文系単科大学におけるデータサイエンス教授法試論

綿貫真也*1
Email: Shinya_Watanuki@red.umds.ac.jp

*1：流通科学大学商学部マーケティング学科

©Key Words　文系学部，データ・サイエンス，ドメイン知識，潜在データ・サイエンス抵抗意識

1. はじめに

近年、ビックデータを AI（Artificial Intelligence：人工知能）・機械学習を活用し、ビジネスを展開する巨大プラットフォーマー（GAFA: Google/Apple/Facebook/Amazon, BAT: Baido/Alibaba/Tencent）の活躍が注目されている。こうした企業を支え、第４次産業革命と言われる現代社会において新たに生まれた職種がデータ・サイエンティストである。データ・サイエンティストという職種は、2010 年代初頭から出現し始め、「２１世紀でもっともセクシーな仕事」（HBR 2012）と称され脚光を浴びてきた。その役割は、従来の研究開発部門や研究所の研究者とは異なり、実務的な側面が極めて強い。データ・サイエンティストの主な仕事内容は、ビジネス課題の解決のために、AI（Artificial Intelligence：人工知能）・機械学習の手法を用いて、データを活用し、収益化に寄与することであるが、数理統計学、情報科学に関する非常に高度な専門知識が要求されることから、職務の遂行には、当該領域に関連する分野の博士号（Ph.D.）が必須であるとされた。しかし、100 を超える統計学などデータ・サイエンス関連の専門学部、学科、専門プログラムが教育機関によって提供されているアメリカ合衆国と比較して、我が国では、教育機関における、そうした人材育成・供給体制が整っているとは言い難い。

こうしたわが国の状況において、政府は、2019 年 3 月に AI 戦略の具体案（「ＡＩ戦略（有識者提案）及び人間中心のＡＩ社会原則（案）について」）を提示した。その中の「AI 人材の教育と育成」において、「数理・AI・データサイエンス」を２１世紀の「読み・書き・そろばん」と位置づけ、AI 人材の育成を３段階に分類し、具体的な数値目標を掲げている。特に、リテラシーレベルにおいては、文理問わず、大学・高専において、年間 50 万人（１学年）に展開することを数値目標として掲げ、さらに社会人リカレント教育の一環としても展開することを提唱している。その後、2020 年 3 月に、数理・データサイエンス教育強化拠点コンソーシアムから、リテラシーレベルの参考カリキュラムが提案された。その仔細に関しては、コンソーシアムの資料「数理・データサイエンス・ＡＩ教育プログラム認定制度（リテラシーレベル）の創設について」に譲るが、その概要は、データ・サイエンスの理論的な側面に関して軽視はしないものの、「実データを用いて、データ・サイエンスの活用により身近な実務・実社会的な課題が解決されることを学ぶ」というプラクティカルな内容と捉えることができる。

しかし、データ・サイエンス教育を文系学部において実施することは容易ではない。特に、教育困難大学の文系学部における実施は、さらに困難を極めることが予測される。本稿では、筆者担当のマーケティングデータ分析の講義において、そうした困難を克服することを目的として、企画設計し、実施されたカリキュラムに関する紹介とその効果検証について報告を行う。

2. 想定される困難とカリキュラムに与える影響について

文系学部において、データ・サイエンス教育を実施する際には、様々な困難が想定されることは上述したが、本節では、その具体的な障壁とその影響について述べる。障壁には、顕在化されている障壁と顕在化された障壁から発生する障壁がある。これらを、本稿では、前者を顕在的障壁、後者を発生障壁と呼ぶ。こうした障壁は、カリキュラム設計に影響を与える。以下に、具体的な障壁とそのカリキュラム設計への影響について詳述する。

