学習データとテストデータ

手持ちのデータ
（100個）

学習に使う

テスト・評価に使う

学習データ＝テストデータ

Hold-out

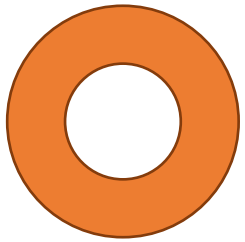
手持ちのデータ
（100個）

学習データ

テストデータ

学習データに対する性能
学習誤差

テストデータに対する性能
テスト誤差



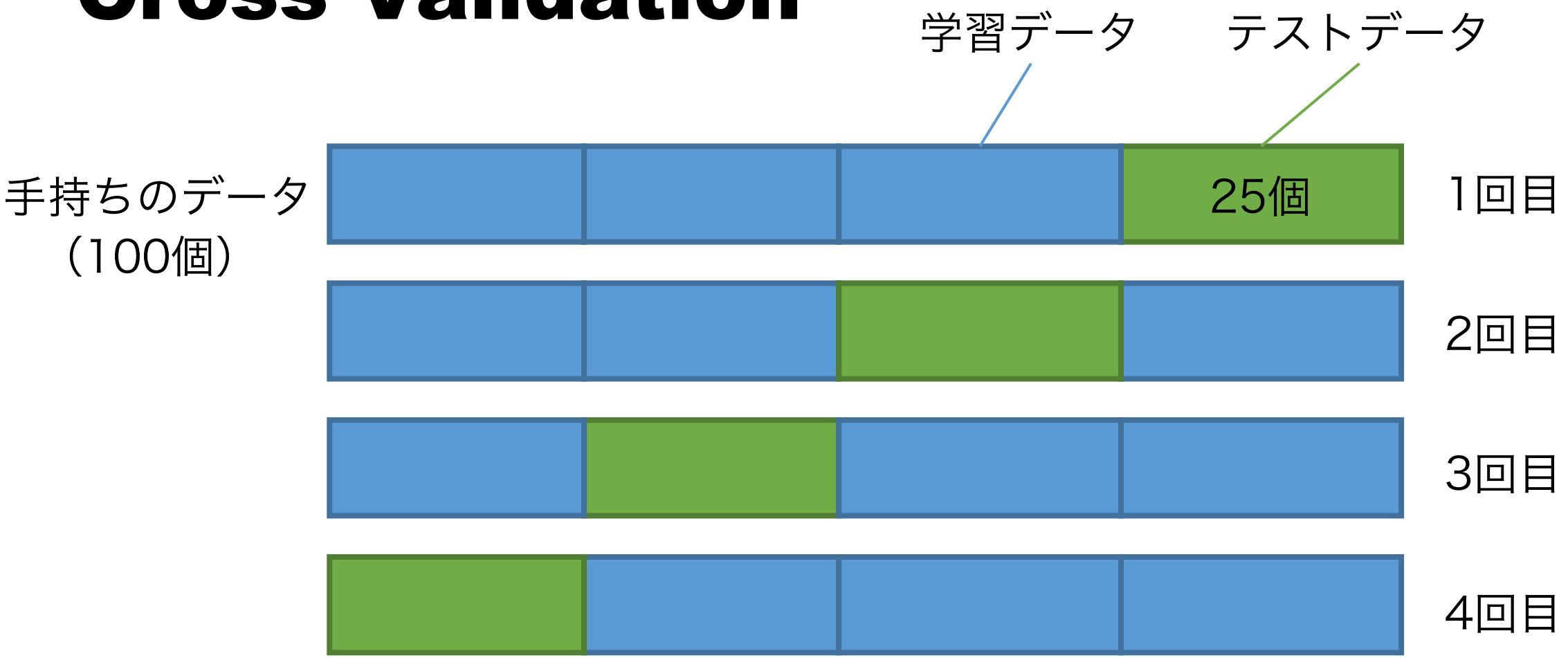
学習データ≠テストデータ

（50:50, 80:20,　etc）

汎化性能
汎化誤差

例：4-fold cross validation

