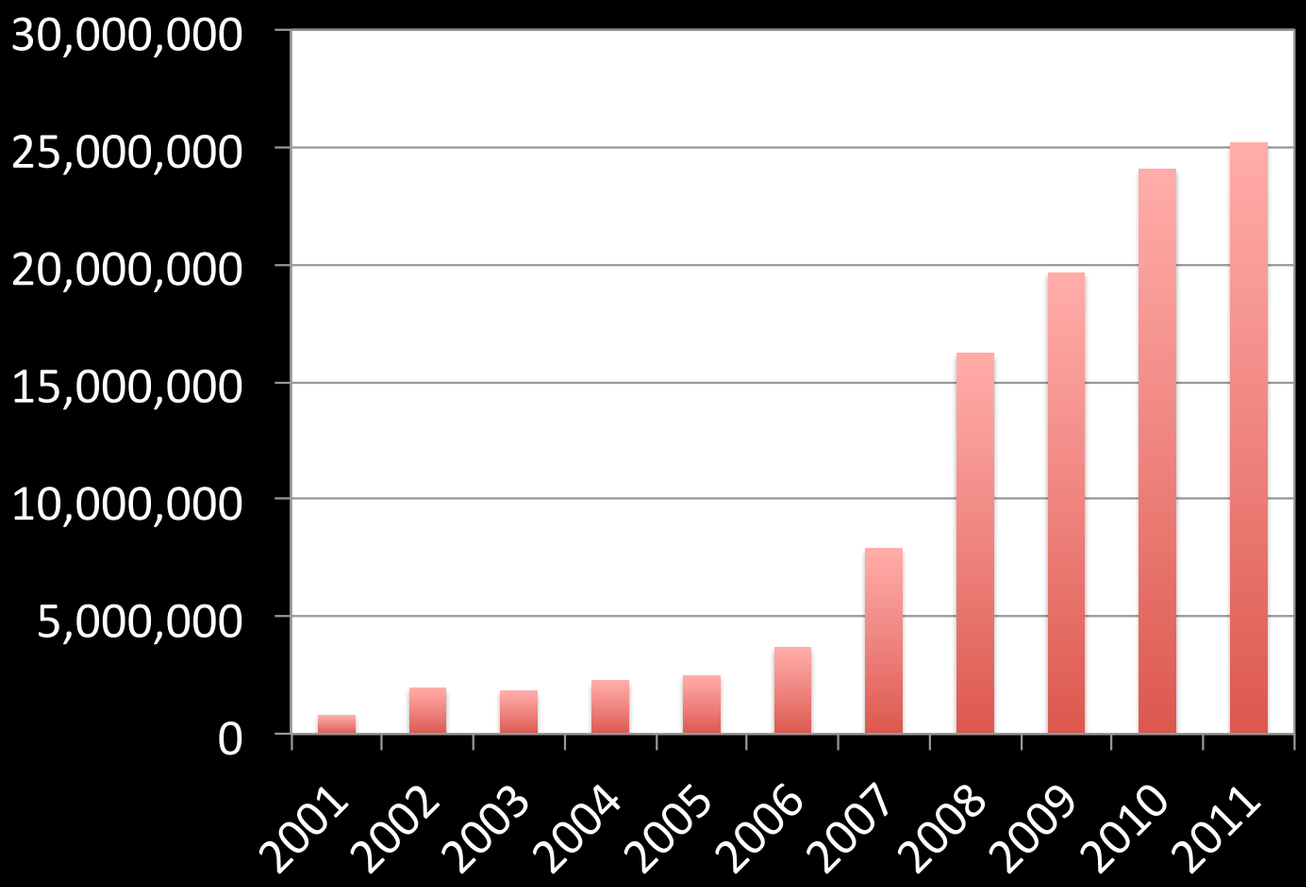
The MAST Archive: 2 minute summary

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

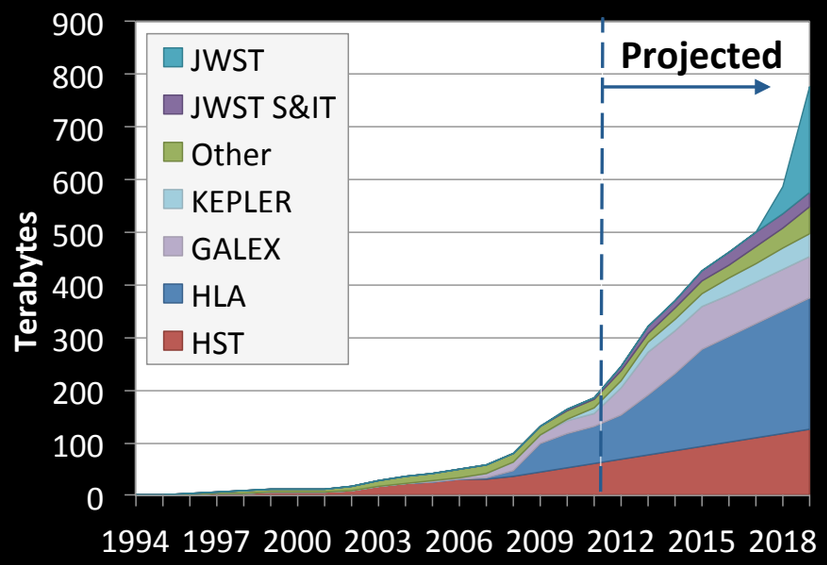


Data holdings

