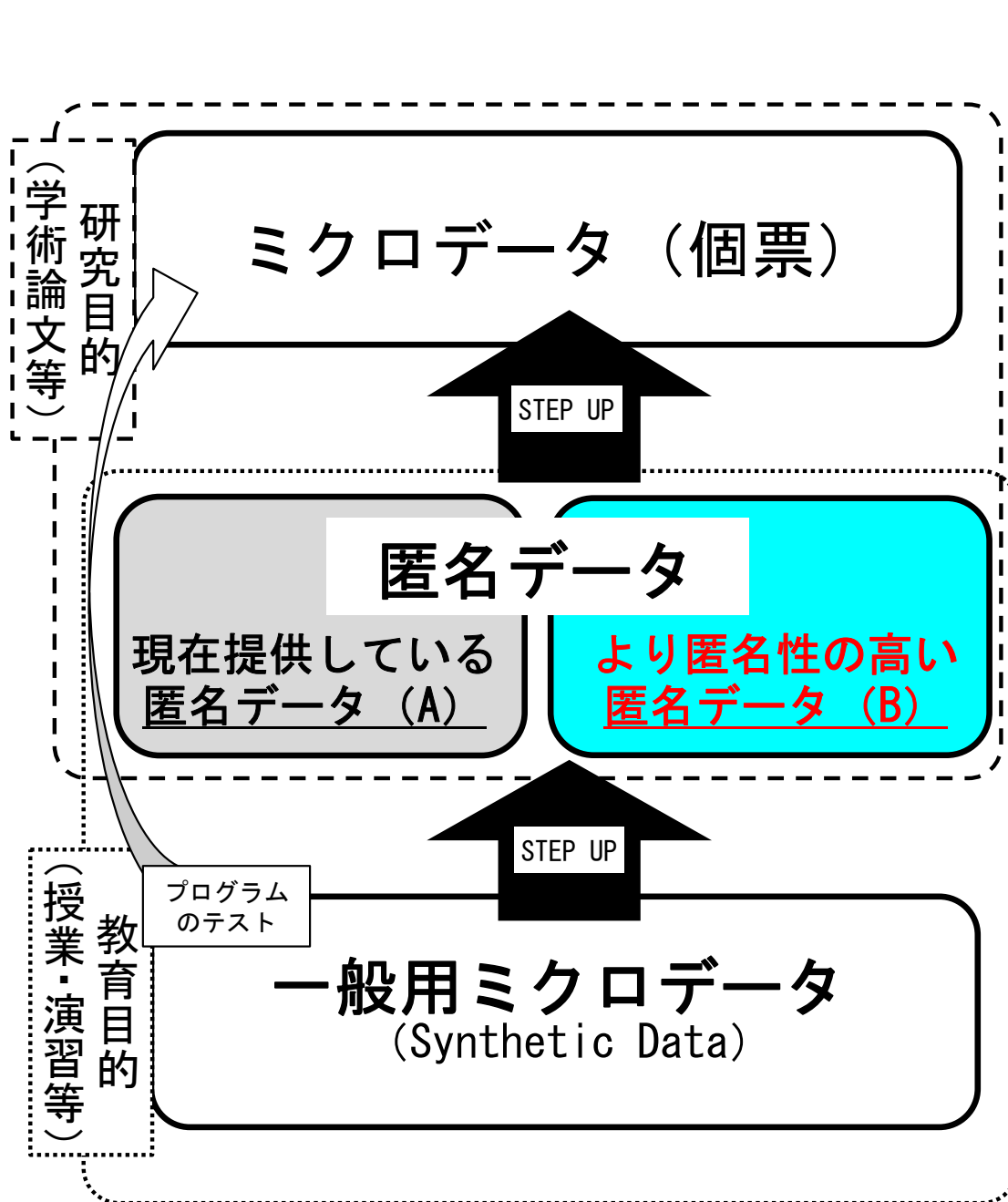




現在提供している
匿名データ (A)

より匿名性の高い
匿名データ (B)

現在提供してい「匿名データ (A)」
を基に、そこにさらに、

- ・ リサンプリング
- ・ リコーディング
- ・ トップ・ボトムコーディング
- ・ 出現頻度が低いレコードの削除

などの手法を適用して、より匿名性の高いデータ (B) を作成

(※まずは攪乱手法を用いない簡易な作成方法について検討していく)

(※各種学会で中間結果の報告を予定)

