電気通信大学紀要 30 巻第 1 号 pp.1-6（300110）〔論文〕

# データサイエンスの人材育成モデル

清　洲　正　勝，　田　村　元　紀

# Human Resources Development Model of Data Science

Masakatsu KIYOSU, Motonori TAMURA

**Abstract**

By the accumulation of big data and improvement of the computer ability, the mathematics theory including machine learning or data mining became a practical use stage. With it, the specialist in data science is highly demanded in the field of various sciences and industry.

We defined the conformity degree index for lecture contents of the data science and tried effective inspection in the human resources development program. As a result, it was shown that the conformity degree index for each student was effective for an index of educational training, and the conformity degree index for each teacher was effective for the teacher or the lecture constitution. These indexes are extremely effective to optimize the human resources development program in the data science.

Key words : data science; human resources development; conformity degree index

## 1　序論

### 1.1　背景

計算機性能の向上と、複雑な計算処理が可能な専用処理装置の登場によって、数理最適化や統計学、機械学習、データマイニング分野で発展してきた膨大な計算量を要する理論は、高い実用的効果が証明されつつある。それに伴い、様々な学術や産業の応用分野において高度なデータサイエンスの専門家が必要とされ、教育研究機関等で研究や産業の人材育成が進んでいる。“Data, Information, Knowledge, and Wisdom”（Gene Bellinger ら，2004）[1]の概念を独自に拡張した、現在のデータサイエンスの概念を表す図を示す。（図 1 ）[2]

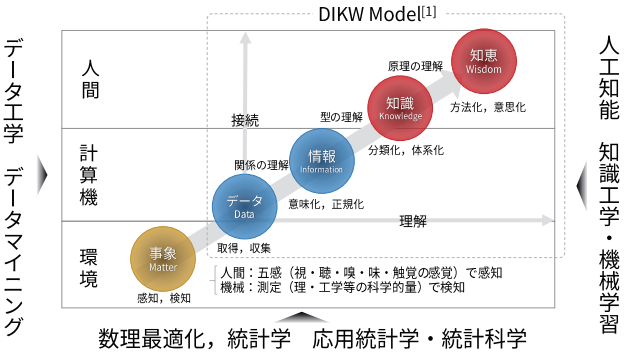


図1．データサイエンス

### 1.2　目的

人材育成における効果や満足度に関する調査研究（星野ら、2005）（山口ら、2014）[3,4] は、凡そ均一的な学生を対象に、効果を定量化し因果関係や有意性を明らかにしている。また、顧客に対する満足度の指数には、政府の成長戦略によって発足した、日本生産性本部サービス産業生産性協議会が策定している顧客満足度指数がある。[5]本研究では、データサイエンスに関する講義内容に対する受講生の適合度指数を新たに定義し、分析することによって講義の改善を行いデータサイエンスにおける最適な人材育成モデルを構築することを目的とする。本研究における人材育成モデルのスキルセットは、“The Data Science Venn Diagram”（Drew Conway、2010）[6]と“Typology of an Entrepreneur”（Jeffry A. Timmons、

Received on September 6, 2017.
産学官連携センター

1989)[7] を包括するものと定義する。（図２）

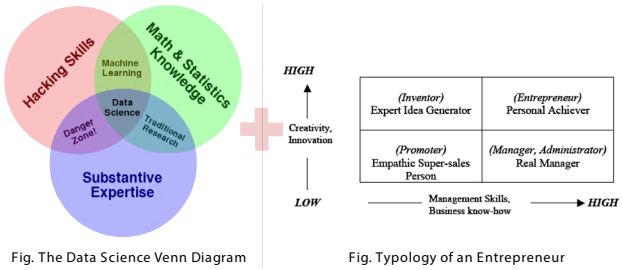


Fig. The Data Science Venn Diagram　　Fig. Typology of an Entrepreneur

図 2.　人材育成モデルのスキルセット

## 2　手法

最適な人材育成モデルを構築するための受講生別及び講義別の適合度指数による評価を行うために、最初に適合度指数と調査票設計、次に講義環境の構築と実施、最後に実施結果と適合度指数の評価と言う３段階で研究を進めた。具体的には、2016年度実施のデータアントレプレナープログラム[8]集中講義の「データアントレプレナー実践論」（以下、実践論科目）及び「データサイエンティスト特論」（以下、特論科目）の２科目を研究対象とした。

