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

データサイエンティストの採用・育成におけるハッカソンの活用

橋本 武彦[†1]

概要：GA technologies は 2013 年 3 月に設立された ReTech（RealEstate（不動産）×Technology の略．以降 ReTech）のスタートアップ企業です．設立から実質 5 年で売上 200 億円を達成し，2018 年 7 月に東証マザーズに上場しています．AI Strategy Center は 2017 年 4 月に不動産業界初の AI・データサイエンス組織として設立されました．設立 2 年弱ですが多くのメディアから取材を頂き，官公庁での登壇や複数の大学で講義を担当するなど注目を集めています．その中でもユニークな活動として，不動産データによるハッカソンを活用した採用・育成施策があります．データサイエンティストの需要に対して供給が追いついていない昨今，データサイエンティストの採用・育成は各社共通の課題です．AI・データサイエンスの組織立ち上げにおける経験や試行錯誤を踏まえ，採用・育成の観点でハッカソンの活用についてまとめました．

キーワード： AI，データサイエンス，データサイエンティスト，ハッカソン，Kaggle，ケーススタディ，OJT

1. はじめに

1.1 株式会社 GA technologies の紹介

株式会社 GA technologies は 2013 年 3 月創業のスタートアップです．

「テクノロジー × イノベーションで、人々に感動を。」という経営理念のもと，中古不動産に特化した流通プラットフォーム「Renosy」[1] の開発・運営や，自社の不動産業務を支援する Tech シリーズの開発や，Tech シリーズのノウハウをもとに他の不動産企業に業務支援の IT サービスを提供しています．他にも，不動産は建設，保険，金融とも密接な関係があり，これらの分野の AI・データ活用にも取り組んでいます．

創業から実質 5 年で売上 200 億円を達成し，2018 年 7 月に東証マザーズ上場，10 月に ReTech の雄である株式会社 ITANDI を M&A するなど，注目を集めています．

1.2 AISC の紹介

AISC は 2017 年 4 月に不動産業界初の AI・データサイエンス組織として設立されました [2].

ミッションは

1. 不動産ビジネスへの貢献
2. 要素技術の R&D
3. IR への貢献

になります．

設立 2 年弱ですが多くのメディアから取材を頂き，官公庁での登壇や複数の大学で講義を担当するなど注目を集めています [3-7].

1.3 組織立ち上げ時の課題

AI Strategy Center (AISC) 発足時はわずか 3 名でのスタートでした．目指す世界を実現するためにはリソースが圧倒的に足らない状況で，採用が急務でした．当時は上場前で情報発信を抑制していたこともあり，知名度が全くない中で採用の苦労が絶えなかったことをよく覚えています．

また，3 名の内 1 名は新卒でしたので，並行して早期に育成を考えていく必要がありました．

そこで着目したのが AI（特に機械学習）ブームに伴い，機械学習のコンペティションサイトとして注目度が上がっていた Kaggle の採用，育成への活用です．

2. ハッカソンとは

2.1 ハッカソンの定義と類型

wikipedeia によるとハッカソンの定義は以下になります．「ハッカソン（英語：hackathon，別名：hack day, hackfest, codefest）とは，ソフトウェア開発分野のプログラマやグラフィックデザイナー，ユーザインタフェース設計者，プロジェクトマネージャらが集中的に作業をするソフトウェア関連プロジェクトのイベントである」

著者はデータ分析のハッカソンは表 1 に示す 2 種に大別できると考えています．

表 1　データ分析のハッカソン

タイプ	用途	例
①課題設定型	物件価格の推定など，参加者が同一の課題設定の条件で，予測の精度などを競うタイプ	Kaggle Signate
②課題探索型	物件の販促施策立案など，同一のテーマやデータの上で，課題の探索から解決策の立案まで考えるタイプ	データ解析コンペティション データビジネス創造コンテスト

2.2 ハッカソンと育成スキル

データサイエンティスト協会では，データサイエンティ

†1 株式会社 GA technologies（連絡先：t_hashimoto@ga-tech.co.jp）

ストに求められるスキルセットを図 1 のように定義しています．また，図 2 にあるように，この 3 つのスキルは課題解決のフェーズによって，中心となるスキルが変化することを述べています [8].

