# New Trends in Data Integration, Analytics, and BI

Mukesh Mohania
IBM India Research

**Acknowledgement: Thanks to all members of IM Research group at New Delhi for their contribution.**

# Agenda

- Information Integration – Definition and architectures
- Existing solutions
- Context Oriented Information Integration
    - SCORE approach
    - EROCS approach
- Integrating Audio Streams with Structured Data
- Data Analytics and BI Applications
    - Improving Semantic Search
    - Preventing Customer Attrition
    - Preventing Information Leakage from Text Documents
    - Improving Cross/Up-Sale
    - Social Network Analysis for Telecom BI
- Conclusions

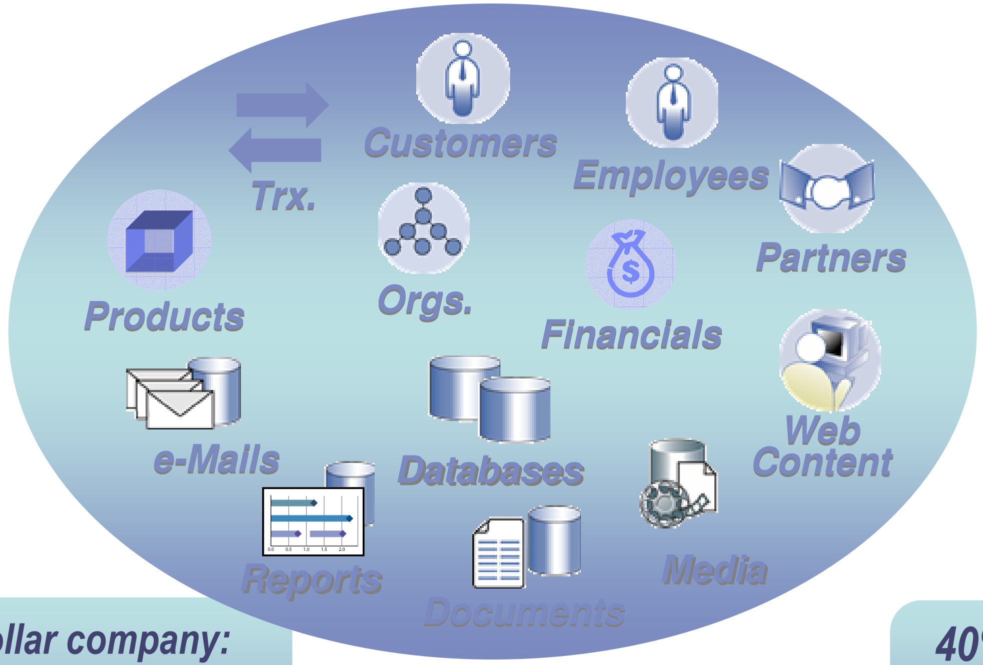
# Market Insights: Information Management Challenges

*60% + of CEOs: Need to do a better job capturing and understanding information rapidly in order to make swift business decisions.*

*Only 1/3rd of CFOs believe that the information is easy to use, tailored, cost effective or integrated.*

*85% of information is unstructured.*

*30-50% of application design time is spent on copy management.*



*42% of transactions are still paper-based.*

*30% of people's time: searching for relevant information.*

*The average billion dollar company:*
*48 disparate financial systems*
*2.7 ERP systems*

*79% of companies have more than two repositories and 25% have more than 15*

*40% of IT budgets may be spent on integration.*

Sources: IBM & Industry Studies, Customer Interviews, Forrester

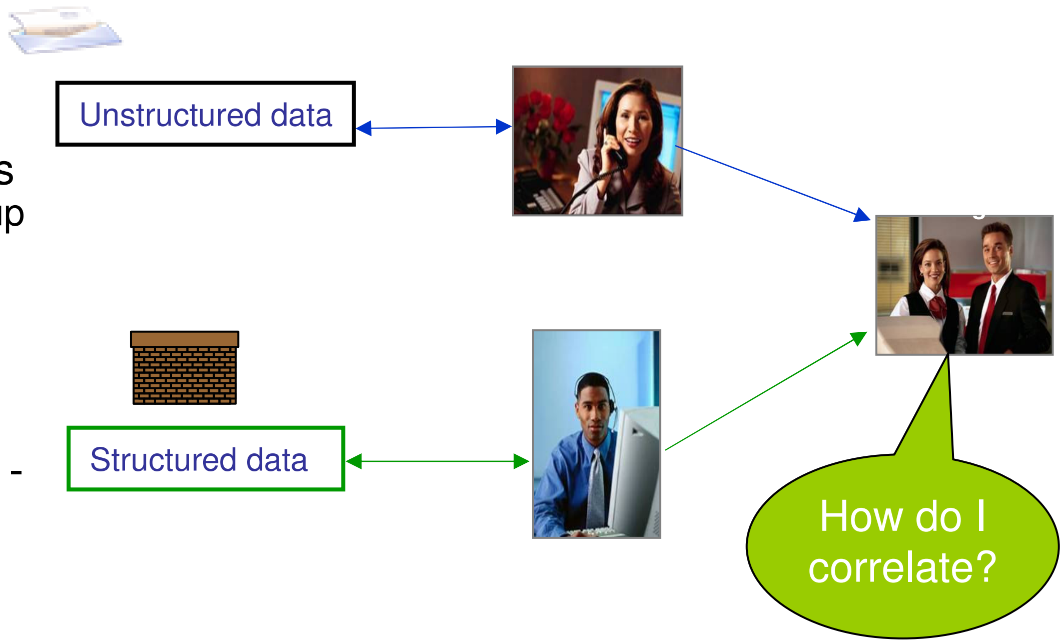
# Valuable Business Information is Buried Under Large Amounts of Unstructured Documents

Structured data:

- Contributes to **20%** of business that is conducted - Gartner Group

Unstructured data:

- **Doubles every three months** - Gartner Group



Unstructured data

Structured data

How do I correlate?

Enterprises are realizing the need to bridge this separation and are demanding **INTEGRATED RETRIEVAL, MANAGEMENT AND ANALYSIS** of both the structured and unstructured content.

**Existing systems do not enable automatic association of the two disparate sources.**

# Improved BI with Email Analysis

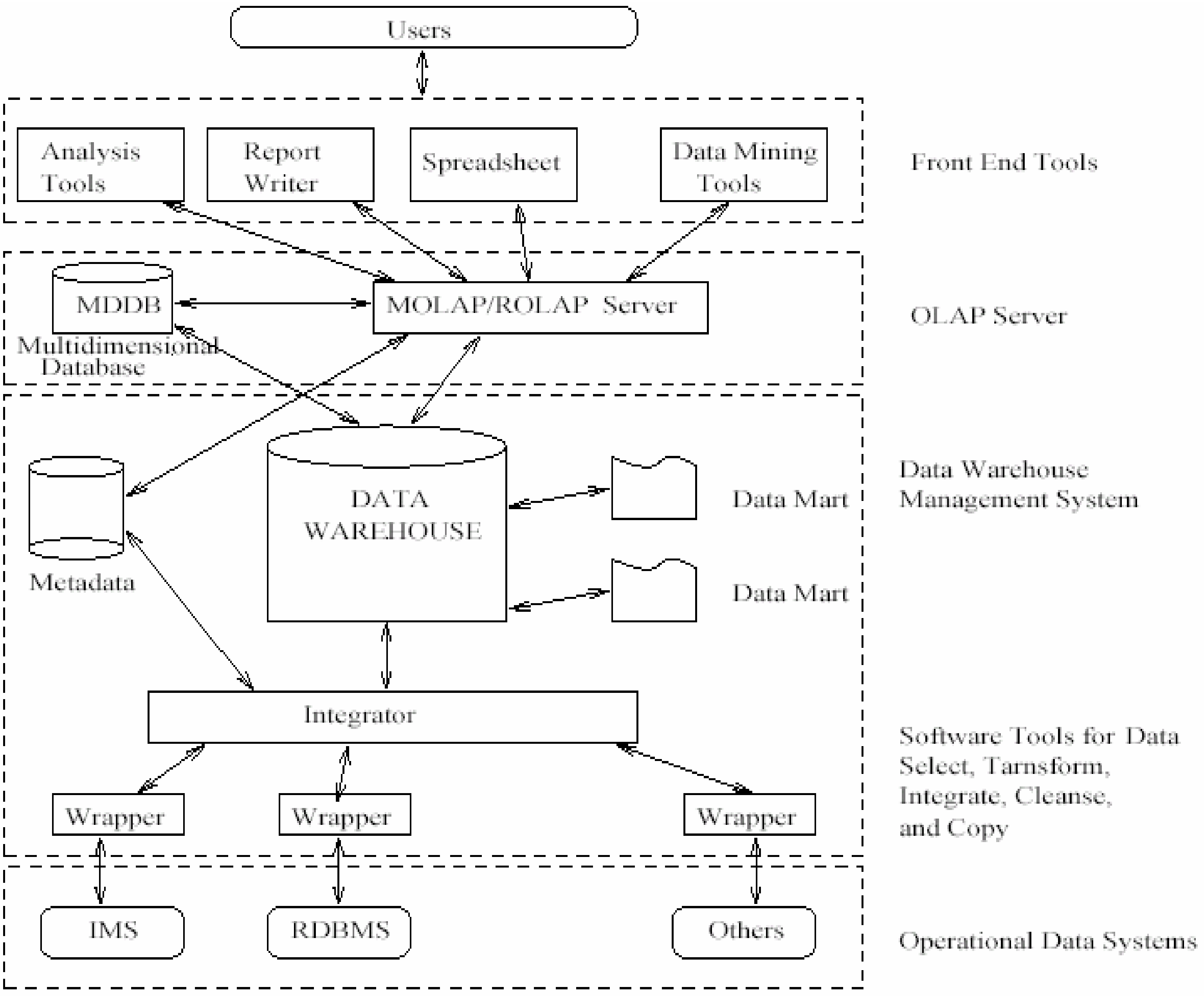
- CRM analysis tends to tell you “what” happened not “why” it happened
  - Customer attrition
    - “ATM usage down by 30% in last quarter” → Likely to leave
  - New product sales
    - Sales 20% less than target
  - Marketing campaigns target selection
    - People who have purchased similar products in the past
- Email analysis may tell you “why”
  - …“very high service charges for loan processing” → customer attrition
  - …“very unhappy with the product quality…” → product sales down
  - …“my wife had called your call center yesterday..” → cross sell opportunity for family products. Better target selection for marketing campaigns
  - …“unable to change my password while traveling…” → customer satisfaction

# Know your Customers Better

- To realize full potential of customer we have to answer certain questions about him/her
  - **Increase share of wallet**: What does a person need? What are his product affinities? What is his opinion? What hinders him from doing more business with us?
  - **Cross-sell/Up-sell** : What products sell best? (Cognos) Will he buy it? When is the right time to sell so that his likelihood of buying is high?
  - **Product Extension**: What features does he like? How can I improve his product experience?
  - **Reduce Churn**: Who is likely to churn? (SPSS) Why is he churning?
  - **Reduce cost to serve**: What is his problem? How can I solve it efficiently?
- Structured Data Analysis and Surveys are employed to answer these questions
  - Structured data analysis can answer only some of these questions
  - Unstructured data can answer more questions which cannot be answered by structured data.

