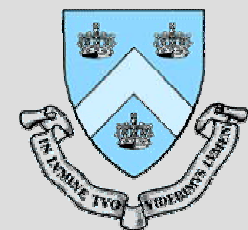




Simplifying Data Access: The Energy Data Collection Project

José Luis Ambite
University of Southern California
Information Science Institute



The Vision: Ask the Government...



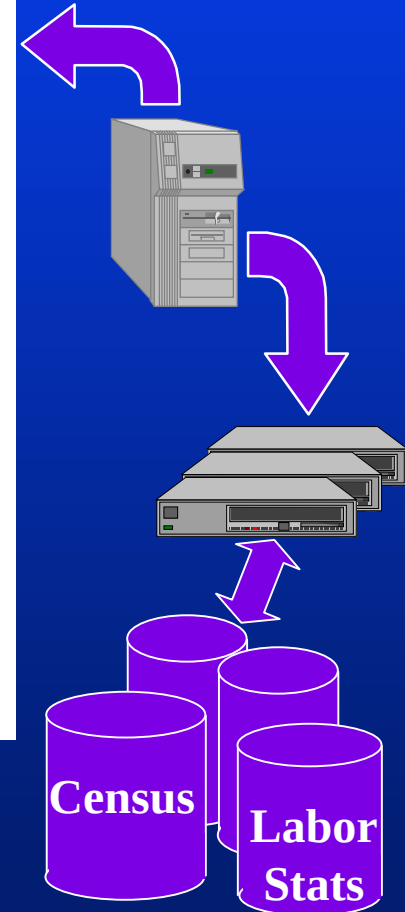
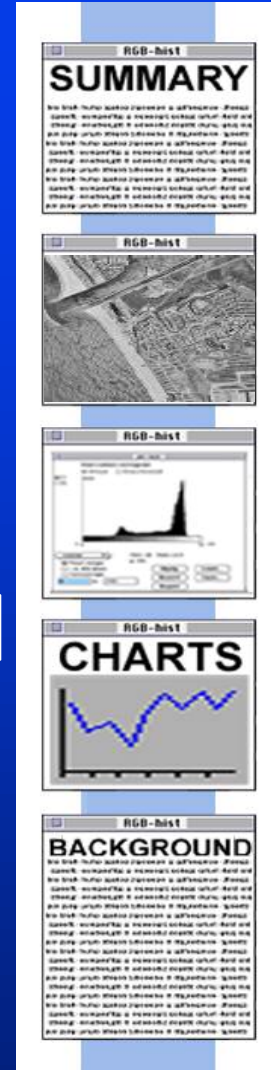
We're thinking of moving to Denver...What are the schools like there?

Query results

How many people had breast cancer in the area over the past 30 years?

Is there an orchestra?
An art gallery? How far are the nightclubs?

How have property values in the area changed over the past decade?



The problem and the solution

- **Problem: FedStats has thousands of databases in over 70 Government agencies:**
 - ◆ data is overlapping, sometimes near-duplicated
 - ◆ even Government officials cannot find it!

- **Solution: Create a system to provide easy standardized access:**
 - ◆ need multi-database access engine,
 - ◆ need powerful user interface,
 - ◆ need terminology standardization mechanism.

Research challenges

- How do you scale to incorporate many databases?
... try to build data models automatically
- How do you integrate their data models, to allow querying across sources and agencies?
... take a large ontology, link the models into it automatically, and provide query reformulator
- How do you incorporate additional information that is available from text sources?
... use language processing tools to extract it
- How do you handle footnotes in the databases?
... extract them from the tables automatically

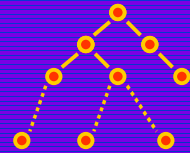
System Architecture

Construction phase:

- Deploy DBs
- Extend ontol.

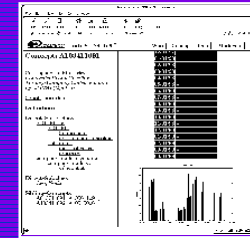
Integrated ontology

- global terminology
- source descriptions
- integration axioms



User Interface

- ontology browser
- query constructor

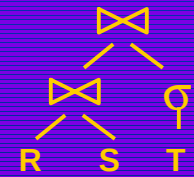


Domain modeling

- DB analysis
- text analysis

Query processor

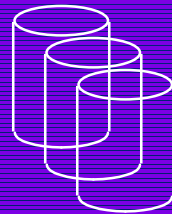
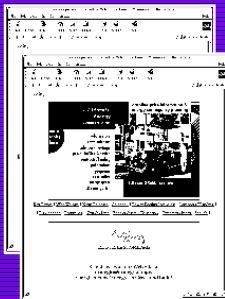
- reformulation
- cost optimization



User phase:

