◆連　載◆

《隔月連載 全５回》　　第４回

R & D 部門における機械学習・AI・生成 AI 活用へのデータ共有の重要性

上島 豊　　（株）キャトルアイ・サイエンス　代表取締役

《PROFILE》

略歴：

1992 年　3 月	大阪大学工学部　原子力工学科　卒業
1997 年　3 月	大阪大学大学院工学研究科　電磁エネルギー工学専攻　博士課程修了
1997 年　4 月	日本原子力研究所　博士研究員
2000 年　4 月	日本原子力研究所　研究職員
2006 年　3 月	日本原子力研究開発機構（旧日本原子力研究所）　退職
2006 年　4 月	キャトルアイ・サイエンス設立　代表取締役　就任

主な参加国家プロジェクト：

文部科学省 e-Japan プロジェクト「ITBL プロジェクト」,「バイオグリッドプロジェクト」
総務省 JGN プロジェクト「JGN を使った遠隔分散環境構築」
文部科学省リーディングプロジェクト「生体細胞機能シミュレーション」

主な受賞歴：

1999 年　6 月	日本原子力研究所　有功賞 「高並列計算機を用いたギガ粒子シミュレーションコードの開発」
2003 年　4 月	第 7 回サイエンス展示・実験ショーアイデアコンテスト文部科学大臣賞「光速の世界へご招待」
2004 年 12 月	第 1 回理研ベンチマークコンテスト　無差別部門　優勝

主な著作：

培風館「PSE book －シミュレーション科学における問題解決のための環境（基礎編）」ISBN：456301558X
培風館「PSE book －シミュレーション科学における問題解決のための環境（応用編）」ISBN：4563015598
培風館『ベタフロップス コンピューティング』ISBN978-4-563-01571-8
臨川書店『視覚とマンガ表現』ISBN978-4-653-04012-5

9　R & D 部門におけるデータ蓄積基盤としてデータベースがなぜ適切か？

前章では，「R & D 部門におけるデータ蓄積を Excel で行えるか？」と題し，R & D 部門におけるデータ蓄積が他の部門のデータ蓄積と大きく異なる点及び Excel でデータ蓄積していく場合の課題に関して，説明を行った。本章では，「R & D 部門におけるデータ蓄積基盤としてデータベースがなぜ適切か？」と題し，R & D 部門のデータ蓄積基盤として，データベースを採用すべき理由を具体的な事例とともに解説する。

データ共有・利活用をするためには，実験，分析の再現ができるレベルの記録すべき項目を明確化し，それらの項目を使って，どのようにデータを探査し，データ処理をするのかを第三者が実施できるレベルの手順書にまとめておく必要がある。実は，これだけでもデータ共有・利活用は可能である。そういう意味でいうと，データベース化は本質でないと言えばその通りである。実際，前章で説明したように Excel を使って，データ蓄積，共有することも可能である。ただ，Excel では，複数の人が同時に値を変更できないばかりか，変更や確認する必要のない実験（行）や項目（列）が表示されており，誤入力，誤編集リスクが大きく，データ品質が維持できない問題を抱えている。

データベースは，Excel のこのような問題を解決し，複数の人が同時編集及び誤入力，誤編集リスク抑止を実現する。実はデータベースの利点はこれだけではない。複数段の工程による実験のデータ蓄積，共有は，Excel では課題がある。例えば，合成工程でエチレン，プロピレンなどを混ぜてポリマーを合成し，配合工程でそれらポリマー３種配合し，配合材料を作成するような実験があったとする。そして，それは確かに次表のように合成工程の Excel と配合工程の Excel で記録できる。

合成 ID	エチレン濃度	プロピレン濃度	ブタン濃度	引張強度
Poly1	10	90	0	100
Poly2	0	30	70	150
Poly3	20	30	50	175
Poly4	50	0	50	175
Poly5	25	25	50	175

配合 ID	Poly1 濃度	Poly2 濃度	Poly3 濃度	Poly4 濃度	Poly5 濃度	配合温度	粘度
Sample1	5	5	90	0	0	100	102
Sample2	10	20	0	0	70	200	203
Sample3	10	0	80	10	0	300	304
Sample4	0	15	80	0	5	400	403
Sample5	0	20	0	40	40	500	505

