

Semantic Analysis for Deduplication of Security Findings in DevOps Security Tool Reports

Abdullah Gulraiz, Mar 14 2022, Kick-off Presentation

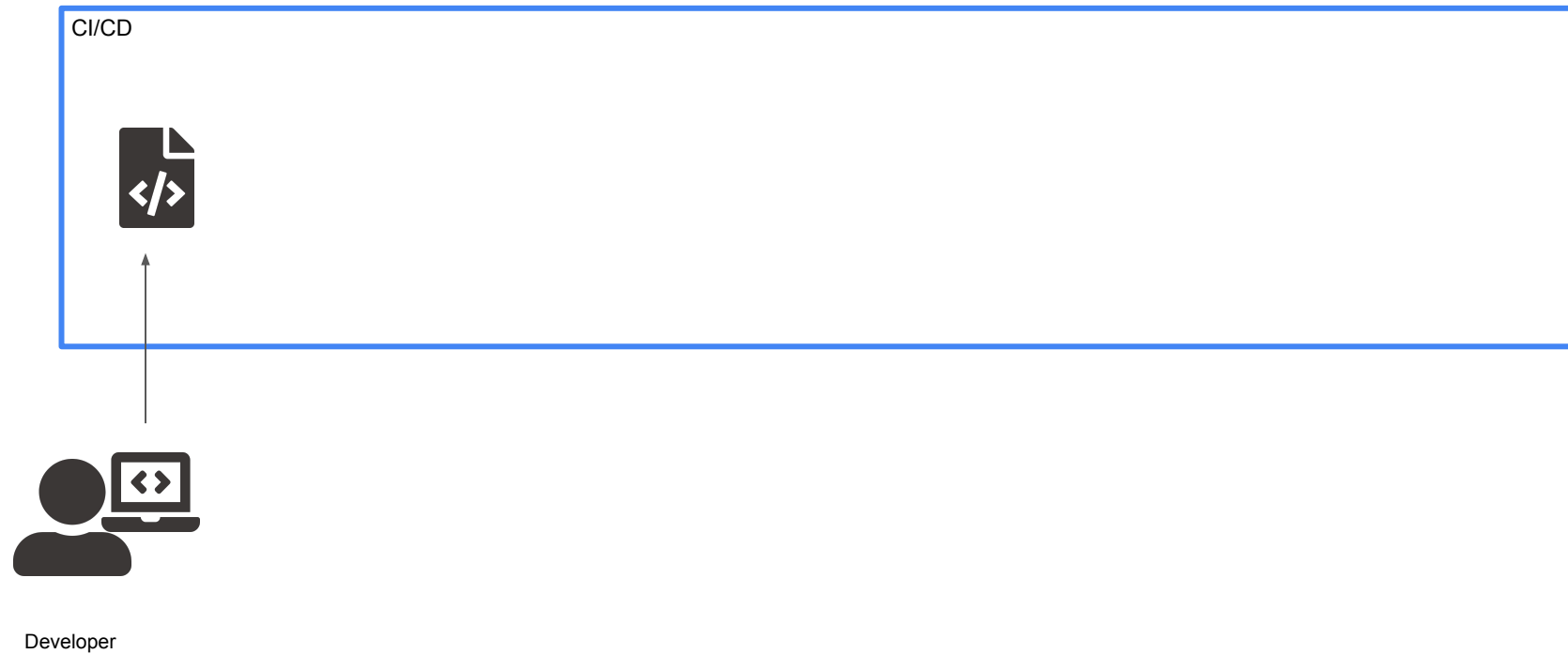
Chair of Software Engineering for Business Information Systems (sebis)
Faculty of Informatics
Technische Universität München
www.matthes.in.tum.de

Outline

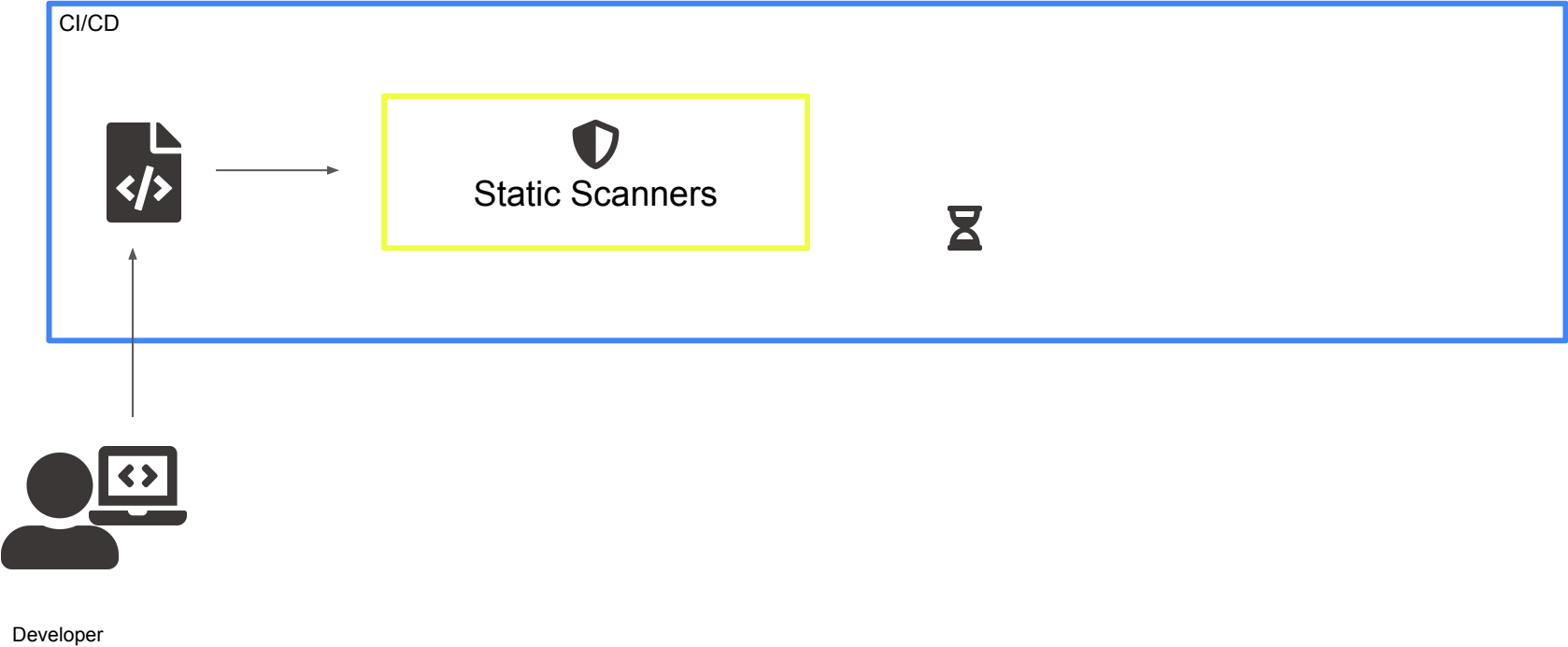


1. Problem illustration
2. NLP-based approach
3. Research questions
4. Progress
5. Timeline

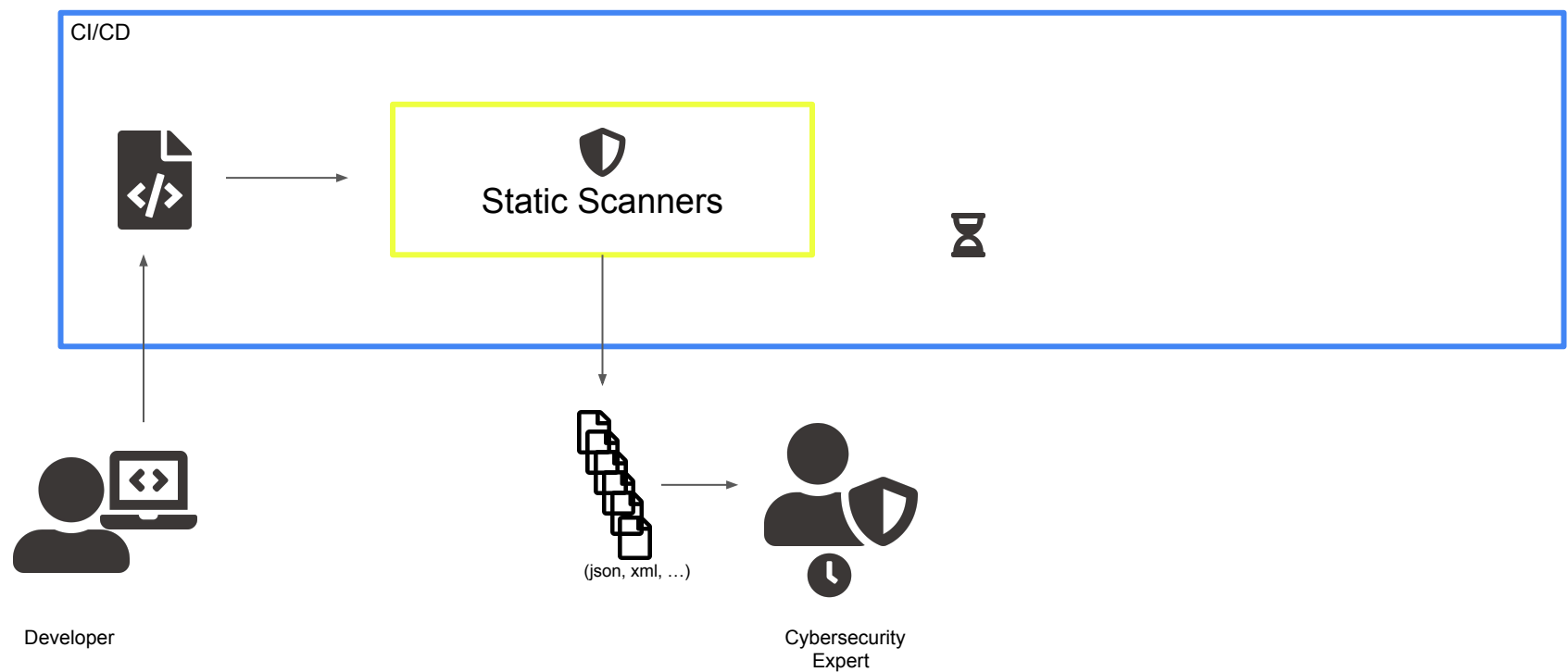
“Security Findings in DevSecOps”



