

# **e-Science & Text Mining**

*a marriage made in heaven?*

**Moustafa M. Ghanem**  
**Department of Computing**  
**Imperial College London**

### Motivations

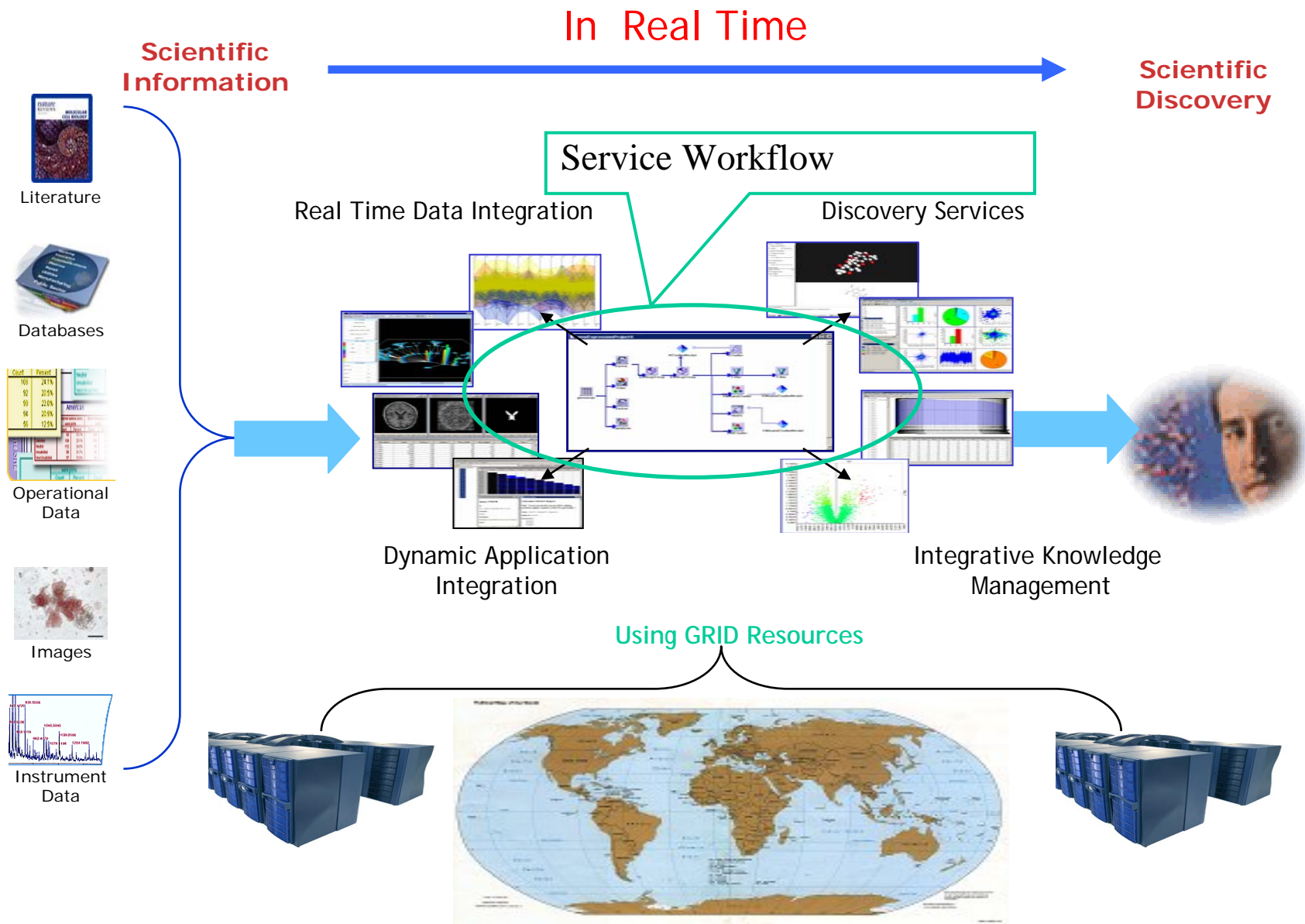
- Add text mining services to e-Science projects
- Use e-Science technologies to build text mining services

### Approach

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

- **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

### 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

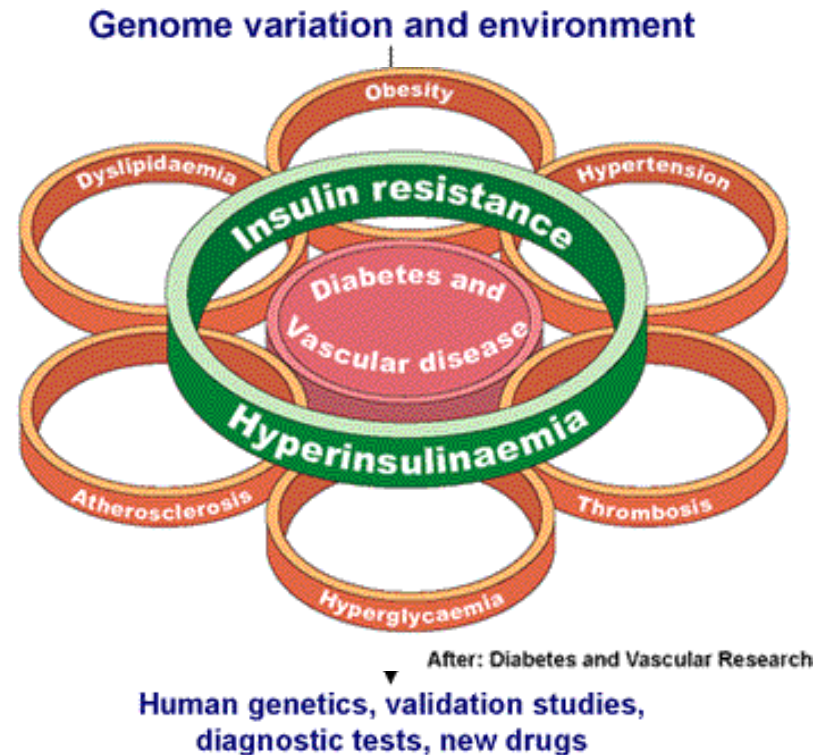
Working in collaboration with scientists working on the £5.5M BAIR project funded by the Wellcome Trust.

