Sharing datasets for Digital Forensic: a novel taxonomy and legal concerns

Frank Breitinger, Alexandre Jotterand

Presentation for DFRWS USA, Baltimore

DFRWS
DIGITAL FORENSIC RESEARCH CONFERENCE

UNIL | Université de Lausanne

id est
avocats

Grajeda, C., Breitinger, F., & Baggili, I. (2017). Availability of datasets for digital forensics–and what is missing. *Digital Investigation*, *22*, S94-S105.

Data problem

Available Data is important for research …

… but often not shared (Grajeda et al. (2017))

Many funding agencies now require sharing data

Data Management Plan (DMP) according to FAIR principles

How can datasets be described to ensure findability? (**RQ1**)

Findability requires a common terminology, taxonomies, and classification schemes to organize datasets (searchability)

Under what circumstances can datasets be shared? (**RQ2**)

Depending on the origin (e.g., real-world data), restrictions may apply which prohibit sharing

Contributions

01

Summarize frequently used terms (common understanding)

02

Propose a new taxonomy (complements existing taxonomies)

03

Outline existing legal restrictions impacting sharing of data

Characteristic Feature

Existing repositories language focuses on the **content of the data**

We propose two additional factors:

organization and origin

Depending on the “origin” legal restrictions apply

Terminology

Organization of data: Structured, semi-structured, and unstructured data

Ground truth data

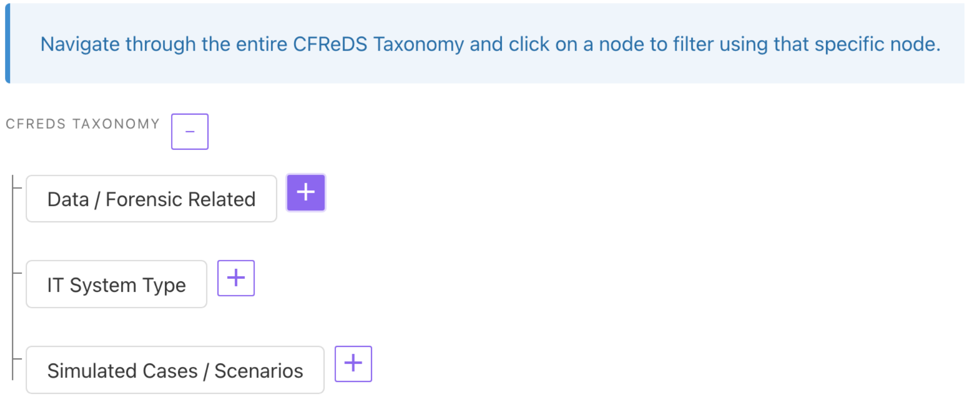
Metadata

Sensitivity of data from a research perspective

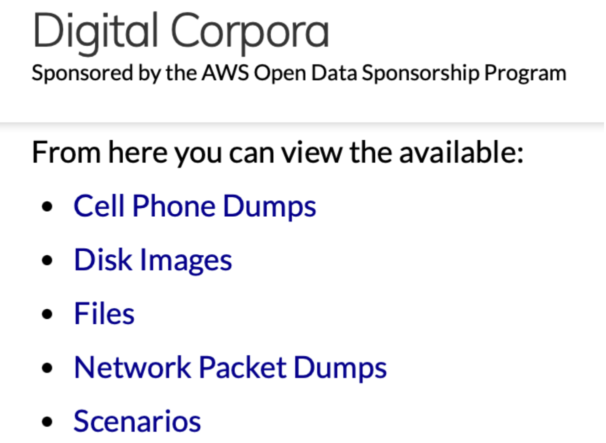
Current situation – Data repositories

Organize data based on the content

e.g., CFReDS uses tags to organize data



Navigate through the entire CFReDS Taxonomy and click on a node to filter using that specific node.



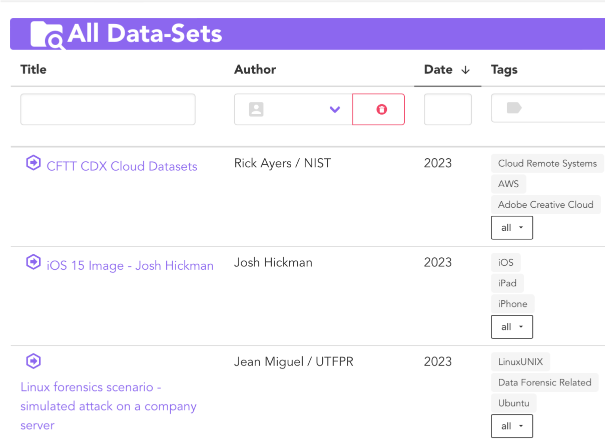
Digital Corpora

Sponsored by the AWS Open Data Sponsorship Program

From here you can view the available:

