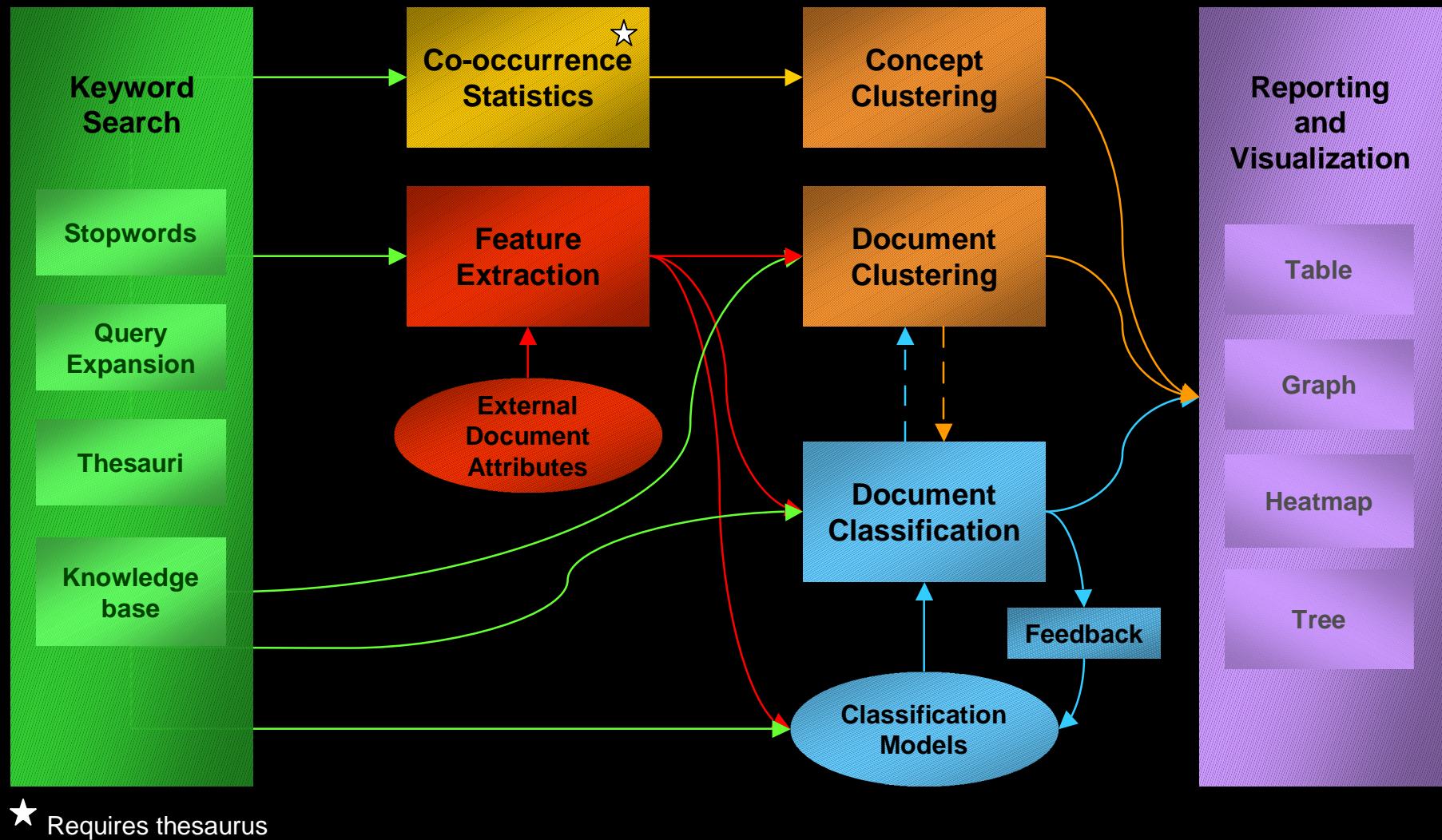


# Oracle Text Mining Demo

## Example Workflow

ORACLE®

# Oracle Text Mining Application



ORACLE®

# Text Mining Tasks

## *Search and Retrieval*

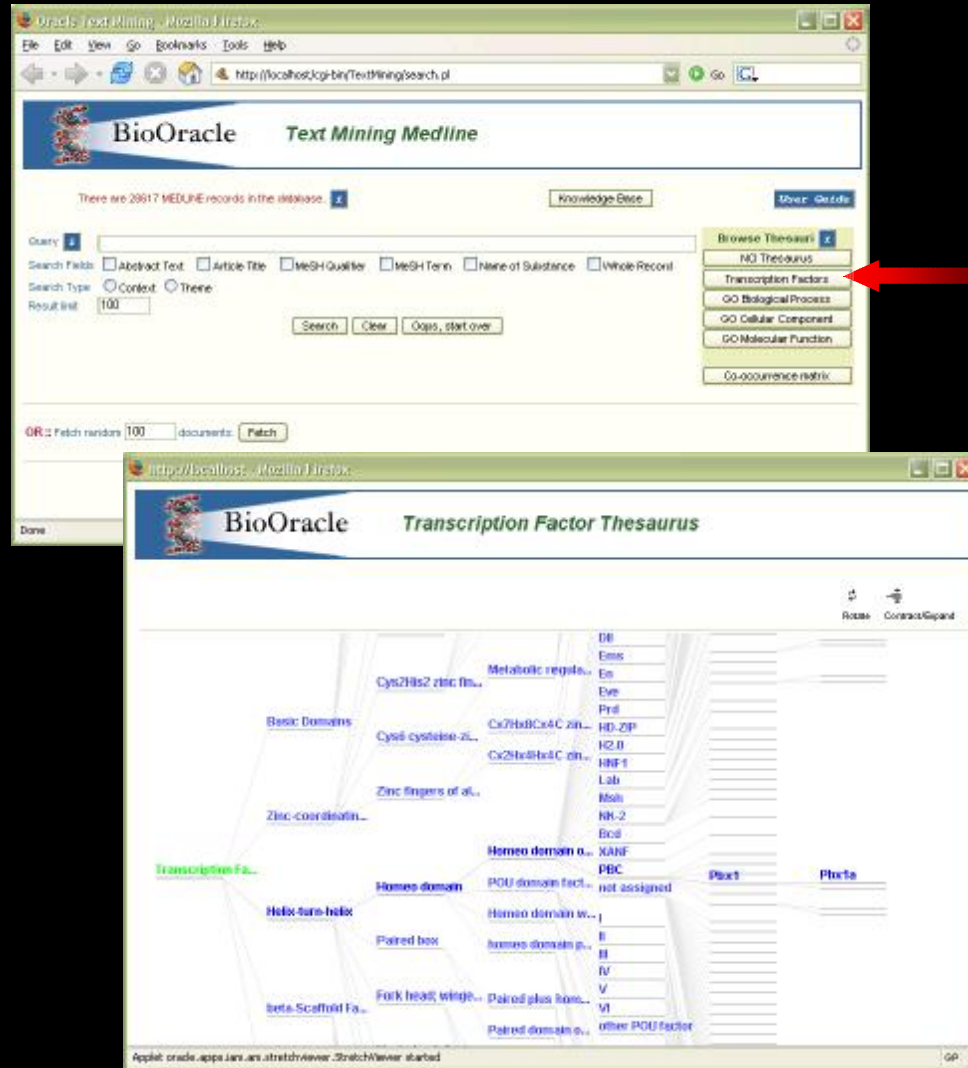
### Search

- keyword, phrase or thesaurus query (context or theme)
  - CONTAINS(text, 'query')
  - CONTAINS(text, 'ABOUT(query)')
- random document fetch (for clustering or classification)
  - dbms\_random.value

### Retrieve

- fetch, highlight and display document and document themes
- ctx\_doc.markup to add HTML style tags to query matches
- ctx\_doc.themes to extract document themes
- ctx\_thes.nt to display expended query terms for a thesaurus query

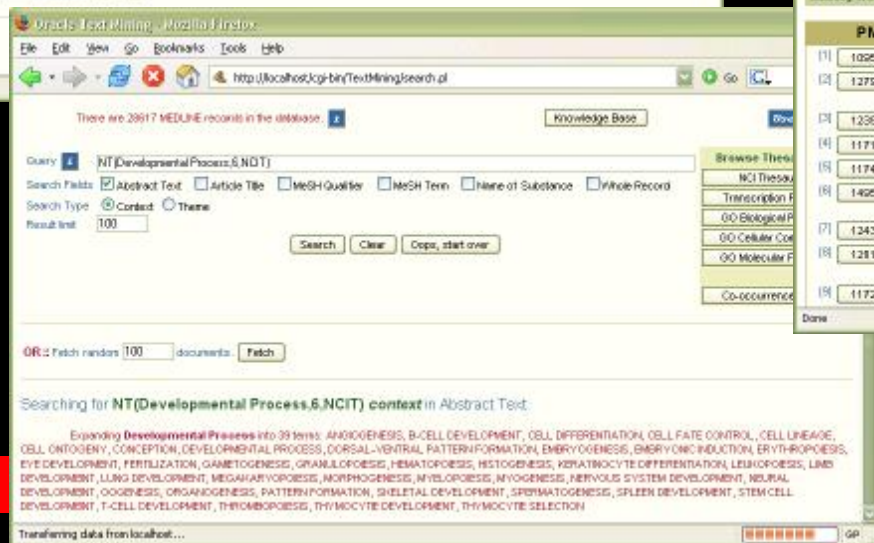
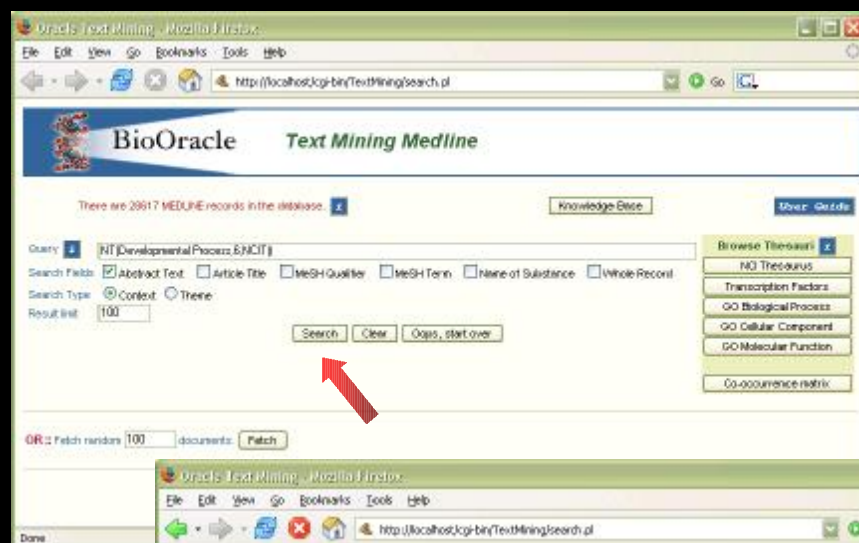
— Open <http://localhost/cgi-bin/TextMining/search.pl>



Click on “Transcription Factors” thesaurus link to display stretchviewer of transcription factor genes

ORACLE

- Type “NT(Developmental Process,6,NCIT)” in the Query window
- Select the “Abstract Text” and “Context” checkbox options
- Click on the “Search” button



ORACLE®



- Click on the highest scoring PMID button to see full document

The screenshot displays the Oracle Text Mining application interface, which is divided into two main windows.

**Left Window: Search Results**

The top section shows "Found 100 matching records". Below this, there are configuration options for "Document Clustering" and "Feature Extraction for full Data Mining". The "Document Clustering" section includes settings for "No. Clusters" (3), "Max Distinct Terms per Doc" (10), "Use Tokens as Features" (checked), and "Use Themes as Features" (unchecked). The "Feature Extraction" section has a "NMF Features" button and a "Data Source" dropdown set to "Thesauri".

The "SVM classification" section includes a "Classify with Special Models" button and a "SVM Model Feedback" button.

The main results table lists 10 records with columns for "PMID", "Score", and "Contents". A red arrow points to the first record, which has a score of 90 and a PMID of 10255791. The "Contents" column for this record reads: "Prognostic value of bone marrow **angiogenesis** in multiple myeloma."

