JSDSS 日本ソーシャルデータサイエンス論文誌
第 3 巻　第 1 号（2019 年 3 月）

特集：データサイエンス人材育成

データソン：統計数理研究所におけるデータ分析ハッカソン

神谷 直樹[†1]　宮園 法明[†1]

概要：我々は，統計数理研究所統計思考院における統計思考力人材育成事業の一環として，データ分析ハッカソン（データソン）を実施している．ハッカソンはコンペティションの一形態として知られているが，人材育成という観点に立つと，アクティブ・ラーニングの一形態といえる．人材育成という観点に立ったときのデータ分析ハッカソンに求められる要件をまとめ，これまでに実施してきたハッカソンにおける課題と対策について報告する．

キーワード：データソン，データ分析ハッカソン，データサイエンティスト，人材育成

1. はじめに

統計数理研究所では 2016 年からデータ分析ハッカソンを実施している [1, 2]. データサイエンティストには，統計学や統計的機械学習などのデータサイエンス力に加え，情報学やソフトウェア工学などのデータエンジニアリング力とビジネスを理解し推進するビジネス力というスキルが必要である．これらのスキルを有機的に獲得するためにはアクティブ・ラーニング型 [3, 4] の育成が不可欠である [5]. 例えば，ハッカソンは Harvard Business Review 誌でデータサイエンティスト [6] が取り上げられてからは特に，インダストリ，アカデミア，非営利組織や自治体・政府機関において様々なテーマで実施されて注目が集まってきている（図 1）．本稿では，これまでに統計数理研究所で実施した計 3 回にわたるデータ分析ハッカソンの取り組みについて紹介する．

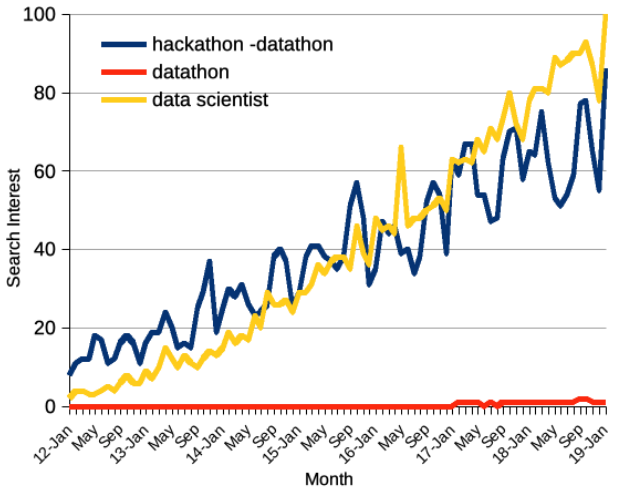

図 1　2012 年 1 月から 2019 年 1 月までの“hackathon（“datathon”を除く）”，“datathon”，“data scientist”に関する毎月の Google トレンド

1.1 ハッカソンとは

ハッカソン (Hackathon)とは Hack と Marathon を掛け合わせた造語で，もともとはソフトウェア開発分野で実施されていたイベントである．課題によっては，様々な呼称をつけられることもある．例えば，マーケティング分野におけるハッカソンはマーカソン (Markathon) と呼ばれる [7]. また，近年はデータサイエンス力やデータエンジニアリング力が必要スキルの中核となるハッカソンをデータソン (Datathon) [8, 9] と呼ぶようになってきている（図 1 参照）．ただし，ハッカソンで扱う課題が異なっても実施形態はほぼ同じで，参加者は数人のチームで，（多くの場合）正解を決めにくい課題に一定期間集中的に取り組み，その成果を競う．

1.2 データサイエンティスト育成に向けたデータ分析ハッカソンのあり方

人材育成に関わると考えられるコンペティションを表 1 にまとめた．統計数理研究所データ分析ハッカソンの参加者は，制限されたデータ分析・開発環境でオープンな課題（正解を決めにくい課題）に取り組んでいる．人材育成に向けたハッカソンの要件の一つは，制限されたデータ分析・開発環境で実施することである．統計学や統計的機械学習などのデータサイエンス力や情報学やソフトウェア工学などのデータエンジニアリング力は，分析・開発環境のハードウェア性能に影響されることが容易に推測される．例えば，「大きな」統計的機械学習モデリングを行うためには相応のハードウェアが必要になり，そのような環境を用意できた参加者が必然的に有利になってしまう．ハッカソンに対する学習ツールとしての注目も集まっている中 [5]，データサイエンス力，データエンジニアリング力のいずれでもない要因が成績に影響しない環境を提供することが必要である．

