マーケティングにおける
データサイエンス研究・教育の共通基盤としてのKIP

加　藤　淳　一

要約

今日，ビッグデータ（Big Data）あるいはデータサイエンティスト（Data Scientist）という言葉が広く知られている。ビジネスの研究者や実務家の間でも，これらビッグデータやデータサイエンティストの重要性が指摘されている。これらに批判的な立場の研究者を含めて，研究者の考えは「データが重要な時代である」という点でビッグデータを支持する人々と共通している。これは言葉を換えれば，ビジネスが対象の時でさえ，工学者がほぼその学問的な進展を先導できる。

このような現実を前にして，本研究はKIPをマーケティング研究・教育の共通基盤として位置づけることを提案した。KIPは，ブログデータから市場セグメンテーションと市場を捉える軸の抽出の手順である。このような研究・教育の共通基盤ができることで，工学者ほどにプログラムを組めなくともこの分野でマーケティング研究・教育者の立場から貢献できる。加えて，個々人の貢献が散逸することなく蓄積されていく。このようなメリットが想定される。本研究はKIPの背景と手順を説明した。

最後に，残された今後の課題を整理した。1つ目の課題は，より使い易いプログラムを作るという観点からの残された課題である。2つ目の課題は，現行のプログラム自体が抱えている課題である。これは現在使用しているプログラムをかなりの程度書き換える必要があるような種類の課題ともいえる。最後の課題は，マーケティング理論など関連研究を一層詳細に吟味して解決を目指さなければならない課題である。これらの問題について現在の見通しを含めて整理した。

キーワード：KIP，データサイエンス，ビッグデータ，マーケティング

1．背景

（1）ビッグデータとデータサイエンスへの関心

今日，ビッグデータ（Big Data）[1]あるいはデータサイエンティスト（Data Scientist）[2]という言葉が広く知られている。ビジネスの研究者や実務家の間でも，これらビッグデータやデータサイエンティストの重要性が指摘されている。

McAfee and Brynjolfsson（2012）は，これまでの分析（analytics）とビッグデータとの違いを（1）量（Volume），（2）速さ（Velocity），そして（3）多様さ（Variety）の3点と指摘している[3]。

量とは，データの膨大さを指している[4]。速さとは，ほぼリアルタイムでデータ入手できることを意味している[5]。そして，多様さとは，ソーシャルネットワークに掲載された文章や画像あるいはGPSをはじめとしたセンサーなどの様々なデータ形態を指している[6]。

こうしたビッグデータの新しさが指摘されると，その新しさは経営上意味のあるものなのかが問題となる。この点に関してMcAfee and Brynjolfsson（2012）はデータ志向の意思決定を行っている企業は競争企業と比較して平均で5％以上高い生産性と6％以上高い収益率であると述べている[7]。

だが単にビッグデータの分析をすればいいのか。Barton and Court（2012）は，十分にデータと分析とを生かすには3つの相互依存的な能力が必要であるという。1つ目は，企業が多様なデータ源を特定でき，組み合わせられ，そして管理できる能力である[8]。2つ目は，予測と成果の最適化のために先進的な分析モデルを組める能力である[9]。最後に3つは，組織を変革していく能力である[10]。これらの能力を指摘している。

こうした能力の中でもとりわけ，ビッグデータ分析の専門家をデータサイエンティストと呼んでいる。Davenport and Patil（2012）は，データサイエンティストを非常に専門性の高く[11]人材の不足がいくつかの産業では制約条件になっていると指摘している[12]。それほどに注目を集めるデータサイエンティストとは具体的にどのような能力を持つものとされているのか。

Davenport and Patil（2012）は，データの熟達者（data hacker），分析家，コミュニケーター，そして信頼された助言者の混成物（hybrid）と述べ，このような人材はきわめて強力であり稀少でもあると指摘している[13]。今日ビジネスの世界で活躍しているデータサイエンティストはコンピュータサイエンス，数学，あるいは経済学の教育を受けた人材であるという[14]。

データサイエンティストは，上述のように多様な能力の混成物であるが故に魅力的でありながら獲得の困難な人材とされている。Davenport and Patil（2012）は，このような人材はかつての金融工学の専門家と類似しているという[15]。そしてそのような人材は時間と共に教育機関で第2世代の人材としてより大量に養成されるようになっていった[16]。

ではデータサイエンティストも同じ経過を辿るのかといえば，Davenport and Patil（2012）はビッグデータの関連事項の進歩発展が緩慢になる兆しはないと否定的な意見を表明している[17]。さらに，もしも人材が手に入らずこの時代の流れに乗れなければ，競争他社が競争優位を確立して利益を獲得すると指摘する[18]。

以上のような議論は，海外のみならず我が国でも行われている。我が国でビッグデータやデータサイエンティストへ注目した記事を見てみる。日本経済新聞（2013）はITだけでなく統計学やビジネスセンスを持ったこれからの企業成長を左右する存在としてデータサイエンティストの重要性を指摘している[19]。

より具体的にデータサイエンティストのイメージを語った記事は週刊ダイヤモンド（2013）にある。週刊ダイヤモンド（2013）は，大量のデータの分析からビジネス上の価値創出のシナリオを描けるようなデータ活用の専門家であるとし，統計・プログラミング・マーケティングなど多様な知識とスキルを兼ね備えた人物と指摘している[20]。

このような人材は今後不足するとの予測も示されている。日本経済新聞（2013）は「データサイ

エンティストについて，米マッキンゼー・アンド・カンパニーは2018年に米国だけで14万～19万人が不足すると予測する」との記事を載せている。

米国だけでなく日本でも，既にデータサイエンティストという職種に就いている人々がいる。日経情報ストラテジー（2013）は10人のデータサイエンティストとして，日本航空，楽天，全日本食品，大阪ガス，エステー，花王，ロイヤリティマーケティング，東芝，アイズファクトリー，遠州鉄道の10社のデータサイエンティストを取り上げている。

データサイエンティストが取り上げられているのは企業だけではない。週刊ダイヤモンド（2013）は，スポーツにおけるデータの活用を取り上げている。スポーツにおけるデータ活用といえば，まずルイス（2006）のマネー・ボールが思いつくところである。だが日本でもデータ活用は既に現実になっている。週刊ダイヤモンド（2013）は，28年ぶりにロンドン五輪で銅メダルを獲得した女子バレーボールのデータアナリストを取り上げている。その他にも，週刊ダイヤモンド（2013）は福岡ソフトバンクホークスでデータの活用に積極的な姿勢を示していると伝えている。

このように，国内外でビッグデータやデータサイエンティストが注目されている。ただ全ての統計学者が，このようなビッグデータへの注目を好意的に評価しているわけではない。次に，ビッグデータを巡る2つの立場を紹介しつつ，彼らの相違点と更に重要な彼らの共通点を明らかにしていく。

（2）データが重要な時代

西内（2013）はこうしたビッグデータへの注目が集まる現状に対して批判的な立場をとる[21]。西内（2013）の主張点はおおむね次のように整理できる。ビッグデータのように全数を集めようとしなくとも，サンプルからでも十分な精度で全体を推定し分析できる[22]。にもかかわらず，ビッグデータの収集分析に多額の出費をすることは合理的ではない。

ビッグデータの技術を使っていても，単に単純集計だけでは全く統計学を生かせていない[23]。西内（2013）は特に因果関係の特定に関連して多数のページを割いて説明している[24]。西内（2013）は，因果関係の特定方法に関連してランダム化実験の重要性[25]，ランダム化できない3つのパターン[26]，さらにランダム化できないときケースコントロール[27]や回帰分析[28]により因果関係を追求できると説明している。つまり，西内（2013）は小規模のサンプルから全体を推測し因果関係を明らかにするというスタンスをとる。

このようなスタンスをこれまでのデータ分析の立場と指摘し，ビッグデータは新しい貢献をすると主張する研究者もいる。例えば，マイヤー＝ショーンベルガー・クキエ（2013）と同じ2人による論文のCukier and Mayer-Schoenberger（2013）である。

彼らは，ビッグデータが活用されるようになり大きく3つの変化が起こるという。彼らの主張にしたがって整理してみる。第1の変化は，小規模のサンプルではなく多くのデータを収集分析することである[29]。サンプルを抽出する方法は全データを収集できないときの便利な簡便法である[30]。だがほぼ完全な全データを収集すれば，収集した後に様々な角度から分析も可能になる[31]。つまり，これまでならばあらかじめ調査目的から緻密に計画したうえで，サンプリングによりデータを収集し

てきた。だが全データを収集する場合，まずデータをかき集めてからどのような人々に注目するのかを考えるような分析が可能となる[32]。

第２の変化は，わずかなきちんと整えられたデータをあきらめて価値のある大量のデータを手に入れるようになることである[33]。サンプルから全体を推定するとき，データに誤りが混入すると分析結果の全体に影響を与えていた。だがビッグデータのように膨大なデータから全体的な傾向を把握できることで十分に有効なケースがある[34]。これは全体的なイメージを把握できるという意味で印象派の絵画に例えられている[35]。

