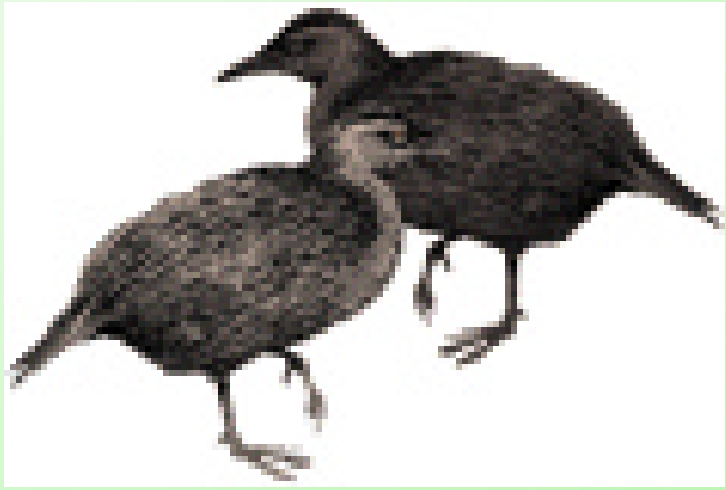
# オープンソースデータマイニングツール WEKA



静岡大学大学院理工学研究科

阿部　秀尚

hidenao@ks.cs.inf.shizuoka.ac.jp

http://panda.cs.inf.shizuoka.ac.jp/~hidenao/

# 自己紹介

阿部秀尚（あべ　ひでなお）
静岡大学大学院　理工学研究科
博士後期課程　設計科学専攻

研究テーマ：
「リポジトリに基づく
帰納アプリケーション構築支援環境の開発」

キーワード：
「リポジトリ」構成型メタ学習」帰納アプリケーション」

WEKAとの関連：
- データマイニング全般の紹介として利用
- WEKAにより提供される選択型メタ学習を実験対象として利用

# 用語説明

- データセット(Data Set)
  - 発表中では表形式のデータを意味する
- 属性(Attribute)
  - 表形式で与えられたデータセットの列（カラム）, 説明変数
- インスタンス(Instance)
  - 表形式で与えられたデータセットの行（ライン）
- スキーム(Scheme)
  - 特定のタスクを実行するアルゴリズムの集まり
- プロセス(Process)
  - 実行経路

# WEKAとは？

- 世界で最も使われているフリーのデータマイニングツール
- ワイカト大学（ニュージーランド）を中心に開発されている
- オープンソース開発手法で開発が進められている
- 現在, Development Versionは3.3.6
- Javaで実装されている（⇐マルチプラットホーム）
- 開発チームが掲げる理念：

