

研究発表

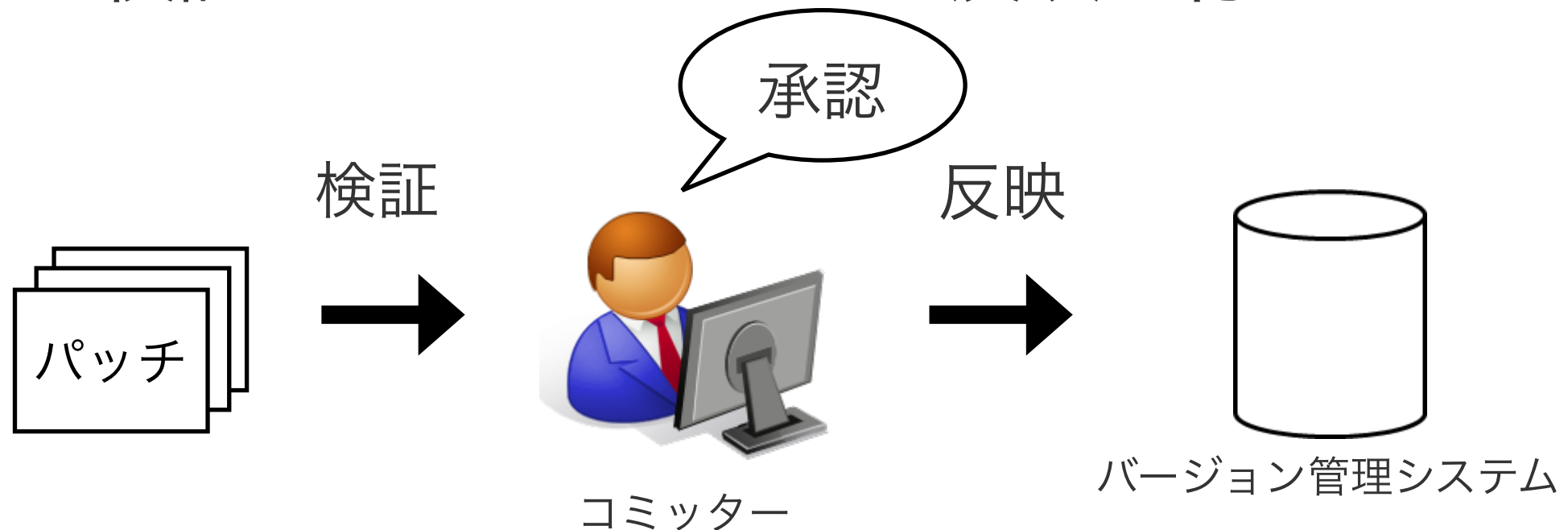
開発者の活動量の経時的変化が コミッター候補者予測に与える 影響の分析

オープンソースソフトウェア工学研究室

60191055 山崎 大輝

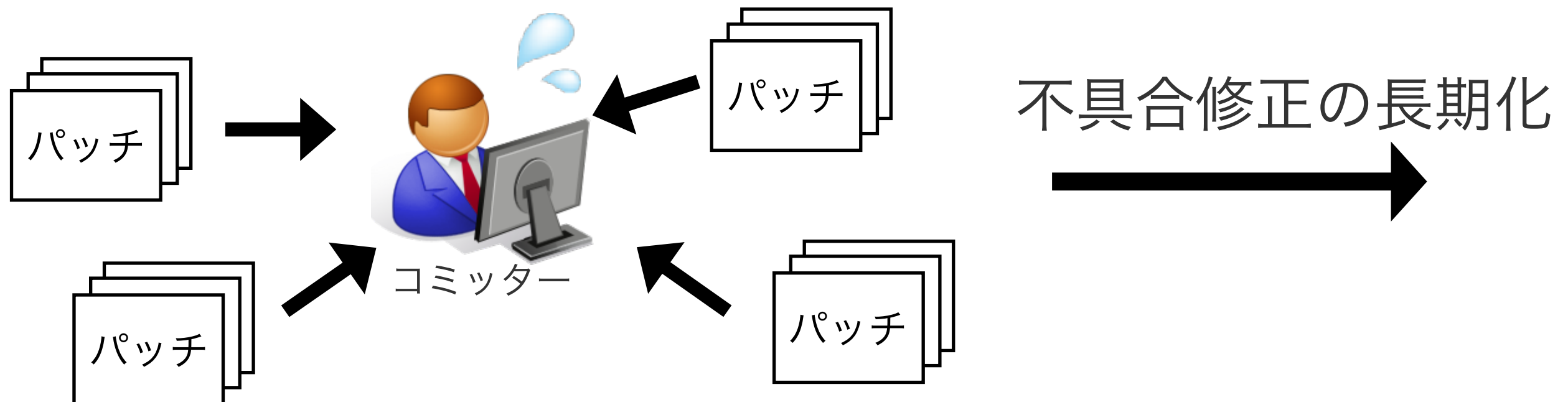
背景

- オープンソースソフトウェア(OSS)が社会に普及するにつれてプロジェクトへの不具合報告が増加している
- コミッタと呼ばれる開発者が
パッチの検証・プロジェクトへの反映を行う



