Big Data in Astronomy: Introduction to Large Surveys

AST 597 B

mjuric@astro.washington.edu

Today

- Administrivia
- Introduction to Large Surveys

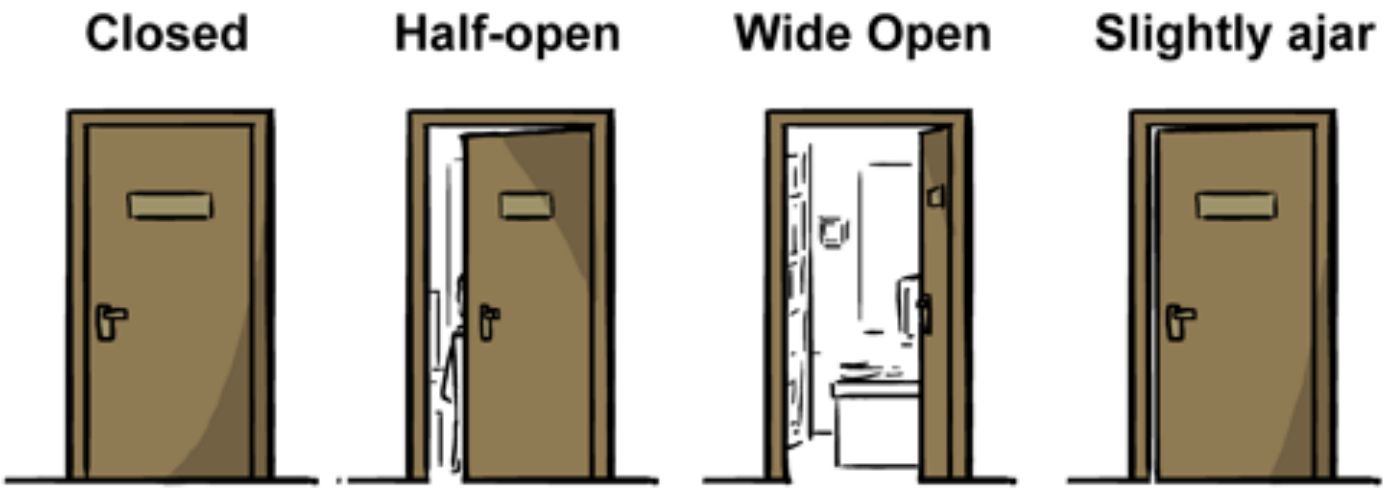
About Me

- Mario Juric (mar-ee-oh you-rich)
- What I do:
  - Science derived from large surveys: Galactic structure, properties of the solar system
  - Astronomical algorithms and software research
  - Optical astronomer (=> some bias towards optical surveys)
- Hats I wear
  - Professor @ Astronomy
  - Sr. Fellow @ UW's eScience Institute
  - Data Management Project Scientist @ LSST

About Me

- Nominal Coordinates:
  - Office: C320
  - E-mail: mjuric@astro.washington.edu
  - Open door policy: stop by any time!

How to read a Professor's door



WWW.PHDCOMICS.COM

About This Course

- The goal of this course is to prepare you for research with large survey data, teach you how to think about such data sets, and give you an overview of what is or soon will be available.
- We will:
  - Learn about how large observational data sets are changing astronomy
  - Introduce tools and techniques for working with large data sets (incl. access, analysis, and visualization)
  - Introduce and work hands-on with existing data sets (including SDSS, WISE, Kepler and others)
  - Learn about major upcoming surveys (including PanSTARRS, Gaia, and LSST)

The Goal

- Late March 2015 (reading astro-ph):

  "Oh, this is an interesting paper! Looks like if I combined IR and optical observations I could find these kinds of objects… But wait, I know there's WISE and SDSS! I'll run a SQL query against those databases, use LSD to join the result, analyze the data in Python using scikits tools, and send the result to my collaborator as an IPython notebook via github.

- You'll know what's out there, what it's good for, and how to use it.

Learning about Learning

- In survey astronomy (and related computing disciplines), things change too quickly to be learned only once – it's a process that never ends.

- The major emphasis of this course will be on showing how to independently find information, evaluate, and use it.

Mailing List

- This is a topics course, and many of you are auditing
- If you're not enrolled, e-mail me your e-mail address so I can add you to the class mailing list
  - mjuric@astro.washington.edu
- All announcements, etc., will go via that list

Lectures

- MW, 3:30pm-4:50pm
  - Note: Next class – merged with ASTR 511, Thursday, 1:45pm
- Q: Does this work for everyone?

- Lecture structure:
  - At most 1hr for introducing new materials, etc.
  - At least 20 minutes for open-ended discussion
  - Interrupt and ask questions at any time!

Course Materials

- I'll be adding most of what we need to the following repository on GitHub:

https://github.com/mjuric/ast597b

Smile, You're On Camera!

- Actually, it's mostly me who's on camera…
- Actually, it's mostly my slides that will be on camera…

- Joining the class live via Google Hangouts: http://goo.gl/9oTASp
- Viewing the recording later, via YouTube
  - I'll e-mail the link after class
  - Note: recorded classess will be private, and won't be kept after this quarter ends.

Syllabus

- https://github.com/mjuric/ast597b

- Three parts
  - Software / Data Analysis Tools (~ January)
  - Major Surveys (~ February)
  - Future Surveys and Data Challenges (~March)

Grading, Homeworks, Etc.

- Two homeworks (50% of the grade):
  - All homeworks will be IPython notebooks, turned in via GitHub
  - HW #1, due Feb 11, focused on software tools
  - HW #2, due Feb 25, focused on survey data analysis

- Paper presentations (25% of the grade):
  - Present a paper on a survey / dataset of your choosing, making sure to cover:
    - The goals of the survey
    - The data it contains (and any caveats about it)
    - What it may be useful for
    - The best ways to obtain it

- Final project (25% of the grade):
  - More about this in a week or so ☺.

Questions?

- Large Survey and Why They're Different

sur·vey

verb
/sərˈvā/

1. (of a person or their eyes) look carefully and thoroughly at (someone or something), especially so as to appraise them.
   "her green eyes surveyed him coolly"
   synonyms: look at, look over, observe, view, contemplate, regard, gaze at, stare at, eye; More
2. examine and record the area and features of (an area of land) so as to construct a map, plan, or description.
   "he surveyed the coasts of New Zealand"

noun
/ˈsərˌvā/

1. a general view, examination, or description of someone or something.
   "the author provides a survey of the relevant literature"
   synonyms: study, review, consideration, overview; More
2. an act of surveying an area of land.
   "the flight involved a detailed aerial survey of military bases"

Translations, word origin, and more definitions

Hipparchus of Rhodes (180-125 BC)

Discovered the precession of the equinoxes.

Measured the length of the year to ~6 minutes.

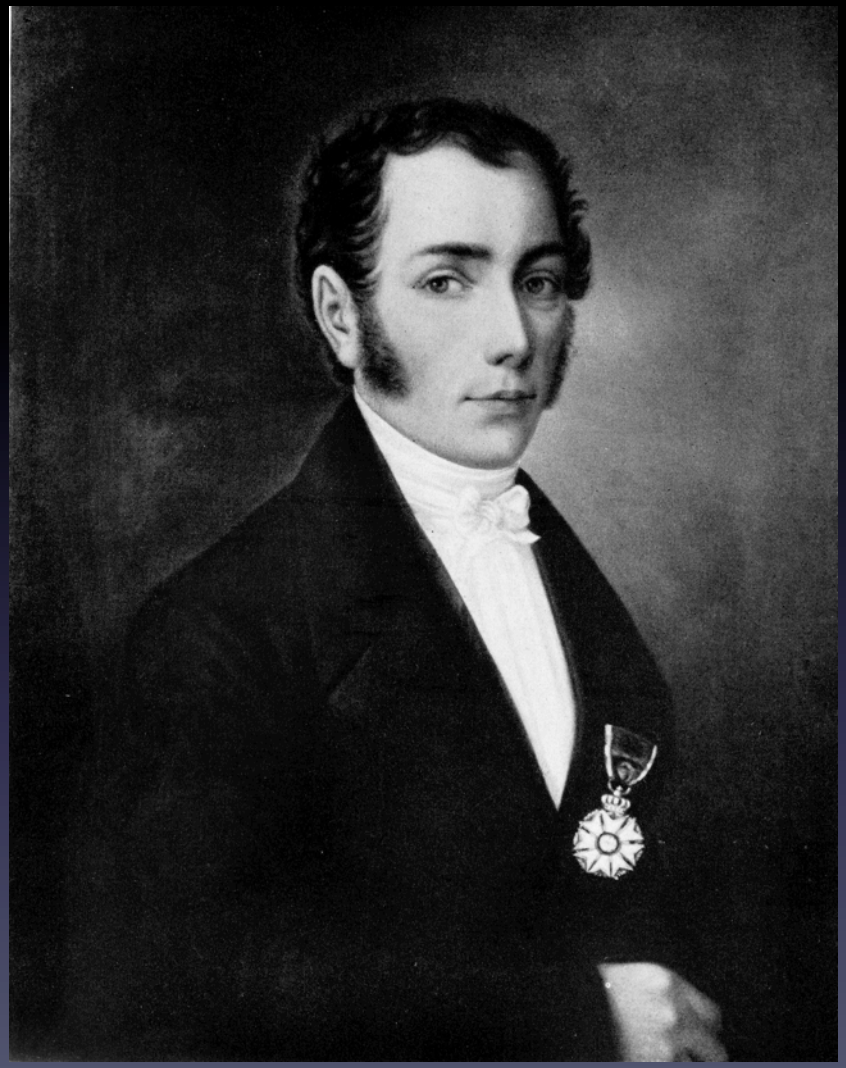
In 129 BC, constructed one of the first star catalogs*, containing about 850 stars.

n.b.: also the one to blame for the magnitude system …

Galileo Galilei (1564-1642)

Researched a variety of topics in physics, but called out here for the introduction of the *Galilean telescope*.

Galileo's telescope allowed us for the first time to *zoom in* on the cosmos, and study the individual objects in great detail.



Joseph von Fraunhofer (1787-1826)

Mounted a prism in front of an objective of a small telescope, and pointed it to the Sun.

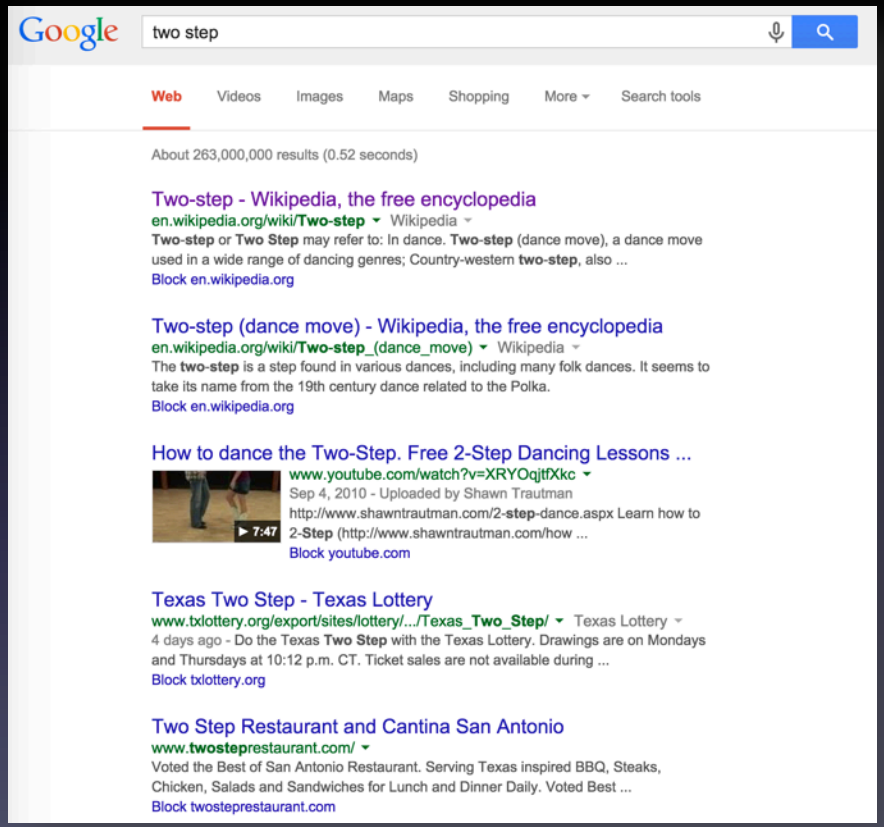
In 1859, Kirchhoff and Bunsen understood what Fraunhofer saw.

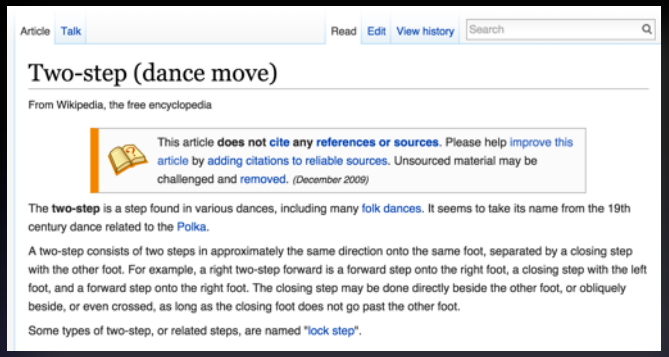
The birth of modern astrophysics!

The Astrophysics Two-Step

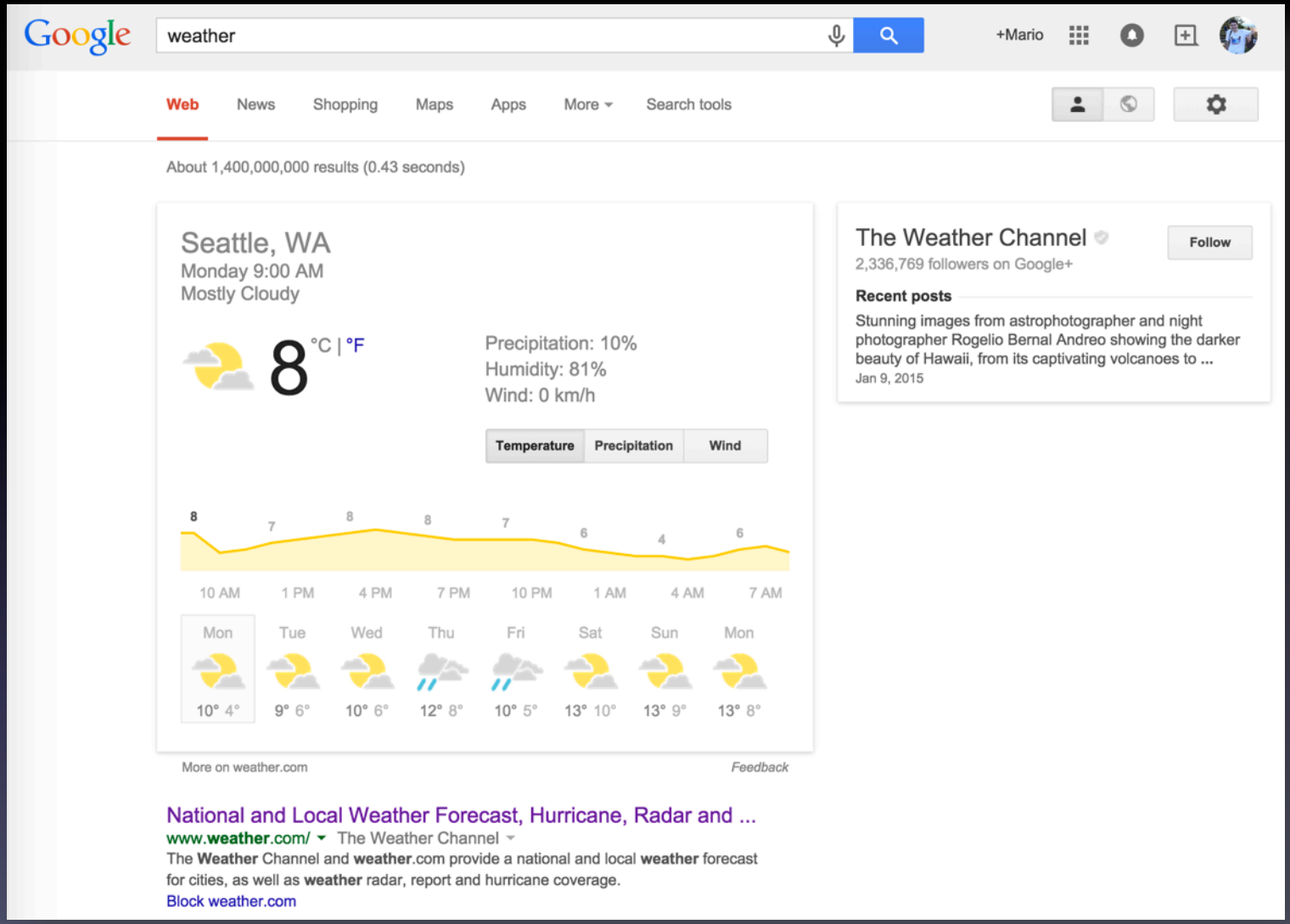
- Surveys
  - Construct catalogs and maps of objects in the sky. Focus on coarse classification and discovering targets for further follow-up.
- Large telescopes
  - Acquire detailed observations of a **few** representative objects. Understand the details of astrophysical processes that govern them, and **extrapolate that understanding to the entire class**.

Analogy: Google Search





Google's index is a catalog of the Web. We use it to "zoom in" on individual entries to find out more.



But, it's more than just a catalog of pointers – more and more, Google itself collects, processes, indexes, visualizes, and serves the actual information we need.

More and more often, our "research" begins and ends with Google!

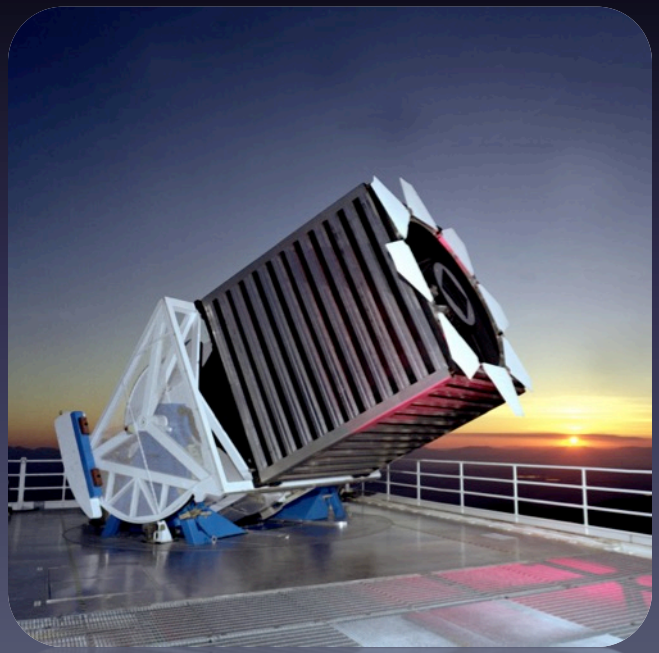
Modern Large Astronomical Surveys are the Googles of the Sky

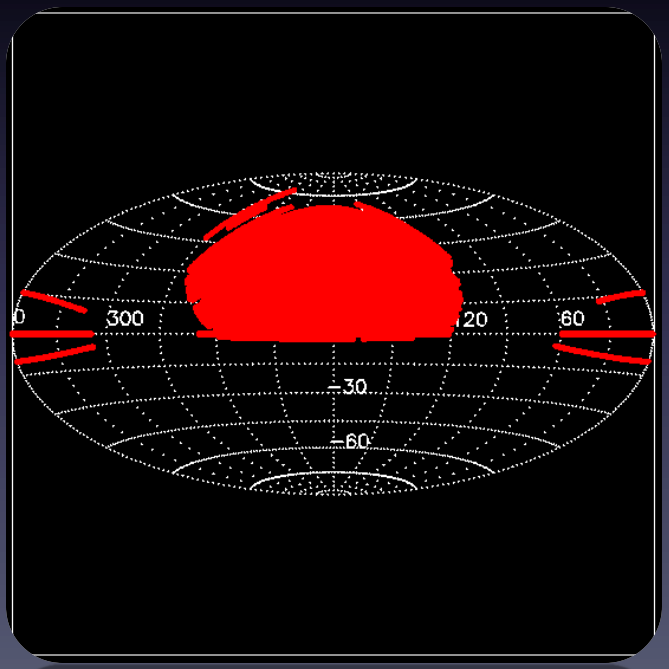
- There's a close parallel with large surveys in astronomy, in scale, quality, and richness of the collected information
  - Scale: We're entering the era when we can catalog the entire sky
  - Quality: Those catalogs will be as precise as the measurements taken with "pointed" observations (used to be ~5-10x worse)
  - Richness: Those catalogs contain not only positions and magnitudes, but also shapes, profiles, and temporal behavior of the objects.
- Quite often, the research begins and ends with the survey.
- This is what makes large surveys of today not just bigger, but different. They're more than just "finding charts".

