解　説

データマイニングの活用事例

髙　松　伸　一

1　はじめに

データマイニングとは，大量のデータに対し統計学や機械学習などのデータ解析技術を適用することで，潜在的な傾向や法則性などの有用な知見を獲得する技術である．これがうまく機能すると，熟練者が経験の中で気付いていた暗黙知，いわゆるカンやコツの領分を定式化して形式知に変えられることもあれば，誰も気付いていなかった事象の関連性を見出すことで，革新の礎になることもある．その名の通りデータ（Data）を採掘（Mining）することで，金脈を掘り当て得るこの技術は，データに溢れた現代社会の中でますます重要性を高めてきている．

この分野がどれほど注目されているかは図１を見れば想像が及ぶだろう．図はこの技術に関連する用語“データマイニング（Data Mining）”，“データアナリティクス（Data Analytics）”，“データサイエンス（Data Science）”と，参考として併記する当社基幹技術“油圧（Hydraulics）”，及び地球人類の見果てぬ夢“不老不死（Immortality）”の合計５用語について，過去５年間各週の検索件数を計上し，期間内最大値を100とした比率を表示したものである．先ず，iPS細胞などを通じて科学的に不老不死を議論できる今生の世にあってなお，どの語も人類の夢を上回る注目を浴びているという現実路線が伺えるほか，1980年代に概念化されたデータマイニングが産業革命時代から続く油圧と同程度に注目されていることがわかる．また，データアナリティクスやデータサイエンスといった類語は目覚ましい上昇傾向を示し，これらデータにまつわる３語は，今や油圧や不老不死に勝るとも劣らない注目を集めているのである．

KYBでは，データすなわち情報を効果的に利用する技術としてこのデータマイニングに着目し，社内への普及と定着を目指して技術適用と人財育成を推し進めている．この取り組みはまだ緒についたばかりではあるが，本報にてその意義を示すとともに，技術的成果の一端を紹介したい．

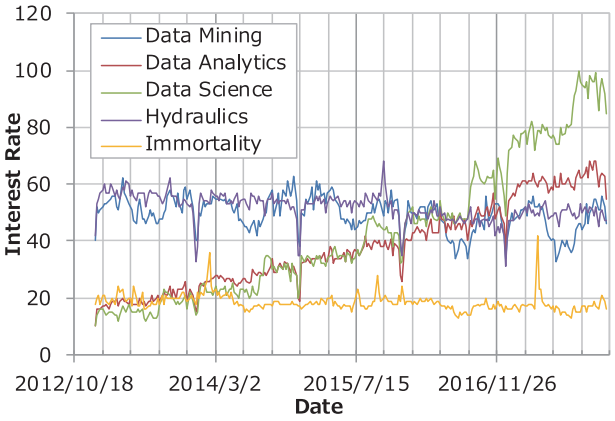


図１　Googleトレンド検索結果（著者の自主調査結果）

2　情報とは何か

データマイニングなどというキーワードに目を留められる方々の中には，“21世紀で最もセクシーな職業”としてデータサイエンティストが挙げられている[1]ことをご存知の方も多いだろう．これは元をただせば2009年にGoogle社のチーフエコノミストだったHal Varian氏による“the sexiest job in the next 10 years will be statisticians.”という言葉[2]に端を発している．データ分析が重要となる世の中の到来を表すこの言葉は，その鮮烈な響きもあって瞬く間に世界に広がり，もはや食傷を感じるほどに減価償却されたレトリックだ．もしこの言葉自体を知らなかったとしても，ビッグデータ，AI（Artificial Intelligence），IoT（Internet of Things）といった語に触れずに現代を生きることは難しく，誰もが何処かでこれらの語を見聞きしていることに疑いの余地は無い．こういった語の背景として横たわる共通概念が情報である．

情報という言葉は，陸軍参謀本部が1876年に訳した“仏国歩兵陣中要務実地演習軌典”にて，敵情報告の略語として著したことを始まりとする軍事用語である[3]．原語である“information”は，“心に形（form）を与えるもの”の意で，ある物事の内容や事情につ

