

OSSコミュニティへの参加支援を目的とした新人開発者のGitHubとStackOverFlowでの活動分析

学籍番号 60226276

氏名 森岡 勇介

背景 | OSSについて

Open Source Software (OSS)

- オペレーティングシステム
- ソフトウェア開発環境
- Webアプリケーション実行環境

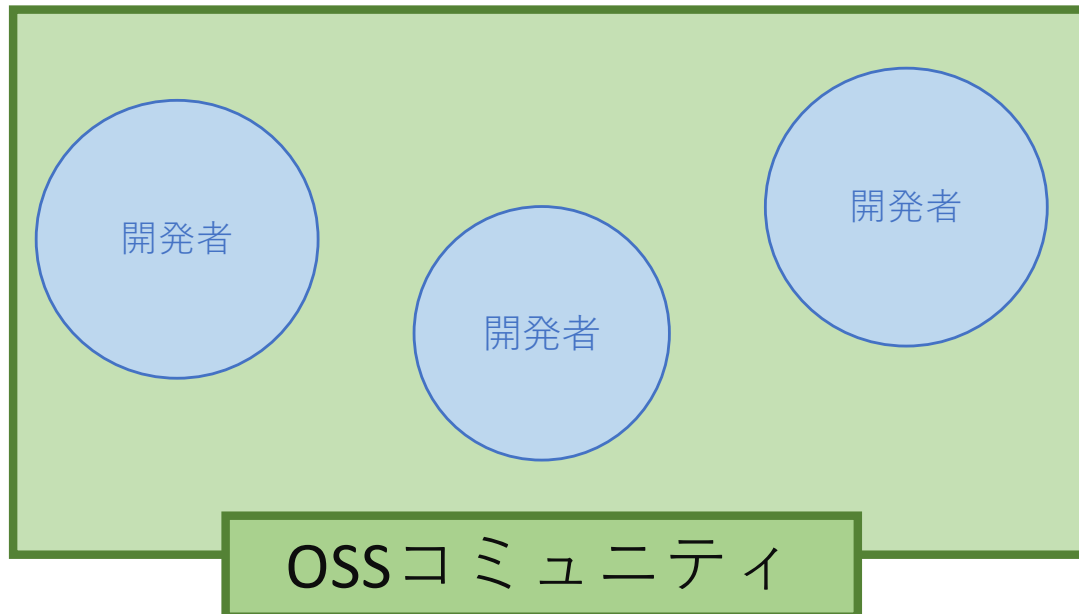
ソフトウェア開発にOSSは必要不可欠

背景 | OSSプロジェクトの持続可能性

Open Source Software (OSS)



