

社会の進歩と データサイエンス・AI

熊本保健科学大学「データサイエンス入門」

熊本保健科学大学 保健科学部 共通教育センター

水本 豪

わたしたちを取り巻く社会の変化

これからの時代は「**データ駆動型社会**」 (Society 5.0)

様々な大規模データ（ビッグデータ）を活用し、課題解決や意思決定に役立てる社会

Society 1.0	Society 2.0	Society 3.0	Society 4.0	Society 5.0
狩猟社会	農耕社会	工業社会	情報社会	データ駆動型社会 超スマート社会

ビッグデータとは？

ビッグデータの特徴（3V）

- ・ Volume（量）： 扱うことのできるデータ量が大きい
- ・ Velocity（速度）： データを計測、記録する速度が速い
- ・ Variety（種類）： 扱うことのできるデータの種類が多い

データ駆動型社会を可能にしたもの

コンピュータ性能やインフラ技術の進歩

	1996年 (デスクトップ)	2022年 (ノートパソコン)
コンピュータの処理能力の向上 CPU (コンピュータの頭脳)	120 MHz	2.66 GHz
保存可能な記憶容量の増大 メモリ (データの一時記憶) ストレージ (データの長期保存)	16 MB 1.6 GB (HDD)	32 GB 1 TB (SSD)
インターネット回線の高速化 最大回線速度	64 Kbps	220 Mbps

参考：単位について K(キロ) – 1024倍 → M(メガ) – 1024倍 → G(ギガ) – 1024倍 → T(テラ)

ちょっとだけ本題から逸れますが...

コンピュータで処理されるデータ (データ表現)

コンピュータで処理するデータは2進数 (0 or 1)

- 2進数の値1つで1 bit (ビット)
 - 1 bit の情報 : 2通り [0/1]
 - 2 bit の情報 : 4通り [00, 01, 10, 11]
 - 3 bit の情報 : 8通り [000, 001, 010, 011, 100, 101, 110, 111] $\Rightarrow$ n bit の情報 : 2^n 通り
- 8 bit の情報 : 2^8 通り = 1 byte (バイト)
 - 1 byte の情報 : 2^8 (256) 通り
 - 2 byte の情報 : 2^{16} (65,536) 通り
 - 3 byte の情報 : 2^{24} (16,777,216通り) $\Rightarrow$ n byte の情報 : $2^{8 \times n}$ 通り

いろいろなものを区別するために必要な情報

- 0~9までの値を区別 : 10通りを区別する必要あり $\Rightarrow$ 4 bit (2^4 通り)
- アルファベットの小文字を区別 : 26通りを区別する必要あり $\Rightarrow$ 5 bit (2^5 通り)
- 色を区別する : 12色 $\Rightarrow$ 4 bit、24色 $\Rightarrow$ 5 bit、プロ用色鉛筆520色 $\Rightarrow$ 10 bit 必要
- 世界中の文字を扱う世界標準規格 Unicode (UTF-8) : 3 byte (2^{24} 通り)

実際のファイルサイズは？

ファイルサイズを確認してみよう

ファイルの上で右クリックし、「プロパティ」を選択するとファイルサイズの確認ができます。

- ・メモ帳に自分の名前を記入しただけのファイル： 18 byte
- ・4年生の国家試験対策資料（Wordファイル、10ページ、画像なし）： 43,565 byte (42.5KB)
- ・大学院の講義で使ったPowerPointのファイル（スライド30枚）： 1,068,368 byte (1.01MB)
- ・20分ほどの授業動画（mp4形式）： 294,579,268 byte (280MB)

ファイルサイズの単位：

1 byte —1024倍→ 1 KB(キロバイト) —1024倍→ 1 MB(メガバイト) —1024倍→
1GB(ギガバイト) —1024倍→ 1 TB(テラバイト)

