第1章 データ加工概論

データ加工について

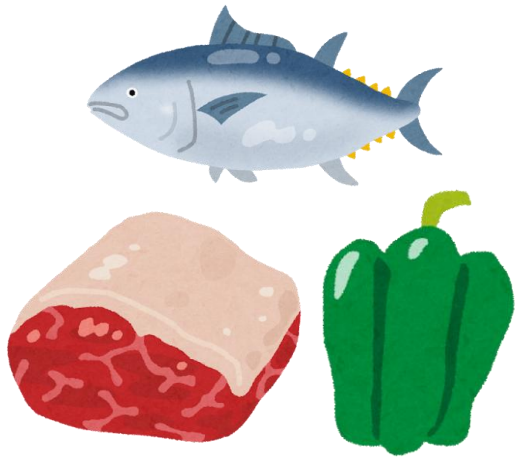
データ加工の目的

- なぜデータ加工が必要なのか？

- 実際にはどのようなことが行われているか？

データ加工の必要性

- 世の中にあるデータは、ノイズがあったりや余分なデータが含まれていることが多い
- 欲しいデータが取り出しやすい形になっていない
- データをある程度加工し適切なデータとなっている必要がある



データ収集
(食材の仕入れ)

データ前処理
(仕込み・下ごしらえ)

データ加工・分析
(調理)

データ可視化・評価
(提供する料理)

データ加工の実情

- データ処理の中で最も時間がかかり、苦労するタスクがデータ前処理だといわれている
- 諸説あるが、一般的に70%から80%のタスクがデータ前処理作業だといわれている
- データ前処理には、データのクリーニング、データの変換、データの結合、データの集約などがある
- データ前処理の出来上がりによって、後処理の結果が大きく変わることもある

名前	社名	住所	電話番号
佐藤　太郎	株式会社パイソン	東京都墨田区〇丁目〇番地	0355550000
山田一郎	株式会社　パイソン	東京都墨田区〇ー〇	03（5555）0000
後藤 頼子	㈱パイソン	墨田区〇ー〇	03-5555-0000
寺田　　学	（株）パイソン	東京都墨田区〇-〇	03ー5555ー0000
次郎石川	パイソン	〇ー〇	0 3 5 5 5 5 0 0 0 0

名前	社名	住所	電話番号
佐藤太郎	株式会社パイソン	東京都墨田区〇-〇	0355550000
山田一郎	株式会社パイソン	東京都墨田区〇-〇	03-5555-0000
後藤頼子	株式会社パイソン	墨田区〇-〇	03-5555-0000
寺田学	株式会社パイソン	東京都墨田区〇-〇	03-5555-0000
次郎石川	パイソン	〇-〇	0355550000

名前	社名	住所	電話番号
佐藤太郎	株式会社パイソン	東京都墨田区〇-〇	03-5555-0000
山田一郎	株式会社パイソン	東京都墨田区〇-〇	03-5555-0000
後藤頼子	株式会社パイソン	東京都墨田区〇-〇	03-5555-0000
寺田学	株式会社パイソン	東京都墨田区〇-〇	03-5555-0000
石川次郎	株式会社パイソン	東京都墨田区〇-〇	03-5555-0000

図1-2　データ加工のイメージ（汚いデータを何回か加工してきれいにしていく）

代表的なデータ形式

データ形式 - 表形式

- データを整理する際に最も使われる形式
- データを行と列の2次元で管理する方法。テーブル形式とも呼ばれる。
- Microsoft Excelなどの表計算ソフトで扱うデータのイメージ
- pandasのDataFrame形式で扱う

スタッフ名簿			
	名前	所属	年数
data1	佐藤	東京	10
data2	伊藤	福岡	3
data3	田中	東京	15
data4	山本	福岡	3
data5	吉田	福岡	1
data6	鈴木	札幌	8
data7	前田	福岡	5

