Indexer Bulletによるビッグデータ解析

IIJ Techweek2013

2013/11/19

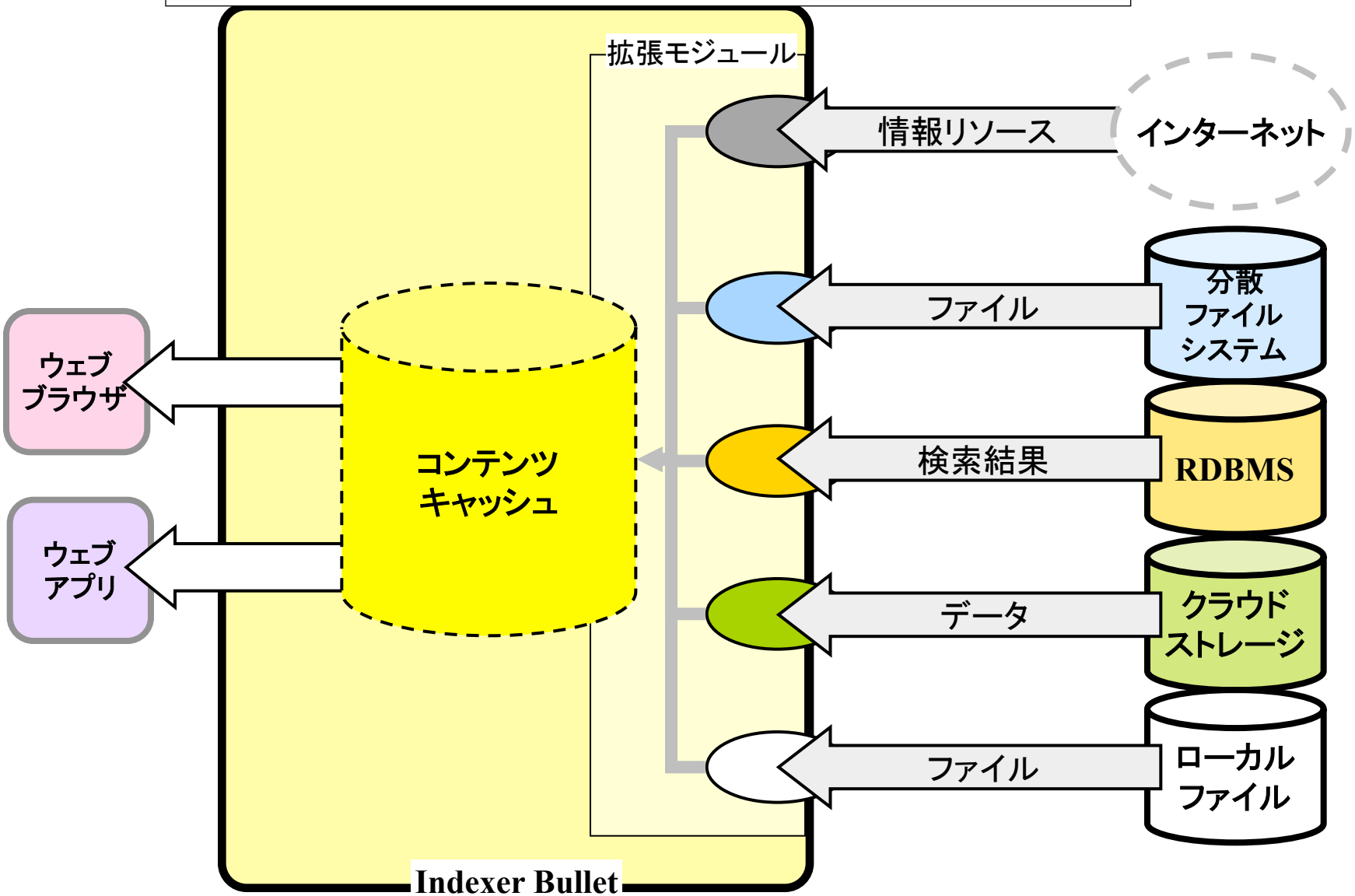
藤田昭人
株式会社IIJイノベーションインスティテュート

はじめに

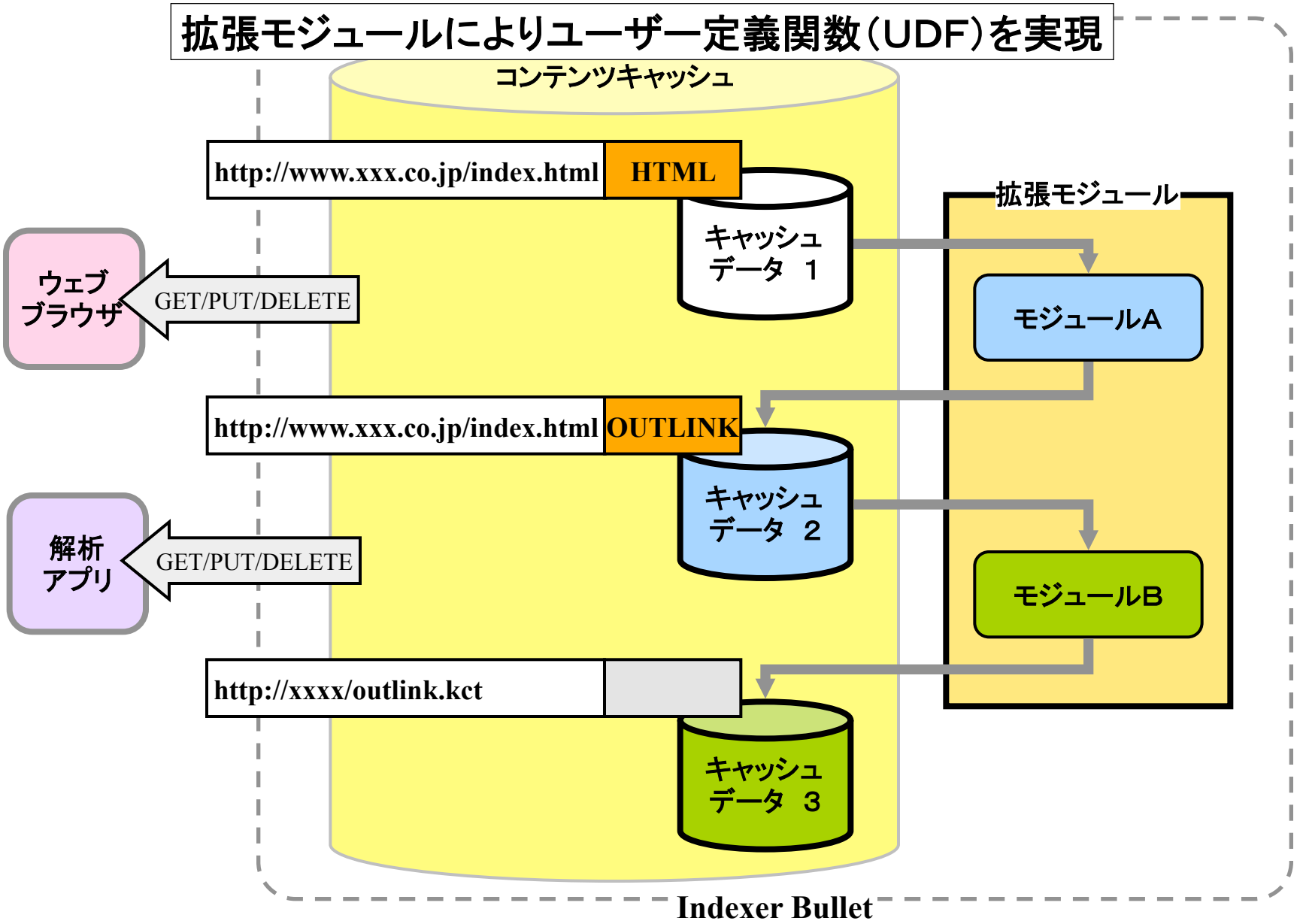
- Indexer Bullet（iBullet）
  - 一般的なビッグデータ解析プラットフォームの実現を目指す
  - インターネットからの情報リソースの取得
    - 情報リソースの統一的な取得手段を提供
  - 各種解析アルゴリズムの利用
    - 解析アルゴリズムの統一的な利用手段を提供
    - 複数の解析アルゴリズムを組み合わせた解析環境
  - 各種クラウドインフラストラクチャの活用
    - 複数のストレージシステムからなるヘテロジニアスな環境
    - 分散処理基盤との連携

Indexer Bullet（1）

様々なストレージに対応するヘテロジニアスなシステム



Indexer Bullet（2）



Indexer Bullet（3）

- 実装の現状
  - ５月　　初期プロトタイプ（Wikipediaランキング向け）
  - １０月　　Wikipediaランキングシステムを iBullet ベースに移行
- 課題
  - 拡張モジュールの仕様
    - より一般性のあるプログラミングインターフェースへの移行
    - ノンブロッキング同期によるマルチスレッド化
      - CompaireAndSwap 命令を活用
      - マルチコアのメリットをより生かせる
  - キャッシュシステムに基づく設計の有意性
    - ビッグデータ解析作業の効率化に寄与する・・・はず
- ビッグデータ解析のプロセスを把握しなければならない

