Biological Data Analysis using InterMine

Rachel Lyne
Daniela Butano
Kevin Reierskog
Sergio Contrino

November 16th 2020

UNIVERSITY OF CAMBRIDGE

InterMine

Workshop Schedule

- Introduction to InterMine
- Walk through interface with hands-on exercises
- Questions (at intervals)
- BlueGenes interface demo
- Use-cases and questions

- We will take a short coffee break at a convenient point in the workshop
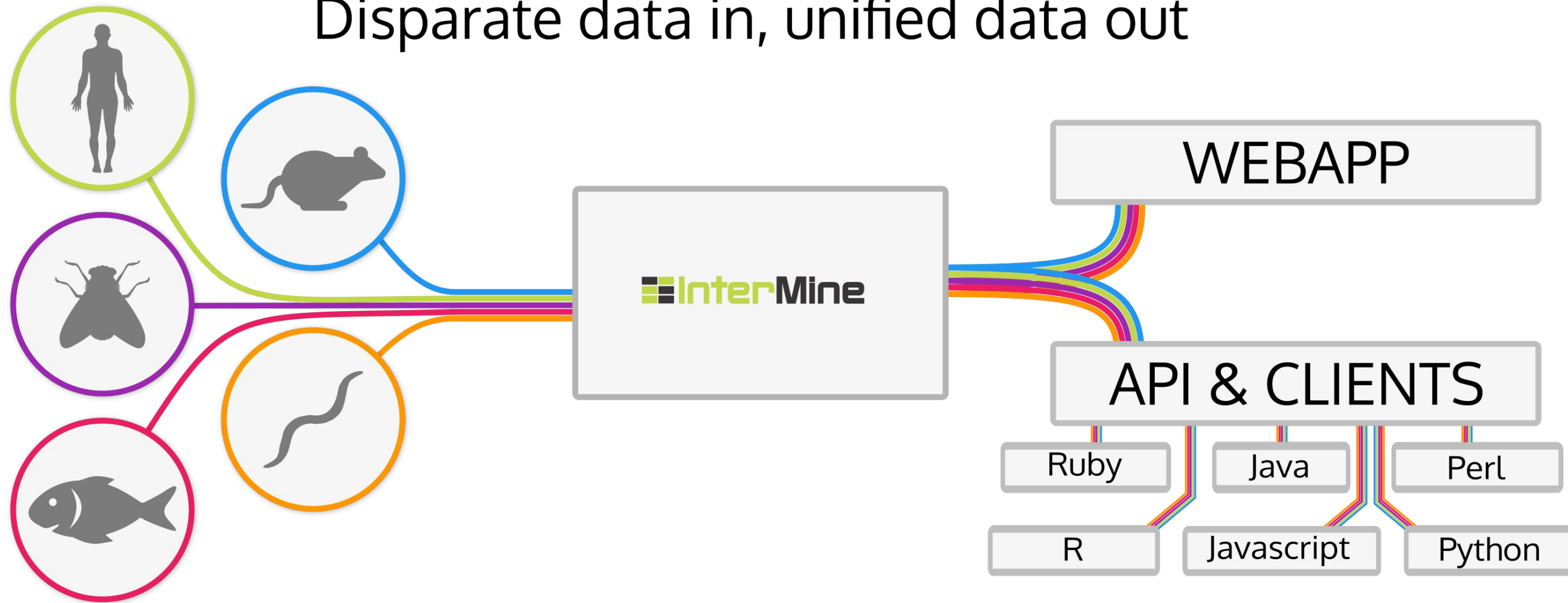
www.flymine.org
www.humanmine.org

Questions

Please add to the Q&A section as we go along.

We will review and answer and select some (or all) to go over in the questions sections.

What is InterMine

Disparate data in, unified data out

WEBAPP

API & CLIENTS

Ruby Java Perl

R Javascript Python

Model organism images Designed by Freepik and distributed by Flaticon

UNIVERSITY OF CAMBRIDGE

InterMine

Who Uses InterMine?

UNIVERSITY OF CAMBRIDGE

InterMine

http://registry.intermine.org/

InterMine Registry All InterMine instances up-to-date information in one place.

	Name	Description	Organisms
Icon	BMAP	Brassicales Map Alignment Project	
	BeanMine	A mine with common bean data from the Legume Info tripal.chado database	A. ipaensis, A. duranensis, A. thalia...
	BovineMine	An integrated data warehouse for the Bovine Genome Database	B. taurus, C. hircus, O. aries
	CHOmine	An integrated database for Cricetulus griseus and CHO cells	C. griseus
	ChickpeaMine	A mine with chickpea data (both desi and kabuli varieties) from the Legume...	A. ipaensis, A. duranensis, A. thalia...
	CowpeaMine	A mine containing both cowpea genetic and genomic data, courtesy UC-Riv...	A. duranensis, A. ipaensis, C. arietin...
	FlyMine	An integrated database for Drosophila genomics	D. melanogaster
	GrapeMine	An integrated database for grapevine data	
	HumanMine	HumanMine integrates many types of data for Homo sapiens and Mus mus...	H. sapiens
	HymenopteraMine	An integrated data warehouse for the Hymenoptera Genome Database	A. dorsata, A. echinatior, A. florea, A...
	IndigoMine	INDIGO enables the integration of annotations for the exploration and analy...	Archae
	LegumeMine	Multi-organism mine integrates data from legume species: string bean, soy...	A. duranensis, A. ipaensis, C arietin...
	MaizeMine	An integrated data warehouse for MaizeGDB	
	MedicMine	MedicMine integrates many types of data for Medicago truncatula. You can...	A. thaliana, M. truncatula, M. trunca...
	MitoMiner	MitoMiner is an integrated web resource of mitochondrial localisation evide...	D. rerio, H. sapiens, M. musculus, R....
	ModMine	A data warehouse for the modENCODE project	
	MouseMine	MouseMine is a powerful new system for online access to mouse data fro...	M. musculus

InterMine

Data

Why use InterMine?

- Query across several data sources at once
- Information without visiting several sites
- Data formatting issues resolved
- Identifier resolution system
- Collate information about items and sets
- Common platform to many organisms and type of data
- Extensive API
- Build you own InterMine

