データサイエンスの忘れ物

松本　渉*

要　旨

データサイエンスが注目されている．しかし，データサイエンスが語られる際，データの分析手法へ目を奪われがちである．データには収集されるプロセスがあり，データサイエンスの真の理解には，収集プロセスの理解も重要である．本稿では，あらためてデータのサイエンスとは何かについて，林知己夫が提唱した「データの科学」の意義を確認することを通じて，分析手法バイアスともいうべき，データ分析手法へ傾倒しがちな風潮に警鐘を鳴らす．さらに，選挙情勢調査と世論調査の現状を検討することを通じて，データ収集プロセスへの真摯な取り組みが必要であることを主張する．

キーワード：データの科学，選挙予測，世論調査

Lost Property of Data Science

Wataru MATSUMOTO**

Abstract

Data science is currently an emerging field. There seems to be a stream that it means data analytical methods. Genuine understanding of data science requires knowledge of both data collection process and data analytical methods. Thus, this study reconsiders the significance of Chikio Hayashi's definition of "data science" and warns against relying entirely on data analysis methods which should be referred to as "analysis method bias." By reviewing the current state of pre-election polls and public opinion research, this study emphasizes that a sincere approach to the data collection processes is of great significance.

Keywords: data science, pre-election poll, public opinion research

*関西大学総合情報学部

**Faculty of Informatics, Kansai University

１．はじめに

大学教育においてデータサイエンスという言葉が広がっている．「データサイエンス学部」が，2017年に滋賀大学，2018年に横浜市立大学，2019年に武蔵野大学，2021年に立正大学において設立されている．関西大学では2021年度から，全学的な取り組みとして，「数理・データサイエンス・AI」に関する知識を習得できる全学共通カリキュラムが開始され，さらに総合情報学部では学部独自の「データサイエンス教育プログラム」も設置されている．個別の大学や学部などで，データサイエンスを冠するカリキュラムが広まっていく理由としては，政府が推進するモデルカリキュラム（教育プログラム）の認定をうける必要性や，受験生や在校生へのアピールという広報目的もあると予想されるが，大きな背景として，データサイエンスそのものの教育ニーズが高まっているという社会の変化があることはいうまでもない．

このことは，上記のような大学教育上の動きが生じるよりも少し前の2013年に，一般社団法人のデータサイエンティスト協会が一般社団法人として発足していたことからも理解できる．実際，同協会のホームページによれば[1)]，データサイエンティスト協会設立の目的は，「データサイエンティスト」についての人材の需給ミスマッチを解消することとされている．

では，データサイエンティストとはどのような人材であろうか．同協会はデータサイエンティストそのものに明確な定義がないという立場をとっているが，*Harvard Business Review* 誌に寄稿された Davenport & Patil（2012）の論文（日本語版は2013年刊行）や，日本学術会議情報学委員会 E−サイエンス・データ中心科学分科会（2014）の提言では，データサイエンティストに期待される主要な能力として，ビッグデータの活用が想定されている．結果として，コードの記述とデータの分析のための人材というのが全体的な論調であるようにも読みとれる．

筆者は，このような人材の重要性を否定するものではない．ビッグデータを取り扱う人材育成上そのようなニーズが高まる事情は理解できる．気になるのは，データサイエンスが語られる時，ややもすれば，どのようにデータを活用するか，あるいは分析するかといった視点に偏っており，データそのものが所与のものであるかのように議論されている点である．実際のデータは，何らかの対象から収集されるというプロセスがあり，収集されたデータから分析を通じて有益な情報が引き出されることを考えれば，データサイエンスにおけるデータ収集プロセスの理解は，分析と同様に重要なはずだからである．

そこで，本稿では，データ収集プロセスを理解することの重要性を強調したい．そのために，まずデータサイエンスとは，そもそも何なのかという原点に立ち返る．あらかじめ述べておくと，林知己夫の「データの科学」の提唱の経緯と考え方を確認する．その上で，古典的な事件が示す教訓も含め，データ収集プロセスの重要性を示す事例に言及する．

1） 設立の背景　https://www.datascientist.or.jp/about/background/

2．データサイエンスの源流

データサイエンス（Data Science）という言葉は，よく使われる二つの単語から構成されるため，誰が最初に口にしたかを厳密には明らかにすることはできない．文章中で明確に用いた人物や用いられた場面ついても諸説あるかもしれない．しかし，数理・データサイエンス・AI教育プログラム認定会議の委員を務めた椿広計氏が『横幹』誌においてデータサイエンスをめぐる大きな潮流をまとめた「システム科学とデータ科学」には重要な手がかりが示されている．そこでは，「データサイエンス」という用語を日本の統計学コミュニティが用いるようになったのは，1996年３月に神戸で開催された国際学会「第５回国際分類学会連合」（IFCS-96）であると指摘されている（椿，2020）．Data Scienceをキャッチフレーズとした当該国際学会（IFCS-96）では，林知己夫による講演「What is Data Science? Fundamental Concepts and a Heuristic Example」（データの科学とは何か――根本理念と一つの説明例――）がなされ，その中でデータの意味や品質も含めたデータの計画（design），データの収集，データの解析の３つからなる新しい概念としてData Scienceが明確に提唱されたのである（林，1996；Hayashi, 1998）．

