

Software Fingerprinting for Supply Chain Security

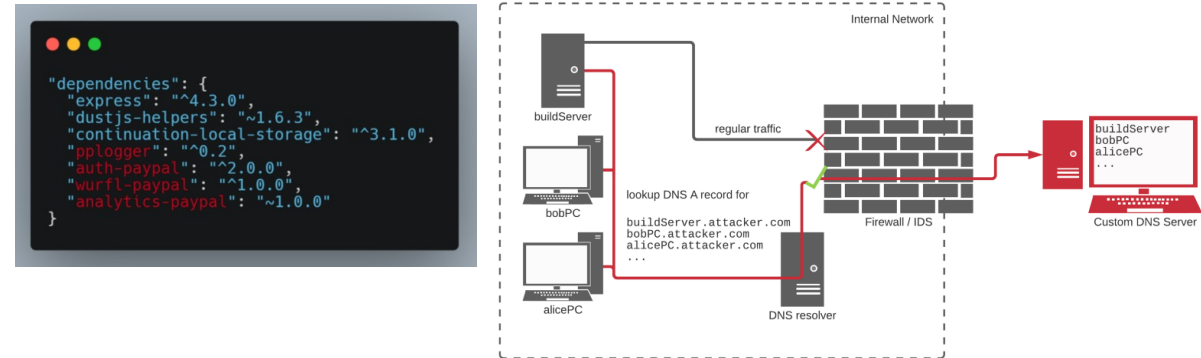
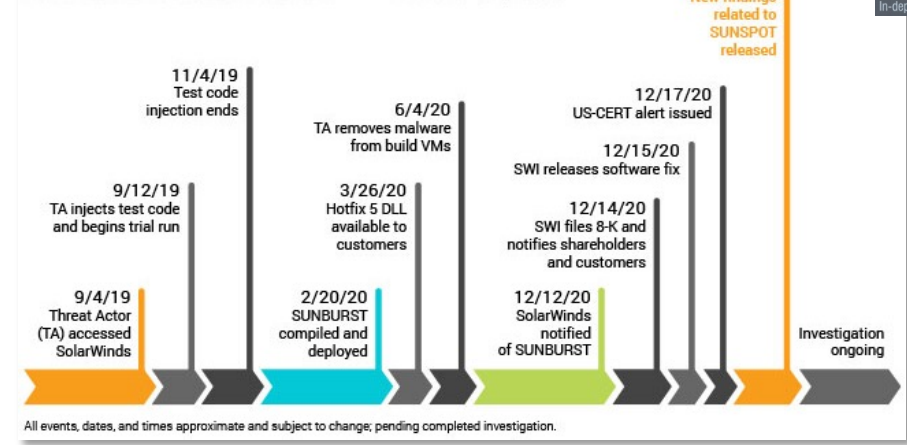
IBM Research Security

—
Ted Habeck
Jiyong Jang
Dhilung Kirat
Ian Molloy
JR Rao
Doug Schales

Supply Chain Attacks

- SolarWinds (2019-2021) **est. cost > \$100B**
 - Malicious code (backdoor) pushed out through updates
- Dependency confusion (Feb 2021)
 - Private vs public packages (npm, PyPi, RubyGems)
- Codecov (Apr 2021)
 - DevOps tool. Vulnerability in CI. Bash uploader modified
- Kaseya (Jul 2021) **ransom \$70M**
 - IT solutions, including VSA (remote monitoring and management software) to deliver REvil ransomware
- Protestware (Mar 2022)
 - Popular NPM package wiped files in Russia and Belarus

