解説

データジャーナリズムとデータ科学

Data Journalism and Data Science

岡崎直観

Abstract

インターネット上で様々なデータがオープンかつ大量に利用可能になる中で，放送・新聞といったメディアにおける報道のあり方が大きく変わりつつある．メディアの社会的使命の一つは，読者・視聴者に対して意思決定に役立つ情報や新たな事実を提供することである．この使命に揺るぎはないが，近年のビッグデータビジネスの進展に伴い，膨大なデータを分析して情報を分かりやすく提供する「データジャーナリズム」が注目を浴びている．そこで，本稿ではデータジャーナリズムを，その事例や背景を交えながら解説する．更に，データジャーナリズムの基盤となる解析技術の一つである自然言語処理の応用事例を紹介し，データ科学をジャーナリズムに応用する際に留意すべきポイントを紹介する．最後に，情報技術を活用した報道に関して，今後の展望を述べる．

キーワード：データジャーナリズム，データ科学，自然言語処理，報道，メディア

1. データジャーナリズムとは

データジャーナリズム（data journalism）[1]とは，文字どおり「データによるジャーナリズム」のことで，データに重心を置いた報道姿勢（journalism）のことを指す．ただ，この定義には不明瞭な点が多い．例えば，株価やスポーツの試合結果などは数値で表される「データ」であるが，これらの数値データを新聞に掲載するだけではデータジャーナリズムと呼べない．「データ科学（data science）をジャーナリズムに適用したもの」という簡潔な説明[2]は分かりやすい．この説明はある程度的を得ているが，やはりデータジャーナリズムを語るには不十分である．データジャーナリズムの正確な定義を与えるのは難しいが，現場レベルでは様々な取組みが行われてきた[3]．そこで，データジャーナリズムの典型例を紹介し，その後データジャーナリズムとは何か，少し掘り下げてみたい．

1.1 データジャーナリズムの代表例

図1は，2010年に米紙The Las Vegas Sunが公開した「Do No Harm」プロジェクト(注1)のひとこまである．このプロジェクトでは，1999年から2009年までにネバダ州の病院で発行された290万件の医療明細を分析し，ラスベガスの病院の「医療品質」を検証している．図1では，2008年から2009年までの間に，病院の入院患者が被った969件の医療ミス（防げたはずの損害）を，病院や種類（褥瘡，落下，感染，けがなど）別に示している．分析結果は，解説記事[4]とともにインタラクティブなWebサイト(注2)で公開された．

医療明細には診療内容の詳細や患者の既往歴は載っていないため，このデータから医療ミスが判明するのは意外なことのように思える．ただ，医療明細は公開情報であり，データそのものはネバダ州の担当部署から入手できる．今回のプロジェクトは，患者の病状が入院前のものであるかを識別するコードが診療明細に付与されていることに目を着け，そこから院内感染やけがなどの医療ミスを明らかにしていった[5]．すなわち，公開データを収集・加工・分析することで，ラスベガスの病院の現状という価値あるデータとニュースを発掘し，社会に大きなインパクトを与えた．「Do No Harm」プロジェクトのインパクトは医療改革にも波及し，後にネバダ州で医療行為の透明性に関する法律が成立した[6]．

岡崎直観　東北大学大学院情報科学研究科システム情報科学専攻
Naoaki OKAZAKI, Nonmember (Graduate School of Information Sciences, Tohoku University, Sendai-shi, 980-8579 Japan).
電子情報通信学会誌　Vol.99 No.4 pp.339-346 2016年4月

（注1）　http://lasvegassun.com/hospital-care/
（注2）　http://lasvegassun.com/hospital-care/events-interactive/

