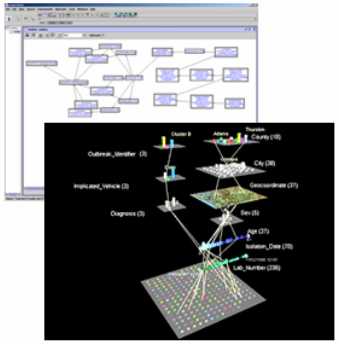
Information to Insight

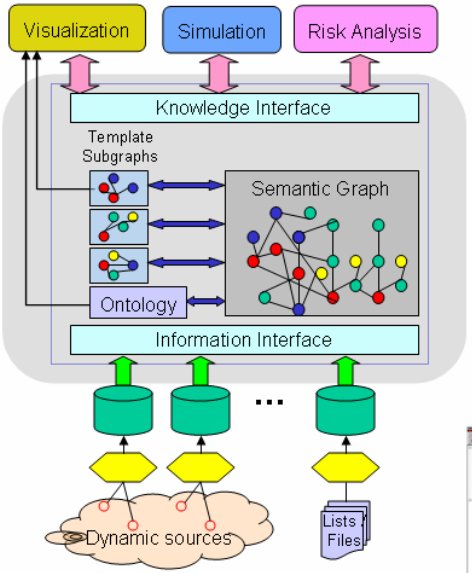
in a Counterterrorism Context

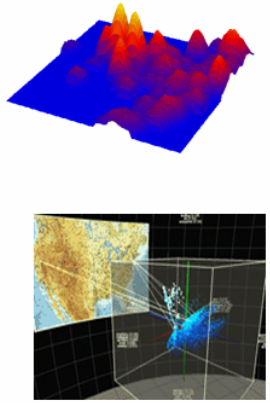
Robert Burleson

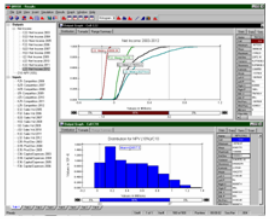
Lawrence Livermore
National Laboratory

UCRL-PRES-211319
UCRL-PRES-211466
UCRL-PRES-211485
UCRL-PRES-211467









This work was performed under the auspices of the U.S. Department of Energy by University of California, Lawrence Livermore National Laboratory under Contract W-7405-Eng-48

We must be able to address the analysts' requirements

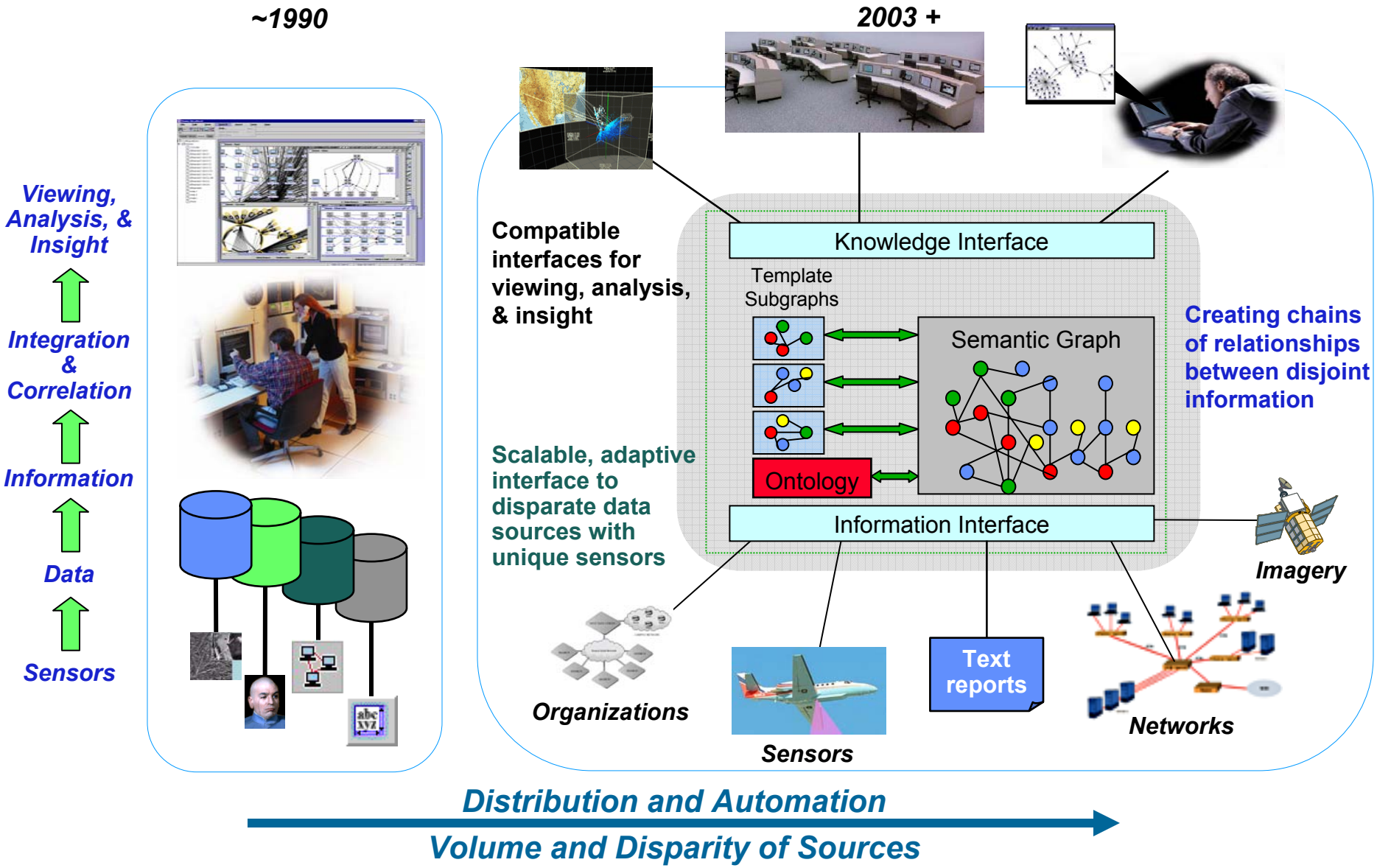
- Strategic analysis
  - See the "big picture" and how to counter terrorism
  - Support decision makers in setting policies and priorities
  - Integral to targeting technical and human source collection

- Tactical analysis
  - Predict and warn of pending attacks
  - Provide an understanding of our adversaries' current intentions and capabilities
  - Allow the United States to act with precision both defensively and offensively

- Both strategic and tactical analysis require a system capable of fusing information obtained from very diverse sources…

The Analysis, Dissemination, Visualization, Insight, and Semantic Enhancement (ADVISE) system is being developed for DHS S&T to meet these requirements

