

データサイエンス分野における人材育成のあり方 ～オープンデータ活用研究部会の活動を通して～

水野 信也^{†1}

1. はじめに

現在、世界的に人口増加が急激に進み、食料や化石燃料不足などが問題視される中、日本においては急激な人口減少が見込まれ、2100年には日本の人口が6400万人になると言われている [1]。2018年は18歳人口の減少が再び始まり、2025年には団塊世代が後期高齢者に達し、若年層人口のさらなる減少とサポートを必要とする人口増が同時に起こる。日本国内では既に人材不足が問題となり、この状況は年々悪化すると考えられる。またICT環境整備が進み、人工知能を中心とした多くの話題提供もあり、人工知能、データサイエンスに注目が集まっている。日本政府としても、未来投資戦略2017を掲げ、図1のようにSociety 5.0の実現に向けた改革を進めている [2]。

I Society 5.0に向けた戦略分野・

1. 健康寿命の延伸
2. 移動革命の実現
3. サプライチェーンの次世代化
4. 快適なインフラ・まちづくり
5. FinTech

A. 価値の源泉の創出・

1. データ利活用基盤の構築、徹底したデータ利活用に向けた制度整備
2. 教育・人材力の技術強化
3. イノベーション・ベンチャーを生み出す好循環システム

図1 日本経済再生本部 これまでの「日本再興戦略」について (抜粋)

これらの技術を現場で利用することで、人材不足解消だけでなく、蓄積されたデータから新しい知見を発見し、イノベーションに繋げていくことが重要となる。同時にその基盤技術を担うデータサイエンス分野の技術者育成も急務となっている。しかしながら、様々なキーワードやバズワードが飛び交い、教育機関、企業、自治体の枠に関係なく、現場での混乱が見られる。本稿では今おかれている環境を一度整理し、今後データサイエンス分野でどのような人材育成が必要かを検討していく。

2. データサイエンスの役割

ここではまずデータサイエンスの技術範囲について確認する。データサイエンスは図2のように、コンピュータの処理性能、ネットワーク帯域の大容量化などの情報基盤の進歩とICTの浸透から蓄積されたデータ活用、そして確率過程、統計、機械学習といった成熟した理論とディープラーニング活用の相乗効果で成り立っている。これらの技術を融合させることで、今まで人が経験値を基に決めてきたルールを、莫大なデータ、成熟した理論、大容量計算を可能にした情報基盤から自動的に算出できるようになった。

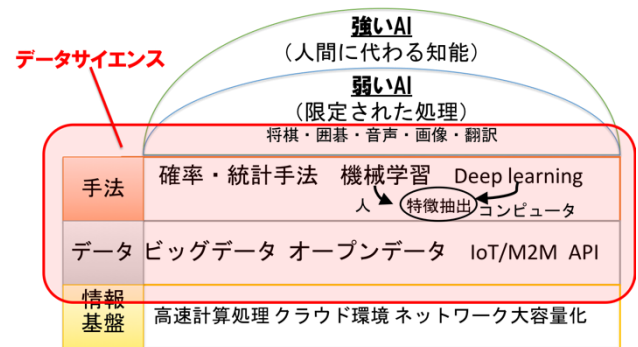


図2 データサイエンスの技術範囲

図2からもAI(人工知能)とデータサイエンスはカバーする技術範囲が重なっているが、AIとデータサイエンスの違いは図3のように考えられる。データサイエンスはデータを利用することで意思決定に寄与する解析結果を提供することを目的とし、アウトプットとしては新しいルール、計算エンジン、戦略といった論理的なものが主である。一方AIは、データサイエンスで得られたルールを自動化し、自律的に運用可能とし、ロボットなどでの物理的な表現まで目的としている。

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

[データサイエンス]

数理モデル、確率過程、統計、機械学習などの手法を用いて
様々なデータを分析し新しい知見を発見することで、
実社会に貢献する

=> **新しいルール、計算エンジン、方針、戦略の形成** **論理的**

[AI 人工知能]

データサイエンス等で得られたルールを実現する手法で、
自律的な学習や行動を物理的に表現する

=> **処理や動作を自動化し、物理的な構成も提供** **物理的**

図3 データサイエンスと人工知能の違い

3. データサイエンス活用：経験値を戦略へ (徹底的な可視化)