† 1 統計数理研究所（連絡先：nkamiya@ism.ac.jp）

表 1　人材育成に関わるコンペティションの分類

	正解のある課題設定	オープンな課題設定
制限された分析・開発環境	国際大学対抗プログラミング・コンテスト	統数研データ分析ハッカソン，enPIT「クラウドアプリケーション開発演習」
オープンな分析・開発環境	The Data Science Bowl, Kaggle など多数	データビジネス創造コンテスト，スポーツデータ解析コンペティションなど

一方で，参加者がデータサイエンスやデータエンジニアリングの最新技術を利用可能な環境を提供することも重要である．技術革新が日進月歩の速さで進む現代においては，参加を通じて最新技術をより使いこなせるようになることが人材育成を目的としたハッカソンに求められる要件と考えられる．

また，参加者が触れることのできるデータに対しても一定の配慮が必要である．日本では，データやモデルの流通に関して制度などの整備 [10, 11] が始まったばかりで，その社会受容が進むかどうかは今後の課題になっている [12]．企業内で実施されビジネス上のインサイトを得ることがハッカソン実施の目的になっていない場合であっても，ハッカソン実施期間中だけでなく終了後において，データが流出しないことが保証できる環境を用意することが必要である．

2. 統計数理研究所におけるデータ分析ハッカソン

1.2 節で述べた要件をふまえて統計数理研究所で実施してきたデータ分析ハッカソンは，第1回とそれ以降では実施形態に変化がある．以下では，それぞれのデータ分析ハッカソンに分けて記述する．

2.1 第1回データ分析ハッカソン

第1回データ分析ハッカソンは文献 [1] で詳細に報告されているので，ここではその概略をまとめる．統計数理研究所では，文部科学省科学技術試験研究委託事業「データサイエンティスト育成ネットワークの形成」を通じインターンシップ・プログラムを活用して，学生に企業におけるデータ分析の実際を経験してもらう試みを推進してきた．これに対して，我々自身が実際に学生にデータ分析の機会を与える方法の1つとして，2016年2月20日から21日にかけての2日間，統計数理研究所においてデータ分析ハッカソンを実施した．参加者は1チーム3名，6チームの学生（大学生，大学院生）であった．

(1) 課題

あるエンターテイメント系企業から提供していただいたおよそ1.5年分，合計約1千万レコードのデータから，売上向上の施策を提案することを課題とした．

(2) 分析環境

分析環境は，統計数理研究所共用クラウド計算システムを利用した．このシステムはDELL社製サーバー69ノード（合計138CPU，1380コア，16.4TBメモリ）と，大規模共有ディスク装置（合計364TB）を中心として構成されている．このクラウド上に同一環境のインスタンスを8つ用意した．参加6チームへインスタンスを1つずつ提供し，運営側の管理・バックアップ用インスタンスとして1つずつ用意した．各インスタンスは，4コア，64GBメモリ，500GB HDDの仮想ハードウェアを持っていた．そして，これらのインスタンスを1つの仮装サブネット内に配置した．参加者は自分のPCからVNC仮想デスクトップ経由で各インスタンスにアクセスした．一方，運営側はSSH経由でこれら全てのインスタンスにアクセスすることができた．

OSはFedora23とし，標準的なLinux環境を用意した．また，データ分析ツールとして，オープンソースで入手できるMySQL Workbench（データベース用のGUI），R（統計分析パッケージ）/ RStudio（そのGUI），Python 2.7 / Anaconda（Python用の統計分析ライブラリをパッケージしたもの）/ Spyder（その統合開発環境）を用意した．各言語の追加パッケージは，参加者からのリクエストに応じて全チームのインスタンスに対して一律にインストールした．さらに，プレゼンテーション用として，オフィス統合環境であるLibreOfficeを提供した．

