ScanCode to reuse FOSS safely, with FOSS

Agenda

AboutCode

- About me, AboutCode and nexB
- Why should we scan code?
  - Licenses change, vulnerabilities present, quality issues, regulations
- How to communicate? Open Standards:
  - PackageURL, Vers, SBOMs (SPDX, CycloneDx), VDRs, VEX
- Problems and solutions
  - Know your full list of exact dependencies
  - what can you do to avoid issues
  - Hard to identify all packages, too much vulnerabilities
- Why use AboutCode?
- Questions

About Ayan

AboutCode

- Core maintainer of ScanCode (scancode-toolkit and scancode.io)
  - also contributes to and helps maintain other AboutCode tools: license-expression licenseDB scancode-workbench PURLdb
- Working primarily on license detection and package identification, data summarization and
- Google Summer of Code Mentor at AboutCode (2021-2025)
  - participant in GSoC2020 and GSoD2019
- FOSS Maintainer at AboutCode (a non-profit org)
  - asmahapatra@aboutcode.org
  - GitHub: https://github.com/AyanSinhaMahapatra/
  - LinkedIn: https://www.linkedin.com/in/ayansinhaju/

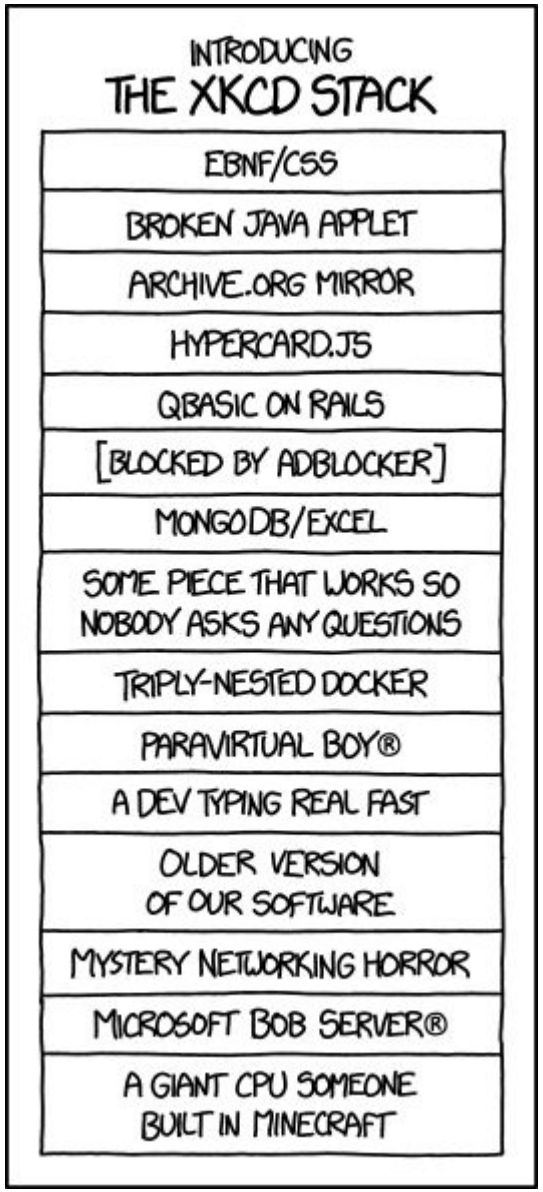
AboutCode and nexB

AboutCode

- AboutCode's FOSS-first mission: FOSS for FOSS
  - Open source tools and open knowledge base (AboutCode stack)
  - Simple and practical standards (Package-URL)
  - Applications for Legal Business users (DejaCode, also FOSS) with APIs
- Trusted experts on Software Composition Analysis (SCA) since 2007
  - Creator of Package-URL: https://github.com/package-url
  - Co-founders of SPDX: https://spdx.org
  - Contributors to CycloneDX: https://cyclonedx.org
  - Co-founders of ClearlyDefined: https://clearlydefined.io
- nexB: professional services for SCA
  - 800+ SCA projects completed to-date
  - Sponsored development for AboutCode projects
  - Technical support and advisory for SCA process, and deployments

Modern software ecosystem

- Ever more FOSS software packages are reused
  - small apps routinely embed 500 FOSS packages
  - large apps: 10,000!
- Everyday you have new vulnerabilities, license problems and package updates in your package dependency trees
  - Impossible to check this manually!
- Each package ecosystem is different
- Goal: Discover the problems and help alleviate the pain



Source: https://xkcd.com/1636/

Why is Software License important?

