# SkyServer Traffic Report: The SQL!

## *10 Years of SkyServer Web and SQL Logs*

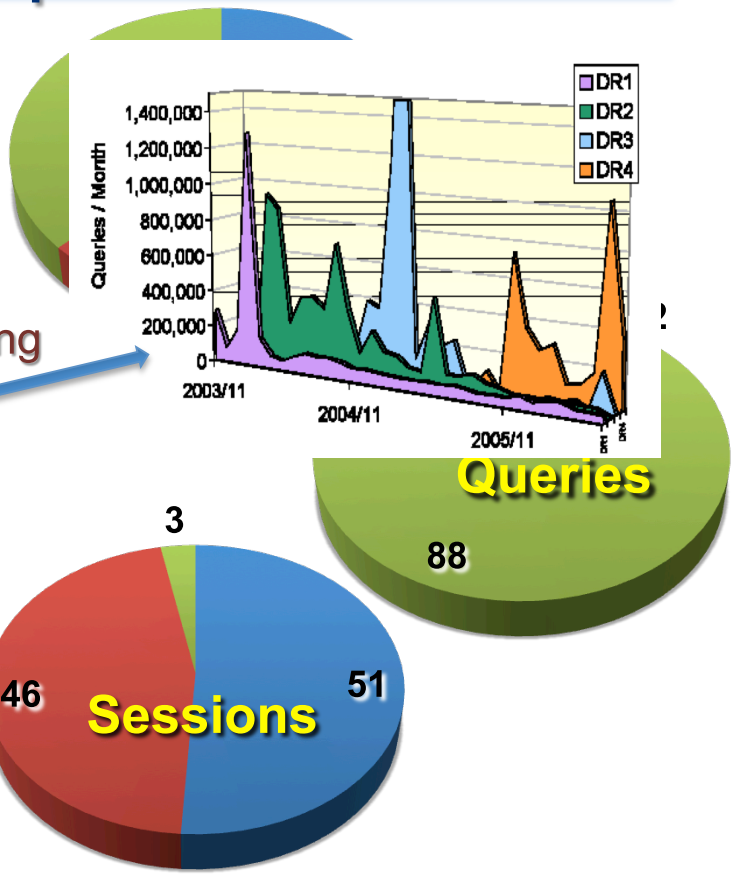
**Ani Thakar**
**Jordan Raddick**
**Alex Szalay**
**(JHU)**

***and***

**Jim Gray**

# First Traffic Report

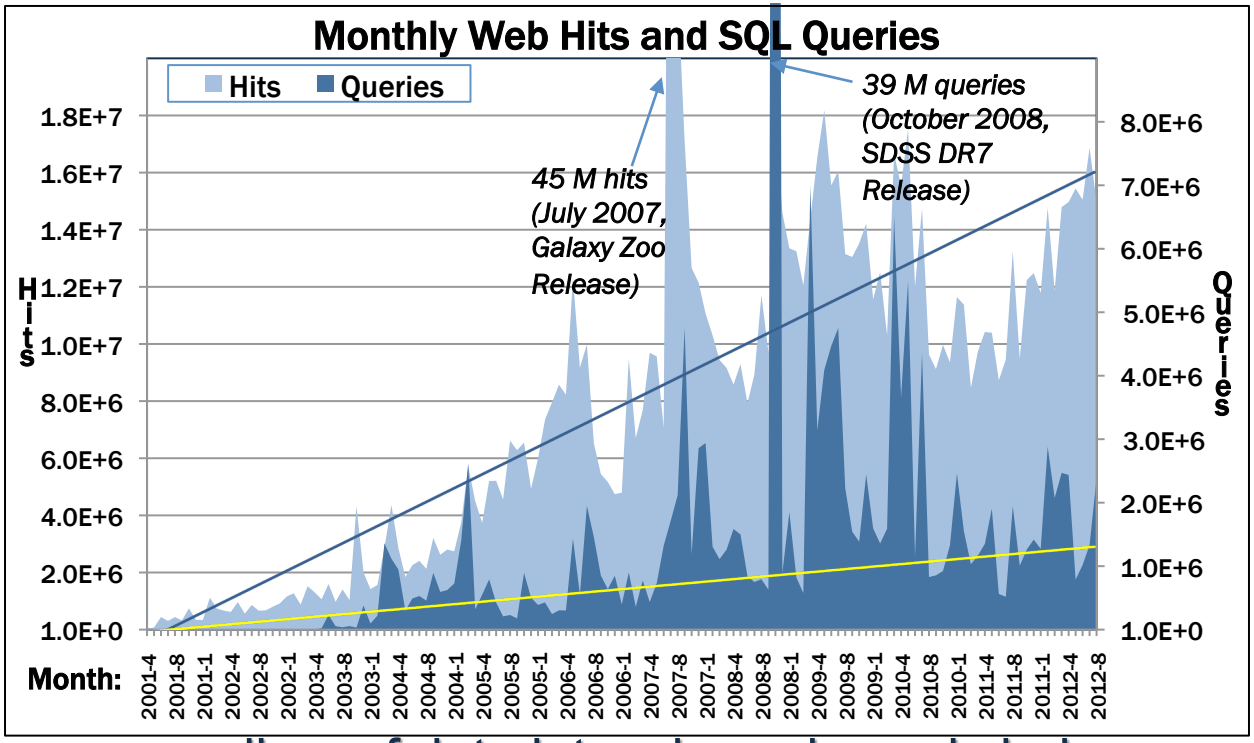
- 5-year report, 2006 MSTR
  - Vik Singh, Jim Gray et al., covered 2001-2006
  - *http://bit.ly/skyserver5years*
- Highlights:
  - Web & SQL traffic doubled every year
  - Hundreds of astronomers "graduated" from using canned and sample queries to free-form SQL
  - Flurry of activity after each release
  - Hard to separate bots & mortals reliably