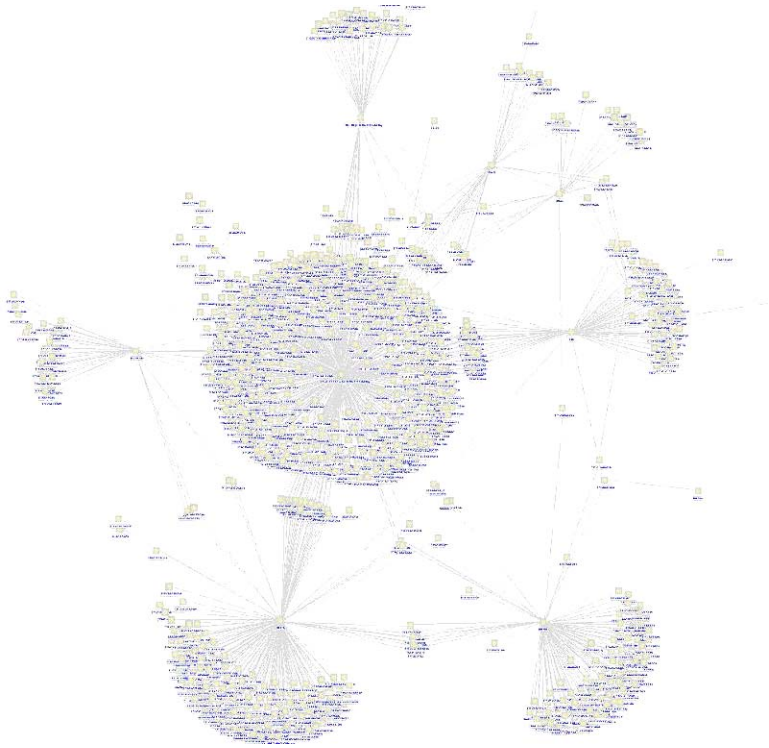
ADVISE lets us understand the information that characterizes our national security challenges



What drives the design…

"Connect the Dots…"

- Scaling to massive data volume
- Ingest information from information s
  - 100's of systems
  - Real-time
  - High-throughput
  - Stove-piped by intent
- Support 100's of analysts
  - Event notification in near-real time
- Control access and Protect privacy
- Responsive to change



What to consider when scaling to massive levels

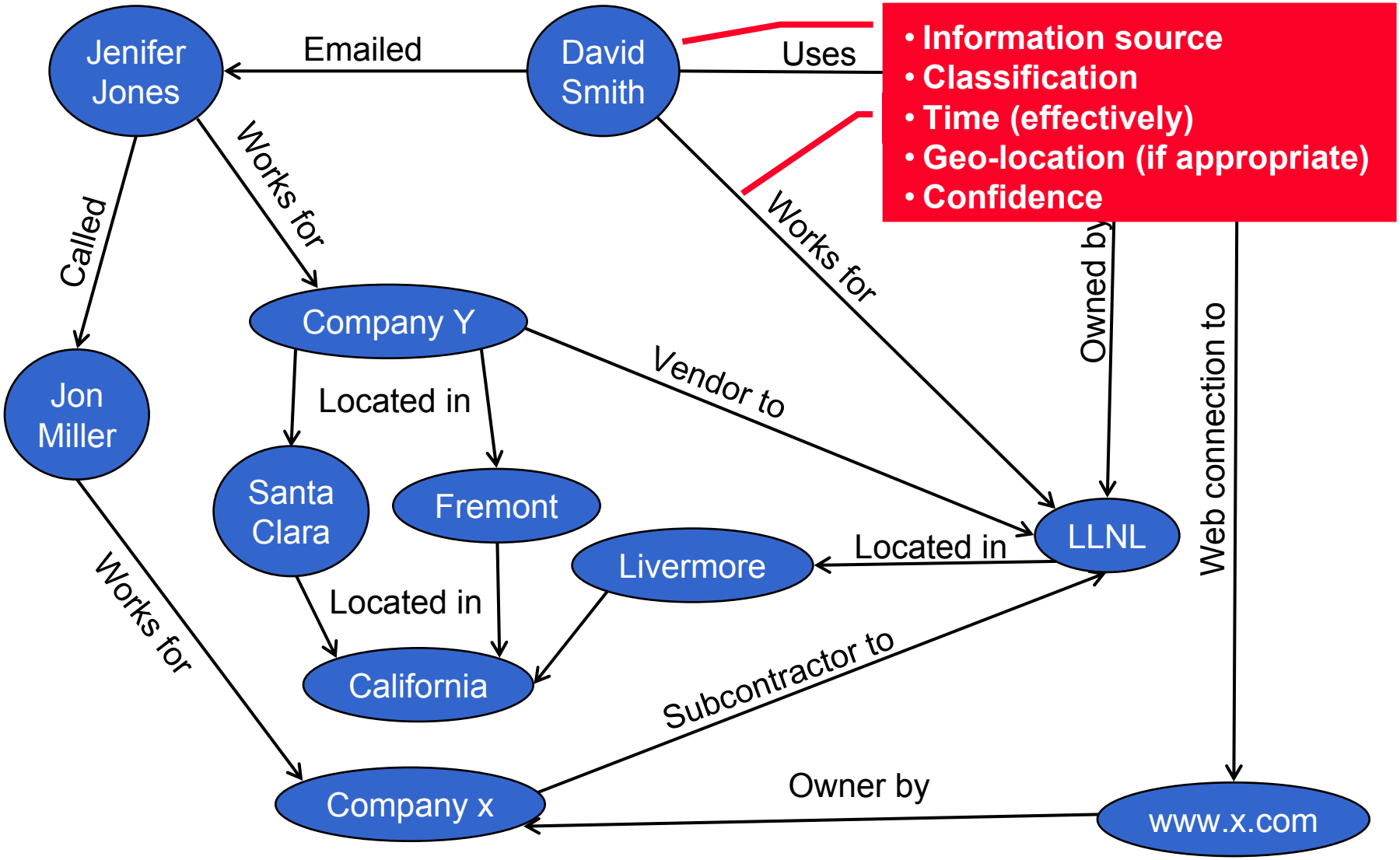
What do we want from the Knowledge Fusion engine?

- Relations between facts (nodes)
  - Individual facts without relations not particularly useful (might as well keep stovepipes)
- Relate facts (build the graph)…
  - …at high ingest rates with results in real-time
- Responsive to change

What is important when scaling to massive sizes?