- FOSS: Freedom
- Freedom and Responsibilities
  - Can we use the software in different scenarios?
  - Can we modify and redistribute freely, under my choice of terms?
  - Give credit, generate attribution
- See License categories for more details
- Copyrights:
  - Copyright notices often have to be included and redistributed
- History of Litigation
  - GPL based court rulings to Distribute Source Code

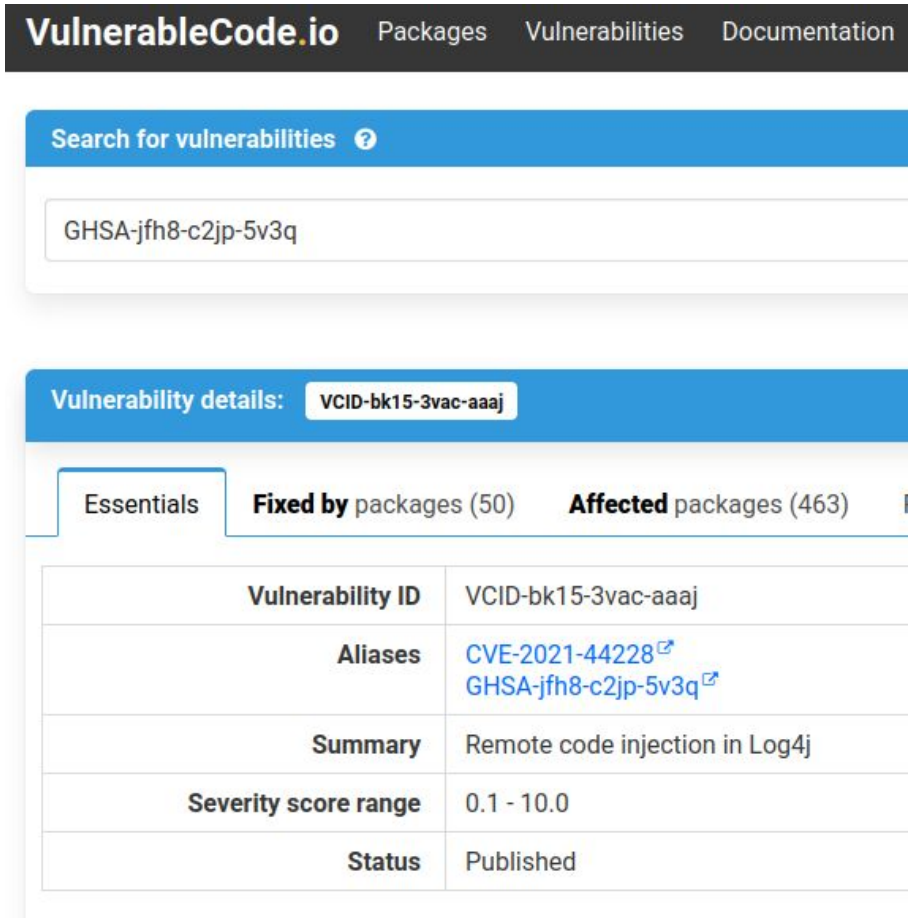
Why is identification important?

- Modifications can be released under different terms
- License could change between versions
  - packages/products often decide to change their license
    - https://redis.com/blog/redis-adopts-dual-source-available-licensing/
    - https://www.elastic.co/blog/elastic-license-update

AboutCode

Why is identification important?

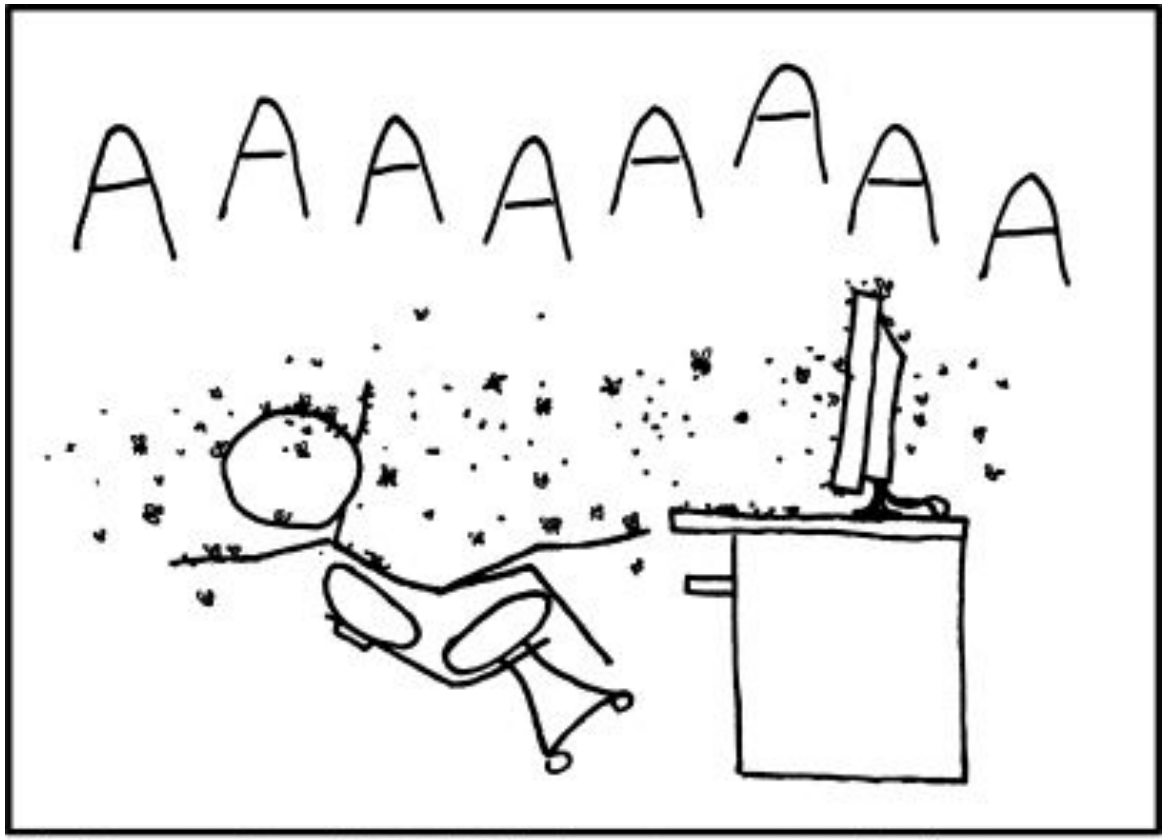
- Vulnerabilities are introduced and fixed by versions (or not!)
- False positives!