**Right Window: Document View**

The right window displays the full document for the selected PMID (10255791). The title is "Prognostic value of bone marrow **angiogenesis** in multiple myeloma." The document is from "Clin Cancer Res 8 (8), Aug 2000". The abstract text is as follows:

We studied the prognostic value of **angiogenesis** grading and microvessel density estimation in newly diagnosed multiple myeloma. Seventy-five patients with newly diagnosed myeloma, treated on Eastern Cooperative Oncology Protocol E9488 and Intergroup study 0141 (S9321) at the Mayo Clinic, were studied. Bone marrow microvessels were examined using immunohistochemical staining for von Willebrand factor. Determination of microvessel density and **angiogenesis** grading was done in a blinded manner. There was a strong correlation between microvessel density and the plasma cell labeling index,  $\rho = 0.42$ ,  $P < 0.001$ . **Angiogenesis** grade was also significantly associated with the plasma cell labeling index. Fifteen % of patients with low-grade **angiogenesis** had a high labeling index ( $> 1\%$ ). In contrast, 47% of patients with intermediate or high-grade **angiogenesis** had high labeling indices ( $P = 0.02$ ). Overall survival was significantly different among those with high-, intermediate-, and low-grade **angiogenesis**, with median times of 2.4, and 4.4 years, respectively ( $P = 0.02$ ). Similarly, patients with microvessel density  $> 50 \times 400$  field had poorer survival compared with those with 50 or fewer microvessels/field, median survival 2.6 versus 5.1 years, respectively ( $P = 0.004$ ). There was a strong association between **angiogenesis** grade and microvessel density ( $P < 0.001$ ). We conclude that bone marrow **angiogenesis** is a predictor of poor survival in newly diagnosed myeloma. **Angiogenesis** is correlated with the plasma cell labeling index but not the bone marrow

ORACLE

# Text Mining Tasks

## *Clustering*

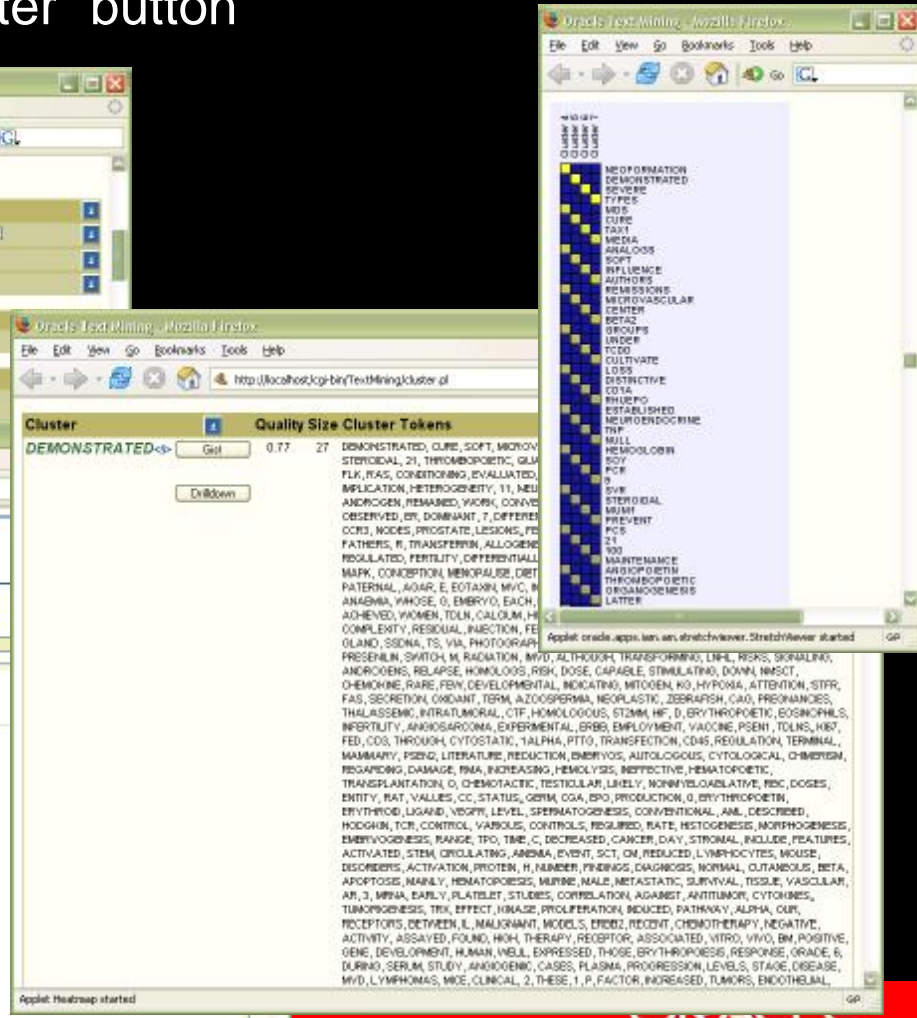
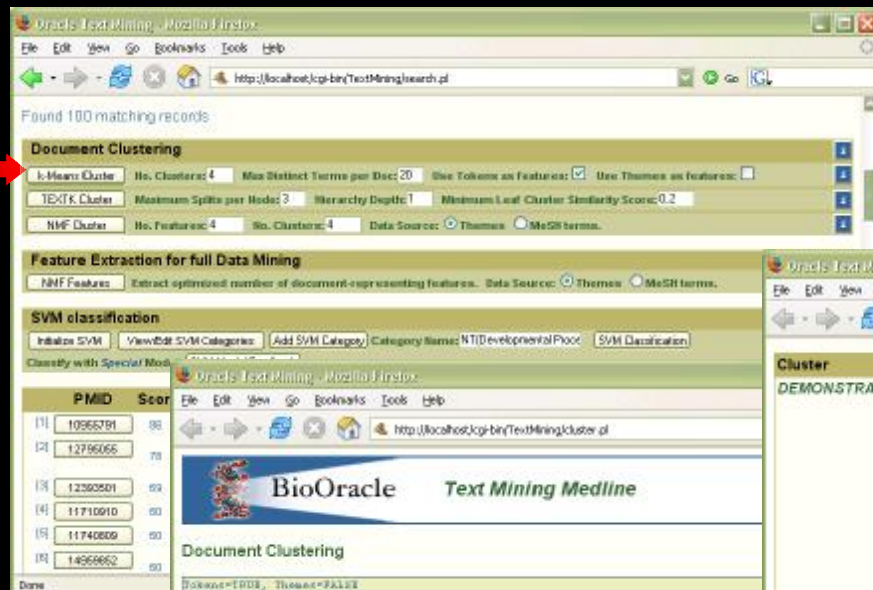
### Oracle Text