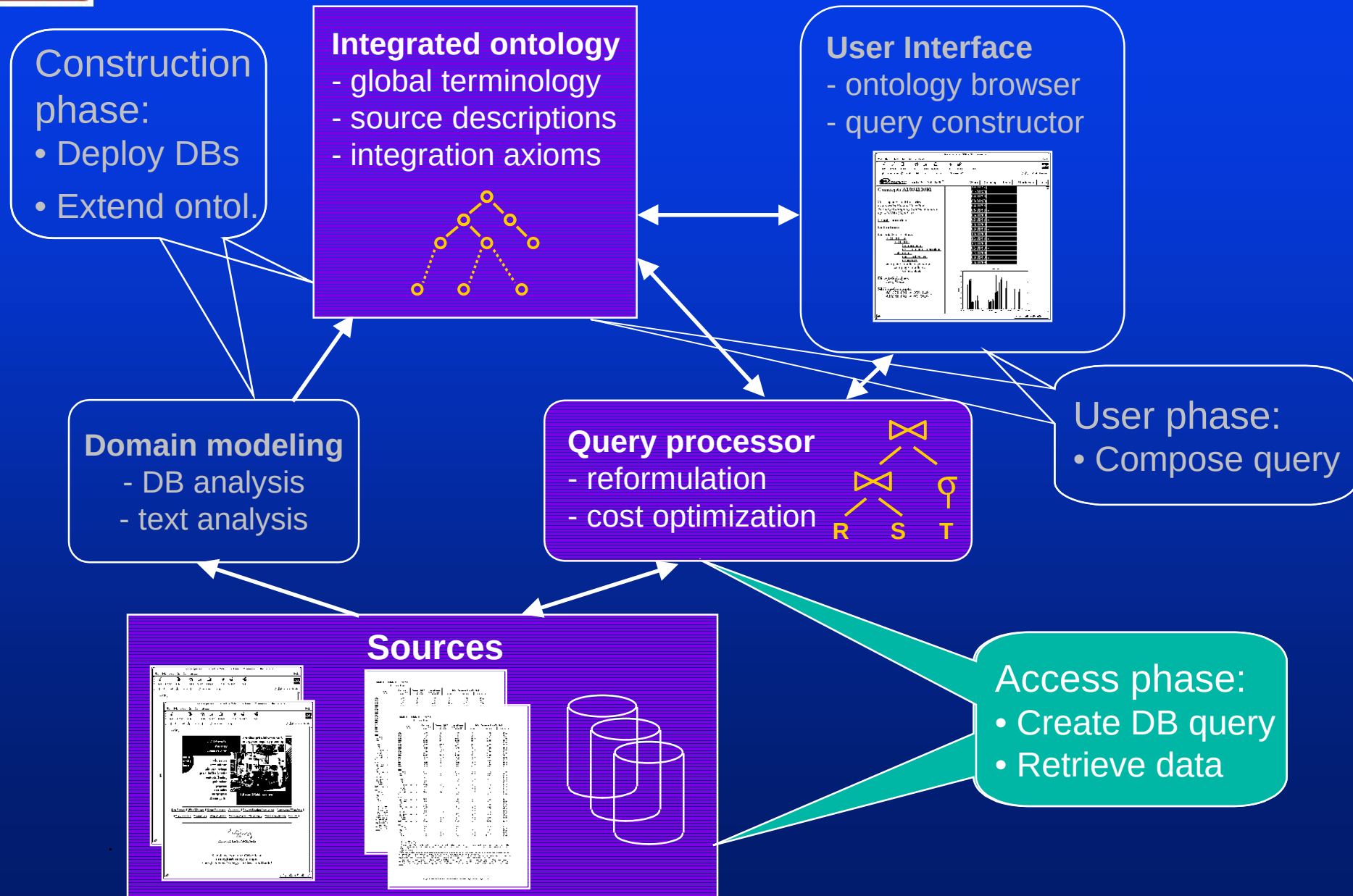
- Compose query

Sources



Access phase:

- Create DB query
- Retrieve data



Information Integration in SIMS

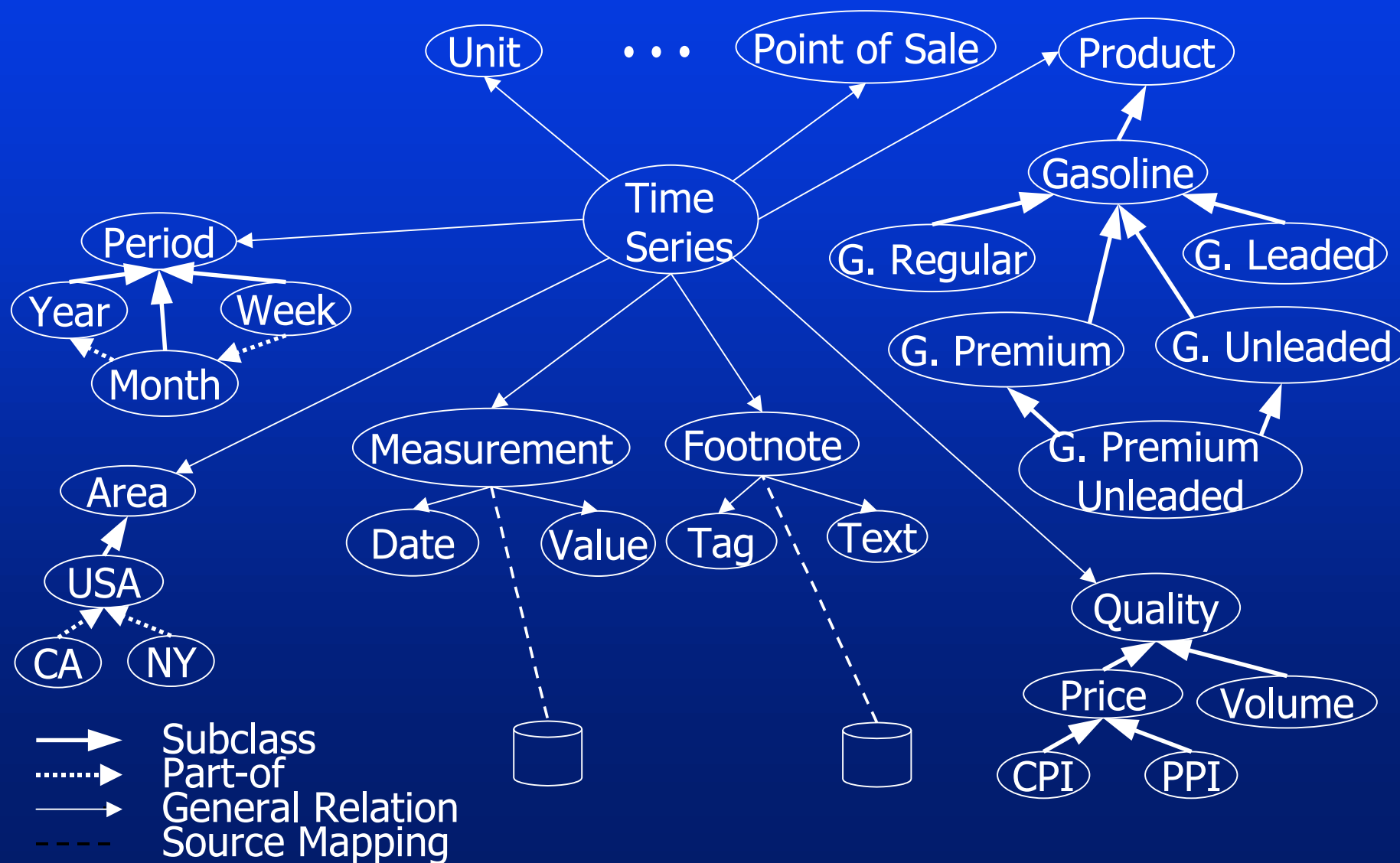
[Arens+96]

[Ambite+00]

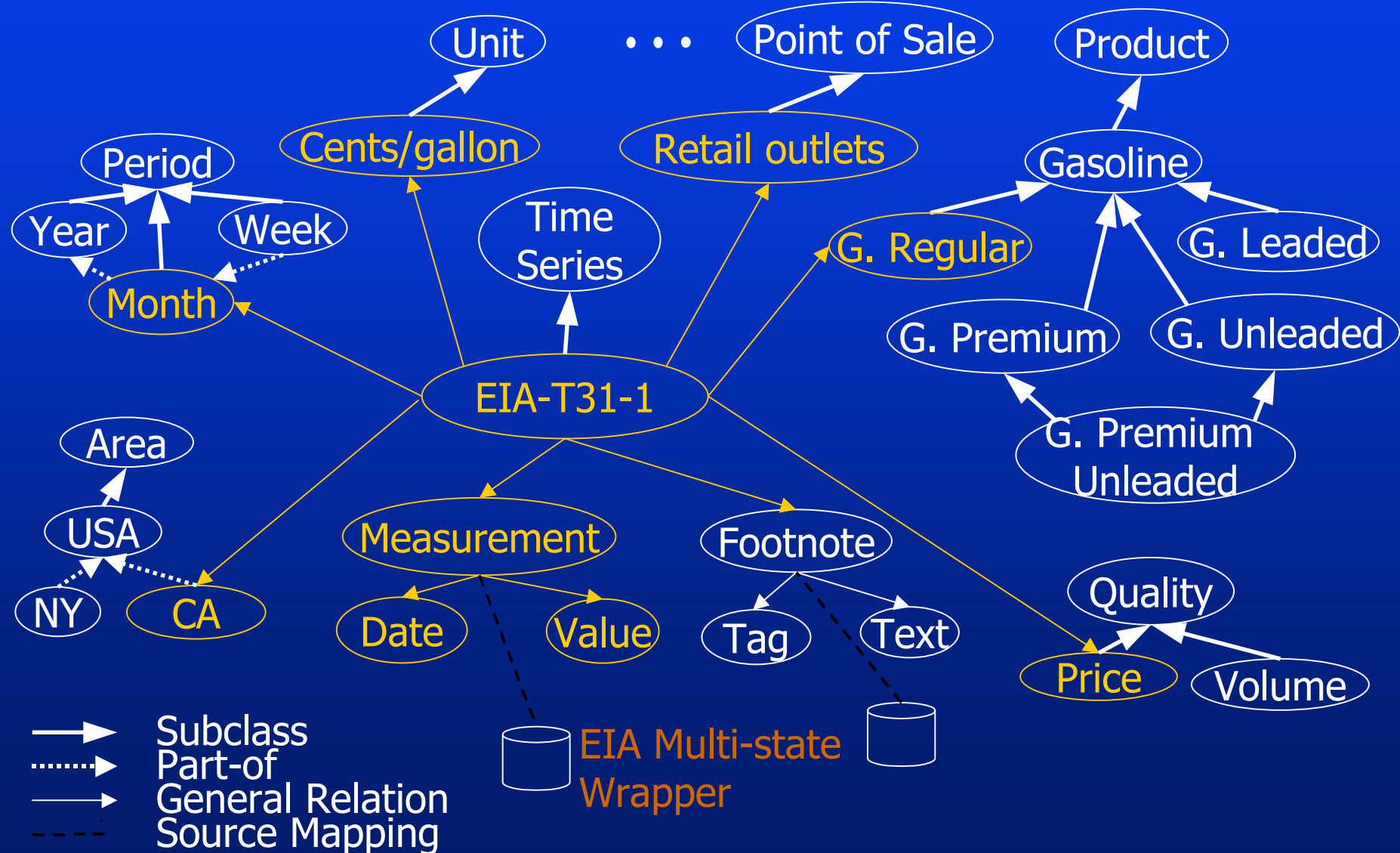
To enable query access, SIMS needs to:

- **address semantic heterogeneity:**
=> describe sources in common *domain model*
- **address syntactic (format) heterogeneity:**
=> standardize access to sources:
 - ◆ Structured (DBMS): Oracle, MS Access ...
 - ◆ Semistructured: *wrappers* for html, text, pdf

Domain Model for Time Series



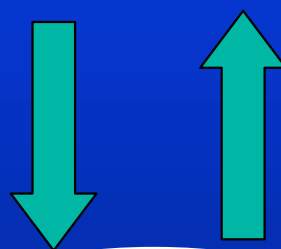
Sales to end users through retail outlets
Measured in cents per gallon



Wrappers

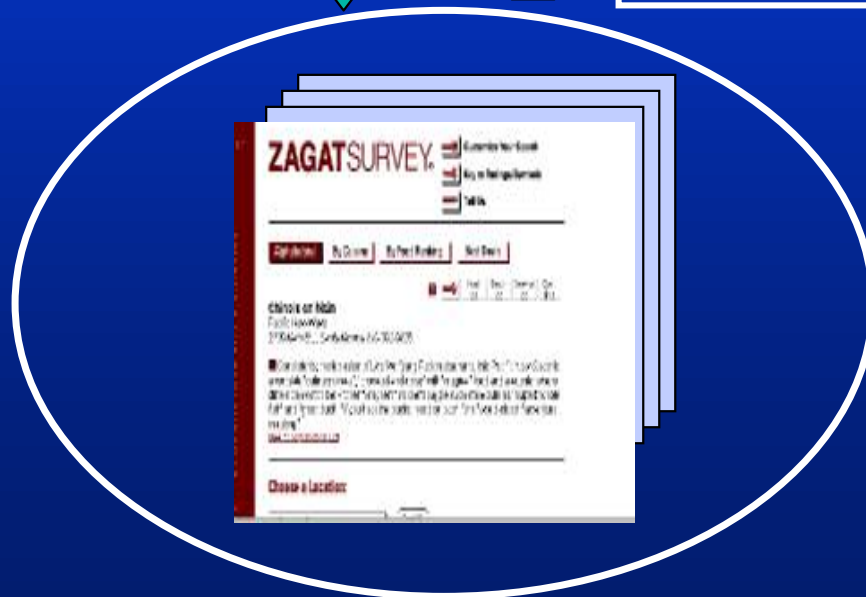
provide uniform mechanism for extracting data from **semi-structured** sources (HTML, text, ...)

Restaurants in
Santa Monica?



Name	Address
Chinois on Main	2709 Main St.
Chao Dara	13 Union Sq
...	...

Wrapper



Wrapper Building Tools

[Muslea+99]

■ Creating Wrappers (semi-)automatically:

- ◆ Demonstration-oriented interface enables users to show system what to extract *by example*
- ◆ System automatically induces extraction rules
- ◆ Common extraction engine

■ Benefits:

- ◆ Rapid wrapper creation
- ◆ Simplified wrapper maintenance

■ Fetch.com

- ◆ Start-up that comercializes the technology

Wrapper Building Tools

Wrapper Builder - detail.trn - [Teaching and Learning]

File Edit Mark-Up Tools View Window Help

New Open Save Fetch Define Teach Learn Verify Link

detail

- Page3.html
 - 3
 - Mostly Cloudy
 - 3
 - 11
 - 2
 - 80%
 - NE/3 km/h
 - 6 km
 - 6
 - 1022.00 mb

Yahoo! Weather - Beijing

[Weather](#) > [Asia](#) > [China](#) > Beijing

[Add to My Yahoo!](#)

Today		F° or C°	Sat	Sun	Mon	Tue
-3°	Mostly Cloudy	Hi 3° at: 10:00pm GMT LO -11°	Wind Chill: -2° Humidity: 80% Wind: NE/3 km/h Visibility: 6 km Dewpoint: -6° Barometer: 1022.00 mb	Partly Cloudy	Partly Cloudy	Partly Cloudy
<10 -10 -5 0 5 10 15 20 25 30 35+		Records & Averages		Hi 8 Lo -6	Hi 8 Lo 1	Hi 10 Lo -4

Current conditions and forecasts provided by Weathernews Inc.

Node Name	Rule Type	Rule Content
temp	Extract	Begin_Rule [0] <code>_FE_</code> End_Rule [1] <code>__ST__ 2 " face = " Arial " >
 &nbsp; <nobr></code>
outlook	Extract	Begin_Rule [0] <code>_FE_</code> End Rule FE

Pages Rules Messages

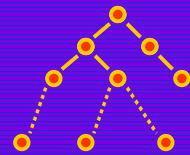
Project: weather; Site: yahoo Markup samples and train. Created on 2/16/2001

Construction phase:

- Deploy DBs
- Extend ontol.

