# 大規模データで手軽にデータマイニング
# Big Data Module 紹介

数理システムユーザーコンファレンス 2013
（株）NTTデータ数理システム データマイニング部
五十嵐 健太

Global IT Innovator NTT DATA Group

NTT DATA

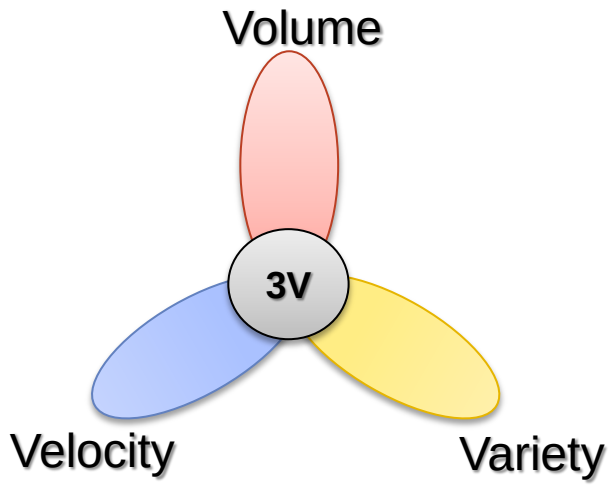
MATHEMATICAL SYSTEMS INC.

# ー ビッグデータのデータマイニング ー

# ビッグデータとは

## ビッグデータとは

- 従来型のシステムでは処理するのが困難な大きなデータ
- 3V
  - 巨大（Volume）
  - データが多様（Variety）
  - 速い蓄積速度（Velocity）




## バズワード？

大きさに対する感覚は，人／分野によって異なる

- カウントだけならテラバイトオーダーのデータでも対して難しくない
- 予測モデルの作成は数10ギガバイトのデータでも難しいかも
- ハードウェア性能、分析内容に大きく依存

単純にデータサイズの大小だけでは語れない

# データマイニングとは

データマイニングとは一言で言うと

「大量のデータから役に立つ情報を抽出すること」

単純な集計だけではなく、機械学習的な複雑な分析アルゴリズムを使用して分析が行われる

典型的なデータマイニング手法

- アソシエーション分析（ルール抽出）
- 分類分析
- 数値予測
- クラスタリング

それぞれの手法に対して、様々な分析アルゴリズムが開発、研究されている

アルゴリズムによって使用できるデータサイズは異なる

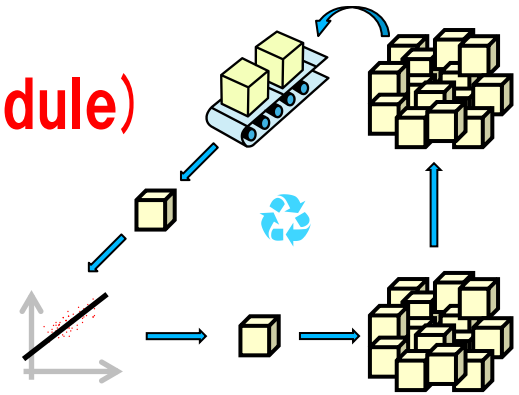
# ビッグデータのデータマイニング

従来のハードウェア、分析ツールでは分析が難しいビッグデータに対して分析を行うには…

- より高性能なマシンを使う

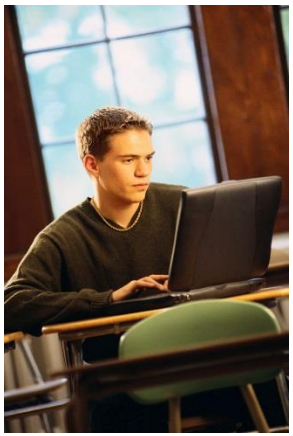
- 分散・並列処理
  - ➢一台のマシン上でのマルチコア並列処理
  - ➢複数マシン上での分散処理（Hadoopなど）

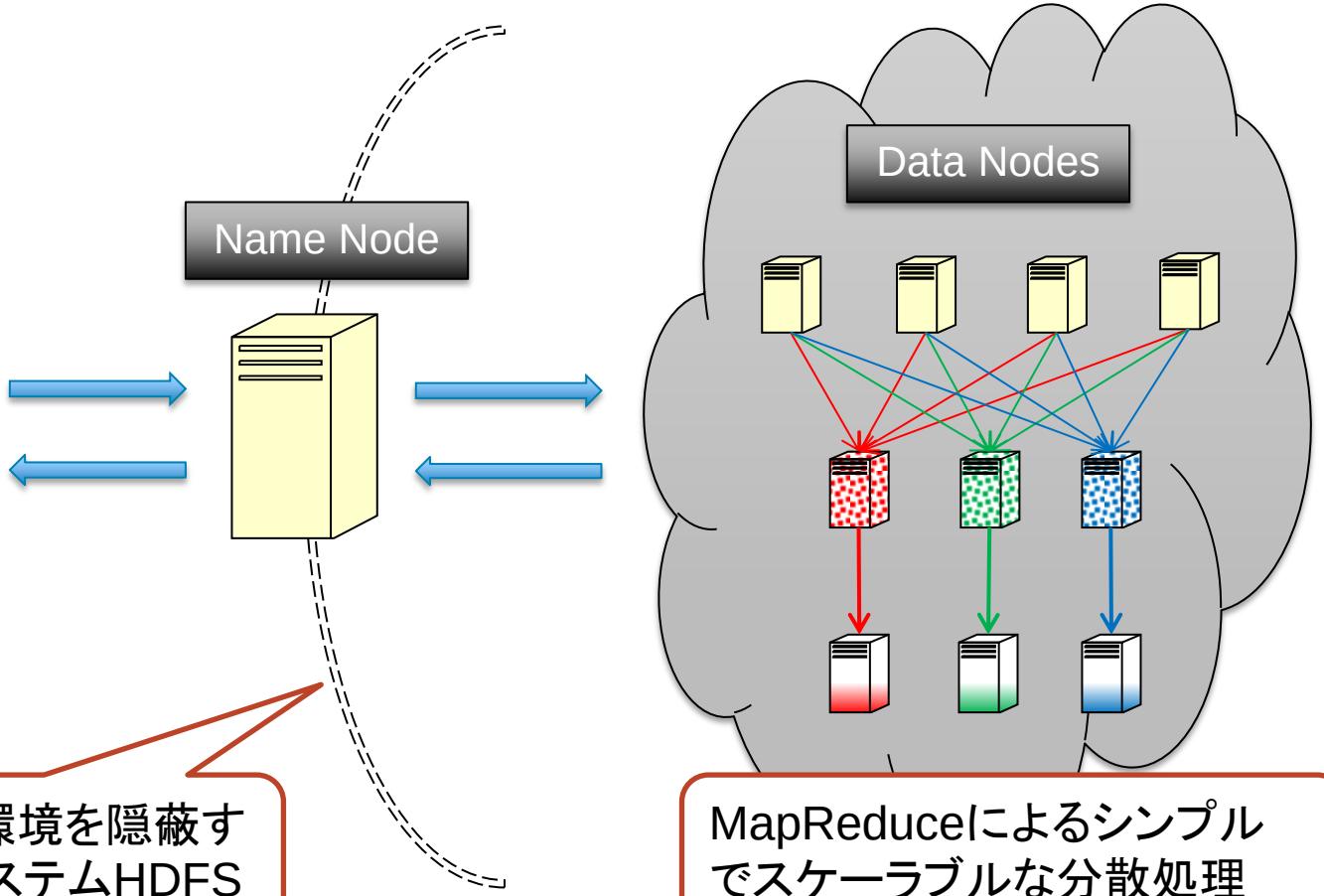
- 大規模データの処理を意識した先進的なアルゴリズムを使う（Big Data Module）
  - ➢データ圧縮技術
  - ➢オンラインアルゴリズム



