サービス経済におけるデータ科学

照井伸彦
東北大学大学院経済学研究科

実験家のためのデータ駆動科学オンラインセミナー
２０２０年5月２８日

目次

1．背景

・経済のサービス化

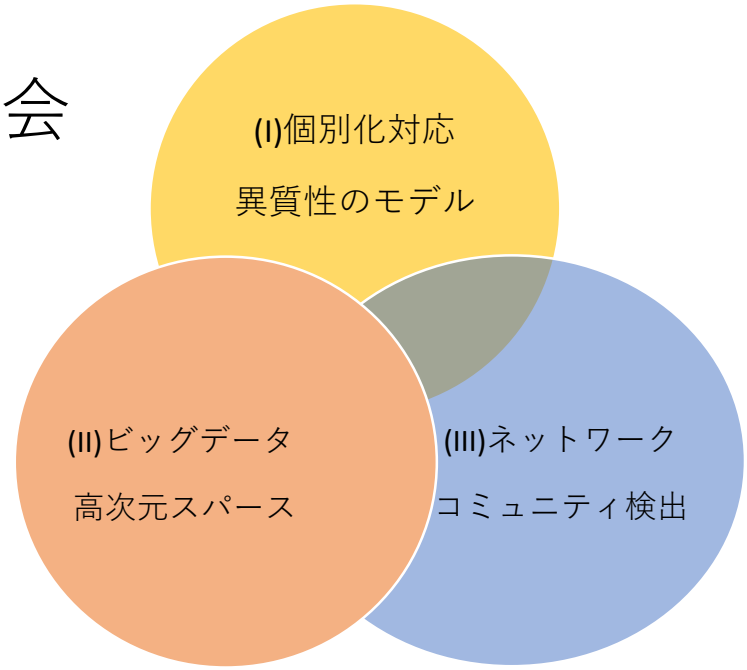
・ビッグデータと超スマート社会

2. 求められるデータ科学手法

(I) 個別化対応
異質性のモデル

(II)ビッグデータ
高次元スパース性

(III)ネットワーク
テキスト解析も利用したコミュニティ検出



ビッグデータの出現と活用

デジタルデータ量の増加予測

- デジタルデータの量が急速に増加している
- ２０２０年までに約４０**ゼタ（40×10^{20}）バイト**
- ２０３０年までに**ヨッタ（10^{24}）バイト**

理由

〇ブログやSNSなど利用者参加型メディア情報

〇IoT(internet of Things)の登場

あらゆるモノがインターネットで繋がる

・GPS端末の位置情報

・交通系ICカードの乗車履歴

・各種センサからの温度・圧力など物理量

・会員カードによる購買履歴

e.t.c.

