

データサイエンスへの誘い

第4回: データ処理の手法

瓜生真也（デザイン型AI教育研究センター・助教）

先週の振り返り

講義内容（予定）

講義に関する資料（スライド、補足資料等）を  GitHubに置いておきます

 <https://github.com/uribo/INNV1250>



ダウンロード可能



- | | |
|---|---|
| <ol style="list-style-type: none">1. ガイダンス、データサイエンスとは何か2. プログラミング基礎3. 再現可能性<li style="background-color: #00a0c0; color: white;">4. データ処理の手法5. データの要約6. データの可視化7. データと確率8. データからの推論 | <ol style="list-style-type: none">9. 複数のデータを比較する10. 統計のウソ11. 統計的モデリング12. 統計的学習13. さまざまなデータサイエンスの手法14. 機械学習と人工知能（AI）15. 期末試験16. 振り返りと統括 |
|---|---|

【お願い】 不要なWi-Fi接続は切断してください

講義中の情報センターJupyterHub環境を快適（少なくとも全員が）に利用できるよう

情報センターJupyterHubの利用時の留意事項

1. Server Optionsで間違った指定されたサーバーを選択した場合の対処法

メニューバーからFile、下の方にある「Hub Control Panel」から「Stop My Server」
再度「個人サーバーに移動」して選択しなおし

2. ファイル名は半角英数字のみにしておく及安全

日本語（漢字、片仮名、平仮名）、全角英数字、スペース、記号等は使わない
`mv 日本語のファイル名.ipynb myfile.ipynb` のように変換が可能

3. ダウンロードしたnotebookファイル(ipynb)は開かない

Jupyter Notebookのファイルの実体はテキストファイルです。

メモ帳、ワード等で開くことが可能ですが、文字の羅列（JSON形式）でノートブックの見た目とは異なります。

Ipynbファイルを編集する際はJupyterHubか自分のコンピュータ内にJupyter環境を用意しましょう。

今日の目標

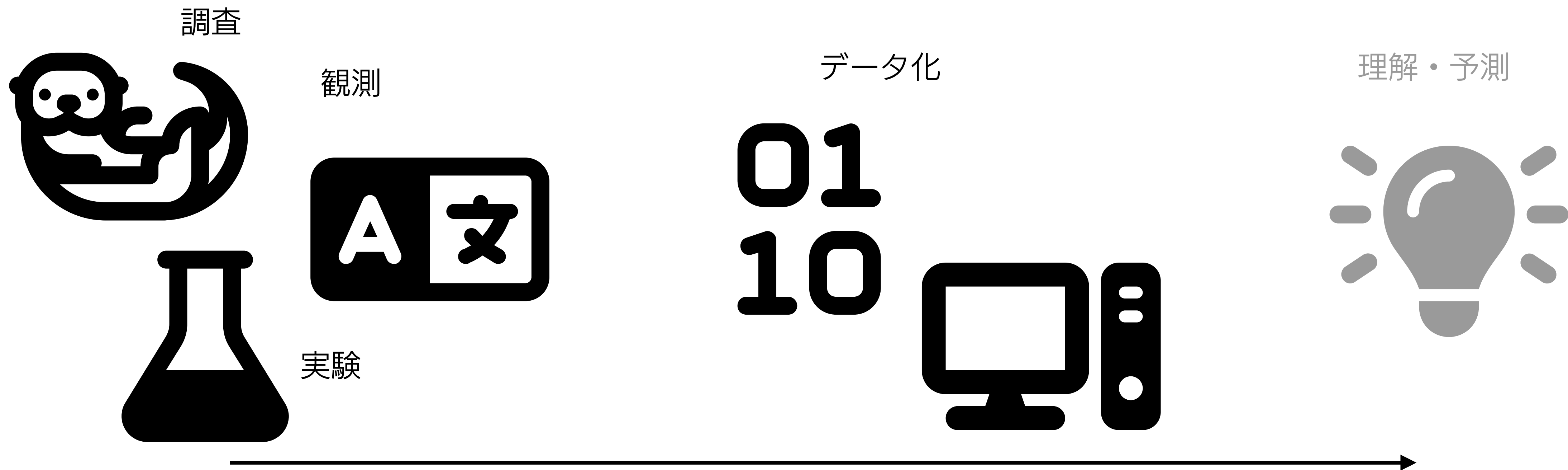
多様な種類のデータへの

理解を深める

データサイエンスとは

(参照) 第一回の講義

あらゆる種類のデータを処理・分析し、有用な情報（価値）を引き出すための学問分野



データ

判断や立論のもとになる資料・情報・事実—『スーパー大辞林』
実在から情報を抽出し、符号化する

【課題】 Rでのデータ表現・操作方法を学ぶ

提出期限: 来週の講義開始前まで

manabaのレポートとして提出してください

GitHubからr_introduction2.ipynbをJupyterHubにアップロードして記載

注意: ファイル名は英数字のみにすること

日本語（漢字、片仮名、平仮名）、全角英数字、スペース、記号等は使わない

ファイルをダウンロードしても開けなくても問題ない（気にしない）

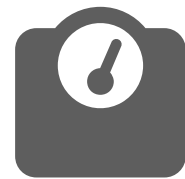
内容の確認、編集はJupyterHub上で行う

データの特徴

変数

共通の手法によって得られた値。対象によって数値が変化する値を意味する

例えば、

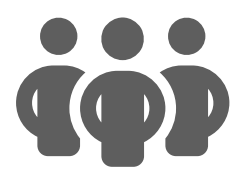


動物の体重

6

3.5

5.4



