Friedrich-Alexander-Universität
Technische Fakultät

Cybercrime and
Forensic Computing
Research Training Group 2475

FAU

A Metrics-Based Look at Disk Images: Insights and Applications

Lena L. Voigt, Felix Freiling and Christopher Hargreaves

Digital Forensics Research Conference Europe | April 3, 2025

Motivation

A Metrics-Based Look at Disk Images

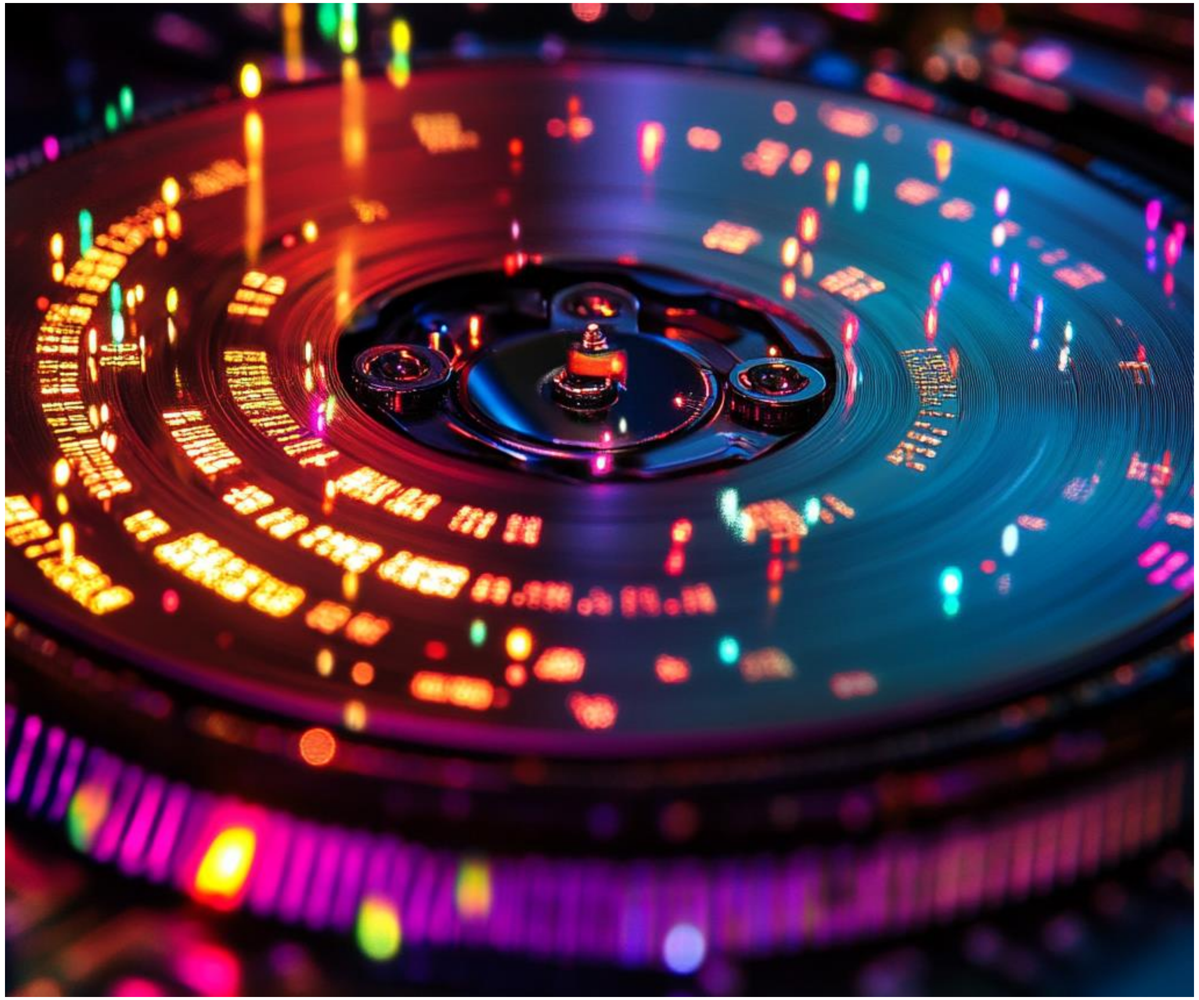


*Image generated using Midjourney

Explanation of Terms Used

Real-World vs. Scenario-Based Synthetic Disk Images

Real-world disk image

- *image of a disk on which regular day-to-day activities were carried out by one or more users without the intention of creating data for digital forensic analysis or investigation.*

Scenario-based synthetic disk image

- *produced with the intention of creating data that can be utilized for digital forensic purposes*
- *created in accordance with a scenario for digital forensic investigation:*
  - *disks from the M-57 scenarios (Digital Corpora)*
  - *forensic CTF disks with a storyline*

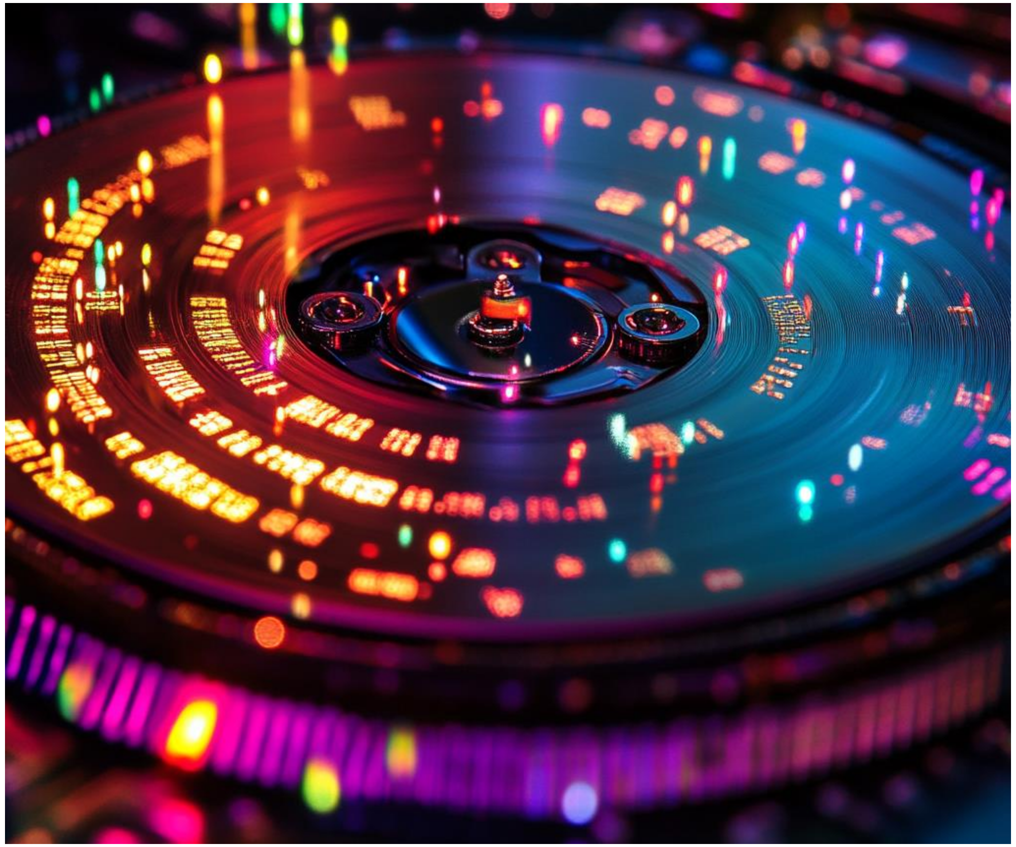


*Image generated using Midjourney

Explanation of Terms Used

Real-World vs. Scenario-Based Synthetic Disk Images

Real-world disk image

- *image of a disk on which regular day-to-day activities were carried out by one or more users without the intention of creating data for digital forensic analysis or investigation.*