いての知らせという意味を持つ．これらのことから想像が付くように，情報は受け取った者に新たな判断基準や考え方をもたらす存在であり，“物事に関する知識の不確定さを減少させるもの[4]”と定義される．

例えば雲一つない晴天の状況では，一時間後も晴天であるという情報は意外性が少なく価値が低いが，一時間後に大雨になるという情報は予想だにしないため価値が高く，予めこの情報を入手できたのなら傘を用意するなど，適確な判断を下す材料になる．

つまり情報は，一見してそのようになるとは思えないこと，換言すれば“発生確率の低いこと”を予見できたときにその価値を高めるため，情報量という概念はいわば驚き度として解釈できるのである[5]．したがって情報量は，容易に見通すことのできない不明瞭性の中に確かな事象を見通せる場合にもっとも高まるが，このような情報を入手するのは容易なことではない．

3　データマイニングの潮流

情報の意義を端的に表すモデルとして知の階層構造を示すDIKWモデル（図２）があるので紹介しよう．

このモデルは一般に言う情報を“データ”，“情報”，“知識”，“知恵”という階層構造に細分化し，上位に至るほど重要な要素として機能することを表している．例えば時刻や気圧，座標のような数値による“データ”を統合し，気圧配置の時間変化を表す“情報”としてまとめれば，低気圧は悪天候を招くという“知識”をあてはめて天気予報を実現し，一定以上の降水率では雨具を用意するという“知恵”に応用できる，といったことを意味するモデルである．

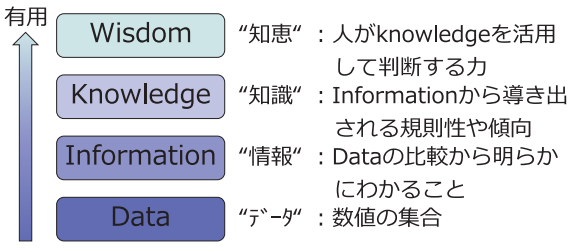


図２　DIKWモデル

このような“知恵”の発揮は，望ましい結果をもたらすための適切な判断を支援するため，“知恵”を発揮できない場合に比べて大きな優位性を与えてくれる．そして有用な“知恵”を獲得するためには，“データ”や“情報”を整理・蓄積し，活用可能な“知識”に昇華させておかなければならず，これこそが情報を入手しておく意義である．

このような情報の重要性は昔から認められていたものの，集めること，分析することの双方が容易ではないため，かつて活用の手段は限られていた．

しかしながら近年では，日進月歩で進む科学の発展によりその障害は大きく取り除かれてきている．図３，４に示すように，計算機の発展によって情報処理速度が飛躍的に向上し，集積回路の小型化技術の発展によって蓄積できる情報の桁数が爆発的に増大した事実に加え，インターネットの発展やセンサの進化がもたらされたことで，かつては集めることも扱うことも不可能だった巨大量の情報を蓄積・分析することが可能になった．然るべくして今，数多と集まる情報の山を効率的に分析する技術としてデータマイニングなどが注目され，その技術者としてデータサイエンティストが求められているのである．

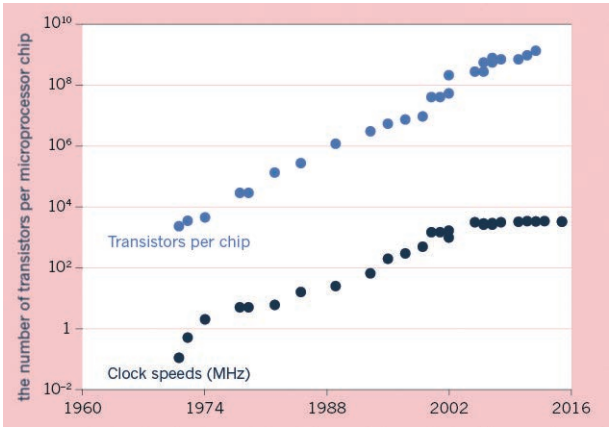


図３　コンピュータの処理能力向上の歩み[6]

