最近のビッグデータ活用事情

Recent Big Data Utilization Circumstances

2014．11．15(SAT)

高井　正三(Shoso Takai)

富山大学総合情報基盤センター

Information Technology Center，University of Toyama

IT Executive Forum IT Japan 2014からの報告，講演者の講演タイトル･内容からのキーワード

- ICT(Information and Communication Technology)の活用
- イノベーションInnovation／変革するInnovate
- **ビッグデータBig Data活用**
  - マツダのモノ造り革新とビッグデータ活用による更なる進化
  - ビッグデータを競争優位につなげる
  - ビッグデータとIoT～不連続な世界～
- DBaaS：Database as a Service
- IoT：Internet of Things
- Digital Marketing
- WhiteColor Productivities
- Work Style
- Leadership

Globalな世界で活躍するために必要なこと

日本GE，GE Capital CEO，　安渕　聖司氏のIT JAPAN 2014
2日目の基調講演「世界で勝つリーダーシップ」から抜粋

- Who are you?
  - あなた誰？
  - 揺るぎない軸･価値基準を持つこと
- Where are you interested?
  - どこに興味がありますか？
  - 強い好奇心を持ち，広く学び続けること
- What do you think?
  - どう思う？
  - 自分の頭で考え，自分なりの意見を持ち，それを堂々と披露すること
- How do you come up with ideas?
  - どのようにアイデアを思い付くのですか？
  - 多様性を受け入れ，積極的に活用し，新しいアイデアや考えを育むこと
- Why don't you speak up?
  - なぜあなたは発言しないのですか？
  - 母国語と英語による積極的でオープンなコミュニケーションに参加すること

目　　次

１．身近なビッグデータ活用例他

１－１　アマゾンAmazon.comの例
１－２　最近の日本経済新聞とIT Proページから
１－３　日経BPから「日経ビッグデータ」創刊
１－４「統計学が最強の学問である」から
１－５「世界でもっとも強力な9つのアルゴリズム」
　　から

１－１　アマゾンAmazon.comの例

「ビッグデータの正体」と言う本を買った後，再度検索してみたところ，・・・

1-1-2　Amazon.comの例２

１－２　最近の日本経済新聞とIT Proページから

（日経新聞：2014.10.07）

1-2-2　最近の日経産業新聞とIT Proページから


日経産業新聞
2014年(平成26年)10月10日(金曜日)

ビッグデータで便利に

トヨタやホンダ

日経産業新聞　2014.10.10

Nikkei IT Pro 2014.11.05

1－3　日経BPから「日経ビッグデータ」創刊

1－4　「統計学が最強の学問である」から

- ◆ビジネス領域における統計学を応用したSolutionのことを“Business Intelligence”という
- ◆統計学を制する者が世界を制する
  - 「疫学の父」ジョン・スノウの活躍
  - 「EBM（Evidence-Based Medicine：科学的根拠に基づく医療）」が最も重要な考え方
  - Googleのチーフ・エコノミストであるハリ・ヴァリアンは“I„keep saying the sexy job in the next ten years will be statisticians.”
- ◆「ビッグデータ」という言葉が流行るわけ
  - ・・・どんな大量のデータでも，どんな計算でもできる技術ができた今，何を計算すべきかを考えると，統計解析以外にはありえない．
  - そしてもし，「統計解析」という地味な言葉がお題目として魅力的でないのなら，「ビッグデータ」とか「ビジネス・インテリジェンス」といった，はやり言葉を生み出せば良い．
- ◆統計家が見たビッグデータ狂想曲/狂想曲をもりあげる専門用語
  - 「Exadata, Hadoop, NoSQL, KVS, Data/Text Mining」

1－5　「世界でもっとも強力な9つのアルゴリズム」から ビッグデータに関係ありそうなもの抜粋

- ◆第2章　インデクシング　Indexing（検索エンジンの）
  - Google Searchで「デジタル　一眼レフ　カメラ」と入力して，検索ボタンを押下すると，「約 752,000 件（0.25 秒）」とヒットを表示する
- ◆第3章　ページランク　Page Rank
  - Google Searchで使用しているページ順位を表示するアルゴリズム
  - もっともよく引用されているページは良いページ？
- ◆第6章　パターン認識　Pattern Recognition
  - 監視カメラでの顔認識やSiriなどの音声認識，・・・
- ◆第7章　データ圧縮　Data Compression
  - ビッグデータをDiskに保管できるようになった
- ◆第8章　データベース　Database
  - 一貫性（複数のアプリケーションが同時にデータベースを更新した結果，片方は更新したが，もう一方は更新されなかったということがないこと）の追求　RDBMS
  - ビッグデータでは一貫性が厳密に維持されていない　NoSQL，・・・

２．ビッグデータに関する最近の話題

◆IBMとTwitterの連携でWatsonが果たす役割
- 2014/10/30 ITpro NOW 松本 敏明＝日経コンピュータ
- 米IBMは2014年10月29日（米国時間），米ツイッターとビジネス向けビッグデータ解析で提携すると発表
- Twitter上のつぶやきを分析し，活用する業務アプリケーションを，銀行や消費財などの各業界に向けて開発
- IBMの人工知能Watsonの分析技術cognitive computing 認知計算で「つぶやき」データをビジネスに応用

◆米メディア界の構造が変わる，俳優ケヴィン・スペイシーが語るビッグデータの破壊力
- 2014/11/04 News 浅川 直輝＝日経コンピュータ
- ビッグデータをメディアビジネスに生かした著名な成功例の一つに，2013年に米ネットフリックスが独自に製作した政治ドラマシリーズ「House of Cards（邦題：ハウス・オブ・カード 野望の階段）」がある

◆「Tポイント」新規約施行へ，個人情報の第三者提供停止には手続きが必要，11月1日から適用．
- 2014/10/30 News 清嶋 直樹＝日経コンピュータ
- 「共同利用」から「第三者提供」に，CCCがT会員規約を大幅改訂

2-1 クイズ王に勝ったコンピューターIBM Watson

2011年2月14日～16日の3日間，アメリカ合衆国の人気クイズ番組「Jeopardy!（ジョパディ！）」で行われたクイズ王対決の最終的な成績は，IBMのSupercomputer Wastonが7万7147ドル，クイズ王のケン・ジェニングス氏は2万4000ドルで，ブラッド・ラッター氏は2万1600ドルだった．

2.6GHzのPower7 CPU Coreを2880個（32Core×90Server）搭載したWatsonは，
1台だと1問に解答するのに2時間を要するが，Watsonは3秒で解答を出すことができるという．

http://blogs.itmedia.co.jp/marron/2011/01/watsonhal-4530.html
http://www.youtube.com/watch?v=qOpW5VN2j2Q&NR=1

2-2. ビッグデータに関する最近の話題２

◆利益が上がる「値決め」をビッグデータで支援，デロイトとSAPが協業
- 2014/11/05 News & Trend 鈴木 慶太＝日経コンピュータ
- デロイト・トーマツ・コンサルティング（DTC）とSAPジャパンは2014年10月，SAP製ERP（統合基幹業務システム）の製造業ユーザー向けに製品や部品・材料の価格決めを支援するソフトとサービスを提供することで協業すると発表
- ERP（Enterprise Resource Planning ）から得られる大量の自社データや競合他社のデータなどを収集・分析し，現場の担当者がタブレットなどを使って最適な価格を設定できる

◆[車載情報端末の未来3]最終形はクルマ全体の高度なコンピュータ化
- 2014/10/31 ITpro Report 竹居 智久＝日経コンピュータ
- 「デジタルコックピット」・・・車内のあちこちに置かれたディスプレーや車内外の状況を検出する多数のセンサーなどを利用しながら，運転や乗車の体験を向上，高度運転支援
- SmartPhone連携，通信機能を備え，新しいサービスを提供

３．格差広げるビッグ・データ１００から活用事例

（第１部）日経コンピュータ　2014.07.24から

３－１　先進事例
知らぬ間に不満を解消

３－２　データ・サイエンティスト Data Scientist
データよりも現場を愛す

３－３　注目製品・サービス
やりたいことは大抵できる

３－４　最新技術
貯めたデータが大化けする

3－1． 先進事例知らぬ間に不満を解消

001 博物館の“人流”をセンサーで全記録
国立科学博物館／乃村工藝社／日立製作所
見学ルートの改善，子供と大人の展示解説を分ける
002 メール分析で男女カップルを増やす
えひめ結婚支援センター
003 家電の利用状況を収集，ピーク時の省エネ支援
三井不動産／東芝
004 ゲーム感覚で学習習熟度をITで分析
DeNA
005 名医の技を学習，外科手術も自動化？
手術支援ロボット「da Vinci」･･･いずれ，ロボットが外科手術
006 がんを早期発見，類似症例を高速検索
富士フィルム

001 博物館の“人流”をカード型センサーとビーコンで全記録し，ルートを改善

3-1-2. 知らぬ間に不満を解消２

◆007 検索キーワードを基にヒット商品を開発
■ カゴメ／アマゾンジャパン･･･リコピンを増やしたプレミアムレッド
◆008 クール機材不足を配送予測で撲滅へ
■ ヤマト運輸
◆009 希望に「マッチ」した就職先を紹介
■ エン･ジャパン
◆010 波浪や天候を読みコンテナ船燃費5%向上
■ 日本郵船
■ コストを200億円削減･･･船齢や季節に応じて，最適な運行方法を船毎に提案
◆011 2ミリ秒でデータ収集，歩留まりを向上
■ オムロン／富士通
◆012 島根原発を徹底監視，故障予兆を検知する
■ 中国電力／NEC