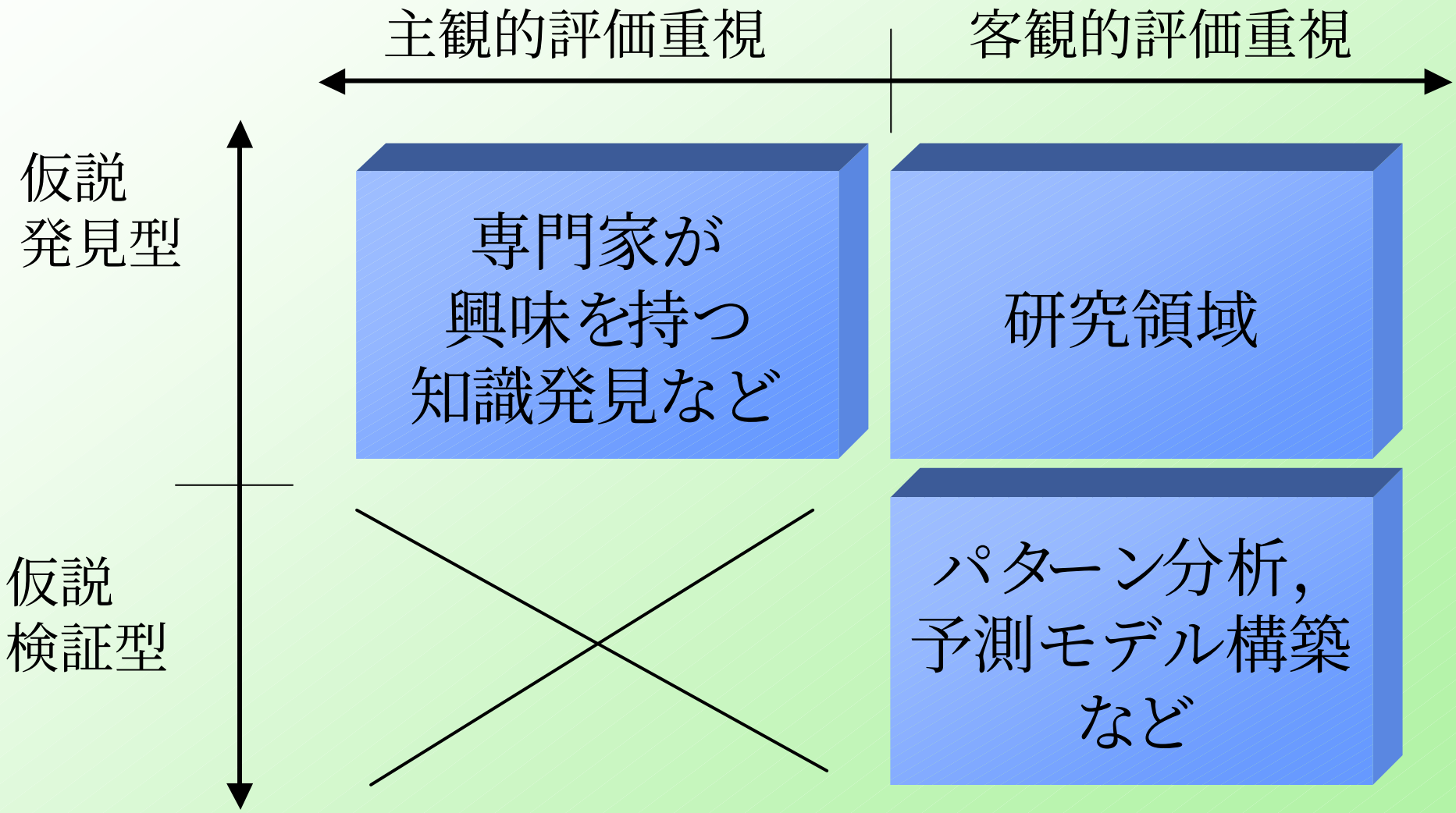
  機械学習の技術を広めること

  機械学習の技術を農業の現実問題に適用すること

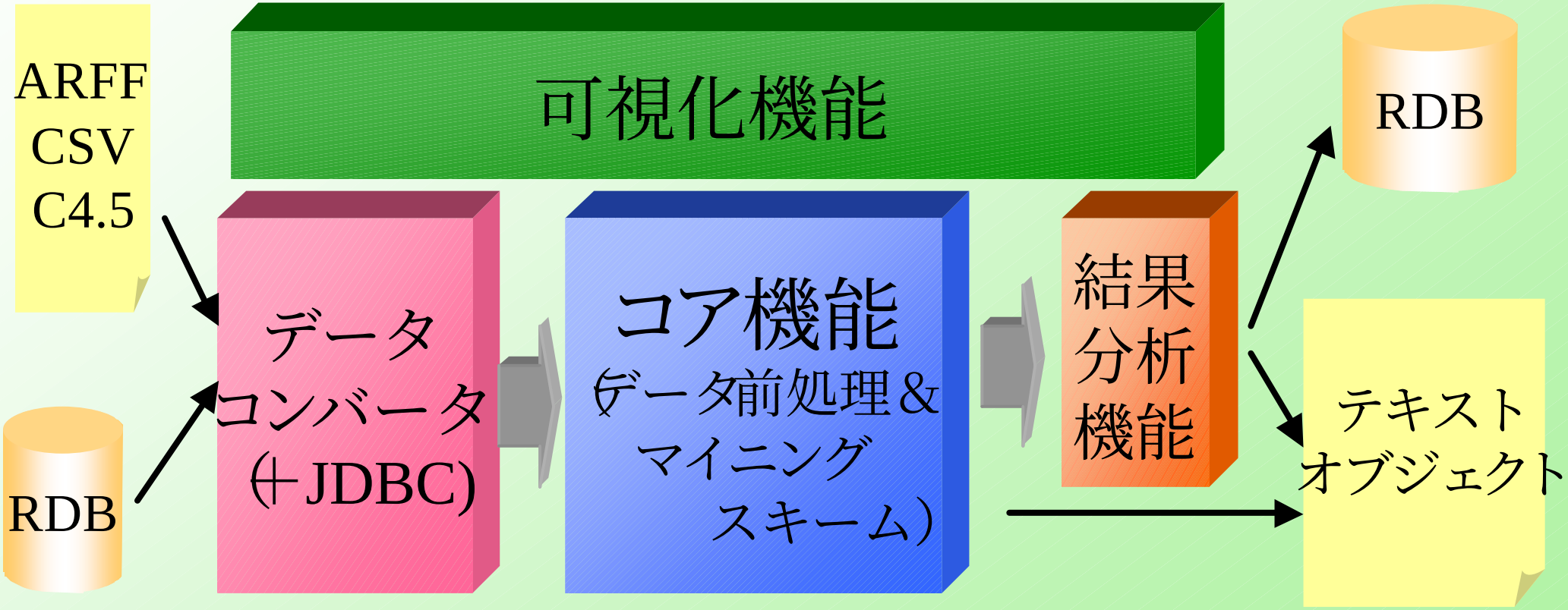
  新たな機械学習アルゴリズムを開発し, 世界に流布すること

  現場に論理的な枠組みを提供すること

