Hadoop Recommendation Machine Learning

2014/7/16
EVF2014

スマートプレイス開発部　吉田　匠

自己紹介

- 吉田　匠
  - スマートプレイス開発部
  - 11 年目

- 検索エンジンを苦節 (?) 8年
  - 社内、コーポレートサイト、ECサイト
  - Apache Solrが主力、Elasticsearch は次世代のホープ！
  - 昔は、FAST ESP / Autonomy IDOL / Verity K2 など

- 今年のテーマはクラウドと機械学習
  - AWS ソリューションアーキテクト取りました！
  - Softlayer は勉強中？
  - Hadoop / Mahout を中心に調査・検証

amazon web services Certified
Solutions Architect - Associate

本日のメニュー

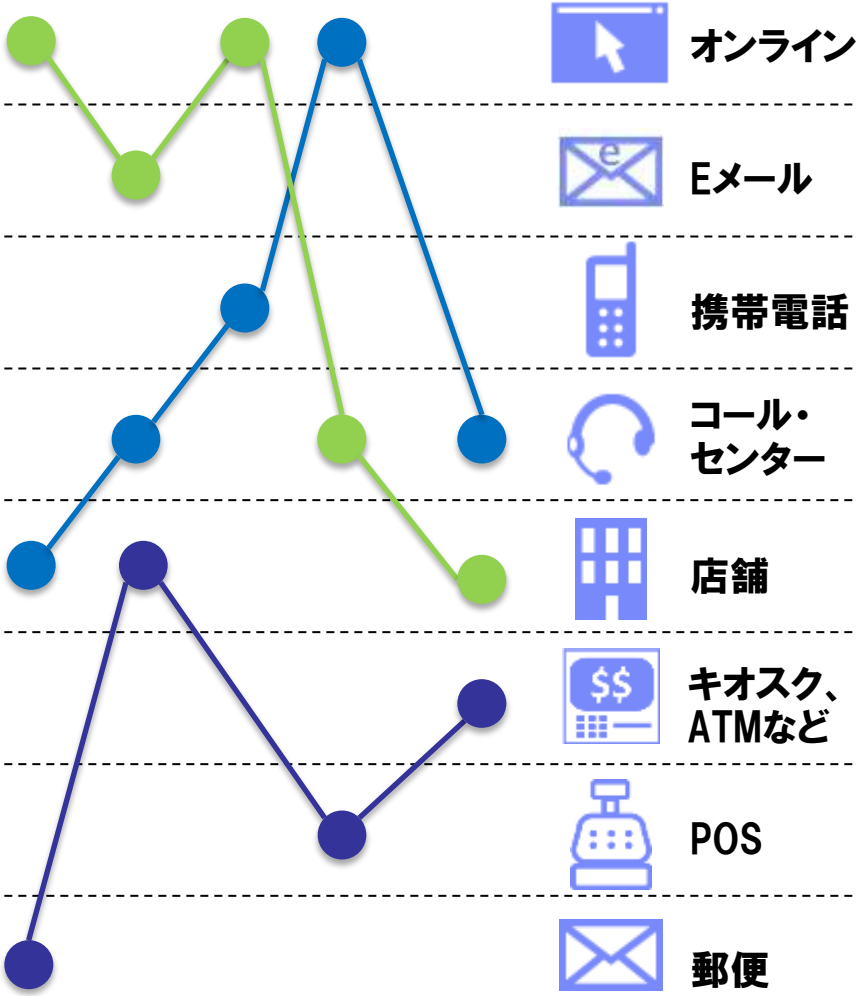
- ことの経緯
  - スマートプレイスについて
  - Hadoop についてのおさらいと最新の動向
- レコメンドとは
  - レコメンドの基本的な仕組みと課題
  - オムニチャネル・レコメンド
- 機械学習への応用
  - 機械学習とは
  - 現実的な課題と皆様へのお願い