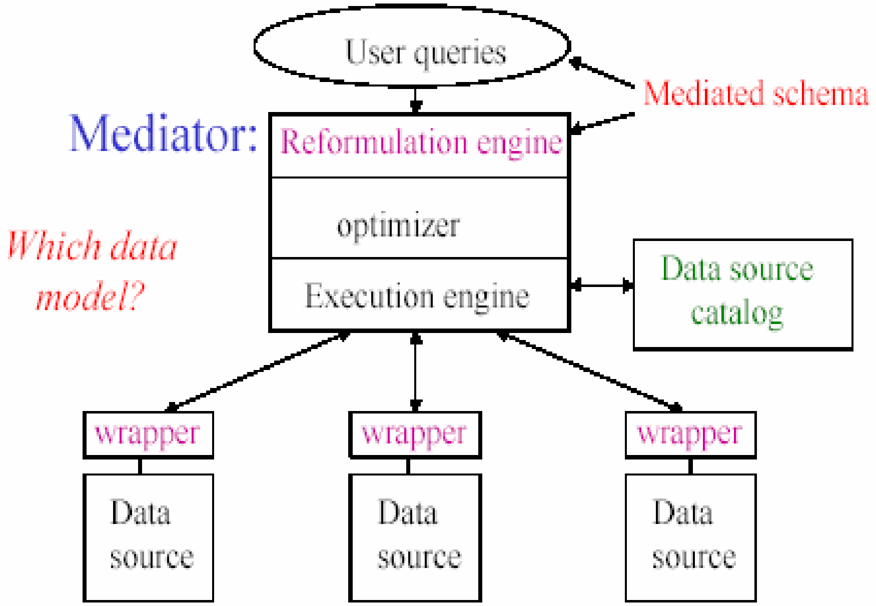
# Existing Solutions

# II Architecture: A Data Warehousing Approach



# II Architecture: Virtualization Layer Approach

- Leave the data in the sources.
- When a query comes in:
  - Determine the relevant sources to the query
  - Break down the query into sub-queries for the sources.
  - Get the answers from the sources, and combine them appropriately.
- Data is fresh. Approach scalable
- Issues:
  - Relating Sources & Mediator
  - Reformulating the query
  - Efficient planning & execution



User queries

Mediated schema

Mediator:

Reformulation engine

optimizer

Execution engine

*Which data model?*

Data source catalog

wrapper

Data source

wrapper

Data source

wrapper

Data source

Garlic [IBM], Hermes[UMD];Tsimmis, InfoMaster[Stanford]; DISCO[INRIA]; Information Manifold [AT&T]; SIMS/Ariadne[USC];Emerac/Havasu[ASU]

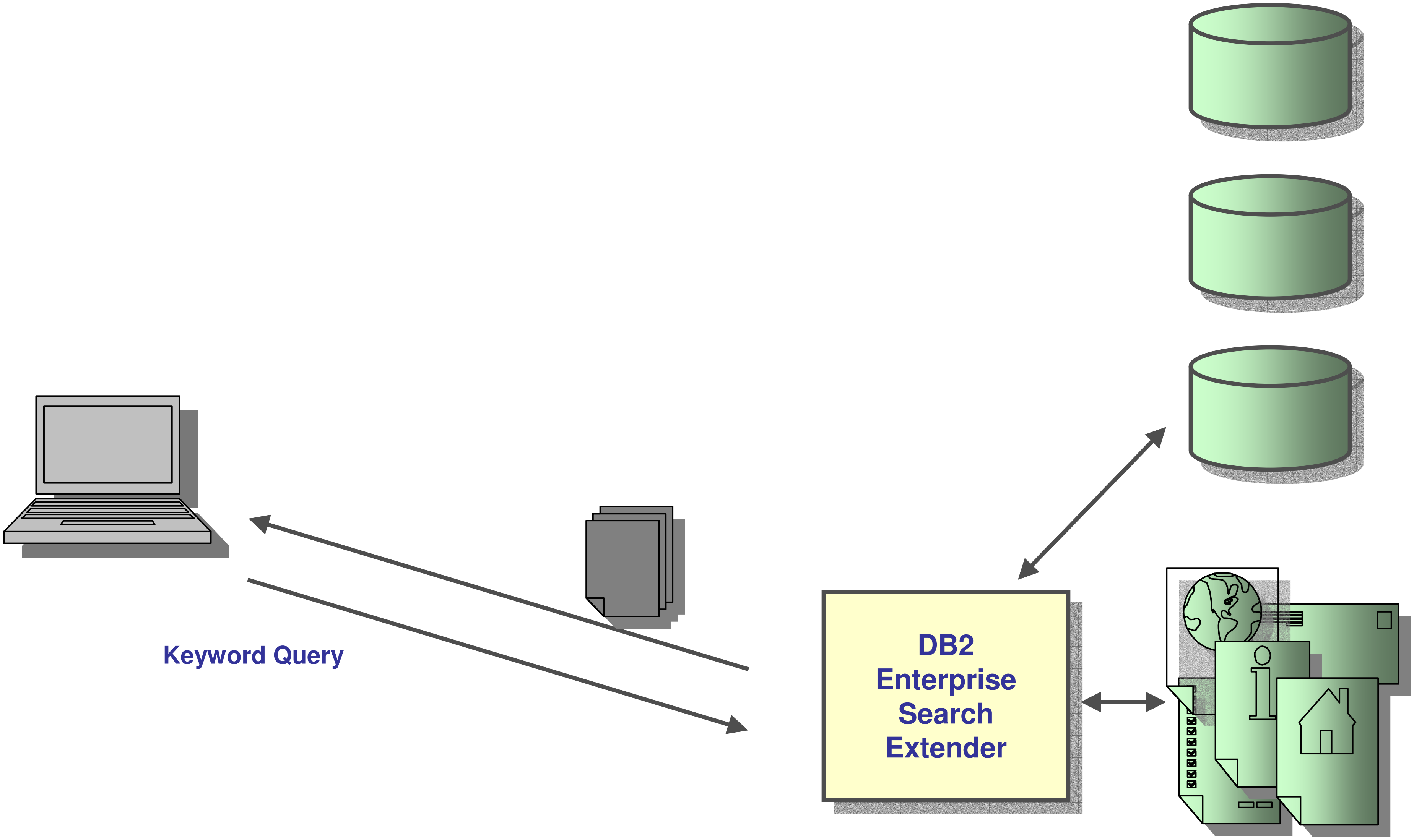
# Structured and Unstructured Information Integration: A Brief Background on Existing Solutions

Existing solutions can be classified in terms of the query paradigm used:

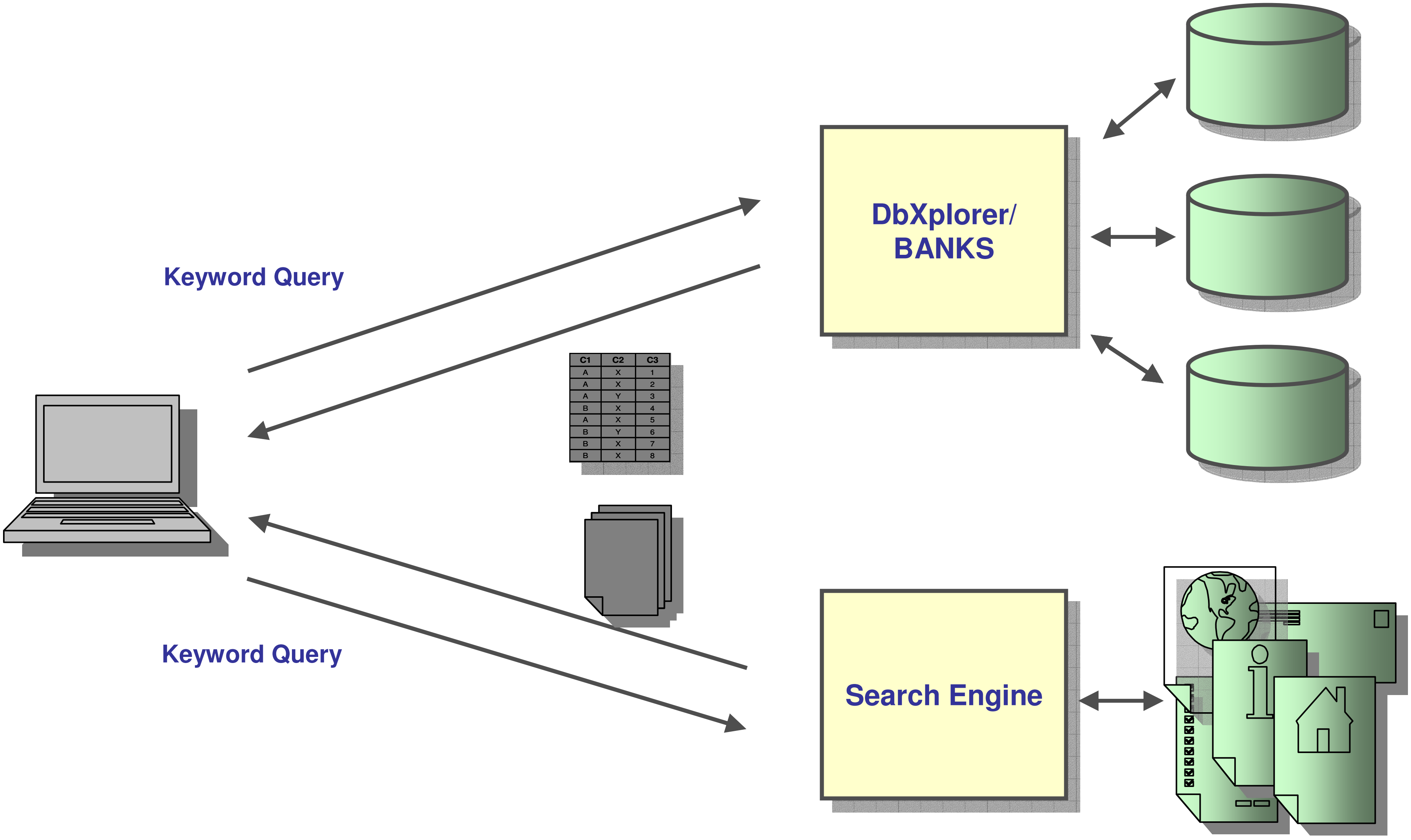
- Keyword Query Based Solutions (DB2 ESE, DbXplorer/BANKS)
    - Relational data exposed to search engine as virtual text documents
    - Query both structured and unstructured information using keywords

- SQL Query Based Solutions (SQL LIKE predicate, DB2 NetSearch Extender)
    - Text data exposed to relational engine as virtual tables with text columns
    - Query both structured and unstructured information using SQL
    - Provide SQL primitives to search text in table columns using a set of keywords

# Keyword Query Based Solution: DB2 ESE



# Keyword Query Based Solution: DbXplorer/BANKS



| C1 | C2 | C3 |
|---|---|---|
| A | X | 1 |
| A | X | 2 |
| A | Y | 3 |
| B | X | 4 |
| A | X | 5 |
| B | Y | 6 |
| B | X | 7 |
| B | X | 8 |

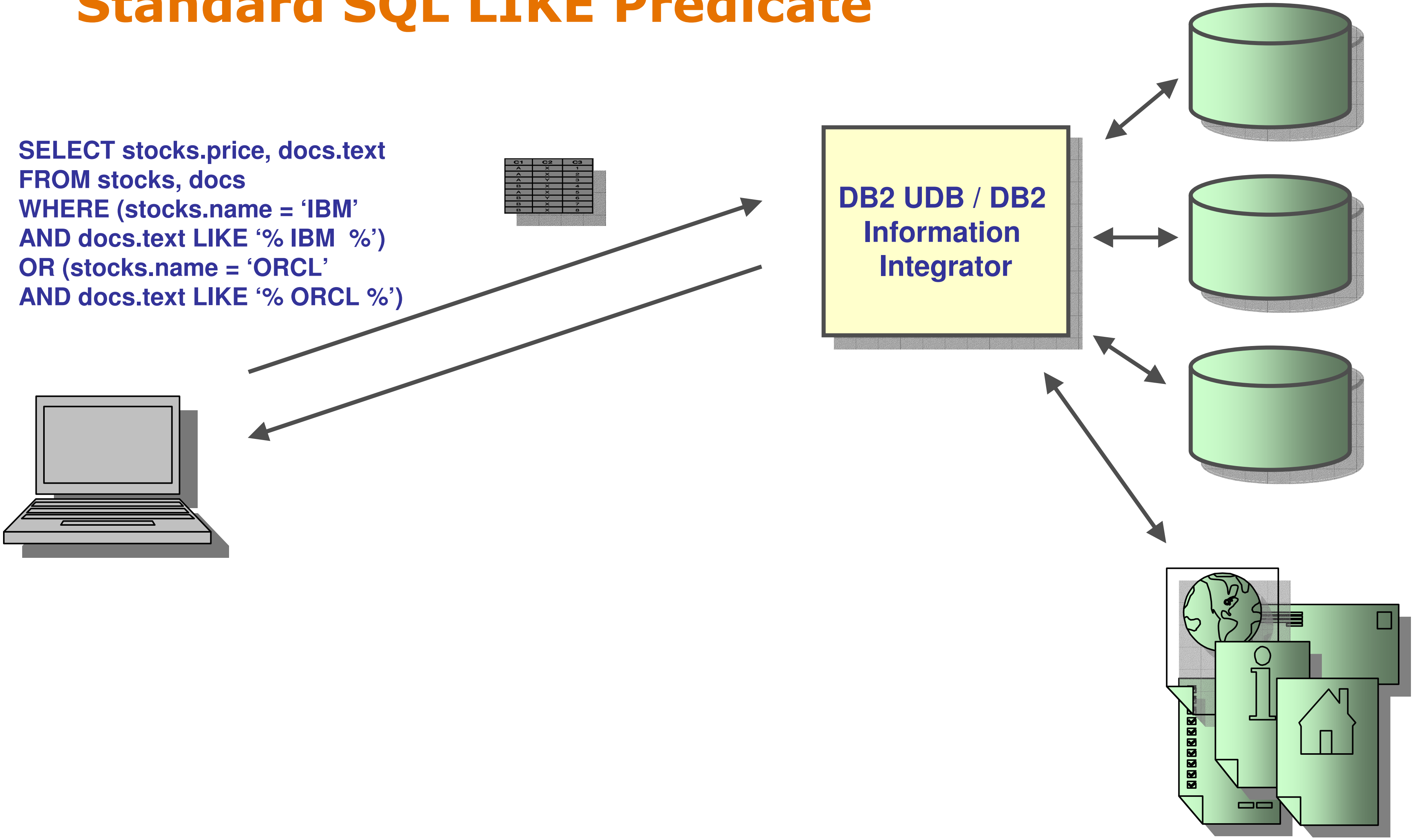
# Keyword Query Based Solutions: Summary

- Advantage: Simplicity!

