東京科学大学（理工学系）
応用基礎レベルの取り組み紹介

特任准教授　奥村 圭司
特任教授　　宮﨑 慧

東京科学大学 データサイエンス・AI全学教育機構

Institute of SCIENCE TOKYO

数理・データサイエンス・AI教育強化拠点コンソーシアム　2024年度 関東ブロック第7回ワークショップ　2025年1月10日（金）

サマリー

- **データサイエンス・AI全学教育プログラムのご紹介**
  - 応用基礎レベル

    取組概要 / カリキュラム / 履修者の声 / 修了証
- **教材提供について**
  - 一部抜粋としても利用可能
    - ➢ 一般公開している資料（スライドPDF版, Jupyterノートブック）
    - ➢ 教員限定での提供資料（スライドPPT版, 課題）
- **プログラミング課題の自動採点について**
  - 学生が解答をセルフチェックする場合の例
- **「応用基礎データサイエンス・AI第一／第二」（2年次以上向け授業）について**
  - 科目の概要 / 講義・演習資料例 / 講義フロー・イメージ / 評価方法 / 講義実施の所感

DS・AI全学教育プログラムのご紹介

DS･AI全学教育プログラム(理工学系)の概要

Institute of SCIENCE TOKYO

- 理工系総合大学・学士修士一貫教育の特徴を活かし、大学院修士レベルの全学向けデータサイエンス・AI特別専門学修プログラムを創設　（令和元年度）
- リテラシーレベルからエキスパートレベルまで一貫した教育をめざし、リテラシーレベル教育プログラムを開始、博士後期課程科目を追加　（令和３年度）
- リテラシーレベルからエキスパートレベルへの橋渡しとなる応用基礎レベル教育プログラムを開始し、博士後期課程科目を充実　（令和４年度）

データサイエンス・AI特別専門学修プログラム【令和元年度～】
データサイエンス・AI全学教育プログラム（エキスパートレベル）【令和５年度～】
データサイエンス・AI全学教育プログラム（エキスパートレベルプラス）【令和6年度～】

共創型エキスパート

エキスパートレベルプラス（令和6年度～）

エキスパートレベル

大学院修士課程博士後期課程

先端系科目群
- 先端DS・AI第一
- 先端DS・AI第三
- 先端DS・AI第二

６学院
物質・情報卓越教育院
超スマート社会卓越教育院
エネルギー・情報卓越教育院

修士博士一貫教育により修士課程において一部科目を履修可能

共創系科目群（DS: データサイエンス）
- DS・AI博士フォーラム
- DS・AIインターンシップ

基盤系科目群
- 基盤DS/発展
- 基盤AI/発展
- 基盤DS/発展 演習
- 基盤AI/発展 演習

応用実践系科目群
- 応用実践DS・AI/発展第一
- 応用実践DS・AI/発展第二
- 応用実践DS・AI/発展第三

６学院
物質・情報卓越教育院
超スマート社会卓越教育院
エネルギー・情報卓越教育院

学士修士一貫教育・修士博士一貫教育により学士課程において大学院科目を修士課程において博士科目を一部履修可能

応用基礎レベル

学士課程（～4年次）

学内６学院における学士課程教育体系に基づいた専門教育
（専門分野に沿った独自のデータサイエンス・AI教育の実施）

６学院：理学院／工学院／物質理工学院
情報理工学院／生命理工学院
環境・社会理工学院

データサイエンス・AI全学教育プログラム（応用基礎レベル）【令和４年度～】

学士課程（2年次）

基礎系科目群
- 応用基礎データサイエンス・AI 第一
- 応用基礎データサイエンス・AI 第二

理工系教養科目(情報)
- コンピュータサイエンス第二

リテラシーレベル

学士課程（1年次）

理工系教養科目(情報)
- コンピュータサイエンス第一
- 情報リテラシ第一
- 情報リテラシ第二

データサイエンス・AI全学教育プログラム（リテラシーレベル）【令和3年度～】

（令和6年10月現在）

応用基礎レベル：取組概要（申請書より）

Institute of SCIENCE TOKYO

教育目標

共創型エキスパート人材育成に向けた応用基礎レベル教育

将来的に学生が進むであろう専門分野に依らず**数理・データサイエンス(DS)・AIを駆使して問題解決ができる能力獲得**に向け、リテラシーレベルの学修を終えて基礎的素養を身につけた学生が、**各種手法の理論やプログミング実践スキルを含むより高度な学習項目を修得**し、大学院課程におけるエキスパートレベルの学修への橋渡しとなる素養を獲得する。

実施体制

データサイエンス・AI全学教育機構運営委員会

学部と大学院を統一した組織である６学院*および卓越大学院プログラムに採択された３卓越教育院**等から選出された委員により構成

全学教育プログラムの運営・実施、改善・進化、自己点検・評価を行う全学委員会

* 理学院、工学院、物質理工学院、情報理工学院、生命理工学院、環境・社会理工学院
** 物質・情報卓越教育院、超スマート社会卓越教育院、エネルギー・情報卓越教育院

理工系教養科目（情報）実施委員会

理工系教養科目(情報)の授業を担当する教員により構成

データサイエンス・AI全学教育機構運営委員会と連携して授業の点検、評価、改善

全学教育プログラム（応用基礎レベル）科目構成

理工系教養科目(情報) 1科目1単位＋情報理工学院共通科目2科目2単位を取得

学士課程（１年次）
理工系教養系科目（情報）

4Q コンピュータサイエンス第二 1単位

全学教育プログラム（リテラシーレベル）科目構成
情報リテラシ第一・第二／コンピュータサイエンス第一／基礎データサイエンス・AI

学士課程（２年次）
情報理工学院共通科目（全学開講）

1Q 応用基礎データサイエンス・AI第一 1単位

2Q 応用基礎データサイエンス・AI第二 1単位

- 1Q: 前学期・第1クォーター、2Q: 前学期・第2クォーター、4Q: 後学期・第4クォーター
- 1年次4Qは複数クラス編成、講義0.5単位＋演習0.5単位で合計1単位
- 1Q・2Qは講義1単位、演習教材を豊富に提供することにより実践スキル向上を促す
- 1年次必修の理工系教養科目（数学）（微分積分学・線形代数学）の履修を前提

科目の特徴

- 理工系総合大学の特徴と学士修士一貫教育システムを活かしたカリキュラム編成
- 座学一辺倒ではなく実例演習も重視し、基礎的なプログラミング実践スキルを修得
- 学院間の垣根を越えたクラス編成と系統的に深く理解しやすい授業の実施
- エキスパートレベル教育への橋渡しの役割

共創型エキスパート人材とは

- DS・AIを駆使できる理論的基盤を身につけ
- DS・AIで専門の境界を越えて多様な人々と交わり
- DS・AIの未来を担う若者を教えられるトップ人材

応用基礎レベル：カリキュラム

Institute of SCIENCE TOKYO

◆コンピュータサイエンス第二

データサイエンス、機械学習の基本を学ぶと共に
より実践的なプログラミングを書けるようになる

第1回　計算とは何か（1）
第2回　計算とは何か（2）
第3回　より実践的なPythonプログラム
第4回　データ分析基礎
第5回　機械学習の基礎
第6回　コンピュータサイエンス、データサイエンス、AIに関するトピックス
第7回　コンピュータサイエンス、データサイエンス、AIに関するトピックス

チューリング機械
計算可能性理論と計算複雑性理論（P・NP）
NumPy, Matplotlib
相関、疑似相関、回帰
教師あり学習（サポートベクタ分類器）
教師あり学習（汎化）
教師なし学習（クラスタリング）

◆応用基礎データサイエンス・AI第一

データサイエンス・AIの基本理論、アルゴリズムを理解
すると共に課題演習を通して実践的スキルを身に付ける

第1回　データサイエンス・AIの基礎と歴史
第2回　データエンジニアリング基礎
第3回　データサイエンス・AI利活用基礎1　（Python言語とライブラリ）
第4回　データサイエンス・AI利活用基礎2　（Python/pandasの利用）
第5回　データサイエンス・AI数学基礎
第6回　データサイエンス基礎1
第7回　データサイエンス基礎2

ビッグデータとデータエンジニアリング、データ表現、データベース
関数、クラス・メソッド、可視化
オープンデータ、データクレンジング
線形代数、確率、確率分布
各種の離散・連続分布、チェビシェフの不等式、中心極限定理、乱数
基本統計量、相関、単回帰・重回帰分析

◆応用基礎データサイエンス・AI第二

将来のエキスパートレベルの学修につながる知識
や実践的スキルの修得を目指した学修内容を提供

第1回　数理統計の基礎1
第2回　数理統計の基礎2
第3回　機械学習1
第4回　機械学習2
第5回　ニューラルネットワークと深層学習1
第6回　ニューラルネットワークと深層学習2
第7回　機械学習・深層学習・AI応用

標本と統計的推定、点推定、区間推定
仮説検定、母分散・母比率の推定と検定、適合度と独立性の検定
教師なし学習：クラスタリング、主成分分析
教師あり学習：各種の回帰、過学習と正則化、分類、性能評価
パーセプトロン、確率的勾配降下法、誤差逆伝播法
畳み込みNN、リカレントNN
強化学習、生成モデル(VAE, GAN, 拡散モデル)、注意機構とトランスフォーマ

