

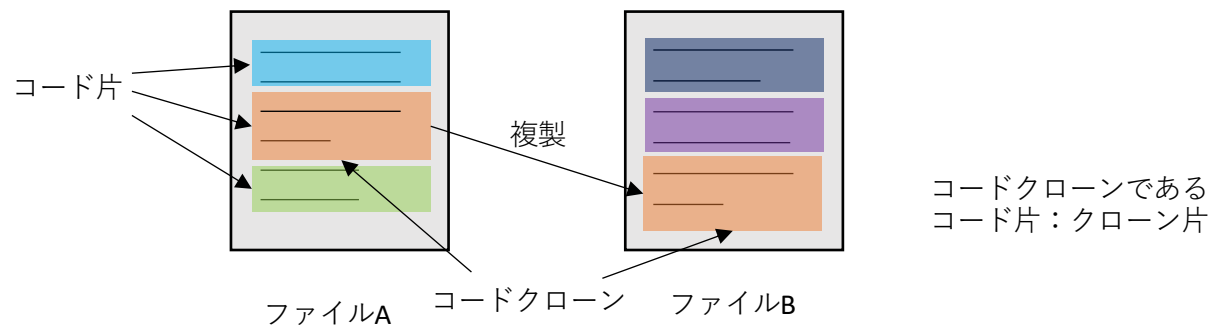
コードクローンに特化した 欠陥混入予測手法の構築

オープンソースソフトウェア工学研究室 M2

60200089 野口 耕二郎

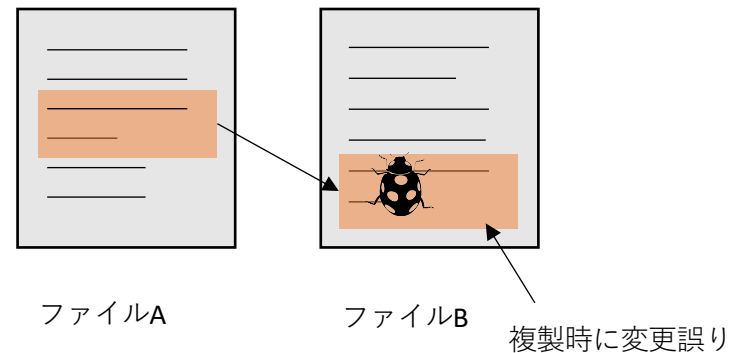
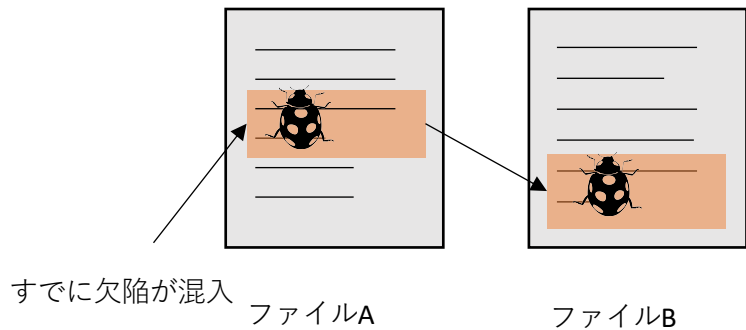
コードクローン

- 「一致」または「類似」するソースコードの断片
- 生産性を向上させるためにコードクローンの作成が頻繁に行われる
- コードクローン作成時に欠陥が複製されたり新たに混入する場合がある



コードクローン生成時に混入する欠陥

- 欠陥が混入しているコードを欠陥が混入したまま複製（左）
- 複製時に適切な変更を加えなかったことで複製先のコードに欠陥が混入（右）



コードクローン作成時の欠陥混入の問題点

- コードクローンの作成時の欠陥の拡散により欠陥の修正難易度が高くなる
 - 修正すべき箇所が増加する
 - どのクローン片に欠陥が混入してるかを見極める必要がある

