

Knowledge
Discovery &
Dissemination

The Requirements

The Problem

- Making decisions in a data rich world of large and diverse data sources

Tools are needed to make *"actionable"* sense out of disparate information

The Solution

- **Knowledge Discovery** to gather, analyze and integrate data and information into knowledge
- **Knowledge Dissemination** to make knowledge actionable for decision making

LANL Leadership For KDD

KDD = Knowledge Discovery and Dissemination

LANL KDD Leadership

Deborah Leishman – leishman@lanl.gov

Nancy David and Cliff Joslyn

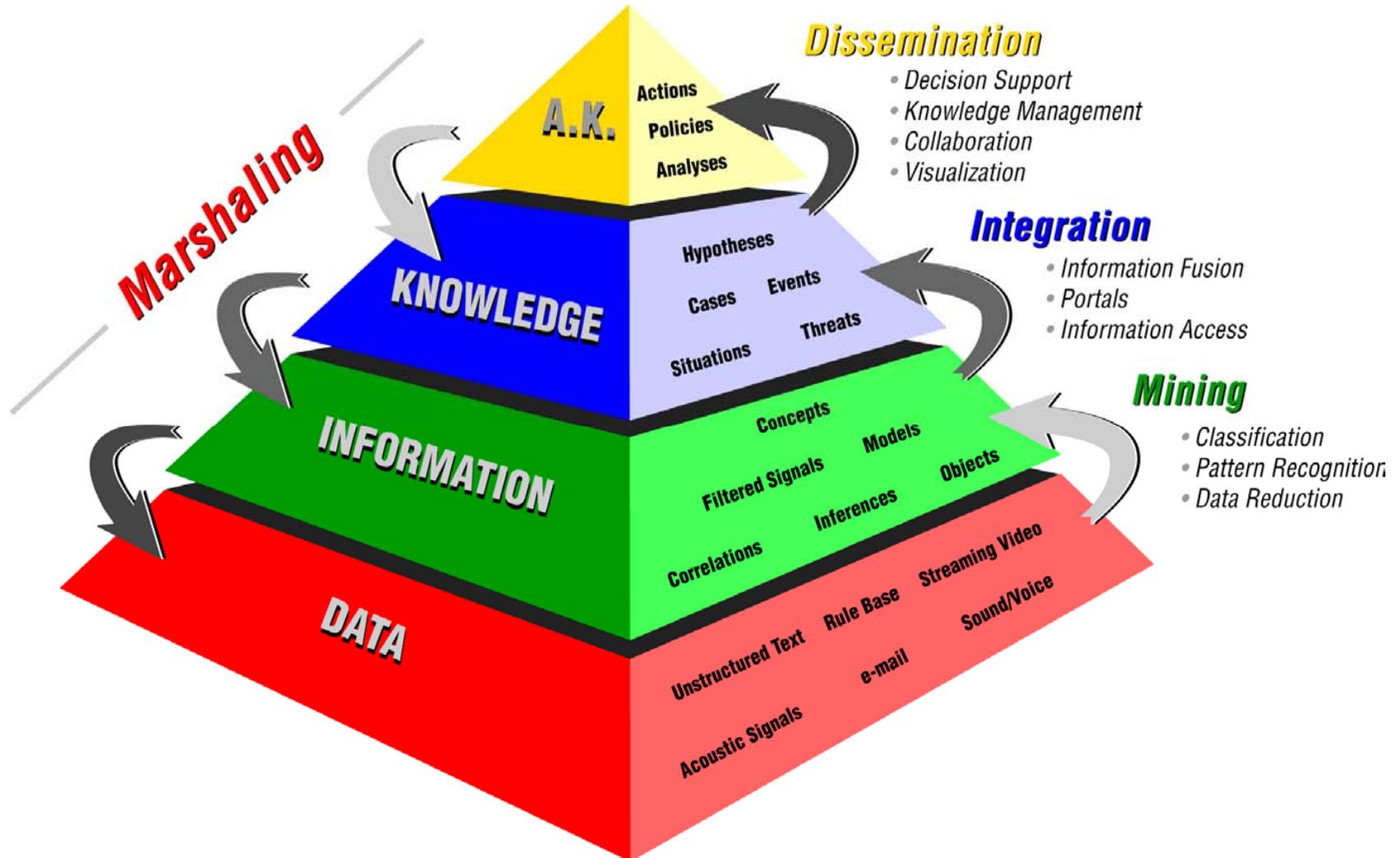
A critical focus in LANL research programs to support strategic objectives in:

- ✓ *Global Situational Awareness***
- ✓ *Stockpile Certification and Archives***
- ✓ *Terrorist Tracking***
- ✓ *Surveillance Systems***
- ✓ *Defense Transformation***
- ✓ *Threat Analysis and Response***

Why KDD at LANL?

- **Pressing needs in Knowledge Discovery and Dissemination (KDD)**
- **Needs:**
 - Support analysts
 - Provide decision makers with meaningful information
 - Glean knowledge from vast proliferation of data sources and analytical methods
- **LANL and KDD:**
 - Interdisciplinary applied science laboratory
 - One of the very few institutions worldwide with appropriate skills necessary

What is KDD



LANL Capabilities Span the Pyramid

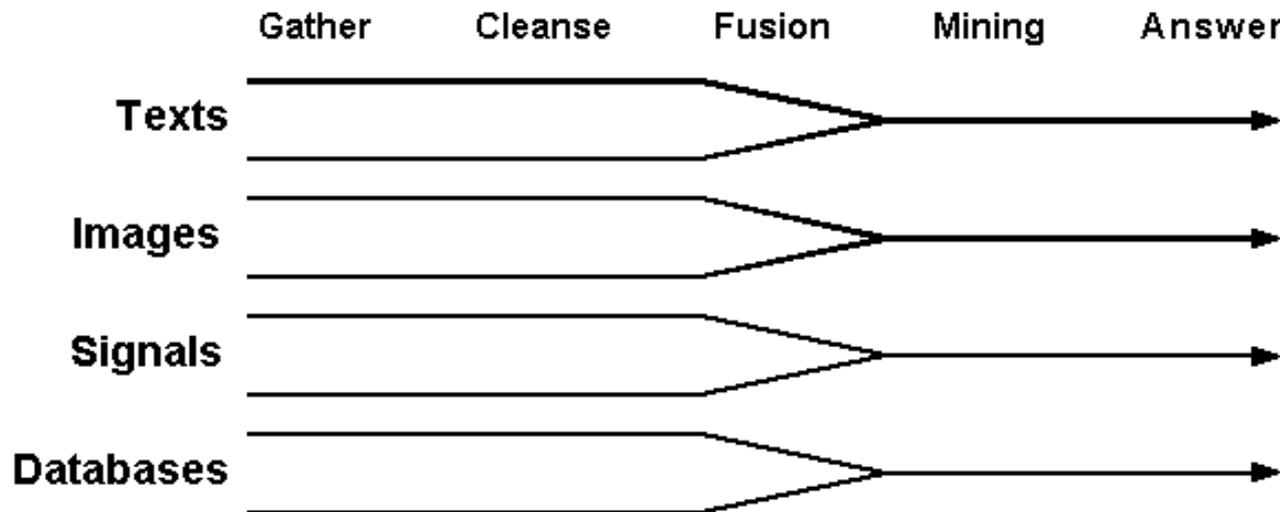
- **Mining**
 - Machine Learning
 - Statistical Techniques
 - Image Understanding
 - Image Compression
 - Text Mining
 - Information Retrieval
- **Dissemination**
 - Decision Support
 - Visualization
- **Marshaling**
 - Evidence and Data Marshaling
- **Integration**
 - Information Integration
 - Knowledge and Social Networks
 - Ontologies
 - Probability and Uncertainty
 - Agents

