第 2 回 SPARC Japan セミナー2015
「科学的研究プロセスと研究環境の新たなパラダイムに向けて
－e-サイエンス，研究データ共有，そして研究データ基盤 -」

オープンなプラットフォームが研究に与える影響を帰納的に考える

池田　大輔
（九州大学システム情報科学研究院）

講演要旨

学術情報流通の分野で、第 4 の科学、オープンサイエンス、データ中心科学等の言葉がよく聞かれるようになったが、その実体はよく分からないという方も多いだろう。発表者は、情報学の研究者として、様々な分野のデータを扱う一方で、データベースや情報検索等のデータや情報の基盤に関する研究も行っている。これらの経験をもとに、本発表では、データやソフトウェア、インターネットなどのオープンなプラットフォームが研究に与える影響を、具体的な例から帰納的に考える。考察の出発点は、上述の全てのキーワードが内包とすると思われる「科学」と「情報処理技術」であり、例えば、情報処理技術は使うが科学ではないものとの違いなどから、これらのキーワードが表す曖昧な概念を明確にしたい。

池田　大輔

九州大学理学部物理学科卒業後、情報系の大学院へ進学。同大の大型計算機センター、附属図書館勤務の後、九州大学大学院システム情報科学研究院（准教授）。マイニング等のデータの利活用及び機関リポジトリ等のデータ基盤に関する研究に従事。博士（理学）

今日の話は、一言で言うと、e-サイエンスとは何かということについてです。私は学部のころは理学部の物理学科で学んでいました。その後、情報系に移ってずっと情報系で働いています。普通の研究者とは少し違うキャリアパスで、最初はコンピューターセンターでしばらく働いていました。コンピューターセンターでは、ハイパフォーマンスコンピューティング、ネットワーク、セキュリティーなどに関わってきました。その後、九州大学の附属図書館で働きました。ここでスカラリーコミュニケーション、学術情報流通のインフラ、機関リポジトリ、自動認識（RFID やバーコード）などの研究に関わってきました。その後、今の九州大学システム情報科学研究院に所属し、現在、バイオインフォマティクス、e-サイエンス、情報検索、データベースなどを研究しています。このようなハイパフォーマンスコンピューティングは e-サイエンスに非常に近く、関係あるところです。

研究内容は主に二つで、一つはデータ解析です。帰納的なアプローチで、データから共通のパターン、知識などを見つけようとしています。もう一つはデータのインフラに関する研究です。情報検索やリポジトリなどの研究をしています。

サイエンスを支える四つの柱

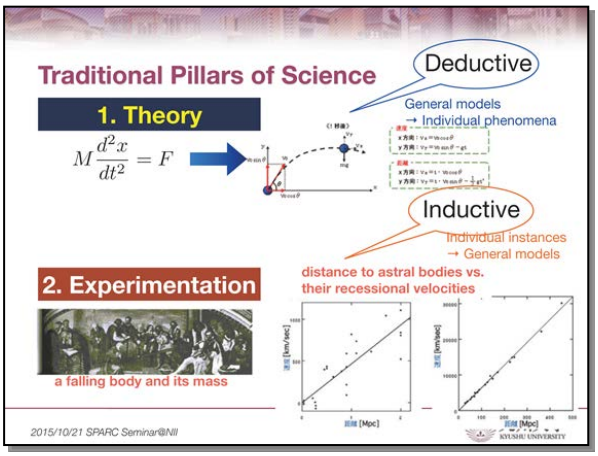
伝統的なサイエンスを支えてきた二つの柱は、理論（theory）と実験（experimentation）です（図 1）。Theory は演繹的（deductive）で、一般的なモデルを仮定することで未来を予測します。例えば、物体の放物線運動では、微分方程式というモデルによって、初期条件を与えたら何秒後にどこの位置にあるということが予測できます。Theory は古代ギリシャのころからされていて、例えば宇宙はアトムでできているといったことを考えていたのですが、当時は実証する人はいませんでした。初めてガリレオのあたりから実験をして確かめるということが出てきて、これが科学を支えるもう一つの大きな柱になりました。こちらは個別の事象から一般的な仮説やモデルを立てようとする帰納的（inductive）なものです。下のグラフは、天体の距離と遠ざかる速度のグラフで、宇宙が膨張しているという仮説をハッブルが得たときに使われたグラフです。個別に観測された点から直線を引いたところが帰納的ということになります。

これが従来の二つの柱でしたが、新しい柱が出てきました（図 2）。一つはコンピューターシミュレーションです。こちらは大きく言うと演繹的な方で、何かモデルを仮定して、一般には時間発展を追って、その現象を再現します。そして、その現象を実際のデータなどと突き合わせて検証します。これは現実には実験がしにくい、原子爆弾、山火事、地震、経済活動といった現象を再現するのに役に立っています。非常に大きいインパクトを与えていて、もう数十年前から科学を支える柱の一つになっています。