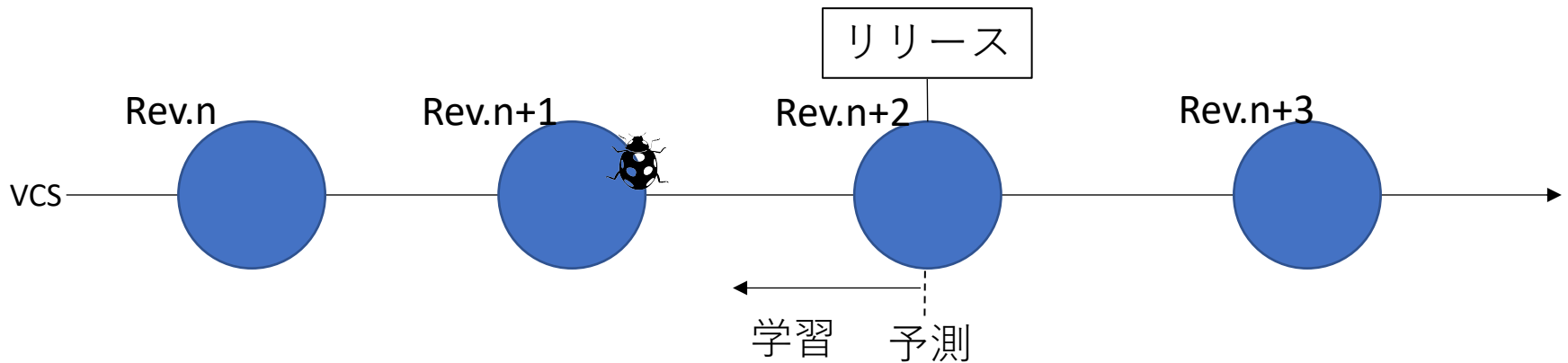
先行研究

- **Just-In-Time欠陥混入予測手法[1][2]**
 - コミット時に欠陥が混入しているかどうかを予測
 - 従来手法に比べて欠陥混入箇所の特定範囲が細かい

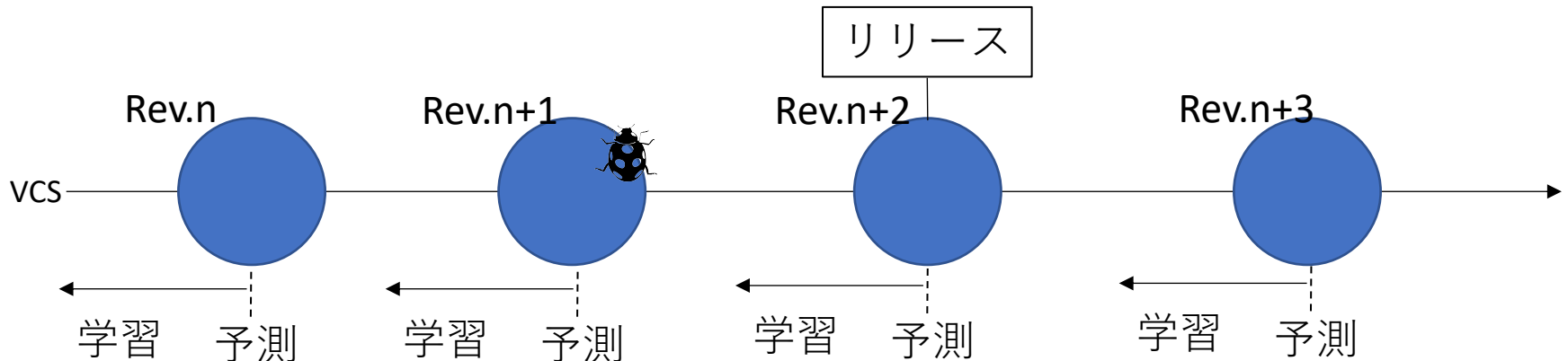
欠陥混入クロンの予測に特化した
手法は構築されていない

Just-In-Time欠陥混入予測手法

- 従来の欠陥混入予測手法



- JIT欠陥混入予測手法



既存手法の問題点

- コードクローンを考慮して手法が構築されていない
 - 欠陥混入クローンの特徴を反映するメトリクスが存在すると仮定する
 - 既存手法では欠陥混入クローンの特徴を反映するメトリクスを考慮して構築されていない
 - 一部の欠陥混入クローンを検出できない可能性がある

