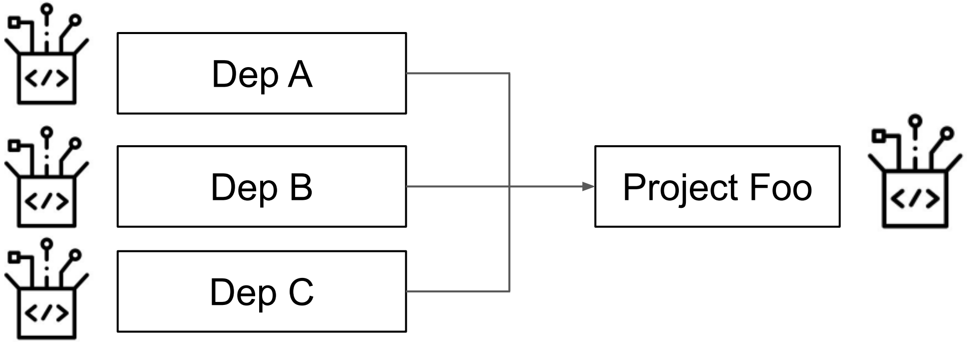
A Comprehensive Study of Bills of Materials for Software Systems

Trevor Stalnaker

July 21, 2023

What is the software supply chain?

- Software components combined to make final product
- Open source software (OSS)
    - Developers don't reinvent the wheel
    - Libraries with core functionality can be shared / distributed

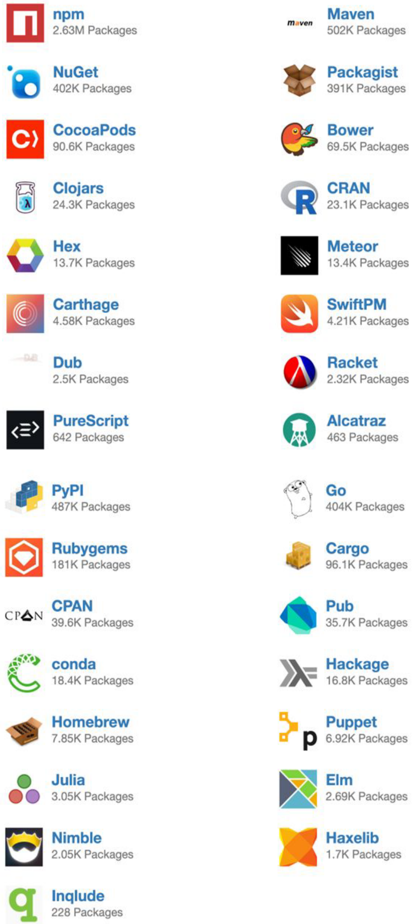


Dep A

Dep B

Dep C

Project Foo



npm
2.63M Packages

NuGet
402K Packages

CocoaPods
90.6K Packages

Clojars
24.3K Packages

Hex
13.7K Packages

Carthage
4.58K Packages

Dub
2.5K Packages

PureScript
642 Packages

PyPI
487K Packages

Rubygems
181K Packages

CPAN
39.6K Packages

conda
18.4K Packages

Homebrew
7.85K Packages

Julia
3.05K Packages

Nimble
2.05K Packages

Inqlude
228 Packages

Maven
502K Packages

Packagist
391K Packages

Bower
69.5K Packages

CRAN
23.1K Packages

Meteor
13.4K Packages

SwiftPM
4.21K Packages

Racket
2.32K Packages

Alcatraz
463 Packages

Go
404K Packages

Cargo
96.1K Packages

Pub
35.7K Packages

Hackage
16.8K Packages

Puppet
6.92K Packages

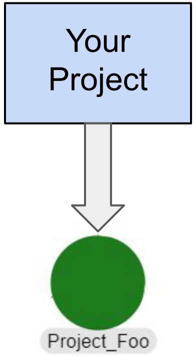
Elm
2.69K Packages

Haxelib
1.7K Packages

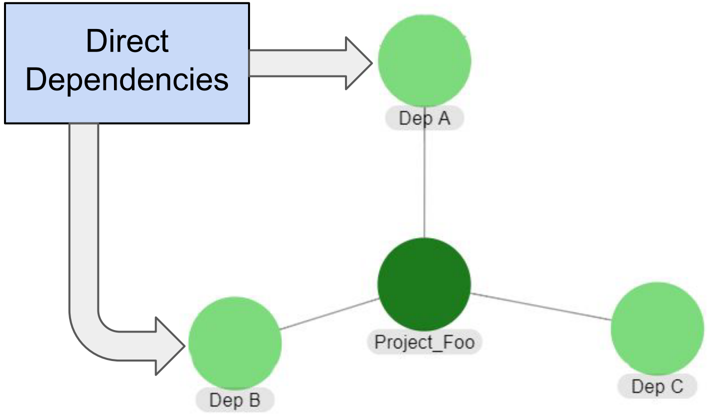
Problems in the

Software Supply Chain

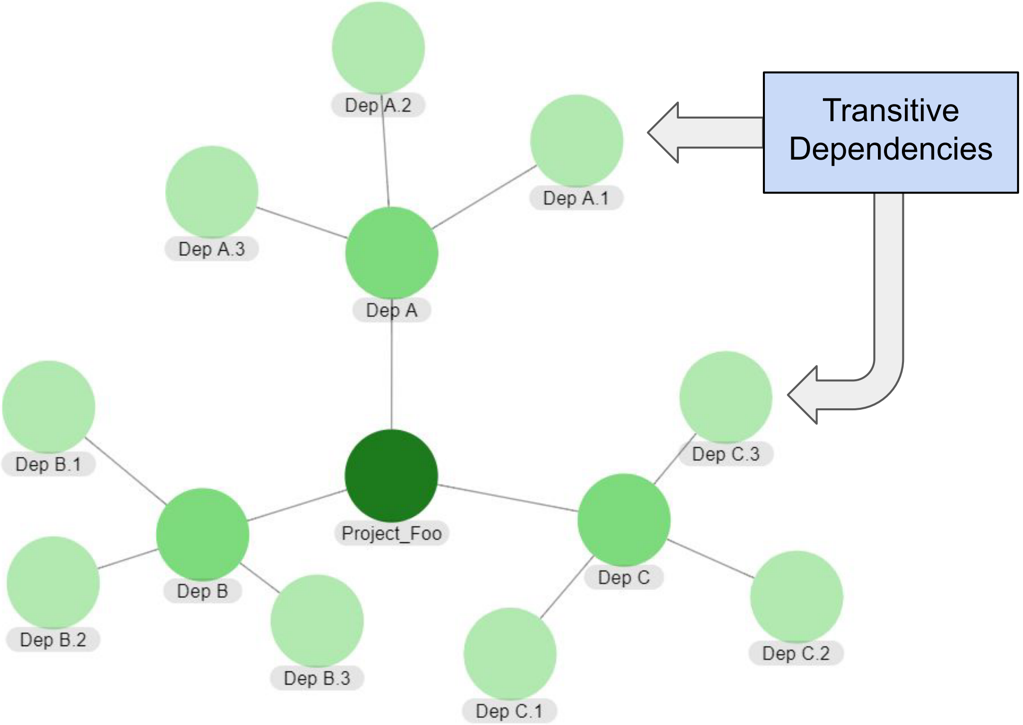
1. Dependency Management: What's in your project?



1. Dependency Management: What's in your project?



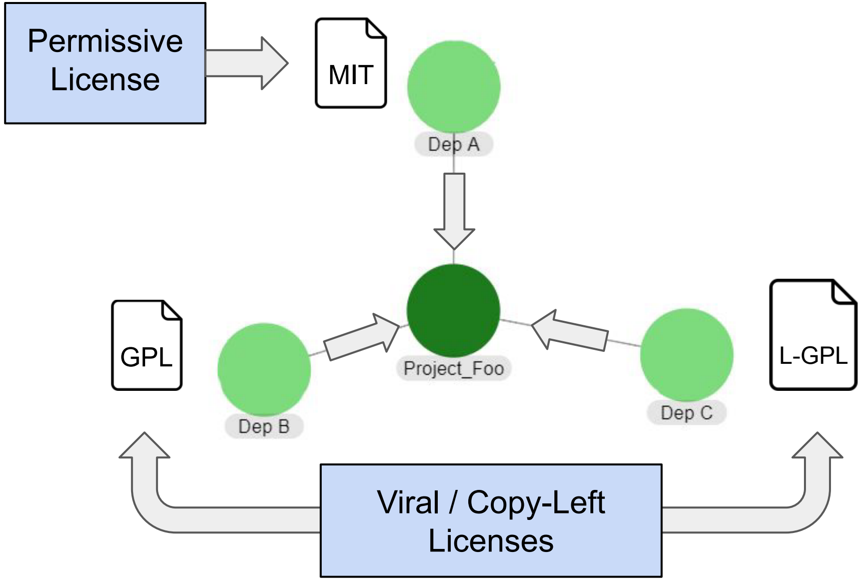
1. Dependency Management: What’s in your project?



Transitive Dependencies

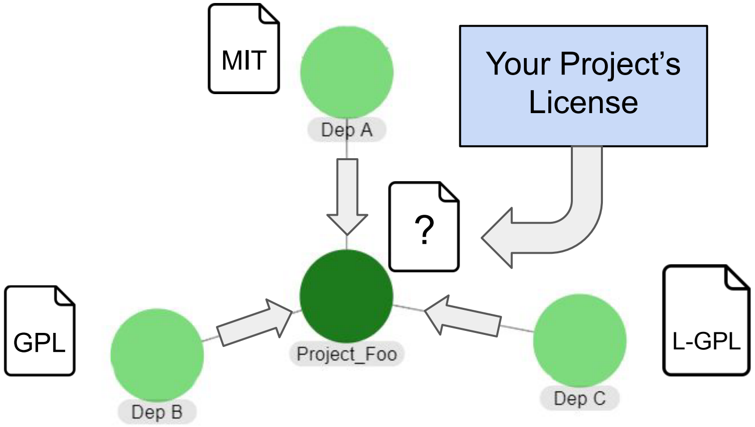
2. License Compliance

Each dependency can have a different license



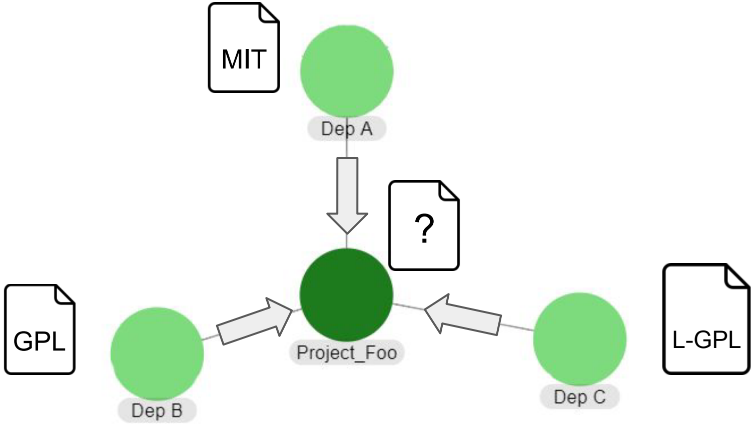
2. License Compliance

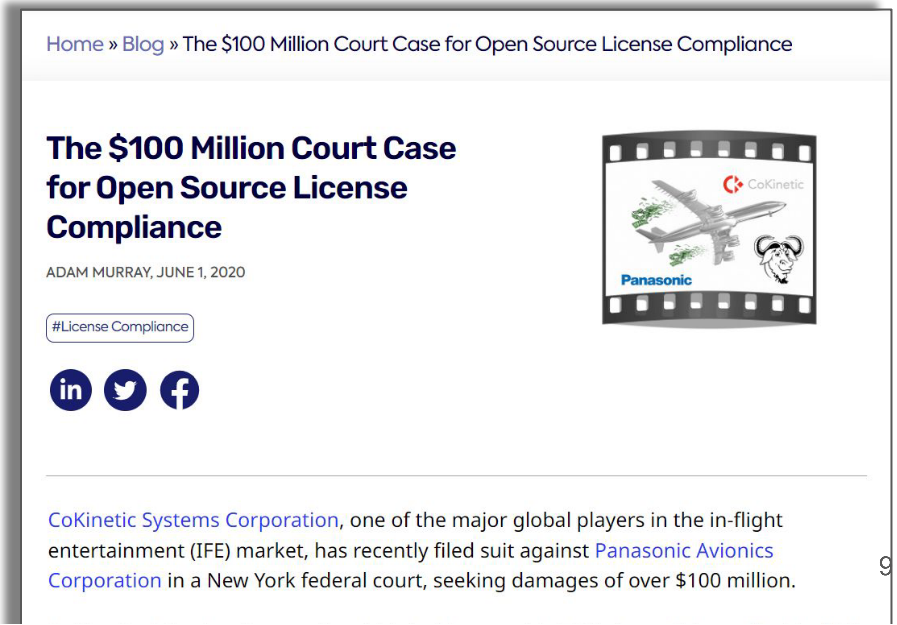
Project license must comply with them all