Sources:
https://public.vulnerablecode.io/vulnerabilities/VCID-bk15-3vac-aaaj?search=GHSA-jfh8-c2jp-5v3q
https://github.com/advisories/GHSA-jfh8-c2jp-5v3q

Why is Software Quality important?

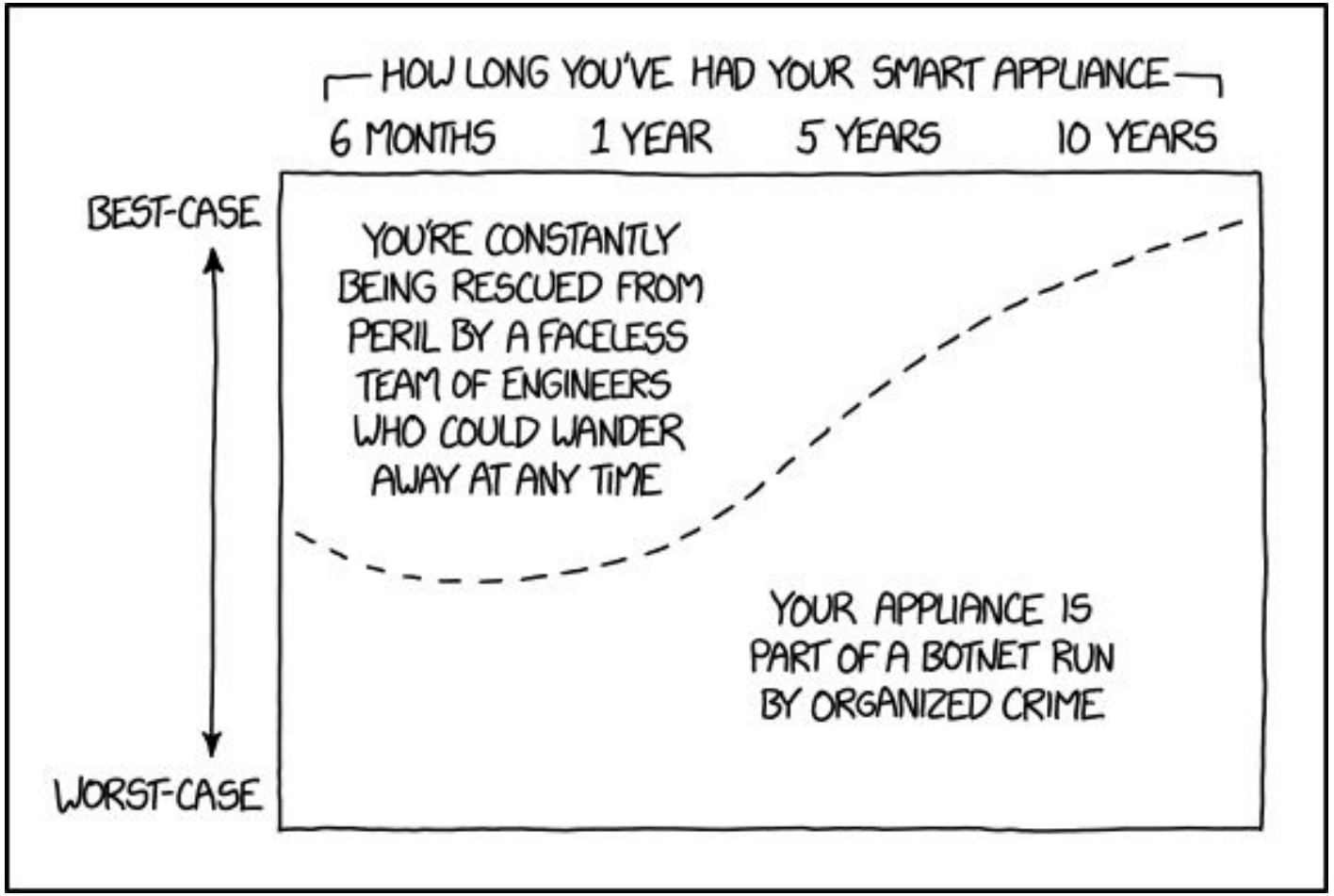
- Better maintained: more secure
  - at least a correlation!
- code review, branch protection and other quality checks are important
- Great FOSS projects on quality:
  - OpenSSF Scorecard
  - endoflife.date
  - CHAOSS: auger, grimorelab