=> 貢献

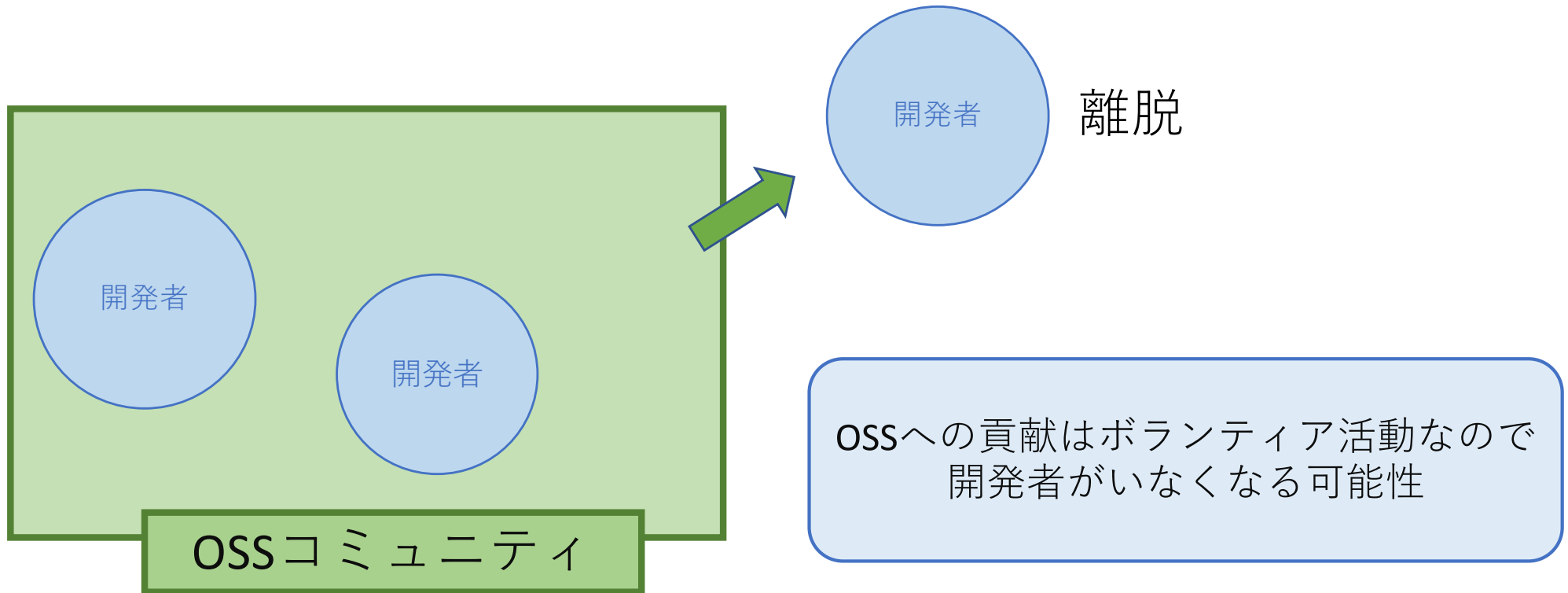


貢献

- ・バグの報告
- ・バグの修正
- ・機能の拡張

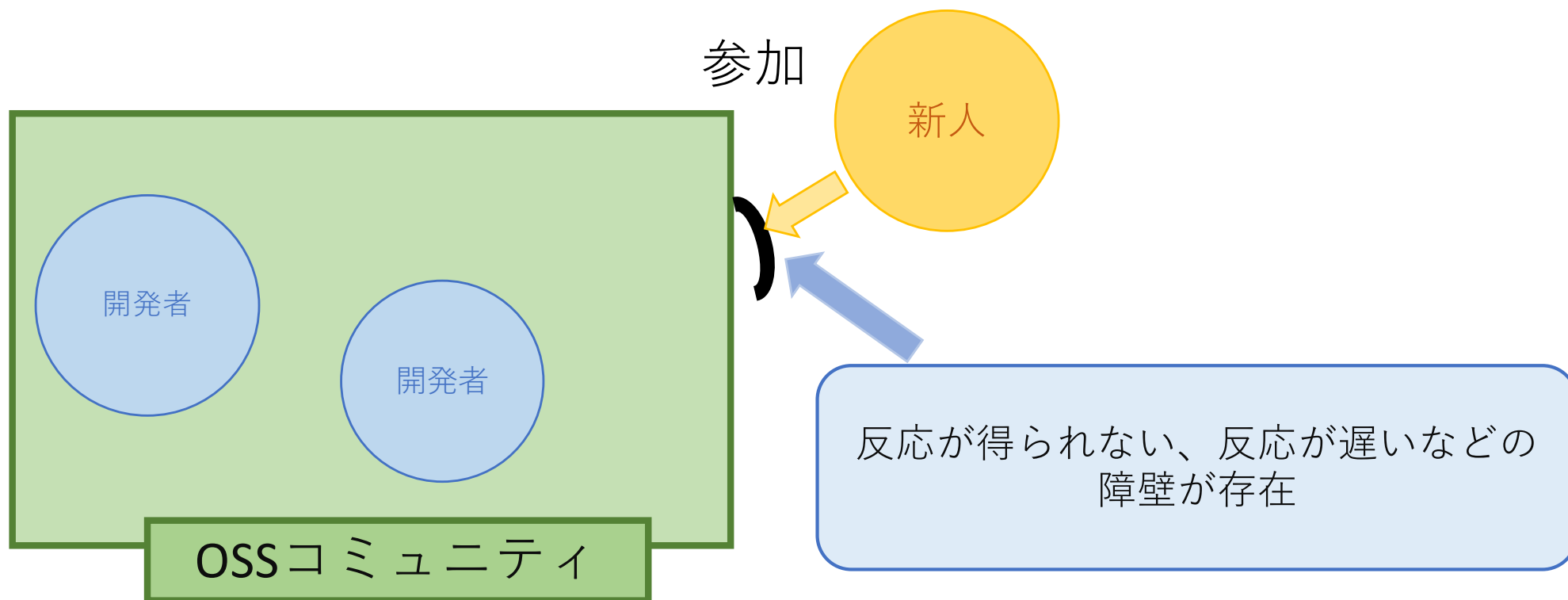
