FlyMine

An integrated database for *Drosophila* and *Anopheles* genomics

Richard Smith
Genetics Department
University of Cambridge

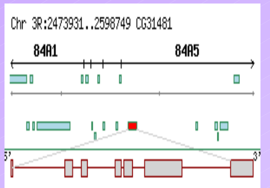
Outline

- What is FlyMine?
- What can you do with it?
- InterMine – generic data warehouse
- FlyMine – integrating biological data
- Data integration examples

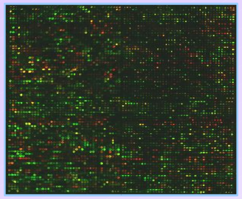
What is FlyMine?

- A data warehouse that integrates several genomic and proteomic data sets in one place.
- Main focus is *Drosophila melanogaster* and *Anopheles gambiae* with cross species comparasons with other organisms
- Website allows users to build arbitrary complex queries across all data.
- Focus is on complex queries in place rather than data access.

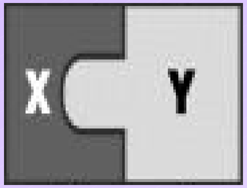
Data Sources



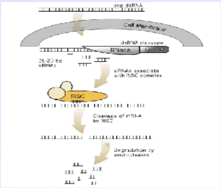
Genome Annotation



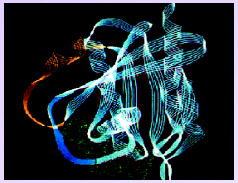
Microarray expression data (soon)



2-hybrid
Protein-Protein
Interactions

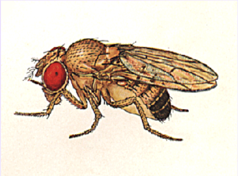
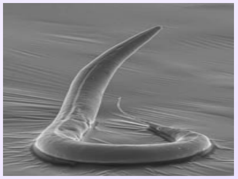


RNAi Phenotypes

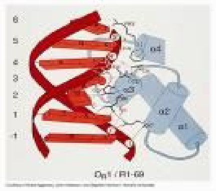


3-D structural
domain predictions

GO Annotation

 ↔ 

Ortholgues/
Paralogues

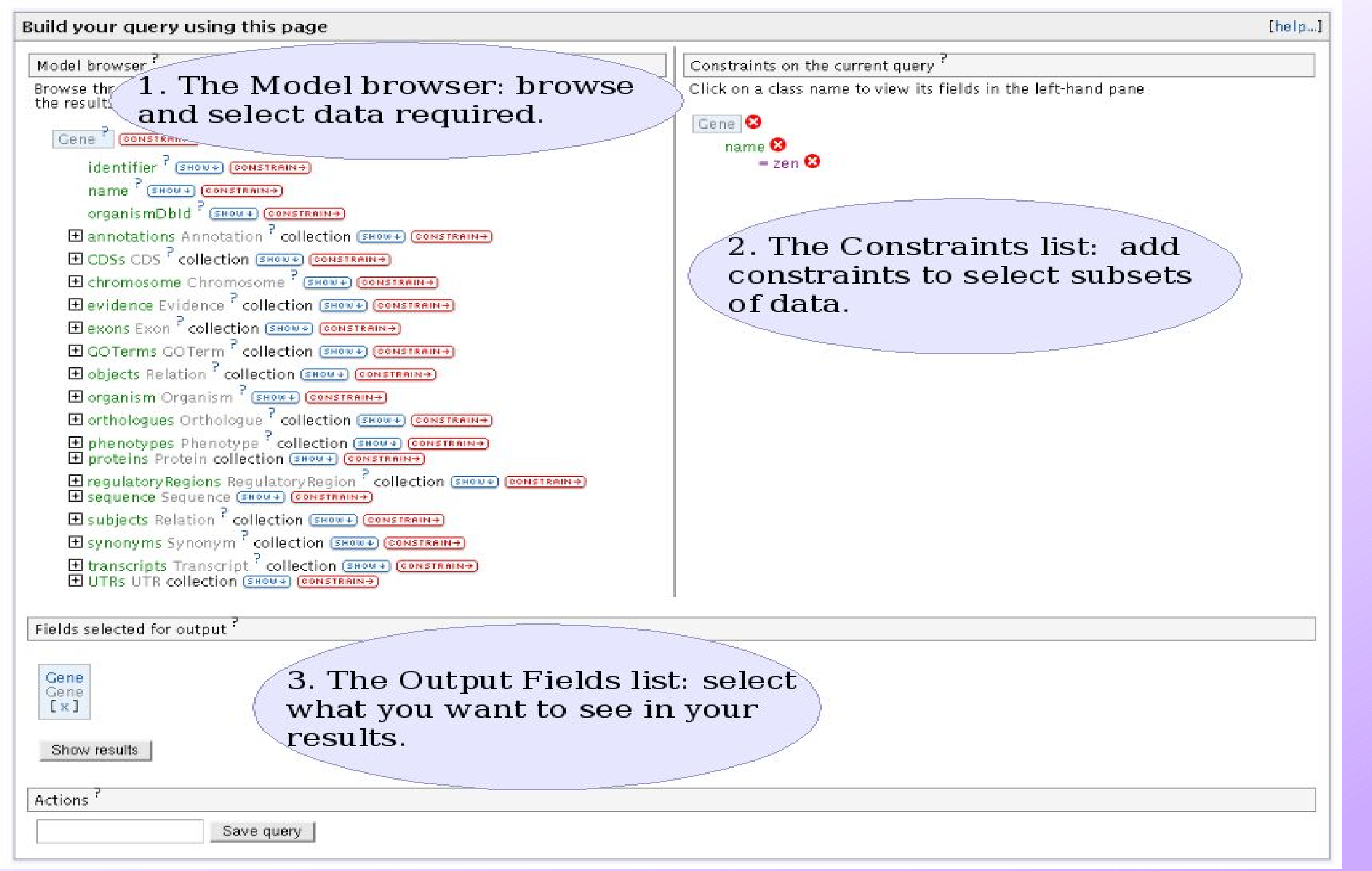


DNAase 1 footprints

Example Queries

- For proteins for which there is evidence that they interact, find whether the orthologues of these proteins also have evidence that they interact (ie search for interologs)
- Search for Drosophila genes associated with a particular transcription factor binding site and show the predicted orthologues for these genes.
- *Determine whether proteins that may interact are also co-expressed under any conditions*

The Query Builder:



Build your query using this page [help...]

Model browser ?

