

“This Is Damn Slick!”

Estimating the Impact of Tweets on Open Source Project Popularity and New Contributors

Hongbo Fang, Hemank Lamba,
James Herbsleb, Bogdan Vasilescu

2022 International Conference on Software Engineering (ICSE)

発表者：井町 瑠貴

背景

- | 多くのOSSプロジェクトにとって、コミュニティから注目を集めることが課題となっている
- | 先行研究では、GitHubのREADMEファイルの構成、プラットフォームによるプロジェクトの紹介などについて研究されている

背景

- | 前述のように，先行研究では，プラットフォーム内の行動に焦点が当てられている
- | しかし，OSSに関するプラットフォーム外（SNS上）での行動がもたらす影響は大きいと考えられる

本研究では，OSSプロジェクトに関するツイートの有効性を評価する

リサーチクエスチョン

RQ1

ツイートは、人気や新しい貢献者にどのような影響を与えるのか

RQ2

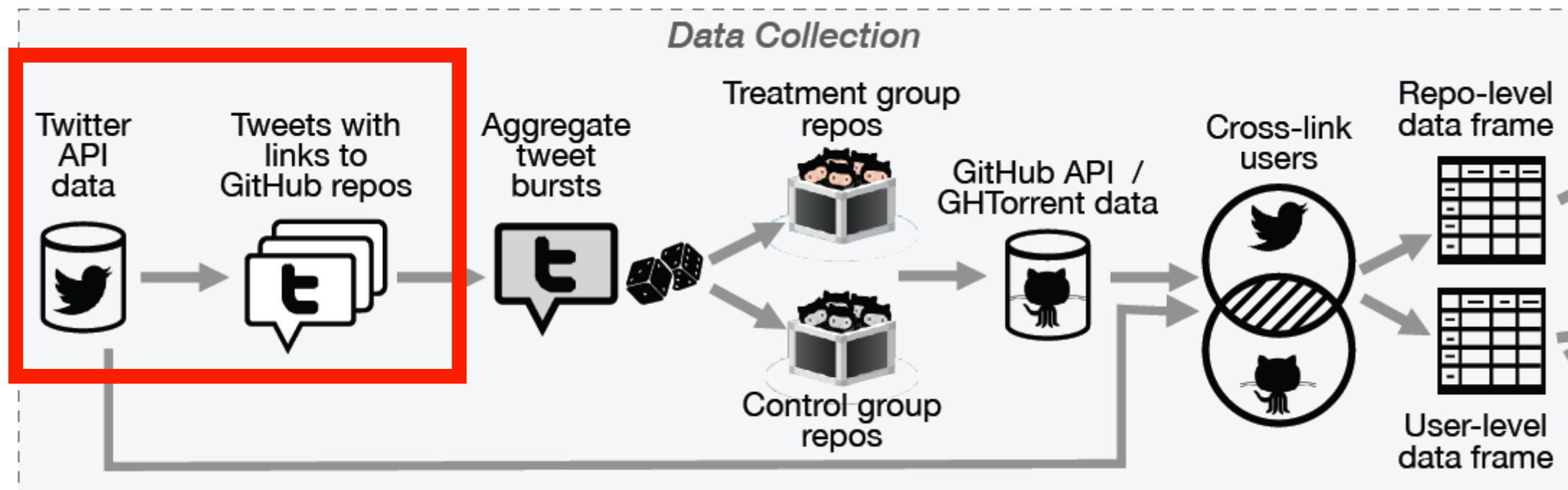
ツイートの影響は、ツイートの特徴の違いによってどのように異なるか

RQ3

ツイートによって集まる可能性が高い貢献者の特徴は何か

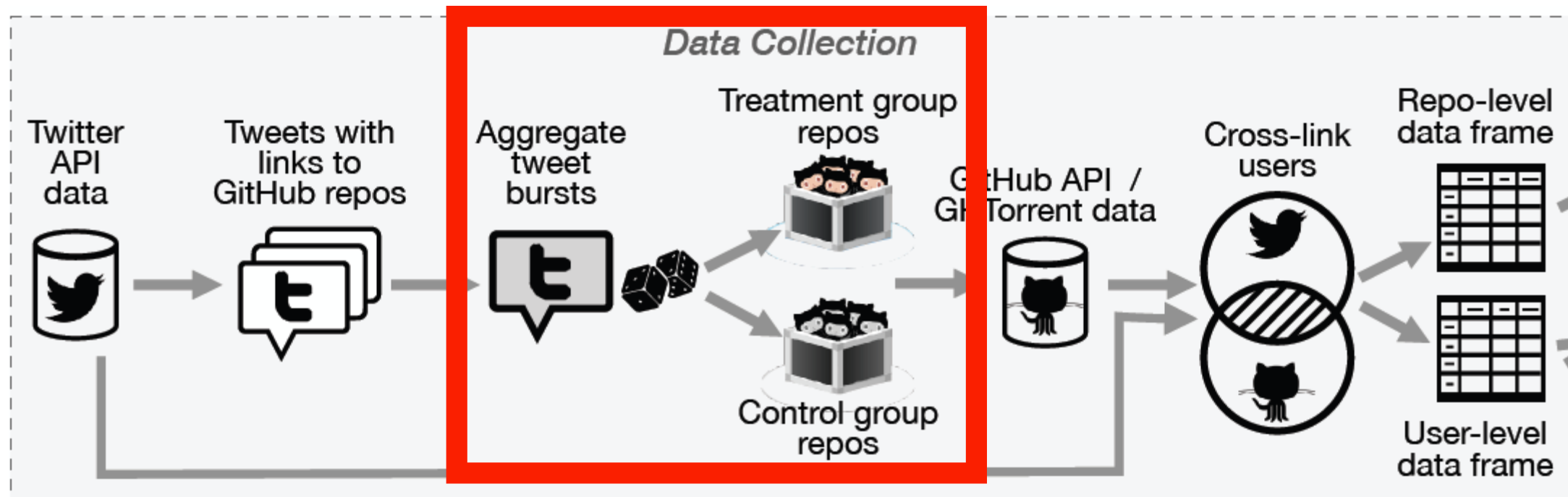
アプローチ | データ収集

- TwitterAPIにより, GitHubの成果物へのリンクを含むツイートを331,627 件収集
- GitHubから除外されているものなどを除外し, 10,837件のツイートを抽出