Source: https://xkcd.com/797/

How to communicate? SBOMs, VDRs

- How to disclose security vulnerabilities in my software?
  - lots of legacy software being used all around us
- What are the software licenses for all the packages used?
- Machine Readable!
- On every release



Source: https://xkcd.com/1966/

And really why?

AboutCode

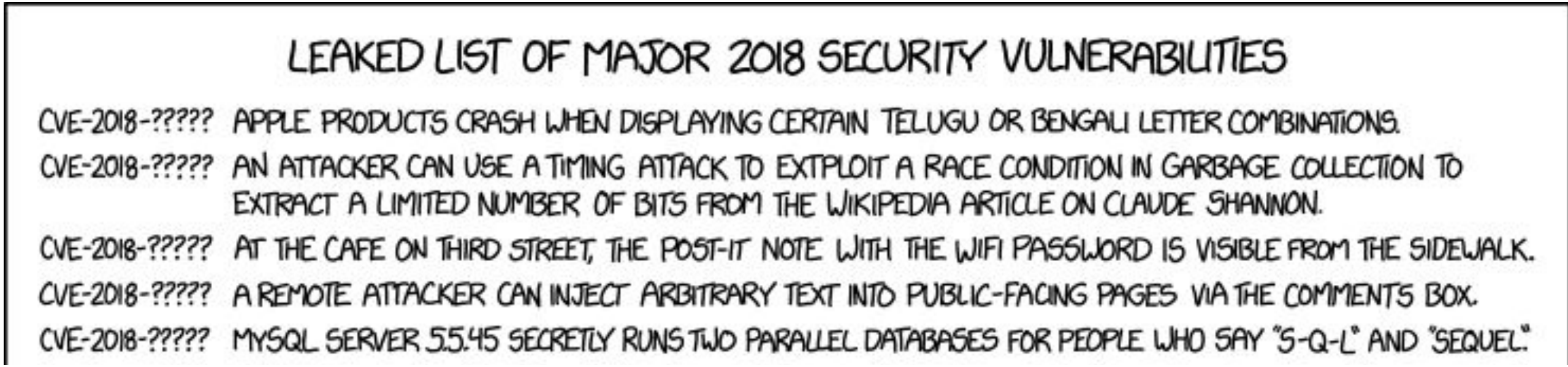
In the US and in Europe, it's the law.

- US: executive order 14028
  - SBOM for any software business with the government.
- Europe:
  - the CRA (Cyber Resilience Act)
  - Maintainers, Open Source Stewards, Manufacturers
  - https://github.com/orcwg/cra-hub/blob/main/faq.md
- India:
  - CERT-In SBOM guidelines
  - CSCRF for SEBI REs
- Often required by companies: using a product/acquiring company
- Similar legislation/requirements likely in everywhere else

What are the key ingredients?

AboutCode

- standard package identifiers: PURL
- standard license identifiers: SPDX
- standard vulnerability identifiers: CVE, GHSA etc
- standard documents to exchange all this information:
  - SBOMs: CycloneDx, SPDX
  - VEX: CycloneDx VEX, OpenVEX, CSAF VEX



Source: https://xkcd.com/1957/

PackageURL

Started in ScanCode to uniquely identify packages.