ことの経緯

スマートプレイスとは？

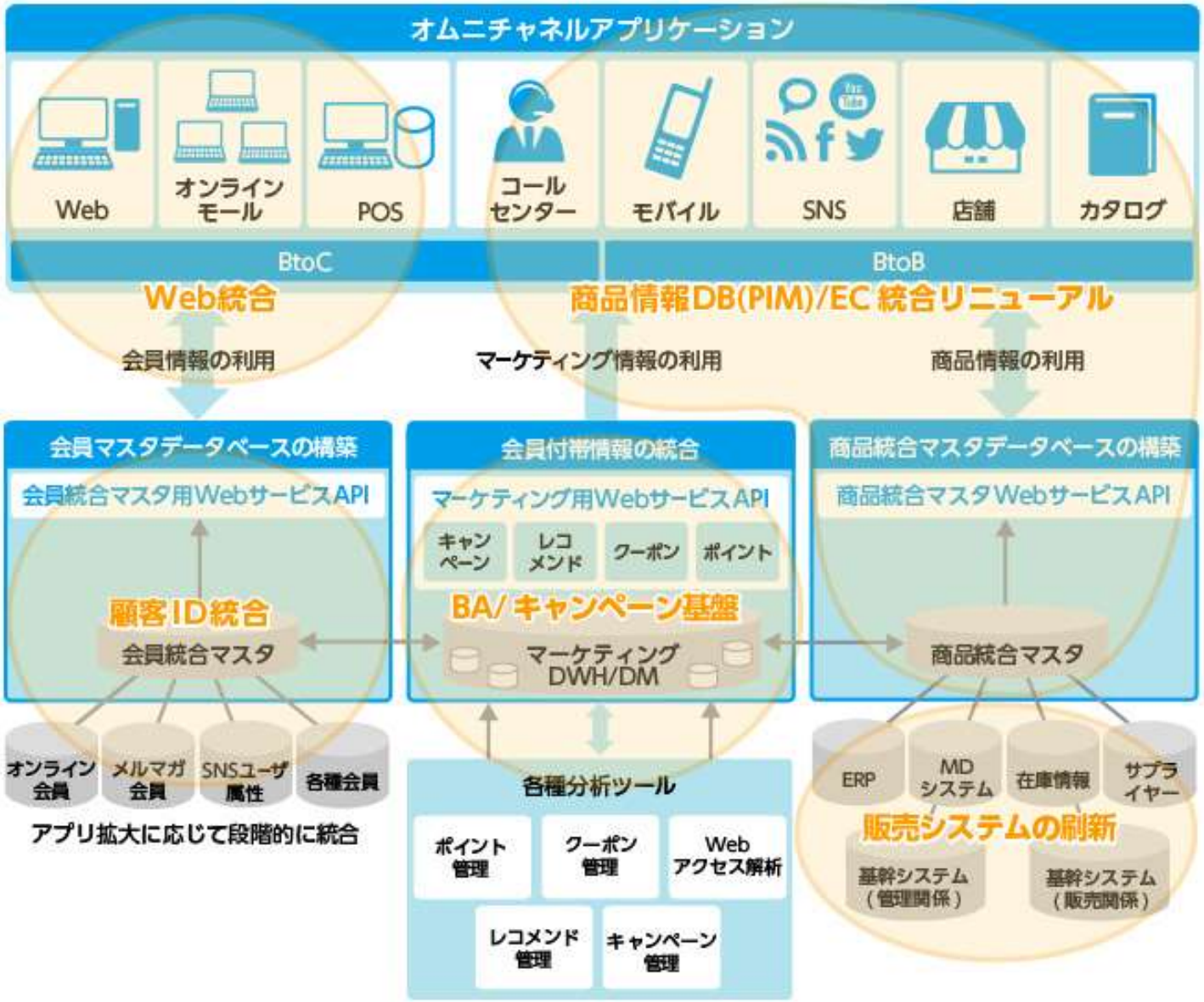
「マルチチャネル」から「オムニチャネル」へ

顧客

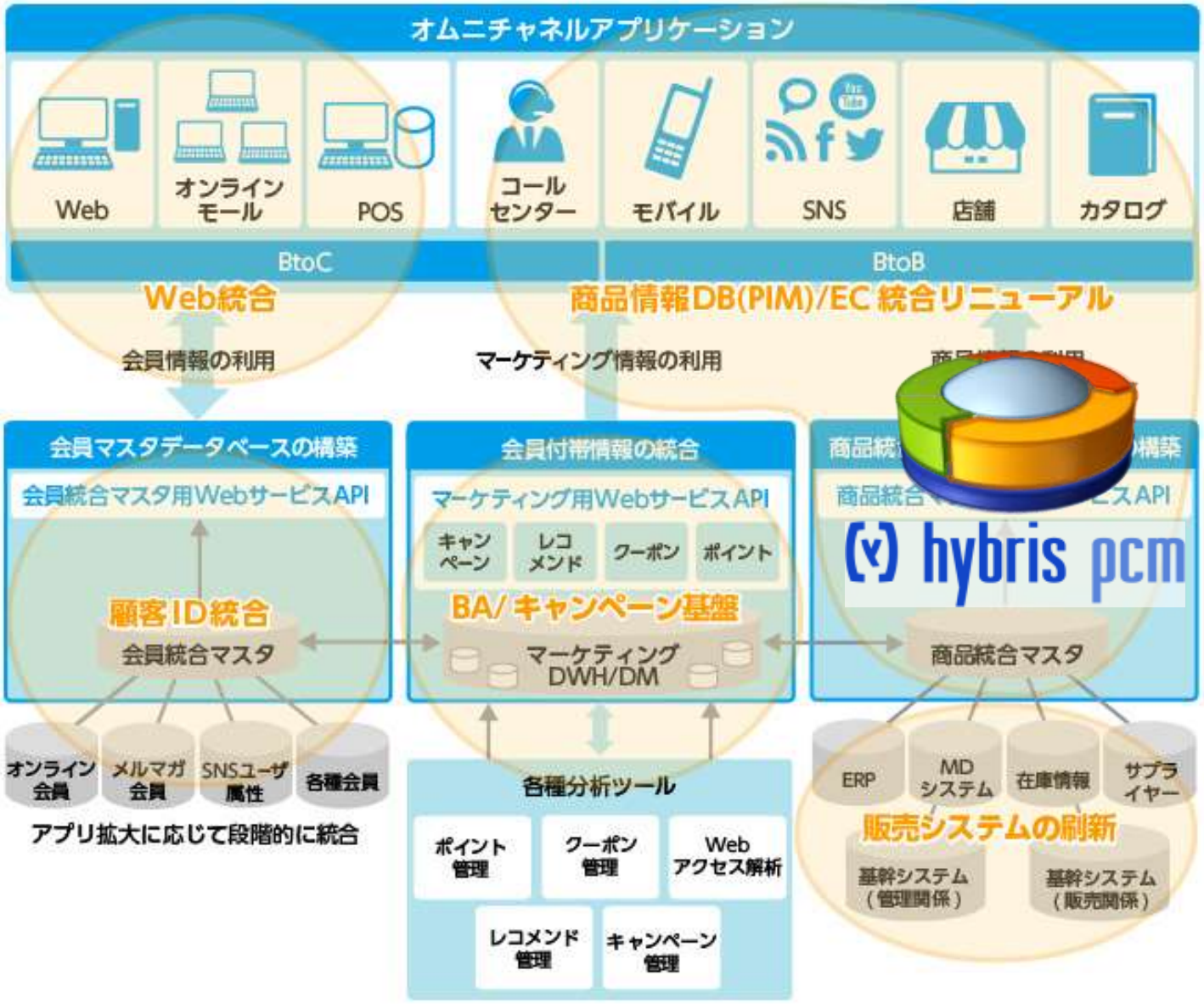


企業

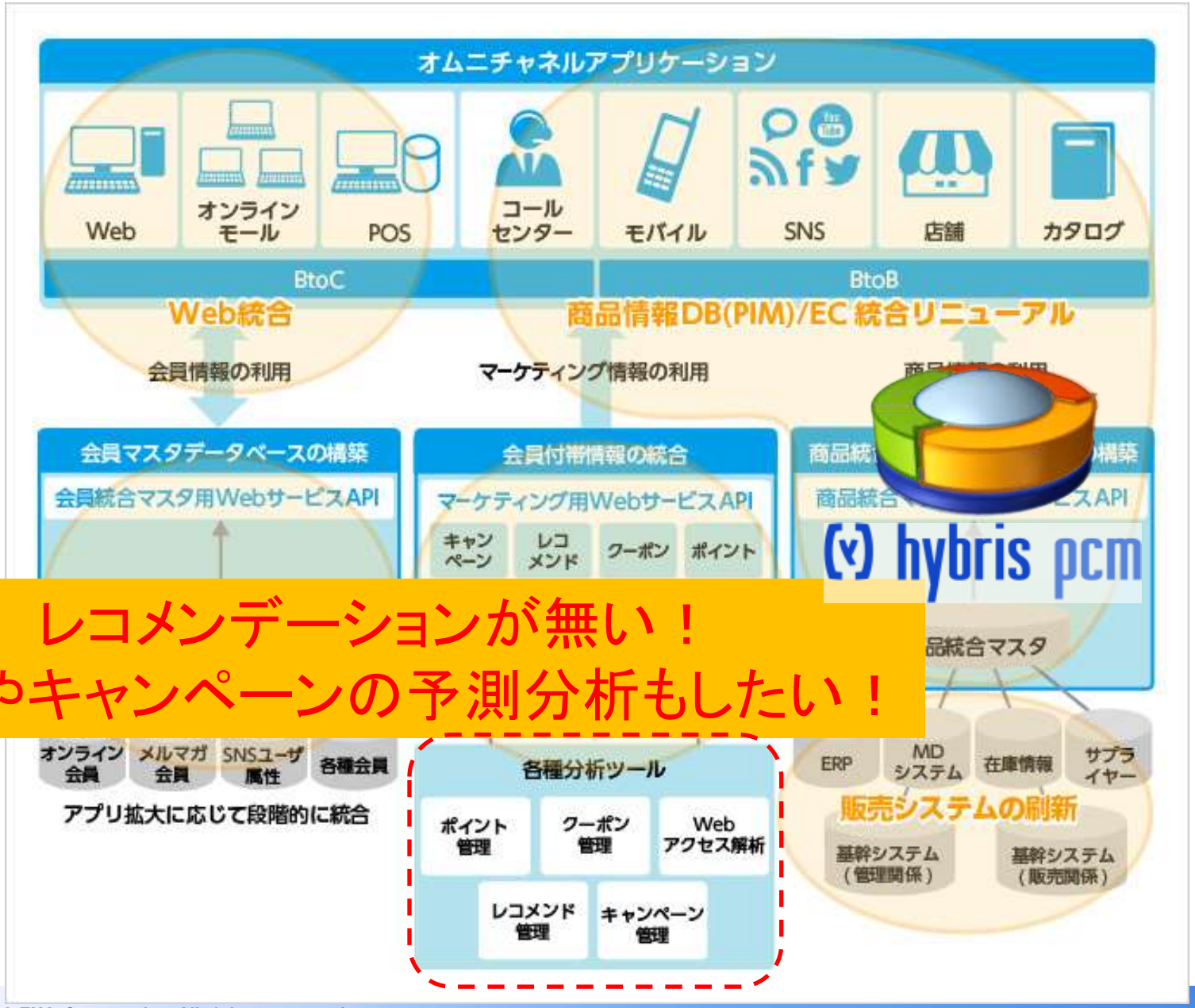
スマートプレイスとは？



スマートプレイスとは？



スマートプレイスとは？



レコメンデーションが無い！
売上やキャンペーンの予測分析もしたい！

って、何だっけ？？

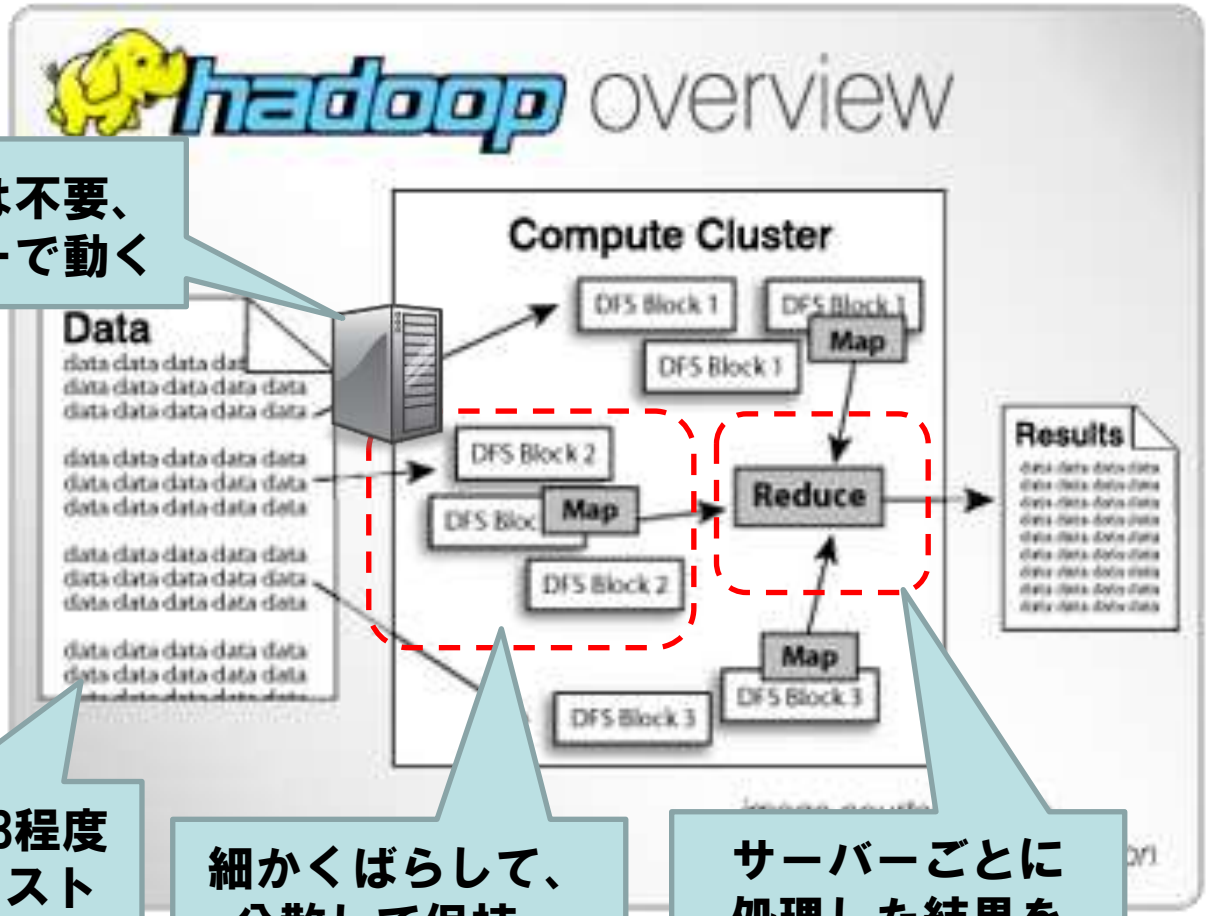
Hadoop とは？

- Hadoop とは、分散ファイルシステム(HDFS)を使用した、高速な**バッチ処理基盤**です。

- 大量のデータを高速に**バッチ処理**できるという効果があります。
  - 多数のマシンに分散することで、高スループットを実現し処理時間を短縮

- MapReduce というフレームワークに沿って処理を記述する（プログラミングする）必要がありました。
  - Hive（SQL）
  - Pig（DSL）
  - Hbase（KVS、DWH）
  - Mahout（機械学習、レコメンド）

Hadoop とは？



特別なハードは不要、
普通のサーバーで動く

数十GBからTB程度
の巨大なテキスト
ファイル

細かくばらして、
分散して保持。
行単位で処理

サーバーごとに
処理した結果を
集計、結合。

出典：How jStart is leveraging distributed computing for business
http://www-01.ibm.com/software/ebusiness/jstart/hadoop/