Sloan Digital Sky Survey

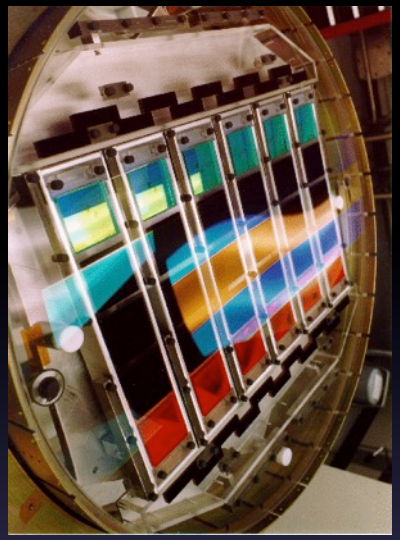
2.5m telescope 10000 deg^2 0.1” astrometry r<22.5 flux limit

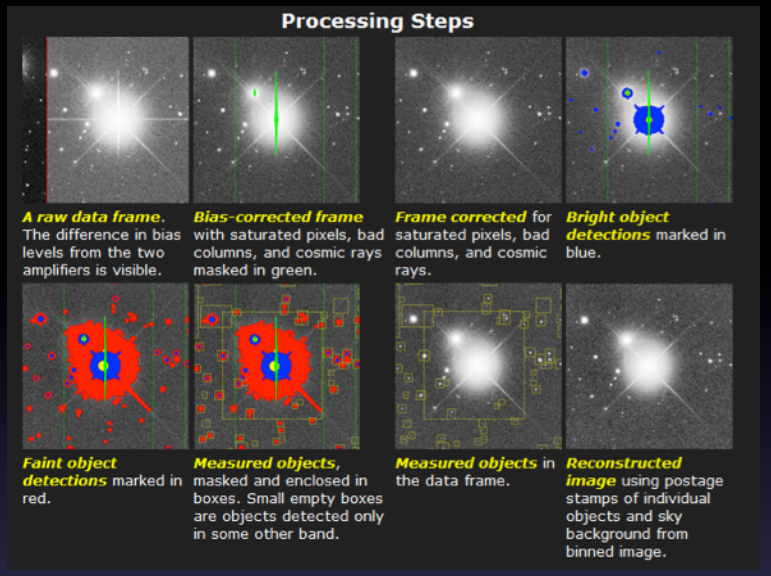
5 band, 2%, photometry for >50M stars, >280k R=2000 stellar spectra

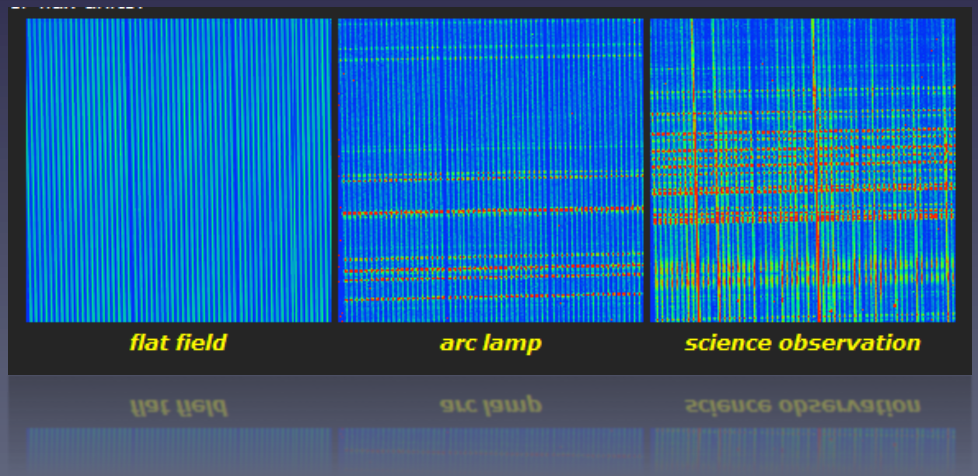


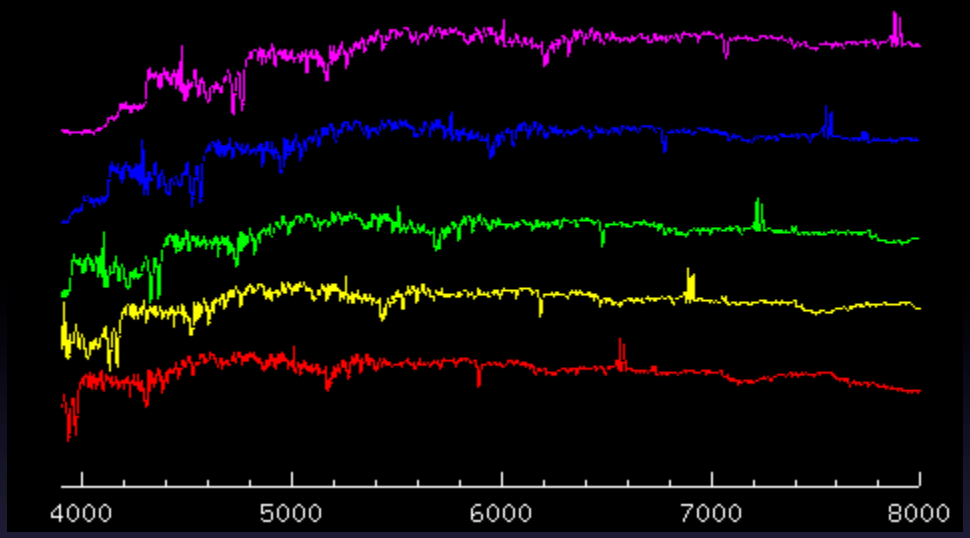


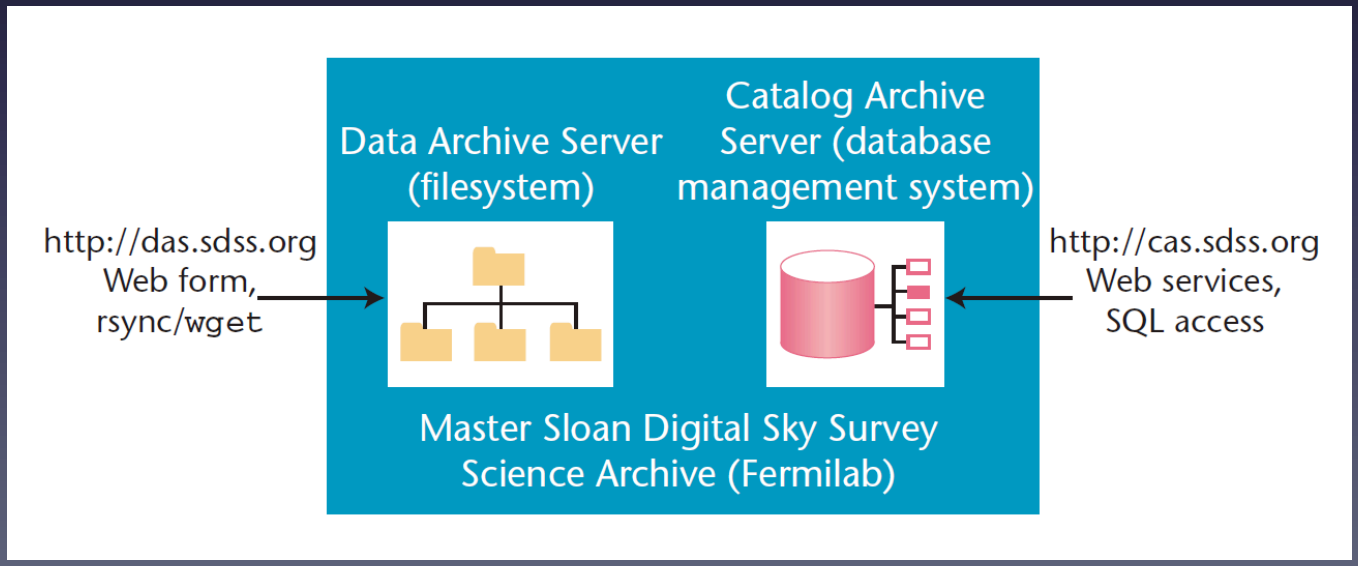
Observing With SDSS








4000
5000
6000
7000
8000


Data Archive Server (filesystem)
Catalog Archive Server (database management system)
http://das.sdss.org
Web form,
rsync/wget
http://cas.sdss.org
Web services,
SQL access
Master Sloan Digital Sky Survey Science Archive (Fermilab)

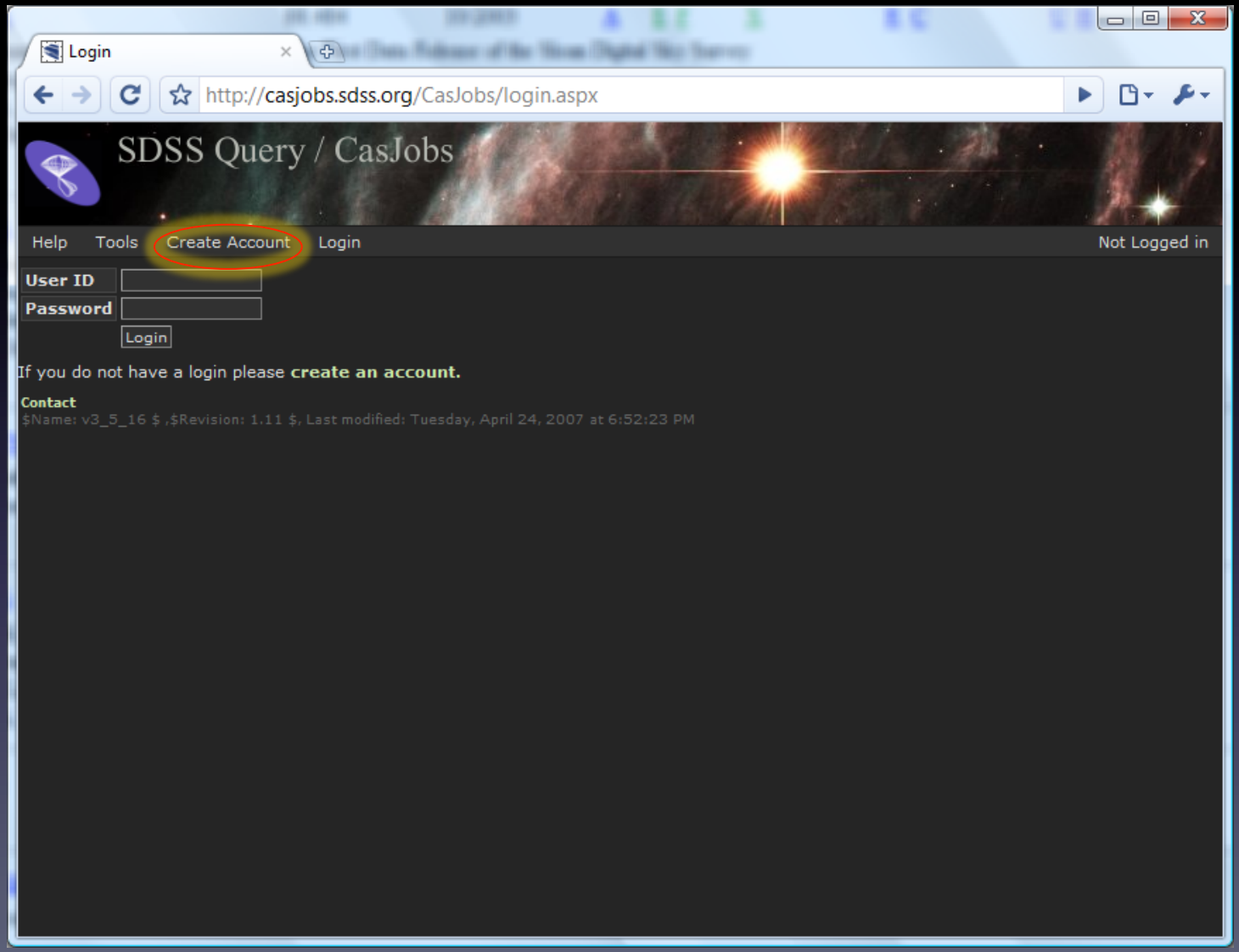
Thakar (2006)

Table 1. Sloan Digital Sky Survey public data releases.

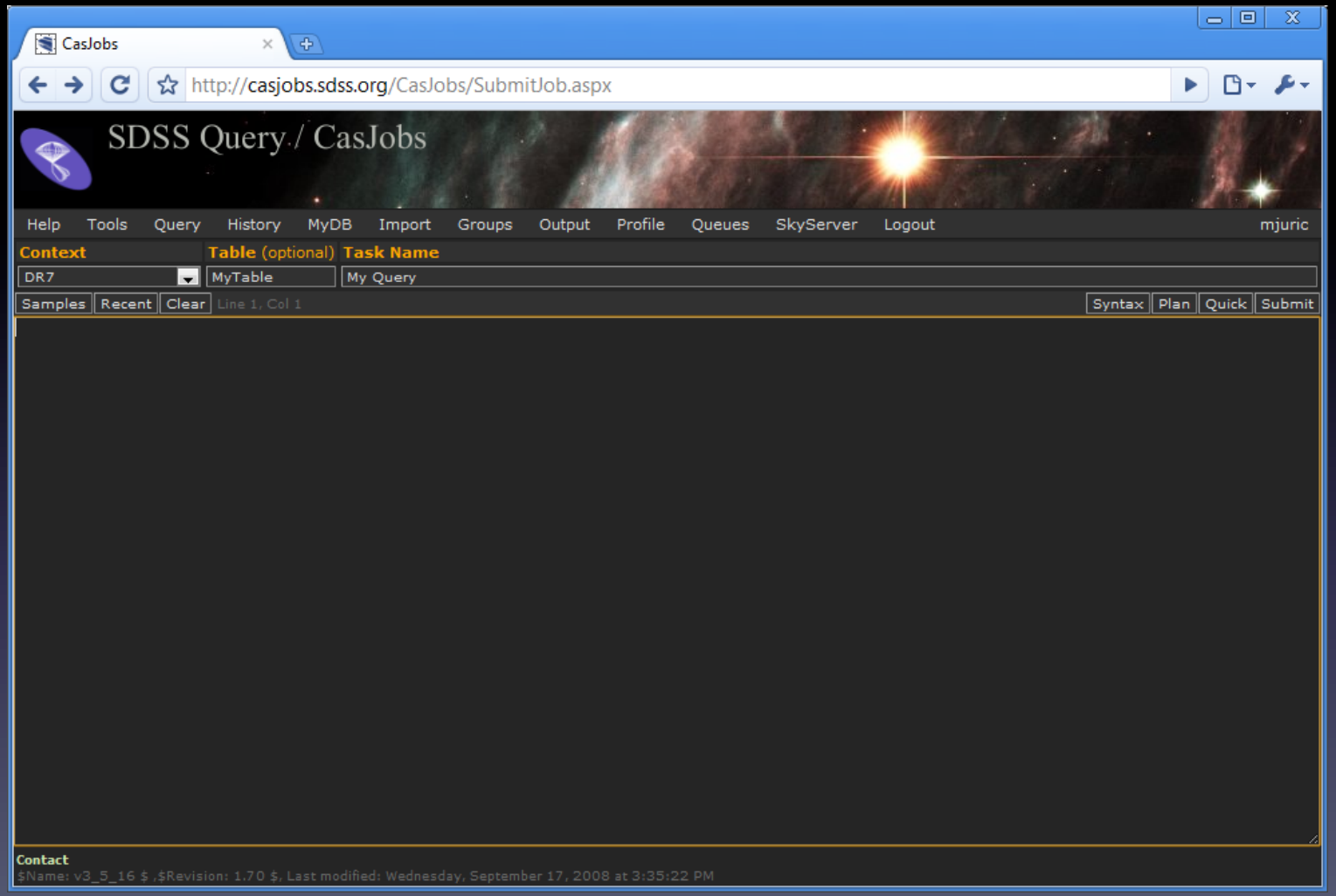
Release	Date	Size (catalogs)	Images (millions)	Spectra (thousands)	Distribution/mirrors
EDR[2]	June 2001	200 Gbytes	14	54	Johns Hopkins Univ. (JHU), San Diego Supercomputing Center (SDSC), UK, Japan
DR1[6]	June 2003	1 Tbyte	53	186	JHU, SDSC, Canadian Astronomical Data Centre (CADC), Univ. of Pittsburgh, UK, Germany, Japan, India
DR2[5]	Mar. 2004	2 Tbytes	88	330	JHU, Univ. of Pittsburgh, SDSC, Germany
DR3[3]	Sept. 2004	3 Tbytes	141	478	JHU, UK, India
DR4[4]	June 2005	4 Tbytes	180	608	JHU, Germany, Hungary, Brazil
DR5[6]	June 2006	5 Tbytes	215	738	JHU, India, Russia, Hungary, Australia
DR6[7]	June 2007	6 Tbytes	287 M	1.27 M	JHU, India, Hungary

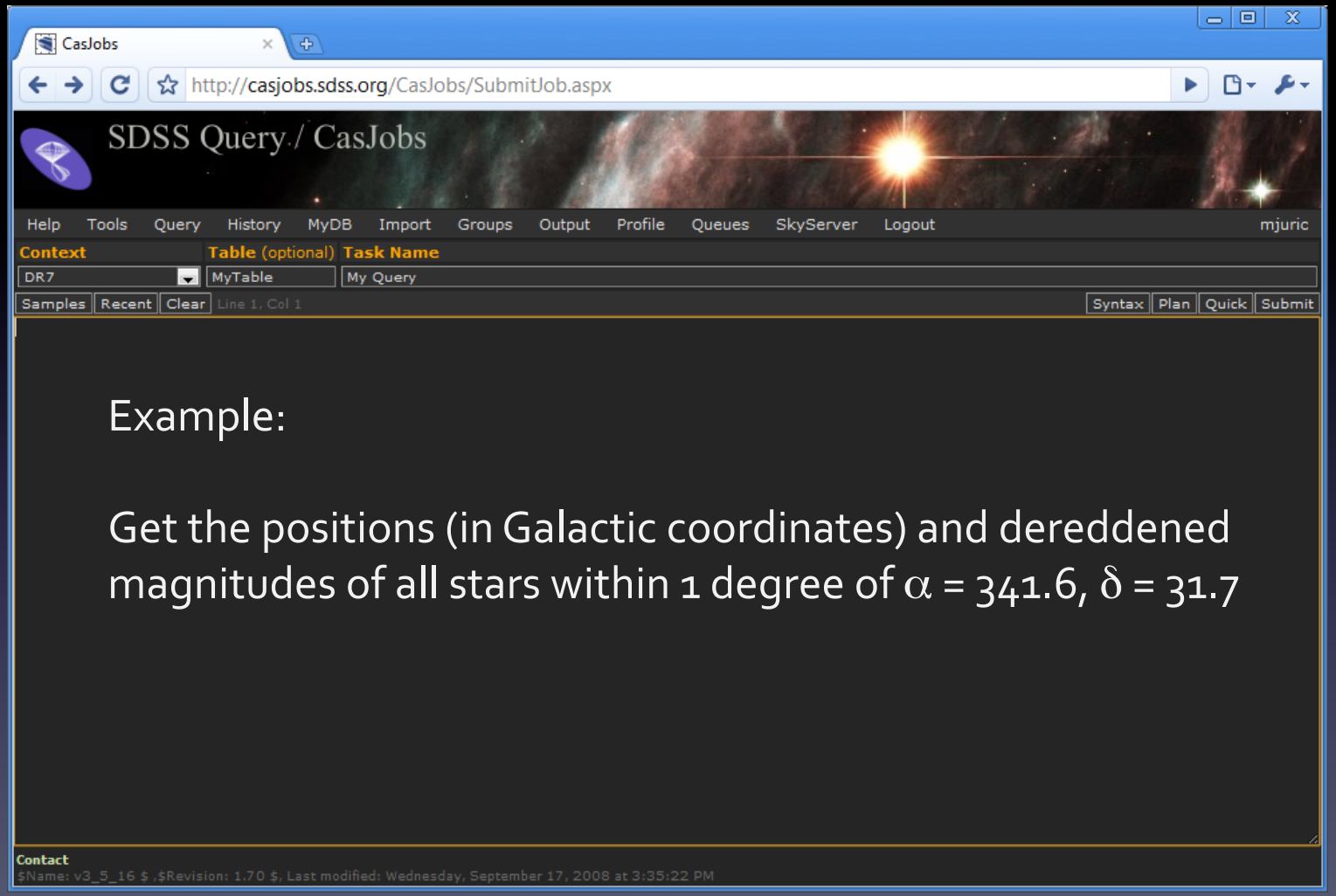

SDSS SkyServer DR7
http://cas.sdss.org/dr7/en/
Sloan Digital Sky Survey / SkyServer
SDSS
Home
Tools
Schema
Projects
Astronomy
SDSS
Contact Us
Download
Site Search
Help
Welcome to the DR7 site!!!
This website presents data from the Sloan Digital Sky Survey, a project to make a map of a large part of the 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...
SDSS is supported by
SkyServer Tools
Famous places
Get images
Visual Tools
Explore
Search
Object Cross-ID
CasJobs
Science Projects
Basic
Advanced
Challenges
For Kids
Games and Contests
Teachers
Links to other projects
Info Links
About Astronomy
About the SDSS
About the SkyServer
SDSS Data Release 7
SDSS Project Website
Open SkyQuery
Images of RC3 Galaxies
Help
Getting Started
FAQ
How To
Glossary
Schema Browser
Sample SQL Queries
Details of SDSS Data
Powered by
Microsoft
Site Traffic
Privacy Policy
Contact Us