# WEKAが適用されるタスク



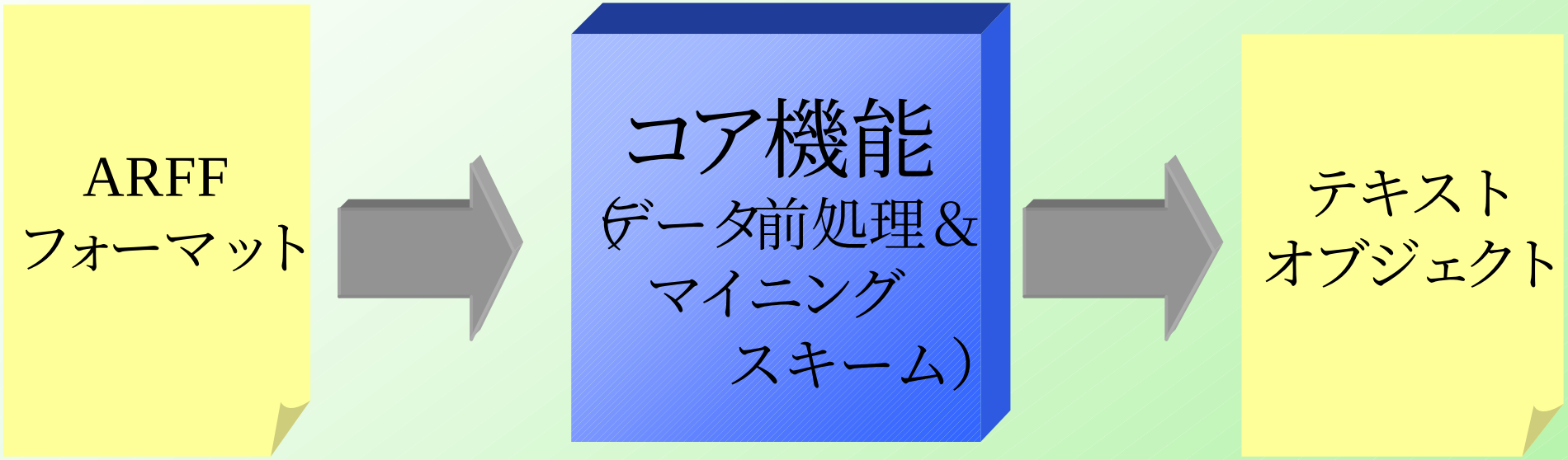
# WEKAの構成


ARFF
CSV
C4.5
RDB
可視化機能
データ
コンバータ
(+JDBC)
コア機能
(データ前処理&
マイニング
スキーム)
結果
分析
機能
RDB
テキスト
オブジェクト

