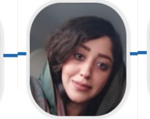
SearchSECO: A Worldwide Index of the Software Ecosystem for Cybersecurity (HiTMaT 2021)

Slinger Jansen, Utrecht University
1-11-2022

Siamak Farshidi
Senior Researcher



Elena Baninemeh
Research Assistant

Fang Hou
Research Assistant

Krishna Kaipa
Junior Researcher

Paul van Vulpen
Researcher

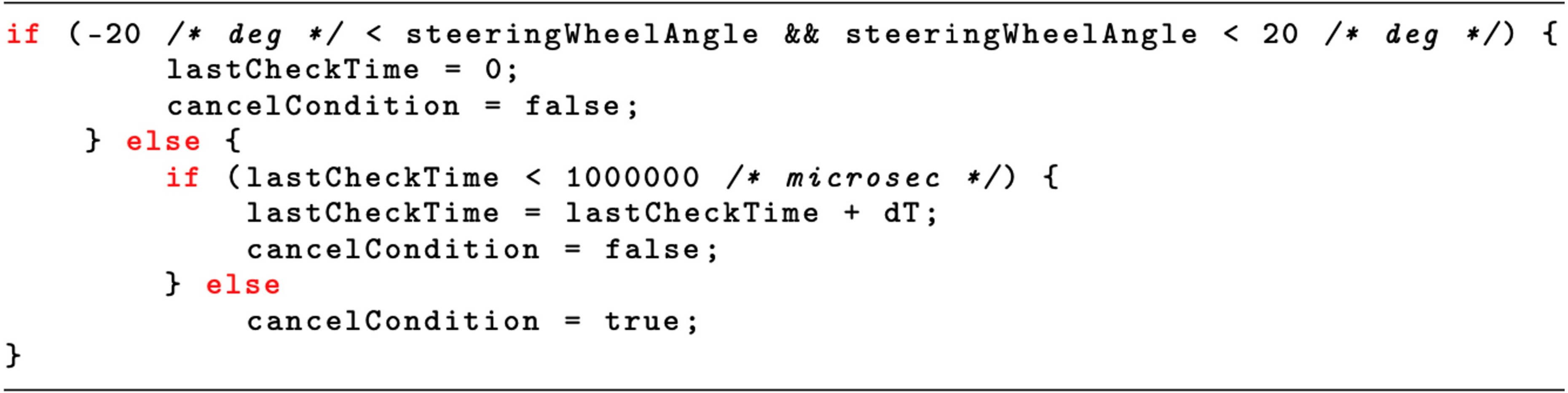
Kate Labunets
Senior Researcher

Slinger Jansen
Primary Investigator

Casper van Schothorst
Researcher

Holland High Tech
Global Challenges, Smart Solutions

Dieselgate: Inspiring Example



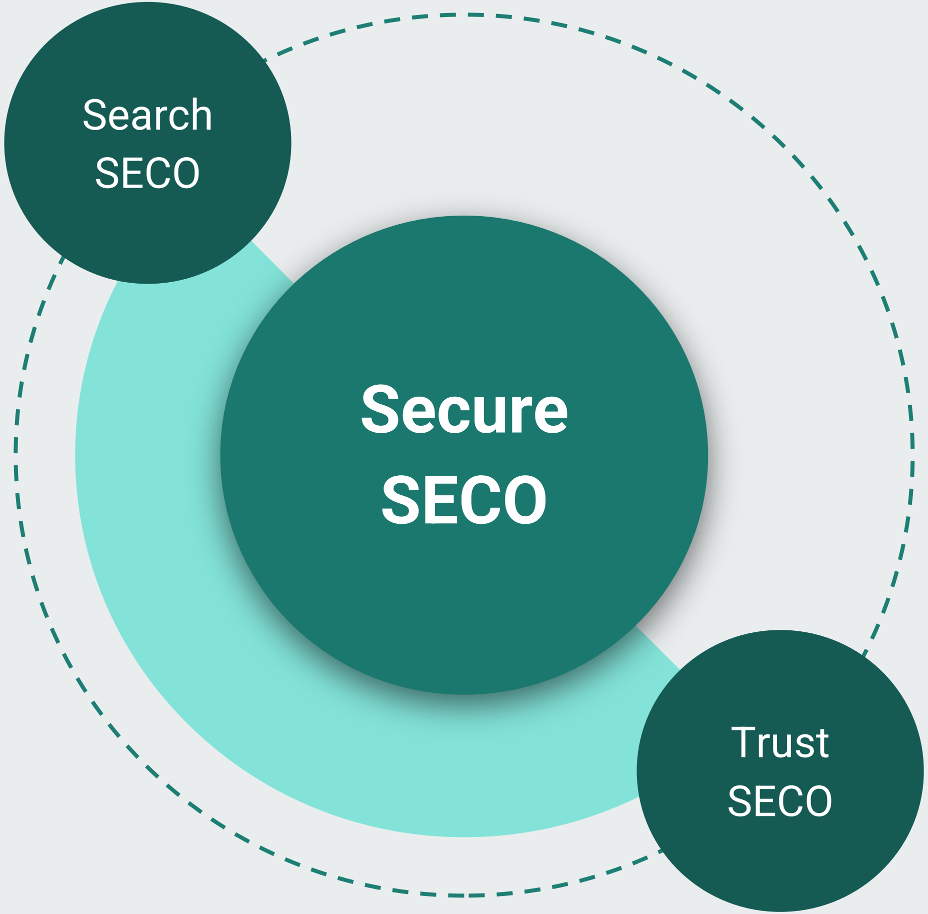
```
if (-20 /* deg */ < steeringWheelAngle && steeringWheelAngle < 20 /* deg */) {