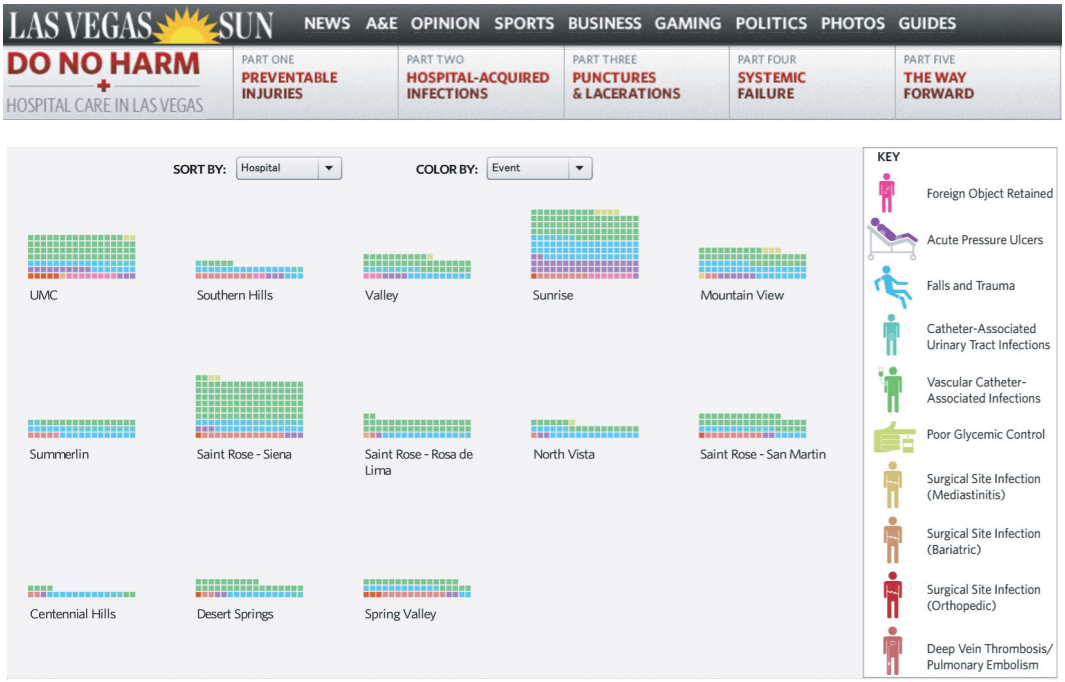

図 1　The Las Vegas Sun の「Do No Harm」プロジェクト　防げたはずの医療ミス（褥瘡，落下，感染，けがなど）の数を病院ごとに集計している．色は医療ミスの種類を表す．

図 2 は，英紙 The Guardian の「Reading the Riots」プロジェクト(注3)で制作された Web アプリのスクリーンショットである．2011 年夏にイギリスで発生した暴動では，真偽不明の情報がソーシャルネットワーク上に流れた．このスクリーンショットでは，「ロンドンアイ（観覧車）に火が放たれた」というデマに関するツイートを可視化している．このデマの発端は，「Oh my god! This can't be happening at London Eye! #Londonriots #Londonriot #Prayforlondon」というメッセージと一緒に，ロンドンアイが燃えているかのような写真が添付されたツイートである．システムの上部は，このデマに関連するツイート数が表示され，ユーザは日時を指定したり，ツイッター上の発言の推移をアニメーションで閲覧できる．ツイートは発言内容に基づいてクラスタ（グループ）化されており，「ロンドンの観覧車に火が放たれた」という情報を支持するクラスタは緑色，否定するクラスタは赤色，この情報に関して追加情報を要求するクラスタは黄色，コメントするクラスタは灰色の円で示されている．円の大きさはクラスタのツイート数を反映しており，クラスタにマウスカーソルを合わせると，実際のツイートを読むことができる．

データ分析方法の詳細は記事(7)や論文(8)にまとめられている．このプロジェクトでは，Twitter 社から特別な承諾を得て，2011 年 8 月 6 日から 17 日までのツイートのうち，54 種類のハッシュタグ（例えば「#riotcleanup」など）のいずれかを含む 260 万件のツイートをデータとして用いた．ジャーナリストたちは分析の対象を 7 件の風評に絞り込み，英国の大学の研究者と共同でツイートを仕分けし，風評ごとにサブコーパスを構築した．ツイートのクラスタは，リツイートや非公式リツイートを編集距離(用語)で認識することで形成されている．クラスタの色分け（支持，否定，要求，コメント）は，3 人の博士課程学生によって行われ，論文によるとその一致率は 89〜96% であった．図 2 のインタラクティブなシステムは，WebGL，HTML5 Canvas，Flash，SVG，VML などの標準的なブラウザ技術により実装され，アプリケーションをインストールしなくても体験できるようになっている．

(注 3)　http://www.theguardian.com/uk/series/reading-the-riots

用　語　解　説

編集距離　二つのテキスト間の類似度を測る尺度．片方のテキストを編集してもう一方のテキストと同一にするために必要な編集操作の数で定義される．

TF*IDF　キーワードのスコア付けのための統計尺度の一つ．文書中で頻繁に出現し，かつ特定の文書のみで出現するキーワードのスコアが高くなるように設計されている．

相互情報量（Point-wise Mutual Information）　二つの事象の共起性を測る統計尺度．キーワード抽出では，コロケーション（単語連接）の強さを測定し，名詞句などのフレーズを認定するために用いられる．

