Novel, Large Scale, High Speed Text and Data Analysis Research

Presented by

Thomas E. Potok, Ph.D

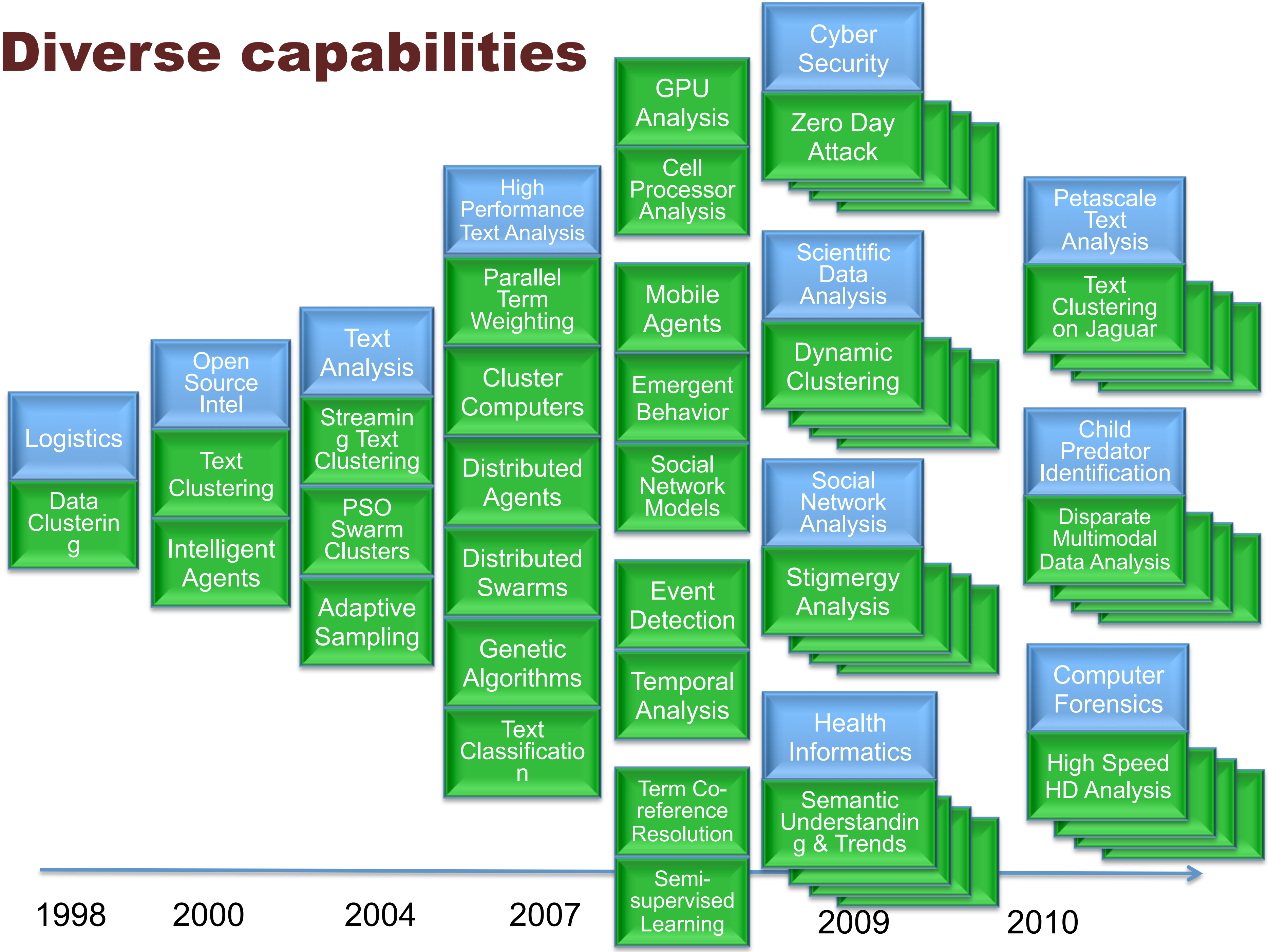
Group Leader
Applied Software Engineering Research

Who we are

ASER | Applied Software Engineering Research Group

- **Group focus**
  - 9 years of research in intelligent agents, emergent behavior, pervasive computing, machine learning, information retrieval, and knowledge discovery
  - In 2009 we organized 7 research workshops and published 26 papers and book chapters
  - Hands-on experience with DHS, Military, IC, and Industry
- **Group success**
  - $15M in research investment
  - 14 staff members, 9 PhDs in computer science
  - R&D 100 Award (Oscars of invention) in 2007

