

Current and Future Intelligent Agent Initiatives



Dr. Thomas E. Potok
Applied Software Engineering
Research Group
Oak Ridge National Laboratory

National Challenge

- Data everywhere
- Sources unreliable
- Difficult to merge
- Cannot be done manually

Sensors

Multimedia

Image

Text

Binary

Data

1970

1980

1990

2000

2010

11010010

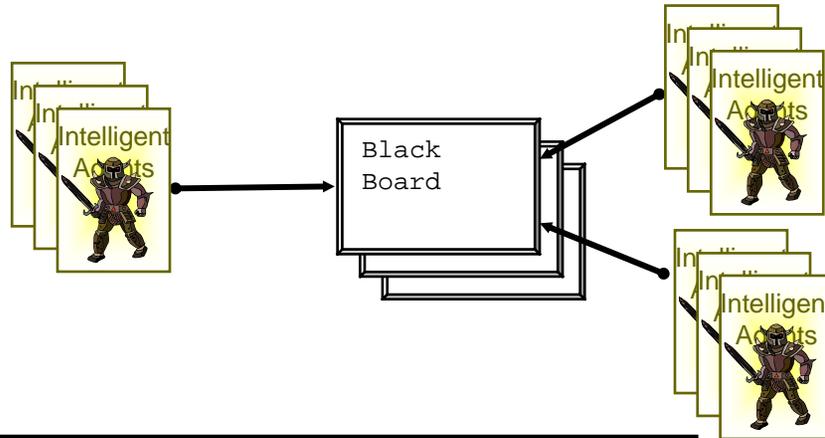
One small
step for
man



National Scope

- How can we run the battle from networked vehicles? US Army - US Army
 - We led a DARPA study on the suitability of intelligent agent technology
- Can software tell us in 3 to 5 minutes if we should destroy an incoming missile? - Missile Defense Agency
 - We serve on an expert panel to assess battle management software capabilities
- What threats exist in the in 10,000 new documents we received today? - Department of Homeland Defense
 - We led a DHS Text Analysis Workshop to set research directions and goals for the department

Key Technologies and Resources



- Intelligent Agents
 - Software processes
 - Can communicate in unstable environments
 - Form teams to solve problems
 - Live, die, and reproduce to solve problems
- High Performance Computing
 - Red/White Oak Clusters
 - 135 Dell Computers
 - Largest cluster computer at ORNL (1.7 TFLOPS, 270 GB Memory, 11.3 TB Disk)

Red Oak / 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**
 - 270 3.2 GHz Processors
 - Effectively 540 Processors (Hyper-Threaded)
 - 270 GB Memory
 - 11.3 TB Disk

The image shows four server racks from the Red Oak / White Oak Clusters. The racks are black and have blue lights on the front panels. The text on the left provides specifications for the Dell 2850s and Dell 1850s servers, and a total summary.

Agenda

- Large Scale Data Mining
 - Approach/Background/Issues
 - TF/ICF
 - Piranha
 - Next Steps
- Intelligent Agent Technology
 - Agent background
 - Swarm Intelligence
 - Evolutionary Agents
 - Gryffin

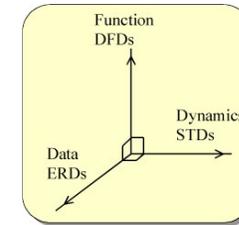
Challenge – What to do with this?



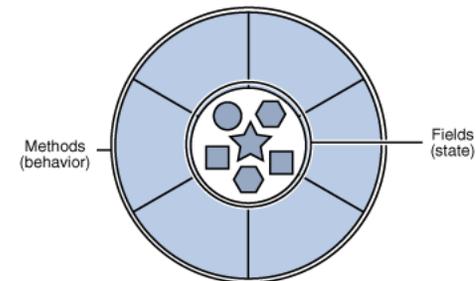
- ❑ What is in there?
- ❑ Are there any threats?
- ❑ What am I missing?

