

Experiences in Enterprise Searching: Tips, Techniques, and Pitfalls

Andy Bair, Senior Security Consultant
KoreLogic, Inc
andy.bair@korelogic.com

Presentation

- Background
- Approach
- Process
- Conclusions



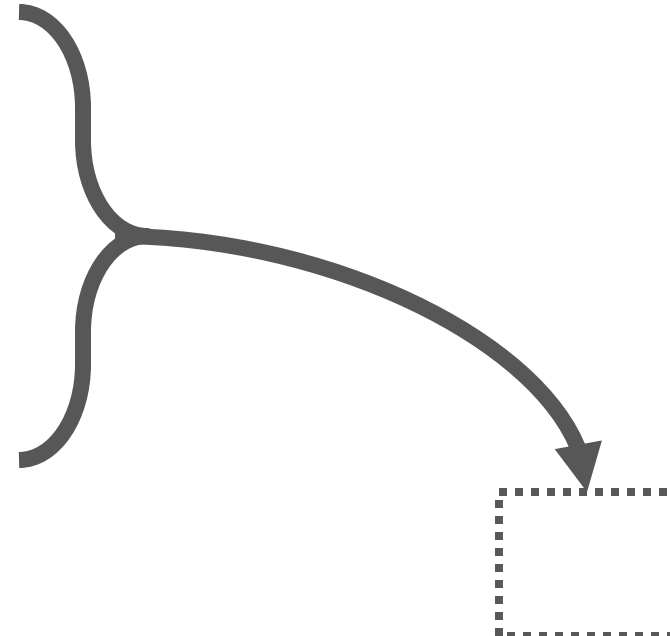
Tips



Techniques



Pitfalls



- Working definition
 - Enterprise-wide vs. handful of servers
 - Large-scale, data could be anywhere
 - Differs from a more traditional/limited forensic analysis
 - Examples:
 - Disclosure of passwords in systems that are SAS/SOX covered
 - Large data breach and data is ubiquitous
- This presentation will ...
 - Describe what we have learned in enterprise searching and various problems that arise
- This presentation will not ...
 - Compare and contrast commercial tools or alternate techniques

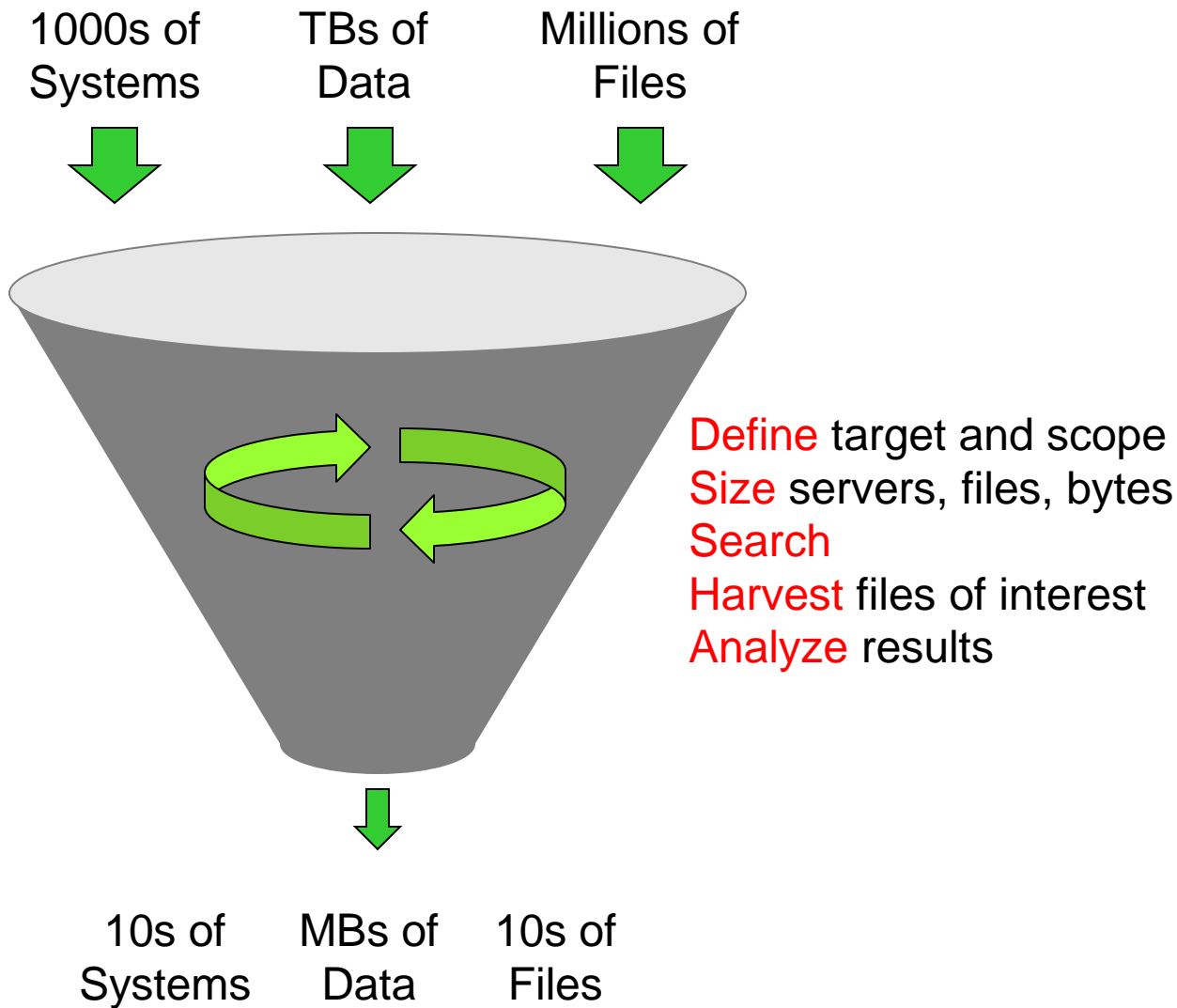
- Case 1: Clear-text password searching (compliance)
- Case 2: Data breach - targeted evidence-based searching

