Addressing a Bottleneck in Data Integration using Automated Learning Techniques

Eduard Hovy
Jose Luis Ambite
Andrew Philpot

DGRC
Information Sciences Institute
University of Southern California

Talk overview

1. The EDC system, as an example of past work
2. Planned research:
   statistical alignment for machine translation
   statistical alignment for data source mapping

EDC: The Vision: Ask the Government...

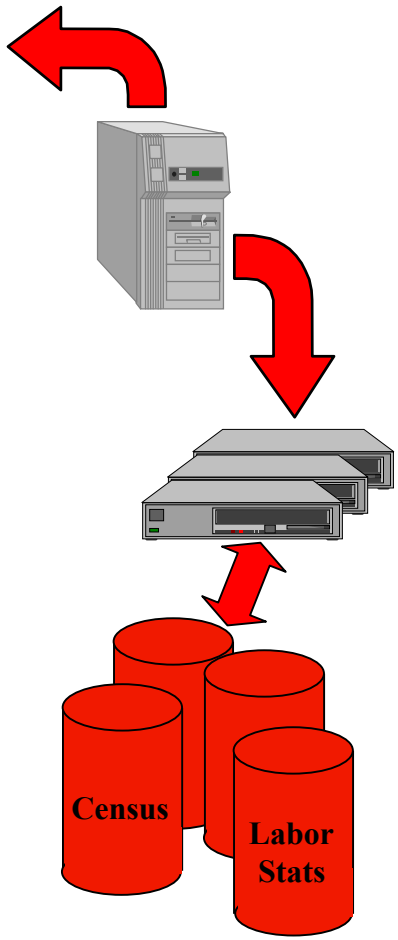
We're thinking of moving to Denver... What are the schools like there?

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

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

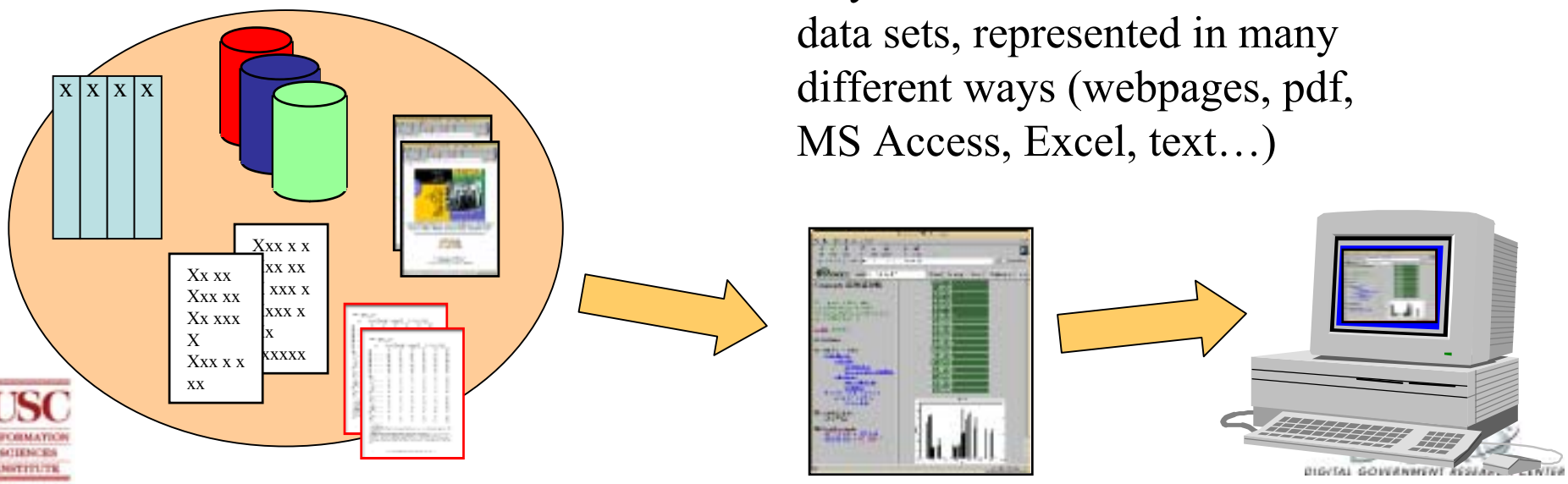
How many people had breast cancer in the area over the past 30 years?