| Users | Web Hits | SQL Queries | Web Sessions |
|---|---|---|---|
| Total | 65M | 16M | 3M |
| Mortals | 27M | 16M | 1.5M |
| Spiders | 14M | 3M | 1.4M |
| Bots | 24M | 14M | 1M |

# 10 Years of SkyServer Logs

- Sequel to the 5-year report
- In preparation
- Extend original analysis to new, larger dataset
- Separate the Web hits and SQL query analysis
- Focus more on the 10-year SQL usage data
- Unique dataset
- Best record of how new paradigm of data intensive science is being embraced by scientists



# Research Questions for SQL Usage

- Who's using SkyServer & CasJobs SQL?
- How often are they using it?
- How are they using it?
- Are they getting better at it?
  - How complex are their queries?
  - How do users learn SQL?
- What type of science is being done?
- Is it meeting the requirements?
- How can we improve the system?
- How effective is our online Help?

# SQL Usage by the Numbers

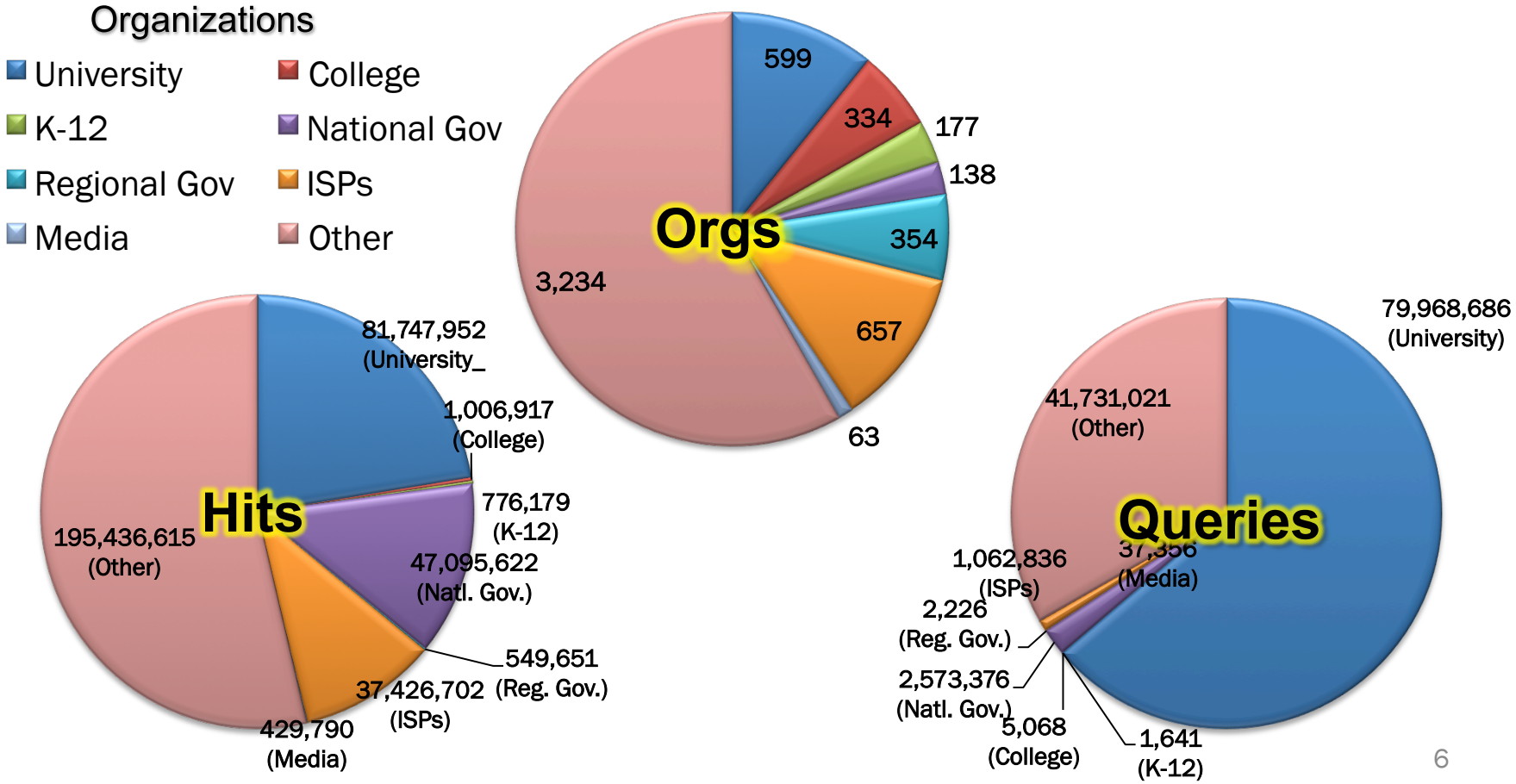
| Total | Unique | Succeeded (error=0) | Failed (error != 0) |
|---|---|---|---|
| 194,023,591 | 67,946,073 | 145,002,755 | 49,020,836 |

- Top 5 SQL users:
  - All bots/programs
  - Big Brother (JHU monitor) is 2nd

| Client IP | Queries |
|---|---|
| UVic/CADC | 44.3 M |
| JHU (BB Monitor) | 14.1 M |
| Berkeley | 12.4 M |
| Japan | 6.0 M |
| NRC (Canada) | 5.4 M |

- Biggest single day: 37,097,351 queries!
  - October 23, 2008 (close to DR7 release)
  - Nearly all from one IP (UVic/CADC)
  - Vast majority *failed*, only 1% actually succeeded (362k)

# Traffic by IP Domain



# Query Complexity Index

- Length of Query (in bytes)
  - Naïve, doesn't necessarily indicate an intelligent or sophisticated query
- JOINs: number and types of JOINs
- GROUP BY / ORDER BY
- CROSS JOINs, CROSS APPLYs, cursors
- Function calls (UDFs)
  - Depends on type(s) of function called
  - Good way to detect science use cases
- Combinations of some/all of the above