KDD Customer Requirements

- **Operational**
 - Gather data from diverse sources, perform mining, integration and dissemination in short tactical timeframe – command post as an example
- **Strategic**
 - Do above but with longer timeframe and more collaboration and integration across analysts
- **Business**
 - Business Intelligence and Customer Relationship Management
- **Cognitive/Automated**
 - Above transitions occur but emphasis is on automation – intelligent agents or automated weapons for example

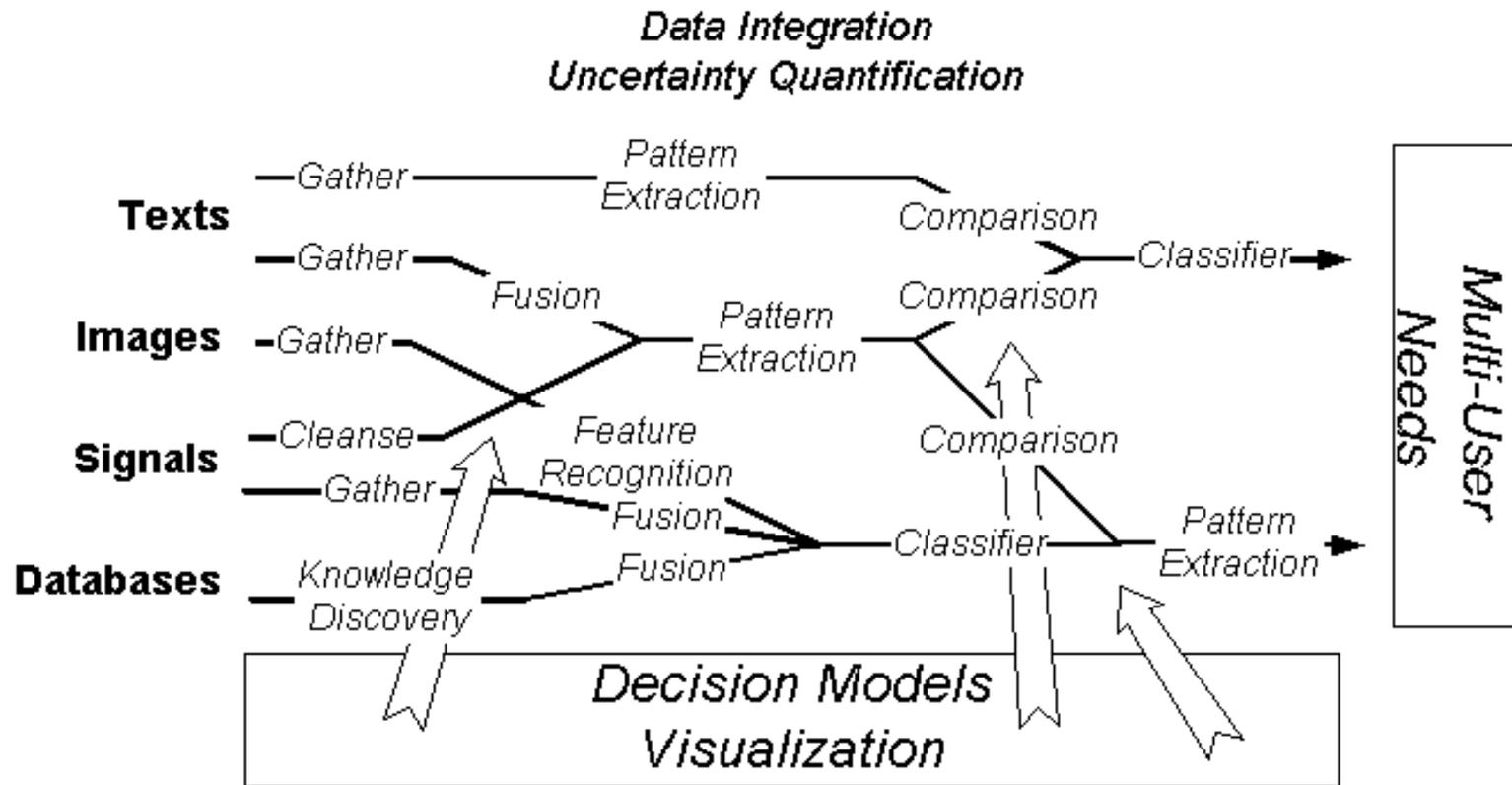
KDD Today

- Single data source
- Single user question
- Particular path through multiple methodologies:
 - Data gathering and cleansing
 - Correlations, Patterns, Trends, Classifications
- Particular repertoire of method:
 - Machine Learning, Statistics, Knowledge Modeling, etc



Our View of KDD

- Hybrid knowledge integration pathways
- User-guided knowledge discovery through marshaling
- Driven by explicit decision models of multi-user needs



Example KDD Projects

- Image Analysis Environment: ISIS
- BioSurveillance Decision Environment
- Intelligence Analyst Support Environment
- Tactical Information Integration System
- Collaborative Forestry Management and Response System

Example KDD Projects

- Image Analysis Environment: ISIS
 - GENIE
 - AFREET
 - POOKA
 - ZEUS
 - ALADIN

ISIS: Intelligent Searching of Images and Signals



- **ISIS Software suite**
 - **GENIE** (genetic algorithm)
 - **AFREET** (support vector machine)
 - **POOKA** (hardware version)
 - **ZEUS** (time series)
 - **ALADDIN** (ENVI user interface)
 - **MAGIC LAMP** (any user interface)



- **Automated Discovery of New Image (and Signal) Analysis Tools based on Expert Assistance**

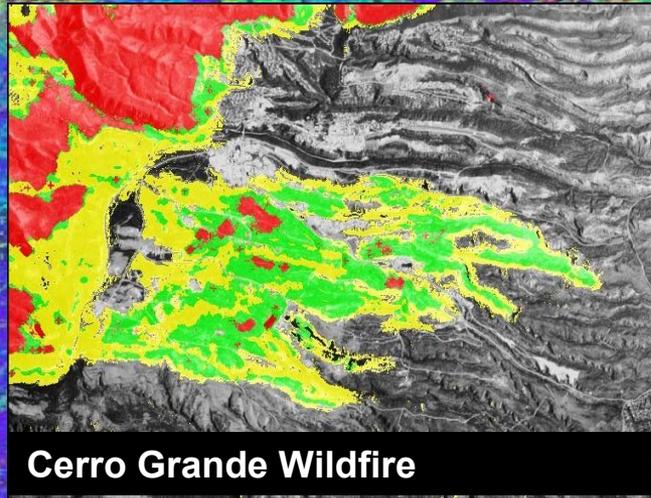


GENIE

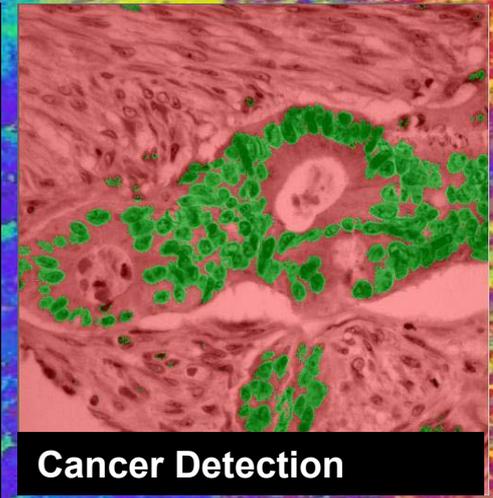
GENetic Imagery Exploitation



New York City



Cerro Grande Wildfire



Cancer Detection

- Machine Learning/Artificial Vision project
- Applications in:
 - Disaster response
 - Environmental monitoring
 - Cartography
 - Medical Imagery
- Winner of a *R&D 100 Award* for 2002