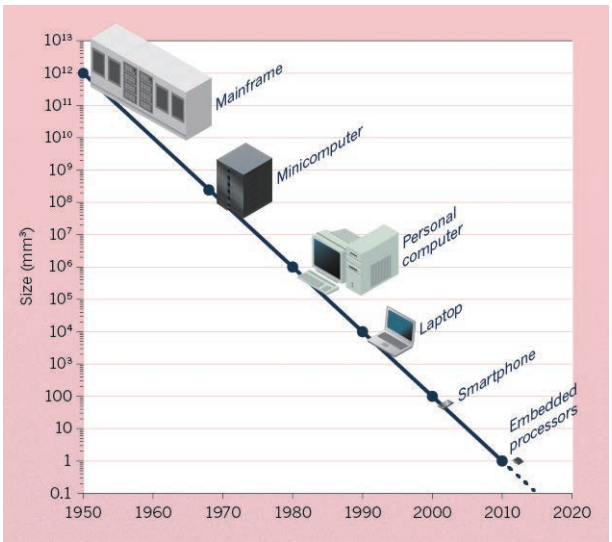


図４　集積回路小型化の歩み[6]

4　物理モデルと統計モデル

従来の工学的手段を用いた製品開発では，対象を物理モデルに置き換えて式を立て，パラメトリック

に解析することで最適値を模索していく．これは条件が単純であればあるほど精度の高い結果が得られるが，実際には実験誤差やモデル化誤差の発生を避けられず，これらがそのまま解析精度の低下に繋がってしまう．このため教科書通りの物理モデルを構築するだけでは不十分で，様々なノウハウを取り入れてモデルに修正を施すのが一般的である．

これに対し，データマイニングで用いられる統計モデルは物理モデルのように真値の生成構造を追求したモデルとは考え方が異なり，数式の構造や単位系に拘らず，入出力の関係を上手く近似することを目的としている．このため統計モデルとして得られる数式は，物理法則に基づいて現象を記述する支配方程式のような理論的説得力を備えておらず，工学的には邪道な取り口であると言えるが，複雑な実現象の中に含まれる隠れた法則性を示唆し，新たな知見を発掘する手法として有効に機能することも多い．

これら物理モデルと統計モデルとの位置付けの違いを対比すると，図５のように表すことができる．物理モデルでは入力を支配方程式に基づいた演算にかけることで演繹的に処理して出力を導く．つまり，“（１＋２）”という入力と演算のセットを利用して“＝３”という出力を求めるような処理をとる．対する統計モデルでは，入力と出力のセットをもとにそれらを結びつける演算を推定するため，結果から帰納的に遡るアプローチをとっている．“（１　２）＝３”という入出力だけを確認し，これらを関係づける“＋”という演算を推定するのである．

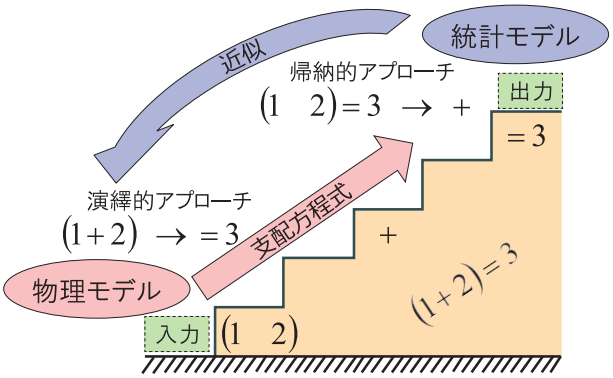


図５　物理モデルと統計モデルの対比

このように物理モデルと統計モデルは，現象に対するアプローチの方法が全く異なっている．従来，社会学や医学，心理学など，再現性の低い現象を扱う分野では統計モデルが活用されることが多かったが，工学分野では確かな法則性が認められる現象を扱ってきたため，専ら物理モデルを適用することが多かった．しかしながら，工学分野の根幹にある支配方程式は，自然現象の中で普遍的に確認できた一部の法則を体系化した式に過ぎず，複雑な実現象をありのままになぞらえて記述しているわけではない．前章で述べたように，近年では情報の蓄積と処理に躍進が起こっているため，実際の入出力という揺るぎ無い事実に対して統計モデルを適用することで，支配方程式に含まれない，ノウハウたるべき領分の知見獲得に繋がることも期待が持てよう．