Diverse capabilities



Logistics

Data Clustering

Open Source Intel

Text Clustering

Intelligent Agents

Text Analysis

Streaming Text Clustering

PSO Swarm Clusters

Adaptive Sampling

High Performance Text Analysis

Parallel Term Weighting

Cluster Computers

Distributed Agents

Distributed Swarms

Genetic Algorithms

Text Classification

GPU Analysis

Cell Processor Analysis

Mobile Agents

Emergent Behavior

Social Network Models

Event Detection

Temporal Analysis

Term Co-reference Resolution

Semi-supervised Learning

Cyber Security

Zero Day Attack

Scientific Data Analysis

Dynamic Clustering

Social Network Analysis

Stigmergy Analysis

Health Informatics

Semantic Understanding & Trends

Petascale Text Analysis

Text Clustering on Jaguar

Child Predator Identification

Disparate Multimodal Data Analysis

Computer Forensics

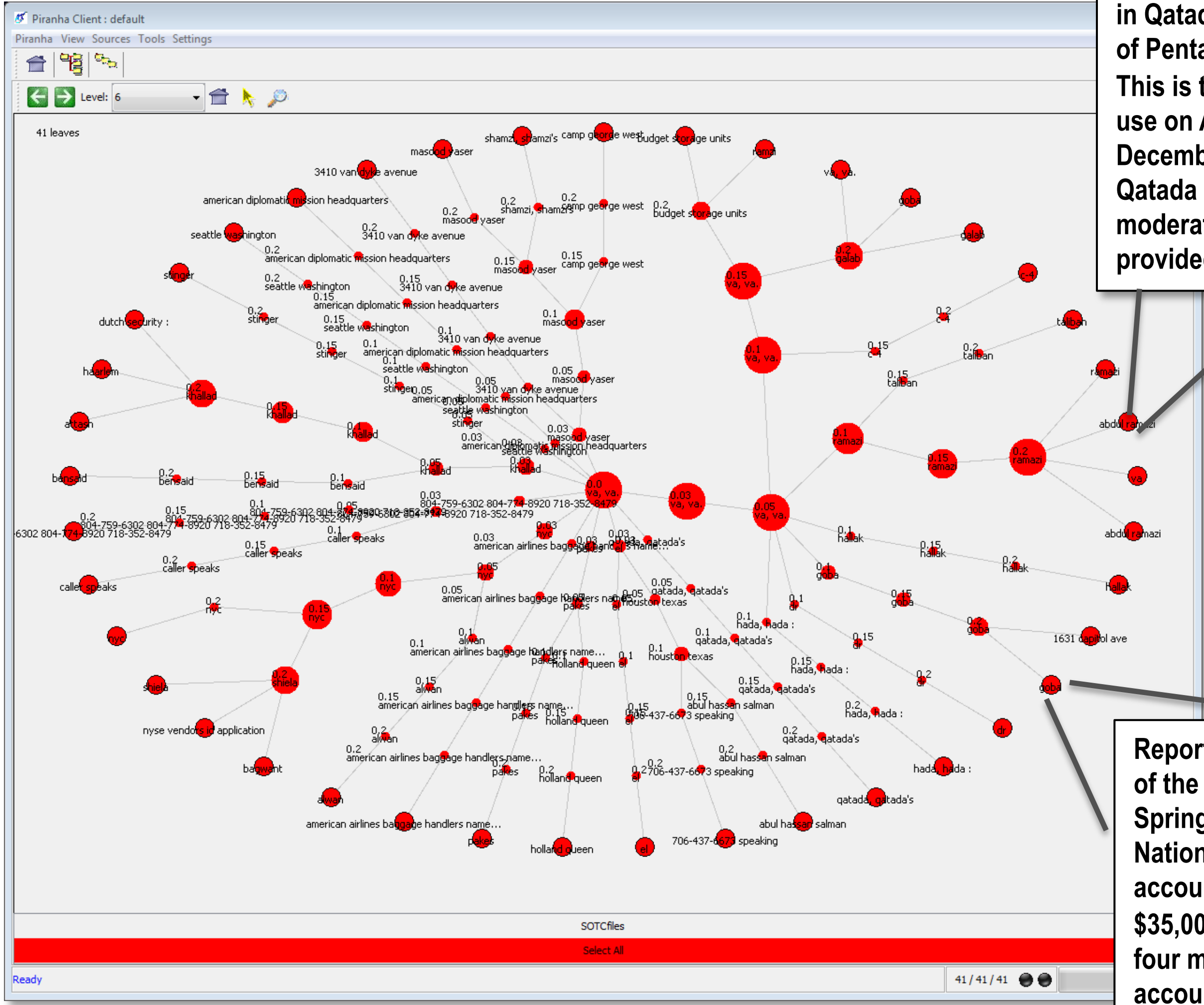
High Speed HD Analysis

1998 2000 2004 2007 2009 2010

Text analysis capability overview

Capability	Capacity in documents	Search Engines	Natural Language Tools	Piranha
Search/Query	100M+	Yes	No	Yes
Unsupervised classification	1M	Some	No	Yes
Supervised classification	1M	No	No	Yes
Clustering (Full document analysis)	100K	No	No	Yes
Term Frequency Analysis (Significant words)	100K	Yes, but not available to user	Yes	Yes
Semantic Extraction (People, places)	1000	No	Yes	Yes

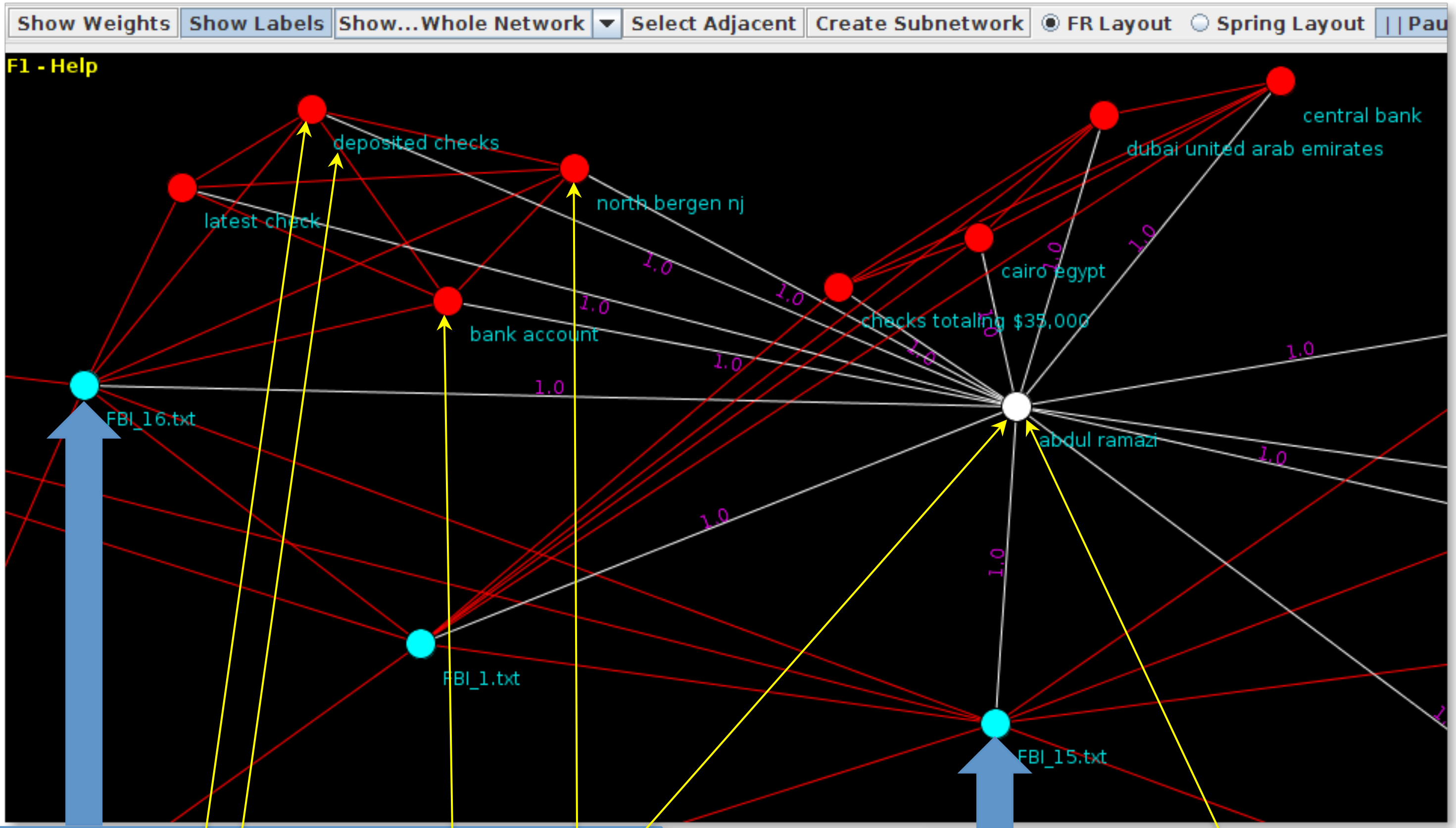
Piranha cluster view



Report Date: 1 April, 2003. CIA [From MI5]: On 30 March, 2003 the British Special Branch arrested Omar Bakri Qatada at his home at #11 St. Mary's Terrace, Paddington, London. Found in Qatada's bedroom was a small carton holding 10 ounces of Pentaerythritol [PETN] and Triacetone Triperoxide [TATP]. This is the same explosive that Richard Reid attempted to use on American Airlines #63 from Paris to Miami on 22 December, 2001. The BSB were alerted to follow and detain Qatada on the basis of information obtained from a respected moderate Moslem cleric in London, whose name was not provided in this report from MI5.

Report Date: 1 April, 2003. FBI: Abdul Ramazi is the owner of the Select Gourmet Foods shop in Springfield Mall, Springfield, VA. [Phone number 703-659-2317]. First Union National Bank lists Select Gourmet Foods as holding account number 1070173749003. Six checks totaling $35,000 have been deposited in this account in the past four months and are recorded as having been drawn on accounts at the Pyramid Bank of Cairo, Egypt and the Central Bank of Dubai, United Arab Emirates. Both of these banks have just been listed as possible conduits in money laundering schemes

Term network with document links



16) Report Date 22 April 2003. FBI: Hani al Hallak, of **North Bergen, NJ**, has **deposited checks** in his **bank account** that were drawn on First Union Bank account number 1070173749003 in Springfield, VA, in the name **Abdul Ramazi**. The latest check is dated 16 April 2003 and was in the amount of $8500.