Browse thr
the result

1. The Model browser: browse and select data required.

Gene ? CONSTRAIN→

identifier ? SHOW↓ CONSTRAIN→
name ? SHOW↓ CONSTRAIN→
organismDbId ? SHOW↓ CONSTRAIN→
⊞ annotations Annotation ? collection SHOW↓ CONSTRAIN→
⊞ CDSs CDS ? collection SHOW↓ CONSTRAIN→
⊞ chromosome Chromosome ? SHOW↓ CONSTRAIN→
⊞ evidence Evidence ? collection SHOW↓ CONSTRAIN→
⊞ exons Exon ? collection SHOW↓ CONSTRAIN→
⊞ GOTerms GOTerm ? collection SHOW↓ CONSTRAIN→
⊞ objects Relation ? collection SHOW↓ CONSTRAIN→
⊞ organism Organism ? SHOW↓ CONSTRAIN→
⊞ orthologues Orthologue ? collection SHOW↓ CONSTRAIN→
⊞ phenotypes Phenotype ? collection SHOW↓ CONSTRAIN→
⊞ proteins Protein collection SHOW↓ CONSTRAIN→
⊞ regulatoryRegions RegulatoryRegion ? collection SHOW↓ CONSTRAIN→
⊞ sequence Sequence SHOW↓ CONSTRAIN→
⊞ subjects Relation ? collection SHOW↓ CONSTRAIN→
⊞ synonyms Synonym ? collection SHOW↓ CONSTRAIN→
⊞ transcripts Transcript ? collection SHOW↓ CONSTRAIN→
⊞ UTRs UTR collection SHOW↓ CONSTRAIN→

Constraints on the current query ?

Click on a class name to view its fields in the left-hand pane

Gene ☒
name ☒
= zen ☒

2. The Constraints list: add constraints to select subsets of data.

Fields selected for output ?

Gene
Gene
[x]

Show results

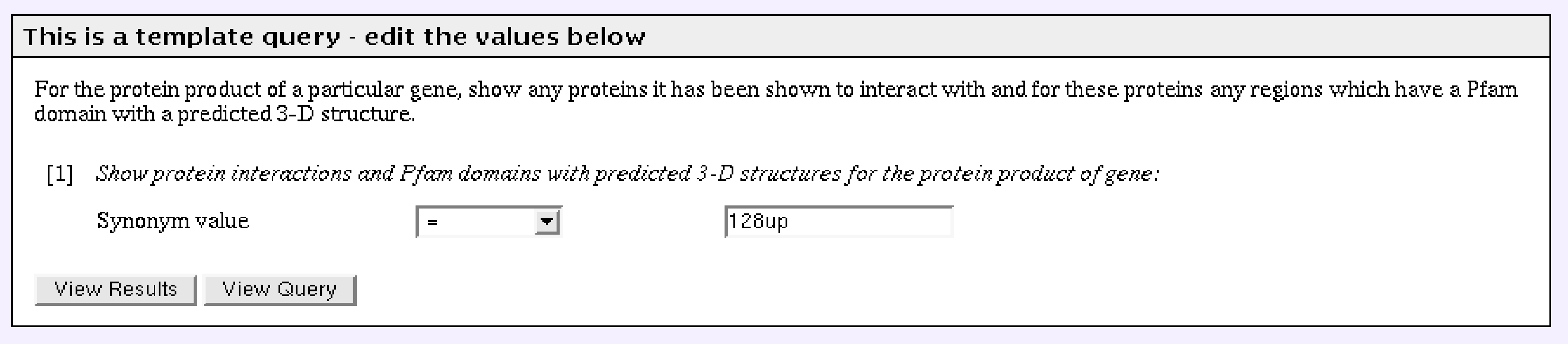
3. The Output Fields list: select what you want to see in your results.

Actions ?

Save query

Template Queries

- For the protein product of a particular gene, show any proteins it has been shown to interact with, the Pfam domains of these proteins and any predicted structure for the domains.



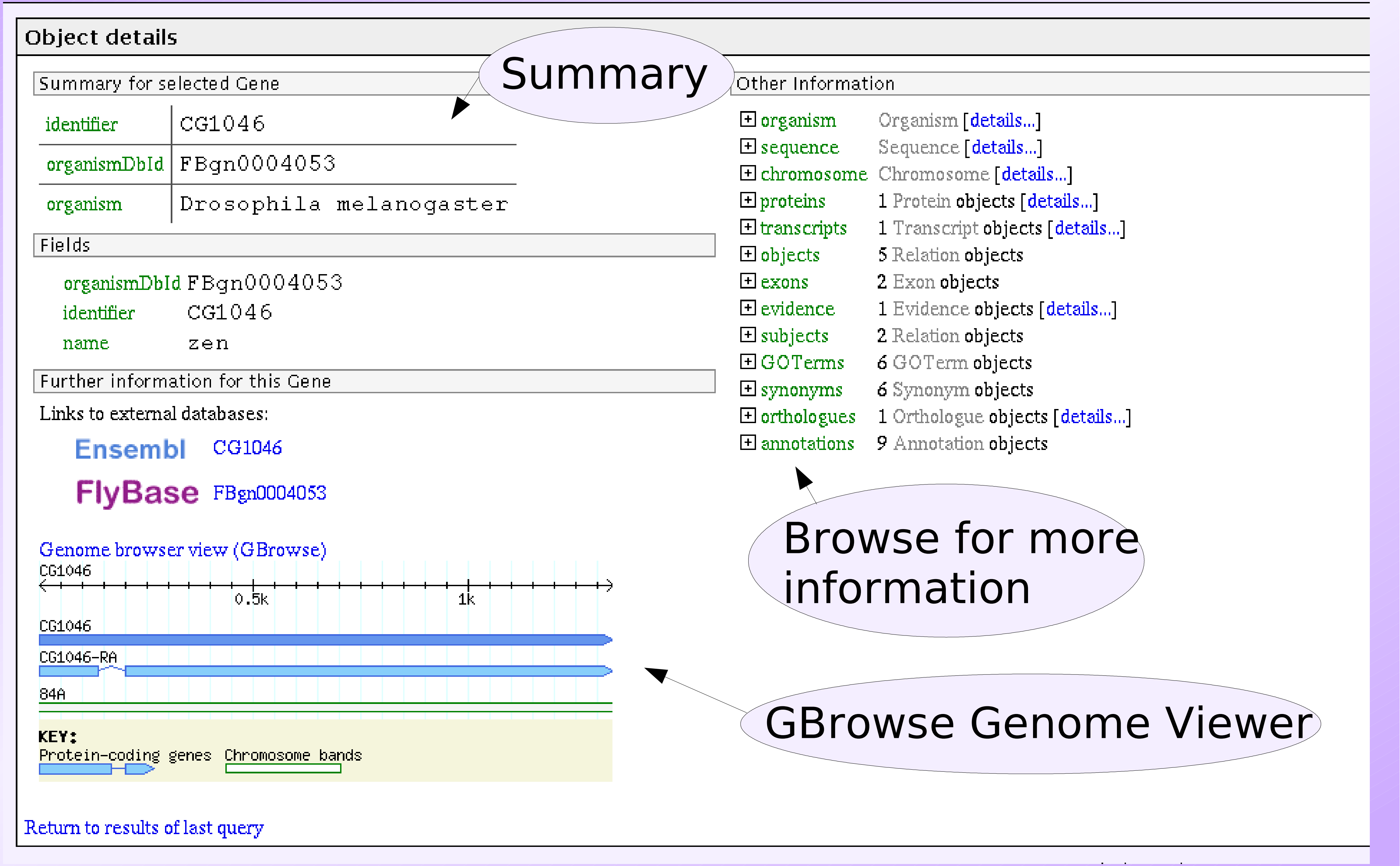
This is a template query - edit the values below

For the protein product of a particular gene, show any proteins it has been shown to interact with and for these proteins any regions which have a Pfam domain with a predicted 3-D structure.

