



Intro to Sessions 3 & 4: Data Management & Data Analysis

Bob Mann

*Wide-Field Astronomy Unit
University of Edinburgh*

Outline

- Data Management Issues
 - Alternatives to monolithic RDBMS model
 - Intercontinental data management
 - More support for individual user data
 - Summary statistics
 - Cross-matching sky surveys
- Data Analysis Issues
 - Types of analysis: “design-goal” & “legacy” science
 - Random sampling
 - Exploratory data analysis

Problems with SDSS model

- JHU developing database for PS1 consortium
 - Data volumes preclude deployment on single server
 - Need to partition large tables over a cluster
 - Need to build supporting infrastructure – GrayWulf
- Little evidence of large-scale astro analysis in RDB
 - Honourable exception: Adrian Pope, Alex Szalay, et al
- Maybe RDBMS is wrong tool for (part of) the job
 - We've been fooled into expressing our problem in its terms

Alternatives to RDMS approach

- *BigTable/MapReduce* – Google
 - Cluster-based: no distinction between data & compute nodes
 - Suitable for a class of problems
 - Use in astro problems pioneered by Ryan Scranton et al
- *SciDB* – new initiative, led by database theorists
 - Generalisation of relational model:
built around multidimensional arrays, not 2D tables
 - Support for access from many programming languages
 - Very strong connections to LSST DM team at SLAC
- Both bring data analysis closer to data management

Intercontinental Data Management

- (Wiki comment from Arun Jagatheesan, iRODS)
- Surveys require consistent data management over multiple institutions – and, often, countries
- WFCAM: home-grown system works quite well, but how far would it scale?
- LSST: prototyping use of iRODS rule-based system – e.g. certain classes of files are copied to Site B as soon as they appear in directory at Site A

Individual User Data

- Related topics being partially addressed in the VO
- Essential issue is the integration of a user's own data with public archives. Two examples:
- "MyDB":
 - SDSS offers users temporary space within a relational database to create tables which can be used in conjunction with SkyServer tables
 - e.g. uploading data for cross-matching, storing intermediate result sets
 - Standardisation foreseen in road map for IVOA VOSpace standard, but some way from implementation

Annotation

- Archive curators prefer not to impose judgements
 - e.g. may perform basic star/galaxy separation, but won't do much more by way of classification
- This is the right attitude – scientific judgement is context-specific – but leaves archives quite basic
 - Interpretation lives in the literature – quite separate
- Value in facilitating third-party annotation of data
 - e.g. identify which transient candidates are real
- Associations: special case of annotation – see later

Summary statistics

- What could/should data centre supply to users?
- Spatial coverage – Tamas' talk later
- Source densities (Richard McMahon wiki comment)
 - e.g. surface density of detections in particular band...
 - ...then divided into flux bins, then into...
 - Use Cases: how accurate do these need to be?

Cross-matching sky surveys

- Associations underpin multiwavelength astronomy
- Balance between making life easier for users and making scientific judgements for them
- One solution: cross-neighbours table – WSA
 - i.e. record every SDSS source which is within a certain matching radius of each UKIDSS source:
 1. Users don't have to do spatial proximity part of x-match
 2. Don't tell them of plausible counterparts is true one
- Different solutions for different situations
 - On-the-fly matching, cross-neighbours,...
 - Could we standardise format of match/neighbour records?

Outline

- Data Management Issues
 - Alternatives to monolithic RDBMS model
 - Intercontinental data management
 - More support for individual user data
 - Summary statistics
 - Cross-matching sky surveys
- Data Analysis Issues
 - Types of analysis: “design-goal” & “legacy” science
 - Random sampling
 - Exploratory data analysis

What kinds of analyses?

- Difference between “design-goal science” and “legacy science”
 - Surveys designed for large-scale statistical analyses: specific access patterns by particular teams
 - Longer-term use more varied & uses other data
- Likely to require different support: can the same data management infrastructure support both?

Design-goal science

- Large-scale statistical analyses. Two basic types:
 - Summarising properties of populations:
could be clustering properties, colour distributions, etc
 - Finding rare objects as outliers from populations
- Both types of analysis use density estimation:
is that something requiring special attention?
 - e.g. k-d trees and other in-memory data structures can provide great scalability

Legacy science

- Expect workload to shift in longer run to extraction by position
 - Image postage stamps
 - Adding multiwavelength data to sources from other bands
- Each job is small, but there may be many of them
 - C.f. few, large-scale statistical analyses
- VO standards exist already for this basic sort of data access, so just need to implement them efficiently

Random sampling

- Astronomers use random sampling widely
 - Initial data analysis – faster/easier to use a subset
 - Error estimation – jack-knifing and boot-strapping
- SQL doesn't readily support random sampling
 - Set-based, it returns all rows with given properties
 - Various techniques exist which can make it work
 - Need to assess which is best – and whether good enough – and how to provide general access to random sampling in a database

Exploratory data analysis

- Much astronomy is about finding patterns in multivariate datasets
- Various data mining algorithms exist to do this sort of thing, but many benefit from guidance
- Want to be able to look for patterns interactively
 - Need to be able to support visualization of summary statistical plots (of random samples), etc

Summary

- Monolithic RDBMS model showing strain: what next?
 - Clustered approach – data analysis added somehow
 - Need to assess likely patterns of use
- Cannot think of individual archives in isolation
 - Distributed data management will be part of their creation
 - Cross-matching and annotation must be supported.
- Need to review science cases for classes of analysis
 - Clearly lots of interest in transient identification and classification
- Data analysis at the data centre
 - What services? - statistical summaries of data sets
 - Implementation: standard services to call, user code,...