機械学習　入門編

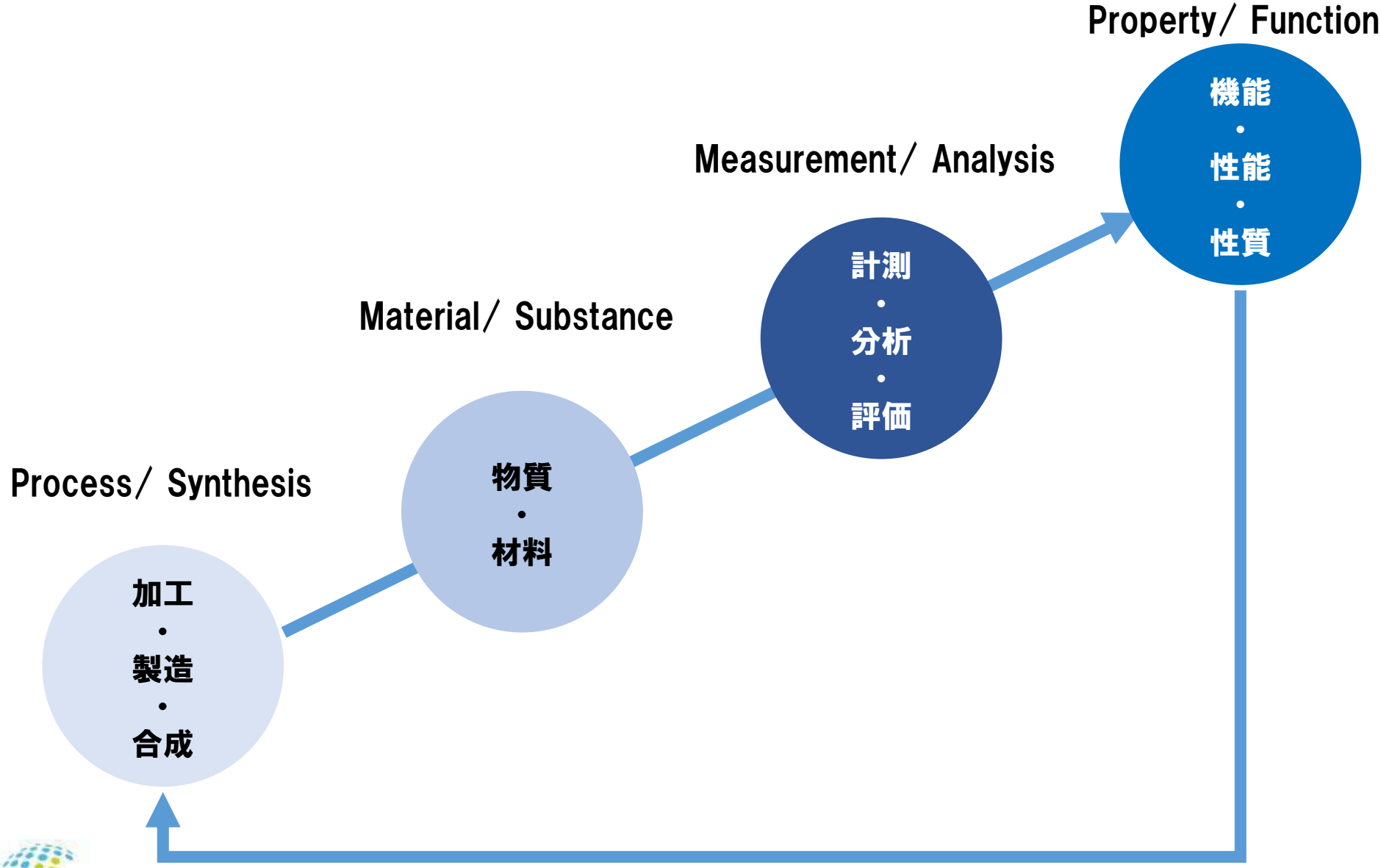
2024年12月4日　ARIMアカデミー　データ構造化ワークショップ（１）

物質・材料研究機構
マテリアル先端リサーチインフラセンターハブ

松波　成行

１．実験系の「データ」とは？

実験系における4つのデータのタイプ



生分解性ポリマーの研究に適用させると

Property/ Function

強度
・
成形性
・
分解性

Measurement/ Analysis

分子量
・
Tg
・
引張強度

Material/ Substance

ポリエーテル系
・
ポリエステル系
・
多糖類系

Process/ Synthesis

モノマー
・
触媒
・
温度/時間