2. License Compliance

If not, you could face serious legal liability




Home » Blog » The $100 Million Court Case for Open Source License Compliance

The $100 Million Court Case for Open Source License Compliance

ADAM MURRAY, JUNE 1, 2020

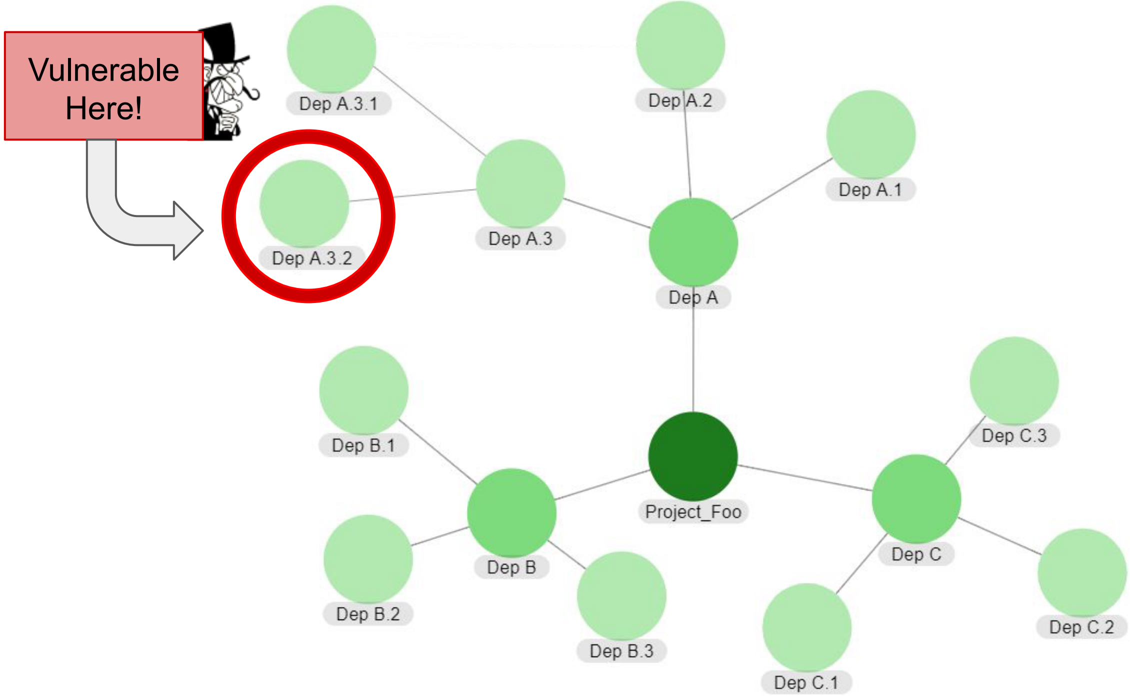
#License Compliance

CoKinetic Systems Corporation, one of the major global players in the in-flight entertainment (IFE) market, has recently filed suit against Panasonic Avionics Corporation in a New York federal court, seeking damages of over $100 million.

3. Security Concerns

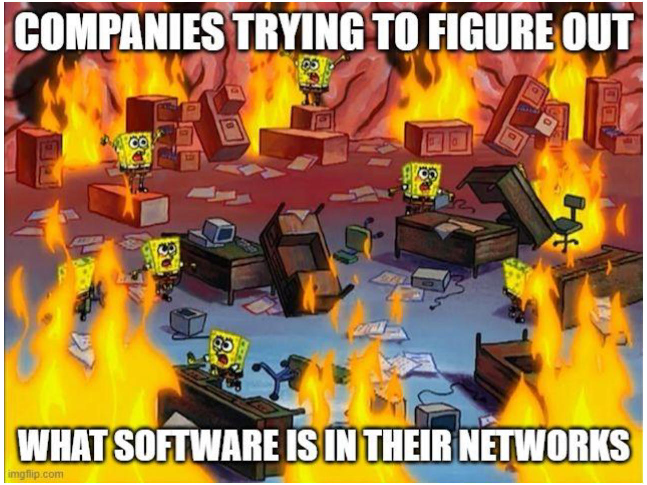
Exploits in dependency can leave project vulnerable

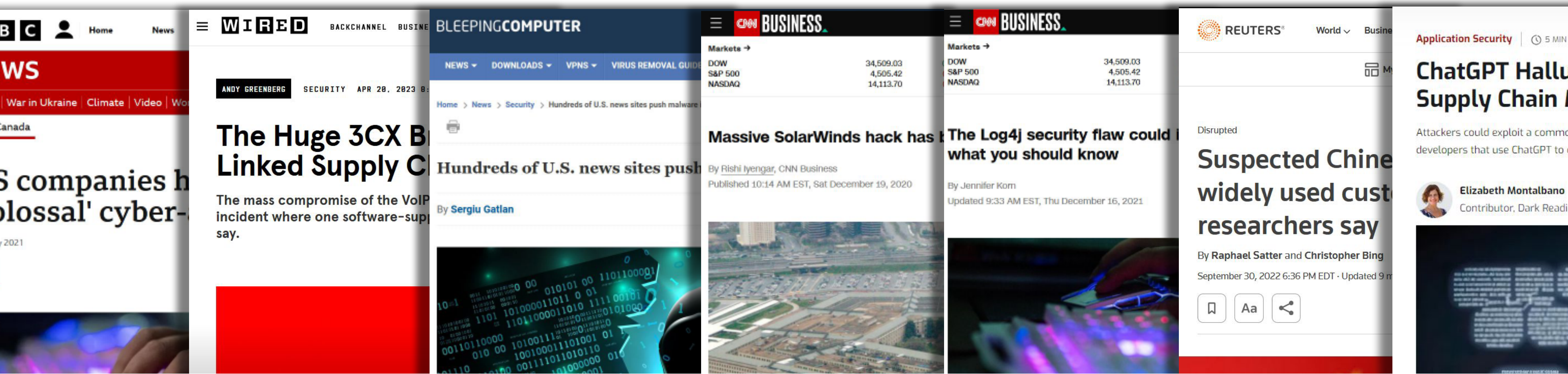
Supply chain attack: Bad actors target dependencies



3. Security Concerns

- In 2022 supply chain attacks impacted over [1]
  - 10 million people
  - 1700 organizations
- Recent examples:
  - SolarWinds breach
  - Log4J critical vulnerability


COMPANIES TRYING TO FIGURE OUT
WHAT SOFTWARE IS IN THEIR NETWORKS
imgflip.com



BC Home News
WS
War in Ukraine | Climate | Video | Wo
Canada
S companies h
olossal' cyber-
2021

WIRED BACKCHANNEL BUSINE
ANDY GREENBERG SECURITY APR 28, 2023 8:
The Huge 3CX Br
Linked Supply Cl
The mass compromise of the VoIP
incident where one software-sup
say.

BLEEPINGCOMPUTER
NEWS DOWNLOADS VPNS VIRUS REMOVAL GUIDE
Home > News > Security > Hundreds of U.S. news sites push malware
Hundreds of U.S. news sites push
By Sergiu Gatlan

CNN BUSINESS
Markets →
DOW 34,509.03
S&P 500 4,505.42
NASDAQ 14,113.70
Massive SolarWinds hack has
By Rishi Iyengar, CNN Business
Published 10:14 AM EST, Sat December 19, 2020

CNN BUSINESS
Markets →
DOW 34,509.03
S&P 500 4,505.42
NASDAQ 14,113.70
The Log4j security flaw could i
what you should know
By Jennifer Korn
Updated 9:33 AM EST, Thu December 16, 2021

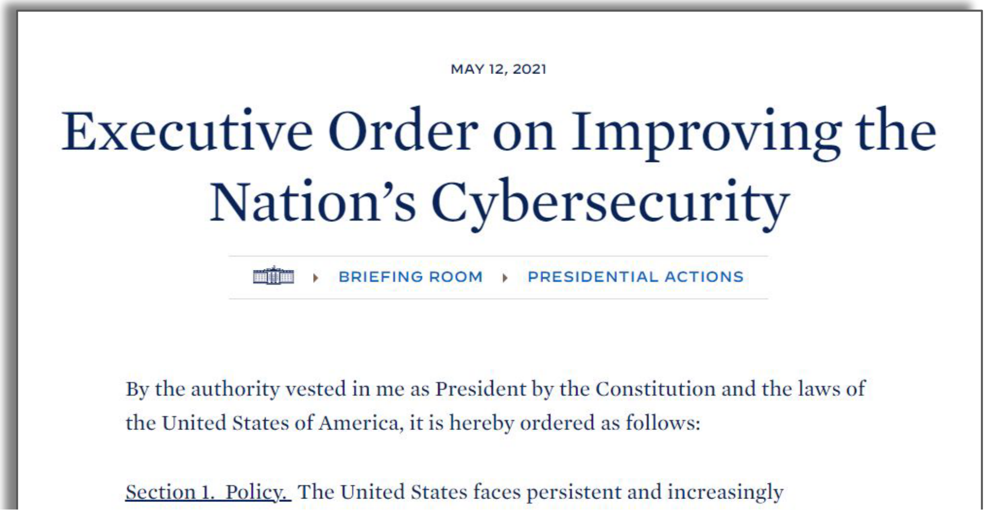
REUTERS World Busine
Disrupted
Suspected Chine
widely used cust
researchers say
By Raphael Satter and Christopher Bing
September 30, 2022 6:36 PM EDT · Updated 9 m
Aa

Application Security | 5 MIN
ChatGPT Hallu
Supply Chain
Attackers could exploit a commo
developers that use ChatGPT to
Elizabeth Montalbano
Contributor, Dark Readi

What’s the solution?

