

Part 3

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/>

統計で見る日本
e-Statは、日本の統計が閲覧できる政府統計ポータルサイトです

お問い合わせ | ヘルプ | English

ログイン 新規登録

統計データを探す 統計データの活用 統計データの高度利用 統計関連情報 リンク集

●統計データを探す (政府統計の調査結果を探します)

その他の絞り込み

利用ガイド

●統計データの高度利用

マイクロデータの利用
公的統計のマイクロデータの利用案内

開発者向け
API、LODで統計データを取得

●統計関連情報

統計分類・調査計画等

●統計データを活用する

グラフ
主要指標をグラフで表示
(統計ダッシュボード)

時系列表
主要指標を時系列表で表示
(統計ダッシュボード)

地図
地図上に統計データを表示

地域
都道府県、市区町村の
主要データを表示

教育用標準データセット（SSDSE）

- 教育用標準データセット（SSDSE; Standarzed 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
<u>SSDSE-市区町村（SSDSE-A）</u>	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
<u>SSDSE-県別推移（SSDSE-B）</u>	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
<u>SSDSE-家計消費（SSDSE-C）</u>	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
<u>SSDSE-社会生活（SSDSE-D）</u>	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
<u>SSDSE-基本素材（SSDSE-E）</u>	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
<u>SSDSE-気候値（SSDSE-F）</u>	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

