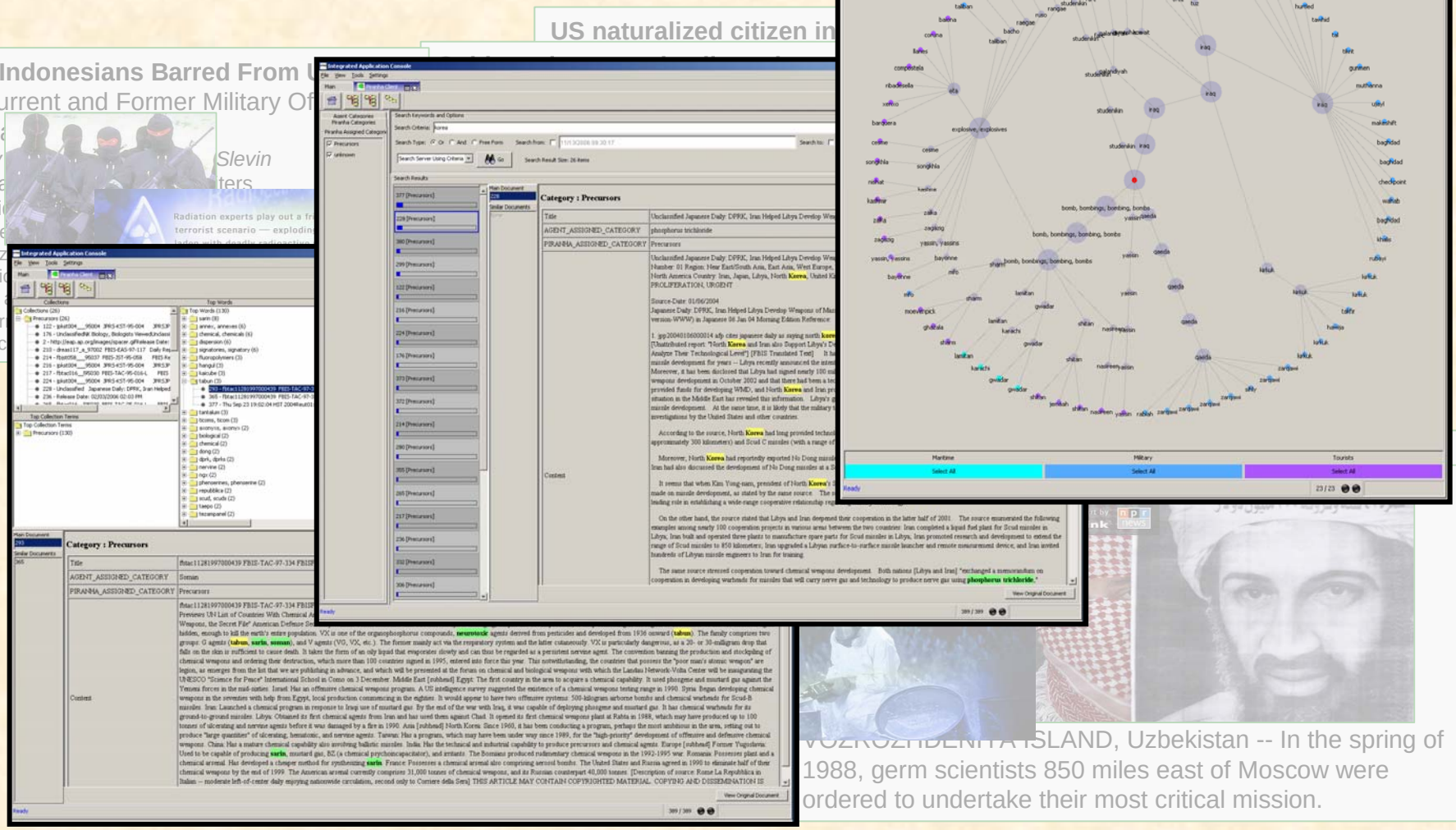
Piranha

Knowledge Discovery Engine



OAK RIDGE NATIONAL LABORATORY
U. S. DEPARTMENT OF ENERGY

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

UT-BATTELLE

Terrorist Bombing

How could you not have seen this coming?

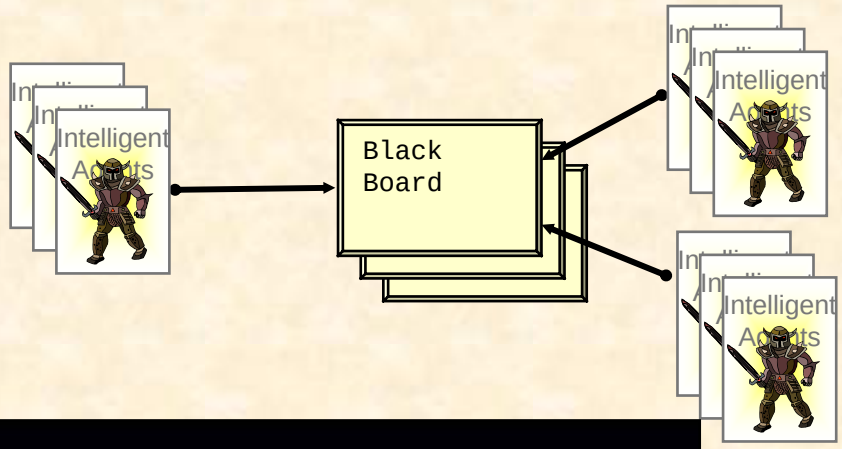
We've traded a stack of paper for a stack of DVDs

Can we trust the Windows OS?
Should we pull out of Iraq?
What should we do about Iran?
What if China destroys key satellites?