Case	Number Systems	Number Files	Bytes Searched	Number Terms	OSs
1	~100	~450,000	~23 TB	~1,800	AIX, Linux, Solaris
2	~1,800	~210,000,000	~59 TB	~50-200	Windows, HPUNIX, Linux, Solaris

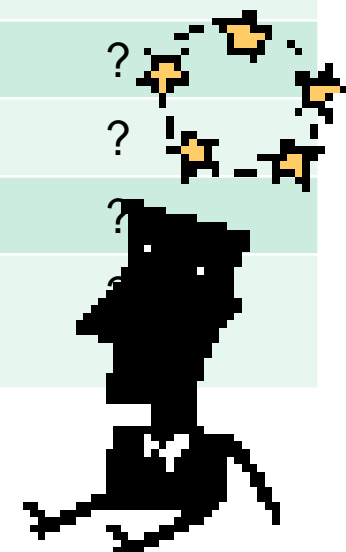
- Smaller search efforts:
 - Search for PII in email folders
 - Search for PII in hard drives from retail store registers

The Goal

Quickly Identify Relevant Hits



Question	Answer
How do you secure search terms, search results, and analysis?	?
How do you limit impact on production serves?	?
How do you pause searching during critical business hours?	?
How do you distill results into something meaningful?	?
How do you develop search terms?	?
How do you search a massive dataset in a timely fashion ?	?
How do you handle “ container ” files?	?
How do you minimize false positives?	?
How do you determine how much data to search ?	?
When you’re searching in a hostile environment, how do you search securely ?	?



INVESTIGATOR: OK, let's get started! We'll just bring those servers down, image the drives, and analyze offline

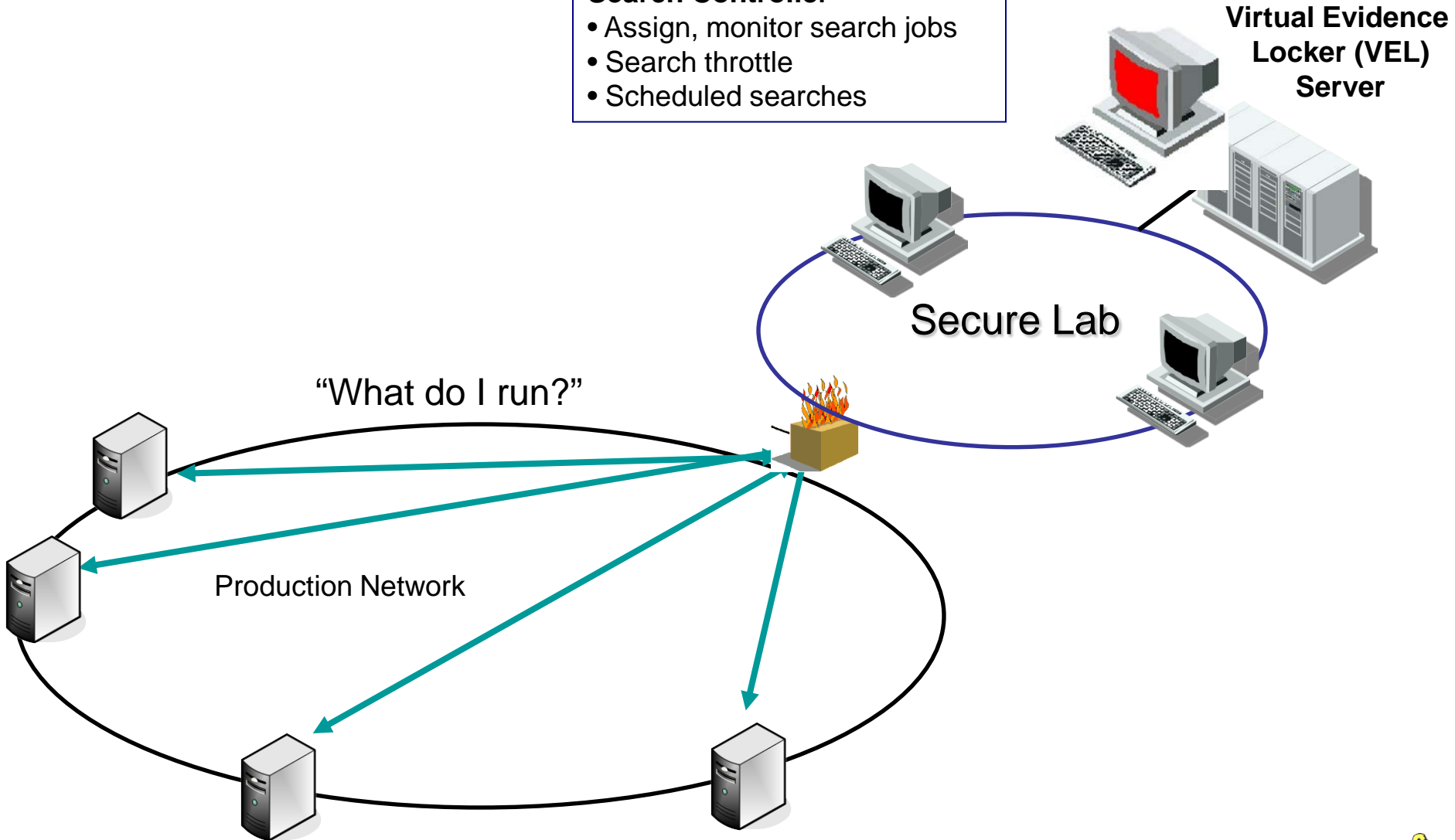
CLIENT: AAAAHHHH, NO!