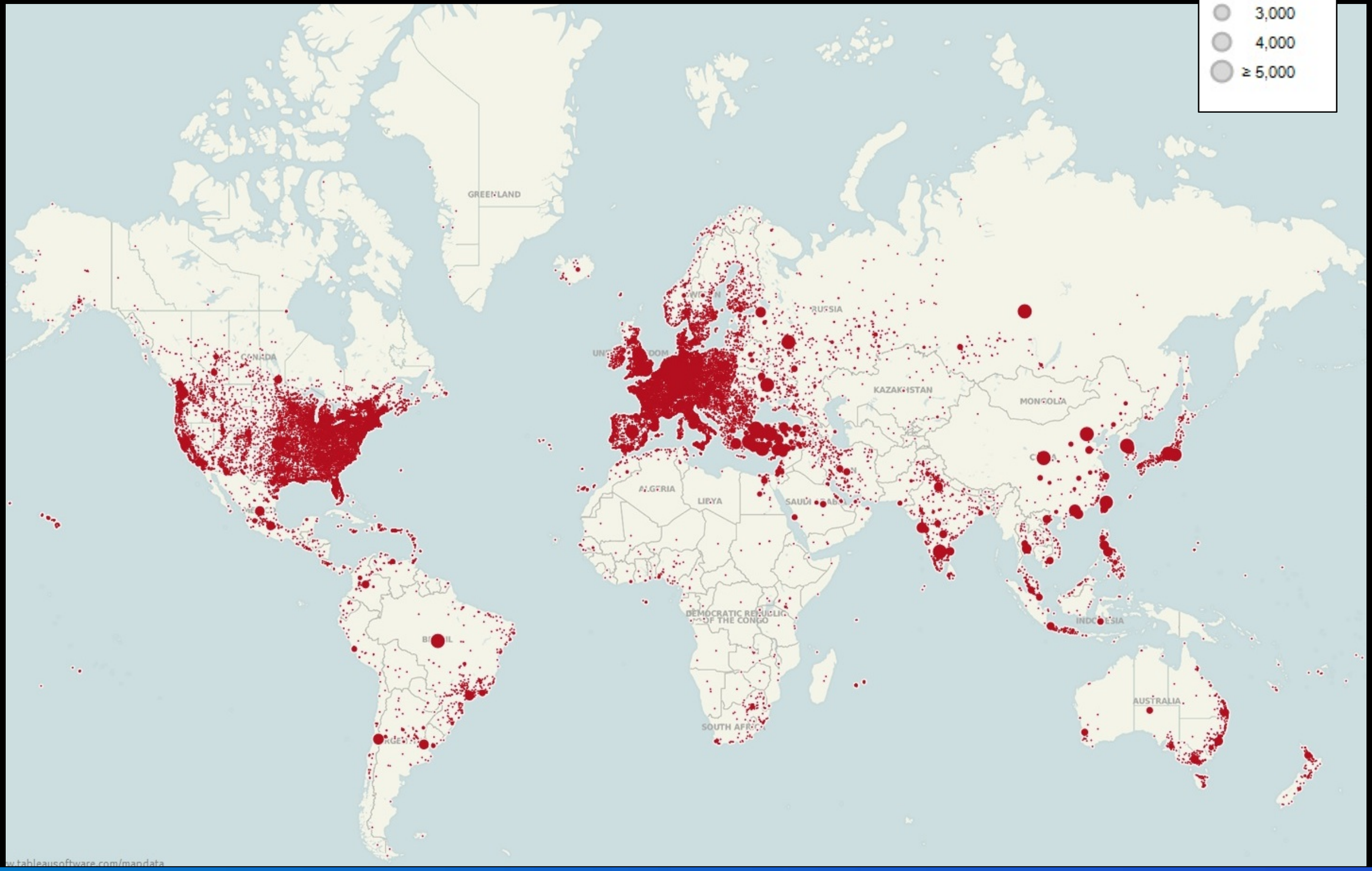
Number of searches per year



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

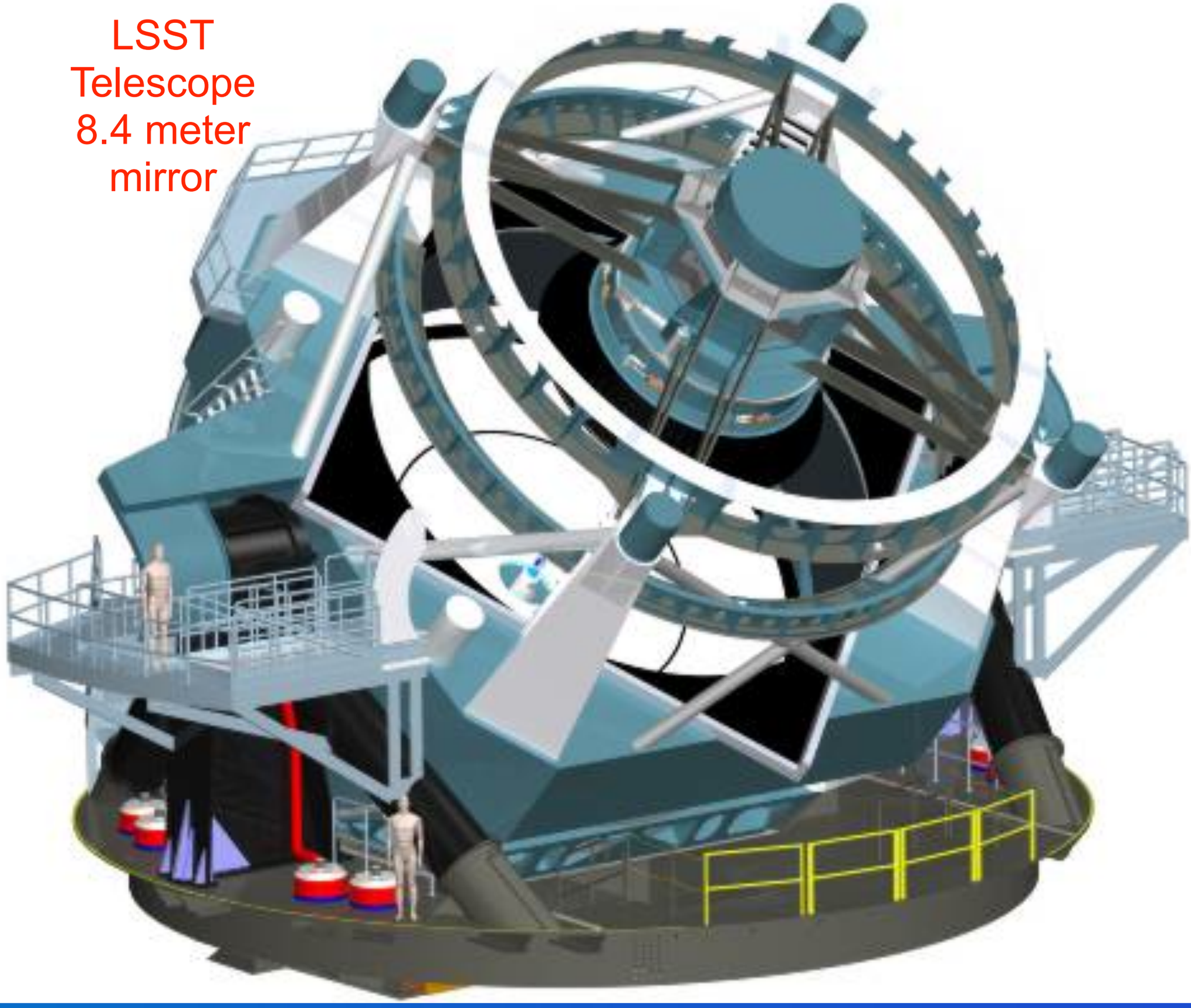
⇒ International Virtual Observatory Alliance

