Modeling Forum 2014資料

FUJITSU
shaping tomorrow with you

分析・学習とモデリング

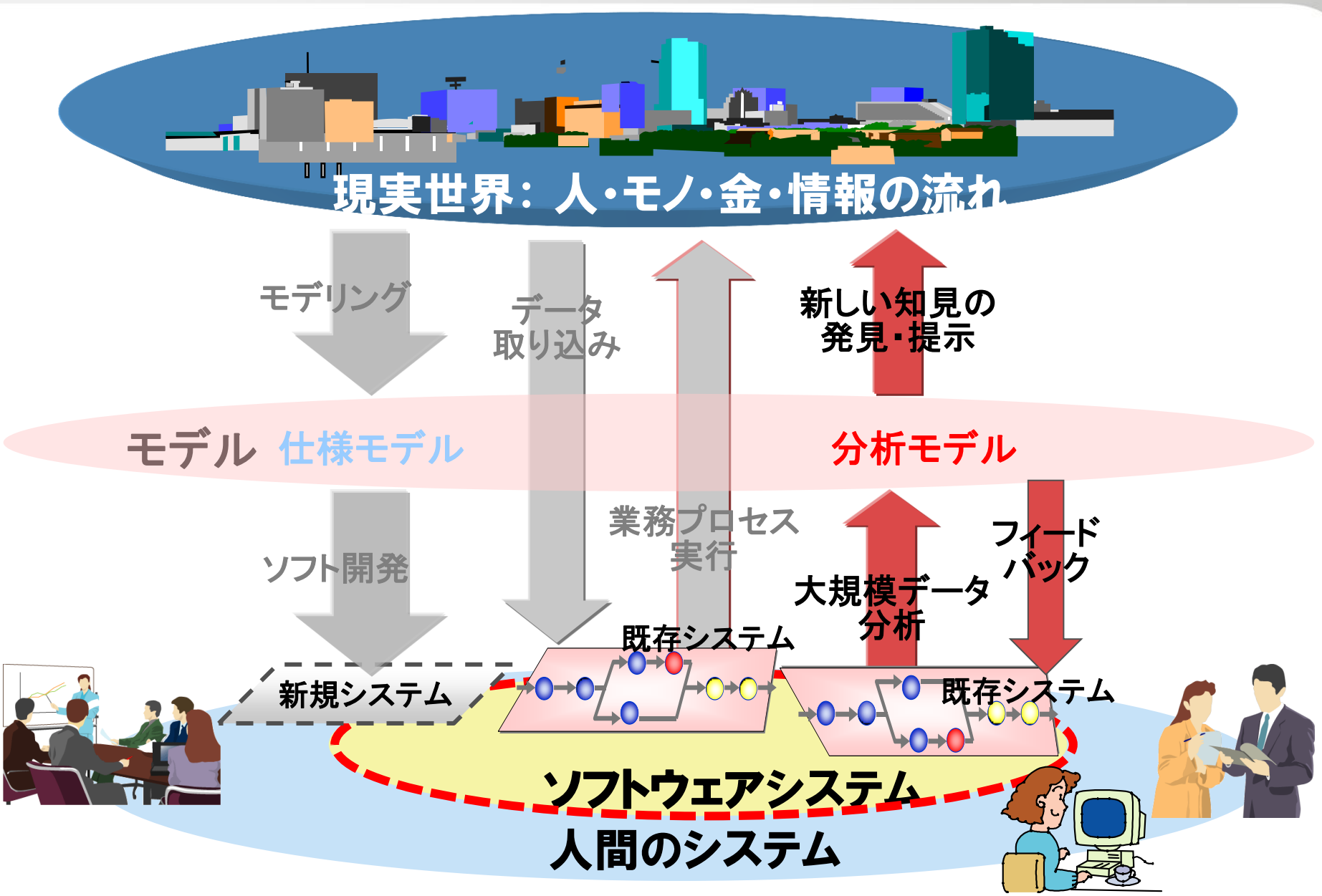
2014年11月19日

富士通研究所R&D戦略本部 特任研究員
人工知能学会 副会長
丸山　文宏

目次

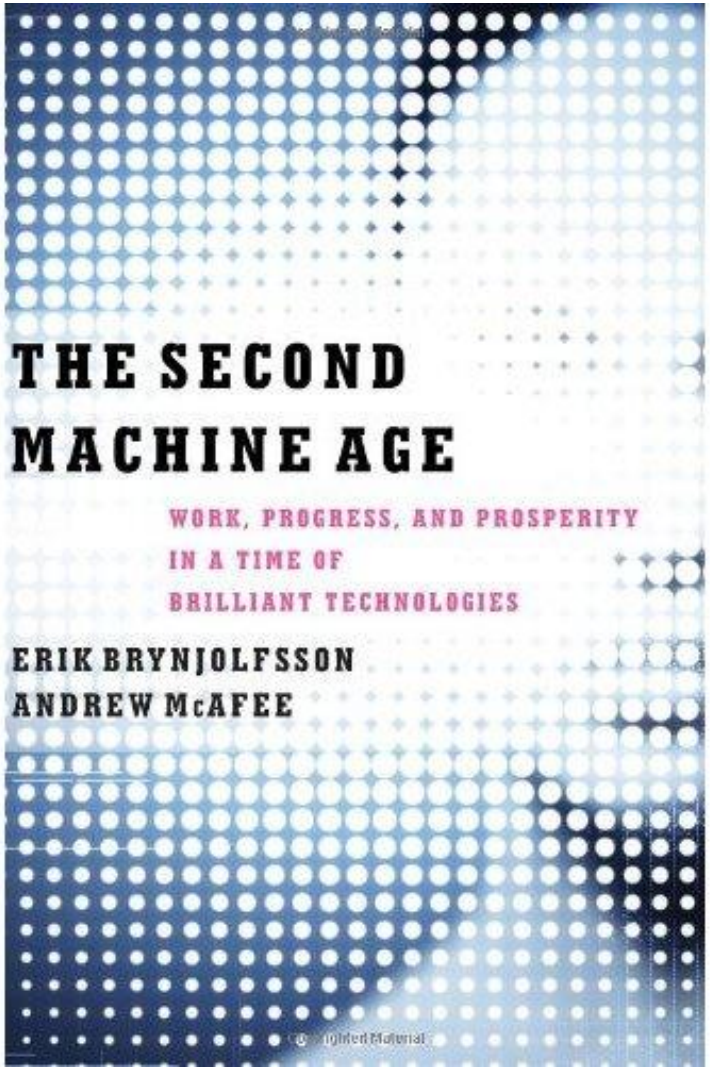
- 分析・学習のモデルとは
- アルゴリズムとモデル
- システムとモデル
- 事例
- まとめ

分析モデル

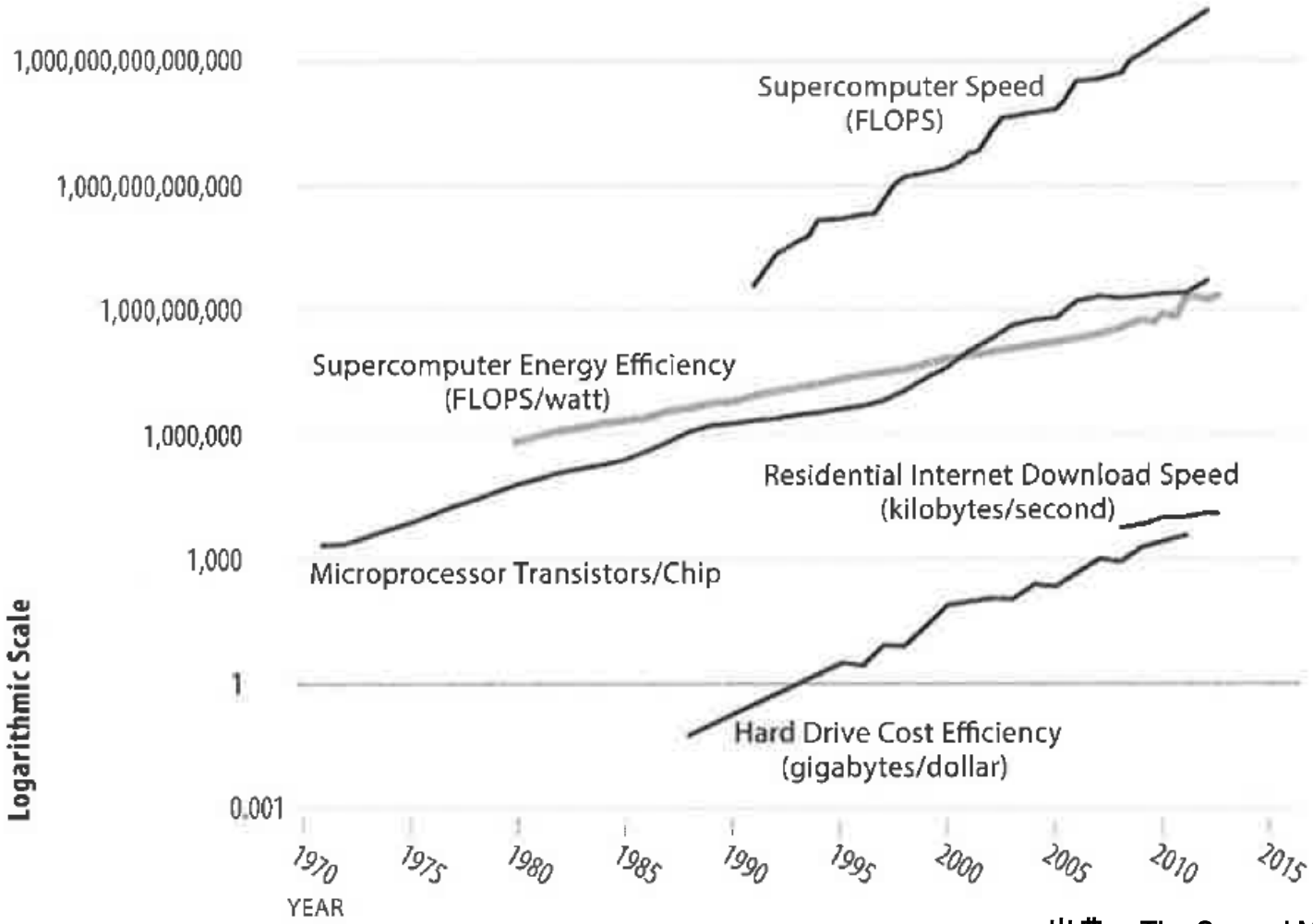


ICTのインパクト

- Exponential（ムーアの法則）
- Digital
  - Non-rival（消費が排他的でない）
  - Marginal cost of reproduction≒0
- Combinatorial
  - アイデアの組合せよるイノベーション
  - Innovation-as-building-block
- Consumerization



ムーアの法則



出典： The Second Machine Age

- 性能／容量が（黙っていても）指数関数的に増大する
- 待っていれば、同じ性能／容量のものが安く手に入る

FUJITSU

アルゴリズムとモデル

相関ルール（Association Rule）

FUJITSU

- 相関ルールとは
  - ある事象Xが起きると別の事象Yが起きる（可能性が高い）というルール

    X（条件部）→Y（帰結部）
- 相関ルールを抽出する問題
  - 支持度（support）＜条件部と帰結部が成立する頻度＞が一定以上
    - レアケースは除外
  - 確信度（confidence）＜条件部が成立する時に帰結部も成立する割合＞が一定以上
- バスケット分析
  - バスケットにアイテム（集合）Xが入っている時、同時に入っている可能性が高い別のアイテム（集合）Yを求める
  - 同時に購入されやすい商品を近くに配置したり、ある商品を購入した人に関連の高い商品を勧める、といった施策が可能

