



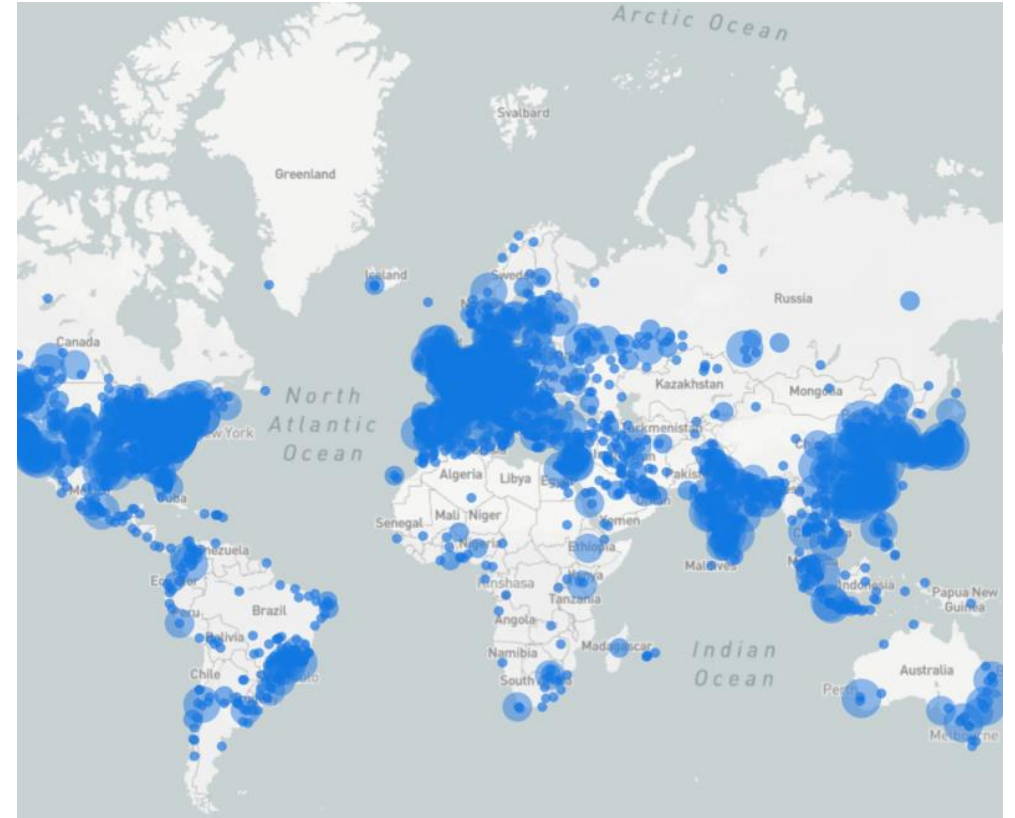
CASSE: Targeted Threat Modeling for Data Management Libraries

Keegan I. H. Sanchez¹, Suren Byna¹, Zhiqiang Lin¹, David Mattson²

¹ The Ohio State University

² Amazon Web Services

- Self-describing high-level data management libraries are used heavily
 - Portability
 - Metadata management
 - Performance
- Examples: HDF5, NetCDF, ADIOS, Zarr
- Numerous areas of science use them