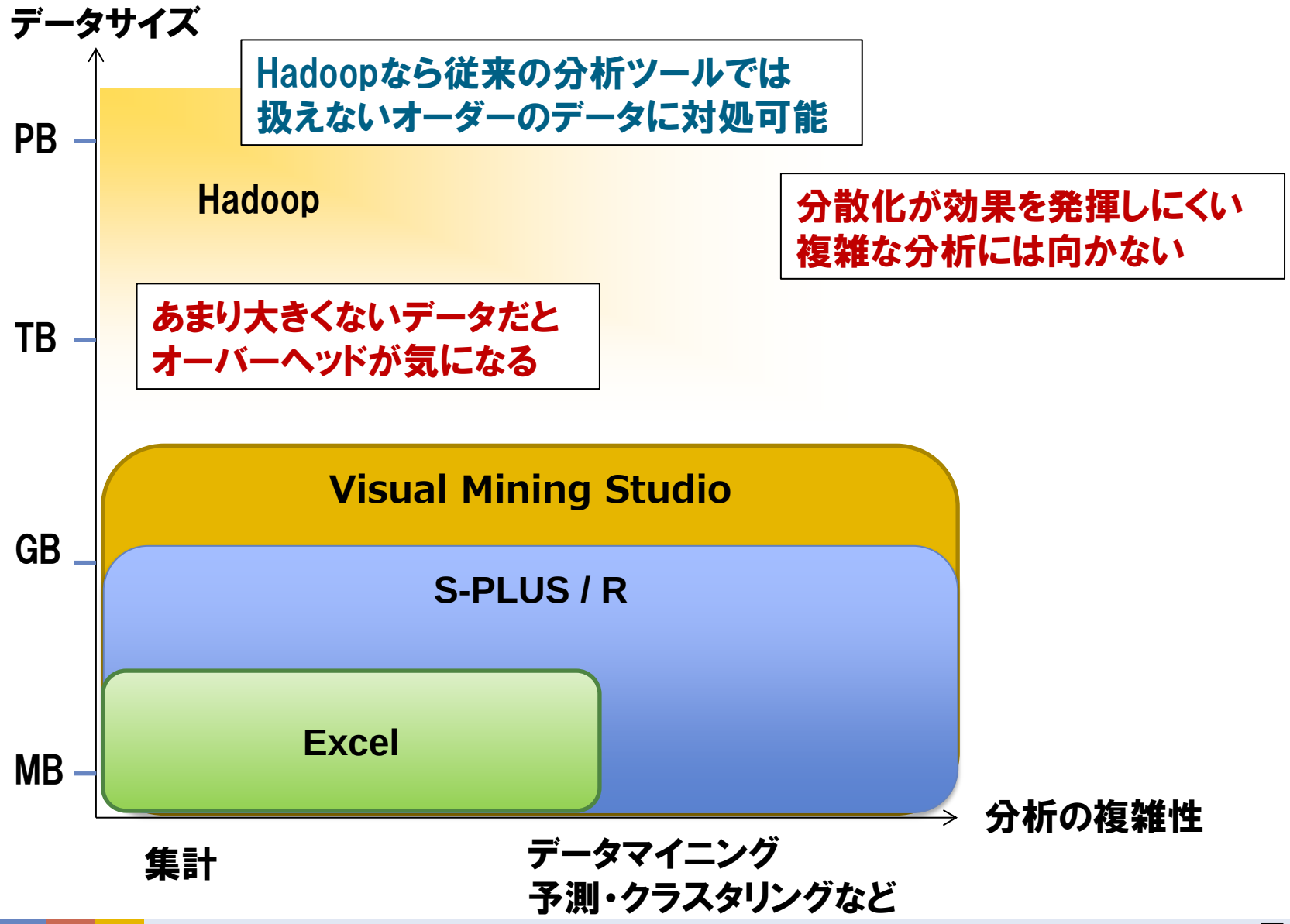
# Hadoopとは

**多数のマシンによる大規模分散バッチ処理フレームワーク**
**超大規模データ（TB～）を扱うジョブの処理に特化**



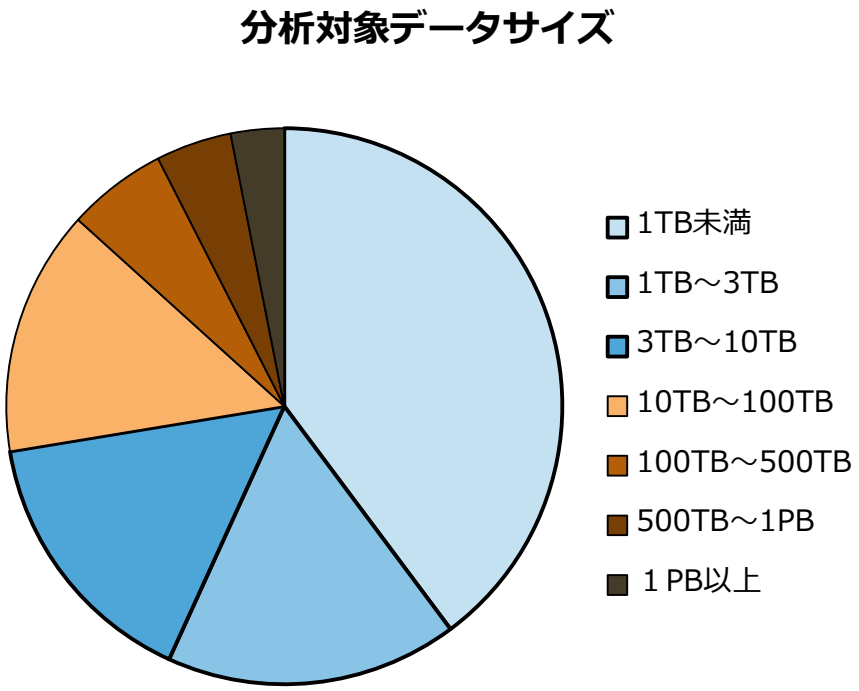


# 分析ツールとデータサイズ



# 分析データサイズ

## 日本企業の大部分の解析対象データは10TB以下




野村総合研究所「企業情報システムとITキーワード調査」
