

日本におけるデータサイエンス教育の課題にどう向き合うか

鈴木寛 (Hiroshi Suzuki)

2023-08-30

1 はじめに

1.1 前提：フォーカス、個人の経験

データサイエンス教育一般ではなく、現在の、日本の大学（高専、短大を含む）において、すべての人のためのリテラシー・レベルのデータサイエンス教育をどのように始めていくかを、考えていきます。

すべてのひとには、理系だけではなく、理系以外の学生にもという意味合いだけではなく、教員や、できれば、職員をも含み、さらに、大学を発信源として、すべての人々に広がっていくようなデータサイエンス教育について考えたいと思います。

データサイエンスは、知識や理論ではなく、実証的な科学で、実際のデータの分析を通して対象物、そして世界を理解することが目的です。しかし、日本の大学において、データサイエンスを日常的に経験している人は、ごく僅かで、教員は多少の知識を持っている人はいても、経験している人はさらにすくなく、経験しているひとであっても、扱うデータは、非常に限られた専門的な分野のものに限られ、すべての人に対するデータサイエンス教育については、ほとんど全員が素人だと思われます。

以下は、これまで、わたしが、学び、教え、経験しながら、考えてきたことを基盤としていますが、データサイエンス教育について、集中的に考え始めてからはまだ数年、上に述べた目的設定も含めて、一般的ではないかもしれないことを、最初にお断りしておきます。

1.1.1 すべてのひと

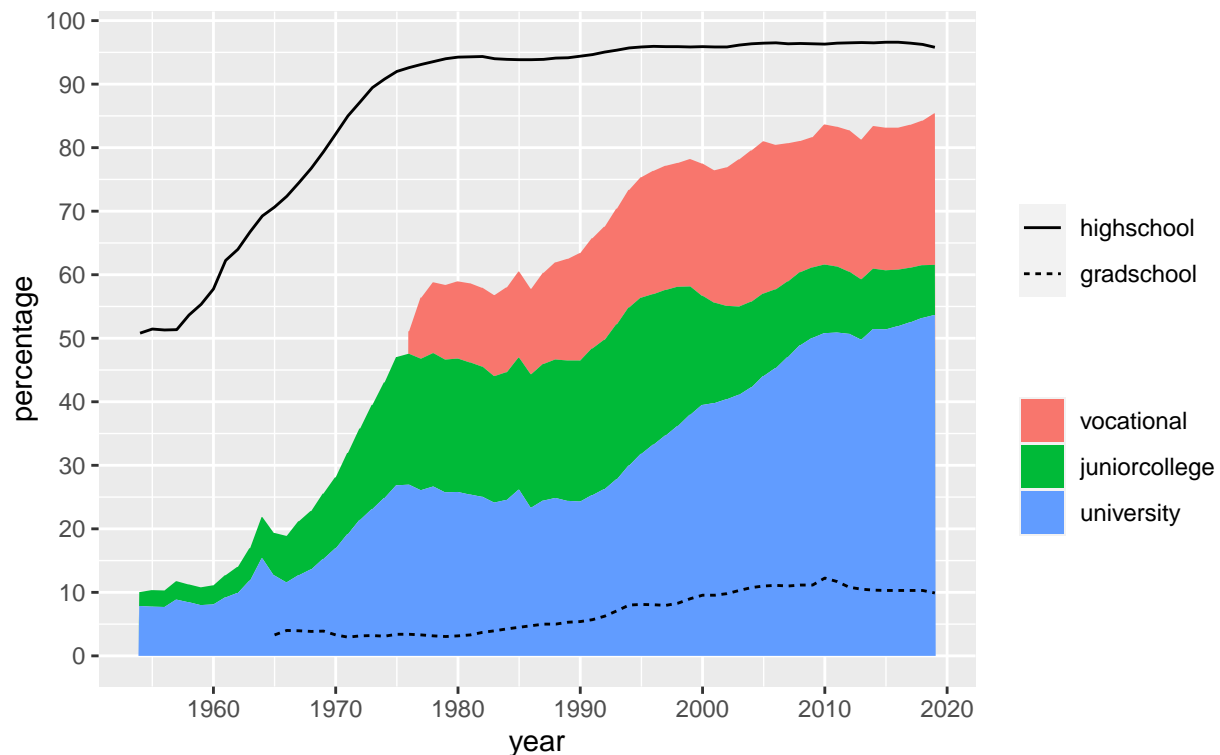
大学における教育を前提としつつも、「すべてのひと」と書きましたが、まずは、学生と教員に絞って、どのような人たちが対象なのかをみてみましょう。下のグラフは、男女共同参画局 のサイトから、データを取得して、学校種別進学率の推移について描いたグラフです。

わたしは、1972 年に大学に入学し、1980 年に大学に就職しました。このグラフでみると、かなりの変化の真っ只中ということになるようです。みなさんは、どうでしょうか。このグラフに現れているかどうかは別として、私たちは、かなりの変化の中にいることは、ご理解いただけたと思います。

グラフの中身に触れることは避けませんが、学生の生きる時代と、教員の生きてきた時代に、かなりの差が生じていることも、みて取れると思います。それは、経験や知識の量の違いではなく、学び方や、生活環境、価値観などの変化にも現れているでしょう。

Tertiary Education についてのグラフになっています。初等教育 (Primary Education)、中等教育 (Secondary Education) の、次のレベルについてのものですが、国際的には、そのレベルを、Tertiary Education と呼んでいます。それは、多くの場合、日本の専門学校を含みます。データサイエンス教育が、すべての人にとって大切ならば、義務教育ではない、高校での教育をどうするか、専門学校での教育をどうするかも、あわせて考えるべきだと思います。日本の教育費への支出は多くないことも確かだと思いますから、むしろ、予算措置についても考える必要があると思います。(OECD:Public Spending on Education)

Tertiary Education After Highschool with Highschool Graduates and Graduate School



1.1.2 個人的経験から

1.1.2.1 児童養護施設でのこと

- 楽しいこと、興味のあることは、いくらでも学ぶ
- ゲームより、魅力的であることも必要
- 英語も、コンピュータも小学校からはじまって、少しずつ変化している

わたしは、最近、児童養護施設に関わり、学習支援や、宿直のボランティアで、子供達と交流する機会を持っています。調査をしているわけではありませんが、明らかに、中高生の男の子は、ゲーム三昧、女の子は、YouTubeなどで、映像を見て、踊ったり、歌ったり。ハングルもある程度読める子もいます。

大学生なども含めて、なにか知りたくて学習するときには、インターネット、YouTubeで探して、解決する場合があります。特に、やり方（使い方・方法）がわからないときは、YouTubeなどで、そのためのビデオを探して学習しているようです。ゲーム感覚で学ぶことができれば、楽しいですし、学習も進むように思います。そのようなことは、不可能でしょうか。どなたか、アイデアはありませんか。