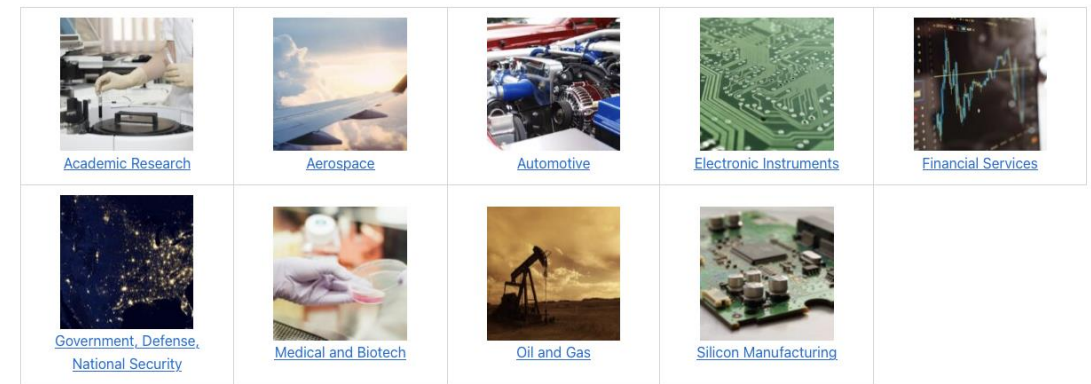
HDF5 downloads from around the world

Data Management Libraries (DML) used heavily

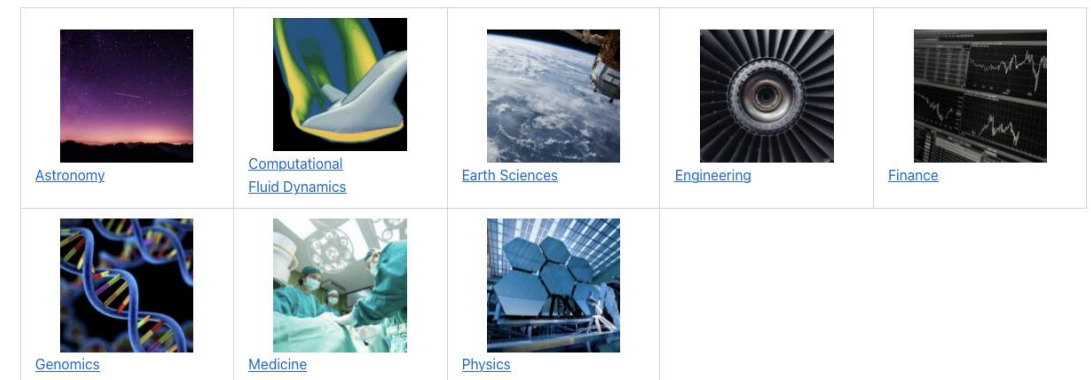
- Self-describing high-level data management libraries are used heavily
 - Portability
 - Metadata management
 - Performance
- Examples: HDF5, NetCDF, ADIOS, Zarr
- Numerous areas of science use them

Who Uses HDF®?

Industries



Scientific Fields





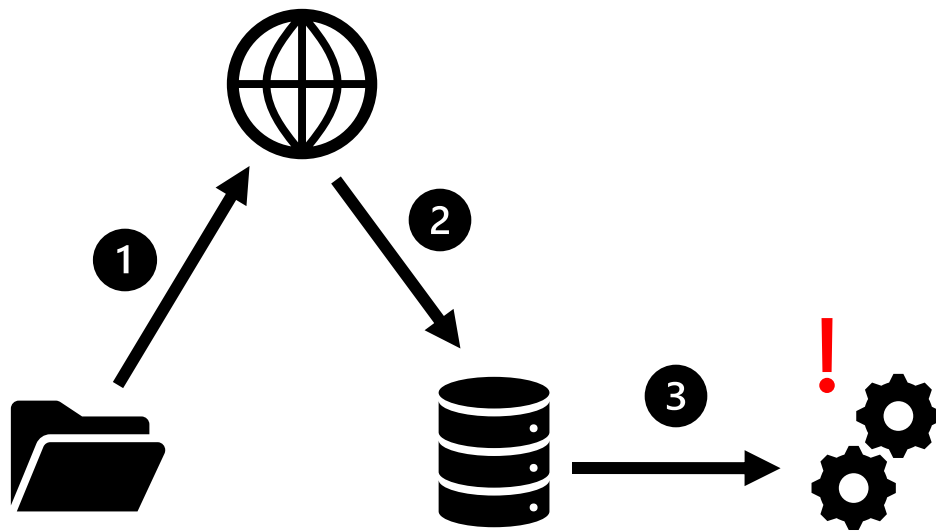
DMLs Developed Before Cybersecurity Concerns

- Numerous vulnerability and attack targets
 - Code
 - Data
 - Metadata
 - External plugins
 - Other data management systems and wrappers
- Lack of a systematic assessment of threats and their impacts
 - How do we increase security?
 - How do we find vulnerabilities?
 - How do we define potential impacts?