(2011年8月～10月、調査運営：日経BPコンサルティング)より

## 実際に1度の分析に適用するデータはさらに小さい

# 分析データサイズ

## Big Data 処理基盤で処理されているデータのサイズは実は大きくない？

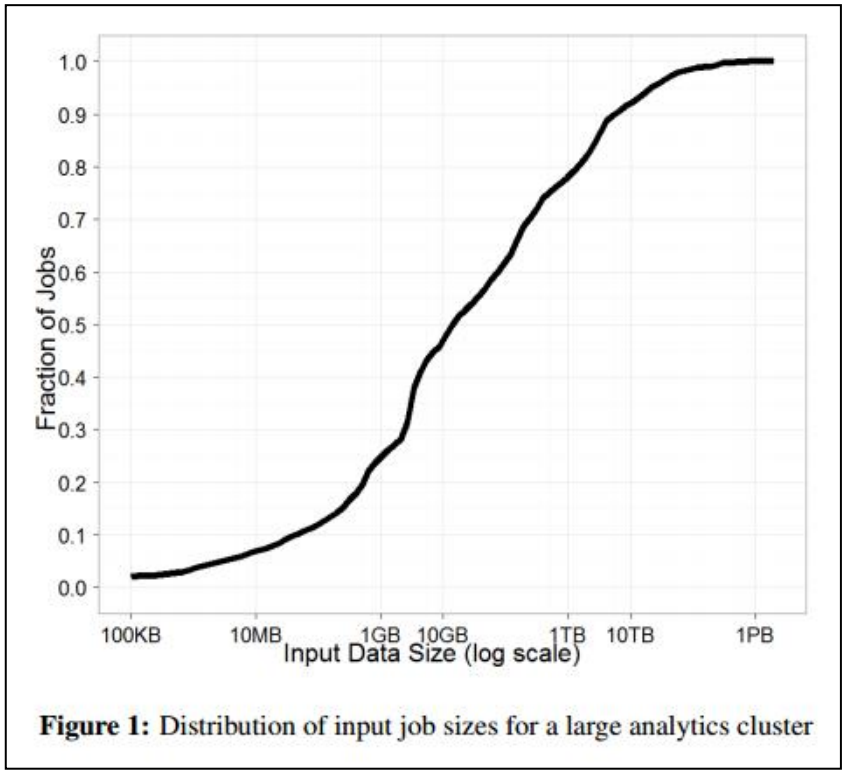
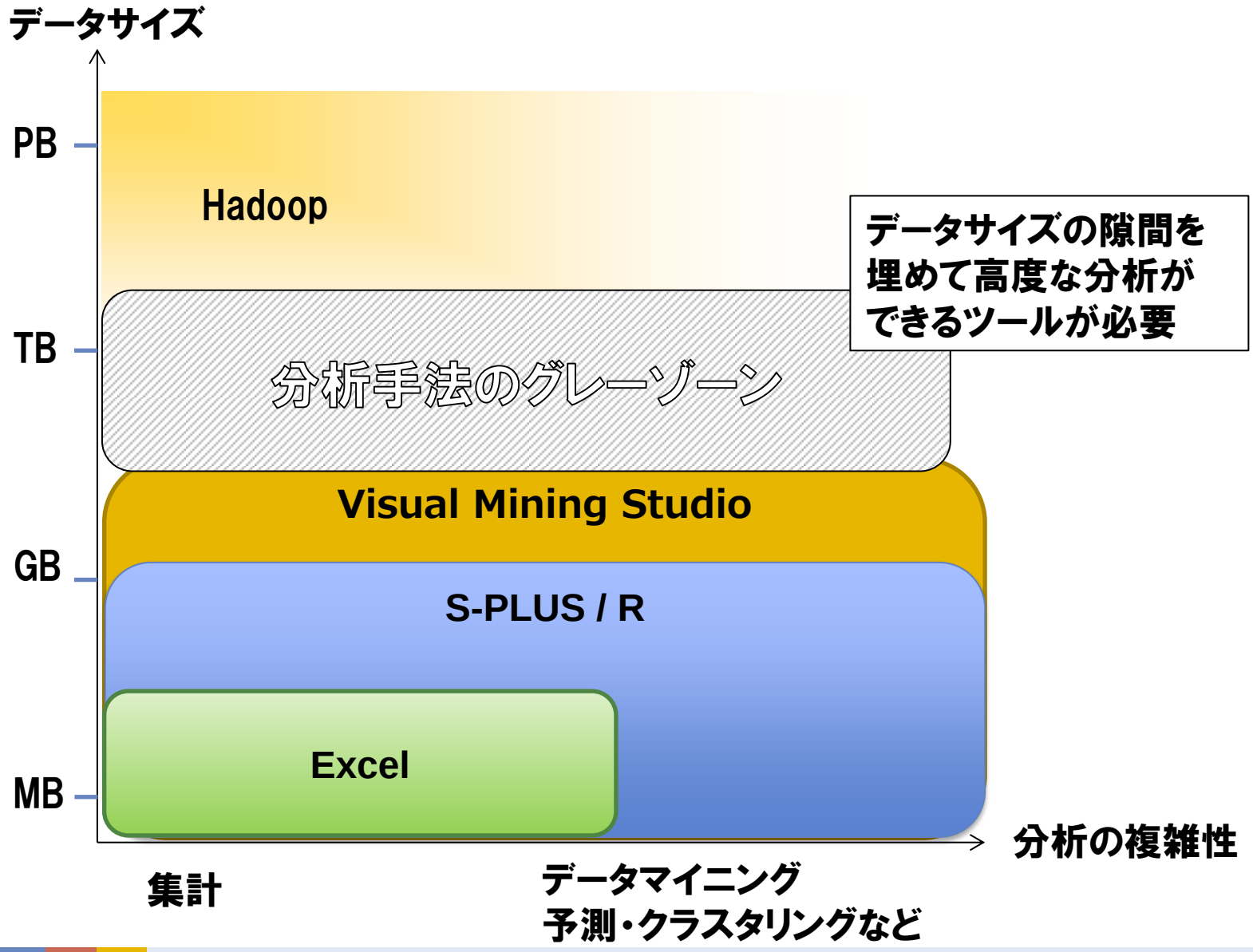