- Disadvantages
  - Less expressive (as compared to SQL)
    - How to ask for the information related to the five best performing stocks in the past week?
  - Need to specify a set of keywords that succinctly encodes the information need
    - Not always easy

# SQL Query Based Solution: Standard SQL LIKE Predicate

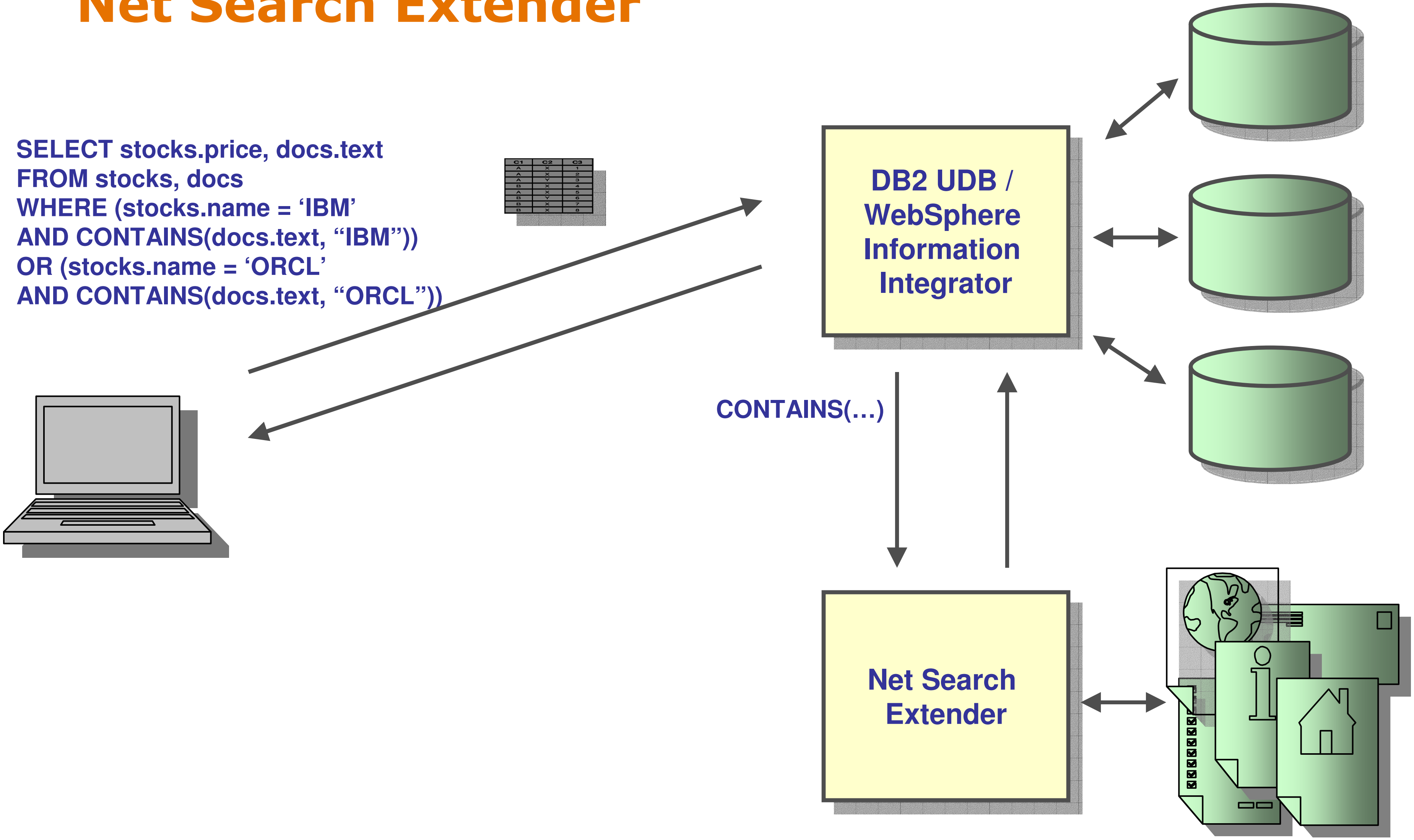
```
SELECT stocks.price, docs.text
FROM stocks, docs
WHERE (stocks.name = ‘IBM’
AND docs.text LIKE ‘% IBM  %’)
OR (stocks.name = ‘ORCL’
AND docs.text LIKE ‘% ORCL %’)
```



DB2 UDB / DB2 Information Integrator

# SQL Query Based Solution:
# Net Search Extender

```
SELECT stocks.price, docs.text
FROM stocks, docs
WHERE (stocks.name = ‘IBM’
AND CONTAINS(docs.text, “IBM”))
OR (stocks.name = ‘ORCL’
AND CONTAINS(docs.text, “ORCL”))
```



# SQL Query Based Solutions: Summary

- Advantages:
  - More expressive – can specify more involved and sophisticated queries

- Disadvantages:
  - The unstructured data is still queried using keywords
  - Need to specify a set of keywords that succinctly encodes the information need
    - Not always easy
  - The SQL query and the embedded keyword query encode the same information need
    - Redundant effort
  - Association of documents with tuples (local context), not with the entire result (global context)
    - Same documents get attached to “IBM” when “IBM” is queried with “ORCL” as when “IBM” is queried with “DELL”

# SCORE Approach –

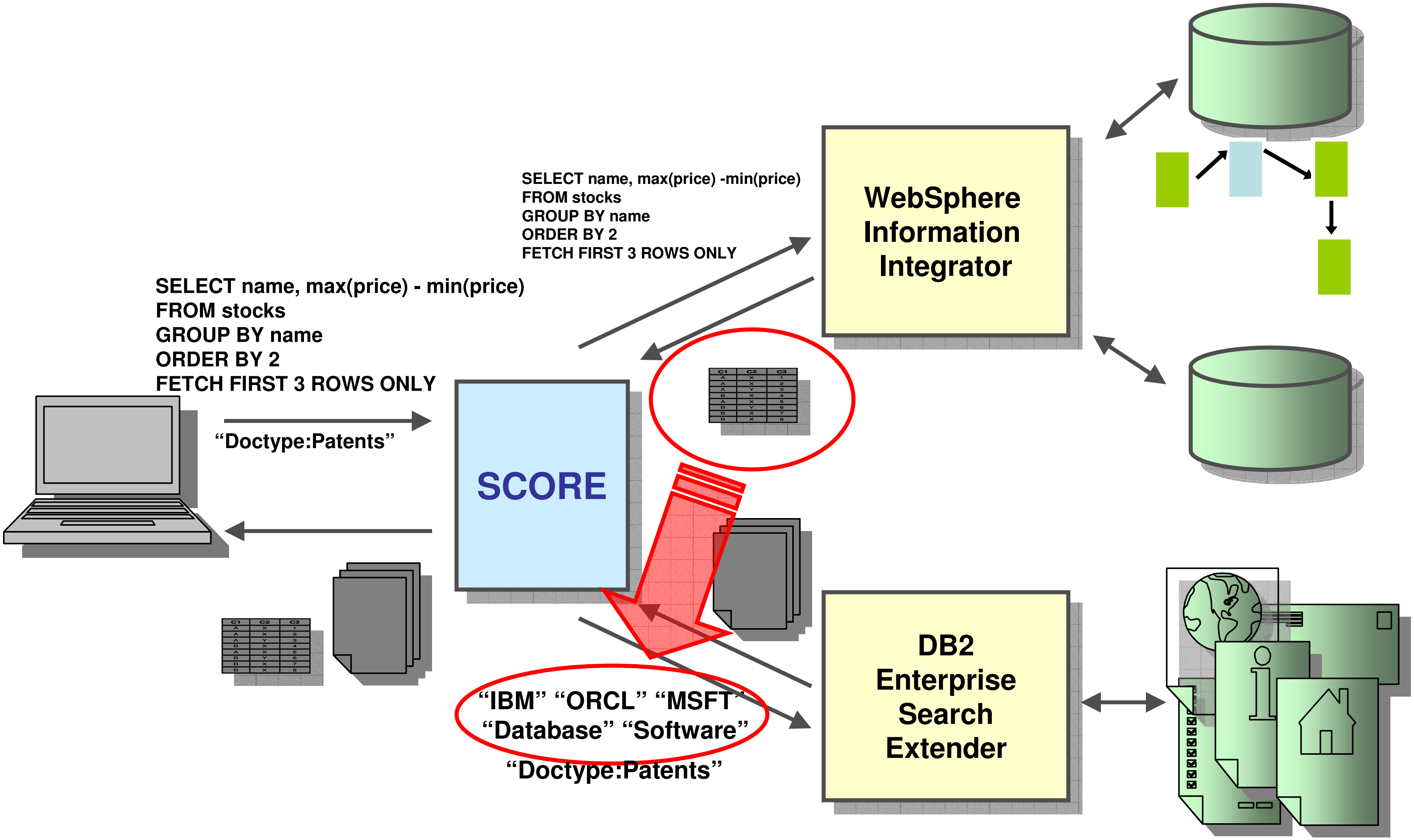
## Associating text Documents with Structured Query Results

**Problem Statement:** Enhance structured data retrieval by associating additional documents relevant to the user context with the query result.

Structured data = relations, schema-based (XML) documents

Unstructured data = schema-less (free-flow) documents, web-pages

# SCORE Overview



```
SELECT name, max(price) -min(price)
FROM stocks
GROUP BY name
ORDER BY 2
FETCH FIRST 3 ROWS ONLY
```

WebSphere Information Integrator

```
SELECT name, max(price) - min(price)
FROM stocks
GROUP BY name
ORDER BY 2
FETCH FIRST 3 ROWS ONLY
```

“Doctype:Patents”

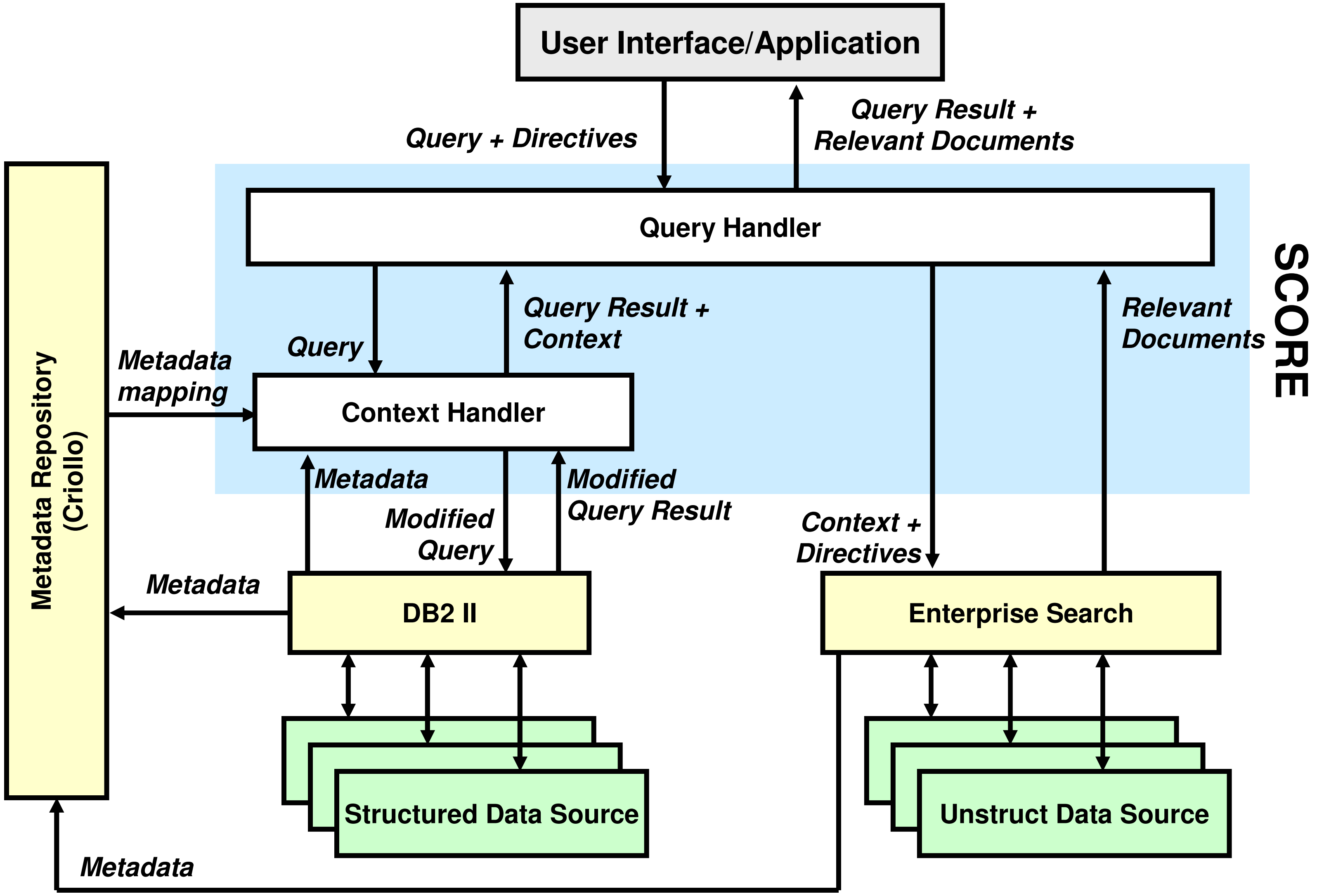
SCORE

“IBM” “ORCL” “MSFT”
“Database” “Software”

