基礎セミナー
データサイエンスのための統計学

名古屋大学 大学院経済学研究科
助教:尚 晋

April 15, 2025

今日のPoint

① 授業のガイダンス

② なぜPythonを使うか？

③ データ・サイエンスに関して

④ 第一回：統計学の基礎

⑤ 実習環境構築と「Hello World!」をプリントアウト

① 授業のガイダンス

- 別紙に参照

② なぜPythonを使うか？

- ほかの分析ツールもある：R、Matlab、Stata…
- Pythonを使う理由
  1. Pythonはプログラミング言語の中でもトップクラスで人気
  2. Python言語はデータサイエンスだけではなく、AI、人工知能やアプリケーションの開発、事務作業の自動化まで幅広く使われている。
  3. ライブラリが豊富で、コードの記述もシンプルでほかの言語よりわかりやすい。

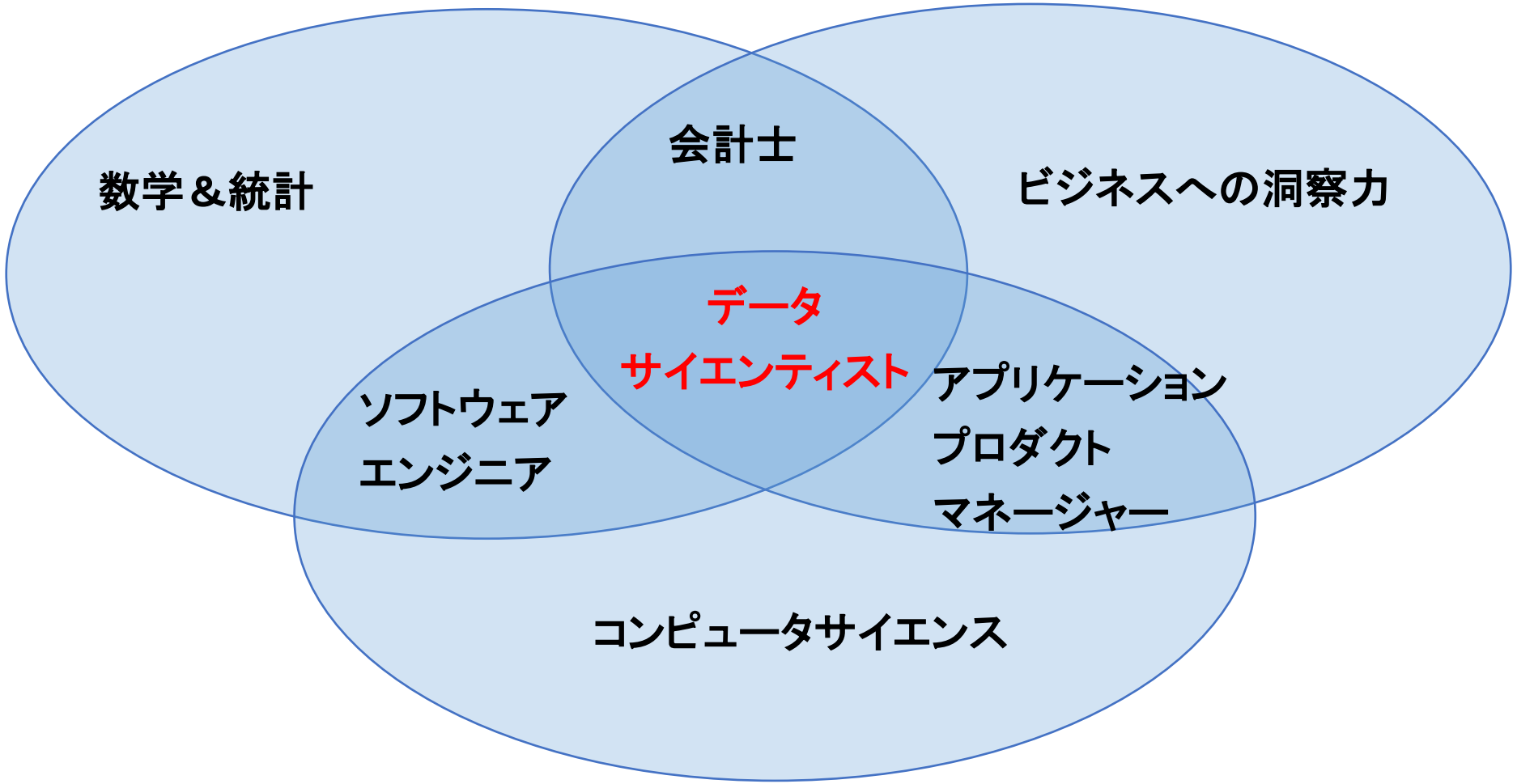
② なぜPythonを使うか？

- Pythonを応用できる分野：
  1. データサイエンス
  2. AI、人工知能、機械学習での活用
  3. アプリケーション開発・IoT開発
  4. Web上の情報収集（Webスクレイピング）
  5. ブロックチェーンの開発
  6. Webサイト、Webアプリケーション開発
  7. 画像処理
  8. 業務効率化・自動化

③　データ・サイエンスとは

- データサイエンスとは、**大規模なデータ**を使って、数学統計モデル、及びプログラミング言語、人工知能AI、機械学習などを**融合したアプローチ**を用いて、
- データに潜在した有用なパターンや情報、関係を洗い出し、
- 意思決定や戦略的な計画を組む、システム構築に有益な洞察を導く研究分野。