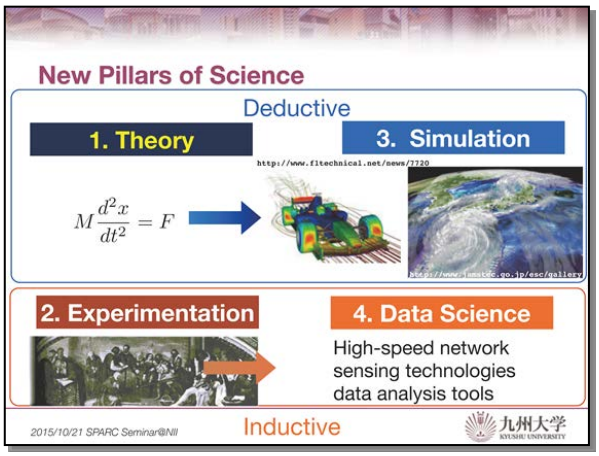
もう一つ、最近いわれているのはデータサイエンスです。帰納的な、データから何か考えるというアプローチです。この背景には、ネットワークが高速になったことや、センサーのテクノロジーが発達したこと、情報系のデータを解析するソフトウエアやツールが充実してきたことがあろうかと思います。

この新しい第 3 と第 4 の柱の例を幾つか考えてみました。第 3 の柱のシミュレーションの例としては、台風、原子爆弾、森林火災、株式市場のコンピューターシミュレーションなどが挙げられます。実験しにくい現象を再現できるので、科学に与えたインパクトは非常に大きいと思います。一方で、研究のスタイルは、従来のクローズドな、研究室単位の研究スタイルからあまり変わっていないのではないかという気がしました。隣の部屋にある大型計算機を使うのか、隣の建物にあるスパコンを使うのか、クラウド上にあるコンピューターを使うのかという違いだけで、研究者自身はあまり変わっていないのではないかというのが、第 3 の柱に対する私の印象です。

第 4 の柱のデータサイエンスの例としては、データから著しい進歩を遂げたような分野を幾つか挙げます。例えばコンピューター将棋は、トッププロを負かすようになっています。これは科学かと言われるとよく分かりませんが、機械学習などの成果が使われています。自動運転もそうです。これも画像認識やパターン認識



（図 1）



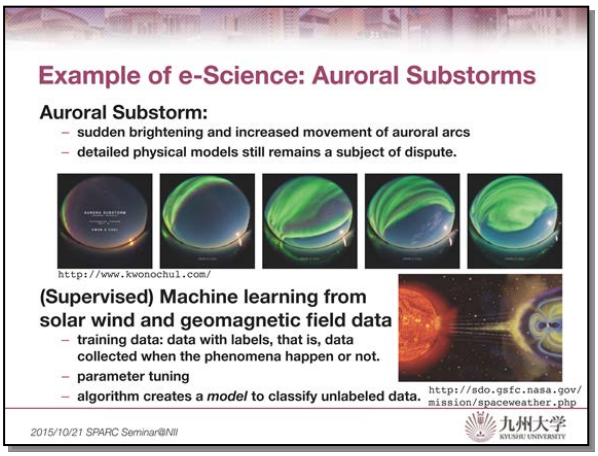
（図 2）

の技術が使われており、人間と同様のことができるようになっています。しかし、これも科学かと言われると分かりません。Google などが、検索履歴からインフルエンザ流行を予測していますが、これもどうなのでしょう。例から幾つか共通するものを抜き出すと、高精度の認識や予測はできる一方で、科学的なメカニズムには踏み込んでいません。例えば、将棋に対してどのような思考が行われるかということは考えず、パターンを学習して、機械がそのまま指すことになっているので、科学ではないのではないかとも思いました。

オープンなプラットフォームが与える影響

e-Science の例として、オーロラサブストーム（オーロラ爆発）という現象を挙げます。これはオーロラが急激に光って、一気に広がる現象で、細かい物理的なメカニズム、モデルはまだきちんと解明されていません。図 3 には約 5 分間の全天画像を幾つか出していますが、実際には数時間にわたるような現象です。

今回私は、機械学習の手法、太陽風と地磁気のデータを使い、いつこの現象が起こったかを手作業で特定しました。トレーニングデータという、現象が起こったときのデータ、現象が起こっていないときのデータをひも付けて、パラメーターをチューニングして、アルゴリズムに与えます。すると、アルゴリズムが勝手にモデルをつくってくれます。これは物理的なモデルというよりも学習のモデルで、その後は未知のデータが与えられたら、これはこのような現象が起こるということを返してくれます。2015 年 8〜9 月にミュンヘンで開かれた IEEE の e-サイエンスに関する国際会議で私がポスター発表してきたときの結果では、79%ほどの精度で予測が当たっていました。



