

ビッグデータの理想と現実

情報通信分野の人材教育に必要なビッグデータの
利活用力の重要性を考える



国立情報学研究所

佐藤一郎

E-mail: ichiro@nii.ac.jp

自己紹介: 佐藤一郎

- 国立情報学研究所・アーキテクチャ科学研究系・教授
 - 国立大学法人総合研究大学院大学・複合科学研究科・情報学専攻・教授
 - 政府IT総合戦略本部「パーソナルデータに関する検討会」委員
& 同検討会技術検討WG主査、総務省行政機関等が保有するパーソナルデータに関する研究会委員他、OECD Research Ethics委員
 - 専門: OSやミドルウェアなどのシステムソフトウェア
- 実証実験
(そごう横浜店, NHK)



国立科学博物館(上野)

ニュースゼロ
(日テレ)



NHK週間ニュース深読み やじうまテレビ
(テレ朝)





講演概要

- ビッグデータとは
 - ビッグデータの利活用：理想と現実
 - ビッグデータと人材育成
 - ビッグデータとパーソナルデータ
-
- 家電、自動車等のあらゆるモノがインターネットを通じて接続される IOT(Internet of Things)の時代を迎え、巷にあふれる膨大で多様なビッグデータをいかに扱い、ビジネスや社会課題の解決に役立てるかの重要性がこれまで以上に増している。このような情報が情報を生む時代の情報通信系分野の人材教育とビッグデータの利活用力の重要性を考える。

▶ ビッグデータとは

- 教科書的な説明では

大量・多様 データ処理

大容量かつ多様な
データを収集・処理

MapReduce/Hadoop
p
NoSQL
Key-Value-Store
非定型データ処理

高度な データ解析

高度な解析手法によ
り、データから特徴
やパターンを抽出

データマイニング
機械学習

高速データ 処理・解析

実世界の様々なデー
タを既知の特徴やパ
ターンと照合

Complex Event Processing
オンメモリデータ処理

ビッグ データ

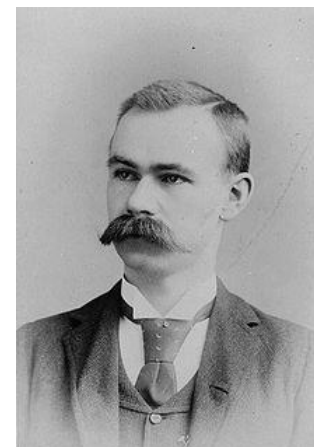
なお、ビッグデータは非定型データを扱うという定義もありますが、
分析するには定型化が必要、手間がかかるのは非定型から定型化する前処理

▶ 世界最古のビッグデータ事例

- おそらく世界最初のビッグデータ事例は、アメリカの国勢調査
 - 1880年国勢調査では集計に7年
 - 1890年国勢調査は移民増により、集計に13年かかると予想
- 備考: 国勢調査は10年おき(憲法の規定)



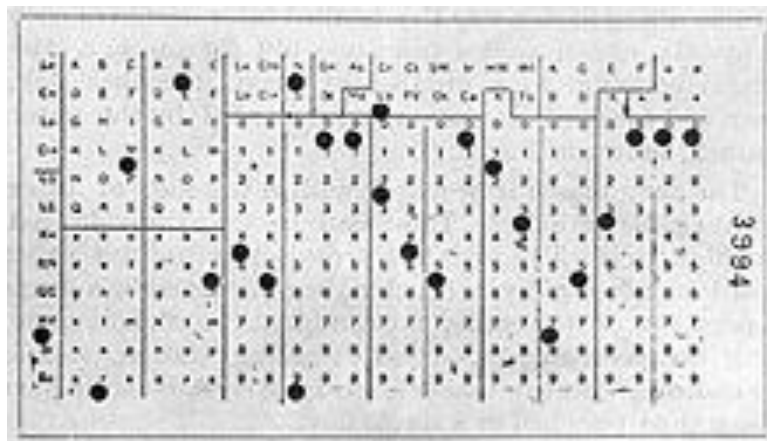
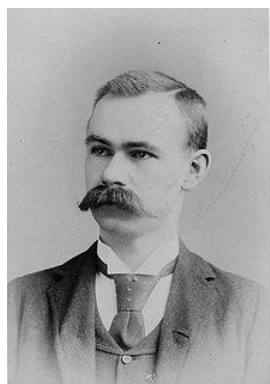
米国政府は集計高速化技術を公募



Herman Hollerith

▶ 古くて新しいビッグデータ

- Herman Hollerithがパンチカードマシン(Tabulating Machine)を発明(1890年)
- 米国政府は同マシンを採用して、国勢調査を1年間で集計



- ビッグデータは決して新しい問題ではない
 - その時点の技術で手に負えないデータがビッグデータ

そしてコンピュータへ

- Hollerithが設立した会社 Tabulating Machine Company はその後、IBMの母体となる(1924)



Hollerith card puncher (1906)



IBM Type 285 tabulators (1936)



