データサイエンスは何を育てるのか

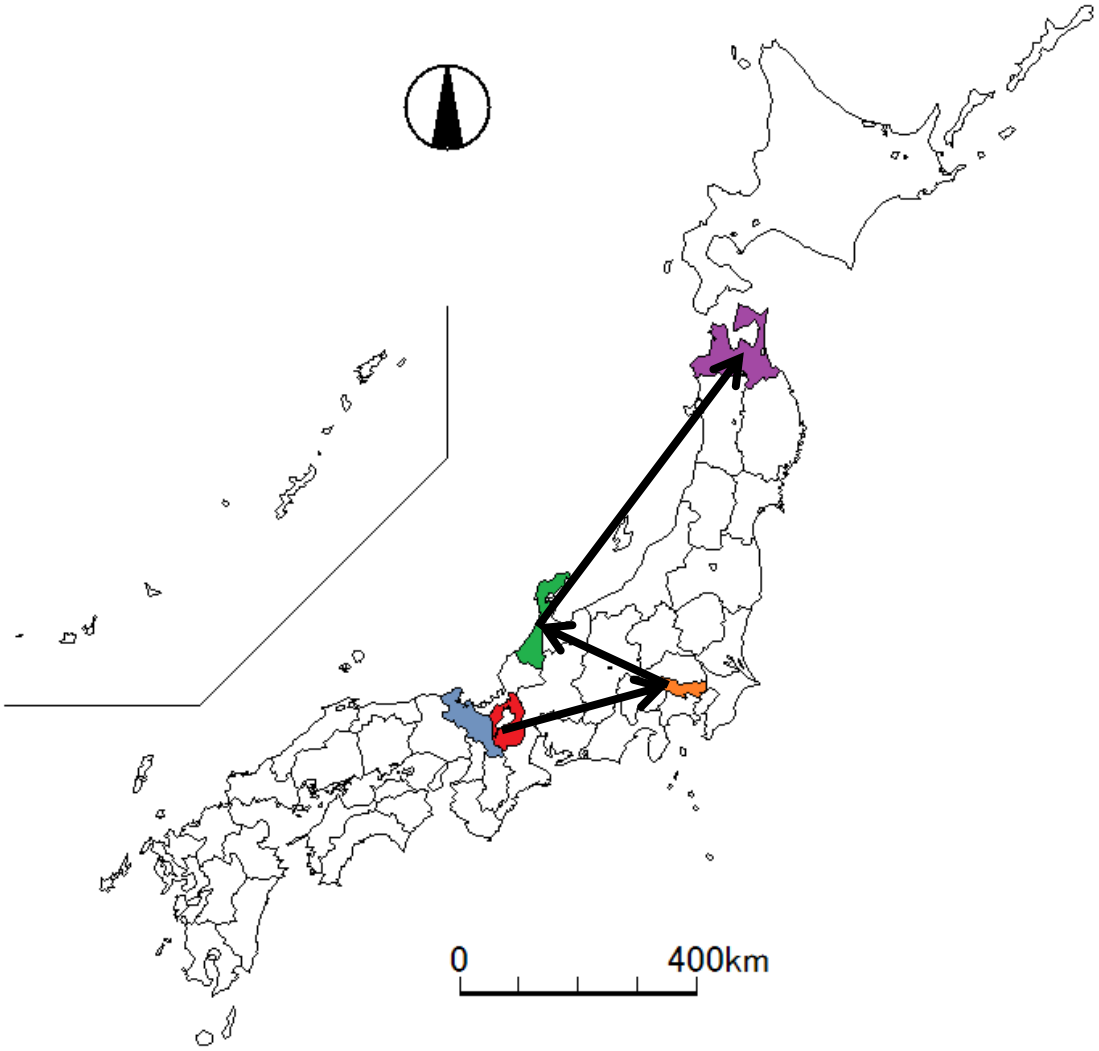
人文社会科学部
花田　真一
地域社会研究科公開セミナー
2022年1月12日

本日の内容のまとめ

- データサイエンスは何を育てるのか？
  - 意思決定に必要な情報を入手可能にする
  - 現在起きていることの延長としての未来の予測精度を上げる
  - 短期的な意思決定において失敗のリスクを下げる
- データサイエンスが発展すればバラ色の未来？
  - あくまで意思決定の「支援」であり、決定の責任は人間にある
  - 現在の状況から大きく変化するような予測には不向き
  - 因果関係よりは相関関係に重点が置かれるため、学術的知見の重要性は損なわれない