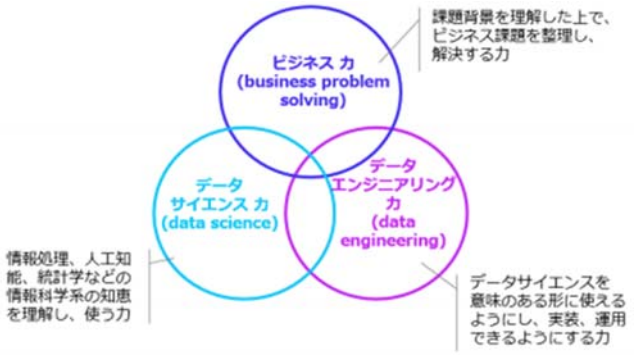

図 1　データサイエンティストに求められるスキルセット

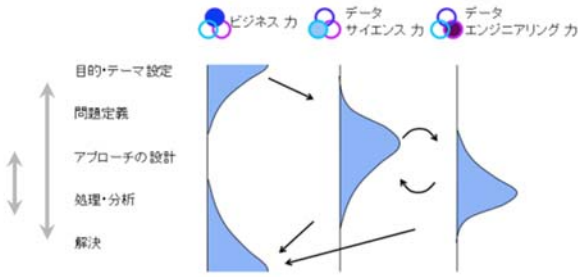

図 2　課題解決の各フェーズで要求されるスキルセットのイメージ

先程あげたデータ分析のハッカソンの 2 種類を図 2 のフェーズに重ねると，ハッカソンのタイプごとに求められるスキル（≒育成スキル）が異なることがわかります．

表 1 の①課題設定型ハッカソンは“アプローチの設計～処理・分析”に対応しておりデータサイエンス力とデータエンジニアリング力の育成が中心です。②課題探索型ハッカソンは“目的・テーマ設定～解決”に対応しておりビジネス力の育成にもつなげていくことができます．

2.3　GA technologies におけるハッカソンの活用

GA technologies では表 3 のハッカソン活用実績があります．

経緯としては新卒向けに行った①課題設定型のハッカソンをコンパクトにして，社内の AI・データ活用促進とエンジニアとの連携強化の狙いで機械学習に関心がある社内のエンジニア向けに実施したところ評判が良かったため，さらに大学での講義や短期インターンなど社外の方にも拡大していきました．

また①課題設定型のハッカソンではカバーが難しいビジネス力の育成や採用時の評価のために，②課題探索型のハッカソンも行っています．

表 3　GA technologies のハッカソン活用領域

目的	ターゲット	用途	ハッカソン
育成	新卒（+若手）	新卒研修	①課題設定型 ②課題探索型 *1
育成	エンジニア	社内勉強会	①　課題設定型
育成	大学生	講義内での演習	①　課題設定型
採用	就職活動生	短期インターン	①　課題設定型
採用	中途 *2	最終面接	②課題探索型

*1：新卒研修の②課題探索型ハッカソンはビジネス力とエンジニアリング力の育成を目的に，「社内向けプロダクト作成（プロトタイプ）」を実施．②課題探索型は受講者のレベルや自身で設定したテーマなどに応じて内容が変化しケースバイケースの要素が多いため，本論において詳細は割愛

*2：中途採用の場合，Kaggle 等のスコアがあれば書類選考時に確認．ただし，あくまで書類選考通過の一材料レベルで，人柄やカルチャーフィットを確認すべく，必ず面接は実施

次節では①課題設定型ハッカソンを例に，内容や実施時のポイントを説明します．

3.　ハッカソン活用事例紹介（採用）

3.1　①課題設定型ハッカソンの事例紹介

当社では現在，AI Booster の名称で 3 日間の短期インターンとして実施しています．下記が実際の募集要項とカリキュラムです．

不動産テーマで Prediction（予測）の練習問題として有名な Kaggle の House Prices [9] を演習課題に設定．3～4 人のチームで課題に取り組み，最終的に予測精度に加えて分析内容と発表内容を競います．

Kaggle に予測結果を投稿するたびに LeaderBoard に予測精度の Score と順位が表示されますので，チームごとに競い合うゲーミフィケーションの効果が発生し，毎回盛り上がっています．