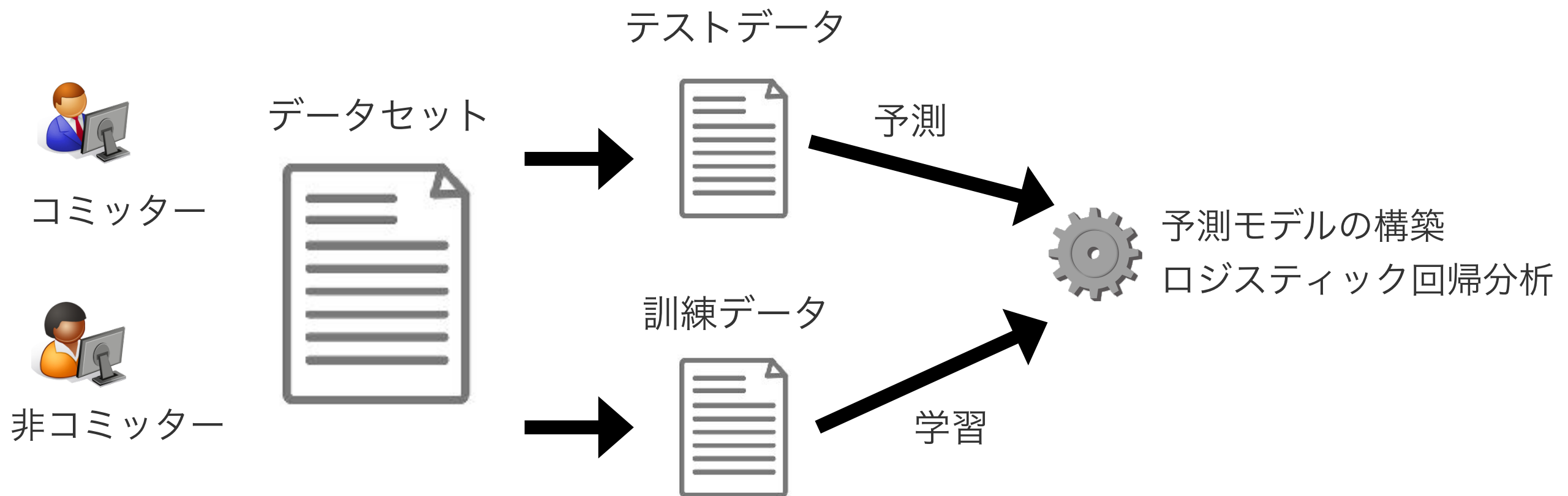
コミッタの現状

- コミッタは開発者の中でもごく一部であり，一人当たりの作業負担が集中している
- パッチの検証・反映が滞り，不具合修正が長期化している



既存研究

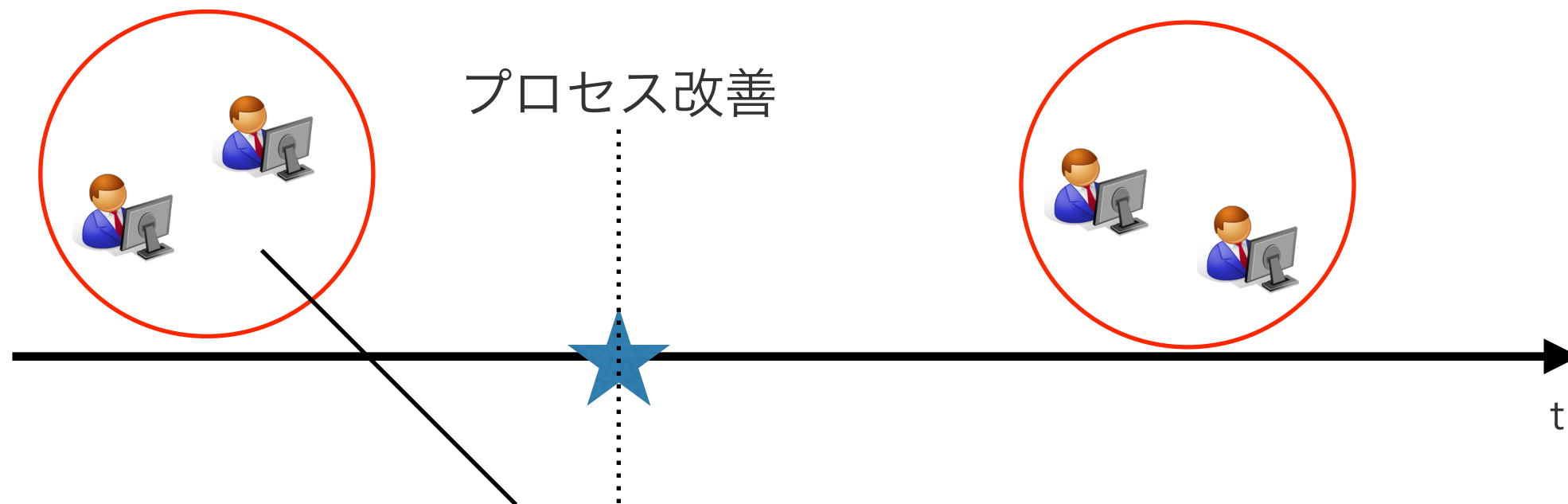
- コミッタと非コミッタの活動量メトリクスを元に
コミッタ候補者を予測するモデルの構築をしている^[1]



[1]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,

問題点

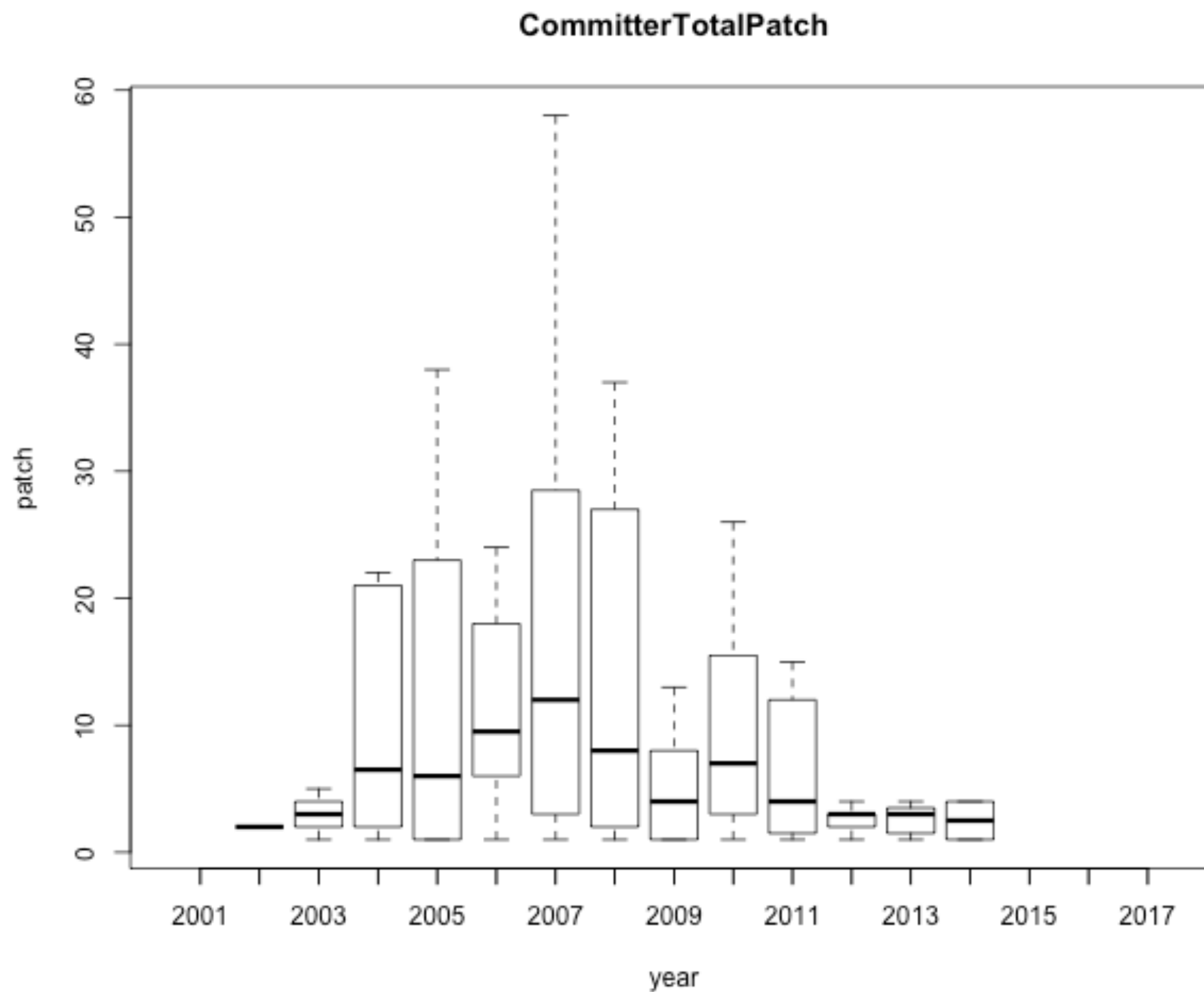
- 長期間のデータを用いてモデルを学習させている
→ プロジェクトのプロセス改善などでデータの性質が変わっている可能性がある^[2]



学習させると予測精度が低下する恐れがある

研究動機

各年のコミッタのパッチ投稿数



目的とアプローチ

- 目的
 - データの経時的変化がコミッタ候補者の予測精度に影響を及ぼすか
- アプローチ
 - データセットの期間を調整して予測することでコミッタ候補者の予測精度がどのように変化するか実験する