この時，ある配合材料の粘度に関して，一番濃度の高いポリマー内のエチレン濃度に対する依存性を知りたいときにどうするのだろうか？配合工程の Excel において，一番濃度の高いポリマーを特定し，当該ポリマーが合成工程のどの実験で作られたものか，つまり，合成 ID を確認し，その後，合成工程の Excel において，その合成 ID の合成樹脂でどの程度濃度のエチレンを合成に使ったのかを確認すれば，必要な情報は得られる。しかし，知りたい「ある配合材料」というのが 100 個あれば，上記 2 つの Excel を行き来する作業を 100 回行わなければならないわけである。そして，当然そこで得た情報は覚えていられないので，別 Excel に一番濃度の高いポリマーのエチレン濃度などの情報を一時記録しておくに違いない。そして，100 回も調べるときっと何回かは間違って一時記録してしまうものもあるはずだ。一番濃度の高いポリマーのエチレン濃度は一時記録しておくといったが，後で他の量との比較やグラフを描いた分析をすることを考えると配合工程の Excel を複写したものに一番濃度の高いポリマーのエチレン濃度の項目を追加するのが，便利なはずである。

実は，同じように一番濃度の高いポリマーのプロピレン濃度やブタン濃度，二番目に濃度の高いポリマーのエチレン濃度やプロピレン濃度やブタン濃度も同じようにして，配合工程の Excel を複写したものに項目追加しておけば，その Excel 一つで他の量との比較やグラフを描いたりし，分析をすることは可能である。それなら最初から工程を分けずに 1 つの Excel で記録すればいいのではないか？と思うかもしれない。

しかし，1 つの Excel で記録しようとすると合成工程で大量に作ったポリマーをたくさんの配合工程で少しずつ配合比率を変えて行う実験を行う場合，固定であるはずの当該ポリマーのエチレン濃度，プロピレン濃度，ブタン濃度を配合実験毎に同じ値を記入していく必要がある。また，配合は，3 種のポリマーを配合するが，エチレン濃度，プロピレン濃度，ブタン濃度と書いてしまうとどちらのポリマーの組成のことを示しているのかがわからなくなるため，高濃度ポリマー内エチレン濃度，高濃度ポリマー内プロピレン濃度，高濃度ポリマー内ブタン濃度，中濃度ポリマー内エチレン濃度，中濃度ポリマー内プロピレン濃度，中濃度ポリマー内ブタン濃度，低濃度ポリマー内エチレン濃度，低濃度ポリマー内プロピレン濃度，低濃度ポリマー内ブタン濃度のような項目にして，記録するようにしなければならない。実際，エチレン濃度，プロピレン濃度，ブタン濃度という項目名だけでは不十分で，何のポリマーの組成の濃度かの情報がないとデータ分析などはできないのである。

図 21　複数工程の実験のデータ分析は結構難しい

実際，「実験の独立変数（実験で原理的に変動させうるパラメータ）は何ですか？」と聞かれた場合，合成工程と配合工程の実験パラメータを列挙して，エチレン濃度，プロピレン濃度，ブタン濃度，高濃度ポリマー濃度，中濃度ポリマー濃度，低濃度ポリマー濃度，配合温度の 7 変数と答えてしまいそうだが，それは間違いである。「実験の独立変数（実験パラメータ）は何ですか？」の正確な答えは，高濃度ポリマー濃度，中濃度ポリマー濃度，低濃度ポリマー濃度，高濃度ポリマー内エチレン濃度，高濃度ポリマー内プロピレン濃度，高濃度ポリマー内ブタン濃度，中濃度ポリマー内エチレン濃度，中濃度ポリマー内プロピレン濃度，中濃度ポリマー内ブタン濃度，低濃度ポリマー内エチレン濃度，低濃度ポリマー内プロピレン濃度，低濃度ポリマー内ブタン濃度，配合温度の 13 変数である。つまり，記録の煩雑さや誤記を減らそうとすると合成工程と配合工程を別々の Excel で記録していく方がいいが，データ絞り込みやデータ分析を行うために独立変数というものを意識しなければならない場合は，配合工程に合成工程を繰り込んだ 1 つの Excel で記録する方が良いのである。

