津田塾大学総合政策学部総合政策学科 | データ政策学

イントロダクション

松本 崇斗(Takato Matsumoto)
takato.matsumoto0114@gmail.com

質問について

- sil.doで受け付けます
  https://www.sli.do/
- Event code
  # SA202
- 挙手していただいても大丈夫です

コンテンツ

1. 講義概要
2. Decision Tree（決定木）
3. Cross Validation (交差検証)
4. Feature Engineering (特徴量作成)

コンテンツ

講義概要

- Q. データサイエンティストになるにはどうしたらいい？
- A. Kaggleをやってみるのをお勧めします！

講義概要

- Kaggleでは，データサイエンティストが行う業務に必要な知識の多くを学ぶことができる
- コンペティションに参加しながら，以下のコンテンツを学んでみましょう
  - Decision Tree（決定木）
  - Cross Validation (交差検証)
  - Feature Engineering (特徴量作成)

講義概要

- Decision Tree（決定木）
  - 決定木を用いた予測モデルの構築を行う
- Cross Validation (交差検証)
  - 予測モデルの評価を行い，汎化性能に優れたモデルを構築する
- Feature Engineering (特徴量作成)
  - 特徴量作成を行い，モデルの精度を高める

講義概要 | Bike Sharing Demand Prediction

Capital Bike Shareが提供するレンタル自転車のデータセットの分析を行う

Capital Bike Share page: https://www.capitalbikeshare.com/

講義概要 | Bike Sharing Demand Prediction

- Bike Sharing Demand
  - https://www.kaggle.com/c/bike-sharing-demand
- 自転車のレンタル数を予測するコンペティション
  - 各時間ごとに何台レンタルされているかを予測
    - 天気や季節などのデータを用いる

コンテンツ

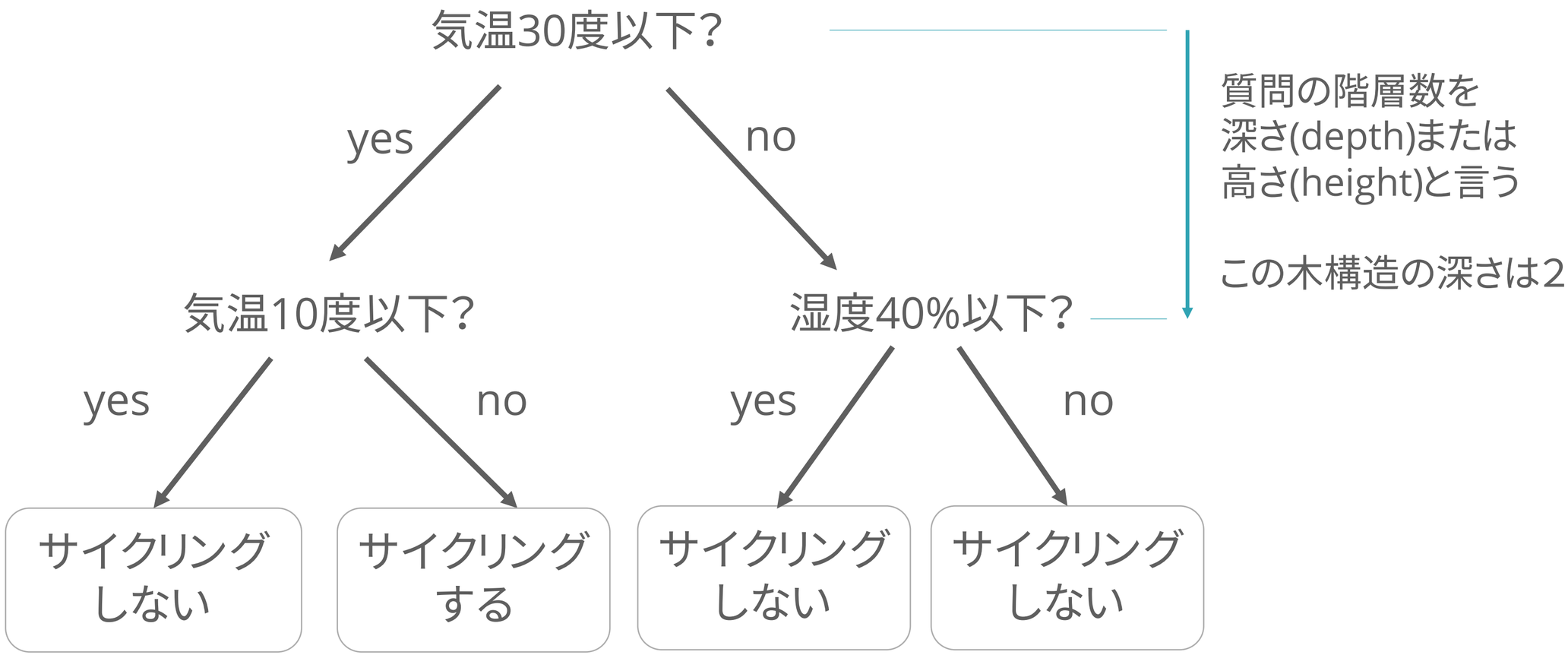
1. 講義概要
2. Decision Tree（決定木）
3. Cross Validation (交差検証)
4. Feature Engineering (特徴量作成)

