Empirical Analysis of Pull Requests for Google Summer of Code

Saheed Popoola

next lives here

University of CINCINNATI

Background

- Internships are key to gain real-world experience and meet expectations of workforce
- Internships need industry partners
- Open source as a viable alternative

Google Summer of Code

next lives here

University of CINCINNATI

Google Summer of Code

19,000 contributors, 18,000 mentors
more than 100 Countries

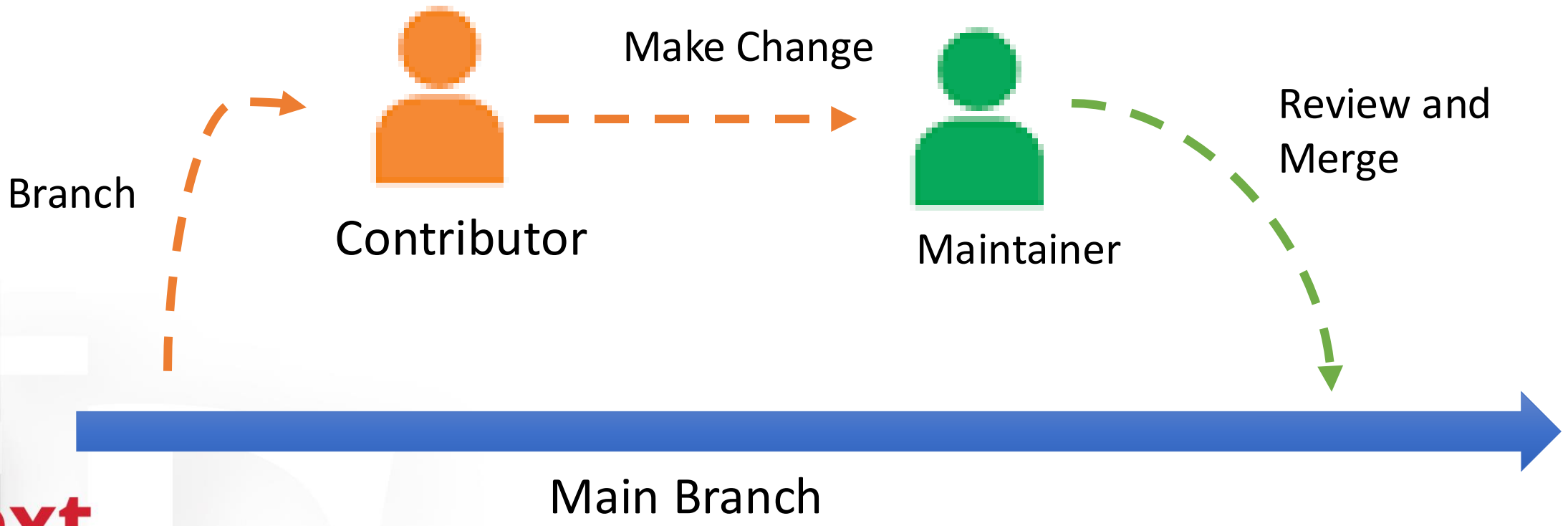
800 open-source organizations

43 million lines of code

next
lives
here

University of
CINCINNATI

Pull Request



next
lives
here

University of
CINCINNATI

Analyze pull requests to understand intern contributions and inform educatior

Contributions

- Publicly available dataset on GSoC pull requests
  - 17,232 PR (200 to 2023)
  - Python scripts for data collection and analysis
- Tasks
- Feedback

next
lives
here

University of
CINCINNATI

Research Questions

- RQ1: GSoC success in encouraging contributions?
- RQ2: What tasks do interns work on?
- RQ3: What feedback do interns receive?

Data Collection

Scrape reflection pages on GSoC Website → Extract Github pages → Github API to extract PR

next lives here

University of CINCINNATI

Methodology

- Topic Modeling
  - Latent Dirichlet Allocation (LDA) - Gensim
  - Manual coding

