資料2
科学技術・学術審議会
戦略的基礎研究部会
数学イノベーション委員会
（第23回）H27.10.30

ビッグデータ時代における情報科学と数学との連携に必要な人材について

田中　譲
北海道大学大学院情報科学研究科
特任教授

ビッグデータとは

- 3V (Volume, Variety, Velocity), 4V (+Veracity), 5V (+Value)
- ミッション駆動型研究からデータ駆動型研究へのパラダイムシフトを象徴する旗印としてキーワード
  - ミッション駆動型：課題→仮説立案→実験（データ収集・解析）→仮説検証
  - データ駆動型：大規模データ収集⇒（（課題）→仮説立案→データ検索・解析→仮説検証）

パラダイムシフトとしてのビッグデータ

- ミッション駆動型研究からデータ駆動型研究へ
- 要因
  - ウェブコンテンツの急増
  - サーチエンジン・サービスの台頭と利用者ログの獲得・蓄積
    → 意図のデータベース
  - モバイル情報の獲得・蓄積（スマートフォン、プローブ・カー・システム）
  - 新世代DNAシーケンサの普及
  - IoT（物のネットワーク）の急速な進歩普及
    - 社会において
      - 多様なセンサーとアクチュエータのネットワーク
    - 先端科学技術研究環境において
      - 計測・観測・分析機器のディジタル化とネットワーク結合
  - IBM Watson

ビッグデータ応用の促進力
①管理検索分析処理技術

- スケーラビリティ
  - Cloudを用いたHadoopなどの大規模データ基盤技術の発展
- 大規模データ検索処理
  - 列志向DBMSやNoSQLなどの新しいDBMS技術の進展
- 分析・可視化
  - 分析・マニング技術と可視化技術の急速な発展
  - 大規模機械学習技術の顕著な発展
- 大規模シミュレーション
  - 京コンピュータに代表される超高速計算技術と大規模シミュレーション技術、観測データとのデータ同化技術の発展

ビッグデータ応用の促進力
②機関によるデータの公開

- オープン化
  - 行政、公的機関、企業などのサイロの中に眠っていたビッグデータの積極的活用を目指したオープン化の動き
- 個人情報の保護と有効活用
  - 個人情報の取り扱いに関する法整備と解釈の基準化
    - 2014.5　**PCAST Report: Big Data and Privacy**

挑戦的ビッグデータ応用（1）

- 安心・安全でサステイナブルな都市基盤サービス
  - 交通／エネルギー／上下水道／ゴミ収集　etc.
  - Social cyber-physical system（SCPS)
    - IBM: Smarter City / Microsoft: Urban Computing
    - 気象や交通、エネルギー消費などの過去データから規則的パターンを見つけ、実時間データに適用して予測を行い、効率的なサービス提供を目指す。
- レジリアントな社会
  - 基盤構造物の維持管理（SIP）
  - 基盤構造物にセンサ群を設置、集中管理によって劣化や異常の検出
- 防災・減災・災害対策・復興支援
  - 地震／洪水／火災／テロ／…
  - 予測／予報／誘導／緊急対応支援／復興支援
  - 日常運用システムの機能拡張としての災害時緊急対応システム
  - 犠牲者・ボランティアを含む市民との連携支援機能
  - 臨機応変な対応、想定外の事態への対応が必要

挑戦的ビッグデータ応用(2)

- ヘルスケア(特にバイオメディカル応用)
  - コホート研究
    - 個々人の病歴データの統合／病院や国の壁を超えたデータの統合管理分析
    - 個々人の遺伝子データ
  - 臨床試験
    - オーダメイド医療を目指した臨床治験・臨床試験データの分析
  - 創薬へのビッグデータ・アプローチ
  - 公開文献と公開データからの新治療法や新薬の発見(トムソン・ロイターのバイオメディカル部門
- 感染症流行予測
  - 流行ウィルス株の予測
  - 流行の地理的拡散予測

