UT-BATTELLE

Knowledge Discovery

Brian Worley, Director
Computational Sciences and Engineering Division

OAK RIDGE NATIONAL LABORATORY
U. S. DEPARTMENT OF ENERGY

Scientific and Technological Motivation for Knowledge Discovery

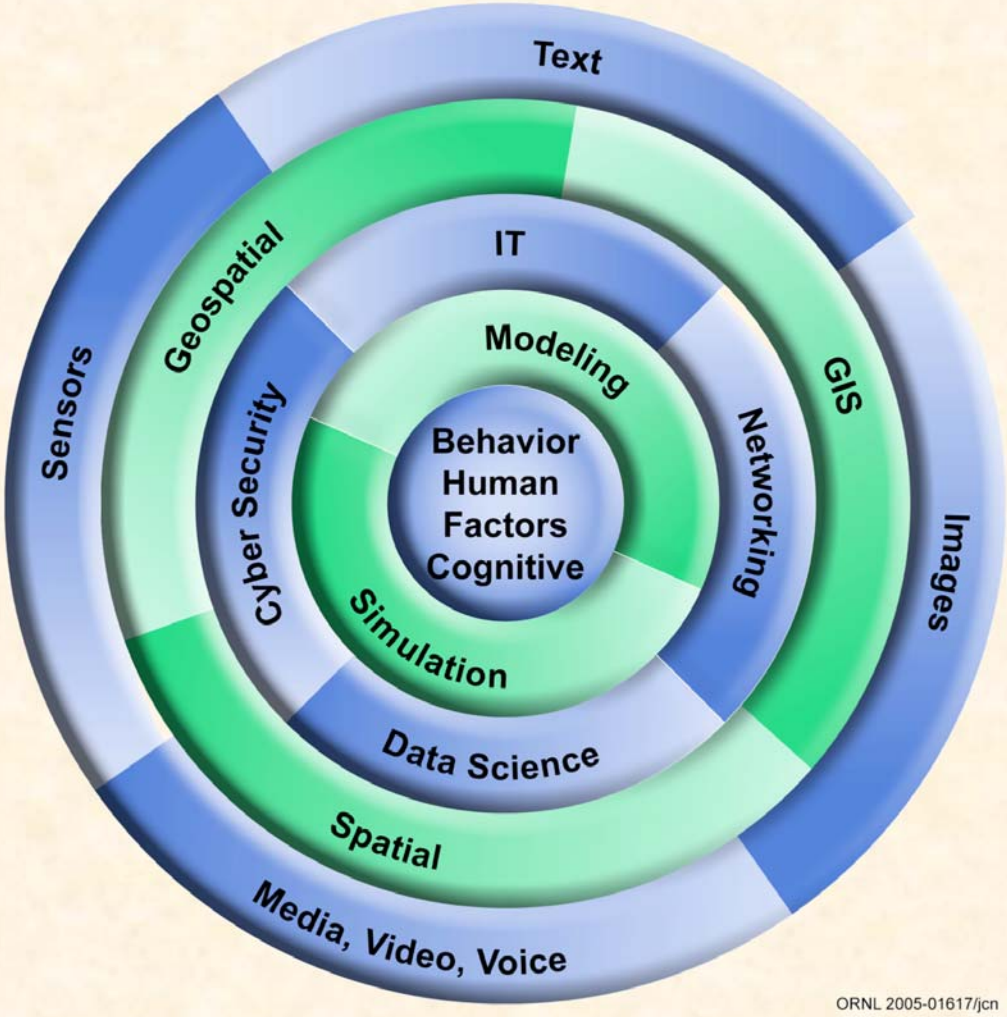
Media type	Tbyte s
Newsletters (40,000 titles)	1
CD-ROMs (850 titles)	1
Scholarly journals(37,609 titles)	6
Books (950,000 titles)	39
DVD videos (4,000 titles)	44
Mass-market periodicals (80,000 titles)	52
Audio CDs (33,443 titles)	58
Newspapers (25,276 titles)	138
Searchable Web	167
Instant messaging	274
Zip disks (1.4 million)	350
Floppy disks (55 million)	800
Office documents (10.75 billion pages)	1,397
Audio minidisks (10.5 million)	1,700

Media type	Tbytes
Flash memory (43 million)	2,200
X-rays (2 billion)	20,000
Motion pictures (10,342)	25,000
Deep Web	91,850
Audio tapes [analog] (128.8 million)	128,800
Digital tapes (5 million)	250,000
Photographs (75 billion)	375,000
E-mails (originals)	440,606
Digital video (115 million)	1,265,000
Video tape (VHS) and camcorder (220 million)	1,340,000
Hard disk drives (44 million)	1,986,000

The searchable Internet (purple) contains only a fraction of the information stored on digital media