**Figure 1:** Distribution of input job sizes for a large analytics cluster

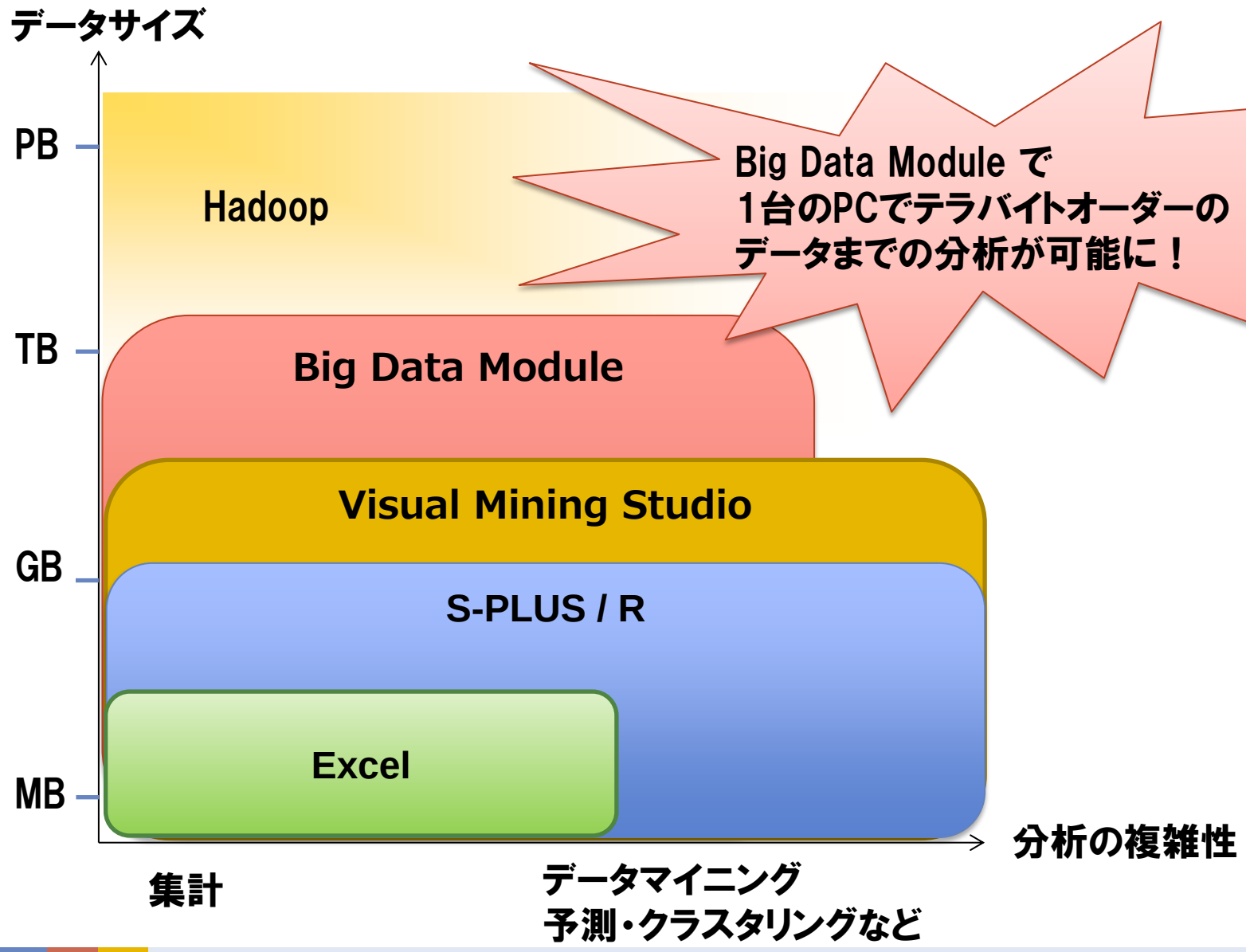
**1GBから1TBが多い**

Microsoftが2011年のある月に解析用コンピュータクラスタ上で実行した
解析ジョブの入力データサイズ分布
(Appuswamy et al., “Nobody ever got fired for buying a cluster”, 2013 から引用)