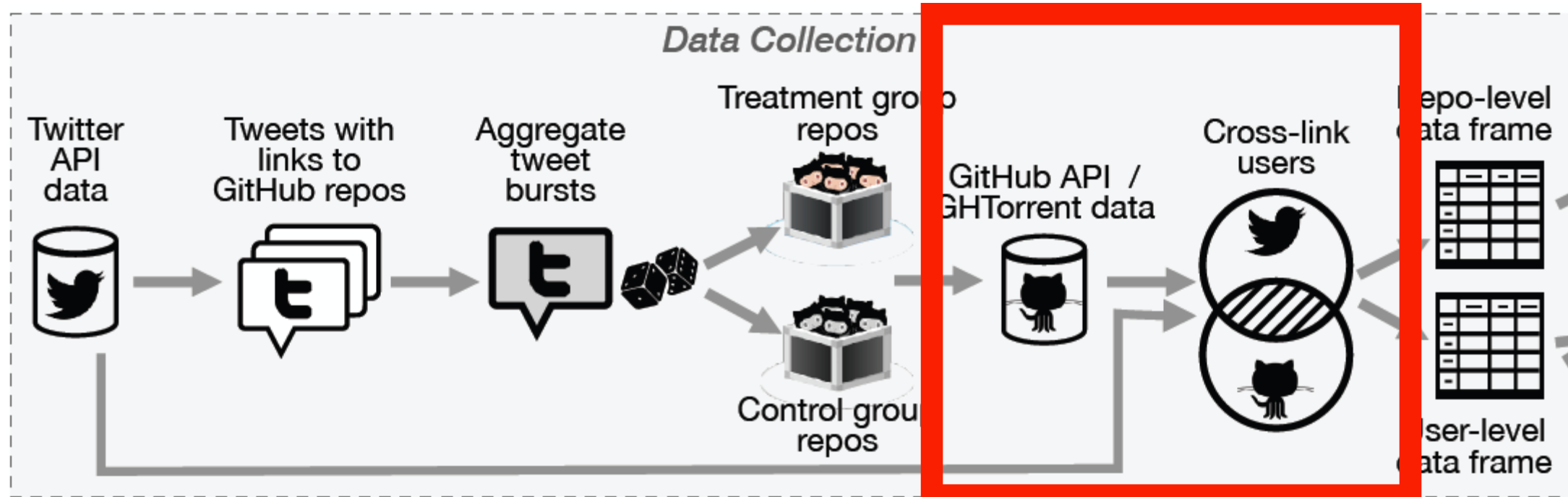
アプローチ | データ収集

- ・ ツイートをバーストとして集約する
- ・ リポジトリを介入グループと，コントロールグループに分類する



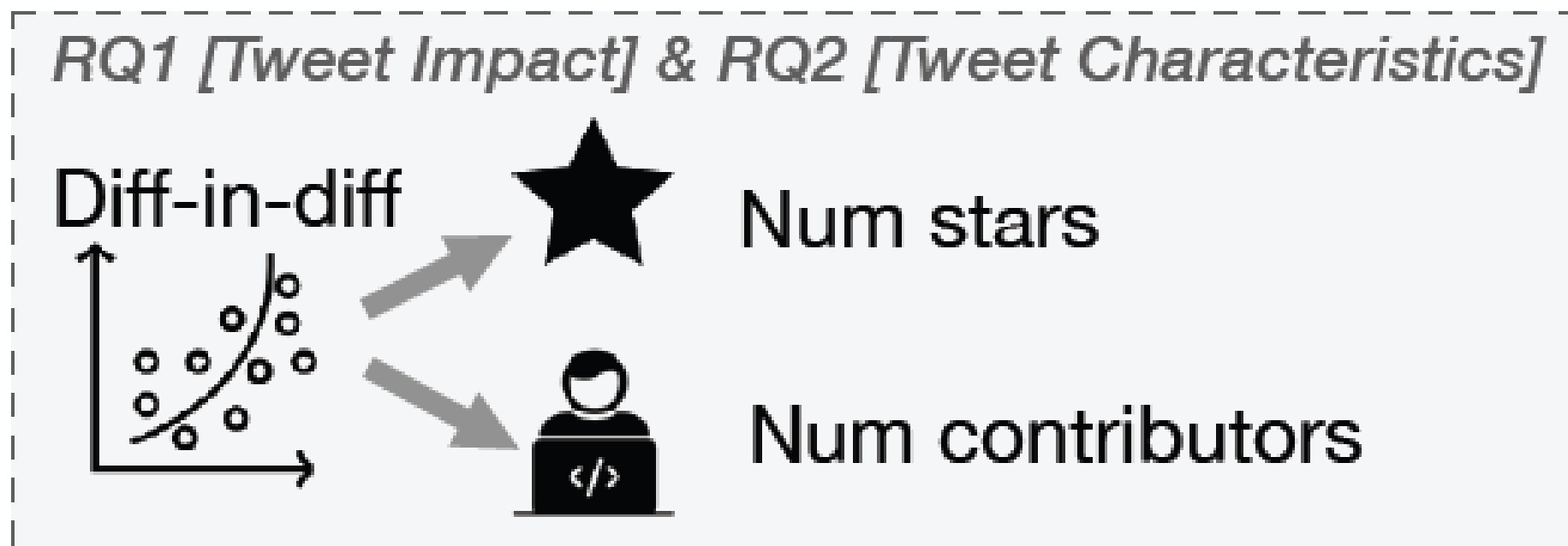
アプローチ | データ収集

- GitHubAPIやGHTorrentにより, コミットやその他の活動データを取得する
- TwitterとGitHubのユーザをリンクさせる



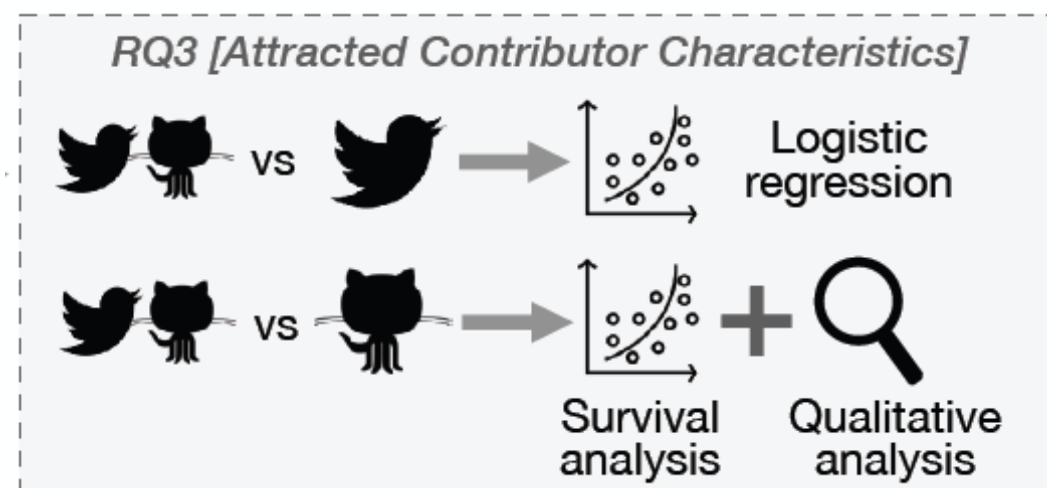
アプローチ | RQ1, RQ2

- RQ1では、差分の差分法を用いて、バーストがリポジトリに与える影響をスター数とコミッター数の変化から分析する
- RQ2では、さらにツイート特性を追加して分析を行う



アプローチ | RQ3

- 新規コミッターについて、定量的な分析と定性的な分析を行う
 - 定量的：他のプロジェクト参加者、同じツイートを見たがコミッターにならなかった開発者と比較
 - 定性的：Twitter, GitHubでの交流を分析



結果分析 | RQ1

RQ1 : ツイートは, 人気や新しい貢献者にどのような影響を与えるのか

- Model I : バースト後に獲得したスター数
 - 平均して約7%のスターの増加がみられる
- Model IV : バースト後に獲得したコミッター数
 - 平均して約2%のコミッター増加がみられる

	Model I	Model IV
<i>Main Treatment Effect</i>		
Is post-treatment	-0.24(0.01)***	-0.08(0.00)***
Is treated group?	0.24(0.02)***	0.01(0.00)*
Is treated group? : Is post-treatment?	0.07(0.02)***	0.02(0.00)***

結果分析 | RQ2

RQ2 : ツイートの影響は, ツイートの特徴の違いによってどのように異なるか

- オリジナルのツイート数が多いほど, スター数への影響が大きい
 - オリジナルツイート数が2倍になると, スターが平均1.8個増加する
- プロジェクトの貢献者によるバーストは, そうでないものに比べて, 獲得スター数が38%少ない
 - ツイートの客観性が低いからと考えられる

結果分析 | RQ3

RQ3 : ツイートによって集まる可能性が高い貢献者の特徴は何か

- 定性的な分析により, 以下の3つのテーマがみられた
 1. TwitterやGitHubでの交流
 2. リーダーや影響力のある開発者をフォローしている
 3. 弱いつながり

