NEXT Co., Ltd.

ビッグデータ活用のオモテとウラ ―ヒトの内面に踏み込むことは許されるか？―

株式会社ネクスト　HOME'S事業本部
プロダクト開発部リッテルラボラトリーユニット
主席研究員　　清田　陽司

九州ICT広域連携シンポジウム2014 ～ビッグデータの利活用と将来展望～
2014年12月12日 TKP天神駅前シティセンター（福岡市中央区天神）

自己紹介

清田　陽司（きよた　ようじ）

- 福岡市出身 1975年生まれ
- 研究分野: 自然言語処理応用、情報検索、情報推薦
- 略歴
  - 京都大学大学院情報学研究科 博士課程 (-2004)
    - 質問応答システム（マイクロソフト日本法人との共同研究）
  - 東京大学情報基盤センター 助教・特任講師 (2004-2012)
    - 図書館ナビゲーションシステム
  - 株式会社リッテル (2007-2011) ※東京大学発ベンチャー
    - 図書館ナビゲーションシステム実用化
    - Hadoopベースの大規模データ処理技術
  - 株式会社ネクスト リッテルラボラトリー (2011-)
    - 情報レコメンデーション研究開発

株式会社ネクストのサービス

日本最大級の不動産・住宅サイト HOME'S を運営

「らしく」住もう。
HOME'S
A NEXT Group Service

※1 No1 総物件数

※2 No1 利用者数

※1 フジサンケイビジネスアイ調べ (2014.3.31掲載)
※2 利用者数 No.1 ニールセンNetView 2013年4月データ (家庭および職場のPCからのアクセス・カテゴリ: 家庭とファッションサブカテゴリ:不動産)

Agenda

1. ビッグデータ処理を支える技術的背景
2. ビッグデータ処理が可能とするヒトの行動理解
3. ビッグデータ活用への懸念と解決策

1. ビッグデータ処理を支える技術的背景

ビッグデータ処理技術を理解するキーワード

- 定型処理から非定型処理へ
- データ源の多様化
  （オープンデータ、センサーデータ、…）
- データ処理技術の高度化
  （ベイズ統計、機械学習、…）

定型処理から非定型処理へ

- 定型処理
  - 給与計算、売上集計、伝票処理など
  - ヒトが介在しない完全な自動化が可能
  - 厳密さが求められる
  - データ量はせいぜいGbytesオーダー
- 非定型処理
  - 統計、検索、データマイニングなど
  - 最終的にはヒトの意思決定に利用される
  - カバレッジ重視 (データ量が重要)
  - データ量はTbytes～Pbytesオーダーになり得る

非定型処理が扱うデータ量

データ量単位	バイト数	例
KB（キロバイト）	10^3	GIF画像1枚
MB（メガバイト）	10^6 (100万)	フロッピーディスク1枚の容量
GB（ギガバイト）	10^9 (10億)	1990年代後半のHDD 1台の容量 DVD1枚の容量 新聞記事10年分
	10^{10}	月間100万PVのWebサイトのログ1年分 RDBMSの1テーブル容量の実用上限界
	10^{11}	1台のRDBMSサーバ容量の実用上限界
TB（テラバイト）	10^{12} (1兆)	現在主流のHDD 1台の容量 **Twitterのツイート1年分**
PB（ペタバイト）	10^{15} (1000兆)	**Googleが1日に処理するデータ量**
EB（エクサバイト）	10^{18} (100京)	1998年に全世界に流通したデータ量
ZB（ゼタバイト）	10^{21}	2010年に全世界に流通したデータ量