図 1-4　表形式の例

データ形式 – ツリー形式

- データを親子関係で管理する方法
- 列方向という概念はない
- PCのフォルダでファイルを管理するイメージ
- JSON形式で扱う

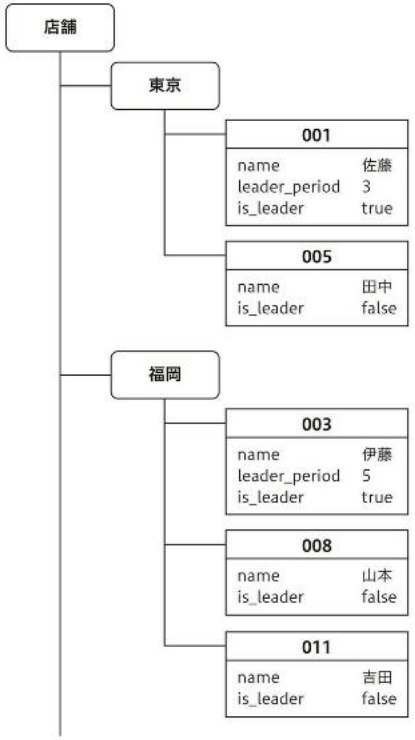


図1-5　ツリー構造の例

データ形式 – リレーション形式

- リレーショナルデータベース(RDB)で扱うデータ
- 表形式の2次元のデータと関連性をもつ
- 複数の表を関連させて、複雑なデータを表現

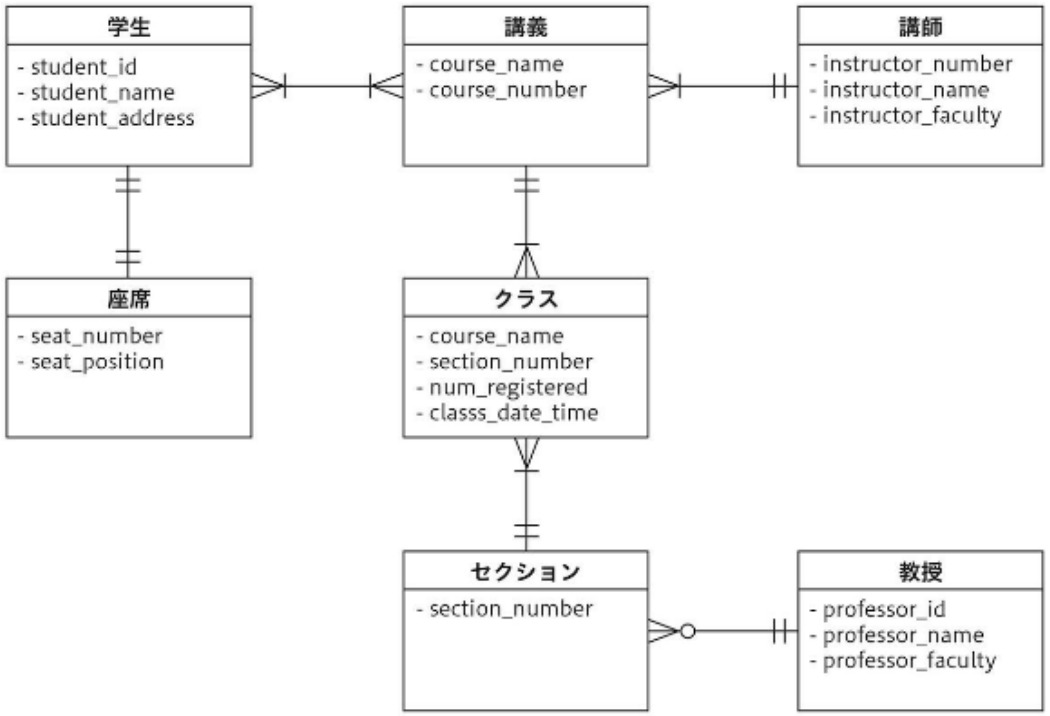


図1-6　リレーションの例

データ形式 – グラフデータ

- データを点と線で表現をする
- 点を「ノード(点)」「エッジ(線)」「プロパティ(属性)」の3つで表現
- データ間の関係性の構造を持つ
- プロパティはグラフの各要素（ノード、エッジ）が持つ情報