そして，第３の変化は，因果関係の探求をあきらめ相関関係を見出すことである[36]。これまでのように仮説を立ててそれを検証するというスタイルではなく，計算により最も相関の高いものをピックアップしてくる方法が有効であると主張する[37]。これらがビッグデータの時代に起こる大きな変化として整理される。

以上で見てきたように，複数のビジネス関連記事でビッグデータおよびデータサイエンティストの重要性が主張されている。確かに西内（2013）のようにビッグデータへの批判的な立場の研究者も存在する。それでもマイヤー＝ショーンベルガー・クキエ（2013）と Cukier and Mayer-Schoenberger（2013）のように，新しい時代の到来を主張する研究者も確かに存在する。

そして何より重要なことは，ビッグデータへ批判的な西内（2013）のような研究者も含めて「データが重要な時代である」という点でビッグデータを支持する人々と共通している[38]。この点は，マーケティング研究者として改めて議論すべき課題である。これは言葉を換えれば，ビジネスが対象の時でさえ，工学者がほぼその学問的な進展を先導できるし現にしているということである。そのことをデータサイエンティストの養成コースから垣間見てみる。

（３）工学者に先導されるデータサイエンス

米国ではデータサイエンティストの養成コースが近年次々と設立されている。コロンビア大学[39]は，修士課程としてデータサイエンティストの養成コースを提供予定であるという[40]。ウォール・ストリート・ジャーナル[41]は，「既にデータマイニングのプログラムを設けている大学には，スタンフォード大学（学士コース），ノースカロライナ州立大学，ノースウエスタン大学，ニューヨーク大学がある。」と伝えている。

これらの大学のうち，例えばコロンビア大学の計画しているカリキュラムを見る。すると開講科目は，アルゴリズム，機械学習，確率論，統計学などである。ビジネス関連科目は，あったとしても選択科目としての扱いである。

このようなカリキュラムは，工学に軸足を置いた人材を養成することになる。言い換えれば，マーケティングのためにデータを分析する状況において，マーケティングに関する知識は選択科目で学べば十分である。このように考えていると伺わせる。

この現実を前にして，我々マーケティング研究者は工学者と共にこの分野に貢献できる立場にいなければならない。このとき特に重要なのは次の点である。我々はマーケティング研究者なのだから，マーケティング先行研究に軸足を置いてビッグデータの分析に貢献しなければ存在意義はない。

図表１　コロンビア大学のカリキュラム

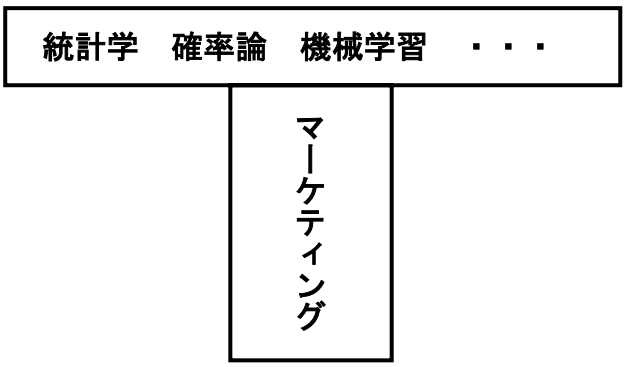
Our proposed MS curriculum will likely be as follows:

- Algorithms for Data Science (CS/IEOR), 3 credits
- Machine Learning for Data Science (CS), 3 credits
- Exploratory Data Analysis for Data Science (STATS), 3 credits
- Data Engineering (CS), 3 credits
- Probability (STATS), 3 credits
- Statistical Inference & Modeling (STATS), 3 credits

出典：http://idse.columbia.edu/masters　2013年９月４日確認

Rob Rodriguez氏の言葉[42]をマーケティングに置き換えて借用すれば，この軸足をおくという考え方はT-shape（Ｔ字）の縦棒をマーケティング先行研究にする。そして，Ｔ字の横棒を幅広く統計学，確率論，機械学習，そしてアルゴリズムなどに広げていくと表現できる。

図表２　Ｔ字型のマーケティングベースのデータサイエンティスト



出典：Rob Rodriguez氏のアイデアに基づき著者描画

マーケティング先行研究に軸足を置いて，その他の多様な知識（統計学，確率論，機械学習，そしてアルゴリズムなど）に広げていく。このようなマーケティングベースのデータサイエンティスト（Data Scientist in Marketing）を具現化していくのに必要とされている研究について検討する。

（４）マーケティングベースのデータサイエンス

我々マーケティング研究者は，自分自身の手で工学者と同等のレベルでビッグデータの分析用プログラムを作成できない。ゼロからビッグデータ分析用プログラムの全体を作成する。これは我々マーケティング研究者が活躍するステージではない。

もちろん我々マーケティング研究者も書籍で学べば，短いプログラムを作成し実際に動かして分析できる。だがビッグデータ分析用プログラムをゼロから作成するとなると敷居は高くなる。なによりこのような工学的な部分での貢献は，先に述べたマーケティング先行研究に則った貢献とも異

なってくる。

より妥当な貢献のあり方を探ってみる。すると，何らかのビッグデータ分析用プログラムを共通基盤とする。そしてマーケティング先行研究の立場から，より望ましいプログラムへとその一部を書き直していく。これが現実的な貢献のあり方となる。

この共通基盤のプログラムという考え方は，我々マーケティング研究者のプログラミング能力が低いと述べているのではない。プログラミングはデータサイエンティストとして必要な能力だが，我々マーケティング研究者の優秀さを示す決定的なポイントではないということである。

より一層重要なことは，こうした共通基盤としてのプログラムの存在が知の蓄積の基盤となることである。共通基盤がなければ散逸してしまうであろう研究成果が共通の基盤へ貢献することで蓄積される。これは研究として重要なことである。

以上から，我々マーケティング研究者にとって受け入れやすい形で作られたビッグデータ分析用プログラムが必要であるという結論に至る。このプログラムとして我々は KIP の利用を提案する。

ブログテキストからの市場セグメンテーション研究が工学者の手により進められていく。KIP は，マーケティング先行研究が全く無視され参考文献にも取り上げられていない状況を批判して提案された[43]。したがって，マーケティング研究者から見て，その提案手順は突飛なものになっていない。

さらに，この手順を基礎にすることで，既に単なるセグメンテーション研究だけでなく他のマーケティング研究へと応用可能であることが示されている。このような特徴からも，マーケティングベースのデータサイエンティストにとって研究蓄積の共通基盤として有益である。

このような前提に立つと，(1) KIP についてマーケティング研究者に理解できる形で整理すること，(2) マーケティング研究における KIP の位置づけを整理しておくことの 2 点が重要である。本研究の目的はここにある。改めて本研究の目的を明示すると，マーケティング研究者に向けて KIP の概要，マーケティング研究における KIP の位置づけ，そして今後の研究に向けて残された課題を整理することである。

2．KIP の概要とマーケティング研究における KIP の位置づけ

本章では，本研究の目的に答えるために 4 点に分けて検討する。まず KIP を提案した背景である。ブログテキストマイニングの手法は様々に提案されている。そのなかで KIP が目指したことを提案の背景から説明する。第 2 に，KIP の概略である。既に KIP の紹介は他の論文の中で行っている。したがって，ここでは使用した指標などを含めた情報を加えつつコンパクトに説明した。

第 3 に，KIP の手順の中でこれまで他の論文で詳しく取り上げていないデータ収集について説明する。特にマーケティング研究者を念頭に置いて，データ収集の方法について説明する。最後に，マーケティング研究の中に KIP を位置づける。本章の背景で説明されるように，KIP は市場セグメンテーションと市場を捉える軸の抽出手順として提案された。したがって素直にその方法として利用できる。だがそれに留まらず，マーケティングの中心的な課題である市場創造とも関連している。この点に関して言及する。

（1）KIPの背景

KIPは，マーケティング先行研究に則してブログテキストデータ[44]により市場セグメンテーションと市場を捉える軸の抽出手順として提案された。Kotler and Keller（2006, p. 240）の定義を引用すれば，「A market segment consists of a group of customers who share a similar set of needs and wants」である。すなわち，市場セグメンテーションは類似のニーズをもった消費者グループである。

ここでの問題は類似のニーズにまとめる基準である。例えば，性別が基準として適切であるならば，男性と女性で消費者をまとめる。すると，男性グループは同じようなニーズを共有しており，女性グループはまた別のニーズを共有している。このようにグループ作成の基準が必要となる。

我々はマーケティング先行研究に則してブログテキストから市場セグメンテーションを行うと主張していた。したがって，我々はマーケティング先行研究から適切な基準を選択せねばならない。