もちろんData Scienceは，何の準備もなしにこのとき唐突に誕生したのではない．この講演で概念として提唱されるに至るまでの経緯があったことにも注意がいる．実はそもそもの発端は，林らが，数理統計学へ対抗したいということでフランスのJ. P. Benzécriらと共鳴して，1987年に東京で第１回日仏セミナーを開催したことにある（林，1996）．このセミナーにおいて，データ解析を発展させることとし，そのための新しい概念としてData Scienceが命名されている（林，1996；Escoufier et al., 1995）．その後，1992年にフランスのモンペリエで，Data Scienceをキャッチフレーズに第２回日仏セミナーが開催されることになり，そこでは「データに関する包括的な方法論」をData Scienceの中心とすることが定められたのである（林，1996）．

第２回日仏セミナーの成果をまとめた書物（Escoufier, Y. et al., 1995）の序文「そのようなアプローチは，その核心にデータについての新しい科学を生み出す．…（中略）…この活動に対して，データの科学data scienceという用語を用いるのが妥当なようである．」（Escoufier, et al., 1995, viii, 26-30行目，訳は筆者）からは，この時点でもまだ概念化の途上であることが伝わってくるが，このような活動を踏まえて，前述のIFCS-96におけるData Science（データの科学）概念の提唱につながったことは確かである．日本学術会議情報学委員会E－サイエンス・データ中心科学分科会（2014）の言い方を借りれば，これを機に（data scienceという）「日本発の用語が国際的に広まる」（p.11）ことになったと言える．

ただし，林知己夫自身，日本語でData Scienceを表現する際は，「データサイエンス」や「データ科学」ではなく，「データの科学」という表現を好んで用いていたことが知られている．この意図は，「データ科学」というと狭い意味の専門用語となっていくことを恐れたためである．ニュアンスを誤解なく伝えるために，『データの科学』（林，2001）の「序にかえて」の記述を

そのまま引用する.

> いわゆる「実験計画法」(experimental design)が「実験の計画法」のごく一部のテクニカルタームとしての特異な方法を指すことになっていることに鑑み,「の」を入れたいと思っている.実験計画法はある特種(原文ママ)な現象に対して意味をもつものとなっており,今日では形骸化して「実験の計画法」とほど遠いものになっている.
>
> (林, 2001, iii, 19-23行目)

ただし,「データ科学」という言い方をことさら毛嫌いしていたというわけではないという点も付記しておきたい.実際,前掲した『データの科学』(林, 2001)の「序にかえて」には,次のような記述がある.

> これは私個人の「思い」であって,本書で「データ科学」と使われていても,今日の状況では一向にさしつかえない.
>
> (林, 2001, iii, 28-29行目)

とはいえ,「データサイエンス」という表現を林知己夫自身が用いた例は確かに少なく,文中ではデータの科学か, Data Science という英語を用いている.しかし,『日経BPムック:データウエアハウスがビジネスを変える』に寄稿している「データ解析からデータサイエンスへ」(林, 1996)のタイトルではカタカナのデータサイエンスを利用している.厳格に避けていたというわけではない.

いずれにせよ,林知己夫のデータの科学/Data Science には,概念としての寛容さが内包されていた.狭い概念に抑制せず,やや広がりのある概念を志向していたのである.この点に注意しつつ, IFCS-96以降の,林(ら)の著作から「データの科学」の概念的特徴を拾っていくこととしよう.

IFCS-96の翌年, 1997年に刊行された『社会調査と数量化(増補版)』(林・鈴木, 1997)では,サブタイトルが初版本(林・鈴木, 1986)の「国際比較におけるデータの解析」から「国際比較におけるデータの科学」に変化しており,「データの科学」提唱の影響が明確に見られる.本文においても,新たな1章が追加され,そこでは「「こうかも知れない,ああかも知れない,ここは間違っている,これで少し見えてきた」というプロセスを通し,データに基づいて(data-driven)探索的に情報を取出す(exploratory approach)」(林・鈴木, 1997, pp. 278-279)ことを望ましいとし,データの科学を,データによって現象を理解することを目的とする「統計学,データ解析,分類・統合やそれに関連した諸方法を統一的に集約する概念」(林・鈴木, 1997, p. 279)と位置づけるようになったのである.