### 2.1　適合度指数と調査票設計

#### 2.1.1　調査票の設計

研究対象のプログラム受講生は、年齢、学歴、職種、産業分野、専門分野、経験等が異なり、多種多様な人材であるため、これまでの統計的手法では最適なモデルの構築は難しいと考えられる。2016年度の応募者の年齢と産業分野を示す。（図３，図４）

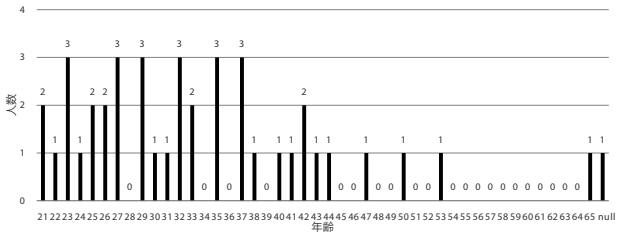


図3.　応募者の年齢

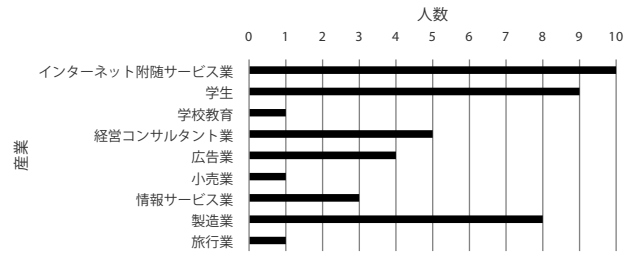


図4.　応募者の産業分類

そこで、講義内容と受講生の適合度指数を定義し調査するため､受講生に大きな負担にならず､かつ十分なデータを収集可能な範囲で、自由回答法を用いた評定法と自由記述法による調査票を設計した。

評定法は、分野と技術の自己分析と、受講後における５つの感度（貢献度、難易度、習熟度、興味度、志向度）、合わせて７つの属性を調査する質問項目とした。（表１）

表 1.　評定法による質問項目と属性

| 項目 | 質問 | 属性 |
|---|---|---|
| 1 | 講義内容の分野は良く知っていましたか？ | 分野 |
| 2 | 講義内容の技術は良く知っていましたか？ | 技術 |
| 3 | ご自身の役に立ちましたか？ | 貢献度 |
| 4 | 分かりやすかったですか？ | 難易度 |
| 5 | 理解できたと思いますか？ | 習熟度 |
| 6 | 講義内容に興味を持ちましたか？ | 興味度 |
| 7 | 講義内容を深く知りたいと思いましたか？ | 志向度 |

７つの属性に対し、５段階のLikert尺度の評定と段階を設定し、単一回答とするようにした。（表２）

表2.　評定段階

| 段階 | 評定 |
|---|---|
| 5 | そう思う |
| 4 | ややそう思う |
| 3 | どちらともいえない |
| 2 | あまりそう思わない |
| 1 | そう思わない |

自由記述法は、講義内容に対する忌憚の無い肯定的及び否定的意見を回答するように促した。

表3.　自由記述法による質問項目と属性

| 項目 | 質問 | 属性 |
|---|---|---|
| 1 | 講義内容の良かった点は何ですか？ | 肯定 |
| 2 | 講義内容の改善点は何ですか？ | 否定 |

#### 2.1.2　適合度指数の設計

評定法の属性より、適合度指数は、受講生個別の適合度指数から講義全体の適合度指数を算出する。既にデータサイエンスに関する情報は多大に存在しており、受講生が既知の内容を扱うことは再確認でしかない。つまり、受講生が知らない分野と技術の事項を扱い、５つの感度が高いものが受講生の知識の幅を広げ、知識量を高める良い講義として定義する。質問項目の文章の性質より、分野と技術の評定段階の値を反転した加重値を$w_a$,$w_t$とすると、全体の重み$W$（0.2 ～ 1.0）は、

$$W = \frac{1}{10}(w_a + w_t) \quad (1)$$

５つの感度、貢献度、難易度、習熟度、興味度、志高

度を評価値の集合として計算する。

$$E^{ds} \ni \{E_c, E_d, E_a, E_{ir}, E_{it}\} \quad (2)$$

講義内容と受講生個別の適合度指数$M_s$（1 ～ 25）は、評価値集合の要素数を$n$とすると、

$$M_s = \sum_{i=1}^{n} w_i E^{ds}{}_i \quad (3)$$

したがって、講義内容と受講生全体の適合度指数$M_l$（1 ～ 25）は、受講生数を$k$とすると、

$$M_l = \frac{1}{k} \sum_{i=1}^{k} M_{s_i} \quad (4)$$

として算出する。適合度指数は高いほど良い。

受講生の調査票提出は、各科目の開催回ごとに行うことで、それぞれの講義がどの程度受講生に適合したかを算出できるようにした。データサイエンスの分野は日進月歩であり、適合度指数や調査票の自由記述を基に、講義内容を改善していくことが、より良い人材育成モデルを構築できると考える。