Query results

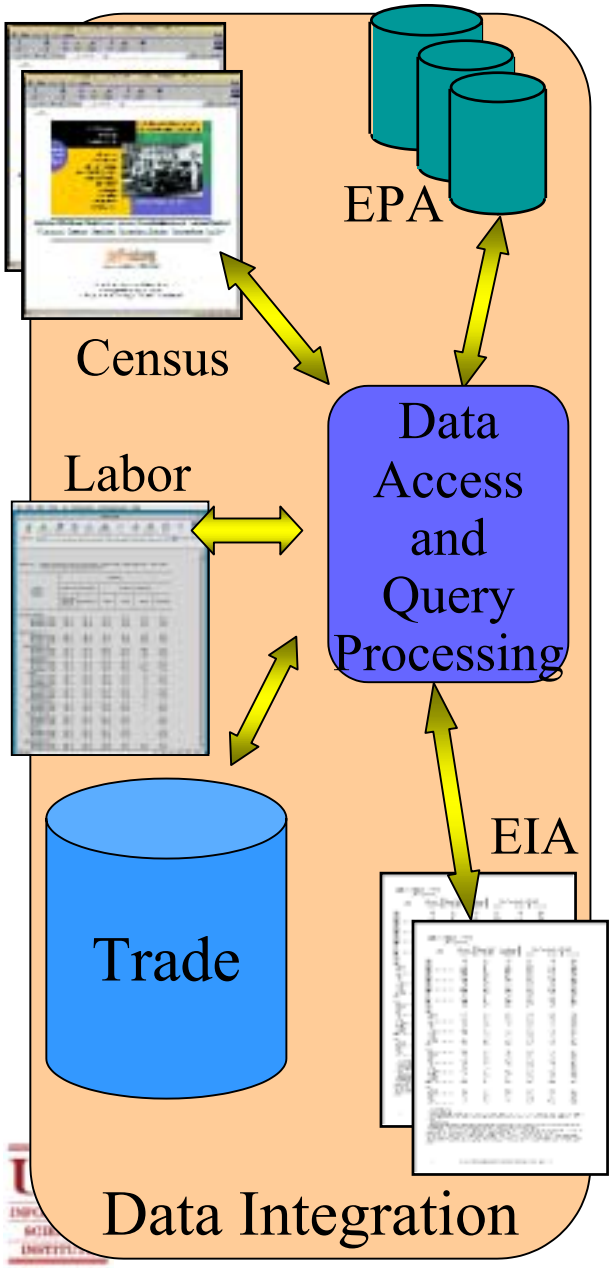


The Energy Data Collection project

- EDC research team
  - Information Sciences Institute, USC
  - Dept of CS, Columbia University
- Government partners
  - Energy Information Admin. (EIA)
  - Bureau of Labor Statistics (BLS)
  - Census Bureau
- Research challenge
  - Make accessible in standardized way the contents of thousands of data sets, represented in many different ways (webpages, pdf, MS Access, Excel, text…)

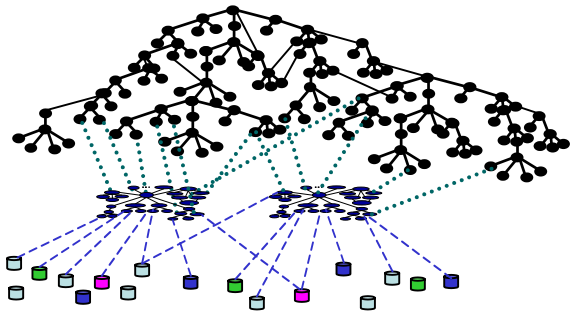


Heterogeneous Data Sources



Information Access

Metadata and Terminology Management

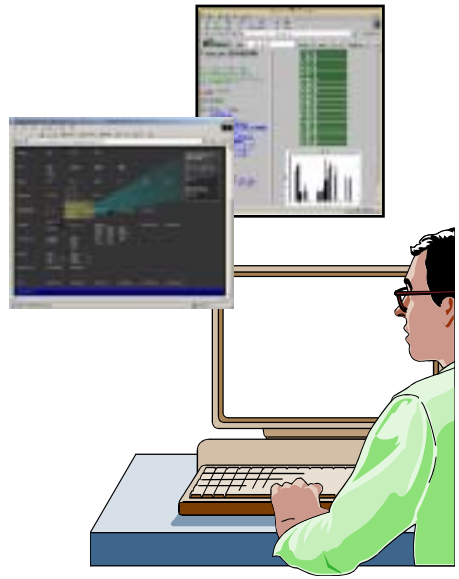


Concept Ontology

Terminology Sources

User Interface

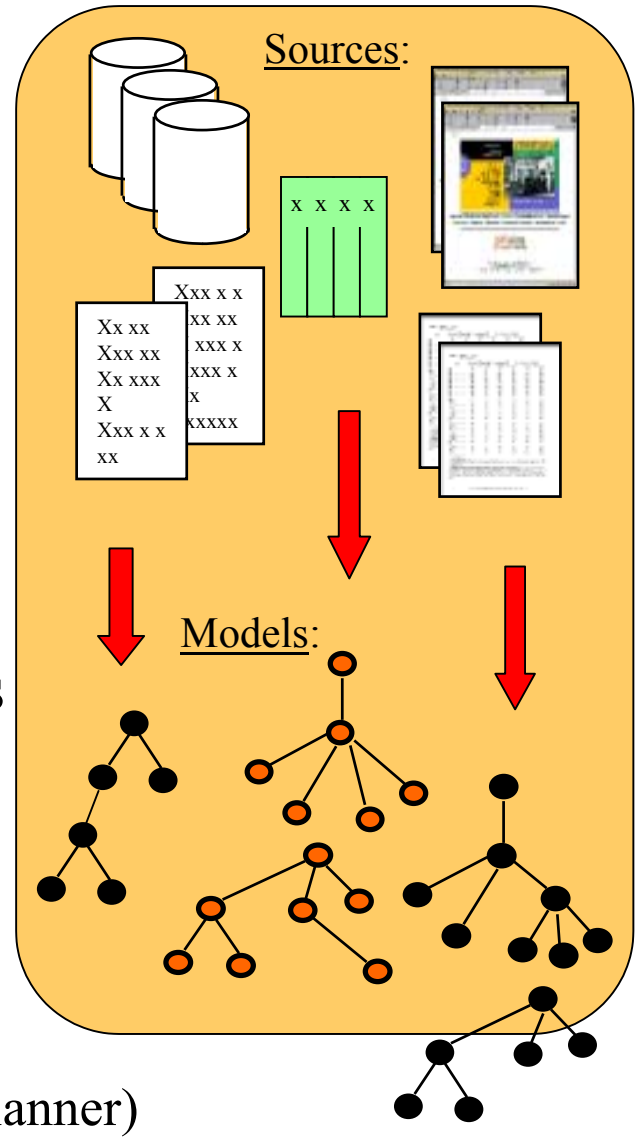
Interface Design and Task-based Evaluation



User Evaluation

Data access using SIMS

- 'Hide' from user details of data sources:
  1. 'Wrap' each source in software that handles access to its data
  2. Record the types of info in each source in a 'Source Model' ≈ metadata schema
  3. Arrange all source models together in the same space—the Domain Model
- SIMS data access planner transforms user's request into individual access queries
- SIMS extracts the right data from the appropriate sources
- Current central domain model:
  - 500 nodes, built manually; for database access planner)
  - LKB: 6000 nodes
    - 1. NL term/info extraction from glossaries (Columbia U)
    - 2. Taxonomization and embedding (ISI)



Wrapping databases

How to wrap sources?