Attack Timeline – Overview



Codecov breach impacted 'hundreds' of customer networks: report

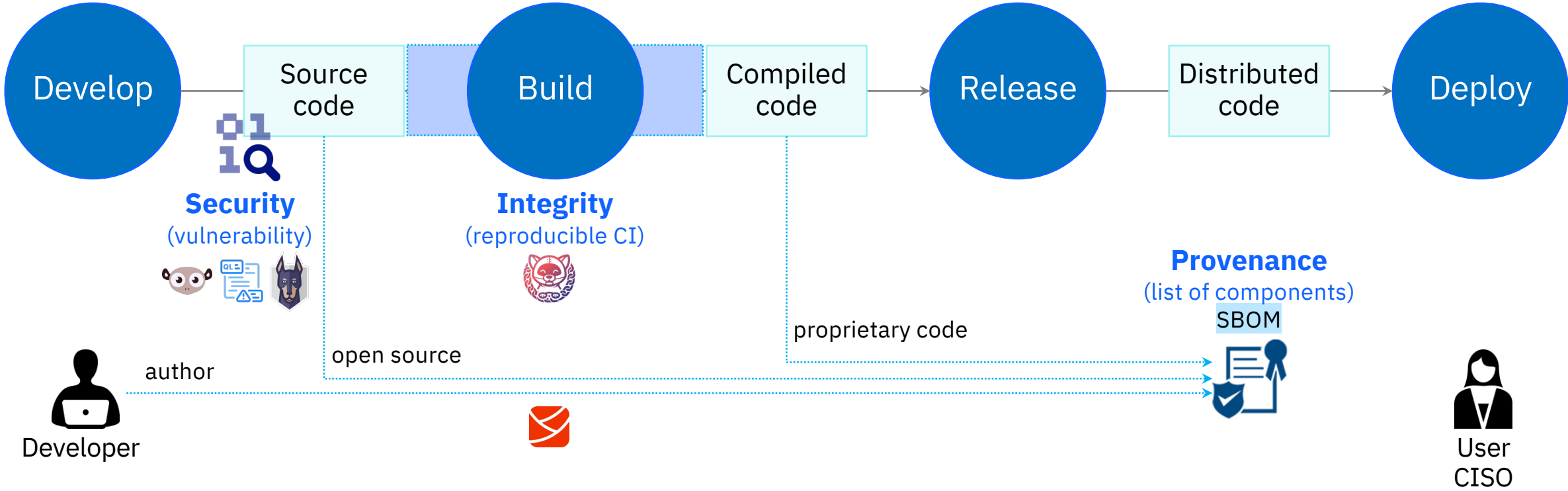
Updated: Reports suggest the initial hack may have led to a more extensive supply chain attack.



CISA-FBI Guidance for MSPs and their Customers Affected by the Kaseya VSA Supply-Chain Ransomware Attack

Supply Chain Security

Industry approach to protecting CI/CD pipelines



Proposal: Software Fingerprints for Software Assurance

- Hashing (e.g., SHA2) verifies *exact* matches in files
 - One-bit changes result in large differences
- Signatures verify integrity assuming root of trust
 - Requires full trust in signer
- Fuzzy Hashing (e.g., **ssdeep**) provides partial matches to known files
 - Handles minor syntactic differences but lacks file semantics

Ken Thompson's *Reflections on Trusting Trust*

- There is a need to verify code and binary integrity and identity
 - Sourced from multiple distributions, compilers, and optimizations
- Identify and verify legacy software in deployments
- Verify software version or backported patches, or code functionality

Continuity across package repositories

ssdeep 2.14

- 1536:HIbwmHxpJjkCt4RxeRGVXE3jETfgx1VH6CTM:HwqJ4Ejgx1VH6CTM
- 768:n5fTR1HCcCaMR+F75fwBcVEpqqnfuAxxXHYjvriZRTqI33oF5Wf+lqhmunaEYcu/:n5tXN4/MufHXHmvrkqII5Q
- 768:XXqchVhV1pGPCNazQSA8W5ie12QSN6CDvrs91c:nqilhszQSUtclQEvrIi