まず、顕在的障壁として、主に次の３つが存在する。1）パソコンスキル 2）理数系基礎学力 3）関与度である。1）のパソコンスキルに関しては、教育困難大学文系学部に限ったことではなく、大学生全般に及ぶ障壁であり、昨今の大学生のパソコンスキルの低さに関しては、具体的な報告が、木村・近藤（2018）よりなされており、原因についても検討されている(1)。パソコンスキルに関する障壁は、カリキュラム設計にあたって、使用するデータ・サイエンスアプリケーションの選定に影響を与える。つまり、複雑な環境設定を必要とする CUI（Character-based User Interface）によるプログラムが必要なアプリケーションの採用は難しいということを意味する。2）の理数系基礎学力に関しては、教育困難ではない大学の文系学部においても障壁が存在するが、教育困難大学においては、さらに障壁が存在する。データ・サイエンス技術の理論的背景である線形代数や基礎解析などの基礎知識を前提とすることが難しいことに加えて、中学校レベルの初等数学の理解を前提とすることも難しい場合が多い。しかし、こうした理数系基礎学力に関する障壁は、社会人リカレント教育においても想定される障壁でもあると考えられ、何かしらの対応と工夫が必要である。つまり、データ・サイエンスに関する基礎的な数学的背景を理解してから、応用をしていくことが理想的ではあるが、限られた 15 回の講義では、そうした理論的背景の理解を目的としたカリキュラム設計は難しいことを意味する。3）の関与度に関しては、文系学部ということから、データ・サイエンス、AI などは理系分野であり、自分には関係ないというステレオタイプに起因する障壁である。この障壁は、文系学部の

表 1：データ・サイエンスアプケーションの選定

	SPSS	R/Python	Google AutoML Tables	MicroSoft Azure	H2O flow	DataRobot	RobotDiver (マクロミル)	Prediction One (ソニー)
導入企業の汎用性	×	△	○	○	△	△	△	△
自宅での操作	×	○	○	○	○	○	○	○
高度な専門知識が必要ない	×	×	○	○	△	○	○	○
日本語操作	○	○	×	△	×	○	○	○
プログラム必要無し	○	×	△	○	△	○	○	○
AUTO ML機能	×	△	○	○	○	○	○	○
主要DSアルゴの搭載	×	△	○	○	○	○	○	△

主な就職先が、銀行、商社、小売などのサービス業を主とする、いわゆる文系就職先であることや、製造業への就職においても、そのほとんどが営業部門への配属ということにも関連すると考えられる。このことは、カリキュラム設計にあたっては、使用するデータやテーマ設定など学修コンテンツに影響を与える。つまり、将来的に自身が就く仕事（社会人リカレント教育の際には、現在のご自身のお仕事、業界）において、データ・サイエンス、AI は無縁のことではなく、役に立ち、必須であることをコンテンツに盛り込み、レリバンス（関連性）の構築が必要であることを意味する。

3. カリキュラム設計について

3.1 カリキュラム方針

前述の考察から、本カリキュラムの方針を以下のように決定をした。1）ドメイン知識の積極的利用、2）手段としてのデータ・サイエンス、AI 技術の位置づけを明確化、3）実データの使用、である。つまり、商学部マーケティング学科学生のドメイン知識であるマーケティングに関する課題、話題を学修コンテンツの柱として、具体的なマーケティング戦略課題をデータ・サイエンスの手法により分析し、解決をしていくという方針である。そして、データ・サイエンスの専門用語の多用は、当該学生の学習への関与度を低めるので、極力使用しないように努めた。

3.2 使用アプリケーションの選定

7つの使用アプリケーションに関する選定基準から8つの検討アプリケーションを評価した（表1）。その結果、クラウド型データ・サイエンスアプリケーションの RobotDiver（マクロミル）、オンプレ型の PredictionOne（ソニー）を選定した。2つのアプリケーションを選定したのは、複数のタイプのアプリケーション操作に触れることで、多くのデータ・サイエンスアプリケーションが提供されている中で、その操作方法は、いずれのアプリケーションを使用したとしても、出来ることや操作方法がそれほど異ならないことを体感してもらうためである。

3.3 カリキュラム構成とバックキャスティング型コンテンツ設計

具体的なカリキュラム構成を表2に示す。4つの学修フェーズを設け、それぞれに、実施目的と期待する効果を事前に明確にすることで、バックキャスティング型のコンテンツ設計を行った。つまり、理解度のチェックを目的とした積み上げ型により、学修コンテンツを決めるのではなく、最終的に狙う効果を明確にし、その効果を最大化させることに適した学修コンテンツを作り上げていく方法である。この効果とは、学修の理解度などとは別に、カリキュラムにより得られる心理的な効果を意味する。