https://tenbou.nies.go.jp/science/description/detail.php?id=54

農業(アグリテック)に適用すると

Property/ Function

甘い
・
健康的
・
大きい

Measurement/ Analysis

糖分
・
水分
・
ミネラル

Material/ Substance

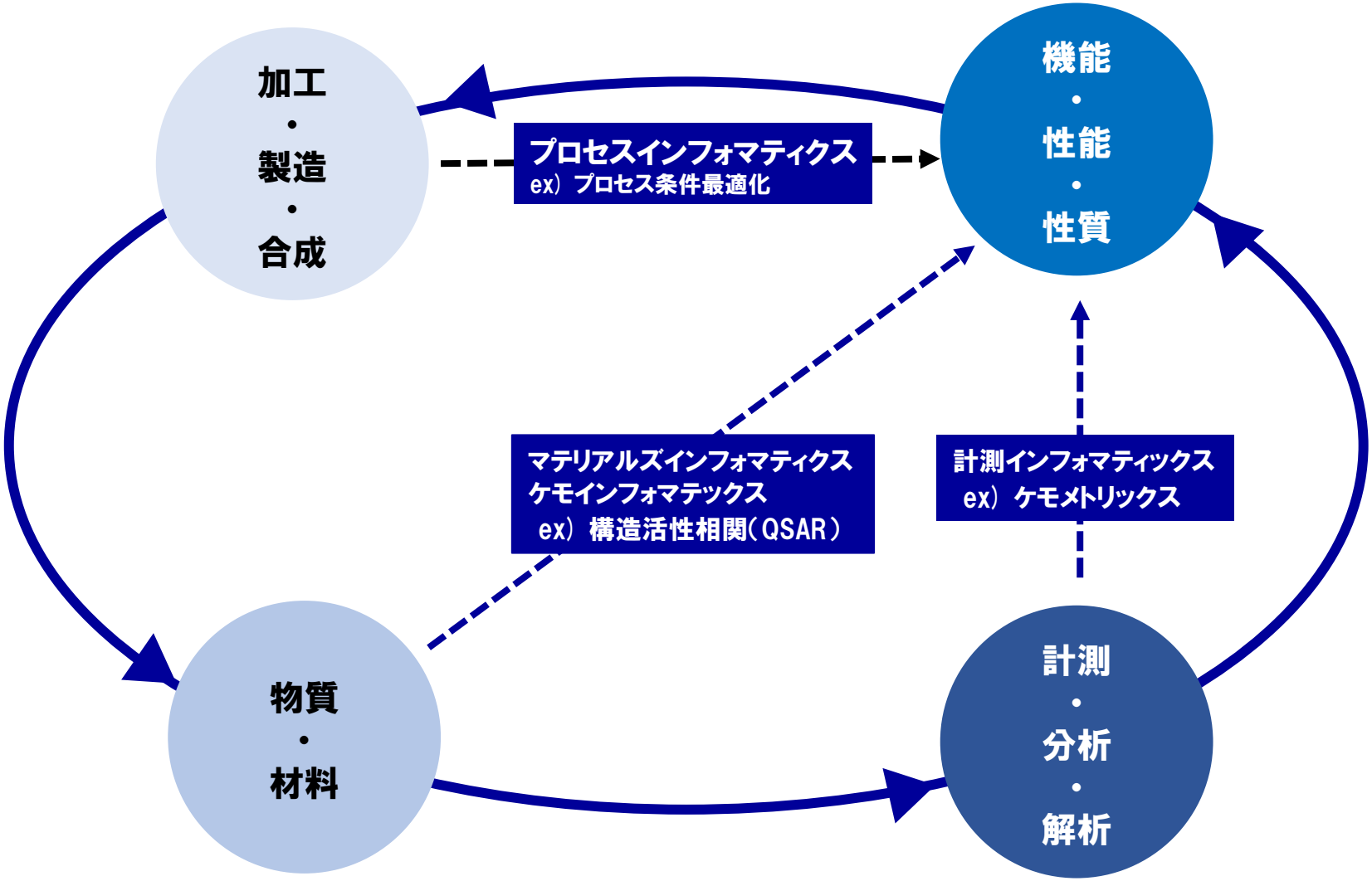
米
・
果物
・
野菜

Process/ Synthesis

土壌
・
肥料
・
天候

マテリアル開発サイクルとデータ活用

■ マテリアルの開発・評価ステージによって、データ構造化の設計は変化する。Informaticsを適用する技術分野ごとで要件を十分に吟味する。