以上の経緯を踏まえ,改めて林知己夫の考えた「データの科学」(林, 1996;林・鈴木, 1997,

Hayashi，1998；林，2001）の概念的特徴を筆者なりに整理すると，次の２つに集約される．

１．仮設発見的立場

データによる現象理解を重視しており，theory-driven よりも data-driven を志向している．仮説検証的立場をとらなかったのは，曖昧で複雑な現象においては，精密科学（exact science）のように因果関係を示せるものではないと考えたためである（林，2001，p.114）．

ただし，これは今日のビッグデータの解説に見られるような相関関係さえ分かれば因果関係は不要といった議論（Mayer-Schönberger and Cukier, 2013）とは全く異なっているので注意がいる．林（2001）自身，（科学には）「因果関係の追求はとりあえず大事」（p.3）と述べており，「因果関係らしきもの」で役に立つ（林，2001，pp.114-116）という風に考えていたからである．

２．計画（design）・実施（collection）・分析（analysis）の三段階による構成

この点は，IFCS-96の講演ですでに示された内容である．計画（design）・実施（collection）については，日本語と英語の対応関係が独特であるが，両者を合わせて，「データをとること」（林，2001，p.21）と表現している場合もある．結局，このことは「データの科学は調査の科学を包含する概念である」（林，2001，p.8）という概念の広さにもつながっている．

林の「データの科学」を踏襲するのであれば，今日のデータサイエンスにおいても，データの分析だけでなく，データの収集にも同等の比重で目をむけるべきである．そこで３節以降では，「データを妄信せず，その由来をたずね」（林，2001，p.6）ることの重要性を（古典的なものも含め），選挙予測に関する選挙情勢調査と世論調査を糸口として，問題提起を行う．

この２つを糸口とするのは，まずは世論調査については，今日内閣支持率などの結果によっては，国政の動向を左右する重要なものとなっていることがあげられる．一方，選挙予測に用いられる選挙情勢調査は，投票者の行動を推測するものであり，現在の有権者の意見を推測する世論調査とは区別されるが，選挙情勢調査には，日頃の世論調査の技量が反映されるので両者は無関係とも言えないことに加え，古くから取り扱うデータが大きかったという事情を考慮したためである．

３．選挙予測と選挙情勢調査

データの収集とデータの分析という両方のプロセスを意図的に行ってきたもののうち，歴史的にも長く取り扱うデータの規模が大きいとされてきたものとしては，選挙結果の予測に関するものがある．例えば2019年度に実施された第25回参議院議員選挙に関する選挙予測は，報道機関によって異なるものの，合計３万人程度の回答者人数に基づいている（細貝，2019）．またこれらの回答者は，報道機関各社によって細かいやり方は異なるものの，主として RDD（Random

Digit Dialing）と呼ばれる抽出方法を用いた電話調査によって得られている．これは，現在では標準的な選挙情勢調査の実施方法である．

日本における選挙予測の方法は，1950年代から60年代にかけて原型が確立した（鈴木，2021）．1964年には，林知己夫らと朝日新聞社世論調査室が共同で進めてきた研究の結果として，「予測に関する実證的研究 ── 選挙予測の方法論 ──」（林・高倉，1964）が発表され，３次曲線を用いた得票率の推定の方法など詳細な手順が明らかにされている．鈴木（2021）によれば，この論文によって朝日新聞の選挙予測方法が公知となり，マスコミの選挙予測の方法として現在に至るまで踏襲されるようになったという．実際には，当時の調査実施手段は，電話調査ではなく，面接調査によるものであり，また調査の対象者も選挙人名簿抄本からの抽出によるものであるから，現在とはデータ収集方法としては異なっていたと考えられるが，基本的な予測の分析手順としては，この当時に確立したということになる．

IFCS-96で Data Science が提唱される30年以上前のことであるが，データの分析だけでなく，データ収集プロセスも含めたデータの理解に基づいて行われた予測手法となっている．実際，選挙予測においては，「データをとること」が特に重要で，歴史的事件ともいえる予測ミスの原因は，データのとり方にある．

古い事件の例としては，1936年アメリカ合衆国大統領選挙において，リテラリー・ダイジェスト（Literary Digest）が237万6523人の回答をもとにカンザス州知事のランドン（Landon）の勝利を予想したが，実際にはルーズベルト（F. Roosevelt）が勝利し，予想を大きく外したことがあげられる（Gallup, 1972）．これは，電話加入者や自動車保有者名簿の掲載者を対象としたため，富裕層に偏ったことが原因とされている（西平，2009）．

その12年後の1948年アメリカ合衆国大統領選挙においても，各調査会社がトールマンの再選を予測できなかったことで知られる（Gallup, 1972；西平，2009）．これは，無作為抽出を用いずに，割当抽出（Quota Sampling）を用いていたことが原因とされている．