コンピュータ性能やインフラ技術の進歩

サイズが大きいデータでも処理可能なコンピュータ

サイズが大きいデータでも転送可能な高速インターネット

データ駆動型社会を可能にしたもの

計測に使用するセンサの進歩

動きや熱、光、磁気、音など多様な情報を正確に捉えることができるセンサの存在
→ いろいろなものをデータ化することが可能になった！

IoT（Internet of Things; モノのインターネット）技術の進歩

コンピュータやセンサなど様々なものをインターネットで接続することが可能
→ インターネットを介してそれぞれに集めたデータを一か所に集積できるようになった！

AIの進歩

- ・データからAI自身が知識を獲得する「機械学習」
 - ・機械学習の中でも特に膨大なデータからデータの特徴や規則性を抽出することに優れた「**深層学習（ディープラーニング）**」の登場
- 集積データを分析できるハードウェア（コンピュータ）とソフトウェア（AI）が整備！

データサイエンスとは？

データの収集・分析により有益な知見を得て、様々な問題の発見や解決に役立てる学問

- ・ コンピュータ関連（データベース、プログラミング、AI、情報処理、セキュリティなど）
- ・ 数学や統計学（データの収集、分析手法を扱う）
- ・ それぞれの専門領域（経済、経営、スポーツ、医療、農業、観光など）

データサイエンスでできること

- (1) 予測する： 過去のデータから法則を見つけ、それをもとに未来のことを予測する
 - (2) 分類する： 膨大なデータから特徴を見つけ、それをもとにグループに分ける
 - (3) 発見する： 膨大なデータに潜む傾向や関係、パターンを見つける
- ➡ これらを単独あるいは組み合わせて用いることで、各分野の問題解決に活かすことが可能

データサイエンスにおける「データ」

調査・実験・観測によって得られたデータから、コンピュータによって自動的に
記録・蓄積されたログデータまでさまざま

データの分類

- ・ 構造化データ： 表の形式で表されるデータ（構造化＝ルールに沿ってきちんと整えること）
→ 表のマス目（セル）に数字や文字がデータとして入っている
- ・ 非構造化データ： 表の形式で表されないデータ（例：画像、動画、音声など）

コンピュータ目線で考えると...

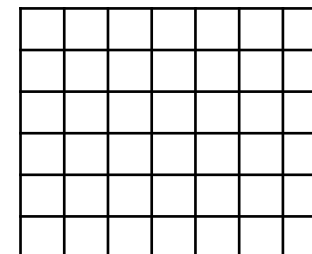
- ・ コンピュータは、0と1で情報を表現する
（＝明確に区別された情報は得意、あいまいな情報は苦手）
- ・ 構造化データは規則的で明確なので得意
- ・ 非構造化データはそのままでは処理できない
（実は、コンピュータで扱えるような処理を行っている＝デジタル処理）

非構造化データのデジタル処理

デジタル処理： 連続した情報を段階的な数値に区切って表す処理（あいまいさをなくす）

画像の場合

- ・ 本来、色は切れ目なく変わる連続的な存在
- ・ 色の名前をつけて分けていくのもある種のデジタル処理



《コンピュータが扱えるようにするためのデジタル処理》

- ・ 画像は小さな画素（ピクセル）が並んでできている

〔参考〕 4K：約829万画素（3840×2160）、iPhone16カメラ：最大4,800万画素(48MP)（6826×7008）

- ・ 色の違いを値で表現

白黒の場合：

黒(0) or 白(255)

2つの値で表現

グレースケールの場合：

黒(0) ----- 白 (255)

256段階の値で表現

カラーの場合：

赤(R)、緑(G)、青(B) の3色について256段階の値で表現

(例) 赤 (255, 0, 0) 、黄色 (255, 255, 0)

黒 (0, 0, 0) 、白 (255, 255, 255)

非構造化データのデジタル処理

動画の場合

《コンピュータが扱えるようにするためのデジタル処理》

- ・ 動画は画像が連続したもの（パラパラ漫画）
- ・ 1秒間に30枚くらいの画像が配置されている（30 fps(frames per second)）
被写体の動きが速い場合には1秒間に60枚くらいで撮影（60 fps）
スローモーション撮影の場合にはもっと細かく撮影（120 fps, 240 fps）

