



企業のデジタルトランスフォーメーション: ビッグデータ利活用に関する活動と課題

2017.1.25

クリエーションライン株式会社 シニアコンサルタント

木内 満歳

自己紹介



- 木内 満歳(きうち みつとし)
- クリエーションライン株式会社 シニアコンサルタント
- Slideshare: <http://www.slideshare.net/mkiuchi4>
- 各種寄稿
 - a. gihyo.jp: “[Mesosphere DCOSでつくるクラウドアプリケーション](#)”
 - b. 日経クラウドファースト2016年6月 “[Azure IoT Suiteの評価](#)”
 - c. Codezine: “[機械学習をクラウドで手軽に体験！ BluemixのApache Sparkで異常なセンサーデータを洗い出す](#)”
- 各種講演
 - a. Developer Summit 2016 Summer
 - b. 日経BP社 “パブリッククラウド導入の企画提案力養成講座”
 - c. Cloudweek Hokkaido 2015/2016
- 専門分野: Apache Mesos, Apache Spark, 分散コンピューティング, クラウドコンピューティング, NoSQL DB, グラフDB
- O’reilly Certified Developer on Apache Spark
- Docker Certified Technical Trainer



会社紹介

クリエーションライン株式会社

2006年1月設立

拠点： 東京都神田佐久間町(秋葉原)

社員数： 35(業務委託・BP含め 60人)

主な業務：

クラウド基盤コンサルティング・アプリケーション開発・運用

IoT/ビッグデータ基盤構築、データ分析サービス

アジャイル開発/DevOps開発/CI/CDに関するコンサルティング



取扱製品群



データ分析基盤



クラウド基盤・アジャイル開発支援

主要取引先

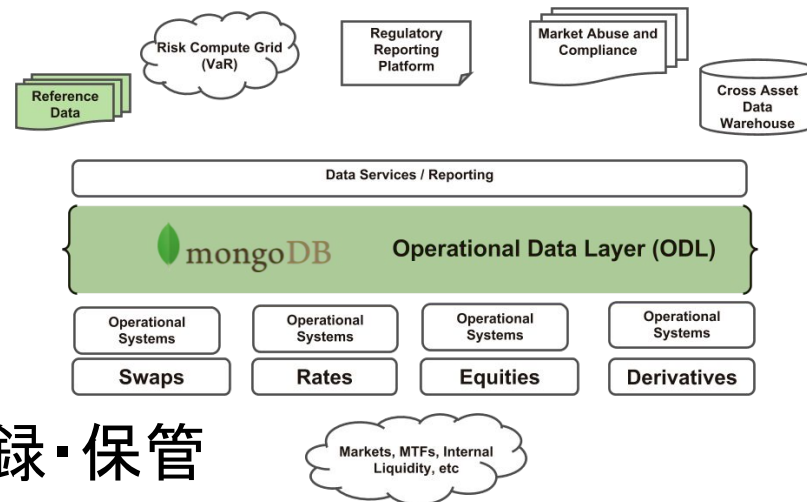
アルパイン株式会社
IzumoBASE株式会社
伊藤忠テクノソリューションズ株式会社
株式会社ウフル
エイベックス・グループ・ホールディングス株式会社
エヌ・ティ・ティ・コミュニケーションズ株式会社
KDDI株式会社
シトリックス・システムズ・ジャパン株式会社
ソフバンク株式会社
TIS株式会社
株式会社電通国際情報サービス
日本アイ・ビー・エム株式会社
ネットワンシステムズ株式会社
パナソニック株式会社
株式会社日立製作所
ブリジストン株式会社
日本マイクロソフト株式会社
三菱電機インフォメーションシステムズ株式会社
三菱電機ビルテクノサービス株式会社
(五十音順)