IBM Watson

2009年、米国のクイズ番組Jeopardy!で優勝

Watsonの裏側

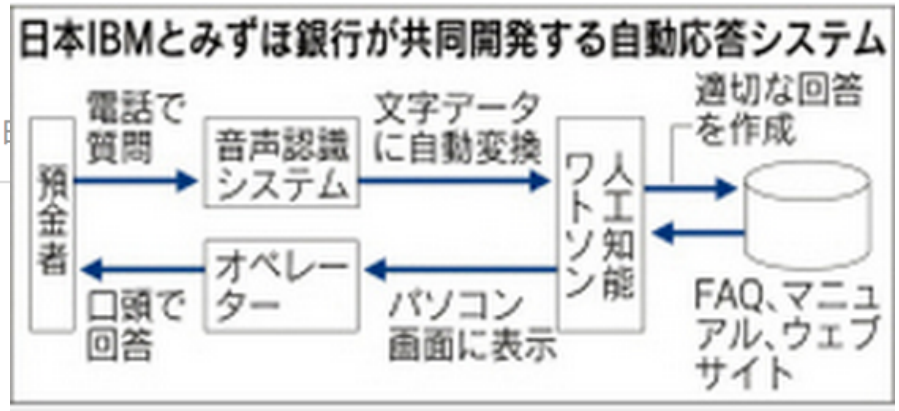
- サーバーの計算能力
  - CPUコア: 2880個
  - 主記憶容量(RAM): 15TBytes
  - 演算速度: 80兆回/秒
- データ（インターネットには接続せず）
  - 2億ページ分の文書（100万冊に相当）
    - 百科事典
    - 書籍
    - 映画の台本

Watsonの応用事例

みずほ銀行、コールセンターに人工知能を導入へ　問い合わせ時間が30分から8分に

The Huffington Post

投稿日: 2014年11月06日 18時43分 JST ｜ 更新: 2014年11月06日



みずほ銀行と日本IBMは11月6日、人工知能コンピューター「Watson（ワトソン）」を利用して、顧客からのコールセンターなどへの問い合わせにタイムリーに回答する世界初のシステムを導入すると発表した。

導入されるのは、問い合わせをしてきた利用者とオペレーターとの会話をシステムが聞き取り、Watsonが適切な回答を見つけるというもの。みずほ銀行の担当者はハフポスト日本版の取材に対し「オペレーター人員が必要無くなるわけではないが、1回の対応時間が平均30分から8分に大幅に短縮できる」と述べた。来年からの導入を検討しているという。

データ源の多様化

- オープンデータの公開・入手・利用が容易に
  - 定義: 「自由に使えて再利用でき、誰でも再配布できる」データ
  - データフォーマットの標準化（機械可読化）
  - ライセンスの整備（オープンライセンス）
- 大量のセンサーデータが入手可能に
  - スマートフォン、タブレットはセンサーのかたまり（タッチ、加速度、位置情報など）
  - Internet of Things (IoT): あらゆるモノからデータがネット経由で収集できる時代