NSF+NASA=>

SDSS Telescope (2.5 meter mirror)

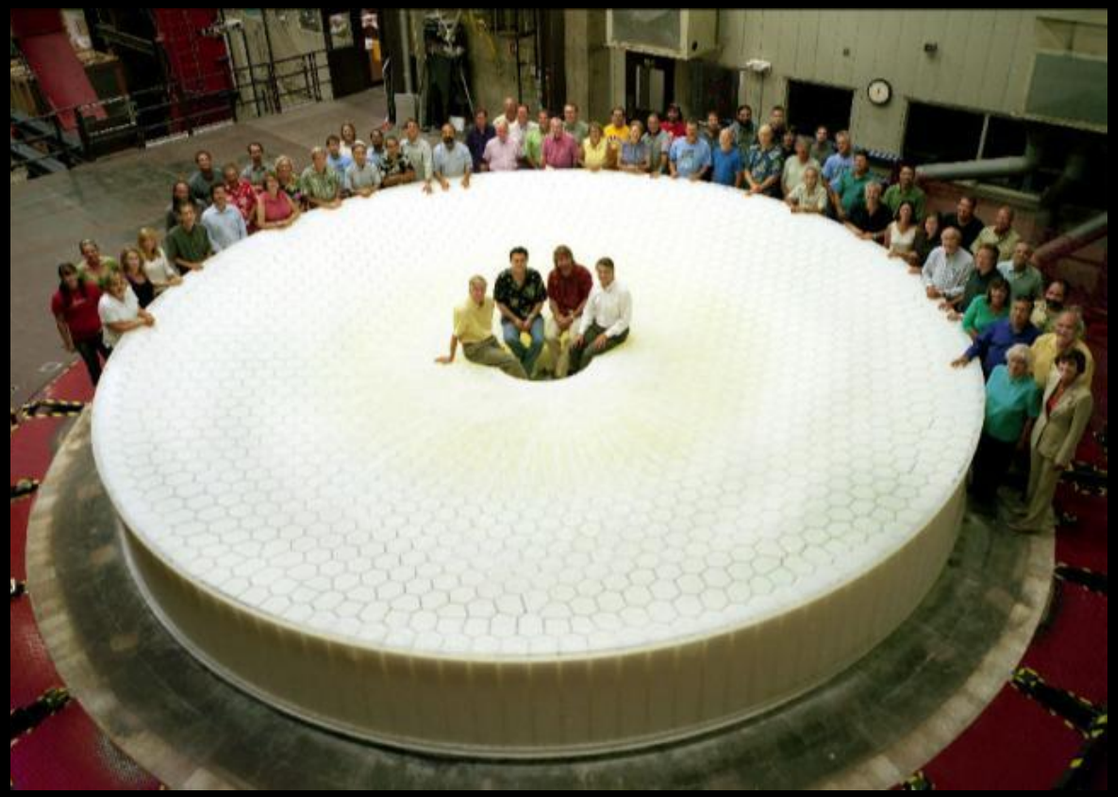
LSST
Telescope
8.4 meter
mirror

SDSS
Telescope
2.5 meter
mirror



Large Synoptic Survey Telescope (LSST)

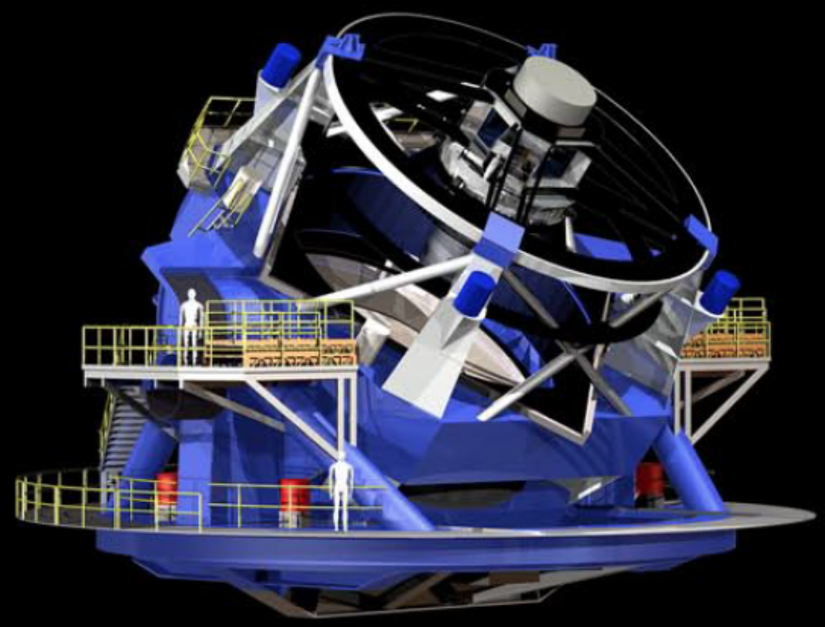
Primary/Tertiary cast from a single borosilicate blank.

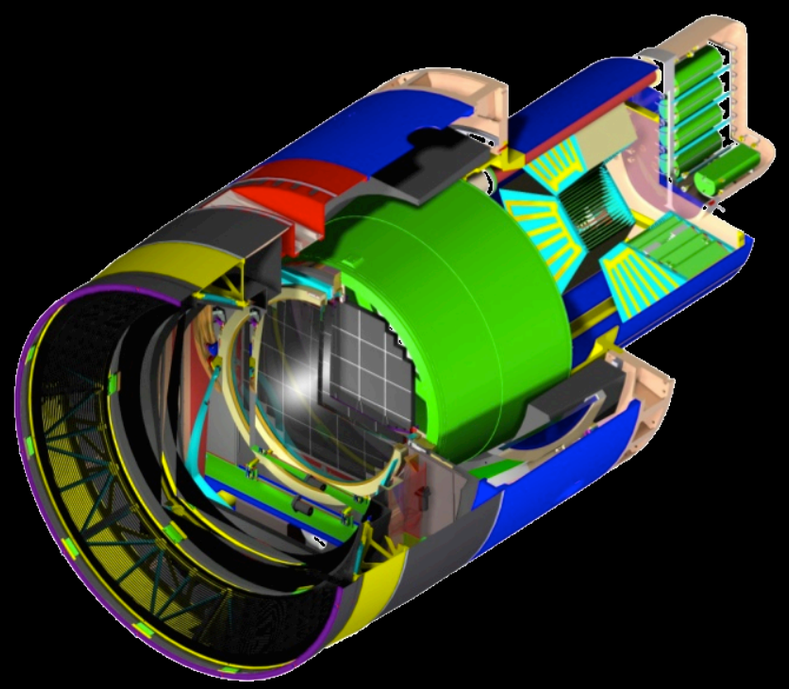


- Primary-Tertiary was cast in the spring of 2008.
- Secondary fabricated by Corning in 2009.