我々の採用した先行研究はWedel and Kamakura（2000）に紹介されている次の図表のような基準である。この基準は，より最近でいえば中村（2008）でも採用されている。Wedel and Kamakura（2000）は，次の図表のようにセグメンテーション基準を整理している[45]。

図表3：市場セグメンテーションの変数

	一般的変数	製品・店舗固有の変数
観測可能	地理変数、デモグラフィックス変数、社会経済変数	使用頻度、ブランドロイヤルティ、店舗ロイヤルティ、採用時期、消費場面
観測不能	パーソナリティ、生活価値、ライフスタイル	プロモーション弾力性、知覚便益、購買意図

出典：中村（2008, p. 6）の図表1-1を使用（原典は，Wedel and Kamakura（2000, p. 7））

この表で変数一つひとつは重要ではない。ここで注目するのは一般的変数と製品・店舗固有の変数である。市場セグメンテーションの分類基準は，この表によると大きく分けて一般的変数と製品・店舗固有の変数という2つに分類されている。

それぞれを大まかに把握すれば，一般的変数は「消費者個人」に関連する変数である。製品・店舗固有の変数は「商品」に関連した変数である。このようなマーケティング先行研究に依拠した基準から，消費者に関連した分類基準と商品に関連した分類基準の2つの分類基準により類似の消費者をグループ化すれば市場セグメンテーションになる。

利用可能なデータは，ブログのテキストを構成する単語（主として名詞）の使用頻度である。このデータから消費者個人と商品を基準にして消費者をグループ化する。その方法として，消費者個人を特徴付けている単語と商品を特徴付けている単語をブログテキストから抽出する。そして，そ

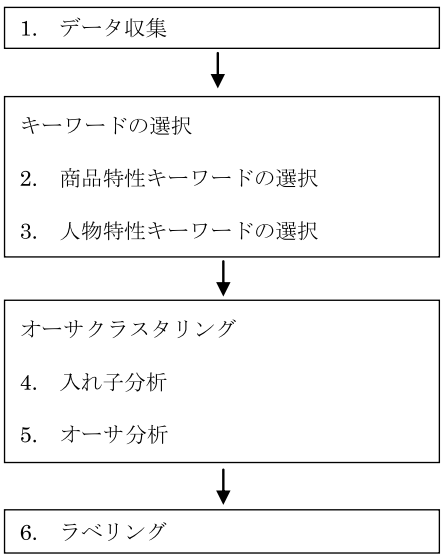
の単語群の使用頻度が類似している消費者（ブログの著者，本研究は，ブログオーサと呼ぶ）をグループにまとめる。

このように，ブログテキストというビッグデータから市場セグメンテーションと市場を捉える軸の抽出手順を示した。KIP は，上述のようにマーケティング先行研究に依拠しつつブログテキストから市場セグメンテーションと市場を捉える軸の抽出を行える。既に述べたように，この手順はマーケティング研究者にとって突飛な手順になっていない。だからこそ，我々はこの手順を共通基盤として提案したい。ここで節を改めて，ブログテキストから市場セグメンテーションを行う手順の概略から説明をはじめる。

（2）KIP の概略

KIP の概略は，既にいくつかの文献で示されてきた。その中でも簡明にまとめられているものとして加藤・石川（2011a）がある。その内容は図表 4 のように図示できる。大きく見ると，(1) データの収集，(2) キーワードの選択，(3) オーサクラスタリング，そして (4) ラベリングとなっている。キーワードの選択は，前節で説明した 2 つの分類基準（消費者個人を特徴付けている単語と商品を特徴付けている単語）をブログ記事（本研究は，ブログエントリと呼ぶ）から抽出するステップである。

図表 4　市場セグメンテーションの手順



出典：加藤・石川（2011b, p.26）の図 2 を使用

オーサクラスタリングは，それら分類基準を用いてブログオーサ（消費者）をグループにまとめるステップである。最後に，ラベリングにより，それらブログオーサ（消費者）のグループを捉える軸を抽出する。このようになる。以下，この図表 4 にしたがって各ステップを説明する。ただし，これら 6 ステップの区分は便宜的なものである。

ステップ 1 は，データの収集である。KIP はブログエントリを収集・分析することを想定している。まず，市場セグメンテーションしたいある特定の商品やサービスを指す単語（名詞）一語を指定する。この単語一語をターゲットキーワードと呼ぶ。次に，ターゲットキーワードを含むブログテキストを検索する。

更に，検索結果に基づいて，一度でもターゲットキーワードを使用したブログオーサを特定し，そのブログオーサの書いた全ブログエントリを収集する。

最後に，収集されたブログテキストから単語（主として名詞のみ）が抽出され，ブログエントリごとに使用した単語とその単語の使用頻度が行列に整理される。この行列が生データとなる。

ステップ 2 は，1 つ目の市場セグメンテーションの基準の選択である。1 つ目の市場セグメンテーションの基準として，ターゲットキーワードと類似度の高い単語がステップ 1 で収集した単語の

中から選択される。これを商品特性キーワードと呼ぶ。類似度の指標としてKIPはコサイン尺度を利用している。

まず一端，全ての生データ行列は，tf・idfの行列に変換される。これは重要度の指標である。tf・idfは，tfという指標とidfという指標の積である。tfはterm frequencyの略であり，各単語が何度使用されたかという単語頻度を指している。idfはinverse document frequencyの略であり，各単語を含むドキュメントの数の逆数を指している。ここでのドキュメント数は，各単語を含むブログエントリ数である。tf・idfの計算式には様々なバージョンが提案されている。例えば金（2009, pp. 60－61）にもその一例が紹介されている。

次に，ターゲットキーワードと各単語のベクトル同士のコサイン尺度が計算される。このコサイン尺度は両者が無関係なほど0に近い値となり，両者の関係があるほど1に近い値となる。したがって，ピアソンの積率相関係数と同様な解釈が可能である。この値により，閾値以上の類似度の高い単語を商品特性キーワードとして採用する。コサイン尺度について，例えば金（2009, p. 161）に説明されている。なお，tf・idfもコサイン尺度も共に一般的な指標であり，他の指標を用いても問題ない。

ステップ3は，2つ目の市場セグメンテーションの基準の選択である。2つ目の市場セグメンテーションの基準として，ブログオーサを特徴付ける単語がステップ1で収集した単語の中から選択される。これを人物特性キーワードと呼ぶ。人物を特徴付けている程度の指標として，tf・idfが使われている。ここでのtf・idfの値は単語とブログオーサの行列を作り，ドキュメント数をブログオーサ数として求める。このようにtf・idfの値を計算し，その値が閾値以上に大きい単語を人物特性キーワードとして採用する。

ステップ4は，ターゲットキーワードと類似した単語（商品特性キーワード）とオーサを特徴付ける単語（人物特性キーワード）の2つの基準でブログオーサをグループ化する。まず，商品特性キーワードでブログオーサをグループ化する。次に，商品特性キーワードでグループ化された各グループを，人物特性キーワードで更にグループ化する。これを入れ子分析と呼ぶ。KIPはグループ化にSOM（Self-Organizing Map）を使っている。これもSOMでなければならない必然性はない。例えばK-meansなど他のグループ化の方法を採用しても構わない。グループ化の方法について，例えば金（2009, p. 160－176）に説明されている。なお，ステップ5でオーサ分析を行うならば，ステップ4を独立に実施する必要はない。

入れ子分析により，市場セグメンテーションできる。だがマーケティングとしては，ロイヤリティの高いブログオーサと低いブログオーサに分けて分析するとより有益である。そこでステップ5では，ロイヤリティの高いロイヤルオーサと低いロングテイルオーサに分け，ブログオーサをグループ化する。

まず，商品特性キーワードでブログオーサをグループ化する。次に，グループ化されたブログオーサからロイヤルオーサとロングテイルオーサを選択する。最後に，選択されたブログオーサを人物特性キーワードによりグループ化する。これをオーサ分析という。

ステップ6では，市場を表す軸を抽出する。ロイヤルオーサとロングテイルオーサのそれぞれの

オーサ分析によりグループ化された。これが市場セグメンテーションである。だがKIPは，市場セグメンテーションで終わらず主成分分析によりこのグループ化された消費者（つまり，市場）を捉える軸を抽出する。この軸は主成分分析の結果にもとづいてラベルを貼られ，軸により市場を捉えられることになる。これをラベリングという。なお，主成分分析の利用の仕方は既に加藤（2013）で説明した。これらがブログ分析の手順となる。

ブログテキストマイニングは，その分析において様々な指標や手法を利用する。KIPは，いくつかあり得る指標や手法の中で，広く用いられている指標や手法を使用している。現在特定の指標や手法を採用しているとしても，その指標や手法でなければならないという必然性はない。したがって，一つひとつの指標の詳細はテキストブックで学べる。その反面，これらの手法が突飛というような状況はない。