※あくまで個人の感想です
※講演者の性質上、経済学的文脈にやや偏っています

簡単な自己紹介



1981年生まれ
大津市立瀬田東小学校
私立洛星中学校・高校
2000.4～東京大学
　法学部（私法コース）（学士）
　公共政策大学院
　　（経済政策コース）（中退）
　経済学研究科
　　（現代経済専攻）（修士・博士）
2012.4～2013.3
　東京大学特任研究員
2013.4～2018.3
　金沢星稜大学経済学部
2018.4～
　弘前大学人文社会科学部

研究テーマ

- **実証産業組織論**
  - 政策や技術・社会情勢など**環境の変化**
  →産業を**企業・顧客の「組織」**と捉え変化を研究

- 中でも…
  - 経済主体・政策の**「シグナル」の効果**
  - **再生可能エネルギー**普及政策の経済評価
  - **技術制約を緩和する経済的手法**

データサイエンス？

The Federalist の著者は誰か？

- *The Federalist Papers*：アメリカ合衆国憲法批准の際に書かれた、憲法に関する論文集（全85篇）
  - 著者はAlexander Hamilton, James Madison Jr., John Jayの3人
  - 各論文は匿名で書かれており、分担については秘匿されていた
  - 後年、HamiltonとMadisonは各論文の執筆者リストを公表したが、食い違いがあった（互いに自分の著作としていた論文があった）
  - 様々な文学的な研究が行われ、おおよその区分はされたが、決め手に欠けた
    - 例えば、Jayが書いたことがほぼ確実なものもHamiltonは自分の著作としており、当事者間の対立以外の部分でMadisonのほうが信頼性がたかった

The Federalist の著者は誰か？

- MostellerとWallaceは、単語の使用頻度に注目して統計分析を行った
  - 著者について論争のない文献から、HamiltonとMadisonの言葉の選択確率を計算
    - 例えば、「…である、一方で」というときに、Hamiltonはwhileを使う傾向にあり、Madisonはwhilstを使う傾向にある
  - 著者不明の文献について、「普段の言葉の使用確率を守った場合、当該文献の使用頻度になる確率はどれくらいか」を計算
  - より確率の高いほうが著者であると考えられる
  - 実際には200近くの単語について重みづけをして同時に確率計算している
  - 計算の結果、Madisonのリストのほうが正しいことが分かった

- データサイエンスの用語を使うと自然言語処理、テキストマイニング、類似度判定、教師あり学習 etc.

The Federalistの著者は誰か？

- 実は、MostellerとWallaceの研究が発表されたのは**1963年**！
  - Mosteller, F., and D.L. Wallace (1963) “Inference in an Authorship Problem: A Comparative Study of Discrimination Methods Applied to the Authorship of the Disputed Federalist Papers”, *Journal of the American Statistical Associations*, 58(302), pp.275-309
  - 今から60年近く前に、すでに人文学の領域でデータサイエンスを用いた研究が行われている

疑問）データサイエンスは本当に「最新の」領域？

データサイエンスの定義？

- 「データサイエンス」という言葉自体、共通の定義が存在しない一種のbuzzword的な言葉である印象が強い
  - 言葉自体は1960年代にはすでに存在していたと考えられる
  - 政府の資料などにも「データサイエンティストの育成」「データサイエンス教育」などの言葉は使われるが、説明されている内容は主にデータリテラシー、統計学、プログラミング、AIなど以前からある内容の寄せ集め
  - 個人、文脈、分野、によって使われ方も捉えられ方も異なる

- データサイエンスという言葉の定義からアプローチするのは難しそう

データサイエンス今昔物語
～少しだけ個人の偏見～

- 私が大学院で統計学・計量経済学を学んだ際（2000年代後半）に強調されたこと
  1. 決定係数、p-値などをただ追求するのは不毛
  2. 回帰分析などで示されるのはあくまで相関関係であり、因果関係と勘違いしてはいけない
  3. 理論モデルに基づかないでデータだけを扱う研究は実証研究とは言わない、せいぜいFact Findingである
  - この辺のことが理解できていないとアホやと思われるから気いつけぇや