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

30 m diameter dome

1.2 m diameter atmospheric telescope

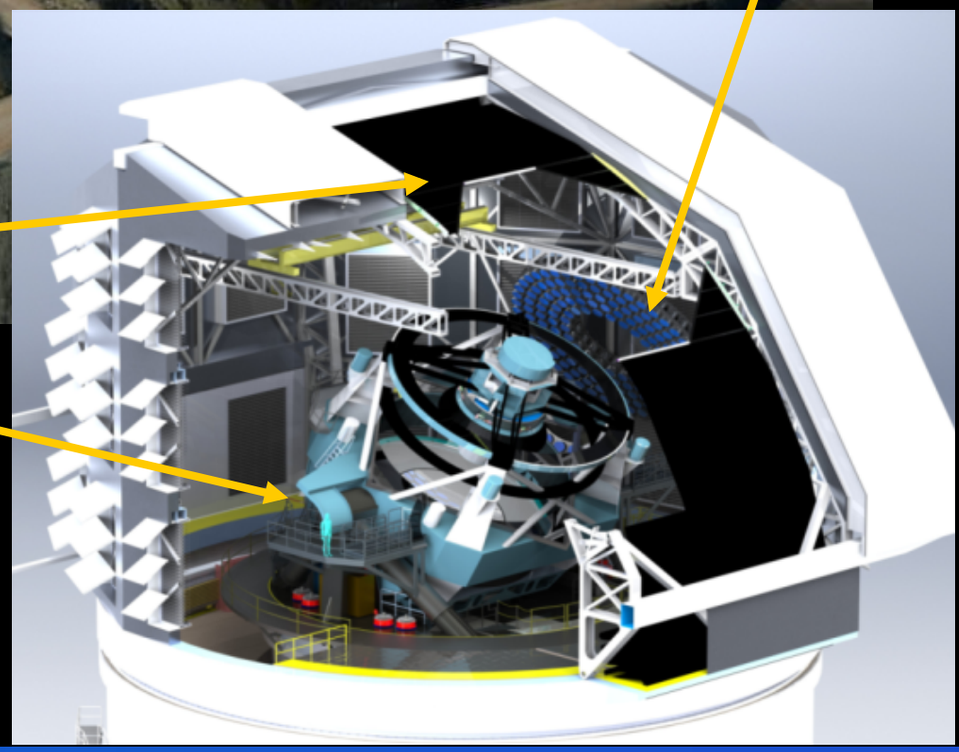
Control room and heat producing equipment (lower level)

1,380 m^2 service and maintenance facility

Calibration Screen

Stray light and Wind Screen

Base Facility



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.

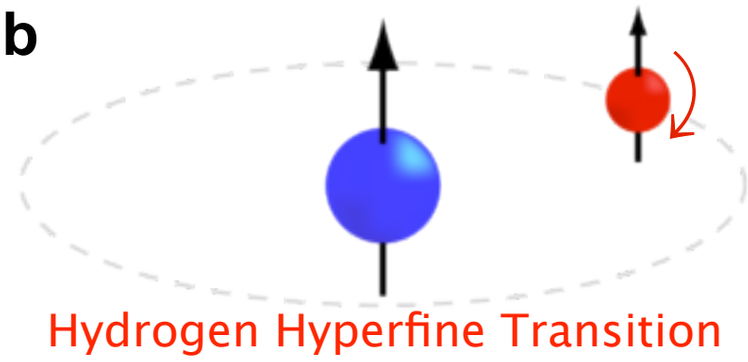
- These tests also provide interesting constraints on other topics in fundamental physics: the nature of inflation, modifications to GR, the masses of neutrinos.

Square Kilometer Array (SKA)

