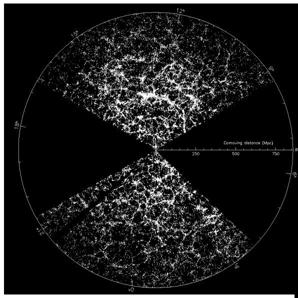
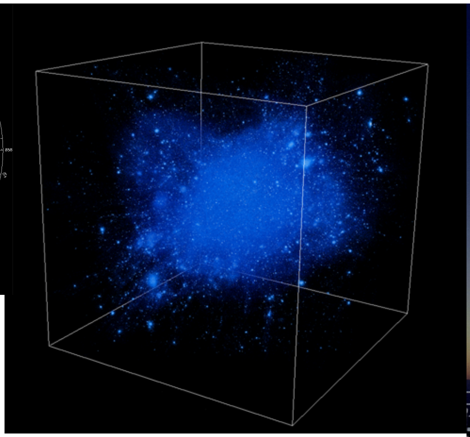
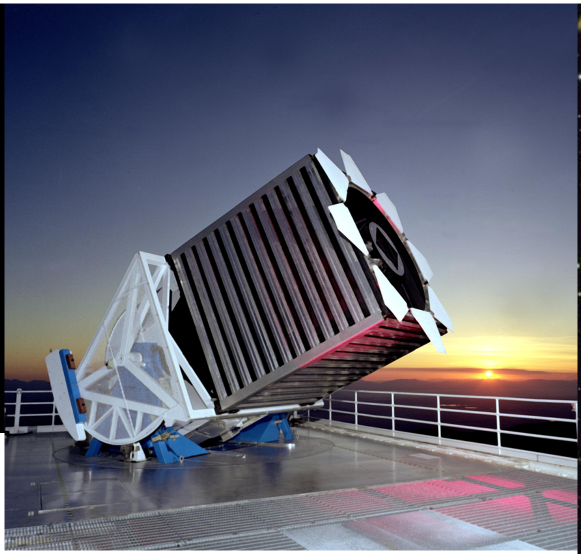






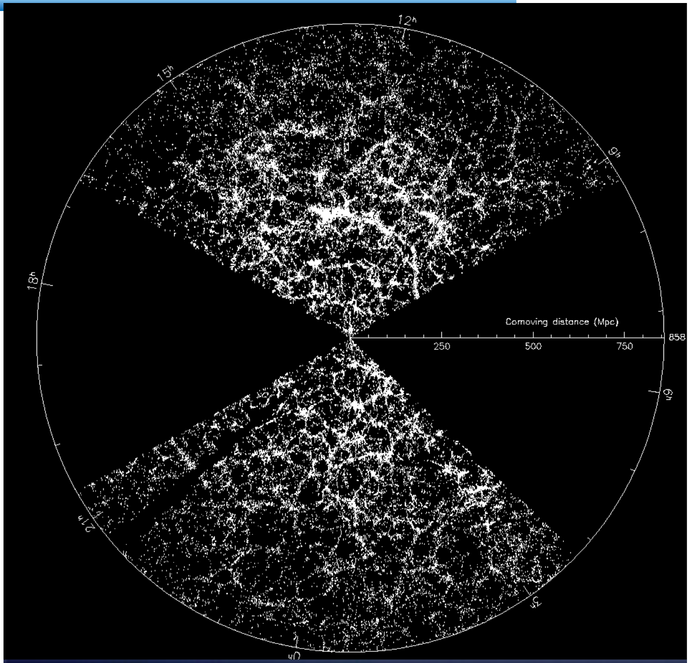
The Legacy Value of Large Public Surveys: the SDSS Archive

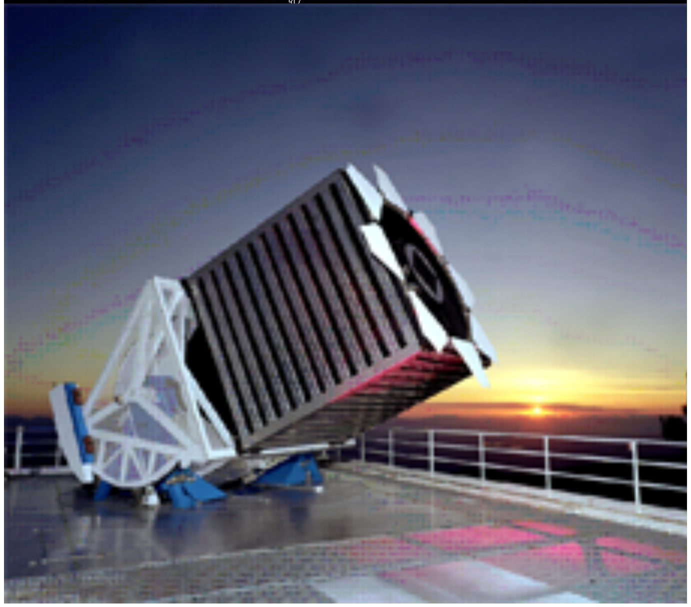
Alexander Szalay
The Johns Hopkins University

Sloan Digital Sky Survey

“The Cosmic Genome Project”

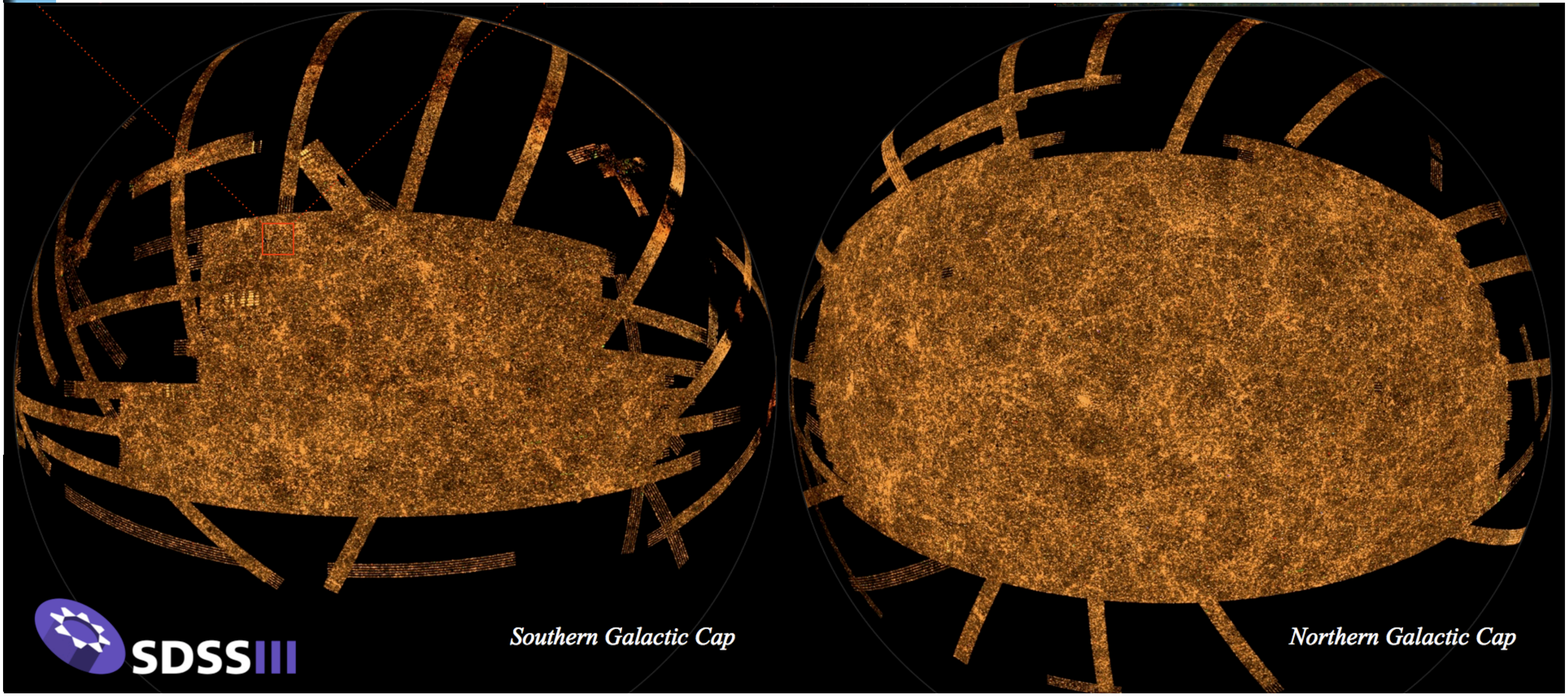
- Started in 1992, finished in 2008
- Data is public
  - 2.5 Terapixels of images => 5 Tpx of sky
  - 10 TB of raw data => 100TB processed
  - 0.5 TB catalogs => 35TB in the end
- Database and spectrograph built at JHU (SkyServer)
- Now SDSS-3/4 data served from JHU



