広島大学

データサイエンス基礎

Fundamental Data Science

第2回 データの取得とオープンデータ，データサイエンスの倫理

広島大学　数理統計学グループ

私たちがナビゲートします！

今日の内容

- **データの取得とオープンデータ**
- **データサイエンスの倫理**

データの取得とオープンデータ

データとは

データサイエンスにはデータが必要!!

データとは…

出典: デジタル大辞泉

❶ 物事の推論の基礎となる事実. また, 参考となる資料・情報.
「データを集める」「確実なデータ」

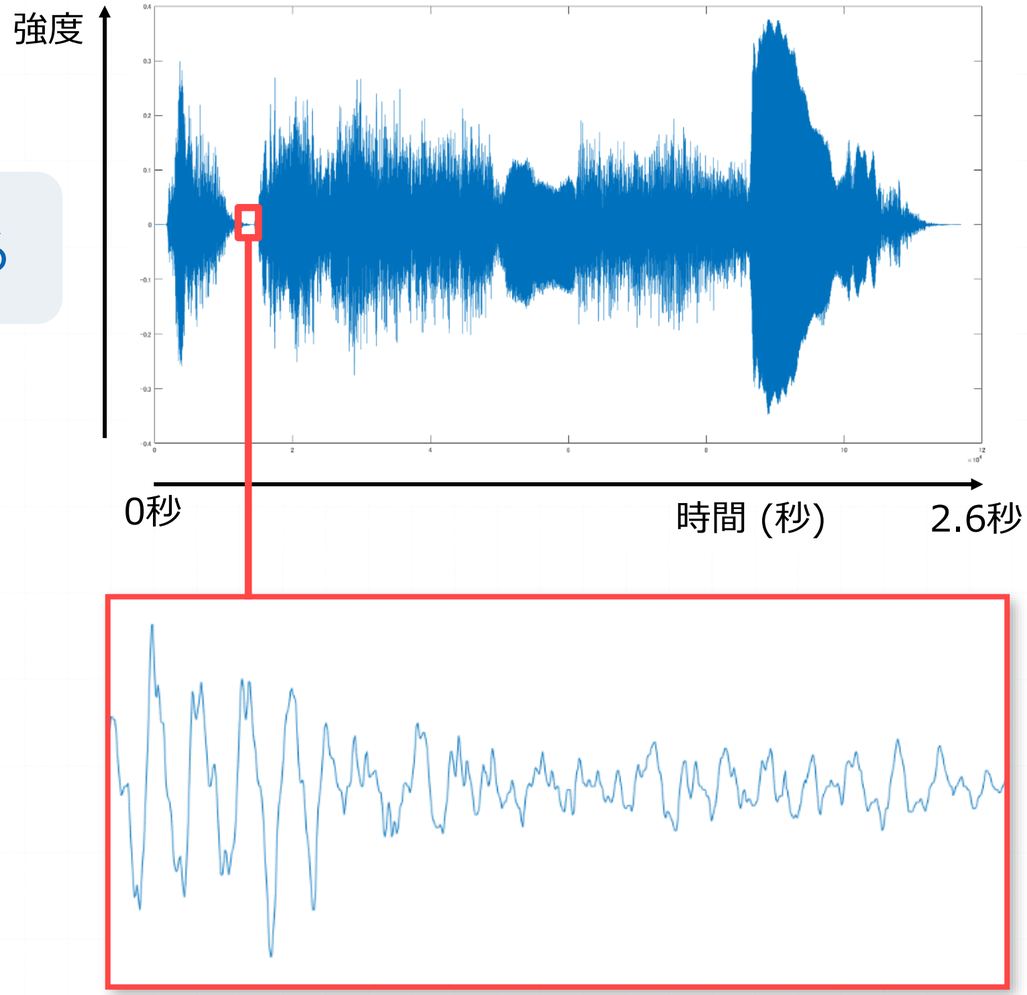
❷ コンピュータで, プログラムを使った処理の対象となる記号化・数字化された資料.

データの例

- 身長, 体重, 血圧, 試験の点数
- 顧客情報 (家族の人数, 商品Aの購入回数, 来店時間, 結婚の有無)
- 音声, 画像

実際のデータ例（音声）

音声は波で表される



強度の数値データ

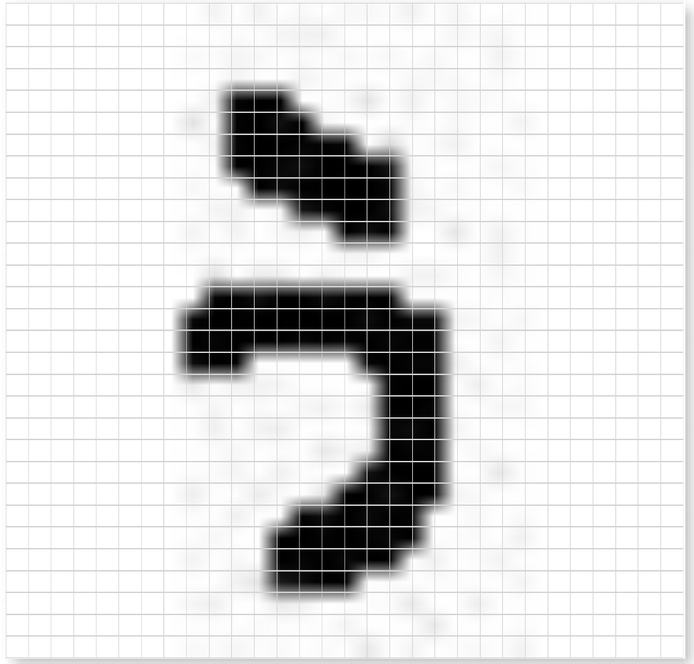
(2.6秒までで約12万個の数値)

