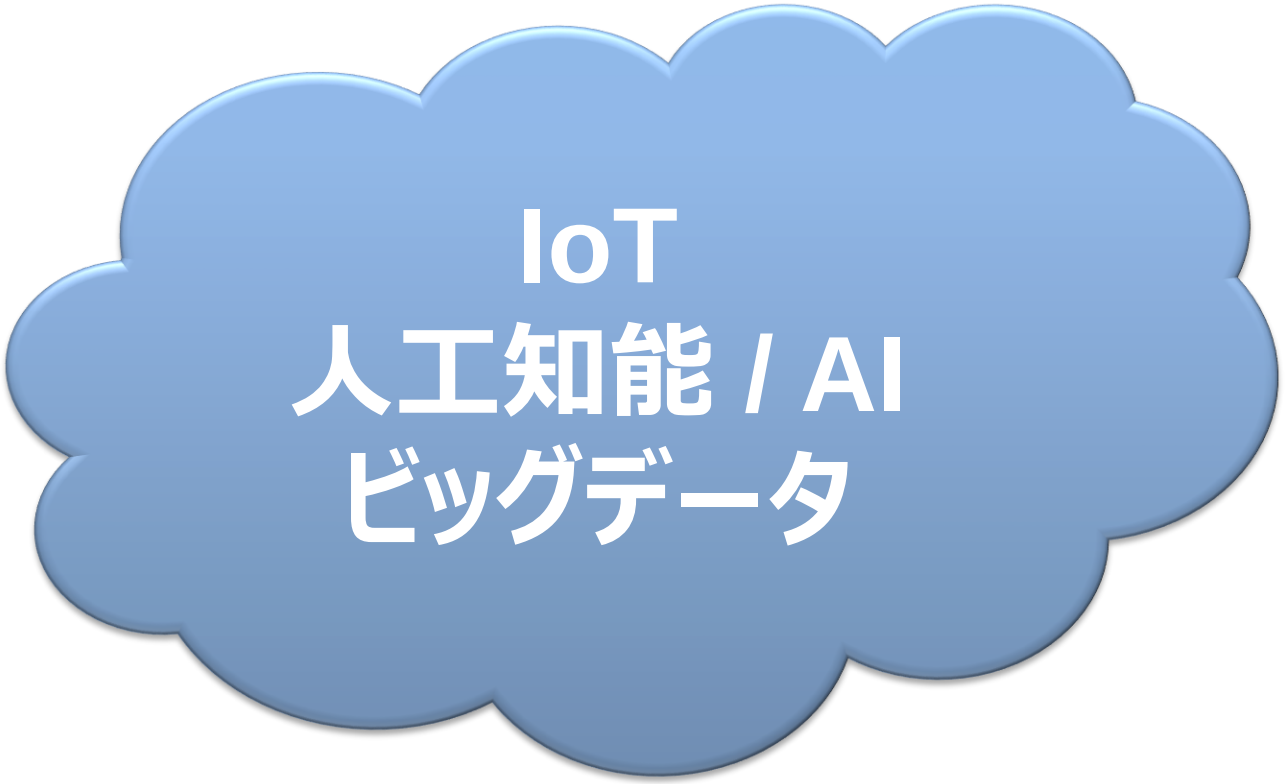
データの本質を読み解くための機械学習
～MATLAB でデータ解析の課題に立ち向かう～

MathWorks Japan
アプリケーションエンジニア部
アプリケーションエンジニア
井原 瑞希

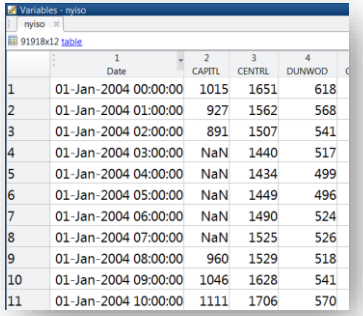
© 2016 The MathWorks, Inc.

Buzzwords…

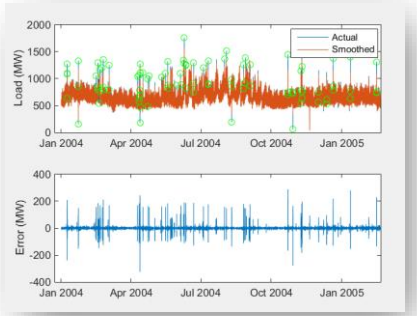


データ解析

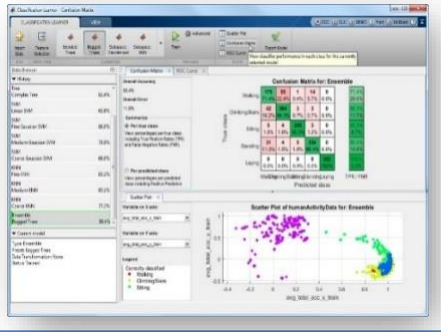
データ解析ワークフロー



データへのアクセスと探索



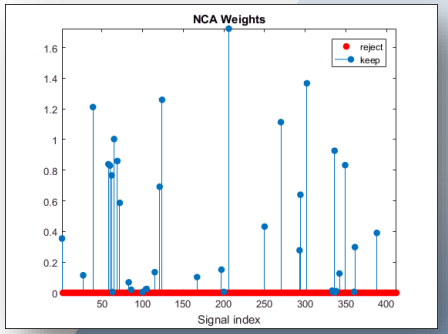
データの前処理



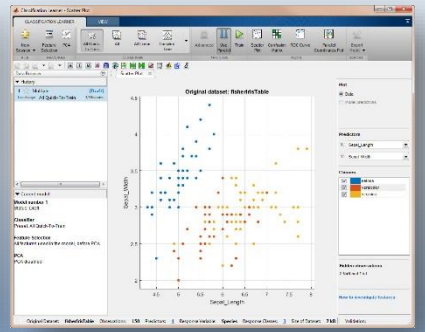
予測モデルの構築

システムへの統合

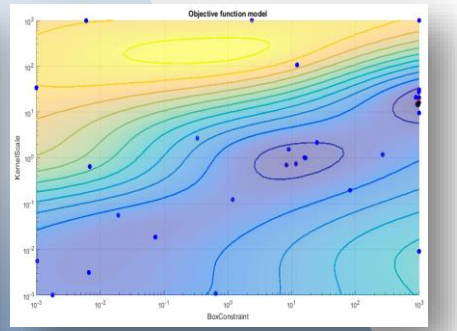
機械学習



特徴選択

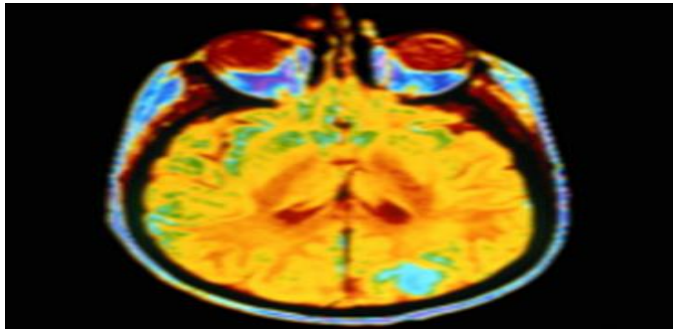
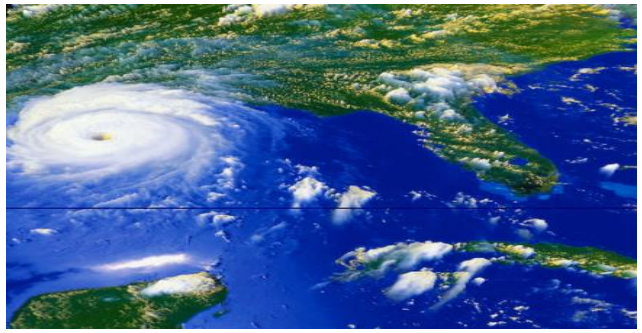
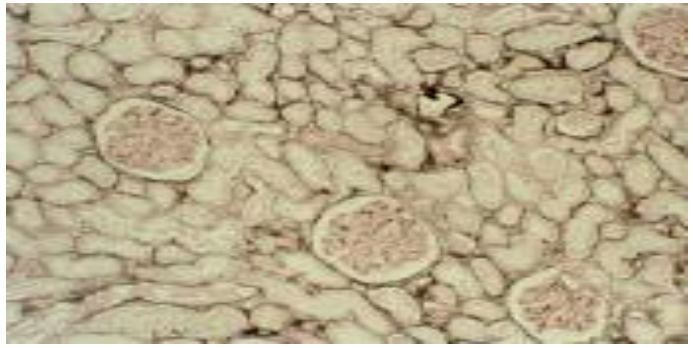