### 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.

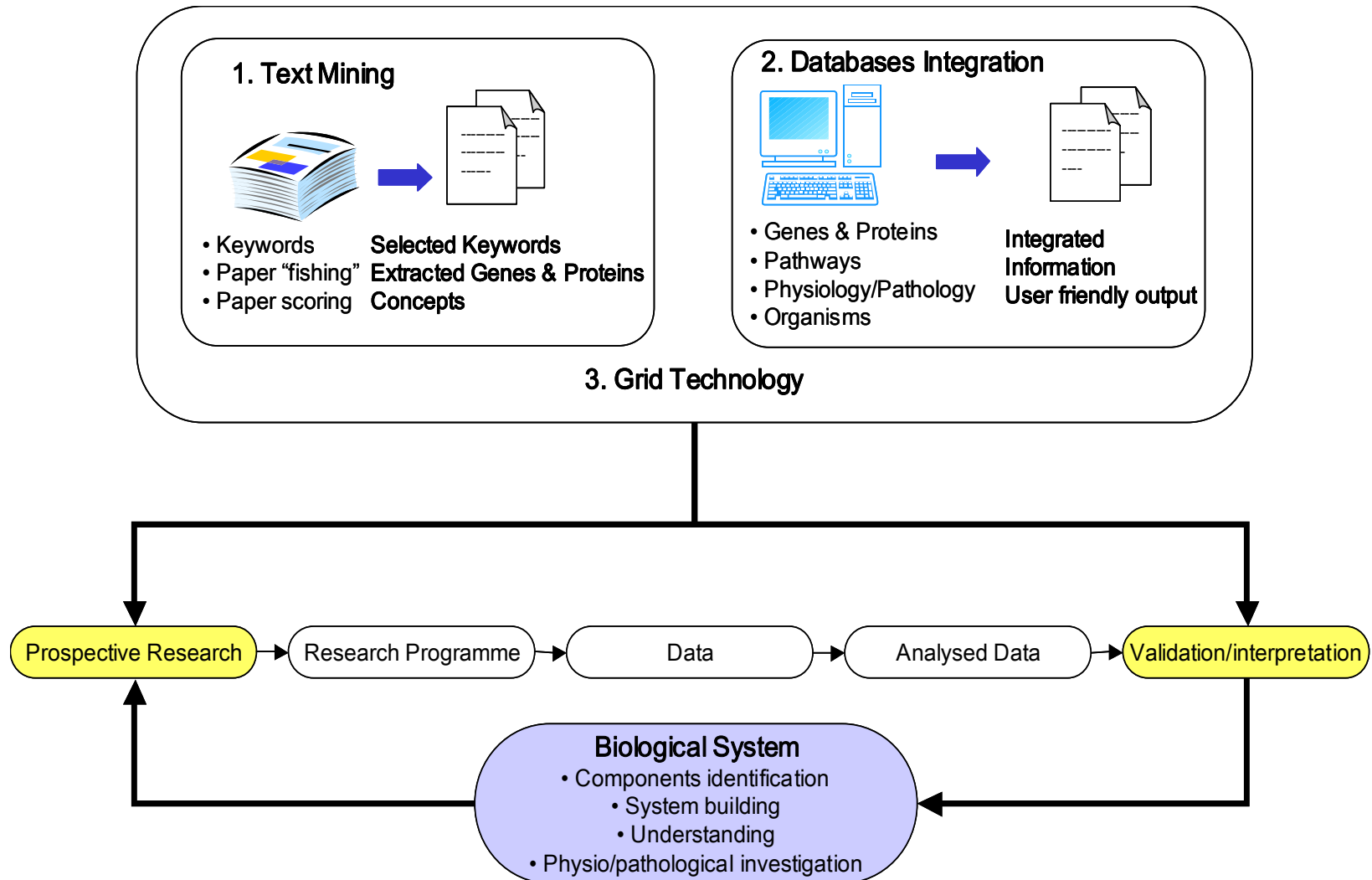
*We have adopted the term “atlas” because it emphasises that our aim is a description of the insulin-action “universe” that incorporates and integrates many different types of information, is flexible enough to cope with continuous updating and revision, and capable of serving as a discovery tool – for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

*Prof James Scott, FRS.*



### What Scientists actually from Text Mining?

### Example from Integrative Systems Biology Studies



### What Scientists actually want from Text Mining?

#### Examples from Integrative Biology Studies

##### Leveraging Literature in Discovery Interpretation

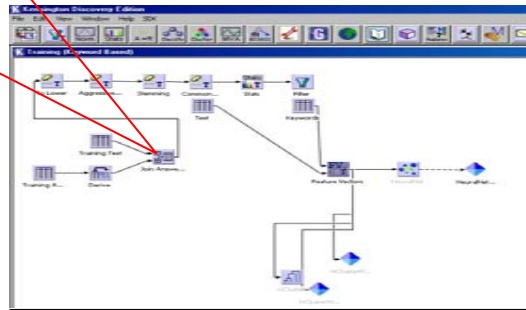
- Automatic Result Annotation
  - (e.g. What is the difference/similarity between microarray results)

##### Generating Prioritised Document Lists

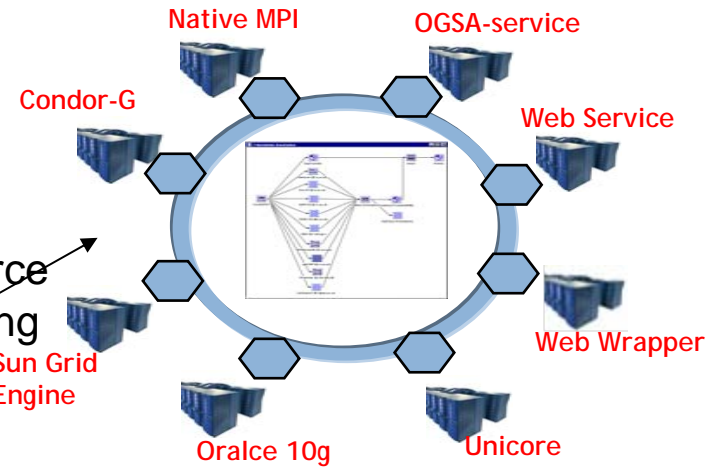
- Automatic Document Categorization
- Topic Maps
- Linking literature to available background knowledge