(3) 審査

各チームに課題についてプレゼンテーションを求め，アカデミア2名，インダストリ2名からなる審査員が審査を行った．審査の基準として，創造性，有用性，技術力，表現力の4つの観点について採点し，それを集計した後，審査員全員の討議を経て，最優秀賞，優秀賞，審査員特別賞の3賞を決定した．

(4) 参加者からのフィードバック

ハッカソン参加者の提出課題，ならびにハッカソン終了後に行った参加者に対するヒアリングの結果は以下の3点に集約することができる．①データ分析は探索的で，手法ありきでは上手くいかない．その分野の定石に従って分析しても，与えられたデータによっては必ずしもうまくいくとは限らない．定石に拘るあまり制限時間内に成果が出せないことがあった．②限られた計算資源を有効に利用することが重要である．各インスタンスに管理者権限を与えず，一つのIDをチームで利用させていたことが原因の一端と考えられるが，提供ツールの利用方法によってはそれがク

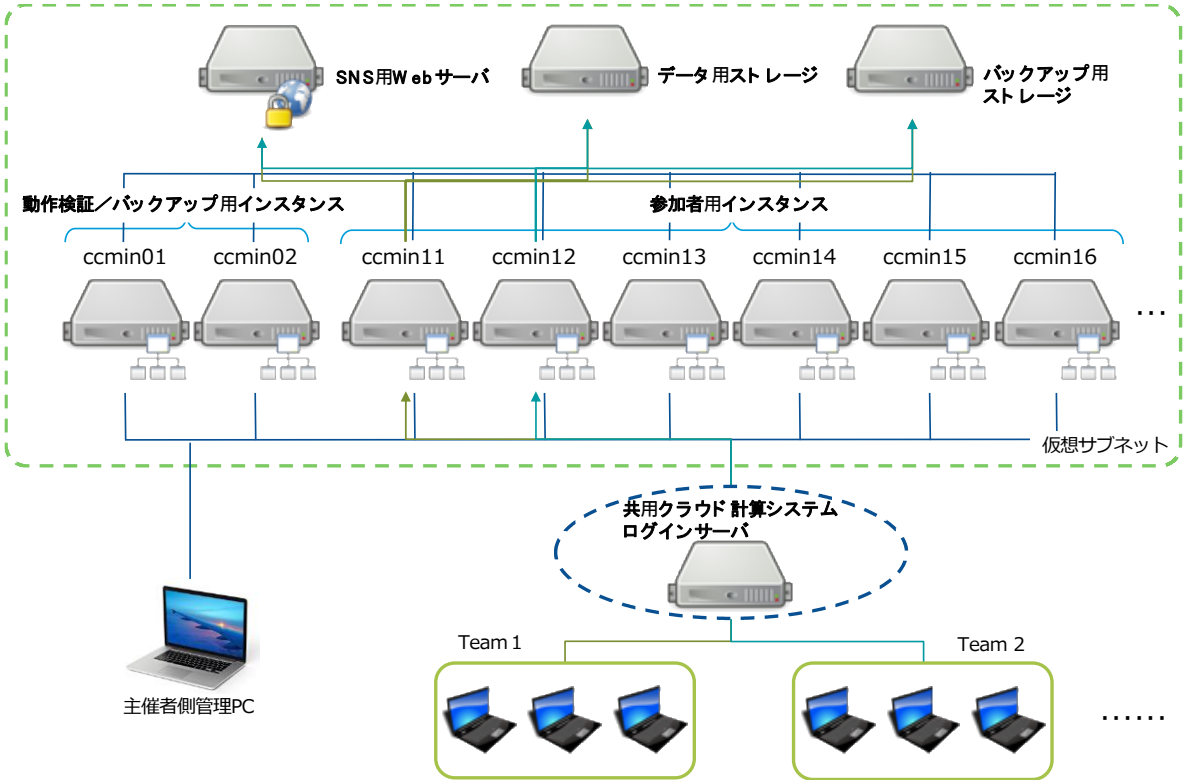

図 2　統計数理研究所データ分析ハッカソンのネットワーク環境

ラッシュしてしまうことがあった．③チーム作業には当然のことであるが，分散して分析・開発するよりもリーダーシップを発揮する人材の存在が重要である．

(5) 改善点

文献[1]には 7 つの改善点がまとめられているが，中でも以下の 2 点の改善が喫緊の課題といえた．一つは，分析ツールの利便性である．例えば，2014 年に IPython から派生した Jupyter Project [13] のプロダクトである Jupyter Notebook は Web ベースのデータ分析・開発ツールとして広く使われているが，ブラウザ設定の制約から 1 チーム 1 人しか使用できなかったため，設定を見直す必要があった．また，さらなるカーネルやパッケージ追加を検討することも必要であった．もう一つは，フォールバック・プランの用意である．特にインフラに障害があったときのデータや各チームのそれまでの作業などをバックアップするシステムを構築して提供していかなくてはならない．

2.2 第 2 回データ分析ハッカソン

第 2 回データ分析ハッカソンは，前節で述べた点を改善するとともに，実施形態そのものも見直した．通常のハッカソンは数日間の短期集中，対面形式で実施される．数日間の短期集中型で実施することは，チーム間のコミュニケーションが促進される環境設定である一方．数日間という実施期間を設定することによって，参加したいが都合がつかないなどの理由によって希望者の参加機会を制限してしまう．そこで，仮想サブネット内に Web サーバーを用意しチーム内限定のチャット・システムを構築して提供し，審査のためのプレゼンテーションを除いた全ての実施をオンライン上に移行した（図 2）．第 2 回データ分析ハッカソンは，2018 年 1 月 22 日から 3 月 15 日にかけて 2 ヶ月弱の間実施し 3 月 20 日に審査を行った．参加者は公募し，1 チーム 3 名，6 チームの学生と社会人が参加した（学生チームは 2 チーム，社会人チームは 4 チーム）．