例：Tポイントカード

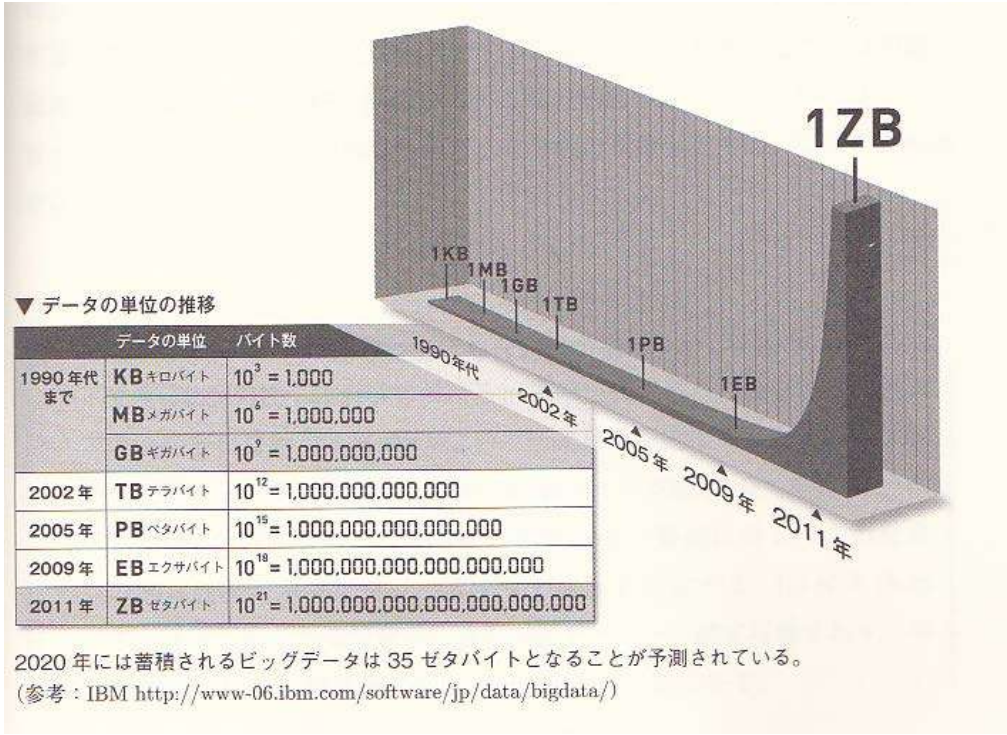
131社・約45万店舗の提携先から

5,556万人の会員の

1)所在地、2)利用ポイント状況、

3)リピート率, 4)購入単価

を提携企業へ提供

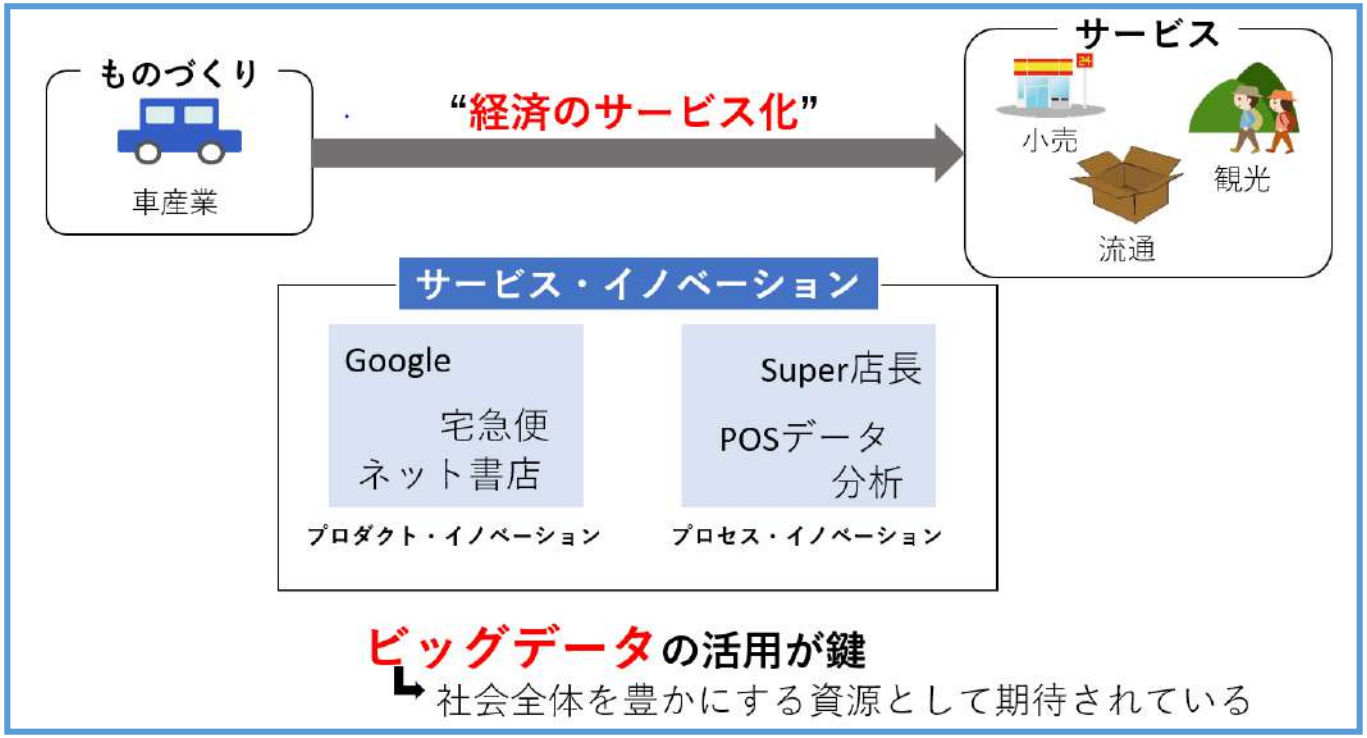


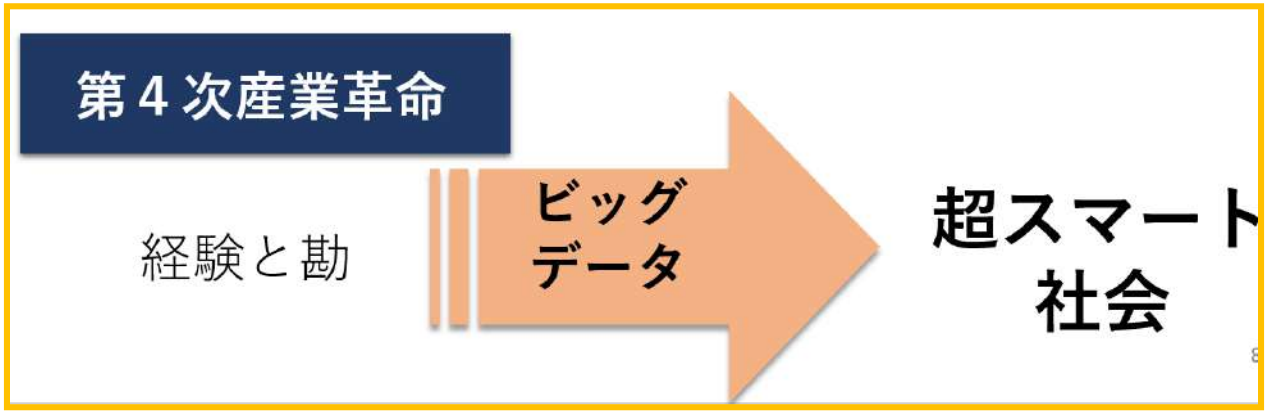
▼ データの単位の推移

	データの単位	バイト数
1990年代まで	KB キロバイト	10^3 = 1,000
	MB メガバイト	10^6 = 1,000,000
	GB ギガバイト	10^9 = 1,000,000,000
2002年	TB テラバイト	10^{12} = 1,000,000,000,000
2005年	PB ペタバイト	10^{15} = 1,000,000,000,000,000
2009年	EB エクサバイト	10^{18} = 1,000,000,000,000,000,000
2011年	ZB ゼタバイト	10^{21} = 1,000,000,000,000,000,000,000

2020年には蓄積されるビッグデータは35ゼタバイトとなることが予測されている。
（参考：IBM http://www-06.ibm.com/software/jp/data/bigdata/）

経済のサービス化とビッグデータ





超スマート社会

―第５期科学技術基本計画（Society 5.0）

出典：科学技術イノベーション総合戦略2015における重点化対象施策について（平成27年9月18日 内閣府政策統括官）

マーケティング：
パーソナライゼーション
個別化対応

定義

個別のシステムが更に高度化し、分野や地域を越えて結びつき、
３次元の地理データ、人間の行動データ、交通データ、環境観測データ、もの作りや農作物等の生産・流通データ等の**多種多様で大量のデータ（ビッグデータ）を適切に収集・解析し、横断的に活用することにより**、
「必要なもの・サービスを、必要な人に、必要な時に、必要なだけ提供し、社会の様々なニーズにきめ細やかに対応でき、あらゆる人が質の高いサービスを受けられ、年齢・性別・地域・言語といった様々な違いを乗り越え、活き活きと快適に暮らすことができる社会」

I.個別化対応（異質性のモデリング）

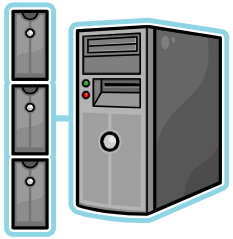
【日本のサービス産業】

- 実質ＧＤＰ・雇用の約７割
- 生産性成長率：低い　小売業の利益率：さらに低い

【生産性向上のために】

- 個人特性を考慮したサービス提供（No！低価格化戦略）
- “個”の理解　⇒　消費者異質性モデル
- ビッグデータを活用した“個”の理解

購買行動に関する
ビッグデータの蓄積



しかし現状は･･･、

データは未活用
社会的損失状態

ビッグデータの活用

“個”客の理解システム

企業生産性向上
顧客満足度向上

個別広告・プロモーションの成功例

amazon.co.jp

こんにちは。おすすめ商品を見るにはサインインしてください。初めての方はこちら。

マイストア　Amazonポイント　ギフト券　お買い得情報　アカウントサービス | ヘルプ

すべてのカテゴリーを見る　検索 和書　GO　カートを見る　ほしい物リスト

和書　詳細検索　ジャンル　新刊・予約　ベストセラー　ハリー・ポッター　雑誌　バーゲン　Amazon で売ろう

Amazonプライム

Amazonプライムへの会員登録はお済みですか？ご登録はこちらから

この商品を火曜日(2月17日)までに受け取りたい場合は、18時間 55分以内にご注文ください。

お急ぎ便配送料無料で今すぐ注文する

詳細

イメージを拡大

自分のイメージを掲載する

出版社,著者の方へ:「なか見！検索」で書籍を紹介しませんか?

ベイズモデリングによるマーケティング分析 (単行本)

照井 伸彦 (著)

まだカスタマーレビューはありません。最初のレビューを書く

価格：￥ 3,570 国内配送料無料(一部例外あり) 詳細

ポイント：35pt (1%) 詳細はこちら

在庫あり。在庫状況について

この商品は、Amazon.co.jp が販売、発送します。ギフト包装を利用できます。

4点在庫あり。ご注文はお早めに。

2009/2/17 火曜日 にお届けします！ 今から18 時間 と55 分以内にレジに進み、「お急ぎ便」オプション(有料)を選択して注文を確定されたご注文が対象です。詳しくはこちら
一部サイズが大きい商品の場合、上記の日付が適用されない場合があります。配送オプションを選択する画面、もしくは注文確定するときに必ず配送予定日をご確認ください。

新品1点 ￥ 3,570より

予約注文・限定版／初回版・特典に関する注意

Click here to see in English.