Decision Tree（決定木）

- 木構造を用いた予測モデル
  - 解釈可能性が高い
    - なぜそのような結果になったかが分かりやすい
    - 説明責任が高いケースに用いられる
  - 分類または回帰モデルがある

Decision Tree（決定木）

■木構造とは



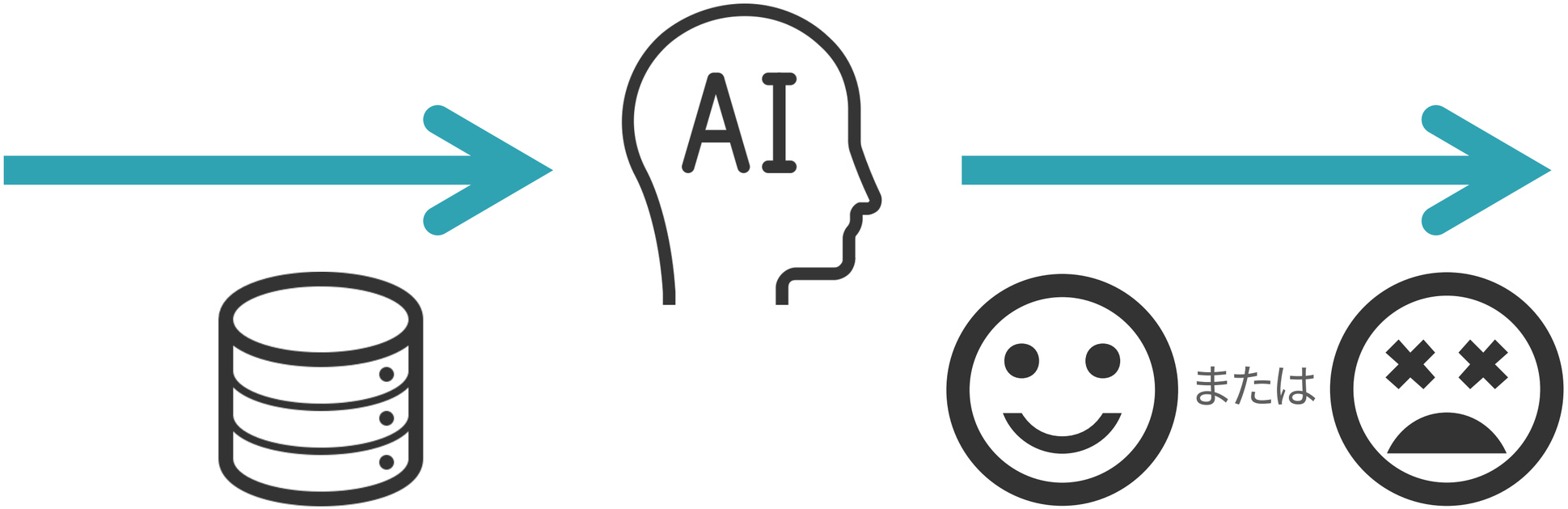
Decision Tree（決定木）

- 解釈可能性が高いためビジネスでよく利用される
  - 要因の把握ができる
    - 顧客の嗜好，購入動機の分析（マーケティング）
    - 顧客の信用度の予測（金融機関など）

- 解釈可能性
  - 予測モデルが，予測を行うプロセスの理解のしやすさのこと
  - Deep Learningなどの手法は，ブラックボックスであるため解釈可能性が低い