目的

1. 既存の欠陥予測手法で欠陥混入クローンを予測できるかを明らかにする
2. コードクローンに特化した欠陥混入予測手法を構築する

アプローチ

1. 既存手法を用いた欠陥混入クローンの予測精度の測定
 - 再現率, 適合率の測定
2. 1で測定した予測精度に応じて分析を実施
 - 再現率, 適合率の大小に応じて追加すべき説明変数の調査や誤検知の原因調査を行う
3. 分析結果を用いる事で欠陥混入クローンに特化したモデルを構築できたかを評価
 - 再現率, 適合率が上昇したかどうかを確かめる

先行研究で用いられている予測モデル

- ロジスティック回帰を使用

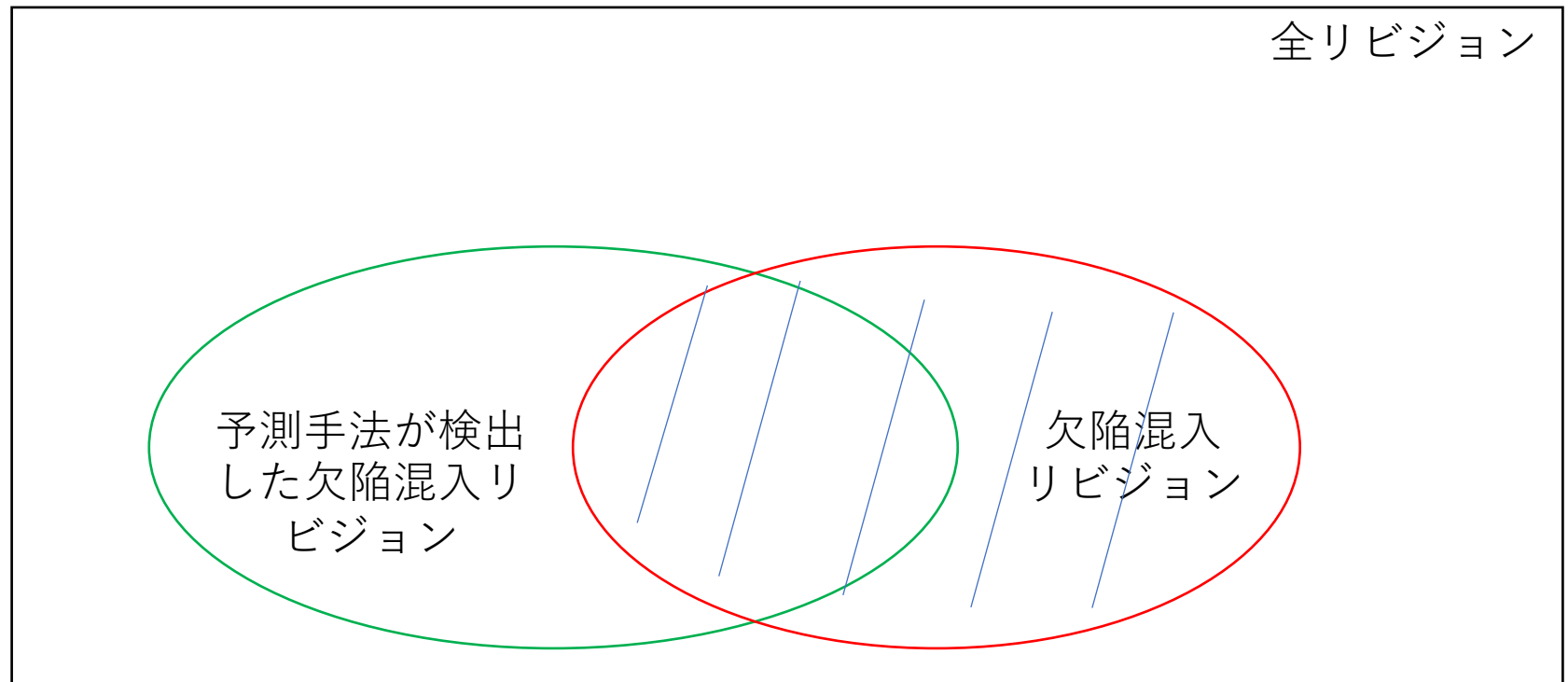
- $$p = \frac{1}{1 + \exp\{-(a_1 \cdot x_1 + a_2 \cdot x_2 + \dots + a_{17} \cdot x_{17} + b)\}}$$
 - p : 目的変数に設定した事象が発生する確率
 - $a_1, a_2, \dots, a_{17}$: 偏回帰係数
 - $x_1, x_2, \dots, x_{17}$: 説明変数
 - b : 定数

