



# Oracle Text Mining

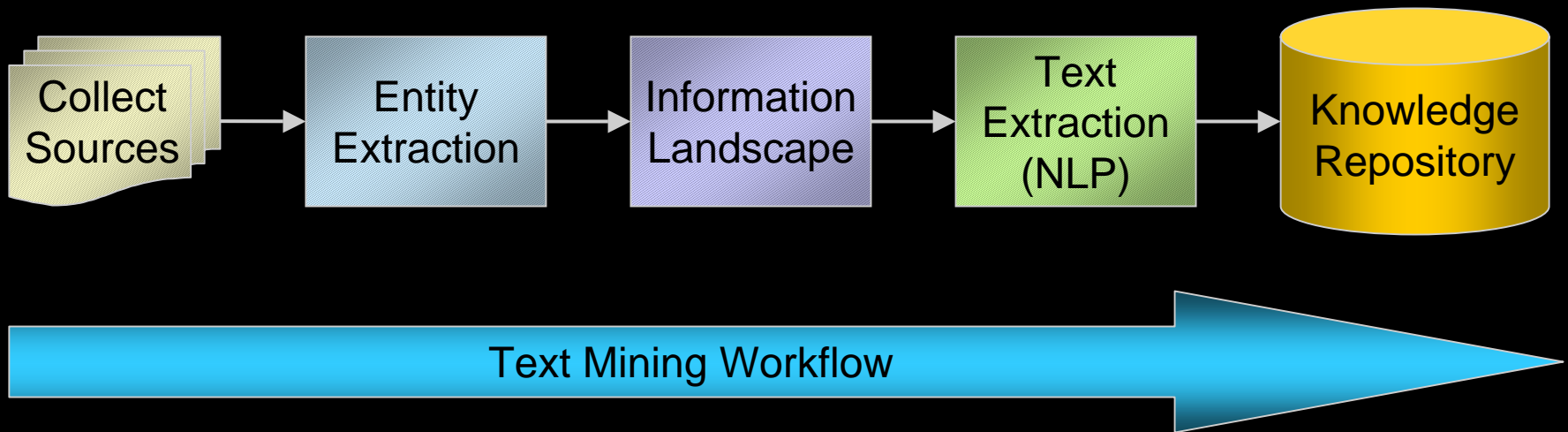
## OLSUG Meeting

16 May 2005

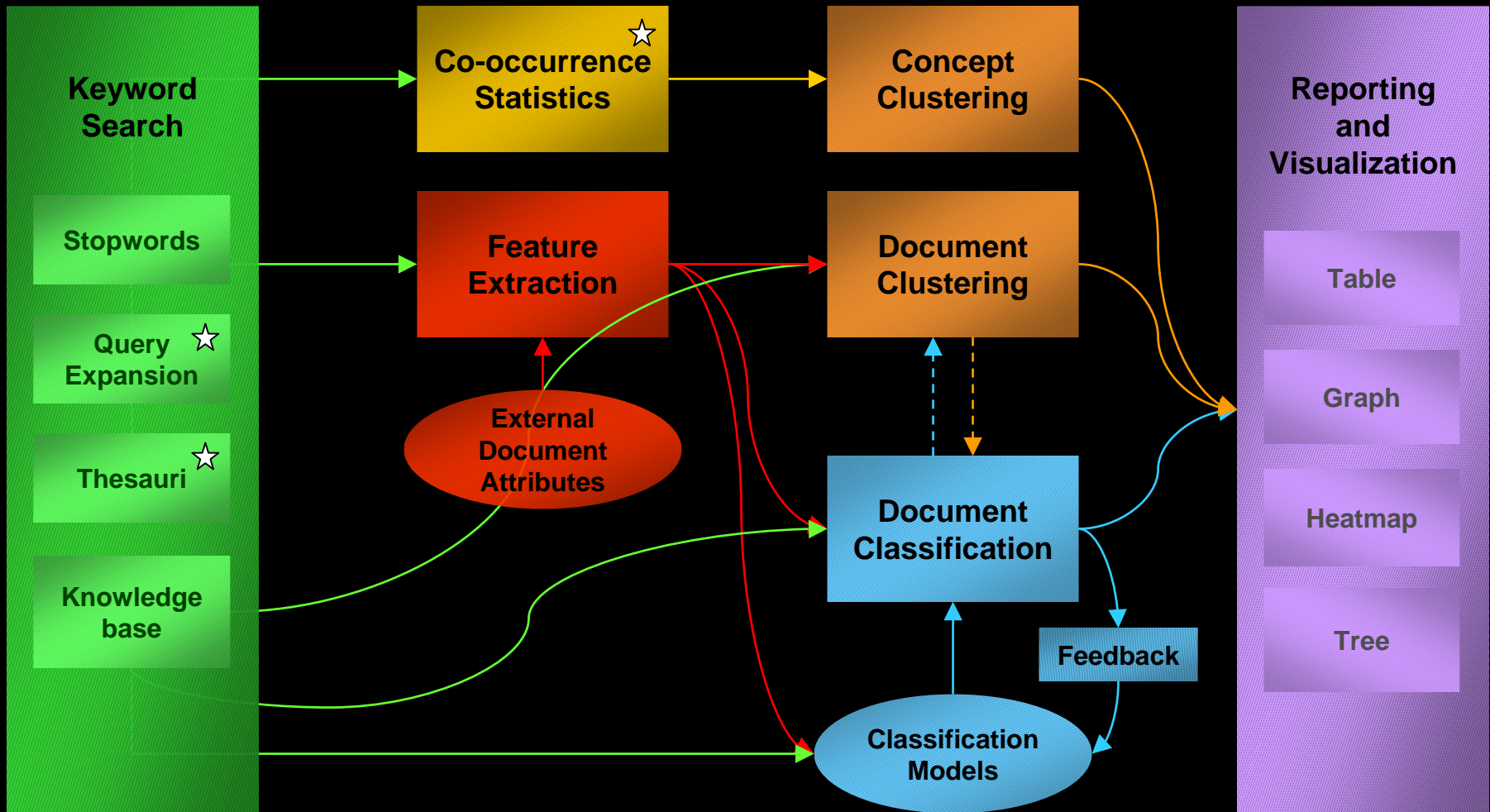
Raf Podowski, Sr. Product Manager, Life Sciences