モデルの選択



ファインチューニング

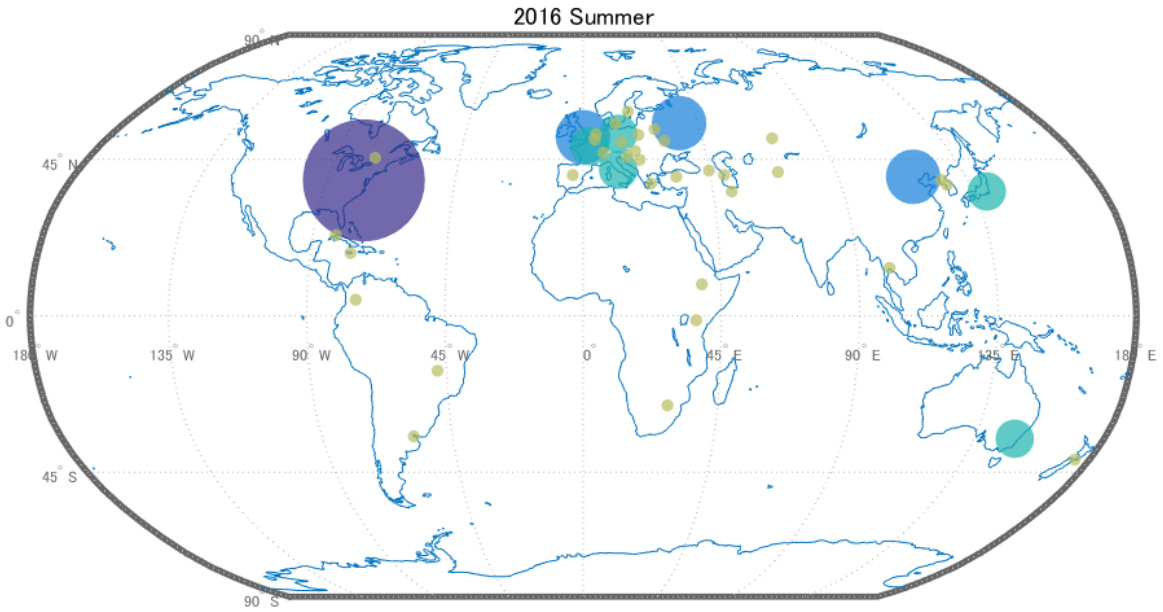
Machine Learning is Everywhere!

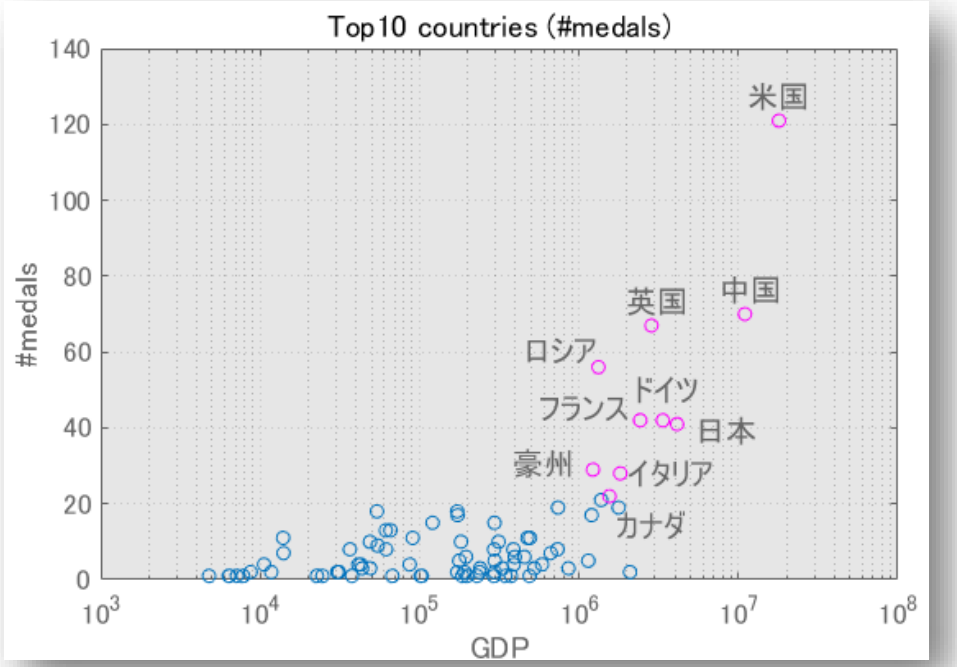








機械学習とは

- 機械学習 ”ではない” データ解析
  - データそのものから直接わかること



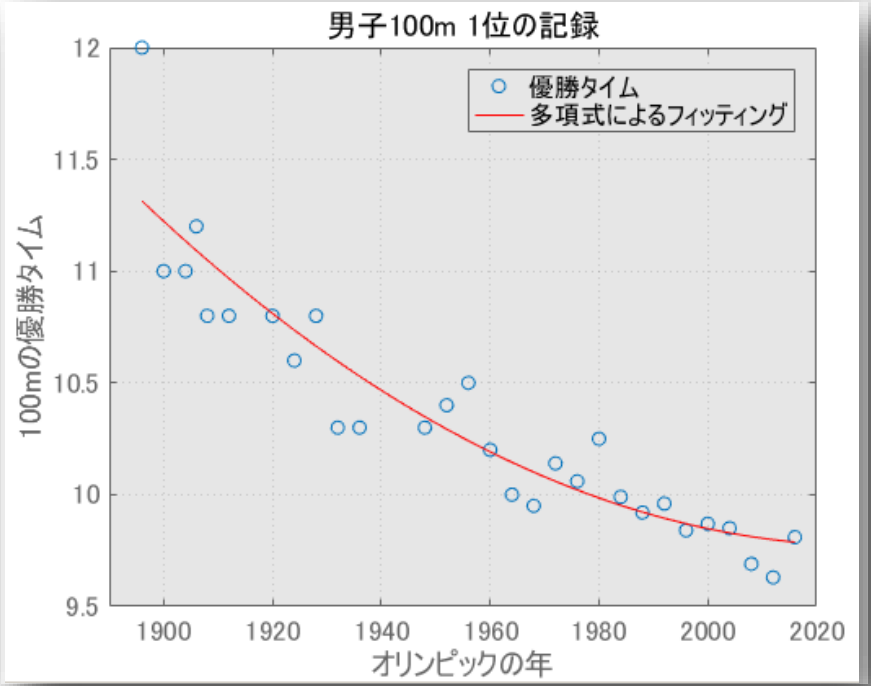



```
>> SimpleDALive
```

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**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



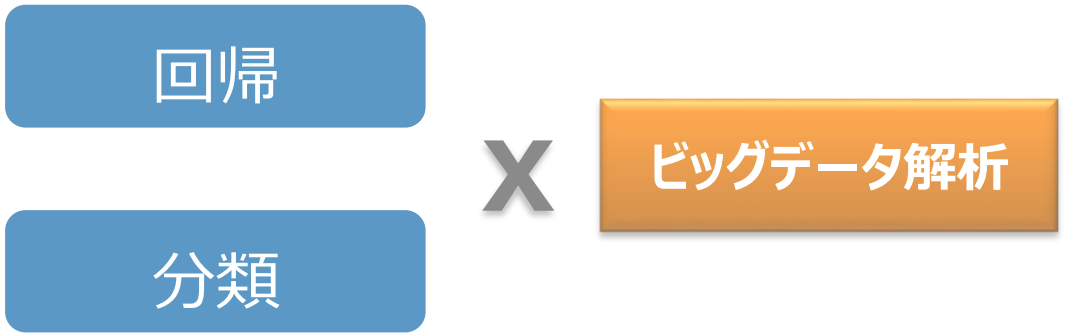
1940年、2020年のそれぞれの男子100mの1位の予想タイム

```
fitresult(1940)

ans = 10.4666

fitresult(2020)

ans = 9.7795
```

機械学習とは

- 機械学習
  - 教師なし学習（ラベルなし）
    - クラスタリング
  - 教師あり学習（ラベルあり）
    - 回帰（連続値）
    - 分類（離散値）

本日のトピック



- **回帰分析**
  - 回帰分析ワークフロー
  - 回帰モデルの見方
  - サンプル数が少ないときの回帰分析
- **分類**
  - 分類のワークフロー
  - コーティングを簡単にする方法
  - 機械学習の課題 – ハイパーパラメータの探索
- **機械学習によるビッグデータ解析**
  - メモリに収まりきらないビッグデータを扱う場合

本日のトピック

回帰

分類

X

ビッグデータ解析

- **回帰分析**
  - 回帰分析ワークフロー
  - 回帰モデルの見方
  - サンプル数が少ないときの回帰分析
- **分類**
  - 分類のワークフロー
  - コーティングを簡単にする方法
  - 機械学習の課題 – ハイパーパラメータの探索
- **機械学習によるビッグデータ解析**
  - メモリに収まりきらないビッグデータを扱う場合

