特集　AI時代の人材育成

数理・データサイエンス教育強化拠点コンソーシアム カリキュラム分科会の取組み －データサイエンス教育の普及に向けて－

東京大学
数理情報教育研究センター　丸山　祐造

1．全大学生がデータサイエンスを学ぶ意義

データサイエンスは産業や科学に大きな影響を及ぼす分野として注目されています。産業界や学術界における研究開発や様々な業務は今後ますますデータ駆動型になります。いわゆる「データサイエンティスト」は、複雑かつ大量のデータを分析することに専念する知識労働者のことです。しかし、データサイエンスはそのような分析に携わる専門家のみに関係する分野ではありません。

web検索において検索語と同時にいくつかの関連語彙がサジェストされたり、Amazonなどのサイトで特定の顧客の嗜好に合わせて商品が推薦されることは、データサイエンスの手法に基づいています。つまり我々はデータサイエンスの成果と隣り合わせで日常を過ごしています。もちろん、データサイエンスを応用し多種多様で大量のデータを分析した成果は、社会に多大なる便益をもたらします。しかし、その一方で、深刻な倫理的な問題を引き起こすことが指摘されてます。

例えば、

- 匿名化したデータベースにおいて、外部データベースとの照合などの操作により、個人が再識別されうること
- 個人が再識別されない場合でも、データ分析の結果の解釈によって、年齢・民族・性などで層別された特定のグループについて、秘密が曝露されたり、差別が引き起こされる場合があること

があげられます。それらが問題であると認識するためには、データサイエンスの基礎を理解する必要があります。

また、データのグラフ（棒グラフ、折れ線グラフ、円グラフ等）により、言葉では伝えるのが難しいデータの重要な特徴を、比較的容易に他者に伝えることができます。そのために、グラフは学術論文や技術文書だけではなく、テレビを始めとするマスメディアで頻繁に登場します。ただし、発信元が予め想定した主張をサポートするために、しばしばバイアスを持って提示されます。その代表例が、縦軸をゼロから始めずに長さを切り詰めた棒グラフ、中心をずらしたり楕円形にゆがめた円グラフ、本来２次元であるグラフを３次元にして、発信元の主張をサポートするように調整された棒グラフや円グラフです。そのようなグラフに騙されないためのリテラシーは、安全・安心に社会生活を送るための基礎であると考えられます。

このように、データサイエンティストのように日常的にデータを分析するような職業につかない場合でも、生じうる倫理的問題やバイアスを持った可視化に代表されるデータサイエンスの負の側面を認識することは、ビッグデータ時代を生きる人間として最低限のデータリテラシーであると考えられます。もちろん冒頭に述べたように、産業界や学術界における研究開発や様々な業務は今後ますますデータ駆動型になるため、データサイエンスを様々なレベルで理解している人材の需要は高まっています。つまり、大学の卒業生に期待されるスキルが変わってきており、高等教育を受けるすべての大学生がデータサイエンスの基礎を学ぶべき時代が到来したと考えられます。さらに、基礎を学んだ学生の中から、新しい時代のデータサイエンスを先導するような研究者や、高度な専

門知識を持って業務にあたる専門家が出てくることが期待されます。そのような裾野の広いデータサイエンス教育が始まろうとしています。

本稿の構成は以下の通りです。２．では、数理・データサイエンス教育強化拠点コンソーシアムの概要を説明します。３．から６．では公開に向けて作業中の標準カリキュラム案について簡単に紹介します。

２．数理・データサイエンス教育強化拠点コンソーシアム

2016年12月、北海道大学、東京大学、滋賀大学、京都大学、大阪大学、九州大学は数理及びデータサイエンス教育の強化に関する懇談会における評価結果を踏まえ、文部科学省より数理及びデータサイエンスに係る教育強化の拠点校として選定されました。拠点校６校は、各大学内での数理・データサイエンス教育の充実に努めるだけでなく、全国の大学に取組み成果の波及を図るため、地域や分野における拠点として他大学の数理・データサイエンス教育の強化に貢献することが期待されています。