Software Bills of Materials (SBOMs)

- 2021 US Presidential Executive Order 14028
- Requires companies selling software to US government to provide SBOM
- Gave momentum to SBOM formalization and adoption



MAY 12, 2021

Executive Order on Improving the Nation's Cybersecurity

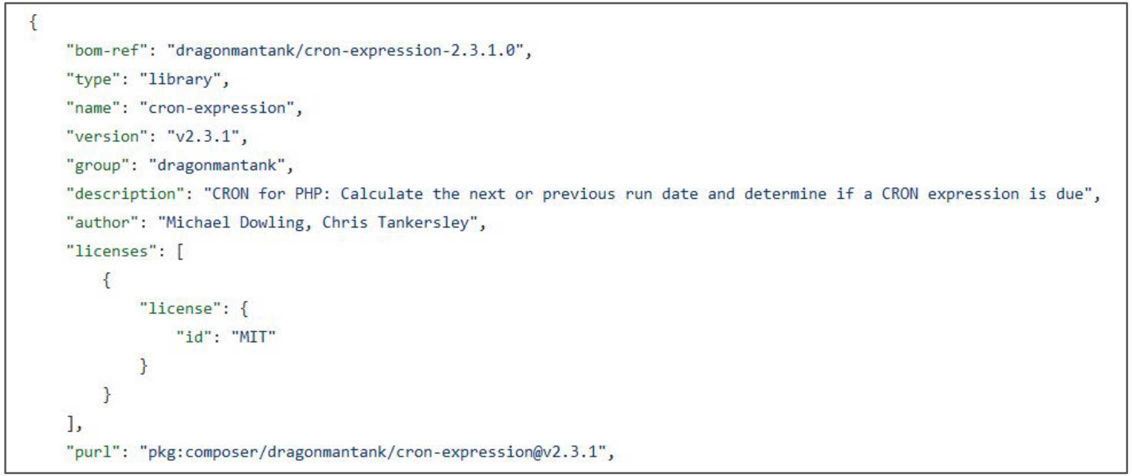
BRIEFING ROOM ▸ PRESIDENTIAL ACTIONS

By the authority vested in me as President by the Constitution and the laws of the United States of America, it is hereby ordered as follows:

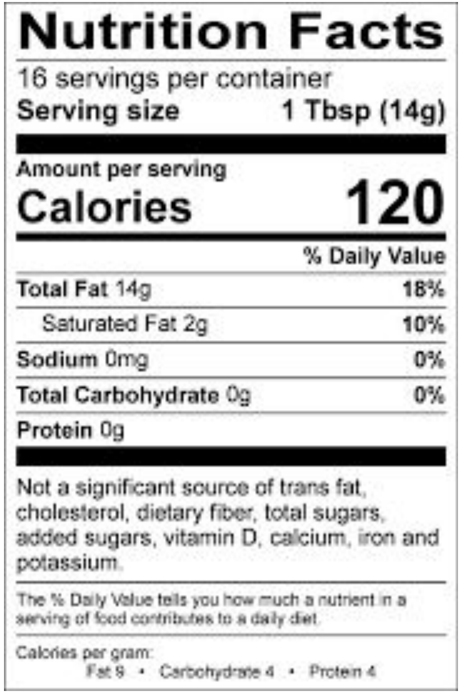
Section 1. Policy. The United States faces persistent and increasingly

What is SBOM?

- Inspired by Bill of Materials (BOM)
    - From manufacturing industry
- Manifest list of components
    - Dependencies, provenance information, licenses, etc
- Machine readable



```
{
    "bom-ref": "dragonmantank/cron-expression-2.3.1.0",
    "type": "library",
    "name": "cron-expression",
    "version": "v2.3.1",
    "group": "dragonmantank",
    "description": "CRON for PHP: Calculate the next or previous run date and determine if a CRON expression is due",
    "author": "Michael Dowling, Chris Tankersley",
    "licenses": [
        {
            "license": {
                "id": "MIT"
            }
        }
    ],
    "purl": "pkg:composer/dragonmantank/cron-expression@v2.3.1",
```



Nutrition Facts

16 servings per container

Serving size	**1 Tbsp (14g)**
Amount per serving **Calories**	**120**
	% Daily Value
Total Fat 14g	**18%**
Saturated Fat 2g	**10%**
Sodium 0mg	**0%**
Total Carbohydrate 0g	**0%**
Protein 0g	

Not a significant source of trans fat, cholesterol, dietary fiber, total sugars, added sugars, vitamin D, calcium, iron and potassium.

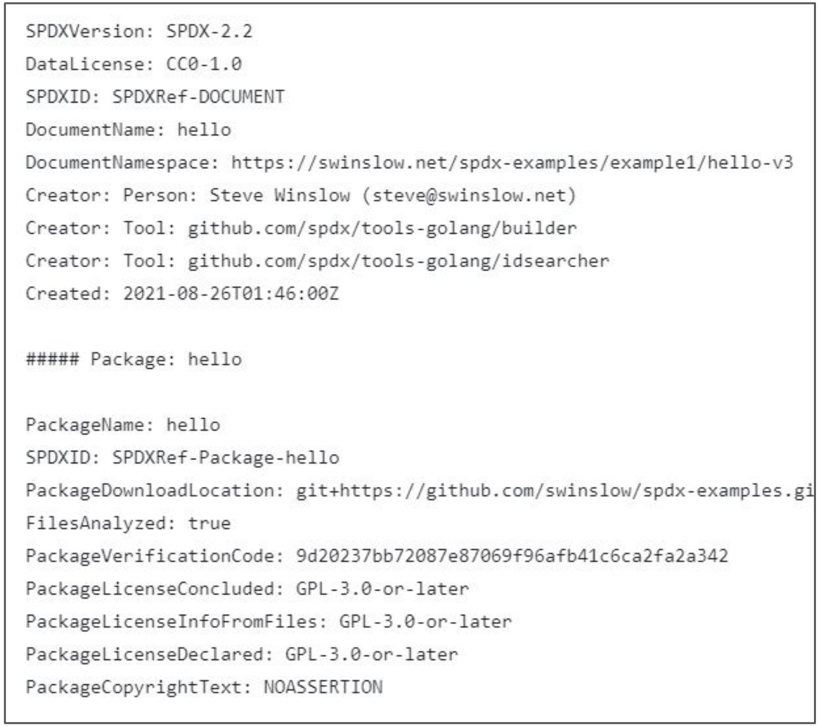
The % Daily Value tells you how much a nutrient in a serving of food contributes to a daily diet.

Calories per gram:
Fat 9 • Carbohydrate 4 • Protein 4

Like this!

SBOM Formats: SPDX

- ISO recognized standard
- Primarily licensing focused
- Promoted by: The Linux Foundation
- Supported file formats:
  - tag/value (.spdx)
  - JSON
  - YAML
  - RDF/XML
  - spreadsheets (.xls)



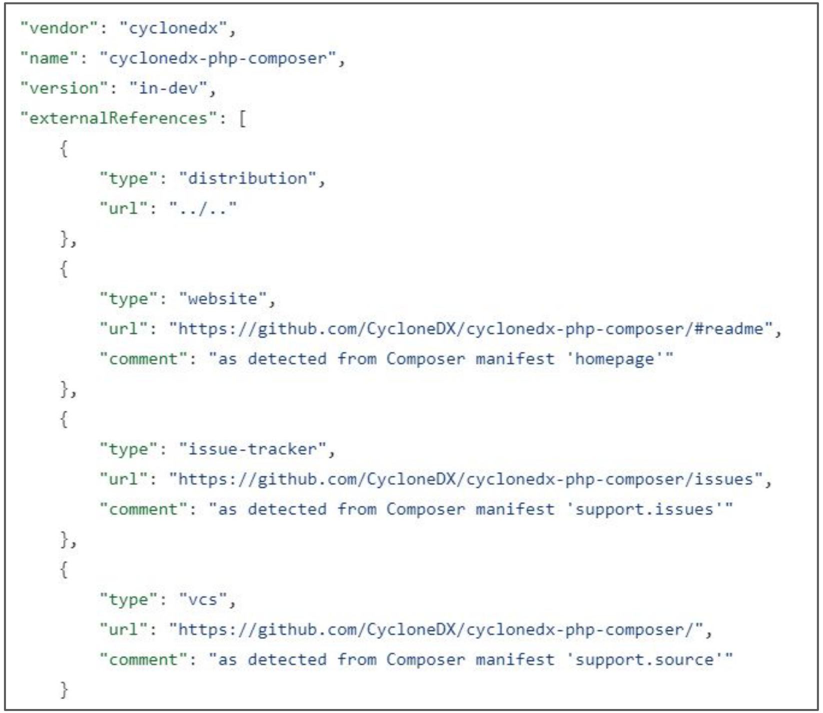
```
SPDXVersion: SPDX-2.2
DataLicense: CC0-1.0
SPDXID: SPDXRef-DOCUMENT
DocumentName: hello
DocumentNamespace: https://swinslow.net/spdx-examples/example1/hello-v3
Creator: Person: Steve Winslow (steve@swinslow.net)
Creator: Tool: github.com/spdx/tools-golang/builder
Creator: Tool: github.com/spdx/tools-golang/idsearcher
Created: 2021-08-26T01:46:00Z

##### Package: hello

PackageName: hello
SPDXID: SPDXRef-Package-hello
PackageDownloadLocation: git+https://github.com/swinslow/spdx-examples.gi
FilesAnalyzed: true
PackageVerificationCode: 9d20237bb72087e87069f96afb41c6ca2fa2a342
PackageLicenseConcluded: GPL-3.0-or-later
PackageLicenseInfoFromFiles: GPL-3.0-or-later
PackageLicenseDeclared: GPL-3.0-or-later
PackageCopyrightText: NOASSERTION
```

SBOM Formats: CycloneDX

- Primarily security focused
- Promoted by: OWASP
- Supported file formats:
  - JSON
  - XML
  - protocol buffers



```
"vendor": "cyclonedx",
"name": "cyclonedx-php-composer",
"version": "in-dev",
"externalReferences": [
    {
        "type": "distribution",
        "url": "../.."
    },
    {
        "type": "website",
        "url": "https://github.com/CycloneDX/cyclonedx-php-composer/#readme",
        "comment": "as detected from Composer manifest 'homepage'"
    },
    {
        "type": "issue-tracker",
        "url": "https://github.com/CycloneDX/cyclonedx-php-composer/issues",
        "comment": "as detected from Composer manifest 'support.issues'"
    },
    {
        "type": "vcs",
        "url": "https://github.com/CycloneDX/cyclonedx-php-composer/",
        "comment": "as detected from Composer manifest 'support.source'"
    }
```

CycloneDX

BOMs for Software Systems

- SBOM (software in general)
- SaaSBOM (services and APIs)
- HBOM (hardware)
- FBOM (firmware)
- OBOM (operational / configuration environments)
- DataBOM (datasets)
- AI / MLBOM (AI models)

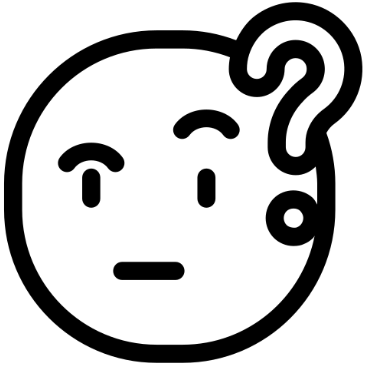
* For simplicity, referred to as SBOM from here unless otherwise noted

BOMs for Software Systems

- **SBOM** (software in general)
- SaaSBOM (services and APIs)
- **HBOM** (hardware)
- FBOM (firmware)
- OBOM (operational / configuration environments)
- **DataBOM** (datasets)
- **AI / MLBOM** (AI models)

* For simplicity, referred to as SBOM from here unless otherwise noted

Sounds great!



What’s the problem?

Major Stakeholder Concerns

- Uncertain / low levels of commitment to SBOM
- Unclear if SBOM benefits will be actualized
- Fears of inaccurate and incomplete SBOM
- Absence of agreement in SBOM content
- Lack of mature tool support for consumption / production
- Unsure when / how SBOM should be used in development processes

The goal of this thesis is to understand…

1. How and to what extent stakeholders currently create and use SBOMs
2. Opportunities / benefits SBOMs offer for different software and stakeholders
3. Specific challenges preventing stakeholders from enjoying SBOM benefits
4. Actionable solutions to overcome challenges and enable new opportunities

Research Questions

RQ1:
SBOM Usage

How do software stakeholders **create and use SBOMs**?