- 2020年代初期に言われている（ように見える）こと
  I. データをバカバカ突っ込めば、どんどんフィットが良くなって予測が当たるぜ、スゲー
  II. AIに学習させれば何でも予言できるようになるで
  III. AIさえ組んでデータを学習させれば、なんやようわからんが結果が出てきて便利や！
  - 違和感を感じるのは、私が古い人間だからなのか？

齟齬の原因＝目的の違い

- 前提条件：数字は嘘をつかないが、数字で嘘はつける
  - 「世の中には3種類の嘘がある。口に出して言う嘘（lie）、黙っている嘘（dammed lie）、そして統計（statistics）だ」～Benjamin Disraeli～
  - 統計学の知識があれば、データ処理と統計手法の組み合わせで同じデータから全く異なる結論を導き出すことが可能
- 学術研究の目的：一般化可能な結論を得ること
  - 「現在」の状況を正しく説明できることは（望ましいが）、必須の条件ではない→1
  - 統計モデルは「前提条件が正しければ」結論が正しい→計算ができること＝前提条件が正しいことではない→2
  - 構造がブラックボックスだと、他の状況に応用可能かどうか判断できない→3
- 実社会の要求：現状を少しでも改善すること
  - そのためには、現状が正しく説明できる必要あり→Ⅰバンザイ！
  - 因果関係がどうでも、結果さえ当たればとりあえず良い→Ⅱバンザイ！
  - 自分が知りたい範囲のことがわかれば、構造自体には興味がない→Ⅲバンザイ！

「データサイエンス」用語のかんたんな説明

- 機械学習：人間がモデルを指定するのではなく、機械（AI）がデータの学習から背後の関連性・モデルを発見する手法
  - 教師あり学習：結果（正解）を表す指標とセットでデータを与え、現実に近い結果指標の判定・予測を行わせる
  - 教師なし学習：データ間の相関などから類似度の判定方法を学習させ、グループ分けを行わせる
  - 強化学習：得られた結果に点数をつけ、できるだけ高い点数をつける予測を学習する方法
- 自然言語処理：人間が使う言語をデータ分析が可能な形に処理する方法
  - 人間は言葉の意味を理解できるが、AIは言葉の意味を理解できない
  - 意味はわからなくても、文字の並び（形態素）やその順番（構文）、辞書との対応（意味）などを認識できれば分類することができる
  - 分類の仕方自体は機械学習を応用する
- The Federalistの分析では、単語の頻度のみを利用しており、意味自体は利用していない→自然言語処理に基づく教師あり学習
- 検索エンジンも、意味を理解してページを探しているわけではない

データサイエンス活用の現状

社会科学（経済学）における データを用いたアプローチ

- 実証研究と課題発見（Fact Finding）
  - 実証研究：理論に基づくモデルを構築し、データと重ね合わせて係数を推定する研究
    - 理論モデルは、一般に状況によって変化する変数（variables）と変数がモデルに与える影響を表す係数（coefficients）で構築される
    - 係数がわかれば、変数が変化したとしても、結果を予測することができる
  - Fact Finding：現在の状況を観察することで、従来知られていなかった状況を把握する研究
    - 理論研究の発展の基礎となるケースがあるが、実務的な要請に答える面が強い
- 顕示選好アプローチと表明選好アプローチ
  - 顕示選好：実際に示された行動を利用するアプローチ
  - 表明選好：アンケートなどで表明された仮想的な行動を利用するアプローチ

実証研究と「データサイエンス」

- 実証研究の課題はデータの入手可能性と実証可能なモデルの構築
  - モデルに合致するデータを入手できないことが多い/複数の候補があり適切なものを選べない
  - 通常は均衡しか観察できないため、まずモデルの均衡を解として得る必要があるが、複雑なモデルは解析的に解けない
- 「データサイエンス」により
  - ビッグデータの整備→よりモデルに合致したデータの利用可能性の拡大
  - 計算能力の向上

  →複雑で解析的に解けないモデルを数値的に解くことが可能になった

  →巨大な連立方程式モデルを計算する時間もかなり短縮された
  - 機械学習によるモデル選択の支援→モデルの定式化の際に、過去の状況により合致したモデルの候補がある程度示される可能性がある

