データ科学の研究者・教育者視点での新情報科の「データサイエンス」

早稲田大学・創造理工学部・経営システム工学科　蓮池 隆
(e-mail: thasuike@waseda.jp)

1. はじめに

現在の実生活において，膨大かつ多様な情報やデータが収集され，分析され，社会のありとあらゆる場面で活用されている．今やデータ数が貧弱な分析による意思決定などは通用しない状況となり，今後この流れはますます加速することが予想される．まさに，「データを持っていない．データを分析・活用できない企業は取り残されていく」時代が既に到来している．

私が現在所属している経営システム工学科においても，学生の詳細の就職先も見据えて，データサイエンティストの育成を念頭においたカリキュラムが提供されており，情報を収集する・分析する・応用する能力を高める教育がなされている(はずである)．本学科を卒業する学生の多くが，企業の経営部門やデータ解析部門，SE や SIER といったシステム開発やコンサル業務に携わることになるため，学生には事あるごとに，情報技術・データ解析技術の取得は英会話能力の取得と類似しており，今後の数多くの業種にとって必須の技術となるだけでなく，継続的な努力をする必要性を説いている．

以上のことからも，情報系科目においてデータを分析・活用する力を若いころから養う重要性を認識している経緯から，新高等学校学習指導要領でも『情報とデータサイエンス』の項目が入ってきたものと考えている．実際に，新高等学校学習指導要領解説(情報編)(以下，指導要領解説)[1]の冒頭にも，「今回の改訂では，「情報の科学的な理解」に裏打ちされた情報活用能力を育むとともに，情報と情報技術を問題の発見・解決に活用するための科学的な考え方等を育むことが求められている」が記載されている．特に今後社会で活躍するデータサイエンスの実践者であるデータサイエンティストには，どのくらいの量・質の情報・データが必要となるのか，どの分析手法を用いればよいのか，社会に還元するためにはどう伝えればよいのか，といった総合的な情報活用能力が必要となる．

よって，本発表では，私見もかなり入ってしまうかもしれないが，データ科学に携わるものとして，新高等学校学習指導要領を考察し，現状との比較や注力すべき点，また近未来の社会をより良くする上での課題について検討していく．

2. データ解析技術の深化と一般化

指導要領解説にも記載されているが，「データに基づく現象のモデル化やデータの処理を行い解釈・表現する方法について理解し技能を身に付けること」が今後社会で活躍する若者にとって重要であることは 1 節でも言及した．実際，ここ 10 数年の計算機技術の目覚ましい発展や，Python をはじめとする汎用的なプログラミング言語，および Microsoft Excel だけでなく商用ソフトウェアの SPSS や SAS，無償の R といた様々なデータ解析用ソフトウ

ェアの普及により，ハード面・ソフト面両方において，情報やデータを取り扱う環境が急速に整備されている．よって、データさえ手元にあれば，誰もがさほどの困難なくかつ安価にデータ分析が可能な時代である．指導要領解説にも記載のある，重回帰分析や因子分析，クラスタリングといった手法もそれらのソフトウェアに基本分析ツールとして導入されていることから，極論をすれば入力データを用意し，ソフトウェアのボタンを 1 度クリックするだけで，結果が得られてしまうような時代となっている．つまり，各手法の詳細な中身を知らない，つまり重回帰分析の理論的背景や出力結果が得られる仕組みを知らずに「重回帰分析をすれば，目的としている出力を得る式がわかる」といった漠然とした理解だけでも，結果は得られるのである．

さらに，ある程度 Python などでプログラミングができるような高校生であれば，ニューラルネットワークの拡張系であるディープラーニング手法や，機械学習手法も実装可能となり，指導要領解説にもある手書き文字や顔認識といった画像処理の学習活動も可能になるであろう．これらの体験は，先端研究の追体験と考えることもでき，高校生にとっても大きな刺激となるであろう．これらが容易に実現できることもひとえに，データ科学に携わる研究者・技術者が，ビッグデータから有益な情報を効率的に発見し，かつ実社会に大いに役立つ分析結果を導出したいという大目的の下で，努力を重ねてきた結果である．