- Manually—for pdf, plain text, footnotes
- Automatically—using parsing rules based on landmarks in html pages (Ariadne project)…this work being incorporated and extended…

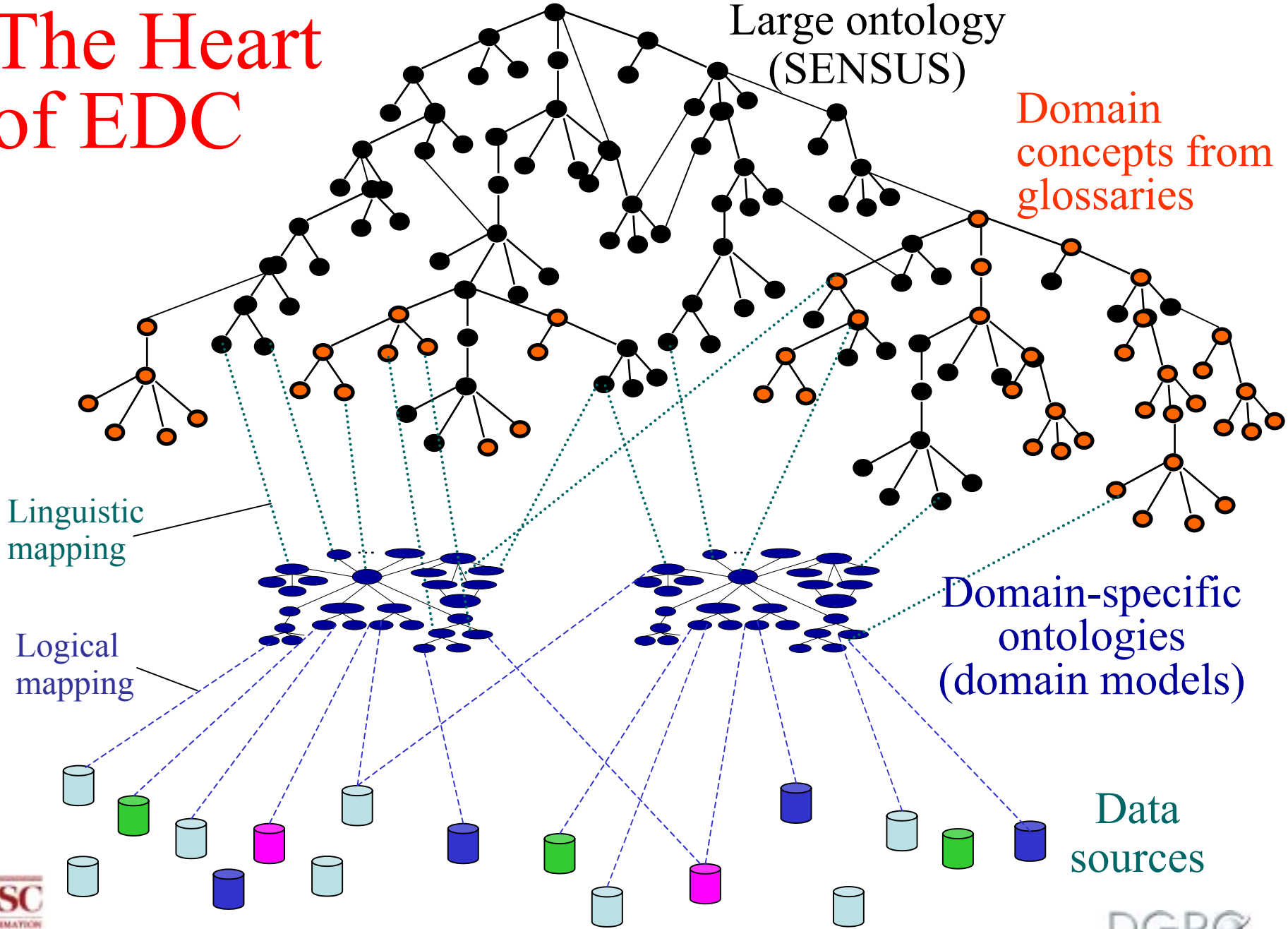
Time series wrapped (W) to date:

- BLS: 58 series, website (html) → (W) ⇒ data is 'live'
- CEC: 3 series, website (html) → (W) ⇒ data is 'live'
- EIA Petroleum Supply Monthly: 16 series, publ. (pdf) → html → (W)
- EIA OGIRS database: 25,000+ series, CD (MS Access) → Oracle → (W)
- EIA tables: 25,000+ series, text tables → (W)

