



E-commerce企業における ビッグデータへの挑戦と課題

-機械学習への期待について-

A horizontal banner with a red diagonal stripe. On the left, a man in a suit. In the center, a woman wearing a colorful hijab. On the right, a man in a suit and a woman with long dark hair. The text 'Empowering Society Through the Internet.' is written in white across the top. The text 'Rakuten aims to become the No. 1 Internet service company in the world.' is written in white on the red stripe.

Empowering Society Through the Internet.

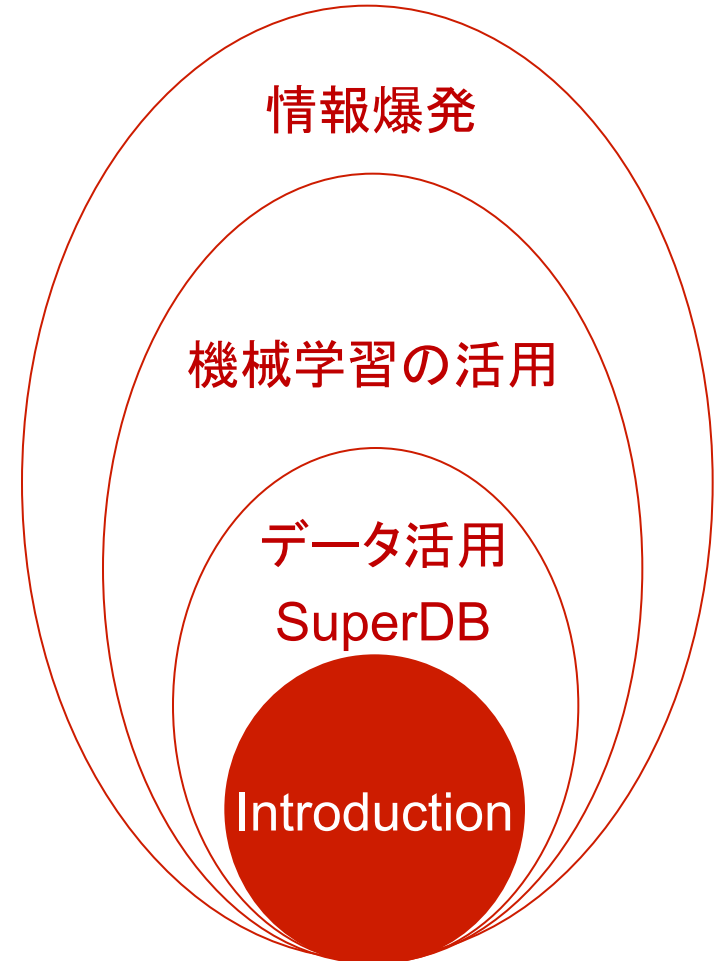
Rakuten aims to become
the **No. 1**
Internet service
company in the world.

Rakuten Inc. RIT. Masaya Mori Nov. 7th, 2012

NII
国立情報学研究所

Introduction

BigDataの時代へ



- ・ **森 正弥（もり まさや）**
- ・ 楽天株式会社 執行役員
- ・ 開発アーキテクチャ部 部長
- ・ ビッグデータ部 副部長
- ・ 楽天技術研究所 所長
- ・ 職掌
 - － 開発部署のマネジメント
 - － 研究開発の推進・統括



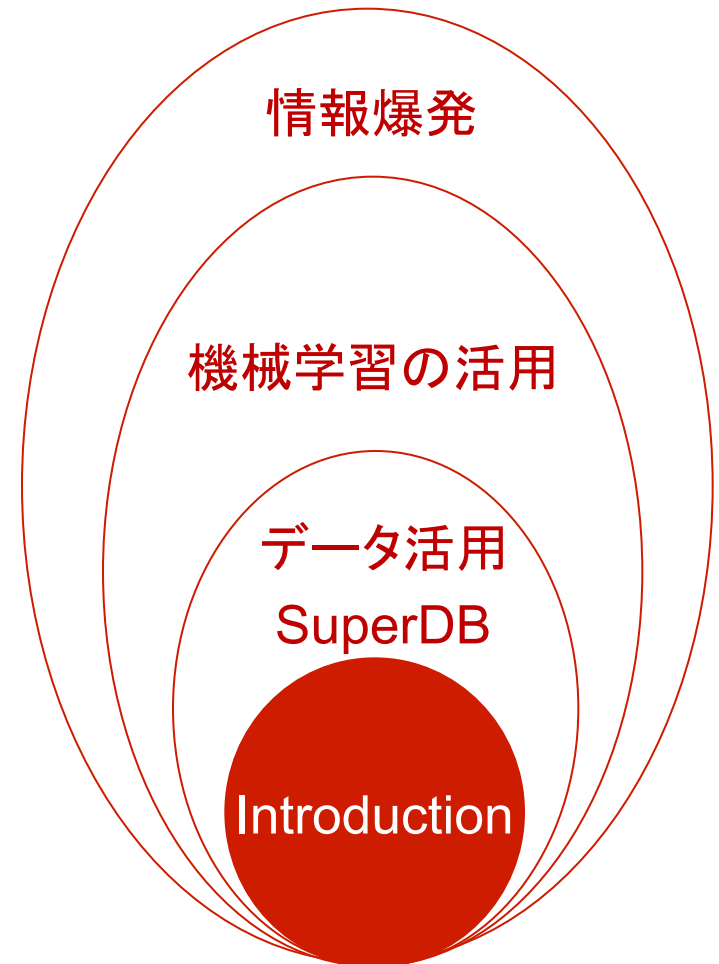
Masaya Mori

Twitter: @emasha



Rakuten Group

BigDataの時代へ



- 代表取締役会長兼社長 三木谷 浩史
- 従業員数 単体3,209人、グループ7,615人
- 設立日 1997年2月17日
- IPO 2000年4月19日(ジャスダック)
- 資本金 1,079億円(2011年12月末現在)
- 連結売上高 3,799億円(2011年度)
- 連結営業利益 756億円(2011年度)



楽天市場(eコマース事業)を中核とした,
総合インターネットサービス企業

海外拠点 & 楽天技術研究所

- ・ 13か国にてサービスを展開
 - － マレーシアにて、EC事業に参入
- ・ 楽天技術研究所は、2拠点 (Tokyo, New York)



Free Cause(USA)



Linkshare(USA)



Tradoria(Germany)

技術の理論面を担うR&D組織

コンセプト

Next Reality

- 来るべき豊かなリアリティを -

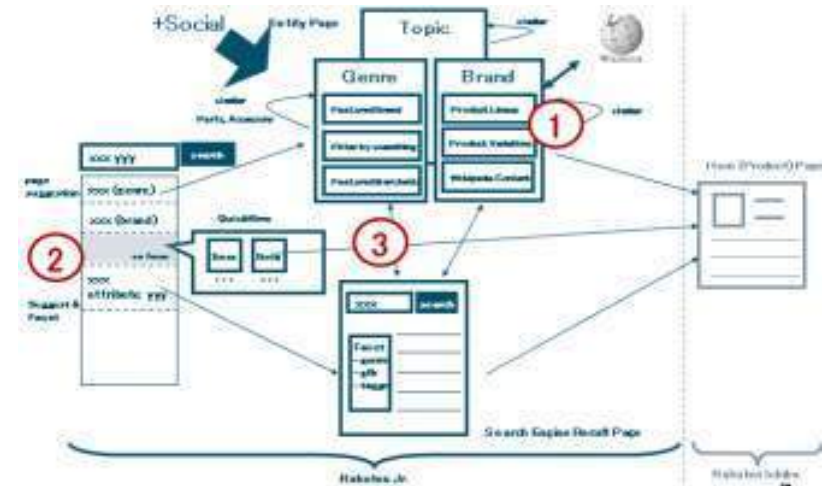
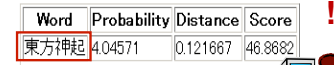
ミッション