Source: MIT Technology Review, 2005

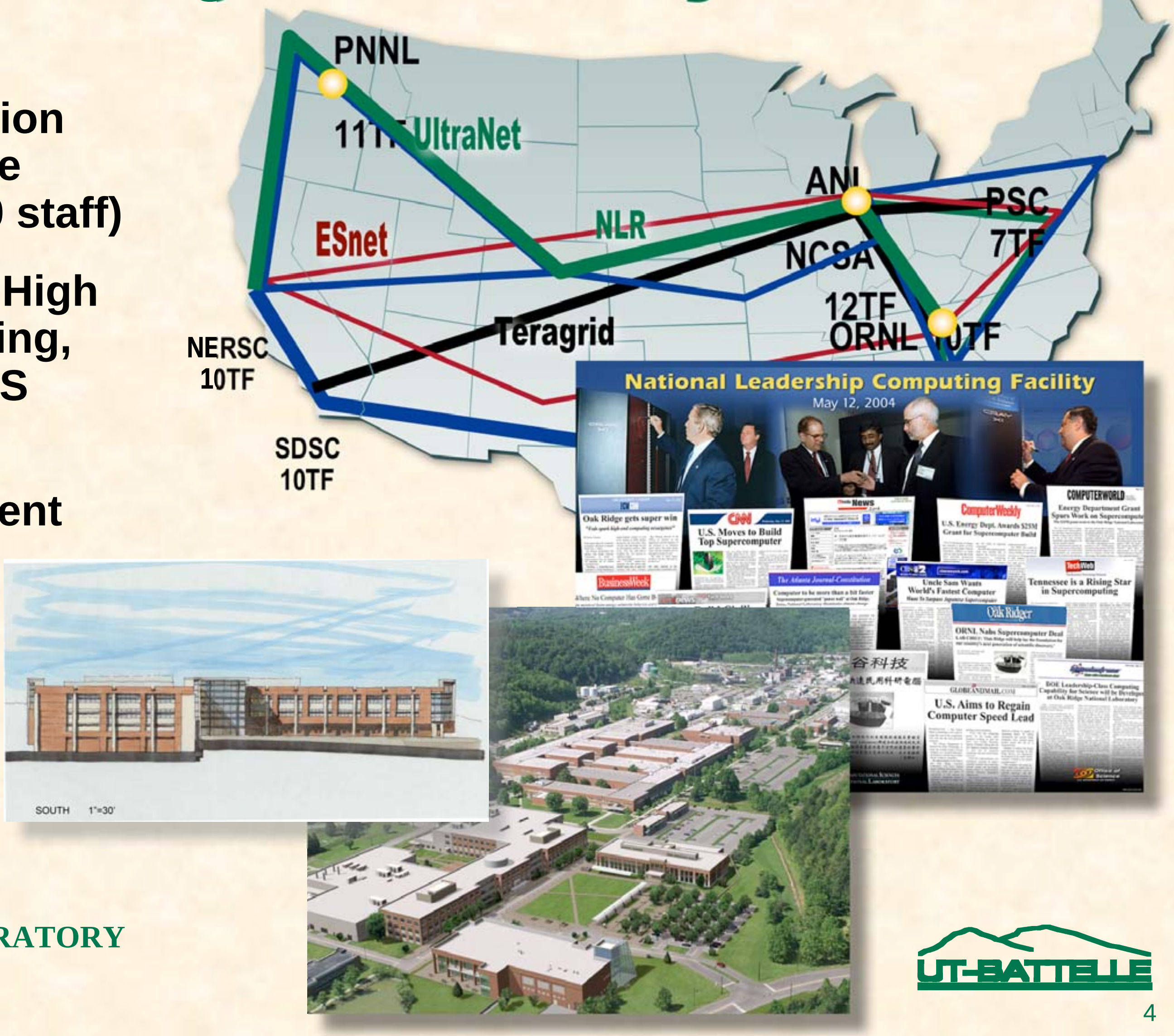
CSED Scientific Thrusts Revolve around Knowledge Discovery from Diverse, Dynamic, Large Datasets



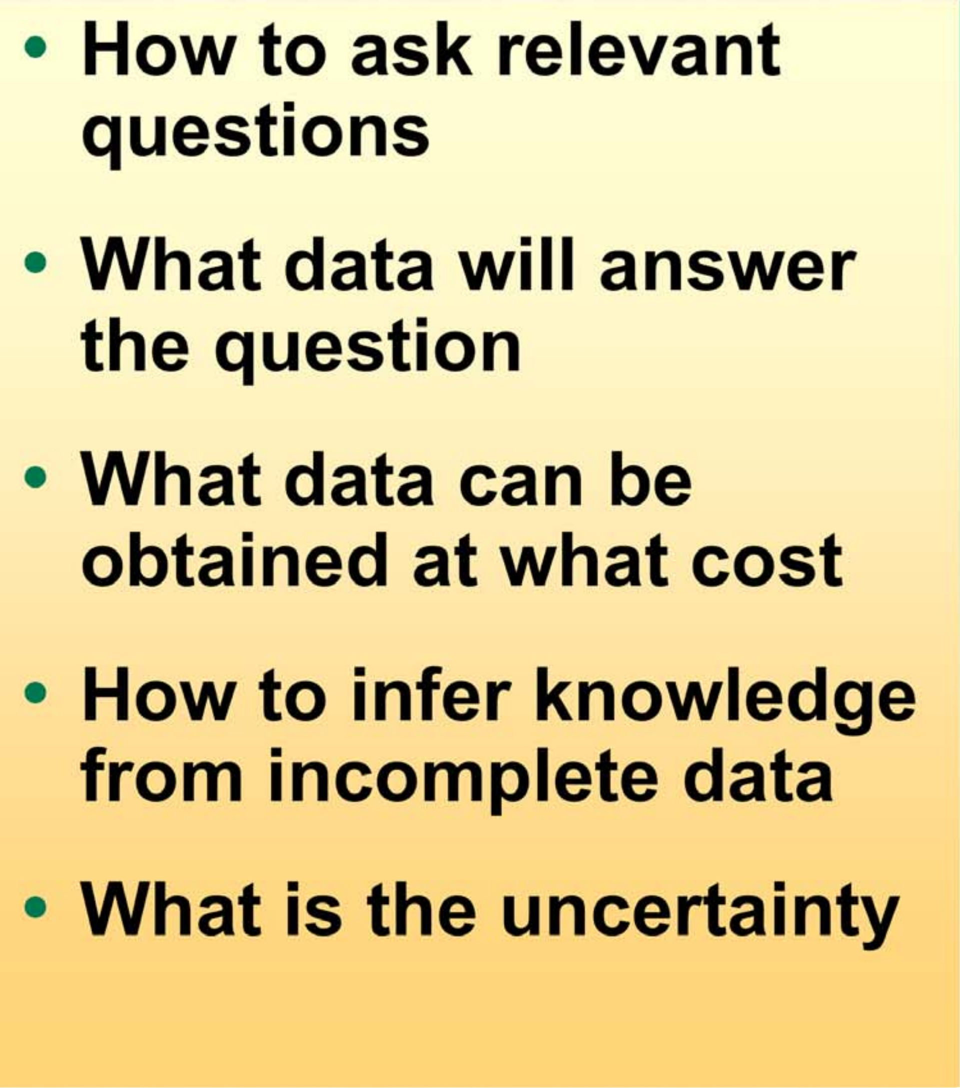
ORNL 2005-01617/jcn

ORNL Commitment to Science for Knowledge Discovery

- **Entire Research Division Devoted to Knowledge Discovery (CSED, 150 staff)**
- **Physical Resources: High Performance Computing, Networking, MRF, JICS**
- **LDRD Initiative in Knowledge Management**


PNNL
11TF
UltraNet
ESnet
NLR
ANL
PSC
7TF
NCSA
12TF
ORNL
10TF
Teragrid
NERSC
10TF
SDSC
10TF
National Leadership Computing Facility
May 12, 2004
Oak Ridge gets super win
U.S. Moves to Build Top Supercomputer
ComputerWeekly
U.S. Energy Dept. Awards $25M Grant for Supercomputer Build
COMPUTERWORLD
Energy Department Grant Spurs Work on Supercomputer
BusinessWeek
The Atlanta Journal-Constitution
Computer to be more than a bit faster
Uncle Sam Wants World's Fastest Computer
TechWeb
Tennessee is a Rising Star in Supercomputing
Oak Ridger
ORNL Nabs Supercomputer Deal
谷科技
GLOBEANDMAIL.COM
U.S. Aims to Regain Computer Speed Lead
DOE Leadership-Class Computing Capability for Science will be Developed at Oak Ridge National Laboratory
SOUTH 1"=30'