Architecture

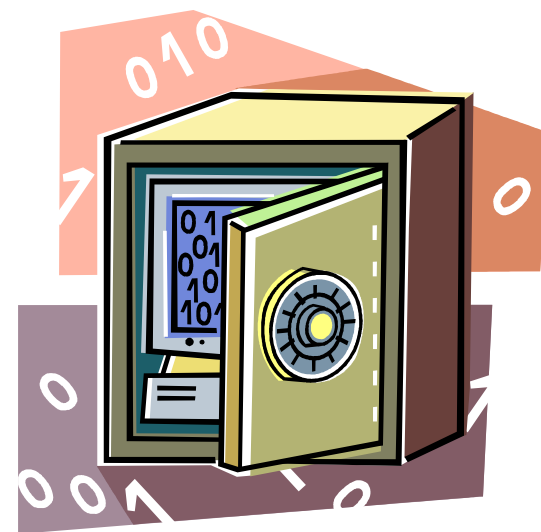
Search Controller

- Assign, monitor search jobs
- Search throttle
- Scheduled searches

Virtual Evidence Locker (VEL) Server



- Needed location to conduct all search operations
 - Schedule, monitor, and digitally sign jobs
 - Aggregate and analyze results
 - Harvest files of interest
- Server must be secure!
 - Encrypted disk partitions
 - Digitally Signed jobs (trusted programs)
 - Limited access, firewall
- Collaborate with team members
 - Secure private and group chat (silc)



It Takes Many Tools ...