この2科目のみで
申請書の様式1が
求める範囲はカバー
（教材提供も可能）

応用基礎レベル：履修者の声・修了証

●履修者の声

- 統計学は統計学、PythonはPython、データ分析はデータ分析、という別々の認識でしたが、これらを関連して理解できました。とても密度が高い授業だと思います。(環境・社会理工学院)
- Pythonのライブラリの使い方、データのクレンジングがかなり自分の思い通りにできるようになった。(情報理工学院)
- データサイエンスにおける処理のなかで数学の知見が大いに活用されていることを実感できた。(工学院)
- 資料、丁寧なご講義、オンデマンド動画があって、繰り返し学習ができますので、本当に有り難いです。一度きりの対面授業では厳しいです。オンデマンドで振り返りができるので、勉強に集中ができます。(環境・社会理工学院)

●修了者全員に「デジタル修了証」（オープンバッジ）を付与

教材提供について

提供資料とライセンス

授業スライド・演習問題・プログラミング課題(自動採点対応)などを公開・提供中

https://www.dsai.titech.ac.jp
Collaboration > 教材提供

一般公開

- スライドPDF版

  CCライセンス(CC-BY-NC-SA 4.0)

- Jupyterノートブック

  CCライセンス+MITライセンス

教員限定提供

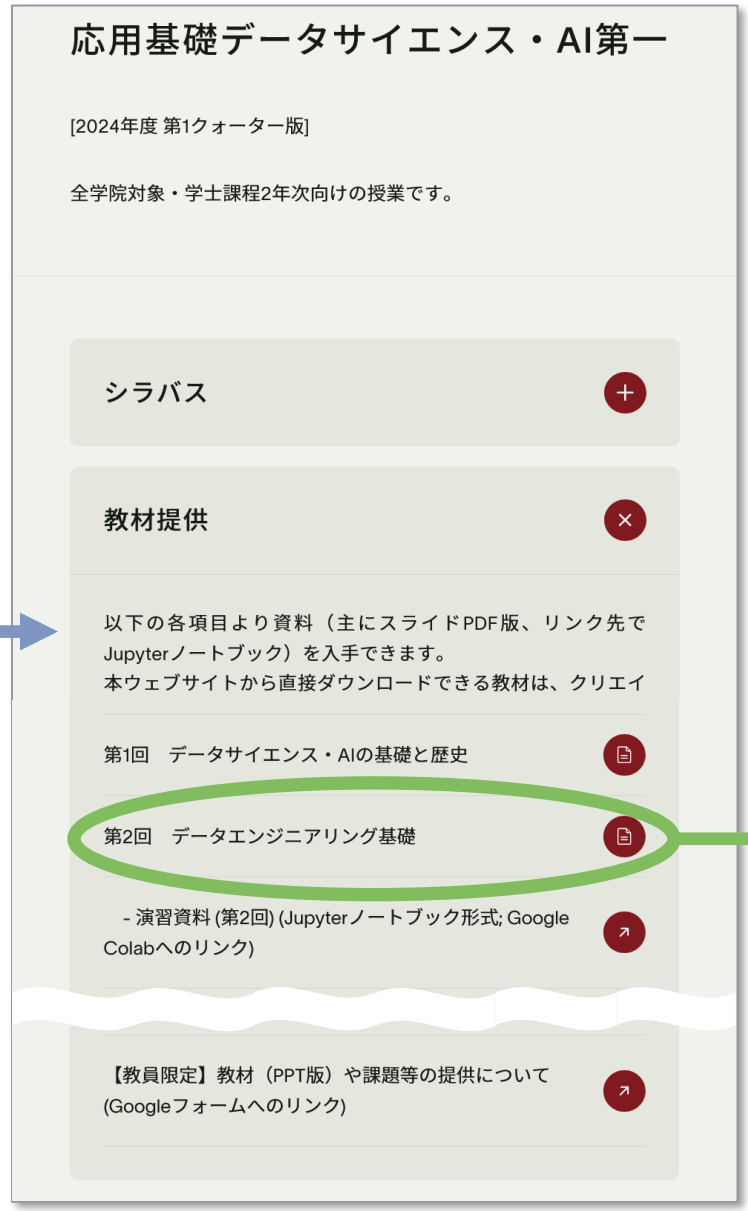
- スライドPPT版・課題など

  教材利用規約への同意により提供

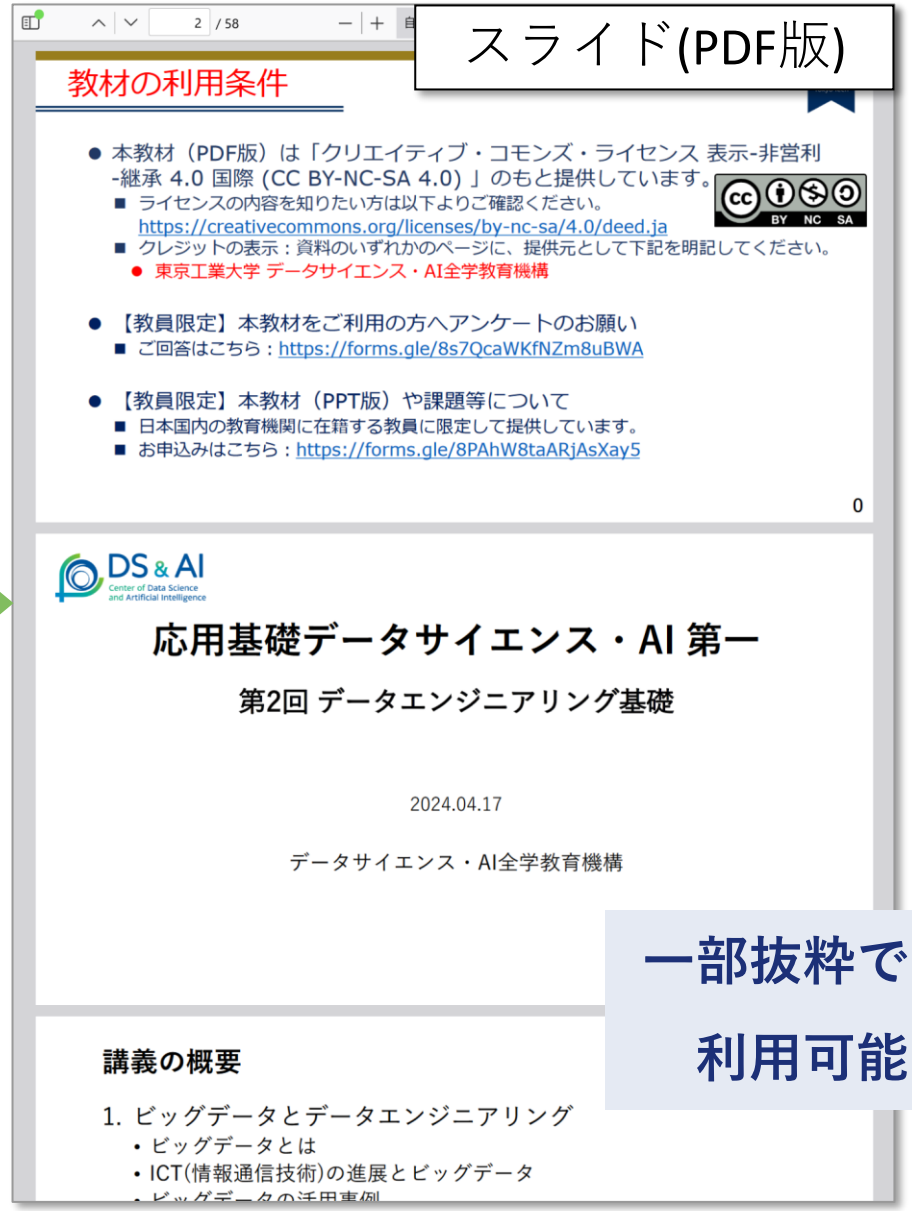
例）一般公開・スライド

Institute of SCIENCE TOKYO

www.dsai.titech.ac.jp



スライド(PDF版)