# WEKAの特徴

- 数多くのマイニングスキームが利用可能
- API, CLI, GUIの各インターフェイスを備える
- 各種の可視化機能が提供される
- 商用データマイニングツールに迫る機能や品質
- 研究段階のアルゴリズムも実行可能
- ユーザの試行錯誤により, 新たなデータマイニングプロセスが実行可能
- ソースコードが公開されているため, アルゴリズムの教育目的に利用可能
- アルゴリズムとアルゴリズム内のパラメータが整理されている　etc…

# データ前処理＆マイニングスキーム




- コマンドラインから利用
- APIから利用

# ARFFフォーマット

```
@relation    weather
@attribute   outlook    {sunny,overcast,rainy}
@attribute   temperature   numeric
@attribute   humidity   integer
@attribute   windy   {true,false}
@attribute   play   {yes,no}

@data
sunny,85,85,FALSE,no
sunny,80,90,TRUE,no
overcast,83,86,FALSE,yes
rainy,70,？,FALSE,yes
```

データセットの名前

属性名を縦に列挙する

数値属性を示す
（実数，整数）

名義属性は属性値
をカンマ区切りで示す

クラスは他の名義属性と
同様に記述（実行時に指定）

これ以降CSV形式の
データ

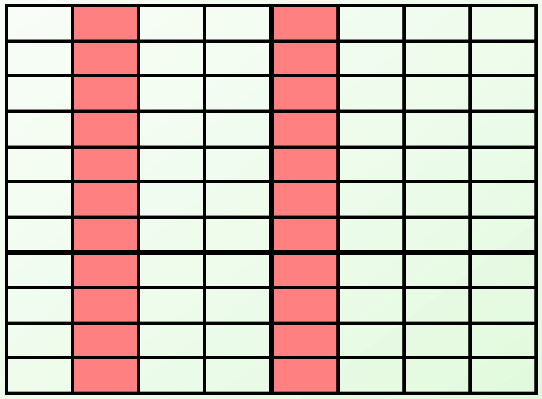
不明値は’?’

# 用意されているデータ前処理・マイニングスキーム(weka-3-3-6)

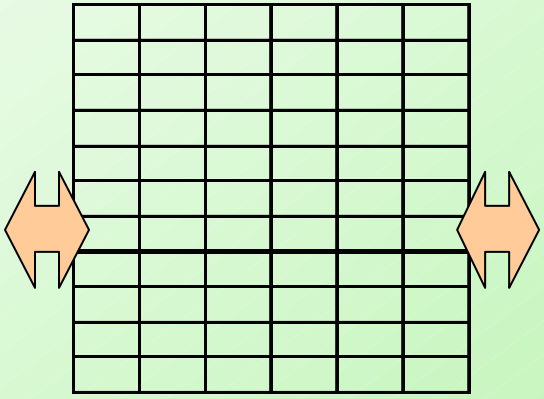
- フィルタ
  - 35種類
- 数値・分類予測スキーム
  - 60種類
- 相関ルール学習スキーム
  - 1種類
- クラスタリング
  - 5種類
- 属性選択
  - 評価法：12種類，探索法：8種類

# フィルタ

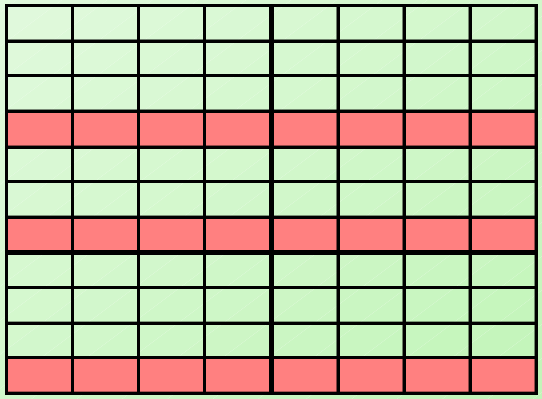
## 属性フィルタ



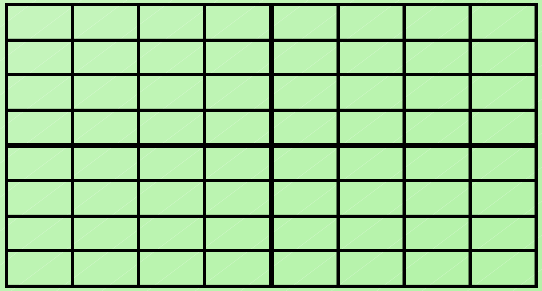
属性に対して操作を加える



## インスタンスフィルタ



インスタンスに対して
操作を加える



# 分類・数値予測スキーム

- ルール(rules)
  - J48決定木からのルール生成など10種類
- 決定木(tree)
  - J48(C4.5)やM5’モデル木など9種類
- 関数型(functions)
  - 線形回帰, ニューラルネットワーク, SVMなど11種類
- インスタンスベース(lazy)
  - K-NNなど5種類

