

Curriculum Guidelines for Undergraduate Programs in Data Science

学部レベルのデータサイエンス・プログラムのカリキュラム 指針

Richard D. De Veaux 他 24 名 *

概要

2016 年の Park City Math Institute 夏期学部教育プログラムは、学部レベルのデータサイエンス・プログラムのカリキュラム指針を作成する目的で開催されました。このプログラムのメンバーは、数学、統計、計算機科学の分野を中心に、米国の様々な大学の学部教育を担当する 25 人の教員から構成されていました。本稿で提案する指針は、データサイエンスの専攻を計画または再編する大学のための枠組みを提供することを目的としています。

* Mahesh Agarwal, Maia Averett, Benjamin S. Baumer, Andrew Bray, Thomas C. Bressoud, Lance Bryant, Lei Z. Cheng, Amanda Francis, Robert Gould, Albert Y. Kim, Matt Kretchmar, Qin Lu, Ann Moskol, Deborah Nolan, Roberto Pelayo, Sean Raleigh, Ricky J. Sethi, Mutiara Sondjaja, Neelesh Tiruvilumala, Paul X. Uhlig, Talitha M. Washington, Curtis L. Wesley, David White, and Ping Ye

日本語訳の公開にあたって

平成 28 年 12 月、北海道大学、東京大学、滋賀大学、京都大学、大阪大学、九州大学は数理及びデータサイエンス教育の強化に関する懇談会における評価結果を踏まえ、文部科学省より数理及びデータサイエンスに係る教育強化の拠点校として選定されました。拠点校は、数理・データサイエンスを中心とした全学的・組織的な教育を行うセンターを整備して、各大学内での数理・データサイエンス教育の充実に努めるだけでなく、全国の大学に取組成果の波及を図るため、地域や分野における拠点として他大学の数理・データサイエンス教育の強化に貢献することが期待されています。そのために東京大学を幹事校として数理・データサイエンス教育強化拠点コンソーシアムを形成して、以下の取り組みを行うことになっています。

- 全国的なモデルとなる標準カリキュラム・教材の作成
- その標準カリキュラム・教材の他大学への普及方策（例えば全国的なシンポジウムの開催等）の検討及び実施
- センターの情報交換等を行うための対話の場の設定（各大学のセンターにおける教育内容・教育方法の好事例を共有し、より取組を発展させるための議論など）
- センターの取組の成果指標や評価方法の検討

特に標準カリキュラム・教材の作成は、当面の最重要ミッションです。

御承知の通り、データサイエンス教育はアメリカの大学で先行しております。2016 年の Park City Math Institute 夏期学部教育プログラムではデータサイエンスの学部プログラムのカリキュラム指針を作成する目的で、アメリカのデータサイエンス関連分野の大学教員が集まって議論しました。本論文「Curriculum Guidelines for Undergraduate Programs in Data Science」[Annual Review of Statistics and Its Application](https://doi.org/10.1146/annurev-statistics-060116-053930) (2017 年 volume 4, p15–30, <https://doi.org/10.1146/annurev-statistics-060116-053930>) は、その成果物です。アメリカと日本では、大学の教育体系や関連分野（数学、統計学、計算機科学）の国内における歴史的背景が違うため、本論文の提案がそのまま役に立つわけではありません。しかし、日本でデータサイエンス教育を強化していく上で重要な資料であることは間違いなく、関係者各位と情報を共有したく論文の日本語訳を公開致します。

なお、学会として本ワークショップに関与したアメリカ統計学会、及び筆頭著者の De Veaux 教授には日本語訳の公開の許可を頂いております。

翻訳 丸山 祐造（東京大学 数理・情報教育研究センター）

1 はじめに

データサイエンスは、科学、産業、行政における複雑かつ詳細なデータの急速な蓄積に刺激されて、想定を上回るペースで急速に成長しています。頻繁に引用される McKinsey の最近の報告書 (McKinsey Global Inst 2011) では、今後十年間にデータサイエンスに関わる人材が数十万人のオーダーで必要になると予測しています。このような産業界からの報告書に半ば煽られる形で、大学関係者はその需要に応えるべくデータサイエンスプログラムをスタートさせています。現在、ウェブサイト <http://datascience.community/colleges> では、世界の 200 以上の大学における、データサイエンス、データ解析、および関連分野の 530 のプログラムが掲載されています。これらの大部分は修士課程のプログラムであり、通常の修士課程の他オンラインで取得できるプログラムも含まれています。データ・サイエンス (またはデータ分析) の博士号プログラムは依然として比較的少数ですが、学部プログラムが研究大学とリベラル・アーツ・カレッジの両方で急速に増えてきました。我々は近い将来、これらの数字が大幅に増加すると予想しています。

国立科学財団 (NSF) とプリンストン高等研究所が主催する Park City Mathematics Institute (PCMI) は、データサイエンスの学士号のカリキュラム指針を策定する作業にフォーカスした **ワークショップ** を 2016 年に開催しました。様々なリベラルアーツ・カレッジや研究大学の計算機科学者、統計学者、数学者からなる 25 人の教員が、学部レベルにおけるデータサイエンス教育において、「学生が身につけるべきと考えられる能力・スキル」や「現在から将来にわたって、データサイエンスプログラムをどのように運営していくか」について、3 週間にわたって議論しました。本稿ではその成果であるカリキュラム指針を報告します。

我々は、専攻の修了認定のためにセメスター科目 (訳注: 2 学期制における週 2 回開講の科目だと思われる) を 10 科目履修すると定められているデータサイエンス専攻を対象として、カリキュラム指針を策定しました。なお 10 科目というのは、リベラルアーツ・カレッジにおいて、典型的な数字です。もちろん、研究大学においてはそれらにいくつかの科目を追加することになるでしょう。我々は、これらの指針に厳格に沿ってプログラムが組まれるべきとは考えておらず、むしろデータサイエンスプログラムの学生が在学中に習得すべきコアスキルを列挙することに意味があると考えています。我々は、

- NSF のデータサイエンス教育ワークショップ (Cassel & Topi 2015)
- AALAC (リベラルアーツ・カレッジ連盟) のカンファレンス「教養教育においてビッグデータをどう教えるか」
- 関連分野である数学、統計、計算機科学の学部教育指針 (下の枠内を参照のこと)

の報告書を検討することから議論を始めました。

関連する学問分野におけるカリキュラム指針

- [2015 CUPM Curriculum Guide to Majors in the Mathematical Sciences \(MAA 2015\)](#)
- [Computer Science Curricula 2013: Curriculum Guidelines for Undergraduate Degree Programs in Computer Science \(ACM/IEEE 2013\)](#)
- [Curriculum Guidelines for Undergraduate Programs in Statistical Science \(ASA 2014b\)](#)

本稿の構成は以下の通りです。第 2 節では、データサイエンスの背景及び我々の考える基本理念をより詳しく議論します。第 3 節では、データサイエンス専攻において学生が身につけるべきスキルを考察します。第 4

節と第 5 節では、肝であるカリキュラムの主題を整理します。第 6 章では、我々が重要だと考えるスキルの多くをカバーしうる既存の科目に基づくカリキュラムの一例をお見せします。ただし、このような既存の科目の寄せ集めでは理想とは程遠いのです。計算機科学、統計学、数学で伝統的に提供されている科目の多くは、効率性、及び相乗効果の観点から、データサイエンスプログラムのために内容を再編成して提供されるべきであると考えています。多くの大学における既存の科目を前提とすると、学生はデータサイエンスに必要なスキルを身につけるために 14 を超える科目を履修する必要があります。既に述べたようにリベラルアーツ・カレッジの専攻で典型的な 10 セメスター科目の中に、我々が重要だと考えるスキルの多くを含めるように設計することが重要です。我々は、抜本的な講義内容・カリキュラムの精選によって、それが可能だと考えています。そのような科目体系の詳細は、付録 A にまとめられています。

