A Glance at Microarray Data Management

Claudio Lottaz

Computational Diagnostics Group
Computational Molecular Biology
Max Planck Institute for
Molecular Genetics

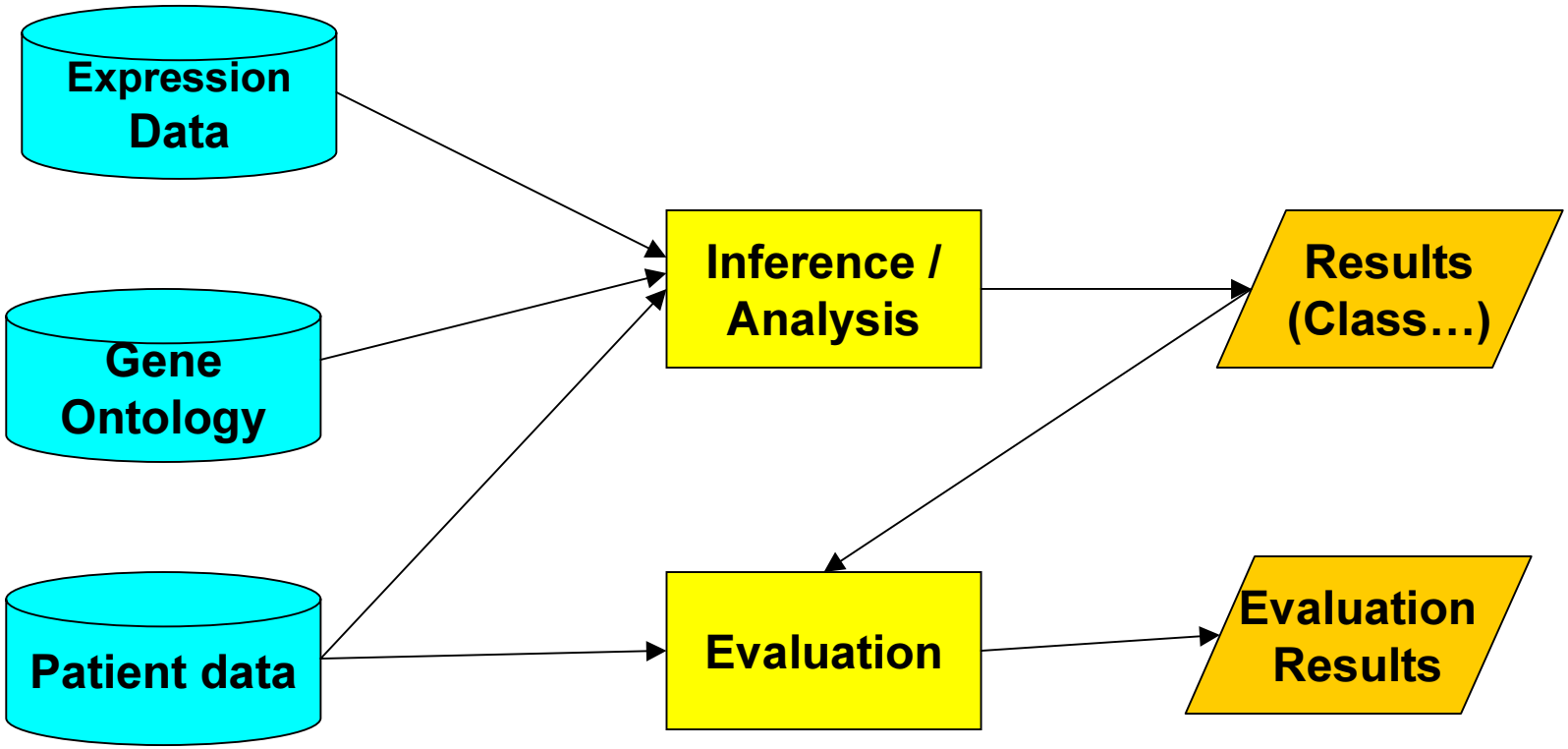
Overview

- Introduction - needs & techniques
- Data models and file formats
- Microarray data management solutions
- Our alternatives
- Summary