[raf.podowski@oracle.com](mailto:raf.podowski@oracle.com)

# Text Mining Workflow



# Oracle Text Mining Application



★ NLP-like capabilities

# Outline

- Loading
- Retrieval
- Indexing
- Searching
- Tokens
- Stopwords
- Results Markup
- Results Highlighting
- Ontologies/Thesauri
- Knowledgebase
- Themes
- Clustering
- Classification
- Demo

# Loading

- Storage choice
  - ◆ XML DB
  - ◆ CLOB
- Loading method
  - ◆ SQL\*Loader
  - ◆ INSERT
  - ◆ UPSERT
- Considerations
  - ◆ Initial Loading
  - ◆ Periodic Updates

# Loading

```
create table medtab (  
    PMID number primary key,  
    text xmltype );
```

```
Load DATA  
  INFILE 'medline.dat'  
  BADFILE 'medline.bad'  
  DISCARDFILE 'medline.discard'  
  INTO TABLE medtab  
  REPLACE  
  FIELDS TERMINATED BY '\t'  
  (pmid, text char(1000000))
```

```
select PMID,  
       extractValue(text,  
 ' /MedlineCitation/Article/Abstract/AbstractText') Abstract  
from medtab  
where PMID='15129431';
```

# Retrieval

```
select pmid,  
       extractValue(text, '/MedlineCitation/Article/ArticleTitle') Title,  
       extractValue(text, '/MedlineCitation/Article/Abstract/AbstractText') Abstract  
from medtab  
where PMID='3298569';
```

3298569

Estrogen receptor immunoreactivity in meningiomas. Comparison with the binding activity of estrogen, progesterone, and androgen receptors.

Estrogen receptor (ER) analysis was performed in 70 meningioma samples by means of two assays: an enzyme immunoassay that used monoclonal antibodies against human ER protein (estrophilin), and a sensitive radioligand binding assay that used iodine-125-labeled estradiol as the radioligand. Low levels of ER immunoreactivity were found in tumors from 51% of patients, whereas ER binding activity was demonstrated in 40% of the meningiomas examined. In eight (11%) of the tissue samples, multiple binding sites for estradiol were observed. The immunoreactive binding sites corresponded to those of the classic high-affinity ER. In ligand binding studies, however, measurement of classic ER was...

# Retrieval

```
select extract(text,  
  '/MedlineCitation/MeshHeadingList/MeshHeading/DescriptorName').getStringVal()  
from medtab  
where PMID='3298569';
```

```
<DescriptorName MajorTopicYN="N">Comparative Study</DescriptorName>  
<DescriptorName MajorTopicYN="N">Female</DescriptorName>  
<DescriptorName MajorTopicYN="N">Human</DescriptorName>  
<DescriptorName MajorTopicYN="N">Immunoenzyme Techniques</DescriptorName>  
<DescriptorName MajorTopicYN="N">Male</DescriptorName>  
<DescriptorName MajorTopicYN="N">Meningeal Neoplasms</DescriptorName>  
<DescriptorName MajorTopicYN="N">Meningioma</DescriptorName>  
<DescriptorName MajorTopicYN="N">Middle Aged</DescriptorName>  
<DescriptorName MajorTopicYN="N">Radioligand Assay</DescriptorName>  
<DescriptorName MajorTopicYN="N">Receptors, Androgen</DescriptorName>  
<DescriptorName MajorTopicYN="N">Receptors, Estrogen</DescriptorName>  
<DescriptorName MajorTopicYN="N">Receptors, Progesterone</DescriptorName>
```

# Retrieval

```
select extract(text,  
  '/MedlineCitation/ChemicalList/Chemical/NameOfSubstance').getStringVal()  
from medtab  
where PMID='3298569';
```

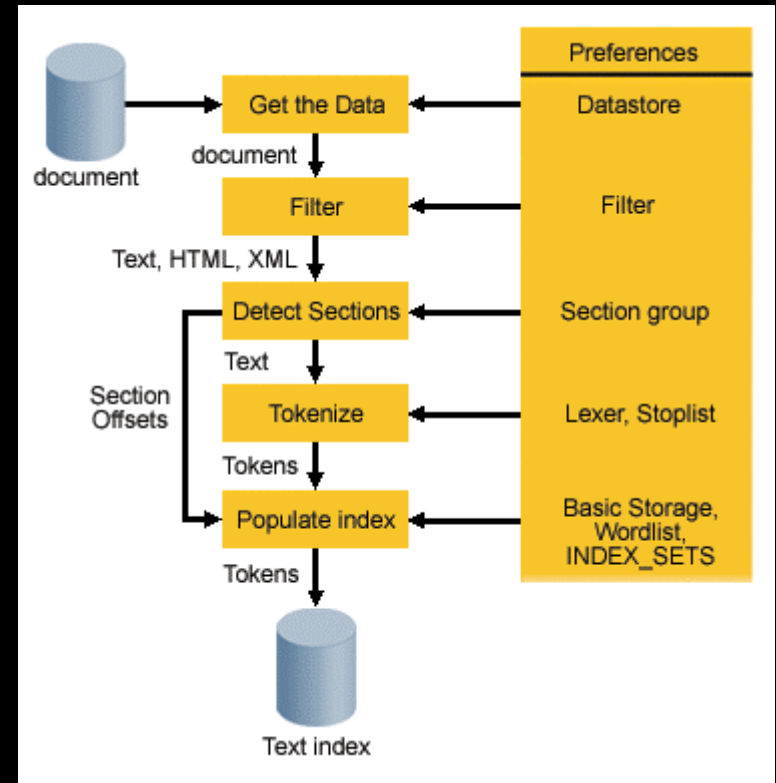
```
<NameOfSubstance>Receptors, Androgen</NameOfSubstance>
```

```
<NameOfSubstance>Receptors, Estrogen</NameOfSubstance>
```

```
<NameOfSubstance>Receptors, Progesterone</NameOfSubstance>
```