(1) 課題

株式会社 LIFULL から提供していただいた日本全国の賃貸物件データ（1 年分，賃料，面積，立地，築年数，間取り，建物構造，諸設備などのデータ）から新しい不動産テック，あるいは現状の不動産テックにおける課題を解決する方法を提案することを課題とした．

(2) 分析環境

分析環境は第 1 回データ分析ハッカソンと同等とした．ただし，分析環境の OS などは最新のバージョンで提供し，ブラウザ設定を見直して Jupyter Notebook などの Web ベースのデータ分析・開発ツールをチーム全員で使用できるようにした．各言語の追加カーネルやパッケージのインストールは，分析環境のデザインも含めてチームごとの特色を出すために，リクエストをしたチームのイン

スタンスごとに行った．前回のデータ分析ハッカソンではリクエストに応じて全チームのインスタンスに対して一律にインストールしたが，自分たちがリクエストしたパッケージでないものを利用することは稀であった．

フォールバック・プランとして 2 種類用意した．一つは，データを各インスタンスにインストールして提供するのではなく，各チームのインスタンスとは別に用意したストレージに集約した．各チームにはこのデータ用ストレージにアクセスし，分析・開発に必要なデータを自分たちのインスタンスにコピーして使用するように教示した．さらに，実施期間が 2 ヶ月弱と，通常のハッカソンに比べて長期であるため，1 週間ごとにバックアップ用ストレージに各インスタンスの作業内容をバックアップした．なお，このバックアップ用ストレージにはチームごとの領域を割り当てたので，参加者は過去の作業内容の参照やチーム内でのファイルのやりとりも可能であった．

(3) 審査

各チームに課題についてプレゼンテーションを求め，アカデミア 2 名，インダストリ 2 名からなる審査員が審査を行った．審査の基準として，前回の審査基準である創造性，有用性，技術力，表現力に加え，チームワーク（リーダーの元でチームとして機能していたか）を加えた 5 つの観点について採点し，それを集計した後，審査員全員の討議を経て，最優秀賞，優秀賞，審査員特別賞の 3 賞を決定した．チームワークについては，チャット・システム内のやりとりを分析することを想定していたが，このチャット・システムの利用は特定のチームに偏っていた．したがって，その分析結果は比較のための指標になり得ず，プレゼンテーション内容から想像するに留まってしまった．