2 背景と理念

我々は、データサイエンスの他の学問分野における重要性の高まりや、来たるべきデータ駆動型経済におけるデータサイエンティストへの強い需要に応えるべく、カリキュラム指針をまとめます。

2.1 科学としてのデータサイエンス

データサイエンスを厳密に定義することは難しいのですが、我々は NSF の数物科学委員会の小委員会である StatSNSF での [コンセンサス](#) を立脚点として議論を始めました。それは、データサイエンスは、データの設計、取得、管理、解析、そしてデータからの推論を目指す科学である、というものです。学部レベルの教育ではデータサイエンスを、現実を記述するためにデータを用いる点を強調して、工学のような応用分野とみなします。現在のところ、理論的基盤は主に、統計学・計算機科学・数学という確立した学問分野に由来します。ただしデータというのは、「どの学問分野においてどのような文脈でデータが生じているか」が重要です。その各学問分野の文脈においてデータを解釈することにより、データの現実世界での実際の意味がわかります。このため学部プログラムでは、我々は工学や計算機科学の分野にならって、ケーススタディや実践を重視します。

2.2 データサイエンスの学際性

データサイエンスは本質的に学際的です。データを扱うには、統計、計算機科学、数学の分野に関連する多くのスキルや概念を熟知している必要があります。データサイエンスは、3 つの学問分野の教育内容を融合したものでありますが、それらの単純な共通部分でも上位集合でもありません（下のコラムを参照して下さい）。データを扱うという文脈において各学問分野から必要な概念を抽出してそれに専念することにより、カリキュラムを大幅に合理化し、強化することができます。データにフォーカスして科目の統合・融合を行うことが効果的なデータサイエンスプログラム設計の本質であり、課題解決への相乗的なアプローチをもたらします。

本稿では、データサイエンス専攻の学生が習得すべき主要な知識と方法論を概説します。もちろん理想的にはデータサイエンスのためにカリキュラムを大幅に合理化し、強化された科目群が新規に設定されることが効率性と相乗効果の観点から望ましいと考えます。しかし、全ての大学がすぐに多くの新しい科目群を作成できるわけではないので、どの既存の科目がデータサイエンス専攻での基本的なトピックをカバーしうるかを説明します。また将来におけるカリキュラム設計図として利用可能なデータサイエンスカリキュラムの一つのモデルを提案します。

データサイエンスは「科学」なのか？

データサイエンスの「サイエンス」とは何かという議論は今でも続いています。しかし [David Donoho](#) や [Michael Jordan](#) といった著名な科学者達は、データサイエンスには確固としたサイエンスと言えるものがあり、また、データサイエンスはさらに進化し続けるだろう、と考えています。例えば、[Donoho 2015, p. 10](#) は、「幸いにして、『データサイエンス』なる科学の分野が生まれるということに対する十分な論拠がある。データサイエンスは真の科学となっていくだろう。永続的に取り組まれていくような本質的な課題があり、科学的に厳密な技術でそれらに取り組む。これは正しく科学である。」と述べています。

もちろん、データサイエンスのサイエンスとしての進展をめぐるコンセンサスの有無に関わらず（あるいは、それがなくても）学部レベルのデータサイエンスプログラムは、3つの既存の学問分野（統計、数学、計算機科学）の成果を活用しながら相乗効果を生む、問題解決のアプローチを提供するので、学生が修了後どのような進路を取るにせよ（産業界に就職する、修士課程などでさらに学問を続ける、など）、データサイエンスプログラムが学生の役に立つと我々は確信しています。

2.3 データ中心主義

データの取得、データの前処理、データ・キュレーション、データ・マネジメント、データ・プロセッシング、探索的データ解析、研究開発する上での課題の定義、データ解析の実行、結果を共有し分かりやすく伝えること、これら一連の作業の繰り返しこそが、データサイエンスのコアを成しています。学生は、このデータサイクルの全てのステップを理解してさらに実行する訓練を必要とします。そのような理解、訓練があってこそ、現実の研究課題に適切に携わることが出来るようになります。Google の [Diane Lambert](#) の言葉を借りれば、学生には「データと共に思考する」能力が必要というわけです。データと共に思考する体験は、初級科目から上級科目、卒業実習まで全ての科目において中心的な役割を果たす必要があります。また様々なソースからの生データの取り扱いや、さらに解析の前処理としてデータのクリーニング、データ変換、データの構造化という一連のプロセスも体験する必要があります。さらにデータの出处を検討したり、それがデータから導かれる結論にどのような情報を与えるのか、についても学ぶ必要があります。このようにデータサイエンスは必然的に高度に経験的な学問であり、実践での経験から開発されていった技術を含んでいます。学生は、実地で現実世界の応用例と常に向き合う必要があり、その過程で実データに基礎的なアルゴリズムやモデルを適用します。ところで数学の学部カリキュラム委員会 ([MAA 2004, 2015](#)) は全ての数理科学専攻の学生にとって、実データを解析することの重要性を強調しています。彼らは、以下のように述べています。

データ解析により、学生は抽象化とデータの生じている現実とを行き来する経験が出来るわけです。このことは数理科学専攻の学生にとって決定的に重要です。そのようなデータ解析の経験は、卒業後すぐに産業界で働く学生、関連する学問分野に興味を持つ学生、中学校高校の数学教師となる準備をする学生にとっても、重要です。

実は統計学者も同じことを考えています。[2002 年の ASA の報告書](#)では、学部の統計学のカリキュラムでは、データ解析に重点を置く（さらに「データ解析」において「解析」よりも「データ」に重点）ことを推奨しました。同じロジックで、学生はデータサイエンスを試行錯誤で課題をこなしながら学んでいくのです。既に述べたデータサイクルを何度も繰り返すことが、データサイエンスの学びの特徴的な要素なのです。また、グルー

プ学習及びプレゼンテーションを含む実習がカリキュラムの至るところで扱われることが必要でしょう。さらに、卒業実習も学部の中に経験すべき重要な要素であり、インターンシップは、データサイエンスプログラムに必然的にフィットするでしょう。

2.4 解析的（計算論的・統計的）思考

Breiman (2001) は、データから何らかの判断を下す場合に2つの学派があることを指摘しました。2つの学派とは、「algorithmic (computational) model」と「data (statistical) model」のそれぞれを金科玉条とする学派です。ただし、Donoho 2015 は、それぞれ「predictive model」と「inferential model」と改名した方が実態を反映しているとししました。データサイエンスは、課題解決のために、そのような計算論的思考、統計的思考のうち一方を他方より強調するということではなく、両方を統合的に使う学問です。興味深いことに、これらは新しいアイデアではありません。過去半世紀において、多くの統計学者は「data (statistical) model」をより好んでいたのですが、Wilkinson (2008) が指摘するように、確立された優れた方法論をいくつも提案した Tukey は、「algorithmic (computational) model」も等しく統計学において重要なモデルであると認識していました。計算論的思考、統計的思考という2本の柱は別々に教えられるべきではないのです。科目によって両者の重みが多少ブレることがあっても、両者が必ず共存してこそ、効率的・効果的な授業と言えるのです。

