データサイエンス・オンライン講座 特別セミナー

今日から始められる、Pythonによるビジネスデータ解析プログラミング講座

2024/12/14

野口　怜

総務省統計局 × DSI DataScienceInstitute

Agenda

■ Part１：今日から始めるための Pythonプログラミングの基本（30min）

- ✓ Google Colaboratoryの使い方（10min）
  - Google Colab使い方演習
- ✓ Pythonプログラミングの基本（20min）
  - Python基本文法演習

■ Part２：ビジネスデータ解析に 必要なスキル演習（2hr00min）

- ✓ データ分析の進め方（5min）
- ✓ 分析に必要なデータサイエンス力とデータエンジニアリング力
  - Pythonでできる分析実体験：分析デモンストレーション（5min）
  - 記述統計学の基礎（15min）
    - Pythonによる記述統計量算出演習
  - データ観察の基本（25min）
    - Pythonによるグラフ描画演習
  - データ加工の基本（20min）
    - Pythonによるデータ加工演習（ダミー変数化、外れ値・欠損値の処理）
  - データ分析（モデル構築）の基本（50min）
    - Pythonによるモデル構築演習（単回帰分析、重回帰分析）

■ Part３：実際のビジネスデータ解析 に向けた実践演習（2.5hr）

- ✓ e-Stat, SSDSEの紹介と使い方（15min）
- ✓ データ分析の実践演習（2hr）
  - 決定木分析によるビジネス意思決定サポート
  - 階層的クラスタリングによるデータ層別と戦略ポイントの検討
- ✓ 今後の継続学習、実践活用のためのポイント（15min）

※青字はGoogle Colaboratory による演習

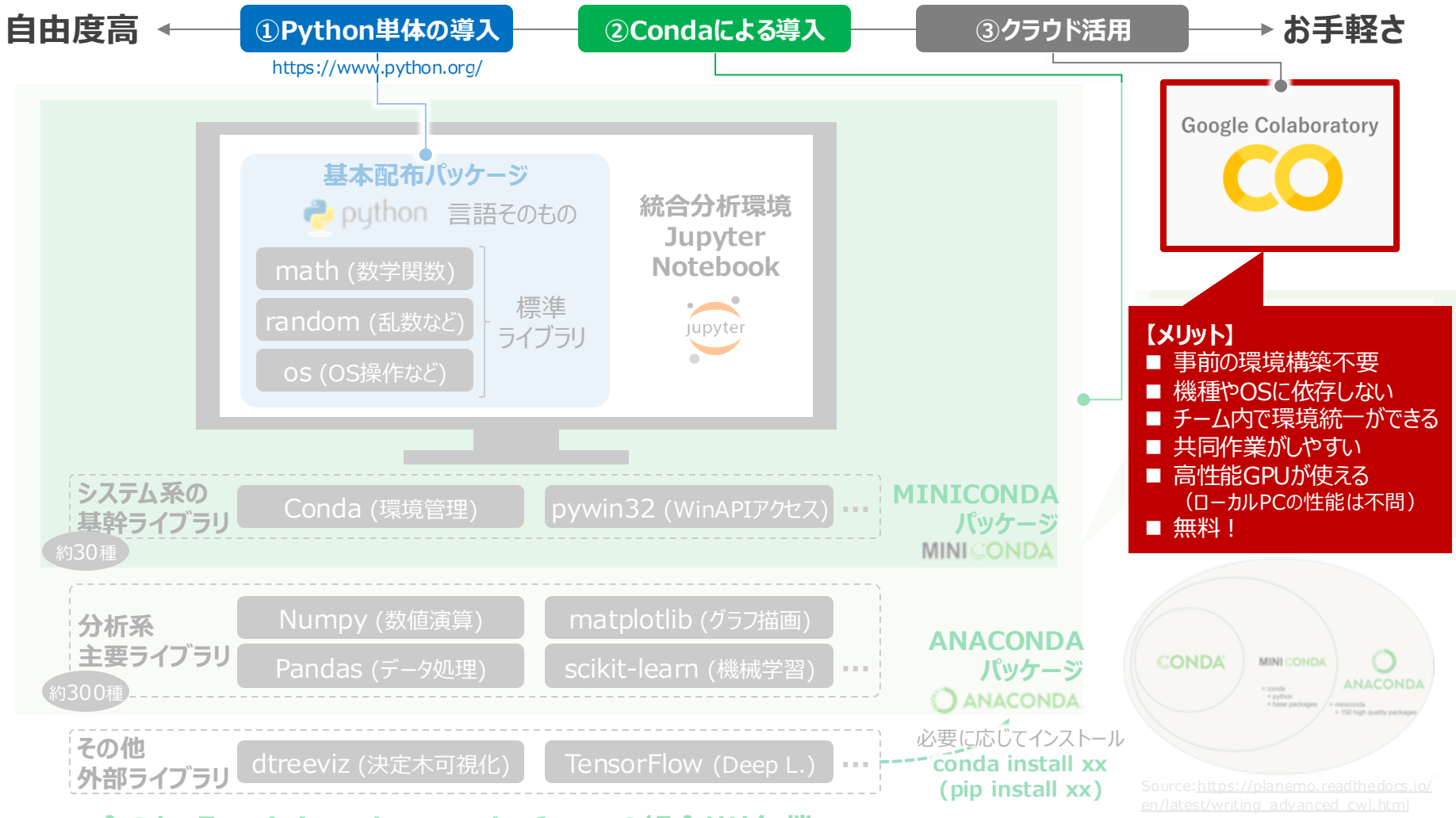
Part 1

今日から始めるための Pythonプログラミングの基本

- ✓Pythonプログラミング環境：Google Colaboratoryの使い方
- ✓Pythonプログラミングの基本

Python導入方法の3パターン とGoogle Colaboratory

■ Pythonの初期環境構築には、①Python単体の導入、②Condaによる導入、③構築済のクラウド活用（Google Colaboratory）の大きく3パターンが存在。環境統一や共同作業を行う場合は、③が適している



※今のところ、miniconda+conda-forgeの組合せは無償！

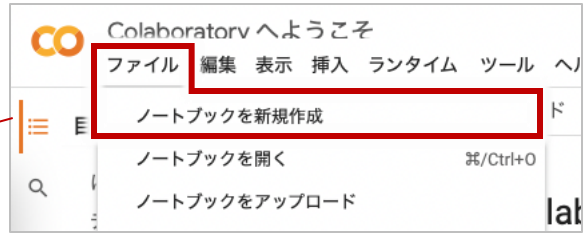
Google Colaboratoryへのアクセス方法とノートブックの作成

- Google Colaboratory はインストール不要のPython開発環境であり、Googleのサーバーにインターネット接続（クラウド環境）することで、ブラウザ上でコーディング、実行ができる
- 必要なデータも随時、Googleのサーバーにアップする必要がある

1 Google Colaboratory へアクセス
https://colab.research.google.com

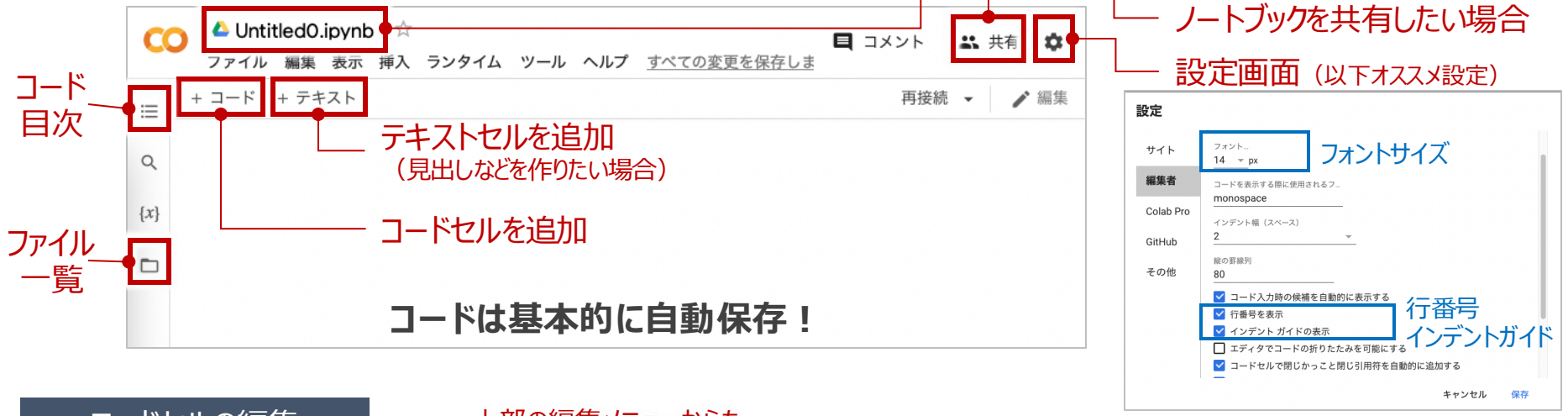
2 右上の [ログイン] ボタンより、Googleアカウントでログイン

3 左上の [ファイル] → [ノートブックを新規作成]

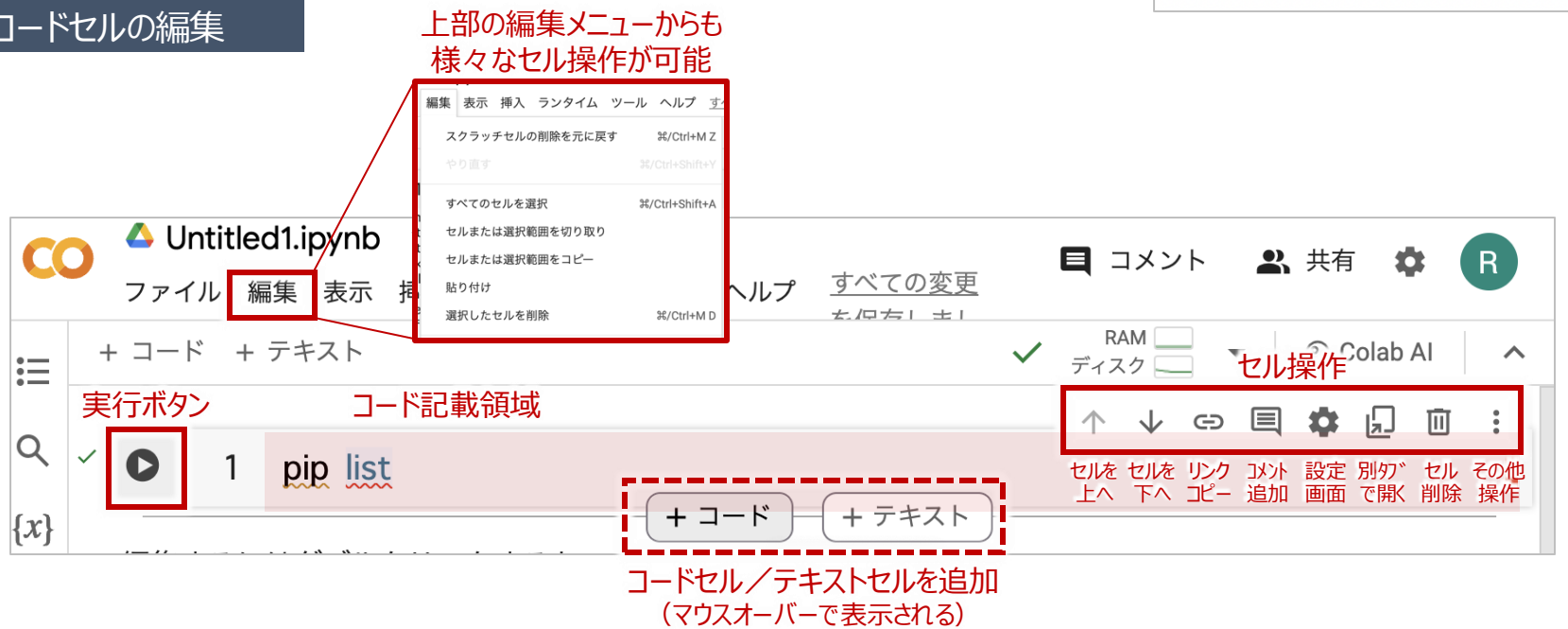


Google Colaboratory画面の概要

画面構成



コードセルの編集



Google Colaboratory上での
レクチャー＆演習

「変数」とは

■ 変数とは、「データを入れておく箱」のことであり、数値に限らず、日付型や文字列型のデータも格納しておくことができる。変数同士の演算も可能で、実際の分析では変数を用いて行う

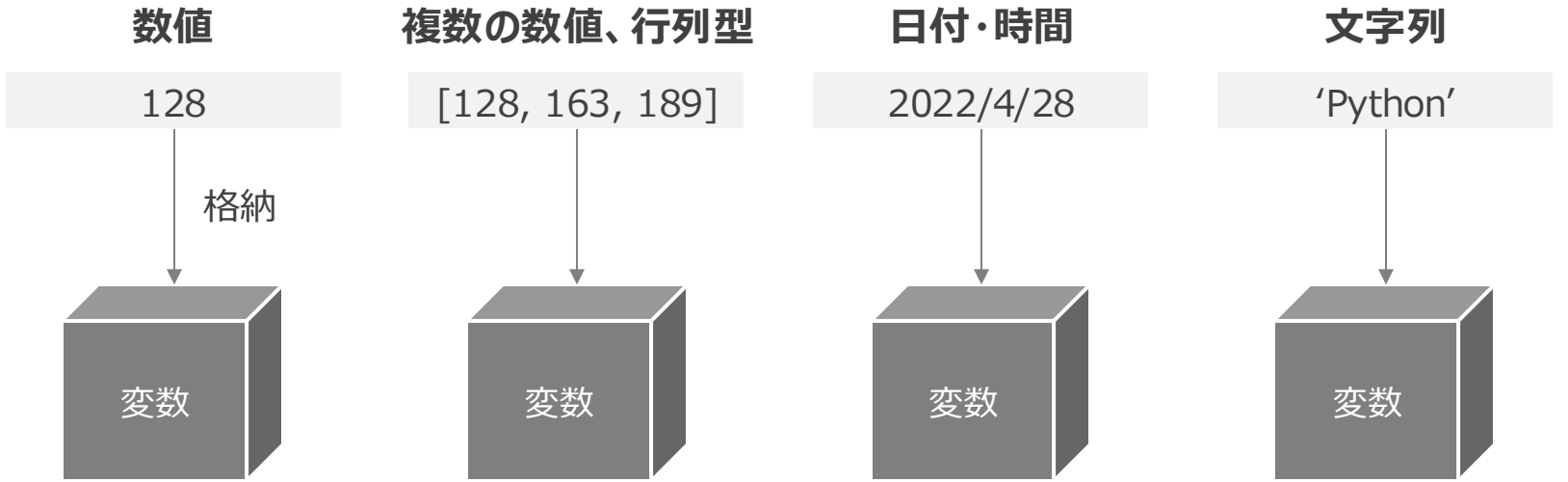
■ 変数名は自由に設定できるが、使用可能な文字には制限があるため、注意が必要

▼プログラミングでは「変数」と呼ばれる「箱」にデータを入れる

例：数値を入れる場合
```
num = 10
print(num)
>> 10
```

例：文字を入れる場合
```
str = "Hello"
print(str)
>> Hello
```

例：演算（足し算）の処理
```
num1 = 1
num2 = 2
num1 + num2
>>3
```



▼変数名の制約

OK

- **アルファベット（大文字・小文字は区別）**
- ひらがな／カタカナ／漢字 **※非推奨**
- 数字（ただし頭文字は不可）

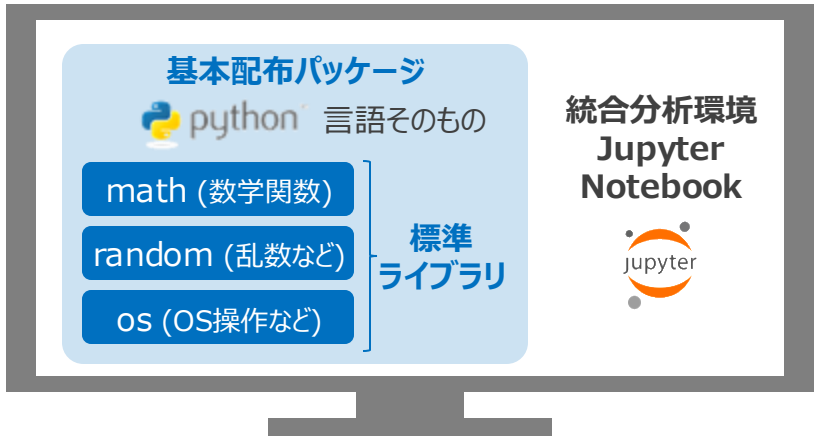
NG

- 記号（@ ！？ – など。アンダーバー _ はOK）
- 環境依存文字（① ㌔ など）
- 予約語や関数名

Google Colaboratory上での
レクチャー＆演習

「ライブラリ」について

- Pythonでは、よく使う関数や便利な機能をひとまとめにした「**ライブラリ**」が豊富に提供されており、コード作成の際は、随時必要なライブラリを import して用いるのが基本
- インストール不要な標準ライブラリの他、別途インストールが必要なライブラリもある



外部ライブラリ
（本来は別途インストールが必要）

システム系の基幹ライブラリ
- Conda (環境管理)
- pywin32 (WinAPIアクセス) …

分析系主要ライブラリ
- Numpy (数値演算)
- matplotlib (グラフ描画)
- Pandas (データ処理)
- scikit-learn (機械学習) …

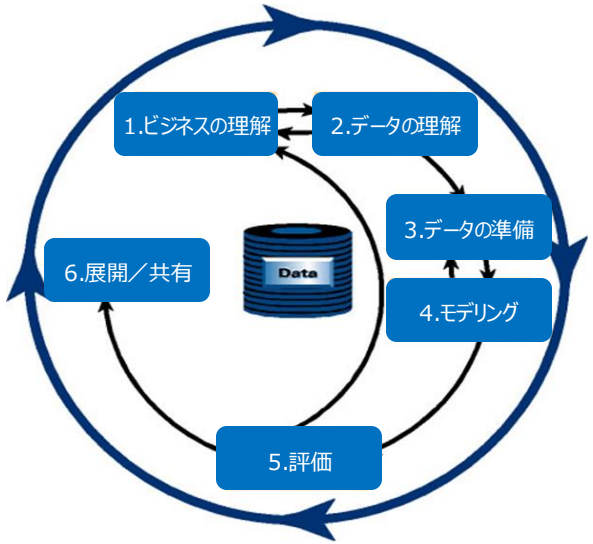
その他外部ライブラリ
- dtreeviz (決定木可視化)
- TensorFlow (Deep L.) …

Google Colaboratory
Google Colaboratoryでは主要なライブラリはインストール済み

ただし、一部の外部ライブラリはインストール必要
（別途ご案内）

ビッグデータ分析／機械学習でよく用いられるライブラリ

ライブラリ名（パッケージ名）	用途
matplotlib	基本グラフ描画
Seaborn	拡張グラフ描画
Numpy	数値演算、ベクトル/行列演算
Pandas	テーブル形式データのデータ処理
scikit-learn	機械学習モデル



CRISP-DM

1.ビジネスの理解

2.データの理解

3.データの準備

4.モデル構築

5.評価

6.展開／共有

matplotlib
Seaborn

Pandas

NumPy

scikit learn

表データの処理ライブラリ：pandas

- pandasは、**表形式のデータ**を処理する上で必須のライブラリである
- 表データをそのまま読み込んだ**DataFrame型**と、一部の列（行）のみを抽出した**Series型**とが存在する

表形式のデータ（テーブルデータ）

Excel, csvファイル など

顧客番号	性別	…	残高	契約有無
000011	男性	…	2,520,000	契約
000012	女性	…	12,353,000	契約
000013	女性	…	4,531,000	未契約
⋮	⋮	⋮	⋮	⋮

Pandas

○Excel
○CSV
○テキスト

Pandasのデータフレーム型

インデックス　カラム名／列名

	顧客番号	性別	残高	契約有無
0	000011	男性	2520000	契約
1	000012	女性	12353000	契約
2	000013	男性	4531000	未契約

Series型

DataFrame型

行 (row)

列 (column)

標準ライブラリのみで読み込むと・・・

×Excel
○CSV
○テキスト

ネストされた2重のリスト
（2次元配列とも呼ばれる）

```
[['000011', '男性', 2520000, '契約'],
 ['000012', '女性', 12353000, '契約'],
 ['000013', '男性', 4531000, '未契約']]
```