1.1.2.2 データサイエンスの授業を担当して

- Data Analysis for Researchers 研究者のためのデータ分析：大学院生一般（英語）
- Introduction to R 70分 x 2：経済を学んでいる大学院生（英語）
- Rではじめるデータサイエンス 70分 x 3 + 課題：中級マクロ経済学の学生（日本語）
- はじめてのデータサイエンス 70分 x 3 + 課題：初級マクロ経済学の学生（日本語）

研究者のためのデータ分析は、ある程度長く関わっていますが、中心的な部分を責任をもって教えるようになったのは、ここ4年です。JICAの働きの一つの、アジアの若手公務員の研修制度（JDS）の学生、ロータリ平和プログラムの学生と、一般の大学院生、約20人のクラスです。

データサイエンスに関しては、研究者のためのデータ分析を最初に教えたことが理由でもあります。あとで、紹介する、世界銀行（World Bank）、経済協力開発機構（OECD）、国際連合（United Nations）などが提供するデータを使

いました。その経験から、データベースが非常によく整備されていることを知ったことも幸運でした。

実は、似た分析を日本のデータベースを活用してしようとする、非常に手間がかかって、学生にクラスの中で、自分が興味のあるデータを見つけてもらって、分析するというレベルには、なかなか、たどり着けないということも、経験しました。学生が、自分で分析したいデータを扱えるようになることが目標で、分析技術を学ぶことはその一部でしかないからです。

世界銀行などのデータからはじめることができたのは、非常に幸運だったと考えています。

1.2 なぜ、今、データサイエンスか

課題をリストする前に、まずは、なぜ、いま、データサイエンスか。さらに、なぜ、すべての人、特に、大学などで学ぶ理系、文系を問わず、学生が、学ばなければならないかを短くまとめると次のようになると思います。

AI（人工知能）で、大きく変化しつつある社会において、すべての人が、どのように生きていくかを、個人で、そして、協力して、批判的思考を養い、根拠を確かめながら考え、意思決定に結びつけていくために、AI の背後にある、データサイエンスと、その考え方の基本を学ぶことが必須である。

そして、これに付け加えると、

日本は、国や社会のデジタル化、データを根拠とした議論、さらに、教育機関において、理系・文系の枠を超えた、AI やデータサイエンス教育が、非常に遅れていると認識されている。

ということだと思います。みなさんは、どう思われますか。

背後には、市民性が未発達で、一人一人が社会の一員として、課題に取り組んでいくという意識が十分醸成されていないというより大きな問題があるように思いますが、それについては、少しずつ、述べていきたいと思っています。

1.2.1 ハンス・ロスリング - Hans Rosling

- Factfulness (ファクトフルネス) 10 の思い込みを乗り越え、データを基に世界を正しく見る習慣: <https://bookplus.nikkei.com/atcl/catalog/19/P89600/>
- Gapminder: <https://www.gapminder.org/>
- How not to be ignorant about the world | Hans and Ola Rosling
- The best stats you've ever seen, Hans Rosling

「ファクトフルネス」は、四年ほど前に、注目を浴びた本で、ご存知の方もおられるかと思います。ハンス・ロスリングは、公衆衛生が専門の医師ですが、アフリカのいくつかの国での経験もふまえて、ひとが、多くの思い込み（バイアス）に支配されていることを示し、実際のデータを確認して、世界を正しくみる習慣をつけなければいけないと説いています。

その思い込みを乗り越えるためとして、立ち上げたのが、Gapminder というサイトで、最初に、10 個の間があり、時系列データをアニメーションで動かす、バブルチャートを自分でも操作することができるようにしています。後でも紹介する、ダッシュボードと呼ばれるものの一つの形でもあります。Google の支援もうけ、Google の Public Data Explorer というサイトでも、同様のものを使うことができるようになっています。（言語は英語しないと使えません）

ハンス・ロスリングは 2017 年に死亡していますが、心配している 5 つのリスクとして、感染症の世界的な流行、金融危機、世界大戦、地球温暖化、極度の貧困をあげています。生のデータをしっかりみることの大切さでしょうか。それが、思い込み（バイアス）を、少しでも乗り越える大切なリソースになるのだと思います。

1.2.2 デミス・ハサビス - Demis Hassabis

デミス・ハサビスは、AI の研究者で、さまざまな貢献をしています。若い頃は Chess プレイヤーで、英国の代表チームのメンバーにもなりましたが、その後、ゲームの開発をし、Theme Park などのシミュレーションゲームを開発、二十歳ごろに、研究の分野に入り脳神経科学の分野で、海馬の研究をし、海馬が萎縮するなどして、認知症を起こすと、過去のことを忘れるだけではなく、将来のことも、予測できなくなる傾向が強いことを示し、過去のデータから、未来を予測するメカニズムを研究して、強化学習 (reinforcement learning) に生かしたと言われています。

彼は、Atari のゲームを、49 個、同じアルゴリズムで攻略するソフトを開発して、披露していたときに、Google に認められ、彼の会社を買収されたとのことです。目標は、AGI (Artificial General Intelligence)。一つのことに特化した人間を凌ぐ AI ではなく、何にでもできる、AI です。これが、現在の、AI の一つの大きな目標になっています。

彼を有名にしたのは、囲碁の AI である、Alpha Go - Alpha Zero を開発した、DeepMind（現在は Google 傘下）の創始者だということでしょう。Alpha Go で、世界チャンピオンにもなったことのある韓国人棋士イ・セドルと対局して、第一局プロのだれも予想していなかった、素晴らしい手を打って、勝利をおさめた直後のインタビュー記事で、次のように語っています。

私が、AI に関して、本当にエキサイティングだと思うのは、科学をより早く進化させることを期待できることです。私は AI に助けられて進歩していく科学を見たいと思います。AI が、多くのつまらない労働をサポートしてくれるとともに、興味深いことを教えてくれ、山のようなデータから構造を見つけ、人間のエキスパートや研究者がブレイクスルーをもっと素早く達成できるように、助けてくれることです。数カ月前に CERN の研究者と話す機会がありましたが、彼らは、データの量が多すぎて膨大で処理できないようなデータと格闘しています。AI が膨大なデータの中から新しい何かを見つけてくれる未来はクールだと思います。（DeepMind founder Demis Hassabis on how AI will shape the future）