- An optimal model
  - Use relations between facts (connectivity) to extract knowledge from data
- Query performance
  - Key for high-complexity algorithms

Semantic graphs provide the basis for these massive knowledge relationships



The fused graph reveals connections and gaps not immediately apparent

Existing search tools can find documents that contain a given connection:

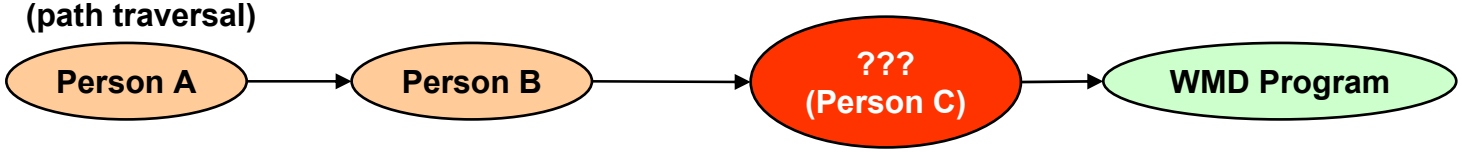
"Person A" & "Person B"

"Person A" & "Country X"

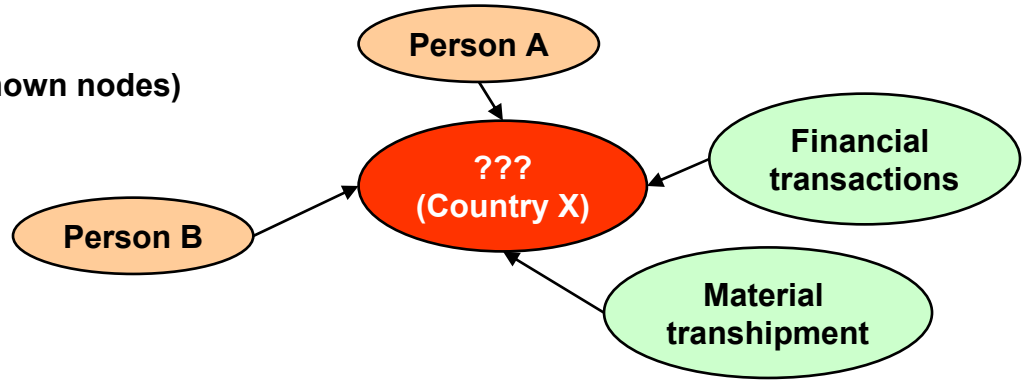
"Person A" & "City Y"

Graph identifies connections that span several messages (sources):

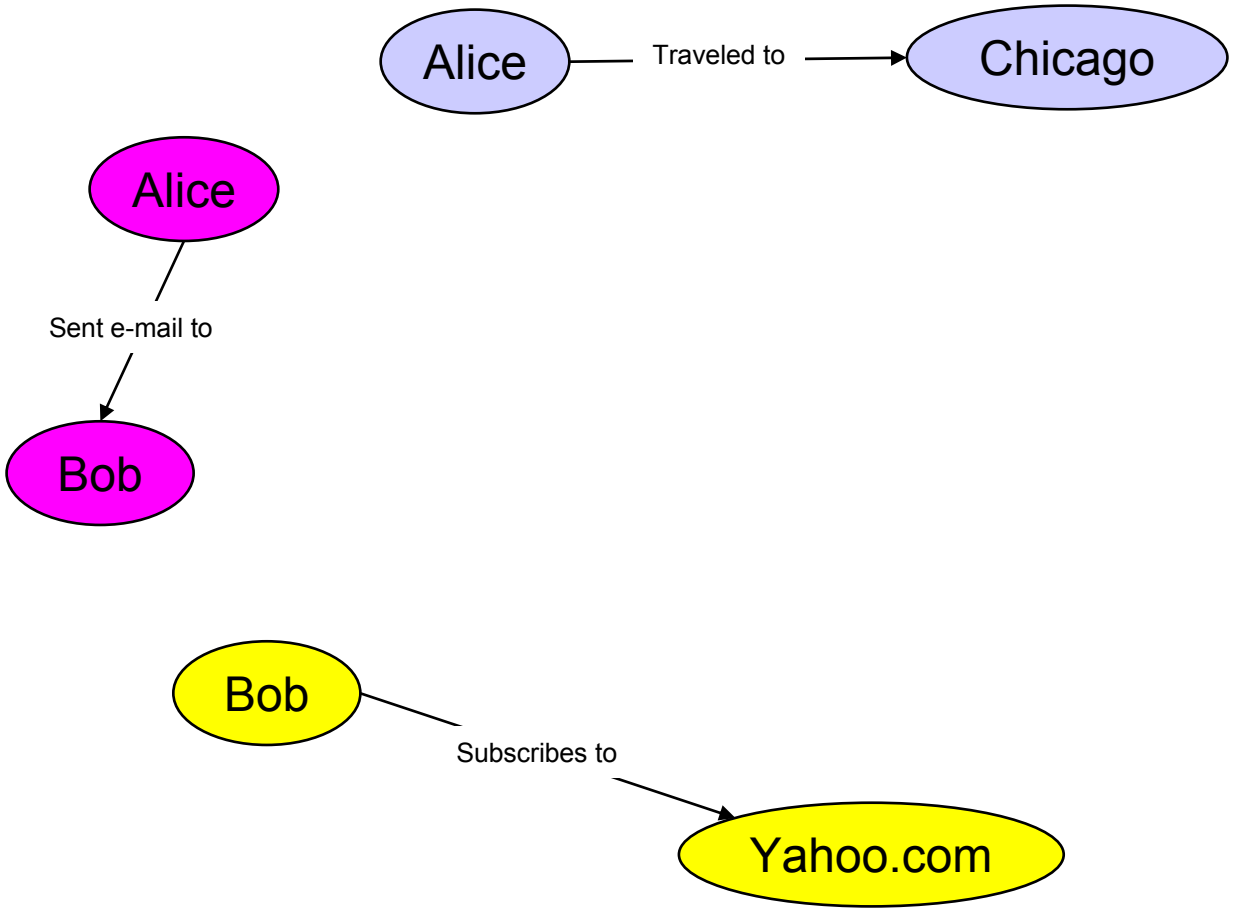
Previously unknown "Middlemen" (path traversal)



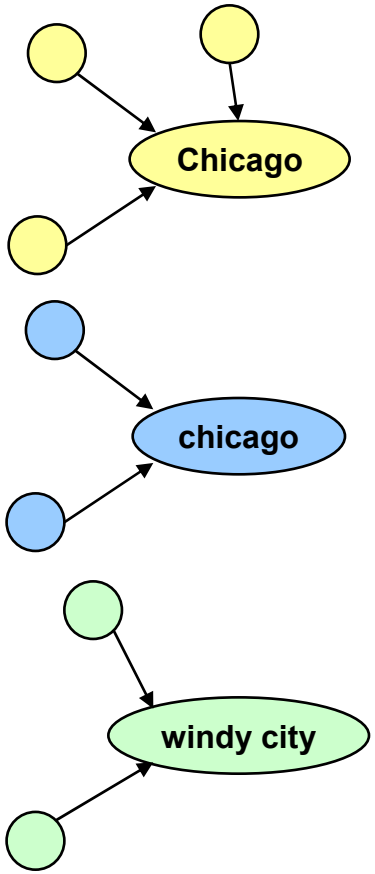
Hidden common connection (unknown nodes)

