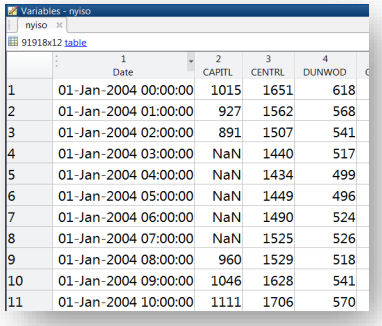
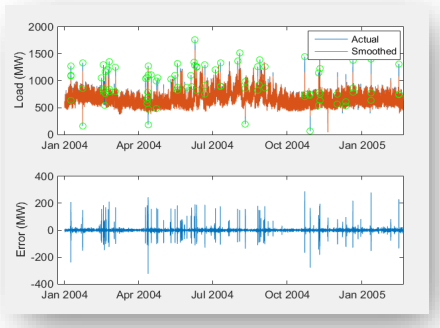
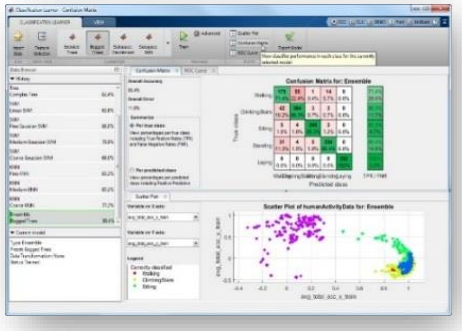
データ活用を成功させるための解析ワークフロー

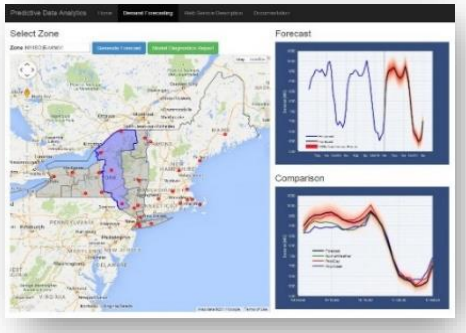
MathWorks® Japan
アプリケーションエンジニアリング部
アプリケーションエンジニア
井原 瑞希

