Discovery Net

Unifying the World's Knowledge

e-Science & Text Mining

a marriage made in heaven?

Moustafa M. Ghanem
Department of Computing
Imperial College London

e-Science & Text Mining

Motivations

- Add text mining services to e-Science projects
- Use e-Science technologies to build text mining services

Approach

- Work with end user Scientists
- Architecture for distributed text mining based on Discovery Net

e-Science & Text Mining
Projects at Imperial College

- Discovery Net:
  - £2.2M EPSRC Pilot Project, ends in March 05
  - Service-based infrastructure
  - Meta data models for mixed data mining / text mining

- Real-time Text Mining:
  - £125K EPSRC Best Practice, start Oct 04
  - Collaboration with myGrid: Service interoperability
  - Automatic Annotation of Medline Documents with GO codes

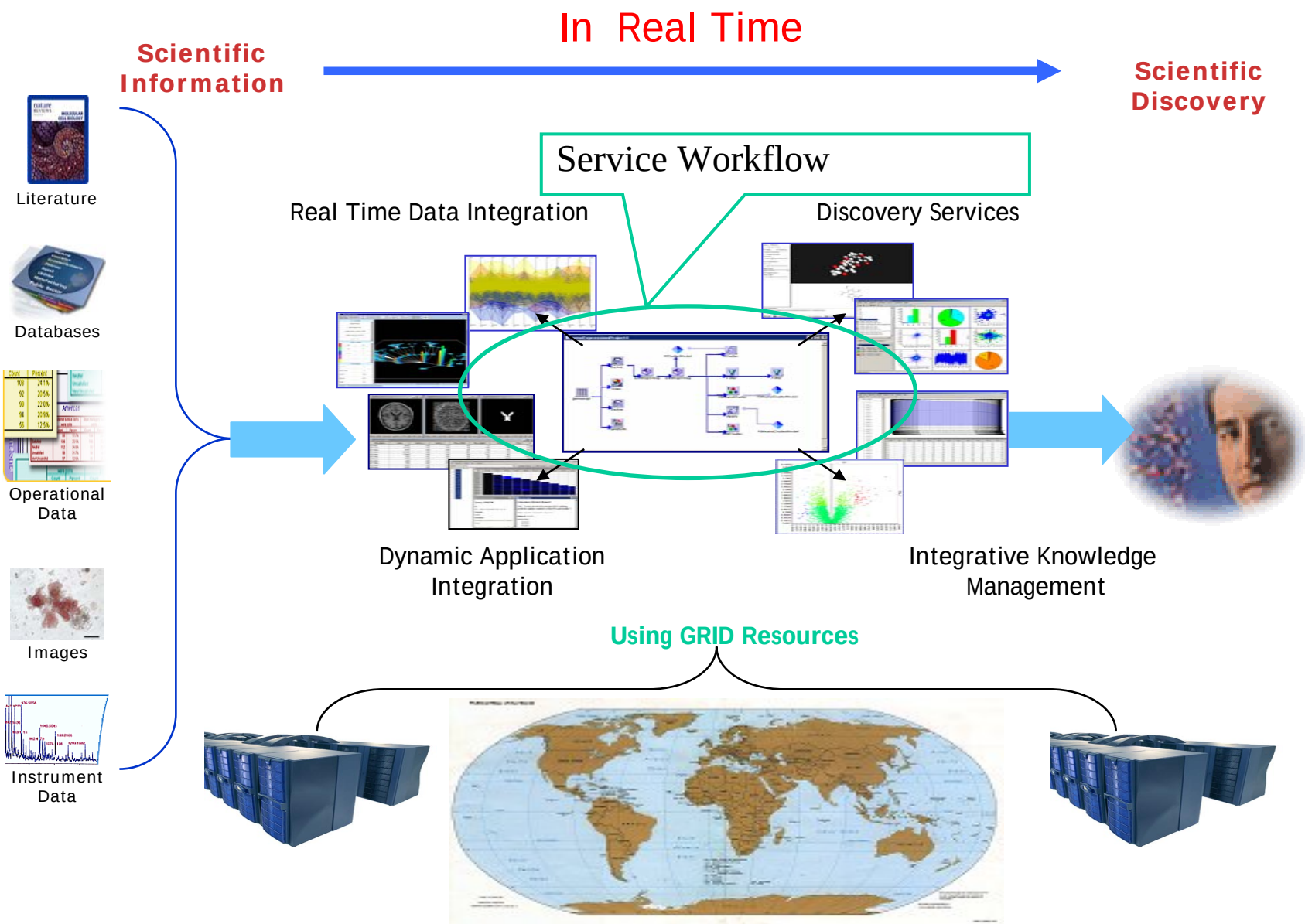
- Integrative Biology in silico: Applications of Advanced Informatics to Systems Biology
  - £550K BBSRC BEPII Project, start Oct 04
  - Using Text Mining to interpret Discovery Results
  - Insulin Signalling

Discovery Net

Unifying the World's Knowledge

Discovery Net e-science infrastructure

Infrastructure for Global Knowledge Discovery Services



Challenge:
Meeting Scientists' Requirements

Scientist's Requirements:

- I want an easy a user-friendly tool that helps solve my scientific problem. There is a wealth of knowledge in the literature and I want to tap in. I heard e-Science text mining is great.

What Computer Scientists say:

- **Information Retrieval:** Google, PubMed, ….
- **NLP:** Entity/Information Extraction, Lexicons, Anaphora Resolution, Semantic Ambiguities,
- **Data Mining:** Machine Learning, Statistics, Classification, Clustering, etc
- **Grid Computing:** Large Data-sets, Distributed Data Sources, Compute-intensive Tasks, ---- Condor, Globus, OGSA, GT3, GT4, GT2010..
- **e-Science:** Workflow, Service Computing, Ontologies, Metadata, Information Integration, Semantic Grid, …

Scientists say:

- ??? , *Can't I just have a better Google!!*

Challenge:

- Fill Gap between what scientists want and what technology provides

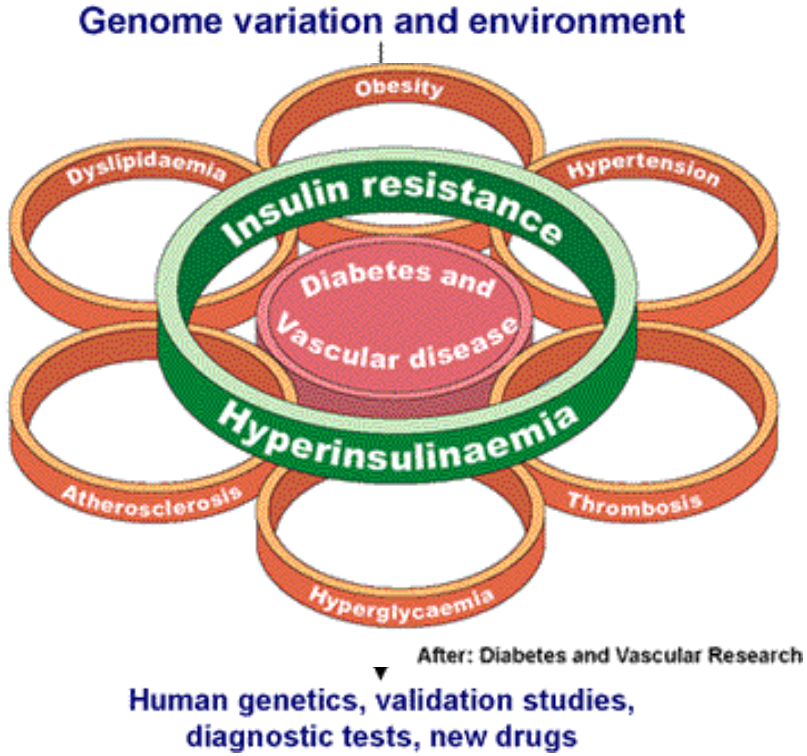
Application Context: Biological Atlas of Insulin Resistance

http://www.bair.org.uk/

Working in collaboration with scientists working on the £5.5M BAIR project funded by the Wellcome Trust.