CasJobs: Running long and complex queries



SDSS CasJobs Service

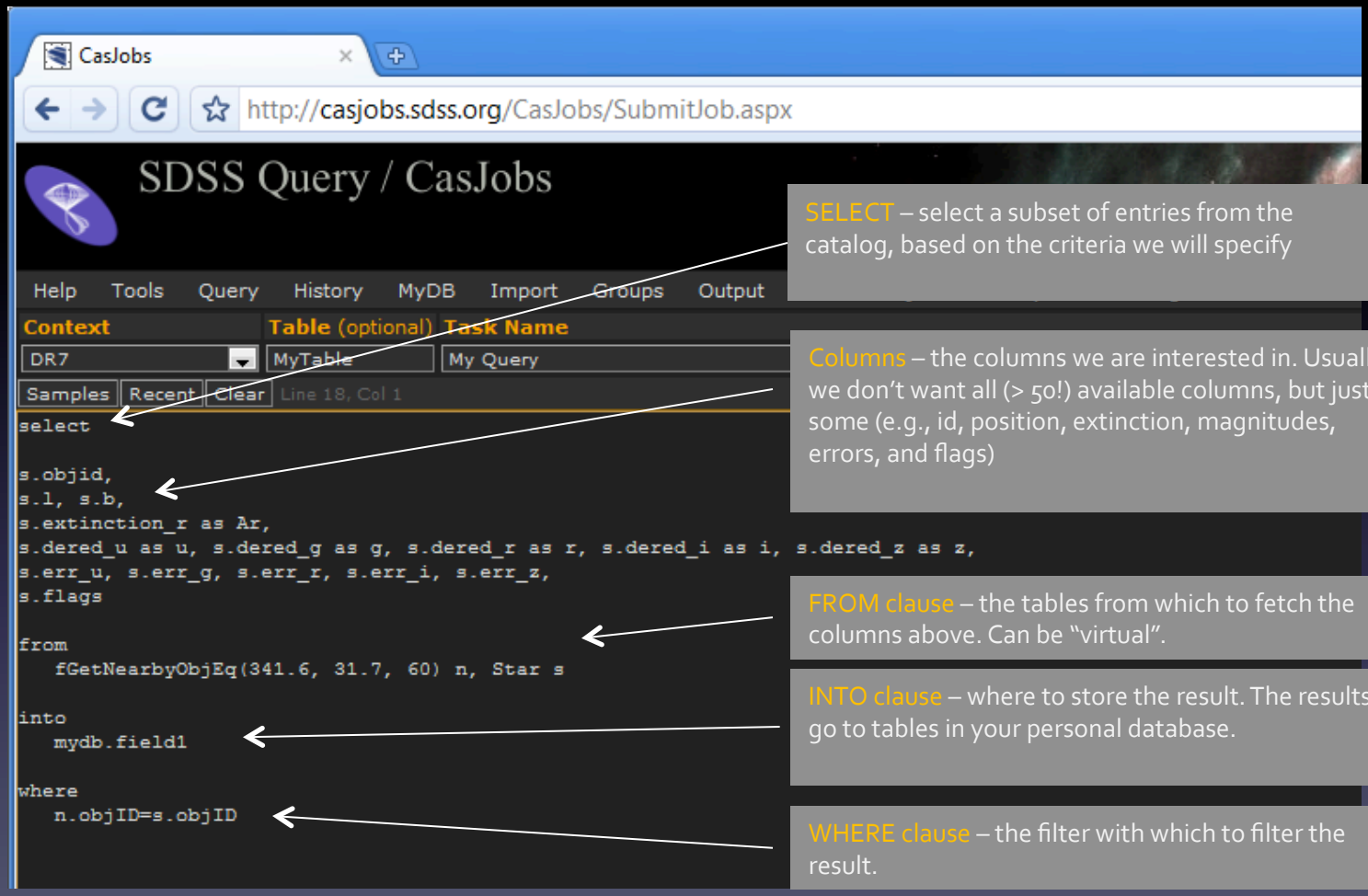



CasJobs
http://casjobs.sdss.org/CasJobs/SubmitJob.aspx
SDSS Query / CasJobs
Help Tools Query History MyDB Import Groups Output Profile Queues SkyServer Logout mjuric
Context
Table (optional)
Task Name
DR7
MyTable
My Query
Samples Recent Clear Line 1, Col 1
Syntax Plan Quick Submit
Example:
Get the positions (in Galactic coordinates) and dereddened magnitudes of all stars within 1 degree of α = 341.6, δ = 31.7
Contact
$Name: v3_5_16 $,$Revision: 1.70 $, Last modified: Wednesday, September 17, 2008 at 3:35:22 PM

An SQL Query

SQL – Structured Query Language
A language for specifying and extracting subsets of data from large tabular databases.
Allows you to select only the data that satisfy a certain set of (possibly very complicated) restrictions that you specify

Anatomy of an SQL Query



SELECT – select a subset of entries from the catalog, based on the criteria we will specify

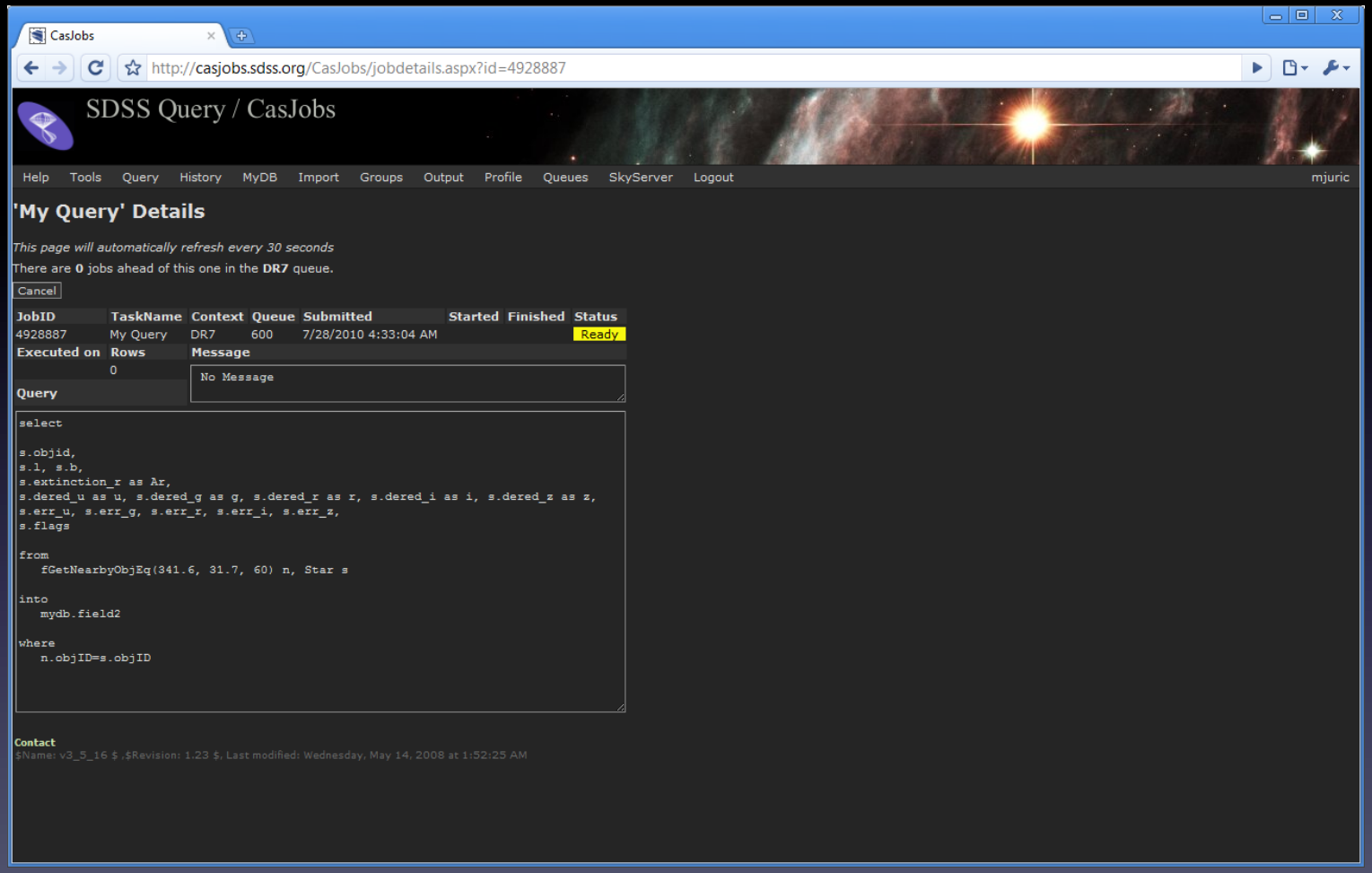
Columns – the columns we are interested in. Usually we don't want all (> 50!) available columns, but just some (e.g., id, position, extinction, magnitudes, errors, and flags)

FROM clause – the tables from which to fetch the columns above. Can be "virtual".

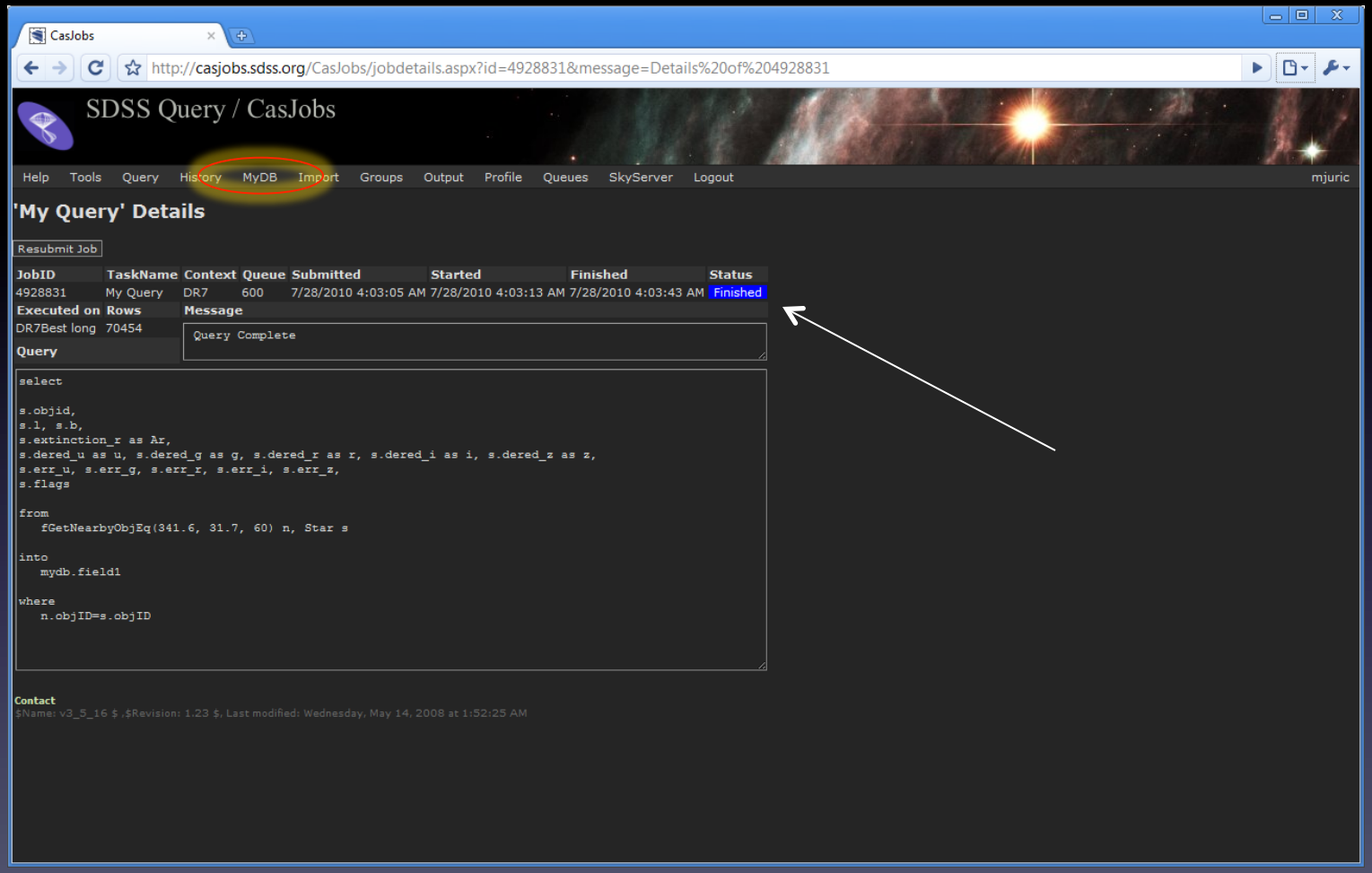
INTO clause – where to store the result. The results go to tables in your personal database.

WHERE clause – the filter with which to filter the result.

Submitted Job

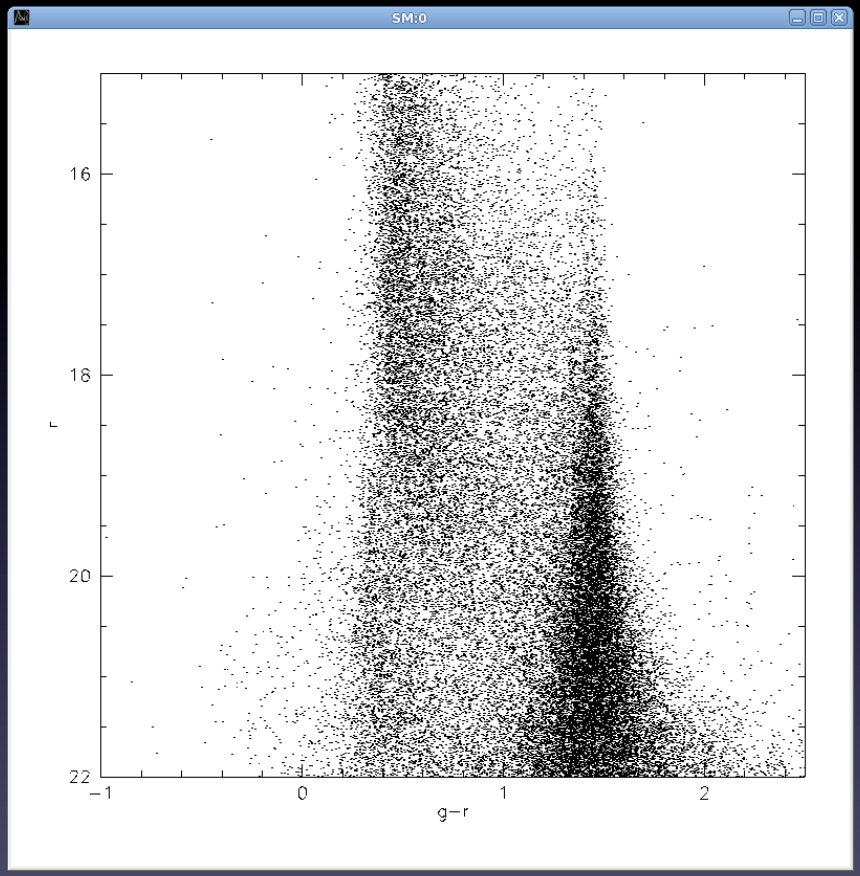


Completed Job




MyDB
miner - Google Search
http://casjobs.sdss.org/CasJobs/MyDB.aspx
SDSS Query / CasJobs
Help Tools Query History MyDB Import Groups Output Profile Queues SkyServer Logout mjuric
MyDB
Local Only
Views
Tables
Functions
Procedures
Sort by...
All selected...
Rows kB Name
70,454 9,160 field1
Mario Juric 's MyDB
9,624 kB of 2,000,000 kB used
From this page you can get various information about the contents of both your MyDB and shared tables within your groups. Click the left table links to get information about a specific table, such as rows, columns or size. From the table pages you can also perform various table-specific tasks, such as:
Download a table
Mangage your group tables
Rename a table
Drop a table
Sizes are approximations only.
Row counts are approximations only. For exact value run a count.
There's always some overhead, even empty MyDB's take up space.
Group tables do not count towards your MyDB size limit.
Contact
$Name: v3_5_16 $,$Revision: 1.64 $, Last modified: Tuesday, January 27, 2009 at 3:19:32 PM

Basic Data Analysis: A Color-Magnitude Diagram (CMD)



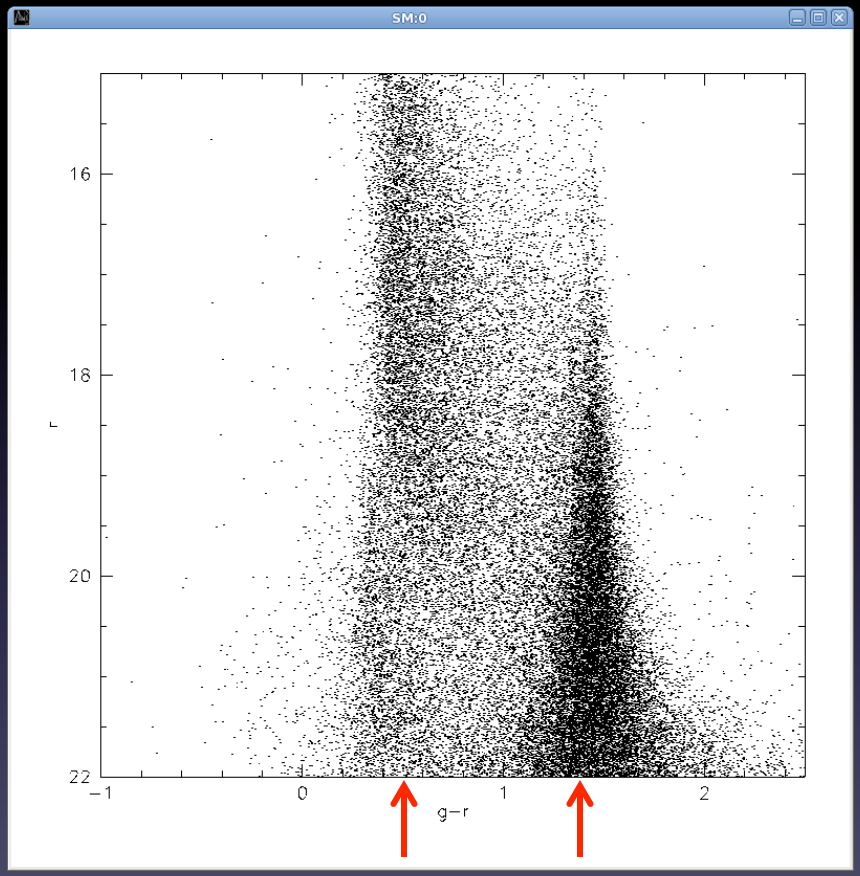
Untitled - Notepad

File Edit Format View Help

```
data "field1.csv"
read <1 2 b 3 Ar 4 u 5 g 6 r 7 i 8 z 9>

set gr = g-r
lweight 2
expand 1.001
limits -1 2.5 22 15
ptype 0 0
erase
box
points gr r
xlabel g-r
ylabel r
```