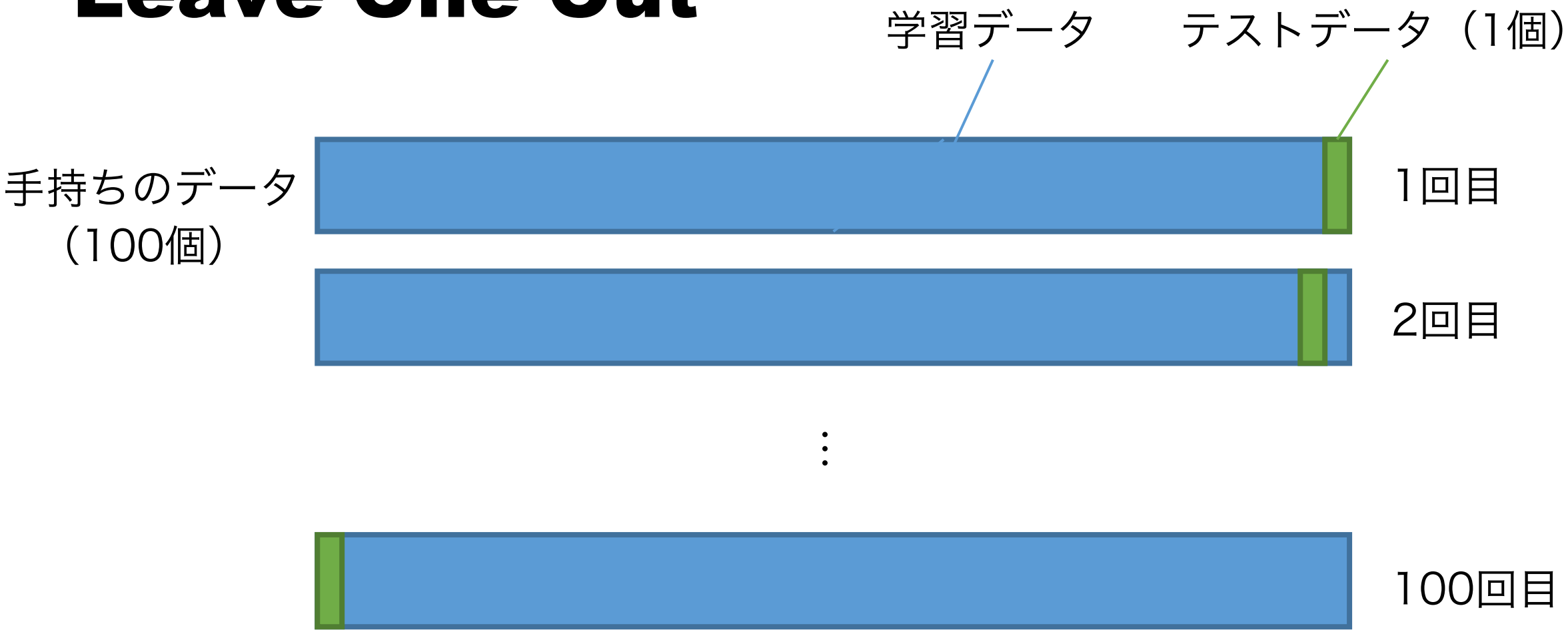
Cross Validation



一般的にはK-fold CV（K=10, 5, 3, …）

CV，交差検定，交差検証，交差確認

Leave One Out

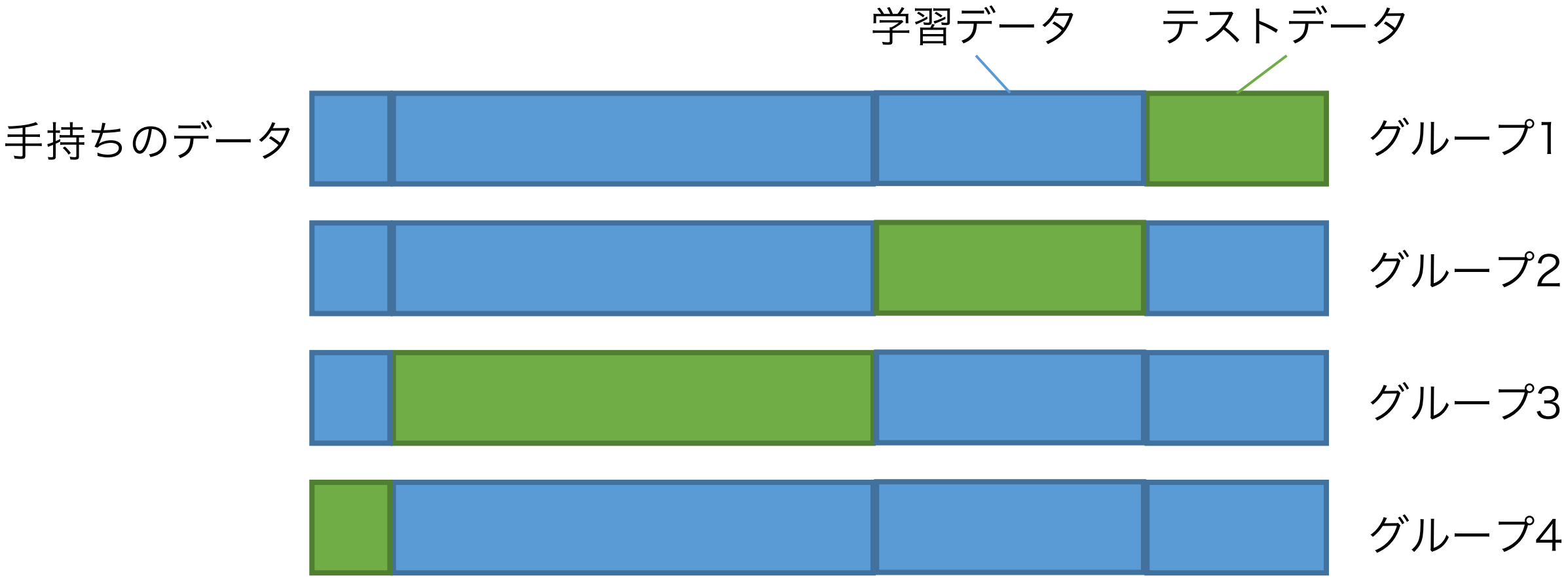


Leave-one-out = N-fold CV

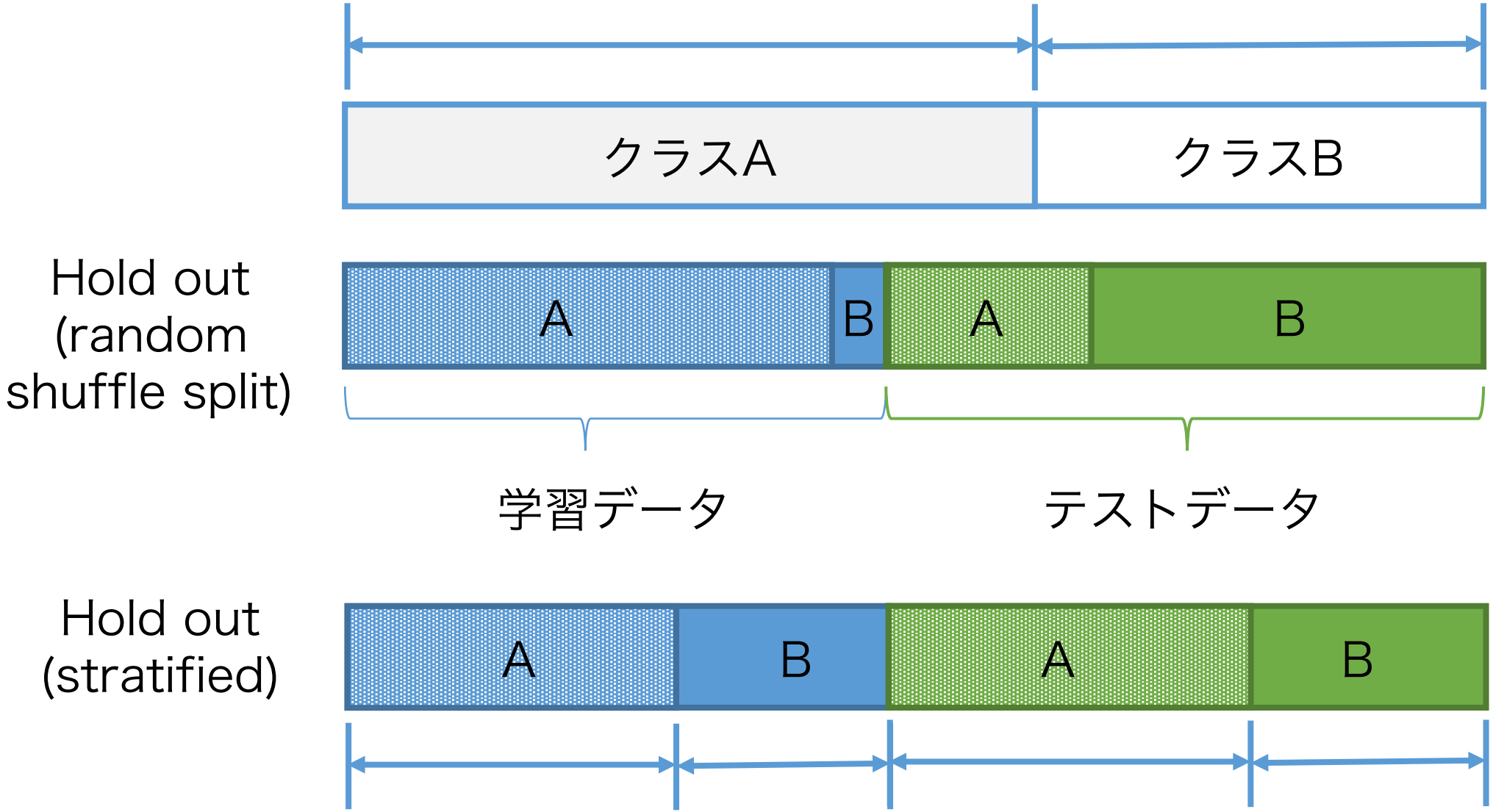
LOO, LOOCV
（一つ抜き法，ジャックナイフ法）

例：4グループ

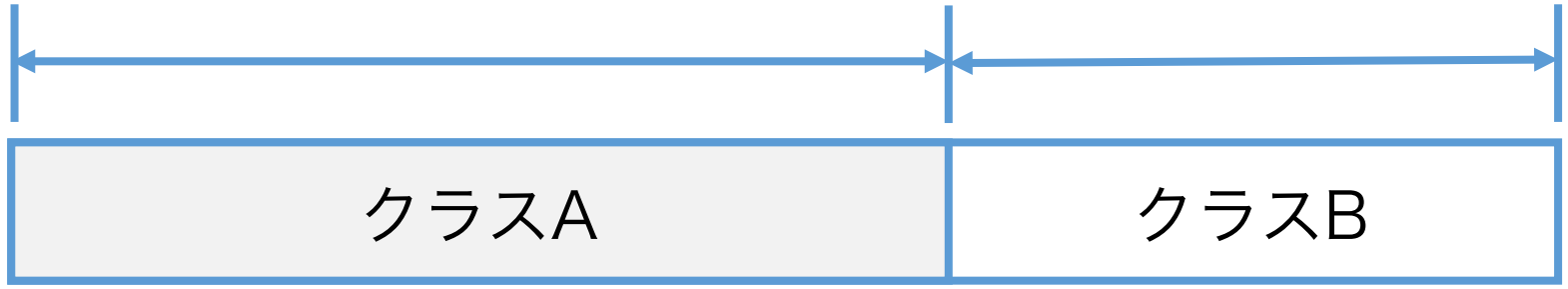