##### 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)



Resource Mapping

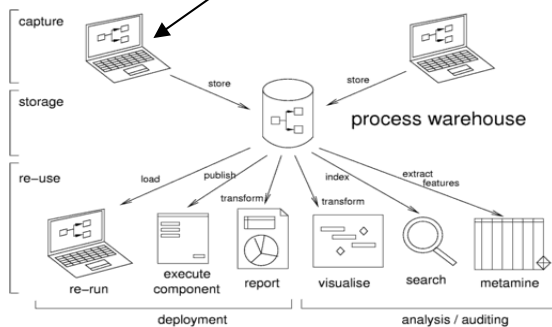


**Workflow Execution**  
A compositional GRID

**Workflow Authoring**  
**Composing services**

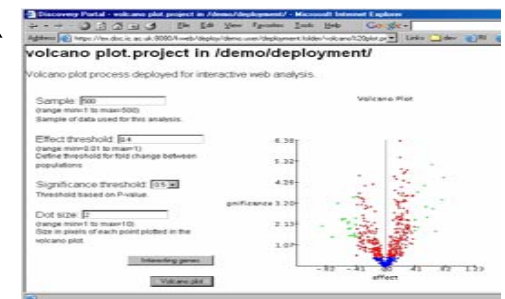
Workflow  
Warehousing

Service  
Abstraction



**Workflow Management**

**Collaborative Knowledge Management**

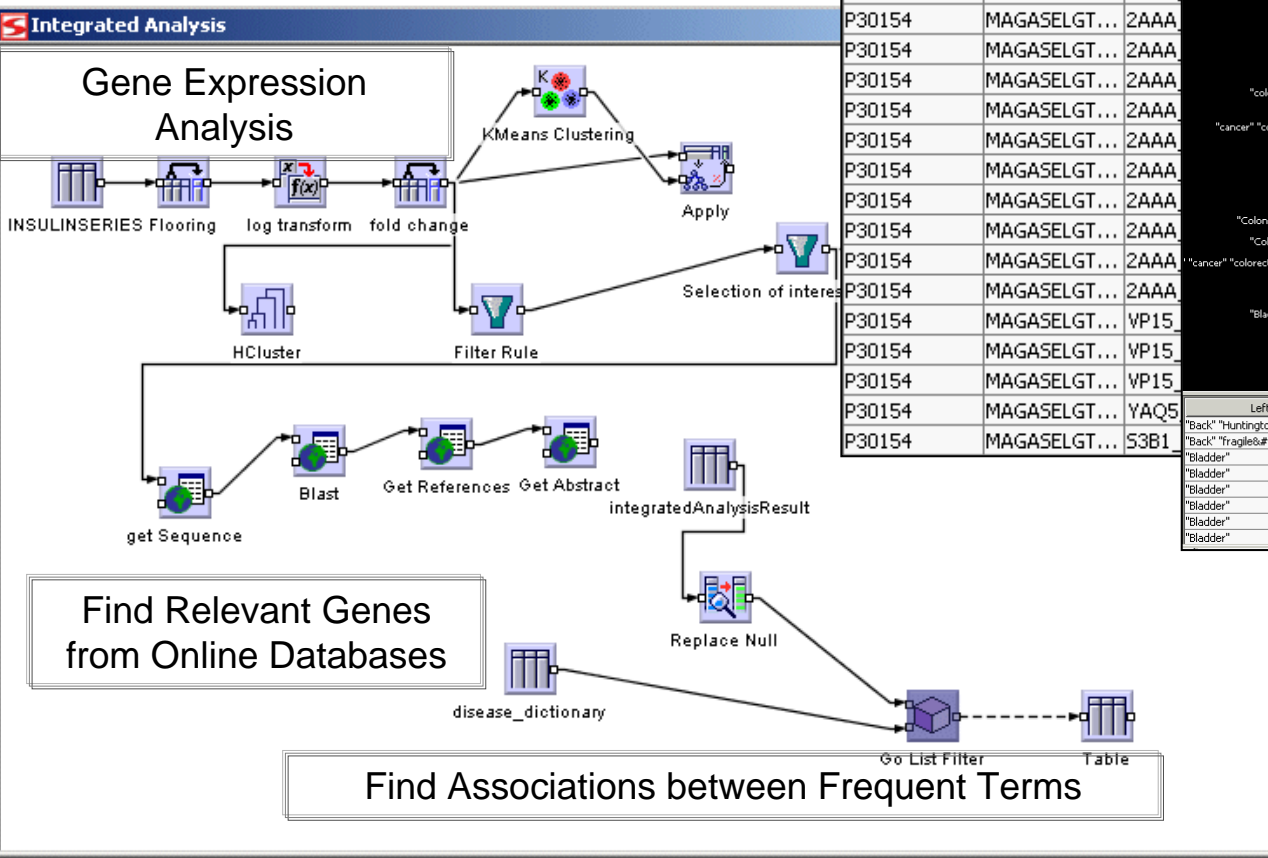


**Workflow Deployment:**  
**Grid Service and Portal**

## 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...	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			

"gastric"

"colorectal cancer"

"colorectal"

"cancer" "colorectal cancer"

"cancer"

"brain"

"Face" "paralysis"

"Face" "gastric"

"Colon" "colorectal cancer"

"Colon" "cancer" "finger"

"cancer" "colorectal" "colorectal cancer"

"Colon"

"Cervix" "breast"

"Bladder" "breast" "stomach"

"infection" "paralysis"

"cancer" "colorectal" "colorectal cancer"

"Face" "infection" "paralysis"

"Colon" "cancer" "finger"

"Cervix" "breast"

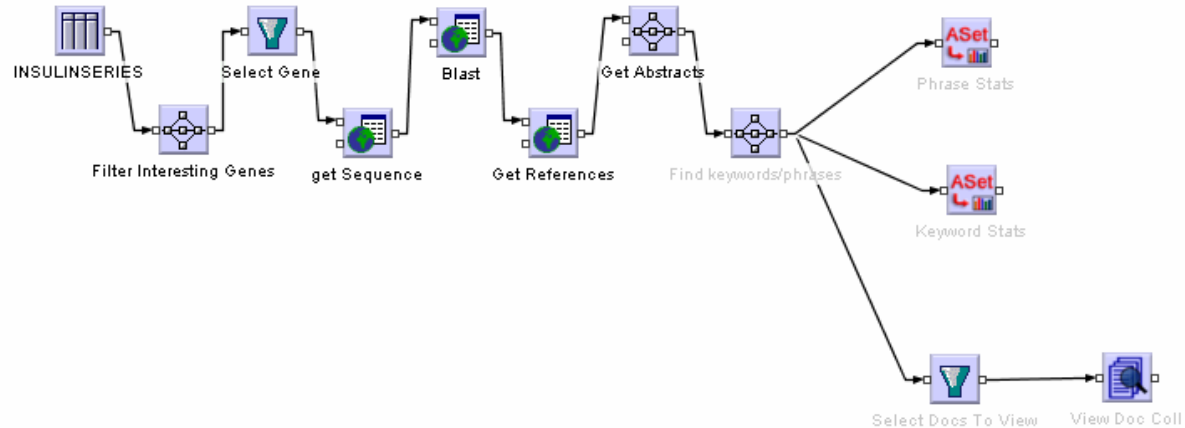
Left side	Right side	Confidence	Support
"Back" "Huntington&apos;s#32;..."	"Fragile&#32;X#32;syndrome"	1.0	0.023256
"Back" "Fragile&#32;X#32;syn..."	"Huntington&apos;s#32;disease"	1.0	0.023256
"Bladder"	"Cervix"	1.0	0.023256
"Bladder"	"Cervix" "breast"	1.0	0.023256
"Bladder"	"Cervix" "breast" "stomach"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



