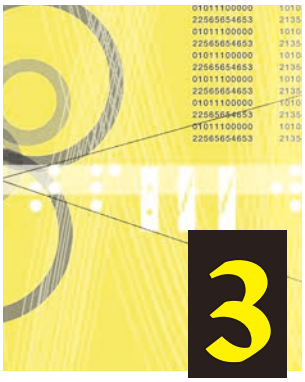


公共データ活用とデータサイエンティストのかかわり

―オープンデータ利活用の発展に必要な要素―

応用

工藤卓哉（アクセンチュア（株））

オープンデータ活用促進の鍵

❖ データ解析とオープンデータの出会い

企業内データの活用や，分析レベルの成熟度が進むにつれ，米国の新興企業などでは，公開されているオープンデータを結合して新規事業を立ち上げる動きが加速してきている．リアルタイムに生成される大規模ログデータを，従来の企業内構造化データとかけ合わせ，そこに専門性の高い解析担当者やオペレーションズ・リサーチの担当者の技術力がそのデータに出会うことで，何かこれまでには想像もしなかった新しい知見やビジネスモデルが得られるのではないだろうかという期待の高さの表れなのかもしれない．

❖ 促進の鍵は解析担当者の「企画力」

米国では，Google 社の社員が設立したインターネット保険会社 The Climate Corporation が，米国国立気象局（National Weather Service）と，米国農務省（U.S. Department of Agriculture）が公開しているオープンデータ等を活用した農業向けのリスクヘッジ保険商品を提供している．ここで忘れてはいけないのは，単にデータがあれば何かできるという類のものではないということだ．利用する側の着眼点・企画力と，正しい統計学的な数理モデルの手法を適切に適用できる技術力を持った人材＝データサイエンティストの存在が不可欠である．昨今，脚光を集めているデータサイエンティストは，ある特定の数学や統計，IT の分野に詳しいだけではなく，解析する目的をきちんと把握した上で仮説を立て，データから導いた意思決定の判断基準を，最適化プロセスを通じて実務にまで落とすことが求められる．また，情報処理基盤や解析手法についての知見に加え，適用する業界についての知識や卓越したコミュニケーション能力も必要だ．

自由度の高いオープンデータを実業界で有効に利活用させていくためには，提供する行政側の努力に加えて，それを形にする利用者側の工夫，企画力と技術力，さらにスケーラブルな分析処理基盤が融合する必要がある．それで初めて，データが商用向けに使える価値になるのだ．

❖ 利便性と透明性の確保が必須

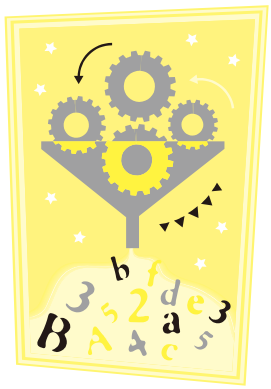
筆者は米国ニューヨーク州ニューヨーク市で複数年にわたり，公共政策立案に携わった．公共データの提供では，米国が，やはり一歩進んでいる．米国では単にデータをオープンにしている分野が幅広いだけではなく，公共データとしてオープンにする際のさまざまな配慮がなされている．

たとえば，統計値や生データの利用者には，そのデータがどのように収集され，計算処理を加えられたものであるのかを深掘りして知りたいというニーズがある．そうしたニーズに応えるべく，然るべき法規制に基づいて（Freedom of Information Act. 以下 FOIA とする），どのようにデータを収集・集約・判定処理したのかを開示要求できる道筋を提供しており，筆者も幾度かその開示要求を受けて応えたことがある（ただし，国防関連や FOIA の Section 552(b) に規定されている秘匿性の高い情報については，一部開示要求から免れる例外を設けている）．つまり，開示する側には透明性を担保する責任を持つことが求められており，都合の良い統計値だけ出

すことはできないということだ．出す以上は使う側から要求されれば，その源泉データも含めて開示要求に応えるという説明責任が伴う．これにより，利用者はデータに対する信憑性を確保できる．このようにして利活用へのインセンティブが保たれるよう配慮がなされている．

さらに連邦政府の提供するサイト data.gov では，2013 年 9 月時点で 100,000 弱のデータセットを公開している．通常のダウンロードはもちろん，オンライン上で視覚化ツールが充実していたり，パワーユーザのために，API（Application Programming Interface）を提供して，外部インタフェース構築による情報活用を意図したアプリケーションの開発も容易になるようにしている．