今後大きく成長する技術のシーズから、
インターネットを活用した人々の生活(リアリティ)
を豊かにする
新しいサービス・事業の可能性を創出する

Tokyo & NY

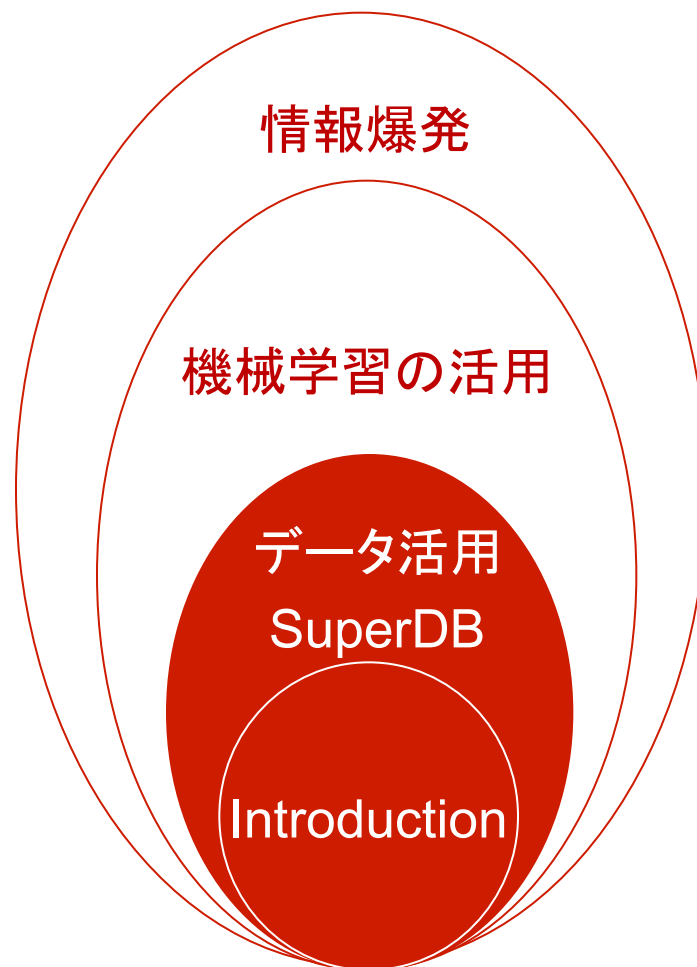


Search Tech



企業における データの活用

BigDataの時代へ



・情報の活用例



・Amazon, 楽天

- ・協調フィルタリングの典型的な適用
- ・その価値の発見

・Pandora Radio

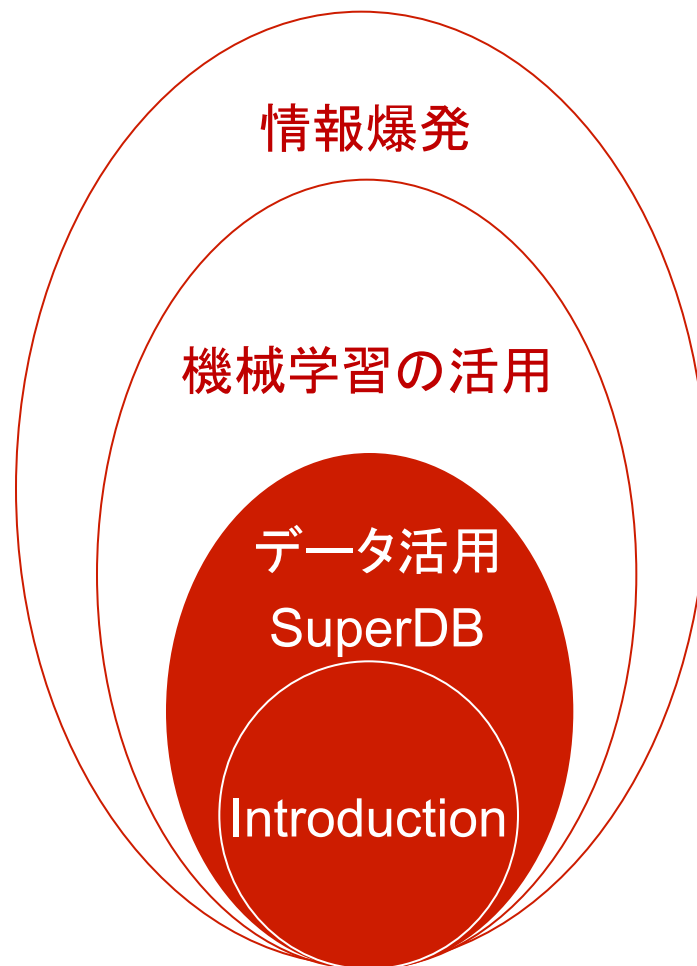
- ・ユーザデータに加え、曲の旋律、歌詞、声質、楽器、拍子、コード進行等を細かく解析
- ・ユーザの評価を組み合わせ、よりよい曲を提供
- ・1億人以上のユーザを持ち、成長を続ける



- ・情報解析のビジネス価値の発見
- ・インターネット企業においては、データ活用は主たる基幹機能
- ・データ分析の結果をオンラインサービスに直接反映。レスポンスをすぐに受け、短いPDCAで改善ができる、ビジネスのコア

SuperDB

BigDataの時代へ



E-Commerce



Portal and Media



Securities



Banking



Travel



Credit Card



Telecommunications



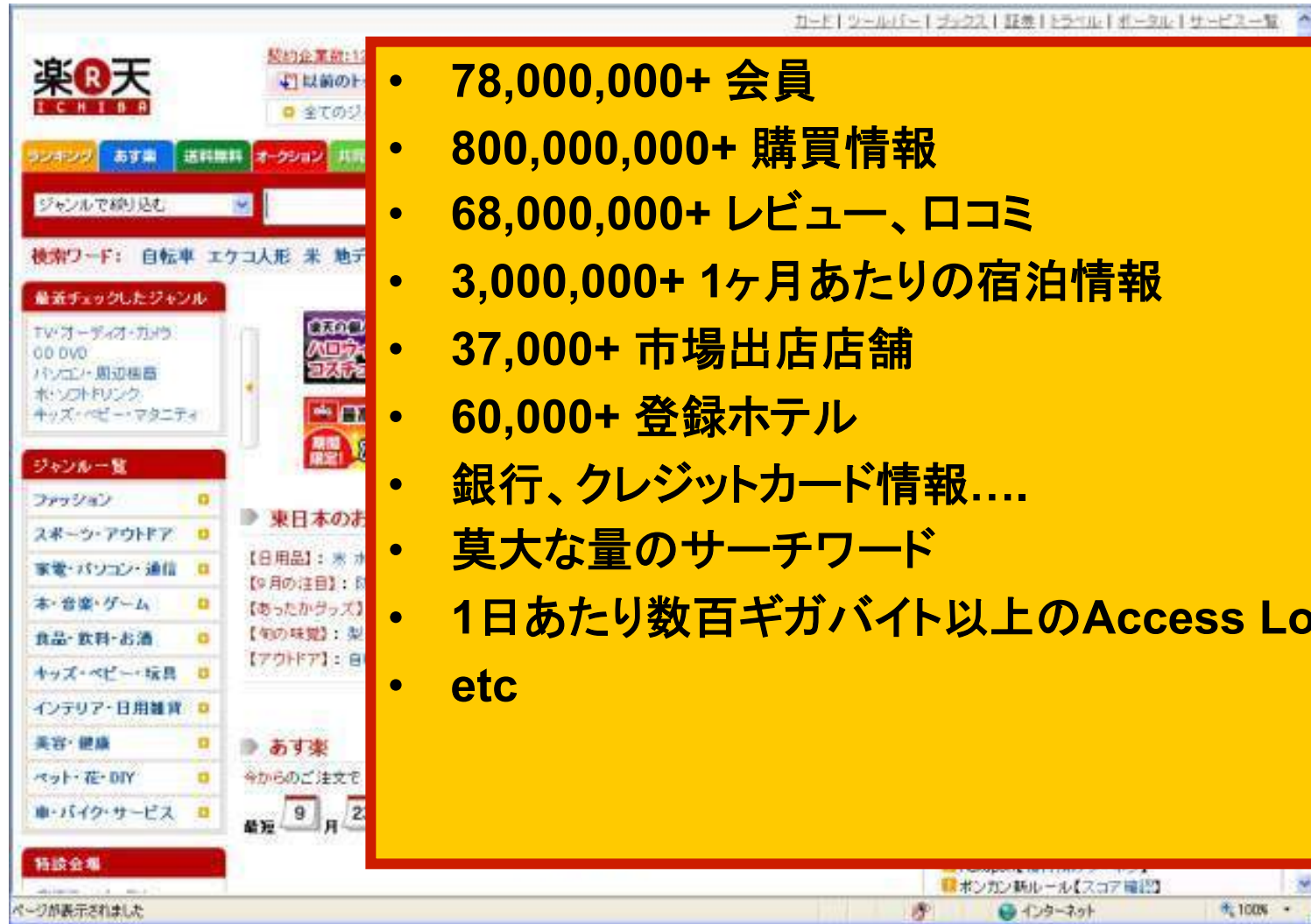
Professional Sports



E-money



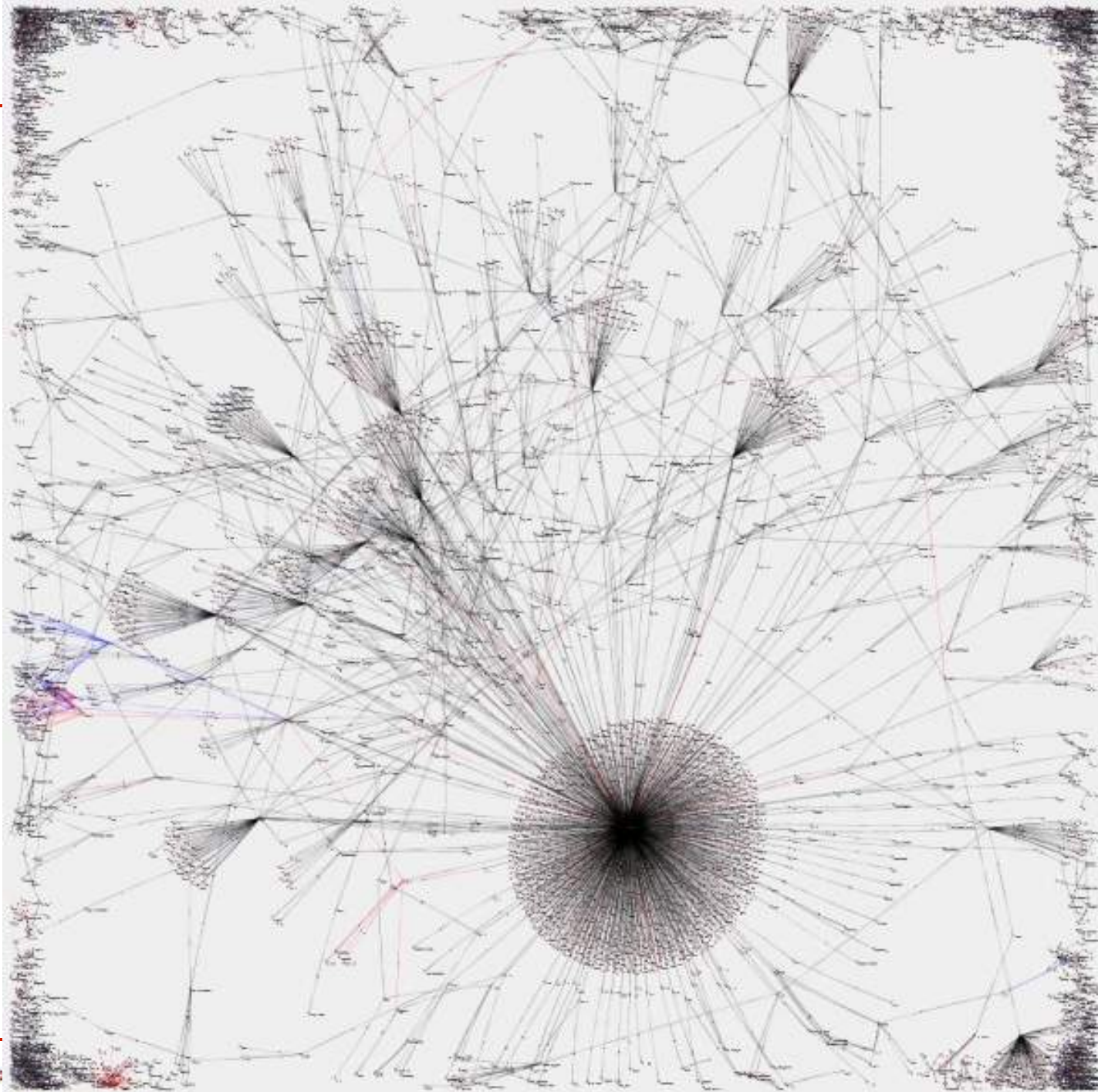
楽天にある大量のデータ



The screenshot shows the Rakuten website with a search bar, navigation tabs (ランディング, おすすめ, 送料無料, オークション, 共有), and a sidebar with category lists like ファッション, スポーツ・アウトドア, 家電・パソコン・通信, etc. The main content area displays search results for '自転車' (bicycle).