Userspace: //demo@localhost:1099

- demo
  - GeneExpression
    - Integrated Analysis (clean)
    - Integrated Analysis
    - VolcanoPlot Complete
  - disease\_dictionary
  - INSULINSERIES
  - WILD\_VS\_TREATED

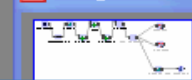
## Integrated Analysis (clean)



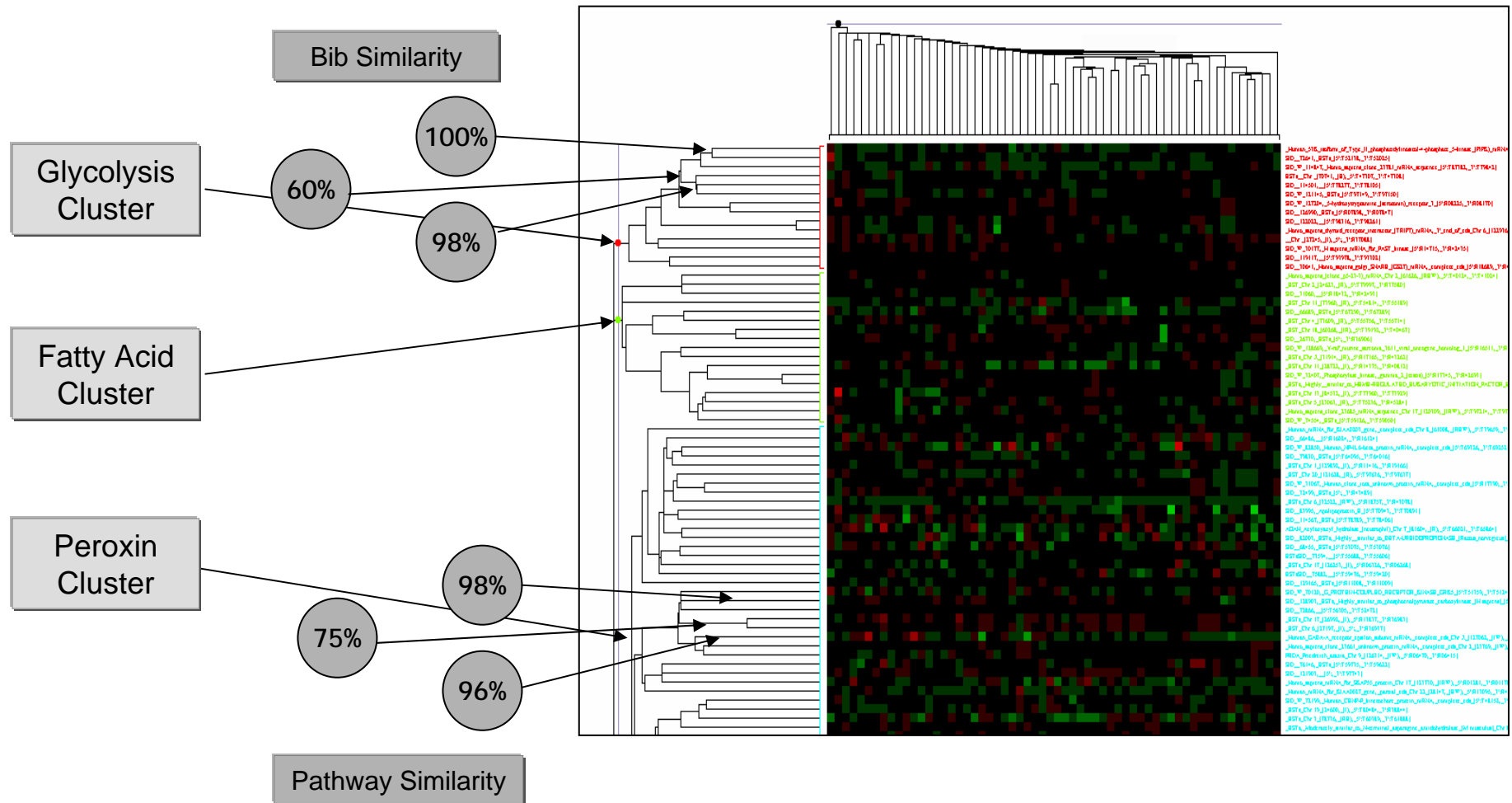