プロジェクトの存続には
貢献が重要

背景 | OSSプロジェクトの持続可能性



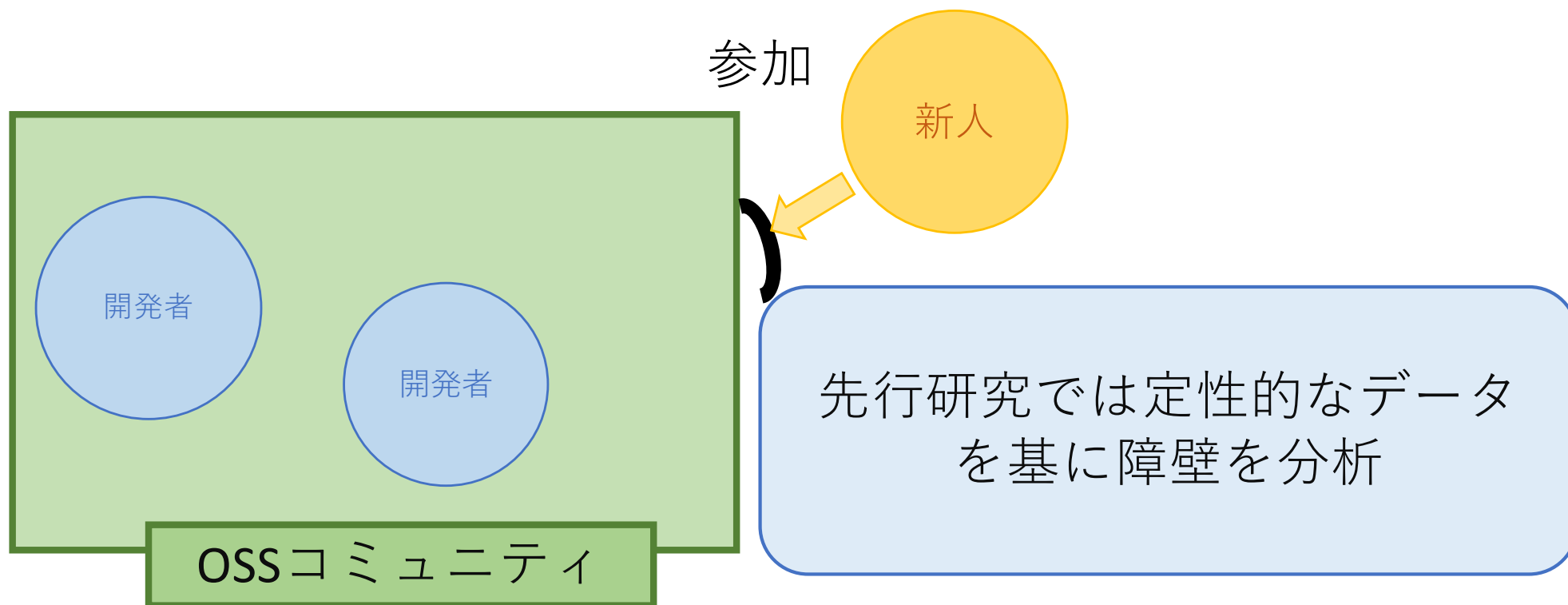
OSSプロジェクトの存続にはコミュニティに
新しい開発者が必要
(以降新人と呼称)