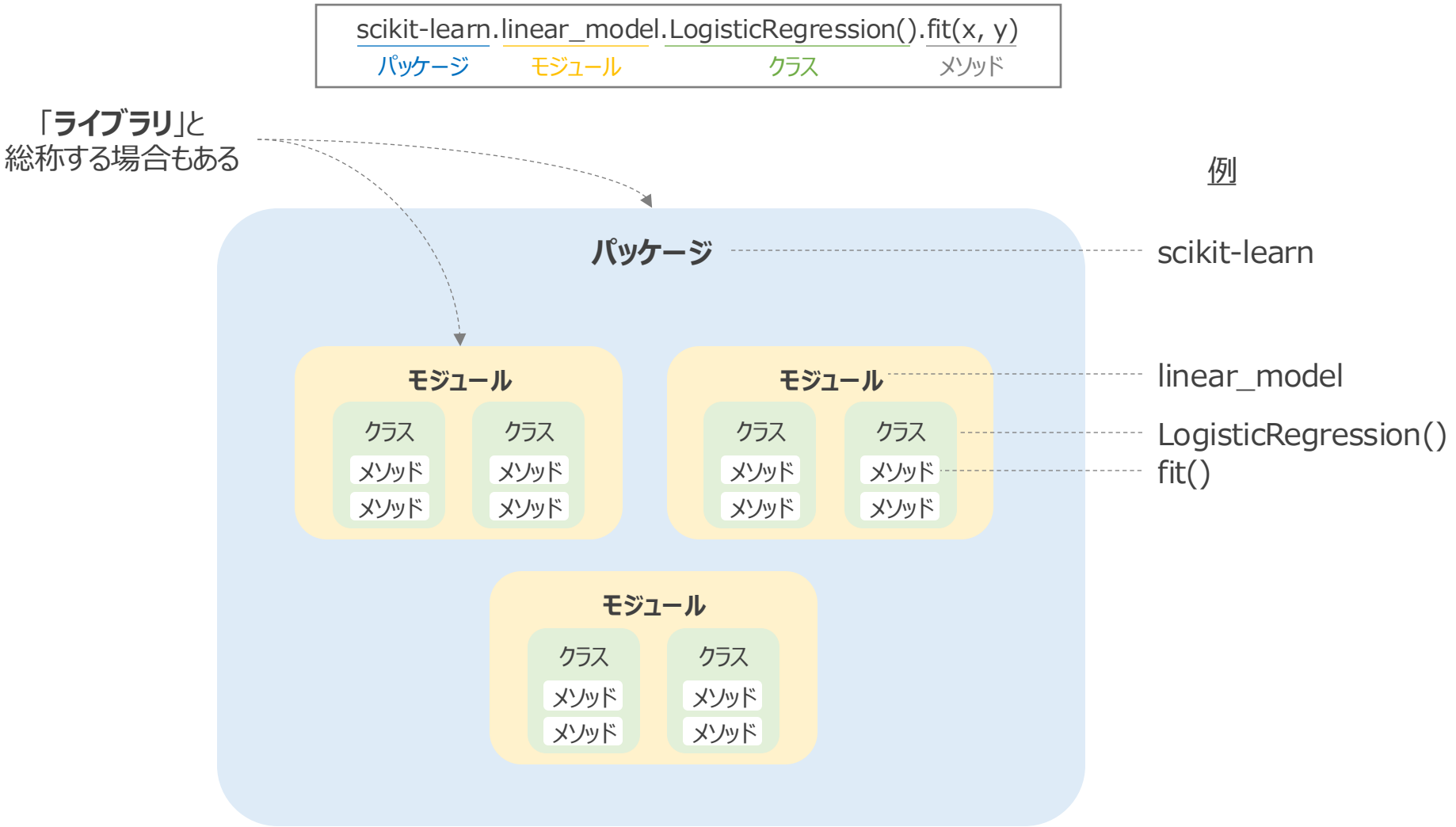
余談：pandasの由来

諸説ありますが・・・下記が有力です
Python **and** **d**ata **a**nalysis
panel **da**ta

画像出典：Adobe Stock
https://stock.adobe.com/jp

Google Colaboratory上での
レクチャー＆演習

（参考）用語の整理：「ライブラリ」「パッケージ」「モジュール」「クラス」



Part2

ビジネスデータ解析に必要なスキルを学ぶ

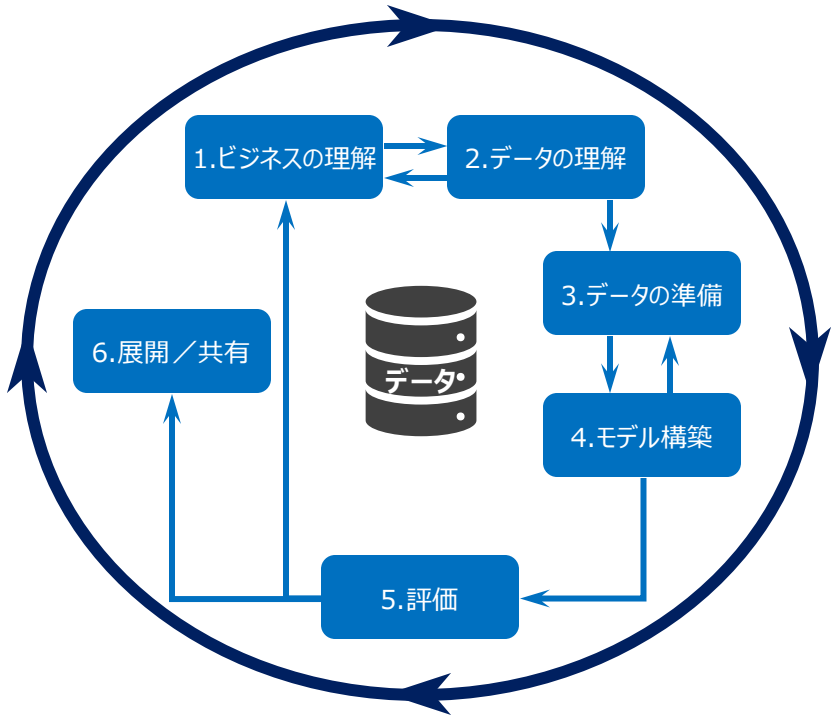
✓データ分析の進め方
✓分析に必要なビジネス力
✓分析に必要なデータサイエンス力とデータエンジニアリング力

データ分析の進め方

■ データ分析の進め方に関する方法論「**CRISP-DM**」に基づいて、分析と評価を繰り返して試行錯誤しながら進めるのが一般的である

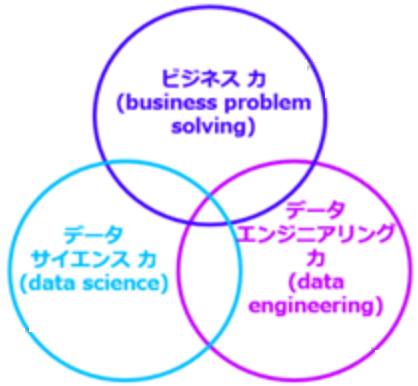
CRISP-DM: 分析方法論

(**CR**oss **I**ndustry **S**tandard **P**rocess for **D**ata **M**ining)



1.ビジネスの理解	•ビジネス、分析目標の決定 •プロジェクトの立ち上げ	ビジネス
2.データの理解	•データの収集 •データの調査 •データ品質の検証	データサイエンス
3.データの準備	•データの選択や除外 •データのクリーニング •データの構築や統合	ﾃﾞｰﾀｴﾝｼﾞﾆｱﾘﾝｸﾞ
4.モデル構築	•モデリング手法の選択 •モデルの作成 •モデルの評価	データサイエンス
5.評価	•分析結果の評価 •プロセスの見直し •実行可能なアクションリストの作成	ﾃﾞｰﾀｴﾝｼﾞﾆｱﾘﾝｸﾞ
6.展開／共有	•業務への導入計画 •モニタリング、メンテナンスの計画	ビジネス

一般社団法人 データサイエンティスト協会



分析に必要なビジネス力

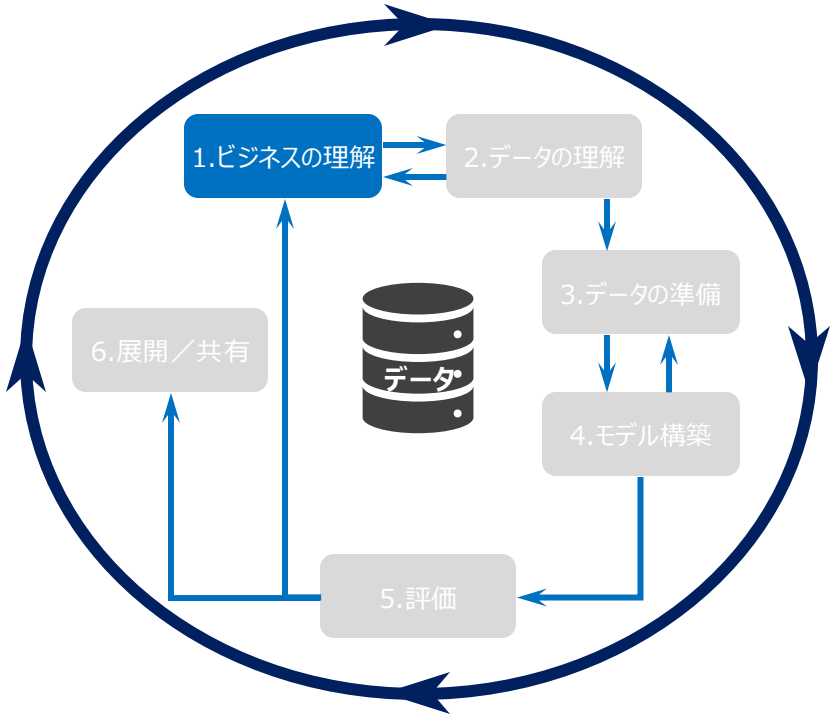
✓ 分析テーマの明確化
✓ 仮説思考力の重要性

データ分析の進め方

■ データ分析の進め方に関する方法論「CRISP-DM」に基づいて、分析と評価を繰り返して試行錯誤しながら進めるのが一般的である

CRISP-DM: データマイニング方法論

(**CR**oss **I**ndustry **S**tandard **P**rocess for **D**ata **M**ining)



1.ビジネスの理解	•ビジネス、データマイニング目標の決定 •プロジェクトの立ち上げ
2.データの理解	•データの収集 •データの調査 •データ品質の検証
3.データの準備	•データの選択や除外 •データのクリーニング •データの構築や統合
4.モデル構築	•モデリング手法の選択 •モデルの作成 •モデルの評価
5.評価	•データマイニングの結果の評価 •プロセスの見直し •実行可能なアクションリストの作成
6.展開／共有	•業務への導入計画 •モニタリング、メンテナンスの計画

「なんかわかるでしょ？」の注意

■ データ分析のきっかけは、大体ぼんやりとした依頼から始まることが多い・・・

最近ほら、AIとか流行ってるでしょ。
ウチも売上データ使ってなんかできない？
データがあればなんかわかるよね？

売り上げを増やしたいの？
顧客を増やしたいの？

- **分析対象（目的変数）の明確化**

なんとか貸し倒れを減らしたいんだけど、
データ分析して原因を見つけられる？

今どのくらいの貸し倒れ？
どこまで改善目指す？

- **現状の把握**
- **分析目標（ゴール）の設定**

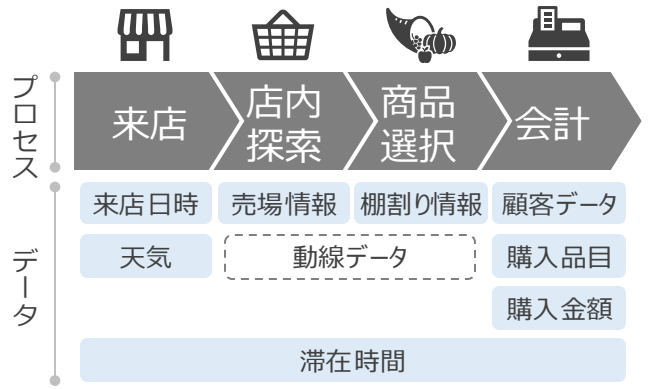
（参考）ビジネスの理解＝「プロセス」と「構造」の理解

■ データを触る前に、分析対象ビジネスの「プロセス」と「構造」を理解しておくことが重要である
■ これらと並行して、分析スコープと分析目的の検討、データの理解も行う

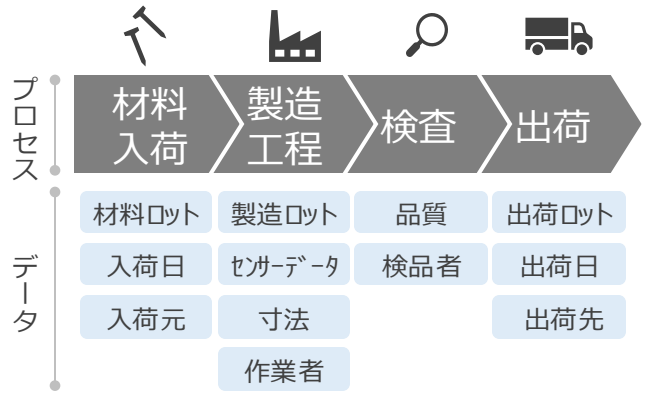
1 「ビジネスプロセス」を理解する

例

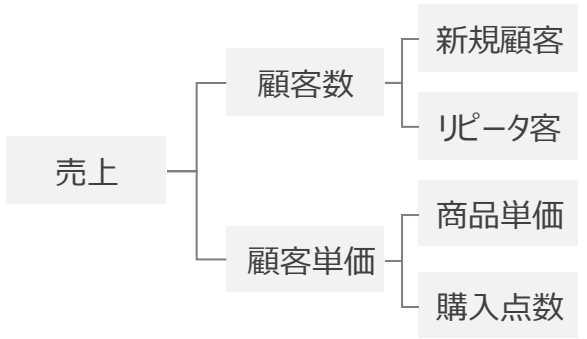
小売業

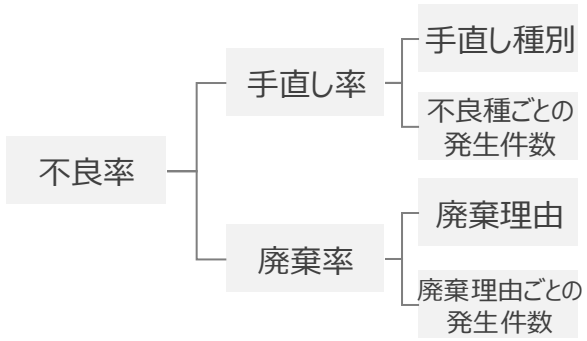


製造業



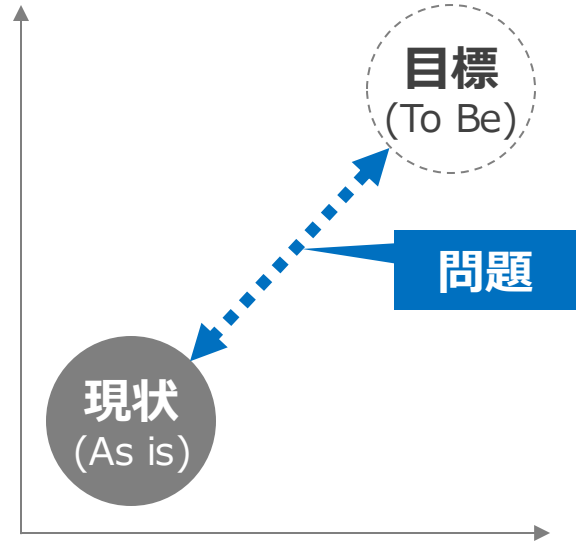
2 「ビジネス構造」を理解する





（参考）現状把握と目標の設定

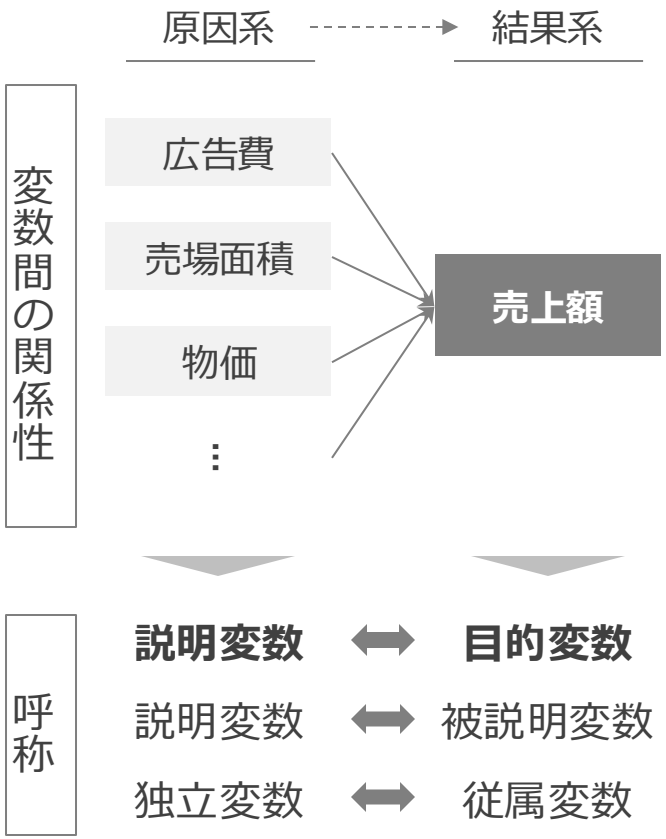
- 理解したビジネスのプロセス・構造に基づいて、現状の把握・分析目標の設定を行ない、**データ分析により解くべき問題を明確化**する
- （特にコスト削減のケースでは）**可能な限り定量的**な現状把握、目標設定を行う



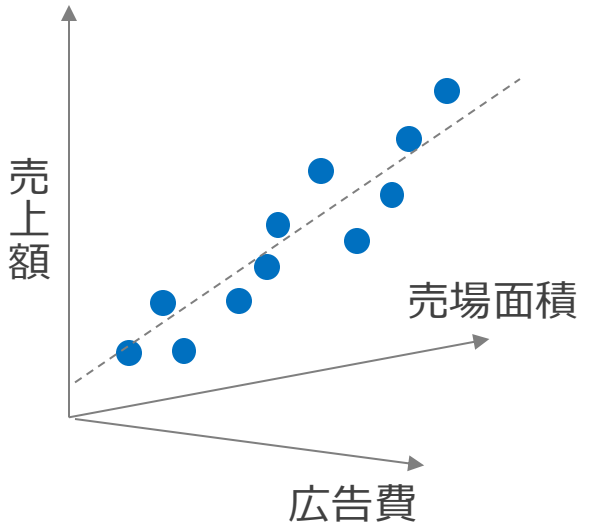
	例1：売上向上のケース （分析前には必ずしも定量化できない） 小売業の例	例2：リスク低減のケース （定量化可能） 銀行の例
目標	▪ 最適な商品陳列により、売上アップを見込む	▪ 貸し倒れしやすい顧客の特性を絞り込み、**貸し倒れ率xx%以下**を目指す ▪ **xx百万/月**の損失低減目標
問題	**現状と目標のギャップ（＝問題）を データ分析により、「打ち手」につなげる**	
現状	▪ 商品陳列の工夫ができておらず、販売機会損失が発生している可能性	▪ 貸し倒れによる損失が、20xx年x月の1ヶ月間でxx件発生（貸し倒れ率xx%） ▪ 金額ベースでxx百万/月のロス

目的変数・説明変数とは

- 説明変数は四則演算などで必要に応じて新規変数を作る
- 目的変数は、必要に応じてフラグ値化／カテゴリー値化などの加工をする
  (分布の形を見てビンの幅や閾値を判断する)



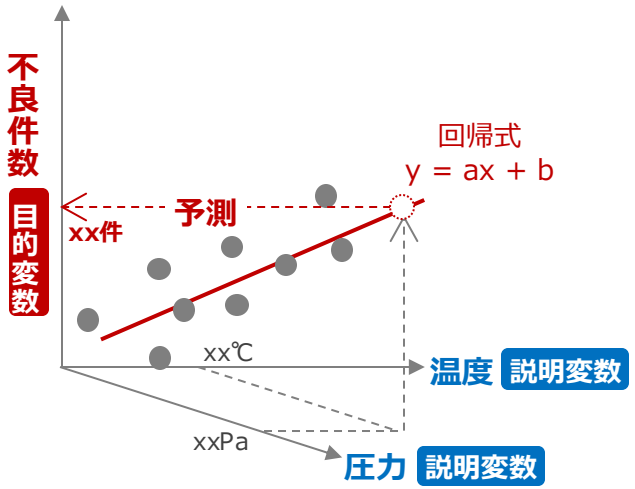
重回帰の場合：
売上額 = a_1・広告費 + a_2・売場面積 + a_3・物価 +・・・



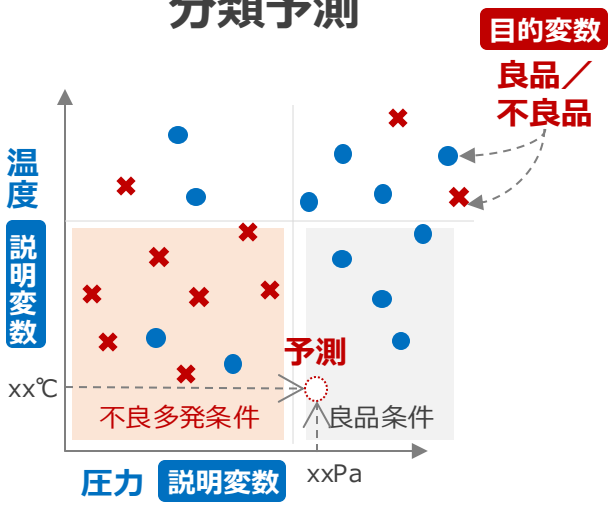
目的変数・説明変数の検討

■ 分析目的やデータ形式に応じて、目的変数や説明変数の形式は異なる
「どんなデータを用いてどんな分析をしたいか」を明確にすることが重要

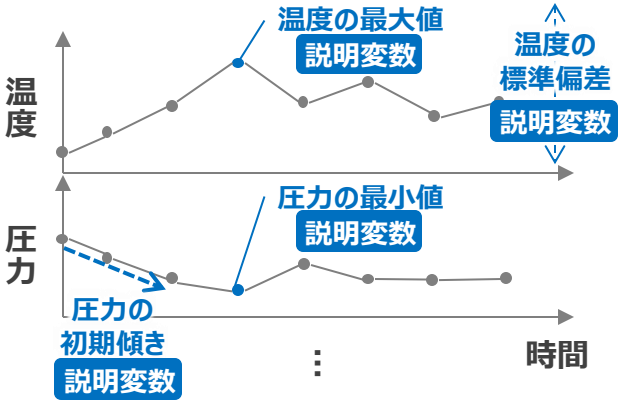
数値予測



分類予測



時系列分析



目的変数 良品／不良品

分析例

- ■ 小売：売上金額の予測
- ■ 小売：需要予測
- ■ 製造：不良発生率の予測

- ■ 小売：購入／非購入顧客の分類
- ■ 医療：生存条件の分析
- ■ 製造：故障種類の分類

- ■ 設備稼働の異常検知
- ■ リアルタイム不良・不具合検出

統計学の基礎

✓ 記述統計学 と 推測統計学
✓ 要約統計量（代表値／ばらつき指標／順序統計量）

統計の落とし穴

■ 以下は、日本国内でのある統計調査の結果です
■ みなさん、こんな危険な食べ物、即刻禁止にすべきと思いませんか？

- **心筋梗塞による死亡者の95%以上**が生前ずっとこの食べ物を食べていた
- **がん患者の98%**がこの食べ物を摂取していた
- **強盗や殺人などの凶悪犯の70%以上**が犯行の24時間以内にこの食べ物を口にしている

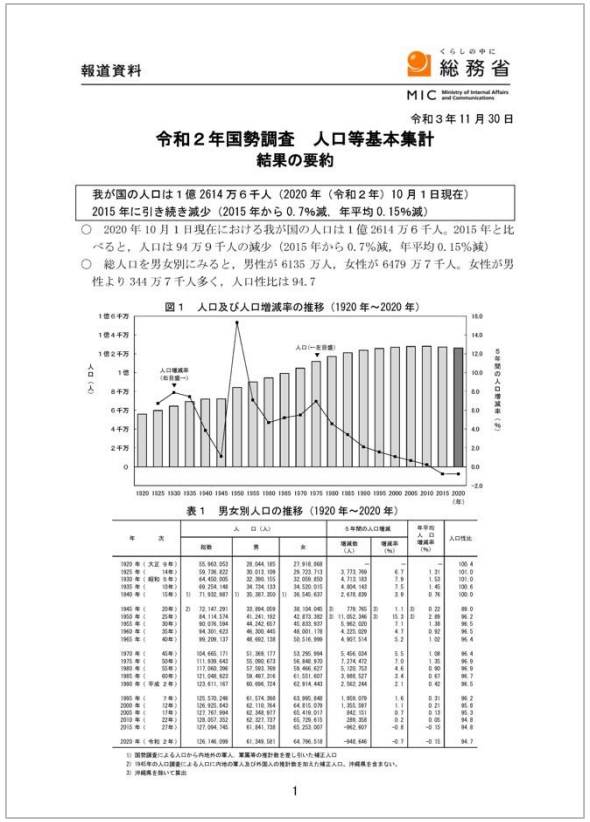
西内 啓「統計学が最強の学問である」より

「**お米**」禁止しますか？

2つの「統計学」

■ 国勢調査と視聴率調査は何が違う？

国勢調査



報道資料

総務省 MIC Ministry of Internal Affairs and Communications

令和３年11月30日

令和２年国勢調査　人口等基本集計
結果の要約

我が国の人口は１億2614万６千人（2020年（令和２年）10月１日現在）
2015年に引き続き減少（2015年から0.7%減，年平均0.15%減）

○ 2020年10月１日現在における我が国の人口は１億2614万６千人。2015年と比べると，人口は94万９千人の減少（2015年から0.7%減，年平均0.15%減）
○ 総人口を男女別にみると，男性が6135万人，女性が6479万７千人。女性が男性より344万７千人多く，人口性比は94.7