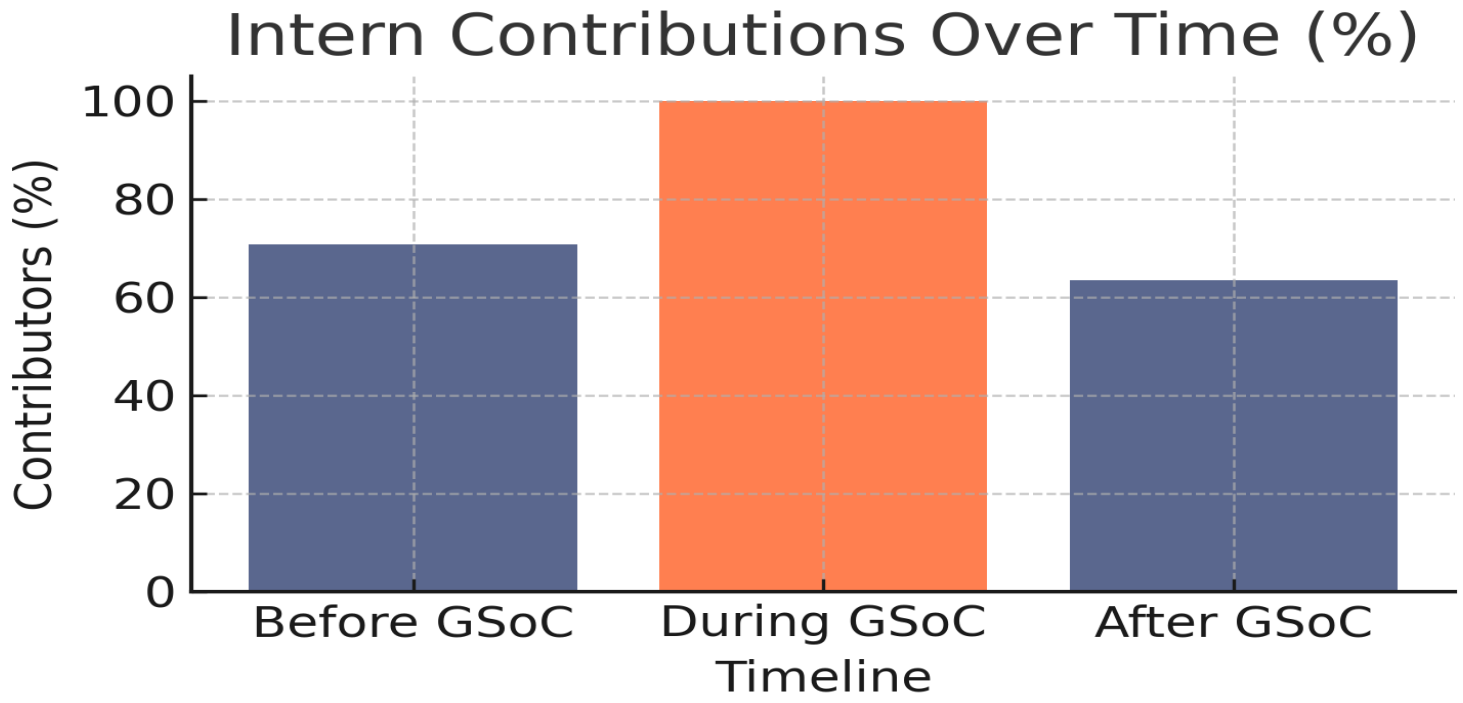
next lives here

University of CINCINNATI

Results

- RQ1: GSoC success in encouraging contributions?
- RQ2: What tasks do interns work on?
- RQ3: What feedback do interns receive?

GSoC Success (RQ1)



next lives here

- 85.7% of PRs merged
- 63.5% continued contributing after GSoC

→ GSoC effectively fosters sustained OSS engagement

University of CINCINNATI

RQ2: Tasks & Contributions

- New feature & API development
- Corrective maintenance (bug fixes)
- Project setup & dependencies
- Documentation updates
- Code refactoring

RQ3: Feedback & Review Comments

- Code structure & organization
- Best practices & style
- Functionality & logic
- Testing coverage
- Error handling
- Readability & clarity

