A New Approach to Digital Forensic Methodology

And !!BUSTED!! Case studies

David C. Smith

Samuel Petreski

Introductions

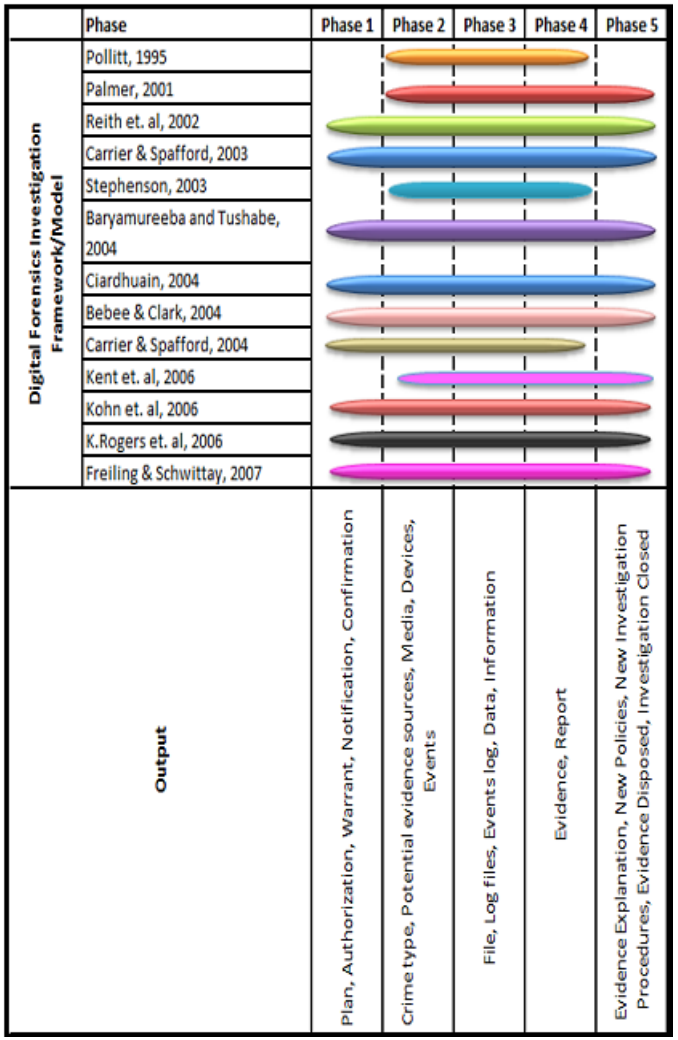
- David C. Smith, Georgetown University, & HCP Forensic Services
  - David works as the CSO for Georgetown University and a co-owner of HCP Forensic Services providing information security programs, digital forensics, and expert witness testimony. He has been in the technical field for over 20 years and enjoys engaging in complex technical problems.
- Samuel Petreski, Georgetown University
  - Samuel Petreski works as a Senior Security Analyst for Georgetown University and an owner of Remote IT Consulting. Samuel has worked mostly in higher-ed focusing on network architecture and administration, as well as building and administering scalable network security solutions. He possesses over 10 years of experience in the IT field working in very diverse environments.

The IDEA

- Read “Mapping Process of Digital Forensic Investigation Frameworks” – Selamat, Yusof, and Sahib [IJCSNS Vol 8 No 10, Oct 2008]
- Thought: Lots of methodologies out there, but none were what I “taught” or saw as issues when running a forensic team.
- Why not make my processes and methods into a new, practical methodology?

Mapping Process of Digital Forensic Investigation Framework

No	Digital Forensic Investigation Framework	No of Phases
1	Computer Forensic Process (M.Pollitt, 1995)	4 processes
2	Generic Investigative Process (Palmer, 2001)	7 classes
3	Abstract Model of the Digital Forensic Procedure (Reith, Carr, & Gunsch, 2002)	9 components
4	An Integrated Digital Investigation Process (Carrier & Spafford, 2003)	17 phases
5	End-to-End Digital Investigation (Stephenson, 2003)	9 steps
6	Enhance Integrated Digital Investigation Process (Baryamureeba & Tushabe, 2004)	21 phases
7	Extended Model of Cybercrime Investigations (Ciardhuain, 2004)	13 activities
8	Hierarchical, Objective-based Framework (Beebe & Clark, 2004)	6 phases
9	Event-based Digital Forensic Investigation Framework (Carrier & Spafford, 2004)	16 phases
10	Forensic Process (Kent K. , Chevalier, Grance, & Dang, 2006)	4 processes
11	Investigation Framework (Kohn, Eloff, & Oliver, 2006)	3stages
12	Computer Forensics Field Triage Process Model (K.Rogers, Goldman, Mislan, Wedge, & Debrota, 2006)	4 phases
13	Investigative Process Model (Freiling & Schwittay, 2007)	4 phases