図１　人口及び人口増減率の推移（1920年～2020年）

表１　男女別人口の推移（1920年～2020年）

年次	人口（人） 総数	男	女	5年間の人口増減 増減数（人）	増減率（%）	年平均人口増減率（%）	人口性比
1920年（大正９年）	55,963,053	28,044,185	27,918,868	－	－	－	100.4
1925年（14年）	59,736,822	30,013,199	29,723,713	3,773,769	6.7	1.31	101.0
1930年（昭和５年）	64,450,005	32,390,155	32,059,850	4,713,183	7.9	1.53	101.0
1935年（10年）	69,254,148	34,734,133	34,520,015	4,804,143	7.5	1.45	100.6
1940年（15年）	1) 71,932,987	1) 35,987,350	1) 36,945,637	2,678,839	3.9	0.76	100.0
1945年（20年）	2) 72,147,291	33,894,059	38,104,045	3) 776,765	3) 1.1	3) 0.22	89.0
1950年（25年）	84,114,574	41,241,192	42,873,382	3) 11,052,346	3) 15.3	3) 2.89	96.2
1955年（30年）	90,076,594	44,242,657	45,833,937	5,962,020	7.1	1.38	96.5
1960年（35年）	94,301,623	46,300,445	48,001,178	4,225,029	4.7	0.92	96.5
1965年（40年）	99,209,137	48,692,138	50,516,999	4,907,514	5.2	1.02	96.4
1970年（45年）	104,665,171	51,368,177	53,296,994	5,456,034	5.5	1.06	96.4
1975年（50年）	111,939,643	55,060,673	56,878,970	7,274,472	7.0	1.35	96.9
1980年（55年）	117,060,396	57,553,769	59,506,627	5,120,753	4.6	0.90	96.7
1985年（60年）	121,048,923	59,467,316	61,581,607	3,988,527	3.4	0.67	96.7
1990年（平成２年）	123,611,167	60,696,724	62,914,443	2,562,244	2.1	0.42	96.5
1995年（７年）	125,570,246	61,574,368	63,995,878	1,959,079	1.6	0.31	96.2
2000年（12年）	126,925,843	62,110,764	64,815,079	1,355,597	1.1	0.21	95.8
2005年（17年）	127,767,994	62,348,817	65,419,017	842,151	0.7	0.13	95.3
2010年（22年）	128,057,352	62,327,737	65,729,615	289,358	0.2	0.05	94.8
2015年（27年）	127,094,745	61,841,738	65,253,007	-962,607	-0.8	-0.15	94.8
2020年（令和２年）	126,146,099	61,349,581	64,796,518	-948,646	-0.7	-0.15	94.7

1) 国勢調査による人口から内地外の軍人，軍属等の推計数を差し引いた補正人口
2) 1945年の人口調査による人口に内地の軍人及び外国人の推計数を加えた補正人口。沖縄県を含まない。
3) 沖縄県を除いて算出

1

Source: https://www.stat.go.jp/data/kokusei/2020/kekka/pdf/summary_01.pdf

視聴率調査

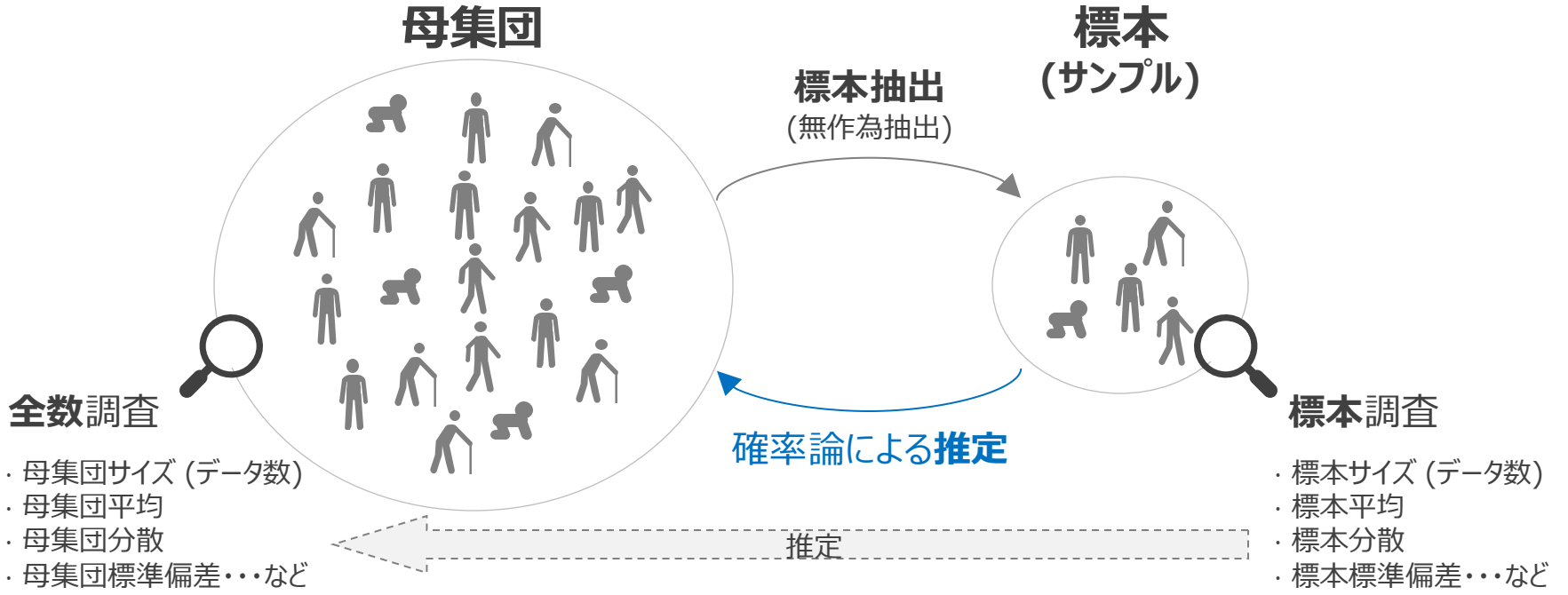
順位	番組タイトル	放送局	視聴率
1	連続テレビ小説・らんまん（7/14）	NHK総合	17.5 %
2	ザワつく！金曜日（7/14）	テレビ朝日	13.1 %
3	ポツンと一軒家 3時間SP（7/16）	テレビ朝日	12.3 %
4	大相撲 2023年 名古屋場所・8日目（7/16 17:06～）	NHK総合	12.1 %
5	日曜劇場・VIVANT（7/16）	TBS	11.5 %
6	ブラタモリ（7/15）	NHK総合	11.3 %
7	1億人の大質問!?笑ってコラえて！ 27周年2時間SP（7/12）	日本テレビ	11.1 %
8	世界まる見え！テレビ特捜部 2時間SP（7/10）	日本テレビ	10.6 %
9	木曜ドラマ・ハヤブサ消防団（7/13）	テレビ朝日	10.5 %
10	音楽の日2023（7/15 19:00～）	TBS	10.1 %

2023年7月10日(月)～7月16日(日) 視聴率ランキング

Source: https://www.tvguide.or.jp/tvranking/tvrank-20230728/

2つの「統計学」– 記述統計学 と 推測統計学

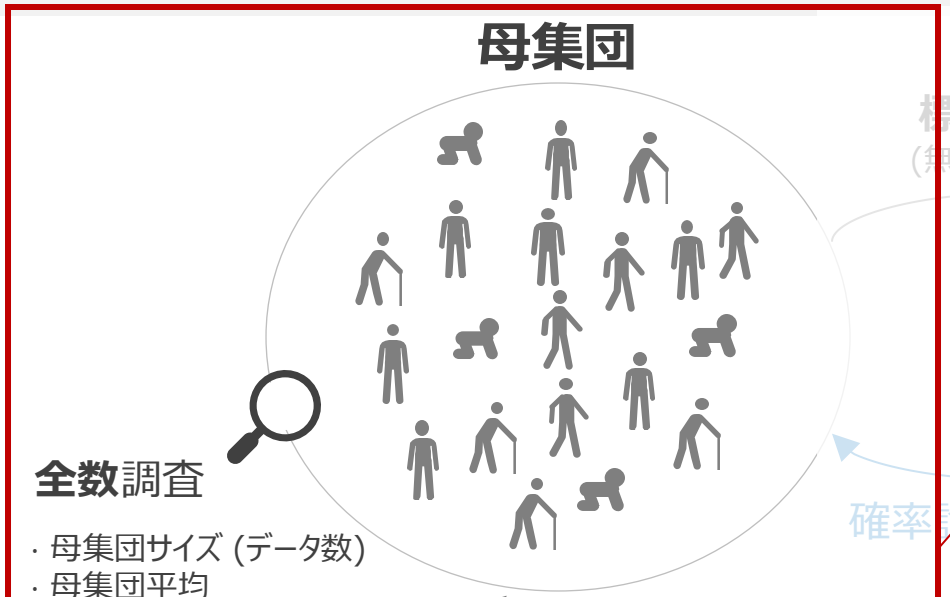
■ データ全体を調べてデータの特徴や傾向を把握する「記述統計学」と、
一部のサンプルを抽出して調べ、全体の特徴や傾向を推測する「推測統計学」がある



	記述統計学	推測統計学
概要	全数調査により観測された**データ全体の特徴を調べたり、データを要約する**アプローチのこと	データ全体（母集団）を調べあげるのが困難な場合に、抽出したデータの一部（標本）の特徴から、**データ全体の特徴を確率的に推測**するアプローチのこと
例	国勢調査、(学校の)健康診断・定期テスト など	世論調査、視聴率、街角アンケート など
例えば企業では…	（CRM内の）顧客データの分析 など	顧客の一部へのアンケート調査分析 など

2つの「統計学」– 記述統計学 と 推測統計学

■ データ全体を調べてデータの特徴や傾向を把握する「記述統計学」と、
一部のサンプルを抽出して調べ、全体の特徴や傾向を推測する「推測統計学」がある



全数調査

- 母集団サイズ (データ数)
- 母集団平均
- 母集団分散
- 母集団標準偏差・・・など

標本抽出
(無作為抽出)

標本
(サンプル)

データ利活用に向けての
最低限の知識として
今回はこちらをメインに扱います

確率論による**推定**

推定

標本調査

- 標本サイズ (データ数)
- 標本平均
- 標本分散
- 標本標準偏差・・・など

記述統計学

概要	全数調査により観測された**データ全体の特徴を調べたり、データを要約する**アプローチのこと
例	国勢調査、(学校の)健康診断・定期テスト など
例えば企業では…	（CRM内の）顧客データの分析 など

推測統計学

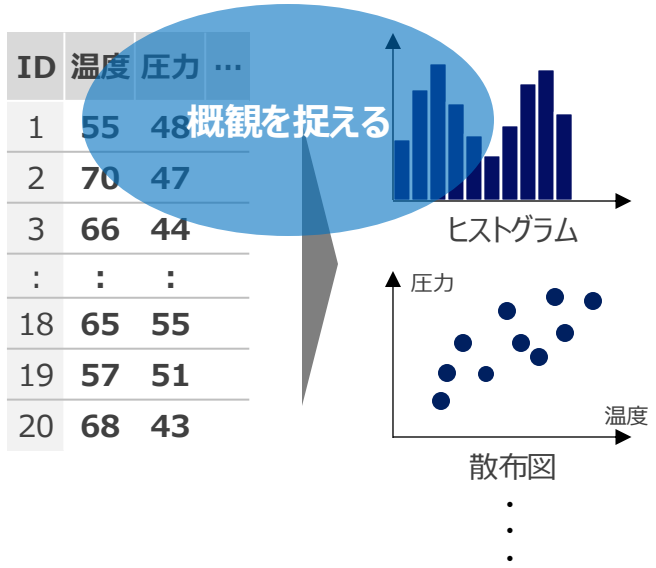
概要	データ全体（母集団）を調べあげるのが困難な場合に、抽出したデータの一部（標本）の特徴から、**データ全体の特徴を確率的に推測**するアプローチのこと
例	世論調査、視聴率、街角アンケート など
例えば企業では…	顧客の一部へのアンケート調査分析 など

記述統計学の基本：データの特徴を捉える

■ 個々のデータをくまなく見るのは難しいため、グラフ（ヒストグラムや散布図）や
要約統計量（平均値や標準偏差） を用いて全体傾向を把握する

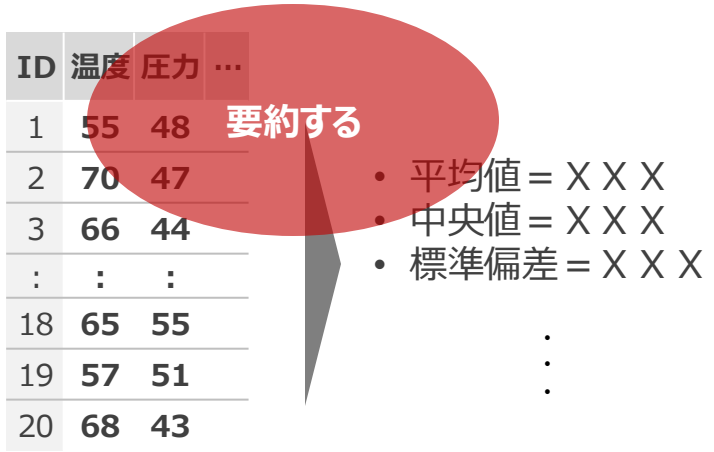
可視化（グラフ化）

視覚的にデータの特徴や傾向を把握



数値化（データ要約）

データの特徴を示す値に要約し
比較可能な客観的傾向を掴む
（**要約統計量**）



記述統計学の基本：データの特徴を捉える

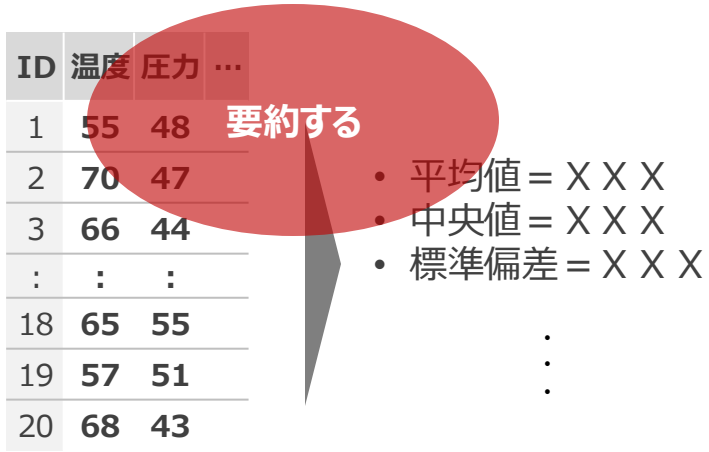
■ 個々のデータをくまなく見るのは難しいため、グラフ（ヒストグラムや散布図）や
要約統計量（平均値や標準偏差） を用いて全体傾向を把握する

可視化（グラフ化）

視覚的にデータの特徴や傾向を把握

数値化（データ要約）

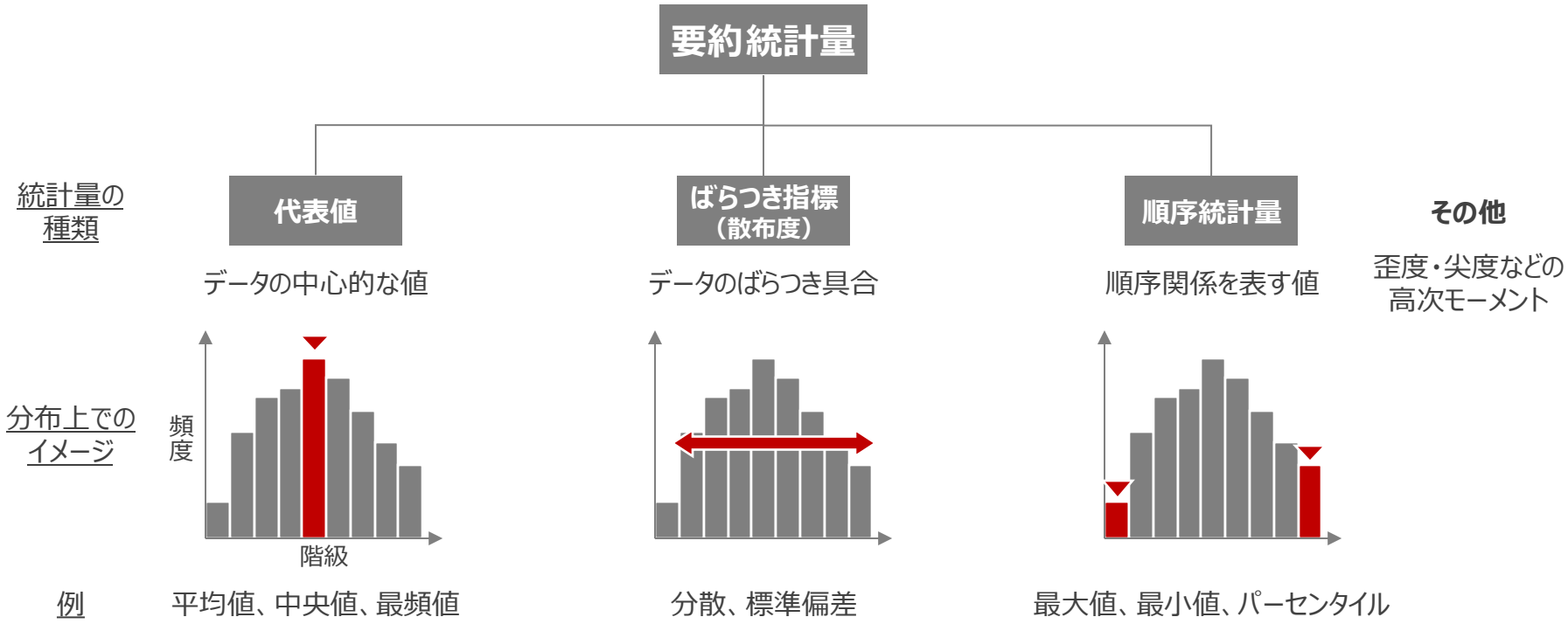
データの特徴を示す値に要約し
比較可能な客観的傾向を掴む
（**要約統計量**）



※可視化については、別途「データの理解（可視化）」のプロセスで取り扱う

数値化（データ要約）

■ 全データをくまなく見るのは難しいため、「要約した値（＝要約統計量）」を用いて全体傾向を把握するのが一般的



代表値（平均値）

■ 平均値は最もよく用いられる代表値であるが、分布の形によっては思わぬ落とし穴がある

ある会社の社員9名の年収

社員	年収
A	200万円
B	200万円
C	200万円
D	400万円
E	400万円
F	500万円
G	700万円
H	900万円
I	1億円

平均値の算出

$$\text{平均値} = \frac{200\text{万円} + 200\text{万円} + \cdots + 1\text{億円}}{9\text{名}}$$

$$= \mathbf{1{,}500\text{万円}}$$

この平均値、直感と合っている？

代表値（中央値・最頻値）

- 中央値：データを値の小さい順に並べたときに中央に位置する値
- 最頻値：データの中で最も多く出現する値。複数存在するケースもある

ある会社の社員9名の年収

社員	年収
A	200万円
B	200万円
C	200万円
D	400万円
E	400万円
F	500万円
G	700万円
H	900万円
I	1億円

中央値と最頻値

代表値だけではつかめない特徴

■ 代表値はすべて同じ。どちらの会社に入っても待遇は同じ！と言ってよいのか？

		A社	B社
A社/B社 社員の年収		年収	年収
		300万円	0万円
		400万円	100万円
		400万円	200万円
		500万円	300万円
		500万円	500万円
		500万円	500万円
		500万円	700万円
		600万円	800万円
		600万円	900万円
		700万円	1000万円
代表値	平均値	500万円	500万円
	中央値	500万円	500万円
	最頻値	500万円	500万円

代表値だけではつかめない特徴

■ ヒストグラムで見ると、社員によって年収が大きく異なっていることがわかる

A社/B社
社員の年収

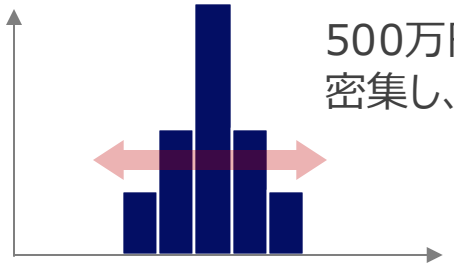
A社

年収
300万円
400万円
400万円
500万円
500万円
500万円
500万円
600万円
600万円
700万円

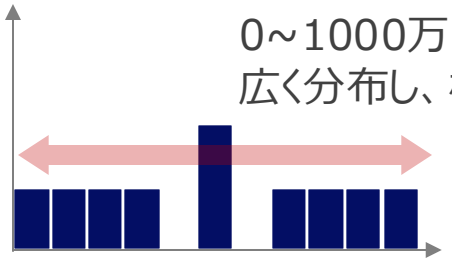
B社

年収
0万円
100万円
200万円
300万円
500万円
500万円
700万円
800万円
900万円
1000万円

ヒスト
グラム



500万円付近に
密集し、格差が小さい



0~1000万円まで
広く分布し、格差が大きい

データの「ばらつき」の表し方

■ データのばらつきは、「平均からのズレ」＝偏差 として捉えるのが基本的な考え方

A社/B社

社員の年収

A社

年収	平均値との差
300万円	-200万円
400万円	-100万円
400万円	-100万円
500万円	0万円
500万円	0万円
500万円	0万円
500万円	0万円
600万円	+100万円
600万円	+100万円
700万円	+200万円

平均年収
500万円

B社

年収	平均値との差
0万円	-500万円
100万円	-400万円
200万円	-300万円
300万円	-200万円
500万円	0万円
500万円	0万円
700万円	+200万円
800万円	+300万円
900万円	+400万円
1000万円	+500万円

平均年収
500万円

データの「ばらつき」の計算

- 分散：平均からのズレを、（正負の符号を消すために）2乗して平均した値
- 標準偏差：元のデータと単位を揃えるために、分散の値を変換した値

	A社		B社	
	年収	平均値との差	年収	平均値との差
A社/B社	300万円	-200万円	0万円	-500万円
社員の年収	400万円	-100万円	100万円	-400万円
	400万円	-100万円	200万円	-300万円
	500万円	0万円	300万円	-200万円
	500万円	0万円	500万円	0万円

$$\text{A社年収の}\textbf{分散} = \frac{(-200\text{万円})^2 + (-100\text{万円})^2 + \cdots + (100\text{万円})^2 + (200\text{万円})^2}{10\text{名}}$$

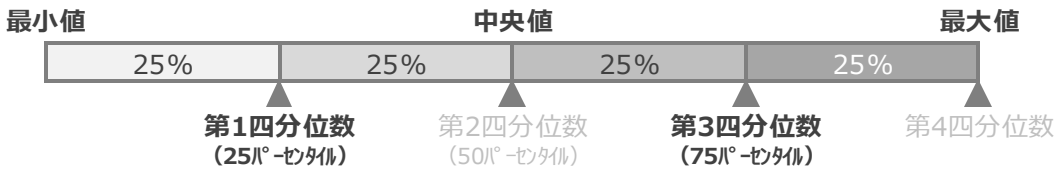
$$= \text{A社の年収}\ \textbf{標準偏差} \times \text{A社の年収}\ \textbf{標準偏差}$$