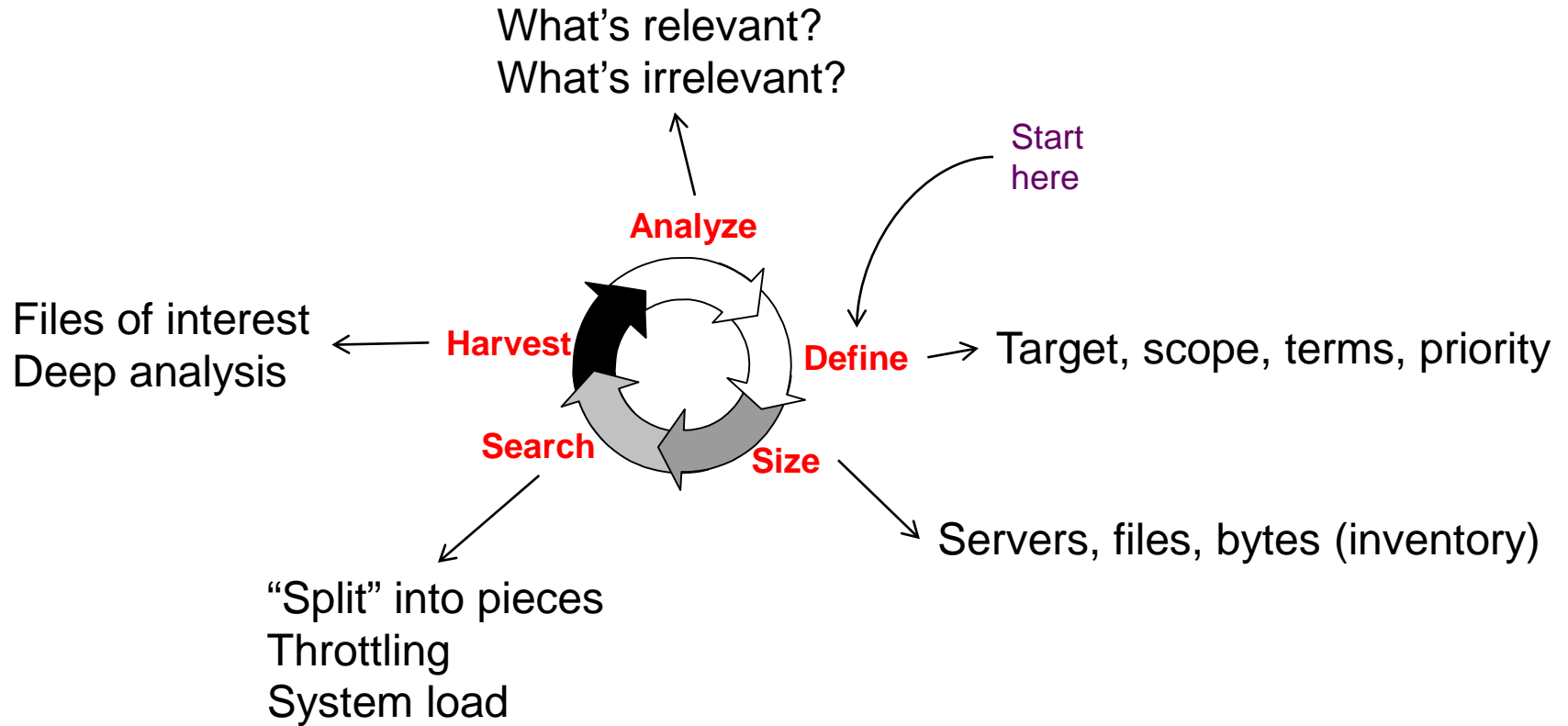
Requirements	Tools / Technologies
Generic framework to run jobs securely and collect information	WebJob, SSH
Tool to perform search	FTimes
Tool to validate results	Ftimes-check-ssn, ...
Database to store results and perform analysis	MySQL
High-performance, read-only remote file system mount	F-Secure
Malware sandbox to run unknown binaries	Malware sandnet
Utilities to convert container files to clear text for searching	Conversion utilities, various OSs and POUND
Encrypted partitions for sensitive data	FreeBSD geli, Linux loop AES
Ability to stop and start searching	Splits
And so on ...	



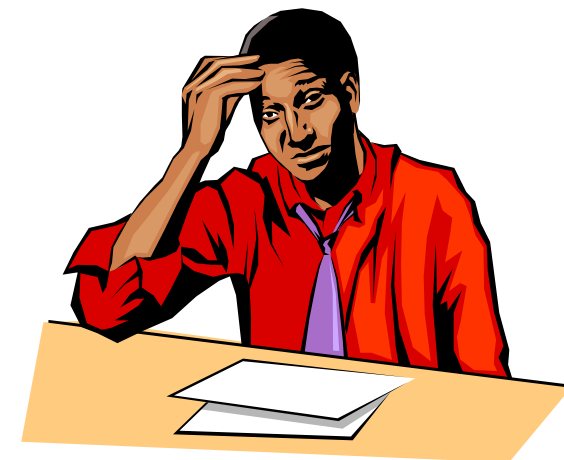
What to do ... What to do ...



Overview



- Questions
 - What is the goal of the search effort?
 - What systems / locations do we search?
 - What search terms do we use?
- It's simple, right? Just go search for the data ...
 - Scale presented problems
 - Hours/days to collect basic file metadata
 - Realized quickly must breakup massive search space
 - We must protect search terms, results, and harvested files throughout the process



Define: Search Terms

- Some things to consider
 - Literal vs. Regular Expressions (regex)
 - Literal faster than complex regex, but regex more flexible
 - False positive rate vs. data volume & validation/analysis time
 - High false positive rate may be acceptable if data volume and validation are manageable
 - All terms vs. sampling of terms
 - For quicker searches may need to use a sampling of terms to identify files for further examination
 - System load
 - Consider how searching will impact end system
- Benchmark search techniques
 - Could mean the difference between 2 weeks and 2 months



Define: Search Terms

- “Sweet spot”
 - “Best” terms to **minimize search & analysis** time while **maximizing hit confidence**
- Characteristics you want ...
 - Unique or long terms
 - Common delimiters (e.g., “1-2-3”, “1.2.3”, and “1 2 3”)
 - Unicode
- Characteristics you want to avoid ...
 - Short terms (<5 characters) such as “dog”
 - False-positive hits
- Additional difficulty ...
 - Container files (e.g. zip, xls, doc, jar, gz, tgz, ...)