オープンデータ

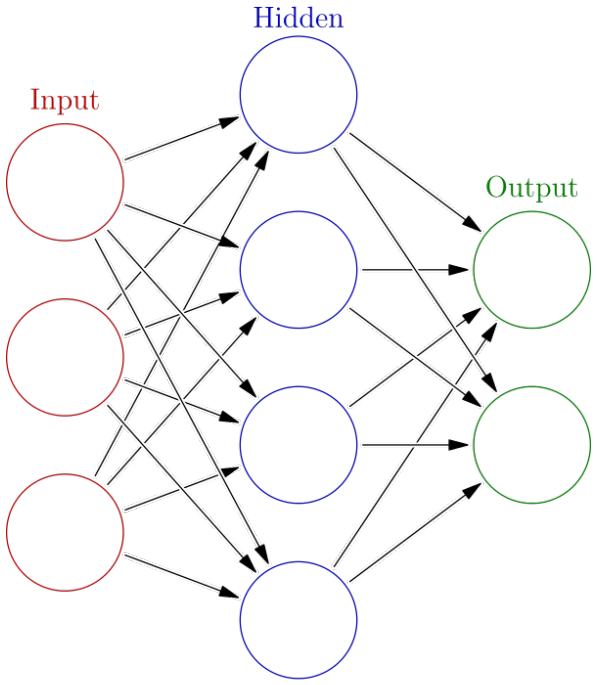
データ処理技術の高度化

コンピューターの性能向上によって、膨大な計算を必要とするアルゴリズムが利用可能に

- ベイズ統計
  - ヒトの「主観」を確率的に表現
- 機械学習
  - 大量のデータの中に埋もれている「規則」を発見する

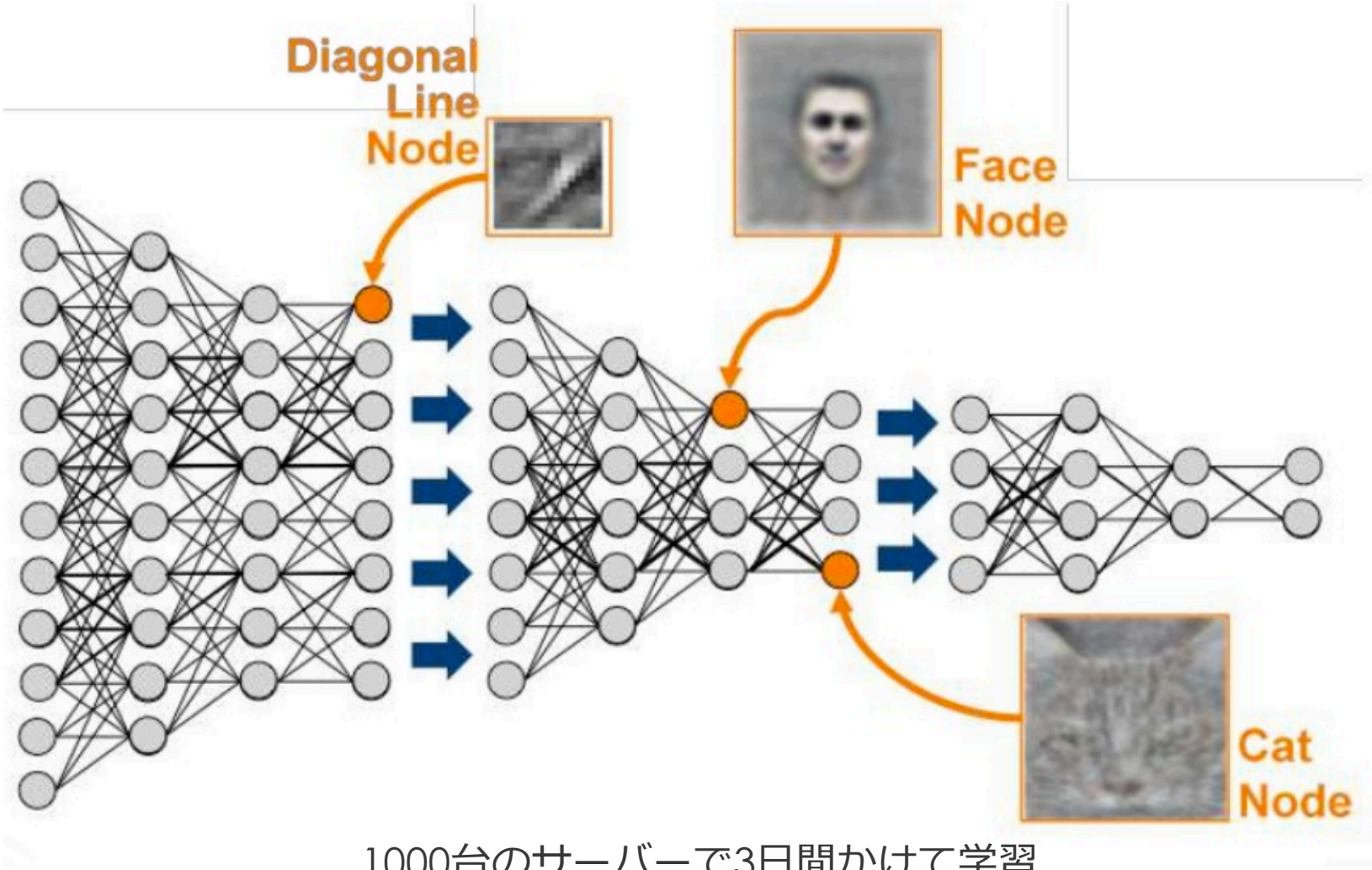
深層学習（Deep Learning）

- 1970年代から研究されているニューラルネットワークのブレークスルー
  - ニューラルネットワーク＝脳神経回路の模倣
    - 一般的に、入力層、中間層、出力層の3層
  - 層の数を増やせば表現能力が上がるが、学習が困難だった
  - 多層でも学習可能な方法が提案された (Hinton 2006)
- 各種の機械学習コンペティションで、他の手法を大きく上回る精度を達成