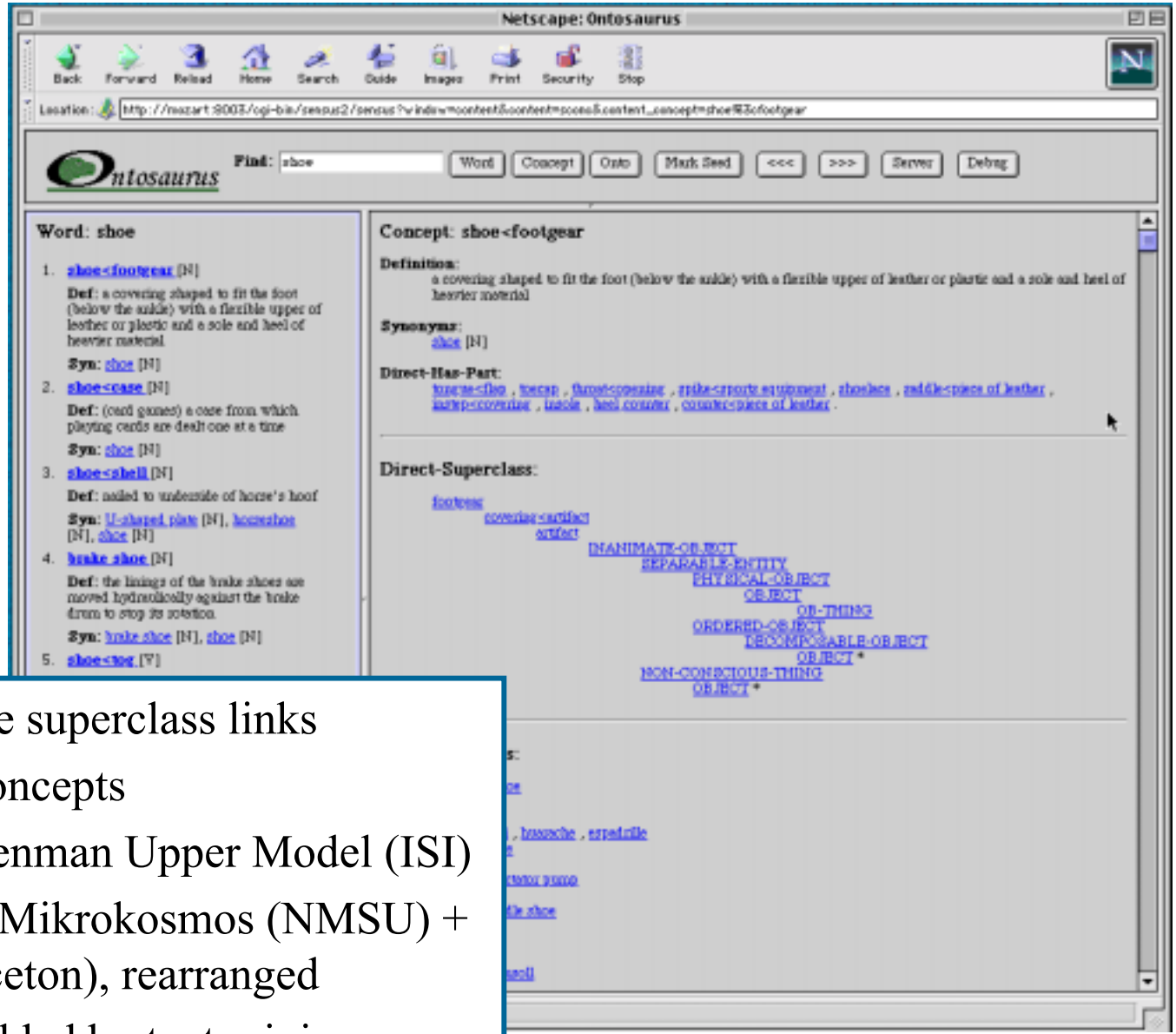
Procedure for most recent set:

- Analyze format; break table into header, body, footnotes
- Manually create 'access paths' for queries (via landmarks)
- Can then extract data and save in any format (html, MS Access, etc.)