本題:ビッグデータ利活用に関する活動と課題

Agenda

- 海外におけるビッグデータ活用事例
- 国内企業の動向（主観的視点から）

海外におけるビッグデータ活用事例



金融における取引データ(証跡)の記録・保管

- データソースによるスキーマの違いを吸収

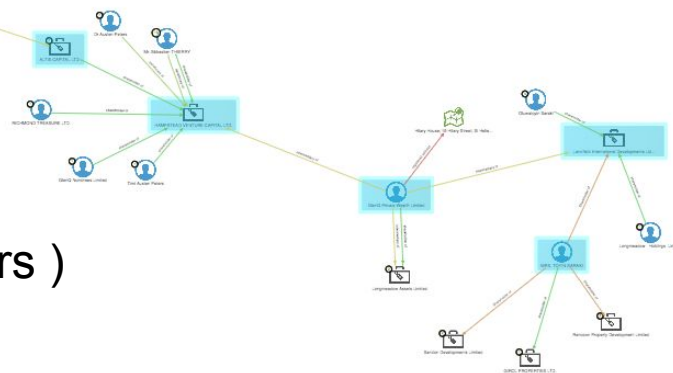
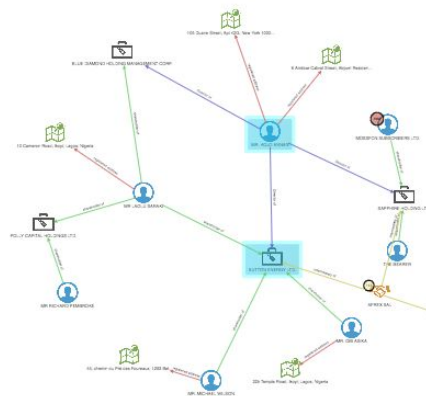
数百個の正規化テーブルを1つのデータベースに統合

- 厳密な型定義(int, long, float, double, string...)

[成果] 顧客の取引履歴を記録・保管し、バーゼルIIIに準拠

[成果] 複雑性とオペレーションリスクの低減

海外におけるビッグデータ活用事例



グラフデータによる不正取引の抽出

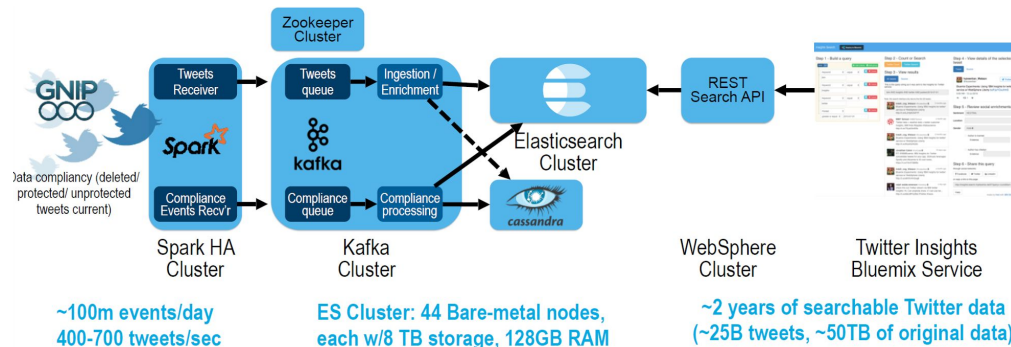
○ ICIJ Offshore Leak Database (含 Panama Papers)

○ 2.6TB, 1,150万ドキュメントの非定型データから

関係性をグラフデータとして構築

[成果] 140人以上の政治家、多数の企業家が不正取引に関与していたことを解明

海外におけるビッグデータ活用事例



ソーシャルネットワークの動向把握

- 1億ツイート/日(400~700ツイート/秒)のツイッター投稿を2年以上アーカイブ
- 自然語での検索、データマイニング、トレンドの視覚化

[成果] キーワードに対する自動アラーム、感情分析、個人の性格・嗜好などを分析

海外におけるビッグデータ活用事例

[そのほか]

- 先行事例検索データベース
- ホテル価格決定
- 配送ルート効率化
- 商品リコmend
- ユーザ権限監査
- ダークファイバルルート探索
- 請求統合



- オンラインチャット分析
- リアルタイム広告オークション
- コンピュータウイルストレンド検知
- センサー情報蓄積・分析
- 都市の移動データ蓄積・可視化

国内企業の動向(主観的な視点から)

- ビッグデータ利活用に関する興味は高い
- 海外の動向(ニュースになるような事例)は継続的に注視している
- いくつかの企業では実証実験(Proof of Concept, PoC)を始めている
- 実業務に取り入れて成果を出している企業は限られている