Stability across versions

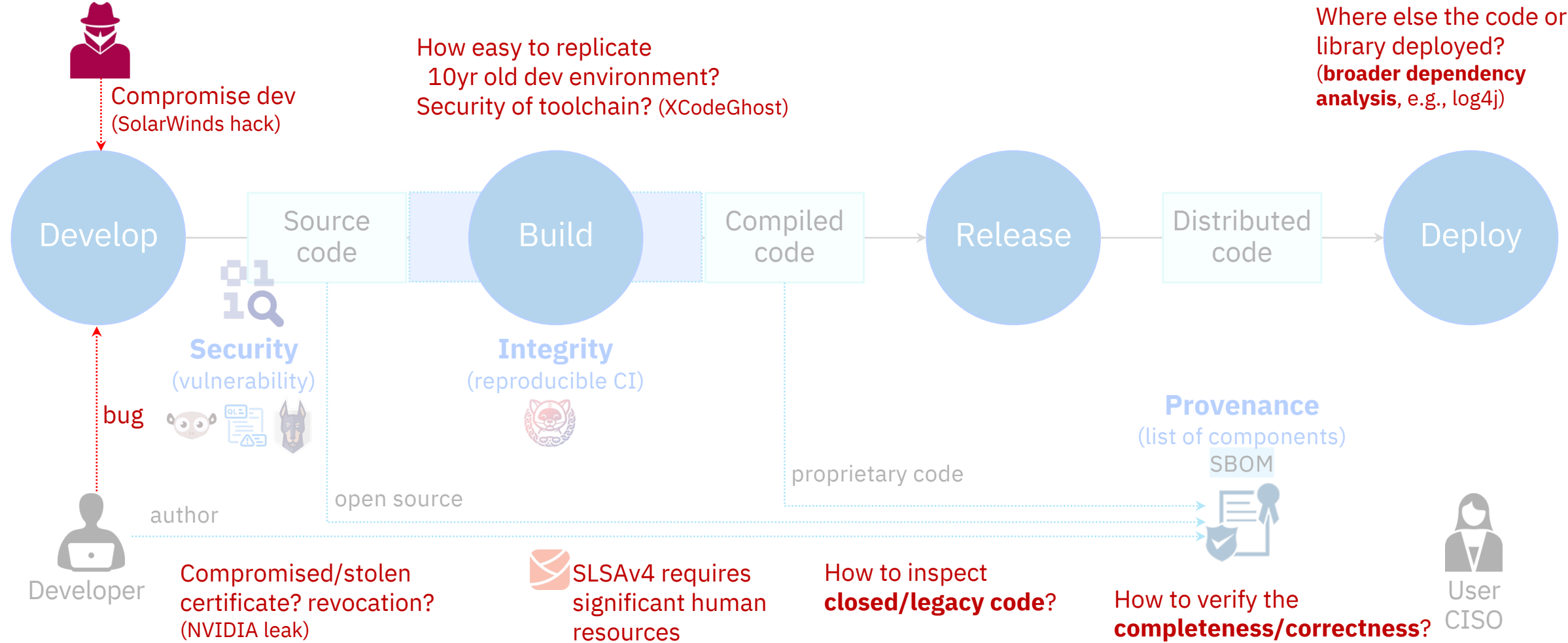


Robustness for legacy



Supply Chain Security

Open security issues and residual risks

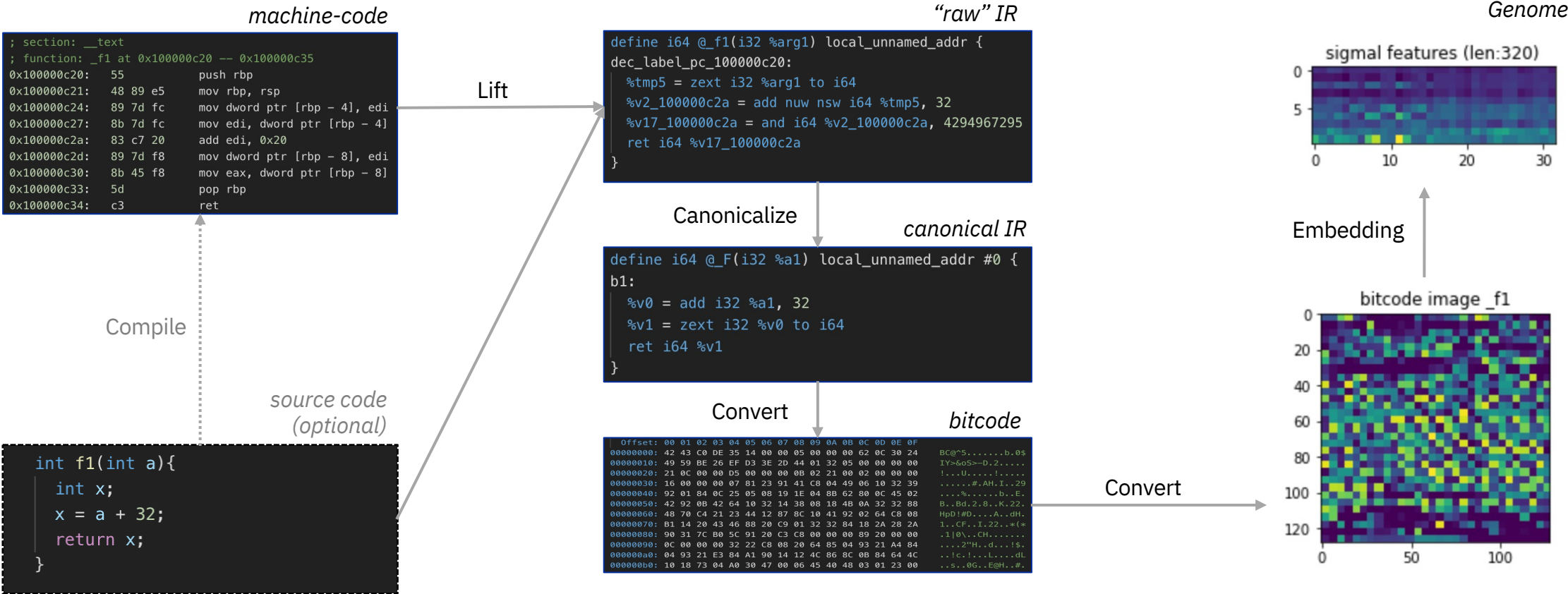


Code Genome: Semantically meaningful fingerprint

- Across multiple *architectures* (x86, ARM, ...)
- Across multiple *compilers* (gcc, clang, ...)
- Across multiple *optimization levels*
- Handling *obfuscation*



Key Idea: Code Genome construction



Genome can be constructed from *closed-source/legacy code* where source code is not easily available.