Mapping Process of Digital Forensic Investigation Framework, Selamat, Yusof, Sahib, 2008

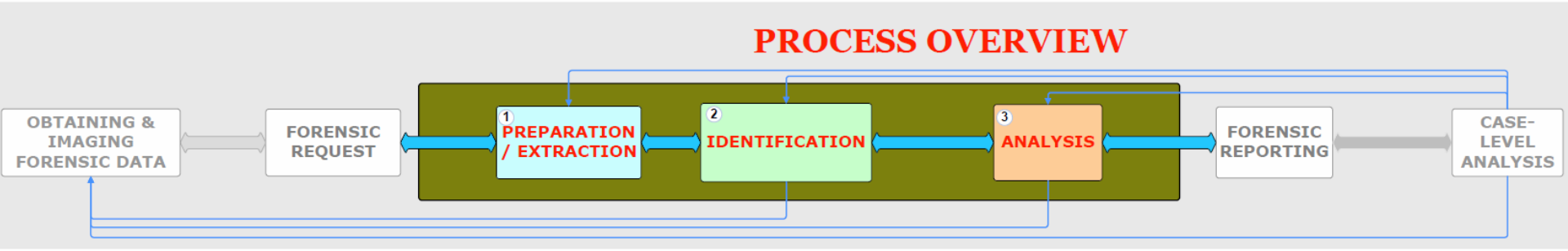
Typical Digital Investigation Methodologies

- Most frameworks cover the range from acquisition to reporting
- Such as:
  - Obtaining authorization for investigations
  - Determining evidence locations
  - Determining and validating techniques to find and interpret significant data
  - Summarize and provide explanation of conclusions

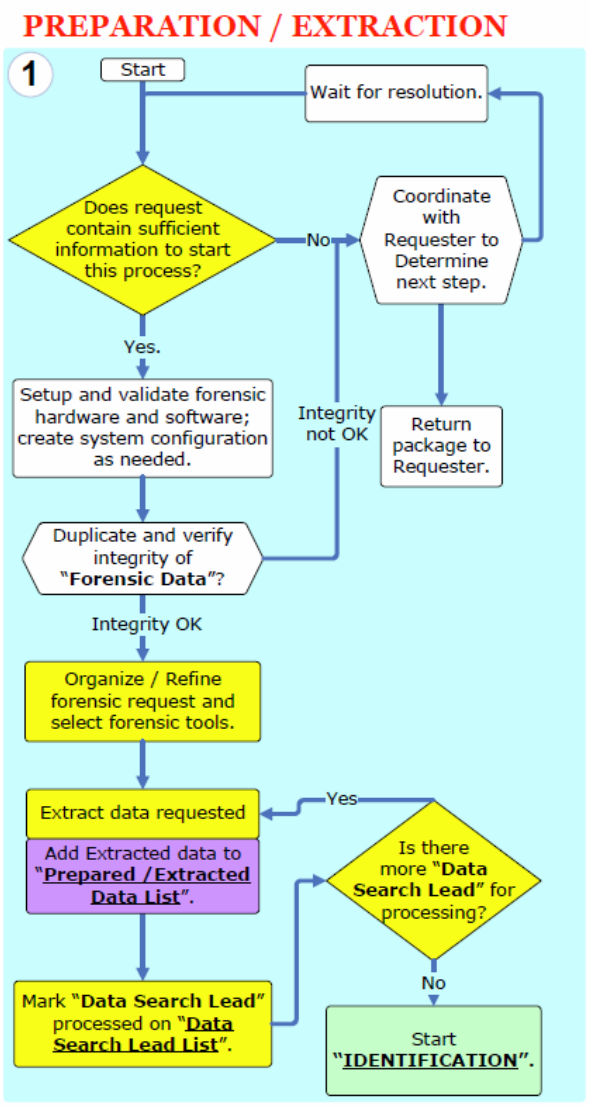
What is a Digital Forensic Methodology?