（3）データの収集方法

ここでは，KIPで独自の工夫をしたデータの収集方法についてマーケティング研究者を想定して説明する。KIPで収集したデータはブログエントリであった。このようなウェブ上のデータを収集する分野はクローリング（web crawling）といわれており，情報理工学の専門の分野として研究されている。例えばManning et al.（2008）において，クローラが必ず満たさなければならないことと満たすべきことを示した上で，基本的なクローラの仕組みを説明している。

このような専門分野での研究蓄積がある中で，KIPは次のような方法でブログを収集する。まず，少なくとも一度はターゲットキーワードを使用したブログオーサを特定する。そして，そのブログオーサが書いた全ブログエントリを収集する。このような手法を採用している。データの収集方法の詳細は，既に石川等（2012）でとり上げている。だが，石川等（2012）はブログ検索機能を利用した方法の説明になっていない点で，KIPで採用している方法の説明と若干異なる。そこで，石川等（2012）に依拠しつつ，この点を含めてマーケティング研究者に理解できるように説明する。まず，KIPでのブログエントリの収集の手順を箇条書きにまとめると次のようになる。

⑴ ターゲットキーワードを名詞一語で決定する。
⑵ gooブログ検索により，（手順⑴で決定した）ターゲットキーワードをブログエントリ内で一度でも使ったブログエントリを検索する。
⑶（手順⑵の）gooブログ検索でヒットした全ブログエントリの全ブログオーサを重複無く特定する。
⑷（手順⑶で）特定された全ブログオーサがこれまでに書いた全ブログエントリを収集する。

このような手順になる。KIPはgooブログというブログサイトからブログエントリを収集している。ブログエントリを含めたウェブサイトはXMLといわれる言語により記述されている。我々ユーザは，それをInternet Exploreのようなウェブブラウザというソフトウェアで開くことにより閲覧できる。

したがって，メモ帳に「あいうえお」と書くのと同じ意味で，書いたままが表示されているのではない。書式情報など（中央揃えか，太字かなど）の多様な情報を含めて記述し，それをウェブブラウザというソフトウェアが読み込み処理する。こうすることで，通常見ているようなウェブページとして表示される。

ここで問題は，ブログサイトごとにその構造が異なることである。例えばInternet Exploreのようなウェブブラウザで表示したとき，画面右側にブログ記事を表示するのか，中央に表示するのか，あるいはタイトルや日付などの情報の後ろにブログ記事を表示するのかなどブログサイトごとに異なる。

これではどこに存在するテキストを抽出すれば，ブログテキストを抽出できるのか一般化できない。そこで，特定のブログサイトに絞ってデータを収集することで，より適切にブログテキストのみを抽出できる。KIPはgooブログに絞ってデータを収集している[46]。

まず，KIPはgooブログ検索の機能を利用して，ターゲットキーワードを使用したブログ記事を探し出す。これはGoogleやYahooといったポータルサイトからウェブサイトを検索するのと同様のことをプログラムで自動的に行っている。

次に，ターゲットキーワードを使用したブログエントリが特定できたら，そのブログエントリの中に記述されているブログオーサを識別するID（以後，ブログオーサIDと呼ぶ）を抽出する。この情報はInternet Exploreのようなウェブブラウザからは見えないものの，XML言語では書いてある。具体的には次のような記述から抽出する。

〈rdf:li red:resource=http://blog.goo.ne.jp/[blog id]/e/[entry id]/〉

この［blog id］が，ブログオーサIDを指している。［blog id］の抽出は，少なくともブログエントリの中で一度はターゲットキーワードを使ったことのあるブログオーサの発見を意味している。プログラムにより，上記の「blog id」の部分のみを抽出する。なお，「entry id」の部分は，各ブログエントリを識別するIDが記述されている。

最後に，この［blog id］のブログオーサがこれまでに書いた全てのブログエントリを手に入れられれば，KIPの目指すデータを手に入れられたことになる。ブログエントリは最新のものだけでなく，過去の記事も蓄積されている。一般に，この過去の記事を一覧できるページが用意されている。gooブログの場合，次のURLに置かれている。

http://blog.goo.ne.jp/[blog id]/arcv

この一覧ページを利用することで，ブログオーサIDさえあれば，そのブログオーサが過去に書いた全ブログエントリを総なめにして取得できる。以上のような方法により，KIPは分析対象となったブログオーサが過去に書いた全ブログエントリを収集し生データの取得を行っている。

ここまでで，（1）KIPの背景となったマーケティング研究とそこから導かれた2つの基準につい

て，（2）KIPの手順の概略について，そして（3）KIPでのデータ収集方法について説明してきた。この章の最後に，マーケティング研究の中でのKIPの位置づけを整理する。

これによりKIPが単に市場をセグメンテーションと市場を捉える軸の抽出だけでなく，マーケティングの中心的課題の1つである市場創造とどのように関わるのかを説明する。これはKIPをマーケティングにおけるデータサイエンス研究・教育の共通基盤として用いる上で重要である。なお，既発表のKIPを用いた経験研究などの個別研究[47]は取り上げない。

（4）市場創造の分析：マーケティング研究の中でのKIPの位置づけ

既に説明してきたように，KIPは自然言語をデータとしてマーケティング既存研究に則して市場セグメンテーションと市場を捉える軸の抽出手順として提案された。したがって，素直に市場セグメンテーションと市場を捉える軸の抽出に使える。だがそれだけではない。ここではマーケティングの中心的課題となっている市場創造の分析とKIPの関係について整理する。

石井（1993）に指摘されて以降，広く認識されてきたことにマーケティングにおける異文化性の媒介と創造に関連した問題がある[48]。ここで市場における異文化性の媒介と創造とは，広く市場で用いられている見方・考え方の変化を指している[49]。とりわけ重要なのは，この見方・考え方が「変化する」という点にある。

この見方・考え方の変化を明らかにする手法として，石井（1993）において解釈学的アプローチの重要性を主張している[50]。マーケティング研究においてインタビューに代表されるような定性的手法の研究が蓄積されるようになり現在に至っている。なお，ケーススタディについては，イン（1996）が広く知られている。インタビュー（特にフォーカス・グループ・インタビュー）についてはヴォーン等（1999）に詳しい。

定性的手法の研究は，主として自然言語をデータとして利用する。KIPもブログテキストという自然言語をデータとして用いる。また解釈学的アプローチでは理解を重視する。KIPも，最終的に分析結果を解釈しなければならない。これらの点でKIPはこれまでの研究と関係しているといえる。これらの関係を大まかに比べるならば，これまでの定性的手法の研究が一次資料の分析であるとすれば，KIPは大量の二次資料の分析に例えられる。

これら両者の関係は別途論じるべきだが，ここで確認しておきたいのは両者の詳細な関係ではない。KIPは最終的に解釈の部分を持つにしても可能な限り客観的な指標と手順に従う。つまり，KIPは見方・考え方の変化を可能な限り客観的な指標を使って定義し，その客観的指標から見方・考え方の変化を明らかにする。Kato and Ninomiya （2013）とKato（2013）はその方法として2分割型KIP（2 separated type KIP） を提案した。

2分割型KIP（2 separated type KIP）は，見方・考え方の変化（市場創造）を軸の変化と定義する。補足すれば，見方・考え方の変化を同一の軸の水準の変化ではなく軸自体の変化と定義する。KIPの概要で説明したように，KIPはラベリングとして主成分分析により軸を抽出する。市場創造の定義から，この軸が時間の経過の中で別の軸へ変化すれば，その時点で見方・考え方の変化（市場創造）が起こったといえる。この市場創造の時点の特定は，二宮（1977）で提案されたStepwise

Chow Test によって行える。

Stepwise Chow Test は，計量経済学において時系列あるいは順序のあるデータから回帰モデルを仮定して事前の知識なしに市場構造の転換点を明らかにできる手法として提案された。例えば，本田（1990），Takeuchi（1991），Nomura（1991）などで用いられている。また，Riddle（1978）と二宮（2009）で方法についての検討も行われている。本研究に関連して，特に重要なのは（1）「回帰モデルを仮定して」と（2）「事前の知識なしで」という 2 点にある。

「回帰モデルを仮定して」とは，Stepwise Chow Test の実行に当たり市場を捉える回帰モデルを構築しなければならないことを指している。回帰モデルの被説明変数は，マーケティングの成果を表す変数となる。例えば，商品やサービスの売上額，テレビ番組の視聴率，あるいは商品・サービスの普及率などがある。

他方で，説明変数は KIP の主成分分析により抽出された軸を用いる。この軸を説明変数にすることで，2 つのことがいえる。1 つに，説明変数の偏回帰係数の変化として市場創造の時点を特定できる。これは軸それ自体の変化を市場創造と呼ぶという定義と一貫したものになる。これに加えて，2 つに主成分軸同士は独立の関係にあり，理論上は説明変数間の相関が無くなり多重共線性が無くなる。