どちらも教科書に掲載されている古典的な事件であるが，データ収集プロセスが明確であるからこそ，原因が特定されているといえる．

2016年アメリカ合衆国大統領選挙もトランプ（Trump）の当選によって選挙予測が外れたことで知られている．米国世論調査学会の委員会は，トランプに対する投票選好が現れるのが遅かったことと，（クリントン支持の）大卒者の過剰な代表性が生じていることへの調整が広範囲に失敗したことを主な原因としている（Kennedy et al., 2018）．しかし，米国の選挙情勢調査では１ケタ台の回収率，あるいは回収率が計算できないようなオプト・イン型の調査が広まっている（Callegaro & DiSogra, 2008）．日本の調査関係者からすれば，もはや予測で用いようがない領域の調査結果を用いており，外す・外さない以前である．これは，「データをとること」の問題が根本にある．

この点，日本の選挙情勢調査は，少なくとも大手マスコミによる調査であれば，「データをと

ること」について慎重に議論されてきた．しかし近年，オートコール[2]と称する自動音声を用いた電話調査（outbound 型の Interactive Voice Responses）が（体感的にであるが）増え始め，状況が変わりつつある．増えているかどうか明確に書けないのは，多くの調査結果が直接的に公表されず，その結果明確な統計がないからである．そのこと自体問題であるが，これは次のような事情による．

まず，公職選挙法の第138条の３で人気投票の公表を禁止している[3]．しかし，人気投票に該当するかどうかは，投票に似ているかどうかで判断されるので，いわゆる従来の RDD による電話調査の場合など，口頭で回答を得る方法の場合は，これまで人気投票に該当しないものとされてきた（安田・荒川，2009）．一方，オートコール調査の場合は，自動音声で質問がなされるが，回答する側は，プッシュボタンなどで回答を選ぶ形をとっている．投票に似ているかどうかは，意見が分かれるところであろう．この場合，調査実施者は，予測の数字はともかく，その結果を明確に公表しがたい．直接的な調査結果を公表すると人気投票を公表したとみなされる可能性があるからである．そのためオートコール調査の場合は実施してもその調査結果自体は秘匿されがちになることになる．さらに，政党などが実施する場合は，選挙戦略上の秘密からより他者に知られたくないので実施したこと自体を明らかにしないようにする可能性が高い．調査結果はもちろん，予測結果についても秘匿されることになる．

オートコール調査は，このように実施状況が不明瞭といった問題に加え，実施管理上の問題も多い．RDD の場合は，一般世帯を調査対象とする電話調査では，企業や学校の電話回線は事前に架電の対象から除外するか，仮につながっても調査対象外として除外するのが普通である．しかし，オートコール調査では現実に大学の研究室に架電が生じる場合も珍しくない．発信側では留守電でつながった場合と途中で切られた場合との区別がつかないので，かけ直すといった手間をかけることが難しい．好意的な調査対象者が，タイミングが悪い場合にかけ直してほしいという要望をだすわけにもいかない．つまり，コールバックができないという設計上の問題がある．また電話調査では一般的に，選挙区の区分けに電話番号をあらかじめ（完全に）振り分けることはできない．特に衆議院選挙のように選挙区が狭い場合は問題[4]で，該当しない選挙区に架電するという現象が起きかねない．オートコール調査だと選挙区外の無関係の人に当該選挙区の質問をするという無駄が生じるため，無駄な架電がまき散らかされることになる．

2） この手法の特徴は，コール（架電）の自動化ではなく，自動音声の利用にあるので，和製英語としてもオートコールという表現は必ずしも適切ではない．オペレータが会話する場合でも，既存のコンピュータ支援型電話調査（CATI, Computer-Assisted Telephone Interviewing）であれば，ある程度架電システムの自動化は可能だからである．本稿では，実情に合わせて便宜上オートコールと呼んでいる．

3）「何人も，選挙に関し，公職に就くべき者（衆議院比例代表選出議員の選挙にあつては政党その他の政治団体に係る公職に就くべき者又はその数，参議院比例代表選出議員の選挙にあつては政党その他の政治団体に係る公職に就くべき者又はその数若しくは公職に就くべき順位）を予想する人気投票の経過又は結果を公表してはならない」（公職選挙法第138条の３）．

4） 異なる選挙区に転居した場合でも，固定電話の番号が変わらないようなことが起きるためである．

要するにオートコール調査とは，融通が利かない調査であり，得られるデータの質保証についても疑念が残る調査といえる．対象者の重複，無関係な対象者による誤差を考慮しておらず，データサイズさえ大きければ良いという発想がうかがえる．「データを妄信せず，その由来をたずね」（林，2001，p.6）なければ，歴史の二の轍を踏むことになるのではないか．