“Doctype:Patents”

DB2 Enterprise Search Extender

# Overall Architecture



# EROCS Approach –

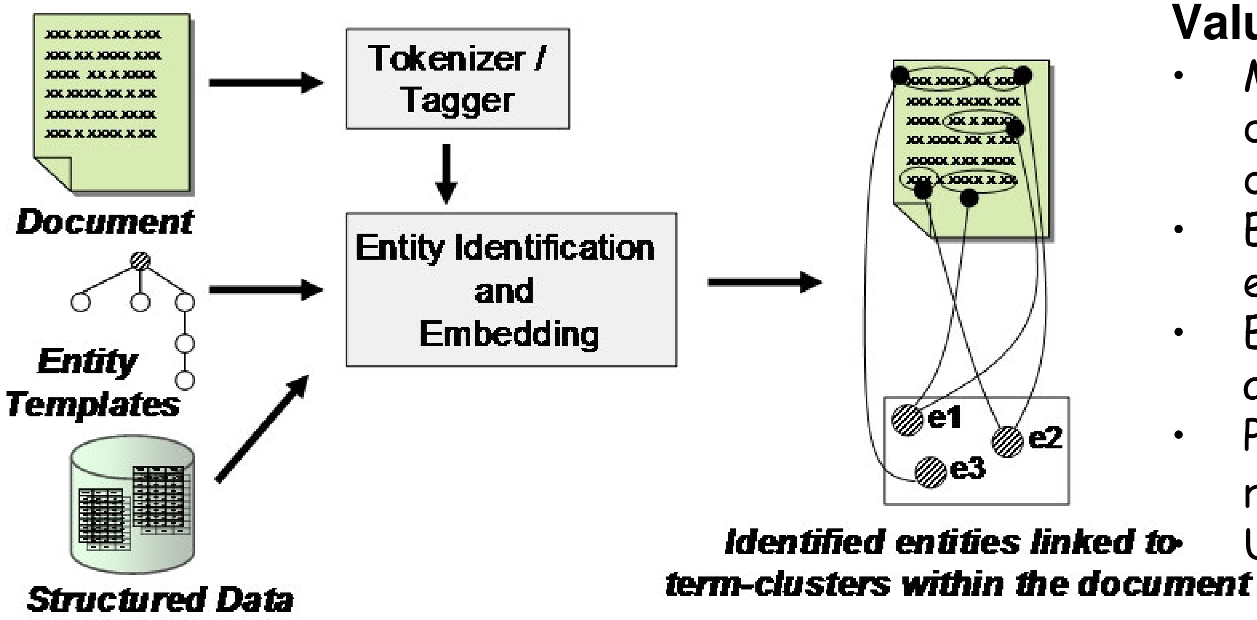
## Associating Relevant Structured Data with Text Documents

# Linkage Discovery (EROCS): Efficiently Linking Diverse Data

- Exploit partial information contained in a document to automatically identify and link relevant structured data

**Main Idea**

- View the structured data as a set of pre-defined “entities”
- Identify the entities from this set that best match the document, and also find embeddings of the identified entities in the document


Document
Entity Templates
Structured Data
Tokenizer / Tagger
Entity Identification and Embedding
e1
e2
e3
Identified entities linked to term-clusters within the document

**Value Proposition**

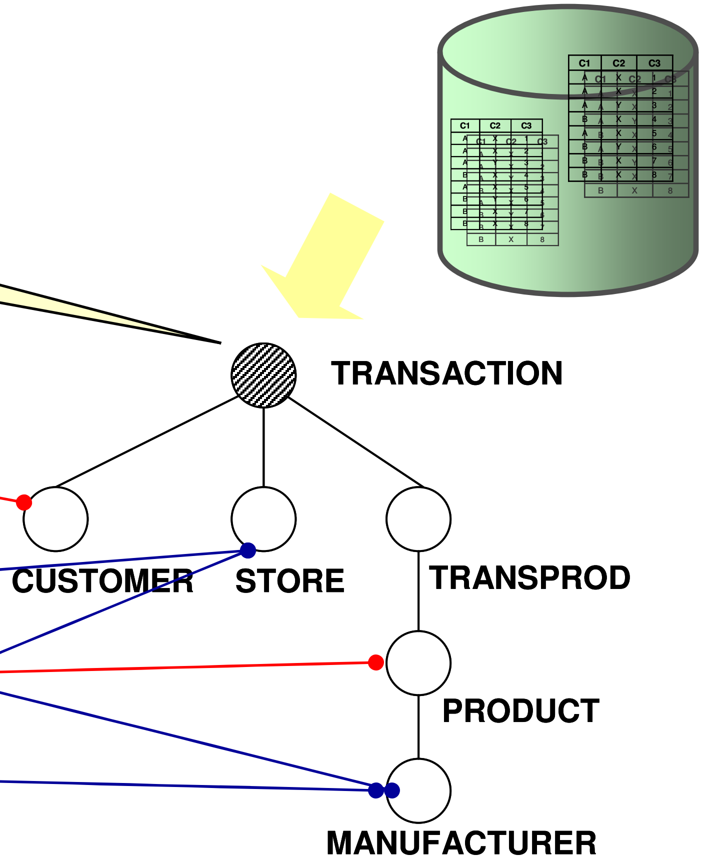
- Metadata Extraction (linking documents with structured data)
- Enhance semantic Search by exploiting this Metadata
- Enable BI across Structured and Unstructured Data
- Providing more metadata and richer text search in CM UIMA Annotator

**VLDB 2006, SIGMOD 2007, PODS 2007, ICDE 2008 (Demo)**

# Example

Find the transaction that best matches the context

I am <Name> John Smith </Name>
….
…… bought a
<Company>Sony</Company>
<product> DVD player </product>
….
from <Company>JK Electronics
</Company> ……..

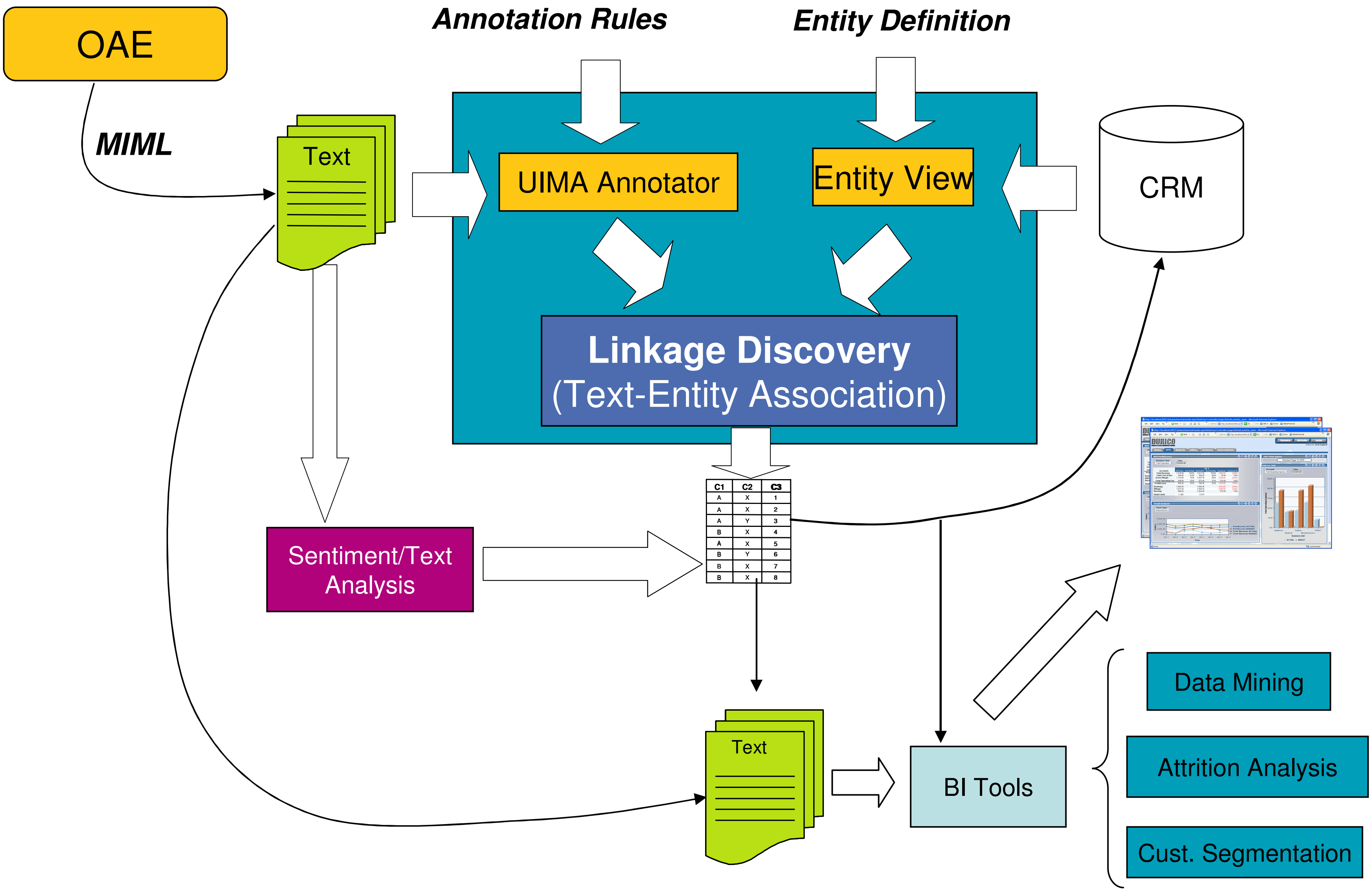


| CustId | StoreId | Payment | Discount | Terms |
|---|---|---|---|---|
| A756K9 | S8976 | Card (AMEX) | Promo# 1236 | NOREFND |

| CustId | Name | Loyalty | Club | Addr |
|---|---|---|---|---|
| A756K9 | John W Smith | Platinum | IBM | Chicago, IL |

Additional "sidebar" information available as a result of the annotation

# Linkage Discovery Architecture



| C1 | C2 | C3 |
|---|---|---|
| A | X | 1 |
| A | X | 2 |
| A | Y | 3 |
| B | X | 4 |
| A | X | 5 |
| B | Y | 6 |
| B | X | 7 |
| B | X | 8 |

# Issues:

- Connecting People, Products, etc in a Database To Text document
  - Learn the key differences between finding names of people and products versus connecting those names to database ID's
  - Understand the three major difficulties encountered in linking:
    - Ambiguity of reference, unknown aliases, and ambiguity of type
  - Explore state-of-the-art approaches to resolving the problem which span machine learning to rule based systems

# CallAssist: Integrating Real-time Audio with Databases

**CallAssist**

- A novel system for linking audio conversations with relevant structured data automatically in real-time,
- Suggests informative queries to narrow down the context