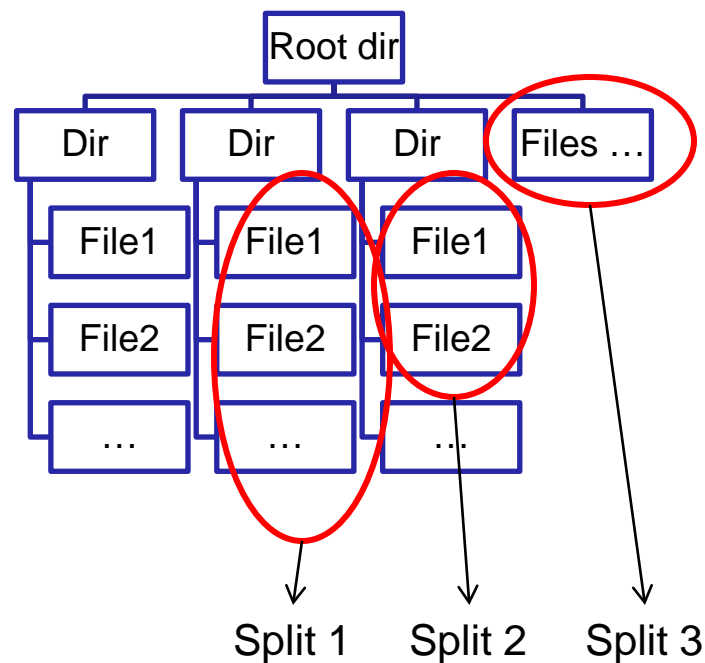
Size: Inventory

- Can't just search whole system
 - Production window
 - Failed searches
 - Failed network connections
- Collect file metadata (no hashes)
 - Relatively quick
 - Determine how big your search space is
 - Help in search estimation
 - Help prioritize files/locations to search
 - Determine optimal search size (“splits”)
- Be aware of file system drift (added, deleted, modified)



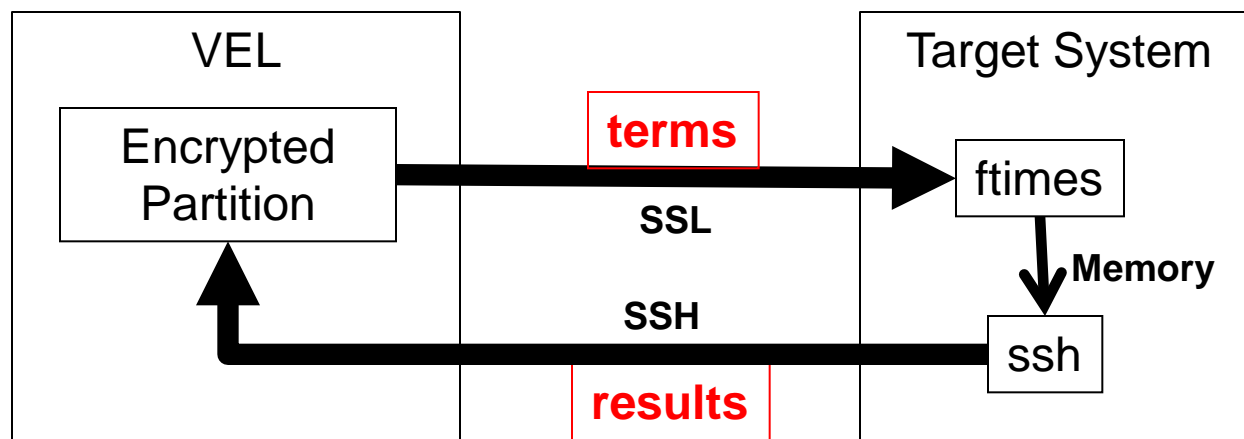
Size: Splits

- Broke file system search components into thousands of **mini-searches** a.k.a. “splits”
- Each split:
 - < ~1 hour
 - < 10,000 files
 - < 2GB
- Queued splits on VEL
- Enabled freezing/thawing queues
- Allowed for search throttling
- Added robustness & monitoring



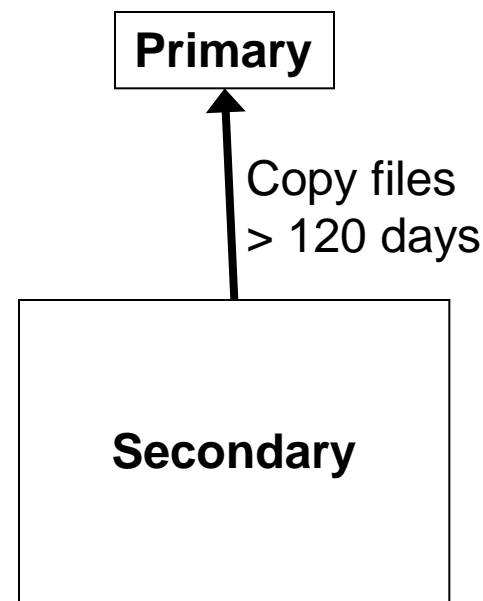
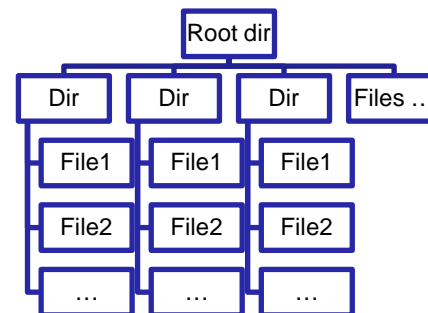
Search: Security