アプリオリ（Apriori）アルゴリズム

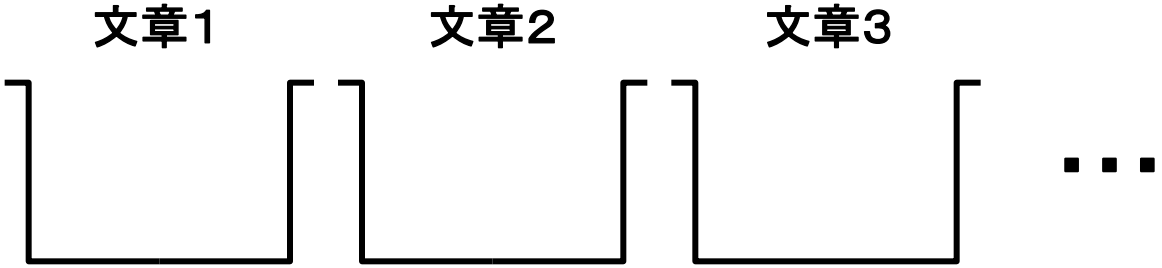
FUJITSU

- 相関ルールの効率的な抽出
  - 2ステップによる抽出
    1. 最小支持度を満たすアイテム集合をすべて含む「頻出アイテム集合」を生成
    2. 頻出アイテム集合から最小確信度を満たす相関ルールを抽出
  - 第1ステップではアイテム総数の指数状の数の候補を検査する必要がある
  - 「アイテム集合の支持度はその部分アイテム集合の支持度を上回らない」という性質を使って計算を省略
- 頻出アイテム集合を生成するアルゴリズムの概略
  - すべての単体アイテムの支持度を計算し、最小支持度を上回るもので頻出アイテム集合を作る
  - kを2から始めて以下の処理を実行し、終了するまでk←k+1として繰り返す
    - k-1個のアイテムから成る頻出アイテム集合の要素にアイテムをひとつ追加した、k個のアイテムから成るアイテム集合の支持度を計算し、最小支持度を上回るものだけを頻出アイテム集合に追加する
    - 最小支持度を上回るk個のアイテムから成るアイテム集合がなくなったら、k個以上のアイテムから成るアイテム集合で最小支持度を上回るものはないため、終了

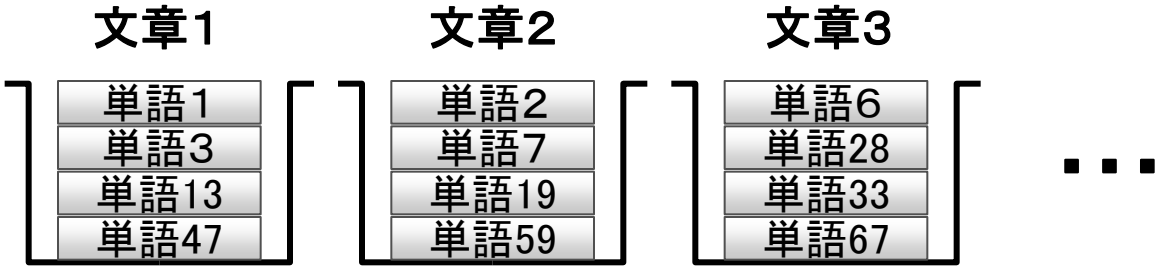
バスケット分析による共起分析

FUJITSU

- 文章をバスケットと見なす



- 文章に現れる単語をそのバスケットに入っているアイテムと見なす

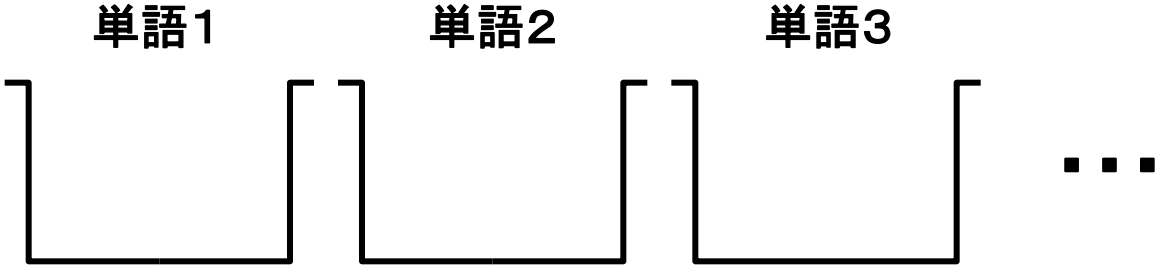


- バスケットにアイテムXが入っている時、同時に入っている可能性が高い別のアイテムYを求める
  - 単語Xが現れる文章に同時に現れる可能性が高い別の単語Yを求める
  - 確信度の高い相関ルールX→Yは単語の共起関係を示す

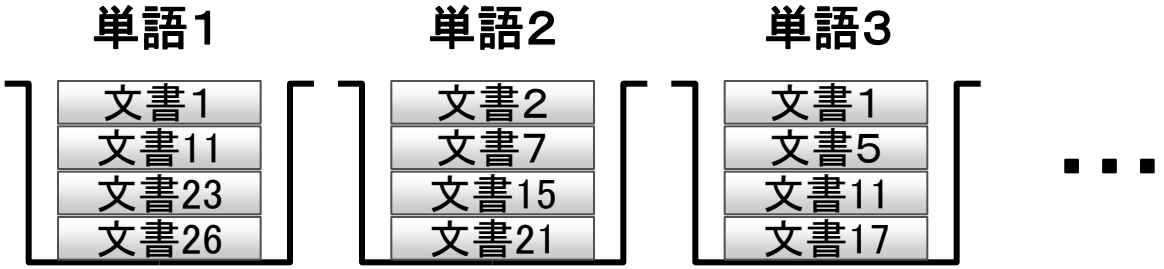
バスケット分析による剽窃検出

FUJITSU

- 文書中に現れる単語をバスケットと見なす



- 単語が現れる文書をそのバスケットに入っているアイテムと見なす

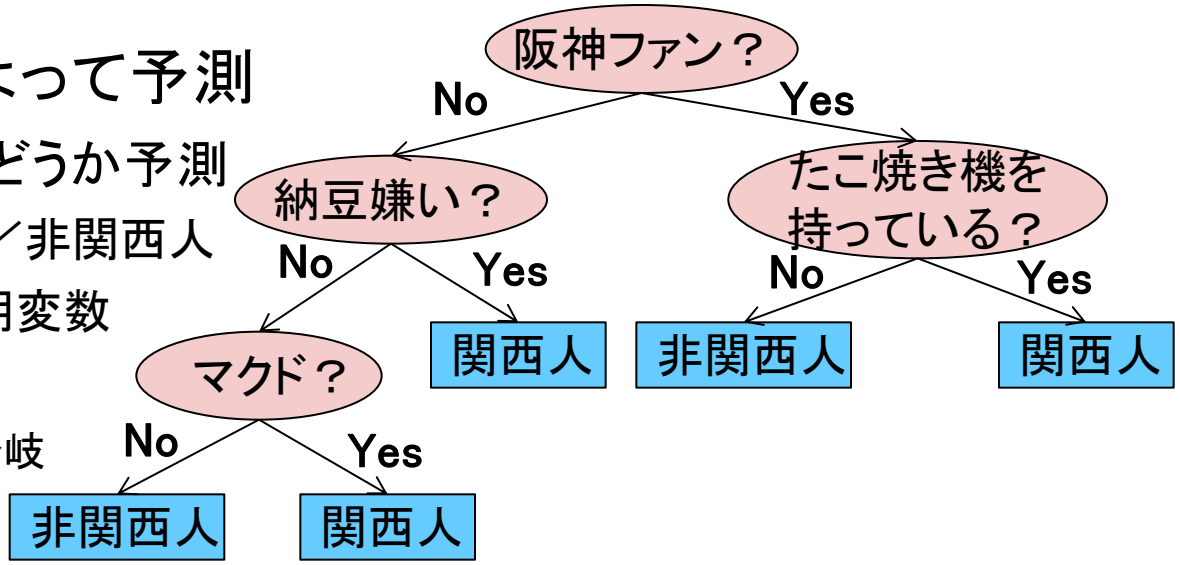


- バスケットにアイテムXが入っている時、同時に入っている可能性が高い別のアイテムYを求める
  - 単語が文書Xに現れる時、その単語が現れる可能性が高い別の文書Yを求める
  - 確信度の高い相関ルールX→Yは剽窃の可能性を示唆する

決定木（Decision Tree）

FUJITSU

- 木構造の分岐条件によって予測
  - 右の決定木は関西人かどうか予測
    - 目的変数の値は関西人／非関西人
    - 分岐に使われるのが説明変数
    - 変数は数値型も可能
      - 数値型では値の範囲で分岐
    - 多分岐も可能



- 決定木の学習には訓練例の集合を用いる
  - 分岐することによって得られる情報量が大きくなるように分岐
  - 分岐の条件に従って訓練例の集合も分割し、その先の部分木を学習

最小記述長基準(MDL基準)

FUJITSU

- 学習には過剰適合(overfitting)の問題がある
  - 訓練データに対しては学習されているが、未知データに対しては適合できない、汎化できていない状態
- 最小記述量(MLD: Minimum Description Length)基準

$$\mathrm{MDL} = -\log L + \frac{k \log n}{2} \quad \text{を最小化}$$

L: モデルの最大尤度
k: モデル次数(自由パラメータの数)
n: データ数

  - 実際の観測データをもとに、複数のモデルのうち、
    (モデルと観測データのずれ)＋(モデル自身の複雑さ)
    が最小になるモデルを最適なモデルとする
  - 観測データに過度にフィットした複雑なモデルや、簡単すぎて観測データを説明できないモデルは排除され、比較的単純で観測データをよく説明できるモデルが選択できる

パターン・クラスタリング

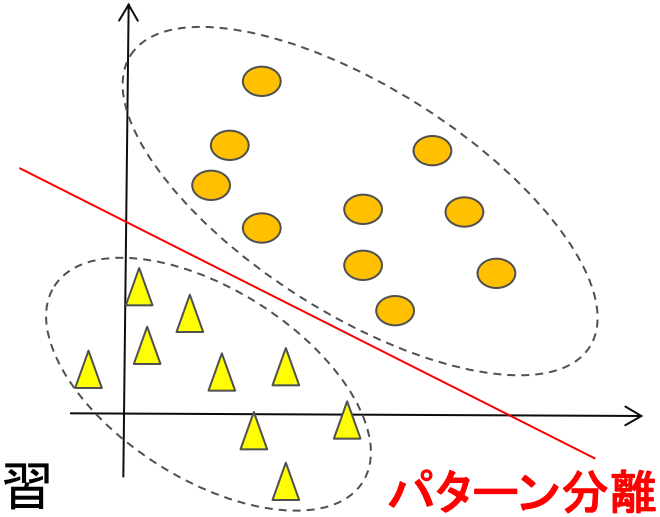
FUJITSU

- パターンを複数のクラスタ（固まり）に分類する
  - 教師なし学習
- ベクトル空間モデル
  - パターンやクラスタの内部表現形式として特徴ベクトルを採用
  - ベクトル間の類似度
    - 余弦（cosine）
    - ハミング距離
      - ビットベクトルの場合