[1] *Show protein interactions and Pfam domains with predicted 3-D structures for the protein product of gene:*

Synonym value = 128up

Browsable Details Page:



Object details

Summary for selected Gene

identifier	CG1046
organismDbId	FBgn0004053
organism	Drosophila melanogaster

Fields

organismDbId	FBgn0004053
identifier	CG1046
name	zen

Further information for this Gene

Links to external databases:

Ensembl	CG1046
FlyBase	FBgn0004053

Genome browser view (GBrowse)

Return to results of last query

Other Information

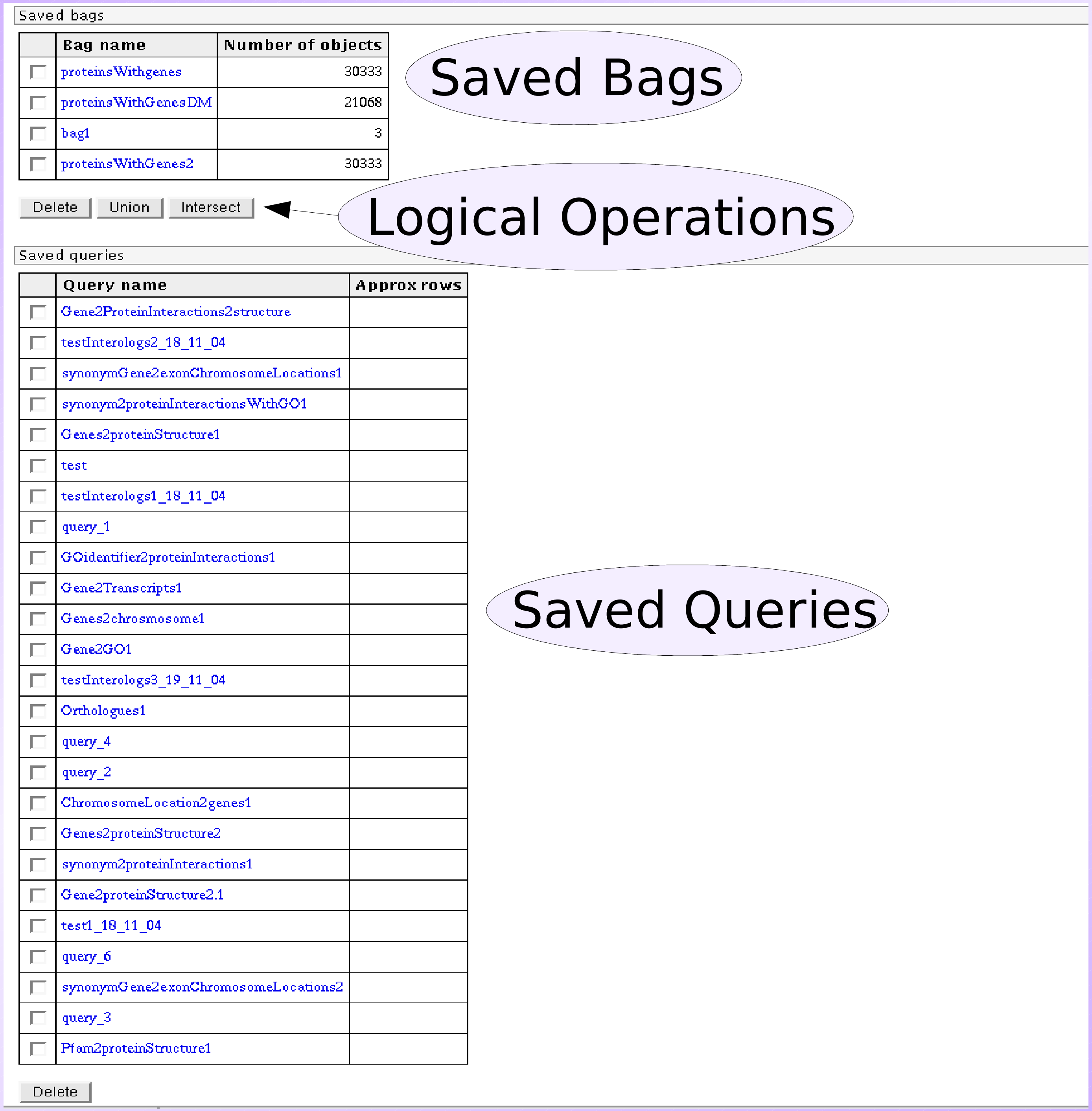
⊞ organism	Organism [details...]
⊞ sequence	Sequence [details...]
⊞ chromosome	Chromosome [details...]
⊞ proteins	1 Protein objects [details...]
⊞ transcripts	1 Transcript objects [details...]
⊞ objects	5 Relation objects
⊞ exons	2 Exon objects
⊞ evidence	1 Evidence objects [details...]
⊞ subjects	2 Relation objects
⊞ GOTerms	6 GOTerm objects
⊞ synonyms	6 Synonym objects
⊞ orthologues	1 Orthologue objects [details...]
⊞ annotations	9 Annotation objects

Summary

Browse for more information

GBrowse Genome Viewer

Other Features:



Saved bags

	Bag name	Number of objects
□	proteinsWithgenes	30333
□	proteinsWithGenesDM	21068
□	bag1	3
□	proteinsWithGenes2	30333