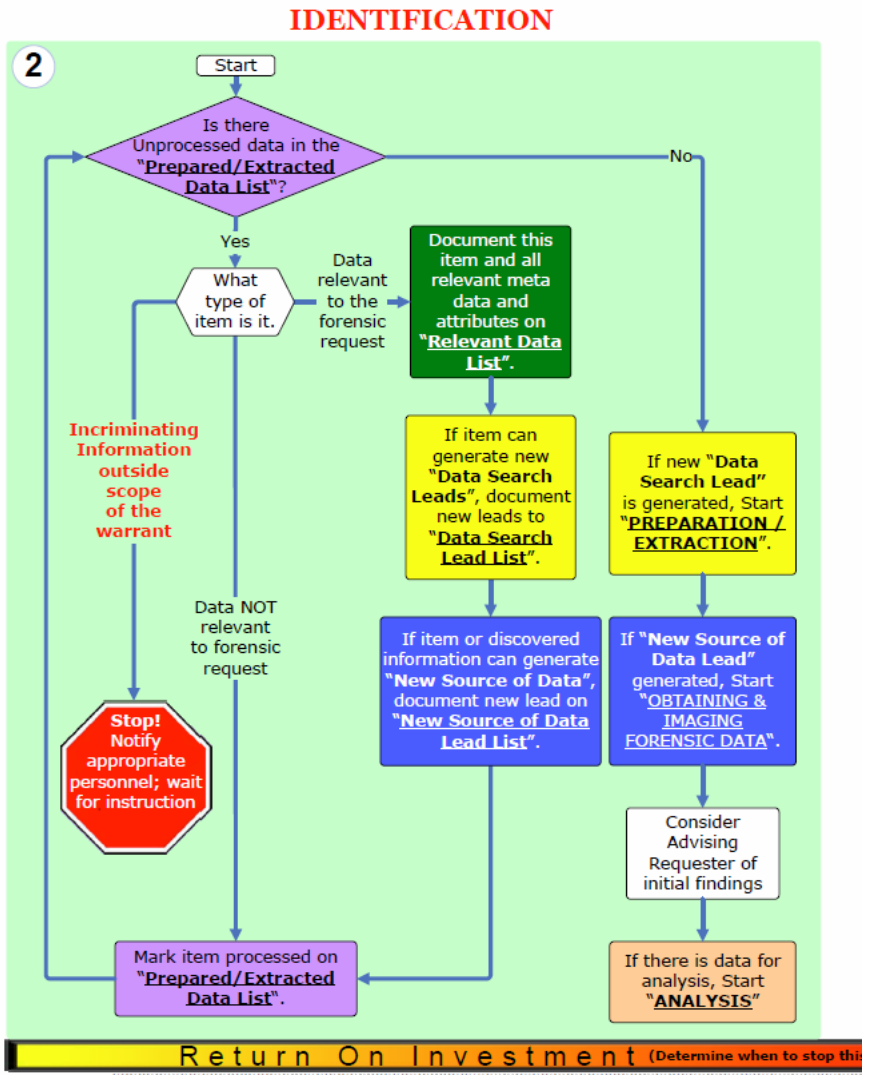
DIGITAL FORENSIC ANALYSIS METHODOLOGY

Last Updated: August 22, 2007

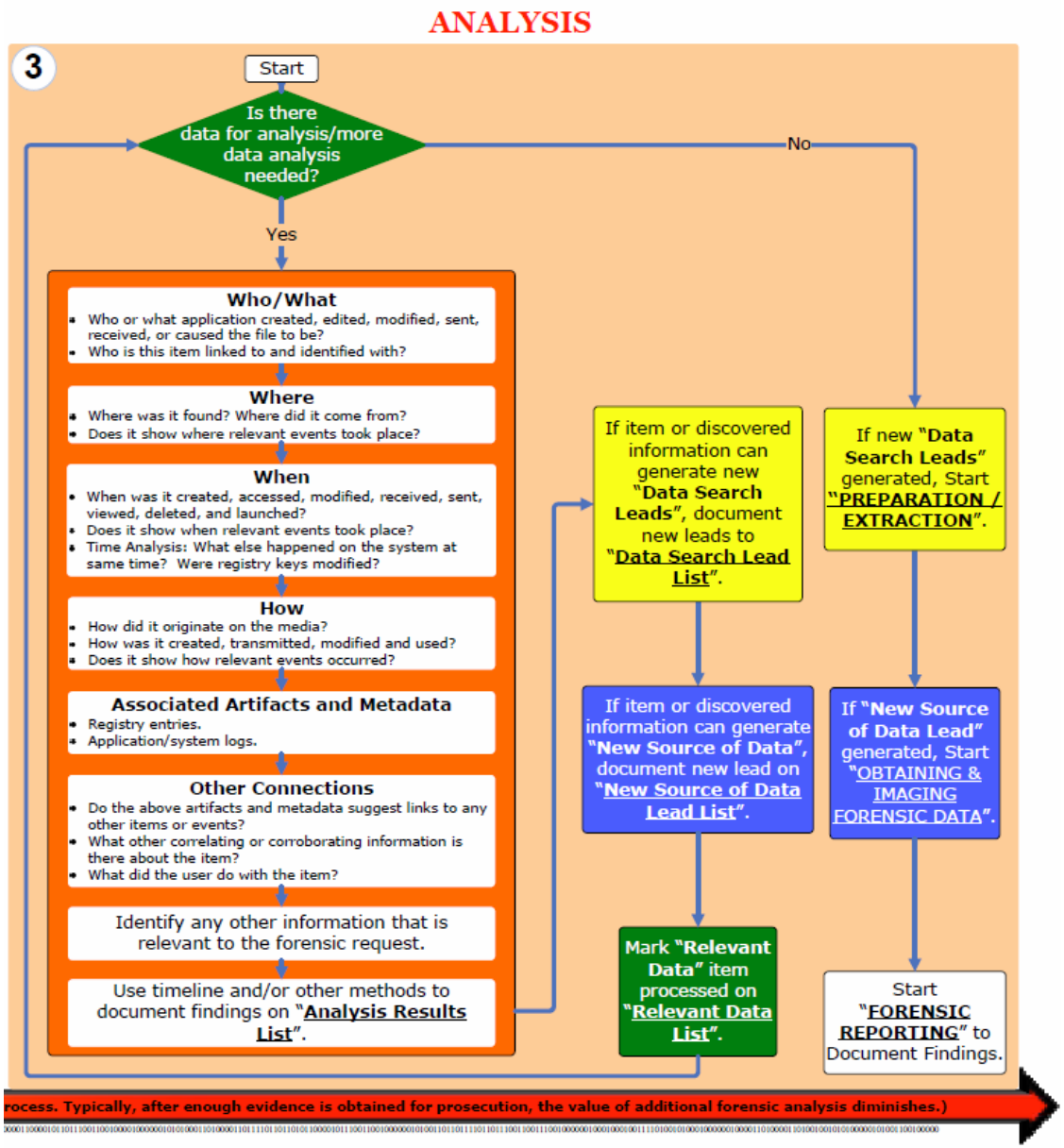


DOJ Methodology (1)





DOJ Methodology (2)



Stages / Processes / Phases

- There are some really good methodologies out there
- Integrated Digital Investigation process (IDIP), Digital Forensic Research Workshop (DFRWS)
  - Identification, preservation, examination, analysis, presentation, and decision
- Enhanced IDIP includes a “Dynamite” Phase

Integrated Digital Investigation Process, Carrier & Spafford, 2003
Enhanced Integrated Digital Investigation Process, Baryamureeba & Tushabe, 2004

Introduction to the “Problem”

- Problems with learning and performing digital Forensic Investigations
  - Open solution set, many ways to find or approximate the “answer”
  - A lot of self-teaching & “sit and do it”
  - Patience, learning to “stay on target”, and having to learn new techniques while performing an investigation
- All of these things improve over time as an analyst gains experience

Open Solution Set

- Last cup of coffee[1]
  - You arrive in the break room and find 5 individuals drinking coffee and the pot empty. You want to determine who drank the last cup.
  - How many ways can you determine who drank the last cup?

[1] B. Carrier, A Brief Introduction To the Computer History Model, 2008

Determine Who Drank the Last Cup!

- Measure the amount of coffee in each cup
- Measure temperature of each coffee
- Measure strength of each coffee (stronger on bottom of pot?)
- Amount of coffee grounds in each cup
- Interview individuals, analysis for truthfulness
- Interview group, analysis for truthfulness
- Develop timeline for coffee drinkers (internal and external)
- Measure the temperature of the cup (heat loss) vs. the temperature of the coffee

Who Drank the Last Cup!

- A little off the wall…
  - Gain a history of known and previous convicted last cup takers
  - Coffee on breath
  - Offer reward to rat them out!
  - Dust for the fresh fingerprints
  - Are there cameras in break room? Hallways?
  - Interview of last trip to the bathroom, hold everyone until they have to go

The Point is

- Is there a combination of methods that produces a higher probability answer?
- To be efficient the investigator needs to choose the optimal method(s) to draw conclusions
- This is what experts in the field do from experience and instinct

Thought Experiment

- 3 digital forensic analysts of different skill levels are give an identical assignment
  - Allowed to interact with requestor
  - Requested to develop an estimate of time
  - Off they go…
  - Performance based on total findings, time to process, and estimation of time.
- Based on the way we do things now, what results could we expect?