3-1-3. 知らぬ間に不満を解消３

013 70万社の取引を分析，優良銘柄を集中支援
帝国データバンク／中小企業庁
014 渋滞予測精度を向上，幹線道路以外も対象に
NTTドコモ／パイオニア
015 価動きある銘柄予想，投資家の強い味方に
カブドットコム証券
016 100万台の複合機から毎日数千万件収集
キャノン･･･目的は故障を素早く察知して保守業務を効率化
017 犯罪発生地点を事前に予測
米サンタクルーズSanta Cruz市警･･･予測警備，高い的中率

017　犯罪発生地点を事前に予測
「プレディクティブ・ポリシング」を導入したサンタクルーズ市警

写真1●「プレディクティブ・ポリシング」を導入したサンタクルーズ市警

写真2●サンタクルーズ市警で副署長（Deputy Chief of Police）を務めるスティーブ・クラーク（Steve Clark）氏

写真3●サンタクルーズ市警のパトカー車内

Nikkei IT Pro
http://itpro.nikkeibp.co.jp/atcl/watcher/14/334361/080100020/?SS=imgview&FD=1124500606&ST=bigdata

017-2 プレディクティブ・ポリシング
Predictive Policing=予測警備

- 2011年7月．米カリフォルニア州サンタクルーズ市で不思議な現象が起こった．犯罪が発生する前に，犯罪現場に警察官が現れるようになったのである．
- それから3年，同市では実際に犯罪発生件数が17%も減少したという．
- Repeat Victimization（一度被害にあった場所で2週間以内に被害が再発するという傾向）
- Near Repeats（犯罪が発生した近郊で犯罪が再発しやすいという傾向）
- サンタクルーズ市警は2011年7月に，モラー博士らが開発した予測モデルを搭載した犯罪予測システム「**PredPol**」を導入した．

017-3 プレディクティブ・ポリシング
Predictive Policing=予測警備

「車上荒らし（Vehicle Burglary）」「住居への強盗（Burglary）」「自動車窃盗（Auto Theft）」「拳銃やナイフを使った犯罪（DW Assault，DWはDeadly Weaponの略）」「拳銃などを使わない暴行（Battery）」といった犯罪が昨日どこで発生し，これからどこで発生しそうか地図上に表示する。

写真4●カリフォルニア大学ロサンゼルス校（UCLA）のジェフ・ブランティンガム（Jeff Brantingham）博士が開発した犯罪予測モデル

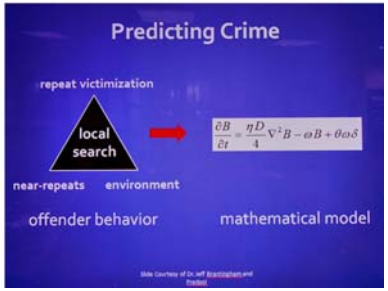
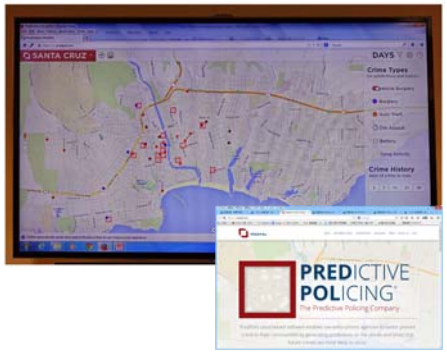


写真5●「PredPol」による犯罪発生予測画面



３－２．Data Scientistデータよりも現場を愛す

018　現場との対話が気付きを生む
　花王
019　ゲームの遊ばれ方，BIツールで可視化
　セガ
020　マーケ部隊が分析，強い売り場を作る
　ファイミリーマート
021　個別最適なカタログ，購買分析で送付
　ディノス・セシール
022　自販機オペレーターにデータの活用を促す
　JR東日本ウォータービジネス
023　システム部門の仕事にデータ分析で本業強化
　LIXIL

3-2-2. データよりも現場を愛す

024 ソーシャルメディア分析，独自にデータを蓄積
データセクション
025 独自ビッグデータ，人海戦術で生成
エム・データ
026 24万人のブログを分析，購買行動を探る
富士通総研
027 「宝の山」はどこだ．異業種で共有・活用
データエクスチェンジ・コンソーシアム
028 経産省が活用後押し，情報共有の協議会発足
データ駆動型イノベーション創出戦略協議会
029 人材育成が急務としてサイエンティストを定義
CSAJ（コンピューター・ソフトウェア協会）

3-2-3. データよりも現場を愛す

030 データサイエンティスト，職種として確立目指す
データサイエンティスト協会
031 データ分析チームを立ち上げるサービス
EMCジャパン
032 分析スキルを習得する複数の育成コース
日本サード・パーティ
033 5日間の人材育成コース，青果物見て資格授与
チェンジ
034 買い手と売り手が同席，POSデータの分析講座
コープさっぽろ
035 人材育成が前提の無料の統計学講座
NTTドコモ／NTTナレッジ・スクエア

3－3．注目製品・サービスやりたいことは大抵できる

高速分析編

036 統計解析の王道，多彩な分析できる
■ SAS9.4／SAS Institute Japan
037 知りたいことを予測して分析
■ SAP InfiniteInsight／SAPジャパン
038 統計解析ソフトを無償で利用可能に
■ R／日本オラクルなど
039 老舗の統計解析ソフト，学術分野で利用進む
■ SPSS／日本IBM
040 S言語の商用実装，豊富なデータ解析機能
■ S-PLUS／NTTデータ数理システム

3-3-2 注目製品・サービスやりたいことは大抵できる

人工知能編

041 ドライバーの目的地を予測する
■ T-Connect／トヨタ自動車
042 怪しいメールを機械が監視する
■ Lit i View EMAIL AUDITOR／UBIC（内部監査支援サービス）
043 人工知能がアプリ開発を支援
■ AppDataBank／メタップス
044 機械学習を容易にできる
■ Azure Machine Learning／日本マイクロソフト
045 人と会話ができるコンピューター
■ IBM Watson／日本IBM
046 日本生まれの機械学習ソフト
■ Jubatus／プリファード インフラストラクチャー

3-3-3 注目製品・サービスやりたいことは大抵できる

可視化編／Business Intelligence（BI） Tools

047 ビジュアルな表現力が特徴
- Tableau／Tableau Japan

048 異なるデータの連想結合が強み
- Qlik View／クリックテック・ジャパン

049 インメモリーDBと連携
- Business Object／SAP ジャパン

050 小田急電鉄が利用
- Cognos／日本IBM

051 スシローが30億件分析
- MotionBoard／ウィングアーク1st

052 タブレットから操作できる
- Oracle BI／日本オラクル

3-3-4 注目製品・サービスやりたいことは大抵できる

可視化編／Business Intelligence（BI） Tools 2

053 独立したOSS（Open Source Software）を統合
- Pentaho／米ペンタホ

054 Excelのアドオン
- Power BI／日本マイクロソフト

055 専門知識は不要
- SAS Visual Analytics／SAS Institute Japan

056 ログデータ分析が得意
- Splunk／Splunk Services Japan

057 研究開発分野で導入進む
- TIBCO Spotfire／日本ティブコソフトウェア

058 定型クエリーが豊富
- WebQuery／システムコンサルタント

059 クラウド型BIサービス
- xica adelie／サイカ

3-3-5 注目製品・サービスやりたいことは大抵できる

主なビッグ・データ・サービス

060 購買履歴データを提供
- Audience Operation System／アクシオム ジャパン

061 ピンポイントの天気が分かる
- HalexDream!／ハレックス

062 SNSから評判を収集する
- InsightCather／NTTコムウェア

主なデータ・ウェアハウス・サービス（DWH）製品

063 良品計画などが活用
- Amazon Redshift／アマゾン・データ・サービス・ジャパン

064 バッチ，ストリームに両対応
- Google Cloud Dataflow／米グーグル

065 FPGA（Field Programmable Gate Array）を活用
- IBM PureData System for Analytics／日本IBM

3-3-6 注目製品・サービスやりたいことは大抵できる

主なデータ・ウェアハウス・サービス

066 大企業での導入実績増やす
- Oracle Exadata／日本オラクル

067 佐川急便が導入
- Pivotal Greenplum Database／Pivotal ジャパン

068 Hadoopと連携しやすい
- SAP Sybase IQ／SAPジャパン

069 RDBが進化
- SQL Server 2014／日本マイクロソフト

070 米ウォルマートで実績
- Teradata／日本テラデータ

071 クラウド型DWH（Data WareHouse）
- Tresure Data Sevice／米トレジャー データ

3－4．貯めたデータが大化けする

◆072　あらゆるデータを活用可能な状態で貯める
- データレイク

◆073 機械が人間に頼らず，全自動で知識を習得
- ディープラーニング

◆074　大量データを“探検”，道の事実を発見
- データエクスプロレーション

◆075　機械単独で猫を発見，単語で画像検索OK
- 画像認識

◆076　火星探索ロボット，地図なしで自律走行
- ロボット制御／NASA

◆077　単語を機械が理解，王－男＋女＝？
- 自然言語処理／米Google

3-4-2．貯めたデータが大化けする

◆078　顔向きが変わっても同じ顔として認識
- 3次元顔認識／Facebookの「DeepFace」

◆079 人の会話を認識，同時通訳まで行う
- 音声認識／米Microsoft

◆080　稼働情報や気温から電力効率を予測
- データセンターの電力制御／米Google

◆081　MapReduce以外のビッグデータ処理に対応
- Hadoop2／Apache

◆082　Hadoop2の心臓部，タスク管理を司る
- YARN／Yet-Another-Resource-Negotiator

3-4-3．貯めたデータが大化けする
Hadoop2向けの主なリアルタイムSQLエンジン

083　BIツールとの連携が容易
- Big SQL・・・米IBMが開発するMPP（超並列処理）型SQLエンジン
- ApplicationからDBを利用するための「JDBC」「ODBC」インターフェースを備えており，既存のBIツールから利用するのが容易である．