LAUR 03-7185



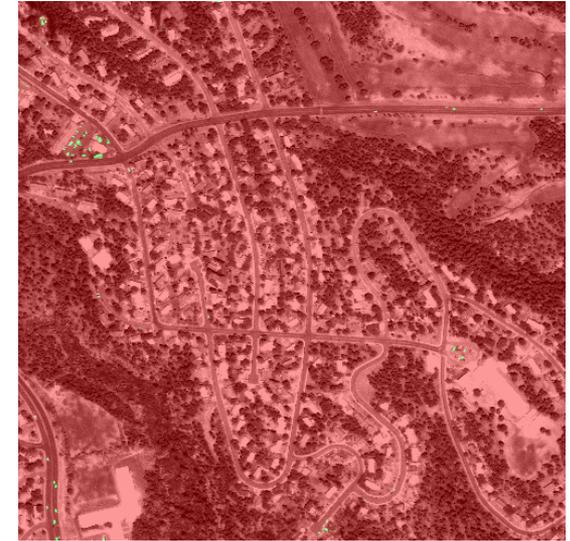
Finding Vehicles



Image data



Training Data (markup)



Results

- Finding vehicles in multispectral imagery
- Analyst marks (a small piece of) image data
 - green is "true" -- red is "false"
- ML algorithm finds classifier that distinguishes "true" from "false"
- Classifier is applied to the rest of the image (or to other images)

POOKA and VOOKA - Accelerating ISIS tools

Deployable data analysis for large volume/high throughput sensors

POOKA - Reconfigurable hardware system inspired by GENIE

- Prototype deployed at NIMA and CIA for accelerating BAS

VOOKA - Extension of POOKA to real-time video exploitation

- High performance hardware for generating data products in real-time

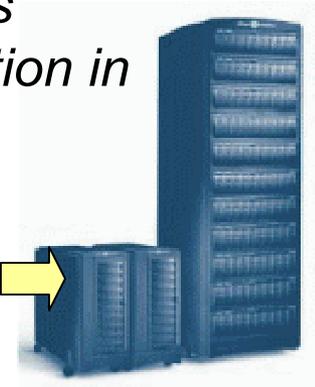
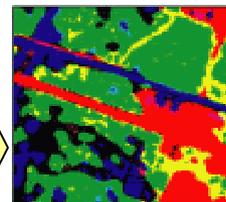
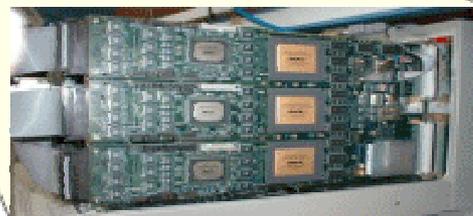
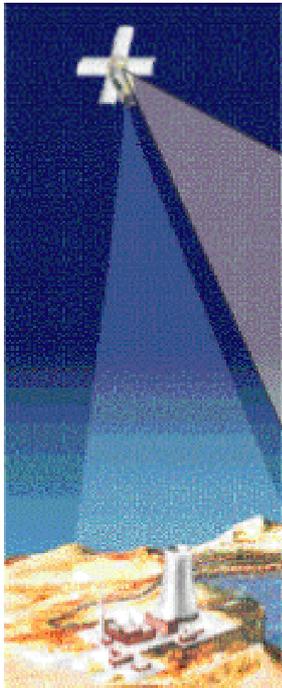
REQUIREMENT

Correlate objects seen in imagery with simple text

Example: Predator

Need ontology for both representations

Hard problem: bringing image information in



Example KDD Projects

- BioSurveillance Decision Environment

BioSurveillance Environment

- Operations Center to support collaborative Decision Making across several agencies – *local, state, fed*
- Decisions here are threat levels – *standby, alert, incident, recovery*
- Data → Data Integration → Decisions → Actions → Data
- Disparate Data and Information with uncertainty
- Integration and Quantification of the Information – Computationally and Statistically
- System of Systems Approach

Current state:

INCIDENT

Frame will be updated every 60 seconds.



Current date and time:

Thu 02/20/2003 14:23:27

Current login: John Martinez

Environmental Monitoring

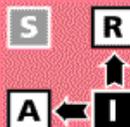
- [Airport](#) ✓ on time
- [TIC](#) ✓ on time

Medical Surveillance

- [Bsafer](#) ● 10.7 hrs late
- [Essence](#) ✓ on time
- [RSVP](#) ✓ on time

Models

- [NARAC](#) ✓ on time
- [HPAC](#) ✓ on time



Force Update

View Log

View GIS Map

New Urgent Messages

[View All](#)



Received: 02/20/2003 13:38:11

From: Laboratory - Manager
To: NMDoH Office of Epidemiology/State Epidemiologist, TOC Supervisor and Deputy, LLNL Senior Microbiologist List

Subject: [Confirmed presence of Bacillus anthracis](#)

Message: **NEW** A confirmatory positive PCR result for anthrax (*B. anthracis*) has been obtained from a sample retrieved from the Del Norte sampling station, at San Mateo and Montgomery, at Del Norte High School.
NEW Positive readings were obtained on each of the five new loci tested.

The aerosol sample tested was drawn some time between Day 0, 1900 and Day 1, 0700.

Threshold crossing times indicate that the sample contains a relatively large quantity of genetic material.

No other positive results have been obtained to date for this or any other agent

Operations for Incident State

- ✓ 1. [Update/send situation report.](#)
- ✓ 2. [Review appropriate subsystem data.](#)
- 3. [Change state to ALERT if the message is denied or brought into question.](#)
- 4. [Change state to RECOVERY if incident has been sufficiently addressed.](#)
- 5. [Perform INCIDENT state conference call.](#)

Consider performing the following actions if appropriate:

- ✓ 6. [Contact subject matter experts.](#)
- 7. [Request sensor time-resolution testing.](#)
- 8. [Request sampling rate increase of sensors.](#)
- 9. [Request sampling rate increase of medical surveillance.](#)
- ✓ 10. [Request plume model.](#)
- 11. [Request epidemiological model.](#)

LINKS

- [Traffic Cam - Big I](#)
- [Traffic Reports](#)
- [Weather - Local](#)
- [City Links](#)

ASSOCIATED PRESS

- 02/20/2003 12:51:34
- [Kids Evacuated From Ivory Coast City](#)
- [India Troops Seek to Prevent Riot](#)
- [Rumsfeld: Germany Has to Repair Ties](#)
- [Gunmen Kill 7 in Pakistan](#)
- [U.S., Canada Praise Iraq Dossier](#)
- [Israel Destroys 3 Palestinian Homes](#)

LOCAL NEWS



KOB-TV

CNN



AFP PHOTO

Log - TOC Integration Console - Netscape

File Edit View Go Communicator Help

Operation Scheduled conference call (see details next item).
02/21/2003 17:38:06

Time Call 1: 2000
Time Call 2: 2200
Participants: DOH, Epidemiology, FBI, Lab Director

Operation Completed ALERT state conference call.
02/21/2003 17:38:15

Message: 11 - URGENT
Dates: 02/21/2003 17:38:23 received; 02/21/2003 17:38:23 stamped
From: Laboratory - Manager
To: NMDoH Office of Epidemiology/State Epidemiologist, TOC Supervisor and Deputy, LLNL Senior Microbiologist List
Subject: [Confirmed presence of Bacillus anthracis](#)
Message: **NEW** A confirmatory positive PCR result for anthrax (*B. anthracis*) has been obtained from a sample retrieved from the Del Norte sampling station, at San Mateo and Montgomery, at Del Norte High School.
NEW Positive readings were obtained on each of the five new loci tested.
The aerosol sample tested was drawn some time between Day 0, 1900 and Day 1, 0700.
Threshold crossing times indicate that the sample contains a relatively large quantity of genetic material.
No other positive results have been obtained to date for this or any other agent, from any sample.
NEW We are proceeding with tests from the same sample set to obtain better time resolution. These tests should be complete about 1700, and will determine the collection time to within a one-hour window.
NEW A presumptive positive result was reported earlier for this sample at 1201.