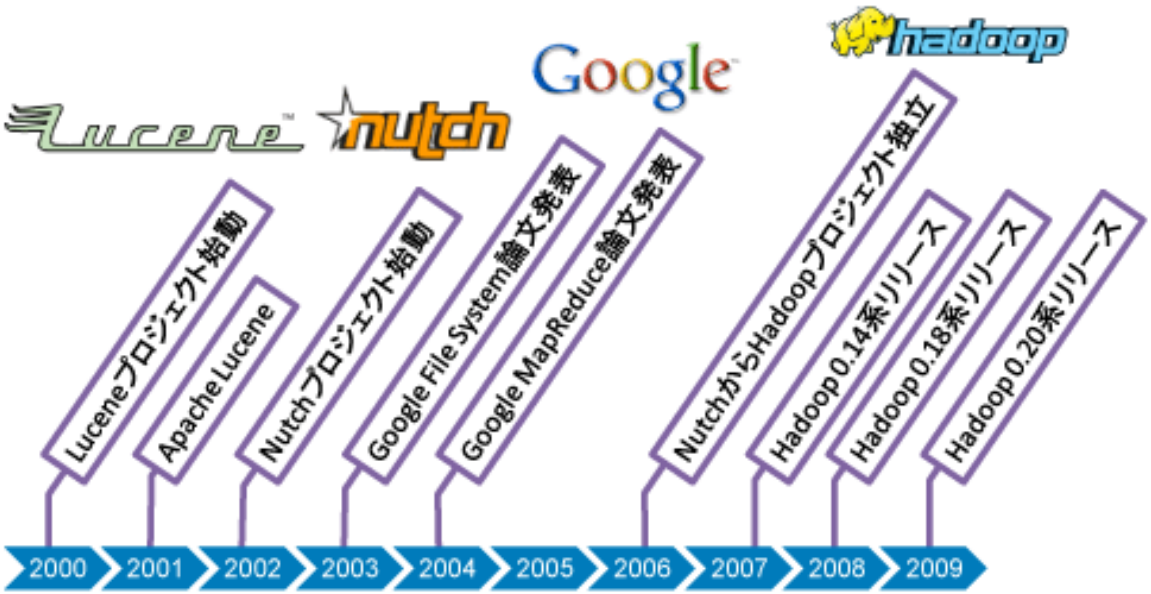
Hadoop Conference Japan 2014

- 7月8日（火）
  @ベルサール汐留
- 参加者: 1299人
- 私服組（Web系企業）が多いが、メッセージボードを見るとSIerも多数参加。
- YARN（MRv2）、Sparkおよび関連アプリ、機械学習のセッションが中心。

最近の Hadoop の動向

- 開発開始からおよそ１０年。安定、**普及期へ**。
  - Web系だけでなく、一般企業も検討を始めている

- MapReduce からの脱却、パフォーマンスが劇的に向上。**リアルタイム性**をうたいはじめる。
  - 2013年10月 - Hadoop 2 リリース（MRv2 / YARN）
  - 2014年 5月 - Apache Spark リリース



出典：[第2回]Hadoopの生い立ち
http://itpro.nikkeibp.co.jp/article/COLUMN/20120215/381721/

大規模リアルタイムデータ処理

Data Source / Connector / Crawler	Queuing	Processing	Store
Apache Flume	Apache Kafka	Hadoop MapReduce	Hadoop HDFS
fluentd	RabbitMQ	Apache Spark	NoSQL
	Apache Storm		Elasticsearch

クラウド

Amazon SQS	Amazon Kinesis	Amazon S3
	Amazon EMR	Amazon Redshift
		Amazon Dynamo
		Treasure Data

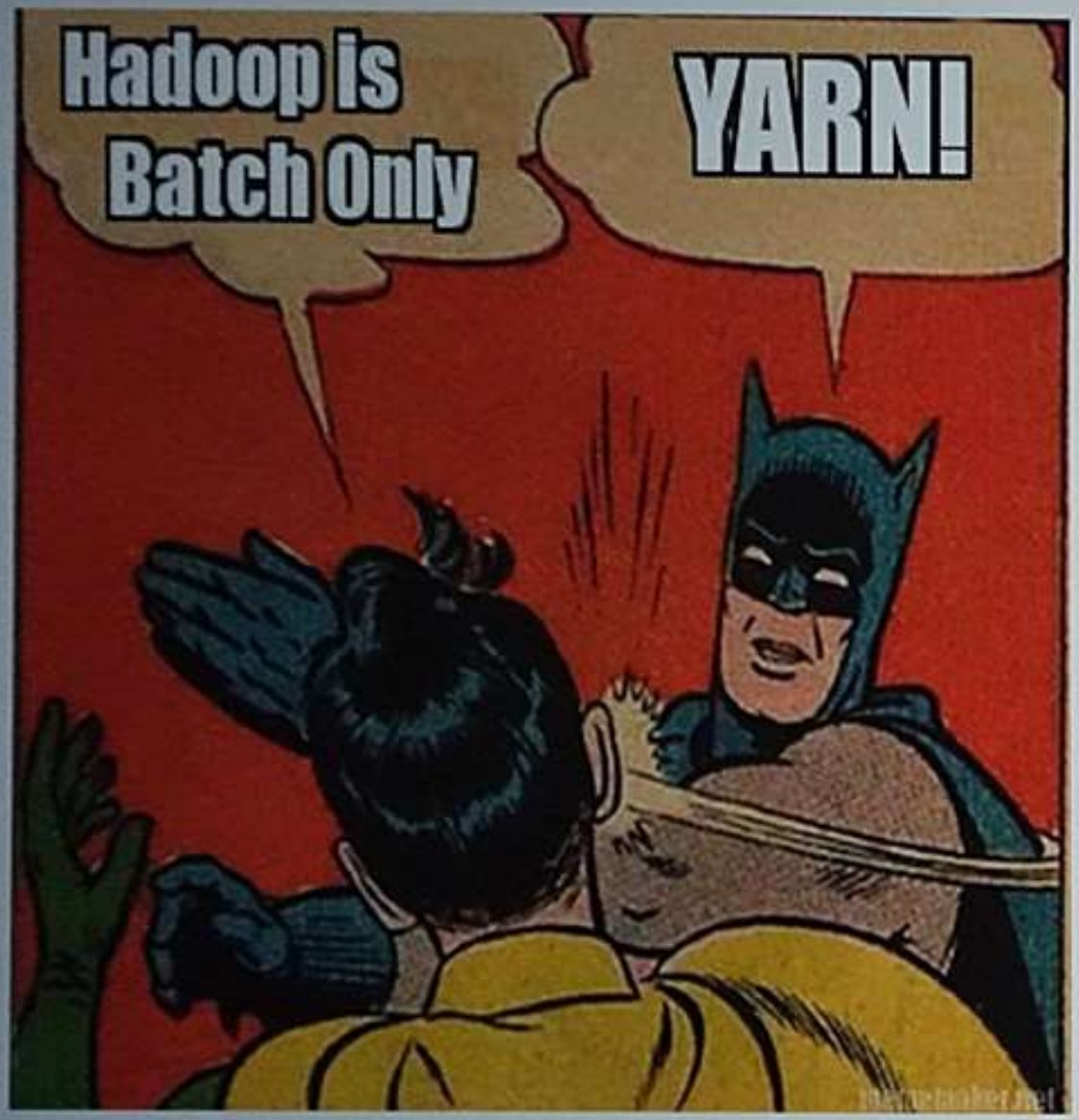
大規模リアルタイムデータ処理

速いは正義！

- 大量のデータに、繰り返し集計や統計処理を行い、業務上、優位なデータを生成する

- Hadoop の MapReduce を使ったソリューションがバッチ処理であった点に対し、処理を分散して高速化する優位性を、リアルタイムに 実現する

- Applications
  - 売上やアクセス数の一時間あたりの速報値（KPI ダッシュボード）
  - センサーデータ分析
  - リアルタイム・レコメンド （生放送のコメントに連動したレコメンド）