挑戦的ビッグデータ応用（3）

- 農業の効率化・高収益化（e-agriculture)　（→　漁業・林業）
  - 生産・収穫支援／流通販売支援
- サイバー・セキュリティ
- 設計ビッグデータ
  - 基本手法
    - 膨大な設計事例の蓄積／部品・合成部品ライブラリの拡充
      - 部品ライブラリ中の部品のあらゆる組み合わせとその構造を予め網羅的に作成し、ライブラリを大規模化
    - 構造と性能・機能の関係を学習し、特定性能や機能の検索を可能にする
      - 合成部品の性能や機能を既知の部品組み合わせにおける構造と性能・機能との関連を用いて得られた学習結果に基づき推定
      - 必要な性能・機能を持った合成部品をライブラリから検索
  - 大規模システム
    - 航空機設計／超並列マルチコア・システムの設計
  - 機能材料設計
    - 特定機能を持つ分子構造やメゾスコピック構造の設計（有機／無機）
  - 創薬

挑戦的ビッグデータ応用(4)

- 科学技術基盤システム
  - e-science / data intensive science / data centric science
  - 個別科学におけるビッグデータ応用

    宇宙物理学／地球環境学／高エネルギー物理学、核物理学／分子生物学、進化生物学／薬学／感染症疫学／材料物性学／…
  - 科学技術研究開発共通基盤システム技術
    - 多様なデータに基づく仮説設定／仮説検定の繰り返し過程の支援　←　分析法を見つけること自体が研究対象
    - 大規模な文献情報を推論可能な知識表現に直し、有益な知識を推論により発見(IBM　Watsonの機能を拡張発展)
      - トムソン・ロイターは公表された文献とデータのみを用いる知識発見を軸に、バイオメディアカル事業を立ち上げ
      - IBM は、Watsonのバイオメディカル応用を重視

挑戦的ビッグデータ応用の特徴

- Volumeに重点を置いた単一種類大規模データの分析
- **対象系はモデリング可能(と仮定)**

- Varietyに重点を置いた、多種類データを関連付けた分析による価値創生
- **対象系は複雑系(System of systems)**:個々の要素系は異なるモデルに従う → **単一モデルによる全系のモデリングは不可**
  - 複数異種データを関連付けた分析
  - 異なるパターンや規則性に従う異種混合データの分析
  - 特定分析毎に目的に応じた適切なデータセグメンテーションが必要
  - 分析シナリオが定石として確立していない分野への応用
    - 先端科学技術分野や、戦略的意思決定を必要とする分野
    - 試行錯誤的・即興的分析過程の支援が必要

分析対象が複雑化

- 単一モデルでモデリング可能な対象
  - POS データ, カード利用ログ
  - ウェブ情報, Google 検索ログ
  - etc.
- 異種混合系と考えられる複雑系が対象(個々の要素系は単一モデルでモデリング可能)
  - 個人化医療
  - 都市規模の社会サービスの最適化
  - etc.

分析シナリオを見つけること自体も研究対象

典型的アプローチ

- 数学モデリング可
  - ⇒ 異なる多数のパラメータ値で**シミュレーション**
  - ＋ **実データとデータ同化によりシミュレーション結果を選択**
  - ＝ **予測（動的システム）**（例：気象予測）
    - or
    - **パラメータ値同定（静的システム）**(例：航空機設計）
  - **適切なモデリングが最も重要**
- 数学モデリング不可
  - ⇒ 個々の対象を特徴づける**特徴量集合を定義**
    - e.g., Catalyst Genome
  - ⇒ 個々のオブジェクトを**特徴量の多次元ベクトル表現**
  - ⇒ **統計解析、クラスタリング、マイニング、機械学習**等々が可

  適切な特徴量の定義が最も重要
  特徴量の創出に数学的モデリングによる対象現象の解釈が必要

応用と基盤技術の間の溝

- 実応用から本質的な課題を明確に抽出し、数学モデリング
  - → 対話の場が限定されている
  - → 数学モデリングが行える人材の不足
  - → 専門用語の壁
- 実応用に即し、種々の最先端の手法の中から最適なものを選び最適な分析シナリオを構築して適用
  - → データサイエンティストの不足
  - → 各種分析手法の意味理解が必須
  - → 課題に即した分析シナリオ構築能力が必須

