A Survey of Visualization Systems for Malware Analysis

Abdulrahman Alshammari

March 14, 2017

University of Delaware

Abstract

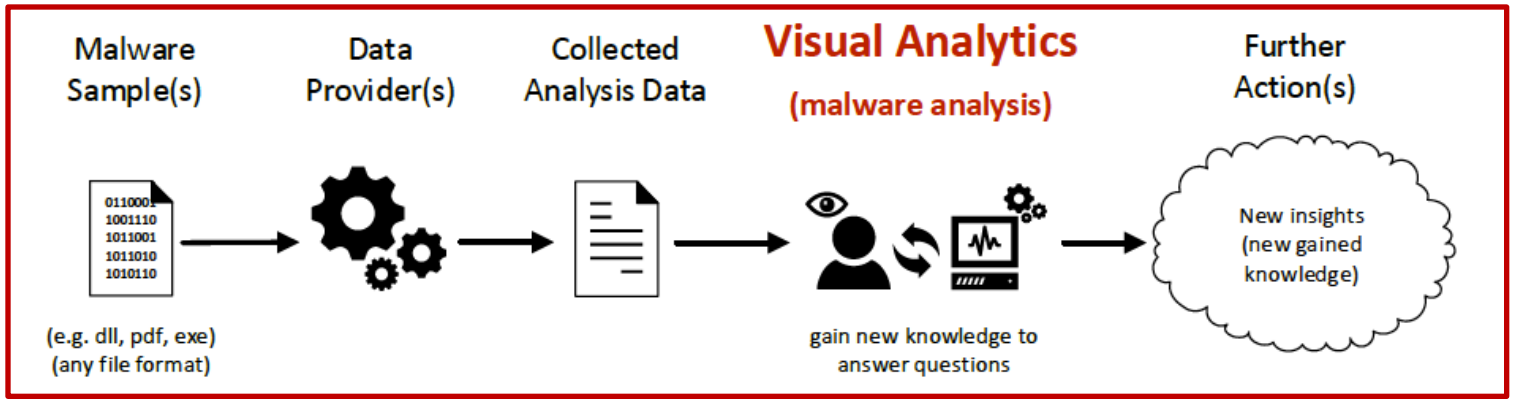
- Problem : increasing of malware
- Automatic approaches for malware detection
- The paper proposed systematic overview and categorization of malware visualization systems
- Review on Data Providers

Introduction

- What is Malware?
- Malware Analysis.
- In the third quarter of 2014 alone, 20 million new samples were discovered
  - Automated data analysis is not enough.

Introduction

- Patterns and categorize Malware.
- Fact: Visualization speeds up Malware detection.
- Definition of Visual analytics.



Related Work

- Beside this paper, no academic works support malware analysis from visualization perspective.
- Network and mobile security
- No detailed overview available in the field of visual analytics for malware analysis

Data Providers

- Definition of Data Providers
- Data Providers output => input for visualization tools.
- Static vs Dynamic analysis
- The need of Analysis Environment

Data Providers

- Base Data describes the type of data monitored and logged by a provider.
  - Virus definition
  - Packer information
  - File and header information
  - Library imports
  - CPU instruction
  - System and API calls
  - File system operations
  - Registry and network operation

Data Providers Examples

- Anubis : Automated dynamic analysis tool
- Cuckoo Sandbox: open source for automating the dynamic analysis
- CWSandbox: Dynamic Malware Analysis
- FireEye Malware Analysis System

Data Providers Examples

- Joe Sanbox: Dynamic Malware analysis
- Process Monitor: Free file system monitoring tool
- API Monitor: Free Tool
- Generic dissembler
- Generic debugger