音声・音楽の場合

《コンピュータが扱えるようにするためのデジタル処理》

- ・ 1秒間に16,000回音声を取り込む（サンプリング処理；16kHz）
音楽の場合にはもっと多い（スタジオ録音だと48,000回、96,000回、192,000回など）
- ・ 取り込んだ音の振幅（音の強さ）を数値化する（量子化処理；bitで表現）
8 bit = 256段階、16 bit = 65,536段階、24 bit = 16,777,216段階
- ・ データの大きさは、サンプリング周波数×量子化ビット数で表される
- ・ サンプリングと量子化の数が大きくなるほど高精度だがファイルサイズは膨大になる

データサイエンスでできること

データサイエンスの手法を学ぶ前に、何ができるのかを知っておく

- ・ 予測する： 過去のデータから法則を見つけ、それをもとに未来のことを予測する
(例) 天気予報、台風の進路予測、経済予測 など
- ・ 分類する： 膨大なデータから特徴を見つけ、それをもとにグループに分ける
(例) 画像認識・音声認識、迷惑メールの判定 など
- ・ 発見する： 膨大なデータに潜む傾向や関係、パターンを見つける
(例) 故障や破損の発見、おすすめ商品の提示 など

これらを単独あるいは組み合わせることで、各分野の問題解決に活かす！

実はAIが行うことも基本的に同じ

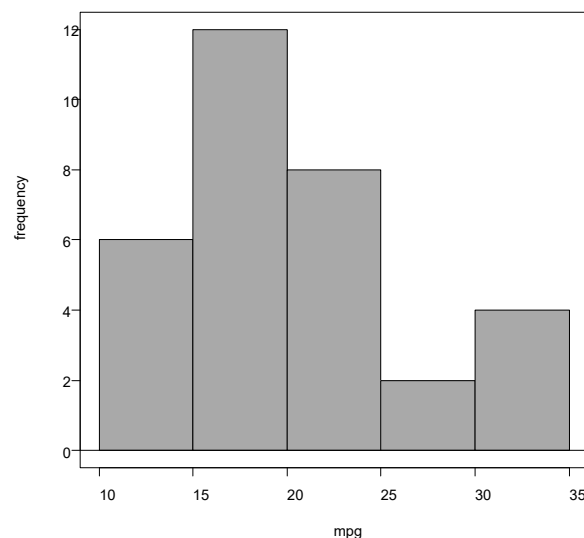
データサイエンスの技術・方法

データの可視化

データの特徴をわかりやすく視覚的に示すための方法（表、グラフ）

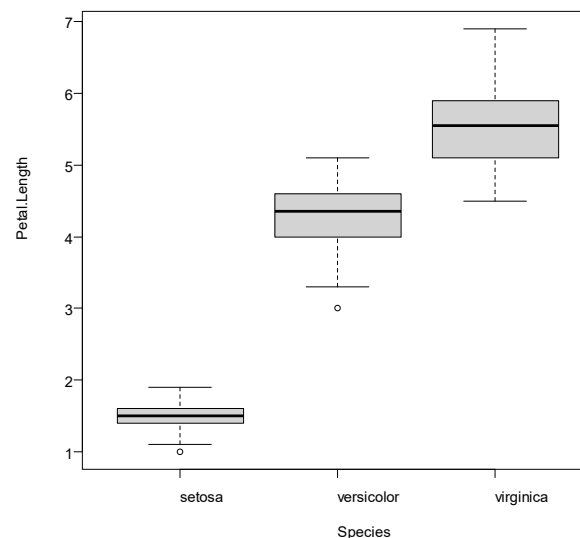
- ・ 量的データの分布の可視化

ヒストグラム



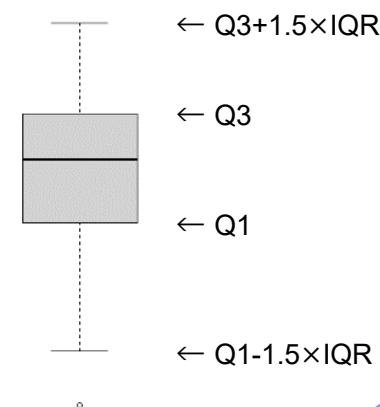
横軸に示す区間にどれだけのデータがあるか（度数）を表す。

箱ひげ図



種類別にデータの分布を表したもの。箱が全データの50%、箱の中の線は中央値。

参考：箱ひげ図で表される情報



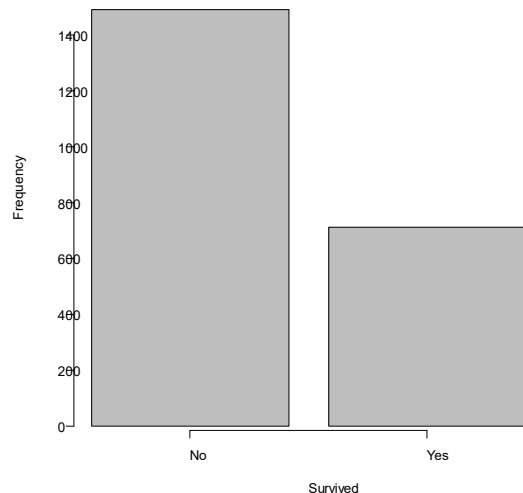
Q1（第1四分位数）：最小値—中央値の半分（下から1/4）の値
Q3（第3四分位数）：中央値—最大値の半分（上から1/4）の値
IQR（四分位範囲）： $Q3 - Q1$