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

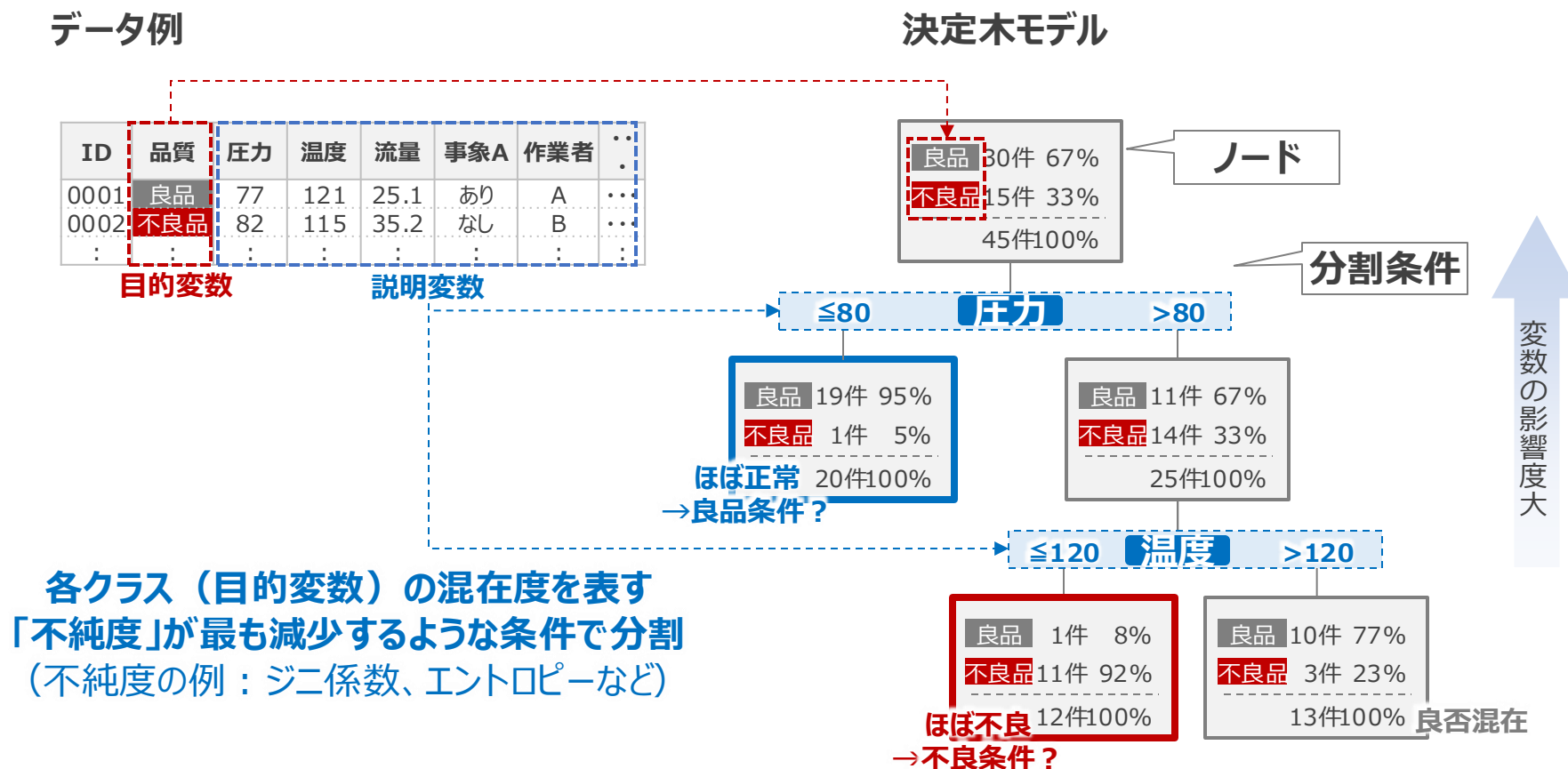
SSDSE-基本素材（SSDSE-E）

名称	内容
<u>SSDSE-市区町村（SSDSE-A）</u>	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
<u>SSDSE-県別推移（SSDSE-B）</u>	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
<u>SSDSE-家計消費（SSDSE-C）</u>	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
<u>SSDSE-社会生活（SSDSE-D）</u>	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
<u>SSDSE-基本素材（SSDSE-E）</u>	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
<u>SSDSE-気候値（SSDSE-F）</u>	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



※分岐の縦方向は「AND」条件、横方向は「OR」条件



Google Colaboratory上での レクチャー&演習

Part 3 : Google Colab.リンク

https://colab.research.google.com/drive/1roO0rhXNjHOPccpK21p3_UF-oF4I9cKk?usp=sharing

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

		予測結果		
		不良品	良品	
実データ	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)	

適合率
(陽性的中率)
予測

陰性的中率

実際



誤報

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*)
*Positive Predictive Value

$$= \frac{1}{1 + 2 + 3 + 4}$$

誤報の少なさ

↔ **陰性的中率**
(NPV*)
*Negative Predictive Value

$$= \frac{4}{1 + 2 + 3 + 4}$$

負例の誤報
の少なさ

再現率, 感度, 真陽性率
(Recall, Sensitivity)

$$= \frac{1}{1 + 2 + 3 + 4}$$

見逃し (取りこぼし)
の少なさ

↔ **特異度**, 真陰性率
(Specificity)

$$= \frac{4}{1 + 2 + 3 + 4}$$

負例の見逃し
の少なさ

F値, F1スコア
(F-measure)

$$= \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

Precision と Recall の調和平均

誤報と見逃しの
少なさのバランス



Google Colaboratory上での レクチャー&演習

Part 3 : Google Colab.リンク

https://colab.research.google.com/drive/1roO0rhXNjHOPccpK21p3_UF-oF4I9cKk?usp=sharing

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

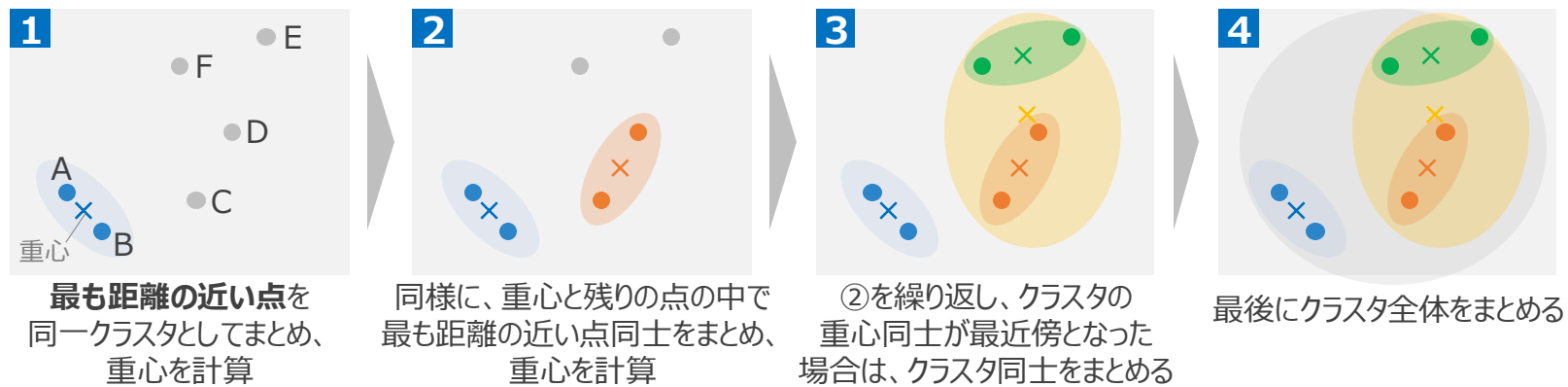
分析テーマ	SSDSEのデータを用いて、 消費傾向が似ている都道府県をグルーピングする
使用データ	SSDSE-家計消費（SSDSE-C）