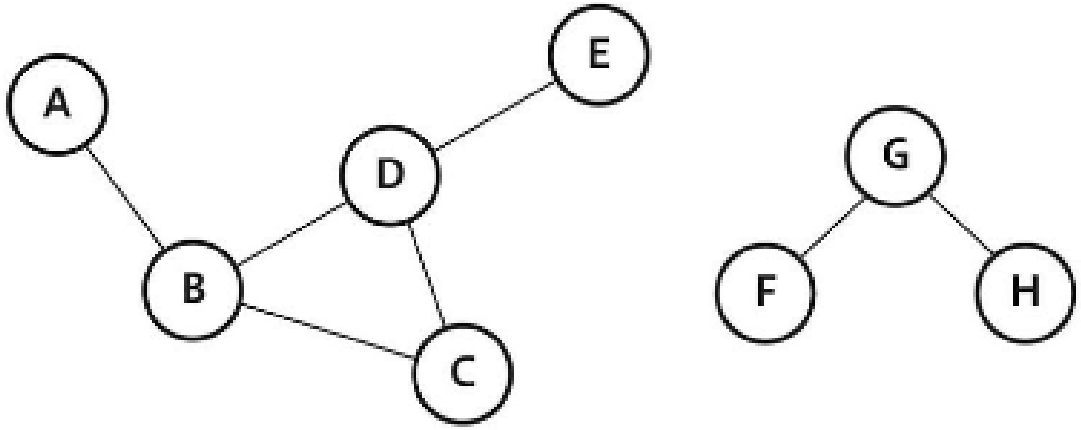


図1-7　グラフデータの例

データ形式 – 文書データ

- 文章のテキスト形式のデータ
- 文字の羅列や定型フォーマットなど，自由に作成されている
- 人間が読みやすい形式となっている

リスト1-1　サンプル文書

文書管理番号：S-2022-192
2022年8月21日

お客様各位

新サービスのご案内

株式会社パイソンED
担当：寺田　学

来月より、新サービスを開始いたします。このサービスをご利用いただくには事前の登録が必要です。
下記に示す.....

主なライブラリ

主なライブラリ

- 主なライブラリの機能

表 1-1　主なライブラリ

ライブラリ名	主要機能	説明	機能分類
NumPy	配列構造	配列の演算や加工などの機能を提供	データ型、演算、加工
pandas	データ解析／操作	1・2次元のデータ構造を提供し、加工やデータ読み書きなどの機能を提供	データ型、加工
Matplotlib	可視化	静的画像への可視化	可視化
Plotly	可視化	インタラクティブなグラフなど高度な可視化機能を提供	可視化
scikit-learn	機械学習	各種アルゴリズム実装や機械学習に必要な機能を提供	機械学習、アルゴリズム実装
SciPy	科学技術計算	高度な演算	演算、アルゴリズム実装

主なライブラリ - NumPy

- 配列構造をデータ型と提供し、数値演算や可能の機能を持つ
- Pythonにおけるデータ分析の基礎となるライブラリ

表1-2　NumPyが持つ機能

データ型	演算	加工	読み書き	可視化	機械学習	深層学習
◎	○	△				

◎：主機能　○：機能あり　△：一部の機能のみ

主なライブラリ - pandas

- データ加工ライブラリとして、1・2次元のデータ構造を提供する
- データ加工やデータ読み書きなどの豊富な機能を持ち
- Pythonでデータ分析を行うときに最も使われるライブラリ