## Properties editor

Properties editor

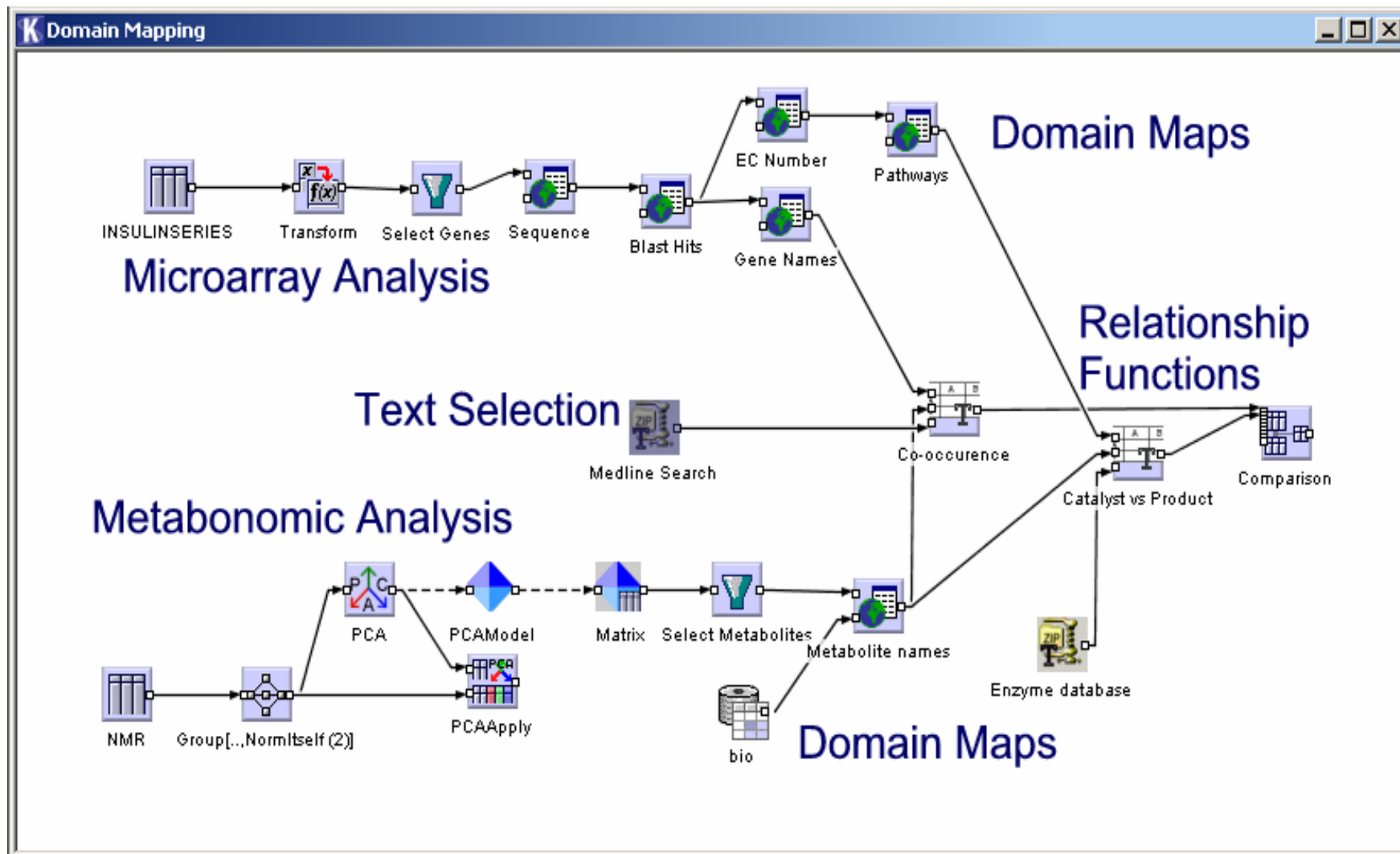
## Navigator



### 1. Traditional Data Mining Methods Generates Gene Clusters



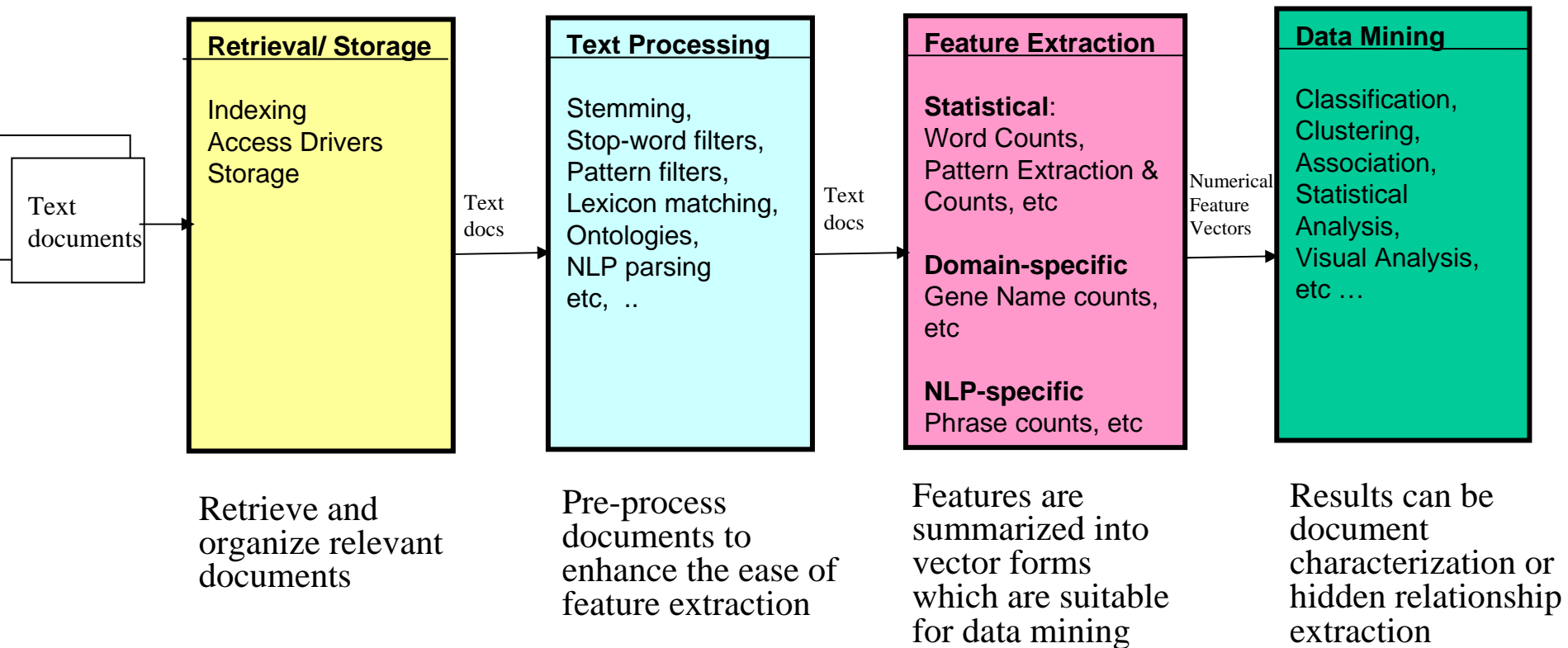
### 2. Validate and Score Clusters using Text Information