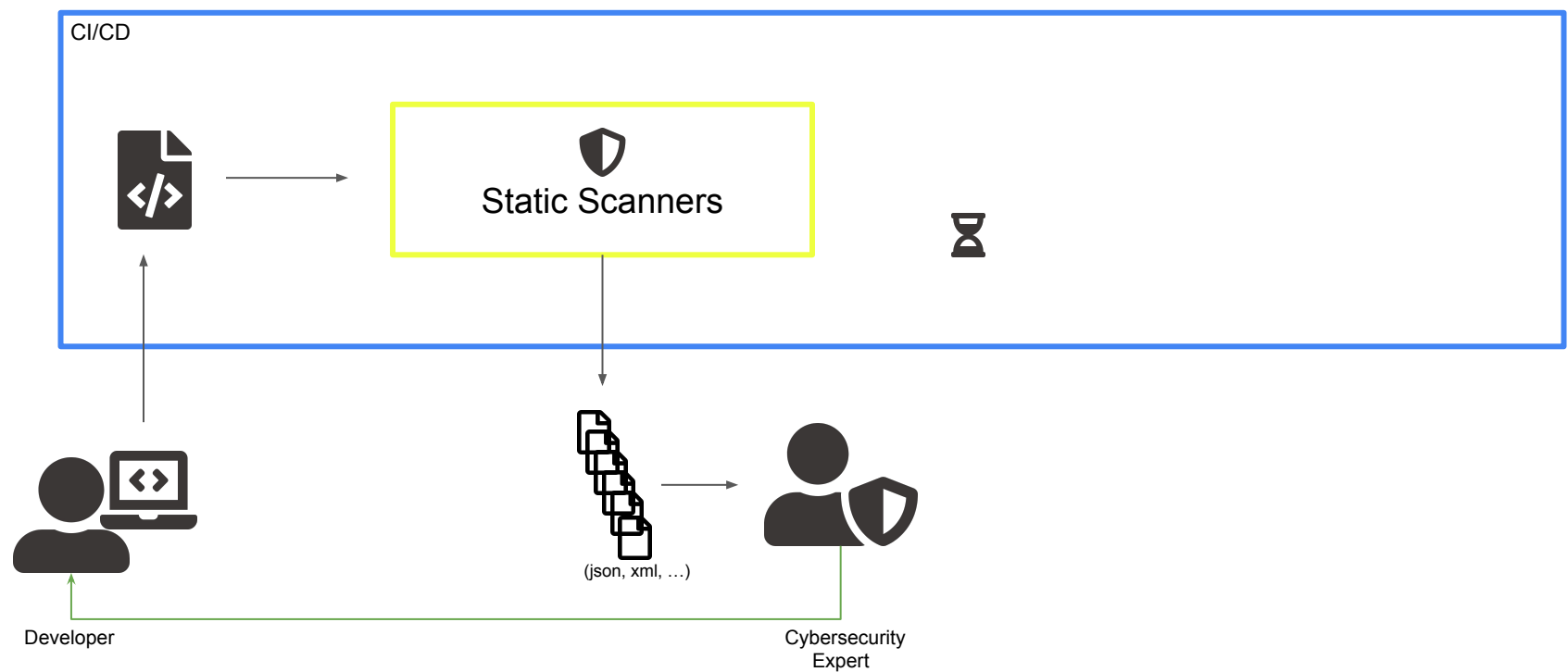
“Security Findings in DevSecOps”



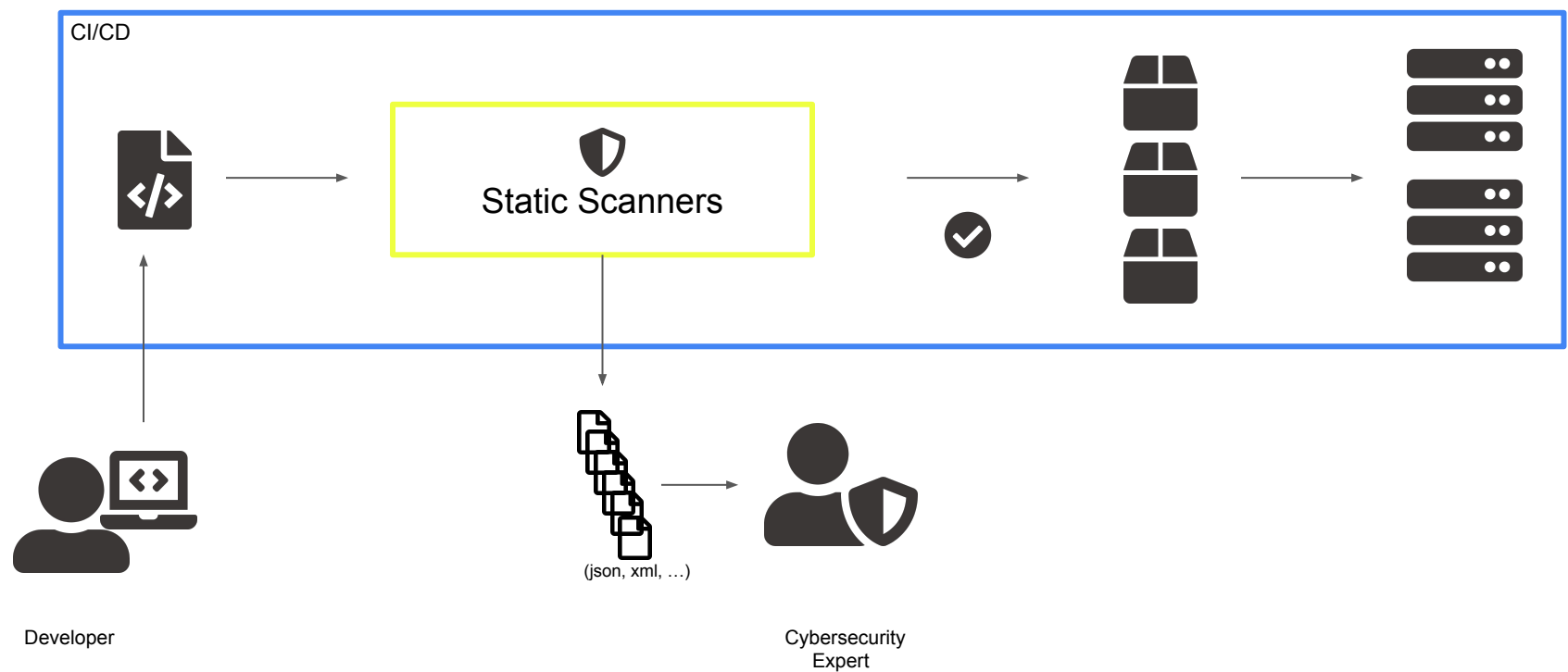
“Security Findings in DevSecOps”



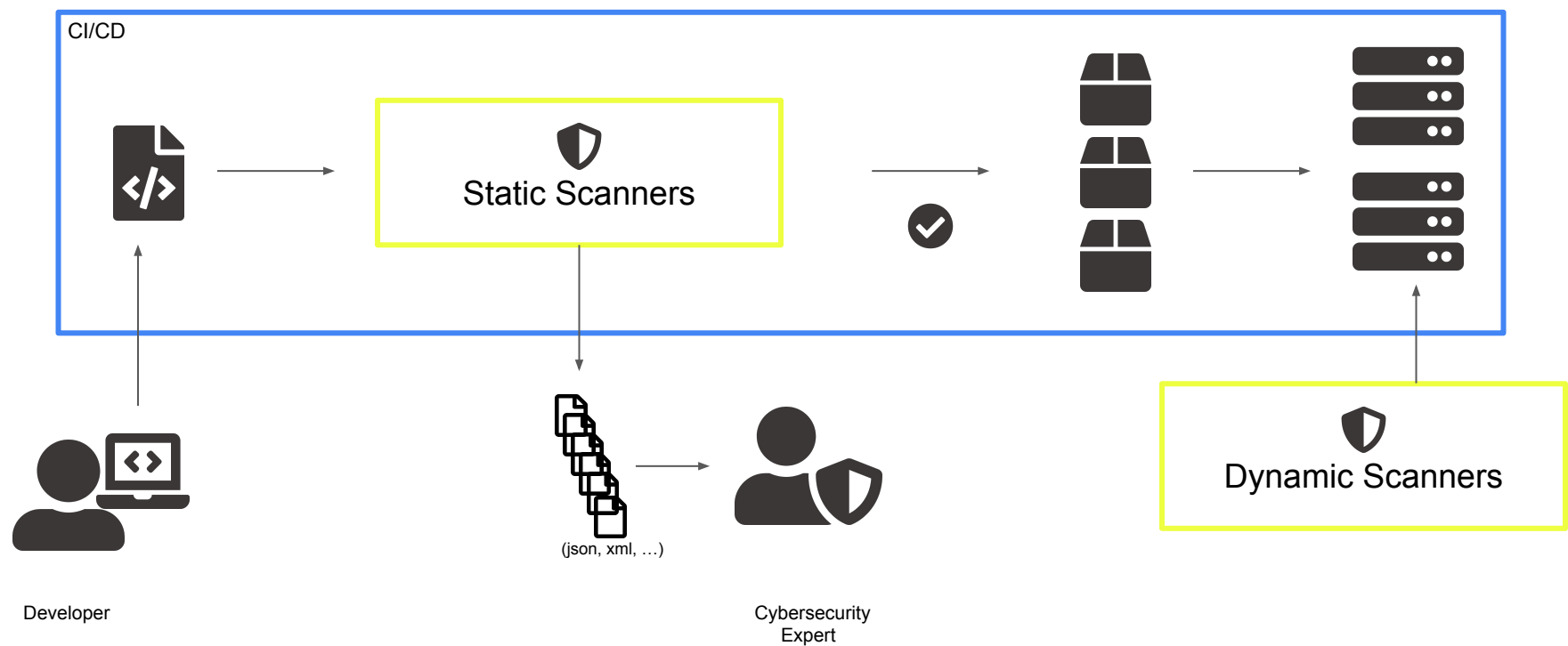
“Security Findings in DevSecOps”