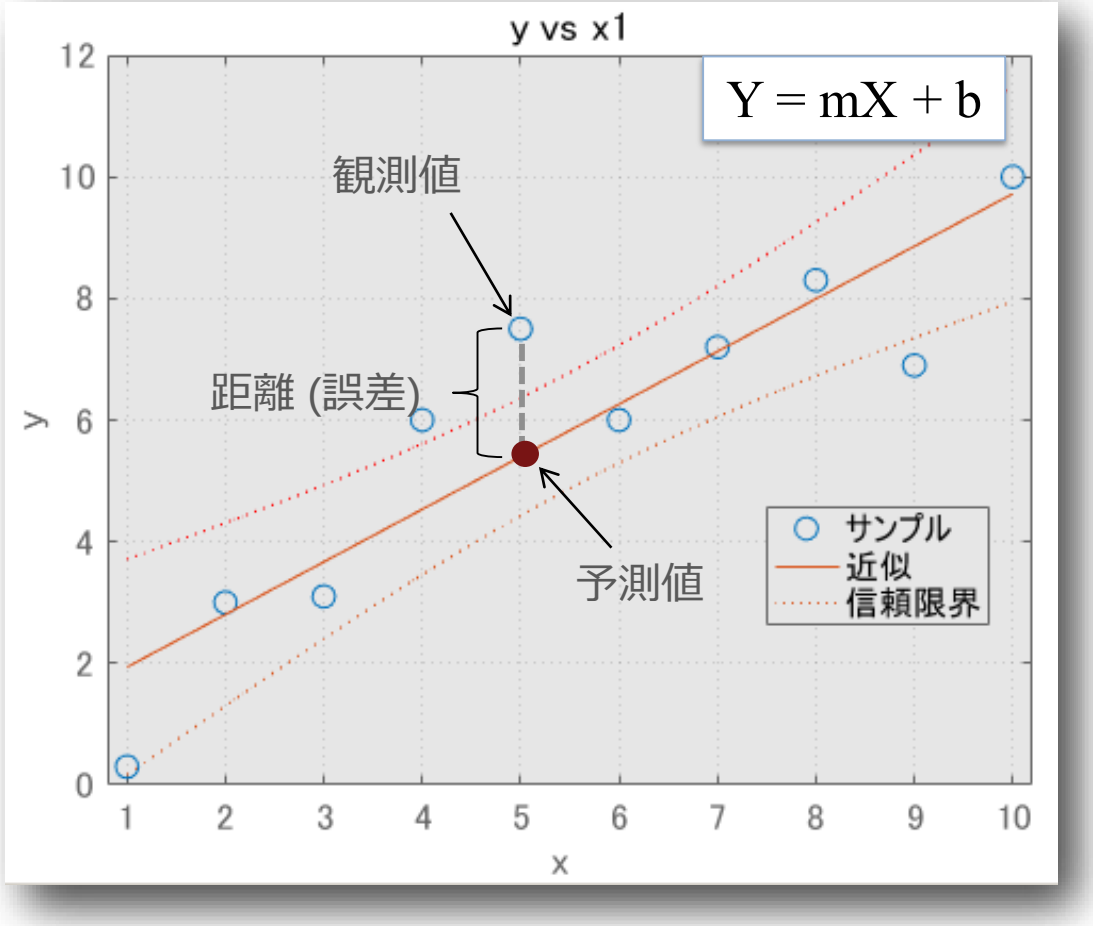
回帰分析とは

- 観測可能なデータから変数間の関係性をモデリング
- **説明変数 X** の関数として**目的変数 Y** を説明する変数の関係性モデルを構築

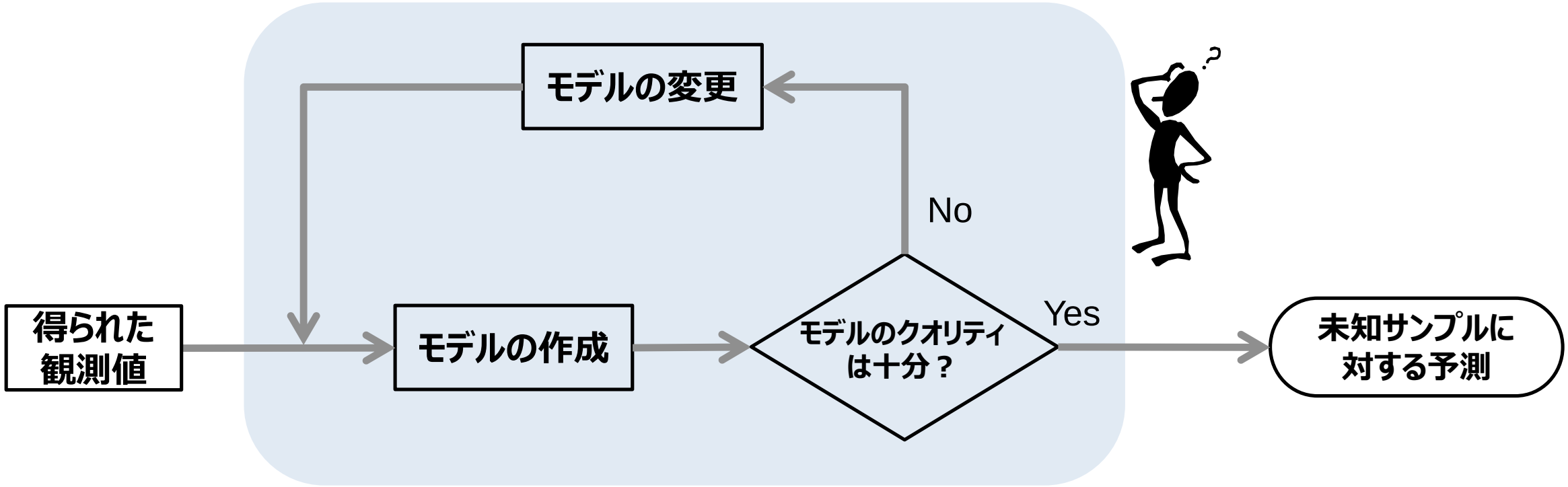
$$Y = f(X)$$

- 観測値と予測値の差を最小化する係数を推定
- 入力 X の未知の出力 Y を予測可能

例: 二乗誤差の和を最小化

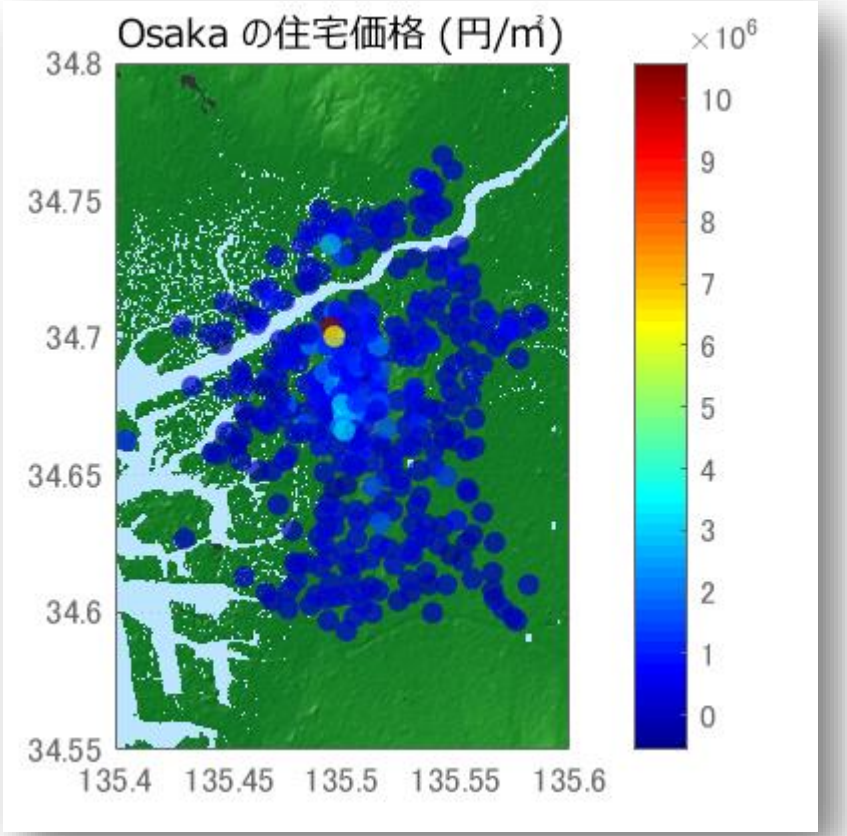


回帰分析の流れ



例: 大阪の住宅価格の推定

- 推定対象 (目的変数)
  - 住宅価格
- 既知の情報 (説明変数)
  - 位置情報
  - 部屋の形状
  - 地積
  - 階数
  - 駅からの距離など
- 目的
  - どの説明変数が住宅価格に関わっているか調べる
  - 正確な住宅価格予測モデルの作成



商業施設の場所の決定、経済指標として利用

```
>> HousingPriceEst_Osaka
```

回帰モデルの推定結果

- 線形回帰モデル:

  y ~ 1 + x1

- 推定された係数:

t値: 説明変数が与える影響

p値: 係数の有意確率
(極端な値を取る確率)

	Estimate	SE	tStat	pValue
(Intercept)	-3749.5	51.154	-73.298	1.188e-21
x1	2.0608	0.025507	80.794	2.5117e-22

