研究室でのデータサイエンス教育に関する取り組み ～データサイエンスの卒研事例～

高見 友幸

大阪電気通信大学 総合情報学部 デジタルゲーム学科

takami@osakac.ac.jp

キーワード　データサイエンス教育，卒業研究事例，Python プログラミング学

1　はじめに

本稿では，著者の研究室の卒業研究（以下，卒研）の事例のうち，データサイエンスに関連する研究テーマの概要について紹介する。これらの研究は卒研としての取り組みであり，運営する側（教員）も学修する側（学生）もデータサイエンス教育という意識は持っていない。しかしながら，結果的にはデータサイエンスを学ぶ過程でのゼミの実施と各自の学修であり，それらの成果は研究の成果と直結している。データサイエンス教育の方法論と有効性を探る上での一事例になるものと考える。

卒研への誘導は，著者が担当する授業科目に沿って組み立てられており，その多くが最終的には Python プログラミングへと流れ着くようになっている。これは，研究室の研究テーマの大半に Python プログラミングが必須となるからである。具体的には次の 3 つの流れがある。

- C 言語プログラミング → C++プログラミング → Python プログラミング
- 微積分の数学 → Python プログラミング
- ハードウェアプログラミング → Python プログラミング

卒研のテーマとしてデータサイエンス自体に興味をもつ場合は，著者の研究室を選ぶことになるが，Python プログラミングを活用する研究題材のひとつとしてテーマをみている場合は，第一義的には Python プログラミング技術の追求，第二義的にデータサイエンス研究となっている場合もあろう。

2　卒研事例

研究室への配属は 3 回生前期の中頃に確定するが，卒研テーマへの本格的な取り組みは 4 回生からである。配属となる学生のほぼ全員は Python プログラミングの基礎を 2 回生までの著者の授業で習得済みであり，3 回生の卒研ゼミでは，Python プログラミングの様々な活用について学修する。以下では，Python プログラミングの要点を述べた後，データサイエンスの卒研事例を紹介する。

2.1　Python プログラミング

Python プログラミングの特徴として次の 3 点を強調して卒業研究への導入を行っている。

- 人間の思考を「Python プログラムに翻訳する」という考え方。
- ひとつの問題について幾通りものプログラム例を作るという学修の方法論。
- 「プログラミングしない」という意識（クラスやメソッドをできるかぎり利用する）。

上記の「Python プログラムに翻訳する」という考え方を「整数のリストの中に偶数はいくつあるか」という単純な問題を例に説明したい。以下のプログラム 1 に 3 つの解答例 ans1，ans2，ans3 を示した。どの解答例も 1 行のプログラムである。

```
data = [23, 34, 56, 78, 89, 13, 67]

ans1 = len([i for i in data if i%2==0])

ans2 = sum(i%2==0 for i in data)

ans3 = sum(str(i)[-1] in "02468" for i in data)

print(ans1)
print(ans2)
print(ans3)
```

プログラム 1．リストの中に偶数はいくつあるか.

2 の剰余を使うのが, 通常の解答であり, ans1 と ans2 がこれに当たる（複数のコーディング例として提示）。ans3 は人が偶数を判断する場合の思考方法であり，その思考

(1 桁目が偶数かどうかを見る）を Python に「翻訳」したプログラム例として提示している。

2.2　卒研テーマ

卒研テーマとして案内している中で，データサイエンスの関連分野（あるいは研究題材として使用するデータベース）は，今年度の場合，次の 6 件である。

1. イオノゾンデ観測／地磁気観測の観測データ
   多彩なプログラミングを駆使できるという点，データ収集からはじめて独自データベースの公開までを一貫して学修できる点で，データサイエンス教育に適した教材である。

2. MU レーダーの観測生データ
   観測した受信データのうち，本来は除去するべき雑音部分（スペースデブリや宇宙背景放射の解析）に注目するのが興味深い。

3. DNA の主成分分析
   データは 4 つの塩基配列 GATC だけで構成される単純さながらも巨大なデータ容量であり，かつ明解な結果となる点が面白い。

4. 前方後円墳のデータベースの解析
   個別データを収集し，その上で一括した独自のデータベースを自分で構築できる点が面白い。

5. データベース天文学／古代の天文学
   古天文の領域ではデータ処理が難解であるが，データ解析の考え方に融通性があって，目指す結果をある程度まで引き出すことができる。