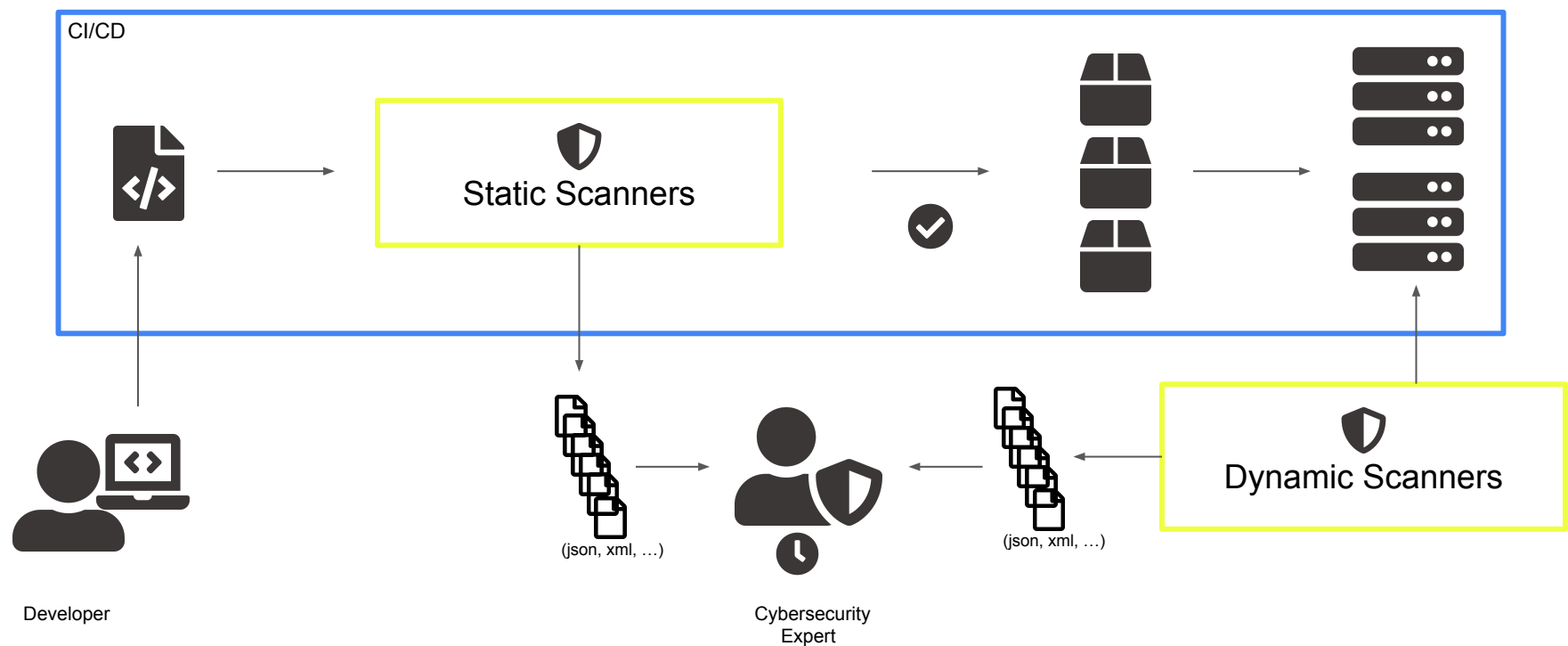
“Security Findings in DevSecOps”



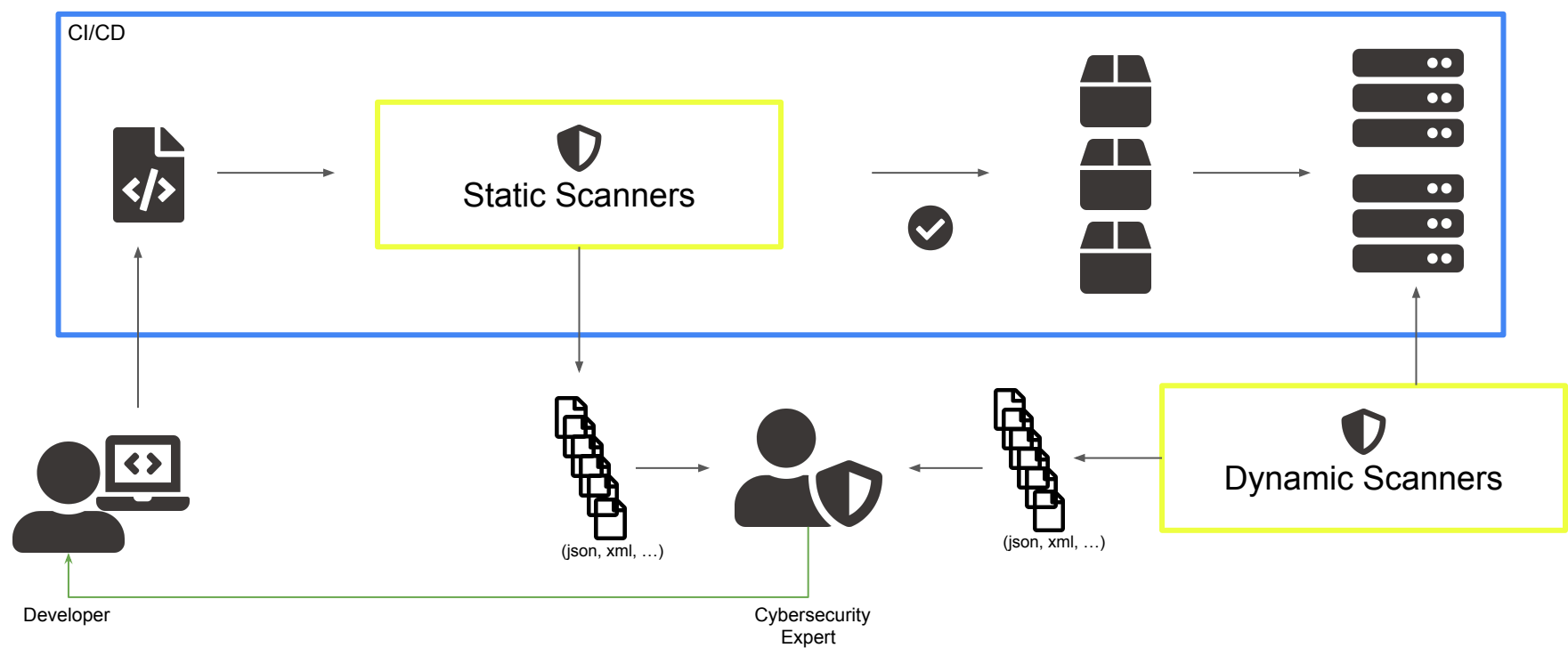
“Security Findings in DevSecOps”



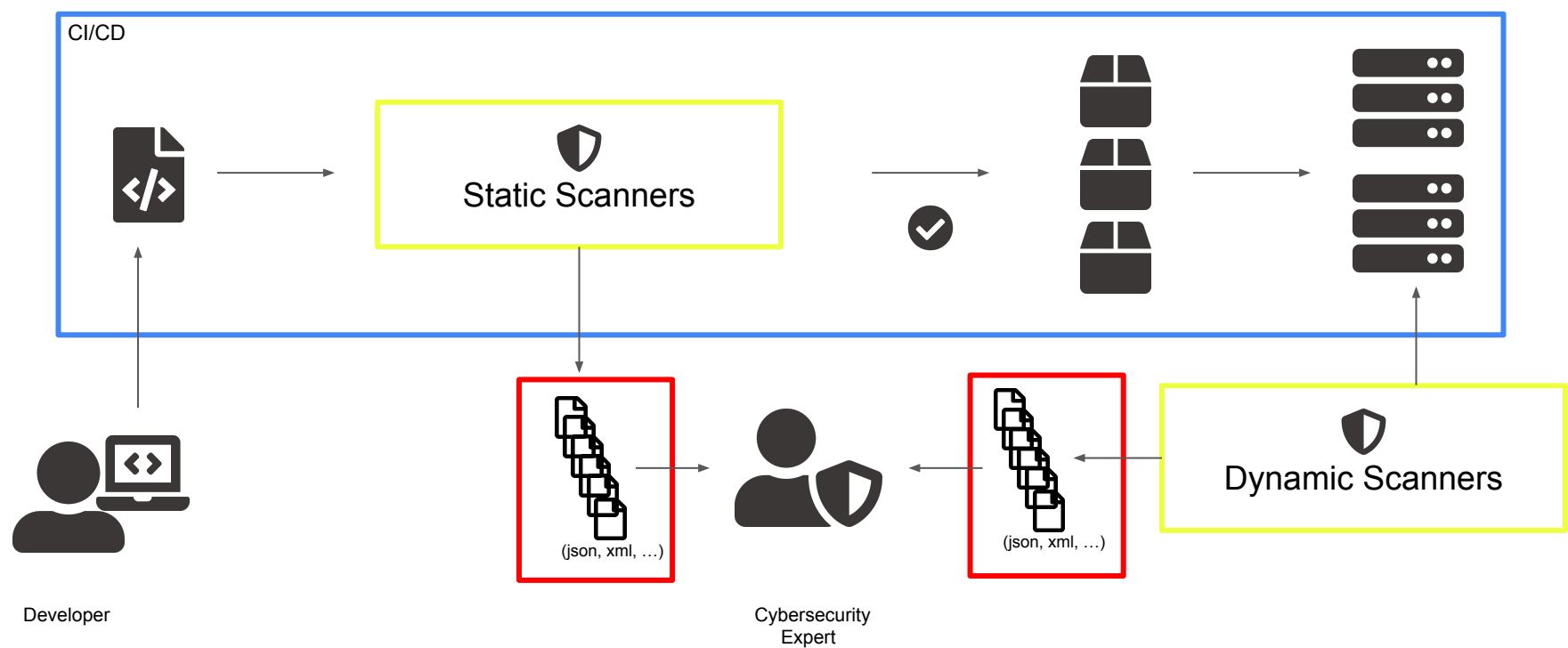
“Security Findings in DevSecOps”