	Anubis	Cuckoo	CWSandbox	FireEye MAS	Joe Sandbox	ProcMon	APIMon	Generic disassembler	Generic debugger
Analysis mode and environment									
Static analysis support		✓✓		✓	✓✓			✓✓	
Dynamic analysis support	✓✓	✓✓	✓✓	✓✓	✓✓	✓✓	✓✓		✓✓
Native analysis environment			✓✓		✓✓	✓✓	✓✓	✓✓	✓✓
Virtual machine environment		✓✓	✓✓	✓✓	✓✓	✓✓	✓✓	✓✓	✓✓
Emulation environment	✓✓					✓✓	✓✓	✓✓	✓✓
(Simulated) Internet access	✓✓	✓✓	✓✓	✓✓	✓✓	✓✓	✓✓		
(Simulated) LAN services	✓✓		✓✓			✓✓	✓✓		
Interface									
Command line interface	✓✓	✓✓		✓✓	✓				
Graphical (web) interface (GUI)	✓✓	✓✓	✓✓	✓✓	✓✓	✓✓	✓✓	✓✓	✓✓
Sample input									
Single file submission	✓✓	✓✓	✓✓	✓✓	✓✓		✓✓	✓✓	✓✓
Folder submission	(✓)	(✓)		✓✓	✓✓				
URL/URI	✓✓	✓✓	✓✓	✓✓	✓✓				
Batch processing	(✓)	(✓)	(✓)	✓✓	(✓)	(✓)	(✓)	(✓)	(✓)
Interactive on-demand analysis	✓					✓✓	✓✓	✓✓	✓✓
Supported input file formats									
Windows executables (.exe)	✓✓	✓✓	✓✓	✓✓	✓✓	✓✓	✓✓	✓✓	✓✓
Windows libraries (.dll)	✓✓	✓✓	✓✓	✓✓	✓✓		✓	✓✓	✓
Microsoft Office files	✓✓	✓✓	✓✓	✓✓	✓✓				
Portable document format (.pdf)	✓✓	✓✓	✓✓	✓✓	✓✓				
Malicious URL scan	✓✓	✓✓	✓✓	✓✓	✓✓				
PHP files (.php)					✓✓				
Java file (.jar)	✓✓	✓✓	✓✓	✓✓	✓✓			✓✓	
Visual Basic scripts (.vbs)					✓✓				
Image files (.jpg, .png,...)		✓✓		✓✓					
Video files (.wmv, .flv,...)	✓	✓		✓✓	✓				
ZIP archive (.zip)	✓✓	✓✓		✓✓	✓✓			✓✓	

	Anubis	Cuckoo	CWSandbox	FireEye MAS	Joe Sandbox	ProcMon	APIMon	Generic disassembler	Generic debugger
Base data									
Virus definition/Malware name	✓✓	✓✓	✓✓	✓✓					
Behavior classification	✓				✓✓				
Packer information					✓			✓	✓
File information/File header		✓✓						✓✓	✓
Library imports/loads	✓✓	✓✓	✓✓		✓✓		✓	✓✓	✓
CPU instructions/assembly					✓✓			✓✓	✓✓
API calls	✓	✓✓	✓	✓	✓	✓	✓✓	✓✓	✓✓
System calls	✓	✓✓	✓	✓	✓	✓	✓✓	✓✓	✓✓
File system operations	✓✓	✓✓	✓✓	✓✓	✓✓	✓✓	✓	✓	✓
Registry operations	✓✓	✓✓	✓✓	✓✓	✓✓	✓✓	✓	✓	✓
Process/thread information	✓✓	✓✓	✓✓	✓✓	✓✓	✓✓	✓	✓	✓
Network activity	✓✓	✓✓	✓✓	✓✓	✓✓	✓	✓	✓	✓
Report output									
PDF report	✓✓		✓✓						
HTML report	✓✓	✓✓	✓✓		✓✓				
XML report	✓✓		✓✓	✓✓	✓✓	✓✓			
TXT report	✓✓			✓✓			(✓)		✓✓
CSV report				✓✓		✓✓			
Native/Proprietary format						✓✓	✓✓	✓✓	✓✓
PCAP network dump	✓✓	✓✓	✓✓	✓✓	✓✓				
JSON report			✓✓	✓✓	✓✓				
Memory dumps		✓✓		✓✓	✓✓				
String dumps		✓✓			✓✓				
Screenshots		✓✓			✓✓				