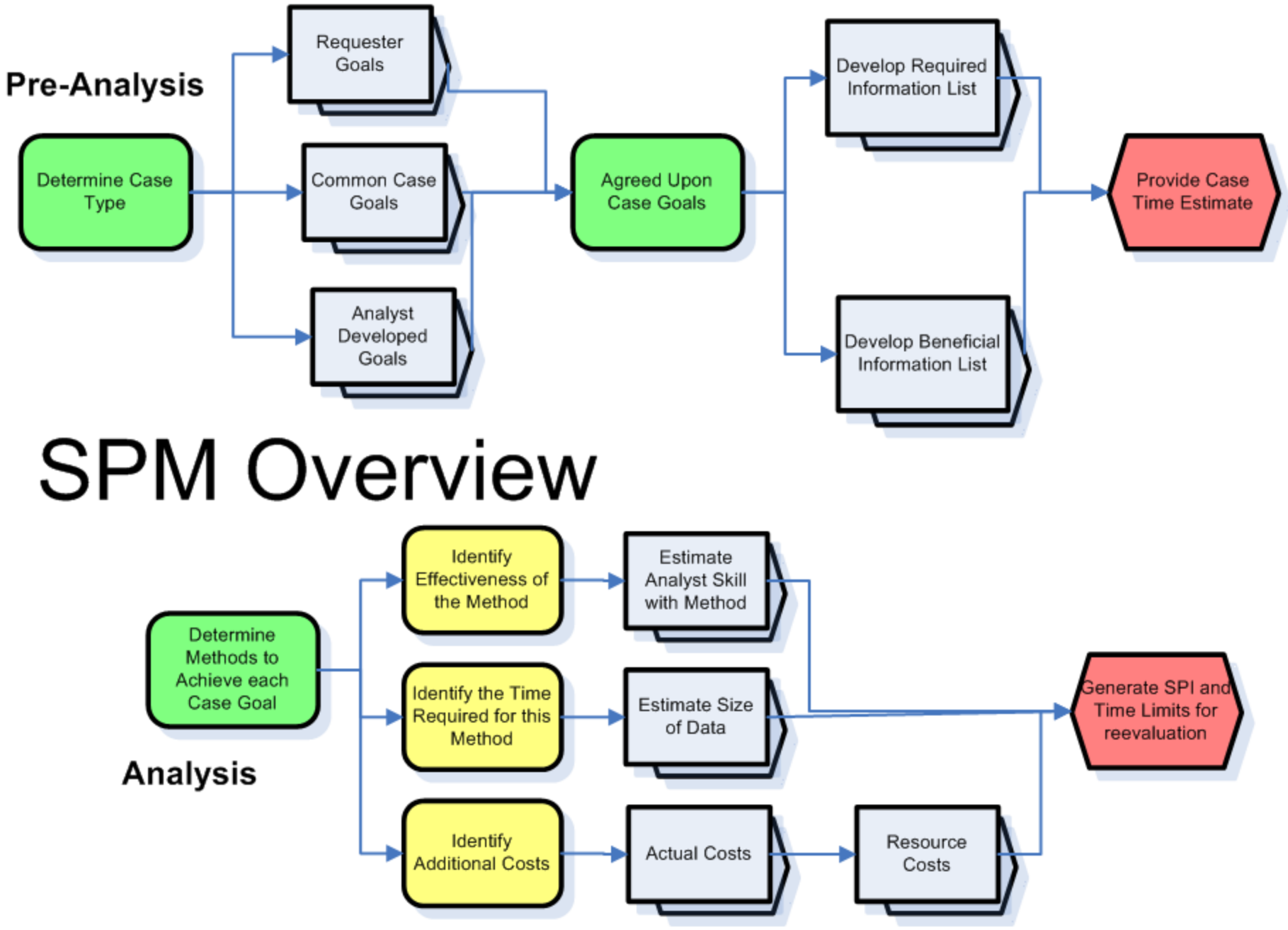
Thought Experiment (2)

- What if we limited it to 20 hours?
  - Reduced findings?
  - More varied results?
- What about 8 hours?
  - Partial results?
  - Experts only?

Questions

- In the ANALYSIS phase of your favorite comprehensive digital forensics methodology
  - How do we do a better job of maximizing our time with the requester?
  - How do we do a better job of estimating the time it takes to solve open solution set problems?
  - How do we optimize the methods we use to develop conclusions for the case goals?
  - Can we achieve consistent results in the field?