# 分類 •数値予測スキーム 続き)

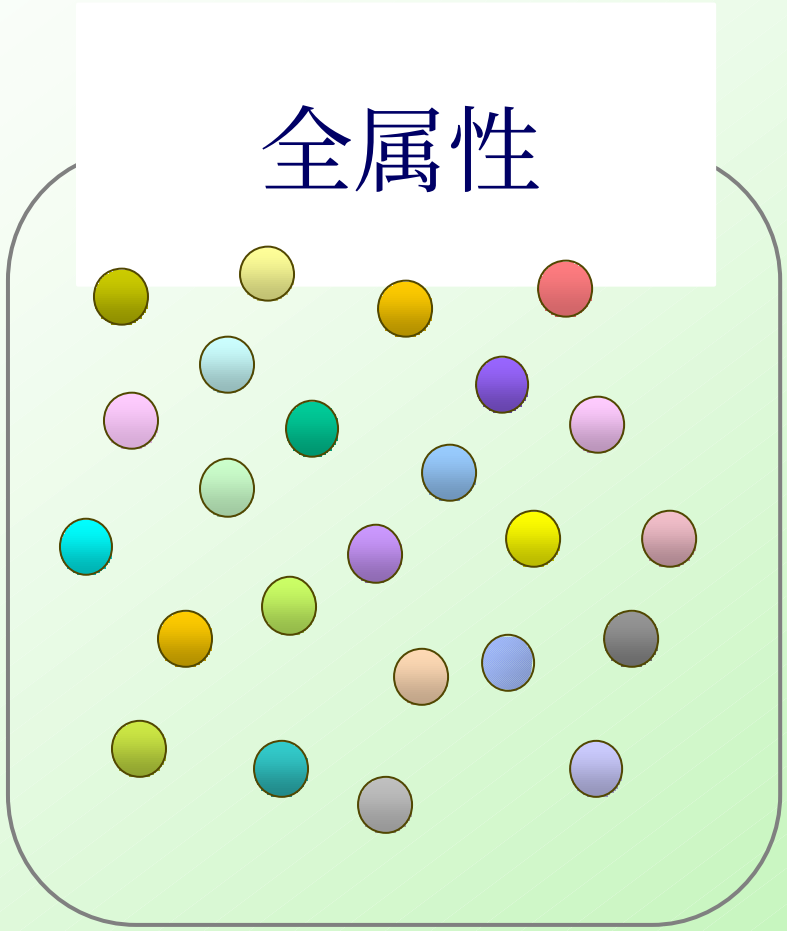
- ベイズ(Bayes)
  - NaiveBayes, ベイジアンネットk2など6種類
- メタスキーム(meta)
  - Bagging,Boosting,クラス分割など22種類
- その他(misc)
  - HyperPipesなど2種類

# 相関ルール •クラスタリング

- 相関ルール(Association Rules)
  - アプリオリアルゴリズム
- クラスタリング(Clusters)
  - K-means
  - EMアルゴリズム
  - CobWeb
  - FathestFirst