- pkg:type/namespace/name@version?qualifiers#subpath
  - Specification: https://github.com/package-url/purl-spec
- PURL examples:
  - pkg:deb/debian/curl@7.50.3-1?arch=i386&distro=jessie
  - pkg:github/package-url/purl-spec
  - pkg:pypi/django@1.11.1
  - pkg:rpm/fedora/curl@7.50.3-1.fc25
  - pkg:golang/google.golang.org/genproto#googleapis/api/annotations
- Vers: https://github.com/aboutcode-org/univers/

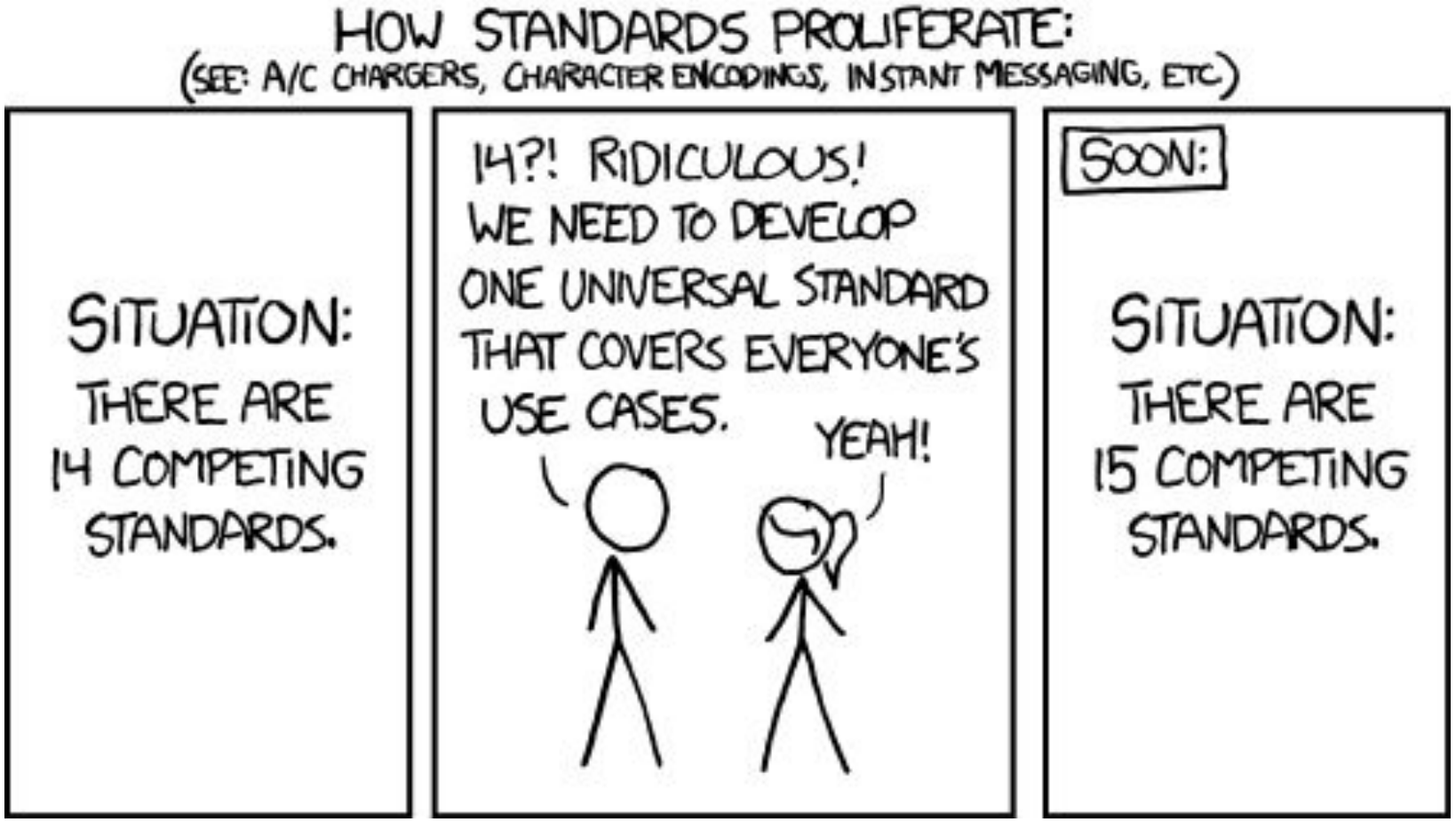
Are these just more standards?

PackageURL (PURL):

An identifier to uniquely identify and download packages

Vers:

Version range specification for package requirements



Source: https://xkcd.com/927/

Who is using PackageURL and Vers?

Everyone!