- Cell Phone Dumps
- Disk Images
- Files
- Network Packet Dumps
- Scenarios

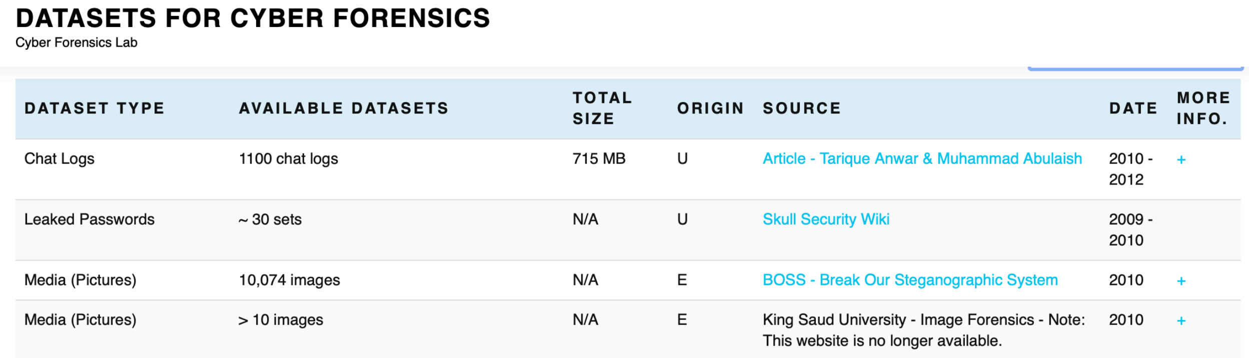
digitalcorpora.org



All Data-Sets

Title	Author	Date ↓	Tags
CFTT CDX Cloud Datasets	Rick Ayers / NIST	2023	Cloud Remote Systems AWS Adobe Creative Cloud all
iOS 15 Image - Josh Hickman	Josh Hickman	2023	iOS iPad iPhone all
Linux forensics scenario - simulated attack on a company server	Jean Miguel / UTFPR	2023	LinuxUNIX Data Forensic Related Ubuntu all

cfreds.nist.gov



DATASETS FOR CYBER FORENSICS

Cyber Forensics Lab

DATASET TYPE	AVAILABLE DATASETS	TOTAL SIZE	ORIGIN	SOURCE	DATE	MORE INFO.
Chat Logs	1100 chat logs	715 MB	U	Article - Tarique Anwar & Muhammad Abulaish	2010 - 2012	+
Leaked Passwords	~ 30 sets	N/A	U	Skull Security Wiki	2009 - 2010	
Media (Pictures)	10,074 images	N/A	E	BOSS - Break Our Steganographic System	2010	+
Media (Pictures)	> 10 images	N/A	E	King Saud University - Image Forensics - Note: This website is no longer available.	2010	+

