データ活用型社会創成プラットフォーム計画について

情報基盤センター長　田浦健次朗

1　概要

データ活用型社会創成プラットフォーム(以下データプラットフォーム)計画は、データ中心的な研究分野や、データ解析・データ科学的手法への期待が高い分野のための情報基盤を提供するとともに、研究分野間や産学間の連携を促進することを目指す計画である。

東京大学は、国立情報学研究所との綿密な連携の元、2018 年初頭から構想を始め、本学内の多くの部局、全国の大学との連携・協力関係を築きつつある。本学では、本部直轄の、「未来社会協創推進本部」の下に、「データプラットフォーム推進タスクフォース(座長:相原博昭 副学長)」が設置され、全学的な推進体制が敷かれている。

その中にあって情報基盤センターは、情報基盤の設計で中心的役割を果たすと共に、学内の情報系部局と、人文社会を含む非情報系の多くの部局との連携・協力の推進、全国の大学や研究所と連携した全国的な運営、研究コミュニティ作りなどの役割を果たしていく。

2　背景: データ中心科学やデータ活用を取り巻く状況

様々な研究分野で、データが研究における重要な資産となっている。

背景には、(1)データが大量に取得・利用可能になったこと、(2)大量のデータからの高次な情報抽出が、特に機械学習技術の進展により汎用的かつ高精度に行えるようになったこと、の相乗効果がある。さらに、(3)研究成果が少数の法則の発見ではなく大量のデータそのものである、という分野、(4)従来の物理法則の忠実なシミュレーション(演繹的な計算)と、大量のデータから計算結果を推論する手法の相乗効果、(5)オープンサイエンスという考え方の元、論文だけでなくデータそのものも成果として共有する考え方の浸透、などが一体となって現在の潮流が作られている。

ウェブページやソーシャルネットワークなどのテキスト・画像データが容易に利用可能になったのは二昔ほど前のことである。また、地球科学分野などで全国的、地球的規模でデータを蓄積することは古くから行われている。それらに加え、かつてはユビキタス、現在は IoT と呼ばれている、センサネットワークからの観測データが用途に応じて取得可能になったこと、主に行政データなどの公のデータが公開されるようになったことなどから、さらに様々なデータが利用可能になっている。そして、深層学習を代表としたデータ解析技術との相乗効果で、データの価値が二重に高まっている。二重という意味は、機械学習によってデータから高水準な、解釈可能な情報を得るのが容易になったということ、データが機械学習の訓練データとして使われる

ため、質の高い大量のデータが機械学習の結果を高める、ということである。また、ヒトゲノムプロジェクト、ヒューマン・ブレイン・プロジェクト、ヒト細胞アトラスなどに代表されるように、複雑な系を対象にして、少数の法則に還元されることを目指すのではなく(もとよりそれが期待できないので)、結果をデータとして蓄積・アウトプットすること自身が目標となる分野が生まれている。さらに、従来物理法則を忠実に、高性能に再現することで、より解像度が高く、より長期に渡るシミュレーションを可能にしてきた分野(分子シミュレーションや気象シミュレーションなど)で、シミュレーション結果を学習データとして計算結果を予測する手法が発展し、従来の計算科学的手法とデータ科学的手法の相乗効果が起きている。

3 計画の骨子

データ活用型社会創成プラットフォーム計画は、これらの潮流に基づき、広くは大学・研究機関、その中で情報系研究者、さらにその中で情報基盤センターのような機関が何をすべきかを考えて生まれた。本計画は以下を具体的なアクションとしている。

1 情報基盤の設計と提供
2 全国的研究コミュニティの創出
3 将来像としての、大学間が有機的に連携した産・地域との連携

3.1 情報基盤の設計と提供　大学・研究機関が連携して新しい情報基盤を作る意義

まずは情報基盤の設計と提供について述べる。情報基盤センターはこれまで、全国共同利用設備としてスーパーコンピュータの導入、運用を長年に渡って行っている。また、7 大学(北大、東北大、東工大、名古屋大、京大、阪大、九大)と共に、共同利用・共同研究拠点(学際大規模情報基盤共同利用・共同研究拠点:JHPCN)として活動し、全国の研究者(一部海外を含む)と共同研究を実施している。その共同研究実施のために、大学のスーパーコンピュータを始めとした設備を全国の研究コミュニティに提供している。