表1-3　pandasが持つ機能

データ型	演算	加工	読み書き	可視化	機械学習	深層学習
○	○	◎	○	○		

◎：主機能　○：機能あり　△：一部の機能のみ

主なライブラリ - Matplotlib

- Pythonの可視化ライブラリで、グラフを描画する

表1-4　Matplotlibが持つ機能

データ型	演算	加工	読み書き	可視化	機械学習	深層学習
△			△	◎		

◎：主機能　○：機能あり　△：一部の機能のみ

主なライブラリ - Plotly

- インタラクティブなグラフ描画を得意とする可視化ライブラリ

表 1-5　Plotly が持つ機能

データ型	演算	加工	読み書き	可視化	機械学習	深層学習
△			△	◎		

◎：主機能　○：機能あり　△：一部の機能のみ

主なライブラリ - scikit-learn

- オープンソースの機械学習ライブラリ
- 機械学習ならscikit-learnからはじめるのがおススメ

表 1-5　Plotlyが持つ機能

データ型	演算	加工	読み書き	可視化	機械学習	深層学習
△			△	◎		

◎：主機能　○：機能あり　△：一部の機能のみ

主なライブラリ - SciPy

- 科学技術計算ライブラリ
- 微分方程式の数値解析やフーリエ変換など演算・計算をサポート

表1-7　SciPyが持つ機能

データ型	演算	加工	読み書き	可視化	機械学習	深層学習
	◎	△			○	

◎：主機能　○：機能あり　△：一部の機能のみ

EDA (Exploratory Data Analysis)

EDAとは？

- EDAは「Exploratory Data Analysis」の略
- 日本語では「探索的データ分析」と呼ばれている
- EDAは、データの特徴や構造を理解し、仮説を生成、検証するための一連の手法
- データを探索し、視覚化や統計を活用してデータの特徴や傾向を理解する

EDAの主な目的

1. データの全体像を把握する
2. 変数間の関係性を明らかにする
3. 異常値や欠損値を発見する
4. データの質を評価する
5. 仮説を生成し検証する

- EDAは、データクリーニング、データ変換、統計的分析、可視化など、多岐にわたる手法を組み合わせて行われいる
- 他のデータ分析手法と比べ、EDAは特に探索的な性質が強く、データの “理解” に重点を置いている

オープンデータ

オープンデータ

オープンデータとは、自由に使えて再利用もでき、かつ誰でも再配布できるようなデータのこと。従うべきはせいぜい「**作者のクレジットを残す**」「**同条件で配布する**」程度

出典： http://opendefinition.org/

- **オープンなライセンス**
- **オープンなアクセス**
  - 複製のための適切な費用以上の価格が課せられてはならず、インターネットを通じ無償ダウンロード可能であることが望まれる
- **オープンな形式**
  - 更新可能で簡便な形式、機械判読・一括利用が可能
  - 利用に制限や料金がかからず、自由に公開利用可能
- **単に「公表されたデータ」ではなく「開放資料」であること**
  - Open Definition 2.0 by Open Knowledge CC-BY4.0
    - Qiita で公開されているnyampire氏の翻訳を参照　https://qiita.com/nyampire/items/5aa1b3fc91890b132ab9
    - ビッグデータ・オープンデータ活用の現状と国土交通分野　～オープンデータ活用編～ https://www.mlit.go.jp/pri/kouenkai/syousai/pdf/b-141217_3.pdf

オープンデータとは、国や地方公共団体などが保有するデータを、営利・非営利を問わず、誰もが自由に利用（加工、編集、再配布など）できるように公開する取り組みのこと。

具体的には、機械判読に適した形式で、インターネットを通じて公開されれているデータ。

オープンデータの主な特徴

- 二次利用の促進
  - 公開されたデータを自由に二次利用（加工、編集、再配布など）できることが前提
- 機械判読に適した形式
  - コンピューターがデータを効率的に処理できる形式で公開
- 無償公開
  - 誰でも無料で利用可能
- 多様な活用
  - 学術研究、教育、ビジネスなど、様々な分野で活用可能