        lastCheckTime = 0;
        cancelCondition = false;
    } else {
        if (lastCheckTime < 1000000 /* microsec */) {
            lastCheckTime = lastCheckTime + dT;
            cancelCondition = false;
        } else
            cancelCondition = true;
}
```

Goal of the SecureSECO Initiative

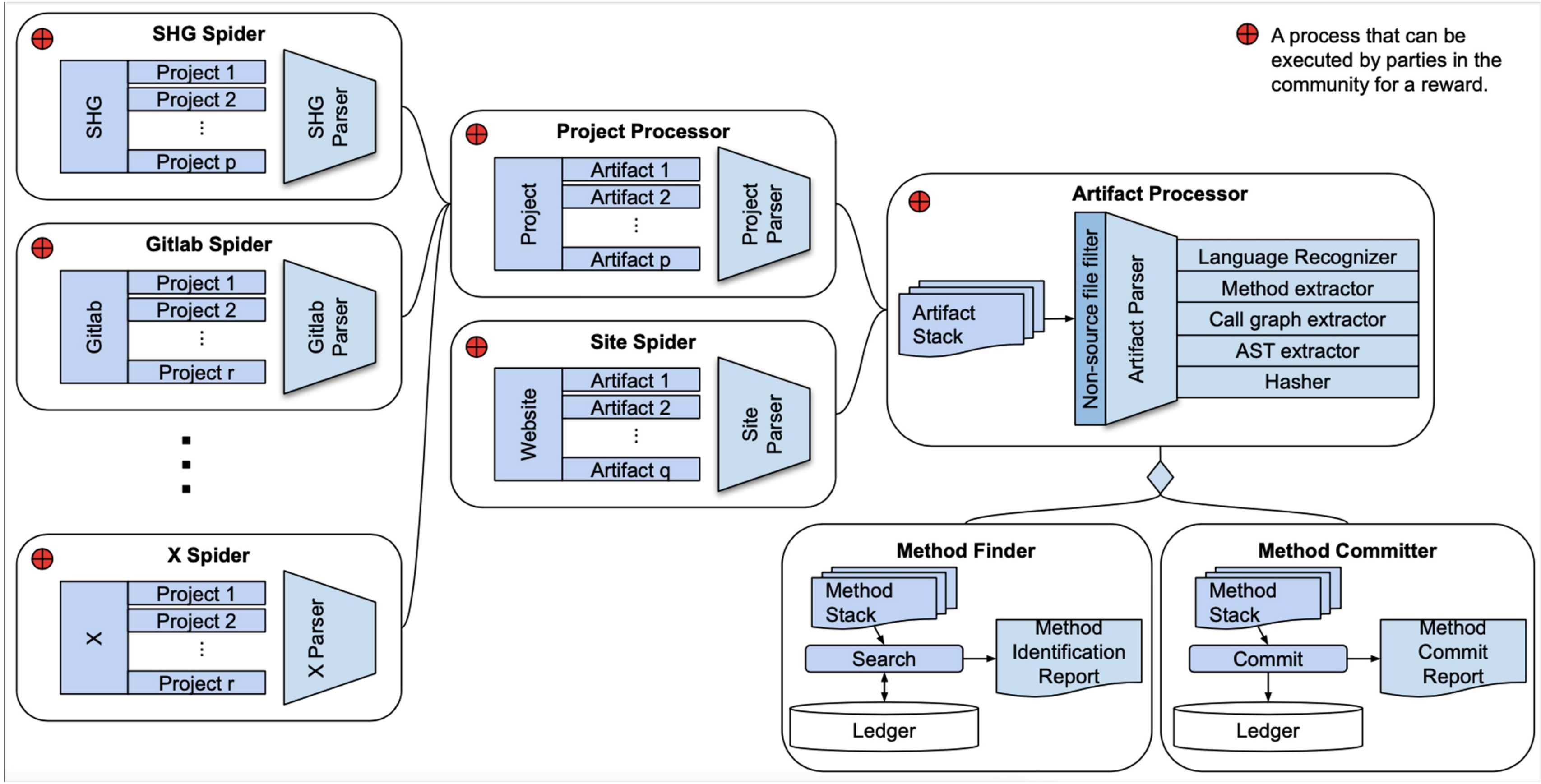
Goal: The *Secure Software Ecosystems* initiative takes **a worldwide scope** on the produced **software that society depends on**. We hunt for software vulnerabilities and propose methods for eliminating them. We collect and analyze data on **why software is to be trusted**.