The Biological Atlas of Insulin Resistance

The knowledge in the Atlas will lead to a new and fundamental understanding and classification of the causes of insulin resistance and the processes leading to its development. This information will be used as a platform for studies of the causes of insulin resistance in humans, and the basis for more rational and effective strategies for its prevention and treatment than are currently available.



We have adopted the term "atlas" because it emphasises that our aim is a description of the insulin-action "universe" that incorporates and integrates many different types of information, is flexible enough to cope with continuous updating and revision, and capable of serving as a discovery tool – for example, allowing insulin-resistant states of unknown aetiology to be "mapped" onto stable metabolic coordinates.

Prof James Scott, FRS.

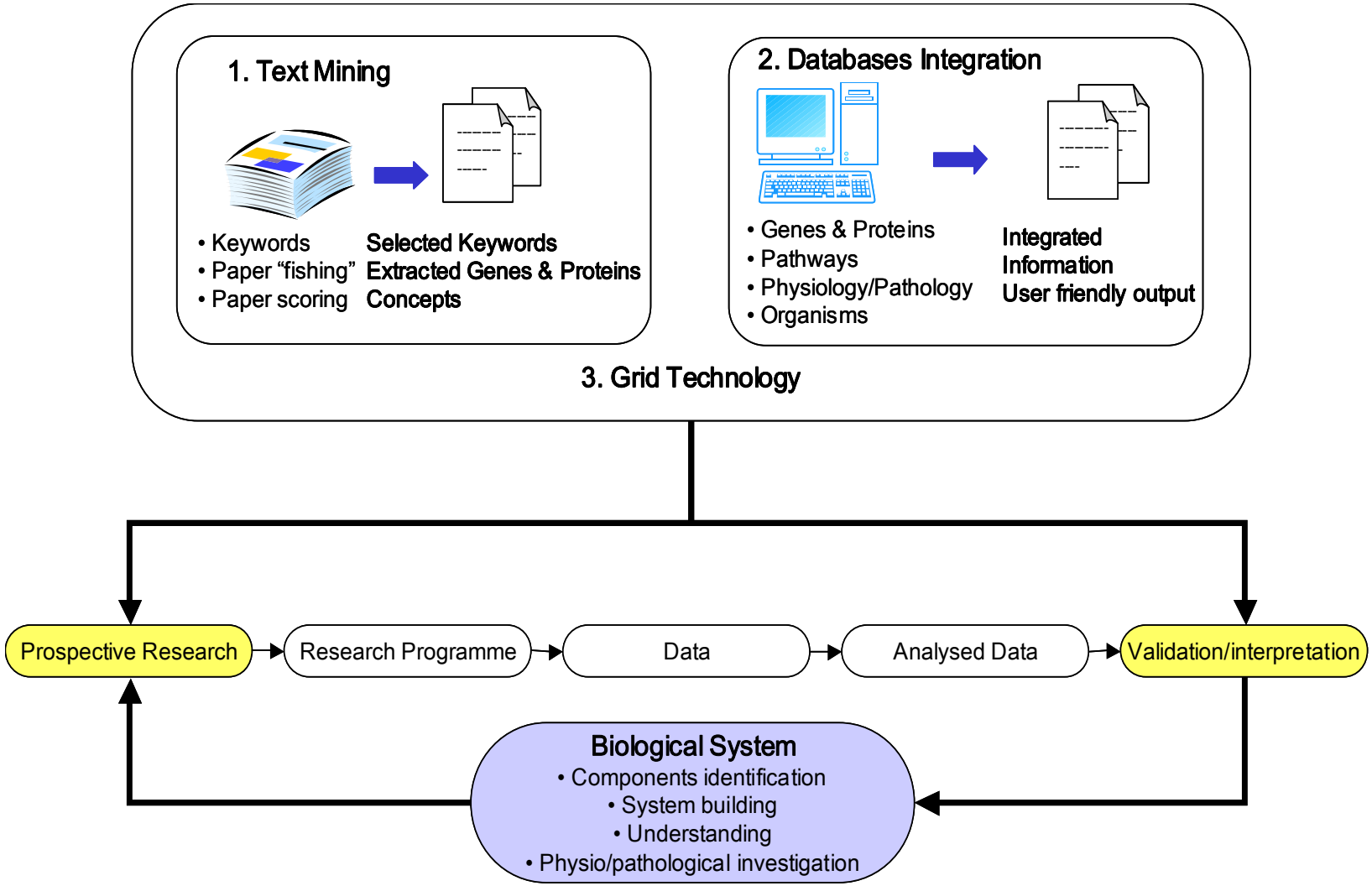
Discovery Net

Unifying the World's Knowledge

Integrative Biology Example

What Scientists actually from Text Mining?

Example from Integrative Systems Biology Studies



Summarising the Requirements

What Scientists actually want from Text Mining?

Examples from Integrative Biology Studies

Leveraging Literature in Discovery Interpretation

- Automatic Result Annotation
  - (e.g. What is the difference/similarity between microarray results)

Generating Prioritised Document Lists

- Automatic Document Categorization
- Topic Maps
- Linking literature to available background knowledge

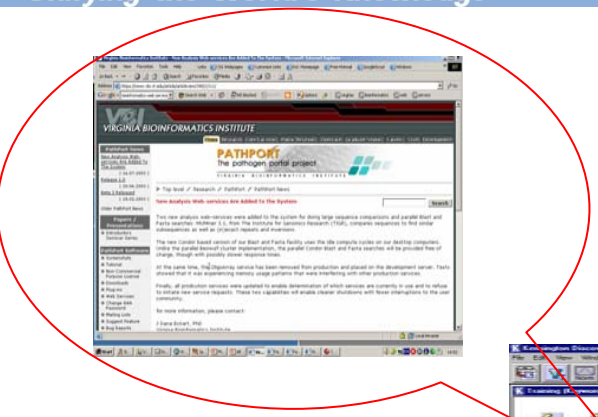
Generating New Knowledge

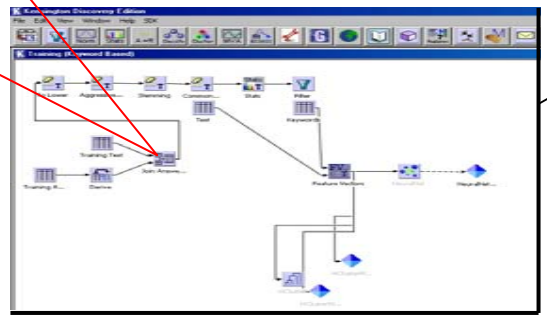
- Information Extraction/Database Curation
  - (e.g. protein-protein interactions)
- Summarising/Comparing IR Query Results
  - (e.g. compare result sets for disease in two stages)