RQ2:
Challenges

What are the **challenges** of creating and using SBOMs?

RQ3:
Solutions

What are actionable **solutions** to SBOM challenges?

Populations

SBOM Community and Adopters (**SBOM C&A**)

Producers, Consumers, Tool Makers, Educators, Standard Makers

Contributors of Critical OSS Projects (**Critical Projects**)

AI/**ML** Developers and Researchers

Cyber Physical Systems (**CPS**) Developers and Researchers

Legal Practitioners

Participant Identification

SBOM C&A

- Keyword-based GitHub mining

  Tags, issues, commits, etc. looking for evidence of SBOM usage
- Repositories dependent on SBOM tools and repos
- Mailing lists & newsletters
- Industry contacts

Critical Projects

- OSSF workgroup on Securing Critical Projects: 102 critical projects (564 repositories)
- Mined top-10 contributors of repositories

Participant Identification

ML

- Machine learning projects on GitHub with 100+ stars
- Professional network

CPS

- Professional network

Legal

- Professional network

Survey Design - Quick Stats

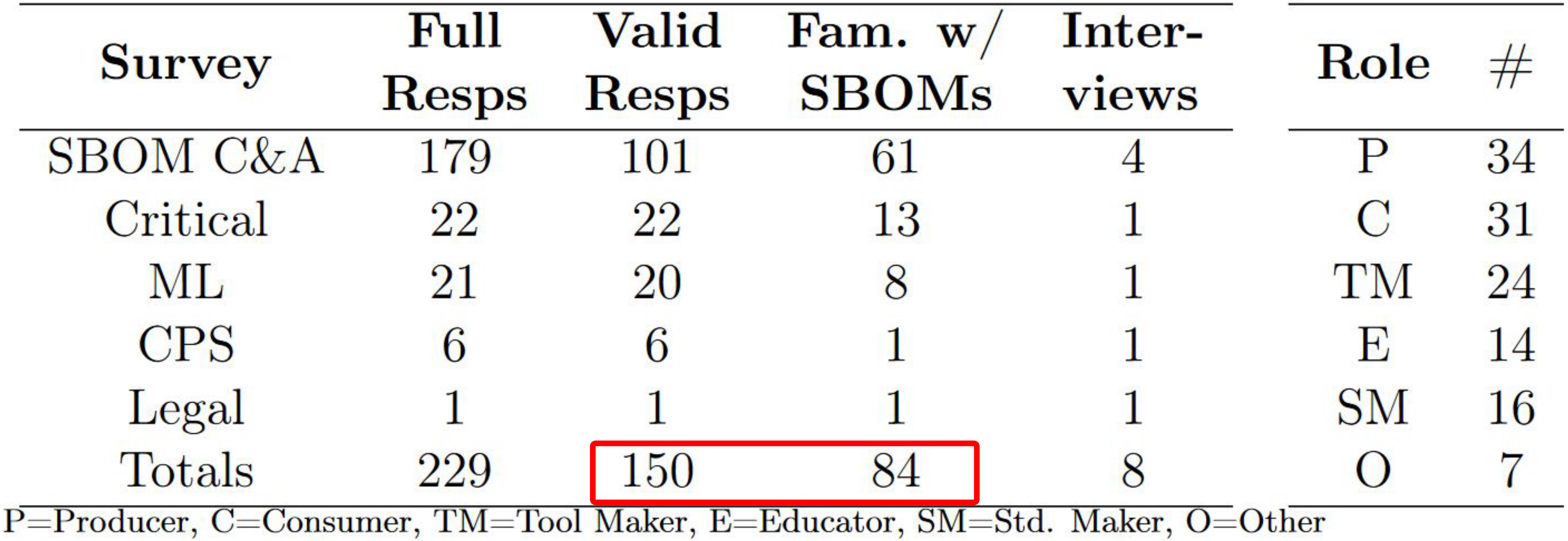
- Platform: Qualtrics
- Completion time: 20-30 minutes
- Waves: 3
- Compensation: lottery for $50 gift card
- Questions types:
  - Likert scale
  - Multiple choice
  - Ranking
  - Short answer

Survey Design - Questions

- Clear and concise
  - Avoided language that would bias responses
- Broken into logical sections
- Based on
  - Literature and prior works
  - General points of interest
  - Early findings from SBOM C&A survey

Final Response Count

- Contacted over 4.4K individuals via email
- Responses from 16/102 Critical Projects



Survey	Full Resps	Valid Resps	Fam. w/ SBOMs	Inter-views
SBOM C&A	179	101	61	4
Critical	22	22	13	1
ML	21	20	8	1
CPS	6	6	1	1
Legal	1	1	1	1
Totals	229	150	84	8

Role	#
P	34
C	31
TM	24
E	14
SM	16
O	7

P=Producer, C=Consumer, TM=Tool Maker, E=Educator, SM=Std. Maker, O=Other

Response Annotation

- Employed open coding methodology
- Two authors annotated all responses
  - Shared Google Sheet with evolving list of codes
- Codes tagged participant responses
  - More than one code could be applied to responses
- Authors met and reconciled codes
  - 3rd author was brought in to resolve disputes

Example:
Question: What issues have you faced when consuming SBOMs?
Response: “In most of the cases, we receive SBOMs in a proprietary format with varying quality.”
Codes: [DIFFERENT STANDARDS], [POOR QUALITY SBOMS]

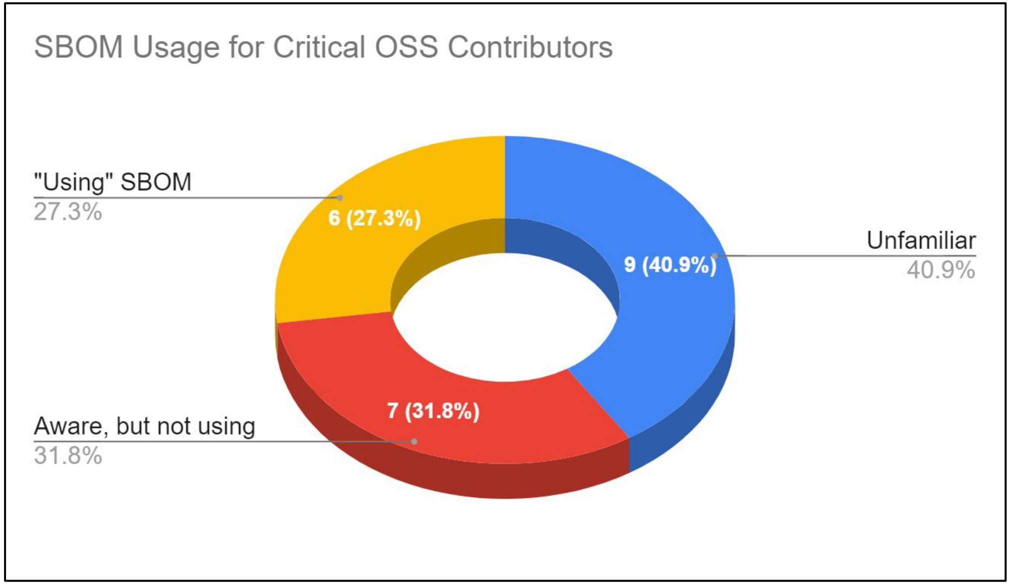
Interviews

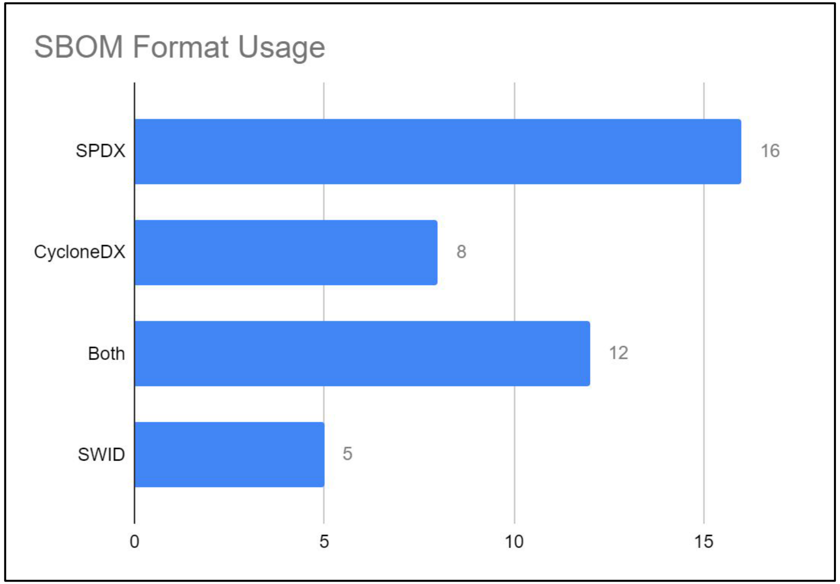
- Platform: Zoom (recorded)
- Duration: One hour
- Format: semi-structured
- Compensation: $50 Amazon gift cards
- Interview count: 8
- Question types:
    - Follow up and clarification
    - Domain specific
    - Derived from survey results
- Analysis: Open-coding of transcription

Results

SBOM Awareness

56% of all participants familiar with SBOM





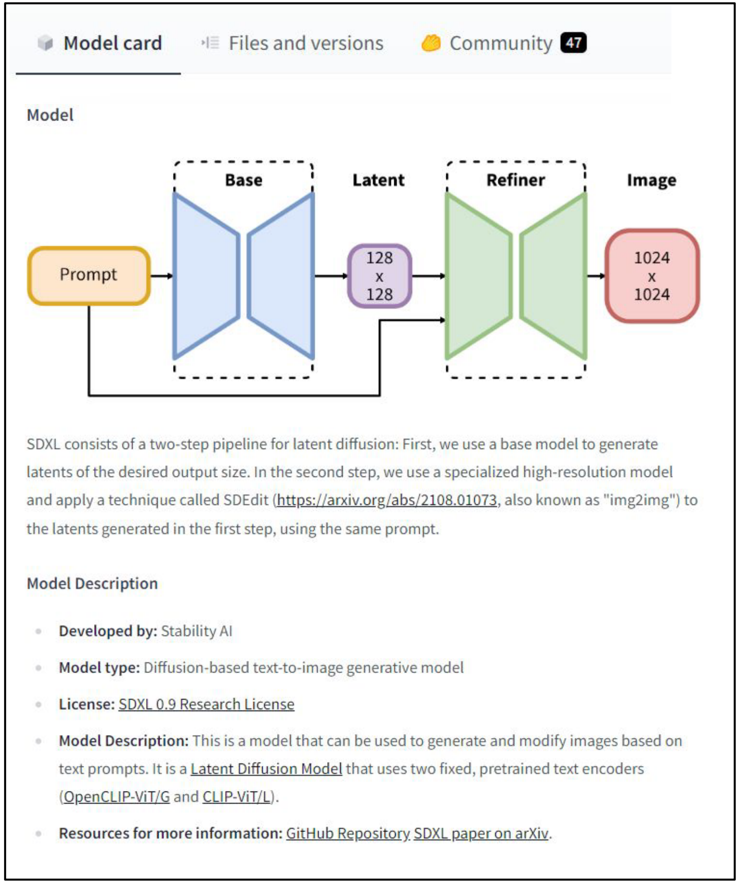
SBOM Awareness

CPS:

- Familiar with HBOM: 3/6
- Used BOM: 2/6
  - Bespoke formats

ML practitioners:

- Unaware of BOM formats for AI or datasets
- Quasi-AIBOM
  - Hugging Face data and model cards



Model

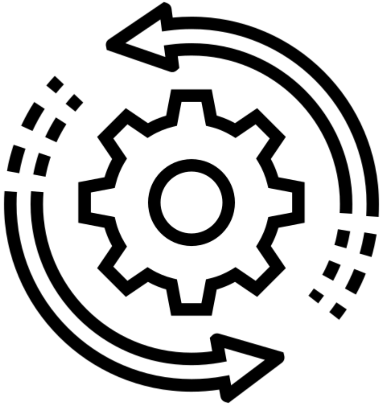
SDXL consists of a two-step pipeline for latent diffusion: First, we use a base model to generate latents of the desired output size. In the second step, we use a specialized high-resolution model and apply a technique called SDEdit (https://arxiv.org/abs/2108.01073, also known as "img2img") to the latents generated in the first step, using the same prompt.