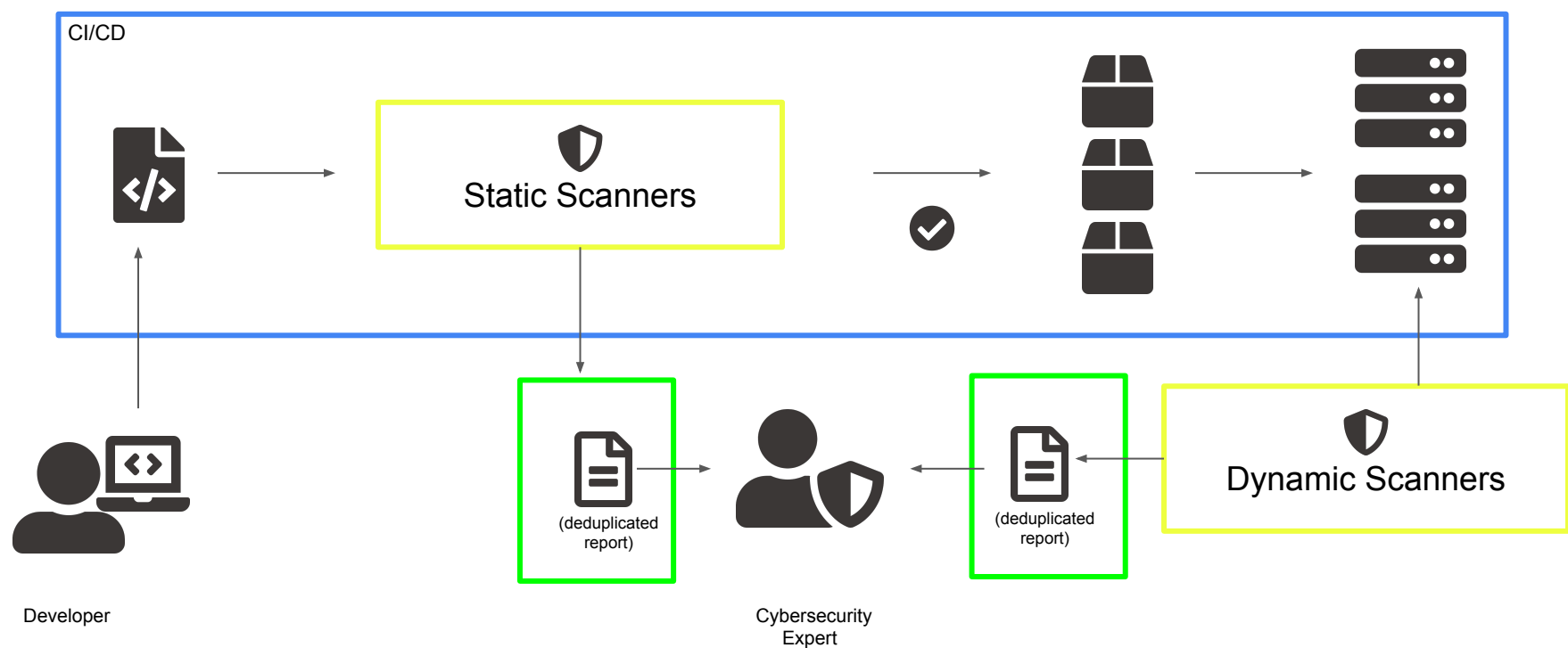
“Security Findings in DevSecOps”



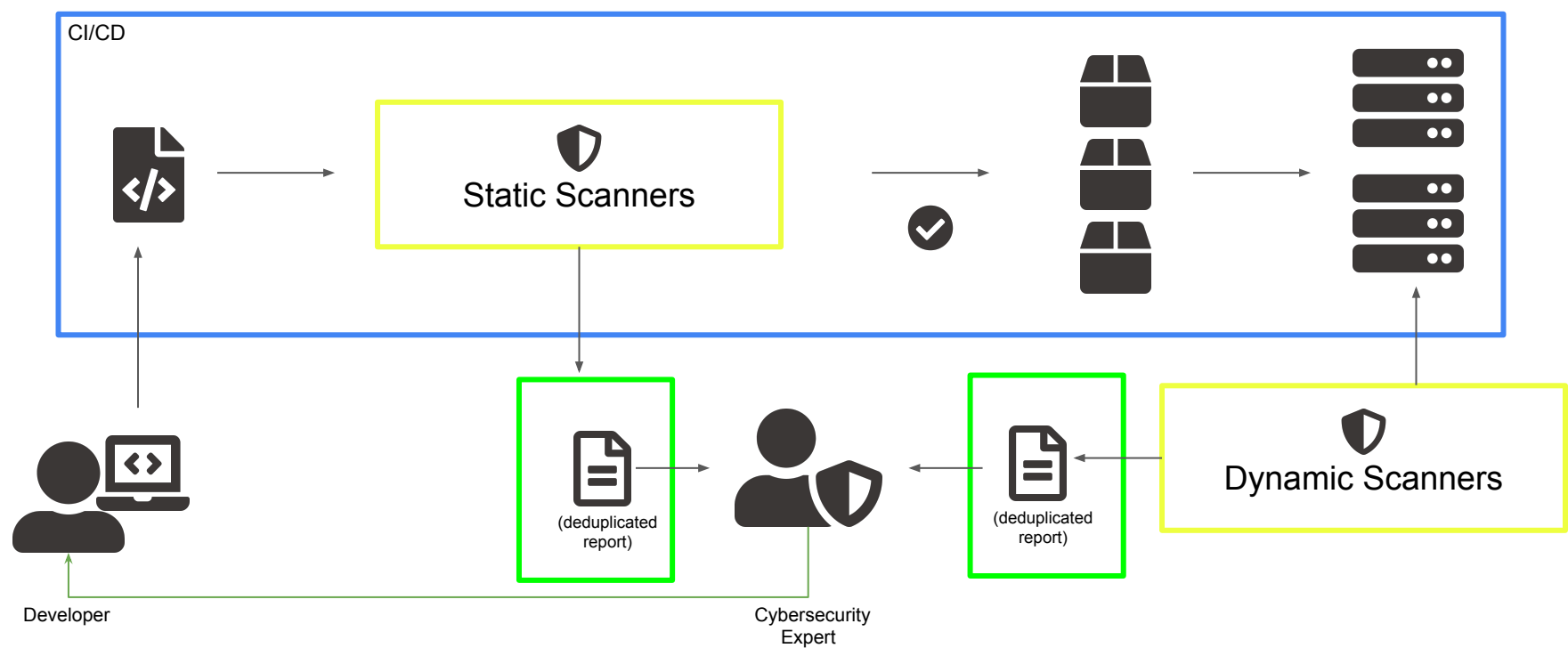
“Security Findings in DevSecOps”



“Security Findings in DevSecOps”



“Security Findings in DevSecOps”



Problem illustration

Finding reports:

- Contain natural language descriptive texts
- Result from tools with overlapping use cases and hence contain duplicates
- Contain multiple and different fields to analyze text from
- Have very domain-specific knowledge

```
{
  "pluginid": "10098",
  "alert": "Cross-Domain Misconfiguration",
  "name": "Cross-Domain Misconfiguration",
  "riskcode": "2",
  "confidence": "2",
  "riskdesc": "Medium (Medium)",
  "desc": "<p>Web browser data loading may be possible, due to a Cross Origin Resource Sharing (CORS) misconfiguration on the web server</p>",
  "instances": [
  ],
  "count": "30",
  "solution": "<p>Ensure that sensitive data is not available in an unauthenticated manner (using IP address white-listing, for instance).</p><p>Configure the \"Access-Control-Allow-Origin\" HTTP header to a more restrictive set of domains, or remove all CORS headers entirely, to allow the web browser to enforce the Same Origin Policy (SOP) in a more restrictive manner.</p>",
  "otherinfo": "<p>The CORS misconfiguration on the web server permits cross-domain read requests from arbitrary third party domains, using unauthenticated APIs on this domain. Web browser implementations do not permit arbitrary third parties to read the response from authenticated APIs, however. This reduces the risk somewhat. This misconfiguration could be used by an attacker to access data that is available in an unauthenticated manner, but which uses some other form of security, such as IP address white-listing.</p>",
  "reference": "<p>http://www.hpenterprisesecurity.com/vulncat/en/vulncat/vb/html5_overly_permissive_cors_policy.html</p>",
  "cweid": "264",
}
```

ZAP

```
{
  "name": "Insecure 'Access-Control-Allow-Origin' header",
  "description": "\n_Cross Origin Resource Sharing (CORS)_ is an HTML5 technology which gives modern\nweb browsers the ability to bypass restrictions implemented by the _Same Origin Policy_.\nThe _Same Origin Policy_ requires that both the JavaScript and the page are loaded\nfrom the same domain in order to allow JavaScript to interact with the page. This\nin turn prevents malicious JavaScript being executed when loaded from external domains.\n\nThe CORS policy allows the application to specify exceptions to the protections\nimplemented by the browser, and allows the developer to whitelist domains for\nwhich external JavaScript is permitted to execute and interact with the page.\n\nA weak CORS policy is one which whitelists all domains using a wildcard (*),\nwhich will allow any externally loaded JavaScript resource to interact with the\naffected page. This can severely increase the risk of attacks such as Cross Site Scripting etc.\n\nArachni detected that the CORS policy being set by the server was weak, and used\na wildcard value. This is evident by the 'Access-Control-Allow-Origin' header being set to '*'.",
  "references": {
    "OWASP": "https://www.owasp.org/index.php/CORS_OriginHeaderScrutiny",
    "Mozilla Developer Network": "https://developer.mozilla.org/en-US/docs/Web/HTTP/Access_control_CORS"
  },
  "severity": "low",
  "remedy_guidance": "\nIt is important that weak CORS policies are not used. Policies can be hardened by\nremoving the wildcard and individually specifying the domains where the trusted\nJavaScript resources are located. If the list of hosts for externally hosted\nJavaScript resources is excessive, then a whole top level domain can be whitelisted\nby using a combination of the wildcard and the domain (example: '*.arachni-scanner.com').",
}
```

Arachni

Problem illustration

Finding reports:

- Contain natural language descriptive texts
- Result from tools with overlapping use cases and hence contain duplicates
- Contain multiple and different fields to analyze text from
- Have very domain-specific knowledge

*We intend to solve this problem using
Natural Language Processing*

```
{
  "pluginid": "10098",
  "alert": "Cross-Domain Misconfiguration",
  "name": "Cross-Domain Misconfiguration",
  "riskcode": "2",
  "confidence": "2",
  "riskdesc": "Medium (Medium)",
  "desc": "<p>Web browser data loading may be possible, due to a Cross Origin Resource Sharing (CORS) misconfiguration on the web server</p>",
  "instances": [
  ],
  "count": "30",
  "solution": "<p>Ensure that sensitive data is not available in an unauthenticated manner (using IP address white-listing, for instance).</p><p>Configure the \"Access-Control-Allow-Origin\" HTTP header to a more restrictive set of domains, or remove all CORS headers entirely, to allow the web browser to enforce the Same Origin Policy (SOP) in a more restrictive manner.</p>",
  "otherinfo": "<p>The CORS misconfiguration on the web server permits cross-domain read requests from arbitrary third party domains, using unauthenticated APIs on this domain. Web browser implementations do not permit arbitrary third parties to read the response from authenticated APIs, however. This reduces the risk somewhat. This misconfiguration could be used by an attacker to access data that is available in an unauthenticated manner, but which uses some other form of security, such as IP address white-listing.</p>",
  "reference": "<p>http://www.hpenterprisesecurity.com/vulncat/en/vulncat/vb/html5_overly_permissive_cors_policy.html</p>",
  "cweid": "264",
}
```