084　1万ノードに対応
- Drill・・・米Apacheソフトウェア財団ASFが開発するMPP型SQLエンジン（開発の中心は米MapR Technologies）
- 米Googleが開発した「Dremel」が手本．最大1万ノードで稼働可能．

085　DWHベンダーが開発
- HAWQ・・・米Pivotalが開発するMPP型SQLエンジン
- 「PostgreSQL」ベースのDWH「GreenPlum」で使われているMPP型SQLエンジンをHadoop向けにしたもので，クエリーの実行最適化アルゴリズムなどがGreenPlumと同じ．

086　開発言語は「C++」
- Impala・・・HadoopのDistribution Venderである米クラウデアが開発したOSSのMPP型SQLエンジン．
- 米Googleが開発した「Dremel」が手本．既存の「Hive」との文法互換性を維持する．

3-4-3-2．貯めたデータが大化けする２
Hadoop2向けの主なリアルタイムSQLエンジン

087　HBase用に特化
- Phoenix・・・ASFが開発．分散データベース「Hbase」用SQLエンジン
- 元々の開発は米SaleForce.comで，ソース・コードをASFに寄贈．
- HbaseでSQLを利用できるようにした．
- Hbaseは低遅延で一貫性のあるデータ更新ができる．

088　Facebookが毎日利用
- Presto・・・米Facebookが開発するOSSのMPP型SQLエンジン
- Facebookでは2013年から1000ノード以上で使用しており，1日に処理しているクエリー件数は，データ1PBにつき3万件に達する．

089　開発元は韓国スタートアップ
- Tajo・・・ASFが開発するMPP型SQLエンジン
- SQLクエリーを機械語にJIT(Just In Time)コンパイルすることで処理を高速化した．
- 元々の開発元はかんこくのスタートアップであるグルター

3-4-4. 貯めたデータが大化けする

090 処理を細分化し同時実行．次世代並列処理の本命
- DAG（有向非循環グラフDirected Acyclic Graph）＝トラック輸送
- ○データを小分けにして運び（処理し），運び終わるまでの時間が短い
- ○交通状況などに応じて最適なルートを選べる
  （パイプライン処理の最適化が可能）

091 インメモリー処理に強み．Yahoo!，1200台で利用
- Spark（DAGエンジンの一つ）

092 既存プログラムを高速化．MSが開発に協力
- Tez（HadoopのDistribution（検証済みパッケージ）を手がける
- 米ホートンワークスが開発）

093 ビッグデータの元祖．Googleの技術が手本
- MapReduce＝貨物列車輸送
- ×データを一気に運ぶ（処理する）が，時間がかかる
- ×予め敷設された線路上しか走行できない
- （パイプライン処理の最適化が難しい）

3-4-5. 貯めたデータが大化けする SparkやTezで動く周辺機能

094 汎用的なJavaプログラムを実行
- Cascading・・・MapReduceを使わない汎用的なJavaプログラムをHadoop上で実行するための仕組み
- 複雑なパイプライン処理を記述するためのAPIを備える．

095 グラフ処理をメモリー上で実行
- GraphX・・・SNSにおける友人同士のつながりの強さなどを分析する「グラフ処理」をSpark上で実行するための仕組み．
- Sparkが備えるインメモリー機能を使う．

096 Tezで100倍高速化するSQL
- Hive・・・SQLクエリをMapReduceに変換して実行する仕組み
- 元々は米Facebookが開発した．MapReduceに変換する処理性能は良くなかったが，Tezによって最大100倍高速化する．

097 機械学習をインメモリー処理
- MLib・・・Spark用機械学習ライブラリ
- 機械学習では同じデータを繰り返し読み込んで処理する．
- Sparkが備えるインメモリー処理を使うため，Diskを使うMapReduceよりも高速．

3-4-5-2. 貯めたデータが大化けする SparkやTezで動く周辺機能

098 複雑なデータフロー処理に向く
- Pig・・・独自のクエリーをMapReduceに変換して実行する仕組み．元々米Yahoo!が開発した．
- 複雑なデータフロー処理を効率化できる．
- Tezによって処理速度が大幅に高速化する

099 Sparkを使う分散SQLエンジン
- Spark SQL・・・Spark上に実装した分散SQLエンジン．
- Sparkが備えるインメモリー機能を使用するため動作が高速．
- 既存のHive用SQLクエリーも利用できる

100 ストリーム処理をSpark上で実行
- Spark Streaming・・・ストリーム処理をSpark上で実行する仕組み
- Sparkが備えるインメモリー機能を使用する．
- ストリーム処理とは，新しく発生したデータをディスクに書き込まずにメモリー上で即時に処理すること

4．ビッグデータとは

4－1 ビッグデータと言われる前
4－2 ビッグデータとは
4－3 ビッグデータの量
4－4 医療ビッグデータ
4－5 ビッグデータを支える技術
4－6 ビッグデータ背景にデータ・サイエンティスト

4-1 「Big Data」と言われる前

◆PointカードとPOS端末

◆ソーシャル・メディア・リスニングSocial Media Listening

- 2011年　富山県内では三協立山(株)が既に実施
- Social Media Listeningとは，Facebook，Twitter上で展開される企業や商品に関する生活者の口コミ情報を収集／分析すること
- Facebook以上に情報が入手しやすいTwitterがターゲット
- Twitterの情報はフリーの分析サイトや，「見える化エンジン」を提供しているプラスアルファ・コンサルティング，Facebookも同様のBuzz FinderやTrue Tellerの他，Salesforce.comのRadian6などのテキスト・マイニング分析システムによって，つぶやき情報，アカウント情報，アクセス解析情報などから分析
- 自社のアカウント／ブランディング／キャンペーン／競合分析，関連ワードや発言者分析などが行われ，企業の商品やサービスの戦略に利用
- 企業のFacebook活用事例として，米国ではナイキやコカコーラ，スターバックスが，国内ではSatisfaction Guaranteed，ユニクロ，無印良品，楽天市場などが「ファンページ」を開設し，その情報を分析して，マーケティングを行っている

４－２．ビッグデータとは

◆総務省情報通信白書でのビッグデータの定義

- 鈴木良介著「ビッグデータビジネスの時代」を参照し，ビッグデータとは，「事業に役立つ知見を導出するためのデータ」
- ビッグデータ・ビジネスの定義とは，「ビッグデータを用いて社会・経済の問題解決や，業務の付加価値向上を行う，あるいは支援する事業」

◆「ビッグデータの正体」p.18から

- 「小規模ではなしえないことを，大きな規模で実行し，新たな知の抽出や価値の創出によって，市場，組織，さらには市民と政府の関係などを変えること」．それがビッグデータである．

◆2012年2月発行のThe Economist誌

- 特集“The data deluge「データ大洪水」”が契機
- 「ビッグデータとは，既存の一般的な技術(RDBMSなど)では管理するのが困難な大量のデータ群である」
- ビッグデータの特性は3V(Volume, Velocity, Variety)で示される

4-2-2 Big Dataの定義 (Gartner)

◆Definition with US Wikipedia(日本版はこれを日本語訳)

- Big data usually includes data sets with sizes beyond the ability of commonly used software tools to capture, curate, manage, and process data within a tolerable elapsed time.
- Big data "size" is a constantly moving target, as of 2012 ranging from a few dozen terabytes to many petabytes of data.
- Big data is a set of techniques and technologies that require new forms of integration to uncover large hidden values from large datasets that are diverse, complex, and of a massive scale.
- In a 2001 research report and related lectures, META Group (now Gartner) analyst **Doug Laney** defined data growth challenges and opportunities as being three-dimensional, i.e. increasing **volume** (amount of data), **velocity** (speed of data in and out), and **variety** (range of data types and sources).
- Gartner, and now much of the industry, continue to use this "3Vs" model for describing big data.
- In 2012, Gartner updated its definition as follows: "**Big data is high volume, high velocity, and/or high variety information assets** that require new forms of processing to enable **enhanced decision making, insight discovery and process optimization**.