Saved Bags

Logical Operations

Saved queries

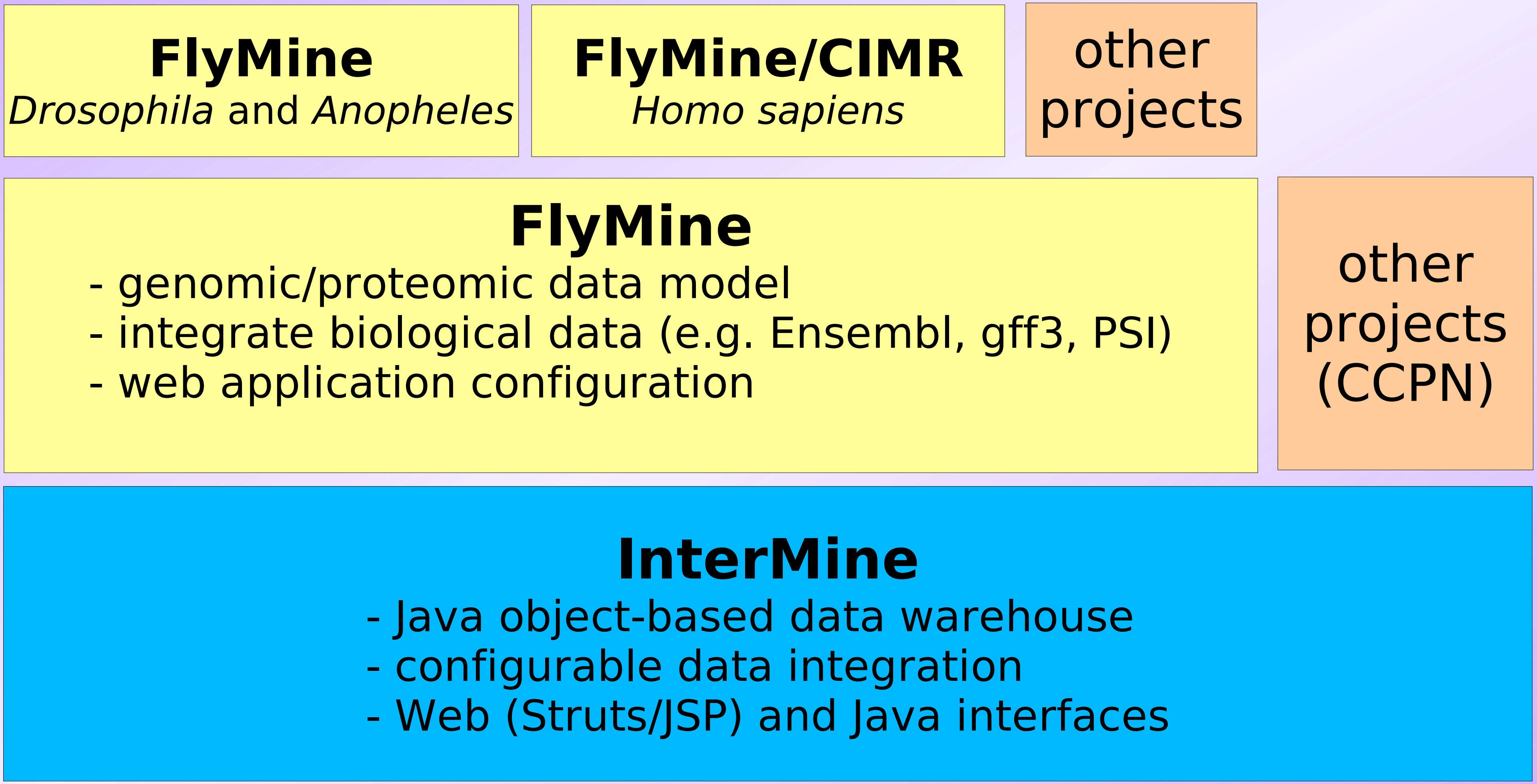
	Query name	Approx rows
□	Gene2ProteinInteractions2structure	
□	testInterologs2_18_11_04	
□	synonymGene2exonChromosomeLocations1	
□	synonym2proteinInteractionsWithGO1	
□	Genes2proteinStructure1	
□	test	
□	testInterologs1_18_11_04	
□	query_1	
□	GOidentifier2proteinInteractions1	
□	Gene2Transcripts1	
□	Genes2chrosmosome1	
□	Gene2GO1	
□	testInterologs3_19_11_04	
□	Orthologues1	
□	query_4	
□	query_2	
□	ChromosomeLocation2genes1	
□	Genes2proteinStructure2	
□	synonym2proteinInteractions1	
□	Gene2proteinStructure2.1	
□	test1_18_11_04	
□	query_6	
□	synonymGene2exonChromosomeLocations2	
□	query_3	
□	Pfam2proteinStructure1	

Saved Queries

Software Goals

- Integrate multiple data sets into a single data warehouse
- Flexible - continual addition of new types of data
- Powerful - allow complex queries not known at design time to run on the data warehouse
- Robust – reject queries that would take too long

FlyMine Software



FlyMine
Drosophila and *Anopheles*

FlyMine/CIMR
Homo sapiens

other projects

FlyMine

- genomic/proteomic data model
- integrate biological data (e.g. Ensembl, gff3, PSI)
- web application configuration

other projects (CCPN)

InterMine

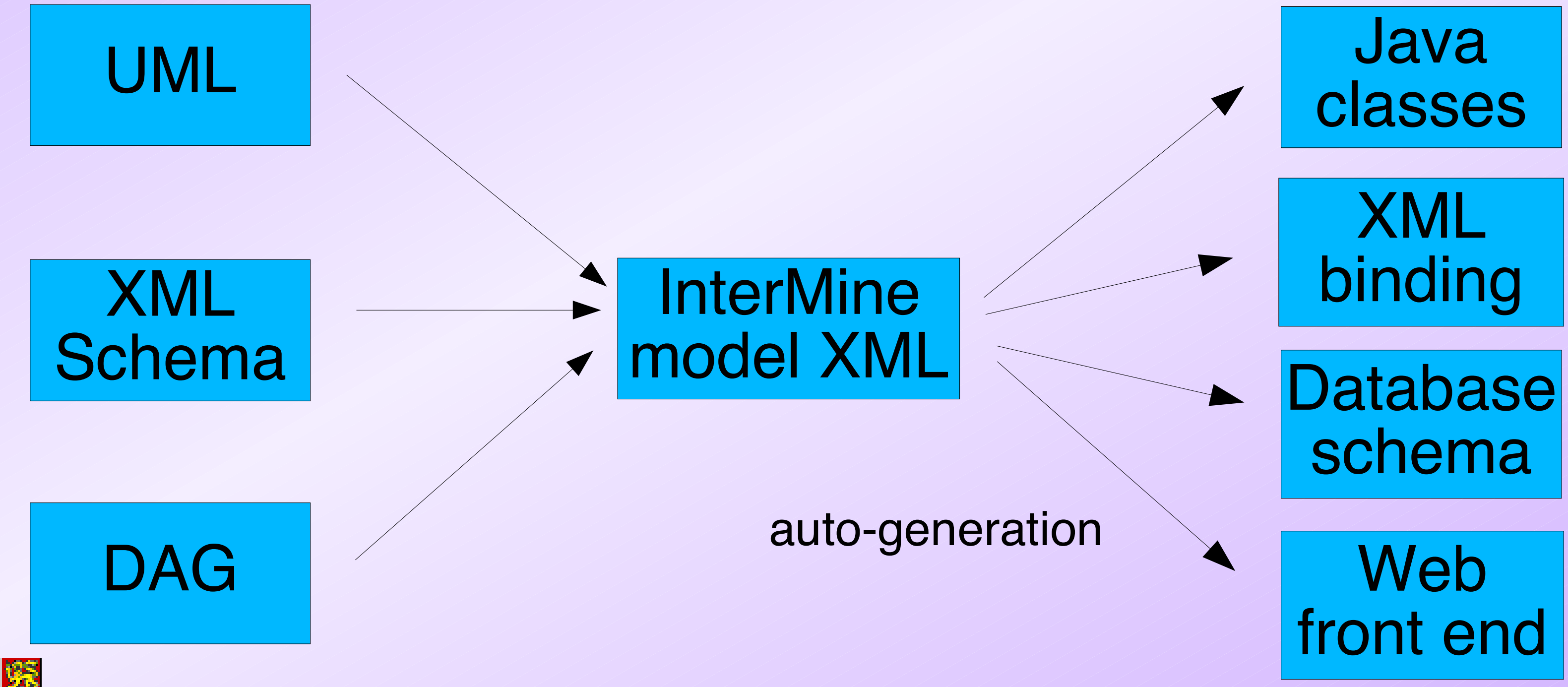
- Java object-based data warehouse
- configurable data integration
- Web (Struts/JSP) and Java interfaces