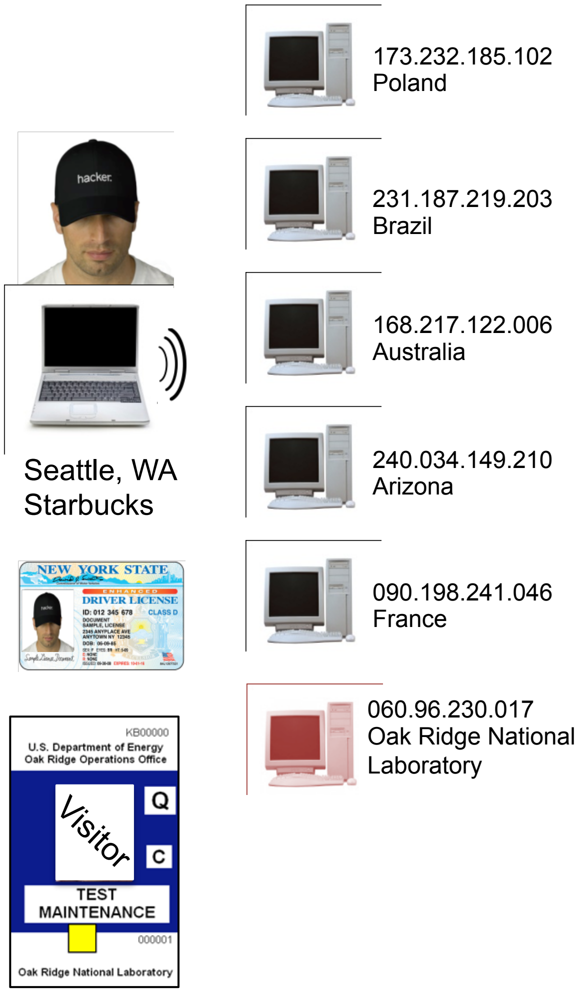
15) Report Date 20 April 2003. FBI: Mukhtar Galab has an account at the Virginia National Bank in Charlottesville, VA. Bank records say he has deposited several checks in the last three months, totaling $13,000, drawn on account number 1070173749003 held by **Abdul Ramazi** at the First Union Bank in Springfield, VA

Petascale text analysis

- ORNL’s Jaguar is the fastest computer in the world
  - 255,000 cores -10PB (13,400 1TB drives) of Storage -362TB of memory
- Google has indexed 1 Trillion unique URLs, but has not analyzed the content of the information
- We are currently developing petascale text analysis techniques to cluster (deep analysis) of 1 trillion documents using Jaguar

Exfiltrate the data from real targets

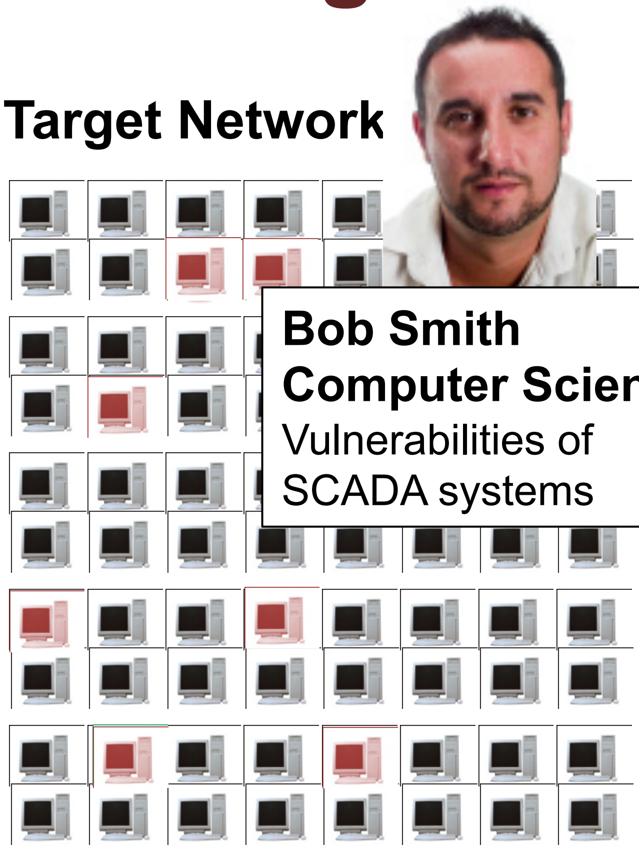
Hacker's Machines



Seattle, WA
Starbucks

Jane Doe
Visitor Control
Names, SSN, DOB, dates of visits, clearances, authorized buildings