一方で、共通教科「情報」として高校生に対して，「「情報の科学的な理解」に裏打ちされた情報活用能力を育む」ことを念頭に置けば，ソフトウェアの利用に偏ってしまっては，目的が達成するとは言い難い．例えば，ソフトウェアを利用して，データ間の相関関係を分析する際に，何も考えず・何も知識を持たず，ソフトウェアの結果だけで，本来相関があるはずもない 2 つの要因間の擬似相関を，本当に相関があると信じてしまうと，例えば「かき氷の売り上げ増加が熱中症患者増加の原因である」のような誤った結論を導きかねない．

しかし，上記にも挙げた重回帰分析や相関分析，クラスタリングなどの理論背景や出力結果の意味を教えるだけ，もしくは非常に簡単な数値例で試してみるだけでは，机上の空論となり，実社会の膨大なデータに立ち向かう力を養うことは難しいであろう．以上のことからも，理論とソフトウェアを利用した分析の両輪をどのようにバランスを取るか，最初は試行錯誤かもしれないが，実践で様々な知見を集積していく必要があるだろう．

3. データ分析は「データの質」次第

2 節でも述べたように，データ分析手法の深化は急激に進んでいる一方で，「データの質」に関しては，ここ 2〜3 年でようやく本格的に言及され始めている状況である．ビッグデータ解析という言葉が急速に広まっていった時期には，データさえ大量に集めることができれば，何らかの有益な知見が得られるだろうと，量重視の考え方が想像以上に拡大しすぎていた(今でも，こういった考えの下で，データだけ持ち込んで，何かいい結果を出してほしいと共同研究を持ちかけてくる企業は少なくない).

しかし，いくらデータがあっても，目的にあったデータ出ないと，何ら良い結果などは出

るはずはない．例えば，若者の趣味嗜好を，日頃の食料品や飲料品の購買履歴から分析したいと考えたとしても，データが男女のみの区別しかないものや，あるいは食料品や飲料品の1日に使う金額のみのデータで，何に利用したかわからないようなデータからでは，目的に合った分析はほぼ不可能である．

さらに，目的に合ったデータであるとしても，質が伴わない場合においては，もちろん良い結果など出てこない．例えば，購買履歴のデータから，どのような特徴を持つ人物が，特定の商品を購買しやすいか解析する場合においても，本来なら 1 ユーザに対してユニークでなければならない購買履歴情報が複数あったとしたら，その時点で現在あるデータの信憑性が著しく低下してしまうであろう．

これらのようにデータが分析に耐えられない状況は，実社会でのデータ解析において，いまだに頻繁に起こるものであり，克服するためには，指導要領解説にも記載の「多様かつ大量のデータの存在やデータ活用の有用性，データサイエンスが社会に果たす役割について理解し，目的に応じた適切なデータの収集や整理，整形について理解し技能を身に付け」た人材の育成が必要不可欠となる．特に，「信頼性の高いデータを収集し適切に問題解決に活用するために必要なデータの整理や整形，データを収集する際に存在する様々なバイアスやデータの入手元の違いによる信頼性を含めたデータの特性について判断する力を養う」能力は，データサイエンスを志す若い人々の必須リテラシーになると感じている．

以上 2 節，3 節で述べてきたように，高校生の段階で，『データサイエンスにおける情報活用能力≒質の高いデータ収集力＋適切なデータ分析手法適用力＋分析結果を正しく理解し，伝える能力』を養うことは，重要である一方，試行錯誤の積み重ねであると感じている．(積み重ねられた知見を分析し，次の施策を立てることも立派なデータサイエンスであろう．)

4. Society5.0 の世界で活躍するために

新高等学校学習指導要領とは多少論点がずれるかもしれないが，情報化社会・情報の将来の在り方を考えるうえで理解しておくべき考え方に，政府提唱の国家戦略：Society5.0 が挙げられる．Society5.0 とは「サイバー空間(仮想空間)とフィジカル空間(現実空間)を高度に融合させたシステムにより，経済発展と社会的課題の解決を両立する，人間中心の社会」[2]を示し，狩猟社会(Society 1.0)，農耕社会(Society 2.0)，工業社会(Society 3.0)，情報社会(Society 4.0)に続く，新たな社会を指すもので，第 5 期科学技術基本計画において我が国が目指すべき未来社会の姿として初めて提唱された．

これまでの情報社会(Society 4.0)において，クラウドサービスが普及し，情報をクラウド上に置くことにより，クラウドに関わる人々の間では，大量の情報の共有が可能となった．しかし，分野横断的な連携・情報共有という観点では，まだ不十分である．また，クラウドに置かれた情報の多くは，人が自ら提供するという操作を必要としている．さらに，人が行う能力にも限界があることから，膨大にあふれる情報から必要な情報を見つけて分析する

作業への対策が求められていた.