データセットの概要



期間 2003/01 - 2014/12

コミッタ 195人

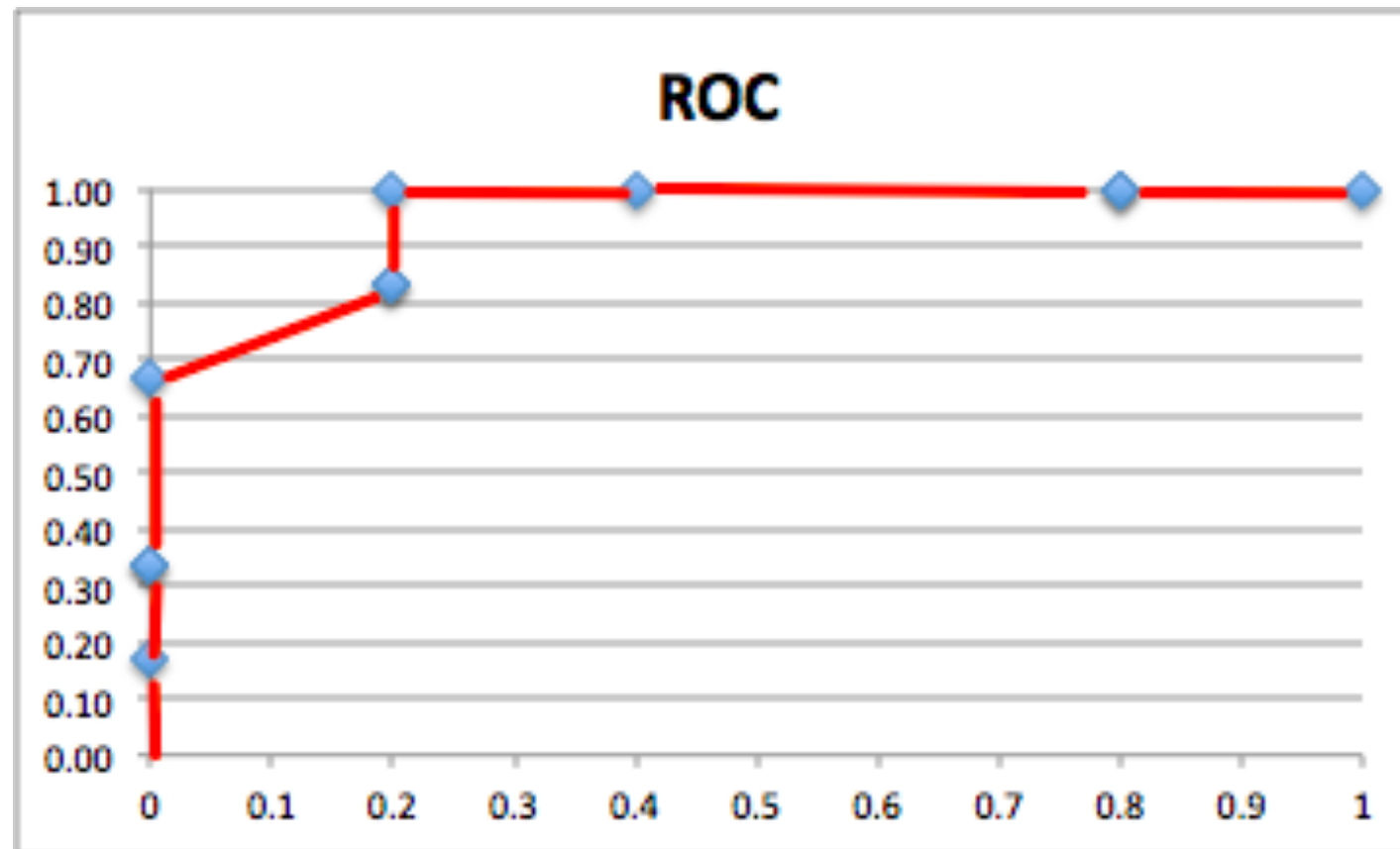
非コミッタ 13,297人

活動量メトリクス

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

評価指標

- ROC曲線の AUC (Area Under an ROC Curve)で評価

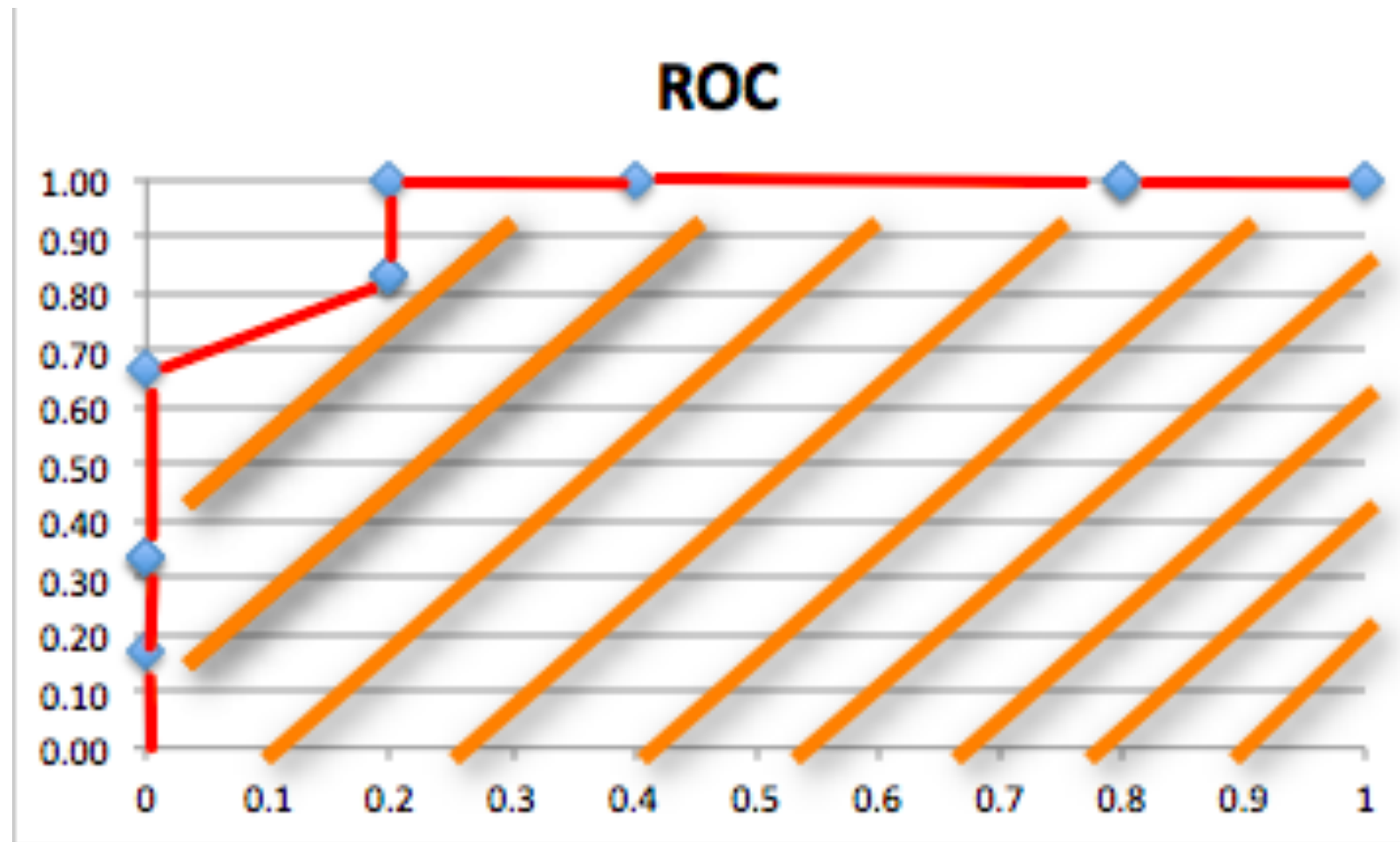


X軸:偽陽性率(間違って陽性と判定する割合)

Y軸:真陽性率(正解をできるだけ多く陽性と判定)

評価指標

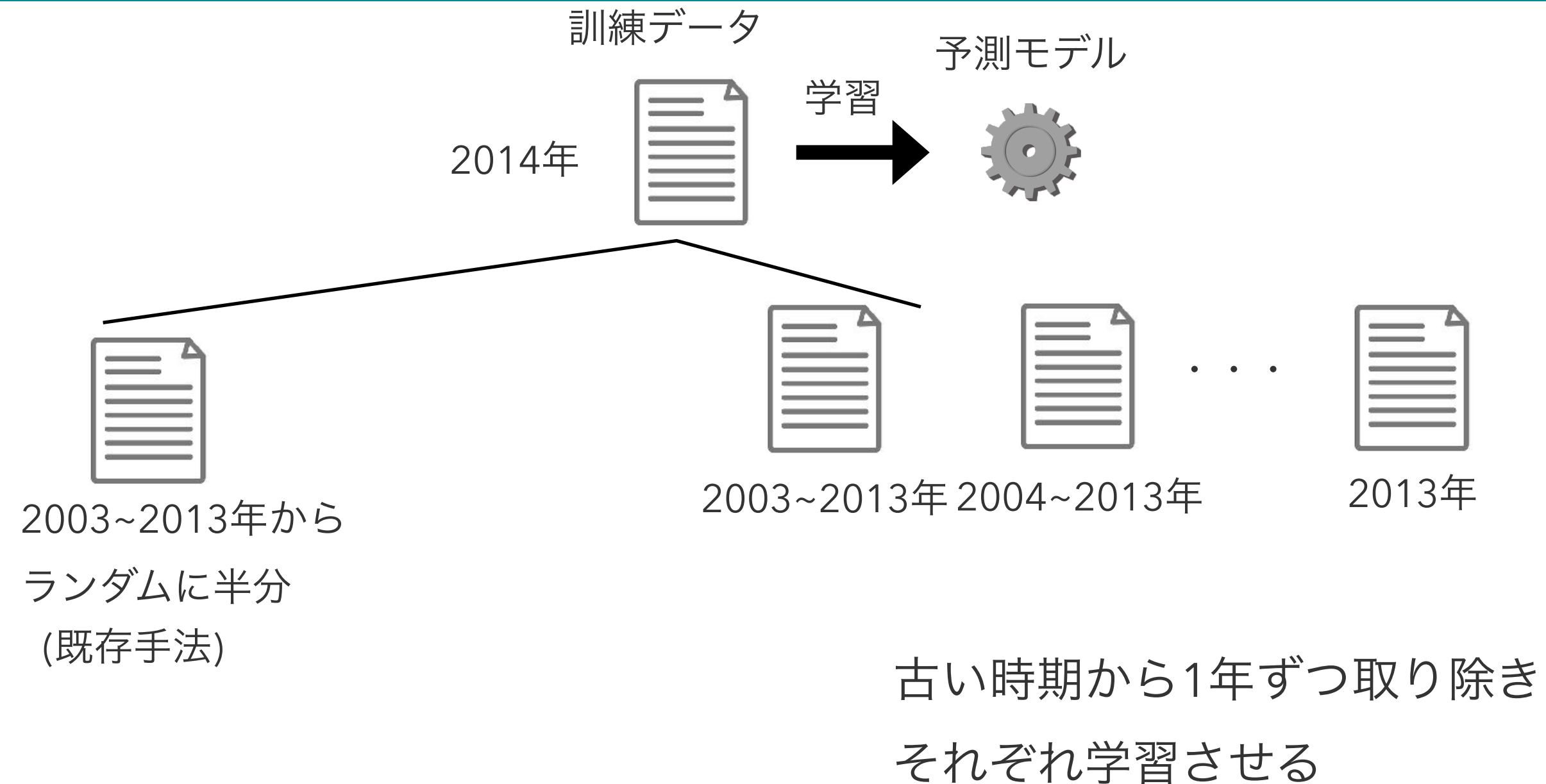
- ROC曲線の AUC (Area Under an ROC Curve)で評価