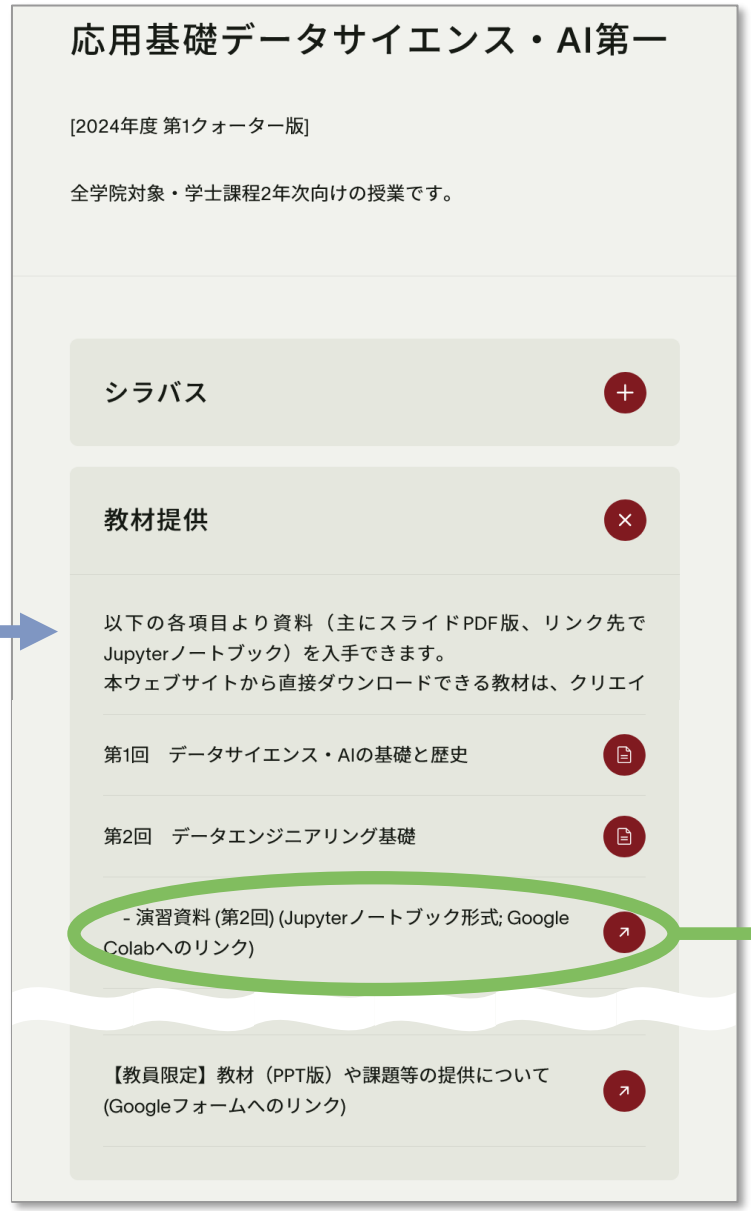


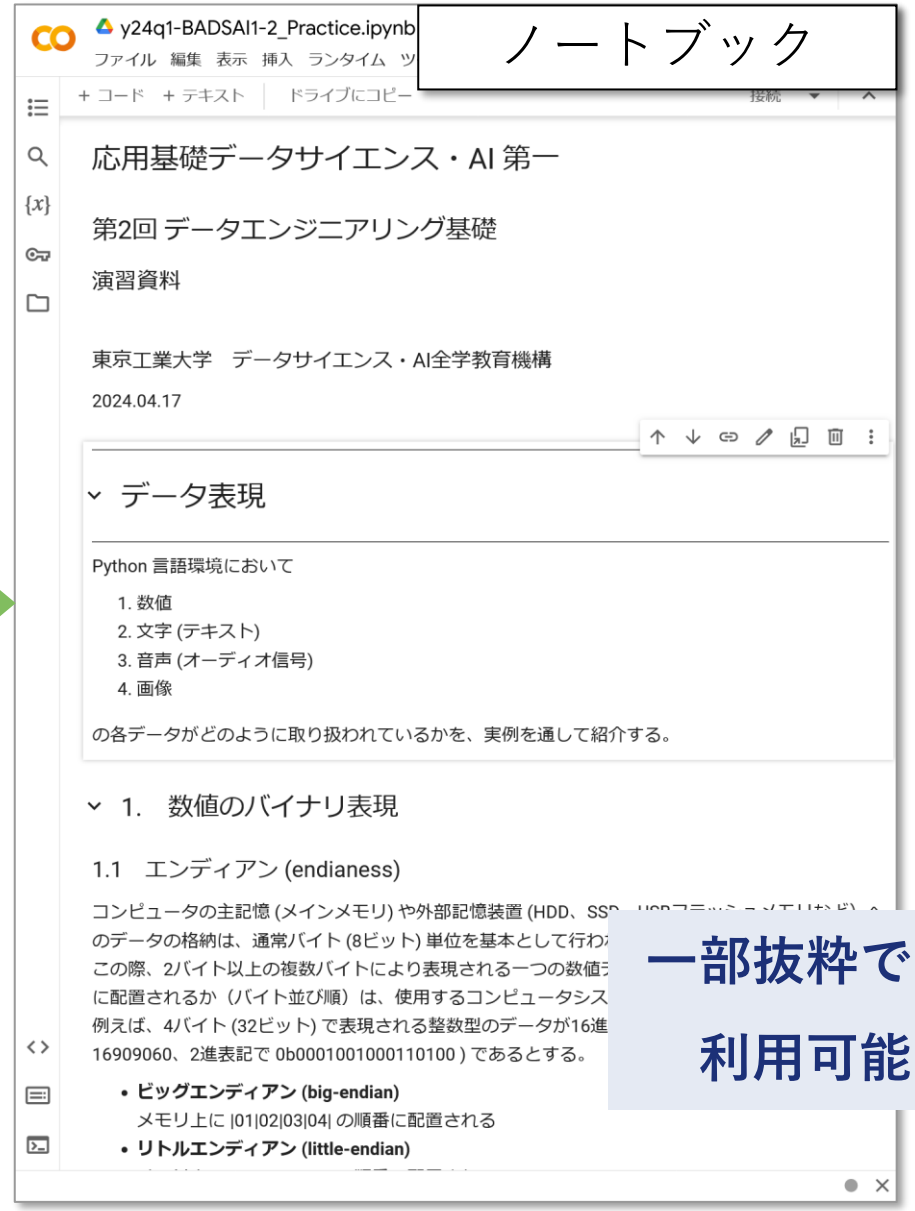
一部抜粋でも利用可能

例）一般公開・ノートブック

Institute of SCIENCE TOKYO

www.dsai.titech.ac.jp





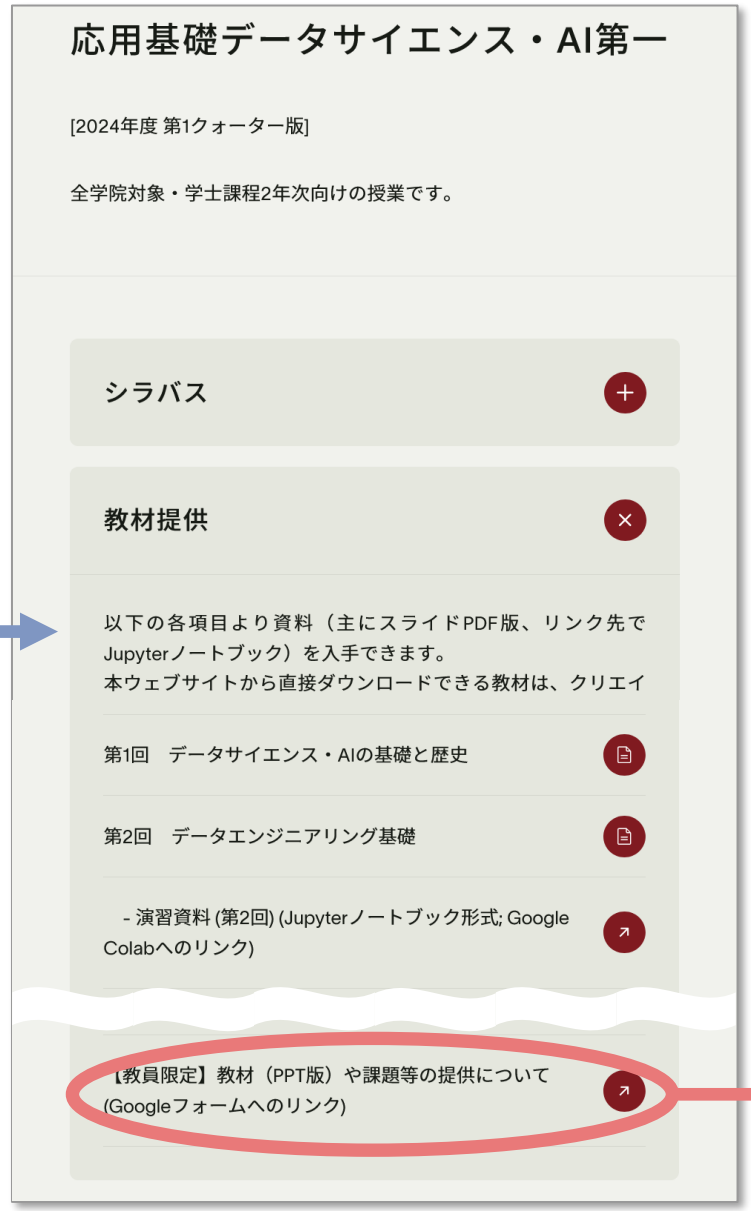
ノートブック

一部抜粋でも利用可能

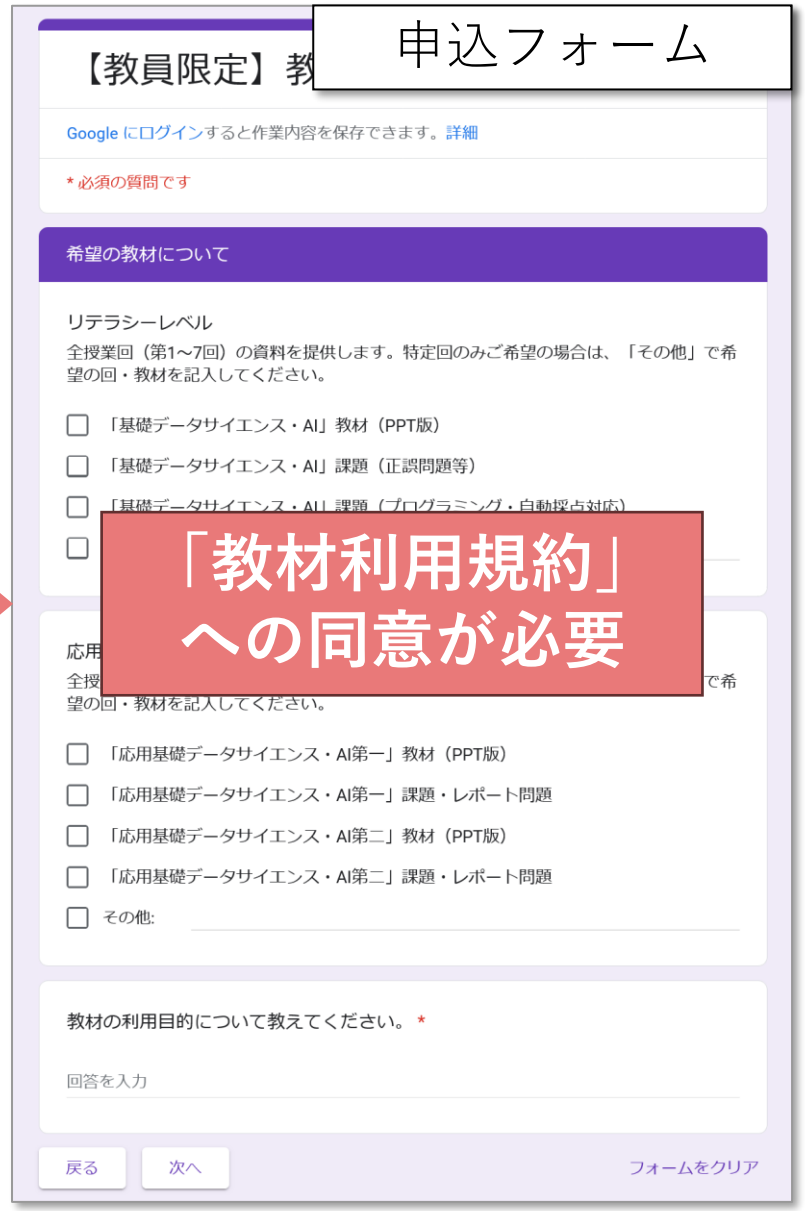
例）教員限定資料の申し込み

Institute of SCIENCE TOKYO

www.dsai.titech.ac.jp



申込フォーム



「教材利用規約」への同意が必要

教材利用規約（抜粋）

- 演習問題・課題等を教員限定で提供する場合の規約

www.dsai.titech.ac.jp/terms-of-use/

Terms of Use

教材利用規約

本教材利用規約は、国立大学法人東京工業大学（以下、「本学」といいます。）が、データサイエンス・AI教育の普及及び促進のために、日本国内の学校その他の教育機関に在籍する常勤・非常勤の教員（以下、「利用者」といいます。）に対して提供する教材の利用条件について定めるものです。利用者は、本教材利用規約に同意の上、本教材利用規約に定めに従って、教材を利用することができます。

3. 複製、配布、公開の範囲

利用者は、所属教育機関に在籍する学生及び職員に対してのみ、教材の複製、提供、配布、公開をすることができます。

5. 改変、翻訳、編集の許諾