表 2：カリキュラム構成

学修フェーズ	講義タイトル	内容	目的	狙っている効果
1.導入	1. ガイダンス＋これからのマーケティング情報環境を知る	活用事例解説とニュース映像の視聴	DSが実務・社会課題の解決において身近であることの理解	親近感と関与度の向上（レリバンス）
2.基礎	2. ビジネス・データサイエンスの作法と手法 3. 戦略課題と取り扱うデータについて 4. 問題の理解と分析結果の見通し１ 5. 問題の理解と分析結果の見通し2	マーケティング戦略と当該戦略で活用されるアルゴリズムの解説	マーケティング戦略とデータサイエンス手法の関連性の理解	
3.実習	6. 「当てる」アルゴリズムのオペレーション１ 7. 「当てる」アルゴリズムのオペレーション２ 8. 「当てる」アルゴリズムの戦略的分析（Strategic Analysis） 9. 「当てる」アルゴリズムの分析とトレーニング 10. 学習済「当てる」アルゴリズムによる予測１ 11. 学習済「当てる」アルゴリズムによる予測２ 12. アルゴリズムの連携 13. 当てる」アルゴリズム「プライシング」 14. 別のDSアプリで、「当てる」アルゴリズムで解析	実データとアプリケーションによる実習	慣れ	拒否感の低減
4.まとめ	15. まとめと最終レポート	レポート	習熟度チェック	動機付け

3.4 具体的な学修コンテンツについて

実施したカリキュラムの具体的なコンテンツについては、紙幅の都合で詳述することは難しいが、学修フェーズ2「基礎」のマーケティング戦略とデータ・サイエンス手法の理解を目的としたコンテンツについて、特徴的な箇所についてのみ解説を行う。通常、統計学、機械学習などの学修コンテンツの実施は、平均、分散、ヒストグラムなどの記述統計や取り扱う尺度の説明から入ることが多いが、本カリキュラムでは、最初に、具体的なマーケティング戦略を示し、その戦略の具体的な課題の解説を行い、当該戦略課題の解決をデータ・サイエンス手法が担うと位置付け、コンテンツを作成した（図1）。さらに、データ・サイエンス手法を、「教師あり学習モデル」「教師なし学習モデル」などのデータ・サイエンスの専門用語ではなく、

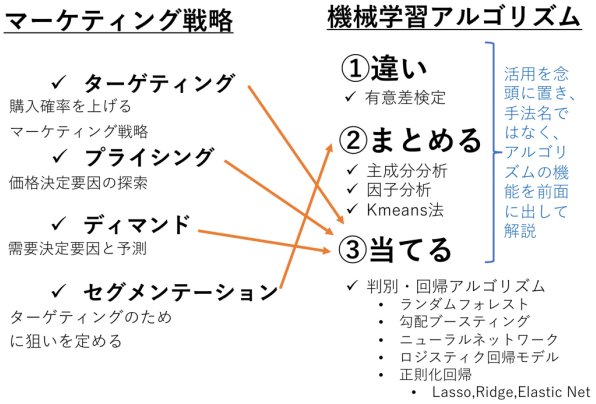


図 1：マーケティング戦略とアルゴリズムの対応

「1.違い」「2.まとめる」「3.当てる」というように言い換えることで、拒否反応の低減と直感的な理解を狙った。例えば、ターゲティング戦略に関しては、「ブランドスイッチ」「顧客行動予測」「ロイヤルカスタマー判定」などの具体的な戦略課題をテーマとして取り上げて、このターゲティング戦略課題の解決には、「当てる」アルゴリズムの適用が有効であることを解説した（図2右）。また、課題の問題構造を図示することで、アルゴリズム適用の直感的な理解を狙った（図2左）。