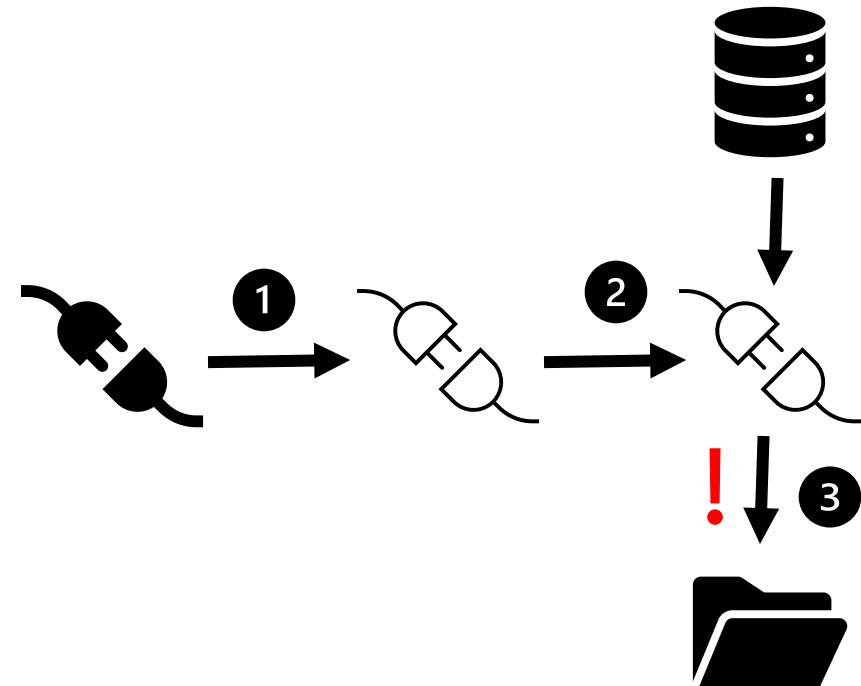


Some Attack Examples

Scenario 1: Poisoned ML Training Data



Scenario 2: Altered Compression Plugin





Current Security Approaches

- **Libraries**
 - Vulnerability Libraries
 - Attack libraries
- **Threat Modeling**
 - STRIDE
 - PASTA
 - LINDDUN
- Can we apply pre-existing approaches?