InterMine

Cross-organism analysis

RatMine

FlyMine

WormMine

MouseMine

SGD YeastMine

HumanMine

ZebrafishMine

UNIVERSITY OF CAMBRIDGE
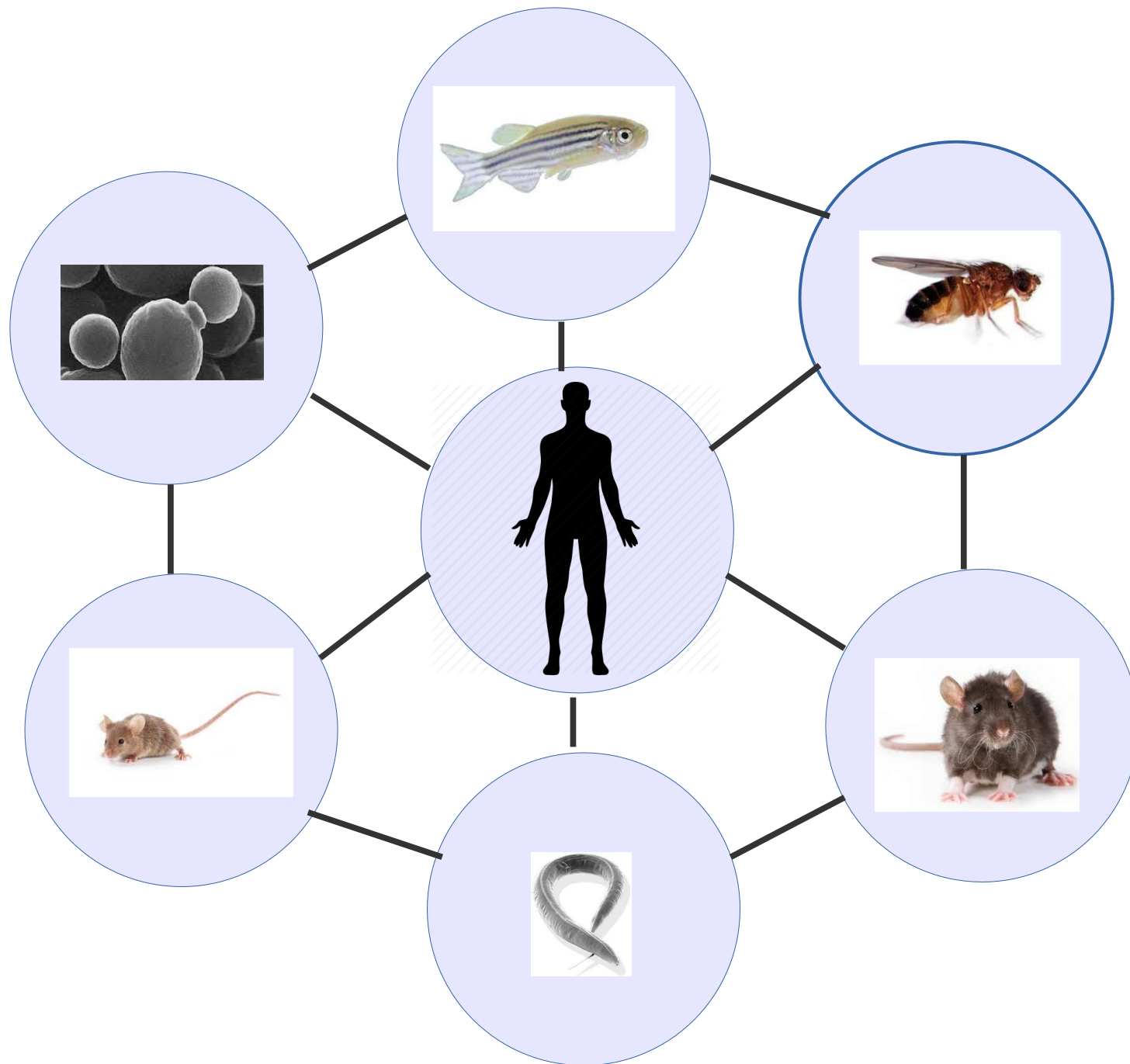
InterMine

InterMine Data Integration

Your own InterMine:

- Sophisticated build system
- Set integration keys
- Core library of data loaders
- Custom data loaders - own data

Help docs, mailing lists, etc: http://www.intermine.org/

Twitter: @intermineorg

Github: http://www.github.com/intermine

UNIVERSITY OF
CAMBRIDGE

InterMine

The Future: Widening Use
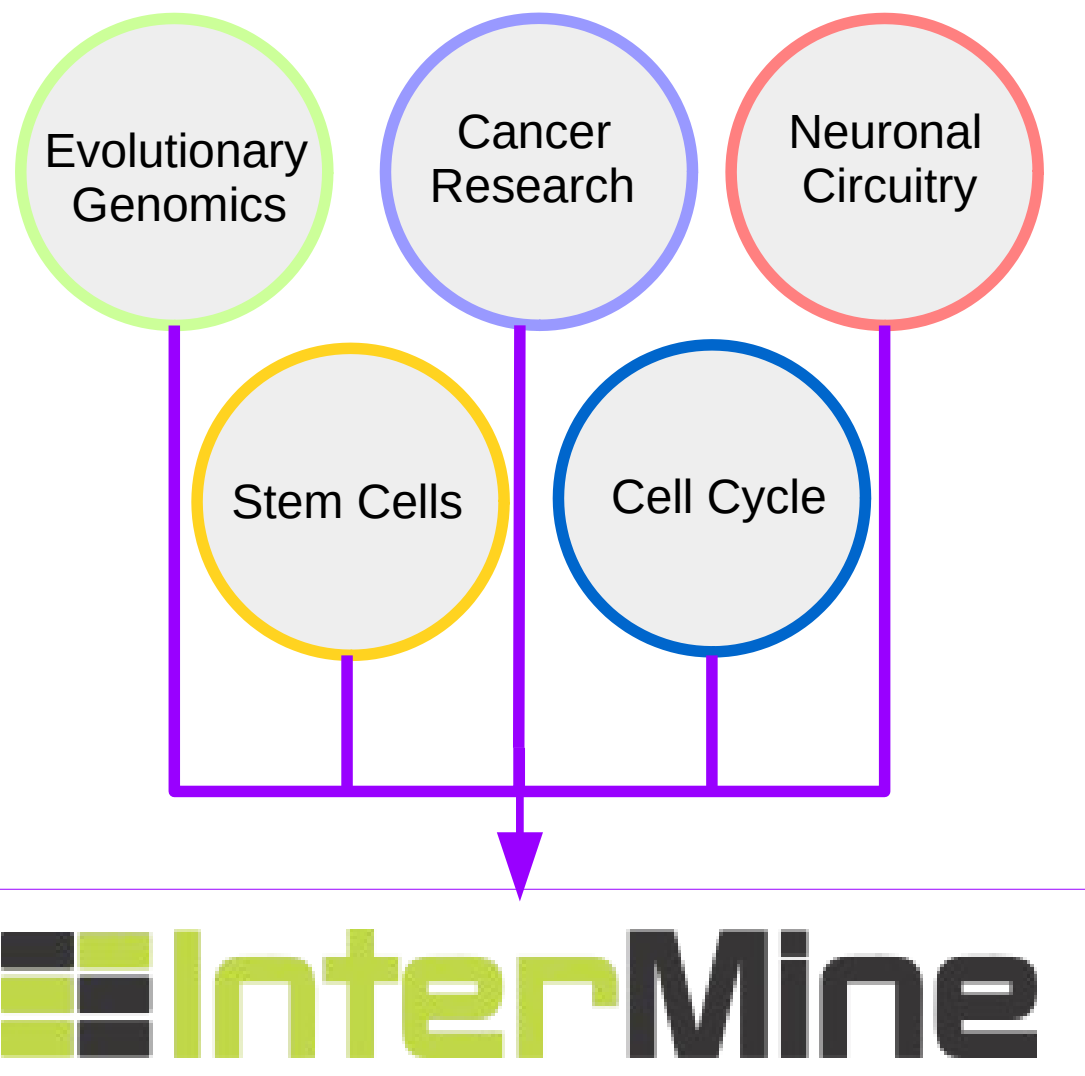
- "One click install"
- Using the cloud / docker

- Integrated with additional data sources
- New analysis and visualisation tools
- FAIR

InterMine Accounts

InterMine is free to use without creating an account.

Creating an account allows you to save lists and searches permanently and share lists with your colleagues.

At the moment you have to make a separate account for each InterMine database

UNIVERSITY OF CAMBRIDGE

InterMine

Using InterMine

Search

Explore

Analyse

UNIVERSITY OF
CAMBRIDGE

InterMine

The Web Interface

Inputs

Single Item

List

Search Interfaces

Keyword Search

Template Searches

Query Builder

Region Search

Outputs

Report Pages

List Analysis Pages

Results Tables

UNIVERSITY OF CAMBRIDGE

InterMine

Keyword Search

Inputs

Single Item

List

Search Interfaces

Keyword Search

Template Searches

Query Builder

Region Search

Outputs

Report Pages

List Analysis Pages

Results Tables

UNIVERSITY OF CAMBRIDGE

InterMine

Search and Explore

Faceted Keyword Search:

-Specific e.g. pax6, pparg
→ Report pages
-Exploratory e.g. Insulin, Diabetes
→ Report pages
→ Lists
-Filter results by data type