Scenario-based synthetic disk image

- *produced with the intention of creating data that can be utilized for digital forensic purposes*
- *created in accordance with a scenario for digital forensic investigation:*
  - *disks from the M-57 scenarios (Digital Corpora)*
  - *forensic CTF disks with a storyline*

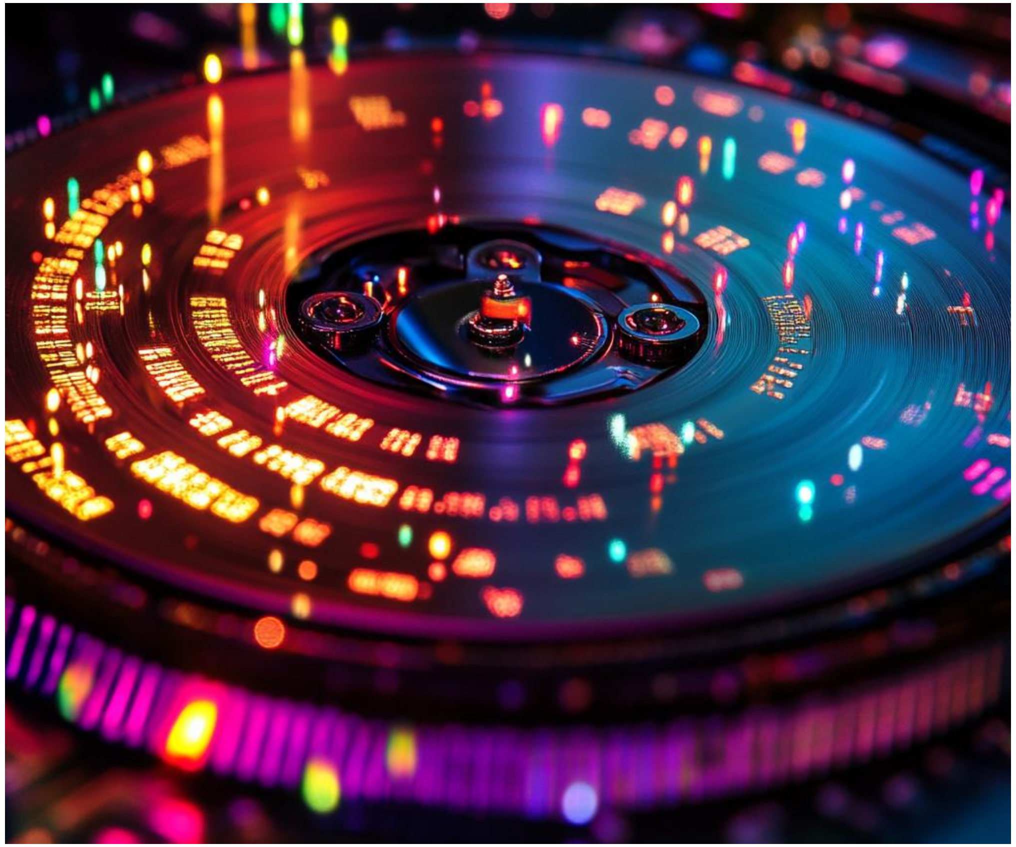


*Image generated using Midjourney

Explanation of Terms Used

Real-World vs. Scenario-Based Synthetic Disk Images

Real-world disk image

- *image of a disk on which regular day-to-day activities were carried out by one or more users without the intention of creating data for digital forensic analysis or investigation.*

Scenario-based synthetic disk image

- *produced with the intention of creating data that can be utilized for digital forensic purposes*
- *created in accordance with a scenario for digital forensic investigation:*
  - *disks from the M-57 scenarios (Digital Corpora)*
  - *forensic CTF disks with a storyline*

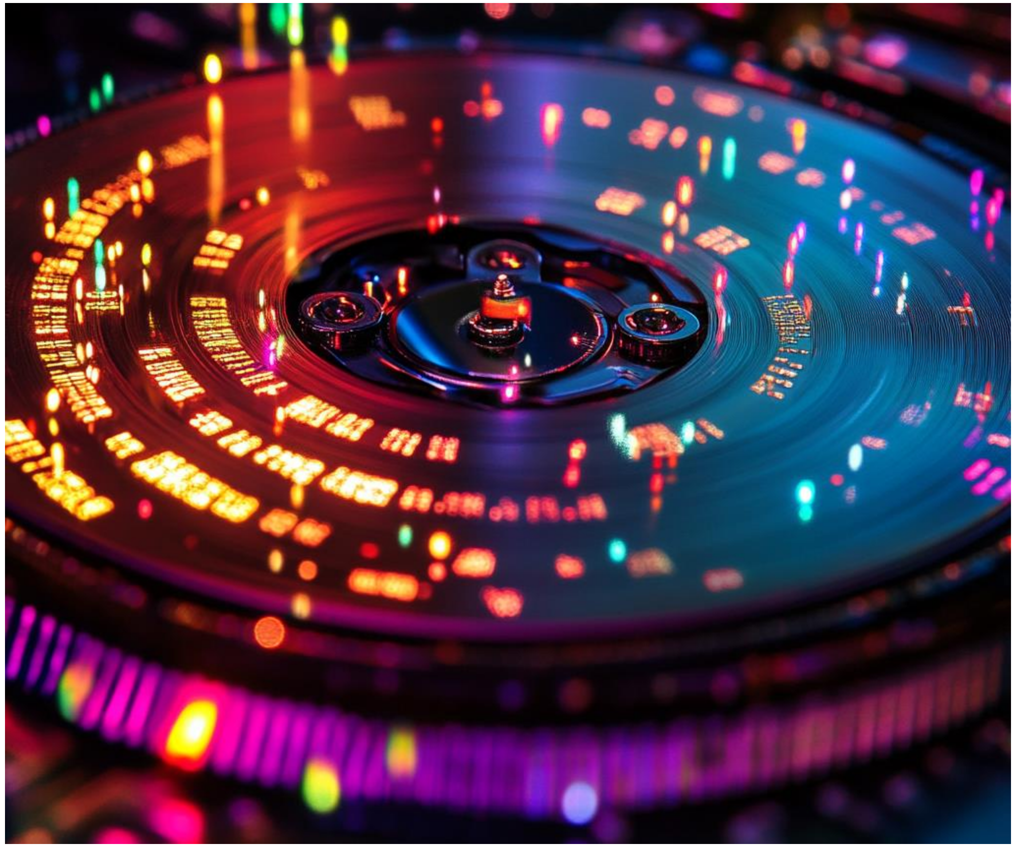


*Image generated using Midjourney

Realism of Synthetic Disk Images

Methodology

1. A **Formal Definition** of *Realism* of Synthetic Disk Images
2. **Compiling Datasets** of Synthetic and Real-World Disk Images
3. **Collecting Metrics** from Disk Image Datasets
4. Insights from **a Comparison of Synthetic and Real-world** Disk Images

Realism of Synthetic Disk Images

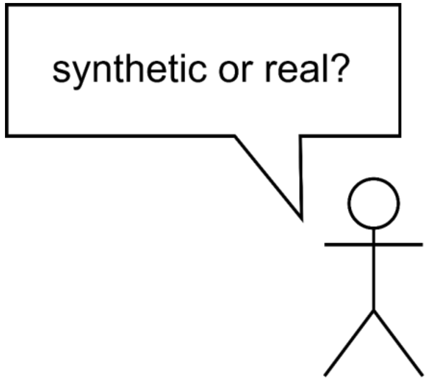
A Formal Definition

Intuitive Definition

- A realistic synthetic disk image is *indistinguishable* from a real-world disk image.