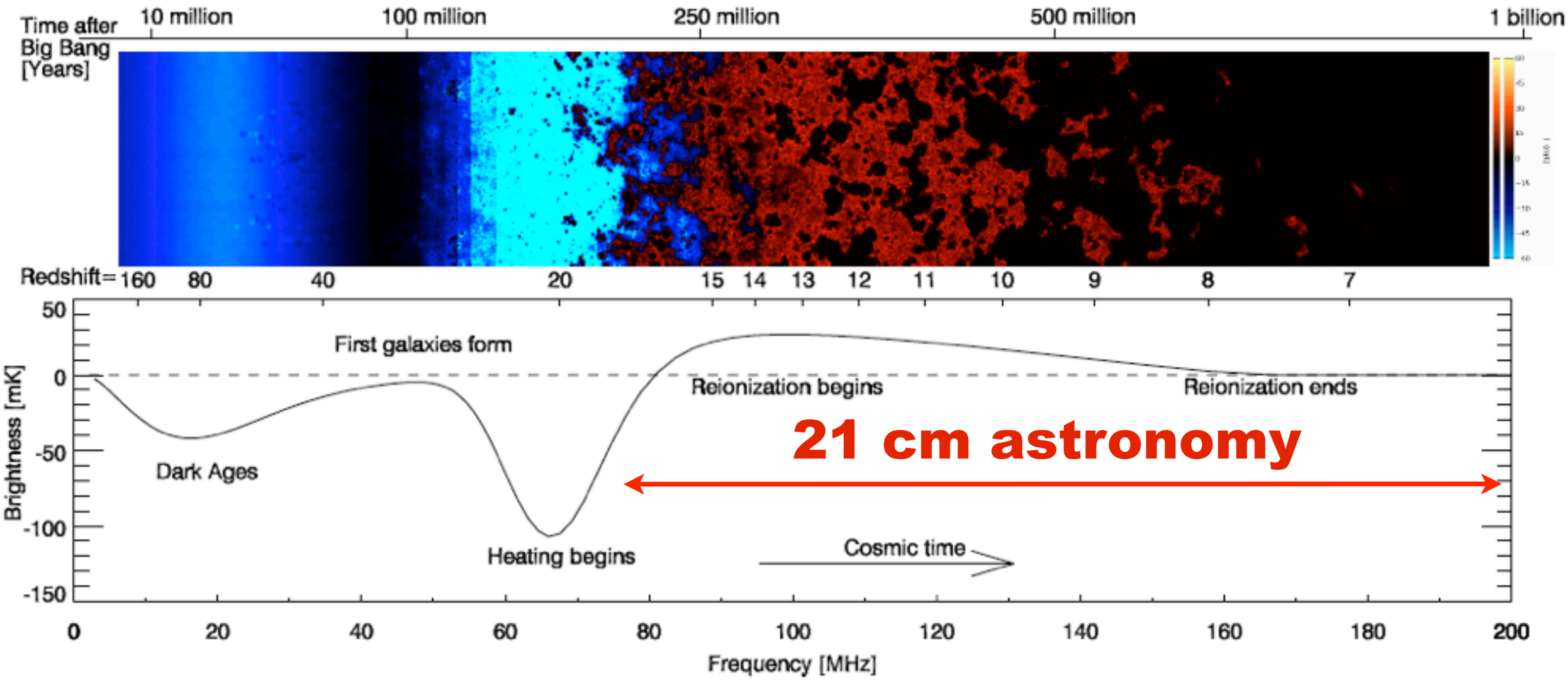
21 cm Cosmology in the 21st Century

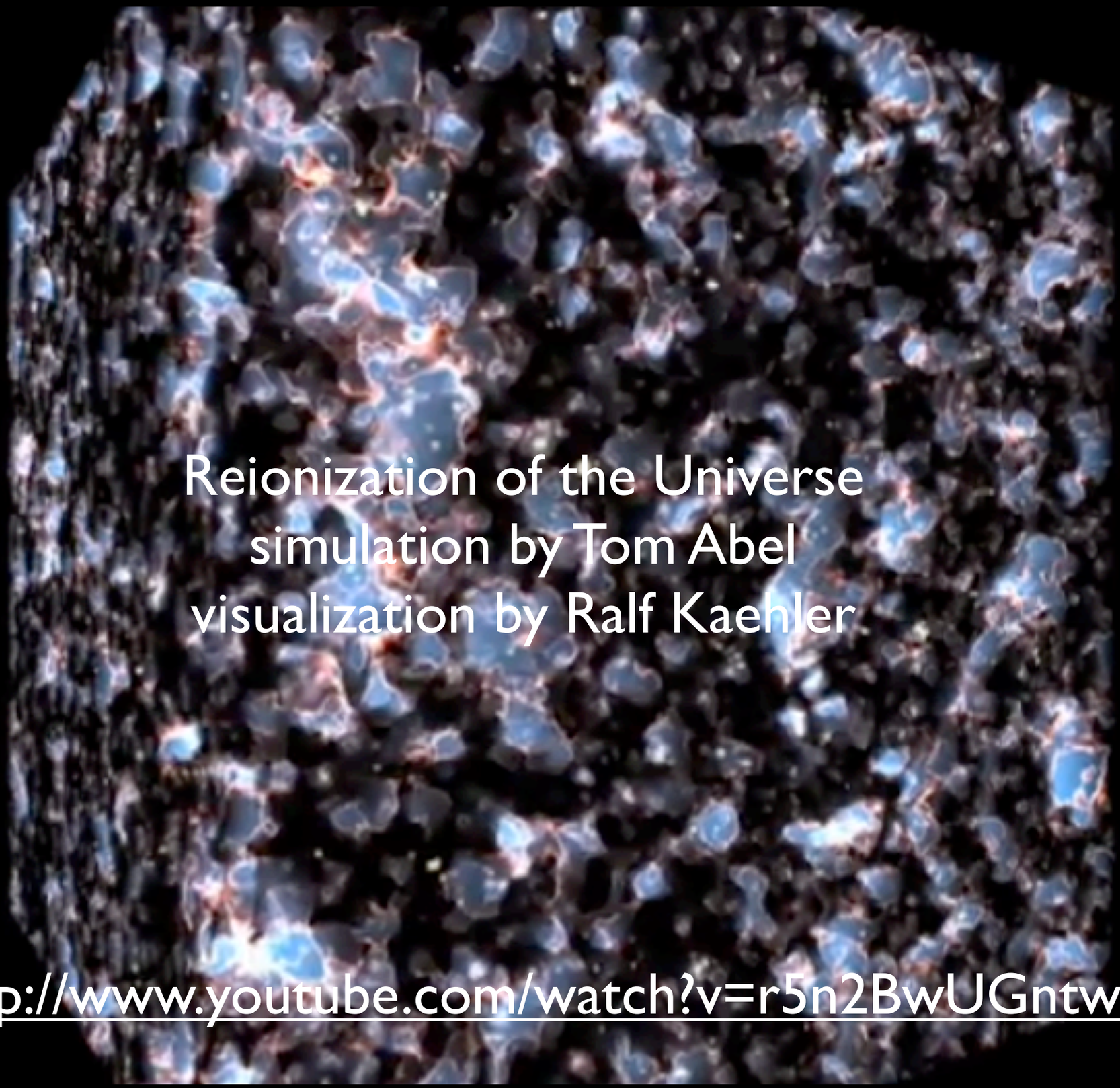
Jonathan R. Pritchard & Abraham Loeb

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



The First Billion Years




Reionization of the Universe
simulation by Tom Abel
visualization by Ralf Kaehler
http://www.youtube.com/watch?v=r5n2BwUGntw

Tom Abel
Ralf Kahler

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
2011	Establish SKA organisation as a legal entity
2008-12	Telescope conceptual design
2006	Short listing of suitable sites
1991	Concept

South Africa
& Western Australia

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

The configuration of the SKA core and remote stations throughout Africa.

