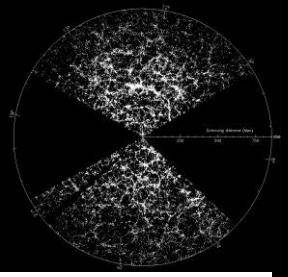
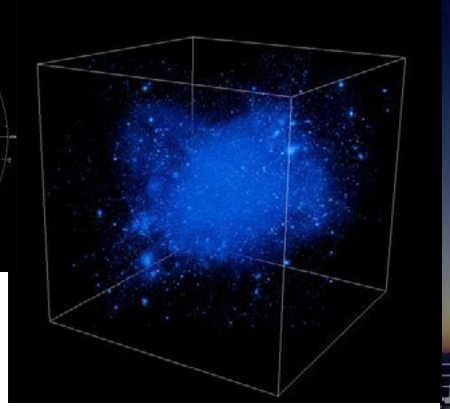
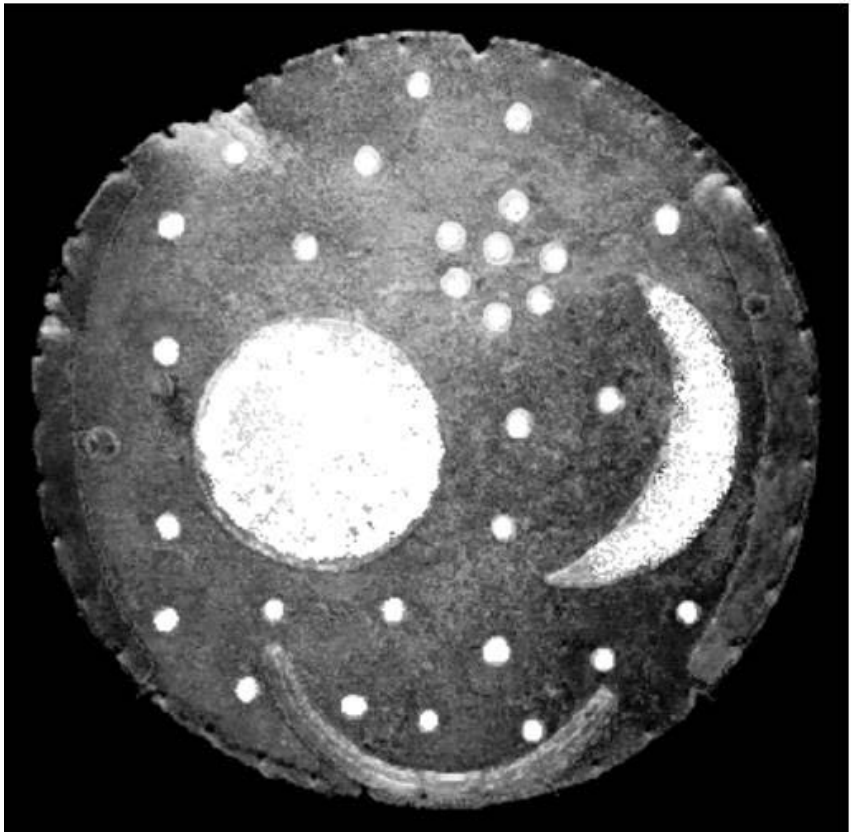



Mapping the Universe

Alexander Szalay
The Johns Hopkins University

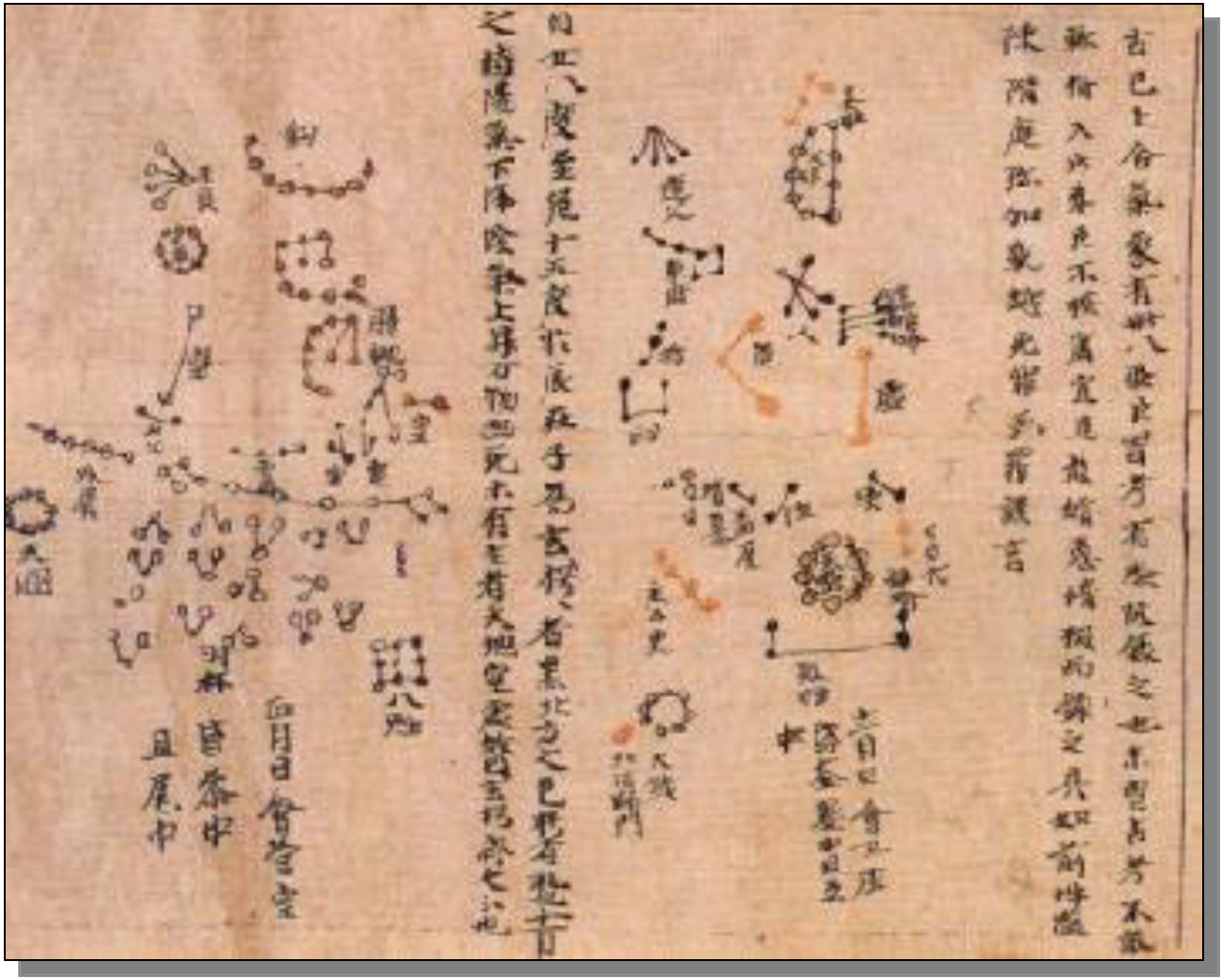
The Oldest Star Charts

30,000 B.C., Orion

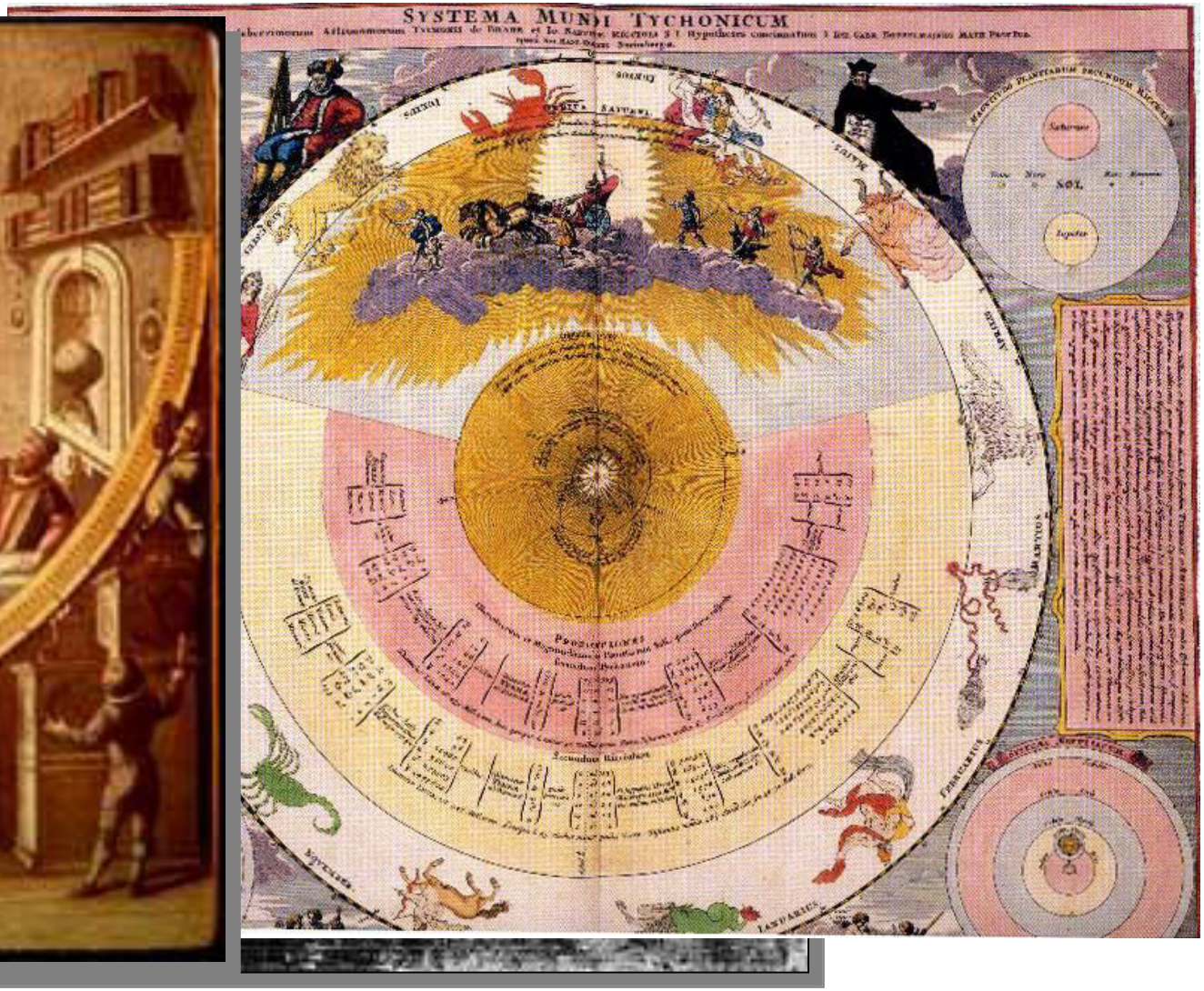


1600 B.C., Pleiades

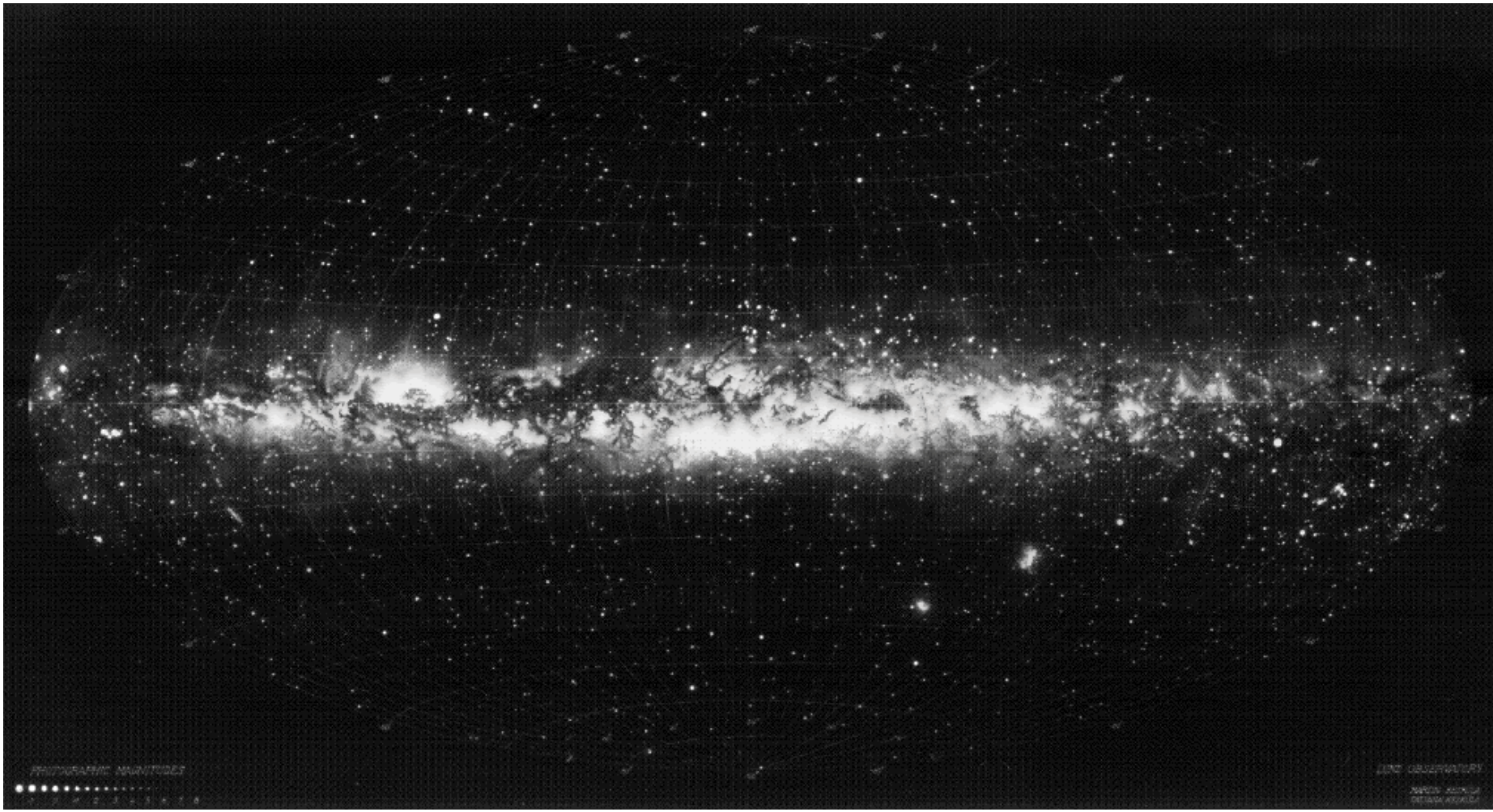
China, 940 A.D.



Tycho Brahe 1600 A.D.