ROC曲線は右にいくほど下がることはない

偽陽性率の値が小さい時点で、高い真陽性率を達成しているモデルほど良い

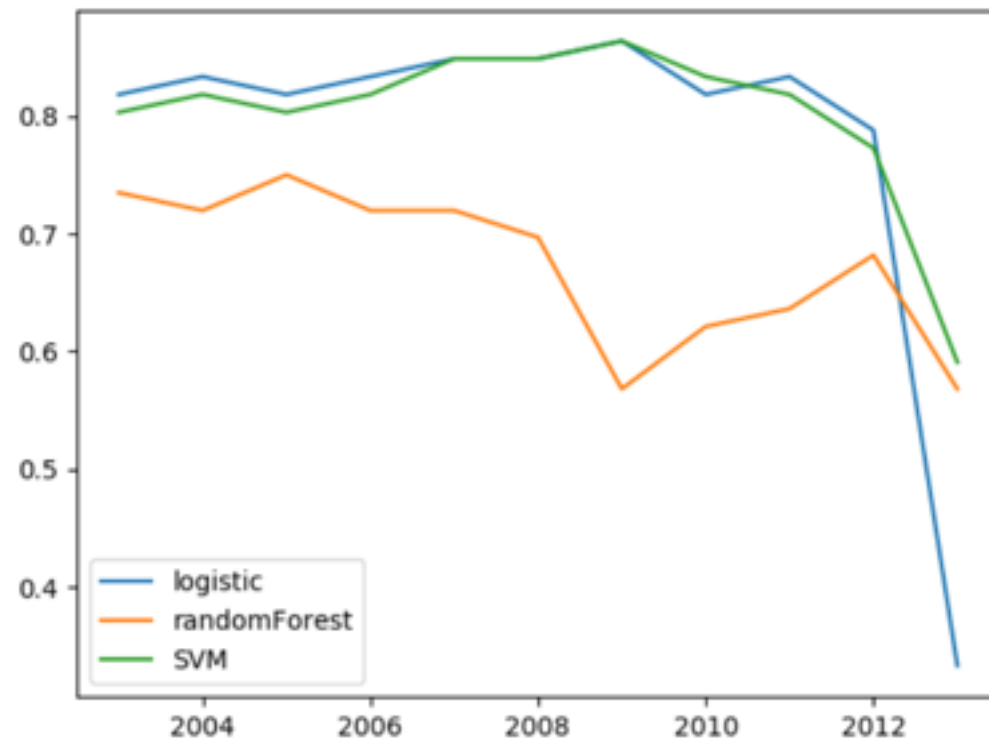
実験の手順



2014年だけでなく2013,2012・・・と予測

Eclipseの実験結果

2014年の開発者の予測



2011年の開発者の予測

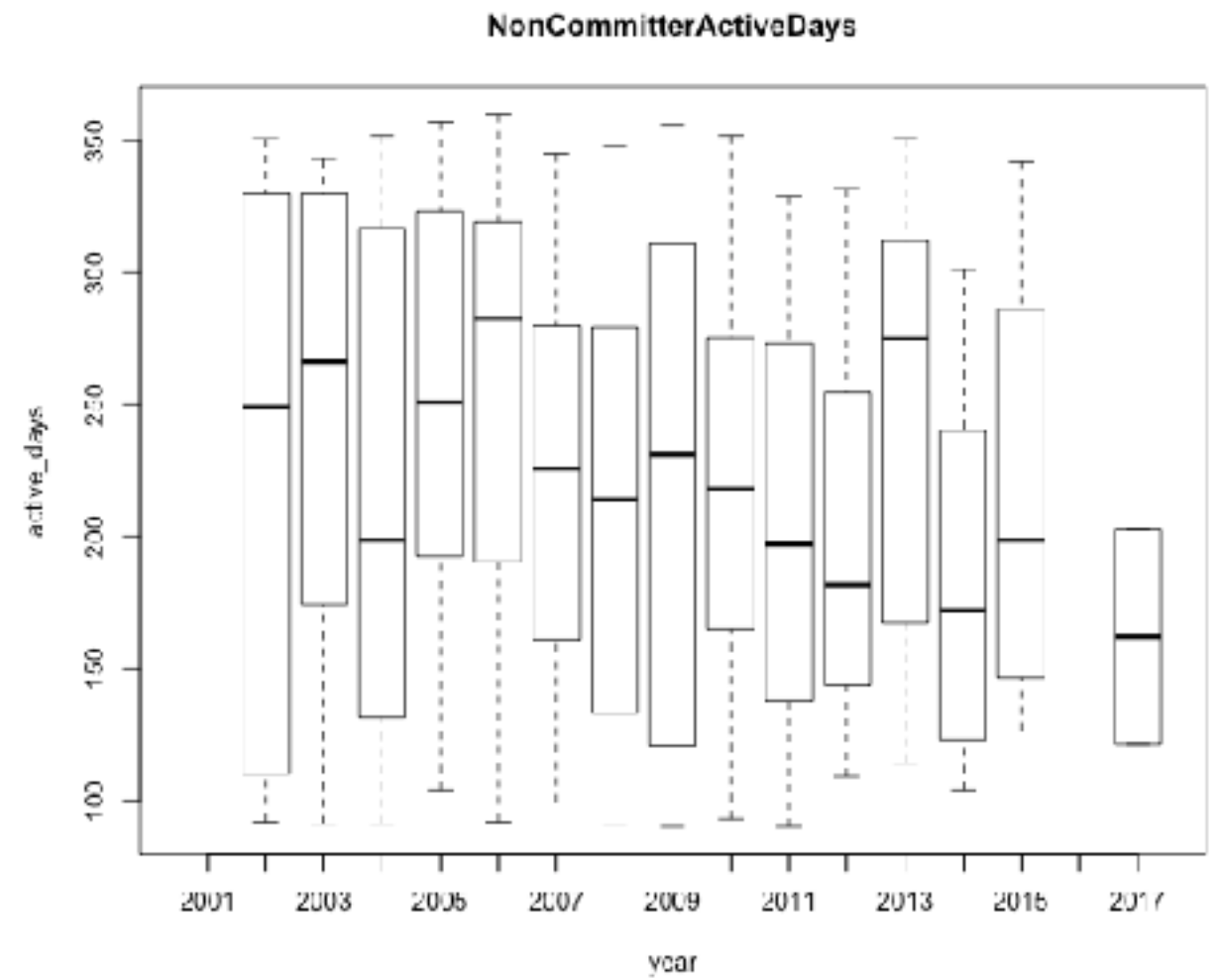
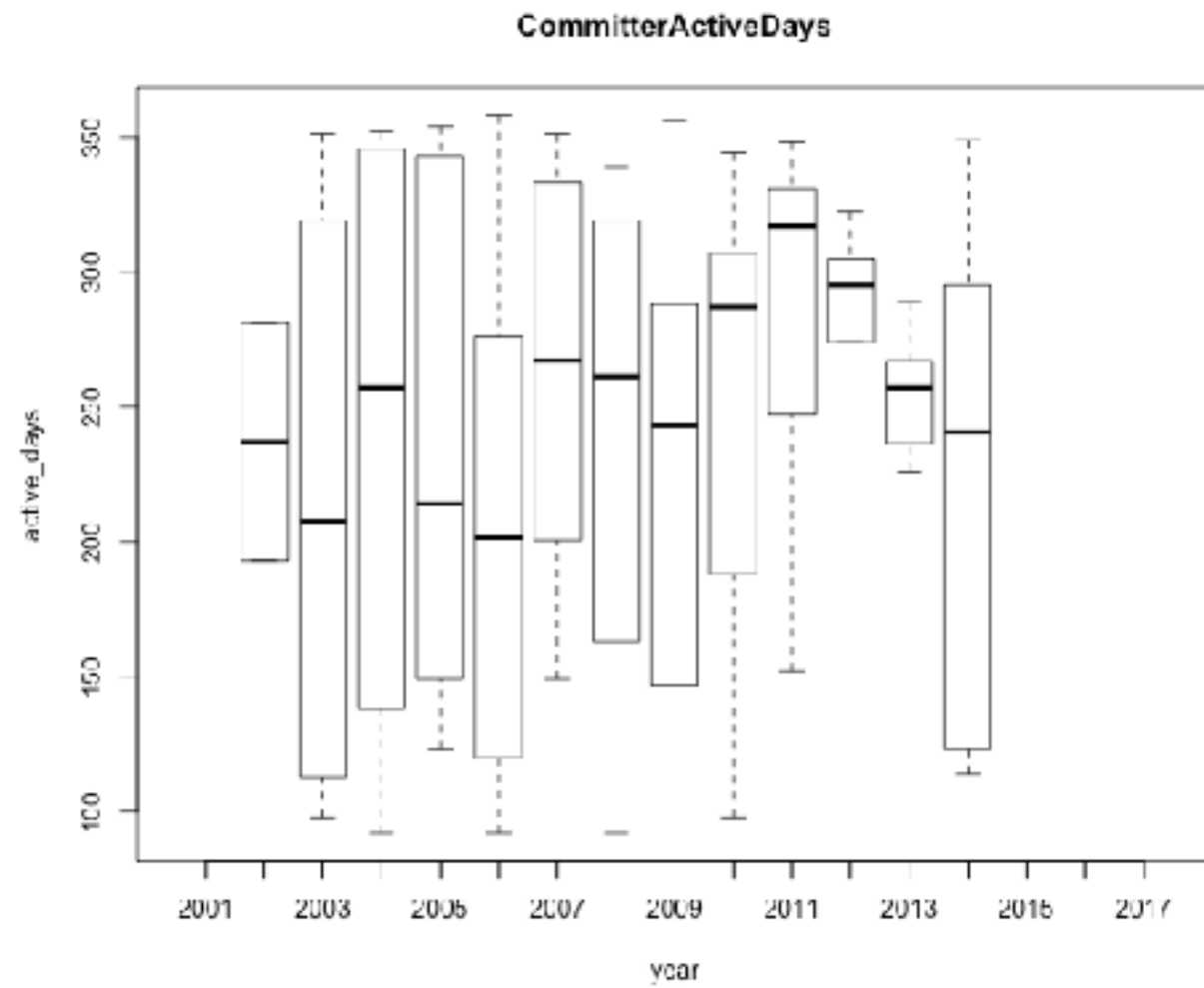