パーセプトロン

FUJITSU

- パターン分類（与えられたパターンをカテゴリに分類）
  - 教師信号がある場合



- 基本パーセプトロンは分離平面を学習
  - 分離平面を表現する結合荷重としきい値のセットがモデル

結合荷重　$w(t) = \{w_0(t)=\theta, w_1(t), w_2(t), \cdots, w_n(t)\}$　（θ：しきい値）

パターンベクトル　$x(t) = \{-1, x_1(t), x_2(t), \cdots, x_n(t)\}$

基本パーセプトロンの出力は

0 if $w(t)\cdot x(t) \leqq 0$

1 if $w(t)\cdot x(t) > 0$

分類が正しく行われなかった場合のみ、以下のように学習

正しい分類が1だった場合　$w(t+1) \leftarrow w(t) + c\,x(t)$　（cは0に近い正の定数）

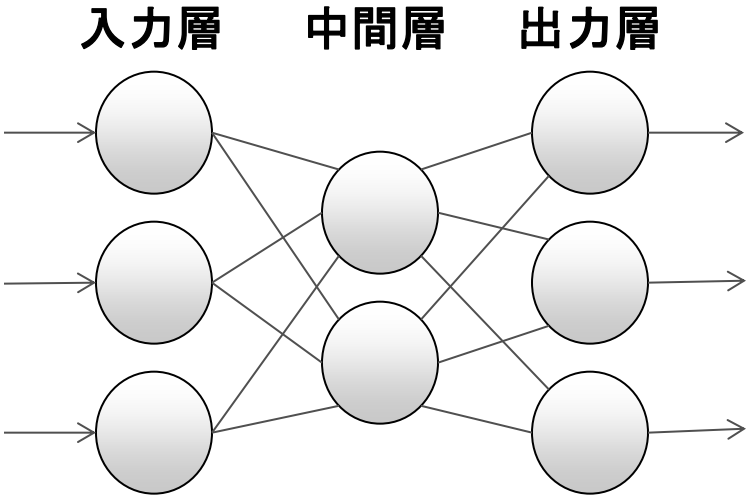
正しい分類が0だった場合　$w(t+1) \leftarrow w(t) - c\,x(t)$　（cは0に近い正の定数）

  - 基本パーセプトロンは有限回の学習で正しく分離できるようになる（収束定理）
    - ただし、線形分離可能な場合に限る

ニューラルネットワーク

FUJITSU

階層型ネットワーク



$y_k = f(\Sigma w_{ki} x_i)$

y_k: 出力
x_i: 前段からの入力
w_{ki}: 結合荷重

$f(x) = \dfrac{1}{1 + e^{-x}}$ （シグモイド関数）

- 結合荷重のセットがモデル

バックプロパゲーション（誤差逆伝播法）による学習

- 誤差が小さくなるように、最急降下法（結合荷重で偏微分）で結合荷重を修正

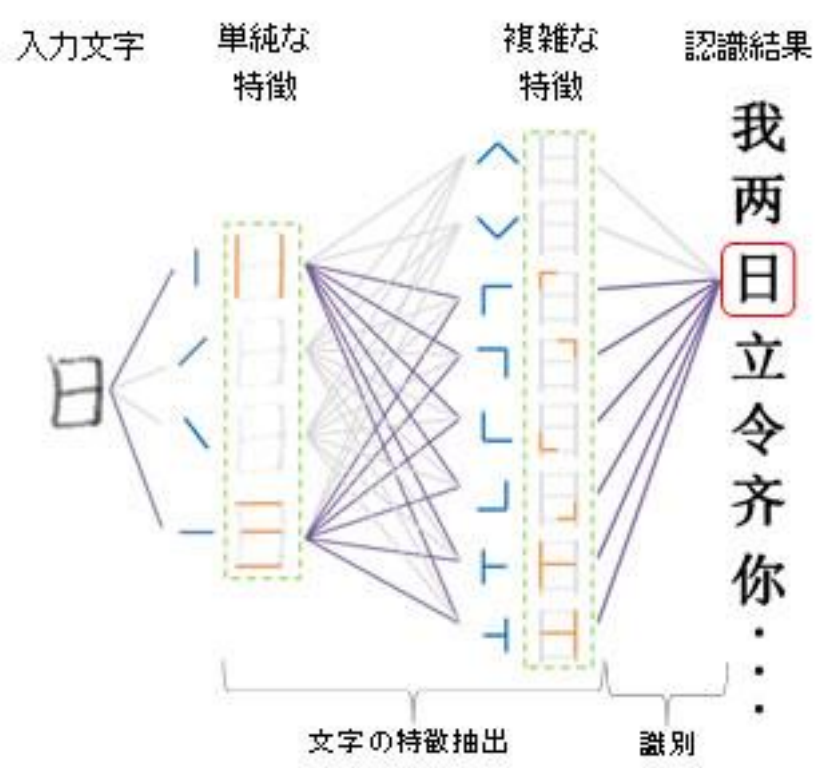
誤差 $2E = \Sigma (t_i - O_i)^2$

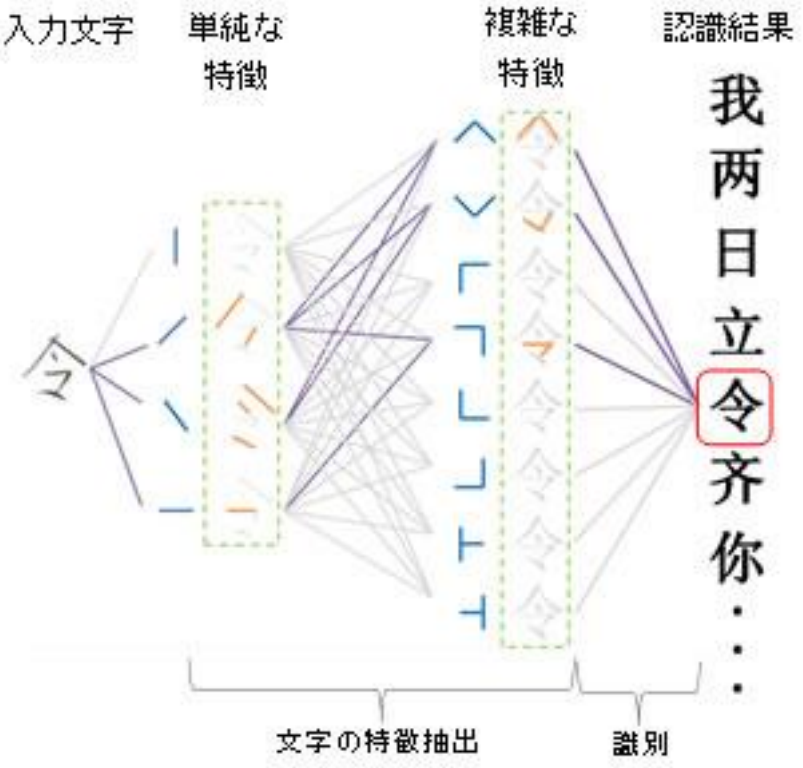
t_i: 教師信号
O_i: 出力層の出力

機械学習のブレイクスルー：Deep Learning

FUJITSU

- Deep Learningをベースに手書き文字認識の高精度化を実現
- ハードウェア活用（GPU）により学習時間を短縮（17倍）





- **ICDAR 2013（International Conference on Document Analysis and Recognition）の手書き文字（中国語）認識コンテストで1位**

http://pr.fujitsu.com/jp/news/2013/08/21.html

FUJITSU

システムとモデル

第3回将棋電王戦

FUJITSU

■ **現役プロ棋士対コンピュータ将棋**
- **5対5の団体戦**
- **2014年3月15日～4月12日**
- **DENSOロボットアーム「電王手くん」採用**

コンピュータ将棋の４勝1敗

プロ棋士	勝敗		コンピュータ
菅井竜也五段	×	○	習 甦
佐藤紳哉六段	×	○	やねうら王
豊島将之七段	○	×	YSS
森下卓九段	×	○	ツツカナ
屋敷伸之九段	×	○	ponanza