表 4　AI Booster 募集要項

Title と概要	『AI Booster 2017（物件価格の予測に挑戦！）』 X-Tech 系の中で最も熱いリアルエステートテック（不動産 Tech）領域の物件価格の予測問題に挑戦してもらいます. 事業会社におけるデータ解析の実体験の機会を得られます.
対象者	大学以上（学士・修士・博士） ※理系，文系は問いません. ※就学状況は特に限定しないので 18 卒，19 卒，既卒，どなたでも応募可能です.
実施期間	2017 年 8 月 30 日（水）～9 月 1 日（金） 10:00−19:00
募集期間	7 月 14 日（金）～8 月 18 日（金） ※選考者の発表は随時
必要な（もしくはあると良い）スキル・経験	データ解析の経験，もしくは某かのプログラミング経験
場所	株式会社 GA technologies　セミナールーム
募集人数	4～16 名
エントリー方法	https://www.ga-tech.co.jp/news/27/
エントリー後のフロー	書類選考 ⬇ 選考通過者のみインターンシップ参加
体験できる職種	データサイエンティスト，AI エンジニア
参加にあたって	ノートパソコンをご持参頂ける方 （機種，OS は問いません）
こんな方におすすめ	1. 事業会社におけるデータサイエンス・AI 活用に関心がある方 2. データサイエンス・AI 活用のサービス企画に関心がある方 3. X-Tech 系の中で最も熱いリアルエステートテック（不動産 Tech）に関心がある方 4. 少量データから精度の高い意思決定という難易度の高いテーマにチャレンジしたい方 5. ウェブで完結するのでなく，リアルに影響を及ぼすビジネスを手がけたい方

カリキュラムは以下になります.

表 5　AI Booster カリキュラム

Time	1 日目		2 日目		3 日目	
	Program	講師	Program	講師	Program	講師
10:00	Internship の説明 参加者自己紹介 会社紹介 AI Strategy Center の説明	橋本 小林	作業方針討議	全員	House Prices 演習/資料作成	全員
11:00	ReTech（不動産 Tech）の動向	樋口				
12:00	Kaggle と今回の課題・データ説明 Python 基礎（入出力，集計，可視化）	橋本				
13:00	昼食		昼食		昼食	
14:00	Python 基礎（モデリング，クレンジング，欠損値処理，変数選択，アルゴリズム選択）	橋本	作業方針発表&Feedback House Prices 演習/資料作成	全員	House Prices 演習/資料作成	全員
15:00						
16:00	House Prices 演習 ※デフォルトの変数で Submit し，順位を共有	全員 *1				
17:00					各チームプレゼン（10 分×各チーム） 講師評 総評	全員
18:00					参加者打ち上げ	

*1 受講者数によるが講師以外に 2～3 名の TA を配置

3.2 重視点

開催時に重視しているのは以下のポイントです.

- 正しい分析プロセスの理解と体得
  - CRISP-DM [10]や PPDAC [11]など様々なデータによる問題解決のプロセスが提案されています. 講義内では，問題解決の一連のプロセスを回すことを最重視しています.
  - 小さなテーマでよいので，プロセスのサイクルを回すことを何度も繰り返し行うことが，問題解決のプロセス習得の近道と考えています.
- 座学でなく実践（“論より RUN”）
  - データサイエンティストとして目指すレベルによりますが，（特に初学者には）理論・数式といった座学から入るよりも実データを使って実践を繰り返すのが良いと考えています.
  - 1 日目の講義では，一連の問題解決のプロセスを説明した後，対応する最低限の Python コードを提供し，実行するにとどめています.
  - アメリカの著名データサイエンティストのネイト・シルバーも“So, getting your hands dirty with the data set is, I think, far and away better than spending too much time doing reading and so forth.”と述べています．また，筆者の知人は「論より RUN」と述べており，筆者も同意見です[12].
- 目的と手段を取り違えない
  - 受講者の中には Python ができる，機械学習ができる，ディープラーニングを使っている，などを価値と考える方もいます．“Python”や“機械学習”は問題解決のための手段に過ぎず，「ビジネスに如何に貢献するか」が価値であることを繰り返し伝えるようにしています.
  - 他にも「基本統計量や可視化，モデルの出力を算出することは作業であり，その結果から何を読み解くか重要」と伝え，例えば講義内でグラフを書く度にそこから何を読み取ったかなどを必ず発表してもらうようにしています.
- データ理解と前処理の重要性
  - 受講者の中にはデータ理解を十分に行う前に，すぐに機械学習のアルゴリズムを試したがる方も多いです. こういうタイプの受講者は，初期はスコアが伸びますが，一定ラインで頭打ちになります.
  - そういった際に「普段の仕事を振り返って考えるとデータの理解，加工が大半. 我々は実はデータサイエンティストでなく，データマエショリストです」と，データ理解やデータ加工など分析の前処理の重要性を伝えるようにしています.
- 分析の楽しさを知ってもらう
  - データサイエンティストの定義が浸透しきっていない影響もあるかと思いますが，多くの場合でデータサイエンティストに求めるレベルが高く，難易度設定が適切でないケースも見受けられます.
  - 少なくとも初学者には難易度を過度にインフレさせずに，まずは分析の楽しさを実感してもらうことを重視し，今後，データサイエンス・AI 領域に関わるきっかけとなればといつも考えています.