Research Methodology

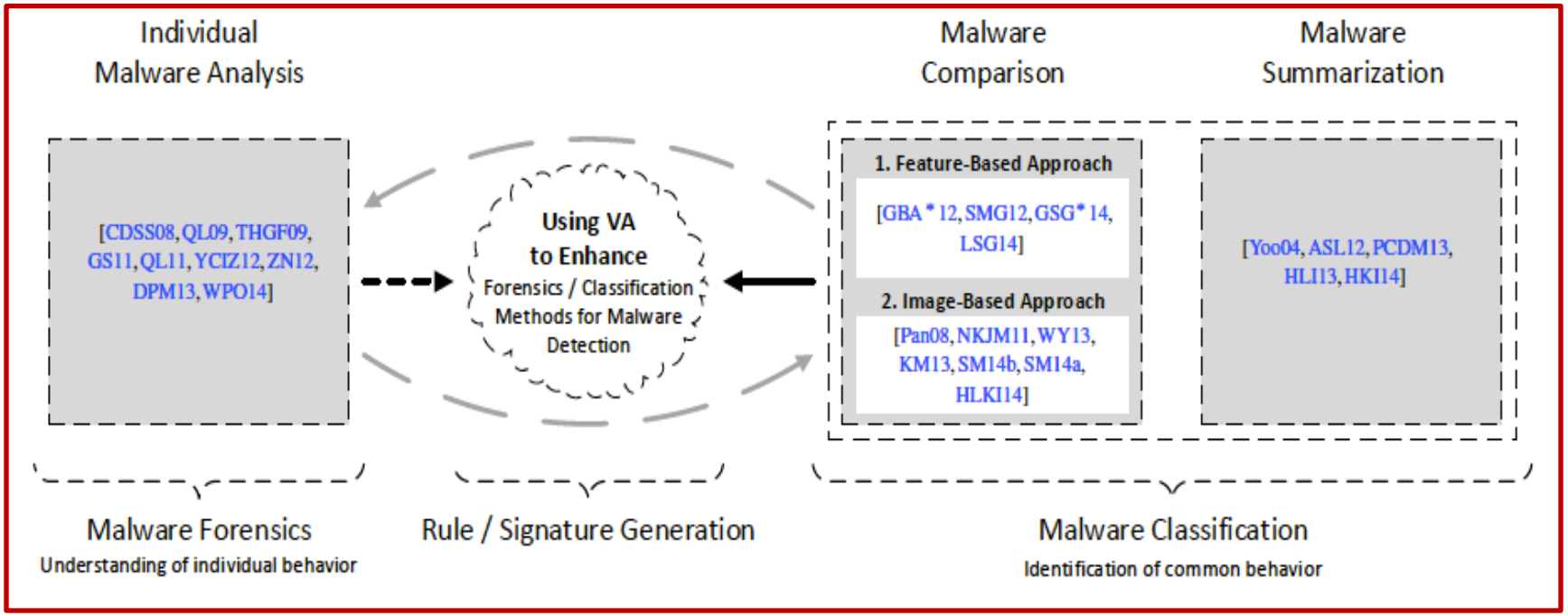
1. Used many digital libraries
2. Identified authors of most relevant papers
3. Published in *VizSec* (Visualization for Cyber Security)

- Result: 25 papers matching Malware Visualization Systems.
  - For Reference: (http://malware.dbvis.de/)

Visualization for Malware Analysis

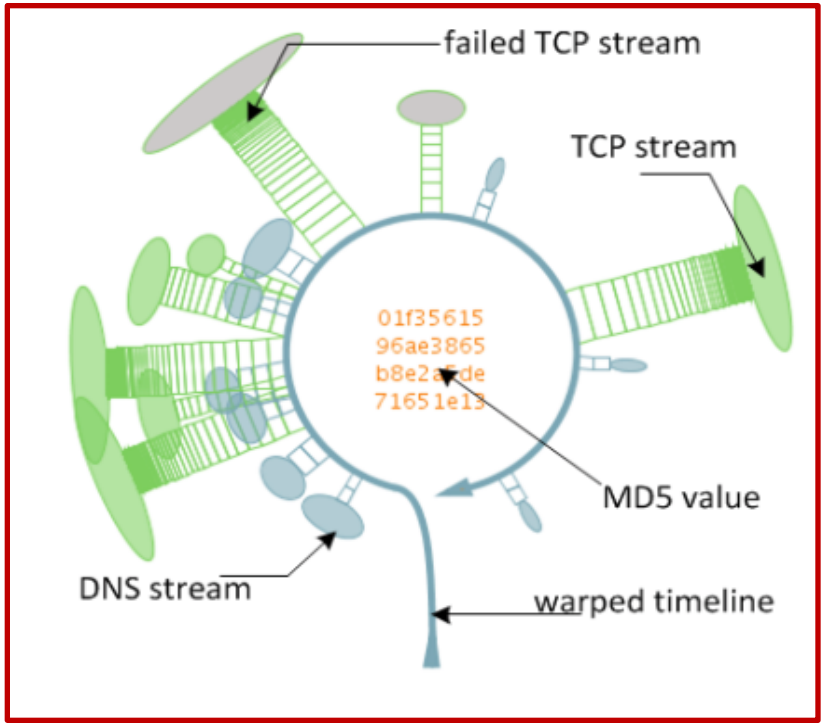
- Main goals of Malware Visualization Systems
- Malware Visualization Taxonomy
  1. Individual Malware Analysis
  2. Malware Comparison
     1. Featured-based approach
     2. Image -based approach
  3. Malware Summarization