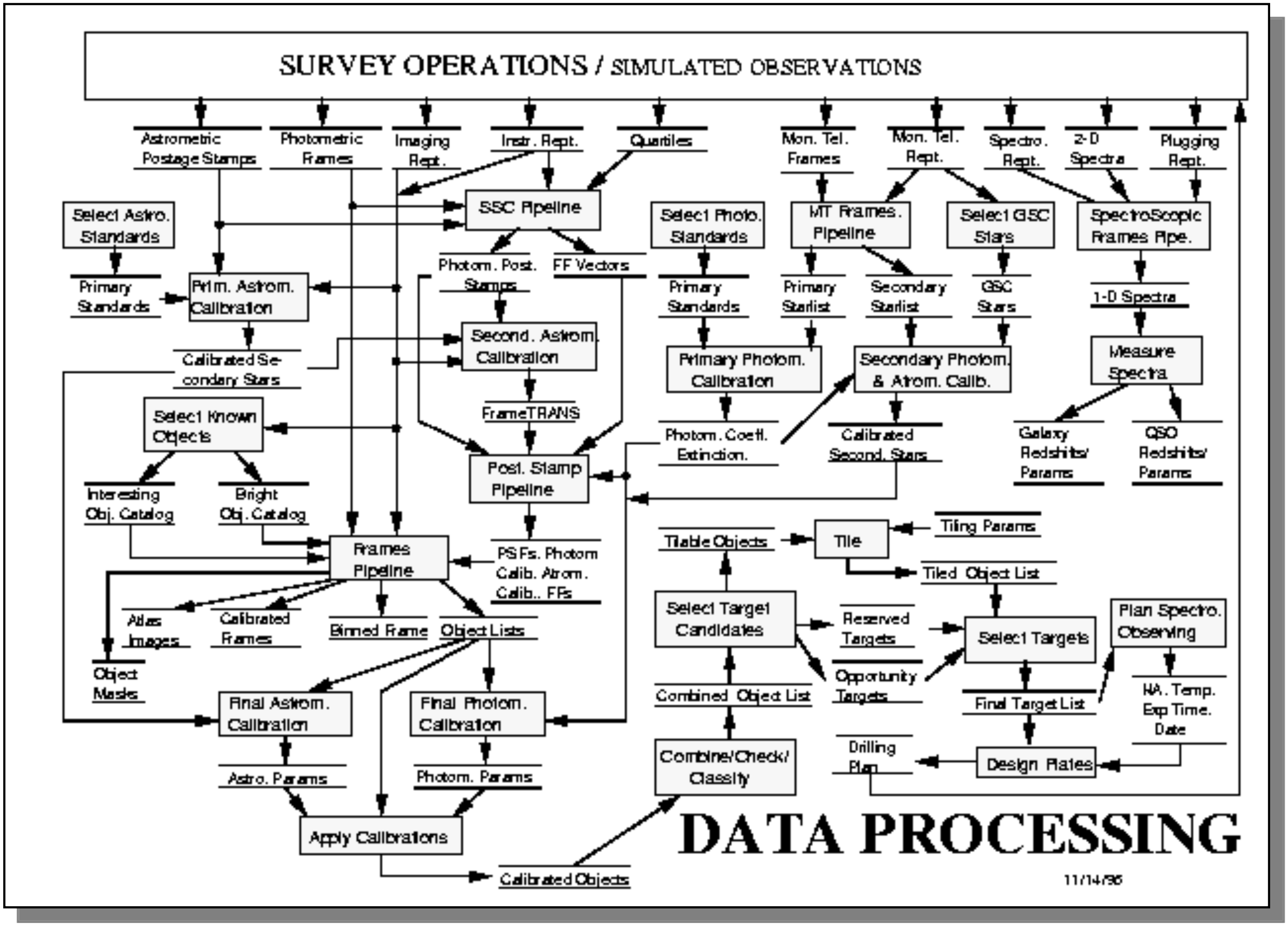


SDSS III

14,555 square degrees
2,674,200 spectra



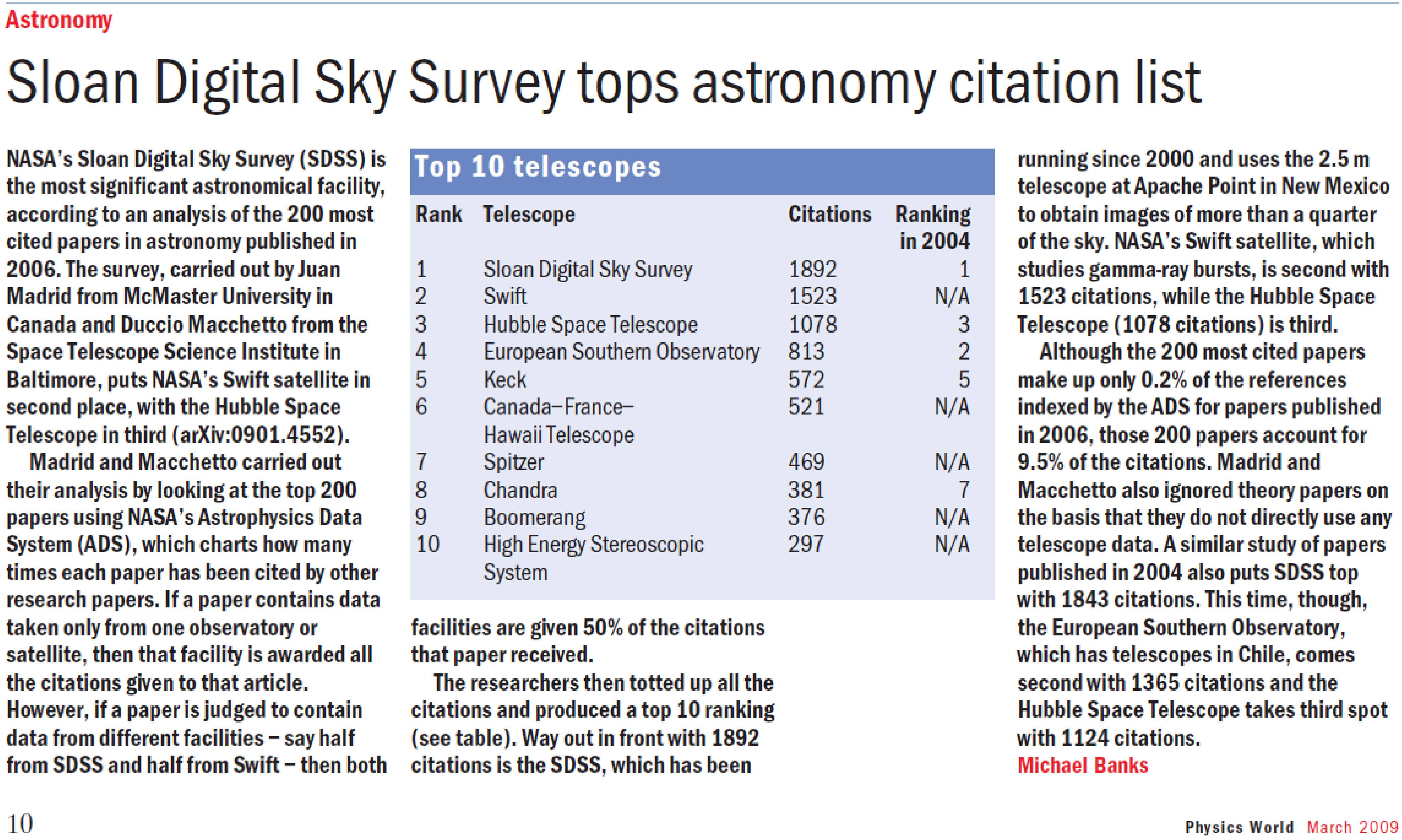
Data Processing Pipelines



Wide Range of Science

- 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

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.