Model Description

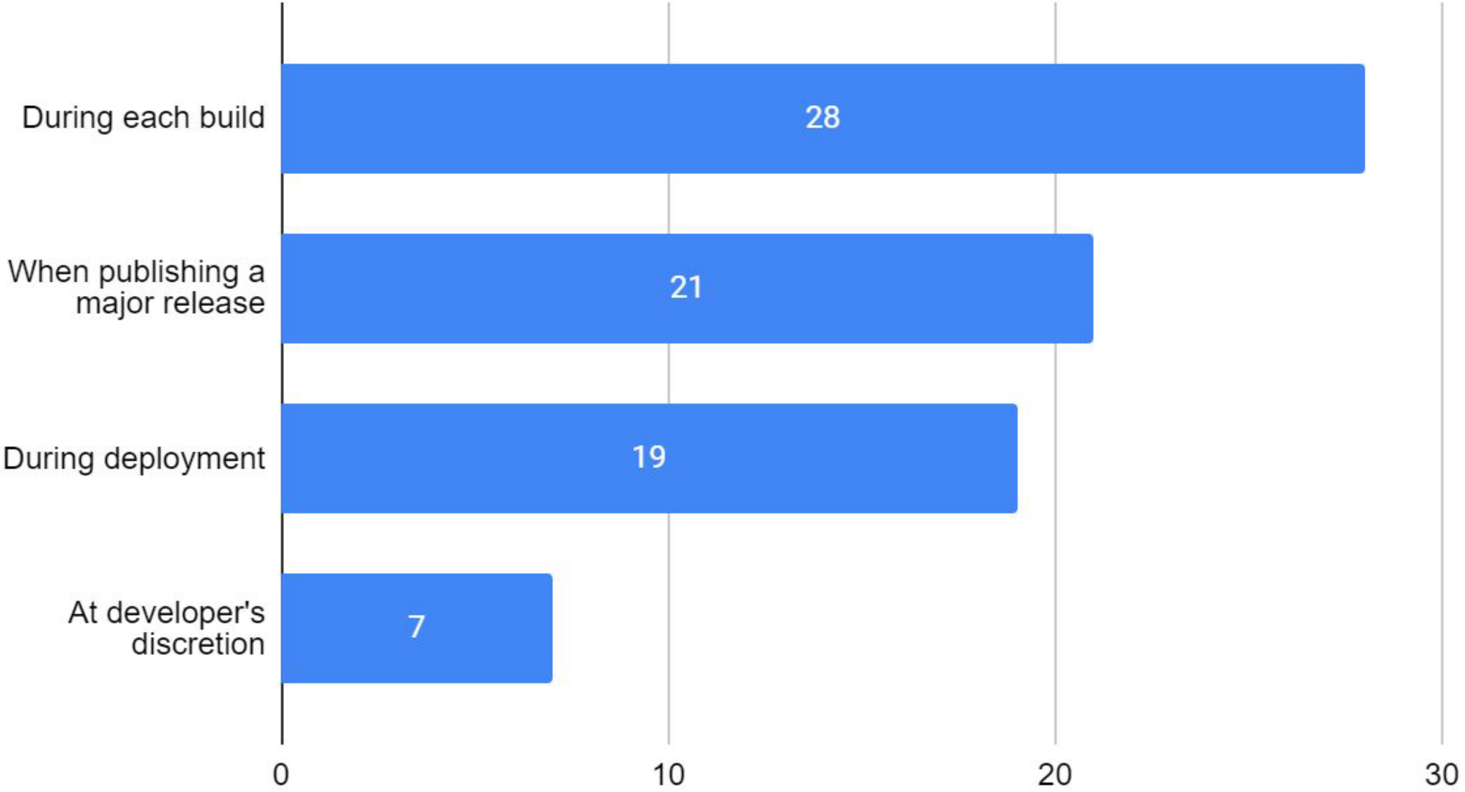
- **Developed by:** Stability AI
- **Model type:** Diffusion-based text-to-image generative model
- **License:** SDXL 0.9 Research License
- **Model Description:** This is a model that can be used to generate and modify images based on text prompts. It is a Latent Diffusion Model that uses two fixed, pretrained text encoders (OpenCLIP-ViT/G and CLIP-ViT/L).
- **Resources for more information:** GitHub Repository SDXL paper on arXiv.

SBOM Creation

- Pressure largely felt at end of supply chain
- Little incentive for projects at beginning to produce SBOM
  - Some have no dependencies to manage
- Leads to consumers producing SBOM for their dependencies
  - Can result in missing something or inaccurate SBOM

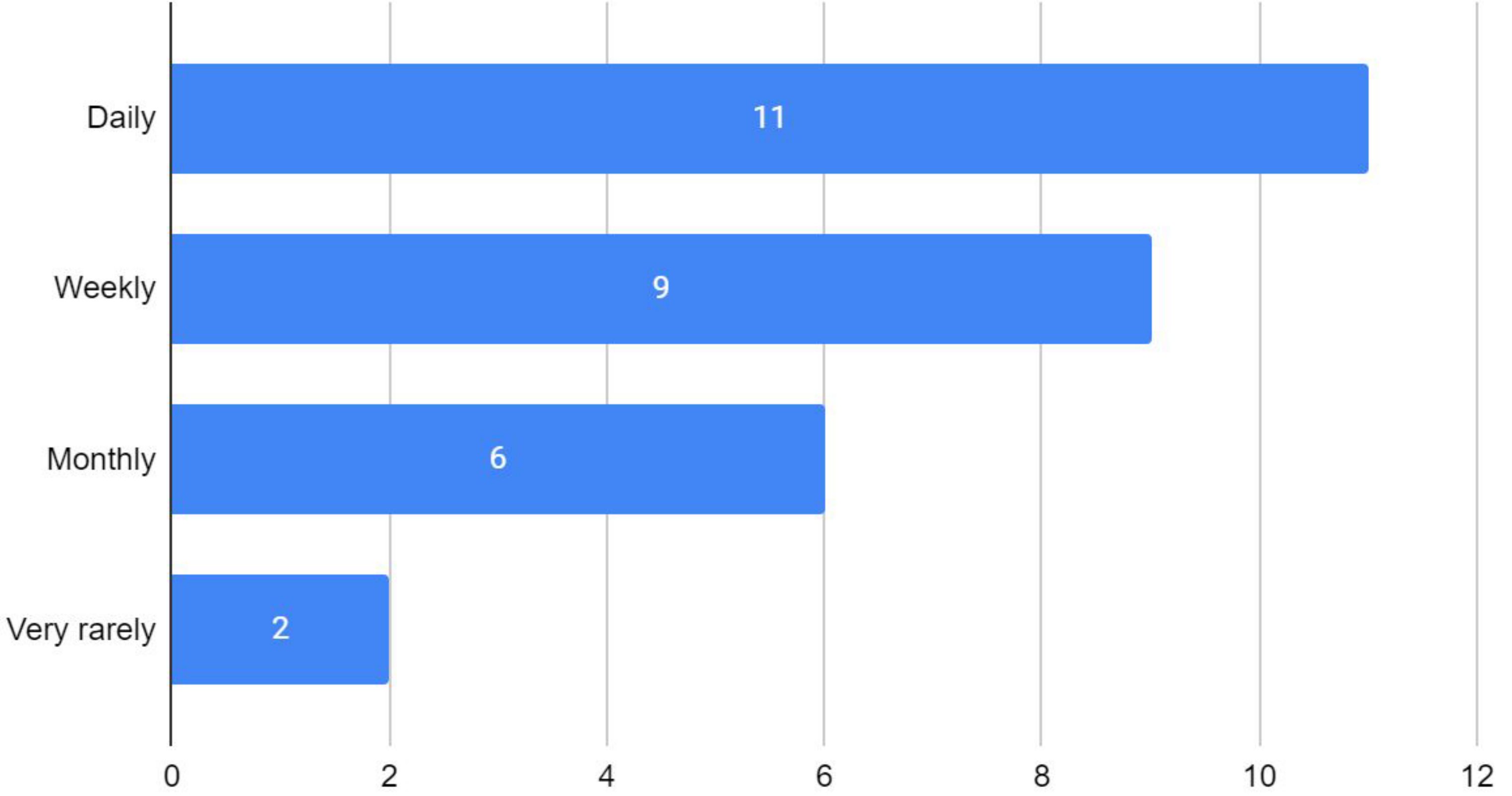


When should SBOM be generated?



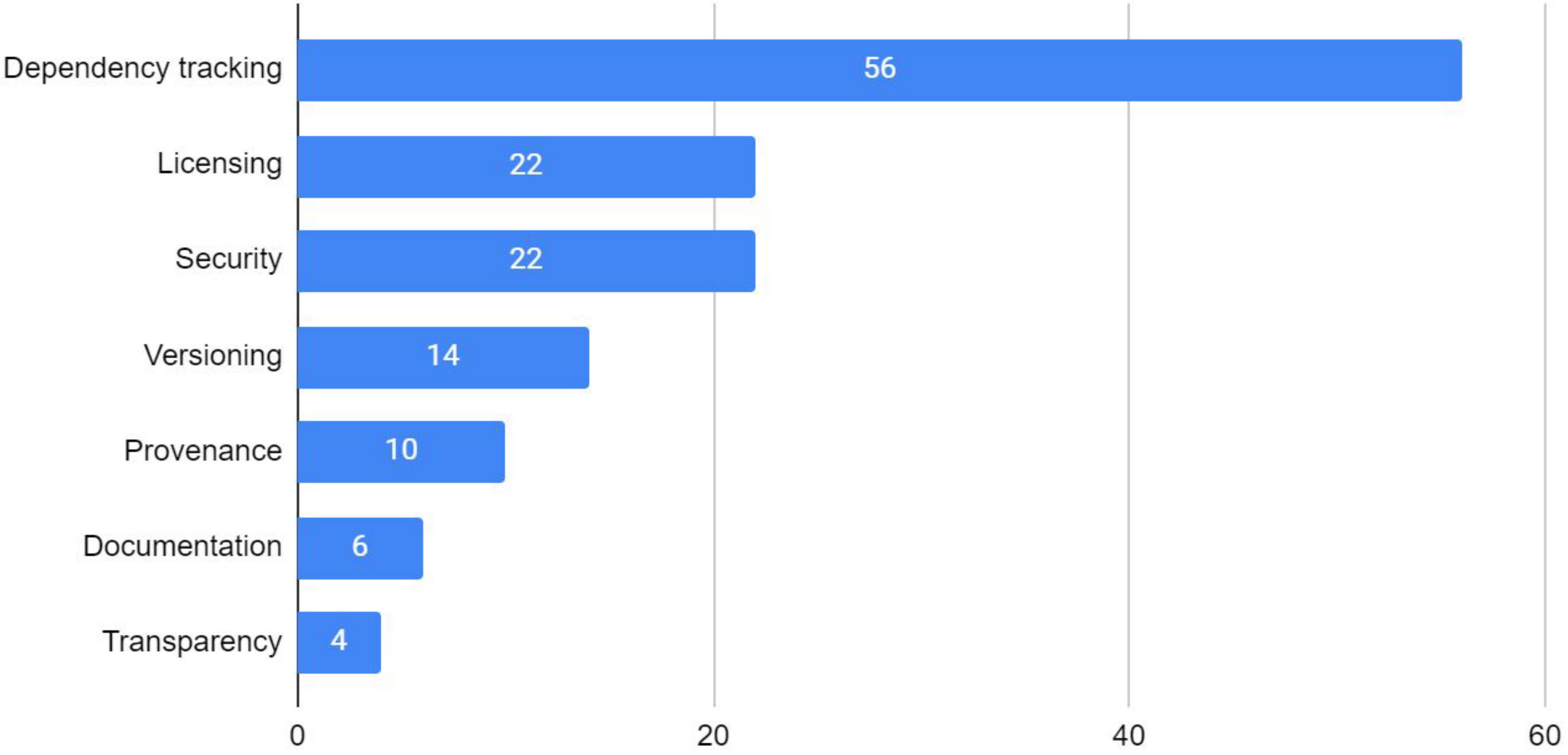
*34 producers polled

How often are SBOMs consumed?