ひげ部分の上端・下端から外側の値がある場合は、外れ値（○）として扱われる。※ひげ部分の上端・下端は最大値・最小値のこともある（その場合は外れ値なし）。

データサイエンスの技術・方法

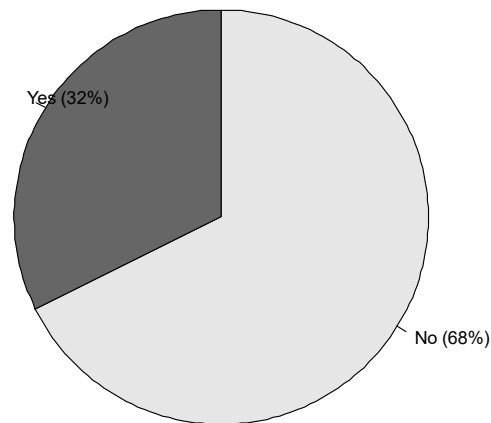
- ・ 質的データの分布の可視化

棒グラフ



横軸に示す区間にどれだけのデータがあるか（度数）を表す。

円グラフ

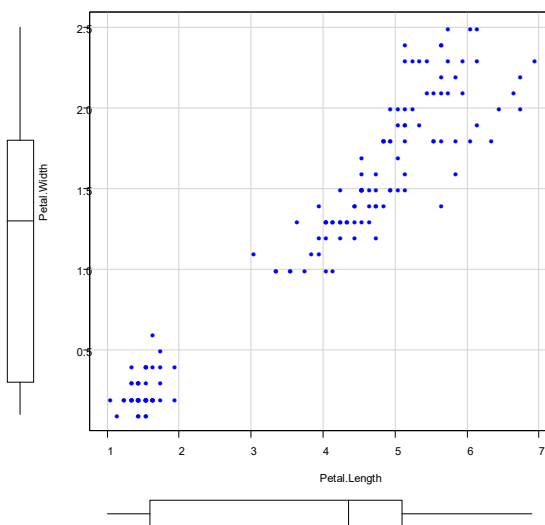


区分別のデータの割合を表す。
帯状に表したものは帯グラフ。

データサイエンスの技術・方法

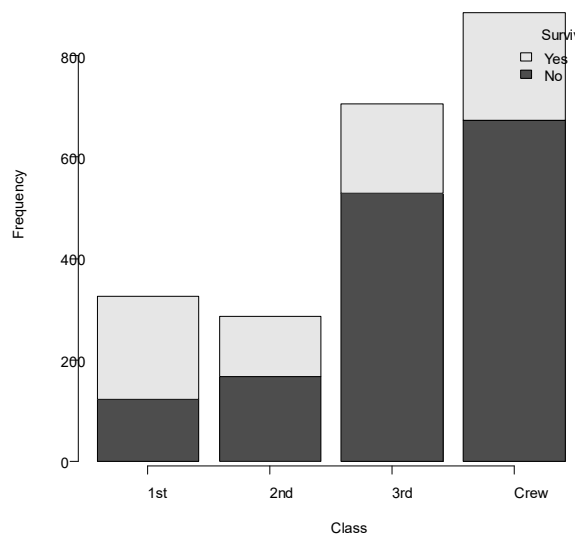
- 2変数（以上）の関係の可視化

散布図



縦軸と横軸に該当する箇所に点をプロットして表したグラフ。

積み上げ棒グラフ



2つの質的変数の頻度を表す。
表の形で表すとクロス集計表。

クロス集計表

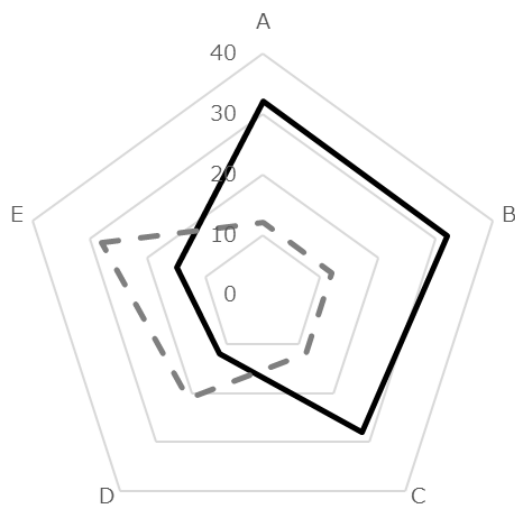
(2つの項目を掛け合わせて集計)

