データ分析の準備

広島大学 AI・データイノベーション教育研究センター
稲垣知宏

目標

データ分析のための準備について検討できるようになる。

この授業で紹介すること

- データクレンジングとアノテーション
- 作業の自動化

キーワード

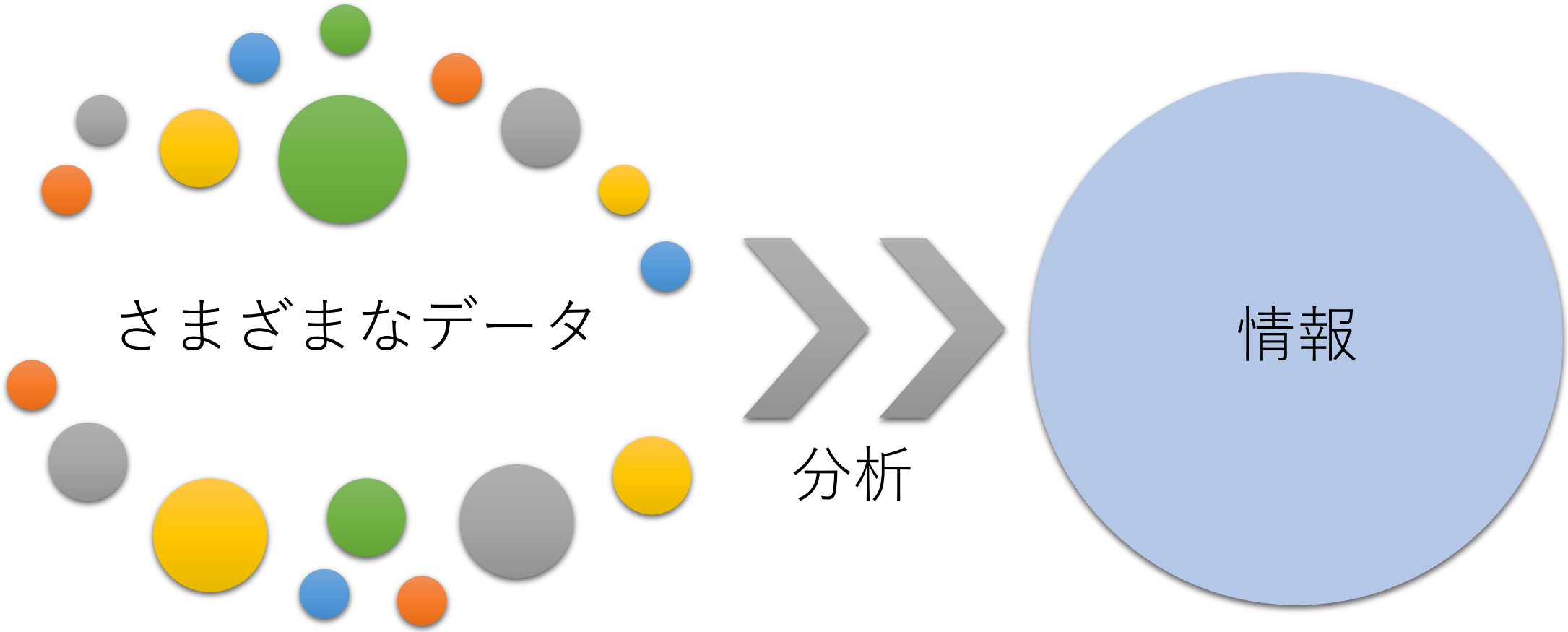
構造化データ、データクレンジング、非構造化データ、アノテーション

こんなことはありませんか？

Aさんのお父さんは、Aさんの所属する
サッカーチームの試合を、毎回、ビデオ
で撮影しています。

Aさんは、これを分析することでより良い
作戦やプレーに繋げたいと思っています。

分析するデータの形式はさまざま



構造化データ

例）リレーショナルデータベース

商品番号	品名	分類	価格	在庫
320001	甘い飴	菓子	580	23
320002	甘くない雨	菓子	680	-300
320003	赤い飴		420	67
…	…	…	…	…
…	…	…	…	…
…	…	…	…	…
…	…	…	…	…

構造化データ

列と行（フィールドとレコード）という構造を持つデータを構造化データと呼びます。

例）リレーショナルデータベース

商品番号	品名	分類	価格	在庫
320001	甘い飴	菓子	580	23
320002	甘くない飴	菓子	680	-300
320003	赤い飴		420	67
…	…	…	…	…
…	…	…	…	…
…	…	…	…	…
…	…	…	…	…

レコード

フィールド

分析の前に

商品番号	品名	分類	価格	在庫
320001	甘い飴	菓子	580	23
320002	甘くない雨	菓子	680	-300
320003	赤い飴		420	67
…	…	…	…	…
…	…	…	…	…
…	…	…	…	…
…	…	…	…	…