*31 consumers polled

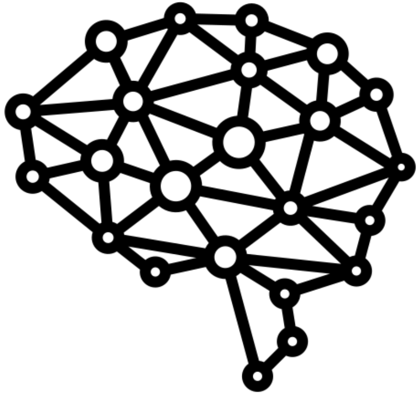
SBOM Use Cases



*61 practitioners polled

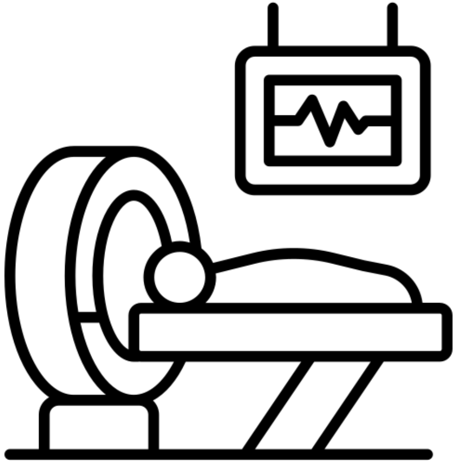
Use Cases - Machine Learning

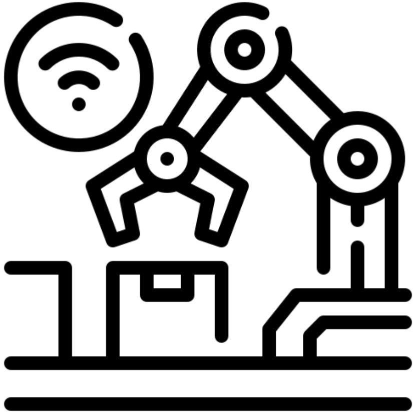
- Facilitate model reproducibility
- Help to identify and verify datasets across academic papers
- AIBOM
  - Provide transparency into how model was trained
    - Information on architecture
    - Hyper-parameters
    - Pre-trained base models used
- DataBOM
  - Identify poisoned, biased, or illegally sourced dataset



Use Cases - CPS

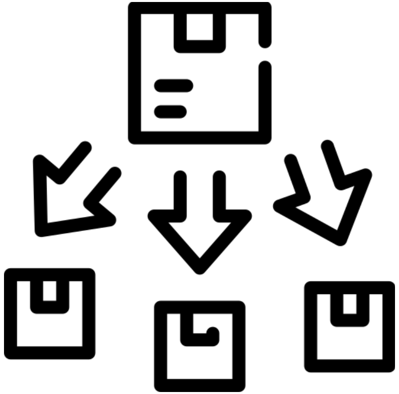
- Serve as regulatory documents
  - Facilitate review and approval of devices (consistent with prior work)
- Increase transparency and reproducibility of research results





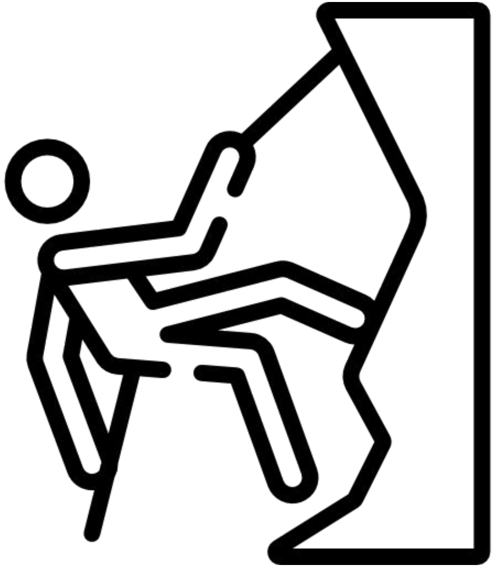
Tooling and Distribution

- Little consistency between respondents
  - Mix of in-house, commercial, and open-source tools
- No agreed upon method of distributing SBOM
- Expectation that developers of software are responsible for SBOM
  - Creation
  - Maintenance
  - Distribution
- Distribution is a challenge moving forward
  - Critical Project contributors (5/12)



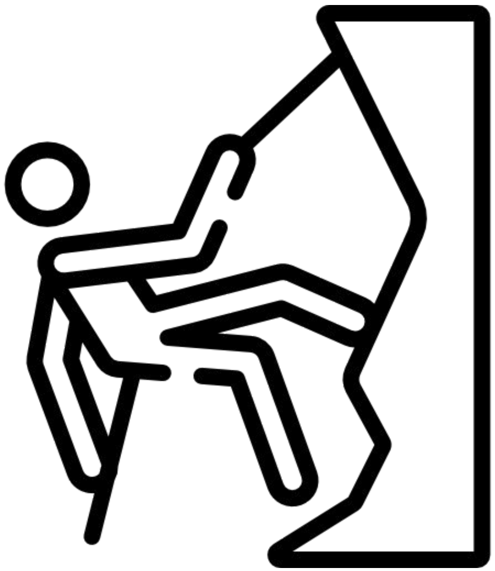
Identified Challenges

C1: Complexity of SBOM specifications
C2: Determining data fields to include in SBOMs
C3: Interoperability between SBOM standards
C4: Keeping SBOM up to date
C5: Insufficient SBOM tooling
C6: Inaccurate and incomplete SBOM
C7: Verifying SBOM accuracy and completeness
C8: Differences across ecosystems and communities
C9: SBOM completeness and data privacy trade-off
C10: SBOMs for legacy packages and repositories
C11: Inability to locate dependencies for SBOM
C12: Unclear SBOM direction
C13: Generating global software IDs
C14: Managing SBOM versions



Identified Challenges

C1: Complexity of SBOM specifications
C2: Determining data fields to include in SBOMs
C3: Interoperability between SBOM standards
C4: Keeping SBOM up to date
C5: Insufficient SBOM tooling
C6: Inaccurate and incomplete SBOM
C7: Verifying SBOM accuracy and completeness
C8: Differences across ecosystems and communities
C9: SBOM completeness and data privacy trade-off
C10: SBOMs for legacy packages and repositories
C11: Inability to locate dependencies for SBOM
C12: Unclear SBOM direction
C13: Generating global software IDs
C14: Managing SBOM versions



C1 - Complexity of SBOM specifications

- Struggling to understand / use the spec
- Every supported use case makes the spec more complicated

“If all you’re interested in is licensing, [...] [you] don’t want to have to learn [about other domains like security] just to be able to use the spec.”

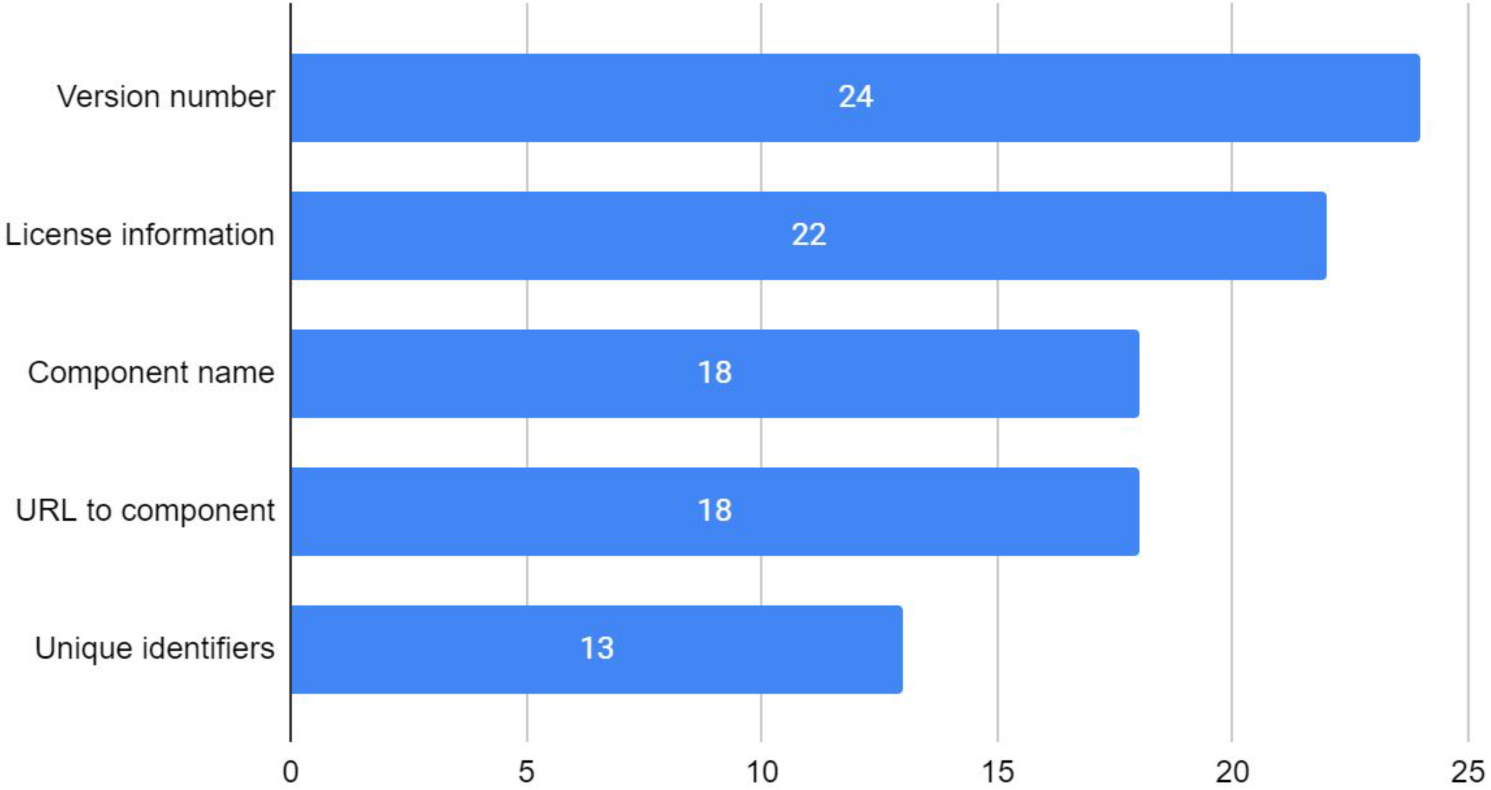
"[...] one core issue [...] is definitely a tension between use case coverage and the complexity of the spec."