結果分析 | RQ3

RQ3 : ツイートによって集まる可能性が高い貢献者の特徴は何か

Table 3: Characterizing the developers attracted by tweets

	Model VII (is attracted)	Model VIII (30-day commits)	Model IX (disengagement)
<i>Project controls</i>			
Project age (log)		-0.03 (0.01)**	0.13 (0.03)***
Project contribs (log)		-0.08 (0.01)***	0.11 (0.03)***
Project commits (log)		0.07 (0.01)***	-0.17 (0.02)***
<i>Developer Characteristics</i>			
GITHUB tenure (log)	-0.36 (0.13)**	-0.00 (0.01)	-0.02 (0.04)
GITHUB commits (log)	0.01 (0.05)	-0.02 (0.01)**	0.01 (0.01)
Has GITHUB collab.	1.47 (0.13)***	0.06 (0.02)**	-0.34 (0.05)***
Has Twitter interact.		0.07 (0.03)*	0.01 (0.07)
Twitter tenure (log)	0.15 (0.10)		
Num. tweets (log)	-0.44 (0.07)***		
Ratio original tweets	0.84 (0.28)**		
Num. obs.	1, 914	2, 192	2, 281

* $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$

結論

- | GitHub リポジトリを Twitter で紹介することで、プロジェクトの人気度が向上し、新規開発者の獲得につながることを実証
- | スター数の獲得に比べて、新規開発者の獲得に対する効果は少ない
これは、技術的貢献をするためのハードルが高いためである

所感

- メンテナはプロジェクトが人気になると、ユーザから多くの要求が発生して、手に負えなくなる可能性があるということを理解する必要があると感じた
- 注目を集めることも重要だが、技術的貢献のハードルの高さの方が問題に感じた

補足資料

補足 | 用いた変数の定義

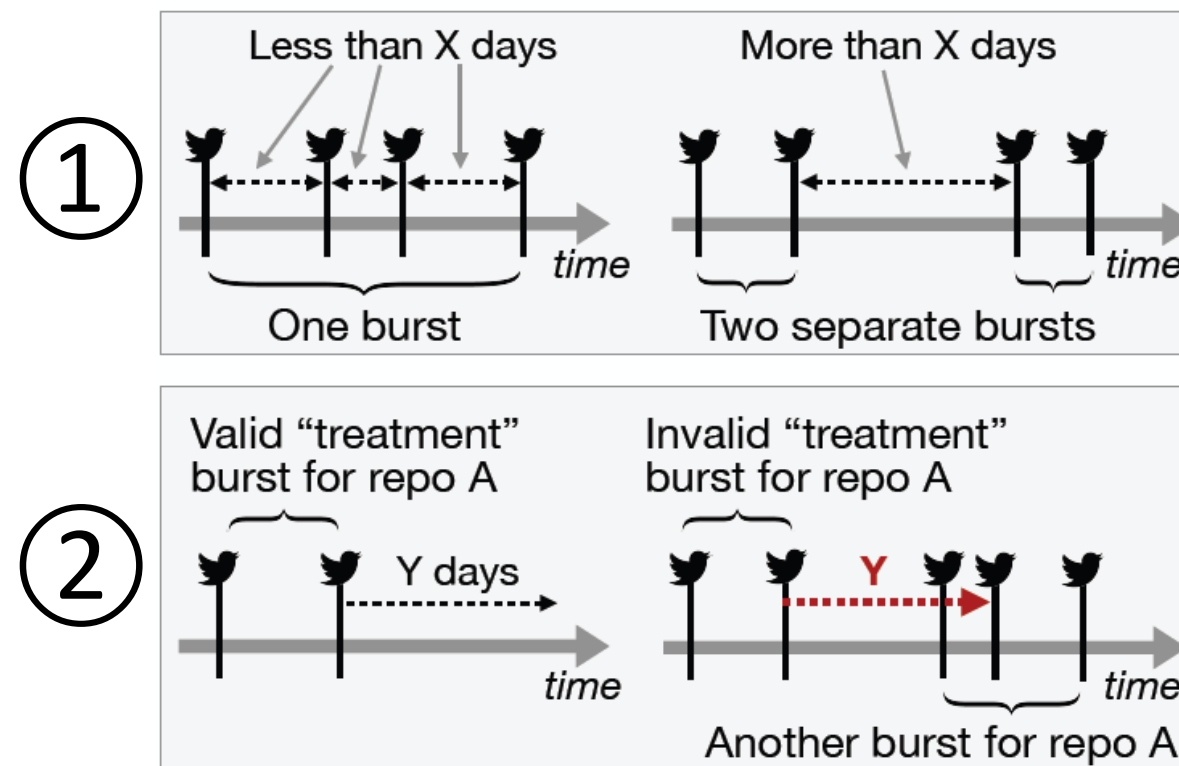
Table 1: The definitions of the variables in our models.