- 78,000,000+ 会員
- 800,000,000+ 購買情報
- 68,000,000+ レビュー、口コミ
- 3,000,000+ 1ヶ月あたりの宿泊情報
- 37,000+ 市場出店店舗
- 60,000+ 登録ホテル
- 銀行、クレジットカード情報....
- 莫大な量のサーチワード
- 1日あたり数百ギガバイト以上のAccess Log
- etc

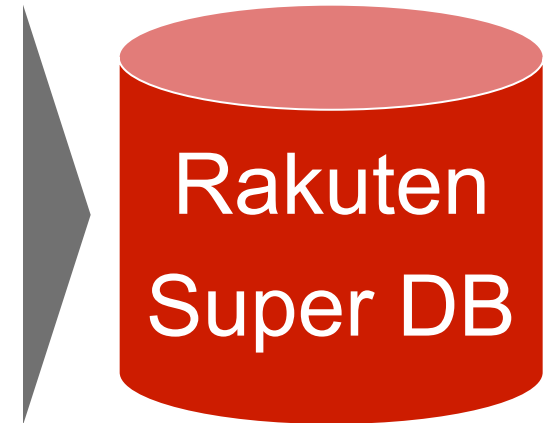
ユーザ・店舗間の関連の全体像

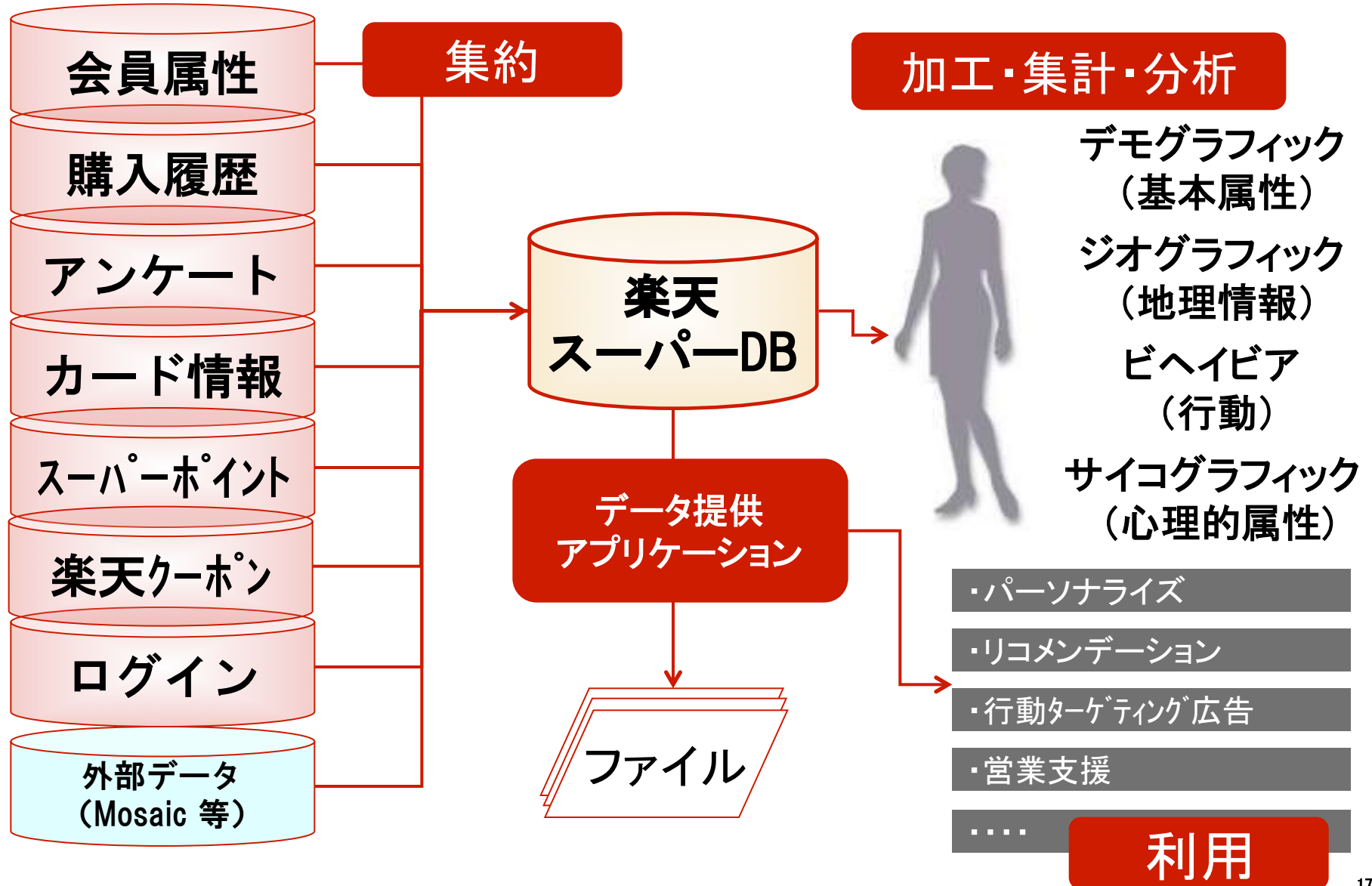


DWH

- Rakuten has tons of businesses, and so have many kinds of business data. It's diversified.
- We aggregate such data into one big dataware house.

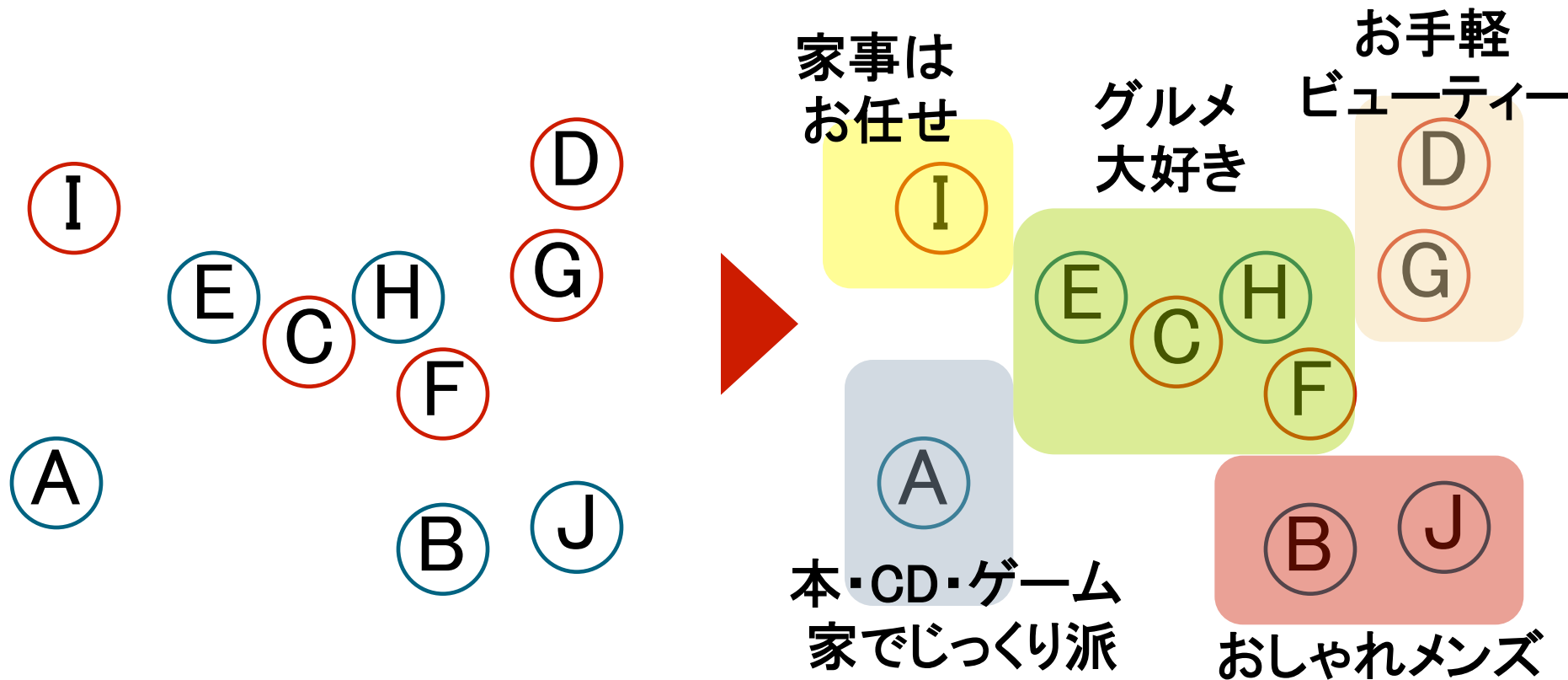
多様なビジネスデータ





顧客クラスタリングの考え方

顧客の属性データや購買履歴を利用し、顧客をいくつかのグループに分類する。→楽天会員全員をクラスタリング



パーソナライズ(バナー)