次に，「事前の知識なしに」とは，時系列データの変化点についての仮説を準備しなくてよいという意味である。したがって，時系列データさえあれば，そのデータからどこに市場創造の時点があるのかを明らかにできる。この特徴により，データから探索的に市場創造の時点を特定できる。

市場創造へ迫る手順を簡単にまとめると次のようになる。まず，収集したブログデータの全期間を対象にして，KIP により市場セグメンテーションと市場を捉える主成分軸を明らかにする。次に，この主成分軸を説明変数に，マーケティング成果変数を被説明変数にした回帰モデルを構築し，Stepwise Chow Test により探索的に市場創造の時点を特定する。

最後に，特定された市場創造時点の前後でブログデータを 2 分割して，それぞれの期間のブログデータに対して独立に KIP を実行し，それぞれの期間の市場を捉える主成分軸を抽出する。この 2 期間それぞれの市場を捉える軸の変化により市場創造を捉える。

これがマーケティングの中心的な課題の 1 つである市場創造へのアプローチとなる。このようにして，市場セグメンテーションと市場を捉える軸の抽出の方法として素直に用いるだけでなく，KIP はマーケティング研究の中に市場創造を扱える手法として位置づけられている。したがって，マーケティングにおけるデータサイエンス研究・教育の共通基盤として市場創造の研究・教育に生かせる方法となっている。

ここまでで，現状の KIP を巡る議論は一端整理できた。ただ KIP は既に完成した手法ではなく，今後解決すべき多くの課題を抱えた手法である。本研究の最後に，これまでのまとめと残された多数の課題を整理したい。

3．まとめと残された課題

（1）まとめ

本研究は，ここまででビッグデータやデータサイエンティストが注目されるようになってきたビジネスの現状に始まり，この現実への統計学者の立場の違いとデータを重視する共通の考え方を整理した。次に，米国におけるデータサイエンティスト養成教育プログラムが工学者の養成を想定しており，マーケティングの研究が無視されている点を指摘した。最後に，マーケティング既存研究に立脚したデータサイエンス研究・教育の共通基盤として KIP の利用を提案した。

このようなビッグデータを巡る背景を確認した上で，本研究の目的は次の 3 点に整理された。第 1 に，マーケティング研究者に向けて KIP の概要を整理する。第 2 に，マーケティング研究における KIP の位置づけを示す。第 3 に，今後の研究に向けて残された課題を整理する。これらであった。

これらの目的に対して，まずマーケティング既存研究に則したブログテキストマイニングの手順として提案された KIP の背景，その手順の概略，そして手順の中でも特にデータ収集方法について整理した。次に，市場創造の分析との関連から，マーケティング研究の中に KIP を位置づけた。これにより KIP は素直に市場セグメンテーションと市場を捉える軸の抽出の方法として利用できるだけでなく，マーケティングの中心的な課題の一つである市場創造の分析にも利用できることを示した。

このような整理により，本研究の 3 つの目的のうち 2 つにこたえた。だがKIPは既に十分に完成された手法であるというよりもむしろ，今後の更なる改善を必要とする手法である。以後，本研究の締めくくりとして，今後取り組むべき課題を整理する。

（2）利便性の改善

KIP に残された課題は大きく 3 つある。1 つ目の課題は，より使い易いプログラムを作るという観点からの残された課題である。2 つ目の課題は，現行のプログラム自体が抱えている課題である。これは現在使用しているプログラムをかなりの程度書き換える必要があるような種類の課題ともいえる。最後の課題は，マーケティング理論など関連研究を一層詳細に吟味して解決を目指さなければならない課題である。順に説明する。

1 つ目の課題は，より使い易いプログラムを作るという観点からの残された課題である。現在の KIP プログラムは事前の設定をきちんとすれば Shell script という言語により全自動で分析をするように作られている。だがこの事前の設定の部分がより簡素にできる。端的にはターゲットキーワードを複数の箇所に入力する必要があるなどの問題である。

この課題に分類される問題としてもう一つ具体的に示すと，Shell script により全自動化されているということはコマンド入力により実行する CUI であるという意味である。これを GUI に変更することで，より PC に詳しくない経験研究の蓄積をはかりたいマーケティング研究者にも使えるプログラムになる。これは利用者の裾野を広げるという観点から重要な貢献になる。

（3）分析プログラムの改善

2つ目の課題は，現行のプログラム自体が抱えている課題である。この2つ目の課題に属する課題は自覚的なものだけでもかなりの数に上る。その分だけ多くの優秀なマーケティング研究者の活躍の余地がまだまだ残されているともいえる。

第1に，データ収集に関して2つの課題がある。1つは，既にKIPの概要とデータ収集でも説明したように，gooブログからのみブログテキストデータを収集している。これを多様なデータソースから収集できるように改善する必要がある。gooブログに限定している背景にはその構造が独特であるという点を指摘した。すくなくとも，開設ブログ数の多いブログサイトから順に，データ収集の対象としていく必要がある。

2つは，少なくとも一度ターゲットキーワードを使用したブログオーサのみをデータ収集対象としている。その結果，ビックデータと呼ぶにはブログオーサ数が限られている。今後，ターゲットキーワードを使用したブログオーサだけでなく，ターゲットキーワードを使用していないブログオーサも含めて分析できなければならない。

このとき手がかりとなりそうなのは，石川等（2012）が示したブログデータの収集方法と現行のKIPのデータ収集方法を組み合わせることである。まず石川等（2012）の網羅的なデータ収集方法に加えて，ターゲットキーワードを使ったブログオーサのデータを入手する。次に，この両者のデータに対して，テキストマイニングで広く知られている著者の特定技術を利用する。著者の特定技術については，例えば金（2009，pp.178－202）で説明されている。

これにより，網羅的に収集したブログオーサ一人ひとりについて，ターゲットキーワードを用いたブログオーサに分類可能な程度を指標でとらえる。この指標を用いて，網羅的に収集したブログオーサの使用した単語頻度へ重み付けする。このようにすると，分析の対象とできるブログオーサ数を大きくできる可能性がある。

第2に，KIPの概要で説明したように，主成分分析により市場を捉える軸を抽出している。この主成分分析は軸にラベルを貼るために解釈をしなければならない。解釈には，当然のことだが主観が入る。この主観の余地を少しでも減らし，解釈を容易にする工夫をしなければならない。

1つのアイデアとして，主成分分析の結果として出てきた単語でブログテキストを検索する。この検索にヒットしたブログテキストを要約技術で要約する。このような方法がありえる。この方法を使うと，主成分軸を単語から解釈するのではなく自然言語で書かれた文章から解釈できる。

例えば，情報処理学会の学会誌「情報処理」の14巻12号ではテキスト自動要約の特集号が組まれている。その中でも特に，難波・奥村（2002）は複数テキストを対象とした要約技術について取り上げている。複数テキストは一種類だけのテキストを要約するのではなく，複数種類のテキストから一つの要約文を作成する。これらは複数のブログテキストから要約文を作成したいKIPと関連の深い技術である。

第3に，KIPの概要で説明したように，KIPはブログオーサのグループ化においてSOMを用いている。市場創造の分析のように2時点にブログデータを分割し，市場創造時点前後で分析を繰り返す。この方法でグループ化すると，市場を捉える軸に変化があったから新しい軸が出てきたのか，

それとも単に繰り返し SOM を実行した結果として前回と違う軸が出てきたのか判然としない可能性がある。この辺りについてはより詳細な検討を必要としている。

第４に，ブログテキストデータを市場創造時点前後で分割したとき，ターゲットキーワードを使用していないブログ記事のみの期間も出現する。そのとき，現行の方法はターゲットキーワードとの類似度の計算ができず，商品特性キーワードの選択ができない。これをターゲットキーワードの使用に関係なく分析できるようにするために，普遍的に利用できるターゲットキーワードベクトルを用いて分析できるようにする必要がある。

（４）マーケティング研究としての改善

最後の課題は，マーケティング理論など関連研究を一層詳細に吟味して解決を目指さなければならない課題である。この課題に分類される問題として，セグメンテーション基準と Stepwise Chow Test における回帰モデルについての２つがある。１つ目は，セグメンテーション基準についてである。現行の KIP は商品特性キーワードと人物特性キーワードの２つの基準で分類している。この基準での分類が最も望ましいのか，あるいは他の分類基準を設定することが望ましいのかを検討しなければならない。これは，市場セグメンテーションに関して，マーケティング既存研究を徹底的にレビューしなければならない。

２つ目は，市場創造の分析の中で出てきた回帰モデルについてである。現行の２分割型 KIP（2 separated type KIP）は全期間のブログテキストデータから抽出された主成分軸を説明変数にした回帰モデルを用いている。ただこのような回帰モデルは，マーケティング既存研究の視点から見て妥当なのかを吟味しなければならない。そのためには，特にマーケティングサイエンスやマーケティングモデリングといわれている計量経済学を基礎としたマーケティング既存研究の徹底したレビューが必要とされている。