コンピュータ将棋の棋力 ＝

ビッグデータ（棋譜）	×	アルゴリズム（機械学習、探索）	×	計算機パワー（クラスタ）
6万局以上		1億パラメータ以上		2億手/秒

機械学習（アルゴリズム）の威力

FUJITSU

これまでの将棋ソフト

・プログラミングができて
将棋の強い人

開発者の固定観念、先入観、
主観などにより、パラメータ設定

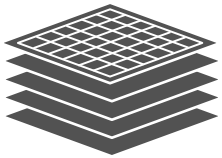
約500パラメータ

アマチュア
有段者レベル

機械学習を用いた将棋ソフト（2005年 *Bonanza～）

・プロの棋譜

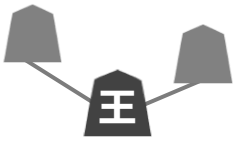
6万局



・各駒の価値

例

・駒と駒の位置関係



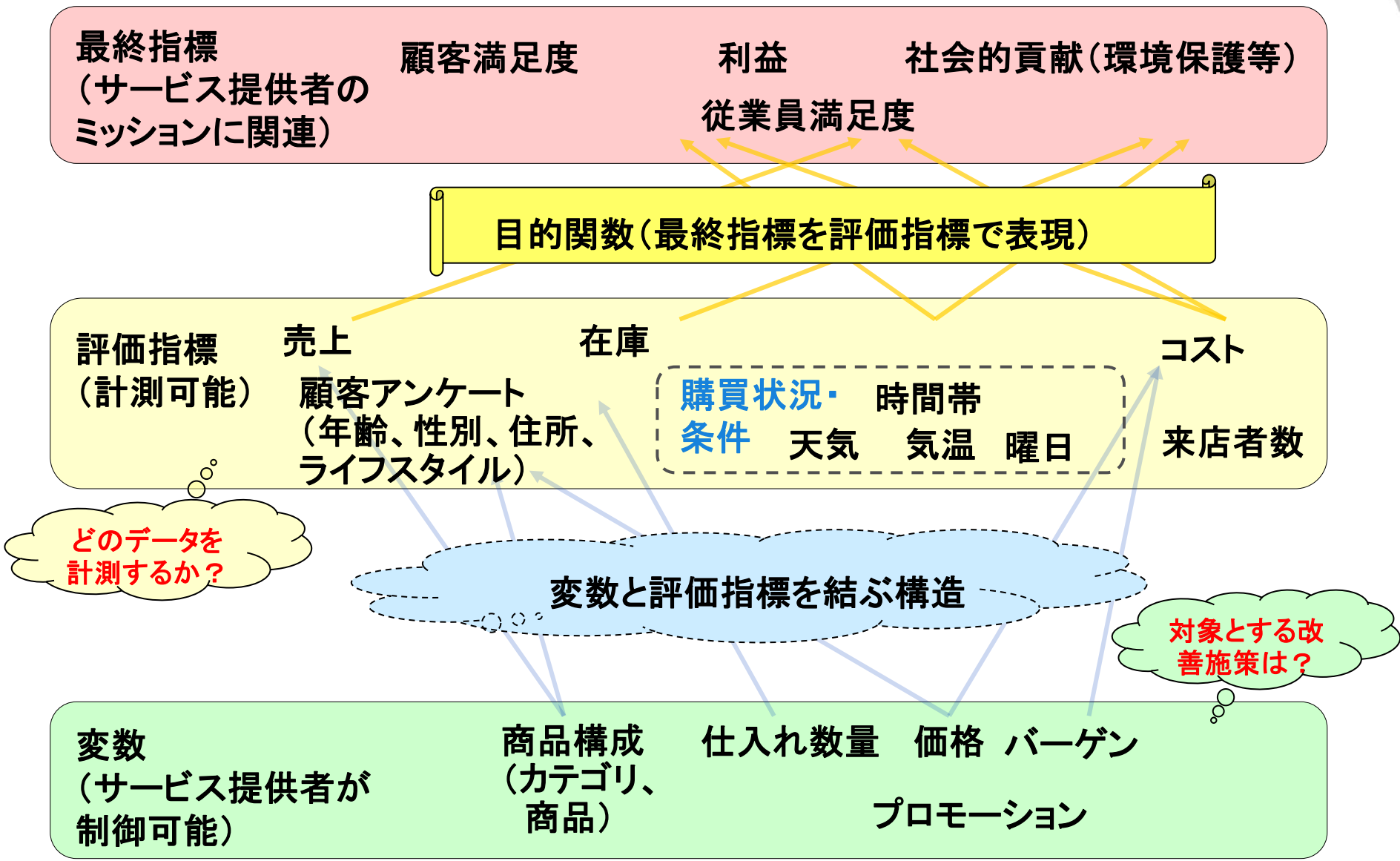
「局面評価関数」の
最適なパラメータを自動学習

約1億パラメータ

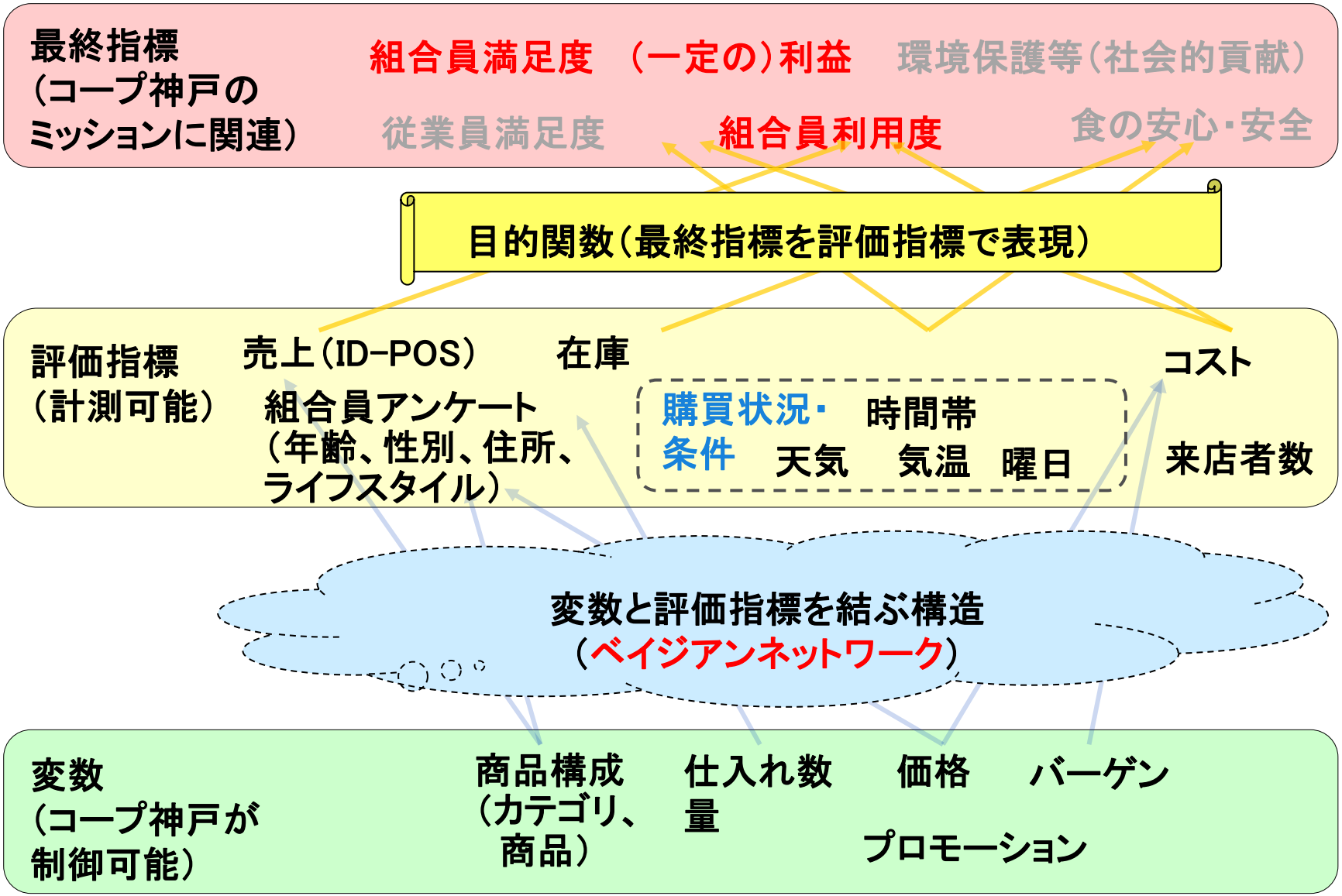
プロレベルの
棋力を実現

※「Bonanza」は、保木邦仁先生（現　電気通信大学特任助教）が開発したコンピュータ将棋ソフトです

最適化モデル（小売業の例）

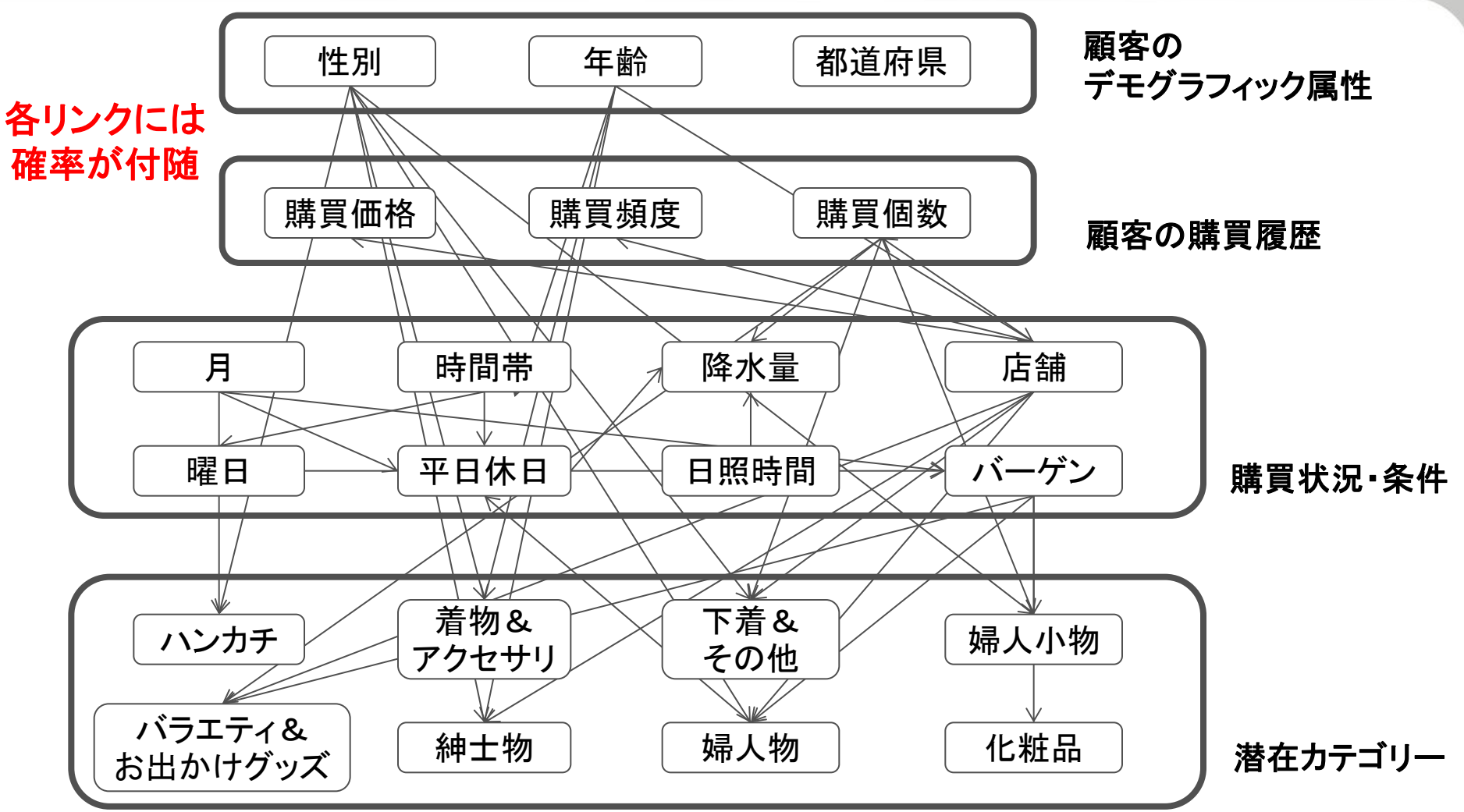


生活協同組合コープ神戸（産総研）



顧客行動モデル

FUJITSU

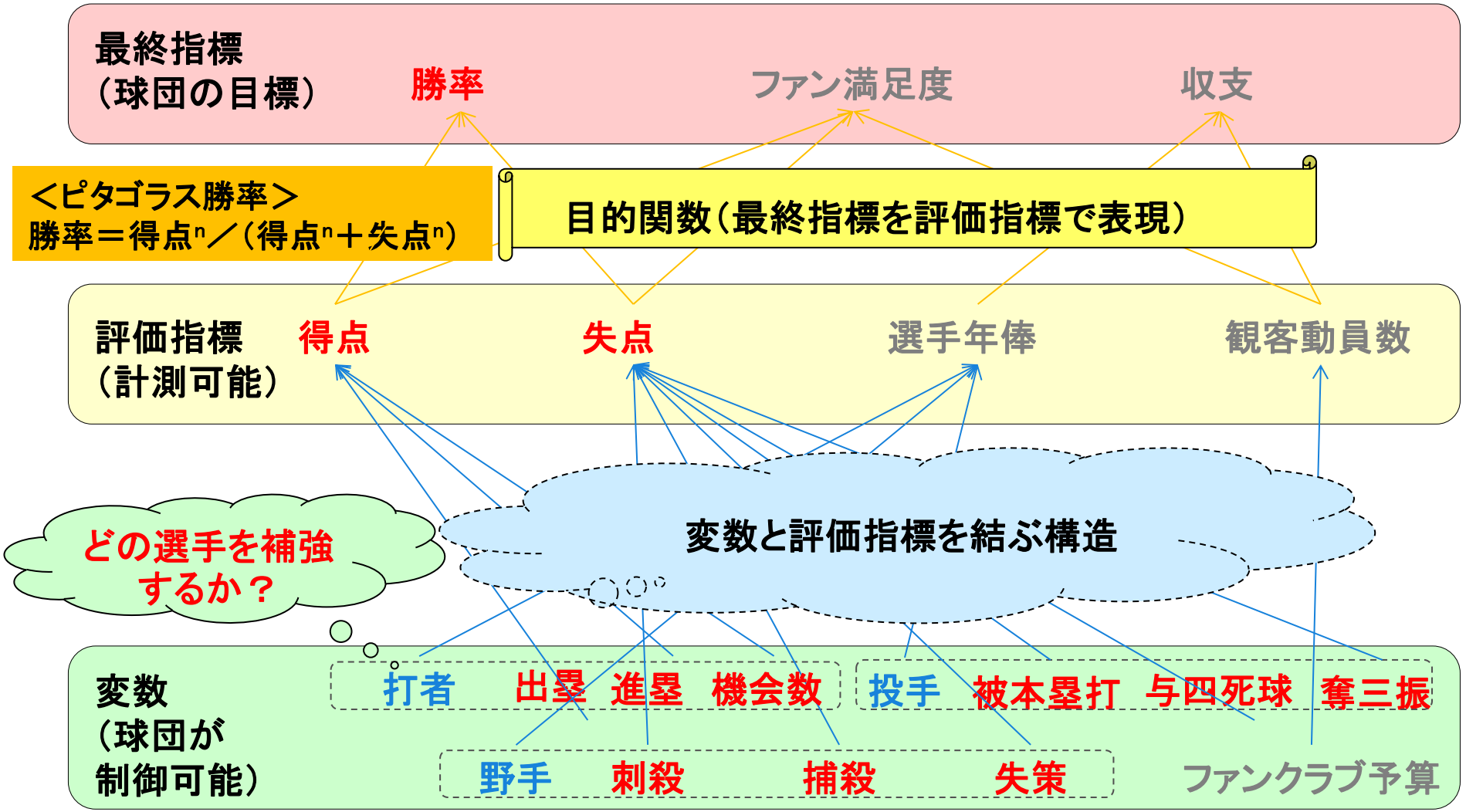


・大量のデータを用いた学習によりベイジアンネットワークを精製
・出来上がったベイジアンネットワークのうえで確率推論を実行

野球のセイバーメトリクス

FUJITSU

■ チーム編成（マクロ）の観点



野球のセイバーメトリクス

FUJITSU

- 戦術（ミクロ）の観点
- 各イニングのBase/Out状態ごとの勝利確率が算出可能
- 以下の表はホームチームから見た1回表の各状態の勝利確率

塁の状態			アウト数	ホームチームから見た点差				
1B	2B	3B	Out	-4	-3	-2	-1	Tie
-	-	-	0	0.182	0.246	0.322	0.409	0.500
-	-	-	1	0.194	0.261	0.341	0.430	0.524
-	-	-	2	0.202	0.272	0.354	0.445	0.540
1B	-	-	0	0.165	0.224	0.296	0.377	0.466
1B	-	-	1	0.181	0.245	0.322	0.408	0.500
1B	-	-	2	0.196	0.264	0.345	0.434	0.529
-	2B	-	0	0.152	0.208	0.276	0.356	0.444
-	2B	-	1	0.173	0.235	0.309	0.394	0.485
-	2B	-	2	0.191	0.257	0.337	0.425	0.519
-	-	3B	0	0.139	0.191	0.256	0.333	0.420
-	-	3B	1	0.158	0.216	0.286	0.369	0.459
-	-	3B	2	0.188	0.254	0.333	0.421	0.515
1B	2B	-	0	0.139	0.191	0.255	0.329	0.413
1B	2B	-	1	0.164	0.223	0.294	0.375	0.464
1B	2B	-	2	0.186	0.251	0.329	0.417	0.509
1B	-	3B	0	0.125	0.172	0.232	0.305	0.386
1B	-	3B	1	0.150	0.205	0.272	0.351	0.439
1B	-	3B	2	0.182	0.246	0.323	0.409	0.502
-	2B	3B	0	0.116	0.162	0.219	0.288	0.368
-	2B	3B	1	0.141	0.194	0.259	0.335	0.421
-	2B	3B	2	0.179	0.242	0.318	0.403	0.494
1B	2B	3B	0	0.108	0.149	0.203	0.268	0.343
1B	2B	3B	1	0.136	0.186	0.249	0.323	0.405
1B	2B	3B	2	0.173	0.234	0.307	0.390	0.480

FUJITSU

事例

Watsonの概要

FUJITSU

- IBMのQA Systemで、Jeopardy!という番組でクイズ王に勝利
- Jeopardy!の問題への解答のために必要な情報をWeb上で収集
  - 主にFactoidというタイプのQAが対象
    - 名詞や動詞といった単語レベルで答えられるものが主な対象
  - 過去問の解答の約95%がWikipedia/Wiktionaryのタイトルだったらしく、Wikipedia/Wiktionaryを中心に知識を生成・拡張した模様
- 収集した情報を知識化し、質問に答えられるようにする
  - 言語処理の要素技術（品詞タグ付け、固有表現抽出、構文解析、照応解析など）を用いて、テキストを解析
  - 解析結果を基に蓄積された知識から、質問に対する解答を検索
    - ベースラインは、質問文を解析して、Wikipediaを検索し、そのタイトルを答えの候補とする方法
  - 解答の9割以上はテキストから
  - 構造データからは精度は高い解答が得られるがカバレッジが低い

質問文解析から解答まで

FUJITSU

1 質問文解析

Who is the leading actor of *Ocean's Eleven*?

関係：leading-actor(?, Ocean's Eleven)
解答タイプ：leading actor
固有名詞：*Ocean's Eleven*：*映画*