The Heart of EDC



Omega ontology and Ontosaurus and DINO browsers



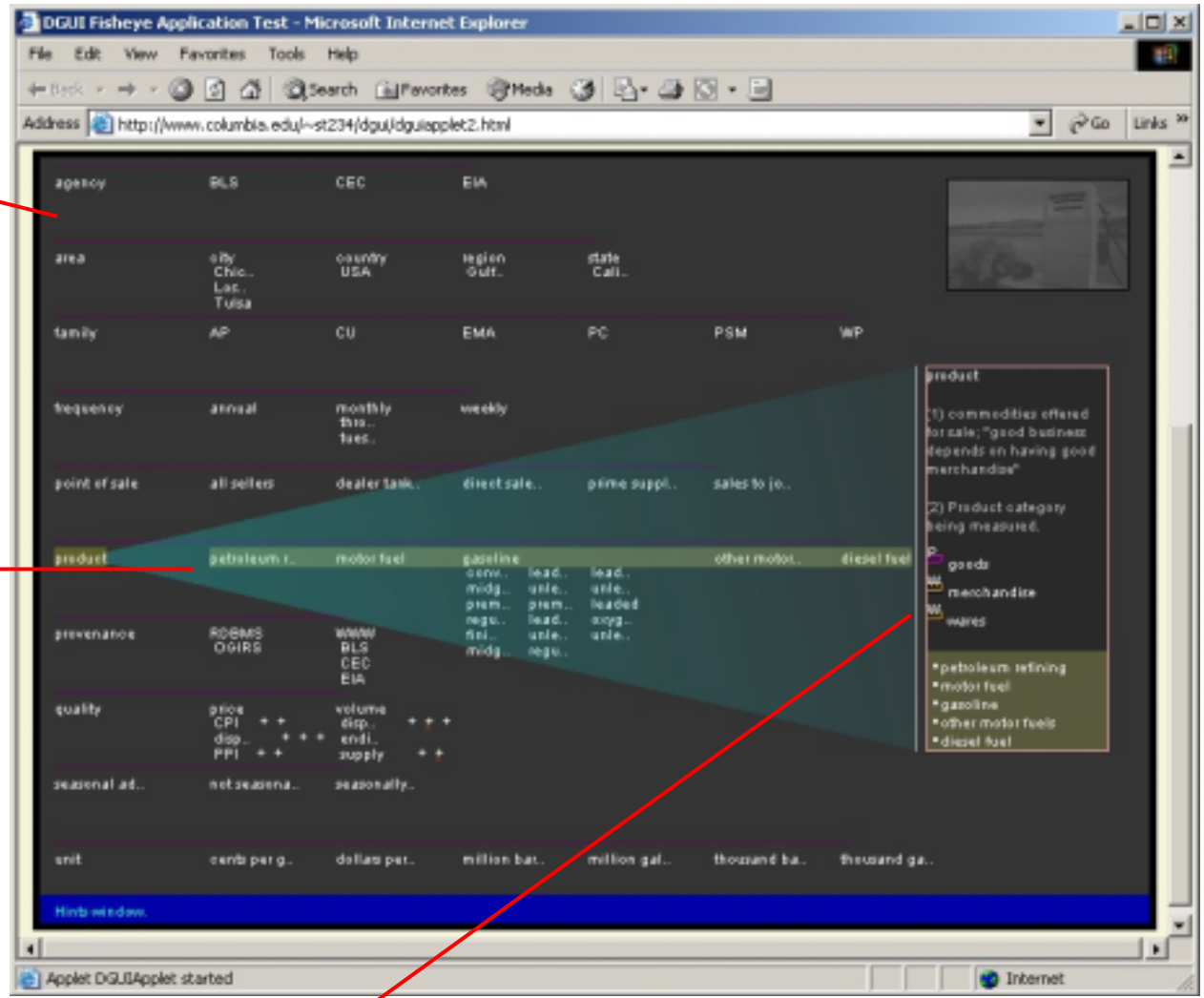
- Taxonomy, multiple superclass links
- Approx. 110,000 concepts
- Top level: new + Penman Upper Model (ISI)
- Body: own work + Mikrokosmos (NMSU) + WordNet 1.7 (Princeton), rearranged
- New information added by text mining
- Used at ISI for machine translation, text summarization, database access

http://mozart.isi.edu:8003/sensus2/

User interface testbed

(Feiner et al., Columbia)

- Menu presented as grid of alternating rows and columns
- Ontology entry shown in beam for selected item
  - Located as near as possible
  - Color coding shows parental and semantic relationships