数量: 1

ショッピングカートに入れる

または

1-Clickで注文する場合は、サインインをしてください。

こちらからも買えますよ

この商品を安く買いたい！

中古商品を予約する

この商品をお持ちですか？

マーケットプレイスに出品する

ほしい物リストに追加する

お友だちに知らせる

この商品を買った人はこんな商品も買っています

ページ: 1 / 17

ベイズ統計データ分析—R & WinBUGS (統計ライブラリー) 古谷 知之

(3) ￥ 3,990

マルコフ連鎖モンテカルロ法 (統計ライブラリー) 豊田 秀樹

￥ 4,410

データマイニング入門 豊田 秀樹

(1) ￥ 3,570

消費者行動論体系 田中 洋

(2) ￥ 3,045

入門ベイズ統計—意思決定の理論と発展 松原 望

(3) ￥ 3,360

Rによるテキストマイニング入門 石田 基広

￥ 2,940

戦術モデルと戦略モデル

ルールベースの意思決定は**戦術モデル**

計画を可能とする**戦略モデルへ**

因果に基づく計量モデル
制御モデル: Y=f(X)

(1) 個性の推定/個別化対応
(2)大規模データ対応：次元圧縮
(3)逆問題への対応：結果から原因の推論
(4)情報更新/融合して新しい知へ
＝＞**基盤技術としてのベイズ統計**

ビッグデータはごみの山

情報量は浅い(Shallow)
a)人間の個性や行動の振れ幅が大きい
b) 社会経済システムに大局的・安定的物理定数は想定しがたい

複雑な事象に潜伏する構造のモデリング⇔実在論的アプローチ

有用な情報を取りだす機能(網,フィルタ,プリズム)必要

(1) 先験的知識（事前分布）:経験, 理論, 合意
(2) 経験データ（尤度）
＝＞両者を統合して複雑事象の認識（事後分布): 情報更新
＝＞検証,予測,制御

I.個別化対応（異質性のモデリング）

ID付POSデータ:メンバーシップ顧客の購買履歴データ

○データ情報

データに記録される顧客行動データ（尤度関数）

＝＞個別対応するには情報不足

$$Y_h = f\left(X_h \beta_h\right) + \varepsilon_h \quad h = 1, \ldots, H$$

○事前情報

「各消費者は異質ではあるが共通する部分もある」

共通の知見を利用（事前分布）

$$\beta_h = \overline{\beta} + \varepsilon_h$$

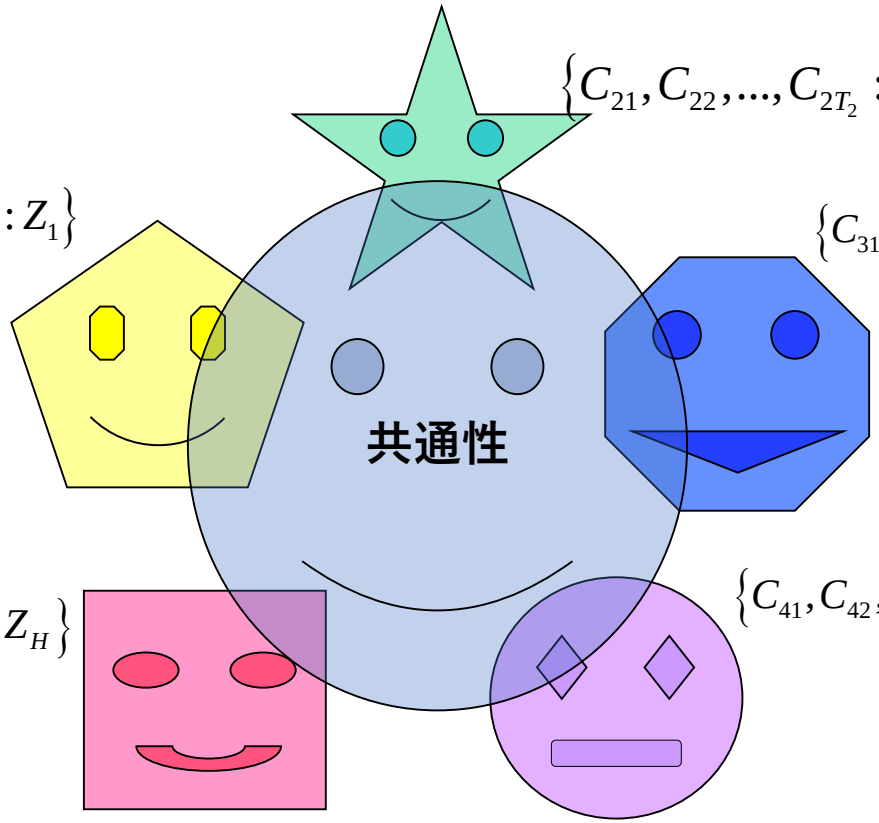
○データの持つ情報を「異質性」と「共通性」とにバランスよく分配し、異質性を推定するのに不足する情報を共通性として消費者全体をプールした情報で補う

＝＞ **階層ベイズモデル** の利用：事後分布による情報更新

I.個別化対応（異質性のモデリング）

共通性と異質性

C:行動データ
Z:属性データ



マーケティングのデータ：
➔多くの意思決定主体に関する情報
（パネル，サーベイ）
➔ 各主体のデータが少ない
➔ 全体で集計
➔ 各主体の異質性を無視

➔ 各主体間の情報をプールするモデル
例えば　ランダム効果モデル

$$\beta_h = \bar{\beta} + \varepsilon_h, \qquad \beta_h = Z_h\theta + \varepsilon_h,$$

・典型的構造
（１）主体内行動　➔ 尤度関数
（２）主体間行動　➔ 異質性の分布
（３）意思決定モデル

II.高次元スパースデータ

次元圧縮

・因子(主成分)モデル

・トピックモデル

＝＞次元圧縮後の密な情報空間で構造推定

（1）ＰＯＳデータ（商品×時間）

目的：大規模市場反応の計測と店舗マネジメント

(i)自然言語処理の手法を売上数量データに適用（トピックモデル）

=>購買の文脈を考慮しながらマーケットバスケットを構成（次元圧縮）

(ii)マーケットバスケットの中でNP問題を解きながら効果的戦略を探る

（因子空間での階層回帰モデルを推定し高次元の元空間へ還元）

（2）ID-POSデータ（顧客×商品×時間）

目的：個人の選択行動モデルの大規模化とマーケティング個別対応

(i)トピックモデルによる次元圧縮

(ii)変分ベイズによる高次元積分評価近似

(1)POSデータの高度利用研究

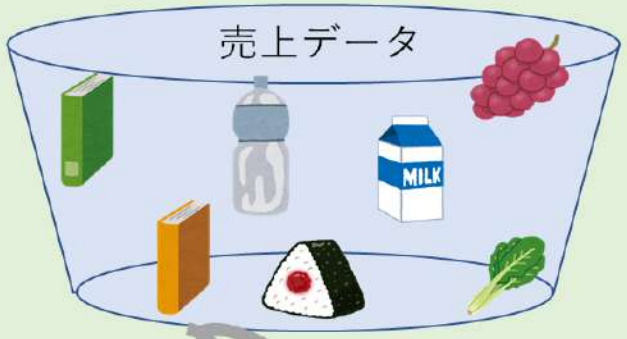
大規模市場反応の計測と店舗マネジメント

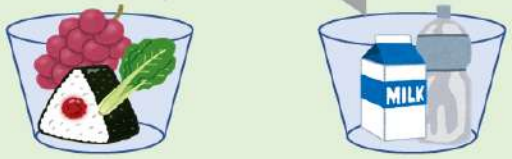
Terui and Li (2019) , Journal of Forecasting, vol.38, 440-458