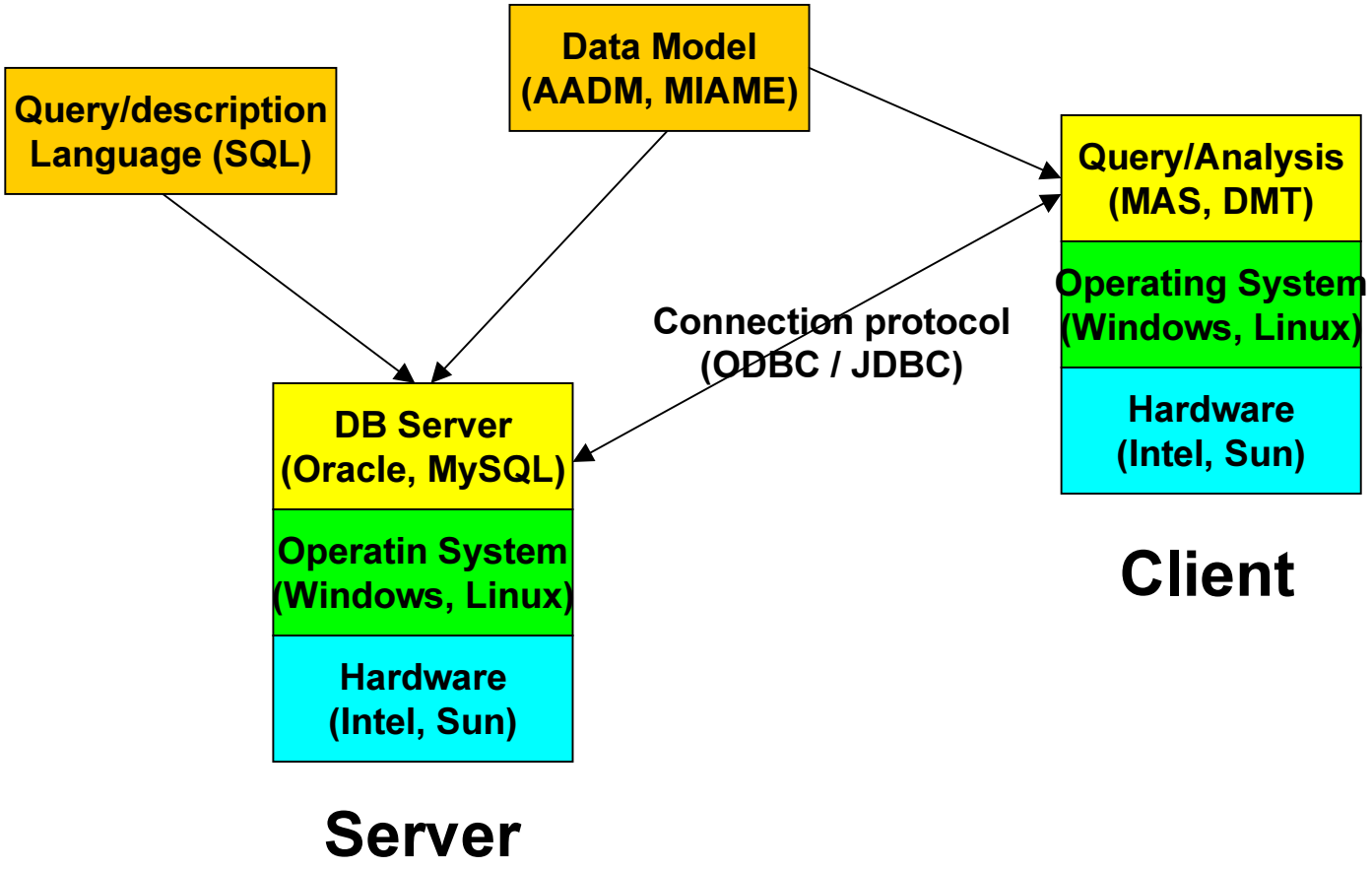
Some of Our Needs

- Exchange data with project partners
- Store considerable amount of data:
  - centralised, accessible for all interested partners
  - structured for mostly simple queries
- Compatibility to analysis software development packages (Matlab, R, Perl, Java...)