観測数: 18、誤差の自由度: 16

二乗平均平方根誤差: 0.561

決定係数: 0.998、自由度調整済み決定係数 0.997

F 統計量と一定のモデルの比較: 6.53e+03、p 値は 2.51e-22

説明変数が、目的変数の変化
をどれくらい説明できているか

回帰の種類

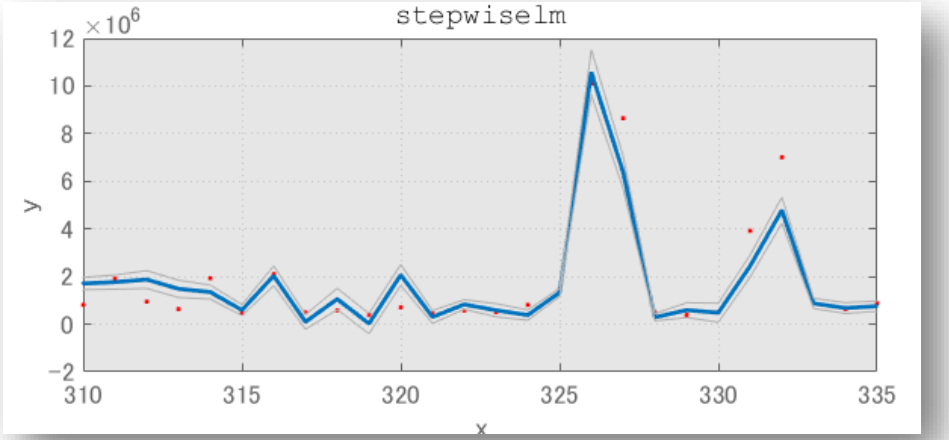
- パラメトリック回帰
  例) 線形回帰、ステップワイズ回帰、…
  - モデル式を仮定して、データにフィットするようなパラメタを探索
    データ = 確定的な成分 + ランダムな誤差
  - データの傾向や関数がある程度わかっている場合に有効

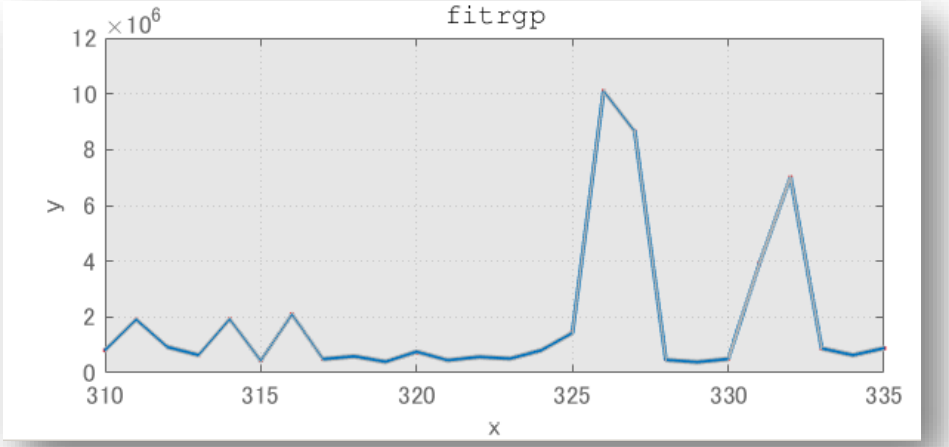
- ノンパラメトリック回帰
  - 例) ガウス過程回帰、決定木、…
  - 関数の形を定めない
  - 事前に関数がわからない場合や**観測可能なサンプル数が少ない場合に有効**

**学習サンプルが少なく
推定結果が良くない場合は？**

ガウス過程回帰

- ガウス過程回帰 (Gaussian Processes, Kriging)
  - ノンパラメトリックな**確率**モデル
  - 訓練データに近ければ分散が小、離れると分散大のガウス分布から確率的に生成されていると仮定
    （近傍とのなめらかな遷移を仮定）
  - 利点
    - サンプル数が少ない場合にも高い予測精度
    - 途中でサンプルの傾向が変わった場合に対応可能
  - 欠点
    - 高次元のデータでは予測精度が高くない

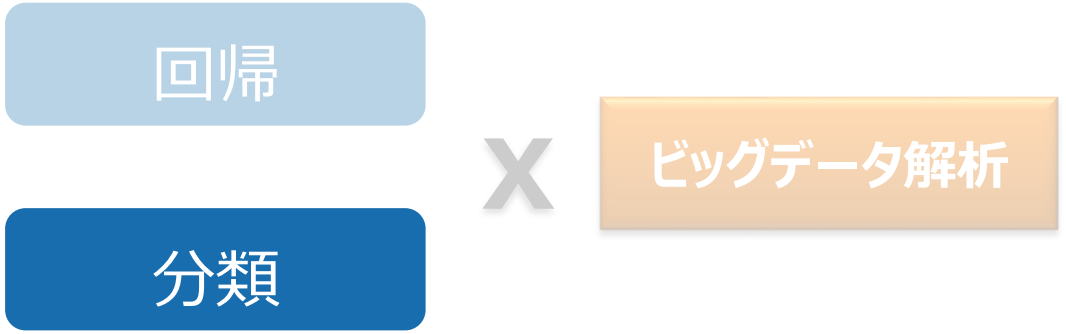




回帰分析のまとめ

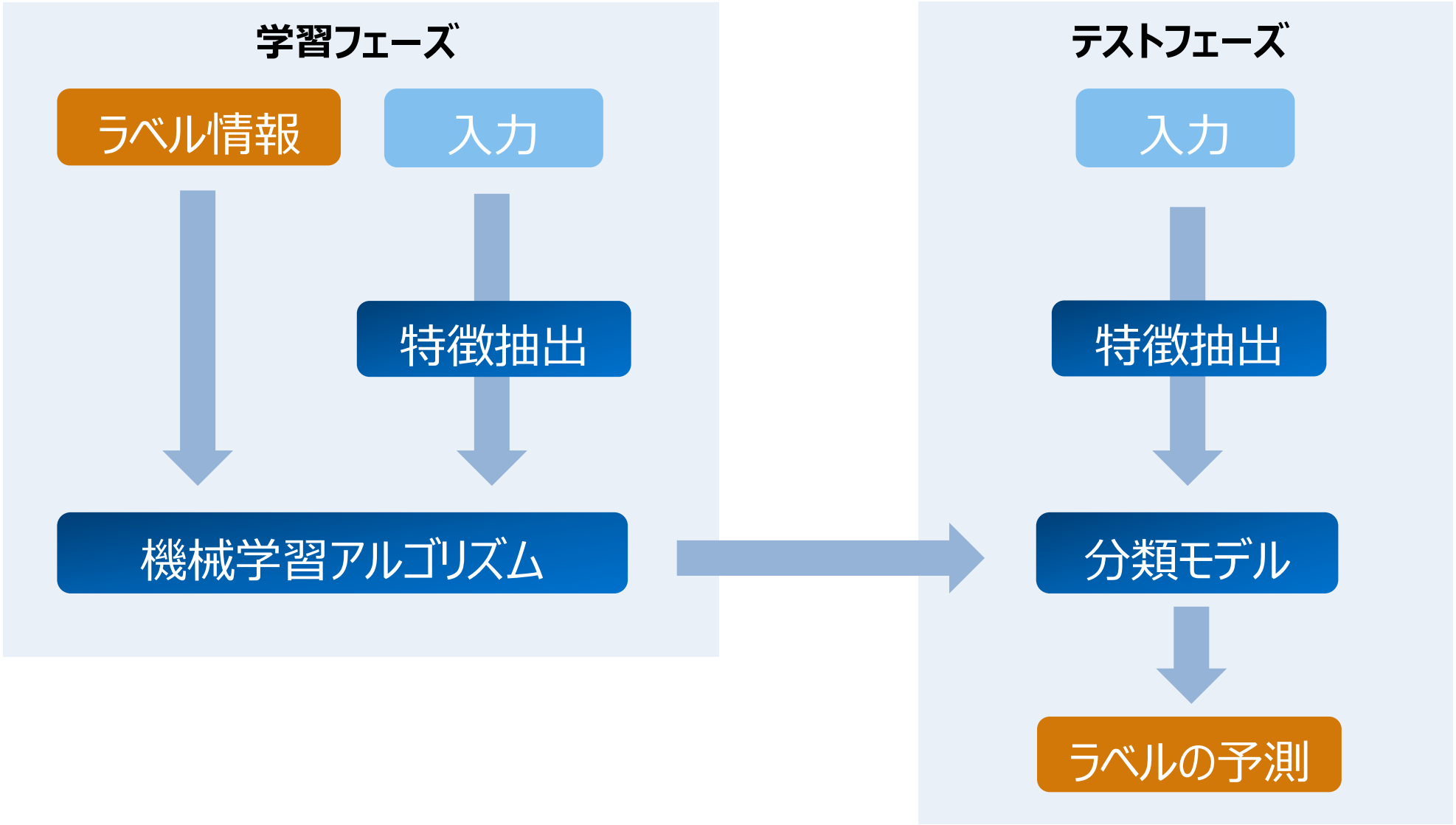
モデル		フィッティング関数	クラス名
パラメトリック回帰	線形回帰	fitlm	LinearModel
	一般化線形回帰	fitglm	GeneralizedLinearModel
	非線形回帰	fitnlm	NonLinearModel
ノンパラメトリック回帰	サポートベクタ回帰	fitrsvm	RegressionSVM
	ガウス過程回帰	fitrgp	RegressionGP
	回帰木	fitrtree	RegressionTree
	アンサンブル学習 (回帰)	fitensemble	RegressionEnsemble
	ニューラルネットワーク	train	---