# 分析ツールとデータサイズ



# 分析ツールとデータサイズ



# ー Big Data Module 紹介 ー

# Big Data Moduleとは

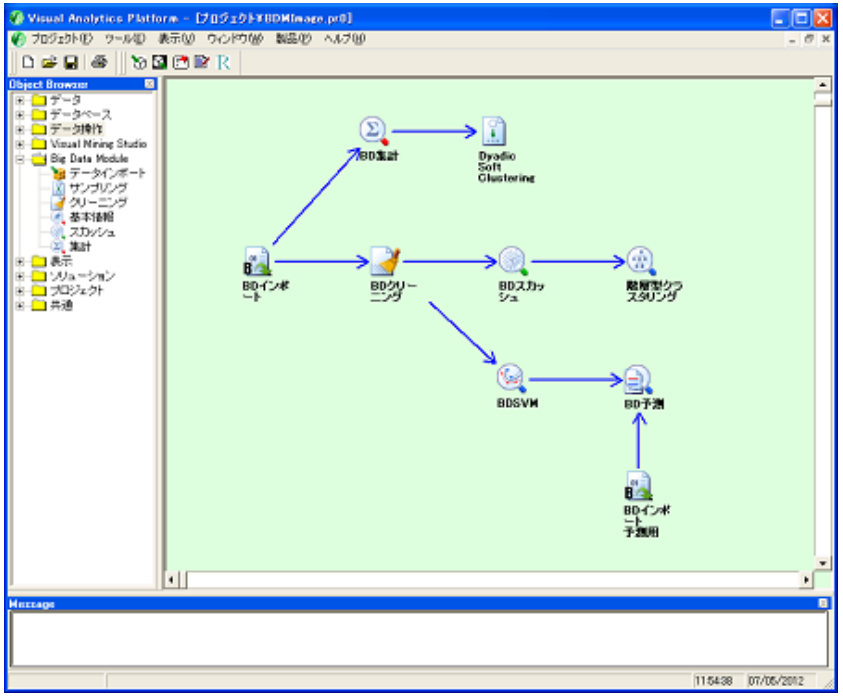
**数理システム自社開発**

**簡単な操作で高度なビッグデータ分析が可能**

**特殊な分析専用マシンは不要**

**市販のマシンを1台用意すれば**

**それだけで分析が実行可能**



# 数値予測

売上予測

株価予測

電力需要予測

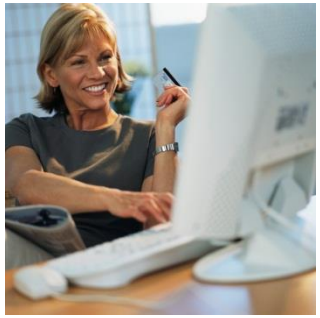
予測精度を高めるには…
- データ数を増やす
- 説明変数を増やす

→

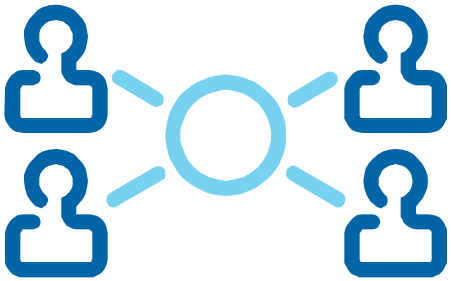
計算時間の
爆発的な増加で
計算不可能

Big Data Moduleなら…
SGDを使った**オンライン線形回帰**でビッグデータでも予測可能
- データ数の線形オーダーの計算時間→**超高速**
- データ数に依存しないメモリ使用量→**超省メモリ**

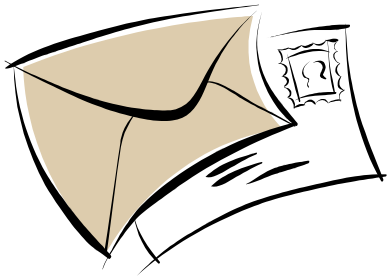
# レコメンデーション



ECサイト



SNS



ダイレクトメール

膨大なログの中から
おすすめのアイテムを
発見する

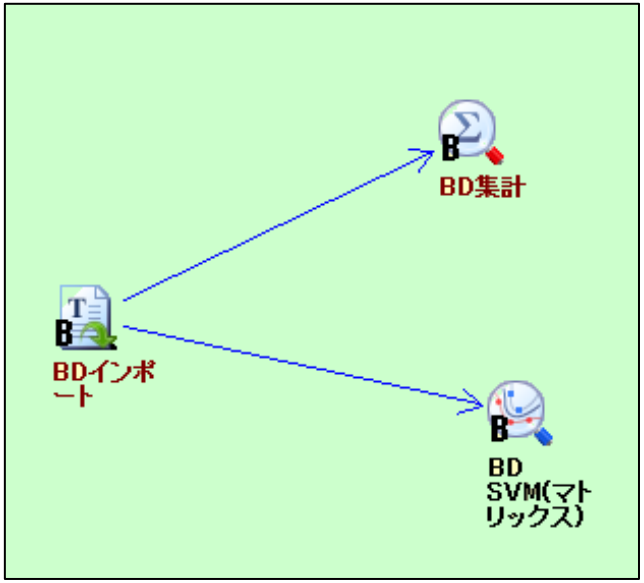
時間をかけて高精度を目指す？
単純なモデルでリアルタイム性？

Big Data Moduleなら…
**オンライン行列分解**で協調フィルタリングによる
レコメンデーション
**高速**かつ**高精度**なレコメンデーション