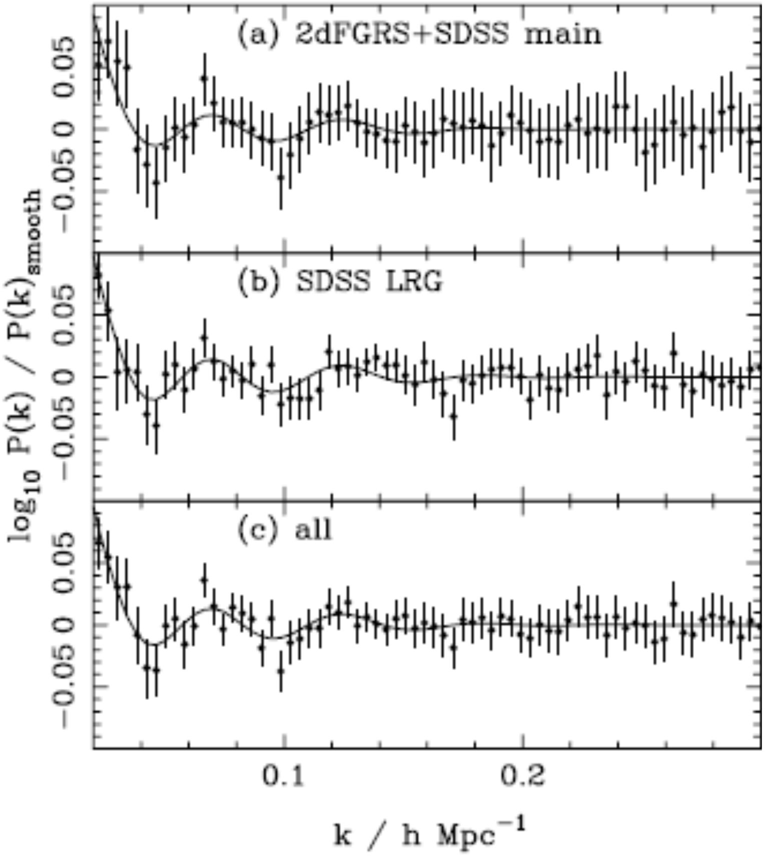
Michael Banks

Primordial Sound Waves in SDSS

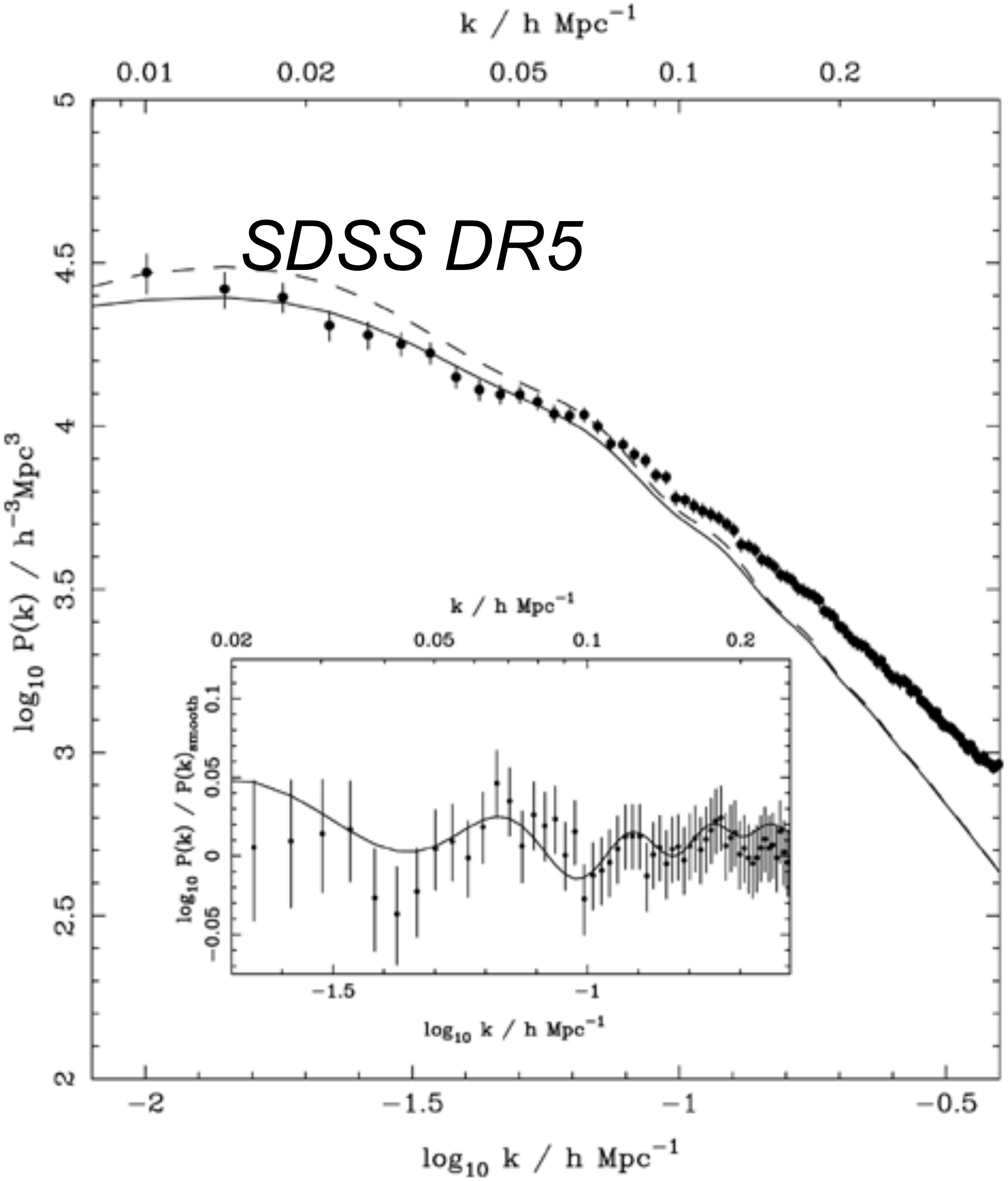
Power Spectrum

(Percival et al 2006, 2007)