## Value Proposition

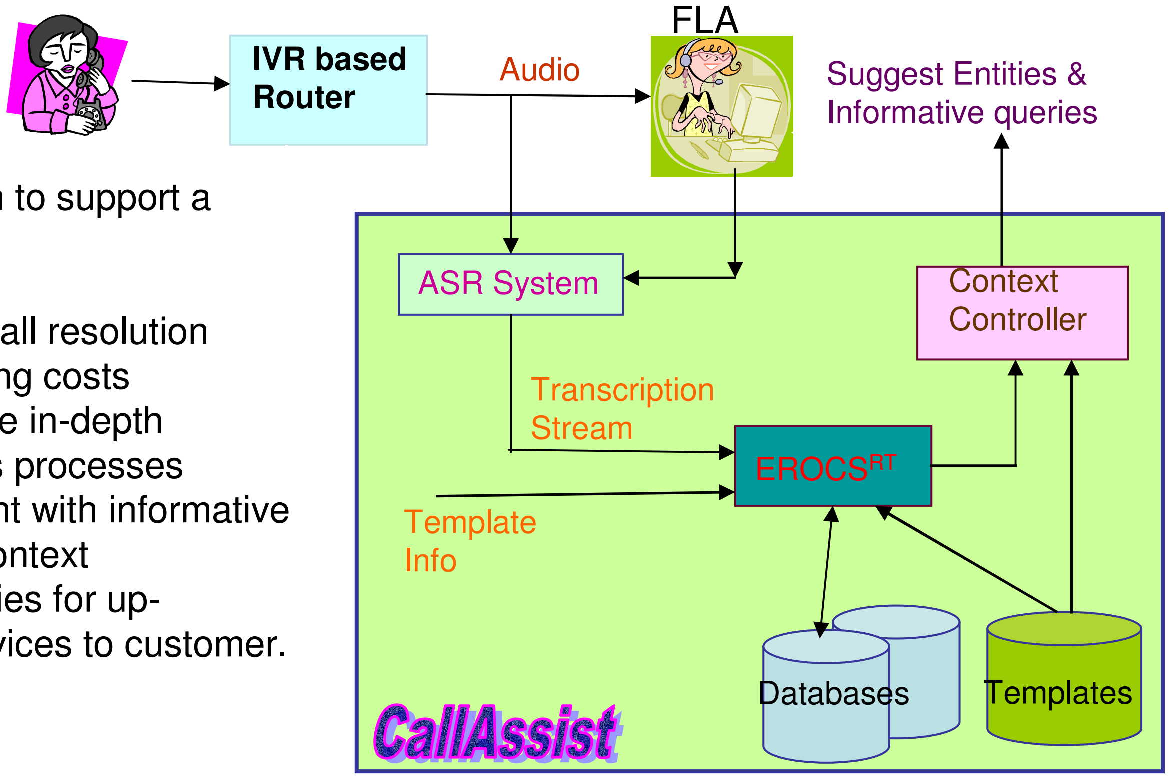
Cost Reduction: the end-to-end time taken to support a customer will be reduced.

- Reduced call escalation rates
  - Lower average time for call resolution
  - Reduction in agent training costs

    -- Agents need not be have in-depth understanding of business processes

    -- System will prompt agent with informative queries that can narrow context

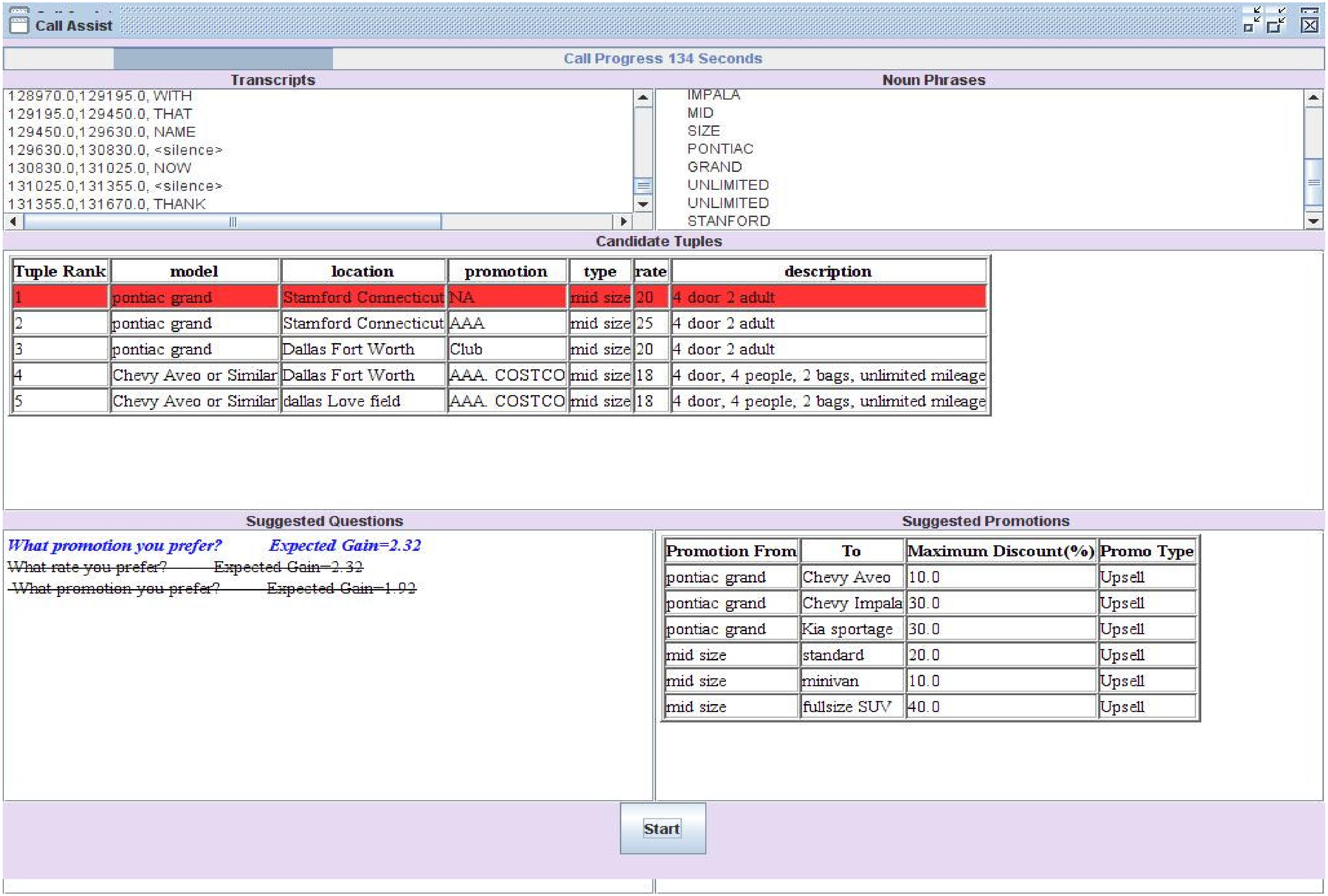
Potential Sales: By suggesting opportunities for up-selling/cross-selling products and services to customer.



**Potential Applications**

Automate Q&A for Call Center
Automated Solution Provider
Dynamic Learning/Self-help Program
Integrated Trouble Ticket cum Diagnosis Generator
Extend Customer Data Integration tools to support real-time audio

**VLDB 2007 (Demo), SIGMOD 2008**



Transcripts

```
128970.0,129195.0, WITH
129195.0,129450.0, THAT
129450.0,129630.0, NAME
129630.0,130830.0, <silence>
130830.0,131025.0, NOW
131025.0,131355.0, <silence>
131355.0,131670.0, THANK
```

Noun Phrases

IMPALA
MID
SIZE
PONTIAC
GRAND
UNLIMITED
UNLIMITED
STANFORD

Candidate Tuples

| Tuple Rank | model | location | promotion | type | rate | description |
|---|---|---|---|---|---|---|
| 1 | pontiac grand | Stamford Connecticut | NA | mid size | 20 | 4 door 2 adult |
| 2 | pontiac grand | Stamford Connecticut | AAA | mid size | 25 | 4 door 2 adult |
| 3 | pontiac grand | Dallas Fort Worth | Club | mid size | 20 | 4 door 2 adult |
| 4 | Chevy Aveo or Similar | Dallas Fort Worth | AAA. COSTCO | mid size | 18 | 4 door, 4 people, 2 bags, unlimited mileage |
| 5 | Chevy Aveo or Similar | dallas Love field | AAA. COSTCO | mid size | 18 | 4 door, 4 people, 2 bags, unlimited mileage |

Suggested Questions

***What promotion you prefer?*** ***Expected Gain=2.32***

~~What rate you prefer? Expected Gain=2.32~~

~~What promotion you prefer? Expected Gain=1.92~~

Suggested Promotions

| Promotion From | To | Maximum Discount(%) | Promo Type |
|---|---|---|---|
| pontiac grand | Chevy Aveo | 10.0 | Upsell |
| pontiac grand | Chevy Impala | 30.0 | Upsell |
| pontiac grand | Kia sportage | 30.0 | Upsell |
| mid size | standard | 20.0 | Upsell |
| mid size | minivan | 10.0 | Upsell |
| mid size | fullsize SUV | 40.0 | Upsell |

# Data Analytics and BI Applications

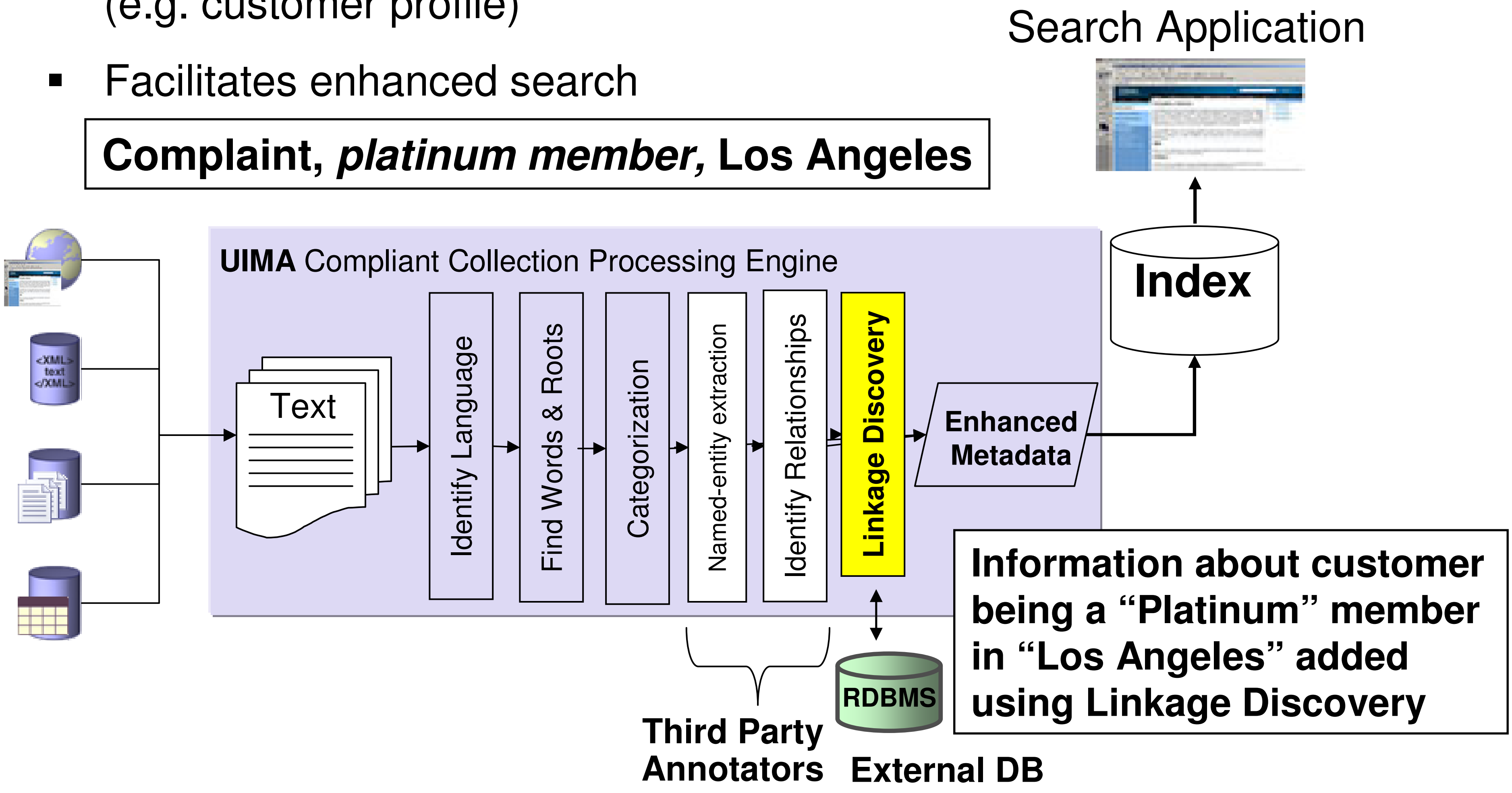
# Example Scenario: *Semantic Search*

- More types (e.g. transaction)
- Attributes from structured data (e.g. customer profile)
- Facilitates enhanced search