Operation Reviewed appropriate subsystem data.
02/21/2003 17:38:45

Operation Dialed BDI subsystem manager to request sensor time-resolution testing.
02/21/2003 17:39:00

INCIDENT **State changed to INCIDENT**
02/21/2003 17:39:06

Operation Updated and sent situation report for INCIDENT state (see details next item).
02/21/2003 17:39:53

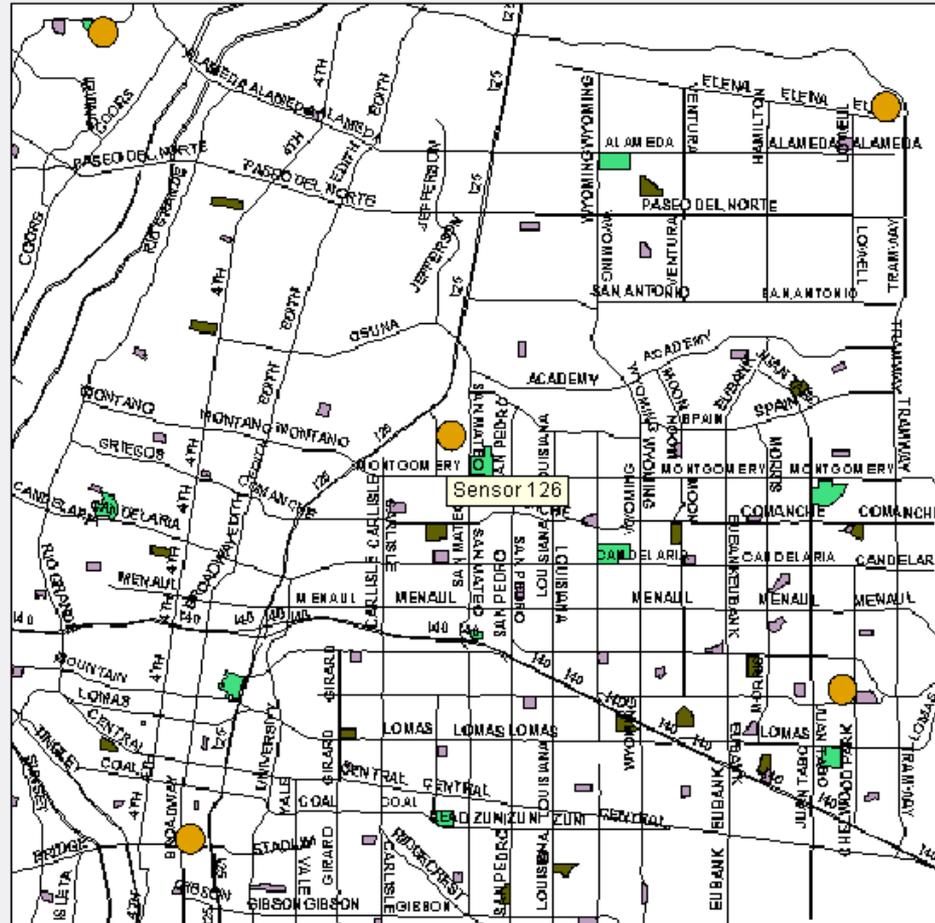
Summary: confirmed anthrax
Information: time resolution still large requesting more time resolution
Follow up:
Notes:

Document: Done



Click on sensor for latest status

Close Window



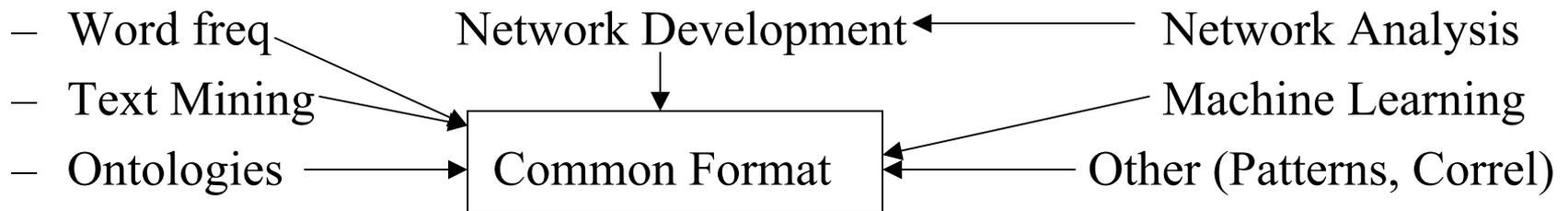
[Go to Aerial Photo of Map](#)
[Go to Other Mappings](#)

Example KDD Projects

- Intelligence Analyst Support Environment
 - Collaborative Notes-based Infrastructure
 - Ontology Centered
 - Text Mining
 - Network/Link Analysis

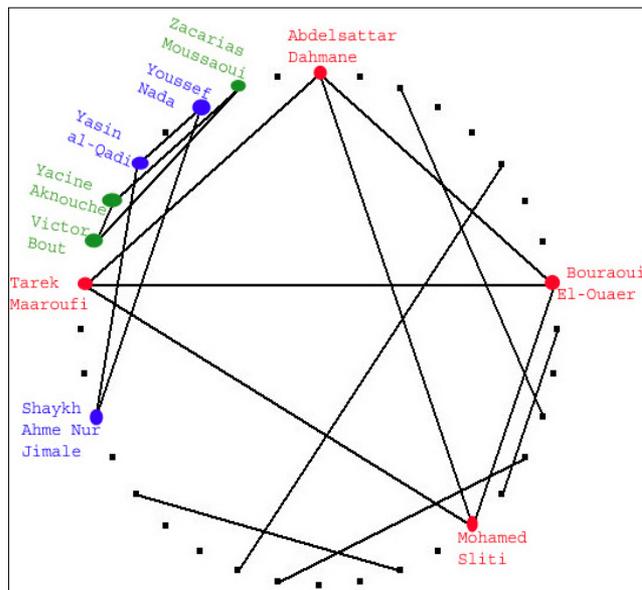
KDD Intelligence Project

- Intelligence analysts have need for tools to support data mining and integration of multiple data sources – *today they do this by hand*
- **Intelligence Analyst Workflow:**
 - Data → Filter → Database → Annotate → Analyze → Threats and Reports
- **Techniques to Apply:**

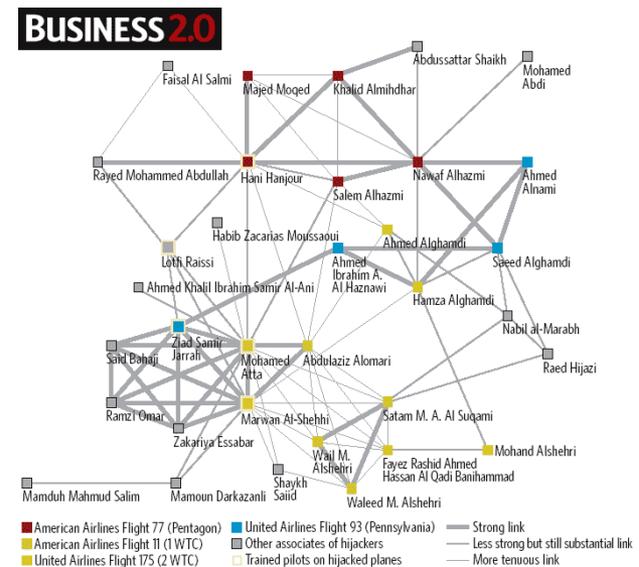


Social Network Analysis

- Identify terrorist social networks
- Predict network evolution
- Find implicit, latent connections
- Model disruptions of social and infrastructure networks
- Adaptive retrieval, recommendation from knowledge networks