3.3 受講者の評価

受講者の評価は総じて高く，下図の受講者アンケート（無記名）の総合満足度評価は満足計（かなり＋やや満足）で 98%，TopBox（かなり満足）で 75%となっています.

開催時期	参加者数	かなり満足	やや満足	どちらともいえない	やや不満	かなり不満
全体	103	75%		23%		2%
2018/12	20	75%		20%		5%
2018/08_2	16	81%		19%		
2018/08_1	18	72%		28%		
2018/03	22	77%		23%		
2018/02	10	70%		20%		10%
2017/12	10	90%				10%
2017/08	7	43%			57%	

図 3　受講者アンケート結果（総合満足度：5 段階）

満足度評価の理由（自由回答）をみると，主な評価点は，

- Kaggle を通して分析の一連の流れを経験できた
- 機械学習，Python を実践的に体得できた
- 前処理の重要性を改めて理解できた
- 3 日間データに溺れることができ楽しかった
- 講師やメンターの指導やサポートがよかった
- チーム内で協力しあい，他チームと競い合えるのがよい

といった意見があった一方で，改善要望としては，

- 参加者間の機械学習，プログラミングレベルのバラツキを軽減すべく，事前課題を用意してほしい
- 時間が足りないので日数を延長してほしい

などが挙げられていました.

評価点の内容を見るに，当社が重視していることが受講者に伝わっていると考えています.

3.4 成果と課題

取組みの一環が学生の間に口コミで評判が広がっていったこともあり，参加者数も順調に増加しており，当初は集客施策に労力をかけていましたが，2018 年以降はほぼ集客

施策なしでも応募が集まるようになっています.

その結果，人数については社外秘のため伏せますが，本インターンを通じて採用目標人数を上回る方に入社いただけることになり，GA technologies AISC の新卒採用の施策として定着しています.

今後の課題として House Prices が扱う Prediction（予測）以外の分類などのテーマや，より実務に近いデータ，より多量のデータ（House Prices のデータは約 3000 件）を準備していくなどの必要があると考えています.

4. ハッカソン活用事例紹介（育成）

4.1 ①課題設定型ハッカソンの事例紹介

続いて育成における事例を紹介します．一昨年，昨年と配属された新卒に対し，短期インターンと同じく，House Prices をテーマに課題を設定しました.

前述の短期インターンとの違いは以下になります.

- チームでなく個人で取り組み
- 期間は 2 週間からで，分析結果を部内にレビューし，合格判定が出るまで課題に取り組んでもらいます.
  （合格までの期間は過去実績で 3 週間～6 週間）

4.2 重視点

短期インターンでの重視点に加え，以下を重視します.

- 分析の内容理解
  - 単に精度がよいだけでなく，分析内容の理解も重視します. なぜその処理か，なぜそのアルゴリズムを選定したかなどを説明してもらいます. よくわからないけどやったら精度が良くなったというのは，この場では評価されません.
- 発表に対するフィードバックの吸収
  - 中間レビューや合格をかけた最終レビューにおいて，レビュアーから多くの指摘があります. さまざまな指摘をどう受け止め，次回に反映できているかを見ています.
- 課題へ取り組む目線の高さ
  - 分析はやり始めるとキリがない部分もあります. 課題に対して，自身で Goal をどこに設定しているか？　当初設定した Goal に対する進捗の現状をどのように受け止めているか？　最終的に Goal に到達しなかった場合残課題をどう認識しているかなども見ています.