Target Network

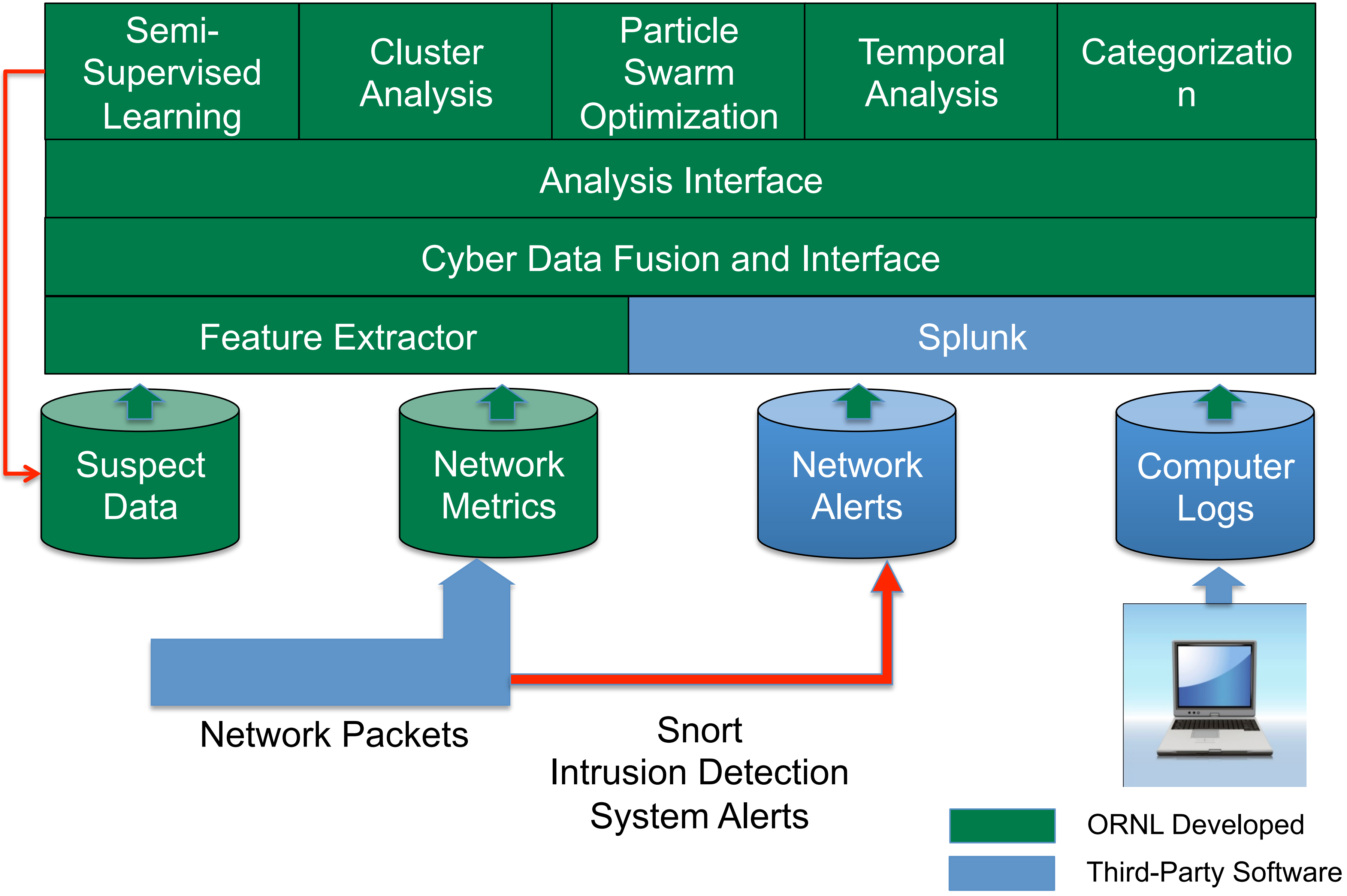


Bob Smith
Computer Scientist
Vulnerabilities of SCADA systems

- **Download all files from Smith and Doe's computers**
- **Sell Smith's data to highest bidder – Iran or North Korea**
- **Find open visitor requests for a visitor with a clearance**
- **Forge a driver license**
- **Pick up the badge at Visitor Control**

New approach

ORCA
OAK RIDGE CYBER ANALYTICS



Semi-Supervised Learning

Cluster Analysis

Particle Swarm Optimization

Temporal Analysis

Categorization

Analysis Interface

Cyber Data Fusion and Interface

Feature Extractor

Splunk

Suspect Data

Network Metrics

Network Alerts

Computer Logs

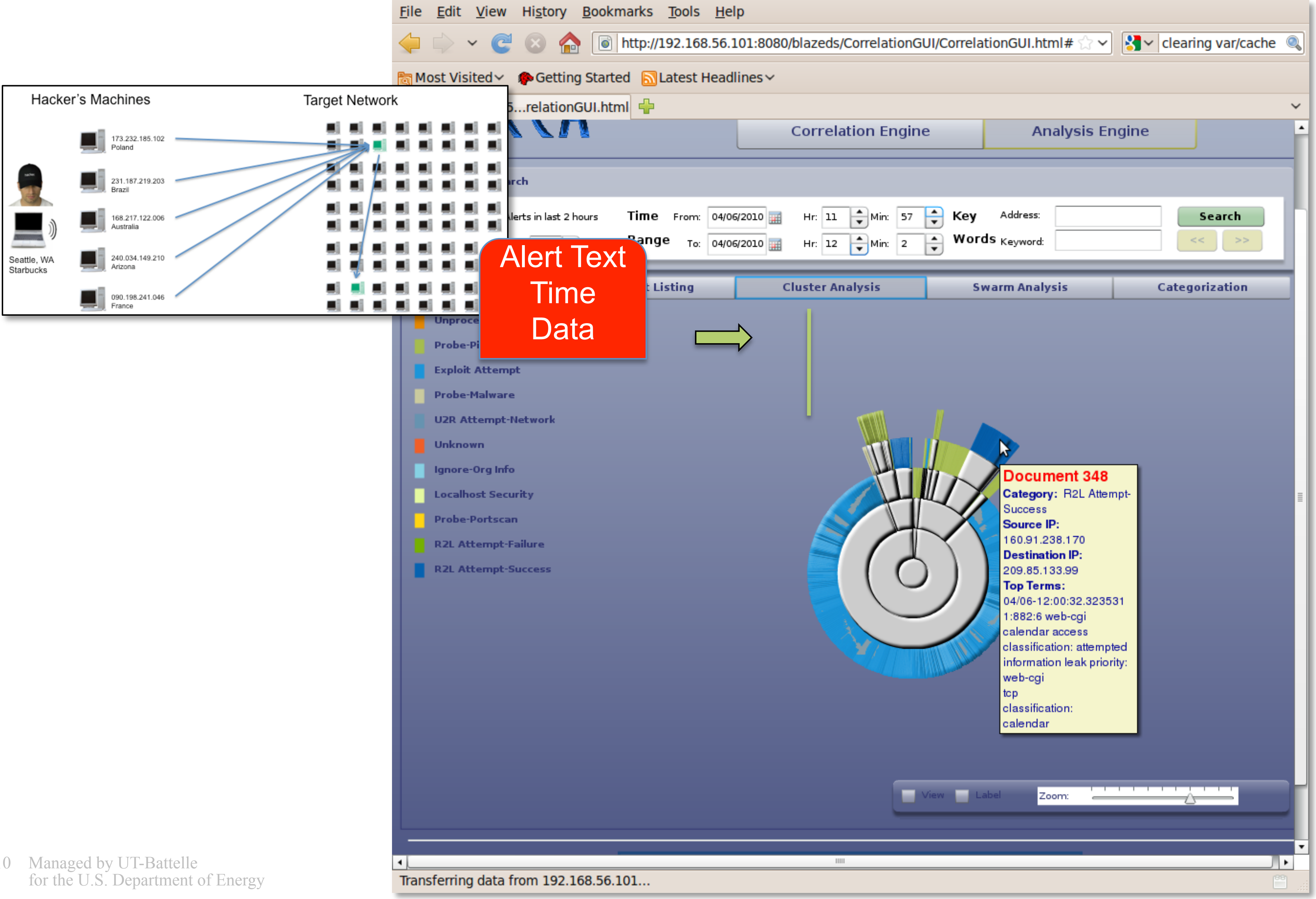
Network Packets

Snort
Intrusion Detection
System Alerts

🟩 ORNL Developed

🟦 Third-Party Software

Find attacks on vulnerable computers



From: 04/06/2010 Hr: 11 Min: 57
To: 04/06/2010 Hr: 12 Min: 2
Address:
Keyword:

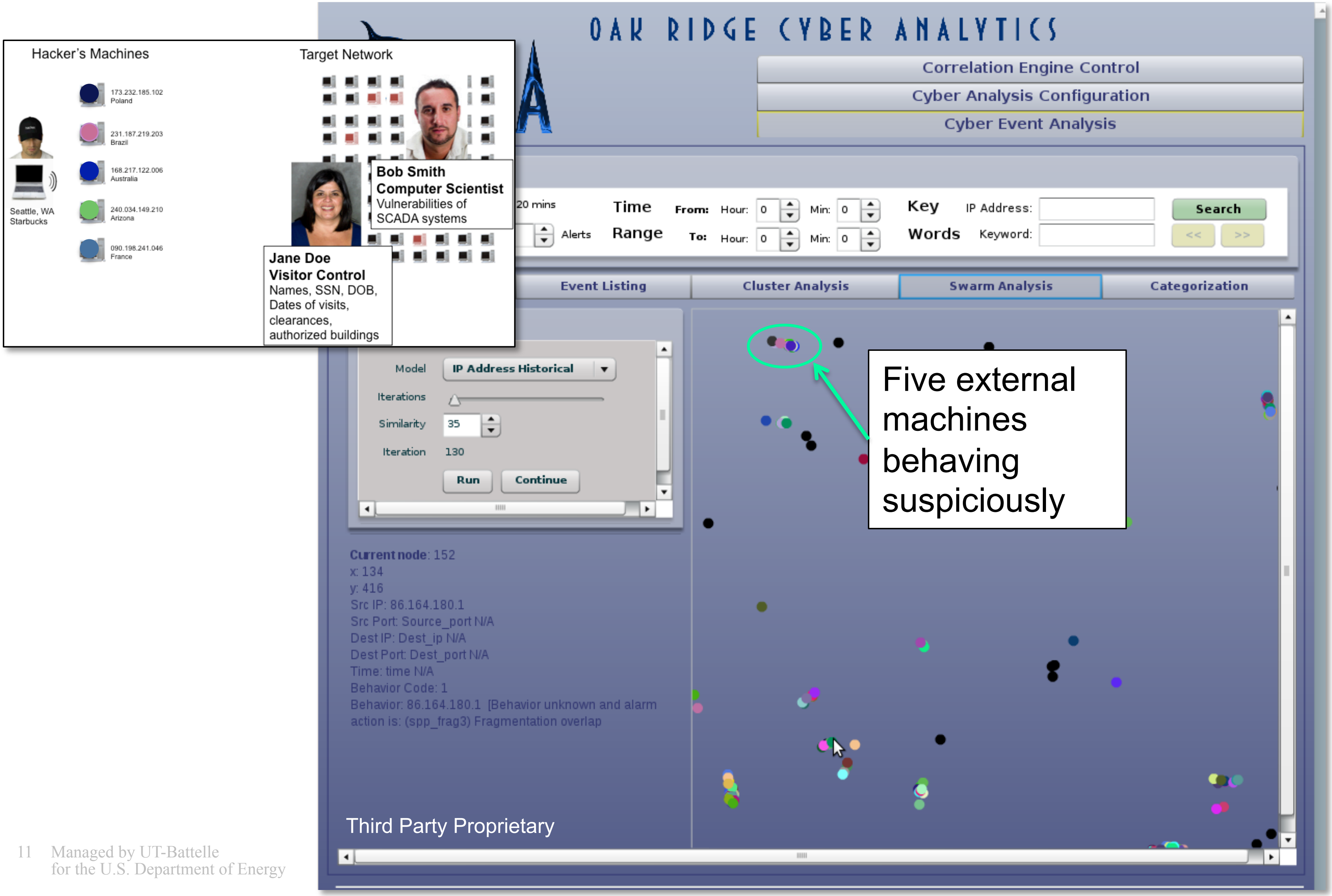
- 🟧 Unproce
- 🟩 Probe-Pi
- 🟦 Exploit Attempt
- ⬜ Probe-Malware
- 🟦 U2R Attempt-Network
- 🟥 Unknown
- 🟦 Ignore-Org Info
- 🟨 Localhost Security
- 🟨 Probe-Portscan
- 🟩 R2L Attempt-Failure
- 🟦 R2L Attempt-Success

Document 348
Category: R2L Attempt-Success
Source IP: 160.91.238.170
Destination IP: 209.85.133.99
Top Terms:
04/06-12:00:32.323531
1:882:6 web-cgi
calendar access
classification: attempted
information leak priority:
web-cgi
tcp
classification:
calendar

Transferring data from 192.168.56.101...

Alert Text
Time
Data

Find exploits



Bob Smith
Computer Scientist
Vulnerabilities of
SCADA systems

Jane Doe
Visitor Control
Names, SSN, DOB,
Dates of visits,
clearances,
authorized buildings

Model: IP Address Historical
Iterations
Similarity: 35
Iteration: 130

```
Current node: 152
x: 134
y: 416
Src IP: 86.164.180.1
Src Port: Source_port N/A
Dest IP: Dest_ip N/A
Dest Port: Dest_port N/A
Time: time N/A
Behavior Code: 1
Behavior: 86.164.180.1  [Behavior unknown and alarm
action is: (spp_frag3) Fragmentation overlap
```

Five external machines behaving suspiciously

Third Party Proprietary

Current capabilities

- TRL 7+ in 18 month
  - Live Sprint and AT&T feeds
  - 3+ terabytes of network traffic daily
- Extraction of network features – Alerts plus network traffic metrics
- Machine learning – Incorporate new insights into analysis
- Cluster analysis – Group similar alerts
- Particle swarm optimization – Group similar computer behavior
- Temporal analysis – View time vs. IP Addresses vs. alert type
- Categorization – Group alerts by defined category

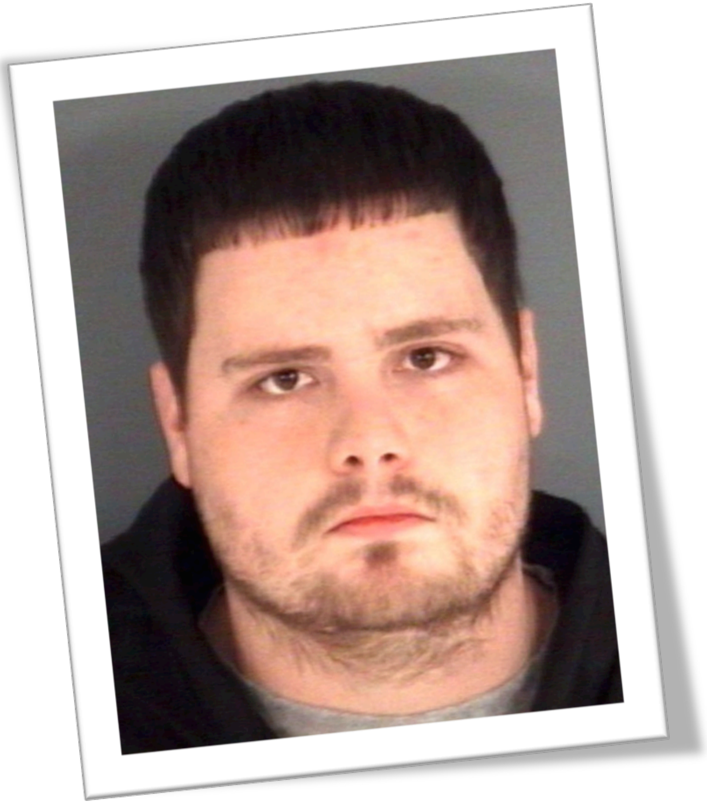
Child predators

Somer Thompson

2002 - 2009

- This is Somer. She was born April 5th 2002 in Columbia, South Carolina.
- She enjoyed swimming and dancing. She often thought of herself as becoming a ballerina.
- On October 19th 2009 at 2:45pm Somer left Grove Park Elementary school to meet up with her brother and sister for their daily mile walk home.
- She ran ahead of her siblings and got separated from the group
- This was the very last time Somer Thompson was seen alive.
- Her body was found in a Georgia landfill 2 days later. At that time, there were no suspects.

Child predators (continued)

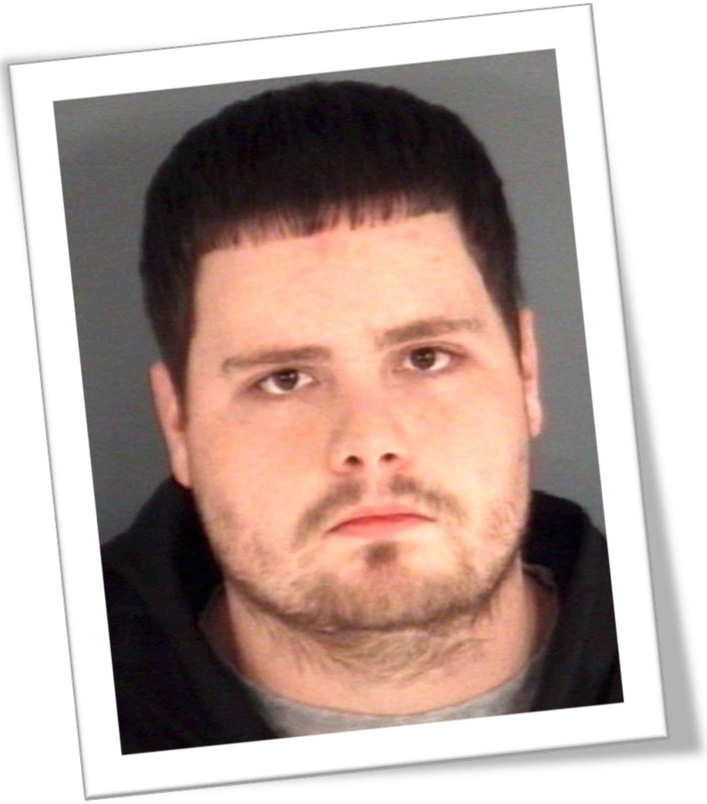


Jarred Harrell

He abducted Somer that day.
He was charged with raping
and murdering her.