Use Case 1: Finding Log4j

Legacy Software Discovery

- Software deployments are a turducken
 - `zip`, `tar`, container image, `jar`, etc.
 - Dependencies often wrapped up
- No good provenance or CMDB
- Can't rely on standard directories, filenames, or hashes
- Code can be repackaged

- Applies across the board: CICD, DevSecOps and Legacy



Use Case 2: SBOM Verification

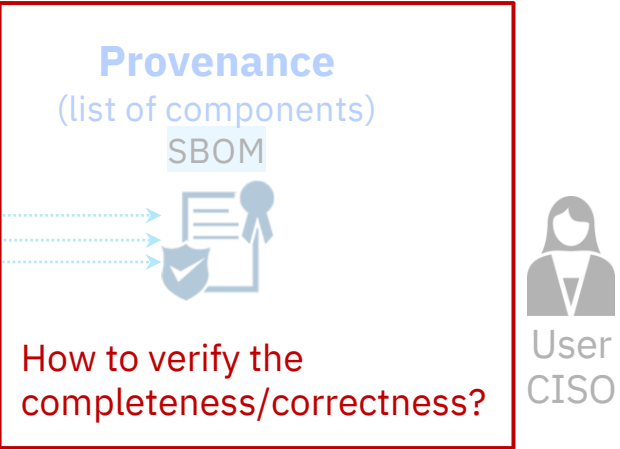
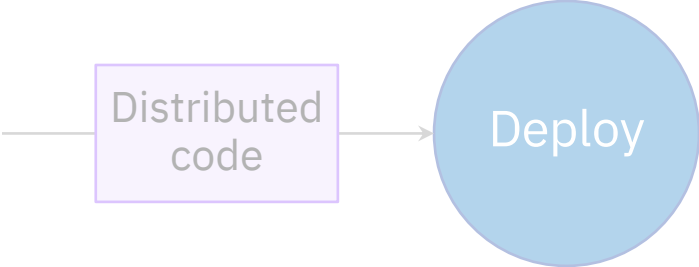
– Problem

- Each vendor creates SBOM of their own software including open-source and closed-source components. How can we verify its correctness (containing incorrect library mistakenly/maliciously) and completeness (missing library)?

– Value

- Given software, we can verify (generate) SBOM
- Support closed-source and legacy software without requiring source code access
- Help developers generate correct SBOM
- Vet software before integrating/deploying into a product

Where else the code or library deployed?
(broader dependency analysis, e.g., log4j)



Trust but Verify SBOM: Metadata vs. Code



CycloneDX SBOM Standard
CycloneDX is a modern standard for the software supply chain. SBOM, SaaSOM, OBOM, Advisories, VEX, and more. CycloneDX is a OWASP Flagship Project.

OWASP <https://cyclonedx.org/> @CycloneDX_Spec Verified

“Unfortunately, some images – such as the [official node image on Docker Hub](#) – incorrectly report the version of OpenSSL that's used by the Node.js runtime.”

```
# Syft detects the image contains OpenSSL 1.1.1 (not vulnerable):
$ syft node | grep openssl
openssl                               1.1.1n-0+deb11u3                deb







# But the node runtime actually uses the vulnerable OpenSSL 3.0.5!
$ docker run node -e "console.log(process.versions)" | grep openssl
openssl: '3.0.5+quic',

# The Chainguard Image for node correctly reports the version of OpenSSL it uses,
# due to its build-time SBOM.
$ cosign download sbom --platform=linux/amd64 cgr.dev/chainguard/node | grep libssl
libssl3-3.0.5-r3

# Syft doesn't have to detect what's in the image.
$ syft cgr.dev/chainguard/node | grep libssl
libssl3                               3.0.5-r3                        apk
```

<https://www.chainguard.dev/unchained/mitigating-critical-openssl-vulnerability-with-chainguard>