工程ごとの項目を分けて記録をすると見かけ上，独立変数が減ってしまうので，実際データ分析をするときは，この 13 項目に展開しないと X-Y プロットさえ描けないのである。当たり前だが，X-Y プロットを書くということは，データが単純な 2 次元表形式になっている必要があり，2 つの表のようなものは，そのままでは X-Y プロットさえ，描けないのである。複数工程実験を複数の表で記録している人は，X-Y プロットを書くたびに複数の表を一つの表にまとめる作業をしているはずである。ただし，全ての項目を一つの表にするのは大変なので，X-Y プロットで使う，X 軸，Y 軸，凡例だけを取り出して，一つの表にまとめる作業をしている人が多いはずである。この作業が面倒な人は，良く使う項目だけをマクロを使って，自動的に一つの表にまとめる機構を作っている人もいると思う。しかし，このようにすることで，良く使う項目以外のデータ絞り込み，データ分析はより億劫になってしまい，分析範囲が無意識に狭まってしまうので，実は，研究としては望ましい状況ではない。

もちろん，最初から高濃度ポリマー濃度，中濃度ポリマー濃度，低濃度ポリマー濃度，高濃度ポリマー内エチレン濃度，高濃度ポリマー内プロピレン濃度，高濃度ポリマー内ブタン濃度，中濃度ポリマー内エチレン濃度，中濃度ポリマー内プロピレン濃度，中濃度ポリマー内ブタン濃度，低濃度ポリマー内エチレン濃度，低濃度ポリマー内プロピレン濃度，低濃度ポリマー内ブタン濃度，配合温度という項目名の 1 工程実験として，データを記録すれば問題ないのだが，通常，そういう記録は煩雑で好まれないはずである。また，2 工程程度ならまだいいのだが，5,6 工程にもなれば，項目名が幾何級数的に増えることになってしまい，恐らく正しく記録していくこと自体が困難になる。

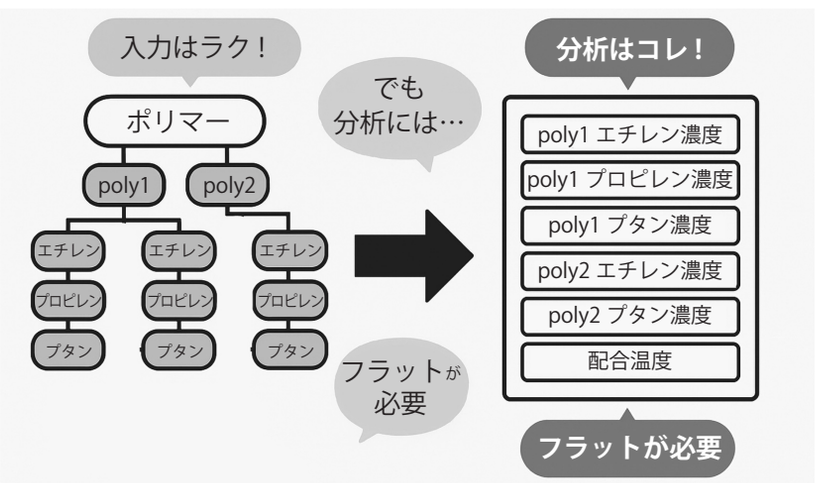

図 22　複数工程の実験のデータ記録と分析では項目名が異なる

つまり，データを共有し，利活用するためには，記録は工程ごとの Excel で，データ絞り込みや分析はそれらを 1 工程に繰り込んだ Excel で行える必要があるのである。そして，複数工程を 1 工程に繰り込むときには，工程間の紐づけ情報を分析し，項目名を内部で自動生成する機構が必要であり，そういう機構がないと複数段の工程のデータ検索，分析はできないのである。データベースというものはそういうことを内部で行えるようになっており，それがデータベースのもうひとつの大きな利点である。つまり，データ共有，利活用を考えると「複数の人が同時編集及び誤入力，誤編集リスク抑止」，「多工程データの紐づけと 1 工程への繰り込み」が最初から考慮されているデータベースを基盤に据えることが適切だということである。ちなみに，Excel でそういうことをできる仕組みを作り込むことも不可能ではないが，これら機構を Excel に作り込むというのは，ある意味，Excel をベースとしてデータベースを開発するのと同じことになり，車輪の再発明になるため費用対効果やソフトウェア品質の観点から止めておいた方が良い。