Google 猫ニューロン

Youtubeから抽出した1000万枚の画像に深層学習を適用



1000台のサーバーで3日間かけて学習
→ 猫の顔、人間の顔に反応するニューラルネットができた

2. ビッグデータ処理が
可能とするヒトの行動理解

ビッグデータ処理が可能とするヒトの行動理解

ヒトの表層的行動のみならず、ヒトの内面へのアプローチも可能になりつつある

- 嗜好・興味
- 性格の類型化
- 行動ターゲティング広告

検索・閲覧履歴によるプロフィールの推定（Google）

https://www.google.com/settings/u/0/ads

Microsoft Bing（検索エンジン）のユーザー行動分析事例

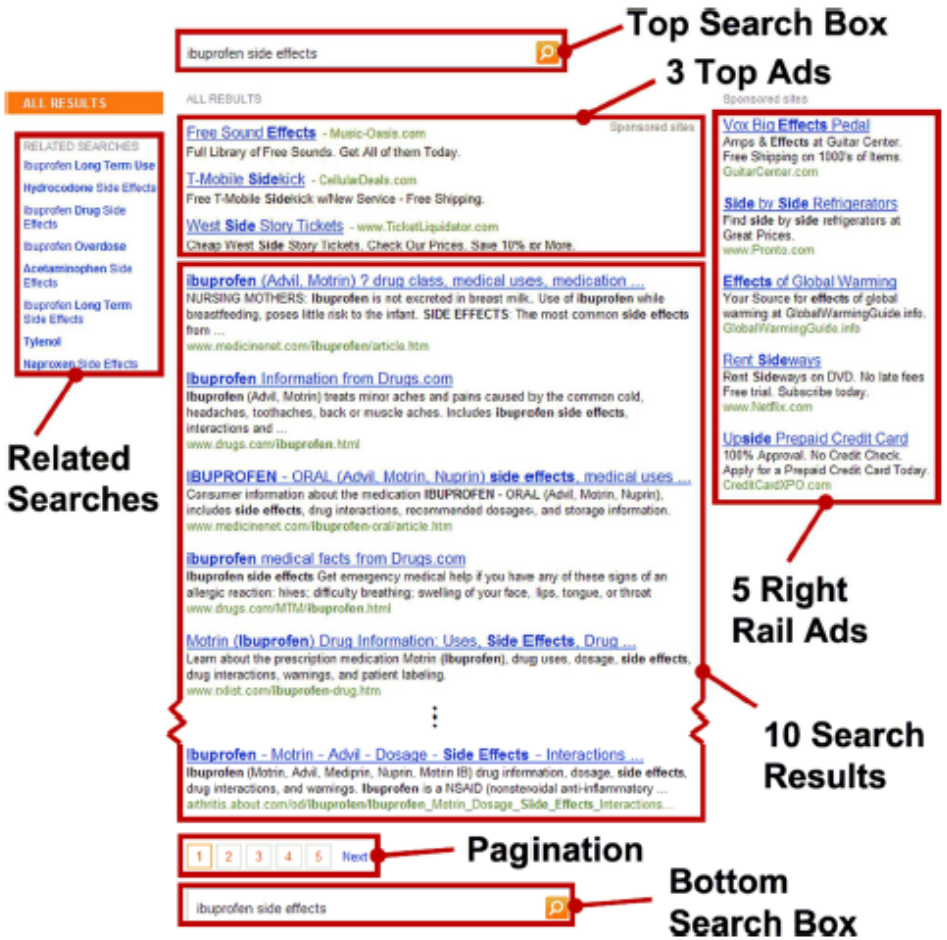


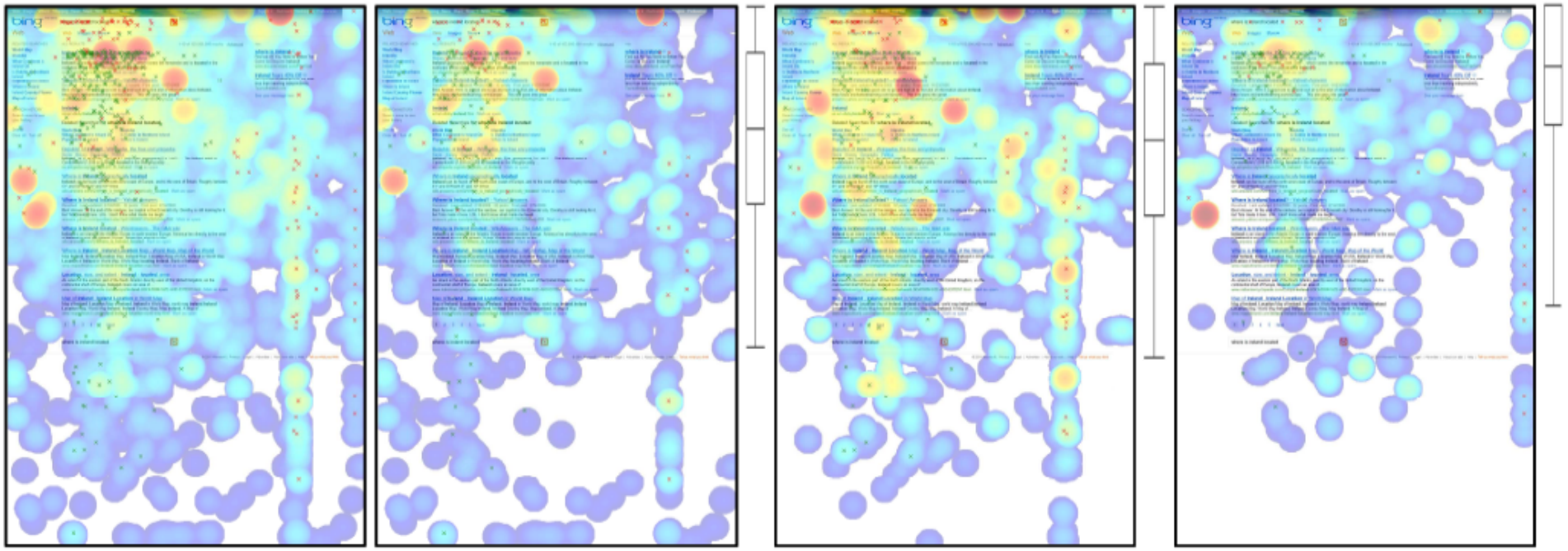
Figure 1. Recorded AOIs on an example SERP.

Georg Buscher et al. (Microsoft Bing)

Large-scale analysis of individual and task differences in search result page examination strategies.

In proceedings of WSDM 2012 (the fifth ACM International Conference on Web Search and Data Mining)

マウス・スクロール軌跡データによるヒートマップ導出



All *Exhaustive-Active* *Exhaustive-Passive* *Economic*

Figure 3. User cluster heat maps. Box and whisker diagrams show the median, first/third quartiles, and min/max scroll depth.