そのために本学を幹事校としてコンソーシアムを形成して、以下を含む様々な取組みを行うことになっています。

- 全国的なモデルとなる標準カリキュラム・教材の作成
- その標準カリキュラム・教材の他大学への普及方策（例えば全国的なシンポジウムの開催等）の検討及び実施

３．カリキュラム分科会の取組み

コンソーシアムの活動を効率的かつ機動的に進めるために、３つの分科会（カリキュラム分科会、教材分科会、教育用データベース分科会）が設けられました。筆者が主査を務めるカリキュラム分科会の最大のミッションは全国的なモデルとなる標準カリキュラムの作成・普及です。2018年４月の分科会の活動開始に当たり、データサイエンス教育で先行するアメリカの取組みを調査しました。都合の良いことに日本の学術会議に相当するThe National Academies of Sciences, Engineering, and Medicineでプロジェクト Data Science for Undergraduatesが立ち上がっており、その中間報告書が公開されていました。その直後の2018年５月には最終報告書が公開されました。プロジェクトの委員会には、大学関係者だけでなくIBMやMicrosoftなど産業界からも委員が加わり、アメリカの産学によるインパクトのある報告書です。

報告書においてはデータサイエンスの学部教育について様々な提言がされています。その中の一つが、データに基づいた妥当な判断を行う能力を学部教育において養うために、以下の10分野が重要というものです。

① Mathematical foundations（数学基礎）
② Computational foundations（計算基礎）
③ Statistical foundations（統計基礎）
④ Data management and curation（データ管理とキュレーション）
⑤ Data description and visualization（データ記述と可視化）
⑥ Data modeling and assessment（データモデリングと評価）
⑦ Workflow and reproducibility（ワークフローと再現性）
⑧ Communication and teamwork（コミュニケーションとチームワーク）
⑨ Domain-specific considerations（ドメイン知識の考慮）
⑩ Ethical problem solving（倫理的な問題への対応）

カリキュラム分科会では、この10分野を参考に⑦、⑧、⑨を除く７分野について学修目標とスキルセットを整備することにしました。学修目標とは「全大学生への数理データサイエンス教育の普及・展開に向けて、リテラシーとして修得すべき内容を文章でまとめたもの」であり、スキルセットとは「データサイエンスのスキルを初級（学修目標・リテラシーレベル）から上級のレベル別に整理したもの」です。なお、本稿は７月下旬におけるベータ版について記述します。専門家のレビューを経て９月に公開される正式版とはいくつかの相違点が予想されます。

なお、３分野を除いた理由は以下の通りです。「⑦ワークフローと再現性」は、作業手順がスクリプトに明記されるRやPython等の利用による再現性確保のスキルなどを想定していますが、リテラシーレベルを超えると判断しました。「⑧コミ

ュニケーションとチームワーク」の重要性は、数理データサイエンス教育に限定されるわけではありません。「⑨ドメイン知識の考慮」は、データサイエンスの学際性に関連しています。ドメイン知識（領域知識）は重要ですが、本プロジェクトでは、各ドメインで共通に役に立つような数理データサイエンスのスキルを整理することに注力することとします。

学修目標とスキルセットでは共通の階層構造を設けています。上記の分野を大分類として、その下に中分類、小分類を設けます。例えば、

大分類　数学基礎
中分類　線形代数
小分類　ベクトル

という具合です。高大接続を強く意識しており、高校の数学や情報の学習内容を積極的に含めています。「小分類 ベクトル」には高校「数学B」の内容である二次元ベクトルが含まれます。

4．スキルセット

前述の通り、スキルセットとはデータサイエンスのスキルを初級(学修目標・リテラシーレベル)から上級のレベル別に表形式に整理したものであり、上級に向けて以下の①、①'、②、③の４つのレベルに分類しています。

① （学修目標コアレベル）専門を問わず、すべての大学生が教養課程あるいは専門課程で学ぶレベル
①'（学修目標レベル）専門を問わず、すべての大学生が教養課程あるいは専門課程で学ぶレベル
② 拠点校、協力校など数理データサイエンス教育を先導する大学の教養レベル
③ 拠点校、協力校など数理データサイエンス教育を先導する大学の専門レベル

ただし、①におけるコアについては「６．コアの設定」を参照して下さい。また、現在のところ②と③は、数学基礎、計算基礎、統計基礎の３分野のみ整備しています。他の分野のリテラシーレベルを超えるスキルセットは来年度に整備予定です。実際のスキルセットは９月に公開される正式版をご覧下さい。

5．学修目標

中分類に対して、その中分類のデータサイエンス全体における意義や重要性を記述し、小分類について文章でスキルを記述しています。ここでは、スペースの都合で中分類、小分類のリストを紹介するにとどめて、最後の「データの法規と倫理」についてのみ中分類の記載事項を紹介します。
(次ページ「表１ 学修目標とスキルセット」を参照して下さい)