InterMine

Data Exploration: report pages
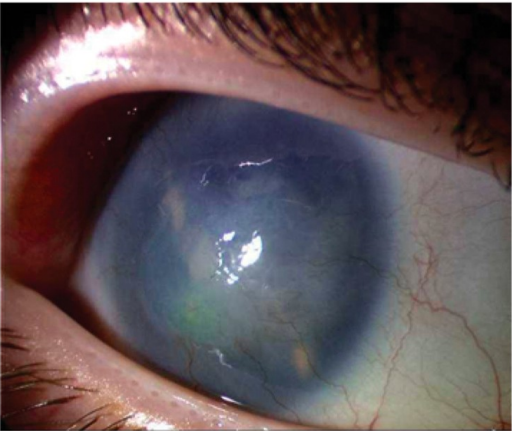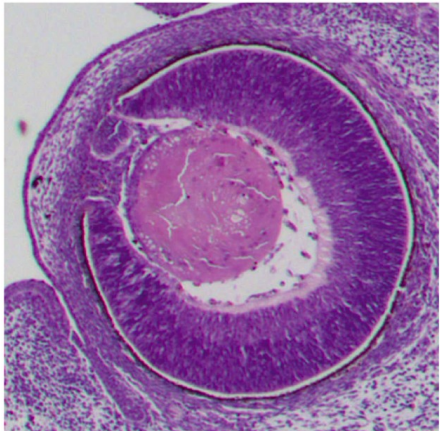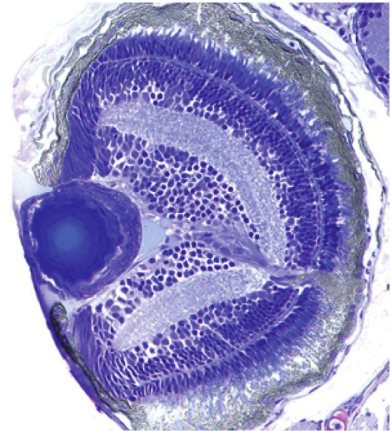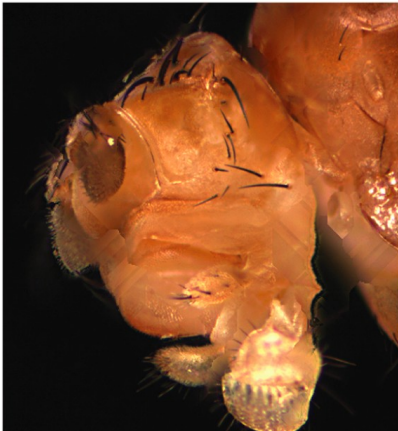
Every object (item) in InterMine has a report page

- Collate all the data available for that object

- Contain a mixture of interactive tables, search results and graphical displays.

- Links to related data

UNIVERSITY OF CAMBRIDGE

InterMine

Explore: Pax6

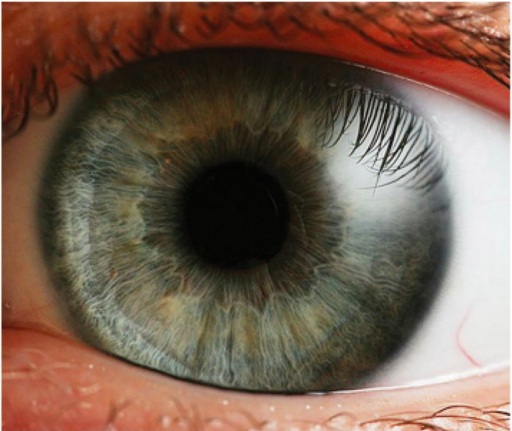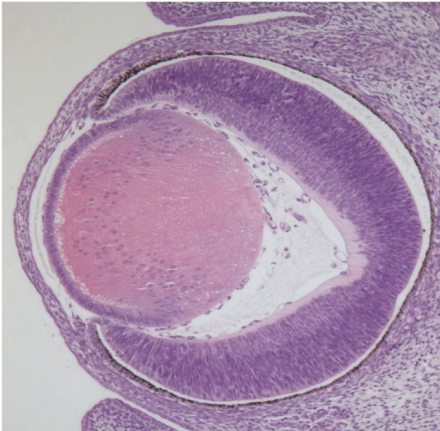| Human	Mouse	Zebrafish	*Drosophila*

WT

mut

PAX6⁺ᐟ⁻	*Pax6*⁻ᐟ⁻	*pax6b*⁻ᐟ⁻	*ey*⁻ᐟ⁻

EQs

cornea opaque
iris absent
retina degenerate
lens opaque
aqueous humor of eyeball
  increased pressure
	eye decreased size
	lens fused_to cornea
	iris morphology
	anterior chamber
	  absent
		eye decreased size
		lens decreased size
		retina malformed
			eye absent

UNIVERSITY OF CAMBRIDGE

InterMine

Exercise 1: Faceted Search

1. Search for one or more of the following in HumanMine:

- *Pax6*
- rs10509540
- *diabetes*

2. Filter and create a list:

- Search for *diabetes*
- Filter for publications
- Make a list of these publications

Note: If you filter for genes note that this is not a comprehensive search for diabetes genes as it does not check functional annotations.

Exercise2: Exploring a Gene

You are interested in the Human PAX6 gene and want to know the following things about it:

1. On which chromosome is *PAX6* located?

2. Can I access the sequence for the *PAX6* gene?

3. With which diseases is *PAX6* associated?

4. In which tissues is *PAX6* most highly expressed?

5. Does the *PAX6* protein have any know isoforms?
6. Does the Pax6 protein have known domains?

7. Is there a *PAX6* orthologue in *D. melanogaster*?

8. Does this orthologue interact with any other Genes/proteins? Identify the interaction type (genetic/physical)

9. For the interaction of Ey with 4E-T, what was the original experiment and publication that determined this interaction
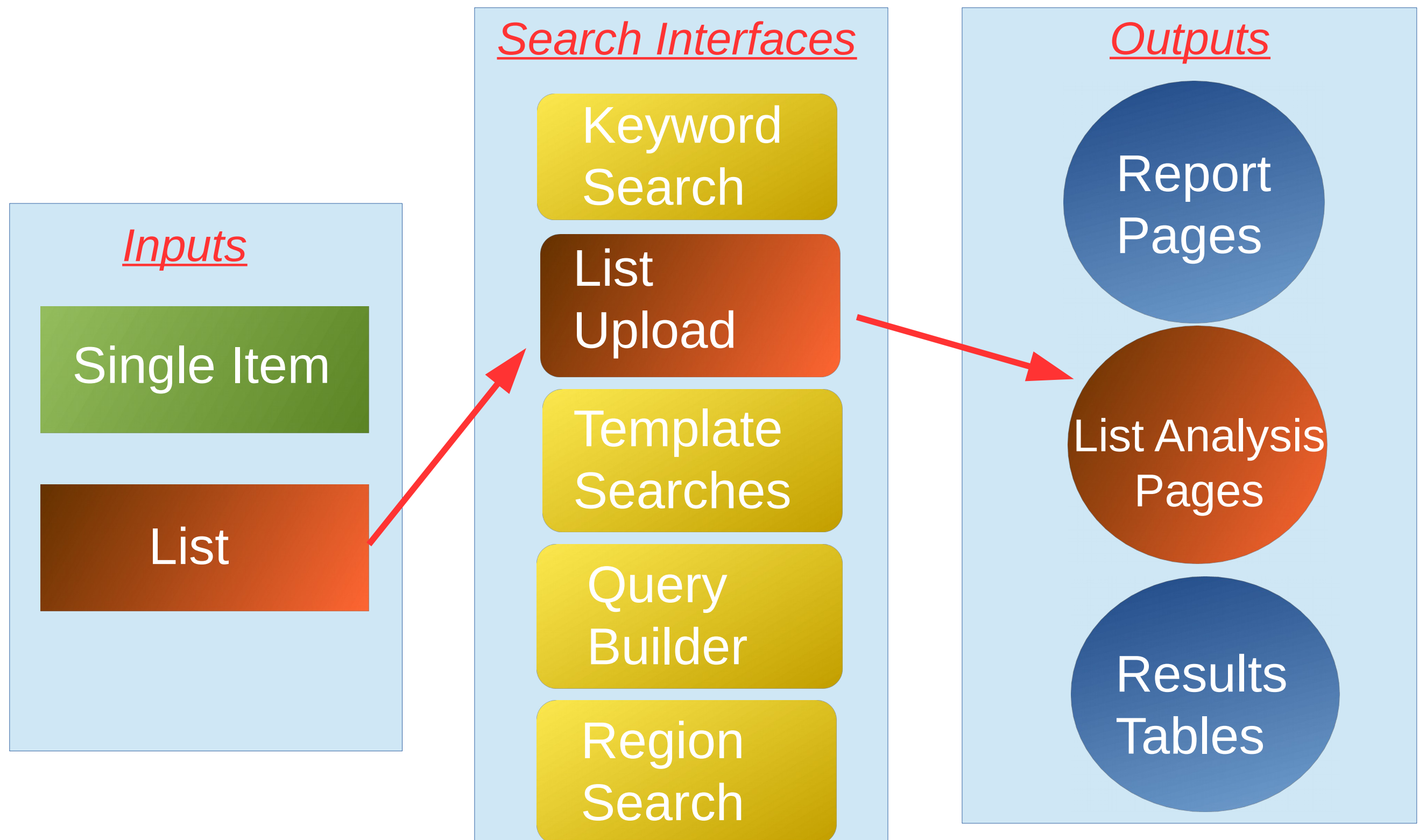
InterMine