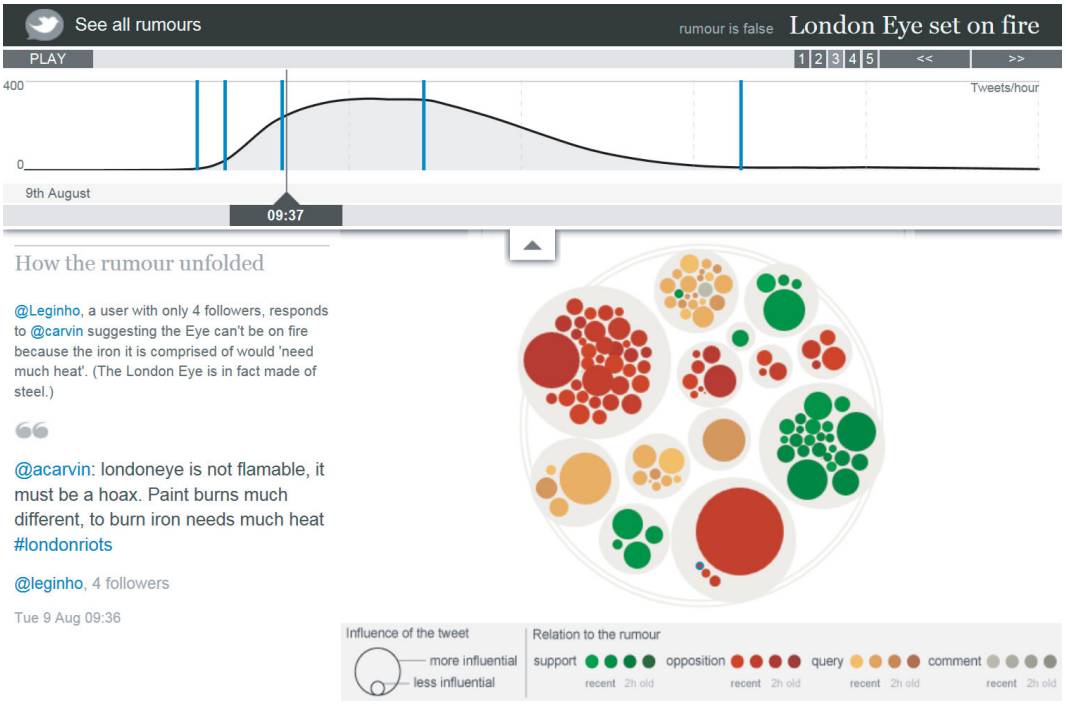

図 2 「Reading the Riots」プロジェクトで「ロンドンアイ（観覧車）に火が放たれた」という誤情報の時系列をインタラクティブに可視化したシステム　誤情報を支持するツイートのクラスタは緑色，誤情報を否定するツイートのクラスタは赤色，誤情報に関する情報を要求するツイートのクラスタは黄色，誤情報にコメントするツイートのクラスタは灰色の円で示されている．円の大きさはクラスタのツイート数を表し，クラスタにマウスカーソルを合わせることで，そのクラスタの典型的なツイートを読むことができる．

1.2　データジャーナリズムへの流れ

これまでの事例で見てきたように，データジャーナリズムはデータを収集，整形，整理，分析，可視化，公表することで，新しいニュースを発見したり，ニュースを分かりやすく伝えようとする報道活動である．データ分析から得られた知見を記事にするだけでなく，データの分析過程を詳細に説明したり，分析に用いたデータやツールを積極的に公開することで，分析及び報道の透明性を高める．また，データや分析結果を単に提示するだけではなく，インフォグラフィック（infographic），インタラクティブな Web アプリケーション，動画像，アニメーションを駆使し，視聴者や読者をデータの世界に引き込んでいく．

英紙 The Guardian が 2009 年に立ち上げた Datablog[注4] は，大手メディアがデータジャーナリズムに大々的に取り組んだ最初の事例として有名である．2010 年には，WikiLeaks が公開したアフガン紛争関連資料やイラク戦争の米軍機密文書を大手報道機関が分析し[9),(10)]，データジャーナリズムは更に注目を浴びた．一方で，分析結果の公開により引き起こされる二次被害（プライバシーの侵害）など，データジャーナリズムに対する新たな懸念も生まれた．

データからニュースを発見するという試み自体は目新しいものではない．Simon Rogers[注5] は，英紙 The Guardian における最初のデータジャーナリズムは 1821 年 5 月 5 日まで遡ると説明している[11),(12)]．この記事では，マンチェスターとサルフォードの学校の生徒数と年間支出から，無料で学校に通っていた生徒数や貧困に苦しむ生徒数を初めて明らかにし，その後の教育制度の改革につながったと言われている．コンピュータの発明前であったため，現在のデータジャーナリズムとはデータ分析の様子は異なるが，データからニュースを発見しようという意気込みは同じである．

また，ニュースの発掘にコンピュータや計量的手法を活用する試みも古くから存在する．1952 年に米テレビ局 CBS が大統領選挙の結果を予測するために，UNIVAC[注6] を活用したことは有名である[13)]．コンピュータを活用してデータを収拾・分析し，ニュースを執筆する手法はコンピュータ支援報道（CAR：Computer-Assisted Reporting）と呼ばれ，コンピュータはニュース制作の現場に取り込まれている．コンピュータ