データ解析ワークフロー



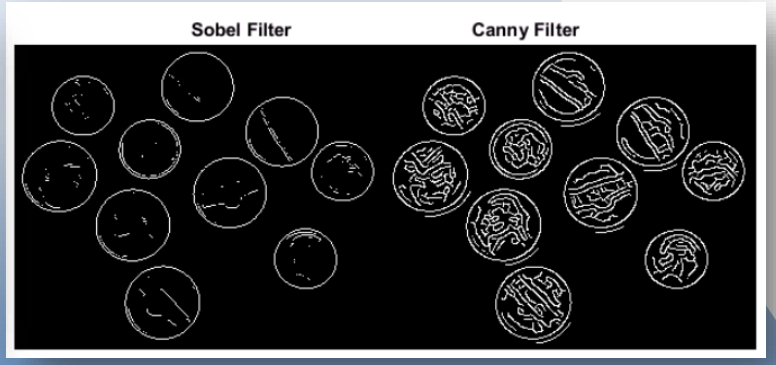




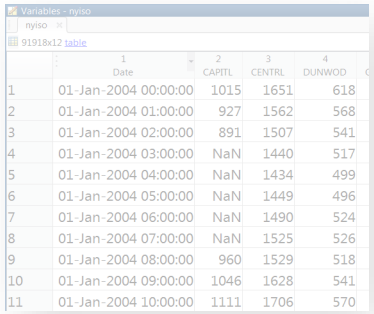


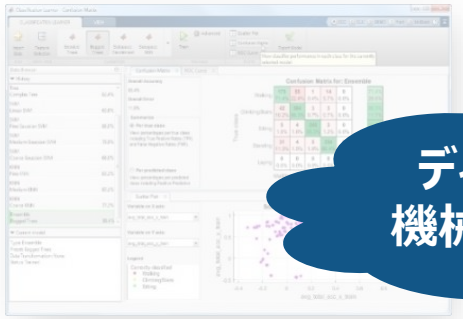
データへのアクセス　データの前処理　予測モデルの構築　システムへの統合

画像/信号処理



データ解析ワークフロー

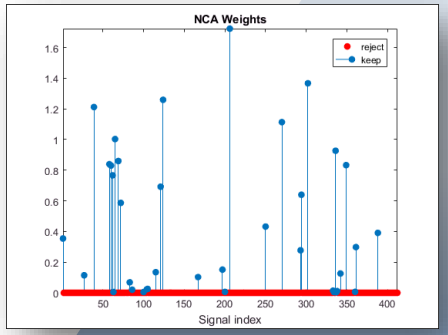


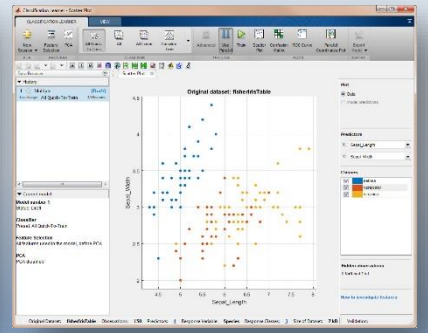


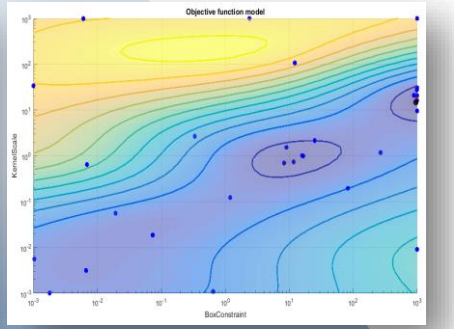
ディープラーニングも
機械学習モデルの一種

データへのアクセス　データの前処理　予測モデルの構築　システムへの統合

機械学習







特徴選択　モデルの選択　ファイン
チューニング

機械学習とは

- 機械学習の定義
  - **データ**から直接観測できないパタンやルールを、
    モデルを元にして**機械**的 (自動的) に**モデル**を**学習**すること

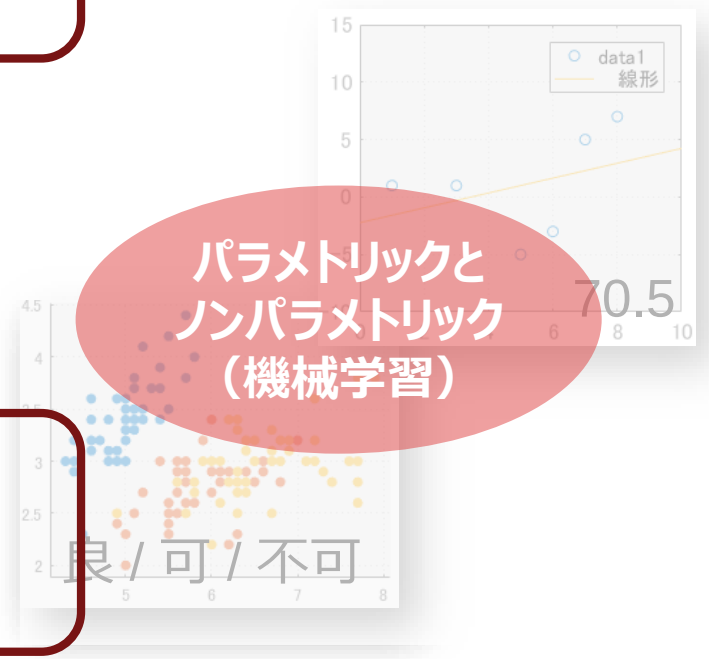
$$y = f(x)$$

  - 現在のデータ (学習データ) から関数 f を学習すること

- なぜ機械学習を使うのか
  - 未知のサンプルに対する予測が可能
  - 予測に必要な情報を残し、冗長な情報を
    省くことが可能

機械学習とは

- 機械学習
  - 教師なし学習
    - data driven
    - 次元削減
    - クラスタリング
  - 教師あり学習
    - task driven
    - 回帰
      - 連続値
    - 分類
      - 離散値



本日のトピック

- **分類**
  - 機械学習ベースの分類ワークフロー
  - コーティングを簡単にする方法
  - 分類のモデル選択
  - 特徴量の重要度判定

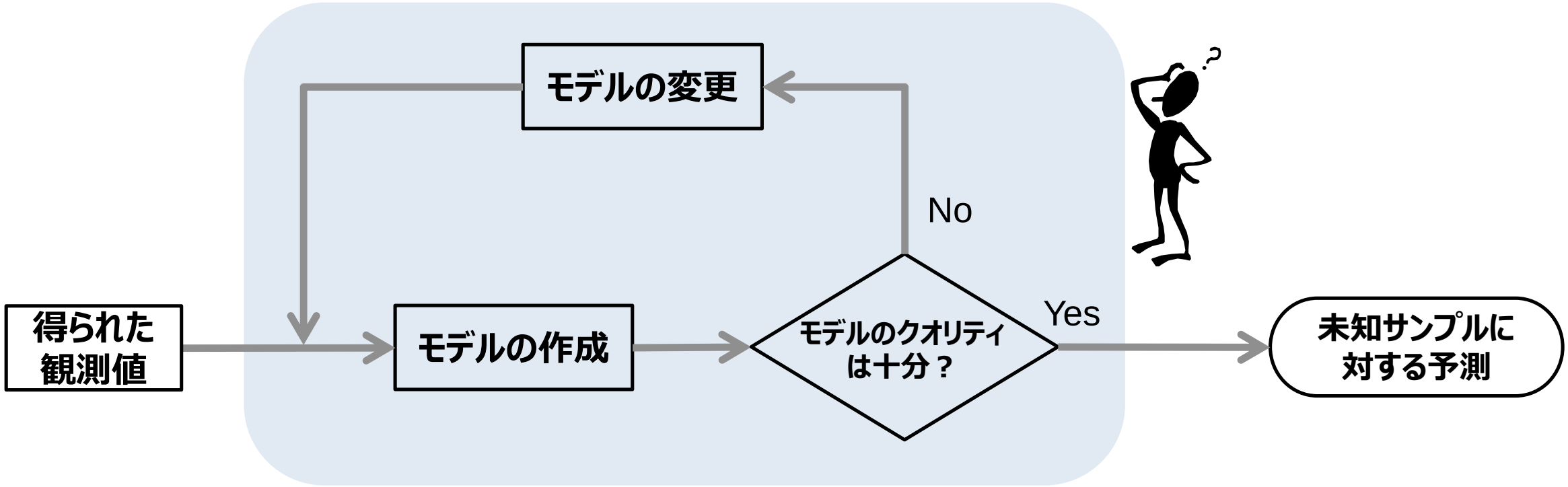
- **テキスト解析（教師なし学習）**
  - テキスト解析のワークフロー
  - テキストデータ解析に特化した難点と解決策
  - 機械学習の課題 – ハイパーパラメータの探索