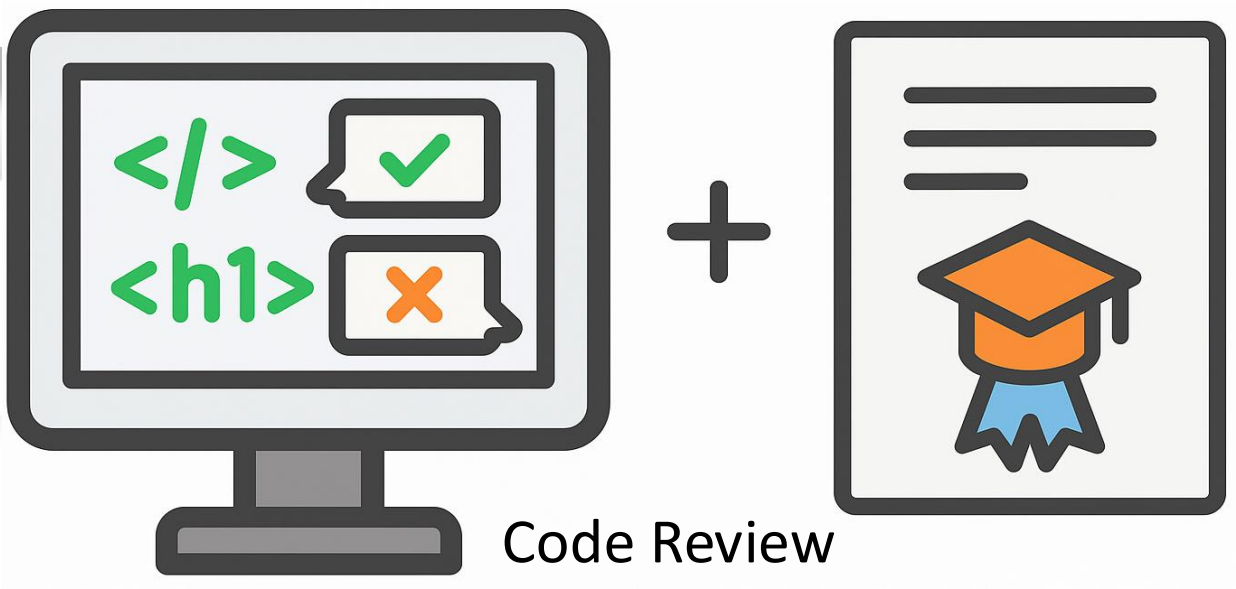
next lives here

University of CINCINNATI

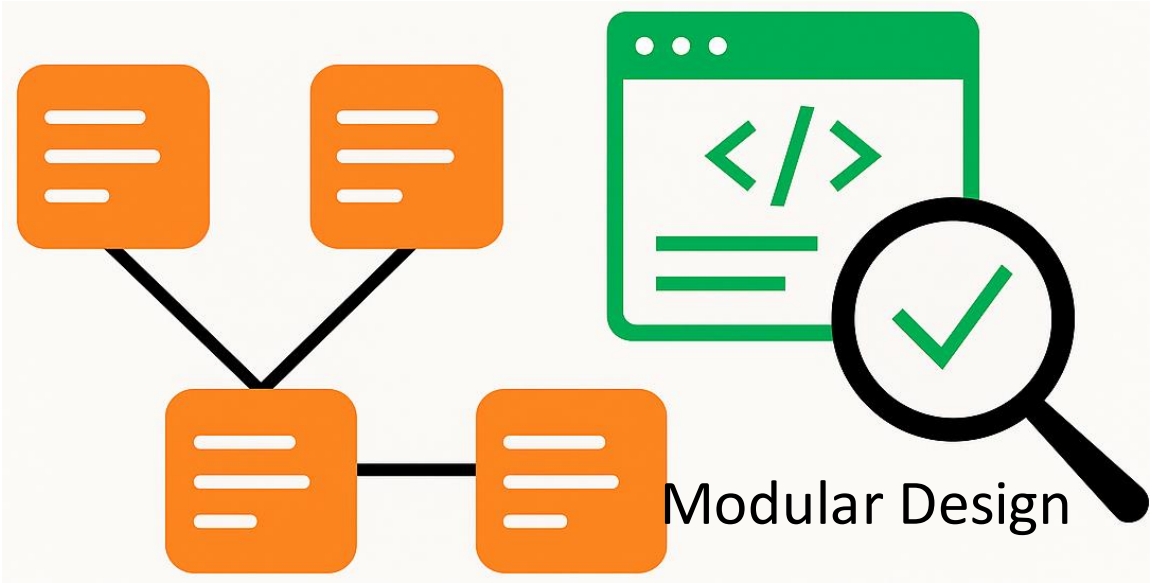
Implications for IT Education

next
lives
here

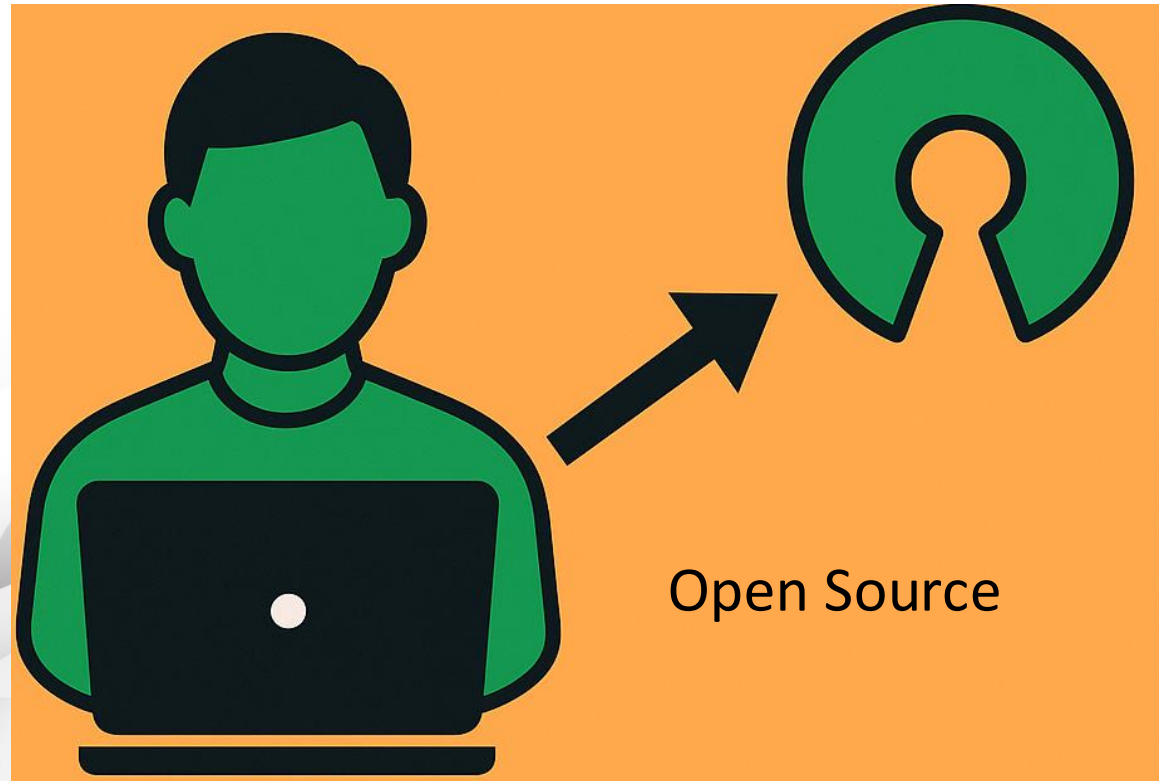
University of
CINCINNATI



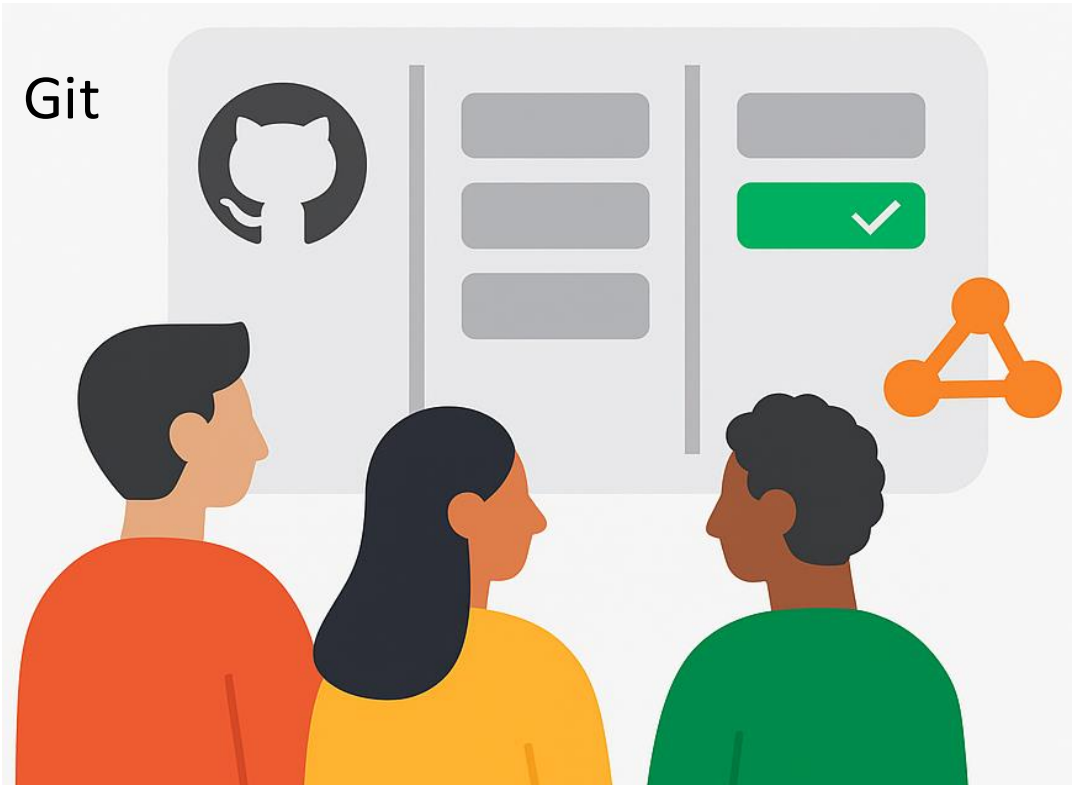
Code Review



Modular Design



Open Source



Git

Conclusion

- 17K PRs analyzed (2020–2023)
- GSoC promotes long-term engagement
- Code and non-code contributions
- Feedback on code organization, additional tests, performance optimizations, and better error handling.
- curriculum emphasis on code review, modular design, testing, and open-source

next lives here

University of CINCINNATI