ASA Statement on The Role of Statistics in Data Science
データサイエンスにおける統計学の役割に関するアメリカ統計学会の声明

日本語訳の公開にあたって

平成 28 年 12 月，北海道大学，東京大学，滋賀大学，京都大学，大阪大学，九州大学は数理及びデータサイエンス教育の強化に関する懇談会における評価結果を踏まえ，文部科学省より数理及びデータサイエンスに係る教育強化の拠点校として選定されました．拠点校は，数理・データサイエンスを中心とした全学的・組織的な教育を行うセンターを整備して，各大学内での数理・データサイエンス教育の充実に努めるだけでなく，全国の大学に取組成果の波及を図るため，地域や分野における拠点として他大学の数理・データサイエンス教育の強化に貢献することが期待されています．

アメリカの統計学の学術団体である American Statistical Association は 2015 年 8 月 8 日に ASA Statement on The Role of Statistics in Data Science「データサイエンスにおける統計学の役割に関する声明」http://www.amstat.org/misc/DataScienceStatement.pdf を発表しました．日本でデータサイエンス教育を強化していく上で重要な参考資料ですので，関係者各位と情報を共有したく声明の日本語訳を公開致します．なおアメリカ統計学会には日本語訳の公開の許可を頂いております．

アメリカ統計学会の news magazine である AMSTAT NEWS の公式ブログによると，声明は以下の 7 人の貢献により作成されました．

David van Dyk, Imperial College (chair)，Montse Fuentes, NCSU，Michael I. Jordan, UC Berkeley，Michael Newton, University of Wisconsin，Bonnie K. Ray, Pegged Software，Duncan Temple Lang, UC Davis，Hadley Wickham, RStudio

翻訳　丸山　祐造（東京大学　数理・情報教育研究センター）

「ビッグデータ」,「データ・アナリティクス」を含む「データサイエンス」の台頭は，様々なメディアに取り上げられて近年非常に注目されています．学術界及び産業界で目を見張る成果が挙がっているためです．このような成功は，主として，急成長するこの分野の特徴である革新的かつ起業家的スピリットの賜物です．それでもなお，データサイエンスがさらなるイノベーションを起こすべく潜在能力を最大限に発揮するためには，本質的に学際的な分野であることを考慮すると，分野を横断する共同研究が必須です．データサイエンスの構成要素が何であるかに関して確固としたコンセンサスは依然としてありませんが，コンピューターサイエンス及び統計学にまたがる以下の 3 つのコミュニティが，データサイエンスの基盤として浮上しています．

1. データベース管理：データ資源の変換，集塊，統合化
2. 統計学と機械学習：データから知識への変換
3. 分散コンピューティング・並列コンピューティング：データ解析を実行するための計算機のインフラ

確かに，データサイエンスは多くの学問分野，研究領域と接点を持ちます．実際，このデータ革命に影響されない科学，産業，商業，行政の分野など一つも想像できません．しかし，我々はデータサイエンスに関わる多くの分野の中で，データベース，統計学，分散コンピューティングこそがコアであり，様々な分野を繋ぐ根幹をなすということを強調したいのです．最も基礎のレベルにおいて，我々はデータサイエンスを，これら 3 つの学問分野のコミュニティ間で相互に有益な共同作業が行える場であると捉えています．同時に，関連する多くの学問分野と意義深い交流を持つための場であるとも考えています．データサイエンスがその潜在能力を十分に発揮するためには，これらのコミュニティ内で最大限かつ多面的な協力関係が不可欠です．

とりわけ統計学と機械学習は，データサイエンスにおいて中心的な役割を果たします．研究課題を統計的に定式化することによって，知識発見及びより良い処方箋を得るためのデータ活用が可能となります．データをランダムネス込みで扱うという統計的推測のセントラルドグマにより，研究者は「背後に確率モデルを想定して研究課題を定式化」及び「その考察や処方箋において不確実性の定量的な提示」が可能となります．また統計的な定式化によって，因果と相関を区別することが出来，結果に変化をもたらす要因を特定することが出来ます．さらに予測と推定のための方法論を確立させ，その信頼性を定量化することが出来ます．予測可能性，再現可能性を担保するアルゴリズムに基づいて，これら全てが実行出来ることも統計的な定式化の利点と言えます．このように統計的手法は，他の研究者によっても再現可能な，また異なったデータソースでも再現可能な，科学的な研究成果を挙げることに照準を合わせています．要するに，統計的手法によって研究者は知識を積み上げていけるのです．

統計学者が，データサイエンティストが直面しているチャレンジングな研究課題を助けるためには，データ構造やデータフロー計算，分散コンピューティングの専門家を巻き込んで

持続的かつ確固とした共同作業を行う必要があります．統計学者は，そのような専門家を巻き込んで，また彼らと学び合いながら，一緒に仕事が出来なくてはなりません．研究者個人，研究者グループ，各大学の学科，そしてその分野全体，どのレベルにおいても歯車が咬み合わないといけません，課題解決のための新たな戦略が必要であり，「フルコース」の一貫したパイプラインを開発しなければなりません．「フルコース」というのは，生データのデータ管理から始まって，使いやすく効率的な統計手法の実装・実行，そして意義ある結果を分かりやすく伝えることまでやり遂げることを意味します．統計教育及びトレーニングは，進化を続ける必要があります．次世代の統計専門家は，より広範なスキルセットを必要とし，また，データベースや分散コンピューティングの専門家と協同で仕事が出来なくてはなりません．既存の教育プログラムに加えて新たに設置されつつある革新的な教育プログラムも含めると，データサイエンス分野全体の学生定員は増えていますが，今後予想されている需要を考えるとさらに増やさなければなりません．次世代においては，伝統的な学問分野である統計学，データベース，分散コンピューティングの垣根を超えて活躍する研究者が数多く出てくることが期待されています．今後は，そのようなマルチリンガルな専門家に対して圧倒的な需要があるのです．

統計家，統計学科，及びその他学会などと連携して，アメリカ統計学会は，

- データサイエンスにおける統計学の役割に関する議論の場の提供
- また日進月歩のデータサイエンス分野における適切な道案内
- さらにデータサイエンティスト（その中には統計家もそうでない方も含むわけですが）のコミュニティの中で情報共有したり，共同研究するためのフォーラムの提供

のようなサポートをする上で適切な立場にあります．アメリカ統計学会は，統計家とデータサイエンティストの共同作業を促進，またそれによって共同作業によらず独自で行うよりもより良い成果を出してもらいたいと考えています．