- K-Means or TEXTK clustering
- operates directly on document text
- index tokens and/or themes to generate document attributes for clustering
- display clusters and documents in stretchviewer applet with hotlinks to fetch individual documents
  - fetched document highlights cluster definition terms found in the text
- generate cluster gist and themes
  - concatenate all cluster documents, index with Oracle Text, extract gist (ctx\_doc.gist) and themes (ctx\_doc.themes)
- launch cluster drilldown (sub operations on a selected cluster)

ORACLE

# Oracle Text K-Means Clustering

- Set “No. Clusters” to 4 and “Max Distinct Terms per Doc” to 20
- Click on the “K-Means Cluster” button





- Double-click on a document PMID in the stretchviewer to see full document with cluster definition terms highlighted



# Text Mining Tasks

## *Clustering drilldown*

### **Filter list of documents**

- keyword search within selected group of documents (from parent cluster)

### **Weigh document terms**

- assign weights (0-100) to selected terms in remaining documents
  - replicate each instance of each weighted term x times ( $x = \text{term weight}$ )

### **Cluster**

- K-Means cluster of remaining, term-weighted documents

### **Classify**

- Classify remaining, term-weighted documents with existing SVM models
  - straight classification score or model feedback option
- Use remaining, term-weighted documents to create new SVM model category

# Oracle Text K-Means Clustering

- Click on the “Drilldown” button of the “Neoformation” cluster
- Enter “myeloma%” in the subsearch window and click “Search” to reduce the document set from 31 to 22

The image displays three sequential screenshots of the Oracle Text Mining Medline application, illustrating the process of refining a document cluster.

**Top Screenshot:** The application window shows the "Neoformation" cluster selected. A red arrow points to the "Drilldown" button. The document list on the right contains 31 documents.

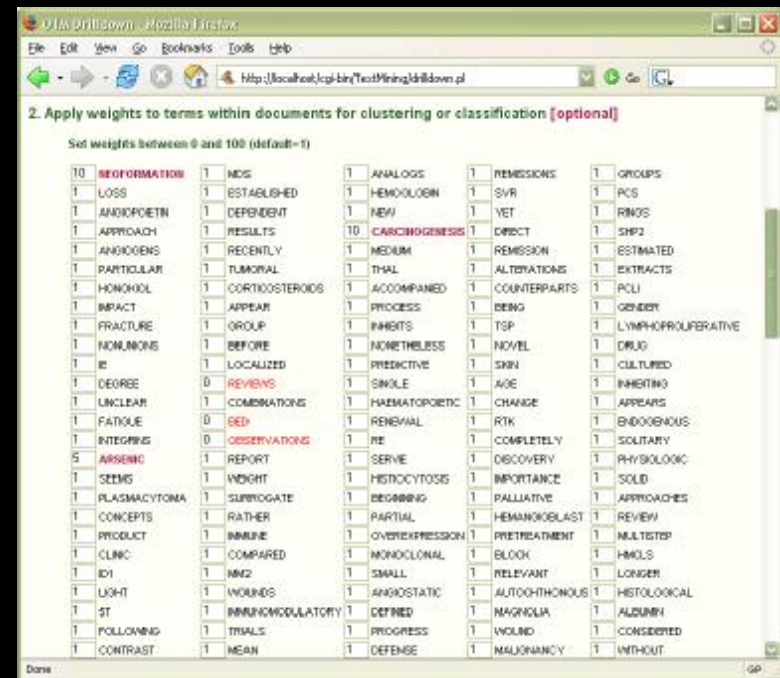
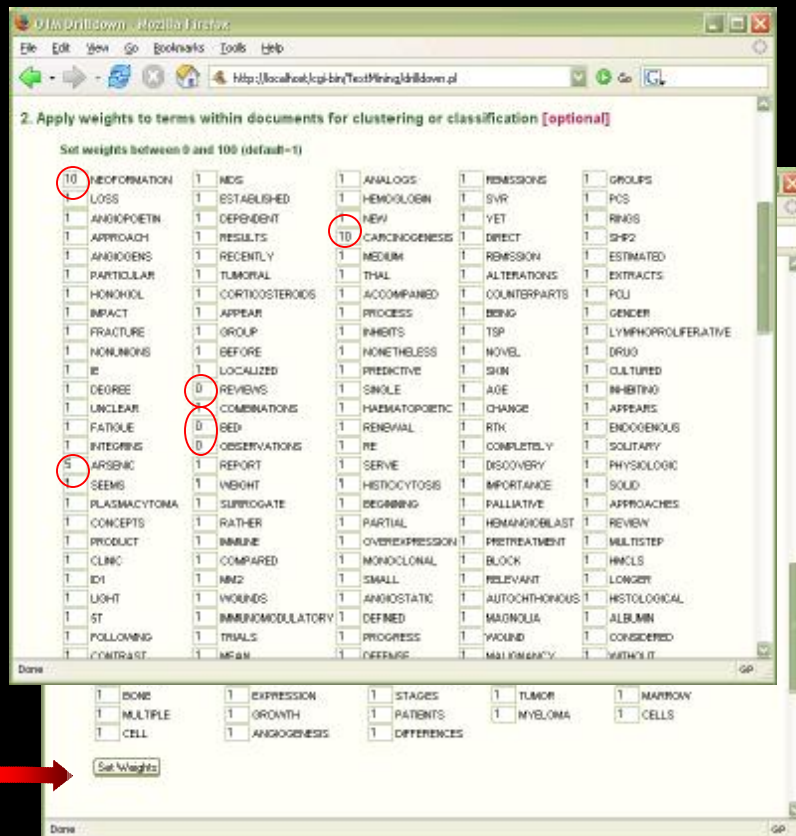
**Middle Screenshot:** The "Document Cluster Drilldown" window is open. A red arrow points to the "Sub-search within documents [optional]" section, where the query "myeloma%" has been entered. The "Search" button is visible.

**Bottom Screenshot:** The application window shows the result of the sub-search. The document list on the right now contains 22 documents, indicating that the search successfully filtered the original set of 31 documents.