# Indexing

- Filter
  - ◆ File formats
- Sectioner
  - ◆ HTML
  - ◆ XML
- Tokenizer
  - ◆ Lexer - tokenize
  - ◆ Stoplists - mask



# Indexing

- Index Types
  - ◆ CONTEXT
    - Text retrieval
    - CONTAINS query operator
  - ◆ CTXCAT
    - Item categories
    - CATSEARCH query operator
  - ◆ CTXRULE
    - Classification rules
    - MATCHES query operator

# Indexing

```
-- Enable theme indexing
exec ctx_ddl.create_preference('mylex','BASIC_LEXER');
exec ctx_ddl.set_attribute('mylex','MIXED_CASE','NO');
exec ctx_ddl.set_attribute('mylex','THEME_LANGUAGE','ENGLISH');
exec ctx_ddl.set_attribute('mylex','index_themes','YES');
exec ctx_ddl.set_attribute('mylex','index_text','YES');

-- Create XML sections
exec ctx_ddl.create_section_group('xmlgroup','auto_section_group');

-- Index column 'text' of table 'medtab' for user 'dmuser'
create index medtab_idx on medtab(text)
indextype is ctxsys.context
parameters('lexer mylex filter ctxsys.null_filter section group
xmlgroup');
```

# Tokens

```
declare
  the_tokens ctx_doc.token_tab;
begin
  ctx_doc.tokens('medtab_idx','9115210',the_tokens);
  for i in 1..the_tokens.count loop
    dbms_output.put_line(the_tokens(i).token || ' ' || the_tokens(i).offset);
  end loop;
end;
```

```
...
MUTATIONS 176
TUMOR 193
SUPPRESSOR 199
GENE 210
PATCHED 215
PTC 224
FOUND 233
HUMAN 242
PATIENTS 248
BASAL 266
CELL 272
NEVUS 277
SYNDROME 283
DISEASE 295
...
```

**Basal cell carcinomas in mice overexpressing sonic hedgehog.**

**Mutations in the tumor suppressor gene PATCHED (PTC) are found in human patients with the basal cell nevus syndrome, a disease** causing developmental defects and tumors, including basal cell carcinomas. Gene regulatory relationships defined in the fruit fly *Drosophila* suggest that overproduction of Sonic hedgehog (SHH), the ligand for PTC, will mimic loss of ptc function. It is shown here that transgenic mice overexpressing SHH in the skin develop many features of basal cell nevus syndrome, demonstrating that SHH is sufficient to induce basal cell carcinomas in mice. These data suggest that SHH may have a role in human tumorigenesis.

# Stop Words

## ◆ CTX\_REPORT.INDEX\_STATS

Query: liver cancer  
No.docs: 9, No.tokens: 702

| <b>Token</b> | <b>TF</b> | <b>DF</b> | <b>TFIDF</b> |
|--------------|-----------|-----------|--------------|
| DNA          | 22        | 2         | 11.00        |
| TISSUE       | 30        | 3         | 10.00        |
| CELLS        | 37        | 5         | 7.40         |
| TUMOR        | 29        | 4         | 7.25         |
| AR           | 40        | 6         | 6.67         |
| ENZYME       | 12        | 2         | 6.00         |
| <b>LIVER</b> | 50        | <b>9</b>  | 5.56         |
| PROTEIN      | 11        | 2         | 5.50         |
| DIETS        | 11        | 2         | 5.50         |

## ◆ Add stopwords to text index

```
ALTER INDEX myindex REBUILD PARAMETERS ('ADD STOPWORD 1');  
ALTER INDEX myindex REBUILD PARAMETERS ('ADD STOPWORD CANCER');  
ALTER INDEX myindex REBUILD PARAMETERS ('ADD STOPWORD LIVER');
```

# Searching

```
COL Title FORMAT a60;
COL S FORMAT 999;
select score(1) s, pmid,
       extractValue(text, '/MedlineCitation/Article/ArticleTitle') Title
from medtab
where CONTAINS(text, 'aldose reductase WITHIN AbstractText', 1) > 0
ORDER BY score(1) DESC;
```

| S  | PMID     | TITLE   |
|----|----------|---|
| 59 | 14768008 | Detection and identification of tumor-associated protein variants in human hepatocellular carcinomas.   |
| 12 | 9537432  | New member of aldose reductase family proteins overexpressed in human hepatocellular carcinoma.   |
| 12 | 10322639 | Developmental expression of urine concentration-associated genes and their altered expression in murine infantile-type polycystic kidney disease. |
| 12 | 9565553  | Identification and characterization of a novel human aldose reductase-like gene.  |
| 12 | 11261885 | Overexpression of aldose reductase in liver cancers may contribute to drug resistance.  |

# Searching

```
select score(1) s, pmid,  
  extractValue(text, '/MedlineCitation/Article/ArticleTitle') Title  
from medtab  
where contains(text,  
  '<query><textquery>aldose reductase WITHIN AbstractText</textquery>  
  <score algorithm="COUNT"/></query>', 1) > 0;
```

```
S          PMID TITLE  
-----
```

```
5    14768008  Detection and identification of tumor-associated protein var  
          iants in human hepatocellular carcinomas.
```

```
1    9537432  New member of aldose reductase family proteins overexpressed  
          in human hepatocellular carcinoma.
```

```
1    10322639  Developmental expression of urine concentration-associated g  
          enes and their altered expression in murine infantile-type p  
          olycystic kidney disease.
```

```
1    9565553  Identification and characterization of a novel human aldose  
          reductase-like gene.
```

```
1    11261885  Overexpression of aldose reductase in liver cancers may cont  
          ribute to drug resistance.
```