Lists

List Analysis: Uploading a list

- Upload your own lists to InterMine

- Powerful identifier resolution system

- Convert old identifiers into an up-to-date set

UNIVERSITY OF CAMBRIDGE

InterMine

Exercise 3: Uploading a list:

1. Use FlyMine: navigate to the lists tab and list upload sub-tab

2. Select the example list (leave type and organism as the default values).

3. Click "Create list".

4. Examine and understand the list page, name and save your list.

UNIVERSITY OF
CAMBRIDGE

Inter**Mine**

Exercise 3: Uploading a list:

- E2f has matched two genes (**duplicates**) - in this case you need to decide which of the two genes you want in your list (or both). The action column allows you to do this.

- Two of the identifiers in the list matched the same gene: FBgn0010433 and ato. This is indicated in the **direct hits**.

- One of the identifiers is a protein identifier (TWIST_DROME). As the associated gene could be identified, this has been added to the list. This is shown under **non-gene identifiers**.

- Two of the identifiers matched a **synonym** (rather than a current identifier). As the synonyms matched only one gene, these are automatically added to the list.

InterMine

Data Exploration: Lists

InterMine allows you to explore data for a whole list of objects

- Uploaded or created from searches
- Identifier resolution
- List analysis pages
- Summary tables and graphs, enrichment statistics and search results
- List set operations - union, intersect, subtract
- Workflows through iterative querying and set operations
- Public lists also available

UNIVERSITY OF CAMBRIDGE

InterMine

List Analysis

Exercise 4: List Analysis:

Examine the HumanMine public list: **PL_Pax6_Targets (319 genes)**

1. What is the most enriched GO term for this list?

2. How many genes in the list are annotated with this GO term?

Note: you could make a sub-list containing only genes from this list annotated with this term by clicking on the matches number

3. Navigate to the MouseMine database to examine the mouse orthologues for this list.

4. How many mouse orthologues are there for this list?

5. Are these mouse genes enriched for any phenotypes?

UNIVERSITY OF CAMBRIDGE

InterMine

Using InterMine

Search

Explore

Analyse

Template Searches

Inputs

Single Item

List

Search Interfaces

Keyword Search

Template Searches

Query Builder

Region Search

Outputs

Report Pages

List Analysis Pages

Results Tables

UNIVERSITY OF CAMBRIDGE

InterMine
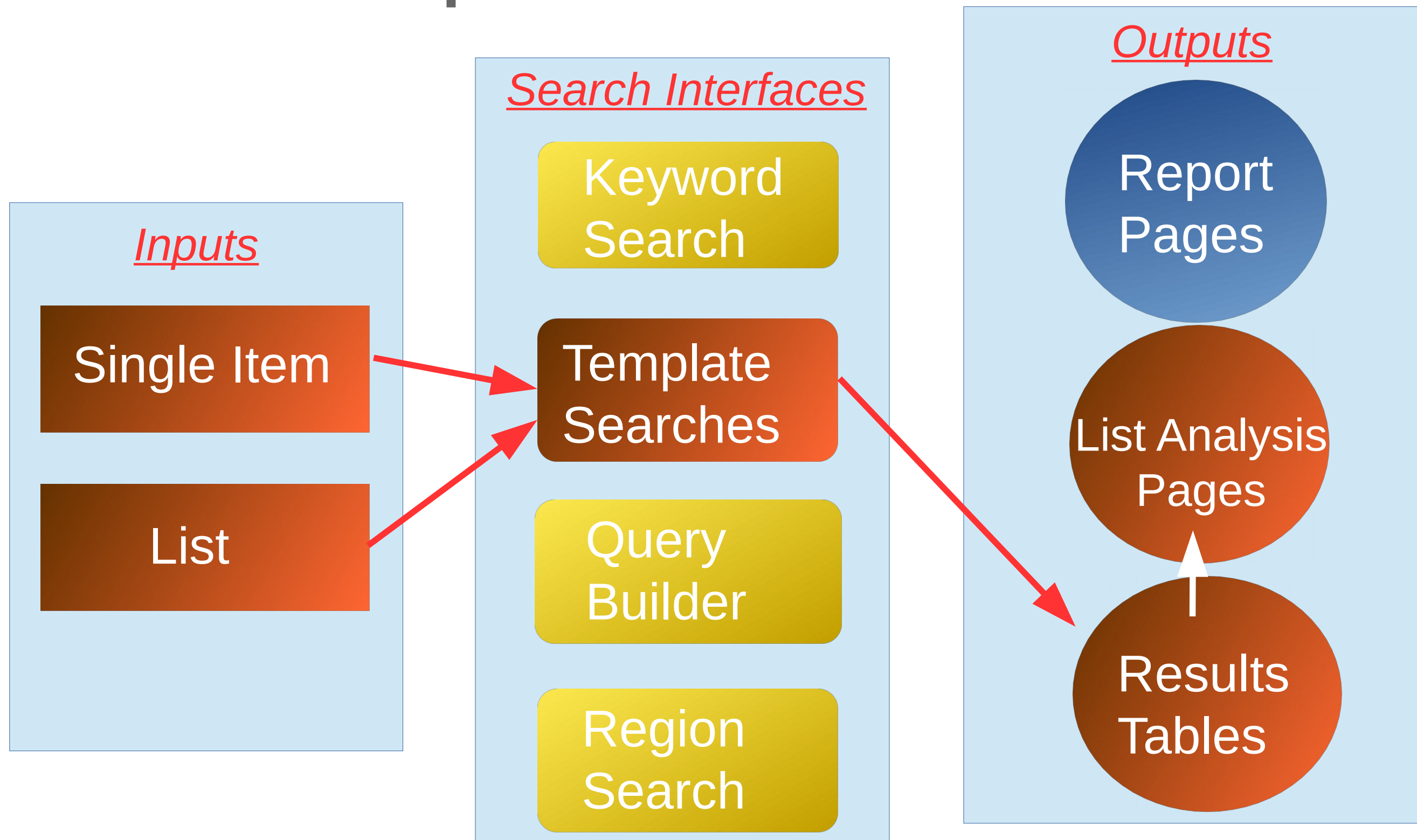
Data Analysis: Template Searches

The template searches allow a more refined search than the keyword search and report pages but are still quick and easy to access.

- Pre-defined searches with simple filters
- Range from simple searches to more complex searches spanning several data types
- Run with single item or list
- Results are returned in sophisticated results tables
- Easy to add - just ask

Many Many Searches.....

- Which other genes have this GO annotation?
- Are there mutant phenotypes for this gene?
- Where is this gene expressed?
- What does this gene interact with?
- Do any of the interacting genes share the mutant phenotypes?
- Does this gene have a human orthologue with a disease association?
- Have any variants been associated with this gene/disease?
- Which organisms have models for this disease/gene?

InterMine

Data Analysis: Template Searches

- What other genes are involved in pancreatic function?

- Are there potential targets of pax6 in pancreatic tissue?

- Have these genes been implicated in pancreatic disease?

- What published data is there about these genes?