以上のように，KIP は完成された手順というよりも，未だ改善しなければならない多数の課題を抱えている。今後これらの課題一つひとつに取り組んでいかなければならない。こうした課題はあるものの，マーケティングにおけるデータサイエンス研究・教育の共通基盤としてこれから知の蓄積をはかる基盤としての役割は期待できる。

（かとう・じゅんいち　メディア社会学科）

註

1　ビッグデータの定義は未だ定まっていない。本研究は McAfee and Brynjolfsson.（2012）の（1）量（Volume），（2）速さ（Velocity），そして（3）多様さ（Variety）という意味で用いている。例えば KIP は厳密にこの３つを満たしていない。この点は本文中の今後の課題としても取り上げられており，今後より改善していくべき課題とされる。

2　データサイエンティストの定義は未だ定まっていない。ただ現在我が国でもデータサイエンティスト協会が設立され，定義や必要要件の検討に向けて準備が進められている。データサイエン

ティスト協会についての情報は次のURL（ http://www.datascientist.or.jp/ 2013年9月29日確認）を参照していただきたい。

3　McAfee and Brynjolfsson. (2012, p. 62) は「Business executives sometimes ask us, “Isn’t ‘big data’ just another way of saying ‘analytics’?” It’s true that they’re related: The big data movement, like analytics before it, seeks to glean intelligence from data and translate that into business advantage. However, there are three key differences:」と述べ，その後に本文中で示した量，速さ，そして多様さの3点を指摘している。

4　McAfee and Brynjolfsson.（2012, p. 62）は，量（Volume）について「As of 2012, about 2.5 exabytes of data are created each day, and that number is doubling every 40 months or so. More data cross the internet every second than were stored in the entire internet just 20 years ago. This gives companies an opportunity to work with many petabytes of data in a single data set – and not just from the internet.」と指摘している。

5　McAfee and Brynjolfsson. (2012, p. 63)は，速さ（Velocity）について「For many applications, the speed of data creation is even more important than the volume. Real-time or nearly real-time information makes it possible for a company to be much more agile than its competitors.」と述べている。

6　McAfee and Brynjolfsson. (2012, p.63)は，多様さ（Variety）について「Big data takes the form of messages, updates, and images posted to social networks; readings from sensors; GPS signals from cell phones, and more. Many of the most important sources of big data are relatively new.」と説明している。

7　McAfee and Brynjolfsson. (2012, p. 64)は，ビッグデータの経営における意義に関連して「In particular, companies in the top third of their industry in the use of data-driven decision making were, on average, 5% more productive and 6% more profitable than their competitors. This performance difference remained robust after accounting for the contributions of labor, capital, purchased services, and traditional IT investment. It was statistically significant and economically important and was reflected in measureable increases in stock market valuations.」と述べている。

8　Barton and Court (2012, p. 80)は，「First, companies must be able to identify, combine, and manage multiple sources of data.」と述べている。また別の箇所でBarton and Court (2012, p. 80)は「Often companies already have the data they need to tackle business problems, but managers simply don’t know how the information can be used for key decisions.」あるいは「Managers also need to get creative about the potential of external and new sources of data.」とも述べている。

9　Barton and Court (2012, p. 80)は，「Second, they need the capability to build advanced analytics models for predicting and optimizing outcomes.」と説明している。また別の箇所でBarton and Court (2012, p. 81) は「Data are essential, but performance improvements and competitive advantage arise from analytics models that allow managers to predict and optimize outcomes.」

とも述べている。

10　Barton and Court (2012, p. 80)は，「Third, and most critical, management must possess the muscle to transform the organization so that the data and models actually yield better decisions. Two important features underpin those activities: a clear strategy for how to use data and analytics to compete, and deployment of the right technology architecture and capabilities.」と述べている。

11　Davenport and Patil (2012, p. 72) は，「the “data scientist”. It’s a high-ranking professional with the training and curiosity to make discoveries in the world of big data.」と述べている。

12　Davenport and Patil (2012, p. 72) は，「While those are important breakthroughs, at least as important are the people with the skill set (and the mind-set) to put them to good use. On this front, demand has raced ahead of supply. Indeed, the shortage of data scientists is becoming a serious constraint in some sectors.」と指摘している。

13　Davenport and Patil (2012, p. 73) は，「What kind of person does all this? What abilities make a data scientist successful? Think of him or her as a hybrid of data hacker, analyst, communicator, and trusted adviser. The combination is extremely powerful – and rare.」と述べている。さらに Davenport and Patil (2012, p. 73) は，「But we would say the dominant trait among data scientists is an intense curiosity – a desire to go beneath the surface of a problem, find the questions at its heart, and distill them into a very clear set of hypotheses that can be tested.」とも説明している。

14　Davenport and Patil (2012, p. 74) は，「A little less surprisingly, many of the data scientists working in business today were formally trained in computer science, math, or economics. They can emerge from any field that has a strong data and computational focus.」と指摘している。

15　Davenport and Patil (2012, p. 76) は，「Data scientists today are akin to Wall Street “quants” of the 1980s and 1990s. In those days people with backgrounds in physics and math streamed to investment banks and hedge funds, where they could devise entirely new algorithms and data strategies.」と指摘している。

16　Davenport and Patil (2012, p. 76) は，「Then a variety of universities developed master’s programs in financial engineering, which churned out a second generation of talent that was more accessible to mainstream firms. The pattern was repeated later in the 1990s with search engineers, whose rarefied skills soon came to be taught in computer science programs.」との過去の金融工学やコンピュータ科学の事例を示している。

17　Davenport and Patil (2012, p.76) は，「One question raised by this is whether some firms would be wise to wait until that second generation of data scientists emerges, and the candidates are more numerous, less expensive, and easier to vet and assimilate in a business setting. 」と述べている。さらに，Davenport and Patil (2012, p. 76) は，「The problem with that reasoning is that the advance of big data shows no signs of slowing.」と指摘している。

18　Davenport and Patil (2012, p. 76) は，「If companies sit out this trend’s early days for lack of

talent, they risk falling behind as competitors and channel partners gain nearly unassailable advantages. Think of big data as an epic wave gathering now, starting to crest. If you want to catch it, you need people who can surf.」と指摘している。

19　日本経済新聞（2013）は「大量のデータを分析して経営に役立てる「データサイエンティスト」という職種が脚光を浴びている。IT（情報技術）のスキルだけでなく，統計学や業務知識，ビジネスのセンスが必要で，これからの企業の成長を左右するともいわれる。」と指摘している。

20　週刊ダイヤモンド（2013，p. 64）は「ここでいう統計家とは「データサイエンティスト」とも呼ばれる職種。大量のデータを分析し，そこから有益な結果を導き出すことで，ビジネス上の価値創出のシナリオまで描けるような，データ活用の専門家である。数理統計などを用いた分析スキルとコンピュータプログラミングのスキルを持つ一方，経営やマーケティングなどに関するビジネススキルも兼ね備えた人物像を指す。」と説明している。

21　週刊エコノミスト（2013，p. 25）は，西内氏の「統計学を知る立場から「ビッグデータにだまされるな」と説明する必要があると考えて『統計学が最強の学問である』を書いた」という記事を掲載している。

22　西内（2013，p. 47）は「もちろん全数調査よりサンプリング調査の方が精度が低いことは間違いない。だが問題となるのは，それによってどの程度精度が低下するのか，そしてその精度が低下した結果，実際に下すべき判断や取るべき行動にどのような影響があるのかということである。逆に言えば，判断や行動に影響しないレベルの精度は無意味で，そのためにかけなければいけないコストはムダだ。対処しきれない量のデータが存在する際に，適切なサンプリングさえすれば，必要な情報を得るためのコストが激減するのは80年前だろうが現代だろうが本質的には変わらない。にもかかわらず，ビッグデータに関心のあるビジネスマンは，しばしばビッグデータをビッグなままで扱うことにしか目が行かないのだ。」と述べている。この主張と，マイヤー＝ショーンベルガー・クキエ（2013）と Cukier and Mayer-Schoenberger（2013）のデータを収集した後から分析の切り口を考えられるとの主張とを比較することで，両者の立場の違いはより鮮明になってくる。加えて，マイヤー＝ショーンベルガー・クキエ（2013，p. 68）のように，ビッグデータを支持する立場から精度を問題としていないとの主張のある点も興味深い。

23　西内（2013，p. 64）は，「ビッグデータ技術を使って全数データを使った単純集計しかしないというのは，最新の技術を２世紀前の手法でしか活用できていないということである。これではまるで，最新のスマートホンを金づち代わりにして犬小屋を作ろうとするようなものではないだろうか。」と述べている。加えて，西内（2013，p. 59）は，統計学の知識を活用することで答えるべき問として(1) 何かの要因が変化すれば利益は向上するのか，(2) そうした変化を起こすような行動は実際に可能なのか，(3) 変化を起こす行動が可能だとしてそのコストは利益を上回るのか，の３つをあげている。

