

研究発表

OSSコミュニティの傾向の変化を 考慮したコミッター候補者予測

山崎 大輝

背景

- OSSが社会に普及するにつれて、コミュニティへの不具合報告が増加
- 不具合を解決するためにパッチ投稿が増加し、それを検証するコミッターへの負担が増えている

コミッターに関して

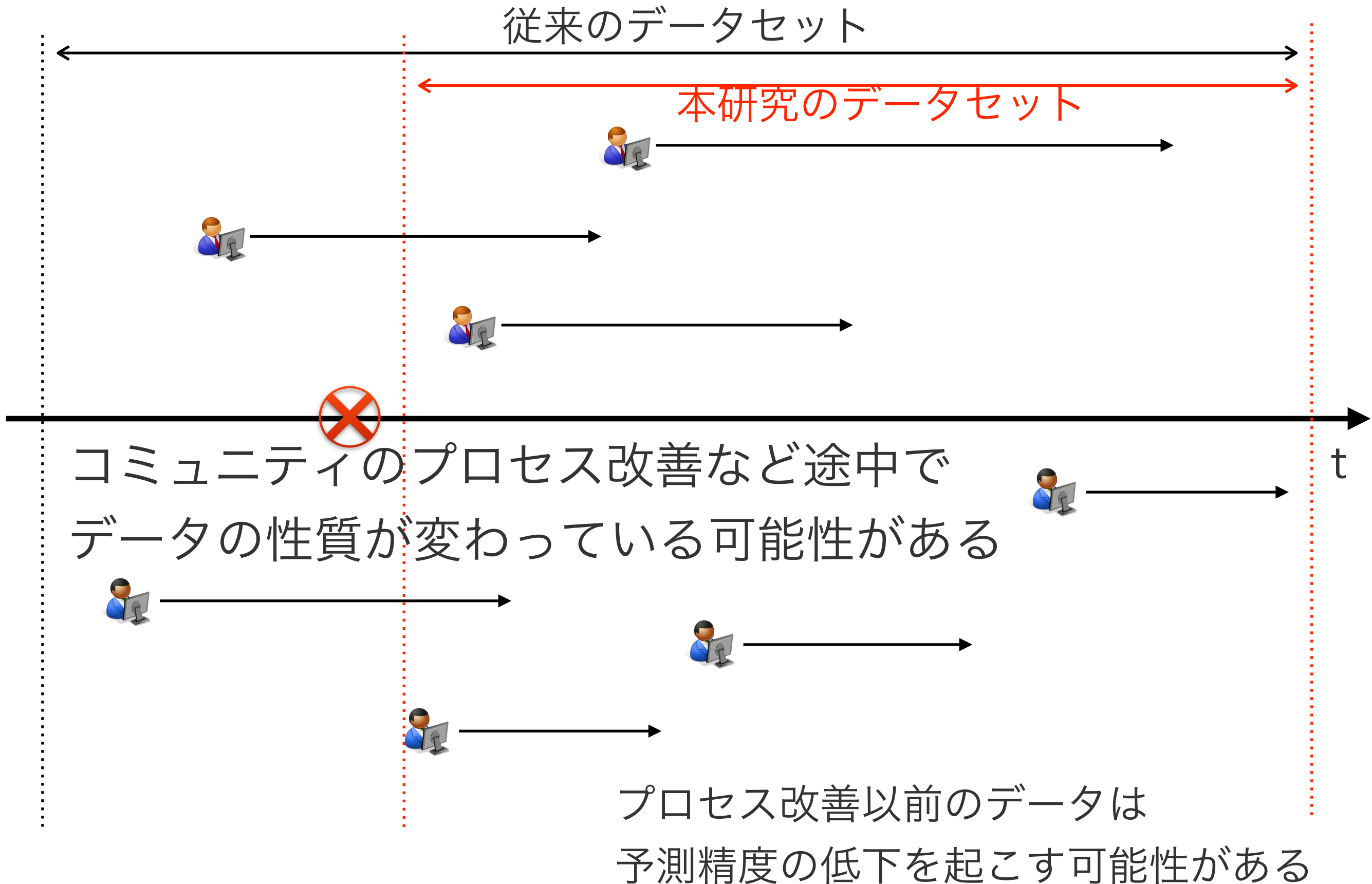
- コミッター候補はプロジェクト管理者や既存コミッターが推薦[1]
- コミッター候補となる人物は、早期に発見しないとプロジェクトを脱退する恐れ
- 早期にコミッター候補となる人物を見つける必要があるが、推薦のためには長期にわたる活動が求められる