研究をしていると、問題によっては、非常に広がり、深さがあるが、あるものは、計算量などは多くても、あまり豊かなものが得られないという経験をすると思います。その違いを、素晴らしい研究者は、見抜くことができるが、そうでないと、なかなか見抜けない。また、素晴らしい研究者であっても、どうしても、自分の経験から判断することが多いので、すこしズレた分野に関しては、豊かな研究分野を見つけることが困難だと思います。それを、AI が、このあたりは、とても豊かな研究になる可能性がある、指摘してくれる、そのようなことが期待できるのではと言っているように思います。そして、それは、まだ、限られた分野ですが、少しずつ現実のものになっているように思います。

AI は、Pretraining（事前に学習する）データが重要ですが、同時に、それから学んだことを使って、ランダムに、広い範囲を探索することもできるからだだと思います。

Alpha Go では、過去の世界のプロの打碁を学習データとして使ったのですが、Alpha Zero では、そのようなデータは使わず、ルールだけを入力し、コンピュータ同士を戦わせることによって、強くしていき、Alpha Go で、イ・セドルを破った一年後には、それを完全に凌駕する Alpha Zero を作り上げています。これも、事前学習データがある程度減らすことができることも示唆している、重要な結果だと思います。

2 データサイエンス教育の課題

課題は、さまざまなまとめ方が可能だと思いますが、この論考で考えたい課題を整理して挙げてみたいと思います。

2.1 課題の整理

- 教育から学習へ
 - － 現在も、学ぶことが中心になっていないが、さらに、教員が教えることが中心だと、教員がわからないことは、学生は学びようがない。
 - － 教員は、専門から離れられず、多様性、社会の変化に適合した、教育には到達しない。
- 世界の課題、人間の課題として、向き合う認識が薄い
 - － ローカルな議論に終始することが多く、世界規模の課題に向き合うために、起業したり、世界の一員として活動することに向かわない。
 - － 同じような課題に世界で、どう向き合っているかを考える視点が欠如している。
- 英語の学習に時間をかけるが使う機会はない
 - － 大学においてすら、教員が使わないので、英語を学習に活用する機会がほとんどない
 - － 綺麗な英語を話したいなどの願望は、一部にあるが、聞き取りはできず、学習のために活用することはない。

まず第一に、教育から学習へと言われ、教育をしても学ばなければ、その教育の価値はないということは、ある程度、受け入れられても、どうしても、教員が教えたいことを教える傾向が強く、学生が学びたいことを学ぶこととの間をどのように折り返すかが難しい。データサイエンス教育では、もっと大きな問題として、教員が教育を受けていない、理解できていない、知らない、使っていないという、先に指摘した問題があると思います。当然で、だれにとっても、新しいものだからです。知らないもの、経験していないことを教えることは、基本的には不可能です。

第二に、わたしの個人的感触かもしれませんが、日本では内向き、AGI や、データサイエンスのような日本が遅れている分野でも、トップレベルの大学から、海外の大学院などに出て行く人あまりいない。さらに、世界で共通の課題を抱えているにもかかわらず、日本の中だけの経験で解決しようとする。数理・データサイエンス・AI 教育を待たずとも、さまざまな分野にある共通の課題なのではないでしょうか。先に述べた、教育費の問題もさまざまな国と比較することも

大切でしょう。所得水準が高くなれば、少子化、晩婚化は一般的ですが、日本より早くに、経験している国々でどのような対策をし、現在どうなっているかについては、ほとんど語られません。防衛費を、GDP の 1 パーセント程度から、2 パーセント程度にあげるというような議論においても、それは、世界において、どのような意味を持つのか、GDP を基準にすることは、どのような意味があるのかなど、殆ど議論されません。殆どすべての課題が世界共通でありながら、世界の状況を見ない。学ぼうとしない。これは、大きな課題であると共に、データサイエンス教育が関係する課題だと思います。

第三に、英語の問題です。最近、小学校から、コンピュータやプログラミングと共に、英語教育も始まり、使える英語というコンセプトで、英語教育が進んでいます。プログラミング教育や、英語教育が、どうなっていくのか、興味深いですが、少なくとも、現在の大学生、これから、10 年間に入学してくる学生は、そのような教育を十分には受けていない学生です。コンピュータの翻訳機能もどんどん進んでいますから、それを活用すれば、ネットで、英語が出てきたら、そこから先は、見ないというようなことを避けることが訓練によってできるはずです。さらに、プログラミング言語は、世界共通ですが、やはり英語がベースとなっている面もあります。世界の人たちと協力しながら、課題と向き合っていくには、プログラミング言語を介し、英語をも、コンピュータの機能を使いながら、活用して行くことが、有効なように見えます。

2.1.1 英語の問題

実は、今回のコースデザインで一番、苦労したのは、英語の問題です。

もう少し、詳しく述べておきたいと思います。

データを取得し、コード（短いプログラム）を書いて、分析を始めようとする、すぐ出会うのが英語の問題だと思います。なかなか実感できないかもしれませんが、世界のデータを取得しようとする、英語が必要になってくることはある程度理解できると思います。しかし、それだけではありません。プログラミング言語は、世界共通言語ですから、不明な点、エラーなどが生じたときに、英語でのコミュニケーションであれば、多くの支援を受けられますが、日本語の世界に止まっていれば、支援はその、百分の一程度にとどまるでしょう。あとでまた述べますが、英語では、無償で、公開された、誰でも使えるさまざまなリソースがふんだんにあります。パブリックドメインという考え方が、日本では未発達であることも関係していると思いますが、プログラムも含め、だれでも利用できるようになっているのが、プログラミングの世界です。英語が出てきた途端に、扉を閉じてしまっている、それ以上先に進めません。コンピュータツールを使うことに慣れれば、簡単にこのハードルを超えることができます。正しい、文法で、綺麗な英語などが、求められるわけではありません。ほとんどの人が、母語以外の言語として使っているのですから。ただ、日本人よりは圧倒的になれていると思います。

英語でのリソースは、圧倒的で、質も非常に高い。しかし、現状では、簡単には、勧められません。どのように回避しつつ改善していったら良いでしょうか。

自動翻訳を用いることで、デジタル化された文字情報は、かなり利用できます。しかし、ビデオなどのコンテンツが中心の Coursera や、edX など MOOCs、AI などの説明や、データサイエンスに関するコンテンツは、YouTube などふくめて、良い質のものがあありますが、英語の音声日本語で出力する技術は、改善は期待するものの、まだ、十分ではなく、現時点では推奨できません。

一番よいのは、聞いて理解する部分に少しずつ慣れながら、学習することですが、現時点では、その負荷を最初から、かけることは不可能であるとの結論にいたり、その理解のもとで、考えたというのが実情です。