# Results Markup

```
DECLARE
  mklob clob;
  dx number := 80;
  line varchar2(80);
  length number;
BEGIN
  ctx_doc.markup('medtab_idx', '9837785', 'apoptosis', mklob, FALSE,
    'HTML_DEFAULT', '<span id="mkp">', '</span>');
  length := floor(dbms_lob.getlength(mklob) / dx);
  for i in 0..length loop
    dbms_lob.read(mklob, dx, i*dx+1, line);
    dbms_output.put_line(line);
  end loop;
END;
```

<AbstractText>The marked increases in p53 and p21/WAF1 levels that occur during Epstein-Barr virus (EBV) infection and the generation of immortal B lymphoblastoid cell lines (LCL) do not lead to growth arrest or **apoptosis**, although increasing wild-type (wt) p53 levels in EBV-infected cells by transfection or DNA damage induce these effects. We hypothesized that the concentration of p53 relative to that of LMP1 determines whether EBV-infected B cells undergo growth arrest and **apoptosis**. Cell cycle arrest and **apoptosis** were evaluated in LCL expressing varying p53 levels achieved by treating the cells with increasing concentrations of cisplatin, and we supplemented this approach with experiments in EBV-infected Burkitt's lymphoma (BL) cells transfected with a temperature-sensitive (ts) mutant human p53 and studies in LCL infected with recombinant adenoviruses expressing wt and ts mutant p53. Small increases in p53 and p21/WAF1 led to cell cycle arrest at the G2/M boundary, but not to **apoptosis**; moderate increases resulted in growth arrest at the G1/S boundary, also without **apoptosis**; and large increases also induced **apoptosis**. These results confirm the hypothesis and reveal unanticipated complexities in cell cycle regulation by p53.</AbstractText>

# Results Highlighting

```
create table hightab(query_id number, offset number, length number);
BEGIN
  ctx_doc.highlight('medtab_idx', '15088555',
    'NEAR((thalidomide,inhibit%,angiogenesis),10,TRUE) WITHIN AbstractText',
    'hightab', 0, FALSE);
END;
```

```
select * from hightab;
```

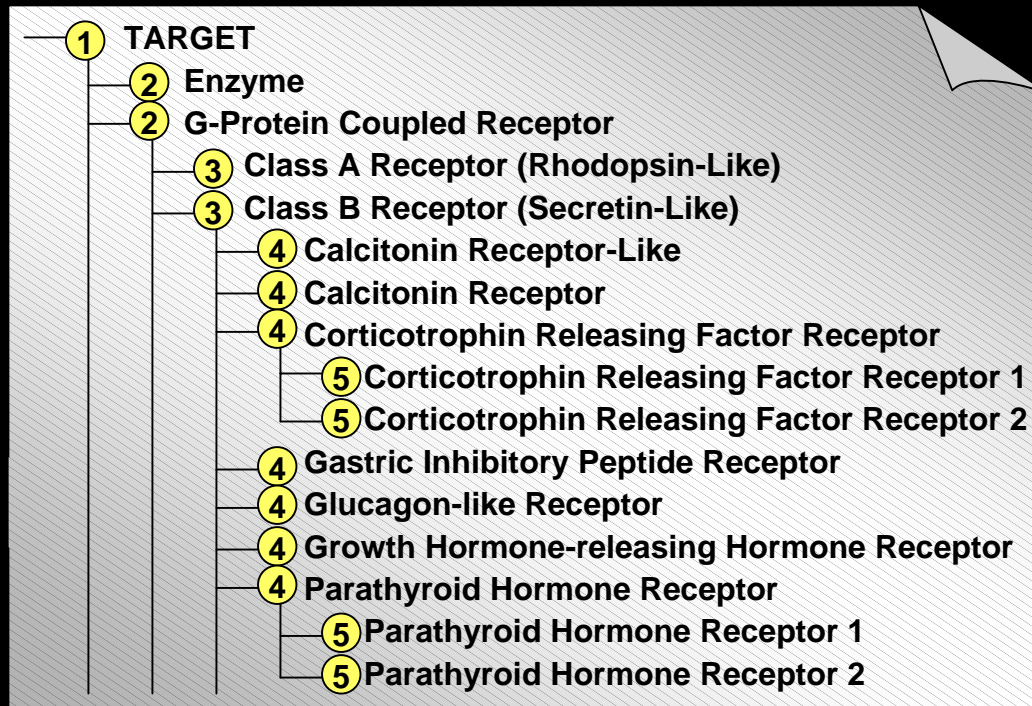
| QUERY_ID | OFFSET | LENGTH |
|----------|--------|--------|
| 0        | 769    | 11     |
| 0        | 781    | 8      |
| 0        | 798    | 12     |

```
select SUBSTR(text,769,798+12-769) txt from medtab where PMID=15088555;
```

```
TXT
```

```
-----  
Thalidomide inhibits tumoral angiogenesis
```

# Ontologies



**Ontologies → Thesauri → Knowledge Base → Themes**

# Thesauri

```
begin
  ctx_thes.create_thesaurus('TARGETS',FALSE);
end;
```

```
Begin
  ctx_thes.create_phrase('TARGETS','Target');
  ctx_thes.create_phrase('TARGETS','Enzyme');
  ctx_thes.create_phrase('TARGETS','G-Protein Coupled Receptor');
end;
```

```
begin
  ctx_thes.create_phrase('TARGETS','Class A Receptor (Rhodopsin-Like)');
  ctx_thes.create_phrase('TARGETS','Class B Receptor (Secretin-Like)');
  ctx_thes.create_phrase('TARGETS','Calcitonin Receptor-Like');
  ctx_thes.create_phrase('TARGETS','Calcitonin Receptor');
  ctx_thes.create_phrase('TARGETS','Corticotrophin Releasing Factor Receptor');
  ctx_thes.create_phrase('TARGETS','Corticotrophin Releasing Factor Receptor 1');
  ctx_thes.create_phrase('TARGETS','Corticotrophin Releasing Factor Receptor 2');
  ctx_thes.create_phrase('TARGETS','Gastric Inhibitory Peptide Receptor');
  ctx_thes.create_phrase('TARGETS','Glucagon-like Receptor');
end;
```