４．世論調査の行方

近年の日本の世論調査に関する大きな事件の一つに，2008年８月の福田改造内閣の支持率が報道機関各社で大きく乖離したことがあげられる．この問題の根幹は，各社の報道する支持率が単に異なっていただけでなく，朝日新聞社（24%）と読売新聞社（41.3%）の間で過去最大級ともいえる17%もの開きが生じたことにある（鈴木，2009）[5]．各方面からの批判的追及があり，各社で異例の情報交換をした結果（鈴木，2021）[6]，各社で運用方法が異なっていたことが明らかになった．鈴木（2009）の記述をもとに整理すると，具体的には，

１．内閣改造の事実を伝えたうえで尋ねると，支持率が高い．

２．重ね聞きを行うと支持率が高い．

の２点が原因と考えられたのである．

ワーディングの違いにより結果が変わる可能性があることは，社会調査法においては基本事項である．古典的な事例として，林（1970）が，共通する質問文「ある会社につぎのような２人の課長がいます．もしあなたが使われるとしたら，どちらの課長につかわれる方がよいと思いますか，どちらか一つあげて下さい？」の選択肢の文章の前半と後半を入れ替えただけで，回答結果が大きく変わることを二種類の面接調査から示したものが知られている[7]．しかし，本件では，もともとは各社が独自に運用をしていたので，異例の情報交換によって具体的な状況が判明した面が大きい．データの収集プロセスの重要性や知識に対する理解がないとなかなか原因究明までたどり着きにくいことがらである．「データを妄信せず，その由来をたずね」（林，2001，p.6）ることの重要性を示すもうひとつのできごとである．

５） 従来までは，各社によって内閣支持率が異なる理由は，調査主体刺激仮説（調査主体によって調査協力層が異なるために回答分布が異なる）が有力な説明と考えられてきたが，それだけでは説明がつきにくい乖離であった．

６） 結論は2009年時点に出ているが，初めて情報交換があった事実は2021年に公知となった．

７） 第４次国民性調査（1968年実施）と1967年に実施された東京都23区の有権者を対象とした調査で示されたもので，選択肢については，前者では，「１ 規則をまげてまで，無理な仕事をさせることはありませんが，仕事以外のことでは人のめんどうを見ません 12%」「２ 時には規則をまげて，無理な仕事をさせることもありますが，仕事のこと以外でも人のめんどうをよく見ます 81%」を使用したのに対し，後者では，「１ 甲課長 仕事以外のことでは人のめんどうを見ませんが，規則をまげてまで，無理な仕事をさせることはありません 48%」「２ 乙課長 仕事のこと以外でも人のめんどうをよく見ますが，時には規則をまげて，無理な仕事をさせることもあります 47%」と異なっていた（数字は，選択肢が選ばれた%）．

一方で，世論調査の手法も岐路に立っている．ウェブ調査の隆盛をうけ，日本の世論調査へのウェブ調査導入の試みる動きがでてきたからである．典型的には，NHK 放送文化研究所の「参院選後の政治意識・2016」調査のように，住民基本台帳から無作為に抽出した人を対象に郵便を用いてウェブ調査へ依頼する方法である（萩原・村田・吉藤・広川，2018）．この方法は，ウェブ調査でありながらも従来の無作為抽出に基づく方法に擬していて手続きとしては好ましい．しかし，ウェブ調査としての回収率は 2 割程度にとどまっており，世論調査の手法として十分機能できている状況ではない．

またこのようなウェブ調査はむしろまれである．むしろ多くのウェブ調査は，パネル標本を各社がプールしておいてそこからサンプリングする，いわゆる公募型パネルの利用が主流となっている．無作為割当による認知実験の場面などに用いるのに有効と思われるが，これを世論調査に適用するのには，標本構成の問題から無理がある．世論調査と区別される時系列的な推移を把握するための世論観測的手段がせいぜい許容可能な利用法と考えられる．しかし現状では，その場合でもいくつか超えるべきハードルがある．

まず実施を請け負う調査会社は，標本に関する統計的な情報を積極的に開示すべきである．現状のウェブ調査の情報公開が不十分だからである．回収された標本構成とは別に，計画標本にあたる情報を開示していない調査会社が多い．電話調査における配信数に該当する情報であるだけに，その具体的構成を示すべきである．そうでなければ調査の質的な評価をくだすのが難しい．ウェブ調査を担う調査会社の多くは，一般社団法人日本マーケティング・リサーチ協会（JMRA）に加盟しているが，JMRA のマーケティング・リサーチ綱領第 6 条には「リサーチャーは，リサーチプロジェクトについて，クライアントに適切，かつ詳細な技術情報を提供しなければならない．また，クライアントからの要請があった場合，データの収集および加工についての品質チェックの機会を提供するよう努めなければならない．」とある．正当な理由なく情報開示に応じられない場合は同条違反と考えられる．

次に，調査結果の集計にあたっては，品質管理を入念に行うことも必要である．経験的には，スピーダーと呼ばれる異常に短い時間で調査を完了する回答者や，Satisficer のように異常に同一回答が多い回答者がウェブ調査には必ず存在にするので，これらをあらかじめ除去する工夫が必要である．