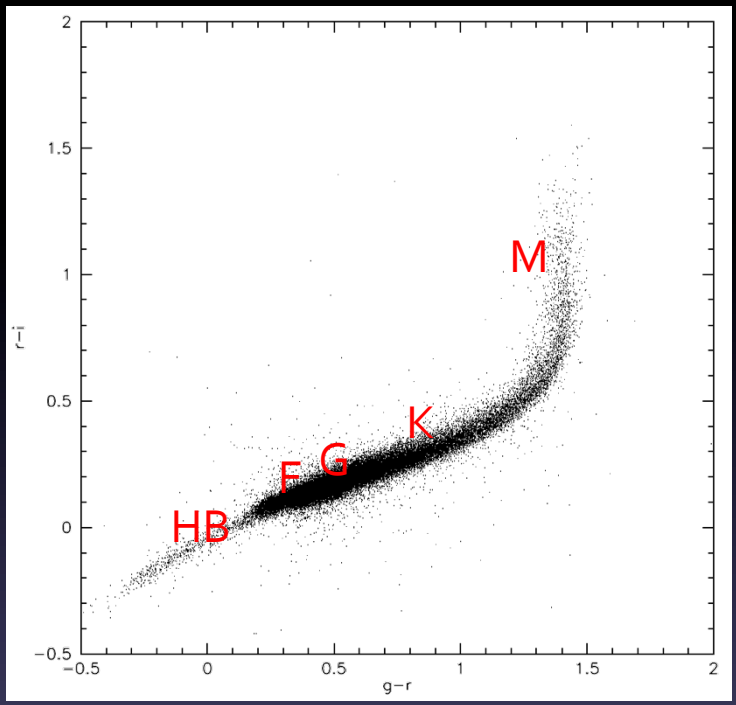
Interpreting Stellar CMDs



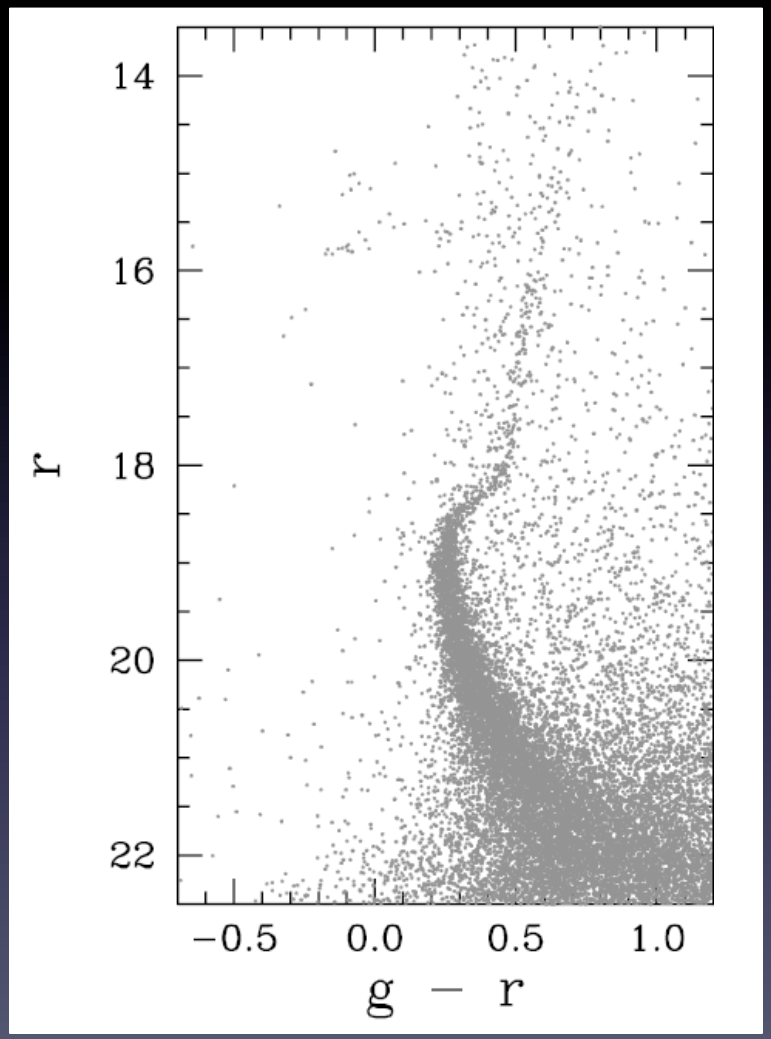
F stars M stars

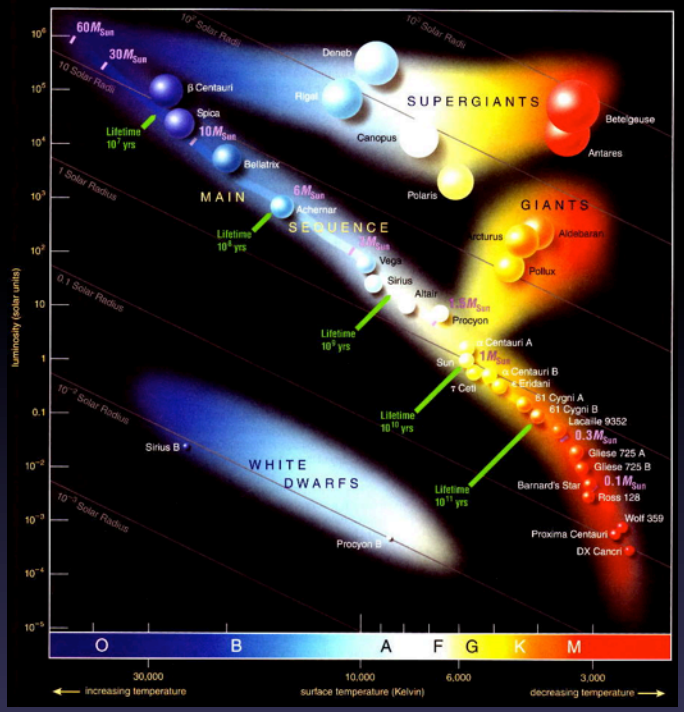
← More Luminous

Less Luminous →



A CMD of a Globular Cluster

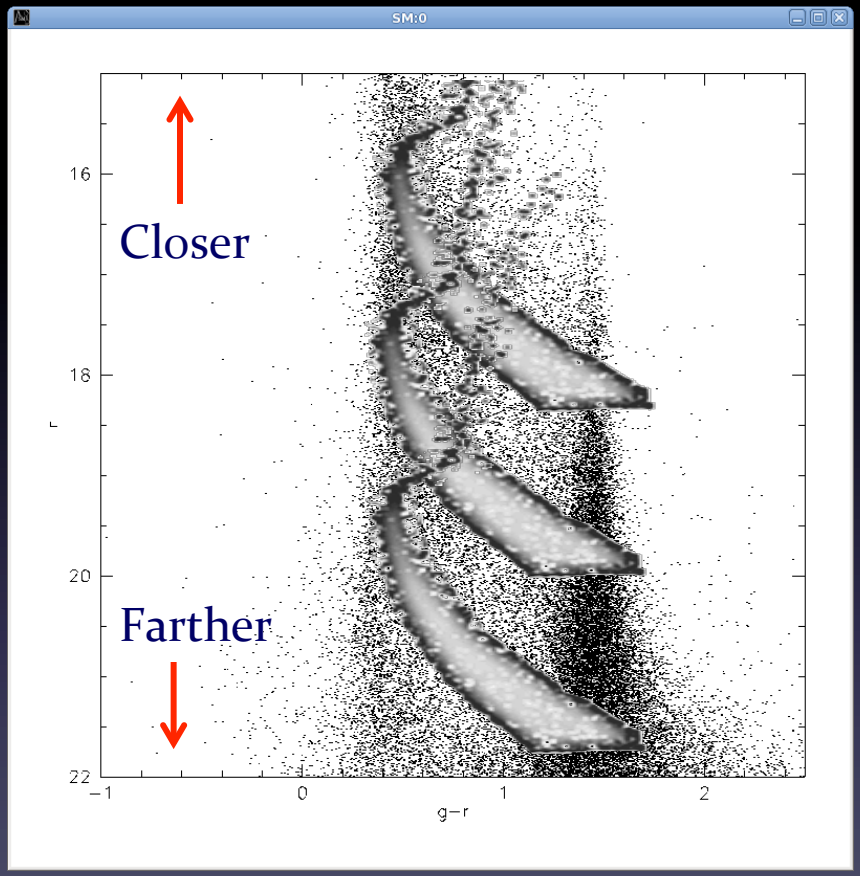


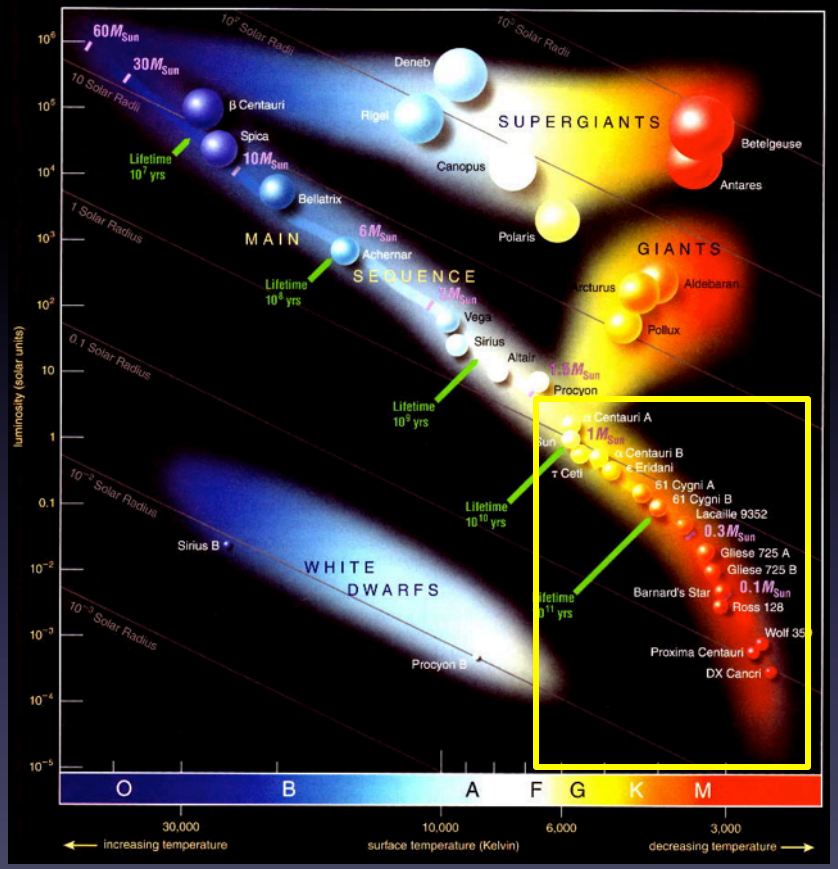


A globular cluster is a simple population (same age, same metallicity), at the same distance.

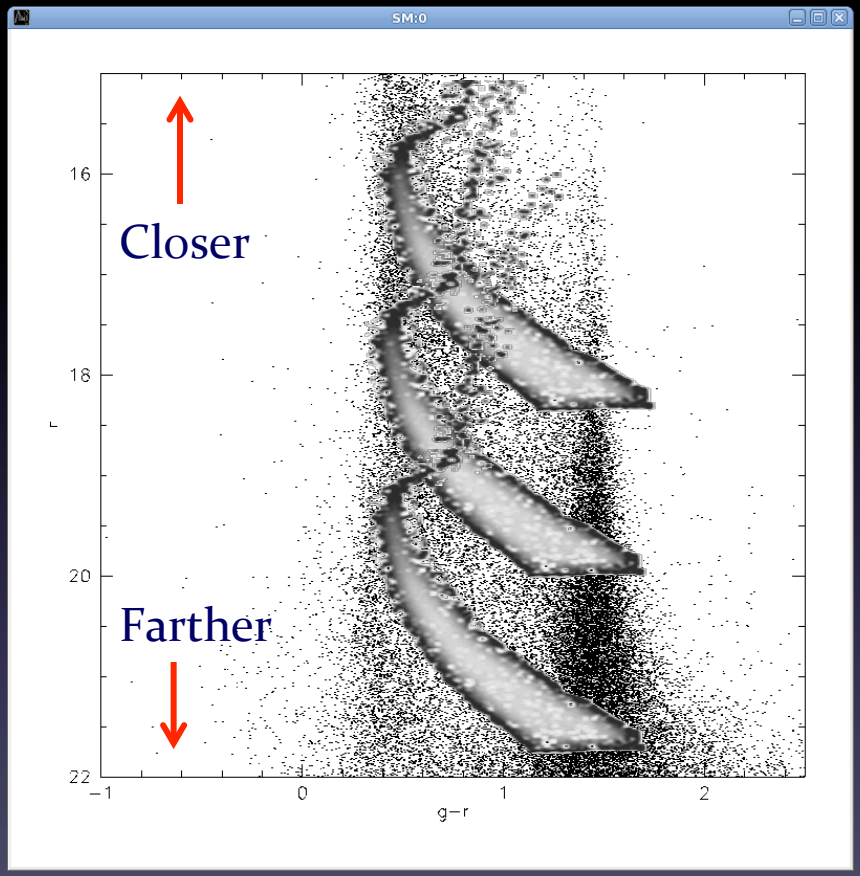
$$m = M + 5\log_{10}\left(d/10\text{kpc}\right)$$

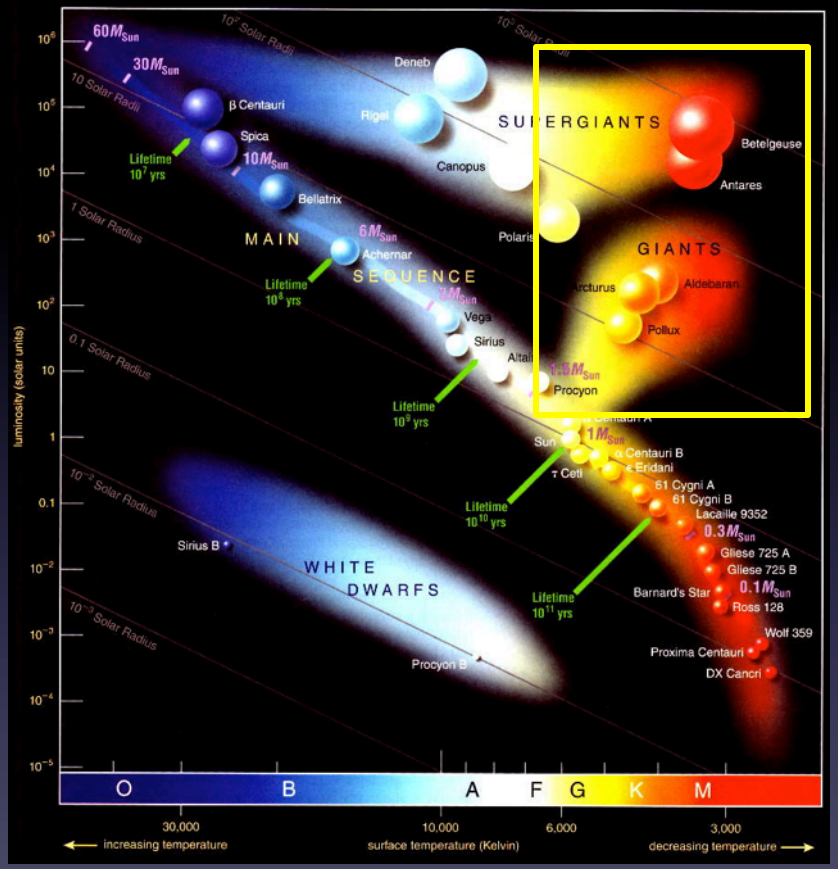
Interpreting "Field" CMDs





Interpreting "Field" CMDs

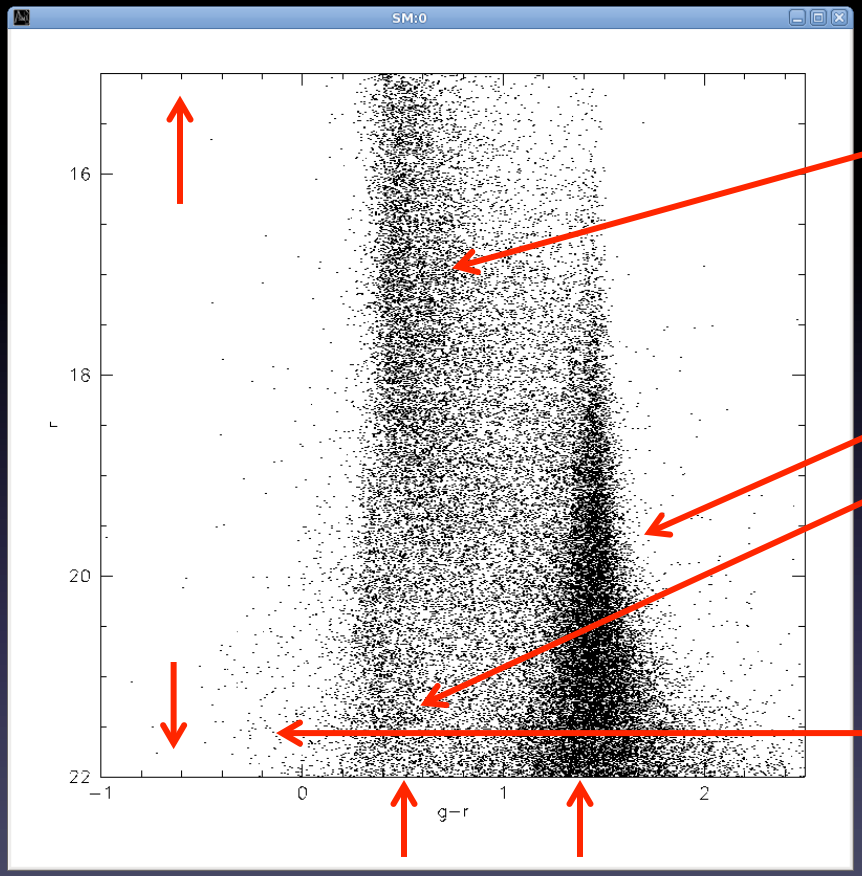




What about the giants? Yes, they are there, but significantly less numerous (<5%) than the main sequence stars.

The same applies for early type stars, white dwarfs, etc.