先ほど話をした，多工程データを 2 次元表に射影するには，どのような機構が必要なのかを少し説明しておく。例えば，モノマー調整工程，樹脂合成工程，配合工程，測定工程の 4 工程からなる実験を考える。例えば，測定工程の配合材料において，内部で使われているモノマーの分子量を表記したり，条件指定したい場合，同じモノマーが異なる樹脂に使われていたり，同じ樹脂が異なる配合材に使われていたりすると，計測に使った配合材内のどういう樹脂に使われたモノマーの分子量ということを指定できる方法（指定できる項目名）が必要になる。したがって，『計測で使われた計測使用配合材 ID ＝“F01”の中で使われた配合内樹脂 ID ＝“P02”の中で使われた樹脂内モノマー ID ＝“M03”の分子量』と指定することで初めて，項目として意味のあるものになるのである。

計測使用配合材｜F01｜配合内樹脂｜P02｜樹脂内モノマー｜M03｜分子量

したがって，R & D のデータ蓄積基盤では，上記のような項目名を自動生成し，表現できる必要があるということになる。

10 そもそもデータ蓄積，共有は何のために行うのか？

前章では，「R & D 部門におけるデータ蓄積基盤としてデータベースがなぜ適切か？」と題し，R & D 部門におけるデータ蓄積，利活用の観点からデータ蓄積基盤として，Excel では限界があり，データベースを採用すべき理由を説明した。本章では，「そもそもデータ蓄積，共有は何のために行うのか？」と題し，データ蓄積，共有の動機に関して，考察を行う。

まず，「なぜ，データを蓄積し，共有しようと考えたのか？」に戻って動機を見つめなおしてみよう。それは，恐らく，「自分の記憶の実験結果だけでは得られない有用な情報が，他者の実験や記憶から消えてしまっている自分の過去実験にあるのではないか？」，「そういう実験結果を簡単に確実に閲覧し，分析できれば，新しい傾向が見つかり，研究が加速するのではないか？」といったものではないだろうか？双方，漠然と参照できるデータが増えれば，「研究にポジティブインパクトがあるのではないか？」とか，「機械学習などの MI が使えるようになり，それらから良いアイデアが提示されるのではないか？」と思っているのではないだろうか？それは，大きな認識間違いである。参照できるデータが増えれば，自分の考えや推測に合う好都合な情報に見えるデータを選び出すことがより容易になり，それはある意味データを間違って解釈してしまう可能性を増やすことにつながる。機械学習などの MI に関しては，4 章で話をしたように必要とされる予測精度を確保できる適用領域を明確化すること自体が難しく，「良いアイデアが提示されるのではないか？」と思って，利用すると期待を裏切られ，利用者は強い MI アレルギーになってしまいかねない。

「有用な情報」とか，「良いアイデア」というのは，結果論であり，「他者の実験や記憶から消えてしまっている自分の過去実験」自身や「機械学習などの MI の予測値」自身は，常に「有用な情報」とか，「良いアイデア」であるわけではない。つまり，「他者の実験や記憶から消えてしまっている自分の過去実験」や「機械学習などの MI の予測値」が「有用な情報」とか，「良いアイデア」かを判定する装置なようなものがあれば，もちろん，データ共有は研究進展を確実に改善できるのであるが，そんな判定装置がない以上，そういうことに大きな期待を寄せてはいけないのである。それであれば，データ共有では