0.054262
0.037141
0.033052
0.033387
-0.0034181
-0.031831
-0.036073
-0.036653
-0.059999
-0.09238
-0.071169
-0.06357
-0.085177
-0.011414
0.04416
0.0053713
0.056154
0.12592

音声引用: https://soundeffect-lab.info/sound/animal/

実際のデータの例（画像）

手書き文字



画像を細かい四角形に
分けて色の濃淡を数値化

255	255	255	255	255	255	255	255	255	251	255	254	255	248	255	255	252	255	250	255	254	255	255	254	255	255	255	255	255
255	255	255	255	255	255	255	255	255	255	253	255	255	255	246	255	254	253	252	255	255	251	255	251	255	255	255	255	255
255	255	255	255	255	255	255	255	252	255	250	254	255	254	254	255	253	255	255	252	254	255	247	255	255	255	255	255	255
255	255	255	255	255	255	255	255	255	255	255	255	255	251	255	255	254	255	250	255	255	254	255	253	255	255	255	255	255
255	255	255	255	255	255	255	255	255	247	0	0	0	255	252	255	255	249	255	255	248	249	255	255	255	255	255	255	255
255	255	255	255	255	255	255	255	249	255	2	7	3	0	255	250	255	255	252	255	255	255	248	254	255	255	255	255	255
255	255	255	255	255	255	255	255	254	255	0	0	0	5	4	1	251	253	250	254	255	253	255	253	255	255	255	255	255
255	255	255	255	255	255	255	255	255	255	0	5	5	0	4	0	0	5	255	253	255	254	251	255	255	255	255	255	255
255	255	255	255	255	255	255	255	255	255	255	0	0	0	0	0	0	5	250	255	255	255	255	255	255	255	255	255	255
255	255	255	255	255	255	255	255	255	247	255	253	255	0	5	0	3	0	254	255	251	255	254	253	255	255	255	255	255
255	255	255	255	255	255	255	255	251	255	249	255	251	251	254	1	2	0	255	255	255	255	255	255	255	255	255	255	255
255	255	255	255	255	255	255	255	255	251	252	255	255	255	255	252	253	250	255	255	252	255	250	255	255	255	255	255	255
255	255	255	255	255	255	255	255	252	253	255	254	252	255	253	255	255	255	254	252	254	255	253	255	255	255	255	255	255
255	255	255	255	255	255	255	255	255	4	0	0	2	0	0	0	0	0	255	252	255	255	249	255	255	255	255	255	255
255	255	255	255	255	255	255	255	0	0	0	0	6	2	0	2	1	2	0	8	249	253	255	251	255	255	255	255	255
255	255	255	255	255	255	255	255	1	0	5	1	0	0	6	0	1	0	8	0	255	255	252	255	255	255	255	255	255
255	255	255	255	255	255	255	255	1	1	0	254	253	255	254	253	2	0	2	1	251	255	255	253	255	255	255	255	255
255	255	255	255	255	255	255	255	250	250	255	251	255	255	255	255	255	0	0	1	255	248	255	255	255	255	255	255	255
255	255	255	255	255	255	255	255	255	255	255	255	255	247	254	254	253	0	4	0	253	255	251	252	255	255	255	255	255
255	255	255	255	255	255	255	255	255	254	255	249	254	255	255	255	253	7	0	0	255	255	255	255	255	255	255	255	255
255	255	255	255	255	255	255	255	255	252	254	255	255	252	255	250	253	0	4	0	255	255	255	250	255	255	255	255	255
255	255	255	255	255	255	255	255	255	251	255	255	247	255	255	255	5	1	0	0	252	255	254	255	255	255	255	255	255
255	255	255	255	255	255	255	255	248	255	255	244	255	255	252	0	0	1	5	0	255	255	250	255	255	255	255	255	255
255	255	255	255	255	255	255	255	255	255	249	255	251	0	1	1	1	0	0	255	255	249	255	251	255	255	255	255	255
255	255	255	255	255	255	255	255	255	254	254	255	0	1	1	0	0	3	0	255	255	251	255	252	255	255	255	255	255
255	255	255	255	255	255	255	255	250	255	255	251	4	0	0	5	2	0	254	252	250	255	255	255	255	255	255	255	255
255	255	255	255	255	255	255	255	255	253	249	255	0	5	1	0	254	255	253	255	254	255	251	255	255	255	255	255	255
255	255	255	255	255	255	255	255	255	253	255	255	253	250	255	252	255	242	255	255	255	254	255	252	255	255	255	255	255
255	255	255	255	255	255	255	255	255	254	254	250	255	255	255	255	255	255	255	248	251	254	255	255	255	255	255	255	255
255	255	255	255	255	255	255	255	254	255	255	255	253	254	251	255	255	255	248	255	255	253	251	253	255	255	255	255	255