③　データ・サイエンスの仕事とは



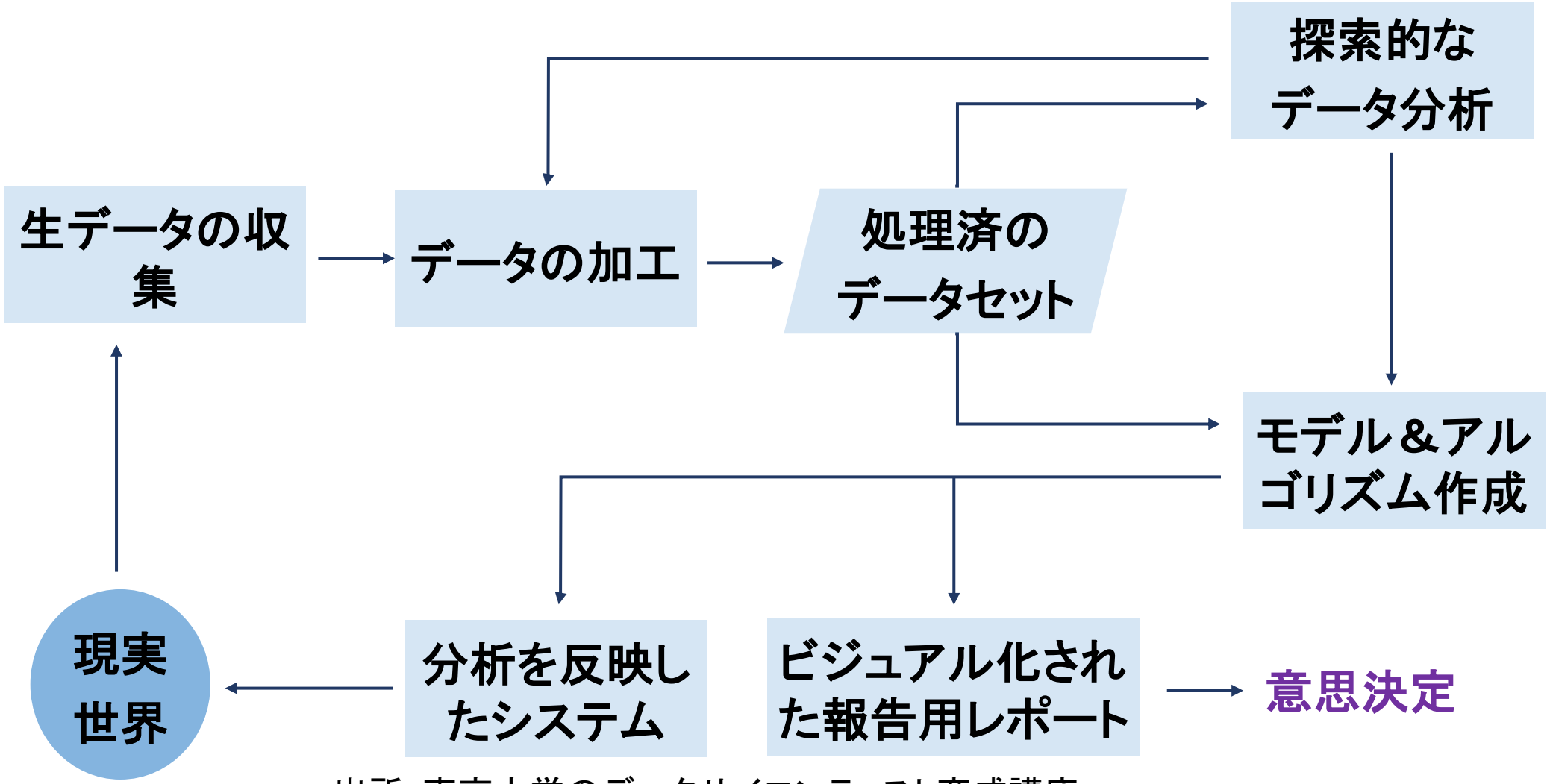
出所：東京大学のデータサイエンティスト育成講座

③　データといえば？

■データといえば、思いついたのは何かありますか？

- □ある店の売り上げという取引データ
- □ある店の顧客データ、顧客の購入履歴
- □GPSから取得した位置情報や、速度・温度などを感知・計測したセンサーデータ
- □SNSやブログの記事とか、新聞記事など、店のロコミとかというテキストデータ
- □画像と動画など

③　データサイエンスのプロセス



出所：東京大学のデータサイエンティスト育成講座

③　データサイエンスの応用

➢データサイエンスができること、()の中は例：

1. 予測分析（株価の動向予測）
2. パターンやトレンドの発見（商品やサービスの評価）
3. データマイニング（主成分分析、センチメント分析）
4. データの可視化（リアルタイムの世界各地空気汚染指数、気温）
5. 意思決定のサポート（投資のポートフォリオの最適化）
6. 最短ルート（ナビ、交通アプリのルート）
7. 画像解析（医療用の画像解析、ガンや悪性腫瘍を検出）
8. 詐欺防止と不正行為検出（不正取引、脱税）
9. Fintech（信用報告書、モバイル決済）

③　将来の就職に役立つ認定試験や資格情報：