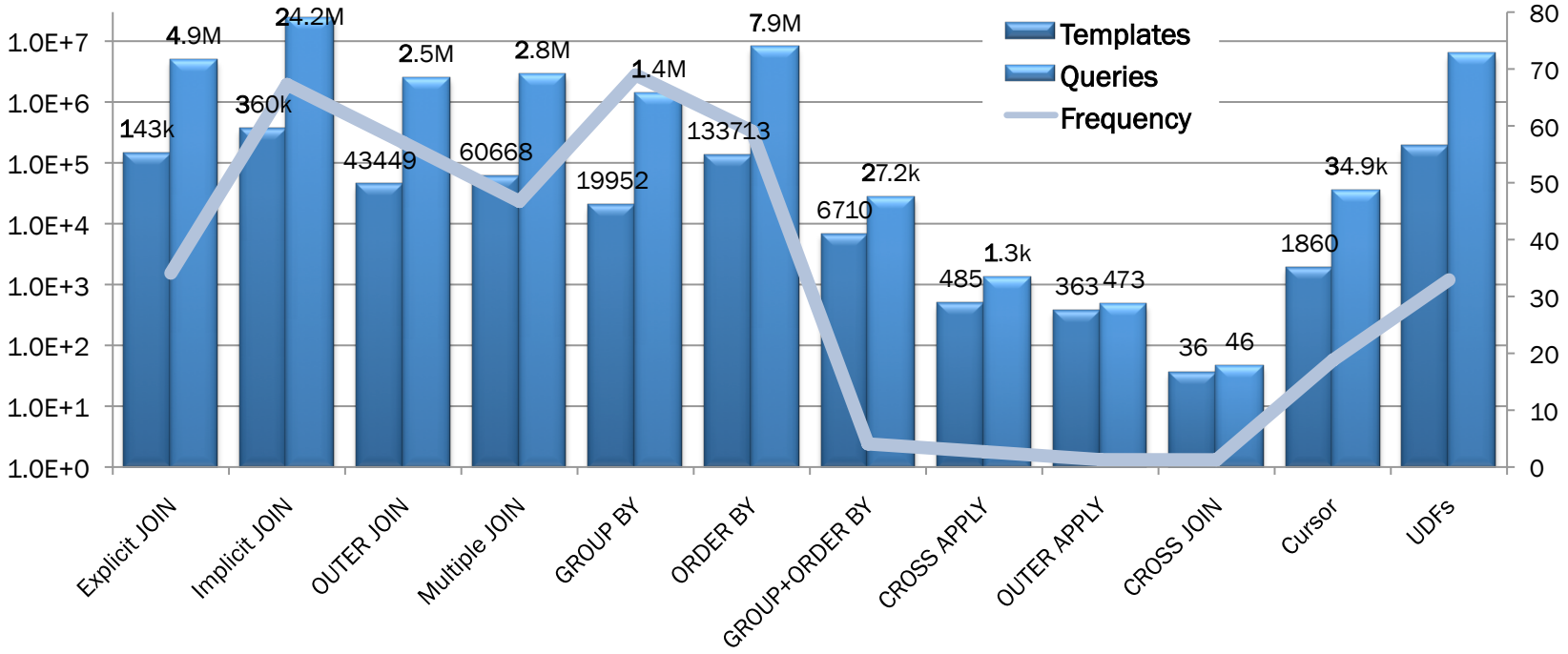
# SQL Templates

- Divide queries into templates:
  - Focus on successful queries only (69M)
  - Regexp replace of all numbers with "#"
  - SELECT DISTINCT SQL statements
  - Assign each template a number (templateID)
- Just under a million query templates
  - From ~ 69 M unique queries
  - Derive a crude "complexity index"
  - Based on presence of SQL elements
    - JOINs (implicit/explicit/CROSS/OUTER,multiple)
    - GROUP BY, ORDER BY, cursors
- Faster to query based on templates
  - Avoids expensive text search through entire DB

# Template Creation

```
INSERT SqlTemplate
 SELECT
  dbo.RegExReplace(
    dbo.RegExReplace(
       dbo.RegExReplace(
          dbo.RegExReplace(
             dbo.RegExReplace(
                SUBSTRING(statement,
                   PATINDEX('%select%',
                                statement),9999),
                   '(?<char>\W)(0x[0-9A-Fa-f]+|[+|-]?((\.\d+)|(\d+(\.\d*)*)))',
                '${char}#'
              )                                          -- squeeze out numbers (but not identifiers eg DR1
             ,'\s+',' ')                                 -- squeeze out white space
          ,'/\/\*(.*\n)*\*\/g',''),                      -- remove multi-line comments
       '/^[ \t\f\v]*--.*\r\n/gm',''),                    -- remove isolated single-line comments
    '/--[^\r^\n]*/g',''),                                -- remove embedded single-line comments
    0 AS hits,
    COUNT(*) AS queries
FROM sqlstatement
WHERE
                statement LIKE '%select%from%‘ AND NOT
                (statement LIKE '%delete%'
                                OR statement LIKE '%drop%'
                                OR statement LIKE '%create%'
                                OR statement LIKE '%parseonly%'
                                OR statement LIKE '%<a%'
                                OR statement LIKE '%up_name%'
                                OR statement LIKE '%batch%‘)
GROUP BY {first column in SELECT]
```