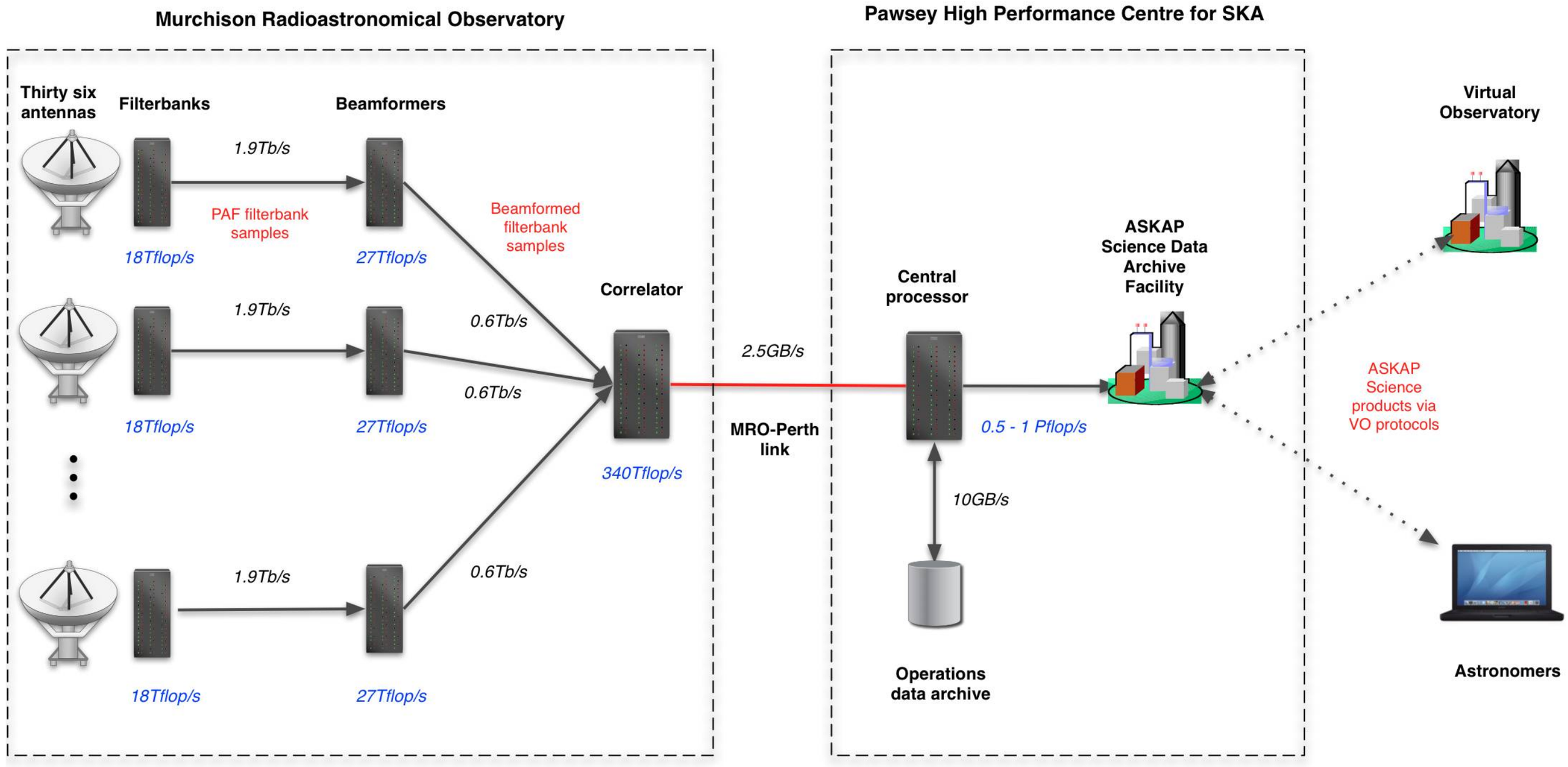
Square Kilometer Array Antenna Types

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) 2012

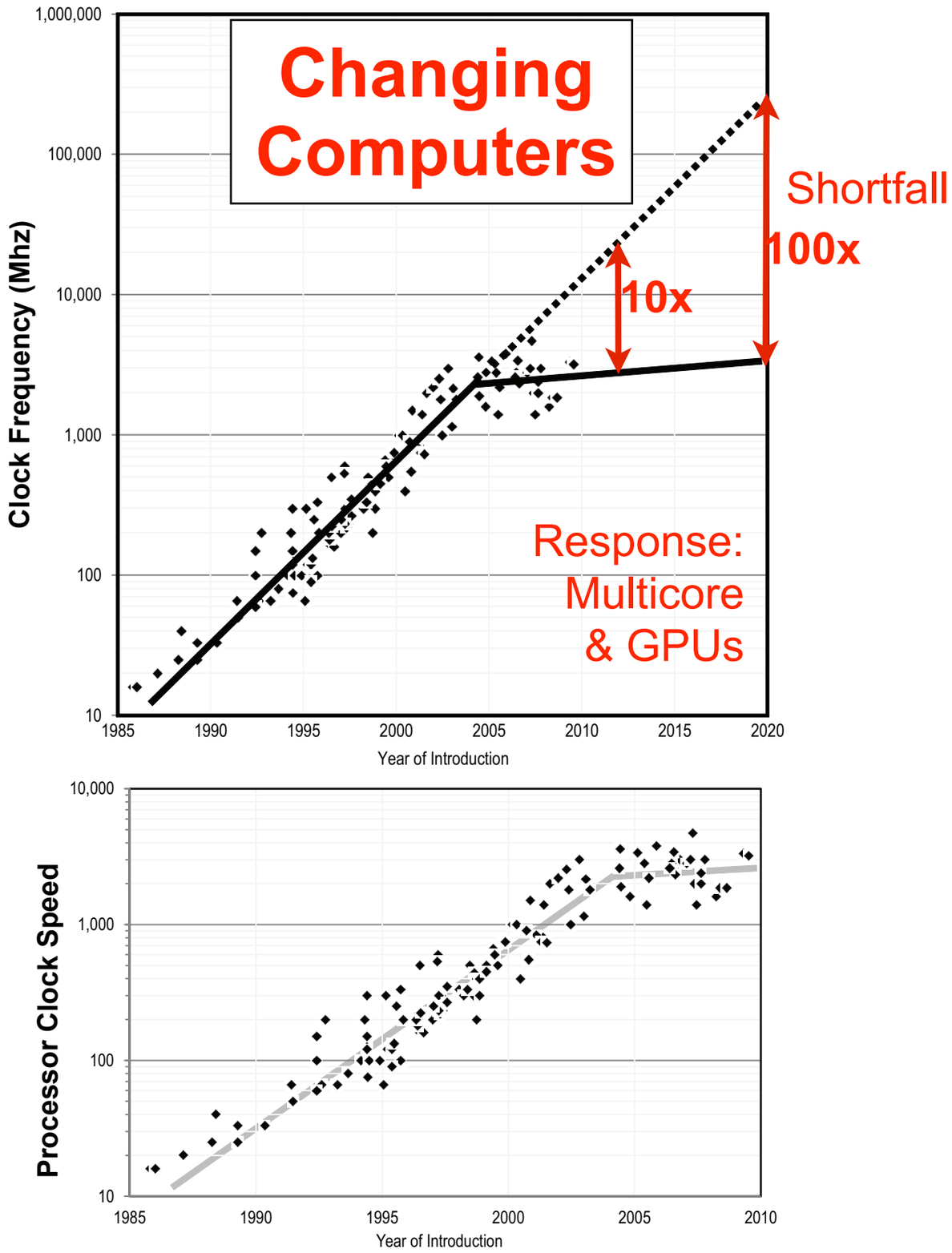
185 Tb of images
25 Tb/year ingest rate
>100 Tb/year retrieval rate