National scope of our work

- **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

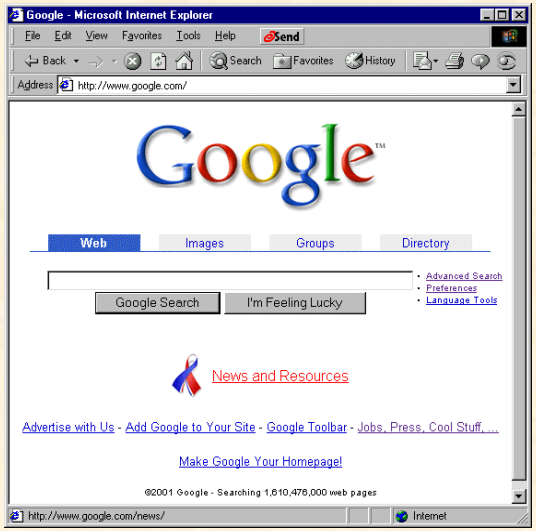


- **Intelligent Agents**
  - **Software processes**
  - **Can communicate in unstable environments**
  - **Form teams to solve problems**
  - **Live, die, and reproduce to solve problems**
- **High Performance Computing**
  - **Red/White Oak Clusters**
  - **135 Dell Computers**
  - **Largest cluster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

Challenge – What to do with this?

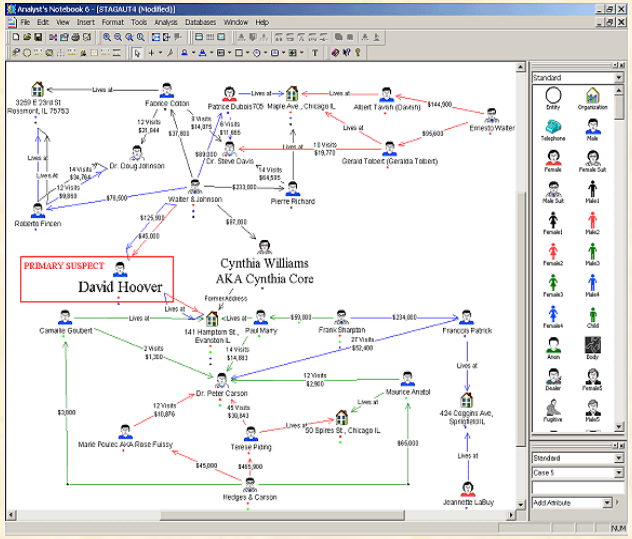
- **What is in there?**
- **Are there any threats?**
- **What am I missing?**

Process



Minutes

Months



Minutes

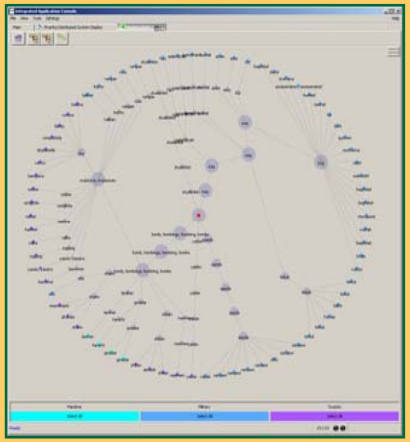
Process

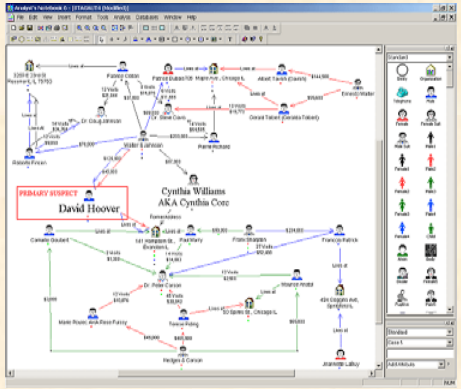
Piranha

Organize the Information

Connect the Information

- Nuclear materials
- Chemical Weapons
- Threats
- Potential Targets
- Money Laundering



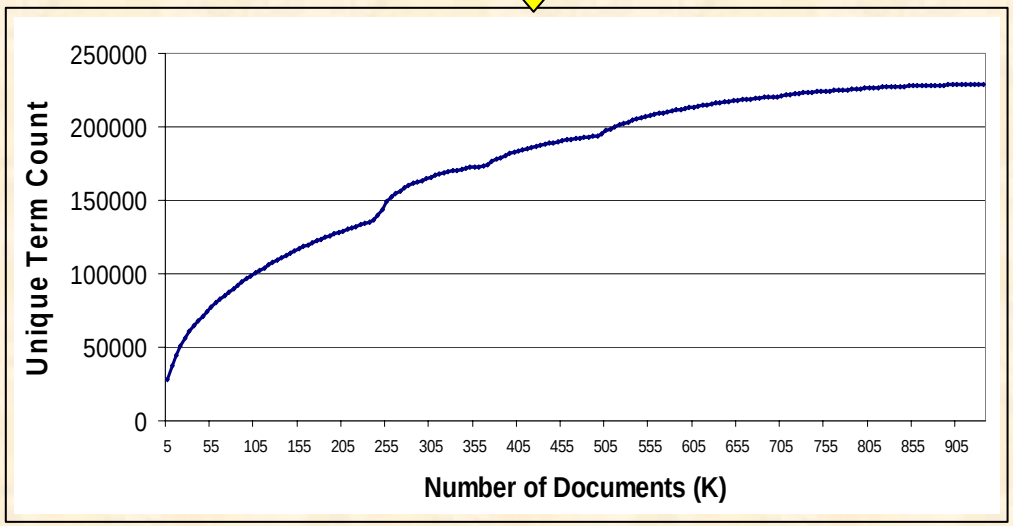


Minutes *Minutes* ~~*Months*~~ *Days* *Minutes*

Breakthrough: Inverse Corpus Frequency

- Look at the forest, not the trees
  - We analyzed near 1 million documents from 6 major research corpora
  - We found 229,023 unique terms (A large dictionary contains around 70,000 terms)
  - We use this term frequency distribution as our "global" term frequency