LM Rocha, "Proximity and Semi-Metric Analysis of Social Networks", LAUR, 2002



TA Stewart, "Six Degrees of Mohammed Atta", *Business 2.0*, Dec. 2001



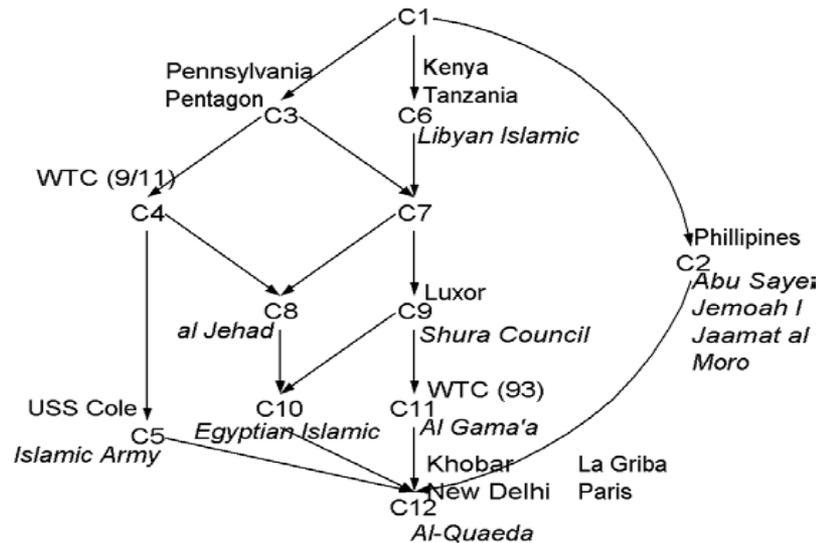
Formal Concept Analysis

- Galois lattices show hierarchical sub-relations in complex datasets
- Unbiased, graphical, visual representation
- Hypothesis and rule generation and evaluation
- Support intelligence analysts in knowledge management

W T N P
L T a e e
W a K C n w n
K P T G h L z P D t
e h C r P C o u 9 a a e a
n i i e o b x / n r l g
y l 9 b n l a o l i i h o
a l 3 a n e r r l a s i n

GROUPS	EVENTS																			
1	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
3	X				X			X	X	X										X
4	X																			
6	X																			
7	X				X			X		X										X
8	X				X	X		X												X
9	X				X			X	X											X
11	X	X	X		X			X	X											X
12	X									X										
16	X																			
26	X																			

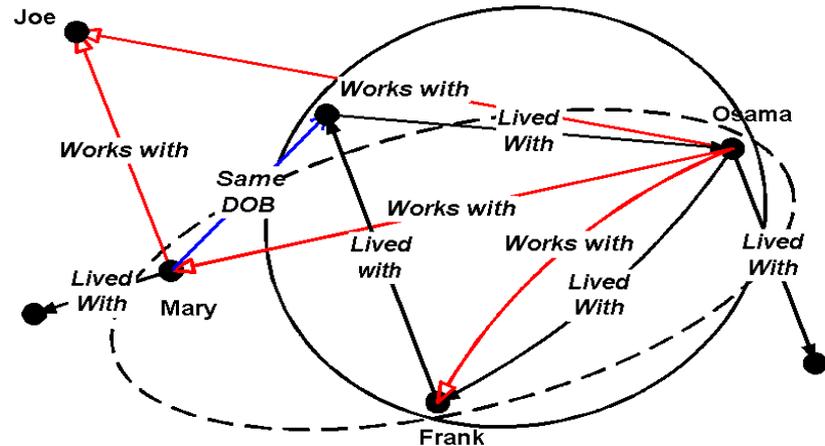
Al Qaeda
Eg. Islamic Jihad
Abu Sayyaf
Jemmah Islamiyah
Shurah Council
Islamic Army of Aden
al Jihad
Al-Gama'a al Islamiyya
Libyan Islmaic Group
Jaamat al-Islamie
Moro Islamic Lib Front



Voss, Susan and Cliff Joslyn (2002): "Advanced Knowledge Integration in Assessing Terrorist Threats", LAUR 02-7867 <ftp://ftp.c3.lanl.gov/pub/users/joslyn/knowint.pdf>

Multi-Dimensional Link Analysis

- Methods for guided knowledge discovery in computationally huge, heterogeneous databases
- Typed-link representations of social meta-networks
- Assist analysts to identify relevant fields, locate areas of high local structure in databases
- Identification of terrorist social networks



$$D_{\{f\}}, \{w, y, z\} \rightarrow D'_{\{g\}}, \{x, w, z\}$$

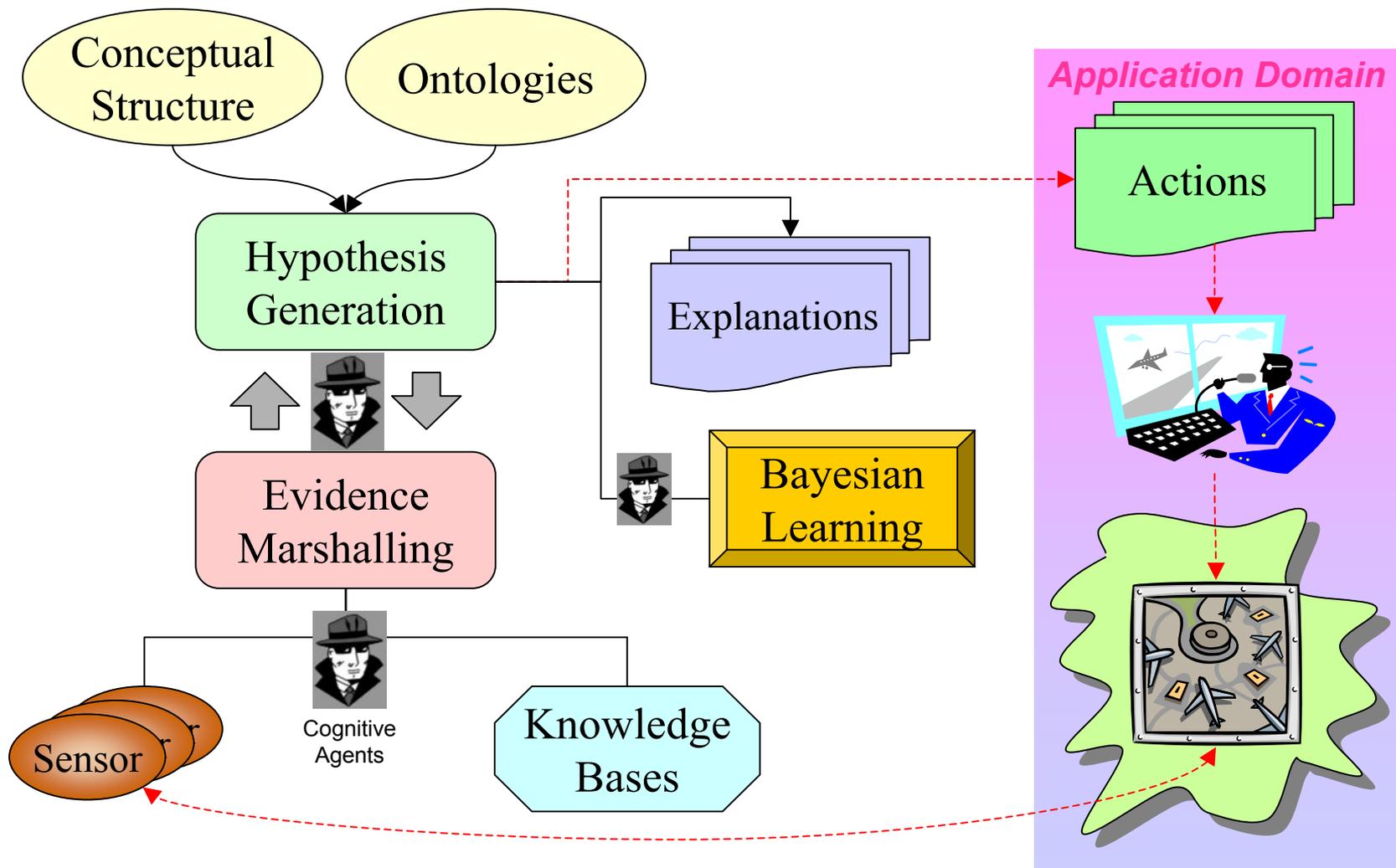
Joslyn, CA: (2002) "Link Analysis of Social Meta-Networks", *Proc. CASOS 2002*