Discovery Net

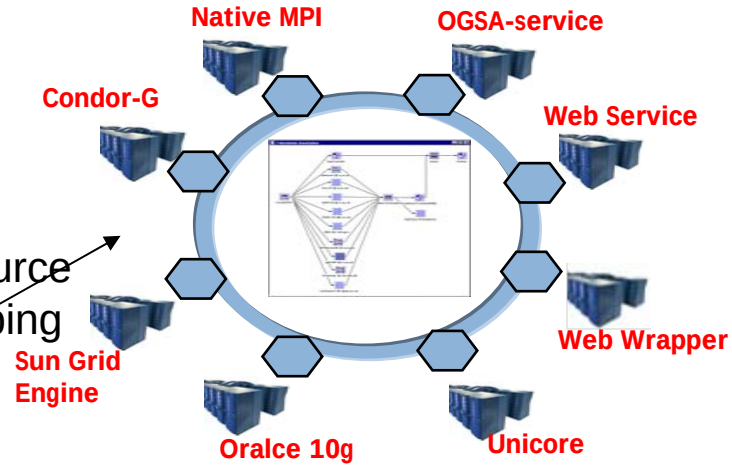
Unifying the World's Knowledge

Discovery Net Model





Resource Mapping



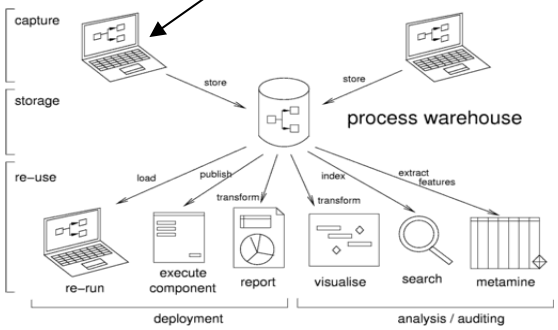
Workflow Execution

A compositional GRID

Workflow Authoring

Composing services

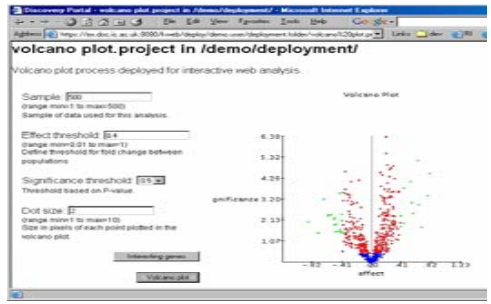
Workflow Warehousing



Workflow Management

Collaborative Knowledge Management

Service Abstraction



Workflow Deployment:

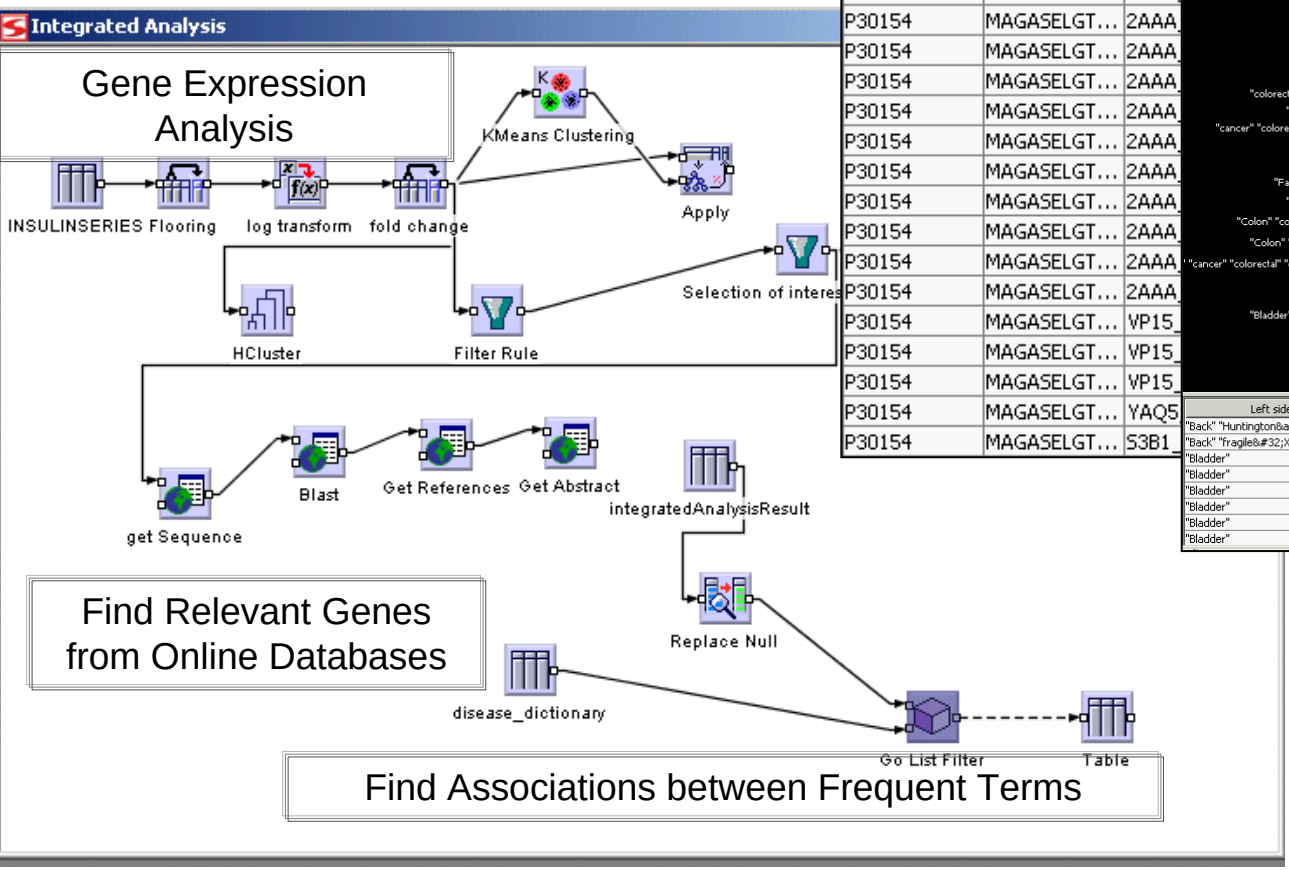
Grid Service and Portal

Leveraging Literature in Discovery Interpretation

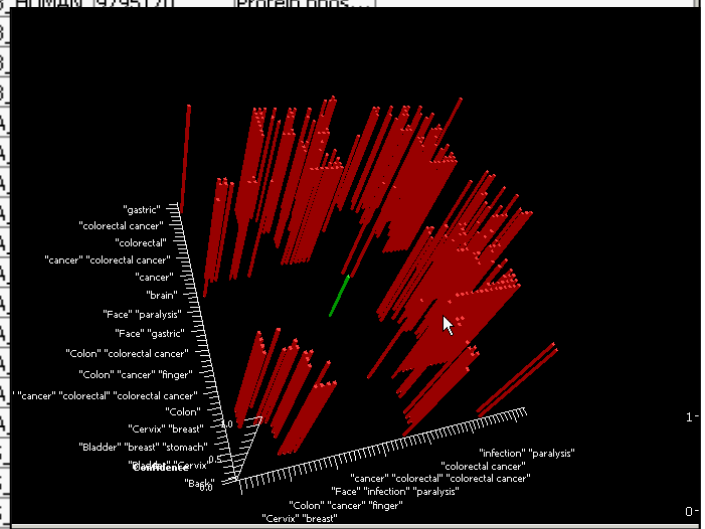
Adding text mining services to e-Science projects

Gene Disease Analysis

- Dynamic access to online data



Acc	seq	hits	pmid	abstract	Diseases
P30154	MAGASELGT...	2AAB_HUMAN	9795170	Protein phos...	
P30154	MAGASELGT...	2AAB_			
P30154	MAGASELGT...	2AAB_			
P30154	MAGASELGT...	2AAB_			
P30154	MAGASELGT...	2AAB_			
P30154	MAGASELGT...	2AAA_			
P30154	MAGASELGT...	2AAA_			
P30154	MAGASELGT...	2AAA_			
P30154	MAGASELGT...	2AAA_			
P30154	MAGASELGT...	2AAA_			
P30154	MAGASELGT...	2AAA_			
P30154	MAGASELGT...	2AAA_			
P30154	MAGASELGT...	2AAA_			
P30154	MAGASELGT...	2AAA_			
P30154	MAGASELGT...	2AAA_			
P30154	MAGASELGT...	VP15_			
P30154	MAGASELGT...	VP15_			
P30154	MAGASELGT...	VP15_			
P30154	MAGASELGT...	YAQ5_			
P30154	MAGASELGT...	S3B1_			