# Big Data Module の主な特長 (1)

## GUIベースの簡単操作

アイコンを矢印でつなぐだけの
簡単操作で、高度な分析機能を
実行可能




## オンラインアルゴリズム

ビッグデータに適した分析アルゴリズムを
搭載

## 並列処理

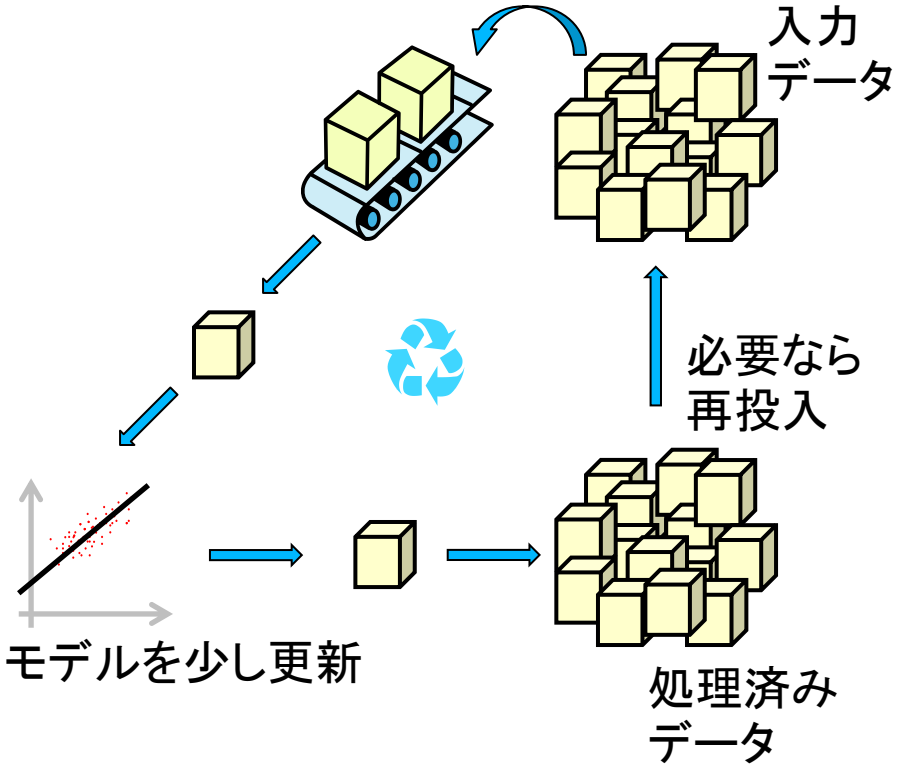
並列処理をアイコンベースで簡単に実行可能
マルチコアをフルに生かして分析を実行

# オンラインアルゴリズム

データを一つずつ読み込み、モデルを逐次更新

データをためず、**必要最低限のメモリ使用量**

計算時間は処理データ数の**線形オーダー**

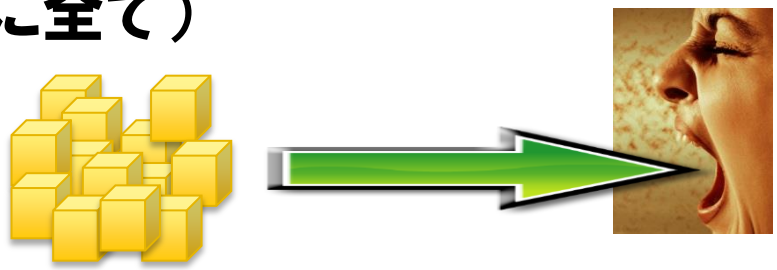



※「オンライン」はチューリングマシンの理論に
由来した用語で、「ネットワークにつないで」
といったような意味ではありません。

# バッチアルゴリズム/オンラインアルゴリズム

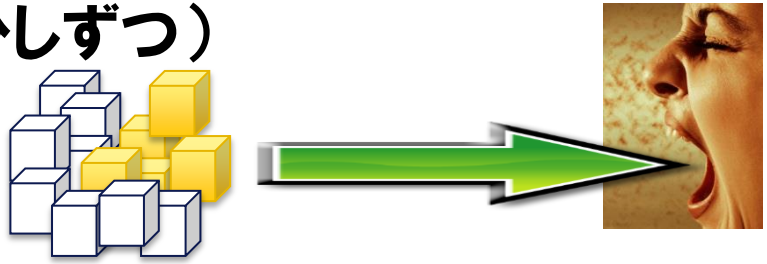
MATHEMATICAL SYSTEMS INC.