- Longterm maybe: WWW connection
  - data acquisition
  - analysis presentation

Kinds of Data



Client/Server Database Architecture



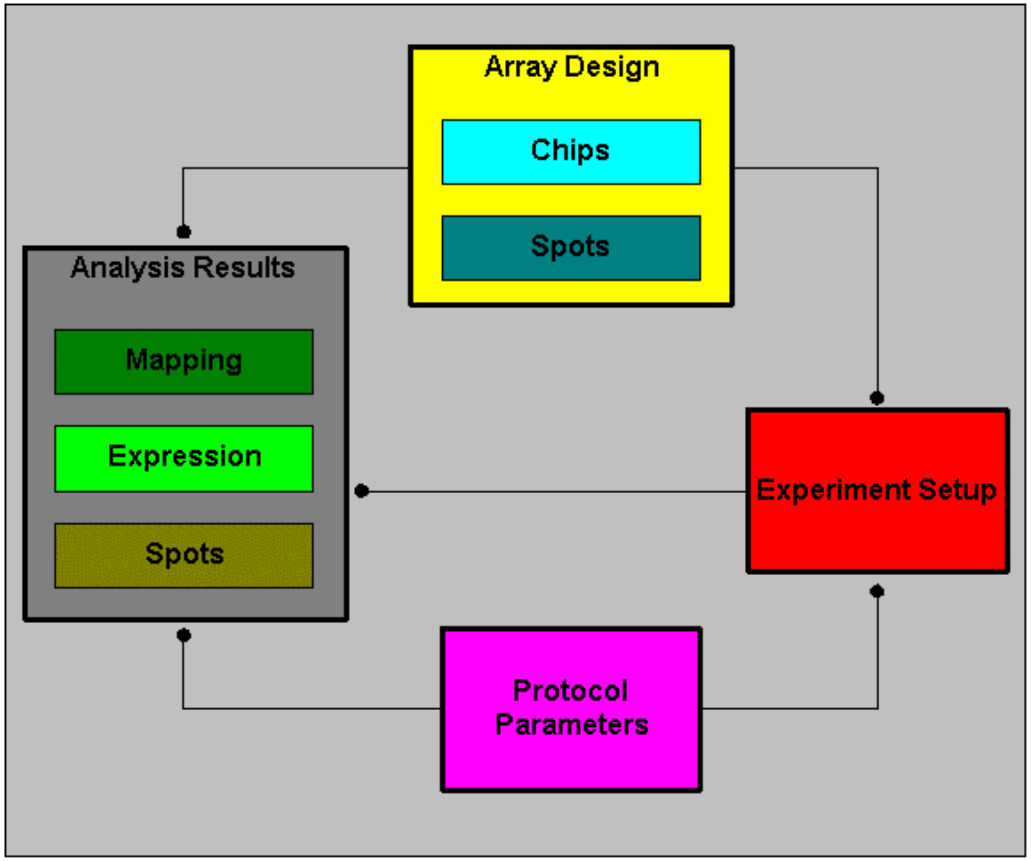
Relational Data Models

- Data entities
- Attributes
- Relations
  - one to one
  - one to many
  - many to many