5　事例紹介

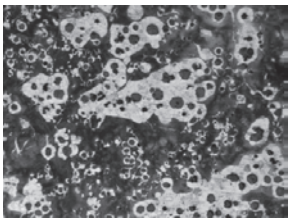
当社では，製造業としての取り組みの中にデータマイニングを活用することの意義を考え，社内に技術を普及・定着させるための活動を進めている．社内各所から様々な案件を集めて技術適用を試みるとともに，その結果を専門部署の技術者と検討することで効率的に新たな知見を見出し，これを練磨することで固有技術に転化させていく活動である．

この活動の中で携わった技術適用事例には，従来技術では解明できなかったであろう案件を解決に導いた事例や，対策案の早期確立に寄与した事例等，事業活動に貢献する様々な成果が挙がっているため，以下にこれらのうちいくつかを抜粋して紹介する．

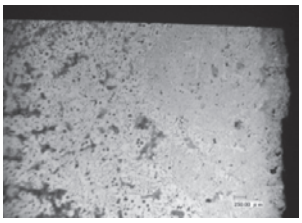
5.1　鋳造製品の異常組織対策

鋳造製品は材料の成分や形状，温度など様々な因子が絡み合い，品質の安定化が難しい製品の一つである．当社では，関係会社であるKYB-YS㈱の工場にIoTを駆使したQuality Traceability System（QTS）導入ラインを設けており，個々の鋳造製品に対し数百以上のデータが蓄積される体制が整っている．

このKYB-YSでは，ある時期において鋳造製品に異常組織（写真１）が多発し，廃却によるコストを看過できない事態となった．



(a)正常組織（球状黒鉛）



(b)異常組織（片状黒鉛）

写真１　鋳造製品の正常／異常組織

この不良問題について，鋳造技術の専門家を交えた技術検討では，異常発生部位が限定的であったことから，その周辺で発生する燃焼ガスに着目し，正常組織となるための黒鉛の球状化を阻害する元素の存在を疑うなど，専門の技術的知見に基づいた仮説立案を進めていた．しかしながら，これらの仮説に基づいた対策に一定の効果はあったものの抜本的な

解決には至らず，QTSデータ全体を俯瞰的に見るデータマイニングに白羽の矢が立った．

初動期に行った各種分析では，鋳造時の溶湯（溶融金属）の温度や，含有されている各元素成分の量など，様々な情報を含むQTSデータを加工し，定量的な分析に処すことのできる110の変数に整理したが，明らかな不良発生要因を見出すことはできなかった．しかしながら，鋳造は生き物であると言われるほど，気温や湿度などその日ごとの変化に対して諸条件の微調整を加える玄妙な技術であり，この観点を踏まえると異なる取り口が見えてくる．時期や製品ごとにデータを層別し，決定木分析など，不良発生有無の分岐条件を調査可能な分析手法を適用していくことで，次第に異常組織が発現する条件が明らかになってきた．それは各種の詳細条件を省いて単純化すれば，図６のようにまとめられる．図は層別を進めたある時期の製品に対し，注湯工程にかける時間と注湯温度とを二軸にとって散布図を描いた結果である．不良品は，良品よりもやや図中左上に集中する傾向があるため，注湯時間が短く温度が高いことが不良の発生率を高めていることが伺える．