(4) 参加者からのフィードバック

ハッカソン参加者の提出課題，ならびにハッカソン終了後に行った参加者に対するヒアリングの結果は以下の 2 点に集約することができる．①作業の進捗についてしっかり時間管理しなければ，参加者にとって効果的なコンペティションになり得ない．審査時に統計数理研究所（東京都立川市）に集まってもらうことを除いて，データ分析や開発作業は参加者の所在地を問わず実施できていたが，逆にいつでも作業を行うことができるという環境がチームによっては作業の遅延を生じさせ，成果に結びつかなかった．②チームワークを評価するための指標が十分ではなかった．通常，データ分析関連業務はチームで行われることが多い．コンペティションを学習ツールとみなす場合 [5]，その効果測定のためにチーム内の作業やチームの構造を分析する必要がある．

(5) 改善点

参加者を一箇所に集めて実施する通常のハッカソンでは，運営側が参加者の作業進捗状況を常に確認することができる．しかし，オンライン上で実施する場合には，どのタイミングで作業を促せば良いのか，作業ログからだけでは容易に判断できない．したがって，参加者間の交流も含めて，作業進捗を運営側が対面で確認する機会を設ける必要がある．また，今回から追加したチームワークという審査項目について，チャット・システム内のやりとりの分析は比較のための指標になり得なかったので，チーム内の作業やチームの構造を分析する手段を再考する必要がある．

2.3 第 3 回データ分析ハッカソン

第 3 回データ分析ハッカソンは前回と同様オンライン上で開催した．2018 年 8 月 8 日から 10 月 31 日にかけて 3 ヶ月弱の間実施し，11 月 9 日に審査を行った．また，9 月 26 日には参加チームのうち希望したチームを統計数理研究所に集め，作業内容確認などのために中間報告会を実施した．この中間報告会には審査員 3 名も参加し，中間時点での作業内容や方向性について議論した．なお，参加者は，1 チーム 3 名，7 チームの学生と社会人であった（学生チームは 1 チーム，社会人チームは 6 チーム）．

(1) 課題

株式会社インサイトテックが運営する Web サービス「不満買取センター」に一般ユーザが投稿した様々な不満に関するデータ（約 2 年 3 ヶ月分，約 130 万件の本文とカテゴリデータ，約 7 万人分の投稿者プロフィール情報も付随）とカテゴリ別不満特徴語辞書を用意した．これに加えて，株式会社日本データ取引所が提供しているデータカタログ内のデータを用意した．これらのデータから，データ分析に基づいて，多くのステークホルダーが利益（金銭に限らない）を得られる（相利共生）方法を提案することを課題とした．

(2) 分析環境

分析環境は第 2 回データ分析ハッカソンと同等とした．ただし，分析環境の OS などは最新のバージョンで提供した．

(3) 審査

各チームに課題についてプレゼンテーションを求め，アカデミア 3 名，インダストリ 2 名からなる審査員が審査を行った．審査の基準としては，前回の審査基準である創造性，有用性，技術力，表現力，チームワークという 5 つの観点について採点し，それを集計した後，審査員全員の討議を経て，最優秀賞，優秀賞，審査員特別賞の 3 賞を決定した．チームワークについては，チャット・システム内のやりとりを分析することを想定していたが，今回も利用が特定のチームに偏っていたため，その分析結果は比較のための指標になり得ず，プレゼンテーション内容から想像するに留まった．チームワークに対する評価

は審査参考資料として扱った．

(4) 参加者からのフィードバック

ハッカソン参加者の提出課題，ならびにハッカソン終了後に行った参加者に対するヒアリングの結果を以下の 3 点に集約することができる．①計画的に取り組めたチームとそうでないチームに分かれた．特に社会人参加者の場合，業務の合間を縫って参加することになったが，役割分担をすることで効率的にできたというコメントがあった一方で，役割分担はしていたが約 2 ヶ月間は作業していなかったため課題提出期限に間に合わなかったというチームがあった．前者のようなチームは中間報告会に参加したチームの典型で，後者のようなチームは参加しなかったチームの典型であった．②使い慣れたアプリケーションを使いたい／インスタンスの管理者権限を参加者に与えて欲しい．③第 2 回データ分析ハッカソンに続き，チームワークを評価するための指標が十分ではなかった．仮想サブネット内においたチャット・システムのマニュアルを見直すなど，その利用を促す措置を施したが十分に機能しなかった．参加チームの多くは対面で集まって議論したり，Skype などの他のコミュニケーション・ツールを利用していた．