本日のトピック



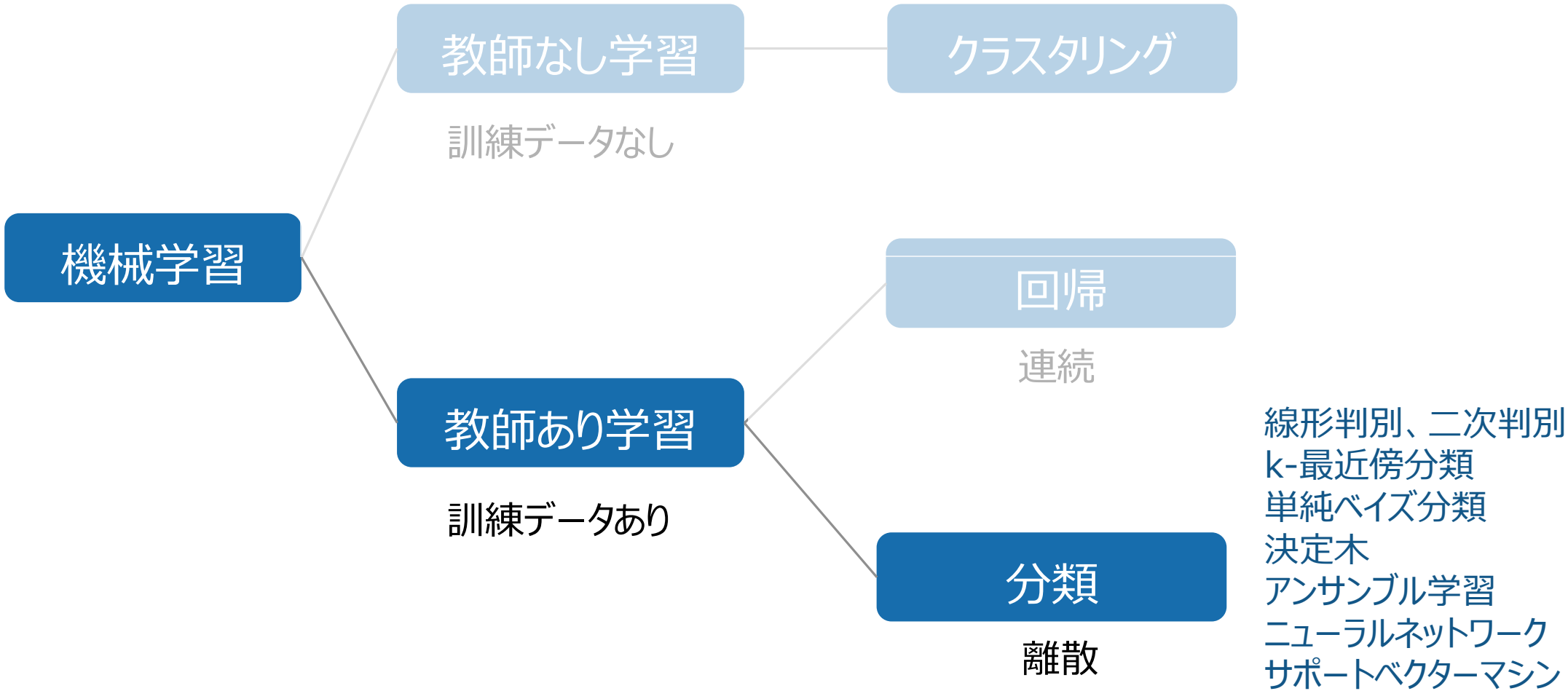
- **回帰分析**
  - 回帰分析ワークフロー
  - 回帰モデルの見方
  - サンプル数が少ないときの回帰分析
- **分類**
  - 分類のワークフロー
  - コーティングを簡単にする方法
  - 機械学習の課題 – ハイパーパラメータの探索
- **機械学習によるビッグデータ解析**
  - メモリに収まりきらないビッグデータを扱う場合

分類の流れ


学習フェーズ
ラベル情報
入力
特徴抽出
機械学習アルゴリズム
テストフェーズ
入力
特徴抽出
分類モデル
ラベルの予測

MATLAB® における機械学習

Statistics and Machine Learning Toolbox
Neural Network Toolbox



アプリを使った機械学習のコード作成

MathWorks R2015a

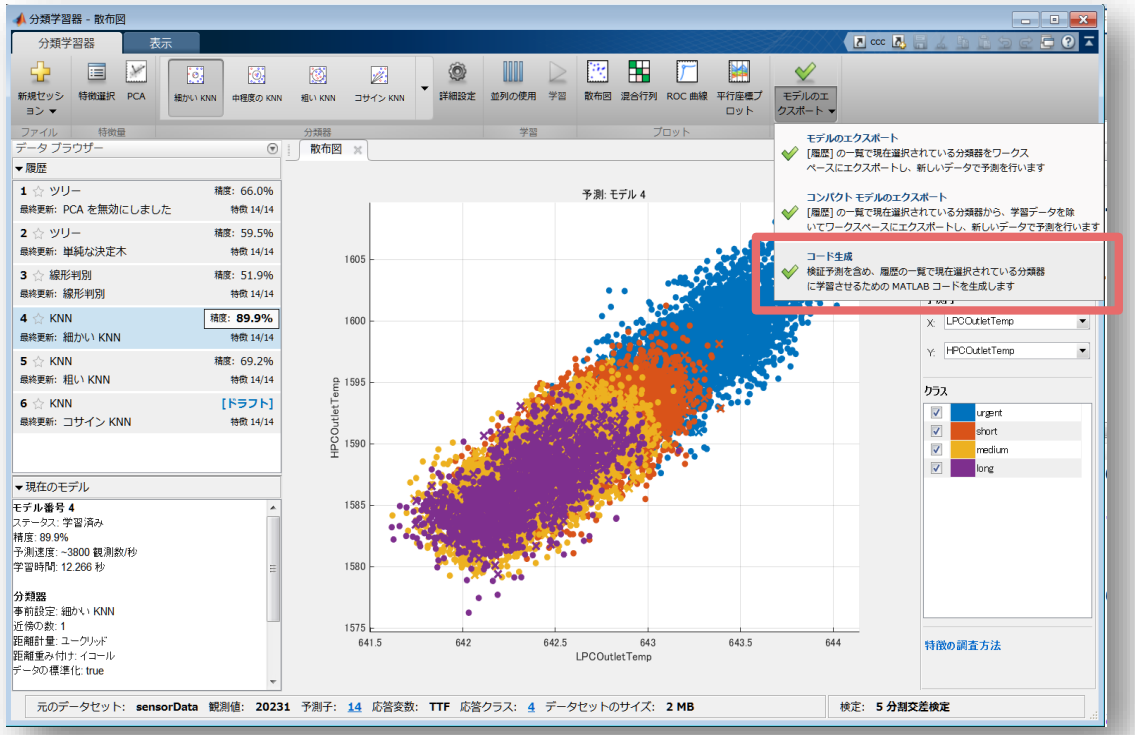
Statistics and Machine Learning Toolbox

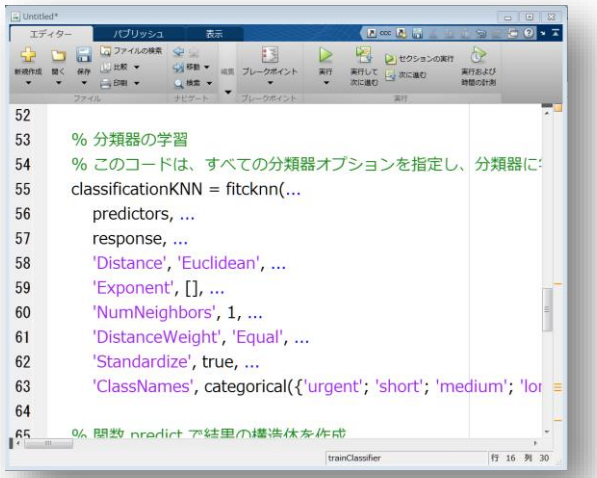
- 分類学習器アプリ
  - データを分類するためのモデル学習 GUI
  - GUI 操作を MATLAB コードとして生成することができる

Parallel Computing Toolbox™

並列モデル学習機能の追加

R2016b





MATLAB プログラムの自動生成

例: New York のタクシーチップカテゴリの分類

- 目的
  - チップの多い乗客の傾向を調べる
- 使用するデータ
  - New York のタクシー利用履歴
    (乗車の時間、人数、位置など)
  - csv ファイル
- 解析ワークフロー
  - 前処理
  - 解析に使用する特徴の選択
  - チップカテゴリ分類モデルの構築 (**分類学習器アプリ**の使用)
  - チップカテゴリ分類モデルの評価


```
>> TaxiTipClassification
```

分類器の種類と使い分け

分類器	すべての予測子が数値	すべての予測子がカテゴリカル	一部がカテゴリカル、一部が数値
決定木	あり	あり	あり
判別分析	あり	なし	なし
ロジスティック回帰	あり	あり	あり
SVM	あり	あり	あり
最近傍	ユークリッド距離のみ	ハミング距離のみ	なし
集団	あり	はい。ただし、部分空間判別を除く	はい。ただし、部分空間判別を除く