- 5種類, 17項目のメトリクスを説明変数とし, 欠陥の有無を目的変数とする

- 変更規模
 - 変更の拡散度合い
 - 歴史
 - 開発者の経験
 - コードレビュー

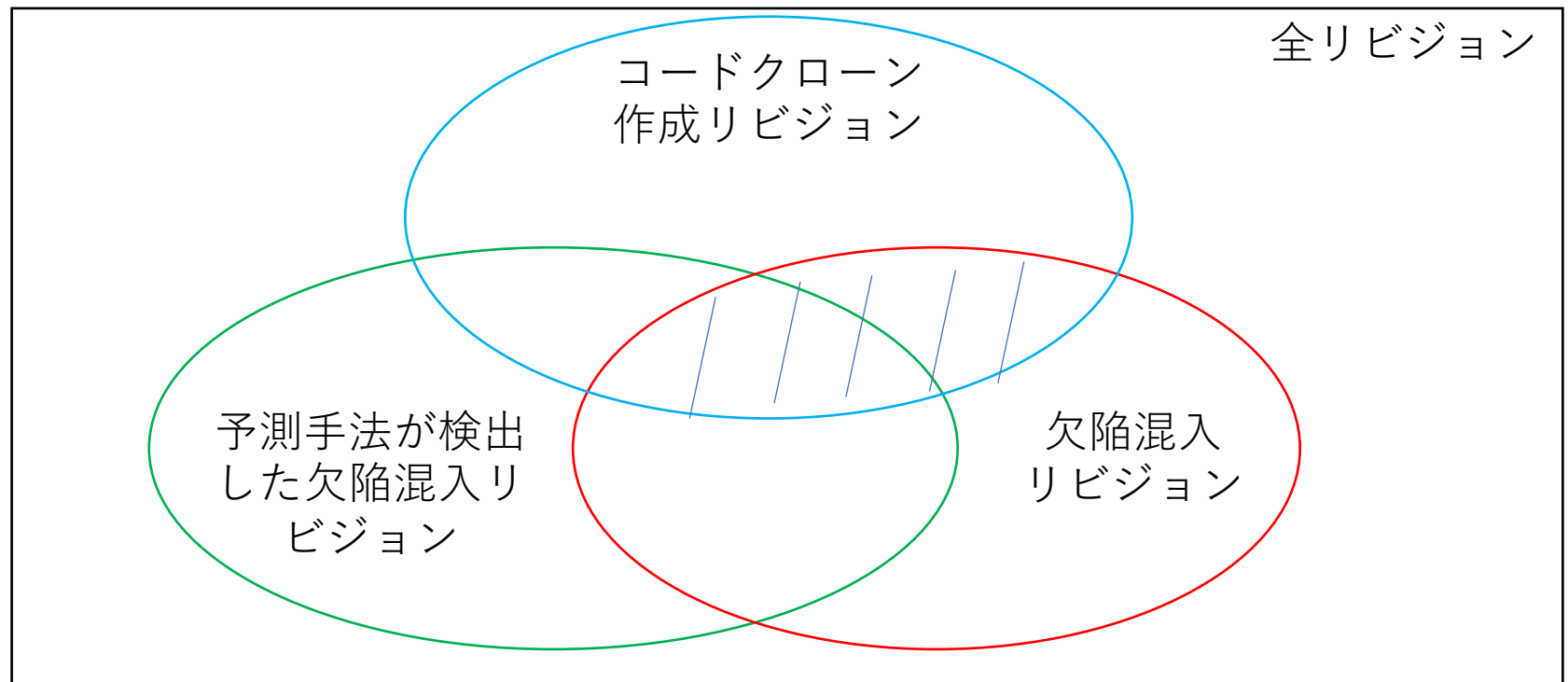
先行研究の予測対象

- 欠陥混入リビジョンを予測対象としている



本研究の予測対象

- 欠陥混入クローン作成リビジョンを予測対象としている

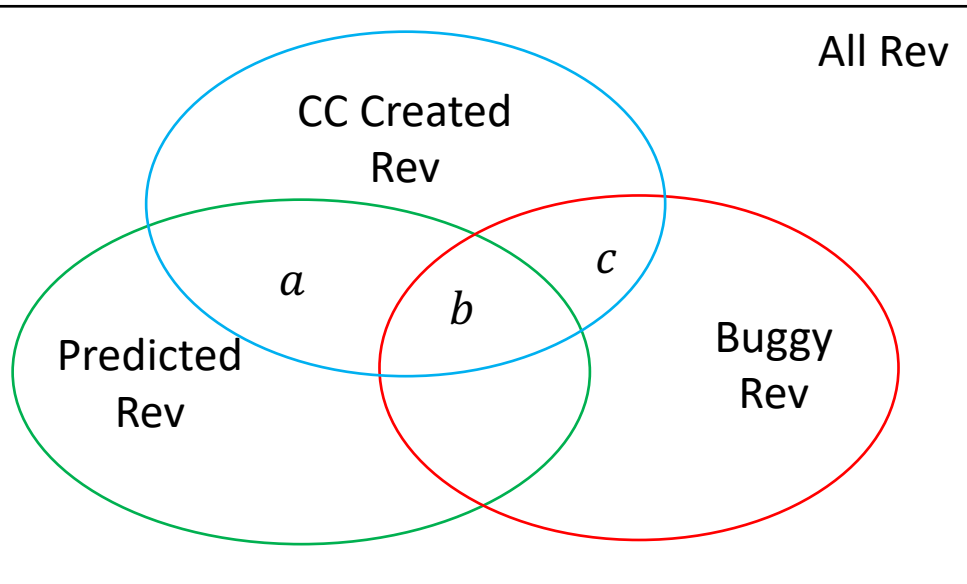


既存手法を用いた予測精度の算出

- a , b , c をもとに再現率, 適合率を算出する

- 再現率 : $\frac{|b|}{|b \cup c|}$