Why Agents

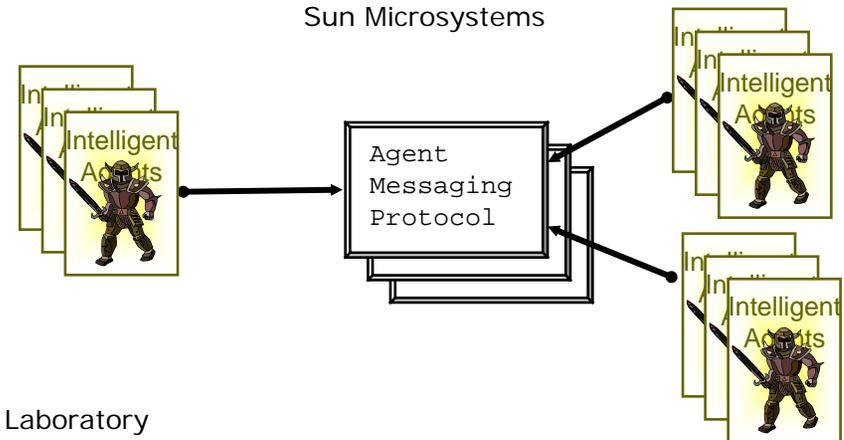
- ❑ Structured Analysis
 - Design data and function separately
- ❑ Object Oriented
 - Design data and function together then communication
 - ❑ Encapsulation
 - ❑ Polymorphism
 - ❑ Inheritance
- ❑ Agents
 - Design data, function, and communication together



CRaG Systems



Sun Microsystems



Agents from a Software Engineering Perspective

- Communication and control aspects of agent systems
 - Peer-to-peer topology
 - Agent coordination models
 - Encapsulated and asynchronous messaging
 - Use of blackboards, and tuple space models and associated pattern-matching
 - High-level messages
 - Agent control language (ACL) such as KQML or the FIPA ACL.
 - These languages provide a structured means of exchanging information and knowledge among agents.

Large Scale Data Mining

□ Problem

- How to effectively reduce the size of a large, streaming set of documents
- “Give me the 10 documents that I need to read, out of the 1000 I received today?”

□ Characteristics

- A steady flow of simple documents
- Need to rapidly organize the documents into subsets
- Select representative documents from the subsets

Approach

- Use IR techniques to convert text to vectors
- Use unsupervised learning/text clustering to organize the documents
- Use adaptive sampling and cluster centroid methods for selecting representative documents
 - R. M. Patton and T. E. Potok, "Adaptive Sampling of Text Documents," *13th International Conference on Intelligent and Adaptive Systems and Software Engineering*, 2004.
 - J. W. Reed, T. E. Potok, and R. M. Patton, "A Multi-Agent System for Distributed Cluster Analysis," *Third International Workshop on Software Engineering for Large-Scale Multi-Agent Systems*, 2004

Standard Information Retrieval

Document 1

The Army needs sensor technology to help find improvised explosive devices

Terms

Army
Sensor
Technology
Help
Find
Improvise
Explosive
device

Document 2

ORNL has developed sensor technology for homeland defense

ORNL
develop
sensor
technology
homeland
defense

Document 3

Mitre has won a contract to develop homeland defense sensors for explosive devices

Mitre
won
contract
develop
homeland
defense
sensor
explosive
devices

Term List

Army
Sensor
Technology
Help
Find
Improvise
Explosive
Device
ORNL
develop
homeland
Defense
Mitre
won
contract

Vector Space Model

	Doc 1	Doc 2	Doc 3
Army	1	0	0
Sensor	1	1	1
Technology	1	1	0
Help	1	0	0
Find	1	0	0
Improvise	1	0	0
Explosive	1	0	1
Device	1	0	1
ORNL	0	1	0
develop	0	1	1
homeland	0	1	1
Defense	0	1	1
Mitre	0	0	1
won	0	0	1
contract	0	0	1