先行研究

- コミッターと非コミッターの活動量を元に
コミッター候補者を予測する^[2]
- OSSコミュニティは時期によって環境が変わると考えられる→データの性質が変わっている可能性が高い
- 時期を限定して予測モデルの説明変数にした方が、より高い精度でコミッター予測をできるのではないか？

[2]伊原彰紀, 亀井靖高, 大平雅雄, 松本健一, 鵜林尚靖: OSSプロジェクトにおける開発者の活動量を用いたコミッター候補者予測, 電子情報通信学会論文誌,

本研究の着眼点



目的とアプローチ

- 目的
 - コミッター候補となるべき人物の早期発見
 - データの数はモデルの予測精度に影響を与えるか？
 - (データに傾向が存在するか？)
- アプローチ
 - データセット期間を調整することでコミッター候補者予測の精度がどう変化するか調べる
 - 他の期間に予測モデルを適用し、本当にその期間に最適な予測モデルになっているかを調べる

データセット



期間 2001/10 - 2012/10

コミッター 195人

一般開発者 13,297人



期間 1999/07 - 2013/01

コミッター 530人

一般開発者 76,931人



期間 2000/12 - 2014/04

コミッター 363人

一般開発者 8,017人

活動量メトリクス

開発期間

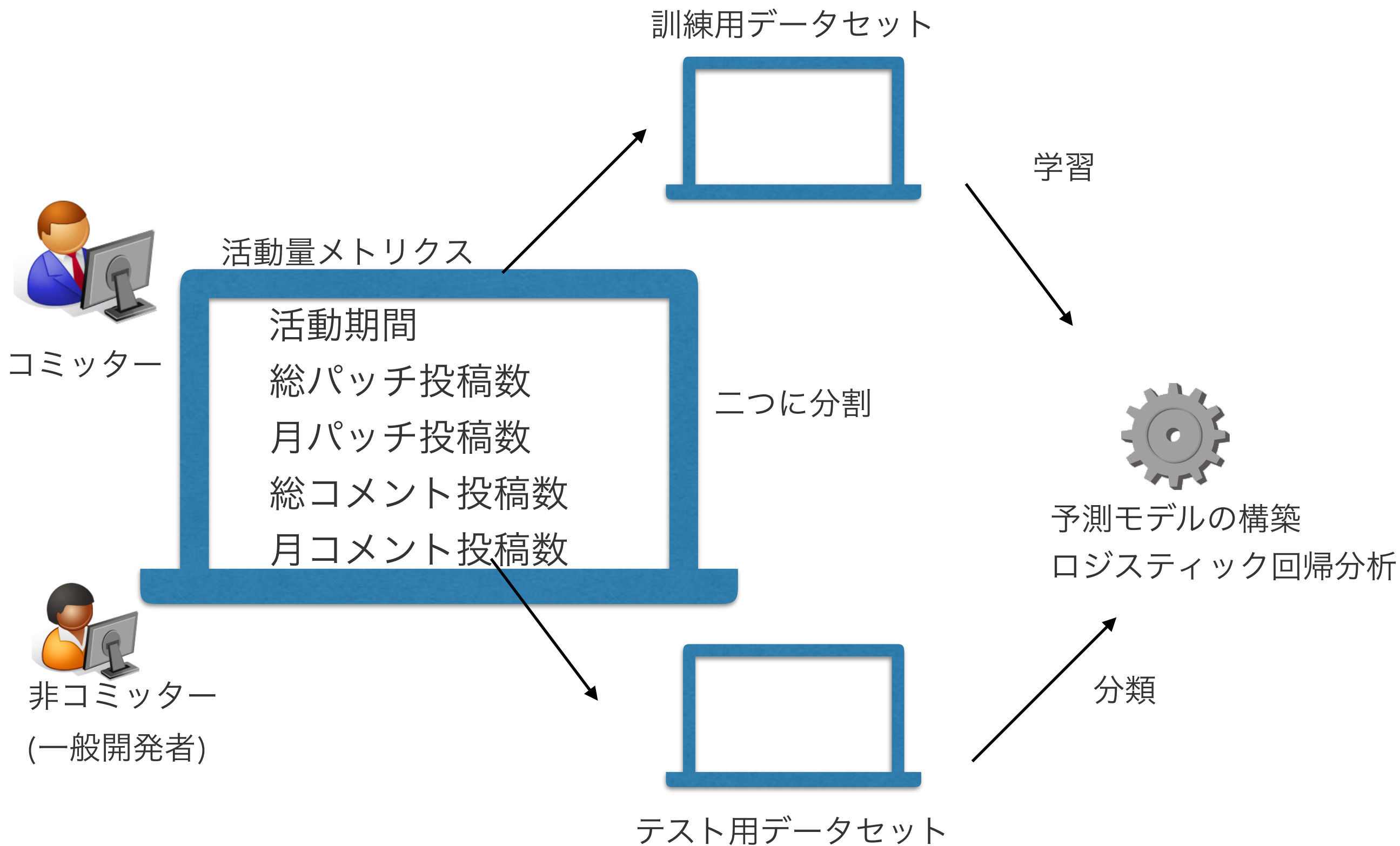