# 属性選択（フィルタアプローチ）

WEKAでは7種類

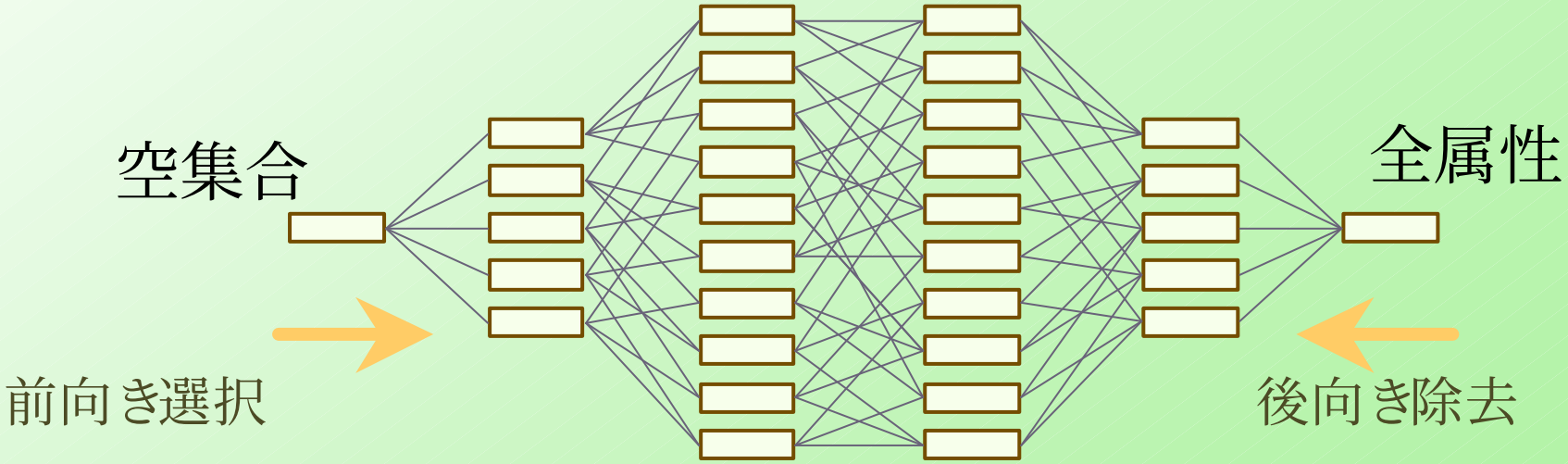



順位付け

- 重み
- エントロピー
- etc…

# 属性選択（ラッパーアプローチ）

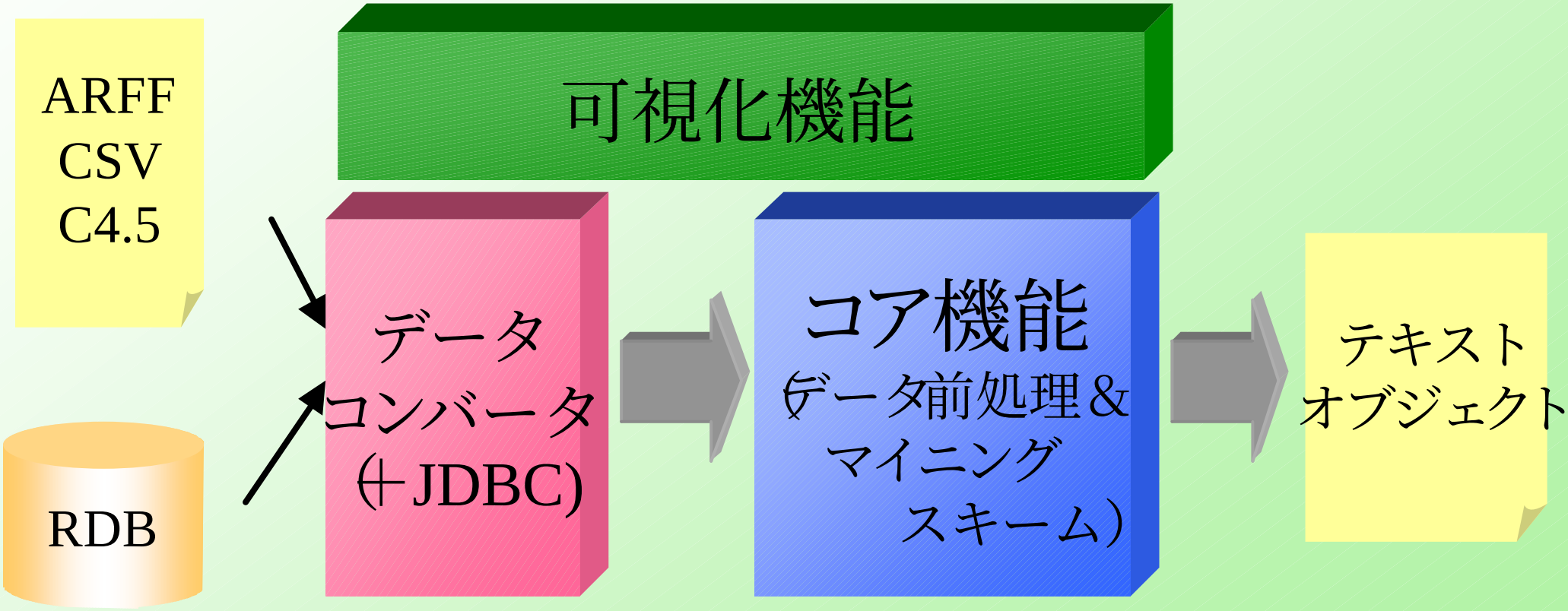
- WEKA では4種類
- 分類・数値予測スキームを実行しながら属性を選択する
  - 初期状態・探索方法:
    空集合から／全属性から／適切な中間点から
  - 評価方法・関数:5-CVによる正解率を用いた評価
  - 終了条件:正解率の収束度合（標準偏差）など