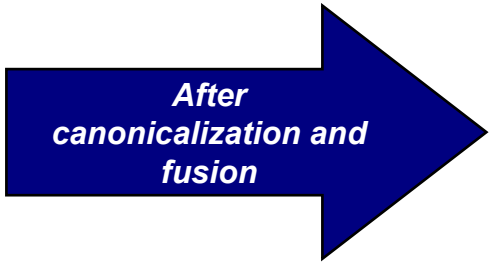


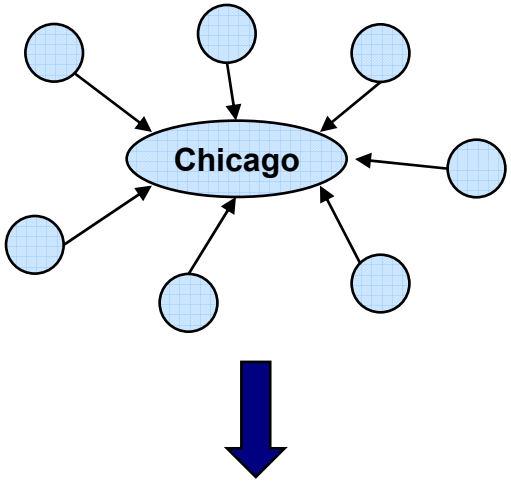
Two facts "fuse" when they contain a common node with identical attribute values

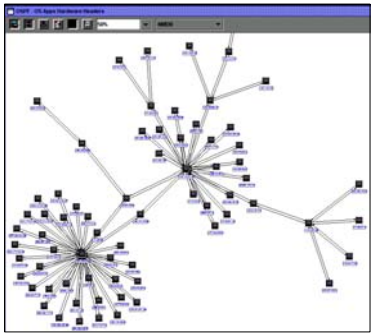


ADVISE canonicalizes data to maximize fusion and improve searches



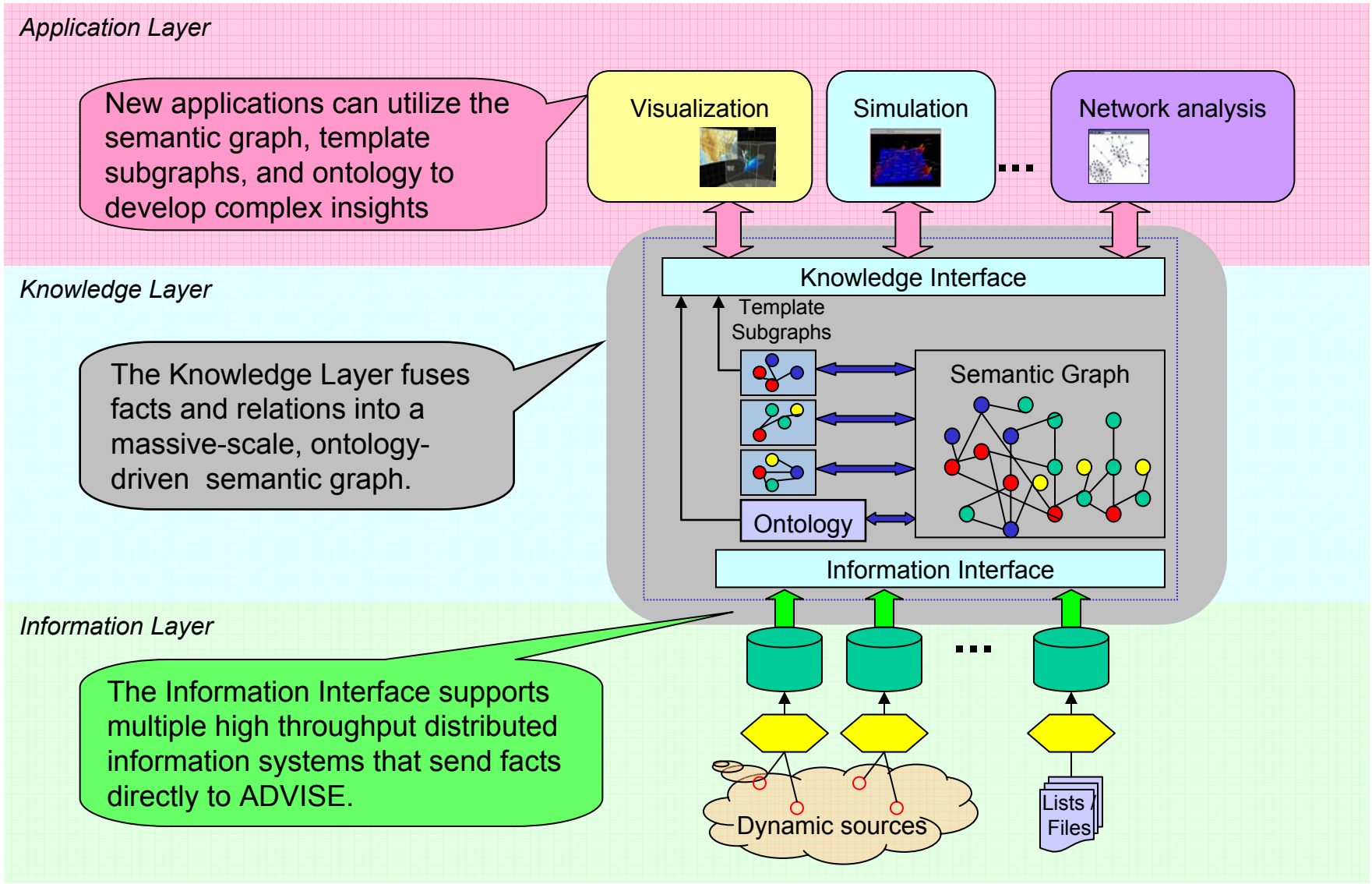






Applications can use any of the organization names to get the same result

The ADVISE system model partitions the design



Creating entities and relationships from free text is critical

"BAGHDAD, Iraq (CNN) -- A hostage shown in a videotape on an Arabic language satellite TV network Wednesday is the American executive who was kidnapped Monday at a construction site in Baghdad, according to a U.S. Embassy official.

Jeffrey Ake, president and chief executive officer of a machine manufacturing firm, was seen in the video being held at gunpoint by militants."

