シミュレーションの世界

データ市場：データ・AI・シミュレータ・人の連結場

大澤　幸生*

Market of Data: Connection of Data, AI, Simulators, and Humans

Yukio Ohsawa*

Key words: Market of Data, Communication, Innovation

データ市場とは，データの利活用により生み出される，あるいは期待される価値を取引する市場である．例えば，国内における過去の気象変化のデータは，ビールの総消費量の気象への依存を考える上で重要な意義を持つから，日本でビールを販売する業者やビアガーデンの経営者にとって大きな価値を持つ．一方，同じデータはパンの消費量に対してさほどの影響をもたらさないから，パンの販売者にとってはさほどの価値を持たない．この場合，このデータをビール業者は高い金を支払ってでも得ようとするから，データそのものに価値があるように見える．しかし，パン業者はこのデータに高い金を出さないのであるから，データそのものに固定の価値があると考えてデータの定価での取引を考えるのは難しい．むしろ，この気象データを用いて得られるビール売り上げの予測情報のほうにビール業者にとっての価値がある（パン売り上げの予測情報もパン業者にとって価値があるが気象データとは関係が薄い）のであり，データそのものの価値は金銭を尺度として定めることができないと考える方がデータにまつわる価値の説明は容易であり，市場に参加して金を実際に支払う人々にとっても納得できる説明となる．このため，データ市場という概念を，データ自体を金銭化する意味でのデータ取引市場とは区別してデータ市場を設計・構成することは，データの価値を評価してデータ取引を成立させるうえでも必須となる．

データ D の価値が仮に定義できるとすると，それは使用する目的 P や状況 S，利用者 H などに応じて定まるため value(D, P, S, H) など多様な引数をもつ函数となるはずである．目的 P はデータを直接的に扱う分析者や収集者の目的に限らず，データを用いて生み出される製品やサービスの顧客の要求に由来することもある．さらに，状況 S は多様である．例えば D が食品店における POS データであり目的 P が商品の売れ行き予測であるならば，状況 S は天候や市場のセンチメントなどの要素を含むことになる．もし SNS の内容からセンチメントを捉え消費者の欲する飲食品が推定できるならば需要予測者にとって相対的に POS データの価値は下がる．逆に，SNS の書き込みに特筆すべき変化もなく気温の上下動も緩やかなら市場変化を予測 / 説明する上で頼りは POS データとなる．いずれの場合も，POS データの価値が状況 S と目的 P によって変わってしまうため安定した価格づけが困難であるが，データによってもたらされる間接的な価値に注目するとこの問題は起きない．すなわち，データ D に対する value(D, P, S, H) ではなく D から派生的に得られる商品やサービス E の価値を value(E, P, S, H) とすると，E は fruit $(D_1, D_2, \ldots D_e, A)$ となる．ここで fruit は e 個のデータセット $D_1, D_2, \ldots D_e$ を組み合わせて構成した製品やサービスなどの成果体であり，A はその構成のために $D_1, D_2, \ldots D_e$ を利用するための分析や予測などの行為をさす．

すなわち，データ利用者 H によっても，H が用いる分析等の行為やそこで用いられる技術によっても，組み合わせるデータセット $D_2, \ldots D_e$ によってもデータ D_1 の発揮する価値すなわち式(1)の値が変わる．[1,2] などにもあるようにデータの二次利用による価値創出の期待は世界的な共通意識であり，既存の材の組み合わせから新たな価値を創出するというシュンペーターの新結合理論[3] とあわせて考えればこの式(1)は特に議

* 東京大学
The University of Tokyo

論の余地はないと考えられる．ただし式(1)は，どの項の引数からも P, S, H を省略した略記であり，D_{e+1} の価値評価はこれらの変数を含む様々な要素に左右される．単独のデータがそれ自体の価値を定めることは前提から外れる市場，それがデータ市場である．データそのものは，定価のつけられない商材になる．

$$\begin{aligned}\mathrm{value}(D_{e+1}) = {} & \mathrm{value}(\mathrm{fruit}(\cup_{i=1,\,e+1} D_i, A)) \\ & - \mathrm{value}(\mathrm{fruit}(\cup_{i=1,\,e} D_i, A)) \qquad (1)\end{aligned}$$

定価の付かないこのデータという商材を取引する限り，その市場の信頼性は参加者のコミュニケーションに委ねられざるを得ない．このコミュニケーションはどのような内容になるだろうか．例えば，天候のデータについては