datasets.fbreitinger.de

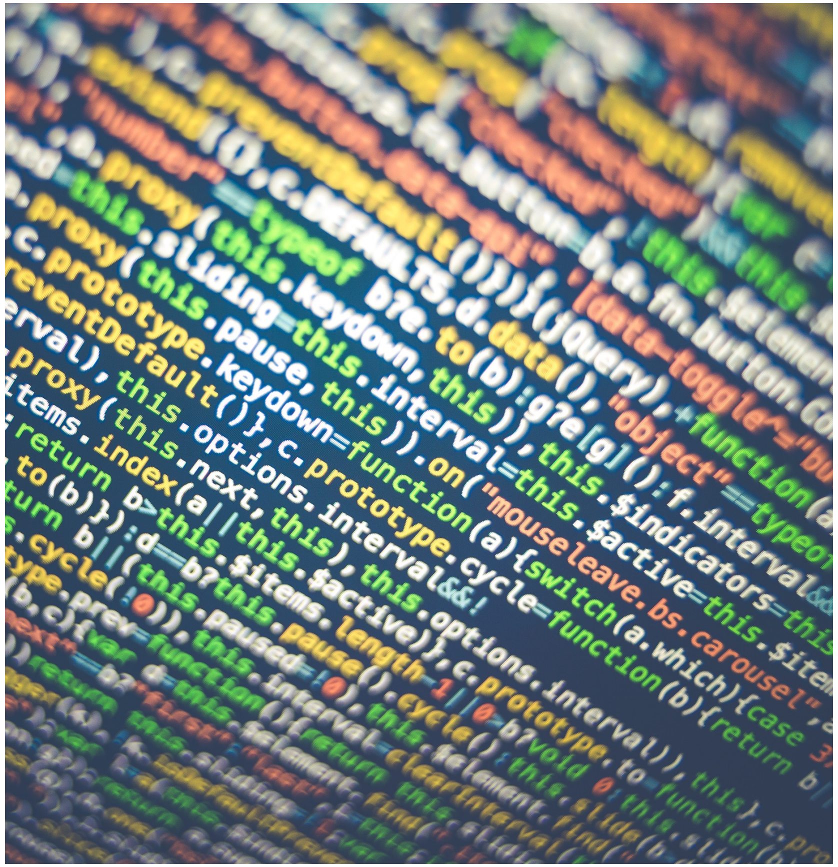
Current situation – Data generation

Data generation frameworks such as TraceGen or ForTrace

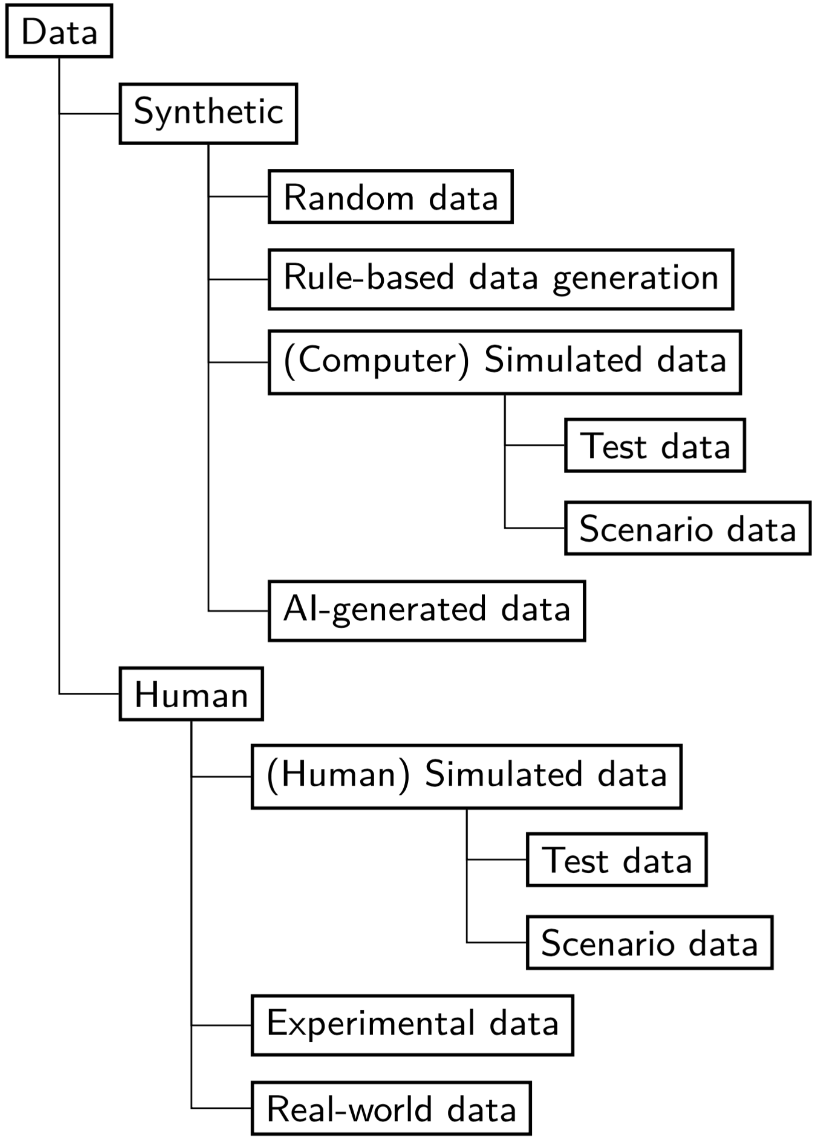
Data categories:

- Test data, sampled data, realistic data, real and restricted data, and real but unrestricted data (Garfinkel et al 2009)
- Real world vs. synthetic (Yannikos et al. 2014)
- Experiment-generated, user-generated, computer-generated (Grajeda et al. 2017)