国内企業の動向(主観的な視点から)

- ビッグデータ利活用に関する興味は高い
- 海外の動向(ニュースになるような事例)は継続的に注視している



情報ソースは概ねニュースサイトやソーシャルフィード

国内企業の動向(主観的な視点から)

- いくつかの企業では実証実験(Proof of Concept, PoC)を始めている
- 実業務に取り入れて成果を出している企業は限られている
 - アドテクノロジー関連企業
 - 人材マッチング企業

PoCをはじめることができない
PoCから先に進むことができない
企業は少なくない

国内企業の動向(主観的な視点から)

- なぜPoCをはじめることができないか
 - 用語がわからない・リテラシーの不足
 - アイデアが思いつかない
 - 情報セキュリティ・プライバシー
- なぜPoCから先に進まないか
 - 評価できない
 - 効果を定量化できない

国内企業の動向(主観的な視点から)

- なぜPoCをはじめることができないか
 - 用語がわからない・リテラシーの不足



国内企業の動向(主観的な視点から)

- なぜPoCをはじめることができないか
 - 用語がわからない・リテラシーの不足

御社の次期販売支援システムに搭載するリコメンドエンジンの提案について協調フィルタリングアルゴリズムを活用したいと思っています。協調フィルタリングアルゴリズムとは Collaborative Filtering、CFのことであり、今回はモデルベースの交差確認によって汎化誤差を推定し、その汎化誤差で予測精度を評価いたします。通常の交差確認では、データを訓練用とテスト用の二つに分けますが、スーパーパラメータがある場合は、厳密には、データを訓練用 (training)、確認用 (validation)、およびテスト用 (test) の三つに分ける必要があり、御社の見積もりデータ及び発注データを使用し、RMSEを用いた精度検証を行いたいと思います。さらに弊社では

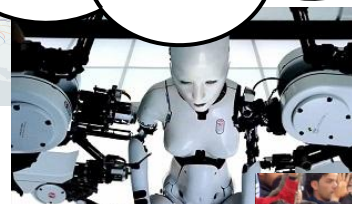
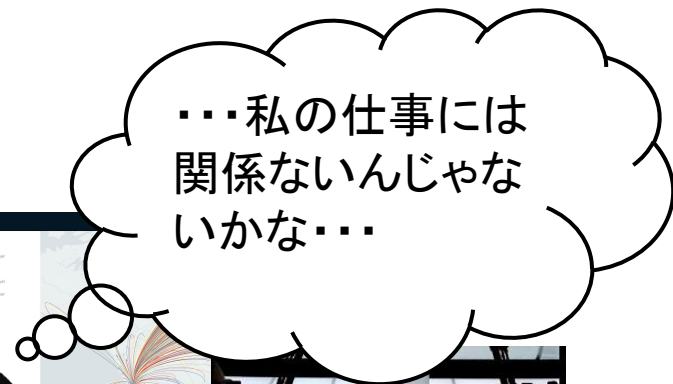
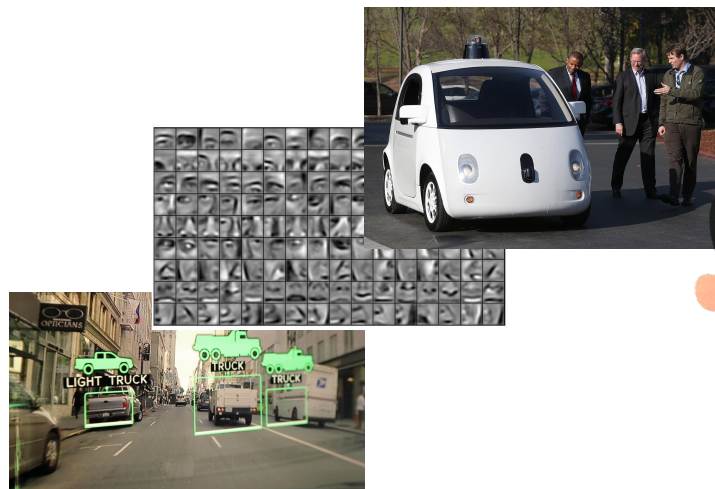