背景 | OSSプロジェクトの参加障壁



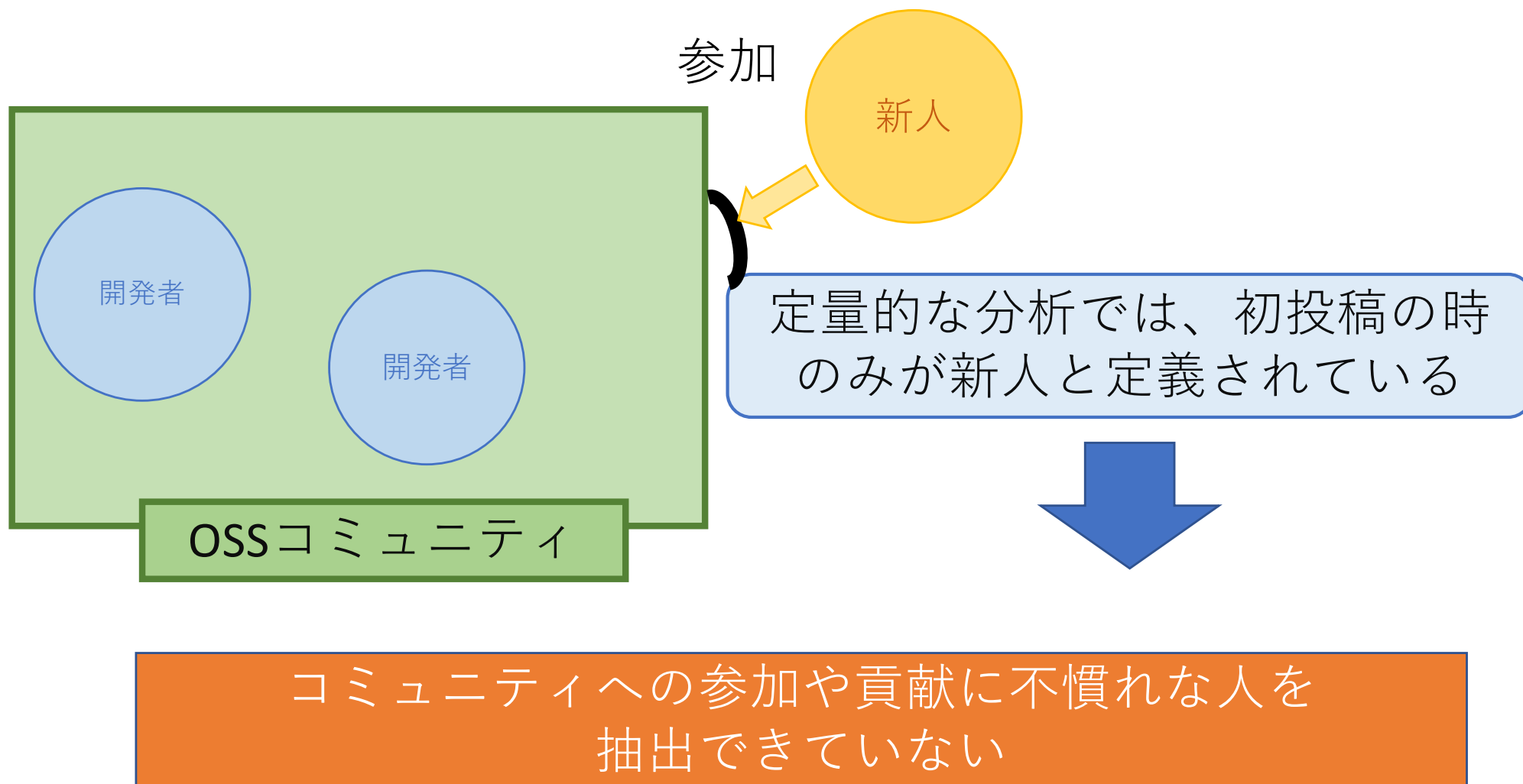
特にOSS開発経験のない初心者は障壁に直面しやすい

背景 | 先行研究

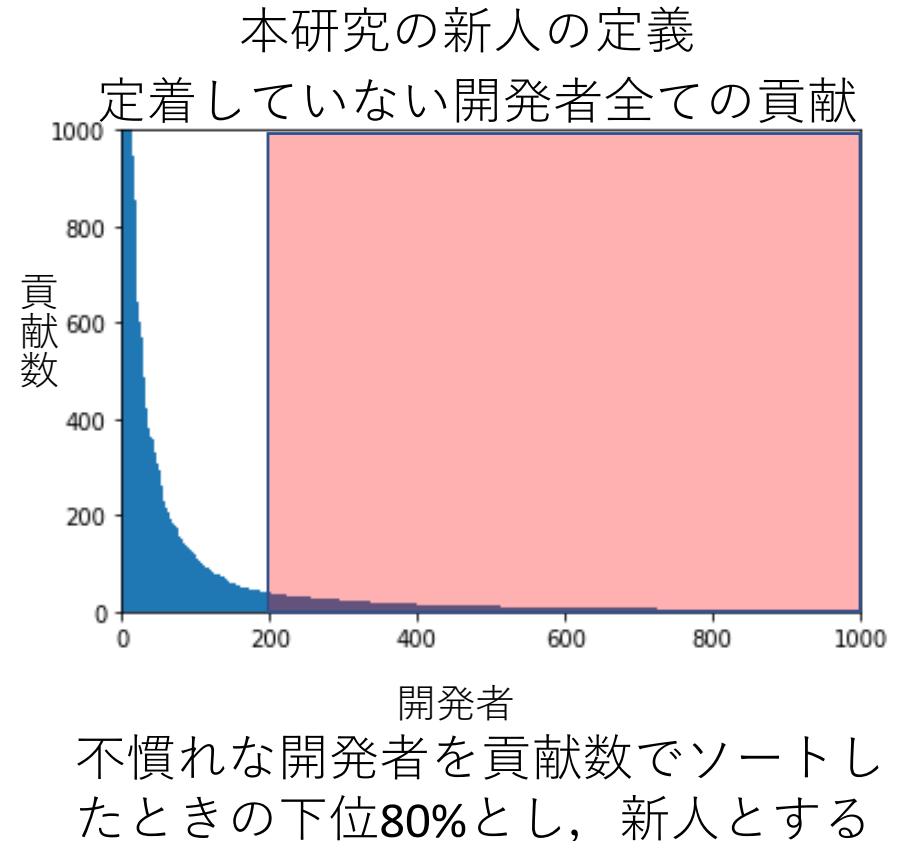
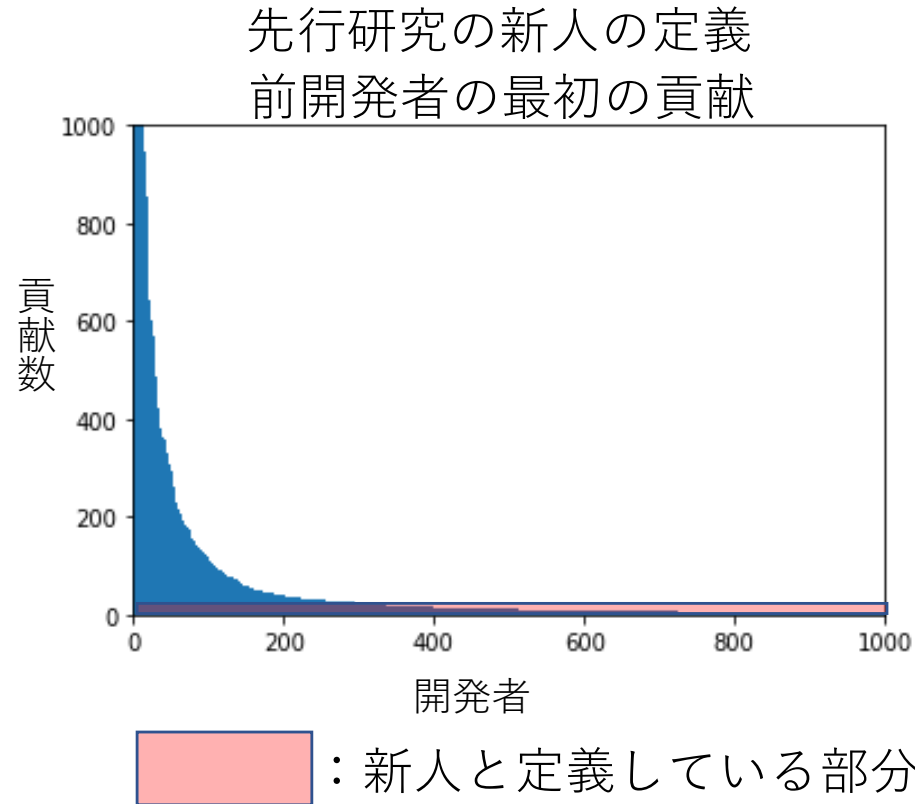