月当たりのコミット数

総コミット数

月当たりのパッチ投稿数

総パッチ数

コミッター候補者予測モデルの概要



活動量メトリクスの詳細

活動期間

開発者が活動(パッチまたはコメント投稿)した月の総数

総パッチ投稿数

対象期間内にパッチを投稿した総数

月パッチ投稿数

1 ヶ月の間でパッチを投稿した数の中央値

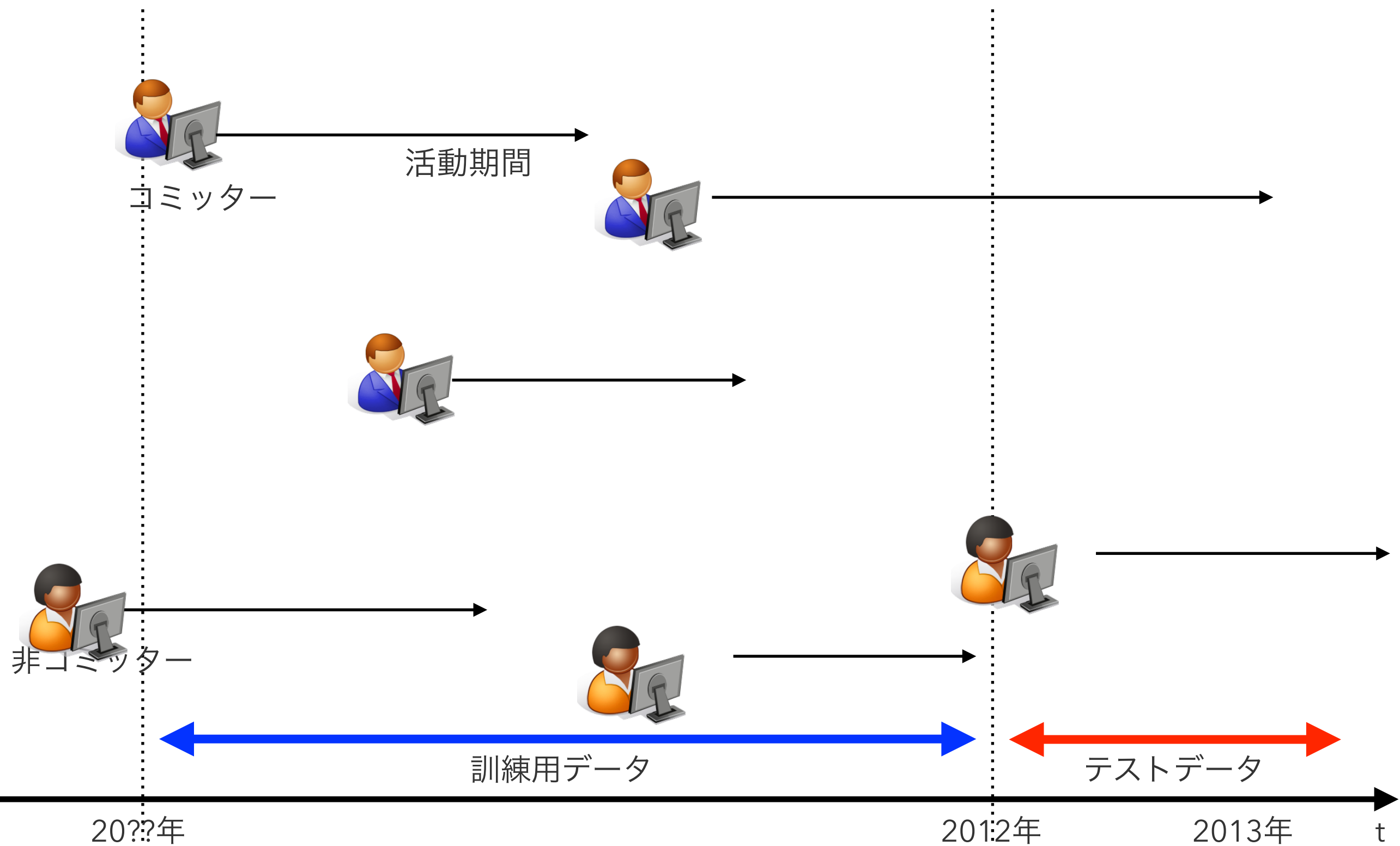
総コメント投稿数