Considerations

- What does *indistinguishable* mean? Aren't two disk images always distinguishable?
    - ➢ We need to compare sets of synthetic and real-world disk images and the distribution of values for different features.
- Some features of the disk image can be 'out of scope' for the analysis, e.g.:
    - Does the disk image exhibit virtualization artifacts?
    - Are there traces of an automation framework used?



Realism of Synthetic Disk Images

A Formal Definition

Concept: Define *Realism* using a cryptography-inspired security game

Data Sets: ***R*** (Real-world disk images), ***S*** (Synthetic disk images)

Security Game:

- Verifier tries to distinguish data from ***R*** and ***S***; can only inspect *allowed* features of the data

Process:

Presentation: Verifier is presented an item (randomly of either ***R*** or ***S***)

Realism of Synthetic Disk Images

A Formal Definition

Concept: Define *Realism* using a cryptography-inspired security game

Data Sets: ***R*** (Real-world disk images), ***S*** (Synthetic disk images)

Security Game:

- Verifier tries to distinguish data from ***R*** and ***S***; can only inspect *allowed* features of the data

Process:

Presentation: Verifier is presented an item (randomly of either ***R*** or ***S***)

1. Query: Verifier selects a feature of the item to inspect
2. Check: Prover verifies if this query is *allowed*
3. Response: Measurement of the feature is provided

Realism of Synthetic Disk Images

A Formal Definition

Concept: Define *Realism* using a cryptography-inspired security game

Data Sets: ***R*** (Real-world disk images), ***S*** (Synthetic disk images)

Security Game:

- Verifier tries to distinguish data from ***R*** and ***S***; can only inspect *allowed* features of the data

Process:

Presentation: Verifier is presented an item (randomly of either ***R*** or ***S***)

1. Query: Verifier selects a feature of the item to inspect
2. Check: Prover verifies if this query is *allowed*
3. Response: Measurement of the feature is provided

Realism of Synthetic Disk Images

A Formal Definition

Concept: Define *Realism* using a cryptography-inspired security game

Data Sets: ***R*** (Real-world disk images), ***S*** (Synthetic disk images)

Security Game:

- Verifier tries to distinguish data from ***R*** and ***S***; can only inspect *allowed* features of the data

Process:

Presentation: Verifier is presented an item (randomly of either ***R*** or ***S***)

1. Query: Verifier selects a feature of the item to inspect
2. Check: Prover verifies if this query is *allowed*
3. Response: Measurement of the feature is provided

Guess: Verifier decides "real-world" or "synthetic"

Realism of Synthetic Disk Images

A Formal Definition

Concept: Define *Realism* using a cryptography-inspired security game

Data Sets: ***R*** (Real-world disk images), ***S*** (Synthetic disk images)

Security Game:

- Verifier tries to distinguish data from ***R*** and ***S***; can only inspect *allowed* features of the data

Process:

Presentation: Verifier is presented an item (randomly of either ***R*** or ***S***)

1. Query: Verifier selects a feature of the item to inspect
2. Check: Prover verifies if this query is *allowed*
3. Response: Measurement of the feature is provided

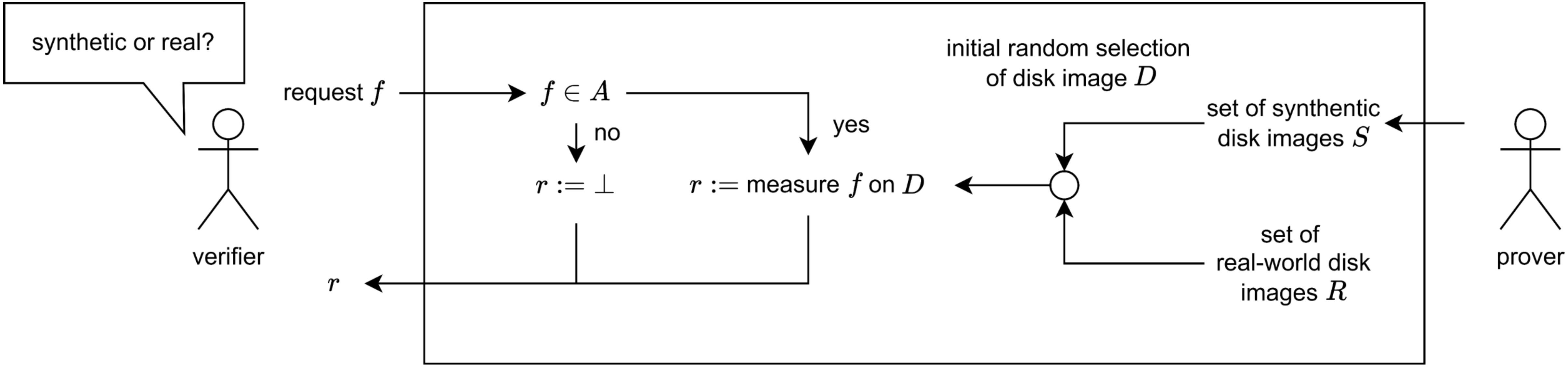
Guess: Verifier decides "real-world" or "synthetic"

n Iterations (with distinct items):

If the Verifier cannot reliably distinguish synthetic and real-world data items, and we call the synthetic forensic data in set ***S*** *realistic* w.r.t. the *allowed* features.

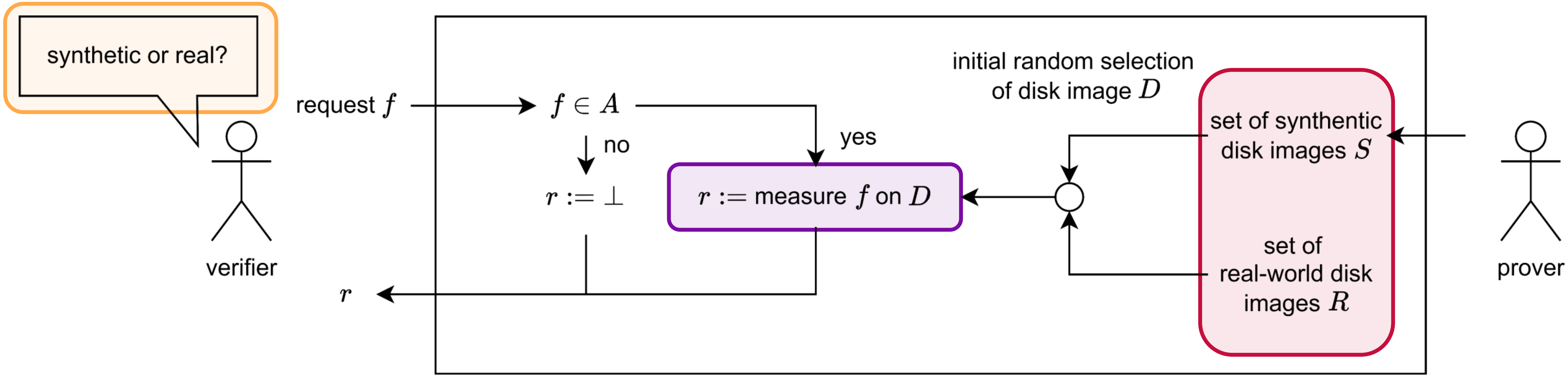
Realism of Synthetic Disk Images

A Formal Definition


synthetic or real?
verifier
request f
$f \in A$
no
yes
$r := \perp$
$r :=$ measure f on D
r
initial random selection of disk image D
set of synthentic disk images S
set of real-world disk images R
prover