SDSS DR6+2dF

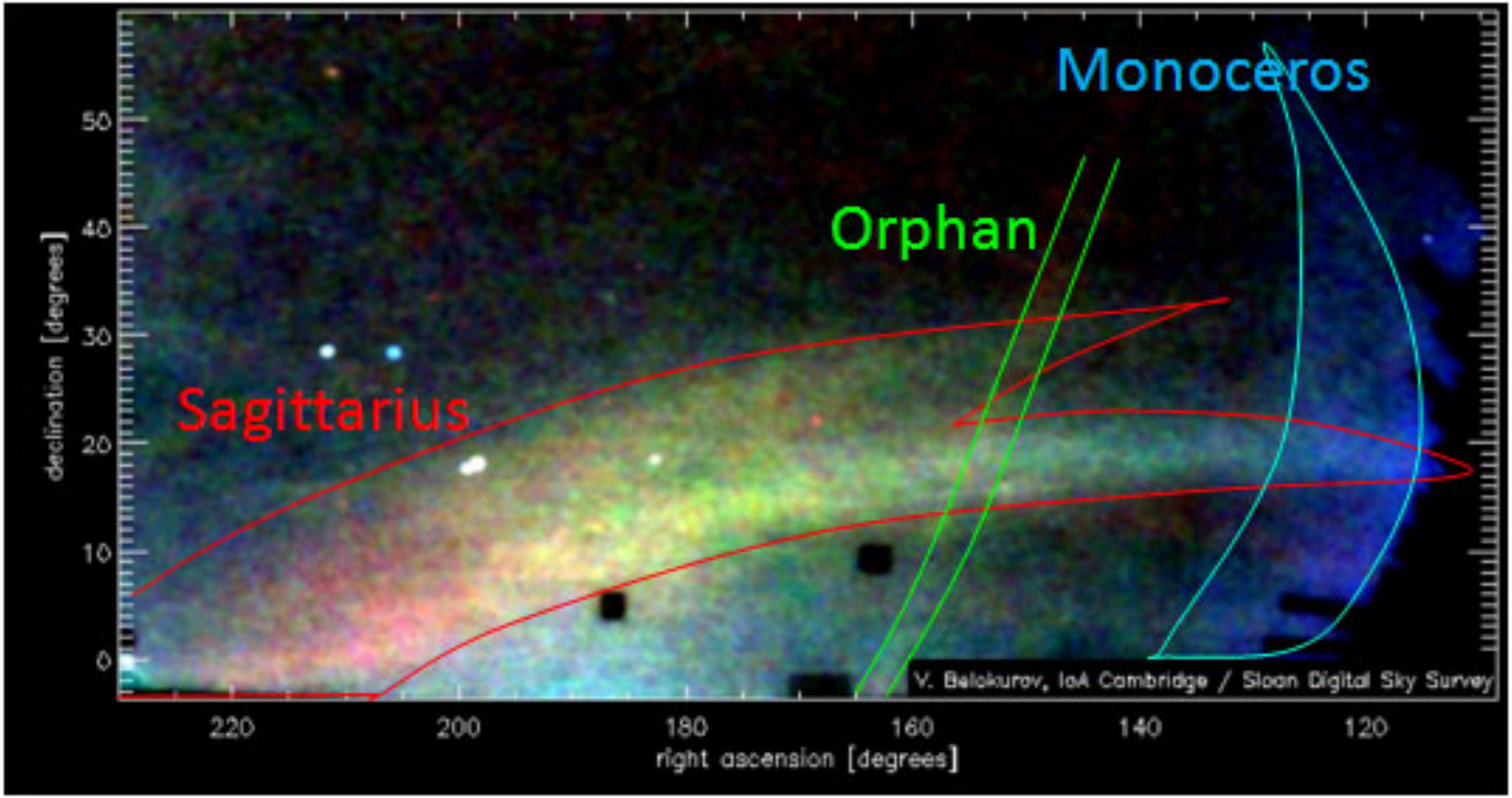


800K galaxies



SDSS DR5

Field of Streams

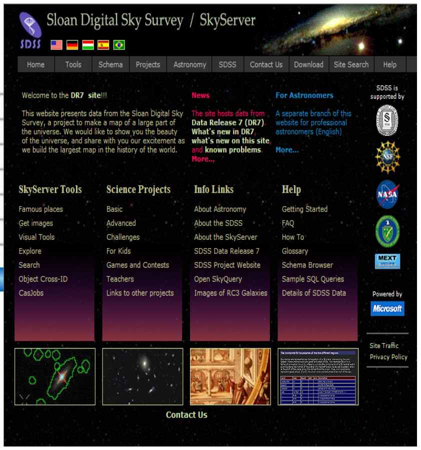


Belokurov et al 2006

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 now 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

Skyserver



Welcome to the **DR7 site**!!!

This website presents data from the Sloan Digital Sky Survey, a project to make a map of a large part of the universe. We would like to show you the beauty of the universe, and share with you our excitement as we build the largest map in the history of the world.

News

The site hosts data from **Data Release 7 (DR7)**. **What's new in DR7**, **what's new on this site**, and **known problems**. **More...**

For Astronomers

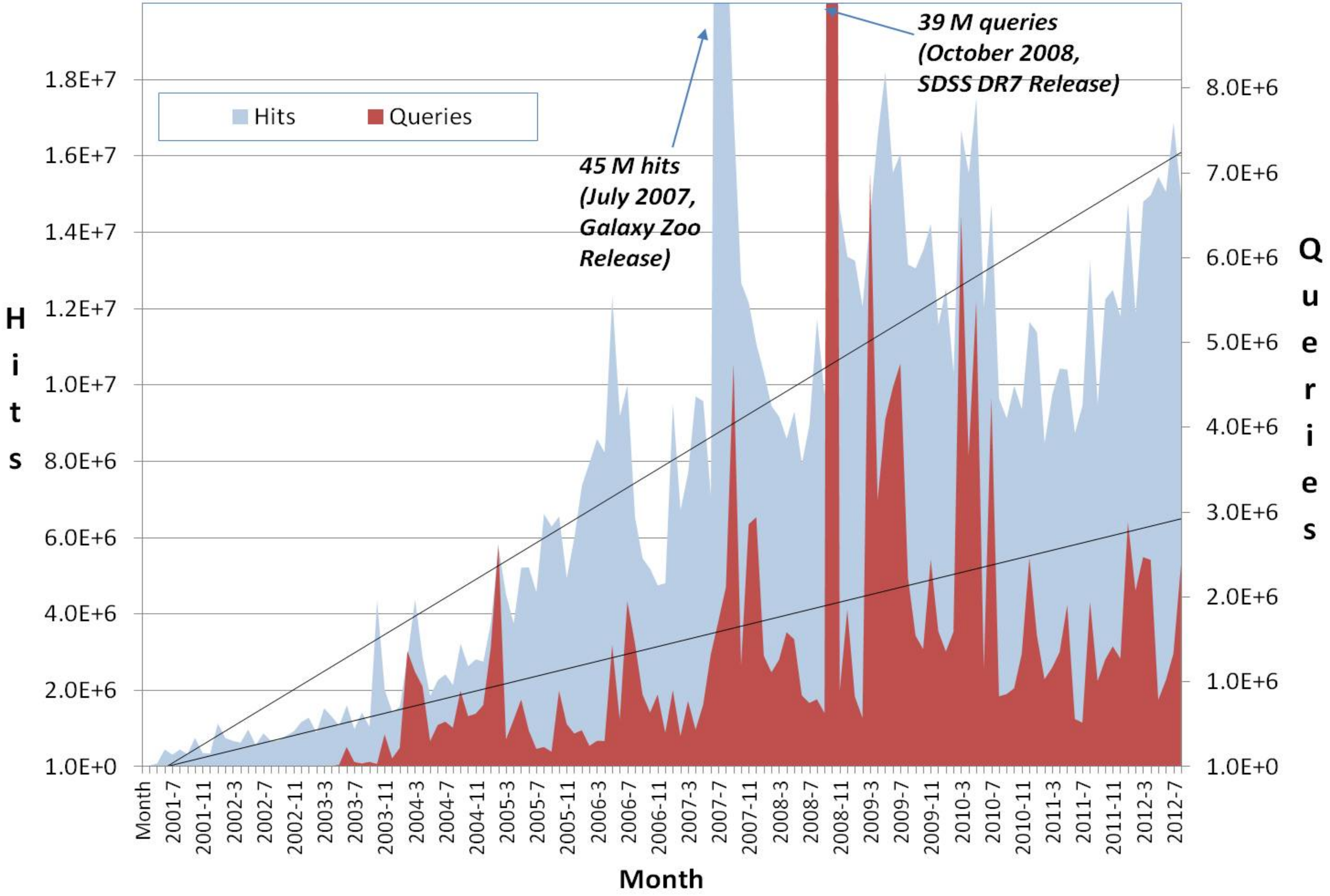
A separate branch of this website for professional astronomers (English)

More...