# Thesauri

```
begin
  ctx_thes.create_relation('TARGETS','Target','NT','Enzyme');
  ctx_thes.create_relation('TARGETS','Target','NT','G-Protein Coupled Receptor');
  ctx_thes.create_relation('TARGETS','G-Protein Coupled Receptor','NT',
                          'Class A Receptor (Rhodopsin-Like)');
  ...
end;
```

```
declare
  synonyms varchar2(2000);
begin
  synonyms := ctx_thes.nt('G-Protein Coupled Receptor',2,'TARGETS');
  dbms_output.put_line('Thesaurus: TARGETS');
  dbms_output.put_line('The synonym expansion for "G-Protein Coupled Receptor"
is: ' || synonyms);
end;
```

Thesaurus: TARGETS

The synonym expansion for "G-Protein Coupled Receptor" is: {G-Protein Coupled Receptor}|{Class A Receptor (Rhodopsin-Like)}|{Class B Receptor (Secretin-Like)}|{Calcitonin Receptor-Like}|{'Calcitonin Receptor'}|{Corticotrophin Releasing Factor Receptor}|{Gastric Inhibitory Peptide Receptor}|{Glucagon-like Receptor}|{...

# Knowledgebase

- Six main branches of the knowledge base:
  - ◆ Branch 1: science and technology
  - ◆ Branch 2: business and economics
  - ◆ Branch 3: government and military
  - ◆ Branch 4: social environment
  - ◆ Branch 5: geography
  - ◆ Branch 6: abstract ideas and concepts
- <http://www.oracle.com/technology/products/text/htdocs/aknow.htm>

# Themes

```
declare
  the_themes ctx_doc.theme_tab;
begin
  ctx_doc.themes('medtab_idx','9115210',the_themes,FALSE,10);
  for i in 1..the_themes.count loop
    dbms_output.put_line(the_themes(i).theme || ' : ' || the_themes(i).weight);
  end loop;
end;
```

```
metabolisms:33
genetics:33
proteins:27
pathology:26
genes:25
United States:23
humankind:20
basal cells:19
SHH:18
suggestion:17
```

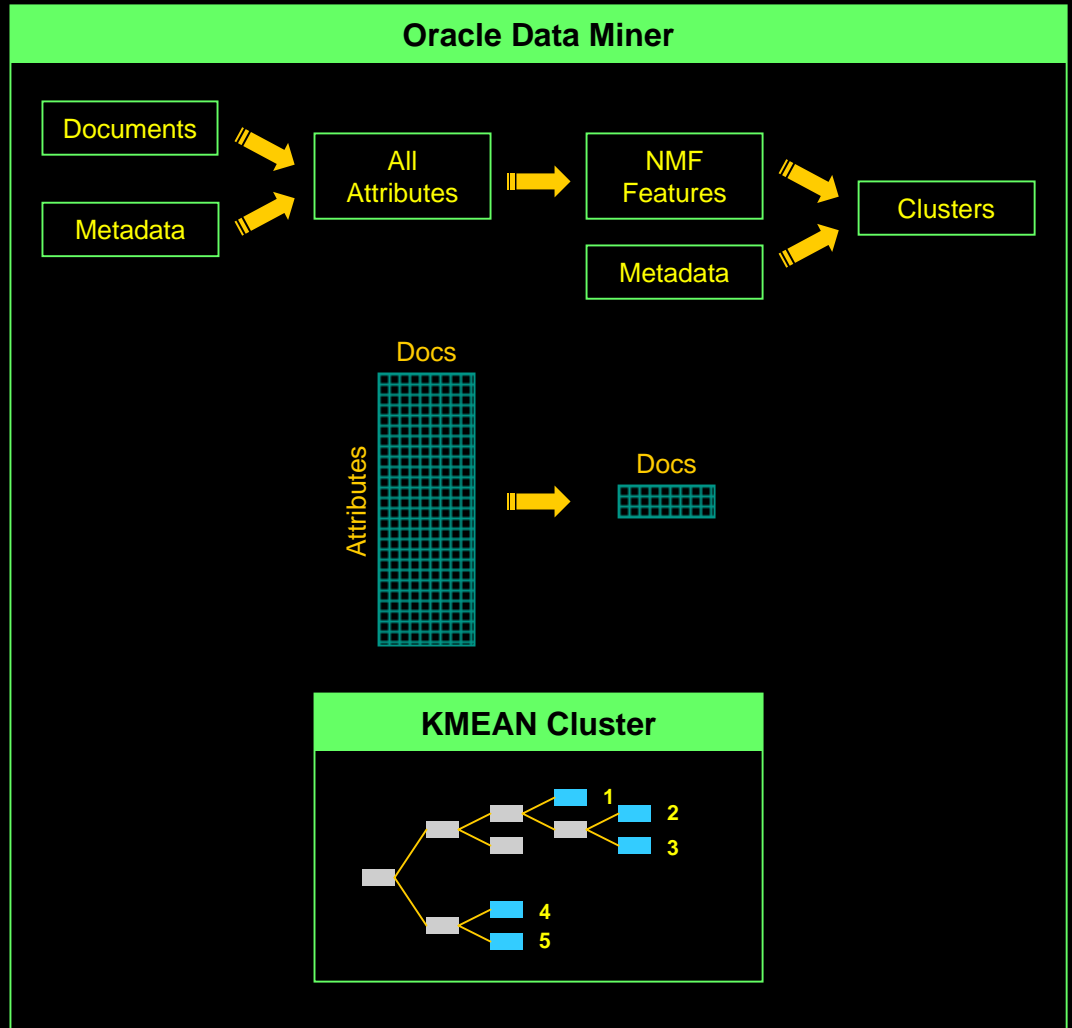
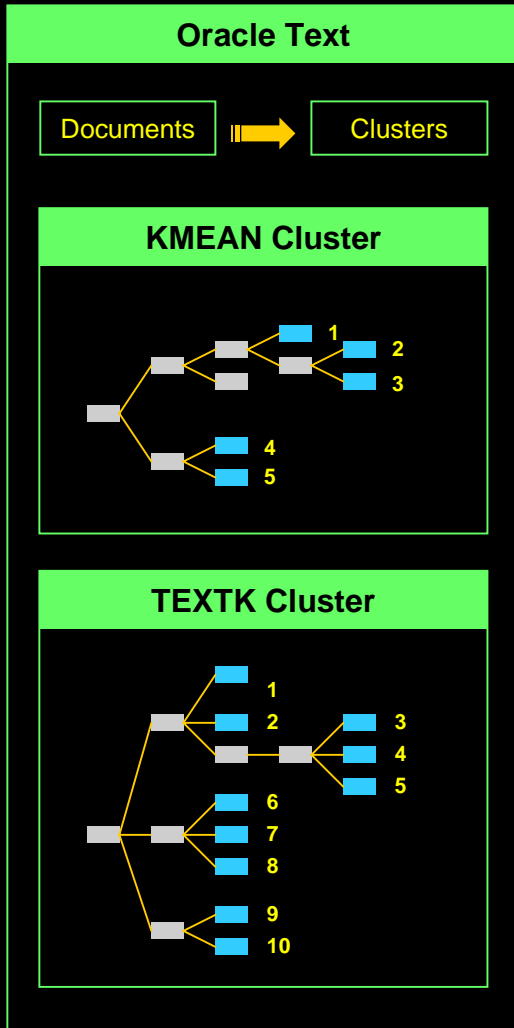
## Basal cell carcinomas in mice overexpressing sonic hedgehog.