本日のトピック

- **分類**
  - 機械学習ベースの分類ワークフロー
  - コーティングを簡単にする方法
  - 分類のモデル選択
  - 特徴量の重要度判定

- **テキスト解析（教師なし学習）**
  - テキスト解析のワークフロー
  - テキストデータ解析に特化した難点と解決策
  - 機械学習の課題 – ハイパーパラメータの探索

機械学習の流れ



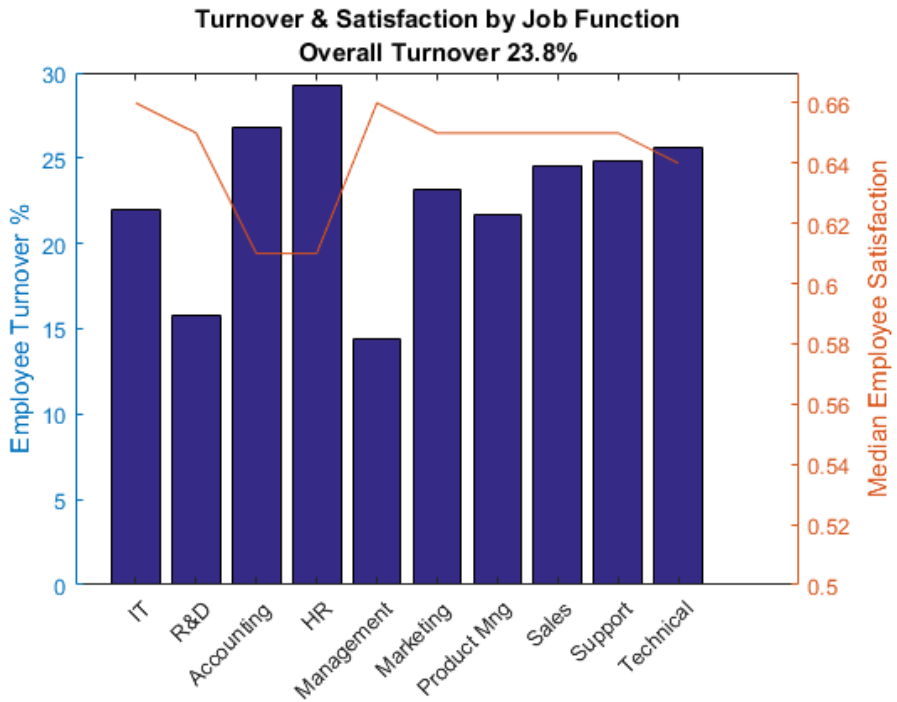
例: 退職理由の分析

目的

- 従業員が退職してしまう理由を分析

アプローチ

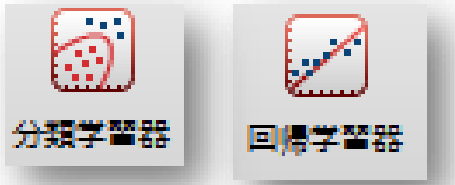
- 人事アンケートデータの読み込み
- データ理解のための可視化
- 機械学習の手法を使用した退職予測
- 退職に関わる重要要素を分析