## 2.2　講義環境の構築と実施

### 2.2.1　講義室環境

講義環境は、講師や受講生が講義内容そのものに集中できるように構築する必要がある。講師の価値の最大化や受講生の負担の最小化のための環境を構築し、講義方針を確立した。このことによって、講義時間内での講義内容の質を向上させ、講師が講義内容に注力し、受講生が講義内容に集中することができる。（図５）

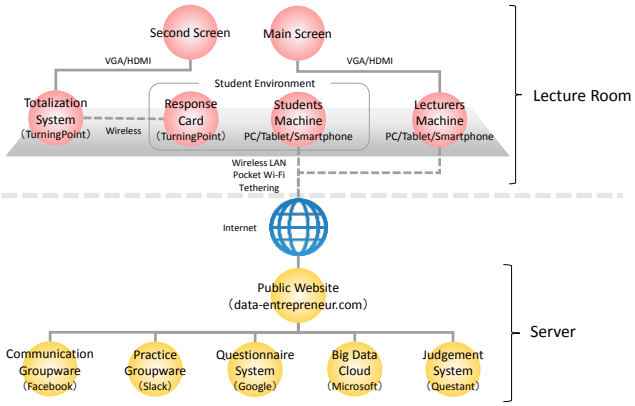


図 5.　講義環境のネットワークトポロジー

### 講義室

講義室は、受講生の座席指定を行い、講義開始時間迄に着席している受講生を出席として扱う。この方法は、出欠点呼の無駄な時間を省略し、遅刻早退の管理を容易にした。座席の机と椅子は、可動式のものを採用し、グループワークでのグループ配置を自在にできるようにした。学内無線LANアカウントを提供し、安定的に分析ができるようにした。

### 受講生の個人分析環境

分析のための個人環境は、受講生自身が使い慣れた個人利用のデータ分析が可能なノートブックPCを持ち込む方法とした。Microsoft Windows またはApple Mac OSいずれかのプラットフォームを推奨した。電源を提供し、給電可能とした。

### 集計システム

講師と受講生の意思疎通を即時的に解決できるように、双方向の集計システム[9]を導入した。受講生に受講生番号を指定し、講義開始前にケースから個別に割り当てられたレスポンスカードを取り着座する。講義が終わると、受講生自身がケースに戻す運用とした。

### レスポンスカード

レスポンスカードの背面には、学内無線LANのアカウントを貼ってあり、この運用でアカウントの持ち帰りによる流出を防止した。講義室のプロジェクターメインスクリーンに、講師PCの講義画面を表示し、レスポンスカードは、直感的に押せる縦５キーの最新のカードを導入した。（図６）


図6.　講義画面とレスポンスカード

### 集計システム画面

講義室のプロジェクターセカンドスクリーンに双方向集計システムを常時表示させることで、講義画面と集計システム画面の切り替えの必要を無くし、リアルタイムに集計結果が分かるようにした。

図６の講義画面の質問に対する集計結果画面を示す。（図７）集計システム画面の上部は、パーセンテージ表示の集計結果を表示し、下部には受講生の氏名と座席をグリッドで表示し、受講生が投票したかどうかを一覧で確認できる。受講生が投票した番号は分からないようになっている。

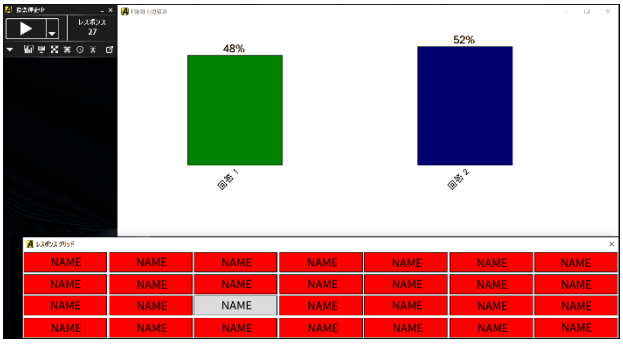


図 7.　集計画面

### 2.2.2　サーバ環境

人材育成プログラムで利用するサーバは、インターネット上にある外部の信頼性の高いCMS（Content Management System）やSNS（Social Networking Service）、ASP（Application Service Provider）、パブリッククラウドを活用した。