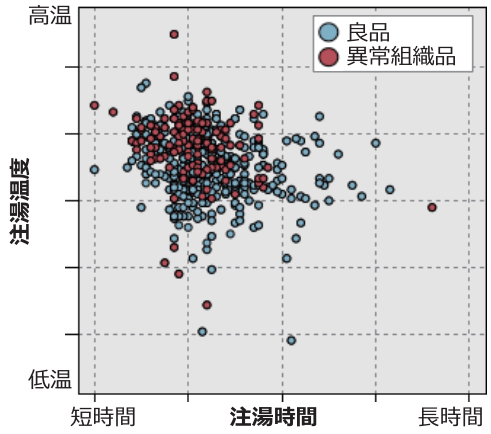


図６　異常組織の発生傾向

ただし，これがわかっただけでは対策には繋がらない．実際のモノづくりでは製造条件を変更することの背反として表れる他の不良現象もあるため，様々な現象のバランスを取って条件を定める必要がある．決定木分析では，その他に影響を及ぼさずに異常組織を減らすというように，条件変更による背反事象の発生をも踏まえた分析を行うことができ，本件ではこれが効果的に機能した．注湯時間や注湯温度のほか，着目すべき所与の因子を効率的に洗い出し，技術的見解を交えて条件変更の均衡点を探ることで，不良を撲滅するに至った事例である．

5.2　コンクリートミキサ車の状態判定

当社の販売製品の一つであるコンクリートミキサ車（以下ミキサ車）は，建設産業にとって欠かせない役割を担う製品であり，その効率的な運用を図ることのメリットは大きい．このためミキサ車をより効率よく運用する目的で，ミキサ車が保有する計測信号を活用して運行状態を把握する試みを進めている．

ミキサ車にもともと実装されているドラム回転数やドラム駆動圧力等のセンサ信号を収集・加工することで，27変数からなる時系列データが構築される．このデータを分析して変数同士の振る舞いを総合的に解釈し，各時刻においてミキサ車がどのような状態にあるかを判断するという課題があった．このデータのイメージは表１のようなものである．

表１　コンクリートミキサ車の状態判定データ

時刻	状態			変数		
T	Y_1	Y_2	…	X_1	…	X_{27}
t_0	TRUE	FALSE	TRUE	…	…	…
t_1	FALSE	FALSE	TRUE	…	…	…
…	…	…	…	…	…	…

試験時にはミキサ車の状態を任意に操作できるため，表１におけるYの状態判定に正しいデータを与えることができるが，実際の稼働時にはこれらの情報は得られないため，センサ信号から得られる変数Xのみを用いて正しい状態を推定する必要がある．しかし，同じ状態にあってもミキサ車の積載する生コンの量が違えば変数の関係は大きく変わり，また作業員ごとの操作量の違いなどによっても変化してしまう．このため状態を正しく判断するための普遍的な基準を設けることが難しく，従来の方法では精度よく判ずる基準を定めることができなかった．

これに対し，データマイニングを用いて判定基準を模索したところ，線形判別分析で約90%，機械学習による非線形判別ではほぼ100%の精度で状態を判定できるようになった．線形判別分析ではどの変数が判定にとって重要であるかといった示唆が得られるため，技術課題の検討材料にすることができるが，一方の機械学習手法では高い判別精度が得られる代わりに判定基準は明らかでなく，工学技術としての検討には向かない．高精度の判定を行いながらも工学的な技術発展に寄与するよう，合目的的に手法を組み合わせて課題を解決した事例である．

5.3　オイルシールの最適製造条件調査

新たな製造技術を適用したオイルシールの開発では，従来の知見にない新製法が品質の不安定化を招き，射出成型によって製造される試作品の中に多数の不良品が発生してしまっていた．射出成型機には各部の温度や圧力等，製造条件に関する50もの状態

量が記録されており，多数の試作品の不良状態とともにデータが蓄積されていたが，思考によって関連性を見出すことができるほど単純な傾向は認められず，品質安定化に向けた製造条件決定は難航した．

これに対し，不良の発生有無を対象として分析を試みた結果が表２である．この表には，全４種の分析手法それぞれで作成された不良有無を判ずるモデルが，50種の状態量のうちどれを重要な変数として着目しているかということについてまとめられている．４種の分析手法はいずれも異なる統計量を基準にモデルを導く手法であり，本件ではどのような基準が適切であるかは不明であったため，とりあえず一通りの分析を行って結果を確認するという手段を取った．このように無暗な取り口であったとしても，４種の分析ごとに選択される変数を集計し，その結果を俯瞰すれば，どの手法でも重視されている変数の存在が自然と見えてくる．