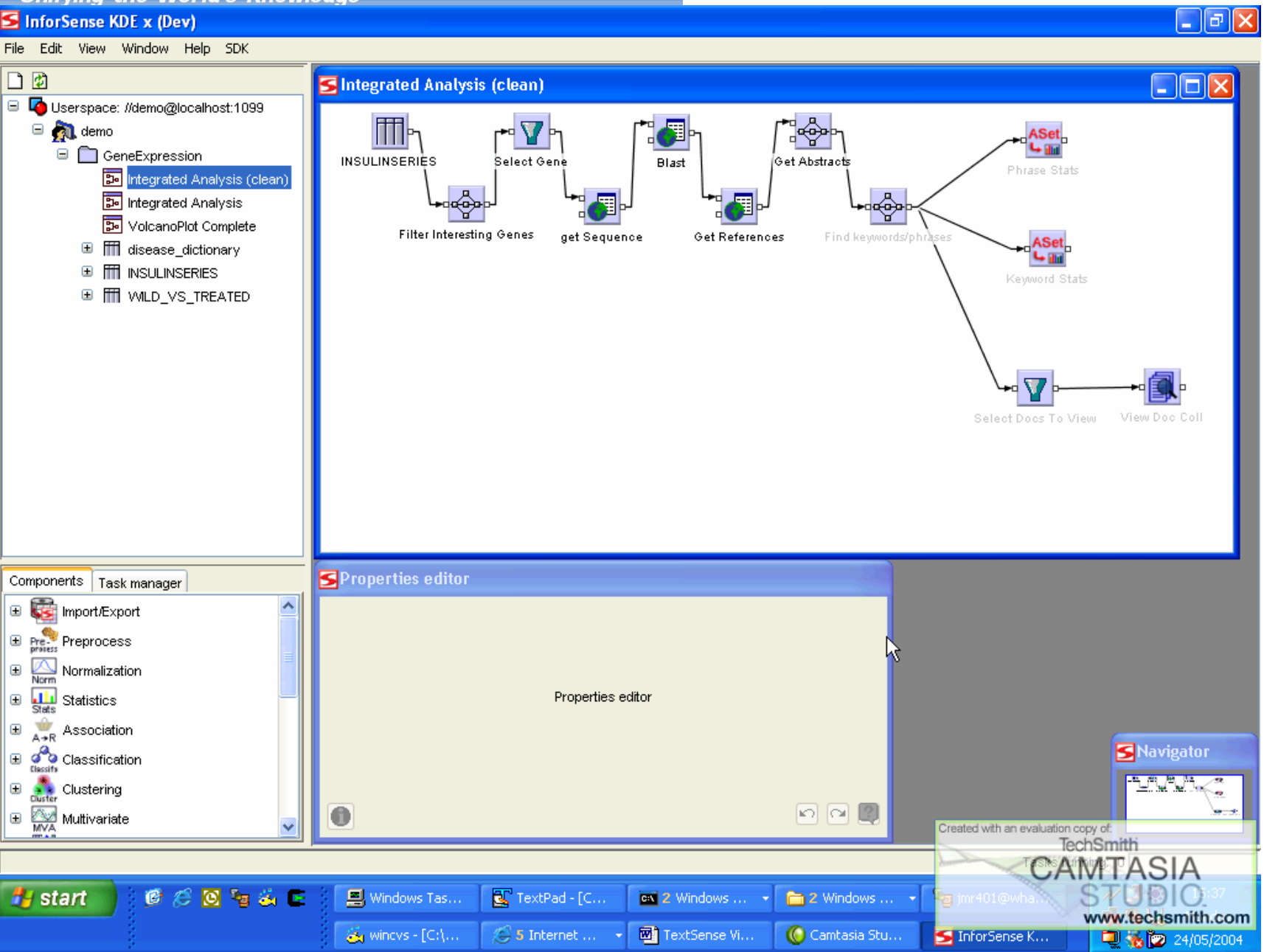


Left side	Right side	Confidence	Support
"Back" "Huntington's&#32...	"fragile X syndrome"	1.0	0.023256
"Back" "fragile X syn...	"Huntington's disease"	1.0	0.023256
"Bladder"	"Cervix"	1.0	0.023256
"Bladder"	"Cervix" "breast"	1.0	0.023256
"Bladder"	"Cervix" "breast" "stomach"	1.0	0.023256
"Bladder"	"Cervix" "stomach"	1.0	0.023256
"Bladder"	"breast"	1.0	0.023256
"Bladder"	"breast" "stomach"	1.0	0.023256