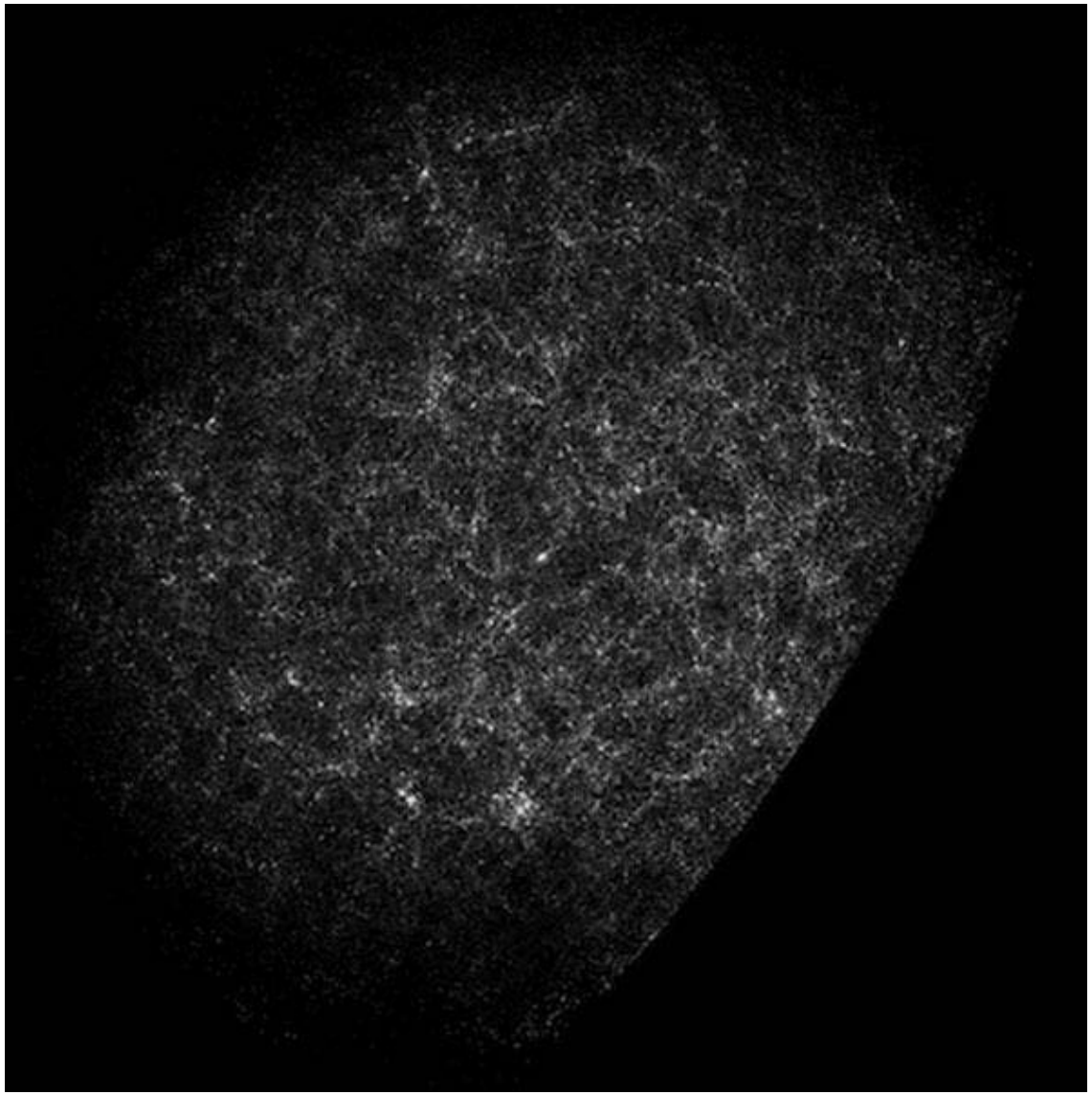


The Lund Map



7000 stars and the Milky Way, hand painted
Supervised by Knut Lundmark, 1950

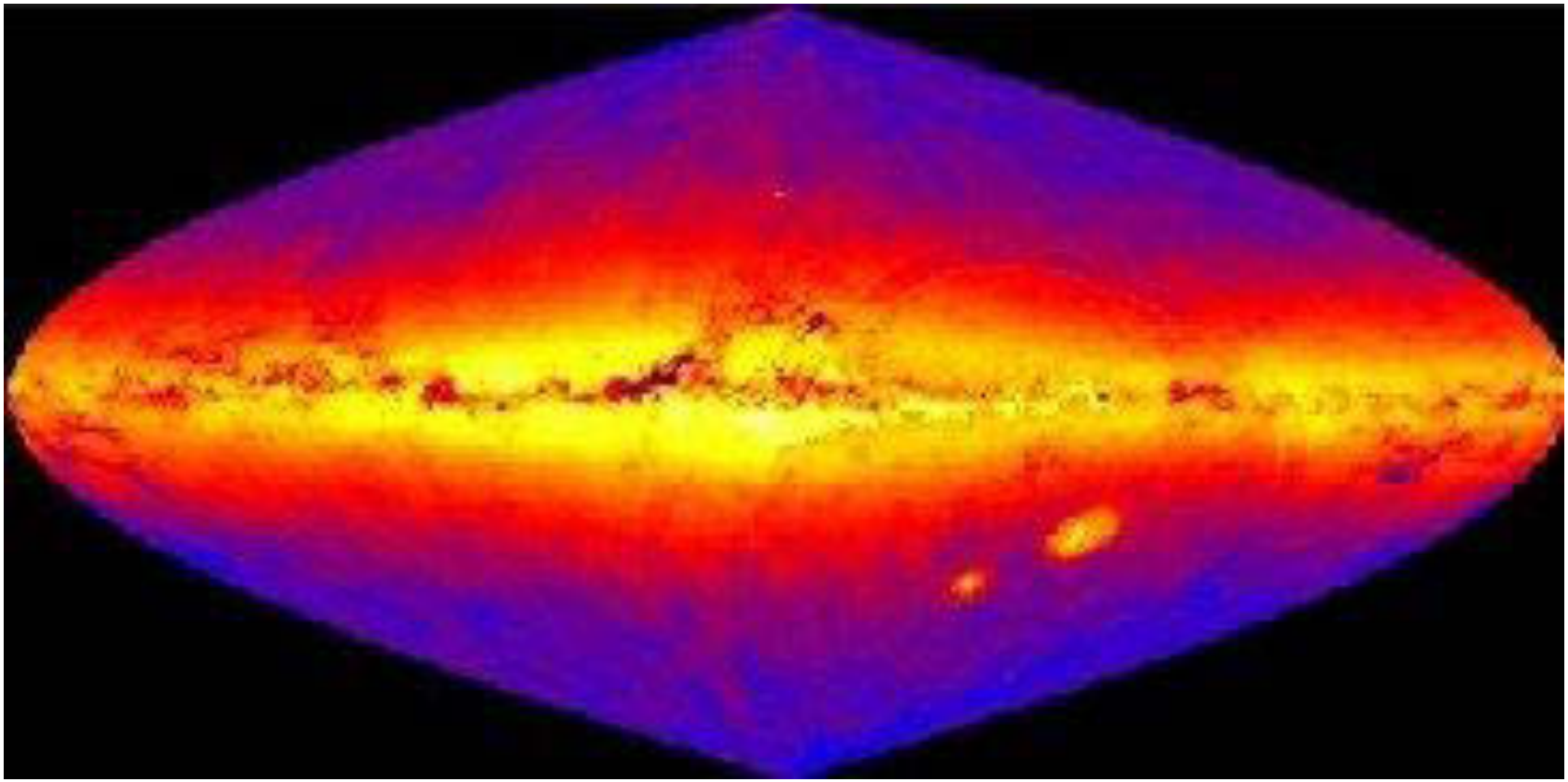
The Lick Catalog



Shane-Wirtanen 1970s, Digitized by Groth and Peebles 1977

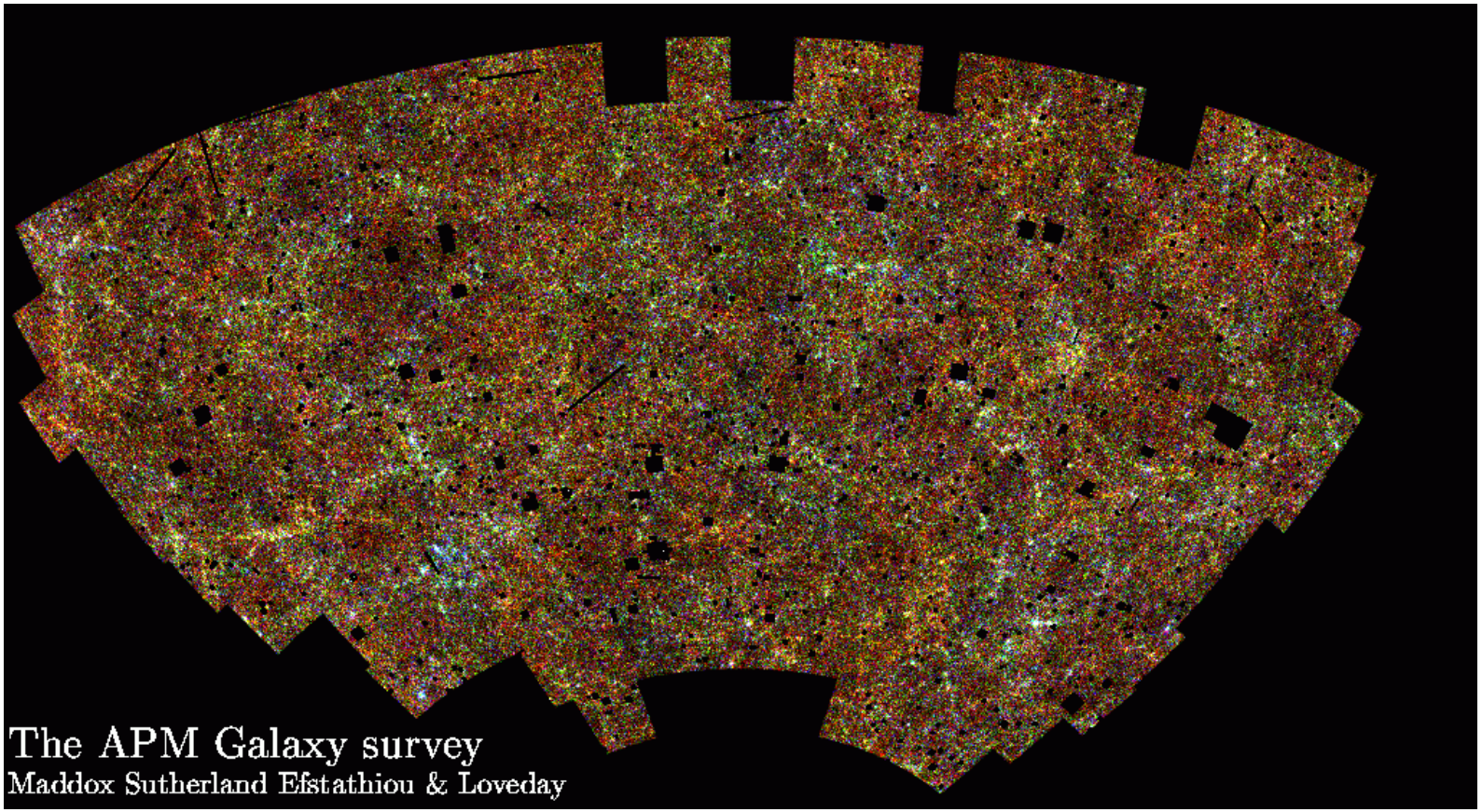
1 million galaxies, counted visually

USNO



From digitized Palomar and UK Schmidt plates1980-90

The APM Galaxy Survey



3M galaxy positions, digitized from UKST plates
Maddox, Efstathiou, Sutherland and Loveday (1990)

The Expanding Universe

- Hubble's law:

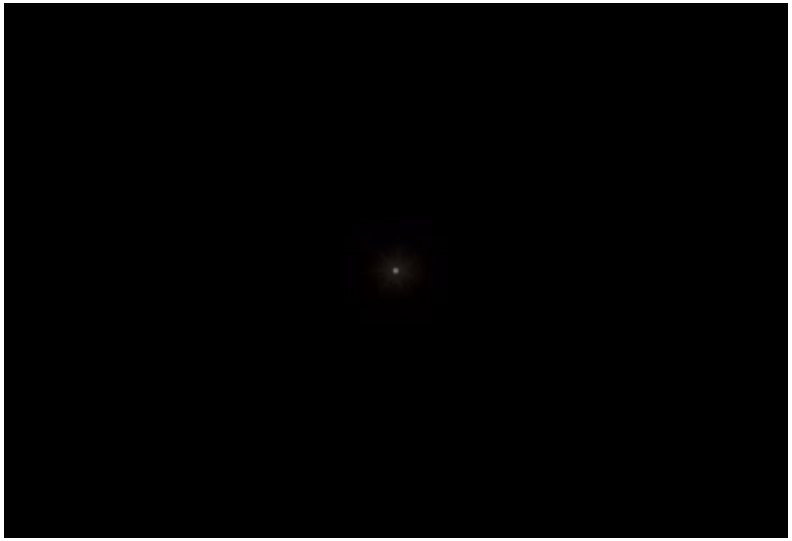
$$v = H_0 r$$

- Uniform expansion of comoving space:

$$r = a(t)x$$

- Redshift: wavelength of light is expanding as well

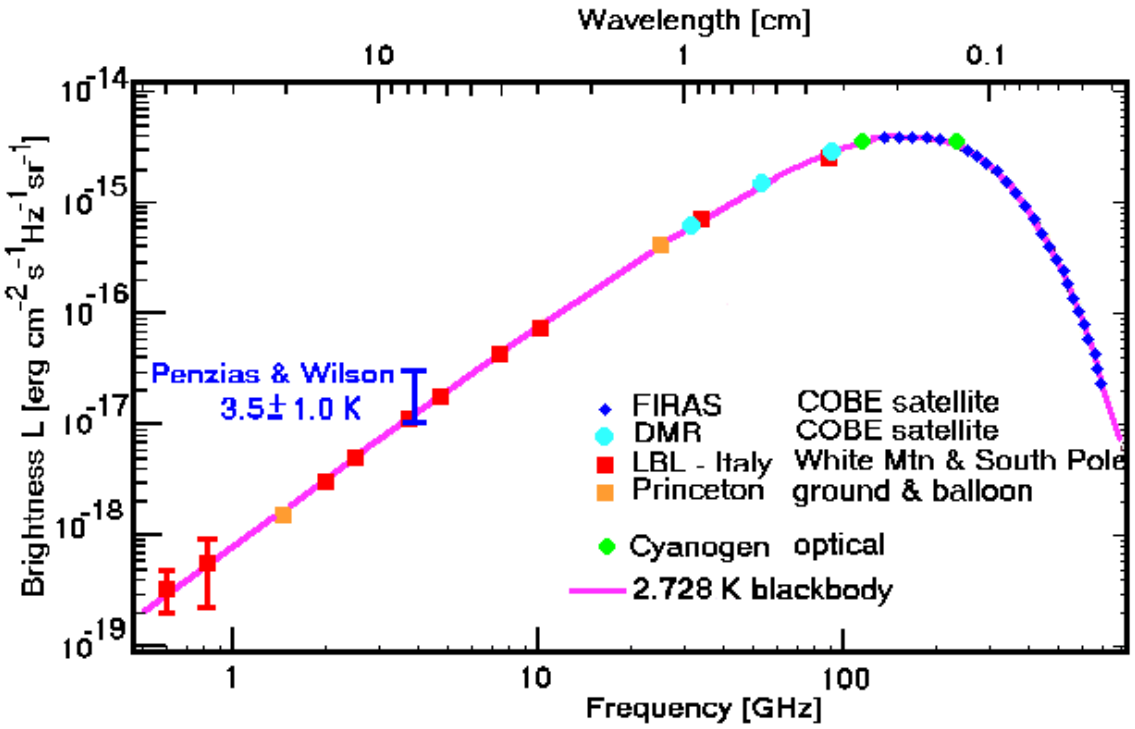
$$\frac{\lambda_{obs}}{\lambda_{em}} = 1 + z = \frac{1}{a(t)}$$



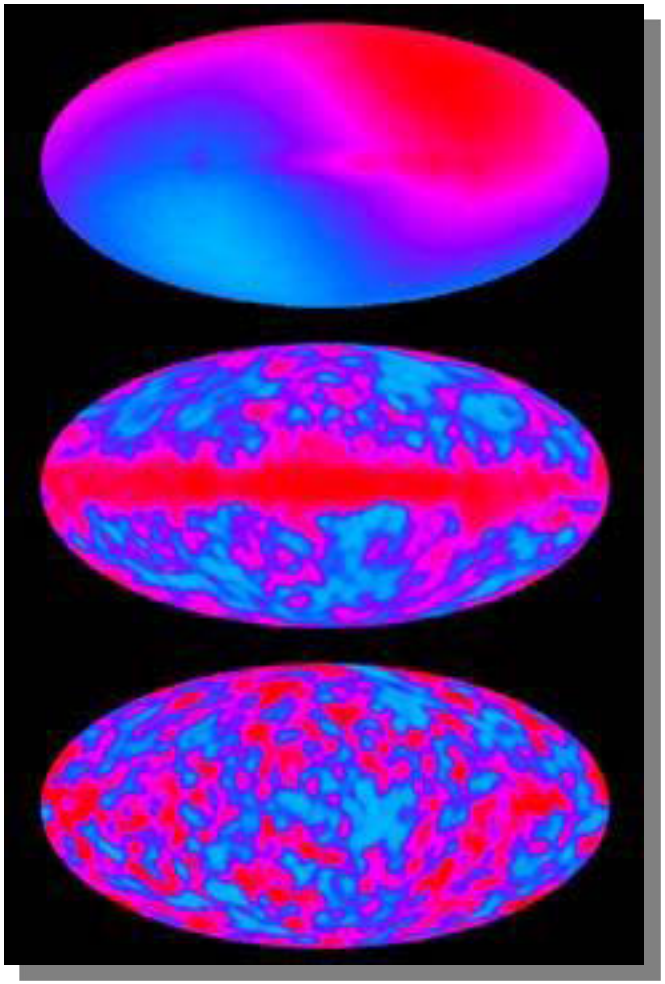
The Microwave Background

- Uniform black-body radiation, detected by Penzias and Wilson
- Temperature T_0=2.725°K today
- Earlier:

$$T = T_0(1+z)$$



COBE, 1990



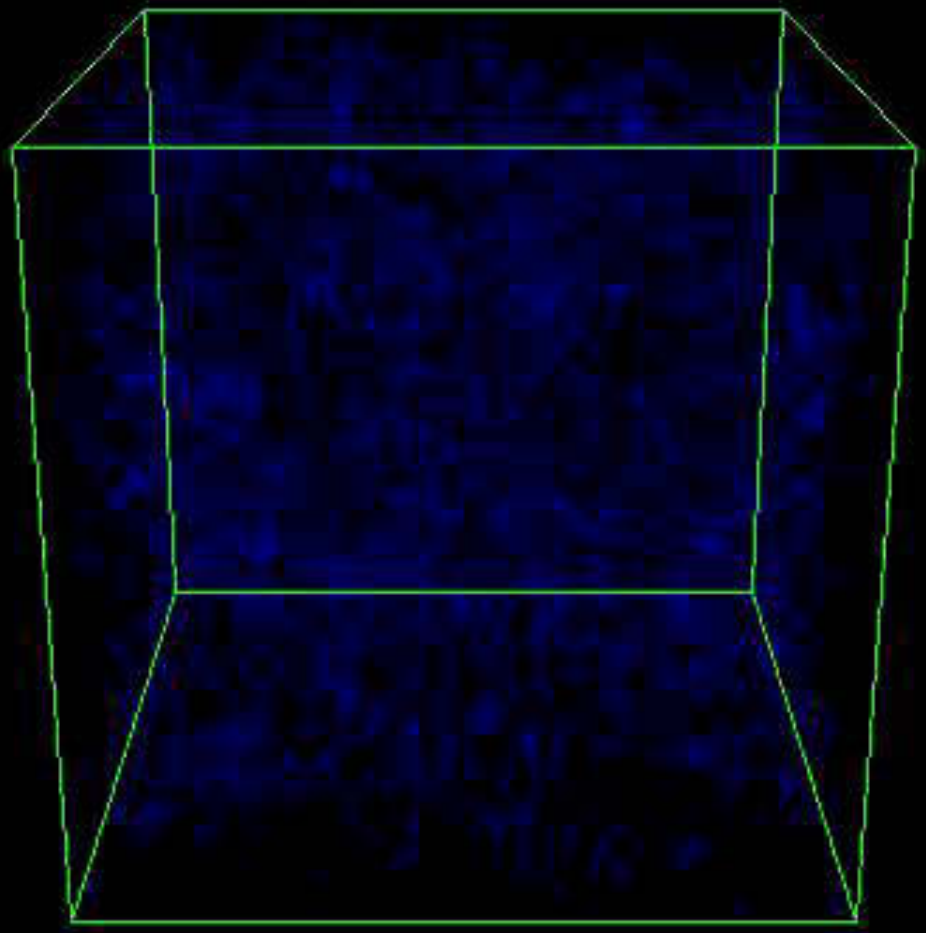
Detection of primordial fluctuations in the Cosmic Microwave Background

There are small ripples on top of the smooth background, leading to the observed large-scale structure (LSS) in the Universe

Evolution of Structure in a Low Omega Universe

200 Mpc across

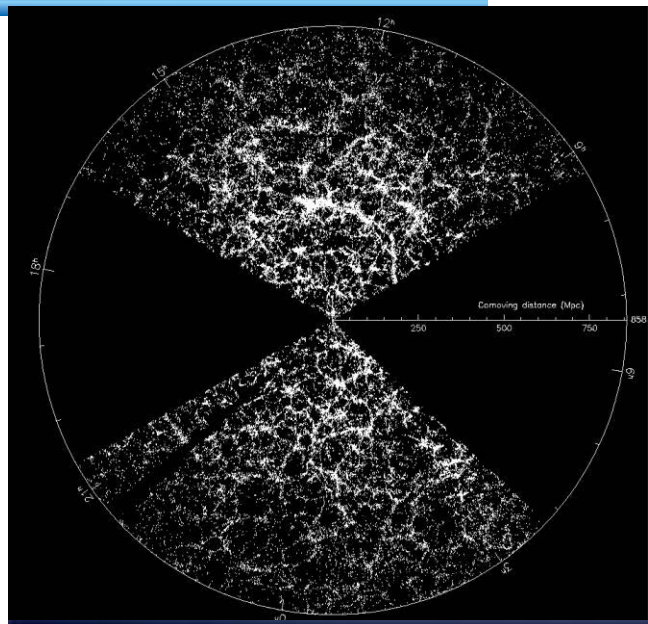
Time = 0.05 Gyr



University of Washington Astronomy

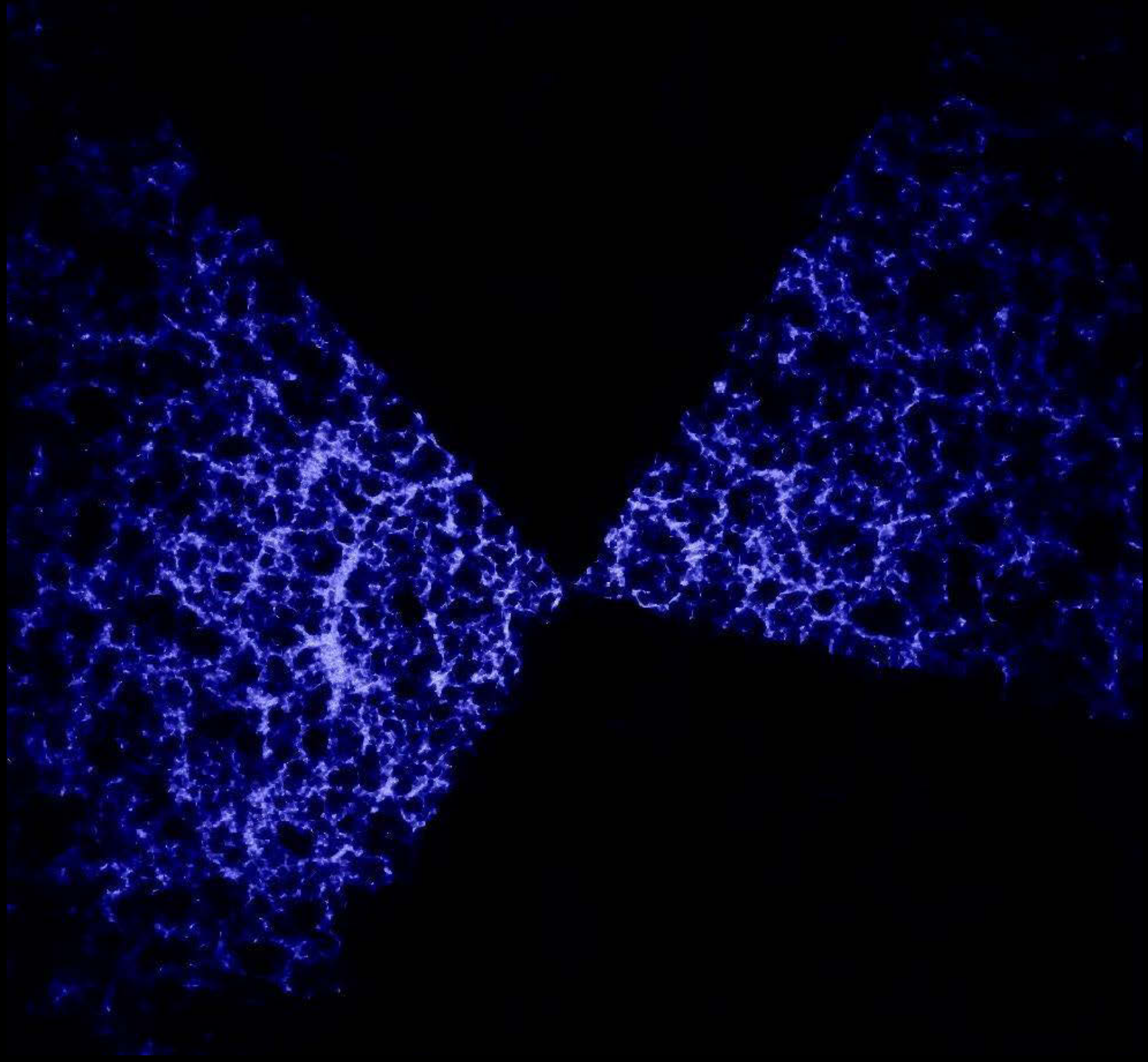
Stadel 1998

Sloan Digital Sky Survey



“The Cosmic Genome Project”

- Started in 1992, finished in 2008
- Data is public
  - 2.5 Terapixels of images => 5 Tpx of sky
  - 10 TB of raw data => 100TB processed
  - 0.5 TB catalogs => 35TB in the end
- Database and spectrograph built at JHU (SkyServer)
- Now SDSS-3, data served from JHU



Features of the SDSS

Special 2.5m telescope, at Apache Point, NM

3 degree field of view

Zero distortion focal plane

Two surveys in one

Photometric survey in 5 bands

Spectroscopic redshift survey

Automated data reduction

Over 120 man-years of development

(Fermilab + collaboration scientists)

Very high data volume

Expect over 40 TB of raw data

About 2 TB processed catalogs

Data made available to the public

Apache Point Observatory

Located in New Mexico,
near White Sands National Monument

The Telescope

Special 2.5m telescope

3 degree field of view

Wind screen moved separately

The Photometric Survey

Continuous data rate of 8 Mbytes/sec

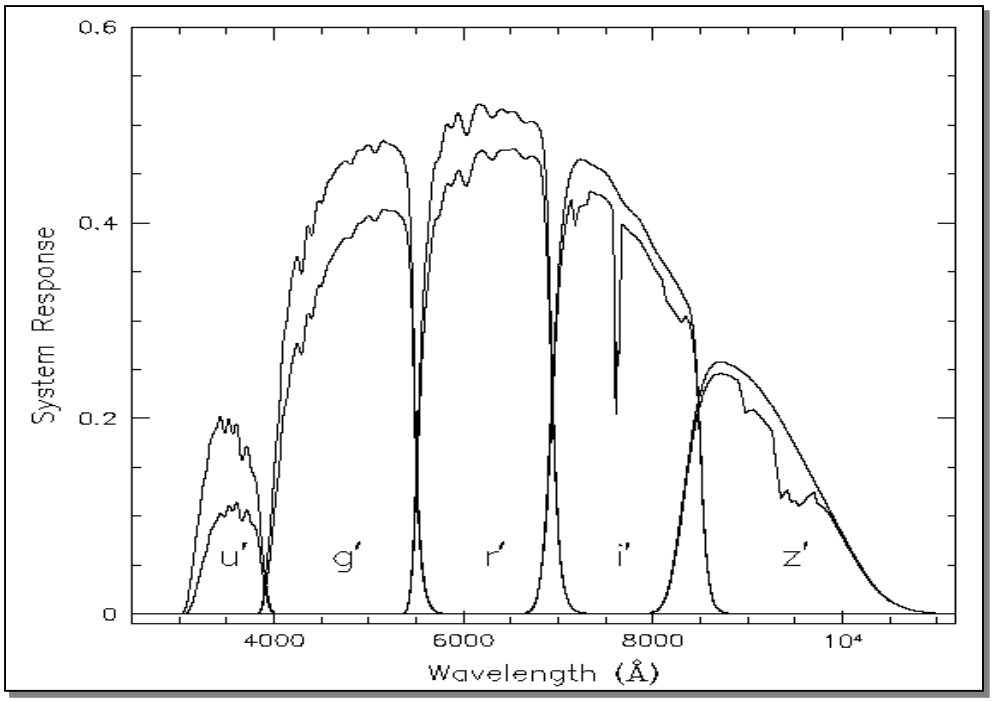
Northern Galactic Cap

drift scan of 10,000 square degrees
5 broad-band filters
exposure time: 55 sec
pixel size: 0.4 arcsec
astrometry: 60 mas
calibration: 2% at r'=19.8
done only in best seeing
(20 nights/year)