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

## Use e-Science technologies to build text mining services

### Text Mining Pipelines



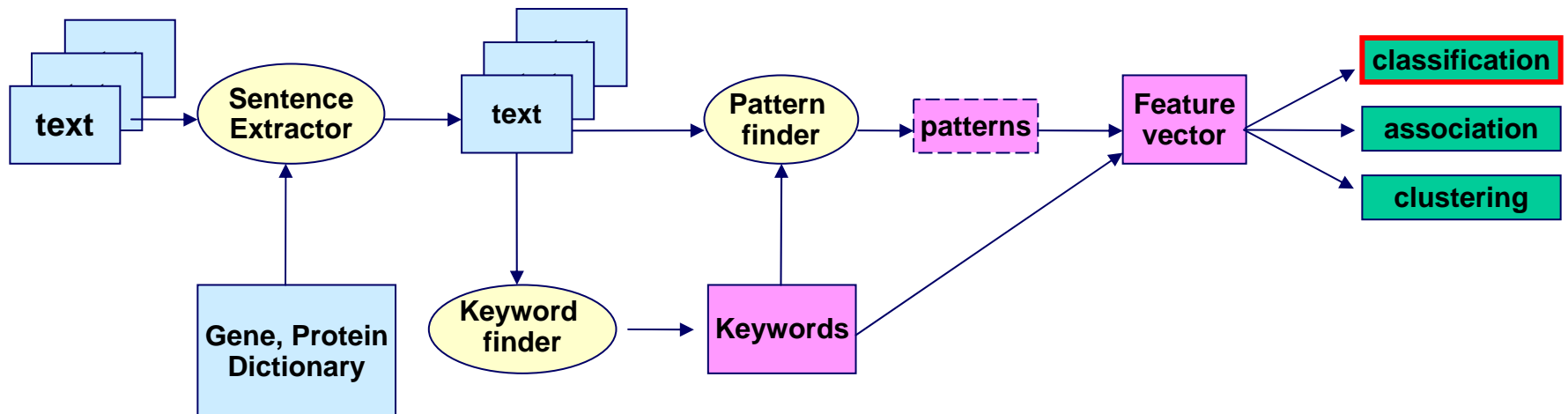
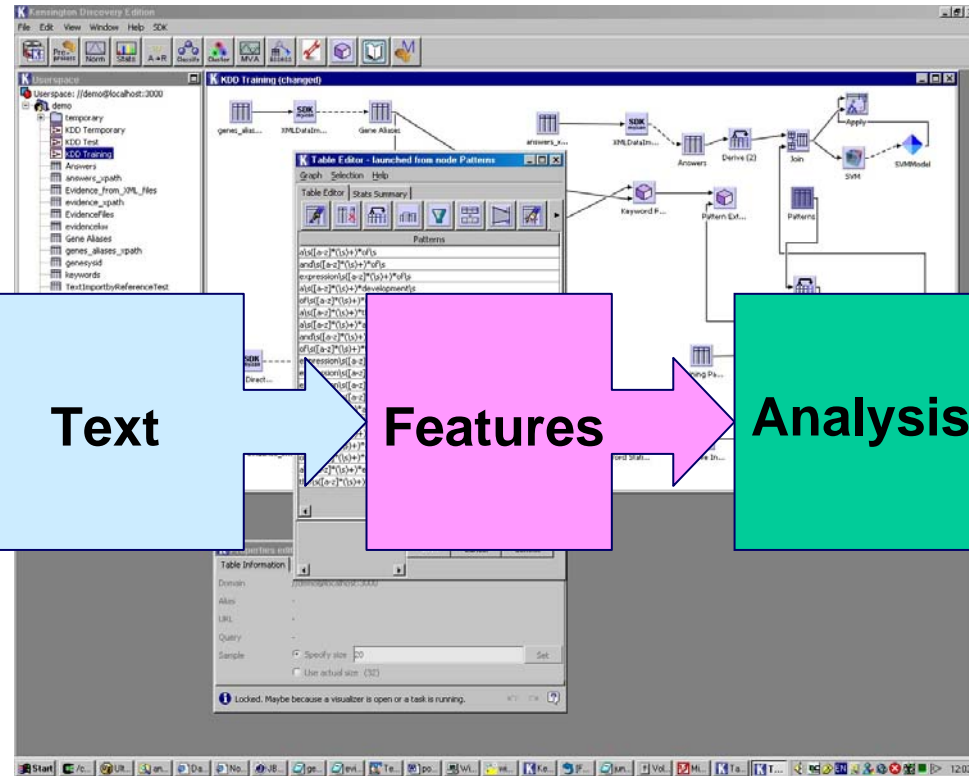
## Example 1: Develop a document classification system (KDD CUP 2002)

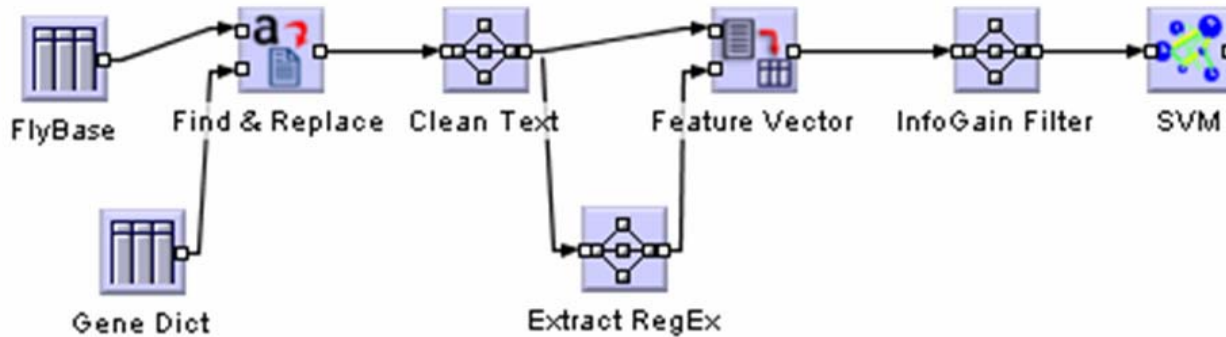
### Training Set:

**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**





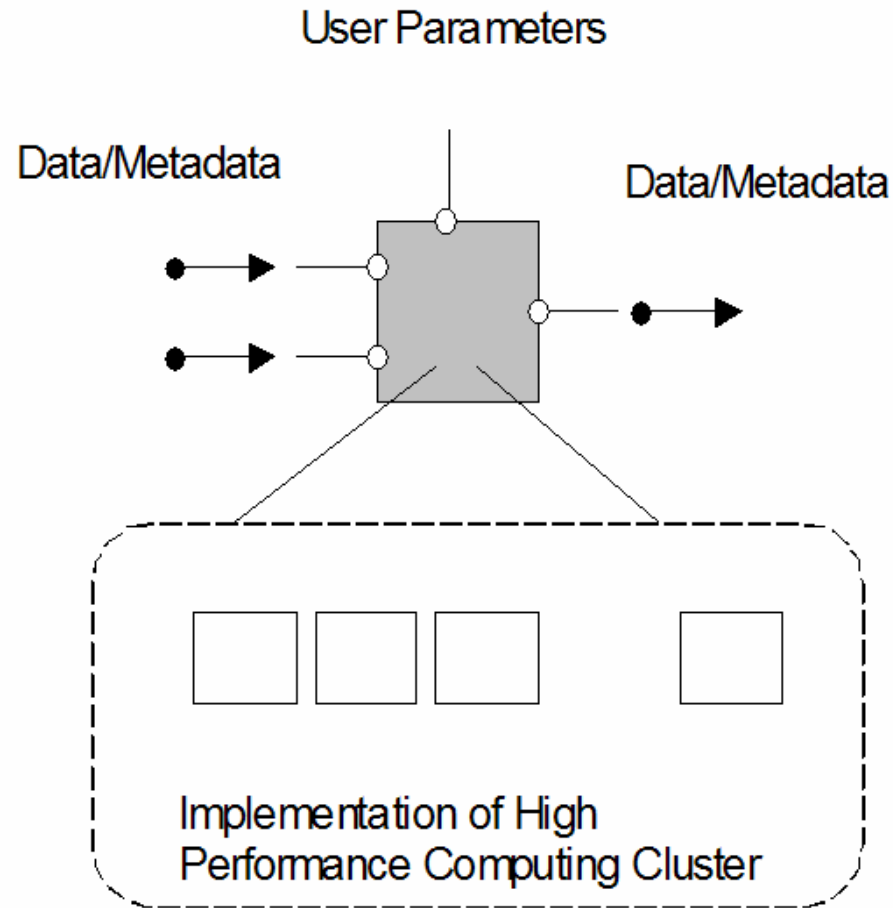
**Predictive Accuracy of Relevance prediction,  
using Support Vector Machine classification**

**Overall accuracy: 84.5%**

**Precision 78.11%**

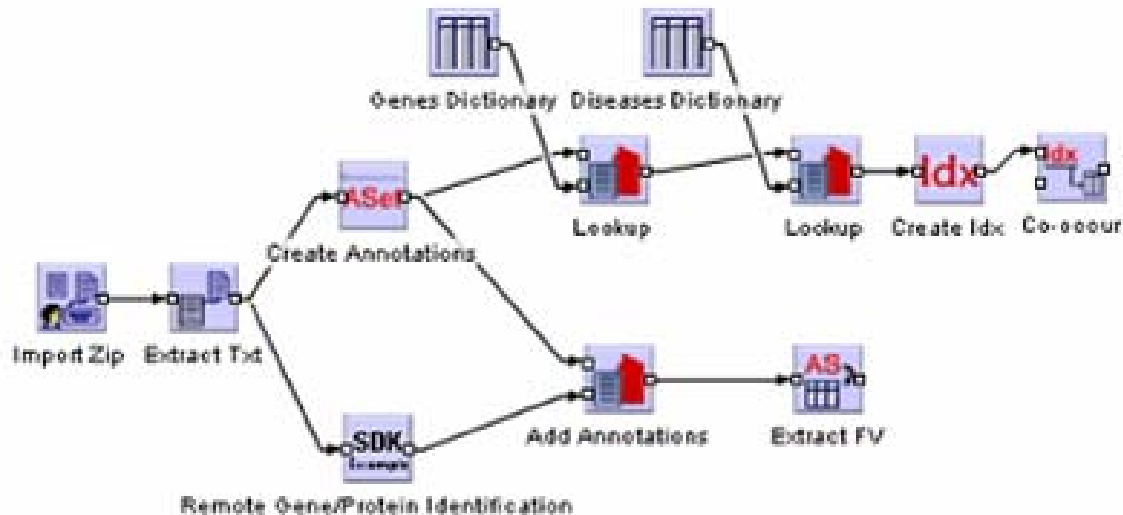
**Recall 73.40%**

## HPC implementation of text mining nodes



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

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

<i>Text</i>	<i>Start</i>	<i>End</i>	<i>Annot. Type</i>	<i>Attributes</i>
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

## 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. phosphorylation -> TSC2

## Use Annotations as Feature Vectors

- Classification
- Clustering

Document Viewer

File Tools Options

Analysis Annotations Rules Query

Phrases Associations

et al [44]  
bgr z [36]  
lgr B [25]  
antimicrobial peptide [23]  
peptide gene [23]  
immune response [19]  
zDmkk bgr z [18]  
**Signaling pathways [15]**  
antibacterial peptide [12]  
antimicrobial peptide gene [12]  
zird5/Dmkk bgr z [12]  
zird5 z [11]  
fat body [10]  
innate immune [10]  
NF- kgr [9]  
Abstract/Full Text [8]  
NF- kgr B [8]  
normal induction [8]  
bacterial growth [7]  
reporter gene [7]  
Lemaître et al [6]  
Materials methods [6]  
new window [6]  
Results Discussion [6]  
UAS-zDmkk bgr [6]  
zDiptericin-lacZ reporter [6]  
3R zshd [5]  
host defense [5]  
View larger version [5]  
3R zshd z 141

R932.txt

homolog of mammalian I&agr;B kinases (IKKs). The zird5z phenotype and sequence suggest that the gene is specifically required for the activation of Relish, a zDrosophilaz NF-&agr;B family member.

