データサイエンス実践の統合支援環境 TRAD

慶應義塾大学　柴田 里程
一橋大学国際企業戦略研究科　横内 大介

1　データサイエンスの実践

次々と統計の看板がデータサイエンスに書き換えられていく中で，なにが変わったのだろうか?「データから新たな価値を創造する」というデータサイエンスを世の中に定着させるためには，今や，研究や教育のスタイルの変革ばかりでなく，その実践によって，その世の中に有用性を確実に示していくことが求められている．

しかし，実践の現場が抱えている，最大の問題は「データを取得し，的確に把握し，解析に適した形まで持っていくのに費やす時間の削減」である．実際，作業時間の 3/4 以上がこのような作業に費やされていることも珍しくない．

これは何もコストの問題だけではない．質の問題に大きくかかわる．いわば準備に時間をとられて肝心のデータ解析で新たな価値を生み出す部分がおろそかになりがちだからである．

本報告では，昨年の，仲・横内・柴田による報告で触れた，現在開発中の次世代データ解析環境 TRAD (TextilePlot, R and DandD) が，どのようにして，このような問題を解決しようとしているのか報告する．

2　統合支援環境　TRAD

TRAD はこれまで報告者らが研究開発してきた，TextilePlot, DandD サーバ，DandD クライエントと R をひとつに統合したデータサイエンス実践支援環境である．

高次元空間のデータの雲を２次元平面上で可視化する TextilePlot をヒューマンインタフェースとして採用したことにより，データの様子や解析の結果を画面上で視覚的に把握することができるようになった．

DandD サーバとクライエントを一体化することにより，通信にかかる時間を節約できるようになっただけでなく，別途 DandD サーバを立てる必要はなくなった．また，CSV ファイルなども，ドラッグするだけで DandD インスタンスに変換するので，TextilePlot 上で不足する情報を補足したり，不必要な２次データなどを除去したりの，いわゆるクレンジングと呼ばれる作業を TextilePlot の助けを借りて，視覚的に楽しくおこなえる．その結果を新たな DandD インスタンスとして保存すれば，他の人に解析を引き継ぐことも容易である．

これまで，DandD はさまざまな場所に散らばった異種のデータを取得し活用することや，そのブラウジングに重点を置いて研究開発を進めてきたため，その先の解析とのつながりがどちらかというと不十分であったが， R との連携により，このような問題は解消した．TextilePlot の画面上で眺めているデータテーブルがそのまま，R 上のデータフレームとして検索リストに登録されるので，解析は自由である．必要な解析結果を TextilePlot 画面上に表示することで，結果が高次元空間を構成していてもその制約にはとらわれずに，その様子を探ることができる．

すでに，R をフロントエンドあるいはバックエンドとして利用するシステムも，いくつか発表されているが，TRAD は，基本的なヒューマンインタフェースとして TextilePlot を採用することで，R での解析を視覚的に支援することができるのが，ひとつの特徴である．また，中核に DandD を据えることにより，データフレームだけでは不十分背景情報，たとえばさまざまな属性情報，意味に属する情報，データ取得時の状況などを多角的に捉えながら，R 上での解析を進めることが可能となった．これは，もちろん解析の質の向上につながる．

3　今後の展開

TRAD は誰でも自由に利用できるフリーソフトウエアとして datascience.jp からダウンロードできるようにしてあるので，ユーザの声を反映しての改良はもちろん，

1. 解析プロセスの DandD インスタンスへの記録
2. TRAD 上での複数のインスタンスの併合
3. DandD インスタンスライブラリの充実

など，さまざまな興味深い研究課題が残っているので，今後はその解決に向け研究を重ねていく．