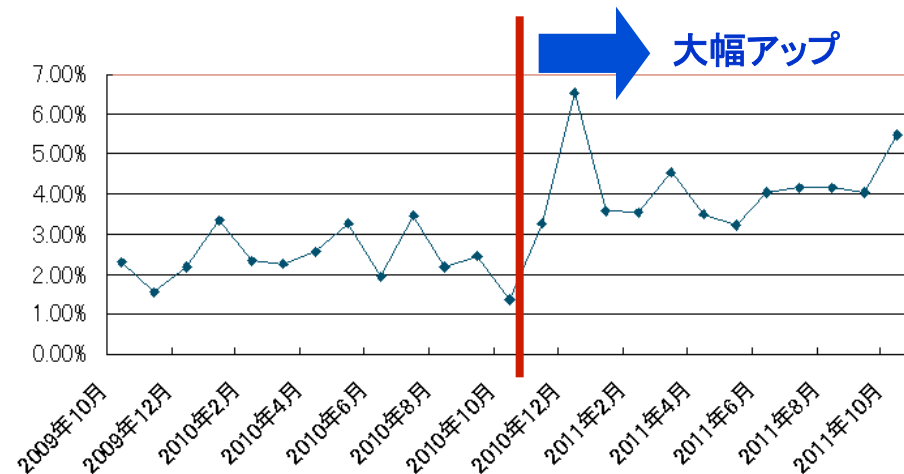
顧客をセグメント化し、適切なバナーを出し分け



カード未保有	条件	10%	20%	30%	40%	50%	60%	70%	80%
未保有		カード	カード	カード	カード	カード	カード	カード	カード
未保有 誕生月	5月生まれ	カード	カード	カード	トラベル	トラベル	ボックス	トラベル	トラベル
未保有 振替・代引		カード	カード	カード	イーバンクのカード	イーバンクのカード	イーバンク	イーバンク	イーバンク
ボックス初回移動	3ヵ月以内に市場初利用	カード	カード	カード	ボックス	ボックス	ボックス	ボックス	ボックス
誕生月	5月生まれ	カード(プレミアム)	トラベル	ボックス	トラベル	トラベル	トラベル	トラベル	トラベル
広島	pref.が広島県	トラベル	トラベル	ボックス	楽天銀行	楽天銀行	楽天銀行	楽天銀行	楽天銀行
GORA	GORA	トラベル	トラベル	トラベル	トラベル	トラベル	トラベル	トラベル	トラベル
ビットワレット	ゴールド以上でツールバー	トラベル	トラベル	ボックス	楽天銀行	楽天銀行	楽天銀行	楽天銀行	楽天銀行
競馬	toto買った人	楽天競馬	楽天競馬	楽天競馬	楽天競馬	楽天競馬	楽天競馬	楽天競馬	楽天競馬
写真館・家族写真1冊	キッズ持ち、子あり	写真館	写真館	写真館	写真館	写真館	写真館	写真館	写真館
オート	車特性・PG会員	オート	オート	オート	保険	保険	保険	保険	保険
決済	イーバンク口座保有者	ボックス	ボックス	ボックス	楽天銀行	楽天銀行	楽天銀行	楽天銀行	楽天銀行
Tポイント	Tポイント交換者	ボックス	ボックス	レンタル	レンタル	レンタル	レンタル	レンタル	レンタル
その他		ボックス	ボックス	o-net	楽天銀行	楽天銀行	楽天銀行	楽天銀行	楽天銀行
	カード新規入会者	初回移動	初回移動	初回移動	ボックス	ボックス	トラベル	トラベル	トラベル

効果:

よりきめ細かいセグメント作成をもとにした、バナーの出し分けを自動化したことにより、手動で作成していた時に比べ、平均CVRが約1%向上。



- # スーパーDB

ユーザー情報

購入履歴

閱覽履歷

ビジネス

DB for service

TOHO

【各種アルゴリズム】

協調フィルタリング

リターゲティング

バスケット分析

グループ^o関連度

コンテンツベース

各種サービスへ
展開



楽^R天 みんなの就職活動日記

▼ご注意▼ 以下の【こんな商品にも興味をもっています】欄の商品は、当ページ商品との機能的関連商品の意味合いはございません。

■この商品を買った人は、こんな商品にも興味をもっています。



楽天市場

★あなたへのおすすめ



DVD レンタル



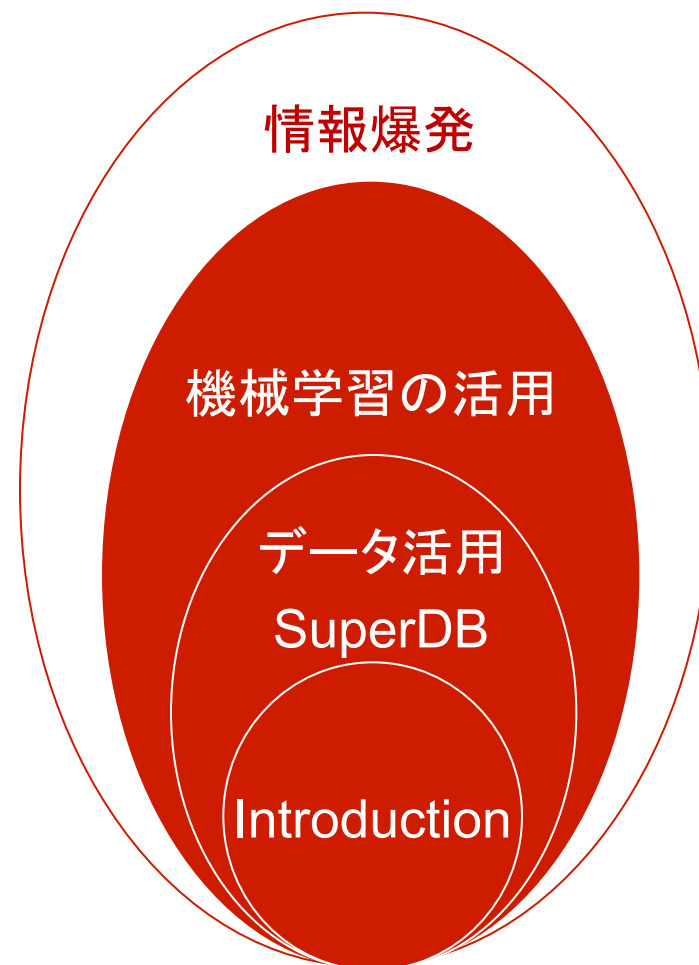
ブックス



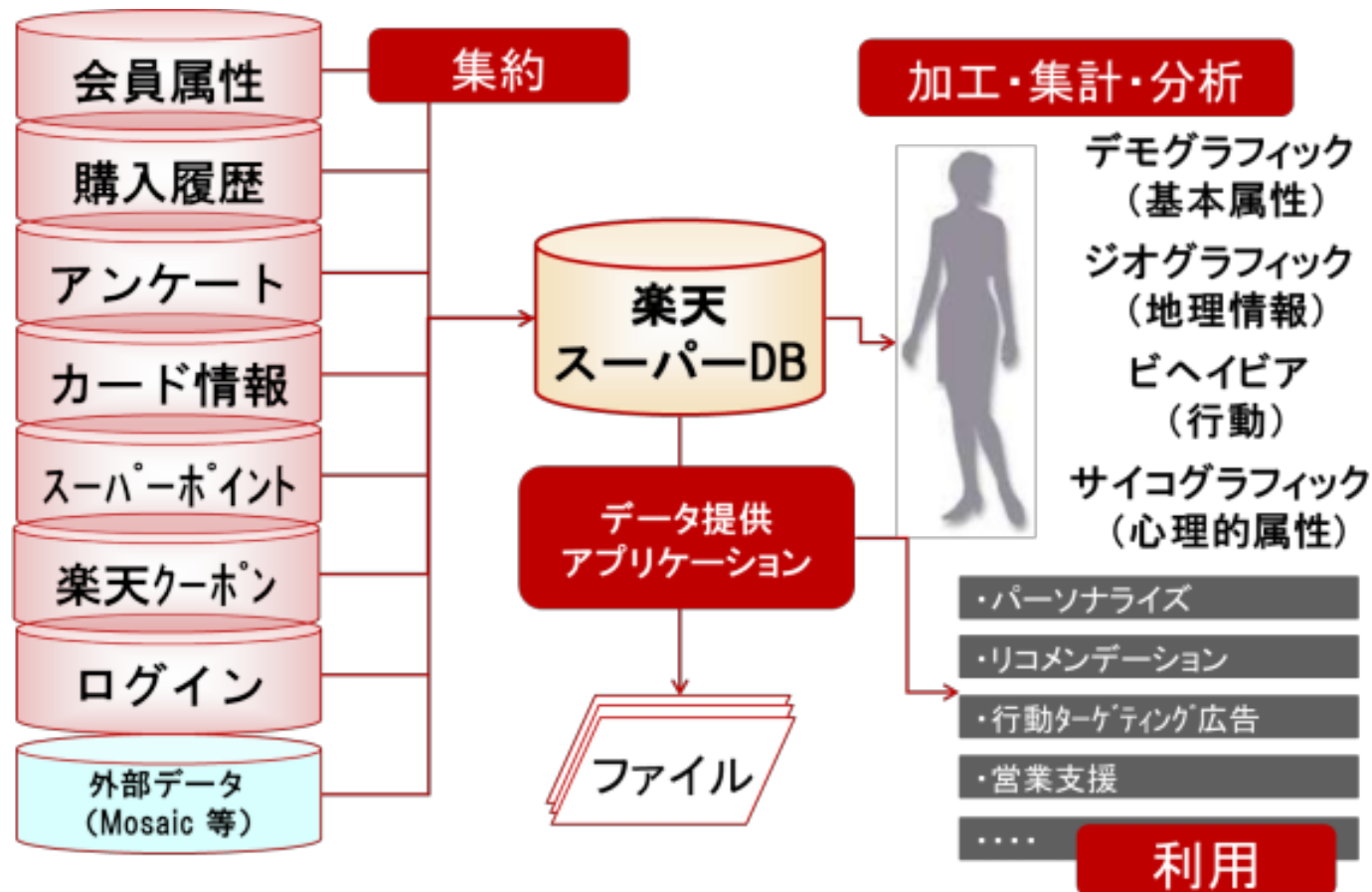
ダウンロード

BigDataの時代へ

「機械学習」の活用



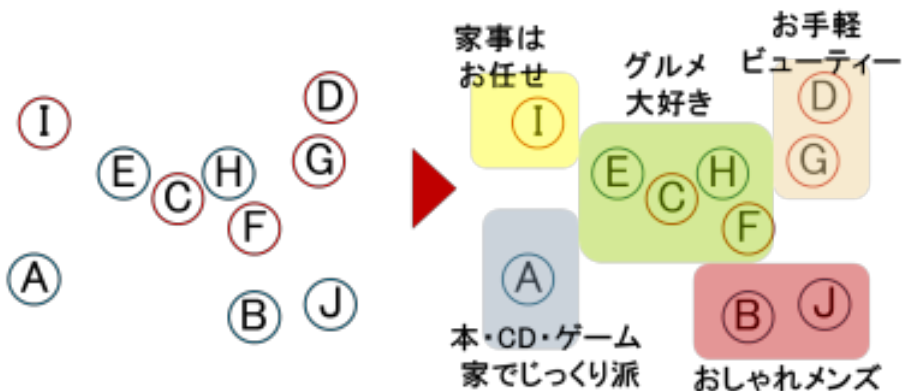
- 多種多様なビジネスデータを格納したスーパーDBの活用より、楽天においてはデータ分析・活用のシーンは幅広く見られるようになった。
- その中では、「機械学習」の活用も主にマーケティングの部署や楽天技術研究所において広まってきている。



