Research Report

データ科学と自然科学

池田思朗 いけだ・しろう
情報・システム研究機構 統計数理研究所 教授、
Kavli IPMU 客員上級科学研究員

「科学」という言葉は「対象に関する知識を深め、広めること」を意味しているとしていいだろう。対象が自然現象であれば自然科学、社会現象であれば社会科学となる。Kavli IPMUの研究対象は物理、そして数学であるから、目指す科学は自然科学、そして形式科学だろう。一方、私の属する統計数理研究所の研究対象はデータの処理・解析の方法である。そのため、我々の科学をデータ科学と呼ぶことがある。研究の目的を丁寧に書けば「データの処理方法、解析方法に関する知識を深め、それを広めること」となる。

最近、様々な学問分野、そして産業界で、統計学、機械学習、人工知能といったデータ科学分野に対する期待が高まっている。一方では計測技術の発達やインターネットの拡大によって取得されるデータが爆発的に増え、他方では応用数学分野の発展と計算機の能力の向上によってデータ科学の方法が格段に進歩している。この二つを組み合わせることにより、新たな発見が期待されているのだろう。

データ科学には、応用数学の理論と実際にデータ解析を行う実践との両面がある。私は理論に関しては、微分幾何学を統計学や情報理論に応用する情報幾何学の研究を行ってきた。データ解析としては、この数年間、天文分野との共同研究を中心に研究している。昨年からはKavli IPMUの客員となり、いくつかのプロジェクトに関わろうとしている。どのプロジェクトのデータも興味深いものばかりである。それぞれをデータ科学の観点から説明する。


2016年8月22日に行われたKavli IPMUの外部評価委員会で報告する筆者。

私は二年前よりJST（科学技術振興機構）のプロジェクト（研究代表：吉田直紀）を通じて、すばる望遠鏡のHSCを用いたサーベイ観測に参加している。このサーベイ観測では、天球上の特定の領域を繰り返し観測しており、ある日突然現れる超新星を画像の差分から発見することが課題のひとつである（図1）。差分画像に現れる「引き残し」は一晩で数万にも及び、その中から超新星と思われる数十個の天体を素早く発見するには、機械学習の技術を用いた自動判別器の実装が不可欠である。これまでにNTTコミュニケーション科学基礎研究所と共同でIa型超新星の自動判別器を開発し、観測に用いている。

神岡で行われている実験についても、研究者と議論をし始めている。T2KとXMASSとでは目的は異なるが、どちらも観測されたデータの中から非常にまれなイベントを見逃さずにとらえることが課題である。物理学のモデルを用いてイベントの生成確率を記述し、素早く計算を行って判別につなげる方法を開発しなければならない。今後、貢献できるように努力したい。

将来計画にあがっているLiteBIRDでは、非常に微弱なCMB（宇宙マイクロ波背景放射）の揺らぎをとらえることが目的である。そのためには、はるかに大きい前景放射を分離して取り除かなければならない。前景放射とCMB、それぞれの統計的性質を基に、信号を分離する方法を考えることになるだろう。Planck衛星の信号処理では、90年代に提案された独立成分分析の方法が用いられている。私が研究者になりたての頃に研究した方法が、ここでまた使えるのかもしれないと考えている。

他の分野の研究者と共同研究を行う際、データだけを送ってもらい、我々が処理して送り返すという形ではうまくいかないことが多い。データ科学に対する過度の期待が原因の場合もあれば、我々の理解不足が原因の場合もある。いずれの場合もコミュニケーション不足なのだ。

データ科学者の役割は、データ処理に関するコンサルティングである。そのために、データの背景を知り、データ科学ができることを説明する。議論を重ねながらプロジェクトに参加するのが理想だ。Kavli IPMUとの共同研究はまだ始まったばかりである。今後、様々なプロジェクトと関わりを持っていきたい。

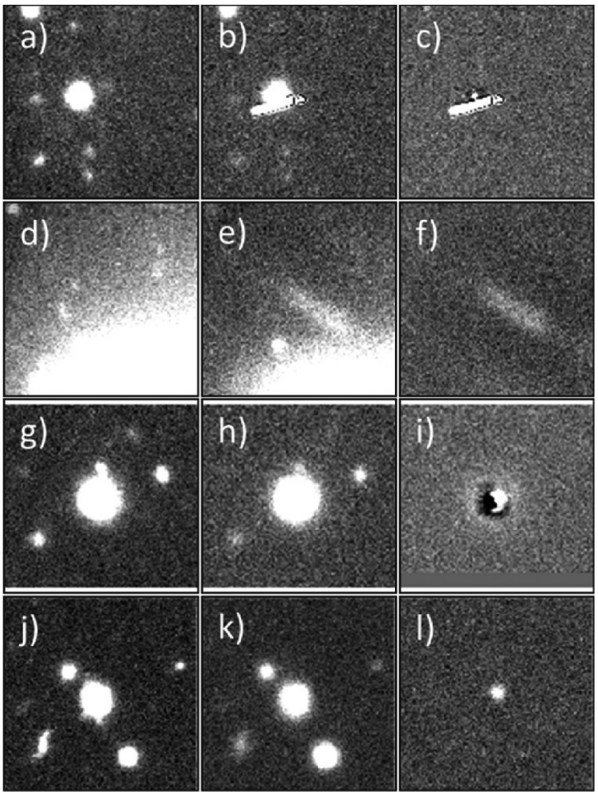

図1 すばる望遠鏡のHSCを用いた観測で得られる変動天体の画像と偽像。各行とも左から右へ、参照画像、新規画像、引き残し画像となる。（a-c）宇宙線が写り込んだ例。（d-f）明るい星の近くの偽像。（g-i）位置ずれによる偽像。（j-l）銀河の中にある変動天体と思われる天体。