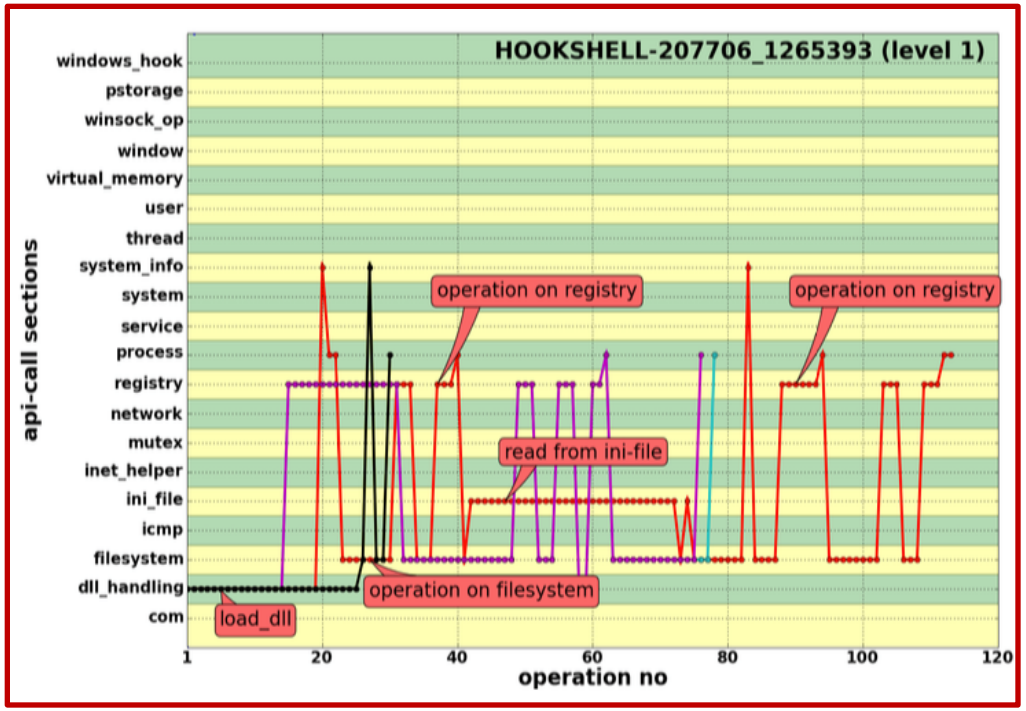
Visualization for Malware Analysis



Visualization for Individual Malware Analysis

- Tools for individual activity
  - Example1: ZN12 focus on network activity of malware
  - Example: THGF09 focus on system calls of malware





Visualization support for Malware comparison

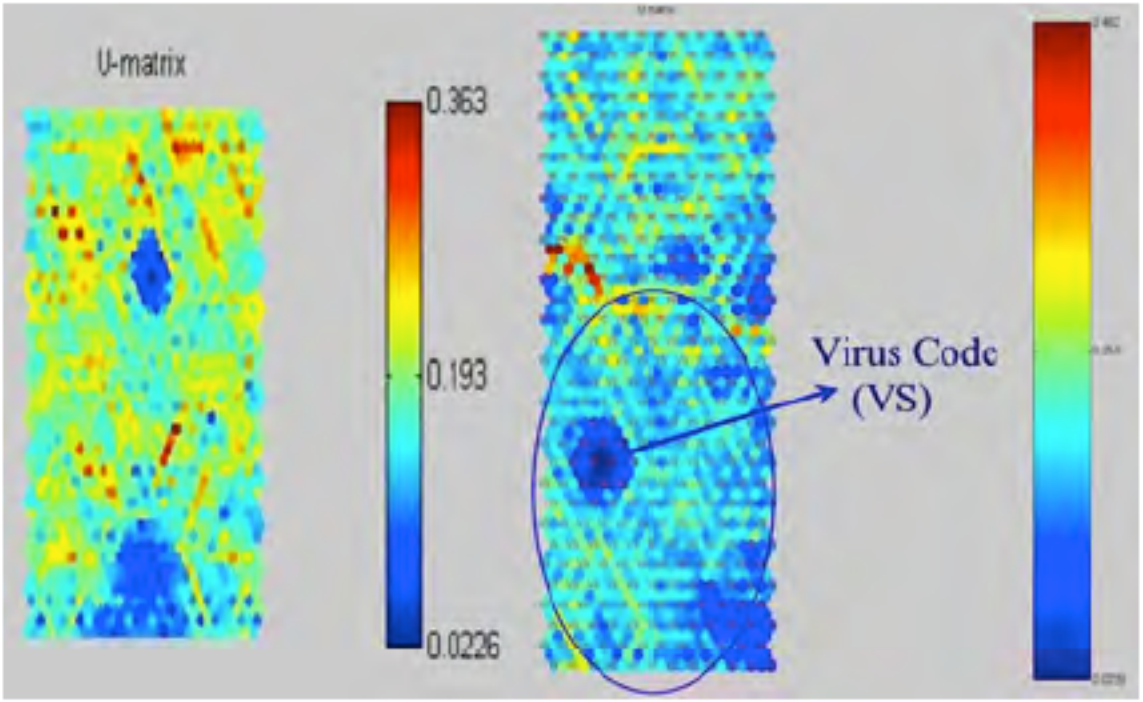
- Important for Malware classification

1. Feature-Based Approach
2. Image-Based Approach

1. Featured Based Approach

Visualization support for Malware summarization

- Summarize and extract a single combined representative out of many malware variants.