実際に標準偏差を計算すると、A社は109.5万円、B社は328.6万円

代表的な要約統計量まとめ

■ どれか一つを算出すれば良いわけではなく、様々な観点でデータの概観を把握することが重要
■ また、時系列データの分析ではこれら統計量自体を「特徴量」として扱うことも多い

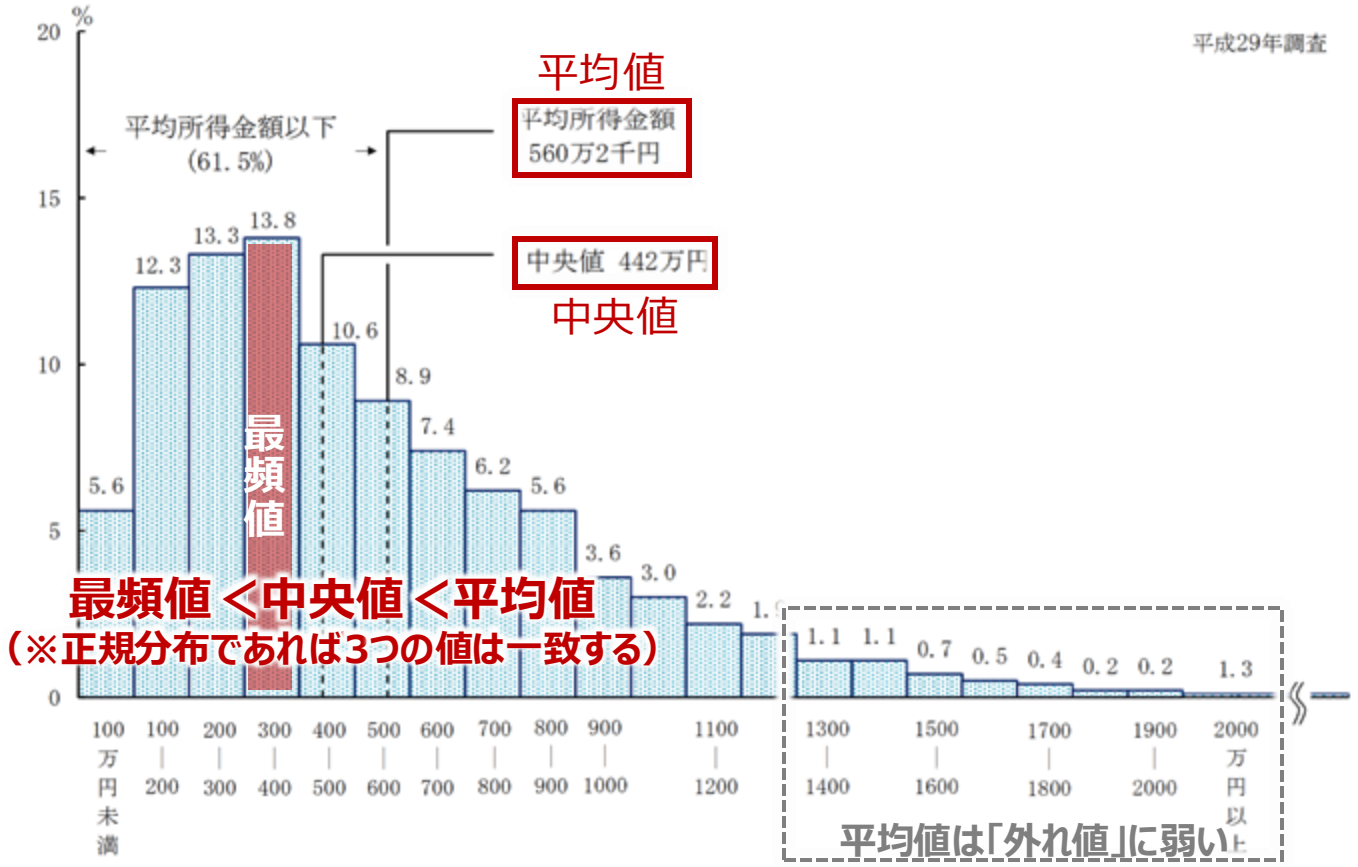
分類	統計量	説明
代表する値（代表値）	平均	データの中心位置を表す値
	中央値（メディアン）	データを小さい順に並べたとき、ちょうど中央の位置にくる値
	最頻値（モード）	データの中で最頻出する値
ばらつき具合を示す値（散布度）	分散	データのばらつき具合を表す値
	標準偏差	データのばらつき具合を表す値 元データと**同じ単位**のため直感的なばらつき把握が可能
順序を表す値（順序統計量）	最大値／最小値	データの中で最も大きい／小さい値
	第1／第3四分位数	データを小さい順に並べたとき、 それぞれ25%／75% の位置にくる値



（参考）年収分布における代表値の差異

■ 正規分布ではない分布では、平均値だけを採用するとデータの特徴を見誤る恐れがある

図９　所得金額階級別世帯数の相対度数分布



出典：厚生労働省　平成 29 年 国民生活基礎調査の概況

データ観察の基本

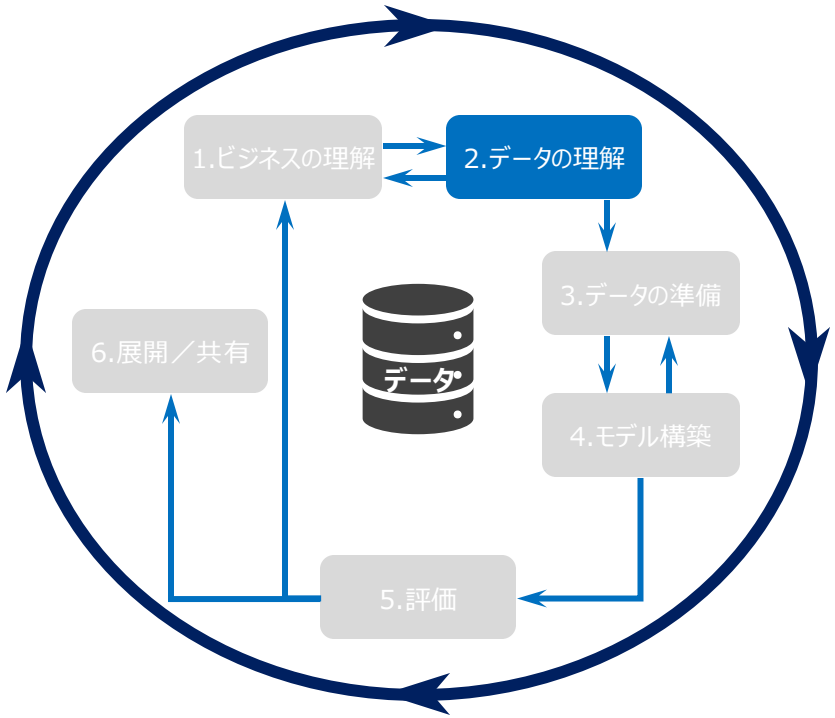
✓グラフと要約統計量によるデータの傾向把握
✓相関分析によるデータ間の関係性把握
✓相関と因果の違い

データ分析の進め方

■ データ分析の進め方に関する方法論「CRISP-DM」に基づいて、分析と評価を繰り返して試行錯誤しながら進めるのが一般的である

CRISP-DM: データマイニング方法論

(**CR**oss **I**ndustry **S**tandard **P**rocess for **D**ata **M**ining)



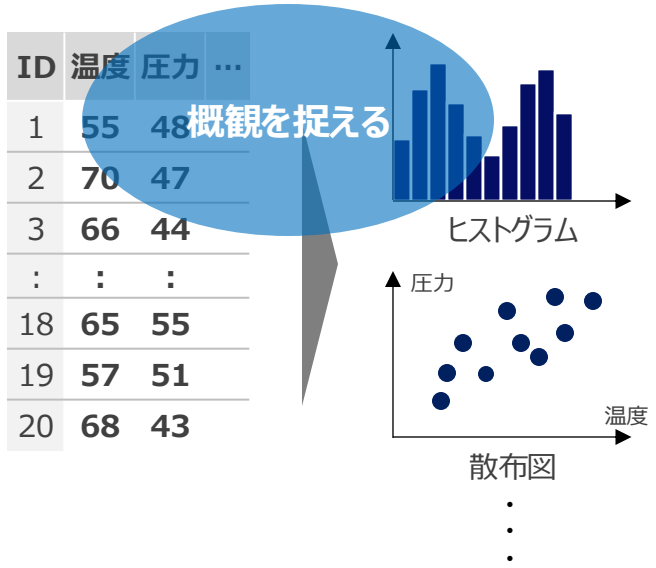
1.ビジネスの理解	•ビジネス、データマイニング目標の決定 •プロジェクトの立ち上げ
2.データの理解	•データの収集 •データの調査 •データ品質の検証
3.データの準備	•データの選択や除外 •データのクリーニング •データの構築や統合
4.モデル構築	•モデリング手法の選択 •モデルの作成 •モデルの評価
5.評価	•データマイニングの結果の評価 •プロセスの見直し •実行可能なアクションリストの作成
6.展開／共有	•業務への導入計画 •モニタリング、メンテナンスの計画

記述統計学の基本：データの特徴を捉える

■ 個々のデータをくまなく見るのは難しいため、グラフ（ヒストグラムや散布図）や
要約統計量（平均値や標準偏差） を用いて全体傾向を把握する

可視化（グラフ化）

視覚的にデータの特徴や傾向を把握



数値化（データ要約）

データの特徴を示す値に要約し
比較可能な客観的傾向を掴む
（**要約統計量**）

集団の比較

- あなたは、「飲むログ」でデートで使うお店を選んでいます。
  候補となったお店は2つ。でもレビュー点数が全く同じです。
- 特徴を捉えるにはどうすれば良い？

レストランA

★★★½☆ 3.45

#	評点
1	3.5
2	3.5
3	4.5
4	3.5
5	3.5
6	2.5
7	3.0
8	3.5
9	4.0
10	3.0

#	評点
11	3.0
12	3.5
13	3.0
14	4.0
15	4.0
16	4.0
17	2.5
18	4.5
19	1.0
20	5.0

どっちも同じ評価だからどっちでもいいのかな・・・

レストランB

★★★½☆ 3.45

#	評点
1	3.0
2	4.5
3	5.0
4	4.5
5	2.5
6	2.5
7	4.0
8	2.0
9	5.0
10	2.0

#	評点
11	1.0
12	2.0
13	4.5
14	2.0
15	5.0
16	4.5
17	4.5
18	4.0
19	2.5
20	4.0

ヒストグラム（度数分布図）

- ヒストグラムは、データを一定間隔（**階級**）ごとに頻度集計（**度数**）をとり、横軸に階級、縦軸に度数をとって柱状（**ビン**）の集合で表したグラフ
- データの分布状況を可視化して、直感的にデータの特徴を捉えやすくする目的がある

元データ

レストランAの
レビューアーごとの評点データ

#	評点
1	3.5
2	3.5
3	4.5
4	3.5
5	3.5
6	2.5
7	3.0
8	3.5
9	4.0
10	3.0

#	評点
11	3.0
12	3.5
13	3.0
14	4.0
15	4.0
16	4.0
17	2.5
18	4.5
19	1.0
20	5.0

度数分布表

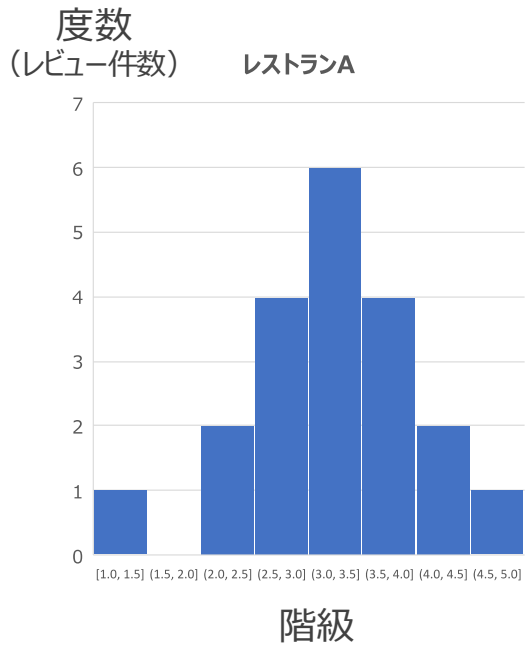
一定間隔（階級）ごとに
レビュー件数（度数）を集計

階級	度数
0.0~0.5	0
0.5~1.0	0
1.0~1.5	1
1.5~2.0	0
2.0~2.5	2
2.5~3.0	4
3.0~3.5	6
3.5~4.0	4
4.0~4.5	2
4.5~5.0	1

※境界値は小さい側の階級に含める
（ただし、1.0は便宜上1.0～1.5に含める）

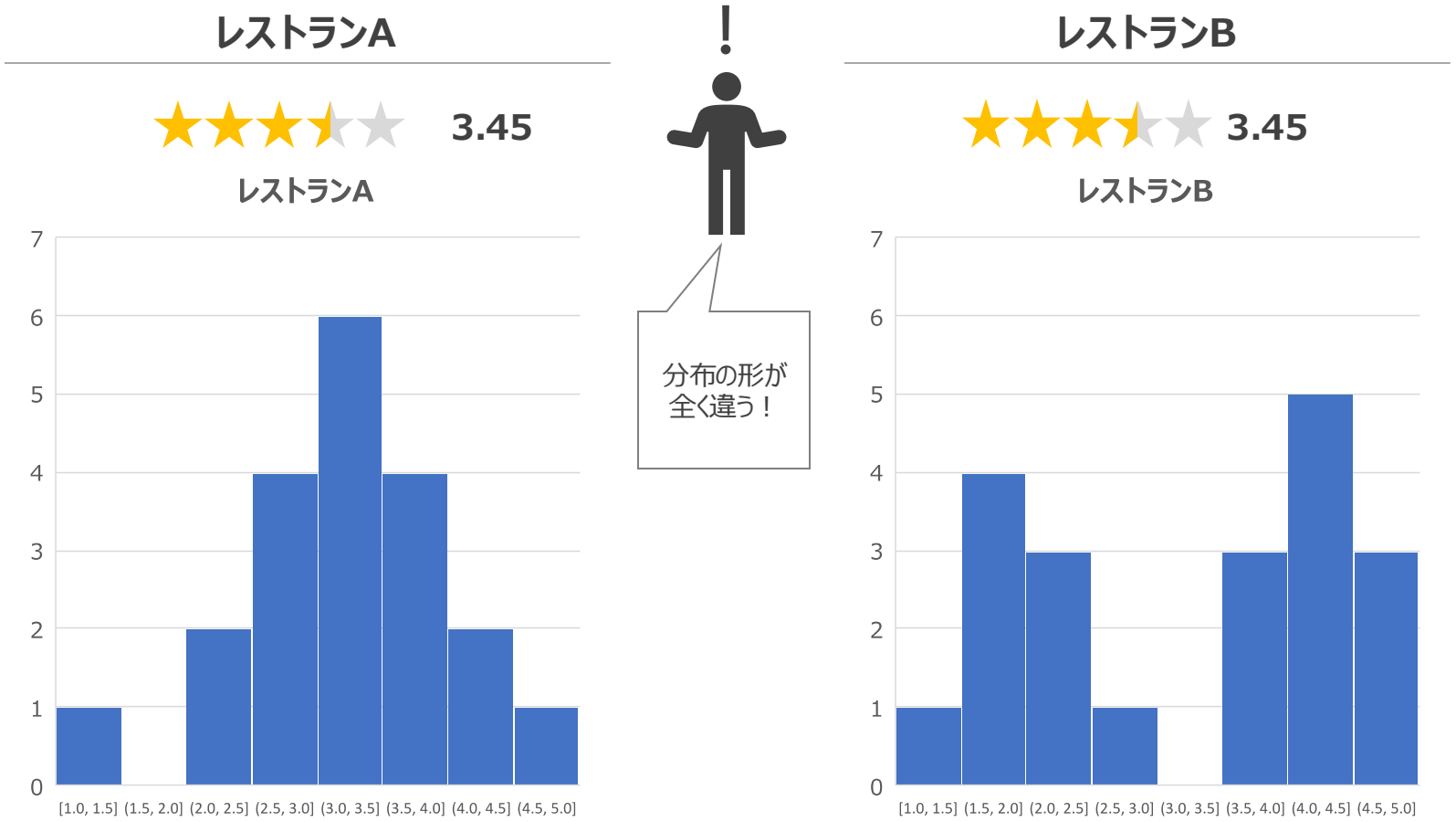
ヒストグラム（度数分布図）

横軸に階級、縦軸に度数を取り、
柱（ビン）の集合で可視化



ヒストグラムの描画と比較

■ ヒストグラムを描くと、「分布の形」（≒データの散らばり具合）が可視化される

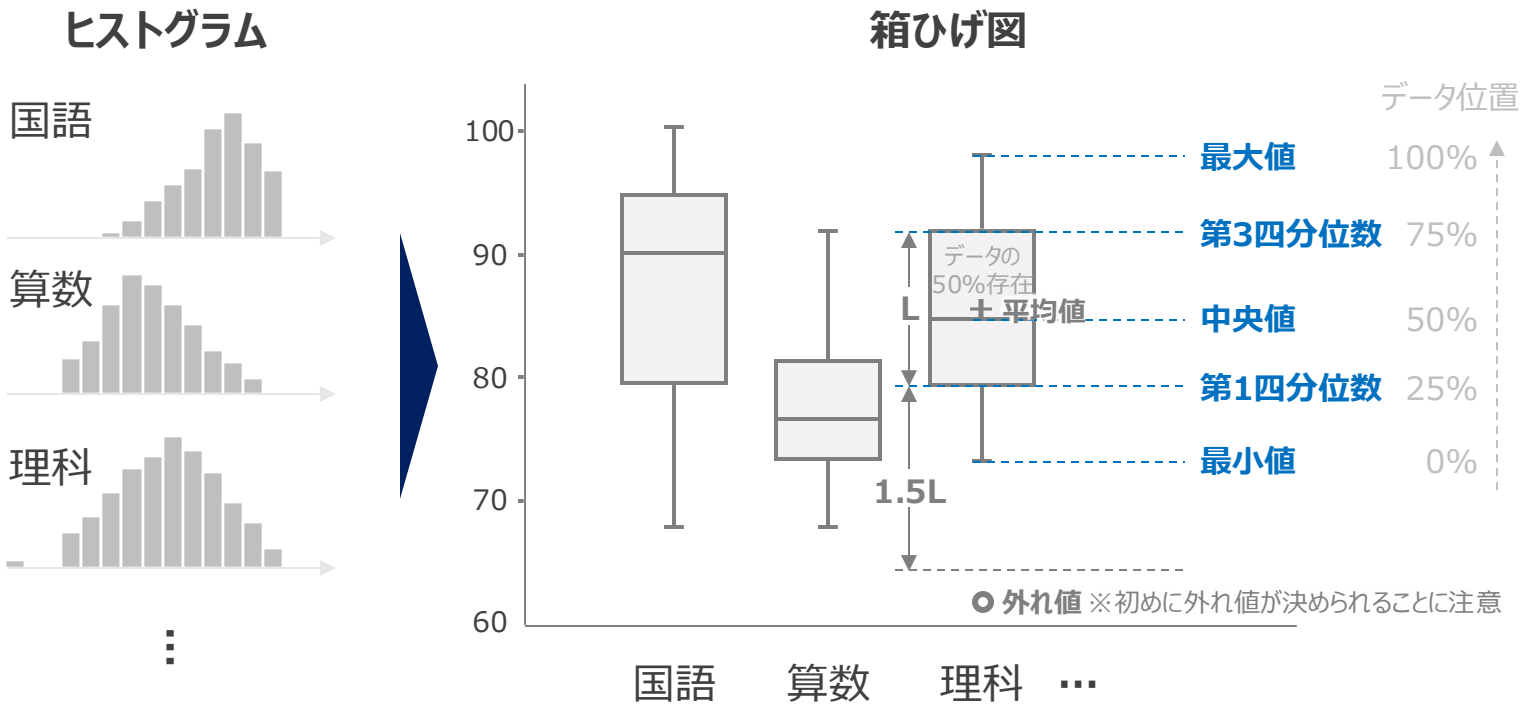


評価は安定している
→ ハズレは少ない？

評価は二つに分かれる
→ ワンチャンに賭けるか？

箱ひげ図

■ 分布形状の把握は極めて重要であるが、主にヒストグラムは1変数、散布図は2変数での分布可視化に適している。一方、箱ひげ図は、分布の概形を**多変数間で比較**するのに適している

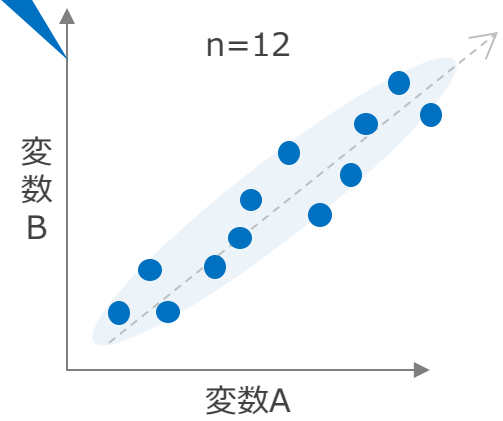


散布図 と 相関

- 二つの値（**変数**）間において、一方が上がれば他方も上がる（or 下がる）ような関係性のことを「**相関**」と呼ぶ
- 各変数を各軸にとってグラフ化した「**散布図**」を描くことで、相関関係の視覚的な把握が可能

散布図

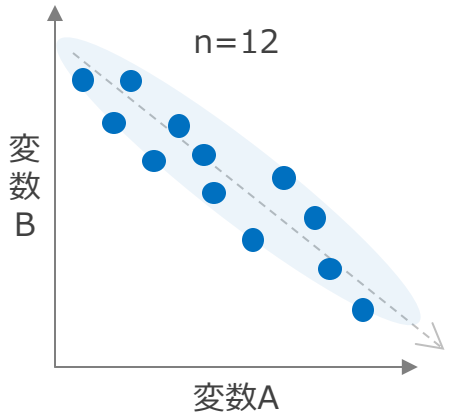
正の相関



一方 (A) が増加すると
他方 (B) も増加
(右上がりの楕円状)

例）A･･･(夏の)平均気温
B･･･エアコンの販売台数

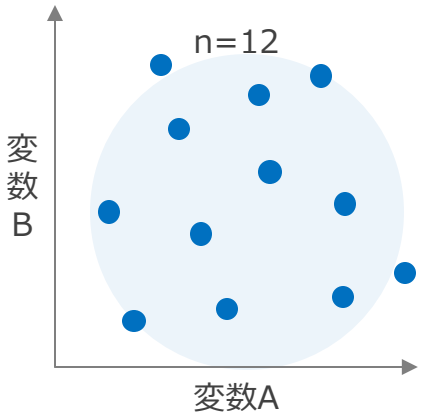
負の相関



一方 (A) が増加すると
他方 (B) は減少
(右下がりの楕円状)