大きく 2 パターン

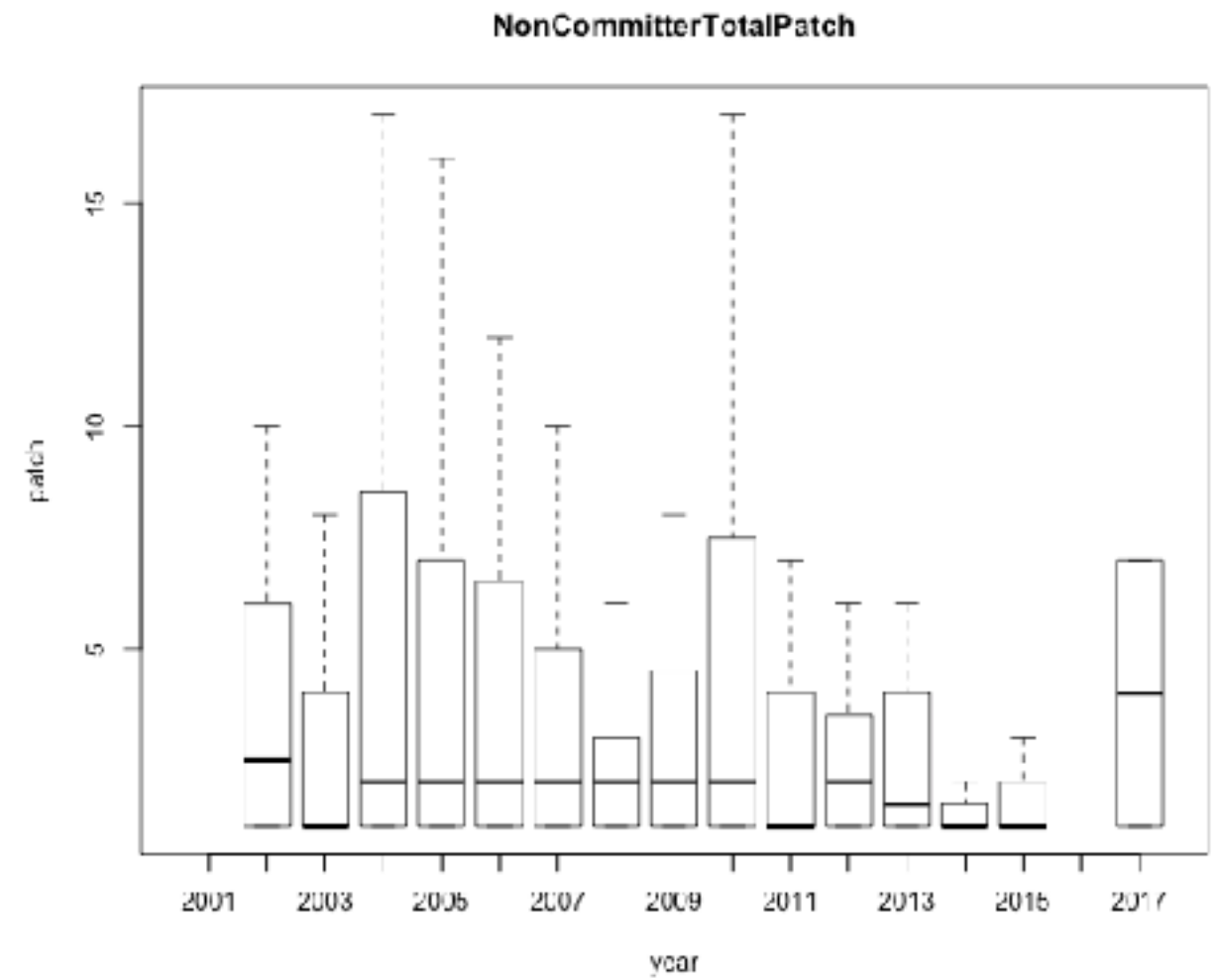
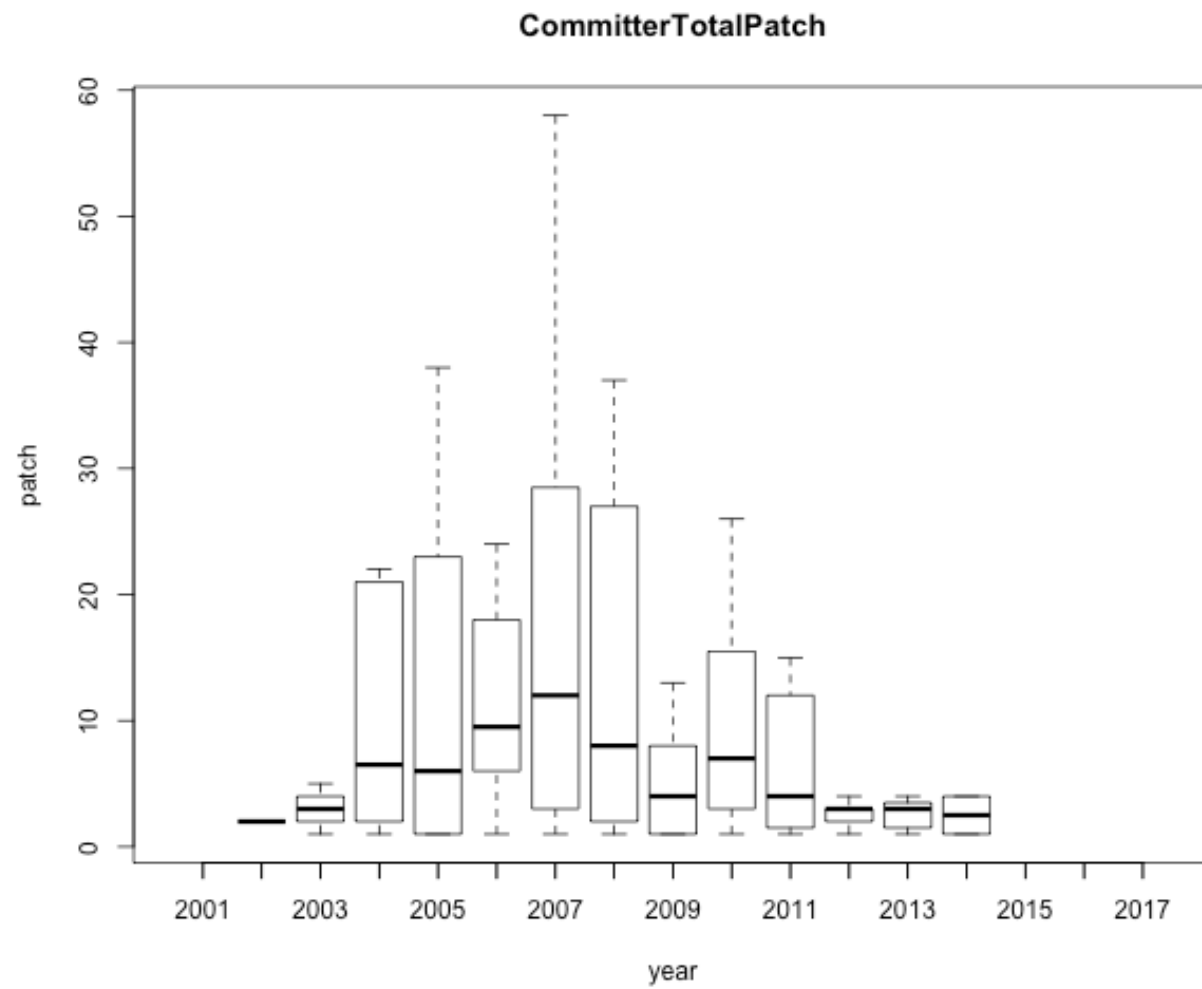
直近データでは ↓ 年数が増えていくと ↑