Prototype in 21st Century data access

- *1.6B web hits in 14 years*
- *280M external SQL queries*
- *5,000+ papers and 200K+ citations*
- *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



45 M hits (July 2007, Galaxy Zoo Release)

39 M queries (October 2008, SDSS DR7 Release)

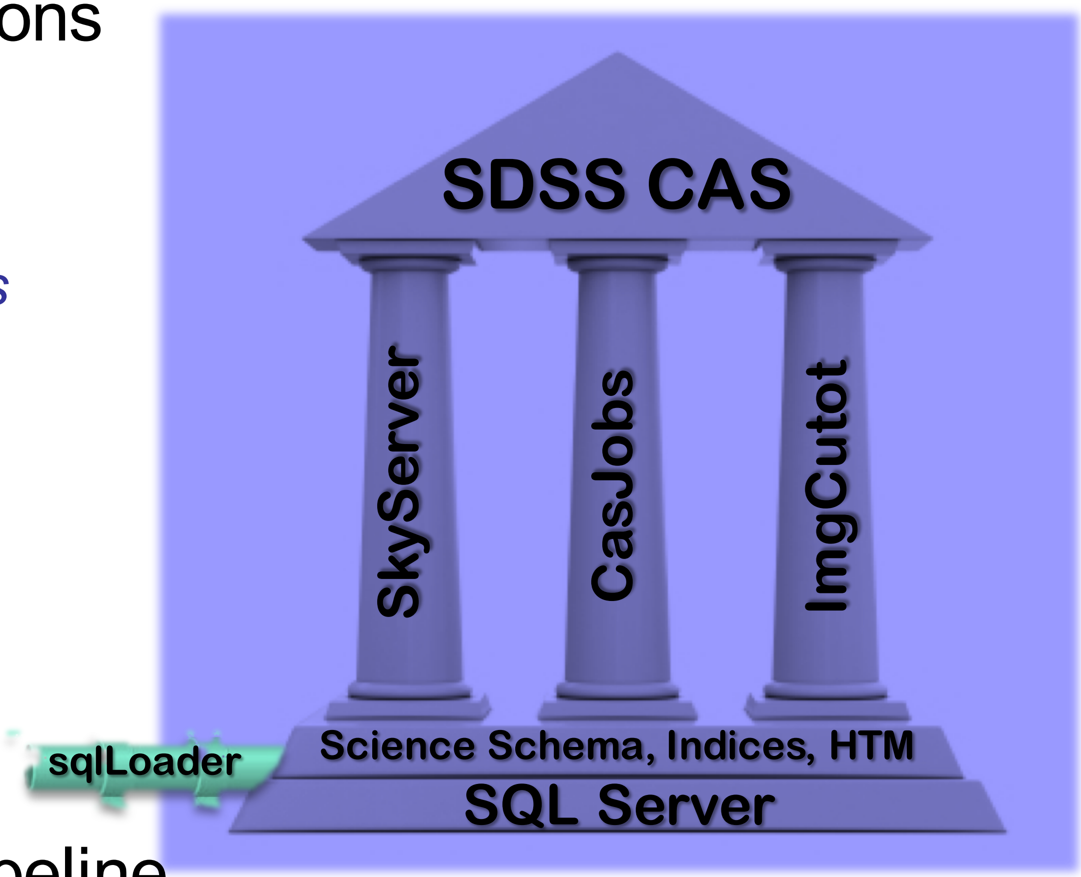
GalaxyZoo

- 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 Database Design

- MS SQL Server + extensions
- Layer of science schema built right into the DB
  - *Extensive use of UDFs/SPs*
  - *HTM spatial index in C# CLR extensions*
- 3 pillars of data access
  - *Synch: SkyServer*
  - *Asynch: CasJobs*
  - *Visual: ImgCutout*
- sqlLoader data loading pipeline



SkyServer Web Interface

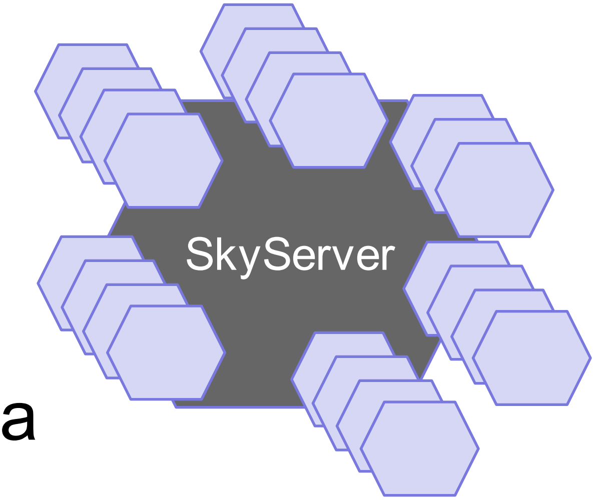
- The public portal to CAS data since 2001
- Supports several levels of user access
  - *Simple to complex form queries*
  - *CrossID search with upload capability*
  - *Visual browsing of individual objects*
  - *Raw SQL queries*
  - *Direct access to raw/calibrated FITS files*
- Service oriented architecture
  - *First web services in science*
  - *Virtual Observatory services (VO standards/protocols)*
- Includes client for ImgCutout service
  - *Finding Chart page, Navigate page, Queryable Image List*
- Integrated Schema Browser, extensive SQL help
- Rich educational projects section (K-12+)

SkyServer Usage Logging

- All web hits and queries logged since day 1 (2001)
- SkyServer traffic page shows up-to-the-hour logs
  - *1.68 billion hits, 282 million SQL queries to date*
  - *Currently averaging 15M hits and 1.5M queries/mo*
- Logging overview document at skyserver.org/doc
- 3 published papers on SkyServer traffic:
  - *"SkyServer Traffic Report – The First Five Years", MS Technical Report (Singh et al. 2006)*
  - *"Ten Years of SkyServer – Tracking Web and SQL e-Science Usage", CiSE (Raddick et al. 2014)*
  - *"Ten Years of SkyServer – How Astronomers and the Public Have Embraced e-Science", CiSE (Raddick et al. 2014)*

CasJobs

- Batch query workbench (launched 08/2003)
- Web application + web service backend
  - *ASP.NET/C# development platform*
- Workhorse of CAS data access
  - *SDSS-II CasJobs: 9300 users (5000 active), 6.1M jobs*
  - *SDSS-III CasJobs: 3800 users (2500 active), 8.8M jobs*
- Every user has their own SQL "MyDB"
  - *Default size 0.5 GB, increased on request*
  - *Users can do anything in their own DB*
- Complete searchable job history
- Schema browser, MyDB table browser
- Data Import, Groups feature to share data