- Don't make the problem worse!
- Safe guard search terms:
 - Encrypted disk partitions
 - Search terms in memory only during searching
 - What if you are in a hostile environment?
- Search smartly
 - Exclude short search terms, large files



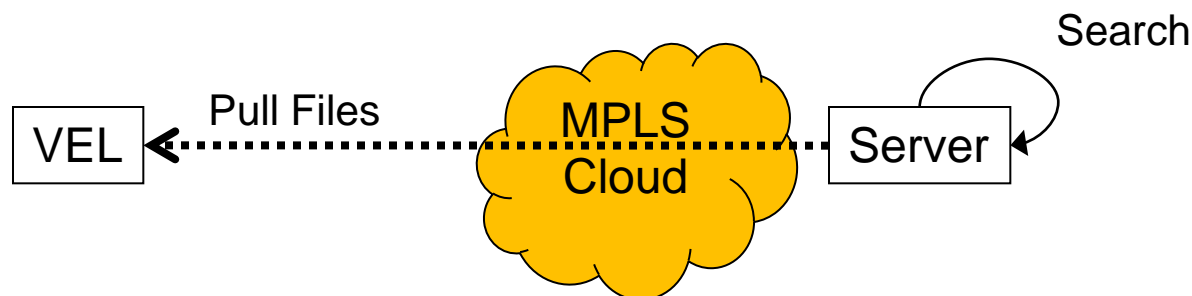
Search: Storage Migration

- Task: Search Windows file server
- Background:
 - File server connected to a storage migration service
 - Service composed of primary and secondary data stores
- Problem:
 - Requested files > 120 days copied from secondary to primary
 - Primary approach full utilization
- Solution:
 - Vendor modified registry to disable copying files > 120 days



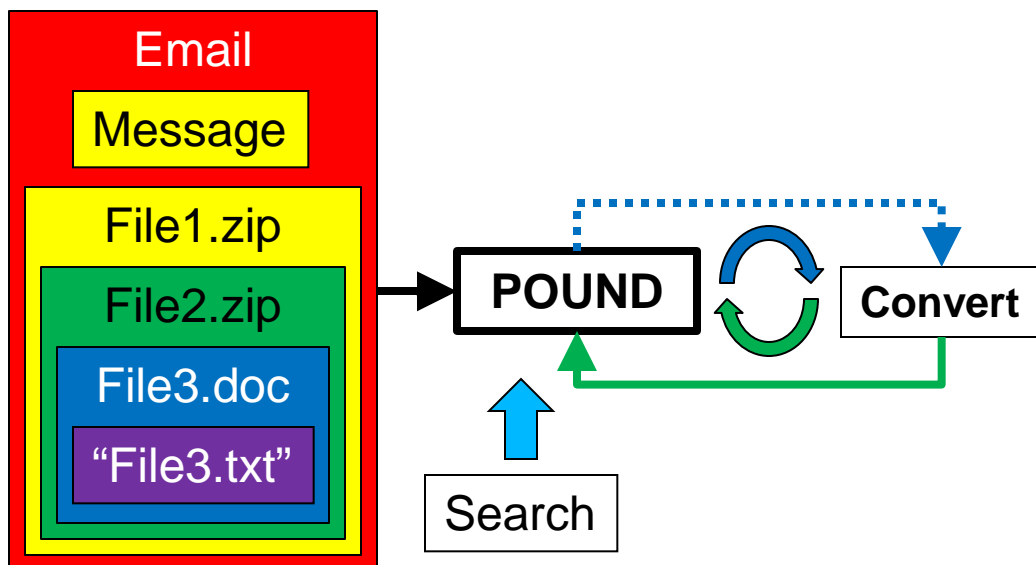
Search: Limited Bandwidth

- Client imposed bandwidth limitations across Multiprotocol Label Switching (MPLS) Cloud
- Client paying for high-performance and searching impact cost
- Would take weeks to do restore over MPLS cloud
- Modified approach to search locally and selectively pull files