Standard Textual Clustering

Vector Space Model

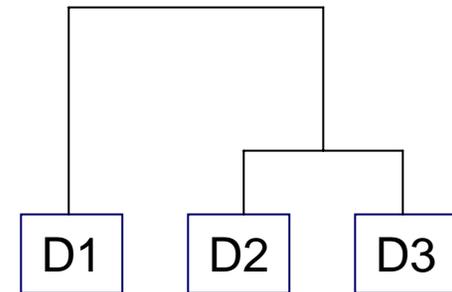
	Doc 1	Doc 2	Doc 3
Army	1	0	0
Sensor	1	1	1
Technology	1	1	0
Help	1	0	0
Find	1	0	0
Improvise	1	0	0
Explosive	1	0	1
Device	1	0	1
ORNL	0	1	0
develop	0	1	1
homeland	0	1	1
Defense	0	1	1
Mitre	0	0	1
won	0	0	1
contract	0	0	1

Similarity Matrix

	Doc 1	Doc 2	Doc 3
Doc 1	100%	17%	21%
Doc 2		100%	36%
Doc 3			100%

Documents to Documents

Cluster Analysis



Most similar documents

TFIDF

$$W_{ij} = \log_2(f_{ij} + 1) * \log_2\left(\frac{N}{n}\right)$$

Euclidean distance

$$d_2(\mathbf{x}_i, \mathbf{x}_j) = \left(\sum_{k=1}^d (x_{i,k} - x_{j,k})^2\right)^{1/2}$$

Time Complexity

$$O(n^2 \text{Log } n)$$

Issues

- Term weights: Every added or removed document from the set requires recalculation of the entire VSM

$$W_{ij} = \log_2(f_{ij} + 1) * \log_2\left(\frac{N}{n}\right)$$

Document Set must be known before VSM can be calculated

- Agent Messaging
 - Agent systems have high message traffic which can limit scalability and performance
 - Vectors generated from text are very large

The TF-ICF Term Weighting

- Can we use the global frequency of a known corpus to approximate that of an unknown data stream?

Reference Corpora	Document Count
TREC-AP	239,055
TREC-FBIS	130,392
TREC-LATIMES	127,732
TREC-SJM	73,748
TREC-WSJ	161,576
RSS News Feeds	202,991
Total	935,494



Statistics:

- # of Unique Terms: 229,023
- # of Common Terms: 28,872
- Average Standard Deviation of the Global Frequency (n/N) of the Six Corpora: $< 0.01\%$



$$W_{ij} = \log_2(f_{ij} + 1) * \log_2\left(\frac{C + 1}{c + 1}\right)$$

Inverse Corpus Frequency¹⁴

TF-ICF Performance Evaluation

Term Weighting Schemes

Name	Term Weighting Scheme
TF-IDF	$w_{ij} = \log(f_{ij}) \times \log(N / n_j)$
MI	$w_{ij} = \log \frac{\frac{f_{ij}}{N}}{\sum_{i=1}^N f_{ij} \times \frac{M}{\sum_{j=1}^M f_{ij}}}$
ATC	$w_{ij} = \frac{\left(0.5 + 0.5 \times \frac{f_{ij}}{\max_f}\right) \log\left(\frac{N}{n_j}\right)}{\sqrt{\sum_{i=1}^N \left[\left(0.5 + 0.5 \times \frac{f_{ij}}{\max_f}\right) \log\left(\frac{N}{n_j}\right)\right]^2}}$
Okapi	$w_{ij} = \left(\frac{f_{ij}}{0.5 + 1.5 \times \frac{dl}{\text{avg_dl}} + f_{ij}} \right) \log\left(\frac{N - n_j + 0.5}{f_{ij} + 0.5}\right)$
LTU	$w_{ij} = \frac{(\log(f_{ij}) + 1.0) \log\left(\frac{N}{n_j}\right)}{0.8 + 0.2 \times \frac{dl}{\text{avg_dl}}}$