We subsequently research, design, and build prototype systems that use trust data for **providing trustworthy reliable systems for consumers and industry**.

These systems are frequently underpinned by open distributed systems that provide guarantees for **reliable open untamperable data**.



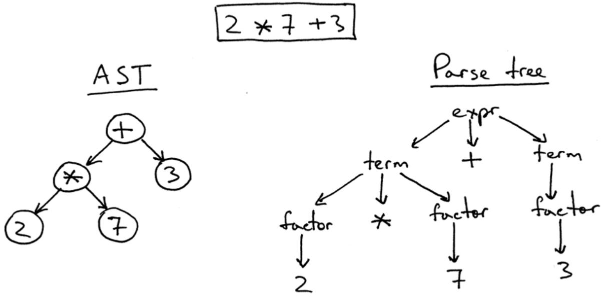
We spider the worldwide software ecosystem

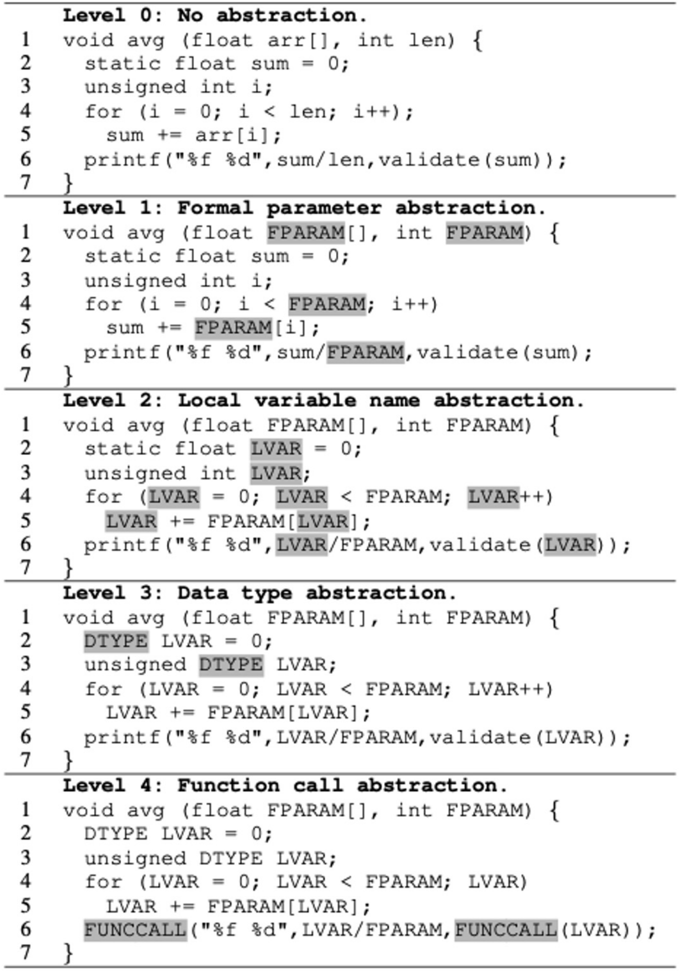


Slinger Jansen, Siamak Farshidi, Georgios Gousios, Joost Visser, Tijs van der Storm, Magiel Bruntink (2020) **SearchSECO: A Worldwide Index of the Open Source Software Ecosystem.** Proceedings of the 19th Belgium-Netherlands Software Evolution Workshop (pdf)