ORACLE

# Oracle Text K-Means Clustering

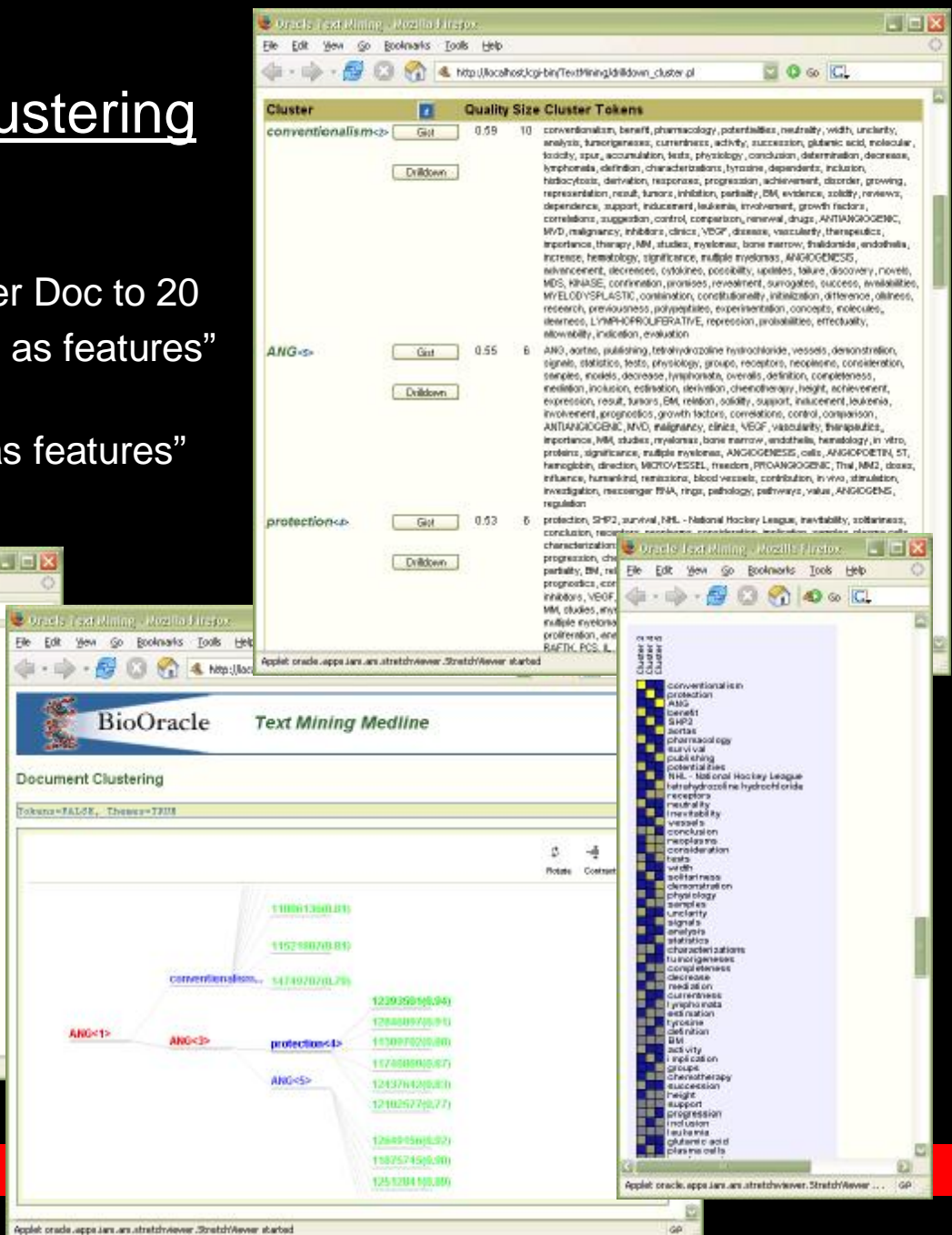
- Change the weights of select terms and click the “Set Weights” button



ORACLE



- Sub-cluster weighted documents
  - Set “No. Clusters” to 3
  - Set Max Distinct Terms per Doc to 20
  - Deselect the “Use Tokens as features” checkbox
  - Select the “Use Themes as features” checkbox



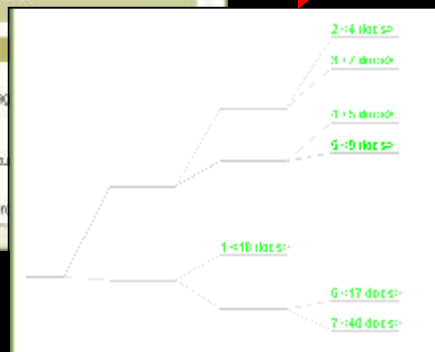
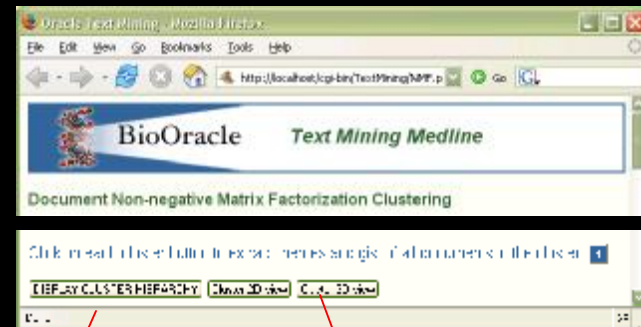
# Text Mining Tasks

## *Clustering*

### ODM

- generate document attributes with one of the following:
  - document term frequencies (not used in demo to speed things up)
  - document themes and theme weights
  - MESH terms binary data (MESH term metadata consists of phrases from a controlled vocabulary assigned to documents; can be used as attributes or as document categories)
- generate a number of NMF document features based on above input data
- generate a specified number of document clusters with K-Means based on NMF feature vectors
- generate cluster gist and themes
  - concatenate all cluster documents, index with Oracle Text, extract gist (ctx\_doc.gist) and themes (ctx\_doc.themes)
- group documents within a cluster based on Kendall's Tau Correlation Coefficient to spot similar or form documents
  - single linkage based on predefined value of  $\text{corr\_k}(\text{doc\_A}, \text{doc\_B})$