Voss, Susan and Cliff Joslyn (2002): "Advanced Knowledge Integration in Assessing Terrorist Threats", LAUR 02-7867
<ftp://ftp.c3.lanl.gov/pub/users/joslyn/knowint.pdf>

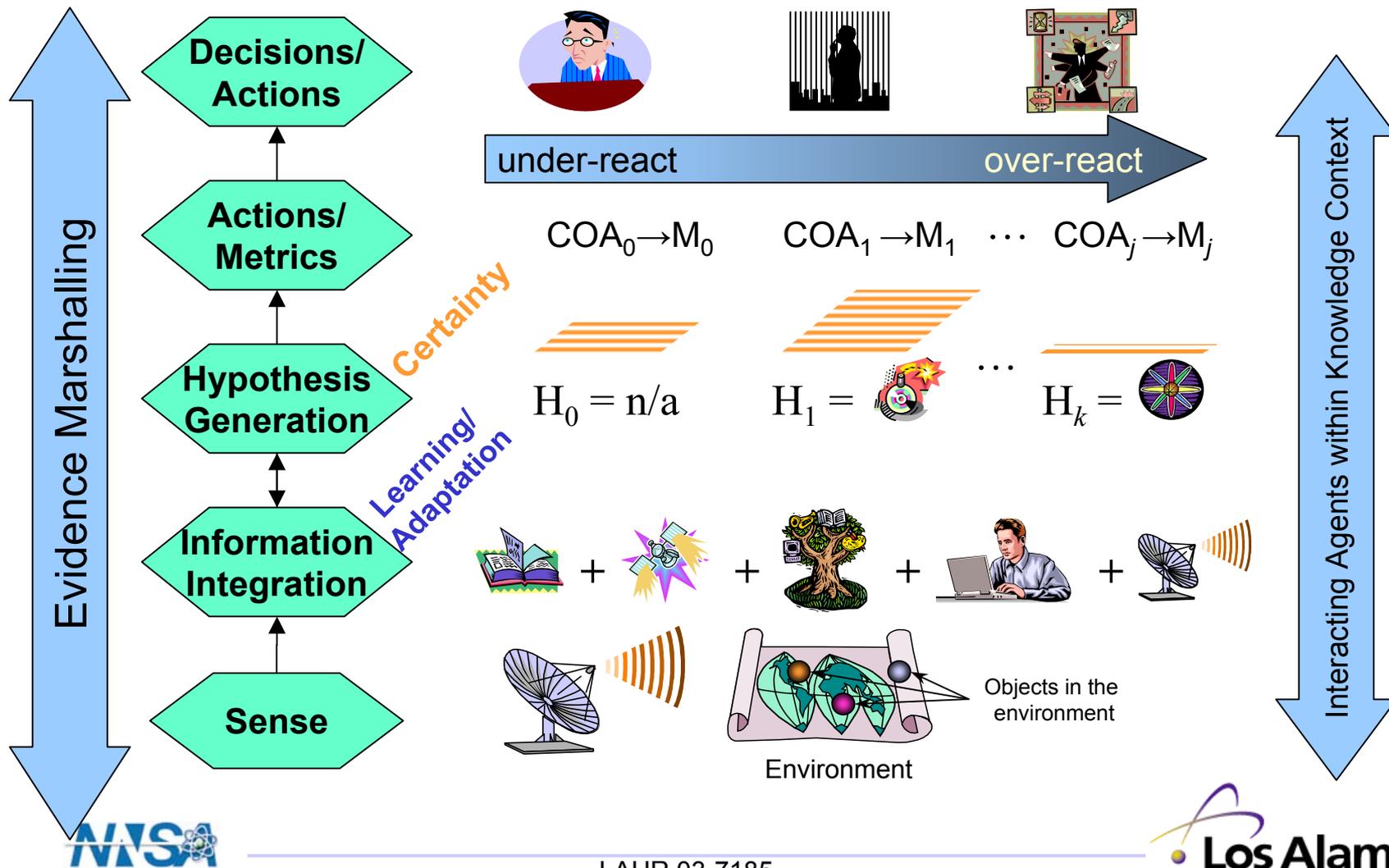
Example KDD Projects

- Tactical Information Integration System

Decisions in Time Critical Environments



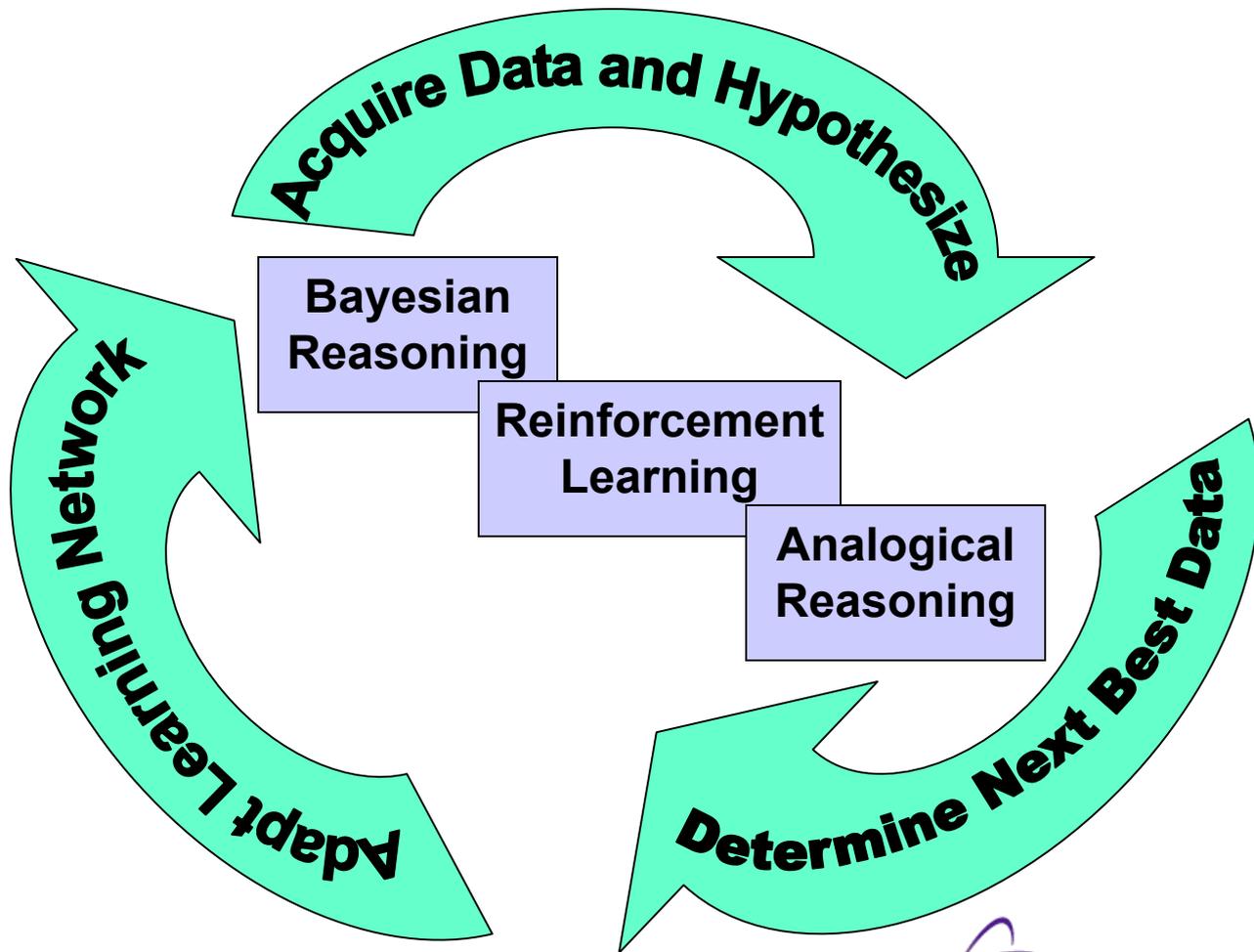
Decision Hypotheses through Evidence Marshalling