Decision Tree（決定木）| 金融機関の例

予測モデル



顧客データ

収入，年齢，資産，取引履歴

顧客の貸し倒れリスク

お金を貸しても大丈夫か否か

Decision Tree（決定木）| 金融機関の例

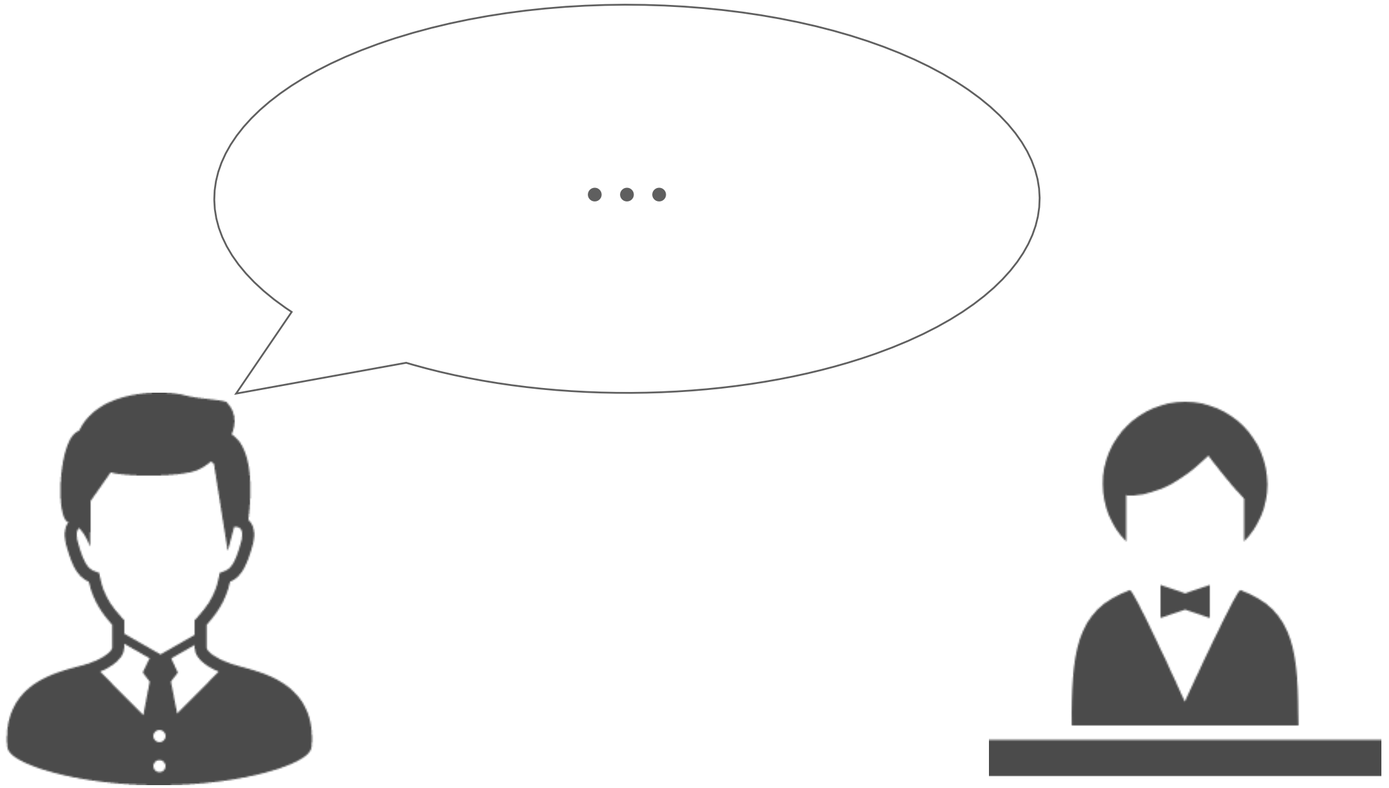
解釈可能性が低い予測モデルを使った場合

Decision Tree（決定木）| 金融機関の例

解釈可能性が低い予測モデルを使った場合

Decision Tree（決定木）| 金融機関の例

解釈可能性が低い予測モデルを使った場合



Decision Tree（決定木）