Data Providers in Visual Analysis

	[Yoo04]	[Pan08]	[CDSS08]	[QL09]	[THGF09]	[NKJM11]	[GS11]	[QL11]	[YCIZ12]	[GBA*12]	[ZN12]	[SMG12]	[ASL12]	[PCDM13]	[HLI13]	[WY13]	[KM13]	[DPM13]	[SM14b]	[HLKI14]	[HKI14]	[SM14a]	[GSG*14]	[WPO14]	[LSG14]
Raw virus definition	✓	✓	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Raw file (direct input)	✓	✓	✓	-	-	✓	-	-	-	-	-	-	✓	✓	✓	-	✓	✓	-	✓	✓	-	-	-	✓
Packer information	-	-	-	-	-	-	-	-	-	-	-	-	✓	✓	-	-	-	-	-	-	-	-	-	-	-
File information/File header	✓	✓	-	-	-	✓	-	-	-	-	-	-	✓	✓	-	-	✓	✓	-	✓	-	-	✓	-	-
Library imports/loads	-	✓	-	-	✓	-	-	-	-	-	-	-	-	✓	-	-	-	-	-	-	-	-	✓	-	-
CPU instructions/assembly	-	✓	-	✓	-	-	-	✓	✓	-	-	-	✓	-	✓	-	-	-	-	-	✓	-	-	-	-
API calls	-	-	-	-	-	-	-	-	-	✓	-	-	-	-	✓	-	-	-	✓	-	✓	✓	✓	✓	-
System calls	-	-	-	-	-	-	✓	-	-	✓	-	-	✓	-	✓	✓	-	-	-	-	✓	-	✓	✓	-
File system operations	-	-	-	✓	✓	-	✓	✓	-	✓	-	✓	-	-	-	-	-	-	-	-	-	-	-	✓	-
Registry operations	-	-	-	-	✓	-	✓	-	-	✓	-	✓	-	-	-	-	-	-	-	-	-	-	-	✓	-
Process/thread information	-	-	-	-	✓	-	✓	-	-	✓	-	✓	-	-	-	-	-	-	-	-	-	-	-	✓	-
Network activity	-	-	-	-	✓	-	✓	-	-	✓	✓	✓	-	-	-	-	-	-	-	-	-	-	✓	✓	-
Resource utilization	-	-	-	-	-	-	-	-	-	✓	-	-	-	-	-	-	-	-	-	-	-	-	-	✓	-
Memory/driver I/O	-	-	-	✓	-	-	-	✓	-	-	-	-	-	-	-	✓	-	-	-	-	-	-	-	-	-