名称	内容
<u>SSDSE-市区町村（SSDSE-A）</u>	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
<u>SSDSE-県別推移（SSDSE-B）</u>	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
<u>SSDSE-家計消費（SSDSE-C）</u>	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
<u>SSDSE-社会生活（SSDSE-D）</u>	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
<u>SSDSE-基本素材（SSDSE-E）</u>	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
<u>SSDSE-気候値（SSDSE-F）</u>	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）

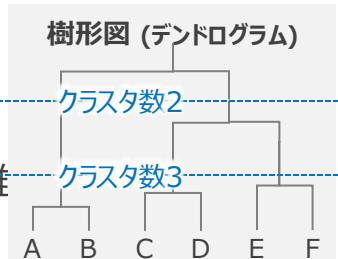
- 凝集型階層クラスタリングは、距離に応じて小さいクラスタを束ねて階層的に分類する手法
- クラスタ数は自動的に決定してくれる他、分類過程を可視化した**樹形図**（デンドログラム）も同時に出力されるので、結果の解釈やクラスタ数の決定に役立つ



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

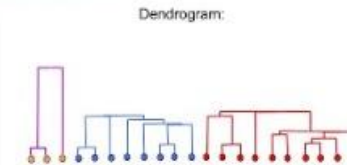
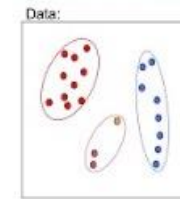
メリット
デメリット

- 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
- 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



Iteration m-3

Builds up a sequence of clusters ("hierarchical")



In matlab: "linkage" function (stats toolbox)