**Enables Semantic Search on Text + External DB**

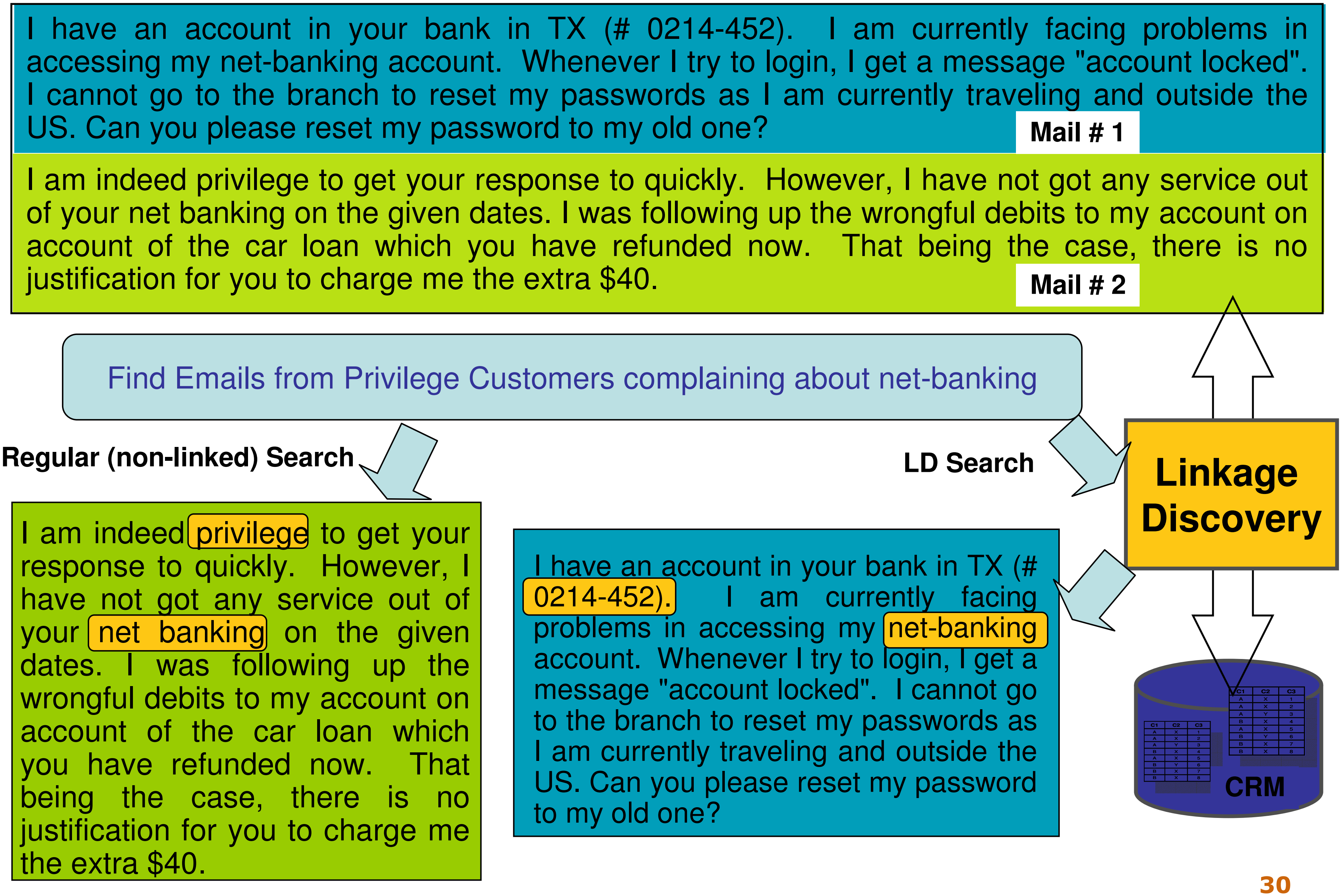
**Improves precision/recall**

**Complaint, *platinum member,* Los Angeles**



**Information about customer being a "Platinum" member in "Los Angeles" added using Linkage Discovery**

# Semantic Search Example



I have an account in your bank in TX (# 0214-452). I am currently facing problems in accessing my net-banking account. Whenever I try to login, I get a message "account locked". I cannot go to the branch to reset my passwords as I am currently traveling and outside the US. Can you please reset my password to my old one?

**Mail # 1**

I am indeed privilege to get your response to quickly. However, I have not got any service out of your net banking on the given dates. I was following up the wrongful debits to my account on account of the car loan which you have refunded now. That being the case, there is no justification for you to charge me the extra $40.

**Mail # 2**

Find Emails from Privilege Customers complaining about net-banking

**Regular (non-linked) Search**

I am indeed privilege to get your response to quickly. However, I have not got any service out of your net banking on the given dates. I was following up the wrongful debits to my account on account of the car loan which you have refunded now. That being the case, there is no justification for you to charge me the extra $40.

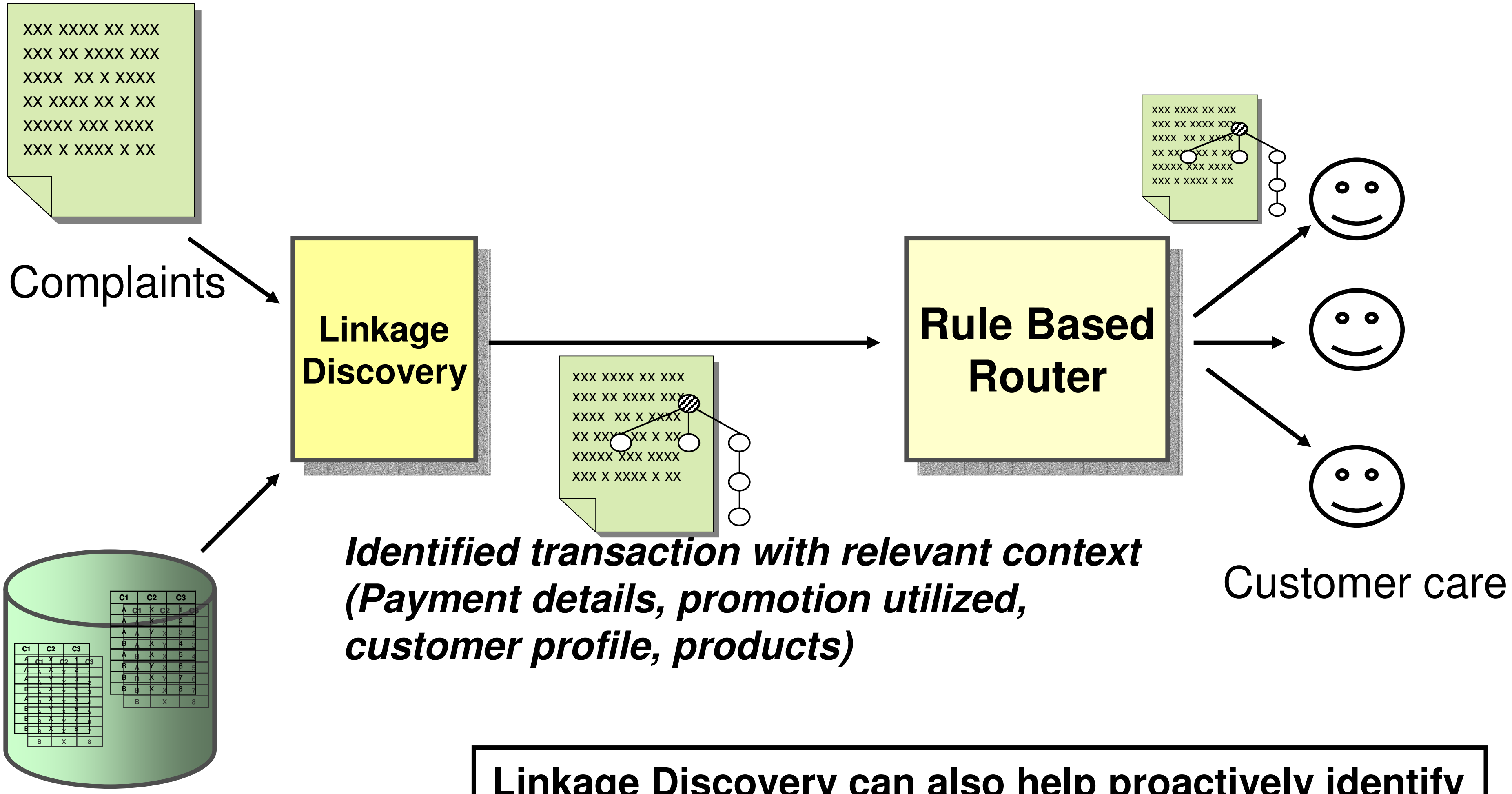
**LD Search**

**Linkage Discovery**

I have an account in your bank in TX (# 0214-452). I am currently facing problems in accessing my net-banking account. Whenever I try to login, I get a message "account locked". I cannot go to the branch to reset my passwords as I am currently traveling and outside the US. Can you please reset my password to my old one?

**CRM**

# Example: Complaint Routing



Complaints

Linkage Discovery

Rule Based Router

Customer care

***Identified transaction with relevant context (Payment details, promotion utilized, customer profile, products)***

Transactions data

**Linkage Discovery can also help proactively identify the *minimal* additional information needed from the customer to identify the relevant transaction**

# Improved BI spanning both on Content and Data

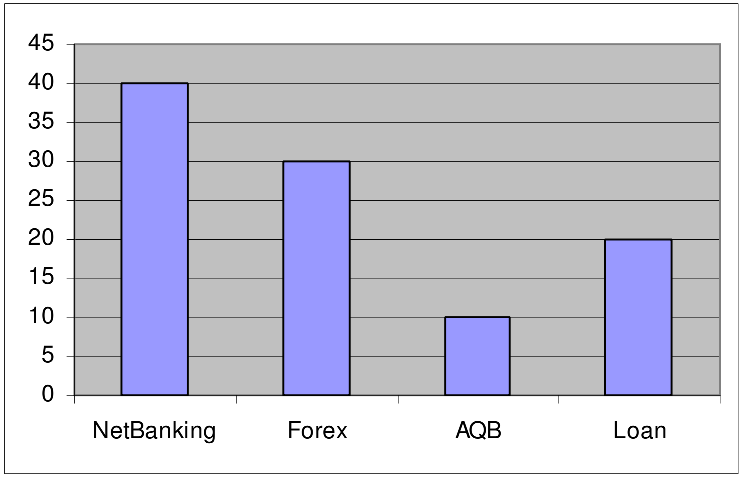
Show me the top 4 **pain points** of my most **privileged customers from North region** who have reduced their balance by more than **50% in the last quarter**.

Information from Emails

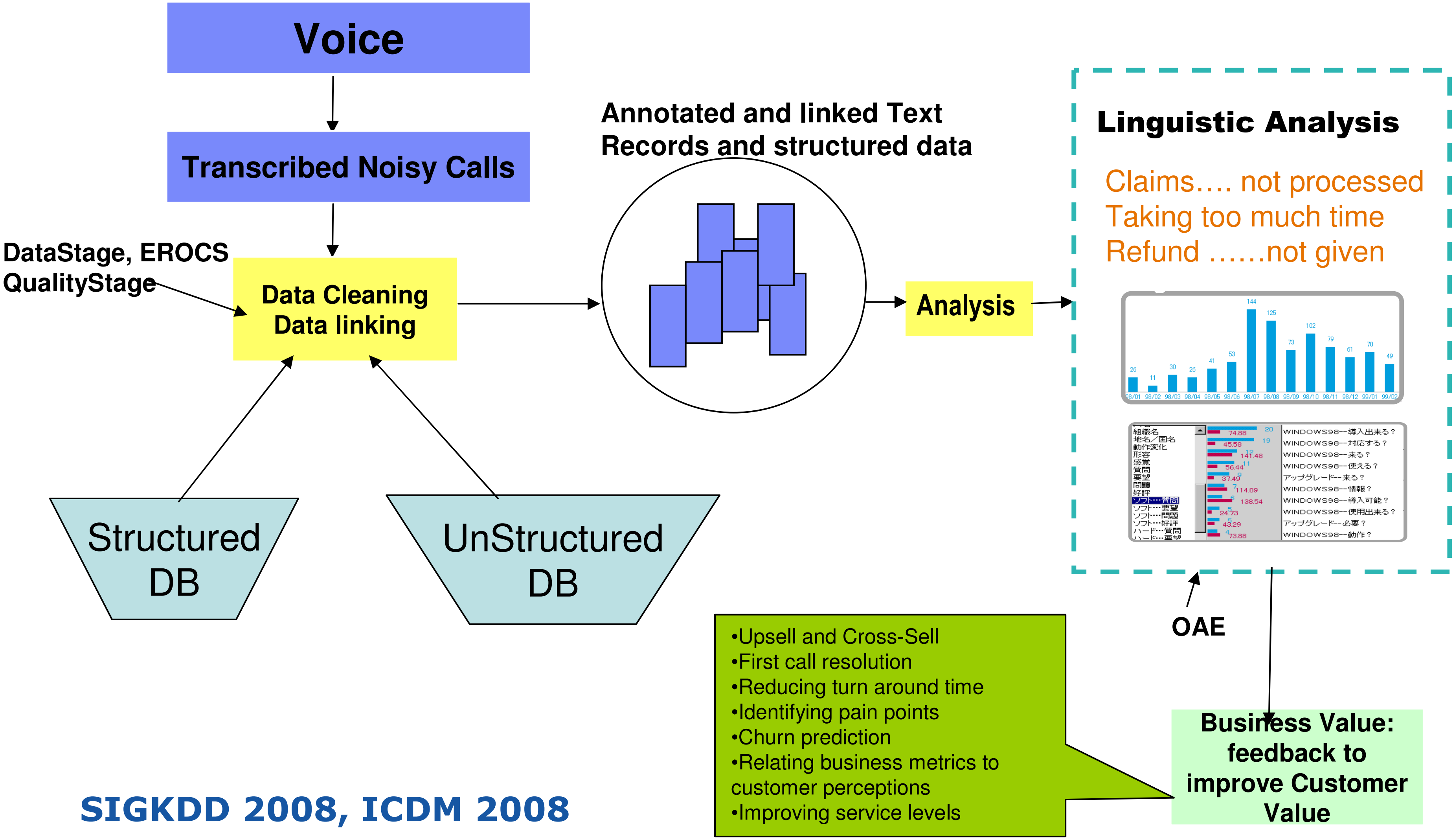
Information from CRM Data

- Traditional BI systems cannot answer this type of hybrid query requiring manual analysis

- OmniFind Analytics Edition with Linkage Discovery can.