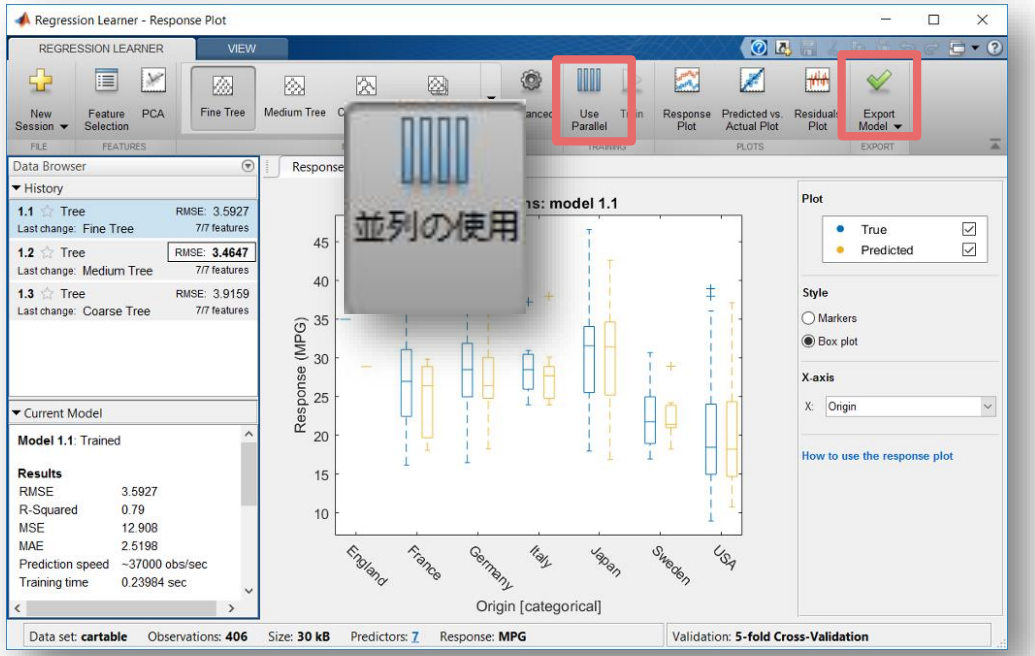
アプリから機械学習の MATLAB® コード作成

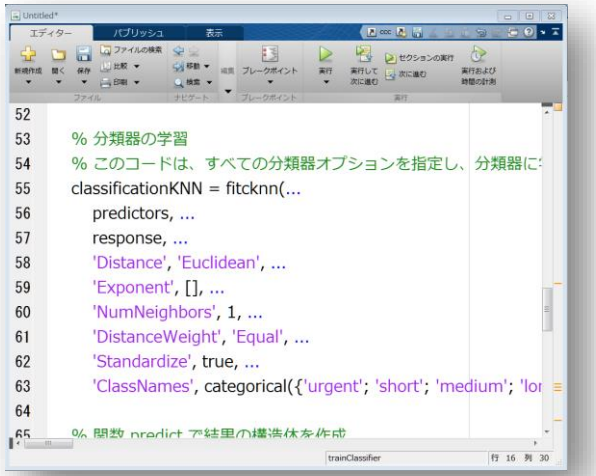
Statistics and Machine Learning Toolbox
Parallel Computing Toolbox™

R2015a R2017a



- 分類/回帰学習器アプリ
  - データを分類するためのモデル学習 GUI
  - GUI 操作を MATLAB コードとして生成可能
  - 複数モデルを並列に学習可能