Where our taxonomy starts



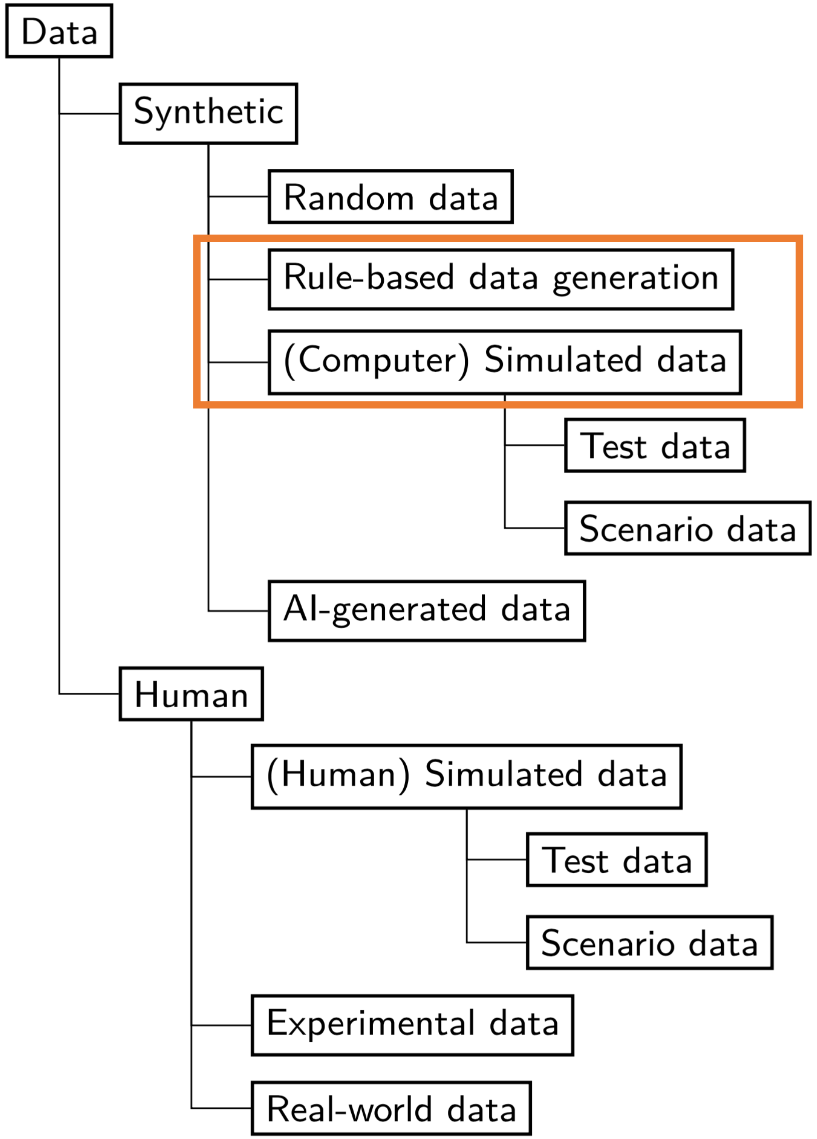
Forensic dataset taxonomy



- Data
  - Synthetic
    - Random data
    - Rule-based data generation
    - (Computer) Simulated data
      - Test data
      - Scenario data
    - AI-generated data
  - Human
    - (Human) Simulated data
      - Test data
      - Scenario data
    - Experimental data
    - Real-world data

Rule-based data generation (a.k.a. generation) follows strict rules and thus process is often deterministic (i.e., outputs will have identical hashes)