例）A･･･(冬の)平均気温
B･･･ストーブの販売台数

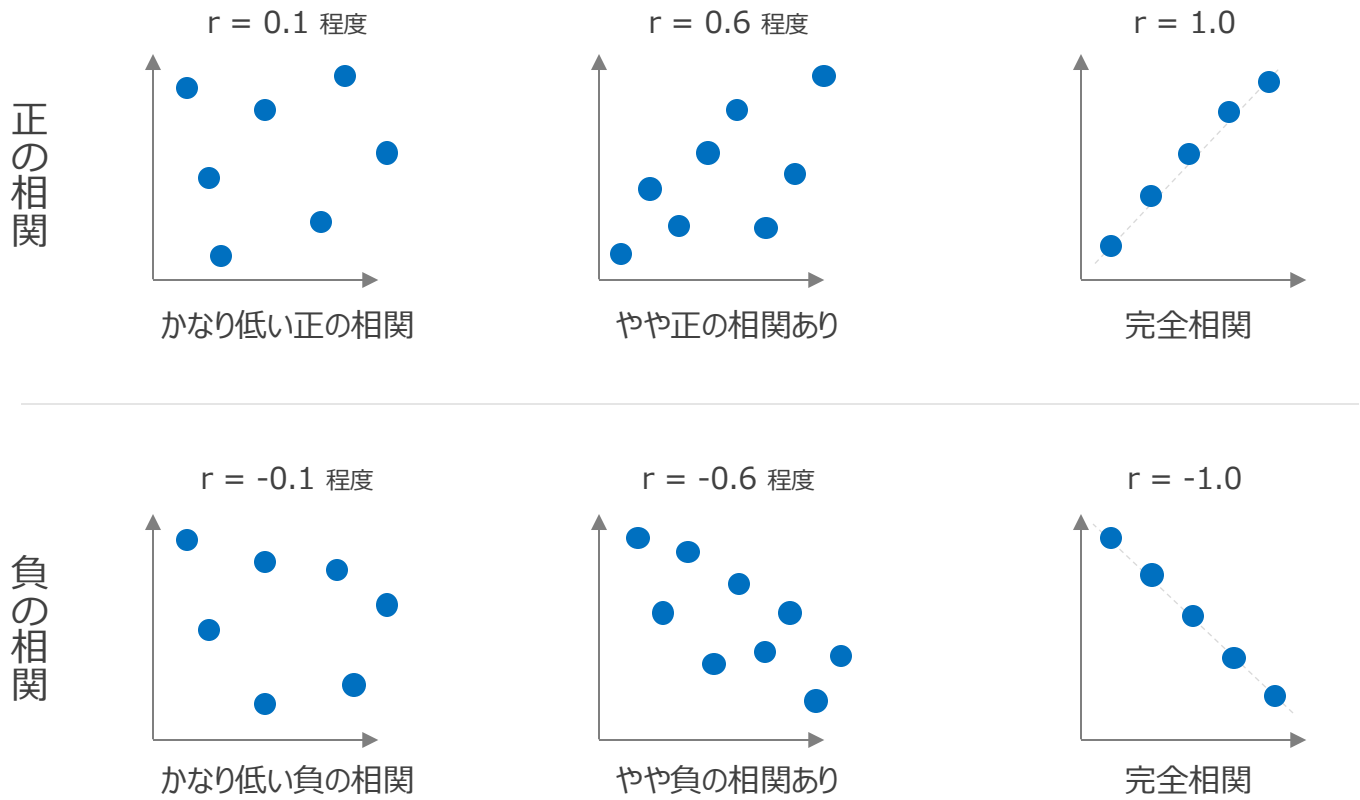
無相関



双方の変数間に
関連性なし
(まばらな円状)

相関係数

■ 相関係数r (correlation coefficient) とは、2つの変数間の相関の度合いを表す指標
■ -1 ≦ r ≦ 1 の値を取り、正の場合は正相関、負の場合は負相関、0の場合は無相関



（参考）**相関**係数の目安

相関係数（絶対値）	相関の強弱
1.0	**強い相関**
0.7	**やや相関あり**
0.4	**弱い相関**
0.2	**ほぼ相関なし**
0	

相関行列による網羅的な変数間の関係性把握

- 事前に各変数同士の相関係数を総当たりで調べておくと、後々の結果解釈に役立つ（**相関行列**）
- また、**共線性が高い変数** (相関の高い) が複数混ざっていると、その変数の影響を強く受け、偏った分析結果になることがある。この場合、共線性が高い変数は除外することが有効

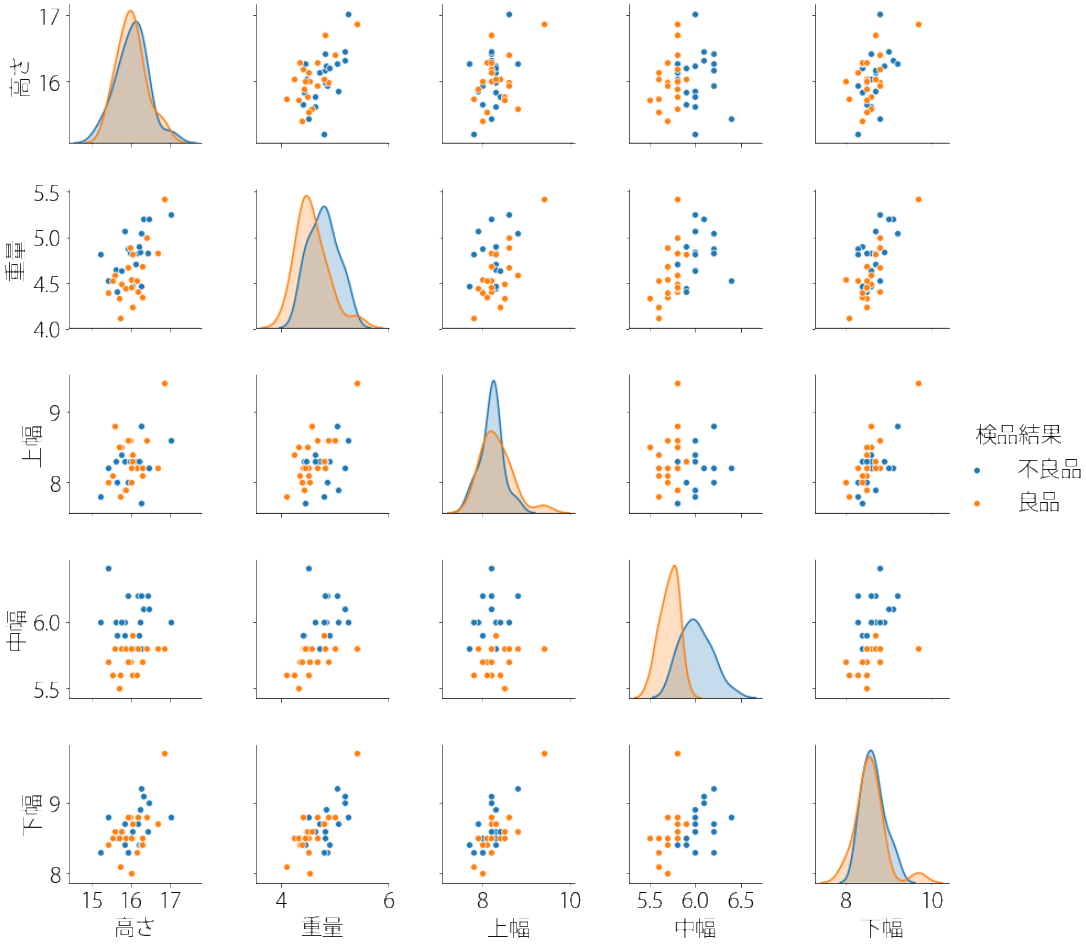
	打率	試合	打席数	打数	安打	本塁打	打点	盗
打率	1.00							
試合	-0.22	1.00						
打席数	0.02	0.85	1.00					
打数	-0.03	0.83	0.95	1.00				
安打	0.56	0.58	0.80	0.81	1.00			
本塁打	0.00	0.34	0.48	0.38	0.30	1.00		
打点	0.07	0.40	0.48	0.39	0.36	0.92	1.00	
盗塁	[illegible]	0.28	[illegible]	[illegible]	[illegible]	[illegible]	-0.17	
四球	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.60	
死球	0.20	0.14	[illegible]	[illegible]	0.11	-0.16	-0.02	
三振	-0.34	0.44	0.47	0.41	0.13	0.63	0.57	
犠打	-0.33	0.12	0.05	0.09	-0.11	-0.49	-0.54	
併殺打	-0.06	0.08	0.10	0.12	0.05	0.32	0.46	
出塁率	0.73	-0.10	0.10	-0.14	0.32	0.31	0.34	
長打率	0.38	0.11	0.32	0.19	0.38	0.90	0.86	
OPS	0.55	0.04	0.27	0.09	0.40	0.77	0.76	
RC27	0.63	0.04	0.28	0.08	0.43	0.65	0.64	
XR27	0.59	0.07	0.30	0.09	0.42	0.70	0.67	

各変数同士の**総当たり形式**で相関係数を算出
(相関行列)

散布図行列と色分け可視化

■ 各変数を個別に見るだけでなく、「**散布図行列**」として網羅的に各変数の関係性を俯瞰したり、
目的変数に応じて色分けすることで、影響因子を仮説立てできることもある

目的変数で色分けした散布図行列の例



Google Colaboratory上での
レクチャー＆演習

（参考）変数の尺度（名義尺度・順序尺度・間隔尺度・比例尺度）

- 変数の種類は大きく「**質的データ（カテゴリーデータ）**」と「**量的データ（数量データ）**」に分けられ、それぞれの特性に合わせて扱う必要がある
- 数量データは先述の記述統計量、グラフが有効であるが、カテゴリーデータはクロス集計が有効

クロス集計（次頁）が有効！

種類	変数の尺度	概要	データの例	扱い方 大小 (A<B)	差分 (A-B)	比率 (A/B)
質的データ（カテゴリーデータ）	名義尺度	単にデータを区別するための分類ラベル。演算不可で、順序も意味をなさない	▪性別、血液型、顧客ID ▪作業者、個品ID、良品/不良品	－	－	－
	順序尺度	**順序**（大小関係）にのみ意味がある尺度。したがって、平均値は意味を持たないが、順序統計量（最大・最小など）は算出可能	▪顧客満足度、震度 ▪書道の「段」や検定の「級」 ▪不良レベル、工程順序	●	－	－
量的データ（数量データ）	間隔尺度	数値演算可能だが、**値の差**のみに意味がある尺度。0はあくまで相対的な位置関係でしかない	▪年齢、西暦、偏差値 ▪温度（℃）、製造日時	●	●	－
	比例尺度	数値演算可能で、値の差に加え、**値の比**にも意味がある尺度。0が「何もない」という絶対的な意味を持つ	▪身長、売上金額 ▪寸法、圧力、作業時間、絶対温度	●	●	●

※集計によるカウントのみ可能

平均値に意味あり

（参考）クロス集計とは

- カテゴリーデータは、平均値や標準偏差などの記述統計量を算出してもあまり意味がないため、当該データ項目に登場する**カテゴリごとの頻度集計を算出**することが多い
- 特に、複数のデータ項目を掛合せて算出（例えば年代x性別）した頻度集計表を**クロス集計**と呼ぶ

▼ある血液検査の結果

年代	性別	血清ナトリウム	…
70	男性	136	…
50	男性	136	…
60	男性	129	…
50	男性	137	…
60	女性	116	…
90	男性	132	…
70	男性	137	…
60	男性	131	…
60	女性	138	…
80	男性	133	…
70	男性	131	…
60	男性	140	…
40	男性	137	…
50	男性	137	…
40	女性	138	…
80	男性	136	…
80	男性	140	…
⋮	⋮	⋮	

カテゴリーデータのため平均値などの統計量に意味を持たない

▼性別ごとの頻度集計

性別	人数
男性	194
女性	105
合計	**299**

▼年代ごとの頻度集計

年代	人数
40	47
50	82
60	93
70	52
80	19
90	6
合計	**299**

複数データ項目を掛け合わせる

▼性別x年代のクロス集計表

年代	性別		人数
	男性	女性	合計
40	27	20	**47**
50	56	26	**82**
60	58	35	**93**
70	34	18	**52**
80	15	4	**19**
90	4	2	**6**
合計	**194**	**105**	**299**

▼応用編：他項目の「平均値」を表示

年代	性別		血清ナトリウム
	男性	女性	平均値
40	137.1	135.6	136.4
50	137.3	136.9	137.1
60	136.0	137.1	136.4
70	136.7	136.6	136.7
80	135.5	138.0	136.1
90	133.0	141.0	135.7
血清ナトリウム平均値	**136.5**	**136.8**	**136.6**

Excelのピボットテーブルのようなイメージ

（参考）「データの理解」でよく用いる関数

確認内容		関数	使用例
基本情報	データの先頭/末尾確認	データフレーム.head(行数) / tail(行数)	df.head(10) / df.tail(10)
	データ件数確認	データフレーム.shape	df.shape
	各列のデータ型などの確認	データフレーム.info()	df.info()
	各列の欠損値確認	データフレーム.isnull().sum()	df.isnull().sum()
統計値	要約統計量の算出	データフレーム.describe()	df.describe()
	相関行列	データフレーム.corr()	df.corr(numeric_only=True) #文字列の列がある場合”numeric_only=True”の指定が必要
集計値	クロス集計	pd.crosstab(行見出し, 列見出し)	pd.crosstab(df['役割'], df['[成約]契約まで'])
グラフ	散布図行列	sns.pairplot(df, hue='色分け対象列名') ※seabornのインポート必要	sns.pairplot(df, hue='[成約]契約まで')
	ヒストグラム	データフレーム.hist()	df.hist()
	箱ひげ図	データフレーム.boxplot()	df.boxplot()
	棒グラフ （縦棒／横棒／積み上げ）	縦棒：データフレーム.plot.bar() 横棒：データフレーム.plot.barh()	df.plot.bar(), df.plot.barh() ※積上棒グラフは df.plot.bar(**stacked=True**)
	その他	過去資料を参考に	

データ加工の基本

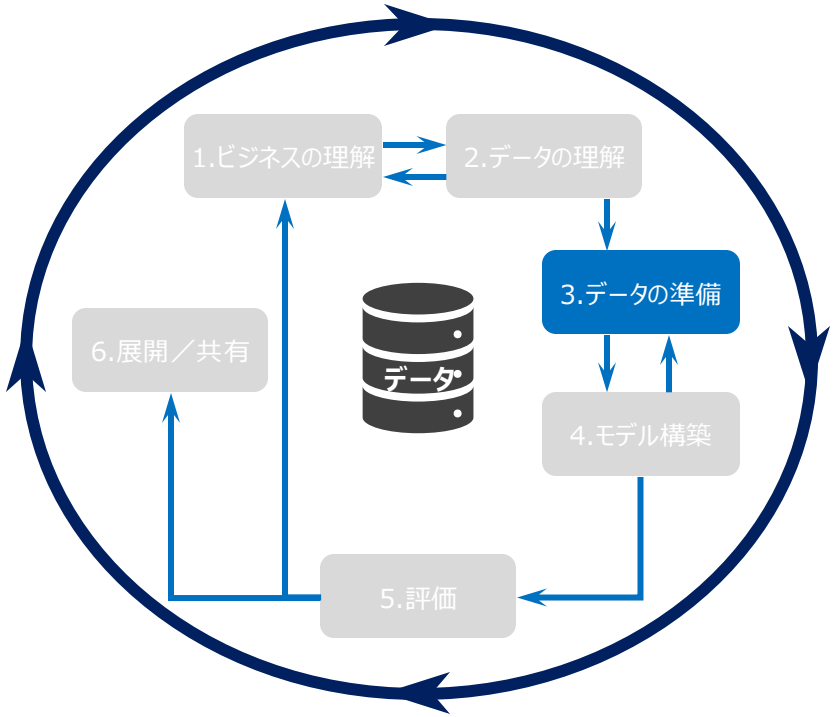
✓ 外れ値／異常値／欠損値の確認と対応
✓ カテゴリーデータとダミー変数化

データ分析の進め方

■ データ分析の進め方に関する方法論「CRISP-DM」に基づいて、分析と評価を繰り返して試行錯誤しながら進めるのが一般的である

CRISP-DM: データマイニング方法論

(**CR**oss **I**ndustry **S**tandard **P**rocess for **D**ata **M**ining)



1.ビジネスの理解	•ビジネス、データマイニング目標の決定 •プロジェクトの立ち上げ
2.データの理解	•データの収集 •データの調査 •データ品質の検証
3.データの準備	**•データの選択や除外 •データのクリーニング •データの構築や統合**
4.モデル構築	•モデリング手法の選択 •モデルの作成 •モデルの評価
5.評価	•データマイニングの結果の評価 •プロセスの見直し •実行可能なアクションリストの作成
6.展開／共有	•業務への導入計画 •モニタリング、メンテナンスの計画

「データの準備」で主に行うこと

■ 前段のデータの理解や後段のモデル構築と並行しながら、必要に応じてデータ加工を行う

レコード方向の増減

クレンジング

今回の演習のターゲット

- 欠損値を含むレコードの除外／補完
- 異常値を含むレコードの除外／修正

「データ観察」で行うことが多い

サンプリング

- 学習データ／テストデータの分割
- 不均衡データのバランシング（アンダー／オーバーサンプリング）

「モデル構築」で行うことが多い

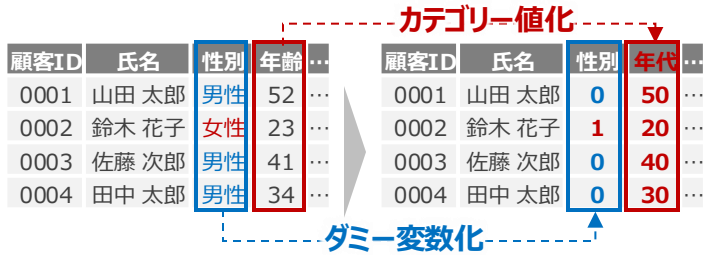
カラム方向の増減

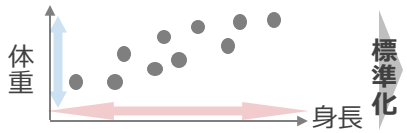
データの統合

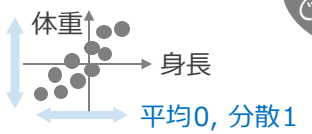
- 複数データファイルの結合

変数の加工
（説明変数・目的変数）

- カテゴリーデータのダミー変数化
- 数値データのカテゴリー値化
- 数値データの標準化（スケーリング）
- 四則演算による変数加工







「モデル構築」で行うことが多い

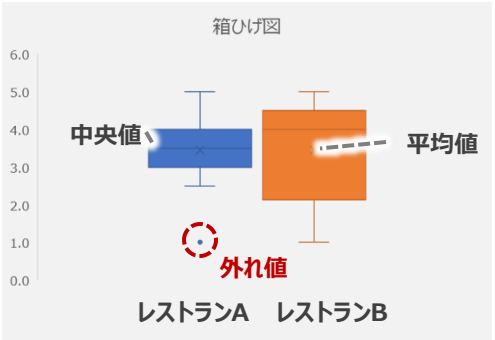
外れ値／異常値／欠損値の確認と対応

■ データに外れ値や異常値、欠損値が存在した場合、**分析結果に悪影響を与える可能性**があるため、事前に存在を確かめ、対応方法を検討しておくことが重要

外れ値と異常値

データの大部分から大きく外れた値

外れ値のうち、原因がハッキリしているもの
(測定ミス、記録ミスなど)



箱ひげ図の描画により外れ値の検出が可能

欠損値

欠損値

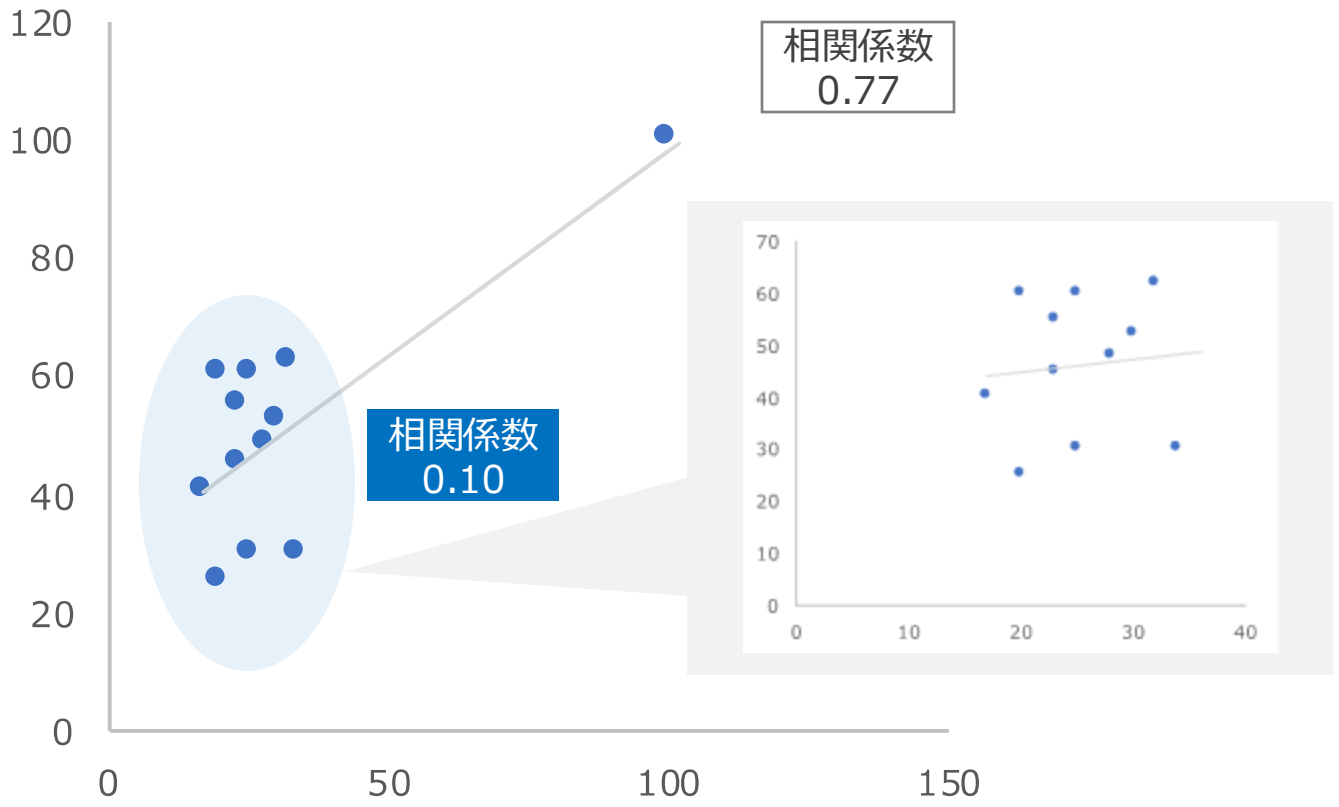
A	B	C
121	15	187
	21	175
120	12	177
115	25	169
⋮	⋮	⋮