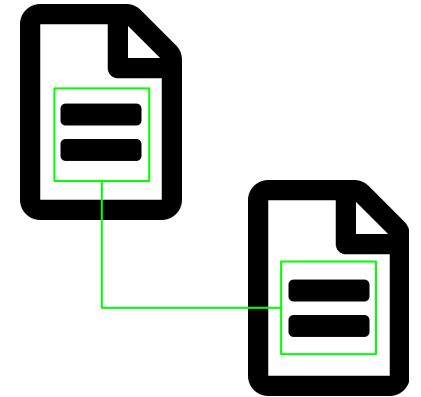
ZAP

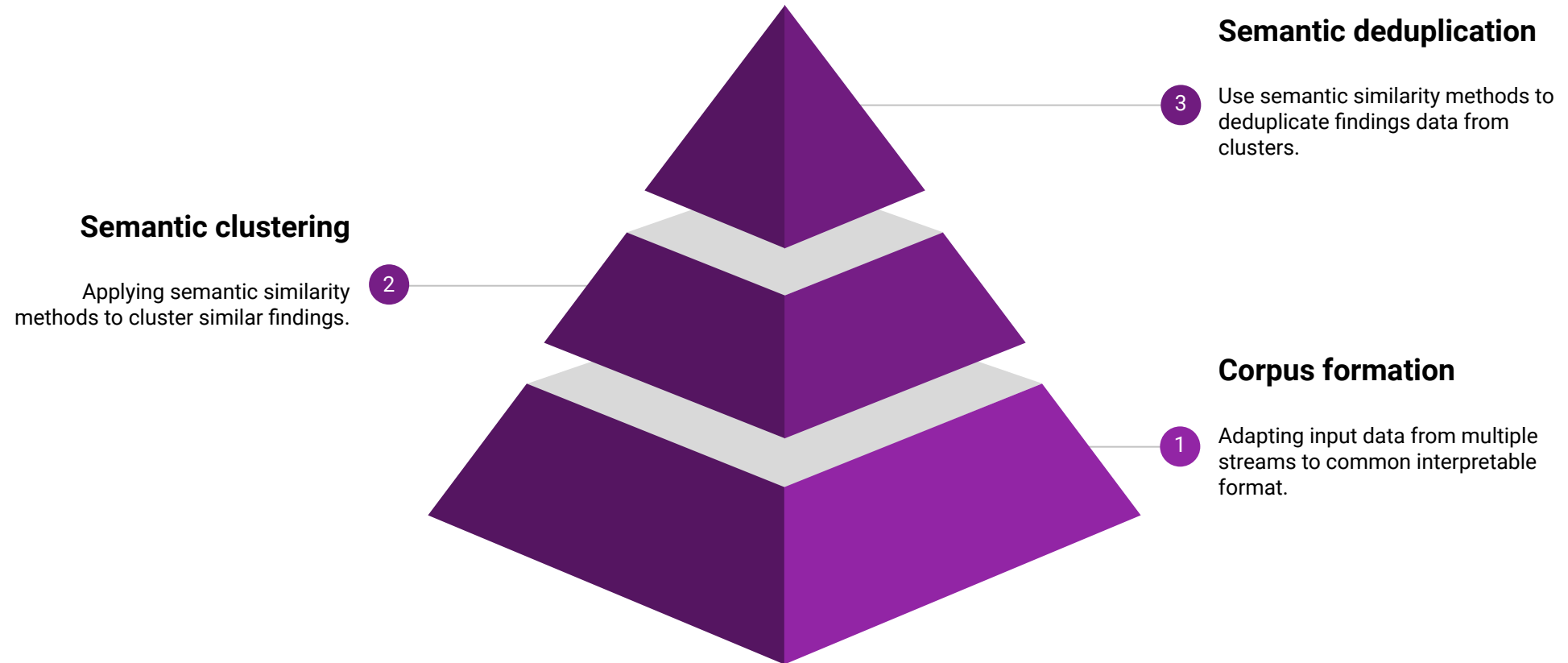
```
{
  "name": "Insecure 'Access-Control-Allow-Origin' header",
  "description": "\n_Cross Origin Resource Sharing (CORS)_ is an HTML5 technology which gives modern\nweb browsers the ability to bypass restrictions implemented by the _Same Origin Policy_.\nThe _Same Origin Policy_ requires that both the JavaScript and the page are loaded\nfrom the same domain in order to allow JavaScript to interact with the page. This\nin turn prevents malicious JavaScript being executed when loaded from external domains.\n\nThe CORS policy allows the application to specify exceptions to the protections\nimplemented by the browser, and allows the developer to whitelist domains for\nwhich external JavaScript is permitted to execute and interact with the page.\n\nA weak CORS policy is one which whitelists all domains using a wildcard (*),\nwhich will allow any externally loaded JavaScript resource to interact with the\naffected page. This can severely increase the risk of attacks such as Cross Site Scripting etc.\n\nArachni detected that the CORS policy being set by the server was weak, and used\na wildcard value. This is evident by the 'Access-Control-Allow-Origin' header being set to '*'.",
  "references": {
    "OWASP": "https://www.owasp.org/index.php/CORS_OriginHeaderScrutiny",
    "Mozilla Developer Network": "https://developer.mozilla.org/en-US/docs/Web/HTTP/Access_control_CORS"
  },
  "severity": "low",
  "remedy_guidance": "\nIt is important that weak CORS policies are not used. Policies can be hardened by\nremoving the wildcard and individually specifying the domains where the trusted\nJavaScript resources are located. If the list of hosts for externally hosted\nJavaScript resources is excessive, then a whole top level domain can be whitelisted\nby using a combination of the wildcard and the domain (example: '*.arachni-scanner.com').",
}
```

Arachni

“Semantic Similarity”

- The task of determining how similar a set of terms or documents are, in terms of what they mean.
- Scores the relationship between texts or documents using a defined metric
- Multiple approaches exist
 - Knowledge-based, corpus-based, deep neural network-based, hybrid

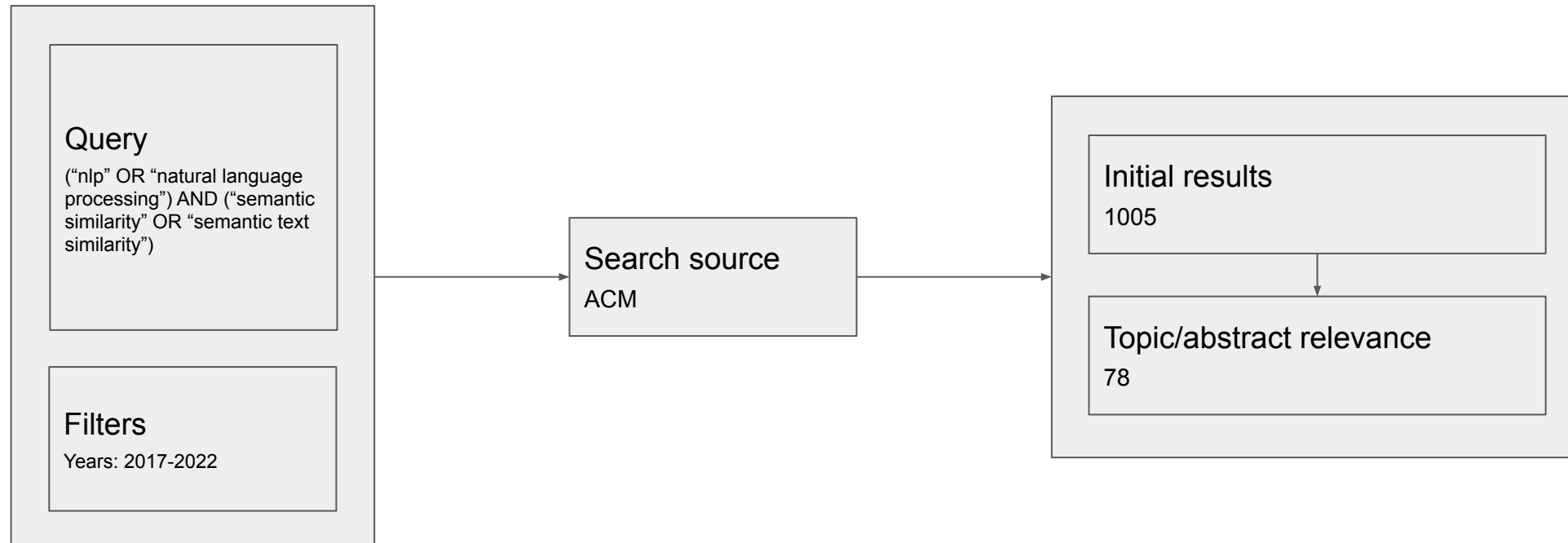




How can we use Semantic Similarity methods to deduplicate Security Findings in reports from DevOps Security Tools?

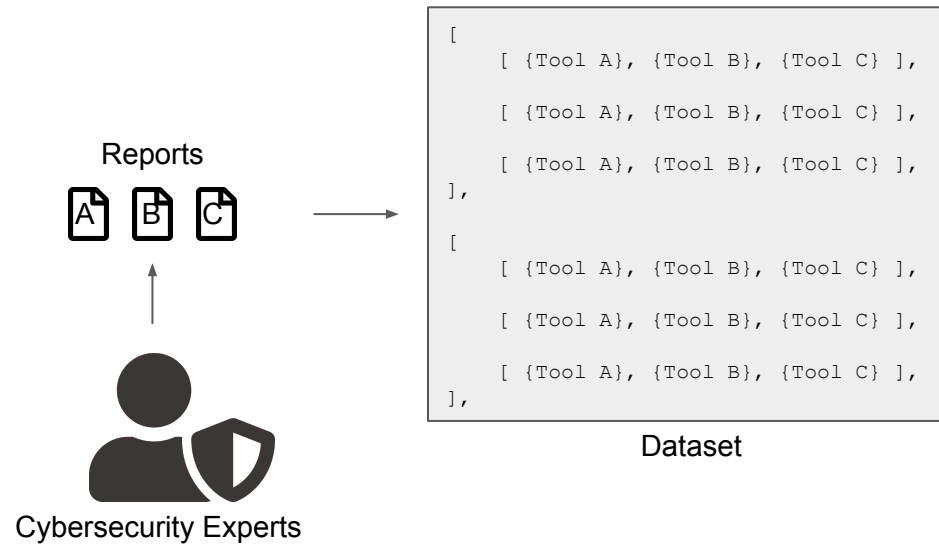
1. What semantic similarity methods that have been proposed in literature?
2. How do we construct a suitable corpus from security tool reports?
3. What methods are applicable to find semantic clusters in security tool reports?
4. How can we find semantic duplicates from clusters of security tool reports?

RQ1: What semantic similarity methods that have been proposed in literature?