数値データ

0: 黒
～
255: 白

データの取得

データを集める方法

自ら集める

例

- アンケートを行う
- 実験・観察を行う

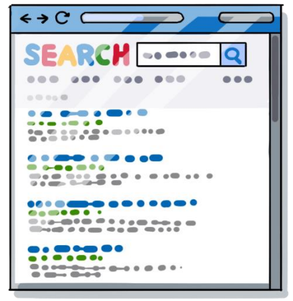
ほしいデータが集まる

だけど **労力大!!**

既に誰かが集めたデータを使う

例

- 資料請求する
- ウェブ上で公開されているデータをダウンロード



労力小!! だけど

ほしいデータがない or **もらえない** かも

オープンデータ

オープンデータ

国, 地方公共団体及び事業者が保有する官民データのうち, 国民誰もがインターネット等を通じて容易に利用 (加工, 編集, 再配布等) できるよう, 次のいずれの項目にも該当する形で**公開されたデータ**を**オープンデータ**と定義する

---

1. 営利目的, 非営利目的を問わず**二次利用可能**なルールが適用されたもの
2. **機械判読**※に適したもの
3. **無償**で利用できるもの

ざっくりというと,
「誰でも簡単に使えるデータ」
といえます

引用: オープンデータ基本指針　https://cio.go.jp/sites/default/files/uploads/documents/data_shishin.pdf　データサイエンス基礎, データ取得とオープンデータ,データサイエンスの倫理: P.8

機械判読（補足）

機械判読に適したもの

コンピュータプログラムが自動的にデータを加工, 編集等しやすいファイル形式のこと

データ分析では, 表形式ファイル (XLS, CSV ファイルなど) が扱いやすい (R では CSV ファイルをよく用いる)

※ CSV ファイルでも中身を整理しないとデータ分析で使えない

※ 機械判読可能な形式へ変換するのもデータサイエンスの仕事

よく使うファイル形式について
覚えておきましょう

オープンデータの意義

❶ 国民参加・官民協働の推進を通じた諸課題の解決，
経済活性化

新事業・新サービスの創出

❷ 行政の高度化・効率化

公共サービスの充実

❸ 透明性・信頼の向上

オープンデータの提供は
国の発展にも
役立つということですね

引用: オープンデータ基本指針　https://cio.go.jp/sites/default/files/uploads/documents/data_shishin.pdf　データサイエンス基礎, データ取得とオープンデータ,データサイエンスの倫理: P.10

オープンデータ活用例1

ココゆれ (大和ハウス工業株式会社)

家を建てる全ての人が抱える
地震が来たら危ないかな…
という不安

「ココゆれ」はあなたの家が建つ場所の地震発生確率や予測震度のリスクを教える評価ツールです。
(2012年11月サービス開始)

お客さまの居住地（都道府県・市区町村・町丁目）を選択

ご希望の結果表示期間（５年間～５０年間：５年単位）を選択

震度の発生確率，付近の断層情報（予測震度・距離・深さ・マグニチュード・発生確率）を表示

約30秒で専門知識なしで地震リスクを調べられる

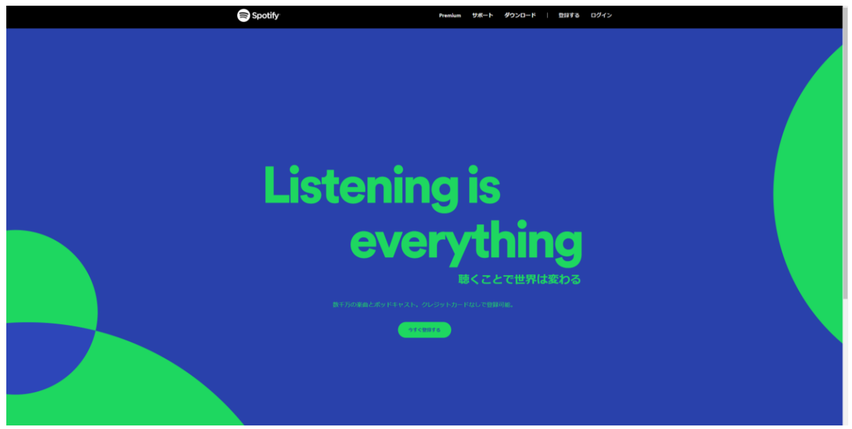
引用: オープンデータ100事例資料
https://cio.go.jp/sites/default/files/uploads/documents/od100_cases_08_kokoyure.pptx

オープンデータ活用例2

Spotify (スポティファイ・テクノロジー)

音楽ストリーミングサービス.