利用者は、利用目的に則った利用のために、教材を改変、翻訳又は編集をすることができます。ただし、これらの改変、翻訳又は編集は、原著作者の表示を維持し、著作権を尊重する形で行われなければなりません。また、改変、翻訳又は編集後の二次的著作物は、本教材利用規約の定めに従って利用しなければなりません。

8. 利用者の責任

1. 利用者は、本教材利用規約に従って教材を利用する責任を負います。利用者が本教材利用規約に違反した場合には、本学の指示に従って、教材の利用停止、返還、消去等に応じなければなりません。

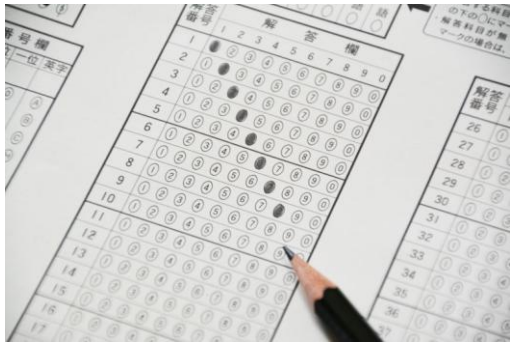
2. 利用者は、教材の修正、訂正、更新等に伴い、本学が、教材の利用停止、差し替え、消去等の要請を行ったときは、直ちに、かかる要請に応じなければなりません。

プログラミング課題の自動採点について

プログラム課題の自動採点とは？①

最も身近なイメージは、「マークシート方式」

マークシート方式
（選択式）で回答



大量の解答を
短時間で
正確に採点できる💡

番号を読み取り、膨大なデータを集計できるため、
大学入学共通テストや資格試験など、多くの場面で使用されている。

プログラム課題の自動採点とは？②

Institute of SCIENCE TOKYO

先の考え方の進化系として、プログラミング（ex. Pythonのコード）課題の採点に応用

Pythonのコードの穴あき部分に
解答（コードの自由記述）を記入

実行キーひとつで

即時に
正確に採点ができる💡

（例）
```
def tashizan(a,b):
    return [          ] ←穴あき部分に解答を記入　（正答は、a+b）
```

- 教員の採点負担を軽減する。
- 学生の自習に活用することで、セルフチェック機能としても有効。
- 自分の書いたコードをチェックしたり、正解へのヒントを得られる。