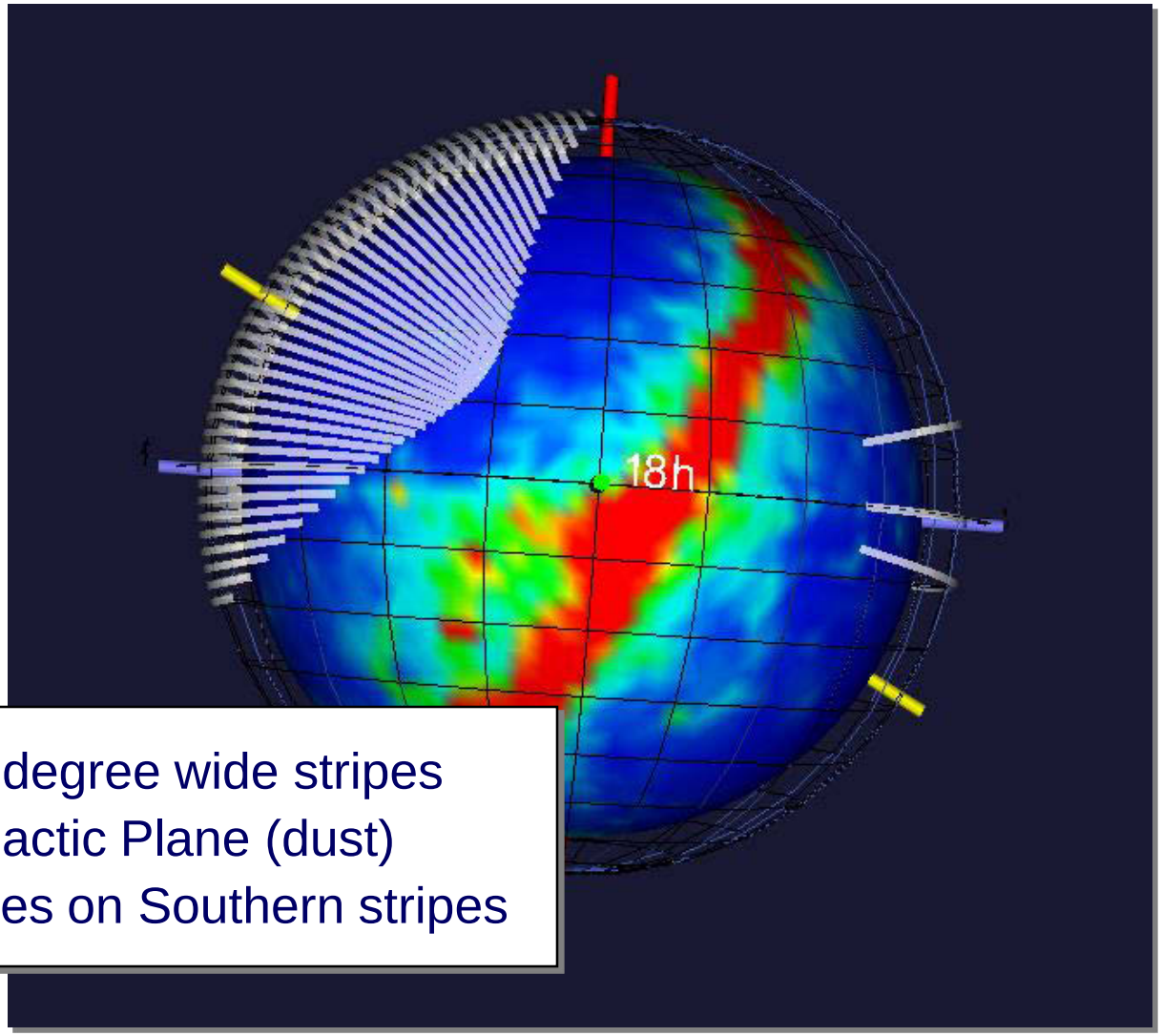
Southern Galactic Cap

multiple scans (> 30 times)
of the same stripe

u'	*g'*	*r'*	*i'*	*z'*
22.3	*23.3*	*23.1*	*22.3*	*20.8*

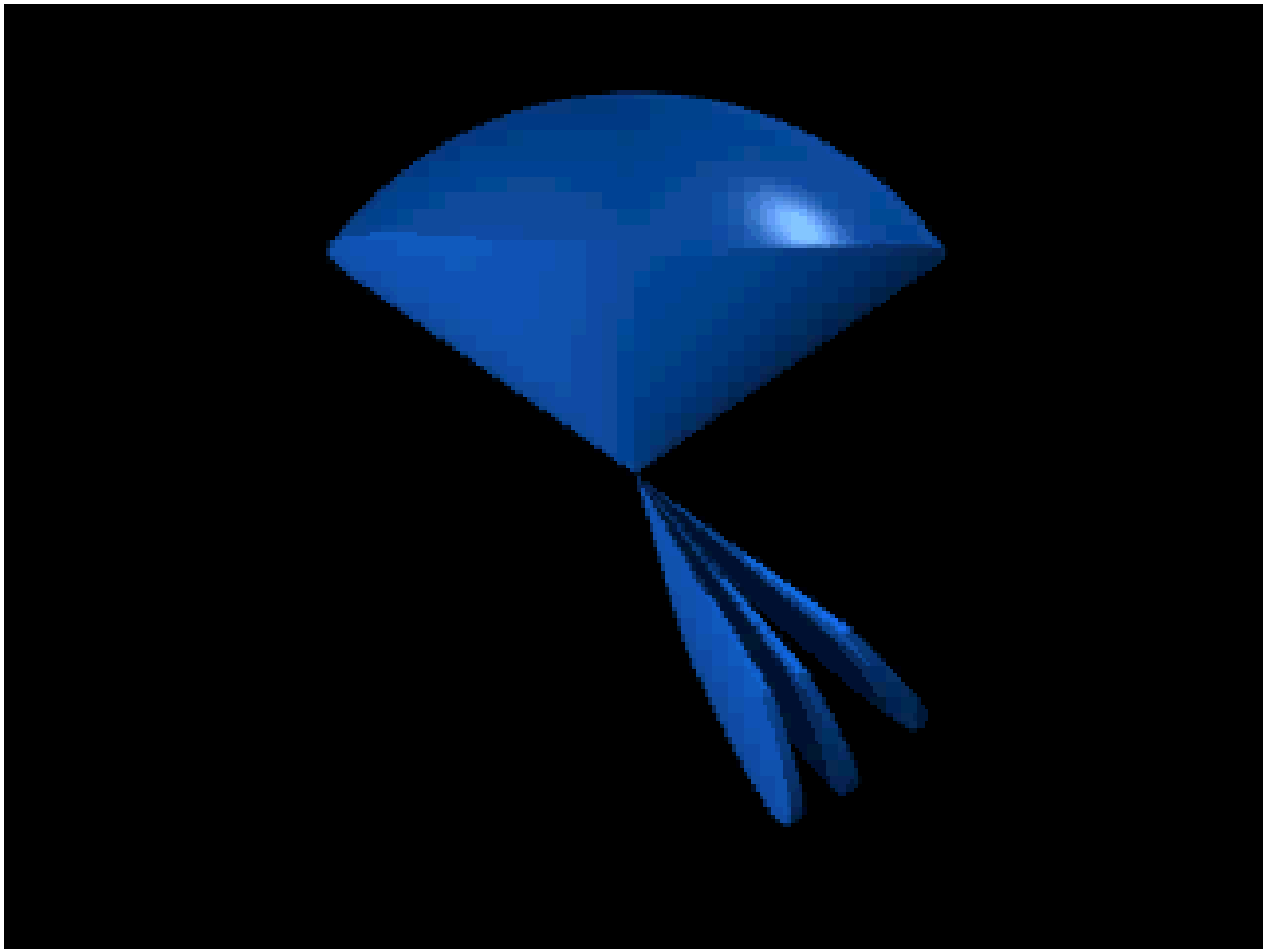


Survey Strategy



Overlapping 2.5 degree wide stripes
Avoiding the Galactic Plane (dust)
Multiple exposures on Southern stripes

The Footprint of the Survey



The Spectroscopic Survey

SDSS Redshift Survey

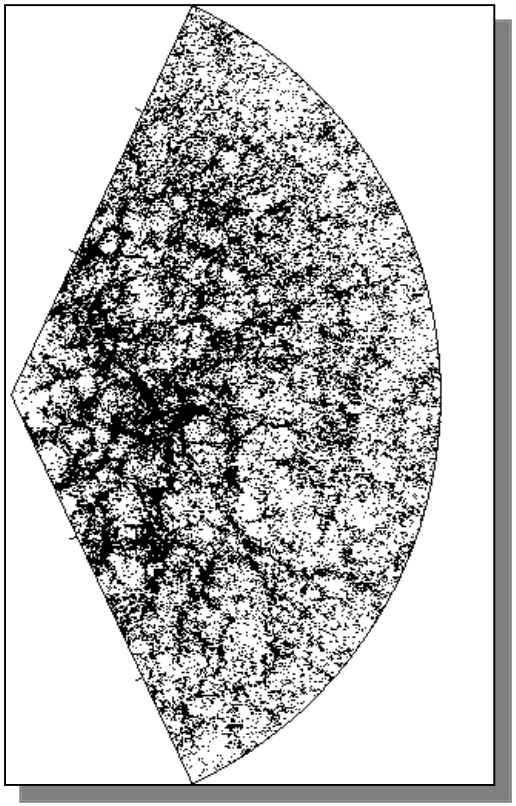
- *1 million galaxies*
  - *900,000 r' limited*
  - *100,000 red galaxies*
    - *volume limited to z=0.45*
- *100,000 quasars*
- *100,000 stars*

Two high throughput spectrographs

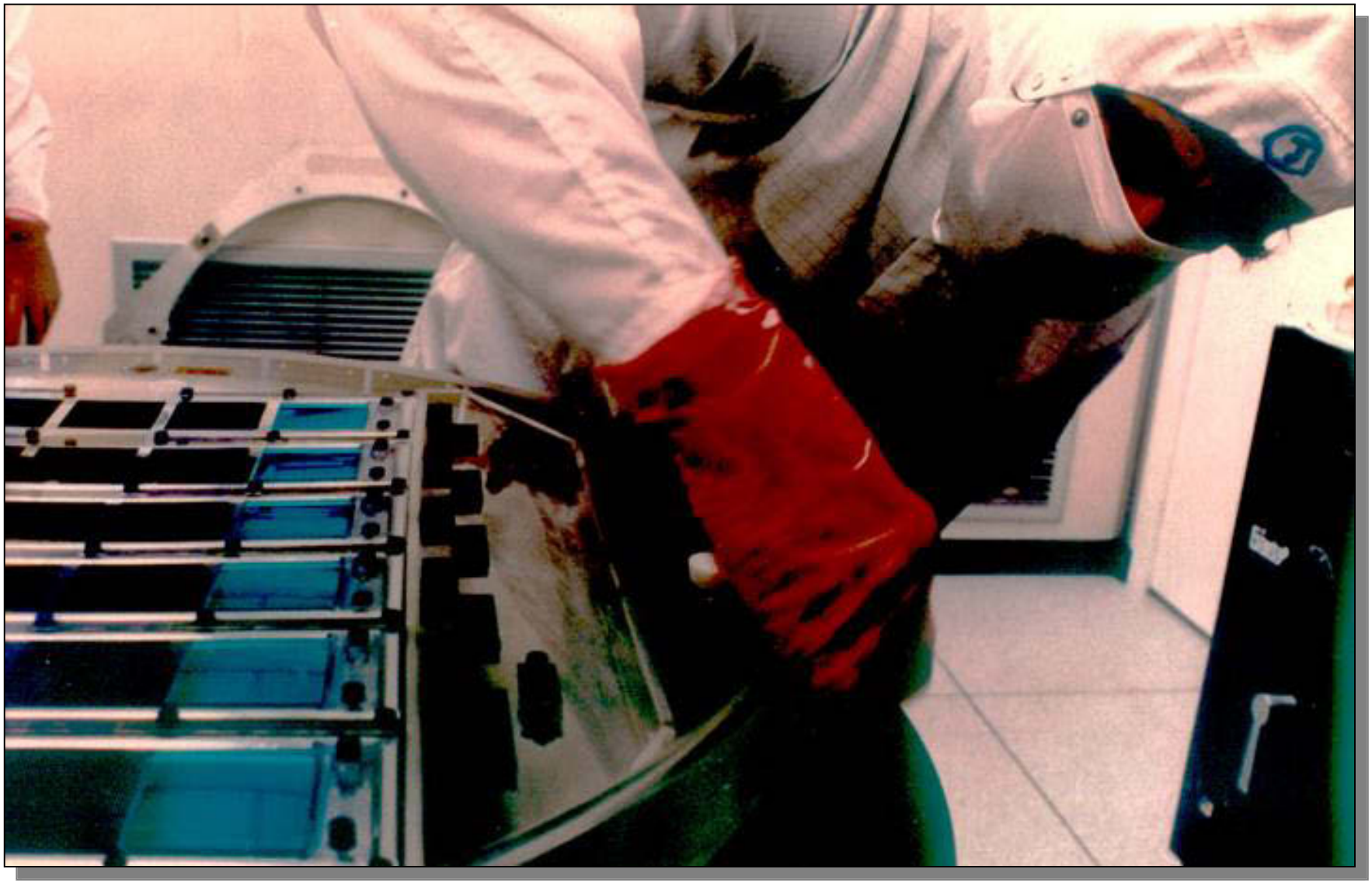
- *spectral range 3900-9200 Å*
- *640 spectra simultaneously*
- *R=2000 resolution, 1.3 Å*

Features

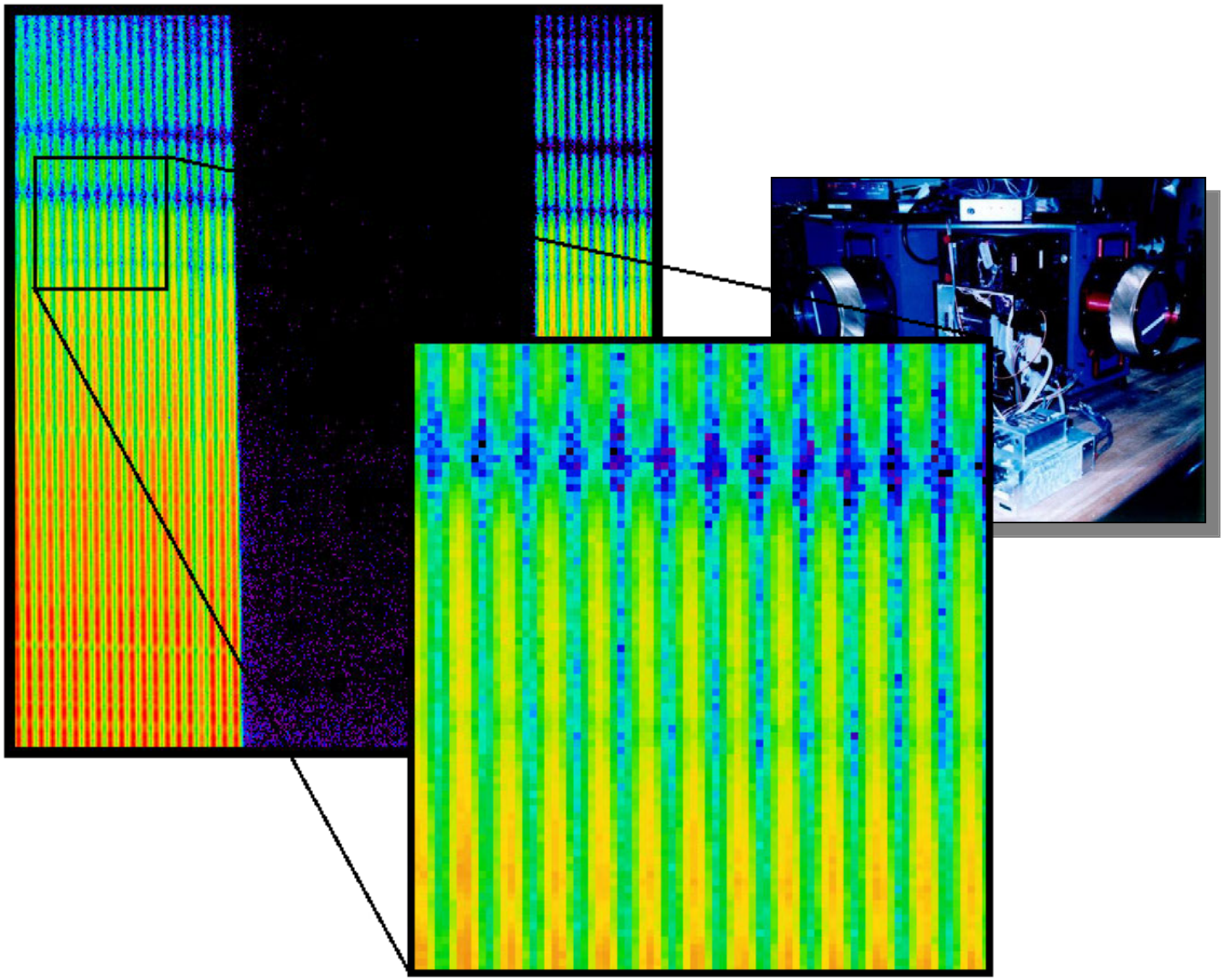
- *Automated reduction of spectra*
- *Very high sampling density and completeness*
- *Objects in other catalogs also targeted*