**コンテンツマネジメントシステム**

講義に関する全ての情報は、CMSのWordPressを使った人材育成プログラムの公式ウェブサイト[8]に集約した。公式ウェブサイト[7]は、スマートフォンやタブレットの画面に対応したレスポンシブウェブデザインを採用しており、PCだけでなく、これらの端末でも閲覧できるようになっている。

**受講生専用ページ**

受講生は全ての配布物をCMSからデータファイルでダウンロードできるようにした。講義資料は、講義開始前にPDF形式でアップロードしておき、受講生が事前に確認できるようにした。受講生は通学等の隙間時間を有効に活用でき、講義時間の資料配布の時間を省略し、受講生の質問事項の質を上げることができた。

**講師専用ページ**

講師には、専用ページから講義室や受講生の座席情報等を確認できるようにした。講師への連絡事項もこのページに集約されている。

**グループウェア**

受講生への一般的な情報を提供するグループウェアは、コミュニティ作成と継続のために、SNSのFacebook Groupsを活用した。国内外の学会の紹介やデータサイエンスの技術を紹介している。また、実習のための作業用のグループウェアはプロジェクトマネジメントで利用されるChatのSlackを活用した。

講義の一つ「データサイエンス論」では、ブックレビューサイトBooklogにて専門書を紹介した。数理最適化、統計学、機械学習、データマイニング等のデータサイエンスの技術は、産業応用のための専門書が多数出版されており、事前知識として講義を補完する形で選書[10]として纏め、受講生向けに公開し、更新を続けている。（図８）

図 8.　データサイエンス選書

**調査票**

調査票は事前にアンケートシステムのGoogle Formsに準備しておき、毎回の講義終了後１週間以内にウェブフォームから入力する形で提出を行うようにした。

**ビッグデータクラウド**

ビッグデータを処理するサーバは、オンプレミスのシステムではなく、処理性能がスケールアウト可能なクラウドMicrosoft Azureを利用した。クラウド上にビッグデータを置き、データベース言語SQLを使いデータセットを作成後にローカルにダウンロードし、表計算ソフトウェアExcelやプログラミング言語Python等の個人環境によって分析すると言うプロセスとした。

**審査システム**

データ分析の実習では、学内外の審査員８名により、グループ毎のデータ分析結果の審査を行った。その審査と集計にASPのQuestantを利用した。

実践論科目の講義の一つ「データ分析実践手法」では、Pythonの開発環境で有用なJupyter Notebook上でのアルゴリズムの使い方を具体的に講義した。Jupyter Notebookは、プログラミング結果の図表を即時に描画することができ、受講生はPCの画面を確認しながらグループワークを行い、分析資料を纏める作業を行った。以上の講義環境の構築と講義の実施により、講師と受講生が講義時間を十分に活用できることができた。

### 2.3　実施結果と適合度指数の評価

得られた調査票のデータから、実践論科目の講義の１

つ「データサイエンス論」の受講生個別の適合度指数$M_S$を算出した。それぞれの評価値の総和を$E_S$とする。

$$E_s = \sum_{i=1}^{5} E^{ds}{}_i \qquad (5)$$

受講生に課題を与え講師による成績評価を行った。成績は、本学の成績評価基準に則り、０～４（数値が高い方が優秀）までの相対評価を付けた。これらを纏め受講生別の適合度指数$M_S$を昇順ソートした表を示す。（表４）

表 4. 受講生別適合度指数

| 受講生 | $W$ | $E_s$ | $M_s$ | 成績評価 |
|---|---|---|---|---|
| A | 0.20 | 22 | 4.4 | 4 |
| B | 0.20 | 22 | 4.4 | 2 |
| C | 0.40 | 18 | 7.2 | 2 |
| D | 0.50 | 16 | 8.0 | 2 |
| E | 0.60 | 14 | 8.4 | 1 |
| F | 0.50 | 17 | 8.5 | 2 |
| G | 0.80 | 11 | 8.8 | 1 |
| H | 0.40 | 23 | 9.2 | 3 |
| I | 0.50 | 19 | 9.5 | 4 |
| J | 0.40 | 24 | 9.6 | 2 |
| K | 0.50 | 20 | 10.0 | 2 |
| L | 0.50 | 23 | 11.5 | 2 |
| M | 0.50 | 25 | 12.5 | 3 |
| N | 0.60 | 21 | 12.6 | 3 |
| O | 0.60 | 22 | 13.2 | 3 |
| P | 0.70 | 19 | 13.3 | 2 |
| Q | 0.60 | 23 | 13.8 | 3 |
| R | 0.60 | 23 | 13.8 | 3 |
| S | 0.60 | 23 | 13.8 | 3 |
| T | 0.60 | 23 | 13.8 | 2 |
| U | 0.60 | 24 | 14.4 | 3 |
| V | 0.60 | 24 | 14.4 | 4 |
| W | 0.70 | 22 | 15.4 | 2 |
| X | 0.80 | 20 | 16.0 | 1 |
| Y | 0.80 | 23 | 18.4 | 3 |
| Z | 0.80 | 23 | 18.4 | 2 |
| AA | 0.80 | 25 | 20.0 | 2 |
| 平均 | 0.57 | 21 | 12.0 | - |