C2 - Determining data fields to include in SBOMs

- What information should be required in an SBOM?
- What information should be optional?
- Adding too many required fields clutters the spec and SBOM document
  - Too many fields may also slow down SBOM generation
- We asked practitioners what fields they thought were necessary…

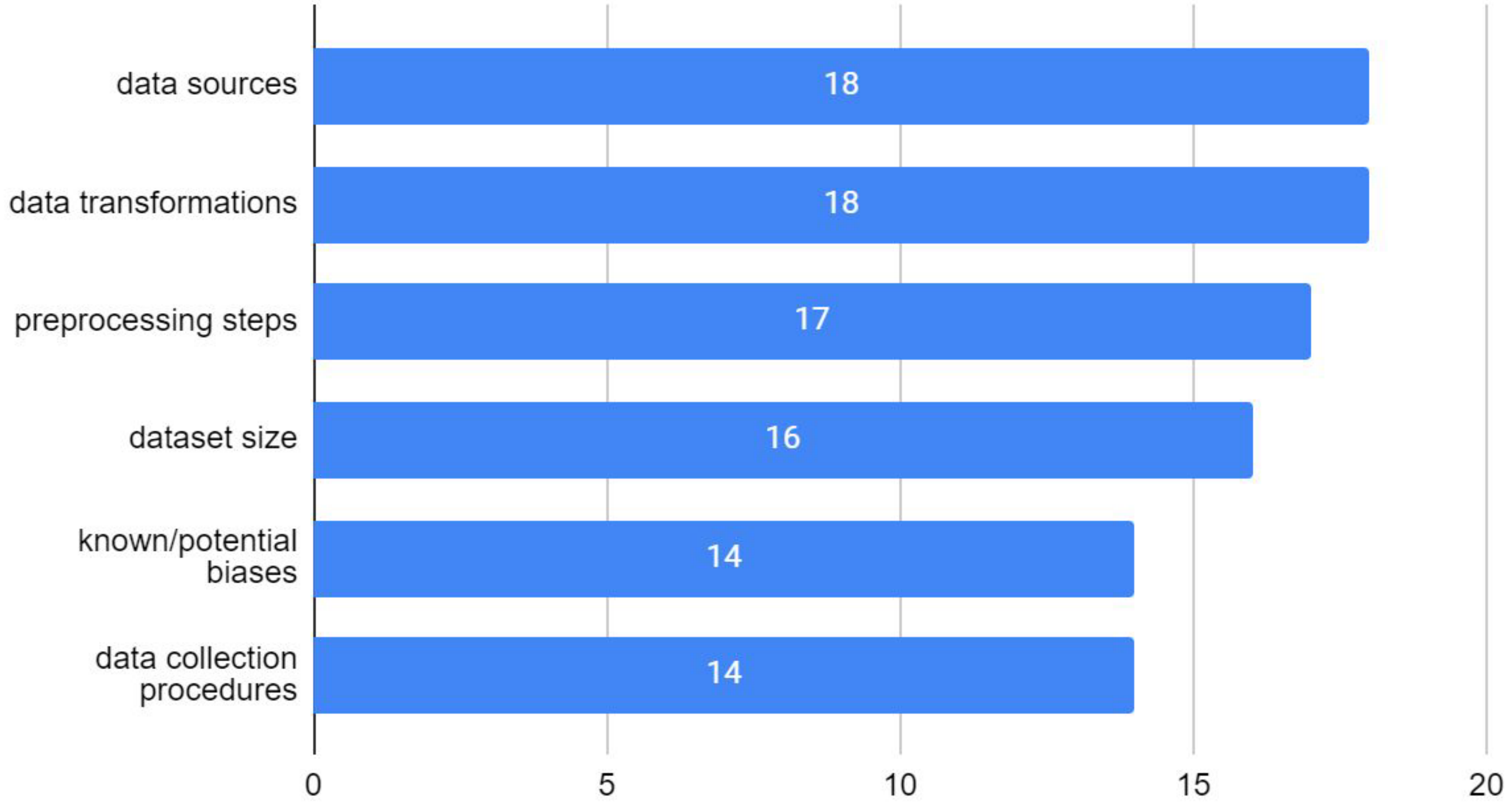
> “There’s a lot of data that’s included in the SBOM that I [don’t] necessarily need, and if some of that data is expensive to calculate, then the tool that gives me the SBOM would run a lot faster if [I didn’t need to include those fields].”