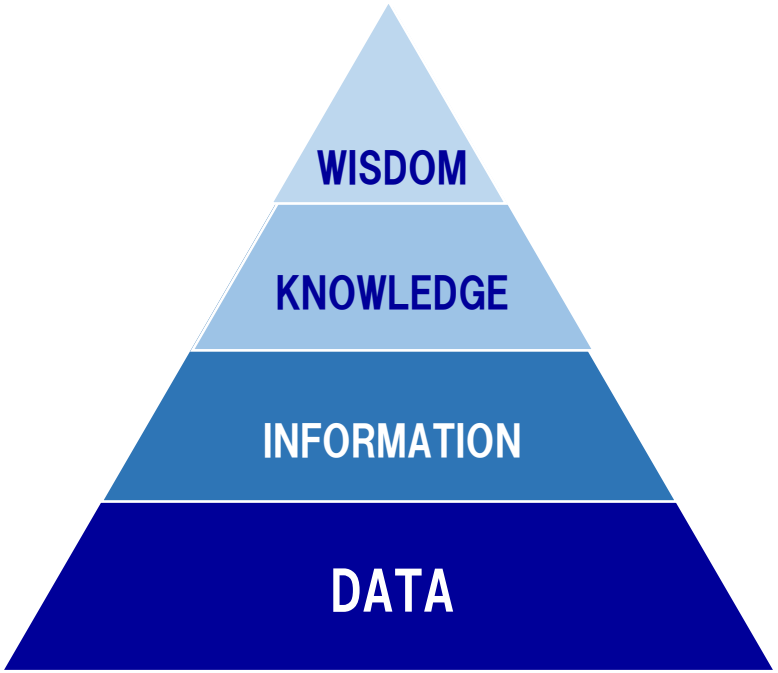


ARIM Japan

２．データの利活用の要諦

① データからパターン見つけること

②パターンを情報→知識→叡智へと昇華させること



叡智

智識

情報

データ

DIKW pyramid

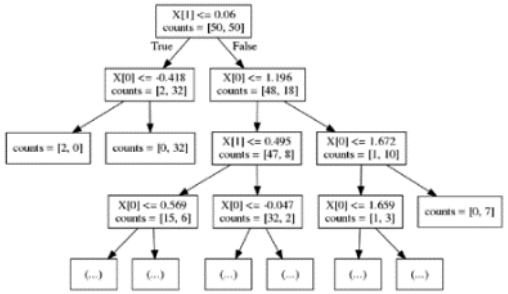
パターン認識からの帰納法的プロセス

３．パターン認識はどのように？

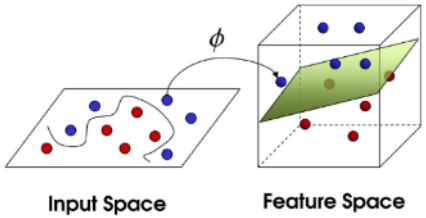
パターン認識で使われるモデル

- 線形型
  - 単回帰
  - 多変量解析
- 非線形型
  - 機械学習
  - 深層学習（画像認識・画像生成）