音楽に関するメタデータ (アーティスト名, タイトル, 言語など) を MusicBrainz から取得して配信している



参考: オープンデータを活用した アプリケーション等に関する調査研究報告書
https://www.soumu.go.jp/main_content/000618472.pdf

オープンデータ活用例3

花粉くん（株式会社博報堂アイ・スタジオ）

独自の体感ポイント KTP（カフン・ツライ・ポイント）や, 総合花粉情報をオリジナルキャラクター「花粉くん」が毎日お知らせするアプリ

統計分析ソフト「R」を使った花粉飛散地周辺ツイートの即時解析・言語解析をオープンデータと組み合わせることで, 独自の指標を計算

引用: オープンデータ100事例資料
https://cio.go.jp/sites/default/files/uploads/documents/od100_cases_07_kafun.pptx

オープンデータ取得例

オープンデータが取得可能なウェブサイト例

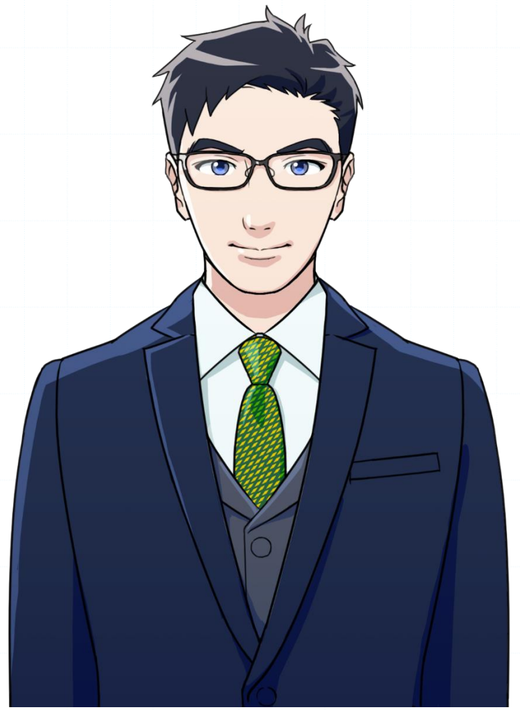
e-Stat
政府統計の総合窓口

https://www.e-stat.go.jp

e-GOV

https://data.e-gov.go.jp/info/ja

オープンデータが入手できる
サイトを紹介します



オープンデータ取得例


e-Stat
政府統計の総合窓口
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)
その他の絞込
すべて
政府統計一覧の中から探します
分野
17の統計分野から探します
組織
統計を作成した府省等から探します
キーワード検索：
例：国勢調査
検　索
●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
利用ガイド
●統計データの高度利用
ミクロデータの利用
公的統計のミ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
Data Science Online Course
社会人のための
データサイエンス演習
社会人・大学生に向けたデータサイエンス・オンライン講座
データ分析の手法を実践的に学べる講座です。
令和５年10月３日開講（受講料無料）


e-GOV データポータル
Language
ログイン
新規登録
中央行政のオープンデータポータルです。
オープンデータをご自身のプロジェクトや業務にご活用ください。
DATASET
データセット
データセットとは、ファイルやURLなどの「オープンデータ」が登録された入れ物を指します。
データポータルでは、複数の切り口からデータセットを探すことができます。
データセットをキーワードで検索

オープンデータ取得例

オープンデータが取得可能なウェブサイト例

e-Stat 政府統計の総合窓口

https://www.e-stat.go.jp

e-GOV

https://data.e-gov.go.jp/info/ja

解析事例

オープンデータ100（政府 CIO ポータル）

https://cio.go.jp/opendata100

データサイエンスの倫理

データ収集

明日のニュースで広島県民の広島県知事に対する
支持率を報道したい

広島駅で道行く人にお願いして
支持率報道の目的も伝えてアンケート調査!!

(報道後) 報道内容に嘘がないことを示すため,
アンケートデータを HP で公開しよう

公開データ

	A	B	C	D	E
1	連番	氏名	氏名（カタカナ）	性別	支持: 1, 支持しない: 0
2	1	小柳清一郎	コヤナギセイイチロウ	男	1
3	2	野中香奈子	ノナカカナコ	女	0
4	3	茂木範久	モギノリヒサ	男	1
5	4	柳沢英人	ヤナギサワヒデト	男	1
6	5	長田紗路	ナガタシャロ	女	0

何が問題･･･?