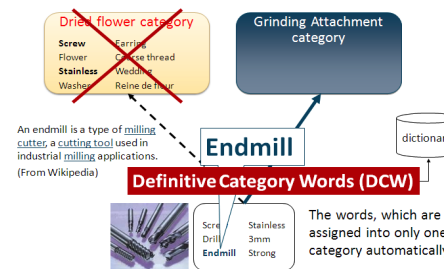
- ・コンピュータが過去のデータから情報処理方法を学習、未知の問題解決を行う
- ・データに対するアプローチや基本姿勢として業務の現場において実践

- ・量的なあるいは時系列的なデータの分布から
- ・クラスタリング(教師なし学習)を行うのは当たり前
- ・統計やNLPの活用分野も多く、教師ありも多い
 - ・統計の回帰はいたるところで広く用いられており、これも機械学習の一種と言えれば一種

スーパーDBのクラスタリング

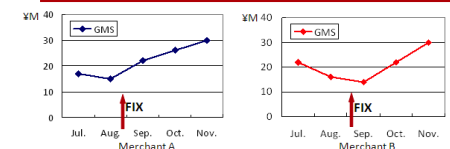


Global Catalogue Creation Noise Detection

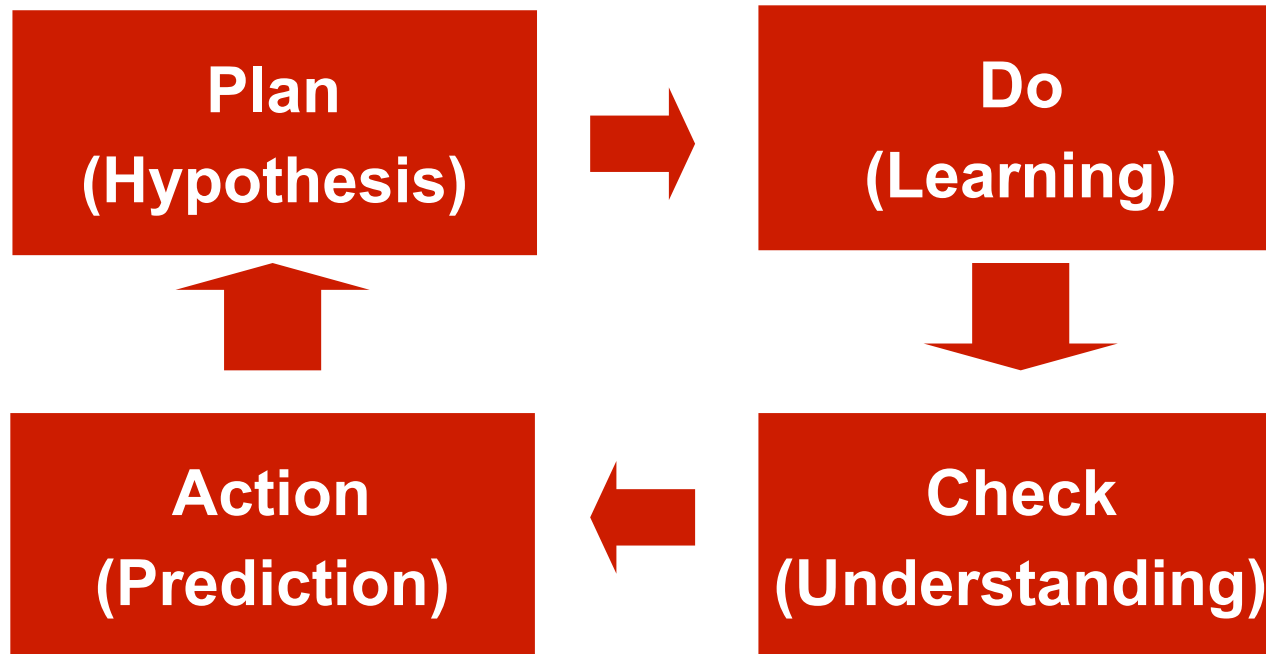


- (1) Constructing a DCW dictionary for 2 big merchants
- Both merchants provide over 2 million products
 - Manually annotated over 7k entries

Detected and Fixed over 2 million noise products!

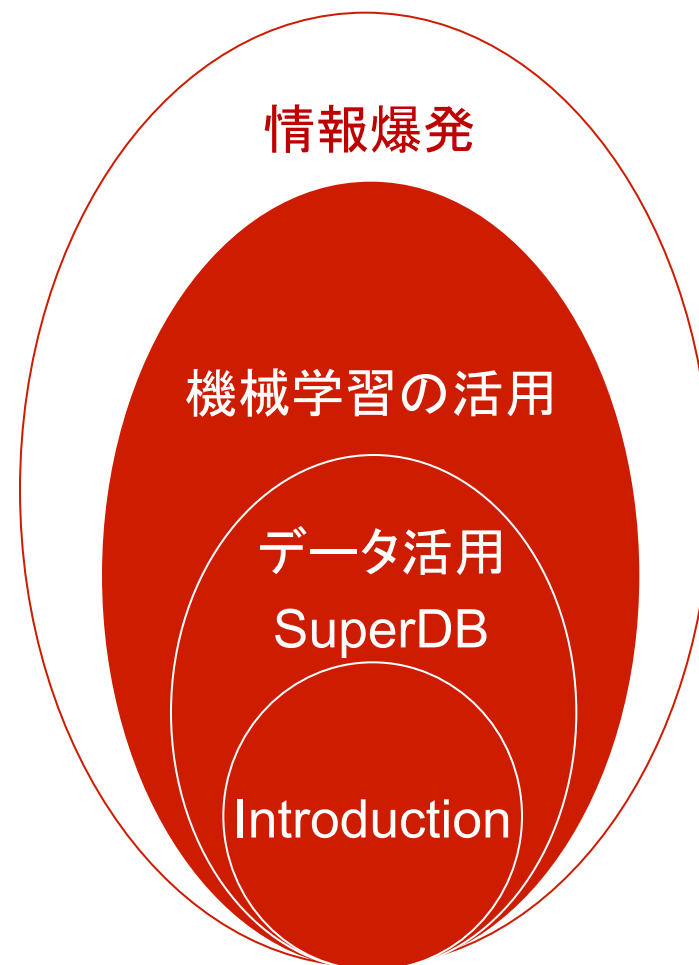


- 学習させ、例えば分類器を作って、PDCAサイクルにのせる
 - 全体やトレンドの理解、未知の事実の発見をしたりした結果を
 - フィードバックとしてビジネス施策やサービスアプリケーションに実装
- 結果、ユーザの反応がダイレクトに得られ
- PDCAサイクルによる継続的な改善活動を、短期間で回すことができる
 - インターネットサービスの大きな特徴であり、醍醐味



BigDataの時代へ

「機械学習」活用事例



- レコメンデーションでの活用。
- K-Means、pLSI (ソフトクラスタリングの一種)。
- 嗜好の似たユーザーかどうかの分析でLSH (Locally Sensitive Hash)。

5ten

- 似た人が買った商品を推薦
- 意外性のある商品が推薦できる



「併売」
を促す目的

Collaborative Filtering
Basket Analysis

0-HO

- 商品名/商品説明が類似する商品を推薦
- ジャンルがほぼ同じ商品を推薦できる



「回遊」
を促す目的

Text Matching
Clustering

萬 gekyo

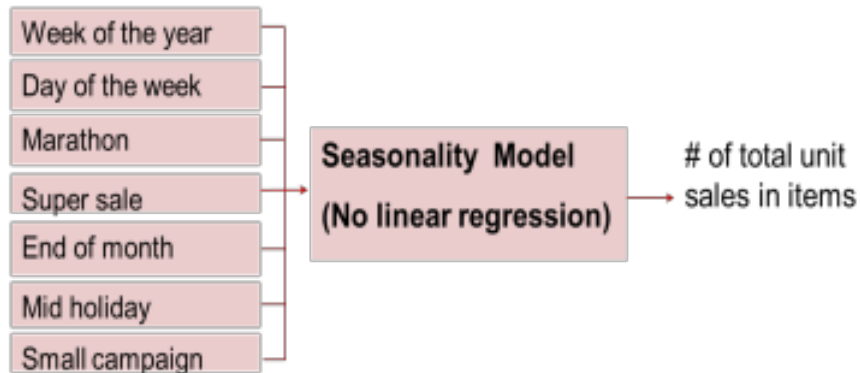
- 顧客を複数のクラスタに分類し、それぞれの顧客クラスタごとに相性のよい/よく買われる商品を推薦
- 購買履歴がなくても商品を推薦できる



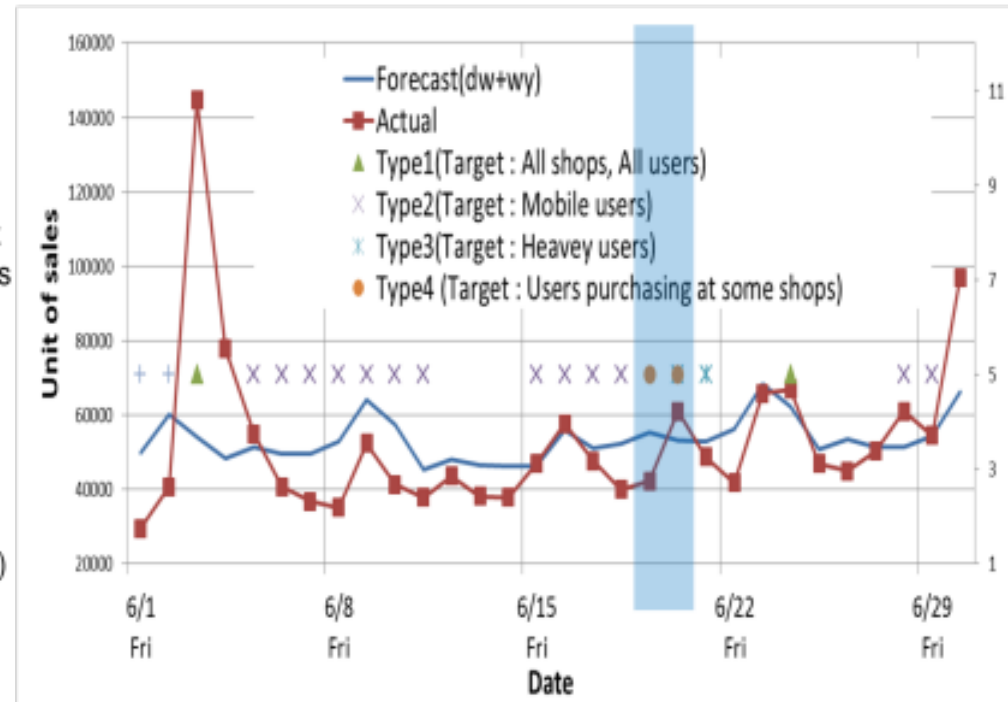
Cluster Coefficient

- 季節性・イベント等を加味した、商品販売量の予測
- 教師あり学習（非線形回帰）
 - 全体の販売量を被説明変数
 - 週、月、キャンペーン、月末、連休、温度、等を説明変数に
- 在庫・価格の最適検討に関するインプットに

- Overview of the model is as follows.