SPM Overview



Smith-Petreski Methodology

- SPM Details
  - Developed for the analysis phase of digital investigations
  - Organized by the classification of case types
  - Development of goals by case type
  - Evaluation and quantification of methods to determine optimal paths
  - Implementation of a time management framework.
  - Part expert system with processes to better develop case goals, identify ideal methods, and set time goals

Smith-Petreski Methodology (2)

- Methodology Goals
  - Better development of pre-analysis information
  - Achieve better estimation of investigation required
  - Optimize time to achieve case goals
  - Provide more consistent results from teams of digital forensic investigators
  - Provide a framework to predict analysis time, resources, and costs

Introduction to the Methodology

- Three Components
  - Pre-analysis
    - Defined case types with in-depth descriptions, common cases goals, typical goals for each case type, and case type requirements
  - Analysis
    - Selection of optimal methods to achieve case goals
  - Structured time management
    - Recommended allocation of time based on methods, case time given, and allows for the re-evaluation of methods based on results

Pre-Analysis

- Two basic request methods
  - Meet with the requester
    - Determination of what the requester wants or believes to want
    - Opportunity to fine-tune the agreed upon goals
  - Request form based
    - Less interaction means more detailed forms or requests
    - Larger shops typically require more complex procedures and processes to maintain the same value in digital forensic analysis

Pre-Analysis (2)

- Sources of case goals
  - Direct and derive case goals from the initial request
    - Find out how this machine was compromised (requester)
    - Determine what the attacker did (analyst)
  - Common goals based on case time
    - Determine the vulnerability or exploit; use this information to identify what other systems may be compromised or at risk
  - Case goals generated by analyst
    - Could be anything, but an example is that the attacker searched for “Star Wars Systems,” so a follow-up case goal would be to identify documents related to “Star Wars Systems”

Pre-Analysis (3)

- SPM also includes a structure to determine what information should be collected during requester meeting based upon the case type
  - Required information
    - Hard drive of compromised workstation
    - Logs from other systems
  - Beneficial information
    - Network packet captures
    - Known vulnerabilities

Pre-Analysis (4)

- Improved guidance for estimating processing time based upon the case goals and type
  - The primary data points are case size, skill level of examiner, and resources available.
  - Our determination for the case type “Malicious Activity” with fairly standard goals is 4.2 methods with a 20% overhead of total time
- Again, we consider this to be consistent with internal dialog that experts use in the field
  - I can normally solve this case type using methods w,x,y and sometimes z, but I need 4 hours to import, 16 hours to process, and 2 hours for reporting – plus some buffer…

Case Goal Estimation Time

- Generated by specific case type and the number of goals
  - Generated by case type and the number of case goal agreed upon
    - 1 to 3 goals, Malicious Activity case type, 12 hours + process time
    - 4-6 goals, Malicious Activity case type, 18 hours + process time
  - This tries to replicate expert internal dialog

Analysis

- Now it is time to sit in front of the computer with your tools…
- Goal: Achieve case goals in an optimal time frame
- Smith Petreski Index (SPI) is an algorithm to assist in determining method or methods with the highest probability of achieving case goals

SPI Algorithm

- SPI is generated using the following data points
  - Effectiveness, how likely will this method achieve your goal in a percentage
  - Level of effort / resources, estimated time to perform this method based on small, medium, large estimates
  - Compatibility of toolsets, the amount of time in minutes to adjust, purchase, or install the prerequisites for this method
  - Familiarity with method and toolset base on descriptions of novice, experienced, and expert (in this toolset)

Smith Petreski Index (SPI) DataFields

- For Methods
  - Short Description
  - Long Description
  - Base effectiveness to case goal (Novice, Experienced, Expert)
  - Analysis time in minutes for dataset size (small, med, large)
  - Machine time in minutes for dataset size (small, med, large)
  - Additional costs are “converted” to minutes to adjust methods that require a purchase, additional set-up time, or resources
- From this SPI and total time are derived

Goal of Generating SPI

- Choosing methods that produce the “best bang for the buck” to solve case goals
- We’ve developed software to provide the hard values, estimates, initial method sets, and to generate the SPI
- What we mean by “methods”
  - Specifically not tool based
  - Description such as “Generate Web Histories”
  - We don’t want to lock in to specific tools or operating systems

