2ND ANNUAL CONFERENCE

The Sleuth Kit
& Open Source Digital Forensics Conference

The Sleuth Kit

June 13, 2011 TUTORIALS / June 14, 2011 CONFERENCE
HILTON MCLEAN TYSONS CORNER, MCLEAN, VA

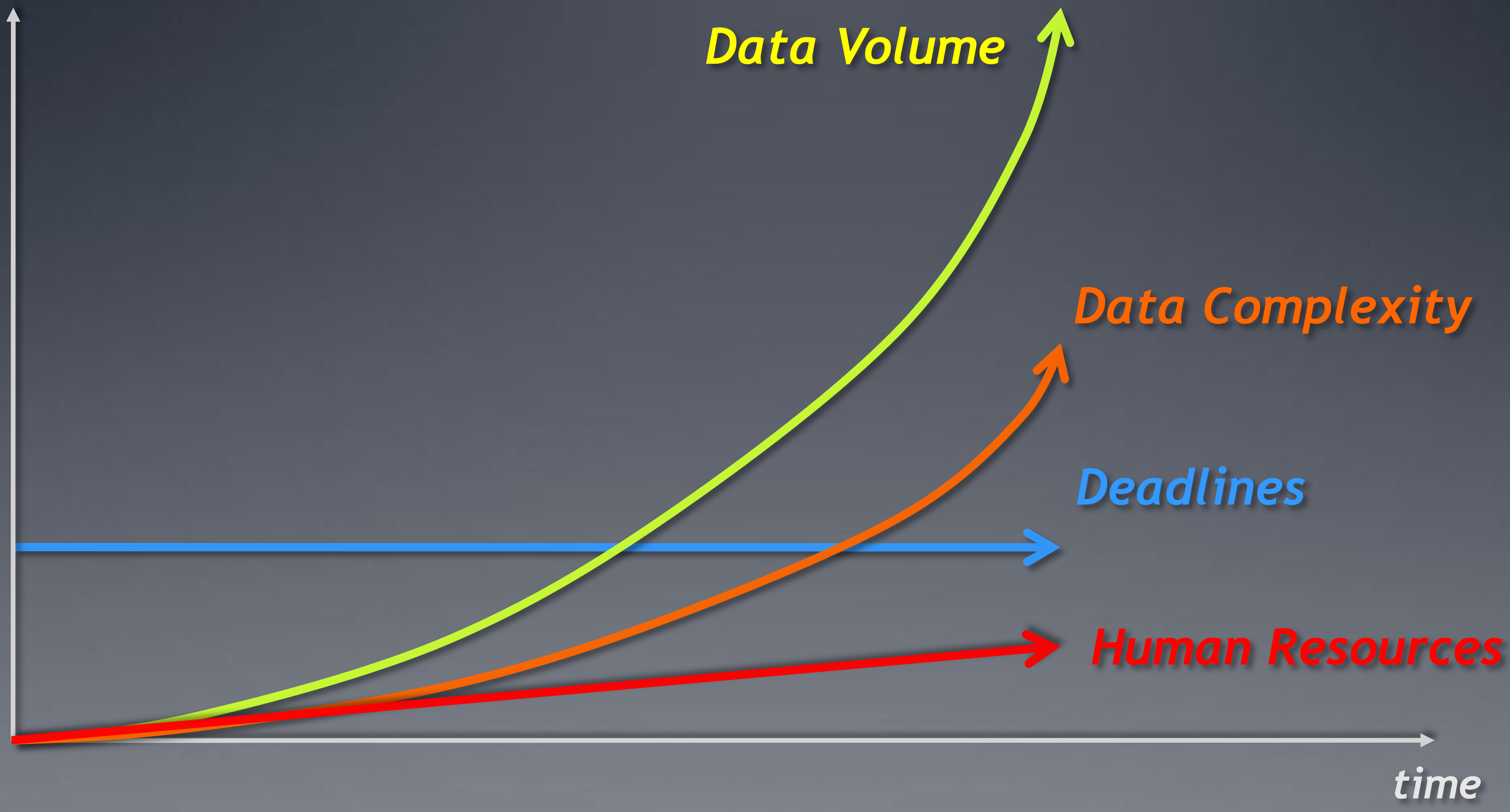
The Gorilla Approach to Scaling & Integrating Open Source Forensic Tools: *Learning from the Web*