Basic Data Analysis: A Color-Magnitude Diagram

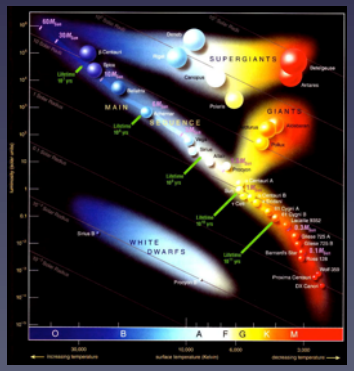


Disk G/F stars

Nearby Disk K/M stars

Halo stars

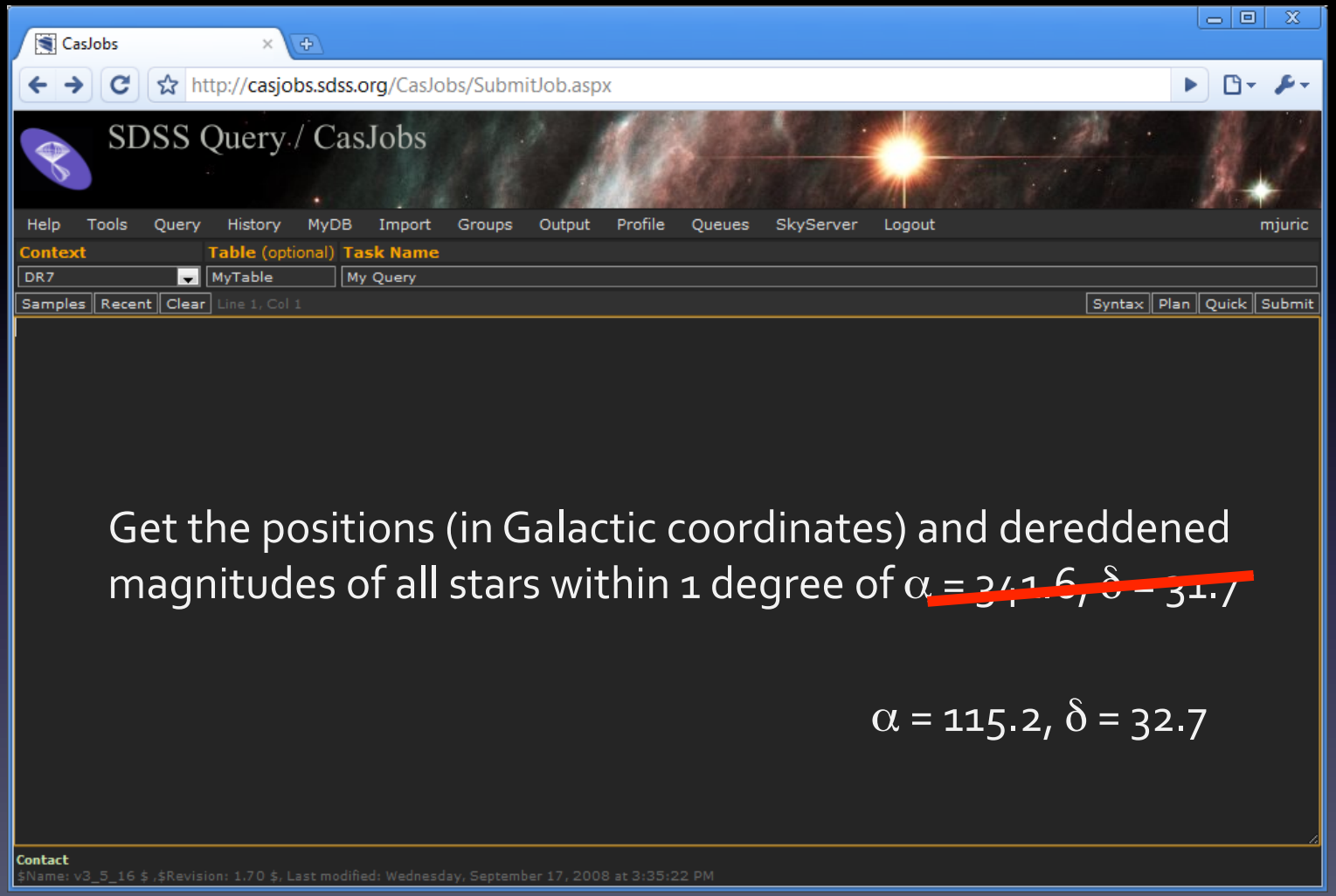
QSOs, RRLyrae, WDs, errors...



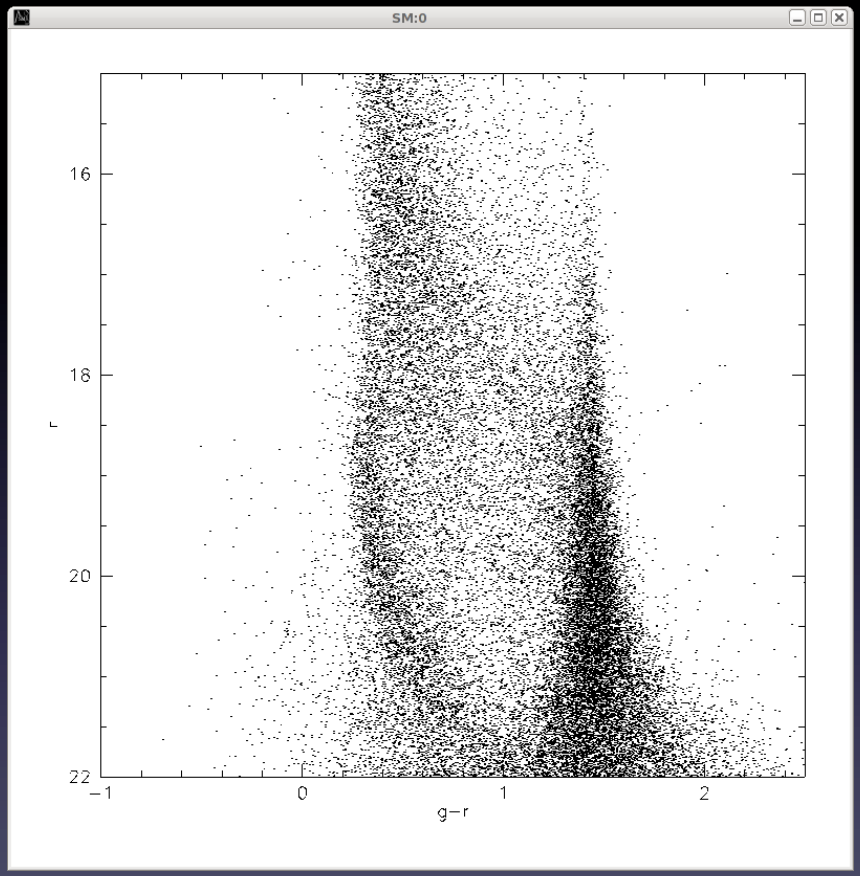
F stars M stars

← More Luminous

Less Luminous →


CasJobs
http://casjobs.sdss.org/CasJobs/SubmitJob.aspx
SDSS Query / CasJobs
Help Tools Query History MyDB Import Groups Output Profile Queues SkyServer Logout mjuric
Context
Table (optional)
Task Name
DR7
MyTable
My Query
Samples Recent Clear Line 1, Col 1
Syntax Plan Quick Submit
Get the positions (in Galactic coordinates) and dereddened magnitudes of all stars within 1 degree of α = 341.6, δ = 31.7
α = 115.2, δ = 32.7
Contact
$Name: v3_5_16 $,$Revision: 1.70 $, Last modified: Wednesday, September 17, 2008 at 3:35:22 PM

Color-Magnitude Diagram



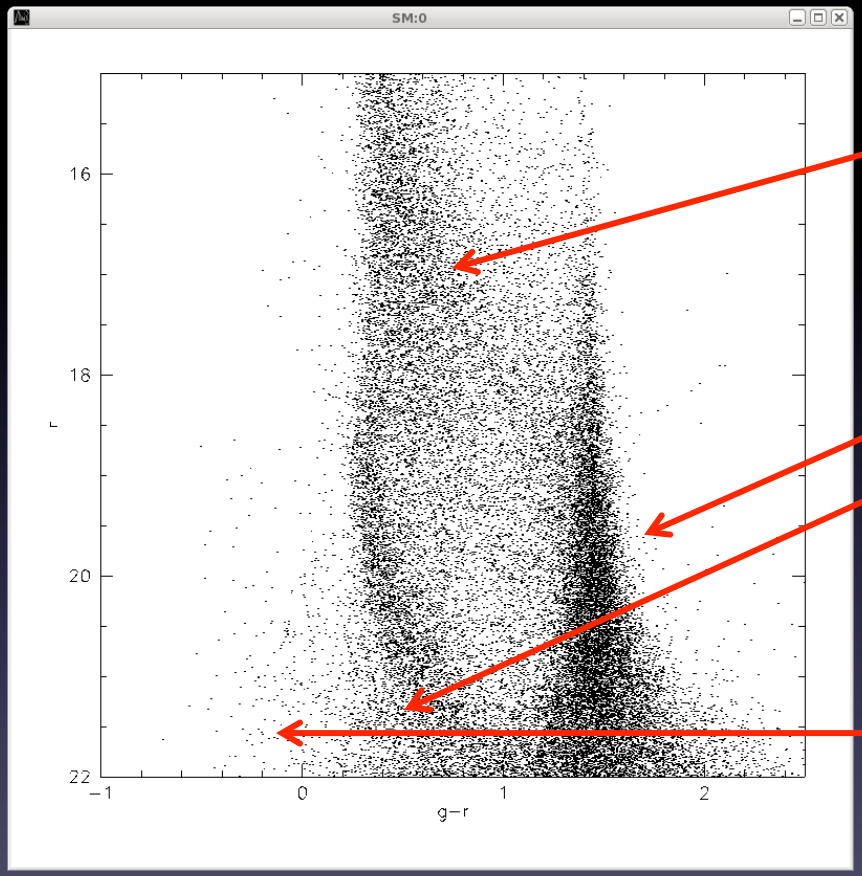
Untitled - Notepad

File Edit Format View Help

```
data "field2.csv"
read <l 2 b 3 Ar 4 u 5 g 6 r 7 i 8 z 9>

set gr = g-r
lweight 2
expand 1.001
limits -1 2.5 22 15
ptype 0 0
erase
box
points gr r
xlabel g-r
ylabel r
```

Color-Magnitude Diagram



Disk G/F stars

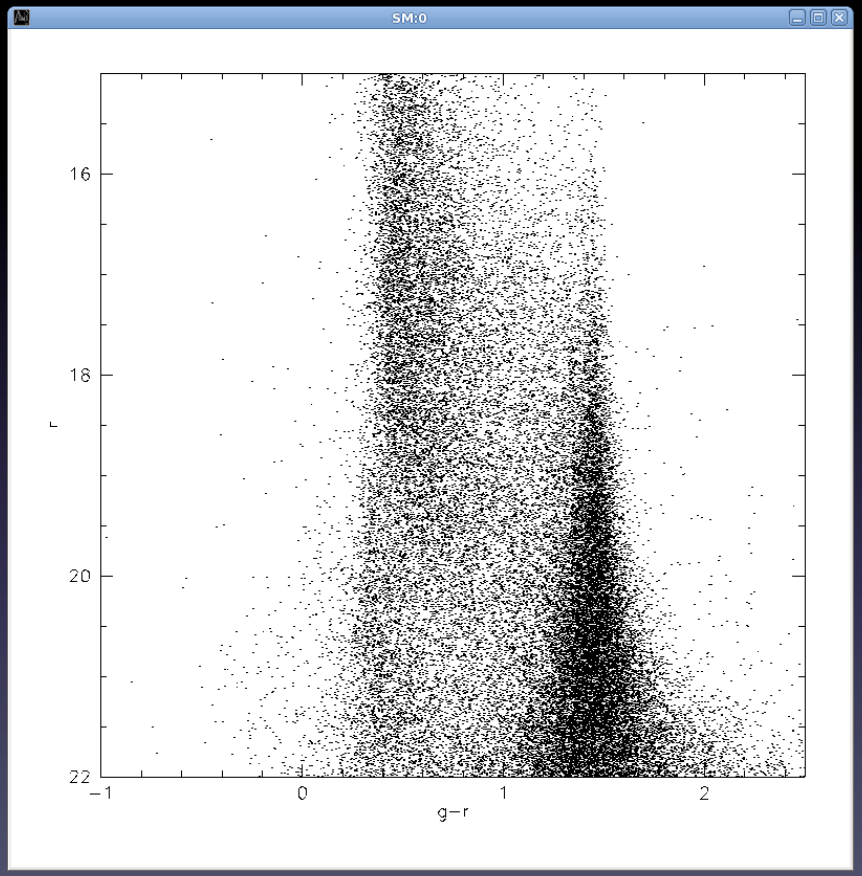
Nearby Disk K/M stars

Halo stars

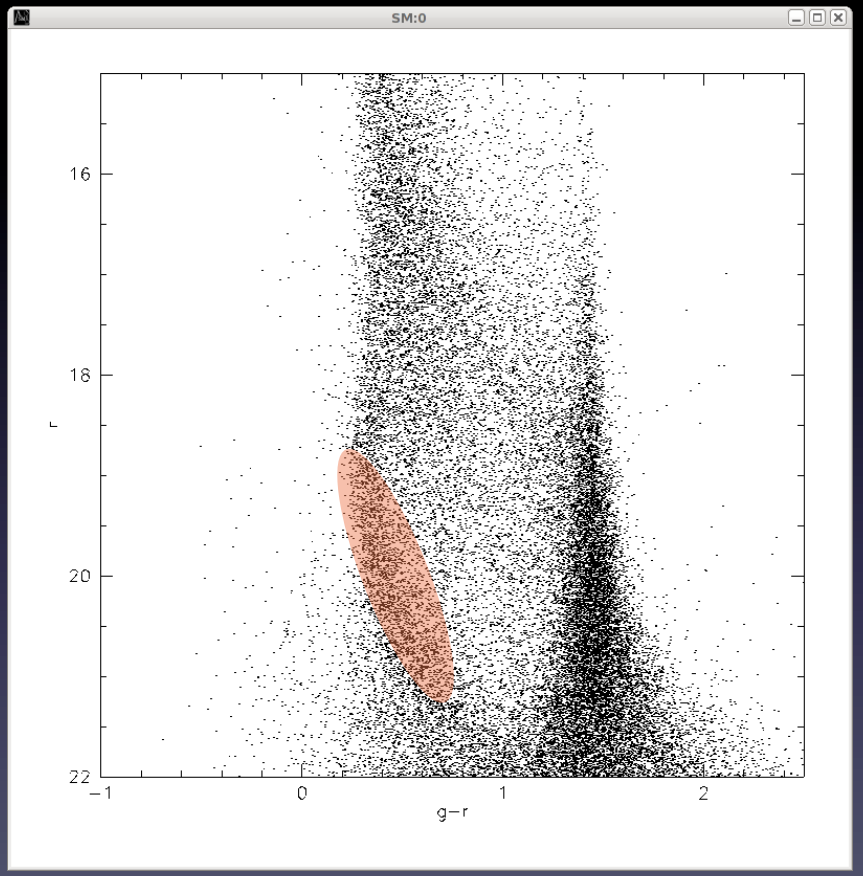
QSOs, RRLyrae,
WDs, errors...

Notice Anything Different…?

Field #1



Field #2



Monoceros Stream (Newberg et al. 2002)

THE GHOST OF SAGITTARIUS AND LUMPS IN THE HALO OF THE MILKY WAY

HEIDI JO NEWBERG,[1,2] BRIAN YANNY,[1,3] CONNIE ROCKOSI,[4] EVA K. GREBEL,[5] HANS-WALTER RIX,[5] JON BRINKMANN,[6] ISTVAN CSABAI,[7] GREG HENNESSY,[8] ROBERT B. HINDSLEY,[8] RODRIGO IBATA,[9] ZELJKO IVEZIĆ,[10] DON LAMB,[4] E. THOMAS NASH,[3] MICHAEL ODENKIRCHEN,[5] HEATHER A. RAVE,[2] D. P. SCHNEIDER,[11] J. ALLYN SMITH,[12] ANDREA STOLTE,[5] AND DONALD G. YORK[4]

Received 2001 June 18; accepted 2001 December 5

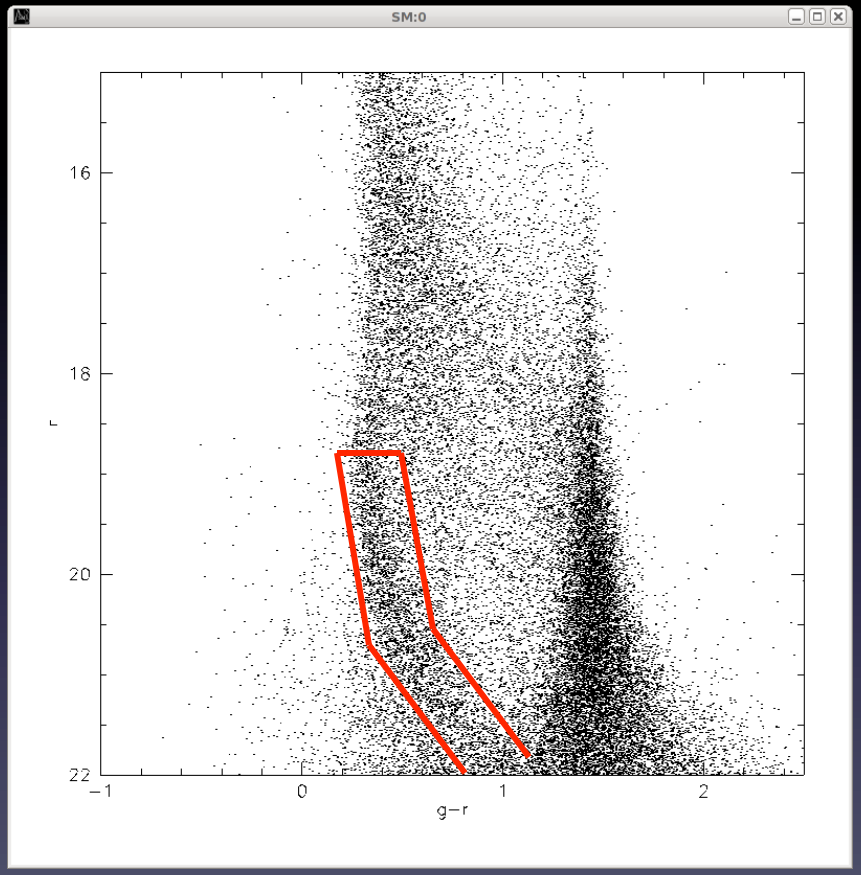
ABSTRACT

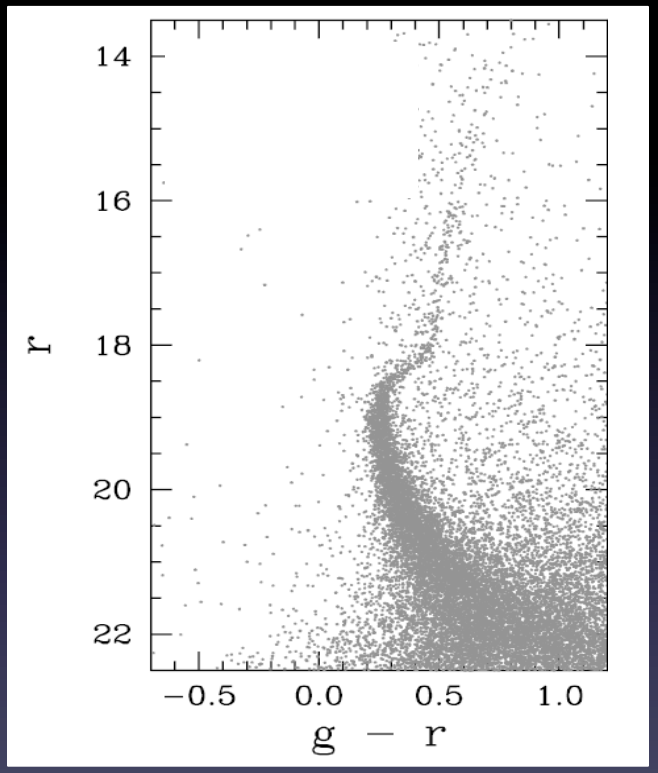
We identify new structures in the halo of the Milky Way from positions, colors, and magnitudes of five million stars detected in the Sloan Digital Sky Survey. Most of these stars are within 1°.26 of the celestial equator. We present color-magnitude diagrams (CMDs) for stars in two previously discovered, tidally disrupted structures. The CMDs and turnoff colors are consistent with those of the Sagittarius dwarf galaxy, as had been predicted. In one direction, we are even able to detect a clump of red stars, similar to that of the Sagittarius dwarf, from stars spread across 110 deg^2 of sky. Focusing on stars with the colors of F turnoff objects, we identify at least five additional overdensities of stars. Four of these may be pieces of the same halo structure, which would cover a region of the sky at least 40° in diameter, at a distance of 11 kpc from the Sun (18 kpc from the center of the Galaxy). The turnoff is significantly bluer than that of thick-disk stars, yet the stars lie closer to the Galactic plane than a power-law spheroid predicts. We suggest two models to explain this new structure. One possibility is that this new structure could be a new dwarf satellite of the Milky Way, hidden in the Galactic plane and in the process of being tidally disrupted. The other possibility is that it could be part of a disklike distribution of stars which is metal-poor, with a scale height of approximately 2 kpc and a scale length of approximately 10 kpc. The fifth overdensity, which is 20 kpc away, is some distance from the Sagittarius dwarf streamer orbit and is not associated with any known Galactic structure. We have tentatively identified a sixth overdensity in the halo. If this sixth structure is instead part of a smooth distribution of halo stars (the spheroid), then the spheroid must be very flattened, with axial ratio $q = 0.5$. It is likely that there are many smaller streams of stars in the Galactic halo.