The Mosaic Camera

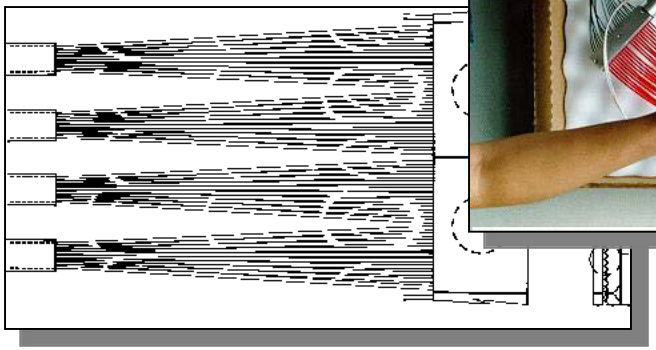


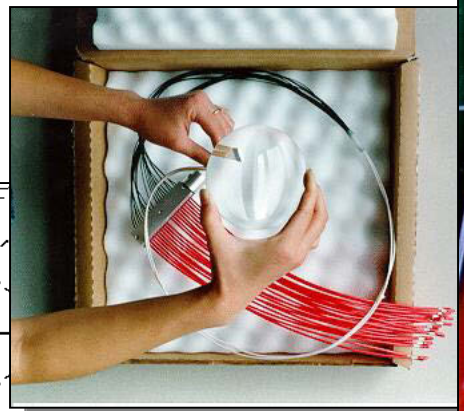
The Spectrographs

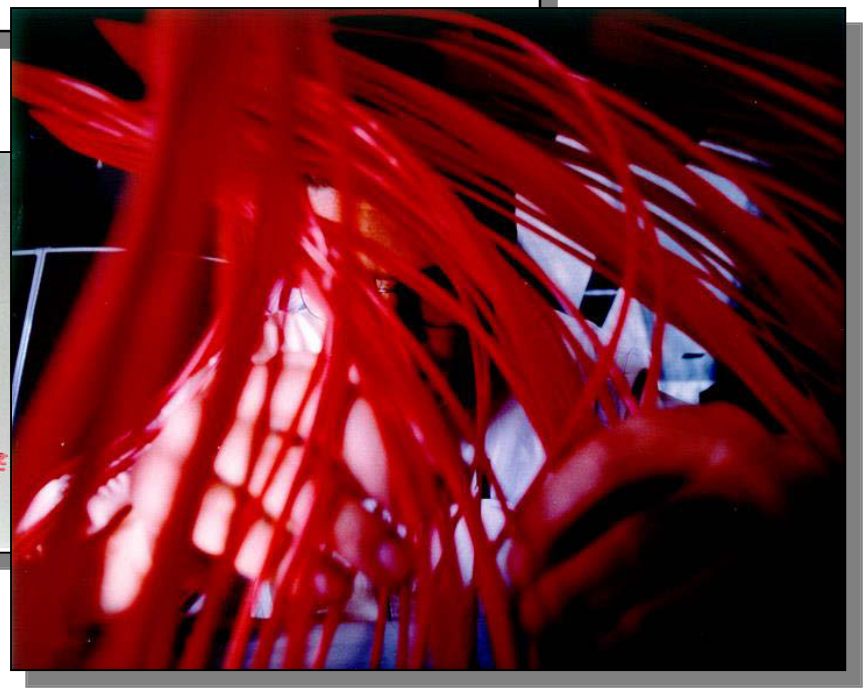


The Fiber Feed System

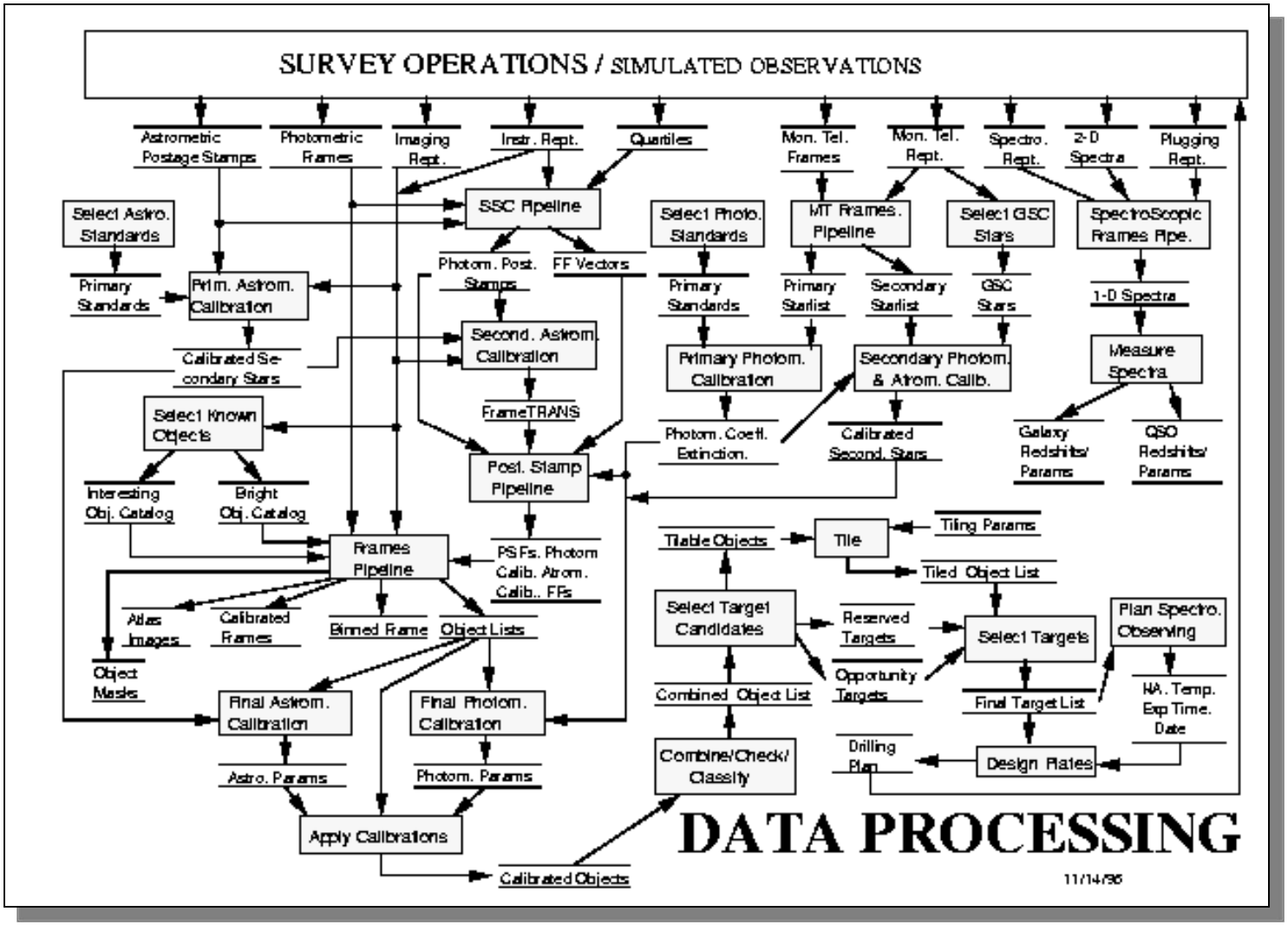
Galaxy images are captured by optical fibers
lined up on the spectrograph slit
Manually plugged during the day into Al plugboards
640 fibers in each bundle







Data Processing Pipelines



First Light Images

Telescope

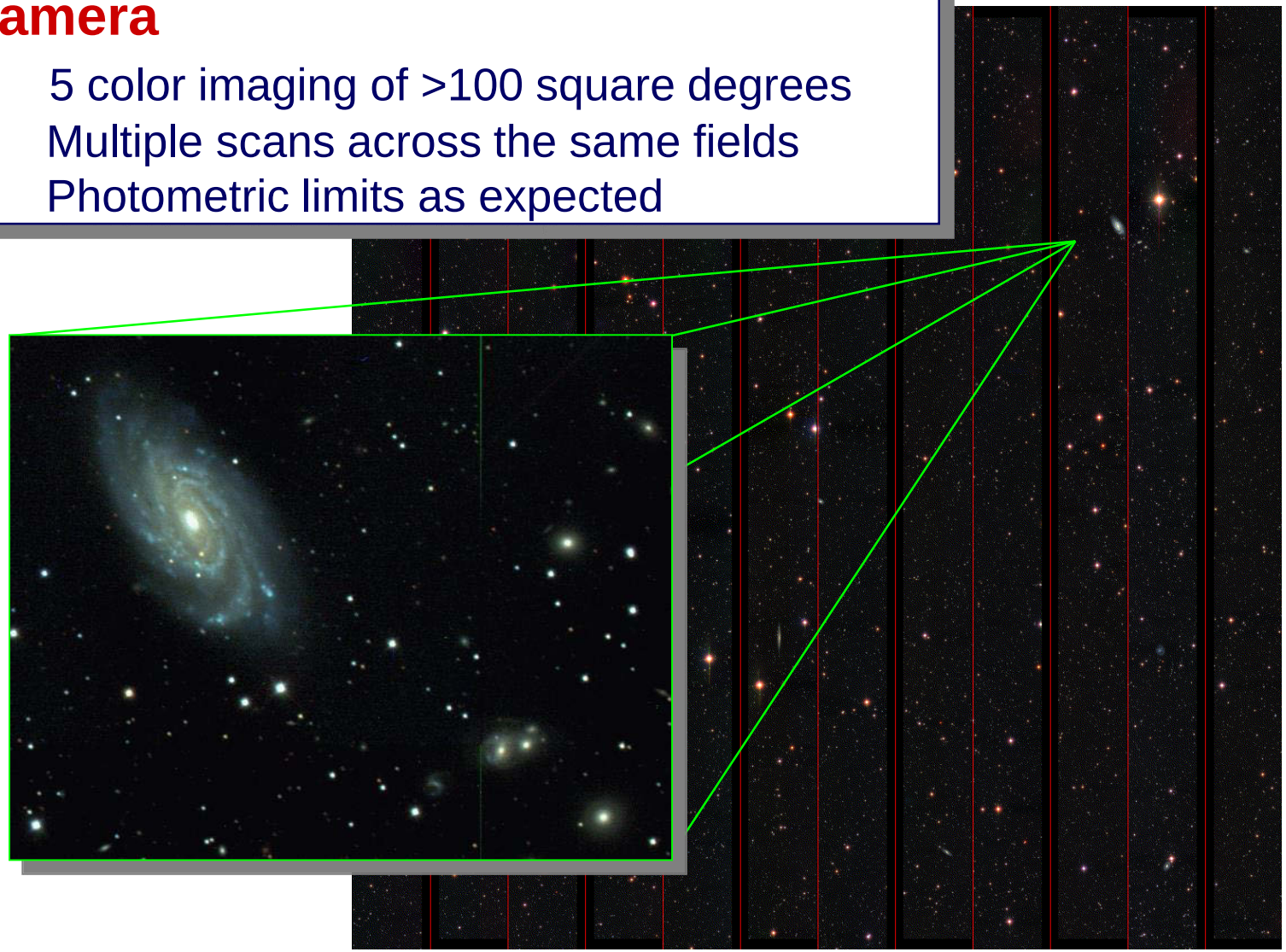
First light May 9th 1998

Equatorial scans

The First Stripes

Camera

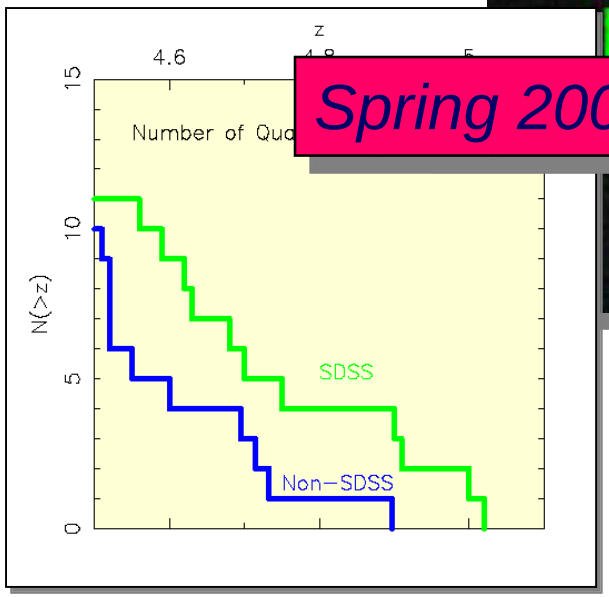
5 color imaging of >100 square degrees
Multiple scans across the same fields
Photometric limits as expected

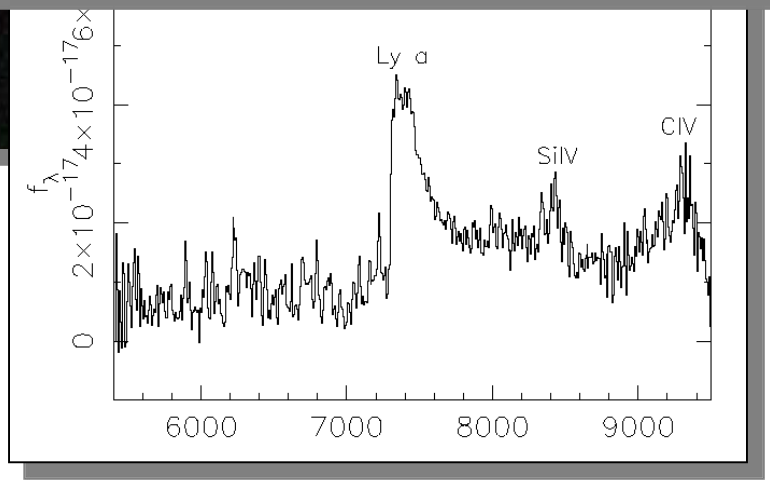


The First Quasars

The four highest redshift quasars at the time have been found in the first SDSS test data !

Spring 2000: a 5.3 and a 5.8 QSO found!

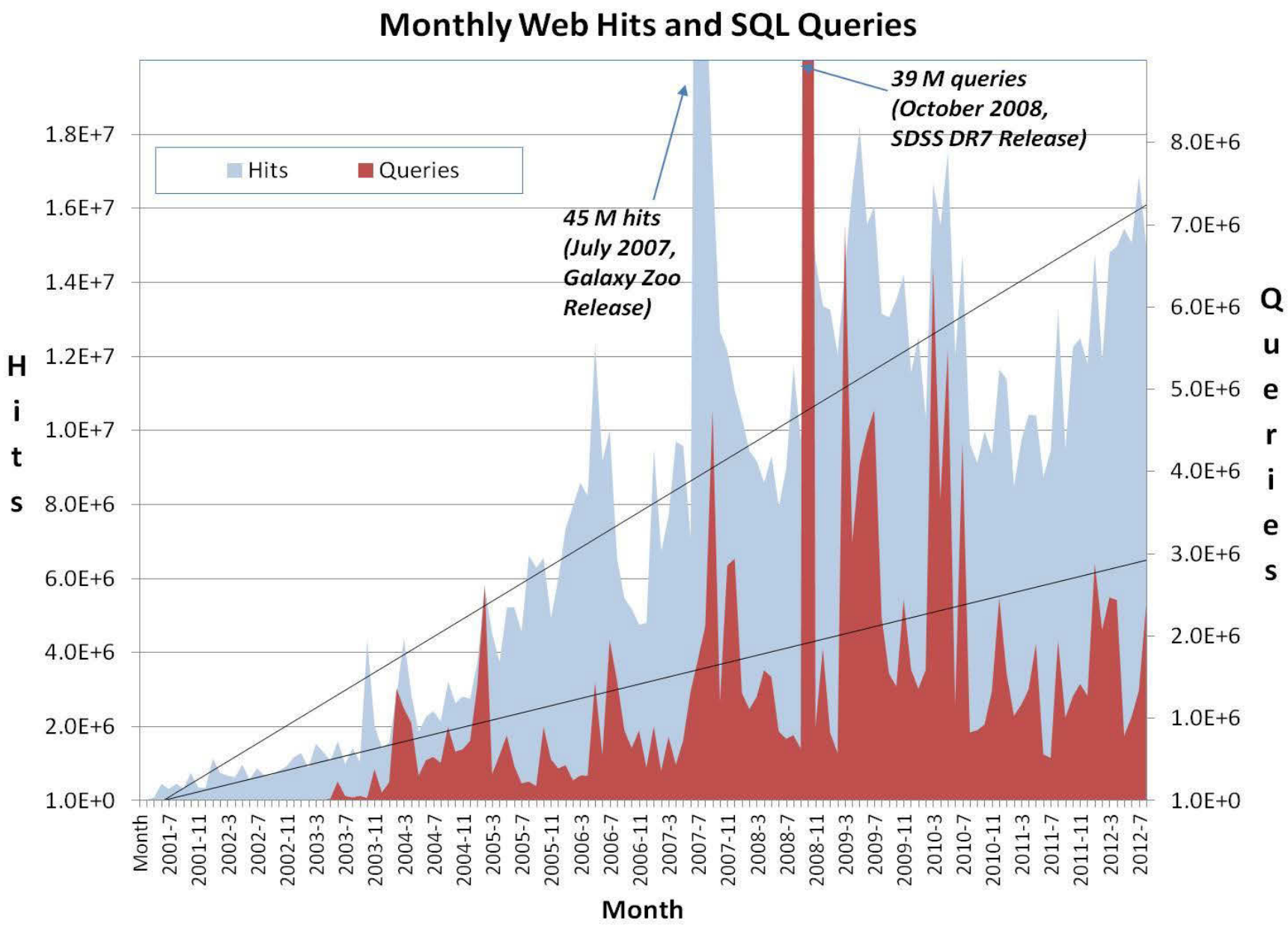




Skyserver

Prototype in 21st Century data access

- *1.2B web hits in 12 years*
- *200M external SQL queries*
- *4,000,000 distinct users vs. 15,000 astronomers*
- *The emergence of the "Internet Scientist"*
- *The world's most used astronomy facility today*
- *Collaborative server-side analysis done by 7K astronomers*


Monthly Web Hits and SQL Queries
Hits
Queries
45 M hits
(July 2007,
Galaxy Zoo
Release)
39 M queries
(October 2008,
SDSS DR7 Release)
1.8E+7
1.6E+7
1.4E+7
1.2E+7
1.0E+7
8.0E+6
6.0E+6
4.0E+6
2.0E+6
1.0E+0
8.0E+6
7.0E+6
6.0E+6
5.0E+6
4.0E+6
3.0E+6
2.0E+6
1.0E+6
1.0E+0
Hits
Queries
Month
Month
2001-7
2001-11
2002-3
2002-7
2002-11
2003-3
2003-7
2003-11
2004-3
2004-7
2004-11
2005-3
2005-7
2005-11
2006-3
2006-7
2006-11
2007-3
2007-7
2007-11
2008-3
2008-7
2008-11
2009-3
2009-7
2009-11
2010-3
2010-7
2010-11
2011-3
2011-7
2011-11
2012-3
2012-7

Impact of Sky Surveys

Astronomy

Sloan Digital Sky Survey tops astronomy citation list

NASA's Sloan Digital Sky Survey (SDSS) is the most significant astronomical facility, according to an analysis of the 200 most cited papers in astronomy published in 2006. The survey, carried out by Juan Madrid from McMaster University in Canada and Duccio Macchetto from the Space Telescope Science Institute in Baltimore, puts NASA's Swift satellite in second place, with the Hubble Space Telescope in third (arXiv:0901.4552).

Madrid and Macchetto carried out their analysis by looking at the top 200 papers using NASA's Astrophysics Data System (ADS), which charts how many times each paper has been cited by other research papers. If a paper contains data taken only from one observatory or satellite, then that facility is awarded all the citations given to that article. However, if a paper is judged to contain data from different facilities – say half from SDSS and half from Swift – then both facilities are given 50% of the citations that paper received.

Top 10 telescopes

Rank	Telescope	Citations	Ranking in 2004
1	Sloan Digital Sky Survey	1892	1
2	Swift	1523	N/A
3	Hubble Space Telescope	1078	3
4	European Southern Observatory	813	2
5	Keck	572	5
6	Canada–France–Hawaii Telescope	521	N/A
7	Spitzer	469	N/A
8	Chandra	381	7
9	Boomerang	376	N/A
10	High Energy Stereoscopic System	297	N/A

The researchers then totted up all the citations and produced a top 10 ranking (see table). Way out in front with 1892 citations is the SDSS, which has been running since 2000 and uses the 2.5 m telescope at Apache Point in New Mexico to obtain images of more than a quarter of the sky. NASA's Swift satellite, which studies gamma-ray bursts, is second with 1523 citations, while the Hubble Space Telescope (1078 citations) is third.

Although the 200 most cited papers make up only 0.2% of the references indexed by the ADS for papers published in 2006, those 200 papers account for 9.5% of the citations. Madrid and Macchetto also ignored theory papers on the basis that they do not directly use any telescope data. A similar study of papers published in 2004 also puts SDSS top with 1843 citations. This time, though, the European Southern Observatory, which has telescopes in Chile, comes second with 1365 citations and the Hubble Space Telescope takes third spot with 1124 citations.