ビッグデータ解析上の問題点

1. ゼロが多い＝＞スパース（疎）性
2. 変数が多い＝＞NP問題

スパース性





トピック1
$Y_1 = F(X_1) + \varepsilon_1$

トピック2
$Y_2 = F(X_2) + \varepsilon_2$

・・・

トピック10
$Y_{10} = F(X_{10}) + \varepsilon_{10}$

P個

N個

$$Y_1 = \beta_1 X_{1,1} + \cdots + \beta_P X_{P,1}$$
$$Y_2 = \beta_1 X_{1,2} + \cdots + \beta_P X_{P,2}$$
$$\vdots$$
$$Y_N = \beta_1 X_{1,N} + \cdots + \beta_P X_{P,N}$$

$N < P$問題

(1)POSデータの高度利用研究

大規模市場反応の計測と店舗マネジメント

Terui and Li (2019) , Journal of Forecasting, vol.38, 440-458

- スパース性とNP問題をどう解決するか

$N = 363$　：365日
$P = 39{,}560$：7,912商品×5マーケティング変数

高次元スパース回帰　$Y = FX + \varepsilon$

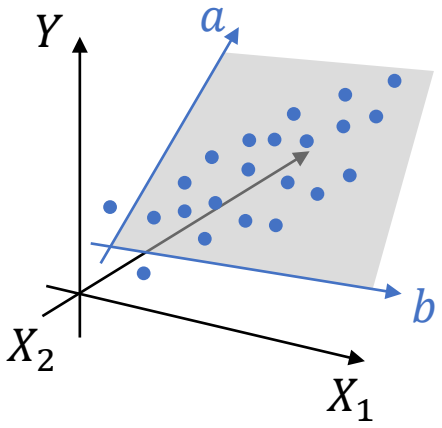
因子モデル
$$Y = Ua + \eta_y$$
$$X = Vb + \eta_x$$

階層回帰モデル　$a = Hb + e$

1. 次元削減

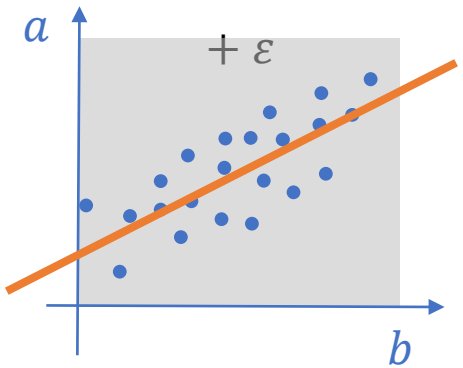
トピックモデル＋因子モデル

$Y \to a, X \to b$



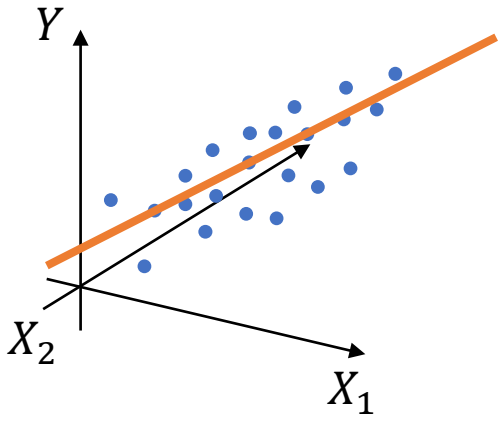
2. 密な空間での推定

$a = Hb + e$



3. 元空間への復元

$F = f(H, a, b)$



(1)POSデータの高度利用研究

大規模市場反応の計測と店舗マネジメント

Terui and Li (2019) , Journal of Forecasting, vol.38, 440-458

・結果の解釈

-トピック９（９月～１１月）

牛乳の売上

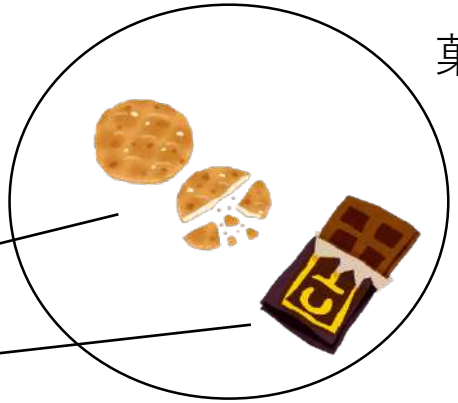
-0.96

-0.64

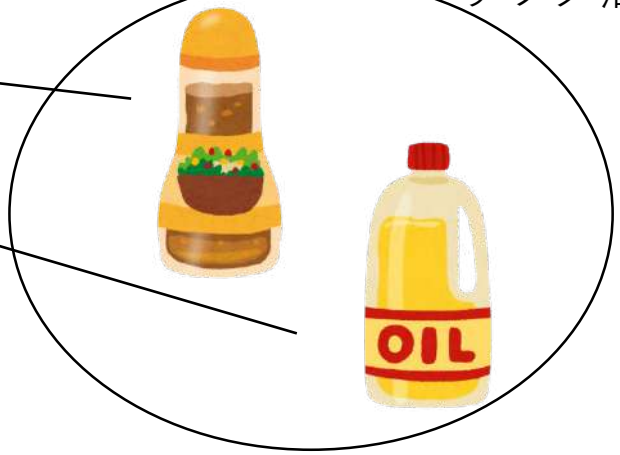
0.98

1.19

菓子類の販促



ドレッシング・サラダ油の売上



サラダ油の売上が1%上がると牛乳の売上が1.19%上がる

意外な組み合わせの発見

＋関係性の定量把握

70年代の第1次人工知能ブーム

米国：（紙おむつ）と（ビール）の同時購買

意外な組み合わせのみ

(2) ID付きPOSデータ:
マーケティングの大規模パーソナライゼーション

Ishigaki, Sato, Allenby and Terui(2019), *International Journal of Data Science, vol.5, 223-248*

【ID付きPOSデータ】

- 誰が,いつ,何を,何個,いくらで購買したか
- 購買・非購買のカウントデータ
- 顧客数、取り扱い商品数が大
- データ量は多い

【データ空間　>>　データ量】

- 「顧客」×「商品」×「時間」の3次元空間
  - ほとんどの顧客がほとんどの商品を買わない
  - データ空間の大きさに対して、購買データはスカスカ

大規模スパースデータ

(2) ID付きPOSデータ: マーケティングの大規模パーソナライゼーション

Ishigaki, Sato, Allenby and Terui(2019), *International Journal of Data Science, vol.5, 223-248*

機械学習（トピックモデル）による次元圧縮
- 圧縮空間でパラメータ推定
- 高速な近似アルゴリズム（変分ベイズ法）
- 元空間へ情報を還元し、効用の値を補完

$$\begin{cases} u_{cit}^{(z)} \sim N(\mathbf{x}_{it}^T \boldsymbol{\beta}_{zi}, 1) \\ \boldsymbol{\beta}_{zi} \sim N(\boldsymbol{\mu}_i, V_i) \end{cases}, \quad \begin{cases} u_{cit} > 0, (y_{cit} = 1) \\ u_{cit} \le 0, (y_{cit} = 0) \end{cases}$$

プロビットモデル

z_{cit} ： （顧客c 商品 i 時刻 t）の状態

$u_{cit}^{(z)}$:z_{cit} = zの状態のとき、顧客c が時刻 tで商品 iに対して持つ効用

$p(z_{cit}=z|c)$: 顧客 c がセグメント z に入る確率

$$p(y_{cit} \mid c, \mathbf{x}_{it}) \equiv \sum_{z=1}^{Z} p(u_{cit}^{(z)} \mid z_{cit} = z)\, p(z_{cit} = z \mid c), z_{cit} = \{1, \cdots, Z\}$$