❖ 日本のオープンデータの課題



一方日本政府は，e-Stat などの窓口があるが米国と異なり公開データの多くが著作権で保護されているために，商用目的での自由な使用に対する敷居が高い．しかし経済産業省が 2013 年より立ち上げた Open DATA METI が，2013 年 9 月時点で約 200 程度のデータセットの提供を開始した．総務省も 2013 年 4 月より情報通信白書のオープンデータ化に踏み切った．これまで制約が多く，真の意味での利活用が促進されていなかった公共データの活用に活路が見出される可能性が出てきている．また，Wikipedia のようなコミュニティデータポータルとしての“the Datahub”はオープンデータを推進している英国の非営利団体によるプロジェクトで，2013 年 9 月末時点で，14,000 以上のデータセットが登録されて日々増加している．また，海外のデータサイエンティストからは「英語でのデータ提供はないのか？」という指摘を受けている．利便性の提供が国内利用者だけに向いていては取り組み自体が小さいものになってしまう．海外からの利活用を促進する仕掛け作りも必要になろう．日本のオープンデータもこうした各国のオープンデータの動向に足並みを合わせ，常に信憑性や利便性を高めていくことが求められる．

産学官連携で新市場形成へ

❖ BtoB や BtoC 等の軸で利活用を模索

オープンデータといえば前述した政府や自治体の公開データが論じられがちであるが，広義には制限はあれど API が提供されクロールで取得が可能なソーシャルネットワークのデータも含めることは可能である．BtoB と BtoC では，その利用シーンが異なる可能性がある．

BtoC におけるオープンデータ利用の代表例はソーシャルメディアに関するデータの利活用である．第三者向けのオープン API には検索対象期間や時間当たりのクロール可能取得データ量に制限がある場合もあり，また Twitter が提供する Firehose というデータ取得の仕組み（日本では NTT データを通じて再販を受ける）のように有償サービスとなっているものもあるが，さまざま提供されている．筆者らがソーシャルメディア分析サービスを提供する中では，源泉データに「癖」もあることが分かってきている．すべてが万事に使えるという類のものではなく，オープンな情報の信憑性が求められることも少なくないため，慎重にその源泉を調査し，情報の粒度，鮮度，カバレッジの広さをまとめた通知表のような，ソーシャルメディアのクロール先に対する評価結果も合わせて提供している．源泉によっては即時性が高い反面，信憑性や，正確性，粒度の細かさが低いなどさまざまである．

こうしたオープンデータがカバーする範囲は広いものの，各企業内のサービス対応履歴や，保守を伴うような複合機等の機器センサログ，スマートメータ，稼働消費電力量，顧客データベースや ATM 取引明細，顧客のカード ID が付加された，POS ジャーナルデータである ID-POS や，コールセンタ等の

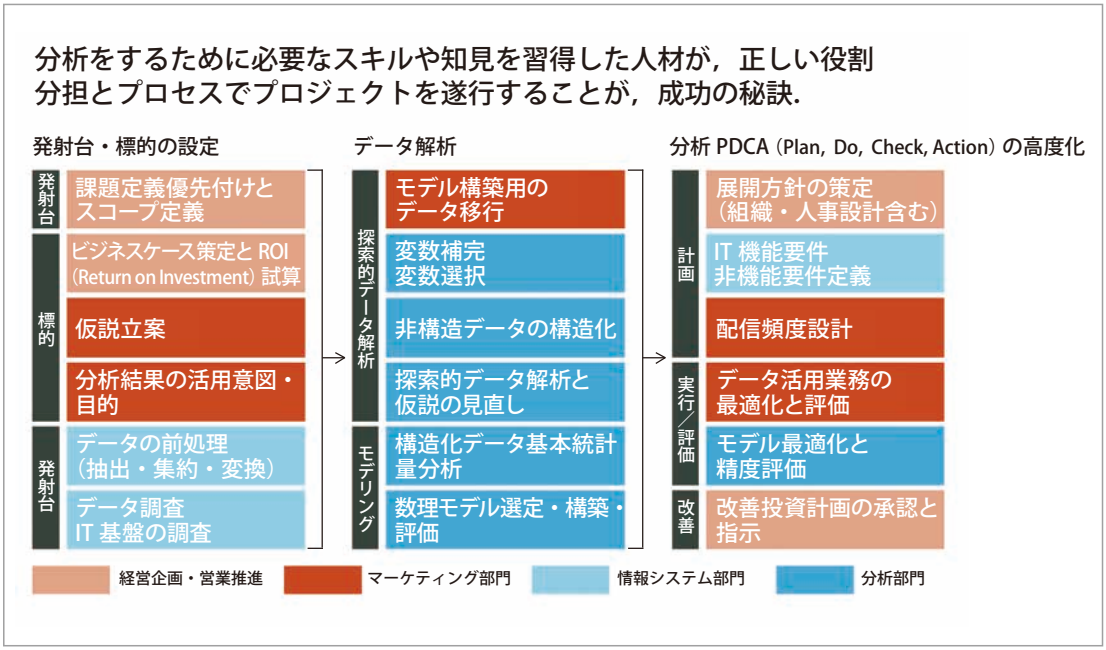


図 -1
マーケティングデータの分析プロジェクトの全体像

ヒアリングデータなど，各社が保有している構造化データに劣ることは否めない.