小型カメラ付き自販機（JR東日本ウォータービジネス）

狙い：自販機1台当たりの利益最大化

取得

- マーケティングデータを収集
  - 購入時刻　• 購入商品
  - 属性（年代と性別）　• リピート購入の有無

次世代機＝約30台（上部の顔認識小型カメラで性別と年代を判定）

- センサーで確認した属性に合う飲料をその場で「おすすめ」する機能もある

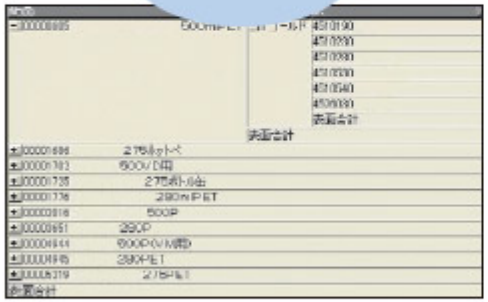
電子マネー対応機（VT-10）＝約2500台

- スイカのIDデータ、「Suicaポイントクラブ」に登録済みのデータで属性判別

分析

- 主に「利用時間帯」「年代」「性別」を軸に顧客動向を分析

高速データ処理エンジン



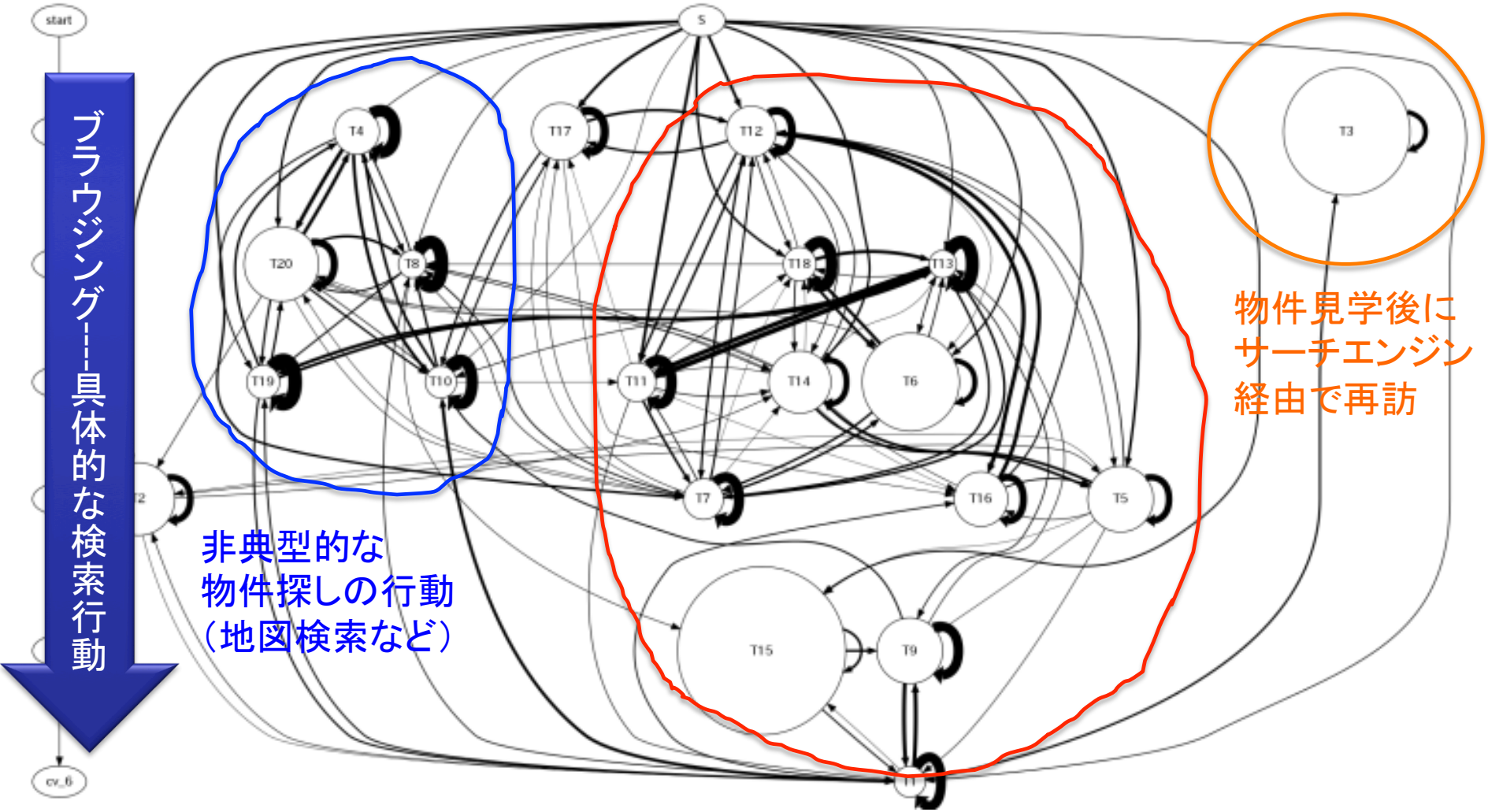
実行

- 自販機立地別の品ぞろえ
  - 繁華街　• オフィス街
  - 郊外の住宅街　など…
- 対象顧客層を絞った独自商品の開発

属性分析の結果、「若い女性」「男性会社員」などに的を絞ってパッケージをデザイン

IT Pro 2012/02/06 「自動販売機が“顧客センサー”に／JR東日本ウォータービジネス」
http://itpro.nikkeibp.co.jp/article/COLUMN/20120123/379107/

ユーザー行動の可視化結果（HOME'S賃貸ユーザーの例）



典型的な物件探しの行動
（路線・駅選択 or 地域選択⇒リスト表示⇒物件閲覧）

3. ビッグデータ活用への懸念と解決策

ビッグデータとプライバシー

- 豊富なユーザーの行動データを収集・活用する流れは加速中
- 一方で、収集・活用がユーザー側との紛争を引き起こす事例も増えている
  - iPhoneのUDID問題
  - 図書館でのポイントカード活用
  - Suicaのデータ販売

プライバシーとは？