顕示選好アプローチと「データサイエンス」

- 顕示選好アプローチの最大の課題は、「実際の行動」の情報を得ることが難しいこと
  - いわゆるミクロ的なモデルは、各個人の行動がベースになる
  - しかし、個票データ（各個人の行動に関するデータ）の入手は、観察可能性や費用の問題から困難だった
  - そこで、個票データよりは入手可能性の高い集計データ（人々の行動の合計や平均）を本来個人のデータに基づくモデルに当てはめる手法が研究された→計量経済学
  - あるいは、アンケート調査などによる表明選好アプローチを採用した
- 「データサイエンス」により
  - ビッグデータ関連技術の発達・整備により、個票データの利用可能性が拡大した
  - 機械学習により、観察できたデータが従来より少数でも、ある程度正確な結果を得られる可能性が高まった
  - 自然言語処理、テキストマイニングなどにより、ツイッター、ブログ、日記などからデータの観測可能性が高まった

表明選好アプローチと「データサイエンス」

- 表明選好は、特に市場が存在しない（＝観察が難しい）行動についてよく用いられてきた
  - 実際の行動に比べると裏付けの不足が課題
  - 質問の仕方による結果の変化なども課題だった
- 「データサイエンス」により
  - 情報通信技術の発達により、従来より低費用で広範囲の表明選好の収集が可能になった
  - 機械学習により、表明選好と顕示選好が同時に観察された怜などを利用して表明選好の正確さを評価可能になった

つまり・・・

- データサイエンスの発展は、
  - 従来の手法の利用可能性・精緻化という意味で有益であった
  - 新しい概念を生み出すようなものでは今のところない

計量経済学のアプローチと「データサイエンス」

- 「データを扱う方法の研究」という意味では、計量経済学（Econometrics）という分野がある
  - 人文社会科学ではたいてい計量〇〇学（計量言語学、計量歴史学、etc）という分野があり、データを扱う手法を応用して研究が行われてきた
  - 基本的には数理統計学の知見を扱う分野の特徴、利用可能なデータの制約に合わせて調整する
    - 個票データの利用が困難→集計データを利用する手法の開発、など
- 計量経済学の分野では、線形/非線形、パラメトリック/ノンパラメトリック、古典統計/ベイズ統計、のようにいくつかの対比された手法が見られる

線形/非線形アプローチと「データサイエンス」

- 線形アプローチ：データの関係を概ね直線的であると仮定して分析
  - 途中で屈折したり、一定の性質を持つ曲線（二次関数など）で表したりしたとしても、部分的には「線形」であるとみなされる
  - 回帰分析が代表的な手法
- 非線形アプローチ：データの関係に仮定を置かない（どのような形でもよい）として分析
  - 実際には完全に不定形なモデルは推定できないので、モデルを細かく区切りものすごく狭い直線の連続として分析する
- 「データサイエンス」により、非線形アプローチの利用可能性が拡大した
  - 推定のためには狭い区間に一定以上のデータが必要
  - ビッグデータの発達などにより、区間を細かく区切ることが可能になった

パラメトリック/ノンパラメトリックと「データサイエンス」

- パラメトリック：モデルの特徴を表すパラメータを仮定し、その値をデータから推測する
  - 回帰分析における係数のような場合や、モーメント（平均、分散など）を合わせるようにする要素など
  - パラメータの値が決まれば、変数の変化に応じた結果の変化が予測できる
- ノンパラメトリック：モデルの特徴を表すパラメータを仮定せず、データの形自体がモデルそのものだと仮定して分析する
  - ものすごく細かい区間を区切り、その中でのデータの形をつなぎ合わせて全体の形と考える
  - パラメータが存在しないので、現実にデータが存在しない部分についての予測は難しい
- 「データサイエンス」により、ノンパラメトリックなアプローチの可能性が高まった
  - 局所的に膨大なデータが必要になるが、ビッグデータの利用により可能に
  - 計算能力が必要になるが、パソコンの情報処理速度の向上で可能に