2 解答候補取得

構造データ検索用クエリ
leading-actor(?, Ocean's Eleven)
テキスト検索クエリ：
"leading actor" & "*Ocean's Eleven*"

3 スコアリング

George Clooney
映画DBから発見=1
質問文との類似度=0.9
解答タイプとの意味の近さ=0.8

Brad Pitt
映画DBから発見=0
質問文との類似度=0.5
解答タイプとの意味の近さ=0.8

Legends of the Fall
映画DBから発見=0
質問文との類似度=0.2
解答タイプとの意味の近さ=0.1

手掛かりを生成しスコアリング。
スコアで解答選択

George Clooney
3.5 = 0.2 * 1+1* 0.9 + 3 * 0.8

Brad Pitt
2.9 = 0.2 * 0+1* 0.5 + 3 * 0.8

Legends of the Fall
0.5 = 0.2 * 0+1* 0.2 + 3 * 0.1

A 構造データ

テキストはWebや新聞など。
解答候補は結果中の単語

映画DB

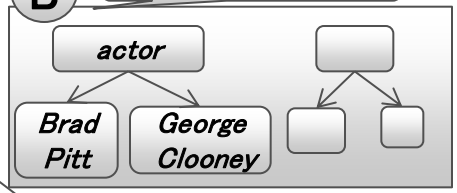
関係	人物	映画
leading-actor	George Clooney	Ocean's Eleven
:		

非構造データ

George Clooney was the leading actor of Ocean's Eleven.

Brad Pitt was the leading actor of Legends of the Fall.
He had a role in Ocean's Eleven.

B 語彙知識など



C 機械学習で重要度を学習

手掛かり	重要度
映画DBから発見	0.2
質問文との類似度	1
解答タイプとの意味の近さ	3
:	

Watsonの金融への応用

FUJITSU

- Citibankでは2012年からWatsonの評価を行っている
  - リスク・マネジメント
  - コールセンターで顧客のために最適な金融商品を選択
- 金融商品選択への応用
  - Watsonに学習させているということだが、まだロールアウトしていない
  - それぞれの商品に付随する複雑な条件を考慮する必要があり、クイズのQAほど単純ではない
    - 数理最適化の機能を組み込むことも必要になるかもしれない
    - ロールアウトされるとしても、Watson本来の機能がメインになるか疑問

Watsonの医療への応用

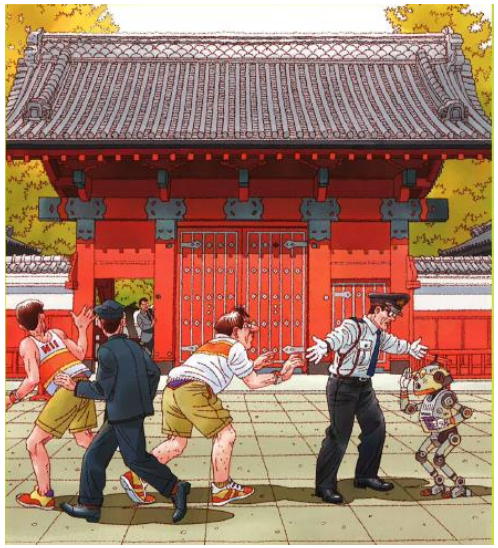
FUJITSU

- IBMではWatsonの医療への応用も進めている
  - 症状から病名を診断することはQAと同様の枠組みでできると考えられるため
- しかし、医療への応用はクイズのQAほど単純ではない
  - 解答に対する明確な根拠を示さないと医師が採用できない
    - 1970年代に開発された血液伝染病の診断を行うエキスパートシステムMycinには、結論の根拠として適用されたルールの系列を提示する機能があった
    - Watsonの根拠（類似度の重み付き和）は薄弱
    - 他の選択肢を否定するための消去法の手段があれば、医師は安心できる
  - 診断に使われる情報がクイズの質問文より複雑
    - バイタル・データ、症状（自訴、検査結果、兆候）、持病／既往症、薬、生活習慣
    - 情報の構造が重要
      - 時間的な前後関係が重要
      - 潜伏期間などの定量情報が重要
  - 検討すべき選択肢に漏れがないよう支援することには意味があるが、選択肢を絞り込む機能がなければGoogle検索（無料）と変わらない
- ただ、最新の医学文献の情報を知識として取り込む機能はこれまでのシステムにないもの
  - 知識を抽出するための学習手法が重要
  - 知識の表現形式も重要

「ロボットは東大に入れるか」

FUJITSU

- プロジェクトチーム（リーダー：新井紀子国立情報学研究所教授）
  - 国立情報学研究所（全般）
  - JSTさきがけ（社会）
  - 富士通研究所（数学）
  - 名古屋大学（国語）
  - 代々木ゼミナール（模試）



- Yes/Noを判定する問題が重要と主張
  - 歴史科目の正誤判定問題など
  - Factoid型質問応答システムには難しい
  - Yes/Noを判定する手法は選択肢を消去する手段として有効
- 含意関係認識を重要なタスクとして特定
  - 「川端康成は、『雪国』などの作品でノーベル文学賞を受賞した」は「川端康成は、『雪国』の作者である」を含意
  - 任意の主張が蓄積された知識から含意されるか否かを判定する手法は、Yes/Noを判定するためのブレークスルーになりうる
    - Google検索からWatsonまでが依拠してきた類似度アプローチからの新展開

認知システム（Cognitive System）の応用動向

FUJITSU

- 反復型のタスクで応用が進んでいる
  - 何回も反復して起きる（それほど難しくない）タスクの処理を自動化
    - 具体的な内容は毎回同じではないため、ルーチン・タスクではない
    - 企業として投資しやすい状況
  - 銀行のカスタマーサービスに適用
    - 顧客からの質問文書（テキストによる問合せ。例えば、残高照会）に回答
  - 交通違反の法的処理で弁護士を支援するシステム
  - 保険のクレーム処理を自動化するシステム
  - 与えられたデータをベースに新聞記事を作成するシステム
    - Yahoo Sportsでは試合の結果についての記事を自動作成
- 発見型のタスクへの応用はこれから