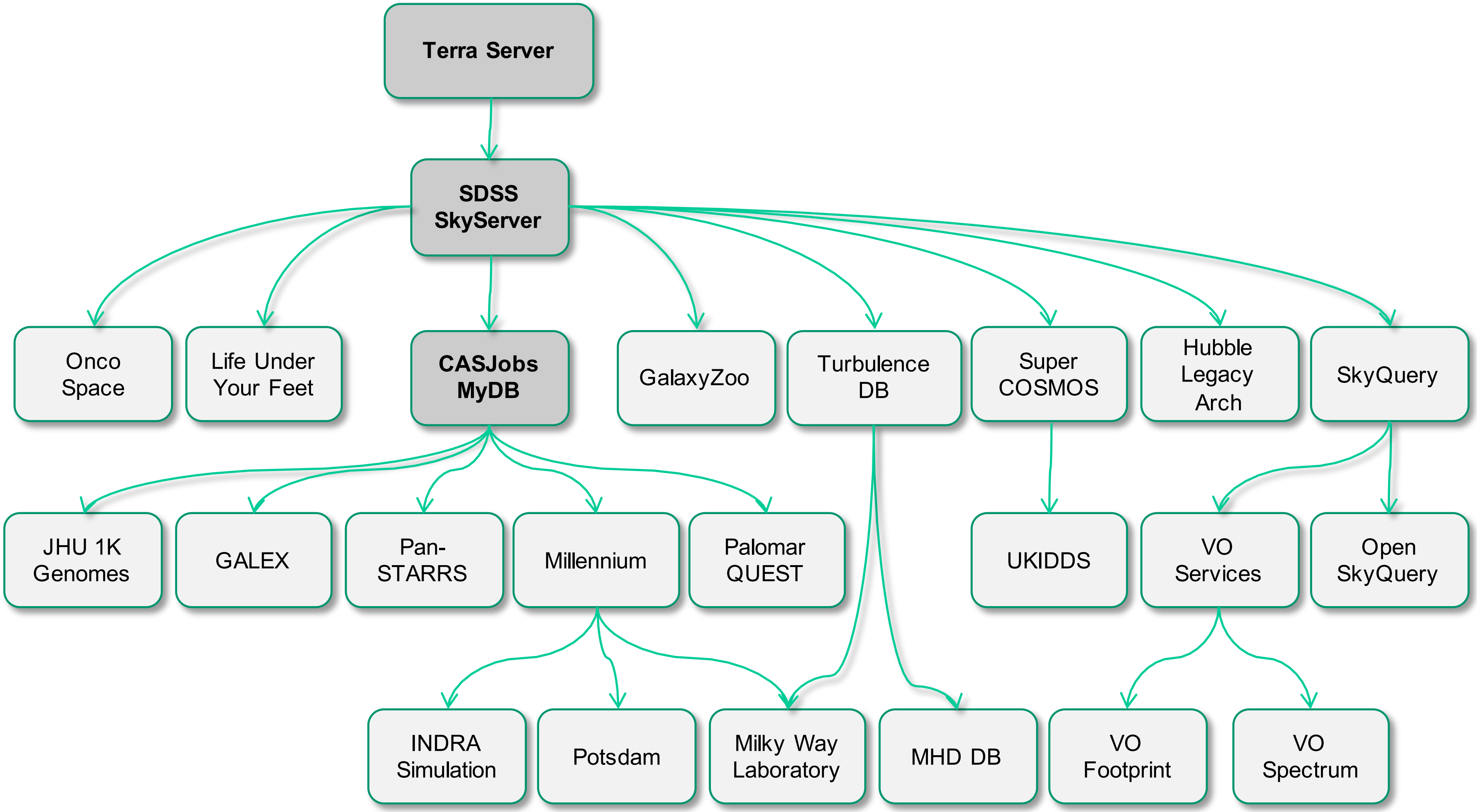
SphericalLib .NET

- 8,500 lines of C# code – 20k total
  - *OS independent (Windows, Un*x w/ Mono)*
  - *Documentation via Sandcastle*
  - *Automatic conversion into C++*
- Great performance
  - *Fast Boolean operations, e.g., intersect, union, difference*
  - *Almost 1000 times faster than MS Spatial for certain queries*
  - *Correct handling of degenerate cases and other imprecision of representation*
  - *Fast point in shape searches*
  - *GDI+ based visualization*
  - *Exact areas*

Data Storage

- Yearly Data Releases, from DR1— DR12 all live
- Several SDSS projects (I-II-III-IV)
- Production CAS moved FNAL ⇒ JHU for SDSS-III
  - *Cluster architecture/ops similar to SDSS-II at FNAL*
- Multiple instances of each DR
  - *For redundancy, load-balancing and performance*
    - Workload segregation
  - *As many as 6 copies of most active DRx!*
    - SkyServer, Quick CasJobs, Long Public CasJobs, Long Collab CasJobs, Imgcutout and development/backup/restore copy
- Hardware choices tend to be conservative
  - *Tradeoffs for reliability vs performance*
- Currently ~ 120 TB of just DR8-DR11 DBs

The SDSS Genealogy

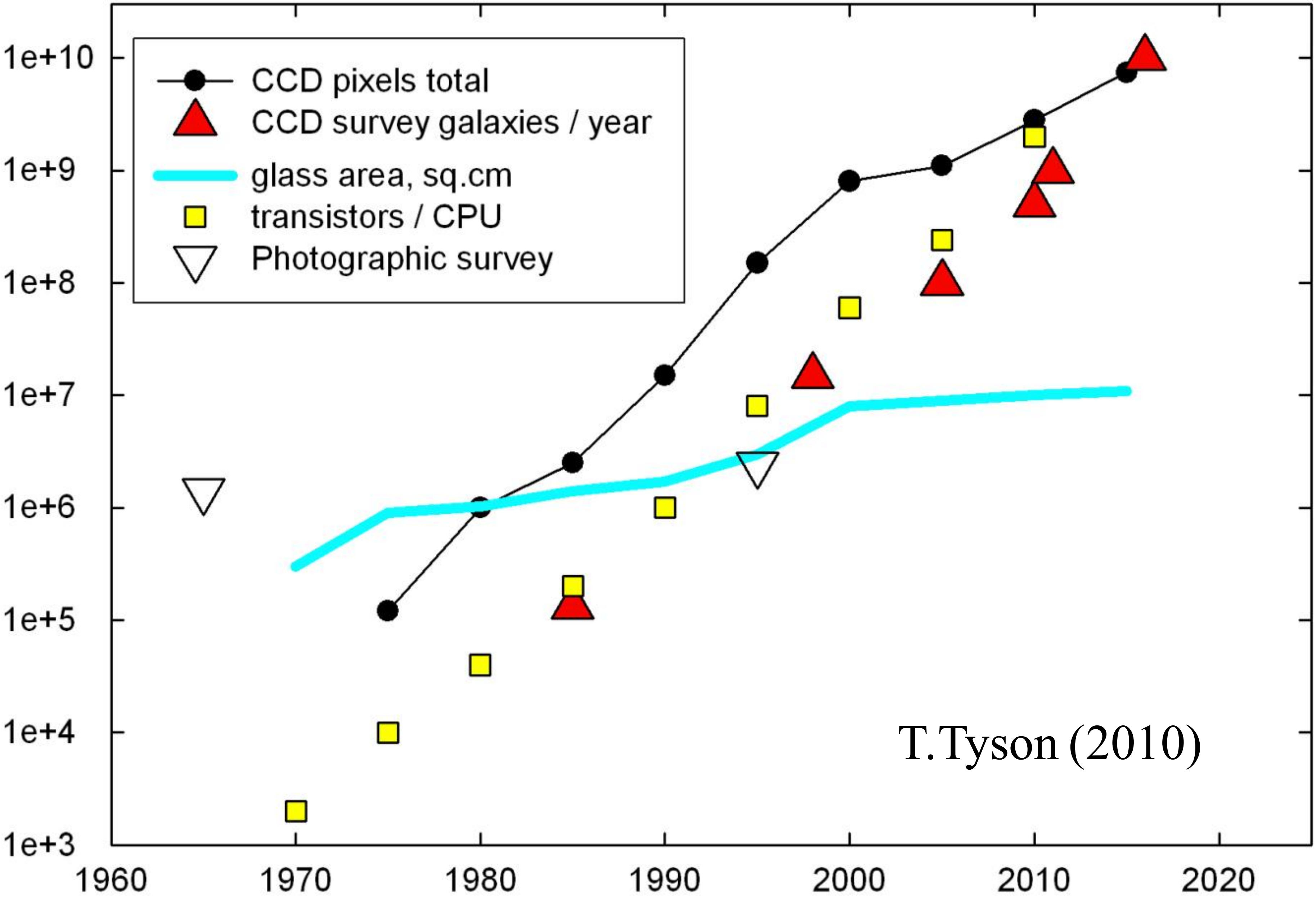


The Evolving SkyServer

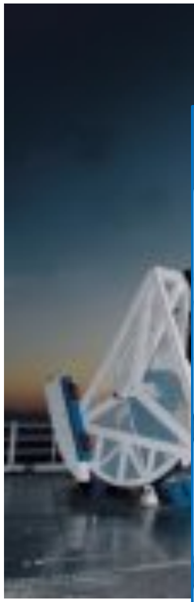
- Service oriented smart data
- More collaborative features added
- System captures interactivity of science well
- Read-only, secure core – free-for-all in MyDB
- Increasingly complex analysis patterns
- Extensive use by other disciplines
- But: signs of service lifecycle after 15 years