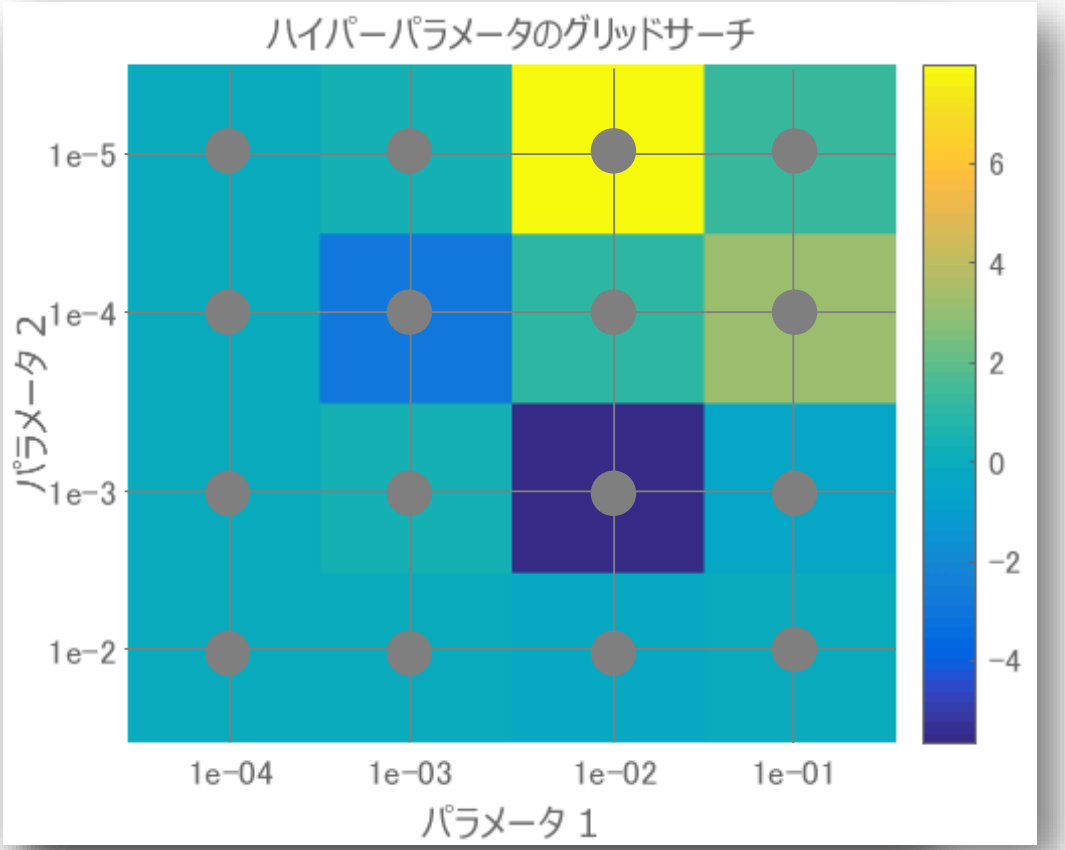
分類器	予測速度	メモリ使用量	解釈可能性
決定木 	高	小	容易
判別分析 	高	線形の場合は小、二次の場合は大	容易
ロジスティック回帰 	高	中	容易
サポート ベクター マシン 	線形の場合は中。他の場合は低	線形の場合は中。他のすべて: マルチクラスの場合は中、バイナリの場合は大	線形 SVM の場合は容易。他のすべてのカーネル タイプの場合は困難。
最近傍分類器 	3 次元の場合は低。他の場合は中	中	困難
アンサンブル分類器 	アルゴリズムの選択によって高から中	アルゴリズムの選択によって低から高	困難

分類器オプションの選択
http://jp.mathworks.com/help/stats/choose-a-classifier.html

機械学習手法における課題

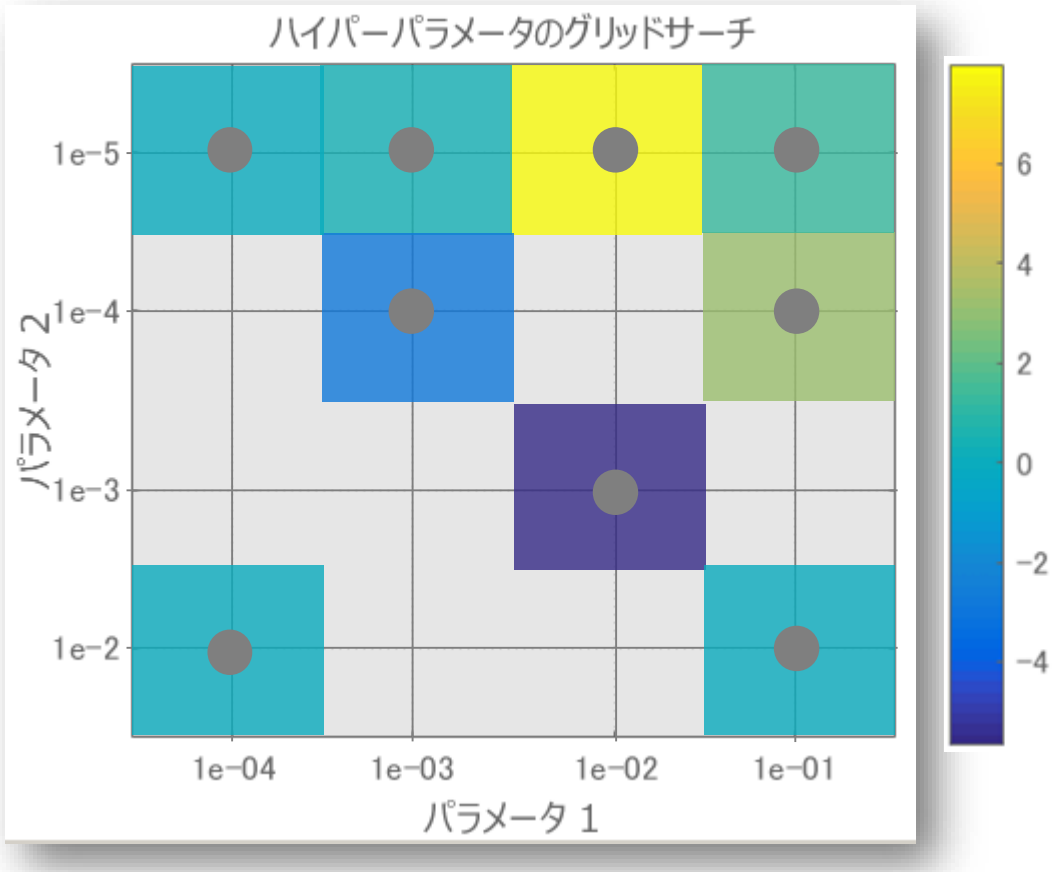
- 課題
  - ハイパーパラメータの調節
- ハイパーパラメータ
  - データから直接決めることのできないパラメータ
  - ユーザがあらかじめ決めておく必要あり

- **グリッドサーチ** (従来の方法)
  - 格子状の空間で最適なパラメータを探索
  - 課題: ある範囲を総当りするため
    計算時間がかかる



機械学習手法における課題: ハイパーパラメータの調節

- **ベイズ最適化** R2016b
  - あるハイパーパラメータでの学習器の精度を目的関数として定義
  - この目的関数を最大化するパラメータを推定
    - ガウス過程回帰でモデル化
    - 精度が上がりやすそうな方向を確率的に推定
  - 条件
    - 低次元データ
    - 目的関数の評価に時間がかかる
    - 低精度
    - 大域的な解を求めたい
    - ハイパーパラメータの決定



ベイズ最適化によるパラメータチューニング

MathWorks
R2016b
Statistics and Machine Learning Toolbox

- 機械学習アルゴリズムのハイパーパラメータ推定の自動化

1. “**fit**” 関数の **OptimizeHyperparameters** オプションを追加

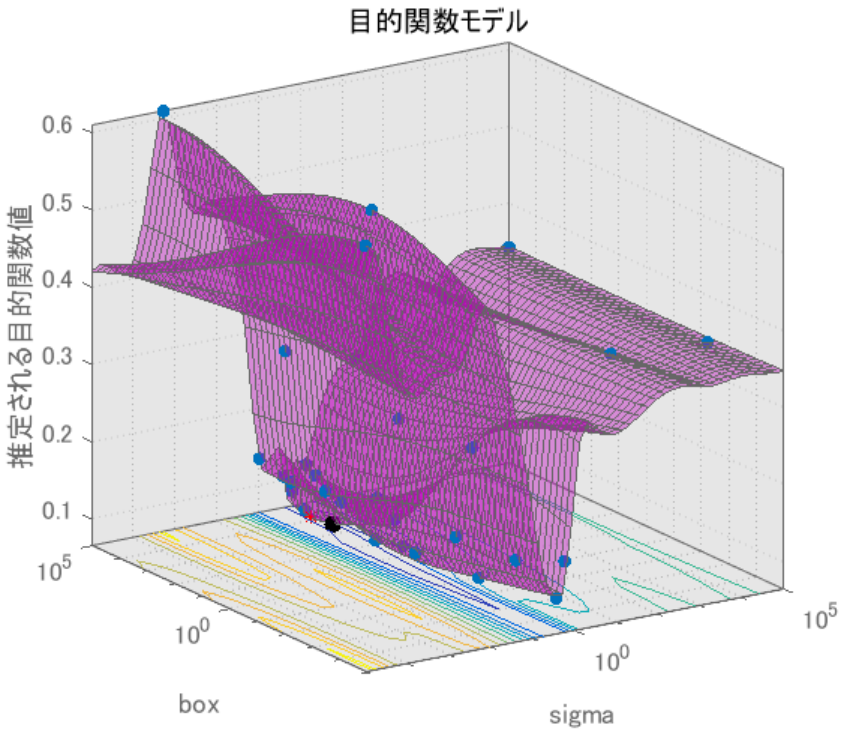
   パラメータ値固定

```
ctree = fitctree(x,y,'MaxNumSplits', 100);
```