Evidence Marshalling and Adaptive Agents

Technical Goals:

- Evidence Marshalling
- Adaptation
- Information Integration
- Learning and Reasoning



Approach: Bayesian Reasoning

- **Model-driven**

- Framework that supports probabilistic integration of diverse data sources
- Define network dependencies and joint probability distributions through knowledge modeling
- Reasoning = update local variable distributions or network structure with new data
- **Build on LANL Information Integration work**

- **Innovation**

- Realtime data inputs to drive local probability distribution
- Evidence Marshalling to reduce uncertainties by retrieving relevant data
- Network nodes as semi-autonomous agents for control and computation

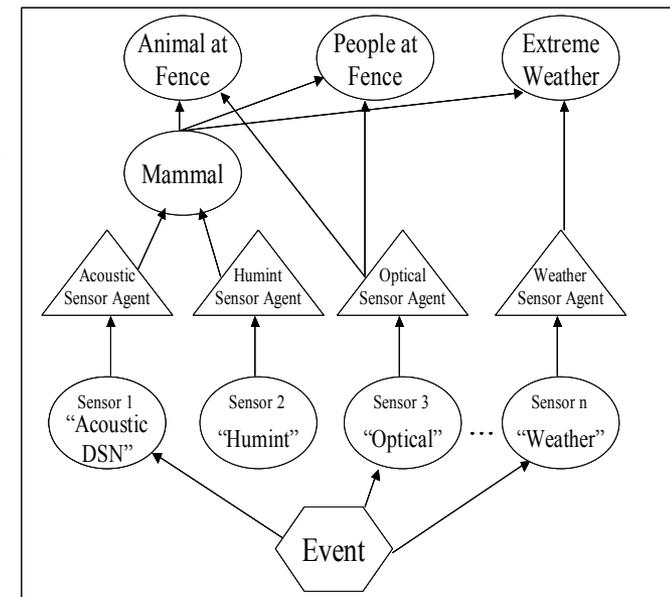


Figure 1 – Simplified Physical Site Security Example

Approach: Reinforcement Learning

- **Example-driven learning**
 - Agents use local strategies to choose actions that optimize global utility functions
 - Agents follow actions of other successful agents
 - Learning = learn best strategies and learn network structure through agent interactions
 - **Build on LANL agent collectives work**
- **Innovation**
 - Realtime data inputs
 - Extends simple minority market game into complex threat assessment domain
 - Addition of hypothesis flow through hierarchy of agents
 - Combine with Bayesian Reasoning for increased adaptation

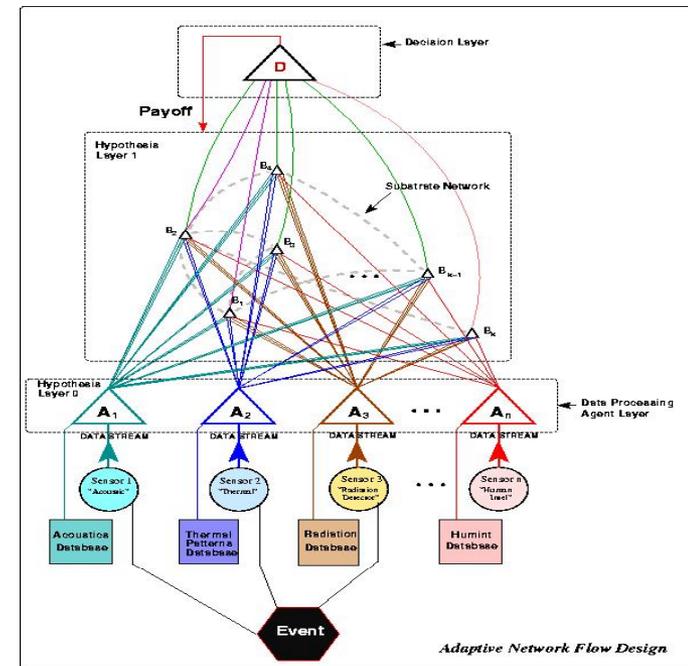


Figure 3. Agent hierarchy in network flow design

Analogical Reasoning

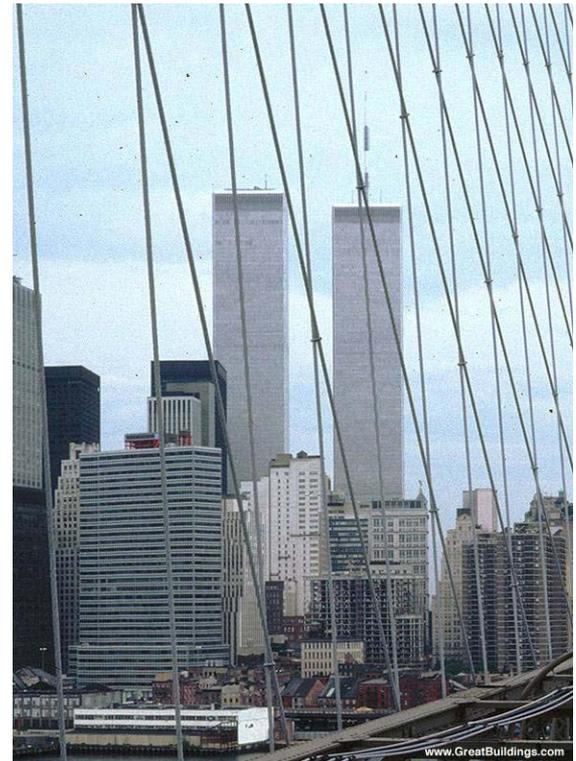
- **Example-driven reasoning**
 - Determine similarities related to network structure and semantic concepts
 - Reason related to similarities and differences
 - Learning = store and retrieve closest analogs
 - **Build on LANL learning and reasoning by analogy research**

- **Innovation**
 - Realtime data inputs
 - Use Analogues to add new hypotheses
 - Combine with Bayesian and Reinforcement Learning for full adaptation

Example KDD Projects

- Collaborative Forestry Management and Response System

The Forest Fire Problem



French Forest Fire Incident

Europe - AP

Four Killed in French Forest Fires

Tue Jul 29, 10:42 AM ET

By PHILIPPE FORTIN, Associated Press Writer

SAINTE-MAXIME, France - Forest fires swept through the French Riviera on Tuesday, killing at least four people, devastating scenic woods and forcing the evacuation of thousands of people.