Technique Used: method hashing

- Methods are similar when they have the same parse tree
- We hash an abstract representation of a method, to enable fast searching through the worldwide software ecosystem
- We find clones on a worldwide scale, far larger than the typical project scope





Level 0: No abstraction.

```
void avg (float arr[], int len) {
  static float sum = 0;
  unsigned int i;
  for (i = 0; i < len; i++);
    sum += arr[i];
  printf("%f %d",sum/len,validate(sum));
}
```

Level 1: Formal parameter abstraction.

```
void avg (float FPARAM[], int FPARAM) {
  static float sum = 0;
  unsigned int i;
  for (i = 0; i < FPARAM; i++)
    sum += FPARAM[i];
  printf("%f %d",sum/FPARAM,validate(sum);
}
```

Level 2: Local variable name abstraction.

```
void avg (float FPARAM[], int FPARAM) {
  static float LVAR = 0;
  unsigned int LVAR;
  for (LVAR = 0; LVAR < FPARAM; LVAR++)
    LVAR += FPARAM[LVAR];
  printf("%f %d",LVAR/FPARAM,validate(LVAR));
}
```

Level 3: Data type abstraction.

```
void avg (float FPARAM[], int FPARAM) {
  DTYPE LVAR = 0;
  unsigned DTYPE LVAR;
  for (LVAR = 0; LVAR < FPARAM; LVAR++)
    LVAR += FPARAM[LVAR];
  printf("%f %d",LVAR/FPARAM,validate(LVAR));
}
```

Level 4: Function call abstraction.

```
void avg (float FPARAM[], int FPARAM) {
  DTYPE LVAR = 0;
  unsigned DTYPE LVAR;
  for (LVAR = 0; LVAR < FPARAM; LVAR)
    LVAR += FPARAM[LVAR];
  FUNCCALL("%f %d",LVAR/FPARAM,FUNCCALL(LVAR));
}
```

Fig. 2: Level-by-level application of abstraction schemes on a sample function.

Our Affordances

Relationships between **methods**

- Study method co-evolution across projects
- Weaknesses tracked, fixes propagated

Relationships between **authors**

- Fine grained authorship
- Copy-paste behavior (StackOverflow)

Relationships between **software projects**

- Establish package dependencies and cohesion
- License violations

Current Progress of SearchSECO

- Over 40M methods indexed (10M unique), nearing 70k projects on GitHub

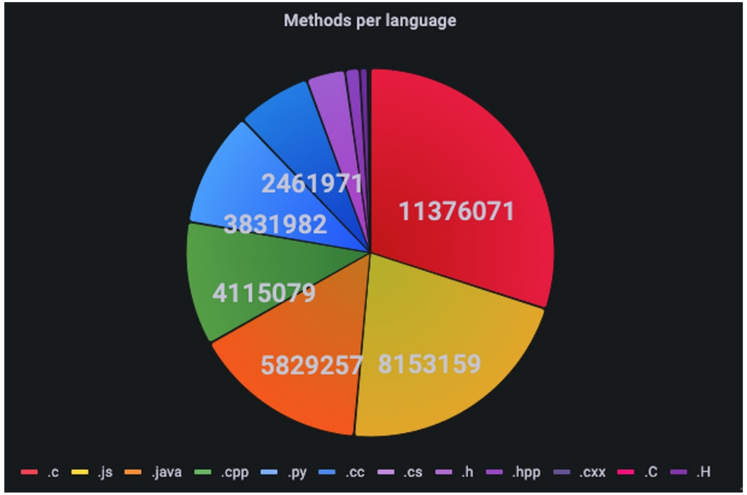
- Including author and license information