動物園の来園者数

320

615

1024



動物の分類群

食肉類

鳥類

食肉類

量的変数

質的変数



連続変数



離散変数

データを記録する精度によって小数点以下の値が変わる
Ref) 誤差

とり得る値が一定の間隔によりバラバラ

尺度水準: データの特性による分類

尺度水準に応じて、取り扱い方や用いる分析・表現手法が異なる

例) 名義尺度間での算術演算はできない
間隔尺度と比例尺度では統計量の利用ができる

変数の種類	尺度水準	判断の基準	例
質的変数	名義尺度	対象が他とは異なるか同一か	性別、出身地
質的変数	順序尺度	対象が他より「大きい」、他より「良い」など	健康度、利便性
量的変数	間隔尺度	対象は他よりもある単位によって～だけ多い（少ない）	温度、時刻、偏差値
量的変数	比例尺度	対象は他よりある単位によって～倍だけ多い（少ない）	身長、絶対温度、年齢

低

高

自由度（データの扱いやすさ）

高い水準の尺度を、より低い水準の尺度に変換できる。
例えば名義尺度である性別（「男」「女」と表現）を「男」= 0、「女」= 1のように

誤差：データの観測・測定に伴う変動

個々の測定値 = 正確な値（真の値） + 誤差

(例)  繰り返し計測を行った動物の体重

1. 複数の体重計を使う

わずかに体重計ごとに
正確さのばらつきがあるために生じる

10.460 10.441
10.442

2. 複数人がそれぞれ計測

サバを読む人、
小数点以下の値を無視する人など
記録者の性格、行動により生じる

13.681 11.0

3. 同じ体重計を使う

測定時の環境条件の変化などにより
生じる

10.774 10.763 10.599



データの不確実性、測定誤差など、さまざまな要因によって生じる

データに潜む問題

データ分析で扱うデータにはさまざまな課題が含まれる

欠損値

さまざまな理由により観測・測定されなかったデータを指す

問題: 欠損値を処理しないと統計的計算処理が不可能な場合がある… PCAなど

対処: 削除または補完による対処が求められる

外れ値・異常値

他の観測データに比して著しく乖離したデータ

問題: データ本来の性質とは異なる結果が導かれる可能性がある

対処: 外れ値を検出し、統計的アプローチなどを適用する

データの表現

構造化データと非構造化データ

構造化データ

データの扱いを容易にするため、あらかじめ定められたデータに含まれる値の性質に基づいてデータが記録される。

ルールに従ってデータが扱われるため効果的に処理できる。

データベース、表計算ソフトなど表形式のデータ全般

非構造化データ

特定のルールや並べ方が存在せずに記録されるデータの総称。

データがもつ意味や構造が曖昧であることが多い。

ビッグデータとして扱われるものに多い（文書、画像、音声、動画、センサーログ）

データフレーム: データを表形式で表現

データ分析ではデータフレーム形式でデータを扱うのが一般的



動物についての分類群と名称（種名）、体長と体重の4つの変数を記録

分類群	種名	体長(cm)	体重(km)
食肉類	レッサーパンダ	63.5	6
霊長類	チンパンジー	85.0	60
霊長類	マントヒヒ	80.0	20
食肉類	ライオン	250.0	225
鳥類	フンボルトペンギン	69.0	6

データフレームの見方

行 (row)

食肉類	レッサーパンダ	63.5	6
-----	---------	------	---

対象についてのすべての変数の値を含む

列(column)

- 分類群
- 食肉類
- 霊長類
- 霊長類
- 食肉類
- 鳥類

変数の中に全データの値を含む