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



Reed, Jiao, et al., "TF-ICF: A New Term Weighting Scheme for Clustering Dynamic Data Streams," *The Fifth International Conference on Machine Learning and Applications (2006) to appear*

Reed et al., "Multi-Agent System for Distributed Cluster Analysis," *Third International Workshop on Software Engineering for Large-Scale Multi-Agent Systems (SELMAS'04)*, May 24-25, 2004, Edinburgh, Scotland

TF-ICF Performance Evaluation

Term Weighting Schemes

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

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

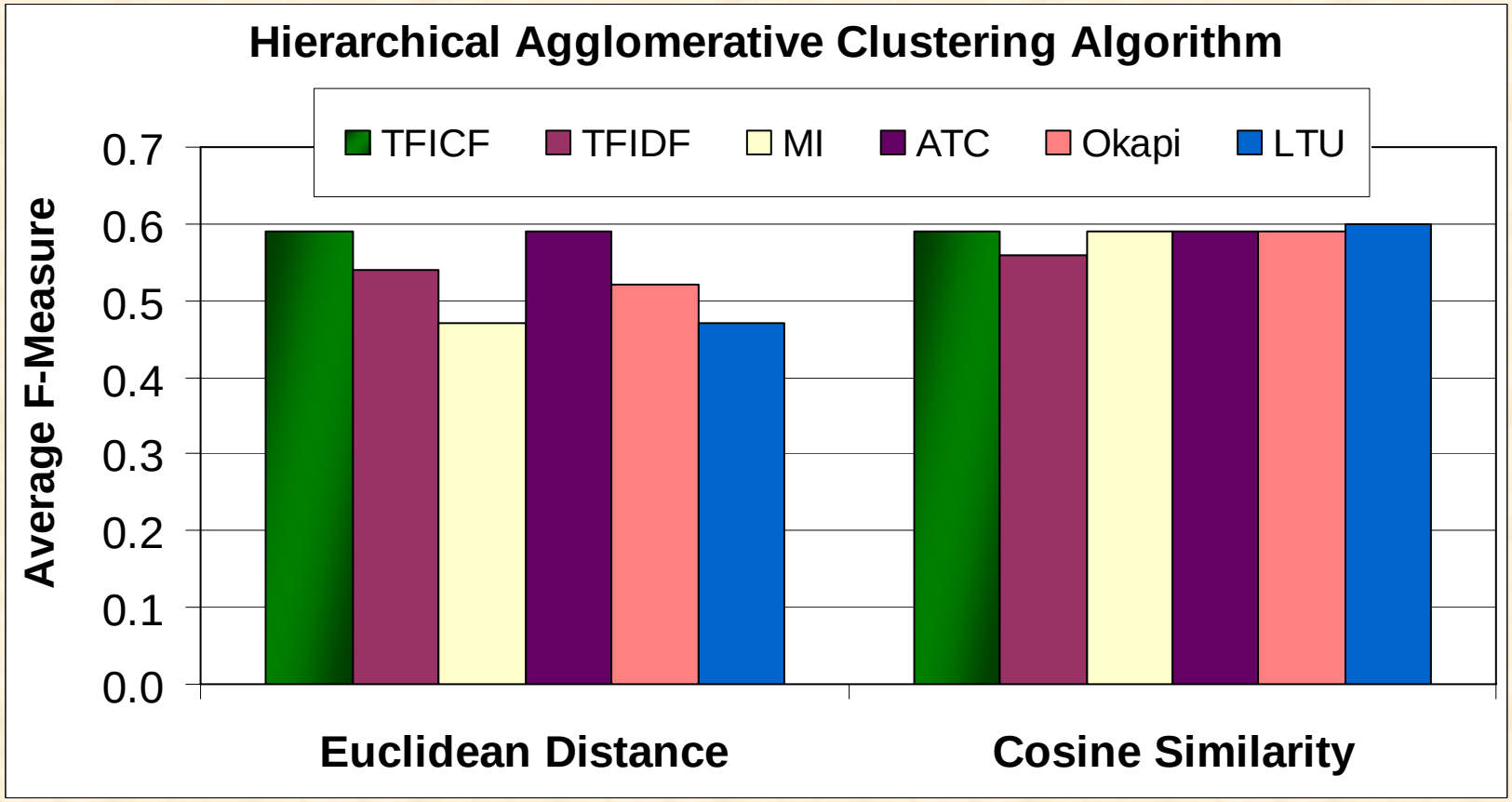
Basic K-Means

Single Link (HAC)

Complete Link (HAC)