Data Providers in Visual Analysis

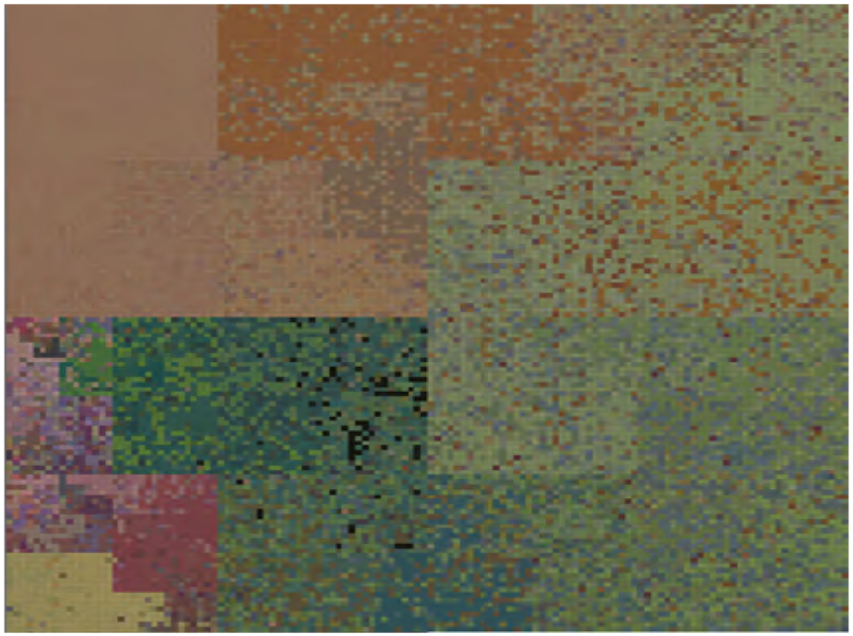
	[Yoo04]	[Pan08]	[CDSS08]	[QL09]	[THGF09]	[NKJM11]	[GS11]	[QL11]	[YCZJ12]	[GBA*12]	[ZN12]	[SMG12]	[ASL12]	[PCDM13]	[HLI13]	[WY13]	[KM13]	[DPM13]	[SM14b]	[HLKI14]	[HKI14]	[SM14a]	[GSG*14]	[WPO14]	[LSG14]
HTML format	-	-	-	-	✓	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
XML format	-	-	-	-	✓	-	-	-	-	-	-	✓	-	✓	-	-	-	-	-	-	-	-	-	-	-
TXT format (plain text)	-	-	-	-	-	-	✓	-	-	✓	-	-	✓	-	✓	-	-	-	✓	-	✓	✓	-	-	-
CSV format	-	-	-	-	-	-	-	-	-	✓	-	✓	✓	-	-	-	-	-	✓	-	-	✓	-	-	-
Native/Proprietary format	✓	✓	-	-	-	-	-	-	-	-	-	✓	✓	-	✓	-	-	-	-	✓	-	-	✓	✓	-
PCAP/network traffic	-	-	-	-	-	-	-	-	-	-	✓	-	-	-	-	-	-	-	-	-	-	-	-	-	-
JSON format	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Raw/binary	✓	✓	✓	✓	-	✓	-	✓	✓	-	-	-	-	✓	✓	✓	✓	✓	-	✓	✓	-	✓	-	✓
Memory dumps (raw)	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	✓	-	-	-	-	-	-	-	-	-
String dumps	-	-	✓	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	✓	-	-
Images (pictures)	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	✓

Visualization Techniques

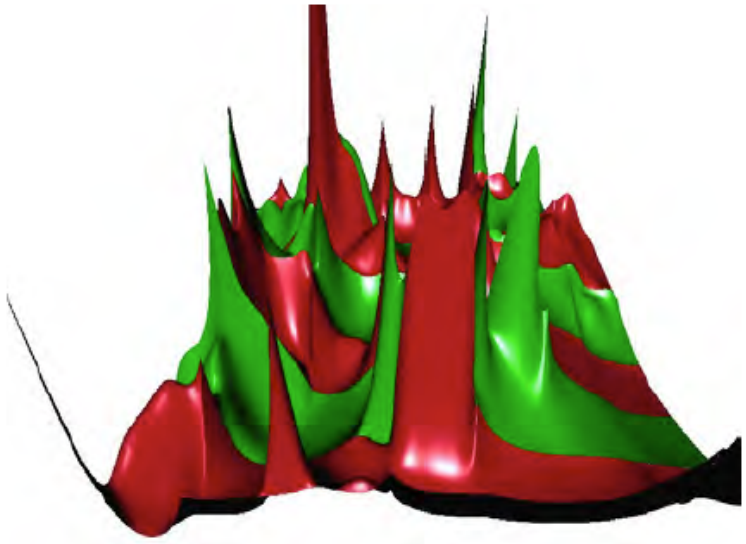
1. Standard 2D/3D Displays
2. Geometrically-transformed Displays
3. Iconic Displays
4. Dense Pixel Displays
5. Stacked Displayed

	[Yoo04]	[Pan08]	[CDSS08]	[QL09]	[THGF09]	[NKJM11]	[GS11]	[QL11]	[YCIZ12]	[GBA*12]	[ZN12]	[SMG12]	[ASL12]	[PCDM13]	[HLI13]	[WY13]	[KM13]	[DPM13]	[SM14b]	[HLKI14]	[HKI14]	[SM14a]	[GSG*14]	[WPO14]	[LSG14]
Standard 2D Display	-	-	-	✓	✓	-	✓	✓	-	✓	-	-	-	-	-	✓	-	✓	✓	✓	-	✓	✓	✓	✓
Standard 3D Display	-	✓	-	✓	-	-	-	✓	-	✓	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Geometrically-transformed Display	✓	✓	-	✓	-	-	✓	✓	✓	-	✓	-	-	✓	-	-	-	-	-	-	-	-	-	✓	✓
Iconic Display	-	-	-	-	-	-	-	-	-	✓	✓	✓	-	-	-	-	-	-	-	-	-	-	✓	-	-
Dense Pixel Display	✓	-	✓	-	-	✓	-	-	-	-	-	✓	✓	-	✓	-	✓	✓	✓	✓	✓	✓	✓	-	-
Stacked Display	-	-	-	-	✓	-	-	-	-	-	-	✓	-	-	-	-	-	-	-	-	-	-	✓	-	-