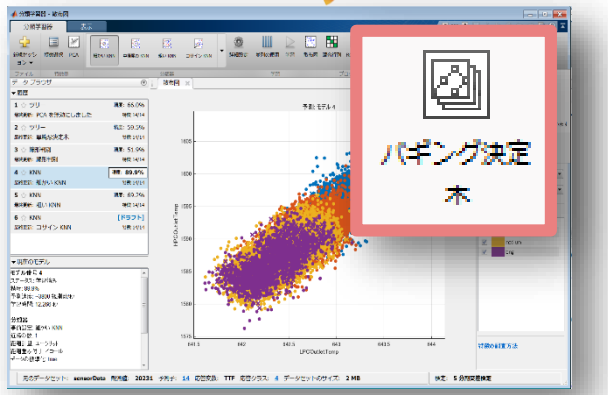




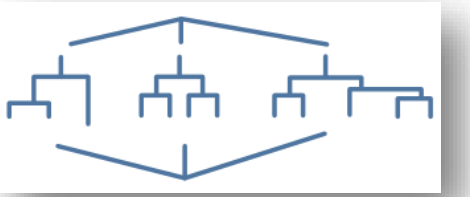
MATLAB プログラムの自動生成

ランダムフォレストとは

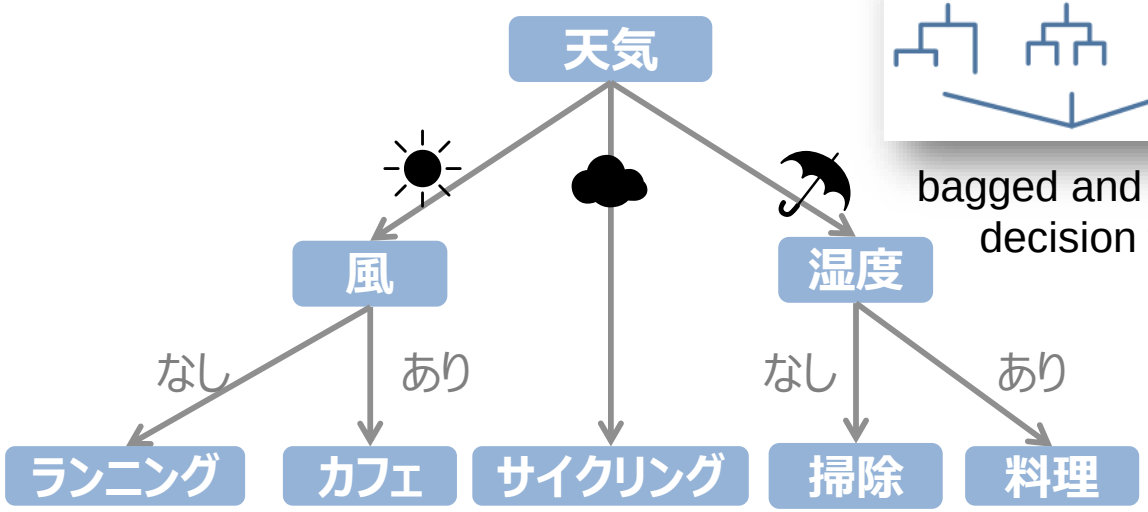
- 決定木
  - 大半は正しい予測が可能
  - 間違った予測も起こる
- バギング
  - データの一部を使って学習することを何度も繰り返す（結果は再利用しない）
- アンサンブル学習
  - 三人寄れば文殊の知恵
  - 多数の決定木 (弱学習器) を組み合わせ



- 決定木 + バギング
  - 利点
    - 並列計算が容易 (効果が出やすい)
    - 特徴量の重要度判定が可能
  - 欠点
    - 他の手法と比較して過学習しがち
    - 木が多いと計算量が増加



bagged and boosted decision trees



ランダムフォレストの良さ

- 並列計算によるモデル学習の高速化

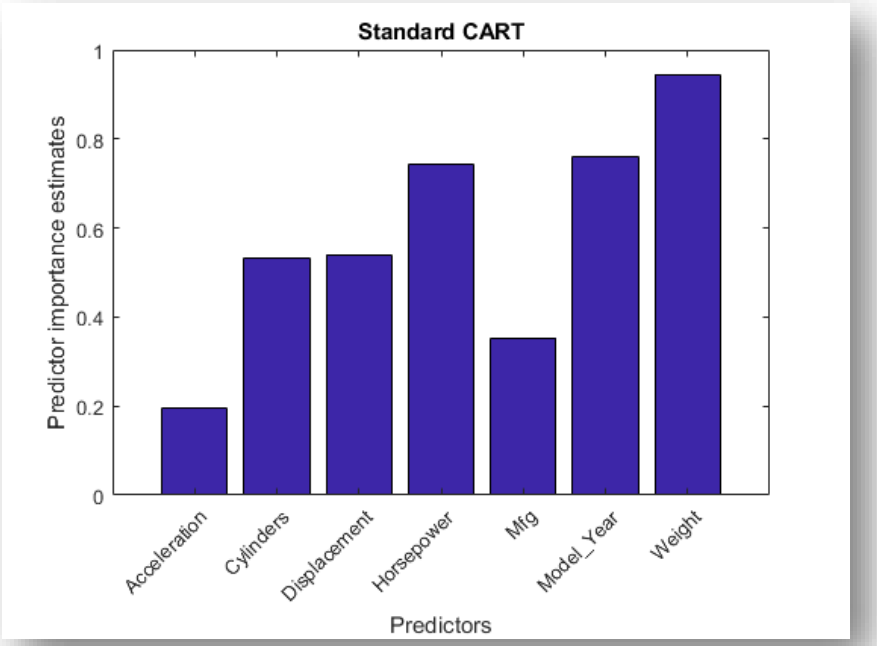
```
>> opt = statset('UseParallel', true);
>> tree = TreeBagger(…, 'Options', opt)
```

- 特徴量の重要度判定
  - 分類結果から要因を分析可能
  - 不要な特徴量を削減
  - 重要度の推定

```
>> predictorImportance(tree)
```

  - Out-of-bag 誤差
    使用されなかったサンプルからの重要度推定

```
>> oobPermutedPredictorImportance(tree)
```



本日のトピック

- **分類**
  - 機械学習ベースの分類ワークフロー
  - コーティングを簡単にする方法
  - 分類のモデル選択
  - 特徴量の重要度判定

- **テキスト解析（教師なし学習）**
  - テキストデータ解析に特化した難点と解決策
  - テキスト解析のワークフロー
  - テキストの教師なしクラスタリング例

テキスト解析の利点とアプリケーション

- アプリケーション例
  - **感情分析**
  - **故障メンテナンス**
  - **ドキュメント分類**

テキストを解析する際の課題

```
bagOfWords with 210 words and 611 documents:
    preventative    maintenance    service    check    noise    ...
               1              1          1        1        1
               0              0          0        0        0
               ...
```