Large Synoptic Survey Telescope (LSST) ~2020

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



The Big Data Future in Astronomy

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

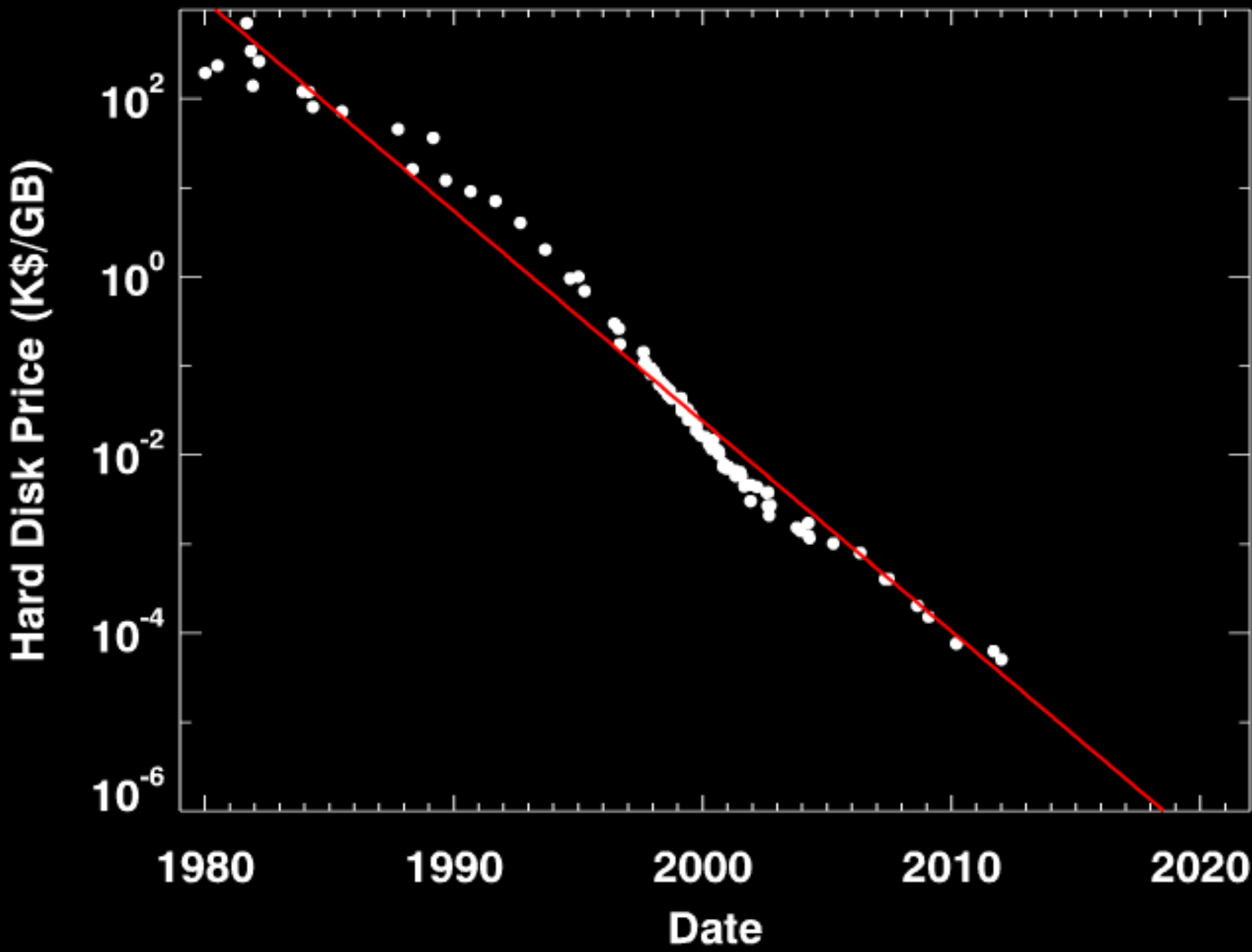
- Ambitious surveys, with massive storage for archives