THE UNIVERSITY of NEW ORLEANS

Vassil Roussev

vassil@cs.uno.edu

Trends in Forensic Analysis

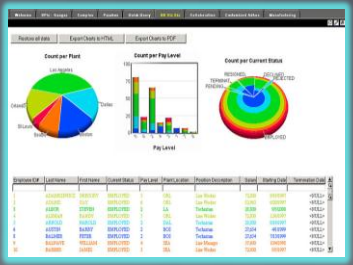


The 4-way Scalability Challenge

Data Scalability

Extensibility

Cost Scalability



UI Scalability

Data Scalability Now

- http://accessdata.com/distributed-processing:

  “**Impressive Test Results!**

  In testing, AccessData fully processed a massive data set, including 62,649,383 items, [...] The compressed size of this data set was 1.28 terabytes. [...] However with AccessData’s distributed processing technology, **it only took 6 days, 5 hours.**”

1.28TB zip ~ 3TB raw

3TB / 129hrs = 23.5MB/s

$$T_{proc} = 5.25 \times T_{HDDclone} \text{ (@130MB/s)}$$

$$T_{proc} = 25.5 \times T_{SSDclone} \text{ (@600MB/s)}$$

Cost Scalability Now

Commercial

- Integrates licensed components
  - Oracle, dtSearch, …
  - ➔ There is a price floor
- Per-CPU pricing
  - 2x CPUs → 2x $$$

➔ User cannot keep up on a fixed budget.

Open Source

- Nominally free, BUT
  - Tools poorly integrated
  - Tool chain incomplete
- Extra cost/resources
  - Higher tech expertise
  - Development/integration costs
  - Testing/validation

➔ No budget guarantees.

Extensibility Now

Commercial

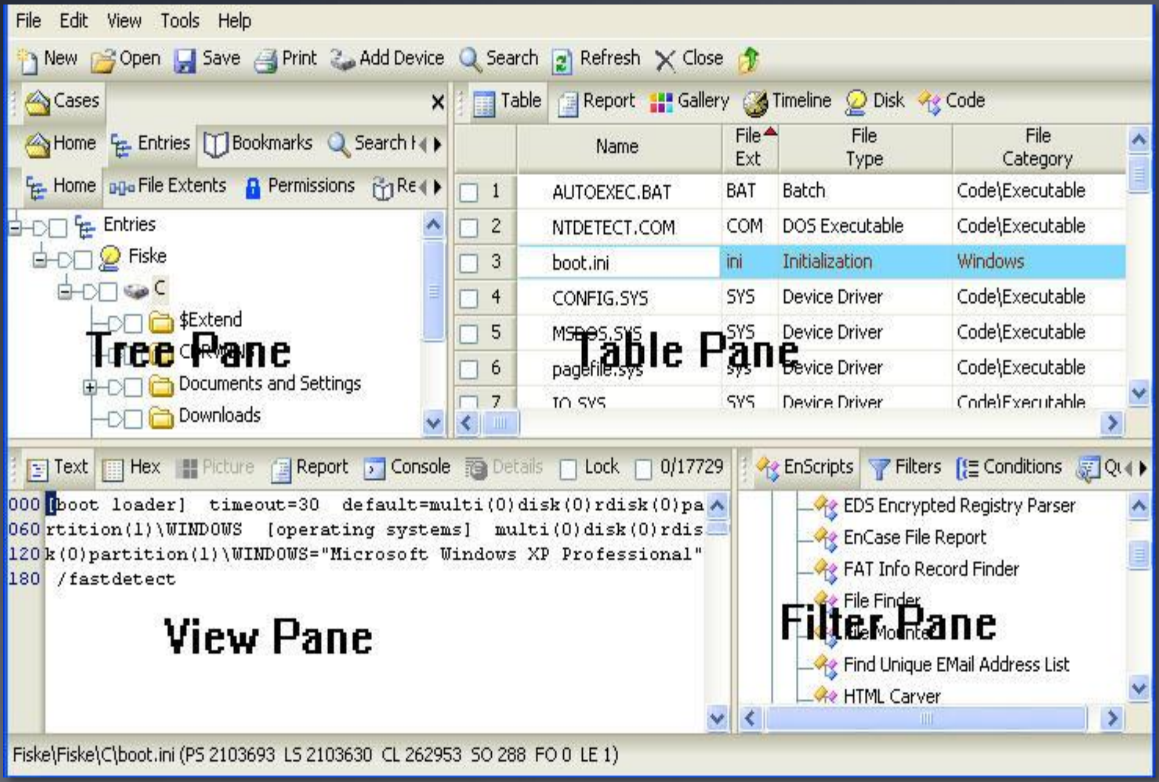
- Some custom scripting/DLL integration
  - E.g., EnCase
- ***Not*** a platform
- How do you extend a black box?