表２　50変数の重要度に関する当たり付け

便宜No.	選択変数	分析手法				重要度（○の数）
		判別分析	決定木①	決定木②	決定木③	
1	X_1	○	○	○	○	4
2	X_3	○			○	2
3	X_4	○	○	○		3
4	X_5			○		1
5	X_7			○		1
6	X_8		○	○	○	3
7	X_{10}				○	1
8	X_{31}	○	○	○	○	4
9	X_{32}		○	○		2
10	X_{33}				○	1
11	X_{46}		○	○		2
12	X_{47}	○	○	○	○	4
13	X_{48}	○		○	○	3
14	X_{49}		○	○		2

この結果は射出成型に関する技術的知見を一切交えず，統計モデルに当てはめたデータ分析のみで見出される傾向であるが，これだけでも50あった状態量を14にまで絞り込むことができ，なおかつそれらについて重要度まで付与できている．

こうして数が絞られ重要度による優先順位の付いた状態量群であれば，個別に技術的な審議を加えることも可能となる．各状態量の是非について技術者に判断して貰うと，不良発生に関連するメカニズムに不整合がないことがわかったため，これら数種の状態量を実験計画法に割り付けて直交実験を行うことでさらに調査を進め，結果として真に重要な製造条件の特定，及び不良率の大幅減を実現できた．データマイニングと実験計画法を連携することで，分析的に発見した傾向を技術的知見に昇華させ，効率的な問題解決に至った事例である．

6　おわりに

6.1　結言

史上初めて中華を統一し不老不死の霊薬を求めた秦の始皇帝は，自身のみを指す一人称として“朕”を用いた．以来，朕は天子のみが用いる一人称となり，自身を朕と称す者は国家全体を導く者と同義である．朕は“兆し”を意味する語であることから，兆しが全体を導く，という意味深長な関係が興味深い．

データマイニングは効率よくこの“兆し”を探す技術と考える．溢れ返る情報の中に微かな兆しを捉えるもので，無垢の財宝を見出すわけではなく，法則として利用できるまでの製錬を要す．特に当社のような製造業では，統計モデルとして得られた結果を技術的に解釈し，その示唆するところを見極めて物理モデルに反映する手腕が欠かせない．これを疎かにすれば，未知の法則を短期的に実利適用できることにはなっても，全容を解明した広範な応用科学には到達しないため，真の技術力は培われず，むしろ長期的には発展の阻害要因にさえなるだろう．

現代を生きる我々が手にする情報の量は，もはや我々自身の手に余るほどに大きなものになり，多くの情報は処理されることなく捨てられている（図７）．

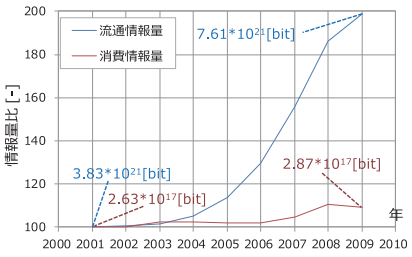


図７　2001年を100とした場合の情報量の推移[7)]

こういった日の目を見ない情報の中には，複雑すぎて容易に解釈することはできずとも，実は再現性高く存在するような法則が埋もれていても不思議は無い．特に，従来は別個の分野として独立して扱われていた領域の情報を統合的に分析したとき，未知の関連性が見出されるというようなことは大いに期待できる．これは元来，発明や発見の定石的パターンでもあるからだ．それ故，IoTをはじめとした情報収集基盤の整備を進めることで情報を蓄積し，データ分析を通じて効率よく新たな発見に触れ，確かなモノづくりの技術に落とし込む，という流れを軌道に乗せることが肝要と考える．この視座に立てば，まだまだ至らぬ点も多く目立つのが当社の実情