IBM 704 (1952)
The first IBM's mainframe



with solid state computing

IBMの製品リスト

List of IBM products - Wikipedia, the free encyclopedia

W http://en.wikipedia.org/wiki/List_of_IBM_products#Unit RSS Google

iCloud News Google Web日記 Java 1.5 Java 6.0 天気・東京 Wikipedia

Tabulators, Accounting machines

- [Hollerith Census Tabulator](#) — 1890^{[8][14][15]}
- [Hollerith Integrating Tabulator](#) — 1896^[15]
- [Hollerith Automatic Feed Tabulator](#) — 1900^[15]
- [Hollerith Type I Tabulator](#) (IBM 090) — 1906^[15]
- [Hollerith Type III Tabulator](#) (IBM 091) — 1921^[15]
- [Hollerith Type 3-S Tabulator](#) — 192x^[15]
- [Hollerith Type IV Tabulator](#) (IBM 301) — 1928^{[9][15]}
- [Columbia Difference Tabulator](#) — 1931^[15]
- [IBM 092](#) — 5-counter Accounting Machine^[7]
- [IBM 093](#) — 5-counter Accounting Machines^[7]
- [IBM 094](#) — Non-print Automatic Checking Machine^[7]
- [IBM 211](#) — Accounting Machine^[7]
- [IBM 212](#) — Accounting Machine^[7]
- [IBM 285](#) — Numerical Accounting Machine; 1927^{[8][2]}
- [IBM 298](#) — Numerical Accounting Machine^[7]
- [IBM 301](#) — see [Hollerith Type IV](#) above
- [IBM 401](#) — Tabulator; 1933^[15]
- [IBM 402](#) — Alphabetic Accounting Machine 1948^{[2][10]}
- [IBM 402](#) — Computing Accounting Machine (with solid state computing device)^{[16][15]}
- [IBM 403](#) — Alphabetic Accounting Machine, 1948^{[2][10]} (MLP - multiple line printing)^[6]
- [IBM 403](#) — Computing Accounting Machine (with solid state computing device)^[16]
- [IBM 404](#) — Accounting Machine
- [IBM 405](#) — Accounting Machine; 1934^{[8][17][15]}
- [IBM 407](#) — Alphabetic Accounting Machine; 1949^{[8][2][10][15]}
- [IBM 407](#) — Computing Accounting Machine (with solid state computing device)^[16]
- [IBM 408](#) — Alphabetic Accounting Machine, 1957^{[2][10]}
- [IBM 409](#) — Accounting Machine; 1959^{[2][18]}
- [IBM 412](#) — Accounting Machine
- [IBM 416](#) — Numerical Accounting Machine^[19] (version of 405)^[6]
- [IBM 417](#) — Numerical Accounting Machine^[7]
- [IBM 418](#) — Numerical Accounting Machine^[6]
- [IBM 419](#) — Numerical Accounting Machine^{[2][10]} (version of 402)^[6]
- [IBM 420](#) — Alphabetical Accounting Machine^[7]

▶ ビッグデータとコンピュータ

コンピュータがビッグデータを可能にしたではなく

ビッグデータがコンピュータを生み出した

■ 結局

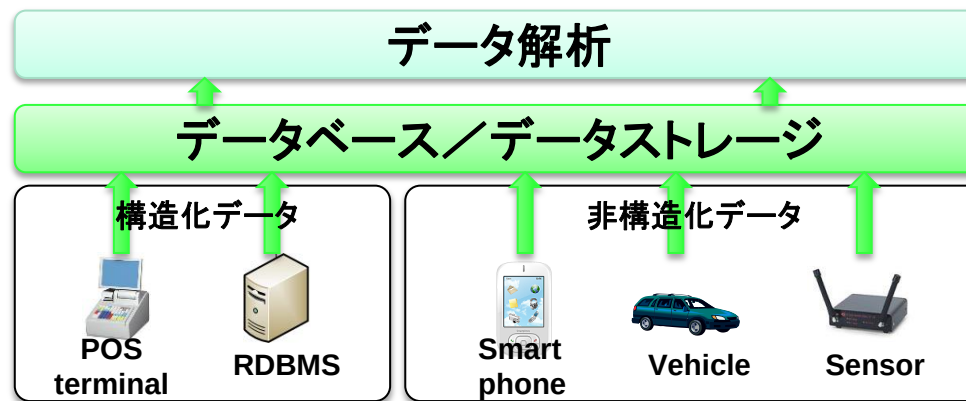
- 現行システムで処理しきれないデータ量・種類を扱う場合のみ、ビッグデータを検討すべき
- データ量は少なくてすむのならば、少ない方がいい
- これまでもビッグデータはあった