- 個人情報保護法 (2005年制定)
  - 個人情報＝個人を特定可能な情報
    - 氏名、性別、生年月日、住所、電話番号、職業、年収、家族構成、IPアドレス、メールアドレス…
  - 「個人情報の有益性に配慮しつつ、個人の権利利益を保護」
- 業界ごとの倫理規程
- 「プライバシー」と「個人情報」
  - 利用履歴、検索キーワード、メールの内容などは「プライバシー」に該当
  - 処理の結果、個人を特定できる可能性があるならば「個人情報」として取り扱う義務あり

プライバシーポリシー

ウェブサイトにおいて、収集した個人情報をどう扱うかについての管理者の姿勢を宣言したもの

- 利用規約の一部として記載されている場合もある
- ウェブサイトの管理者と利用者の間の契約？

プライバシーポリシーをめぐる問題

- そもそも、プライバシーポリシーを熟読する利用者がどれくらいいるのか？
- プライバシーポリシーに書いておきさえすれば、何をやってもいいのか？
  - 炎上につながる事例も

サービス提供者側にとっての2つの視点

- 利用者の意思尊重の視点
  - 「自由」を妨げてはならない
    - 要求されない限り口を出さない
- 専門家としての支援の視点
  - 利用者との間に知識格差があるので、アドバイスをすることが有益
    - 機会がある限り口を出す

リバタリアニズムとパターナリズム

- リバタリアニズム (libertarianism, 完全自由主義)
  - 他人の権利を侵害しない限り、各個人の自由を最大限尊重すべきと考える立場
- パターナリズム (paternalism, 父権主義)
  - 弱い立場の者の利益を保護するために、強い立場の者が介入することを許容する立場

パターナリズムが批判される理由

- 当事者にとっての「最善」が何かを強い立場の者が判断できるとは限らない
- 自己決定権の侵害になりかねない

リバタリアン・パターナリズム (1993年)