（注 4）　http://www.theguardian.com/data

（注 5）　英紙 The Guardian で Datablog の立ち上げと編集長を務める．その後 Twitter 社で初のデータエディター（data editor）に就任し，現在は Google の News Lab チームで data editor を担当．

（注 6）　1951 年にレミントンランド社（現 Unisys）が発売した世界初の商用コンピュータ．

支援報道とデータジャーナリズムに本質的な差があるのかどうか，専門家の間でも意見が分かれており，両者に差はないとの主張もある[1]．ただ，データジャーナリズムはニュースを発見するためにデータを使うだけでなく，分析に用いたデータや分析結果を公開するなど，データそのものを尊重し，読者に伝えようとする思想が強い．オープンデータの推進，ソーシャルネットワークの普及，ビッグデータ分析の発展，Web プラットホームの成熟，マッシュアップによる迅速かつ効果的なデータ連携，報道のディジタル化など，インターネット中心の情報アーキテクチャが，報道のアーキテクチャにも大きな変化をもたらし，データジャーナリズムの原動力となっている．

データジャーナリズムの新展開としてセンサジャーナリズム（sensor journalism）[14]も見逃せない．データジャーナリズムでは既存のデータを分析に用いていたが，センサジャーナリズムではスマートフォンや Arduino[注7]，Raspberry Pi[注8]，ドローン（無人航空機）などの機器を「センサ」として活用し，分析したいデータそのものを作り出す．代表例として，米ニューヨークのラジオ局 WNYC の「Cicada Tracker」プロジェクト[注9]が挙げられる．このプロジェクトは，10 数年おきに大量発生するセミの羽化を予測することを目的に掲げたが，セミの幼虫は数年間を地中で過ごすため，その羽化を検出することはほぼ不可能である．代わりに，地表面下 6 インチ（約 15 cm）の温度が華氏 64 度（摂氏 17.8 度）に上昇した時点を羽化のタイミングと仮定し，そのデータを計測・収集する計画を立てた．WNYC が安価な地中温度計測機器[注10]の組立て手順を公開したことで，ラジオのリスナーの間で機器を入手・組み立てる動きが広まり，最終的に約 1,500 件の測定結果を集めた．このプロジェクトは，Internet of Things（IoT）やクラウドソーシングなどの技術革新が報道にも革新をもたらす可能性を示唆しており，ハードとソフトの両面において興味深い．

海外でデータジャーナリズムが花開いた頃，日本は 2011 年の東日本大震災に見舞われた．広範囲かつ甚大な地震・津波被害，予断を許さない東京電力福島第一原子力発電所の事故などで，多種多様な情報ニーズが急上昇し，Google パーソンファインダー[注11]，ホンダの通行可能道路実績マップ[注12]，sinsai.info[注13]など，企業や個人を主体とする情報共有プロジェクトが自然発生した．また，放射線測定プロジェクト「Safecast」[注14]や市民放射能測定所での食品や土壌の放射能測定結果を共有する「みんなのデータサイト」[注15]など，一般市民によるデータ作成・共有の動きも広まった．これらのプロジェクトはストーリー化を見据えたものではなかったが，データ作成・共有やマッシュアップなど，データジャーナリズムやセンサジャーナリズムに通じる要素があった．2012 年秋に開催された東日本大震災ビッグデータワークショップ[注16]は，震災時の情報伝達を振り返り，次に起こる災害に備えるため，震災直後 1 週間の間に発生したデータを分析する試みであった．このワークショップには多くの個人，研究者，企業，メディアが参加し[注17]，日本のデータジャーナリズムの一つの転換点となった．

2013 年以降は，朝日新聞，毎日新聞，日本放送協会などの大手メディアがデータジャーナリズムをけん引してきた．朝日新聞は 2013 年からビリオメディア[注18]という企画を立ち上げ，ソーシャルメディアやビッグデータ分析を活用し，震災・復興や国政選挙に関するツイートを分析した．また，投稿マップ[注19]というサービスは，新聞社がネット上で読者の投稿を収集し，それらを集合知として可視化するという新しいコンセプトに基づいている．毎日新聞も 2013 年頃から国政選挙や政策に関するツイートの分析を行い，その結果を度々紙面に掲載している．日本放送協会は，震災ビッグデータ[注20]，DATAFILE.JPN[注21]，データなび[注22]など，様々なデータ分析に積極的に取り組み，その分析結果をテレビ番組や Web サイト上で発信している．Yahoo! Japan は国政選挙に加え，インフルエンザの感染状況や人々の感情などをデータから分析し，分析ビッグデータレポート[注23]というサイトで紹介している．データジャーナリズムの動きは地方紙や新興メディアにも広がり，2015 年 1 月に日本ジャーナリスト教育センター（JCEJ）が主催した「ジャーナリズム・イノベーション・アワード[注24]」では，37 件のジャーナリズム作品が出品され，最優秀作品「台風リアルタイム・ウォッチャー」[15]を含む 18 件がデータジャーナリズムに関連するものであった[16]．