４－３．ビッグデータの量とは

◆南カリフォルニア大学コミュニケーション学部のマーティン・ヒルバート教授

- 書籍，絵画，メール，写真，音楽，動画(Analog/ Digital)，テレビゲーム，電話通話，カーナビ・システム，放送メディアは視聴率から算出
- 2007年300EB(Exa Bytes，10の18乗)

◆日本IBM

- 2009年の年間，0.8ZB，毎日2.5EBのデータが生成
- 2011年の年間，1.8ZB(Zetta Bytes，10の21乗)
- 2020年の年間，35ZB(予測)

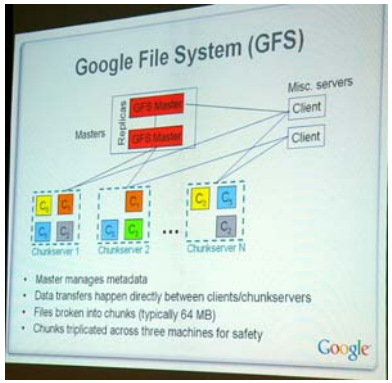
４－４　新たな潮流医療ビッグデータ

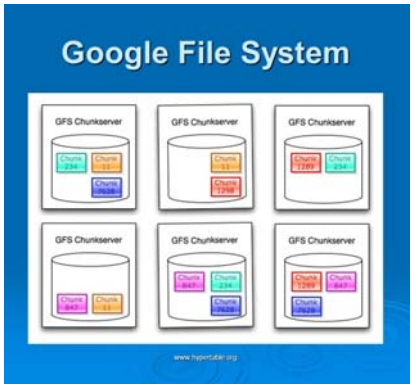
ＮＨＫスペシャル2014. 11. 02, 21:00-21:50

1. 病気を「予知」命を守れ（US Rhode Island州）
   - 新生児集中治療室
   - 感染症を予知
   - オンタリオ工科大学教授のキャサリン・マクグレゴーさん
2. 最先端！ビッグデータ病院（済生会熊本病院）
   - 患者にセンサーを付けて，300項目のデータを収集
   - 早く退院と相関のある3大要素
   - 食事再会の早さ，点滴の期間の短さ，痛みの度合いの少なさ
   - リハビリを早く始め，入院期間を半分に短縮
3. 町ぐるみで「ぜんそく」激減（US Kentucky州）
   - 吸入器を使って，発作の起きた原因を解析，発作の回数が半減
   - 発作のポットスポットを調査，南西の風，
   - 原因を調べるための大気調査を開始

４－５　ビッグデータを支える技術

Google File System（GFS）





GFSの特徴としては同じファイルを異なるマシンに重複して持たせることで，
一部のマシンが故障してもファイルが失われない（全世界に3か所）

4-5-2　ビッグデータを支える技術

（Googleの技術基盤）		（オープン・ソースの技術基盤）	
MapReduce	Big Table	Hadoop MapReduce	HBase
Google File System（GFS）		Hadoop Distributed File System（HDFS）	

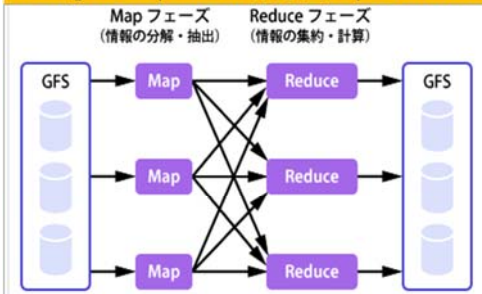
MapReduceは，HPCを多数並べて構成したクラスター・システム上で，大規模なデータ処理を分散実行するためのフレームワーク

Big Tableは，はGoogle File System上に構成されたスケーラブルな分散データベース

GFSは，同じファイルを異なるマシンに重複して持たせることで，一部のマシンが故障しても，他の2つのサーバーから複製をつくる

4-5-3　ビッグデータを支える技術

GoogleのMapReduce概念図http://codezine.jp/article/detail/2448



MapReduceという名前は処理を「Mapフェーズ」と「Reduceフェーズ」という2段階のフェーズに分割することに由来します。Mapフェーズでは大量の情報を分解し、必要な情報を抜き出して出力します。ReduceフェーズではMapフェーズで抽出された情報を集約し、それに対して計算を行い結果を出力します。MapReduceの入出力はGFSからなされることが多い

４－６　ビッグデータ背景に
データ・サイエンティストData Scientist

◆FacebookのData Scientistの記述
- Facebookでは，ビジネス・アナリスト，統計学者，エンジニア，リサーチ・サイエンティストといった従来の肩書きは，私たちのチームにとってまったく魅力的なものではなかった．
- 各役割の作業負荷は多種多様である．ある日の，あるメンバーの行動は，
- 多段階の処理パイプラインをPython（言語）で書き，
- 仮説検定を設計し，
- 統計ソフトウェアRを用いてデータ・サンプルの回帰分析を行い，
- Hadoopで大量のデータを扱う製品やサービスのアルゴリズムを設計して実装し，
- 分析結果を明瞭かつ簡潔な方法で，組織の他のメンバーと話し合う，といった感じだ．
- このように数多くの仕事をこなすのに必要なスキルー式を著すために，私たちは“Data Scientist”という肩書きを作りだした．
- “ビューティフルデータBeautiful Data”，Toby Segaran, Jeff Hammerbacher編，堀内，真鍋，苅谷，小俣，篠崎共訳，オライリー・ジャパン，2011.2.28, ISBN978-4-87311-1489-7, ￥3,400+TAX

4-6-2　Data Scientistに求められるスキル

◆Skill技能
- Computer Science・・・HadoopやMahoutなどの大規模並列処理技術や機械学習，Database, RDBMSとSQL, Python/PHPなどのScript言語，修士号/博士号または同等の職務に4年以上の経験
- 数学，統計，データマイニング・・・統計パッケージSPSS, SASなどの技術の他，OSSのプログラミング言語R
- データの可視化・・・SAS,MATLAB,R, Infographics

◆FacebookのData Scientistに対する求人票の内容
- 職務内容：
- 重要なプロダクトの課題を同定し，対処するために，Product Engneering Teamと密接に連携して職務にあたる．
- データに対して，適切な統計テクニックを適用し，課題解決を図る
- 結論をProduct ManagerとEngineerに伝える
- 新規データの収集と既存のデータソースの改良を推進する
- Productの実験結果を分析・解明する
- 計測・実験方法のBest Practiceを開発し，Product Engneering Teamに伝える
- 資質：コミュニケーション能力，起業家精神，好奇心

５．ビッグデータ「３つの大変化」

５－１　第１の変化「すべてのデータを扱う」
「N=全部」の世界

５－２　第２の変化「精度は重要ではない」
量は質を凌駕する

５－３　第３の変化「因果から相関への世界」
答えが分かれば，理由はいらない

５－１　第１の変化「すべてのデータを扱う」
「N=全部」の世界

◆ビッグデータは人々の意識に3つの大きな変化をもたらした．

■あるテーマに関して，一部のデータや統計的なサンプルで済まさず，すべてのデータを分析できるようになった．

■正確さにこだわり続けるのではなく，現実世界の乱雑なデータに，真っ直ぐ向き合おうとする意欲である．

■つかみ所のない因果関係を追い求めるよりも，相関関係を積極的に受け入れる発送の転換である．

5-1-2 第１の変化「すべてのデータを扱う」
「N=全部」の世界

◆無作為抽出という革命
- 無作為抽出した1,100人の標本があれば97%以上の精度で，母集団の動向を言い当てることができる．
- 400人無作為データでは，95%の確率で，１万人から，10万人，100万人，1,000万人，１億人の意見が分かる．

◆標本作成の失敗例
- 1936，当時存在した有力週刊誌「Reader Digest」が，大統領選を前に有権者200万人を対象に調査を実施，共和党候補の圧勝を予測したが，これが大外れで，Franklin D. Rooseveltが523対8で圧勝した．
- 無作為性が甘かった．同誌は購読者リストと電話帳により調査対象者を選んだのだが，当時としては裕福者で，共和党支持者が多かった．

5-1-3 第１の変化「すべてのデータを扱う」
「N=全部」の世界

◆八百長試合を探せ
- 角界を揺るがす八百長疑惑。日本相撲協会の放駒理事長は2011年2月2日の会見で「過去には一切なかった」と述べたが，
- シカゴ大学のスティーブン・レビット教授等は，１９８９年から２０００年までの十両以上の力士２８１人の取組3万2千回以上を調べた．
- 千秋楽に7勝7敗の力士が8勝6敗の力士と対戦した際の，勝率の「からくり」過去の対戦結果から出した計算では，７勝７敗の力士の勝率は48・7%だが，
- ７勝７敗で迎えた力士の千秋楽での勝率は７９．６%にもなった．
- この確率は，次の場所で両者とも勝ち越し問題が生じない場合，７勝７敗の力士の勝率は４０%にダウン．その次の場所では約５０%と，元の勝率に近づくという．

5-1-3-2 第１の変化「すべてのデータを扱う」
「N=全部」の世界

◆八百長試合を探せ２
- レビット教授と同僚のマーク･ダガン教授は，過去11年分，延べ6万4000番の取組データを基に異常を探し出した．目論見は当たった。確かに八百長試合らしき動きがあったが，誰も注目しないような取組だった．

  「ビッグデータの正体」p.49-51では64,000の取り組みとなっている．
- この奇想天外な研究論文は，学術誌の「American Economic Review」に掲載され，後に「Freakonimics（邦訳『やばい経済学』共著，東洋経済新報社）」として出版され，ベストセラーになっている．

５－２　第２の変化「精度は重要ではない」
量は質を凌駕する

◆「乱暴な方が正確になる」時代

■文法チェッカー（Microsoft）
- 2000年MS Researchのミシェル・バンコとエリック・ブリルがMS Wordの文法チェッカーの改良を模索
- 既存のアルゴリズムで，データ量を増やすことを確かめる．通常は100万語のコーパス（Corpus:実際の文例DB）
- 2人は4つのアルゴリズムを用意し，1000万語，1億語，10億語でトライ．
- 50万語で最低の成績だった単純なアルゴリズムでは，10億語で，文法ミスを見つけ出して修正する正答率が75%から95%以上に跳ね上がった．
- 最高のアルゴリズムでも正答率は86%から94%に改善された．

◆Googleは１兆語で，Google翻訳に挑んでいる．

5-2-2　第２の変化「精度は重要ではない」

量は質を凌駕する

■機械翻訳（IBM）

- 1954年，IBM701で250語の言葉のペアと6つの文法ルールを登録し，ロシア語の60フレーズを英語に，円滑に翻訳した．
- 1990年代後半，IBMの「キャンディード」プロジェクトでは，英語とフランス語で発行されているカナダ議会の議事録から10年分に及ぶ翻訳，およそ300万センテンスを利用して，機械翻訳をおこなった．成果は今ひとつだった．