Michael Banks

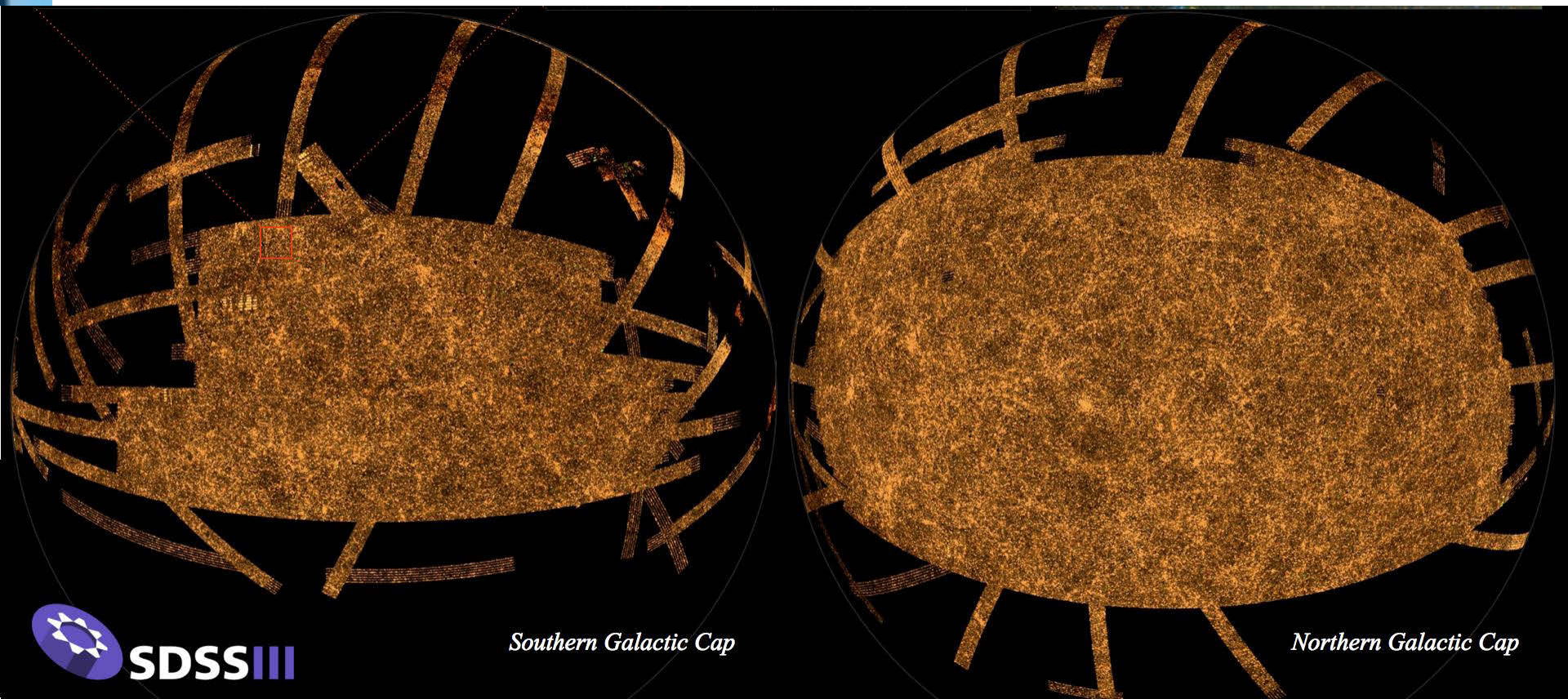
GalaxyZoo

- 40 million visual galaxy classifications by the public
- Good publicity (CNN, Times, Washington Post, BBC)
- 300,000 people participating, blogs, poems…
- Original discoveries by the public
  (Voorwerp, Green Peas)

Chris Lintott et al

SDSS III

14,555 square degrees
2,674,200 spectra



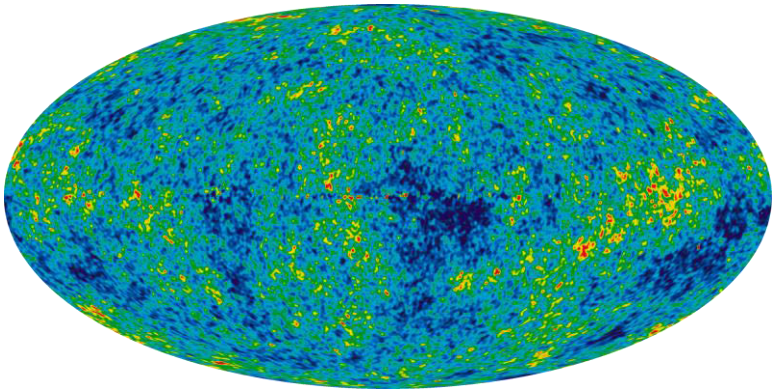
SDSSIII

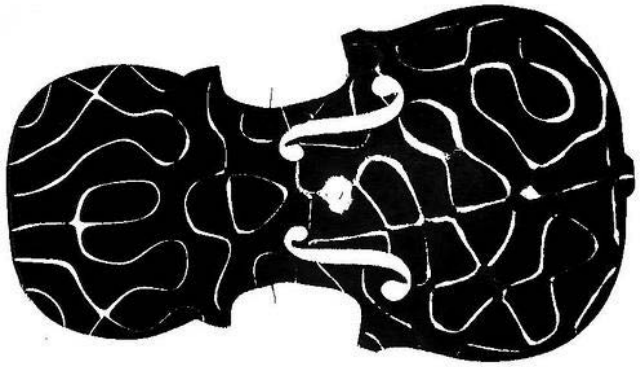
Numerous Science Projects

- 5,000 publications, 200,000 citations
- More papers from outside the collaboration
- From cosmology/LSS to galaxy evolution, quasars, stellar evolution, even time-domain
- Combination of 5-band photometry and matching spectroscopy provided unique synergy
- Overall, seeing not as good as originally hoped for, but systematic errors extremely well understood
- Very uniform, statistically complete data sets
- Photometry entirely redone for DR9, using cross-scans to calibrate the zero points across the stripes

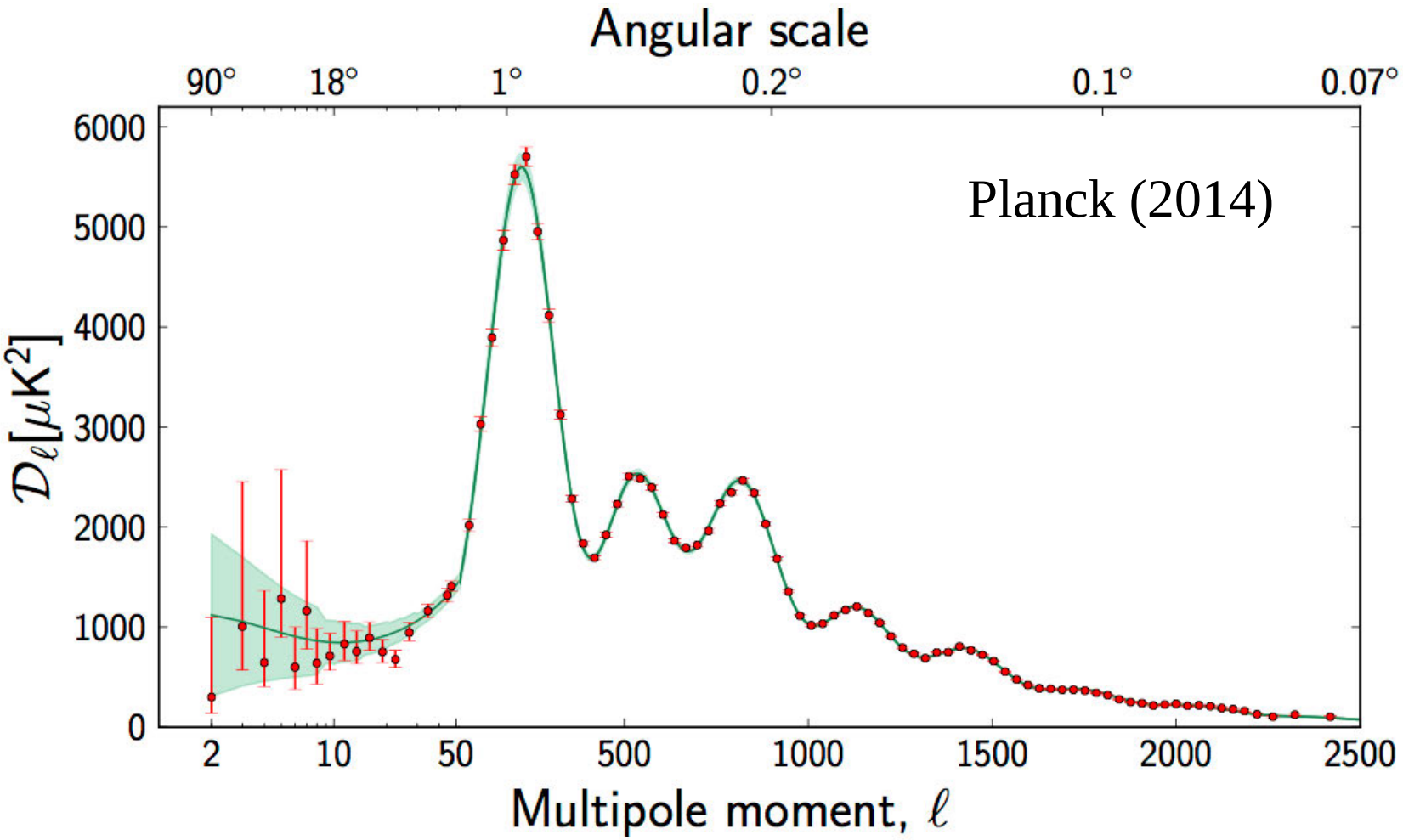
Baryon Acoustic Oscillations

- The Early Universe behaves like a resonant cavity (A. Sakharov)
- At 300,000 years the oscillating pattern “freezes”
- This provides the seeds of galaxy formation
- Observed in the Cosmic Microwave Background



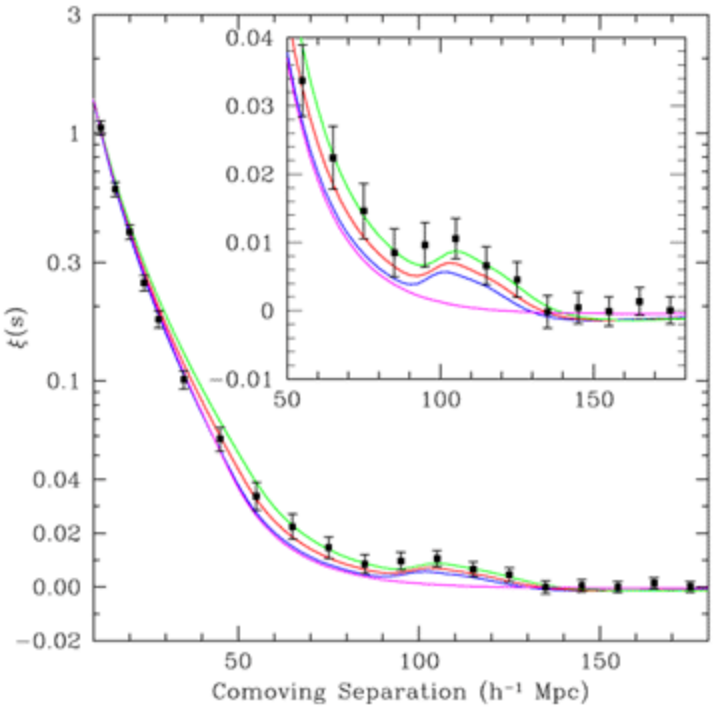


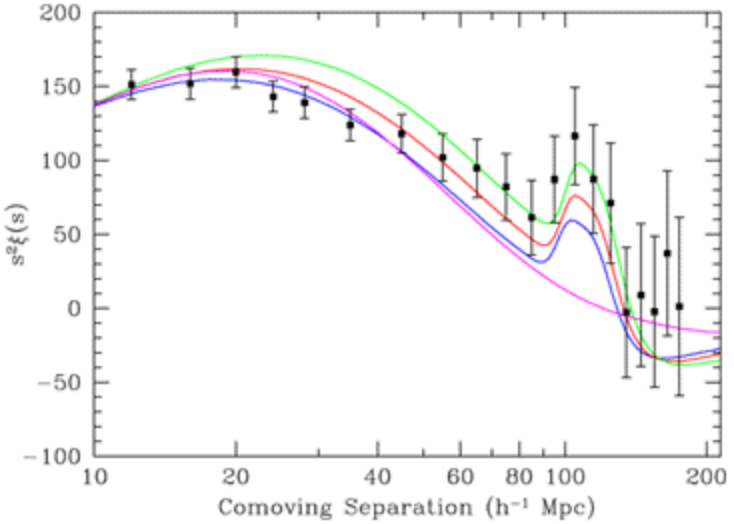
The Resonance Frequencies



Detecting the BAO in SDSS

- Eisenstein et al (2005) – DR4 LRG sample





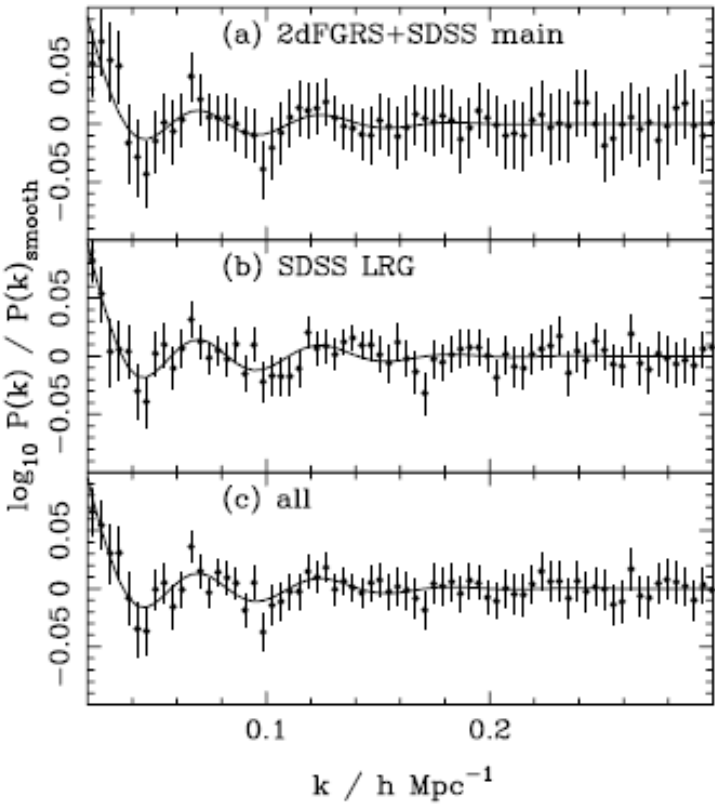
Correlation function

Primordial Sound Waves in SDSS

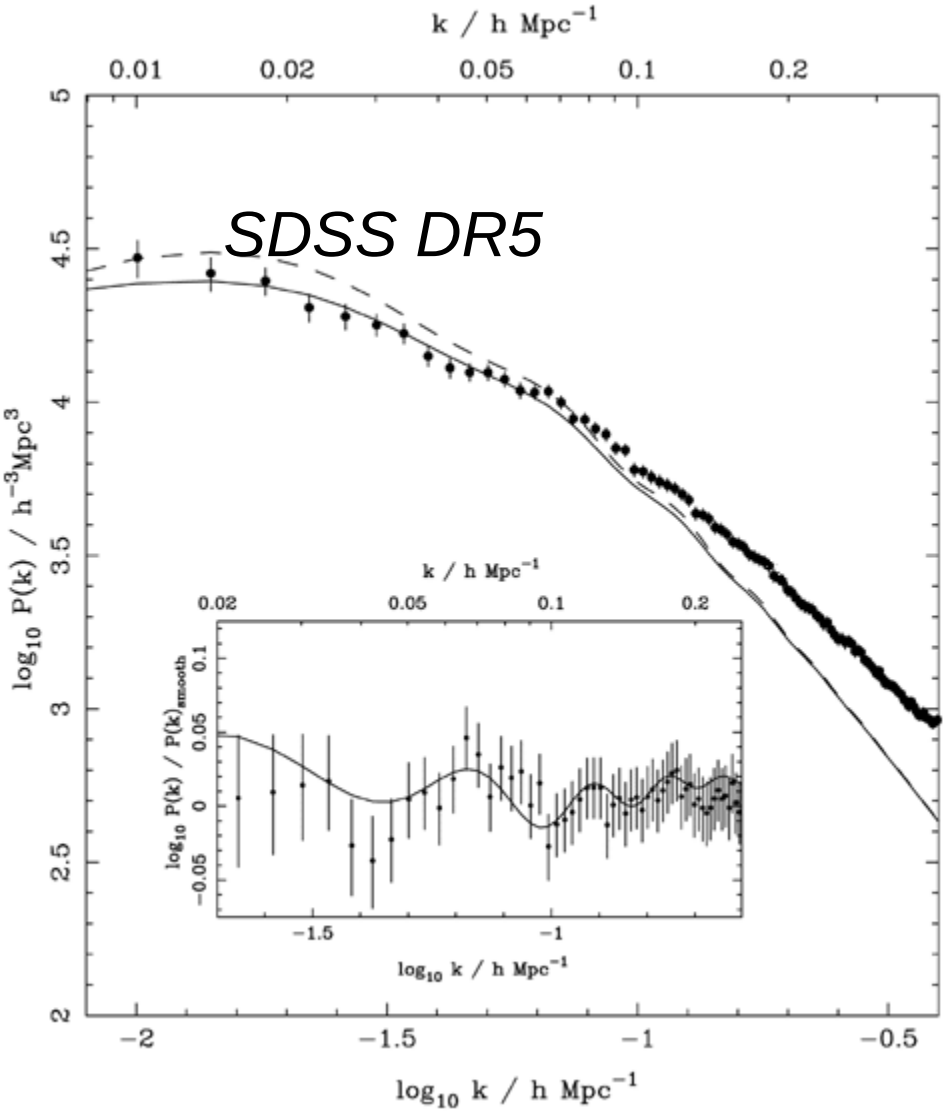
Power Spectrum

(Percival et al 2006, 2007)