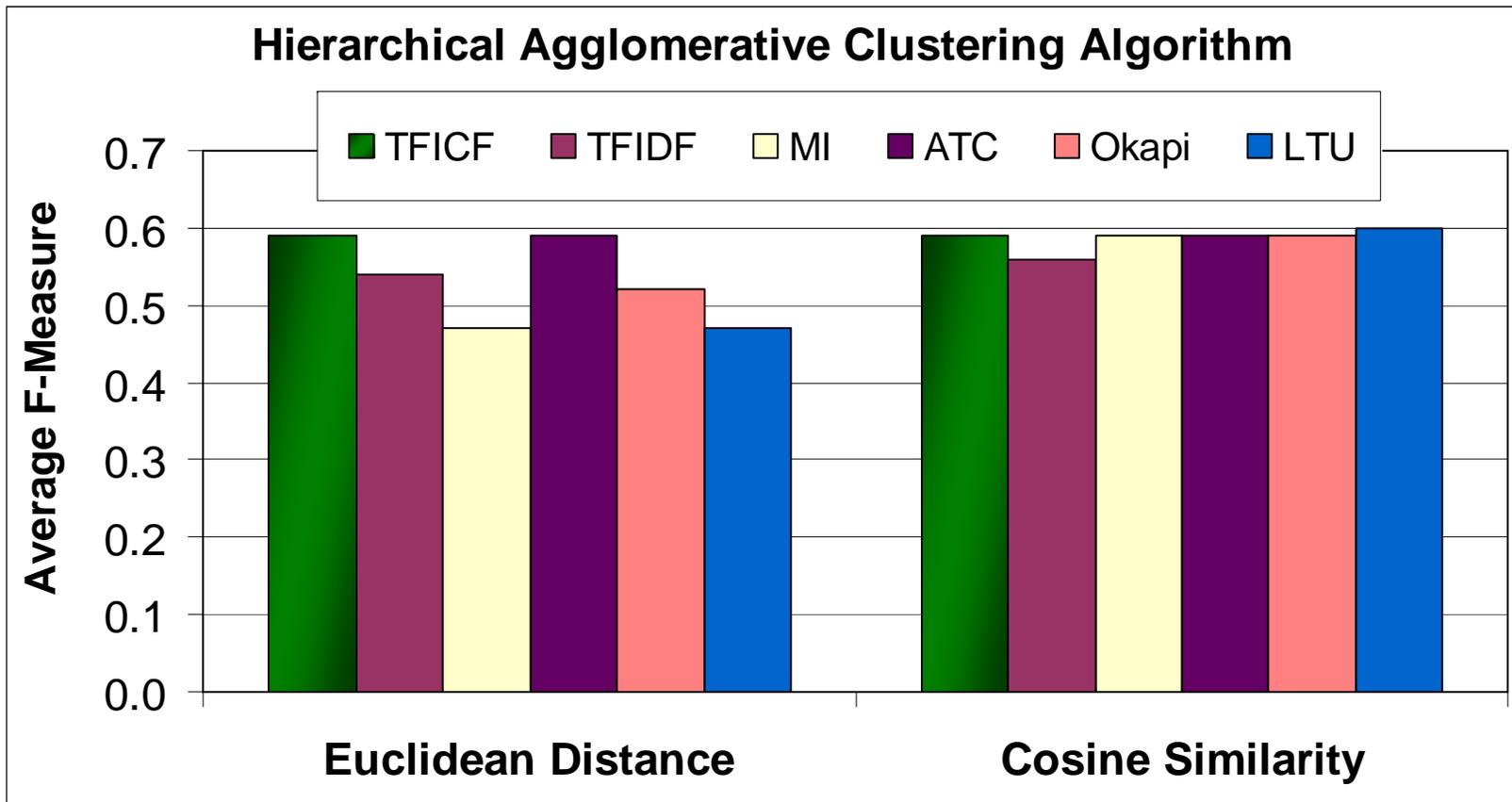
Test Data Sets

Data Set	# of Docs	# of Classes	Largest Class	Smallest Class
Reuters	2349	58	1041	1
SMART	3891	3	1460	1033
20 News	4650	12	399	385

Test Algorithms

Basic K-Means
Single Link (HAC)
Complete Link (HAC)
UPGMA (HAC)