カテゴリカル分布

The number of customers

The number of products

$$\begin{bmatrix} P(y_{11t}=1|c=1) & \cdots & P(y_{C1t}=1|c=C) \\ \vdots & \ddots & \\ P(y_{1it}=1|c=1) & \cdots & P(y_{Cit}=1|c=C) \end{bmatrix}$$

Large-scale matrix

The number of products

=

The number of latent class

$$\begin{bmatrix} P(u_{1it}^{(z)} \mid z_{1it}) \\ \vdots \\ P(u_{Cit}^{(z)} \mid z_{Cit}) \end{bmatrix}$$

The number of customers

$$\begin{bmatrix} P(z_{1it} \mid c) \\ \vdots \\ P(z_{Cit} \mid c) \end{bmatrix}$$

The number of latent class

特異値分解による次元圧縮

(2) ID付きPOSデータ: マーケティングの大規模パーソナライゼーション

Ishigaki, Sato, Allenby and Terui(2019), *International Journal of Data Science, vol.5, 223-248*

【“個”客の理解】

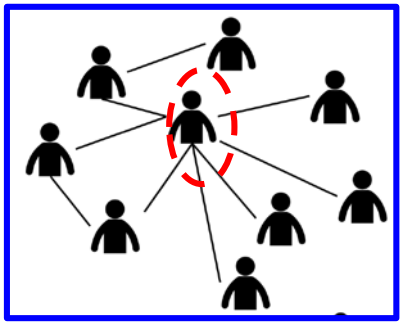
- 95% 信頼区間での有意な係数
- 個人毎、商品毎に有効な施策（個別化）を選択可能に
- 全体の最適化へ

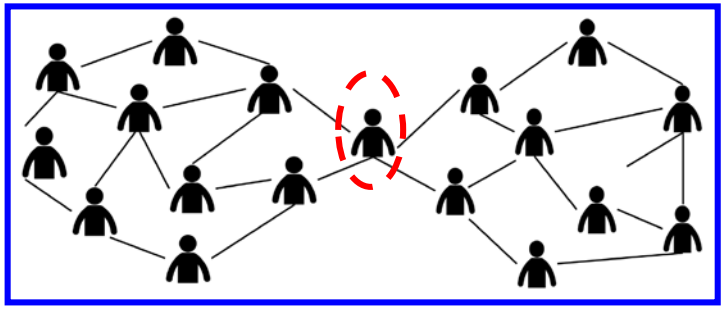
価格							エンド陳列							チラシ掲載						
		商品番号						商品番号							商品番号					
		18	110	253	318	742		18	110	253	318	742			18	110	253	318	742	
顧客	(a)	*	*	*	*	*	顧客	*	*	*		*	顧客	(a)	*	*		*		
	(b)	*	*	*		*			*	*		*		(b)		*		*		
	(c)	*	*			*		*	*	*				(c)		*		*		
	(d)	*	*			*		*	*	*		*		(d)		*		*		
	(e)	*	*		*	*		*	*	*				(e)		*		*		

III.ネットワークとテキスト情報

ネットワークデータとテキストデータの同時利用

- 解釈可能なコミュニティの検出
- インフルエンサー発見
- ブローカー,ストラクチャーホール(SH)発見
- “弱いつながりの力(SWT)”の可視化





◎経営のソーシャルネットワーク研究

知識の探索(弱いつながり)＝＞イノベーション創発

知識の深化(強いつながり)＝＞実践・製品化

III.ネットワークとテキスト情報

マーケティングにおける自然言語処理の役割

(Berger et al. 2020)

- **消費者　＜＝＞　消費者**
  - ➢ カスタマーレビュー、ソーシャルメディア・ブログの書き込み
- **企業　＝＞　消費者**
  - ➢ 広告、エンタメ商品（映画、音楽、本）
- **消費者　＝＞　企業**
  - ➢ 不満の声、アンケート調査