<i>Outcome Variables</i>		<i>Controls for Concurrent Events</i>	
Number of new stars	The number of stars received by the mentioned repository during the observation period.	Had official release	True if the repository had an official release on GITHUB during the period of tweet burst or within 24-hour after the start of the burst.
Number of new committers	Analogous to stars. New committers are defined as developers never made commit or sent pull request to the focal repository before the tweet mention.	Is GITHUB trending	True if the repository appeared on the GITHUB trending page during the period of tweet burst or within 24-hour after the start of the burst. We collect historical records of trending repositories from a public dataset [52].
<i>Tweet Burst Characteristics</i>		<i>Repository Characteristics</i>	
Number of likes	Sum of Twitter likes across all tweets in a burst.	Number of Google search results	The number of Google search results containing the exact repository slug and originating outside of the twitter.com domain during the period of tweet burst; serves as a proxy for the scale of promotion repositories may have received outside of Twitter.
Number of original tweets	Tweet count in a burst, excluding retweets.	Burst duration	The duration of tweet burst (in hours). The longer the duration, the longer the period under observation.
Number of retweets	Retweet count in a burst.	<i>Attracted Developer Characteristics</i>	
Average tweet length	Measured characters; longer, more descriptive tweets can convey more information and may have stronger effects. The VIF test revealed collinearity with Number of original tweets and it was thus removed in the final model.	Project age	Number of days since the first commit.
Has hashtags	True if more than half of the tweets in the burst contain hashtags (we also use 'any tweet in the burst contains hashtags' as alternative construction). Hashtags are known to impact information diffusion [58] and are expected to increase the visibility of tweets.	Project contribs	Total number of commit authors, excluding obvious bots [21].
Is from committers	True if more than half of the tweets in the burst is authored by a current or past project commit author (we also use 'at least one tweet authored by current or past project committers' as alternative construction). In general, the identity of the creators has a big impact on the popularity of social media content [6].	Project commits	Total number of commits by the previous authors.
Is promotional	True if more than half of the tweet burst contain an URL pointing to a repository homepage. Following Fang et al. [26] we assume that such tweets are primarily intended for promotional purposes, while tweets with links to other GITHUB artifacts, most often specific issues, tend to serve other purposes. See Appendix for the validation of this approach.	GITHUB tenure	Number of days since the new contributor's first commit anywhere on GITHUB.
		GITHUB commits	Total number of commits by the new contributor anywhere on GITHUB.
		Has GITHUB collab	True if the new contributor ever committed to the same repository as any of the focal project's (i.e., the project mentioned by tweets) committers in the past.
		Twitter tenure	Number of days since the new contributor's first tweet.
		Num. tweets	Total number of original tweets (i.e., excluding retweets) authored by the new contributor in the past year.
		Ratio original tweets	The fraction of original tweets to all tweets authored by the new contributor in the past year.
		Has Twitter interaction	True if the new contributor ever @-mentioned any of the authors of tweets in the focal burst in the past year.