倫理に配慮したデータ収集と利活用

データ収集と利活用で倫理上配慮すべきこと

データ収集の目的，データの活用方法を**インフォームドコンセント** (説明・納得・同意) により，データ収集対象者に説明したうえで同意を得る

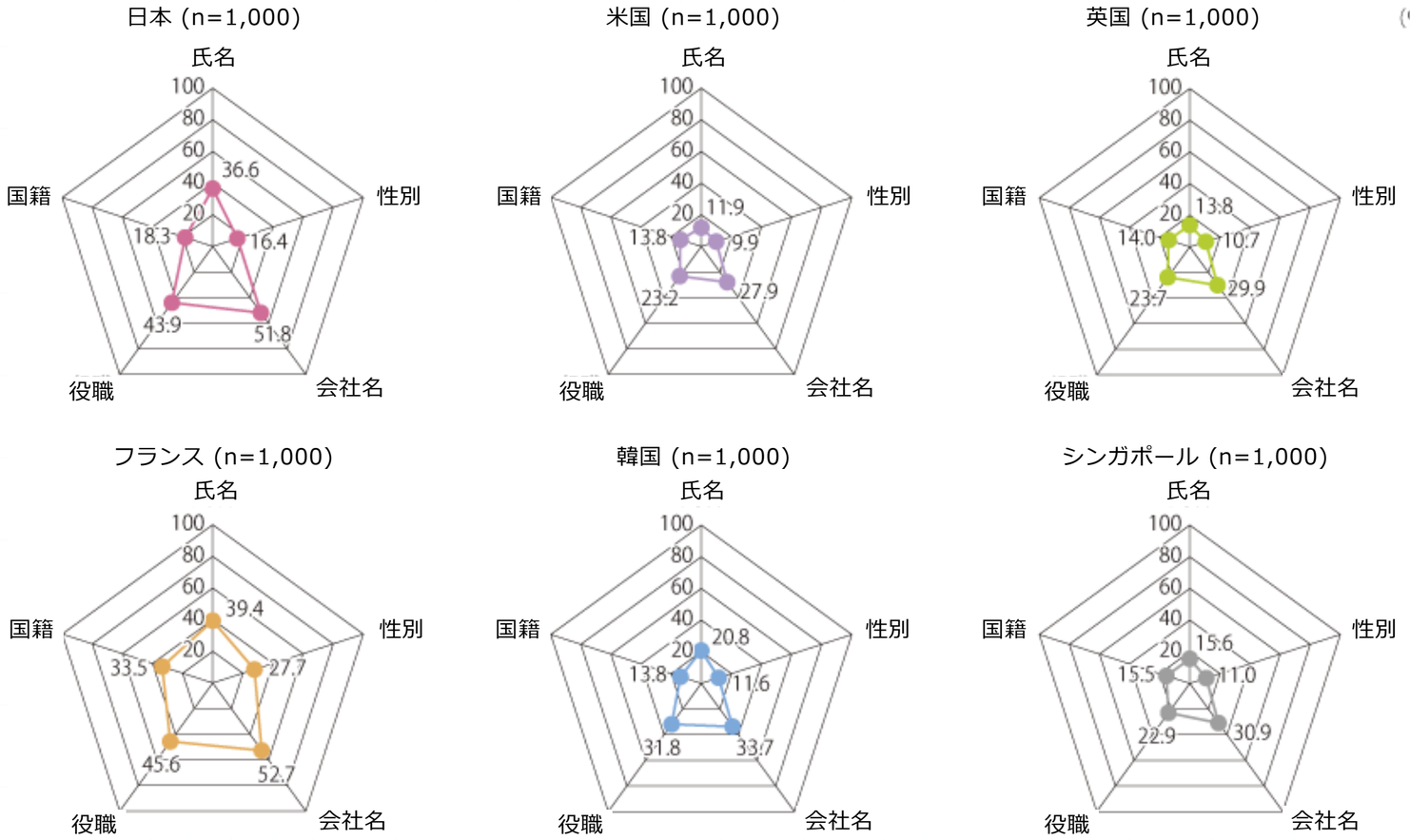
インフォームドコンセント
(医療を例に) 医師等が医療を提供するに当たり適切な説明を行い，患者が理解し同意することをいう
(精神保健及び精神障害者福祉に関する法律より)