Mutations in the tumor suppressor gene PATCHED (PTC) are found in human patients with the basal cell nevus syndrome, a disease causing developmental defects and tumors, including basal cell carcinomas. Gene regulatory relationships defined in the fruit fly Drosophila suggest that overproduction of Sonic hedgehog (SHH), the ligand for PTC, will mimic loss of ptc function. It is shown here that transgenic mice overexpressing SHH in the skin develop many features of basal cell nevus syndrome, demonstrating that SHH is sufficient to induce basal cell carcinomas in mice. These data suggest that SHH may have a role in human tumorigenesis.

# Document Clustering

- Unsupervised Classification
  - ◆ No training needed
  - ◆ Good for initial overview of a group of documents
  - ◆ Identifies shared attributes
- CTX\_CLS.CLUSTERING
  - ◆ KMEAN
    - requires setting the number of clusters
  - ◆ TEXTK
    - experimental hierarchical clustering
- ODM NMF feature k-Mean clustering

# Document Clustering



# Document Clustering

- Prepare database objects
  - ◆ Collection table
  - ◆ Clusters table
  - ◆ Document results table
- Set clustering preferences
- Populate and index collection table
- Run clustering

# Clustering

- ◆ Collection table

```
create table collection (id number primary key, text clob);
```

- ◆ Clusters table

```
create table clusters (  
    clusterid NUMBER,  
    descript varchar2(4000),  
    label varchar2(200),  
    sze NUMBER,  
    quality_score NUMBER,  
    parent NUMBER);
```

- ◆ Document results table

```
create table restab (  
    docid NUMBER,  
    clusterid NUMBER,  
    score NIMBER);
```

# Clustering

## ◆ Index collection table

```
create index collectionx on collection(text)
  indextype is ctxsys.context
  parameters('STOPLIST CTXSYS.DEFAULT_STOPLIST');
```

## ◆ Examine token TF and DF

```
Declare
  x clob := null;
Begin
  ctx_report.index_stats('collectionx',x);
  insert into collection_stats values (x);
  commit;
  dbms_lob.freetemporary(x);
end;

Select * from collection_stats;
```

# Clustering

- ◆ Set clustering preferences for k-Means

```
BEGIN
  ctx_ddl.drop_preference('my_cluster');
  ctx_ddl.create_preference('my_cluster','KMEAN_CLUSTERING');
  ctx_ddl.set_attribute('my_cluster','CLUSTER_NUM',5);
  ctx_ddl.set_attribute('my_cluster','MAX_FEATURES',200);
  ctx_ddl.set_attribute('my_cluster','MAX_DOCTERMS',20);
END;
```

- ◆ Cluster collection

```
BEGIN
  ctx_cls.clustering('collectionx','id','restab','clusters','my_cluster');
END;
```

# Document Classification

- Supervised Classification
  - ◆ Needs training
  - ◆ Can be applied to any document
- Rule-based
  - ◆ Manual rule creation
- Decision Trees
  - ◆ Automatic rule creation (editable)
- SVM
  - ◆ Automatic rule creation (opaque)

# SVM Classification

- Training Documents
  - ◆ Create and populate training document table
  - ◆ Generate CONTEXT index on documents
- Categories
  - ◆ Assign documents to categories
- Set classifier preferences
  - ◆ MAX\_FEATURES
- Train Classifier
  - ◆ Create Rules Table
  - ◆ Train
  - ◆ Generate a CTXRULE index on the rules table
- Classify New Documents

# SVM Classification

- ◆ Create and populate training documents table

```
create table svmtrain (  
    docid number primary key,  
    text clob);
```

- ◆ Create training document index

```
create index svmtrainx on svmtrain(text)  
    indextype is ctxsys.context  
    parameters('STOPLIST CTXSYS.DEFAULT_STOPLIST');
```

# SVM Classification

- ◆ Create and populate the category table

```
create table svmcats (  
    docid    number,  
    cat_id   number,  
    catname  varchar2(250));
```

- ◆ Create the rules table

```
create table svmtab (  
    cat_id   number,  
    type     number(3) NOT NULL,  
    rule     blob);
```

# SVM Classification

- ◆ Set SVM classifier preferences

```
begin
  ctx_ddl.set_attribute('mysvm', 'MAX_FEATURES', '100');
end;
```