Questions to Data



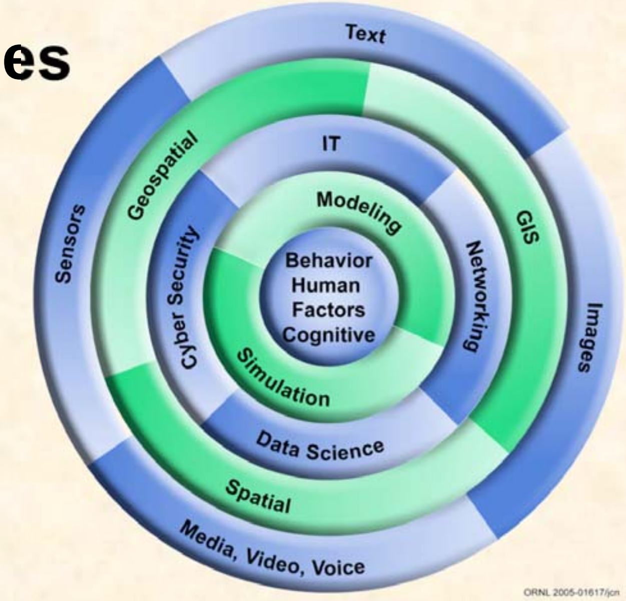
- How to ask relevant questions
- What data will answer the question
- What data can be obtained at what cost
- How to infer knowledge from incomplete data
- What is the uncertainty

Data to Questions

- What useful data do we have or can obtain
- What trends or patterns exist in the data
- What is the class of questions that can be answered from this data
- What additional data would be most useful

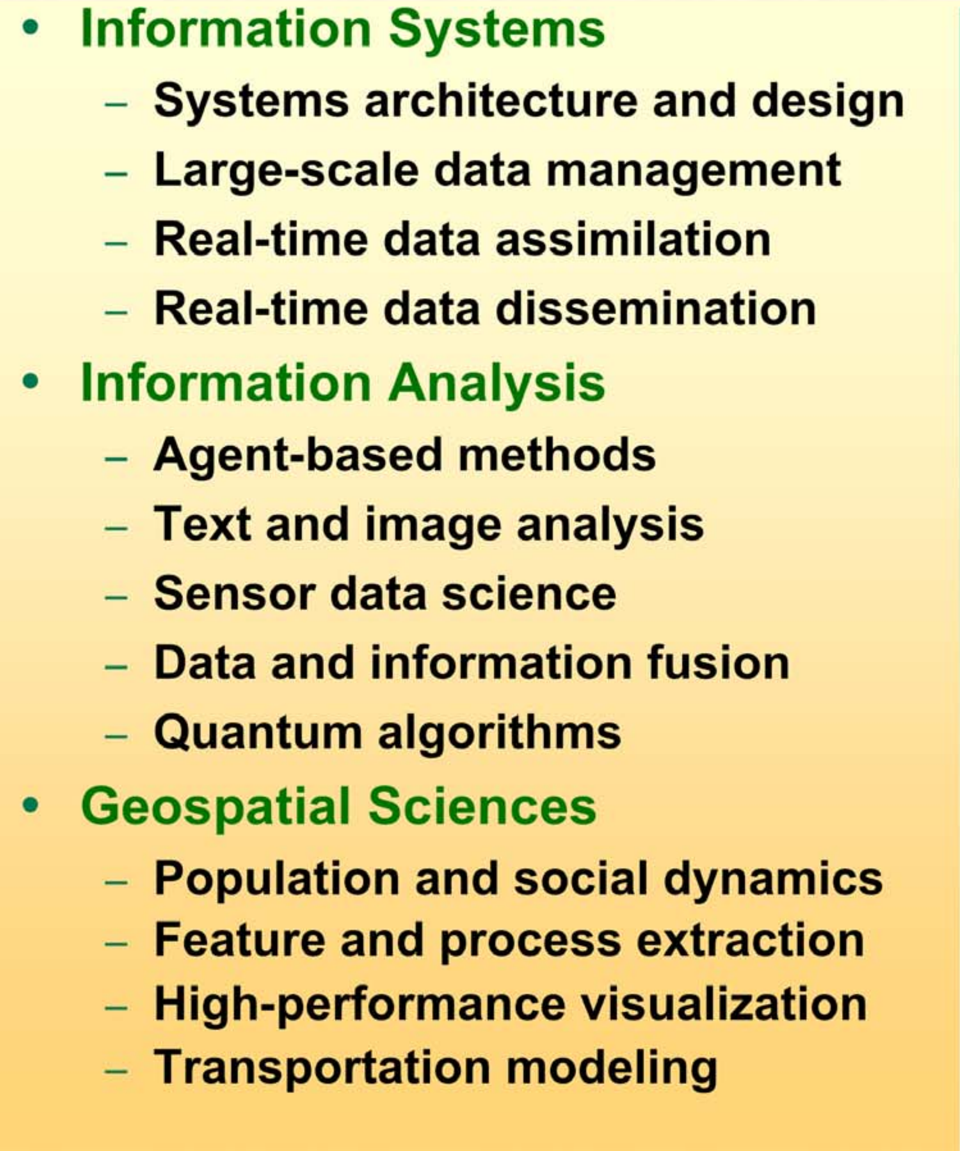
CSED Scientific Thrusts Revolve around Knowledge Discovery from Diverse, Dynamic, Large Datasets

- **Data Sources: data integrity, security, and policy**
- **Data Assimilation: open and/or covert**
- **Geospatial and temporal attribution**
- **Data management; networking issues**
- **Modeling and analysis**
- **Interpretation**
- **Data Dissemination**
- **Additional decision support**



ORNL 2005-01617/jcn

CSED Core Research Areas



- **Information Systems**
  - Systems architecture and design
  - Large-scale data management
  - Real-time data assimilation
  - Real-time data dissemination
- **Information Analysis**
  - Agent-based methods
  - Text and image analysis
  - Sensor data science
  - Data and information fusion
  - Quantum algorithms
- **Geospatial Sciences**
  - Population and social dynamics
  - Feature and process extraction
  - High-performance visualization
  - Transportation modeling

- **Information Security**
  - Multi-level access
  - Authentication and Trust
  - Information Assurance
  - Quantum Information Systems
- **Decision Sciences**
  - Man/Machine Interfaces
  - Time-critical mission support
  - Behavioral Sciences
  - Cognitive Inference
- **Modeling and Simulation**
  - Discrete event simulations
  - Predictive simulations
  - Inverse problems simulations
  - System integration simulations
  - Complex nonlinear systems