そのような中で提唱されたのが Society5.0 である. Society 5.0 で実現する社会においては,「IoT(Internet of Things)で全ての人とモノがつながり，様々な知識や情報が共有され，今までにない新たな価値を生み出すことで，様々な課題や困難を克服できる」とされている．また，人工知能(AI)も含めたデータサイエンスの技術の進化と深化により，「必要な情報が必要な時に提供され，ロボットや自動走行車が実現し，少子高齢化，地方の過疎化，貧富の格差などの課題が克服される」と期待されている(図 1 はそのイメージ資料で内閣府作成).

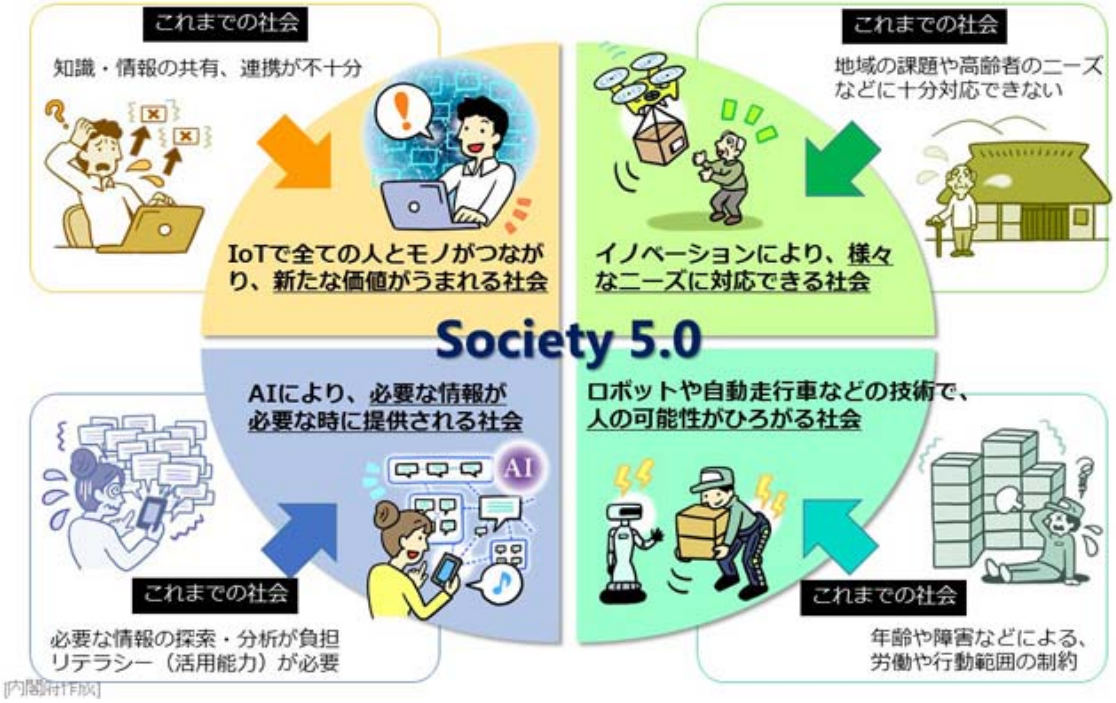

図 1　Society5.0 により実現される社会(内閣府作成資料より抜粋)[2]

Society 5.0 の実現に欠かせないことが，サイバー空間(仮想空間)とフィジカル空間(現実空間)の融合システムの開発である．前述のように，現在の情報社会(Society 4.0)では，サイバー空間に存在するクラウドサービスに，ユーザ自らがインターネットを経由してアクセスし，情報やデータを入手・分析を行ってきた．Society 5.0 では、実社会にばらまかれたセンサーからの膨大な情報がサイバー空間に集積される．サイバー空間では，このビッグデータを人工知能が解析し，その解析結果がフィジカル空間の人間に様々な形でフィードバックされる．つまり，今までの情報社会で行われてきた，人間が情報を解析し価値を生み出す過程が，Society 5.0 では、人手を介さない人工知能がロボットなどを通して人間にフィードバックされることで，これまでには出来なかった新たな価値が産業や社会にもたらされることが期待されている(図 2).