対象期間内にコメントを投稿した総数

月コメント投稿数

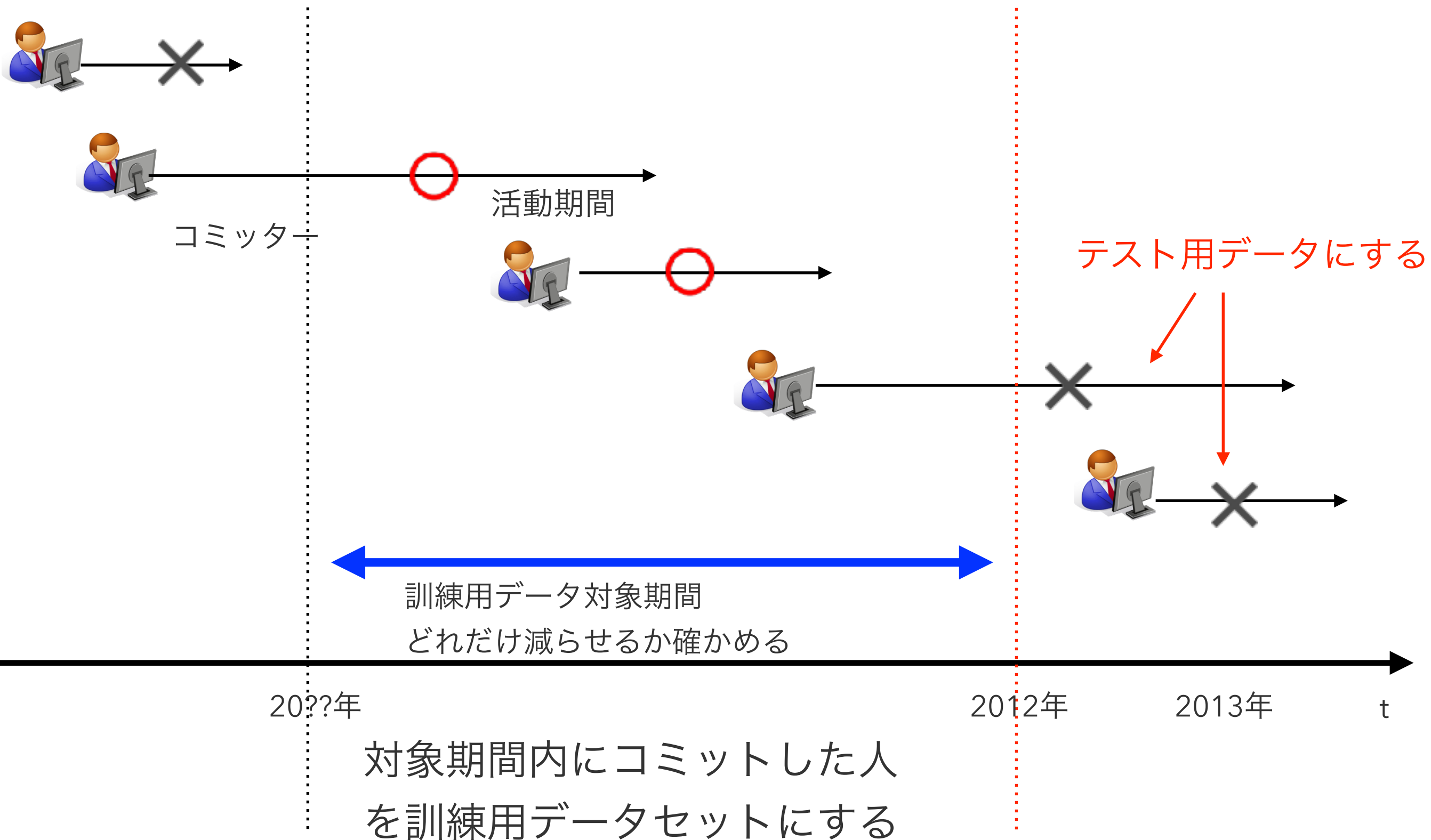
1 ヶ月の間でコメントを投稿した数の中央値

データセットの内訳

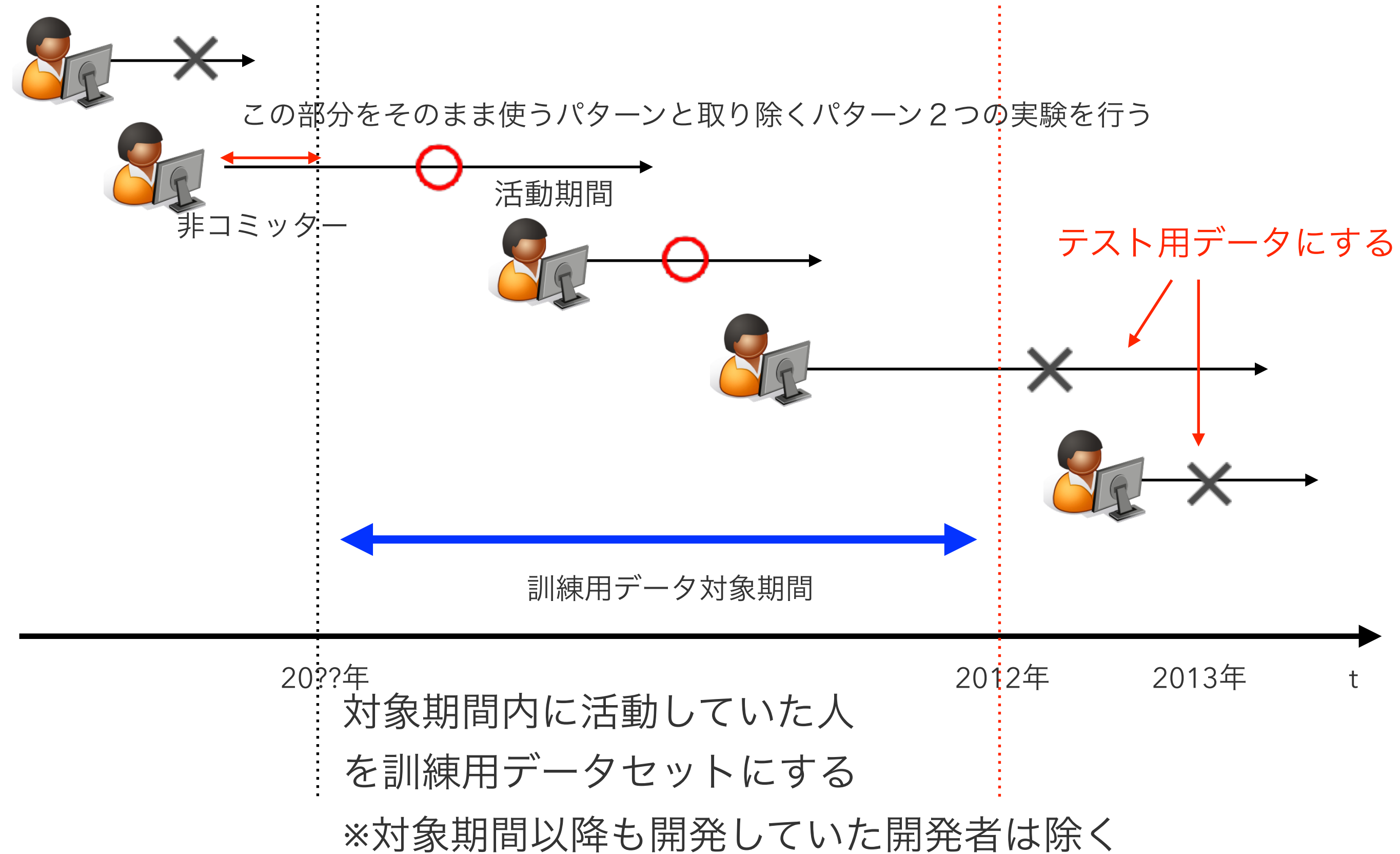


3ヶ月以上活動 & 総パッチ投稿数1以上の開発者のみ使用

訓練用データの対象(コミッター)



訓練用データの対象(非コミッター)