位置情報活用基盤：SPATIOWL（スペーシオウル）

FUJITSU

■ 日々変化する道路状況をリアルタイムに把握し、住民へ情報提供

移動・位置情報収集

工事・規制情報登録

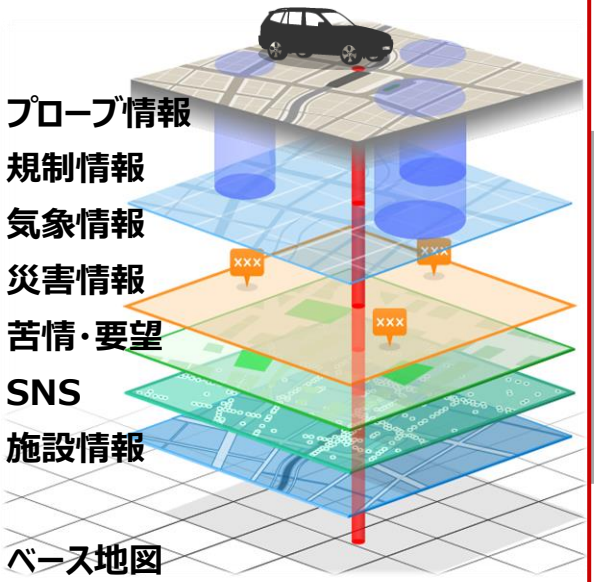
・自治体職員

スマートフォンアプリで現地から道路状況登録

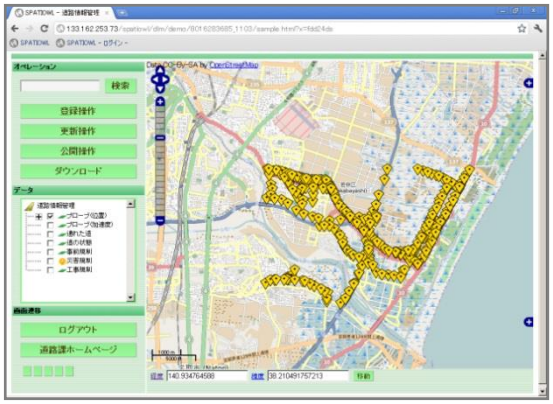
・自治体職員
・住民
・ボランティア

要望・クレーム

道路情報管理システム



SPATIOWL基盤
時空間データベース

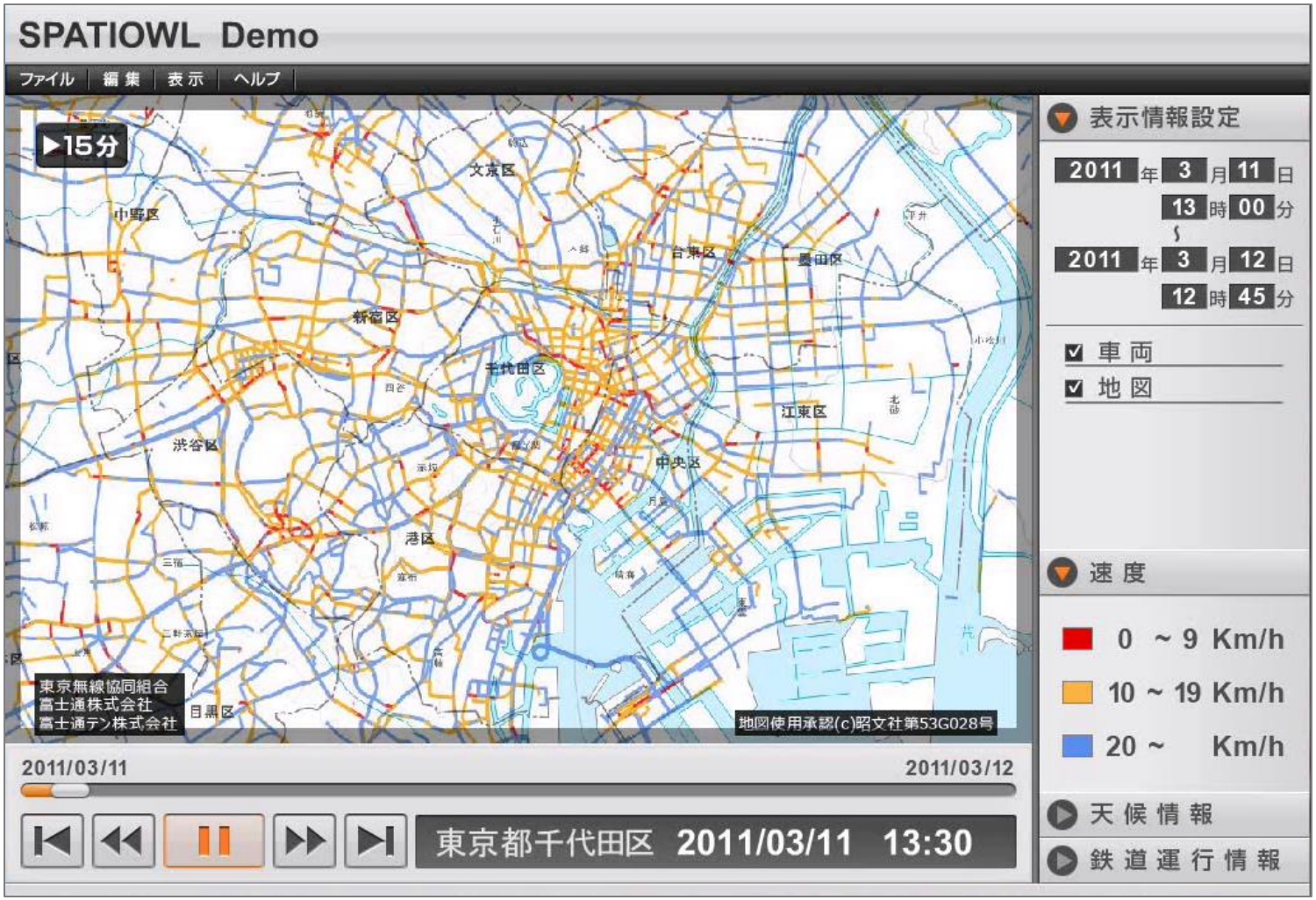


通行可能/不可能道路やその範囲（エリア）を、地図上で確認。

- ■住民アナウンス
- ■ボランティア活動支援
- ■企業の運行計画
- ■都市計画、復興計画

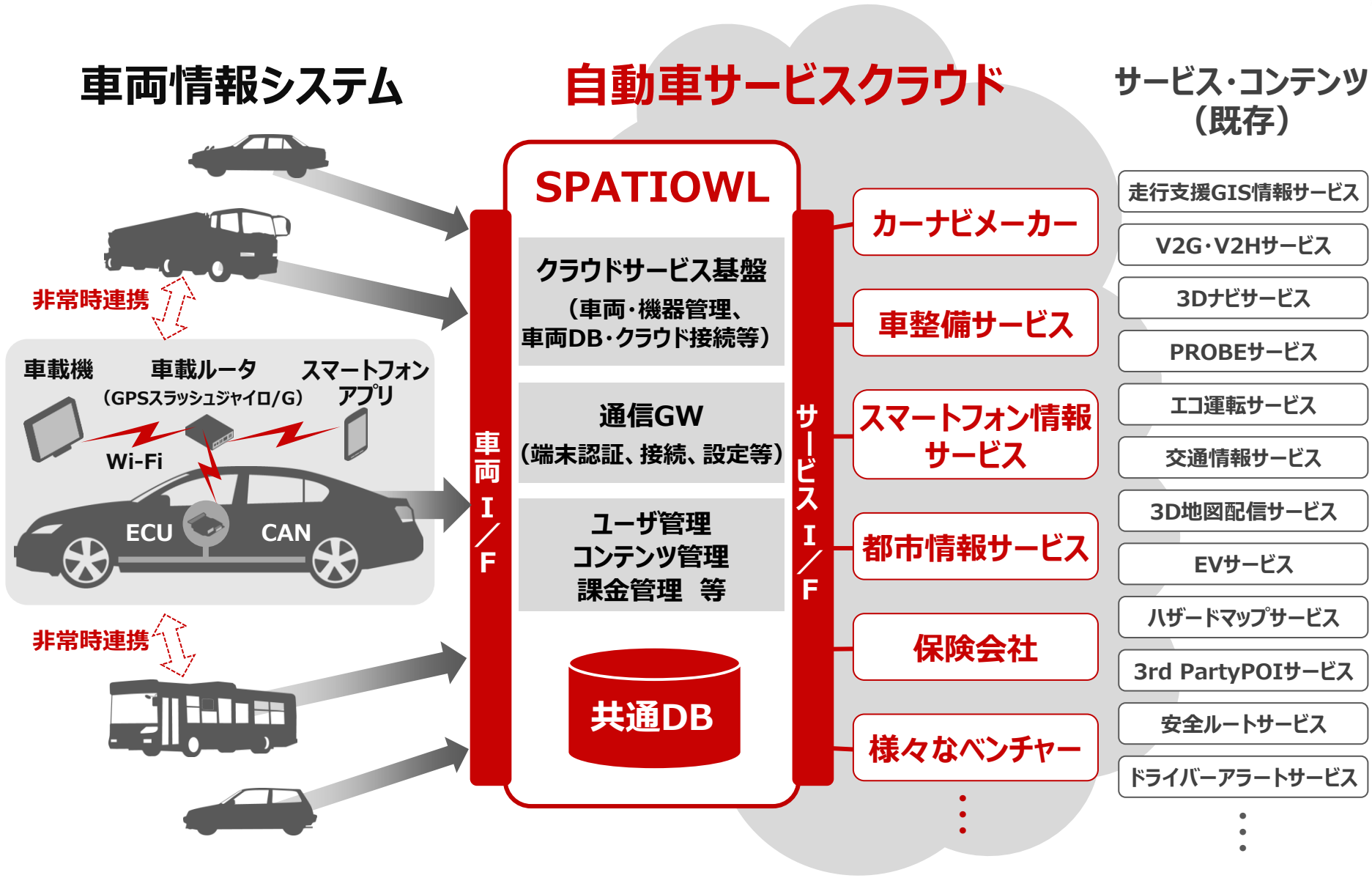
震災時の東京の交通状況（蓄積データ）

震災直後30分～1時間で交通マヒになり、翌朝4時頃になるまで解消していないことがわかる。



自動車サービスクラウド

FUJITSU



予測モデルの構築

大規模データ処理　（並列・分散処理）

膨大なパラメータを最適化し、最高精度の予測モデルを作る

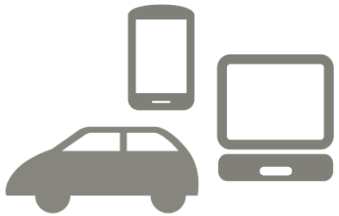
キュレーター

FUJITSU

- **2011年1月、BI/BA、コンサルタント、分析アルゴリズム研究者等を集約した組織を設立。データキュレーションサービスを2012年4月から提供開始**

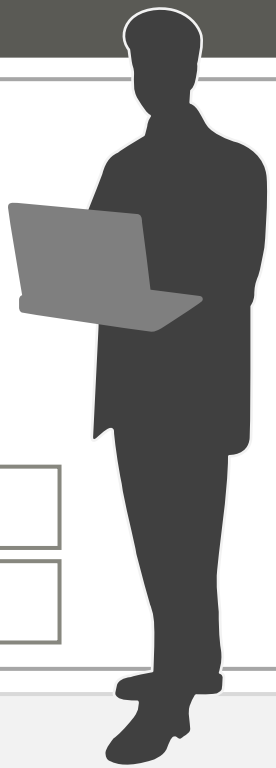
データに語らせる

キュレーターのやること



・時系列データ
・イベント連鎖
（A→B→C）

・方程式
・予測モデル
・評価関数
・KPI 等



相関関係/因果関係発見	パラメータ最適化	アルゴリズム選択
イベントパターン発見	予測シミュレーション	ダイナミック最適化

キュレーターの専門スキル

モデリング
数学、統計学、自然科学 等

×

アナリティクス
多変量解析、機械学習、最適化 等

×

システムデザイン
並列分散処理、CEP 等

新しいデータ活用の担い手

FUJITSU

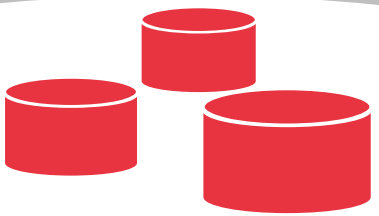
キュレーター

ビッグデータ分析の専門家

2011年1月：組織化
2012年4月：サービス商品化

意思決定者

業務や領域の専門家



活用

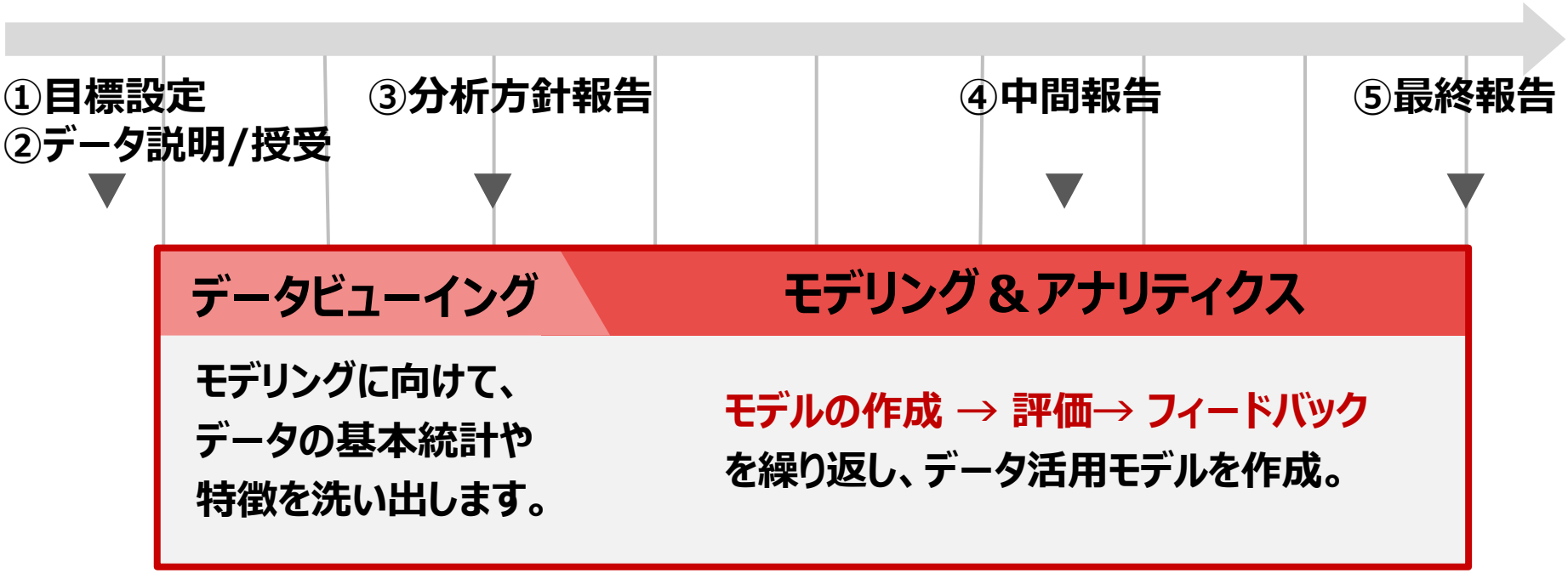
業務/商品/社会

データによる価値創造のサイクル

データキュレーションサービス

FUJITSU

- データオリエンテッドな分析アプローチによりデータ活用モデルを作成/検証します
- 業務刷新や新商品開発に向けたPoCに適したサービスです
- データをお預かりして、最新のキュレーション環境で分析します（最短8週間）



2012年4月より提供中

データキュレーションの適用事例（抜粋）

FUJITSU

テーマ	データ活用モデル
新ビジネス開発	・疾病リスク予測 ・運転新評価指標作成