Discovery Net

Unifying the World's Knowledge

Short Animation

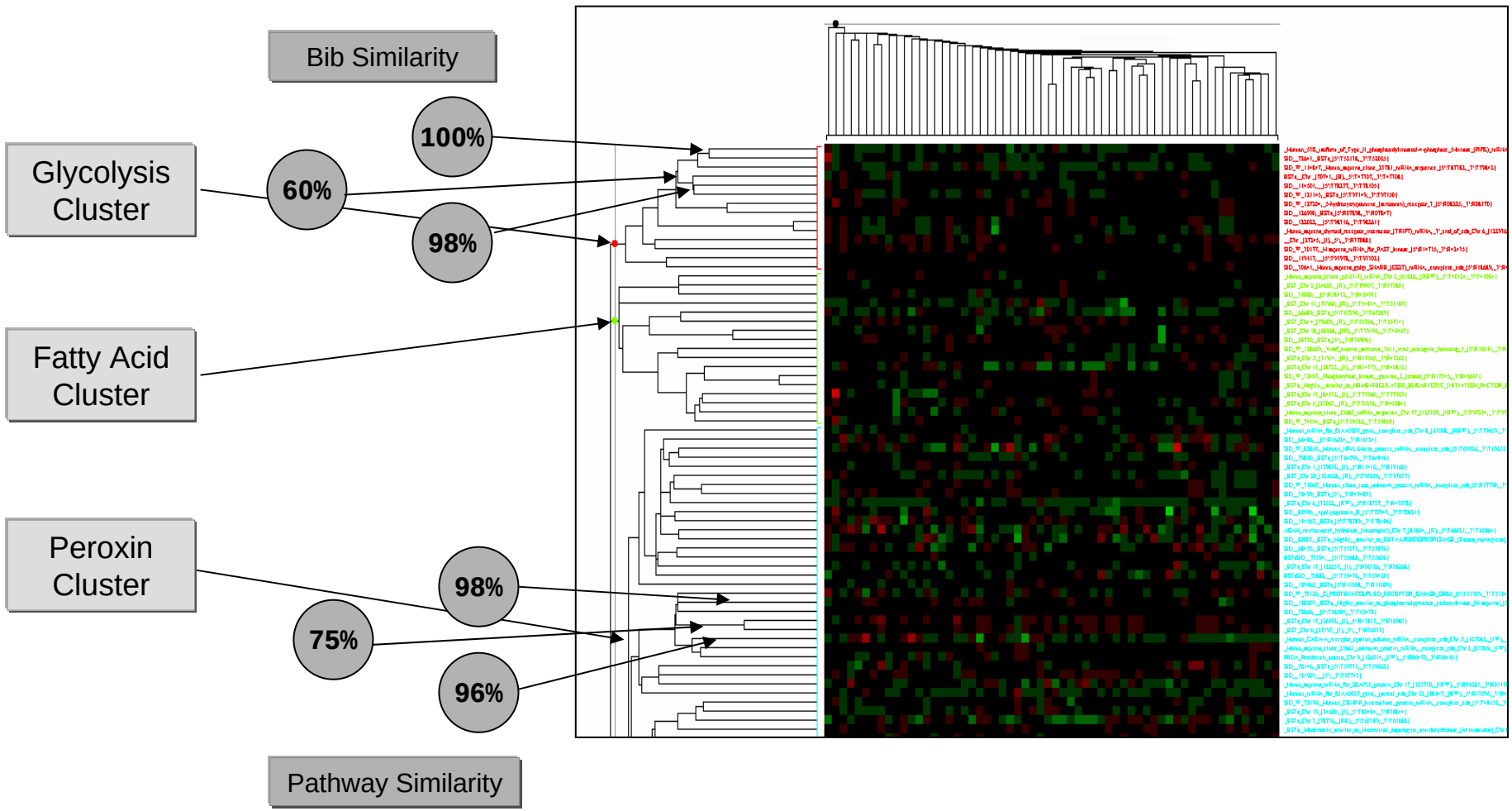


Discovery Net

Unifying the World's Knowledge

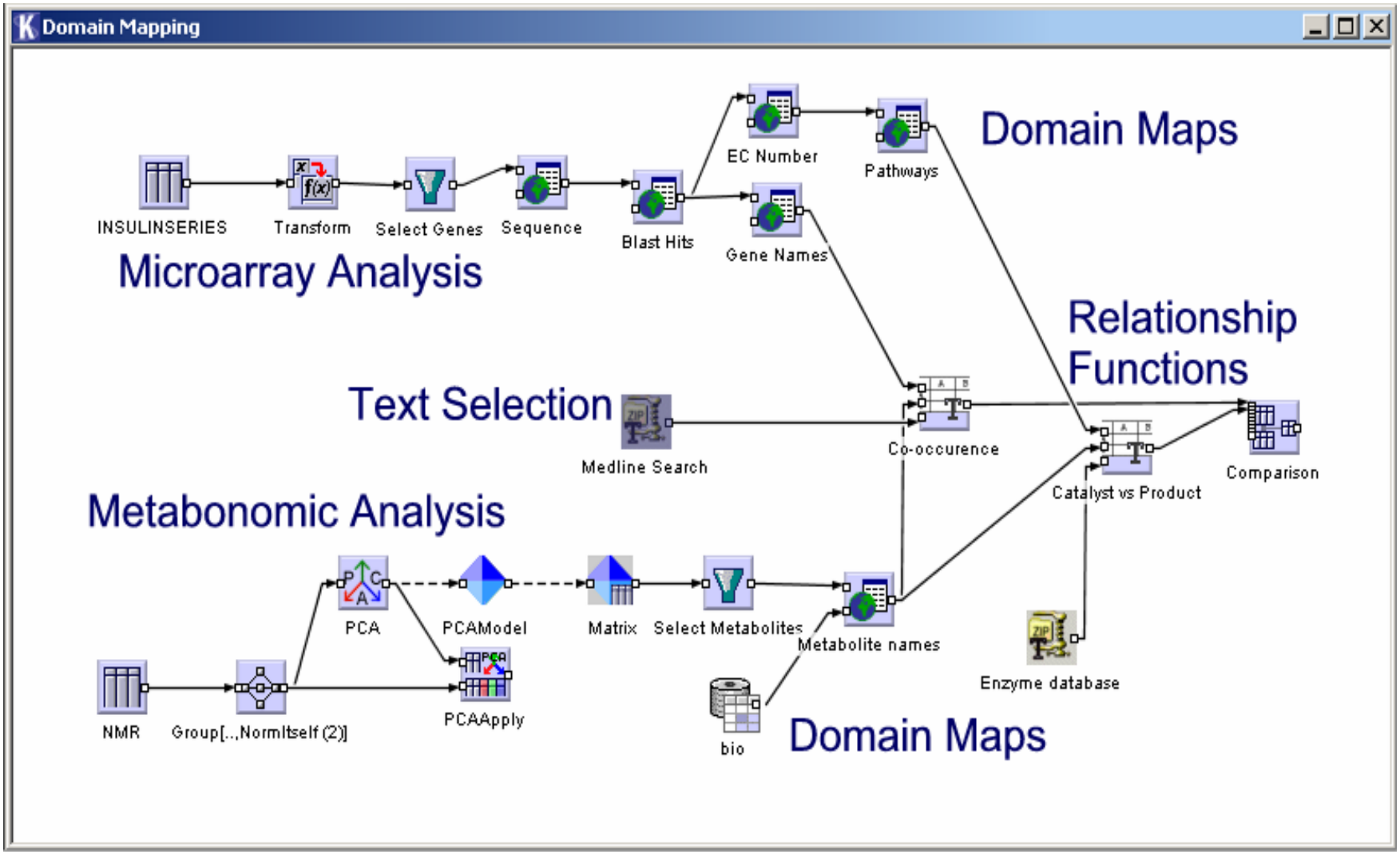
Leveraging Literature in Discovery Interpretation: Gene Group Interpretation

1. Traditional Data Mining Methods Generates Gene Clusters



2. Validate and Score Clusters using Text Information

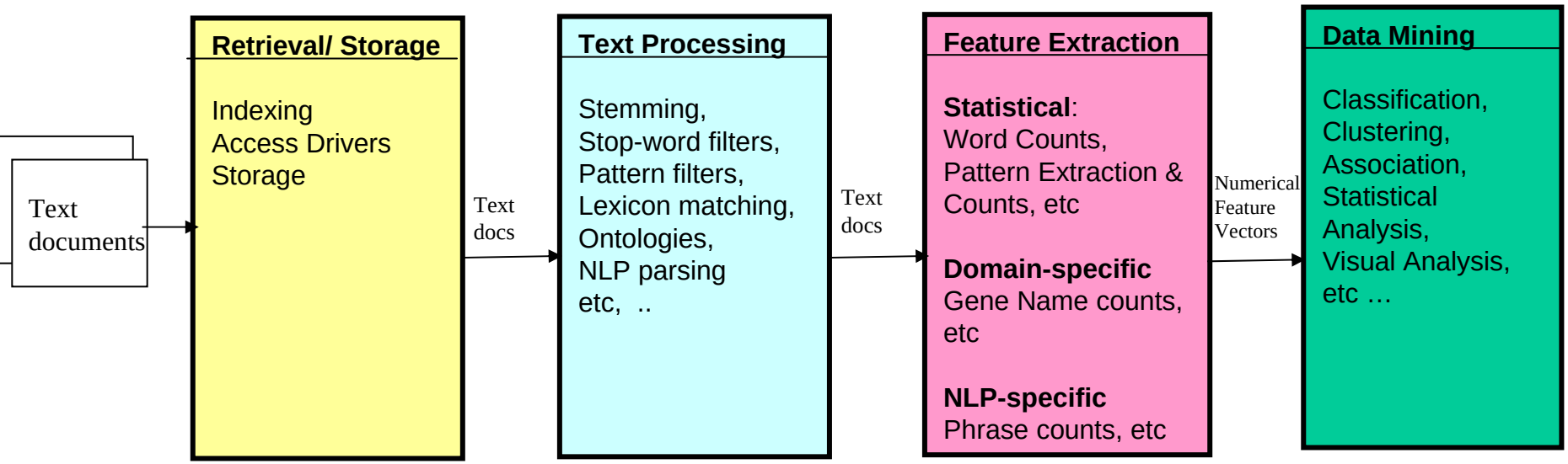
Leveraging Literature in Discovery Interpretation: Domain Mapping



Using literature analysis for the interpretation of gene expression and metabonomics data

Use e-Science technologies to build text mining services

Text Mining Pipelines



Retrieve and organize relevant documents

Pre-process documents to enhance the ease of feature extraction

Features are summarized into vector forms which are suitable for data mining

Results can be document characterization or hidden relationship extraction

Prioritised Document Lists
Example: Document Classification

Example 1: Develop a document classification system (KDD CUP 2002)