- 適合率 : $\frac{|b|}{|b \cup a|}$



$b \cup c$: 欠陥混入クローン作成リビジョン

b : 既存手法で検出できた欠陥混入クローン作成リビジョン

c : 既存手法で検出できなかった欠陥混入クローン作成リビジョン

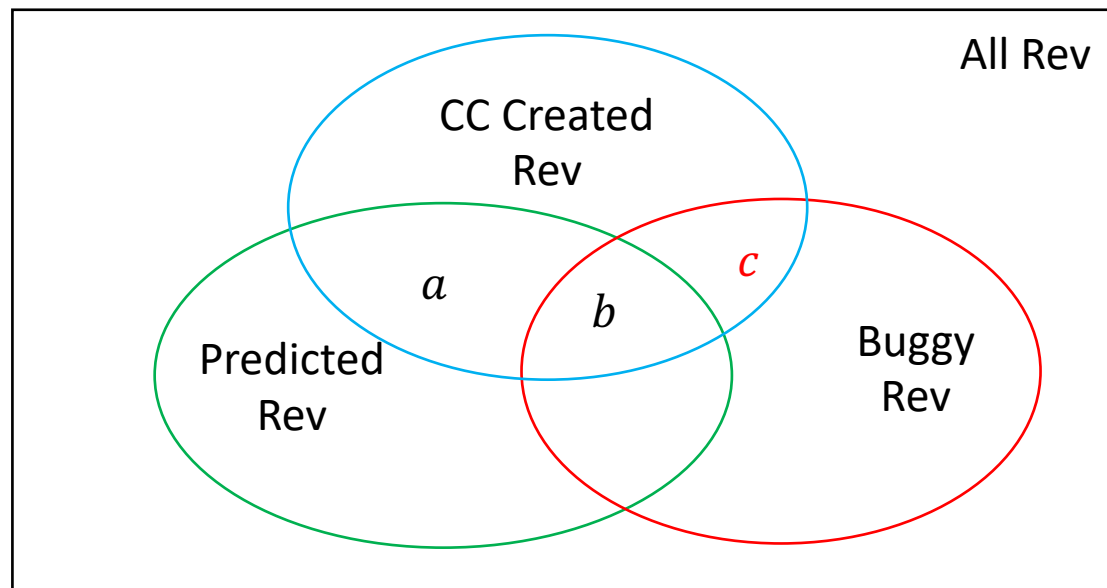
a : 既存手法が誤って検出したクローン生成リビジョン

予測精度に基づく分析

- 欠陥混入リビジョンのみ予測した場合に比べて,
 - 再現率が小さい：欠陥混入クローンの検出に特化した説明変数を調査
 - 適合率が小さい：誤検知の原因の調査
 - どちらも同程度あるいは大きい：既存手法で用いている説明変数の中で欠陥混入クローンの検出に向いているものを調査