最後に，調査結果の公表にあたっては，公募型パネルで構成される調査である以上，「そうでない調査」とは明確に区別すべきである．米国調査学会の Condemned Survey Practices の第 5 項に同様の記載があり[8)]，倫理的にはこれに準拠するのが望ましいと考えられる．

新しい世論調査の手法としては，萩原（2021）のように，メタアナリシス的手法の導入や，空気感を含めた多様なメディアを用いた可視化の提案もある．前者のメタアナリシス的手法の導入については，既存の世論調査の結果を統合するものであれば，その中身は現存の世論調査

8） https://www.aapor.org/Standards-Ethics/Survey-Practices-that-AAPOR-Condemns.aspx

の延長にすぎず大きな問題ではない．報道機関各社の利害調整の必要性の検証や運用の問題が考えられる程度である．後者の具体的な手法は，ニュース記事サイトのコメントや投票などを用いて世論調査結果の代用品を可視化するものであるが，これについてはいくつか危惧される点がある．

このアイデアは，GDP 以外に景気の判断指標が多様に活用されている点に準じたものであるとされるが（萩原，2021），景気の判断指標の多くは，手順が昔から確立している．しかし，ネット上のコメントや投票は，ダブルアカウントで同じ人が複数回回答することやコメントすることが可能であり，測定の方法としてはやや不安な点がある．ビジネス目的であればかけられるコストに限界もあるので回答やコメントを何度も行うとは考え難いが，操作可能な世論によって政治や政策の意思決定を動かせるのであれば，コメントや投票といった形の世論を歪めようとする動きが出ないとは限らない．実際，2016年にはグラセゲールとクロゲルスが，スイスの『ダス・マガジン』誌で指摘したケンブリッジ・アナリティカ社に対する疑惑も生じている[9)]．

以上から，萩原（2021）の提案にあった後者の可視化手法は，仮にそれが既存の世論調査の結果を反映するような結果が得られたとしても，得られたから直ちに手法としてすぐれていることにはならない．ある程度手続きとしてもしっかりと管理されていて，その上で他の世論調査の結果とも符合する必要があると考えられる．

５．展望：世論を測る

世論調査や選挙情勢調査で，代替技術が試みられるのは，回収率が低下しているのではないかという危惧が広がっているためと考えられる．しかし，直近の回収率はそれほど下がってはいない．もちろん50年ぐらい前と比べたら大きく下がっているのは確かである．しかし，ここ数年という範囲で考えた場合，やや下げ止まっているか，下がり方が弱くなっているかとみなした方がよい．日本人の国民性調査は，56％（2003）⇒52％（2008）⇒50％（2013）と下がり気味であるが，JGSS においては，52.6％（2015），50.6％（2016），55.6％（2017），54.3％（2018）であり下がっているとは言えない．NHK「「日本人の意識」調査」に関しては，56.9％（2013）⇒50.9％（2018）と下がっているが，内閣府「社会意識に関する世論調査」に関しては，60.9（2013）⇒61.9（2014）⇒58.8％（2016）⇒59.9（2017）⇒57.4（2018）⇒54.4（2019）⇒53.9％（2020）と推移しているものの，５割台中盤で踏みとどまっている．

そのような現状を踏まえると，必ずしも急速に代替手段に変更しなくてもよいともいえる．回収率は頑張れば上がるという現実もあるからである．例えば，関西大学総合情報学部の社会

9） ケンブリッジ・アナリティカ社がトランプ陣営に雇用されて，計量心理学の手法により，フェイスブックのいいね！などから人々のパーソナリティーを測定し，広告を通じてヒラリー・クリントンへの潜在的投票者を投票に行かないように仕向けたのではないかとされる疑惑．同社は，FB のデータ利用，投票を思い留まらせる活動等を否定している（Grassegger & Krogerus, 2017）．

調査実習における郵送調査は，６割ぐらいの回収率を維持している（松本・李，2015）．例えば，この調査では，８月・９月の間に調査を実施しているが，必ずお盆休みが確実に終わった時期に行開始するようにするなどの他は，標準的な社会調査法のやり方（松本，2021）を遵守しているにすぎない．調査そのものをしっかり管理して改善していくことが重要と考えられる．

この他に，各調査における情報開示も重要である．2008年８月の福田改造内閣の際，各社が情報交換して，内閣支持率の不一致の原因が明らかになってきたことは一つの教訓といえる．選挙情勢調査に関しては，各社のノウハウがあるので公開しにくいと思われるが，20年ぐらい経過した段階で，この選挙ではどういう文言で聞いて支持率がどういう数字だったかなどを公開するということも考えられる[10]．そのような情報があれば，（当時の）考え方はどうだったのかということを事後的に検証できる余地が広がると考えられる．