Subject headings: Galaxy: halo — Galaxy: structure

Learning about structures from CMDs

Monoceros stream CMD

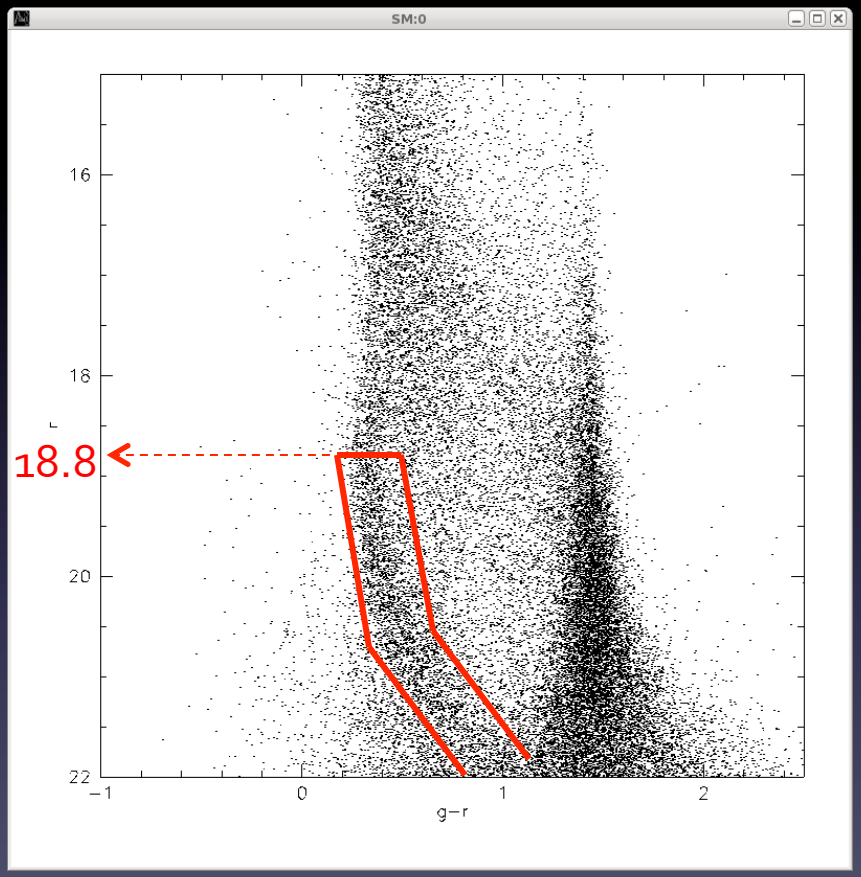




⇨ Monoceros stream is a localized structure

Learning about structures from CMDs

Monoceros stream CMD



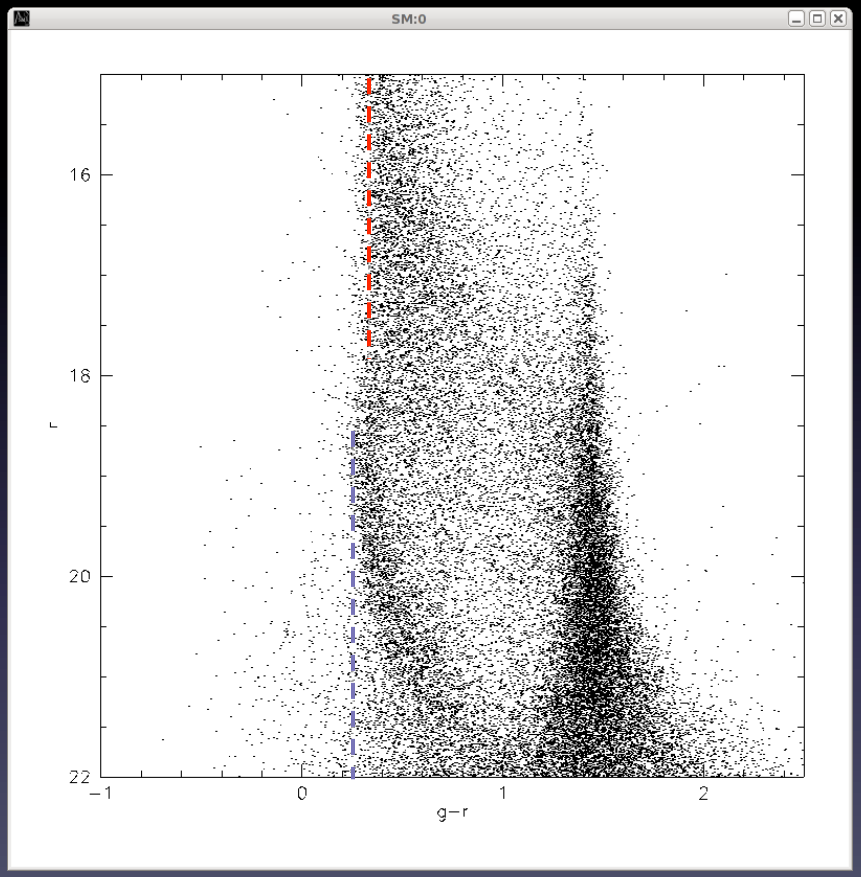
$$M_{MSTO} \approx 3.5$$

$$d = 10^{\frac{m-M}{5}+1}$$

⇨ *The distance to the Monoceros stream has to be around ~11.5kpc*

Learning about structures from CMDs

Monoceros stream CMD



Turn-off color of the Monoceros stream is bluer than that of the disk

⇨ Mon. stars are either more metal poor, or younger than disk stars

Monoceros Stream (Newberg et al. 2002)

THE GHOST OF SAGITTARIUS AND LUMPS IN THE HALO OF THE MILKY WAY

HEIDI JO NEWBERG,[1,2] BRIAN YANNY,[1,3] CONNIE ROCKOSI,[4] EVA K. GREBEL,[5] HANS-WALTER RIX,[5] JON BRINKMANN,[6] ISTVAN CSABAI,[7] GREG HENNESSY,[8] ROBERT B. HINDSLEY,[8] RODRIGO IBATA,[9] ZELJKO IVEZIĆ,[10] DON LAMB,[4] E. THOMAS NASH,[3] MICHAEL ODENKIRCHEN,[5] HEATHER A. RAVE,[2] D. P. SCHNEIDER,[11] J. ALLYN SMITH,[12] ANDREA STOLTE,[5] AND DONALD G. YORK[4]

Received 2001 June 18; accepted 2001 December 5

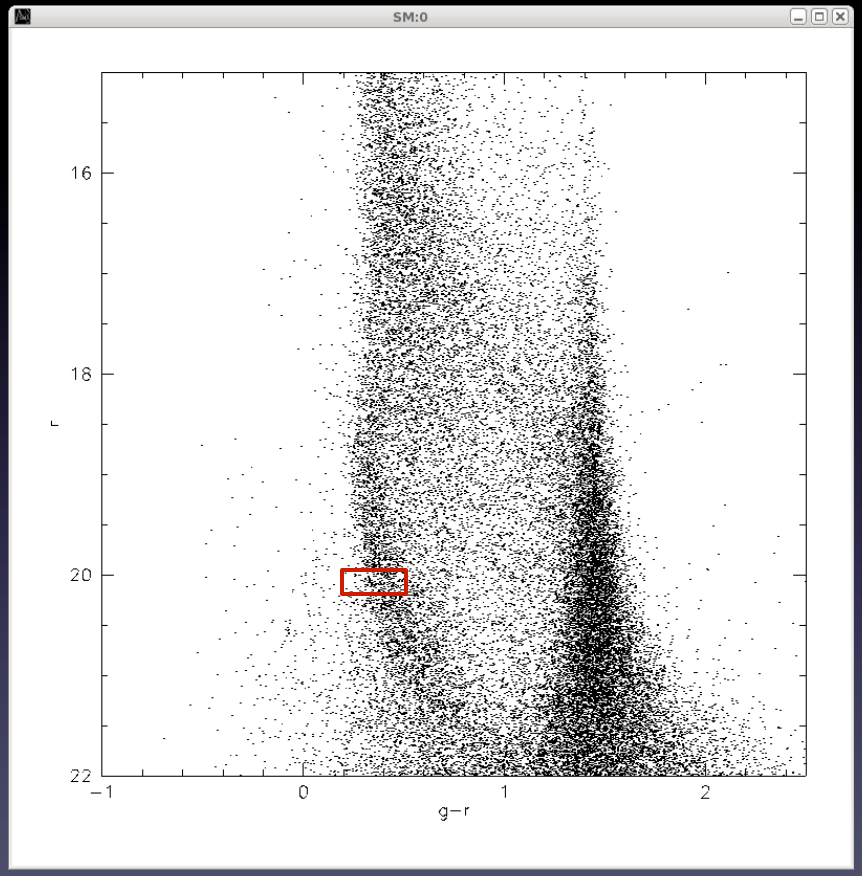
ABSTRACT

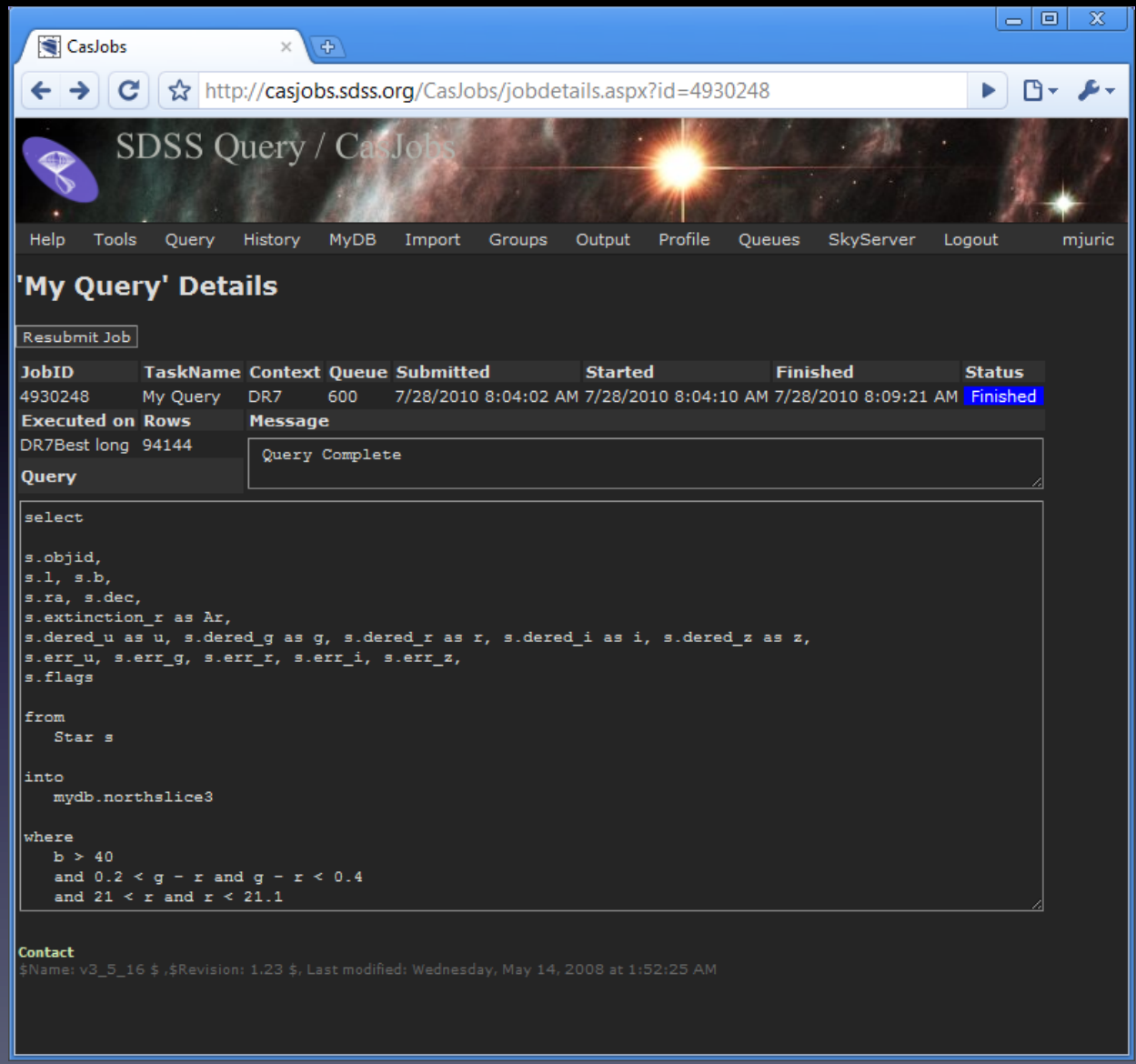
We identify new structures in the halo of the Milky Way from positions, colors, and magnitudes of five million stars detected in the Sloan Digital Sky Survey. Most of these stars are within 1°.26 of the celestial equator. We present color-magnitude diagrams (CMDs) for stars in two previously discovered, tidally disrupted structures. The CMDs and turnoff colors are consistent with those of the Sagittarius dwarf galaxy, as had been predicted. In one direction, we are even able to detect a clump of red stars, similar to that of the Sagittarius dwarf, from stars spread across 110 deg^2 of sky. Focusing on stars with the colors of F turnoff objects, we identify at least five additional overdensities of stars. Four of these may be pieces of the same halo structure, which would cover a region of the sky at least 40° in diameter, at a distance of 11 kpc from the Sun (18 kpc from the center of the Galaxy). The turnoff is significantly bluer than that of thick-disk stars, yet the stars lie closer to the Galactic plane than a power-law spheroid predicts. We suggest two models to explain this new structure. One possibility is that this new structure could be a new dwarf satellite of the Milky Way, hidden in the Galactic plane and in the process of being tidally disrupted. The other possibility is that it could be part of a disklike distribution of stars which is metal-poor, with a scale height of approximately 2 kpc and a scale length of approximately 10 kpc. The fifth overdensity, which is 20 kpc away, is some distance from the Sagittarius dwarf streamer orbit and is not associated with any known Galactic structure. We have tentatively identified a sixth overdensity in the halo. If this sixth structure is instead part of a smooth distribution of halo stars (the spheroid), then the spheroid must be very flattened, with axial ratio $q = 0.5$. It is likely that there are many smaller streams of stars in the Galactic halo.

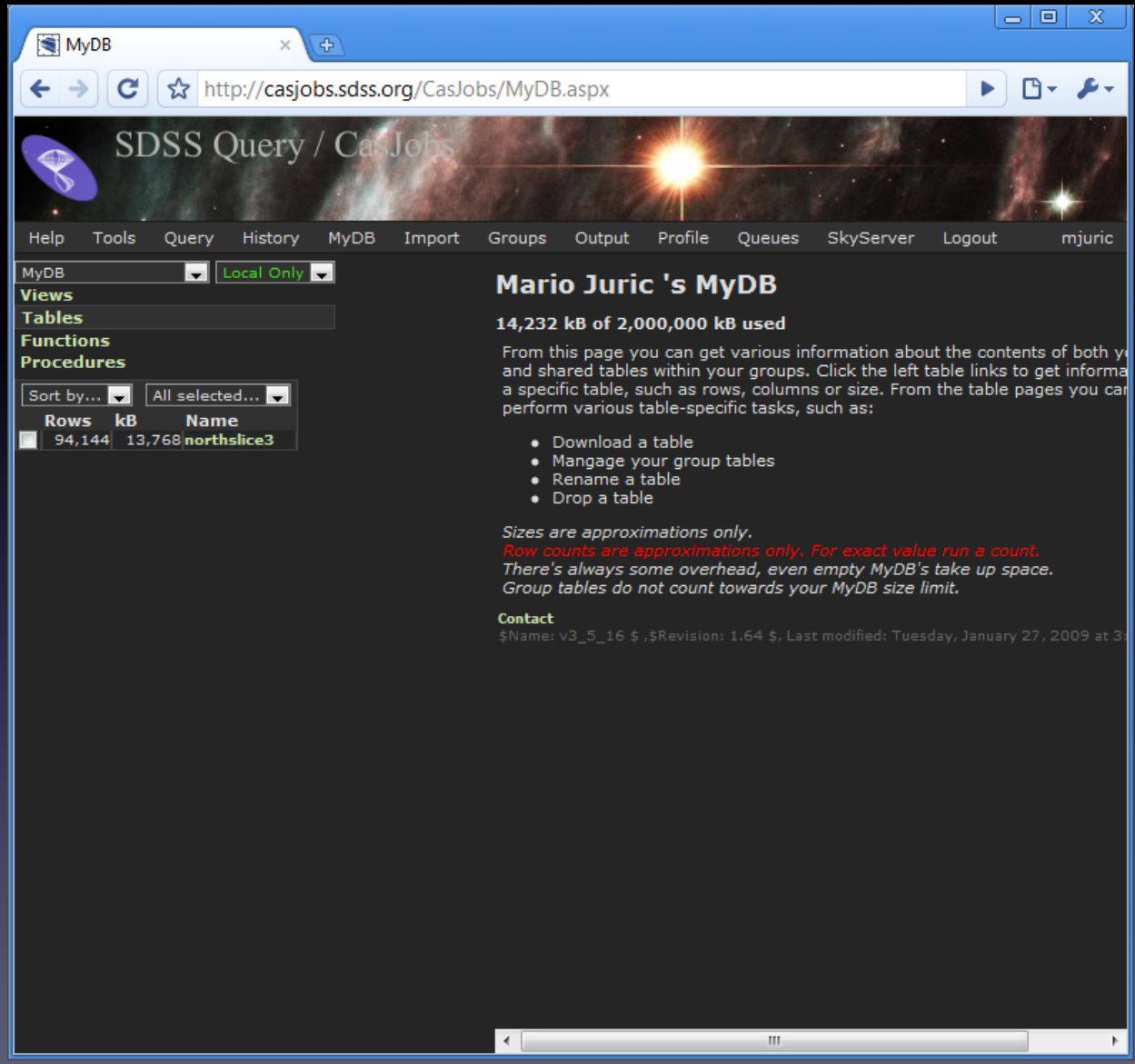
Subject headings: Galaxy: halo — Galaxy: structure

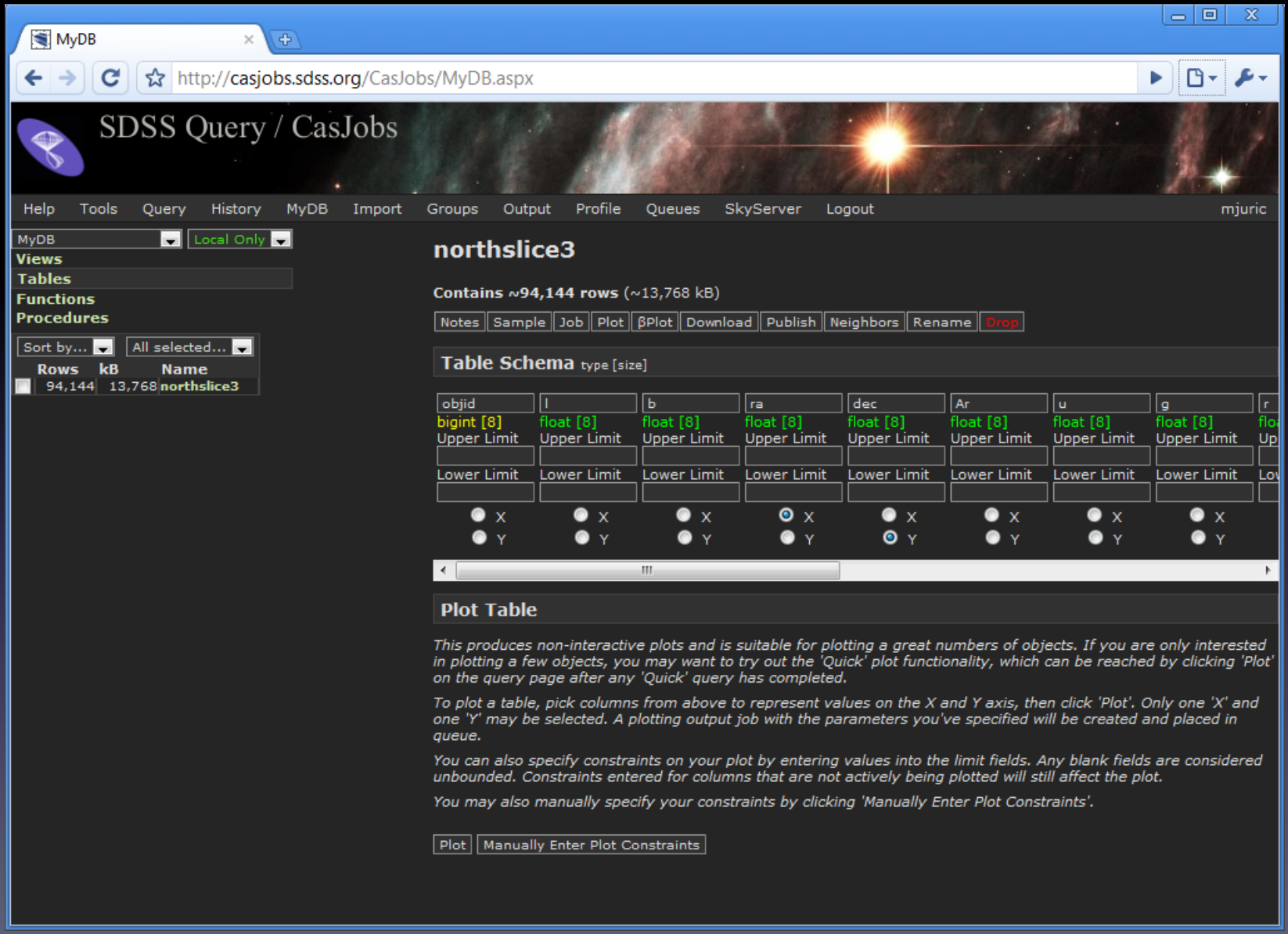
Hmmm… How do I do this for the whole sky?