「過去 50 年間の天候データをいくらで買いますか？」

「そうだな，私はパン屋だが天候データが何の役に立つかな？」

「パンの売り上げ予測にはあまり役に立ちませんね．でも，小麦と酵母の状態は夏が良いようです．もっとも，食べる側は蒸し暑い夏日にパンは敬遠するかも知れませんが」

「では，夏に作られたパンを涼しくなる日の前に大量に仕入れて翌日売ろうか」

「しかし，それは需要がそれほど多くないのに大量に供給する無意味な戦略では？」

この例でも見られるように，データに関する会話はデータそのものの話題に閉じずにデータを使った製品やその販売の価値評価に向かって進んでゆく．天候データとパンの売り上げの関係を学習させることはありえるが，そのような手元にあるデータから得た目星だけで進めたデータ分析の結果は実行性も市場性も乏しい「無意味な戦略」になってしまう．よってデータの価値は，必ず value（fruit）を含めて検討しなければならない．このように，消費者の目に吹きさらされた製品・サービス市場がデータの価値を決める手順でデータの価格を設定しなければデータ取引は不適正化し，市場は空洞化してゆくのであるから，データ取引市場の安定運営は容易ではない．**図 1** は，異業種のデータ取引を司る事業者の開業件数が近年，大幅な増加と下降を繰り返しており，安定性に欠けるデータ市場発展の様子を示している．このような不安定さを受け入れながら，データ市場の参加者は式(1)の略記において省略した人 H としての役割を果たし，value(D_{e+1})の評価において取り入れる必要が状況 S や目的 P を表出化するコミュニケーションを行うことが求められるのである．

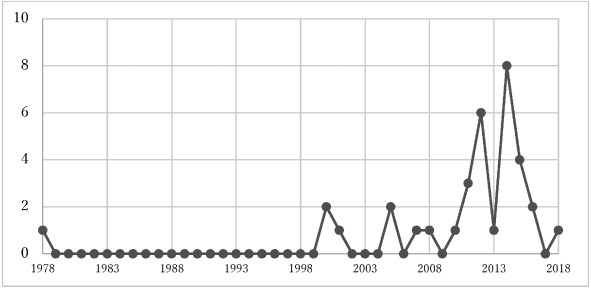

図 1　データ市場運営事業者の開業件数の変遷([4]の調査から．ただしこの図では累積ではなく各年の開業件数を示した)

さらに，式(1)を注意深く見てみよう．左辺の value(D_{e+1})を，他のデータセット value(D_1)から value(D_e)までの全てに書き換え式(1)に書き下して左右の辺をそれぞれ合計して比較すると，引数としては行動 A だけが右辺に余ることになる．すなわち，実際には value(D_{e+1})は変数 P, S, H を考える以前にデータユーザの行動 A を含む value(D_{e+1}, A, P, S, H) と記されるべきものであるが，A は市場参加者，特に fruit の利用者から理解困難なデータ利活用に用いる AI などの技術を用いる行為であるため value の評価は通常困難となる．ひとつの考え方は，式（1）の右辺の各項で value の前に $\max_A$ を挿入することであるが，A を枚挙することはやはり困難である．

以上のようなことを視野に入れて，既に存在するデータ取引市場の運営事業者は，データ市場への参入者にとっても窓口として機能するため，後述の様に検討手法も導入しつつある．運営事業者の業態は，㈱日本データ取引所のようにデータ取引市場の運用に専念する企業のほか，エブリセンス㈱，オムロン㈱のようにデータ取引以外の事業を行う企業がデータ取引市場を運営するなど様々である．しかし，一般の企業買収について買収側が被買収側の所有したデータを目当てとしたというような解釈は適切ではない，実際の買収者はデータのみならず，そのデータを用いた被買収者のサービスそのものも活かすからである．データだけを切り離してビジネスを解釈する考え方は，データ市場のダイナミクスを見誤る要因となる．データ市場でのコミュニケーションには多様なステークホルダーが参加し，データの有意義な組み合わせを実現するために上流からも下流からも多様なデータを受け入れてデータ利活用目的(P)と組み合わせ案を検討する必要がある．この意味で，データ市場はデータの取引に意識を置きすぎず，情報の粘着性[5] を抑制しつつイノベーションを興すコミュニケーションの場として構成してゆくべきである（**図 2**）．データ市場は古くからある概念であり[6] 国際競争が進んでいるだけに，この