表４において、受講生個別に見た適合度指数$M_S$は、4.4～20.0まで広い範囲を示した。この時の講義全体の適合度指数$M_l$は12.0である。評価値の総和$E_S$に対して加重値$W$を乗算することによって、適正な適合度指数を算出できていると考えられる。

次に表４より、加重値$W$と評価値の総和$E_S$の散布図を示す。（図９）

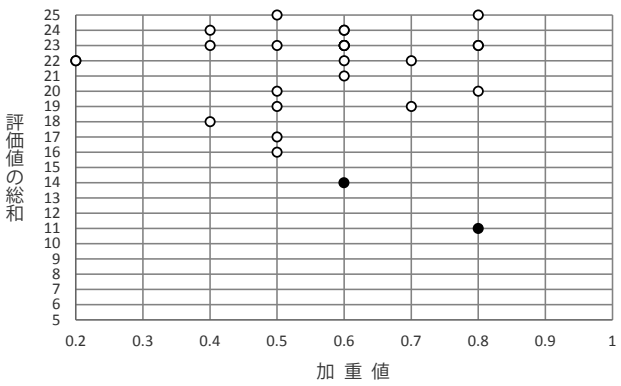


図 9. 加重値と評価値の総和

図９において、黒丸で示されている「講義内容の分野や技術を良く知らず、講義内容の評価が低い」という条件を満たす表４のＥやＧの受講生の意見を参考にすることで、講義の不備が改善されると考える。さらに、「講義内容の分野や技術を良く知っており、講義内容の評価の高い」という条件を満たすＡやＢの受講生の自由記述の意見を取り入れることで、講義を高度な内容に発展させることができる。

表４より、受講生別適合度指数と評定の散布図を示す。（図10）

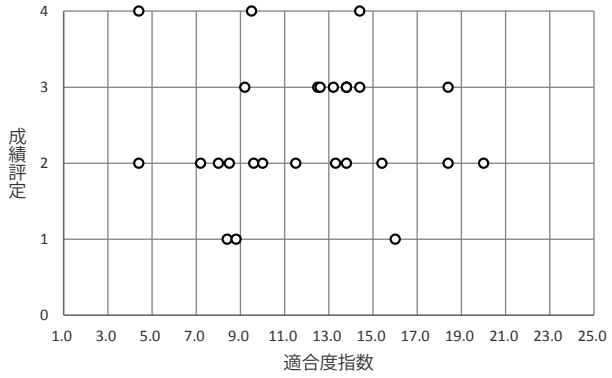


図 10. 受講生個別適合度指数と成績評価

図10において、結果が全体的に分散しており、適合度指数と成績評価は無関係であり、公平な成績評価が行われたと言える。

実践論科目は10回のオムニバス形式の講義を実施した。この講義別の適合度指数は、実施された講義内容や調査票の自由記述の結果と伴に、講義に対する評価基準とすることができる。（表５）

表 5. 実践論科目の講義別適合度指数

| 講義 | $\overline{W}$ | $\overline{E_s}$ | $M_l$ |
|---|---|---|---|
| E1 | 0.57 | 21.1 | 12.0 |
| E2 | 0.50 | 20.5 | 10.3 |
| E3 | 0.49 | 22.4 | 10.9 |
| E4 | 0.72 | 20.5 | 14.6 |
| E5 | 0.62 | 22.9 | 14.1 |
| E6 | 0.55 | 22.4 | 12.2 |
| E7 | 0.45 | 21.8 | 9.9 |
| E8 | 0.45 | 18.6 | 8.4 |
| E9 | 0.45 | 20.6 | 9.3 |
| E10 | 0.59 | 20.8 | 12.3 |