# Discovering Customer Needs to Maximize Customer Value



Voice

Transcribed Noisy Calls

DataStage, EROCS QualityStage

Data Cleaning Data linking

Structured DB

UnStructured DB

Annotated and linked Text Records and structured data

Analysis

**Linguistic Analysis**

Claims…. not processed
Taking too much time
Refund ……not given

OAE

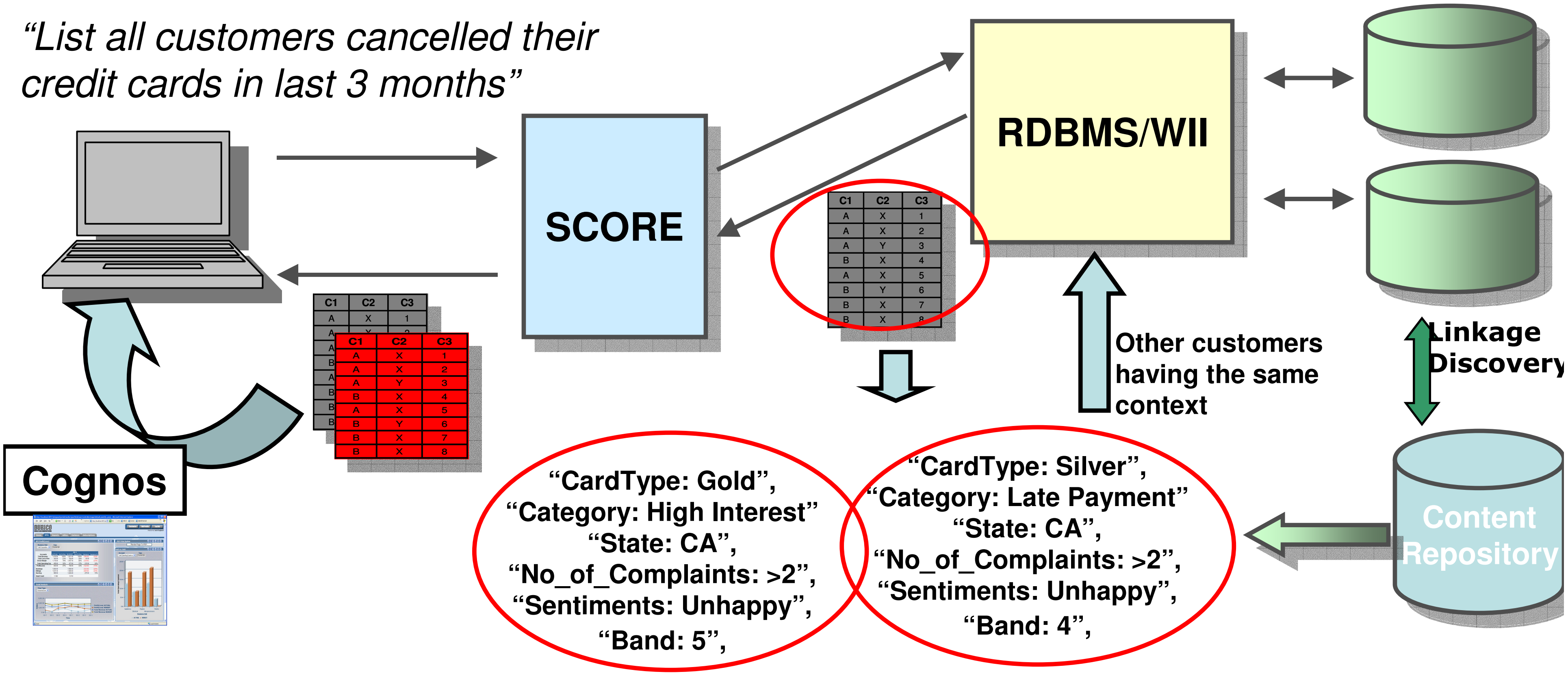
- •Upsell and Cross-Sell
- •First call resolution
- •Reducing turn around time
- •Identifying pain points
- •Churn prediction
- •Relating business metrics to customer perceptions
- •Improving service levels

**Business Value: feedback to improve Customer Value**

**SIGKDD 2008, ICDM 2008**

# Customer Relationship Management: Churn Prediction

- Attrition Prevention
  - Use SCORE to deduce common features of the set of customers who have cancelled their credit card
  - Prioritize customer retention campaign for remaining customers exhibiting these characteristics



*"List all customers cancelled their credit cards in last 3 months"*

SCORE

RDBMS/WII

| C1 | C2 | C3 |
|---|---|---|
| A | X | 1 |
| A | X | 2 |
| A | Y | 3 |
| B | X | 4 |
| A | X | 5 |
| B | Y | 6 |
| B | X | 7 |
| B | X | 8 |

| C1 | C2 | C3 |
|---|---|---|
| A | X | 1 |
| A | X | 2 |
| A | Y | 3 |
| B | X | 4 |
| A | X | 5 |
| B | Y | 6 |
| B | X | 7 |
| B | X | 8 |

Cognos

"CardType: Gold",
"Category: High Interest"
"State: CA",
"No_of_Complaints: >2",
"Sentiments: Unhappy",
"Band: 5",

"CardType: Silver",
"Category: Late Payment"
"State: CA",
"No_of_Complaints: >2",
"Sentiments: Unhappy",
"Band: 4",

Other customers having the same context

Linkage Discovery

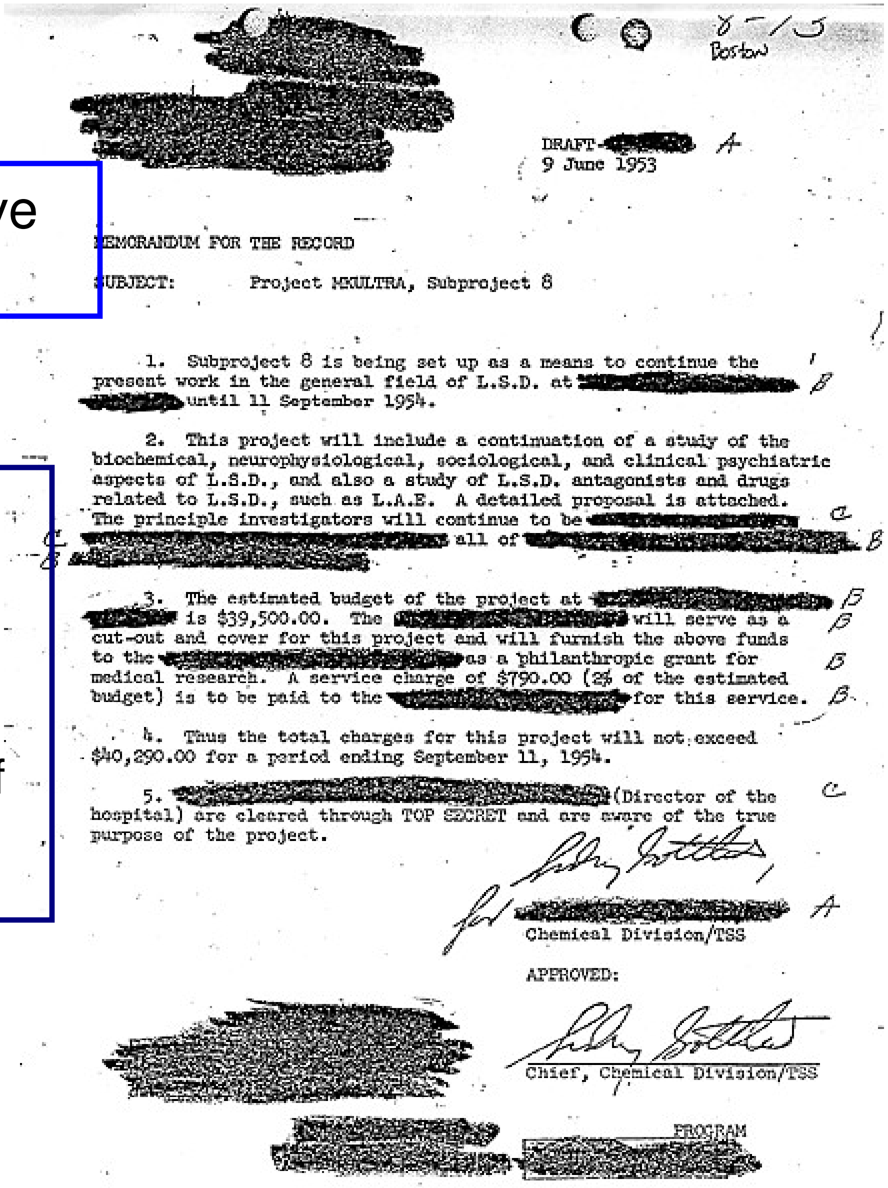
Content Repository

# Document Sanitization: Preventing Information Leakage

Sanitization involves removing sensitive information from a document.

## Problem Statement

- ➢Given a document D and a parameter K, delete a minimum number of terms so that the remaining document T is K-safe.
  - ➢ K-safety: A set of terms T is K-safe, if for any entity e, at least K other entities contain $T \cap C(e)$ in their context



[redacted] 8-15 Boston

DRAFT-[redacted] A
9 June 1953

MEMORANDUM FOR THE RECORD

SUBJECT: Project MKULTRA, Subproject 8

1. Subproject 8 is being set up as a means to continue the present work in the general field of L.S.D. at [redacted] B until 11 September 1954.

2. This project will include a continuation of a study of the biochemical, neurophysiological, sociological, and clinical psychiatric aspects of L.S.D., and also a study of L.S.D. antagonists and drugs related to L.S.D., such as L.A.E. A detailed proposal is attached. The principle investigators will continue to be [redacted] C all of [redacted] B

3. The estimated budget of the project at [redacted] B is $39,500.00. The [redacted] B will serve as a cut-out and cover for this project and will furnish the above funds to the [redacted] B as a philanthropic grant for medical research. A service charge of $790.00 (2% of the estimated budget) is to be paid to the [redacted] B for this service.