SPI Algorithm

- Probability based
  - Measures effectiveness of a method balanced against how long it takes execute that method in terms of both person and machine time, as well as additional costs.
  - Function
    - $f(x) = \log_2(1/1\text{-effectiveness}) * \text{Inflator} – (\text{machineTime} + 2 * \text{personTime} + 1.5 * \text{additionalCost})$
  - Excel / Open Office / Google Apps
    - =LOG((1/(1-effectiveness)),2)*1000-((machineTime)+(2*(personTime))+(1.5*additionalCost))

Couldn’t address everything with SPI

- Additional considerations
  - Willingness and ability to purchase additional tools
  - Specific expertise and skills of the analyst
    - A scripting heavy method may be more effective for an expert scripter than the SPI predicts
  - Type of environments needed for specific methods
    - Such as mobile examination or a windows only shop

SPI vs. Expert

- Again, the expert has experience with success and failures of methods, missed deadlines, and empirical data on the processing time required for various methods.
- Determining the “best bang for the buck” has become second nature
- Already has an intuitive understanding of the best methods for the specific case goals

Framework for Structured Analysis Time

- Two factors in time estimation component of SPI:
  - Data size, e.g. web history small is under 1000 relevant records @ 1 hour
  - Skill level, with a choice of novice, experienced, and expert for each method
- Provides the ability to budget time based on expected results
  - If time exceeds the estimate by 20%, then this should force a revaluation of the method used.
- Provides a systematic time management strategy unique to the case

Case Studies using SPM

- You've made it past the dry methodology, so hopefully this is more entertaining
- Case studies are made up of real cases, sanitized and cleared by our lawyers

- Should represent the value of SPM as we walk through the phases of the case

Intellectual Property

- Case Background
  - Employee left and started a competing business
  - Employee hire dates and “last date”
  - Employee was assigned workstation
- Case Type
  - Intellectual Property case type includes analysis of systems, media, and network traffic for the use and misuse of proprietary data and is usually associated to the identification and verifications of documents, ideas, and concepts of the requesting organization. While it is not the analyst’s responsibility to interpret laws that determine unfair business practices or the violation of regulations, the analyst may be required to make the associations of proprietary information and derivative work. This case type typically includes keyword searches for key terms, system use analysis, and discovery of method that may have been used to transfer information. Intellectual property case types can also include external sources in conjunction with protective orders and analysis of similar work products that may be derivative of other work.

Meeting with Requester

- Initial Meeting
  - Requester wants to know if any business protected information was taken
    - Specifically contacts and vendor lists
- SPM Common Case Goals for Intellectual Property
  - Identification of specific documents
  - Identification of specific parts of documents
  - Identification of system use based around documents or time
  - Identification of external transfer methods, such as USB drives or network uploading.
  - Identification of documents based on keyword searches for ideas, concepts, and known terms
  - Validation and opinion of derivative work

Agreement of Goals

- Agreement of goals
  - Emails to and from identified contacts or mail domains
  - Identification of USB devices that have attached to the system
  - Identify system usage for selected time periods
    - Link files, registry files, timelines of use
  - Locate all copies of selected documents
    - Both full copies and selected parts of documents
  - Identify documents based on keywords
    - Keywords provided by requester

Analyst's Potential Additional Goals

- Extract instant message logs
- Recover deleted files
- Memory Analysis
- Convert identified persons of interest into common usernames (instant massage, personal email account, etc)

Case Information

- Based on Case Goals
  - Required information
    - Keywords / mail domains for email analysis
    - Keywords for document identification
    - Documents to be located
    - Copies of documents to be searched for
    - System Images
  - Beneficial Information
    - Full case background or timeline of events
    - Work-product names / external associated names
    - Specific dates and times

Pre-Analysis Time Estimation

- 5 Goals, removing some duplication
  - Email analysis
  - Registry analysis
  - Identification of files
  - Extract files for analysis, recover delete files
  - Identify system usage
- Pre-analysis estimate of 28.5 hours required

Common Methods

- Intellectual Property Common Methods include
  - Hash files for matches
  - Fuzzy hash for partial matches
  - Extract files from container files
  - Extract mail
  - Registry analysis for system usage
  - Registry timeline
  - System usage timeline (super timeline, log files)
  - USB analysis
  - Network PCAP analysis
  - Extract metadata
  - Recover deleted files
  - Keyword index and analysis
  - Extract IM
  - OCR graphic formats for text indexing

Methods to Goals IP Case

- Case Goals -> Methods
  - Extraction of emails for analysis
    - SPI: 3,222 estimated time 80 minutes
  - Hash files for identification and location
    - SPI: 2,457 estimated time 260 minutes
  - Fuzzy Hash files for identification and location
    - SPI: 2,643 estimated time 280 minutes
  - Recover deleted files
    - SPI 3,052 estimated time 225 minutes

Methods to Goals IP Case

- Case Goals -> Methods
  - Identification of system usage
    - Registry analysis, SPI 2,707 @ est 45 minutes
    - Super Timeline Analysis, SPI 2,257 @ est 5 hours
    - Link File analysis, SPI 1,395 @ est 75 minutes
    - Web History analysis, SPI 2,527 @ est 115 minutes
    - Analysis of IM / Carve IM logs, SPI 1,410 @ est 60 minutes

So, what did we find?