誤字

あり得ない値

欠損

より正確な分析のために、誤入力や欠損、重複などを修正します。これをデータクレンジングと呼びます。

構造化データの分析

商品番号	品名	分類	価格	在庫
320001	甘い飴	菓子	580	23
320002	甘くない飴	菓子	680	58
320003	赤い飴	菓子	420	67
…	…	…	…	…
…	…	…	…	…
…	…	…	…	…
…	…	…	…	…
…	…	…	…	…

クレンジング後のデータ

表計算
統計分析

可視化

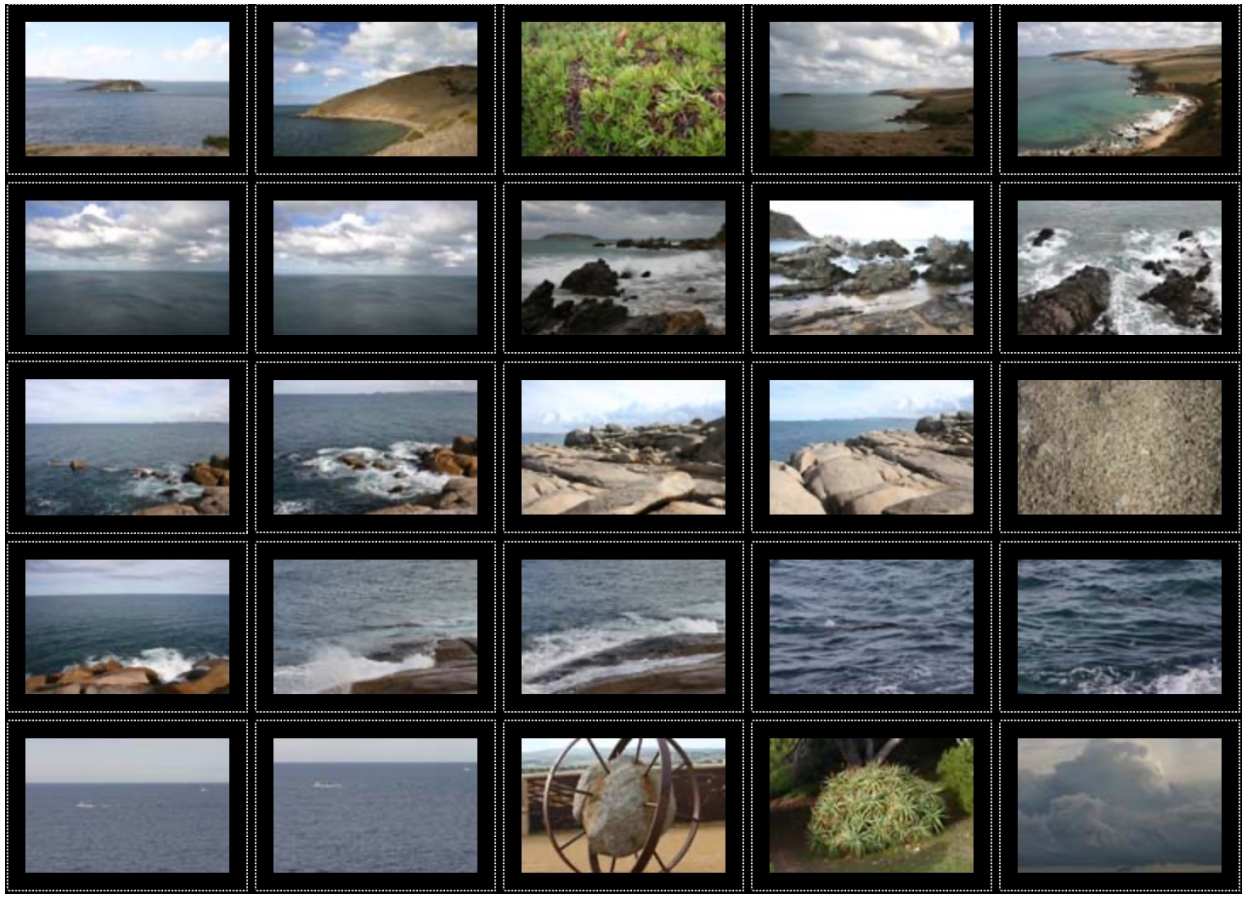
非構造化データ

構造化データにできる？

必要なデータを抽出して構造化データに

回答日	回答１	回答２	回答３	回答３
4/1	1	3	4	2
4/1	2	2	5	1
4/2	1	4	4	3
…	…	…	…	…
…	…	…	…	…
…	…	…	…	…
…	…	…	…	…