分析方法

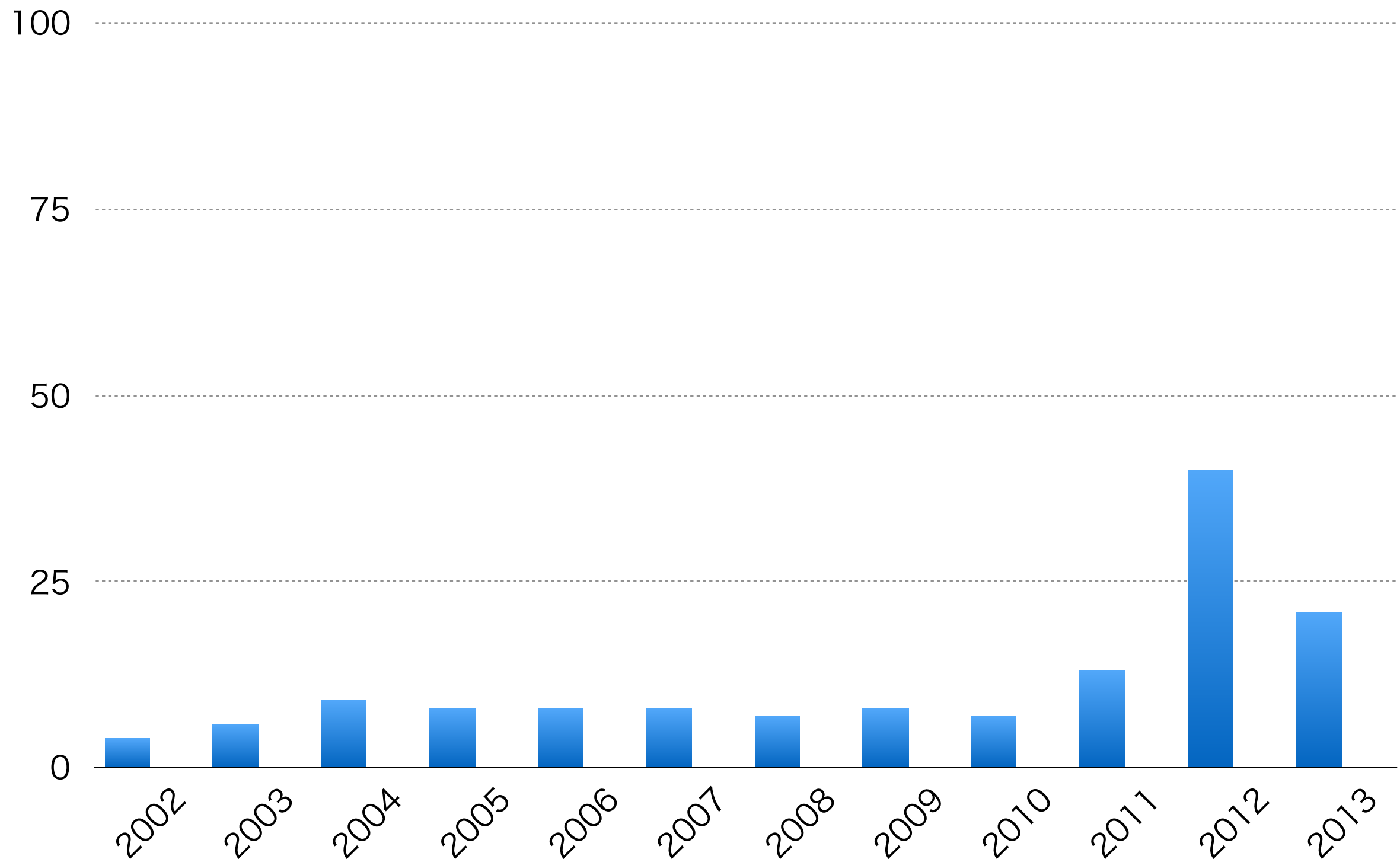
- 訓練用データをもとに予測モデルの構築
- コミッター,非コミッターの数が違うので数を揃える
- 予測モデルとしてロジスティック回帰分析を使用
 - Rの関数を使用, Thresholdは0.5に設定
 - なぜ回帰分析か？
 - . . .
 - なぜロジスティック回帰か？
 - . . .

やること(やっていること)