4.3 成果と課題

過去の受講者は卒業までの期間にばらつきはあるものの，一定の基準を満たし，全員合格することができています. 配属後，研修時の内容を忘れていることも度々ありますが，その際に「研修時に習っている」ことを指摘すると自身で振り返ることができ，両者の共通指針となっています.

課題としては①と同じくより実務に近いデータの用意と，レビュアー間での指摘レベルやフィードバックのやり方の統一などがあげられます.

5. まとめ

5.1 ハッカソンのタイプ別の特徴考察

過去の経験を踏まえタイプ別に特徴を考察したのが下表になります.

表 6　ハッカソンのタイプ別特徴

タイプ	PROS	CONS
①課題設定型	● データサイエンス力育成 ● ゲーミフィケーションの要素で競い合い，盛り上る	● ビジネス力の育成につながらない ● 課題作成に工夫が必要（リーク配慮） ● 提供可能なデータ準備が困難
②課題探索型	● データサイエンス力 ＋ ビジネス力の育成 ● PJ 配属前のケーススタディとして活用可能	● 短期実施に向かない ● 超初心者には向かない（受講者が一定のレベルに達している必要） ● 準備や運営が手間

5.2 今後に向けて

一番直近に開催した電気通信大学のデータサイエンティスト特論では上記考察を踏まえ，よりデータ件数を増やした House Prices 以外の Washington D.C.の Property Data（10 万件超）を用意し，Kaggle Inclass の形で開催し，好評を博しました.

①課題設定型はあくまで設定された課題をいかに解くかが主眼であり，何を解くべきか（Issue 選定）は育成スコープ外になります. 本来のビジネス成果を考えると両者の組み合わせが理想です. まず①課題設定型ハッカソンを短期間で実施し，その後可能であれば実際の実務データを用いて②課題探索型ハッカソンを長期間行い，データサイエンティストに求められるスキルの土台を統合的に養っていくのが理想と思っています.

日本のデータサイエンティスト・AI 人材の不足に対し，企業の立場から実践的なデータサイエンス・AI の経験機会を提供し、人材育成に貢献できればと考えています.

参考文献

[1] 中古不動産流通プラットフォーム Renosy（リノシー）
https://www.renosy.com/
[2] 『AI 戦略室』を新設，不動産業界で初となる自社内研究開発組織
https://www.ga-tech.co.jp/news/news/64/
[3] GA technologies 首都大学東京との共同研究を実施. 産学連携で不動産業界の業務効率化を目指し，業務時間最大 55%削減に成功
https://prtimes.jp/main/html/rd/p/000000017.000021066.html
[4] 総務省のパーソナルデータ活用に関する検討会に, データサイエンティストの橋本が登壇
https://www.ga-tech.co.jp/news/news/95/
[5] 「滋賀大学データサイエンス学部パンフレット」
https://www.ds.shiga-u.ac.jp/ds_ms_2018/wp-content/uploads/2018/06/7315b3a3af254876c4ef9148b95f1868.pdf
[6] 電気通信大学 データ関連人材育成のための研修プログラム開発・実施の受託とデータサイエンティスト特論(Advanced Data Scientist) 登壇のお知らせ
https://www.ga-tech.co.jp/news/news/1294/
https://www.ga-tech.co.jp/news/news/141/
[7] 人材不足とは無縁，採用が順調な IT ベンチャーの謎
https://tech.nikkeibp.co.jp/atcl/nxt/column/18/00138/080200119/
[8] データサイエンティスト協会, “データサイエンティストのミッション，スキルセット，定義，スキルレベルを発表,”
http://www.datascientist.or.jp/news/2014/pdf/1210.pdf
[9] House Prices: Advanced Regression Techniques
https://www.kaggle.com/c/house-prices-advanced-regression-techniques
[10] 機械学習によるデータ分析プロセス
http://topse.or.jp/wp-content/uploads/2015/06/420ed85f4cf9e3522861a6656e8ce3b5.pdf
[11] gacco 社会人のためのデータサイエンス演習（総務省統計局）
https://lms.gacco.org/courses/coursev1:gacco+ga063+2019_05/about
[12] Frick, W., “Nate Silver on Finding a Mentor, Teaching Yourself Statistics, and Not Settling in Your Career,” *Harvard Business Review*,
https://hbr.org/2013/09/nate-silver-on-finding-a-mentor-teaching-yourself-statistics-and-not-settling-in-your-career