むしろいまビッグデータが注目される理由が重要

なぜいまビッグデータが話題になっているかが重要

■ 情報の多様化

- カメラを含むセンサーが増えるとともに、その測定データが増加

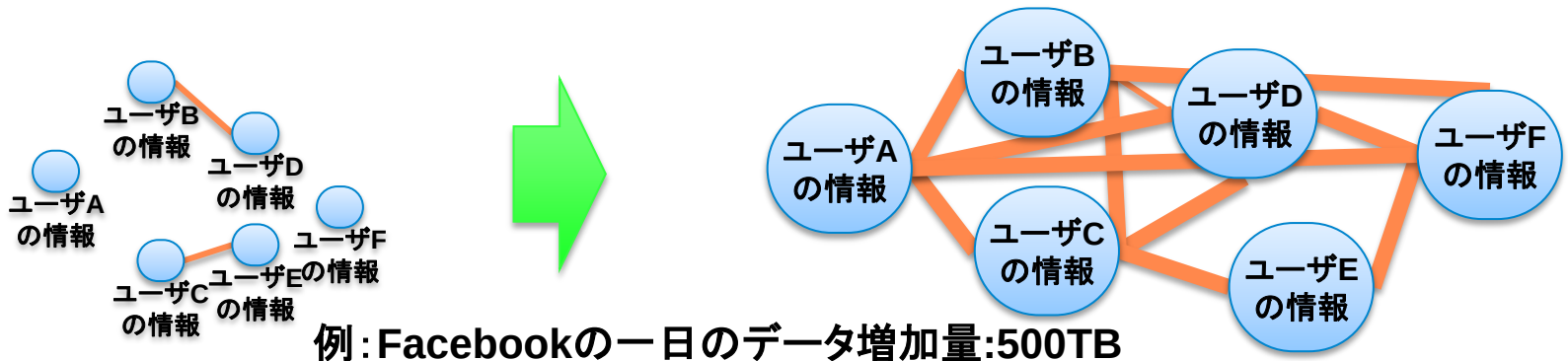


Ciscoの予測では、2015年、スマートフォンが生成するデータ量は1か月で1800ペタバイト(2011年の5倍)

なぜいまビッグデータが話題になっているかが重要

■ 情報が情報を生む

- SNSでは個人が入力したデータ量よりも、プロファイリング量が多い



■ サンプリングで分析できない対象が増えた

- 画一化から多様化へ、ロングテール化した対象は、個々の事例を追う必要がある

▶ マーケティングの変化

- ビジネス環境の変化
 - 消費者の影響力増大、グローバル化、ネット化
 - 消費者は、企業よりも他の消費者を信頼
- マス市場を前提としたマーケティングは通用しない
 - 個々の消費者行動の分析が必須

	これまでのマーケティング	いまのマーケティング
製品管理	4P (製品、価格、流通、プロモーション)	協創
顧客管理	STP(セグメンテーション、ターゲティング、ポジショニング)	コミュニティ化
ブランド管理	ブランド構築	キャラクターの構築

一部のネット通販サイトはマウスの動き(≒視線の動き)やリンクを開く時間も収集

なぜビッグデータが話題になるのか

■ マーケティングの変化

- ネット情報、SNS、消費者の影響力増大
- 消費者は、企業のマス広告よりも他の消費者を信頼

個々の消費者
行動の分析が
必須



■ ディメンションデータからファクトデータへ

- 従来: コンビニ1店舗のアイテム数: 3,000個 → 3000個の日販数データ
- 現在: 各コンビニ店の一日来店数: 1000人で、平均購入数: 3個

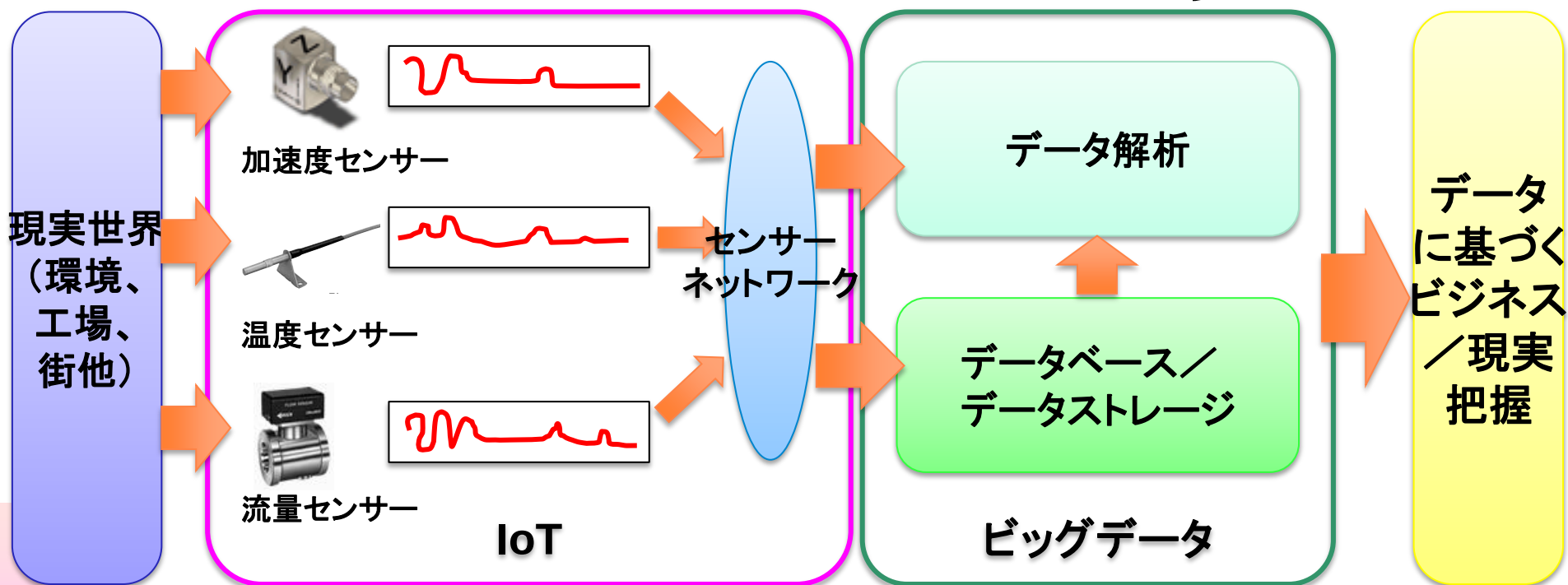
1000 × (アイテム1 / 1 + アイテム2 / 2 + アイテム3 / 3)