6. 古典籍のテキストマイニング
   日本書紀や古事記への適用が，解読に定説がないため面白い教材となろう。

2.3　テーマの実際

上記 6 件の卒研テーマでは，いずれのテーマにおいても既存のデータを用いる。データサイエンス教育という観点から見た場合，それらは教材に相当する。卒研では，それらのデータから有用な特徴をどのように抽出するかということに重点が置かれる。そのため，データ解析に先立ち，データをどのように分類し，どのように事前処理し，関連する周辺データを収集し，どのような相関を取るかという作業を検討することになる。こうした作業は，データサイエンスの本質的な作業である。

ところで，データサイエンス教育として紹介される多数の事例では，教育過程の大部分が，データ解析のときに使用する基礎数学，使用するプログラミングリテラシー，定型化された処理（統計解析やデータ可視化等々）の演習に当てられている。これらの講義や演習は，データサイエンスの本質のトレーニングという観点から見れば，さほど重要ではないというのが著者の考えである。

上記 1) のイオノゾンデ観測データから得られた結果の一例を図 1 に示した。電離圏の平均日変動である。しかし，この平均は収集したデータを単に平均操作するだけでは特徴を抽出することができない。試行錯誤の上，データに内在する性質を掴み出す感覚的な（経験的な）何かが必要となろう。

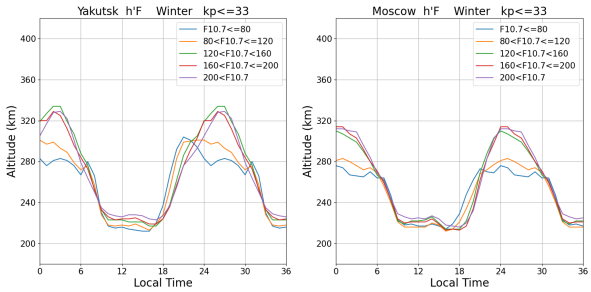

図 1.　Yaktsk と Moscow の平均日変動．F10.7=120 を境として日変動の様相が大きく変わることがわかる．

上記の 6 教材は卒研事例であるが，大学院の授業としての実施にも向いている。また，Python プログラミングの適用題材としての有効性も大きいと感じる。ただし，１年間程度の Python プログラミングのトレーニングが必要である。

データサイエンス学の実践的なトレーニングを想定する場合，どのようなデータベースを教材に用いるかという点が重要になろう。理系の場合，その選択肢は多数ありそうだが，文系の場合，たとえば，ビジネスに対するデータサイエンスの適用事例（成功した事例）がうまく見つかるのかどうか。

3　まとめ

データ解析は，言うなれば，砂金採りが砂金を探すような作業であって，その点がデータサイエンスの面白さであろう。コードが書けないと独自のデータ解析はむずかしくデータサイエンスはできないと考える。既存のデータベースを採用し，それを定型処理するだけのツールを使う限り，導かれる結果はほぼ同じである。

1) データ収集 → 2) データベース構築 → 3) データ解析／データ可視化 → 4) データ公開という過程をデー

タサイエンスの全過程と考えたとき，近年始動している「データサイエンス教育」は，この全過程を教育の範囲として考えているのかどうか。ある目的のために，どのようなデータを収集し（1），どのように分類し（2），どのような切り口で提示するか（4）が，データ解析手法（3）と同等程度に重要なように感じる。

データサイエンス教育として1）～4）の全過程を想定しない場合，では，データサイエンティストとはどの部分の専門家を指すのであろう。たとえば，ゲームクリエイタが，ゲームプログラマ，ゲームデザイナー，ゲームプランナーの共同作業で成り立つように，データサイエンティストにはいくつかの分業があり，それらの共同作業の上に成り立つのであろうか。非専門家の立場からの気ままな原稿をご容赦いただきたく思う。