データサイエンスを実施する時、まず手がけるのがデータの可視化である。データの可視化によってデータがわかりやすくなり、様々な角度からディスカッションが可能になる。このディスカッション機会を提供することで、それぞれの立場から得られた経験値が意見に反映され、今後の方向性を見出すヒントとなる。ここで得られた仮説が今後の戦略の要素となり、仮説の証明を次のステップで実施していく。また可視化したデータが当然のものとして捉えられてしまう場合も多々あるが、当然と思われていたことがデータで証明されたことになり、根拠を示したことになる。この可視化を実施することで、今までの状況評価を実施することと同様であり、今後の戦略を実施したときの評価指標にもなる。

[データサイエンスの役割]

- ・ 可視化によるディスカッションの場を提供
- ・ 勘や経験値を可視化
- ・ 様々な視点から仮説を構築
- ・ 意思決定に寄与する解析結果を提供

データサイエンス活用例は次のようなものがある。図4は麻酔科医の視線をアイトラッキング装置で評価したものである。アイトラッキングで可視化をすることで、特定の症状が発生した場合の医師の経験値によって着目点に差が生じることが明確になった[3]。

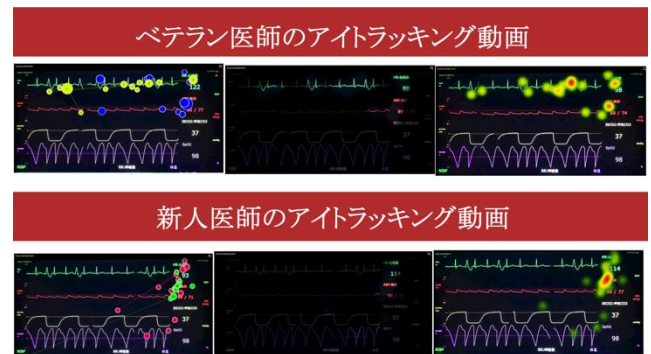


図4 データサイエンス活用例：アイトラッキング

図5はオープンデータの活用する場合のフローと境界線情報を可視化したものである。オープンデータは主となる解析データを補助する役割を担い、非常に効果的である。図6は浜名湖ガーデンパークを、Unityを用いて仮想空間上に作成したものである。ここで実際に訪問者の動きや動線を仮想空間でシミュレートすることで、幅広い立場の人達にモデルを容易に説明できる。この他にも、情報基盤におけるネットワーク量、アクセス頻度、セキュリティ脅威情報を可視化することで、安全なネットワーク運用を行うヒントとなる。

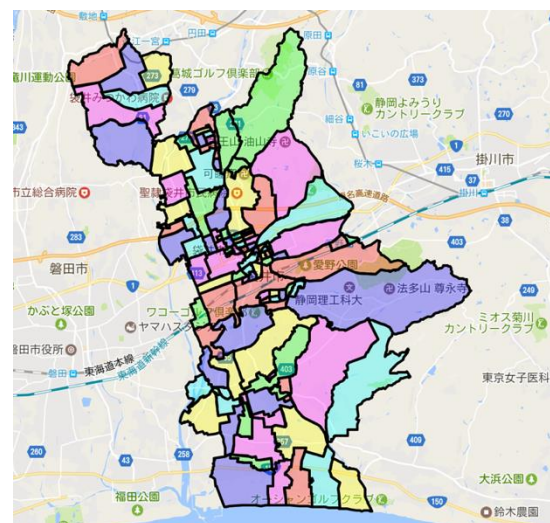
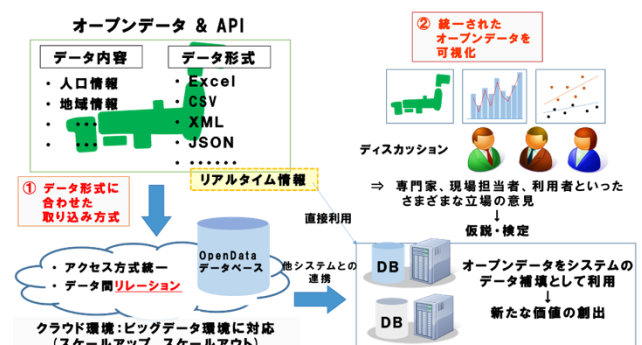