2.5 数学基礎

データサイエンティストは、モデルを用いて現実世界を理解しようとし、また数学はそのモデル化のための共通言語です。従って、データサイエンティストとして活躍するには、確固とした数学の基礎を身につける必要があります。一方で、伝統的な数学のカリキュラムは、「抽象的な数学」と「興味深いデータを含んでいたとしても、扱いにくく、時に数学的にきちんと設定されていない現実の問題」の繋がりについて学ぶことを重視しません。我々は、学部レベルのデータサイエンスに本質的な必要な数学だけを抽出するという新たなアプローチを提案します。数学において4科目以上を課すことなく、データ駆動型問題をモデリングするために2つの数学科目だけでデータサイエンスに本質的な必要な数学を学べるようにします。つまり伝統的な数学で重視される理論・導出・証明よりもデータサイエンスにフォーカスして、数学カリキュラムを効率化することです。とくに数理的な開発ツールのための動機付けとして、(アルゴリズム的及び統計的な) モデリングを提案します。それらは、現実世界の問題を解くために必要になるわけで、そのコンセプトを導入します。また学生に対する動機付けとして、線形代数は線形システムを解くため、微分は最適化問題や sensitivity analysis の理解のため、積分は期待値など確率論における応用問題、という風に位置づけることが出来ます。このようにモデリングやアプリケーションへのシフトは微積分学の最近のトレンドと一致していますが、我々のアプローチは、(理論的な) 数学へのアプローチ、各トピックを学ぶ順番、学ぶべきトピックの取捨選択において画期的と言えます。

もちろん、我々のアプローチを数学専攻では取り得ないという意見もあるでしょう。一方で、我々のアプローチによって数学を学んだ学生が(より理論的な) 数学を勉強したいと考えるかもしれないのです。そのような学生が、既存の数理科学の修士課程に進みたいのであれば、いくつかの理論的な科目を追加的に取る必要があるでしょう。

2.6 柔軟性

我々は、現段階で存在しない新たなテクニックや方法論を自ら学ぶように、学生に覚悟させる必要があります。彼らは、将来において非常に多様なデータを扱いながら仕事をしていく必要があります、さもないと将来において職場で必要とされなくなります。データサイエンスプログラムのコアとしている数学的・計算論的・統計的な思考能力と演習が、(時代の要請に合うように) 実用的で重要なデータサイエンスのスキルを含んでいるのか、我々は注意を払わねばなりません。データサイエンスは、全てのレベルにおいて、急速に進化・変化を続けています。多くの大学においては、既存の学問分野における現行の科目によってデータサイエンス専攻を運営していくでしょう。将来的には **付録 A** で説明するような、より統合的な科目に移行するケースもあるかもしれません。大学が現在及び将来の学生のニーズを満たすべくコース設計をする場合にこのカリキュラム指針が役に立つことを、我々は願っています。データサイエンスは急速に変化している分野ですので、定期的にプログラムを見直し、必要に応じて新たに再編することが望ましいでしょう。

3 データサイエンス専攻で身に付ける能力

データサイエンス専攻の卒業生は研究開発の全ての段階において、データと関わり共に思考する準備ができていなければなりません。また、研究開発チームの一員として活躍することが期待されています。データサイエンスの学部卒業生に期待される能力は以下の枠の通りです。

学部レベルのデータサイエンス専攻で重要な能力

- 計算論的、統計的思考能力
- 数学基礎
- モデリングとその性能評価
- アルゴリズムとソフトウェア基礎
- データキュレーション
- 知見を他者に伝える能力 – コミュニケーション能力と責任感

3.1 解析的（計算論的・統計的）思考

データサイエンスは、データから知見が導き出されるような実践的な状況に於いて用いられる課題解決のためのアプローチです。このアプローチは統計学、計算機科学、数学における思考方法を合わせたものと言えます。

3.1.1 大量データ環境下での統計的思考

統計的思考は、データを通じて世界を理解するためのアプローチであり、「課題設定から結論まで」全てに関与します (Wild & Pfannkuch 1999, p. 1 による)。データサイエンティストは、基礎的な統計理論を理解していなければなりません。学生は「データ解析」、「データ収集」、「データモデリング」、及び「統計的推測」の基礎的な統計概念を理解するべきなのです。基礎的な理論をきちんと理解すれば、自分たちのモデルに基づく解析結果及びそのモデルの限界がわかります。理想的な卒業生には、「問題設定」、「新規データ取得のため

のプランニング』あるいは『既存の関連するデータの収集』，そして「知見を得るための解析」のために統計学の知識及び計算機科学の技能を使えることが期待されます。

3.1.2 計算論的思考

データを扱う場合には，計算機に関する幅広いスキルが必要となります。データサイエンスの学生は，職場や研究室にあるデータに対して，様々なスキルを用いて仕事をするようになります。例えば，データベースのデータにアクセス，データの整理・管理，ウェブサイトからのデータの抽出，テキストデータの解析可能なフォーマットへの処理，秘匿性の高いデータの安全の担保，これら全てにおいて，計算機に関する幅広いスキルが必要となるわけです。このような課題解決のための計算機的スキルは，データサイエンティストのワークフローにおいて繰り返し必要となります。Wing (2006, p. 2) が述べているように，「コンピュータサイエンティスト的思考は単に『プログラミングが出来る』ことだけを意味するのではないのです。多層にも及ぶ抽象化をしながら思考することが必要なのです。」データサイエンスの卒業生は，多くの基礎的なソフトウェアの技能，計算機科学分野におけるアルゴリズム的・計算機な課題解決方法に習熟している必要があります。データサイエンティストとしてのキャリアを積んでいくためには，さらに，統計解析ソフトウェアパッケージにも習熟して使いこなせる必要があります。また，そのようなソフトウェアパッケージの基礎にあるプログラミング原理及びアルゴリズム的な課題解決方法をきちんと理解することも必要です。

3.1.3 様々なアプローチを融合統合する能力

学部レベルのデータサイエンスは，確固とした数学的基礎に裏付けられた計算論的・統計学的思考の集合体です。以下に述べる能力に加えて，データサイエンスの卒業生は，既存の学問分野の概念の関係を理解して説明できなければなりません。彼らは，どんな問題に対しても対処できるよう様々なスキルや課題解決アプローチをきちんと把握して，また与えられた状況に於いてどのスキルが適しているのか判断できなくてはなりません。多種多様なツールを扱える，また新しいツールを自力で学べる，必要があれば新しいツールの開発にも貢献できるような器用さ，柔軟さを備えていなければなりません。計算機環境はソフトウェア，ハードウェア両面において，急速かつ頻繁に変化します。そして，それらの変化はデータ構造，データストレージ，計算効率の改善に及びます。データサイエンティストは，そのような変化にスムーズに対応できなければなりません。また，計算機的な困難，モデリングの困難，また2つの困難がどのように関連しているかも理解する必要があります。例えば，ある特定のデータセットを特定のモデルにフィッティングしようとしたときに，計算機的困難が新たに生じたとします。そのときに，データサイエンティストには，課題解決のためにモデリングの調整・改良，あるいは計算機環境の改良，あるいはその両方を実行できる腕前がなければなりません。データサイエンティストには，クライアントに解を提示し，多くの現代的な統計的モデリング方法を理解し，科学的な再現性を達成する能力が必要です。統計的思考に裏打ちされることにより，計算論的思考は，データサイエンティストのそのような能力を大きく向上させるのです。

3.2 数学基礎

学部のデータサイエンス専攻において，数学的側面で重点が置かれるべきなのは，数理モデルの選択，数理モデルへのフィッティング，数理モデルを使いこなす能力の育成です。データ駆動型問題が，数学的に整理されておらず，また問題へのアプローチが数学的に一見明快でないというのはよくあることです。従って，学生は体系化された数理的な問題解決スキルを身につけることによって，そのような問題に対しても数学的構造