Top Required Fields: SBOM C&A



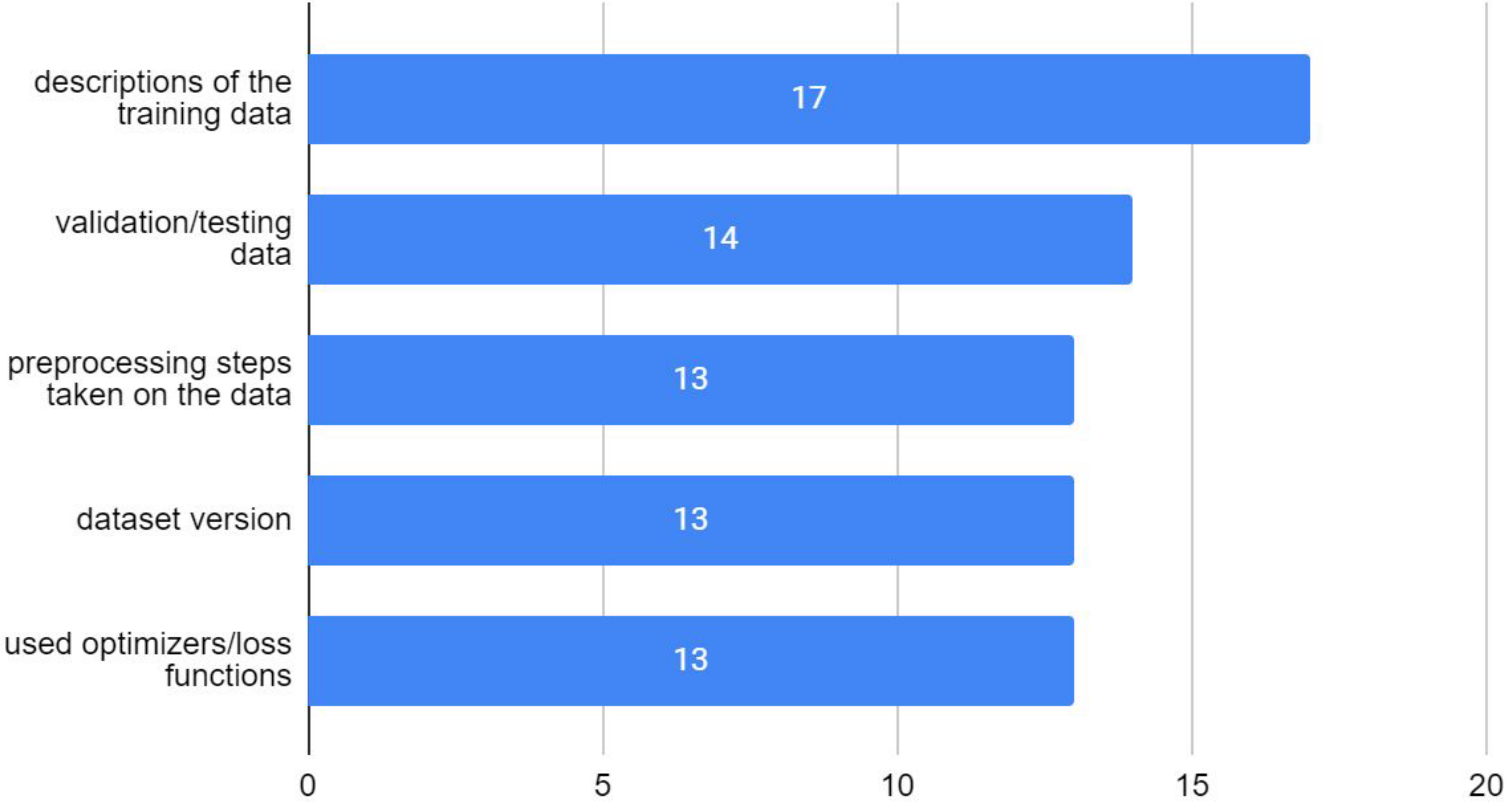
*41 practitioners polled

Top Required Fields: DataBOM



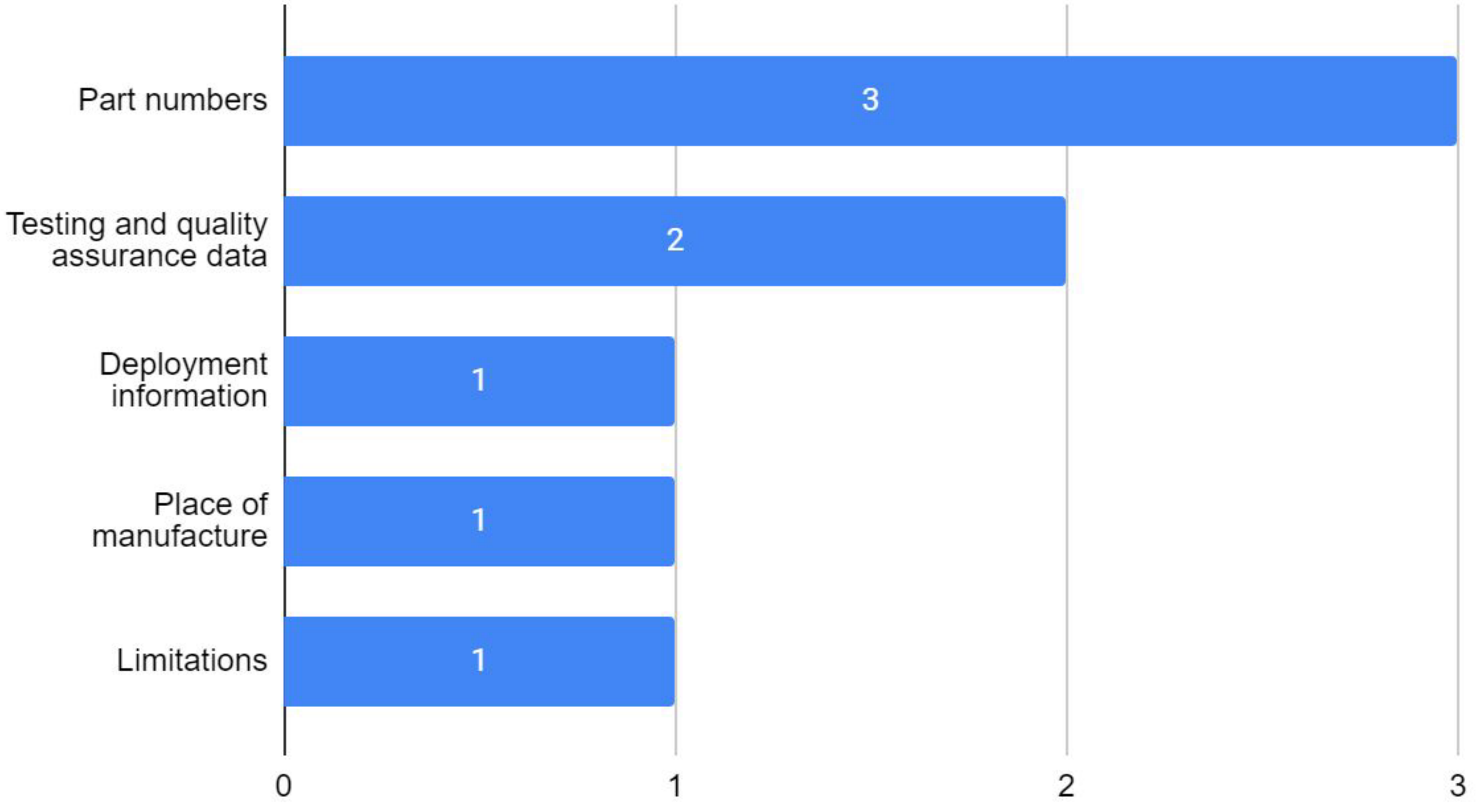
*20 practitioners polled

Top Required Fields: AIBOM



*20 practitioners polled

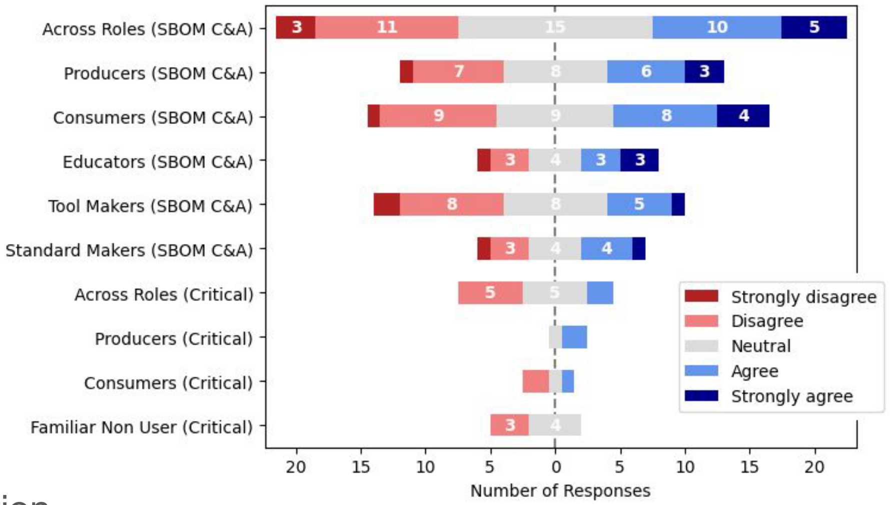
Top Required Fields: CPS



*6 practitioners polled

C5 - Insufficient SBOM tooling

- Lack of consensus among participants
  - Tool makers slightly more negative
- Current tool support is insufficient
  - Lack of multi-language support
  - Poor performance on large projects
  - Comparative lack of tools for SBOM consumption
- ML respondents mostly unaware of appropriate tool support
- Only one CPS practitioner aware of existing tools
  - Suggests tooling does not exist, is insufficient, or is obscure



C8 - Differences across ecosystems and communities

- Varying level of support across ecosystems
  - Python, JavaScript, Ruby, C / C++
- Difficult for languages with no package managers (e.g. C, C++)
- Tools from same standard can vary in quality across languages

> "a big part of the bottleneck is just retrieving all the information that needs to go into the SBOM and getting it from different sources [...] some language communities do a better job of capturing the metadata [to] include in the SBOM."

C10 - SBOMs for legacy packages and repositories

- Challenge generating SBOMs for legacy software
    - Systems no longer maintained
    - Original source code is unavailable
    - Written in older, less common language (e.g. COBOL)
- For existing systems,
    - Should SBOM be created for older release versions?
        - Some software may still be relying on them

> “If ecosystems did start to publish SBOMs, [...] it would be great to see [centralized repository maintainers] go back in time, generate SBOMs for older packages.”

Proposed Solutions

S1: Multi-dimensional SBOM specifications

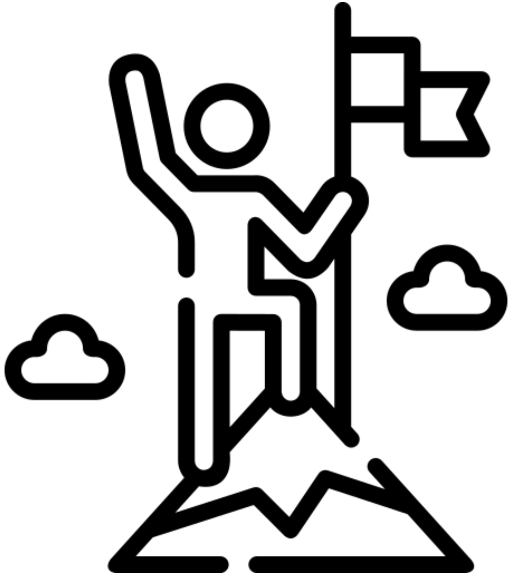
S2: Enhanced SBOM tooling and build system support

S3: Strategies for SBOM verification

S4: Increasing incentives for SBOM adoption

S5: Improving documentation

S6: Techniques for generating software IDs



Proposed Solutions

S1: Multi-dimensional SBOM specifications

S2: Enhanced SBOM tooling and build system support

S3: Strategies for SBOM verification

S4: Increasing incentives for SBOM adoption

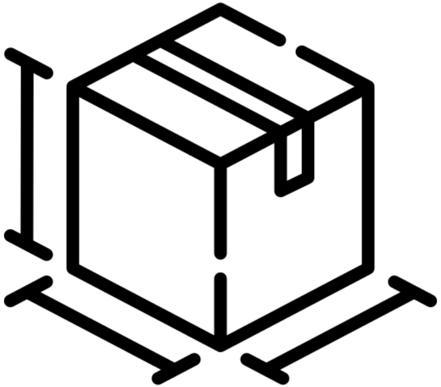
S5: Improving documentation

S6: Techniques for generating software IDs

S1 - Multi-dimensional SBOM specifications

- Use-case
  - Security
  - Licensing
  - Dependency tracking
  - etc.
- Type of software
  - Machine learning model
  - Embedded system
  - Cloud service
  - etc.

- Amount of information documented
  - Granular and detailed
  - High-level and cursory
  - etc.



S1 - Anticipated Benefits

- Make specs easier to understand and reference (C1)
- Make BOM documents shorter and more readable (C1)
- Specify fields to include in particular SBOM (C2, C6)
- Provide indication of expected quality (C6, C9)

> "Even though the minimum requirements that have been provided [...] seem to be or could be construed as daunting, the essence of what needs to be provided in SBOM can be surprisingly simple."

S2 - Enhanced SBOM tooling

- Better language agnostic libraries
  - Foundation for developing SBOM tools
- Language specific SBOM tooling
  - Create tools for different ecosystems
  - Community effort
- ML libraries (e.g. TensorFlow) can
  - Generate AIBOM
  - Provided required information

> “Part of it is just [...] being willing to get in and help out with the quality of those tools.”

> “I imagine [...] that eventually there’ll be [...] something built into TensorFlow or PyTorch [...] that outputs a document in a JSON file [...] that tells you the key elements [like] the hyper-parameters.”

S2 - Enhanced SBOM build system support

- Make build systems SBOM-aware
- Integrate SBOM into package managers
  - Generate SBOM along with or instead of quasi-SBOM
  - Store SBOM with other information
  - Make queryable through API
- Make generating SBOM the default

> "When the recommended way of doing something is the default, then it gets done more often."

S2 - Anticipated Benefits

- Smoother development by using shared libraries / frameworks (C5)
- Additional language support (C5, C8)
- Improved SBOM output and tooling capabilities (C6, C7)
- Easier updating and managing of SBOMs (C4, C14)

> “Increased investment in open source libraries that can be incorporated in end user commercial and open source tools [can address current deficiencies in tooling].”

Conclusion

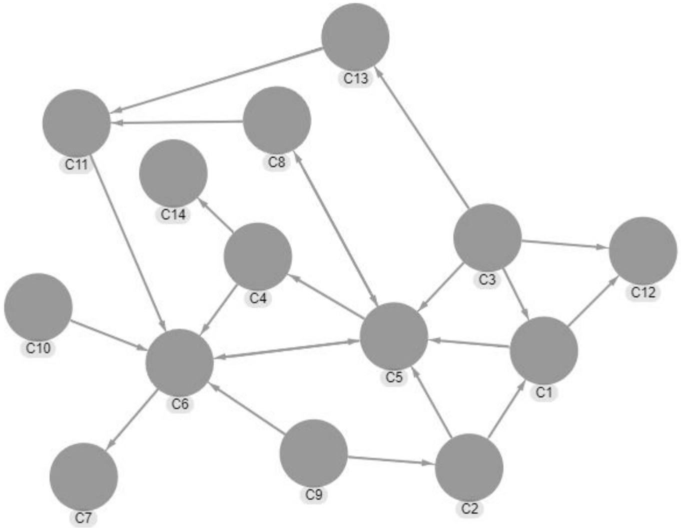
Summary

- SBOMs offer promising solution to problems in supply chain
- But SBOM is a still a young technology with challenges
- To explore these we
  - Conducted 5 surveys with different stakeholder groups
  - Organized 8 follow-up interviews
- To discover
  - How SBOMs are used in practice
  - Challenges encountered by stakeholder groups
  - Actionable solutions to those challenges

Final Thoughts

- Widespread SBOM usage will make software products better
- Before stakeholders fully enjoy benefits many challenges must be overcome
- These challenges are complicated and feed into each other
- We’ve proposed solutions, but implementing will take time, effort, research

Graph of challenge relationships



Thesis Contributions

- Provides a clearer picture of
    - How and why SBOMs are used in practice
    - What use cases are still unmet
- Considers 5 stakeholder groups and 4 BOM types
    - Furthers discussion on AI / DataBOM requirements
- Explores 14 main challenges to SBOM
- Brings light to 12 novel issues not mentioned in prior works
- Proposes 6 actionable solutions to identified challenges

Bibliographical Note

- Paper supporting the content of this thesis was written in collaboration with
    - members of the SEMERU research lab at William & Mary
    - researchers from the University of Sannio and the University of Victoria
- It is currently under review for publication at ICSE.

Stalnaker, T., Wintersgill, N., Chaparro, O., Penta, M., German D., & Poshyvanyk, D. (2023, March). *BOMs Away! Inside the Minds of Stakeholders: A Comprehensive Study of Bills of Materials for Software Systems*. Under Second Round Review.

Questions?

References and Image Credits

[1] https://www.idtheftcenter.org/wp-content/uploads/2023/01/ITRC_2022-Data-Breach-Report_Final-1.pdf

Freekpik

Spam, Nonsense, Blank, Survey, Selection, Detective, Confused, Gavel, Climbing, Anxiety, Summit, Multi-dimensional, Thinking

monkik (https://www.flaticon.com/authors/monkik)

Puzzle, Build system

Eucalypp (https://www.flaticon.com/authors/eucalyp)

Incentive, Puzzle lightbulb

wanicon (https://www.flaticon.com/authors/wanicon)

Distribution, Robo arm

Reference and Image Credits

Copy Paste	Tempo_doloe	https://www.flaticon.com/authors/tempo-doloe
Repeats	Dewi Sari	https://www.flaticon.com/authors/dewi-sari
ChatGPT	Zulfa Mahendra	https://www.flaticon.com/authors/zulfa-mahendra
Giftcard	Boris farias	https://www.flaticon.com/authors/boris-farias
Evil face	Fir3Ghost	https://www.flaticon.com/authors/fir3ghost
Lightbulb	Bartama Graphic	https://www.flaticon.com/authors/bartama-graphic
Cycle Gear	Icon home	https://www.flaticon.com/authors/icon-home
Floppy disk	Those Icons	https://www.flaticon.com/authors/those-icons
MRI	Design Circle	https://www.flaticon.com/authors/design-circle
AI Brain	imaginationlol	https://www.flaticon.com/authors/imaginationlol
Annotation	justicon	https://www.flaticon.com/authors/justicon
Thinking	Prashanth Rapolu 15	https://www.flaticon.com/authors/prashanth-rapolu-15
Nutshell	Mihimihi	https://www.flaticon.com/authors/mihimihi
Benefits	zero_wing	https://www.flaticon.com/authors/zero-wing