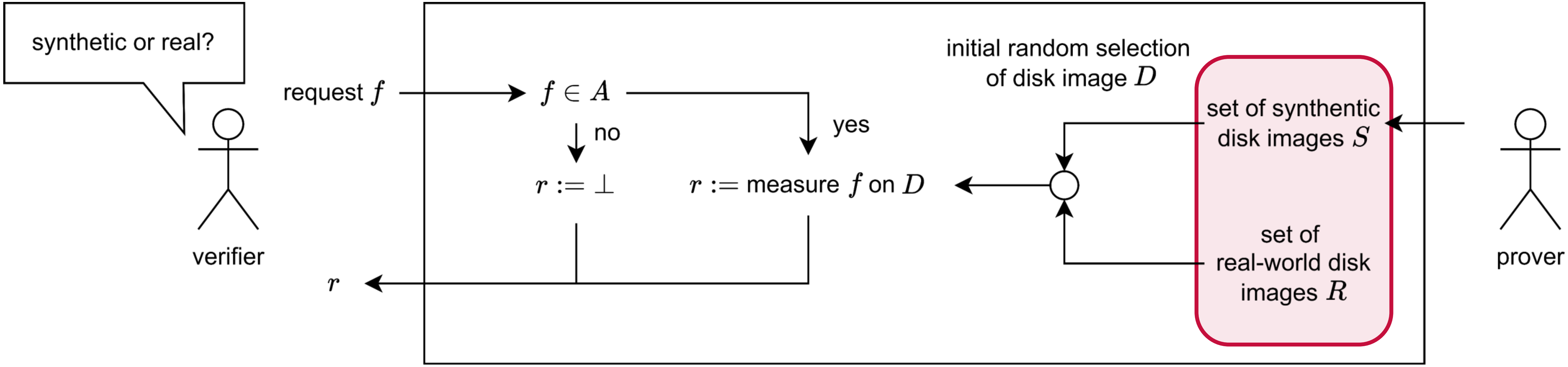
Realism of Synthetic Disk Images

Compiling Datasets of Synthetic and Real-World Disk Images


synthetic or real?
verifier
request f
$f \in A$
no
$r := \perp$
yes
$r :=$ measure f on D
initial random selection of disk image D
set of synthentic disk images S
set of real-world disk images R
prover
r

Realism of Synthetic Disk Images

Compiling Datasets of Synthetic and Real-World Disk Images


synthetic or real?
verifier
request f
$f \in A$
no
yes
$r := \perp$
$r :=$ measure f on D
r
initial random selection of disk image D
set of synthentic disk images S
set of real-world disk images R
prover

Realism of Synthetic Disk Images

Compiling Datasets of Synthetic and Real-World Disk Images

Disk Image Collection

Conducted in September 2024, only Windows systems

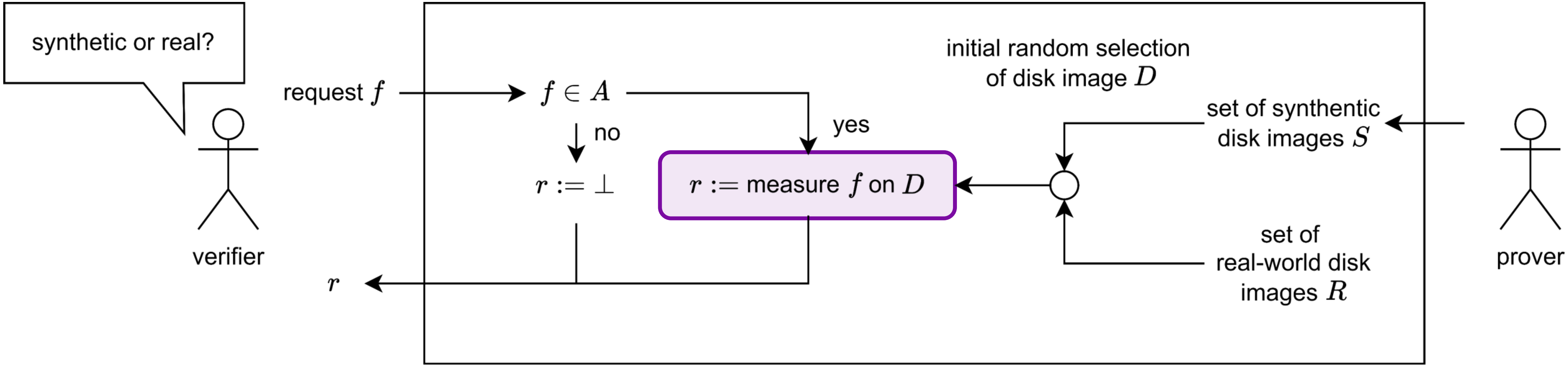
- **Public[1]**: publ. available (Digital Corpora, CFReDS, etc.)
- **Internal[1]**: from five different institutions
- **Real-World:** drives from personal computers, in use between June 2012 and September 2024

Windows Version	Public[2]	Internal	Real-world
Windows 11	3	-	2
Windows 10	8	11	5
Windows 8.1	3	-	-
Windows 7	2	19	3
Windows Vista	1	1	-
Windows XP	6	5	1
Win. Server 2008	1	-	-
Win. Server 2022	1	-	-
Win. Server 2019	-	1	-
Gesamt	**25**	**37**	**11**

[1] scenario-based, synthetic

Realism of Synthetic Disk Images

Collecting Metrics from Disk Image Datasets


synthetic or real?
verifier
request f
$f \in A$
no
$r := \perp$
yes
$r :=$ measure f on D
initial random selection of disk image D
set of synthentic disk images S
set of real-world disk images R
prover
r