IoTとビッグデータ

- IoTとビッグデータは方向性は同じ

- データを集めるところに着目すればIoT
- データを分析するところに着目すればビッグデータ

両者は両輪



必要な人材: 現実世界の知見+IT

ビッグデータの利活用：理想と現実



国立情報学研究所

佐藤一郎

E-mail: ichiro@nii.ac.jp

理想と現実

- 大量データを保有しているのは少数
 - 非常に成功しているWebサービス事業者ぐらい
 - メディアに登場する事例はスモールデータ
- データは目的に応じて収集される(そして捨てられる)
 - 当初目的外利用は、データ精度やエラー率などで問題がおきやすい
 - 例: 旅客機のジェットエンジンに搭載された各種センサーや制御システムは1時間に10TBのデータを生成
 - 異常検出目的であり、正常値(大部分)は不要
- 予測通りの結果しか出ない
 - 思いもしなかったデータ特性はまずない(仮にあっても、本来、気付くべき特性を見過ごしているだけのケースは多い)
 - 研究的に興味深い特性と役に立つ特性は違う

▶ 売上拡大よりも損失縮小

■ 売上拡大手法として

- 他のユーザ行動から、商品を推奨
 - Amazonなどの推薦機能
- ユーザ行動を先回りして商品を提示

データ分析結果が興味深くても
収益拡大につながるとは限らない

■ 損失縮小手法として

- 不正利用監視
 - クレジットカードユーザの行動パターンを抽出して、不正を発見
- 医療データから患者の状態、病気の前兆を発見

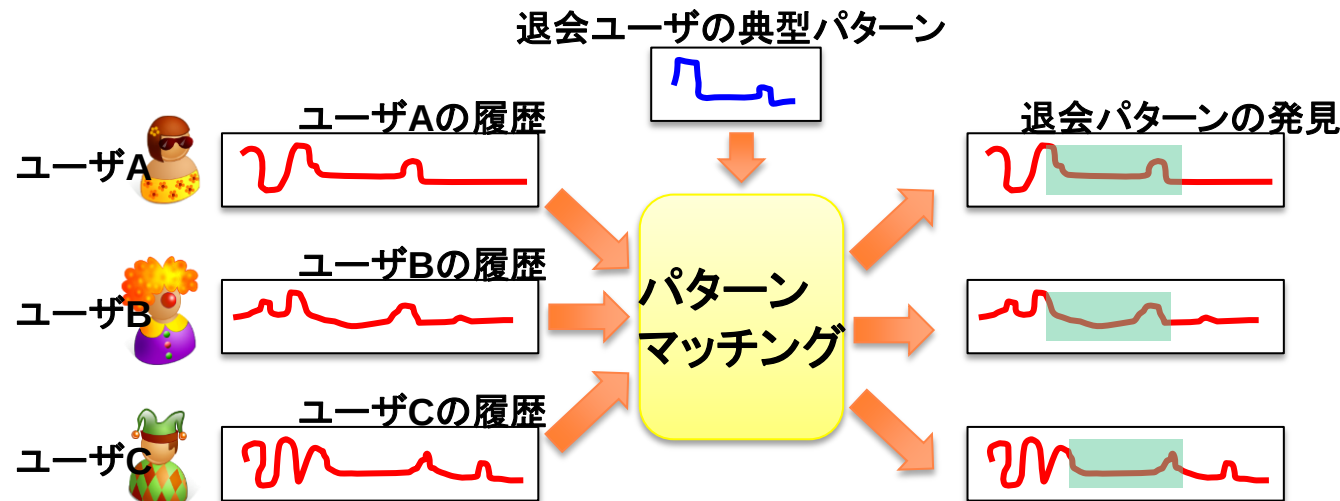
■ 短期的には損失縮小の方が確実&効果的

- 儲けにつながるデータ特性は未知、損につながるデータ特性は既知

応用事例：ネットゲームのユーザサポート

ビッグデータの主要応用先は売上拡大よりも損出削減

- 退会しそうなユーザを発見
 - 退会ユーザには事前に典型的な行動パターンをとる
 - 例：アクセスが減る、他のユーザとの通信が減る
 - 退会しそうなユーザに特典付与、新規ゲームを提案



- 仮にビッグデータを教育に役立てるとしたら、教育効果の高い学習者への支援よりも、低い学習者への支援では？

遠隔監視

づまりや部品の磨耗、
トナー交換、用紙切れなど

- 事例：某事務機器メーカー
 - コピー／プリンターの利用状況をモニタリング
 - 顧客からのトラブル対応依頼電話を受ける前に状況を把握
→ 事前に電話または営業員派遣

傾向：

- ・ ビッグデータを実際に利用している企業は、目的名称や処理内容名称で呼び、**ビッグデータとは呼ばない**
- ・ 「ウチはビッグデータをやっている」という企業の話は信じな...