InterMine

- Object/relational query system
- Data model independent
- Query optmisation system – pre-computed tables/query re-writing
- Flexible data integration
- Retrieve data from RDBMS, XML, flat files
- Web application (Struts/JSP)

InterMine - Model Independent

- auto-generation from XML model definition
- single model or merge multiple



FlyMine

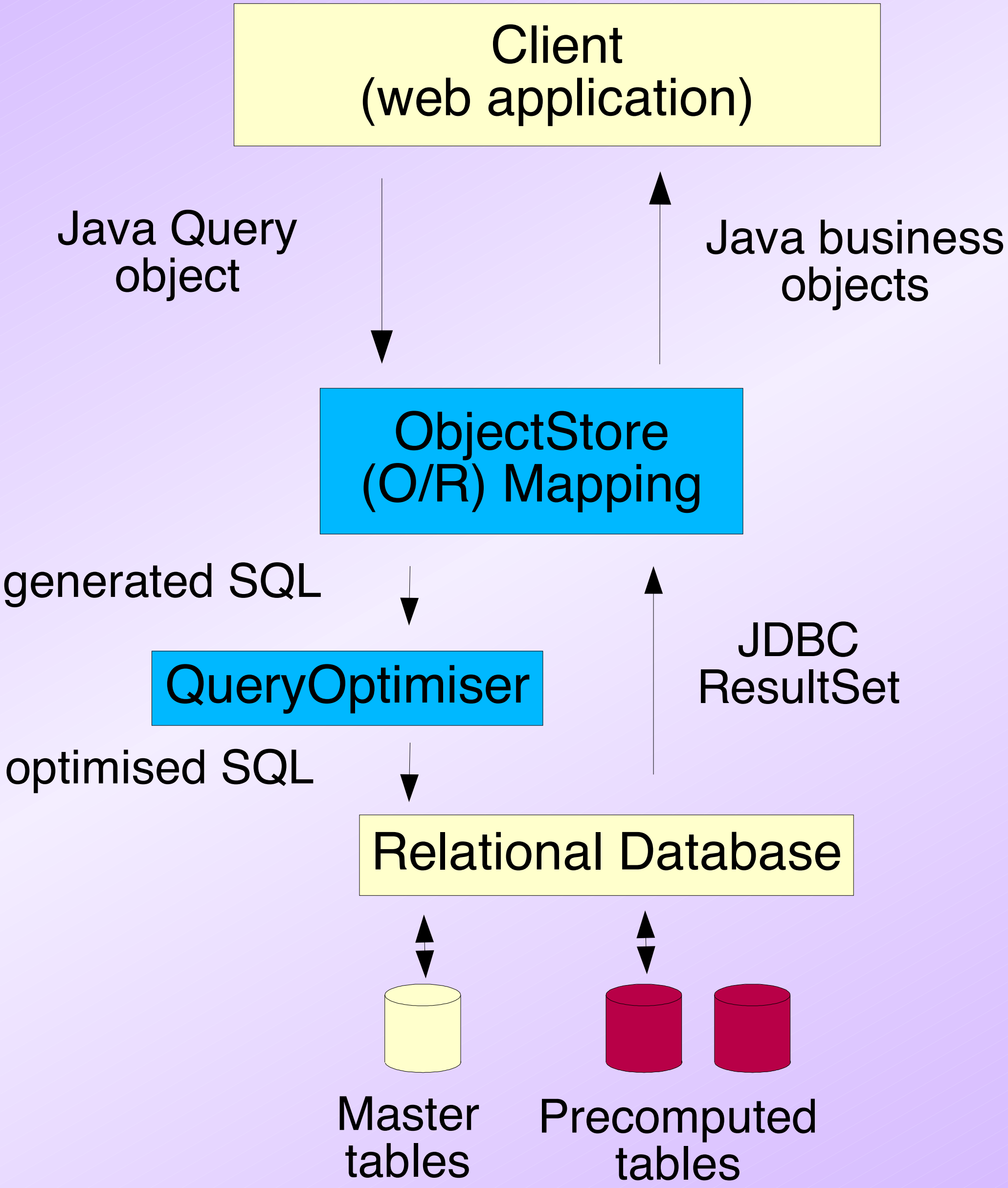
Arbitrary queries – problems

- Badly formed queries may overload database server
- Difficult to optimise database for all queries
  - Which indexes to use?
  - Slow response to complex queries involving multi-table joins

Arbitrary queries – solutions

- Close relationship with database server query planner
  - ask how long a query will take before attempting to run it (~3ms)
  - Disallow queries that will take longer than a certain threshold
- Store data massively redundantly in “precomputed tables” and rewrite incoming queries on-the-fly.