TF-ICF Performance Evaluation

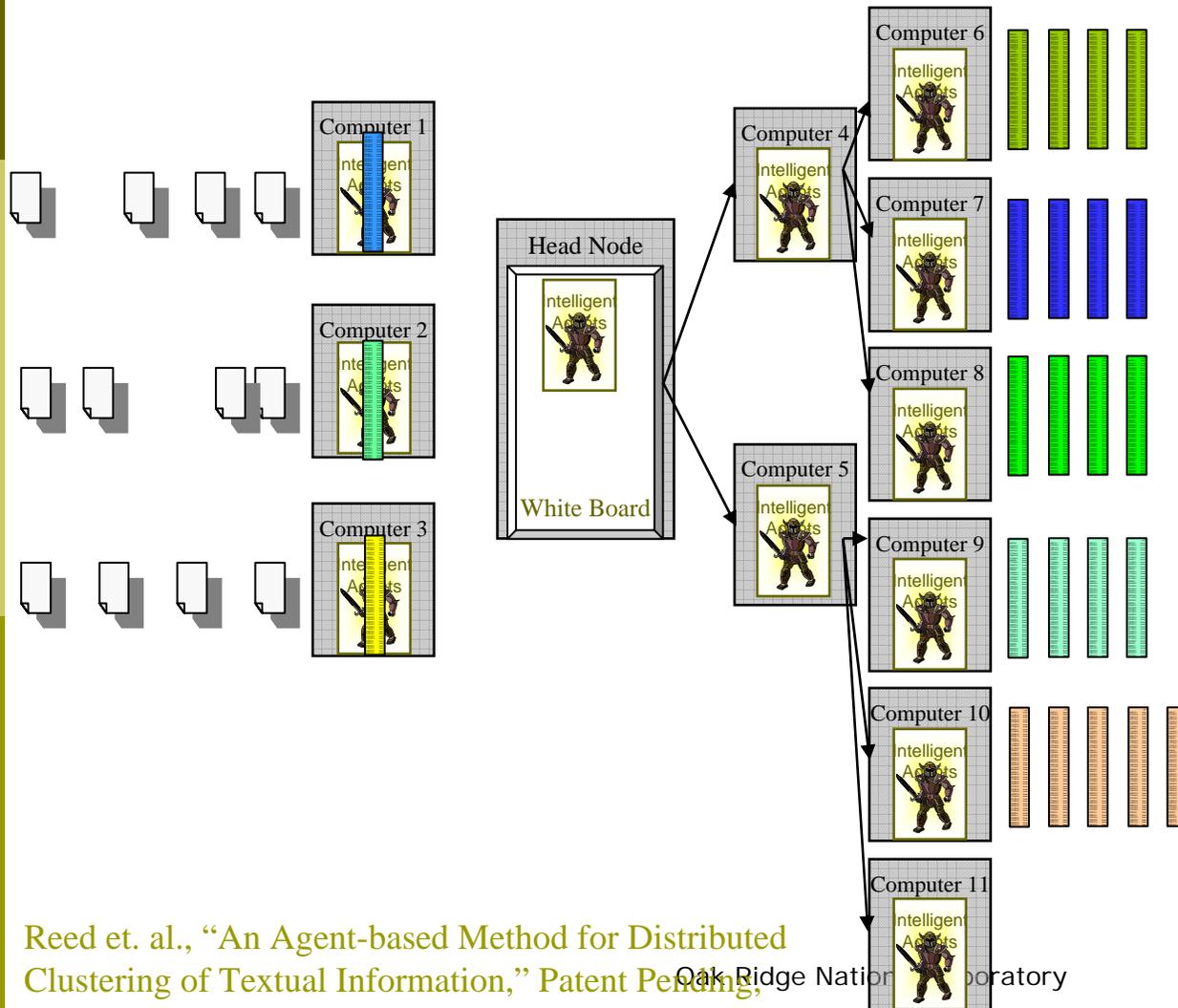


Reed, Jiao, et al., "TF-ICF: A New Term Weighting Scheme for Clustering Dynamic Data Streams," International Conference on Machine Learning and Applications, Orlando, FL, Dec. 14-16, 2006.