■Google翻訳

- 2006年，Googleが誇る１兆語Corpusに収録されている英語センテンスは，品質は怪しいが，950億語を達成．
- 翻訳サービスは，精度も高く，最もうまくいっている．
- 2012年半ばには，対象言語が60に拡大，14言語では音声入力でも，円滑な翻訳が可能．

5-2-3　第２の変化「精度は重要ではない」

量は質を凌駕する

■量は質を凌駕する

- ビッグデータの世界に足を踏み入れるためには，「正確=メリット」という考え方を改める必要がある．

◆「ビリオンプライス・プロジェクト」

- 米労働統計局は，消費者物価指数の算出に，全米90都市の小売店や企業を対象に，数百人もの職員が日々，電話，ファクス，直接訪問による聞き取り調査を実施．
- トマトの料金からタクシー料金まで，8万点の価格を，年間2億5千万ドル（250億円）を使って，数週間かけて報告書としてまとめていた．
- MITの経済学アルベルト・カバロ教授とロベルト・リゴボン教授はビッグデータを使って物価調査を実施．
- Web上のデータを自動的に集めるソフトを駆使し，毎日50万点の価格を収集する．
- このビッグデータに，ある分析を加えた結果，2008年9月のリーマンショック後のデフレ兆候を見抜いた．

５－３　第３の変化「因果から相関への世界」

答えが分かれば，理由はいらない

◆書評家を敗北させたアマゾン

- Washington大学大学院で人工知能を研究していたグレッグ・リンデンGreg Linden（24）が，1997年に休学し，オープンから2年のAmazon.comで働くことにした．
- 同社のWeb siteに，当時の競争力の源泉でもあった「アマゾンの声」という書籍紹介コーナーがあった．
- 同社CEOのジェフ・ベゾスがある有望なアイデアの実験に乗り出す．「個々の顧客の購入履歴や好みのデータに基づいて書籍を推薦する仕組み」
- 顧客の膨大なデータ（「最後まで迷ったが，購入に至らなかった書籍」「どれくらいの時間チェックしていたか」「一緒に購入したのはどの書籍か」）を蓄積
- このデータを従来の方法「標本データを分析し，顧客全体の共通項を探る」で加工していた．
- その結果，「前回の購入書と大差ない書籍を延々と紹介し続けた．客にしてみれば，はた迷惑な店員につきまとわれながら買い物をしているようなものだった」（当時の書評委員：ジェームズ・マーカス）

5-3-2　第３の変化「因果から相関への世界」

答えが分かれば，理由はいらない

◆書評家を敗北させたアマゾン－２

- Greg Lindenは，顧客全体の買い物内容から共通項を探る機能は，商品推薦システムに不要だと気付き
- 重要なのは，一見関係なさそうな商品同士の相関関係を見つけることだった．
- Linden等は，「商品間」の強調フィルタリング技術で特許申請し，この手法に切り替えたことが転換点となった
- 相関関係の計算は予め済ませておけるので，お勧め商品は即座に表示でき，汎用性も高く，商品カテゴリーにまたがるお勧めも可能になった．
- 次は提示する内容．専属の書評委員による書評か，それともコンピューターがはじき出した顧客別のお勧めやベストセラー・リストか．
- 書評委員の言葉を信じるか，蓄積されたクリックの“声”を信じるか
- Lindenは，この両者から販売に繋がったケースを比較．差は歴然で，コンピューターのデータから導出したコンテンツが100倍も大きな売り上げを生み出していた

5-3-3　第３の変化「因果から相関への世界」

答えが分かれば，理由はいらない

◆書評家を敗北させたアマゾン－３

- 百田尚樹を読んだ後に，なぜjQueryの本を買いたいとおもったのか，コンピューターは知る由もない．
- それは重要ではなく，ともかく売れたことが事実である
- やがて，人間の手による書評がオンラインで公開されるたびに，書評委員らに正確な売り上げデータが突きつけられた
- ついに書評チームは解散を余儀なくされた．
- Lindenは，「書評チームが負けたことはとても残念だった．しかしデータは嘘をつかない．コストも非常に高かった．」
- 現在，Amazon.comの売上げ全体の１／３は，この「おすすめ」とパーソナル化のシステムから生み出されているという
- Lindenの技術は，Online販売の世界に革命をもたらした

5-3-4　第３の変化「因果から相関への世界」

答えが分かれば，理由はいらない

◆Online DVDレンタルのネットフリックスNetflix, Inc. では，新規受注の3/4が推奨作品である

◆ビッグデータの先駆者–ウォルマート

- ハリケーンの到来が近づくと，懐中電灯と「ポップターツ」の売上げが増加するという事実が判明．そこでハリケーン対策用品コーナーに「ポップターツ」も大量に陳列し，大いに売上げを増大した

◆主役に躍り出た「相関分析」

◆購入品目から女性客の妊娠まで予測

◆各方面に応用される「予測分析」

◆因果関係はそこまで重要なのか

◆オレンジ色のクルマはなぜ欠陥が少ないのか?

◆人間とマンホールの戦い

◆理論は終焉するのか

- ペタバイトのデータがあれば，「相関で十分」と言える．・・・

６．データフィケーションDatafication

「すべてのもの」がデータ化され，ビジネスになる時代

◆航海の姿を変えてしまった男

- 19世紀末の米海軍士官，マシュー・フォンテーン・モーリーMatthew Fontaine Mauryは，33歳のとき，駅馬車の横転事故で足が不自由になり，海軍の事務職に転じ，海図・計器補給廠の責任者に．
- 大西洋航路に関する地図，海図を改良するため，過去の艦長全員が記した航海日誌から，位置・日付ごとの風，波，天候などの情報を抽出・集計した．
- 全体をながめて，一定のパターンを解明し，効果的なルートを探した．そして過去に例のない航海図を完成した．
- Mauryは航海日誌の記入法を標準化し，上陸時の提出を義務化
- 日付，位置，風，海流を記したメモを空き瓶に入れて海に流させるBottle Messageで，情報交換させ，情報を入手する方法を編み出した
- 1855年Mauryは著書“The Physical Geography of the Sea”で，「若き船乗りでも，経験を積むまで手探り状態を続ける必要はない．本書があれば，まるで経験豊富な航海士がすぐそばまで案内してくれるような安心感を覚えることだろう．」と記している

6-2 データフィケーションDatafication２

「すべてのもの」がデータ化され，ビジネスになる時代

◆「座り方」のデータが有望なビジネスに変身

- 産業技術大学院大学の越水重臣准教授は，人間の臀部の形状を科学的に捉える研究に取り組む
- 着座したときの尻の形，姿勢，重量分布を数値化・集計することで，座り方自体が情報になるという
- 自動車のシートに360個の圧力センサーを取り付け，着座時の圧力を256段階で測定し，臀部をデータ化している
- この得られたデータは1人ひとり違うことが分かり，実験では，数人の被験者を98%の精度で識別できた
- この技術を，自動車盗難防止システムの開発に応用し，登録ドライバー以外が運転席に座ると，パスワードが求められ，認証に失敗するとエンジンはかからない．
- この技術の応用は，運転時のドライバーの姿勢も記録されるので，交通事故を防ぐための自動ブレーキかけや，ひき逃げなどの同定，危険防止の警告鳴らしなどに使えるという

6-3 データフィケーションDatafication３

「すべてのもの」がデータ化され，ビジネスになる時代

◆位置もデータに変わる･･･人間の行動を逐一記録するアプリケーション

- GoogleのStreetViewは，街の写真を撮影する際に，近隣から電波が漏れ出ているWiFiルーター情報も収集している
- iPhoneには位置情報とWiFiデータを取得してAppleに送り機能が入っていた（AndroidやMSの携帯向けも同様）
- 米大手運送会社UPSは保有車両にセンサー，無線モジュール，GPSを取り付け･･･システムに知恵や洞察力が生まれる
  - エンジン故障を未然に予測.
  - 配送遅延の有無やドライバーの状況チェック
  - 過去の輸送・配送データから最も効率的な最適ルートの作成で，2011年に，総合情報基盤センター距離4,800万km，ガソリン600万リットル，3万トンのCO2削減に成功
  - 交差点での右左折の少ないルートをアルゴリズムで同定し，安全性や業務効率の向上

6-4 データフィケーションDatafication４

「すべてのもの」がデータ化され，ビジネスになる時代

◆その他のDatafication

- 「Foursqure」というアプリでは，指定された場所を訪れた印として「check-in」ボタンを押すとPointがもらえる
- Foursqure側には客を運んだ謝礼として，各種ポイント・サービスやレストラン案内サービスなど位置情報関連サービスから報酬が支払われる仕組み
- Amazon.comでのショッピング，クリック，カスタマーレビュー
- Googleの様々なサービスでのクリック
- Facebookでの投稿や「いいね」の他，人間関係をグラフ化する「Social Graph」
- TwitterでのtweetやRetweetから「心の動き」をデータ化
- LinkedInでも，･･･Google+でも，Tumblr, Pinterestでも，･･･

７．ビジネス・モデルの大変化（その１）

ーただのデータに新たな価値が宿るー

◆キャプチャCAPTCHA技術を考えたルイス・フォン・アーン

- 1990年代末，メール・アドレス検索ロボット「SPAM Bot」が，オンライン掲示板などに勝手に入り込み，メール・アドレスらしき情報をかき集めていく迷惑行為が蔓延
- 大学を卒業したばかりのフォン・アーンは，「ロボットでなく，人間であることを利用者自身に証明させればいい」ので，「人間には簡単でも，機械には難しい」ことは何かを考えた.
- 入会時やログイン時に，ぐにゃっと曲がって，読みにくい文字列を表示して，読みを入力させるというアイデアを思いつく
- これをCAPTCHAと名付け，1晩で「SPAM　Bot」問題が解決した.
- GoogleはGoogle Books用，書籍のデジタル化を支援する無償のAnti-Bot-Serviceとして，ネットの向こうにいる膨大なユーザーに，10秒間だけ肩代わりさせている.
- このOCRで読み取りにくかった文字列とゆがんだ文字の2つを読み取らせる技術を「reCAPTCHA」という
- その5年後，毎日2億件近く入力されるまでになった.