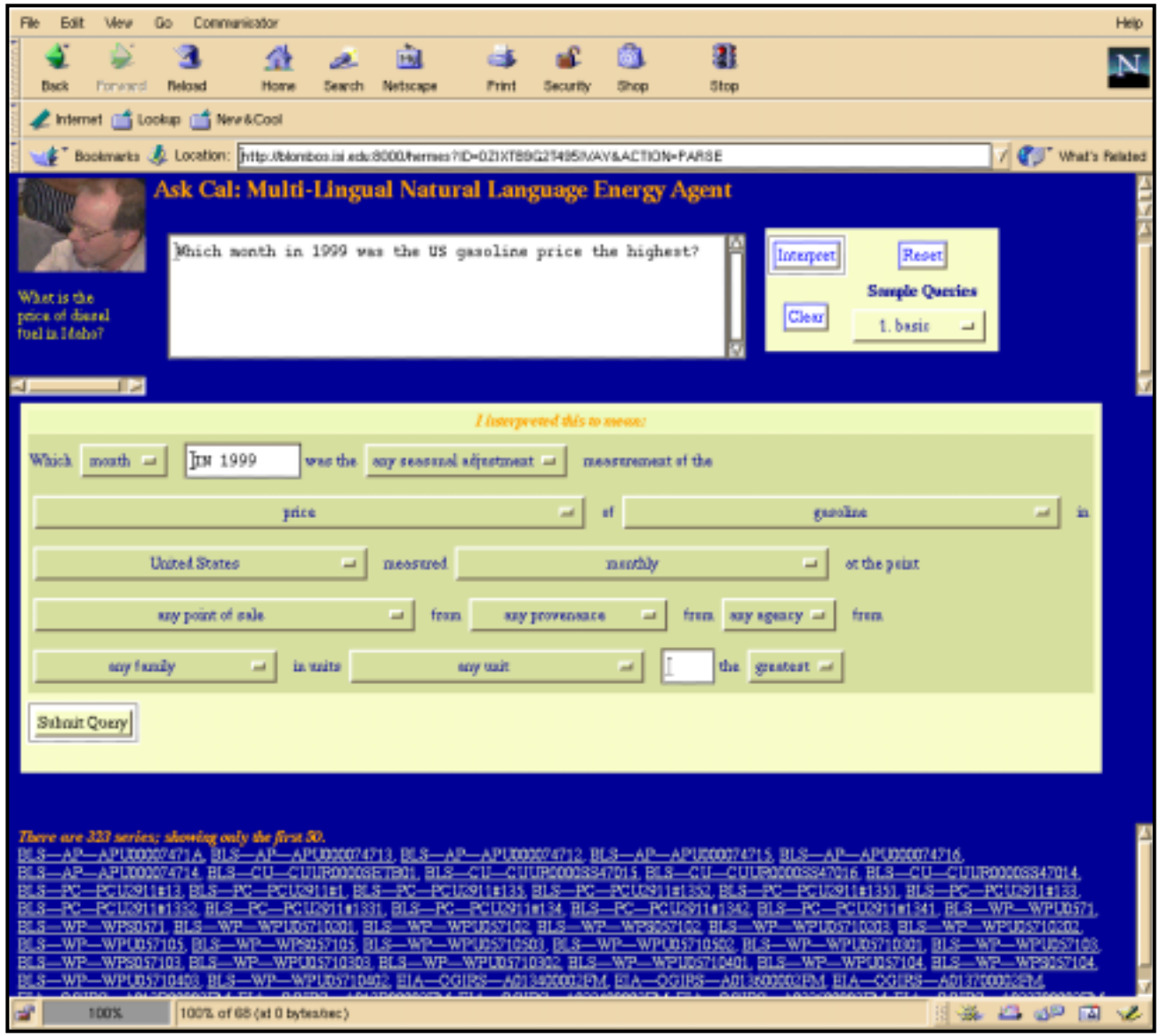


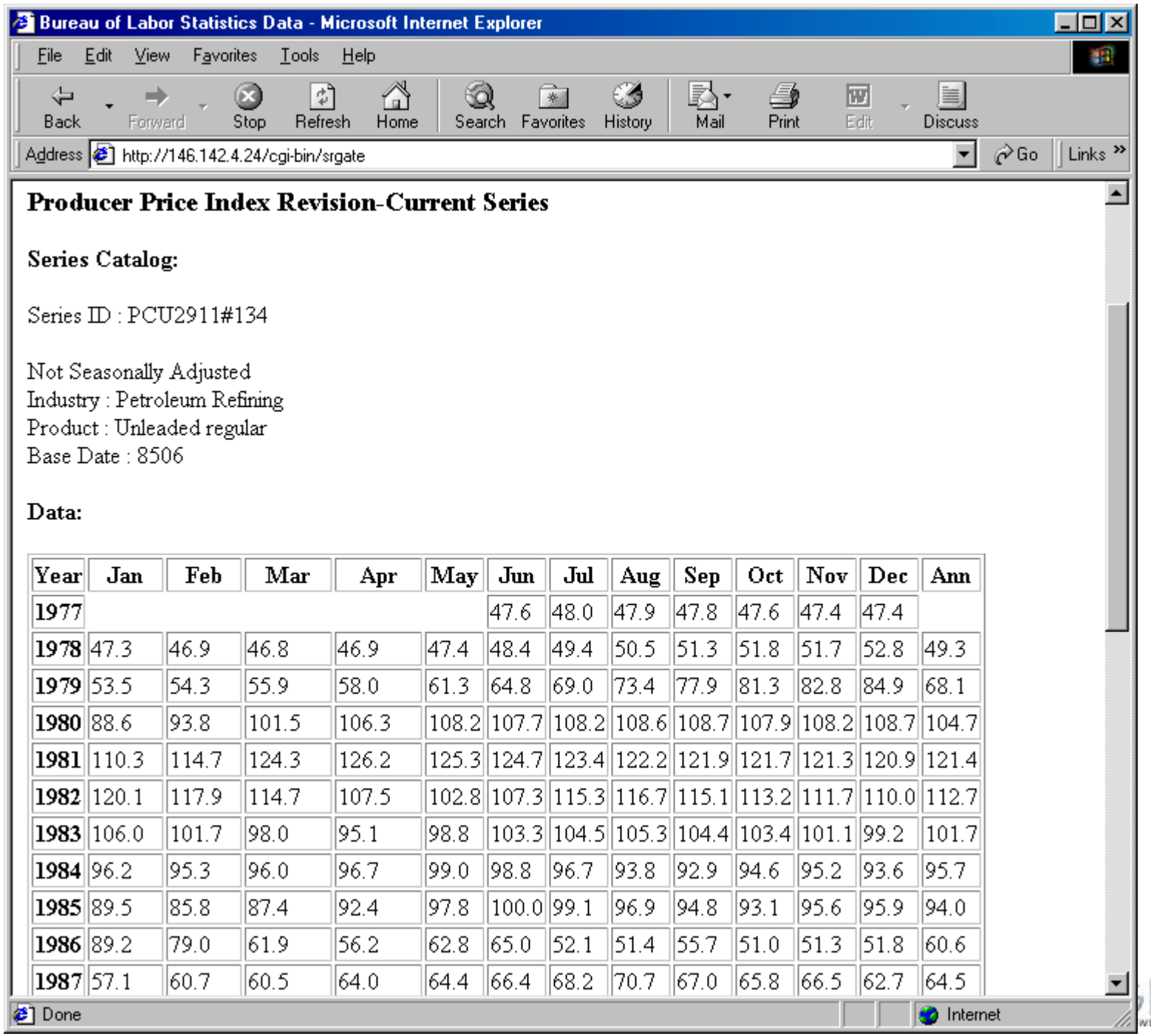
- Fisheye magnification of region of interest
  - Magnified group laid out to avoid internal overlap

USC INFORMATION SCIENCES INSTITUTE

AskCal: User requests in English

- ATN:
  - 341 nodes
  - 14 question types
- Automated paraphrase to confirm
- Dialogue continues via menus for detailed selection




Bureau of Labor Statistics Data - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Forward Stop Refresh Home Search Favorites History Mail Print Edit Discuss
Address http://146.142.4.24/cgi-bin/srgate
Go Links
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
Done
Internet

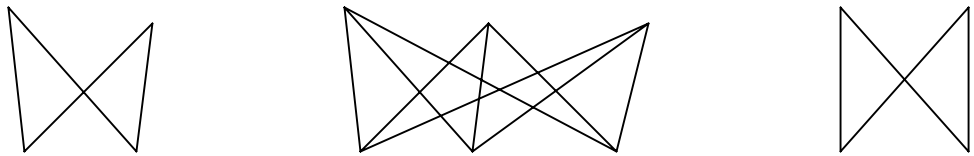
Part 2: Planned research

- 1985-95: A revolution in machine translation
  - Have statistical systems count correspondences between parallel/aligned French-English (Canadian Hansards)
  - Systems induce lexical and grammatical mappings
  - Currently applied to Chinese, Arabic, Japanese, etc. as well – highly promising