①行ごと除外
※あまりに欠損の多い列は列ごと除外

②補完
（平均値／中央値、予測値など）

外れ値の影響例

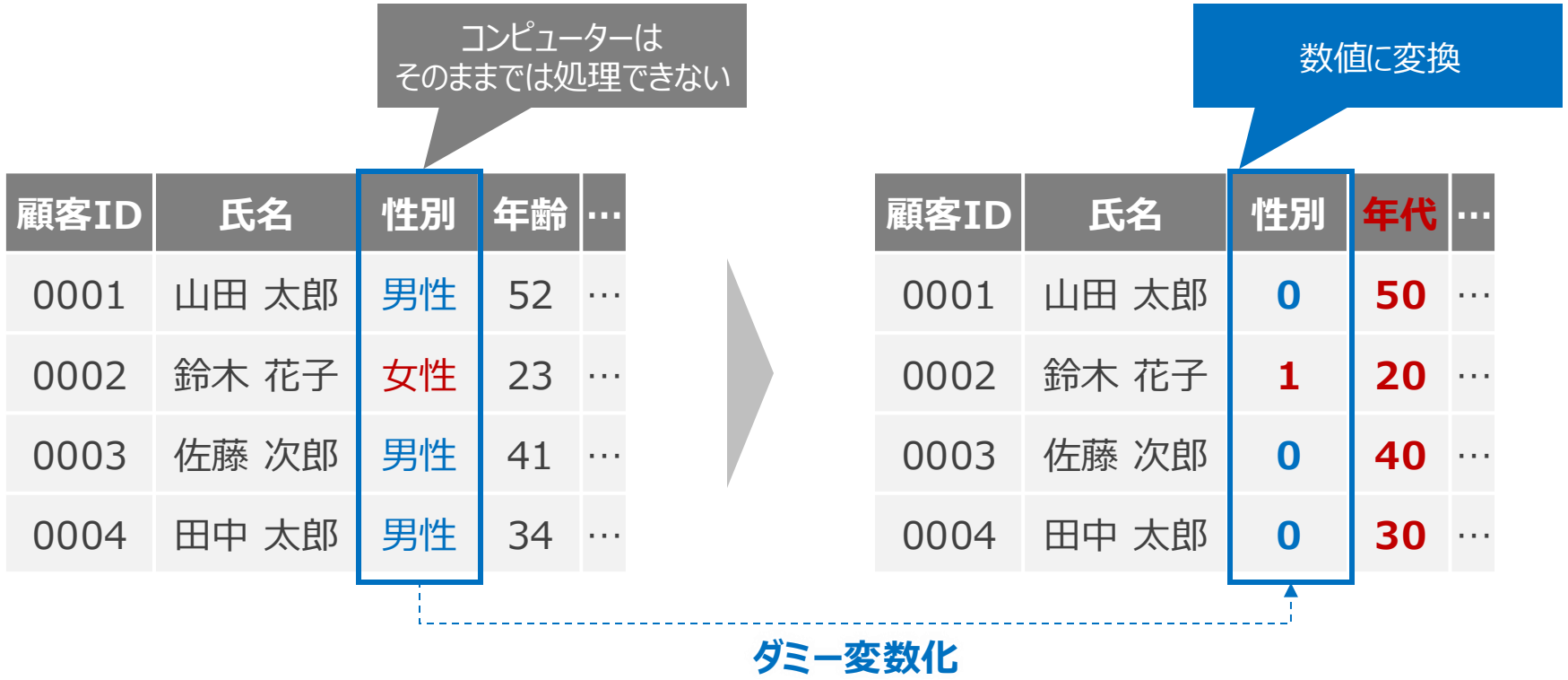
■ 相関係数は外れ値の影響を大きく受けるため、数字だけに惑わされぬよう、
散布図の確認も併せて行うことが重要である



Google Colaboratory上での
レクチャー＆演習

カテゴリデータ（文字列データ）と ダミー変数化

- 通常の数値データに対して、血液型（A, B, O, AB）や部署（経営企画部、DX推進部、・・・）、性別など、分類上の区別に用いられる変数を「**カテゴリデータ**」と呼ぶ
- カテゴリデータは通常文字列で表されるため、擬似的に数値変換する「**ダミー変数化**」が必要



ダミー変数化の種類

■ ダミー変数化のメジャーな考え方は2つあり、男性／女性、契約／未契約のような**2カテゴリのデータは連続値化**をし、職種や部署のような**カテゴリ間に連続性がない場合には横持ち化**を行う

❶ 連続値化（Label/Ordinal Encoding）

2カテゴリ（男／女）やアンケートのような連続的な尺度に有効

元データ → ダミー変数化

- 男性 → **0**
- 女性 → **1**

- 不満である → **0**
- どちらでもない → **1**
- 満足している → **2**

❷ 横持ち化（One Hot Encoding）

カテゴリ間に直接的な連続性がない場合に有効

元データ → ダミー変数化

…	職種
…	会社員
…	自営業
…	会社員
…	会社員
…	会社員
…	医師
…	⋮

…	自営業	医師	会社員
…			1
…	1		
…			1
…			1
…			1
…		1	
…	⋮		

- 自営業 → **0**
- 会社員 → **1**
- 医師 → **2**

この場合、連続値化は不適

Google Colaboratory上での
レクチャー＆演習

データ分析（モデル構築）の基本

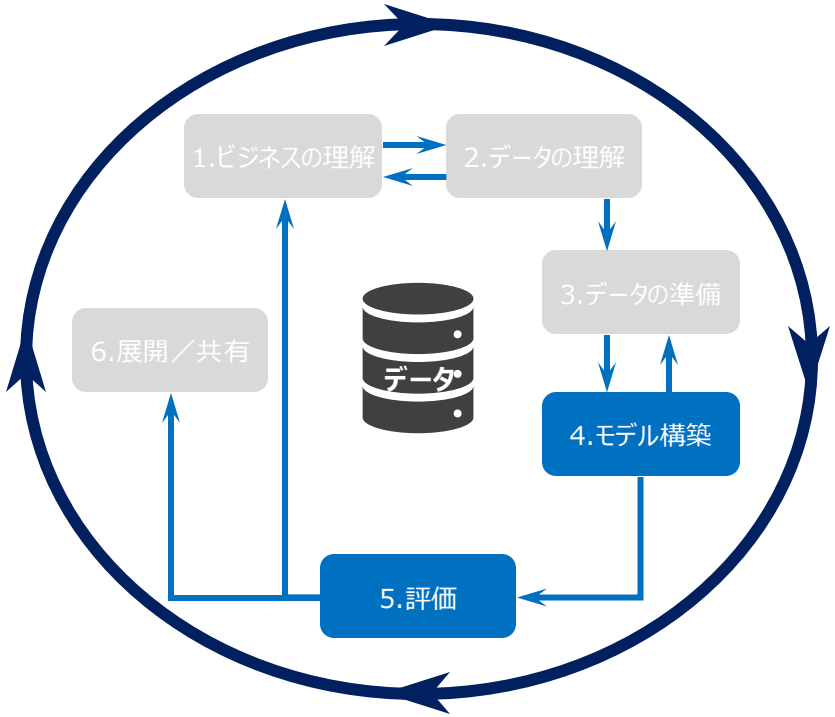
✓ 機械学習手法の基本
✓ 回帰分析による数値予測と影響因子の調査

データ分析の進め方

■ データ分析の進め方に関する方法論「CRISP-DM」に基づいて、分析と評価を繰り返して試行錯誤しながら進めるのが一般的である

CRISP-DM: データマイニング方法論

(**CR**oss **I**ndustry **S**tandard **P**rocess for **D**ata **M**ining)

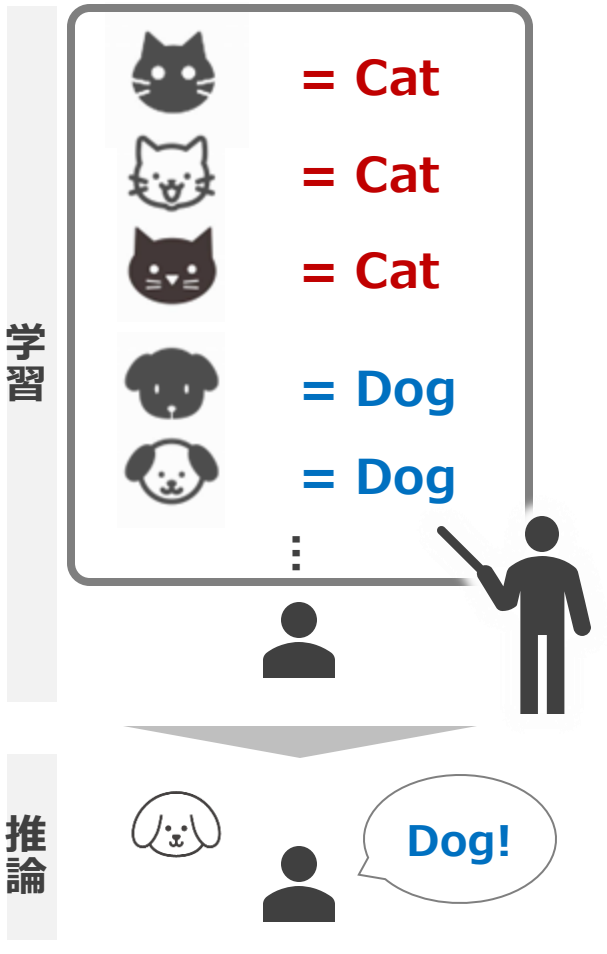


1.ビジネスの理解	•ビジネス、データマイニング目標の決定 •プロジェクトの立ち上げ
2.データの理解	•データの収集 •データの調査 •データ品質の検証
3.データの準備	•データの選択や除外 •データのクリーニング •データの構築や統合
4.モデル構築	•モデリング手法の選択 •モデルの作成 •モデルの評価
5.評価	•データマイニングの結果の評価 •プロセスの見直し •実行可能なアクションリストの作成
6.展開／共有	•業務への導入計画 •モニタリング、メンテナンスの計画

機械学習の分類

教師あり学習

あらかじめ「正解」を与えて
各データと正解の関係を学習させる



教師なし学習

「正解」を与えずに、各データ
のパターンから自分で学習する

※分類されたグループの意味づけは人が行う

教師あり学習のイメージ（数値予測とクラス分類）

■ 各顧客レコードに対して数値もしくはカテゴリー値（クラス）の**解答**を与え関係性を学習

数値予測
（回帰）

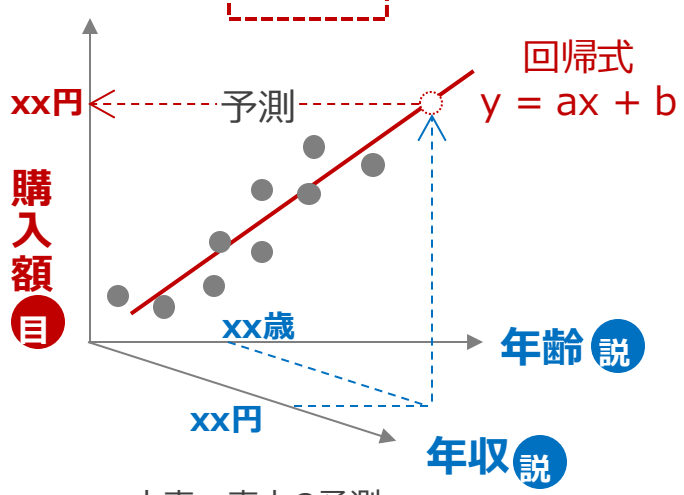
クラス分類
（2クラス or 多クラス分類）

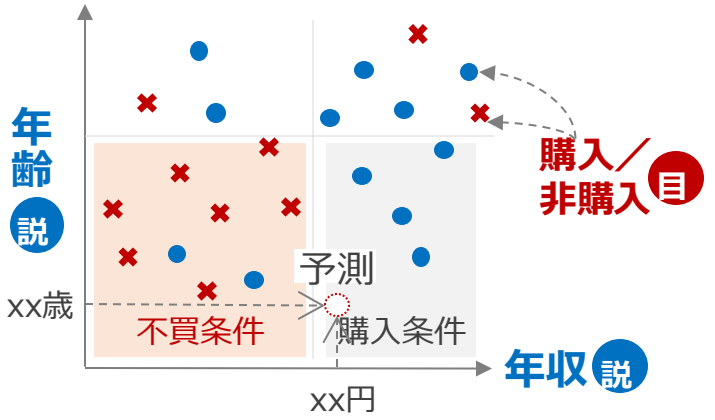
データ例

説明変数 **目的変数**

顧客ID	名前	年齢	年収	購入額	購入有無	…
0001	xx	25	300万	35,000	購入	…
0002	xx	35	600万	68,000	購入	…
0003	xx	18	120万	0	非購入	…
0004	xx	42	820万	85,000	購入	…
⋮	⋮	⋮	⋮	⋮	⋮	…

イメージ





分析例

- 小売：売上の予測
- 小売：需要予測
- 製造：不良発生率の予測

- 小売：購入／非購入顧客の分類
- 医療：生存条件の分析
- 製造：故障種類の分類

教師なし学習のイメージ（クラスタリング）

■ 各データ間の距離に基づき、近接データ（＝類似度が高いデータ）同士のグループ（**クラスタ**）を作り、データを分類する手法。**学習データなし**でデータを大きく層別したい場合に有効

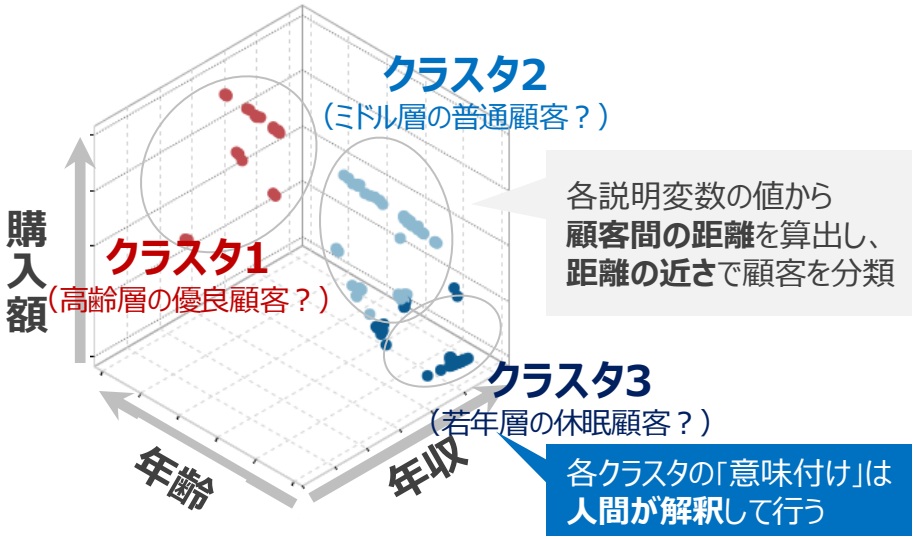
データ例

顧客ID	名前	年齢	年収	購入額	購入有無	…
0001	xx	25	300万	35,000	購入	…
0002	xx	35	600万	68,000	購入	…
0003	xx	18	120万	0	非購入	…
0004	xx	42	820万	85,000	購入	…
⋮	⋮	⋮	⋮	⋮	⋮	…

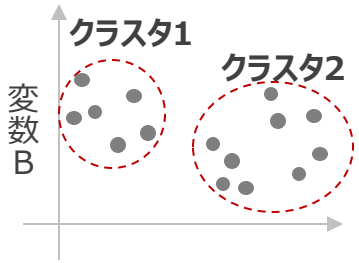
説明変数

※目的変数は無し

クラスタリング



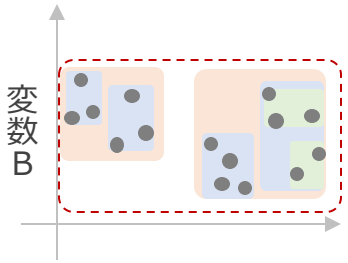
非階層的クラスタリング

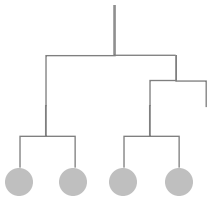


主な手法

- **k-means法**（k平均法）
- 混合ガウス

階層的クラスタリング





主な手法

- 最短距離法
- 最長距離法
- 群平均法
- ウォード法

（参考）代表的な機械学習手法

■ 目的やデータの特徴に合わせて手法を選ぶが、適宜、**手法の試行錯誤**が必要となることも多い

	手法		概要	主なアルゴリズム
教師あり（答え合わせあり）	数値予測	数値予測	連続的な数値に対して、いくつかの影響因子を用いてある関数モデル（y=f(x)）を当てはめる	■ 線形回帰／非線形回帰（2変数：単回帰、多変数：重回帰） ■ 回帰木 ■ ニューラルネットワーク
		時系列予測	継時的な変動に対して、過去の履歴やトレンドから将来の動きを予測	■ ARモデル（自己回帰モデル） ■ ARIMAモデル（自己回帰移動平均モデル）
	分類	クラス分類（多クラス/ニクラス）	あらかじめ分類のラベル付けがされたデータに対して、いくつかの影響因子を用いて、分類境界を抽出する	■ ロジスティック回帰 ■ 決定木分析 ■ 判別分析 ■ ニューラルネットワーク ■ サポートベクタマシン（SVM）
教師なし（答え合わせなし）		クラスタリング（グルーピング）	データに含まれる各変数の特徴から、類似データをグルーピングする	■ K-means ■ 階層的クラスタリング
		異常値検出	データの分布に基づき、分布から外れる異常な値を検出する	■ k-近傍法（k-NN） ■ Local Outlier Factor（LOF）
	パターン抽出（アソシエーションルール分析）		頻出するデータのパターン（データの組み合わせ）を抽出する（例：オムツとビールの併売）	■ アプリオリ ■ シーケンス
	データ要約（次元削減）		高次元（多変数）のデータに対して、より少ない変数で説明できる変数や軸を抽出する	■ 主成分分析 ■ 因子分析
強化学習（報酬）	最適化		試行錯誤を通じてある目的達成のための最適な行動を学習（例：自動運転、アルファ碁）	■ Q-Learning ■ SARSA（State–action–reward–state–action） ■ モンテカルロ法

（参考）代表的な機械学習手法

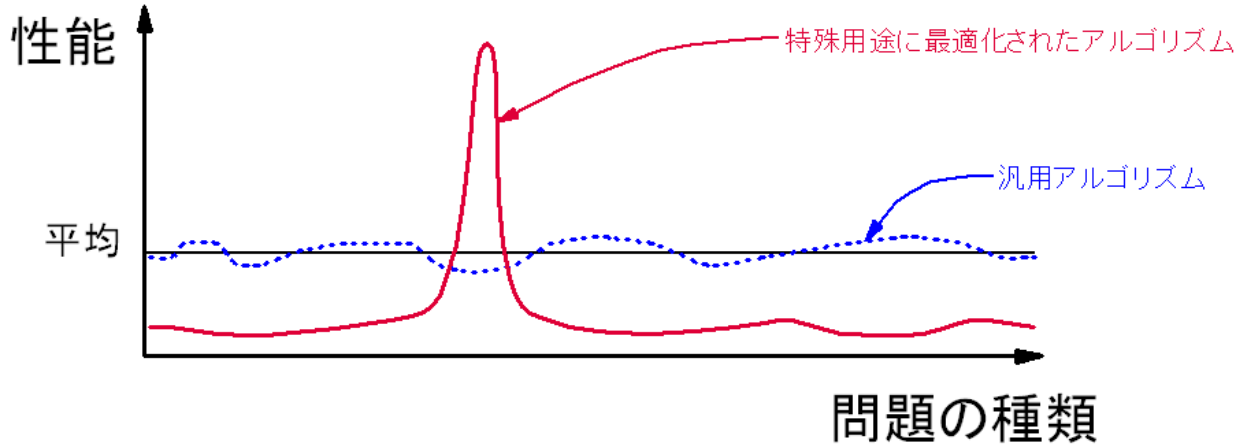
■ 目的やデータの特徴に合わせて手法を選ぶが、適宜、**手法の試行錯誤**が必要となることも多い

	手法		概要	主なアルゴリズム
教師あり（答え合わせあり）	数値予測	数値予測	連続的な数値に対して、いくつかの影響因子を用いてある関数モデル（y=f(x)）を当てはめる	■ **線形回帰**／非線形回帰（2変数：単回帰、多変数：重回帰） ■ 回帰木 ■ ニューラルネットワーク
		時系列予測	継時的な変動に対して、過去の履歴やトレンドの変動の要因	■ ARモデル（自己回帰モデル） ■ ARIMAモデル（自己回帰移動平均モデル）
	分類	クラス分類（多クラス/二クラス）	あらかじめ分類のラベル付けがされたデータに対して、いくつかの影響因子を用いて、分類境界を抽出する	■ ロジスティック回帰 ■ **決定木分析** ■ 判別分析 ■ ニューラルネットワーク ■ サポートベクタマシン（SVM）
教師なし（答え合わせなし）		クラスタリング（グルーピング）	データに含まれる各変数の特徴から、類似データをグルーピングする	■ K-means ■ **階層的クラスタリング**
		異常値検出	データの分布に基づき、分布から外れる異常な値を検出する	■ k-近傍法（k-NN） ■ Local Outlier Factor（LOF）
	パターン抽出（アソシエーションルール分析）		頻出するデータのパターン（データの組み合わせ）を抽出する（例：オムツとビールの併売）	■ アプリオリ ■ シーケンス
	データ要約（次元削減）		高次元（多変数）のデータに対して、より少ない変数で説明できる変数や軸を抽出する	■ 主成分分析 ■ 因子分析
強化学習（報酬）	最適化		試行錯誤を通じてある目的達成のための最適な行動を学習（例：自動運転、アルファ碁）	■ Q-Learning ■ SARSA（State-action-reward-state-action） ■ モンテカルロ法

本セミナーで扱う手法

No Free Lunch定理：タダ飯は無い！

- あらゆる問題で高性能を示すモデルは理論上、構築不可能であることが示されている
- **問題設定や分析目的に応じて、最適なアルゴリズムの検討が必要**である



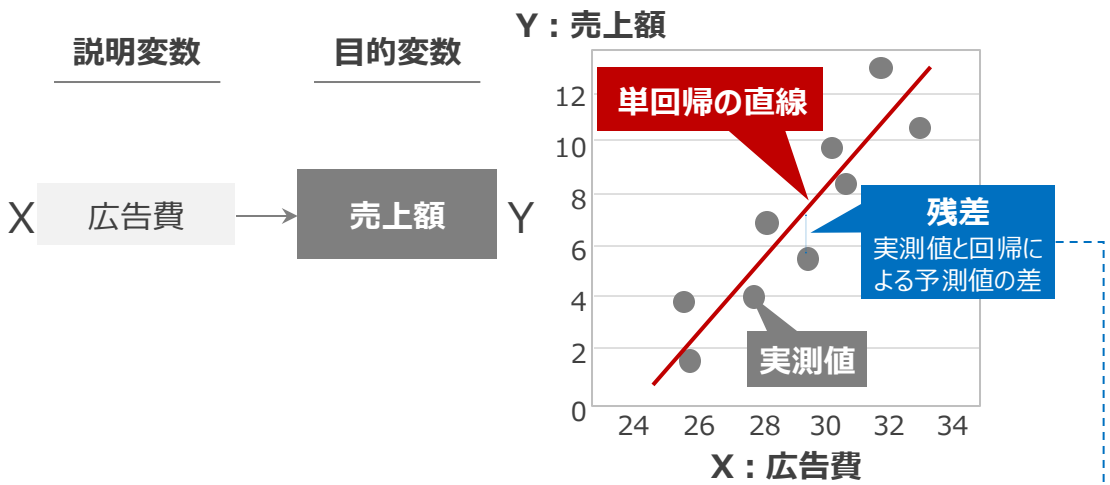
*画像出典：Wikipedia

No-free-lunch theorem: コスト関数の極値を探索するあらゆるアルゴリズムは、
全ての可能なコスト関数に適用した結果を平均すると同じ性能となる（Wolpert and Macready, 1995年）