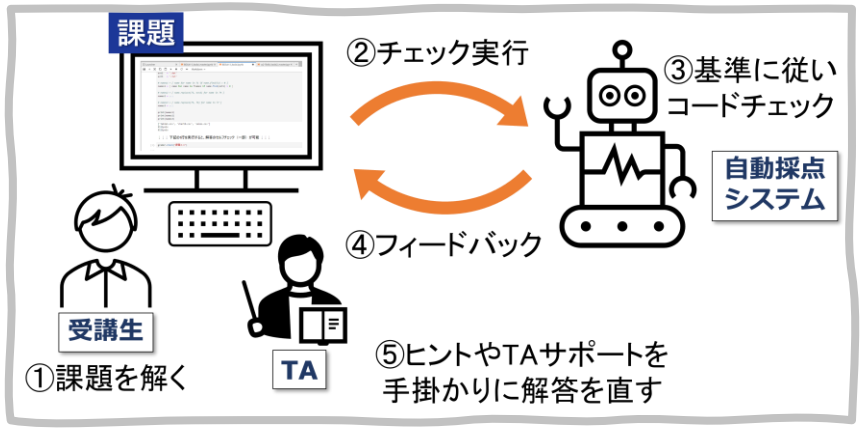
日本国内の大学への普及

自動採点システムを本格的に開発・運用し、
世界でも先進的な取り組みを実施している
カリフォルニア大学バークレー校（UC Berkeley）と連携

考え方自体はシンプル

しかし、ソフトウェアは
開発よりもメンテナンス
が大変

「基礎データサイエンス・AI」など理工学系の授業に
自動採点システム **Otter-Grader** を導入
Jupyterノートブックによる課題へ適用



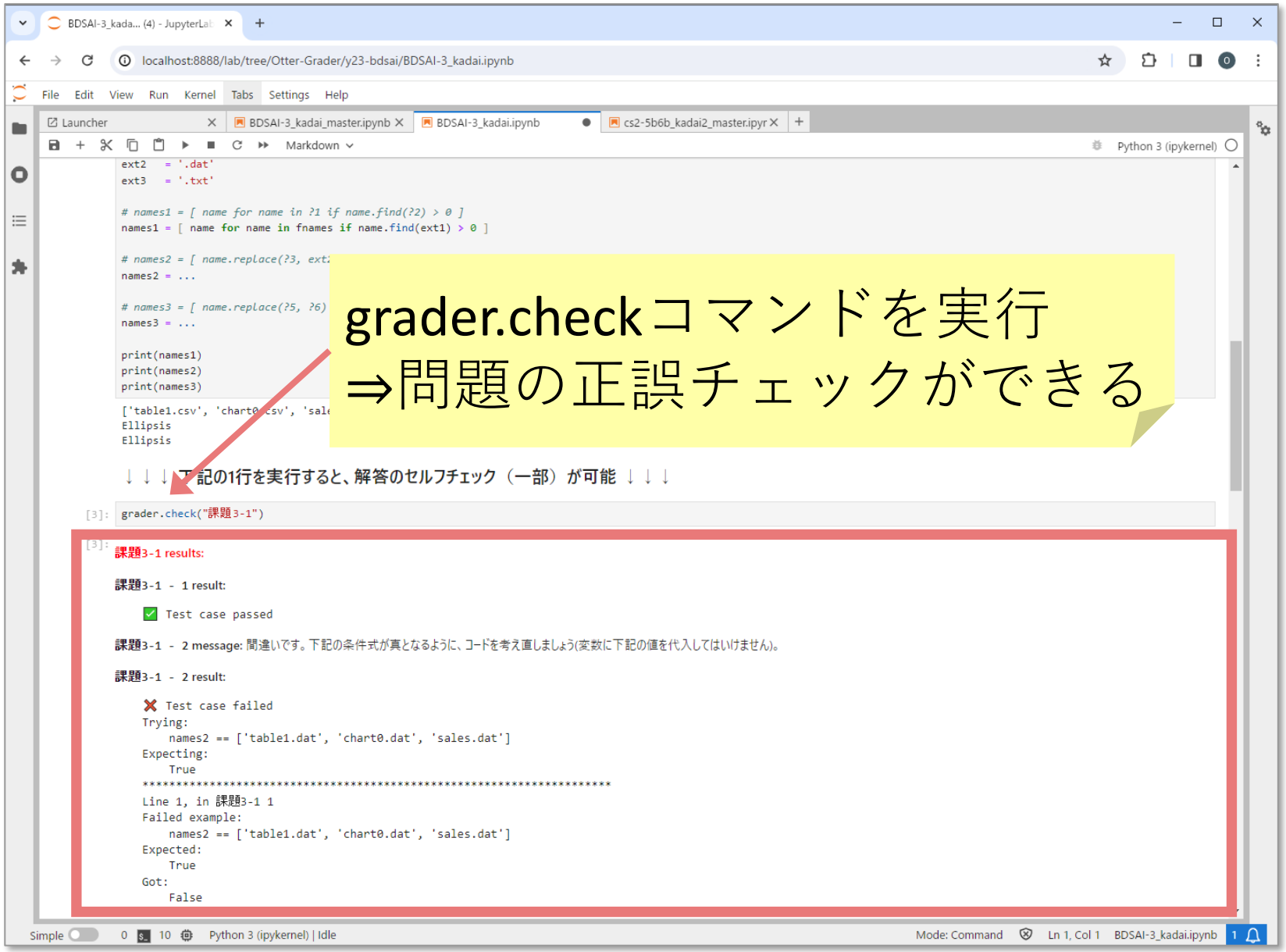
日本国内の様々な大学への普及を推進

- 「数理・データサイエンス・AI教育強化拠点コンソーシアム」の拠点校として、他大学への普及を推進
- 東京科学大学の発足に伴い、同大学医歯学系への普及を推進

提供教材の中に
導入マニュアルや対応課題あり

例）学生が解答をセルフチェックする場合

Institute of SCIENCE TOKYO



TAによるサポート
との二本柱

モデルカリキュラム中心科目
「応用基礎データサイエンス・AI第一／第二」
（2年次以上向け授業）
について

科目の概要

- 科目名
  - 応用基礎データサイエンス・AI 第一（以下第一と略記）

    第１Q（計7回）・各回100分・単位数：1（講義：0.5　演習：0.5）
  - 応用基礎データサイエンス・AI 第二（以下第二と略記）

    第２Q（計7回）・各回100分・単位数：1（講義：0.5　演習：0.5）
- 到達目標
  - データサイエンス、データエンジニアリング、AI分野の基礎理論や手法を学び、研究に応用できる基礎力を養う
  - Python環境でのプログラミングスキル習得とライブラリ活用
  - アルゴリズム実装演習を行い、実例を通じて実践的な課題解決力を向上

応用基礎データサイエンス・AI第一　授業計画

Institute of SCIENCE TOKYO

	授業計画	課題
第1回	データサイエンス・AIの基礎と歴史	データサイエンス・AIの基礎、歴史、役割
第2回	データエンジニアリング基礎	ビッグデータとデータエンジニアリング、データ表現、データベース
第3回	データサイエンス・AI利活用基礎1 （Python言語とライブラリ）	関数、クラス・メソッド、可視化
第4回	データサイエンス・AI利活用基礎2 (Python/pandasの利用)	オープンデータ、データクレンジング
第5回	データサイエンス・AI数学基礎	線形代数、確率、確率分布
第6回	データサイエンス基礎1	各種の離散・連続分布、チェビシェフの不等式、中心極限定理、乱数
第7回	データサイエンス基礎2	基本統計量、相関、単回帰・重回帰分析

応用基礎データサイエンス・AI第二　授業計画

Institute of SCIENCE TOKYO

	授業計画	課題
第1回	数理統計の基礎1	標本と統計的推定、点推定、区間推定
第2回	数理統計の基礎2	仮説検定、母分散・母比率の推定と検定、適合度と独立性の検定
第3回	機械学習1	教師なし学習：クラスタリング、主成分分析
第4回	機械学習2	教師あり学習：各種の回帰、過学習と正則化、分類、性能評価
第5回	ニューラルネットワークと深層学習1	パーセプトロン、確率的勾配降下法、誤差逆伝播法
第6回	ニューラルネットワークと深層学習2	畳み込みNN、リカレントNN
第7回	機械学習・深層学習・AI応用	強化学習、生成モデル(VAE, GAN, 拡散モデル)、注意機構とトランスフォーマ

Institute of SCIENCE TOKYO

ビッグデータとは

- 大量のディジタルデータ
  - 人間の能力では処理しきれない膨大な量のデータ

SNS
YouTube, Twitter, FaceBook, Instagram, Email, TikTok, …

ニュース・ディジタルアーカイブ
テキスト，画像，音声，動画，…

POS (Point of Sales) データ
コンビニやスーパーの購買行動，Eコマース取引履歴情報，…

人流データ
携帯電話位置情報，Wi-Fi接続情報，交通系ICカード情報，…

観測データ
気象観測データ，自然現象観測データ，リモートセンシングデータ，
各種実験データ，…

毎分当たりの流通データ量*
YouTube投稿：500時間分の動画
Twitter共有：347,200ツイート
FaceBook共有：1,700,000件
Instagram共有：66,000枚の写真
Email送信：231,400,000件
日本における情報通信量**
総務省情報通信統計データベース

* Data Never Sleeps, https://www.domo.com/data-never-sleeps#
** https://www.soumu.go.jp/johotsusintokei/index.html

Institute of SCIENCE TOKYO

ビッグデー

- 大量のデ
  - 人間の能
    - SNS
      - YouTu
    - ニュース
      - テキス
    - POS (P
      - コンビ
    - 人流デー
      - 携帯電
    - 観測デー
      - 気象観
      - 各種実

Tokyo Tech, 2024

データの表現：画像（色の表現）

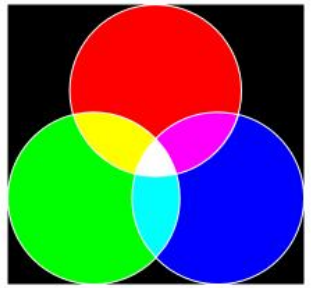
- RGB
  - 三原色：赤 (Red■), 緑 (Green■), 青 (Blue■)
    三原色の組合せで色情報を表現

色の表現としてRGBの他にもHSV*, CMYK**などが使われている

- 色情報の量子化
  - 24ビットカラー
    R/G/B 各成分の明度を256段階 (8ビット符号なし整数)で表し，
    それらの組合せで異なる 256 × 256 × 256 = 16,777,216 色を表現

例）

R:G:B=0:0:0	R:G:B=128:128:128
R:G:B=255:128:0	R:G:B=0:255:128
R:G:B=128:0:255	R:G:B=128:128:255
R:G:B=192:128:64	R:G:B=255:255:255