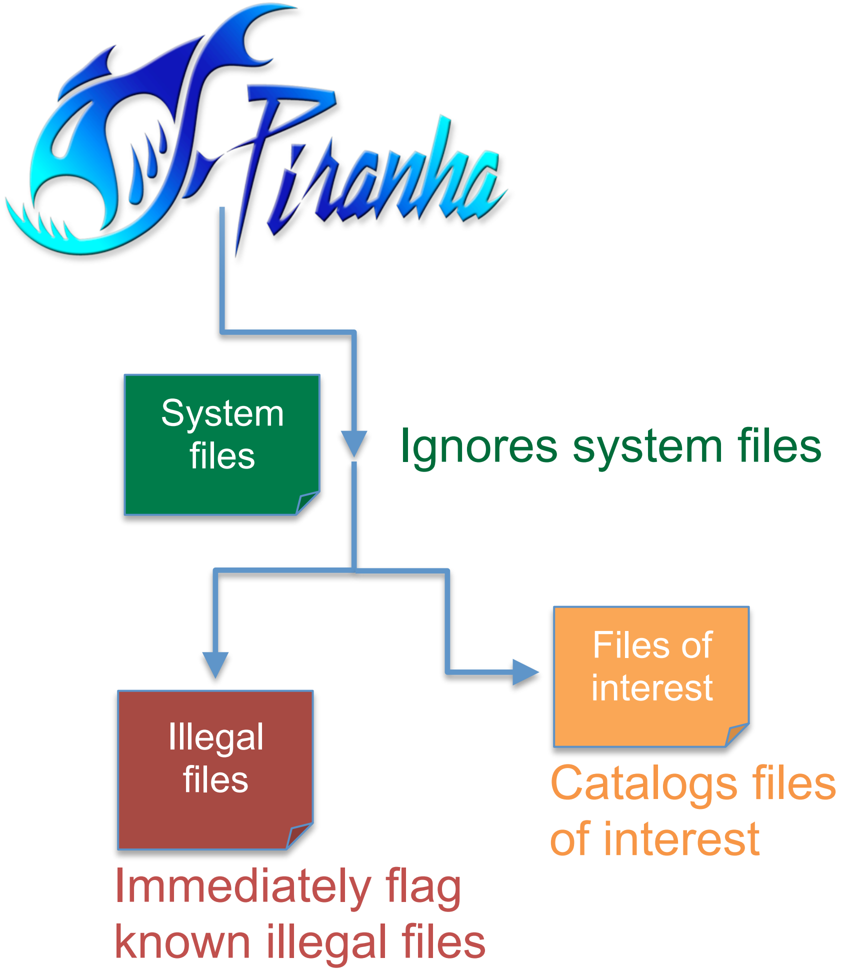
"…Police had possession of Harrell's computer and camera two months before Thompson disappeared, …'they found Harrell's camera containing a video with horrifying images of Harrell molesting a three year-old child'," reports ABC. - Saturday, 17 April 2010 10:54 ABC News' 20/20

Fact



If Jarred Harrell's computer and camera had been forensically analyzed in a timely manner, Somer would still be alive.

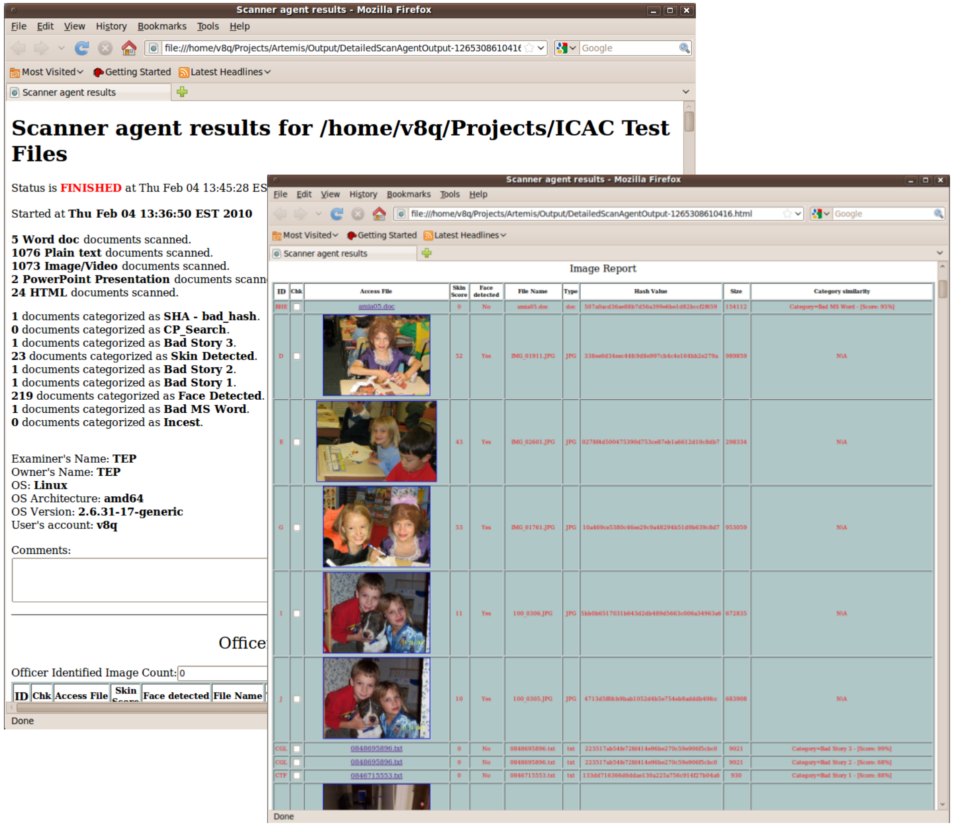
Sampling



- Sample suspect's disk to find a predator immediately
- Assuming illegal files are grouped together becomes a search problem
- Full analysis can be completed later

Artemis – 30 child pornography arrests in Knoxville, TN

- Working with the *Protect* and *Internet Crimes Against Children* (ICAC) Task Force in Knoxville, TN
- A prototype thumb drive system to rapidly analyze information on a hard drive and identify illegal material
- When booted loads Knoppix and Artemis to allows the officer to scan specific directories or the entire disk for images and documents.
- Image analysis looks for high percentage of skin tones, faces, or an illegal hash values
- Text analysis looks for documents that are similar to a provided set.