- ◆ Train SVM classifier

```
begin
  ctx_cls.train('svmtrainx',
               'docid',
               'svmcats',
               'docid',
               'cat_id',
               'svmtab',
               'mysvm');
end;
```

# SVM Classification

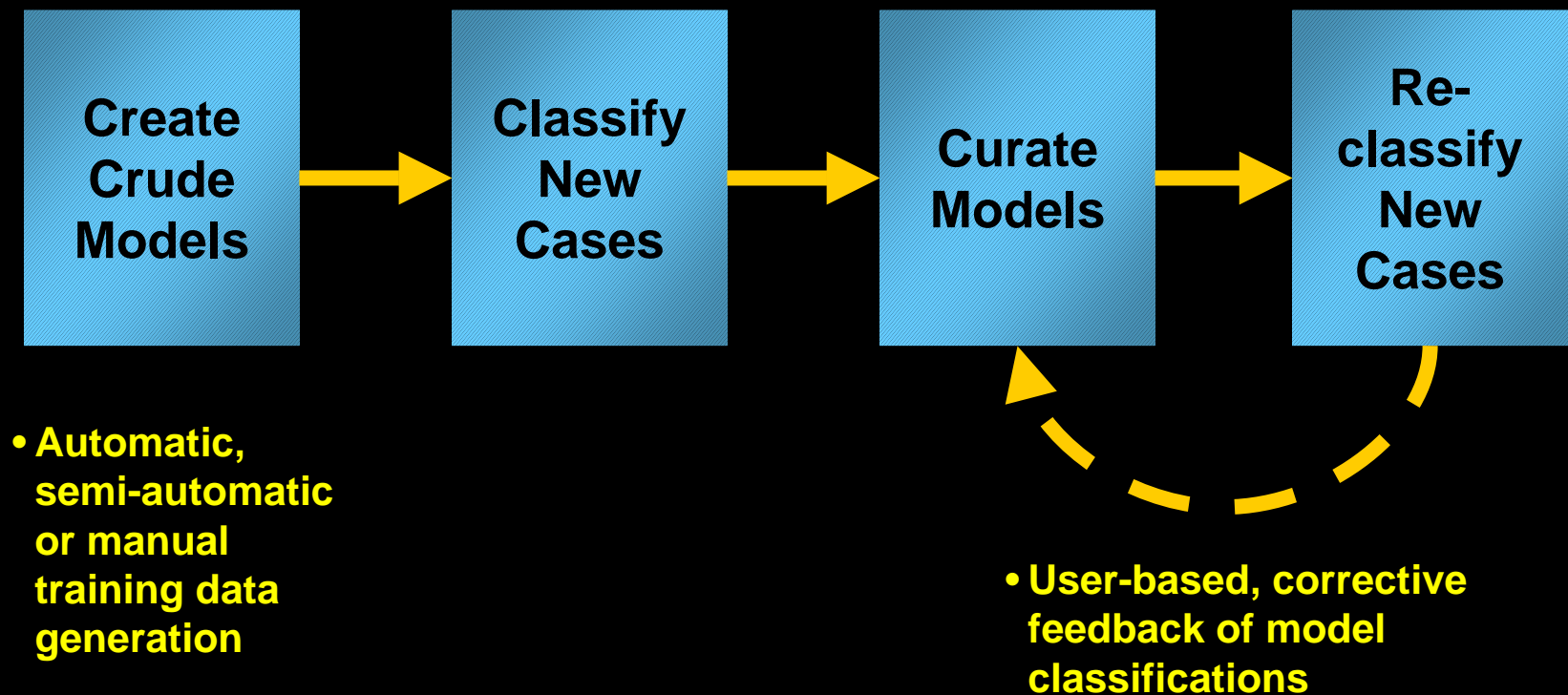
- ◆ Create rules index

```
create index svmx on svmtab(rule)
  indextype is ctxsys.ctxrule
  parameters ('filter svmfilter classifier mysvm');
```

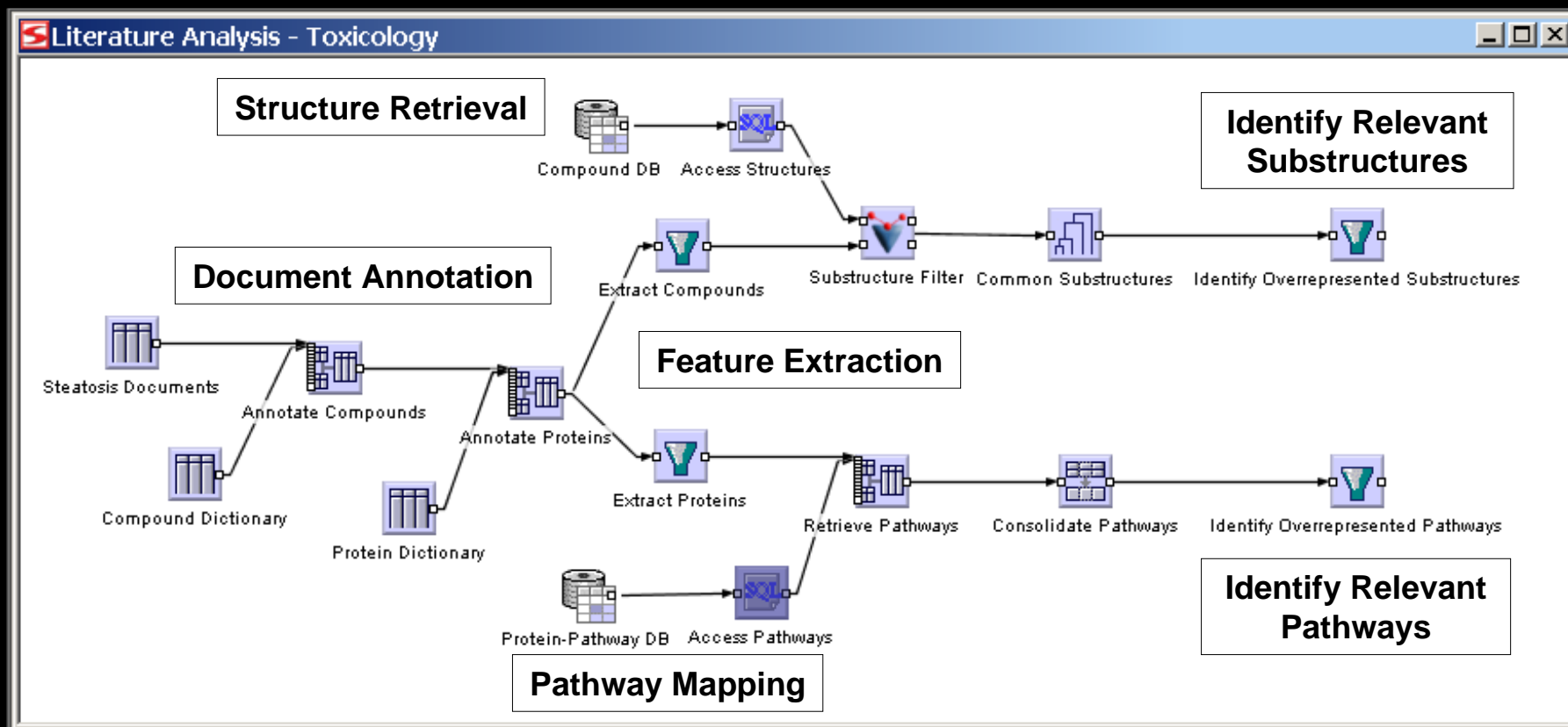
- ◆ Classify unknown documents

```
select cat_id, match_score(1) SCORE
from svmtab
where matches(rule, (
  select extractValue(text, '/MedlineCitation/Article/ArticleTitle')||' '||
    extractValue(text, /MedlineCitation/Article/Abstract/AbstractText')
  from medtab
  where pmid='7587903'), 1) > 50;
```