2.2 現場の課題

- 一般に、コンピュータは、ホームページの検索、閲覧、ワープロとしての利用でとまっている
 - － 教員が十分使えないので、学生も趣味の範囲でしか使わない
 - － ワープロにちょっと表計算のようなものが限度
- 携帯端末は、SNS やゲーム、写真、音楽プレーヤーどまり
 - － 趣味以上のものにはならない
- 事実や、データを根拠にした、批判的思考になじめない
 - － 人格を尊重し、相手に配慮することはたいせつだが、丁寧に、論拠を確認しながら、ここは、ただし、ここは間違い、ここは、確認できない、他の視点から考えるなどの訓練がされていない
- 数学を含めて学問が閉じた世界になってしまっている
 - － 教員が楽しめることと、実際に使われることとの差が大きい

- － 教員は、楽しい、自分にとって有用と思えることしか教えない
 - － 学生は、楽しいと思えないもの、自分にとって有用だと思えないことは学ばない
- 時代の変化の中で、教育の中でどう組み立てていくかが考えられない。
 - － 教育は、次の世代を担う人たちのためであるはずだが、若い世代の人たちが生きる世界を考えずに教育に携わっている
 - － 世の中の変化のスピードが加速しているという考え方は受け入れられない

上に書いたことは、あまり説明がいらないと思います。しかし、大学でのカリキュラムを考えるときには、もう一つ重要な要素があると思います。それは、人的資源の問題です。

誰が教えるか、担当するかという問題です。ある程度、データ分析の経験のある教員が加わることは必要ですが、これから提案するコースでは、あまり教えず、学生と一緒に学んでいくことが大切だという前提で進めます。さらに、最初に学生も教員も学ぶことを書いていますから、すべての教員に関与してもらうことがたいせつだと思います。実際には、すべての教員というのは、さまざまな理由から不可能かもしれませんが、データサイエンス、または、データを元にした、根拠（Evidence）を明確にした、議論をもとに、意思決定して行くことの基礎を学ぶことを考えると、すべての教員、ほとんどすべての教員の関与が必要です。

データサイエンスに、あまり興味のない教員にも、加わって頂かないといけない。このことを、前提としてしっかり、理解しておくことが重要だと思います。

ついつい、コンピュータの使い方を知っているという理由から、情報科学や、理工系の先生が中心になって、カリキュラムを作成する。すると、わたしにはできません。あとは、お願いしますという感じで、それ以外の先生は、関与しないことになります。数学の先生が加わる場合も、統計的な部分なら少し勉強して教えられますなどという感じで、数学の授業の延長のようになってしまう。数学や統計は、データサイエンスに不必要ではありませんが、すべてのひとが学ぶデータサイエンスにおいては、不要といってもよい部分だと思います。必要が出てきたときに、わかりやすい説明をすることは大切ですが。

ですから、ある程度中心となる人は必要ですが、企画段階から、全学の教員に加わっていただいて、カリキュラムを策定することが必要でしょう。

3 学生と教員がともに学ぶデータサイエンス

すべてのひとに大切だとして、全学の学生が履修するデータサイエンス。むろん、教員が共に学ぶ必要があるはずです。

3.1 学生と、一緒に学び始めませんか

数学者、数学の教育者としてからいったん離れ、ひとりの人間として現実を見つめ、将来について考え、共に学ぶ姿勢をもつ。ここでは、以下の学びに焦点を当てる。

1. AI とどう付き合ったら良いか経験しながら考える
2. 時間的、地域的、個人的なバイアスから少しでも、自由になるためにデータから学ぶ
3. 多様な、価値観も異なる他者と協力しながら、考え、学ぶ価値を経験を通して学ぶ

このなかで、数学など、自分が学んできたことの価値を考え、活用していくかを、個人として考える

データサイエンスは、これらのために、非常に適した学びだと考えています。

カリキュラムの内容を考えるときは

数学も、統計学も、コンピュータ科学も一旦忘れて、必要に応じて考える

3.2 主たる学習内容

簡単に目標とする、コースの目的を書き、具体的な内容について、述べ、それから、どのようにそれを実現して行くかを考えたいと思います。

3.2.1 目的

以下で、考えるコースでは、コンピュータを使い、世界のデータを用いて、その分析をすることを学び、自分で興味を持ったデータを自分で分析し、世界の課題について、考えることができるようにすること。また、AI についての基本的な理解を得、AI と向き合う時の課題について考える。

3.2.2 学習方法

データサイエンスで最もたいせつなことは、データを見ることに慣れることだと思います。日本でも、コロナ感染症の影響を受け始めたころ、覚えておられるかもしれませんが、最初は、厚生労働省発表の数字だけが発表され、数が増えだすと、それはとだえ、都道府県のデータとなりました。報道では、Johns Hopkins University (JHU) から、発せられるデータや、そのダッシュボードが日本でも示されました。(WHO でもほとんど同じものが出ていたのですが、どういいうわけか (アメリカの影響が大きいと思いますが) それではなく、やはり WHO のデータを元にした、JHU のデータが使われました)

少しして、2020 年の夏前頃からでしょうか、NHK を始め、さまざまな日本の機関や団体が、さまざまなデータを公開しだしました。しかし、このように、さまざまな集計の仕方のデータが公開されることで、巷でも、それぞれの流行の波の予測なども、語られるようになりました。

まずは、データを見ることに慣れること。そして、次に、自分がみたい方法で集計したデータを見、そのデータについて理解することです。問いをもち、仮説を立てて、検証するようなことも、出てくるかもしれません。しかし、そのような予測・検証といった、プロセスに集中するのではなく、まずは、データからさまざまなグラフを作成し視覚化して、それから、データを理解することを学ぶこと、それを目標とすべきだと思います。

最後の予測は、分野ごとにさまざまな課題があり、なにを目的とするかによって、利用する、モデルも変わってくるからもあります。すぐ、p-値や、t-検定について、語りたい分野もいくつかあるようですが、それは、最後の方で、少し触れれば十分だと思っています。コロナのときも、そのようなことはほとんど語られませんでした。それは、たとえば、流行の波は、正規分布かどうかのテストをすれば、ほぼ正規分布になっていますが、実際には、そうではない部分、増加より減少のほうが緩やかであることを見たりすること、どこでピークアウトするかの予測は、簡単ではないことを見て取るなどです。実際のデータは複雑です。

3.3 リソースの確認 (I)

- コンピュータ - ハード・ソフト共飛躍的な進歩
 - コンピュータの進化で、膨大なデータ (Big Data) を活用できるようになっている
 - 利用者は (コンピュータサイエンスの) 専門家だけではなく、大きな広がりがある
 - コンピュータ言語: R、Python など、特に R
 - * これらは、世界中のひとの共通語で、国や地域に依存していない
 - ソフトウェア開発も、世界中の人が協力して行っている
 - * 抽象概念が、パッケージ、モジュール化されて利用できるようになっている
 - * Public Domain での共有が一般的
- データ - Big Data and Public Data
 - センサーや、インターネットの発達で、膨大なデータ (Big Data) が時々刻々と集められる、取得できるようになっている
 - **Open Data** という、人間全体の財産としての、データを適切に共有する動きが、特に国際機関などで進んでいる
 - 世界の中では、この認識に格差もあり、データの利用が十分考えられてない場合もまだある