InterMine- Query Optimisation



Client
(web application)

Java Query
object

Java business
objects

ObjectStore
(O/R) Mapping

generated SQL

QueryOptimiser

JDBC
ResultSet

optimised SQL

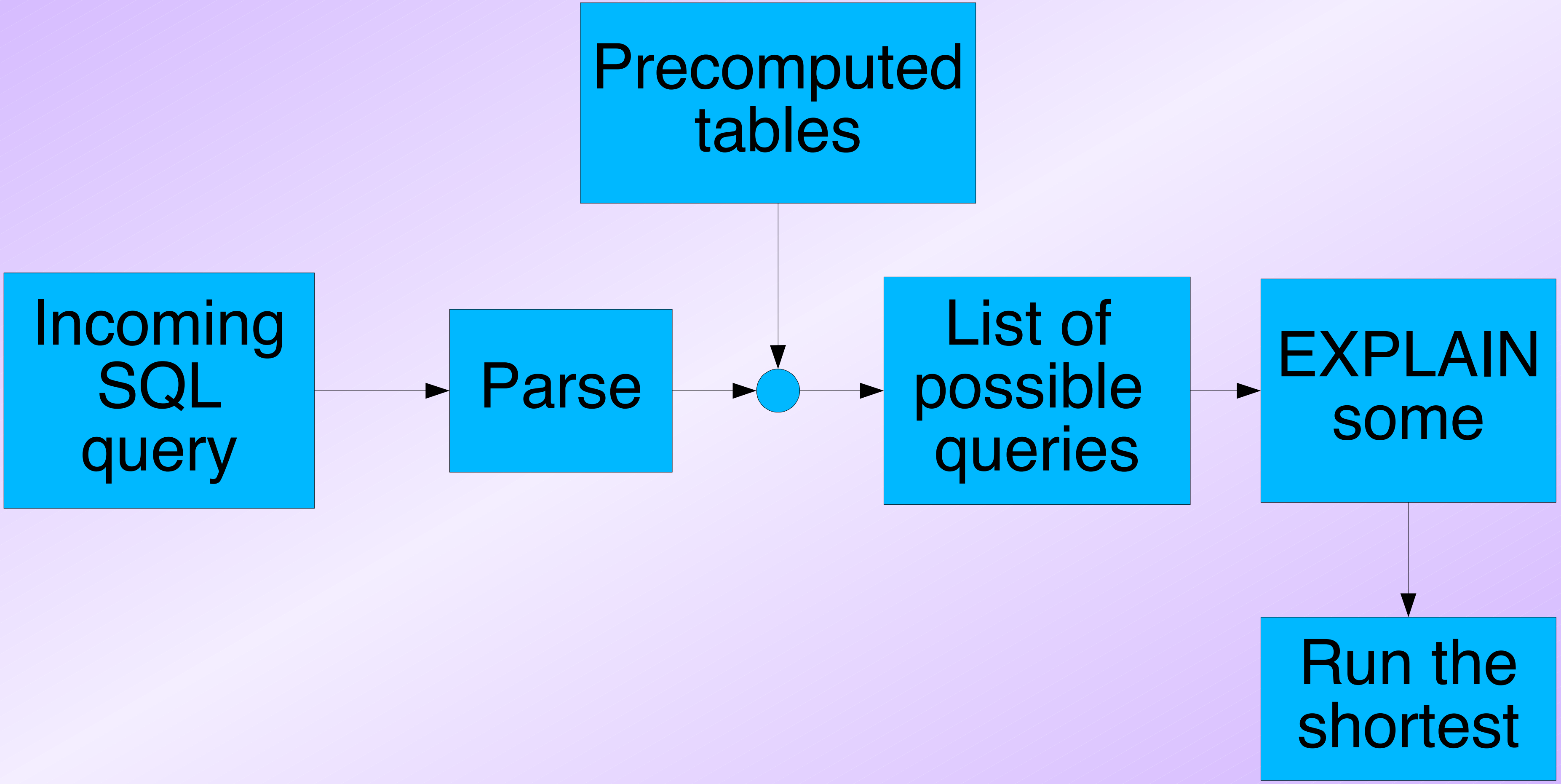
Relational Database

Master
tables

Precomputed
tables

FlyMine

Query optimisation



O/R Data Warehouse

- No need to design a query optimised schema
- Query optimisation moved away from business model
- Pre-computed tables can be added any time after build – **adapt query optimisation to usage**
- Pre-compute template queries

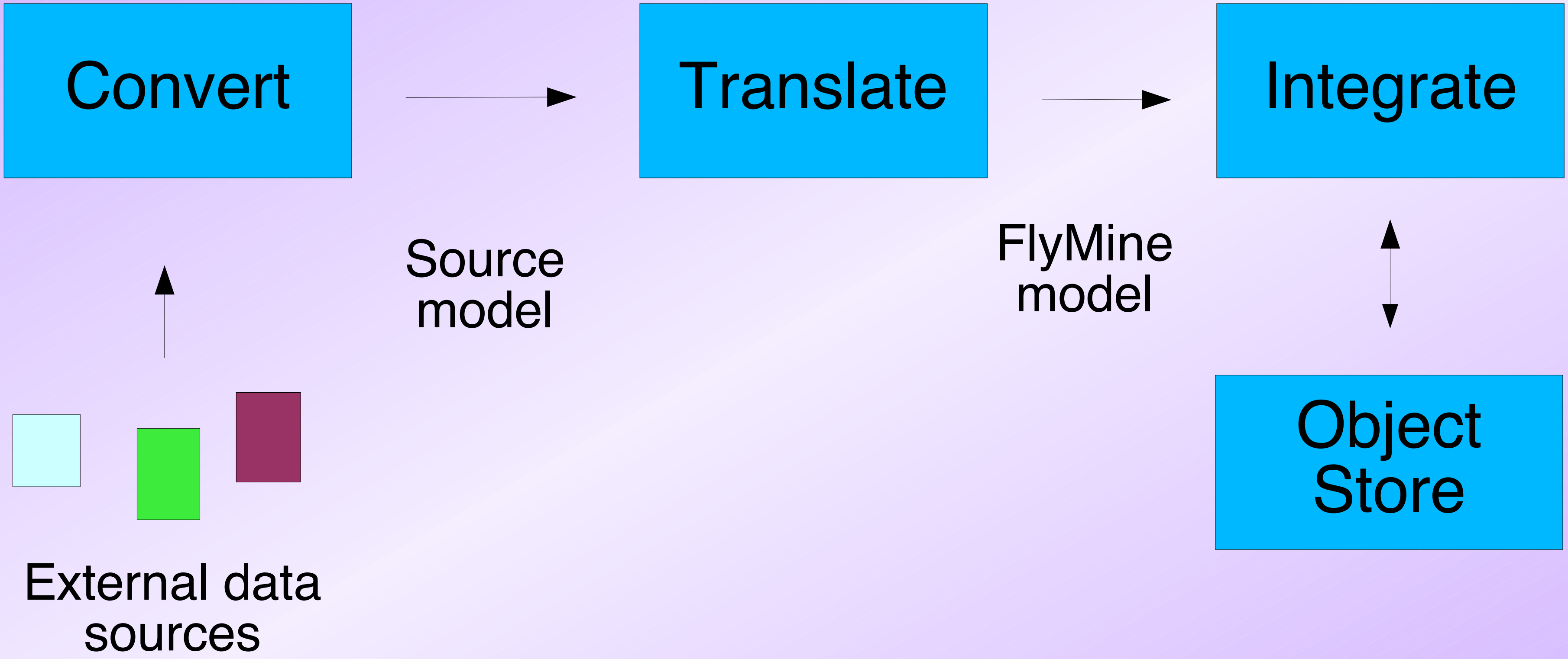
FlyMine – biological model

- Avoid creating a “schema of everything”
- Model based on SOFA
- Data sets add classes to model
  - e.g. orthologue data adds Orthologue/Paralogue classes
- Data sets add fields to model
  - e.g. ensembl adds length field to Contig