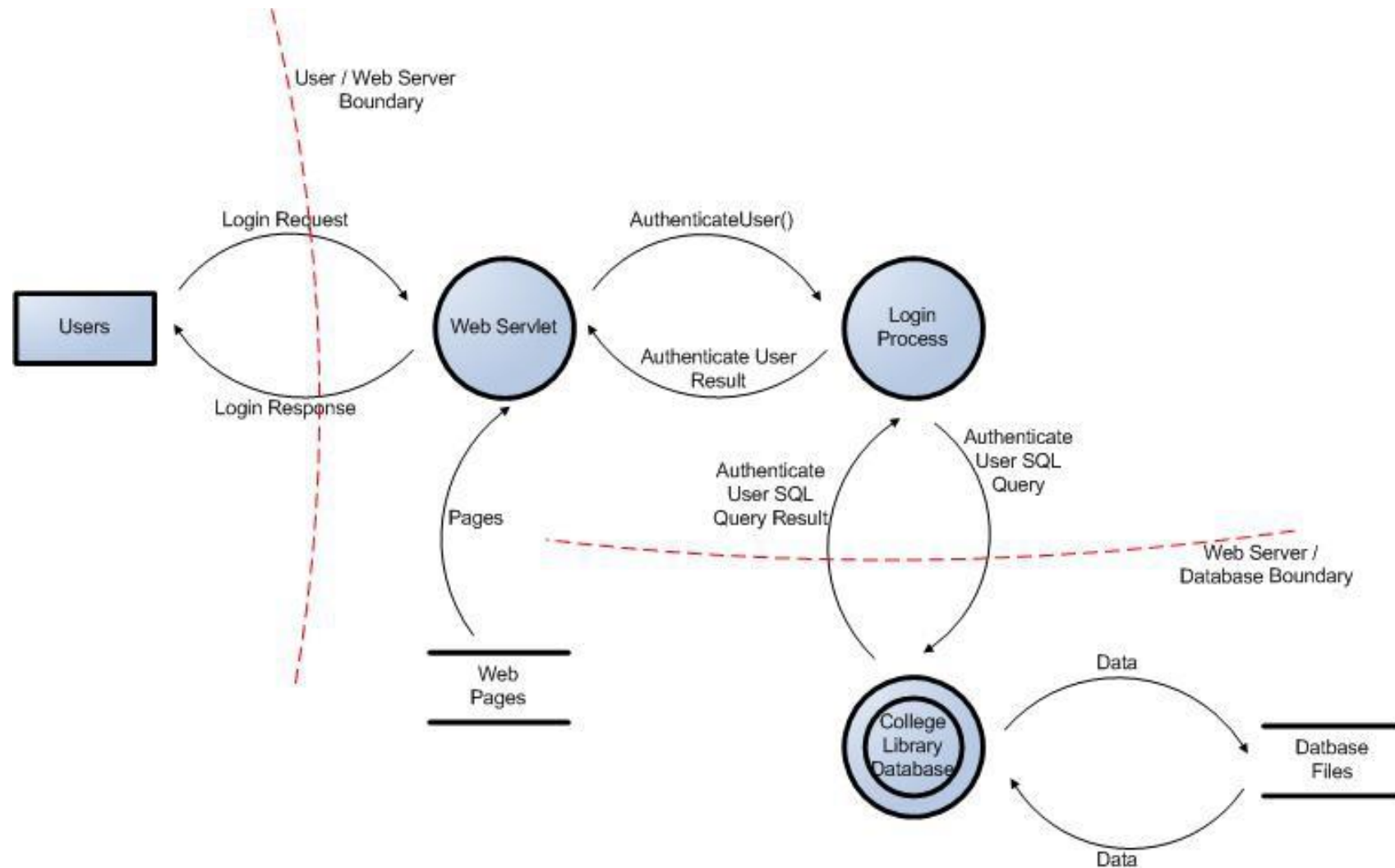


STRIDE in Depth

- Combines *data flow diagram* with *attack taxonomy*
- **Spoofing**
 - Masking identity
- **Tampering**
 - Modifying data
- **Repudiation**
 - Denying responsibility
- **Information Disclosure**
 - Leaking data
- **Denial of Service**
 - Preventing resource usage
- **Elevation of Privilege**
 - Obtaining restricted privileges



What are Dataflow Diagrams?

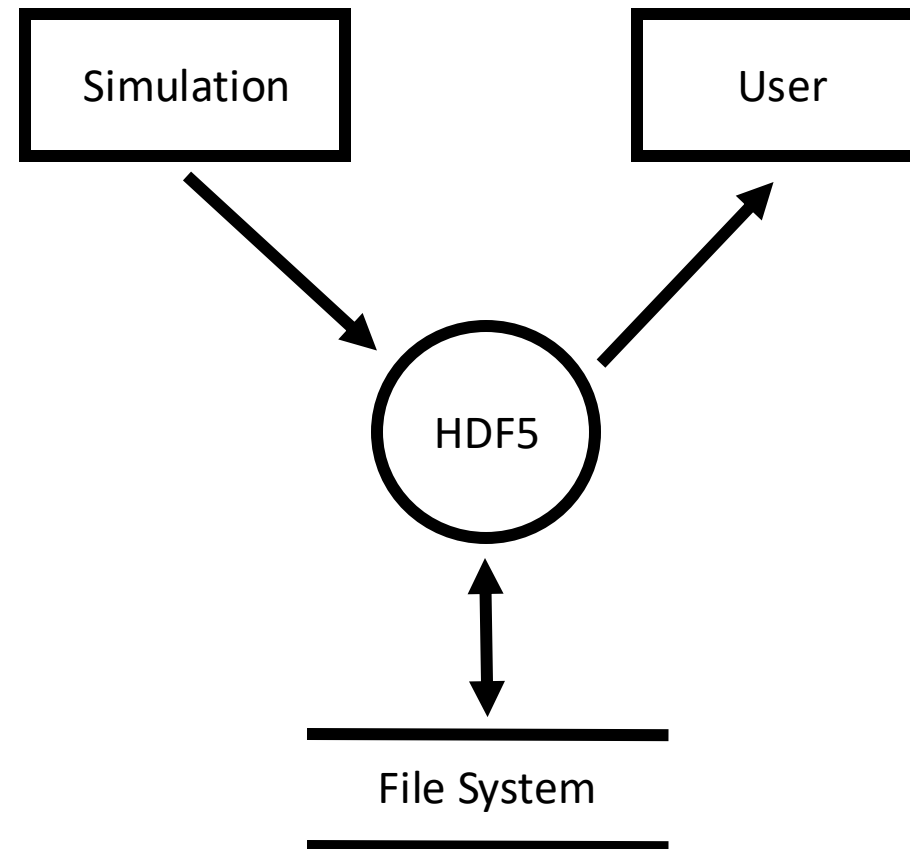


Data Flow Diagrams highlight how data passes through a system.