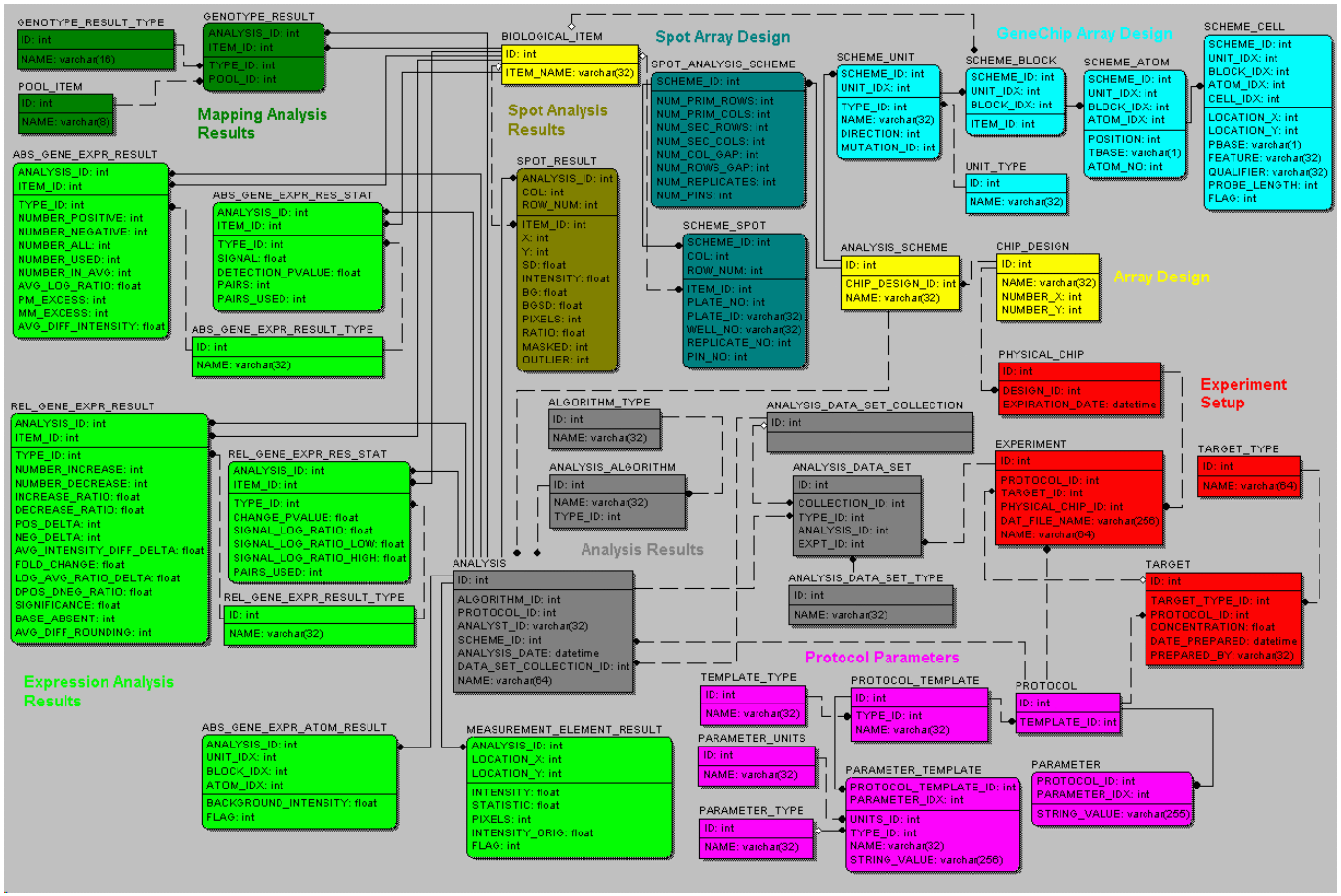
AADM: Affymetrix Analysis Data Model - Characteristics