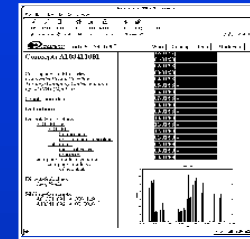
Integrated ontology

- global terminology
- source descriptions
- integration axioms



User Interface

- ontology browser
- query constructor

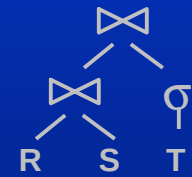


Domain modeling

- DB analysis
- text analysis

Query processor

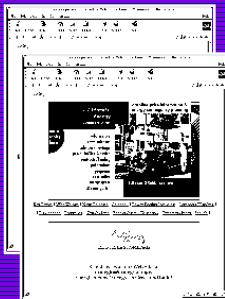
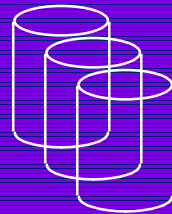
- reformulation
- cost optimization



User phase:

- Compose query

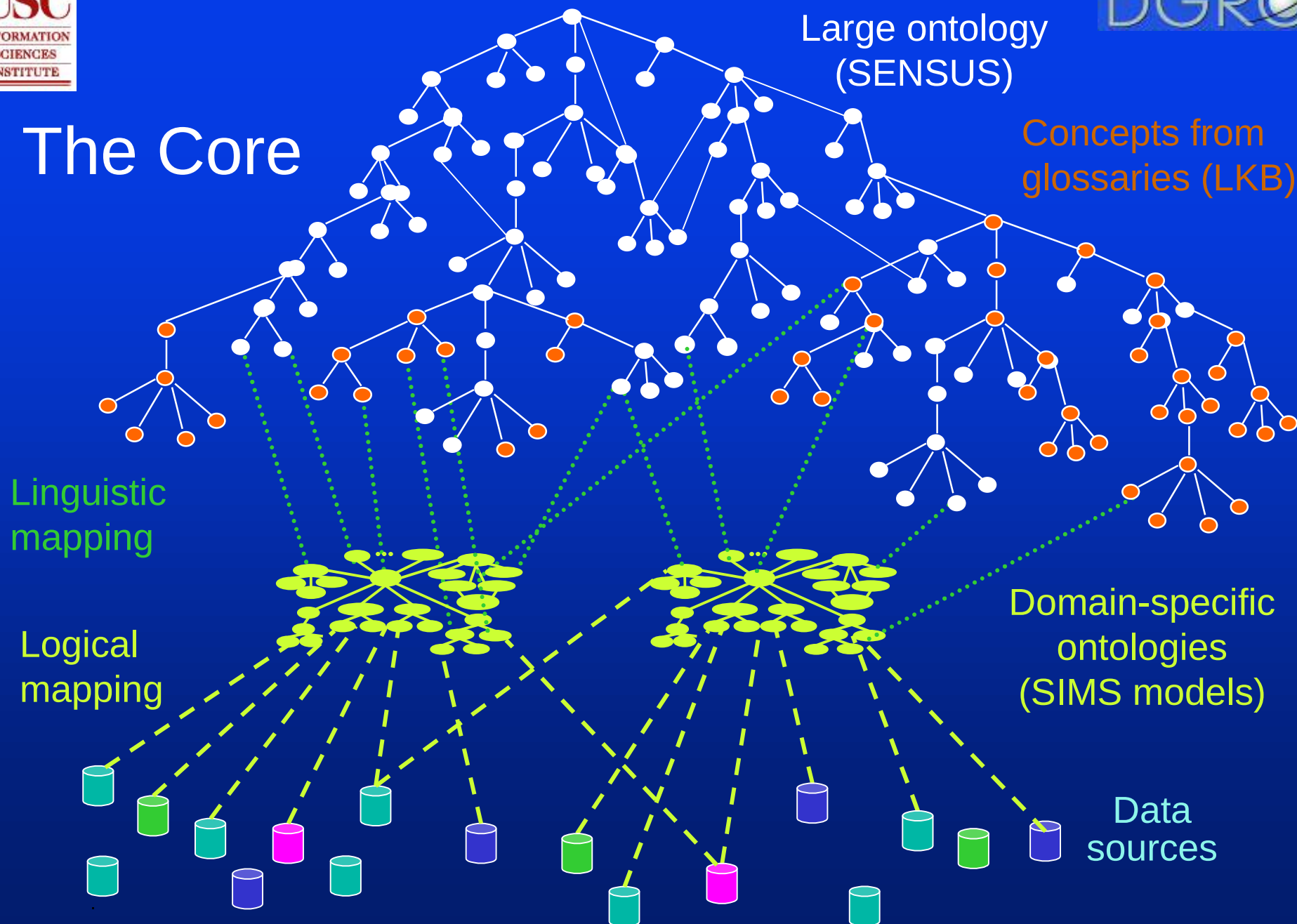
Sources

Access phase:

- Create DB query
- Retrieve data

The Core



Ontology, models, data

- SENSUS ontology (90000 concepts)
- LKB glossary extracts (6000 concepts)
 - ◆ Purpose: provide rich domain knowledge
 - ◆ Sources: Census, EIA, EPA texts
- SIMS domain model (500 dimension conce, 50000 series)
 - ◆ Purpose: describe data, query planning and optimization
 - ◆ Sources: database metadata, extracted by hand
- Database sources (25000 time series)
 - ◆ EIA OGIRS: CD with MS Access database; no footnotes
- Web data sources (25000 time series)
 - ◆ EIA Multi-State: 25000 time series; formatted text
 - ◆ BLS: 60 time series; web form; 40 have footnotes
 - ◆ EIA Petroleum Supply Monthly: 20 time series in PDF; converted to HTML and then wrapped; all with footnotes
 - ◆ CEC: 12 series; static HTML tables; no footnotes