意識改革は急務である．

多様な組織や個人を巻き込むデータ市場を運営する上での障壁としては，データの価値評価の不安定さのほかにも，①データを供出する難しさ　②データ利活用方法を検討する難しさ　③②で発案した利活用を実施することの難しさ　④③におけるデータ分析の結果見出したビジネスシナリオを実施する難しさ等がある．①には，個人情報保護法や欧州における GDPR 等のようにデータ所有者のおかれた社会的あるいは制度的な制約と，他社へのデータ供出による機会損失に対する警戒という最低二つの理由があるが，実は共に②に関係している．なぜなら，個人情報の取り扱いに関する法的制約においては通常，データの利用目的をデータ提供者に事前に説明することが義務付けられているが，実際にはデータ利活用方法はデータ提供を受けたのちに見いだされることが多く，完全に事前説明することは難しいからである．また，供出したデータを潜在的競争相手が利用するシナリオが不明のままでは，機会損失への過剰な警戒が起きビジネスの障害となる．そもそも，データの利活用方法ならずとも，新たな着想は組織内で合意されにくい．ゆえに③④の根源は②にある．

このようなわけで，②すなわちデータ利活用の目的や方法を効率的かつ多視点で表出化する手法，特に異種データの組み合わせを含むプロセスが導入されるようになった．例えばデータジャケット（DJ[7,8]）は，データの内容や含まれる変数（属性），利用価値への期待感を端的に記載したテキストである．データの内容を供出できない所有者は DJ だけを書いてデータ市場に提供する．集まった様々な DJ 間の結合可能性が可視化される（**図 3**）ので，データ市場参加者らはデータの結合と利活用の方法を考えて商談を行うことが可能となる．すなわち，式(1)におけるデータセット D_1, D_2, ..., De について，データジャケット $\mathrm{DJ}_1(D_1)$, ..., $\mathrm{DJ}_{n1}(D_1)$, $\mathrm{DJ}_1(D_2)$, ..., $\mathrm{DJ}_{n2}(D_2)$, $\mathrm{DJ}_1(D_e)$, ..., $\mathrm{DJ}_{ne}(D_e)$ を記してゆく．先に述べたようにデータ市場はコミュニケーションの場であるから，各 D_i に対して多様な視点から説明するデータジャケットが複数（n_i）個書かれてもよいことを積極的に説明する必要がある．さらに，データを利活用するための技術も明示的にデータ市場におけるコミュニケーションの俎上に加えるため，様々な技術 $T_1, T_2, \cdots, T_f$ についてツールジャケット $\mathrm{TJ}_1(T_1)$, .., $\mathrm{TJ}_{n1}(T_1)$, $\mathrm{TJ}_1(T_2)$, ..., $\mathrm{TJ}_{n2}(T_2)$, $\mathrm{TJ}_1(T_f)$, .., $\mathrm{TJ}_{nf}(T_f)$ を技術の有識者から供給してもらっている．特に，データを単に過去データに基づく機械学習のための分析素材とするだけではなく，データ利活用者の未来の事業や社会を生み出してゆく構成素材とするためには，TJ に記載する技術群の中にデータから得たモデルに基づくシミュレーション技術を投入することも必要となる．現在までに，交通シミュレータ MATES[9] など社会システム構築にかかわる技術が，可変幅深の深層学習，クラスタエントロピーに基づく変化説明および予兆検出などとともにツールジャケットとして登録されている（https://sites.google.com/site/datajackets/）．

2019 年 2 月 20 日に，データ流通推進協議会（Data Trading Association：DTA）は産業界におけるデータカタログの標準仕様のガイドラインを公開した．DTA は 2017 年 11 月に産業界が起点となって立ち上がり，データ取引市場における（したがって正確には上記のデータ市場とは異なる）運営者の規則，技術基盤，データ利活用戦略について検討する協議会である．この仕様においては，DJ の一部を明示的に盛り込んでおり，

図 2　データジャケットを用いたデータ市場におけるワークショップの様子（Innovators Marketplace on Data Jackets[7,8]）．左はデータ駆動スポーツ支援のためサッカー，テニス，ラグビーのコーチを含めた実施．右はインド・Amity大学における実施（IEEE SPIN2017）．