を見出せるようになる必要があります。さらに、「統計的学習や機械学習で典型的に用いられるモデルに共通する数理的構造」及び「関連するアルゴリズムの最適化と収束性」を理解するのに十分な数学能力を身につける必要があります。そのための必要なツールは解析学、線形代数、確率論、離散数学といった数学科目に含まれます。ただし、数学以外にも学ぶべき事柄が多いデータサイエンスプログラムの学生たちの学習時間の確保のために、それらの科目におけるトピックの抜本的な配置換えや簡略化を念頭に置いています。

3.3 モデル構築とその評価

3.3.1 Informal modeling

統計モデルは、プロセスの記述、予測、説明のために用いられるだけでなく、知見を他者と共有するためや、将来のデータに対するモデルの基礎としても使われます。Informal modeling は、変動の潜在的な要因の特定、確率的変動・決定論的変動の識別、それらが数学的にまた計算機上でどのようにモデル化されるかへの理解、などを含んでいます。データサイエンスの卒業生はデータの可視化にも熟練しなければなりません、何故ならそれが Informal modeling の重要なツールであり、他人と意思疎通を取り、また提案されたモデルの弱点を特定するのに使われるからです。

3.3.2 Formal modeling

データサイエンスの卒業生は、統計モデル・機械学習モデルの構築、その性能の評価、種々の formal inference procedures の利用、解析からの適切な結論の導出、それらを全て実行できなくてはなりません。例えば、データ収集方法、データに存在する偏りやばらつきなどの問題が解析結果、その解釈、さらに統計的知見の総括にどのように影響を与えるのかを理解することです。卒業生は、特にビッグデータのスケールの視点からデータ解析が可能である計算機環境を理解できなければなりません。

3.4 アルゴリズムとソフトウェア基礎

データサイエンスの卒業生は、アルゴリズム的な課題解決が不可欠な業務に従事することになります。その業務としては、課題・問題に対して、明確に要件定義し、問題を分解し、アルゴリズム的な解法に至る効率的な戦略を使い、さらに適切な高水準言語によるプログラミングにより、解法を実装するということです。卒業生は、実装したプログラムの構造とソフトウェア、及びそれらが使用するライブラリやパッケージのメモリと実行パフォーマンスを理解する必要があります。データサイエンスの卒業生は、プログラムのドキュメンテーションやプログラムの構造において優れた実践を知り、活用し、ソフトウェアを維持するための適切なツールを使用できるようになるべきです。また既存のパッケージやツールを活用して、計算上の課題を解決出来なくてはなりません。

3.5 データキュレーション

データキュレーションとは、課題解決のプロセス全体におけるデータ管理に関することです。2つの主要な段階があります。

3.5.1 データの前処理

卒業生は、様々な出処やフォーマットのデータを扱える必要があります。ウェブページやデータベース、動画のストリームから取得されるデータもあります。また画像、音声、動画、数値、テキストなどその形式は様々です。十分に制御された実験におけるデータもあれば、人為的、能動的な介入を伴わない観察データ、あるいはセンサーなどの自動装置から偶然観測されたような、計画的な記録方法に基づかないデータもあります。一旦データセットが与えられたら、卒業生は様々な統計手法やモデリングが適用可能な形式にデータを前処理出来なくてはなりません。また、データの品質や取得手段が結論に及ぼす影響を認識することも重要です。

3.5.2 データ管理

データサイエンティストは、解析のためにデータを前処理するだけではなく、解析の全ての段階でデータを完全に管理できなくてはなりません。このために、SQLのような関係データベースを用いての作業、バージョン管理、複数のソースからデータがマージされる時のデータの出処のトラッキングが必要です。

3.6 知見の伝達

データは孤立しては存在せず、なんらかの特定の文脈から生じます。そのような文脈の背景を知っておくことがデータを解析する上では必要です。従って学部生は統計学、計算機科学、数学というデータサイエンスのコアの外でデータサイエンスの知見を活用する経験が必要です。データサイエンスプログラムは、物理科学、生命科学、ビジネス、社会科学、人文学、美術のような分野でデータに基づく調査、研究をすることにも重点をおきます。学部生が訓練すべき重要な能力が2つあります。

3.6.1 コミュニケーション能力

専門知識を分かりやすく伝えることを含め他者と上手にコミュニケーションを取ることは、データサイエンティストの中核的なスキルです。研究開発チームの一員として、データサイエンティストはグループメンバーだけでなく、プロジェクトの詳細について必ずしも詳しくない人にも、分かりやすく伝える必要があります。静的及び動的なデータ可視化手法で一般の方に分かりやすく伝えることも、データサイエンティストの素養として今後ますます必要となります。データサイエンスプログラムは、そのようなコミュニケーション能力を育成する機会を随所に取り入れるように配慮されなければなりません。また、コミュニケーション能力育成のために特別なクラスを設けるというのではなく、カリキュラム全体を通じて身につけていくことが望ましいです。学生は、口頭、文書、可視化手法などにより、様々なレベルの相手に対して分かりやすく伝える経験をすべきなのです。

3.6.2 研究倫理と再現性

データサイエンスの興隆により、新たな倫理的問題も出てきます。データサイエンスプログラムでは、引用、データ所有権、データの安全性・機密性、データ分析の結果とプライバシーの問題、研究開発の透明性と再現性などについて、起こり得る不正行為と倫理的トレーニングを考慮しなければなりません。

4 データサイエンス専攻のカリキュラム内容

我々のカリキュラムの目的は、データから知見が得られるような一連のサイクルの中に学生を引き込んで、これまでの節で述べてきたようなスキルを学生が習得出来るように手伝えることです。そのためには、関連するはずのトピックや学問分野が伝統的には積極的に繋がりを持つとしなかった状態を繋げて融合することが必要となります。我々の提案するデータサイエンスプログラムは、以下で説明するように 6 つの主要分野からなっています。我々はそれらの 6 つの分野を学生に学ばせる科目を提案します。なお **付録 A** では、科目体系をより詳しく説明します。以下では、これらの分野のために設計された科目の要約を説明します。

データサイエンス専攻の 6 つの主題目

- データ記述とデータキュレーション
- 数学基礎
- 計算論的思考
- 統計的思考
- データモデリング
- 他者との知見の共有、再現性、倫理

データサイエンス専攻の科目概略

1. データサイエンス入門
 - データサイエンス入門 I
 - データサイエンス入門 II
2. 数学基礎
 - データサイエンスのための数学 I
 - データサイエンスのための数学 II
3. 計算論的思考
 - アルゴリズム・ソフトウェア基礎
 - データキュレーション データベースとデータ管理
4. 統計的思考
 - 統計的モデリング基礎
 - 統計的学習・機械学習
5. 他専攻の科目
6. 卒業実習

4.1 科目体系の概観

データサイエンス入門 I,II

学生は、探索的にデータに触れ、データを可視化し、データに由来する研究課題をもたらすために高水準言語を使います。第 2 学期においては、よりアルゴリズム的な言語を習得します。それによって第 1 学期に学んだ高水準言語の関数の背後にある考え方とその構造を理解できるようになります。

- 高水準言語入門
- データの扱いと基本的な処理
- 関数と基本的なコーディング
- モデリングの基礎：決定論的モデリング，確率論的モデリング
- プロジェクトとコードの管理
- データベース
- データ収集の基礎，統計的推測の基礎

数学基礎 I,II

データサイエンスの学生は，数学的な道具で現実世界の問題にアプローチします．純粋数学では理論を構築し，命題を証明します．データサイエンスでは，そのような純粋数学と違って数理的手法の価値をその限界を理解しつつ追求します．またデータサイエンスの学生は，数理的トレーニングを通じて幾何的・直感的・視覚的な思考を身につけなければなりません．我々は，解析学 1 を履修するレベルの学生にデータサイエンスのための 2 学期に渡る数学科目を提案します．この 2 科目においては，数理モデリング，特に線形モデル，多項式モデルを重視します．（詳しくは [付録 A](#) を参照して下さい）また以下のトピックスを扱います．