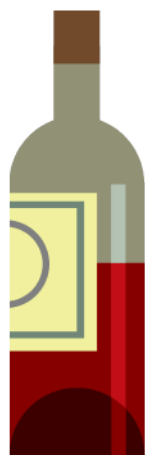
- We use the no-linear regression because
 - Shape of the sales seems to be no-linear (Huge spike at some events)



- 系列ラベリング問題として捉え、CRF(条件付確率場)で解く
- ブートストラップ法(半教師:少量教師+大量教師なし)での生成

WINE > CHILE

Montes Alpha M 2009



Montes Alpha M is a blend of Cabernet Sauvignon, Merlot, Cabernet Franc, and Petit Verdot.

A powerful wine with very good level of soft and rounded tannins. Intense dark red color. The wine is elegant and has a very well defined character. ...

Item page



Attribute	Value
Color	Red
County	Chile
Grape	Cabernet sauvignon, Merlot, Cabernet franc, Petit verdot
Vintage	2009
Company	Montes

Structured data

日本は ロングテール

- ・地方毎に特産品がある、コアな趣味を持つ人が多い
- ・販売者もロングテール、購入者もロングテール
- ・常識ではかれないものが飛ぶように売れる

昨シーズンは18万本を販売
電子レンジで約3分

国産

あったか〜いやきいも!

黄金色の焼き芋

共同購入 最終価格 **263円**



ジャぱらドリンク



じゃばら

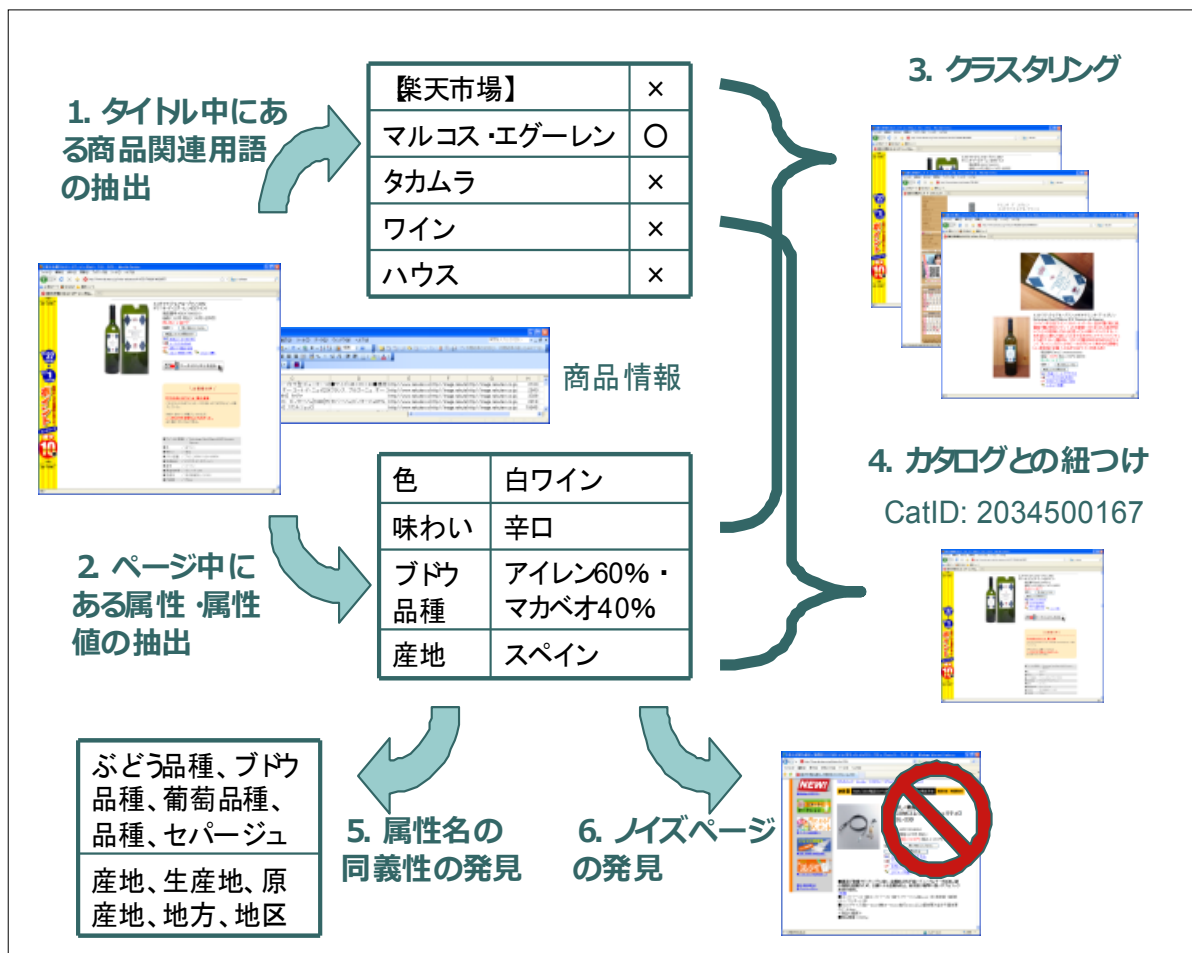
信州伊那谷のたまごやさん

産直農産物 かんぱん 安い物がこ ござい・送料 食品一覧 オークション 共

きっかけは
たまごの通販
それどうよ!?
って感じでした

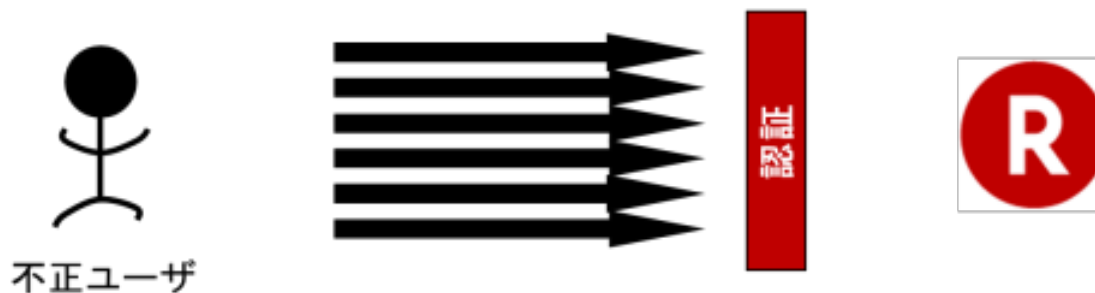


- ・系列ラベリング問題として捉え、CRF(条件付確率場)で解く
- ・ブートストラップ法(半教師:少量教師+大量教師なし)での生成



- 教師なし**: ログインアタックの検知の際に, どのようなアタックパターンがあるのかを検知するためにクラスタリングを利用
- 教師あり**: IP, 検索語の長さ、検索を行ったサービス、国、等を素性としてモデルに学習させて分類機を作成。(SVM/ Passive aggressiveを使用)

ログイン試行を繰り返すケースは比較的容易に防御可能.



他人のID・PWを不正に手に入れ, 認証を一発ですり抜けるケースはどうするのか?



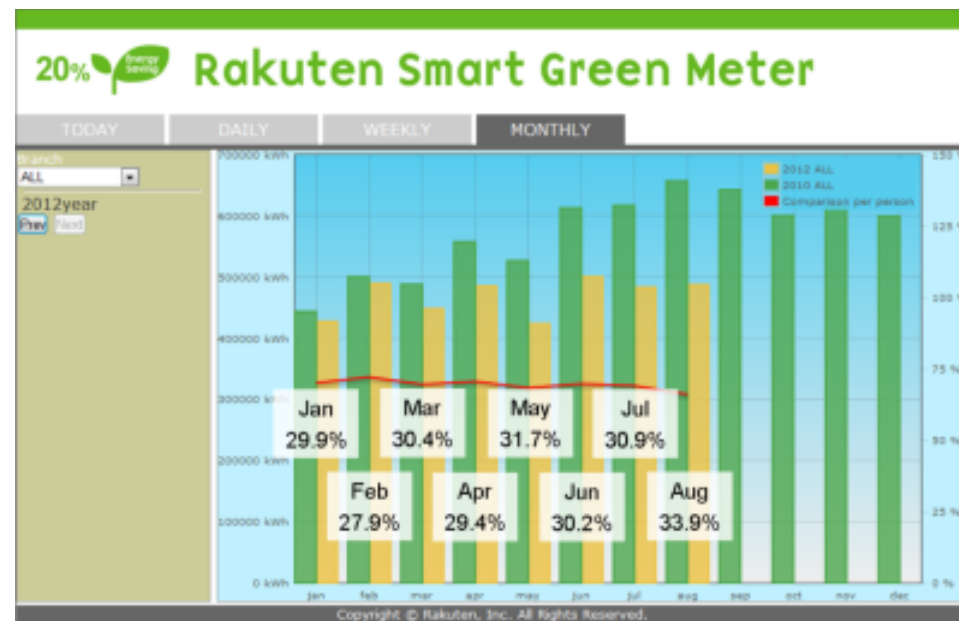
・ログイン後の行動パターンを検知し, 実被害を抑える.

・ユーザIDとIPアドレスの関連性を考える.