Algorithmic Complexity: $O(m^3 \log m) + (m-3) \cdot O(m \log m) +$

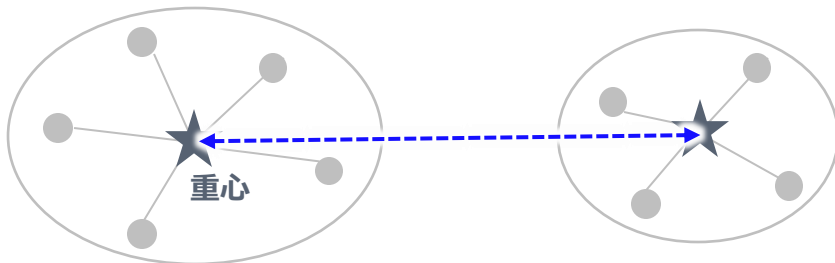
出典: <https://youtu.be/OcoE71bXvY>

(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

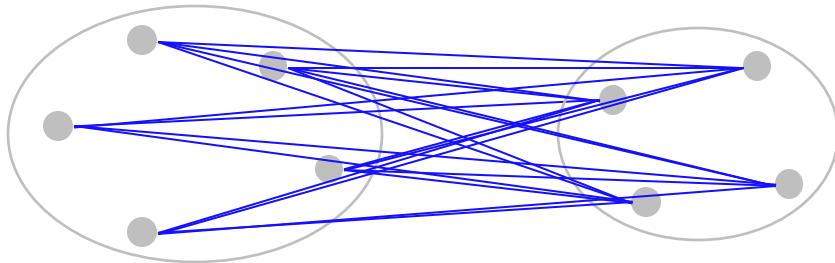
重心法

重心間の距離が近いクラスタを結合



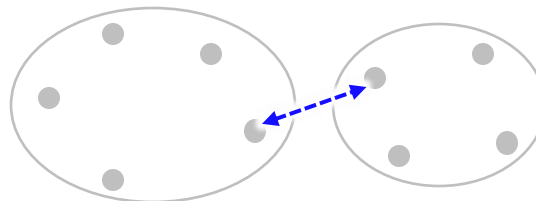
群平均法

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



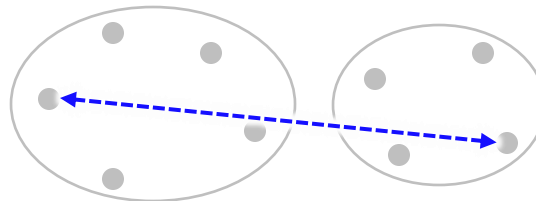
最短距離法

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



最長距離法

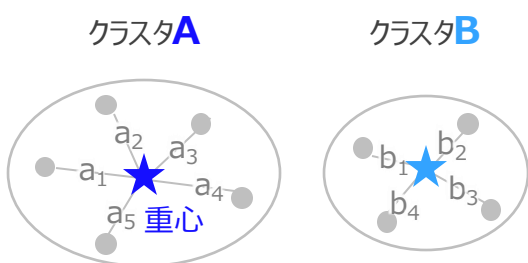
2つのクラスタ間で**最遠方**のデータをクラスタ間距離として採用



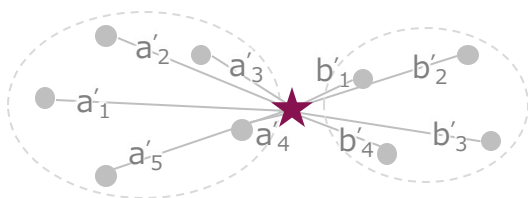
(参考) Ward法の考え方

- Ward法*は最もよく用いられる手法であり、計算量が多いが、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

$$A = a_1^2 + a_2^2 + a_3^2 + a_4^2 + a_5^2$$

クラスタBの場合

$$B = b_1^2 + b_2^2 + b_3^2 + b_4^2$$

- 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

$$AB = a_1'^2 + a_2'^2 + a_3'^2 + a_4'^2 + a_5'^2 + b_1'^2 + b_2'^2 + b_3'^2 + b_4'^2$$

- 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

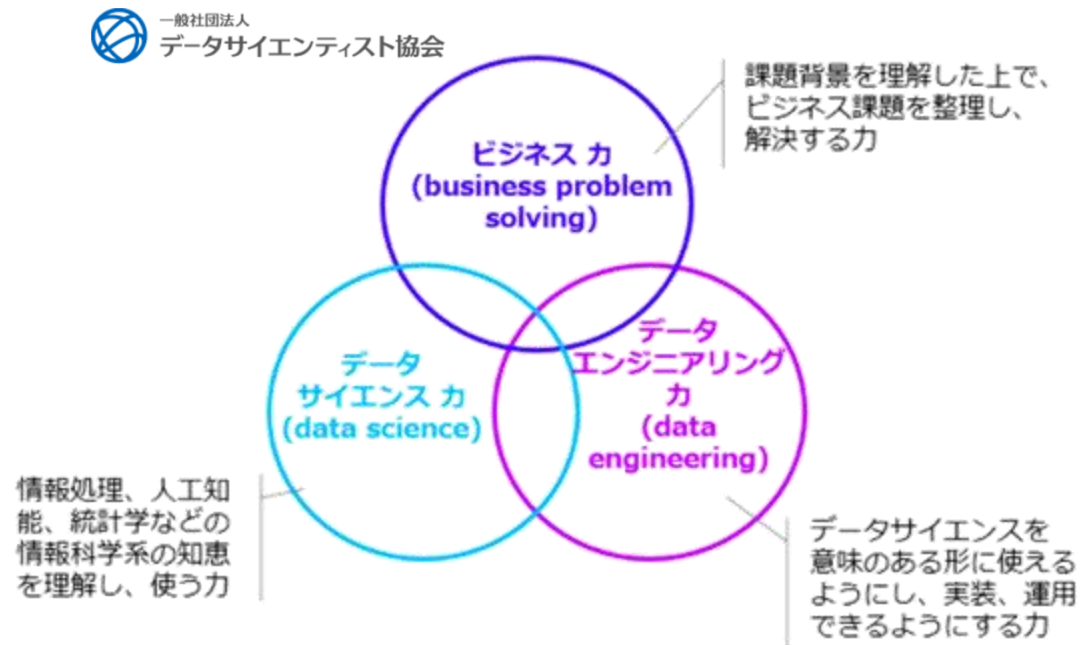
Part 3 : Google Colab.リンク

https://colab.research.google.com/drive/1roO0rhXNjHOPccpK21p3_UF-oF4I9cKk?usp=sharing