Harvest

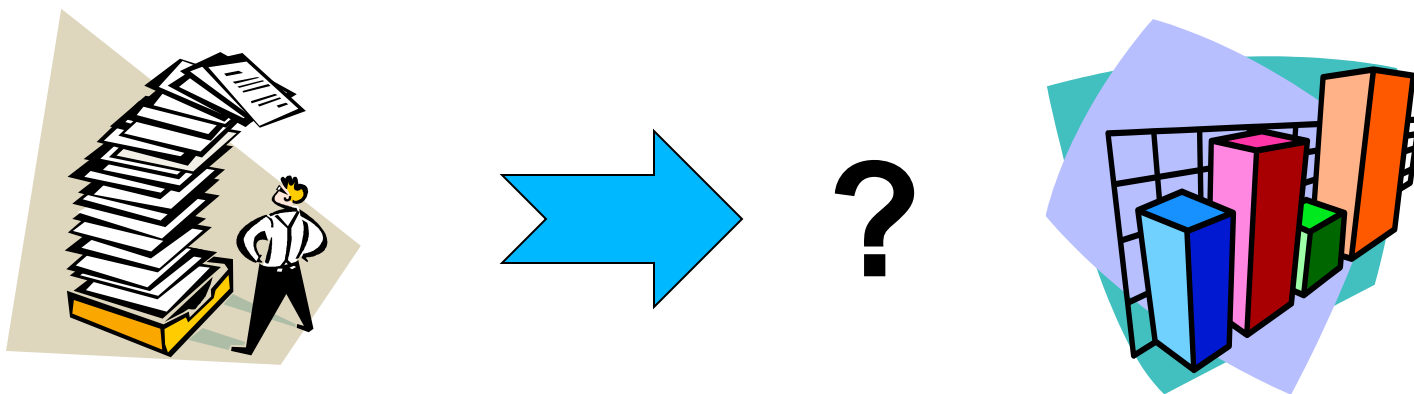
- We have hits, now what? Are they valid?
- Schedule jobs to bring files back to VEL for analysis and further searching
- **POUND = Pile Of Uniquely Named Data**
 - Store & unpack files for efficient searching
 - Password cracking ...



Cycle	Inserted into POUND
Start	Email
1	Message & Attachment (File1.zip)
2	File2.zip
3	File3.doc
4	"File3.txt"

Analyze

- The end of automation ...
- Can't just hand over millions of results
- Have we found what we are looking for?
- Does this look legitimate or is this a valid hit?
- How do we convey the results?



Analyze: Reduce the Pile

- Proximity: A+B+C all within X bytes of each other
- Validate results (Some examples)
 - Social Security Numbers (SSNs) ***
 - Developed tool to validate based on issuance numbers
 - Helps distinguish between other 9-digit numbers (CUSIPs)
 - AAA-GG-SSSS
 - AAA = Area number (e.g., 212-220 is for Maryland)
 - GG = Group number (O 01-09, E 10-98, E 02-08, O 11-99)
 - SSSS = 0001 -> 9999
 - Credit Card Numbers (CCNs)
 - Applied LUHN Test



Analyze: Hits, Files, Servers

System	Hits	Files	Servers
Server 1	1	1	1
Server 2	0	0	0
Server 3	35	19	1

User	Hits	Files	Servers
User 1	7	3	1
User 2	20	3	2
User 3	0	0	0

File Type	Hits	Files	Servers
Script	0	0	0
Source	9	9	1
Log	3	1	1

- Slice-and-dice search results based on system, user, file type, etc.
- Helpful for identifying objects of interest