[Knight+95]
[Hovy+00]

- Taxonomy, multiple superclass links
- Approx. 90,000 concepts
- Top level: Penman Upper Model (ISI)
- Body: WordNet (Princeton), rearranged
- Used at ISI for machine translation, text summarization, database access

Extracting metadata from text

[Klavans+01]

Problems:

- ◆ Proliferation of terms in domain
- ◆ Agencies define terms differently
- ◆ Many refer to the same or related entity
- ◆ Lengthy term definitions often contain important information which is buried

Example input:

Motor Gasoline Blending Components: Naphthas (e.g., straight-run gasoline, alkylate, reformate, benzene, toluene, xylene) used for blending or compounding into finished motor gasoline. These components include reformulated gasoline blendstock for oxygenate blending (RBOB) but exclude oxygenates (alcohols, ethers), butane, and pentanes plus. Note: Oxygenates are reported as individual components and are included in the total for other hydrocarbons, hydrogens, and oxygenates.

Lexical Knowledge Base (LKB) Tool

Definition:

Regular Gasoline: Gasoline having an antiknock index, i.e., octane rating, greater than or equal to 85 and less than 88. Note: Octane requirements may vary by altitude. See Gasoline Grades.

Analysis

Head Term:

Regular Gasoline

Cross-reference

Gasoline Grades

Genus Term:

Gasoline

Head Genus Word:

Gasoline

Properties:

having : an antiknock index

Excludes-Includes:

Quantifiers:

less than 88

Note: Octane requirements may vary by altitude. See Gasoline Grades.

```
(Regular Gasoline
  (isa "Gasoline")
  (contains "an antiknock index")
)
```

Combines statistical and linguistic methods:

- trainable Finite State Recognizer
- identifies topics with high accuracy
- provides complete coverage
- useful for any subject area
- produced over 6,000 gasoline concepts from EIA and BLS

Linking Domain Models to Ontology

[Hovy+01]

Match DM concepts against SENSUS concepts

1. Name Match

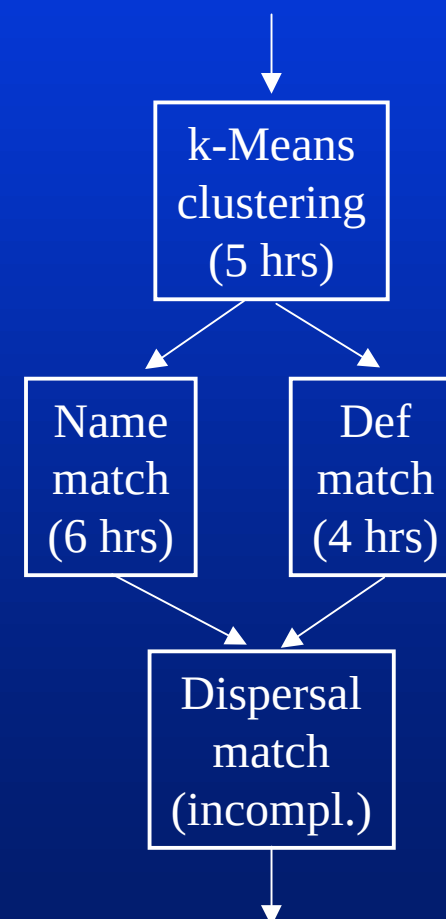
- Variations: substrings; reward for contiguous endings; stemming; variant word forms; case sensitivity; etc.
- Implemented string match from molecular biology for DNA matching