直近データでは ↑ 年数が増えていくと ↓

メトリクスの分布：活動期間

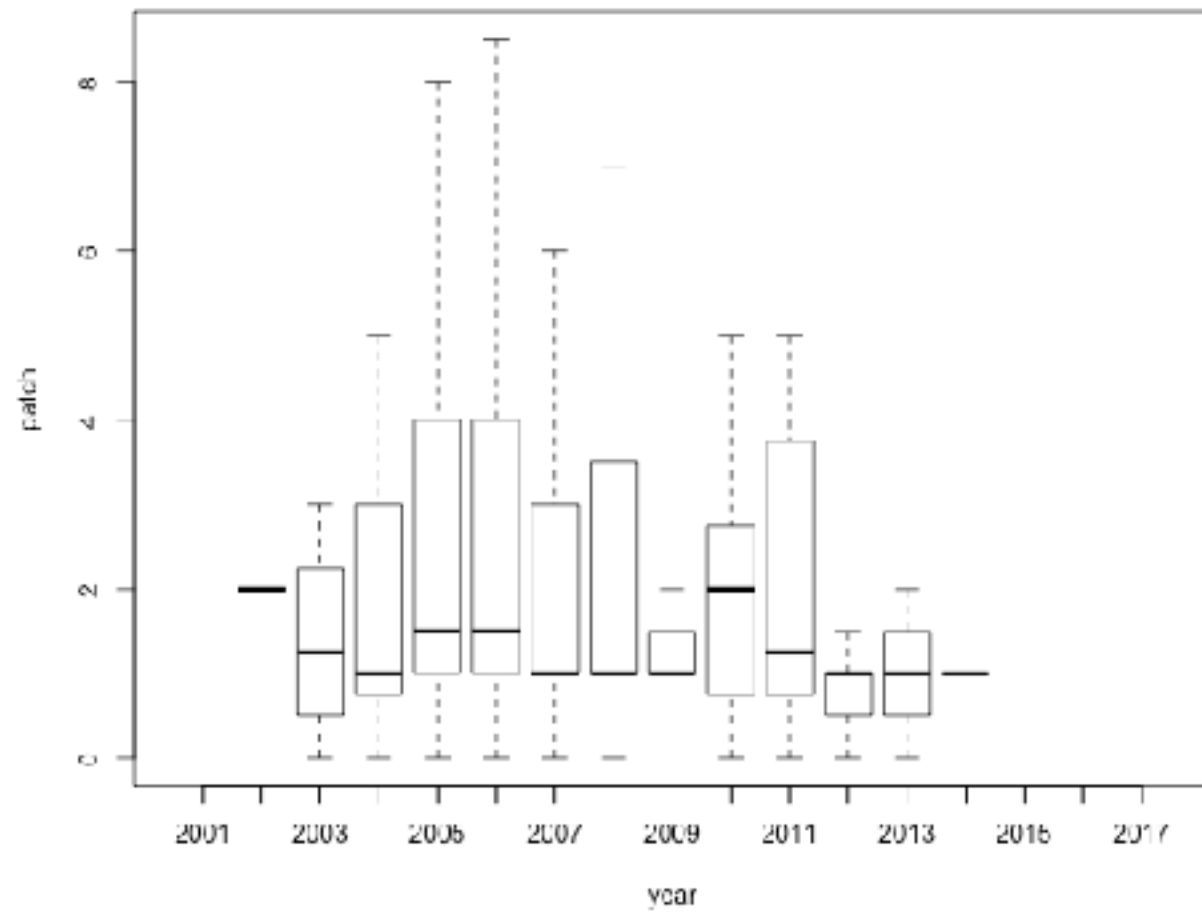


メトリクスの分布：総パッチ投稿数

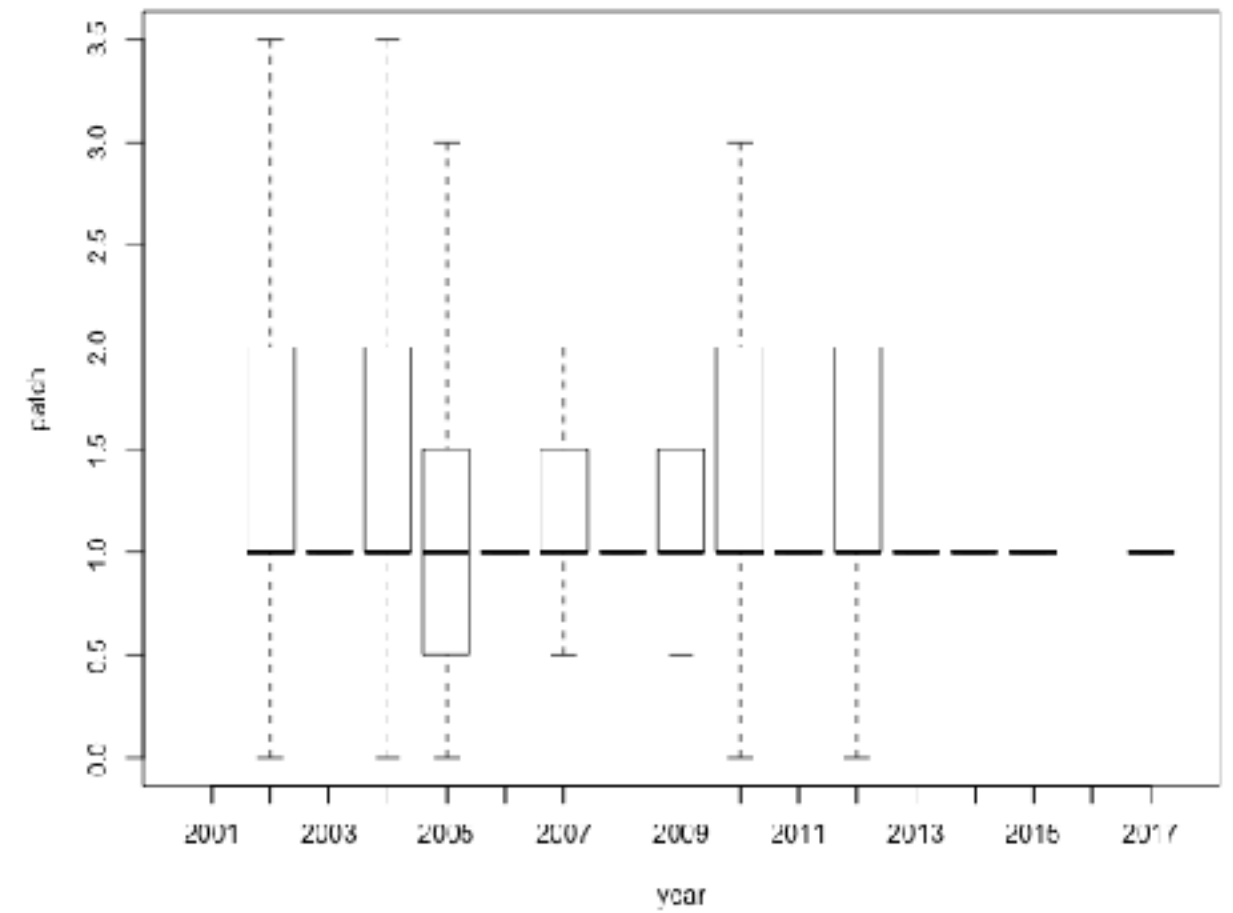