YARN Has Fundamentally Changed Hadoop

It Enables…

- **More Workloads**
  From batch to interactive & real-time
- **More Data**
  Multiple data sets of varying types and structures
- **More Value**
  Hosting multiple business cases in a single Hadoop cluster

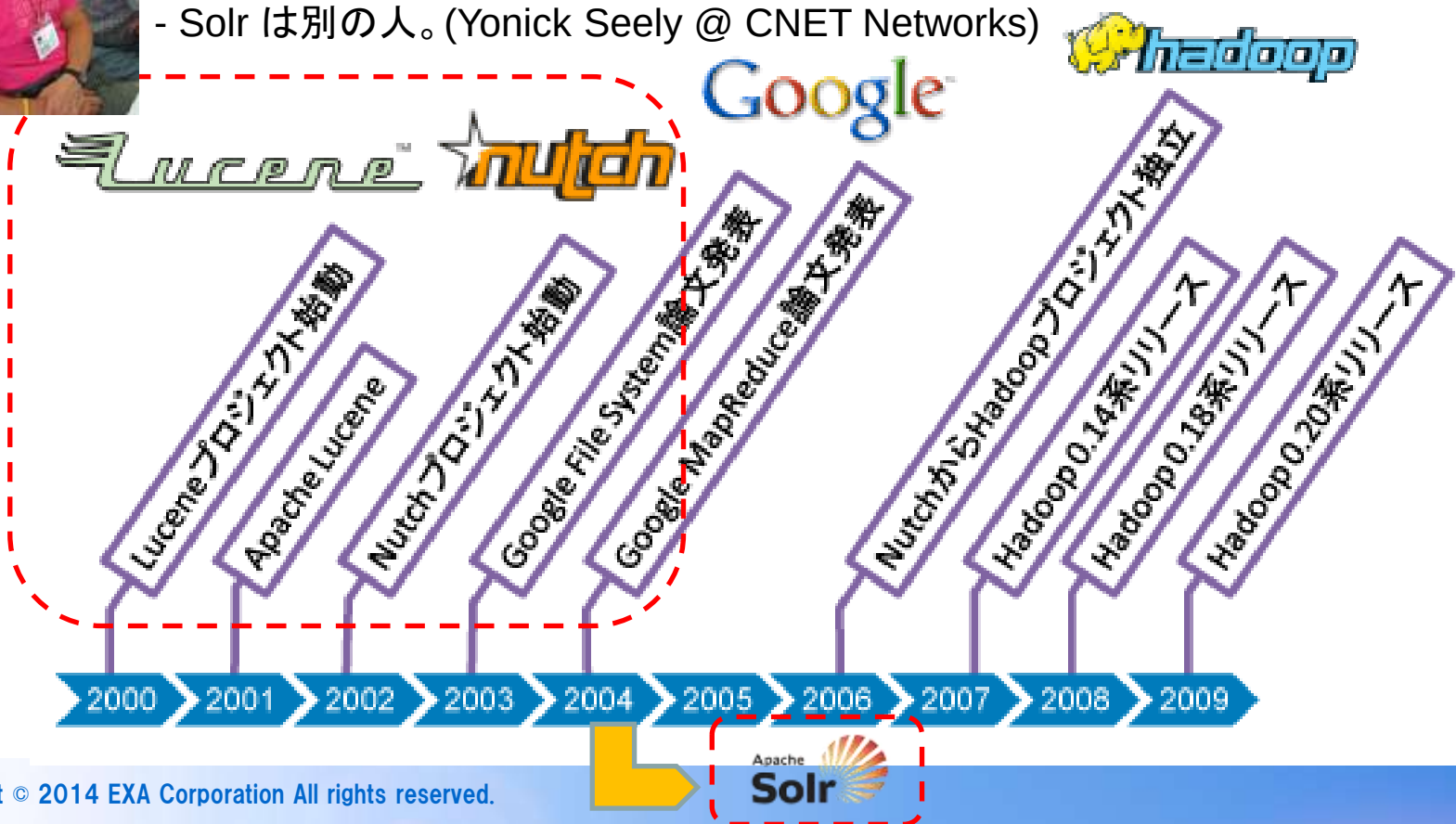
Hortonworks

出典：Hadoop Summit 2014でホートンワークスが示したスライドより
http://itpro.nikkeibp.co.jp/article/COLUMN/20140619/565326/

余談ですが・・・

- 検索ライブラリ Apache Lucene と生みの親が同じ
- 検索エンジンも、Hadoopと同じく大量のデータを扱う

Dug Cutting
- Lucene
- Nutch
- Hadoop
- Solr は別の人。(Yonick Seely @ CNET Networks)



レコメンドとは

Amazon の例

出典：Amazon.co.jp

Amazon の例

出典：Amazon.co.jp

レコメンドの仕組み

- 利用者の動線に基づいた共起分析
  - 次に何をおすすめしますか？

			
安藤さん	○	○	
佐藤さん			○
	○		

オムニチャネル時代のレコメンド

- さまざまなチャネルでの利用履歴を元にしたレコメンド
  - オンラインでの購入履歴を元に、店舗でおすすめ（タブレットと合わせてコンシェルジュ）
- パーソナライゼーション
- 速さ（早さ）は正義！
  - 今日、今、おすすめすべき商品は何か？