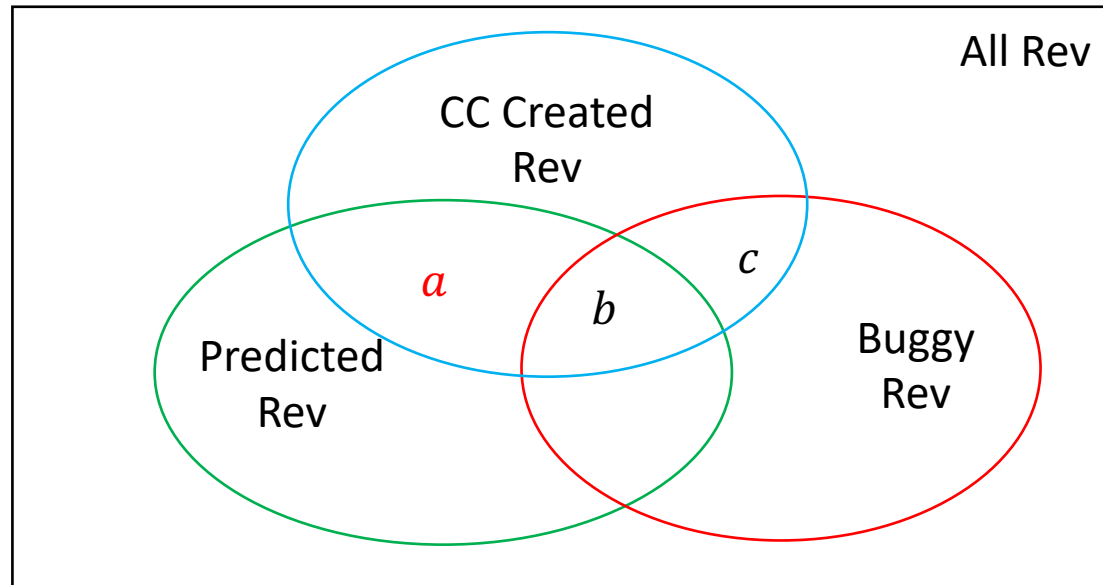
再現率が小さい

- 既存手法で検出できなかった c の部分の欠陥混入クローンの特徴を分析
 - c で作成された欠陥混入クローンの内容から特徴を分析



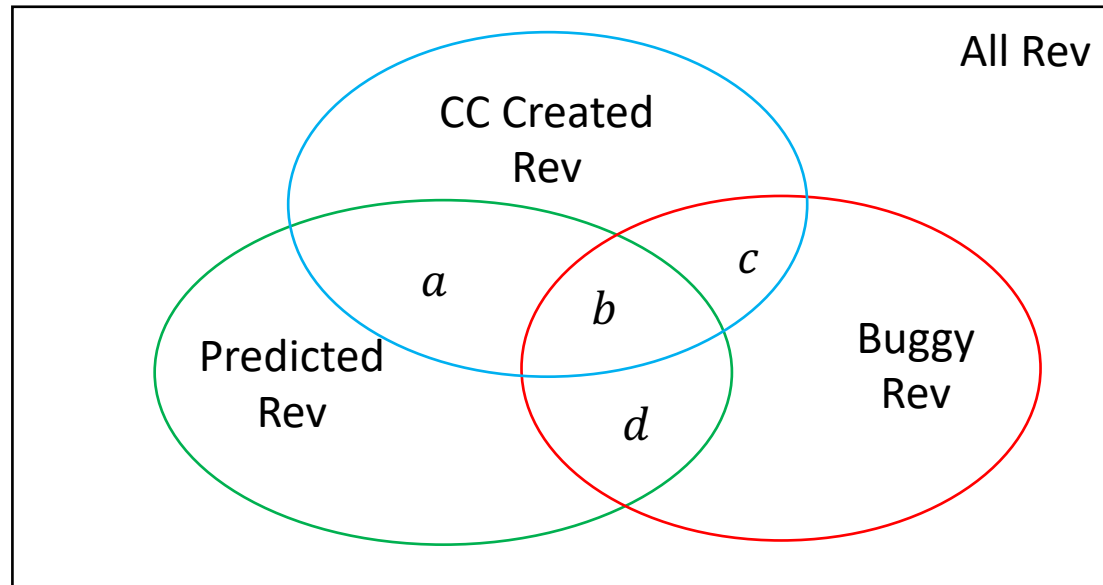
適合率が小さい

- 誤って検出した a が増加した原因を調査
 - 不要な説明変数の調査



どちらも同程度あるいは大きい

- 欠陥混入クローンを検出するのに寄与している既存メトリクス調査
 - b と d のメトリクスの値の違いを見る



今後の予定

- openstackサブプロジェクト 5 つで欠陥混入クロンの抽出
- 既存手法を用いて欠陥混入クロンを予測した際の精度を計測
- 計測結果に基づいた分析
- 分析結果を基に予測手法構築