Why this matters

- ❑ We can now generate an accurate vector directly from a text document
- ❑ That document vector can be generated where the document resides
- ❑ We can now use agents to create vectors from documents over a broad range of computers

How Piranha Works



- Standard Approach
 - 11.5 Days
- Agent approach
 - 8 minutes 24 Seconds!
 - 2000 times faster
 - With no loss of accuracy

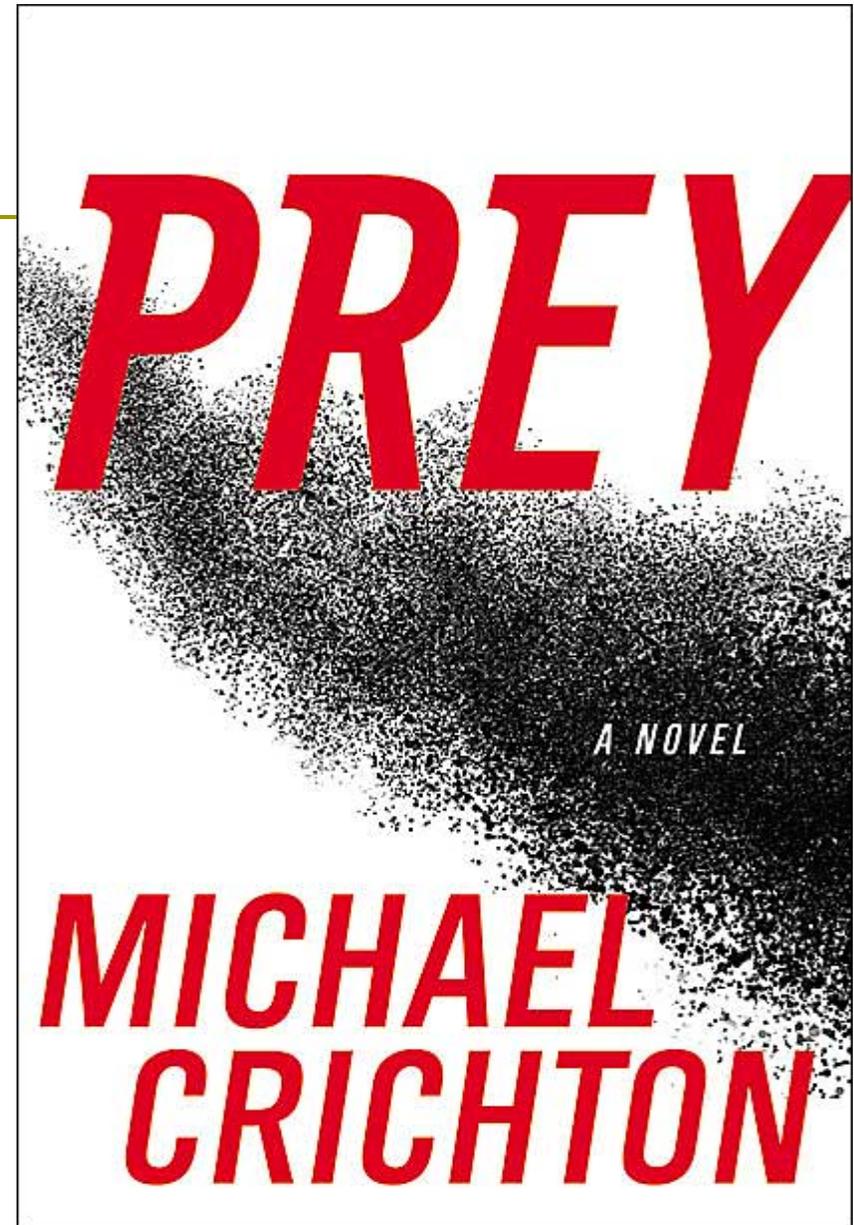
Reed et. al., "An Agent-based Method for Distributed Clustering of Textual Information," Patent Pending, Oak Ridge National Laboratory, licensed to industry

Messaging Approach

- Reduce the dimensionality of the vectors
- Explore software architectures that Limit messaging

Agents:

- Issue:
 - How to go beyond messaging as a means of problem solving
- Approach
 - Emergent behavior
 - Extend current agent framework to include:
 - Ability for agents to swarm
 - Ability for agents to reproduce



Swarm Intelligence (More to follow)

- Swarm intelligence is an emerging field of biologically-

Simple parts, properly connected into a swarm, can yield smart results.

