Labeling Software Security Vulnerabilities

Irena Bojanova, NIST; John Guerrerio, Dartmouth College; Eduard Pinconschi, University of Porto

Motivation

- Crucial need for systematic comprehensive labeling of the more than 228 000 publicly disclosed cybersecurity CVE vulnerabilities to enable advances in modern AI cybersecurity research.

Objective

- Utilize the Bugs Framework (BF) formalism for BF-CWE-CVE mappings.

CWE Labels for CVEs

- NVD, with input from security community, labels CVEs with CWEs
- Challenging, as CWEs can be too specific, ambiguous, or overlapping – e.g., CVE-2023-38435 is labeled CWE-787, while CWE-121 (Stack-based Buffer Overflow) fits it better

CWE-ID	CWE Name
CWE-787	Out-of-bounds Write

- As AI labels, CWEs require extra processing of their unstructured textual information to determine causes, operations, consequences, attributes, and their types, if possible at all.

CWE2BF & CVEs Pre-Annotation

- BF's formalism allows specifying each CWE as a (cause, operation, consequence) weakness triple or a chain of such weakness triples
- We focus on the 60 memory-related CWEs, as a vast number of memory-related CVEs (approx. 61 000) are mapped to them – 48 distinct BF weaknesses or chains of weaknesses
- 7 CWEs share a BF causing chain with other CWEs – e.g.,

CWE-127		(Under Bounds Pointer, Read, Buffer Under-Read)
CWE-786	*(Erroneous Code, Calculate, Wrong Result)* ↪ ↪ *Wrong Index, Reposition, Under Bounds Pointer)*	(Under Bounds Pointer, Read/ Write, Buffer Underflow/ Buffer Under-Read)
CWE-124		(Under Bounds Pointer, Write, Buffer Underflow)

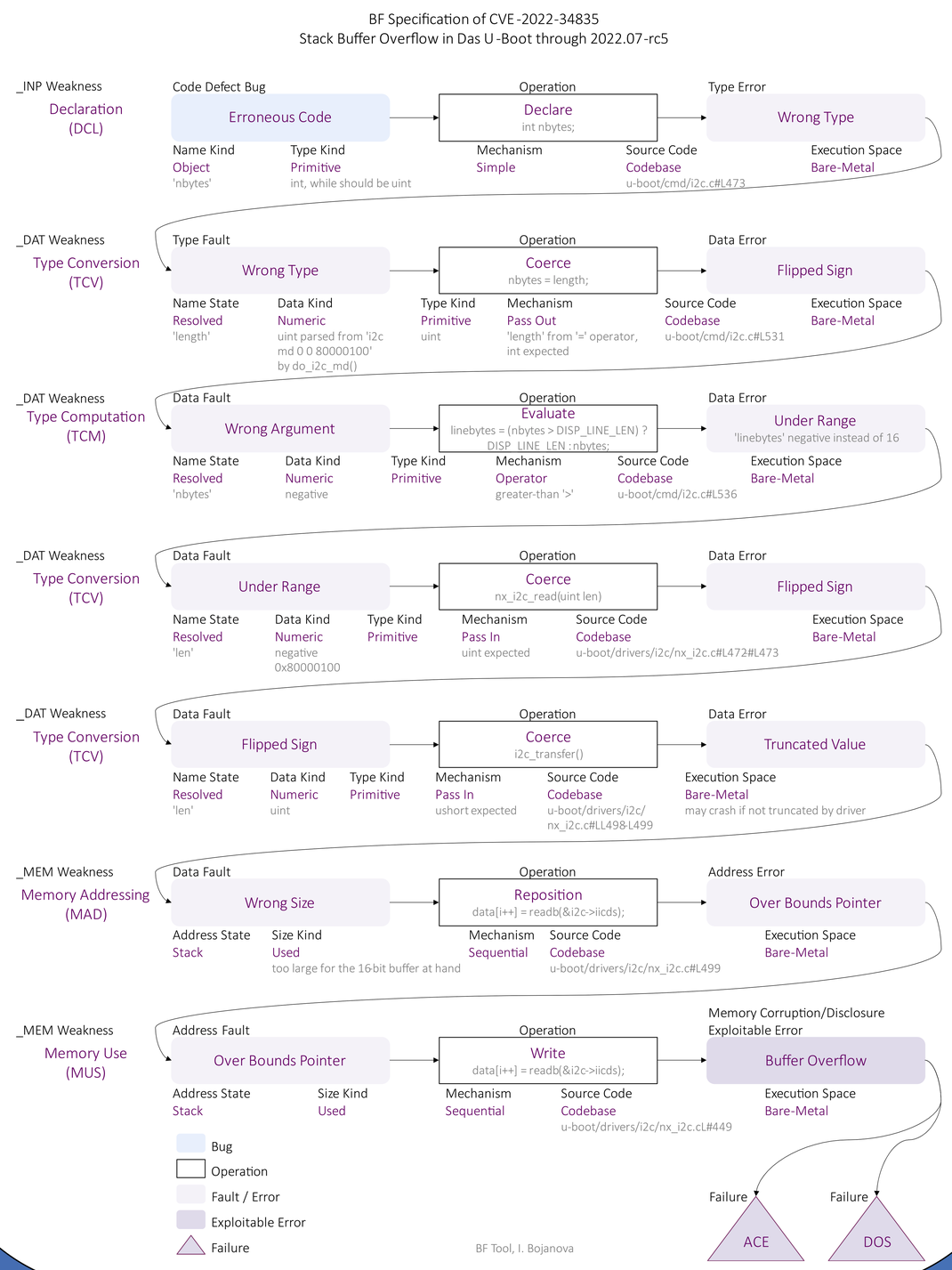
- 11 groups of CWEs map to the same BF weakness triple(s) – e.g.,

CWE-415 CWE-1341	(Erroneous Code, Deallocate, Double Free)

- CWE2BF mappings at least partially fit many CVEs and can aid in their annotation
- CWEs' main weaknesses are mostly correct by operation and final error
- CWEs' identifiable causing weaknesses can be overly specific and sporadic – e.g., CVE-2023-38435's cause is Erroneous Declare followed by other type-related weaknesses, while its assigned CWE-787 lists only Erroneous Calculate as a causing weakness
- CWEs not assigned to any CVEs, may fit better than the assigned CWEs with identical BF triple(s), which is indicative of ambiguity – e.g., CVE-2023-38434 is mapped to CWE-415 (memory-related double free) but better fits CWE-1341 (multiple releases of same resource).

BF Labels for CVEs

- Bugs Framework (BF) allows formal unambiguous specification of CVE vulnerabilities
- BF's rich, precise, unambiguous set of tokens for types and values of bugs, faults, errors, weaknesses, etc. can be used as comprehensive AI labels without additional processing – e.g., the BF CVE-2023-38435 specification's value labels (in purple) and type labels (in black):



Potential Impact

- Comprehensively labeled BF CVE datasets for cybersecurity research, education, and guidance
- CWE2BF specifications for use in software testing tool reports.