データサイエンスにどうしても必要なのが、コンピュータと、データです。上に簡単にまとめたように、どちらにおいても、画期的な進歩を遂げたことが、データサイエンスや、AI の進歩につながっていることは、誰もが認めることでしょう。

コンピュータとデータ、それぞれについて一つずつ付け加えておきます。

コンピュータのソフトウェア開発において、特定の処理をすることに関して、その処理が汎用的に使えるように、抽象化された概念として、理解され、それが、モジュール、またはパッケージとなって、難しいアルゴリズムなどの議論なしに、組み合わせたり、組み込むことで実行できるようになったことが挙げられます。この考え方は、数学の手法だということもできると思います。同時に、深い数学は考えずに、使えるようになったと表現できる面もあります。最近では、作

業手順を、日常語で語ると、AI が、それぞれのプログラム言語で、コード（プログラム）を書いてくれるようになってきています。間違いも指摘し、修正もしてくれます。むろん、最先端は、難しいですが、基本的なことは、ほんの少し、勉強するだけで使えるようになってきています。

データサイエンスや、AI が画期的に進んだ背景に、Big Dataと言われる、大規模データが、どんどん生み出され、集められ、分析できるようになってきているということがあります。それは、たとえば、Google の検索データや、Twitter (社名が X に変わったようですが) などのデータ、さらに、さまざまなセンサーによって、取得し、収集されるデータも増えています。しかし、もう一つ大事なこととして、これらを含む形で、オープンデータ (Open Data) というデータ公開が進んでいることです。そして、その一つとして、人々の活動のデータを、公的機関などでの公開が進んでいるということが挙げられると思います。それは、パブリックデータ (公的データ、人々の活動のデータ) を、オープンデータとして、だれでも使えるように公開することで、社会の改善に利用していこうという方向性です。それを、公的機関は、使命、責任として行うところが増えてきています。背景には、データは、さまざまな人が、さまざまな分析を通して、違った目でみていくことによって、新たな、価値を生み出すという認識がひろがっているからだと思います。

そこで、専門家でもない、我々に、さまざまなデータがオープンデータとして提供され、コンピュータも、専門家に任せなくても、分析できるようになっていることを踏まえて、違った視点から、分析し、社会の改善に生かすことが可能になっており、一市民としての責任にもなっているということだと思います。

3.4 リソースの確認 (II)

- 教育、学習コンテンツ
 - 無償または非常に安価に提供されている膨大なコンテンツがある MOOCs (Coursera, edX, Udacity など)、ビジネス化されているものもある
 - 無償または非常に安価に、電子書籍などが膨大に出版され、共有されている
- AI の実用化が進んでいる
 - AI を使った教育が進んでいる。
 - Duolingo のような、無償での利用も可能だが、巨大になっているものもある。
 - * ミッション: We're here to develop the best education in the world and make it universally available.
 - * 私訳: すべてのひとに世界最高の教育を提供することがわたしたちの使命です。
 - * Personalized (Adapted) education 一人ひとりそれぞれに相応しい教育の提供
 - Chat GPT が有名ですが、Google 以外に、より複雑な検索を扱う、Perplexity もあり、今後、続々と、特徴的な、AI が公開され、有償だけでなく、無償でも利用できるようになっていくと思われる
- 自動翻訳の進化 - Google 翻訳は手軽に使い、ある程度実用的に使える段階にある - DeepL のように、十分な質と思われる自動翻訳も無償でも提供され、活用できるようになっている。

データサイエンスについては、無償の教育および学習コンテンツがほとんど、無限といってもよいほど提供されています。しかし、ほとんどが英語です。MOOCs (Massive Open Online Courses) の中で、一般的な、データサイエンスのコースが充実してきたのは、5 年ほど前からだと思います。学習コンテンツ、無償で公開されている電子書籍は、bookdown (R のパッケージ) などが整備されたことも大きいと思います。ここでは、十分説明できませんが、bookdown (<https://bookdown.org>) のサイトに行き、その中の、Archive (<https://bookdown.org/home/archive/>) を見れば、どれだけ、デジタルブックが無償で公開されているかを確認することができますでしょう。日本語のものもほんのわずかに公開されていますが、非常に限定的です。

AI 支援による学習 (AI Assisted Education/Learning) の例として、Duolingo (<https://www.duolingo.com>) をあげておきました。モバイルアプリ (Apple Store, Google Play) で無償で利用できるもので、語学が中心ですが、昨年末に算数 (Math) のアプリも登場しました。上に引用したように、すべてのひとに世界最高の教育を提供することを使命としていますから、これから、語学だけではなく、他の分野にも広がっていく計画のようです。語学では、CEFR (ヨーロッパ言語共通参照枠) の、B2 レベルを達成できるように、開発していると書かれています。このレベルまでできているのは、英語やヨーロッパ言語のいくつかだけですが、読み・書き・聞き・話すのいずれにも対応しています。わたしは、現在 10 言語で試していますが、素晴らしいのは、一人一人にあった学習プログラム (Adapted Learning) が AI によって、支援されていることですが、優れているのは、ゲーム化 (gamification) がなされ、続けることをさまざまな形で促す仕組みが非常によくできていることだと思います。ダウンロード数も、2 億を越え、365 日以上続けている人が、200 万人を超えているとのこと。日本の大学でも、英語能力の証明に使うところが出てきていますし、無償で使っていると広告を見ることがありますが、日本の大学の広告も増えてきているようです。Duolingo を使いながら、学習のためにどのような工夫がされており、AI がどのように使われているかを考えるのは、AI の理解のためにも、語学の学習

のためにも、とてもよいと思います。

三つ目は、自動翻訳の精度が上がっていること。数年前から、DeepL (<https://www.deepl.com/ja/translator>) が群を抜いていましたが、いまは、ChatGPT などの、LLM (Large Language Model) の、翻訳精度が高くなり、Google Chrome などを含む、ブラウザの翻訳精度も上がっているようです。比較データは持っていませんが、十分使用可能なレベルになっています。すなわち、電子的に提供されているコンテンツであれば、これらの自動翻訳機能を使うことで、英語に対する、ハードルをかなり下げることができるということです。