定量的なデータに基づいた参加障壁の分析は少ない

背景 | 先行研究の問題点

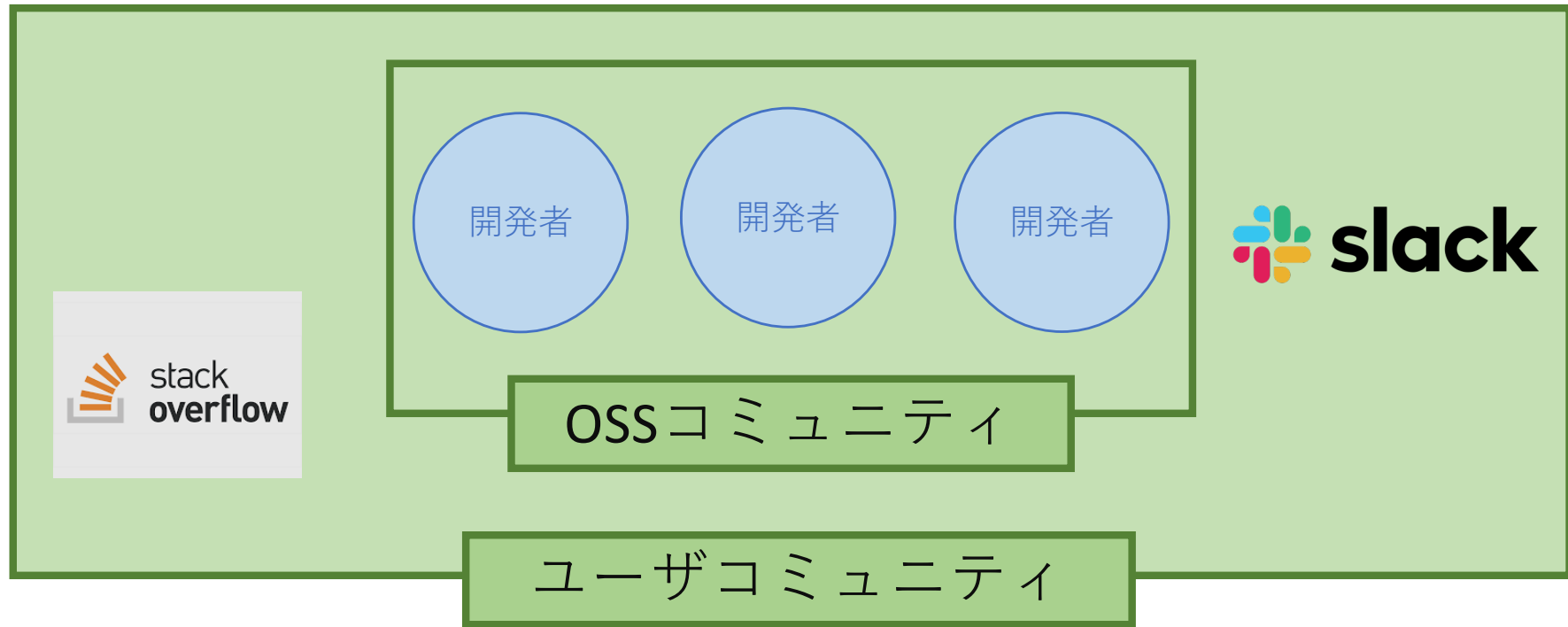


本研究の着眼点



新人の定義を変更して分析を行う

本研究の着眼点



ユーザコミュニティに新人が参加しやすいかどうかは
開発コミュニティを維持するためにも重要

ユーザコミュニティへの参加障壁については
分析されていない

リサーチクエスト(1)