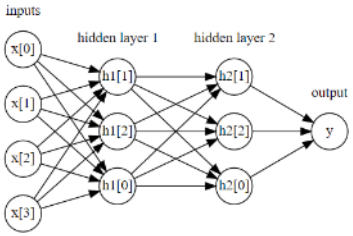
多変量解析・機械学習モデル



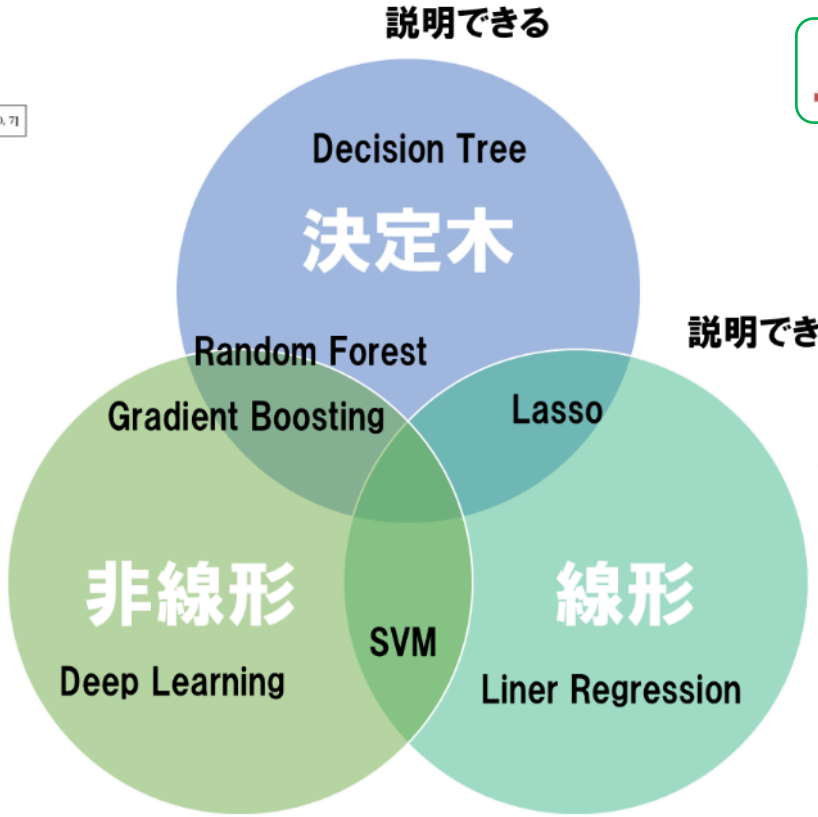
決定木

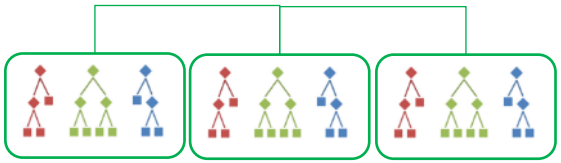


SVM



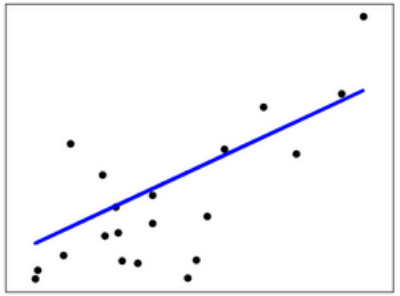
Deep learning





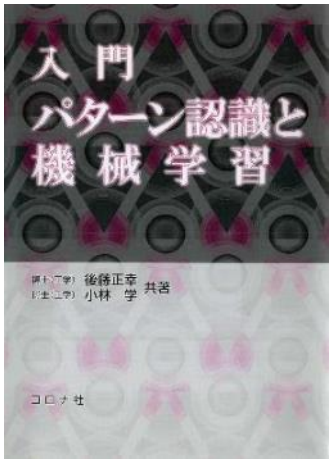
RandomForest
GradirntBoosting

$$y = b_1x_1 + b_2x_2 + b_3x_3 + \cdots + b_kx_k + e$$



線形モデル