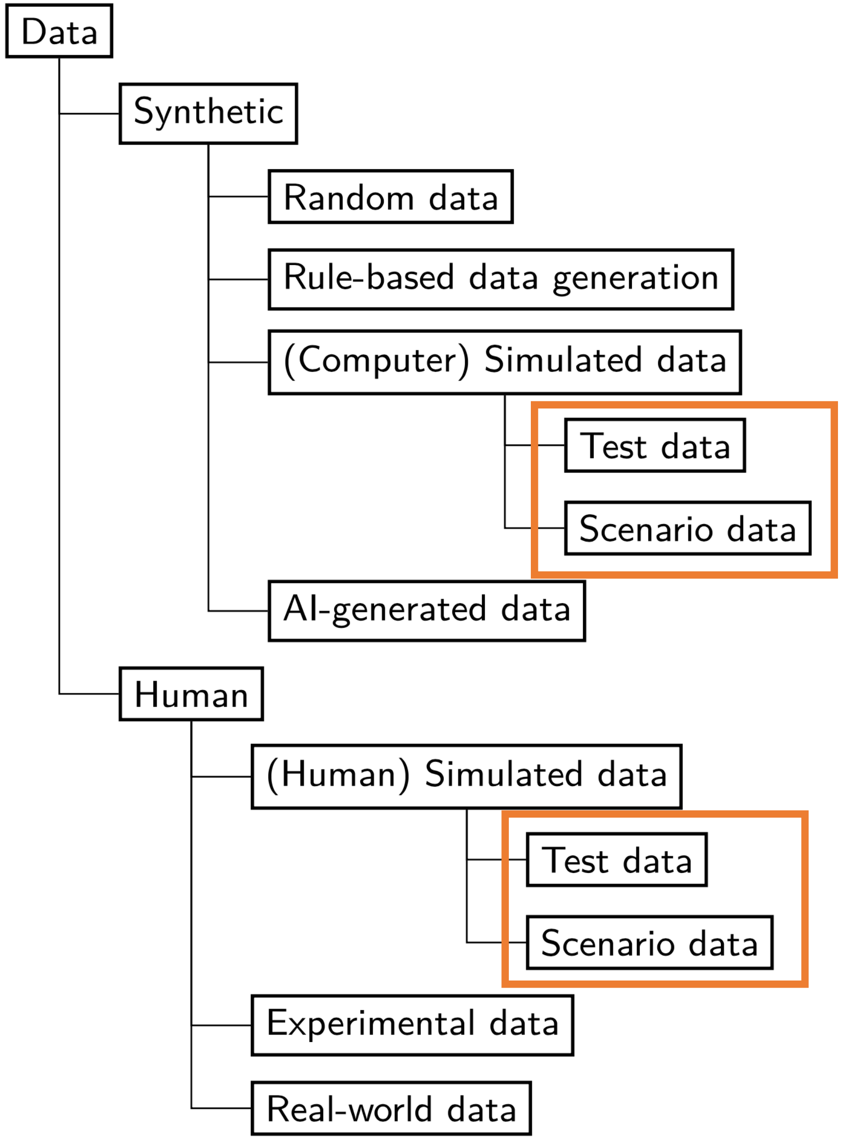
Simulation is often non-deterministic (ground truth is vaguer); simulators rely on other functionality that they call in sequence



Test vs. scenario

One may only produce specific data (e.g., a network capture) or complex scenarios comprising several different artifacts (e.g., network capture, memory dump, and disk images)

Test data: only one type (format) is returned



```
{ ”description”: {
    ”organization” : ”unstructured”,
    ”origin” : ”Real-world data”,
    ”content-tags” : [”memory”, ”windows xp”, ”image”]
   }
  ....
}
```

Example in JSON format
… increases findability

How to use it?

Varies between countries…

Not my area of expertise…

Legal barriers to sharing data

Taxonomy supports understanding legal barriers

Data ownership, control, and sharing restrictions

01

Data may be protected by copyright or another special law (e.g., patent)

02

Contractual provisions may restrict how the data can be shared

03

Privacy or data protection laws may impose rules and restrictions

Copyright & Intellectual Property Rights

- Copyright is **automatically granted** and gives the copyright holder moral and exclusive property rights for a relatively long period (50+ years)

- Work created without sufficient human intervention will not be protected by copyright → **no copyright on synthetic data**

- Use and sharing of copyrighted materials may, be legally permitted in some circumstances without consent, e.g., ‘**fair use**’ doctrine

Personal data

- Personal data is **information relating to an identified or identifiable natural person**
  - Data may be anonymous information for one entity, but personal data for another

- **Anonymization and data pseudonymization**: processing personal data that it can no longer be attributed to a specific data subject
  - Anonymization → irreversibly removed
  - Pseudonymization → reversibly removed (e.g., key, rule)

Data can be shared unless it contains personal data, is protected by a special law, or is subject to contractual restrictions

Any **synthetic data** and **human simulated** data generated by research can generally be shared without legal barriers

Under what circumstances can datasets be shared? (RQ2) https://open-power-system-data.org/legal#:~:text=If%20data%20are%20published%20without,stating%20“all%20rights%20reserved”

Sharing restricted content

Can I use data found on the Internet?

Default: “all rights reserved”

Copyright only applies to human-created data

likely unknown how the data was created

Researcher: Ask for permission

Creator: Add a license when published

Sharing data under GDPR

Experimental data: may contain personal data due to human error

advise: ask for consent/authorization from third parties

Real-world data: if consent can be obtained data can be shared

Can we anonymize real-world data? **No**

Take home messages

Sharing datasets is essential to progress and to allow the comparison of results

Common language, taxonomies, classifications allow granular searches and thus contribute to findability and FAIR

Based on the origin (taxonomy), we highlighted legal considerations and conclude that the dataset creator should obtain/provide consent and be careful with special laws protecting data, e.g., copyright or licensing

Questions

Thank you!

vCard

Frank Breitinger

Frank.Breitinger@unil.ch