- なぜそうなったかといった説明責任が高いケースでは解釈可能性が高いモデルを使用することが推奨される
- 解釈可能性が高いモデルを使うことで，どう言った理由で顧客にお金を貸すことができないかを説明できる
  - 収入が，希望の貸金に比べ低い
  - 過去に貸し倒れたことがある，など

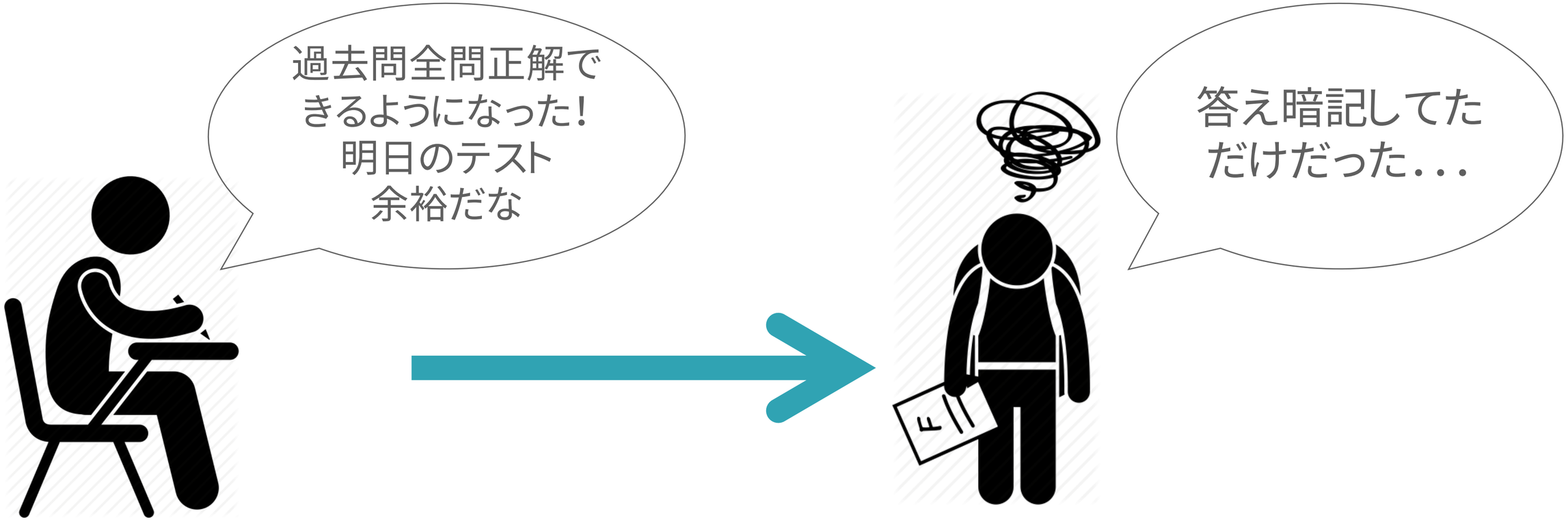
コンテンツ

1. 講義概要
2. Decision Tree（決定木）
3. Cross Validation (交差検証)
4. Feature Engineering (特徴量作成)