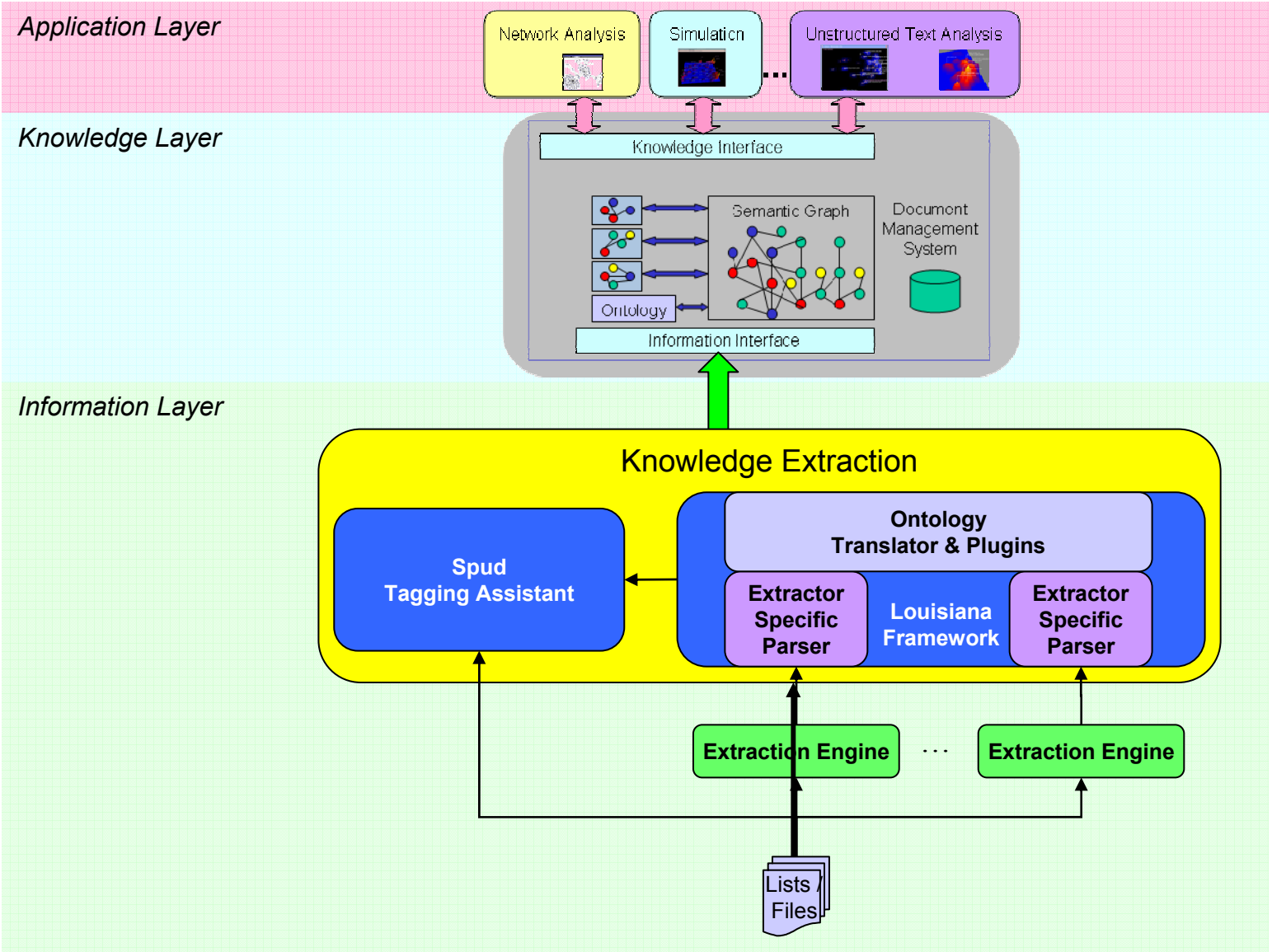
HARD

```
Country:  Iraq
City:     Baghdad, Iraq
Location: construction site
Person:   U.S. Embassy official
Person:   Jeffrey Ake

Relation:  LOCATED_IN
  Locatee: construction site
  Locator: Baghdad, Iraq

Event: KIDNAPPING
  Victim:      Jeffrey Ake
  Perpetrator: militants
  Location:    construction site
```

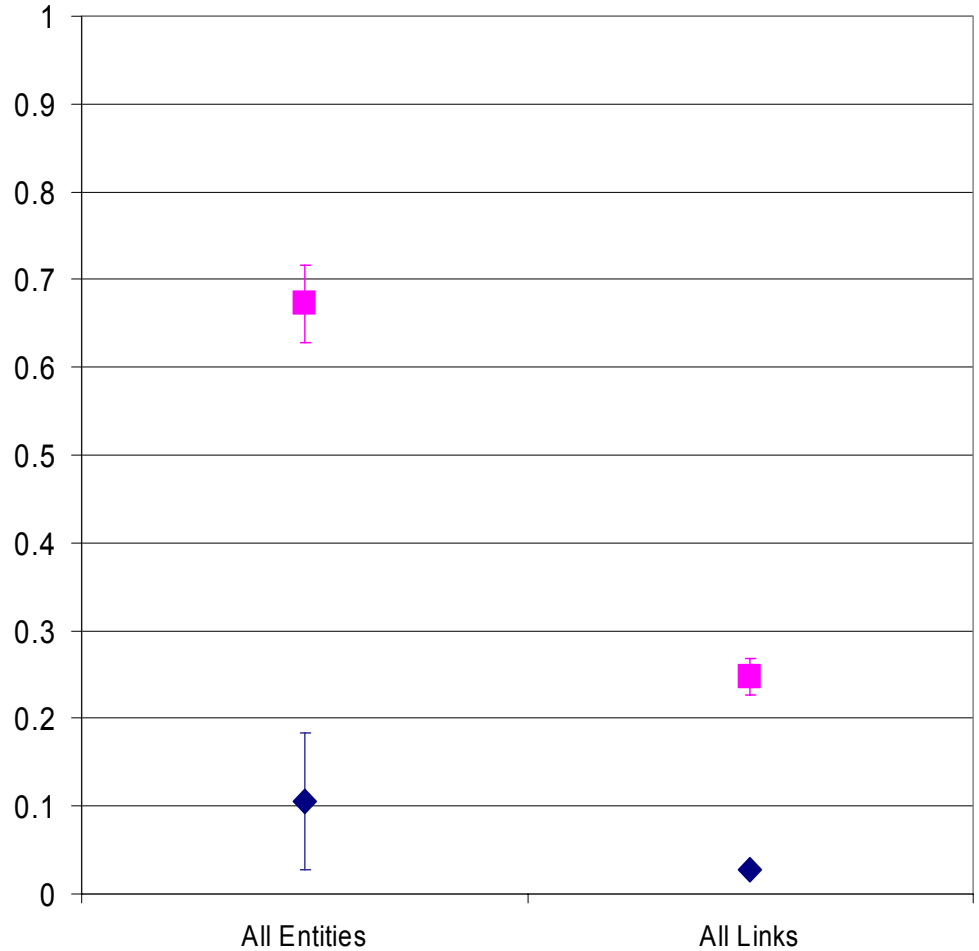
Integrating knowledge extraction into ADVISE