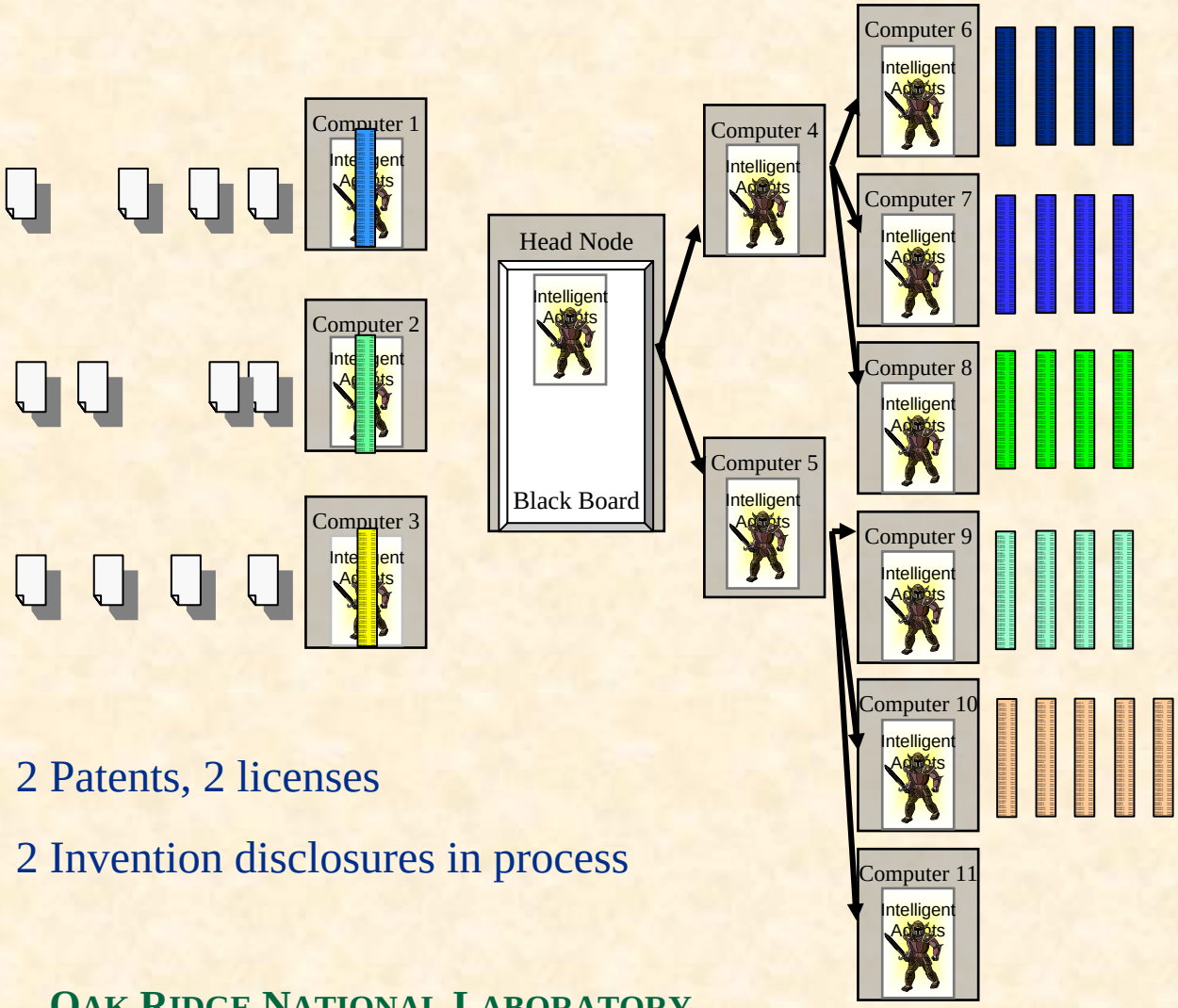
UPGMA (HAC)

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.

How Piranha Works

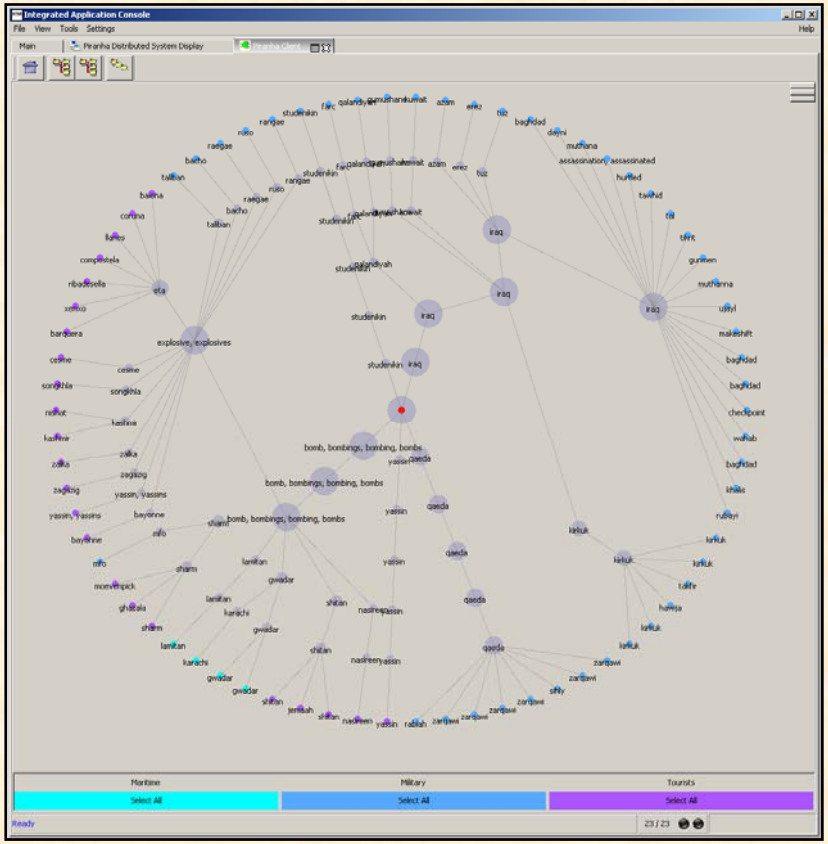


- Standard Approach
  - 11.5 Days
- Agent approach
  - 8 minutes 24 Seconds!
  - 2000 times faster
  - With no loss of accuracy

2 Patents, 2 licenses

2 Invention disclosures in process

Piranha Text Analysis



US PACOM

"Tremendously successful project"

"Software agents … lead to substantially improved analytical products."