  => *NSF DIBBS grant*
- Comprehensive overhaul under the hood
- Now adding iPython scripting, running on pool of VMs
- Single sign-on to CASJOBS + SciDrive (Dropbox+VOSpace)

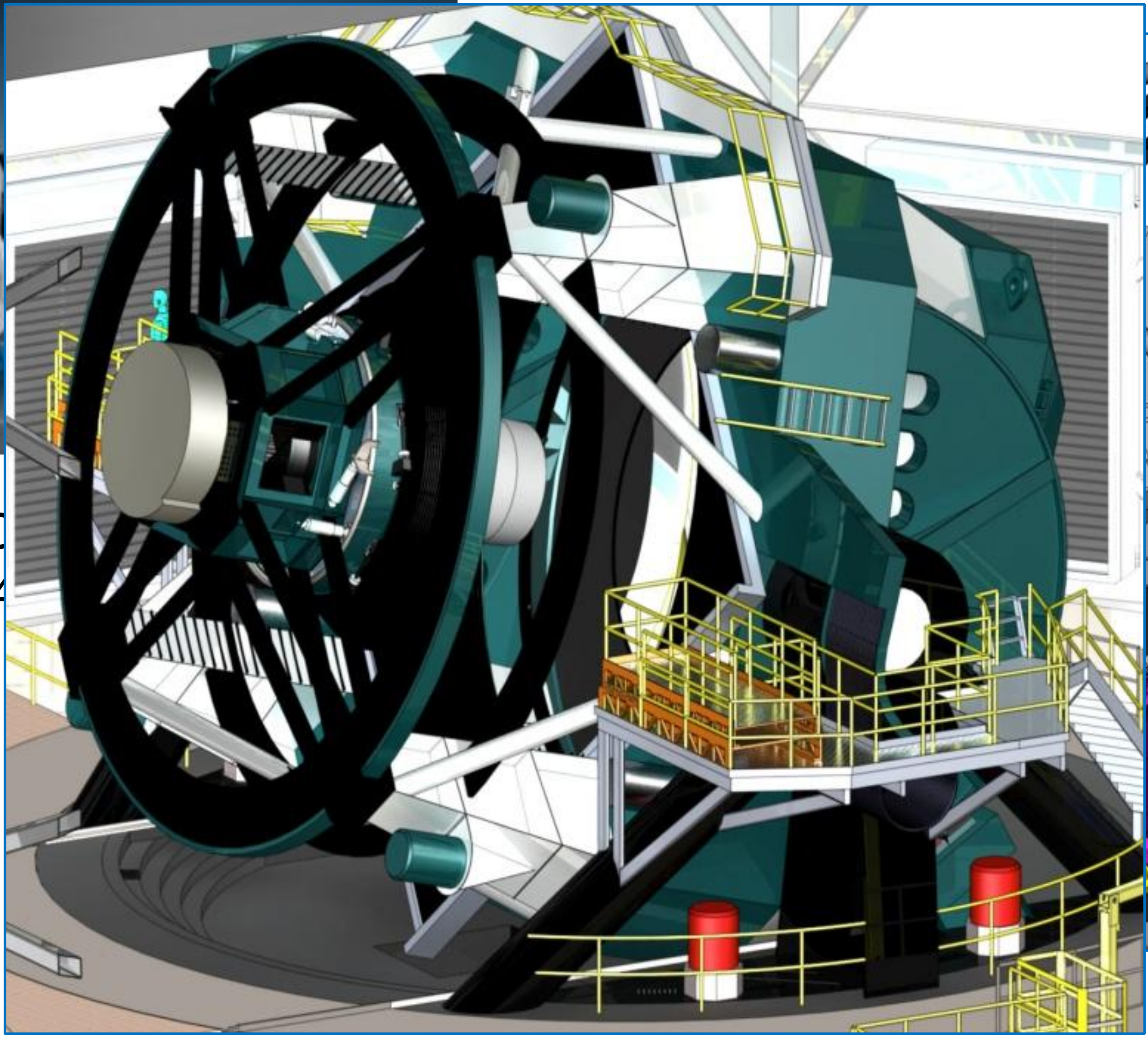
Survey Trends

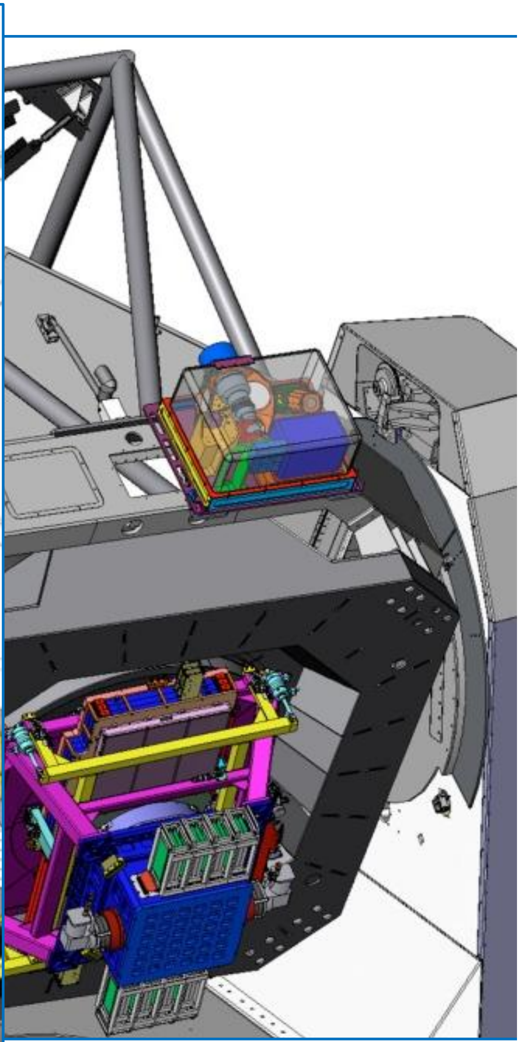


T.Tyson (2010)



SD
2.4





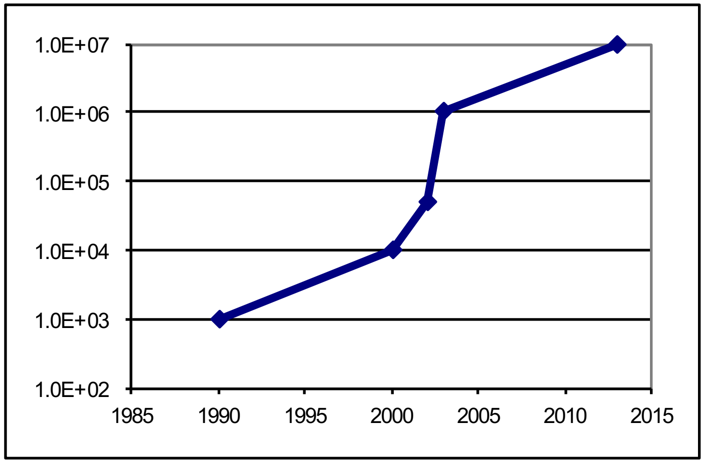
LSST
8.4m 3.2Gpixel

PanSTARRS
1.8m 1.4Gpixel

CMB Maps

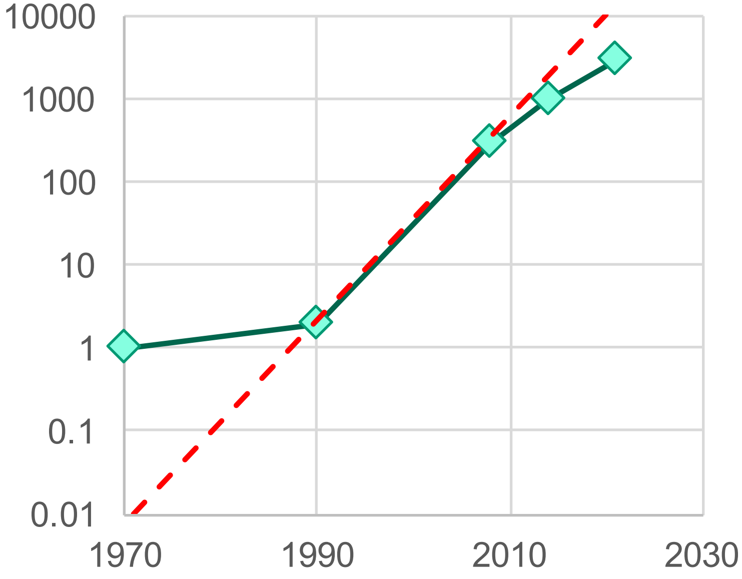
1990	COBE	1000
2000	Boomerang	10,000
2002	CBI	50,000
2003	WMAP	1 Million
2013	Planck	10 Million