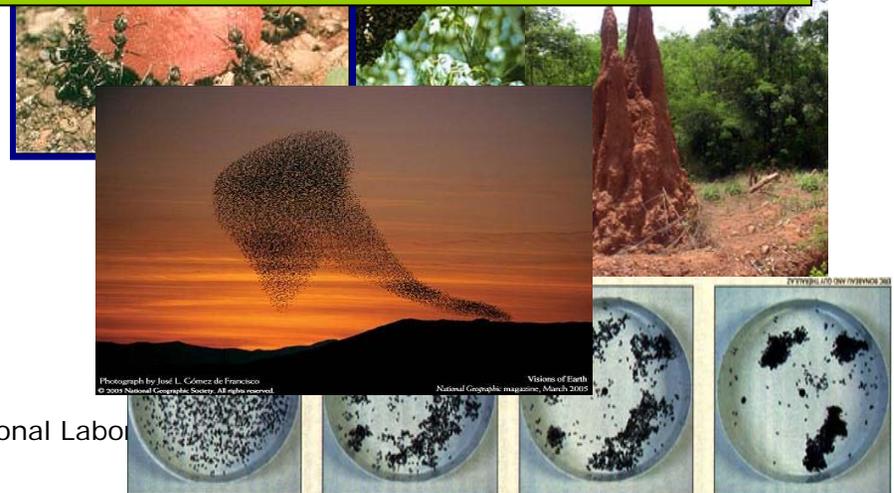
Bird Flocking Model

Ant Clustering Model

Ant Colony Optimization model

Particle Swarm Optimization model

Oak Ridge National Labor



Add DNA and Reproduction to Agents

Endurance Speed Agility Strength



Group



Method



Target



Location



Speed/Strength



Group/Method



Target/Location

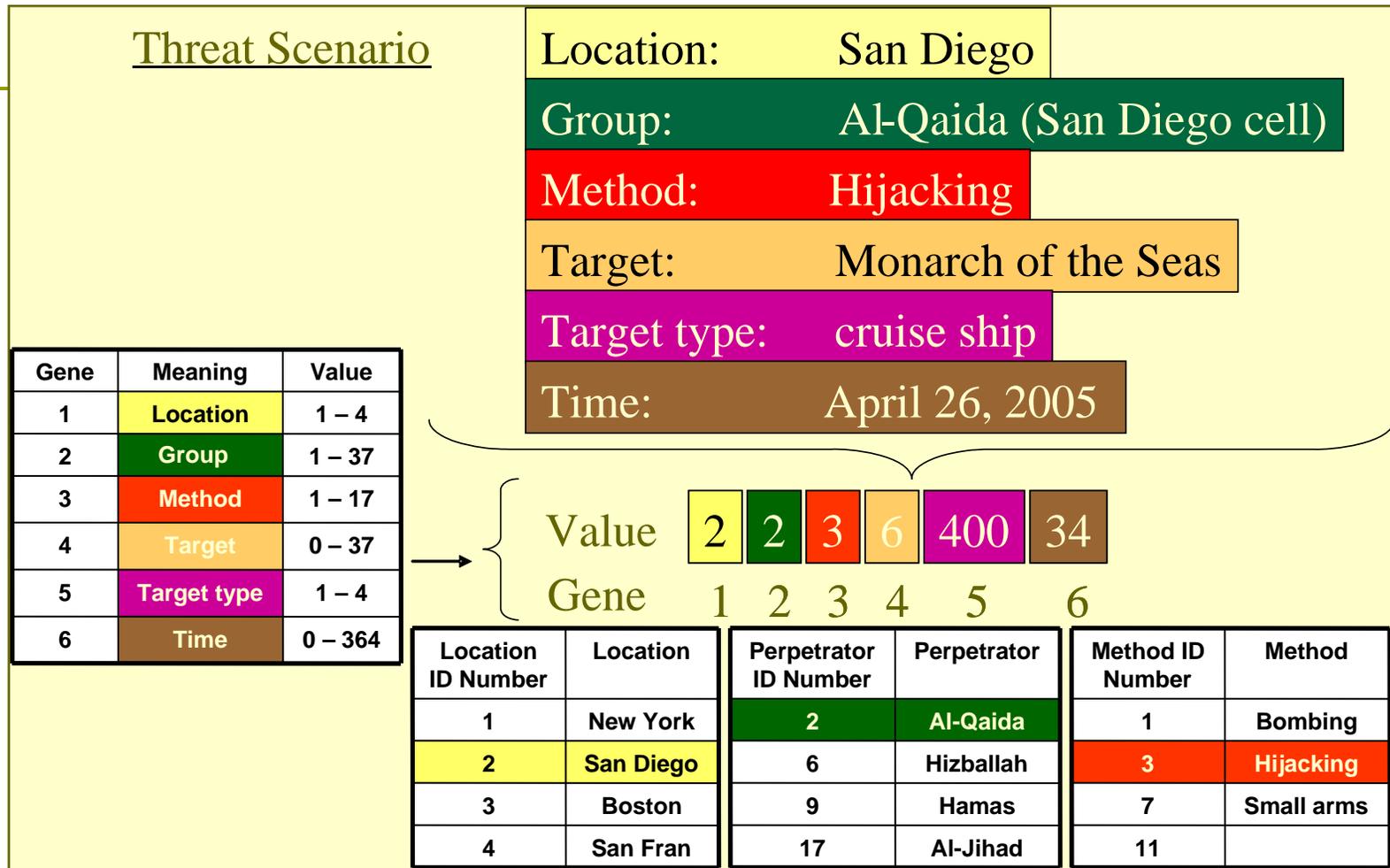


Group/Method/Target/Location



Oak Ridge Nation

Genetic Algorithm DNA Encoding



$$\text{Fitness}(i) = \frac{(\text{Likelihood}(i) \times \text{Intensity}(i))^y}{\text{Niche Size}(p, i)}$$

How Piranha is Used: DHS

Threats

6 Indonesians Barred From U.S.
Current and Former Military Officers



Slevin
Waters

Cold war leaves a deadly anthrax legacy



Miller

US naturalized citizen indicted for hiding



S



Radiation experts play out a frightening terrorist scenario — exploding a bomb laden with deadly radioactive materials. The... dozen current officers, including... on a watch list... barring them from entering the United States, according to U.S. government officials.