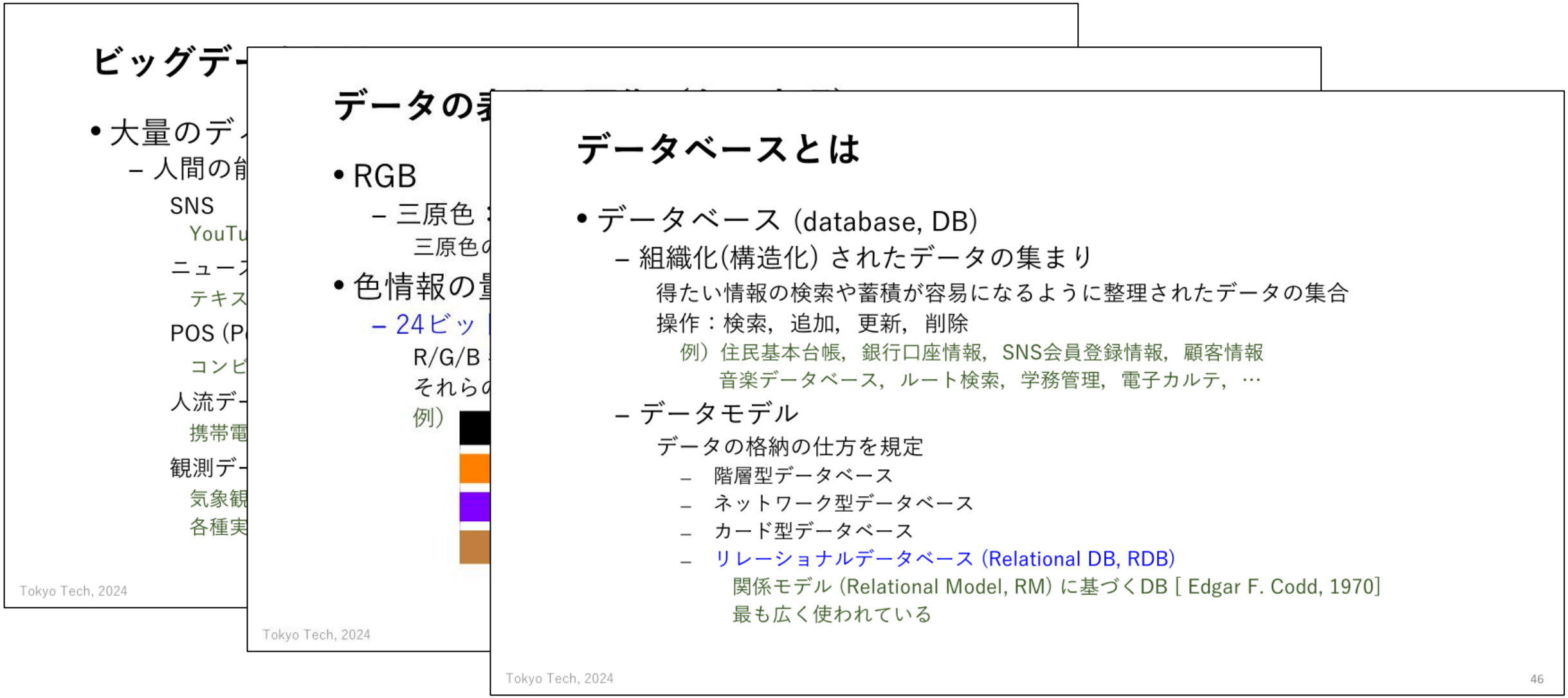
カラー画像

* 色相 (Hue)/彩度 (Saturation)/明度 (Value); ** Cyan/Magenta/Yellow/Key plate

Tokyo Tech, 2024　27

講義資料例：第一　第2回　データエンジニアリング基礎

Institute of SCIENCE TOKYO



データエンジニアリングの基礎

講義資料例：第一　第7回　データサイエンス基礎2

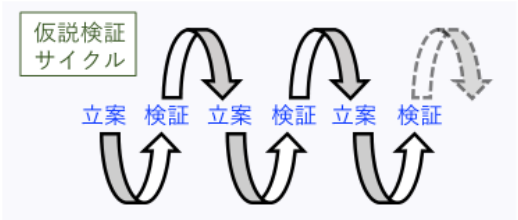
Institute of SCIENCE TOKYO

データ分析と仮説検証サイクル

- データ分析の目的
  - ■ データの利活用
    - データを分析し，有用な情報を抽出
    - 得られた発見・知見を問題解決や付加価値創出に利用
  - – 仮説検証サイクル
    - ■ 仮説の立案とその検証の繰り返し
      1. 仮説の立案（発見）
      2. 仮説の検証
  - – 仮説の発見
    - ■ 調査設計とデータ収集
    - ■ 良い仮説を発見するための手段
      - データ要約・可視化，特徴量抽出，クラスタリング，次元削減，etc.
  - – 仮説の検証
    - ■ 仮説が正しいかどうかを判断するための手段
      - 相関分析，回帰分析，仮説検定，etc.

仮説の発見：例えば統計データやビッグデータの中から有用な傾向や規則性を見出す

・仮説検証の結果，仮説が正しいと言えない場合は，新たな仮説を立て，その検証を行う
・この反復（仮説検証サイクル）を確からしい結論が得られるまで繰り返す
・立てた仮説の検証のために，適切なデータを新たに収集する必要が生じる場合もある



講義資料例：第一　第7回　データサイエンス基礎2

Institute of SCIENCE TOKYO

データ分析

- データ分
  - データ
    - デー
    - 得
  - 仮説検証
    - 仮説の
      1. 仮
      2. 仮
  - 仮説の
    - 調査設
    - 良い仮
      - デー
        クラ
  - 仮説の検
    - 仮説
      - 相関

Tokyo Tech, 2024

量的データの可視化

- 基本的な可視化手法
  - ヒストグラム (度数分布図, histogram)
    - 度数分布から作成する棒/柱状グラフ
      - 適切な階級幅を選択することが重要
  - 円グラフ (pie chart, circle chart)
    - 全体に対する割合を扇型の中心角で表現
  - 箱ひげ図 (box plot)
    - データの分布の代表値を箱とひげ (直線) で表現
      - ヒストグラムの簡易版としての分布比較
  - 散布図 (scatter plot)
    - 量的変数同士の関係の可視化
      - 横軸・縦軸がそれぞれ異なる量的変数の値に相当
      - 各標本のそれぞれの変数の値を2次元平面の1点として表現
  - ヒートマップ (heat map)
    - 二次元データの値の大小を色により表現

Tokyo Tech, 2024　6

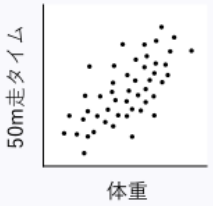
講義資料例：第一　第7回　データサイエンス基礎2

Institute of SCIENCE TOKYO

データ分析の基礎知識：相関関係と因果関係

- 因果関係 (因果性) (causality)
  - 「原因・要因」と「結果」の関係
  - 因果関係が成り立つ条件
    - 影響の単一方向性・時間関係 → 原因が先で結果が後
    - 共変関係 → 一方の変化と連動して他方も変化＝相関関係
    - 原因と結果に共通の要因がない
- 相関関係
  - 相関関係は因果関係を包含しない
    - 相関関係があっても因果関係があるとは限らない
  - 影響の方向性はわからない
    - 二つの事象 X, Y： $X \longrightarrow Y$, $X \longleftarrow Y$, or $X \longleftrightarrow Y$
  - 疑似相関 (spurious correlation/relationship)
    - 二つの事象に共通の要因 Z があり因果関係があるように見える： $Z \longrightarrow X, Z \longrightarrow Y \Longrightarrow X \longrightarrow Y$
  - 相関係数 $\fallingdotseq 0$ の場合
    - 線形関係にはないが無関係とは限らない

疑似相関例：小学生の計測結果 (仮想データ)



明らかに誤った因果関係の解釈：
「体重が重いほど50m走が速くなる」

共通の要因は交絡因子(confounding factor) と呼ばれ
その変数は交絡変数または第三の変数と呼ばれる

Tokyo Tech, 2024　17

Tokyo Tech, 2024

Tokyo Tech, 2024

データ分析の進め方の基礎

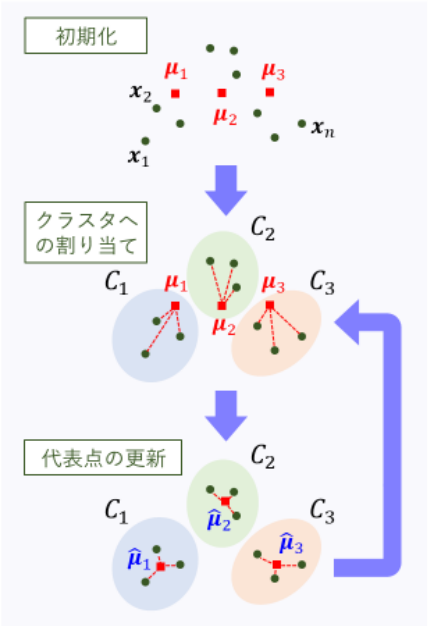
K-means クラスタリング：手法の概要

- 非階層的クラスタリングの代表的手法
  - クラスタリングの問題設定
    - データ：$\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n \in \mathbb{R}^L$
    - データ間の距離：$d(\boldsymbol{x}_i, \boldsymbol{x}_j) \stackrel{\text{def}}{=} \|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2$
    - クラスタ数 K を事前に決めてクラスタリング
  - K-means アルゴリズム
    - 初期化
      - クラスタ $C_1, C_2, \cdots, C_K$ の代表点の初期値を選ぶ

        $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \cdots, \boldsymbol{\mu}_K \in \mathbb{R}^L$
    - 繰り返し
      - 各データを最も近い代表点を持つクラスタに割り当てる

        $C_k = \{\boldsymbol{x}_i : k = \arg\min_j d(\boldsymbol{x}_i, \boldsymbol{\mu}_j)\},\ i = 1, \cdots, n,\ j = 1, \cdots, K$
      - クラスタの代表点を更新する

        $\boldsymbol{\mu}_k = \arg\min_{\boldsymbol{y}} \sum_{\boldsymbol{x}_i \in C_k} d(\boldsymbol{x}_i, \boldsymbol{y}),\quad k = 1, \cdots, K$
      - 代表点が変化しなくなったら繰り返しを終了