Distributed Data Management:

Information Technology Infrastructure Design and Development

Thrusts:

- **Scientific Data Archives for DOE and NASA (CDIAC, ARM, DAAC)**
- **Scientific Data Archive for USGS (NBII)**
- **Universal Communication and Investigation Consortium (UCIC)**
- **Protective Security Analysis Center (PSAC)**

Mercury Provides a Portal to Distributed Data

Developed for NASA by ORNL

NASA Software Product of the Year – 2001 Runner Up

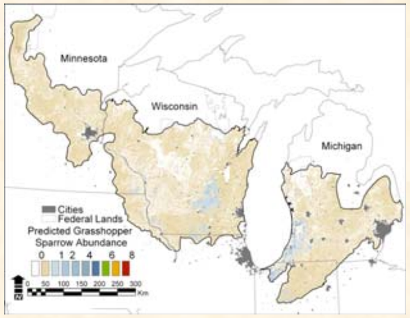
"For the creative development of a technological contribution which has been determined to be of significant value in the advancement of the space and aeronautical activities of NASA, and is entitled: Mercury - A Web Based Metadata Search and Data Retrieval System"

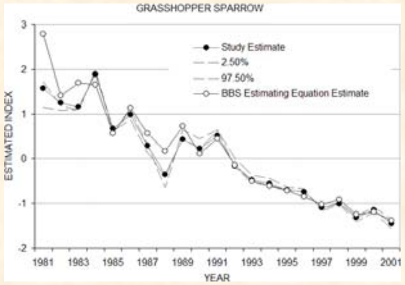
What is the NBII?

- **NBII stands for the National Biological Information Infrastructure**
- **A broad, distributed, collaborative program to provide access to data and information to biological resources.**
- **NBII originates from the United States Geological Survey (USGS), a part of the U. S. Department of the Interior**
- **The NBII Web site is /w**

NBII Goals and Objectives

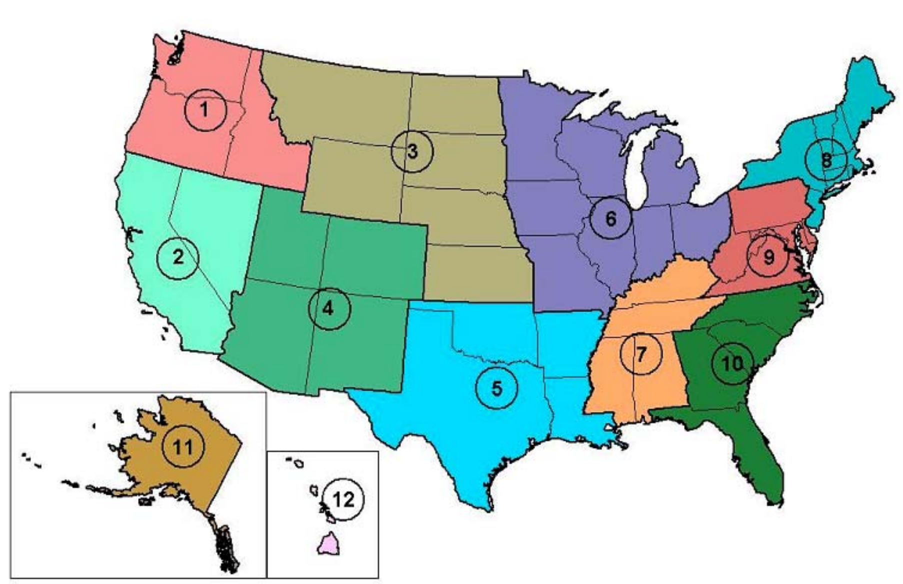
- **The NBII is a broad, collaborative program to** *provide increased access to data and information* **on the nation's biological resources.**
- **The NBII** *links diverse, high-quality biological databases, information products, and analytical tools* **maintained by NBII partners and other contributors in government agencies, academic institutions, non-government organizations, and private industry.**
- **NBII partners and collaborators also work on** *new standards, tools, and technologies* **that make it easier to find, integrate, and apply biological resources information.**
- **Resource managers, scientists, educators, and the general public use the NBII to** *answer a wide range of questions* **related to the management, use, or conservation of this nation's biological resources.**





NBII Regional Nodes

- California
- Central Southwest/Gulf Coast
- Great Basin
- Mid Atlantic
- Mountain Prairie
- Northeast
- Pacific Basin
- Pacific Northwest
- Southern Appalachian
- Southwest



What Does ORNL do for NBII?

- **Provides the NBII Metadata Clearinghouse**
- **UDDI server**
- **Specimen web service**
- **Thesauri web services**
- **Gazetteer web services**
- **Leverages ORNL OGC membership**

- ORNL is an OGC Technical Committee Member

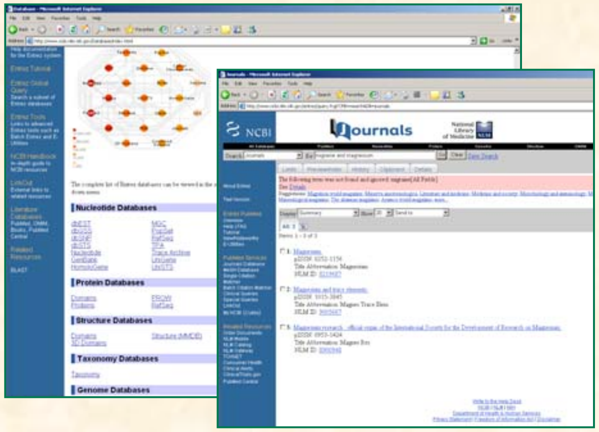
Knowledge Discovery from Text: A Success Story

Biomedical Journals

Low Cross-reference within disciplines

PubMed Archives

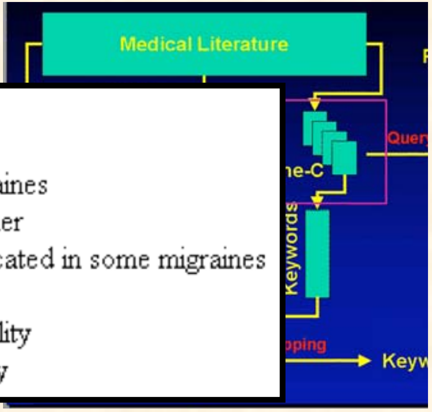
Online Archive of Medical Journals



Question:

What causes migraine headaches?