会員/顧客管理	・会員の休眠/退会予測 ・コールセンターの入電数予測 ・ロイヤルカスタマーの特徴抽出 ・マーケティングの新指標作成
商品の売上予測	・商品の売上/欠品予測 ・店舗属性別の売上予測
製造・生産プロセス/品質管理	・製造・生産品質分析による品質指標作成 ・歩留まりの改善
営業活動評価	・売上予測 ・売上の構成要因の分析 ・営業施策の効果分析 ・自動発注/欠品予測 ・リアル/バーチャル最適化 ・Web、広告、営業活動の評価指標作成

FUJITSU Big Data Initiative

「FUJITSU Big Data Initiative」としてサービス・プロダクトを体系化
お客様の課題に対応したオファリング・実装モデルをメニュー化

オファリング

業務プロセス改革
- リアルタイム経営
- エネルギーマネジメント
- 設備メンテナンス高度化
- 頻発停止発生予測
- 社会インフラ維持・管理

マーケティング高度化
- CX（顧客体験価値）
- パーソナライズ化
- 人的リソース最適配置

商品・サービス強化
- センサーM2M

サプライチェーン最適化
- 需要予測

アプリケーション

業種・業務・情報系アプリケーション

ビッグデータアナリティクス

アナリティクス
- データキュレーションサービス
- ビッグデータ人材育成サービス
- Interstage 、ISV (SAP, SAS, Oracle)

データ利用（コンバージェンス）
- 外部データ連携「DataPlaza」
- 位置情報活用基盤サービス「SPATIOWL」
- オープンデータ検索連携「Linked Open Data」

ビッグデータエンジン

インメモリ | カラムストア | 並列分散処理 | 複合イベント処理 | マスターデータ管理

ビッグデータプラットフォーム

スマートデバイス | サーバ、ストレージ | PCクラスタ | クラウドサービス

センシングネットワーク

売上と連動する因子を解明する

FUJITSU

お客様の課題 要望	・商品の売上に影響する要因の仮説が成り立たなくなっている ・各商品の売上がどんな要因と連動するのか知りたい

■ 因子（約1300種類）

お客様のデータ

・商品群
・商品価格

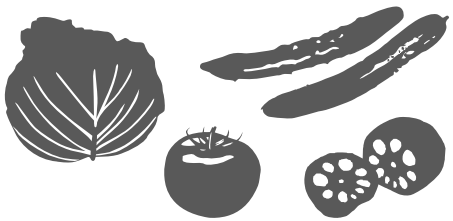
＋

食品に関するデータ（例）

・種類　・価格　・流通量

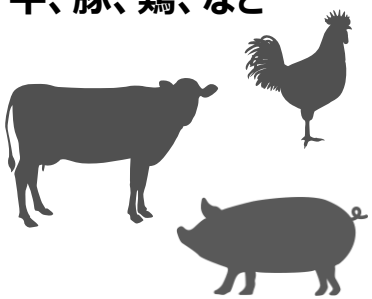
野菜/果物

トマト、キャベツ、きゅうり、
れんこん、じゃがいも、など

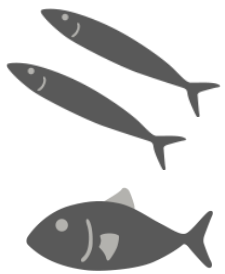


精肉

牛、豚、鶏、など



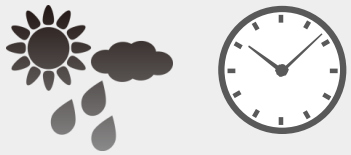
鮮魚



…

その他のデータ

・気象　・時間帯　・地域



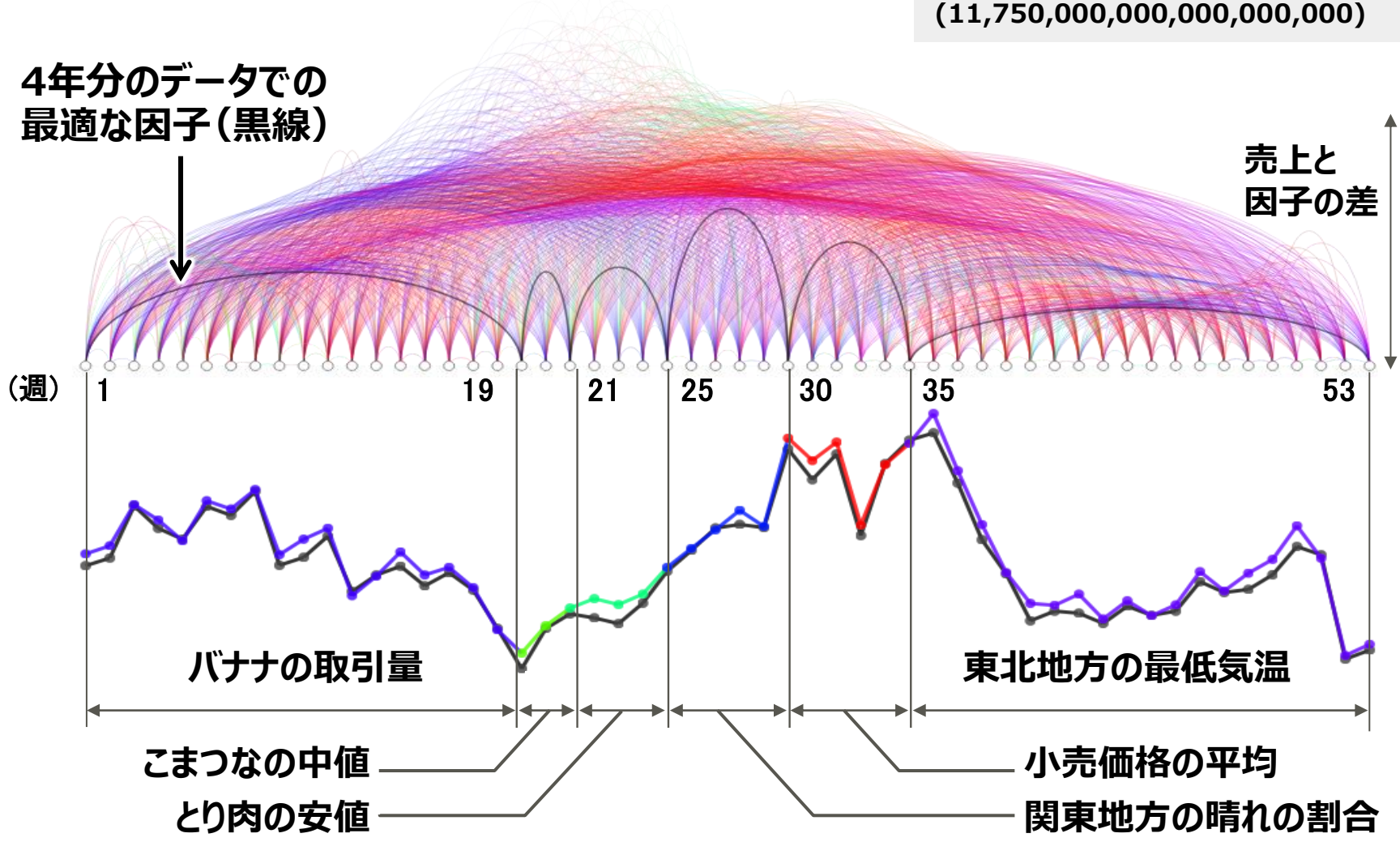
…

売上と連動する因子

FUJITSU

■ 各週の売上と因子の連動性のモデルを構築
→連動性の最も高い因子を求める

因子　1304種類
因子の列の総数 1175京通り
(11,750,000,000,000,000,000)