- 数学的構造（関数，集合，関係，論理）
- 線形モデリングと行列計算（行列代数，行列分解，固有値，固有ベクトル，射影，最小二乗法）
- 最適化（微分に関連する解析学の概念）
- 多変量（多変数関数の微分・積分の概念，その数値計算）
- 確率，確率モデル（数え上げ原理，1 変量確率分布，多変量確率分布，独立性，計算機によるシミュレーション）

アルゴリズム・ソフトウェア基礎

確固とした計算機科学の能力をつけるために，学生は計算機科学の基本的なトピックを学習し，アルゴリズム的な課題解決能力，及びソフトウェア開発，プログラミング能力を身につけます．

- アルゴリズム・デザイン：学生は，課題を理解し，扱いやすい要素に分解し，色々な課題解決法を検討しつつ，効率的に課題解決するアルゴリズムに到達する，そのような一連の作業が出来る能力を身につけます．
- プログラムの概念とデータ構造：学生は「手続き型プログラミング言語，関数型プログラミング言語の両方で」「リスト，ベクトル，データフレーム，ディクショナリ，ツリー，グラフなどのデータ構造を用いた」アルゴリズムを実装出来なくてはなりません．
- ツールと環境：学生は，利用可能なツールやパッケージを適切に使えなければなりません．そのようなパッケージにより「データサービスにアクセスしてプログラムにより入出力」，「データ変換，データ探索，データ可視化，及びデータ解析」そして「開発環境や，バージョン管理とトラッキングのためのツールを含むソフトウェアの開発と管理」が可能になります．
- ビッグデータの処理：データやデータサイエンスに関連する処理の規模が大きくなり続けているので，学生はビッグデータを扱う能力を身につけなければなりません．データの並列計算を実行するシステム構築をするためのプログラミング能力が必要です．さらに既に述べたデータマネジメントの一部として

分散ファイルシステムの現行形式や新しい形式を扱えることも必要です。またストリーミングデータの取り扱いも知っていなければなりません。

データキュレーション－データベースとデータ管理

学生はデータマネジメントの原理を理解して応用できなければなりません。これは伝統的なデータベース管理よりも格段に広範であって、ビッグデータに起因するデータ量、データの生成される速度をサポートするシステムも含みます。従って、データサイエンスの学生はデータ問い合わせ言語を身につけて関係データベースやSQLだけではなく新たな巨大データシステム（NoSQL）を扱える必要があります。また、ウェブサービスを通じたあまり構造化されていないシステムからのデータや、インターネットでのより低レベルなデータへのアクセス、そしてストリームから供給されるデータも利用できなければなりません。データが収集された後は、既に述べた知識、スキルを用いてデータクリーニング、前処理を行うこともデータマネジメントに含まれます。さらに、データ探索、データ可視化、データ解析が出来るようにデータを構造化された形式に変換することもデータマネジメントの一部です。

統計的モデリング基礎

統計的データ解析の導入及び統計的推測の枠組みを導入する科目です。線形モデルを基礎としますが、非線形アプローチも扱います。1年目のデータサイエンスの科目では全ての統計解析の基礎となる重要な概念が導入されるわけですが、この科目ではそれを足場に統計モデルを学びます。学習する全ての主題、概念は、実社会の実データに根付いており、またそのようなデータからインスパイアされているのです。

- 探索的データ解析，視覚的なデータ解析法
- 統計的推定，統計的検定，中心極限定理，大数の法則などの統計手法，ブートストラップ・リサンプリングなどのアルゴリズム的手法，統計手法・アルゴリズム的手法をもとにした点推定・信頼区間・仮説検定，尤度原理，ベイズ法
- シミュレーションとリサンプリング，確率システム下でのモンテカルロ・シミュレーション，ブートストラップ，ジャックナイフ，パーミュテーションなどのリサンプリングを用いた推測手法，研究・調査・実験の計画の基礎知識（無作為割当，任意抽出，データ収集，効率性），因果推論におけるバイアス，因果，交絡，偶然の一致など
- モデリング入門：単回帰モデル，重回帰，一般化線形モデル，アルゴリズム的なモデル（回帰木や k 近隣法），教師なし学習（クラスタリング）
- モデル選択とその効果：正則化法，ケチの原理，バイアス・バリエーションのトレードオフ，損失関数とモデル選択（交差検証法，罰則付き回帰，リッジ回帰）

統計的学習・機械学習

この科目には、計算機科学での機械学習に見られるアルゴリズム的視点と統計的思考の予測的な視点の両方が混ざっていてその融合を目指します。機械学習の一般的な方法論及びその様々な分野の問題への適用に焦点を当てます。学生は単に統計的学習の理論を理解するだけではなく、科学や産業界で生じる新しい課題に適切に適用するために必要となる実用的なスキルも身に付けることになります。

- 伝統的な回帰分析や（線形）判別にとって代わる手法を深く探求する

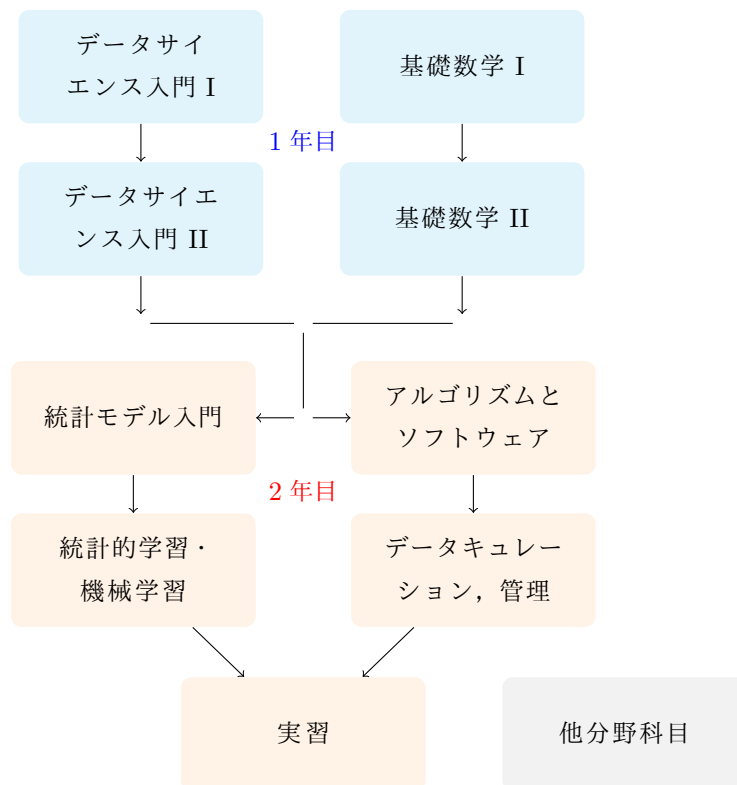


図 1 データサイエンス専攻における履修パスの一例のフローチャート

- モデルのアルゴリズム的な解析。スケーラビリティや実行可能性の問題にも触れる
- 性能測定指標，予測，交差検証法
- データ変換：新たな変数や特徴の作成，主成分分析などの次元削減法，スムージング手法，アグリゲーション，教師あり学習と教師なし学習
- ブースティング，バギング，モデル・アベレージングなどのアンサンブル法