線形回帰分析による数値予測

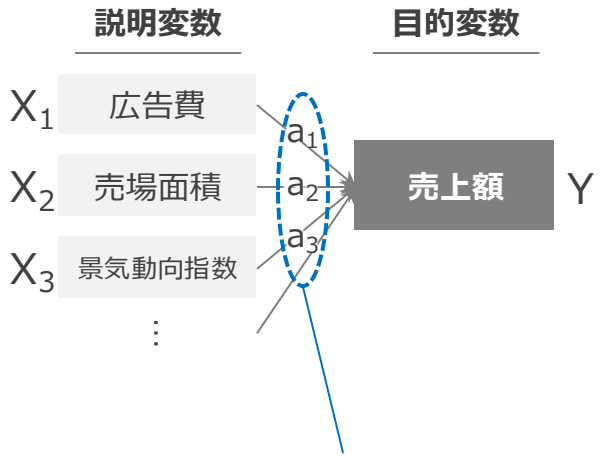
- 目的変数Yとその影響因子である説明変数Xとで関数 (Y=f(X)) に当てはめることを**回帰**という
- 特にf(X)が線形関数の場合、**線形回帰**と呼び、1変数では**単回帰**、多変数では**重回帰**という

単回帰



残差の2乗和が最小になるように
係数a, 定数項b を決める（最小二乗法）

重回帰



各回帰係数は、
各説明変数の影響の大きさを表す

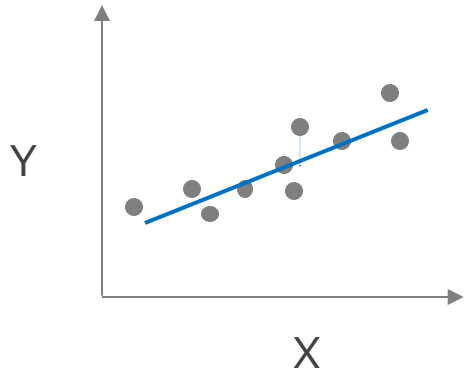
売上額 ＝ a×**広告費** ＋ b

目的変数（Y）	**回帰係数**（傾き）	**説明変数**（X）	**定数項**（Y切片）
結果となる変数	**Xの影響の大きさ**	原因となる変数	広告費=0のときの売上額

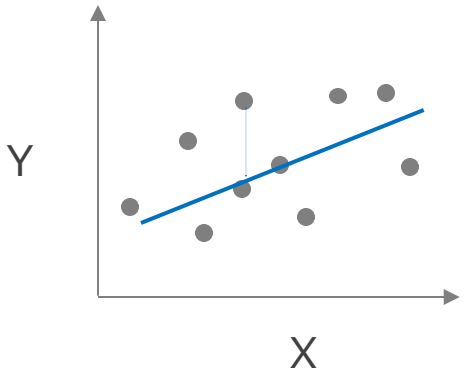
決定係数によるモデルの当てはまり評価

■ モデルの当てはまりの良さは、「決定係数」によって表すことが多い

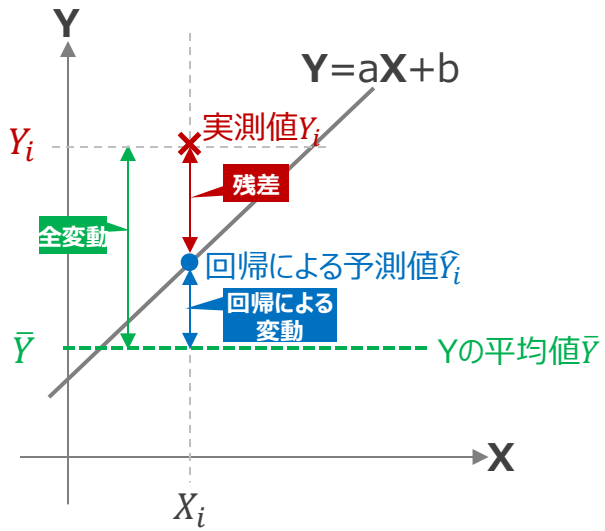
決定係数 大



決定係数 小



▼参考：決定係数の定義イメージ



決定係数 R^2 ＝ **回帰による変動（平方和）** / **全変動（平方和）**

（全変動のうち、回帰式で説明できる変動の割合）

- **0~1**の値をとる
- （一般的な定義では）**相関係数の2乗**に等しい

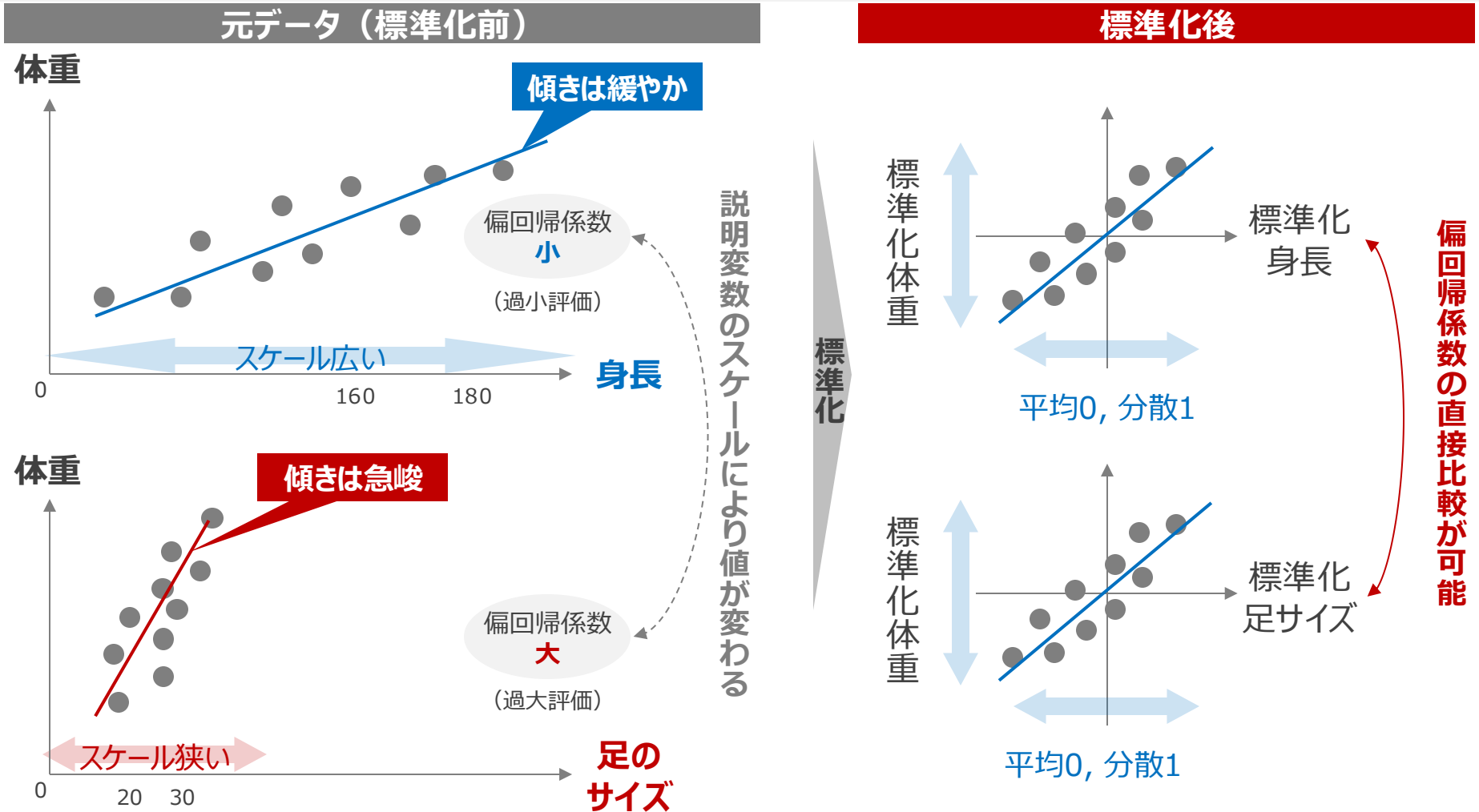
（参考）**相関**係数の目安

相関係数（絶対値）	相関の強弱
1.0	
	強い相関
0.7	
	やや相関あり
0.4	
	弱い相関
0.2	
	ほぼ相関なし
0	

Google Colaboratory上での
レクチャー＆演習

回帰の注意点：データの標準化

- 重回帰分析の偏回帰係数は、目的変数に対する各説明変数の「傾き」を表しており、**その大きさは各変数のスケールに依存**するため、そのままでは直接比較が困難
- 予め**データを標準化**しておくことで、偏回帰係数の直接比較が可能となる

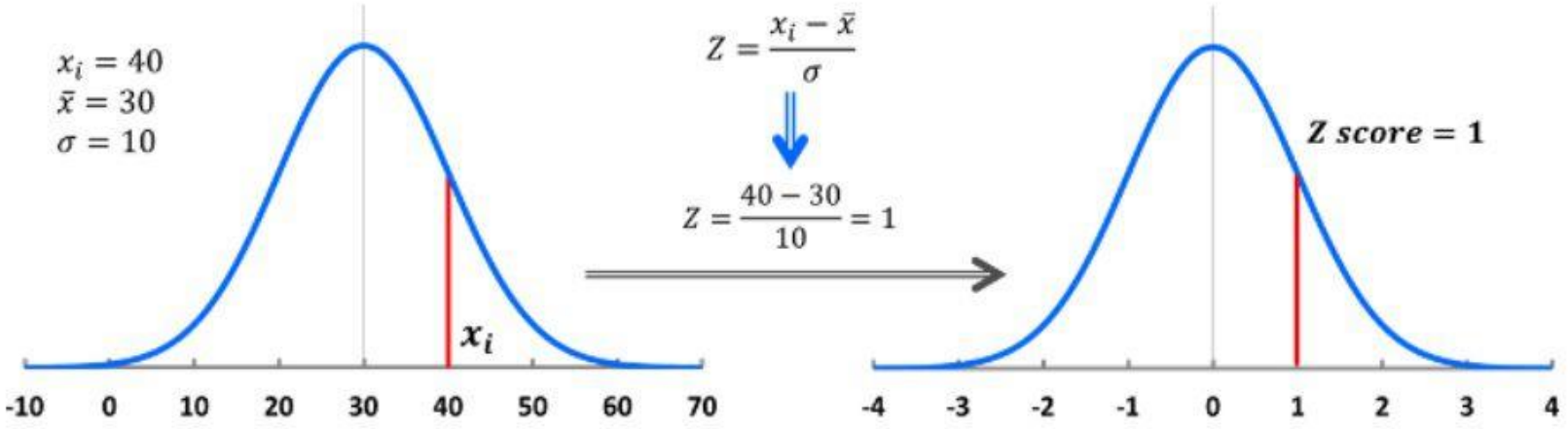


標準化の方法（Zスコア）

- 重回帰分析の偏回帰係数は、各変数のスケールの大小に依存するため、直接比較が困難
- **Zスコア**を算出することでデータの標準化が行え、偏回帰係数の直接比較が可能となる

$$Z\ score = \frac{x_i - \mu}{\sigma}$$

x_i ＝母集団を構成する要素iの値
μ ＝母平均
σ ＝標準偏差



出典：https://www.monodukuri.com/gihou/article/703

（参考）偏差値

偏差値は、「**平均が50、標準偏差が10**」になるように変換したもの： 50 + 10 x Zスコア

Google Colaboratory上での
レクチャー＆演習

Part3

実際のビジネスデータ解析に向けた実践演習

- ✓e-Stat, SSDSEの紹介と使い方
- ✓データ分析の実践演習
- ✓今後の継続学習、実践活用のためのポイント

政府統計のオープンデータサイト “e-Stat”

- オープンデータサイト “e-Stat” では、様々な政府統計を参照・ダウンロードすることができる
  *各府省の統計データを一元的にみれるポータルサイト。総務省統計局が整備し、独立行政法人統計センターが運用管理している
- 国民経済の状況を把握する上で不可欠かつ、唯一無二なデータも多い

■政府統計の総合窓口 e-Stat
https://www.e-stat.go.jp/

教育用標準データセット（SSDSE）

- 教育用標準データセット（SSDSE; Standardized Statistical Data Set for Education）は、、独立行政法人統計センターが作成・公開している、データ分析教育用の統計データ
- 主要な公的統計を地域別に一覧できる表形式のデータセットで、直ちにデータ分析に利用可能

▼SSDSEデータセット一覧

名称	内容
SSDSE-市区町村（SSDSE-A）	**1741市区町村**×多分野125項目 全国の全市区町村の、人口、経済、教育、労働、医療、福祉など、様々な分野の統計データを収録
SSDSE-県別推移（SSDSE-B）	**47都道府県**×12年次×多分野109項目 人口、経済、教育、労働、医療、福祉など、様々な分野の統計データを、12年分の時系列で収録
SSDSE-家計消費（SSDSE-C）	**全国・47都道府県庁所在市**×家計消費226項目 １世帯当たりの食料の年間支出金額（消費額）を、魚介、肉、野菜、果物、菓子、飲料などに分類し、それぞれ詳細な品目別にデータを収録
SSDSE-社会生活（SSDSE-D）	**全国・47都道府県**×男女別×社会生活121項目 男女別に、スポーツ、趣味・娯楽、ボランティアなどの詳細な活動データや、１日の睡眠、食事、学業、家事、仕事、趣味・娯楽などの時間配分データを収録
SSDSE-基本素材（SSDSE-E）	**全国・47都道府県**×多分野90項目 人口、経済、教育、文化、医療、福祉など、様々な分野の統計データを収録し、初学者にも扱いやすいデータセット
SSDSE-気候値（SSDSE-F）	**47都道府県庁所在市**×月・年×気象42項目 気温、気圧、風速、日照、降水、降雪など、様々な気象データについて、月・年別の平年値を収録

出典：独立行政法人 統計センター SSDSE-基本素材（https://www.nstac.go.jp/use/literacy/SSDSE/）

実践演習①：決定木分析によるビジネス意思決定サポート

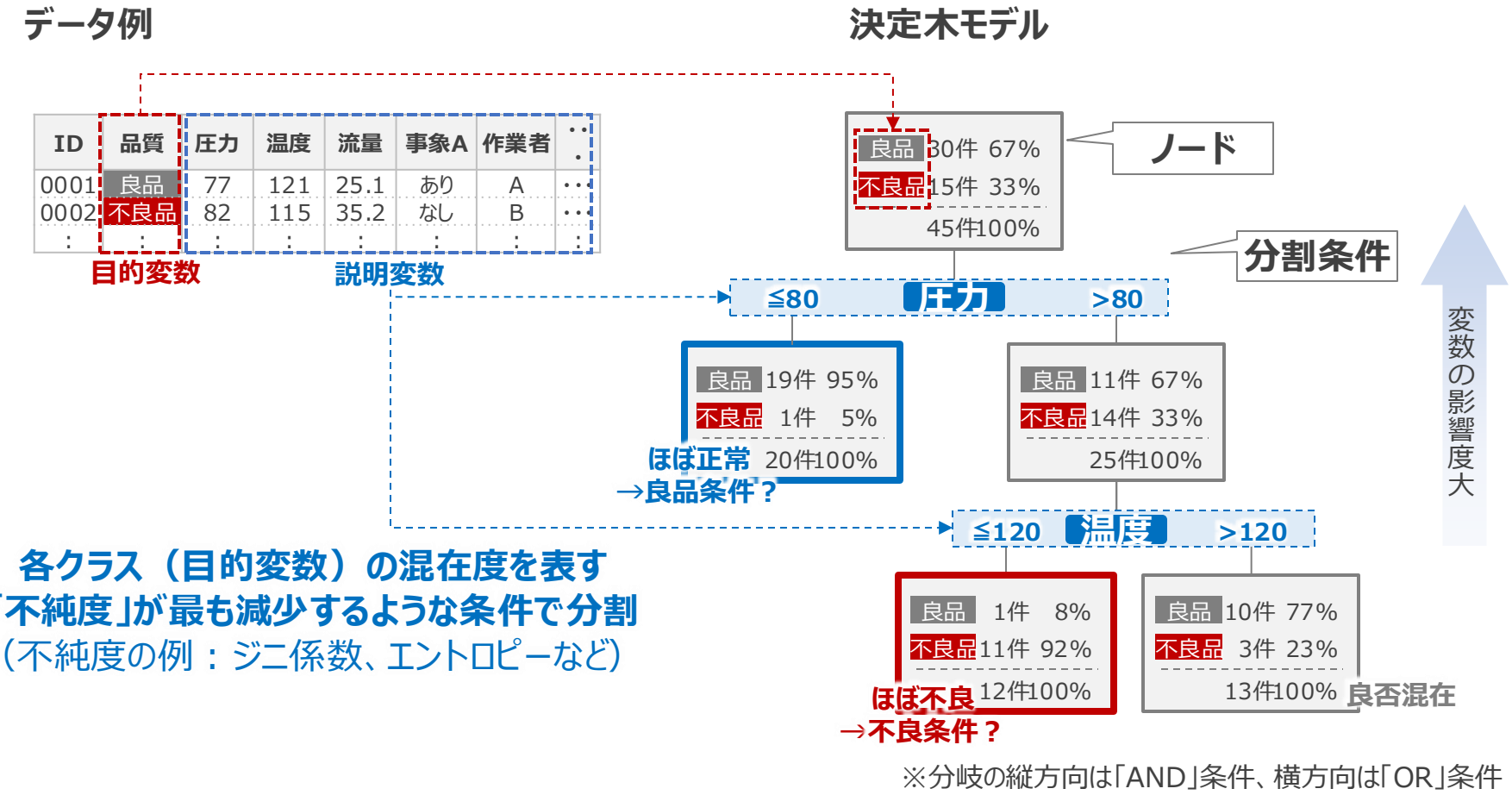
分析テーマ	**SSDSEのデータを用いて、「出生率」が低い都道府県と高い都道府県の違いを見出し、施策検討に繋げる**
使用データ	**SSDSE-基本素材（SSDSE-E）**

名称	内容
SSDSE-市区町村（SSDSE-A）	**1741市区町村**×多分野125項目 全国の全市区町村の、人口、経済、教育、労働、医療、福祉など、様々な分野の統計データを収録
SSDSE-県別推移（SSDSE-B）	**47都道府県**×12年次×多分野109項目 人口、経済、教育、労働、医療、福祉など、様々な分野の統計データを、12年分の時系列で収録
SSDSE-家計消費（SSDSE-C）	**全国・47都道府県庁所在市**×家計消費226項目 １世帯当たりの食料の年間支出金額（消費額）を、魚介、肉、野菜、果物、菓子、飲料などに分類し、それぞれ詳細な品目別にデータを収録
SSDSE-社会生活（SSDSE-D）	**全国・47都道府県**×男女別×社会生活121項目 男女別に、スポーツ、趣味・娯楽、ボランティアなどの詳細な活動データや、１日の睡眠、食事、学業、家事、仕事、趣味・娯楽などの時間配分データを収録
SSDSE-基本素材（SSDSE-E）	**全国・47都道府県**×多分野90項目 人口、経済、教育、文化、医療、福祉など、様々な分野の統計データを収録し、初学者にも扱いやすいデータセット
SSDSE-気候値（SSDSE-F）	**47都道府県庁所在市**×月・年×気象42項目 気温、気圧、風速、日照、降水、降雪など、様々な気象データについて、月・年別の平年値を収録

※SSDSE-Eデータセット（基本素材）の解説PDF：
https://www.nstac.go.jp/sys/files/kaisetsu-E-2024.pdf

決定木分析によるクラス分類

- 決定木分析は、与えられた分類情報に合致するように分割条件を導いていく手法
- 分析結果の解釈性が非常に高く技術的な解釈が容易なため、製造業の分析でも頻用される



Google Colaboratory上での
レクチャー＆演習

モデルの評価指標：Precision と Recall

- モデル評価は、**分析対象や目的に応じて**最適な指標を選択したり、複数の指標を用いて行う
- 例えばマーケティングではDM配布先を限定してコスト抑制するためにPrecisionを、医療では重大所見を見逃すと命に関わるためRecallを重視するなど、目的によって重視する指標は異なる

混同行列

※予測対象が「**不良品**」の場合

実データ＼予測結果	不良品	良品	
不良品	1 正答 (True Positive)	3 誤り：見逃し (False Negative)	再現率
良品	2 誤り：誤報 (False Positive)	4 正答 (True Negative)	特異度
	適合率 (陽性的中率)	陰性的中率	

予測 / 実際

- 災害アラート → 誤報
- アラート無し → 見逃し

評価指標

正解率 (Accuracy) $= \frac{1+4}{1+2+3+4}$

適合率, 精度, 陽性的中率 (**Precision**, PPV*) $= \frac{1}{1+2}$ 誤報の少なさ

*Positive Predictive Value

↔ **陰性的中率** (NPV*) $= \frac{4}{3+4}$ 負例の誤報の少なさ

*Negative Predictive Value

再現率, 感度, 真陽性率 (**Recall**, Sensitivity) $= \frac{1}{1+3}$ 見逃し (取りこぼし) の少なさ

↔ **特異度**, 真陰性率 (Specificity) $= \frac{4}{2+4}$ 負例の見逃しの少なさ

F値, F1スコア (F-measure) $= \frac{2 \cdot Precision \cdot Recall}{Precision + Recall}$ 誤報と見逃しの少なさのバランス

Precision と Recall の調和平均

Google Colaboratory上での
レクチャー＆演習

実践演習②：階層的クラスタリングによるデータ層別と戦略ポイントの検討

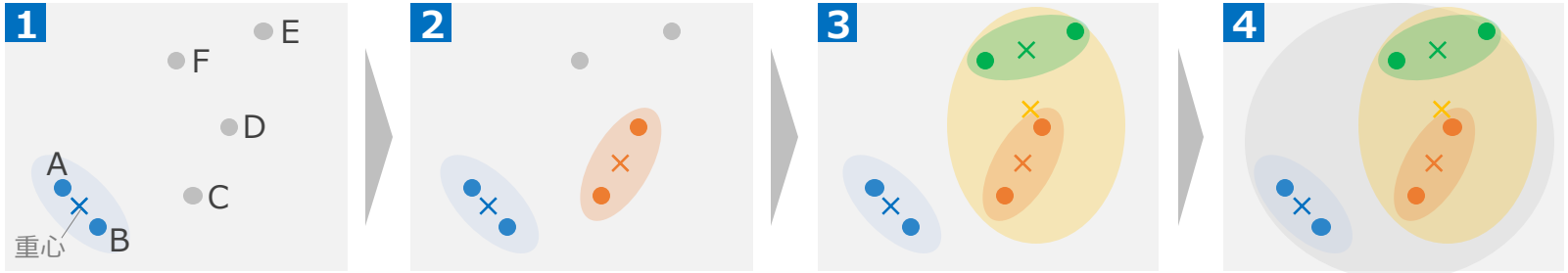
分析テーマ	**SSDSEのデータを用いて、 消費傾向が似ている都道府県をグルーピングする**
使用データ	**SSDSE-家計消費（SSDSE-C）**