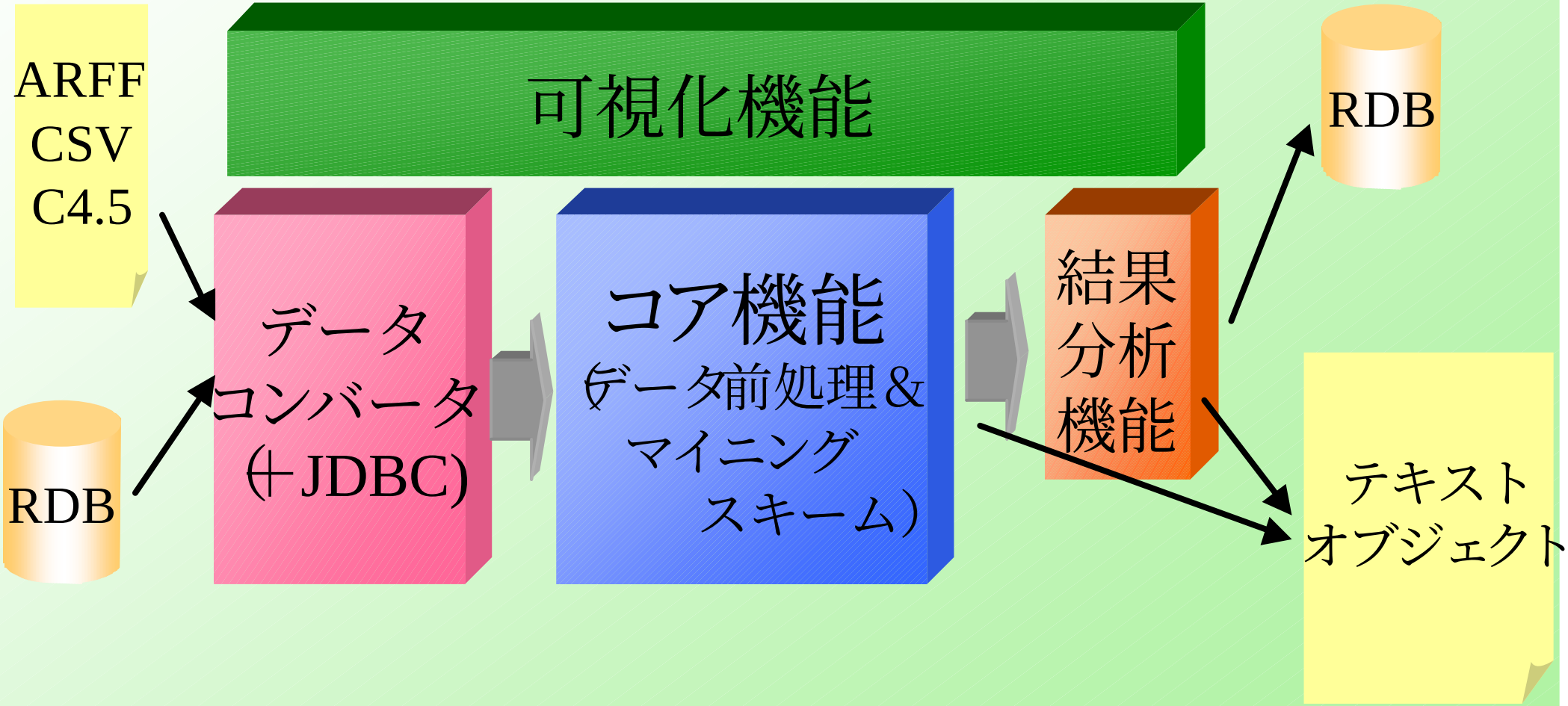
# WEKAのGUI

- Knowledge Explorer
  - マイニングプロセスを試行錯誤で作成
- Experiment Environment
  - 複数のマイニングプロセスを複数のデータセットに対して実行し, 分析を行う
- Knowledge Flow
  - 上記の 2つのGUIでの操作の過程を可視化

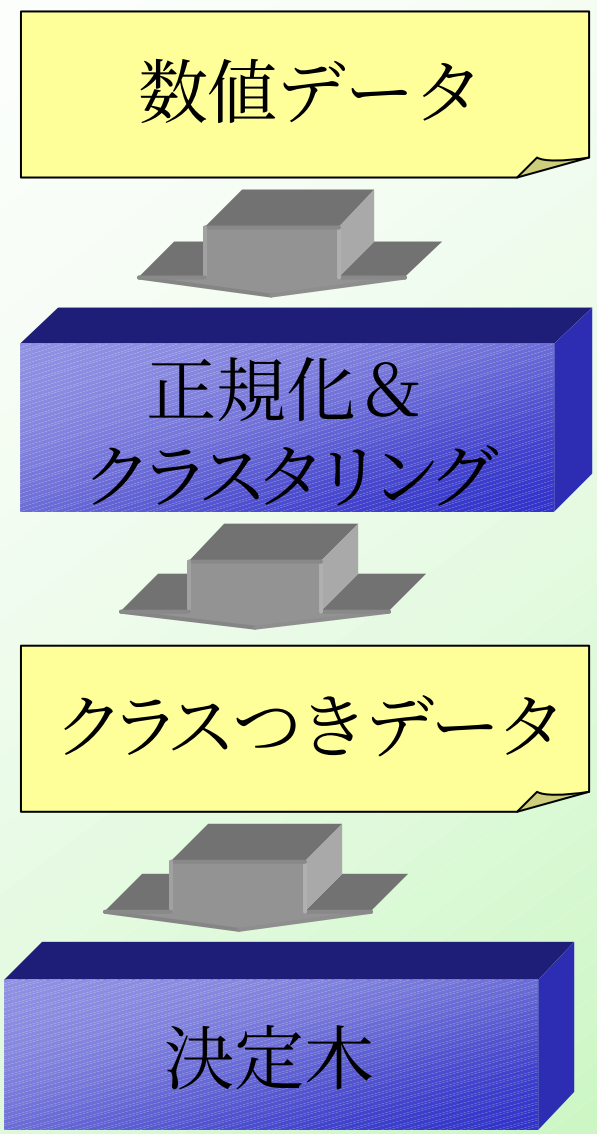
# Knowledge Explorerでの機能



# Experiment Environmentでの機能



# デモンストレーション




Knowledge Explorerでのデモ
(Linux上)

Knowledge Flowでのデモ
(Windows上)

# 関連ソフトウェア

- WEKAのAPIを利用した拡張
  - Automatic Knowledge Minor
- ユーザによるデータマイニングスキーム選択をサポートする機能を拡張
  - WEKA Metal

# まとめ

- 一般的なデータ前処理 •マイニングアルゴリズムが整理されているツール
- 研究段階のデータ前処理 •マイニングアルゴリズムも使用できるツール
- 入力データに適合したマイニングプロセスの作成が可能なツール
- マイニングプロセスの作成を積極的に支援するツールではない

# 最後に

- 公式Webページ
  http://www.cs.waikato.ac.jp/~ml/
- 書籍
  I.H. Witten, E. Frank: “Data Mining”, MORGAN KAUFMANN
  ISBN 1-55860-552-5
- 日本語による情報（非公式）
  http://panda.cs.inf.shizuoka.ac.jp/~hidenao/work/weka/
  weka-jp@ks.cs.inf.shizuoka.ac.jp
- Weka3.3.6を再配布します．声をかけてください．