研究進展を確実に改善できることはないのだろうか？

結論から言うと，データ共有で研究進展を確実に改善できることはある。それはデータ探査，データ処理に網羅性，均質性，再現性を付与することである。そして，まさしく，データベースの価値はここに見出すべきである。逆に言うと，研究過程において網羅性，均質性，再現性のないデータ探査，データ処理は，研究の品質を著しく損ねるので，行ってはいけないはずである。素粒子実験や天体観測などの大規模実験では，「網羅性，均質性，再現性のないデータ探査，データ処理」は厳しく禁止され，監査されている。それは「網羅性，均質性，再現性のない」データ探査，データ処理の結果から導き出された仮説から次の実験を行うと無駄実験になる可能性が高いからであり，大規模実験では無駄実験は金額的にも相当大きい無駄になるからである。しかしながら，これは大規模実験にのみ適用されるべきことというわけではない，例え，再実験が安価なものであったとしても本当に間違ったデータ分析から導かれた無駄実験をしてもいいのだろうか？もちろん，データ分析時点でのデータが少なく，結果的に無駄実験になってしまうのは致し方なしであるが，実際にはデータが十分にあったり，全く同じ実験をしているのにデータ抽出，分析方法を間違ったために，そのことに気が付かず，無駄実験をすることはあってはならないはずである。小規模実験では，実際実験をする過程で，一番高価なものは研究者の時間のはずである。無駄実験で研究者の時間は相当無駄に消費されることになり，それがそもそも研究開発の進展速度を遅らせている大きな原因の一つのはずである。データベースにより網羅性，均質性，再現性を担保されたデータ探査，データ処理を使い，研究を実施することで，このような無駄実験はなくすことができるはずなのである。

データベース化は，これ以外にも作業者負担軽減，作業時間短縮というメリットもある。手動処理でも網羅性，均質性，再現性を簡単に達成できるような業務ではこれが主たるメリットであるが，R & D 部門に限って言うと，手動処理（ばらばらに保存されているデータが記録されたファイルをフォルダを駆け回りながら探し，ファイルを開いて・・・）で網羅性，均質性，再現性を達成できることはまずないため，作業者負担軽減，作業時間短縮はあくまで副次的なメリットであり，データベース化の目的は，データ探査，データ処理に網羅性，均質性，再現性をもたらすことと考えるべきである。以下で，網羅性，均質性，再現性に関して，具体的にどういうものか？研究にどのように影響してくるか？の解説を加えておく。

網羅性：手動で全データを対象に探査，処理を行うことは現実的ではないが，データベースだと全データを網羅した探査，処理が可能である。

例）熱伝導度が○○以上の材料を探したい場合に，記憶からいつぐらいの実験にあったはずと目星をつけて，いくつかのデータを確認する。しかし，目星が間違っていることもあるし，そもそも「熱伝導度が○○以上の全ての材料」に目星が付けられるわけではない。このような網羅的でない抽出データは傾向が偏っている可能性もあり，それを分析した結論は，信頼性を大きく棄損している。