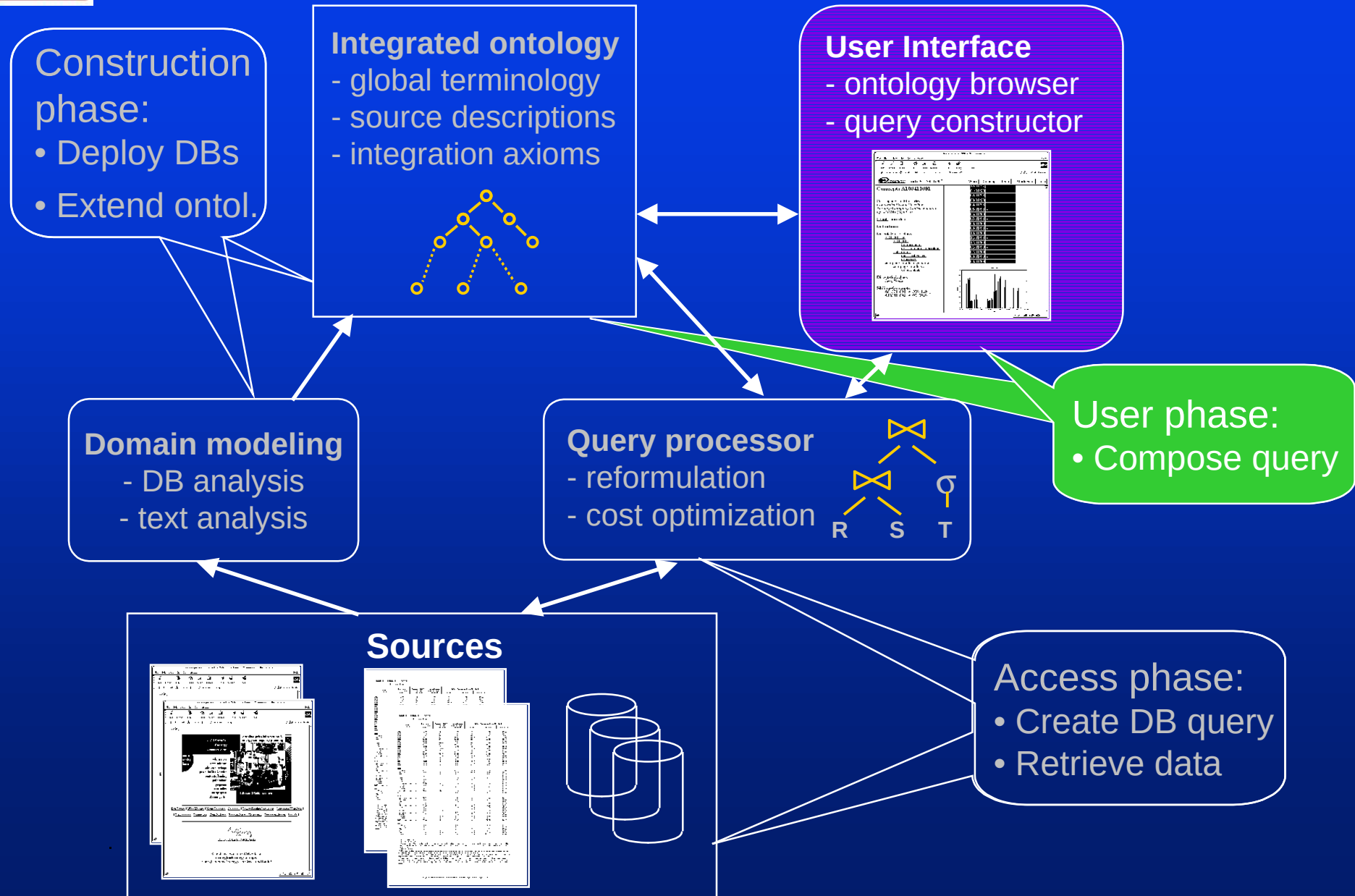
2. Definition Match

- Variations: stemming; stop words; words or letter trigrams; word weights (*itf*, *tf.idf*), etc.
- Used IR vector space; cosine measure

3. Dispersal Match

- Def: Given a cluster of concepts, match them all and then select one match for each concept to form tightest cluster in SENSUS
- Variations: greedy search; number of candidates; weightings





Select Attributes and Sub-Attributes

- agency Choose value
- area Choose value
- family Choose value
- frequency Choose value
- point of sale Choose value
- product Choose value
- provenance Choose value
- quality Choose value
- seasonal adjustment Choose value
- unit Choose value

Definition

Submit Query Reset

Suggested Queries 2. PPI Revision Curr... ▼

Select Attributes and Sub-Attributes

agency	BLS	Definition Geographical scope of measured data: Los Angeles, CA, USA
area	USA	
family	city ▶ Chicago country ▶ Los Angeles region ▶ Tulsa state ▶	
frequency		
point of sale		
product	unleaded regular	
provenance	BLS	
quality	revision current series	
seasonal adjustment	not seasonally adjusted	
unit	dollars per gallon	
Submit Query Reset		
Suggested Queries 2. PPI Revision Curr... ▼		

History

Query 1

Terms Used

BLS
 USA
 PC
 tuesday 13
 point of sale
 unleaded regular
 BLS
 revision current series
 not seasonally adjusted
 dollars per gallon

Sources

Original Source Data

Results and Footnotes

19881100	53.2
19881200	45.4
19990000	70.3
19990100	49.5
19990200	47.0
19990400	69.4
19990500	71.2
19990600	68.9
19990800	81.8
19990900	85.6
19991000	80.4
19991200	82.6
20000100	83.0(P)
20000200	93.4(P)
20000300	107.4(P)
20000400	101.6(P)

P : Preliminary. All indexes are subject to revision four months after original publication.

[\[Accessibility Information\]](#)



[Information and Help](#)

[BLS Program and Survey Specials](#)

[Series ID Formats](#)

Enter series id(s) below:

PCU2911#134

Producer Price Index Revision-Current Series

Series Catalog:

Series ID : PCU2911#134

Not Seasonally Adjusted
Industry : Petroleum Refining
Product : Unleaded regular
Base Date : 8506

Data:

Year	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Ann
1977						47.6	48.0	47.9	47.8	47.6	47.4	47.4	
1978	47.3	46.9	46.8	46.9	47.4	48.4	49.4	50.5	51.3	51.8	51.7	52.8	49.3
1979	53.5	54.3	55.9	58.0	61.3	64.8	69.0	73.4	77.9	81.3	82.8	84.9	68.1
1980	88.6	93.8	101.5	106.3	108.2	107.7	108.2	108.6	108.7	107.9	108.2	108.7	104.7
1981	110.3	114.7	124.3	126.2	125.3	124.7	123.4	122.2	121.9	121.7	121.3	120.9	121.4
1982	120.1	117.9	114.7	107.5	102.8	107.3	115.3	116.7	115.1	113.2	111.7	110.0	112.7
1983	106.0	101.7	98.0	95.1	98.8	103.3	104.5	105.3	104.4	103.4	101.1	99.2	101.7
1984	96.2	95.3	96.0	96.7	99.0	98.8	96.7	93.8	92.9	94.6	95.2	93.6	95.7
1985	89.5	85.8	87.4	92.4	97.8	100.0	99.1	96.9	94.8	93.1	95.6	95.9	94.0
1986	89.2	79.0	61.9	56.2	62.8	65.0	52.1	51.4	55.7	51.0	51.3	51.8	60.6
1987	57.1	60.7	60.5	64.0	64.4	66.4	68.2	70.7	67.0	65.8	66.5	62.7	64.5

Search For Term: unleaded regular gasoline

Last Term Seen:

Search Back

Search For Term:

Last Term Seen:

Search Back

Pick a Concept

Your search returned several concepts.
Please select one.

product.gasoline.unleaded regular^SIMS : product.gasoline.unleaded regular (from SIMS)

unleaded regular gasoline^EIA GU : unleaded regular gasoline (from EIA GU)

Submit

Warning: Applet Window

Search For Term: Search
Last Term Seen: Back

unleaded regular gasoline (from EIA&G)

Gasoline having an antiknock index (R+M/2) greater than or equal to 85 and less than 88, and containing not more than 0.05 grams of lead or 0.005 grams of phosphorus per gallon. OPI: EI-40 Sources: EIA-782A, EIA-782B, EIA-782C, DOE/EIA-0487

Core Definition

Gasoline having an antiknock index (R+M/2) greater than or equal to 85 and less than 88, and containing not more than 0.05 grams of lead or 0.005 grams of phosphorus per gallon

Source Agency
EIA

unleaded regular gasoline (from EIA&G)

word: unleaded regular gasoline

remainder of definition: OPI: EI-40 Sources: EIA-782A, EIA-782B, EIA-782C, DOE/EIA-0487

see also: product:gasoline.unleaded regular (from SIMS)

source URL: http://www.cs.columbia.edu/~bwhitman/fkb/eia/eia_large_piece.html

source nickname: EIA&G

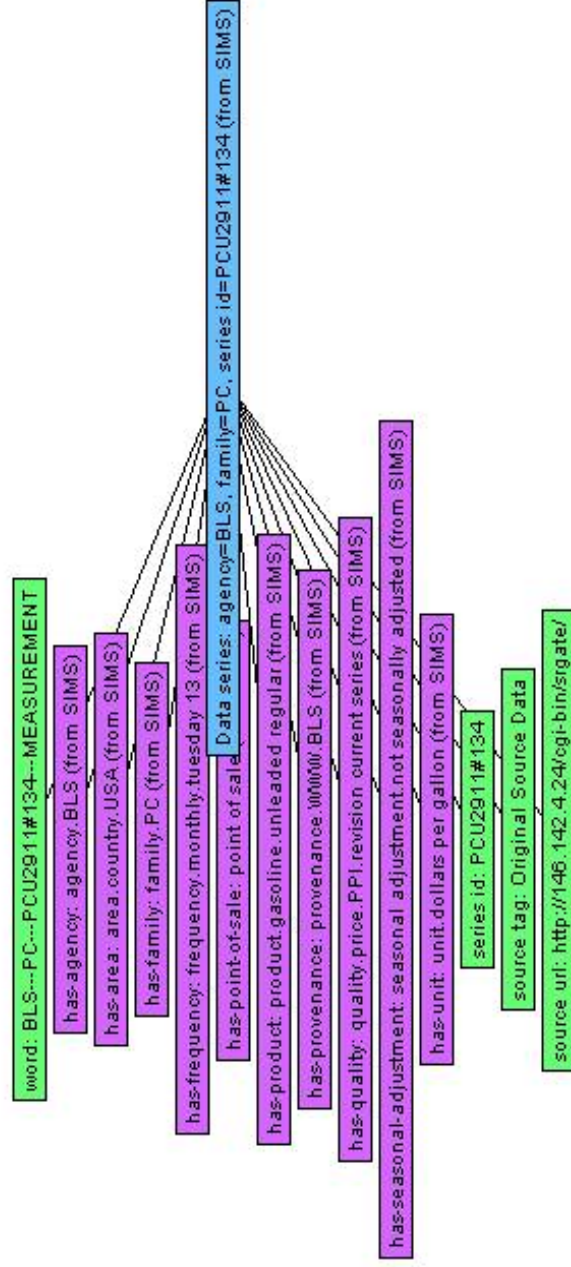
term: unleaded regular gasoline

Search For Term:

Last Term Seen: **product.gasoline.unlead...** Search Back

Data series: agency=BLS, family=PC, series id=PCU2911#134 (from SIMS)

Source PCU2911#134: BLS PC revision current series about gasoline.unleaded regular in USA



Conclusion

Large-scale information integration

■ Major research challenges:

- ◆ Semi-automated domain modeling, information extraction, linkage to ontology
- ◆ Efficient query processing
- ◆ Footnote extraction and display

■ Major practical challenges:

- ◆ Understanding users' needs: interface, information ...
- ◆ Getting a lot of data into the system