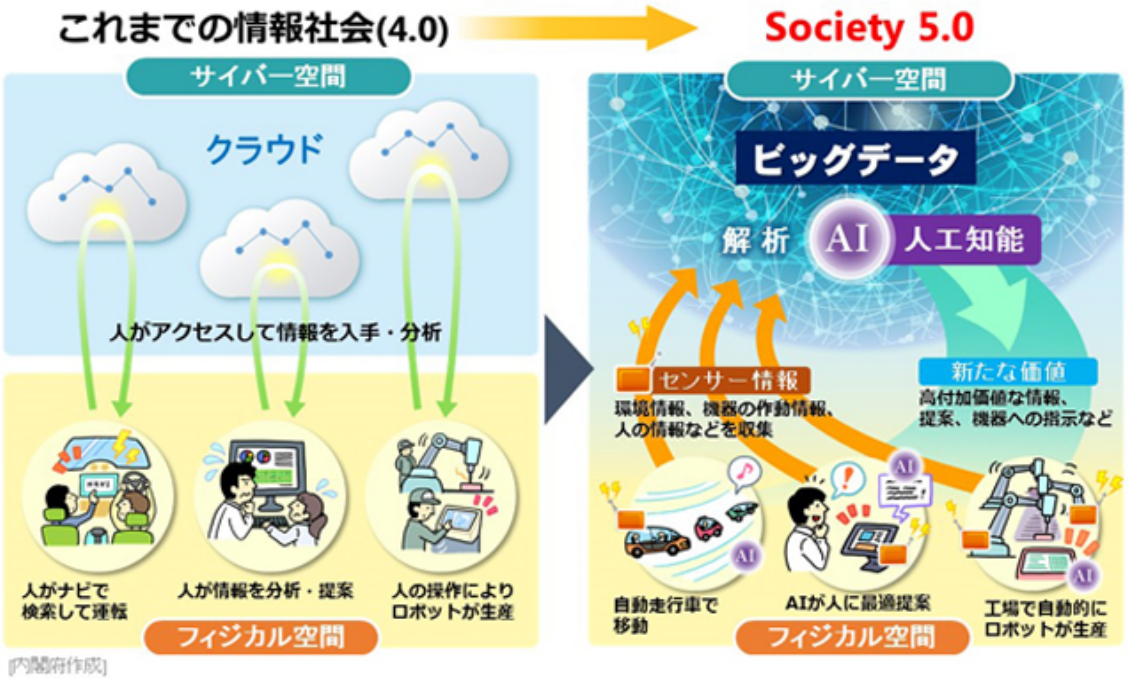

図 2　Society4.0 と Society5.0 の差異(内閣府作成資料より抜粋)[2]

このように，将来実現されるであろう Society5.0 を念頭に，高校生に対して情報技術・情報リテラシーを教育することが重要であろう．特に，IoT の実現により，今よりも桁違いの情報が集積されることになる．その中には，3 節に述べたように，ありとあらゆる情報が得られることから，目的に合ったデータが取得できていないということはあり得ない世界になるかもしれないが，その質については玉石混淆になる可能性が高いであろう．さらに目的に合ったデータの最適な組合せが何かを考える能力を養うことも必要であろう．

5. まとめ

本発表では，新高等学校学習指導要領内にある『情報とデータサイエンス』に関わる項目に関して，データ科学の研究者・教育者の一視点で留意点の考察を行った．やはり，ここ数年の社会における情報教育・データ科学の最も重要な役割としては，「目的に合った情報を質が高い状態で収集できるか」という点だと感じている．データサイエンスは，実社会の多様な意思決定とダイレクトにつながってくる分野であり，なぜそのような結果になったのかという説明責任も伴うものになると考える．よって，もちろんデータ解析ソフトウェアに頼る部分が多くなると思われるが，ソフトウェア任せのブラックボックスではなく，情報やデータの出どころ・信頼性や信憑性，データ解析手法の詳細をできる限り知ったホワイトボックスでの解析ができる能力を養うことが重要であり，これらを実現・実践するために新高等学校学習指導要領を活かしていかなければならない．

参考文献

[1] 文部科学省・高等学校学習指導要領解説(情報編)
http://www.mext.go.jp/component/a_menu/education/micro_detail/__icsFiles/afieldfile/2018/07/13/1407073_11.pdf (最終アクセス日：2018 年 10 月 3 日)

[2] 内閣府・Society5.0 http://www8.cao.go.jp/cstp/society5_0/index.html (最終アクセス日：2018 年 10 月 3 日)