"A grand slam home run!"

Software Agents "working at HQ USCINCPAC operationally."

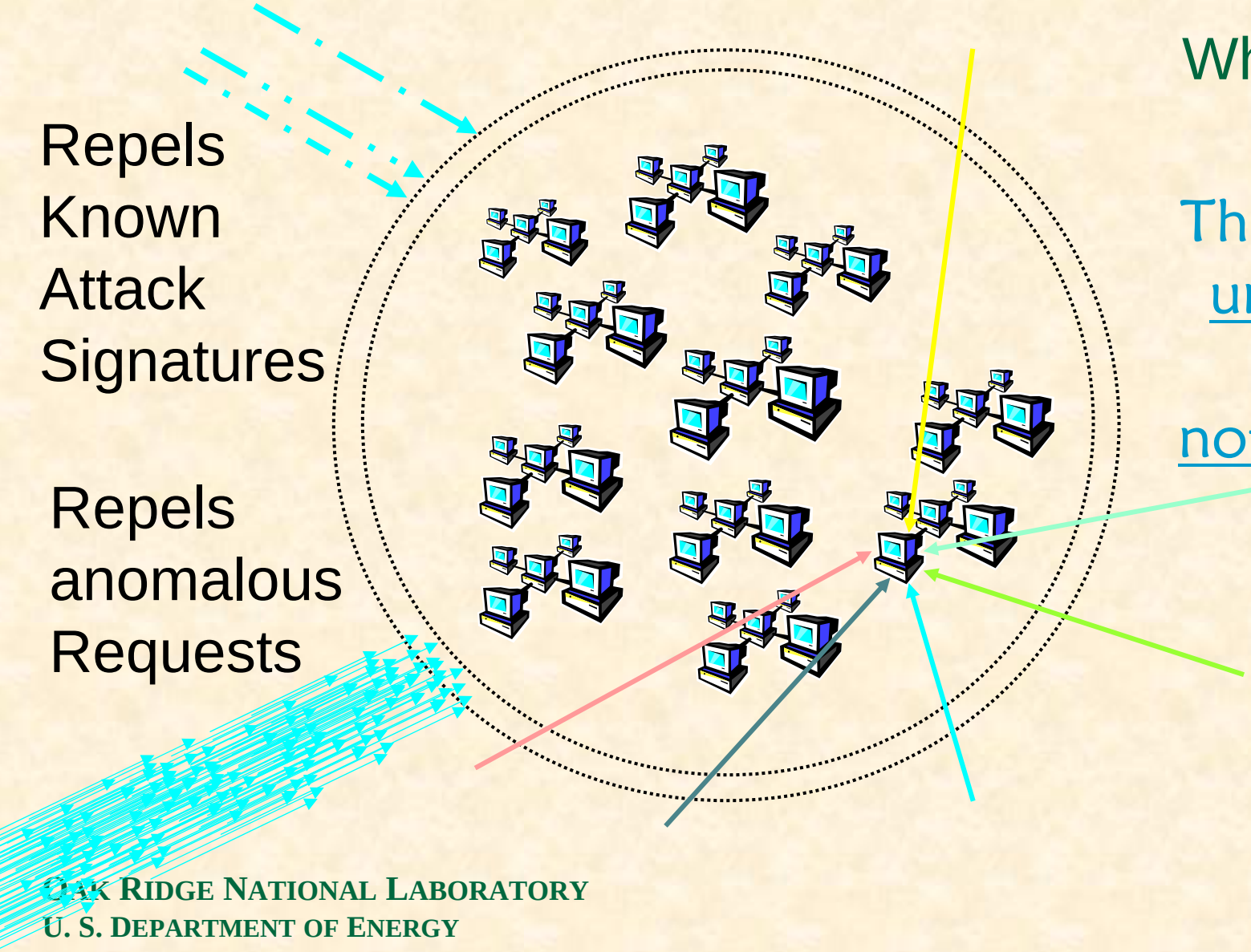
Potok et al., "VIPAR: Advanced Information Agents discovering knowledge in an open and changing environment," *Proc. 7th World Multiconference on Systemics, Cybernetics and Informatics Special Session on Agent-Based Computing*, 2003. Awarded Best Paper

Elmore, et al., "Dynamic Data Fusion Using An Ontology-Based Software Agent System," *7th World Multiconference on Systemics, Cybernetics and Informatics*, 2003.