1. テキストをそのまま扱うことが難しい
   - 数値 (行列) に置き換えて解析

2. 出現する単語の数が膨大なため、大規模かつスパースな特徴行列
   - 特徴量の数が膨大でも扱えるようなアルゴリズムの使用
   - 次元削減

3. 周辺の単語によってコンテキストが変化
   - 単語の分散表現 (word2vec など)

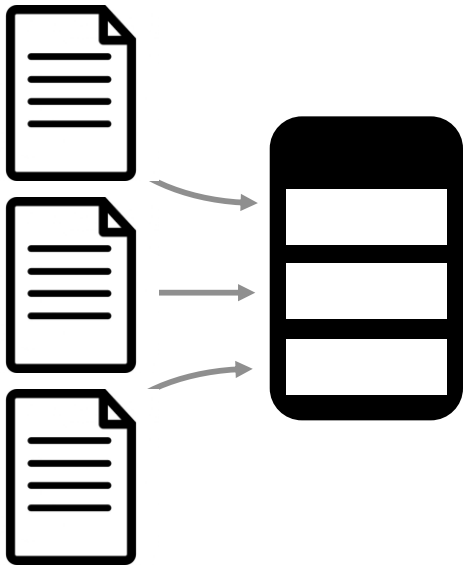
```
king = word2vec(emb, "king");
man = word2vec(emb, "man");
woman = word2vec(emb, "woman");

word = vec2word(emb, king - man + woman)
```

```
word = "queen"
```

テキスト解析のワークフロー

データアクセス → データの前処理 → 予測モデルの構築

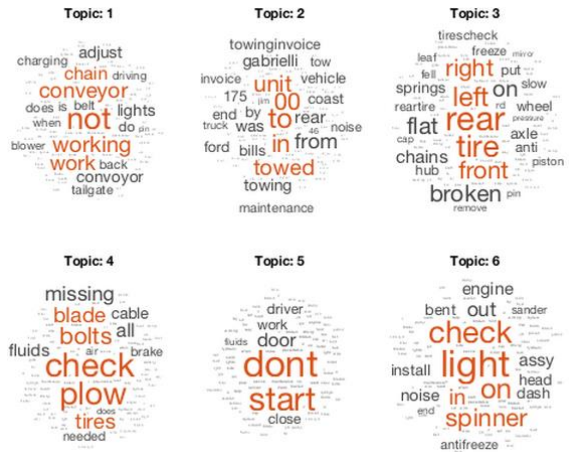
テキストの整形 → テキストの数値化



GAS IN FUEL TANKHYD
LINES UNDER
CABSTROB LIGHTS

Gas fuel tankhyd lines
cabstrob lights

	gas	fuel	tank	hyd
doc1	1	0	1	0
doc2	1	1	0	1



例: 車の修理データの解析

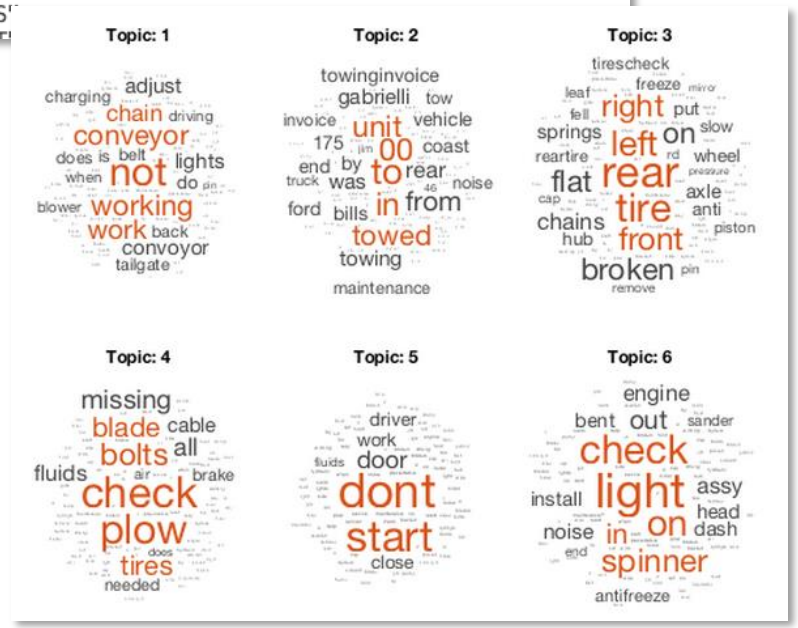
目的

- 車の修理ログから、メンテナンスを実施した理由を分析

アプローチ

- 車両修理データの読み込み
- テキストを分析しやすくするための整形
- 分野に特化した用語に対する前処理
- 機械学習の手法を使用してメンテナンス実施の主な理由を特定

```
repairNotes = 617×1 string array
    "PM SERVICE, CHECK TURN SIGNAL, CLUNKING NOISE WHEN DRIVING"
    "SERVICEROB,EXT,5604"
    "NEED 4 PLOW PINS"
    "INSTALL SPINNER ASSY"
    "DONT START"
    "DOG BONE PIN BROKEN"
    "NEED SERVICE, CHECK BRAKES"
    "HYD CAP CHECK ENGINE LIGHT ON"
    "TARP VALVE STICKINGRIGHT SIDE MIRROR BRACKET BROKEN"
    "HANDLES IN CAB LOOSE"
    "NO PLOW LIGHTS
```