- Parsing Python, C(x), Java, and Javascript

- Currently improving performance, so we can “mine” even faster

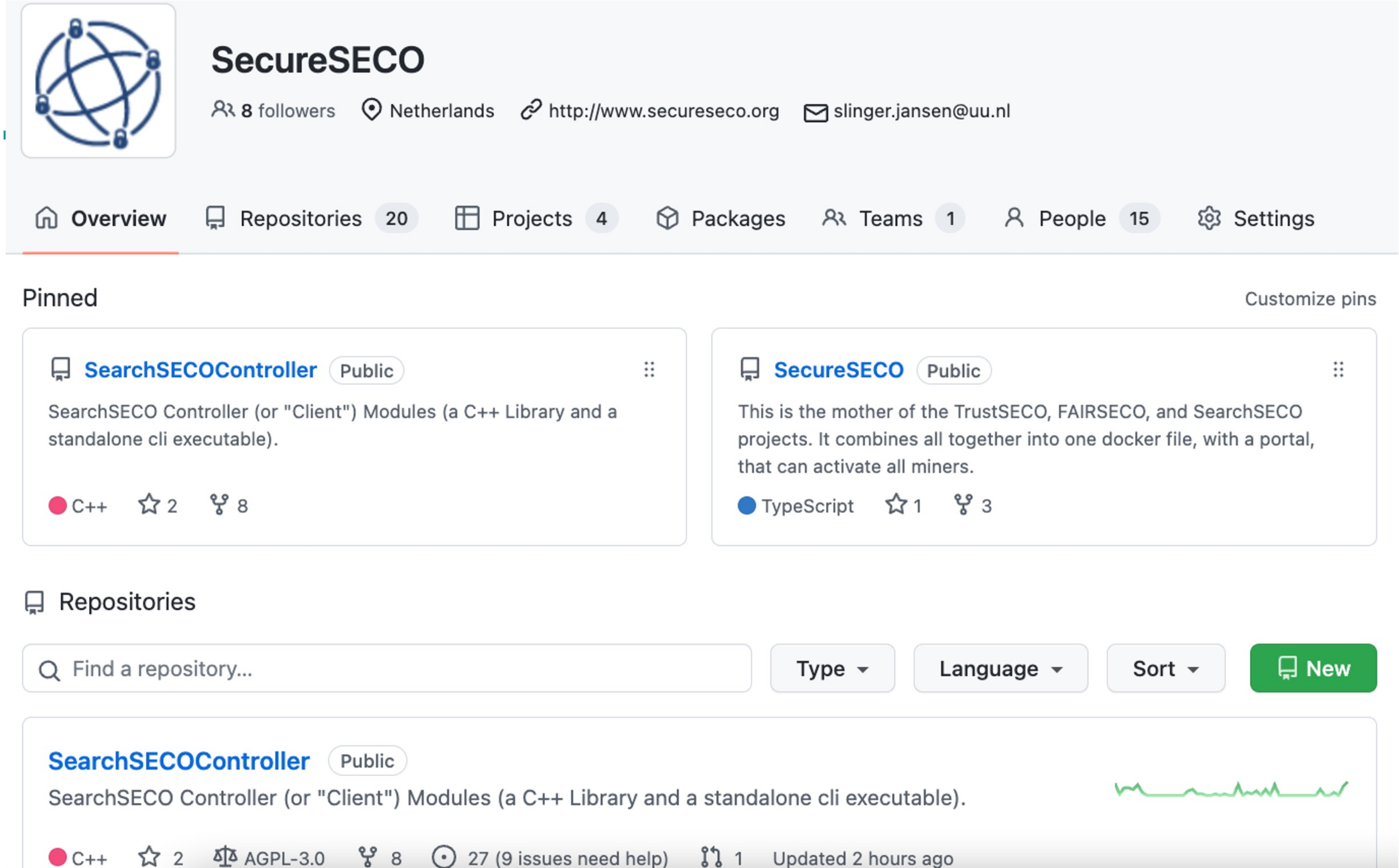
- Thorough consideration of the ethical questions around developer data collection

- Roadmap includes:
  - Advanced method-level license checking
  - “Business model” so that use of the data is paid for with more mining



Follow our progress on https://github.com/SecureSECO/

Active on Github



SecureSECO

8 followers · Netherlands · http://www.secureseco.org · slinger.jansen@uu.nl

Pinned

SearchSECOController Public

SearchSECO Controller (or "Client") Modules (a C++ Library and a standalone cli executable).

● C++ ☆ 2 Forks 8

SecureSECO Public

This is the mother of the TrustSECO, FAIRSECO, and SearchSECO projects. It combines all together into one docker file, with a portal, that can activate all miners.