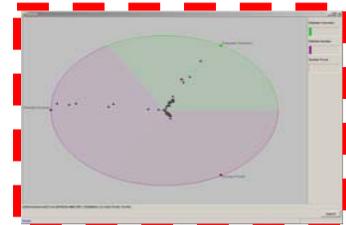
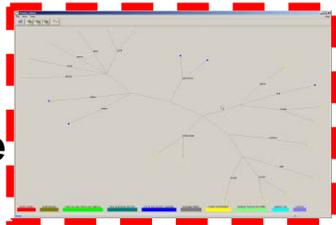
VOZROZHENIYA ISLAND, Uzbekistan -- In the spring of 1988, germ scientists 850 miles east of Moscow were ordered to undertake their most critical mission.

Agents On HPC



Analysis

- Swarm Intelligence
- Artificial Life
- Bayesian Statistics
- Machine Learning



Decision



Synthesizing and Disseminating Information

Oak Ridge Nation

Summary

- ❑ Current technology cannot solve emerging national challenges
- ❑ Intelligent software agents are a significant breakthrough technology
- ❑ Results indicate high-potential to help solve these national challenges
- ❑ We have a progression of significantly successfully deployed agent systems and research to our credit

Contact Information

□ Contact Information

Thomas E. Potok, Ph.D.
Potokte@ornl.gov
865-574-0834