Pinned

 specification Public Software Bill of Material (SBOM) standard designed for use in application security contexts and supply chain component analysis XSLT ☆ 168 📄 34	 cyclonedx-dotnet Public Creates CycloneDX Software Bill of Materials (SBOM) from .NET Projects C# ☆ 82 📄 41
 cyclonedx-python Public Creates CycloneDX Software Bill of Materials (SBOM) from Python projects and environments. Python ☆ 91 📄 35	 cyclonedx-maven-plugin Public Creates CycloneDX Software Bill of Materials (SBOM) from Maven projects Java ☆ 125 📄 37
 cyclonedx-node-module Public Creates CycloneDX Software Bill of Materials (SBOM) from Node.js projects JavaScript ☆ 86 📄 67	 cyclonedx-cli Public CycloneDX CLI tool for SBOM analysis, merging, diffs and format conversions. C# ☆ 85 📄 21

This project provides a runnable Python-based application for generating CycloneDX bill-of-material documents from either:

- Your current Python Environment
- Your project's manifest (e.g. Pipfile.lock, poetry.lock or requirements.txt)
- Conda as a Package Manager

\$ sbom generation tools

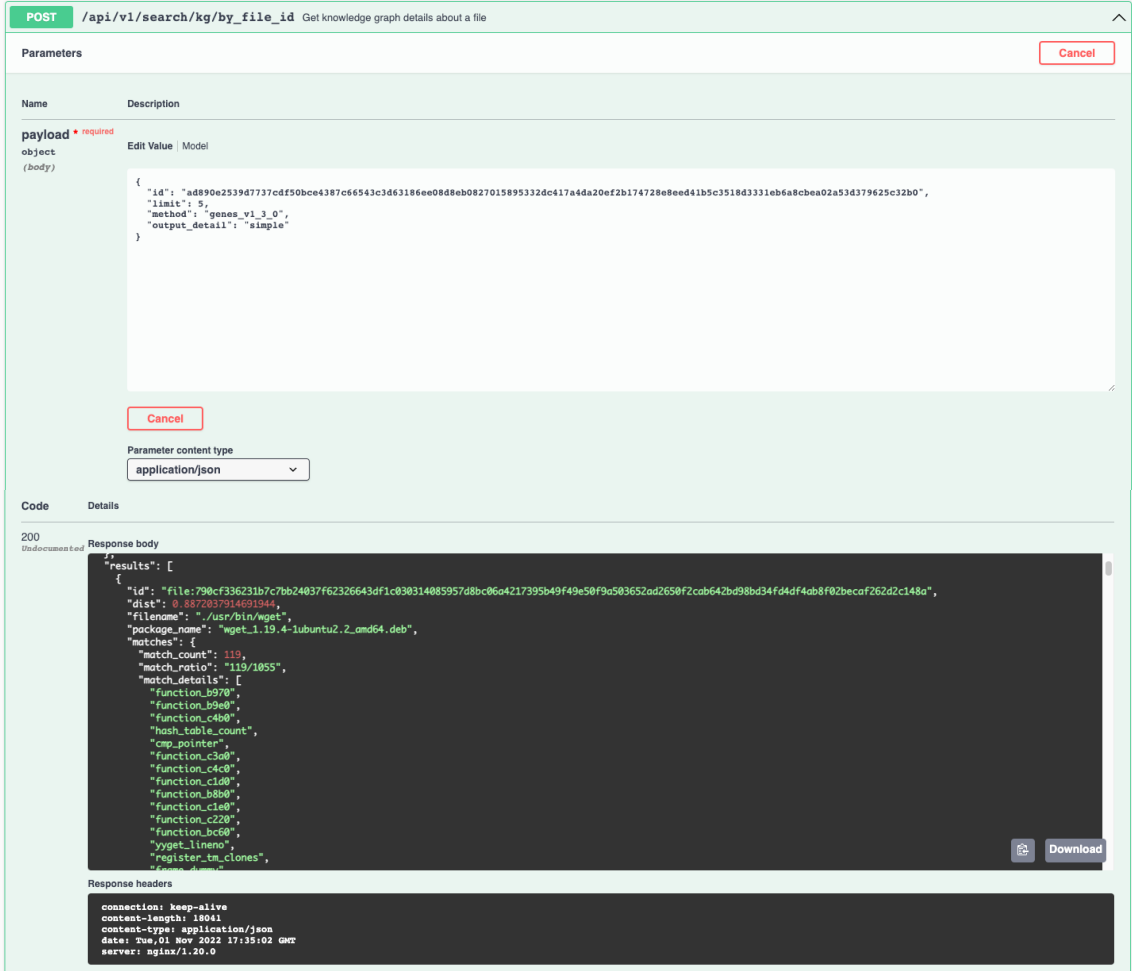
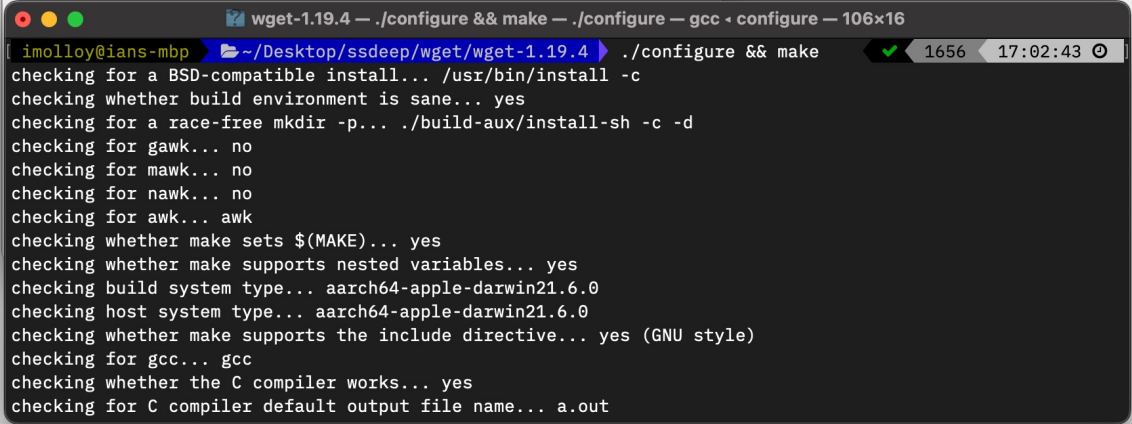
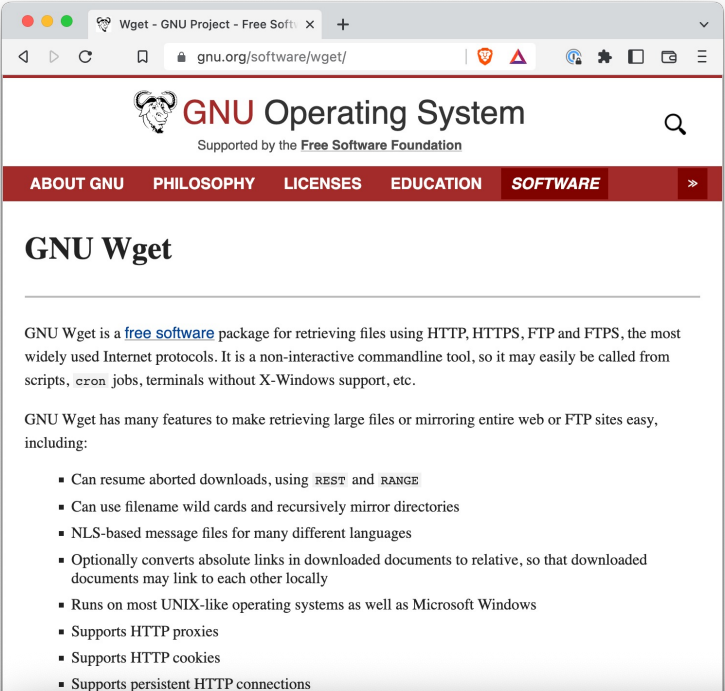
```
Dockerfile > ...
1 FROM ubuntu:focal
2
3 RUN apt-get update
4 RUN apt-get install -y wget
5
6 RUN apt-get update
7
```

```
"bom-ref": "pkg:wget@1.20.3-1ubuntu2",
"type": "library",
"name": "wget",
"version": "1.20.3-1ubuntu2",
"licenses": [
  {
    "license": {
      "name": "UNKNOWN"
    }
  }
],
```