24　西内（2013）は，58頁から210頁までの３章にわたって因果関係の特定に関連した主張に割いている。

25　ランダム化実験の重要性に関連して，西内（2013，p. 116）は「ランダム化してしまえば，比

較したい両グループの諸条件が平均的にはほぼ揃う。そして揃っていない最後の条件は実験で制御しようとした肥料だけであり，その状態で両グループの収穫量に「誤差とは考え難い差」が生じたのであれば，それはすなわち「肥料が原因で収穫量に差が出る結果になった」という因果関係がほぼ実証できたと言えるだろう。」と述べ，さらに西内（2013，p. 116）は「倫理的にも予算的にも実験が許されるものである限り，ごちゃごちゃ理屈を唱えるよりもとりあえず研究参加者をランダムに分けて，異なる状況を設定し，その差を統計学的に分析してしまえばいいのだから，これほど分かりやすく強力な研究方法はない。」と述べている。

26　ランダム化できない３つのパターンとして，西内（2013，p. 126）は「世の中にはランダム化を行うこと自体が不可能な場合，行うことが許されない場合，そして行うこと自体は本来何の問題もないはずだが，やると明らかに大損をする場合，という３つの壁がある。」と述べている。

27　ランダム化ができないときでも因果関係を特定する方法の１つとして，西内（2013，p. 140）はケースコントロールについて「疫学におけるケースとは症例すなわち関心のある病気となった事例（患者）のこと。そしてコントロールとはその比較対照のことである（ちなみに「比較対照」は疫学の専門用語。「比較対象」ではない。）比較対照には「関心のある疾患とリスク要因の有無以外は条件がよく似た人」が選ばれる。「よく似た」の定義は研究によってさまざまだが，関心のあるリスク要因以外は考える限りすべての条件について同等であることが望ましい。だからドールとヒルは，喫煙というリスク要因以外の肺がんと関連しうる条件である，性別・年代・社会階層・居住地域といったものについて，調査対象とした患者と同様の人間を集めて男女別や年代別で区切ったグループごとに比較（専門用語でこれを層別解析と呼ぶ）すれば，ランダム化をしなくても「フェアな比較」ができるというのである。」と説明している。

28　ランダム化ができない状況下でもフェアな比較（因果関係の特定）ができる方法として，西内（2013）は回帰分析について次のように述べている。西内（2013，p. 185－6）は「性別の違いにより平均で何点違うか，という回帰係数と，高校によって平均で何点違うか，という複数の回帰係数を同時に推定するのが重回帰分析である。性別の違いによって，「平均で何点違うか」という影響の度合いが推定できれば，男子同士・女子同士という層別で比較しなくても，「もし仮にこの男子が全員女子だったら」と仮想的に条件を揃えた状態でフェアな比較をしていることになるのだ。これが重回帰分析によってフェアな比較が行われたということである。こうしたやり方であれば，多少条件が増えたとしても莫大な数の層に分ける必要はない。複数の回帰係数は「お互いに相乗効果がなかったとすれば」という仮定のもと，説明変数が結果変数にどの程度の影響を与えるかを示している。」と述べ，ランダム化できない状況でのフェアな比較方法としての回帰分析の有効性を説明している。

29　Cukier and Mayer-Schoenberger (2013, p. 29)は「The first is to collect and use a lot of data rather than settle for small amounts or samples, as statisticians have done for well over a century.」と指摘している。

30　マイヤー＝ショーンベルガー・クキエ（2013，p. 51）は「長い間，無作為標本は，なかなか優れた簡便法だった。」と述べている。

31　マイヤー＝ショーンベルガー・クキエ（2013，p. 51）は「デジタル化以前の時代に大量データの分析を可能にした実績がある。しかし，デジタルの画像や楽曲のファイルサイズを小さくするために圧縮する作業と同じで，標本作成の際には情報が抜け落ちる。一方，完全（あるいはほぼ完全）なデータセットなら，もっと自由に調査できるし，別の角度からデータを眺めたり，特定部分をクローズアップしたりすることも可能だ。」と述べている。

32　データを集めてからどのような人に注目するのか考えられるという点に関連して，Cukier and Mayer-Schoenberger (2013, p. 30) は「But it falls apart when we want to drill down into subgroups within the sample. What if a pollster wants to know which candidate single women under 30 are most likely to vote for? How about university-educated, single Asian American women under 30? Suddenly, the random sample is largely useless, since there may be only a couple of people with those characteristics in the sample, too few to make a meaningful assessment of how the entire subpopulation will vote. But if we collect all the data -"n=all," to use the terminology of statistics- the problem disappears.」と述べている。

33　Cukier and Mayer-Schoenberger (2013, p. 29) は，「The second is to shed our preference for highly curated and pristine data and instead accept messiness: in an increasing number of situations, a bit of inaccuracy can be tolerated, because the benefits of using vastly more data of variable quality outweigh the costs of using smaller amounts of very exact data.」と指摘している。

34　マイヤー＝ショーンベルガー・クキエ(2013，p. 68)は「膨大なデータがある場合，全般的な傾向が推測できさえすれば，精度や正確さはもはや最終ゴールではないケースもある。周囲を見回せば，そんな皮肉な現象はいくらでもある。」と指摘している。具体的な例として，マイヤー＝ショーンベルガー・クキエ（2013，p. 64－65）はGoogleの翻訳の仕方を示している。

35　マイヤー＝ショーンベルガー・クキエ（2013，p. 78）は「それはまるで印象派絵画と同じで，一筆一筆をこと細かに見れば乱雑であっても，一歩引いて眺めれば壮大な全体像が浮かび上がってくる。」と指摘している。

36　Cukier and Mayer-Schoenberger (2013, p. 29) は，「Third, in many instances, we will need to give up our quest to discover the cause of things, in return for accepting correlations.」と指摘している。更に同じp. 29では「Big data helps answer what, not why, and often that's good enough.」とも述べている。これらの主張は因果関係の発見を否定しているものではない。発見できるなら良いがそれは困難だし，発見できなくとも有益であり得るという考えを示している。これはCukier and Mayer-Schoenberger (2013, p. 32) の「Of course, knowing the causes behind things is desirable. The problem is that causes are often extremely hard to figure out, and many times, when we think we have identified them, it is nothing more than a self-congratulatory illusion.」との主張からも理解できる。

37　マイヤー＝ショーンベルガー・クキエ(2013，p. 89)は「このように仮説を立てては試行錯誤の繰り返しで人類の知は進化を遂げてきた。煩わしいことこの上ないプロセスだが，スモールデータの世界ではこれで通用していたのだ。ビッグデータの時代になれば，「もしや」というひらめき

から出発し，特定の変数同士をピックアップして検証するといった手順はもはや不可能だ。データ集合があまりに大きすぎるし，検討対象となる分野も恐らくずっと複雑になる。幸いなことに，かつて仮説主導型にせざるを得なかった制約も，今はほとんどない。これほど大量のデータが利用でき，高度な計算処理能力があるのだから，わざわざ手作業で相関のありそうな数値を勘でピックアップして個別に検証する必要などない。高度な計算解析を駆使すれば，最も相関の高い数値を特定できるのだ。」と述べている。

38　もちろんこうしたデータがものをいう時代を無批判に肯定しているわけではない。例えば，マイヤー＝ショーンベルガー・クキエ（2013）は第８章でリスク―ビッグデータのマイナス面として危険性を指摘しているし，Cukier and Mayer-Schoenberger.（2013, p. 37–40）はデータへの依存の危険性を指摘している。

39　http://idse.columbia.edu/　2013年９月４日確認

40　http://idse.columbia.edu/masters　2013年９月４日確認

41　http://jp.wsj.com/article/SB10001424127887323779204579041962427698246.html
2013年９月４日確認

42　2013年９月２日㈪に開催された，統計教育大学間連携ネットワーク FD 講演会「ビッグデータ時代の統計教育」における「Challenges and Opportunities for Statisticians in the Era of Big Data」と題された講演

43　例えば，Chen et al.（2009）はセグメンテーションといいつつも，マーケティング研究を参考文献にあげていない。

44　マーケティング研究で広く用いられてきた質問紙法との比較検討については例えば加藤・石川（2011b）を参照していただきたい。

45　Kato（2010）はそれぞれの変数を中村（2008）に依拠してより詳細に説明した上で，図表中の変数一つひとつを更に詳細に表（p. 4とp. 28, Table 2-2）にまとめているので参照していただきたい。

46　石川等（2012，p. 43）は，goo ブログ記事に限定する意義と他のブログサイトにおいても基本的に同様の手法が応用できることを述べている。

47　KIPを用いた個別研究は次のURLに整理されている。
（https://sites.google.com/site/junichikatopapers/　2013年10月１日確認）

48　石井（1993，p. 249）は，「利益や価値の源泉は，これまで考えられてきたように，果たして生産活動や労働活動に帰着できるものなのだろうか。むしろ商人あるいはマーケターとして，共時的・通時的な異文化性を媒介とすることによって，そして恐らくは自ら異文化を創り出すことによって，初めて差異が生まれ，そこに差額，つまり利益や価値が生み出されるのではないだろうか。その活動を担うのは，いうまでもなくマーケティングであり，その意味でマーケティングの課題はすぐれて異文化間で偶然的に始まる「最初の交換」の完成にあるということもできる。経済学であればともかく，マーケティングの世界において，定常状態での静態的な交換モデルに留まってはいられない理由はそこにある。」として市場における異文化性の媒介と創造に言及してい