- Proprietary but published
  (http:/www.affymetrix.com/support/aadm/aadm.html)
- Initialisation scripts for Oracle and MS-SQL Server
- Affymetrix software exchanges data using AADM
  - Instruments write to the DB
  - Microarray Suite (MAS) reads/writes in DB
  - Data Mining Tool (DMT) reads from DB

AADM: Affymetrix Analysis Data Model - General Structure



AADM: Affymetrix Analysis Data Model - Detailed Structure



AADM: Affymetrix Analysis Data Model - Sample Query

- Return CEL intensity for A28102_at of analysis T1_r1

```
Select
    mer.location_x, mer.location_y, mer.intensity, mer.statistic, mer.pixels
  from
    measurement_element_result mer, analysis aCel, analysis_data_set dsCel,
    experiment e, physical_chip pc, analysis_scheme ans, scheme_block sb,
    biological_item bi, scheme_cell sc
  where
    mer.analysis_id = aCel.id and
    aCel.data_set_collection_id = dsCel.collection_id and
    dsCel.expt_id = e.id and e.physical_chip_id = pc.id and
    pc.design_id = ans.id and ans.id = sb.scheme_id and
    sb.item_id = bi.id and sb.unit_idx = sc.unit_idx and
    sb.block_idx = sc.block_idx and sb.scheme_id = sc.scheme_id and
    sc.location_x = mer.location_x and sc.location_y = mer.location_y and
    aCel.name = 'T1_r1' and bi.item_name = 'A28102_at'
  order by
    sc.location_y, sc.location_x
```

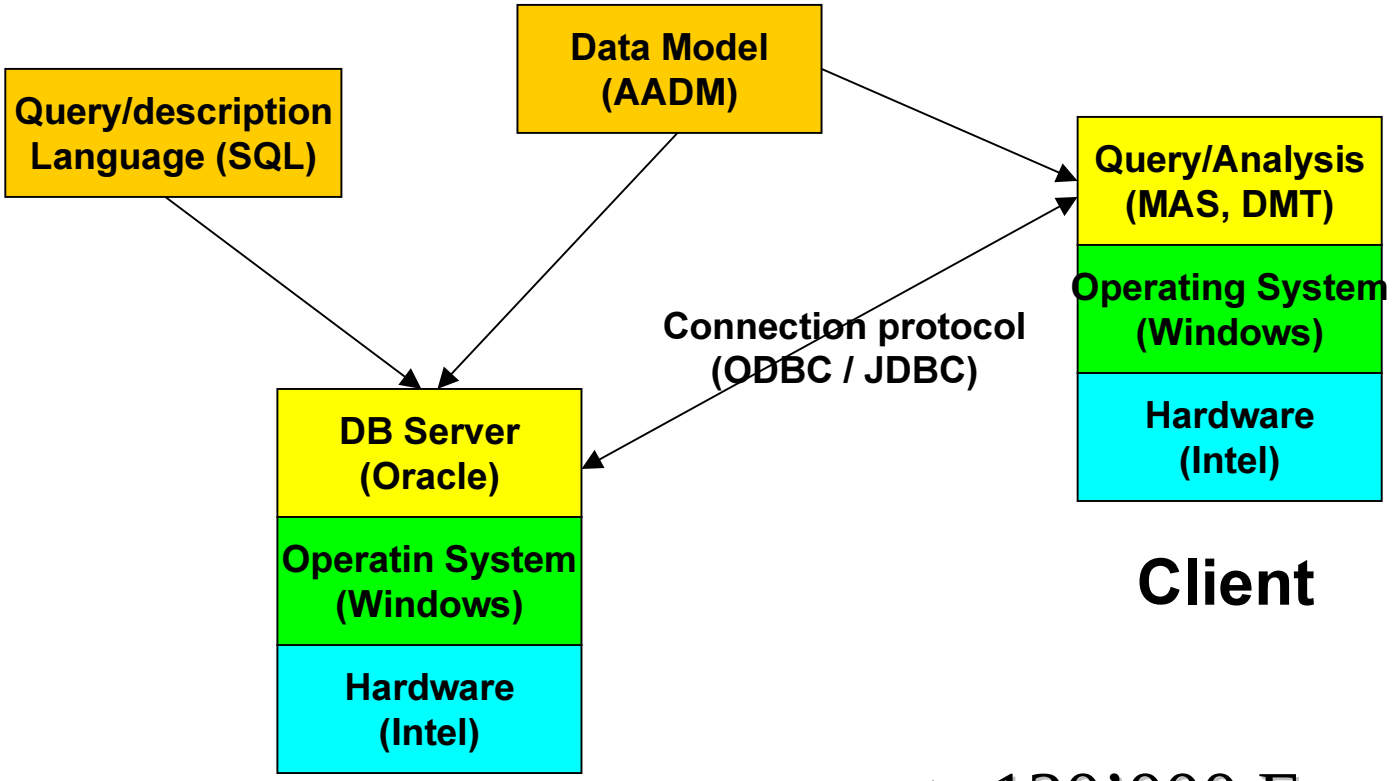
MIAME: Minimal Information about Microarry Experiments