ざqwxせrcdvftybgんふいm
じよ、kp。l;@えxrcdvftybgんふ
いmじよ、kplwxれybgんふいえ
srdcftvygbんふいじもk、え
xsrdcvf

たちの悪いコンサルタントは専門用語を並べ立てて
ユーザをけむに巻こうとします

国内企業の動向(主観的な視点から)

- なぜPoCをはじめることができないか
 - アイデアが思いつかない



ニュースに出てくる事例は日々のビジネスにあまり関連のない
遠い世界の出来事 →目の前の業務に応用する方法を思いつくことができない

国内企業の動向(主観的な視点から)

- なぜPoCをはじめることができないか
 - 情報セキュリティ・プライバシー

セキュリティの制約からデータ
セットは出せません・・・
データサイエンティストを常駐さ
せていただけますか？



いわゆる2000個問題

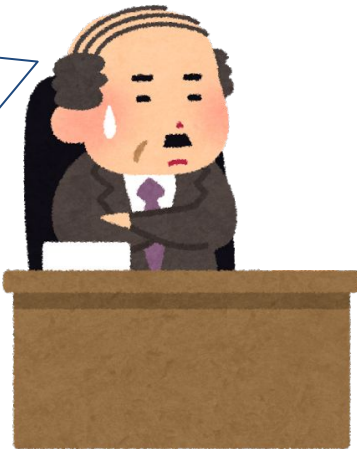
うちにはそんなに人材が
いないんですよ・・・



情報セキュリティ・プライバシーの制約からデータ分析業務を委託
できない場合が多く、エンジニアの常駐・常勤を求められる場合
が多い。一方データ分析を請け負う会社ではデータサイエンティ
ストが不足してるケースが多い

国内企業の動向(主観的な視点から)

- なぜPoCから先に進まないか
 - 評価できない
 - 効果を定量化できない



ビッグデータを分析
すると、どれだけ売
り上げが上がる
の？



さあ...

本番システムの
要件からは外そうか
な...

定量的な成果が証明できないと予算がつかないことが多い→取り組みが増えない
→実例が増えない→効果を上げるプラクティスが出てこない
...という負のスパイラル

まとめ

- 企業におけるビッグデータ利活用に関する興味は高く海外の動向(ニュースになるような事例)は継続的に注視しているが、身近な事例に乏しいため実際の利活用に至るケースはまだ多くない
- ビッグデータ利活用の前提となる、分析に関するリテラシーはまだまだ不足している。データとともに分析リテラシーを広めていく必要がある
- 情報セキュリティ・プライバシーの制約からデータ分析業務を委託できない場合が多く、エンジニアの常駐・常勤を求められる場合が多い。一方データ分析を請け負う会社ではデータサイエンティストが不足してるケースが多い
- いくつかの企業では実証実験 (Proof of Concept, PoC)を始めているが、明確なKPIを設定することができないために本格的な実用に踏み切っている例は少ない。現時点ではビッグデータ利活用において徒に費用対効果を追い求めるマイクロマネジメント的指向は多様な成長を妨げるかもしれない

(参考1) 弊社PoC実績のご紹介

アルパイン株式会社 (2015年9月～2016年2月)



ビッグデータ解析のオープンソースプロジェクトとして注目を集める「Apache Spark」を活用したビッグデータ分析の実証実験を共同で実施。

[成果]

