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
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

Dissemination
- Decision Support
- Knowledge Management
- Collaboration
- Visualization

Integration
- Information Fusion
- Portals
- Information Access

Mining
- Classification
- Pattern Recognition
- Data Reduction

Marshaling

DATA
- Unstructured Text
- Rule Base
- Streaming Video
- Sound/Voice
- Acoustic Signals

INFORMATION
- Concepts
- Models
- Inferences
- Objects
- Correlations
- Filtered Signals

KNOWLEDGE
- Hypotheses
- Cases
- Events
- Threats
- Situations
- Analyses
- Policies
- Actions

LAUR 03-7185
LANL Capabilities Span the Pyramid

• **Mining**
  – Machine Learning
  – Statistical Techniques
  – Image Understanding
  – Image Compression
  – Text Mining
  – Information Retrieval

• **Integration**
  – Information Integration
  – Knowledge and Social Networks
  – Ontologies
  – Probability and Uncertainty
  – Agents

• **Dissemination**
  – Decision Support
  – Visualization

• **Marshaling**
  – Evidence and Data Marshaling

LAUR 03-7185
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

Data Integration
Uncertainty Quantification

Multi-User Needs

Decision Models Visualization

Texts
- Gather
- Pattern Extraction
- Comparison
- Classifier

Images
- Gather
- Fusion
- Pattern Extraction
- Comparison
- Classifier

Signals
- Gather
- Feature Recognition
- Fusion
- Classifier

Databases
- Knowledge Discovery
- Pattern Extraction
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
  – AFREEET
  – 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
• Machine Learning/Artificial Vision project
• Applications in:
  – Disaster response
  – Environmental monitoring
  – Cartography
  – Medical Imagery
• Winner of a *R&D 100 Award* for 2002
Finding Vehicles

- 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
INCIDENT

Current state:

Frame will be updated every 60 seconds.

Environmental Monitoring
- Airport: on time
- TC: on time

Medical Surveillance
- Ender: on time
- Essence: on time
- ESRP: on time

Models
- NARR: on time
- HPAC: on time

New Urgent Messages

Received: 02/29/2003 13:38:11

From: Laboratory - Manager
To: NMDCH Office of Epidemiology/State Epidemiologist, TOC Supervisor and Deputy, LLNL Senior Microbiologist List
Subject: Confirmed presence of Bacillus anthracis
Message: 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.
- Positive readings were obtained on each of the five new loci tested.
- The aerosol sample tested was drawn some time between Day 9, 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 models.
**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:** NM DoH 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, 1000 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**

Dialled 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:**
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:
  – Word freq
  – Text Mining
  – Ontologies

  ![Diagram](attachment:diagram.png)
Analyst Environment

- Computationally vast, heterogeneous databases of social meta-networks:
  - People, aliases, places, organizations, events, times. . .
- **Two-Dimensional:** Find structure in pre-defined variables
- **Multi-Dimensional:** Find collections of relevant variables
- Complementary relational and object-network views
- User-guided knowledge discovery and visualization

User-Guided Relational Knowledge Discovery

![Image of database relationships and network analysis](image_url)

LAUR 03-7185
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

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

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

\[ D_{\{ij\},\{w,y,z\}} \rightarrow D'_{\{g\},\{x,w,z\}} \]


Example KDD Projects

- Tactical Information Integration System
Decisions in Time Critical Environments

- Conceptual Structure
- Ontologies
- Hypothesis Generation
- Evidence Marshalling
- Knowledge Bases
- Bayesian Learning
- Explanations
- Applications Domain
- Actions

Sensor

Cognitive Agents

Los Alamos

LAUR 03-7185
Decision Hypotheses through Evidence Marshalling

- Decisions/Actions
- Actions/Metrics
- Hypothesis Generation
- Information Integration
- Sense

Evidence Marshalling

Certainty

Learning/Adaptation

COA_0 \rightarrow M_0 \quad COA_1 \rightarrow M_1 \quad \cdots \quad COA_j \rightarrow M_j

H_0 = \text{n/a} \quad H_1 \quad \cdots \quad H_k =

under-react \quad over-react

Interacting Agents within Knowledge Context

Objects in the environment

Environment

+ + + +

+ + + +

+ + + +

+ + + +

NWSA

Los Alamos

LAUR 03-7185
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](image-url)
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
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

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

Management Prediction
Mitigation
Detect and Respond

Multiple User Types
Law Enforcement Data
Arson Behavior Prediction Model
Critical Infrastructure Analysis
Fire Simulation
GIS

Collaborative Analytic Space

Los Alamos
LAUR 03-7185
Locate Hot Incident Spots

Example output
Proposed Fire Plan Maps Integrated with Human Caused Fire Risk Rating

![Map of National Fire Plan Maps with legend showing different risk ratings: Very High, Extremely High, High Arson Risk.](map_image)