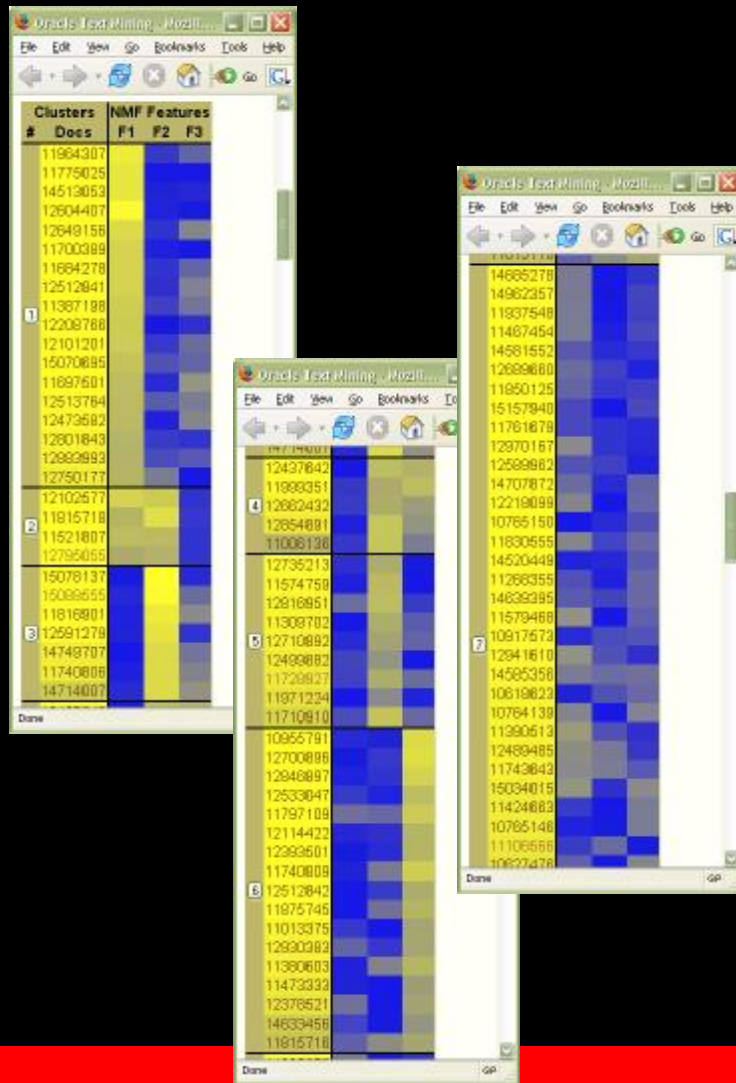
- Set “No. Features” to 3
- Set “No. Clusters” to 7
- Select “Data Source” as “Themes”
- Click the “NMF Cluster” button



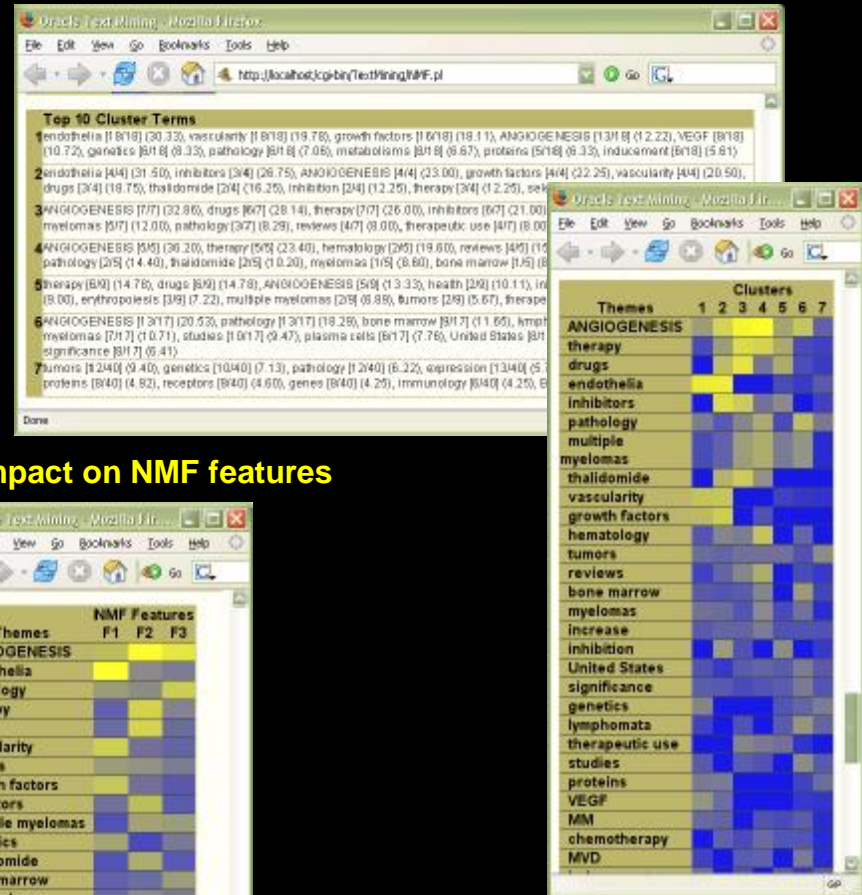


# Oracle Data Mining NMF and K-Means Clustering

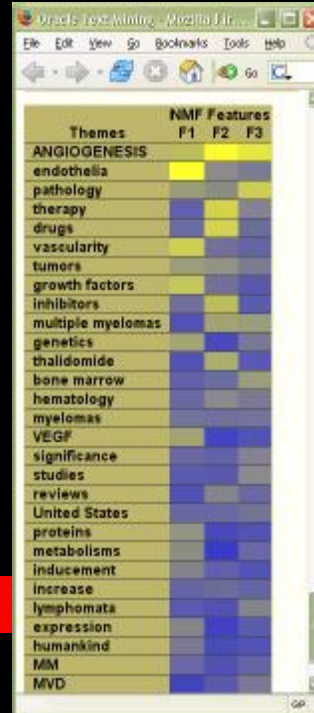
## K-Means clusters based on NMF features