Institute of SCIENCE TOKYO

K-means

- 非階層的
  - クラス
    - デー
    - デー
    - クラ
  - K-mean
    - 初期化
      - クラ
      - $\boldsymbol{\mu}_1, \boldsymbol{\mu}$
    - 繰り返
      - 各デ
      - $C_k =$
      - クラ
      - $\boldsymbol{\mu}_k =$
      - 代表

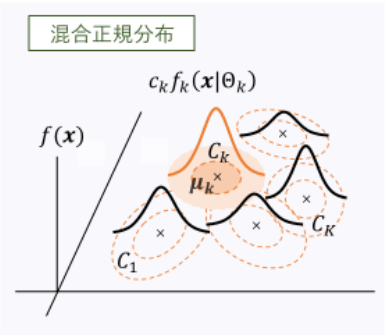
Tokyo Tech, 2024

混合ガウスモデル：モデルパラメータ

- 混合正規分布
  - 確率密度関数
    - k 番目の混合成分の正規分布 $f_k(\boldsymbol{x}|\Theta_k)$

      $f_k(\boldsymbol{x}|\Theta_k) \stackrel{\text{def}}{=} N(\boldsymbol{x}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k), \quad \Theta_k = \{\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}$
    - 混合重み c_k，$c_1 + c_2 + \cdots + c_K = 1$

      $f(\boldsymbol{x}) = \sum_{k=1}^{K} c_k N(\boldsymbol{x}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) = \sum_{k=1}^{K} c_k f_k(\boldsymbol{x}|\Theta_k)$

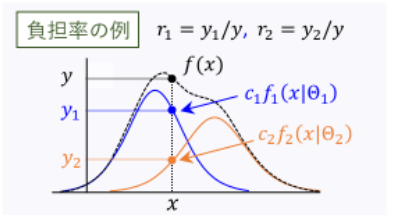


  - 負担率 (responsibility)
    - データ $\boldsymbol{x}$ が観測されたとき，$\boldsymbol{x}$ が k 番目の混合成分から生成された確率 (事後確率)

      $r_k = \mathrm{P_r}(C_k|\boldsymbol{x}) = \dfrac{\mathrm{P_r}(C_k)\mathrm{P_r}(\boldsymbol{x}|C_k)}{\mathrm{P_r}(\boldsymbol{x})} = \dfrac{c_k f_k(\boldsymbol{x}|\Theta_k)}{f(\boldsymbol{x})} = \dfrac{c_k f_k(\boldsymbol{x}|\Theta_k)}{\sum_{k=1}^{K} c_k f_k(\boldsymbol{x}|\Theta_k)}$

    ← ベイズの定理より
  - GMMのモデルパラメータ
    - 混合正規分布のパラメータ
      - 混合重み，正規分布の平均と分散共分散行列の組の集合

        $\boldsymbol{\Theta} \stackrel{\text{def}}{=} \{\{c_k\}, \{\Theta_k\}\} = \{c_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}$



Tokyo Tech, 2024　20

講義資料例：第二　第3回　教師なし学習

Institute of SCIENCE TOKYO

K-means

- 非階層的
  - クラ
    - デー
    - デー
    - クラ
  - K-mean
    - 初期化
      - クラ
    - 繰り返
      - 各デ
      - クラ
      - 代表

Tokyo Tech, 2024

混合ガウス

- 混合正規
  - 確率密
    - k 番
    - 混合
  - 負担率
    - デー
  - GMMの
    - 混合正
      - 混合

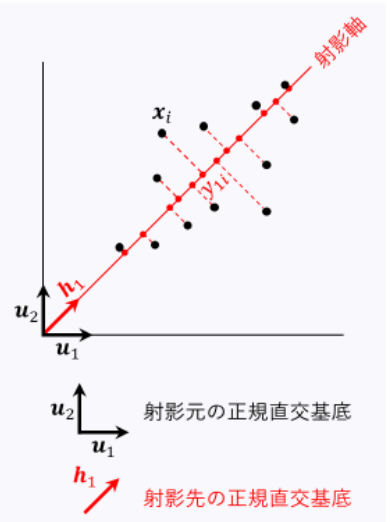
Tokyo Tech, 2024

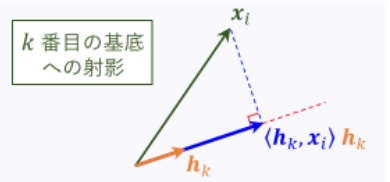
主成分分析の考え方

- データの射影
  - 射影元
    - L 次元データ：$\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n \in \mathbb{R}^L$
    - 正規直交基底：$\boldsymbol{u}_1, \boldsymbol{u}_2, \cdots, \boldsymbol{u}_L \in \mathbb{R}^L,\ \langle \boldsymbol{u}_k, \boldsymbol{u}_l \rangle = \begin{cases} 1, & k = l \\ 0, & k \neq l \end{cases}$
  - 射影先
    - $M(\leq L)$ 次元データ：$\boldsymbol{y}_1, \boldsymbol{y}_2, \cdots, \boldsymbol{y}_n \in \mathbb{R}^M$
    - 正規直交基底：$\boldsymbol{h}_1, \boldsymbol{h}_2, \cdots, \boldsymbol{h}_M \in \mathbb{R}^L,\ \langle \boldsymbol{h}_k, \boldsymbol{h}_l \rangle = \begin{cases} 1, & k = l \\ 0, & k \neq l \end{cases}$
  - 射影データ

$$\boldsymbol{y}_i = [\boldsymbol{h}_1, \boldsymbol{h}_2, \cdots, \boldsymbol{h}_M]^{\mathrm{T}} \boldsymbol{x}_i = \begin{bmatrix} \boldsymbol{h}_1^{\mathrm{T}} \boldsymbol{x}_i \\ \boldsymbol{h}_2^{\mathrm{T}} \boldsymbol{x}_i \\ \vdots \\ \boldsymbol{h}_M^{\mathrm{T}} \boldsymbol{x}_i \end{bmatrix} = \begin{bmatrix} y_{1i} \\ y_{2i} \\ \vdots \\ y_{Mi} \end{bmatrix}$$

$$y_{ki} = \langle \boldsymbol{h}_k, \boldsymbol{x}_i \rangle = \boldsymbol{h}_k^{\mathrm{T}} \boldsymbol{x}_i$$





Tokyo Tech, 2024　33

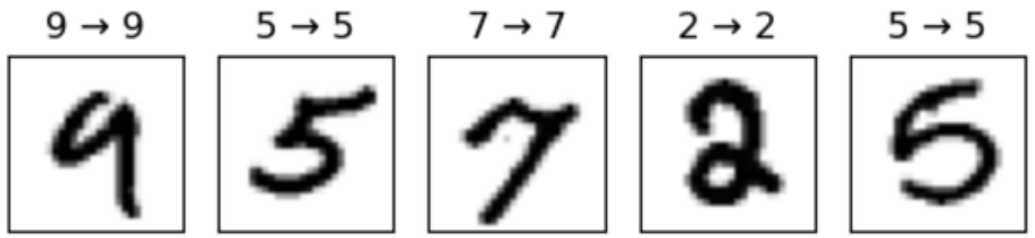
各手法の理論的な背景まで

演習資料例：第二　第5・6回　ニューラルネットワーク・深層学習

テストデータの中からランダムに入力画像を5サンプル選び、その正解ラベルと認識結果を表示した結果を以下に示す。

```
import random

fig, axs = plt.subplots(1,5,figsize=(6,2))
idx = random.sample(range(X_test.shape[0]),5)
for i in range(5):
    image, label = X_test[idx[i]].reshape(28,28), np.argmax(y_test[idx[i]])
    result = np.argmax(z[idx[i]])
    # Test画像と認識結果表示
    axs[i].imshow(image, cmap='gray_r')
    axs[i].set_title(f'{label} → {result}')
    axs[i].get_xaxis().set_visible(False)
    axs[i].get_yaxis().set_visible(False)
plt.show()
```



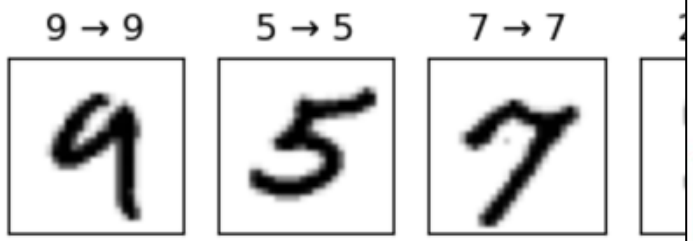
演習資料例：第二　第5・6回　ニューラルネットワーク・深層学習

Institute of SCIENCE TOKYO

テストデータの中からランダムに入力画像を5サンプル選び、その正解ラベルと認識結果を表示した結果を以下に示す。

```
import random

fig, axs = plt.subplots(1,5,figsize=(6,2))
idx = random.sample(range(X_test.shape[0]),5
for i in range(5):
    image, label = X_test[idx[i]].reshape(28
    result = np.argmax(z[idx[i]])
    # Test画像と認識結果表示
    axs[i].imshow(image, cmap='gray_r')
    axs[i].set_title(f'{label} → {result}')
    axs[i].get_xaxis().set_visible(False)
    axs[i].get_yaxis().set_visible(False)
plt.show()
```




```
import random
import numpy as np
import matplotlib.pyplot as plt

# テストデータからランダムに5個選んで画像とカテゴリを表示
classes = ('airplane','automobile','bird','cat','deer','dog','frog','horse','ship','truck')
fig, axs = plt.subplots(1,5,figsize=(6,2))
idx = random.sample(range(data_test.data.shape[0]),5)
for i in range(5):
    img, lbl = data_test[idx[i]]
    img_np = img.numpy().copy()          # テンソルからndarrayへ変換
    img_np = np.transpose(img,(1,2,0)) # [ch,row,col] -> [row,col,ch]
    imgc = (img_np + 1) / 2              # 画素値の範囲を[0,1]に戻す
    axs[i].imshow(imgc)
    axs[i].set_title(f'{classes[lbl]}')
    axs[i].axis('off')
plt.show()
```