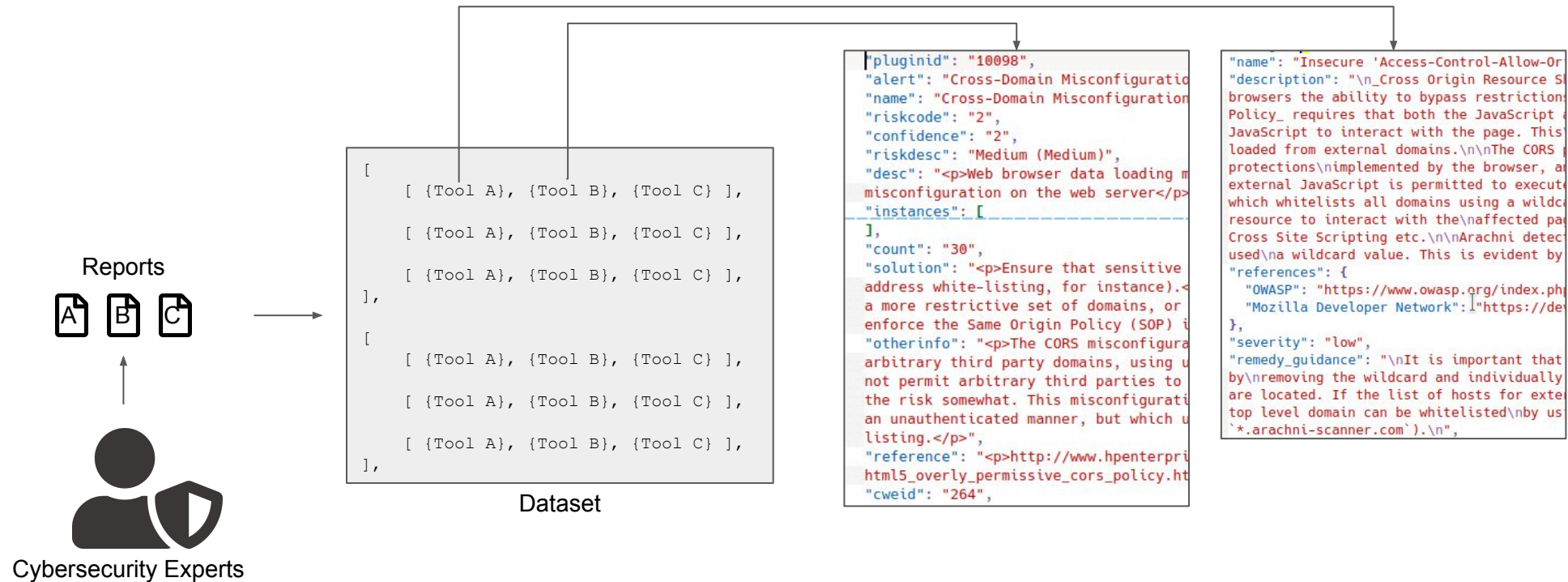


Raharjana, I. K., Siahaan, D., & Fatichah, C. (2021). User stories and natural language processing: A systematic literature review. IEEE Access, 9, 53811-53826.

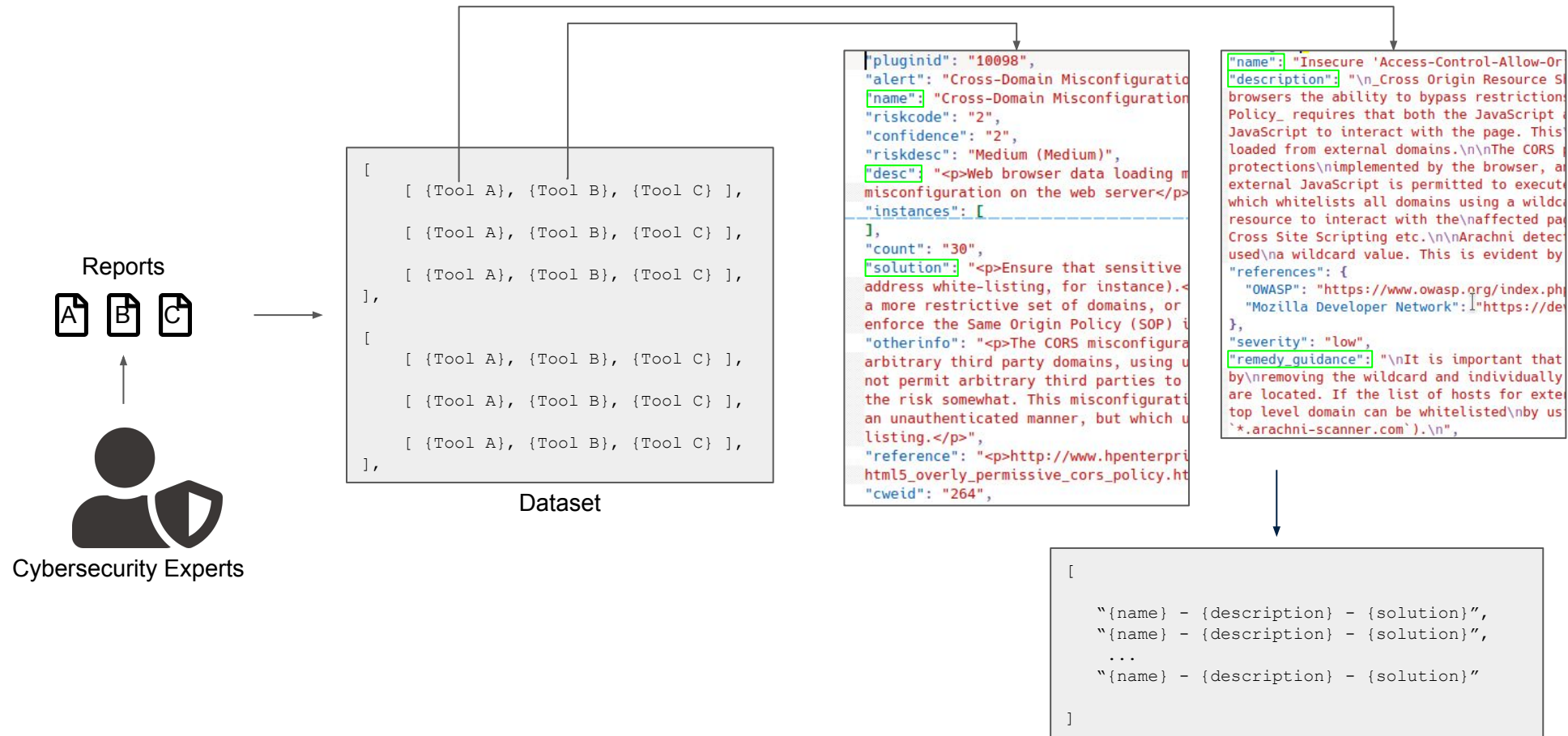
RQ2: How do we construct a suitable corpus from security tool reports?



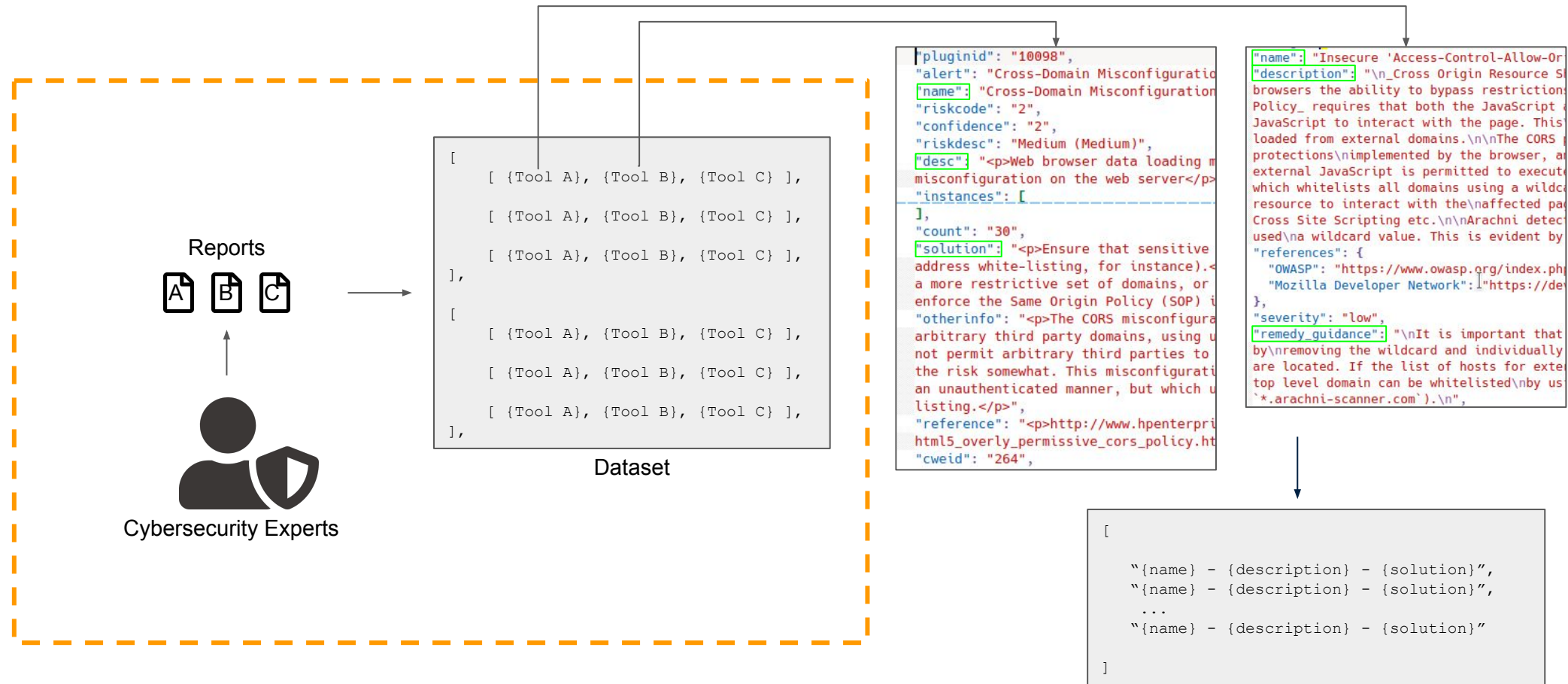
RQ2: How do we construct a suitable corpus from security tool reports?



RQ2: How do we construct a suitable corpus from security tool reports?



RQ2: How do we construct a suitable corpus from security tool reports?



RQ2: How do we construct a suitable corpus from security tool reports?

SeFiLa Label

{Se}curity {Fi}ndings {La}beler

I would like to...

Label →

Collect findings from
different security tools in
one place.

Evaluate →

Compare de-duplication
results of a technique.
(Coming soon)

RQ2: How do we construct a suitable corpus from security tool reports?