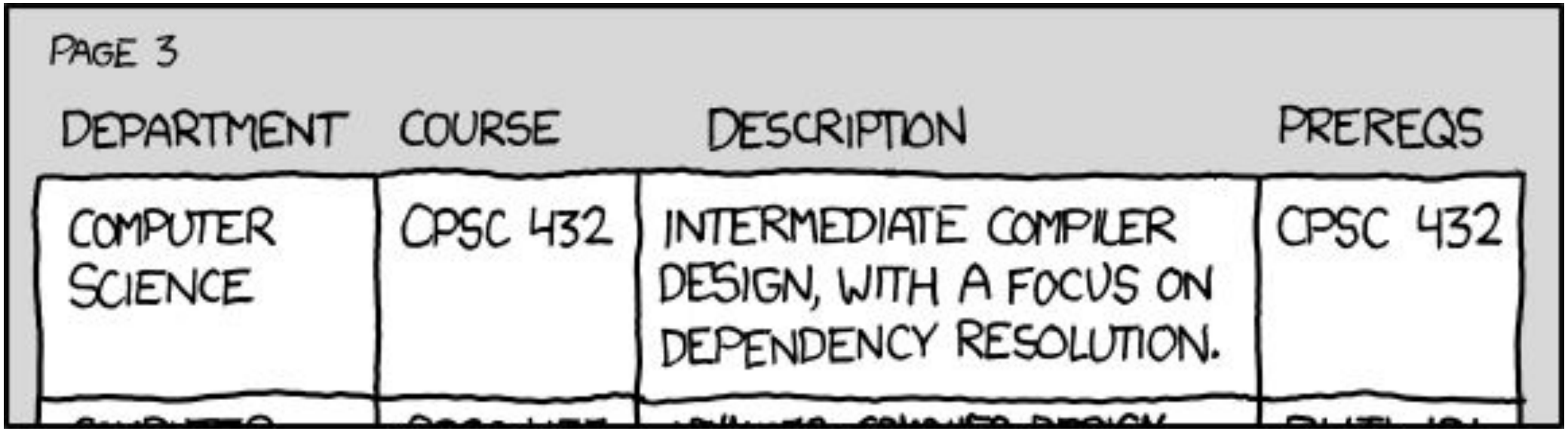
- GitHub Dependency Submission API
- OWASP Dependency-Track
- Two major SBOM standards: CycloneDX and SPDX
- OSS Index
- OSV Schema and OSV.dev (Google)
- AboutCode tools: Scancode Toolkit scancode.io dejacode vulnerablecode
- ORT: OSS Review Toolkit, Osselot
- Anchore, Trivy, Microsoft and GitHub, Chainguard, Snyk
- cve.org, NVD (soon)
- Vers is used at vulnerablecode, Google OSV, AppThreat vulnerability-db

4 Fs of Open Source

- The Three Fs of Open Source Puppy Care: Michael Winser
- relationships with dependencies are important!
- Complete dependency graph
- Fix
  - Engage with maintainers, open Issues/PRs
- Fork
  - Maintain your version, apply patches
- Forget
  - Use something else!
- Fund!

Dependency resolution issues

- Different package versions for the same requirements
- Different results across algorithm/time/environment
- could be useful! Non-vulnerable dependency resolution
- Use Lockfiles!
- Build systems have access to exact dependencies



Source: https://xkcd.com/754/

Package identification can be hard

- Most scanners only scan package manifests!
- Code included from different origins
  - vendored (copied partially/fully)
  - distributed with binaries (maven uberjars, jars inside jars)
  - Code matching (MatchCode and PurlDB)
    - Exact archive and file matching
    - Exact and approximate file tree and subtrees matching
- Finding source repo is not trivial:
  - metadata on source repo missing/incorrect
  - many binaries are compiled from the same source package/monorepo
- Customized build systems + metadata formats together