Python notebookを用いて講義で説明した内容の演習

講義フロー・イメージ

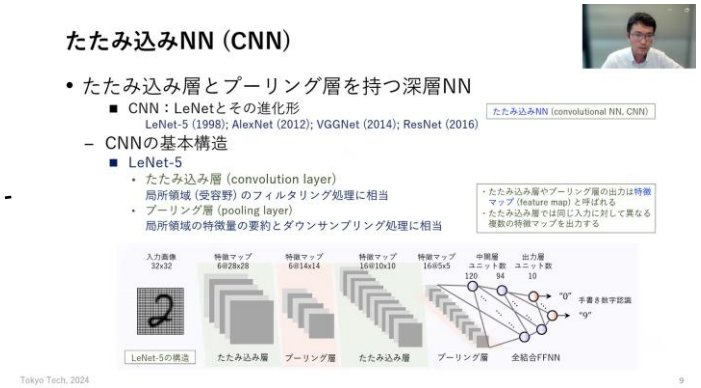
- zoom リアルタイム配信かつ演習室を用意

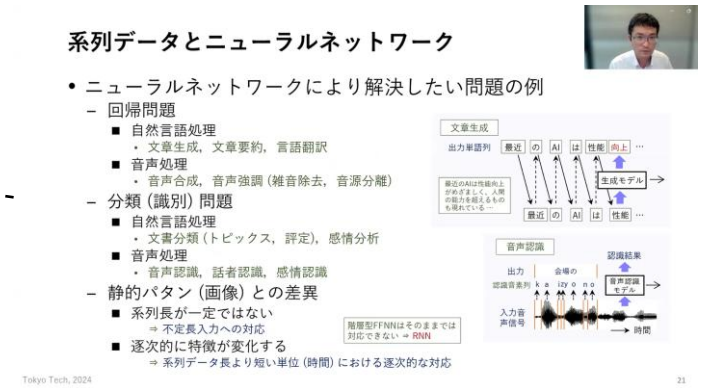
講義 (30分)
畳み込みNN

演習(50分)
畳み込みNN

講義 (20分)
リカレントNN

TAによるサポート

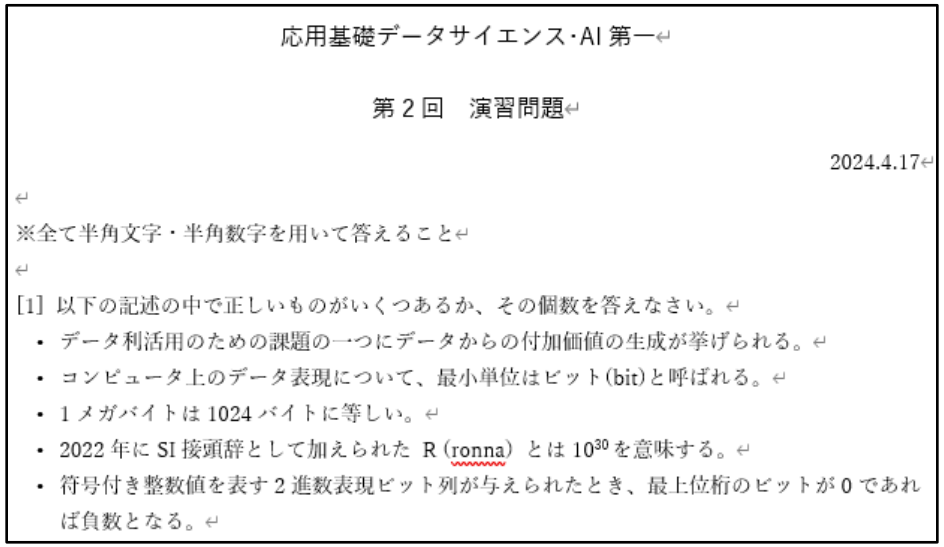




- 講義動画を後日受講生に配布

評価方法

- 講義後に学内の学修管理システムを利用し、各回で多肢選択問題の小テストを実施
- または Pythonを用いた計算問題を回答しJupyter notebookファイルを提出
  - 上記2種の課題については自動採点を利用
- 期末課題：講義で配布したJupyter NotebookのPythonサンプルプログラムを参考に、独自に設定した処理を行うPythonプログラムを作成し、実行結果を提出



応用基礎データサイエンス・AI 第一

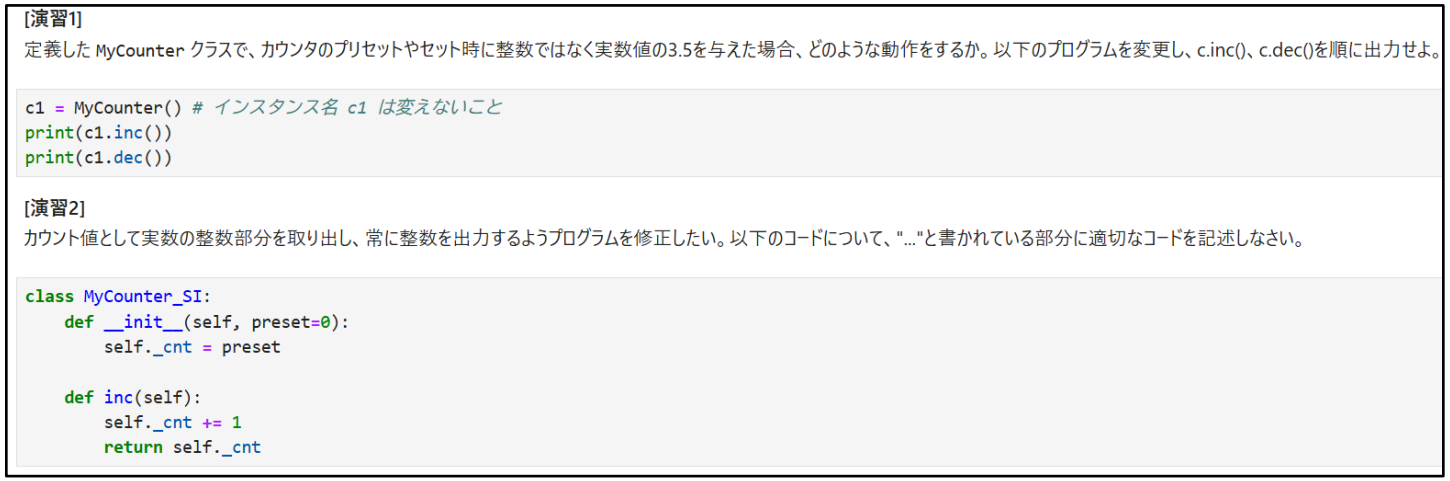
第 2 回　演習問題

2024.4.17

※全て半角文字・半角数字を用いて答えること

[1] 以下の記述の中で正しいものがいくつあるか、その個数を答えなさい。

- データ利活用のための課題の一つにデータからの付加価値の生成が挙げられる。
- コンピュータ上のデータ表現について、最小単位はビット(bit)と呼ばれる。
- 1 メガバイトは 1024 バイトに等しい。
- 2022 年に SI 接頭辞として加えられた R (ronna) とは 10^{30} を意味する。
- 符号付き整数値を表す 2 進数表現ビット列が与えられたとき、最上位桁のビットが 0 であれば負数となる。



[演習1]

定義した MyCounter クラスで、カウンタのプリセットやセット時に整数ではなく実数値の3.5を与えた場合、どのような動作をするか。以下のプログラムを変更し、c.inc()、c.dec()を順に出力せよ。

```
c1 = MyCounter() # インスタンス名 c1 は変えないこと
print(c1.inc())
print(c1.dec())
```

[演習2]

カウント値として実数の整数部分を取り出し、常に整数を出力するようプログラムを修正したい。以下のコードについて、"..."と書かれている部分に適切なコードを記述しなさい。

```
class MyCounter_SI:
    def __init__(self, preset=0):
        self._cnt = preset

    def inc(self):
        self._cnt += 1
        return self._cnt
```

講義実施の所感

- 講義資料と演習資料ともに詳しい解説
- 受講生の予習や復習に適している
- 小テストと演習問題について自動採点を活用し、授業担当者の負担を軽減可能