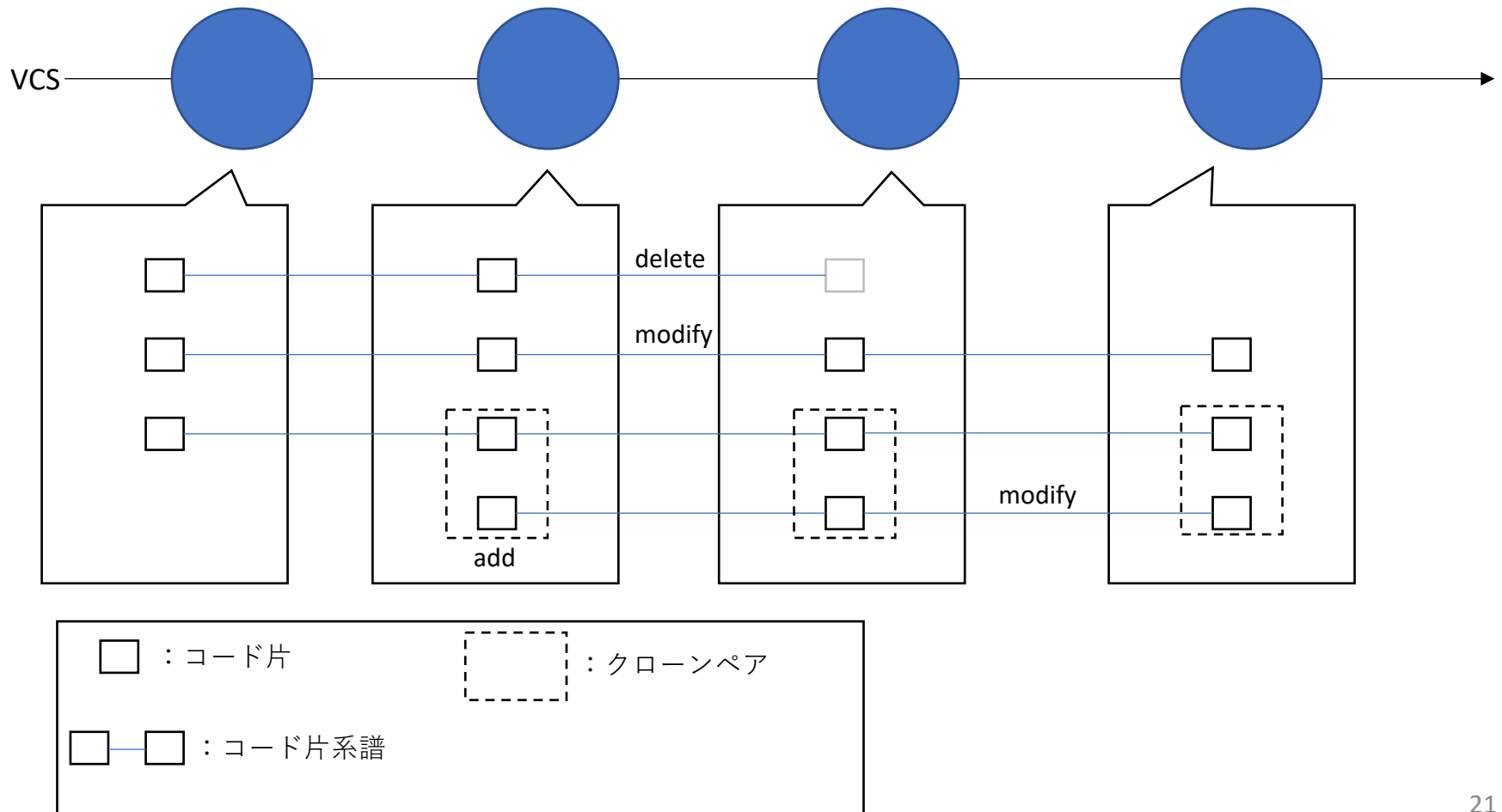
補足資料

欠陥混入クローンの特定

1. リビジョンの親子関係の抽出
2. 変更行数の特定
3. コード片系譜の抽出
4. クローン系譜の抽出
5. 欠陥混入リビジョンで追加されたクローン系譜の特定

欠陥混入クローンの特徴

- クローン系譜の抽出



欠陥混入クローンの特定