Analyze: Hotspots

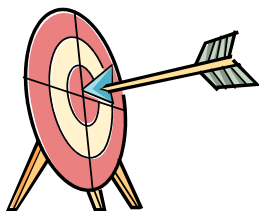
Search Sets

	1	2	3
Server 1	50%	70%	80%
Server 2	70%	80%	90%
Server 3	80%	90%	95%

Search Sets

	1	2	3
File 1	50%	70%	80%
File 2	70%	80%	90%
File 3	80%	90%	95%

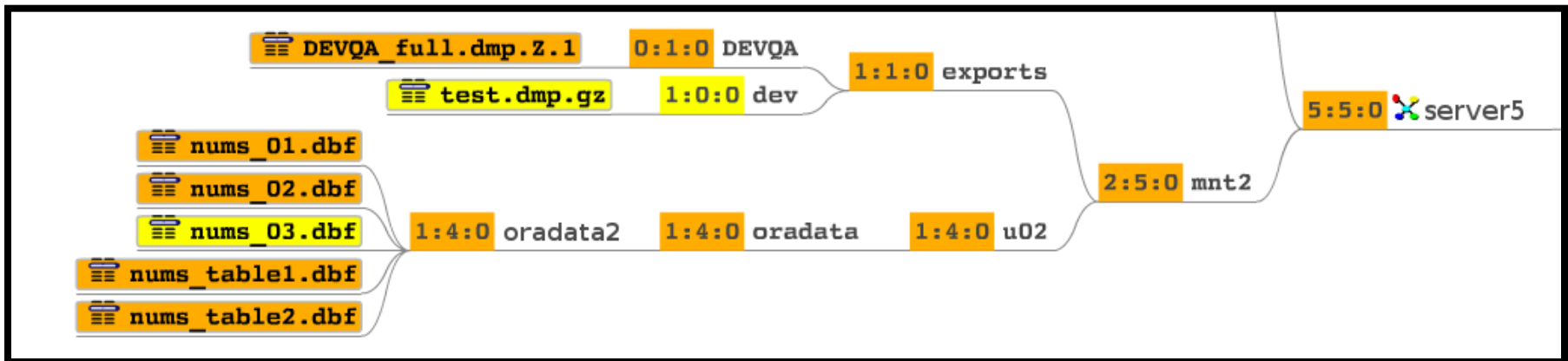
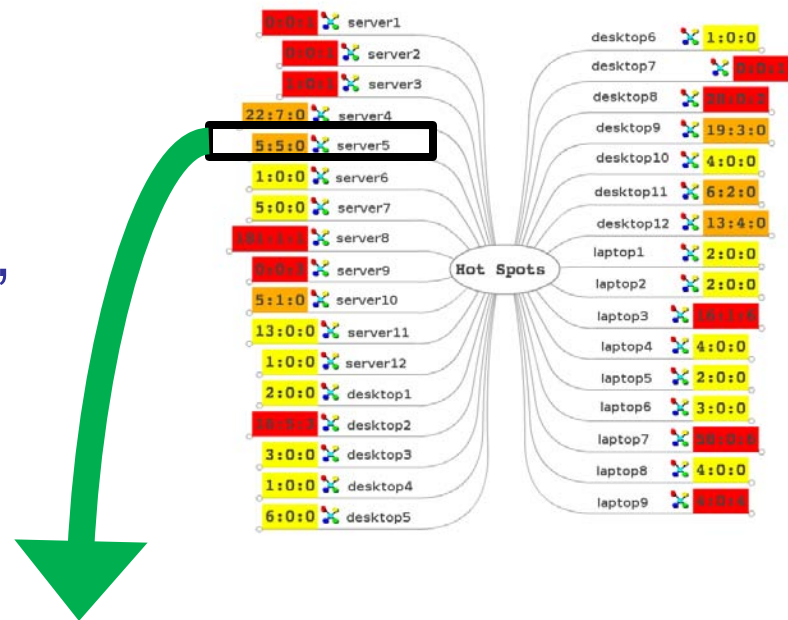
- Columns: Sets of related search terms
- Rows: Servers/Files of interest
- Cells: “Scores” = Percentage of set search terms in a server/file
- Created XLS files with colorized backgrounds based on score
- Servers and files 2 and 3 should be examined first



Analyze: Mind Maps

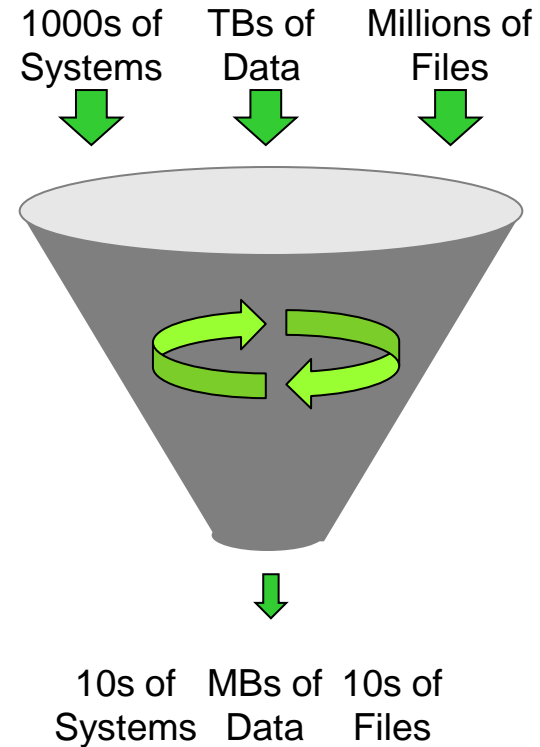
Background
Approach
Process
Define
Size
Search
Harvest
Analyze
Conclusions

- Data Visualization
- Quickly drill down thru search hits
- Colorization and “roll-up”
Low/Medium/High



Things to Think About

- Automate as much as you can but manual analysis may still be necessary
- How do you convey technical results to non-technical people?
- How much analysis do you do before delivering results?
- Delivering improperly analyzed results could waste time



- Lessons Learned
 - Large-scale breaks conventional techniques
 - Secure sensitive data through whole process
- Expect unexpected
 - Manage limitations – bandwidth, system load, production times
 - Be able to improvise & create own tools
 - Will require multiple tools to get the job done
- Be enterprise search ready
 - Conduct dry-run with key personnel to identify readiness & gaps
 - Link this to incident response and E-discovery readiness
 - Conduct search project in lab environment to refine strategy

Question	KoreLogic Approach
How do you secure search terms, search results, and analysis?	Secure Lab, VEL
How do you limit impact on production serves?	Splits
How do you pause searching during critical business hours?	VEL, Splits
How do you distill results into something meaningful?	HFS, Hotspots, Mind Maps
How do you develop search terms?	Benchmarking, Experience
How do you search a massive dataset in a timely fashion ?	WebJob search architecture
How do you handle “ container ” files?	Harvest and pound
How do you minimize false positives?	Search terms
How do you determine how much data to search ?	Client & inventory
When you’re searching in a hostile environment, how do you search securely ?	WebJob, SSH, encryption

Questions and Discussion ...