- SNSサービスにおける、●●系画像を抽出するフィルタの作成
- SOM(教師なし。自己組織化マップ)で分類しようとしたら時間かかりすぎ
- 画像の色に関する分布をEMアルゴリズム(教師なし)で楕円フィッティング
- ニューラルネットワークで抽出したパラメーターと画像のOK/NG 対応の学習
 - 教師あり: FFNN(フィードフォワード)、多層パーセプトロン

No Image

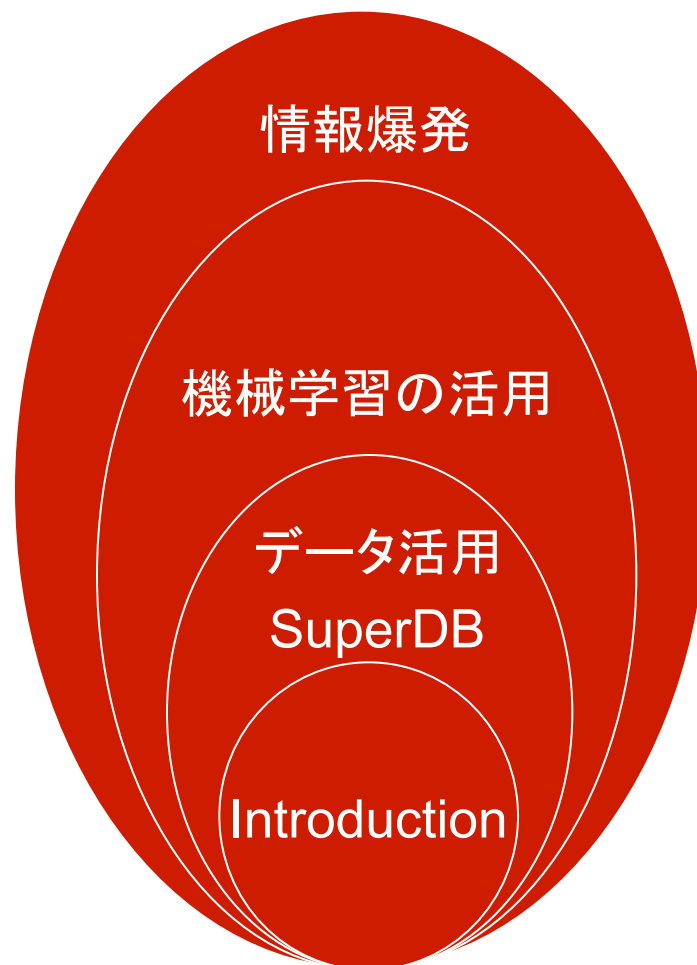
- RSGP (楽天スマートグリーンプロジェクト)
- センサーをつけ、計測。過去データより学習し、電力使用量の予測も。
- 単純な回帰。シーズナリティにいきたいが。。。



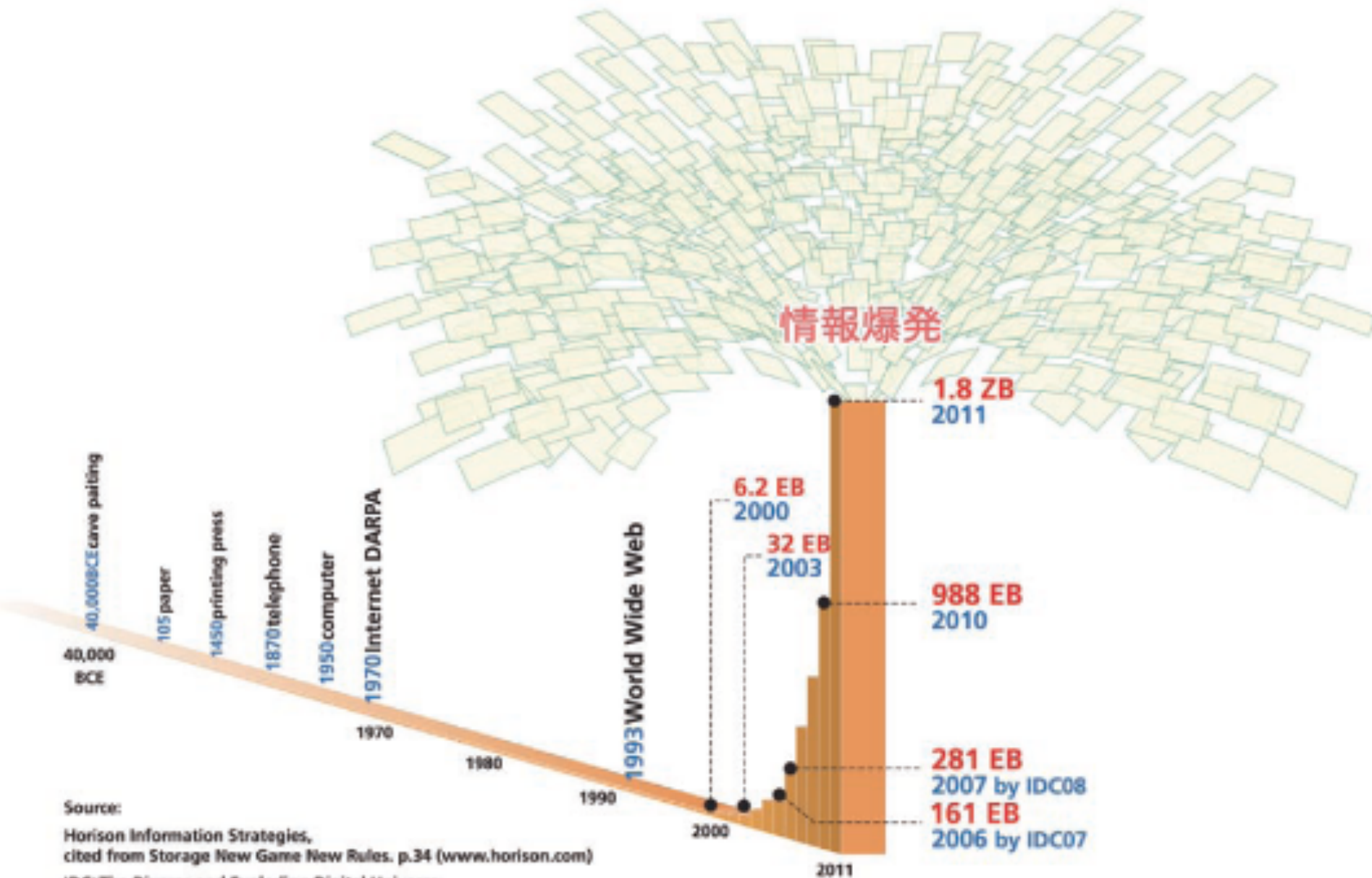
- ・何も高度な技術を適用していない。
- ・何(データ・ログ)を対象とするか。どう適用するか。
 - ・データ・ログ収集・基礎集計の戦略。
- ・変数に入れるためのデータの加工の仕方。
 - ・データ・ログの標準化・統一化。
- ・あるいはそもそもの仮説の立て方。
- ・それらこそが重要。
- ・**その基本を外さないこと**こそが、「機械学習」活用において肝要。

情報爆発 BigData

BigDataの時代へ



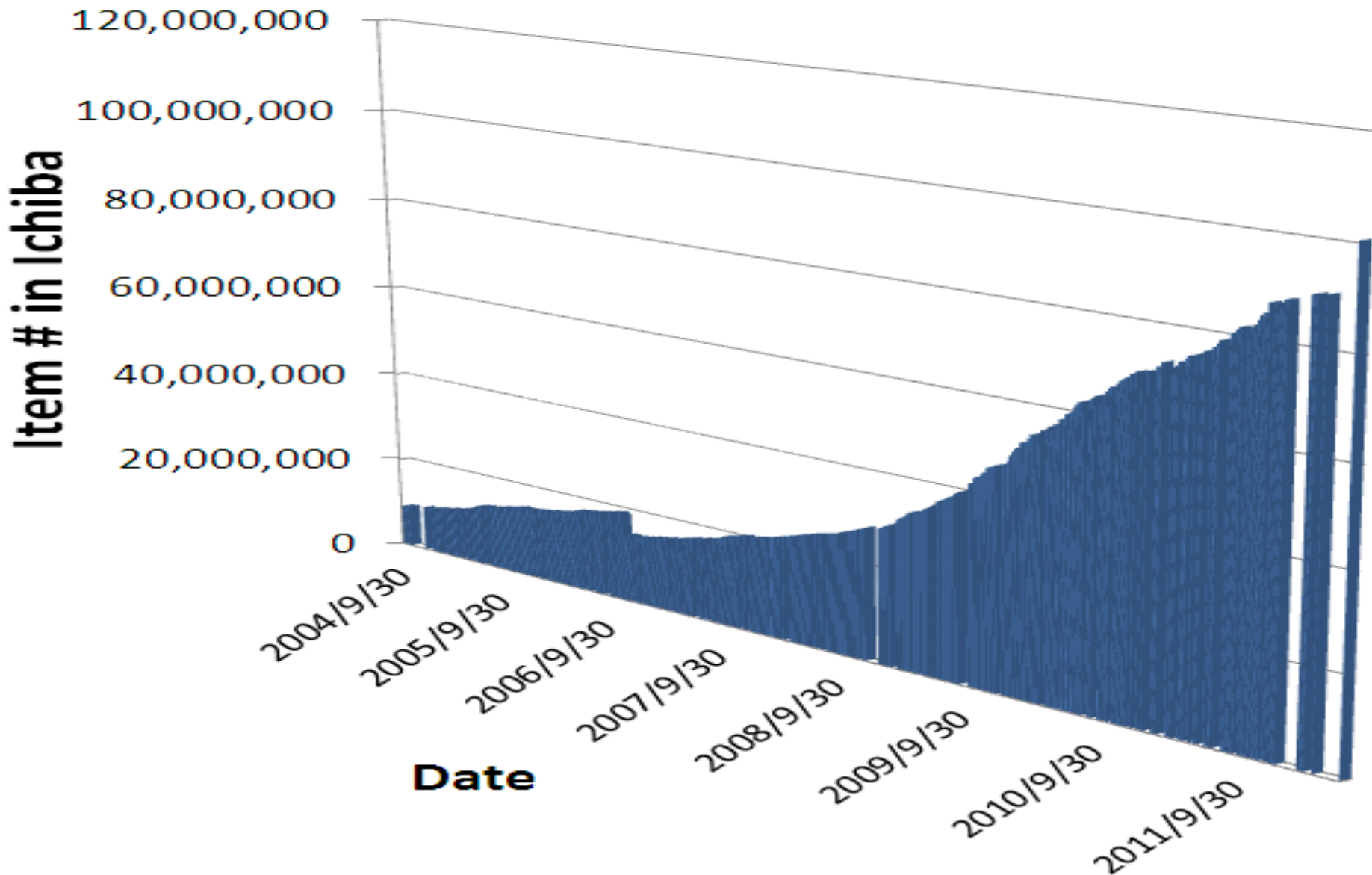
情報爆発



Source:
Horison Information Strategies,
cited from Storage New Game New Rules, p.34 (www.horison.com)
IDC: The Diverse and Exploding Digital Universe

Along with this, we are increasingly getting difficulty of processing data.