## バッチ(一度に全て)

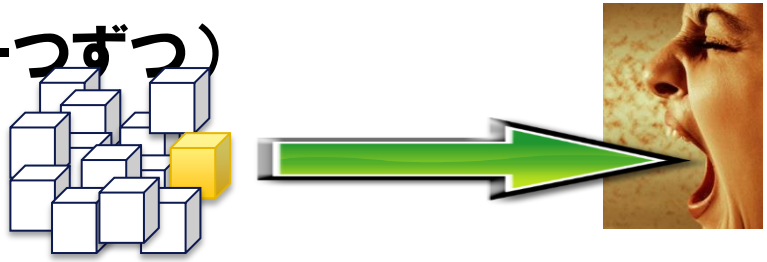


従来のアルゴリズムの多くはバッチ

## ミニバッチ(少しずつ)



## オンライン(一つずつ)



小回りが利き、多くの場合、超高速かつ超省メモリ

# Big Data Module の主な特長 (2)

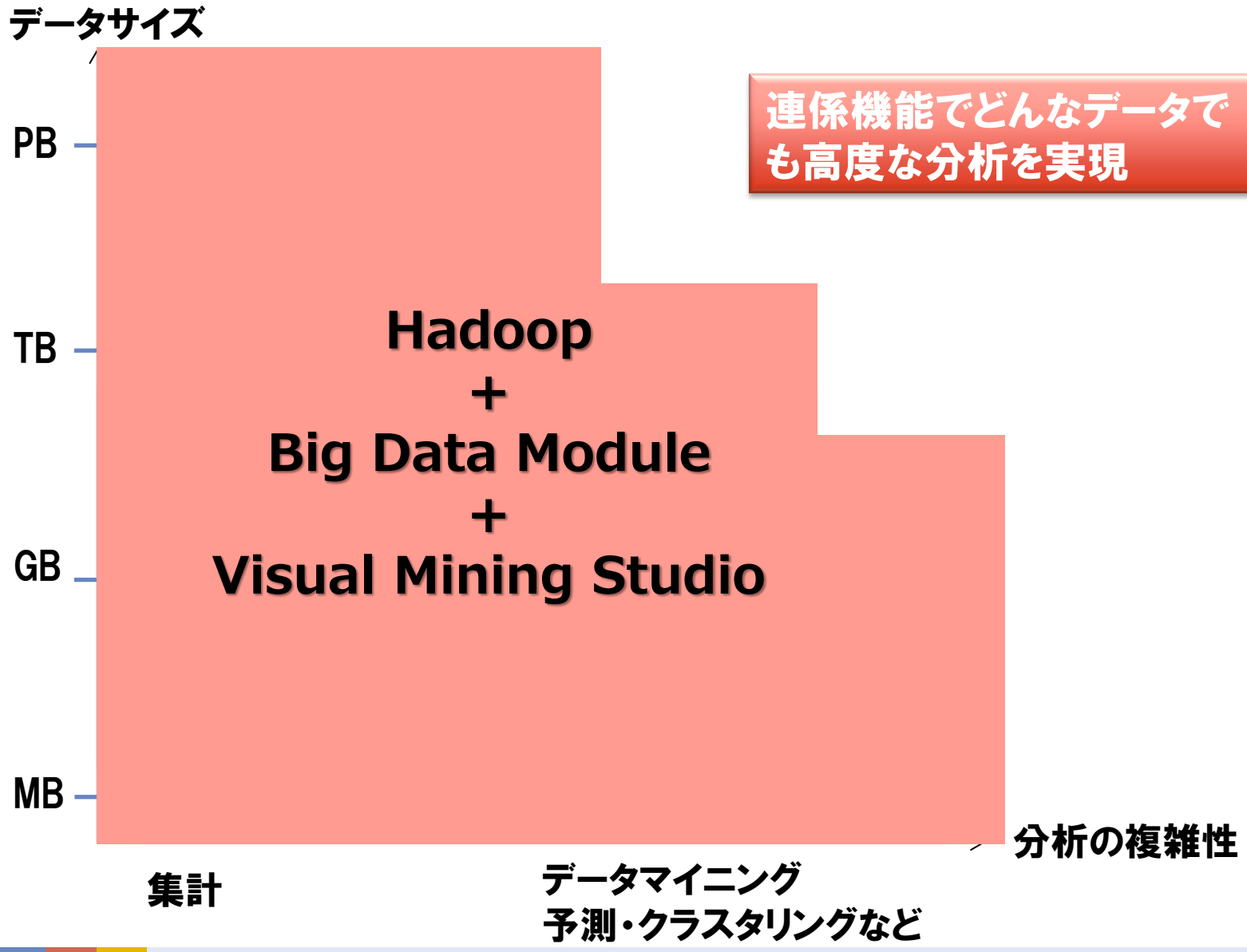
## Visual Mining Studio との連携

データマイニングツール Visual Mining Studioとシームレス連係
Big Data Module でデータの前処理、圧縮などを行った結果を
Visual Mining Studio の多彩な分析機能を使用して分析


## Hadoop との連携

Hadoopと連携することで、1台のマシンでは取り扱えないほどの
超大規模データの処理が可能

# Hadoop連係、Visual Mining Studio 連係




**連係機能でどんなデータでも高度な分析を実現**

# Big Data Module の機能一覧

## 前処理・データ加工

- データインポート
- クリーニング
- サンプリング
- スカッシング
- コアセット抽出
- データハンドリング

## 統計量

- 基本情報
- 集計

## 分析

- オンラインSVM
- オンライン線形回帰
- オンラインロジスティック回帰
- 予測
- 検証
- オンラインk-means
- オンライン行列分解
- レコメンデーション