機械学習（パターン認識）のための専門書



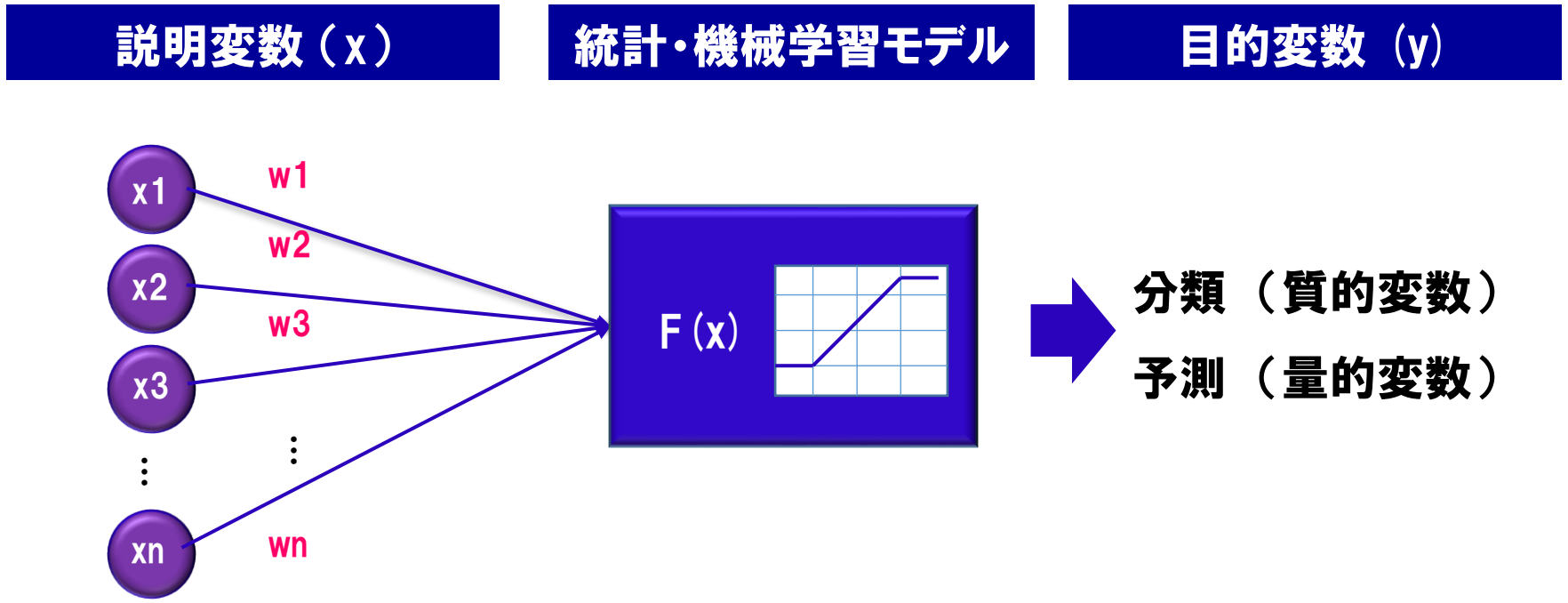
個別のアルゴリズムは専門書から学習してください

４．データの尺度と機械学習

様々なデータ尺度

	尺度名	定　義	例
質的変数（離散性）	名義尺度	分類のための単なる名前や識別ラベル	• 名前 • 性別 • 単語（文字列）
	順序尺度	順序関係を表す。 加減算が意味がない	• アンケートの5段階評価 • 世代区分 • カテゴリー区分
量的変数（連続性）	間隔尺度	一定の単位で量られた量。 原点はあっても「無」ではない 等間隔性がある 加減算が意味を持つ	•年月、時間 •試験の成績 •摂氏、華氏温度
	比例尺度	原点が定まっている 割り算（比）が意味をもつ	•身長や体重 •絶対温度