図5 オープンデータを利用した可視化



図 6 浜名湖ガーデンパークの仮想空間(Unity で作成)

4. データサイエンス活用：個人から組織へ (明確な評価:PDCA)

データサイエンスを効果的に実施する場合、重要な点がいくつか考えられる。例えば、データサイエンスで得られた解析結果はそれだけでは一時的なものではない。解析結果を意思決定に反映させ、さらに実行に移していくには組織体が必要となる。図7はデータサイエンスを行う場合、必要と思われる点である。解析者は結果を示す場合、ストーリー性を持たせることで、説得力のある結果を示すことができる。

データ分析にストーリーが必要

単発のデータ分析ではなく、広い視点での分析を行う
目的のあるデータ分析が理想的

データサイエンスを担当する人材と組織の構築

継続性、正確性、公平性を担保できる体制
意思決定は「人」、実行は「組織」

データサイエンスの仕組みを構築しPDCAを回す

戦略立案(P) → 各部署で実施(D) → IR部門が調査・分析(C)
→ FD部門で改善活動(A)
スモールスタートから確実なエビデンスを得て行く

データマイニングからデータサイエンスへ

データ取得の戦略
部分最適から全体最適へ

図 7 データサイエンスに重要な点

そしてデータサイエンスを利用し、PDCA サイクルから組織内の戦略を評価する。図8のように、このプロセスと人材の育成・登用を並列的に進め、イノベーションへと発展させていく。

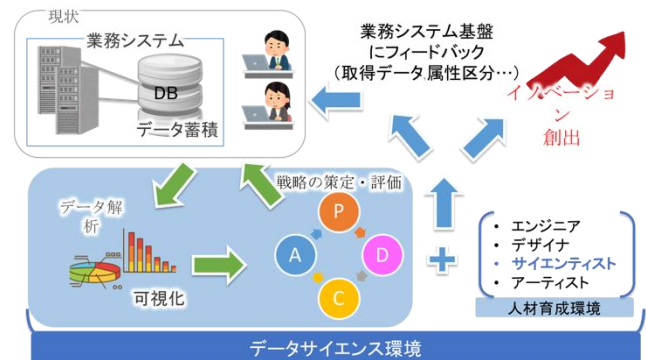


図 8 データサイエンスからイノベーションへ

データサイエンス技術者を育てるためには、次のような流れが考えられる。

【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 → 数値計画
- 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

また解析手法の選択は解析対象によって使い分けないとはいけない。一般的な解析手法 [4] は表1のようになる。

表 1 解析手法の選択と分析

手法	モデル	手法(レベル1)	手法(レベル2)	手法(レベル3)
可視化	教師なし	・k-means法 ・相関分析 ・主成分分析 ・因子分析 ・コレスポンデンス分析	・クラスタリング ・グラフィカルモデル ・構造方程式 ・ベイジアンネットワーク	・自己組織化マップ(SOM)
分類・予測	教師あり	・単回帰分析 ・重回帰分析 ・決定木	・ランダムフォレスト ・ニューラルネットワーク ・k近傍法	・サポートベクターマシン ・勾配ブースティング ・ディープラーニング ・協調フィルタリング

データサイエンスを中心とした高度 ICT 技術者を育成するためには、これまで述べた手法・プロセスに加え、新規分野との連携がイノベーションの観点で必要となる。特に VR (Virtual Reality) といった新しい表現方法は、広い分野で影響を与えている。図 9 のようなデータサイエンスを基軸とした育成カリキュラムに VR, デザイン思考, 語学力といった分野を連携させていくことが必要である。