（注 7） https://www.arduino.cc/
（注 8） https://www.raspberrypi.org
（注 9） http://project.wnyc.org/cicadas/
（注 10） Arduino Uno をベースに約 80 ドル，29 ステップで組み立てることが可能である．
（注 11） https://google.org/personfinder
（注 12） http://www.honda.co.jp/internavi/service//
（注 13） http://www.sinsai.info/
（注 14） http://safecast.jp/
（注 15） http://www.minnanods.net/
（注 16） https://sites.google.com/site/prj311/home
（注 17） https://sites.google.com/site/prj311/event/presentation-session
（注 18） http://www.asahi.com/special/billiomedia/
（注 19） http://www.asahi.com/topics/特定秘密法．html（※ページ下部「投稿マップ」参照）
（注 20） http://www.nhk.or.jp/datajournalism/
（注 21） http://www3.nhk.or.jp/news/dj/datafile.jpn/
（注 22） http://www.nhk.or.jp/d-navi/
（注 23） http://docs.yahoo.co.jp/info/bigdata/
（注 24） http://jcej.info/jia/

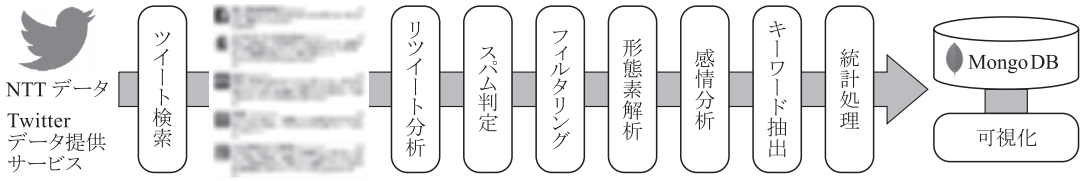

図 3　ツイッター分析の標準的なワークフロー

2. データ科学から見たデータジャーナリズム

ジャーナリズムの使命の一つは，読者・視聴者に対して意思決定に役立つ情報や新たな事実を提供することである．ゆえに，データジャーナリズムの目標の一つは，データから意思決定に役立つ情報を引き出すことである．このデータから意思決定を支援するというアイデア自体は，データマイニング（data mining）や知識発見（knowledge discovery）と呼ばれ，データ科学の分野では当然のように研究・実践されてきた．では，データマイニングに精通している研究者・エンジニアが，データジャーナリズムに取り組む際に押さえておくべきポイントはどのようなものか？　ここでは，筆者らがテキストマイニング（text mining）をデータジャーナリズムに適用した事例を紹介しながら，五つのポイントを説明したい．

2.1　テキストマイニングのデータジャーナリズムへの応用

テキストマイニングとは，テキストに特化したデータマイニングのことで，テキストデータから価値のある情報・知識を抽出する技術である．テキストマイニングをデータジャーナリズムに応用した事例は，先に紹介した「Reading the Riots」のほか，WikiLeaks が公開したイラク戦争に関する機密文書のネットワークによる可視化[10]，アメリカの大統領選挙のキャンペーンでオバマ陣営が送信したメールの分析[17]，2012 年のハリケーンサンディが洪水を引き起こしたときのツイートの分析[18]など，数多く存在する．筆者らは，朝日新聞や日本放送協会などの大手メディアと，2013 年参院選ツイッター分析[注25]，福島の桃に関する風評の分析[19]，2014 年サッカー W 杯日本戦ツイート分析[注26]などに参加した．誌面の都合上，ここでは 2013 年参院選ツイッター分析に焦点を当てる．

2013 年参院選ツイッター分析では，ネット選挙解禁後初の国政選挙において，有権者及び候補者がツイッター上でどのような議論を展開するのか，分析するプロジェクトであった．ジャーナリストチームは主観を排除し，前提なしにデータを眺め，人々の無意識を可視化することにこだわった．筆者らは手法にとらわれないデータ分析を行うため，既存のテキストマイニングフレームワークを利用せず，様々な分析手法を自前で実装した．様々なデータ分析を実践し，その中には紙面化に至らなかったものもある．図 3 に，結果的に汎用性が高かった分析ワークフローを示した．各コンポーネントの処理内容は以下のとおりである．