7-2　ビジネス・モデルの大変化（その２）

ーただのデータに新たな価値が宿るー

◆データ自体が商品に

- Amazon.comは，購入した書籍だけでなく，単に眺めただけのWeb Pageを記録

◆データが持つ「オプション価値」

- 電気自動車EV･･･普及＝移動手段の成否は，バッテリーを支える充電ステーションのようなインフラ整備にかかっている
- 2012年，IBMは，電力・ガス会社のPacific Gas and Electric Company（CA），自動車メーカーの本田技研工業と手を組み，EVが充電するタイミングや場所，電力供給への影響についての疑問点を解消するため，大量の情報を収集する実証実験に乗り出した
- バッテリーの残量，走行位置，時間帯，近隣の充電ステーションの空き状況など，様々な測定値を基に，IBMはかなり手の込んだ予測モデルを開発
- 一連のデータと，供給電力全体の使用量，過去の電力使用パターンを連動させ，あちこちから集まってくるリアルタイム・データと過去のデータを分析し，充電に適したタイミングと場所の割り出しに成功
- 充電ステーション建設に最適なリッチも明らかになっている

7-3 ビジネス・モデルの大変化（その３）

－ただのデータに新たな価値が宿る－

◆再利用はビッグ･ビジネスにつながる

- データ再利用の画期的な例は，検索語句（検索キーワード）
- Web利用状況調査会社のヒットワイズ（Experian Hitwise）では，検索語句のデータ･マイニングから消費者の好みを調べて，顧客に情報提供している
- マーケティング担当者にとっては，春シーズンにピンクが流行るか，黒の人気が復活するかといった傾向が分かる
- Googleでは
- 検索語句分析を評価用に一般公開している
- スペイン第2位の銀行BBVAと組んで事業予測サービスを立ち上げ，主に観光産業向けに検索データに基づくリアルタイムの経済指標を販売している
- イングランド銀行は，不動産関連の検索データを基に，住宅価格の上昇･下降の傾向を判断している

7-4 ビジネス・モデルの大変化（その４）

－ただのデータに新たな価値が宿る－

◆“データ組み替え”という新手法

- データ内に眠っている価値を引き出す方法は様々
- あるデータ集合と別のデータ集合を結合させたり，組み替えたりすることで，初めて新たな価値が生まれることもある
- ジロウという不動産情報Web Site
- 米国の地域別の地図上に不動産の情報や価格を表示
- その地域の最近の不動産取引や不動産の詳細情報など，複数のソースから寄せ集めた膨大なデータを基に，地域の個々の住宅の価値を予測
- ビジュアル表示したことで，データが分かり易い（Mashup効果）
- 将来の使い回しを想定したデータの収集の好例＝Google
- StreetView製作用の自動車は，建物や道路の写真を撮影しながら，GPSデータも取り込み，地図製作用の情報もチェックし，WiFiネットワークの名前まで収集
- 一つのデータ集号が複数の用途に利用できる．上記のGPSデータはGoogle Mapsの改良に利用されている

7-5 ビジネス・モデルの大変化（その５）

－ただのデータに新たな価値が宿る－

◆データの価値の“原価償却”

- NetflixやAmazonといった企業は，商品購入や商品閲覧，レビュー投稿といったデータを新商品推奨にフル活用
- このため，長期的に何回もデータを使い回そうという誘惑
- 大半のデータは，時間の経過に伴って有用性が低下
- 有効なデータと不適切なデータを選別する高度な仕組みを構築
- 過去の購入履歴データに基づいて推奨した本を閲覧･購入した場合，客の「好み」がまだ反映されているとする．
- 古いデータの有効性にScoreを付けて，情報の“減価償却”

◆スペルミスさえも立派なデータになる

- Microsoft･･･MS Wordのスペルチェッカー
- Google･･･Google Search
- Googleでは毎日30億件の検索が行われている
- その中のスペルミスを再利用している
- 巧みなフィードバックの仕組みを用意している＝「ユーザーが本来入力したかった単語は何か」をシステムに教え込む仕組みを用意

7-6 ビジネス・モデルの大変化（その６）

－ただのデータに新たな価値が宿る－

◆ユーザのデータを徹底的に採集するGoogle

- ユーザーが残していった“デジタルの足跡”を「デジタル排出物」という
- ある言葉やその関連語句が検索されたのは何回か
- あるリンクをクリックし，リンク先に満足できずに検索結果のページに戻ってくる頻度はどれくらいか
- 音声認識，迷惑メール･フィルター，翻訳などのサービスに欠かせない
- 最近Firefoxに「履歴消去」ボタンが付いた

◆米大手書店バーンズ＆ノーブルBarnes & Noble

- 独自端末「ヌーク(Nook)」で，ユーザーのフィードバックを収集，分析したところ，長編ノンフィクションは途中で投げ出しやすい．
- この発見を参考に，「ヌークスナップス」という短編作品の書籍シリーズを生み出した

◆オンライン教育プログラム

- 「ユーダシティUdacity」「コーセラCoursera」「エディックスEdics」などのオンライン大学は，Web上での学生の動きを記録し，教育効果の有無をチェック
- クラスの規模が数万人に上り，とてつもない量のデータが集まってくる．講義の一部を再視聴した学生の割合が大きい場合，講義で理解できなかった部分がわかる

7-7 ビジネス・モデルの大変化（その７）
－ただのデータに新たな価値が宿る－

◆政府が保有する大量のデータの行く末
- 2009年のdata.gov開設時は47点
- 2012年7月時点で，172機関から45万点のデータが公開
- FlyOnTime.usでは天候の変化を基に，航空機の発着遅延の見通しが分かる

◆Facebook社の本当の価値はいくらなのか
- 2009～2011年までに「いいね！」や投稿記事，コメントなどで，“カネのなるコンテンツ”を2兆1000億件も集めていた
- 1件につき5セントの価値があった
- １人当たりの価値が約$100とすると，12.8億人では，・・・

◆データは既に金融商品化している
- データを第3者にライセンスし，対価は定額ではなく，抽出された価値に応じて一定割合をもらう
- 書籍や音楽，映画のように，印税と同じ様に支払われる

８．ビジネス・モデルの大変化（その２）
－データを上手に利用する企業－

◆ビッグデータ企業の3タイプ
- データ型・・・データを実際に保有しているか，少なくともデータを入手できる立場にある企業
  - 代表例：Twitter
- スキル型・・・専門的なノウハウを持ち，実際に業務として分析などを手がけるコンサルティング会社やITベンダー，調査会社が中心だが，データを保有せず，画期的な用途を考案する独創性も欠けていることが多い．
  - 代表例：Walmartのデータ分析を行っているTeraData
- アイデア型・・・ビッグデータ思考とも言える．データなし，ノウハウなしで成功している企業の多くが，このタイプ．
  データから新たな価値を引き出すことにかけて，独創的なアイデアを持つ創業者や従業員がいる．
  - 代表例：ジェットパックJetpackという旅行ガイドサイト（Googleが買収）

8-1 ビッグデータのバリュー・チェーン（米国）
－価値連鎖＝価値を高めていく－

◆航空券予約ネットワークを運営するITAソフトウェア
- 航空運賃予測サービスのフェアキャストにデータを提供しているが，自社では分析作業をしていない．

◆Master Card・・・自社で分析
- 同社のカード会員は210カ国15億人
- Master Card Advisersと呼ばれる部門が，650億件の取引データを集めて分析し，ビジネスと消費者のトレンドを予測する
- このトレンド情報を外部に販売する

◆アクセンチュア
- 様々な業界から委託を受けて，最先端の無線センサーでデータを収集し，分析している．
- ミズーリ州セントルイスの市営バスに無線センサーと取り付け，エンジンをモニタリングし，故障発生の予測や最適な定期保守の判断に役立てた．
- この結果車両保有コスト10%を削減，バス1台当たり$1,000の削減
- Washington DCにあるメドスター・ワシントン医療センター

8-1-2 ビッグデータのバリュー・チェーン（米国）
－データを上手に利用する企業－

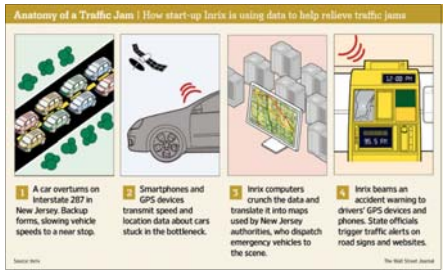
◆Microsoft Research：データ・スペシャリスト企業
- Washington DCにあるメドスター・ワシントン医療センター
- 再入院や感染症を抑えるため，MRに委託して，匿名化した診療記録数年分を分析
- 診療記録には，患者の属性情報，検診結果，診断，治療などが記載されている
- 使用したソフトウェアはMSの「アマルガAmalga」
- 分析の結果，驚くべき相関関係がいくつか見つかった
- 退院後1ヶ月以内に再入院する可能性が高まった条件一覧から，一般に，鬱血性心不全の患者は再入院しやすく，再入院時は治療も難しくなるが，予想外な兆候が見つかった．
- 「憂鬱感」など心痛らしき言葉が含まれていた場合，退院から1か月以内に再入院する確率が著しく高まる．