６．結びに代えて

本稿では，文献レビューを通じて，今日巷間で取りざたされているデータサイエンスの起源に立ち返り，林知己夫の提唱したデータの科学の特徴を確認した．その結果，データの科学の骨格は，計画（design）・実施（collection）・分析（analysis）の三段階で構成され，単にデータ分析だけの概念ではなく，「調査の科学を包含する概念である」（林，2001，p.８）ことがはっきりした．

そこで，選挙情勢調査と世論調査という社会的に影響の大きいデータの収集に関して，主要な出来事をいくつか言及することで，データの収集プロセスの理解の重要性を確認した．と同時に，選挙情勢調査と世論調査のそれぞれが直面する調査法上のあり方の問題についても議論を行った．

収集されたデータに偏りがあれば，（どんなに複雑な予測式を組み立てたとしても）分析によってうまく予測をするのは難しい．実際，歴史的に語られる一連の選挙予測の失敗は，分析に失敗したのではなく，そもそも偏ったデータではだめだということしか示していない．その一方で，質問の仕方によって，回答の得られ方が変わるということが，世論測定をしばしば混乱させてきたことも事実である．測定の仕方によっていかようにも偏りが生じる可能性もあるため，どのようにデータが収集されてきたか理解を深めることが重要なのである．

この点，ビッグデータの世界では，精度は関係ないという議論があるが（Mayer-Schönberger and Cukier, 2013），もし恣意的な世論操作のような方法でデータが歪められていた場合，誤差を誤差として認識することは重要である．それを見抜くのは，やはりデータ収集場面の丁寧な理解が必要なはずである．

10） ここでは個票レベルの公開ではなく，集計レベルの情報公開である．なお，個票レベルでの単純公開には，マサチューセッツ州の医療データ公開問題のような事例もあり，丁寧な匿名化処理など慎重な検討が必要である（Sweeney, 2002；山口，2015）．

幸いにして，「数理・データサイエンス・AI（リテラシーレベル）モデルカリキュラム」（数理・データサイエンス教育強化拠点コンソーシアム，2020）においてもスキル・セットの中にデータの収集に関わる項目がないわけではない．しかしながら，全体としてみたときにそのような項目はA/Bテストの項目などを含むと考えたとしても少ないと言わざるをえない．調査に関しては，「2－1　データを読む」の項目のキーワードの母集団と標本抽出の括弧書きの中に見られるだけである．

林知己夫が，データ科学やデータサイエンスと呼ぶことにより，当初意図したData Science（データの科学）の意味が〈せまく〉理解されることを警戒していた点は，すでに2節で述べた通りである．その危惧は現実のものとなろうとしていないか．Data Science（データの科学）がデータサイエンスへと変容する中で何を忘れてきたか考える必要がある．

Davenport and Patil（2012）は，データサイエンティストを説明する中で，何回か，かつてのクオンツに似ていると表現している．これを聞いて，ヘッジファンドであるLTCM（Long-Term Capital Management）の破たんやリーマンショックを想起するのが，ただの杞憂であることを望む次第である．

謝辞

2020年度世論調査協会におけるシンポジウム「危機と変革の中の世論調査」で討論者として参加したことと，2021年度春学期に関西大学で開講されたリレー形式の授業「活用法を見聞するAI・データサイエンス」（チャレンジ科目）の担当者の一人となったことが，本稿の執筆のきっかけです．どちらの経験も執筆を進める上で全体として有益であったことを付記し，それぞれの関係者に謝意を表します．

参考文献

Callegaro, M. & DiSogra, C.（2008）Computing Response Metrics for Online Panels. *Public Opinion Quarterly*, 72（5），pp. 1008–1032.

Davenport, T. H., and Patil, D. J.（2012）Data Scientist: The Sexiest Job of the 21st Century. *Harvard Business Review* 90（10），pp. 70–76.（邦訳，トーマス H. ダベンポート，D. J. パティル（2013）「いま最も必要とされているプロフェッショナル データ・サイエンティストほど素敵な仕事はない」『Diamond ハーバード・ビジネス・レビュー』（2013年2月）38（2），pp. 84–95）

Escoufier, Y.（1995）Data science at the Unité de Biométrie-Montpellier-. In Escoufier, Y., Hayashi C., Fichet, B. Ohsumi N., Diday, E., Baba Y. and Lebart L.（Eds.），*Data science and its applications: La science des données et ses applications*（pp. 1–6）. Tokyo: Academic Press/Harcourt Brace.

Escoufier, Y., Hayashi C., Fichet, B. Ohsumi N., Diday, E., Baba Y. & Lebart L.（Eds.）（1995）*Data science and its applications: La science des données et ses applications*. Tokyo: Academic Press/Harcourt Brace.