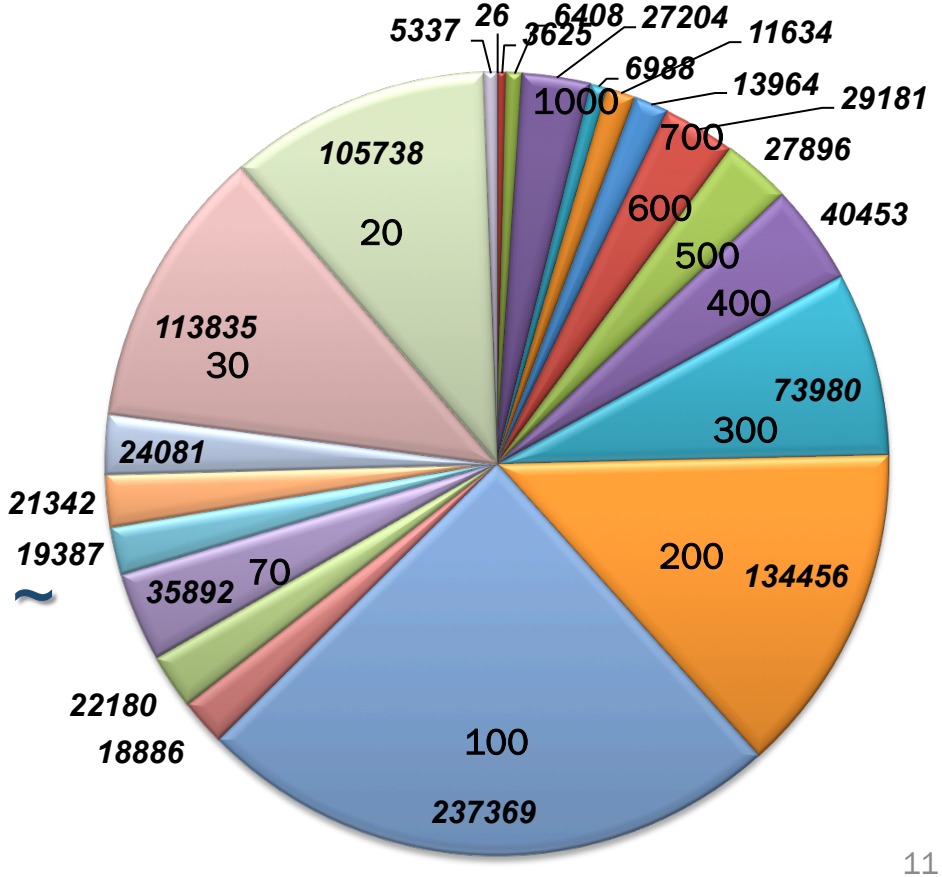
# SQL Constructs in Templates




1) SELECT COUNT(*) FROM SqlTemplate WHERE template LIKE '%join%'
2) SELECT COUNT(*) FROM SqlStatement WHERE TemplateID IN
(SELECT TemplateID FROM SqlTemplate WHERE template LIKE '%join%')

# Length of Query Templates

- Most queries:
  O(100) bytes in length
- ~35k ≥ 1000 bytes
  Less than 5%
- Bot and prog queries are usually small
- Limit on query length ~ 4k in the SkyServer (larger in CasJobs)



# HCI Case Study: SDSS Log Analysis

- Ph.D. Thesis (J. Zhang, Drexel)
- Java SDSS Log Viewer
- Inter-active exploration of SQL logs
  - Color-coded SQL elements
  - Spatial query coverage (SkyMap)
  - Statistics viewer
- Not hooked up to live log DB
  - Connected to downloaded snapshot