データ収集対象者が望めばデータ収集から簡単かつ不利益を被らないように離脱できる仕組みを構築しておく

国や文化によってデータ収集対象者が他者に知られたくないデータは異なることに注意する

どのような場合でも提供・公開したくないデータ

一般パーソナルデータ (プライバシー性が高くないパーソナルデータ) の場合

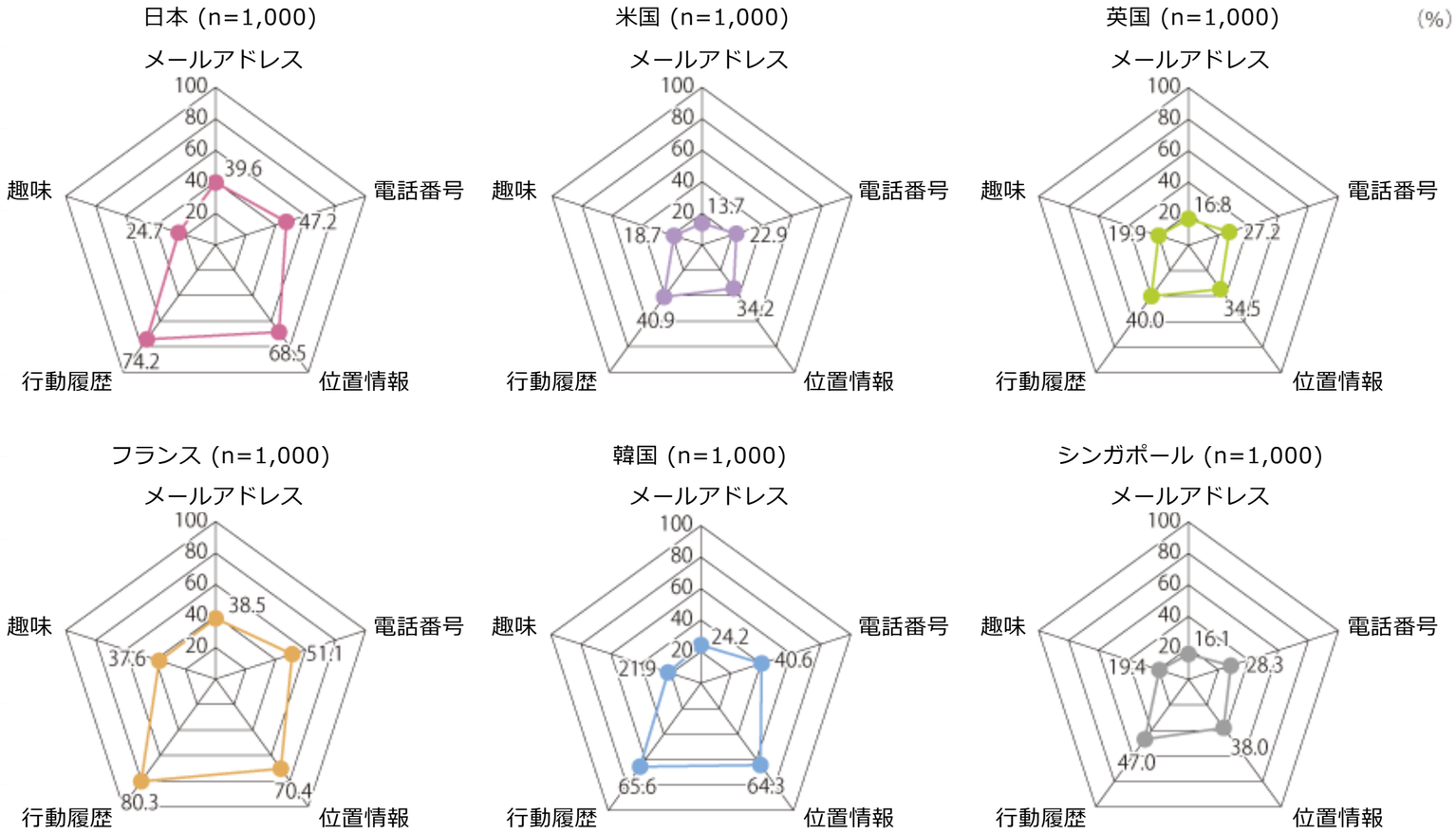


他者に知られたくない
データが,
国によってどう変わるでしょう

図引用: 総務省, https://www.soumu.go.jp/johotsusintokei/whitepaper/ja/h25/html/nc131210.html