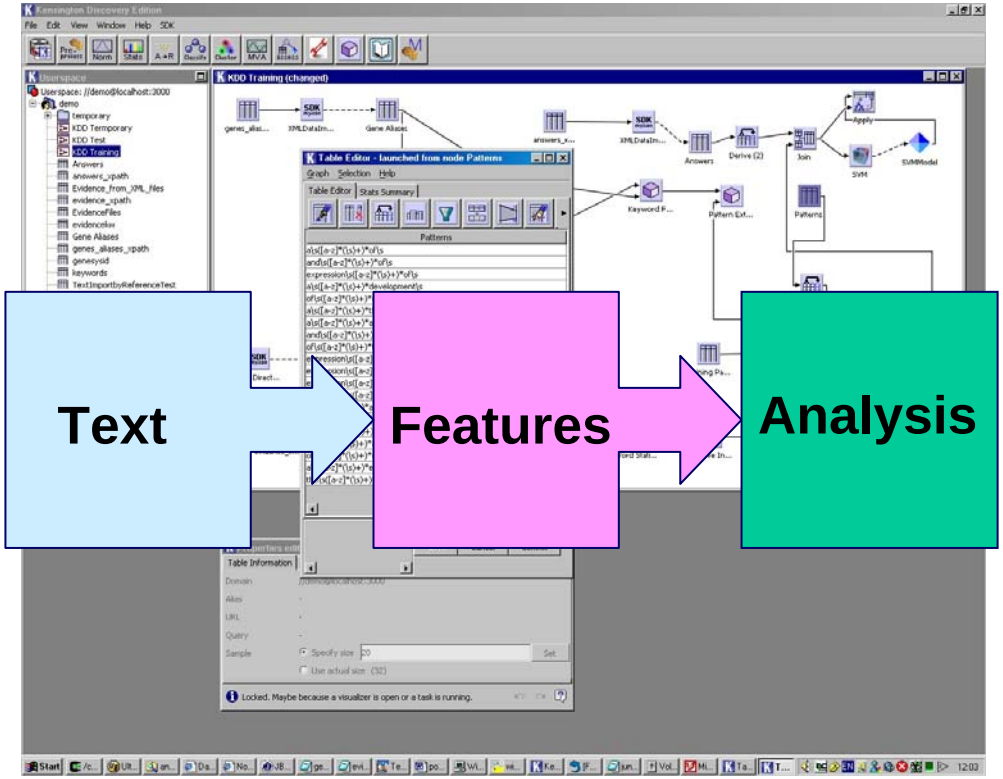
Training Set:

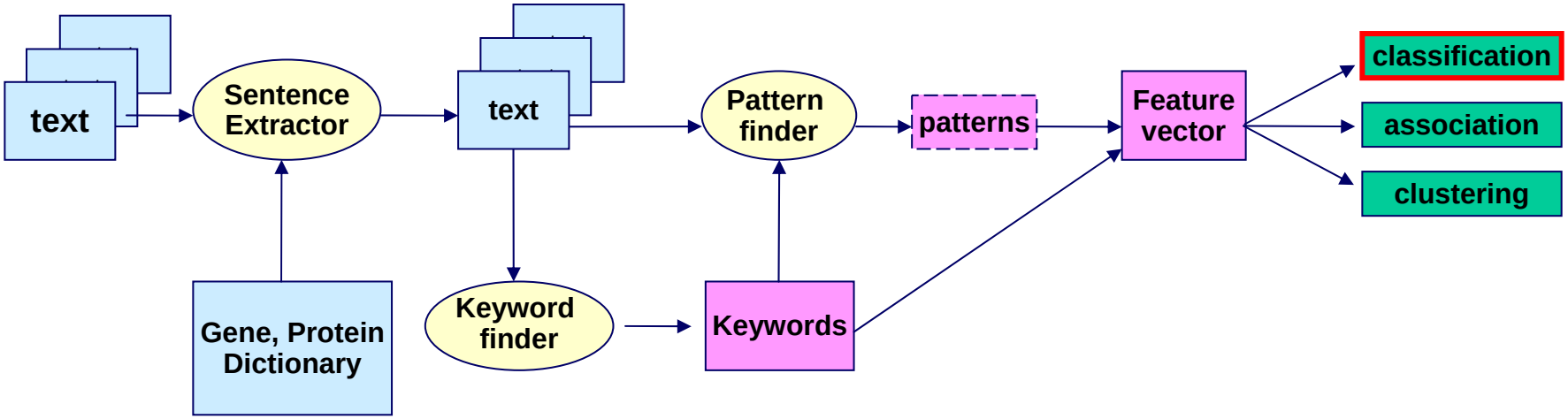
850 papers on Fruit Fly (each with the associated gene names identified in a list) were provided. Each paper is labelled as follows:

1. Whether it meets the curation criteria or not. (Y/N)
2. List of Genes names appearing in paper
3. For each Gene name paper a (RNA and/or Protein) label is provided for two types of gene expression products. (polypeptide, transcription, or both).

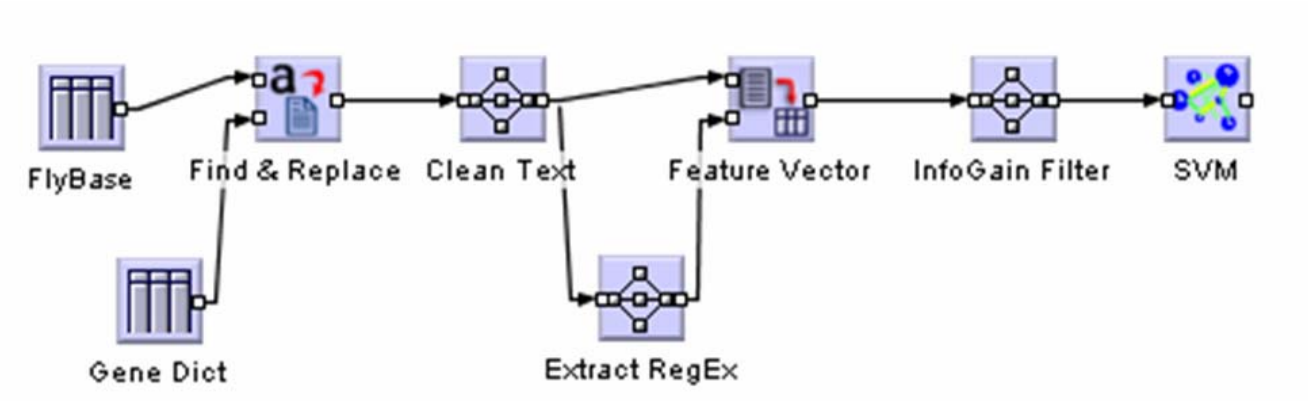
Task: Identify relevant documents from a previously unseen set

Application: Identifying Relevant Literature





Text Mining Workflows



Predictive Accuracy of Relevance prediction, using Support Vector Machine classification