(顧客A)

機器
(顧客B)

利用状況

営業所

遠隔監視センター

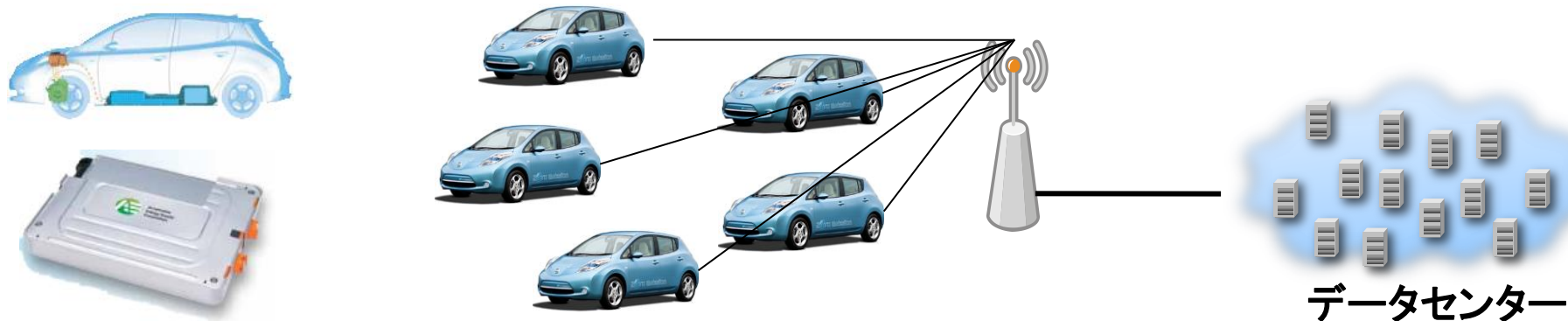
事前配送指示

倉庫

Ichiro Satoh

▶ POSからPOU (Point Of Use)へ

- ビッグデータの情報処理能力があればPOUが可能に
 - 例: EVから定期的にバッテリー状況や位置などをデータセンターに通信
 - 例: 1分おきに送信、充電ステーションの情報を受信(日産Leaf)



- 車体よりもLiイオンバッテリーの方が長持ち
 - 廃車後もバッテリーだけリユース、またはバッテリーだけリース
 - バッテリーのリユースに備えたライフサイクルマネジメント
- ICT人材育成のデータ活用では、育成前後のデータだけでなく、育成過程のデータの利用も可能になる

ビッグデータと人材育成



国立情報学研究所

佐藤一郎

E-mail: ichiro@nii.ac.jp

▶ データサイエンティスト

いまはデータベースや分散システムの技術者の需要が高まっているが、
Hal Varian, chief economist at Google said in *The New York Times*

“I keep saying that the sexy job in the next 10 years will be statisticians”

- 膨大なデータから、調べたい特性に有益なデータを見つけ、その特性とデータにあった解析を方法を選べる人材が必要
 - 統計学や自然科学の実験系の経験・知識のある人材など

ということになっていますが

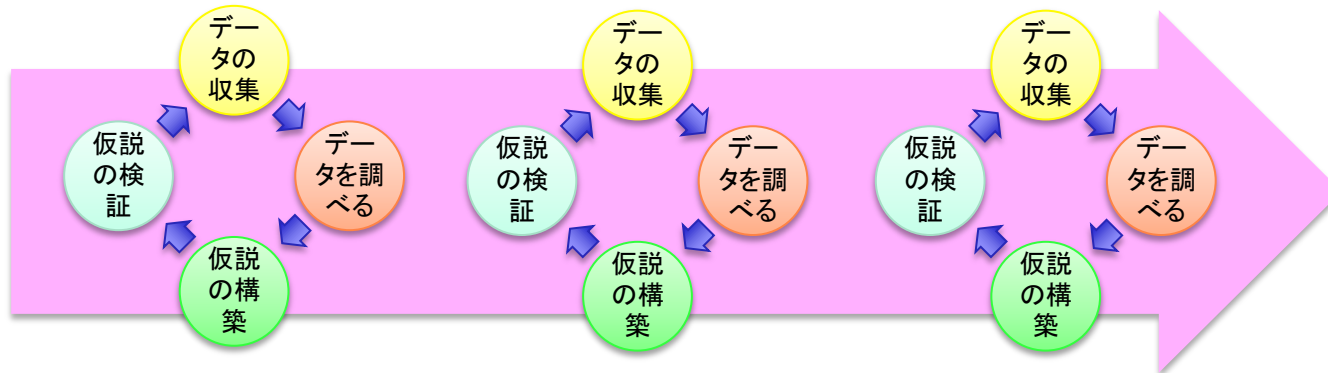
▶ データサイエンティストより現場

- 高度な統計手法は有効とは限らない
 - ビッグデータではデータの品質が悪い(基本的な統計手法で十分)
- 仮説次第
 - 現場の状況を知らなければ仮説が立てられない
- 興味深いデータ特性は現場はうすうす気づいていることが多い
 - 現場の気づきをデータ分析に活かす仕組みが必須
- ビッグデータによる分析を活かすのは現場
 - 現場が分析による知見を活かせるか(現場裁量があるか?)