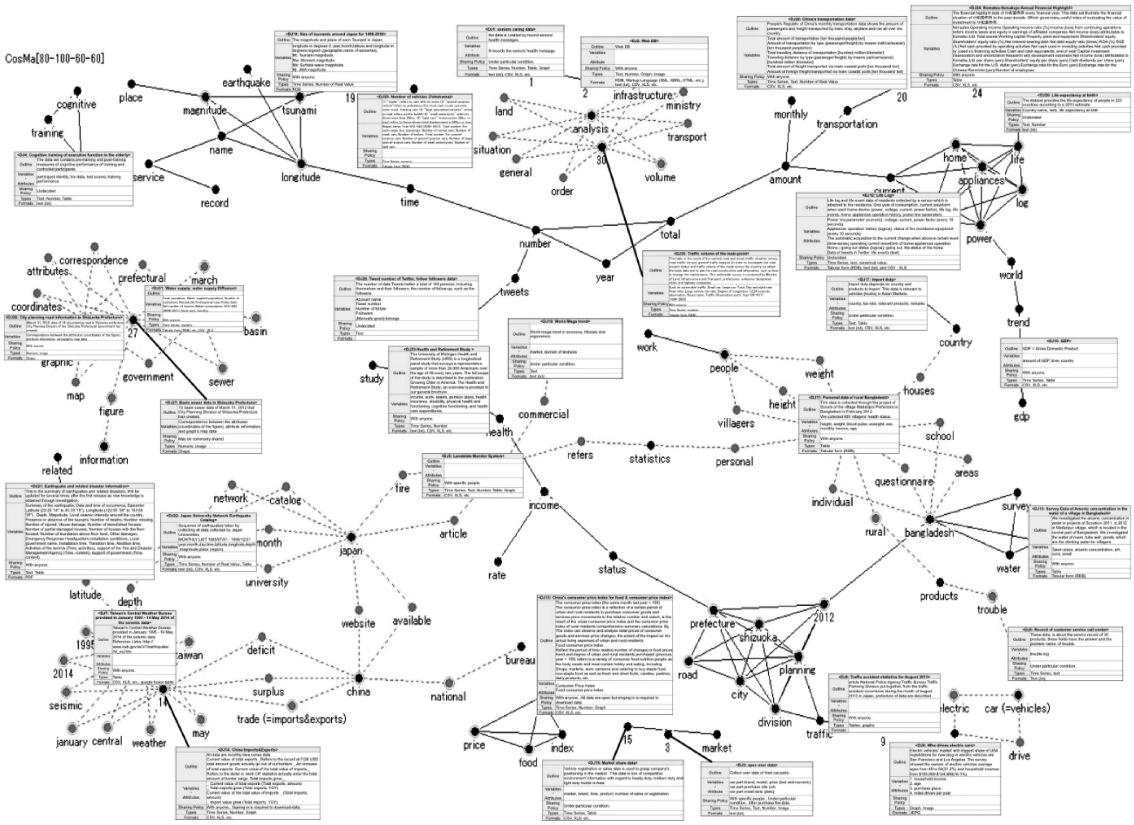

図 3　IMDJで用いるデータジャケットの集合：それぞれの四角形のカードがデータジャケットで，これらの間の結合可能性が KeyGraphで可視化されている[2)]

今後においてもコミュニケーションを活性化させる準備にも配慮している姿勢を察することができる．これまでに産業界からは，DJ をデータ書評のような著作物として扱う発言もあった．この考えが的を得ているのは，人が新しい DJ を創造し公刊するモチベーションを高める行為を通じて，新たなデータを設計してゆく過程も視野に入っていることを示唆しているためである．事実，DJ に基づくデータ市場である Innovators Marketplace on DJ（IMDJ）の副産物は，センサー開発や新規データの取得，これらのデータを利活用目的にあわせて分析するアルゴリズムなど様々である．データジャケットを用いたデータ市場の詳細は文献[8)] に譲る．

参　考　文　献

1）Rabinovich, E., Cheon, S.“Expanding Horizons and Deepening Understanding via the Use of Secondary Data Sources,”Journal of Business Logistics, Vol.32, No.4, pp.303–316, 2011.

2）Ellram, M.L., Tate, L.W.“The Use of Secondary Data in Purchasing and Supply Management（P/SM）Research,”Journal of Purchasing and Supply Management, Vol.22, No.4, pp.250–254, 2016.

3）Schumpeter, J.A., Theorie der wirtschaftlichen Entwicklung, Leipzig, Duncker & Humblot, Berlin（1912）

4）早矢仕晃，小口 裕，飛沢省二，大澤幸生，日本・米・欧州・中国のデータ市場ビジネスの動向，信学技報，Vol.118, No.453, pp.1–4, 2019

5）Hippel, v.E.（1994）.“Sticky Information”and the Locus of Problem Solving: Implications for Innovation. Management Science, 40（4）, 429–439（PDF）

6）New move into data market, Electronics and Power,Vol.26, No.8, pp.619, 1980.

7）大澤幸生：データジャケット ―創造的コミュニケーションのあるデータ市場のために― 人工知能 Vol. 29 No. 6, pp.622–627（2014 年 11 月）

8）大澤幸生（編著），早矢仕晃章，秋元正博，久代紀之，中村潤，寺本正彦，「データ市場」近代科学社

9）Yoshimura, S., Fujii, H.,“MATES: Multi-Agent based Traffic and Environmental Simulator -Core Technologies and Practical Applications-”, 8th Int’l Conf. Computational Methods（ICCM 2017）（2017）.