Cross Validation (交差検証)

- 予測モデルの汎化性能の評価方法
- 汎化性能
  - 未知のテストデータに対する識別能力の高さのこと
    - トレーニングデータに対する予測精度が高いだけでは，良いモデルと言えない

Cross Validation (交差検証) | 汎化性能



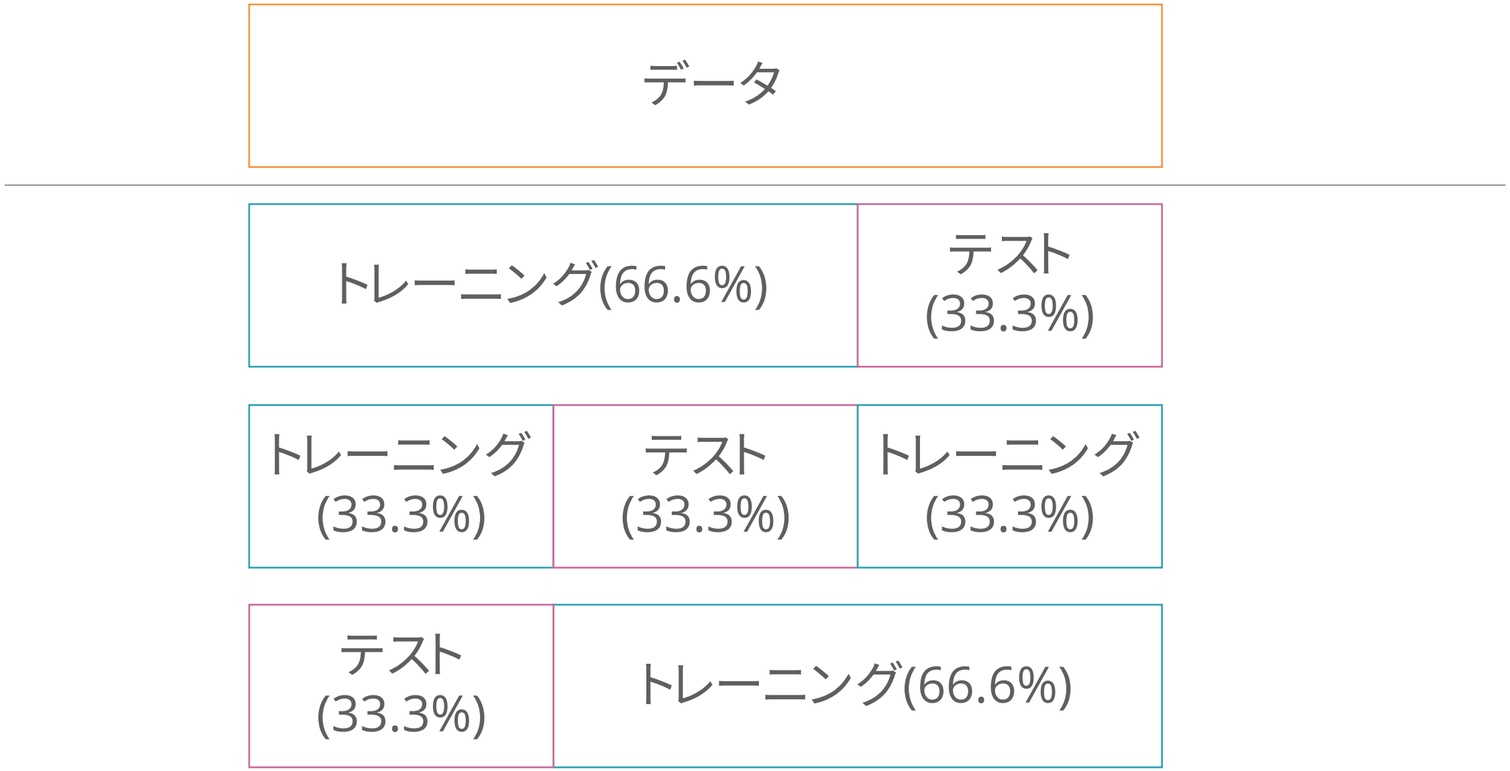
Cross Validation (交差検証)