SeFiLa Label

Generate Dataset

Upload

Report file

Browse...

No file selected.

Tool

▼

Upload

Reports

#	Tool	# of Findings	
No reports yet.			

RQ2: How do we construct a suitable corpus from security tool reports?

SeFiLa Label

Generate Dataset

Upload

Report file

Browse...

arachni.json

Upload

Reports

#	Tool	# of Findings
		No report

Tool

Trivy

Bandit

ZAP

Arachni

Anchore

CodeQL

Semgrep

Horusec

Gitleaks

SonarQube

Dependency Checker

RQ2: How do we construct a suitable corpus from security tool reports?

SeFiLa

Label

Generate Dataset

Upload

Report file


Browse...

arachni.json

Tool

Upload

Reports

#	Tool	# of Findings	
1	Arachni	32	

Next step

RQ2: How do we construct a suitable corpus from security tool reports?

SeFiLa

Label

Label Dataset

Current collection

Total 0 finding(s)

Previous

Next

Change tool

Finding

No findings yet.

Exclude finding →

All findings

Total 32 finding(s), current finding ID 28

Previous

Next

Change tool

Finding

```
{
  "name": "Interesting response",
  "description": "\n\nThe server responded with a non 200 (OK) nor 404 (Not
  "references": {
    "w3.org": "http://www.w3.org/Protocols/rfc2616/rfc2616-sec10.html"
  },
  "severity": "informational",
  "check": {
    "name": "Interesting responses",
    "description": "Logs all non 200 (OK) server responses.",
    "elements": [
      "server"
    ],
    "author": "Tasos \"Zapotek\" Laskos <tasos.laskos@arachni-scanner.com>
    "version": "0.2.1",
    "max_issues": 25,
    "shortname": "interesting_responses"
  },
  "vector": {
    "class": "Arachni::Element::Server",
    "type": "server".
  }
}
```

← Include finding

All collections

ID	Name	# Findings
No collections yet.		

RQ2: How do we construct a suitable corpus from security tool reports?

SeFiLa

Label

Label Dataset

Current collection

Total 0 finding(s)

Previous

Next

Change tool

Finding

No findings yet.

Exclude finding →

All findings

Total 32 finding(s), current finding ID 28

Previous

Next

Change tool

Finding

```
{
  "name": "Interesting response",
  "description": "\n\nThe server responded with a non 200 (OK) nor 404 (Not
  "references": {
    "w3.org": "http://www.w3.org/Protocols/rfc2616/rfc2616-sec10.html"
  },
  "severity": "informational",
  "check": {
    "name": "Interesting responses",
    "description": "Logs all non 200 (OK) server responses.",
    "elements": [
      "server"
    ],
    "author": "Tasos \"Zapotek\" Laskos <tasos.laskos@arachni-scanner.com>
    "version": "0.2.1",
    "max_issues": 25,
    "shortname": "interesting_responses"
  },
  "vector": {
    "class": "Arachni::Element::Server",
    "type": "server".
  }
}
```

← Include finding

All collections

ID	Name	# Findings
No collections yet.		

RQ2: How do we construct a suitable corpus from security tool reports?

SeFiLa

Label

Label Dataset

Current collection

Total 3 finding(s), current finding ID 28

Previous

Next

Change tool

Finding

```
{
  "name": "Interesting response",
  "description": "\n\nThe server responded with a non 200 (OK) nor 404 (Not
  "references": {
    "w3.org": "http://www.w3.org/Protocols/rfc2616/rfc2616-sec10.html"
  },
  "severity": "informational",
  "check": {
    "name": "Interesting responses",
    "description": "Logs all non 200 (OK) server responses.",
    "elements": [
      "server"
    ],
    "author": "Tasos \"Zapotek\" Laskos <tasos.laskos@arachni-scanner.com>
    "version": "0.2.1",
    "max_issues": 25,
    "shortname": "interesting_responses"
  },
  "vector": {
    "class": "Arachni::Element::Server",
    "type": "server".
  }
}
```

Interesting response

Exclude finding →

Give your collection a name.

Save collection

All findings

Total 29 finding(s), current finding ID 31

Previous

Next

Change tool

Finding

```
{
  "name": "Interesting response",
  "description": "\n\nThe server responded with a non 200 (OK) nor 404 (Not
  "references": {
    "w3.org": "http://www.w3.org/Protocols/rfc2616/rfc2616-sec10.html"
  },
  "severity": "informational",
  "check": {
    "name": "Interesting responses",
    "description": "Logs all non 200 (OK) server responses.",
    "elements": [
      "server"
    ],
    "author": "Tasos \"Zapotek\" Laskos <tasos.laskos@arachni-scanner.com>
    "version": "0.2.1",
    "max_issues": 25,
    "shortname": "interesting_responses"
  },
  "vector": {
    "class": "Arachni::Element::Server",
    "type": "server".
  }
}
```

← Include finding

All collections

ID	Name	# Findings
No collections yet.		

RQ2: How do we construct a suitable corpus from security tool reports?

SeFiLa

Label

Label Dataset

Current collection

Total 3 finding(s), current finding ID 28

Previous

Next

Change tool

All findings

Total 29 finding(s), current finding ID 31

Previous

Next

Change tool

Pretty code

Finding

```
{
  "name": "Interesting response",
  "description": "\n\nThe server responded with a non 200 (OK) nor 404 (Not
  "references": {
    "w3.org": "http://www.w3.org/Protocols/rfc2616/rfc2616-sec10.html"
  },
  "severity": "informational",
  "check": {
    "name": "Interesting responses",
    "description": "Logs all non 200 (OK) server responses.",
    "elements": [
      "server"
    ],
    "author": "Tasos \"Zapotek\" Laskos <tasos.laskos@arachni-scanner.com>
    "version": "0.2.1",
    "max_issues": 25,
    "shortname": "interesting_responses"
  },
  "vector": {
    "class": "Arachni::Element::Server",
    "tvpe": "server".
  }
}
```

Interesting response

Give your collection a name.

Save collection

Exclude finding →

← Include finding

All collections

ID	Name	# Findings
No collections yet.		

Abdullah Gulraiz, Mar 14 2022, Kick-off Presentation

© sebis 31

RQ2: How do we construct a suitable corpus from security tool reports?

Serial

Label Dataset

Current collection

Total 0 finding(s)

PreviousNext

Change tool

Finding

No findings yet.

Exclude finding →

All findings

Total 29 finding(s), current finding ID 5

PreviousNext



Change tool

Finding

```
{
  "name": "Common sensitive file",
  "description": "\nWeb applications are often made up of multiple files
  "references": {
    "Apache.org": "http://httpd.apache.org/docs/2.0/mod/mod_access.html"
  },
  "severity": "low",
  "remedy_guidance": "\nIf files are unreferenced then they should be rem
  "check": {
    "name": "Common files",
    "description": "Tries to find common sensitive files on the server.",
    "elements": [
      "server"
    ],
  },
  "author": "Tasos \"Zapotek\" Laskos <tasos.laskos@arachni-scanner.com
  "version": "0.2.4",
  "exempt_platforms": [
    "rack",
    "rails",
    "cakephp",
    "svmfonv".
  ]
}
```

← Include finding

All collections

ID	Name	# Findings	
0	Interesting response	3	 

Download dataset

RQ2: How do we construct a suitable corpus from security tool reports?

Serial

Label Dataset

Current collection

Total 0 finding(s)

Previous

Next

Change tool

Finding

No findings yet.

Exclude finding →

All findings

Total 29 finding(s), current finding ID 5

Previous

Next



Change tool

Finding

```
{
  "name": "Common sensitive file",
  "description": "\nWeb applications are often made up of multiple files
  "references": {
    "Apache.org": "http://httpd.apache.org/docs/2.0/mod/mod_access.html"
  },
  "severity": "low",
  "remedy_guidance": "\nIf files are unreferenced then they should be rem
  "check": {
    "name": "Common files",
    "description": "Tries to find common sensitive files on the server.",
    "elements": [
      "server"
    ],
  },
  "author": "Tasos \"Zapotek\" Laskos <tasos.laskos@arachni-scanner.com
  "version": "0.2.4",
  "exempt_platforms": [
    "rack",
    "rails",
    "cakephp",
    "svmfonv".
  ]
}
```

← Include finding

All collections

ID	Name	# Findings	
0	Interesting response	3	 

Download dataset

RQ2: How do we construct a suitable corpus from security tool reports?

```
166 {
167   "id": 1,
168   "name": "Interesting response",
169   "findings": [
2694 ]
2695 },
2696 {
2697   "id": 2,
2698   "name": "TRACE Config",
2699   "findings": [
2700     {
2701       "id": 26,
2702       "finding": {
2806       },
2807       "tool": "arachni"
2808     }
2809   ]
2810 },
2811 {
2812   "id": 3,
2813   "name": "Wildcard in CORS",
2814   "findings": [
2815     {
2816       "id": 27,
2817       "finding": {
2934       },
2935       "tool": "zap"
2936     },
2937     {
2938       "id": 28,
2939       "finding": {
3034       },
3035       "tool": "arachni"
3036     }
3037   ]
3038 }
```

RQ3: What methods are applicable to find semantic clusters in security tool reports?