Evaluating extraction engines

- Qualitative: Show resultant graph to analysts
  - They hate it

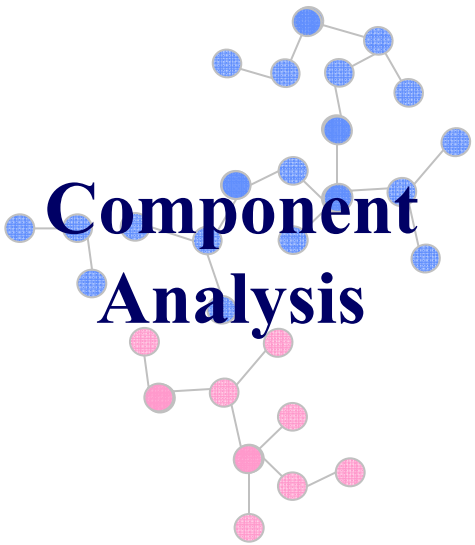
- Quantitative: Compare engine output to an answer key
  - Modified GATE to evaluate extraction engine results against one another or against a hand-annotated answer key
  - Hand-annotated some documents (not fun)
  - Can use documents entered via Spud

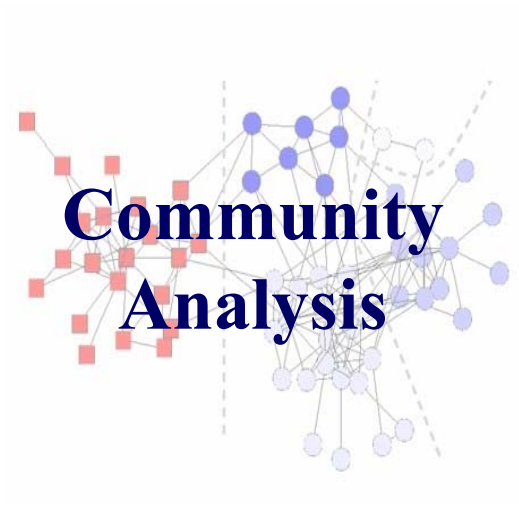


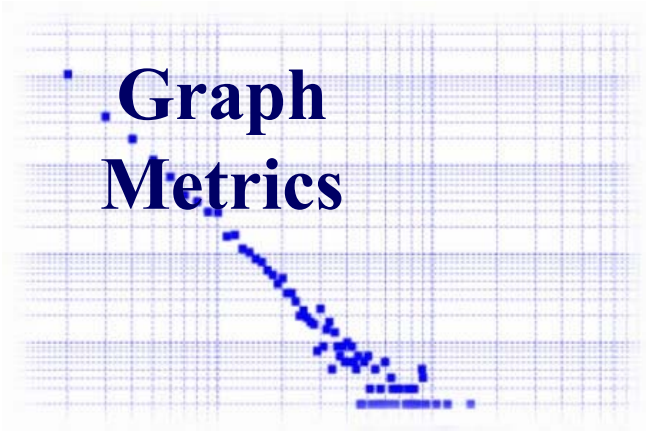
Current direction for text extraction

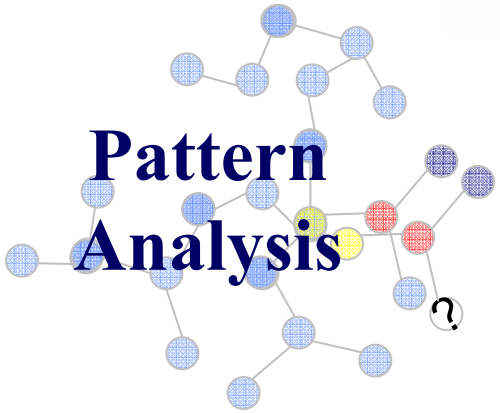
- Integration
  - Improve usability of Louisiana
  - Add graph interactivity to Spud
  - Work on merging results from multiple engines
- Evaluation
  - Evaluate more engines
    - AeroText and ClearForest on deck
  - Look for applicable pre-tagged document corpora
  - Build graph-comparison capability in ADVISE
- Collaboration

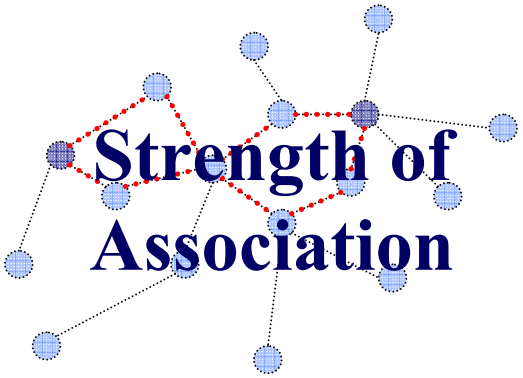
Graph analysis environment





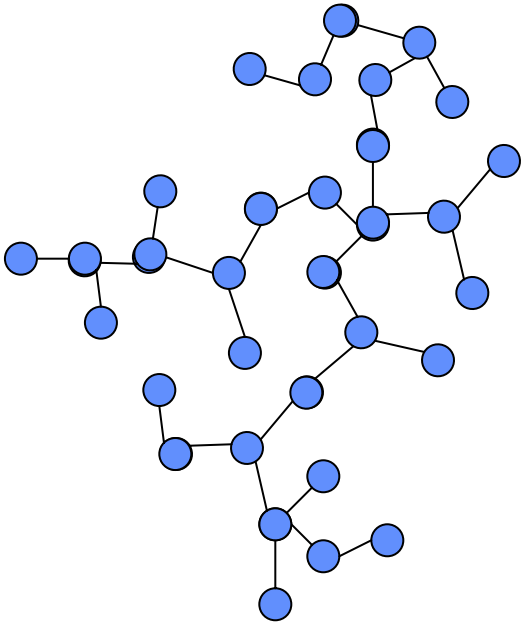






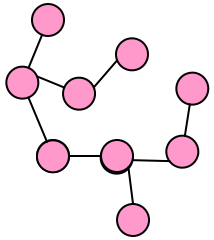
Component Analysis assists in the understanding of how graphs fuse

We build a semantic graph from various information sources

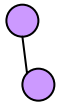


The graph is based on an ontology, which only allows certain relationships

Some data will fail to fuse

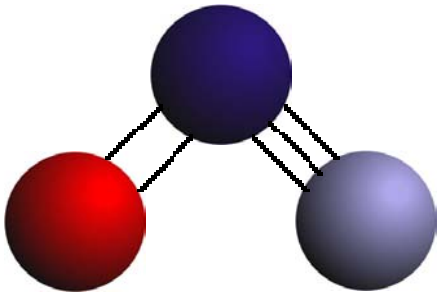


Analyzing resulting components can provide us valuable information about data fusion



Community Analysis partitions the graph into clusters of "related" nodes

- Measure the "betweenness" of each link
- Eliminate the link with highest "betweenness"