Realism of Synthetic Disk Images

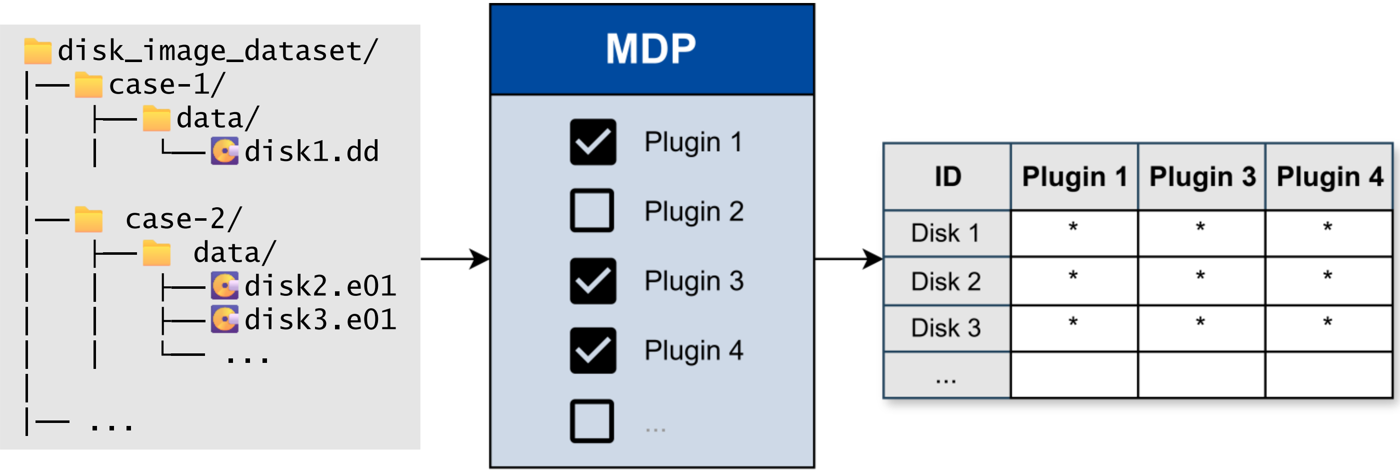
Collecting Metrics from Disk Image Datasets

The Mass Disk Processor (MDP)

- ➢ Automating the collection of metrics from large sets of disk images and summarizing the results
- ➢ Allows for extraction of privacy-preserving high-level metrics

Impementation

- pytsk/libewf wrapper for disk image access
- plugin-based architecture
- *optional* preprocessing
  - ➢ File signature extraction, sha1 calculation
- *optional* integration of existing tools
  - ➢ pyregistry, python-evtx, Plaso



```
disk_image_dataset/
|── case-1/
|     ├── data/
|     |     └── disk1.dd
|
|──  case-2/
|     ├──  data/
|     |     ├── disk2.e01
|     |     ├── disk3.e01
|     |     └── ...
|
|── ...
```

MDP

- ☑ Plugin 1
- □ Plugin 2
- ☑ Plugin 3
- ☑ Plugin 4
- □ ...

ID	Plugin 1	Plugin 3	Plugin 4
Disk 1	*	*	*
Disk 2	*	*	*
Disk 3	*	*	*
...			

Realism of Synthetic Disk Images

Collecting Metrics from Disk Image Datasets

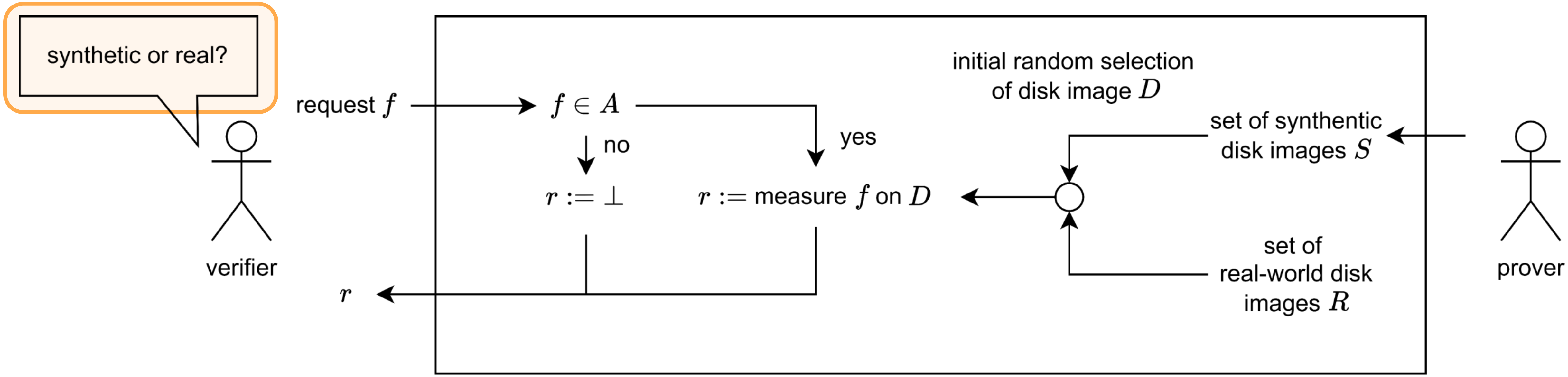
The Mass Disk Processor (MDP)

Category	MDP Metric Name	Metric Description	Value[1]
Configuration	disk_size	Size of the disk *(converted to GB)*	40
	win_build_inferred_os, win_build	Windows Version and Build	Windows 10 (Build: 17763)
Longevity	windows_install	Windows Install Time	2019-03
	windows_last_shutdown	Windows Last Shutdown	2024-03
	win_os_lifetime	Windows OS Lifetime *(in days)*	1809.04
Activity	win_total_login_count	Windows Login Count	33
	no_start_menu_lnk_total	Number of Startmenue Lnk Files	49
	chrome_history_entries_total	Chrome History Entries	0
Volume	no_non_nsrl_files	Number of non-NSRL Files	139864
	audio_files	Number of Audio Files	279
	video_files	Number of Video Files	31
	office_files	Number of Office Files	43
Notables	no_signature_mismatches	Number of Signature Mismatches	4428
	evtx_clock_change_4616	Number of Clock Change Events	9

[1] Values for DFRWS EU 2024 – Rodeo Image (Bytebusters)

Realism of Synthetic Disk Images

Insights from a Metrics-Based Comparison


synthetic or real?
verifier
request f
$f \in A$
no
yes
$r := \perp$
$r :=$ measure f on D
r
initial random selection of disk image D
set of synthentic disk images S
set of real-world disk images R
prover

Synthetic Scenario-Based Disk Images vs. Real-World Disk Images

- Configuration:
  - Disk size, operating system(s) installed, number of users, applications installed, default browsers
- Longevity
- Activity
- Volume

[1] source: microsoft.com/edge
[2] source: google.com/chrome
[3] source: mozilla.org/de/firefox