CYPHER: Current Cyber security Approach

Repels Known Attack Signatures

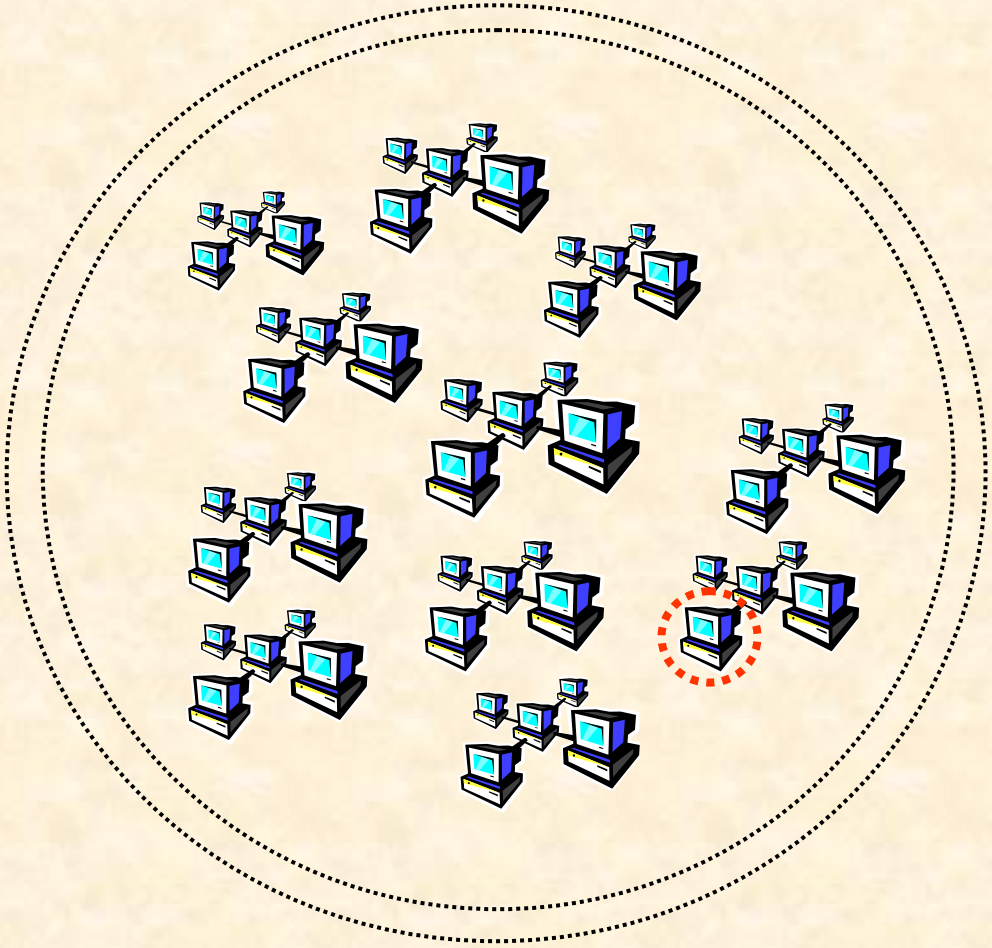
Repels anomalous Requests



What if …

The attack is unknown, and not unusual?

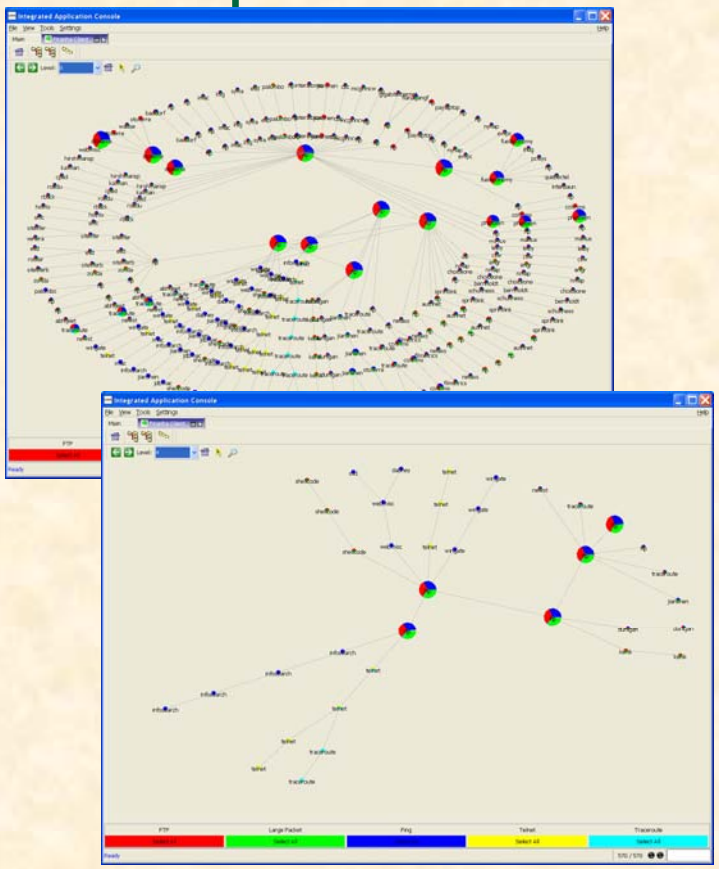
CYPHER: New Approach



Find who is going after the "gold"

CYPHER: Cybersecurity Example

- **1,000,000 records daily**
- **Critical Decisions**
  - **Is someone trying to hack into the system?**
  - **Are attack from "script kiddies" or state sponsored?**
  - **What are they going after and why?**



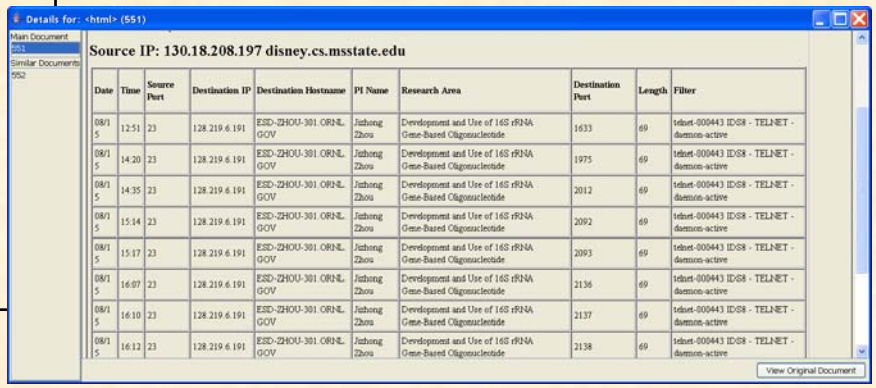
NetReg Database
63.76.192.107
John Doe
BN 123456