Overall accuracy: 84.5%

Precision 78.11%

Recall 73.40%

HPC implementation of text mining nodes

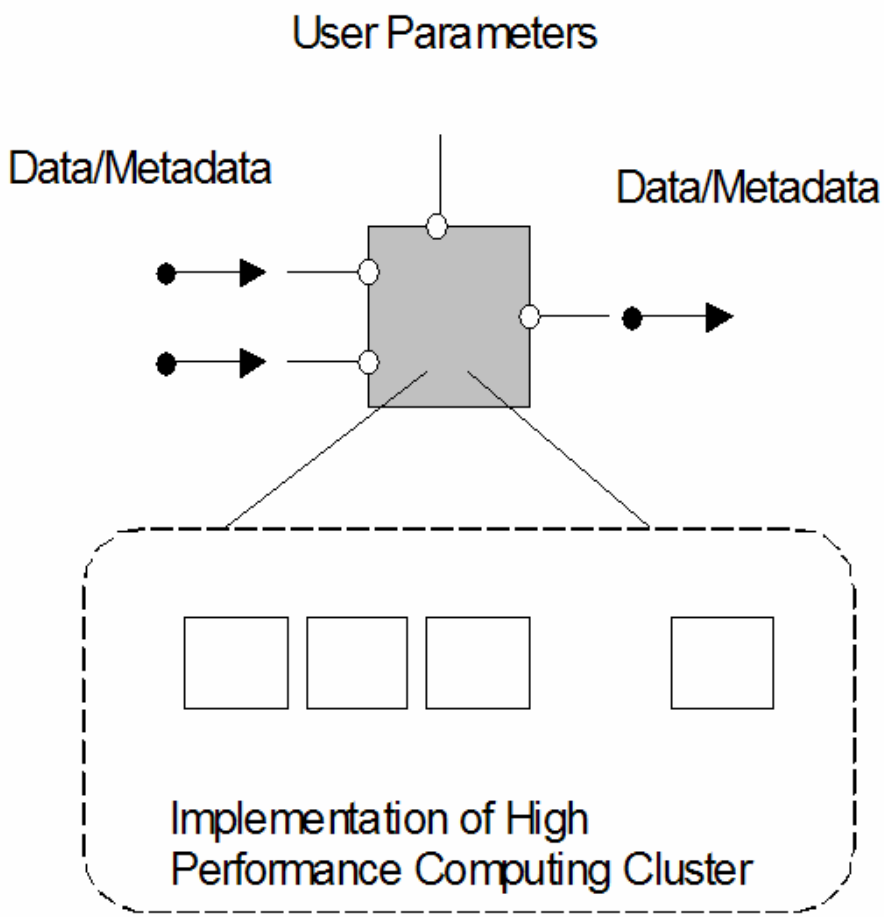
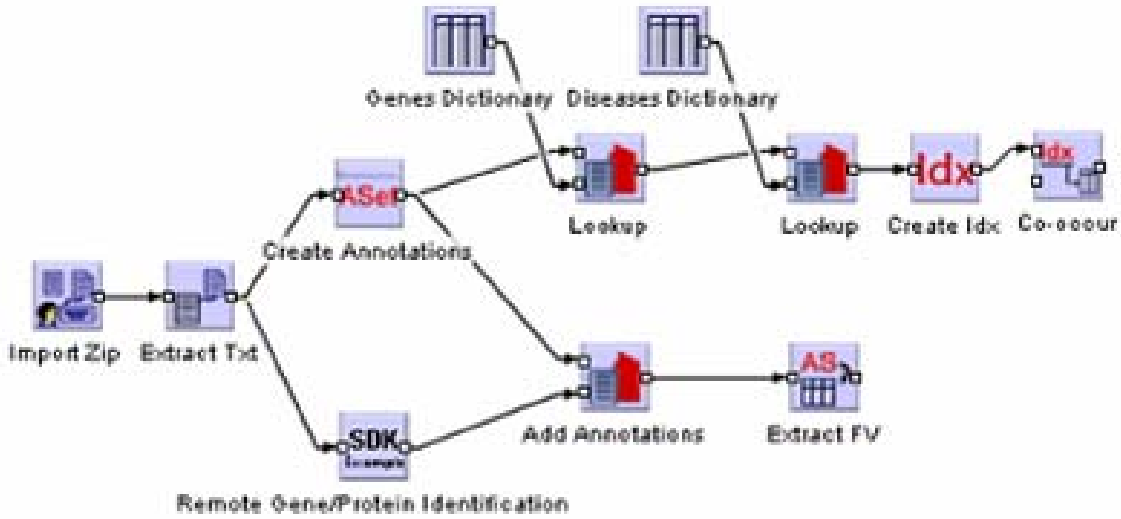


Figure 3: Web service interface / HPC implementation of a text mining component

Text Meta Data Model

Service Interfaces only tell you how to invoke remote service but it is up to you to decide what information flows between services !



Tipster-based model for in document annotation

Text	*Start*	*End*	*Annot. Type*	*Attributes*
Insulin	1	7	token	pos:noun, stem:insulin
resistance	9	18	token	pos:noun, stem:resist
Insulin resistance	1	18	compound token	disease:insulin resistance
plays	20	24	token	pos:verb, stem:plai
major	26	30	token	pos:adj, stem:major
role	32	35	Token	pos:noun, stem:role

Loose model allows co-operating text mining services, allows multiple parsing.

Why Tipster

Mining Over Annotated Data

Identify & Extract Biological/Chemical Entities from Documents

- Genes, Diseases, Compounds
- User-defined entities
- …

Query Entity Relationships

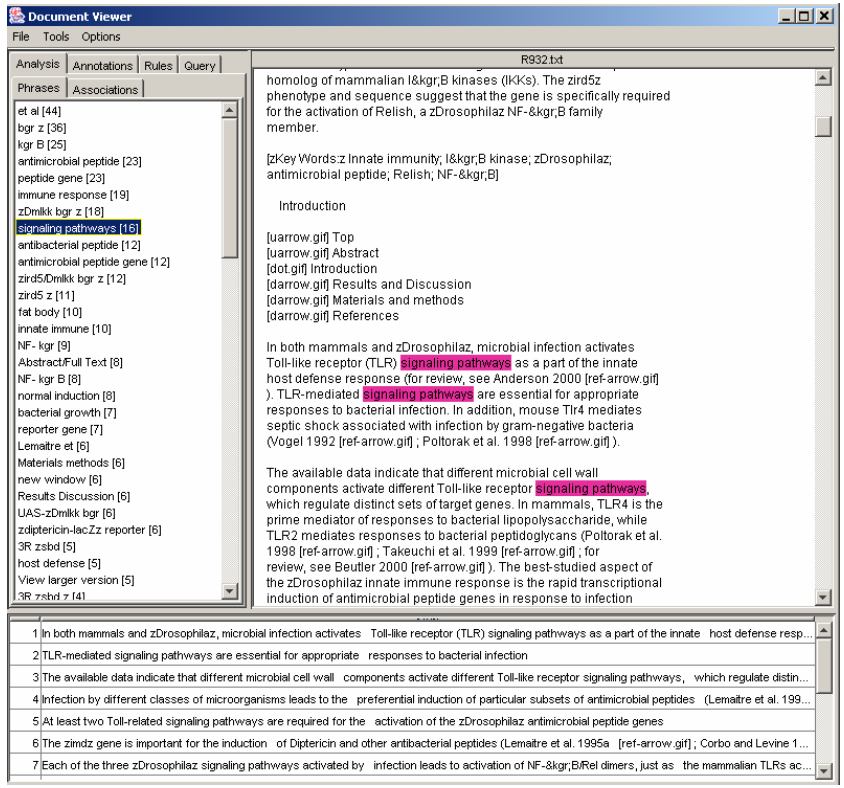
- Interactively Select Sentences/Patterns that satisfy annotation constraints
  - Base Entities: e.g. Disease and Gene
  - Entity Properties: e.g. Sentence, Verb, ..