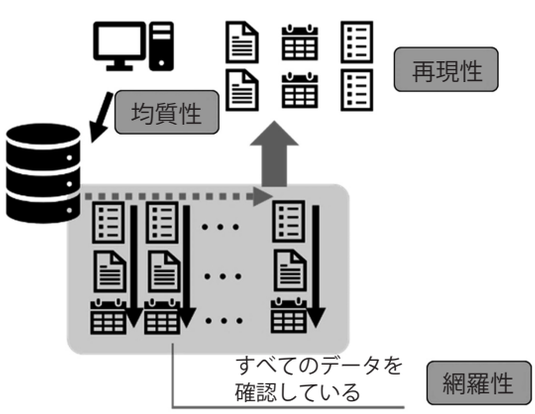

図 23　データベースは，網羅性，均質性，再現性を保証する

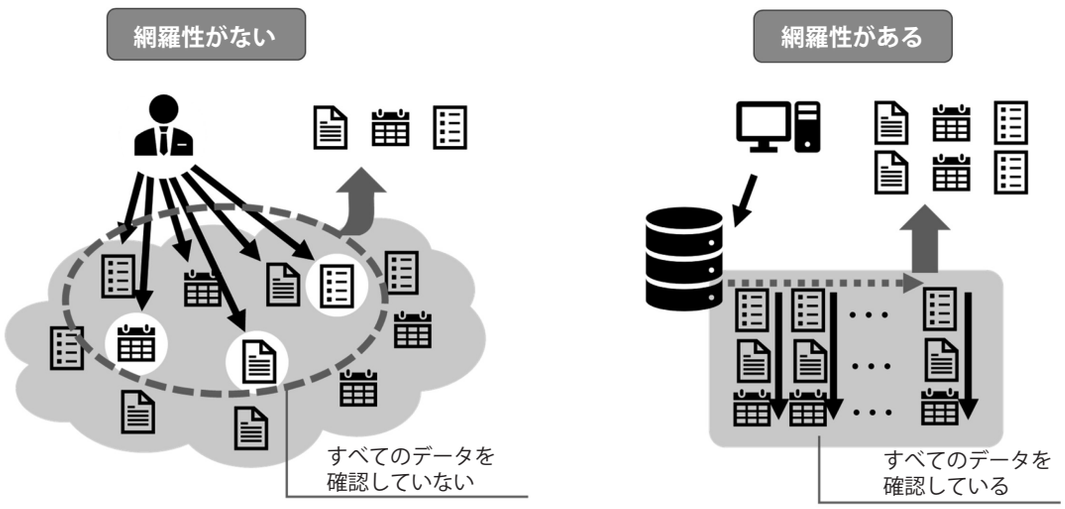

図 24　データベースの利点の網羅性とは

均質性：手動で探査，処理すると意識バイアスがかかったり，初期と終期での探査，処理の同一性が維持できないが，データベースだと完全に均質な探査，処理が可能である。

例）耐電圧が 500V 以上の材料を探している場合に，いくつかのデータを見つけた後，あまりにも該当データが少なかったので，480V 以上のものも含めるようにした。ただ，最初の方から探査をやり直していないので，いくつかのデータは取りこぼしている可能性がある。また，明らかに他の特性が合致しないデータは，500V 以上でも抽出しないことにしたが，明示的なルールで一貫しているかといわれると自信をもって Yes と言い切れないのではないだろうか。このような均質でない抽出データを分析した結論は，信頼性を大きく棄損している。

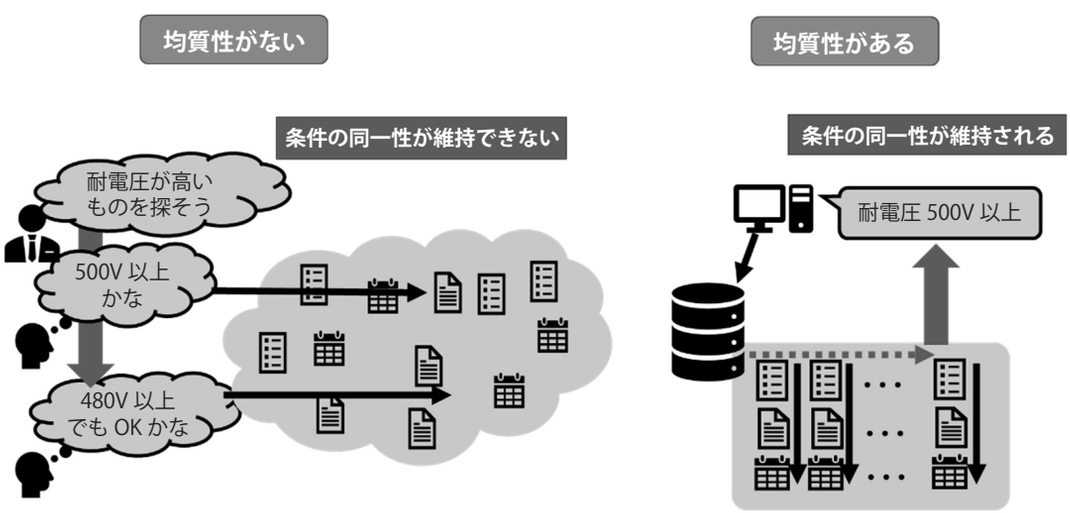

図 25　データベースの利点の均質性とは

再現性：半年前のデータ探査や処理は，手動ではまず再現できないが，データベースであれば，データ探査や処理の再現が可能である。

例）手動処理では，上記で述べたように網羅性も均質性もほとんどの場合で担保されていないので，当然，再現性も期待できない。こういう再現性のない抽出データを分析した結論は，信頼性を大きく棄損している。