Scanner agent results - Mozilla Firefox
File Edit View History Bookmarks Tools Help
file:///home/v8q/Projects/Artemis/Output/DetailedScanAgentOutput-126530861041
Google
Most Visited Getting Started Latest Headlines
Scanner agent results
Done

Scanner agent results for /home/v8q/Projects/ICAC Test Files

Status is FINISHED at Thu Feb 04 13:45:28 ES

Started at **Thu Feb 04 13:36:50 EST 2010**

5 Word doc documents scanned.
1076 Plain text documents scanned.
1073 Image/Video documents scanned.
2 PowerPoint Presentation documents scann
24 HTML documents scanned.

1 documents categorized as **SHA - bad_hash**.
0 documents categorized as **CP_Search**.
1 documents categorized as **Bad Story 3**.
23 documents categorized as **Skin Detected**.
1 documents categorized as **Bad Story 2**.
1 documents categorized as **Bad Story 1**.
219 documents categorized as **Face Detected**.
1 documents categorized as **Bad MS Word**.
0 documents categorized as **Incest**.

Examiner's Name: **TEP**
Owner's Name: **TEP**
OS: **Linux**
OS Architecture: **amd64**
OS Version: **2.6.31-17-generic**
User's account: **v8q**

Comments:

Office

Officer Identified Image Count: 0

ID	Chk	Access File	Skin Score	Face detected	File Name

Scanner agent results - Mozilla Firefox
File Edit View History Bookmarks Tools Help
file:///home/v8q/Projects/Artemis/Output/DetailedScanAgentOutput-1265308610416.html
Google
Most Visited Getting Started Latest Headlines
Scanner agent results
Done

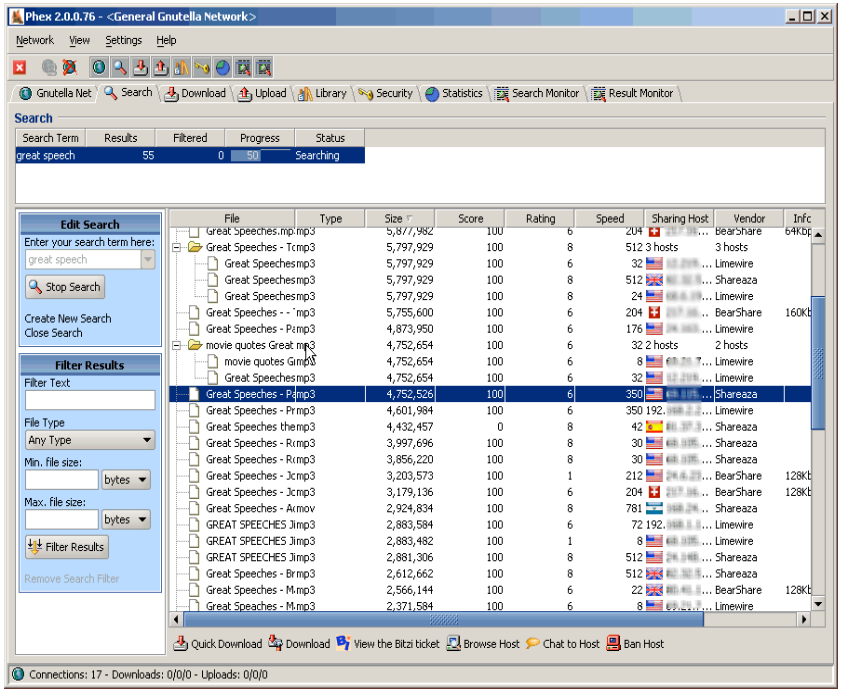
Image Report

ID	Chk	Access File	Skin Score	Face detected	File Name	Type	Hash Value	Size	Category similarity
BHE	□	amia05.doc	0	No	amia05.doc	doc	507a0acd36ae88b7d56a399e6be1d82bccf2f659	154112	Category=Bad MS Word - [Score: 95%]
D	□		52	Yes	IMG_01911.JPG	JPG	338ee0d34eec44fc9d8e997cb4c4e164bb2e279a	989859	N\A
E	□		43	Yes	IMG_02601.JPG	JPG	0278f4d500475390d753ce87eb1a6612d10c8db7	298334	N\A
G	□		53	Yes	IMG_01761.JPG	JPG	10a469ce5380c46ee29c9a48294b51d9b639c8d7	953059	N\A
I	□		11	Yes	100_0306.JPG	JPG	5bb0b6517031b643d2db489d5663c006a34963a6	672835	N\A
J	□		10	Yes	100_0305.JPG	JPG	4713d5f8fcb9bab1052d4b5e754eb8adddb49fcc	683908	N\A
CGL	□	0848695896.txt	0	No	0848695896.txt	txt	223517ab54fe72fd414e96be270c59e906f5cbc0	9021	Category=Bad Story 3 - [Score: 99%]
CGL	□	0848695896.txt	0	No	0848695896.txt	txt	223517ab54fe72fd414e96be270c59e906f5cbc0	9021	Category=Bad Story 2 - [Score: 68%]
CTF	□	0846715553.txt	0	No	0846715553.txt	txt	133dd716366d6ddae130a225a756c914f27b04a6	930	Category=Bad Story 1 - [Score: 88%]

ASLAN – Graph Analysis of Large file sharing network

- You can make part of your computer public, so that others can download your files
- The software keeps track of the IP addresses of the computers that download the files
- Social networks can be built to see who is sharing from whom

File Sharing Software



Phex 2.0.0.76 - <General Gnutella Network>

Network View Settings Help

Gnutella Net | Search | Download | Upload | Library | Security | Statistics | Search Monitor | Result Monitor

Search

Search Term	Results	Filtered	Progress	Status
great speech	55	0	50	Searching

Edit Search

Enter your search term here: great speech

Stop Search

Create New Search

Close Search

Filter Results

Filter Text:

File Type: Any Type

Min. file size: bytes

Max. file size: bytes

Filter Results

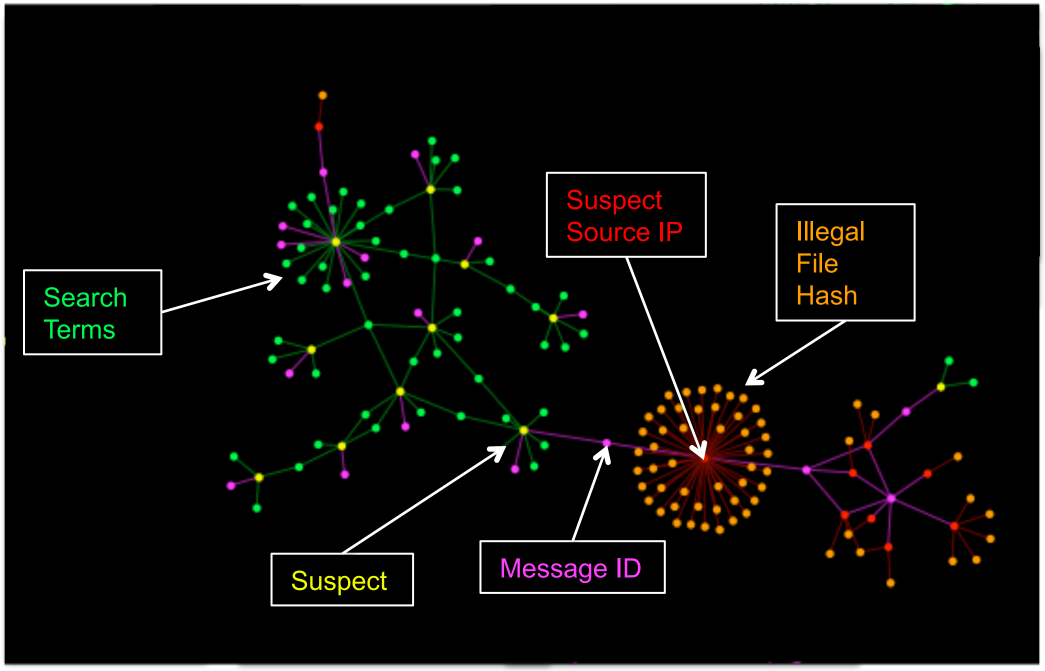
Remove Search Filter

File	Type	Size	Score	Rating	Speed	Sharing Host	Vendor	Info
Great Speeches.mp:	mp3	5,877,982	100	6	204	[redacted]…	BearShare	64Kbp
Great Speeches - Tc	mp3	5,797,929	100	8	512	3 hosts	3 hosts	
Great Speeches	mp3	5,797,929	100	6	32	[redacted]…	Limewire	
Great Speeches	mp3	5,797,929	100	8	512	[redacted]…	Shareaza	
Great Speeches	mp3	5,797,929	100	8	24	[redacted]…	Limewire	
Great Speeches - - ˈ	mp3	5,755,600	100	6	204	[redacted]..	BearShare	160Kb
Great Speeches - Pa	mp3	4,873,950	100	6	176	[redacted]…	Limewire	
movie quotes Great	mp3	4,752,654	100	6	32	2 hosts	2 hosts	
movie quotes Gi	mp3	4,752,654	100	6	8	[redacted]…	Limewire	
Great Speeches	mp3	4,752,654	100	6	32	[redacted]…	Limewire	
Great Speeches - Pa	mp3	4,752,526	100	6	350	[redacted]…	Shareaza	
Great Speeches - Pr	mp3	4,601,984	100	6	350	192.[redacted]…	Limewire	
Great Speeches the	mp3	4,432,457	0	8	42	[redacted]…	Shareaza	
Great Speeches - Rc	mp3	3,997,696	100	8	30	[redacted]…	Shareaza	
Great Speeches - Rc	mp3	3,856,220	100	8	30	[redacted]…	Shareaza	
Great Speeches - Jc	mp3	3,203,573	100	1	212	[redacted]…	BearShare	128Kb
Great Speeches - Jc	mp3	3,179,136	100	6	204	[redacted]..	BearShare	128Kb
Great Speeches - Ac	mov	2,924,834	100	8	781	[redacted]..	Shareaza	
GREAT SPEECHES Ji	mp3	2,883,584	100	6	72	192.[redacted]…	Limewire	
GREAT SPEECHES Ji	mp3	2,883,482	100	1	8	[redacted]…	Limewire	
GREAT SPEECHES Ji	mp3	2,881,306	100	8	512	[redacted]…	Shareaza	
Great Speeches - Br	mp3	2,612,662	100	8	512	[redacted]…	Shareaza	
Great Speeches - M.	mp3	2,566,144	100	6	22	[redacted]…	BearShare	128Kb
Great Speaches - M.	mp3	2,371,584	100	6	8	[redacted]…	Limewire	

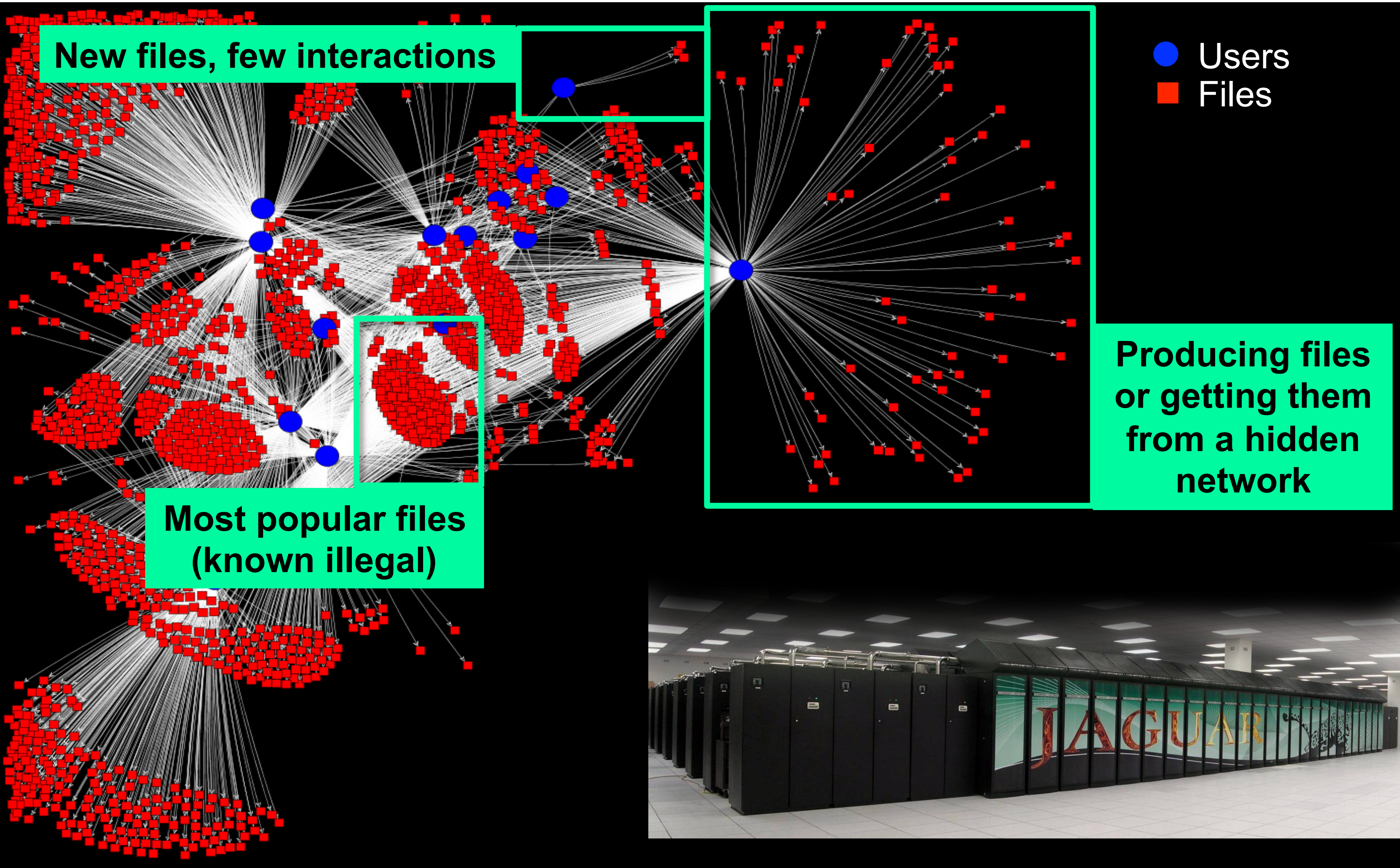
Quick Download | Download | View the Bitzi ticket | Browse Host | Chat to Host | Ban Host

Connections: 17 - Downloads: 0/0/0 - Uploads: 0/0/0

Social Network Graph



Use supercomputers to find the worst of the worst


● Users
🟥 Files
New files, few interactions
Producing files or getting them from a hidden network
Most popular files (known illegal)
JAGUAR

Summary

- Difficult national issues
- The challenge is in analyzing large volumes of unstructured data
- We are proven experts in analyzing large volumes of unstructured data
- Piranha and ORCA provides new ways of analyzing and correlating data
- This capability is applicable to any intelligence gathering and analyzing activity

Contact

Thomas E. Potok, Ph.D

Group Leader
Applied Software Engineering Research
potokte@ornl.gov