	1st	2nd	3rd	Crew
Yes				
No				

データサイエンスの技術・方法

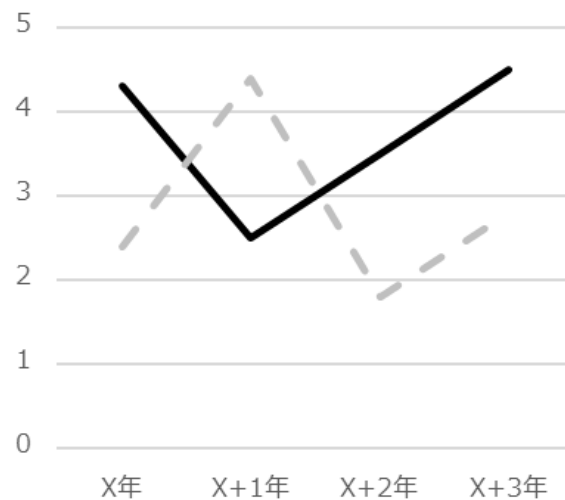
- ・その他可視化の例

レーダーチャート



異なる変数を正多角形内に軸状で表し、変数のバランスや特徴を表したグラフ。

折れ線グラフ

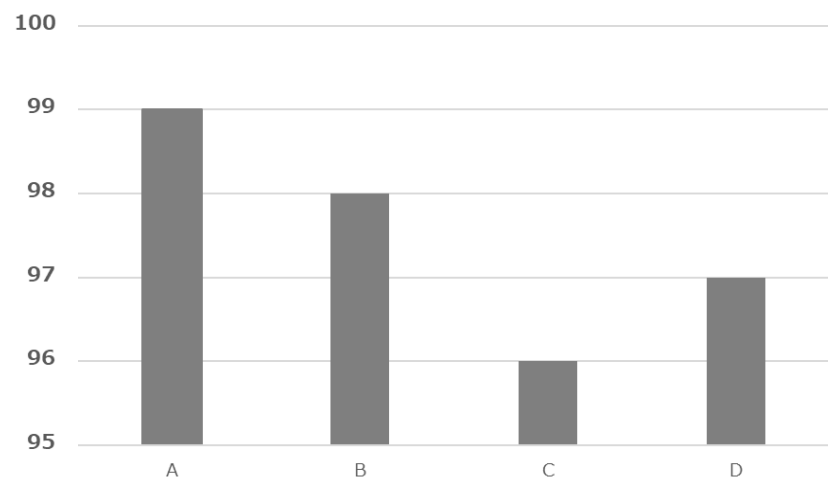


横軸に時間を取り、データの時間的変化を表したグラフ。

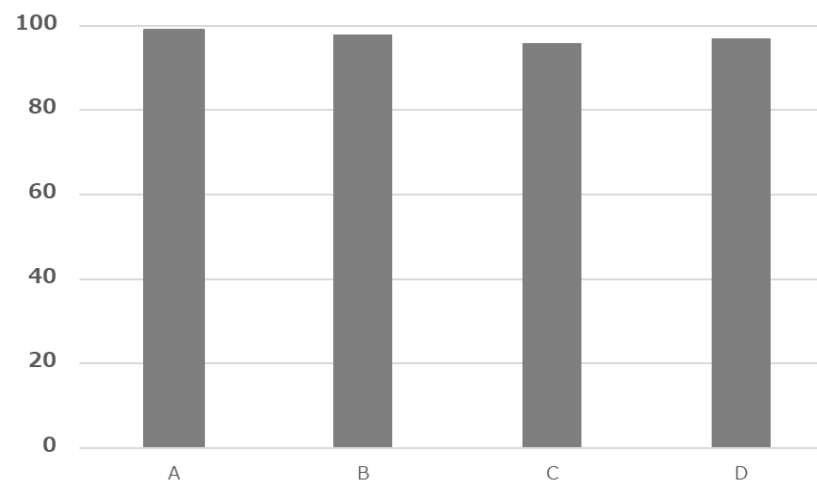
不適切な可視化の例

- ・ 不適切な目盛の設定

不適切



適切

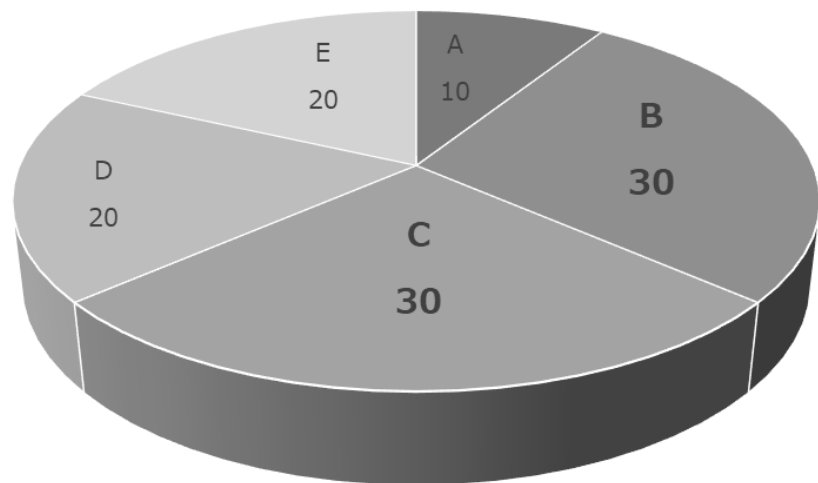


左右のグラフは同じ値を表しているが、左側は軸の最小値が95となっていることで、実際よりも大きな差があるように見えてしまう。

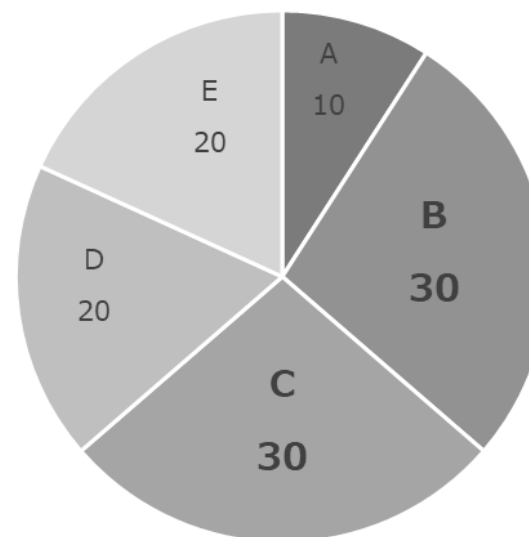
不適切な可視化の例

- ・ 不適切なグラフの配置

不適切



適切



左右のグラフにおけるBとCは同じ値であるが、立体グラフで手前に配置されると遠近法の働きによりCの方が大きな値に見えてしまう。

自分でグラフを作ったら、不適切なグラフになっていないかの確認