講義別適合度指数$M_l$が、８～11までの講義については、講師や講義内容の再検討を十分に行った。

特論科目は、データサイエンスの実習を実施した。特論科目の各回の適合度指数を示す。（表６）

表 6. 特論科目の講義別適合度指数

| 講義 | $\overline{W}$ | $\overline{E_s}$ | $M_l$ |
|---|---|---|---|
| S1 | 0.51 | 21.1 | 10.9 |
| S2 | 0.55 | 22.0 | 12.1 |
| S3 | 0.61 | 20.0 | 12.2 |
| S4 | 0.58 | 19.7 | 11.4 |
| S5 | 0.52 | 19.8 | 10.2 |
| S6 | 0.54 | 19.8 | 10.8 |
| S7 | 0.53 | 19.3 | 10.2 |
| S8 | 0.54 | 18.9 | 10.4 |
| S9 | 0.51 | 18.4 | 9.4 |
| S10 | 0.49 | 18.9 | 9.4 |
| S11 | 0.50 | 19.7 | 10.0 |

特論科目は、グループワークのデータサイエンスを行い、前半が座学、中盤から後半にかけてデータ分析実習と中間発表、最後に最終発表と言うプロセスを経た。難易度が高い課題を出したため、データ解析の行き詰まりによって、後半に行くに従い講義別適合度指数$M_l$が減少していると考えられる。

対象とした2016年度の人材育成プログラムは、受講生の選抜が行われ、ある程度の素養を持つ大学生、大学院生及び卒業の社会人が合格した。この受講生らによる調査票によって、精度の高い結果が得られたと考える。

## 3　結論

データサイエンスに関する講義内容に対する受講生の適合度指数を新たに定義し、分析した。

講義環境の構築は、講師や受講生の講義とは関係のない負担を軽減し不満を解消することで、講師と受講生が講義内容に対して集中力を維持することができた。受講生別適合度指数を分析することで、受講生個別の指導が行い易くなり、高い教育効果を得ることができると考える。講義別適合度指数は、講師や講義構成を考える上での指標となった。

以上の分析から、データサイエンスの最適な人材育成モデルを構築し、改善し続けるライフサイクルを形成することができたと考える。この結果、2017年度より、研究対象とした２科目は、大学院の正規履修科目として認められた。今後もさらなる研究を進め、より良い人材育成プログラムに発展させ、充実した人材育成を推進して行く。

## 謝辞

本研究の一部は、公益社団法人住友電工グループ社会貢献基金大学講座寄付の助成による。データアントレプレナープログラムに関わる運営委員の皆様、協力組織の皆様に厚く御礼を申し上げる。

この論文は、一般社団法人人工知能学会 第80回先進的学習科学と工学研究会 SIG-ALST-B507（2017/7）に掲載されたものである。

## 参考文献

[1] Gene Bellinger, Durval Castro, and Anthony Mills, “Data, Information, Knowledge, and Wisdom”, http://www.systems-thinking.org/dikw/dikw.htm, (2004)
[2] 清洲 正勝,「データサイエンス論」, 電気通信大学, p.10, (2016)
[3] 星野 敦子,「大学の授業における諸要因の相互作用と授業満足度の因果関係」, 日本教育工学会論文誌 Vol. 29 No. 4 p. 463-473, (2005)
[4] 山口 巧 他,「実務実習による学習意欲向上効果とその要因」, YAKUGAKU ZASSHI Vol. 134 No. 11 p. 1227-1235, (2014)
[5] 公益財団法人日本生産性本部 サービス産業生産性協議会,「顧客満足度指数（JCSI）」, http://consul.jpc-net.jp/jcsi/jcsi_index.html, (2007)
[6] Drew Conway, “The Data Science Venn Diagram”, http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram, (2010)
[7] Jeffry A. Timmons, “The Entrepreneurial Mind”, 1989. Fig. from Entrepreneurship in Atlantic Canadian University Environments Part I “Understanding Entrepreneurs: An Examination of the Literature”, (2011)
[8] 田村 元紀, 清洲 正勝,「データアントレプレナープログラム」, 電気通信大学, http://data-entrepreneur.com/, (2015)
[9] Turning Technologies, LLC., “TurningPoint”, “ResponseCard LT”, https://www.turningtechnologies.com/, (2016)
[10] 清洲 正勝,「データサイエンス選書」, 電気通信大学, http://booklog.jp/users/dep/, (2016)