名称	内容
SSDSE-市区町村（SSDSE-A）	**1741市区町村**×多分野125項目 全国の全市区町村の、人口、経済、教育、労働、医療、福祉など、様々な分野の統計データを収録
SSDSE-県別推移（SSDSE-B）	**47都道府県**×12年次×多分野109項目 人口、経済、教育、労働、医療、福祉など、様々な分野の統計データを、12年分の時系列で収録
SSDSE-家計消費（SSDSE-C）	**全国・47都道府県庁所在市**×家計消費226項目 １世帯当たりの食料の年間支出金額（消費額）を、魚介、肉、野菜、果物、菓子、飲料などに分類し、それぞれ詳細な品目別にデータを収録
SSDSE-社会生活（SSDSE-D）	**全国・47都道府県**×男女別×社会生活121項目 男女別に、スポーツ、趣味・娯楽、ボランティアなどの詳細な活動データや、１日の睡眠、食事、学業、家事、仕事、趣味・娯楽などの時間配分データを収録
SSDSE-基本素材（SSDSE-E）	**全国・47都道府県**×多分野90項目 人口、経済、教育、文化、医療、福祉など、様々な分野の統計データを収録し、初学者にも扱いやすいデータセット
SSDSE-気候値（SSDSE-F）	**47都道府県庁所在市**×月・年×気象42項目 気温、気圧、風速、日照、降水、降雪など、様々な気象データについて、月・年別の平年値を収録

※SSDSE-Cデータセット（家計消費）の解説PDF：
https://www.nstac.go.jp/sys/files/kaisetsu-C-2024.pdf

階層的クラスタリング（凝集型階層クラスタリング）

- 凝集型階層クラスタリングは、距離に応じて小さいクラスタを束ねて階層的に分類する手法
- クラスタ数は自動的に決定してくれる他、分類過程を可視化した**樹形図（デンドログラム）**も同時に出力されるので、結果の解釈やクラスタ数の決定に役立つ



最も距離の近い点を同一クラスタとしてまとめ、重心を計算

同様に、重心と残りの点の中で最も距離の近い点同士をまとめ、重心を計算

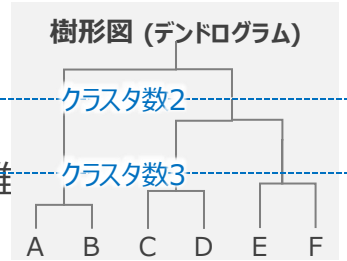
②を繰り返し、クラスタの重心同士が最近傍となった場合は、クラスタ同士をまとめる

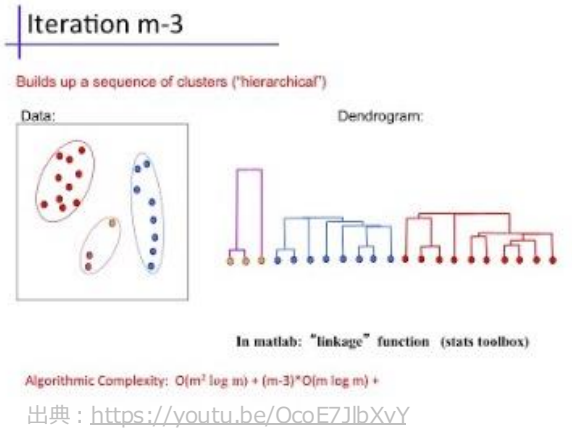
最後にクラスタ全体をまとめる

凝集型階層的クラスタリング
(agglomerative hierarchical clustering)

メリット
- クラスタ数は自動決定
- 樹形図により**分類過程が可視化**されることで、妥当なクラスター数を人が判断可能

デメリット
- 計算量が膨大
- データ量が多い場合、樹形図が複雑となり、解釈が困難になる





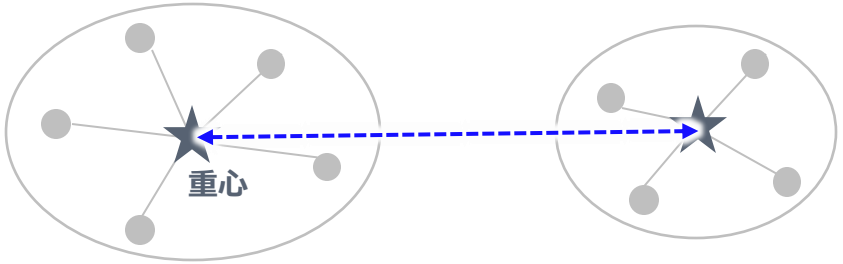
出典：https://youtu.be/OcoE7JlbXvY

（参考）クラスタ間「近さ」の評価尺度バリエーション

- クラスタ間の「近さ」を測る指標には様々あるが、一概にどれが良いとは言えないため、**複数試して比較**するのが一般的である。ただし、一般には、群平均法やWard法（次頁）が頻用される
- 最短距離／最長距離法は、計算量が少なくて済む反面、1点の影響を大きく受けやすい

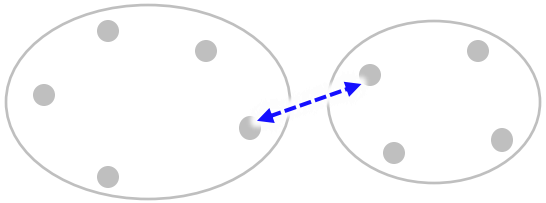
重心法

重心間の距離が近いクラスタを結合



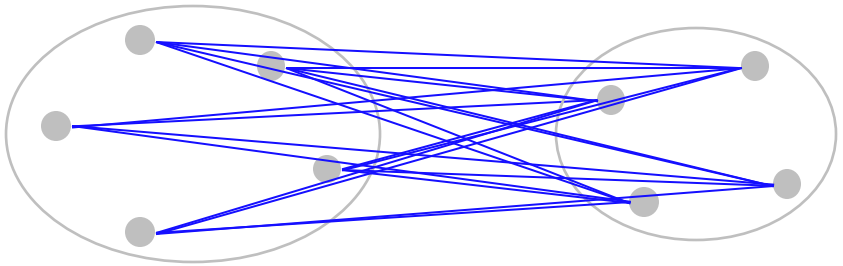
最短距離法

2つのクラスタ間で**最近傍**のデータをクラスタ間距離として採用



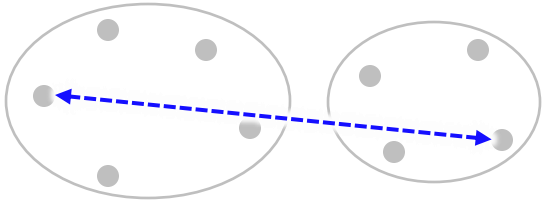
群平均法

クラスタ間で全データ間の距離を算出し、その**平均値**が近いクラスタを結合



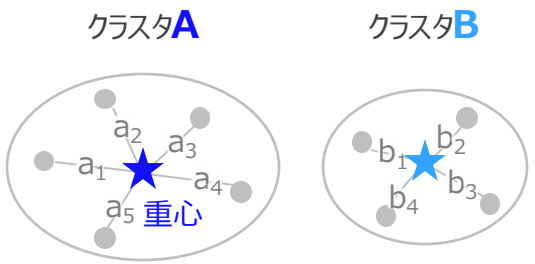
最長距離法

2つのクラスタ間で**最遠方**のデータをクラスタ間距離として採用

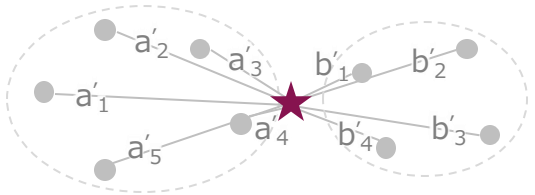


（参考）Ward法の考え方

■ Ward法*は最もよく用いられる手法であり、計算量は多いが、各データ点とクラスタ重心との関係性まで評価しているため、他手法に比べ、**分類感度が高い**とされる
*米国の統計学者Joe H. Ward, Jr.が1963年に発表した論文にちなむ



A, Bの結合を仮定した場合のクラスタ**AB**



1 「クラスタ重心」と、「当該クラスタ内の各データ」との距離の総和（二乗和）をクラスタごとに算出

クラスタ**A**の場合

$$\mathbf{A} = a_1^2 + a_2^2 + a_3^2 + a_4^2 + a_5^2$$

クラスタ**B**の場合

$$\mathbf{B} = b_1^2 + b_2^2 + b_3^2 + b_4^2$$

2 注目する2つのクラスタを結合した場合を仮定し、「結合後のクラスタ重心」と「当該クラスタ内の各データ」との距離の総和（二乗和）を算出

$$\mathbf{AB} = a'_1{}^2 + a'_2{}^2 + a'_3{}^2 + a'_4{}^2 + a'_5{}^2 + b'_1{}^2 + b'_2{}^2 + b'_3{}^2 + b'_4{}^2$$

3 **1** と **2** の差、つまり、**AB** − (**A**+**B**) が**最小**となるクラスター結合を採用
（結合前後でクラスタ内のばらつきに変化なし→統合してもOKと判定）

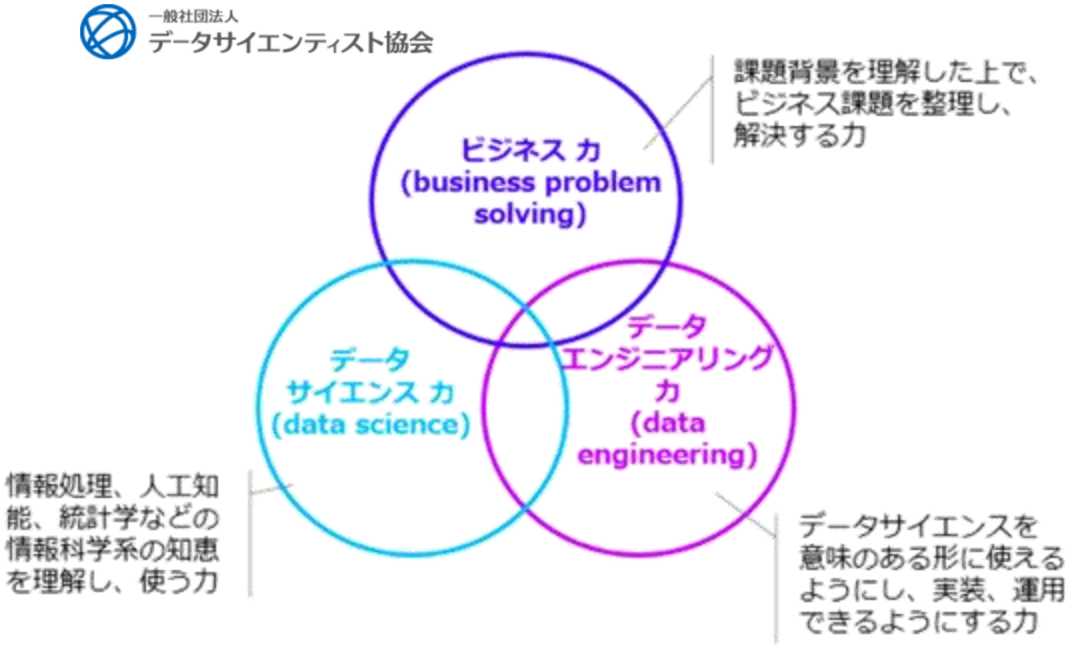
※近くにあり、ばらつきの小さいクラスタ同士が結合しやすい

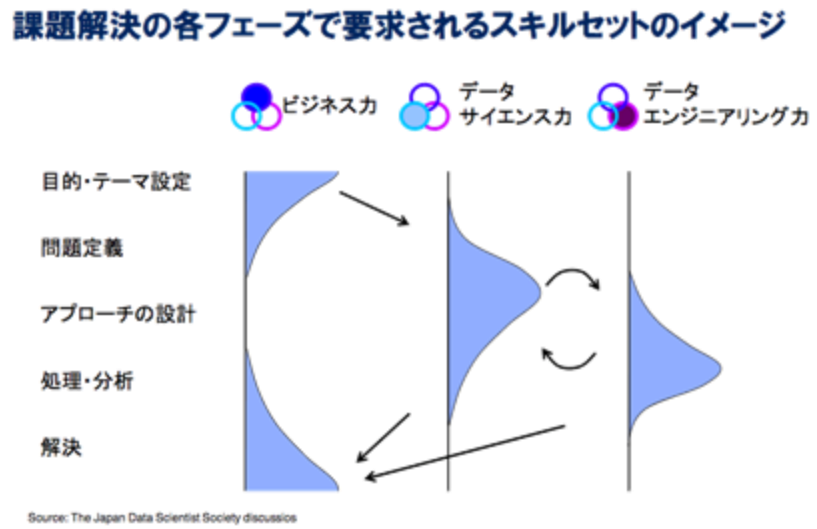
Google Colaboratory上での
レクチャー＆演習

今後の継続学習、
実践活用のためのポイント

再掲 データサイエンティストに求められる知識・スキルセット

- データサイエンティスト協会の定義：「データサイエンティストとは、**データサイエンスカ、データエンジニアリングカ**をベースにデータから価値を創出し、**ビジネス**課題に答えを出すプロフェッショナル」
- これら3スキルはどれも不可欠で、分析フェーズによって中心となるスキルが変化する、としている





出展：データサイエンティスト協会資料
https://www.datascientist.or.jp/news/n-news/post-255/

それぞれのスキルの伸ばし方 – ビジネス力

■自業務における課題について、問題点を整理してみる

■仮説思考力の高め方のコツ

✓起こった事象や自分の思考結果に対して、

- 「**だから何なのか？**」（**so what**）と
- 「**なぜそうなるのか？**」（**how**）を繰り返す
  参考：なぜなぜ分析（トヨタ生産方式から生まれたフレームワーク）

✓日常生活の中で将来予測をする癖をつける

- 新聞記事・ニュース
- 職場での会話
- 日常会話

✓以下はバイブルとして知られる本です
（これ以外にも今は色々出ていると思うので、本屋などで自身に合うものを探してみてください）

- 仮説思考 BCG流 問題発見・解決の発想法 | 内田 和成 https://www.amazon.co.jp/dp/4492555552
- イシューからはじめよ――知的生産の「シンプルな本質」| 安宅和人 https://www.amazon.co.jp/dp/4862760856
- ロジカル・シンキング Best solution | 照屋 華子, 岡田恵子 https://www.amazon.co.jp/dp/B00978ZQOG
- 入門 考える技術・書く技術 | 山﨑 康司 https://www.amazon.co.jp/dp/B0081WMQ4W

それぞれのスキルの伸ばし方 – データサイエンス力／データエンジニアリング力

■実際に自身の**課題感を持って実践的な分析を行う**ことが最も効果的です

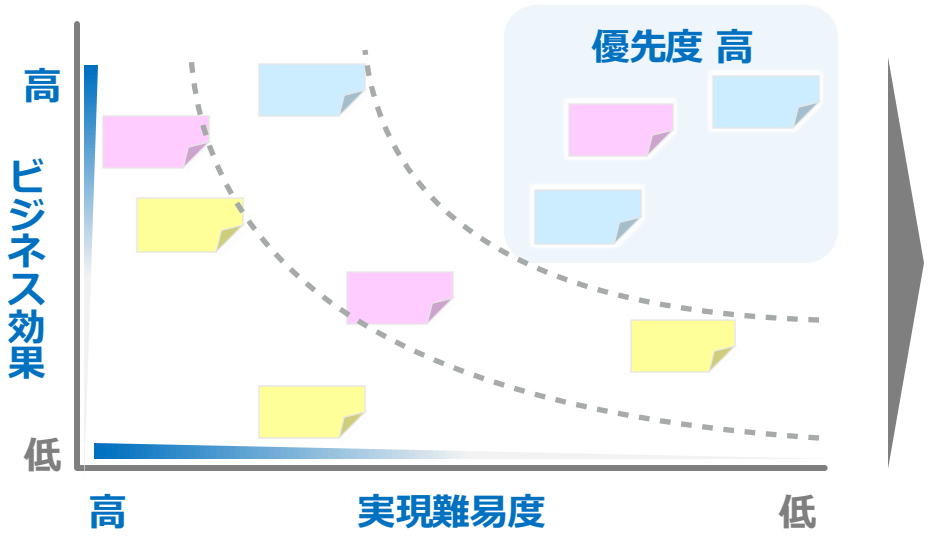
- ✓**自業務におけるデータ**を活用する
- ✓**自業務における切迫した課題**を分析テーマとする
- ✓**オープンデータ**を用いて、自分なりに気になる社会課題などについて分析する
  - e-Stat（政府統計データサイト）：https://www.e-stat.go.jp/
- ✓**コンペティション**に参加する
  - Kaggle（米国）：https://www.kaggle.com/
  - SIGNATE（日本）：https://signate.jp

■モチベーション維持や相互学習のために、**勉強会**などで誰かと一緒に取り組むことも極めて有効です

分析課題の選定・具体化

■ 社内課題を洗い出し、「ビジネス効果」と「実現難易度」の観点で整理して
優先度の高いテーマを選定・具体化する

分析テーマ候補の整理



(実現期間、社員スキル、コスト、データ取得可否・・・etc)

分析テーマの具体化 (例)

分析テーマ	
テーマ概要	
使用データ	
目的変数	
説明変数	
ビジネス効果	
分析ゴール	

Sample

分野を超えた手法の適用

- 同じデータを扱うという意味では分野の壁はなく、様々な分野の最新事例、動向にアンテナを張りながら、積極的に自分の分野に取り入れていくことが重要
- 異分野との交流会は極めて貴重な機会であり、積極的に参加していくべきである

マーケティング領域で用いられる「アソシエーション分析」（教師なし学習手法）

個客の購買データ

購買1件あたり1レコード

データ構造化

購入ID	牛肉	ワイン	リンゴ	みかん	…
0001	✔	✔	✔		…
0002			✔	✔	…
0003		✔			…
0004	✔	✔	✔	✔	…
0005	✔	✔	✔		…
:	:	:	:	:	…

牛肉とワインが**同時購入**される割合が高い

ルール抽出

大量データの中から共起する頻出パターン（=**ルール**）を抽出

条件　⇒　帰結

牛肉　⇒　ワイン

活用例

商品陳列の工夫（同時購入されやすい製品を隣接させる）

レコメンデーション（「条件」側の製品がカゴに入ったら、「帰結」側の製品も推薦）

製造・医療への応用

製造

製品ID	割れ・欠け	シワ	キズ	異物混入	…
0001	✔	✔	✔		…
0002		✔	✔	✔	…
0003		✔	✔		…
0004	✔		✔	✔	…
0005		✔	✔		…
:	:	:	:	:	…

シワとキズが**同時発生**する割合が高い

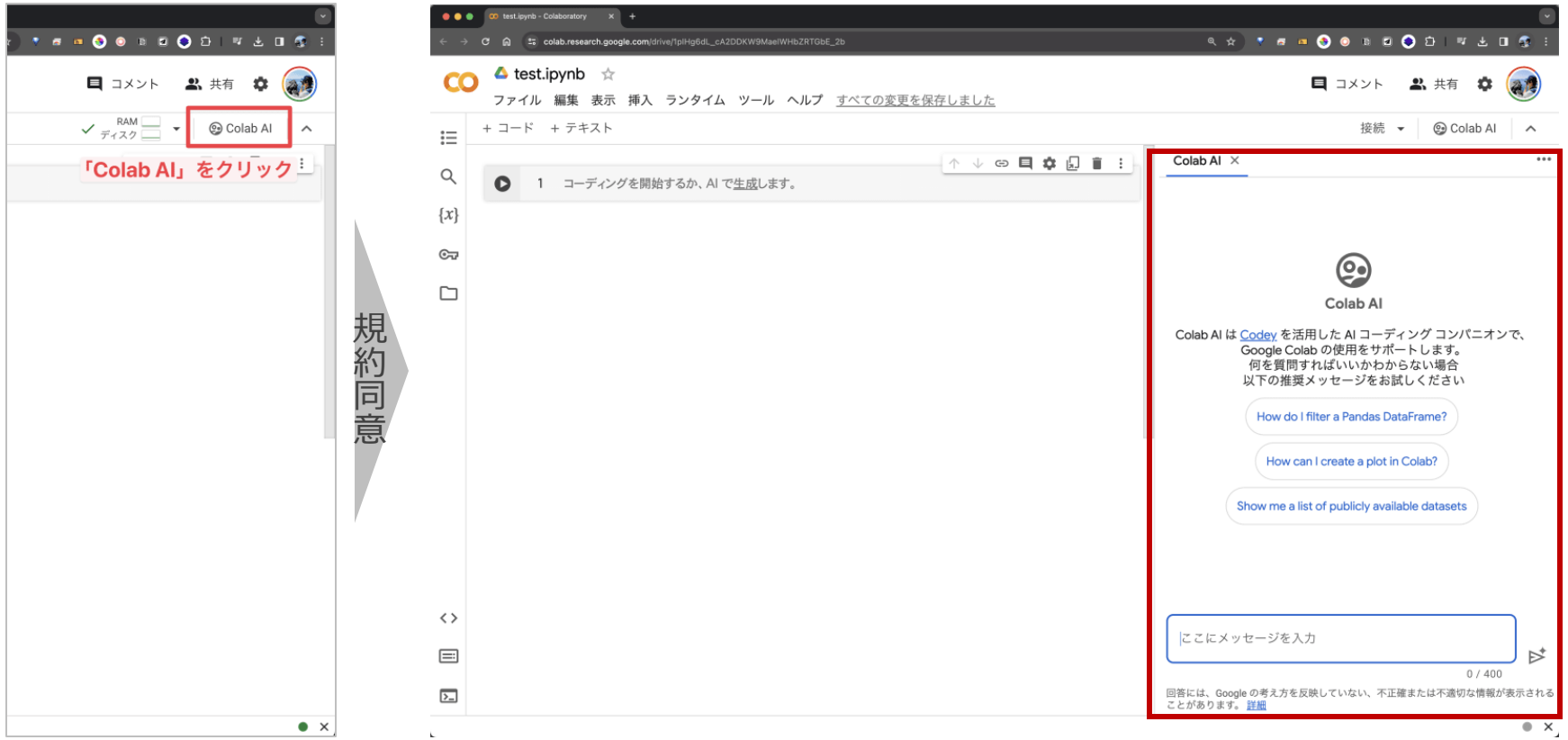
医療

症例ID	虚血性心疾患	糖尿病	高血圧	脂質異常症	…
0001	✔	✔	✔		…
0002			✔	✔	…
0003		✔			…
0004	✔	✔	✔	✔	…
0005	✔	✔	✔		…
:	:	:	:	:	…

虚血性心疾患と糖尿病が**同時発症**する割合が高い

（参考）Colab AI（生成AIによる支援機能）の活用

- Google Colaboratory では、生成AIによるコード作成支援機能として、”Colab AI” が無料で活用できる
- 簡単なコード作成からデバッグなど、様々な活用でき、初心者でも効率的にコード作成ができる
  ただし、スキルアップ／定着の観点では、初めのうちこそ、できる限り、自力で作成することが望ましい



画像出典：https://zenn.dev/umi_mori/books/ai-native-programming/viewer/google_colab_ai

▼参考：プログラミングでの生成AI活用
（前回講演資料抜粋）

- 活用例①：簡単なコード作成
- 活用例②：コードのデバッグ（エラー修正）
- 活用例③：コードレビュー
- 活用例④：コードの説明
- 活用例⑤：コードを別の言語に変換
- 活用例⑥：変数名や関数名を作成
- 活用例⑦：ダミーデータを作成