ビッグデータは仮説検証の繰り返し

■ 仮説検証

- 関心事によって分析手法は違う
- 何らかの特性を予測して、その特性があることのデータから調べる
- 分析してみないと仮説が正しいかはわからない
- 実ビジネスを含めて仮説を検証できる組織(失敗を許容)



100個の仮説を立てて、1個当たればいい世界

■ ビッグデータ処理システムへの要求

- 仮説検証のサイクルを早く回すこと
 - データ分析処理が早いシステム
 - データの可視化(仮説構築や仮説検証を支援)

大量データ向け問い合わせ／ データ分析技術

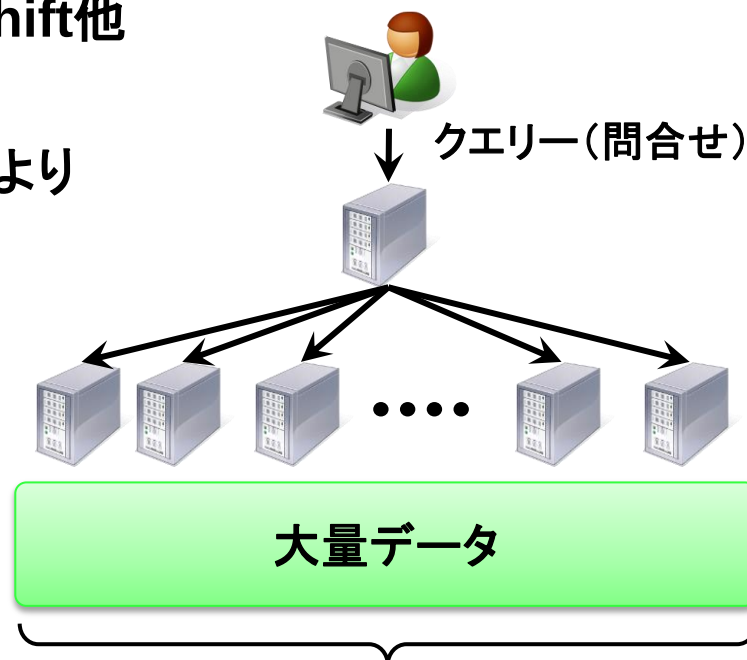
- 詳細分析をする長時間バッチ処理より、仮説検証を早く回す方向へ

- 対話型データ分析向け技術・ソフトウェア

- 例: Google BigQuery(Dremel)、Cloudera Impala、
Apache Drill、Amazon Redshift他

- 共通する方法

- **カラム指向データベース技術**により
大量データ読み込みを高速化
- **クエリーの並列実行**
- **データ分割 (Data Sharding)**
- **SQL記述 (RDBMSと同様の
クエリー(問い合わせ)言語
をサポート)**



数十~数百台のサーバで問合せを分散処理

既存BIやデータ分析との違い

- ビッグデータでは、データを選ぶ、組合せが重要

- コース料理(BI: Business Intelligence)からビュッフェ形式へ



コース形式
(既存データ分析):
与えられた少量の
料理(データ)を
最大限に楽しむ



ビュッフェ形式
(ビッグデータ):
多様な料理(デー
タ)
から選ぶ(摘み食
い)

- 分析精度

- 分析対象のデータが増えることで、分析精度が向上
 - 正確さが低いデータでも大量に集まれば価値につながる(こともある)

データ分析の前処理で大幅にデータ量が減ることが多い

▶ データの母集団と分析精度

■ ユーザ行動解析には、相応データ量が必要

■ 例：イオン

■ のWAON、セブン&アイのNanacoカード

- 発行手数料300円では採算に疑問(さらにポイント付与)
 - データを集めるのはコストがかかる
- 発行側のメリット
 - 顧客囲い込み
 - 顧客行動の把握(最新購買日、購買頻度、購買金額)
- おそらく300万枚程度を発行しないと解析は難しい



■ NanacoはWAONよりユーザ行動分析精度が高い？

- Nanacoは発行時に氏名・年齢・住所などを登録(WAONは無記名)
 - Nanacoはコンビニ向け → 顧客層が広い → 分析が難しい
 - 正確な登録情報は少ないといわれる

▶ 何を分析するのか

- データ分析以前に、区別する対象を決める必要がある
 - 区別しなくても対象までも区別 → データ量は爆発
 - 区別すべき対象を区別できない → 必要なデータ分析はできない
- コンピュータは現実世界をそのまま認識できない
 - 現実世界の対象にID(付番)で区別
 - データ分析ではID設計が肝
 - 例: JANコードは通常品と増量品を区別できるとは限らない

宣伝:

佐藤一郎著「IDの秘密」(丸善出版、2012年)
おそらく唯一のID付けに関する書籍
(業務執筆なので印税はいりません)



Ichiro Satoh

▶ データの種類はほどほどに

- ビッグデータは単一データより、相違なデータの組合せから知見を得る
- しかし、データの種類はほどほどに
 - データ種類が増える(次元が増える)と分析コストは指数的に増えていく
 - 結局、処理が終わらない
 - 次元が多すぎて何を調べればいいのかもわからない