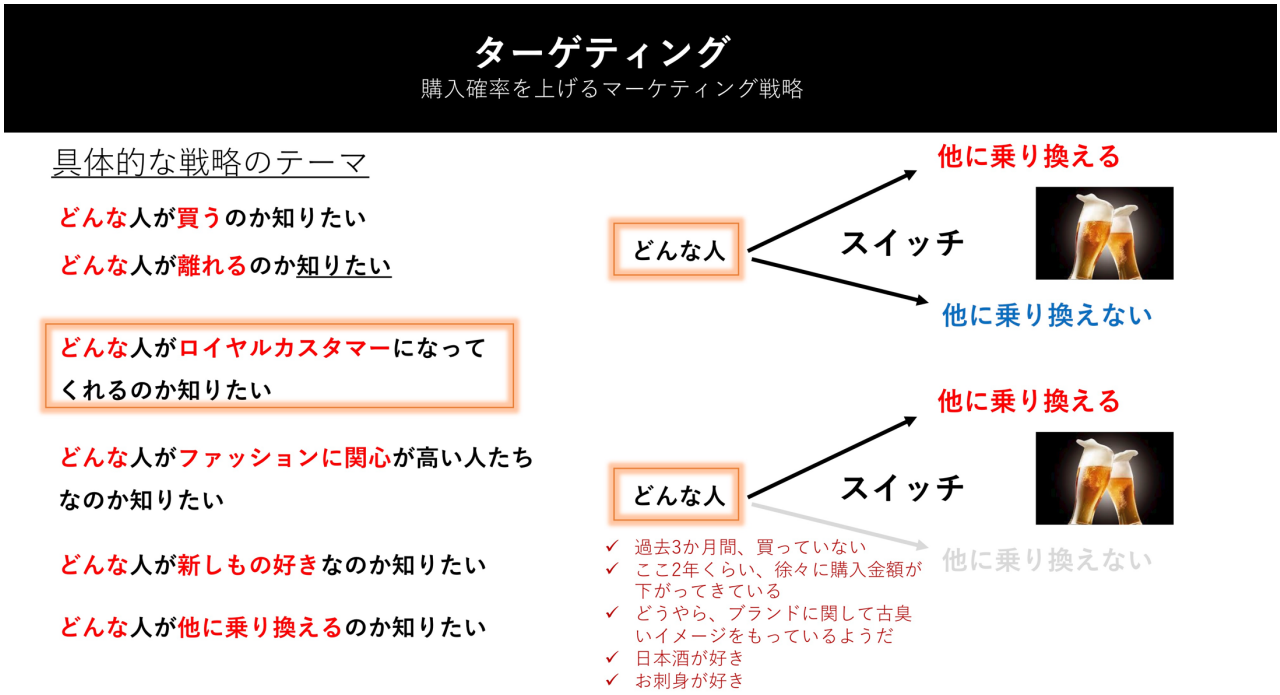


図 2：具体的な学修コンテンツ例

4. おわりに～カリキュラム効果測定結果と限界

効果の測定は、マーケティングデータ分析の受講者に対して実施した。実施時点は、全15回のうち、第1回（2019年9月25日）と第14回（2020年1月10日）の2時点において実施した。効果測定対象者は、両実施回に出席していた受講者５６名である。測定項目は、潜在抵抗意識は、難易感と自己効力感[2]から構成されると考えた。難易感は「Q１：AI（人工知能）・機械学習アルゴリズムは自分には難しそうだ。当てはまると思う番号に〇をしてくださ

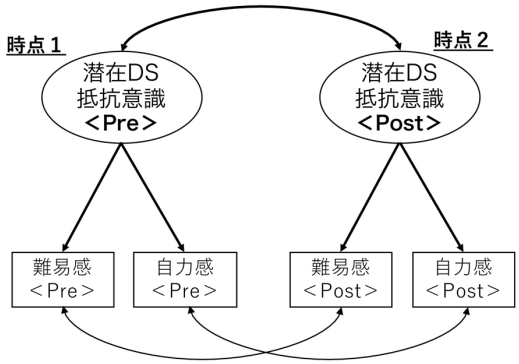
表 3：基礎統計

	時点1		時点2	
	Q1	Q2	Q1	Q2
平均	3.12	2.88	2.82	2.43
標準偏差	0.47	0.74	0.51	0.68
N	56	56	56	56

い」に関して、「とても簡単そうだ」から「とても、難しそうだ」まで、４段階のリッカート尺度で測定した。また、自己効力感は「Q２：今、AI（人工知能）・機械学習アルゴリズムを自分一人で使うことが出来る気がすると思いますか。お気持ちに近いと思う番号に〇をしてください」に関して、「１．今でも、出来る気がする」から「４．まったく出来る気がしない。」まで、４段階のリッカート尺度で測定した。得点が低いほど、難易感が減じ、自己効力感が増すことを意味する。集計結果を表３に示す。

この効果測定は、測定データが２時点間の縦断データであることから、通常の有意差検定は、自己共分散が考慮されていないために適していない。そのために、本研究では、平均共分散構造モデルを構築し、その因子平均の有意差により、効果を検討した。図３に検証のための概念モデルと解析モデルを示す。図３左が概念モデルであり、図３右が概念モデルを数理モデル化した解析モデルである。図３左のDSは「データ・サイエンス」の略である。本稿では、２時点間の因子平均の差に関心がある。そこで、平均構造の比較のためには、測定不変モデルの成立が条件となる。特に、強測定不変モデルの成立が要求されるために、本稿では、強測定不変モデル、厳密測定不変モデルを構築した[3]。強測定不変モデルとは、比較するグループの因子負荷量（図３右のλ）が等しく、さらに、観測変数（図３右のy）の誤差分散（図３右のυ）も等しいという制約を

