多々益々弁ず?

NHK のある AI の番組冒頭でマツコ・デラックスが「正しいデータにもとづいているのだから結果も正しいんです」と何度も叫んでいたが，この一言でその番組をそれ以上視る気は失せた．彼（彼女）は，シナリオライターあるいは番組制作者に言わされているだけかもしれないが，とんでもない発言である．裁量労働制データ問題を意識した一種の反語あるいはギャグだったのだろうか．

ここまでひどい論理でなくとも似たような論理はよく耳にする．「可能な限りすべてのデータを集めて得られた結果なので間違いありません」はマーケティングでよく用いられるフレーズである．確かに経理処理などではもれなくデータを集めることが重要であるが，データ解析では，どのようにして結論を導いたのかのほうが結果の正当性を左右する大きな要因となることを忘れてはならない．拙著『データ分析とデータサイエンス』でも引用させていただいた塩野七生著『ローマ亡き後の地中海世界 1』（新潮文庫）82 ページの 1 センテンスを再度引用させて頂けば，

> **情報とは，量が多ければそれをもとにして下す判断もより正確度を増す，とは，全くの誤解である．情報は，たとえ与えられる量が少なくても，その意味を素早く正確に読み取る能力を持った人の手に渡ったときに，初めて生きる．**

このフレーズは，中世前期にイスラム勢力が，手に入る僅かな情報からキリスト教世界の東方と西方の共闘の可能性はますます希薄になるであろうと推測した故事の説明に用いられているが，「情報」を「データ」に置き換え，「量が多ければ」を「正しければ」に置き換えても同じである．

「多々益々弁ず」確かである．データ量が多ければ多い分，そこからなにか発見できる可能性は広がる．しかし，それが正しい結論，役立つ結果につながるとは限らない．かえって，量が多い分，動きが鈍くなり深い解析を妨げる可能性も増える．大雑把に言って，結論の正確性はデータ量の平方根にしか比例しない．つまり，データが 100 倍になっても正確性は 10 倍にしか増えない．苦労して 1 万も 2 万も記録を集めても，結果は 100 記録や 200 記録と一桁しか違わない．ビックデータの罠はここにある．「データ量が多ければ多いほど役立つ結果が得られるに違いない」という思い込みだけではどうにもならない．まずは，数百の記録をサンプリングしてさまざまな側面からデータを探り，どんな結果が得られそうか，データの背後にどんなメカニズムが潜んでいるか見当をつけることから始めてみてはどうだろうか？　もちろん，そこではデータの正しさ，質を見極める必要もある．これについては別途お話しすることしよう．

2018-03-04 里程