データ尺度とパターン認識



1 変数（特徴量）：どのような変数（特徴量）を選定するか

2 モデル：　どのようなアルゴリズム・関数を用いるか

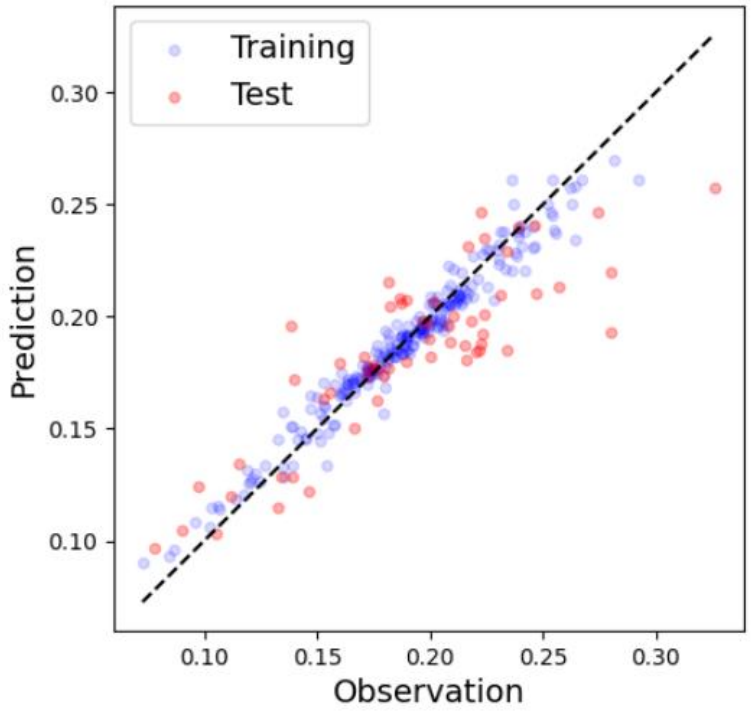
3 パラメータ：　どのようFitting係数・ハイパーパラメータを調整するか

分類の概要

	分　類	予　測
概　要	分類は、入力データを事前に定義されたクラスやカテゴリに分類するタスク。 データポイントを複数のグループに割り当てることを目的とします。	予測は、数値や連続値の予測を行うタスク。 与えられた入力に基づいて、数値の予測や連続値の予測を行います。
目的変数	質的変数	量的変数
使用例	メールが「スパム」または「非スパム」のどちらに分類されるかを判断する	住宅価格の予測や売上予測などが予測する
代表的な機械学習モデル	• ロジスティック回帰 • ランダムフォレスト • サポートベクターマシン • 人工ニューラルネットワーク など。	• 線形重回帰 • サポートベクターマシン • 決定木回帰 • 人工ニューラルネットワーク など。

予測型の機械学習モデルの構築と評価

識別機	ハイパーパラメータ
スパースモデル Rige, Lasso, ElasticNet	alpha(正則化パラメータ)alphaが大きいと単純なモデル分類
決定木	事前枝刈込　max_depth、max_leaf_nodes、min_samples_leaf
ランダムフォレスト	n_estimator(大きければよい)、max_features(小さいと過剰適合が低減される)、事前枝刈込のmax_depth
勾配ブースティング	n_estimator(大きいと複雑なモデルになり過剰学習になる)、learning_rate(決定木の誤りを訂正)、事前枝刈込のmax_depth
サポートベクターマシーン	C(正則化パラメータ)Cが小さいと単純なモデル、kernel (RBFではgammaも調整)
ニューラルネットワーク	hiddon_layer_size(隠れ層)activation(非線形活性化関数)、alpha(l2正則化)alphaが大きいと単純なモデル。