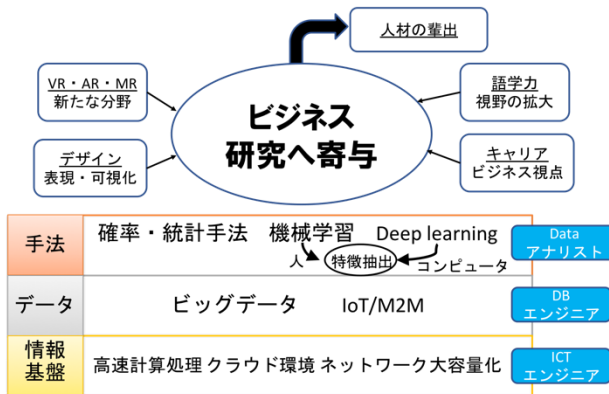


図 9 高度 ICT 技術者育成環境

データサイエンスを推進していく場合、データ分析、戦略立案、評価を一元的にできる図 10 のような組織体が見たい。そこにはデータ収集・分析に欠かせない情報基盤と連携することで、実施がスムーズになると期待できる。現在やられているデータサイエンスは現実的に蓄積されたデータから新しい知見を見出すデータマイニングに近いものである。データサイエンスは結果を得るためにはどのようなデータが必要かといったデータ取得のための戦略も重要である。これにはデータ取得を担う情報基盤が欠かせない。また人材育成と連携することで、組織内でデータサイエンスに関する活性化が生まれ、組織の好循環が期待できる。

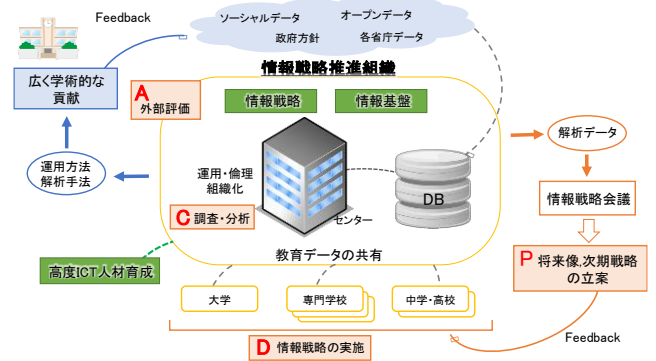


図 10 データサイエンスと情報基盤、人材育成との連携

5. おわりに

今後の変わりゆく社会で、蓄積したデータを解析し、意思決定に寄与する解析結果を提供するデータサイエンスの役割は大きいと考えられる。組織内でデータサイエンスを始めるには本稿で述べたデータの徹底的な可視化を実施することで、組織内でのディスカッションが生まれ、それぞれの経験値、様々な視点から効果的な仮説が導ける。仮説の証明、戦略立案、戦略評価のPDCA サイクルを回し、人材育成と共にイノベーションへ繋げていく。また情報基盤との連携でどのようなデータを取得すべきかといった要求も反映でき、データサイエンスの戦略が明確になる。このようにデータサイエンスを動かす組織体、そこから生まれる情報戦略、そして情報基盤と人材育成の三位一体で連携することで、今後の変わりゆく社会に対応し、特徴ある組織として活躍できると期待できる。

参考文献

- [1] 国立社会保障・人口問題研究所. “将来人口推計とは —その役割と仕組み—,” <http://www.mhlw.go.jp/shingi/2006/06/dl/s0630-4h.pdf> (2018年2月28日アクセス)
- [2] 日本経済再生本部. “未来投資戦略2017—Society5.0の実現に向けた改革—,” https://www.kantei.go.jp/jp/singi/keizaisaisei/pdf/miraitousi2017_t.pdf (2018年2月28日アクセス)
- [3] Ohba, H. and Mizuno, S., “Monitoring Assessment of Anesthesiologists during Operations Using Eye Tracking,” *ISMI 2018 & IEEE SMILE 2018 International Symposium on Semiconductor Manufacturing and Intelligence & IEEE International Conference on Smart Manufacturing, Industrial & Logistics Engineering, February 7-9, 2018, Hsinchu, Taiwan, Program and Abstract Book*, p.50 (2017).
- [4] 加藤エルテス聡志, 機械脳の時代—データサイエンスは戦略・組織・仕事をどう変えるのか?—ダイヤモンド社 (2017).
- [5] 雑賀美明, 「人工知能×仮想現実の衝撃 第4次産業革命からシンギュラリティまで」, マルジュ社 (2017).
- [6] 大角暢之, 「革命の衝撃」, 東洋経済新報社 (2016).