- Dense Pixel Displays



- 3D Examples



- Representation Spaces and Mapping to Time

	[Yoo04]	[Pan08]	[CDSS08]	[QL09]	[THGF09]	[NKJM11]	[GS11]	[QL11]	[YCIZ12]	[GBA*12]	[ZN12]	[SMG12]	[ASL12]	[PCDM13]	[HLI13]	[WY13]	[KM13]	[DPM13]	[SM14b]	[HLKI14]	[HKI14]	[SM14a]	[GSG*14]	[WPO14]	[LSG14]
Mapping ► Static	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Mapping ► Dynamic	-	-	-	-	-	-	-	-	✓	-	-	-	-	-	-	-	-	-	-	-	-	-	-	✓	-
Dimensionality ► 2D	✓	-	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Dimensionality ► 3D	-	✓	-	✓	-	-	-	✓	-	✓	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-

Temporal Aspect

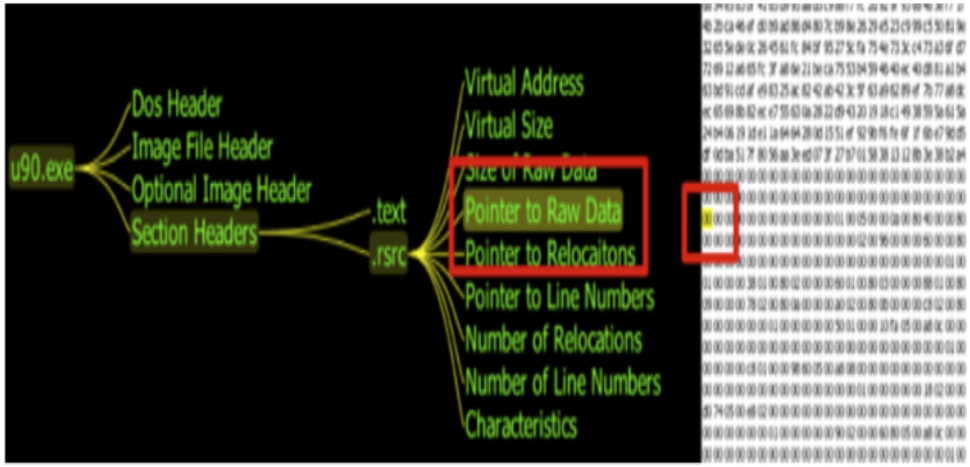
1. Scale (Ordinal, Discrete, Continuous)
2. Arrangement (Linear, Cyclic)
3. Granularity and Calendars (None, Single, Multiple)
4. Time Primitives (Instant, Interval, Span)

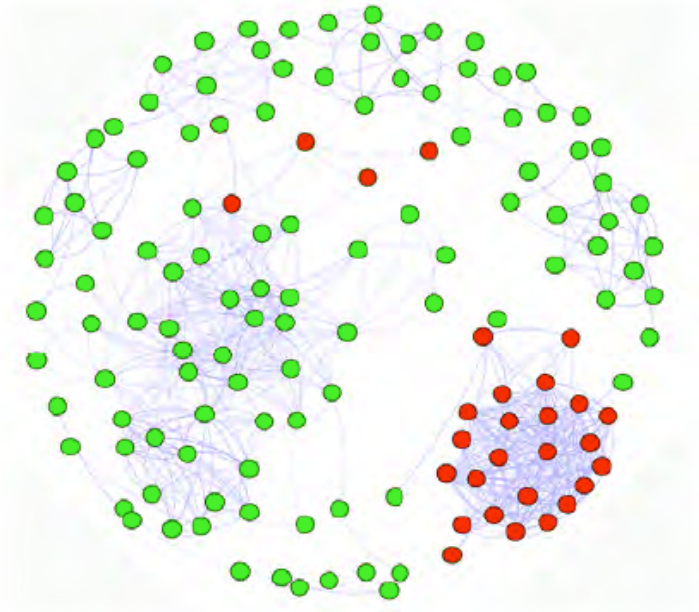
	[Yoo04]	[Pan08]	[CDSS08]	[QL09]	[THGF09]	[NKJM11]	[GS11]	[QL11]	[YCIZ12]	[GBA*12]	[ZN12]	[SMG12]	[ASL12]	[PCDM13]	[HLI13]	[WY13]	[KM13]	[DPM13]	[SM14b]	[HLKI14]	[HKI14]	[SM14a]	[GSG*14]	[WPO14]	[LSG14]
Scale ► Ordinal	-	-	-	✓	✓	-	✓	✓	✓	✓	✓	-	-	-	✓	-	-	-	✓	-	✓	✓	-	-	-
Scale ► Discrete	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	✓	-
Scale ► Continuous	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Arrangement ► Linear	-	-	-	✓	✓	-	✓	✓	✓	✓	✓	-	-	-	✓	-	-	-	✓	-	✓	✓	-	✓	-
Arrangement ► Cyclic	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Granularity ► None	-	-	-	✓	✓	-	✓	✓	✓	✓	-	-	-	-	✓	-	-	-	✓	-	✓	✓	-	-	-
Granularity ► Single	-	-	-	-	-	-	-	-	-	-	✓	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Granularity ► Multiple	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	✓	-
Time primitives ► Instant	-	-	-	✓	✓	-	✓	✓	✓	✓	-	-	-	-	✓	-	-	-	✓	-	✓	✓	-	✓	-
Time primitives ► Interval	-	-	-	-	-	-	-	-	-	-	✓	-	-	-	-	-	-	-	-	-	-	-	-	✓	-
Time primitives ► Span	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-