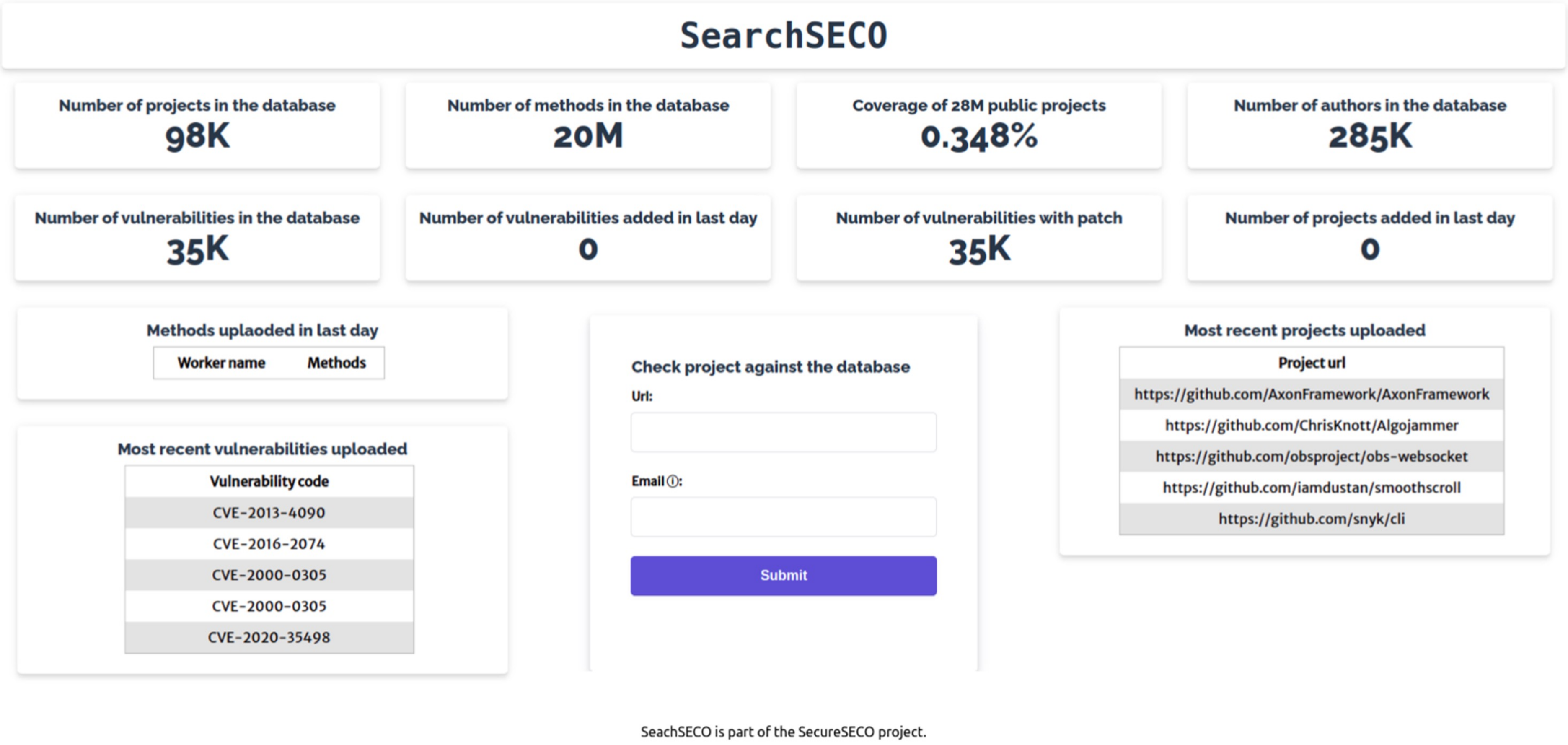
● TypeScript ☆ 1 Forks 3

Repositories

SearchSECOController Public

SearchSECO Controller (or "Client") Modules (a C++ Library and a standalone cli executable).