各種の識別機やハイパーパラメータ調整を行い、機械学習モデルを構築する

ARIM Japan

５．大隅先生の３ステップ

【再掲】これからの時代に生きるために

（大隅昇，統計数理研究所・名誉教授）

探索的データ科学のススメ

「目的にあったデータの取得方法」が必要。そのためのデータ主導型の解析過程が必要

考え方：

現象解析の本質は「データ」にある。データによる現象理解を前提として統計学、分類操作などを背景として統合的に現象解析をすすめる。

方法論：

① Experimental Design： データをどう計画的に取得するか

② Data Collection Mode： データを具体的にどう集めるか

③ Analyzing： 問題とする現象解析に適した解析法はどうあるべきか

① Experimental Design： データをどう計画的に取得するか

① Experimental Design： データをどう計画的に取得するか

→ **考えられる説明変数を考察することからはじまります。**

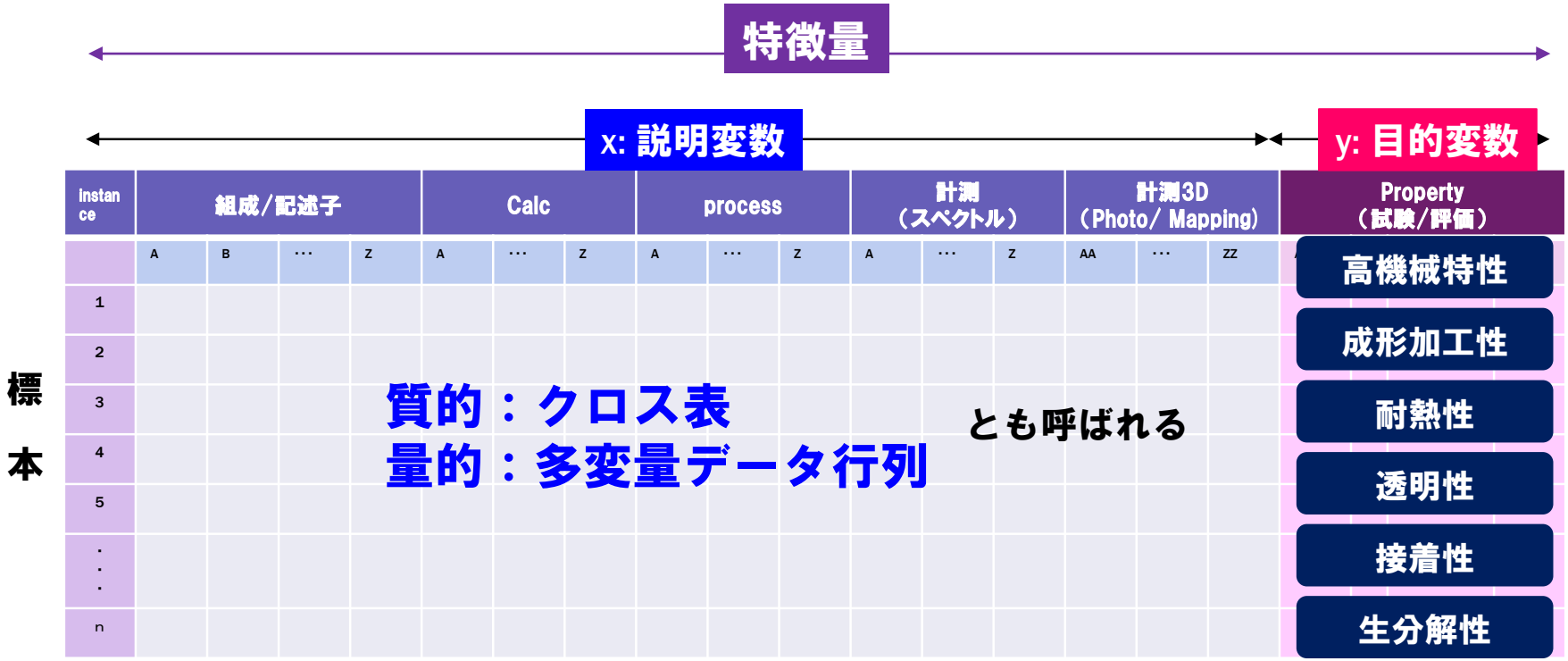
② Data Collection Mode： データを具体的にどう集めるか

→ 科学分野では、

- 既存のデータベース（商用DBを含む）
- 研究室の過去データ
- 実験装置の出力ファイルデータ　← 本題

③ Analyzing： 問題とする現象解析に適した解析法はどうあるべきか

① Experimental Design: 特徴量をつくる



質的変数、量的変数をまとめる表づくりは、
みなさんもExcelで普段行っているはずです

② Data Collection Mode： データを具体的にどう集めるか

① Experimental Design： データをどう計画的に取得するか

→ 考えられる説明変数を考察することからはじまります。

② Data Collection Mode： データを具体的にどう集めるか

→ 科学分野では、

- 既存のデータベース（商用DBを含む）
- 研究室の過去データ
- 実験装置の出力ファイルデータ ← ARIMが注力するところ

③ Analyzing： 問題とする現象解析に適した解析法はどうあるべきか

③ Analyzing： 問題とする現象解析に適した解析法はどうあるべきか

① Experimental Design： データをどう計画的に取得するか

→ 考えられる説明変数を考察することからはじまります。

② Data Collection Mode： データを具体的にどう集めるか

→ 科学分野では、

- 既存のデータベース（商用DBを含む）
- 研究室の過去データ
- 実験装置の出力ファイルデータ

③ Analyzing： 問題とする現象解析に適した解析法はどうあるべきか

← **ワークショップのするところ**

６．予行練習

ミニ演習

設　定

あなたは学生向けアパートの不動産のオーナーです。

課　題

アパートの建築費の借入金を回収するため、入居率を高めなければなりません。
家賃が高すぎると入居率が落ち、収入を得ることができません。一方で家賃が安すぎると、こちらも借入金の返済が長期化します。