古典統計/ベイズ統計と「データサイエンス」

- 古典統計とベイズ統計の違いについては若干議論があるが・・・
  - 古典統計：データの背後に何らかの既知の分布（正規分布、二項分布、ポワソン分布など）を仮定し、その分布のパラメータを利用する形でデータの分析を行う
    - 長所：分布の形が決まっているので、パラメータさえわかれば確率的な評価が容易
    - 短所：仮定した分布の形が現実と異なれば、誤ったデータの評価となる/推定したパラメータが誤っていれば、評価を誤る

      →一定以上のデータが必要
  - ベイズ統計：現状のデータの形そのものを分布として捉え、分析を行う
    - 長所：データの形そのものを分布とするので、少なくとも現実と大きくずれない/データが少なくても、Boostingにより仮想的に大きなデータを作れる
    - 短所：現状説明的な結論になりやすい/データが存在しない部分の評価は難しい
- データサイエンスの発展により、古典的統計は利用可能なデータの拡大の恩恵、ベイズ統計は機械学習による分布の正確な把握の恩恵が、老けられるようになった

つまり・・・

- 「データサイエンス」は非線形/ノンパラメトリック/ベイズ統計アプローチと相性が良い
  - 従来はデータの制約、計算能力の制限、人間に扱える関数の限界、などから仮定に頼らざるを得なかった
    - 直線と仮定すれば簡単な関数で部分的なデータでも全体に延長できる
    - パラメータの形を仮定すれば、その部分の計算に資源を集中することで人間にも処理が可能になる
    - 分布の形を扱いが容易なよく知られているものだと考えれば、確率評価が容易になる
  - データサイエンスはビッグデータ、情報処理能力の向上と効率的なアルゴリズムの開発、機械の特徴である単純だが作業量の多い作業の強み、等によって仮定を置かな異分析を可能にした
- 一方で、より現状説明的な結論に落ち着きやすい
  - 仮定を置かない代わりに「データは基本的に正しい」という前提条件で処理を行う
  - 総当りによる見落としの低下による発見はあるが、完全にデータのない部分の予測はあまり得意ではない

実務におけるデータサイエンスの重要性

- 様々な分野でデータサイエンスは活用されている
- 特に私の研究に関係が近い、政策評価と消費者行動について、簡単に説明する
  - それ以外にも、様々な場面で現実に利用されている
    - 例）迷惑メールフィルタ、など

証拠に基づく政策立案
（EBPM：Evidence Based Policy Making）

- 近年、政府や地方自治体でEBPMの重要性が指摘されている
  - データサイエンス人材の育成の目的の一つでもある
- EBPM：統計データや各種指標など、客観的な証拠・合理的根拠（エビデンス）に基づいて政策目的の明確化や評価を行う
- 本来当たり前に行われるべきものだが・・・
  - 特に日本では、エピソードに基づく政策立案や評価がなされてきた
  - そのため、政策の範囲に偏りが生まれたり、適切な評価が行われないケースが見られた
  - エビデンスに基づくことで、より合理的/効率的な政策を行うことが可能になる

EBPMの各段階と「データサイエンス」

- 政策策定段階
  - 現状把握と将来予測
  - 目的の明確化と評価指標の設定
- 政策実施段階
  - 現状把握による政策の調整
  - 中間的な成果の評価による、経路の把握
  - 目標水準の再設定
- 政策評価段階
  - 評価指標の目標水準との比較
  - 達成度/乖離度の原因の探求
  - 評価指標自体の適切さの評価
- AIの活用などにより、従来よりも幅広い目標指標の設定や精緻な現状把握が可能になる

例）AI×地方創生

- 広井・須藤・福田（2020）『AI×地方創生―データで読み解く地方の未来』、東洋経済新報社
- 京都大学と日立を中心とした共同プロジェクト
- 様々な自治体と共同し、AIを用いた予測と政策立案を支援している
- 特徴はいわゆるデータマイニングではない点
  - 利用するデータだけでなく、データ間のつながりや関係の方向性などは人間が設定し、AIはその枠の中で総当り的にシナリオを作成している
  - 汎用的な知識を持つ研究者と、地域の状況に詳しい自治体職員などが協同することで、より地域の実情にあったモデルの学習が行われている
  - シナリオをいくつかにグループ分けし、その分岐の鍵となる要素を目標に定めて政策立案を行っている

