日本ソーシャルデータサイエンス論文誌
第 3 巻　第 1 号（2019 年 3 月）

特集：データサイエンス人材育成

データサイエンス技術者育成の取組み

水野 信也[†1]

概要：データサイエンス技術者は広い分野で必要とされている．データ解析はどの分野にも必要となる最終的な評価段階であり，このデータ解析の質を上げることで，プロジェクトの意思決定に寄与することが可能となる．しかしながら，このような人材は不足しており，育成することも簡単ではない．本稿では，データサイエンス技術者を育成するための取組みを，実習に関する枠組み，教育の枠組み，そして大学での取組みを紹介する．これらの取組みから，データサイエンス技術者を育成し，現在および将来に対するニーズに応えていく．

キーワード：データサイエンス，人材育成，教育モデル

1. はじめに

データサイエンス技術者は広い分野で必要とされている．データ解析はどの分野にも必要となる最終的な評価段階であり，このデータ解析の質を上げることで，プロジェクトの意思決定に寄与することが可能となる．しかしながら，このような人材は不足しており，育成することも簡単ではない．その理由として，データサイエンス技術者は図 1 のように広く専門的な知識・技術を要求されることが挙げられる．さらに，この知識・技術を修得するには，下記のようなフローが考えられる．

[育成準備]

– 数学, 英語の基礎力の徹底（変わらない知識の獲得）
– データサイエンス像の理解，IT への興味（進み続ける技術の獲得）

[Step1]：ビッグデータの扱いから可視化まで

– データセットの取得解読 → データ構造の決定 → データクリーニング →
– データベース構築 → データの取り出し → データの可視化 →
– 当該分野の専門家とディスカッション

[必要な知識・技術]

• サーバ構築（クラウド環境，データベース運用，計算環境）
• データベース知識（構造決定），プログラミング技術（集計，グラフ作成）
• プレゼンテーション技法

[Step2]：目的に合わせた解析手法の選択と分析(分類・予測)

– 全体の傾向をつかみたい → 統計処理
– シミュレーションをしたい → 確率過程
– 最適化をしたい → 数理計画
– 予測をしたい → 機械学習

⇒ 目的に合わせた手法の選択が必要

[Step3]：結果の検証とモデル改善

– 分析結果 → 専門家とディスカッション → 修正を繰り返す
– 意思決定に寄与する結果 → 必要なデータの提案

この中で特に教育として必要となるのは, Step1 である．データサイエンス分野では様々な分野から，データ解析の依頼がある．医療，製造業，航空，教育など様々である．これらのデータを用いて，意味のある解析を実施するためには，このデータを可視化し，解析者自身が内容を把握するとともに，各分野の専門家とディスカッション出来る材料を揃え，仮説の導きや解析のターゲットを決めていく必要がある．Step1 が疎かになると，データの解釈の点で間違いや曖昧さが残る可能性がある．この Step1 を確実に出来るようにし，Step2 で的確な解析を実施して，Step3 で意思決定に寄与出来る解析フローが重要である．

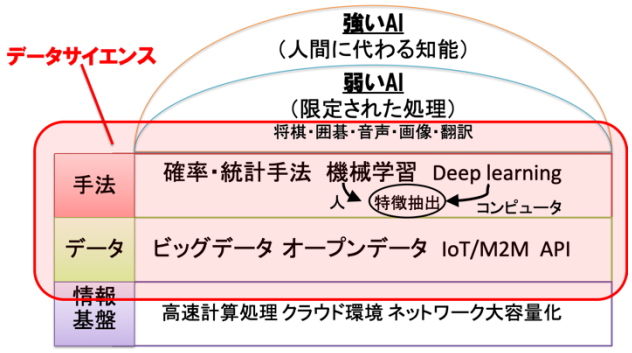

図 1 データサイエンス技術者に必要とされる要素

本稿では，データサイエンス技術者を育成するための取り組みを，実習に関する枠組み，教育の枠組み，そして大学での取組みを紹介する．これらの取組みから，データサイエンス技術者を育成し，現在および将来に対するニーズに応えていく．

†1 静岡理工科大学 （連絡先：mizuno.shinya@sist.ac.jp）

2.　データサイエンス技術者育成における実習内容

データサイエンス技術者育成には，大学だけでなく，企業からの要望も多い．ここでは大学における実習授業や企業内研修で，実際に実施している内容を紹介する．以下は主に企業内研修で実施している内容である．実習環境は現状のニーズに合わせ，Jupyter-Notebook にて，Python と R を両方実行出来る環境で行っている．演習問題を豊富に用意し，得られた知識・技術を活かせるか，確認を行なっている．また数学的知識も必要であることから，必要に応じて紹介をしている．