  … and test it?
- No incentive for change.

Open Source

- Extensibility unlimited, BUT integration is a ***huge*** burden
  - No common platform
  - No common data store
  - Only a few language-centric APIs
  - Integration = scripting
- Need a new approach.

“Analytical” UI Now


File Edit View Tools Help
New Open Save Print Add Device Search Refresh Close
Cases
Home Entries Bookmarks Search I
Home File Extents Permissions Re
Entries
Fiske
C
$Extend
Documents and Settings
Downloads
Tree Pane
Table Report Gallery Timeline Disk Code
Table Pane
Text Hex Picture Report Console Details Lock 0/17729
View Pane
EnScripts Filters Conditions Q
EDS Encrypted Registry Parser
EnCase File Report
FAT Info Record Finder
File Finder
File Mounter
Find Unique EMail Address List
HTML Carver
Filter Pane

	Name	File Ext	File Type	File Category
□ 1	AUTOEXEC.BAT	BAT	Batch	Code\Executable
□ 2	NTDETECT.COM	COM	DOS Executable	Code\Executable
□ 3	boot.ini	ini	Initialization	Windows
□ 4	CONFIG.SYS	SYS	Device Driver	Code\Executable
□ 5	MSDOS.SYS	SYS	Device Driver	Code\Executable
□ 6	pagefile.sys	sys	Device Driver	Code\Executable
□ 7	IO.SYS	SYS	Device Driver	Code\Executable

```
000 [boot loader]  timeout=30  default=multi(0)disk(0)rdisk(0)pa
060 rtition(1)\WINDOWS  [operating systems]  multi(0)disk(0)rdis
120 k(0)partition(1)\WINDOWS="Microsoft Windows XP Professional"
180  /fastdetect
```

Fiske\Fiske\C\boot.ini (PS 2103693 LS 2103630 CL 262953 SO 288 FO 0 LE 1)

- It’s WIMP world …
  - EnCase, FTK, X-Ways, pyFLAG, …
- Does NOT scale
  - More data ➔ more UI data
  - Offloads problem to user
- Does NOT support cognitive process

The 80/20 Rule

At least 80% of forensic processing is NOT forensic-specific.

- So far, we act as if the opposite is true!
  - Forensics is a niche market, this is not sustainable.
- Should we
  a) Continue on the current path, or
  b) Look around for solutions from other areas?

Lessons from the Internet (1)

- Data scalability: ACID vs. BASE
  - BASE scales much better
    - Think Google, Amazon, Facebook, Twitter, etc.
  - ACID is expensive: ~20x slower
  - Forensics does NOT need ACID
    - By definition, all processing must be repeatable!
- Cost scalability:
  - Build an common infrastructure; add proprietary components on top
  - We can use the same data stores that Big Data companies do. FREE!