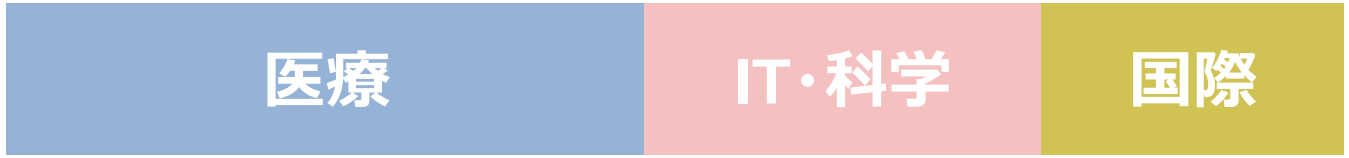



```
>> vehicleRepairAnalysis_jp
```

潜在的ディリクレ配分法 (Latent Dirichlet allocation; LDA) とは

- 文章をトピックの混合と仮定した文章の生成モデル
- 各文章の**トピック分布**と、各トピックの**単語分布**を求める
  - トピック分布

文章 A



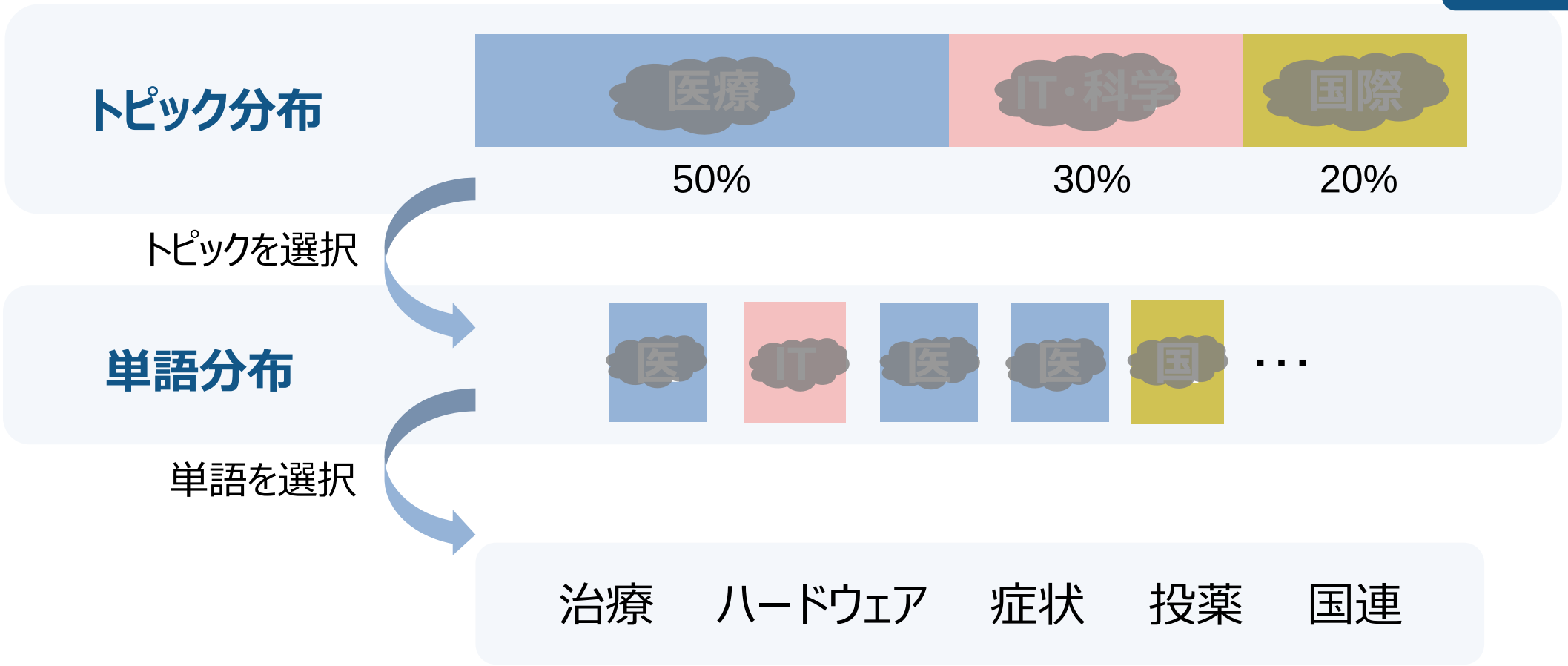
  - 単語分布

医療

治療	診断	症状	血液	…
0.6%	0.2%	0.4%	0.1%	…

潜在的ディリクレ配分法 (Latent Dirichlet allocation; LDA) とは

教師なし学習



Text Analytics Toolbox

ワークフローごとの機能

データへのアクセス → **データの前処理** → **予測モデルの構築**

テキストの整形 → **テキストの数値化**

テキストファイル
スプレッドシート
Web

削除/抽出
正規表現

Word ドキュメント
PDF

高頻度の単語抽出
ステミング (マッチング)
トークン化 (分かち書き)

単語カウント (Bag-of-words)
TF-IDF
単語の分散表現

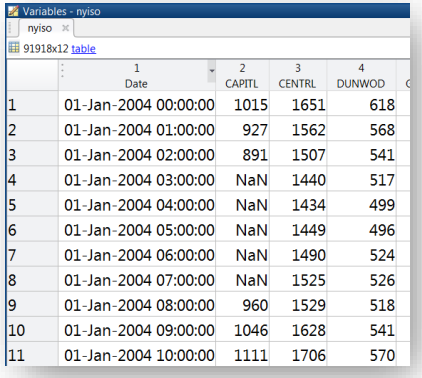