実データを扱う卒業実習

卒業実習では，学生は科学的な研究課題を考えデータを集め分析して，結果を発表します。

科目をどのような順で学ぶかについて一例を図 1 に示します。

5 その他の検討事項

5.1 大学院

大学院で数学，統計学，計算機科学を学びたい学生は，理論に重きを置いた上級科目を履修したほうが良いでしょう。特に，データサイエンスのための数学科目は数学，統計学，計算機科学で博士課程に進学したい学生には十分ではありません。

5.2 コミュニティカレッジからの転入

コミュニティカレッジは学生の多様な目的に合わせようとするために、既存の制度を急には変えません。当面は、既存の科目体系のもとで、学生はカレッジや大学のデータサイエンス専攻に転学する準備をすることになります。日本では該当するケースがないため省略します。

5.3 高校で学んでおくべき概念

全国統一学習到達度テスト（各州共通基礎スタンダード）で統計学を受ける学生もいるので、いくつかの初等的な内容は見直すべきかもしれません。データサイエンスのカリキュラムの準備として、

- 微積分が出来るようになること（微積分を学ぶための前提科目の履修）
- 基礎的な行列代数を理解すること（連立方程式が解ける）
- 散布図、相関の概念、単回帰における直線当てはめを通じて複数の変数間の関係を理解すること
- 分布の中心、散らばり、分布の非対称性などの記述統計の理解をすること

が挙げられます。

5.4 インターンシップと実データを扱う経験

インターンシップや実データを扱う経験も、データサイエンスプログラムにおいて重要です。実データを扱うプロジェクトは、カリキュラム全体を通して頻繁にその機会があるべきであって、また卒業実習（実践的教育プログラムの総仕上げ科目）において特に重要です。

6 既存の科目によるデータサイエンス専攻への移行

我々がこれまで説明してきたのは、3つの学問分野（数学、計算機科学、統計学）にまたがる広範なトピックを融合したカリキュラムです。それらのトピックの多くは、各学問分野の伝統的な科目でカバーされます。以下に太字で記された10科目で、データサイエンスに必要な基本的なスキルを最低限カバーします。ただし、それらを履修するだけではデータサイエンスにおける分野交流の深みには触れることが出来ないのです。

6.1 数学の科目

データサイエンス専攻に最も役に立つ数学科目にするには、「数学」と「実社会の課題・問題、『特にデータが得られることによって導かれる課題』」とが繋がっていることを数学科目を通じて学生に分からせることが重要です。

- 微積分 1
- 微積分 2
- 微積分 3
- 線形代数
- 確率論

- 離散数学

6.2 計算機科学の科目

計算科学的思考のコンテンツは複数の伝統的な計算機科学科目にまたがっています。ただし、下の太字 3 科目にそれらが全て含まれているわけではありません。

- 計算機科学入門
- 計算機科学 2: データ構造, アルゴリズム
- 計算機システムとアーキテクチャ
- 上級アルゴリズム
- データベース
- ソフトウェア工学

6.3 統計学の科目

特に「Introduction to Statistics」の内容は[カレッジにおける統計教育・評価のための指針改訂版](#)に従わなくてはなりません。

- 統計入門
- 統計モデリング, 回帰分析
- 機械学習, データマイニング
- 統計理論 (確率論)

6.4 関連する科目

- 関連分野の入門科目
- 関連分野の中級科目
- データサイエンスの卒業実習, 実践的教育プロジェクト
- ライティング 2 科目 (うち 1 科目はテクニカル・ライティングが望ましい)
- 話術
- 研究倫理

太字にした 10 科目は、データサイエンス専攻に必要最低限の内容をカバーします。ただし、多くのプログラムにおいてはいくつかの科目を追加して内容や実習を充実させたほうが良いでしょう。データサイエンスのために新たに設計した 10 科目に基づくデータサイエンス専攻については、[付録 A](#) をご覧ください。

7 まとめと検討事項

本稿で示したカリキュラム指針が、新しいプログラムを設置する場合や既存のプログラムを見直す場合の議論の叩き台になれば幸いです。

要点

我々の主張の要点をまとめます。

1. データサイエンスは、急速に進化している学問分野であり、データ取得、データキュレーション、データ解析における中心的な役割を果たす学問分野です。
2. 現状では数学、統計学、計算機科学という伝統的な学問分野がデータサイエンス専攻の基盤となっています。
3. 様々なレベルにおいて数学基礎、計算論的思考、統計的思考を統合するようなカリキュラムの再設計により、データサイエンスでキャリアを積んでいく卒業生の役に立つ充実し効率的な科目群を提供できます。

この分野は急速に進化していますが、この文書で説明してきた基礎的な分野は依然として必要とされていると考えます。ところで、我々のミーティングにおいて認識していたが扱えなかった話題として以下の事柄があります。

今後の課題

ファカルティ・デベロップメント 付録 A で説明した科目群は、新しいデータサイエンスのプログラムに向けて大胆に設計しています。効果を発揮するには、多くの大学で繰り返し実施してもらうことが必須です。講義のために講義ノート、実例、ケーススタディ、(また最も重要かもしれませんが) 新たな教科書といったリソースが不可欠です。

2 年制カレッジや高校との連携 データサイエンス専攻は、高校からあるいは 2 年制のカレッジから進学する多くの学生にとって魅力的だと考えられるでしょう。そのような機関と連携して、科目や教育方法を調整、改善することが不可欠です。それにより、4 年制大学への転学も促進されるでしょう。

定期的な見直し 急速に進化するデータサイエンスにおいて具体的なカリキュラム指針の提案は初めての試みです。データサイエンス分野の進化に応じて改訂版が必要となると考えています。フィードバックを頂ければ幸いです。

付録 A Appendix : 提案する科目の詳細

A.1 データサイエンス入門 I

ねらい

終始一貫したデータサイエンスの基礎。学生はデータを扱う一連の作業全てに関与すること。いくつかはグループワークとなる。この授業はデータサイエンス領域の高水準の入門授業であり、予備調査から、データ取得、最終的に得られる結果の効果的なプレゼン方法までの一連のサイクルを繰り返しながら、授業が進むことが期待される。

学習目標

- 必要なデータの調査とデータ収集
- 基本的な関数、コーディング

- データの要約統計量，データ可視化，データ分析
- 決定論的及び確率的な現象のモデリングとシミュレーション
- 論文形式，口頭，ポスター形式での完成したプロジェクトの結果のプレゼンテーション

内容

- データとは何か？
 - － データはどのような形式があるか？データはどのような問いに答えられるか？
 - * 記述，予測，推測
 - * 実装：データはある程度複雑だが，（欠測値や外れ値がなく）比較的扱いやすいものを用いるべきだ．この段階ではデータを学生に（収集させず）与える．また変数の数，サンプルサイズ，複数のデータの表を与えると行った観点から十分大きいデータを扱う．
- 高水準言語の統合開発環境入門（R を推奨）
- データ記述：データの探索的データ解析，データの可視化
 - － 要約統計量，アグリゲーション，平滑化，データの分布
- モデリングと確率（不確実性，シミュレーション，（擬似）乱数生成などの理解）
 - － 数理モデルや関数（線形関数，指数関数など）の概念，プログラミングの概念
 - － ベクトル，表およびデータフレーム，変数
 - － 逐次プログラミング，スクリプト記述
 - － 簡単な関数，簡単な環境，スコーピング規則の定義
 - － 条件式と基本的な繰り返し
- データを用いた（あるいは用いない）シミュレーション：確率的シミュレーション，リサンプリングを用いたシミュレーション
- アルゴリズム
 - － 複雑な問題の小さなステップへの分割
 - － レシピとしてのアルゴリズムの概念
- 実習
 - － コミュニケーション：レポート，口頭発表，グラフの視覚的な効果による発表について学生を評価する
 - － 学問の倫理：コードをコピーすることは剽窃など
 - － 共同作業：GitHubなどで，グループによるプロジェクトの遂行．プロジェクトとコード管理の概念
 - － 応用分野に動機付けされた実習