ビッグデータ時代における情報科学と数学との連携

- 知識発見のための数学
  数学や数理科学の研究者が理論面で活躍
  - 統計数学、組み合わせ離散数学、グラフ理論
    - グラフ理論に基づく大規模グラフマイニングの研究（NII　河原林健一）
    - 非負行列分解や非負テンソル分解による独立成分への分解とトレンド分析
    - ポリトープを用いた近似充足可能性問題の定時間処理アルゴリズム（パリ2大学 Michel de Rougemont)
    - 情報理論における情報スペクトル的方法（電気通信大学：　韓太舜（数理工学出身））
    - MDL原理(Jorma Rissanen, 1978), Information Bottleneck (Naftali Tishby, 1999)
    - etc.
- 対象の数理モデリングのための数学
  - 有用な指標の創出
    - 例：　*The numbers behind numbers: Solving crime with mathematics*
  - 対象の本質を把握するためのモデリング
    - 例：　Stuart kaufmanのrandom binary network (代謝ネットワークのような触媒反応ネットワークの本質的な挙動の説明）
    - 例：　MathematicaのStephen WolframのNew Kind of Science

必要な人材の資質

- データサイエンスの基盤となる数学、数理科学の素養
  - 従来： 統計数学、組み合わせ離散数学、グラフ理論、線形代数　等
  - 今後： ＋？
- 対象システムの数理モデリング能力
  - 対象のマクロなモデリングや、解明されている個々のミクロな機構のモデリングではなく、複雑なシステムをメゾスコピックなレベルでモデリングする能力　（機構は不明で、現象のみが観察可能）
  - 対象ごとにモデリングの基礎となる数学理論は異なる　→　広範な数学の知識と、それらを用いたモデリング・スキルが必要
- 解析数学の素養
  - シミュレーションモデルの構築と解析の能力
  - 離散数学に基づくアルゴリズムだけでなく、解析数学に基づくアルゴリズム(例： SVM)

必要な人材の教育
(欧米との差に注目して)

- KTHのコンピュータ・サイエンスのカリキュラム例
  - 1st yeart
    - Linear algebra 7.5 credits / Analysis 7.5 credits / Discrete mathematics 7.5 credits / Numerical approximations 6 credits
    - 28.5 credits of math (out of 40 total credits)
  - 2nd yeart
    - Logic, 6 credits / Probability theory and statistics, 6 credits / analysis course of 7.5 credits that you can replace with a course in modeling and simulation, 6 credits if you want.
    - 18 or 19.6 credits of math (out of 40)
- 座学（50～60%）＋演習（50～40%）
- 集中講義形式（フランスの科学系のグランゼコールに関しても同様の（週7時間以上の）集中的講義の話を聞いたことがある）
- 多様な分野への応用に関して最新の研究レベルに近い課題を設定して、（ホームワークを含む）演習形式の応用力養成がなされているケースもある。
- Concept並びにProof of Conceptを重視

Concept及びProof of Concept に関する国内教育の現状

- 理論や手法は学習するが、それらが提案された背景や理由など、応用にとって重要な意味や解釈に関する教育は充分になされていない。
- 先端研究のレベルに近い実問題への適用力の養成は殆どなされていない。
  - 研究室配属後に特定研究課題に関していきなり応用が始まる。卒業研究レベルでは、どの理論を適用すべきかを自身で考えることは少ない。
- 重要な語をその概念を理解することなく使用してしまう
  - 哲学用語「知の地平」(horizon of Knowledge)のhorizonの意味は？
  - 心理学用語「感覚与件」とは？認知科学の専門家で答えられない方がいる。英語ではsense data
  - 情報informationは、formを与える、formの中に置くという意味を内在している。このformはshapeとどう違うのか？
  - これらは英語教育の問題でもある。

提言

- 大学１－２年に数学を集中講義形式で教育（週10時間以上）
- 応用演習：先端科学技術分野の研究者を講師として招き、研究レベルに近い課題を学生に課して、数学によるモデリング能力と解析能力を養成し、コンピュータを用いた計算による分析能力を養成する。（1課題につき2週間以上を与え自身で考え解く力を養成。ホームワークを含む。）
- プログラミング・スキルの養成
- 抽象概念の理解力を養い、異分野の研究を理解するための英語教育と、コミュニケーション能力を高めるための英会話教育を実施