- Hashing and filename search results as expected
  - Located the identified documents in emails, on internal and external drives, in LNK files, hits in registry for recent
  - Hash match lead to a zip file with the name “needed for XXXXXXXX.zip” – name of the new competitor
  - Fuzzy hashing found slightly altered copies, including copies with the new competitor’s name and letter head. This lead to directories that contained slightly changed to updated overhauls of company processes and procedures

Email Extraction

- Extracted and processed email was interesting
  - Used original keywords and associated names to develop a dictionary of all individuals associated, abbreviations of the new competitor, and locations from the email threads / IM logs / web mail
  - Developed a timeline that was amazing, from initial contact, follow-up, offer sent in FedEx, last- minute negotiations, discussions of exit strategies and how to approach difficult questions, and status of remaining days before the “last day”
  - Web mail artifacts included discussions of pros and cons with a significant other, purchase of equipment for a new home office on the negotiated “work from home day”.

Deleted Files and Keywords

- Processed deleted files and performed keyword searches.
  - Updated dictionary with IM usernames, personal email addresses, and associated derivatives
  - Mediocre free space results, it is always difficult for me to justify using free space either for searching or to corroborate results from other methods
  - Keyword documents did not generate any follow-up searches or additional analysis

System Usage

- Registry analysis usually has a great SPI!
  - Generated reports on all registries, including restore points.
  - Tons of supporting data for accessing files
  - Tied LNK files to USB drives, showed transfer to external USB keys
- Super Timeline Analysis
  - Didn't exist at the time, but would have a good SPI based on high effectiveness, more machine time than analyst time, and low cost.
  - We didn't have that, so we broke out the sources that were most relevant and custom scripted a merge.

System Usage (2)

- Analyzed web history, recovered deleted histories, rendered HTML cache files.
  - Good approach because it showed a lot of activity not in the interest of the organization
  - A little porn… NOT little people porn!
- Recovered IM chats and carved deleted chats
  - Great conversations trashing the organization, key individuals, and friends
  - Discussions of who to attempt to “take” to the new company

Conclusions

- Met all of the requester's goals
  - Had defendable data and conclusions
  - Rechecked primary findings with multiple tools
  - Happy client, no follow-up required
- Personal Conclusions
  - This guy fills out the ID10T forms in triplicate
  - Does he own a home PC? CCleaner? Eraser?
  - Truecrypt? Zip 8.0 AES encryption?

Judging Your Performance

- Feedback can be shaky sometimes, based on how well you found the answers that were wanted by the requester
- I use the following metrics in weighted order
  - The number of follow-up questions from the data in your report, i.e. how well they understood your presentation of the findings
  - Number of goals the requester really wanted that you were unable to draw out of them
  - Amount of estimated time vs. total time (adjusted for unusual circumstances)
  - Total predicted value vs. actual value to the requester
  - Number of “wrong turns” or undisciplined searching

Presentation Conclusions

- Even if you don't fully invest into the methodology, you will still gain from
  - Defining better case goals with your requester
  - Improved familiarity with common goals of your primary case types
  - Mentally organizing methods with a "best bang for the buck" mentality
  - Developing internal time management for reevaluation of your methods to achieve your case goals

Questions

- Q & A

- Thanks to Kyle Davis, Mickey Lasky, Scott Moulton, and everyone else that contributed to the development of this methodology

- David Smith
  - Email: dcsmith@hcp-fs.com
  - Blog: http://dcinfosec.blogspot.com/
- Samuel Petreski –
  - Email: samuel@petreski.com

Forensic Thoughts

- I like building dictionaries of account names, email addresses, and additional keywords from examinations.
  - Allows an overall priority for additional searching
  - Reduces the temptation to get lost in unguided and unfocused searching
- Keeping the case goals/SPI/common methods handy
  - I like to print them out and scribble status and notes as I go
  - Helps prevent case goal “over-kill” and optimize efforts
- After a couple of cases that had reporting deadlines, I now include raw data as appendixes