A.2 データサイエンス入門 II

ねらい

様々なデータ型とデータソースに慣れること，分析に適した形式に変換するためにデータ・キュレーションの過程を学ぶ．推定，予測，推論における基本概念の導入．計算力や分析力を向上させるために，あまり手入れされていないデータを含むケーススタディを通じてこの科目を教えること．

学習目標

- 関係データベースを含む様々なデータソースに触れる
- 様々なインターフェースからのデータへのアクセス
- 種々のデータ形式からの、データ解析を可能にする構造構築
- 課題を定式化して、推定、予測、推測の基礎概念に繋げる
- データ収集プロセスが推測出来る範囲にどのように影響するかの理解

内容

- 種々のデータ：静的データ，空間データ，時系列データ，テキストデータ，メディアからのデータ
- データソース：関係データベース，web，API，ストリーミング
- データ収集：サンプリング，実験デザイン（観察データ，実験データ）と結果の可視化，モデリングと結果の一般化可能性
- データクリーニング，データ抽出：データの解析可能な形への変換，正規表現，SQL 文
- データ分析，モデリング
 - － 探索的データ解析を通じた課題・問題の定式化
 - － 推定と推測（統計的検定と信頼区間）の入門．シミュレーションとリサンプリングを含む
 - － 推測の可能な範囲
 - － 評価と選択，例えば訓練データとテストデータ
- 実習：包括的なデータ駆動型の実習．課題定義，データの扱いやすい形式への変換，初等的な解析と結論まで．結論の一般化の限界も分からせる
 - － データ収集能力の向上
 - － 解析のための優れたツール，推測の可能な範囲
 - － 結果の共有，チームワーク，研究倫理，プレゼンテーション

A.3 基礎数学 I

ねらい

データサイエンスの学生は、数理モデル、統計モデルを使いこなせる必要があり、そのためには数学の基礎を理解していなければならない。この科目では、線形モデリング、最適化に由来する基礎概念を学び、様々な応用事例から動機づけされる数学の基礎も学ぶことになっている。定理の証明や必要以上の手計算の問題演習よりも、数学を使いこなす、また数学と実社会の概念を関係づけることを重視する。数値計算を通じた具体化によって、数学的な概念がより深く理解出来る。

学習目標

- 一次式，多項式，指数関数などを用いて現実世界の現象をモデル化
- 関数の最適化と結果の解釈
- 線形モデリングと最適化の幾何的理解を養う
- 関数やモデルの局所的振る舞いを理解するための道具（微分）
- 多変量モデルを分析するための行列代数の学習
- インプットの微小変化がモデルに与える影響の理解（感度分析）
- モデル選択の導入．モデルの限界を理解する
- 多様な応用例，多様なツールを用いたモデルの生成

内容

- モデリング入門：関数を使った簡単なモデル（物理の例などを用いる）
- 線形代数と幾何：ベクトル，内積，距離，射影，行列代数
- 多変数を扱う：超平面，連立方程式
- インプットに対するモデルの感度：直線の傾き，超平面の断面，曲線の割線や接線
- 連立方程式の解
 - － 体積測定としての行列式の幾何的アプローチ
 - － 行階段形と幾何的アプローチ
 - － 特異行列，解が存在しないこと，解が存在するための条件
- 微分
 - － 線形近似
 - － データサイエンスの文脈における最適化と回帰
 - － 最急降下法
- 内積を通じた相関係数への幾何的アプローチ
- 行列分解（時間があれば）
 - － LU 分解と QR 分解
 - － 線形変換の幾何的解釈
 - － 固有値，固有ベクトル，行列の対角化
 - － 特異値分解
 - － 主成分分析

A.4 基礎数学 II

ねらい

確率論，また確率分布に繋がるデータ生成過程を学ぶ．経験分布から確率分布，ヒストグラムから密度関数の構成など．確率的なアイデアや離散データから連続関数への移行のプロセスから積分を導入して学ぶ．定理の証明や必要以上の手計算の問題演習よりも，可視化，数値計算を通じた具体化によって，数学的な概念をより明確にさせることを重視する．

学習目標

- データ生成過程におけるランダムネスの理解
- 具体的な例題を解くことで確率の性質やルールを学ぶ
- 基礎的なシミュレーションによりデータを生成して確率を学ぶ
- 確率論における曲線による面積の解釈（連続確率変数の区間の確率 $\Pr(a < X < b)$ ）
- 期待値や分散を含む確率分布の性質を扱えるようになる
- 一次元の確率概念を多次元に拡張する

内容

- ランダムネス，確率変数，経験分布関数
- 確率の概念：標本空間，事象，集合，数え上げ，ベン図
- 確率の性質，大数の法則を体験するためのシミュレーション

- 条件付き確率とベイズの公式
- 二項分布，幾何分布，正規分布．他は，上級科目で必要になった時に学ぶ
- 積分
 - － 確率，期待値，分散に動機づけられた積分論を学ぶ（離散確率変数から連続確率変数へ）
 - － 微積分の基本定理：確率密度関数と累積確率関数の関係
 - － 数値積分とその誤差概念を理解する
- シミュレーションを通じた標本分布と中心極限定理の理解
- 同時確率分布，多重積分，周辺分布
- 級数と級数の離散確率分布における役割

A.5 アルゴリズムとソフトウェア

ねらい

データサイエンス I,II で学んだ計算論的課題解決の理解をさらに深める．データ中心主義を継続し，手続き型プログラミングを使用しながら課題やテクニックの足場を組み，標準的な問題解決戦略を追求し，効率性を評価する．計算機科学の概念を学習するための手段として，統計的アルゴリズムとデータメソッドを活用する．

学習目標

- 課題解決のためのプログラムを完全を書く
- 第二プログラミング言語の習得（Python, C++, Java）
- 新たな IDE（Eclipse など）を使ってみる
- プログラミング言語のより深い理解：変数のスコープ，データの種類，パラメタの受渡しの概念など
- 基礎的な統計アルゴリズムの実装とその効率性を学ぶ（抽象的な視点と低水準な視点の両方）

内容

- データサイエンスにおいて有用な基礎的データ構造の利用
 - － リスト（関数アプリケーション、統計的性質の集約）
 - － スタック（逐次的データ変換のエンコーディング）
 - － キュー（到着プロセスのシミュレーション）
 - － ディクショナリ（グラフ表現，Bag-of-Words の実装）
 - － 行列（因数分解アルゴリズムの実装，グルーピング/集計の実装）
 - － ツリー（再帰的に分割された決定木の実装）
 - － 上記のデータ構造に適用される操作の効率性
 - － 単一ノードを超えた分散を含む，さまざまなデータ構造のスケラビリティ
- アルゴリズムの複雑さ
 - － Big-O 記法，実際の計算効率と理論的な計算効率
 - － 総当たり攻撃と分割統治法
- ソフトウェアの開発プロセス
 - － 正しさ，評価，デバッグ

- コードの可読性, 文書化
- プロジェクトとコード管理 – バージョン管理
- プログラミング言語の比較
 - データサイエンス I,II で学んだ関数プログラミングと, 命令的プログラミング: 証明可能な正当性の価値, 操作をリスト全体にマッピングする能力, システム設計との関連性, 倫理的意義
 - オブジェクト指向プログラミングの基礎: 抽象データ型を使用したソフトウェア設計
- 高性能コンピューティング
 - アルゴリズムのパラダイム (例えば, マップ・リダクション, オンライン/ストリーミング)
 - システム (Hadoop, Spark, NoSQL など)

A.6 データキュレーション, 管理

ねらい

データサイエンス II の学習内容を土台として, データを単一および複数のソースから広範囲に渡って収集しながら, 様々なデータ構造に変換, そして実データを取り込むことによってデータ・キュレーションおよび管理に関する学生の知識を深める. 取得したデータを解析に適した形式に変換する方法を理解する. ビッグデータに対応するためにデータ管理の概念とスキルを拡張する.