メトリクスの分布：月パッチ投稿数

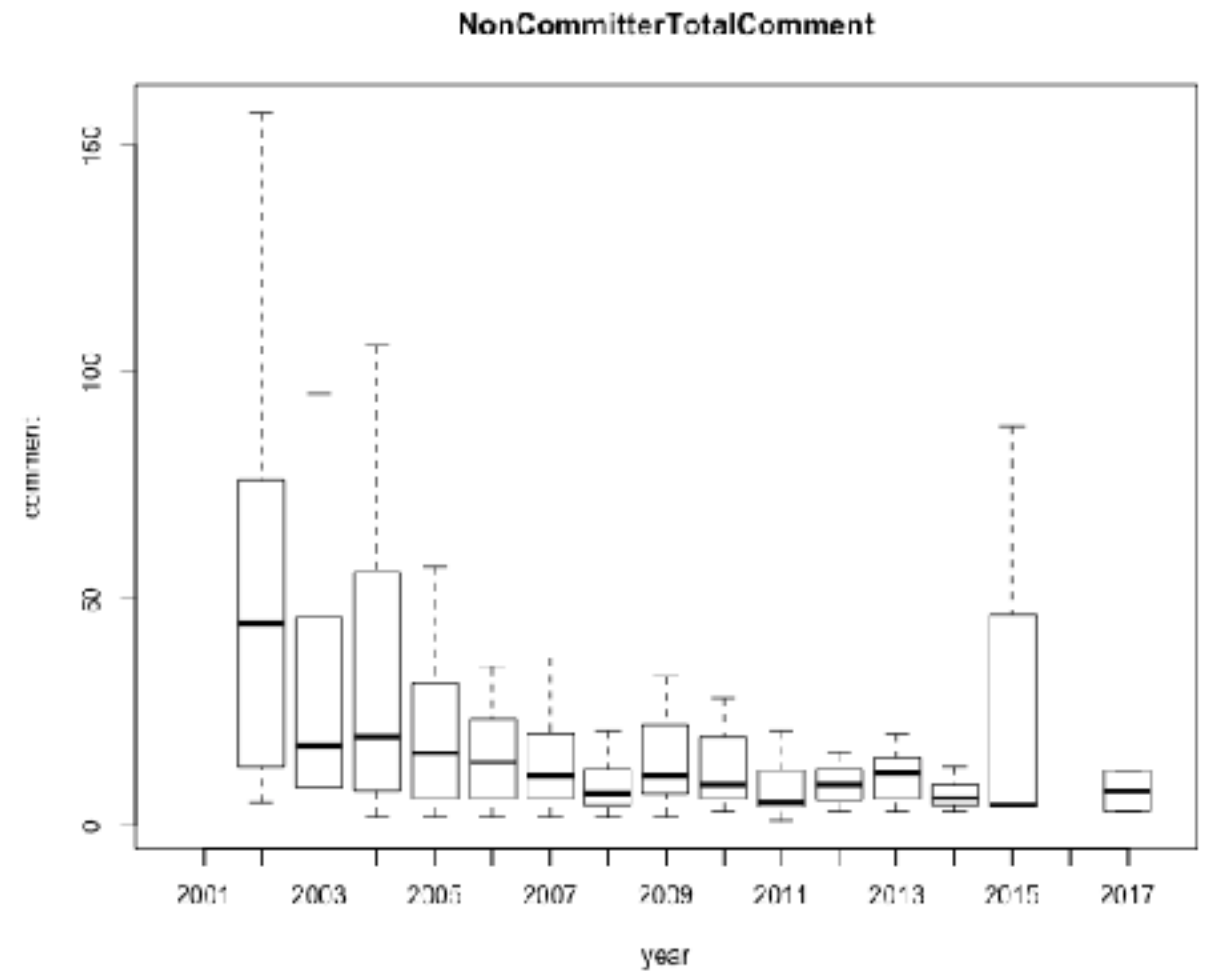
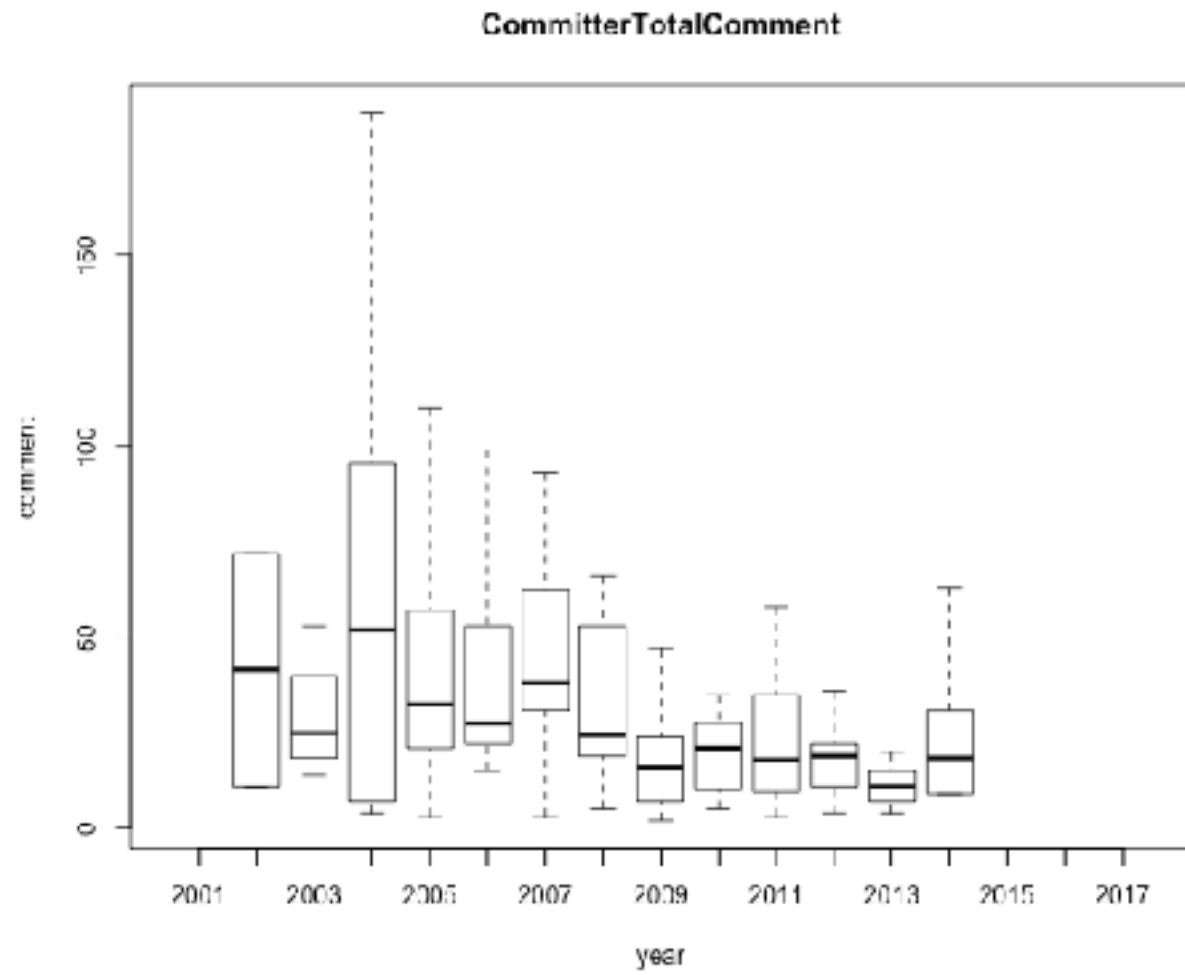
CommitterMedianPatch