EEE InterMine

Data Analysis: Results Tables

Results tables allow further interactive analysis of the data through:

- Column summaries
- Column sorting
- Adding additional columns of data
- Filtering
- List creation
- Export

UNIVERSITY OF CAMBRIDGE

InterMine

The power of column summaries

- Find the number of unique items in a column
- Find the number of items with a particular property
- Filter the table for one or more specific properties

UNIVERSITY OF CAMBRIDGE

InterMine

Exercise 5: Template searches:

1. Browse the template searches in FlyMine and HumanMine - try running a few or changing the filters.

2. Use the search box to find template searches for interactions

3. Filter the FlyMine template searches to show only "expression" templates.

Exercise 6: Using template searches:

We will continue our exploration of the Pax6 gene in Pancreatic tissue. Use the template searches in HumanMine to answer the following question:

Are any of the known targets of Pax6 expressed in the pancreas. You will find a public list of known Pax6 target genes in HumanMine PL_Pax6_Targets).

Use the Protein Atlas dataset for the expression measurement.

Save the list of target genes expressed in the pancreas as a list.

The Query Builder

Inputs

Single Item

List

Search Interfaces

Keyword Search

Template Searches

Query Builder

Region Search

Outputs

Report Pages

List Analysis Pages

Results Tables

UNIVERSITY OF CAMBRIDGE
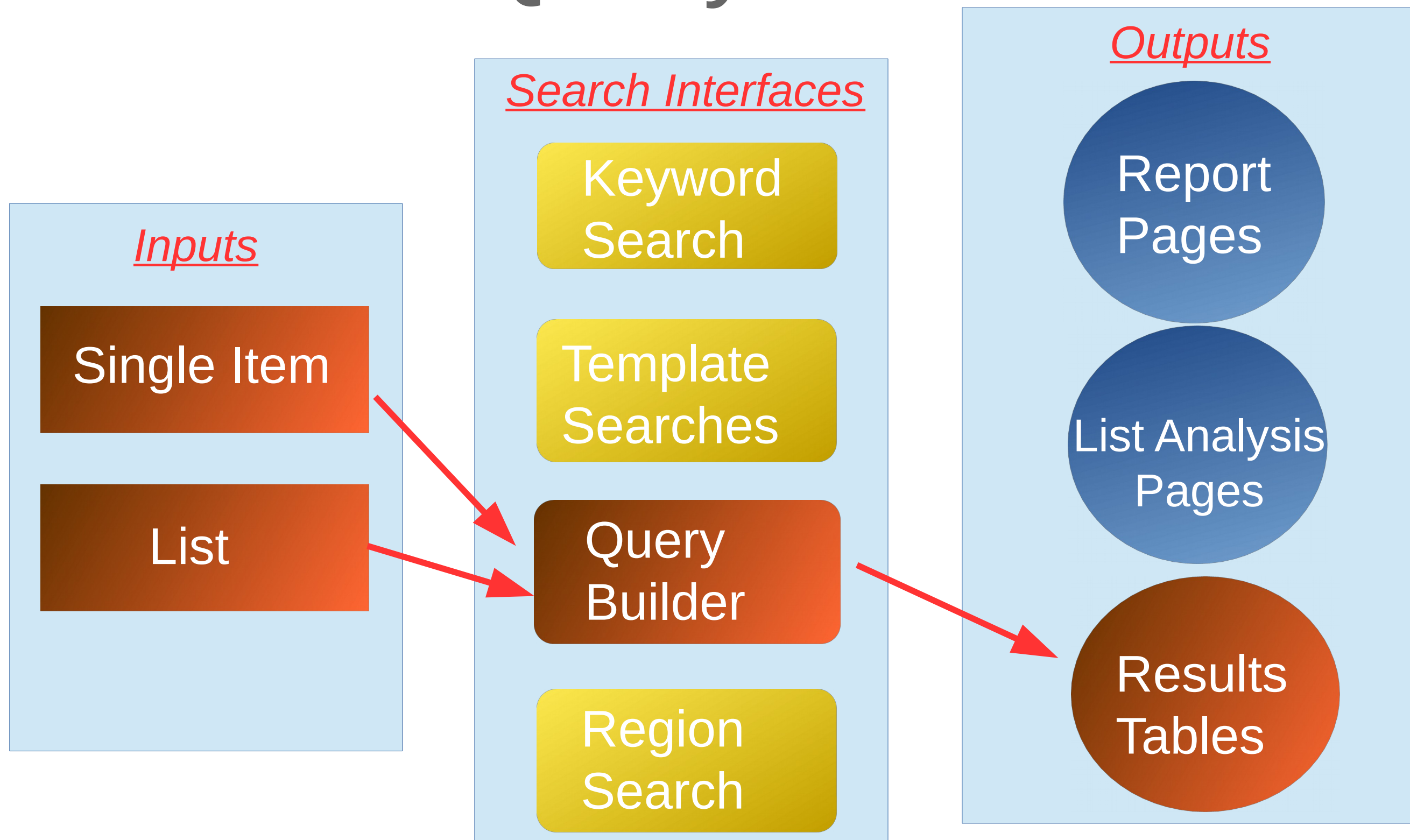
InterMine

Advanced Search: Query Builder

The Query Builder is InterMine's custom query builder, allowing you to create and save your own searches.

- Build your own Searches

- Modify template searches

- Combine any data:

  - And, Or,
  - Intersect; Union

Data Analysis: Query Builder

Three steps to construct a query:

1. Navigate the data model to find the class or attribute you need

2. Add the appropriate constraint (filter) to the class/attribute

3. Decide on the columns you want to view in your results

UNIVERSITY OF CAMBRIDGE

InterMine

Exercise 7: Query Builder:

Using HumanMine: we will build a query to show Human genes and OMIM diseases, and then add a further constraint to show genes associated with all types of Diabetes.

1. Start your query from Gene
2. Constrain "Organism" to Homo Sapiens
3. Add the columns of data we want in our results:

   Gene: Primary identifier and Symbol
   Disease: name

4. Run this search - 'Show results'.
5. Return to the query (Use the "Trail" in the top left) and add a constraint to Disease name for "CONTAINS *Diabetes*"
6. Run the search and save the set of genes