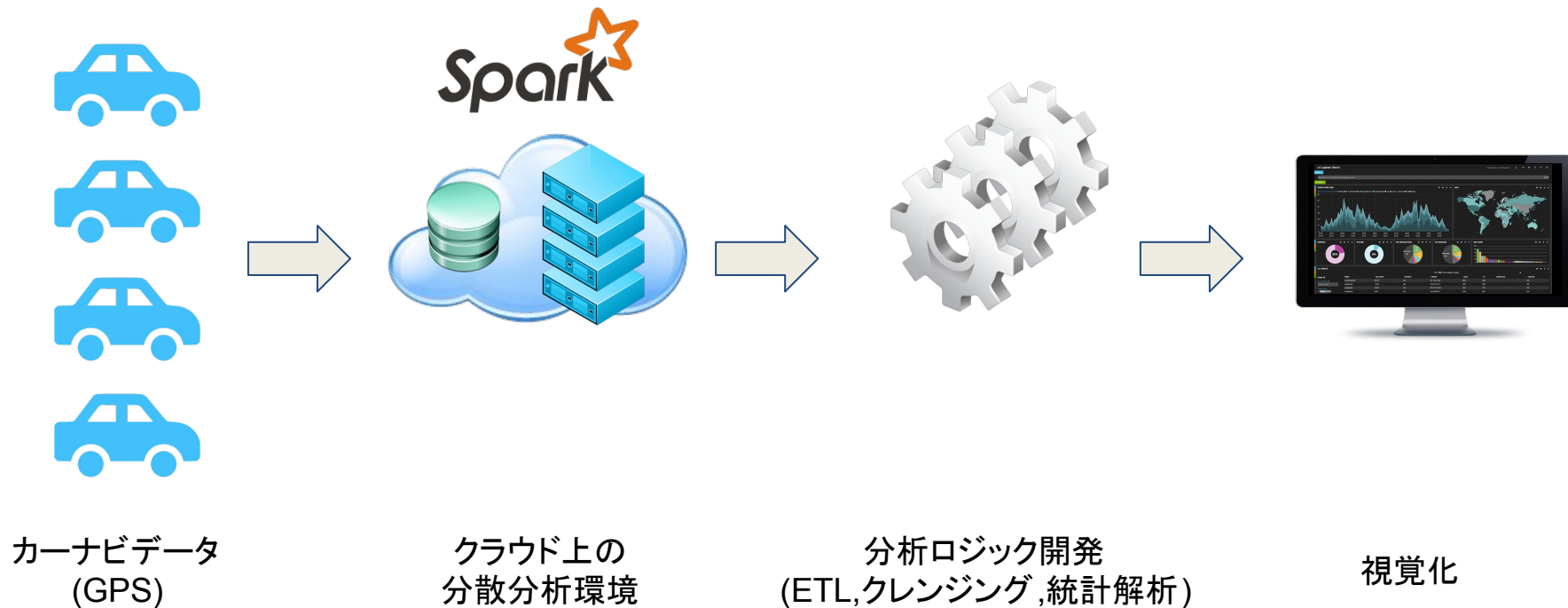
- 走行群データの視覚化
- データからユーザの利用パターンを判別
- 利用パターンの似ているユーザ同士をグループ化

[媒体資料]