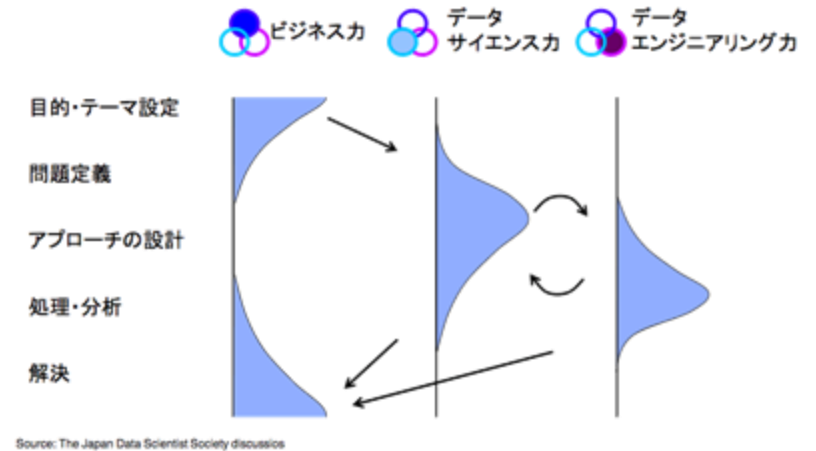
今後の継続学習、
実践活用のためのポイント

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、**データサイエンス力、データエンジニアリング力**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



課題解決の各フェーズで要求されるスキルセットのイメージ



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 | 山崎 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

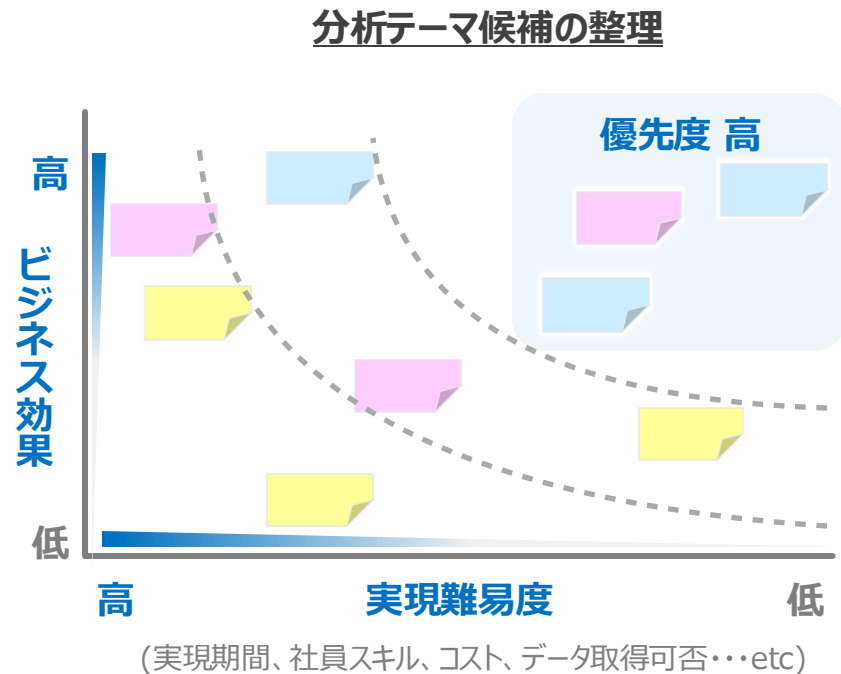
■実際に自身の課題感を持って実践的な分析を行うことが最も効果的です

- ✓ 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		✓	✓		...
:	:	:	:	:	...