- 「R2国勢調査」の使用を検討
 - ・ 項目数がそれほど多くない
 - ・ カテゴリー変数が多い
- 教育利用の方法についても、合わせて検討していく

より匿名性の高い匿名データの教育利用方法の検討

数理・データサイエンス・AI リテラシーレベルの教育方法

➤「導入」「基礎」「心得」「選択」のそれぞれの分類ごとに、推奨される具体的な教育方法を以下のとおりまとめた。

導入	1. 社会におけるデータ・AI利活用 - データ・AI利活用事例を紹介した動画（MOOC等）を使った反転学習を取り入れ、講義ではデータ・AI活用領域の広がりや、技術概要の解説を行うことが望ましい - 学生がデータ・AI利活用事例を調査し発表するグループワーク等を行い、一方通行で事例を話すだけの講義にしないことが望ましい	教育方法（例）※ 1, 2, 3, 4
		1, 4, 5
基礎	2. データリテラシー - 各大学・高専の特徴に応じて適切なテーマを設定し、実データ（あるいは模擬データ）を用いた講義を行うことが望ましい - 実際に手を動かしてデータを可視化する等、学生自身がデータ利活用プロセスの一部を体験できることを推奨する - 必要に応じてデータハンドリングスキルを埋めるためのフォローアップ講義（補講等）を準備することが望ましい	教育方法（例）※ 1, 2, 3, 4
		1, 4
心得	3. データ・AI利活用における留意事項 - 身近で起こったデータ・AI活用における負の事例を通して、データ駆動型社会のリスクを自分ごととして考えさせることが望ましい（必要に応じてMOOC等の活用も検討する） - データ・AIが引き起こす課題についてグループディスカッション等を行い、一方通行で事例を話すだけの講義にしないことが望ましい	教育方法（例）※ 1, 2, 3, 4
		1, 4, 5
選択	4. オプション - 本内容はオプション扱いとし、大学・高専の特徴に応じて学修内容を選択する - 各大学・高専の特徴に応じて適切なテーマを設定し、実データ（あるいは模擬データ）を用いた講義を行うことが望ましい - 学生が希望すれば本内容を受講できるようにしておくことが望ましい（大学間連携等）	教育方法（例）※ 1, 2, 3, 4

※上記の「教育方法」欄の手法・形式 1～4 については次頁以降を参照

■ 令和2年（2020年）国勢調査の匿名データの項目名一覧

- ・ カテゴリ変数が項目の大半（見方によっては全て）を占める
- ・ カテゴリ変数の活用を意識した教育利用方法・匿名化を考える必要がある

項目番号	変数名	項目名
1	V0001	政府統計コード
2	V0002	管理コード
3	V0003	調査年月（西暦）
4	V0004	都道府県
5	V0005	市区町村
6	V0006	レコード一連番号
7	V0007	世帯一連番号
8	V0008	世帯員番号
9	V0009	世帯に関する事項_世帯の種類
10	V0010	世帯に関する事項_住宅の建て方
11	V0011	世帯に関する事項_建物全体の階数
12	V0012	世帯に関する事項_世帯が住んでいる階
13	V0013	世帯に関する事項_住居の種類・住宅の所有の関係
14	V0014	世帯に関する事項_世帯人員
15	V0015	世帯に関する事項_世帯の家族類型
16	V0016	世帯に関する事項_3世代世帯か否か

項目番号	変数名	項目名
17	V0017	個人に関する事項_世帯主との続き柄
18	V0018	個人に関する事項_男女の別
19	V0019	個人に関する事項_年齢
20	V0020	個人に関する事項_配偶者の有無
21	V0021	個人に関する事項_国籍
22	V0022	個人に関する事項_居住期間
23	V0023	個人に関する事項_5年前の常住地
24	V0024	個人に関する事項_教育状態
25	V0025	個人に関する事項_労働力状態
26	V0026	個人に関する事項_利用交通手段
27	V0027	個人に関する事項_従業地・通学地
28	V0028	個人に関する事項_従業上の地位
29	V0029	個人に関する事項_産業（大分類）
30	V0030	個人に関する事項_職業（大分類）

黄色塗りつぶし…連続変数（とみなすことができるもの）

より匿名性の高い匿名データの教育利用方法の検討

2. データリテラシー<スキルセット>

2.データリテラシー	キーワード（知識・スキル）
2-1. データを読む	- データの種類（量的変数、質的変数） - データの分布（ヒストグラム）と代表値（平均値、中央値、最頻値） - 代表値の性質の違い（実社会では平均値＝最頻値でないことが多い） - データのばらつき（分散、標準偏差、偏差値）、外れ値 - 相関と因果（相関係数、擬似相関、交絡） - 観測データに含まれる誤差の扱い - 打ち切りや欠測を含むデータ、層別の必要なデータ - 母集団と標本抽出（国勢調査、アンケート調査、全数調査、単純無作為抽出、層別抽出、多段抽出） - クロス集計表、分割表、相関係数行列、散布図行列 - 統計情報の正しい理解（誇張表現に惑わされない）
2-2. データを説明する	- データ表現（棒グラフ、折線グラフ、散布図、ヒートマップ、箱ひげ図） - データの比較（条件をそろえた比較、処理の前後での比較、A/Bテスト） - 不適切なグラフ表現（チャートジャンク、不必要な視覚的要素） - 優れた可視化事例の紹介（可視化することによって新たな気づきがあった事例など） - 相手に的確かつ正確に情報を伝える技術や考え方（スライド作成、プレゼンテーションなど）
2-3. データを扱う	- データの取得（機械判読可能なデータの作成・表記方法） - データの集計（和、平均） - データの並び替え、ランキング - データ解析ツール（スプレッドシート、BIツール） - 表形式のデータ（csv）

- **カテゴリ変数（離散変数）を想定した分布の可視化などを考える必要がある**
- **クロス集計による小項目の結果（例：性別・年齢別・県別・労働力状態別の人口 など）について分布の可視化等を行う など**

- **スプレッドシート（EXCEL）で扱うことのできるデータ、分析方法**

■ R、Pythonなどのプログラミング演習

- ・データレイアウトの見方
- ・データの入出力
- ・データの前処理
- ・カテゴリ別の集計など

■ 標本調査・標本抽出に関する演習

- ・匿名データを母集団とみなした上での標本の抽出
- ・抽出標本からの母数推定
- ・層化・多段抽出の効果検証など

■ データの要約・可視化に関する演習

- ・基本統計量
- ・ヒストグラム
- ・箱ひげ図
- ・クロス集計結果の可視化など

- **上記の利用に資する限りにおいて、ある程度の情報の損失を認めつつ、更なるリサンプリング、大幅なリコーディング、レコードの削除を実施**