

*Jubatus*勉強会
第二回 特徴ベクトル入門

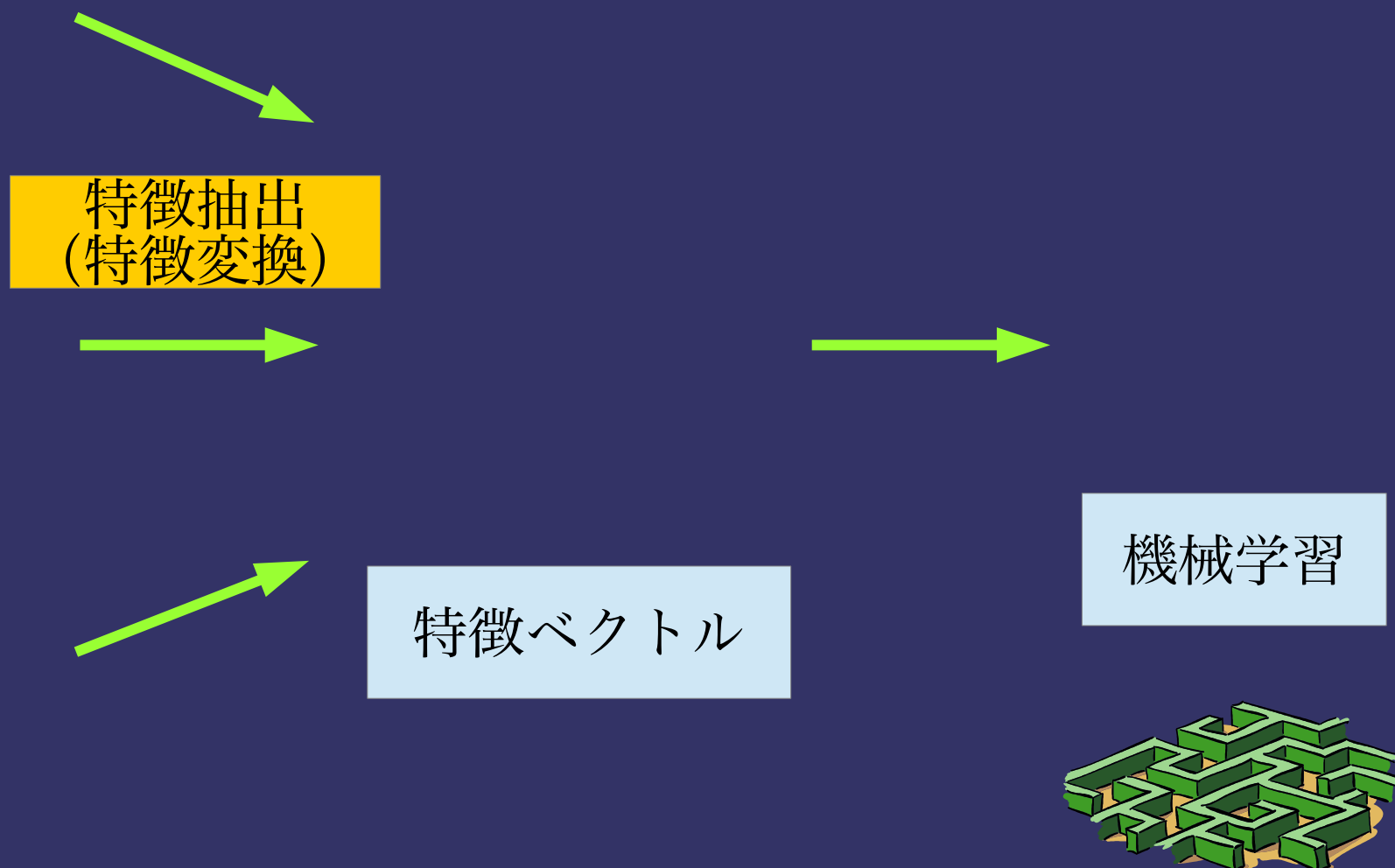


おさらい

- 機械学習に生のデータを入力できない
 - 画像・テキスト・時系列データは変換が必要
 - Jubatusにはコンバータが備わっている



非構造データの変換



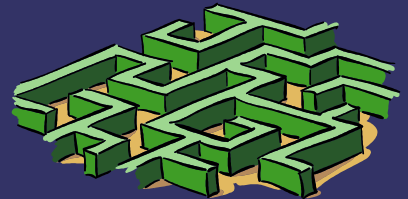
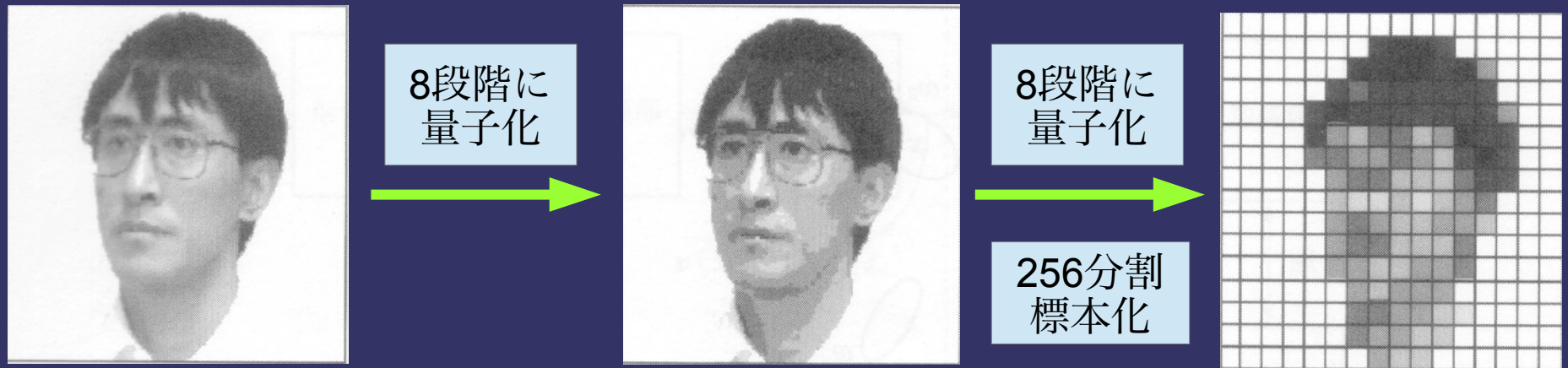
特徴抽出の問題点

- 特徴抽出処理は専門的な知識が必要
 - 画像なら画像の処理に関する知識が必須
- Jubatusでは特徴抽出処理も簡単に使える
 - 専門的な知識を必要とせずに分析可能



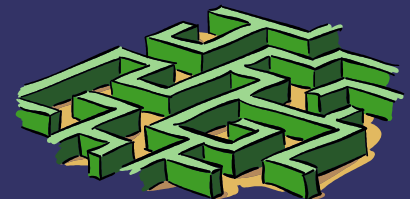
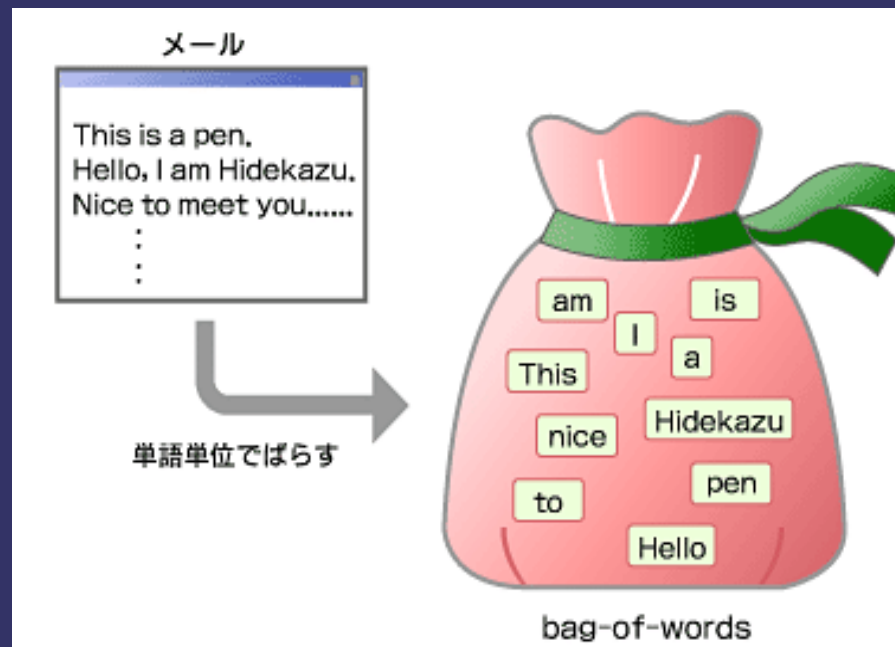
特徴抽出の基本的な考え方

– 画像データに対する特徴抽出



特徴抽出の基本的な考え方

- テキストデータに対する特徴抽出
 - Bag of Wordsがよく使われる
 - 各単語に固有のIDが振り分けられている
 - 単語の数だけ次元が存在する



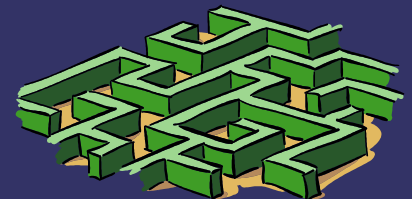
疎ベクトルと密ベクトル

- 再掲：次元数は単語の数だけ存在する
 - 次元数が膨大な数になる！
 - ほとんど0のベクトルを疎ベクトル
 - ほとんどの要素が埋まっているベクトルを密ベクトル



重み付け

- ベクトルの値には何を入れるのか？
 - TF-IDFが一般的に使われる
 - TF (term frequency) : 文字列中に何回出現したのか？
 - 単語の重要性を示す
 - DF (document frequency) : 文書に何回出現したのか？
 - 単語のレア度を示す
- $TF\text{-}IDF = TF * 1/DF$
 - 実際は調整がはいることが一般的



特徴抽出の勘所

- Bowの問題点
 - 同義語を表現できない
 - 単語の順番を考慮していない
- 問題を解決するために研究が行われている