- 予測モデルがトレーニングデータに対して，過剰にフィッティングすることを過学習という
  - 汎化性能を高めるには，過学習を抑える必要がある

- 予測モデルが未知のテストデータに対しても，精度よく予測できるかを評価する必要がある
  - 交差検証は，その評価方法

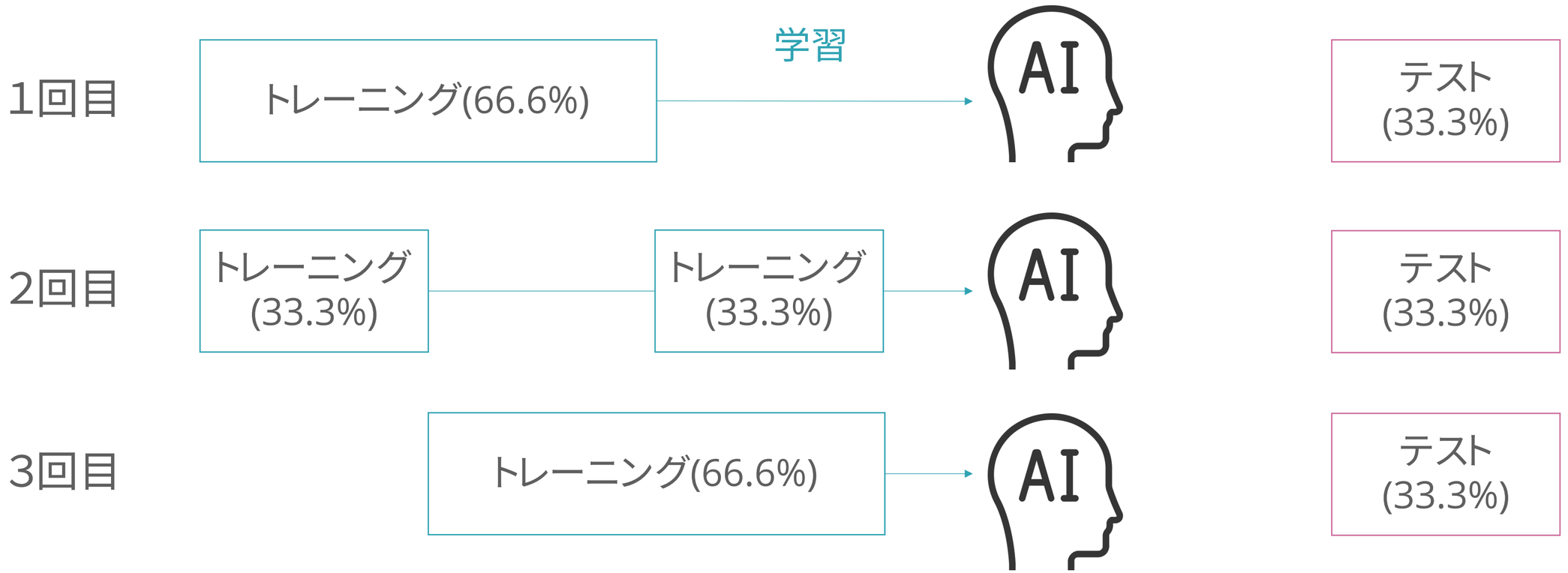
Cross Validation (交差検証) | アルゴリズム

- データを複数個に分割し，トレーニングデータとテストデータを作成する
  - 分割数は，経験的に3~5が良いとされている
  - 下図は3分割の例



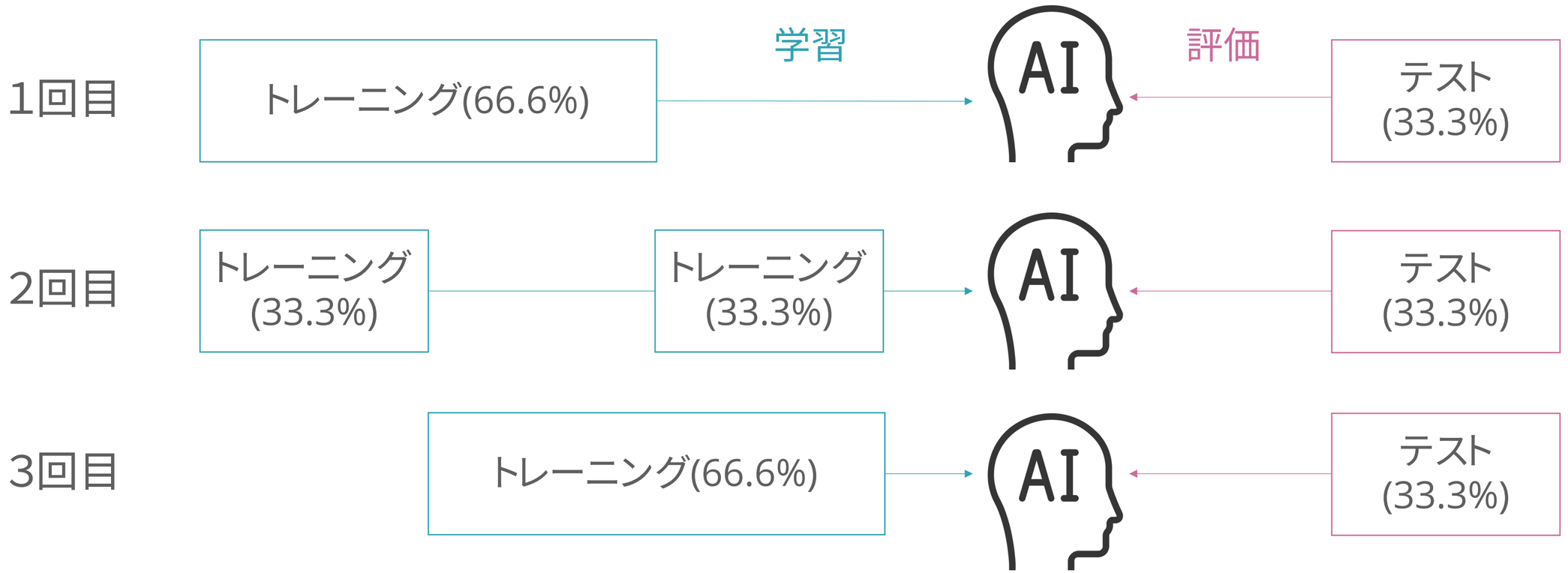
Cross Validation (交差検証) | アルゴリズム

- 各分割のトレーニングデータ(過去問)で，予測モデルを学習
  - テストデータは評価用にとっておく



Cross Validation (交差検証) | アルゴリズム

- 各分割のテストデータ(本試験)で，予測モデルを評価
  - 各分割の平均を，予測モデルの精度とする



コンテンツ

1. 講義概要
2. Decision Tree（決定木）
3. Cross Validation (交差検証)
4. Feature Engineering (特徴量作成)

Feature Engineering (特徴量作成)

- 予測モデルの精度を向上させるため，追加の予測因子（特徴量）を作成してデータセットに追加すること
- 演習で実際に行ってみましょう
  - jupyter notebookでeda.ipynbを開いてください