四つ目として、翻訳でも紹介した、Chat GPT, GPT4, Google Bard, Claude など、さまざまな、LLM (Large Language Model) の生成型 AI が登場し、学習をアシストしてくれるようになりました。まだ、論理的に非常に複雑なことや、新しいニュース、専門的な知識に関しては、幻覚 (hallucination) のような、回答を持っていないために、明らかな間違いを、自信満々に答えるような問題も存在しますが、少なくとも、ネット上で調べられるようなことについては、十分な答えをまとめて回答してくれます。特に、データサイエンスの学習においては、力を発揮し、学習を助けてくれる場面が多いと思います。

3.4.1 最近の MOOCs などからみるデータサイエンスのコースの特徴

- 理論よりプログラミングによる実践 (Theory to Programming)
- 統計よりコンピュータ科学 (Statistics to Computer Science)
- 仮説検定から機械学習 (Hypothesis Testing to Machine Learning)
- オープンデータの活用 (Open Data, Public Domain, Change of Publication Style)
- 探索的データ分析 (Exploratory Data Analysis)
- 視覚化をして特徴を見ることを重視 (Focus on Visualization)
- データも含め、パッケージや、モジュールの活用 (Package, Module)
- Git-GitHub や、R Studio などの IT ツールの活用 (Productivity Tool: Git-GitHub, R Studio)

さらに一つ付け加えておきたいのは、数年前には、MOOCs でのコースの提供が話題となりましたが、いまは、民間の企業で提供するコースや、YouTubers などによるコースなど、非常に多様な学習コンテンツが増え、AI も、データサイエンスのトレーニングも大学の中から大学の外に移っているように見受けられます。

3.5 カリキュラム例の確認

以下では MOOCs の一つである edX の Professional Certificate in Data Science を例にとって、内容をみていきます。

3.5.1 基本情報: HarvardX, through edX

演習も無償で殆どすべて提供

- URL: <https://online-learning.harvard.edu/series/professional-certificate-data-science>
- Book: <https://rafalab.github.io/dsbook/> (R Markdown Document)

履修証明を得るためには、有償登録が必要ですが、学習だけなら、ほとんどの部分が無償で公開されています。また、そこで使われる教科書も、電子書籍として、上で紹介した、Bookdown のサイトに公開されています。

3.5.2 R で学び、Data Camp で Assessment

1. Data Science: R Basics; データ解析ソフト R の基本
2. Data Science: Visualization; データの視覚化
3. Data Science: Probability; 確率・大数の法則
4. Data Science: Inference and Modeling; 推定と数学モデル
5. Data Science: Productivity Tools; Unix, Git, GitHub, R Markdown
6. Data Science: Wrangling; データの整理
7. Data Science: Linear Regression; 線形回帰
8. Data Science: Machine Learning; 機械学習
9. Data Science: Capstone まとめと次のステップへの架け橋

上のような9個のコースからなっています。言語は、R を使いますが、練習問題は、Data Camp が提供する、オンラインで、R を使えるサイトを活用しています。

データは、`dslabs` という R のパッケージを使います。このパッケージは、このコースのために、開発され、利用されるデータをすべて含むものとして、公開されています。

```
library(tidyverse)
library(dslabs)
list_dslabs <- data(package='dslabs')
list_dslabs$results %>% as_tibble() %>% select(3:4)
```

```
## # A tibble: 30 x 2
##   Item                                Title
##   <chr>                             <chr>
## 1 admissions                        Gender bias among graduate school admissions to ~
## 2 brca                             Breast Cancer Wisconsin Diagnostic Dataset from ~
## 3 brexit_polls                     Brexit Poll Data
## 4 death_prob                       2015 US Period Life Table
## 5 divorce_margarine                Divorce rate and margarine consumption data
## 6 gapminder                         Gapminder Data
## 7 greenhouse_gases                 Greenhouse gas concentrations over 2000 years
## 8 heights                          Self-Reported Heights
## 9 historic_co2                     Atmospheric carbon dioxide concentration over 80~
## 10 mice_weights (mice_weights) Mice weights
## # i 20 more rows
```

3.5.3 dslabs に含まれるデータ (訳 DeepL)

1. “UC バークレーへの大学院入学者におけるジェンダーバイアス”, Gender bias among graduate school admissions to UC Berkeley.
2. “UCI 機械学習リポジトリからの乳がんウィスコンシン診断データセット”, Breast Cancer Wisconsin Diagnostic Dataset from UCI Machine Learning Repository
3. 「Brexit Poll Data (ブレグジット世論調査データ)」, Brexit Poll Data
4. “2015 年米国期間生命表”, 2015 US Period Life Table
5. 「離婚率・マーガリン消費データ」“, Divorce rate and margarine consumption data
6. 「ギャップマインダー・データ」, Gapminder Data
7. “2000 年以上の温室効果ガス濃度”, Greenhouse gas concentrations over 2000 years
8. “自己申告制ハイツ”, Self-Reported Heights
9. “80 万年にわたる大気中の二酸化炭素濃度”, Atmospheric carbon dioxide concentration over 800,000 years
10. “MNIST データに基づく機械学習アルゴリズムの説明に役立つ例”, Useful example for illustrating machine learning algorithms based on MNIST data
11. “映画の視聴率”, Movie ratings
12. 「2010 年米国銃殺数 (州別)」, US gun murders by state for 2010
13. “欠損値がいくつかあるカウントデータ”, Count data with some missing values
14. “2010 年ニューヨーク州リージェンツ試験成績”, NYC Regents exams scores 2010
15. 「ギャップマインダー・データ」 “Gapminder Data”
16. “イタリアンオリーブ”, Italian olive
17. 「ギャップマインダー・データ」 “Gapminder Data”

18. “成人男性の身長（フィート）（外れ値あり）”, Adult male heights in feet with outliers
19. “2008 年大統領選の人気投票に関する世論調査データ”, Poll data for popular vote in 2008 presidential election
20. 「Fivethirtyeight 2016 年世論調査データ」, Fivethirtyeight 2016 Poll Data
21. 「オランダの研究費におけるジェンダーバイアス」 “Gender bias in research funding in Netherlands”
22. “自己申告制のハイツ”, Self-reported Heights
23. 「オランダの研究費におけるジェンダーバイアス」 “Gender bias in research funding in Netherlands”
24. “Fivethirtyeight 2016 年世論調査データ”, Fivethirtyeight 2016 Poll Data
25. “星の物理的性質”, Physical Properties of Stars
26. “世界の気温異常と炭素排出量、1751-2018”, Global temperature anomaly and carbon emissions, 1751-2018
27. “7 種類の組織から採取した 189 の生体サンプルの遺伝子発現プロファイル”, Gene expression profiles for 189 biological samples taken from seven different tissue types.
28. “2009 年から 2017 年までのトランプのツイート”, Trump Tweets from 2009 to 2017
29. “米国各州の伝染病データ”, Contagious disease data for US states