学習目標

- 様々な外部システムからのデータ取得: 外部システムとしては, 定義されたスキーマを持つ適切に構成されたシステムから, 十分に構造化されていない API を介したアクセスやスクレイピング/マニピュレーションを要するシステム, 大規模分散ストレージシステムを含む.
 - 複数のソースからのデータ取得. それらの管理統合.
 - 学び方の学習: 新しい API や新しいシステムへの一般化. それらからのデータ取得を可能にする.
 - データソースに固有の統計的な問題 (偏りやランダムネス, 欠測など) と, 解析や一般化におけるそれらの問題の影響を理解すること.
- データ探索, 可視化, 解析を可能にするために必要なデータ構造の理解. そして, その形式ではないデータを取得したときに適切に変換出来ること.
 - 高度なデータのクリーニングと変換
 - いろいろなソースからのツールを使ったデータの結合・分析

内容

- データの指定と変換 (投影, 選択, 結合, 集約/グループ化, 集計など) のための問い合わせ言語と操作
- データのユーザおよび取得者としての構造化/スキーマベースのシステム
 - 関係データベース (SQL), API とプログラムによるアクセス, インデックス作成
 - XML, XPath, 構造化されたデータにアクセスし問い合わせするための API
- データのユーザおよび取得者としての半構造化データシステム
 - 解析して構造化する JSON (JavaScript Object Notation) を生成する API を介したアクセス
- データの取得と構造化における非構造化データシステム

- ウェブスクレイピング
- 構造を与えるためのテキスト/文字列解析/処理
- リモートシステム上のデータへのアクセスの認証と認可に関するセキュリティと倫理的配慮
- ソフトウェア開発ツール (GitHub, version control など)
- 大規模データシステム
 - 分散データストレージのための理論的枠組み
 - example systems (MongoDB, HBase, NoSQL systems) にアクセスする実習
 - * Amazon Web Services (AWS) で提供されるパブリックデータセット (ランドサット, 遺伝学, マルティメディア)

A.7 統計モデル入門

ねらい

回帰分析を題材に統計的推測の枠組みを学ぶ。線形モデルを基礎に置き、非線形アプローチは線形モデルとの比較を通じて学ぶ。初年度のデータサイエンスの科目では様々な統計分析の基礎となる重要な概念を学ぶが、この科目は、それらに基づいてさらに先に進む。学習する概念は全て、実データに根付き、実データから動機づけされたものである。この科目の内容を十分理解することで、統計的学習、機械学習の重要なアルゴリズム及び方法論が共通して持つ構造を理解できるようになる。

学習目標

- 線形モデル及びその利用に関わる数理的構造の理解
- 科学的推論, 理解, 伝達を目的とした実データでの回帰モデルの適用
- モデルの不確実性への理論的アプローチ, シミュレーションによるアプローチ

内容

- 統計的仮説検定, 信頼区間の復習
- 推定 (最尤原理やベイズ理論に基づく)
- 線形モデル
 - 最小二乗法など回帰分析の理論. 推定論
 - 重回帰分析
 - * Box-cox 変換などの変数の変換, モデル選択
 - * 交互作用, ダミー変数, ANOVA
 - ロジスティック回帰など, 一般化線形モデル
- 古典的な回帰モデルの代替手段. 決定木. スムージング手法, スプラインなど
- モデル選択の初歩
 - 正則化法, バイアス・バリエアンスのトレードオフ (ケチの原理, AIC, BIC など)
 - クロスバリデーション (交差検証法)
 - リッジ回帰. lasso を含む罰則付き回帰モデル

A.8 統計的学習・機械学習

ねらい

計算機科学の機械学習におけるアルゴリズム的視点と統計的思考の予測的視点を融合した科目。標準的な機械学習手法と様々な学問分野の問題への応用に焦点を当てる。学生は、統計的学習の理論的基礎を理解するだけでなく、学術や産業における新たな問題に首尾よく適用するのに必要な実際的なスキルを身につける。

学習目標

- 広く使われている標準的な機械学習のアルゴリズムを使ってみる。
- 上記の方法の基礎理論を理解して、推測を行う場合の限界に触れる。
- 実データに上記の方法を適用し、その性能を評価し、その結果を適切に説明出来るようになる。

内容

- アルゴリズム
- スケーラビリティや次元の呪いといったモデルを実装する上での問題点
- 性能測定 (Performance metrics) と予測
 - － 損失関数
 - － 交差検証法, 性能測定 (performance measures), 正則化法といったモデル選択やモデル評価
- データ変換
 - － 次元削減法, 主成分分析, 特徴抽出
 - － 平滑化手法とアグリゲーション手法: カーネル法, 時系列・空間・時空間平滑化, モデル平均法
- 教師あり学習
 - － 回帰分析: 線形モデル, 回帰木
 - － 判別手法: 決定木, ロジスティック回帰, 分離超平面, k 近傍法
- 教師なし学習
 - － クラスタリング: k 平均法, 階層クラスタリング
- アンサンブル法: ブースティング, バギング

実習

A.9 実習

ねらい

卒業実習では履修してきた科目で学んだことを統合する包括的な学習体験をする。学生は3つの学問分野全てから学んできた知識と経験を繋げ、それを他の学問分野へ応用する。この科目で学生は学んできたことを統合し、その知見を込み入った新たな課題に応用する。実際のデータサイエンスの課題解決は、データ収集から始まり、データ処理を経て問題への適切な解析手法の適用、最後に明快で分かりやすく結果を伝える。学生はそのような課題解決の全ての過程に従事する。卒業研究はコアの学問分野や研究対象のデータが生まれた分野の基礎的知識の必要性を十分理解させて、専門家としての努力をしていく準備をさせる。

学習目標

- 最初の課題の定式化から解法の提案まで，さらにその成果の伝達に至るデータサイエンスにおける課題を扱う能力．学生は，実データを収集，前処理，解析して当初の課題への処方箋を提案するというプロセス全てにおいて熟練していく．
- プロジェクトで少なくとも 1 人の他の生徒と協力してチームで働く方法を学ぶ．
- 再現可能で倫理上も問題ない結果とコードを提出する．
- 口頭発表と紙ベースのレポート提出を以下の点に注意して行う．
 1. モチベーション，課題定義，課題への既存のアプローチ
 2. 提案する解法
 3. 結果，結論，現状の解析の限界と今後の研究の方向

卒業実習の例

応用プロジェクト 興味のある応用分野を持つ学生向け：興味ある課題を解決するための学習アルゴリズムをどのように適用するのか探索し，モデルのパフォーマンスとその限界を評価．

アルゴリズムのプロジェクト 既存のアルゴリズムとそれらに関連する文献，または応用分野のアプリケーション領域のいずれかに動機づけられた学生向け：新しい学習アルゴリズムや既存のアルゴリズムの新しい変種の提案，または既存のアルゴリズムを新たな応用分野に適用．

理論プロジェクト 理論においてより野心的な学生向け：新しい学習アルゴリズムあるいは既存の学習アルゴリズムの（理論的な）性質を解析．場合によっては証明．