次世代オムニチャネル・レコメンド

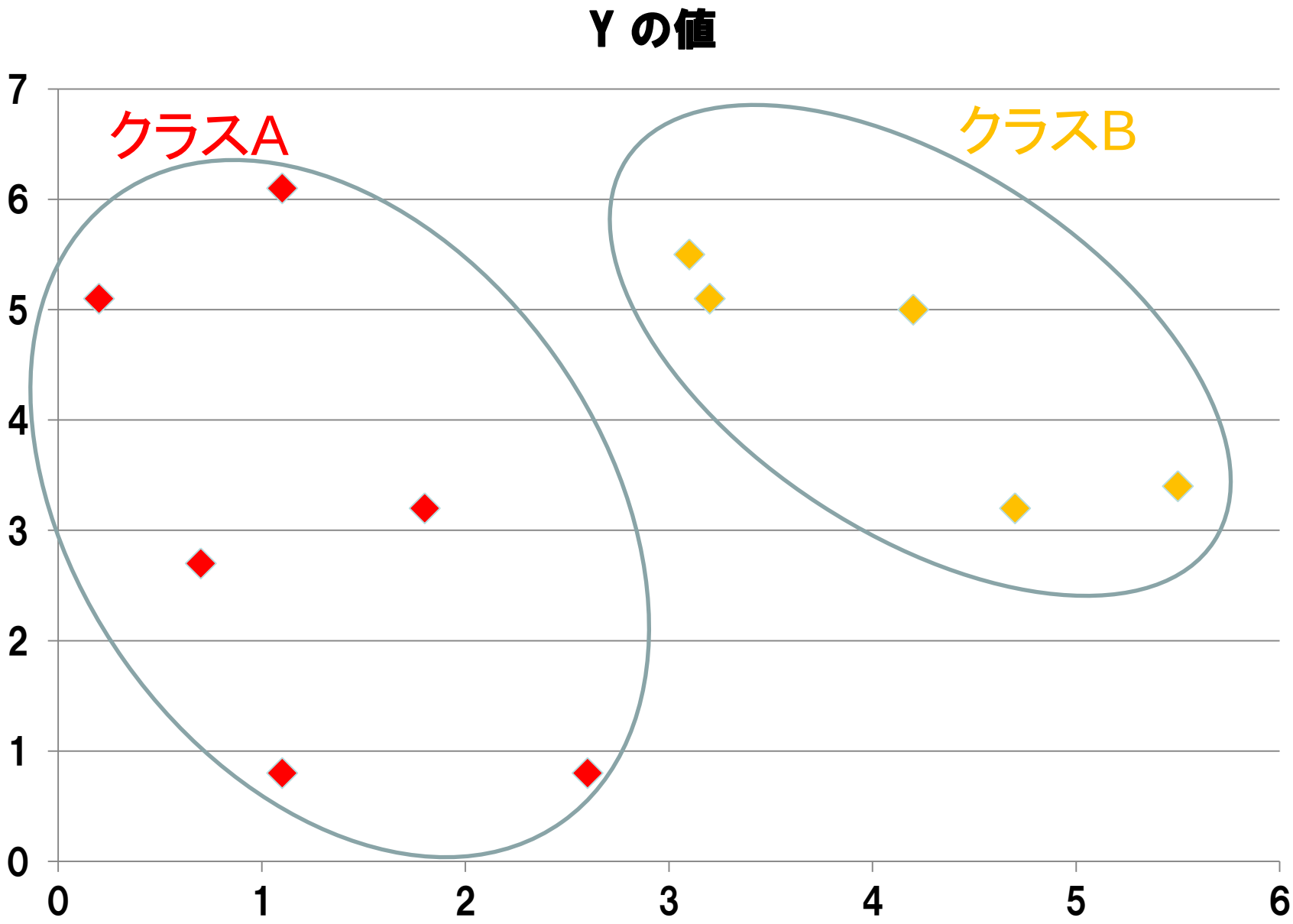
機械学習への応用

機械学習とは？

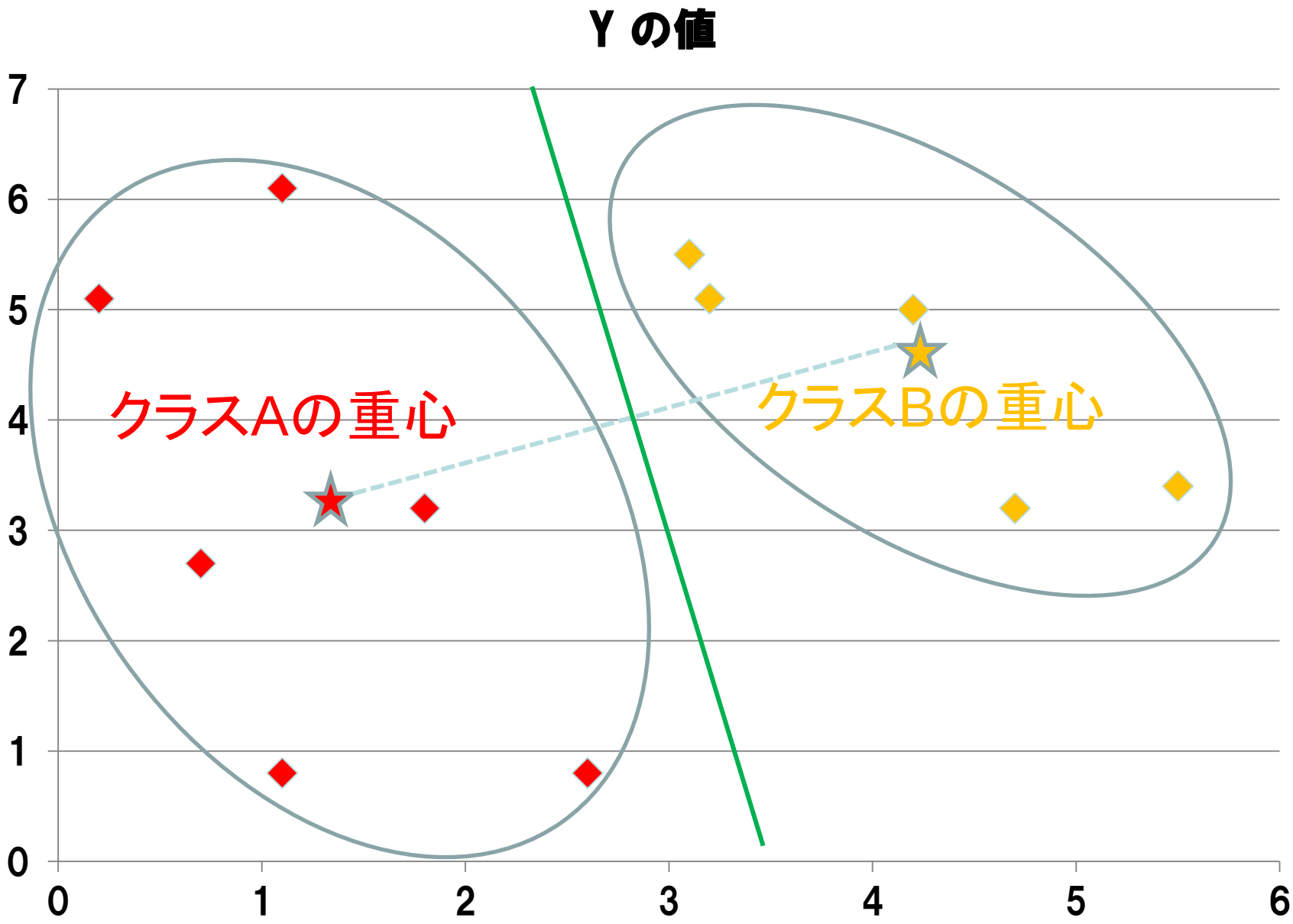
- 機械学習とは「明示的にプログラミングすることなく、コンピュータに行動させるようにする」
  - コンピューターでアルゴリズムを構築し、学習データを読み込ませることで、自動的に今あるデータを分類＆まだ見ぬデータを予測できるようにする
  - ヒトの情報処理能力を超えて、複雑なデータを分類＆予測できるようにする