・消費者の関心（トピック）、購買行動・態度への影響を分析

マーケティング：「personalityを知ること」

➢デモグラフィック・アンケート・行動データが用いられる

➢ソーシャルメディア上における社会ネットワークの形成とコンテンツの生成

➢非構造かつ大規模で情報量豊富なデータ

III.ネットワークとテキスト情報

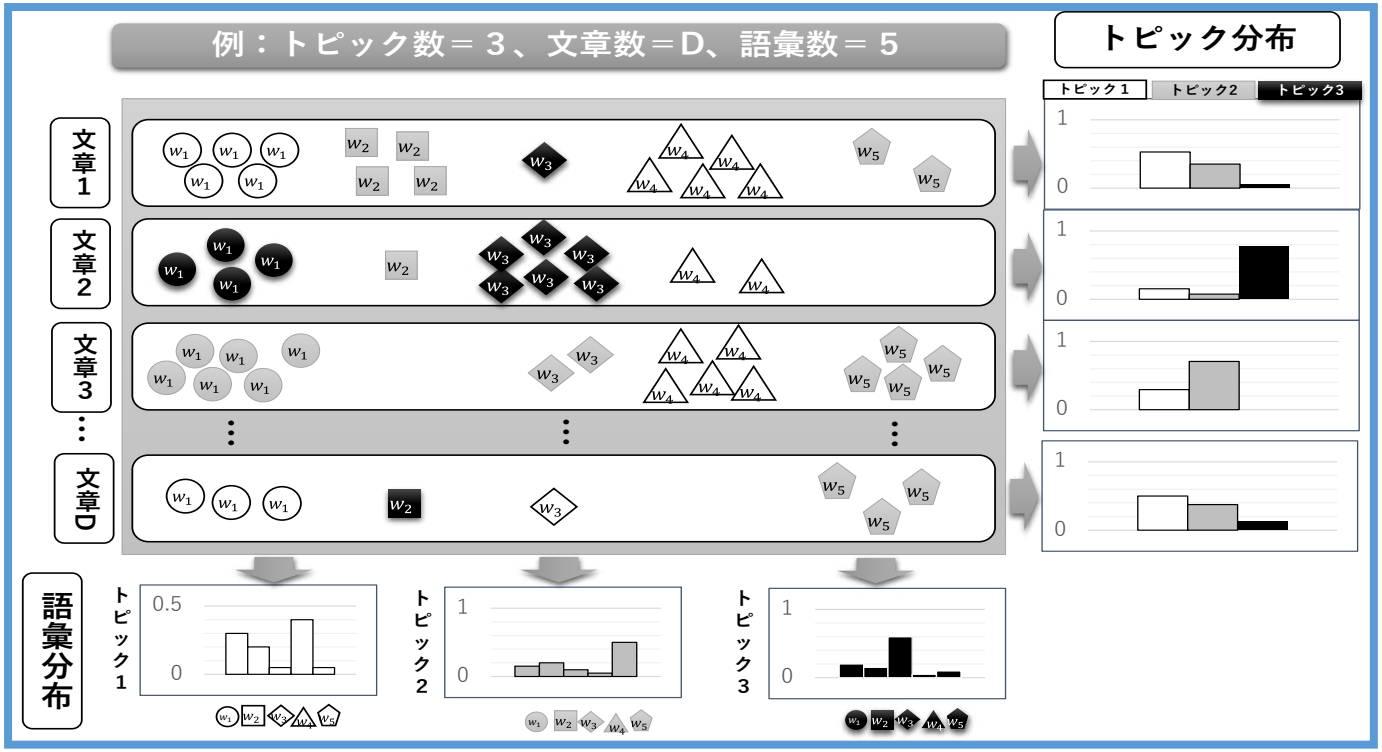
○自然言語処理

・非構造データのテキスト解析

・文書中の単語の同時頻出パターンから背後にある意味を抽出（潜在的意味解析）

・1つの文書に複数のトピックがあり、文書中の各単語にはトピックが確率的に割り当てられる。

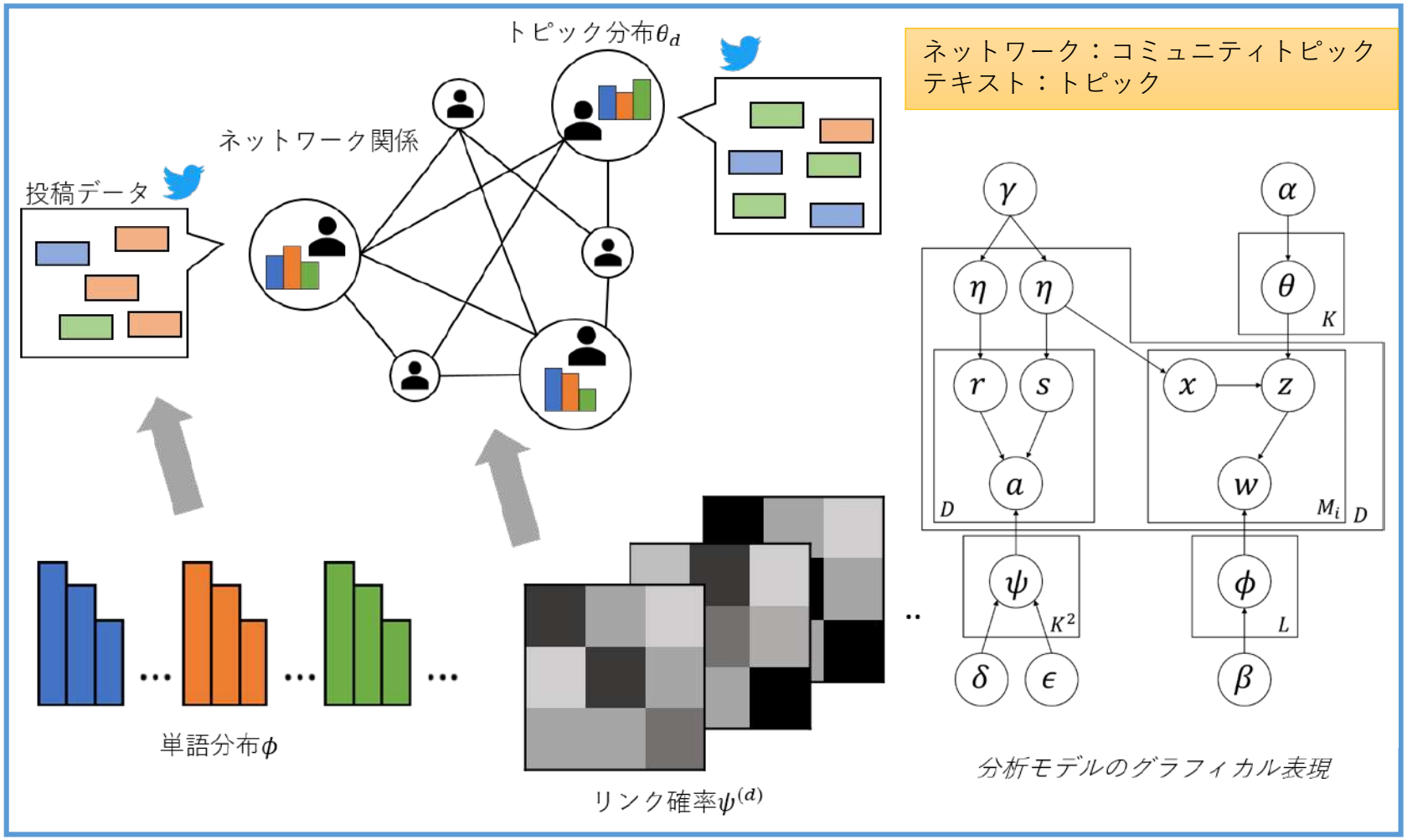
＝＞文書の潜在的意味を機械的に抽出



ネットワーク/テキストデータの同時利用よるコミュニティ検出

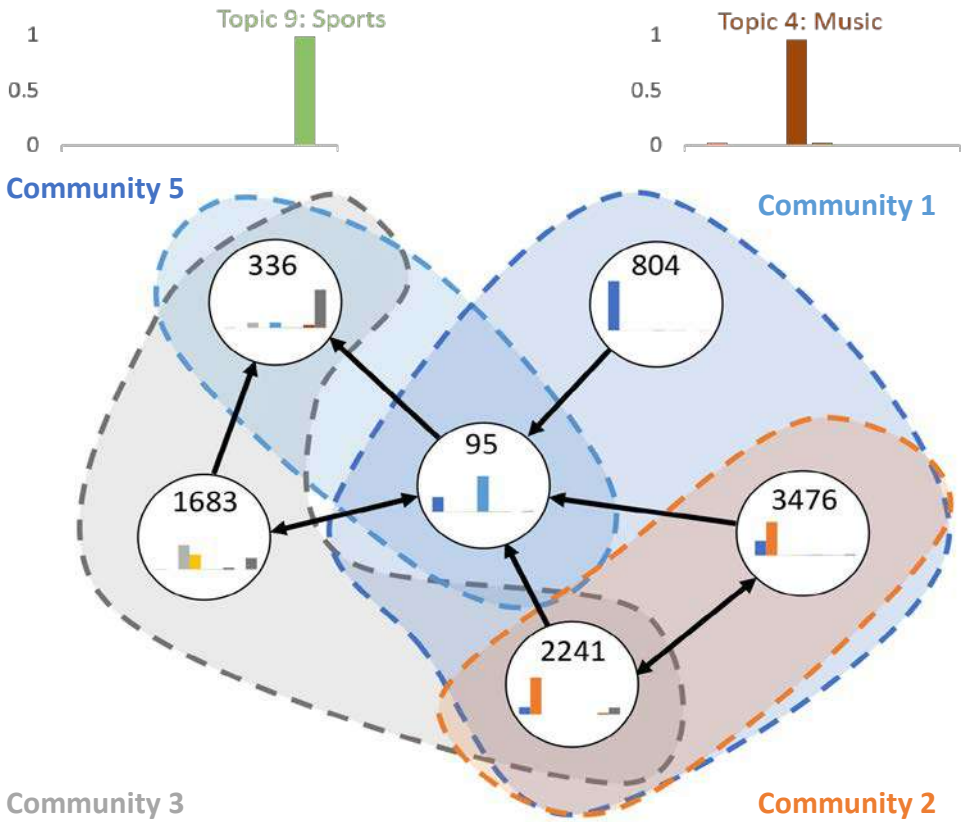
Igarashi and Terui(2020), *Statistics and Computing*