4. Thus the total charges for this project will not exceed $40,290.00 for a period ending September 11, 1954.

5. [redacted] C (Director of the hospital) are cleared through TOP SECRET and are aware of the true purpose of the project.

for [redacted] A
Chemical Division/TSS

APPROVED:

[redacted]

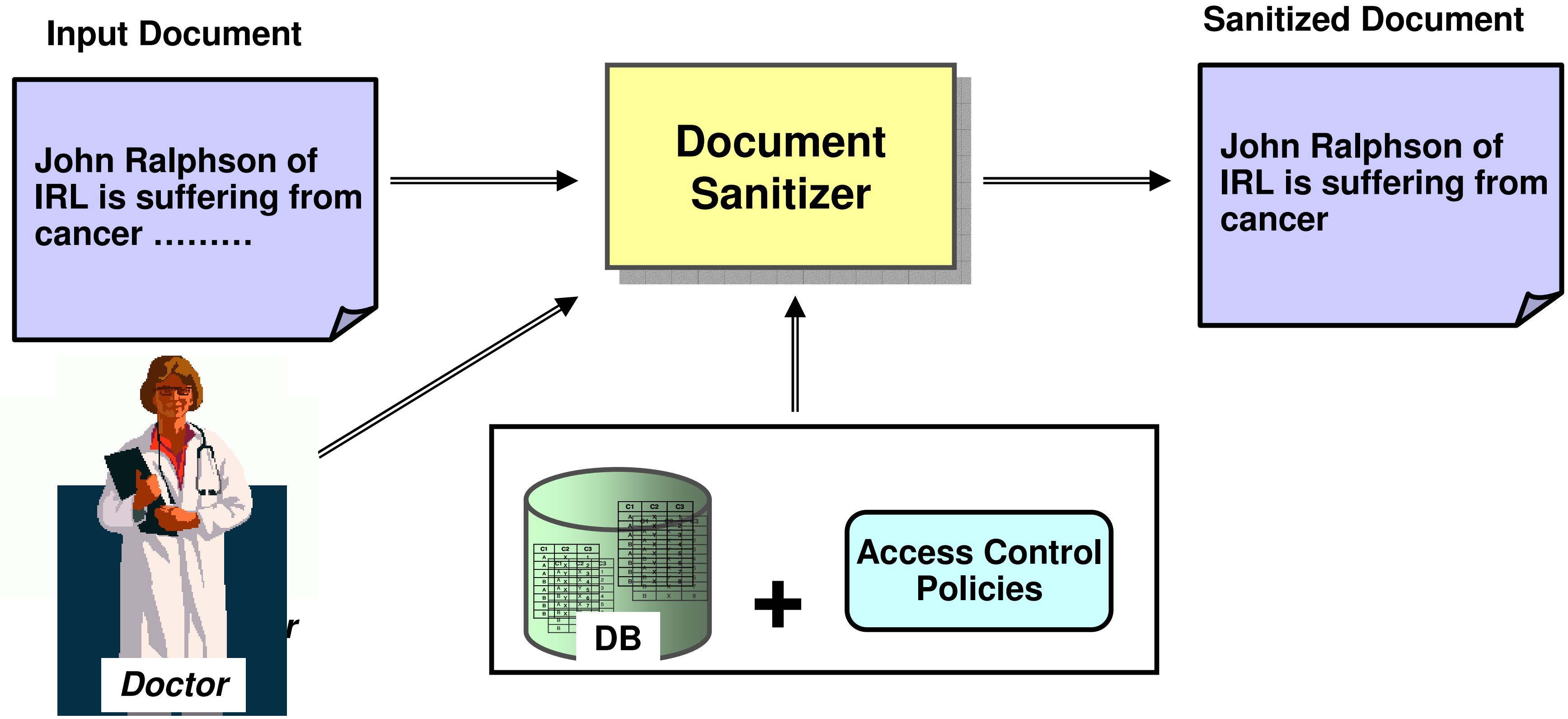
Chief, Chemical Division/TSS

[redacted] PROGRAM [redacted]

CIKM 2008

# Problem Scenario 1: Document Sanitization Based on Access Control Policies

- Dynamically sanitize a document for a specified user, based on his/her access privileges defined on a structured database.
  - Sensitive information hidden from the user in the database should also be removed from the document



Input Document

John Ralphson of IRL is suffering from cancer .........

Document Sanitizer

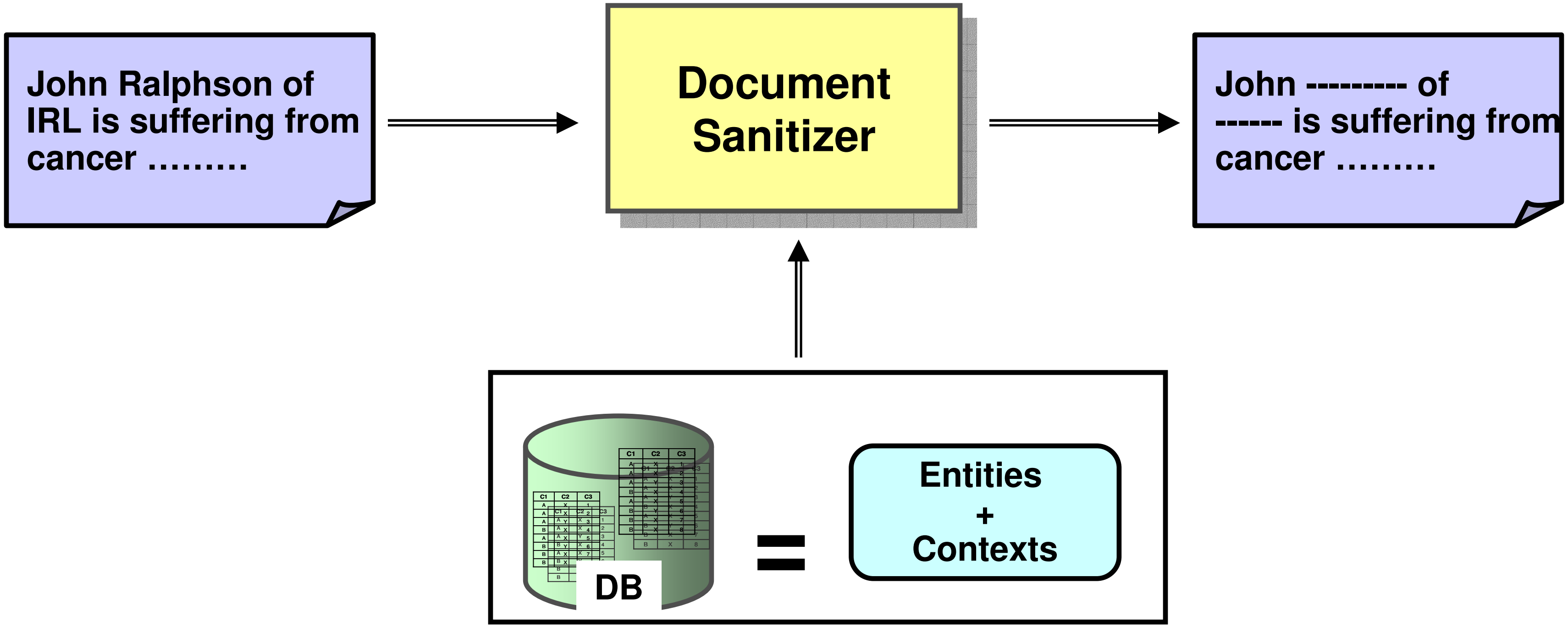
Sanitized Document

John Ralphson of IRL is suffering from cancer

Access Control Policies

# Problem Scenario 2: Document Sanitization for Securing Entities

- Database contains a set of entities
- Each entity e has context C(e) : a set of terms associated with e.
- Sanitization → Hide information from a document so that the entity mentioned in the document cannot be identified.
- Identification: Happens by searching the database and matching terms.



John Ralphson of IRL is suffering from cancer ………

Document Sanitizer

John --------- of ------ is suffering from cancer ………

DB

=

Entities + Contexts

# SNAzzy (Social Network Analysis for Telecom Business Intelligence)

TKDE 2008, WWW 2007

# Technology Overview

- **Goal**
  - Augment the traditional analysis generally utilized by Telcos with Social Network Analyses for improved CRM and Business Intelligence

- **Methodology**
  - Analysis of call and SMS patterns to create a graph where the vertices in the graph represent phone #s, individuals, geographical areas, or communities and the edges in the graph represent their relationships (call duration, friend, acquaintance,…)

- **Focus**
  - **Global Structure Analysis:** Analyzing Call Graphs to help you better understand the underlying behavior of users, in a global context and accordingly plan the marketing services
  - **Customer Social Analysis**: Analyzing the graph to identify the customers of high social influence who should be retained
  - **Churn Analysis:** To predict churn and perform targeted marketing campaigns on potential churners and to identify potential acquisition targets from competitors
  - **Psychographic Analysis:** Analyze the calling patterns to guess the subscriber profile to enable effective customer segmentation leading to targeted campaign management.
  - **Community Discovery:** Analyze the graph to detect communities for improved group targeting and retention

# Case Study

- **Customer: A major Telecom Operator of the world**
- Data:
  - 1 month CDRs (both SMS & Calls) of multiple regions
  - Huge amount of data
    - Graph of 7 million nodes and 35 million edges for 1 region

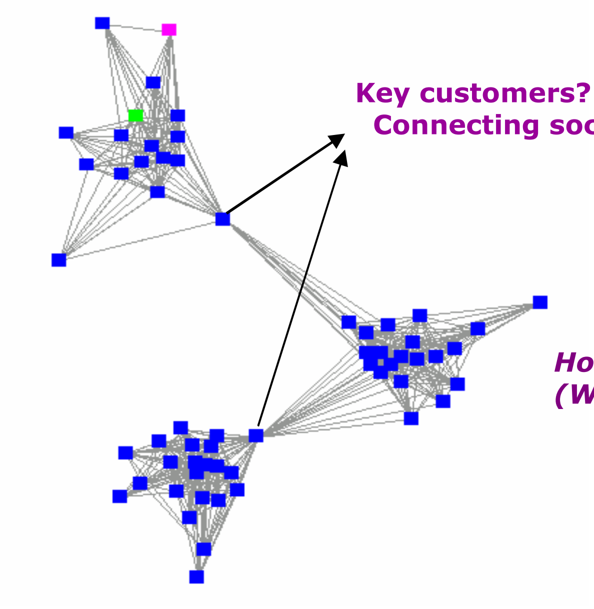
- The CDR graph showed various insights
  - A heavy tailed distribution – very few people know a lot of people, and most people know a few people
  - SMS  is an important medium of communication among certain customer segments
  - SMS is a more social phenomenon than Voice Calls
  - Both SMS graph and Call graph has a very large strongly connected components
    - Can reach a majority of people by traversing links

- Identified various types of interesting communities (see next slide)
  - Cliques (everyone knows everyone else)
  - Clique connectors (people connecting multiple cliques)
  - Competitor's customers as part of cliques signifying that they are external members of a strong community
  - Stars

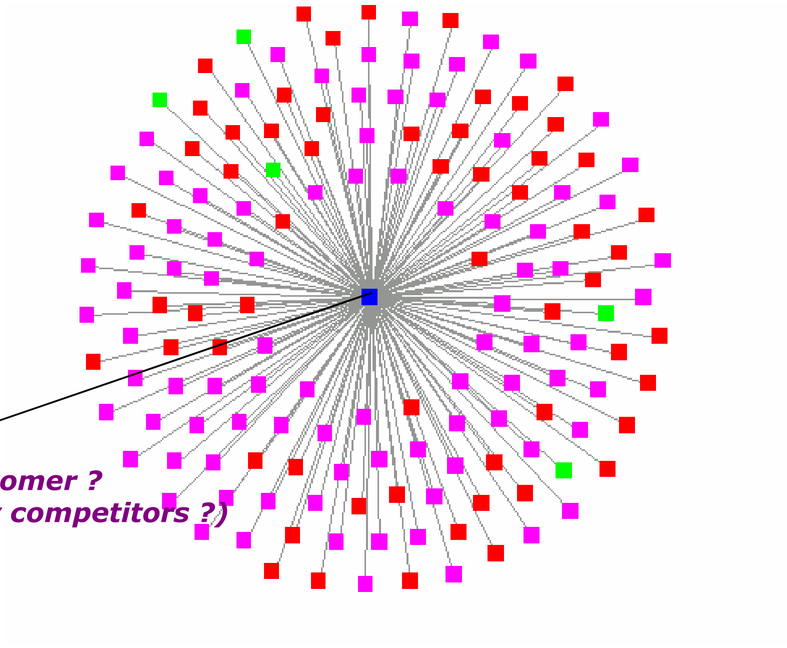
- Analysis shows that people called by churners are more likely to churn

- **In consultation with the Telecom Operator BI and Marketing team to gain more insights on the identified communities and devise campaigns**
  - **GOAL: Integrate SNA with Operator BI System**

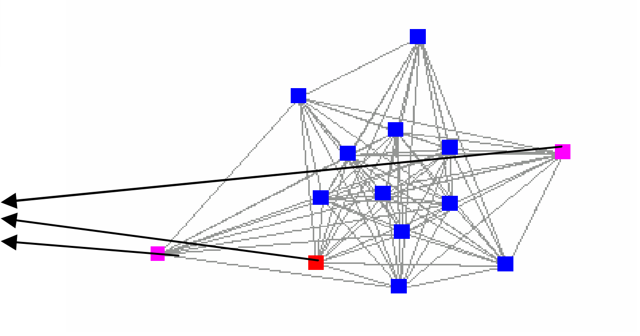
# Some Identified Communities



Key customers?
Connecting social groups



*How do we retain this customer ?*
*(Why is he calling so many competitors ?)*



Potential conversion
targets from
competitors?

🟦 Platinum Customer
🟩 Regular Customer
🟪 CDMA Competitor
🟥 GSM Competitor

# Conclusions

- Information Integration has become widely popular – However a lot needs to be done

- As per Gartner: 80% of data in enterprise is unstructured.  Data which is not integrated with the structured data in the enterprise!

- Huge need for new ways of doing information integration

- Context Oriented Information Integration
  - SCORE: Automatically finds relevant unstructured data for a SQL Query
  - EROCS: Finds links between structured and unstructured data

- Social Network Analysis for Telecom BI
  - Improves churn prediction by analyzing the social behavior of callers, that is, who is calling whom and their calling patterns.
  - Need to consider SNA fact particularly in churn analysis and BI

धन्यवाद

Hindi

# Thank You