開発コミュニティと外側のコミュニティでの
新人の扱われ方の違いを明らかにする

RQ1

Stack OverflowとGitHubで新人と熟練者の活動の
受け入れられやすさに違いはあるか

GitHubの活動：IssueやPullRequestの投稿

StackOverflowの活動：質問や回答

リサーチクエスチョン(2)

外側のコミュニティで活動を行うことで
新人の貢献が受け入れられやすくなるかどうか
を明らかにする

RQ2

事前にStackOverflowで活動をした新人とそうでない
新人では、貢献の受け入れやすさに違いはあるか

貢献の受け入れやすさ：Issueへのコメント数や
PullRequestのマージされる割合

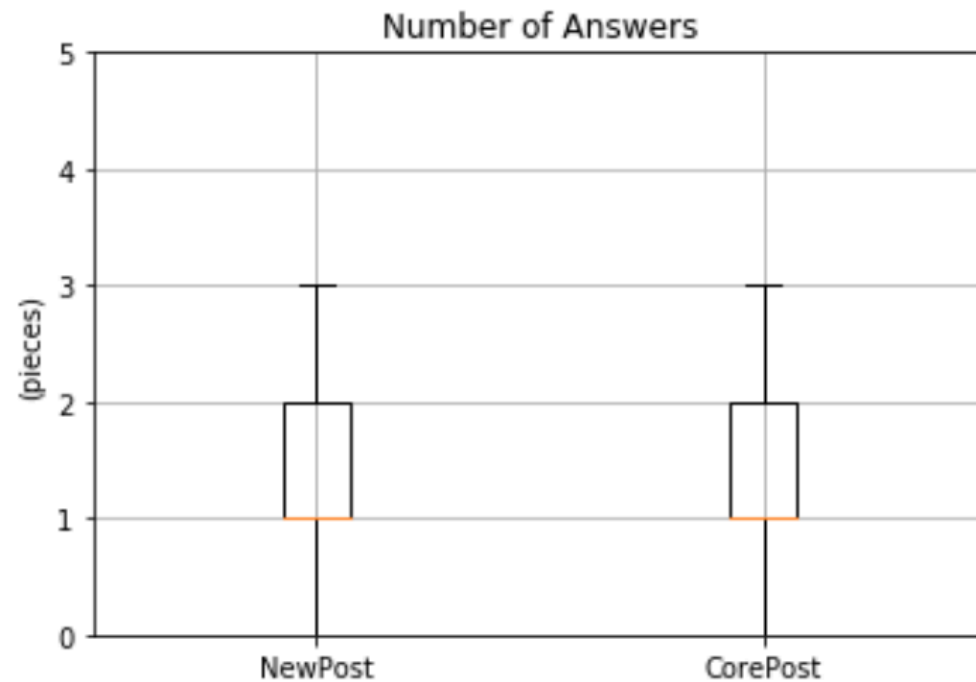
RQ1-1 | SOで質問を投稿したのが新人かどうかと、その質問に対して回答があるかどうかに関係性はあるか

	返信あり	返信なし	合計
新人 度数 期待度数 残差	396587 416565 -92	104132 84153 92	500719
新人以外 度数 期待度数 残差	848103 828124 92	147317 167295 -92	955420
合計	1244690	251449	1496139