  - Numerous toolkits available (alignment, decoding, etc.)
  - ISI has one of most active research groups in the world

Statistical alignments for MT

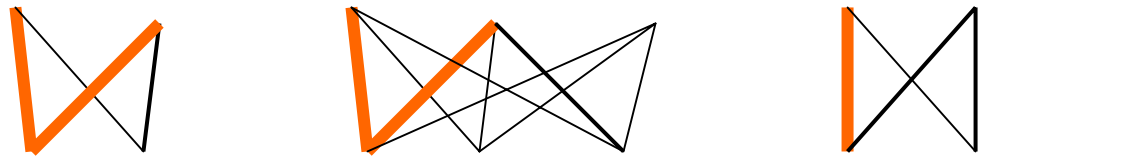
… la maison … la maison bleue … la fleur …



… the house … the blue house … the flower …

“the”→?

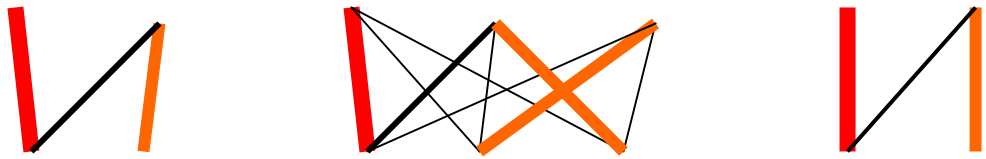
… la maison … la maison bleue … la fleur …



… the house … the blue house … the flower …

“house”→?

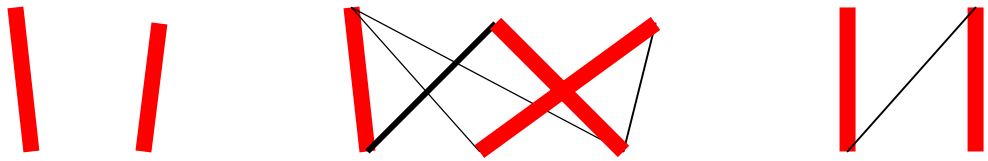
… la maison … la maison bleue … la fleur …



… the house … the blue house … the flower …

Strengthening the tendencies

… la maison … la maison bleue … la fleur …



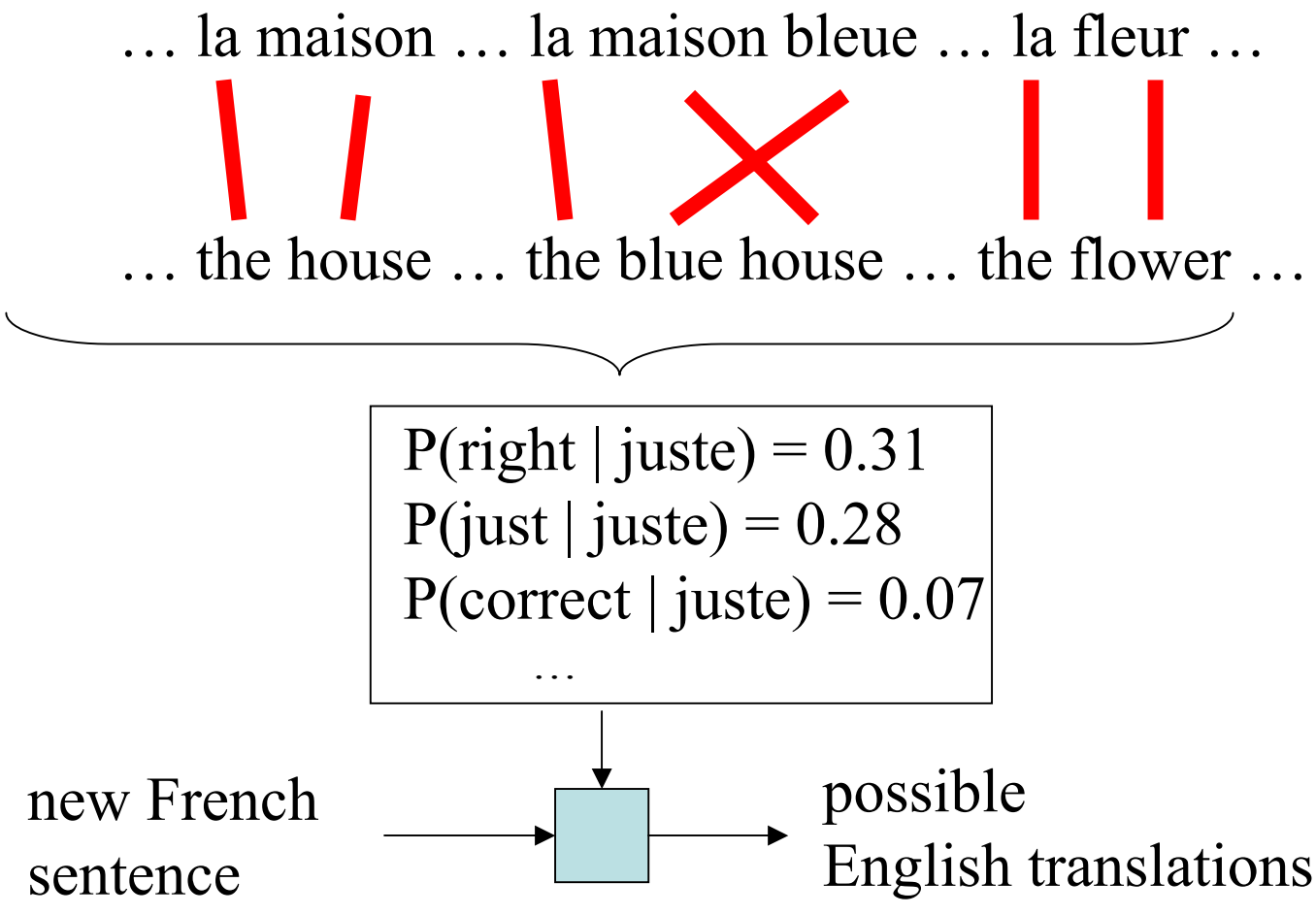
… the house … the blue house … the flower …