他者が作成したグラフを見るときは、不適切なグラフに惑わされていないかの確認

データサイエンスの技術・方法

データの解析

データを統計的に処理することでデータの特徴を見つけたり、予測・分類をする方法

・ 回帰分析

データを関数に当てはめ、あるデータの動きを別のデータの動きから説明する。

量的データを予測する場合： 単回帰分析、重回帰分析

当てはめる関数： 線形関数（直線で表される関数；一次関数）

質的データを予測する場合： ロジスティック回帰分析

当てはめる関数： シグモイド関数

【具体例】 握力、腕立て伏せの連続回数、懸垂の連続回数から球速を予測する。
入院時のいろいろな検査結果からその後（予後）を予測する。

データサイエンスの技術・方法

・判別分析

予め存在しているグループに対し、グループを分けている境界（判別関数）を見つけ、新たなデータに対してどのグループに当てはまるかを予測する。

境界（判別関数）が直線の場合： 線形判別分析

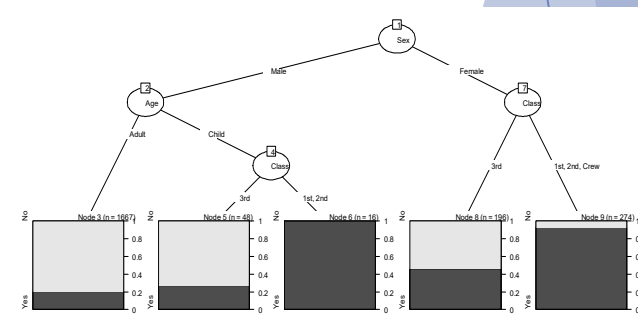
境界（判別関数）が曲線の場合： 非線形判別分析

【具体例】 いくつかの単語の使用頻度から迷惑メールかどうかを判別・予測する。
顧客の購入履歴等から商品を購入するかどうかを判別・予測する。