Text Analysis and Mining



Extracted evidence from titles of articles in the biomedical literature

- stress is associated with migraines
- stress can lead to loss of magnesium
- calcium channel blockers prevent some migraines
- magnesium is a natural calcium channel blocker
- spreading cortical depression (SCD) is implicated in some migraines
- high leveles of magnesium inhibit SCD
- migraine patients have high platelet aggregability
- magnesium can suppress platelet aggregability

Swanson, 1987

Swanson et al., 1991, 1994, 1997

Example from Hearst, 1999

New Hypothesis:

Magnesium Deficiency leads to Migraine

(**New** medical knowledge)

Confirmed by Experts

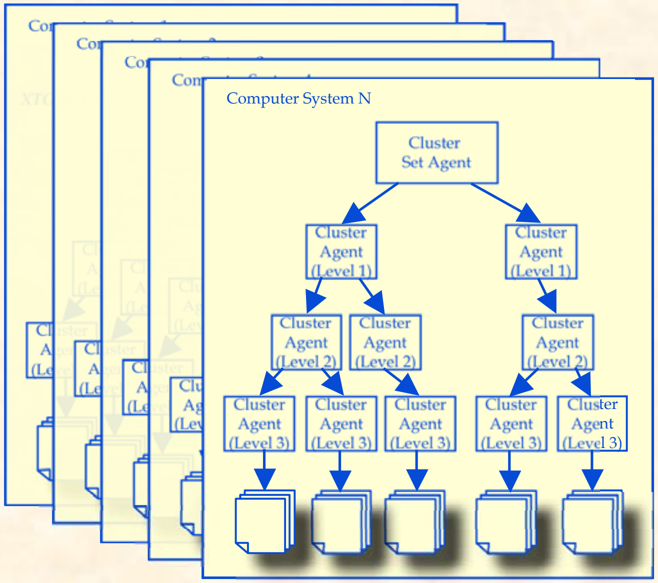
Ramadan et al., 1989

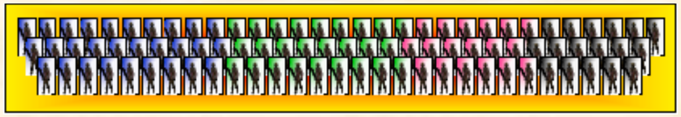
Hypothesis Generation: ***"Chains of causal implication within the medical literature can lead to hypotheses for causes of rare diseases"***

Agent-Based Dynamic Text Analysis

Thrusts:

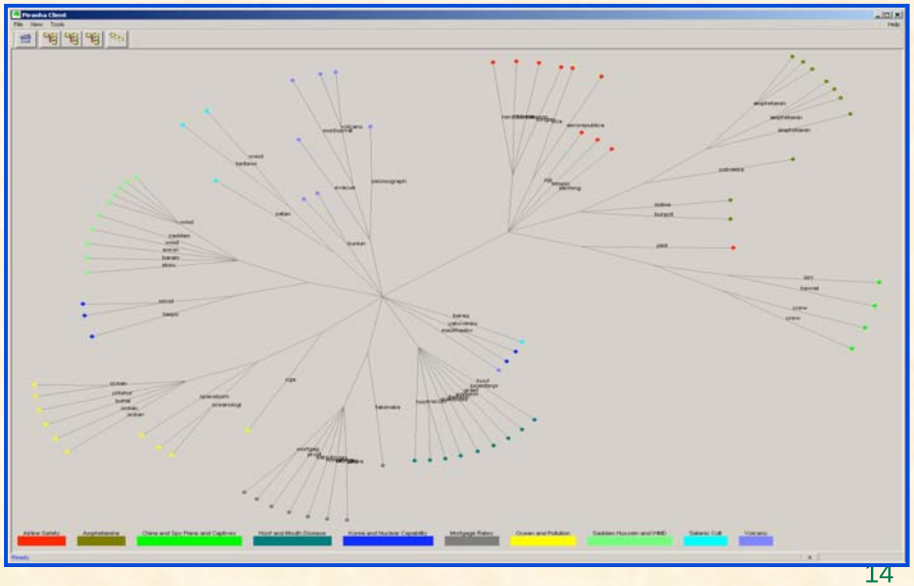
- Piranha – Analyzing high volume, dynamic text data
- VIPAR - DHS Advanced Scientific Computing
- Agent-based Swarming Algorithms





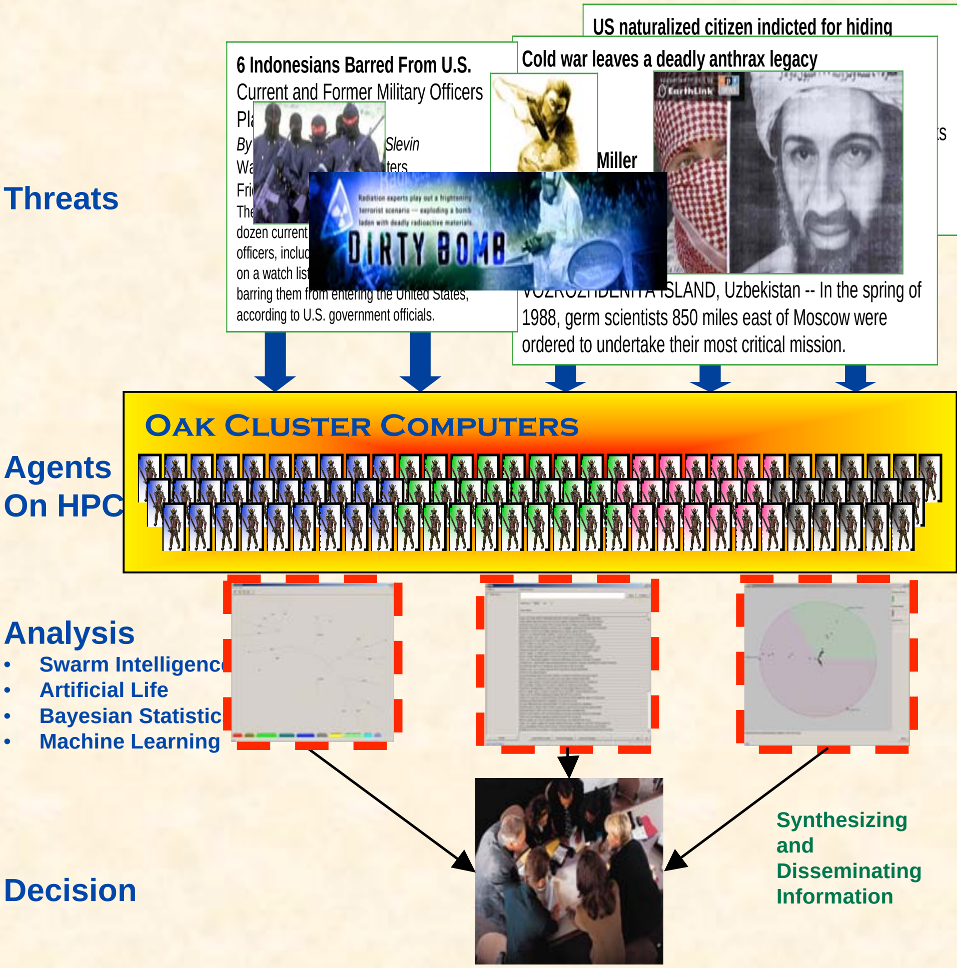
Patents

- J. Reed and T. Potok "An Agent-based Method for Distributed Clustering of Textual Information" submitted (2004)
- T. Potok, J. Reed, M. Elmore, J. Treadwell, N. Samatova, "Method for Gathering and Summarizing Internet Information", application number 20030120639 (2003).