Big Data



・大規模基盤の活用・構築

- ・クラウドの普及、技術のコモディティ化を踏まえ
- ・パブリッククラウドの活用、プライベートの構築
- ・Hadoopクラスターの構築、NoSQLの活用・構築



- ・レコメンド、ターゲティング広告、検索改善、ログ解析での活用
- ・OSSの活用、貢献、開発が進む

楽天プロダクトランキング

■処理内容

- ・ランキングデータ集計
- ・製品ページ用データ生成

■データ量

- ・1日/1億レコード
- ・1日/300GB

■M/R 処理時間

- ・1時間半

■ノード数

- ・70台

サーチエンジン

検索ワード

価格

レビュー

商品件数

購買履歴

売上件数

売上高

売上率

購買時間帯

製品マスタDB

製品名

製品コード

スペック

発売日



家電・AV・カメラ

注目ランキング1位



日本ビクター

Victor・JVC

GZ-HM670-N

最安価格(税込):

¥30,260円

ドラゴンクエスト25周年記念 ファミコン&スーパーファミコン ドラゴンクエストI・II・III スクウェア・エニックス



最安値: 3,780 円

最安ショップで購入する

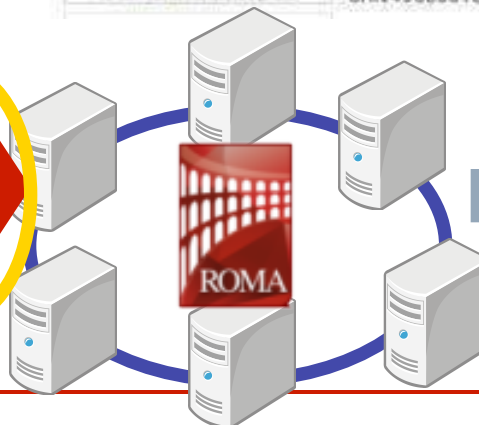
★★★★★4.54 (244件) 売上ランキング1位 注目ランキング1位

お気に入り 34人

JAN4988601007146

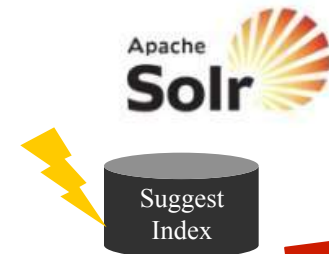
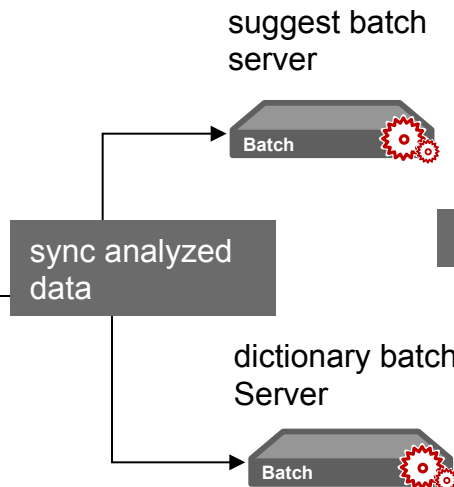
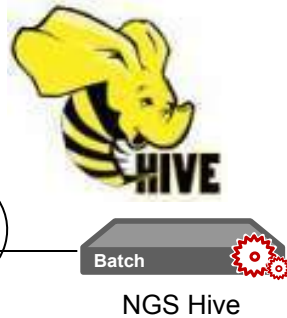
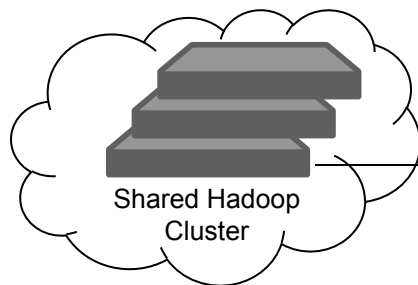


Calculate



- クラスターから検索解析用のHiveに
- 関連語の提示や辞書構築等での活用

日次 300GBの
データを解析



update search index

検索エンジン

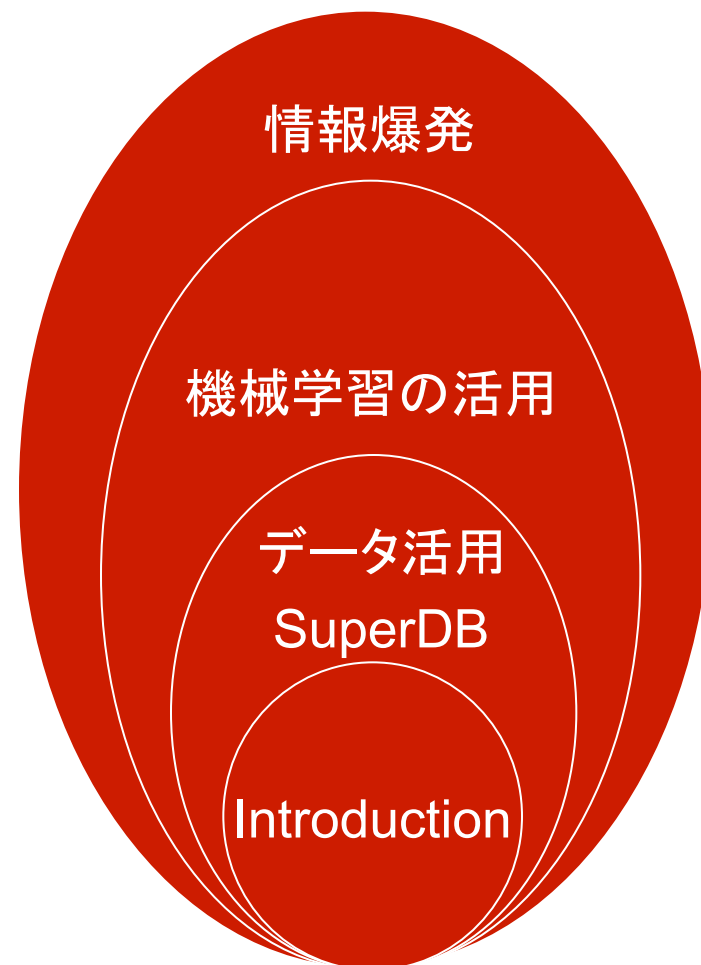


update search index

- ・機械学習の知識は、企業におけるデータ活用において重要な位置を占めつつある。
- ・今後は、データが大規模化。
- ・「機械学習」の知識に加え、**大量データ処理の技術、分散プログラミングの技術**もあわせて必要。
- ・そのため、今後はそのような他のスキルや知識を踏まえた人材が要求される。
- ・しかし、大事なことは「機械学習」の知識の活用機会というのは、ますます広がりとつあるということ。

BigDataの時代へ

For closing



アカデミックへの貢献

楽天の様々なデータを大学、公的研究機関に公開

楽天技術研究所－楽天データ公開

<http://rit.rakuten.co.jp/rdr/index.html>



目的:

- －先進的IT企業として学術的な分野での技術の発展に寄与
- －産学の交流連携をはかりインターネット技術分野の進歩のサイクルを加速
- －楽天研究開発シンポジウムの開催を通じて、独創的な研究を促進

これにより、アカデミアの世界に対して楽天が貢献できると共に、これらの研究を通じて、外部の方により楽天に興味を持って頂くことも期待されます。

<商品データ>

Column	Sample
商品コード	rdk00000000
商品ID	12345678
商品名	[Rakuten 5000] 楽天ロイヤリティ 着心地最高のデニム 赤を基調としたデザイン
商品説明文	上質の素材を使用し、シルエットにもこだわったデザインです。ストリートだけでなく、有名百貨店でも取り扱われるようになりました。人気色上質のデニムです！パーフェクトなシルエットでありながら、年々、性別も問わず愛用していただけるデザインになります。オンにもオフにも、デニムのジャンルにもコーディネートに最適です。動きやすいだけでなく、着心地にもこだわりました。
商品URL	http://www.rakuten.co.jp/rakuten/00000000/12345678/
商品画像	http://i1.rakuten.co.jp/00000000/12345678/00000000/12345678.jpg?one=0&size=12345678
商品価格	4,600
ジャンルID	4000000

<レビュー>

Column	Sample
投稿者	user1
年齢	20
性別	1
商品コード	rdk00000000
商品名	[Rakuten 5000] 楽天ロイヤリティ マグカップ
商品ID	12345678
商品ジャンル	12345678
商品ジャンルID	12345678
商品価格	1500
購入フラグ	1
評価	実用性、デザイン性
目的	自分用
用途	はちめて
評価ポイント	4
レビュータイトル	バズで買いました！
レビュー内容	楽天してから今日で届きました。思ったよりも軽くて持ちやすく、使いやすいです。バズで買ったのですが、思ったよりも軽いため使いやすくなりました。
レビュー登録日時	2010-01-01 00:00:01

<宿泊評価>

Column	Sample
投稿番号	99887766
ニックネームID	12345
目的	レジャー
同伴者	家族
評価1 (立地)	4
評価2 (部屋)	4
評価3 (食事)	0
評価4 (風呂)	4
評価5 (サービス)	3
評価6 (設備)	3
評価7 (総合)	4

<ゴルフ場データ>

Column	Sample
ゴルフ場ID	1234
ゴルフ場名	マンハッタンゴルフコース
郵便番号1	398
郵便番号2	0085
住所	埼玉県楽天市
開場年月日	1998/4/1
施設コメント	河川沿いのフラットなコースで、インターバルも短く、回りやすいコースになっています。川沿いの雄大な形式のなかで、プレーを楽しめます。ビギナーにもおすすめのコースです。お一人様や当日の受付も行っております。
ゴルフ場コメント	毎月1日は女性限定のビギナーズデータを開催しております。詳しくはHPをご覧ください。※毎月第1月曜日は定休日です。
シェーズ指定	2



E-commerce企業における ビッグデータへの挑戦と課題

-機械学習への期待について-

A horizontal banner with a red diagonal stripe. On the left, a man in a suit. In the center, a woman wearing a colorful hijab. On the right, a man in a suit and a woman with long dark hair. The text 'Empowering Society Through the Internet.' is written in white across the top. The text 'Rakuten aims to become the No. 1 Internet service company in the world.' is written in white on the red stripe.

Empowering Society Through the Internet.

Rakuten aims to become
the **No. 1**
Internet service
company in the world.

Rakuten Inc. RIT. Masaya Mori Nov. 7th, 2012