- Guidelines in free text
- Public consortium
- Reproducibility of results is a goal
- MAGE-OM: object model fulfilling MIAME-criteria
- Flexible enough for many techniques
- ArrayExpress follows these guidelines

Data Exchange - File Formats

- Flat text files generated by analysis tools
  - have to be parsed
  - undocumented, may change in new versions
- Backup dumps by databse management systems
  - incompatible among DB management systems
  - incompatible across platforms
- Text files containing SQL commands
  - not the fastest but most standard
  - still incompatibilities, proprietary extensions
  - can DBMSs write such files?

LIMS: Laboratory Information Management System



- > 130’000 Euros

ArrayExpress

- Implements MIAME guidelines and MAGE-OM
- Has a WWW interface
- Provides clustering and visualisation
- Runs on Oracle
- Some scripts are available but use Oracle specific features

GeneX

- Available as a package
- Based on PostgreSQL (freeware)
- WWW interface
- Designed to build worldwide data repository
- Complex to install and use
- Works together with some clustering tools

Charité's Choice

- Various data analysis tools available (also the ones by Affymetrix)
- One database per group on Microsoft SQL servers ("light" version) spread across several machines
- MS Windows is the operating system
- Soon a full version is to be bought, should be able to export into SQL-files
- Everything is entirely hidden behind a fire-wall

Installation of GeneX or Similar under Linux

- Rather complex operation
- Adaptation to GeneChip needed
- GeneX is based on PostgreSQL, we need another database server
- Data access through ODBC and JDBC should be possible from Matlab and others

Installation of Microsoft SQL Server under Windows

- Direct use of MS SQL Server binary database dumps
- An additional Windows machine would be needed
- An additional software license would be needed
- ODBC / JDBC connection across platforms seems feasible but may be tricky

Own Solution based on MySQL

- Either parsing flat files with raw data
  - into our own simplified data model
  - or into a publically available data model
- Or hoping for SQL database dumps from our project partners
- In any case some non-standard work for patient data is needed
- WWW-upload is an alternative to DVDs by mail

Own Solution based on MySQL and netCDF

- Small database on patient, sample and genes
- Treat expression data as matrices
- netCDF provides cross-platform compatible store and retrieve facility for matrices
  - very dense compared to relational database
  - slower retrieve, faster storage than relational DBs
  - interfaces to Matlab, R, Perl, C and others available
- Shouldn’t we keep information about chip design, experiment setup etc.?

Summary

- Server solutions:
  - MySQL under Linux, netCDF optional
  - MS SQL Server under Wndows
  - GeneX or ArrayExpress
- Data models:
  - ADM for direct access to data submitted by partners
  - own simplified model for ease of use on our side
  - GeneX or MAGE-OM (MIAME)
- In any case additional work for patient data and gene ontology is needed