FlyMine – data sets

- Based on standards
  - PSI, MAGE, GFF3, DAG
- And common formats
  - Ensembl, Uniprot XML, Inparanoid output
- Configure generic code, handlers
- Very little is organism specific

Data loading pipeline



Convert

Translate

Integrate

Source model

FlyMine model

Object Store

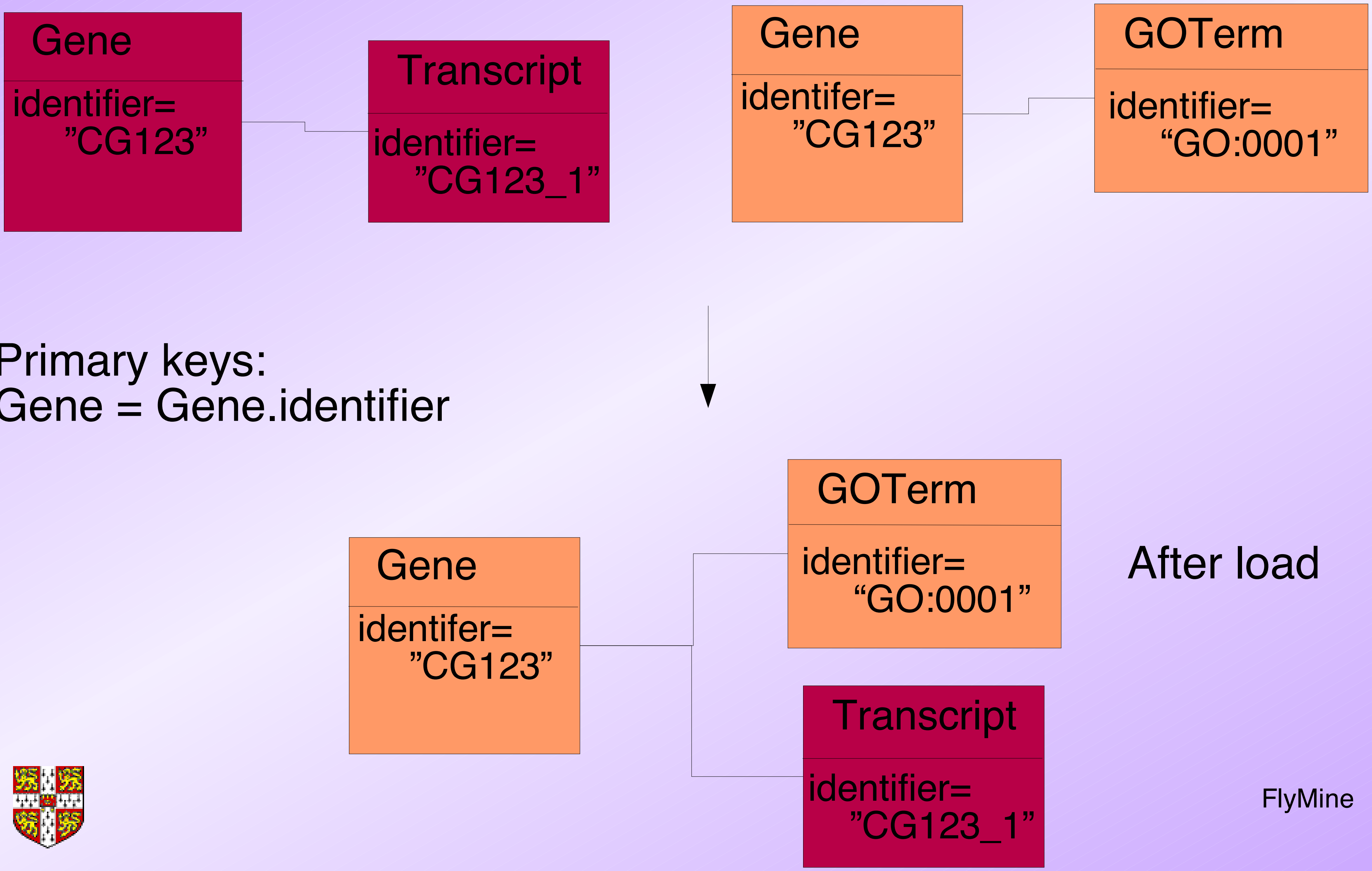
External data sources

Integration

- Define “primary keys” for each class
- Define “primary keys” that each source uses
- Define priorities for fields from different sources
- IntegrationWriter keeps track of originating sources for each field of each object

New objects

Currently in database



Gene

identifier=
”CG123”

Transcript

identifier=
”CG123_1”

Gene

identifer=
”CG123”

GOTerm

identifier=
“GO:0001”

Primary keys:
Gene = Gene.identifier

GOTerm

identifier=
“GO:0001”