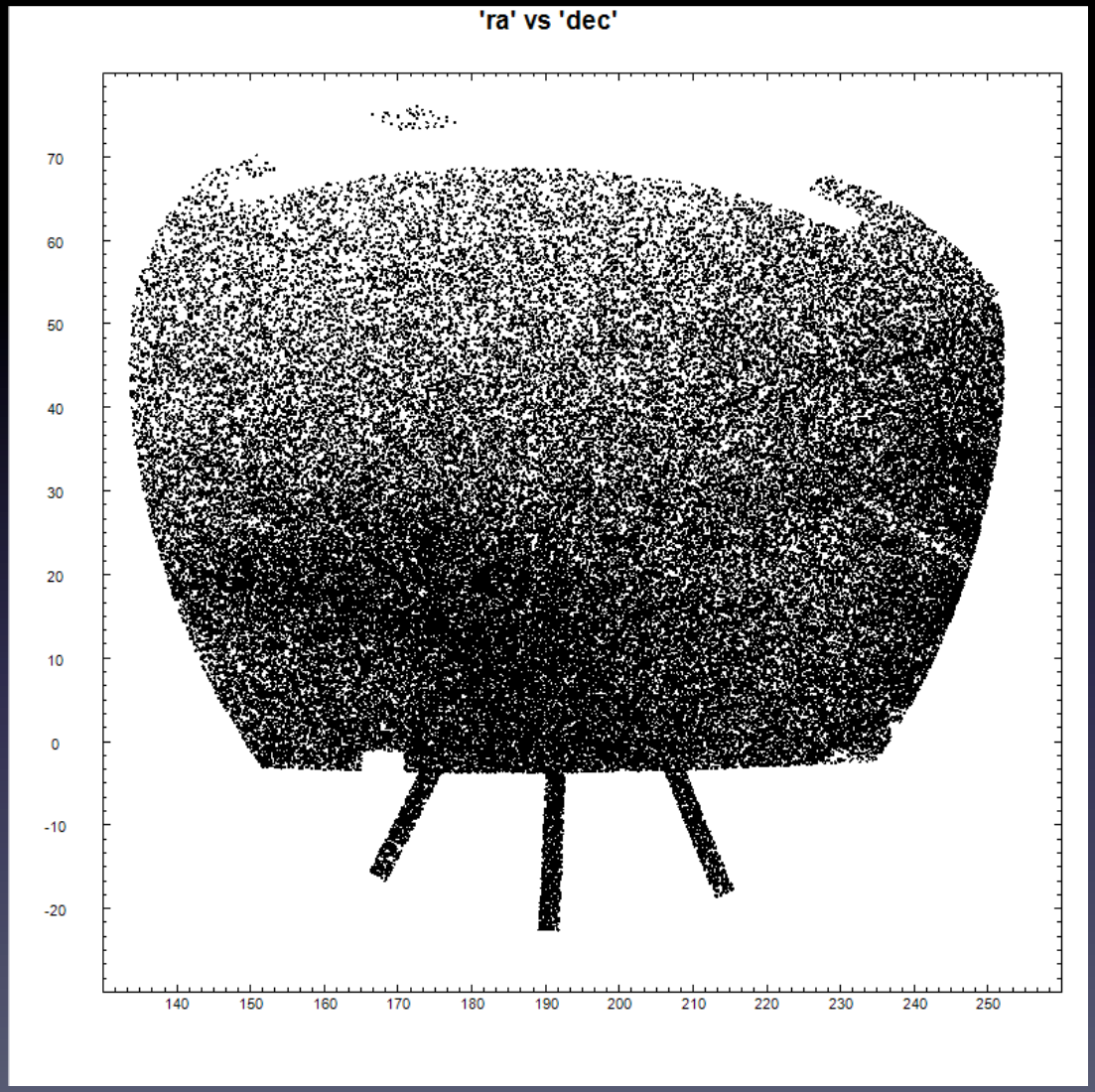
Monoceros stream CMD




CasJobs
http://casjobs.sdss.org/CasJobs/jobdetails.aspx?id=4930248
SDSS Query / CasJobs
Help Tools Query History MyDB Import Groups Output Profile Queues SkyServer Logout mjuric
'My Query' Details
Resubmit Job
JobID TaskName Context Queue Submitted Started Finished Status
4930248 My Query DR7 600 7/28/2010 8:04:02 AM 7/28/2010 8:04:10 AM 7/28/2010 8:09:21 AM Finished
Executed on Rows Message
DR7Best long 94144
Query Complete
Query
select
s.objid,
s.l, s.b,
s.ra, s.dec,
s.extinction_r as Ar,
s.dered_u as u, s.dered_g as g, s.dered_r as r, s.dered_i as i, s.dered_z as z,
s.err_u, s.err_g, s.err_r, s.err_i, s.err_z,
s.flags
from
Star s
into
mydb.northslice3
where
b > 40
and 0.2 < g - r and g - r < 0.4
and 21 < r and r < 21.1
Contact
$Name: v3_5_16 $,$Revision: 1.23 $, Last modified: Wednesday, May 14, 2008 at 1:52:25 AM

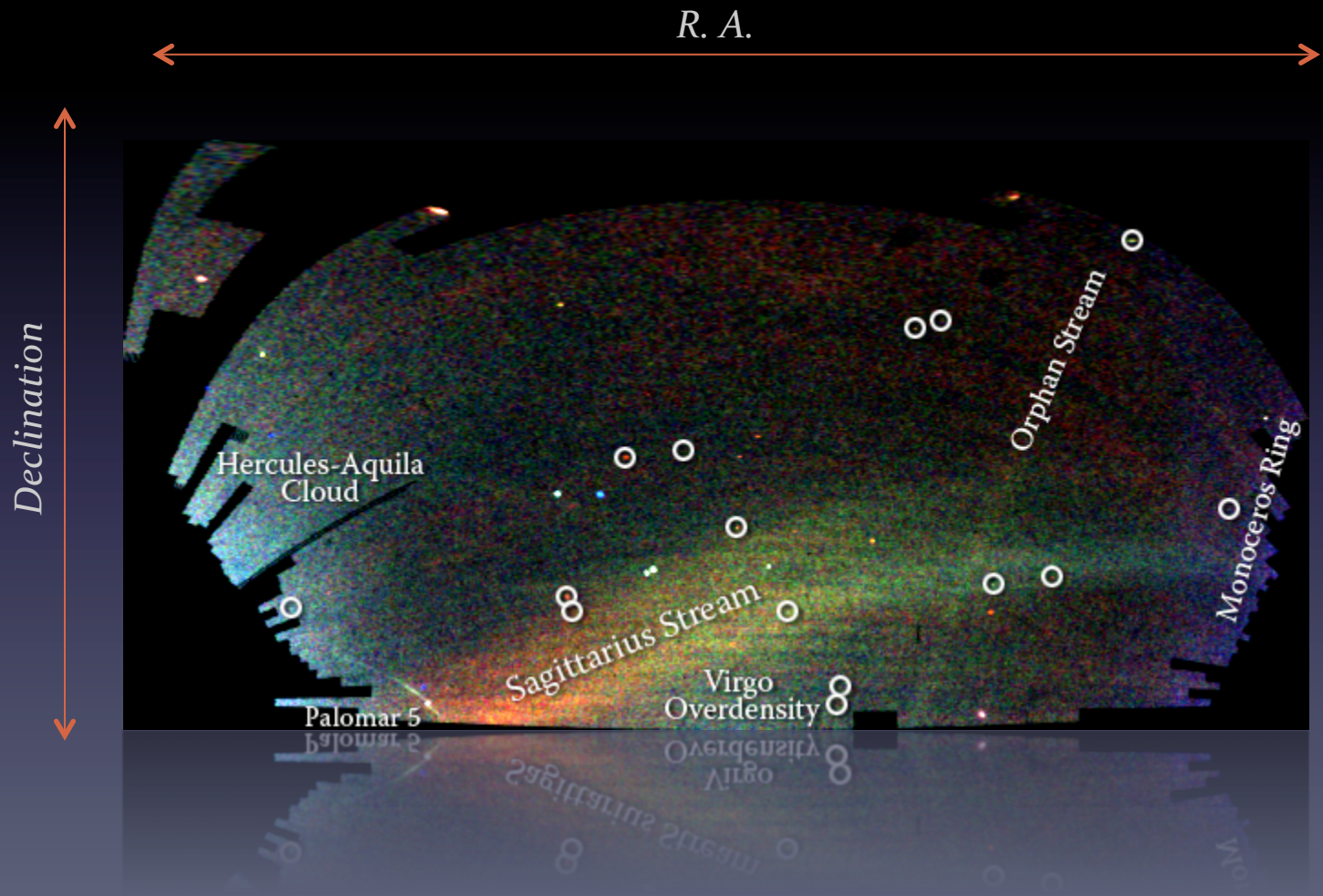

MyDB
http://casjobs.sdss.org/CasJobs/MyDB.aspx
SDSS Query / CasJobs
Help Tools Query History MyDB Import Groups Output Profile Queues SkyServer Logout mjuric
MyDB
Local Only
Views
Tables
Functions
Procedures
Sort by...
All selected...
Rows kB Name
94,144 13,768 northslice3
Mario Juric 's MyDB
14,232 kB of 2,000,000 kB used
From this page you can get various information about the contents of both y
and shared tables within your groups. Click the left table links to get informa
a specific table, such as rows, columns or size. From the table pages you ca
perform various table-specific tasks, such as:
Download a table
Mangage your group tables
Rename a table
Drop a table
Sizes are approximations only.
Row counts are approximations only. For exact value run a count.
There's always some overhead, even empty MyDB's take up space.
Group tables do not count towards your MyDB size limit.
Contact
$Name: v3_5_16 $,$Revision: 1.64 $, Last modified: Tuesday, January 27, 2009 at 3

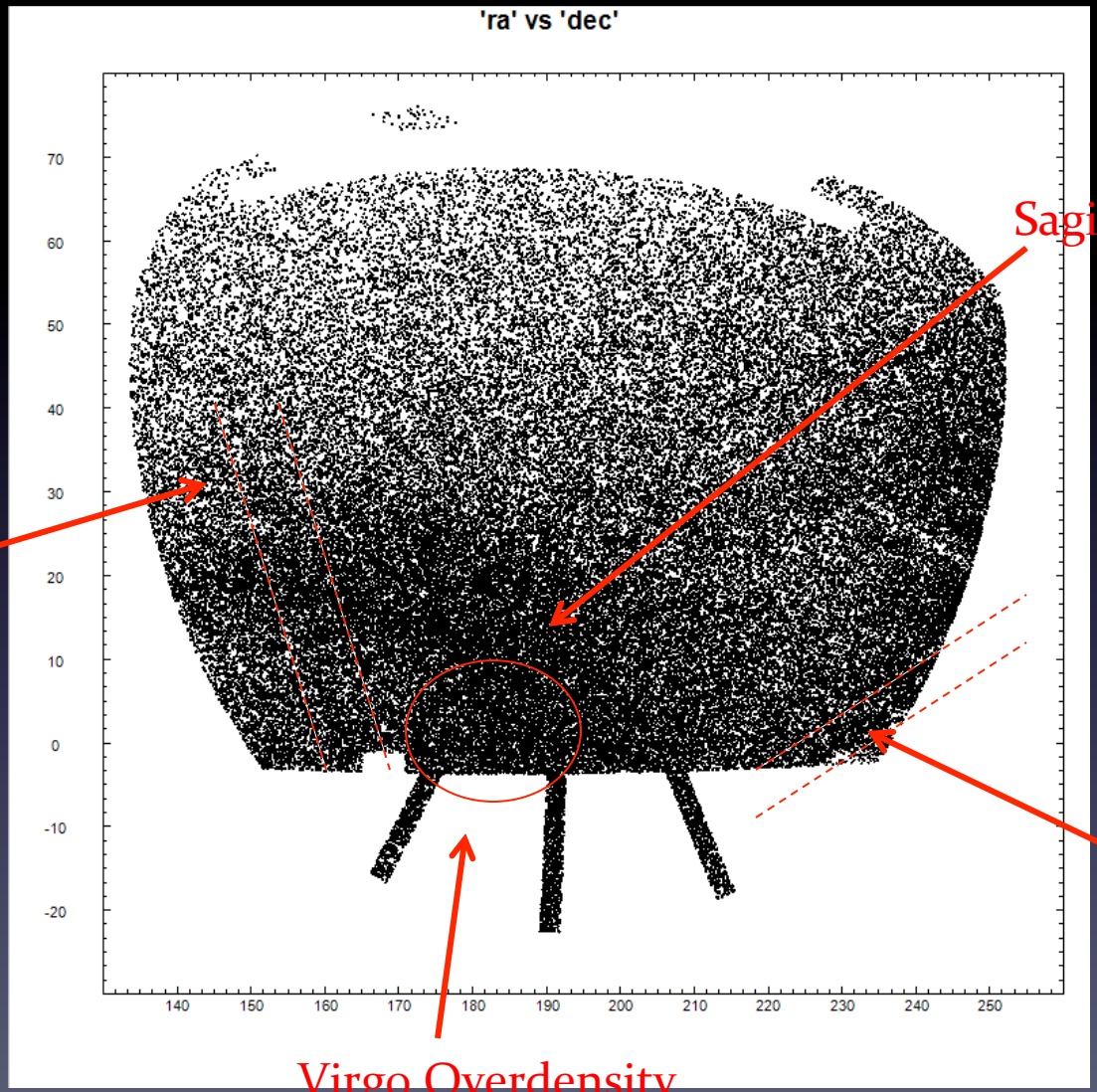

MyDB
http://casjobs.sdss.org/CasJobs/MyDB.aspx
SDSS Query / CasJobs
Help Tools Query History MyDB Import Groups Output Profile Queues SkyServer Logout mjuric
MyDB Local Only
Views
Tables
Functions
Procedures
Sort by... All selected...
Rows kB Name
94,144 13,768 northslice3
northslice3
Contains ~94,144 rows (~13,768 kB)
Notes Sample Job Plot βPlot Download Publish Neighbors Rename Drop
Table Schema type [size]
objid l b ra dec Ar u g r
bigint [8] float [8] float [8] float [8] float [8] float [8] float [8] float [8] flo
Upper Limit
Lower Limit
X Y
Plot Table
This produces non-interactive plots and is suitable for plotting a great numbers of objects. If you are only interested in plotting a few objects, you may want to try out the 'Quick' plot functionality, which can be reached by clicking 'Plot' on the query page after any 'Quick' query has completed.
To plot a table, pick columns from above to represent values on the X and Y axis, then click 'Plot'. Only one 'X' and one 'Y' may be selected. A plotting output job with the parameters you've specified will be created and placed in queue.
You can also specify constraints on your plot by entering values into the limit fields. Any blank fields are considered unbounded. Constraints entered for columns that are not actively being plotted will still affect the plot.
You may also manually specify your constraints by clicking 'Manually Enter Plot Constraints'.
Plot
Manually Enter Plot Constraints


'ra' vs 'dec'
70
60
50
40
30
20
10
0
-10
-20
140
150
160
170
180
190
200
210
220
230
240
250

Credit: Vasily Belokurov and the SDSS

Field of Streams




'ra' vs 'dec'
70
60
50
40
30
20
10
0
-10
-20
140
150
160
170
180
190
200
210
220
230
240
250
Sagittarius stream
Orphan Stream
(Belokurov et al. 2006)
Palomar 5
Virgo Overdensity
(Juric et al. 2005, 2008)

- Let's summarize what we just did:
  - In less than 10 minutes, we "observed" two fields 120deg apart (~4 months), to 22nd magnitude, in 5 bands
  - We "reduced" and extinction-corrected the data (~70,000 stars/field) and analyzed it using color-magnitude plots
  - We discovered a main sequence feature in one of the streams at distance around 11.5kpc, likely metal poor
  - We then proceeded to discover three more streams over the whole sky
- 10 years ago, this would have been a multi-night observing proposal for top telescopes in the world. The challenge is obtaining the data.
- Today it's an SQL query – *the data is already there, the challenge is to ask the right question*!

Whoa!

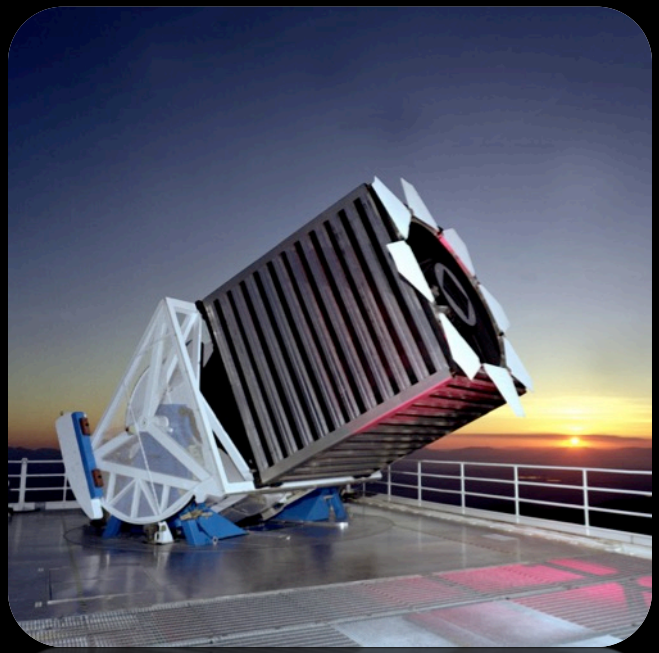
Astronomy in the Age of Large Surveys

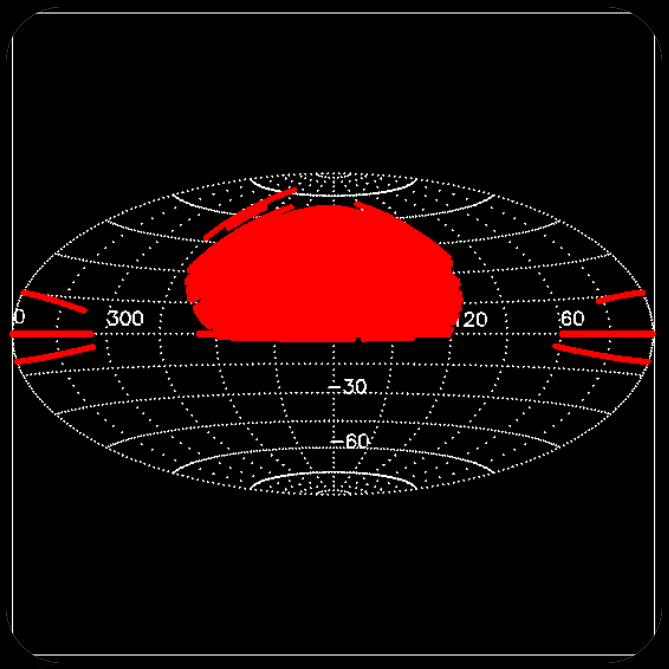
- Traditionally, astronomy was a data-starved science. Our approach to research and our analysis methods were shaped by this environment. Surveys are altering it; data is becoming abundant and of unprecedented quality.