データプラットフォーム計画でも、主旨としては同様の、全国共同利用される情報基盤(固有名称は未定だが本稿ではデータプラットフォームと呼ぶ)を提供する。2020 年度内に稼働する予定である。データプラットフォーム稼働前に、同主旨のパイロットプログラムを開始することも計画している。

設備としてのデータプラットフォームは以下を特徴とする。

3.1.1 セキュア、大容量ストレージと高性能な計算基盤の提供

データ活用型研究を進めるに当たって、しばしば障壁となるのがデータのセキュリティである。セキュアなストレージと言っても広範な意味がある。外部からの侵入などに対して高いセキュリティレベルの運用が出来ているか、そのための責任の所在が明確になっているか、という

運用体制の問題; 個人情報を含むデータの取り扱い、特に民間クラウドを利用した際にデータが国境を越えることに対する懸念; ユーザ間・グループ間のストレージ・通信の隔離がOS、ファイルシステムレベルではなくより下位レイヤ(仮想マシン・コンテナなど)で行われているかという、データ隔離のレベル; データの暗号化がどのように行われているか、どのレイヤで行われているか; ストレージの物理的なセキュリティ(サーバ室への入退室管理やケージなど)が確保されているか; など多数の軸が存在する。データプラットフォームでは従来のスーパーコンピュータ運用時同様のしっかりとした体制に加え、OS・ファイルシステムよりも下のレイヤでストレージやネットワークの隔離を行うことで、プロジェクトごとに個人情報や機密性の高いデータを格納する際の懸念を払拭(軽減)できるようにする。

3.1.2 広域データ収集ネットワークと連携した情報基盤の提供

現在、国立情報学研究所(NII)で各都道府県を 100Gbps 以上の帯域で接続した学術ネットワーク SINET 5 に加え、SINET 5 に接続したモバイル網を「SINET 広域データ収集基盤」として提供している(https://www.sinet.ad.jp/wadci)。これは、各プロジェクトごとに、モバイル仮想閉域網(Mobile Virtual Private Network; Mobile VPN)を提供し、その VPN をクラウドなど計算基盤まで延伸することを可能にしたものである。データプラットフォームもこの広域データ収集ネットワークと連携し、日本全国からのデータの収集から蓄積・処理までをプロジェクトごとに閉じたセキュアな環境で行うことを可能にする。

3.1.3 「プラットフォームのためのプラットフォーム」の提供

データプラットフォームはこれまでのスーパーコンピュータとは大きく異なる分野、異なる利用形態での利用が想定される。スーパーコンピュータは主に大規模な並列計算を高性能に行うためのものである。もちろんスーパーコンピュータはデータ処理性能も優れており、大規模データ処理や深層学習にも適している。そのレベルで計算機の構成を抜本的に変える必然性はないのだが、これまでの運用方法ではサポートできない利用形態も存在する。一つの形態は「プラットフォームのプラットフォーム」、メタプラットフォームとでも言うべき利用形態である。それは、データプラットフォームの「ユーザ」とは、単にデータ処理をするための計算資源としてプラットフォームを使うにとどまらず、分野のデータレポジトリを整備しそれを分野研究者に公開する、「プラットフォーム構築」をするユーザであり得るということである。このようなユーザをサポートするには、これまでのスーパーコンピュータの設計・運用とは異なる要素を取り入れる必要がある。外部ネットワークへの接続、恒久的な資源の割当て、分野プラットフォーム構築のために柔軟にシステムソフトウェアを構成できることが必要で、それと合わせて大規模機械学習処理やデータ同化シミュレーションなどのために、高性能な計算資源と連携できる必要がある。また、分野データプラットフォームとしての利用のためにはこれまでの単年度ごとの申請よりも永続性を持った利用形態にする必要があるし、そもそも何に対して負担金を課すのかという点も検討の余地がある(貴重なデータを共有するインセンティブ

を作るためには、使用ストレージ容量での課金に再考の余地がある)。データプラットフォームは、VPN 構築や仮想化・コンテナなどの技術を組み合わせて、ユーザにカスタマイズ可能なプラットフォームを提供しつつ、運用面の検討を行ってこのような利用モデルをサポートできるようにする。