Forest Fires in Southern France

Forest Fires Rage Through French Riviera

Forest Fires in Southern France Leave Two Dead

Firefighters speculated that the fires — some 30 that broke out nearly simultaneously on Monday — were caused by arson. Molotov cocktails were found in the region, according to radio and television reports. The mayor of Roquebrune-Sur-Argens, Luc Jousse, called the fires "a new form of terrorism."



The Technical Issues

- Acquiring data existing in every digital or analog form from widely distributed, mission specific data sources
- Turning data into knowledge
- Presenting and updating knowledge

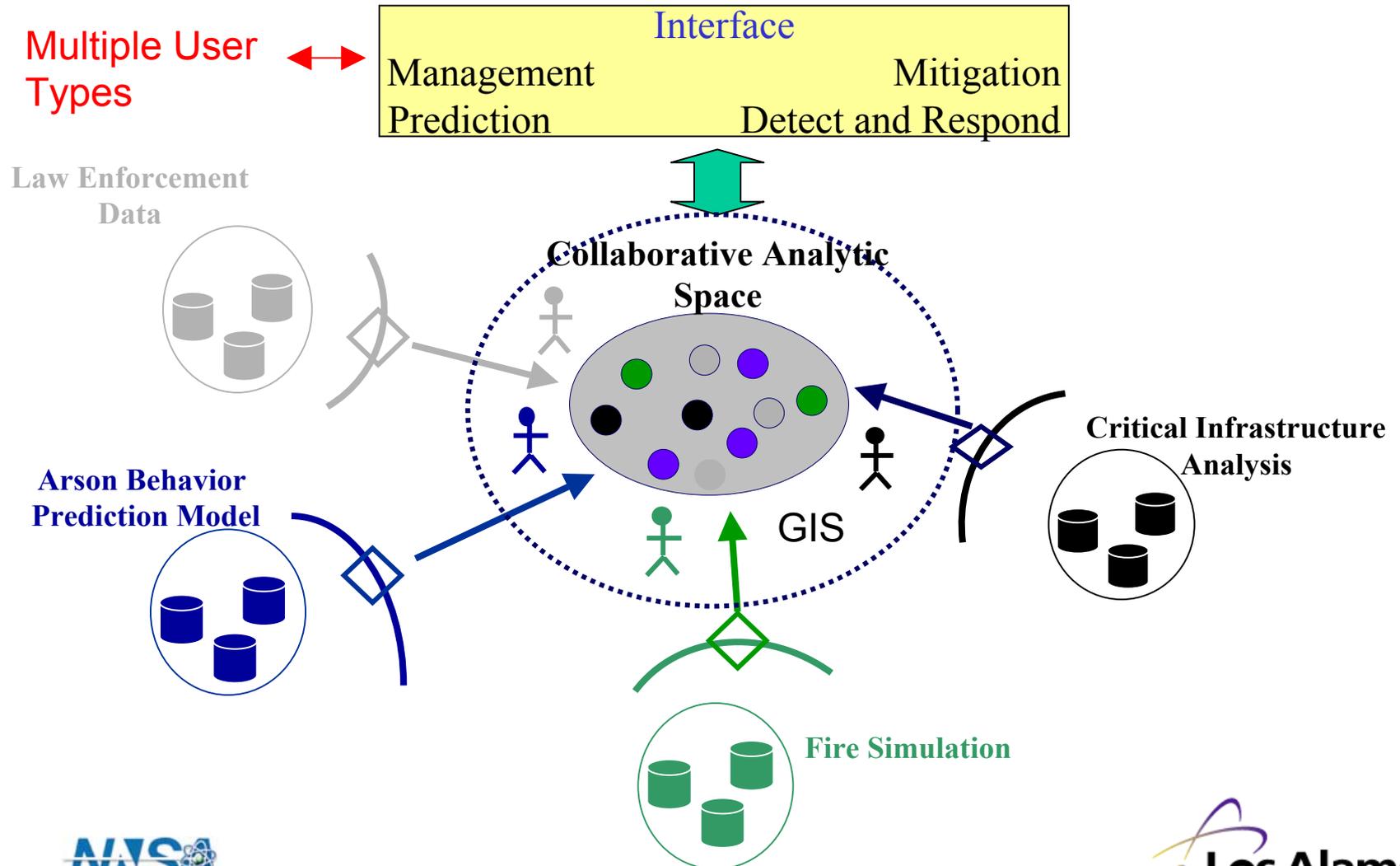
To achieve a global situational awareness

The Policy Issues

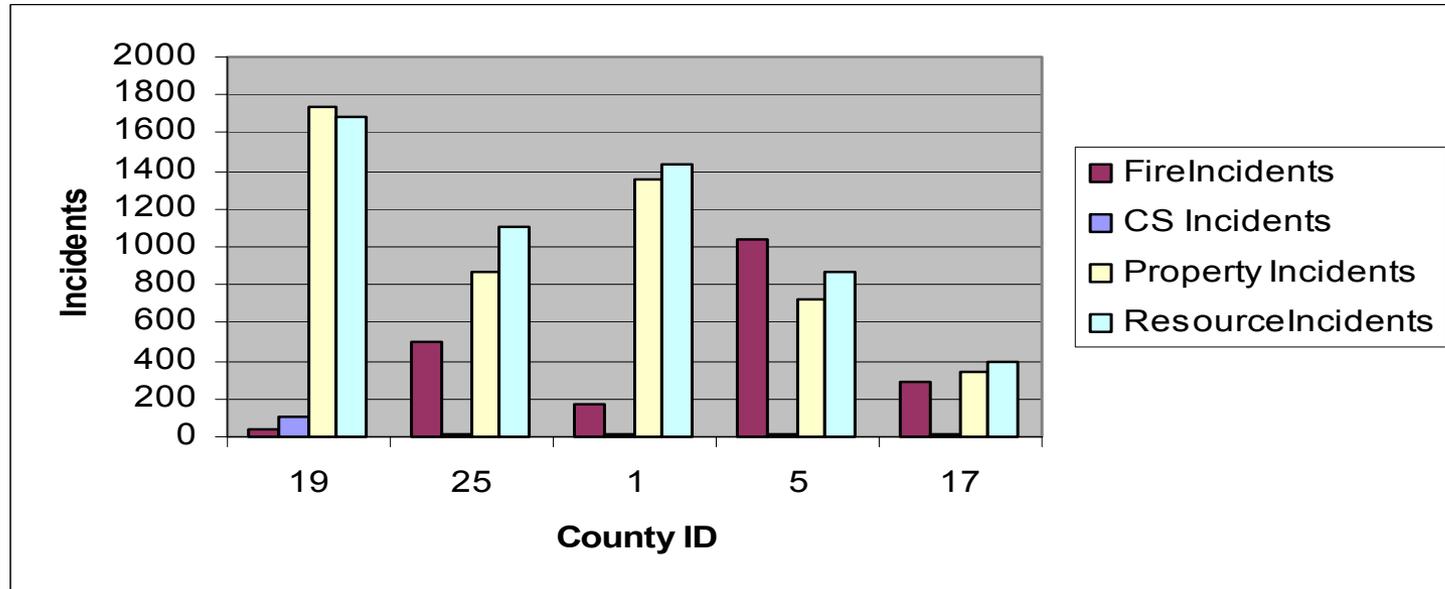
- Achieving cooperation across agencies that do not or may not share information



Forestry Management and Response System



Locate Hot Incident Spots



Example output

Proposed Fire Plan Maps Integrated with Human Caused Fire Risk Rating