どのような場合でも提供・公開したくないデータ

慎重な取り扱いが求められるパーソナルデータの場合

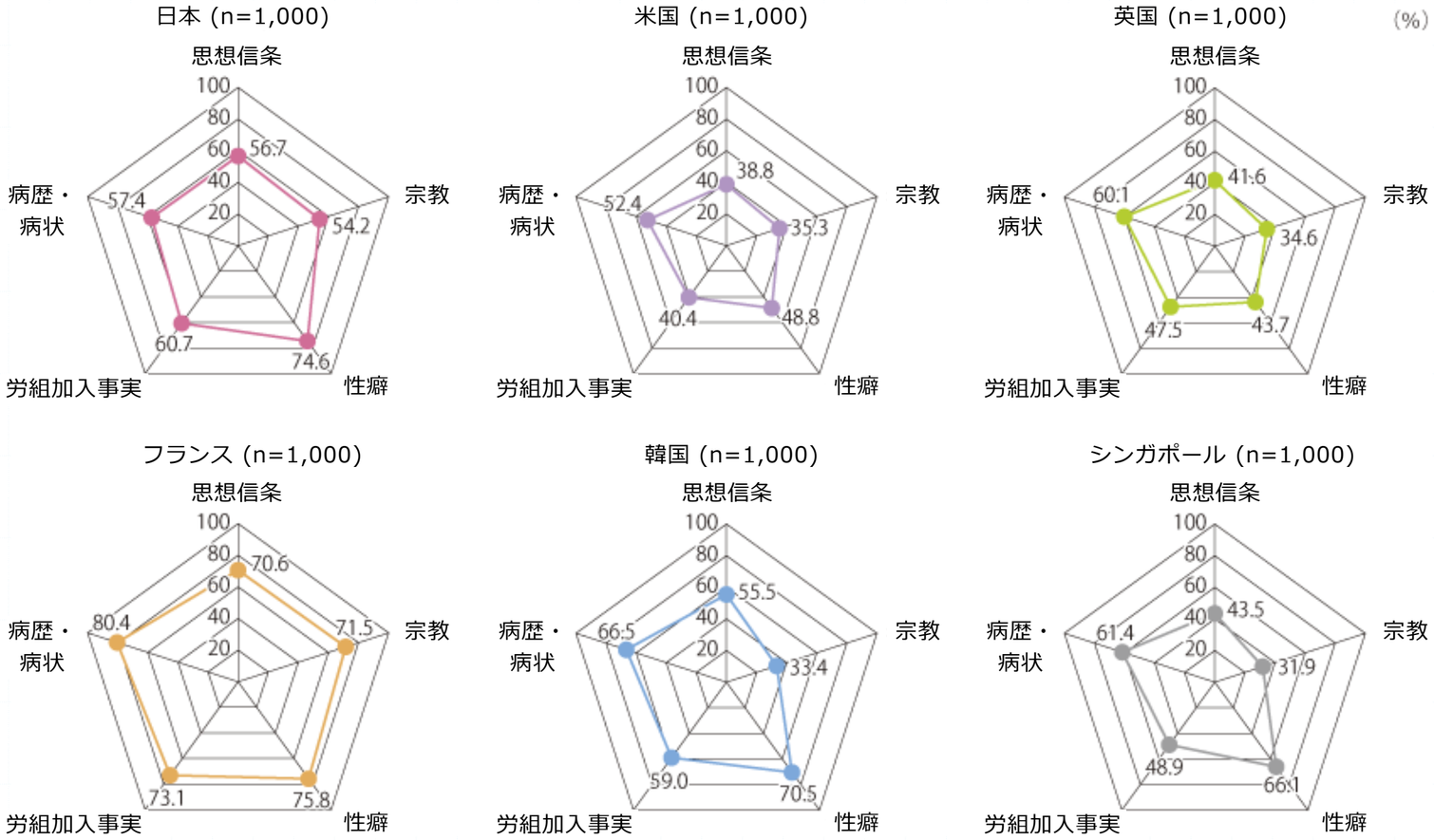


国によって
こんなに価値観が
変わるのか…

図引用: 総務省, https://www.soumu.go.jp/johotsusintokei/whitepaper/ja/h25/html/nc131210.html

どのような場合でも提供・公開したくないデータ

センシティブデータの場合



対象者の文化的な背景にも配慮すべし！ってことですね

図引用: 総務省, https://www.soumu.go.jp/johotsusintokei/whitepaper/ja/h25/html/nc131210.html

データの匿名化と仮名化

匿名化

個人の特定が**不可能**な状態にすること

例 個人を特定可能な情報 (氏名, 住所など) を**削除**する

仮名化

追加情報がないと個人を特定できない状態にすること

例 個人を特定可能な情報 (氏名など) を ID などで**暗号化**する

第三者への提供の違い

匿名化されたデータ: 原則同意がなくても第三者への提供が可能

仮名化されたデータ: 原則同意がないと第三者への提供は不可能

個人情報保護法等: https://www.ppc.go.jp/personalinfo/

データの匿名化と仮名化の例

氏名	年齢	疾患
田中一郎	103	糖尿病
鈴木春子	83	がん
山本二郎	23	肺炎
伊藤夏子	26	肺炎

匿名化

年齢	疾患
103	糖尿病
83	がん
23	肺炎
26	肺炎

個人の特定は不可能

仮名化

ID	年齢	疾患
235125	103	糖尿病
552412	83	がん
100485	23	肺炎
232039	26	肺炎

ID と氏名の紐づけを辿れば個人の特定が可能

データサイエンス基礎, データ取得とオープンデータ,データサイエンスの倫理: P.24

匿名化, 仮名化の注意

匿名化・仮名化したと思っても，個人が特定できてしまうことがある

調査地域内で100歳以上の人が１人しかいない場合

	年齢	疾患
田中	103	糖尿病
鈴木	83	がん
山本	23	肺炎
伊藤	26	肺炎

田中さんの
特定可能

修正

	年齢	疾患
田中	80歳以上	糖尿病
鈴木	80歳以上	がん
山本	20歳代	肺炎
伊藤	20歳代	肺炎

どちらが
田中さんか
特定不可

個人の特定に注意してデータを利活用する!!

何が特に問題?

明日のニュースで広島県民の広島県知事に対する
支持率を報道したい

広島駅で道行く人にお願いして
支持率報道の目的も伝えてアンケート調査!!

(報道後) 報道内容に嘘がないことを示すため,
アンケートデータを HP で公開しよう

公開データ

	A	B	C	D	E
1	連番	氏名	氏名（カタカナ）	性別	支持: 1, 支持しない: 0
2	1	小柳清一郎	コヤナギセイイチロウ	男	1
3	2	野中香奈子	ノナカカナコ	女	0
4	3	茂木範久	モギノリヒサ	男	1
5	4	柳沢英人	ヤナギサワヒデト	男	1
6	5	長田紗路	ナガタシャロ	女	0

2つの問題点が
わかりますか?
考えてみてください

他の問題点

明日のニュースで広島県民の広島県知事に対する
支持率を報道したい

広島駅で道行く人にお願いして
支持率報道の目的も伝えてアンケート調査!!

(報道後) 報道内容に嘘がないことを示すため,
アンケートデータを HP で公開しよう

広島県民の支持率を
知りたいのに, **広島駅**だけで
アンケートを行っている

広島県民全体の意見が反映されていない
= 支持率に偏り (バイアス) がある

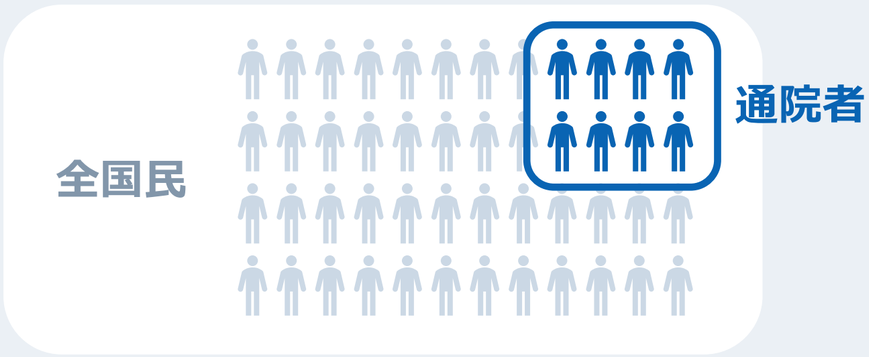
選択バイアス

選択バイアス

誤ったデータ収集方法による偏り.
その結果, 収集したデータが対象の集団を適切に代表できていない状況となる

例 **ある病気に関する全国民の有病率を調べたい**

病院を訪れた人 (通院者) から調査



通院者には病気にかかっている人が多い

有病率が大きくなってしまう

情報バイアス

情報バイアス

情報の取違いや不適切な測定方法による偏り.
その結果, 1方向に偏った測定結果がでてしまうことがある.