（図 3）

この研究プロセスを見ていきます。まずはドメインエキスパートから、どういうデータがあり、それがどういう性質なのかということを聞きます。2 番目に、実際にデータを探します。今回使ったのは、国立極地研究所の all-sky images、NASA の太陽風のデータ、ワールドデータセンターの地磁気のデータでした。3 番目に、トレーニングデータをつくります。学生が一生懸命オーロラの写真をたくさん見て、ここで起こったということを特定します。4 番目に、ここまでできると、あとはかなり機械がやってくれて、実際に機械学習のアルゴリズムにデータを与えます。ここでは LibSVM と呼ばれる有名な機械学習のライブラリーを使っています。最後に、結果が出たら、これはどういう意味かということをドメインエキスパートと話して評価してもらいます。

データと、ソフトウエアも一種オープンプラットフォームと言っていいのではないかと思うのですが、このおかげで今回の成果が得られました。データを取った人、公開している人、LibSVM をつくった人とは直接面識はないけれども、巨人の肩の上に乗って新しい成果を生み出したということになります。これはオープンなコラボレーションと言っていいと思います。

最近、科学系のソフトウエアが非常に充実してきました。少し前のビジネスの世界だと、LAMP（Linux、Apache、MySQL、PHP）によって、ベンチャー企業などが安価にいろいろなシステムをつくれるようになったという大きなインパクトがありましたが、これを使いこなすにはプログラミングスキルが要ります。

一方、最近では、R を含めたプログラミング言語や、科学に特化したライブラリー（線型代数、フーリエ変換、可視化技術、機械学習、データマイニング、統計、自然言語処理）など、他にもいろいろありますが、これらが充実してきました。少し前だと考えられなかっ

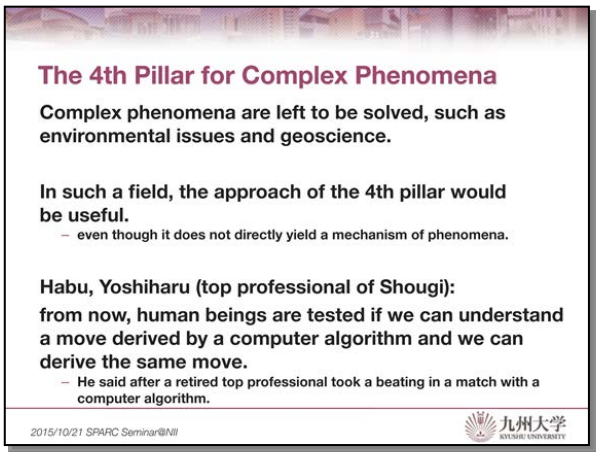
たことに、かなり最新の機械学習などの成果もすぐ実装されて、比較的簡単に使えるようになっています。これは非常に大きなインパクトではないかと思います。最先端のデータアナリシスツールが簡単に使えるようになってきました。

こういうものがあると、粒度の細かい学術情報流通が可能になってきます。例えば、Data Journals というものがあり、これは日本生態学会が発行している英語のジャーナル、「Ecological Research」の中には、論文のタイプとして Data papers というものがあります。これは比較的最近です（図 4）。また、日本情報処理学会のバイオインフォマティクスに関するジャーナルも、オリジナルペーパー、サーベイペーパーに加えて、データベース・ソフトウエアペーパーというものがあります。つまり、何かシステムやソフトウエアを組んだことで論文になるということです。今までは、これらは一つの論文を書くために開発していただけなので、論文をパブリッシュしないと成果にならなかったのですが、これ単体で成果になるという世の中になってきました。こういうものがあると、データを使って、あるいはソフトウエアやデータベースを使って、他の分野の人、つまり知らない人との共同、あるいは時間を越えた共同が簡単になってくると思います。

もう一つ、複雑な現象に対して、第 4 の柱は有効ではないかと思います。いろいろ簡単な現象もだいぶ科学的に解明されてきましたが、複雑な事象、例えば環境問題が残されています。このような分野では、第 4 の柱のデータサイエンスが有望かもしれません。つまり、メカニズムには迫れないのだけれども、出た結果を持って専門家と一緒にそれはどういう意味か考えるということです。図 5 に書いてあるのは羽生善治さんという将棋のトッププロの言葉ですが、有名な元トッププロが負けた後に、「これからは機械が指した一手を人間（プロ）が評価するような時代になるのではないか」とおっしゃっていました。ですから、複雑な現象には第 4 の柱は非常に有効です。科学かどうかは分かりませんが、科学を後押しする、支援することは可能ではないかと考えています。そのためには、エキスパートやデータサイエンティストとの共同が必要だと思います。

データに関するイメージがまだ十分みんなで共有されていないと思います。機関リポジトリのペーパーに対して、分野ごとに非常に違ったイメージがあるので、これをつくるのが大事かと思います。このときに二つのアプローチがあり、データベースのようにきっちり先に抽象化してやる方法と、情報検索のように、あまり抽象化はせずに、Google の検索のように、検索ができればいいというアプローチがあると思います。主に前者をやっているので苦労しているのではないかと考えています。最後に PR ですが、今われわれは科研費をもらって、後者のアプローチをしているので、いつかそのような成果を発表できたらと思っています。

（図 4）



（図 5）