Agent-Based Threat Detection

- **Projects:**
  - **DHS, Military, IC Customers**
- **Recent Publications:**
  - **M. Elmore, J. Reed, T. Potok, "Real-time Document Cluster Analysis for Dynamic Data Sets," IPSI-Amalfi, 2005**
  - **X.Cui, T.Potok, "Tracking non-Stationary Optimal Solution by Particle Swarm Optimizer," ACIS International Conference on Software Engineering, Artificial Intelligence, Networking, and Parallel/Distributed Computing (SNPD2005)**
  - **T. E. Potok, L. Phillips, R. Pollock, and A. Loebl, "Suitability of Agent Technology for Military Command and Control in the Future Combat System Environment," Proceedings of the 8th International Command and Control Research and Technology Symposium, 2003**
- **Resources: Red and White Oak Clusters**
  - **4 Dell 2850s each with**
    - **3.2 GHz Dual Processor**
    - **2 GB Ram**
    - **438 GB Disk**
  - **131 Dell 1850s each with**
    - **3.2 GHz Dual Processor**
    - **2 GB Ram**
    - **73 GB Disk**
  - **Total**
    - **1.7 TFLOPS**
    - **270 GB Memory**
    - **11.3 TB Disk**

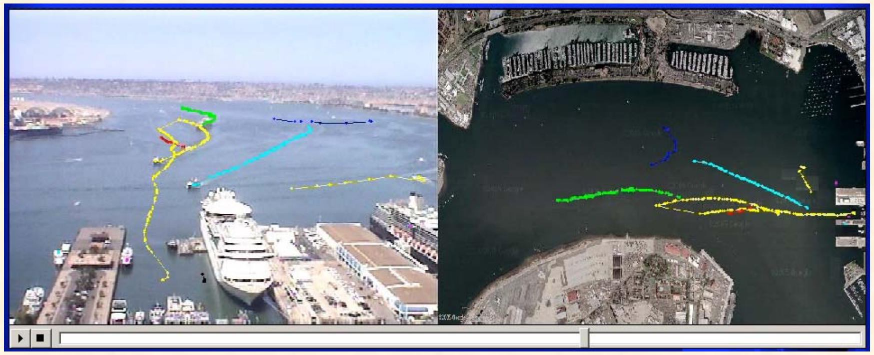

Threats
6 Indonesians Barred From U.S.
Current and Former Military Officers
Pl
By
Slevin
Wa
ters
Fri
The
dozen current
officers, includ
on a watch list
barring them from entering the United States,
according to U.S. government officials.
Radiation experts play out a frightening terrorist scenario -- exploding a bomb laden with deadly radioactive materials.
DIRTY BOMB
US naturalized citizen indicted for hiding
Cold war leaves a deadly anthrax legacy
Miller
VOZROZHDENIYA ISLAND, Uzbekistan -- In the spring of 1988, germ scientists 850 miles east of Moscow were ordered to undertake their most critical mission.
Agents
On HPC
OAK CLUSTER COMPUTERS
Analysis
• Swarm Intelligence
• Artificial Life
• Bayesian Statistic
• Machine Learning
Decision
Synthesizing and Disseminating Information

Video and Image Analysis, Visualization

Thrusts:

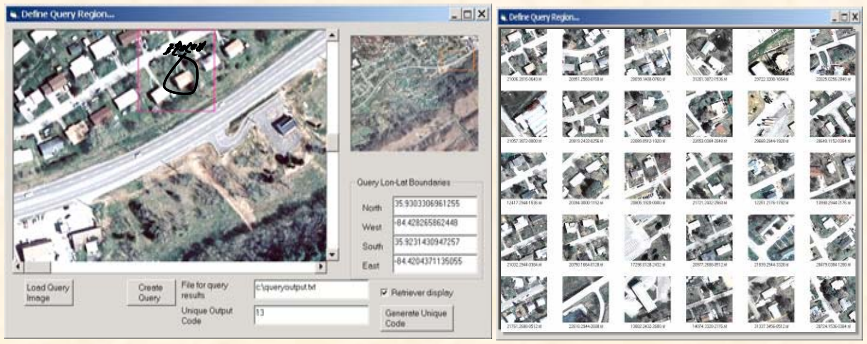
- **Oak Ridge Isochronous Observation Network (ORION)**
  - **Port of San Diego**
  - **Port of Charleston**
- **Image to Intelligence Archive (I2IA)**

ORION



Software Agents Analyze Massive Video, Image Data

Image to Intelligence Archive (I2IA)



Query Lon-Lat Boundaries

North	35.9303306961255
West	-84.428265862448
South	35.9231430947257
East	-84.4204371135055