- ツイート検索：調べたい内容に沿った検索クエリを設計し，そのクエリに合致するツイートを収集する．例えば，自民党に関するツイートを収集する際は，「自民党 or 自由民主党」などの検索クエリが用いられた．正確な分析を期すため，全量のツイートを検索できる NTT データの Twitter データ提供サービス[注27]を用いた．
- リツイート解析：検索 API で取得したツイートのデータには，リツイート元のツイートを特定するための情報が含まれている．ただ，この情報は諸事情により不完全である．そこで，ツイートのテキストの類似度に基づく手法で，データでは捉えられていないリツイート，非公式リツイート，重複（コピー）ツイートの関係を認識する．
- スパム判定：ボットと呼ばれる自動ツイート発信プログラムから投稿されたツイートを除去する．主に，ツイート投稿プログラムの名前やツイート本文中に含まれる URL[注28]を用いてスパム判定を行うが，分析内容に合わせて判定ルールを調整する．
- フィルタリング：調べたい内容とは無関係のツイートを除去する．例えば，日本共産党に関するツイートを収集する際，「共産党」を検索クエリに用いると，「中国共産党」など，日本の共産党とは無関係のツイートが含まれてしまう．この問題は，固有表現の曖昧性に起因するが，この問題の解決を自動化するのは不安が残るため，人手でフィルタリングルールを設計する．

（注 25）　http://www.asahi.com/special/billiomedia/senkyo2013/
（注 26）　http://www.asahi.com/worldcup/2014/special/chart/
（注 27）　https://nazuki-oto.com/twitter/
（注 28）　特定のアフィリエイトサイトへ誘導するツイートをスパムとして認定するため．

・　形態素解析：ツイートのテキストを単語列に分かち書きするため，形態素解析器 MeCab[注29] を用いた．形態素解析の誤りはキーワード抽出に悪影響を及ぼすため，人名や組織名を中心に形態素解析辞書の拡充を行った．2013 年当時は存在しなかったが，今なら Web 上の様々な言語資源から固有表現を拡充した辞書[注30] を活用するとよいかもしれない．ただ，分析のターゲットに応じて人手で形態素解析辞書を管理することは不可欠である．

・　感情分析：ツイートのテキストをポジティブ・ネガティブ・ニュートラルに分類する．感情分析には機械学習に基づく手法を採用しているが，感情分析は分析対象へのドメイン依存度が強いため，分析対象のテキストで訓練事例を作成し，解析精度を向上させた．

・　キーワード抽出：形態素解析の結果に基づき，名詞の連接などのキーワードを抽出する．

・　統計処理：ツイートの数やキーワードの出現回数を集計する．ツイートの数はツイート収集に用いた検索クエリ（例えば「自民党」）のツイッター上での言及数を表し，キーワードの出現回数は検索クエリの関連キーワードの言及数を表す．また，1 日，1 時間，1 分など，時間当りの数を計測し，バースト検出手法を適用することで，検索クエリや関連キーワードの盛り上がりが分かる．

・　可視化：分析結果は MongoDB[注31] などのデータベースに格納し，ブラウザ上で分析結果をすぐに可視化できるようにしてある．分析結果に関連するツイートを確認できるインタフェースを用意し，分析結果の確認と解釈を支援する．分析結果はコンマ区切り形式（CSV）などで新聞社のデザイン担当者に送り，インフォグラフィックスとして紙面や Web サイト上で掲載される．図 4 に，Web サイト上での可視化例[注32] を示した．

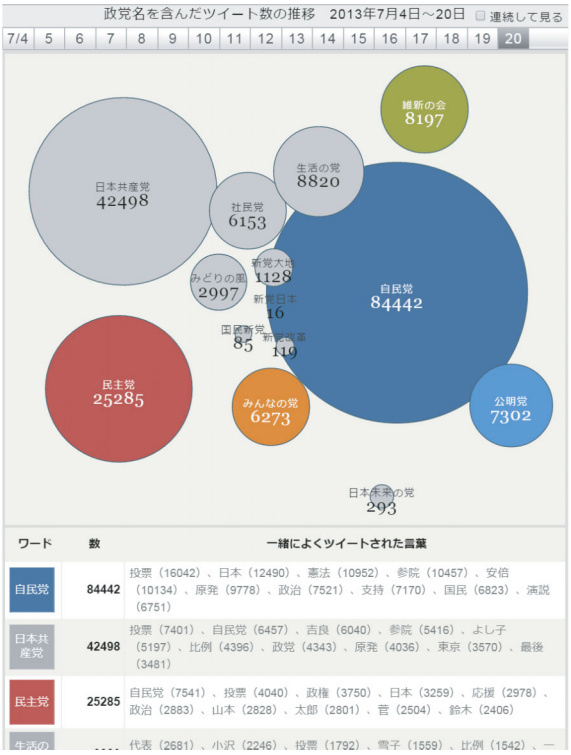

図 4　2013 年参院選ツイッター分析の可視化例

2.2　留意すべき五つのポイント

図 3 で示したワークフローは標準的なもので，テキストマイニングの研究者としては物足りなさを感じるかもしれない．しかし，データ科学者がデータジャーナリズムに取り組む際に留意すべきポイントを考慮すると，このワークフローに収束していく．その留意すべきポイントは，正確性，中立性，了解性，透明性，専門性である．