- **データサイエンティスト検定 リテラシーレベル**：データをどう理解し、処理し、分析し、見せるか。
  https://www.mext.go.jp/a_menu/koutou/suuri_datascience_ai/00002.htm
- **Python 3 エンジニア認定基礎試験**：Pythonプログラミング言語に関する基本知識とスキルを証明する試験
- **データベーススペシャリスト試験**：データモデリングやデータベース設計の基本的な原則、SQLを高度に扱う技術、データベースの性能を向上させる方法
- **G検定**：ディープラーニングの基礎や応用についての知識、「Generalist」を意味し、ビジネスシーンで役立つ資格とされています。
- **E資格**：「Engineer」を意味し、ディープラーニングの理論を深く理解し、実際に実装できる技術を持つ人を対象とした資格です。
- **統計検定**：データサイエンスに関する、データの収集方法、記述統計、推測統計、確率論、回帰分析、実験計画法などが評価の対象です。

第一回：統計学の基礎

① 統計学とは

② 統計データと統計手法

③ 統計データの分析プロセス

いくつかの定義

- 定義：統計学
  - ■データを分析し、有用な情報を取り出す方法論のこと。
- 定義：記述統計学
  - ■データを整理・要約すること。
- 定義：統計的推測
  - ■一部の観察から全体について推測すること。