潜在意味解析 (LSA)
潜在的ディリクレ配分法 (LDA)

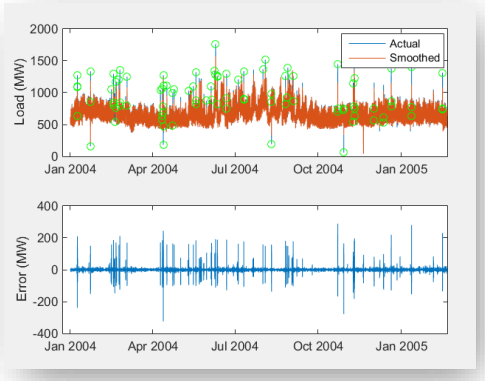
R2017b では英語対応のみ

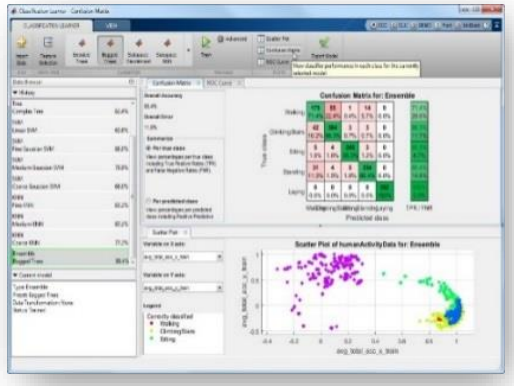
可視化

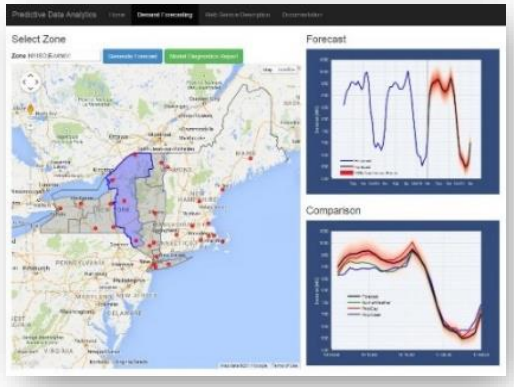
Word Cloud
Text Scatter

Text Analytics Toolbox と他機能の組み合わせ









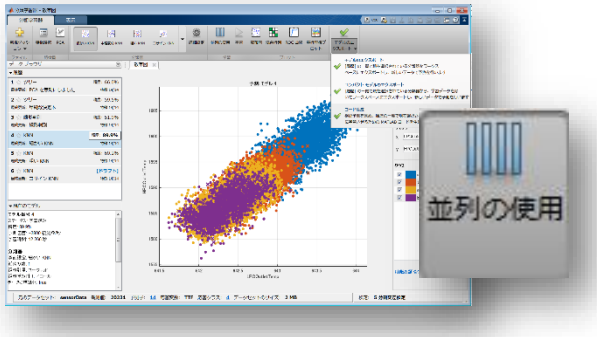
データアクセス → **データの前処理** → **予測モデルの構築** → **システムへの統合**

- データベース
- Deep Learning (RNN, LSTM)
- exe / Web アプリ化などの展開

並列計算

前提条件: Statistics and Machine Learning Toolbox

Key Takeaways



- 機械学習を使用する際のショートカット
  - アプリの利用
  - 並列計算の併用

- 機械学習からさらなる分析へ
  - 機械学習の結果分析

- データ読み込みから予測モデルの構築までの
  テキストデータ解析のワークフローに対応する機能提供
  - Text Analytics Toolbox