Sloan Digital Sky Survey

2.5m telescope >10000 deg^2 0.1" astrometry r<22.5 flux limit

5 band, 2%, photometry for >50M stars
>300k R=2000 stellar spectra





10 years of ops: ~10 TB of imaging

Panoramic Survey Telescope and Rapid Response System

1.8m telescope 30000 deg^2 50mas astrometry r<23 flux limit

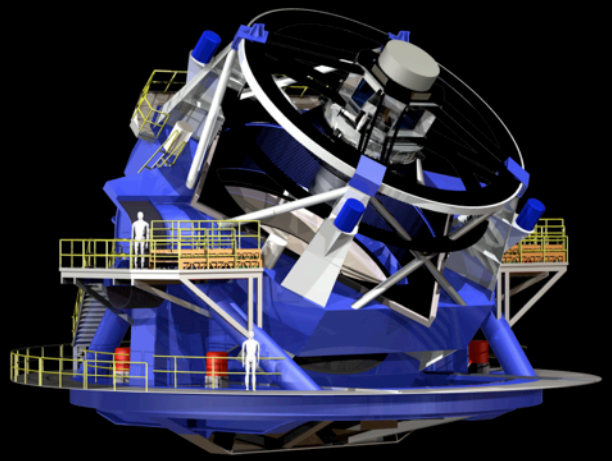
5 band, better than 1% photometry (goal)

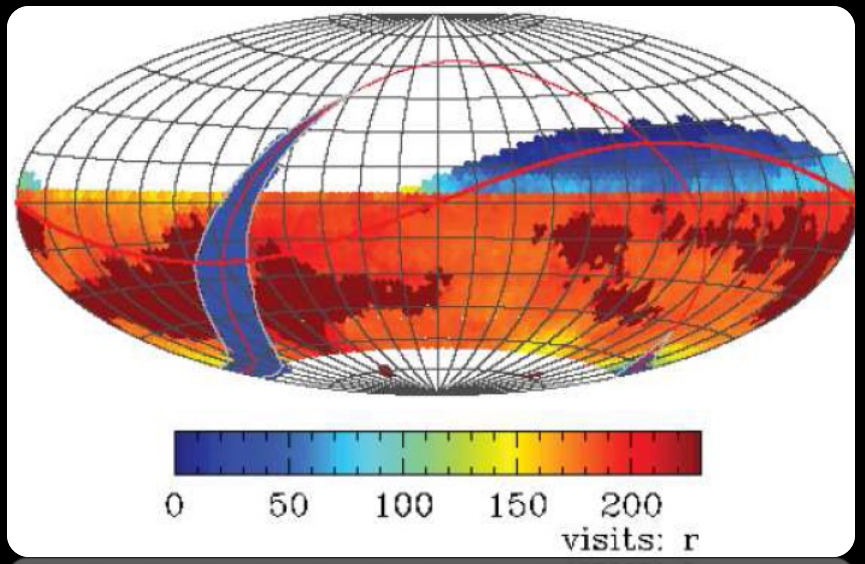
~700 GB/night

LSST: A Deep, Wide, Fast, Optical Sky Survey

8.4m telescope | 18000+ deg^2 | 10mas astrom. | r<24.5 (<27.5@10yr)

ugrizy | 0.5-1% photometry





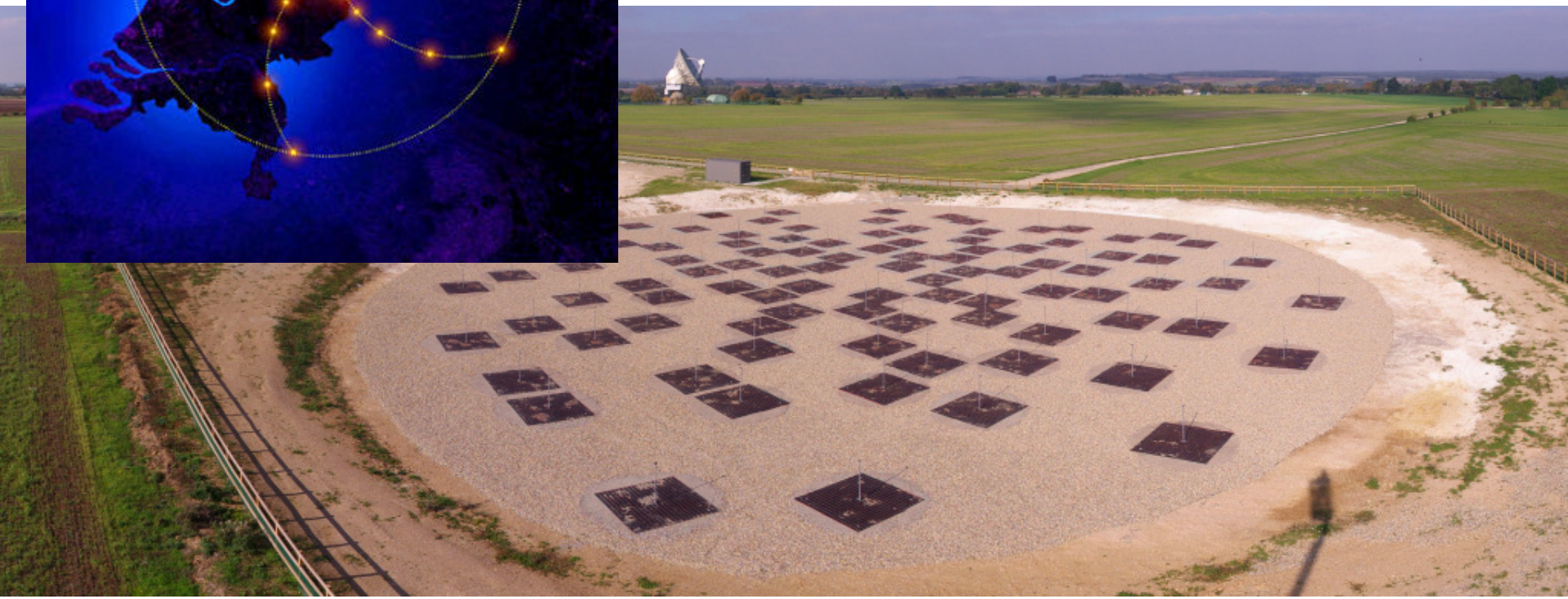
3.2Gpix camera | 30sec exp/4sec rd | 15TB/night | 37 B objects

Imaging the visible sky, once every 3 days, for 10 years (825 revisits)

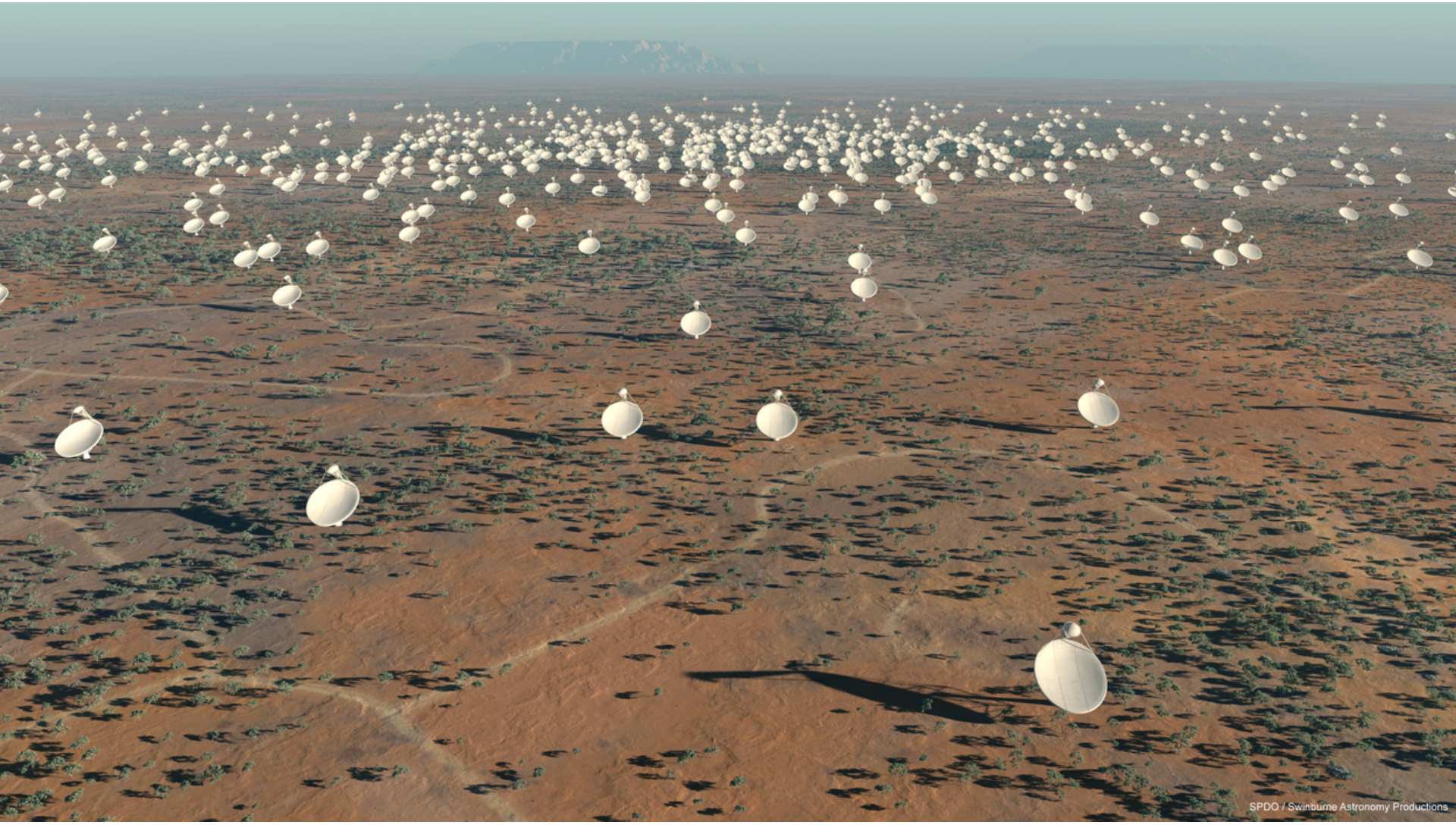
Low-Frequency Array for Radio Astronomy (LOFAR)

Estimated Data Volume

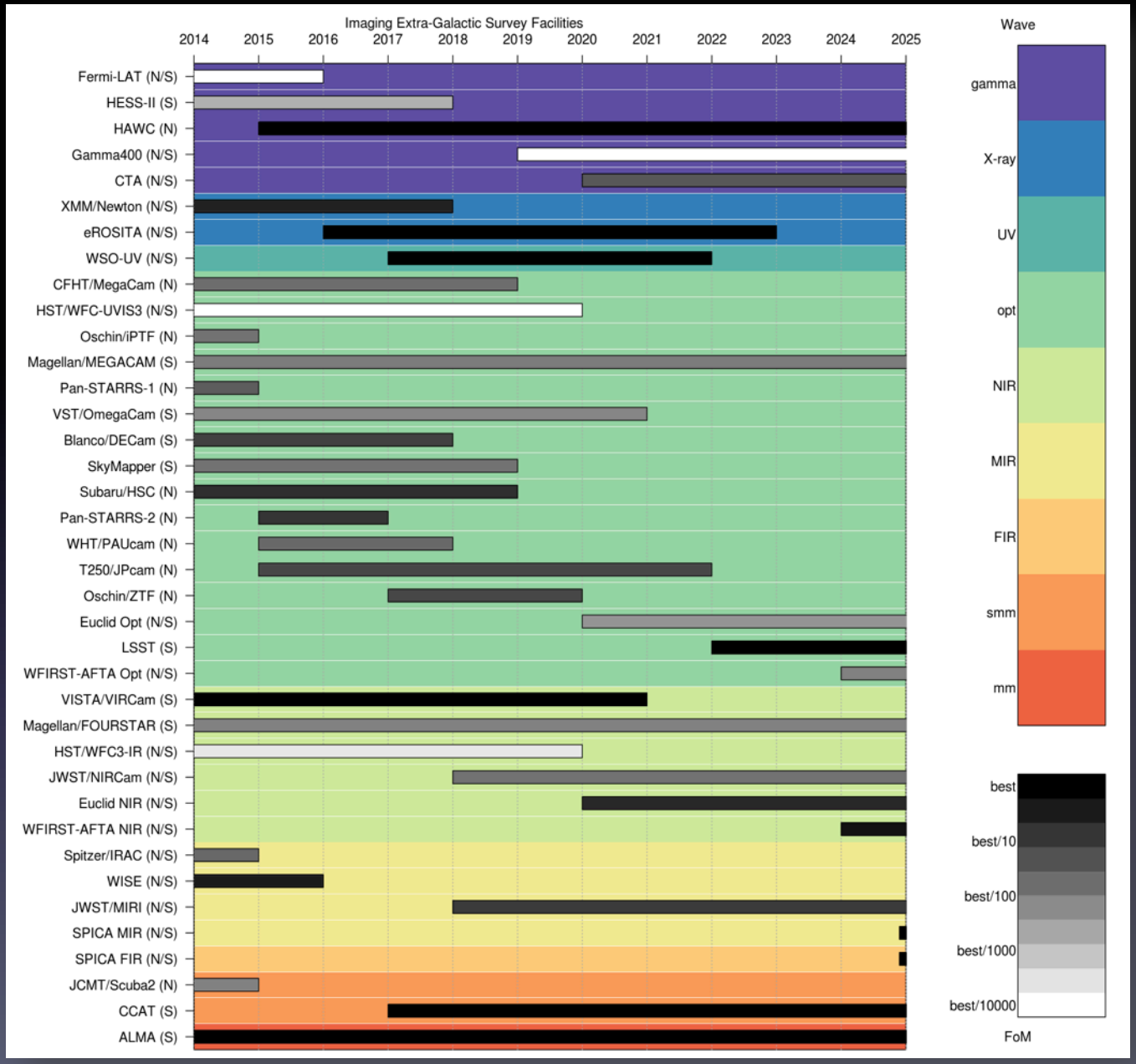
Data Type	Data Rate [MB/s]	Data Rate [GB/h]
Burst Mode	100	360
Monitoring Mode	0.03	0.1
Spectrometer Mode	0.34	1.2



Square Kilometer Array



10 PB / hour



https://asgr.shinyapps.io/ganttshiny/

Astronomy in the Age of Large Surveys

- Traditionally, astronomy was a data-starved science. Our approach to research and our analysis methods were shaped by this environment. Surveys are altering it; data is becoming abundant and of unprecedented quality.

- For optical, LSST will cap this transformation: it will deliver the positions, magnitudes and variability information for virtually *everything* in the southern sky to 24^{th}-27^{th} magnitude, with an order of magnitude better controlled systematics than current surveys.

A Quantitative Difference

Final Image Collection – All DRs	515 PB	All Data Releases Includes Virtual Data (475 PB)
Final Image Collection – DR11	114 PB	Data Release 11 (Year 10) Includes DR11 Virtual Data (88 PB) and all raw images from all years
Final Database	16 PB	Data Release 11 (Year 10) Includes Data, Indexes, and DB Swap
Final Disk Storage	375 PB	Archive and Base Sites
Final Tape Storage	121 PB	Single Site, Single Copy Only
Peak Number of Nodes	1750	Archive and Base Sites Compute and Database Nodes
Number of Alerts Generated	28 billion	Life of survey

- **Virtual Data is data that is dynamically recreated on-demand from provenance information**

A Qualitative Difference

9 hours on CFHT/MegaCam (Munoz et al. 2010)

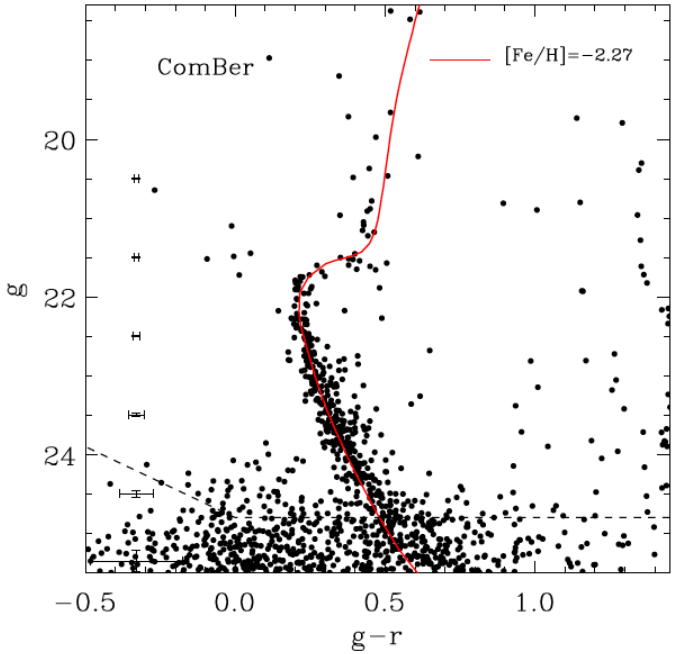
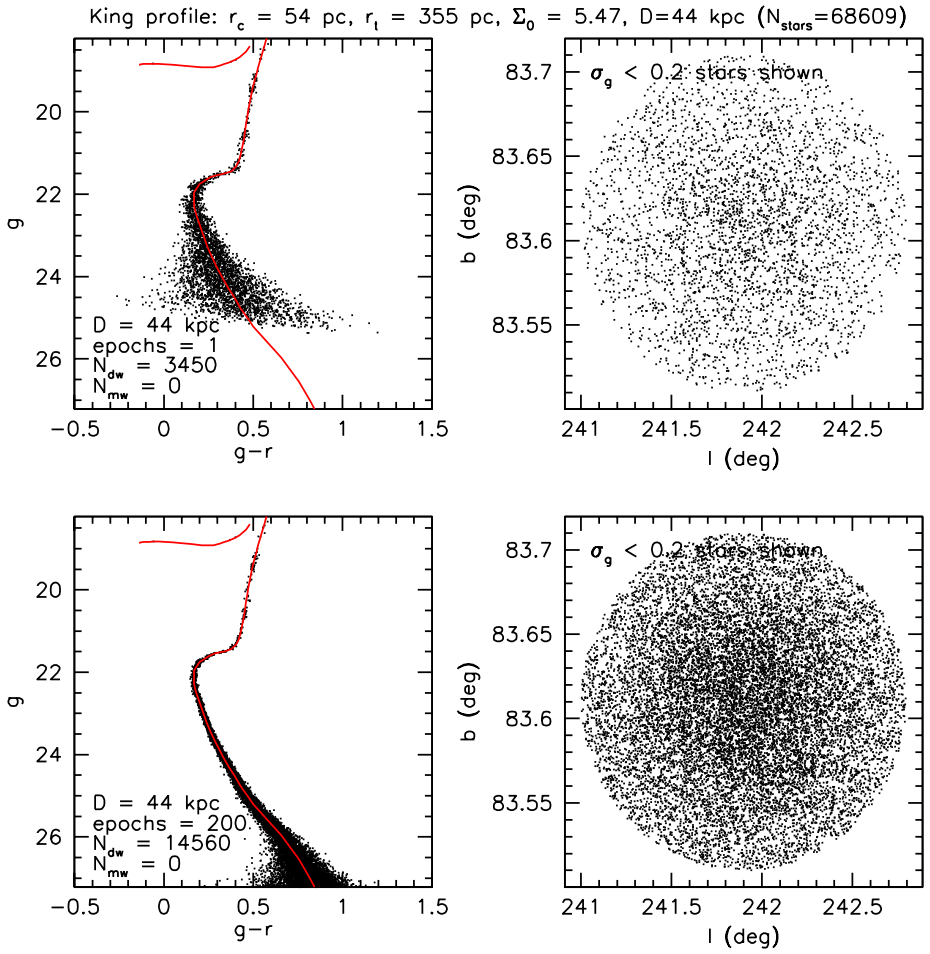


Figure 3. CMD for the inner region ($r < 6'$) of ComBer. The dashed lines mark the 90% completeness level after χ and sharp cuts have been applied to remove non-stellar objects. As it can be seen, our CFHT photometry reaches at least 3 mag below the main sequence turn-off of ComBer. We have complemented our photometry with SDSS data for $g > 20$. The error bars to the left were determined from the artificial star tests and represent the standard deviation of a Gaussian function fitted to the error distribution as a function of magnitude. A theoretical isochrone for a 13 Gyr old, [Fe/H] = −2.27 population is shown with a solid red line (from Girardi et al. 2004).

(A color version of this figure is available in the online journal.)

LSST, single-epoch and full-survey (simulation)



Astronomy in the Age of Large Surveys

- Traditionally, astronomy was a data-starved science. Our approach to research and our analysis methods were shaped by this environment. Surveys are altering it; data is becoming abundant and of unprecedented quality.

- For optical, LSST will cap this transformation: it will deliver the positions, magnitudes and variability information for virtually *everything* in the southern sky to 24th-27th magnitude, with an order of magnitude better controlled systematics than current surveys.

- **We're entering the age of abundance of high quality data. Success in research will depend on the ability to analyze and mine knowledge from that data.**

Next time …

- Software Tools
  - Python
  - IPython
  - Git and github
  - Basic numerical and analysis libraries