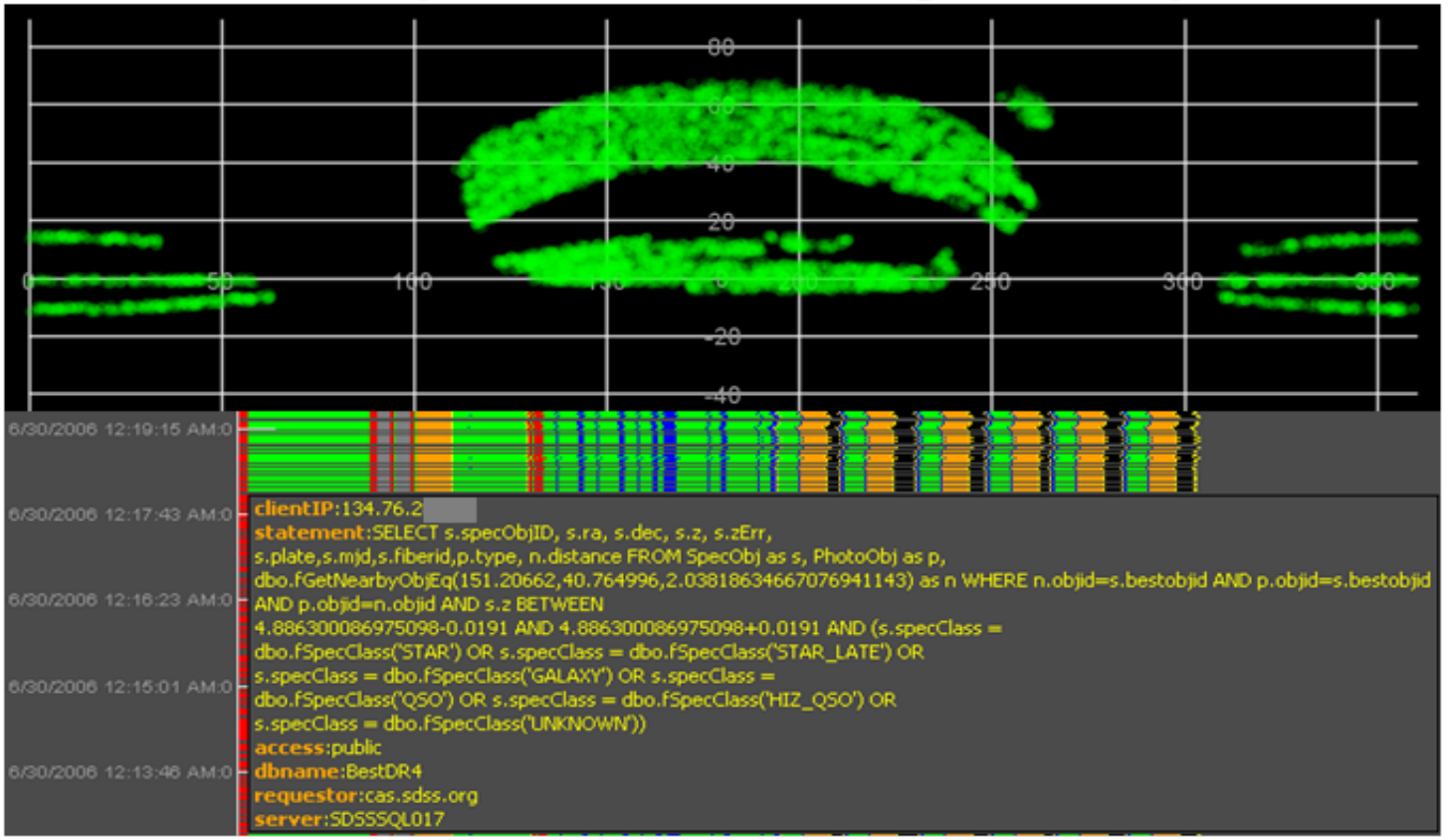



Fig. 11. Typical search queries from a German institute who search SDSS mapped areas only for high redshift objects.

# Provenance and SQL Workflows

- Analysis of CasJobs queries by N.Li (2010)
- Study how users do data-driven analysis
  - Complex queries using MyDB system
- Number of MyDB objects per query
  - How many tables a workflow uses
  - Only ~ 20% of users with > 1 objects/query
- Number of MyDB object (table) dependencies
  - Objects created from queries on other MyDB objects
  - Better measure of workflow/complexity (?)
  - ~38% of users with > 1 dependency
  - These users responsible for 76% of query WFs

# Next Steps

- Get templates for sample queries
  - How useful/effective are sample queries?
  - How many queries are just resubmitted samples?
- Refine complexity index further
  - Nested queries, use of UDFs and SPs
  - Track complexity as function of time
- Track SQL “sessions”
  - More relevant to CasJobs users
  - Multiple queries, use of variables etc.
- Use of built-in indices and HTM
  - Indexed columns, nearby functions, htmID
- More detailed user demographics

# SkyServer Traffic Page

http://skyserver.sdss.org/log/en/traffic/