## Top document themes in K-Means clusters



## Theme impact on NMF features

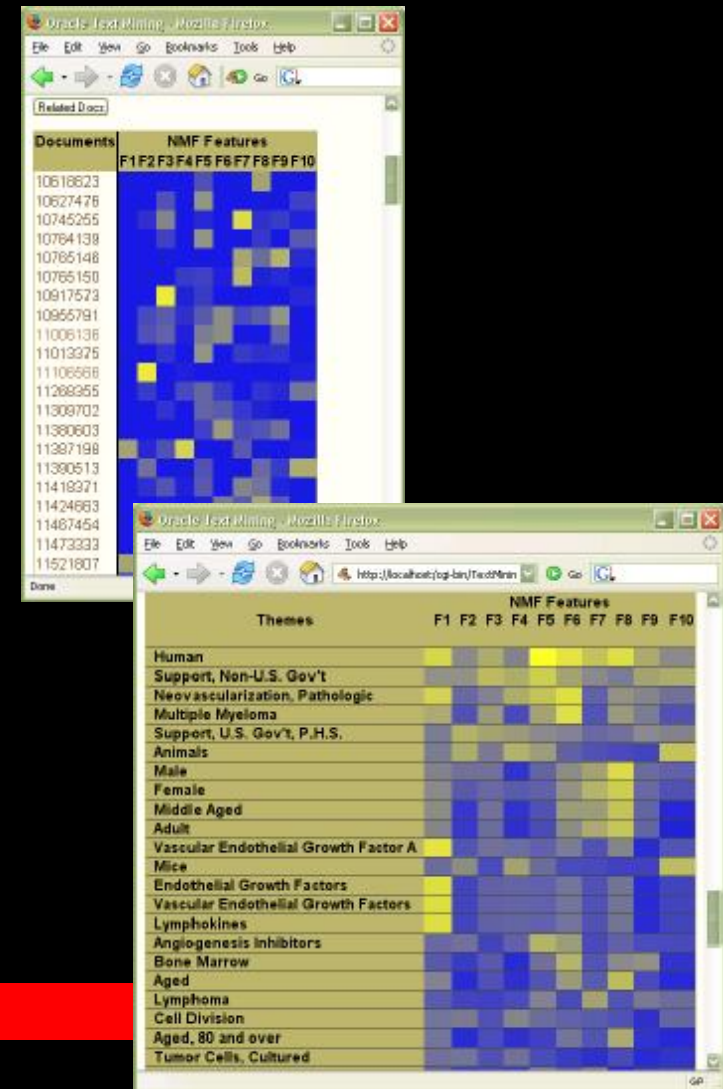


ORACLE



# Oracle Data Mining Document Feature Extraction

- Non-negative Matrix Factorization feature extraction
  - Select “Data Source” as “MeSH terms”
  - Click the “NMF Features” button



# Single Linkage Grouping

- Kendall's Tau Correlation Coefficient based on NMF features
  - Click the "Related Docs" button



# Text Mining Tasks

## *Classification*

### **SVM**

- Use selection of documents to create a new model category
- View existing SVM model categories and delete each if desired
- Classify selected documents with all existing model categories
  - report documents with score>50 for each category
  - cluster each category documents
- Classify selected documents with all existing model categories and use feedback to rebuild model
  - each supportive document can be used as an additional training document for any category
  - each non-supportive document can be checked for presence in a category's training set and removed if present
  - documents can be re-scored with new model categories

ORACLE

# Oracle Text Document Classification

- Clear existing models
- View/Edit existing model classes
- Add new class
- Predict documents with existing model classes

Oracle Text Mining - Mozilla Firefox

Found 100 matching records

**Document Clustering**

☐ k-Means Cluster No. Clusters: 3 Max Distinct Terms per Doc: 10 Use Tokens as features: ☒ Use Themes as features: ☐  
☐ TEXTL Cluster Maximum Splits per Node: 3 Hierarchy Depth: 1 Minimum Leaf Cluster Similarity Score: 0.2  
☐ NMF Cluster No. Features: 4 No. Clusters: 4 Data Source: ☒ Themes ☐ MeSH terms

**Feature Extraction for full Data Mining**

NMF Features Extract optimized number of document-representing features. Data Source: ☒ Themes ☐ MeSH terms

**SVM classification**

Category Named: NT10 Developmental Process

**PMID Score Contents**

PMID	Score	Contents
10925791	80	Prognostic value of bone marrow <b>angiogenesis</b> in multiple myeloma.
12796055	75	Therapeutic potential of selective cyclooxygenase-2 inhibitors in the management of tumor <b>angiogenesis</b>
12309501	69	Prognostic value of <b>angiogenesis</b> in solitary bone plasmacytoma.
11710910	60	The <b>angiogenesis</b> inhibitor vasostatin does not impair wound healing at tumor-inhibiting doses.
11710910	60	density and vascular endothelial growth factor (VEGF)

BioOracle Text Mining Medline

**SVM Classification**

Generating random document model...  
 Populating random category table...  
 Populating random training document table...  
 Removing old rules...  
 Creating training document index...  
 Creating SVM classifier preferences...  
 Training SVM classifier...  
 Creating rules index...  
 Classifying documents...

**Category**

Category	PMID	Score
angiogenesis	12437642	90
	12393901	88
	12700995	86
	14513053	85
	10989595	85
	12513784	84
	11006135	84
	10955791	83
	11075745	82

**Cluster Documents**

No. Clusters: 3 Max Terms per Doc: 10  
 Splits per Node: 3 Hierarchy Depth: 3 Min Cluster Similarity: 0.5  
 No. Features: 4 No. Clusters: 4 Data Source: ☒ Themes ☐ MeSH terms

Oracle Text Mining - Mozilla Firefox

Applet Heatmap started

angiogenesis and lymphoma

PMID	Score	Contents
12437642	90	angiogenesis and lymphoma
12393901	88	angiogenesis and lymphoma
12700995	86	angiogenesis and lymphoma
14513053	85	angiogenesis and lymphoma
10989595	85	angiogenesis and lymphoma
12513784	84	angiogenesis and lymphoma
11006135	84	angiogenesis and lymphoma
10955791	83	angiogenesis and lymphoma
11075745	82	angiogenesis and lymphoma