The Web Interface

Inputs

Single Item

List

Search Interfaces

Keyword Search

Template Searches

Query Builder

Region Search

Outputs

Report Pages

List Analysis Pages

Results Tables

UNIVERSITY OF CAMBRIDGE

InterMine
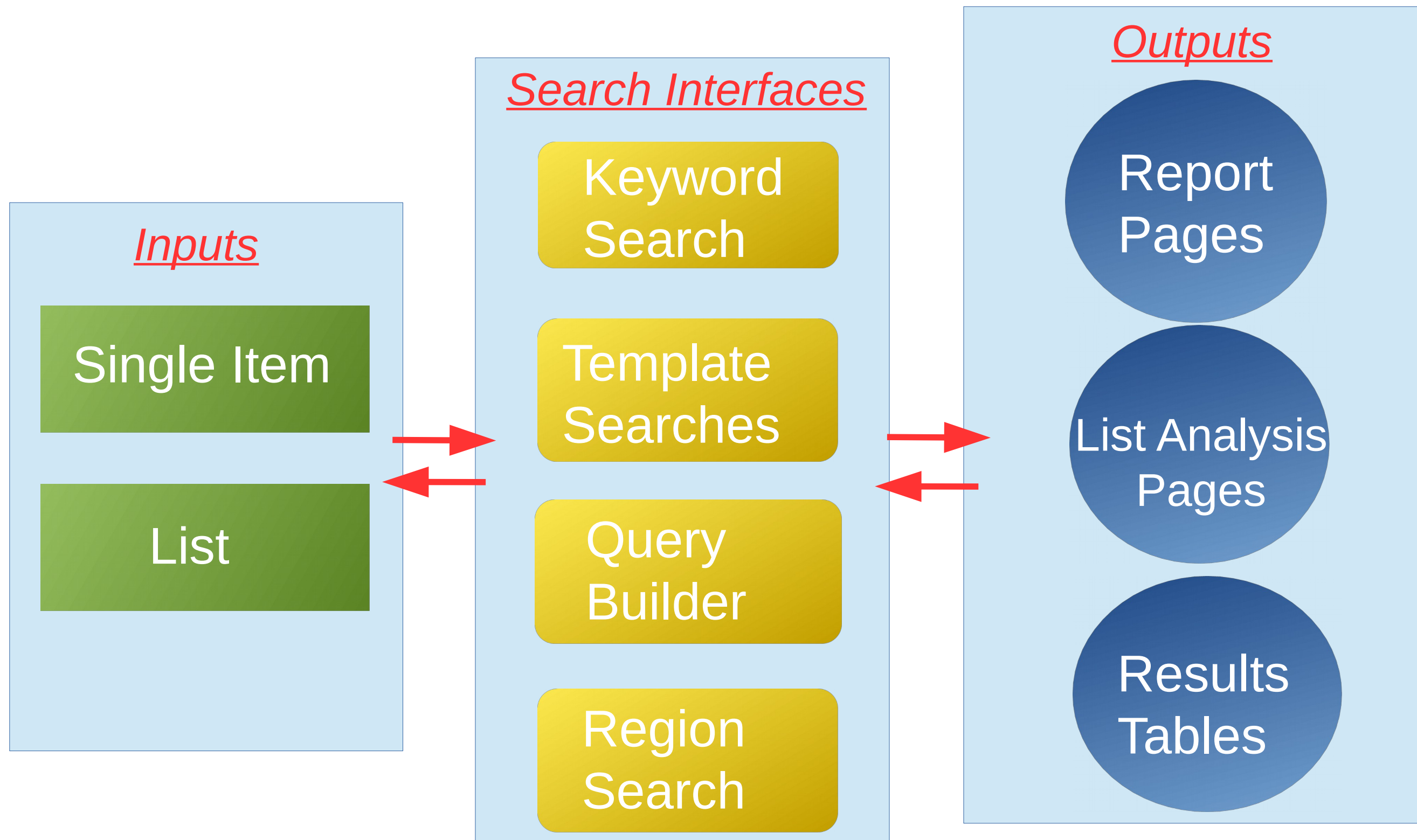
Lists: Set Analysis
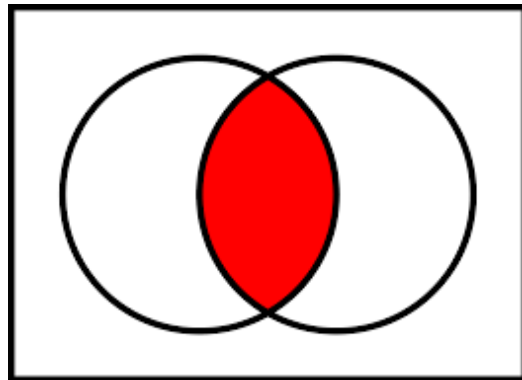
Intersect

Union
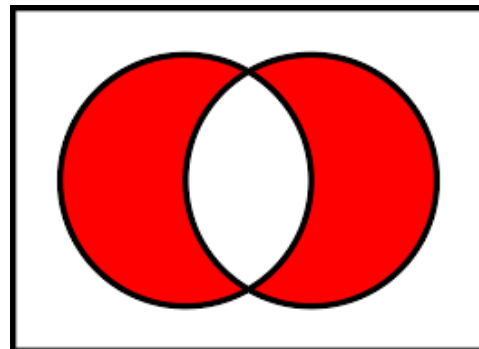
Subtraction

Asymmetric difference

UNIVERSITY OF CAMBRIDGE
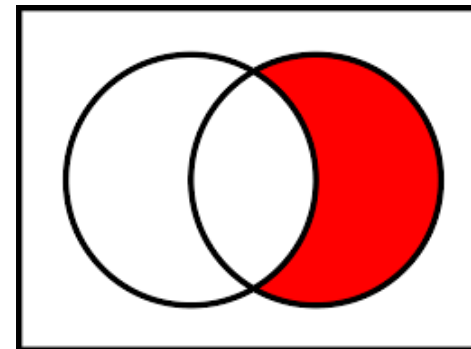
InterMine

Exercise 8: Analysis Workflows

In exercise 6 we ran a template search and saved a set of genes. In this exercise we will combine this set of genes from exercise 6 with our set of diabetes genes from exercise 7. We will then feed the resulting set into a further query.

1. Identify the sets of genes you have created under the lists "view" tab.

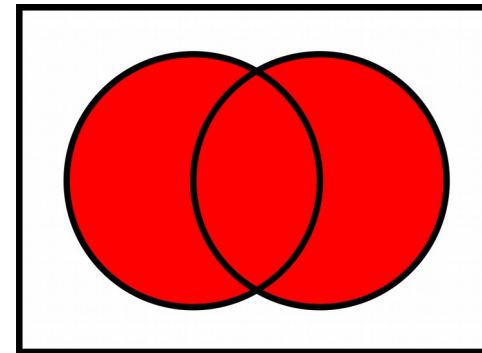2. Use the list set operations available on this page to intersect the list of diabetes genes you created with the query builder with your previous set of genes (Pax6 target genes expressed in the Pancreas) created in exercise 6.

3. We now want to know if any of these genes have been identified in GWAS studies. Run a template on your intersected list to find this out.

4. Use the column summary to find if any of the GWAS phenotypes are related to diabetes.

UNIVERSITY OF CAMBRIDGE

InterMine

Analysis Workflows
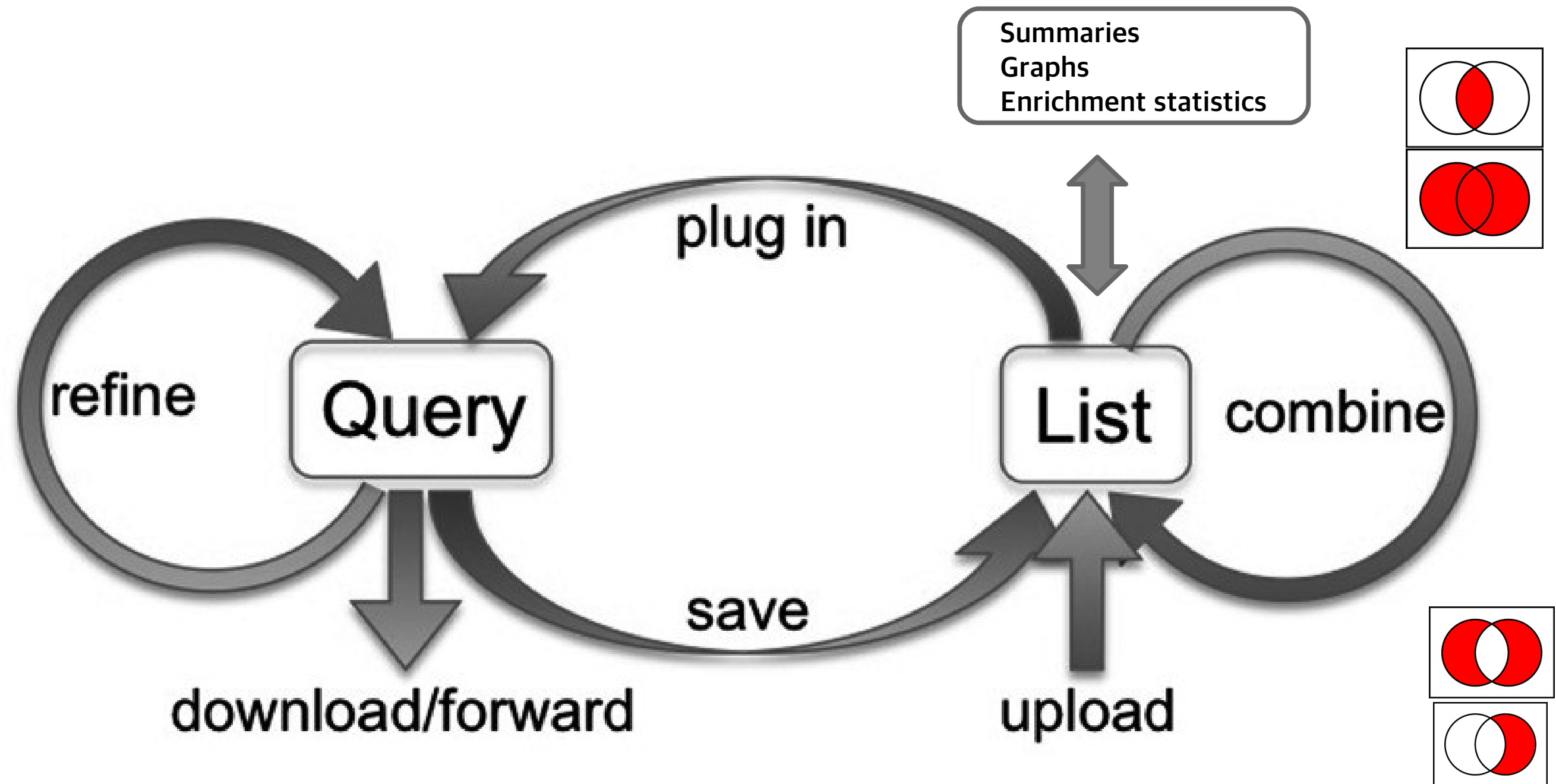
Motenko H, Neuhauser SB, O'Keefe M, Richardson JE. MouseMine: a new data warehouse for MGI. Mamm Genome. 2015 Aug;26(7-8):325-30. doi: 10.1007/s00335-015-9573-z. PubMed PMID: 26092688; PubMed Central PMCID: PMC4534495