Gallup, G.（1972）*The sophisticated poll watcher's guide*. Princeton, NJ: Princeton Opinion Press.（邦訳，G. ギャラップ（1976）『ギャラップの世論調査入門』二木宏二訳，みき書房）

Grassegger, H. & Krogerus M.,（2017）The Data That Turned the World Upside Down: How Cambridge Analytica used your Facebook data to help the Donald Trump campaign in the 2016 election, *Motherboard*, January 28, 2017. https://www.vice.com/en/article/mg 9 vvn/how-our-likes-helped-trump-win（邦訳，ハネス　グラセゲール，ミカエル　クロゲレス（2017）「「私が爆弾を作ったのではない。私は爆弾が存在することを示しただけだ」── 世界をひっくり返したデータとは ──」『世界』坂野正明訳，岩波書店，

pp. 177–191.）
萩原雅之（2021）「マーケティングからみた世論調査について」『よろん』127，pp. 42–45.
林知己夫（1970）「身近な社会」統計数理研究所国民性調査委員会（編）『第２　日本人の国民性』至誠堂，pp. 75–110.
林知己夫（1996）「データ解析からデータサイエンスへ」『日経 BP ムック：データウエアハウスがビジネスを変える』日経 BP 社，pp. 82–87.（再掲：林知己夫著作集編集委員会編『現象をさぐる ── データの科学 ──』勉誠出版，2004，pp. 255–267）
Hayashi, C.（1998）What is Data Science? Fundamental Concepts and a Heuristic Example. In *Data science, classification, and related methods: proceedings of the Fifth Conference of the International Federation of Classification Societies*（IFCS-96），Kobe, Japan, March 27–30, 1996.
林知己夫（2001）『データの科学』朝倉書店.
林知己夫・高倉節子（1964）「予測に関する実證的研究 ── 選挙予測の方法論 ──」『統計数理研究所彙報』12巻 1 号，pp. 9–86.
林知己夫・鈴木達三（1986）『社会調査と数量化 ── 国際比較におけるデータ解析 ──』岩波書店.
林知己夫・鈴木達三（1997）『社会調査と数量化 ── 国際比較におけるデータの科学 ──（増補版）』岩波書店.
細貝亮（2019）「新聞・テレビはどう伝えたか ── 第25回参院選の世論調査報道から ──」『よろん』124 巻，pp. 2–12.
Kennedy, Courtney, Mark Blumenthal, Scott Clement, Joshua D Clinton, Claire Durand, Charles Franklin, Kyley McGeeney, Lee Miringoff, Kristen Olson, Douglas Rivers, Lydia Saad, G Evans Witt, Christopher Wlezien（2018）An Evaluation of the 2016 Election Polls in the United States. *Public Opinion Quarterly*, 82（1），pp. 1–33.
松本渉（2021）『社会調査の方法論』丸善出版.
松本渉・李容玲（2015）「調査実習の事例報告 高品質な郵送調査の実践をめざして：高槻市と連携した関西大学総合情報学部の社会調査実習」『社会と調査』（15），107–111.
Mayer-Schönberger, V. and Cukier, K.（2013）*Big data: a revolution that will transform how we live, work and think*. Boston, MA: Houghton Mifflin Harcourt.（邦訳，ビクター・マイヤー＝ショーンベルガー，ケネス・クキエ（2013）『ビッグデータの正体：情報の産業革命が世界のすべてを変える』斎藤栄一郎訳，講談社）
日本学術会議情報学委員会 E ── サイエンス・データ中心科学分科会（2014）『提言　ビッグデータ時代に対応する人材の育成』（https://www.scj.go.jp/ja/info/kohyo/pdf/kohyo-22-t198-2.pdf）
西平重喜（2009）『世論をさがし求めて ── 陶片追放から選挙予測まで ──』ミネルヴァ書房.
数理・データサイエンス教育強化拠点コンソーシアム（2020）「数理・データサイエンス・AI（リテラシーレベル）モデルカリキュラム ── データ思考の涵養 ──」（2020年 4 月）（http://www.mi.u-tokyo.ac.jp/consortium/pdf/model_literacy.pdf）
鈴木督久（2009）「世論調査の最近の動向」『社会と調査』（3），pp. 13–19.
鈴木督久（2021）『世論調査の真実』日経 BP・日本経済新聞出版本部.
Sweeney, L.（2002）. K-anonymity: A model for protecting privacy. *International Journal of Uncertainty Fuzziness and Knowledge-Based Systems*, 10（5），pp. 557–570.
椿広計（2020）「システム科学とデータ科学」『横幹』14（1），pp. 64–69.
山口利恵（2015）「ビッグデータの利活用とプライバシー保護の難しさ」『映像情報メディア学会誌』69（2），pp. 155–161.
安田充・荒川敦（編著）（2009）『逐条解説公職選挙法』ぎょうせい.