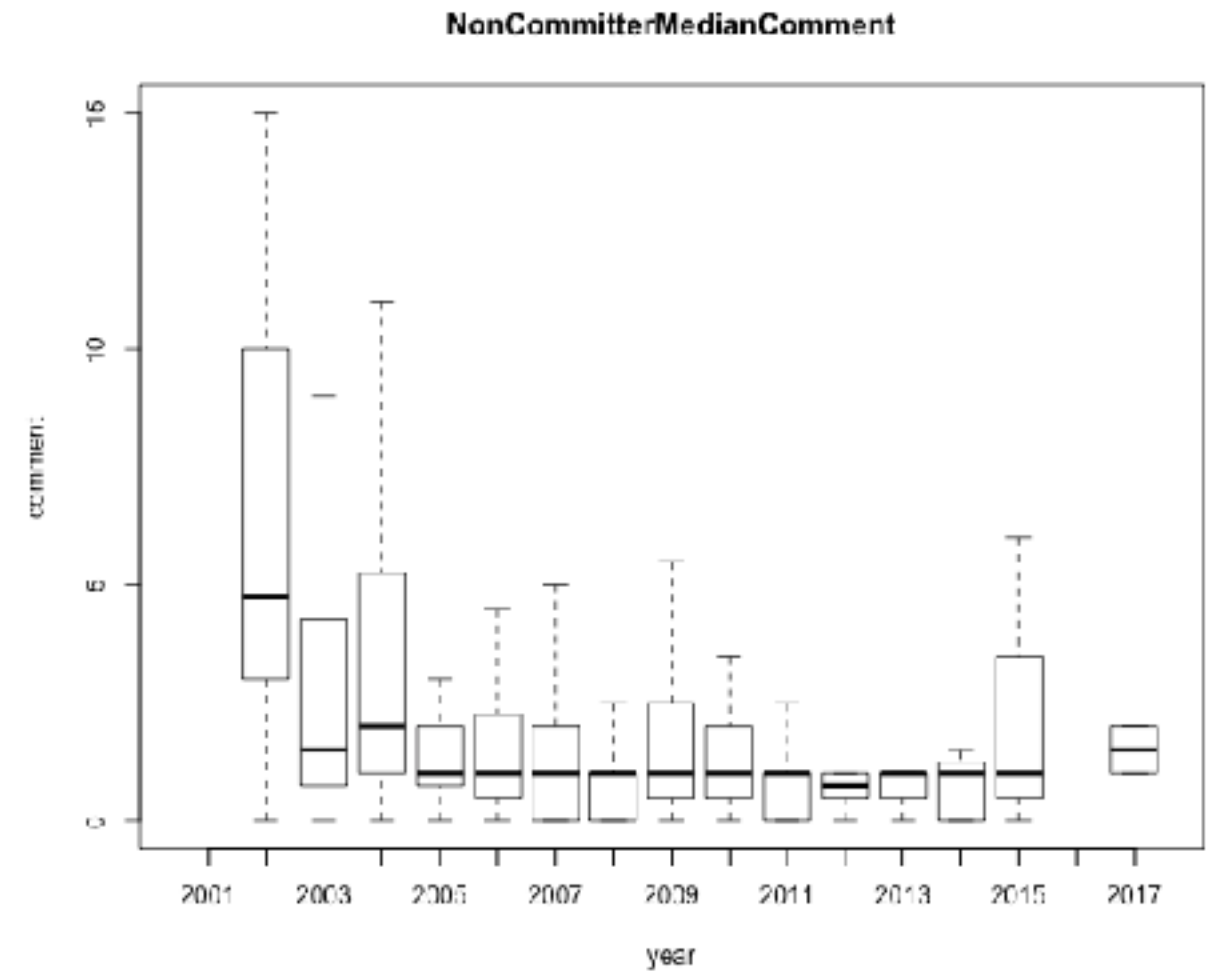
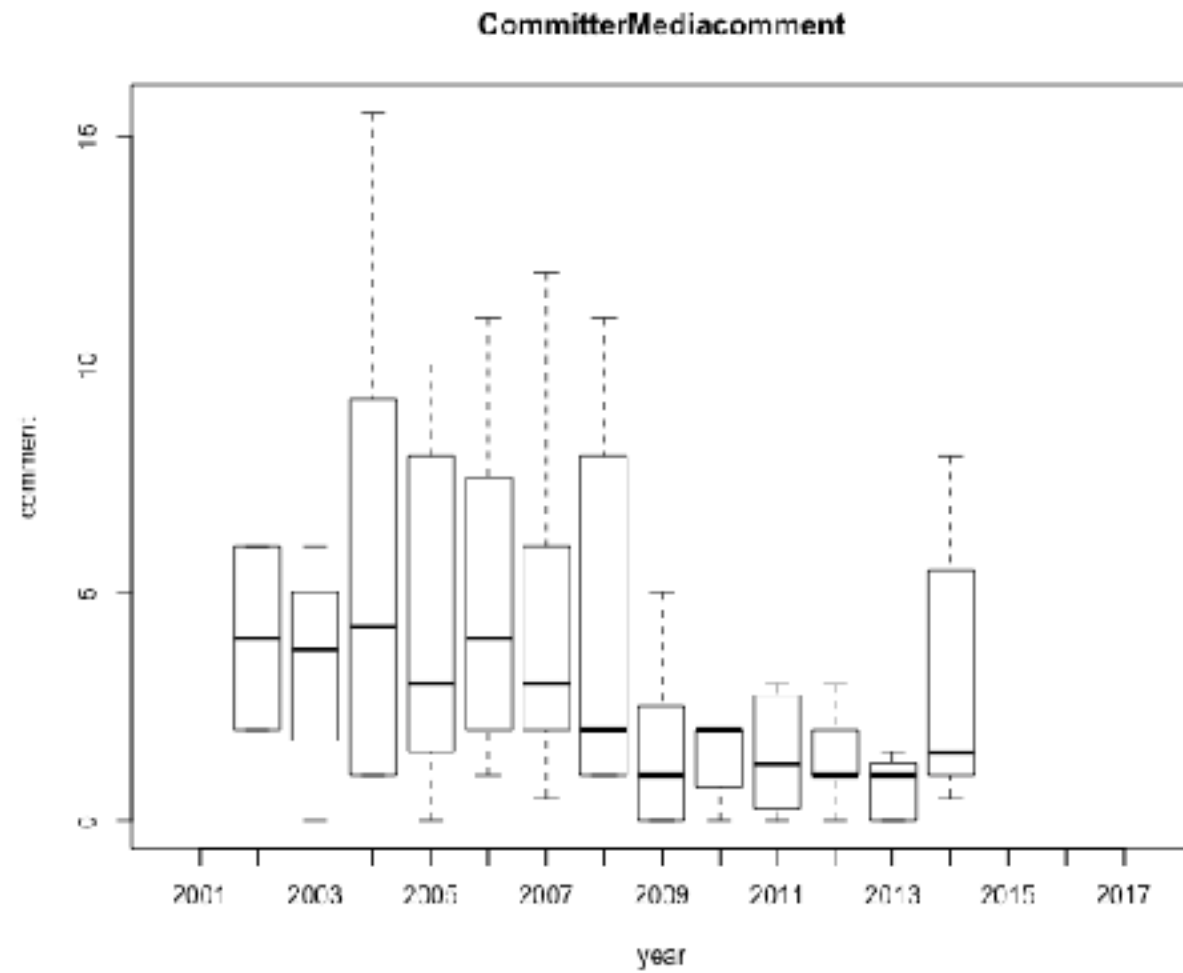
NonCommitterMedianPatch



メトリクスの分布：総コメント投稿数



メトリクスの分布：月コメント投稿数



今後

- メトリクスを見やすい形に表現する
- (実験でのデータの区切り方を変える)
- (実験プロジェクトを増やす)
- (メトリクスの追加)