正確な記事は報道の生命線である．当然，データジャーナリズムでもデータ分析の正確性を保証しなければならない．手法の間違った使い方やバグのある実装は論外として，精度の低い解析手法も避けたほうがよい．更に，データ分析に入り込む（暗黙の）仮定にも注意を払わなければならない．例えば，「自由民主党」というキーワードを含むツイート数を，他の政党名を含むツイート数と比較することを考える．この分析は，自由民主党への注目度を反映するかもしれないが，世間の支持を反映するわけではない．また，ツイッターのユーザ層は有権者の分布とは掛け離れていることにも注意が必要である．

正確性に加え，中立性も報道の大原則である．特定の個人・団体を利するようなデータ分析手法を使ってしまうと，報道への信頼が揺らいでしまう．更に，データ分析結果の了解性，すなわち受け手である視聴者・読者への配慮も重要である．企業で実践されるデータマイニングでは，経営陣やクライアントなど，意思決定権を持つ人物が分析結果を活用できれば十分かもしれない．しかし，報道を活用するのは一般市民である．視聴者・読者を置き去りにするような高度な分析方法を採用してしまうと，データとの距離が縮まるどころか，逆にデータ分析結果に対する了解性が薄れ，データジャーナリズムの意義が損なわれる．

（注 29）　http://taku910.github.io/mecab/
（注 30）　https://github.com/neologd/mecab-ipadic-neologd
（注 31）　https://www.mongodb.org/
（注 32）　http://www.asahi.com/special/billiomedia/senkyo2013/bunseki.html

これらの三つのポイントに対処するために最も有効な戦略は，データ分析の透明性を高めることである．具体的には，複雑な分析手法は使わない，精度が低い分析手法は使わない，一般の視聴者・読者にも説明できる分析手法を検討する，分析に用いたデータや手法はできるだけオープンにする，分析結果を分かりやすく可視化する，などである．透明性を高めておけば，分析手法が置いた仮定も含めて，データ分析の内容と結果を視聴者・読者に伝えることができる．

例えば，図3のワークフローでは，出現頻度を用いて関連キーワードを抽出していた．テキストマイニングの専門家の立場からは，出現頻度ではなくTF*IDF(用語)や相互情報量(用語)などの統計量を活用したくなる．しかしながら，TF*IDFや相互情報量を一般向けに説明するのは大変である．頻度などのシンプルな統計量で分析結果の分かりやすさを優先するか，詳説してまでこれらの統計量を採用する必要があるか，選択を迫られることになる．

最後に忘れてはならないポイントが，取り組んでいる課題に対する専門性である．データジャーナリズムはデータが全てではない．分析結果をストーリー化できなければ，ただのデータ分析になってしまう．しかし，分析結果のストーリー化や，分析方法の検討・改善には，調査対象に関する知識が欠かせない．ジャーナリスト自身がデータ分析にも精通しているのが理想的であるが，データ分析の専門家と報道の専門家で分担するケースも多い．そのようなチーム編成の場合は，報道の専門家とデータ分析の専門家がすぐに議論できるような体制を組み，緊密に連携できるような環境を作ることが，データジャーナリズムのプロジェクトを成功に導く鍵であろう．

3. おわりに

本稿では，データ科学の立場からデータジャーナリズムを概観してきた．データジャーナリズムは一時的な流行なのか，メディアと世界の未来を救うのか，現時点では見当がつかない．しかし，データ科学や情報技術が報道に活用される機会は確実に増えていく．Googleは2015年6月にNews Lab(注33)というサイトを立ち上げ，Google Public Data Explorerなどのツール，Google Trendsなどのデータを報道機関向けに提供している．米AP通信は，Automated Insights(注34)と提携して，記事の自動生成，いわゆる「ロボットジャーナリスト」の実用化を目指している(20)．また，取材資料，記事校正，広告掲載，番組や記事の閲覧履歴など，メディア自身も大量のデータを生産・保有している．視聴者や読者の意見を把握する試み(21)や，メディアが保有するデータやその分析から価値を発掘することで，新たな展開が生まれるだろう．

報道の現場でデータジャーナリズムや情報技術の活用を模索する動きは，アカデミアでも広まりを見せている．例えば，スタンフォード大学(注35)やコロンビア大学(注36)では，データジャーナリズムに特化した講義が行われている．2014年には知識発見に関する国際会議KDDの併設ワークショップとしてNewsKDD(注37)が，2015年にはWebに関する国際会議WWWの併設ワークショップとしてNewsWWW(注38)が開催された．対象とする課題やオーディエンスの違いはあるが，報道も研究も新しい発見を文章や番組にまとめ，世の中にインパクトを与えるという点で，メディアと研究者には共通のゴールがある．データ科学や情報技術という接点で，メディアと研究者の異業種連携が更に進み，新しい研究やビジネスを通じて，ジャーナリズムの未来が切り開かれていくと期待している．