[データ解析入門（企業内研修）] 7 時間×2 日間の構成

1. データ解析入門：AI とデータサイエンスとは？
2. データ可視化のための環境設定
    - Ubuntu 上に Jupyter-Notebook 環境を構築
    - Jupyter 上で Python と R を実習
3. データ可視化：グラフ作成
    - オープンデータを利用
    - 基本グラフの作成（円，棒，積み上げ，散布図など）
4. 統計手法の利用
    - 基本統計量の算出
    - 相関係数行列，主成分分析など
    - 推定，検定
5. データの分類 1：教師なし機械学習
    - クラスタ分析：階層型，k-means
6. データの分類 2：教師あり機械学習
    - SVM，Neural Network
7. データの予測
    - 単回帰，重回帰分析，Stepwise

演習問題

数学的知識

最尤推定，ラグランジュの未定乗数法，最急降下法，最小二乗法，パーセプトロン，ニューラルネットワーク，バックプロパゲーション

特に，このような短期間の講座で重視しているのは，次の点である．

- データ入手から可視化までのプロセスを習得
  データを色々な角度からみることができる
- データ間のリレーションの重要さを実感
  属性を利用した抽出，分析
- データに対応した解析手法の選択
  No Free Lunch 定理（全ての問題に対応したアルゴリズムはない）
- ストーリー性のある分析へ
  意思決定に寄与できる分析精度まで高める
- データマイニングとデータサイエンスの違い
  データマイニング：既存データから価値を算出
  データサイエンス：戦略からデータ取得，仮説を検証

この内容以外にも，Deep Learning を利用したビッグデータに対応した内容，シミュレーション，最適化など特化した内容も提案している．

3.　大学院でのデータサイエンス技術者育成教育モデル

近年のデータサイエンス技術者育成の要望に応えるために，大学院におけるデータサイエンス技術者育成教育モデルを提案した．この教育モデルは，大学が 2018 年問題を抱えて，今後 18 歳人口が減少し，学生の確保が難しくなる課題と，企業が抱えるデータサイエンスに関わる課題を同時に解決する「オンライン教育を活用した産学連携人材育成」である．企業からは，社員を大学院に入学させ，社内課題を修士課程の研究テーマとして，社内課題と人材育成を同時に行う．大学側も近年発達しているオンライン授業形態を提供し，企業側に負担の少ない環境で教育を実施する．この図 2 に示す教育モデルは，ビジネスモデルとしても評価され[1]，社会からのニーズがあることが確認できている．

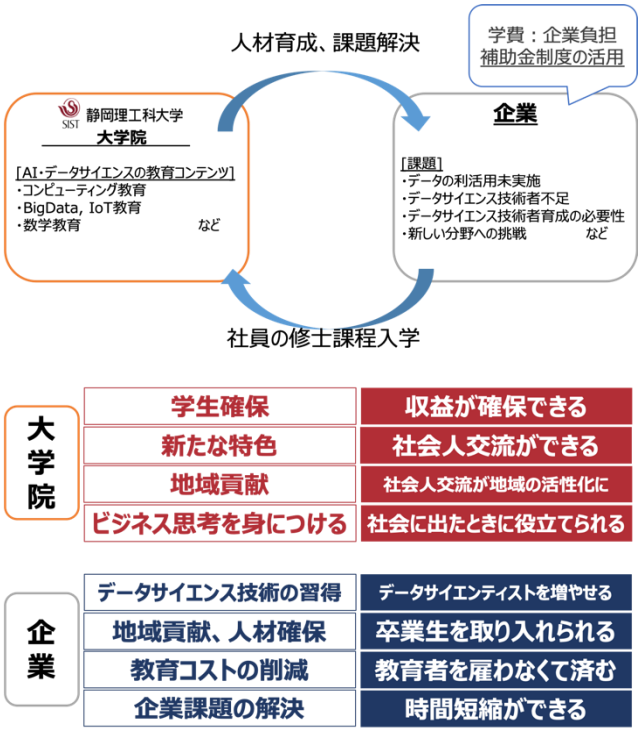

図 2 データサイエンス技術者育成モデルとメリット

4. おわりに

本稿では，データサイエンス技術者育成に関する取り組みを紹介した．ICT 環境の進歩が著しい中，今回述べたような内容もタイミングよく改定していく必要がある．常にニーズに合わせた教育コンテンツの提供が必要である．

私が現在所属している静岡理工科大学では，情報学部コンピュータシステム学科に「データサイエンス専攻」を 2020 年に設置予定である．また大学院にも，理工学研究科システム工学専攻に「データサイエンスコース」を設置予定である．このように大学としての取り込みも活発になっており，産学連携だけでなく，自治体，海外拠点と連携し，常にクオリティの高い教育モデルを提供していくことで，データサイエンス技術者の育成，そして研究・各種事業の活性化が期待できる．

参考文献

[1] 冨永樹哉，大場春佳 : オンライン教育によるデータサイエンス技術の習得と大学の活性化，ビジネスモデル発見&発表会東海大会 2018，東海総合通信局長賞，ICT ビジネス研究会賞