   ベイズ最適化によるパラメータ決定

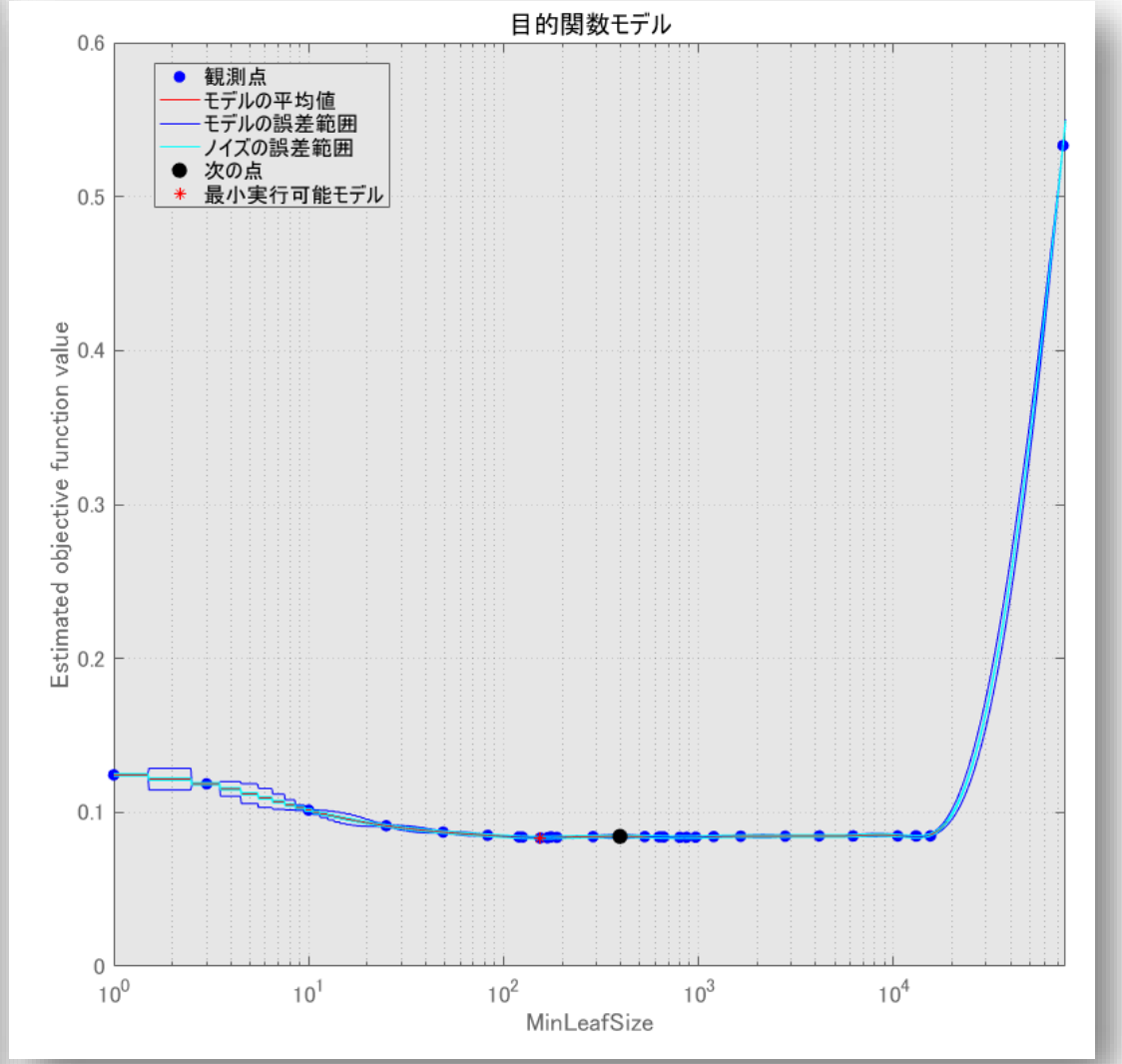
```
ctree_BO = fitctree(x,y,...
    'OptimizeHyperparameters', 'auto');
```

2. 目的関数を指定して **bayesopt** 関数を使用
   - 定義した制約内で最適化

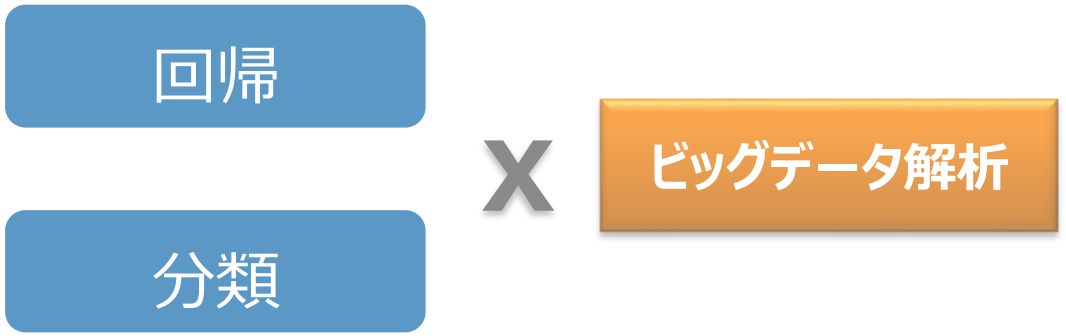


ベイズ最適化最終結果

```
最適化が完了しました。
MaxObjectiveEvaluations の 30 に達しました。
関数評価数の合計: 30
経過時間の合計: 331.7751 秒。
目的関数の評価時間の合計: 312.013

最適な観測実行可能点:
    MinLeafSize
    ___________
    154
観察された目的関数値 = 0.083359
推定される目的関数値 = 0.083665
関数の評価時間 = 10.7994

最適な推定実行可能点 (モデルに基づく):
    MinLeafSize
    ___________
    168
推定される目的関数値 = 0.083665
推定される関数評価時間 = 10.9553

trainedClassifier_BO =
  フィールドをもつ struct:

            predictFcn: [function_handle]
     RequiredVariables: {1×15 cell}
    ClassificationTree: [1×1 ClassificationTree]
                 About: '…'
        HowToPredict: '…'
validationAccuracy_BO =
    0.9164
```



本日のトピック

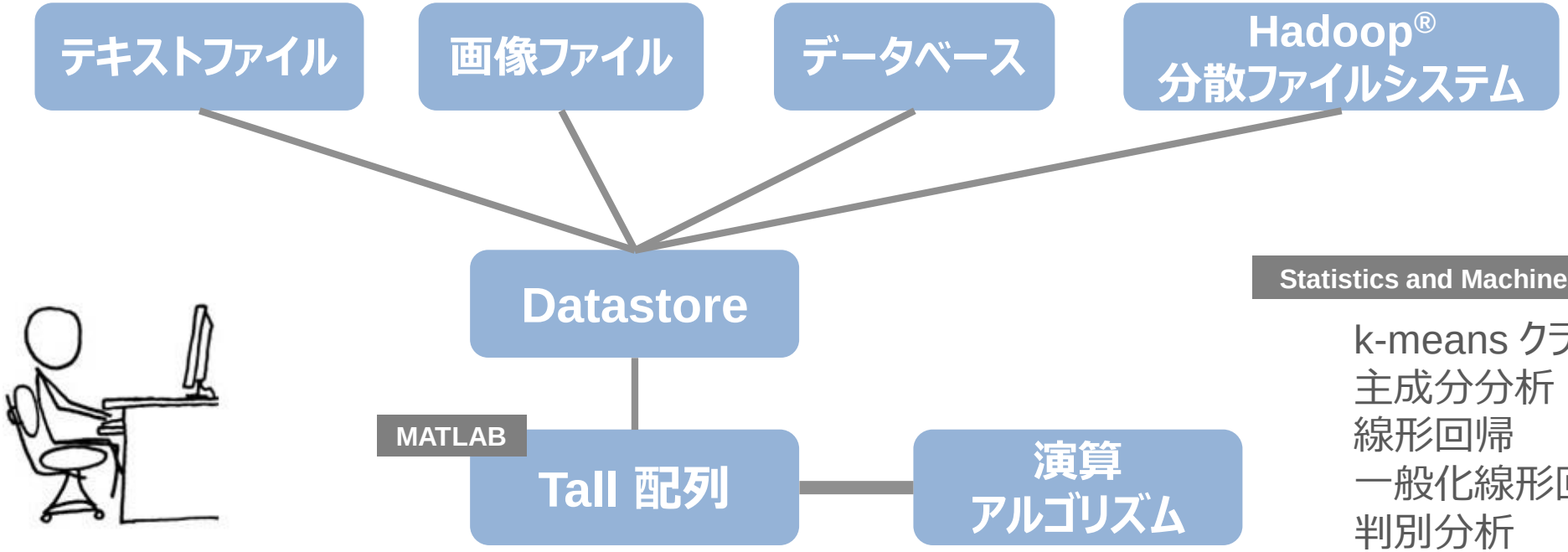


- **回帰分析**
  - 回帰分析ワークフロー
  - 回帰モデルの見方
  - サンプル数が少ないときの回帰分析
- **分類**
  - 分類のワークフロー
  - コーティングを簡単にする方法
  - 機械学習の課題 – ハイパーパラメータの探索
- **機械学習によるビッグデータ解析**
  - メモリに収まりきらないビッグデータを扱う場合