(5) 改善点

計画的にデータ分析ハッカソンに参加することは効果的なスキルアップにつながると考えられる．参加者が計画的に取り組めるように各チームにメンターを配置することが最も効果的かもしれないが，中間報告会を実施するだけでも参加者の作業進捗を管理する上で効果があるので，中間報告会への参加を義務付けることを検討していきたい．

分析環境について，参加者が希望を伝えやすい環境を検討する必要がある．参加者には，事前にインスタンスの概要ともに，各種分析ツールの詳細を伝えてあった．さらに，これまでのデータ分析ハッカソンでは常に，データからデモグラフィなどを抽出するサンプル・プログラムを提供してきた．しかし，通常業務場面や学習場面で商用ツールしか使用していない参加者に対しては，オープンソースで入手できる分析・開発ツールの使用方法のインストラクションをより充実させる必要があるかもしれない．

第 2 回データ分析ハッカソンに続いて，チャット・システム内のやりとりの分析は比較のための指標になり得なかった．チーム内の作業やチームの構造を分析する別の手段を講じなくてはならないことが明らかになった．

3. 人材育成を目的としたハッカソンの課題

ハッカソンは，アクティブ・ラーニング型人材育成の一形態に分類できる[5]．アクティブ・ラーニングとは学習者が能動的に学習に取り組む学習方法の総称であり，いわゆる座学による一方的な教授方法の対極に位置する．これまで述べてきたハッカソンは，参加者にオープンな課題を与えて解決させているので，アクティブ・ラーニングに分類される問題解決学習 (Problem Based Learning; PBL)，あるいは課題解決学習 (Project Based Learning; PBL)といえる．

アクティブ・ラーニングについては，概念的な整理 [3, 4] が十分進んでいるが，文献 [14-17] などにあるようにその効果測定は定性的であり続けている．統計数理研究所データ分析ハッカソンは，オンライン上に実施形態を移行させたことによって，効果測定のための客観的なデータを収集可能な状態にある．アクティブ・ラーニング型人材育成において，参加者にどのようにアプローチしていくことが参加者の効果的な学習につながるのか [18] を検討していくことができると考えている．

4. 今後の展望

データサイエンティストに必要と考えられているデータサイエンス力，データエンジニアリング力とビジネス力全てにおいて高水準である人材は稀であり [19]，各自の得意スキルを生かしてチームで業務に当たることの方が多い．統計数理研究所データ分析ハッカソンは，チーム参加を基本としているので，各参加者が得意スキルを伸ばす環境として使えると考えている．また，育成目標のデータサイエンティストとしては，クロスファンクショナル・データサイエンティスト，シニア・データサイエンティストやデータプロダクト・マネージャーなどが考えられる．

謝辞　ハッカソン実施にあたり，参加していただいた学生，社会人各位，そして審査員を務めていただいた皆様に感謝する．参加者の皆様には十分とはいえない環境にも関わらず，実力を発揮し，参加を楽しんでくださった．審査員の皆様には，参加者に対して審査だけにとどまらず，さらなるスキルアップのための示唆を与えていただいた．第 1 回データ分析ハッカソンを指導していただき，我々を運営に携わらせていただいた丸山氏（現在は Preferred Networks 株式会社　PFN フェロー）に感謝する．現在までのデータ分析ハッカソンの企画・運営・実施に対して重要なインサイトを与えてくださった．そして，これまでの全てのデータ分析ハッカソン運営に協力していただいた統計数理研究所の篠崎支援員に感謝する．最後に，データを提供してくださった企業とその担当者の方に最大限の謝意を表する．