・決定木分析（ディジジョンツリー）

分類したい内容と関係の強い項目から順に分類ルールを作って条件を分岐し、フローチャートのような樹形図で表す。

予測の場合： 回帰木 分類の場合： 分類木



【具体例】 ある商品を購入するかどうかを性別、年齢、年収から見つける。

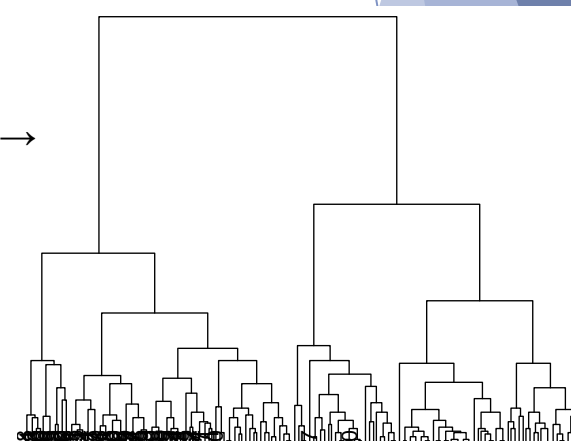
データサイエンスの技術・方法

・クラスター分析（クラスタリング）

似ているデータを同じグループ（タイプ）にまとめる（分類基準は設定せずに行う）
階層的クラスター分析、k-means法（非階層）

【具体例】 成績や学修行動から、学習スタイル（コツコツ型、一夜漬け型など）を分類する。
観光地での観光客の行動パターン（滞在日数、支出金額、移動手段など）から、
観光地を自然志向型、文化・歴史型、ショッピング型などに分類し、プロモーション戦略に役立てる。

階層的クラスター分析の例（デンドログラム）→



・相関分析

データ間の関係の有無や関係の程度を相関係数（や偏相関係数）
によって捉える。（偏相関係数：特定のデータの影響をなくした状態で求めた相関係数）

予測の場合： 回帰木 分類の場合： 分類木

【具体例】 アンケートで得られた顧客満足度と購入額の間関係を捉える。