3.6 dslabs に含まれるデータから見えること

- 学ぶ課題に適したデータが選ばれている
- 受講者の多様な興味に答えられるようなデータが選ばれている
- 環境課題や、公平性を重視するデータが選ばれている
- 比較的新しいデータも含まれるようにしている

背後には、これだけの公開されているデータを具体的に分析して、分析自体を共有することも、社会として、進んでいることが挙げられると思います。日本では、これだけの講義は、可能な人がいないだけでなく、実質的に不可能だと思われる。これは、2017 年から 2018 年には、確立しているコースで、このレベルの内容を日本で提供するの、現在のところ、永遠に不可能であるように思われます。

3.7 MOOCs のコースからの考察

MOOCs で公開されている海外のコースを見てきました。それをそのまま、日本の大学で利用するのも一つの方法であるように思います。語学に関する困難は、翻訳ツールなどを活用して、切り抜けるとして、これらのコースを利用し、学習支援のような形で、学生の学習を支援することは、可能かもしれません。

しかし、土壌が熟成されていない日本では、借り物の域を出ず、ある程度の技術を学ぶことはできると思いますが、最初のにべた目的を達成するには、もう少し違う方法を考える必要があるように思います。

それを以下で考えていきたいと思います。

4 カリキュラムの提案

まずは、ChatGPT に、学生のモチベーションを高め、ワクワクするような、魅力的で、双方向性を大切にするような、カリキュラムの提案をお願いして得られた結果を示します。日本語でも、十分なやりとり（Chat）ができますが、以下は、英語を DeepL で翻訳したものです。元々の英語でのやり取りと、その和訳は以下の URL から読むことができます。

https://icu-hsuzuki.github.io/ds_education/chatgpt.html#dseducation

4.1 Chat GPT の提案

Chat GPT 学生のモチベーションを高め、データサイエンスにワクワクさせるためには、魅力的でインタラクティブなカリキュラムを提供することが重要です。ここでは、そのようなカリキュラムについて提案します。

1. 問題解決型学習：学生が共感でき、データサイエンスの手法を使って解決できる実世界の問題から始める。例：貧困と教育の関係の把握、都市の交通パターンの予測、エネルギー消費の傾向の分析など。

2. インタラクティブな可視化。Tableau、PowerBI、D3.js などのデータ可視化ツールを使用して、データを調査、理解するよう生徒に促します。これらのツールを使用してインタラクティブなダッシュボードやビジュアライゼーションを作成すると、データのパターンや関係性を確認することができます。
3. グループプロジェクト：少人数のグループでデータサイエンス、プロジェクトに取り組むことを奨励します。これにより、協力的で協調的な学習環境が構築され、学生同士が互いに学び合い、学んだことを実践の場で応用する機会を得ることができます。
4. ゲストスピーカー：データサイエンス分野の専門家を招き、学生を対象に講演してもらう。データサイエンスがどのような職業なのか、より深く理解することができます。
5. ハンズオンワークショップ。教室で学んだことを応用できる、実践的なワークショップを提供する。例えば、データ分析に R を使うワークショップや、データの可視化に Tableau を使うワークショップなど。
6. 学生主導のディスカッション：学生がディスカッションをリードし、自分たちの成果をクラスで共有することを奨励する。これにより、協力的な学習環境を構築することができ、学生同士が知識やアイデアを共有する機会を得ることができます。

これらの要素をカリキュラムに取り入れることで、魅力的でインタラクティブな学習環境を構築することができ、データサイエンスについてもっと学びたいという学生の意欲をかき立てることができます。

まとめると以下になるかと思います。

- 実社会の問題と関連づける
- インタラクティブな可視化を重視
- 少人数グループプロジェクト
- 実際のデータサイエンティストをゲストスピーカーとして呼ぶ
- R を使ったり、可視化ツールを使う、ワークショップと連動させる
- 学生主導のディスカッションを奨励する

二番目のインタラクティブな可視化で具体的に挙げられている、Tableau、PowerBI、D3.js は、みな無償で使えるものですが、どちらかというとビジネスで使うものです。PowerBI は、Microsoft が提供しているものですが、いずれも、ある程度学んでからでないと、初心者が使いこなすには少し難しいように思えます。しかし、全体としては、適切なカリキュラム構成の一つの型を提供しているように見えます。

4.2 データ

オープン・データの一例として、世界銀行の世界開発指標を取り上げます。オープンデータは、日本のデータ (e-Stat: <https://www.e-stat.go.jp>)。

4.2.1 世界銀行 (World Bank)

- 世界銀行オープンデータ
- データカタログ
- 世界開発指標 (WDI)

4.2.2 オープンデータの定義 (Open Data Defined)

オープンデータという言葉は、厳密な意味を持っています。データまたはコンテンツは、出所が明示されオープンという性質が維持されれば、誰でも自由に利用、再利用、再配布できるものを言います。

1. データは法的にオープンでなければなりません。つまり、パブリックドメインに置かれ、最小限の制限で自由に使用できなければなりません。
2. データは技術的にオープンでなければなりません。つまり、誰でも自由に使える一般的なソフトウェアツールを使ってデータにアクセスし、機械で読み取ることが可読な電子フォーマットで提供されていなければなりません。パスワードやファイアウォールによる制限を受けずに、公共のサーバーで、だれでもアクセスできなければなりません。また、オープンデータを見つけやすくするために、さまざまな組織がオープンデータカタログを作成し管理して必要があります。

データはだれのものでしょうか？ パブリックデータのパブリックとは？

世界開発指標のサイトには、ダッシュボードと言われる、ある程度、自分で操作して、グラフを作成するようなツールもあります。詳細は、下にもリンクがある、「データサイエンスをはじめましょう」を参考にしてください。

英語版の、Google Public Data Explorer (https://www.google.co.jp/publicdata/directory?hl=en_US&dl=en_US#!) からは、さまざまなデータのダッシュボードが利用できます。トップには、Factfulness の項で紹介した、バブルチャートがあります。また、データベースのリストの最初には、World Development Indicators（世界開発指標）があります。データは最新ではありませんが、折線グラフ（line chart）、棒グラフ（bar chart）、散布図（scatter plot chart）、色付き地図（choropleth map）が作成できるようになっています。