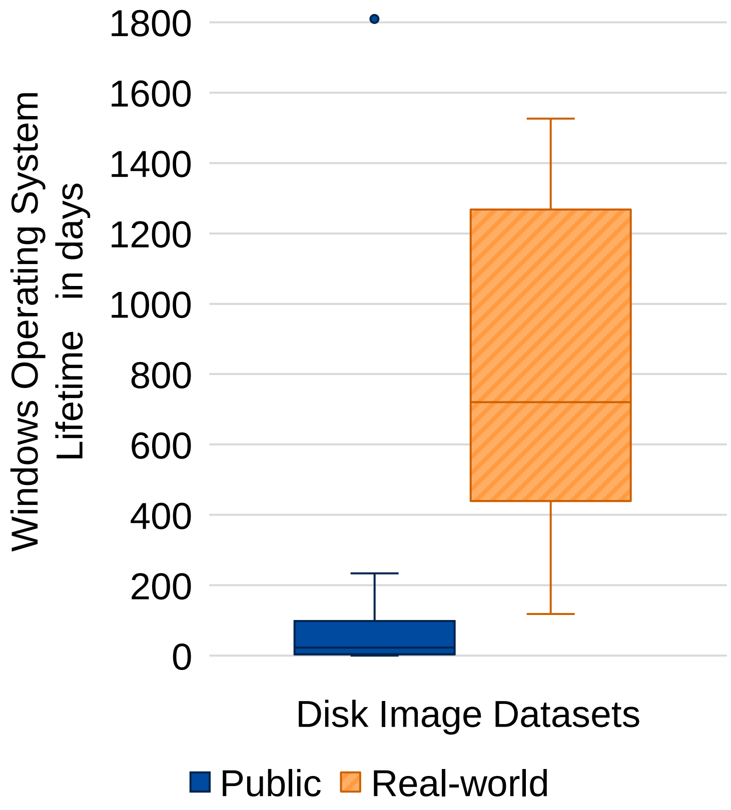
Realism of Synthetic Disk Images

Insights from a Metrics-Based Comparison

Cybercrime and Forensic Computing
Research Training Group 2475

Synthetic Scenario-Based Disk Images vs. Real-World Disk Images

- Configuration
- Longevity:
  - File system lifespan, <u>Windows operating system lifetime</u>
- Activity
- Volume

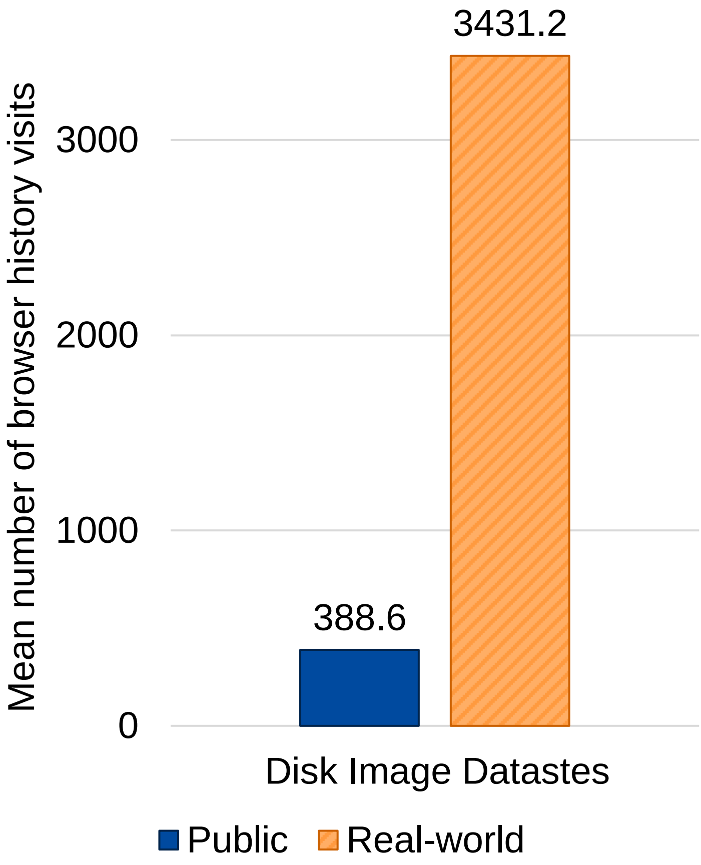


Realism of Synthetic Disk Images

Insights from a Metrics-Based Comparison

Synthetic Scenario-Based Disk Images vs. Real-World Disk Images

- Configuration
- Longevity
- Activity:
  - ➢ Number of logins, number of USB drives attached, number of browser history entries, number of browser searches
- Volume