Vulnerabilities

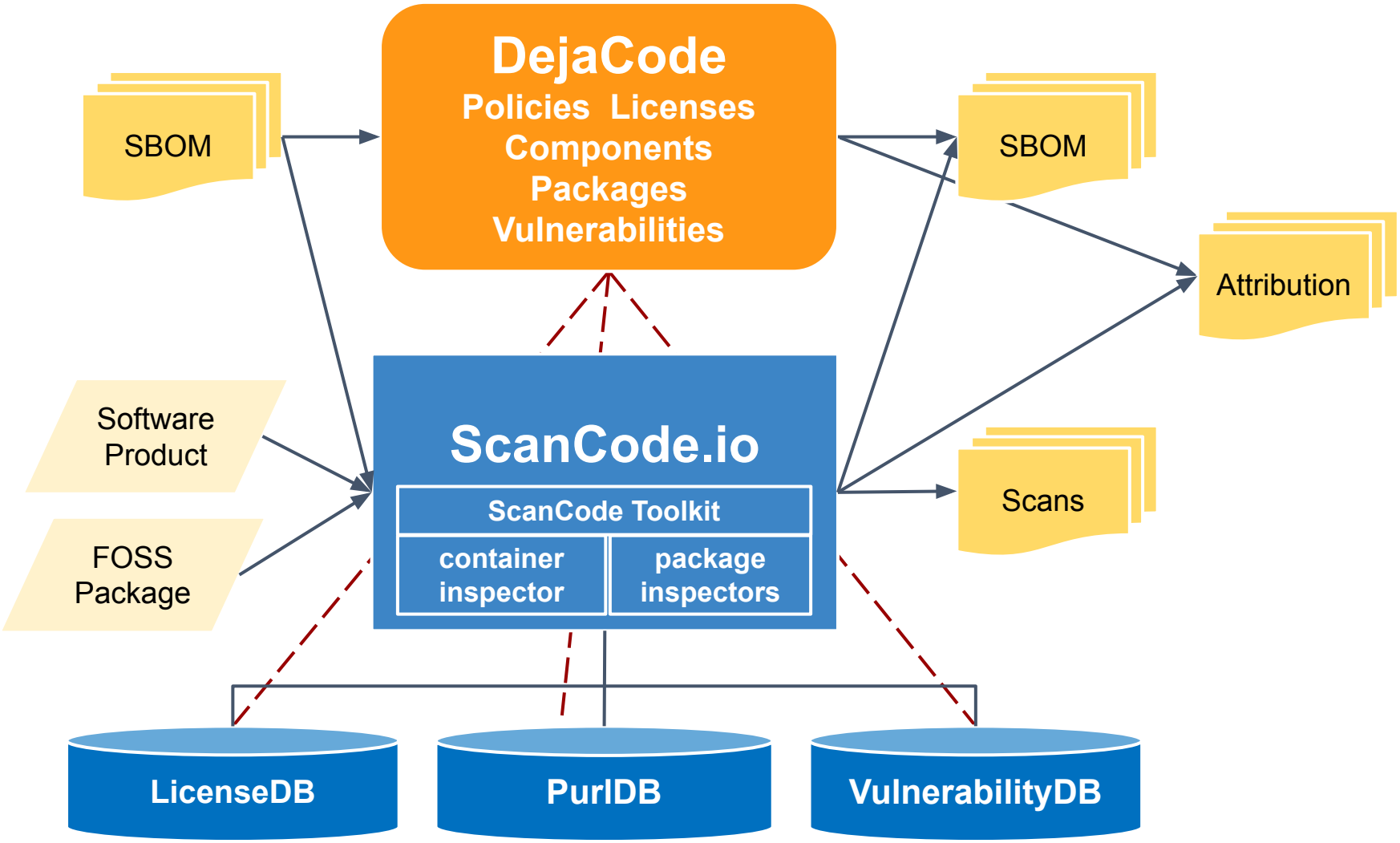
- Transitive dependencies can be exploited
- dependency graph: how was a specific package version introduced
- might not be applicable
  - vulnerable code not used (reachability)
  - not deployed (test/docs/other)
- has this vulnerability been exploited
- score/severity can be good indication, but not absolute
- CI/CD or build dependencies can be problematic too

What AboutCode is doing differently

- Fully open source
- options: CLI tool, Github action, web app, scans: containers, source/binary etc
- supports and working with package ecosystems
  - to build better metadata, more transparency
  - solve ecosystem wide problems at once
- Don’t scan twice
  - Open Data on Licensing, Vulnerabilities
  - Reuse Scan Results
  - Only scan the different parts
- Large community
  - OSPOs, Security, Lawyers, Specifications, Developers
- Developer advocacy

The AboutCode stack:

AboutCode



AboutCode: Who is using it?

(based on public data)

Most FOSS Orgs, many commercial and open source SCA providers use our libraries or standards

- Most FOSS Foundations
- Five of the top big tech companies
- A leading database company, a leading Linux company
- 2 leading code hosting companies
- European and US government agencies
- All major European car manufacturers and most of their vendors
- Major US chip and microprocessor providers
- All SBOM and VEX standards
  - Used to create a database of permissive code to train an open code LLM
  - See https://huggingface.co/blog/starcoder2