● C++ ☆ 2 AGPL-3.0 Forks 8 Issues 27 (9 issues need help) Pull requests 1 Updated 2 hours ago



SearchSECO

Number of projects in the database
98K

Number of methods in the database
20M

Coverage of 28M public projects
0.348%

Number of authors in the database
285K

Number of vulnerabilities in the database
35K

Number of vulnerabilities added in last day
0

Number of vulnerabilities with patch
35K

Number of projects added in last day
0

Methods uplaoded in last day

Worker name	Methods

Most recent vulnerabilities uploaded

Vulnerability code
CVE-2013-4090
CVE-2016-2074
CVE-2000-0305
CVE-2000-0305
CVE-2020-35498

Check project against the database

Url:

Email ⓘ:

Most recent projects uploaded

Project url
https://github.com/AxonFramework/AxonFramework
https://github.com/ChrisKnott/Algojammer
https://github.com/obsproject/obs-websocket
https://github.com/iamdustan/smoothscroll
https://github.com/snyk/cli

SeachSECO is part of the SecureSECO project.

https://secureseco.science.uu.nl/portal/

Our own Ecosystem and Opportunities

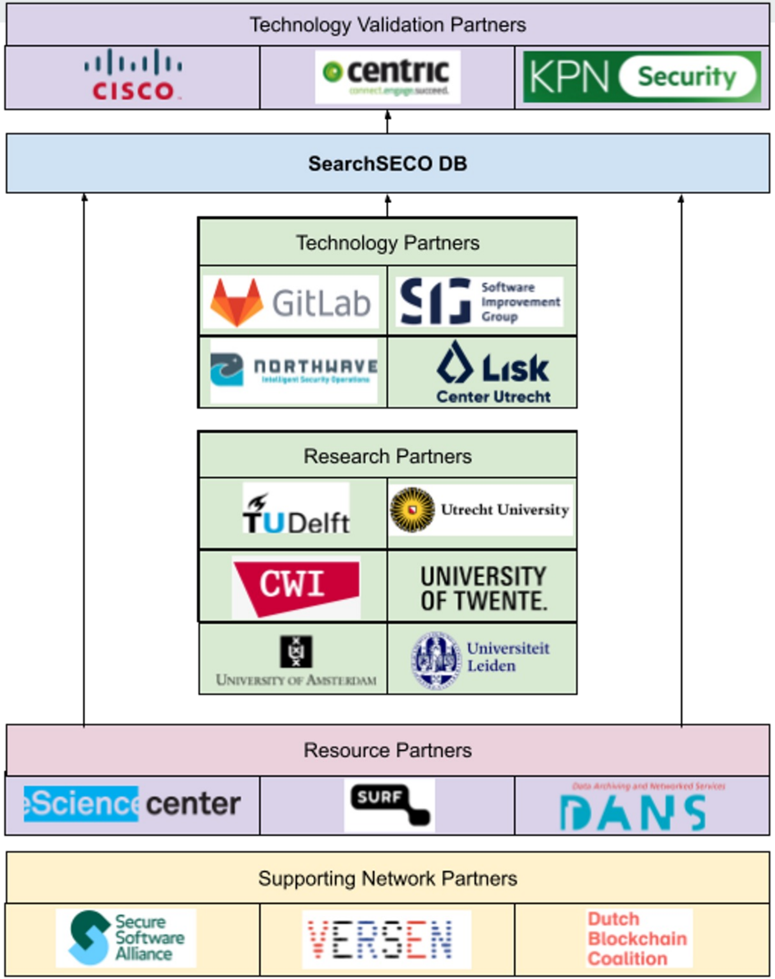
***Opportunity**: Cyber security is increasingly seen as foundational for all computer science students. In relation, there is a windfall coming for the domain of security. We do not want to miss the boat.*

The solutions we offer stem from the research domains of

- empirical software engineering,
- organizational governance,
- distributed ledger technology, and many others.

The research methods we apply are:

- **Qualitative**: Surveys, interviews, case studies, and other empirical methods.
- **Quantitative:** Data analysis, artificial intelligence for big software.