Extract New Knowledge about Relationships between Entities

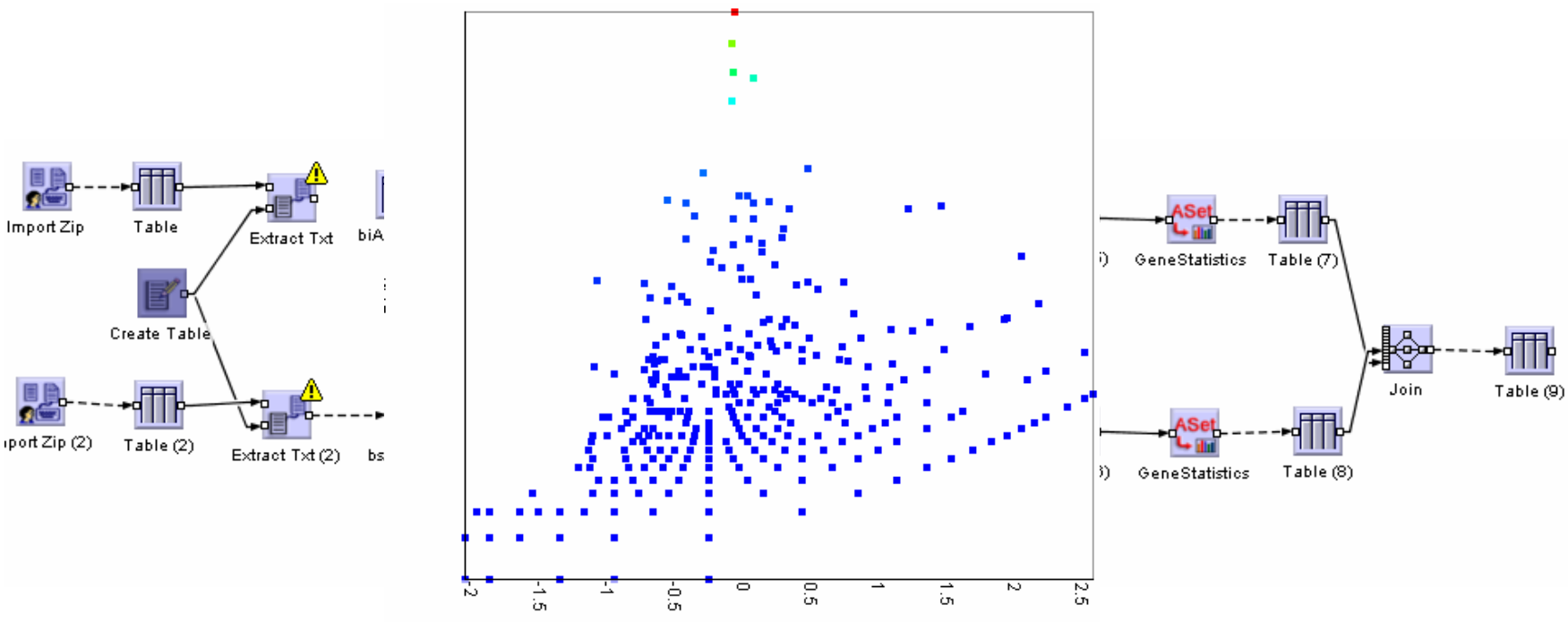
- e.g. association rules between entities or their properties
- e.g. phosphorylastion -> TSC2

Use Annotations as Feature Vectors

- Classification
- Clustering



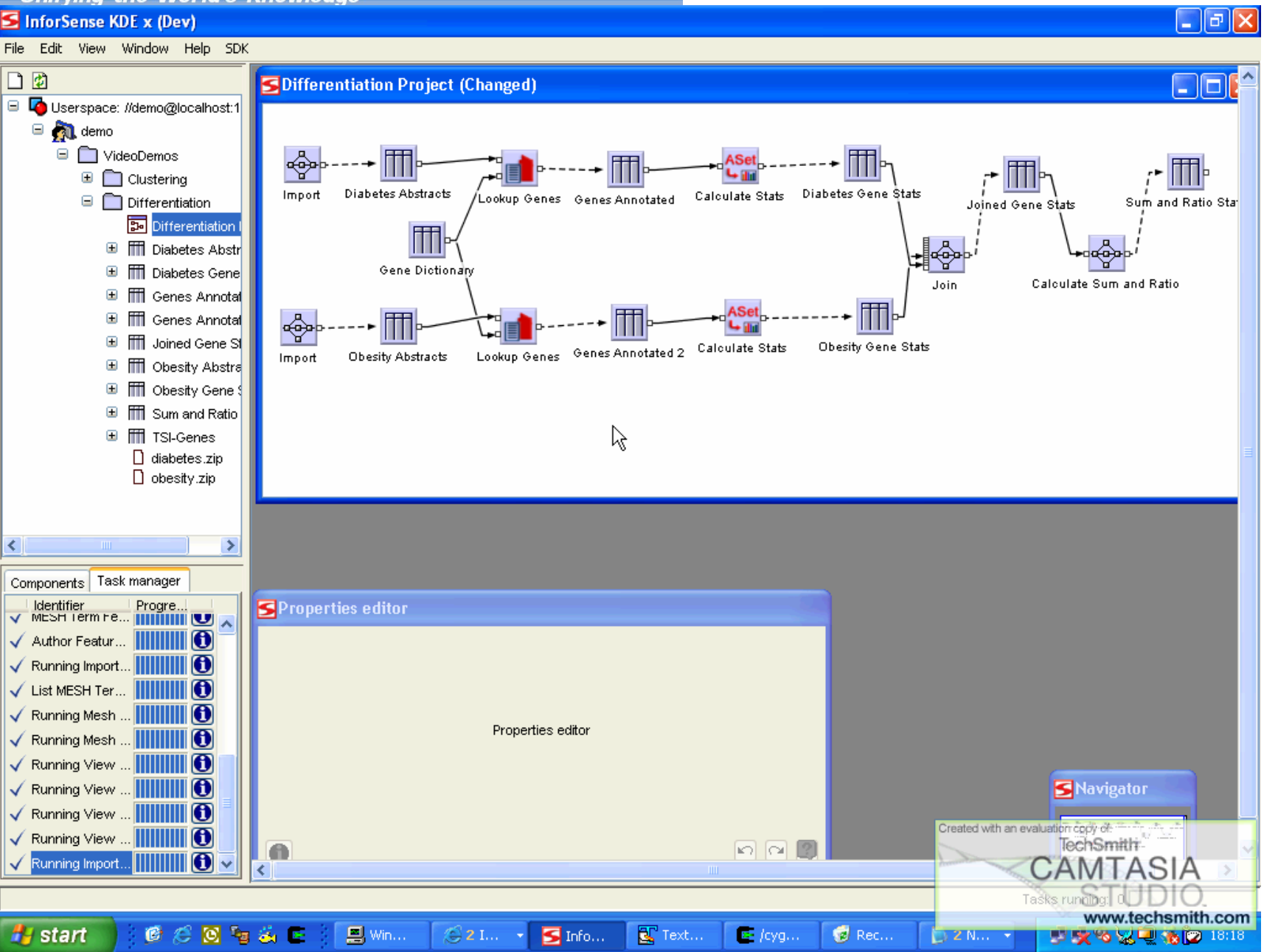
Application: Summarising difference between IR result sets



Discovery Net

Unifying the World's Knowledge

Short Animation



Discovery Net Text Mining Summary

Infrastructure

- Workflow Model for Service Programming
- Annotated Text Data Type
  - Holds both Text and Analysis Results (Annotations)
  - Tipster-based model
- Sparse Feature Vector Data Type
- Integration with DNet data mining tools and algorithms
- Interfaces to Lucene/Oracle Text
- Service Abstraction & Deployment

Components

- Import/Export Nodes (Word, PDF, Bibliographic Formats)
- Text Pre-processing Nodes (Clean, Replace, etc)
- Text Annotation/Analysis Nodes (Mark-up entities, N-grams, collocations, POS, etc)
- Indexing Nodes (Manage Large Data Sets)
- Feature Extraction & Mining Nodes (Stats, Compare, Classify, Cluster, etc)
- Viewer Nodes (Visually explore results)

e-Science & Text Mining:

Marriage made in heaven or hell?

- For application developers/ text mining community
  - Workflows & Service Programming gaining momentum in life sciences
  - Open Up Your Text Mining Tools
  - Agree on Interfaces & a Document Annotation/Metadata Models

- For end user scientists
  - No two scientist have exactly the same needs
    - Rapid prototyping/application development through service integration
  - Useful only when details are hidden
    - Workflows deployed higher level services
    - Workflows accessed from friendly user interfaces

Discovery Net

Unifying the World's Knowledge

Discovery Net e-science infrastructure

Infrastructure for Global Knowledge Discovery Services