File for query results: c:\queryoutput.txt

Unique Output Code: 13

☑ Retriever display

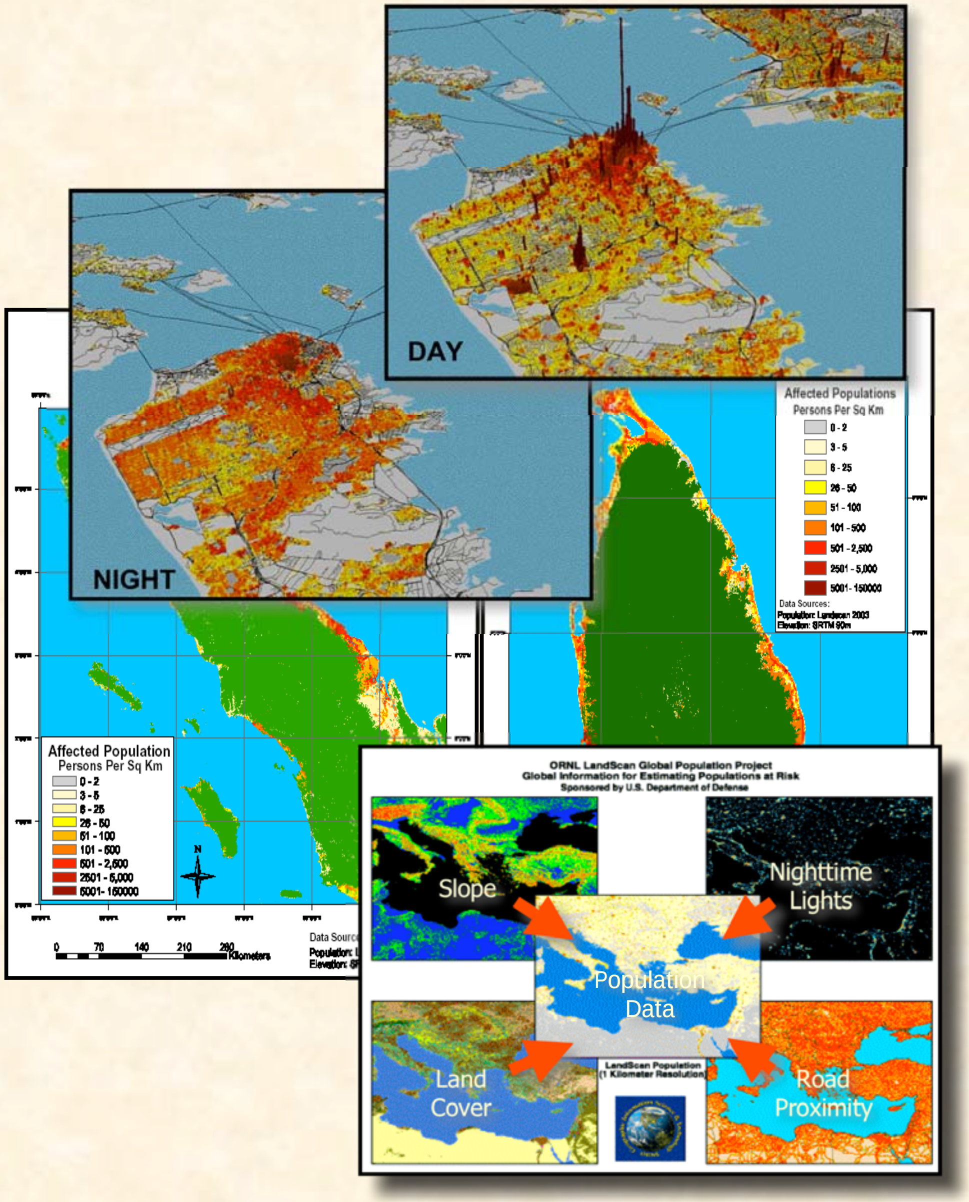
Location
35.930N
-84.428W
35.923S
-84.420E
10 July 03
14:23:04

OAK RIDGE NATIONAL LABORATORY
U. S. DEPARTMENT OF ENERGY

Parallel Discrete Event Simulations: Event-based and Formal Methods for Large Systems

Thrusts:

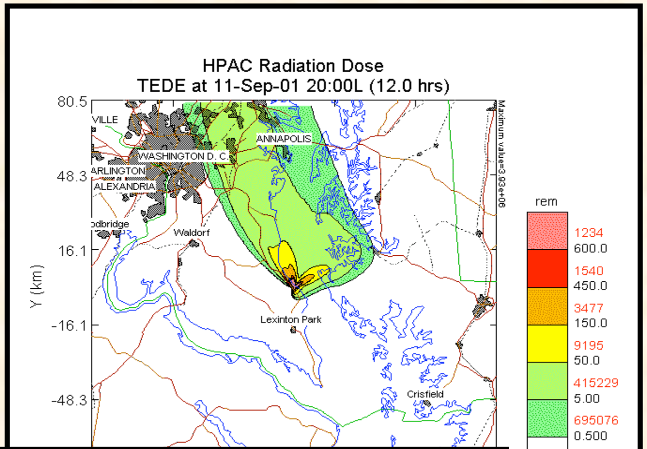
- **LandScan Global and USA Population Modeling**
- **Spatial-Temporal Social Dynamics**
- **Intelligent Consequence Management**
- **Critical Infrastructure Protection Modeling**
- **Multi-modal Transportation Modeling**

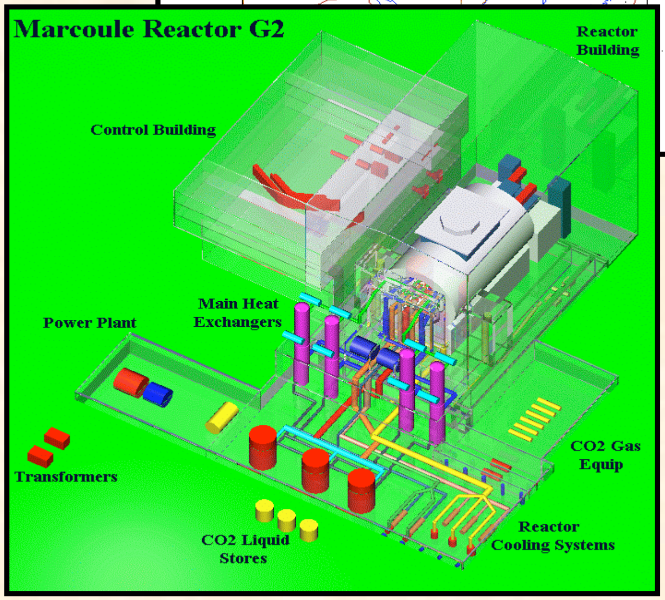


Predictive Simulations: Multi-disciplined Physics-Based Computations

Thrusts:

- **Prediction of Atmospheric transport of hazardous materials**
- **River and estuary water transport**
- **Facility Vulnerability**

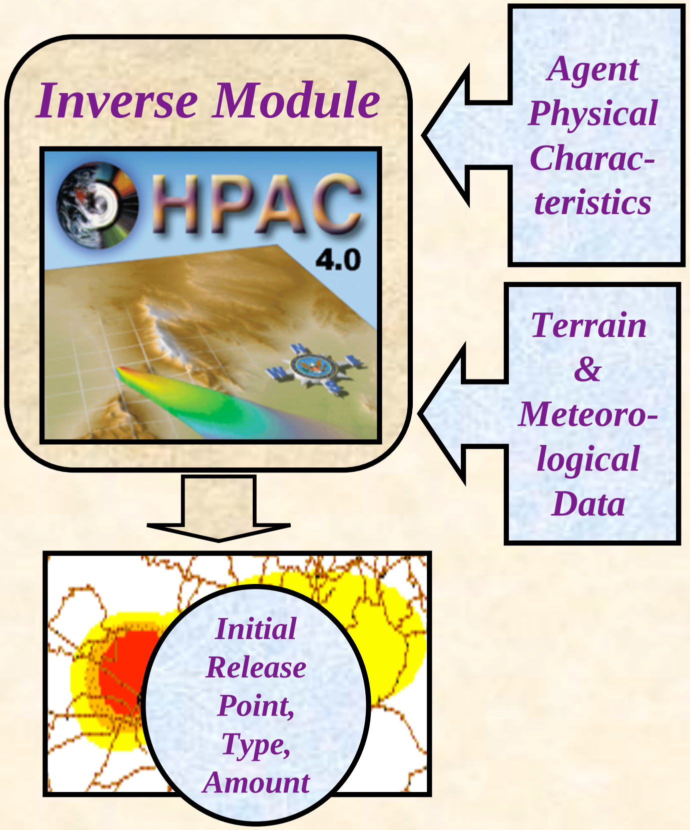




Inverse Problems Simulations: Determining Potential Causes of Known Effects

Thrusts:

- **Prediction of source of observed waterborne hazards**
- **Prediction of source of observed atmospheric hazards**



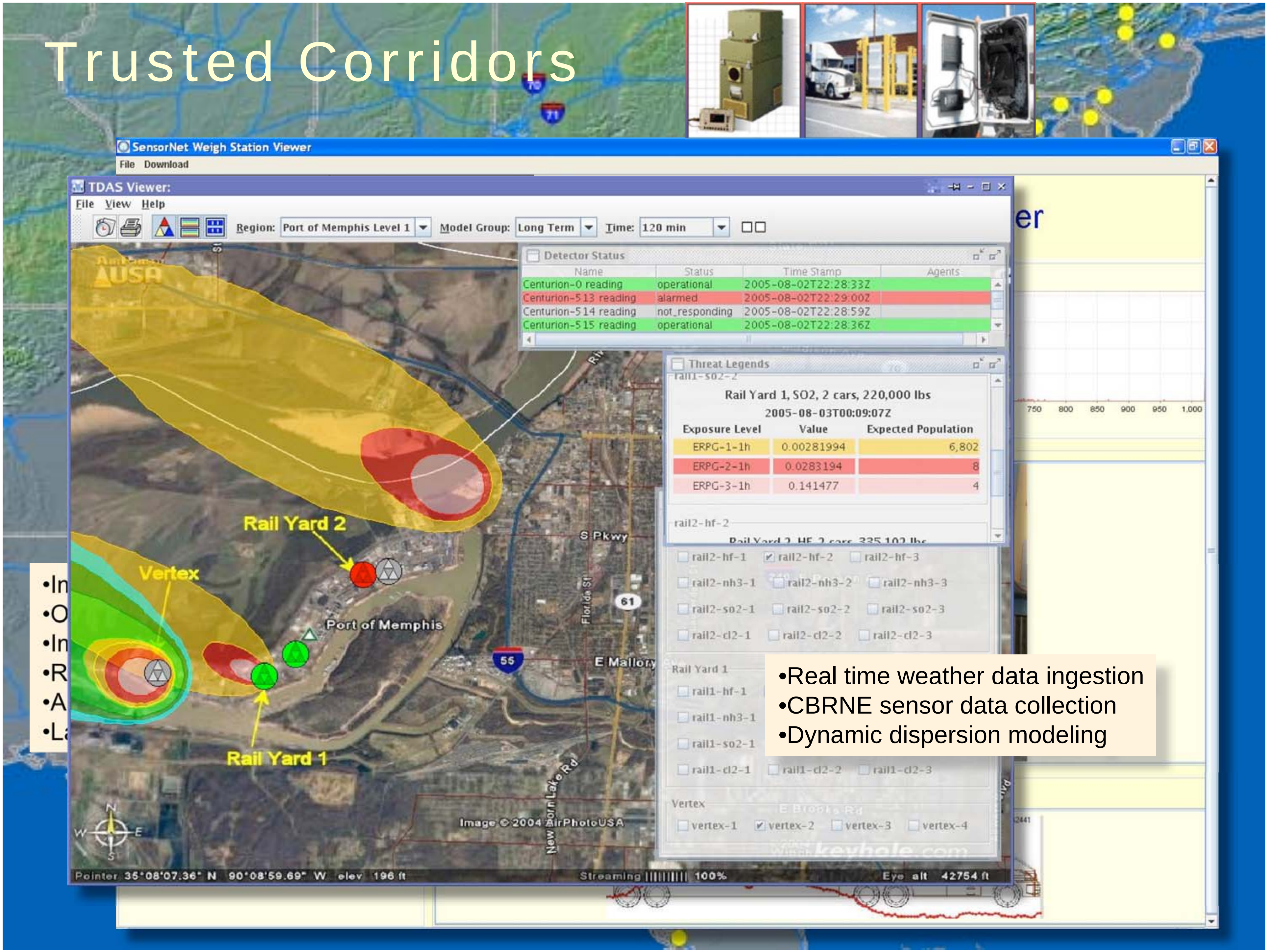
Inverse Module

Agent Physical Characteristics

Terrain & Meteorological Data

Initial Release Point, Type, Amount

Trusted Corridors



Region: Port of Memphis Level 1

Model Group: Long Term

Time: 120 min

Detector Status

Name	Status	Time Stamp	Agents
Centurion-0 reading	operational	2005-08-02T22:28:33Z	
Centurion-513 reading	alarmed	2005-08-02T22:29:00Z	
Centurion-514 reading	not_responding	2005-08-02T22:28:59Z	
Centurion-515 reading	operational	2005-08-02T22:28:36Z	

Threat Legends

rail1-so2-2

Rail Yard 1, SO2, 2 cars, 220,000 lbs

2005-08-03T00:09:07Z

Exposure Level	Value	Expected Population
ERPG-1-1h	0.00281994	6,802
ERPG-2-1h	0.0283194	8
ERPG-3-1h	0.141477	4

rail2-hf-2

Rail Yard 2, HF, 2 cars, 335,102 lbs

☐ rail2-hf-1 ☑ rail2-hf-2 ☐ rail2-hf-3

☐ rail2-nh3-1 ☐ rail2-nh3-2 ☐ rail2-nh3-3

☐ rail2-so2-1 ☐ rail2-so2-2 ☐ rail2-so2-3

☐ rail2-cl2-1 ☐ rail2-cl2-2 ☐ rail2-cl2-3

Rail Yard 1

☐ rail1-hf-1

☐ rail1-nh3-1

☐ rail1-so2-1

☐ rail1-cl2-1 ☐ rail1-cl2-2 ☐ rail1-cl2-3

Vertex

☐ vertex-1 ☑ vertex-2 ☐ vertex-3 ☐ vertex-4

- In
- O
- In
- R
- A
- L

- Real time weather data ingestion
- CBRNE sensor data collection
- Dynamic dispersion modeling