メモリに収まらないデータの扱い

user operation

- `tall`: メモリに収まりきらないデータに対するデータ型



Tall Array 対応関数リスト
https://jp.mathworks.com/help/releases/R2016b/matlab/import_export/functions-that-support-tall-arrays-by-type.html

例: New York のタクシー料金の予測モデルの作成 (回帰分析)

- 使用するデータ
  - 25GB の csv ファイル
- 解析ワークフロー
  - 解析に使用する特徴の選択
  - 前処理
  - データの探索
  - 料金予測モデルの構築
  - 料金モデルの評価

```
>> TallArrayDemo
```

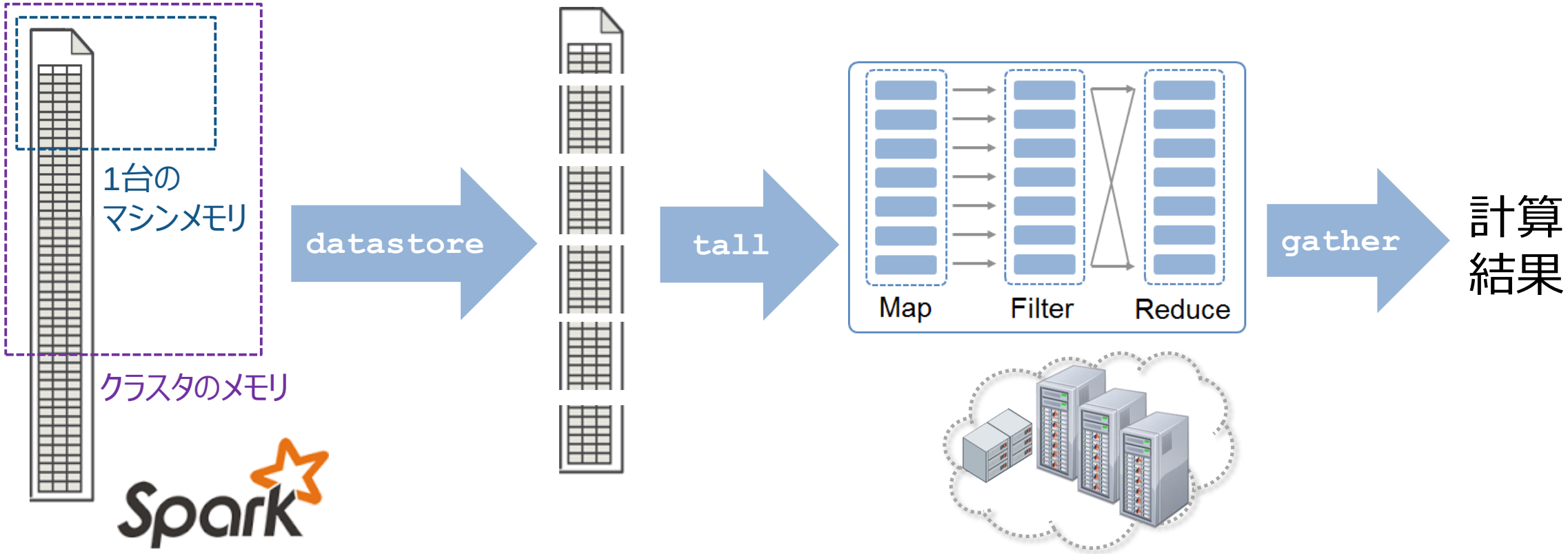
メモリに収まらないデータの扱い

Tall Array のしくみ

MathWorks R2016b

Parallel Computing Toolbox
MATLAB Distributed Computing Server™

- `gather` 関数で初めてすべての演算が実行
  - 演算をまとめることでデータへのアクセスを最適化 (遅延評価)



本日のトピック

回帰

分類

X

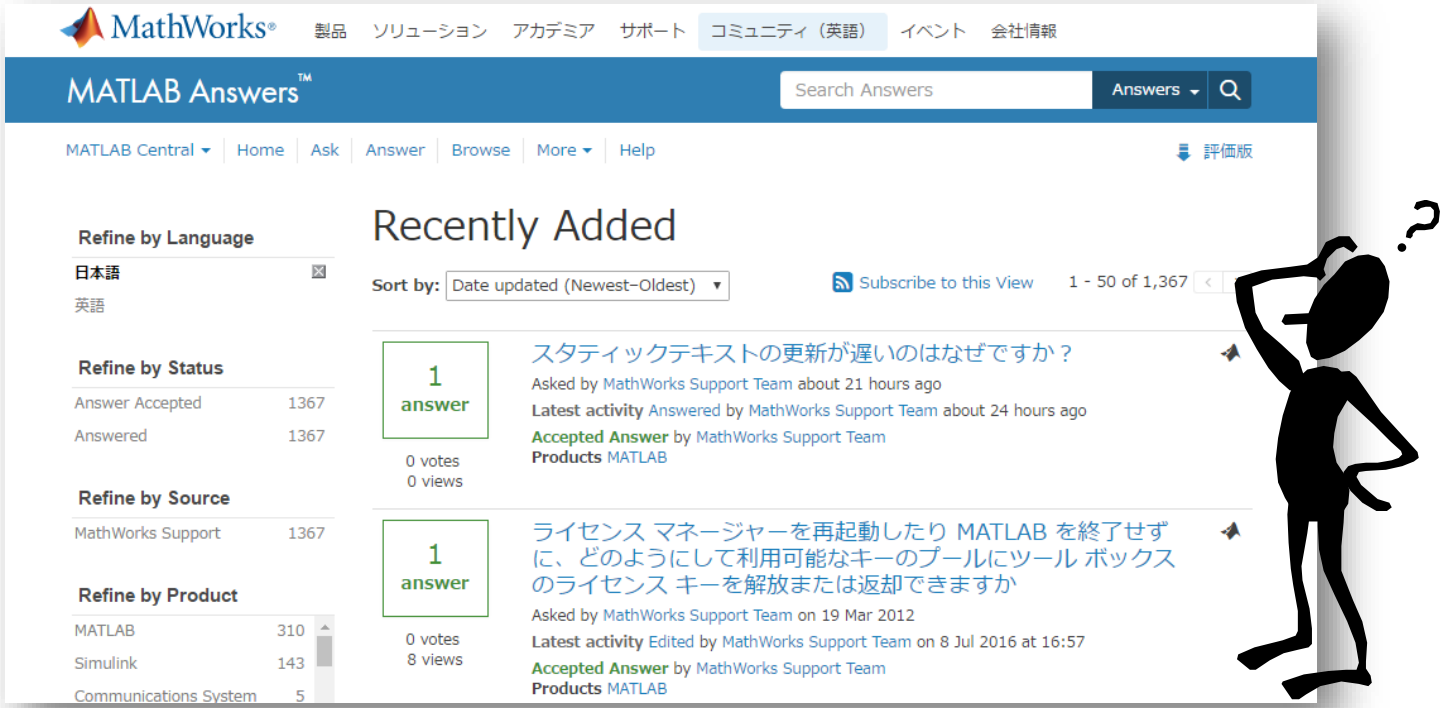
ビッグデータ解析

- **回帰分析**
  - 回帰分析ワークフロー
  - 回帰モデルの見方
  - サンプル数が少ないときの回帰分析: **ガウス過程回帰 R2015b**
- **分類**
  - 分類のワークフロー
  - コーティングを簡単にする方法: **分類学習器アプリ R2015a**
  - ハイパーパラメターの探索: **ベイズ最適化 R2016b**
- **機械学習によるビッグデータ解析**
  - メモリに収まりきらないビッグデータを扱う場合: **Tall 配列 R2016b**

機械学習のトレーニングもあります！

MATLAB Answers
日本語/英語の Q&A サイト

- MATLAB に関する過去の質問＆回答が閲覧可能
- MathWorks アカウントがあれば、誰でも投稿できます！
- 日本語/英語両方に対応
- 得意な分野の質問への回答に挑戦してみませんか？



http://jp.mathworks.com/matlabcentral/answers

© 2016 The MathWorks, Inc. MATLAB and Simulink are registered trademarks of The MathWorks, Inc. See www.mathworks.com/trademarks for a list of additional trademarks. Other product or brand names may be trademarks or registered trademarks of their respective holders.