Other FOSS SCA tools and projects

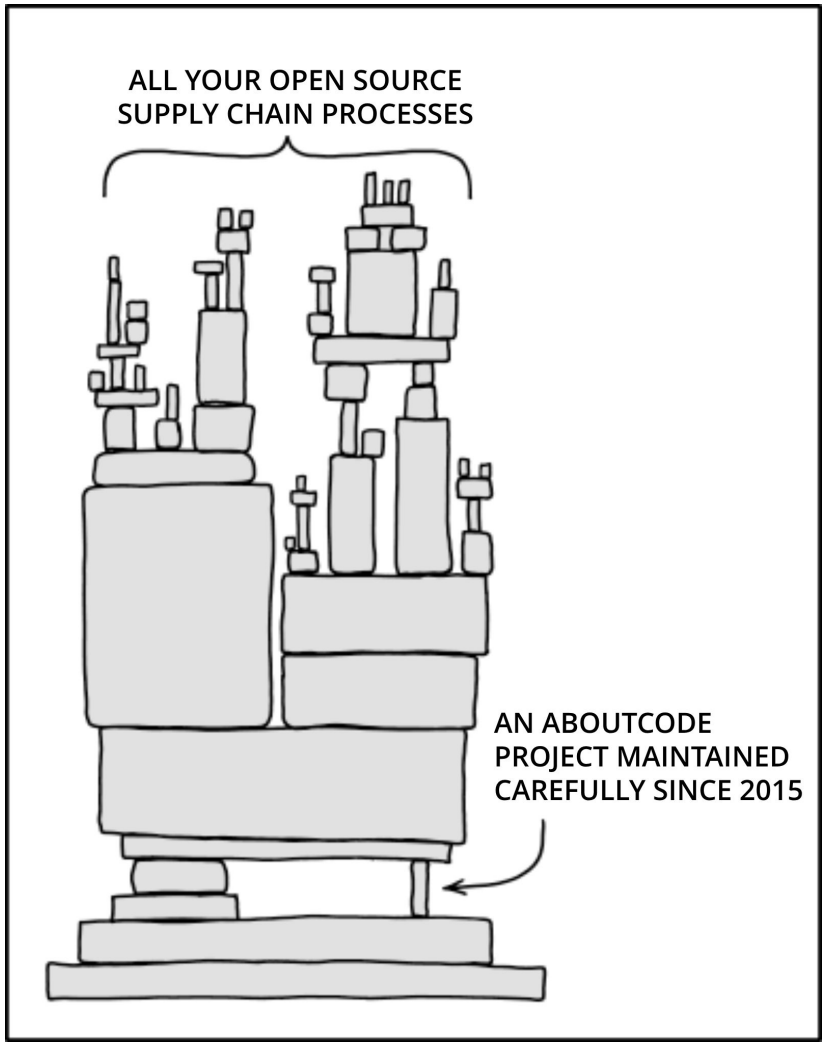
AboutCode

- ORT: OSS Review Toolkit (Uses ScanCode)
- FOSSology (Uses ScanCode)
- TERN (Uses ScanCode)
- ClearlyDefined (Uses scancode)
- OWASP DependencyTrack
- DepScan (and other AppThreat projects)
- CycloneDx cdxgen
- Anchore: syft, grype

AboutCode also needs your help!

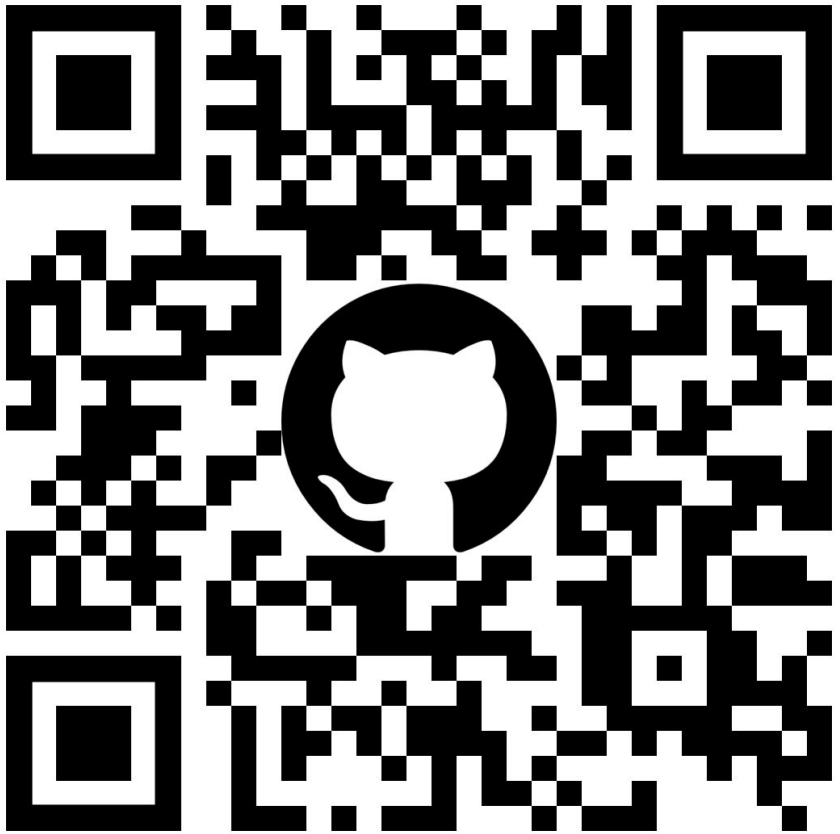
AboutCode

- Contribute to an AboutCode project with code, documentation, use cases, bug reports
  - https://github.com/aboutcode-org
- Sponsor AboutCode project maintainers, accelerate development of new features https://github.com/sponsors/aboutcode-org
- Buy support, implementation, and advisory services from nexB to pay the maintainers
- Join the community:
  - https://www.aboutcode.org/
  - https://matrix.to/aboutcode-org_discuss



"Dependency" by xkcd, Modified text from original

Questions?



Note: QR Codes are without any tracking

Credits

Special thanks to all the people who made and released these excellent free resources:

- All the open source software authors that make AboutCode possible
- xkcd comics under cc-by-nc-2.5
- Presentation template by SlidesCarnival