SDSS DR6+2dF

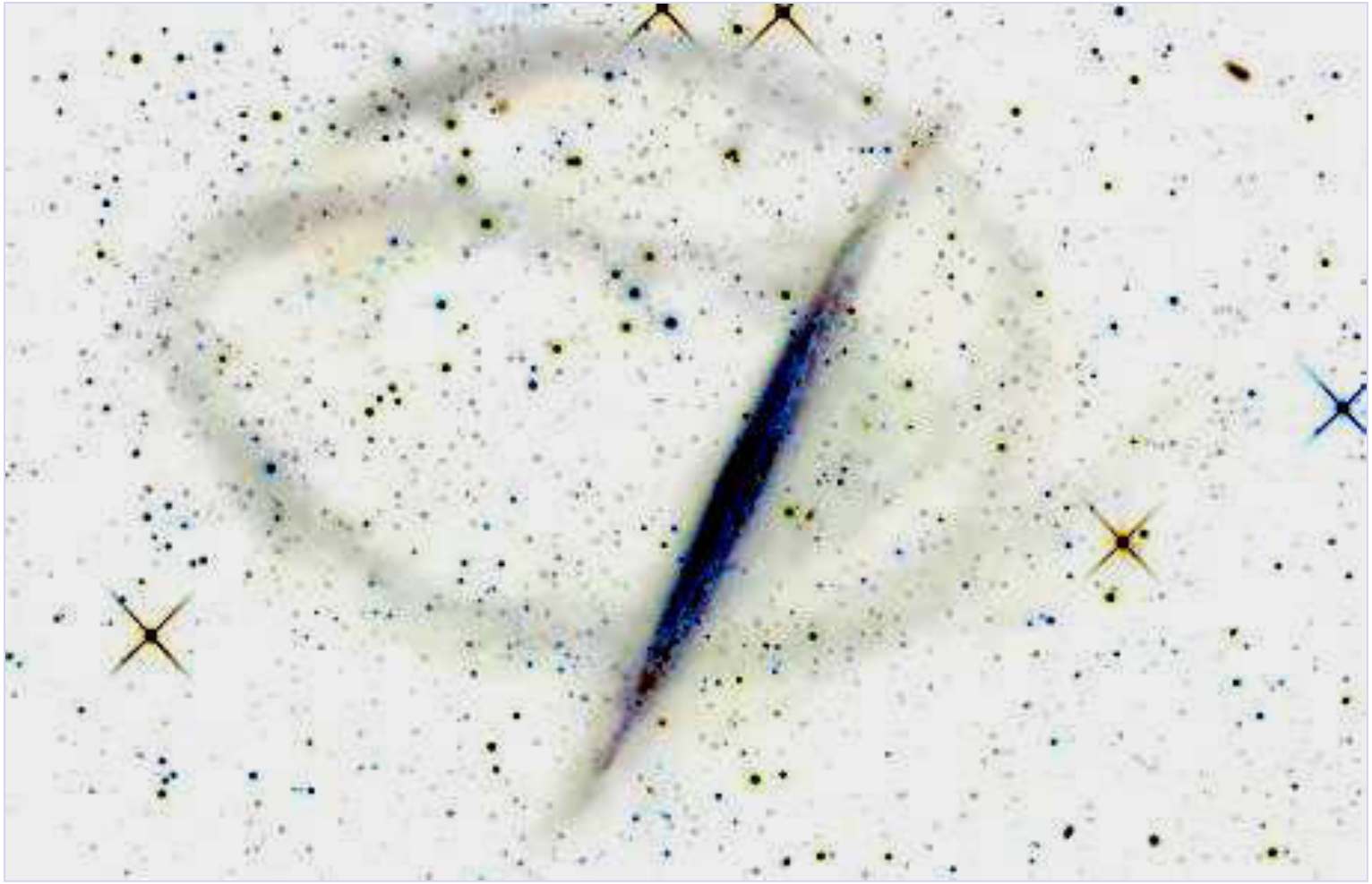


SDSS DR5

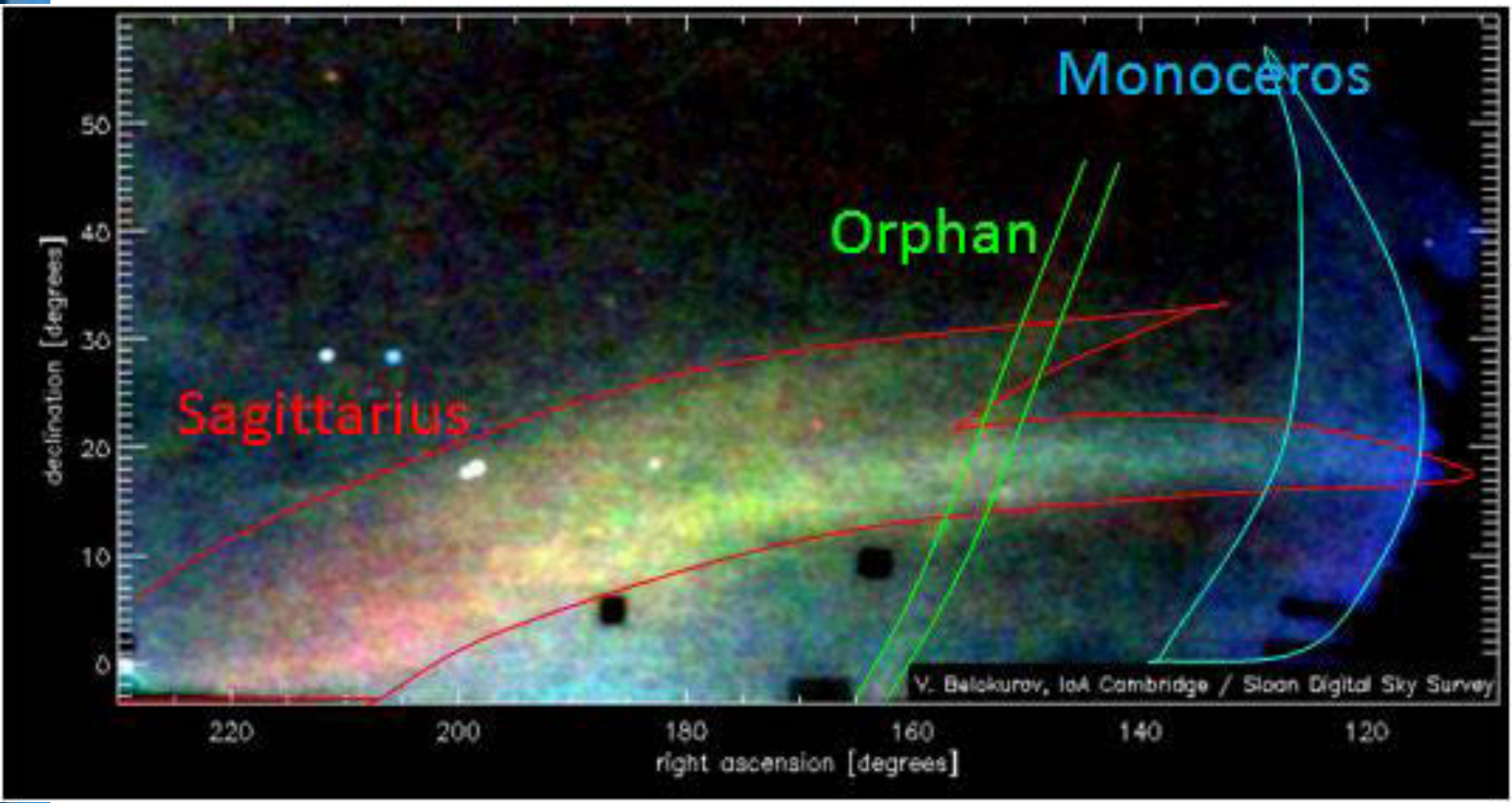


800K galaxies

Galactic Archeology



Field of Streams

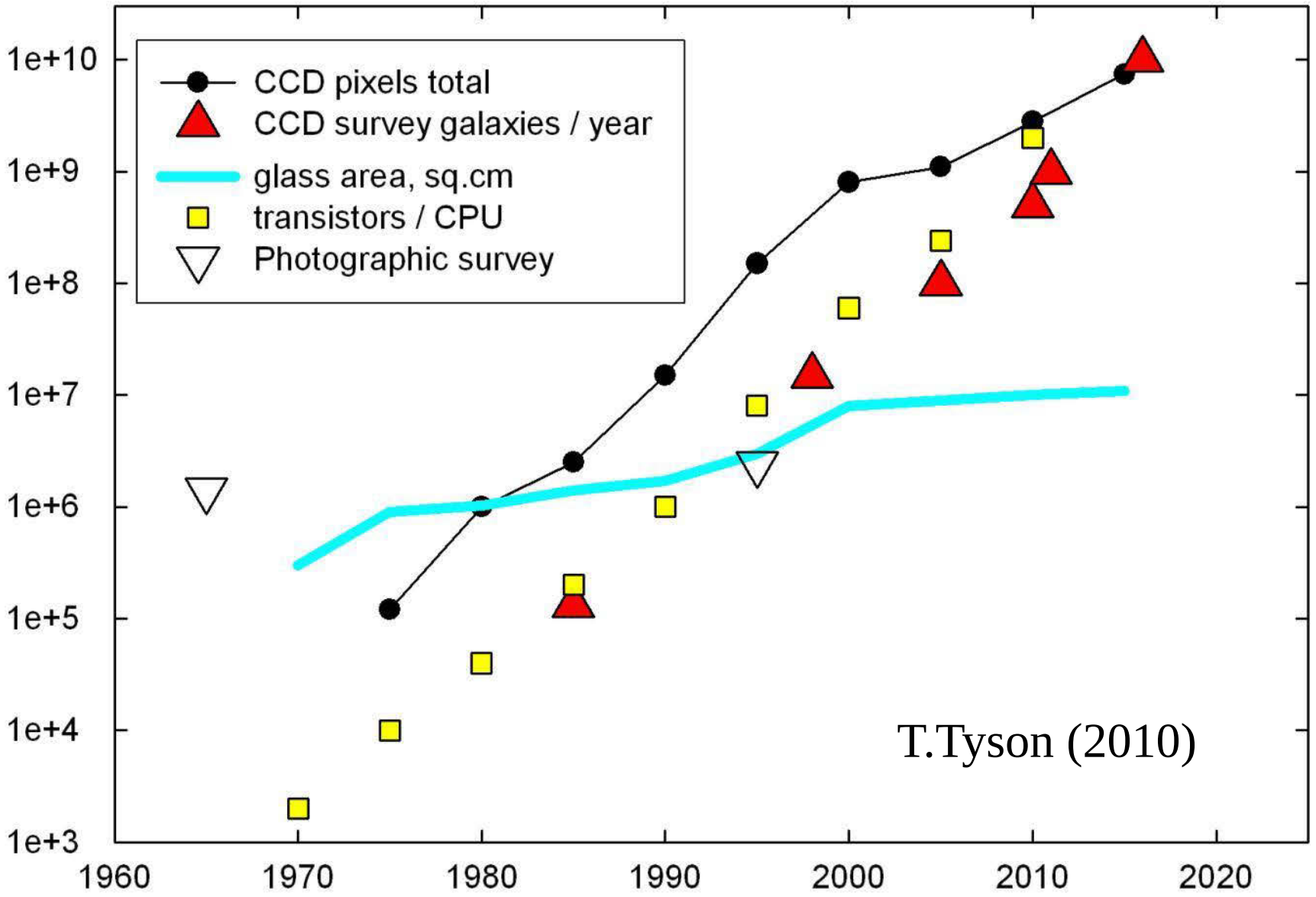


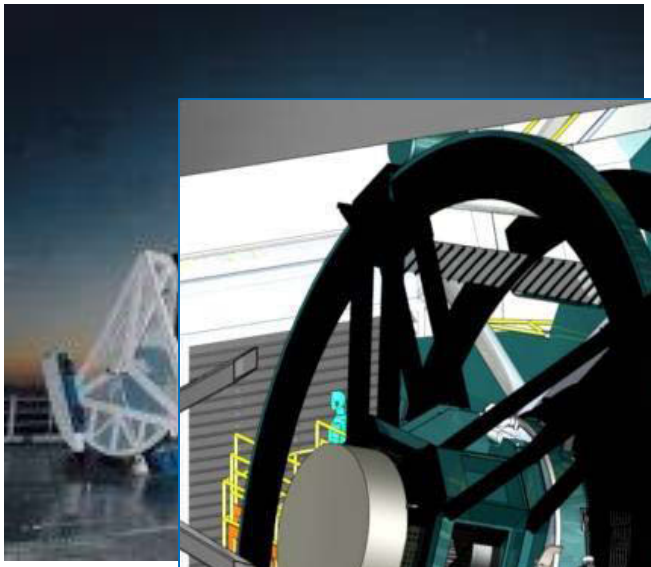
Belokurov et al 2006

The Broad Impact of SDSS

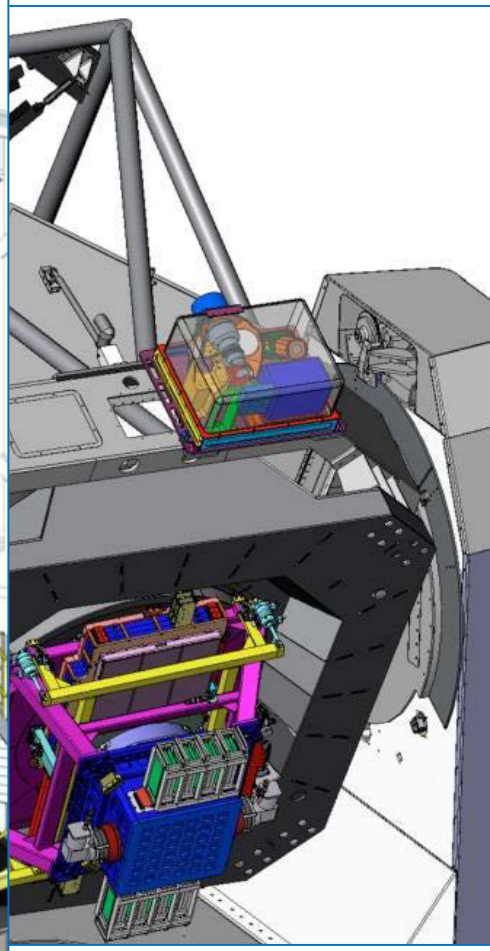
- Changed the way we do astronomy
- Remarkably fast transition seen for the community
- Speeded up the first phase of exploration
- Wide-area statistical queries easy
- Multi-wavelength astronomy is the norm
- SDSS earned the TRUST of the community
- Enormous number of projects, way beyond original vision and expectation
- Many other surveys now follow
- Established expectations for data delivery
- Serves as a model for other communities of science