非構造化データ

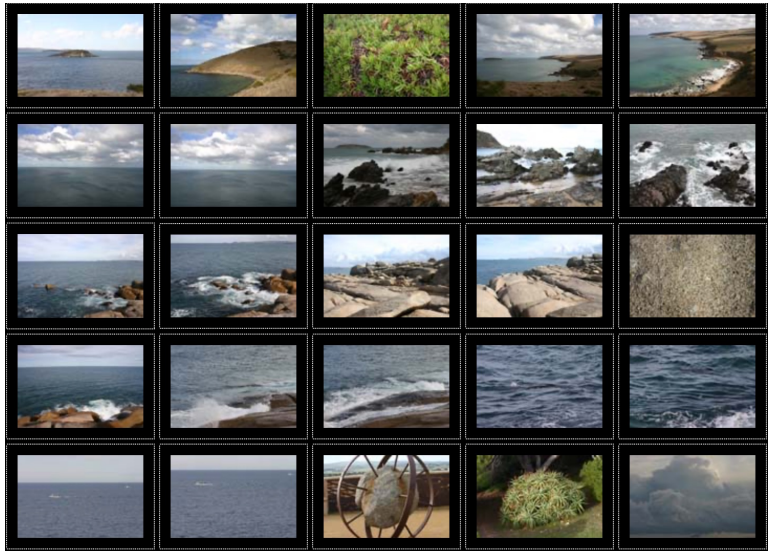


得たい情報
次第だが、

…	…	…	…	…
…	…	…	…	…
…	…	…	…	…
…	…	…	…	…
…	…	…	…	…
…	…	…	…	…
…	…	…	…	…

アノテーション

データにメタ情報を付与することをアノテーションと呼びます。



- 風景、人物、動物、…
- 撮影日、撮影場所、…
- 公開可、友達限定、プラベート
- …

アノテーション

書誌情報のデータベース

書名	著者名	出版社	出版年	頁
…	…	…	…	…
…	…	…	…	…
…	…	…	…	…
…	…	…	…	…
…	…	…	…	…
…	…	…	…	…
…	…	…	…	…

例題

ビデオで撮影したサッカーの試合を分析するには、どの様なメタ情報を付与すると良さそうか検討しなさい。

解説

以下は例ですが、この様なメタ情報が挙がったでしょうか。

- 撮影状況に関するデータ
  - ✓撮影日時、撮影場所等
- 試合成績に関するデータ
  - ✓得点、失点、シュート、フリーキックの数、反則の数など
- 個人の成績に関するデータ
  - ✓シュート、アシストの数、運動量、パスの数など
- 対戦相手に関するデータ
  - ✓チーム名、対成績など

作業の自動化

データ分析の前に行う作業として、データクレンジング、アノテーションを紹介しましたが、データ量によっては、これらを手作業で行うのは困難になります。

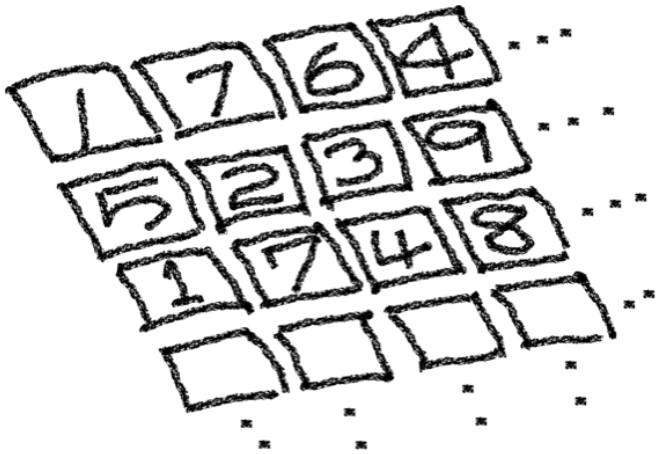
この様な場合には、作業の自動化を検討することになります。

- データクレンジング：プログラムを書いて、あり得ない値を判別、欠損、重複の修正と削除を行うなど。
- アノテーション：スマホやデジカメで撮影した際に画像データに付与されるメタ情報（撮影日時、場所など）を利用する。

機械学習は利用できる？

機械学習を利用することで、手書き文字や、猫と犬の写真を区別するなど、画像分析が可能になっています。機械学習を使うことで、アノテーション作業は自動化できるでしょうか。

教師あり学習では、メタ情報が付与された学習用のデータが必要です。学習用データ作成のために、アノテーションを行います。



1	7	6	4	…
5	2	3	9	…
1	7	4	8	…

問題

Aさんは、ほぼ欠かさず、お昼ご飯の写真を撮っています。

1. 蓄積された写真から、どの様な情報が得られるか考えなさい。
2. データを分析する前の準備として何をすれば良いか検討しなさい。