実務における消費者行動の分析

- 経営学・マーケティングの分野では、従来より一定規模のデータを用いた分析が行われてきた
  - POSデータを利用した購買行動の分析と店舗レイアウトの策定
  - 駐車場利用データを利用した店舗の商圏の把握と広告戦略の策定
  - 過去の契約実績に基づく、営業ターゲットの策定
- 「データサイエンス」の発達により、従来の分析を拡張して活用できるようになった
  - 相関分析により、従来見落とされていた意外な組み合わせ（おむつとタバコ、など）の発見
  - 位置情報などのビッグデータの利用により、会員以外の行動履歴の利用可能性の拡大
  - 決定木分析によるターゲットの設定
  - 機械学習によるレコメンド機能や最適腕識別による広告選択
- ポイントは、実務においては現状最適化が中心課題となる点
  - 学術研究における顕示選好とはやや目的が異なる
  - 完全に新しい製品の可能性などは、人間が評価する必要がある

データサイエンスは何を育てるのか

データサイエンスの強み/弱み

- データサイエンスの強みは以下のようなものではないか？
  - 従来利用できなかった質・量のデータの利用を可能にし、より精緻な分析が可能になる→ビッグデータなど
  - 総当り的なデータの処理が可能なので、見落としが少なくなる→機械学習など
  - 従来は量的に扱うことが難しかった質的データを、量的に処理可能な形式に変換する自由度が拡大した→自然言語処理、プログラミング、アルゴリズムなど
  - 主観を排し、客観的に現状を評価することが可能→機械学習、AI
  - 様々なシナリオを確率的なシミュレーションにより提示し、将来予測を行うことが可能→機械学習、アルゴリズム
  - 入力に対する出力を、現状を説明する形で提示できる→機械学習
- 一方で、データサイエンスには弱点もある
  - 「新しい発見」は総当たりによる見落としの補完の面が強く、その背後の意味などは付随しない
  - データのない部分（学習ができない部分）の精度は保証されていない
  - 相関関係の積み上げで評価しているため、入力が出力になる経路がブラックボックスになり、因果関係も保証しない

データサイエンス時代に必要な人材

- データサイエンス時代に必要な人材≠データを扱える人材ではないか？
  - もちろん、AIを構築したり、データ測定精度を高めたり、といったデータを扱える人材は必要

  →現在推進されているデータサイエンス教育はこの部分
  - 一方で、データサイエンスの「弱み」を補える人材の育成もかなり重要なはずである

  →この部分はややおざなりになっている・・・のか？

データサイエンス時代に必要な人材

- データサイエンスの「弱み」を補うために必要な資質
  - データサイエンスを適用している状況の実情に知見があり、AIが返した答え自体の信憑性を評価できる人材
    - 例）地域の実情に詳しく、AIの返したデータの塊からの答えを地域的に解釈可能な人材
  - データサイエンスを適用している分野の学術的な理論や構造に知見があり、データ分析の方向性を適切に修正したり、解釈できる人材
    - 例）消費者行動の理論に基づいて、入力と出力の因果関係を説明できる/結果を適切に解釈できる
  - データには現れていない状況を想定し、結果の調整が行える人材

→つまり、きちんと「勉強」した人材

データサイエンスを活用すれば・・・

- 人口が少ない地域でも、適切な政策立案や活動が可能になる
  - データサイエンス部分はある程度外注して、地域の実情に合わせてカスタマイズできる人材に集中する
- 政策資源が限られる地域でも、適切な政策立案の可能性が高まる
  - 他地域が行った政策の結果/プレテストの結果のデータから、自地域の結果を予測することができる
  - 先行する事例があるような場合に、それを教訓として次に生かせる
  - そのためにも、やはり地域に関する知見、学術的背景に関する知見は必要になる

データを使える人材になろう！

- AIの発達で人間の仕事がなくなる？→AIと人間は特徴が全く異なるので、置き換えられる部分は意外と少ないのではないか？
  - 作業分担、共同作業のほうが意味としては大きそう
  - AIはあくまでシナリオの提示、見落としのチェックといった意思決定の支援に強みがあり、意思決定自体は人間がしたほうが正確ではないか
  - 煩雑な作業を任せることで、人間はより本質的な部分に注力できる

  →逆に言うと、本質的な部分で発揮できる力の涵養は必要
- 理論や仕組みをある程度理解して、データに使われるのではなく、データを使える人材になりましょう