BEHAVIORAL ECONOMICS, PUBLIC POLICY, AND PATERNALISM†

Libertarian Paternalism

By RICHARD H. THALER AND CASS R. SUNSTEIN*

Many economists are libertarians and consider the term "paternalistic" to be derogatory. Most would think that the phrase libertarian paternalism is an oxymoron. The modest goal of this essay is to encourage economists to rethink their views on paternalism. We believe that the anti-paternalistic fervor expressed by many economists is based on a combination of a false assumption and at least two misconceptions. The false assumption is that people always (usually?) make choices that are in their best interest. This claim is either tautological, and therefore uninteresting, or testable. We claim that it is testable and false—indeed, obviously

appears to be paternalistic, which it is, but would anyone advocate options 2 or 3?

The second misconception is that paternalism always involves coercion. As the cafeteria example illustrates, the choice of which order to present food items does not coerce anyone to do anything, yet one might prefer some orders to others on paternalistic grounds. Would many object to putting the fruit before the desserts at an elementary school cafeteria if the outcome were to increase the consumption ratio of apples to Twinkies? Is this question fundamentally different if the customers are adults? If no coercion is involved, we think that some types of pater-

主旨

- 現実世界では、パターナリズムを完全に避けることはできない
  - だったら、パターナリズムの考え方もうまく取り入れるためのデザインを考えた方が建設的ではないか？
- 要は「デフォルトの設定」をどう活用するか？
  - 利用者にとっての利益を最大化するためには、どういうオプションをデフォルトにするといいのか？

パターナリズムはなぜ避けられないか？

- 学食や社食のカフェテリアコーナー
  - メニューの配置の順番は利用者の選択に影響を与えてしまう
  - そもそも配置の順番には恣意性が絡む
    - デザートの手前にフルーツを置くとか
- 401(k)プラン（確定拠出年金）への勧誘の例
  - オプトイン（デフォルト非加入、意思表示により加入）にするか、オプトアウト（デフォルト加入、意思表示により脱退）にするか？

リバタリアンの要求を満たすパターナリズムの実装

- 利用者にとっての利益を最大化するオプションをデフォルトにしておく
- ただし、利用者の自由意思によってデフォルトオプションを拒否できることは明示されている

利用者にとっての「利益」が自明でないときは？

利益を間接的に表現する代替手段を使う

1. 多数が選択するオプションをデフォルトに
   - 要は市場的アプローチ
     - 多数による選択が informed の状況でない場合が問題
2. 明示的な選択を強制
   - 1がうまくいかない場合に検討の余地あり
     - 利用者の熟慮を引き出せるかどうか？
3. オプトアウトの数を最小化
   - オプトインにしても401(k)の加入者が大多数の場合は、加入をデフォルトにすることは正当化できるだろう

「真の自由」を実現するには？

- カント哲学の視点によるリバタリアニズム批判
  - 「自律的な判断」でない限りは、真に自由が尊重されているとはいえない
    - 「直感のみで判断している状態は他律的」というのがカント哲学の立場
- 信頼関係が成り立っている上での「デフォルトの設定」なしには、真の自由は実現しない？
  - 理性的な判断には多大な労力が必要
  - 労力は有限

考えられるシナリオ

収集・活用されることが既成事実になる

コンシューマー側との
コミュニケーションによる解決

- ユーザ自身がデータの収集および利用の意義を理解できていること
- プライバシーポリシーへの記載だけでなく、ユーザエクスペリエンスの中に埋め込まれている

紛争の激化により、収集・活用が頭打ちに

まとめ

まとめ

人間と社会の情報活用能力向上に、ビッグデータ処理はどう貢献できるか？

- 付加価値の創出と、社会からの要請への対応を両立するためのコミュニケーション
- ユーザー行動の背後にある心理を理解しながら、適切な支援を提供するためのデザイン
  - ログデータの分析・活用には、学際的な知見が必要とされる