Gene

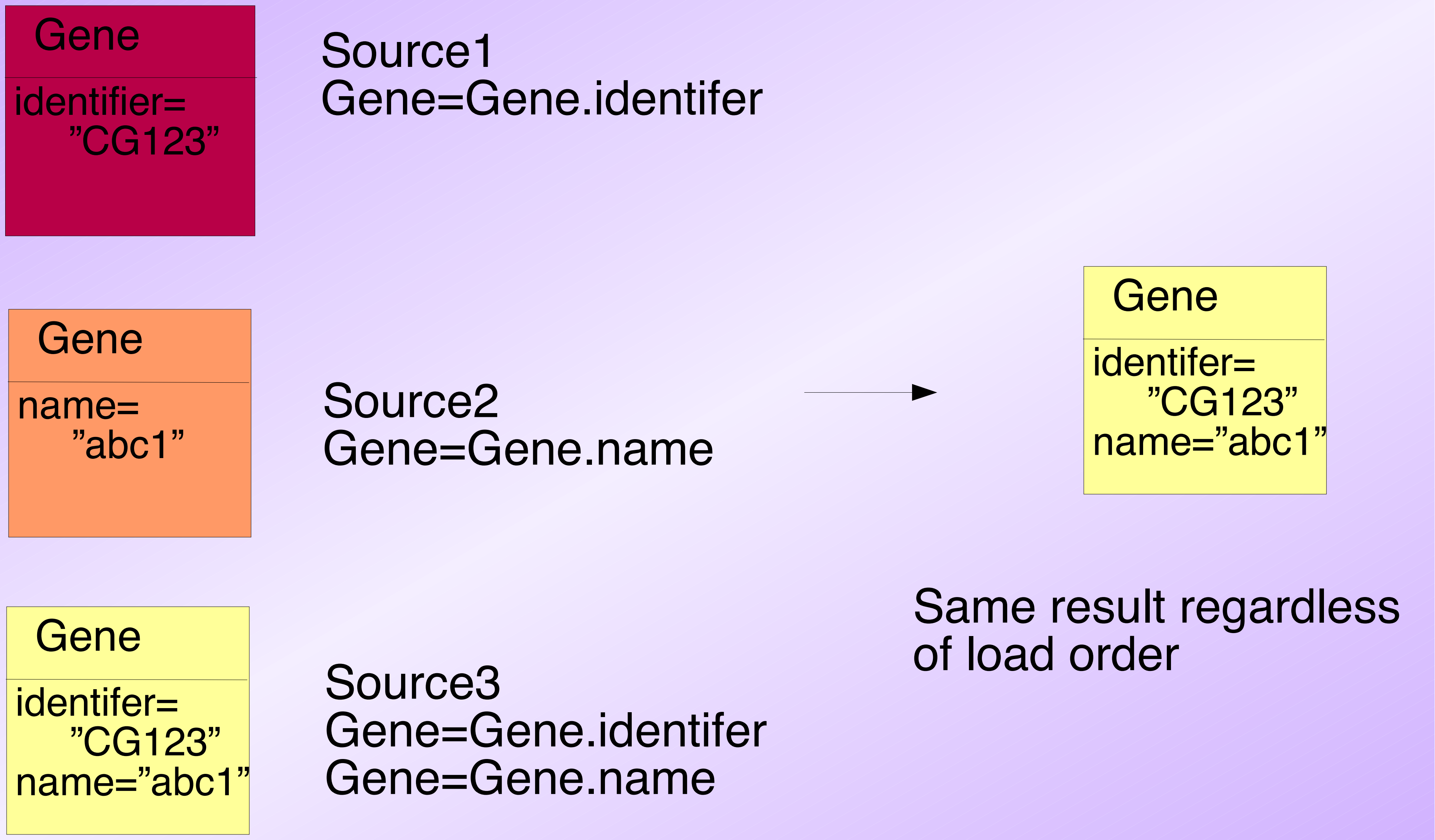
identifer=
”CG123”

Transcript

identifier=
”CG123_1”

After load

FlyMine



Gene

identifier=
"CG123"

Source1
Gene=Gene.identifer

Gene

name=
"abc1"

Source2
Gene=Gene.name

Gene

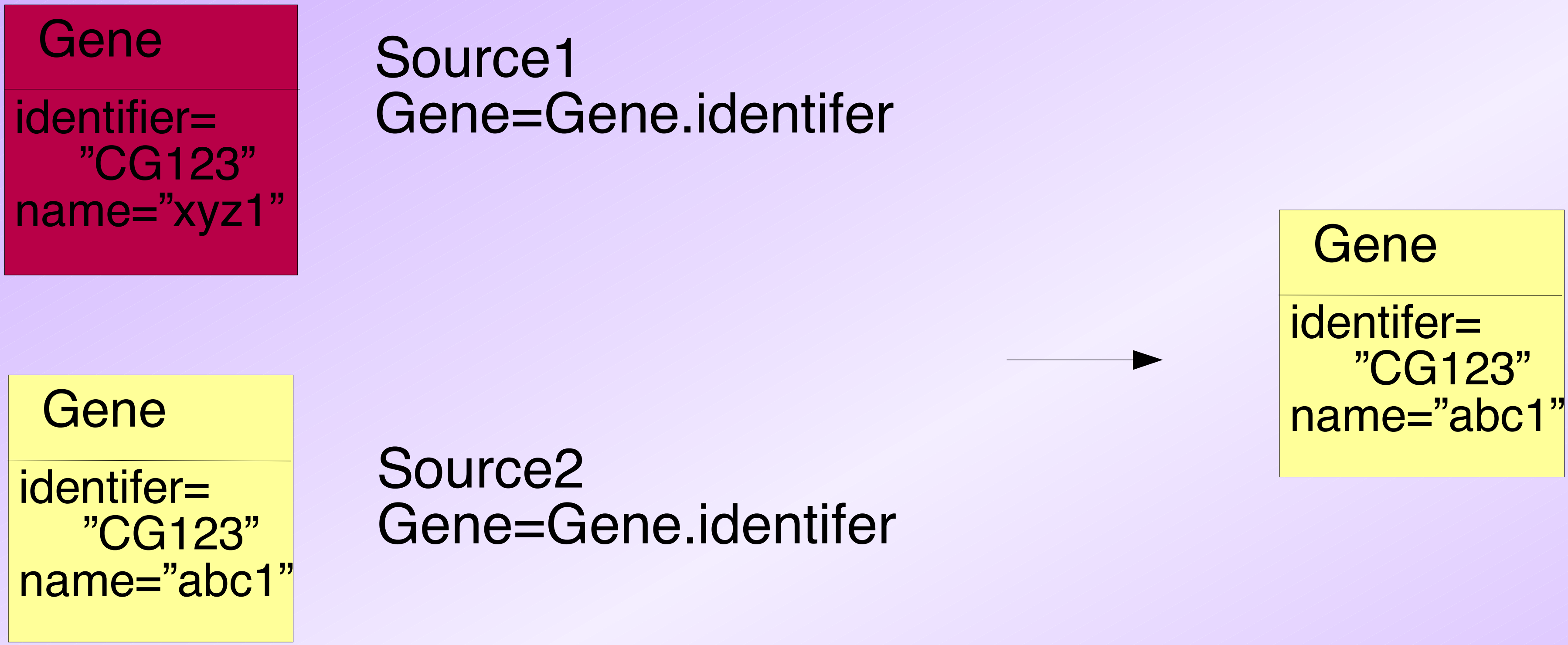
identifer=
"CG123"
name="abc1"

Source3
Gene=Gene.identifer
Gene=Gene.name

Gene

identifer=
"CG123"
name="abc1"

Same result regardless of load order



Gene

identifier=
"CG123"
name="xyz1"

Source1
Gene=Gene.identifer

Gene

identifer=
"CG123"
name="abc1"

Source2
Gene=Gene.identifer

Gene

identifer=
"CG123"
name="abc1"

Same result regardless of load order

Priorities:
Gene.name = Source2, Source1

Acknowledgements

Richard Smith
Kim Rutherford
Matthew Wakeling
Tom Riley
Wenyan Ji

Rachel Lyne
François Guillier
Debashis Rana
Gos Micklem

former
Andy Varley, Mark Woodbridge

More information at www.flymine.org

FlyMine is funded by the Wellcome Trust (grant no. 067205), awarded to M. Ashburner, G. Micklem, S. Russell, K. Lilley and K. Mizuguchi.