ポイントは，各データの持つ良し悪しと，取得にかかわるコストを意識した上で，情報の粒度，鮮度，質等を複合的に理解し，社内データと結合して有益な洞察を抽出するように，解析処理をデプロイできるか否かにかかっている．結局活用されて真価が初めて問われるデータは，膨大なデータをどう目利きして利用すべきかという問題に突き当たる．単純にオープンデータ，企業内データという括りだけでものを考えていては，利活用への道のりは遠のいてしまう.

❖ データサイエンティストの役割

こうした溢れかえるデータを前に，企業や組織はその有効な利活用を模索している．データの質や生成速度，ボリュームを適切に評価し，それを企業内データと結合して利活用を促進する 1 つの方向性として，最適人材の配置，すなわちデータサイエンティストのような目利き力を持つ人材の配置が求められる．**図 -1** は，一般的な分析プロジェクトにおいて，対応を迫られる複数部署と，そのタスクをまとめた全体像である．作業工程は実に複雑多岐にわたるが，まず肝要な点は，分析プロジェクトを開始する際に，その分析プロジェクトによって，どのような意思決定につなげるかの目的を明確にした「標的」を定めることだ．「データはたくさんあるので，そこから何かの知見を導き出してほしい」との声をよく聞くことがある．しかし，目的が明確でないままプロジェクトをスタートさせても，標的が定まっていないがゆえに，往々にして時間とカネの無駄につながってしまうことになりかねない.

また，「標的」を正確に狙うための「発射台」を整備することも忘れてはならない．いかに明確な目的を持ったとしても，それを導くためのデータが欠損していたり，分析するための IT 基盤が整っていないと成功に向けた足元が揺らいでしまうのだ.

分析プロジェクトにおいては，この「標的と発射台」の設定から始まり，実際のデータ分析や，分析で得られたモデルを実務の運用に落とし込み，さらにそれを最適化させる PDCA（Plan，Check，Do，Action）サイクルの高度化にいたるまで，さまざまなプロセスが必要だ．また，異なる部門間との調整も欠かせない．データサイエンティストには，チームワークやコミュニケーション能力もさることながら，プロジェクト全体を俯瞰した上で最新の業界・IT 動向にも注意を払い，その間を繋ぐ目利き力が非常に重要な素養として求められてくる.

たとえば，アクセンチュア（株）の製品保証やサービスサポートにおけるソーシャルメディアアナリティクスサービスでは，企業内コールセンタのサービス別不具合の問合せに対し，複数のオンライン機械学習の線形分類器（例：adaptive regularization of weights algorithm（Koby Crammer 著，2013）など）をスコアリングエンジンとして処理ノードにデプロイしている．一般化線形モデルを適用しているアクセンチュア（株）のモデルでは，そのデータを量的変数に整える前処理が必要となる．また，オープンデータだけで満足いく解析結果が出ることは難しい．まず前処理ではサービス特定をするような分かち書きの基準，表記ゆれの名寄せ，辞書の作成，正規性表現等をある程度標準化し，時に多項移動平均法によるノイズ排除を入れながら，真に意味のある即時性事象を大まかに識別し，不具合や不満の出ているサービスの特定を高い精度で抽出する．しかしその詳細な分析や原因の特定では，企業内データのコールログや，コンタクト履歴の積分値を集計し，長い期間に積みあがった不満等を見抜くことが必要になる．これらの作業にもデータサイエンティストのような存在が必要になろう．

❖オープンデータであれ，企業内データであれ，データの限界点を把握すべき

勘違いしてはいけないのは，データは単にデータであるということだ．Amazon が提供するレコメンデーションの協調フィルタリングのように，データプロダクトが売上に直結している場合を除き，ほとんどの企業においては，データ自体が意思決定自体をすることもなければ，サービスや製品の品質を自動的に向上するものではない．判断するのは人間である．オープンデータが注目され始めて久しいが，利用者側は常にこの視点を忘れてはならないだろう．長い期間に蓄積される積分値のような統計量では，たとえば近年では，商品やサービスに関する苦情や指摘をオープンメディア上で放置し，予兆を見逃してしまい社会的な問題に発展する事例も存在する．それもこれも，人間の判断によるところが大きい．ソーシャルデータが得意とする微分値の波であるバースト分析を通じて検知するだけでは深い分析はできないが，少なくとも捕捉可能な中で，意思決定者がその予兆を重要視するか否かは，判断力にゆだねられる．しかし，極力恣意性を排除し，統計により科学的なアプローチを提供できるようになれば，適切に判断を下すトリガとなり得，オープンデータを活用する価値は十分あると言えよう．今後のオープンデータの発展と，それを支えるデータサイエンティストの市場の成長に期待したい．

（2013 年 9 月 23 日受付）

■工藤卓哉
アクセンチュア（株）経営コンサルティング本部 アクセンチュアアナリティクス日本統括．