カイ二乗値 = 8569.021785596831
P値 = 0.0
Cohen's w = 0.07567974351627675

SOの質問の投稿では新人の方が回答の割合は少ないが差は無視できるほど小さいと言える

RQ1-2 | 新人と新人以外では、SOに投稿した質問に対する回答数に差はあるか



	質問数	標準偏差	最小値	中央値
新人以外	995420	0.785	0	1
新人	500719	0.791	0	1

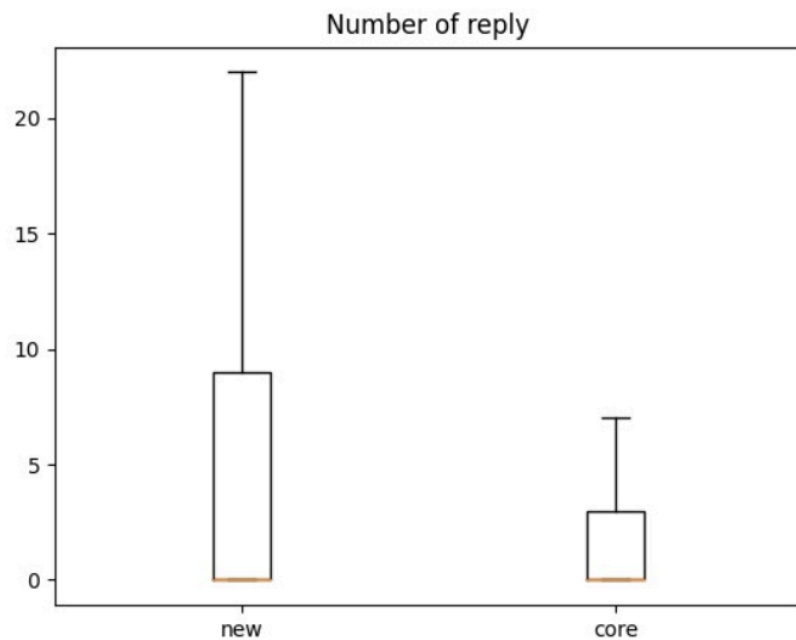
U = 186055266457.0

p = 0.0

Clif's d = 0.07839403863282868

SOの質問の投稿では回答数に差がない

RQ1-3| StackOverflow内で新人と新人以外で最初の回答がくるまでの時間に差があるか



SOの質問に回答が来るまでの時間は新人の方が長い傾向にあるがその差は無視できるほど小さい

	質問数	標準偏差	最小値	中央値
新人以外	764353	9.59	0	0
新人	344220	10.92	0	0

U = 139893813371.5

p = 0.0

cohen's w = 0.06340434311774779

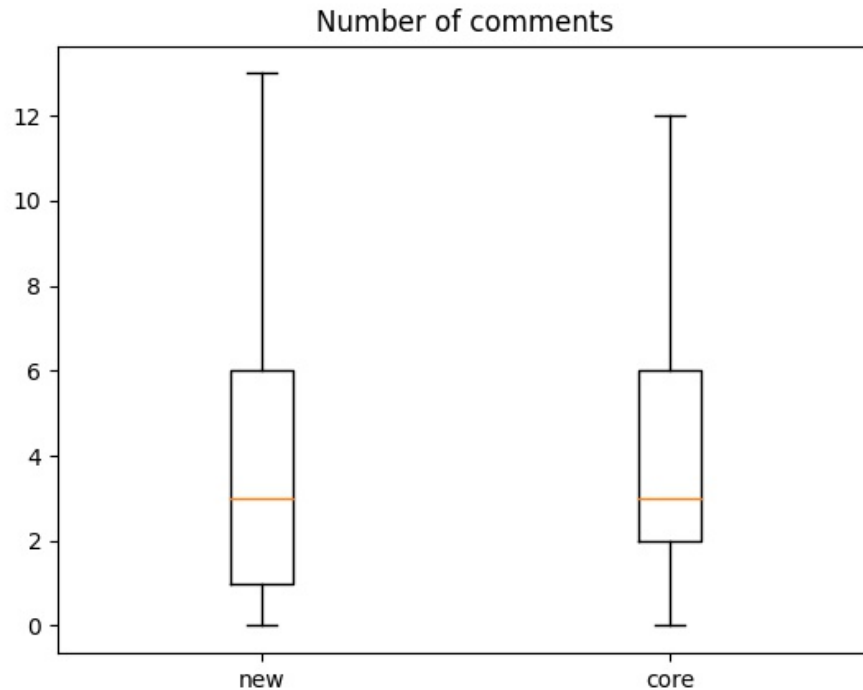
RQ1-4|GHでIssueを投稿したのが新人かどうかと、そのIssueに対してコメントがあるかどうかに関係性はあるか

	返信あり	返信なし	合計
新人	35263	1935	37198
度数	34113	3084	
期待度数	24	-24	
残差			
新人以外	109599	11164	120763
度数	110748	10014	
期待度数	-24	24	
残差			
合計	144862	13099	157961