4.3 コースの提案

簡単に、コースの提案を書きます。コースは、受講学生にあったものを作成する必要がありますので、あくまでも、一つの例です。

レベル：リテラシーレベルを二段階に設定して設計

内容：基本的に下の二つとし、それを融合させた形で行う。

1. AI（人工知能）との付き合い方を学ぶ
2. 世界や、ひとびとの課題について、データサイエンスで学ぶ

形式：グループワーク

1. 少人数で、調べることに共に、問いを精査する場面で協力を促す
2. グループの一員として、AI を活用する

4.4 第一レベル

1. AI に触れる：Duolingo、ChatGPT など
2. WDI: 日本に関するデータ - ダッシュボード
3. WDI: 指標を選択して、課題と向き合う
 - ・ 自分達の興味のあるデータを問題意識をもって調べる
 - ・ バイアスに注意しながら、データに基づいた批判的思考ができるようにする
 - ・ 他者と協力、分担して、調べ、ディスカッション - 協働
4. AI のしくみの基本を（ビデオや外部講師などから）学ぶ
5. AI との付き合い方についての、ディスカッション
6. トピックを決めてグループ発表
 - ・ データ自身の用語、指標の定義が重要
 - ・ 数学・統計学・コンピュータ科学はこの段階では不要
 - ・ 課題への取り組みを見ながら、アドバイスをし、自分も学ぶ

4.5 第二レベル

1. AI とデータサイエンスについてのディスカッション
2. PositCloud 入門、スクリプトの利用
3. R Notebook 入門
4. WDI パッケージの利用
5. テンプレートの利用
6. tidyverse の基本の学び
7. 可視化の学び
8. WDI 以外のデータの取得
9. プロジェクト
10. 発表
 - ・ 探索的データ分析（Exploratory Data Analysis）
 - ・ 統計的概念と意味と活用法は、必要になるまで待ちその都度説明

- 四分位 (Quartile)・平均 (Mean)・標準偏差 (Standard Deviation)
- 分布 (Distribution)・相関 (Correlation)
- 相関係数 (Correlation Coefficient)・回帰直線・p 値・R Squared

4.6 たいせつにしていること

さまざまな生のデータをいろいろな角度から見て、考えることを中心に置き、数学や統計やコンピュータ科学といわれるものは、いったんおいておいて、必要になったときに、少しずつ説明するという姿勢を強調してきました。少しだけ、補足しておきたいと思います。

- 数学・統計学は理解しなくても、たいせつさは、伝わるのでは
 - 背後に多くの、数学・統計学・コンピュータ科学が使われていることは理解できる
 - そのことを理解している親のこどもに、たいせつさが伝わる
 - 現在日本では、自分ではできない。自分には関係ないということだけが残っている
- p-値や、R 二乗値は、グラフを見てその傾向について結論したいときに説明
- 相関は因果にあらず (Correlation is not causation!)
 - 回帰直線の方程式あたりで必ず誤解するひょうが出てくるのでそのとき
- 交絡因子 (Confounder) は、学生が疑問に思う時を待つ
 - その授業で例が出てきた時に、使われているデータを使って説明
- 平均回帰の用語の起源 (Francis Galton)
 - “背の高い親の子の身長は、通常の平均値に向かって減少する傾向がある”
- アンナカレーニナの法則 (AKP)
 - “幸せな家庭はみな似たりよったりだが、不幸な家庭はみなそれぞれに不幸である”
 - 世界のデータに関する傾向や、外れ値や、少数者の理解

4.7 まとめ

午前、午後と 2 回の講演機会をいただきました。講演で使用したスライド、資料および、ビデオは、下の資料にリンクをつけてありますので、参考しにしてください。

特に、午後の講演の、実際の授業内容については、現在デジタルブックとして執筆中の「データサイエンスをはじめましょう」を参考にしてください。大体の内容を書き終わったら、YouTube ビデオだけでなく、できれば、対話型の、練習問題もつけていきたいと考えています。

また、実際に教えたコースの資料や、特別講師として担当した授業の資料も掲載しています。何らかのヒントになれば幸いです。

上に書いたように、英語では、素晴らしい、リソースがたくさんあります。わたしも、それらで学んだのですが、同時に、現在の日本の状況を考えて、今、何をすべきかについて、書かせていただきました。

今後、どのように進んでいくかは不明ですが、一応、全学必修にしていきたいという文部科学省の意向が最初から示されていました。大学教育でそのようなものは、かつてなかったように思います。それを、好機ととらえ、少しでも、一人一人が自分でデータを取得して、それをもとに考え、さまざまな改善点について考えていく、そんな一歩となればと考えています。

日本政府が提供しているデータは、使いやすいとは言えませんが、日本中の学生たち、そして、教員たちが、さらには、社会で、さまざまな人たちが、オープンデータを利用するようになれば、データベースも自然に、改善されていくと思います。

情報科学の専門の先生などに依頼して、コースを最低限提供するのではなく、教員もほとんど経験していないのですから、学生とともに学んでいくことを願っています。

5 資料

わたしのホームページのデータサイエンス関連ページデータサイエンスを学びませんか・データサイエンス教育には次の（短縮した）リンクからも行くことができます。

<https://bit.ly/rims2023>

以下の資料が掲載されています。

- データサイエンスを学びませんか
 - － データサイエンスをはじめましょう（日本語のデジタル教科書：未完成）
 - － はじめてのデータ・サイエンス（マクロ経済学原論での 70 分 x3 回の特別講義（2023 年 6 月 8 日（1 時間）・13 日（2 時間））のために作成したもの：ビデオ・資料）
 - － R ではじめるデータ・サイエンス（中級マクロ経済学での 70 分 x3 回の授業をもとに、2023 年 3 月 9 日の数理科学研究所での講演のために作成したもの）
 - － Introduction to R - A short course（大学院生向けの特別講義（70 分 x2）資料、英語が苦手な方は、Google 翻訳などを ON にして読んでください。）
 - － Data Analysis for Researchers 2022 - Course Material（2022 年冬学期大学院一般向き授業の資料）
 - － Data Analysis for Researchers 2021 - Course Material（2021 年冬学期大学院一般向き授業の資料）
- データサイエンスを教えませんか
 - － データサイエンスを教えませんか（始めたばかりですが少しずつ書いていく予定です）
 - － 日本でのデータサイエンス教育の課題にどう向き合うか・スライド・ビデオ（2023 年 3 月 9 日の数理科学研究所での午前の講演）
 - － 生涯学び続ける基盤を構築するデータサイエンス・コースの開発・スライド・ビデオ（2023 年 3 月 9 日の数理科学研究所での午後の講演前半・午後は「R ではじめるデータ・サイエンス」を利用）
 - － 日本数学会教育委員会主催教育シンポジウム：文理共通して行う数理・データサイエンス教育における講演（時：2019 年 9 月 17 日 於：金沢大学角間キャンパス
「教養としてのデータサイエンス教育～MOOCs の活用を視野に入れて～」
スライド [pdf]、音声付き画面収録ビデオ [Large mp4 (1177M)], [Small mp4 (447M)]
- わたしのデータサイエンス
 - － わたしのデータサイエンス（始めたばかりですが少しずつ書いていく予定です）