Applying STRIDE to a DML – An Example

- Spoofing
 - Spoof the user?
- Tampering
 - Alter during pass through?
- Repudiation
- Information Disclosure
 - Leak stored information?
- Denial of Service
 - Crash HDF5?
- Elevation of Privilege





STRIDE and the example attacks

- Scenario 1: Poisoned ML training files
 - Elevation of privilege? Spoofing?
- Scenario 2:
 - Information Disclosure
- Shortcomings
 - Doesn't fully cover examples
 - Categories don't aid in discovering cause
 - Categories don't aid in finding similar attacks



Shortcomings of Existing Threat Models

- Categories provide little context
 - Want to highlight how attacks are performed
- Categories are not specific to DMLs
 - Some categories are over-represented, some are under-represented
- Lack data modeling component
 - DMLs include complex data
 - Can cause vulnerabilities



Requirements of DML Threat Model

- Add *data modeling* and *taxonomy* in addition to data flow diagrams
- Taxonomy – Cover attack surfaces
 - Sources to be attacked
 - Methods of an attack
 - Targets of an attack
- Data Modeling
 - Aid developers in understanding sometimes complex structures
 - Highlight vulnerabilities



CASSE – A Threat Model focused on DMLs

Three parts highlight the origin of the attack, source, method and target.

- Sources

- Data
- Library

- Methods

- Modification
- Poisoning

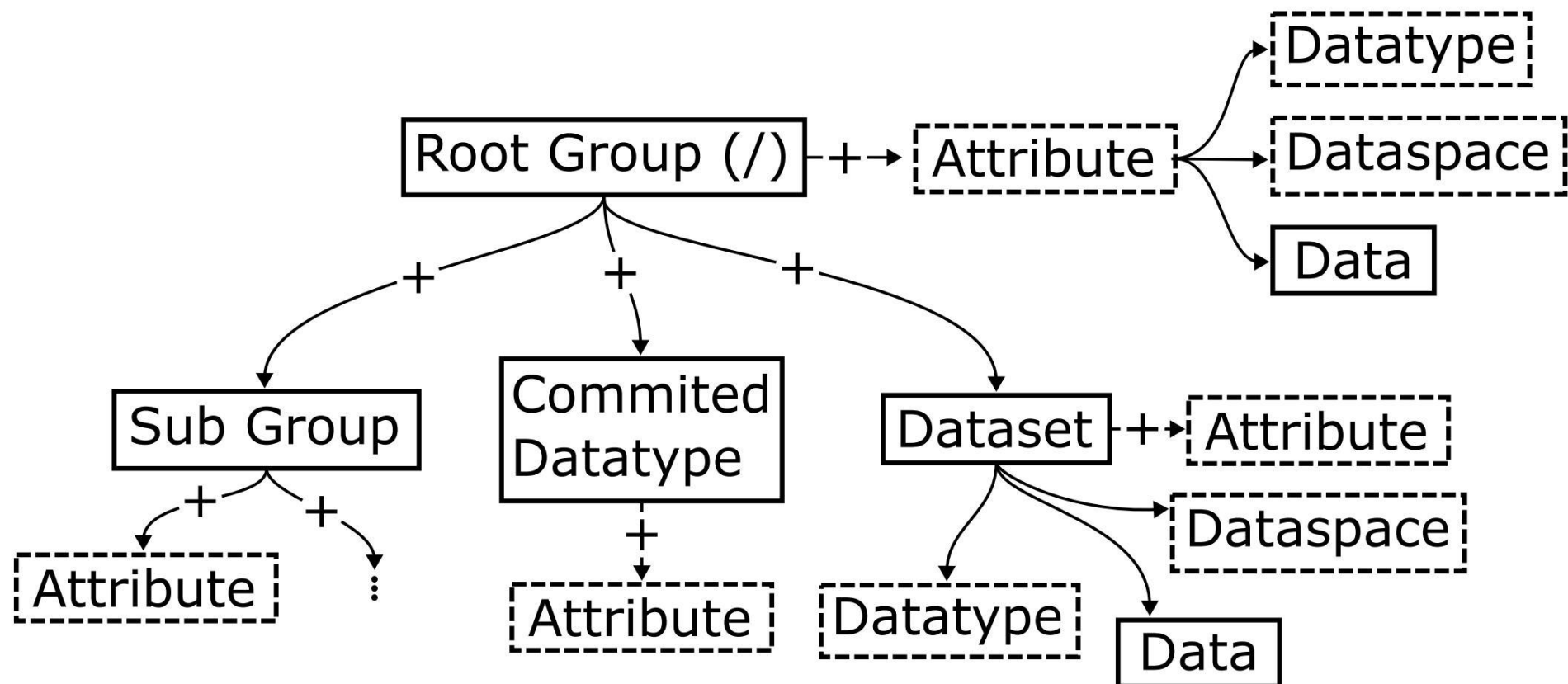
- Targets

- Core Library
- Application
- Storage
- System
- External Library



Modeling Data

HDF5 abstract data model



Developers should be aware of *offsets*, *sizes*, and *pointers*.