3.2 全国的研究コミュニティの創出

データプラットフォームを構築し、研究者に提供する目的は、それを通じて分野を介した共同研究や産学連携を促進することにある。前述したとおり情報基盤センターは JHPCN という全国共同利用・共同研究拠点を全国 7 大学の基盤センター等と共に運営している。共同研究の対象分野は多岐に渡っているがやはり大規模なスーパーコンピュータを用いた計算科学やその萌芽的な研究が中心である。その意義は、単に計算機を提供しているということではなく、計算科学の様々な分野の研究者が、情報学・高性能計算の研究者と、または異なる分野間で交流できることにある。シミュレーションの対象が違っても基づく物理・支配方程式が共通であったり(例: 流体)、支配方程式が違っても共通の計算手法(例: 格子法や粒子法)を使うことはしばしば生ずる上、計算手法を高性能に実装する際の知見も分野間で共通部分が大きい。スーパーコンピュータと高性能な計算手法、高性能システムの専門家を配した拠点には極めて大きな意義がある。

データ中心的な分野、データ活用が期待される分野についてもそのような研究コミュニティの創出が、目指すべきものである。その研究コミュニティのあり方は、精神はこれまで JHPCN が育んできたものと共通でありながら、カバーする範囲や分野間連携の生まれ方が、より広範で、多種多様なものになると期待する。これまで、計算手法や高性能計算のための手法を中心に生まれていた交流が、生データ、curate されたデータ、データ分析手法や機械学習で得られたモデルの共有などを通じた、より幅広いものになることを期待している。もちろんこれまでどおりのシステム・高性能計算の手法を軸とした交流も行われるだろう。特に広域ネットワーク、大規模データ処理や高性能ストレージに関する課題などが分野横断的に議論され、発展することが期待できる。それらの分野を支えるのに必要な情報学の専門性も、これまで情報基盤センターがカバーしてきた、主には計算科学分野とシステムソフトウェア分野だけではカバーしきれない広さが必要になる。したがってこれまでのような、情報基盤センターの専門家と情報系以外の分野の専門家の学際的交流というにとどまらず、機械学習、AI、データ工学分野はもちろんのこと、情報学と他分野の境界的領域の専門家など、情報系の広い分野と、情報系以外の分野専門家との交流が必要になる。

このような交流がどのようにしたら生まれ、持続的に発展するか? 色々な困難が山積しており、試行錯誤が必要であろう。対象となり得る分野が多岐にわたりすぎている上、一部の分野はそれ自体が重厚なディシプリンで、情報や計算機システムの専門性というだけで有益な協力が困難な、いわゆる(情報系の研究者から見たときの)敷居が高い分野も多く存在する。しかしながら昨今のデータ科学や AI への期待を引き合いに出すまでもなく、情報科学的手法と他

分野との協働は多くの分野から強く求められているところであり、また情報を専門とする研究者もそれを求めている。特に、情報技術やデータ科学的手法の重要性を認識しつつも、実践に移すに当たって必要な情報の専門家が近くにいない、という声も、これまでデータプラットフォーム構想について説明している中で多々聞かれるところである。様々な困難を克服した際に得られるものは大きく、また、JHPCN という共同利用・共同研究拠点が培ってきた大学間のネットワークやその運営に関する経験を生かす余地が大いにあると感じている。

3.3 将来像: 大学間が有機的に連携した産・地域との連携体制

本年報の「巻頭言」でも紹介した、豊田長康「科学立国の危機」という本には、研究成果の主たる形である論文の出版数と、GDP との相関が強いことが示されている。大学の研究はそもそも直接的に GDP への貢献を目指して行わているわけではないものも多く存在するが、それでも大学が行う知の追究・人材育成の結果が GDP へプラスに働くことは、研究に携わっている人であれば当然だと感じることであろう。今後大学にはますます、地域創生や産業への貢献、大学外の人も対象とした人材育成などを意識した活動が求められる中、データ科学や AI の専門家を多く擁する大学が果たす役割は大きい。一方で大学は財源の多様化という言葉で、運営費やポストの削減圧力に常にさらされている。もちろんその圧力は本学よりも、より規模の小さい大学において更に強く感じられている。

上述したようなデータ科学的手法、データそのもの、データ処理基盤などを通じて結びついた研究者コミュニティが、大学間で連携して、産業界・地域へ大きく開かれた窓口となることが、本プラットフォームの大きな目標である。