例 **「貯金額はいくら?」と直接的に聞いて調査**

多めに回答する人が
多くなってしまう

例 **「先月風邪を引いたか?」と聞いて調査**

自覚症状のある人のみ
風邪を引いたと回答する

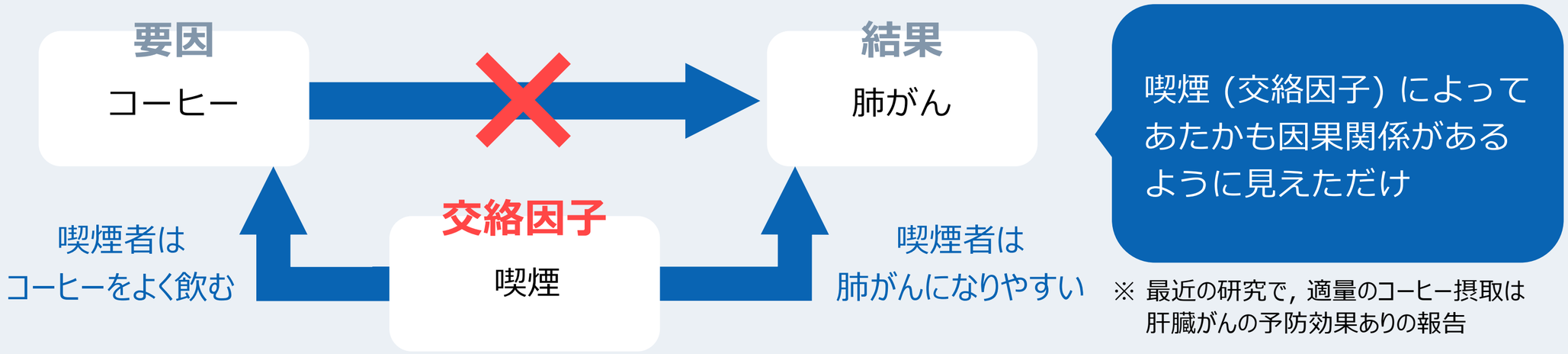
データサイエンス基礎, データ取得とオープンデータ,データサイエンスの倫理: P.29

交絡バイアス

交絡バイアス

要因と結果の両方に影響するもの (**交絡因子**という) による偏り.
その結果, 誤った調査結果がでてしまうことがある.

例 **コーヒーを飲むと肺がんリスクが上がる?**



※ 最近の研究で, 適量のコーヒー摂取は
肝臓がんの予防効果ありの報告

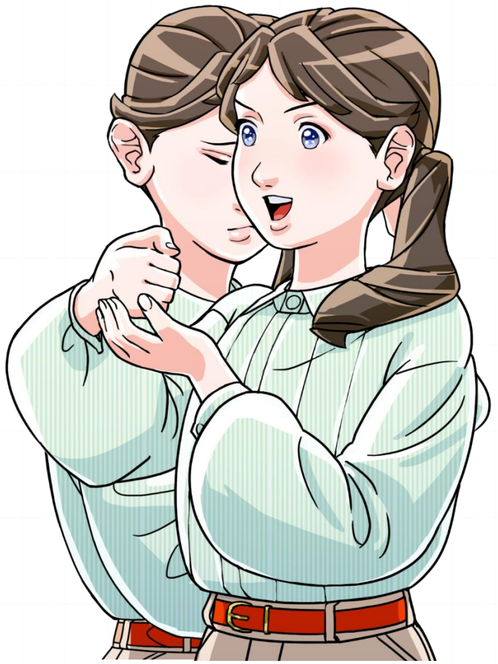
どのバイアス?

明日のニュースで広島県民の広島県知事に対する
支持率を報道したい

広島駅で道行く人にお願いして
支持率報道の目的も伝えてアンケート調査!!

(報道後) 報道内容に嘘がないことを示すため,
アンケートデータを HP で公開しよう

選択バイアス,
情報バイアス,
交絡バイアスの中だと…



ほかにも様々なバイアスが存在する
バイアスを減らすために,
データ収集の仕方を綿密に考える必要がある

データサイエンス基礎, データ取得とオープンデータ,データサイエンスの倫理: P.31

今日のまとめ

▶ データの取得とオープンデータ

データサイエンスを行うには
データが必要不可欠!!

手軽に利用できる**オープンデータ**が便利

▶ データサイエンスの倫理

自分でデータを集めるとき, 利活用するとき,
倫理的配慮, データの**匿名化・仮名化**

様々な**バイアス**に注意!!

第2回の講義は終わりです
お疲れ様でした!