Interactivity

Interaction

No Interaction





	[Yoo04]	[Pan08]	[CDSS08]	[QL09]	[THGF09]	[NKJM11]	[GS11]	[QL11]	[YCZ12]	[GBA*12]	[ZN12]	[SMG12]	[ASL12]	[PCDM13]	[HLI13]	[WY13]	[KM13]	[DPM13]	[SM14b]	[HLKI14]	[HKI14]	[SM14a]	[GSG*14]	[WPO14]	[LSG14]
Interaction	-	-	✓	✓	-	-	✓	✓	✓	✓	✓	✓	-	-	-	✓	-	✓	-	-	-	-	✓	✓	✓
No Interaction	✓	✓	-	-	✓	✓	-	-	-	-	-	-	✓	✓	✓	-	✓	-	✓	✓	✓	✓	-	-	-

Problems/Actions

1. Analyze (Consume, Produce)
2. Search (Lookup, Locate, Browse, Explore)
3. Query (Identify, Compare, Summarize)

	[Yoo04]	[Pan08]	[CDSS08]	[QL09]	[THGF09]	[NKJM11]	[GS11]	[QL11]	[YCIZI12]	[GBA*12]	[ZN12]	[SMG12]	[ASL12]	[PCDM13]	[HLI13]	[WY13]	[KM13]	[DPM13]	[SM14b]	[HLKI14]	[HKI14]	[SM14a]	[GSG*14]	[WPO14]	[LSG14]
Analyze ► Consume ► Discover	-	-	✓	✓	✓	-	✓	✓	✓	✓	✓	✓	-	✓	-	✓	-	✓	✓	✓	-	✓	✓	✓	✓
Analyze ► Consume ► Present	✓	-	✓	✓	✓	✓	✓	✓	-	✓	✓	✓	-	-	-	✓	✓	-	✓	✓	-	✓	✓	✓	✓
Analyze ► Consume ► Enjoy	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Analyze ► Produce ► Annotate	-	-	-	-	-	-	-	-	-	-	-	✓	-	-	-	-	-	-	-	-	-	-	✓	✓	-
Analyze ► Produce ► Record	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Analyze ► Produce ► Derive	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Search ► Lookup	-	-	-	✓	-	-	✓	✓	✓	-	✓	✓	-	-	-	-	-	✓	-	-	-	-	✓	-	-
Search ► Browse	-	-	-	-	-	-	-	-	-	-	✓	-	-	-	-	-	-	-	-	-	-	-	✓	-	✓
Search ► Locate	-	-	✓	✓	✓	-	✓	✓	✓	-	✓	✓	-	-	-	-	-	✓	-	-	-	-	✓	-	-
Search ► Explore	-	-	-	-	-	-	✓	-	-	-	✓	✓	-	-	-	✓	-	-	-	-	-	-	✓	✓	-
Query ► Identify	-	-	-	-	-	-	-	-	-	-	-	✓	-	-	-	-	-	-	-	-	-	-	✓	✓	-
Query ► Compare	✓	✓	-	✓	✓	✓	✓	✓	-	✓	✓	✓	-	✓	✓	✓	✓	-	✓	✓	✓	✓	✓	-	✓
Query ► Summarize	✓	-	-	-	-	-	-	-	-	-	-	✓	✓	✓	✓	-	-	-	-	-	✓	-	-	-	-

Future Challenges

- Bridge between categories
- Integrate different data sources
- Involve expert knowledge through interaction
- Intertwine analytical methods with visualization