- Simulation realism - virtual experiments on the universe

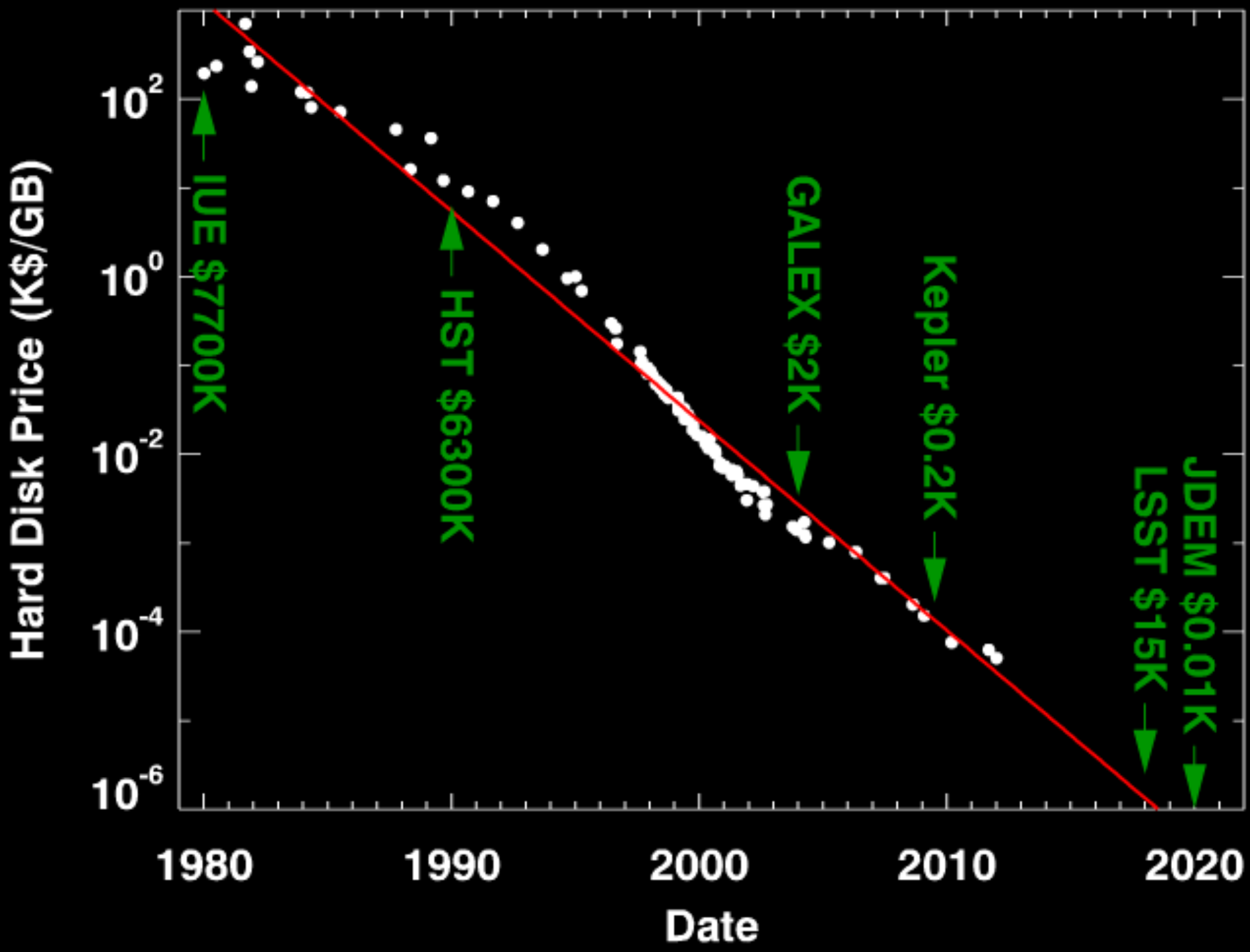
Astronomy is becoming dominated by surveys and simulations

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)

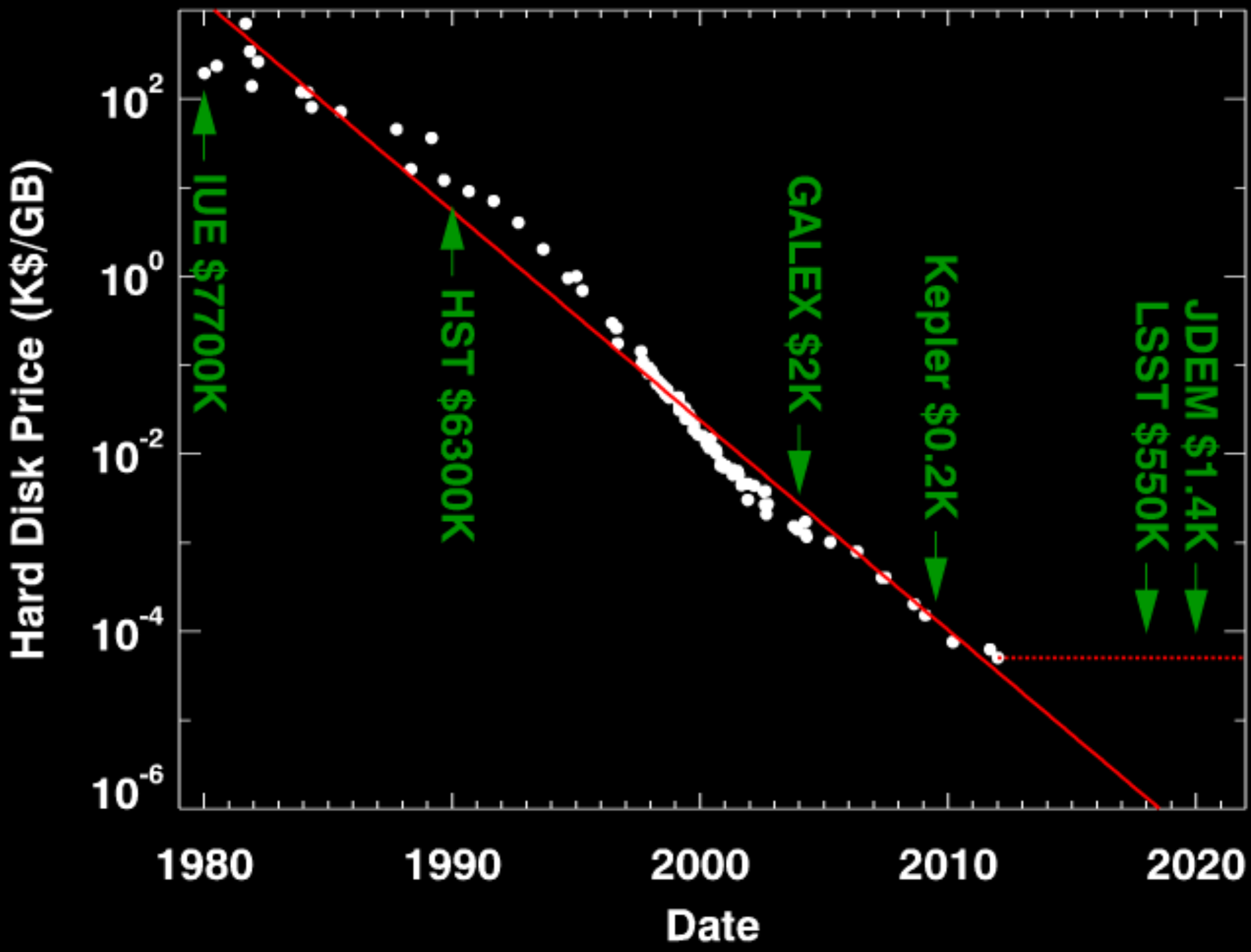
Disk Cost per Gigabyte



Disk Cost per Gigabyte



Disk Cost per Gigabyte



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 and engineers 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

Thanks!

HUAWEI

www.huawei.com