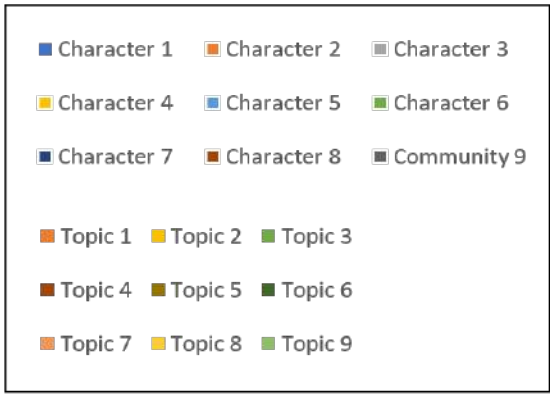
ネットワークデータとUGC (User Genrated Contents)テキストを考慮した
トピックベース・コミュニティ検出モデル



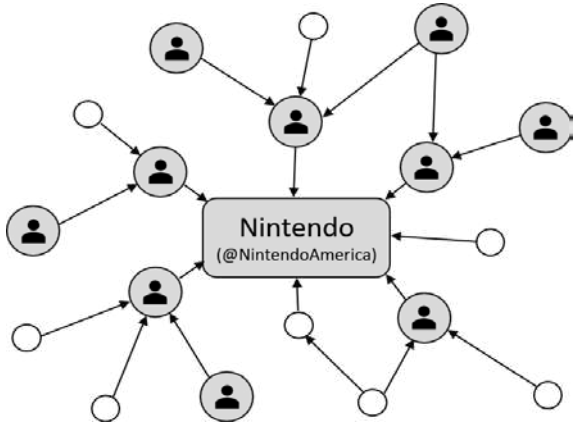
ネットワーク/テキストデータの同時利用よるコミュニティ検出

Igarashi and Terui(2020), *Statistics and Computing*





実証分析で用いるTwitterデータは以下で構成：
1.任天堂アカウントを中心とするネットワーク
（2018年5月1日におけるフォロー関係）
2. タイムライン上に投稿されたTweets
（2017年9月1日から2018年2月28日）