8-1-3 ビッグデータのバリュー・チェーン（米国）
－データを上手に利用する企業－

◆ビッグデータ思考の企業や個人
- Bradford Crossは2009年8月，友人等と「フライト・キャスター・ドットコムFlight Caster.com」を立ち上げた
- すでに公開されている過去10年の全フライトを気象データと組み合わせ、米国内のフライトの遅延予測情報を提供
- プリズマティックPrismatic

◆交通量分析のインリックスInrix

◆・・・



8-1-4 ビッグデータの欧米企業の活用例
－データを上手に利用する企業－

◆イーベイeBay(インターネット・オークション)毎日50TBのデータが生成
- 桁違いのデータ発生速度・・・全世界で2.7億人の登録会員
- 1日に売買されるMP3 Player3,600台，香水4,800個
- 化粧品は2分ごとに，シャンプー・コンディショナーは秒ごと
- イーベイのデータ分析基盤・・・EDW，Singularity，Hadoop

◆ジンガZynga　ゲーム会社の皮をかぶった分析会社
- Facebookで展開されるSocial Gameで上位7／10を独占
- 3クリック・ルール・・・最初の3クリックで続けるか否かが決まる

◆セントリカCentrica　スマートメーター(通信機能を備えた電力メーター)導入によりエネルギー消費パターンを分析
- 英国における電力・ガス料金の請求の実態・・・年2回
- スマートメーター導入のインパクト：1年で2.3万円の節約

◆カタリナ・マーケティングCatalina Marketing
レジ・クーポンで顧客の購買行動をデザイン
- 全世界5.5万店舗，週間約3.6億人の消費者にリーチ可能
- 全世界1億人分以上の購買履歴を蓄積
- 顧客の購買行動を予測し，割引クーポンで店頭消費を動かす

8-2 ビッグデータの日本国内の活用例
－データを上手に利用する企業－

◆コマツのKOMTRAX（コムトラックス）
- 坂根正弘（当時会長，現相談役）ダントツ経営
- 建設機械の稼働状況を遠隔監視できるシステムである
- KOMTRAXは建設機械にGPSや各種センサー等を付けることによって，建機の現在位置，稼働時間，稼働状況，燃料の残量，消耗品の交換時期などのデータを，衛星通信や携帯電話通信などを使って，最終的にInternetを経由して，日本のコマツのサーバーに送信する．
- 世界各地の販売代理店や顧客はコマツのサーバーにアクセスして，自分の地域のデータや顧客が自身データを確認できるシステムで，1999年から稼働．
- 2012年3月末時点で，全世界70カ国で26万台の建機で稼働中．
- GPSにより，どの地域で機械の稼働時間が増加し，どの地域で減少しているかも把握できるため，需要動向を予測し，在庫や生産量を適切にコントロールすることができる．
- 本当は遠隔ロック機能で・・・

8-2-2 ビッグデータの日本国内の活用例２
－データを上手に利用する企業－

◆リクルート
- Hadoopの徹底活用でデータ分析に対する意識改革に成功
- 「SUUMO」「ゼクシィ」「じゃらん」「ホットペッパー」
- 中古車情報サイト「カーセンサーNet」，割引チケット共同購入サイト「ポンパレ」など，企業と人を結び付ける多彩なサイトを運営．
- 「ホットペッパー」では，1週間分のアクセス・ログを処理するのがやっとで，一部の会員8万人にRecommend Mailを送付していたが，Hadoopで，1年半のログを処理し，20万人にRecommend Mailを送付できるようになった．

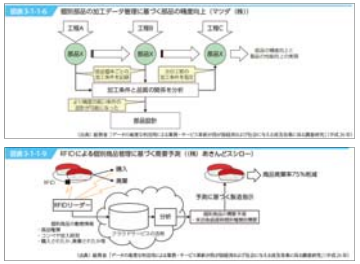
◆GREE・・・2011年第4四半期でDeNAを抜く
- 急成長の原動力となるデータ駆動型アプローチ
- 「1個人のセンスよりも数千万人のデータを信じる」
- GREE AnalyticsというData Mining Toolを独自開発
- ユーザーの登録日，登録経路，利用状況，各イベントの参加率，プレイ率，アイテム別売上げ，ゲーム進捗状況，継続率などのユーザー動向データが，時間単位で把握できる

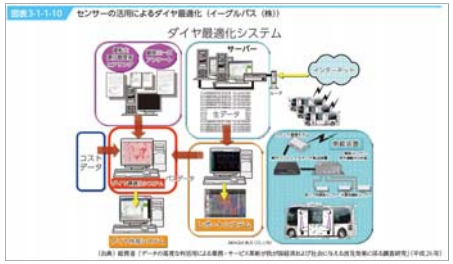
8-2-3 ビッグデータの日本国内の活用例３
－データを上手に利用する企業－

◆マクドナルドのOne to Oneマーケティング
- 携帯電話サイト「トクするケータイサイト」(2003年7月)
- おサイフケータイ対応携帯電話向け「かざすクーポン」(2011年3月)
- 日本マクドナルドは，同社の顧客1人ひとりの購買履歴を詳細に分析し，購買パターンに応じて，1人ひとり内容の異なる割引クーポンをケータイに配信．

8-2-4 ビッグデータの日本国内の活用例４
情報通信白書　H26年版の注目事例から

◆製造業における活用事例・・・マツダ(株)
◆農業における活用事例・・・本川牧場
◆水産業における活用事例・・・(株)グリーン＆ライフイノベーション
◆サービス業における活用事例・・・(株)あきんどスシロー
◆運輸業における活用事例・・・イーグルバス(株)
◆広告業における活用事例・・・(株)マイクロアド





９．ビッグデータのマイナス面

◆Amazon.com・・・ショッピングの好み
◆Google・・・Web Site閲覧の癖
◆Twitter・・・心の動き
◆Facebook・・・心の動き＋交友関係
◆SmartPhone・・・通話相手＋すぐ近くにいる人物
◆街角の監視カメラ・・・移動状況
- プライバシーの保護が困難になる
- プライバシーへの脅威を生み出す
- データ独裁の犠牲者になるリスク

9-2. ビッグデータのマイナス面２

◆プライバシーの麻痺
◆匿名化されたデータでも同定は可能
- AOL事件
- ネットフリックス事件

◆従来のプライバシー保護のために使われてきた3大対策
- 個別の告知と同意
- データ利用拒否を本人が通知できる精度OptOut
- 匿名化

◆捜査のあり方も根底から変わる
- 予防型犯罪捜査
- 映画「Mynority Report」

◆マクナマラの大失敗
◆データの独裁
◆ビッグデータの影

１０．情報洪水時代のルール
ビックデータをコントロールする３要素

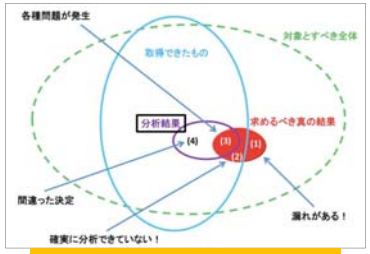
1. 個別同意方式のプライバシー保護から，データ利用者責任制へシフトすること
2. 予測に人間の関与が確実に含まれていること
3. ビッグデータ監査人に相当するアルゴリズミストAlgorithmistの配置すること

10-2　ビッグデータの取り組みで，陥りやすい４つのミス

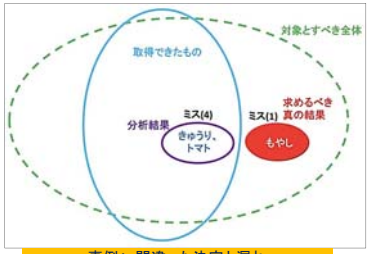
◆IT Leaders http://it.impressbm.co.jp/articles/-/11803　2014年11月4日（火）入江 宏志
- まず認識しておきたいこと＝ビッグデータの収集・分析・活用により“逆転現象”が起こりうる
- 逆転現象とは，これまで不可能だったことが可能になり，後発組が先発の勝ち組に張り合えるようになること
- ビッグデータの目的＝「すべてがビッグデータで予測できる」

◆ビッグデータで扱うデータは統計学とは異なる
- ビッグデータで扱うのは，統計学が扱う「ひな形（全データの代表）」とは異なり，「母数そのもの（全データ）」である
- ビッグデータで陥りやすい以下の4つのミス
  - ミス1：漏れが存在している
  - ミス2：確実に分析できない
  - ミス3：各種問題（局所的やコスト高など）が発生する可能性がある
  - ミス4：間違った決定を下している
- 理解しやすいように3つの事例を紹介（事例A, B, C）
- 事例A：間違った決定と漏れ・・・「感染元は，きゅうり／トマトか，もやしなのか？」
- 事例B：局所的な現象・・・「ビールとおむつは一緒に売れる」
- 事例C：確実に分析できていない・・・「囲碁では人間はコンピューターに負けない」