Final mapping

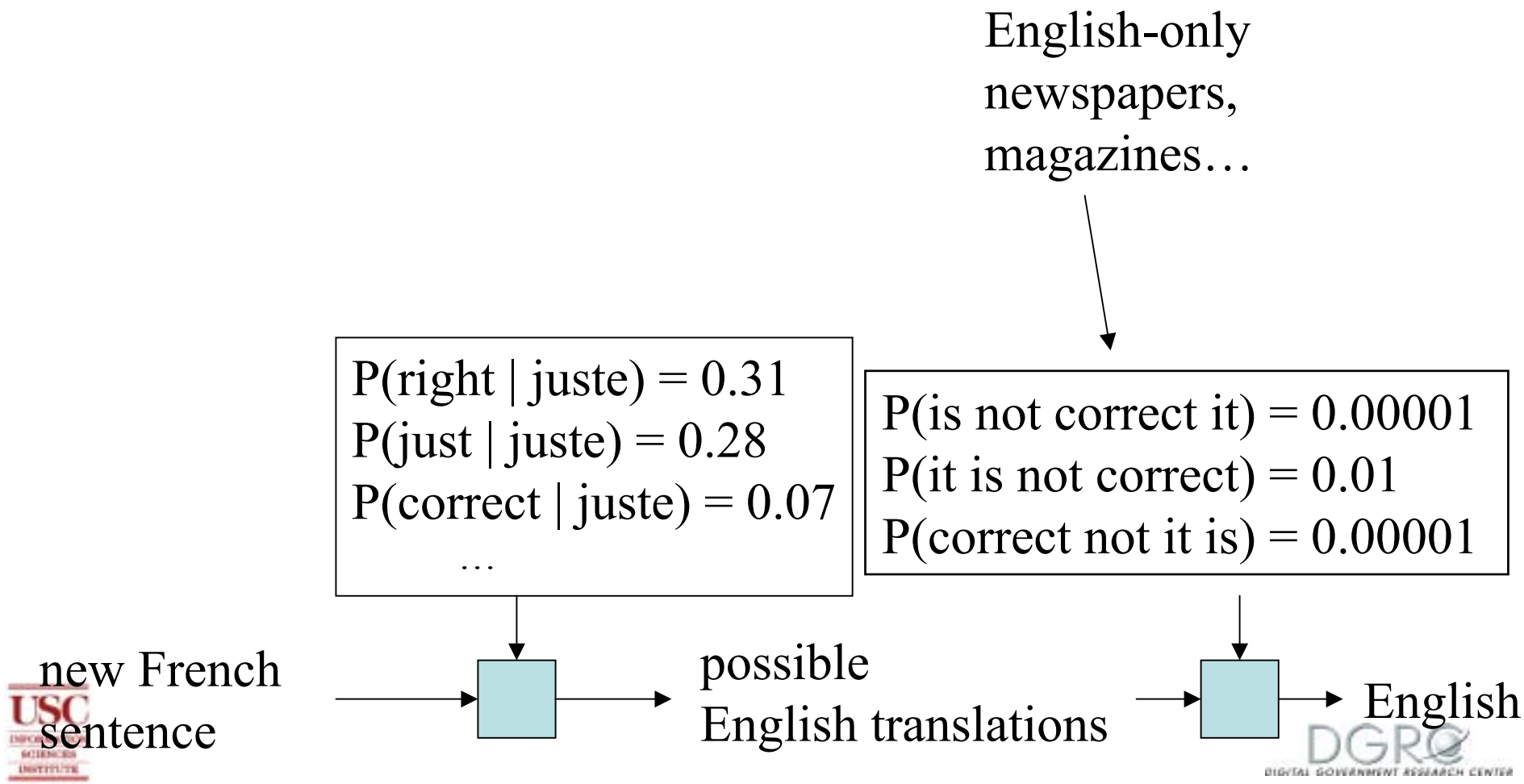
… la maison … la maison bleue … la fleur …

… the house … the blue house … the flower …

Statistical MT: step 1



Statistical MT: step 2



Proposed work

- **Goal**: learn mappings from one database to another
- **Method**:
  - Treat source and target databases as I/O pairs for MT
  - Learn alignments/mappings in layers:
    - Individuals: Cell-to-cell
    - Group of cells: column-to-column
    - Metadata: header-to-header (etc.)
    - Groups of groups: sets of columns -to- sets of columns
  - Similarity metrics: value, orthography (number, text, 2-decimal-places, etc.), value range, ontological type, etc.
- **Approach**: start by hand, then learn

Partners

- California Air Resources Board
  - Michael Benjamin, Emissions Inventor Systems Section, Sacramento
- California Air Quality Management Districts
  - 35 in California, with databases ranging from MS Access databases to Excel spreadsheets
- Later: reach out to other states and Federal EPA
- Possibly: reach out to Mexico and elsewhere

- Fire Emissions (w. Dr. Stefan Falke) – USDA/Forest Service and Dept. of Interior

Thank you!