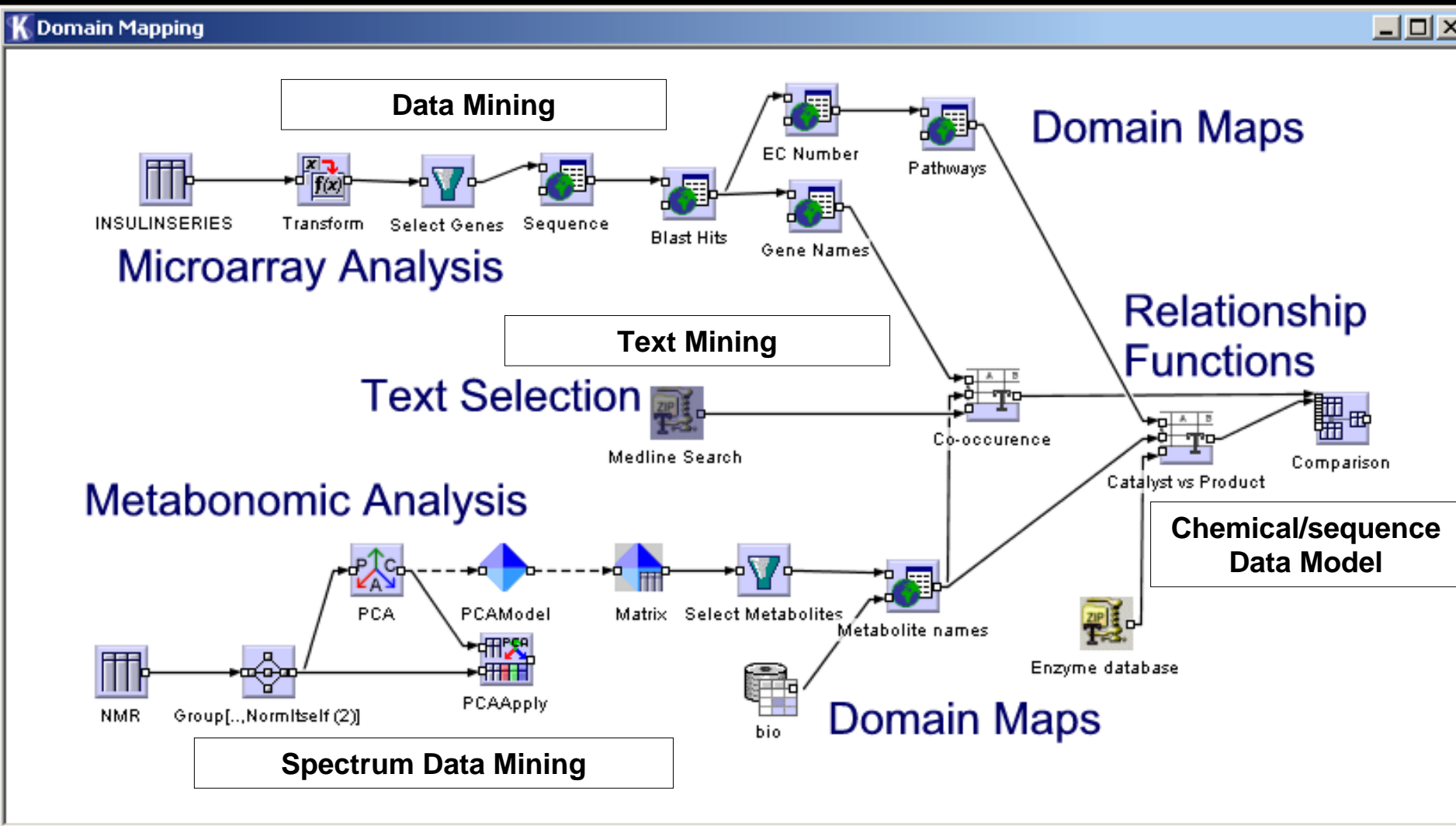
# Classification Model Curation



# Toxicology Literature Analysis



# Cross-domain Data Analysis



*DEMO*



Q U E S T I O N S  
A N S W E R S

ORACLE®