(人工知能の)フレーム問題

- 逆に見えなくてもいいことまでみえることも

ヒッグデータとパーソナルデータ



国立情報学研究所

佐藤一郎

E-mail: ichiro@nii.ac.jp

▶ パーソナルデータ

- ビッグデータとIoTの進化は、プライバシー問題を引き起こしかねない
 - **Volunteered data** (自発的生成データ)
 - 個人が生成し、明示的に共有されるデータ
 - **Observed data** (観測データ)
 - 個人の過去の行動に基づくデータ
 - **Inferred data** (推定データ)
 - 自発的生成データ及び観測データから、推定・プロファイリングされたデータ
- プライバシー保護は
 - 法制度による方法(例: 個人情報保護法)
 - 技術による方法(例: 匿名化、アクセス制御)

} ← IoT

} ← ビッグデータ

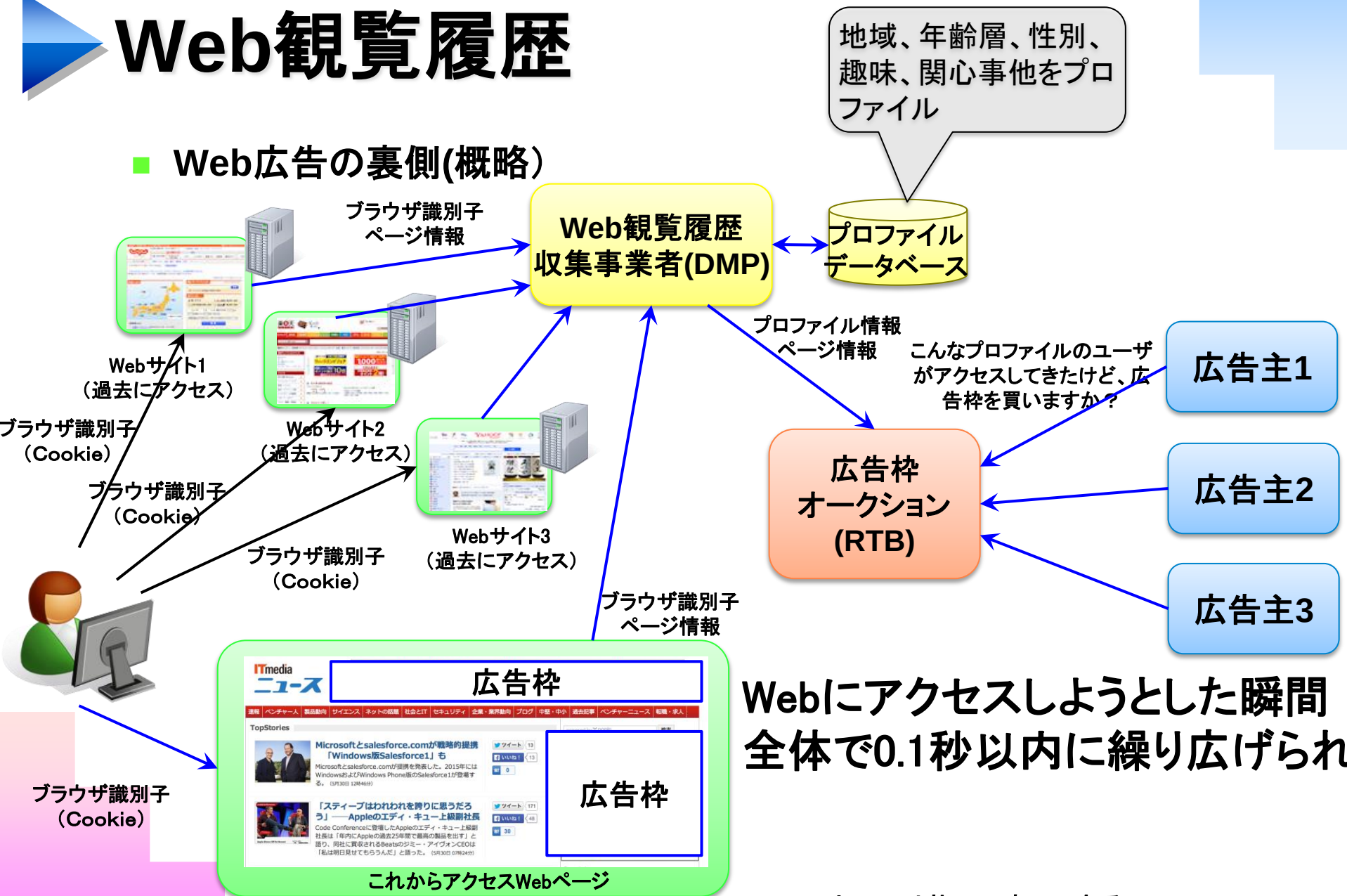
▶ ネットサービスビジネスモデルの変化

- 広告枠を販売（ユーザは顧客ではない）
 - 広告を表示することで、広告主から広告料を稼ぐ
 - 例：既存の多くの無料ネットサービス（Googleを含む）
- サービスを販売（ユーザが顧客）
 - 所定期間・回数・取得情報に応じてユーザから利用料をもらって稼ぐ
 - 例：ネットゲーム、新聞オンライン版、LINE（スタンプ）
- ユーザ情報を販売（ユーザ（の情報）が商品）
 - サービスを通じてユーザに関する情報を収集し、その情報を売って稼ぐ
 - 例：Twitter、無料ヘルスケアサービス

スマートフォンやIoTの時代では、ネット広告効果は下がる

Web観覧履歴

■ Web広告の裏側(概略)