6．コアの設定

全国展開においては、学生の特性や文系・理系の違いを考慮する必要があります。その対応として概ね1.5単位分くらいの内容（「表１」で太字で記した８つの中分類）を特にコアとして推奨することにします。データサイエンス入門のような科目を設置する場合には、コアに各大学の事情に応じて内容を追加する形で、２単位や４単位の講義を設計できると考えます。なお、「データの法規と倫理」の中分類にはやや高度な内容も含まれますが、すべてコアとして推奨します。特にデータサイエンスの倫理の重要性は2019年３月に統合イノベーション戦略推進会議で決定された「人間中心のAI社会原則」でも指摘されています。

7．まとめ

本稿ではデータサイエンス教育の普及に向けた数理・データサイエンス教育強化拠点コンソーシアムのカリキュラム分科会の取組みを紹介しました。学部の数理・データサイエンス教育の重要性は政府の「AI戦略2019」でも指摘されており、大学教育の中でも特に注目されている分野です。2019年度からコンソーシアムは国立大学の中から選定された20の協力校とタッグを組んで、データサイエンス教育の普及を目指します。地域ブロックごとの拠点大学及び協力校の連携による主要な取組みとしてブロックワークショップがあります。ワークショップを通じて、私立大学を含むすべての大学にアプローチして、データサイエンス教育の全国展開を進めていくことになっています。多くの皆様に関心を持って頂ければ幸いです。

表１　学修目標とスキルセット

大分類	中分類	小分類
数学基礎	線形代数	ベクトル、行列
	微積分	初等関数、１変数関数の積分法、１変数関数の微分法、２変数関数の微積分
	線形代数と微積分の計算機演習	線形代数の演習、微積分の演習
	数列	数列
計算基礎	情報、デジタル	数と表現、デジタル化、情報量の単位、文字の表現
	コンピュータの仕組み	集合と命題、論理演算、計算誤差、有効数字
	データ構造	配列・リスト、項目・値形式のデータ
	アルゴリズムとプログラミング	アルゴリズム、プログラミング
統計基礎	**確率と確率分布**	確率、場合の数、順列・組み合わせ、確率分布の概念、主要な確率分布
	データ収集法と確率構造	標本調査、ランダム化比較試験
	統計的推測	統計的モデル、点推定・区間推定、仮説検定
データ管理とキュレーション	**データ取得とオープンデータ**	日本や世界のオープンデータ、オープンデータの取得
	データ管理とデータ形式	代表的なデータ形式、その他のデータ形式、データベース
	データの前処理	データクレンジング　外れ値、異常値、欠損値、データ加工
データ記述と可視化	**データの記述**	種々のデータ、基本統計量、相関関係
	データの可視化	グラフの構成要素、統計グラフ（棒グラフ、折れ線グラフ、円グラフ、帯グラフ）、統計グラフ（チャートジャンク）、分布の統計グラフ（ヒストグラム、箱ひげ図）、散布図
データモデリングと評価	教師あり学習	回帰分析　ロジスティック回帰
	教師なし学習	クラスタリング（k-平均法）階層クラスタリング
	モデルの評価	評価指標　訓練データとテストデータ
データの法規と倫理	**情報倫理、情報セキュリティ**（※１）	情報倫理・関連法規、情報セキュリティ
	データに関連する法律・規制（※２）	個人のデータに関連する法規、統計法
	データサイエンスの倫理（※３）	倫理に配慮したデータ収集と利活用、データの匿名化、データサイエンスに関する様々なバイアス、逸脱事例

（※１）　データサイエンスを学ぶ際に、数学や統計などの理論を身につけるだけではなく、コンピュータで実データを分析する能力やその際の注意点を習得することが重要である。また、ネットワークを介したグループでの共同作業により、データサイエンスの課題解決を導くことも想定される。ネットワークに繋がったコンピュータを扱うにあたり情報倫理の基礎を学び、情報セキュリティへの理解を深める。

（※２）　（改正）個人情報保護法と統計法はデータに関する基本的な法規である。前者について、高度情報化社会において個人の情報・利益の保護と個人情報の有用性のバランスを取ることの重要性を学ぶ。後者について、公的統計が行政利用だけではなく、社会全体で利用される情報基盤として極めて重要であり、根拠に基づく（データに基づく）政策立案の基礎であることを学ぶ。

（※３）　データサイエンスを応用し多種多様で大量のデータを分析した成果は、社会に多大なる便益をもたらす。その一方で、深刻な倫理的な問題を引き起こすことが指摘されている。例えば、個人情報の曝露や年齢・民族・性などで層別された特定のグループやマイノリティへの差別である。ここでは特に、倫理に配慮したデータ収集や利活用、匿名化されたデータであっても個人情報を曝露するリスクがあること、またデータサイエンスに関連する様々な偏り（バイアス）について具体例を通じて理解を深める。