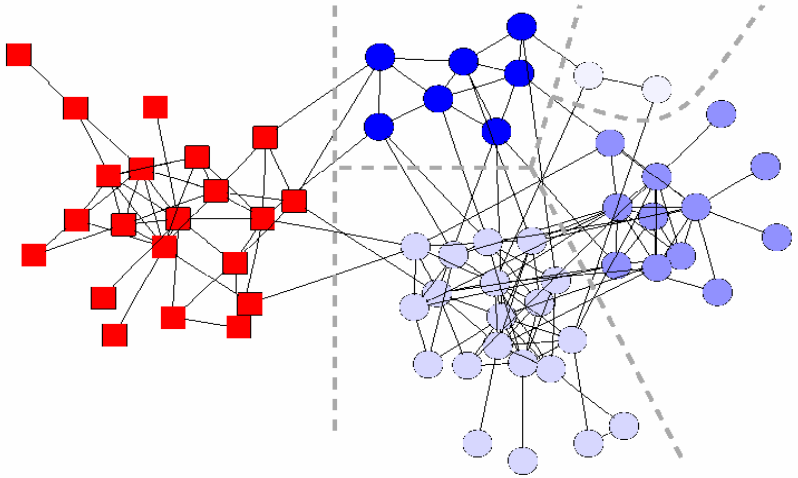
- Stopping criterion – computed at each iteration to determine "ideal" partition

Our stopping criterion measures the density of links within communities relative to the density of links between communities - iterations stop when this is maximized

Community Analysis partitions the graph into clusters that may facilitate knowledge discovery

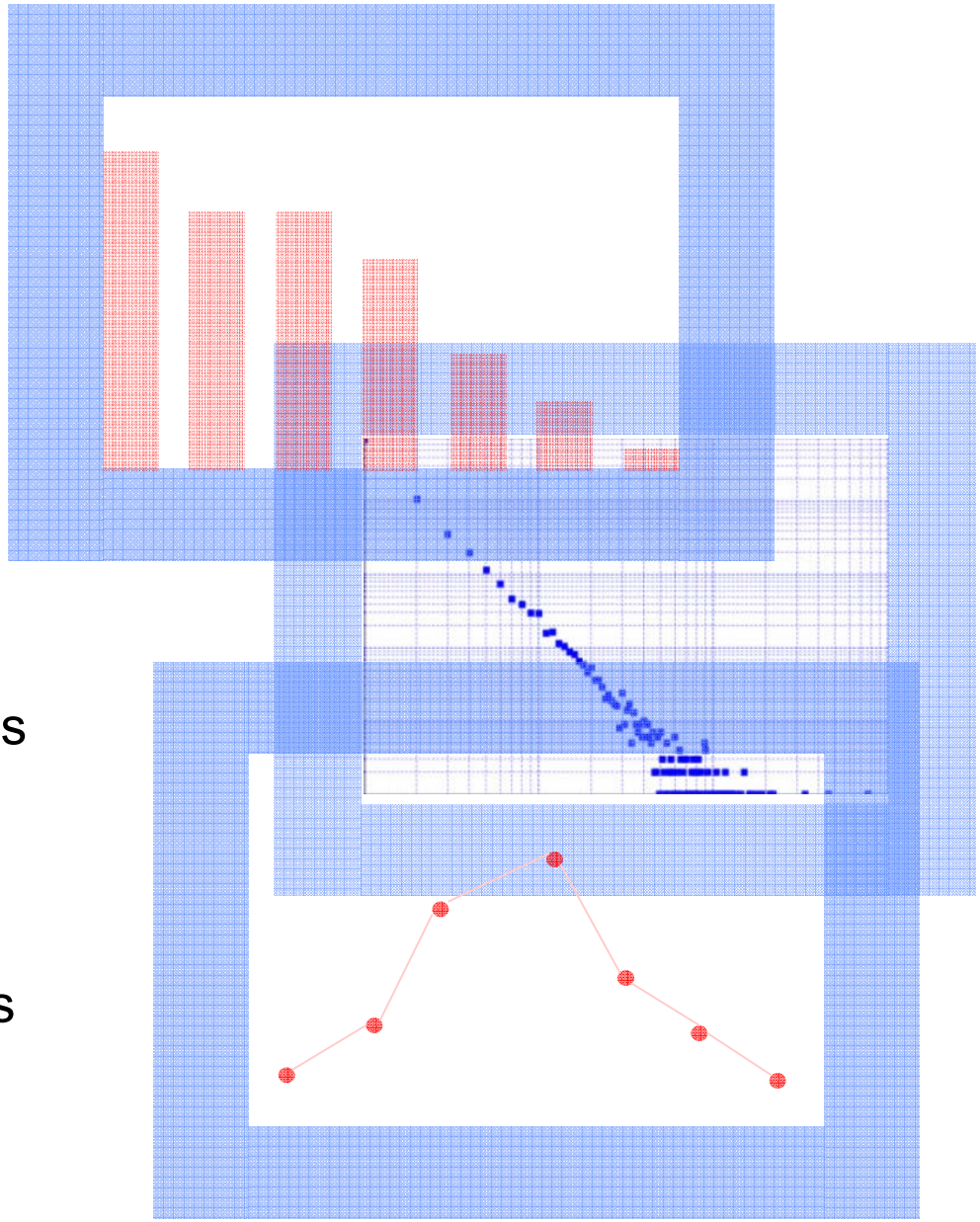
Key Uses for Graph Analysis:

- Examining the semantic graph at varying degrees of granularity
- Trials indicate a tendency to produce semantically homogeneous communities
- Metrics run on communities provide a local and more detailed analysis of a large semantic graph



Graph Metrics helps in the understanding of what is in the graph

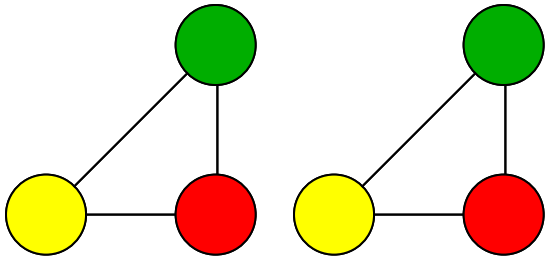
- Our library of graph metrics allows us to:
  - Analyze high-level content
  - Characterize our graph/ communities
  - Measure knowledge extraction performance



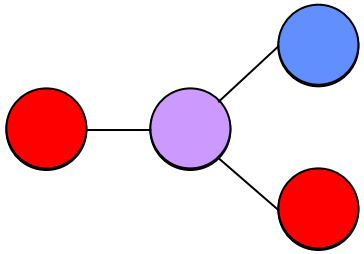
- Node/Link Type Frequencies
- Node Degree Distributions
- Path Analysis
- Ontology Utilization Metrics
- High Degree Node Statistics

Pattern Analysis determines potentially valuable information from patterns in the graph