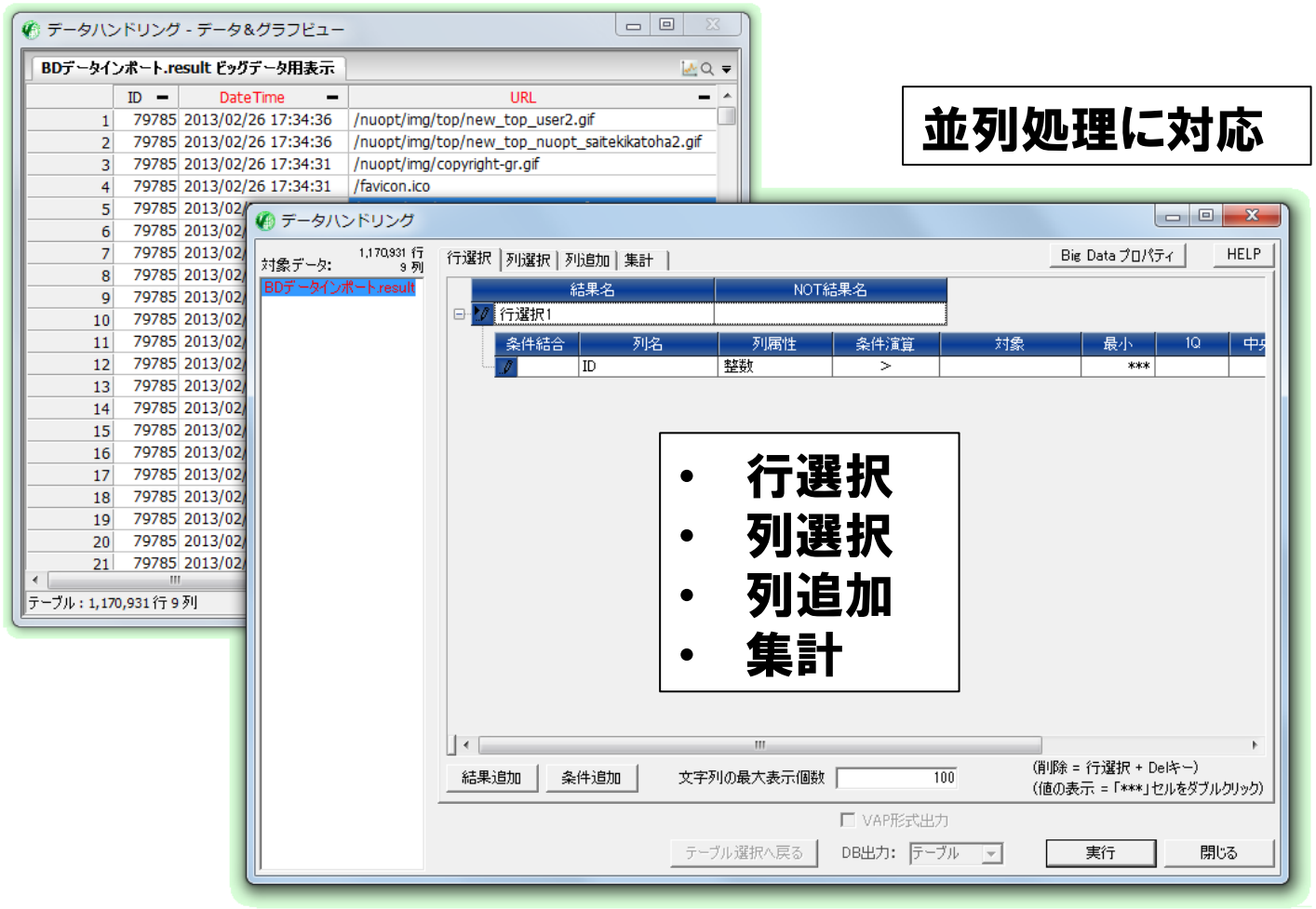
## その他

- データ&グラフビュー
- スクリプト
- バッチ処理

※ 赤字はバージョン1.2の新機能

# 新機能 データハンドリング

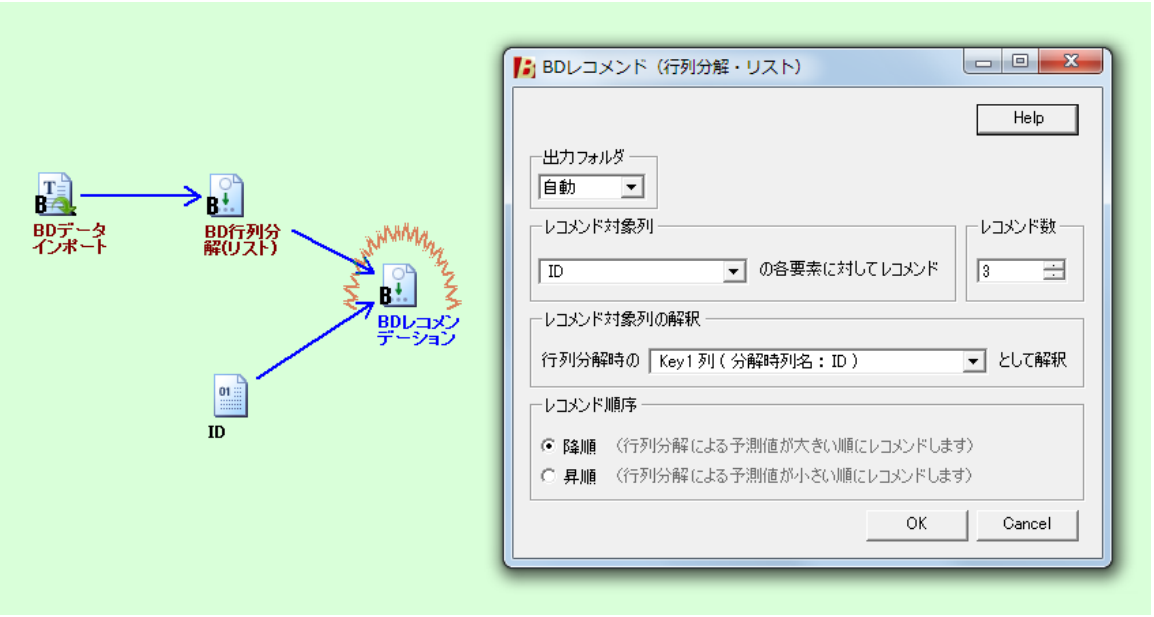
## データハンドリングアイコンがBig Data Module に対応
## GUI上で結果を見ながら簡単データ操作



# 新機能 レコメンデーション

## レコメンデーションアイコンを追加

- 行列分解と組み合わせて、レコメンデーションを行う
- 類似の購買傾向のあるユーザーの情報に基づいておすすめアイテムを抽出する
- 協調フィルタリングによるレコメンデーション



# テスト利用制度もございます
# お気軽にご相談ください


＜お問合わせ先＞
(株) NTTデータ数理システム Big Data Module 担当
〒160-0022
東京都新宿区信濃町35番地 信濃町煉瓦館1階
bigdata-info@msi.co.jp　　http://www.msi.co.jp
Tel :03-3358-6681 [ 営業部直通 ]
Fax :03-3358-1727

変える力を、ともに生み出す。
NTT DATA グループ
NTT DATA