効果測定の概念モデル



効果測定の解析モデル

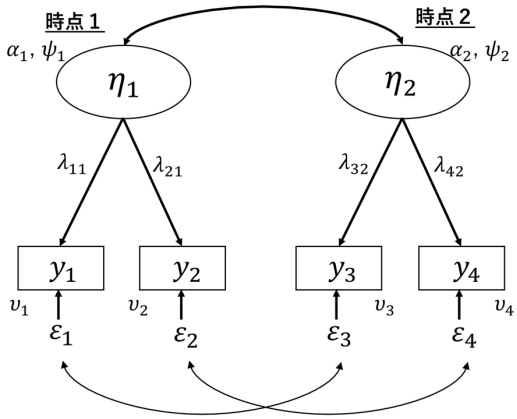


図 3：平均共分散構造モデル

仮定したモデルである。厳密測定不変モデルとは、強測定不変モデルの制約に加えて、誤差も等しいと仮定するモデルである（図３右のε）。両モデルの適合度指標を表４に示す。適合度とは、構築したモデルがデータにどれだけよく当てはまっているのかを示す指標であり、通常、複数の指標から総合的に判断を行う。特に重要な指標では、RMSEA（Root Mean Square Error of Approximation）は１以下で０に近いほど良く、GFI（Goodness of Fit Index）、AGFI（Adjusted Goodness of Fit Index）、CFI（Comparative Fit Index）は１に近いほど適合度が良いことを示す[4]。改めて表４で適合度を確認すると、強測定不変モデルが採択される。よって、強測定不変モデルにより算出された結果により、効果検証を行う。モデルから算出された因子平均（図３右のα）の値と分散（図３右のψ）を表５に示す。時点２における潜在抵抗 DS 意識因子の平均値は、-0.304 で、p<.001 で有意であり、時点１に比して、潜在 DS 抵抗意識因子の因子平均が有意に低下していることがわかる。

表 4：平均共分散構造モデルの適合度指標

	モデル	
	強測定不変	**厳密測定不変**
Chi-sqr	0.134	4.894
df	1.000	2.000
p-value	0.714	0.087
RMSEA	**0.000**	0.161
GFI	**1.000**	0.999
AGFI	0.999	0.990
CFI	**1.000**	0.917
AIC	380.166	382.925
BIC	406.495	407.229

表 5：因子平均と分散（強測定不変モデル）

	潜在DS抵抗意識因子	
	Pre（時点１）	Post（時点２）
平均	0	− 0.304***
分散	0.107**	0.09**

**p<0.05 / ***p<0.001

また、効果量 d を以下の式（1）から算出した。Fm1 は Pre（潜在 DS 抵抗意識因子）の因子平均、Fm2 は Post（潜在 DS 抵抗意識因子）の因子平均、Fv1 は Pre（潜在 DS 抵抗意識因子）の因子分散、Fv2 は Post（潜在 DS 抵抗意識因子）の因子分散、n1 は Pre のサンプル数、n2 は Post のサンプル数である。

$$d = \frac{Fm1 - Fm2}{\sqrt{\frac{n_1 Fv1 + n_2 Fv2}{n_1 + n_2 - 2}}} \qquad (1)$$

d=0.959 であり、標準偏差１程度の差があることを示している。このように、本カリキュラムは、受講学生が、データ・サイエンスに対して、当初抱いていた潜在的な抵抗意識を大きく低減させる効果があったことが認められた。

しかし、本稿では、細かい属性が与える影響については検討をしていない。先天的、後天的を含む理数系能力、高等学校までの数学を中心とした理数系科目の学習状況、大学における教養科目などの影響、受験種別などは、学修効果に影響を与えると考えられる。今後はそうした属性についても統制を行い、検討を重ねていきたい。

参考文献

(1) 木村修平，近藤雪絵：“‘パソコンが使えない大学生’問題はなぜ起こるか～立命館大学大規模調査から考える～”，PC Conference 論文集，pp.179-182（2018）.

(2) A. Bandura: “Self-Efficacy: The Exercise of Control”, p3, New York: W.H. Freeman and Company (1997).

(3) J. T. Newsom：“Longitudinal Structural Equation Modeling~ A Comprehensive Introduction~”, pp.27-52, New York: Routledge (2015).

(4) 豊田秀樹：”共分散構造分析 Amos 編: 構造方程式モデリング”，東京図書（2007）.