る。この「マーケティングの世界において，定常状態での静態的な交換モデルに留まってはいられない」というところからも，石井（1993）が静態的なモデルではなく動態的なとらえ方の重要性を指摘していると理解できる。なお，石井（1993）は既に絶版となっており，代わりに同じ内容で石井（2004），『マーケティングの神話』，岩波文庫が入手可能である。

49　例えば，石井（1993，p. 248）の「異文化のはざまで利益を搾取するだけの昔の「商人」とは違い，現代の「マーケター」は，異文化を媒介するだけでなく，それを積極的に創造する活動にもかかわっていることをとくに強調しておかなければならない。われわれの周囲を見ても，多様な文化カテゴリーが急速に増えてはいないだろうか。現代のマーケターは，ジェンダーをはじめとして，都市生活者，大人と子供，ニューファーミリー，新人類，あるいは学生といった「文化」カテゴリーを「つくりだす」役割も引き受けるという点が一層重要である。それは，文化カテゴリーが次々に生み出されていることを意味している。それら新しい文化と自らの既存の生産組織の文化とのギャップに，新たな利益機会を発見するのである。同じような意味で，未来のニーズを開発することを通じて，未来の市場から利益を得ることも現代のマーケターの役割である。まだ生活に定着していない属性を持った新製品の活発な市場導入や，多くの人がまだ生活にはなじんではいないニュー・ファッションの提案等といった試みは，そのためのものではなかったか。」のような所からも伺える。

50　石井（1993，p. 278）は，解釈主義アプローチに関して「解釈主義者の基本的な立場を実証主義者との比較において示してきたが，両者の最も顕著な差異は，実証主義者の志向が「説明と予測」にある一方で，解釈主義者の基本的立場が「理解」であるということである。」と述べている。

参考文献

Barton, D. and D. Court. (2012), “Making Advanced Analytics Work For You”, *Harvard Business Review,* October, 2012, pp. 79–83.

Chen, L-S., C-C HSU., and M-C. Chen. (2009), “Customer Segmentation and Classification from Blogs by Using Data Mining”, *Cybernetics and Systems,* Vol. 40, No. 7, pp. 608–632.

Cukier, K. and V. Mayer-Schoenberger. (2013), “The Rise of Big Data”, *Foreign Affairs,* Vol. 92, No. 3, pp. 28–40.

Davenport, T. H. and D. J. Patil. (2012), “Data Scientist: The Sexist Job of the 21st Century”, *Harvard Business Review,* October, 2012, pp. 70–76.

本田豊（1990），「Chow テストによる日本経済の構造変化」，『立命館経済学』，第39号 3 巻，pp. 49-73。

石井淳蔵（1993），『マーケティングの神話』，日本経済新聞社。

石川雅弘，池田潔，加藤淳一（2012），「ブログ記事の収集と予備分析」，『研究紀要』，18巻，41-55頁。

金明哲（2009），『テキストデータの統計科学入門』，岩波書店。

Kato, J. (2010), “Procedure of Weblog Data Analysis for Market Segmentation”, つくば国際大学産業社会学

部ワーキングペーパー, No. 5, pp. 1-33. http://www.ktt.ac.jp/tiu/department/is-workingpaper/ProcedureofWeblogDataAnalysisforMarketSegmentation.pdf

加藤淳一，石川雅弘（2011a），「大量ブログ記事をデータとした市場セグメンテーションの半自動的分析手法」，『2011年春季研究発表会アブストラクト集（日本オペレーションズ・リサーチ学会）』，104－105頁。

加藤淳一，石川雅弘（2011b），「iPad の市場セグメンテーション：ブログテキストを用いた消費者ニーズの解明」，『日本 MOT 学会　第 2 回年次研究発表会』，25－28頁。

加藤淳一（2013），「ブログテキストマイニングによる海外観光都市に関する消費者ニーズの探索的調査：モナコ公国を事例に」，『研究紀要』，19巻，35－50頁。

Kato, J. (2013), “Two-separated type KIP: Comparing principal axes before and after the change point of a market”, 『2013年秋季研究発表会アブストラクト集（日本オペレーションズ・リサーチ学会）』, pp.178-179.

Kato, J. and S. Ninomiya. (2013), “Detecting the Changing Points of Multiple-Regression Model on the Basis of the Relations between Audiences’ Rating and the Matching between Needs and Contents”, 『知識共創』, 第 3 号, pp. Ⅲ3-1-6 （6 pages）.
http：//www.jaist.ac.jp/fokcs/papers/G－paper－Kato.pdf
https：//sites.google.com/site/junichikatopapers/home/programs

Kotler, P. and K. Keller. (2006), *Marketing Management,* 12th Edition, Pearson Prentice Hall.

ルイス, M.（2006），『マネー・ボール』，ランダムハウス講談社。

Manning, C. D., P. Raghavan., and H. Schutze（2008）, *Introduction to Information Retrieval,* Cambridge University Press.

McAfee, A. and E. Brynjolfsson. (2012), “Big Data: The Management Revolution”, *Harvard Business Review,* October, 2012, pp. 61–68.

マイヤー＝ショーンベルガー, V，K. クキエ.（2013），『ビッグデータの正体』，講談社。

難波英嗣，奥村学（2002），「ここまで来たテキスト自動要約」，『情報処理』，43巻，12号，1287－1294頁。

西内啓（2013），『統計学が最強の学問である』，ダイヤモンド社。

中村博　編著（2008），『マーケット・セグメンテーション』，白桃書房。

日経情報ストラテジー（2013），「特集：データサイエンティスト」，2013年 6 月，22－35頁，日経 BP 社。

日本経済新聞（2013），「大量データ経営に生かす」，2013年 3 月26日（火曜日），35面。

二宮正司（1977），「Stepwise Chow Test」，『季刊理論経済学』，Vol. 28，No. 1，pp. 50－60。

二宮正司（2009），「Stepwise Chow Test 再論」，『大阪経大論集』，第60巻第 4 号，1－16頁。

Nomura, M. (1991), “The Displacement Effect on Government Expenditure of Two Oil Crises: Japan, the United Kingdom and the United States,” *The Manchester School of Economic and Social Studies,* Volume 59, No. 4, pp. 408-418.

Riddle, W. C. (1978), “The Use of the Stepwise Chow Test.”, *The Economic Studies Quarterly,* Vol. 29, No. 3, pp. 242-247.

週刊エコノミスト（2013），「使える統計学」，2013年6月4日，22-45頁，毎日新聞社。

週刊ダイヤモンド（2013），「特集：最強の武器「統計学」」，2013年03月30日，34-64頁，ダイヤモンド社。

Takeuchi, Y. (1991), “Trend and Structural Changes in Macroeconomic Time Series,” *Journal of the Japan Statistical Society,* Vol. 21, No. 1, pp. 13–25.

ヴォーン，S，J. S. シューム，J. シバグブ（1999），『グループ・インタビューの技法』，慶應義塾大学出版会。

Wedel, M. and W. A. Kamakura (2000), *Market Segmentation,* 2nd Edition, Kluwer Academic Publishers.

イン，R. K.（1996），『ケーススタディの方法』，千倉書房。

KIP as the Platform of Data Science Research and Education in Marketing

Junichi Kato

Nowadays, researchers and business persons pay attention to the big data and data scientists and point out the importance of them. Having nothing to do with their standing-points about big data, researchers share the notion of the data centric era. In other words, engineers lead the academic development of this field even though the field is related to business.

When we face this reality, this paper proposes to set KIP as the platform of data science research and education in marketing. KIP is the procedure to execute market segmentation and to extract the principal axes for explaining the markets by using blogs' texts as data. In this paper, the background and procedure of KIP are explained.

Even if marketing researchers and business persons cannot create programs like engineers, because the platform is already prepared marketers can contribute from their professional fields. In addition, the contributions of marketers are accumulated without scattering and losing them. Such a merit is assumed.

Finally, remaining future research problems are shown as follows. The first problem is the improving of the computer programs for users' facilities. The second problem is the limitations of the programs themselves. In other words, this type of problem needs rewriting of the some parts of the programs. The last problem is to need to scrutinize the existing papers of marketing theory and its related researches. These problems and ideas for solving these problems are summarized.

Keywords: KIP, Data science, Big data, Marketing