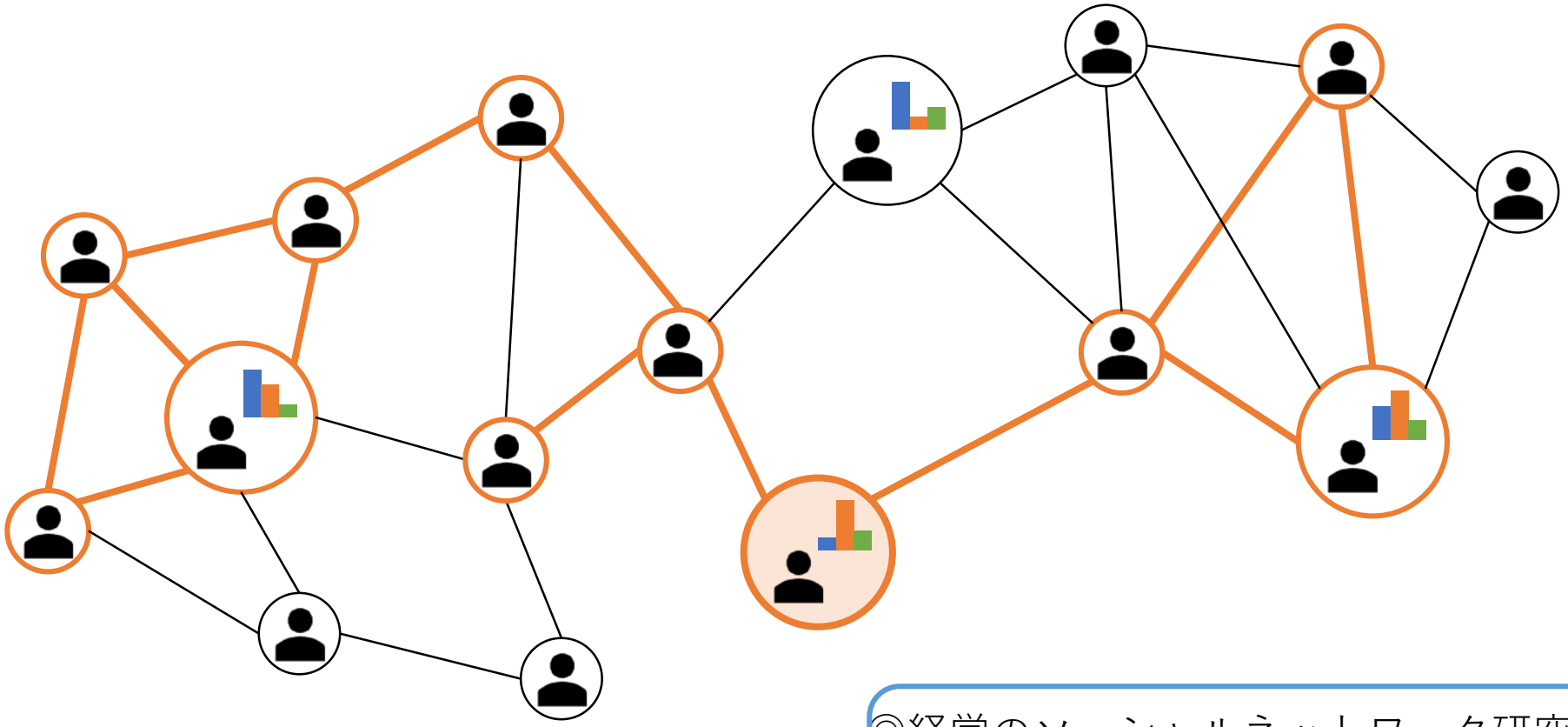


ネットワーク/テキストデータの同時利用よるコミュニティ検出

Igarashi and Terui(2020), *Statistics and Computing*

今後の展開

・ブローカー,ストラクチャーホール（SH)発見
・弱いつながりの力（SWT）の可視化



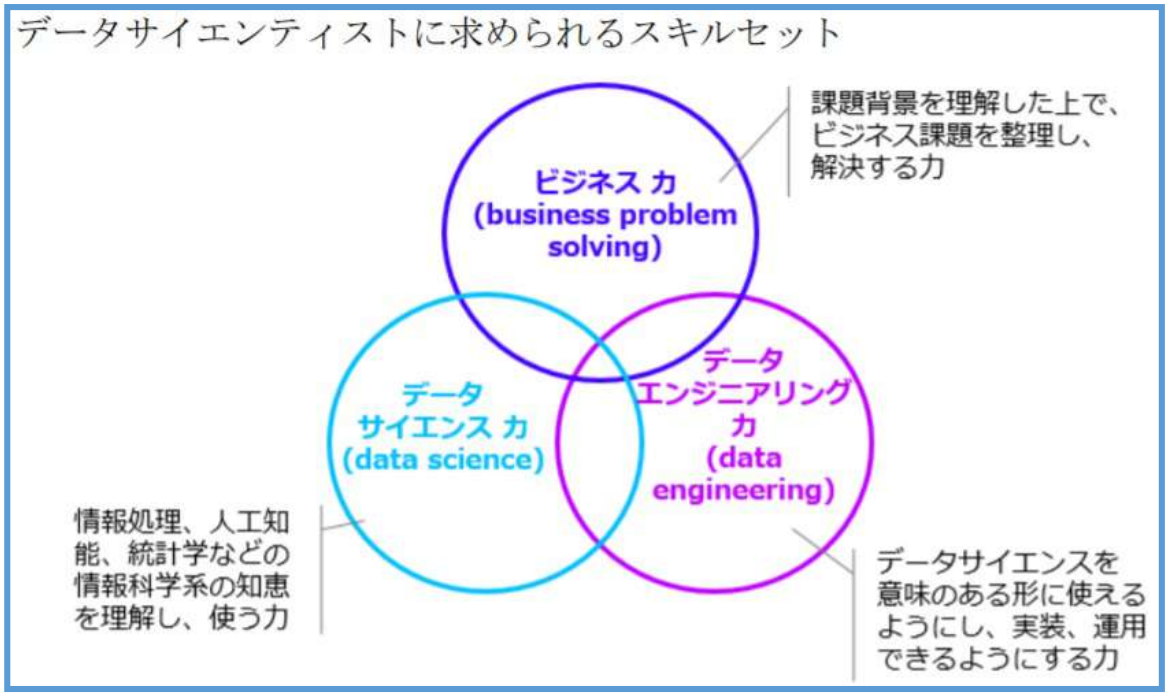
◎経営のソーシャルネットワーク研究

知識の探索（弱いつながり）＝＞イノベーション創発
知識の深化（強いつながり）＝＞実践・製品化

ビッグデータ時代に求められる人材

- IBMによる Thinking by Numbers の推進
- Googleチーフ・エコノミストHal Varian
- 「次の１０年で最もセクシーな仕事は統計家である」

→ビッグデータ時代に必要な“**データサイエンティスト**”



データサイエンティスト協会より抜粋

必要とされる経済・経営がわかるデータ科学人材

データから価値を創出するためには、数学やプログラミングスキル“のみ”では不十分

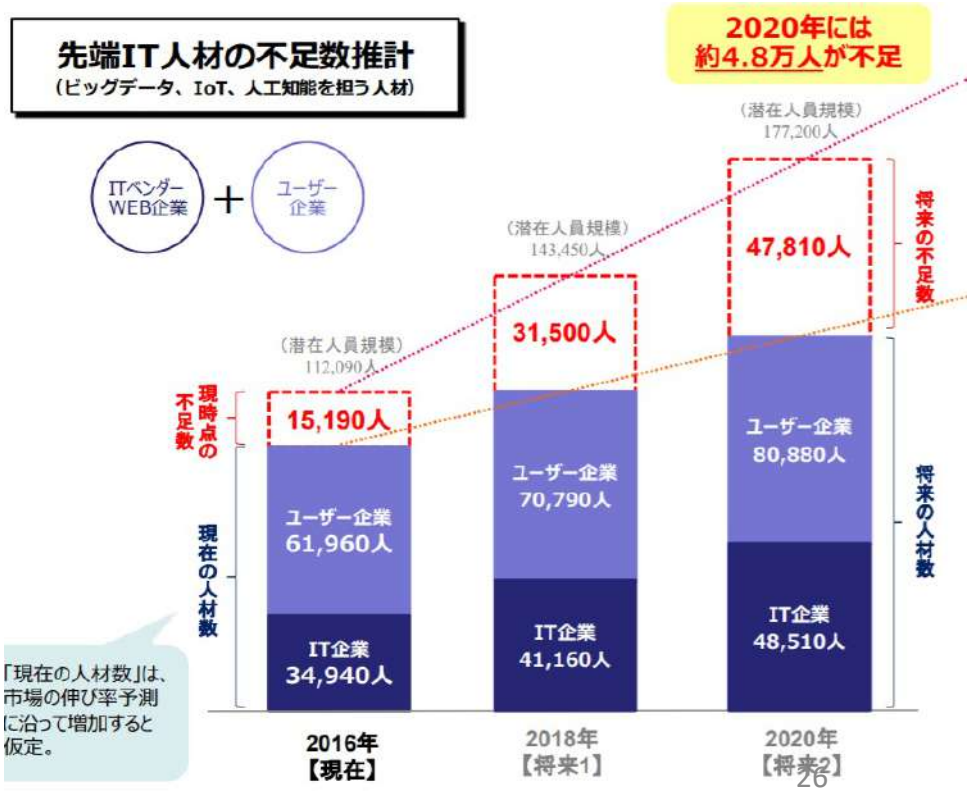
世界的なデータ科学人材の不足

- 特に日本での人材不足は深刻
  - 日本の大学にはデータ科学・統計関連学部がなかった
    - 2017年滋賀大学、18年横浜市立大学でデータサイエンス学部新設

経済・経営＋データ科学の必要性

ビッグデータ関連の人材需要に関しては、ビジネス系の部署、情報系の部署等で生じると考えられ、統計的な高度な専門知識を有するデータサイエンティストのみならず、データ分析をビジネスに活かすビジネス人材（マーケター等）や分析によりビジネス上の価値を発見するアナリスト、ビッグデータのデータ処理等を担うデータエンジニア等職種が活躍すると考えられる。

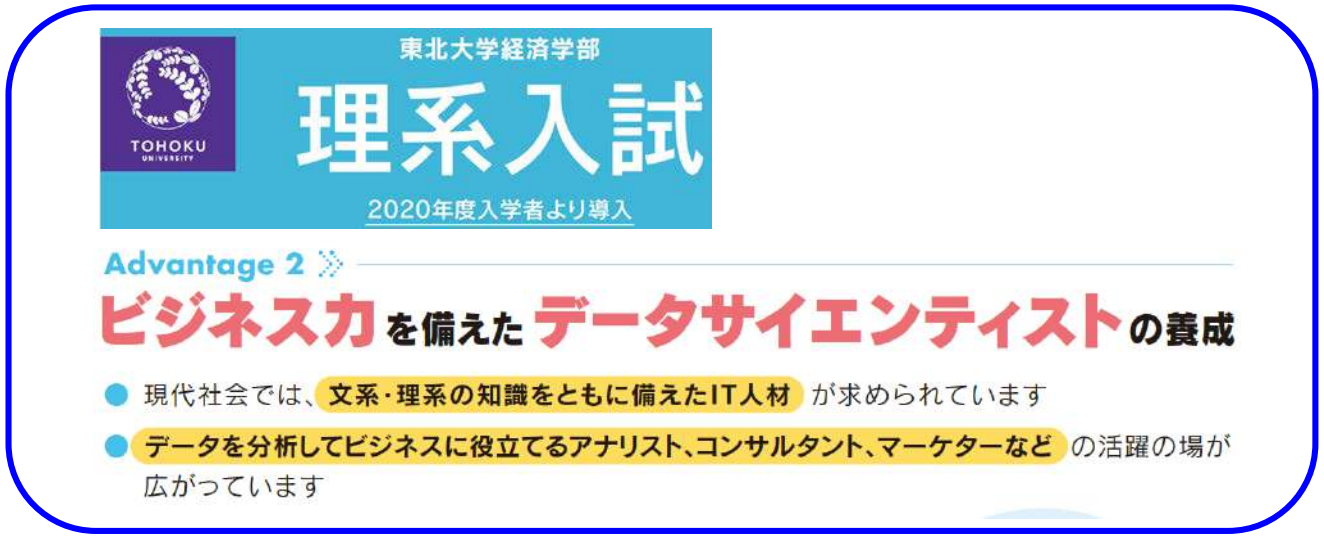
経済産業省、IT人材の最新動向と将来推計に関する調査結果(H28/06)より抜粋



ビッグデータ時代

◎**“データ”**から知識を得て行動する人の数が
“ビッグ”になっていくべき時代

◎**「文理の壁」の打破が必要な時代**



ご清聴ありがとうございました！