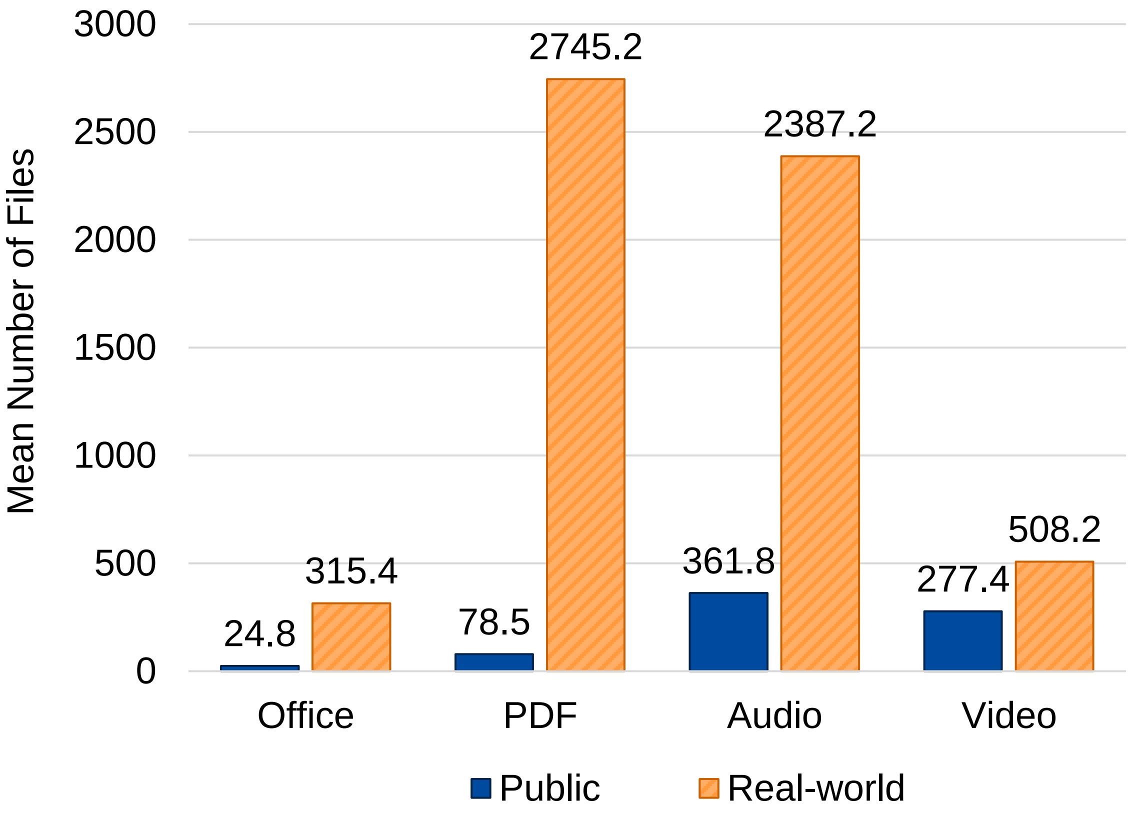


Realism **of Synthetic Disk Images**

Insights from a Metrics-Based Comparison

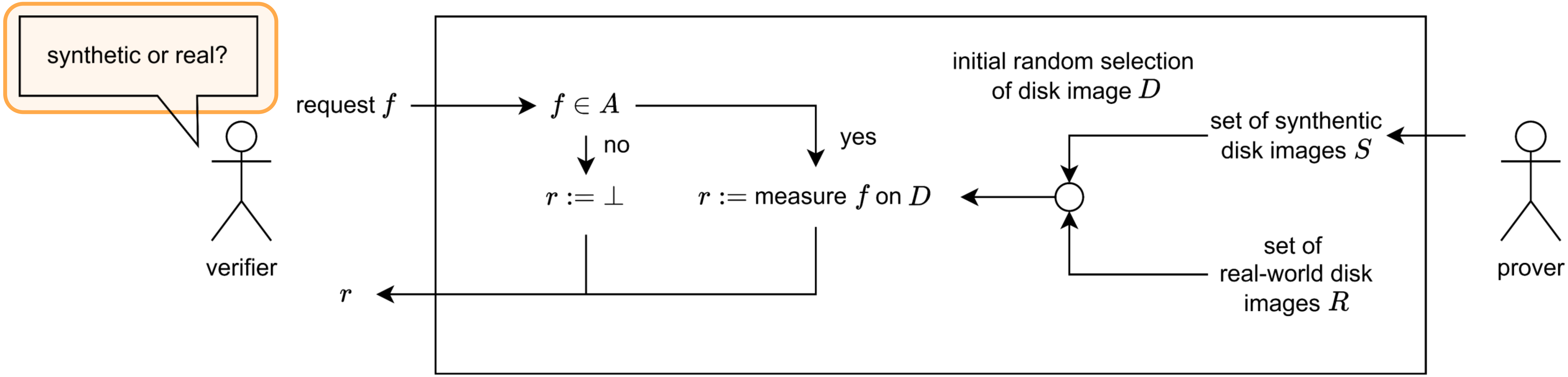
Synthetic Scenario-Based Disk Images vs. Real-World Disk Images

- Configuration
- Longevity
- Activity
- Volume:
  - ➢ Number of files, number of user folder files, number of files per type:
    - ○ Office files
    - ○ PDF files
    - ○ Audio files
    - ○ Video files



Realism of Synthetic Disk Images

Insights from a Metrics-Based Comparison

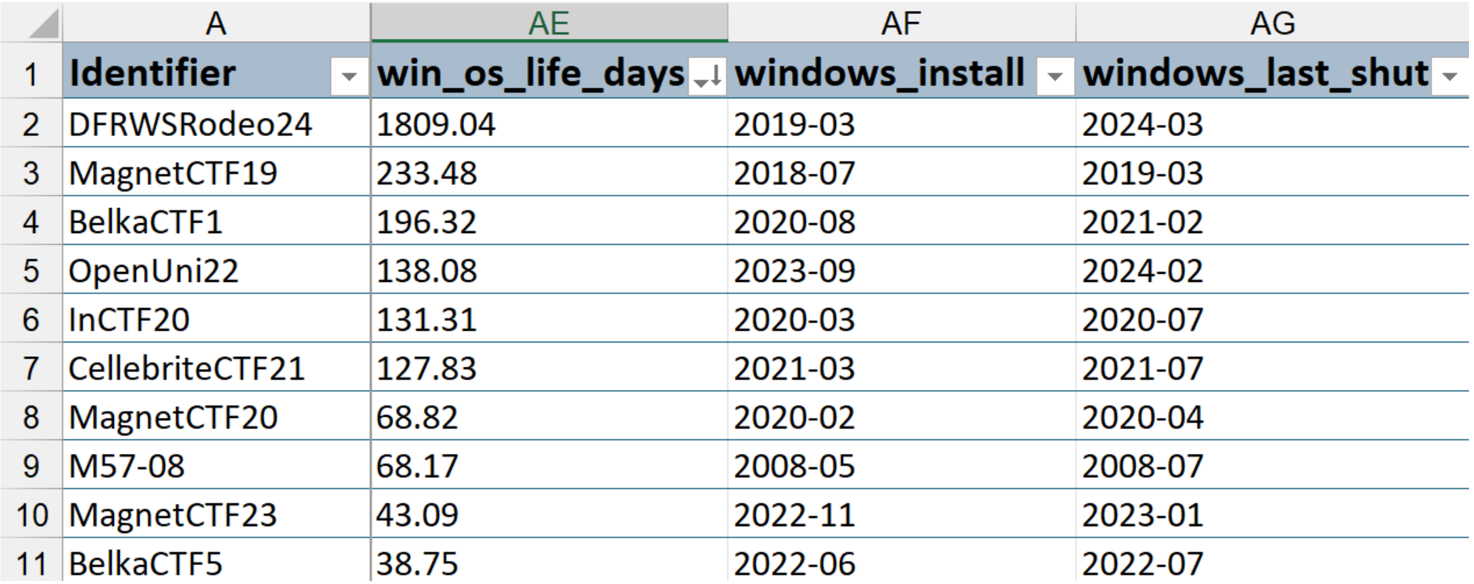

synthetic or real?
verifier
request f
$f \in A$
no
yes
$r := \perp$
$r :=$ measure f on D
initial random selection of disk image D
set of synthentic disk images S
set of real-world disk images R
r
prover

Compiling Datasheets for Disks Images

An Overview of Publicly Available, Scenario-Based Synthetic Disk Images

Providing Metrics for Public Disk Images

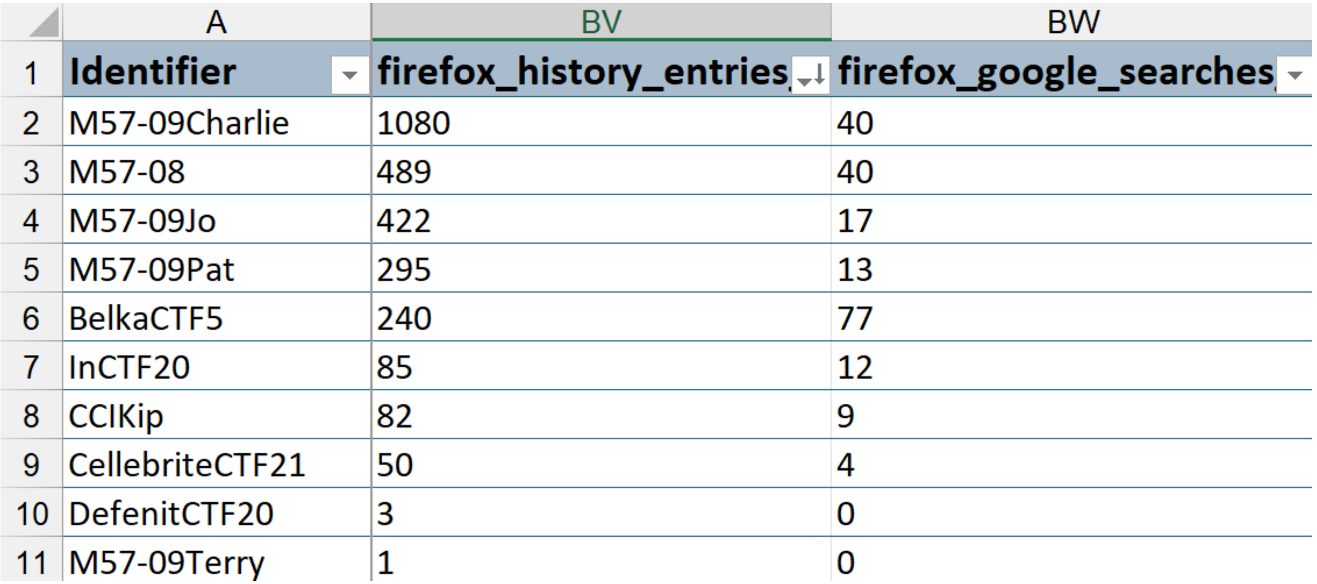
1. Datasheets for individual disk images
2. Summary table for public disk images
   - Facilitate the selection of a public disk image that fits specific needs:
     - ➢ **Long Windows Lifetime**
     - ➢ Rich Firefox browser history
     - ➢ High volume of files