統計学の目的

- 現象の法則性を知るために，データを丹念に調べ，規則性から法則を見出し ← これは記述統計学
- また，一部を観察して，そこから論理性のある推測で全体の法則性の発見 ← これは統計的推測、推測統計学

データの用語

番号	身長（cm）	体重（kg）	性別（女=1）
1	178	63	0
2	165	62	0
3	168	69	0
4	152	41	1
...			
15	168	59	1

- 例：高校生の身長・体重・性別
- データの用語
  - **観測個体**：記録された個体ひとつひとつを観測個体と呼び、代表してiと表す。例：i ＝ 2（2 番目）の個体は、身長 165cm、体重 62kg、男性
  - **変数**：記録されている個体の情報。例：このデータの変数は、身長、体重、性別の三つ
  - **次元**：変数の数。例：このデータは三次元のデータ
  - **サンプル数**：観測個体の総数とサンプル数（或いはサンプルサイズ）呼び、nで表す。例：上のデータのサンプル数はn ＝ 15

データのタイプ

- 量的データと質的データ
  - 量的データ：長さや重さ、金額、温度、時間など。定量的に測られたデータのこと。例：身長と体重。
  - 質的データ：性別（男・女）や学歴（中卒・高卒・大卒）など、個体の属性・状態を示すデータのこと。例：性別。
- 1次元データと多次元データ
  - 1次元データ：例えば一人の学生に対して一つの観測値（身長）だけが得られる
  - 多次元データ：一人の学生に対して身長と体重という二つの観測値が得られる

データのタイプ

- 時系列データ、クロス・セクション・データ、パネル・データ
  - ■時系列データ：同一の対象の異なった時点での観測値からなるデータを時系列データ(time series data)と呼ぶ．例：1994~2024年の日本のマクロ経済データ-GDP。
  - ■クロス・セクション・データ(横断面データ)：ある時点において、複数の個体を観測することで得られるデータ．例：世界各国の2022年のCPI。
  - ■パネル・データ：定めた一定範囲の対象に対して時系列データを集めたもの．例：世界主要国の2010~2022年の石炭生産量。

データのタイプ

- 実験データと調査データ
  - ■実験データ：実験により得られたデータ．主に自然科学の分野
  - ■調査データ：調査により得られたデータ．主に人文・社会科学の分野．
- 全数調査と標本調査
  - ■全数調査：母集団全体を調査すること．例：国勢調査．
  - ■標本調査：標本を調査すること．例：世論調査．

データの分析プロセス

- 統計データの分析はつねにデータの収集に始まる，と考えるのは誤りである．最初に行うべきことは，何を対象にどのようなことを分析するか考えるということでなければならない．分析を行うべき仮説を考えるということである．仮説がないままに，むやみにデータを集めてそれを分析しても何の意味もない．
- 仮説が構築され，分析したい対象が明らかになって，初めてデータが必要となる．もし，分析に必要なデータがもともと存在しない場合には，自らデータ獲得の作業を行わなければならない．この作業は，自然科学の分野では実験experiment，人文・社会科学の分野では調査surveyと呼ばれる．
- データが手に入れば，実際の統計分析に入る．実際の統計分析は，今日ではほとんどの場合，コンピューターを用いて行われている．
- そこで次なる段階として，人間が行う，計算された結果が統計的にどのような意味を持っているかを解釈し，それを適切に表現する手段（プレゼンテーション）を考えるというプロセスが必要となる．
- 一連の統計データ分析はこのようなプロセスを経て行われる．そして多くの場合，1 回目の分析結果を見てさらに仮説を修正するという具合に，結果をフィードバックさせながら，繰り返して分析が行われる．

実習環境構築と「Hello World!」をプリントアウト

- 実習環境の構築：
  1. Anaconda
  2. Colab（おすすめ）

- 「Hello World!」をプリントアウト：
  - 「print(“Hello World!”)」を入力して、Shift+Enterキーを押す