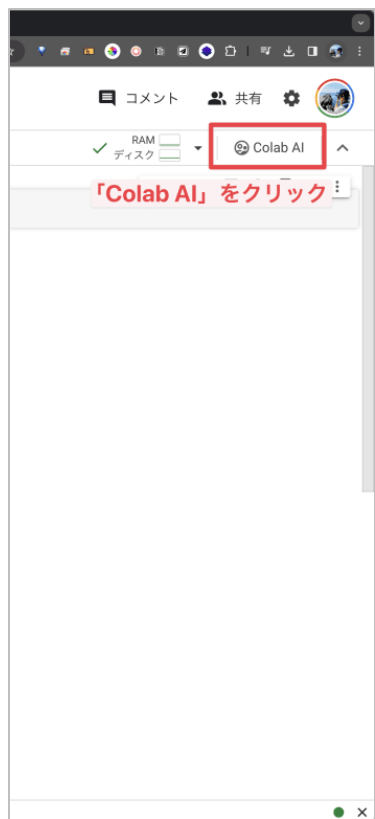
シワとキズが**同時発生**
する割合が高い

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	✓	✓	✓		...
:	:	:	:	:	...

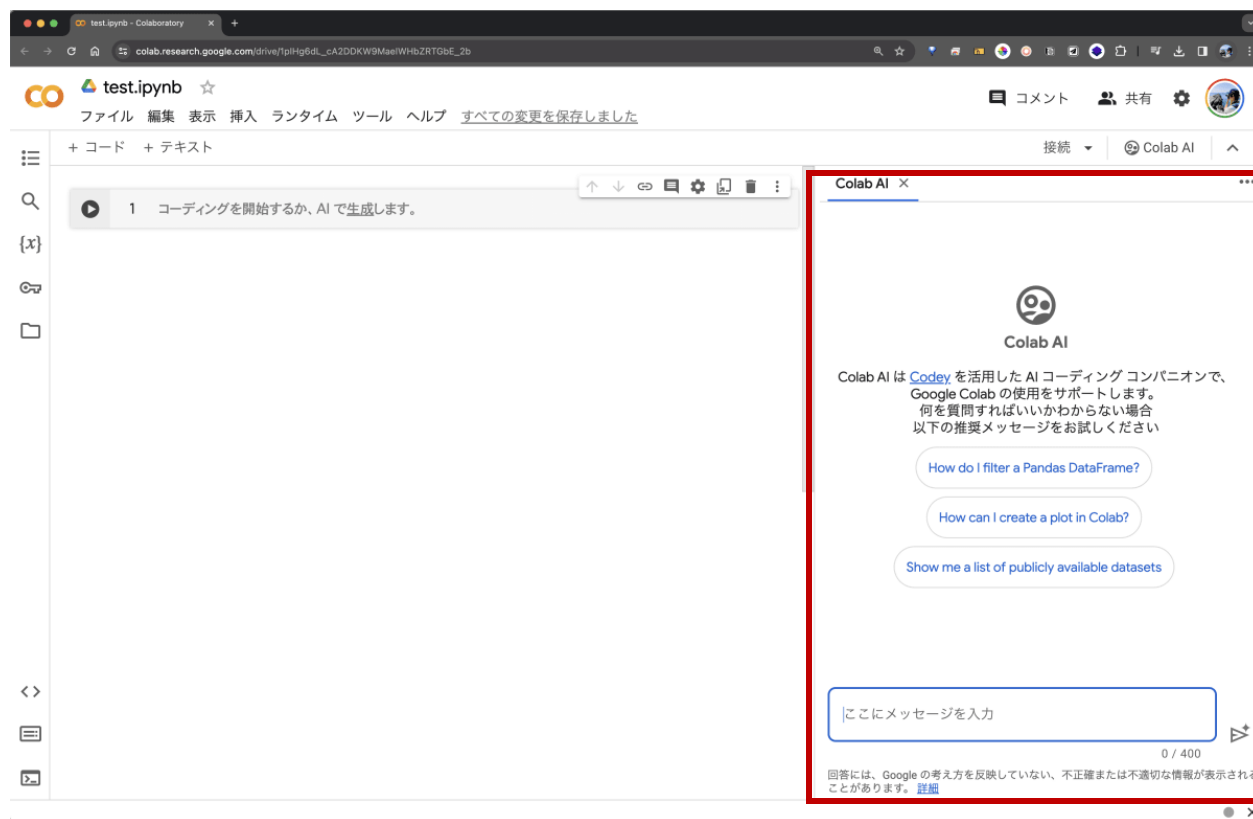
虚血性心疾患と糖尿病が
同時発症する割合が高い

(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



規約同意



▼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