- パターン 2 (対象期間内の活動のみ使用)の実験
 - を行うためにデータセットを取得
- 論文

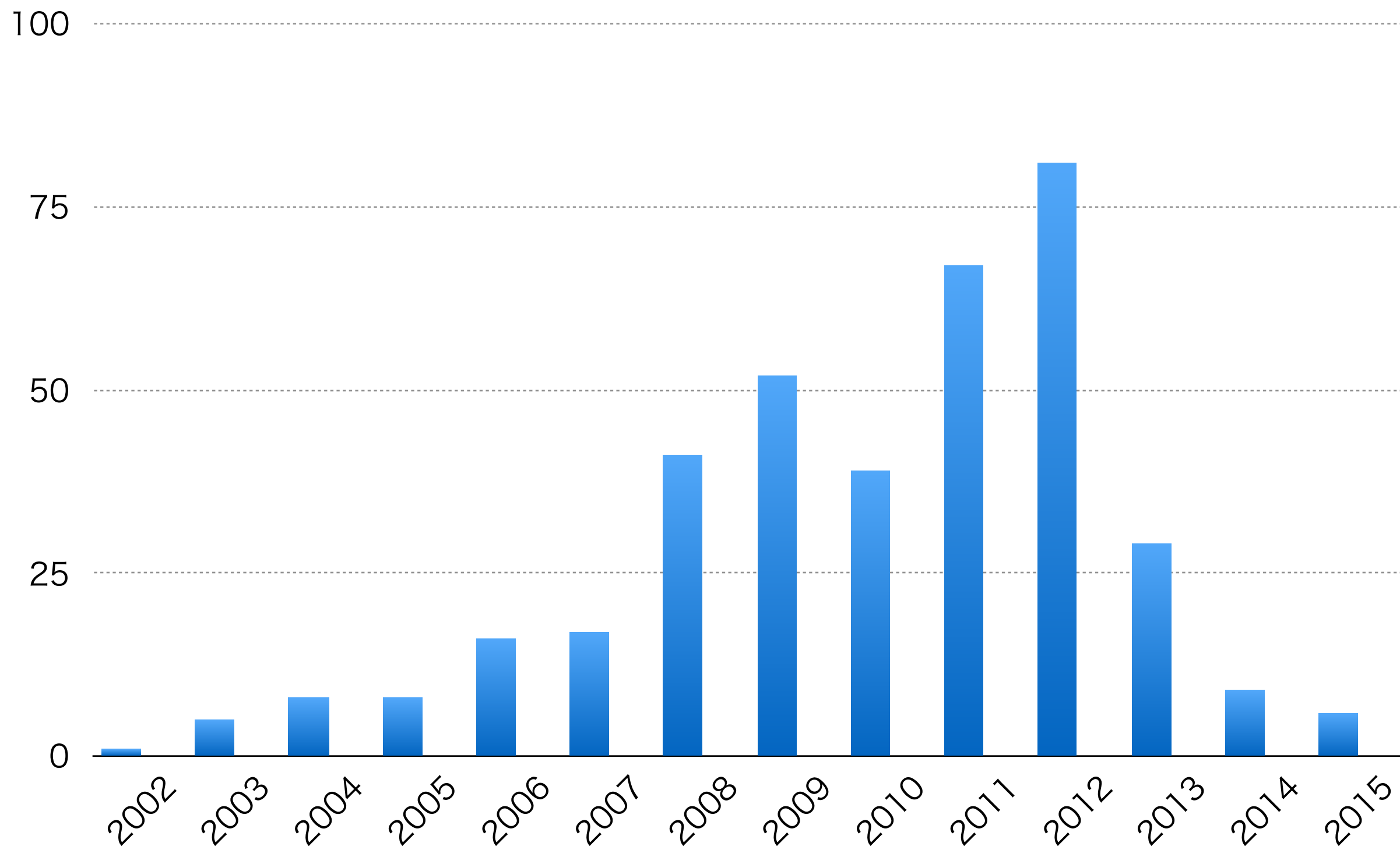
コミッター昇格時期

Eclipse(計139人)



時期ごとのコミッターに昇格した人数

Firefox(計379人)



結果(Firefox)

訓練用データセットが2001~2011年の間

```
[1] "committer : " "288"          " developer : " "478"
[1] "train_committer : " "205"        " train_developer : " "324"
[1] "test_committer : " "52"          " test_developer : " "151"
[1] precision :    0.394736842105263 recall :    0.576923076923077 f1 :
[6] 0.46969696969697
```

訓練用データセットが2006~2011年の間

```
[1] "committer : " "288"          " developer : " "478"
[1] "train_committer : " "171"        " train_developer : " "243"
[1] "test_committer : " "52"          " test_developer : " "151"
[1] precision :    0.420289855072464 recall :    0.576923076923077 f1 :
[6] 0.488188976377953
```

訓練用データセットが2011年の間

```
[1] "committer : " "288"          " developer : " "478"
[1] "train_committer : " "49"        " train_developer : " "71"
[1] "test_committer : " "52"          " test_developer : " "151"
[1] precision :    0.5 recall :    0.730769230769231 f1 :
[6] 0.594202898550725
```

それぞれの評価値ごとにグラフや表を
作って見せた方がわかりやすそう