カイ二乗値 = 611.146525187363
P値 = 6.30054166221563e-135
効果量w = 0.06220105289595013

GHのIssuesの投稿は新人の方がコメントが多い傾向にある

RQ1-5 | 新人と新人以外では、コメントの数に差があるのか



新人と新人以外でコメントの数に
差はない

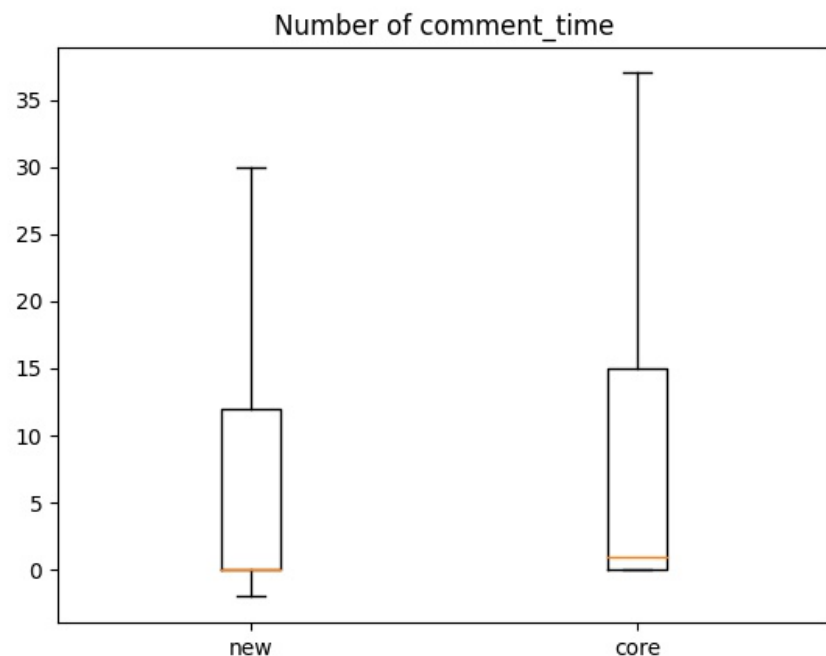
	投稿数	標準偏差	最小値	中央値
新人以外	120763	8.39	0	5
新人	37198	8.94	0	5

U = 2324652477.0

p = 7.45675270040183e-25

d = 0.03498617750975441

RQ1-6|新人と新人以外では、最初のコメントが来るまでの時間に差はあるのか



新人に対しての返信の方が少し早い傾向にあるがその差は無視できるほど小さい

	投稿数	標準偏差	最小値	中央値
新人以外	120763	9.310	0	0
新人	37198	11.453	0	2

U = 2118663616.0

p = 4.991184119664052e-210

cohen's w = 0.10027993727385764

RQ1-7|GHでPRを投稿したのが新人かどうかと、そのPRがマージされるかどうかに関係性はあるか

	マージされた	マージされなかった	合計
新人			
度数	2606	7936	10542
期待度数	5423	5118	
残差	-51	51	
新人以外			
度数	48880	46617	95497
期待度数	46367	49129	
残差	51	-51	
合計	51486	54553	106039

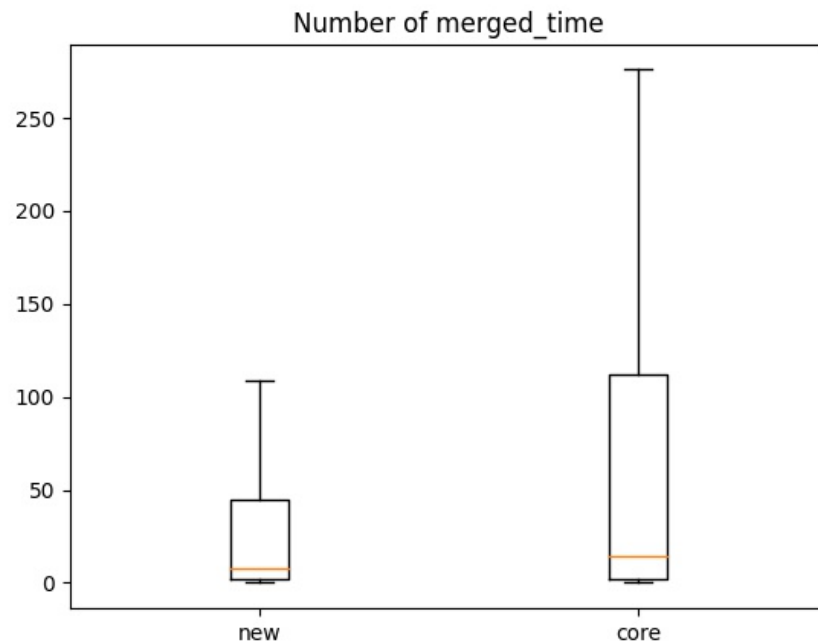
カイ二乗値 = 2661.9757677999537

P値 = 0.0

効果量w = 0.15844160646380934

PRは新人だとマージされない傾向があるが、効果量から無視できるほどの差であると言える

RQ1-8|新人と新人以外では、 マージされるまでの時間に差はあるのか



新人の方がマージされるまでの時間は短い傾向にあるがその差は無視できるほど小さい

	PR数	標準偏差	最小値	中央値
新人以外	95497	6.326	0	14
新人	37198	9.101	0	8

U = 69898773.0

p = 3.2271760380315114e-17

cohen's w = 0.09747323939593011

考察

- 新人と新人以外でIssuesの投稿は返信率, コメント数, 返信時間ともに差はなかった
 - 新人と新人以外で差がないこと自体問題?
 - 他の要素で障壁を感じている可能性
 - コメントが不親切など
- PRでは新人はマージされにくい傾向にあるが, マージされるまでの時間は早い
 - 簡単な変更はすぐにマージされ, 難しい内容に手を出している人がマージされていない

今後の予定

- RQ2を終わらせる
- コメントやPR, Issuesの内容を分析する予定

メモ