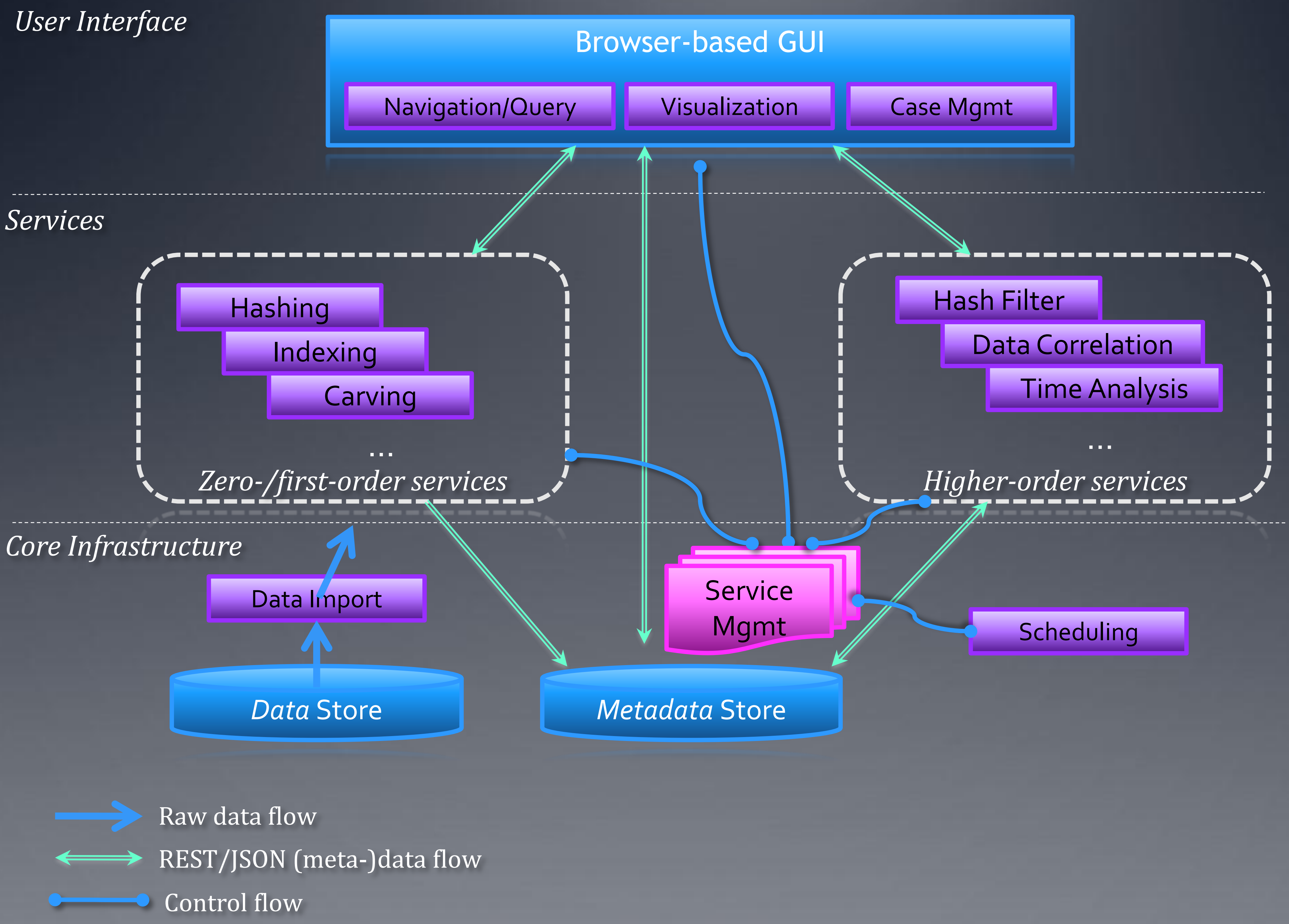
Lessons from the Internet (2)

- Extensibility
  - Simple data-exchange protocols work
    - JSON, Thrift, ProtoBuff, …
  - … XML doesn’t (too much overhead)
  - The one-size-fits-all model is falling apart
    - Performance/scalability
    - Impedance mismatch
  - Schemaless data models fit better than relational ones
    - Hadoop, MongoDB, Cassandra, CouchDB, Redis, Hbase, …

Lessons from the Internet (3)

- UI extensibility
  - Big data is driving data analytics and visualization
    - Plenty of open tools are available ***now***
  - The browser is replacing the desktop; wholesale!
  - Standards won't need proprietary extensions (Flash, Silverlight)
    - HTML5, CSS, JS, WebGL
  - Do we really need "home grown" WIMP interfaces for forensics?

A Solution Sketch



User Interface

Browser-based GUI

Navigation/Query

Visualization

Case Mgmt

Services

Hashing

Indexing

Carving

…

Zero-/first-order services

Hash Filter

Data Correlation

Time Analysis

…

Higher-order services

Core Infrastructure

Data Import

Service Mgmt

Scheduling

Data Store

Metadata Store

→ Raw data flow

↔ REST/JSON (meta-)data flow

●—● Control flow

Quick Demo

Summary

- We need an open and scalable forensics infrastructure to facilitate:
  - Development, instruction, & field work;
  - Research, testing, and validation.
- Current approaches do not work:
  - Commercial: fragmented, expensive, myopic
  - O/S: fragmented, incomplete, not ready for prime time
- We should look to the Web for Big Data answers
  - 80% of the forensics is not unique
  - We share problems/requirements
  - New, robust technology is freely available
  - Need to adopt web-centric standards

Thank You!

- Q & A
- Contact
  - Vassil Roussev
  - vassil@cs.uno.edu
- Come to New Orleans/DFRWS ‘11 (Jul 31—Aug 3)
  - dfrws.org