Analysis Workflow

1. We looked at gene expression of Pax6 targets in the Pancreas and refined this to genes with a high or medium level expression. We saved these as a list – Set A.

refine

Gene → Protein Tissue Localisation
(Protein Atlas data)

SetA

UP regulated in pancreatic tissue

save
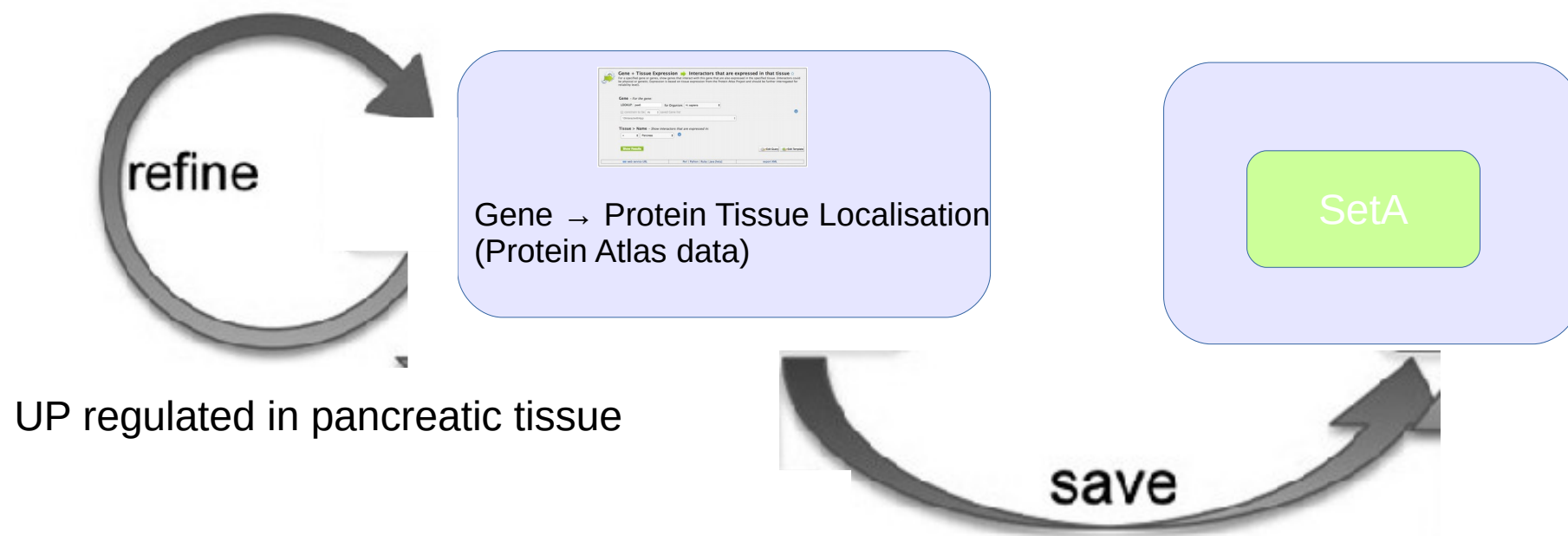
Analysis Workflow

2. A query built through the query builder identified a set of genes involved in Diabetes -> setB

Genes WITH diabetes
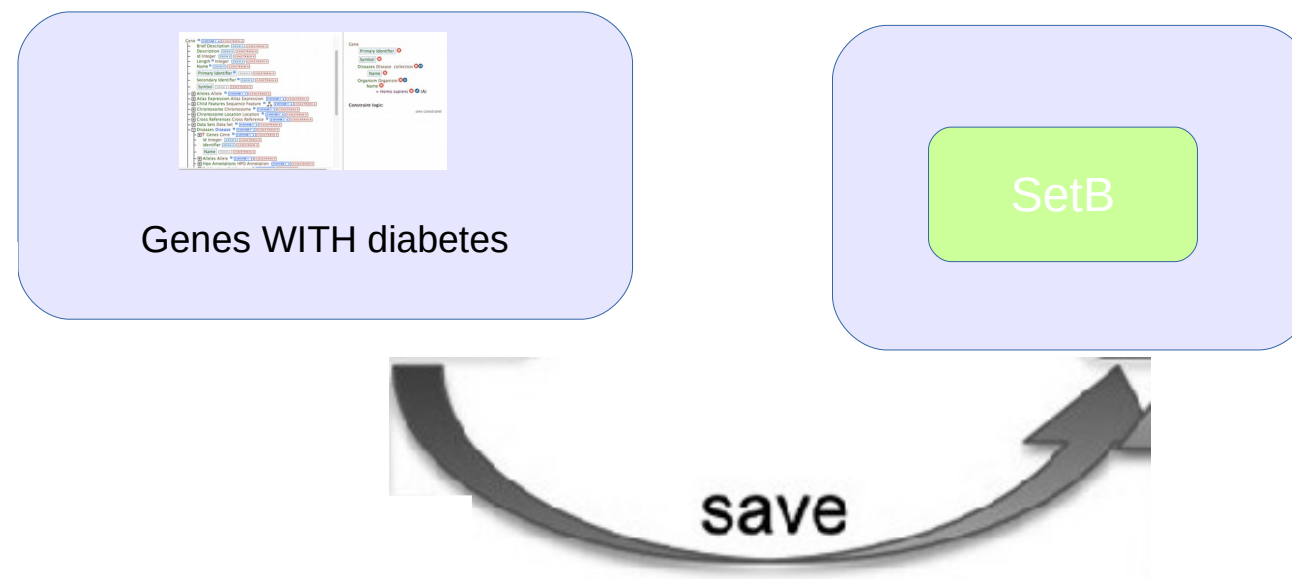
SetB

save

Analysis Workflow

3. A list intersection between setA and setB identified Pax6 target genes that are expressed in the pancreas that also have some association with diabetes → setC

Lists:

SetC

SetA

SetB

A∩B

combine
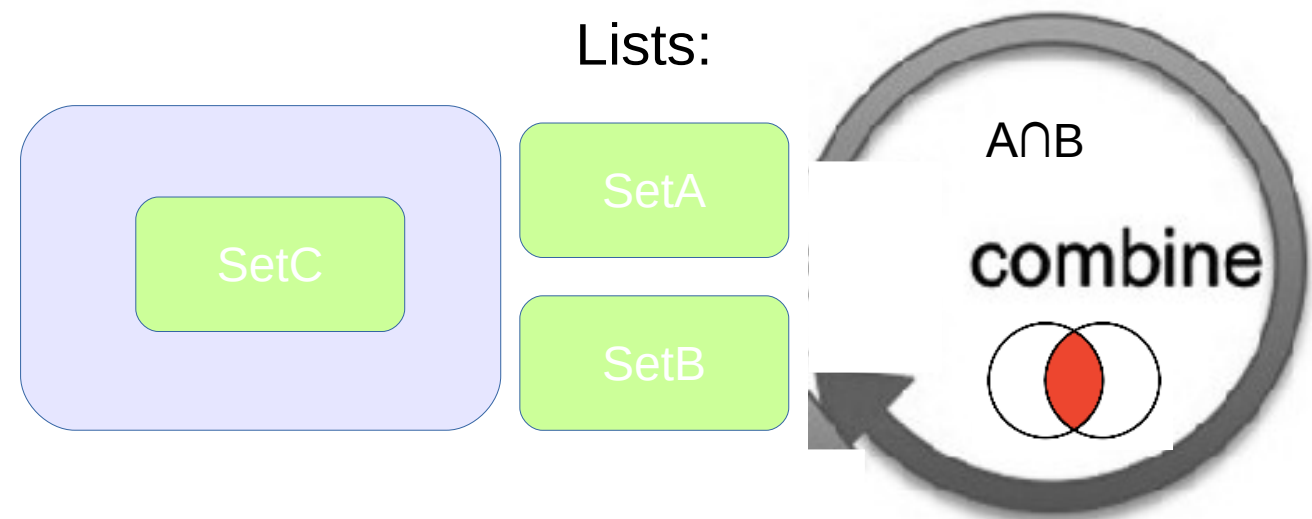
Analysis Workflow

4. Genes saved as SetC were "plugged in" to a second template to identify if there was any association of these genes with diabetes phenotypes according to GWAS studies.
These genes were saved - SetD

Gene -->GWAS hit

SetC

SetD

GWAS studies looking at diabetes/
Phenotypes associated with diabetes

Genes with a diabetes phenotype
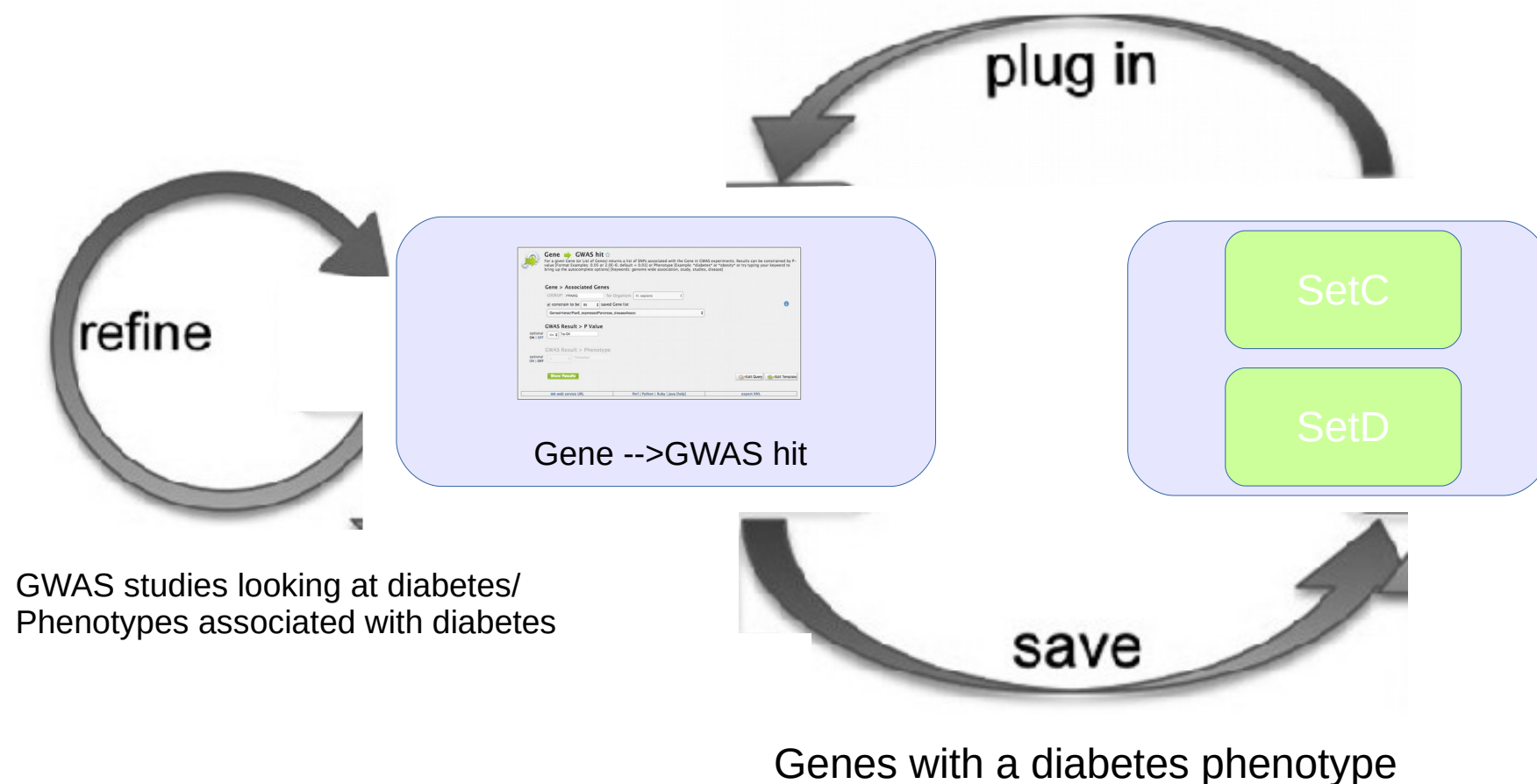
Analysis Workflow

5. ListD contained two genes that we could explore further through their report pages, the list analysis page and through further template searches

Summaries
Graphs
Enrichment Statistics

plug in

refine

Lists:

B∩C

combine

SetC

SetA

SetD

SetB

Gene -->GWAS hit

GWAS studies looking at diabetes/
Phenotypes associated with diabetes

save

Report Page

The Web Interface

Inputs

Single Item

List

Search Interfaces

Keyword Search

Template Searches

Query Builder

Region Search

Outputs

Report Pages

List Analysis Pages

Results Tables

UNIVERSITY OF CAMBRIDGE

InterMine

Data Analysis: RegionSearch

The Region Search allows you to search for features that overlap a list of genome coordinates.

- Any or selected genome features can be searched.
- Accepts base or interbase coordinates
- Region to be searched can be extended upstream and downstream

UNIVERSITY OF CAMBRIDGE

InterMine

Exercise 9: Region Search:

Using **FlyMine:**

1. Select the example set of regions
2. De-select the features and re-select Genes and Regulatory regions

3. Extend the search by 5kb
4. Run the search

Examine the results and:

5. Create a list of all genes found.
6. Create a list of the regulatory regions found in the first genomic span.

Use-Cases

A set of use cases using HumanMine and FlyMine:

https://docs.google.com/document/d/1hkbWNJ_B5i2YTvmuXiRGIz-m99enmgbS2DgOjaGCrUI/edit?usp=sharing

 Hints and answers for the use cases:

https://docs.google.com/document/d/1Dv5kCQRN3YqNj4JYUl1CIlQFN7VGQIYjbEUDzJptavM/edit?usp=sharing

HELP?

- FlyMine - extensive manual and videos under the 'help' link. These apply to all InterMine databases

- Each InterMine has it's own help pages, videos and tutorials

Email Us: Every InterMine page has a 'questions/comments' link.

Support chat: http://chat.intermine.org

Support email: info@intermine.org

Twitter: @intermineorg

UNIVERSITY OF CAMBRIDGE

InterMine

Questions

UNIVERSITY OF CAMBRIDGE

InterMine

The InterMine Team

PI ➝ Gos Micklem

Software Developers ➝ Daniela Butano
Sergio Contrino
Kevin Herald Reierskog

Biologist ➝ Rachel Lyne

wellcome

BBSRC
bioscience for the future

EPSRC
Engineering and Physical Sciences
Research Council

NIH National Human Genome
Research Institute

elixir
UNITED KINGDOM

UNIVERSITY OF CAMBRIDGE

InterMine

Coffee Break