参考文献

[1] 丸山宏，神谷直樹，宮園法明，“クラウド環境を利用したデー

タ分析ハッカソンの計画と実施," *Estrela*, No.272, pp.30-37 (2016).
[2] 神谷直樹, "データ指向キャリアへの効率的支援プログラムとしてのデータ分析ハッカソンの設計・実施." 統計数理研究所公募型共同利用重点型研究「データサイエンス人材育成メソッドの新展開」研究集会, (2019).
[3] Jonassen, D., "Designing Constructivist Learning Environment," in Reigeluth, C. M. (ed.), *Instructional-Design Theories and Models: A New Paradigm of Instructional Theory, Volume II*, pp. 215-239, Lawrence Erlbaum Associates, Mahwah: NJ (1999).
[4] Reigeluth, C. M., Beatty, B. J. and Myers, R. D., *Instructional-Design Theories and Models: The Learner-Centered Paradigm of Education, VolumeIV*, Routledge, New York: NY (2017).
[5] Briscoe, G. and Mulligan, C., "Digital Innovation: The Hackathon Phenomenon," (2014). https://qmro.qmul.ac.uk/xmlui/bitstream/handle/123456789/11418/Briscoe%20Digital%20Innovation:%20The%20Hackathon%20Phenomenon%202014%20Published.pdf?sequence=2（2019 年 02 月 22 日アクセス）
[6] Davenport, T. H. and Patil, D., "Data Scientist: The Sexiest Job of the 21st Century," *Harvard Business Review*, Vol. 90, No. 10, pp.70-76 (2012).
[7] Calco, M. and Veeck, A., "The Markathon: Adapting the Hackathon Model for an Introductory Marketing Class Project," *Marketing Education Review*, Vol. 25, No. 1, pp. 33-38 (2015).
[8] Bigdata Doctor, "The Datathon and How to Make the Most of It," 2016.6.28. Big Data Doctor Website http://bigdata-doctor.com/datathon-how-to-make-the-most-of-it/（2019 年 02 月 22 日アクセス）
[9] Anslow, C., Brosz, J., Maurer, F. and Boyes, M., "Datathons: An Experience Report of Data Hackathons for Data Science Education," *Proceedings of the 47th ACM Technical Symposium on Computing Science Education*, pp. 615-620 (2016).
[10] "官民データ活用推進基本法," http://elaws.e-gov.go.jp/search/elawsSearch/elaws_search/lsg0500/detail?lawId=428AC1000000103（2019 年 02 月 10 日アクセス）
[11] "AI・データの利用に関する契約ガイドライン," http://www.meti.go.jp/press/2018/06/20180615001/20180615001.html（2019 年 02 月 10 日アクセス）
[12] 独立行政法人情報処理推進機構 AI 白書編集委員会,「AI 白書 2019」, 角川アスキー総合研究所 (2018).
[13] "Jupyter Project," https://jupyter.org/ （2019 年 02 月 10 日アクセス）
[14] Prince, M., "Does Active Learning Work? A Review of the Research," *Journal of Engineering Education*, Vol. 93, No. 3, pp. 223-231 (2004).
[15] Felder, R. M. and Spurlin, J. "Applications, Reliability, and Validity of the Index of Learning Styles," *International Journal of Engineering Education*, Vol. 21, No. 1, pp. 103-112 (2005).
[16] Freeman, S., Eddy, S. L., McDonough, M., Smith, M. K., Okoroafor, N., Jordt, H. and Wenderoth, M. P., "Active Learning Increases Student Performance in Science, Engineering, and Mathematics," *Proceedings of the National Academy of Sciences*, Vol. 111, No. 23, pp. 8410-8415 (2014).
[17] Huppenkothen, D., Arendt, A., Hogg, D. W., Ram, K., VanderPlas, J. T. and Rokem, A., "Hack Weeks as a Model for Data Science Education and Collaboration," *Proceedings of the National Academy of Sciences*, Vol. 115, No. 36, pp. 8872-8877 (2018).
[18] Skinner, B. F. *Upon Further Reflection*, Pearson, (1987).
[19] Patil, D. and Mason, H., *Data Driven: Creating a Data Culture*, California, O'Reilly Media, (2015).