<https://thinkit.co.jp/article/9626>



アルパイン様実績



(参考2) オープンデータのより広範な開放と整備のお願い

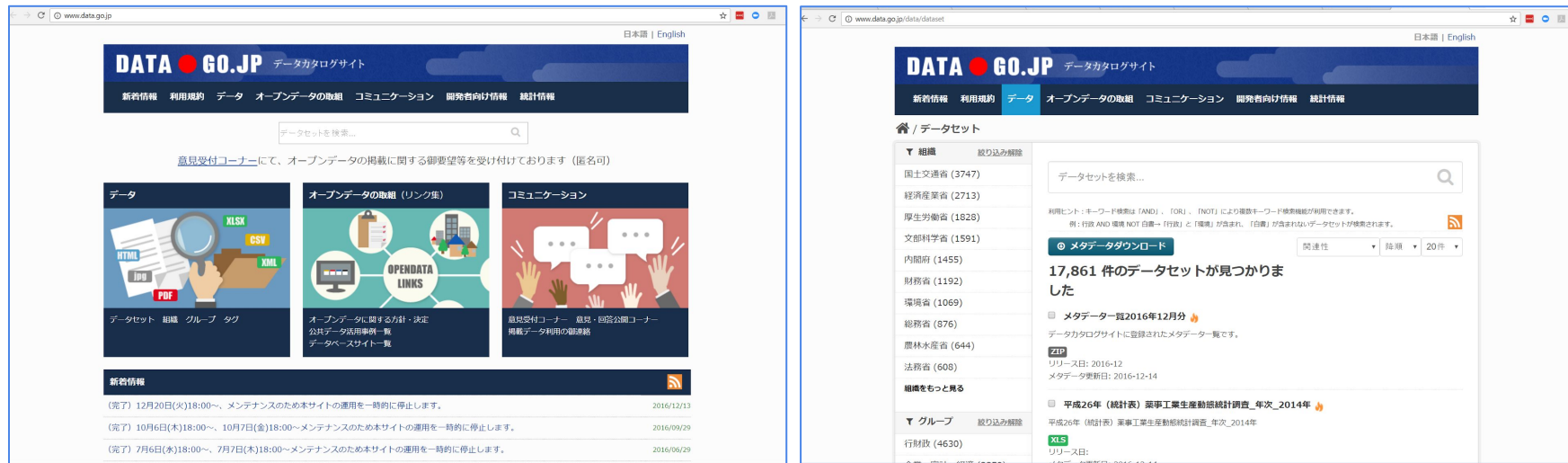
オープンデータのより広範な開放と整備のお願い

この場をお借りして、オープンデータについてお願いしたいことがございます
日本で自由に利用できるデータセットがとても少なく、難儀しています

より広範で自由なデータセット利用と整備についてご検討をお願いいたします

オープンデータのより広範な開放と整備のお願い

(例) DATA.GO.JP



約18,000件のデータセットが利用できるとあるが...

オープンデータのより広範な開放と整備のお願い

(例) DATA.GO.JP



CSVで利用できるデータは全体のわずか4%(約700件)
CSVなど、分析しやすいフォーマットの整備をお願いいたします

オープンデータのより広範な開放と整備のお願い

(例) 国立情報学研究所『情報学研究データリポジトリ』



Yahoo, 楽天などの購買データから、会話コーパス、テストコレクションなど極めて広範かつ有益なデータセット群

オープンデータのより広範な開放と整備のお願い

(例) 国立情報学研究所『情報学研究データリポジトリ』
→各データセット→使用条件

使用条件

国立情報学研究所ではヤフー株式会社との契約に基づき、研究者に対して「Yahoo!知恵袋」のデータを提供します。

提供の対象者は、NTCIR(情報検索システム評価用テストコレクション構築プロジェクト)参加者、及び本プロジェクトに関連する情報検索や自然言語処理などの分野の研究を行っている研究者となります。

利用目的は研究に限ります。また、研究成果の発表等に一定の条件があります。

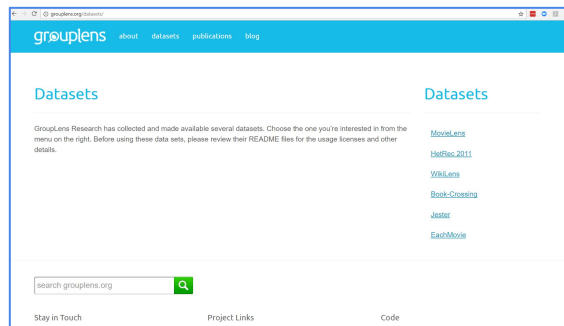
詳細は下記必要書類の「データ取り扱い上の注意事項」及び「覚書(サンプル)」をご覧ください。

・・・提供の対象者は、NTCIR(情報検索システム評価用テストコレクション構築プロジェクト)参加者、及び本プロジェクトに関連する情報検索や自然言語処理などの分野の研究を行っている研究者となります。利用目的は研究に限ります。また、研究成果の発表等に一定の条件があります。