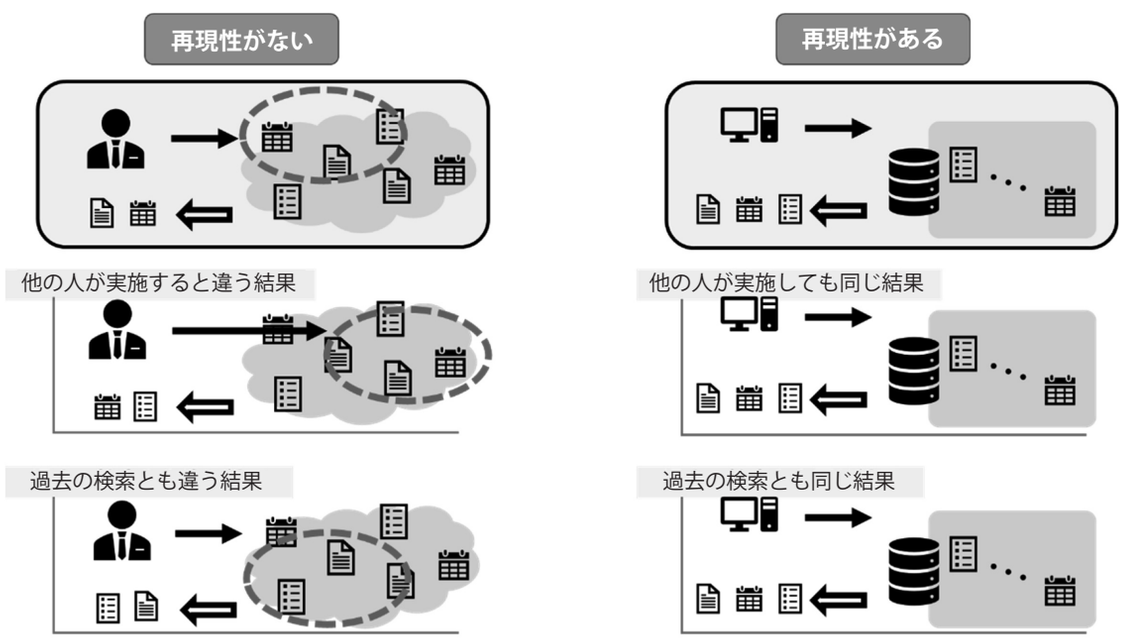

図 26　データベースの利点の再現性とは

参考文献

1）川田重夫，田子精男，梅谷征雄，南多善，上島豊，他
PSE book －シミュレーション科学における問題解決のための環境（応用編），川田重夫，田子精男，梅谷征雄，南多善　共編，培風館，（2005），p69-82

2）谷啓二，奥田洋司，福井義成，上島豊
ペタフロップスコンピューティング，
矢川元基　監修，培風館，（2007），p183-202

3）牧野圭一，上島豊，視覚とマンガ表現，臨川書店，（2007），p1-5，221-229

4）上島豊，月刊「研究開発リーダー」6 月号，技術情報協会，（2023），p63-68

5）上島豊，月刊「研究開発リーダー」7 月号，技術情報協会，（2023），p86-91

6）上島豊，月刊「研究開発リーダー」8 月号，技術情報協会，（2023），p78-82

7）上島豊，月刊「研究開発リーダー」7 月号，技術情報協会，（2024），p77-84

8）上島豊，月刊「研究開発リーダー」9 月号，技術情報協会，（2024），p73-81

9）上島豊，月刊「研究開発リーダー」11 月号，技術情報協会，（2024），p85-96

10）上島豊，月刊「研究開発リーダー」1 月号，技術情報協会，（2025），p74-82

11）上島豊, 月刊「研究開発リーダー」3 月号, 技術情報協会,（2025）, p68-73
12）上島豊，他
研究開発部門への DX 導入による R & D の効率化，実験の短縮化，技術情報協会,（2022）, p195-221
13）上島豊，他
ケムインフォマティックスにおけるデータ収集の最適化と解析手法，技術情報協会,（2023）, p39-74
14）上島豊，他
実験の自動化・自律化による R & D の効率化と運用方法，技術情報協会,（2023）, p159-199
15）上島豊，他
少ないデータによる AI・機械学習の進め方と精度向上，説明可能な AI 開発，技術情報協会,（2024）, p112-127