補足 | バースト

バースト（本研究では $(x, y) = (3\text{day}, 3\text{day})$ ）

- ① 同じリポジトリに X 日以内に言及したツイートは、同じバーストに含まれる
- ② 同じリポジトリに言及した2つのバーストは、少なくとも Y 日離れている



補足 | ツイートの影響

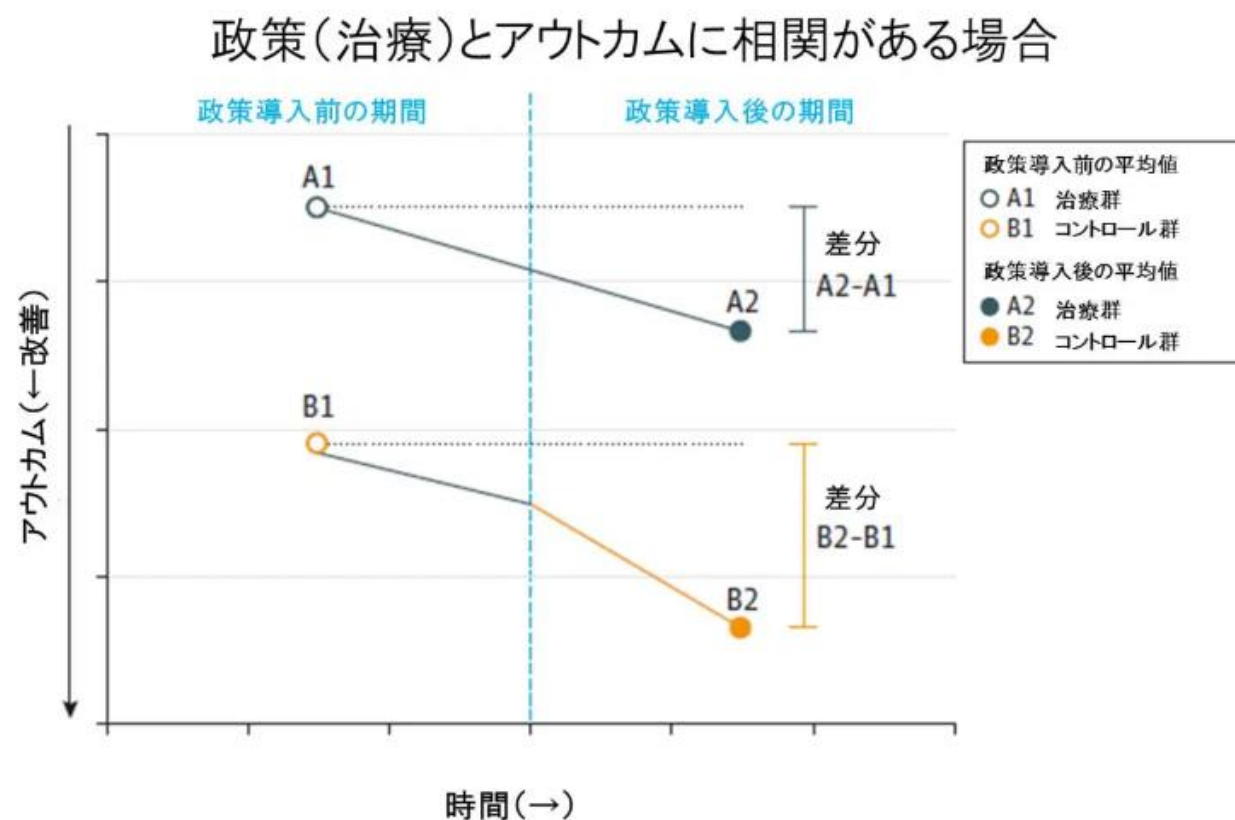
| ツイートに影響を受けたとは？

1. ツイート投稿後3日以内にリポジトリに言及したツイートのいずれかをリツイートした
2. 同期間に言及したりリポジトリにスターを付けた

補足 | 差分の差分法

差分の差分法 (difference in differences)

目的変数の時系列推移が介入群と似ているグループを比較対象として、介入前後の目的変数の変化を比較する



参考

| 差分の差分法について

<https://www.ariseanalytics.com/activities/report/20210423-2/>

<https://healthpolicyhealthecon.com/2015/07/03/difference-in-differences/>