Oracle Text Mining - Mozilla Firefox

Submit Feedback

angiogenesis and lymphoma

PMID	Score	Contents	Feedback
11697501	72	angiogenesis and lymphoma	<input type="radio"/> T <input type="radio"/> F <input type="radio"/> U
11815716	67	angiogenesis and lymphoma	<input type="radio"/> T <input type="radio"/> F <input type="radio"/> U
12489485	66	angiogenesis and lymphoma	<input type="radio"/> T <input type="radio"/> F <input type="radio"/> U
11700369	66	angiogenesis and lymphoma	<input type="radio"/> T <input type="radio"/> F <input type="radio"/> U
11775025	65	angiogenesis and lymphoma	<input type="radio"/> T <input type="radio"/> F <input type="radio"/> U
12604407	53	angiogenesis and lymphoma	<input type="radio"/> T <input type="radio"/> F <input type="radio"/> U
12710892	50	angiogenesis and lymphoma	<input type="radio"/> T <input type="radio"/> F <input type="radio"/> U
12750177	49	angiogenesis and lymphoma	<input type="radio"/> T <input type="radio"/> F <input type="radio"/> U
11064307	49	angiogenesis and lymphoma	<input type="radio"/> T <input type="radio"/> F <input type="radio"/> U
12208766	49	angiogenesis and lymphoma	<input type="radio"/> T <input type="radio"/> F <input type="radio"/> U
11390513	48	angiogenesis and lymphoma	<input type="radio"/> T <input type="radio"/> F <input type="radio"/> U
11579468	46	angiogenesis and lymphoma	<input type="radio"/> T <input type="radio"/> F <input type="radio"/> U
15134368	46	angiogenesis and lymphoma	<input type="radio"/> T <input type="radio"/> F <input type="radio"/> U
11106566	44	angiogenesis and lymphoma	<input type="radio"/> T <input type="radio"/> F <input type="radio"/> U
10764139	42	angiogenesis and lymphoma	<input type="radio"/> T <input type="radio"/> F <input type="radio"/> U
12473582	40	angiogenesis and lymphoma	<input type="radio"/> T <input type="radio"/> F <input type="radio"/> U
11743643	39	angiogenesis and lymphoma	<input type="radio"/> T <input type="radio"/> F <input type="radio"/> U
11830955	38	angiogenesis and lymphoma	<input type="radio"/> T <input type="radio"/> F <input type="radio"/> U
12533047	38	angiogenesis and lymphoma	<input type="radio"/> T <input type="radio"/> F <input type="radio"/> U
14633456	37	angiogenesis and lymphoma	<input type="radio"/> T <input type="radio"/> F <input type="radio"/> U
12735213	36	angiogenesis and lymphoma	<input type="radio"/> T <input type="radio"/> F <input type="radio"/> U
14707872	35	angiogenesis and lymphoma	<input type="radio"/> T <input type="radio"/> F <input type="radio"/> U
12601843	35	angiogenesis and lymphoma	<input type="radio"/> T <input type="radio"/> F <input type="radio"/> U
12591379	34	angiogenesis and lymphoma	<input type="radio"/> T <input type="radio"/> F <input type="radio"/> U

Applet Heatmap loaded



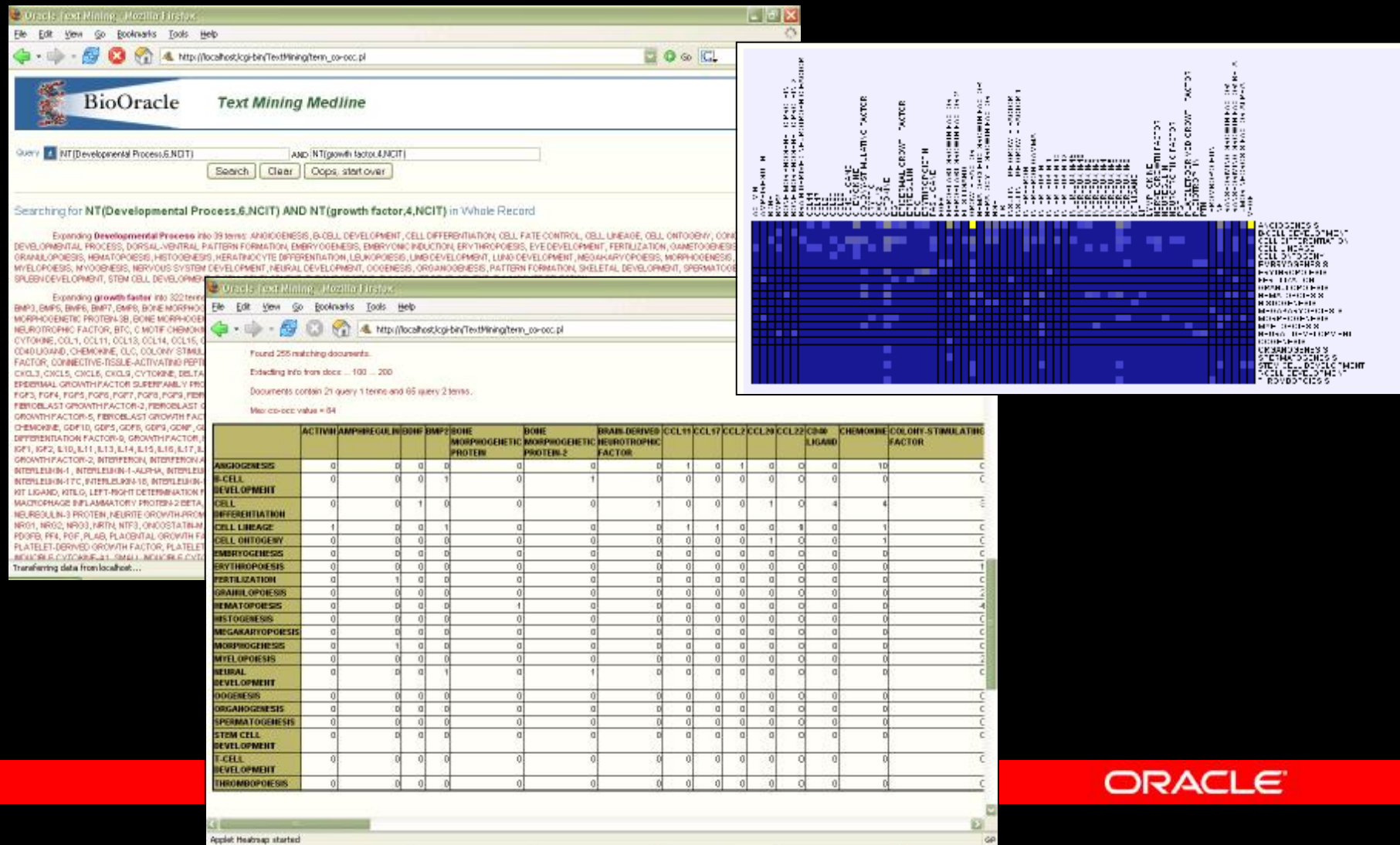
# Oracle Text Knowledgebase

- Query Knowledgebase to check if a specific theme exists in the index
  - Display additional broader and narrower themes indexed



## Thesaurus Co-occurrence

- Compare word lists from 2 thesaurus branches:
  - NT(Developmental Process,6,NCIT) vs. NT(growth factor,4,NCIT)



# Thesaurus Co-occurrence

- View and manipulate graph with Cytoscape
- Save graph as Oracle NDM model