謝辞　本稿の執筆にあたり，朝日新聞社の奥山晶二郎氏，竹下隆一郎氏，野澤博氏から有益なコメントを頂戴しました．また，本稿の執筆の機会を与えて下さった会誌編集委員会の皆様に感謝申し上げます．

（注33）　https://newslab.withgoogle.com/
（注34）　http://automatedinsights.com/
（注35）　http://journalism.stanford.edu/
（注36）　http://www.journalism.columbia.edu/page/1077-specialization-in-data/936
（注37）　http://ailab.ijs.si/~blazf/NewsKDD2014/
（注38）　http://ailab.ijs.si/~marko/NewsWWW/

文　献

（1） J. Gray, L. Chambers, and L. Bounegru, The Data Journalism Handbook, O' Reilly Media, 2012.
（2） A.B. Howard, "The art and science of data-driven journalism," Technical report, The Tow Center for Digital Journalism, Graduate School of Journalism, Columbia University, 2014.
（3） 田中孝宣，"データジャーナリズム・イギリスBBCの取り組み～デジタル時代の新たな報道スタイルの模索～，" 放送研究と調査，vol. 65, no. 2, pp. 46-59, 2015.
（4） A. Richards, "A breakthrough in medical transparency," Las Vegas Sun online article, June 2010. http://lasvegassun.com/news/2010/jun/27/complete-guide-vegas-health-care/
（5） A. Richards, "Billing codes key to data analyzed on infections," Las Vegas Sun online article, Aug. 2010. http://lasvegassun.com/news/2010/aug/08/billing-codes-key-data-analyzed-infections/
（6） D.M. Schwartz, "Governor signs health care transparency bills into law," Las Vegas Sun online article, June 2011. http://lasvegassun.com/news/2011/jun/24/governor-signs-health-care-transparency-bills-law/
（7） A. Dant and J. Richards, "Behind the rumours : how we built our Twitter riots interactive," The Guardian online article, Dec. 2011. http://www.theguardian.com/news/datablog/2011/dec/08/twitter-riots-interactive
（8） R. Procter, F. Vis, and A. Voss, "Reading the riots on Twitter : methodological innovation for the analysis of big data," International Journal of Social Research Methodology, vol. 16, no. 3, pp. 197-214, 2013.

（9） S. Rogers, “Wikileaks Iraq war logs : every death mapped,” 2010. http://www.theguardian.com/world/datablog/interactive/2010/oct/23/wikileaks-iraq-deaths-map
（10） J. Stray, “A full-text visualization of the Iraq war logs,” 2010. http://jonathanstray.com/a-full-text-visualization-of-the-iraq-war-logs
（11） S. Rogers, “The first Guardian data journalism : May 5, 1821,” The Guardian online article, Sept. 2011. http://www.theguardian.com/news/datablog/2011/sep/26/data-journalism-guardian
（12） S. Rogers, Facts are Sacred, Guardian Faber Publishing, 2013.
（13） A. Bochannek, “Have you got a prediction for us, UNIVAC? ,” Computer History Museum online article, 2012. http://www.computerhistory.org/atchm/have-you-got-a-prediction-for-us-univac/
（14） F. Pitt, “Sensors and journalism,” 2014. https://www.gitbook.com/book/towcenter/sensors-and-journalism
（15） 首都大学東京・渡邉英徳研究室，台風リアルタイム・ウォッチャー，2015. http://typhoon.mapping.jp/
（16） 赤倉優蔵，“データジャーナリズム概論　ニュースを変革する新たな報道手法，”情報管理，vol. 58, no. 3, pp.166-175, 2015.
（17） J. Larson, A. Shaw, and L. Beckett, “Message machine : “you probably don’t know janet”,” 2012. http://www.propublica.org/special/message-machine-you-probably-dont-know-janet
（18） M. Graham, “What can Twitter tell us about Hurricane Sandy flooding? visualised,” 2012. http://www.theguardian.com/news/datablog/2012/oct/31/twitter-sandy-flooding
（19） 岡崎直観，佐々木　彬，乾　健太郎，阿部博史，石田　望，“ツイッター分析に基づく福島県産桃に対する風評の実態解明とその対策，”第 26 回日本リスク研究学会年次大会，B-5-3，2013.
（20） K. Finley, “In the future, robots will write news that’s all about you,” 2015. http://www.wired.com/2015/03/future-news-robots-writing-audiences-one/
（21） 小早川　健，“視聴者の意見を把握するための評判分析技術，”NHK 技研 R & D, no. 139, pp. 30-37, 2013.

（平成 27 年 11 月 4 日受付　平成 27 年 11 月 26 日最終受付）

岡崎　直観（おかざき　なおあき）

2007 東大大学院情報理工学研究科博士課程了．2005 英国テキストマイニングセンター・リサーチフェロー，2007 東大大学院情報理工学系研究科・特別研究員を経て，2011 から東北大大学院情報科学研究科准教授．専門は自然言語処理，テキストマイニング，機械学習．