レコメンドも機械学習の一種
ノウハウが貯まれば、Hadoopと合わせて横展開が可能

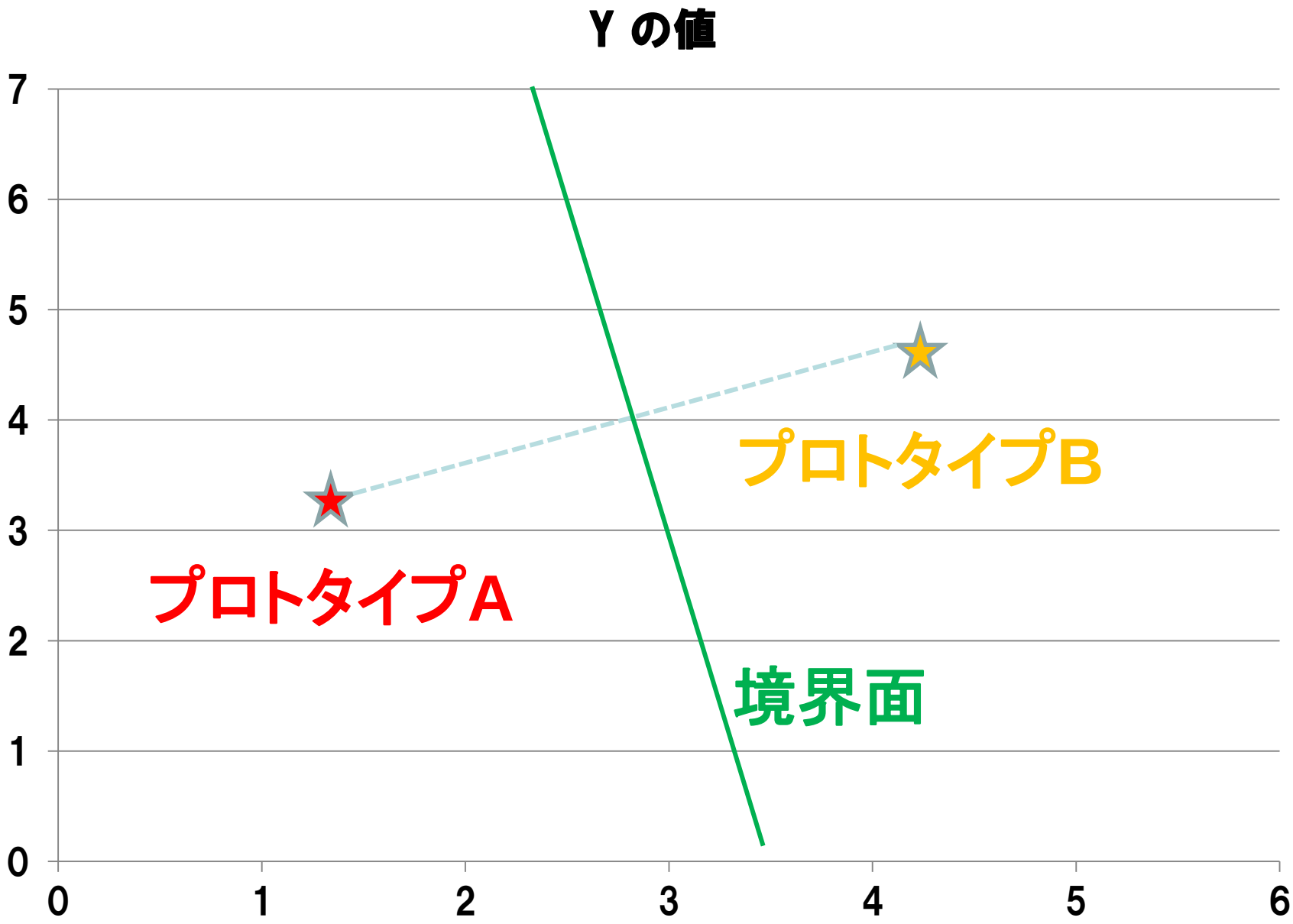
簡単な機械学習の例



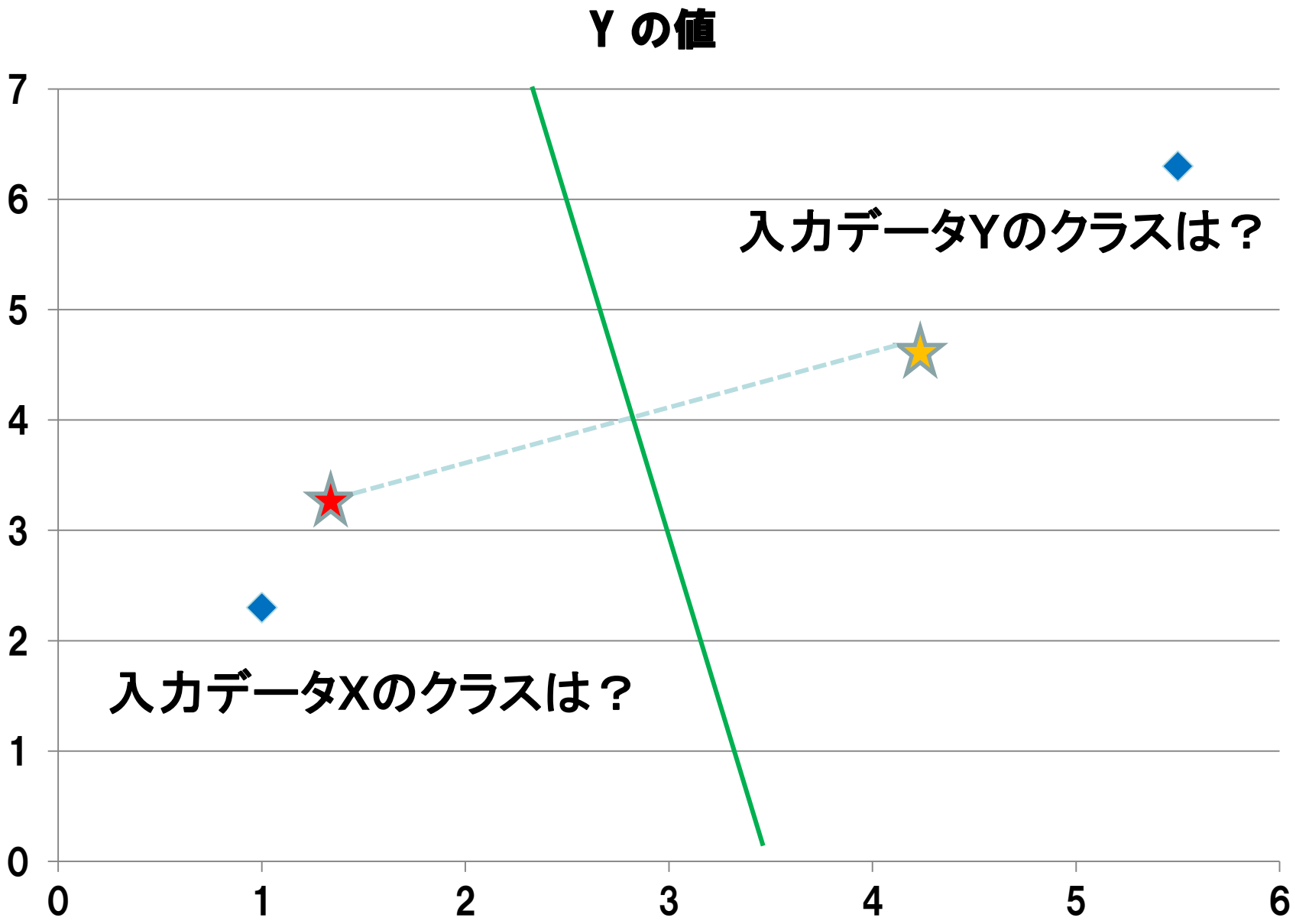
簡単な機械学習の例



簡単な機械学習の例

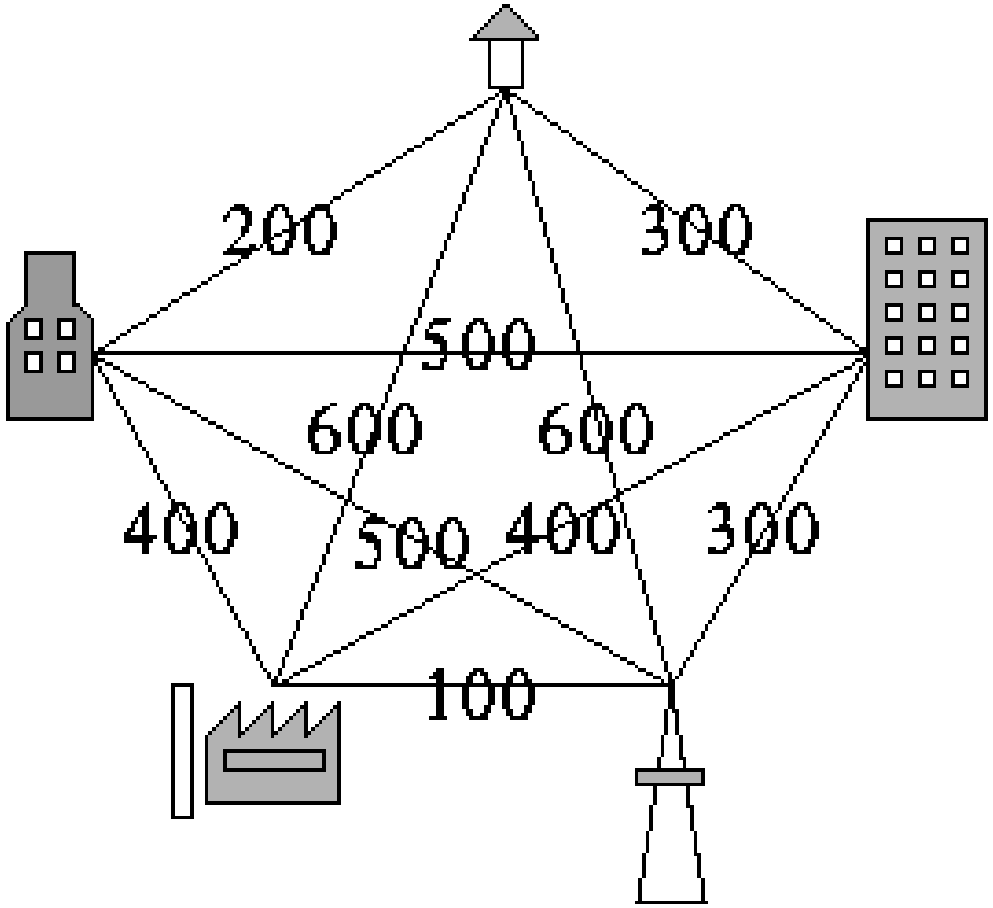


簡単な機械学習の例



応用例：巡回セールスマン問題

- 膨大な組み合わせの中から最適な解を見つけ出す。
  - 全ての街を訪れること。
  - 同じルートは２度通れない。
  - 最も安いルートを通ること。



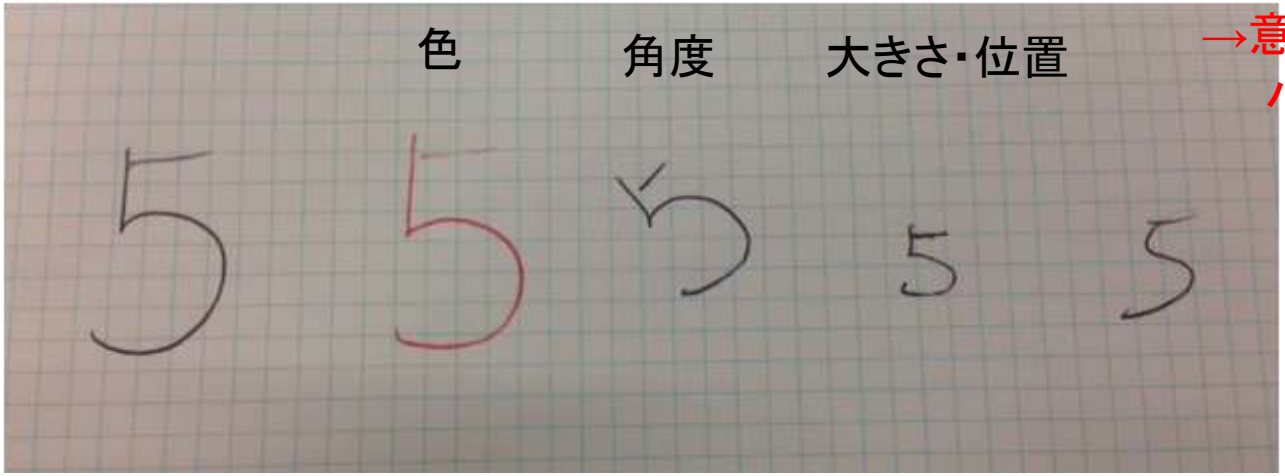
わずか30箇所で、10の30乗！！！

全ての組み合わせ数
　= n! / 2n 通り
5箇所 = 12通り
10箇所 = 181440
30箇所 = 10の30乗

高度な業務領域の知識が必須

- 分析したい事象の特徴を示すパラメータが必要
  - 手書き文字の OCR の例

- 故障に関連するパラメータは？
  - ワット数、電流、電圧、長さや重さ、気温、天気、湿度、温度、使用時間、利用者、使用方法・・・
  - 手法や繰り返し分析することで、関連するパラメータを推定することが可能だが“あたり”をつける必要がある。



ソリューションへの応用

- 異常検知
  - 機器から出力されるパラメータを元に、異常な挙動を検知し、事故や故障を未然に防止する。
  - 例）クレジットカードの不正利用

- 予防保守
  - いつ故障しそうか、過去のデータから推測する。
  - 部品の交換の頻度や、生産ラインを休止する頻度を下げることができる。

まとめ

まとめ

・Hadoop は、安定化・高速化し、Web系から一般企業へ浸透し始めている

・オムニチャネル・コマースにあった、オムニチャネル・レコメンドを提案します。

・機械学習の適用には業務知識が必須エクサには実現するノウハウがあります。

Q&A

ご静聴ありがとうございました。