Activity Log Entry

Filter:	misc-000264 IDS247 - MISC - Large UDP Packet
Date:	07/20
TOD:	00:05:03.171193
Source IP:	63.76.192.107
User	John Doe
Research Area	Nuclear Physics
Source Port:	23882
Target IP:	160.91.64.211
Target Name:	smith.aol.com
Target Port:	6970
Length:	1285

CME Database
Johnathon Doe
BN 123456
Nuclear Physics

DNS Database
63.76.192.107
John Doe
BN 123456



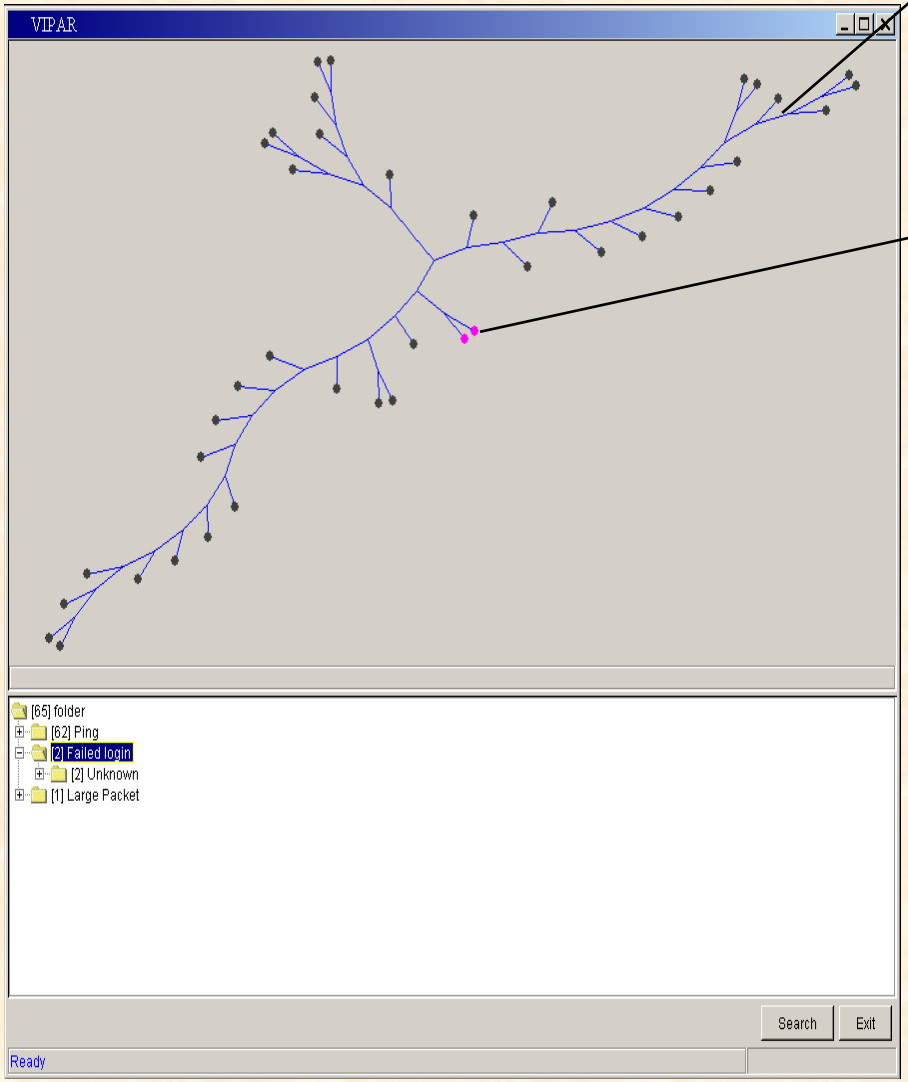
Details for: <html> (551)

Source IP: 130.18.208.197 disney.cs.msstate.edu

Date	Time	Source Port	Destination IP	Destination Hostname	PI Name	Research Area	Destination Port	Length	Filter
08/15	12:51	23	128.219.6.191	ESD-ZHOU-301.ORNL.GOV	Jizhong Zhou	Development and Use of 16S rRNA Gene-Based Oligonucleotide	1633	69	telnet-000443 IDS8 - TELNET - daemon-active
08/15	14:20	23	128.219.6.191	ESD-ZHOU-301.ORNL.GOV	Jizhong Zhou	Development and Use of 16S rRNA Gene-Based Oligonucleotide	1975	69	telnet-000443 IDS8 - TELNET - daemon-active
08/15	14:35	23	128.219.6.191	ESD-ZHOU-301.ORNL.GOV	Jizhong Zhou	Development and Use of 16S rRNA Gene-Based Oligonucleotide	2012	69	telnet-000443 IDS8 - TELNET - daemon-active
08/15	15:14	23	128.219.6.191	ESD-ZHOU-301.ORNL.GOV	Jizhong Zhou	Development and Use of 16S rRNA Gene-Based Oligonucleotide	2092	69	telnet-000443 IDS8 - TELNET - daemon-active
08/15	15:17	23	128.219.6.191	ESD-ZHOU-301.ORNL.GOV	Jizhong Zhou	Development and Use of 16S rRNA Gene-Based Oligonucleotide	2093	69	telnet-000443 IDS8 - TELNET - daemon-active
08/15	16:07	23	128.219.6.191	ESD-ZHOU-301.ORNL.GOV	Jizhong Zhou	Development and Use of 16S rRNA Gene-Based Oligonucleotide	2136	69	telnet-000443 IDS8 - TELNET - daemon-active
08/15	16:10	23	128.219.6.191	ESD-ZHOU-301.ORNL.GOV	Jizhong Zhou	Development and Use of 16S rRNA Gene-Based Oligonucleotide	2137	69	telnet-000443 IDS8 - TELNET - daemon-active
08/15	16:12	23	128.219.6.191	ESD-ZHOU-301.ORNL.GOV	Jizhong Zhou	Development and Use of 16S rRNA Gene-Based Oligonucleotide	2138	69	telnet-000443 IDS8 - TELNET - daemon-active