A factor of 55 per decade



Angular Surveys

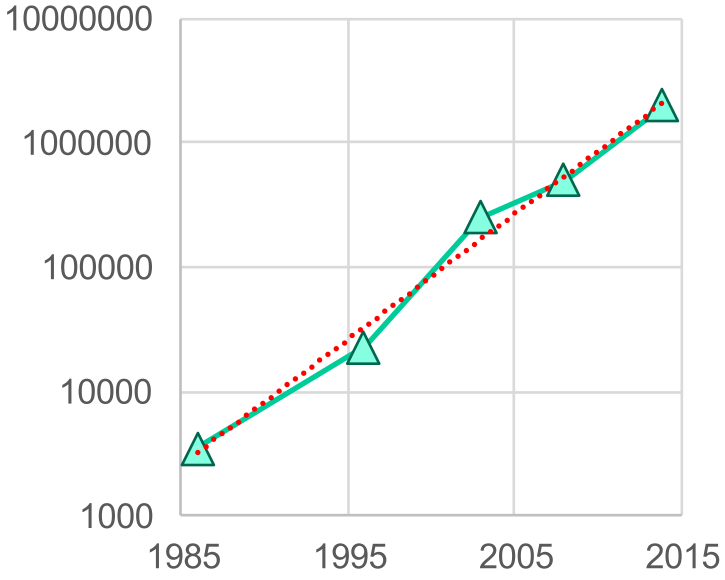
1970	Lick	1M
1990	APM	2M
2008	SDSS	300M
2014	PS1	1000M
2021	LSST	3000M



A factor of 17 per decade

Local Redshift Surveys

1986	CfA	3500
1996	LCRS	23000
2003	2dF	250000
2008	SDSS	500000
2014	BOSS	2000000

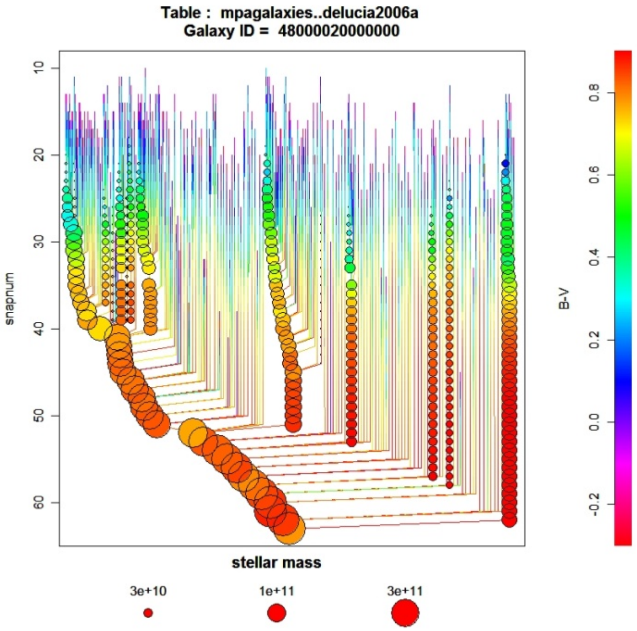


A factor of 10 per decade

Cosmology Simulations

- Simulations are becoming an instrument on their own
- Millennium DB is the poster child/ success story
  - *Built by Gerard Lemson (now at JHU)*
  - *600 registered users, 17.3M queries, 287B rows*

    http://gavo.mpa-garching.mpg.de/Millennium/
  - *Dec 2012 Workshop at MPA: 3 days, 50 people*
- Data size and scalability
  - *PB data sizes, trillion particles of dark mat*
- Value added services
  - *Localized*
  - *Rendering*
  - *Global analytics*



Emerging Challenges

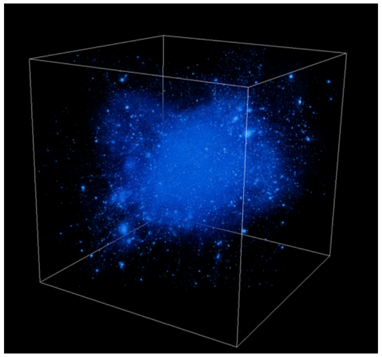
- Data size and scalability
  - *PB, trillion particles, dark matter*
  - *Where is the data located, how does it get there*
- Value added on-demand services
  - *Localized (SED, SAM, star formation history, resimulations)*
  - *Rendering (viz, lensing, DM annihilation, light cones)*
  - *Global analytics (FFT, correlations of subsets, covariances)*
  - *Spatial queries*
- Data representations
  - *Particles vs hydro*
  - *Particle tracking in DM data*
  - *Aggregates, summary of uncertainty quantification (UQ)*
  - *Covariances, ensemble averages*

Science is Changing

THOUSAND YEARS AGO
science was empirical
describing natural phenomena

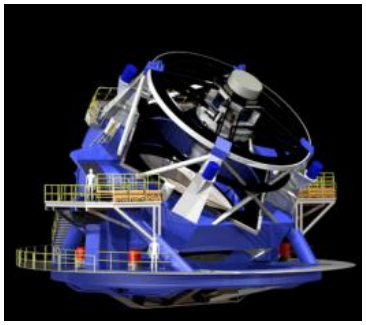
LAST FEW HUNDRED YEARS
theoretical branch using models,
generalizations

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



LAST FEW DECADES
a computational branch simulating
complex phenomena

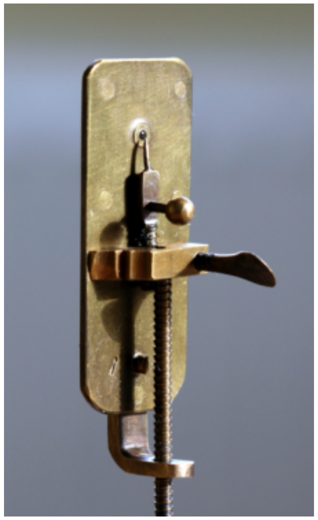
TODAY
data intensive science, synthesizing theory,
experiment and computation with statistics
►new way of thinking required!



Non-Incremental Changes

- Multi-faceted challenges in the analysis as well
- 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



Why Is Astronomy Interesting?

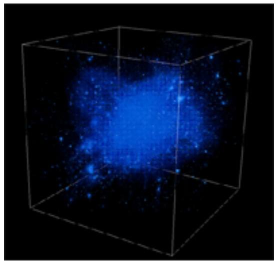
- **Astronomy has always been data-driven…. now this is becoming more accepted in other areas as well**

“Exciting, since it is ***worthless****!”*

— Jim Gray

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)
- Changing sociology – surveys analyzed by individuals
- From hypothesis-driven to data-driven science
- We need new instruments: “microscopes” and “telescopes” for data
- Need to start thinking about how to collect less data…
- There is a major challenge on the “long tail”
- A new, Fourth Paradigm of Science is emerging…
- Sky Surveys have been at the cusp of this transition