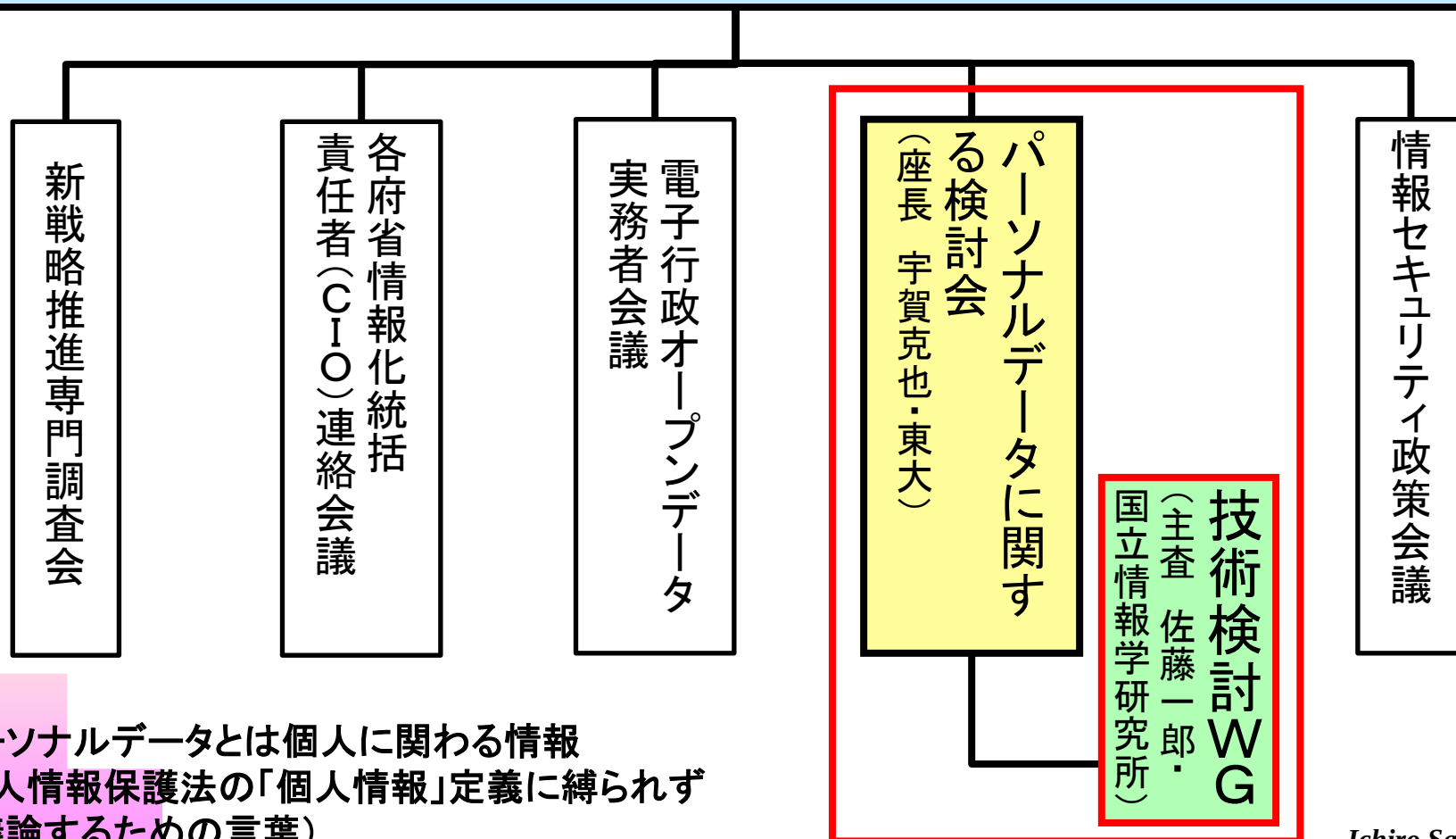
Webにアクセスしようとした瞬間
全体で0.1秒以内に繰り広げられ

▶ ビッグデータとパーソナルデータ

- パーソナルデータに関わる法律は**個人情報保護法**
 - 本来は制定時の付帯決議により2008年に見直しのはずが先延ばし
 - 今週、改正案が閣議決定、通常国会に提出へ
- 法改正の特徴：
 - ビッグデータを前提に、パーソナルデータの利活用と保護
 - 主務大臣制から、第三者機関の設立へ
 - 個人情報個人本人に同意を取らずに第三者提供する枠組みの導入
 - 小規模事業者(5千件)の適用除外は撤廃

▶ パーソナルデータに関する検討会の体制

高度情報通信ネットワーク社会推進戦略本部
(IT総合戦略本部) (本部長: 内閣総理大臣)



パーソナルデータとは個人に関わる情報
(個人情報保護法の「個人情報」定義に縛られず
に議論するための言葉)

▶ まとめ

- ビッグデータ(及びIoT)の対象は現実世界
 - 人材育成ではICT知識だけでなく、現実世界の知見が必要
 - ビッグデータは演繹的より、むしろ機能的
- ビッグデータは魔法ではない
 - 少量データを分析できない組織・人に、大量データの分析は無理
 - データに基づいた判断能力をもつ人材育成が重要
- パーソナルデータなどのビッグデータの負の側面にも着目すべき
 - 匿名化など、プライバシーに関わる分析を防ぐ技術の需要も大きい