なるべく早く借入金を完済するため、適正な家賃を決める必要があります。

問　題

家賃を決めるために必要となる条件を10個書き出してください。

ミニ演習

機械学習で家賃を決めるAIツールを開発を委託発注します。

Q1： 目的変数は何でしょうか？

Q2： 説明変数は何でしょうか？

Q3： それは質的変数ですか、量的変数ですか？

７．はじめてみましょう

機械学習の流れ（Scikit-learnに慣れる）

①探索的データ分析（EDA）	②データ可視化	③機械学習
概要統計量の算出 ペアプロットの作図 相関係数の計算	頻度分布 単回帰	線形重回帰 決定木 非線形回帰

ライブラリ
pandas
matplotlib
seaborn

ライブラリ
scikit-learn
matplotlib

ライブラリ
scikit-learn
matplotlib

【参考】 Pythonの代表的なライブラリ

ライブラリ名	主な機能
NumPy	**高性能の数値計算やデータ処理に特化したPythonのライブラリ** NumPyは多次元の配列や行列を効率的に操作する機能を提供し、科学技術計算やデータ解析の分野で広く使用されています。
Pandas	**データ操作と解析のための高レベルのPythonライブラリ** Pandasはテーブル形式のデータを効率的に処理し、データのフィルタリング、変換、集約、および結合などの機能を提供。データの整形やクリーニング、欠損値の処理などを容易に行うことができます。
Matplotlib	**Pythonでデータを可視化するための強力なライブラリ** Matplotlibはグラフや図を描画するための多様な機能を提供し、折れ線グラフ、ヒストグラム、散布図、バーチャートなどの多くのプロットスタイルをサポート。データの傾向や関係性を視覚的に理解するための強力なツールです。
Scikit-learn	**Pythonで機械学習のタスクを実装するための包括的なライブラリ** Scikit-learnは、分類、回帰、クラスタリング、次元削減などの機械学習アルゴリズムやツールを提供。データの前処理、特徴抽出、モデルの評価などもサポートしており、機械学習の実装を容易にします。

予測の機械学習モデルの練習 （Bonton Housingデータセット）

概要

Boston Housingデータセットは、1970年代初頭にアメリカのマサチューセッツ州ボストン市で収集された住宅価格に関する情報を含むデータセット。

- 506の異なる地域（郊外）の住宅に関する情報が含まれている。
- 各地域には、住宅価格を予測するための13種類の特徴量が示されている。

【利用にあたっての注意】

Boston Housingデータセットは、住宅価格の予測や地域の特徴の関係性の分析など、さまざまな機械学習のタスクに使用されてきました。しかし、「黒人の割合」などの差別的なデータを含んでいるため，Scikit learnのversion 1.0以降から利用は非推奨となりました。

https://colab.research.google.com/github/ARIM-Academy/Advanced_Tutorial_1/blob/main/Scikit-learn-0.ipynb