だが，先ずはデータマイニングの普及・定着を目指して啓蒙活動を進めることで，先々のために露を払い，後の大成に繋がるよう歩みを進めていきたい．

“大事をなさんと欲せば，小なる事を，怠らず勤むべし．小積りて大となればなり．凡そ小人の常，大なる事を欲して，小なることを怠り，出来難きことを憂いて，出来易き事を勤めず．夫れ故，終に大なる事をなす事あたはず．”[8]，まさに至言である．

6.2　謝辞

ここに紙面の一部をお借りし，本技術の普及活動にご尽力頂いている全社横断メンバの方々，その上司の方々，並びに技術適用の題材提供を通じてご協力下さっている多数の方々に謝辞を述べたい．皆様のご協力がなければこの活動は未だ暗中模索の中にあったものと思います．この活動が深みと広がりを持った取り組みとして形付いてきたことに感謝し，ここに御礼申し上げます．

6.3　付記

技術のような自然科学とは境を異にするが，異分野を統合して関連性を見出すことの意義深さになぞらえ，人文科学分野に垣間見える“兆しの意義”についての示唆を紹介しておく．兆しから法則を導くデータマイニングの観点に照らして考えて欲しい．

秦の始皇帝が手にした不死の霊薬は水銀で，彼は霊薬中毒で死したとされる．水銀は古来より神秘に関わる物質とされ，西洋の錬金術においても極めて重要な位置付けにあった．錬金術は金の物質的希少性・安定性に触発され，物質や肉体，精神を完全な存在に昇華させることを目指した学問だが，金アマルガムから蒸発させると金メッキができるという水銀の性質は，比重の概念を知らない古代人を，さぞや完全性への兆しと幻惑させたことだろう．

この錬金術を扱った著名な小説[9]では，前兆に耳を傾けることが道を拓き，夢に至るというテーマが描かれる．この小説は星の王子様[10]と並び称される不朽の名作だが，対する星の王子様に描かれるテーマは“大切なことは目に見えない”であり，ものごとをハートで見る意義が散りばめられている．

以上，意を汲んでいただけただろうか．人文科学分野にも完全性や夢といった無窮の目的があり，自然科学分野と同様，挑戦と失敗に彩られている．また，総じて見つけ難いこれらに至るには，見えざるものを見るための構えを要す，という示唆が万人に共感される名著の核となっており，これは失敗の中に気付きを得ることで大発見に結び付くという，自然科学におけるセレンディピティ[11]の概念に通じる．

最後の蛇足に，“意義”を意味する“significance”は接頭語として“sign”を含んでおり，signの意味は“兆し”でもあることを記して付記を結ぶ．

参　考　文　献

1）Thomas H. Davenport, D.J. Patil: Data Scientist: The Sexiest Job of the 21st Century, Harvard Business Review, (2012/10).
2）Steve Lohr: For Today's Graduate, Just One Word: Statistics, New York Times, (2009/8).
3）佐藤卓己：現代史のリテラシー～書物の宇宙,岩波書店,（2012年).
4）電子通信用語辞典，コロナ社，(1984年).
5）伊庭幸人：「情報」に関する13章―私家版・情報学入門―，物性研究，78-2，(2002年５月).
6）M. Mitchell Waldrop: MORE THAN MOORE, NATURE, Vol. 530, (2016年２月).
7）総務省：我が国の情報通信市場の実態と情報通信量の計量に関する調査研究結果，(2011年).
8）二宮尊徳：報徳要典，内外書房，(1934年).
9）パウロ・コエーリョ：アルケミスト―夢を旅した少年,角川書店，(1997年).
10）アントワーヌ・ド・サン＝テグジュペリ：星の王子さま,岩波書店，(2000年).
11）酒井邦嘉：科学者という仕事―独創性はどのように生まれるか，中央公論新社，(2006年).

著　者

髙松　伸一

2007年入社．技術本部基盤技術研究所要素技術研究室．技術士（機械部門）．振動制御を専門とし，油圧緩衝器等の研究に従事するとともに，統計データ分析の業務応用を推進．