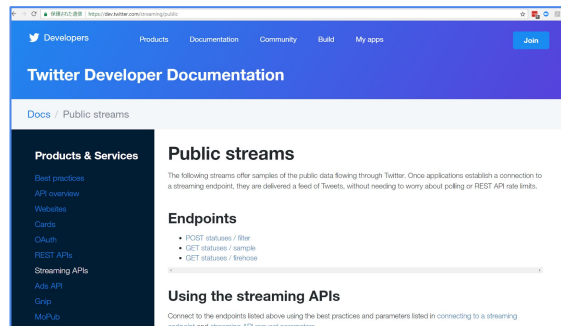
利用できる対象者が極めて限られている
より広いユーザへの開放をお願いいたします

オープンデータのより広範な開放と整備のお願い

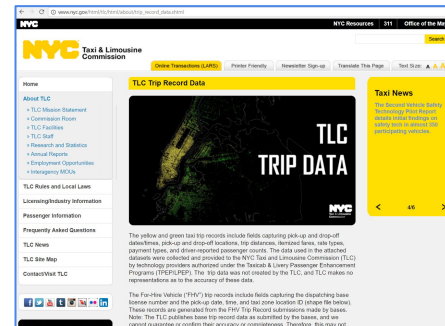
海外では官民間問わず、あまり制限なく利用可能である公開が一般的



Grouplens
(アイテム評価データセット)
(ミネソタ大学)
<http://grouplens.org/>

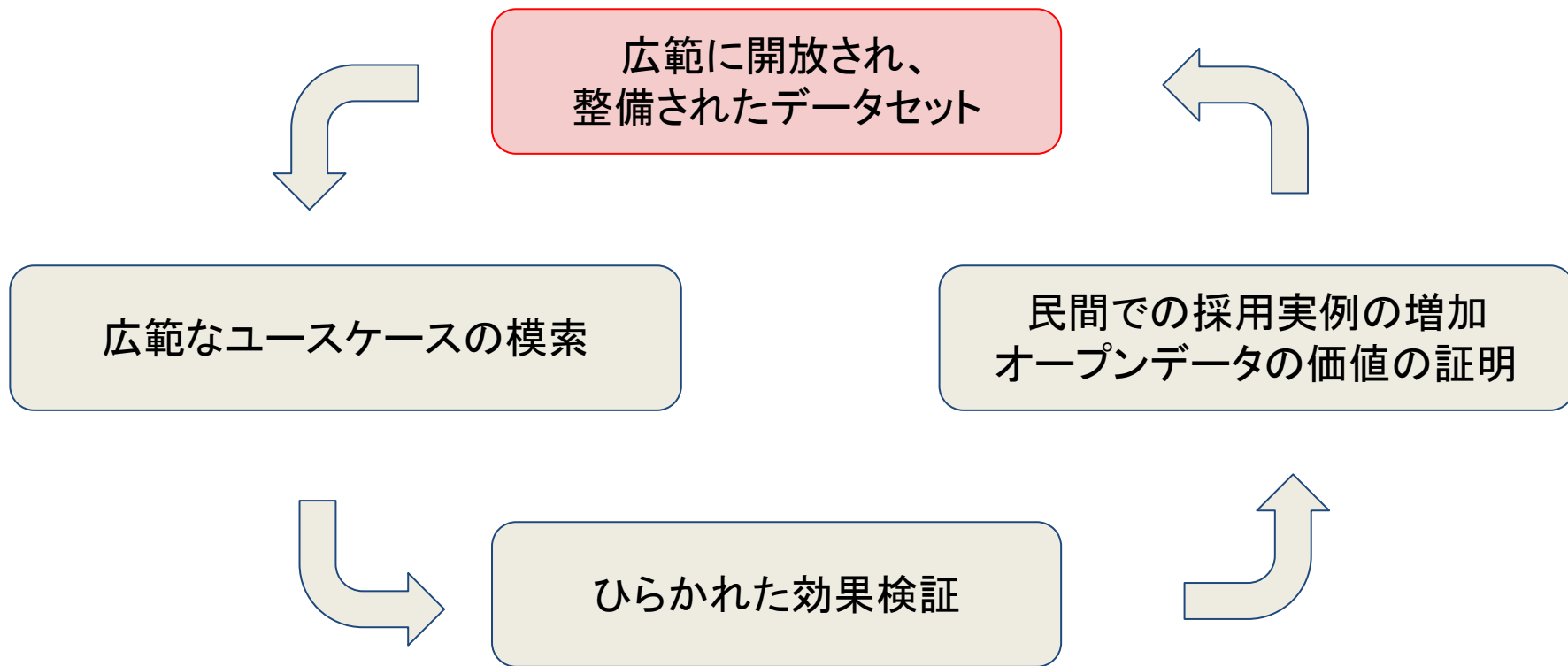


Twitter Public Streams
(Twitter社)
<https://dev.twitter.com/streaming/public>



Newyork Taxi Trip Data
(ニューヨーク市)
http://www.nyc.gov/html/tlc/html/about/trip_record_data.shtml

オープンデータのより広範な開放と整備のお願い





CREATIONLINE, INC.