Leave one group out



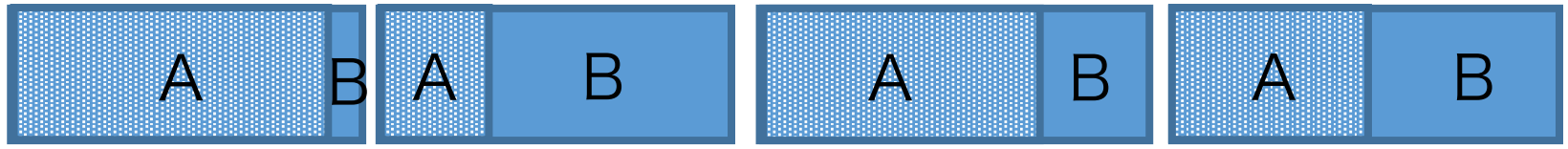
Stratified（層化）



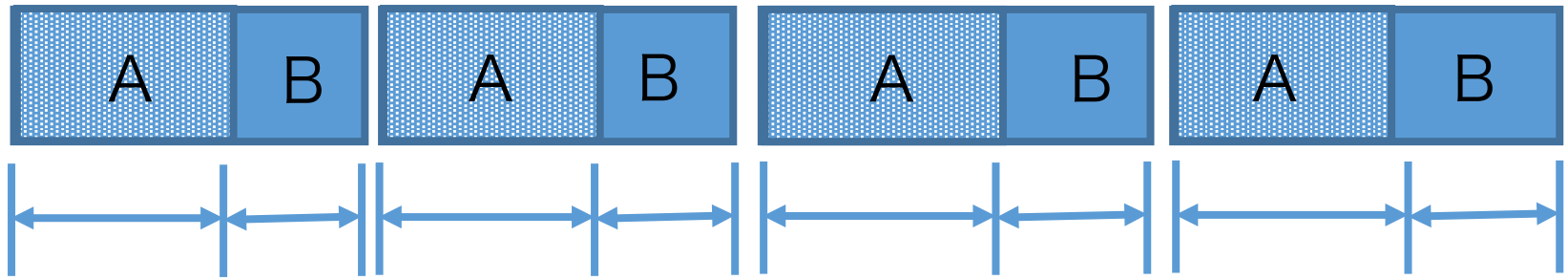
Stratified（層化）



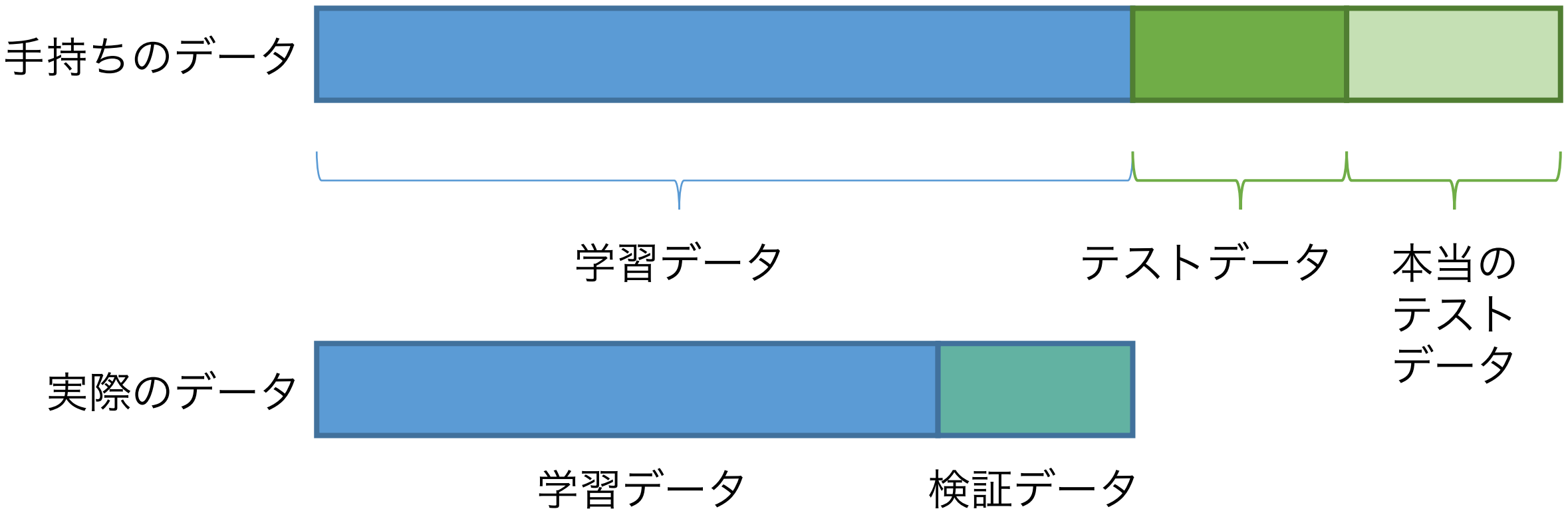
4-fold
(random
shuffle split)



4-fold
(stratified)



学習データ・検証データ・テストデータ



（Hold-outの例．CVでも可）

どのくらいデータがあればよい？

- 学習サンプル数＜10
  - 本当に機械学習が必要？
- 学習サンプル数～100
  - できないことはないが，増やす努力を
  - 性能は悪い
  - LOOCVが可能
- 学習サンプル数～1,000
  - まともな性
  - 10-fold CVで十分
- 学習サンプル数～10,000
  - 良い性能が期待できる
  - K-fold CV, K < 10
  - 計算リソース重要
- 学習サンプル数～100,000
  - 実応用
  - Hold-out以外はムリ
  - かなり工夫が必要
- 学習サンプル数＞100,000
  - 最先端

学習・テストの分割方法はどれがよい？

- 規格，基準などがある
  - コンテストなど
  - データセットに付属
  - それに従う
- 学習サンプル数が少ない
  - 学習サンプル数が多いLOO
- 学習サンプル数がそこそこ
  - 数百～数千
  - 10-fold CV
- 学習サンプルが膨大
  - ディープラーニングなど
  - Hold-outしかない
- Stratifiedは必須
  - 特にクラスバランスが悪い場合は重要
- 特殊な場合はone-group-out
  - 複数の被験者・患者のデータ
- 検証データ
  - 研究比較検討程度なら不要？
  - コンテストなら必須