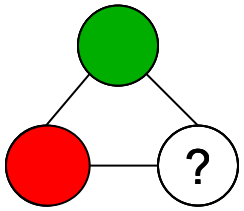
- Identify rare and common patterns

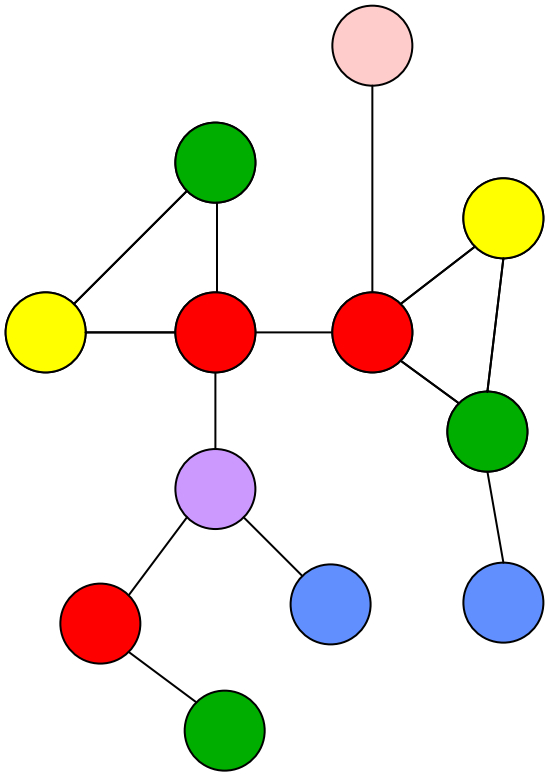


- Pattern matching



- Fuzzy pattern matching





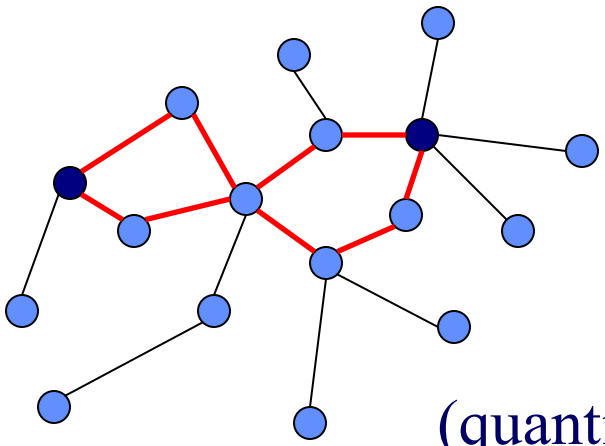
Strength of Association allows nodes to be ranked according to their relative strength

Allow pairs of nodes to be ranked according to their relative strength of association

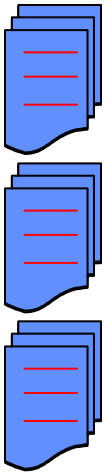
Source-based Weight

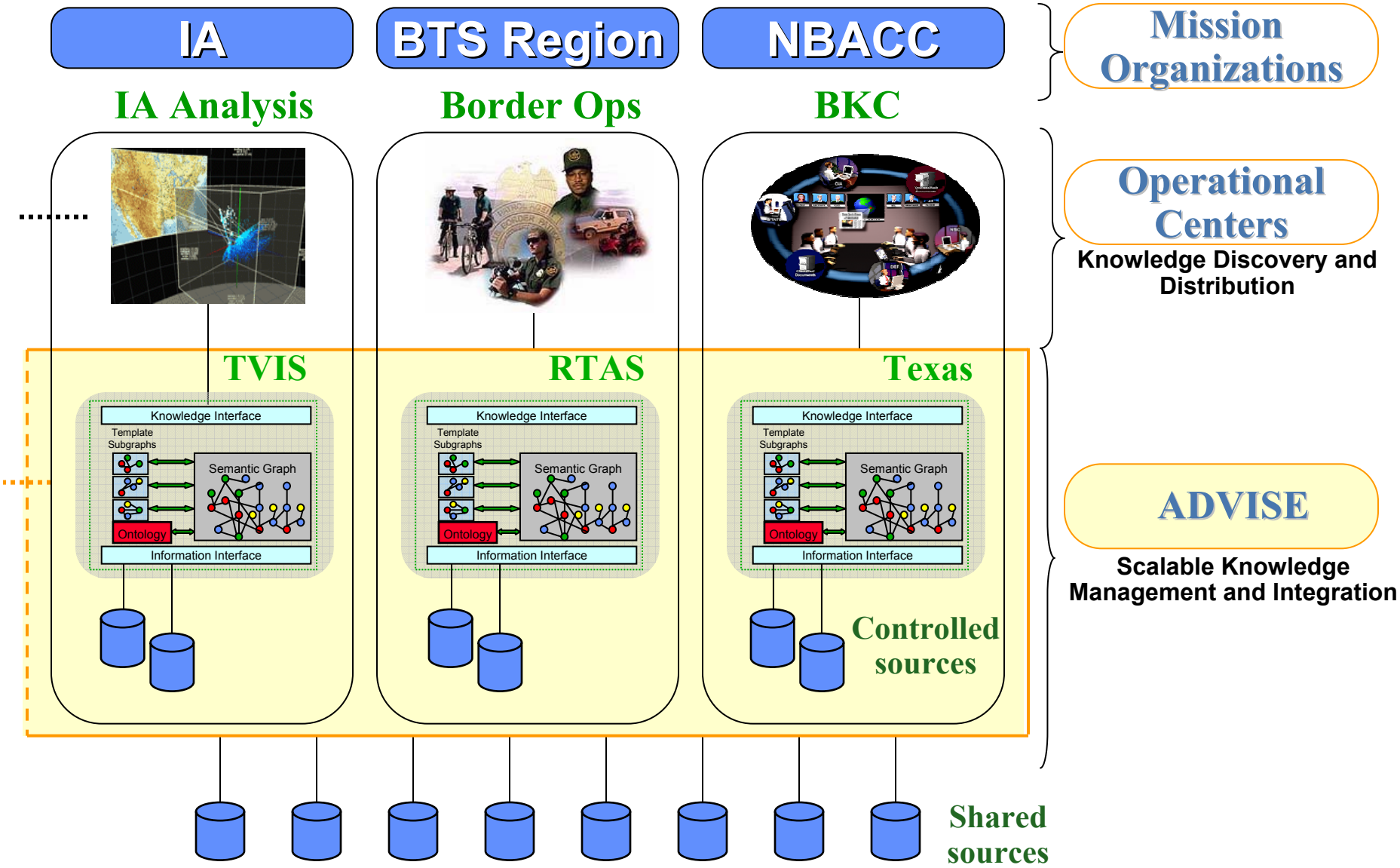
Topological strength

(neighborhood)



(quantify source support)



Allow multiple paths between two nodes to be ranked according to their relative strength

ADVISE supports scalable knowledge management across multiple missions