オープンデータの意義・目的

- 国民参加・官民協働の推進
  - オープンデータは、国民が政策形成や地域課題の解決に参画する機会を増やし、官民が連携してより良い社会を構築するのに役立つ
- 経済の活性化
  - オープンデータを活用することで、新たなビジネスチャンスが生まれ、経済の活性化につながることが期待される
- 行政の高度化・効率化
  - 行政機関は、オープンデータを通じて、より効率的で透明性の高い行政運営を目指すことができる

オープンデータの活用事例

- 公共交通機関の運行状況の可視化
  - オープンデータとして公開されている運行情報を活用し、バスや電車の遅延情報などをリアルタイムで提供するサービスが実現されている
- 防災情報の提供
  - 気象データや避難場所情報をオープンデータとして公開することで、災害時の情報提供や避難行動を支援するサービスが提供されている
- 地域情報の可視化
  - 地域の人口統計や施設情報などをオープンデータとして公開することで、地域課題の解決や活性化に役立てられている

日本のオープンデータカタログ

体表的な日本のオープンデータ

オープンデータ情報

- [デジタル庁 - オープンデータ](https://www.digital.go.jp/resources/open_data/)
- [政府CIOポータル - オープンデータ](https://cio.go.jp/policy-opendata)
- [総務省 - 地方公共団体のオープンデータの推進](https://www.soumu.go.jp/menu_seisaku/ictseisaku/ictriyou/opendata/)

データカタログ

- [オープンデータカタログサイト](https://www.data.go.jp/)
- [e-Govデータポータル](https://data.e-gov.go.jp/info/ja)
- [e-Stat 政府統計の総合窓口](https://www.e-stat.go.jp/)

サービス

- [RESAS - 地域経済分析システム](https://resas.go.jp/)
- [RAIDA - 地方創生データ分析評価プラットフォーム](https://raida.go.jp/)
- デジタル庁
  - [地域幸福度（Well-Being）指標サイト](https://well-being.digital.go.jp/)
    - [SCI-Japan指標サイト](https://www.sci-japan.or.jp/LWCI/index.html)
    - [Japan Dashboard（経済・財政・人口と暮らしに関するダッシュボード）](https://www.digital.go.jp/resources/japandashboard)

自治体標準オープンデータセットの対象となる組織（１）

No	データセット名	格納されている定義書	初めて取り組む基礎自治体	基礎自治体	一部事務組合等*1	都道府県	国	民間
1	公共施設一覧	A	○	○		○	○	
2	文化財一覧	A	○	○		○	○	○
3	指定緊急避難場所一覧	A	○	○		○	○	
4	地域・年齢別人口	A	○	○		○	○	
5	子育て施設一覧	A	○	○		○	○	○
6	オープンデーター一覧	A	○	○	○	○	○	○
7	公衆無線LANアクセスポイント一覧	A		○	○	○	○	○
8	AED設置箇所一覧	A		○		○		○
9	介護サービス事業所一覧	A		○	○	○	○	
10	医療機関一覧	A		○		○		
11	観光施設一覧	A		○	○	○	○	○
12	イベント一覧	A		○	○	○	○	○
13	公衆トイレ一覧	A		○	○	○	○	○
14	消防水利施設一覧	A		○	○			
15	食品等営業許可・届出一覧	A		○		○		
16	学校給食献立情報	A		○	○	○	○	○
17	小中学校通学区域情報	A		○		○		
18	ボーリング柱状図	外部		○		○	○	○
19	都市計画基礎調査情報	外部		○				
20	調達情報	外部		○	○	○	○	
21	標準的なバス情報フォーマット(ある場合)	外部	○	○				○
22	支援制度情報（給付金）	B	○	○	○	○	○	○

*1 一部事務組合等(広域連合など含む)については様々な連携ケースが存在しているため、総務省で想定している広域行政を参考に選択している。https://www.soumu.go.jp/main_content/000658630.pdf