Technology Validation Partners

SearchSECO DB

Technology Partners

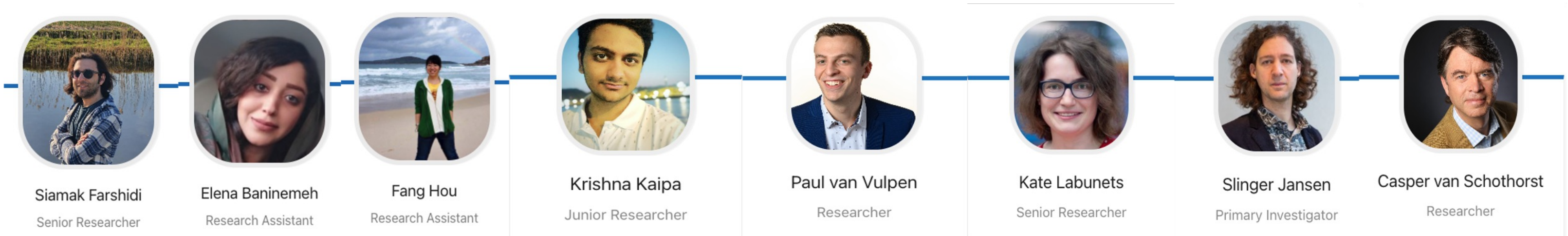
Research Partners

Resource Partners

Supporting Network Partners

In Conclusion

- SecureSECO enables a **safer healthier software ecosystem** for the organizations that take part in it.
- Please go to SecureSECO.org for more information.
- We have an amazing team of PhD, Msc, and Bsc students working on the project.
- Ambitious project goals that create societal and academic impact.
- Actively looking for collaboration and research opportunities.



Current Achievements of the Team

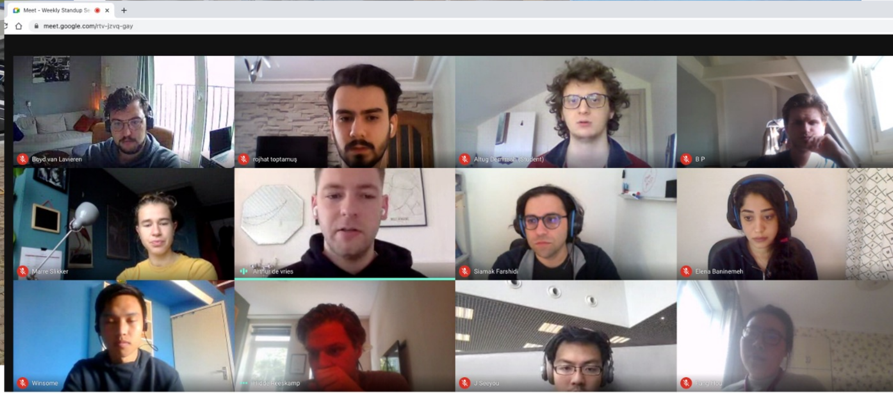
- We have acquired (1)75k euro from TruBlo for exploring trust in the worldwide software ecosystem.
- We have acquired 40k euro funding for exploring whether we can build a startup around this team from NWO
- We have acquired 60k euro funding for conducting several software projects
- We currently have 5 PhD students, 5 Msc students, and 4 Bsc students working on SecureSECO
- Ministry of Defense recently asked *“can you filter out software touched by particular identities?”*
- We have an extensive collaboration with the eScience Center, which recently funded a PhD student
- We won several awards at the Odyssey blockchain hackathon

Recent publications:

- Hou, F., & Jansen, S. (2022). **A Systematic Literature Review on Trust in the Software Ecosystem**. Empirical Software Engineering Journal (draft) (open access)
- Fang Hou, Siamak Farshidi, Slinger Jansen (2021) **TrustSECO: A Distributed Infrastructure for Providing Trust in the Software Ecosystem**. Proceedings of the Workshop on Blockchain for Information Systems Workshop. (open access)
- Slinger Jansen, Siamak Farshidi, Georgios Gousios, Joost Visser, Tijs van der Storm, Magiel Bruntink (2020) **SearchSECO: A Worldwide Index of the Open Source Software Ecosystem**. Proceedings of the 19th Belgium-Netherlands Software Evolution Workshop (pdf) (open access)
- Rowan van Pelt, Slinger Jansen, Sietse Overbeek (2020) **Defining Blockchain Governance: A Framework for Analysis and Comparison**. Information Systems Management. 2020 (pdf) (open access)
- Siamak Farshidi, Slinger Jansen, Sergio Espana, and Jacco Verkleij. **Decision support for blockchain platform selection: Three industry case studies**. IEEE Transactions on Engineering Management, 2020 (pdf). (open access)

And... We have fun!

SECURE SECO
I secured the worldwide software ecosystem and all I got was this lousy t-shirt



Dank voor uw aandacht

✉ slinger.jansen@uu.nl

🔗 www.hollandhightech.nl
www.secureseco.org

NL

Holland High Tech
Global Challenges, Smart Solutions