	A	AE	AF	AG
1	**Identifier**	**win_os_life_days**	**windows_install**	**windows_last_shut**
2	DFRWSRodeo24	1809.04	2019-03	2024-03
3	MagnetCTF19	233.48	2018-07	2019-03
4	BelkaCTF1	196.32	2020-08	2021-02
5	OpenUni22	138.08	2023-09	2024-02
6	InCTF20	131.31	2020-03	2020-07
7	CellebriteCTF21	127.83	2021-03	2021-07
8	MagnetCTF20	68.82	2020-02	2020-04
9	M57-08	68.17	2008-05	2008-07
10	MagnetCTF23	43.09	2022-11	2023-01
11	BelkaCTF5	38.75	2022-06	2022-07

Providing Metrics for Public Disk Images

1. **Datasheets** for individual disk images
2. **Summary table** for public disk images
   - Facilitate the selection of a public disk image that fits specific needs:
     - ➢ Long Windows Lifetime
     - ➢ **Rich Firefox browser history**
     - ➢ High volume of files



	A	BV	BW
1	**Identifier**	**firefox_history_entries**	**firefox_google_searches**
2	M57-09Charlie	1080	40
3	M57-08	489	40
4	M57-09Jo	422	17
5	M57-09Pat	295	13
6	BelkaCTF5	240	77
7	InCTF20	85	12
8	CCIKip	82	9
9	CellebriteCTF21	50	4
10	DefenitCTF20	3	0
11	M57-09Terry	1	0

Compiling Datasheets for Disks Images

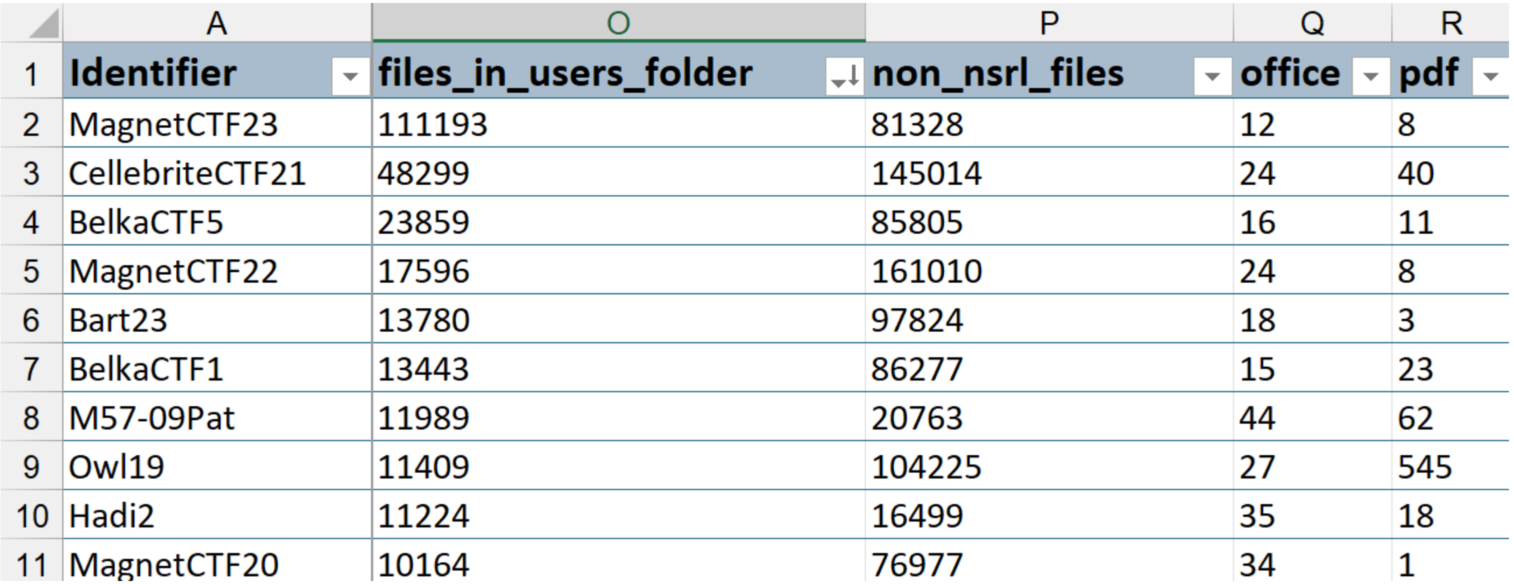
An Overview of Publicly Available, Scenario-Based Synthetic Disk Images

Cybercrime and Forensic Computing
Research Training Group 2475

FAU

Providing Metrics for Public Disk Images

1. **Datasheets** for individual disk images
2. **Summary table** for public disk images
   - Facilitate the selection of a public disk image that fits specific needs:
     - ➢ Long Windows Lifetime
     - ➢ Rich Firefox browser history
     - ➢ **High volume of user files**



	A	O	P	Q	R
1	**Identifier**	**files_in_users_folder**	**non_nsrl_files**	**office**	**pdf**
2	MagnetCTF23	111193	81328	12	8
3	CellebriteCTF21	48299	145014	24	40
4	BelkaCTF5	23859	85805	16	11
5	MagnetCTF22	17596	161010	24	8
6	Bart23	13780	97824	18	3
7	BelkaCTF1	13443	86277	15	23
8	M57-09Pat	11989	20763	44	62
9	Owl19	11409	104225	27	545
10	Hadi2	11224	16499	35	18
11	MagnetCTF20	10164	76977	34	1

Discussion

Are Metrics the Solution? – The Need for Qualitative Assessment

Metrics do not eliminate the **need for qualitative assessment**:

- Complexity of evidence recovery
- Difficulty in reconstructing the scenario
- Coherence of the underlying storyline

Sole focus on metrics can undermine qualitative considerations:

- Longevity ↑: Clock manipulations
- Activity ↑: Arbitrary launching of programs
- Volume ↑: Depositing large numbers of random files

Discussion

Future Work – More Data, More Metrics, More Insights

1. **Larger scale collection of disk metrics**
   - For real-world as well as synthetic disk images
2. **Implementing further metrics**
   - Metrics for different use cases, multiple metrics for the same characteristic
   - Cross-plugin metrics
   - Cross-device metrics for more complex cases
3. **Exploring further applications**
   - Metric Sharing for Non-Shareable Data
   - Evaluation of Synthesis Proposals
   - Lab Metrics (for cost/resource estimation or prioritization)

Conclusion

A Metrics-Based View of Disk Images

Contribution:

- Mass Disk Processor (MDP) Framework: Open-Source Framework for Retrieving Disk Metrics in Bulk
- Formal definition of *Realism* in synthetic disk images
- Comparison of scenario-based and real-world disk images
- Datasheets for public disk images

Further Application Scenarios:

- Sharing of non-shareable data,
- Lab metrics,
- Prioritization, etc.

https://github.com/lenavoigt/mass-disk-processor

- **MDP Code**
- **Summary sheet** of public disk images
- **Individual datasheets** for public disk images
- **Plaso timelines** for public disk images