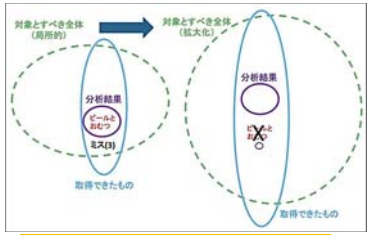
10-2-2 ビッグデータ分析で見られる４つのミス



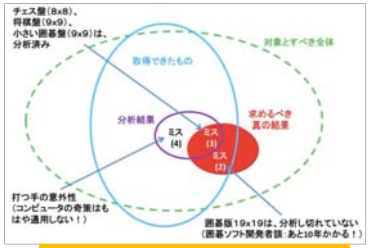
http://it.impressbm.co.jp/articles/-/11803



事例A：間違った決定と漏れ



事例B：局所的な現象

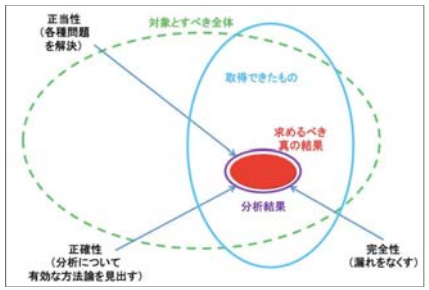


事例C：確実に分析できていない

10-2-3 ビッグデータの世界にGRC（Governance, Risk Management, Compliance）の観点を入れてみよう

100人分のランチボックスを発注した際に遭遇する問題

【GRCの考え方】		【ビッグデータのミス】
信頼性 ＝ 完全性（例：ランチボックスが1人前足りない！）	⇔	ミス1＝漏れ（例：きゅうりでなく、もやし）
× 正確性（例：ランチボックスの中身が足りない！）	⇔	ミス2＝分析が不十分（例：19x19の囲碁盤）
× 正当性（例：一部の口コミでは良い評判だったが、実際に食べると美味しくない！）	⇔	ミス3＝局所的な現象（例：ビールとおむつ）



ビッグデータで陥りやすいミスの（1）漏れがあることは，GRCの世界では「完全性」がないということを表す．
ミス（2）の確実に分析できていないことは「正確性」がないこと．
ミス（3）の各種問題が発生すれば「正当性」がないとなる

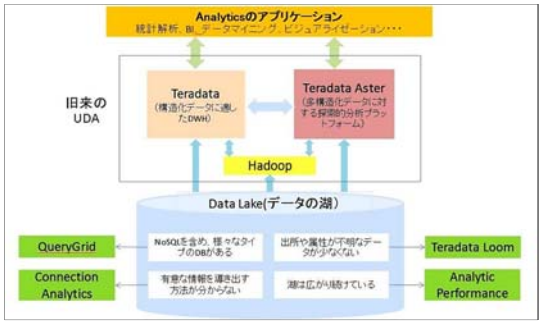
10-3 【PARTNERS 2014報告】「Data Lakeに対応せよ」米テラデータの新製品と新技術の面白味

◆ビッグデータからData Lake（データの湖）へ，そしてAnalytics3.0へ

◆「あなたが所属する企業あるいは組織は，データの価値を十分に引き出し，享受しているか？」
- 「ビッグデータと言われて久しいが，米国企業は本当にそれを分析・活用しているのか」
- 「だとすれば，どんな分析環境を有しているのか」
- 「構造化データと非構造データを統合したこれからのデータ分析（Analytics）環境はどんな姿か」

◆「構造化データ＋ビッグデータ」から「Data lake」へ
- 米テラデータが提唱するUDA（Unified Data Architecture統合データ設計術）はArchitectureとあるが，必ずしも特定の構造や基本設計のことではない
- 「明確な構造を持つリレーショナル・データだけではなく，Webログやマシン生成のデータ，画像データといった非構造データ（「多構造（Multi-Structure）データ」ともいう）も，同じように分析可能にする仕組み，あるいはそれを可能にすること」

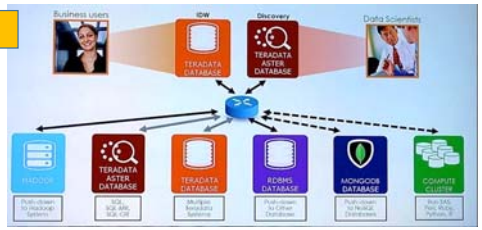
10-3-2 UDA（Unified Data Architecture）全体像



Data Lakeは「いわゆるData Martをペットボトルとすれば，Data Lakeは水源である．
様々な川から流れ込む水が溜まる湖のようなものだ」と説明される

必要なデータを抽出するQueryGrid

ビッグデータと構造化データを分けることなく分析し，顧客対応やマーケティングなど日常の業務プロセスに活かしたり，製品やサービスの開発，経営上の意思決定をサポートしたりすることを，「Analytics 3.0」という



１１．ビッグデータが変える未来

１１－１ ビッグデータが変える医療
- 少子高齢化社会で医療コストの負担を軽減するための「予防医療の推進」・・・電子カルテの標準化，徹底したIT化
- 感染症の予測，伝染病からの被害を最小限に
- DNAの解析から衛生管理・・・不老長寿へ

１１－２ ビッグデータが変える交通インフラ
- 米国自動車保険業界が利用するテレマティクス（遠隔で走行位置や速度などのデータを収集するシステム）を利用し，走行データを分析して，運転状況を保険料に反映
- トヨタの新テレマティクス・サービス「T-Connect」・・・渋滞予測
- ホンダのカーナビ・システム「インターナビ」・・・セーフティマップ
- 航空機遅延情報を提供するFlight Caster.com
- 交通量分析して情報を提供するInrix
- Googleが推進する自動運転システム
- 物流業界での効率的輸送システムでコスト削減
- 渋滞情報，危険回避情報の提供で円滑なトラフィックの確保

11-2 ビッグデータが変える未来２

１１－３ ビッグデータがブラック企業・行政を駆逐
- IT，外食，コンビニ，運輸，スーパー，・・・ブラック企業？
- サービス業でも技術を使ってマージンを生み出し，効率よく仕事ができれば，労働環境と給料アップを両立可能
- 「ブラック化」の危機にある行政，公共部門でビッグデータを活用
- EUの公共部門でビッグデータ技術を使った場合の効果の資産では，15～20%のコスト削減，金額1500億～3000億ユーロ（21.7兆から43兆円）の価値を生み出す
  - 情報アクセスの向上
  - 「見える化」の推進
  - セグメンテーションとパーソナライゼーション
  - 意思決定支援
  - 民間へのデータ提供による新ビジネス創出
- ドイツ連邦労働局の成功事例
  - 失業保険の誤支給，失業対策・雇用促進，他・・・

11-3 ビッグデータが変える未来３

１１－４ ビッグデータが変える「データ都市戦略」

- 不正改造住宅を探せ・・・予測システムで火災のリスクも
- NY市のマイク･フラワーズ率いる分析チーム

１１－５ ビッグデータが変えるエネルギー

- Smart Meterの導入で光熱費の30%のコストを下げる

１１－６ ビッグデータが教育を変える

- TabletとeBook，e-Learningの導入で学生の訪問履歴を収集
- 電子教科書･教材，Virtual Prof.が大学を変える

１１－７ ビッグデータ社会の新しい専門家

- データ･アグリゲーターdata aggregator
- ニーズ高まるデータ･サイエンティスト
- ビッグデータを調査･分析し，公正に評価する「アリゴリズミスト」

◆・・・あらゆる産業が，変わる？・・・

参考図書，雑誌，新聞・・・

◆格差広げるビッグデータ100，日経コンピューター，
2014.07.24号，28-53，日経BP社

◆ビッグデータは人工知能に任せた！，日経コンピューター，
2014.10.02号，22-39，日経BP社

◆IBM Pro Vision，Winter 2012 No.72，
ビッグデータ活用時代 The Era of Big Data Utilization

◆ビッグデータ･ビジネス，鈴木良介著，日経文庫，
2012.10.15，ISBN978-4-532-11268-4，￥860+TAX

◆ビッグデータの正体－情報の産業革命が世界のすべてを変える－，
ビクター･マイヤー=ショーンベルガー，ケネス･クキエ著，斉藤栄一郎訳，
講談社，2013.05.20，ISNB978-4-06-218061-0，¥1,800+TAX

◆ビッグデータの衝撃－巨大なデータが戦略を決める－，城田真琴，東洋経済，2012.07.12，ISBN978-4-492-58096-7，¥1,800+TAX

参考図書，雑誌，新聞２・・・

◆ビッグデータ早わかりA Quick Illustrated Guide to Big Data，
大河原克行著，中経出版，
2013.01.29，ISBN978-4-8061-4620-9，￥1,500+TAX

◆ビッグデータの覇者たち，海部美知著，講談社現代新書，
2013.12.03，ISBN978-4-06-288203-3,¥760+TAX

◆進撃のビッグデータ，牧野武文著，マイナビ新書，
2014.06.30，ISBN978-4-8399-4961-7，￥850+TAX

◆O2O，ビッグデータでお客を呼び込め！ ネットとリアル店舗連携の最前線，松浦由美子著，平凡社新書，
2014.01.15，ISBN978-4-582-85709-2，￥760+TAX

◆統計学が最強の学問である，西内啓著，ダイヤモンド社，
2013.01.24，ISBN978-4-478-02221-4，￥1,600+TAX

◆世界で最も強力な9つのアルゴリズム，ジョン･マコーミック著，長尾高弘訳，日経BP社，2012.07.23，ISBN978-4-8222-8493-0，
￥2,000+TAX

◆日本経済新聞，日経産業新聞

ご清聴ありがとうございました

Thank you for your attention!

最近のビッグデータ活用事情

Recent Big Data Utilization Circumstances

２０１４．１１．１５（SAT）

高井 正三（Shoso Takai）

富山大学総合情報基盤センター

Information Technology Center，University of Toyama