顧客管理: 優良会員化分析

FUJITSU

ユーザー要件

コンテンツビジネス拡大のために優良会員を増やしたい

【現状】

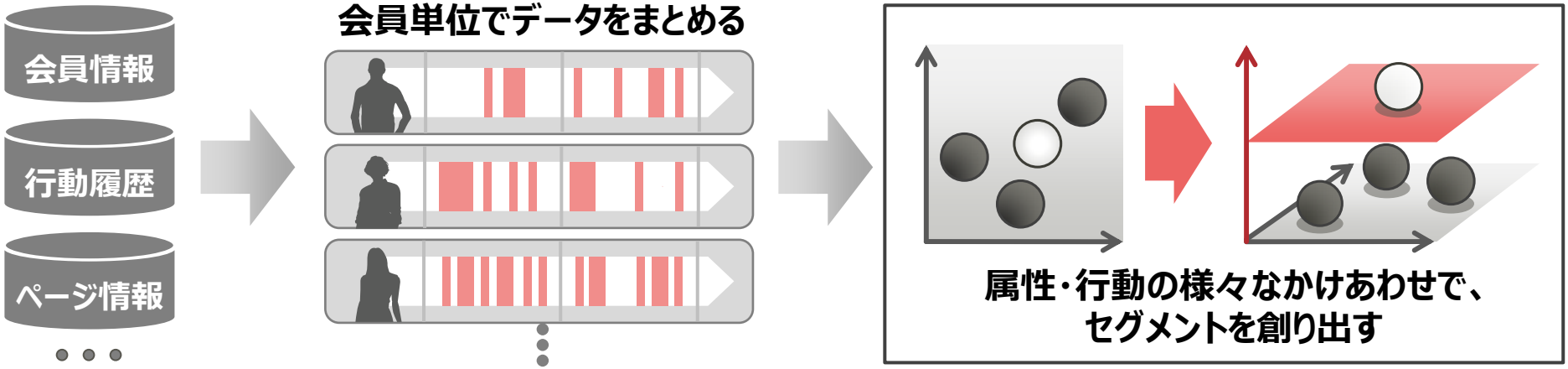
・大量のアクセスログから集計レベルでの分析しかできていない

・優良会員化する行動特徴を経験的に判断し、施策に反映
（人によって考える行動特徴はバラバラ、共有できていない）

→ **優良会員増加のための施策検討のヒントを得たい**

キュレーションのアプローチ

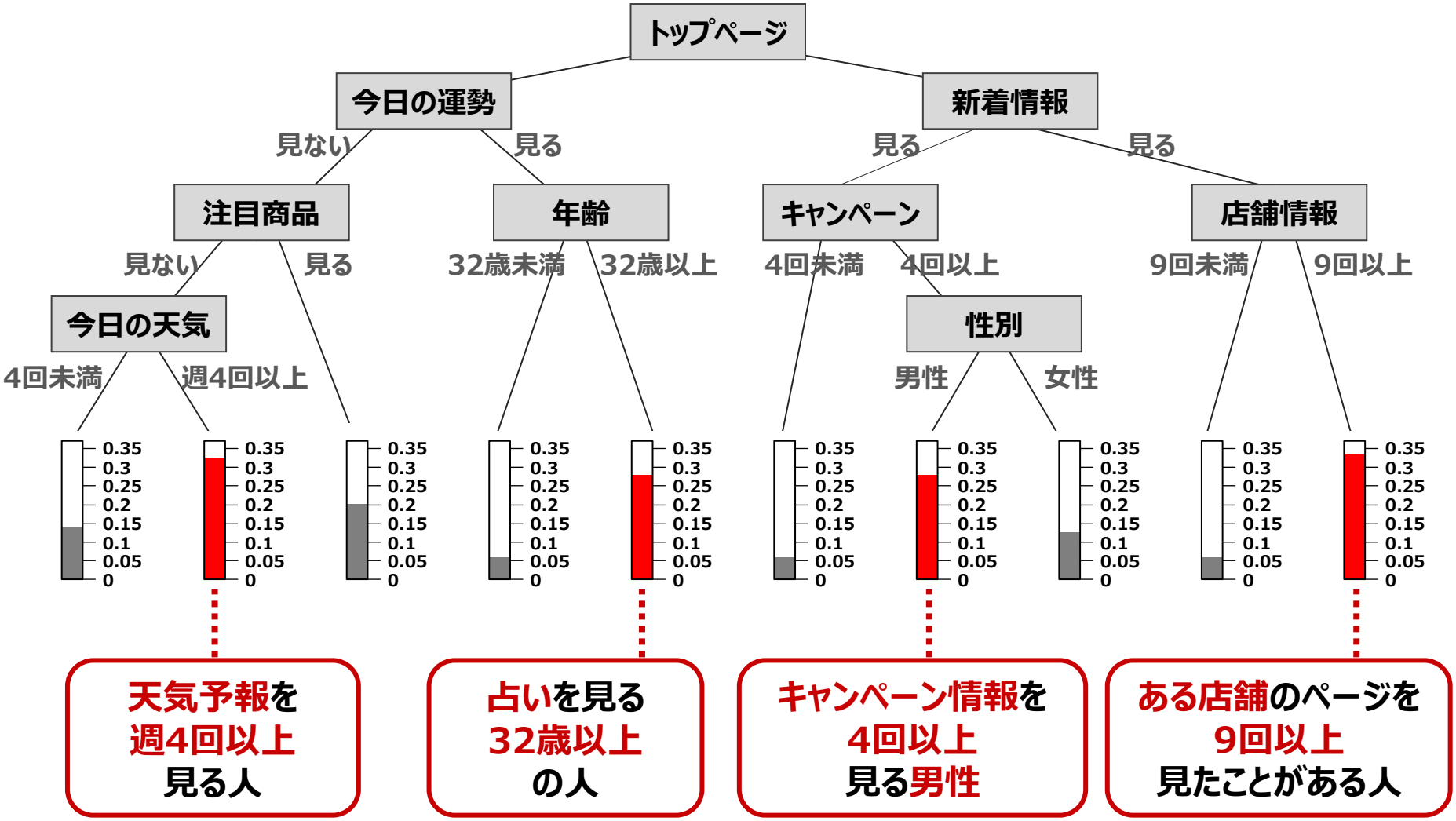
会員属性・行動履歴から新たなセグメントを作り上げる



顧客管理：優良会員化分析

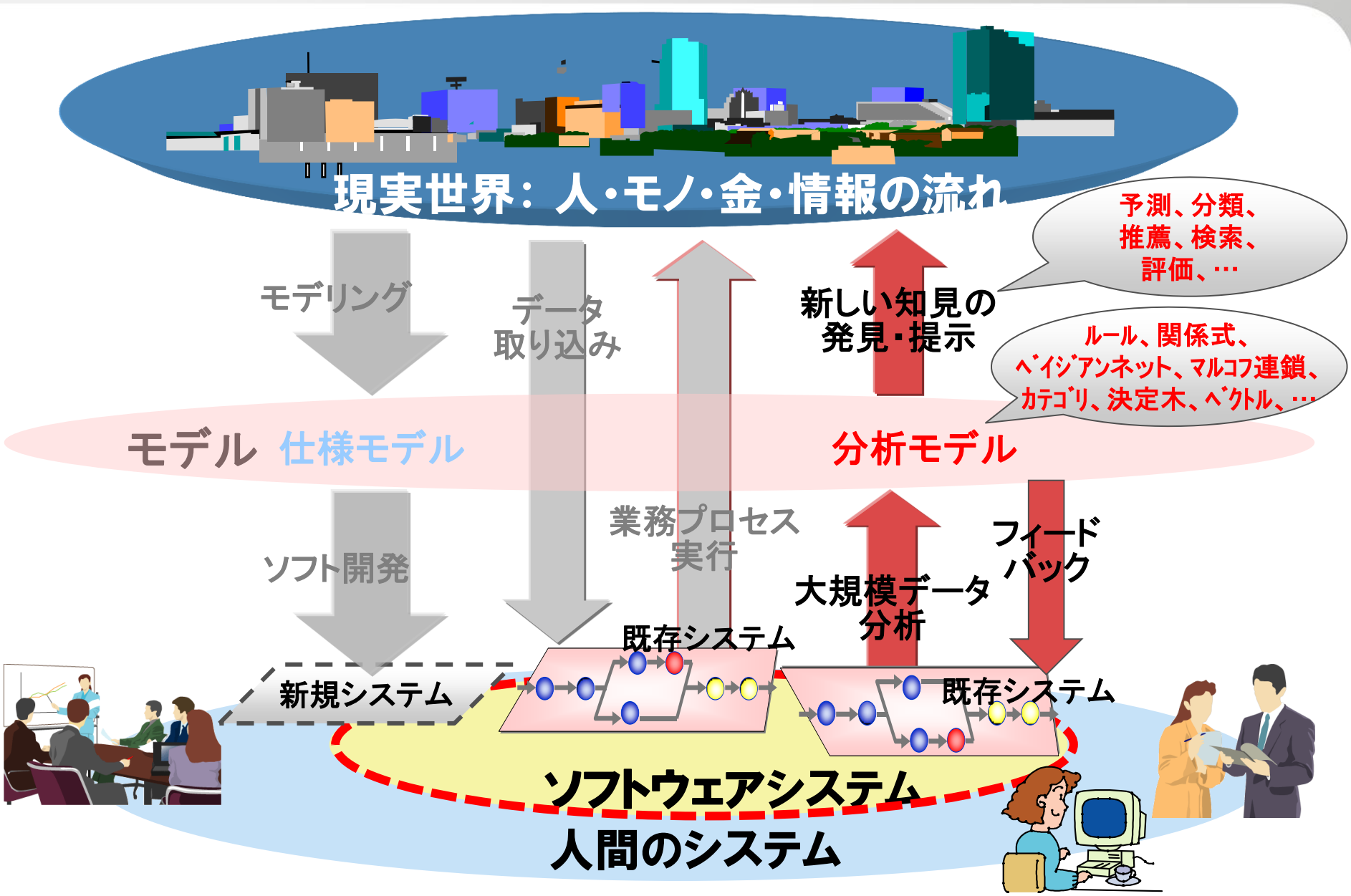
FUJITSU

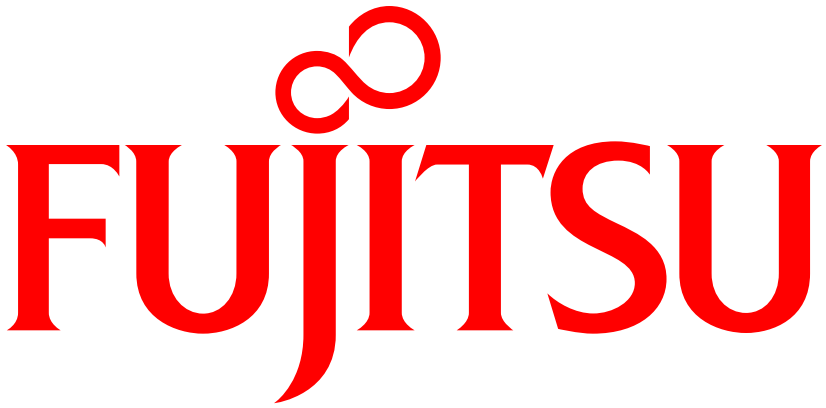
例）優良会員化する会員セグメント



業務に価値ある「会員属性-行動セグメント」を網羅的に作り出す

まとめ





shaping tomorrow with you