View Original Document

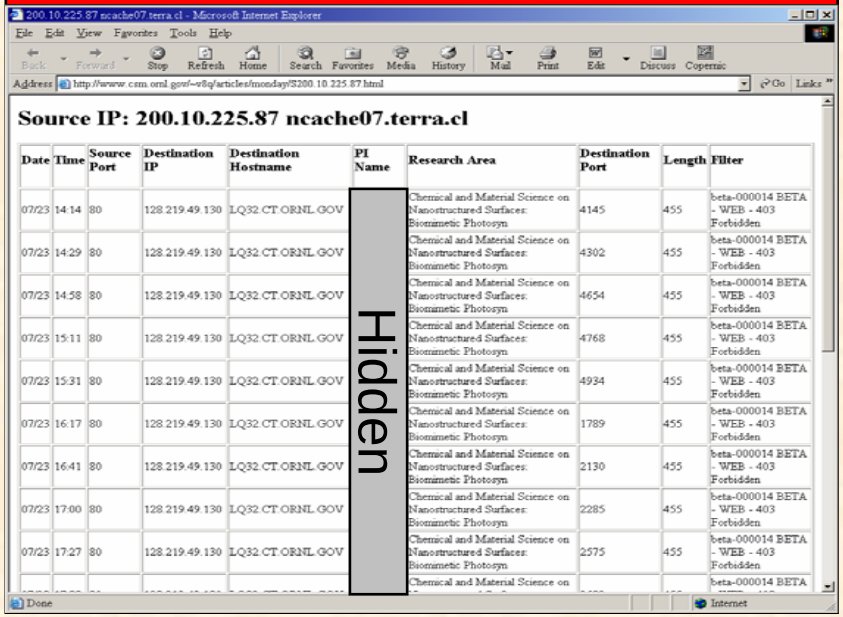
Caught!



45 "low and slow" pings from:

- **Czech Republic, Austria, Hungary, Latvia, France, Chile, and Canada.**

2 attacks on nanotechnology scientists



Source IP: 200.10.225.87 ncache07.terra.cl

Date	Time	Source Port	Destination IP	Destination Hostname	PI Name	Research Area	Destination Port	Length	Filter
07/23	14:14	80	128.219.49.130	LQ32.CT.ORNL.GOV	Hidden	Chemical and Material Science on Nanostructured Surfaces: Biomimetic Photosyn	4145	455	beta-000014 BETA - WEB - 403 Forbidden
07/23	14:29	80	128.219.49.130	LQ32.CT.ORNL.GOV		Chemical and Material Science on Nanostructured Surfaces: Biomimetic Photosyn	4302	455	beta-000014 BETA - WEB - 403 Forbidden
07/23	14:58	80	128.219.49.130	LQ32.CT.ORNL.GOV		Chemical and Material Science on Nanostructured Surfaces: Biomimetic Photosyn	4654	455	beta-000014 BETA - WEB - 403 Forbidden
07/23	15:11	80	128.219.49.130	LQ32.CT.ORNL.GOV		Chemical and Material Science on Nanostructured Surfaces: Biomimetic Photosyn	4768	455	beta-000014 BETA - WEB - 403 Forbidden
07/23	15:31	80	128.219.49.130	LQ32.CT.ORNL.GOV		Chemical and Material Science on Nanostructured Surfaces: Biomimetic Photosyn	4934	455	beta-000014 BETA - WEB - 403 Forbidden
07/23	16:17	80	128.219.49.130	LQ32.CT.ORNL.GOV		Chemical and Material Science on Nanostructured Surfaces: Biomimetic Photosyn	1789	455	beta-000014 BETA - WEB - 403 Forbidden
07/23	16:41	80	128.219.49.130	LQ32.CT.ORNL.GOV		Chemical and Material Science on Nanostructured Surfaces: Biomimetic Photosyn	2130	455	beta-000014 BETA - WEB - 403 Forbidden
07/23	17:00	80	128.219.49.130	LQ32.CT.ORNL.GOV		Chemical and Material Science on Nanostructured Surfaces: Biomimetic Photosyn	2285	455	beta-000014 BETA - WEB - 403 Forbidden
07/23	17:27	80	128.219.49.130	LQ32.CT.ORNL.GOV		Chemical and Material Science on Nanostructured Surfaces: Biomimetic Photosyn	2575	455	beta-000014 BETA - WEB - 403 Forbidden
						Chemical and Material Science on			beta-000014 BETA

Patented

Summary

- **Current technology cannot solve emerging national challenges**
- **Intelligent software agents are a <u>significant</u> breakthrough technology**
- **Results indicate <u>high-potential</u> 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