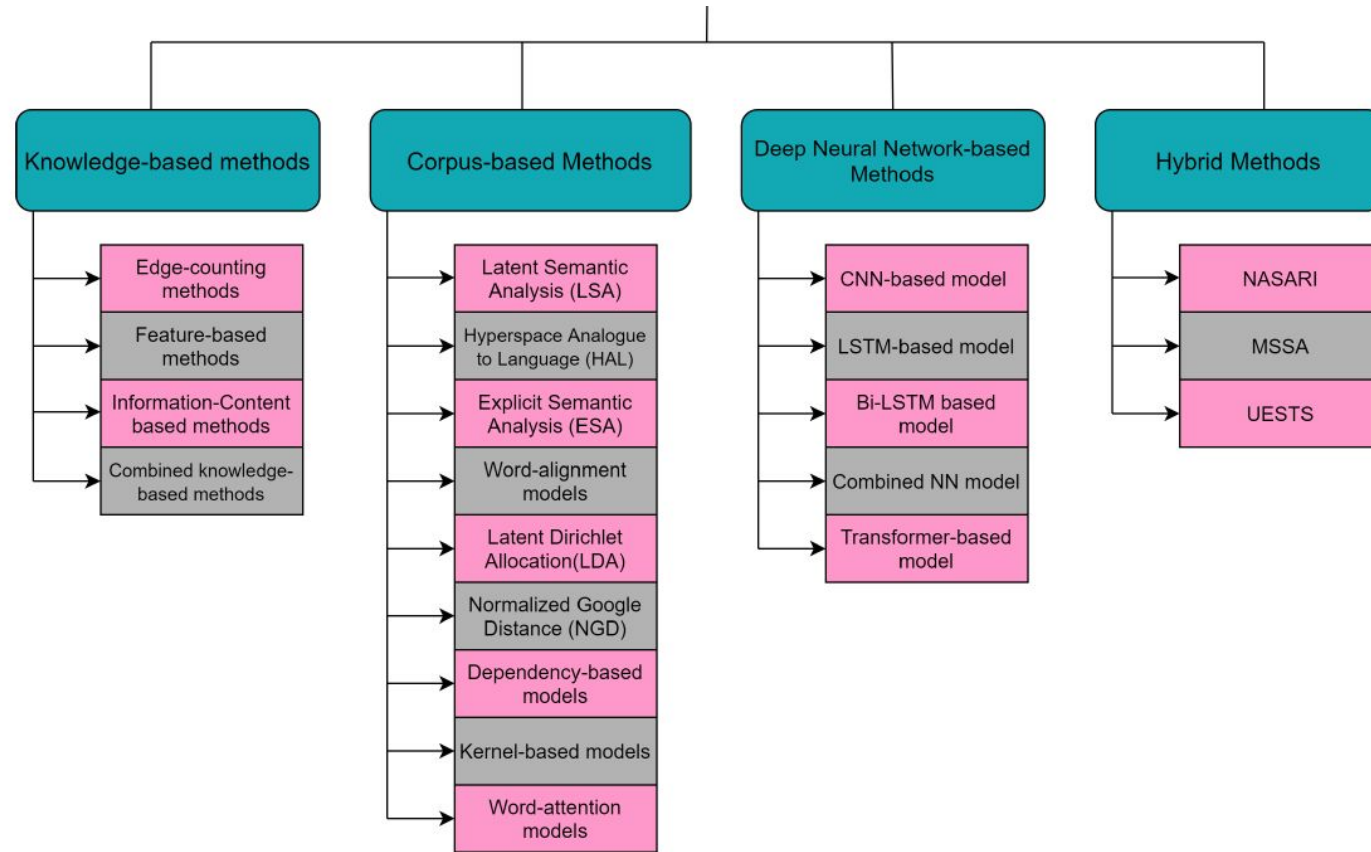


Fig 1, Chandrasekaran, D., & Mago, V. (2021). Evolution of Semantic Similarity—A Survey. *ACM Computing Surveys (CSUR)*, 54(2), 1-37.

RQ3: What methods are applicable to find semantic clusters in security tool reports?

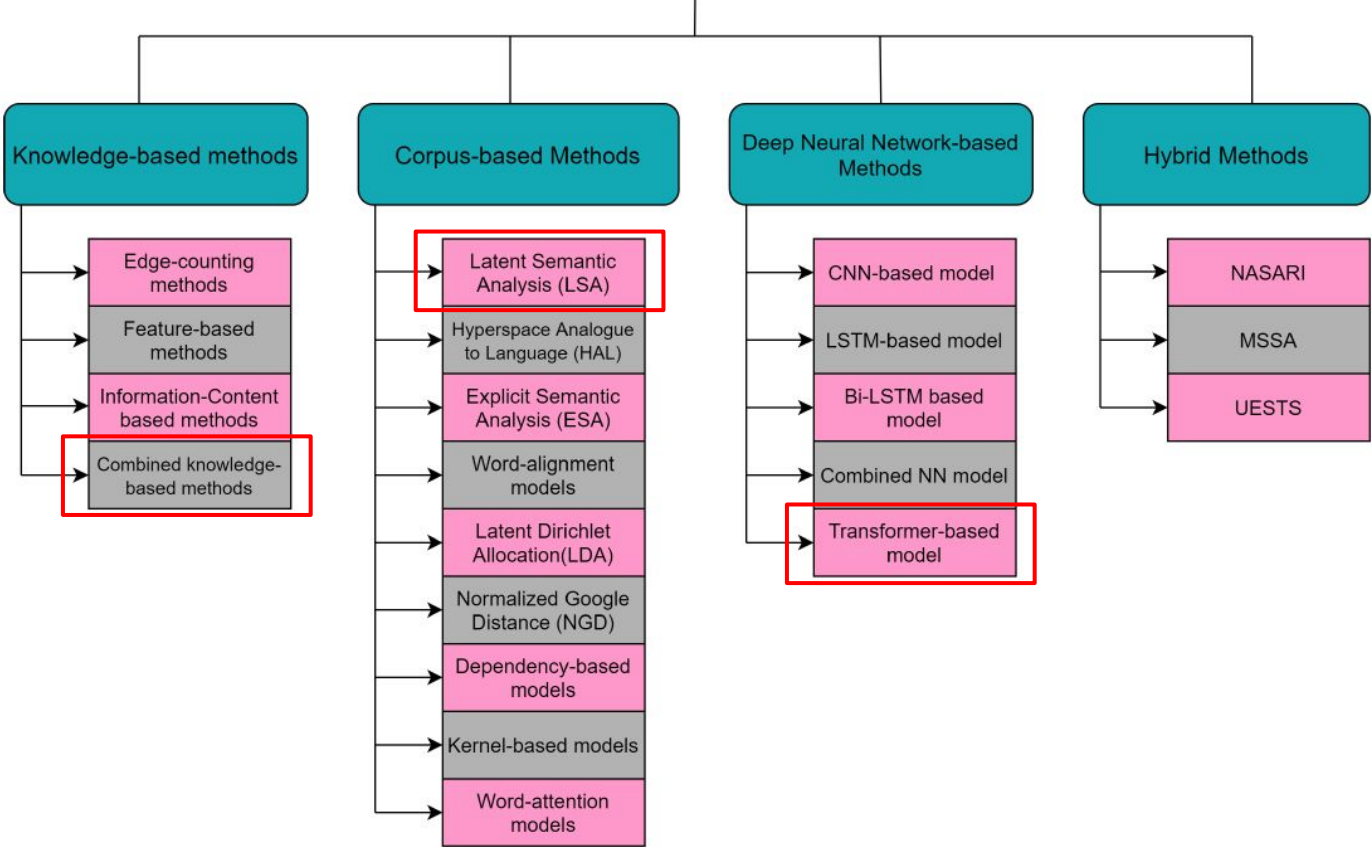
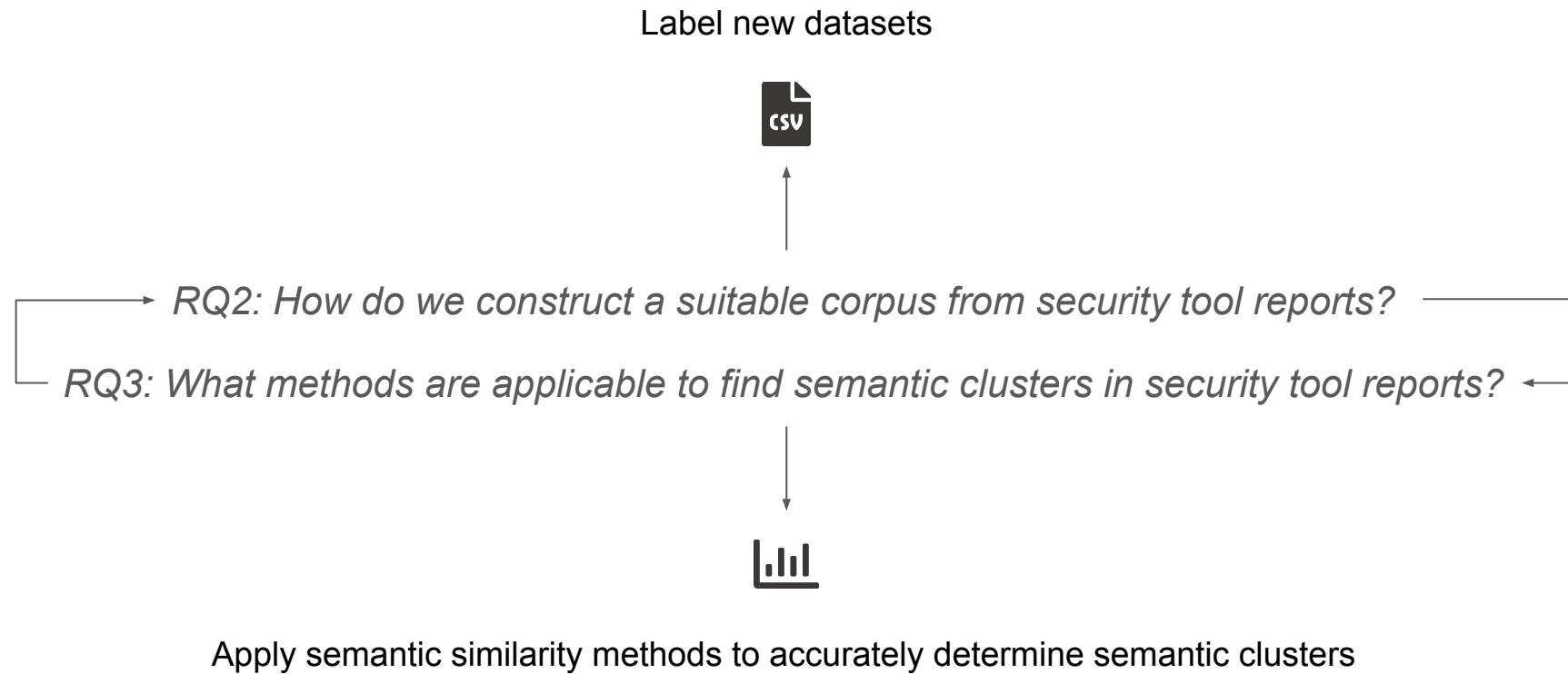


Fig 1, Chandrasekaran, D., & Mago, V. (2021). Evolution of Semantic Similarity—A Survey. ACM Computing Surveys (CSUR), 54(2), 1-37.



Jan	Feb	Mar	Apr	May	Jun
Literature analysis Collect and analyze possible literature relevant to our problem.	Preliminary results Understand and try out basic semantic similarity techniques on cybersecurity findings.	Corpus creation Choose right features over multiple datasets to fine-tune accuracy for.	Semantic clustering Achieve max accuracy for semantic clustering over different datasets.	Semantic deduplication Achieve max accuracy for semantic deduplication of findings in clusters.	Report and wrap-up Write thesis report and present results.



Thank you.

Technische Universität München
Faculty of Informatics
Chair of Software Engineering for Business
Information Systems

Boltzmannstraße 3
85748 Garching bei München

Tel +49.89.289. 17132
Fax +49.89.289.17136

matthes@in.tum.de
www.matthes.in.tum.de