Survey Trends





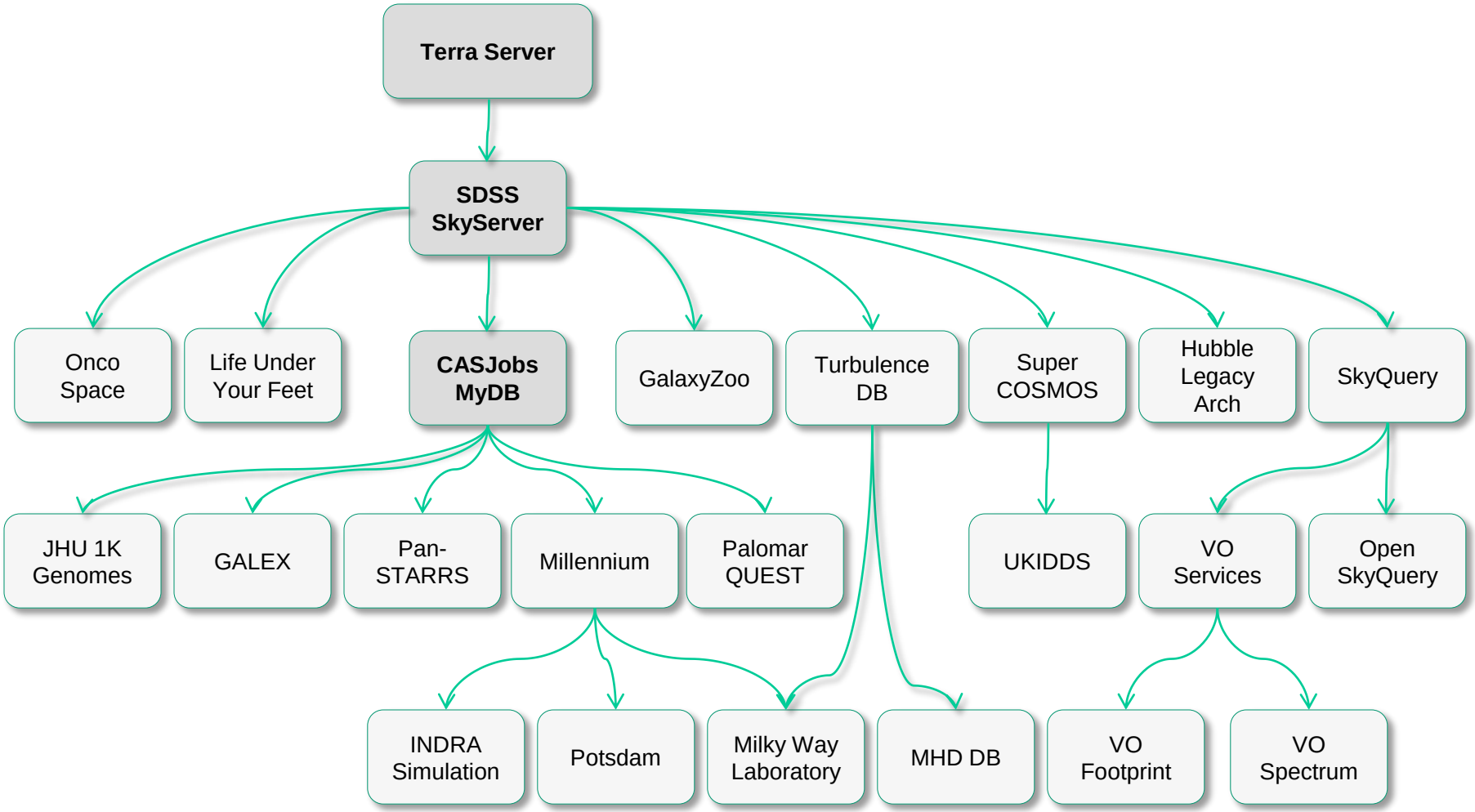
SD
2.4



LSST
8.4m 3.2Gpixel

PanSTARRS
1.8m 1.4Gpixel

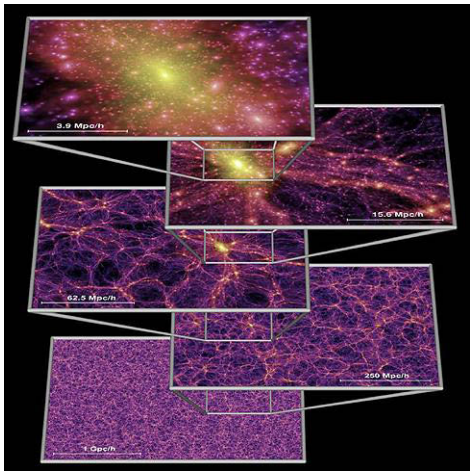
The SDSS Genealogy



Cosmological Simulations

Cosmological simulations have 10^9 particles and produce over 30TB of data (Millennium)

- Build up dark matter halos
- Track merging history of halos
- Use it to assign star formation history
- Combination with spectral synthesis
- Realistic distribution of galaxy types



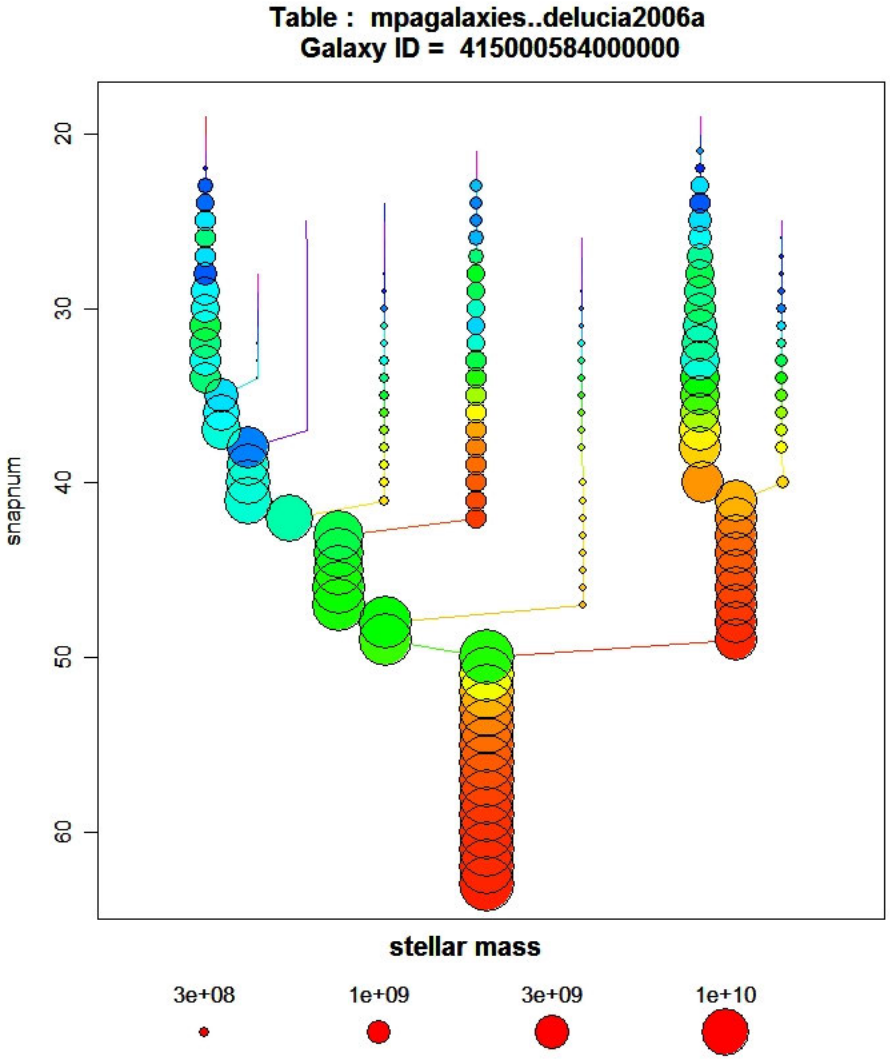
- Hard to analyze the data afterwards -> need DB
- What is the best way to compare to real data?
- Next generation of simulations with 10^{12} particles and 500TB of output are under way (Exascale-Sky)

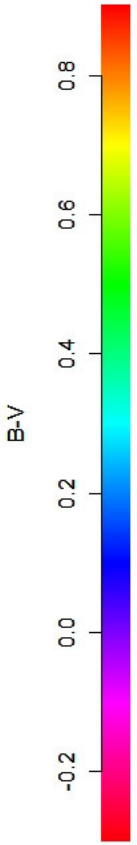
Millennium Database

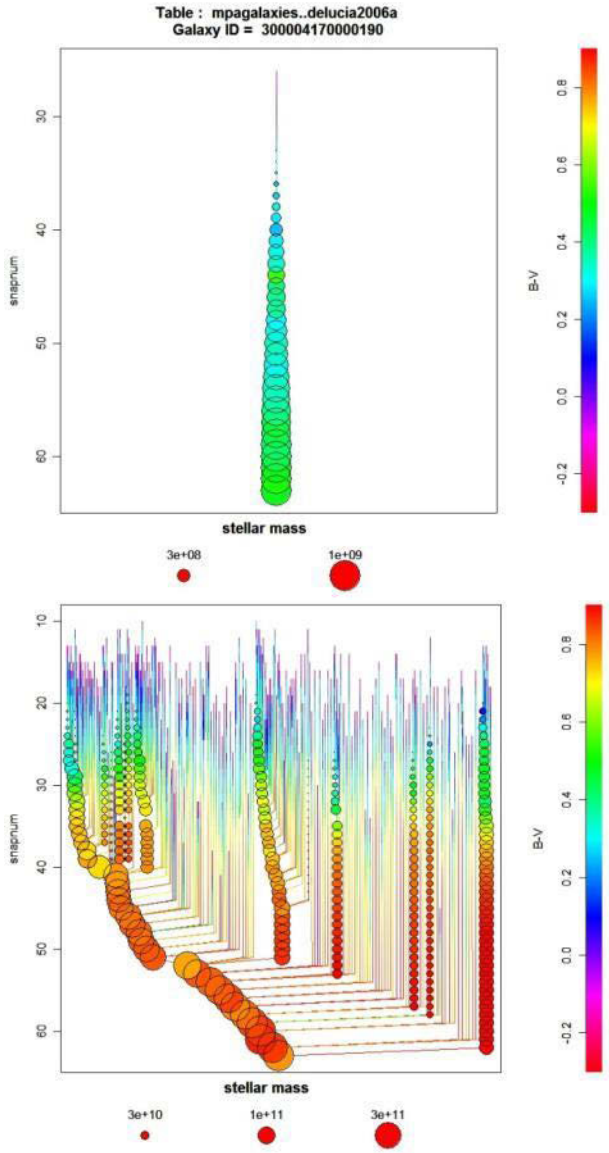
- **Density field on 256^3 mesh**
  - *CIC*
  - *Gaussian smoothed: 1.25,2.5,5,10 Mpc/h*
- Friends-of-Friends (FOF) groups
- SUBFIND Subhalos
- Galaxies from 2 semi-analytical models (SAMs)
  - *MPA (L-Galaxies, DeLucia & Blaizot, 2006)*
  - *Durham (GalForm, Bower et al, 2006)*
- Subhalo and galaxy formation histories: merger trees
- Mock catalogues on light-cone
  - *Pencil beams (Kitzbichler & White, 2006)*
  - *All-sky (depth of SDSS spectral sample)*

Gerard Lemson 2006

Time evolution: merger trees







Big Data in Science

- Data growing exponentially, in all science
- All science is becoming data-driven
- This is happening very rapidly
- Data becoming increasingly open/public
- Non-incremental!
- Convergence of physical and life sciences through Big Data (statistics and computing)
- The “long tail” is important
- A scientific revolution in how discovery takes place

  => a rare and unique opportunity

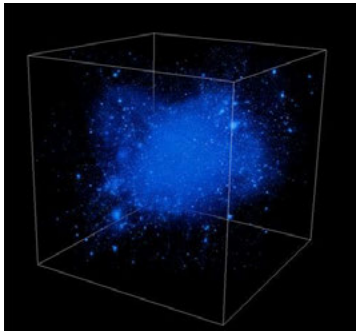
Science is Changing

THOUSAND YEARS AGO
science was empirical
describing natural phenomena

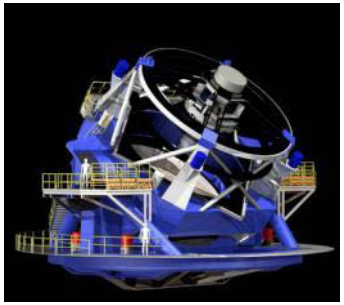
LAST FEW HUNDRED YEARS
theoretical branch using models, generalizations

$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{4\pi G\rho}{3} - \mathrm{K}\frac{c^2}{a^2}$$

LAST FEW DECADES
a computational branch simulating complex phenomena



TODAY
data intensive science, synthesizing theory, experiment and computation with statistics
►new way of thinking required!



Why Is Astronomy Interesting?

Astronomy has always been data-driven…. now this is becoming more accepted in other areas as well

- Important spatio-temporal features
- Very large density contrasts in populations
- Real errors and covariances
- Many signals very subtle, buried in systematics
- Data sets large, pushing scalability
  - *LSST will be 100PB*

*"Exciting, since it is **worthless**!"*

— Jim Gray

Non-Incremental Changes

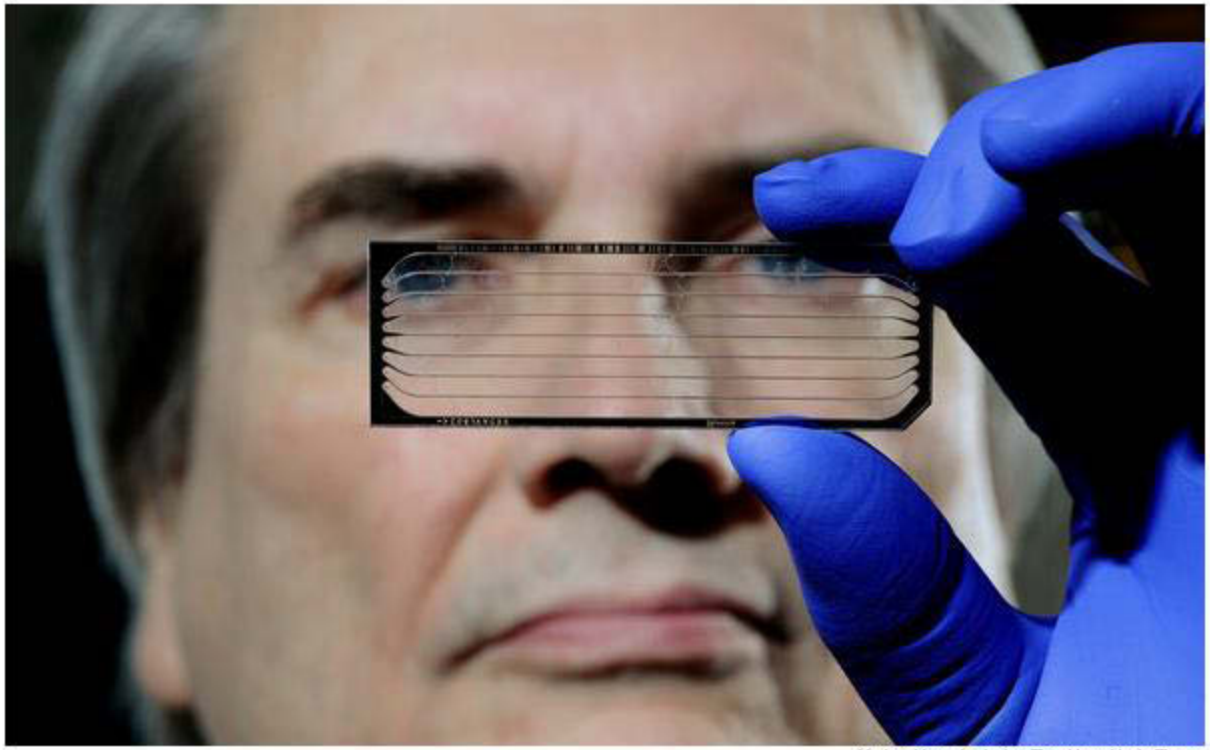
- Multi-faceted challenges
- New computational tools and strategies

  … not just statistics, not just computer science, not just astronomy, not just genomics…
- Science is moving increasingly from hypothesis-driven to data-driven discoveries

The New York Times

Business Day

DNA Sequencing Caught in Deluge of Data



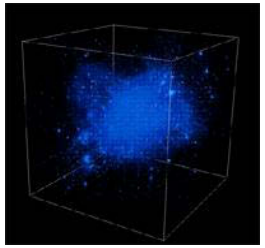
Kathy Kmonicek for The New York Times

W. Richard McCombie, a professor of human genetics at the Cold Spring Harbor Laboratory, examining DNA samples.

By ANDREW POLLACK
Published: November 30, 2011

Trends

- Broad sociological changes
  - *Convergence of Physical and Life Sciences*
  - *Data collection in ever larger collaborations*
  - *Virtual Observatories: CERN, IVOA, NCBI, NEON, OOI,…*
  - *Analysis decoupled, off archived data by smaller groups*
  - *Emergence of the citizen/internet scientist (GalaxyZoo…)*
- Need to start training the next generations
  - *Π-shaped vs I- and T-shaped people*
  - *Early involvement in "Computational thinking"*



Summary

- Science is increasingly driven by data (big and small)
- Surveys analyzed by individuals
- From hypothesis-driven to data-driven science
- “Microscopes” & “Telescopes” for data
- A major challenge on the “long tail”
- A new, Fourth Paradigm of Science is emerging…
- SDSS has been at the cusp of this transition