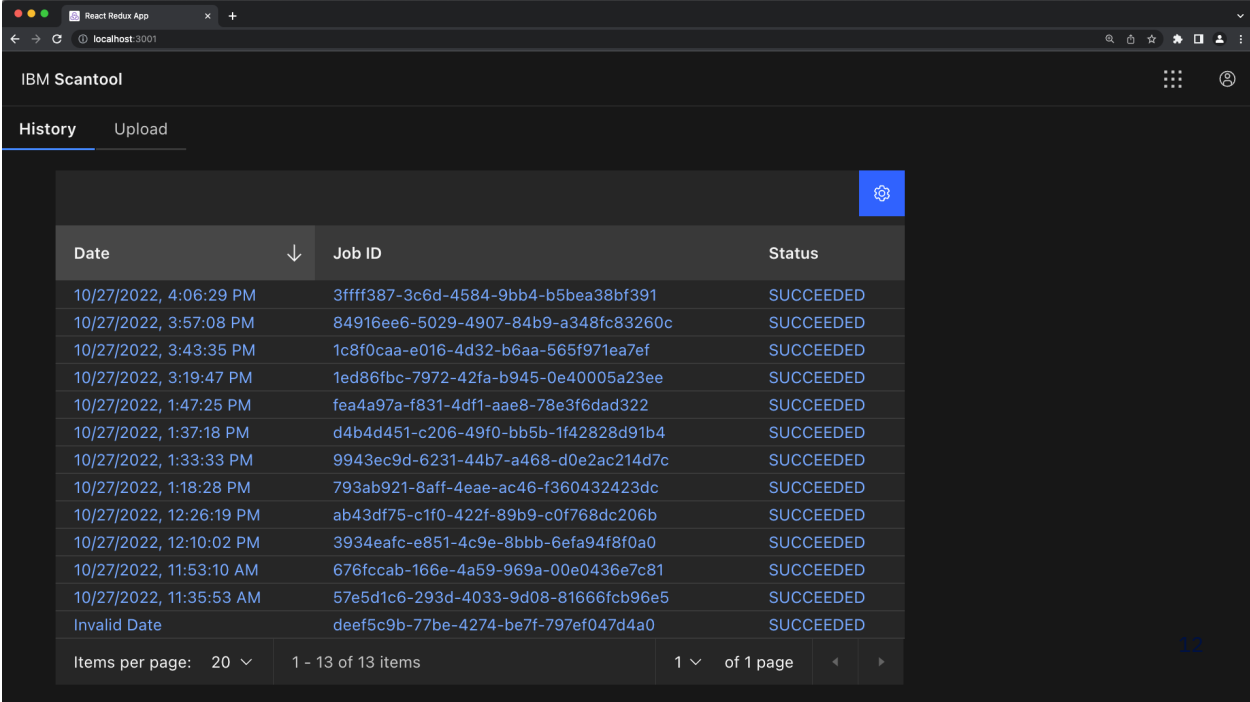
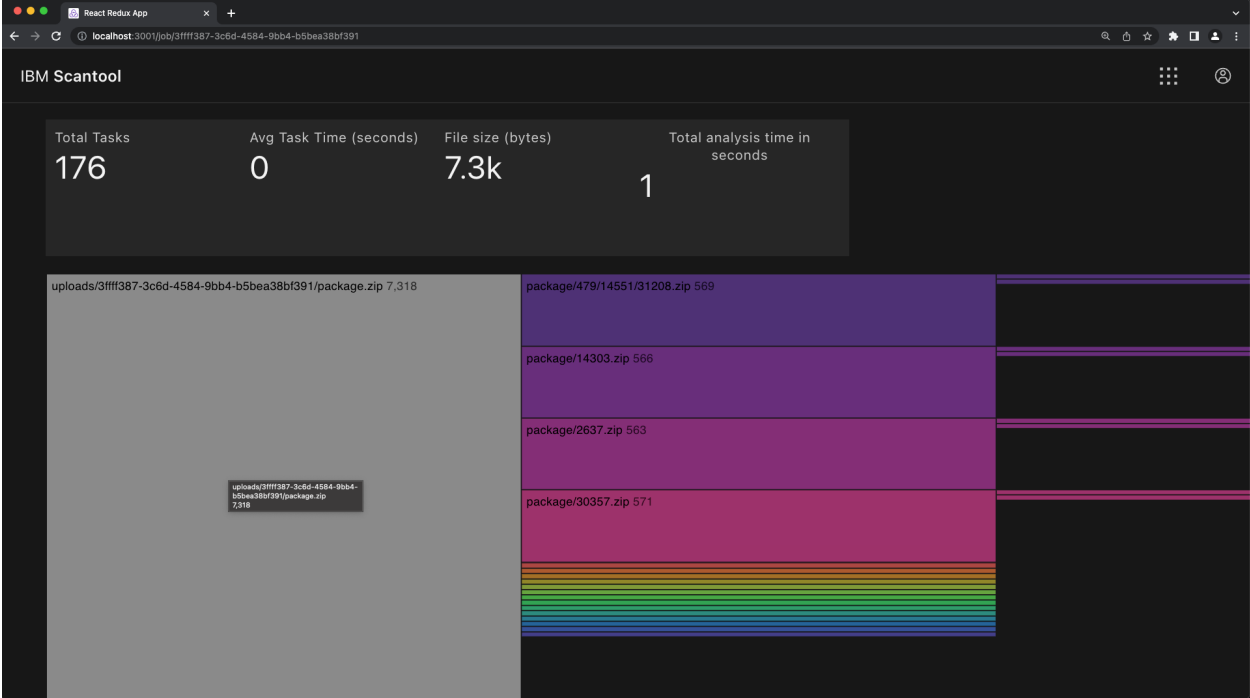
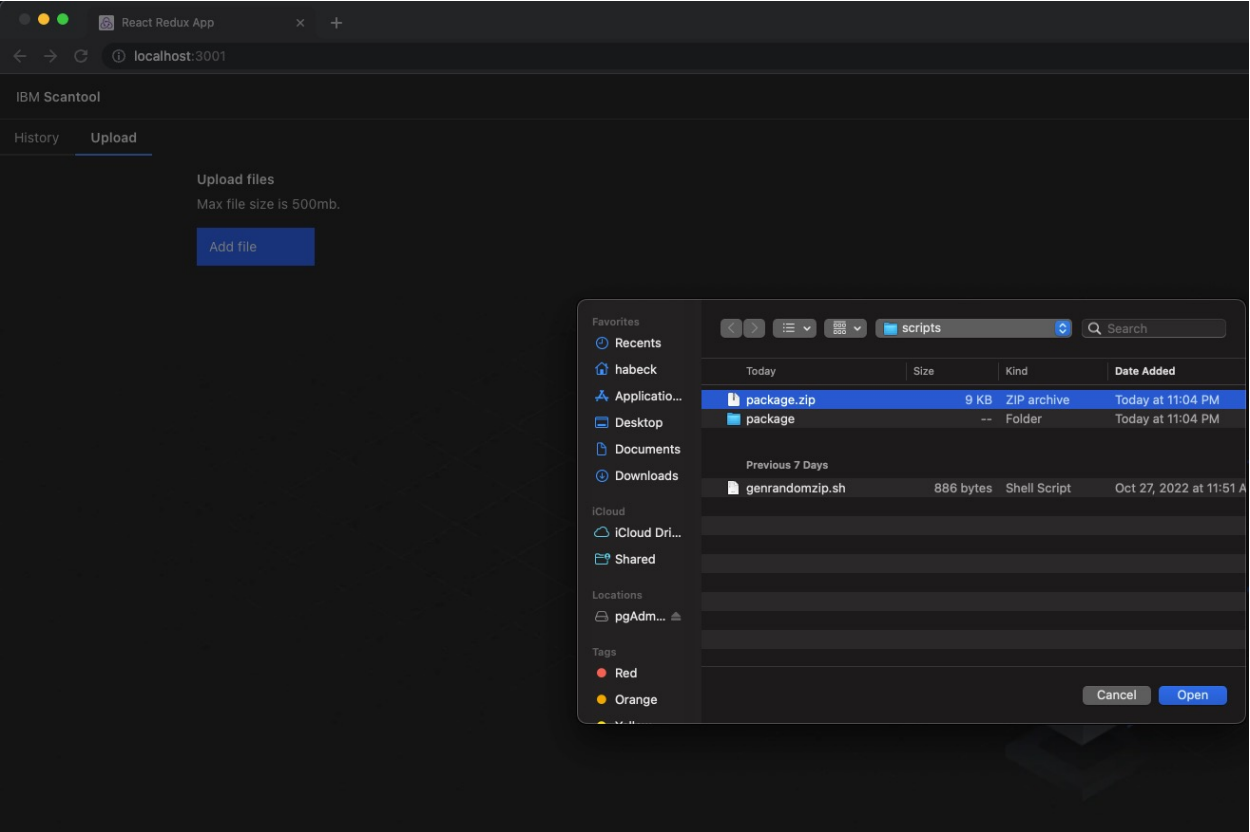
```
Dockerfile > ...
1 FROM ubuntu:focal
2
3 RUN apt-get update
4 RUN apt-get install -y wget
5
6 RUN mv /var/lib/dpkg/status /var/lib/dpkg/status.bak
7 RUN touch /var/lib/dpkg/status
8
9 RUN apt-get update
10
```

```
sbom/docker > grep wget sbom.dpkg.json
sbom/docker >
```

Demo: Verifying wget



UI of POC



Status and Roadmap

Current Status

- Several techniques for computing genes
- Support for binaries, packages, bytecode
- Cloud native application / processing engine
- Currently performing large-scale evaluation
- Additional features and capabilities in development

Plans for Release:

- A version of genome generation
- Service demonstrating technology
- Utilities for handling genomes and querying service

Requests:

- Welcome feedback, support and collaboration
- Insights on capabilities and use cases
- How to complement existing OpenSSF projects