On Relevant Vulnerabilities

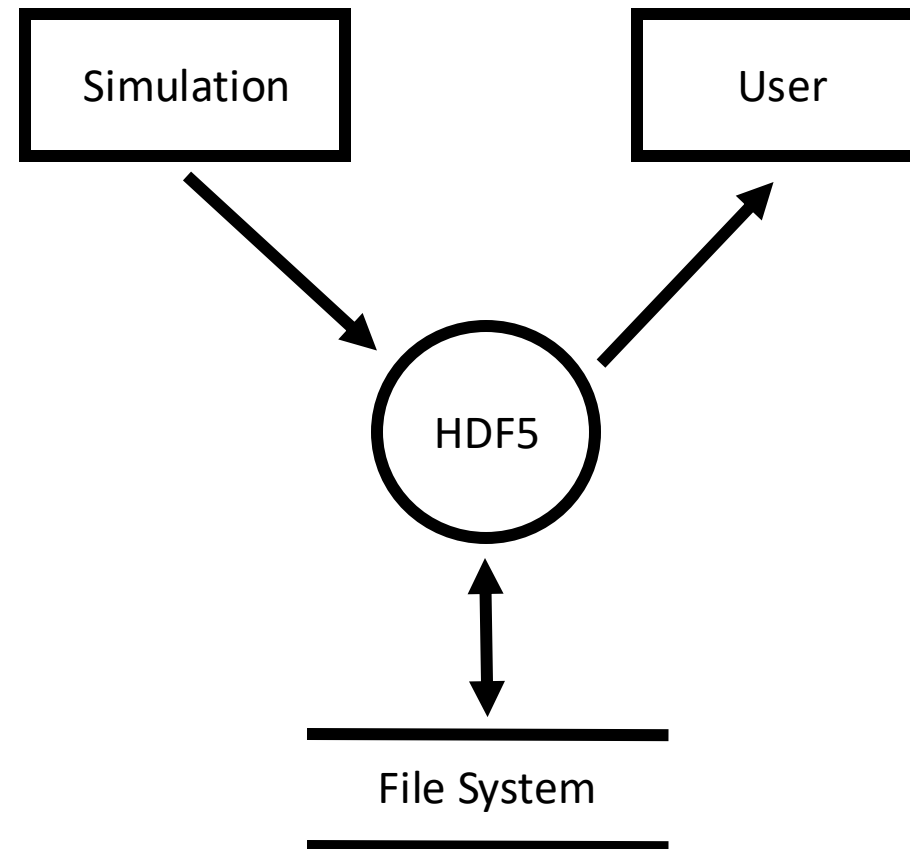
Name	CWE ID	Impact
Out-Of-Bounds Write	CWE-787	●
Out-Of-Bounds Read	CWE-125	○
Use After Free	CWE-416	●
Null Pointer Dereference	CWE-476	○
Integer Overflow or Wraparound	CWE-190	●
Specified Quantity	CWE-1284	●
Specified Index Position	CWE-1285	●
Syntactic Correctness	CWE-1286	●
Specified Type	CWE-1287	●
De-serialization of Untrusted Data	CWE-502	●

Though there is less research into DML based systems, we believe these are common in practice.



Applying CASSE to an Example

- Develop data models for relevant DMLs
- Iterate attack taxonomy and organize by likelihood and impact.





CASSE on attack scenarios

- Scenario 1: Poisoned ML dataset
 - Data Poisoning attack targeting the Core Library
- Scenario 2: Altered compression plugin
 - Library Poisoning attack targeting External Library
- Benefits
 - More detail about where attacks are
 - Provide clear routes for mitigations
 - Highlights similar attacks



Conclusion

While previous threat modeling techniques are general, CASSE directly targets threats surrounding DMLs.

- **Data Models**

- Aid in understanding DMLs
- Highlight vulnerabilities in data structures

- **CASSE Taxonomy**

- Applies directly to DMLs and surrounding systems
- Highlights how attacks are performed

- **Future work**

- Developing a quantification method for outlining severity
- Apply CASSE across more libraries and systems

Thanks to:

NSF CICI program



Project: S2-D2:
Securing Self-
describing Data,
Formats, and
Libraries

Contact: Suren Byna
byna.1@osu.edu