[zKey Words z Innate immunity, I&agr;B kinase, zDrosophilaz, antimicrobial peptide; Relish; NF-&agr;B]

Introduction

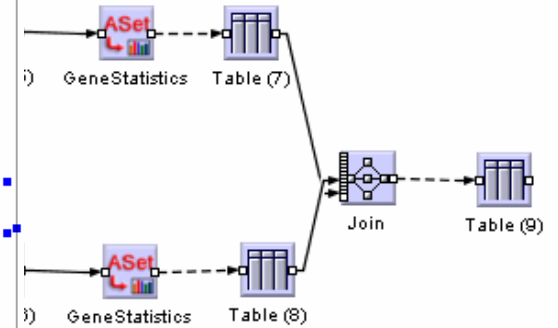
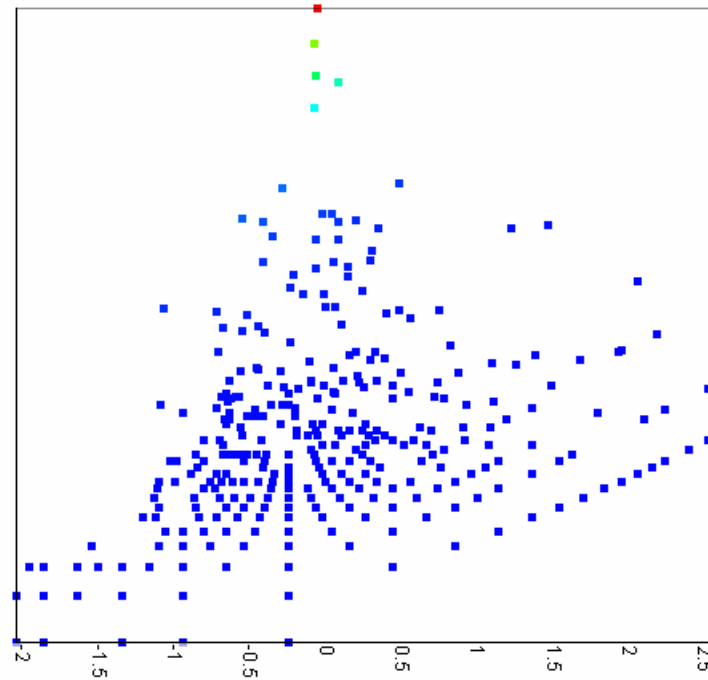
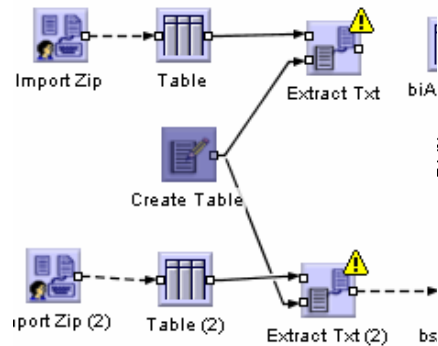
[uarrow.gif] Top  
[uarrow.gif] Abstract  
[dot.gif] Introduction  
[darrow.gif] Results and Discussion  
[darrow.gif] Materials and methods  
[darrow.gif] References

In both mammals and zDrosophilaz, microbial infection activates Toll-like receptor (TLR) **signaling pathways** as a part of the innate host defense response (for review, see Anderson 2000 [ref-arrow.gif]). TLR-mediated **signaling pathways** are essential for appropriate responses to bacterial infection. In addition, mouse Tlr4 mediates septic shock associated with infection by gram-negative bacteria (Vogel 1992 [ref-arrow.gif]; Poltorak et al. 1998 [ref-arrow.gif]).

The available data indicate that different microbial cell wall components activate different Toll-like receptor **signaling pathways**, which regulate distinct sets of target genes. In mammals, TLR4 is the prime mediator of responses to bacterial lipopolysaccharide, while TLR2 mediates responses to bacterial peptidoglycans (Poltorak et al. 1998 [ref-arrow.gif]; Takeuchi et al. 1999 [ref-arrow.gif]), for review, see Beutler 2000 [ref-arrow.gif]). The best-studied aspect of the zDrosophilaz innate immune response is the rapid transcriptional induction of antimicrobial peptide genes in response to infection

1	In both mammals and zDrosophilaz, microbial infection activates Toll-like receptor (TLR) signaling pathways as a part of the innate host defense resp...
2	TLR-mediated signaling pathways are essential for appropriate responses to bacterial infection
3	The available data indicate that different microbial cell wall components activate different Toll-like receptor signaling pathways, which regulate distin...
4	Infection by different classes of microorganisms leads to the preferential induction of particular subsets of antimicrobial peptides (Lemaître et al. 199...
5	At least two Toll-related signaling pathways are required for the activation of the zDrosophilaz antimicrobial peptide genes
6	The zird5 gene is important for the induction of Dipterin and other antibacterial peptides (Lemaître et al. 1995a [ref-arrow.gif]; Corbo and Levine 1...
7	Each of the three zDrosophilaz signaling pathways activated by infection leads to activation of NF-&agr;B/Rel dimers, just as the mammalian TLRs ac...

## Application: Summarising difference between IR result sets





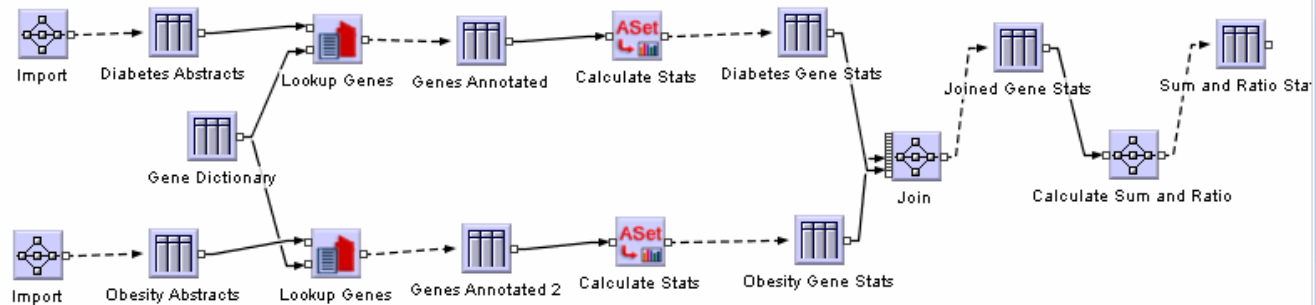
Userspace: //demo@localhost:1

- demo
  - VideoDemos
    - Clustering
    - Differentiation
      - Differentiation I
      - Diabetes Abstr
      - Diabetes Gene
      - Genes Annotat
      - Genes Annotat
      - Joined Gene S
      - Obesity Abstra
      - Obesity Gene S
      - Sum and Ratio
      - TSI-Genes
      - diabetes.zip
      - obesity.zip

Components Task manager

Identifier	Progre...
MESH term fe...	
Author Featur...	
Running Import...	
List MESH Ter...	
Running Mesh ...	
Running Mesh ...	
Running View ...	
Running View ...	
Running View ...	
Running View ...	
Running View ...	
Running Import...	

## Differentiation Project (Changed)



## Properties editor

Properties editor

Navigator

### 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

### Infrastructure for Global Knowledge Discovery Services