分かりやすいと評判のビッグデータ解説

Big data is like teenage sex:
everyone talks about it,
nobody really knows how to do it,
everyone thinks everyone else is doing it,
so every one claims they are doing it…

ビッグデータは10代のセックスに似てる:
みんなが話題にし、
しかし本当は誰も方法を知らず、
みんながやってると思いこみ、
だから自分もやっていると言い張る・・・

URL: https://twitter.com/replore/status/396636923311058944/photo/1

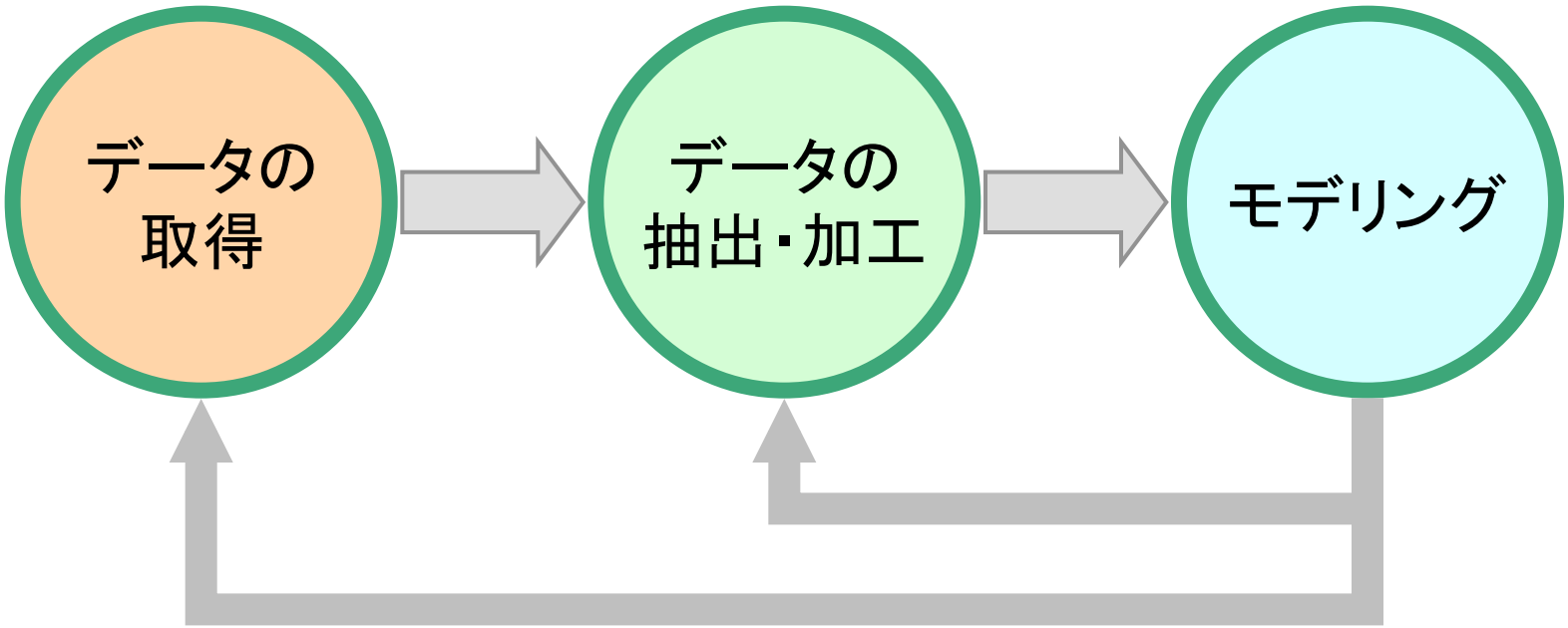
ビッグデータ解析のプロセス（1）

- 解析の結論を統計学的手法に頼るとすると・・・
  - データの関係性（相関など）の特性把握やそれに基づく予測など・・・
  - テキストデータはテキストマイニング的手法（頻度計数など）で数値化

- ビッグデータ解析は・・・
  - 大規模データを対象にしたデータ解析作業
    - 結論を得るまでに頻繁に試行錯誤が発生する
  - 『モデリング』は対話的プロセス
    - 既存のデータ解析ソフトウェアを利用する
  - 『データの抽出・加工』は対象データとモデリングの接続性が重要
    - 抽出・加工は解析目的に従属するが・・・
    - データ解析ソフトウェアに搭載可能なサイズまで絞り込む
    - 解析処理の時間を考慮すると更にサイズを絞り込まなければならない

- ビッグデータ解析では『データの抽出・加工』工程は重要

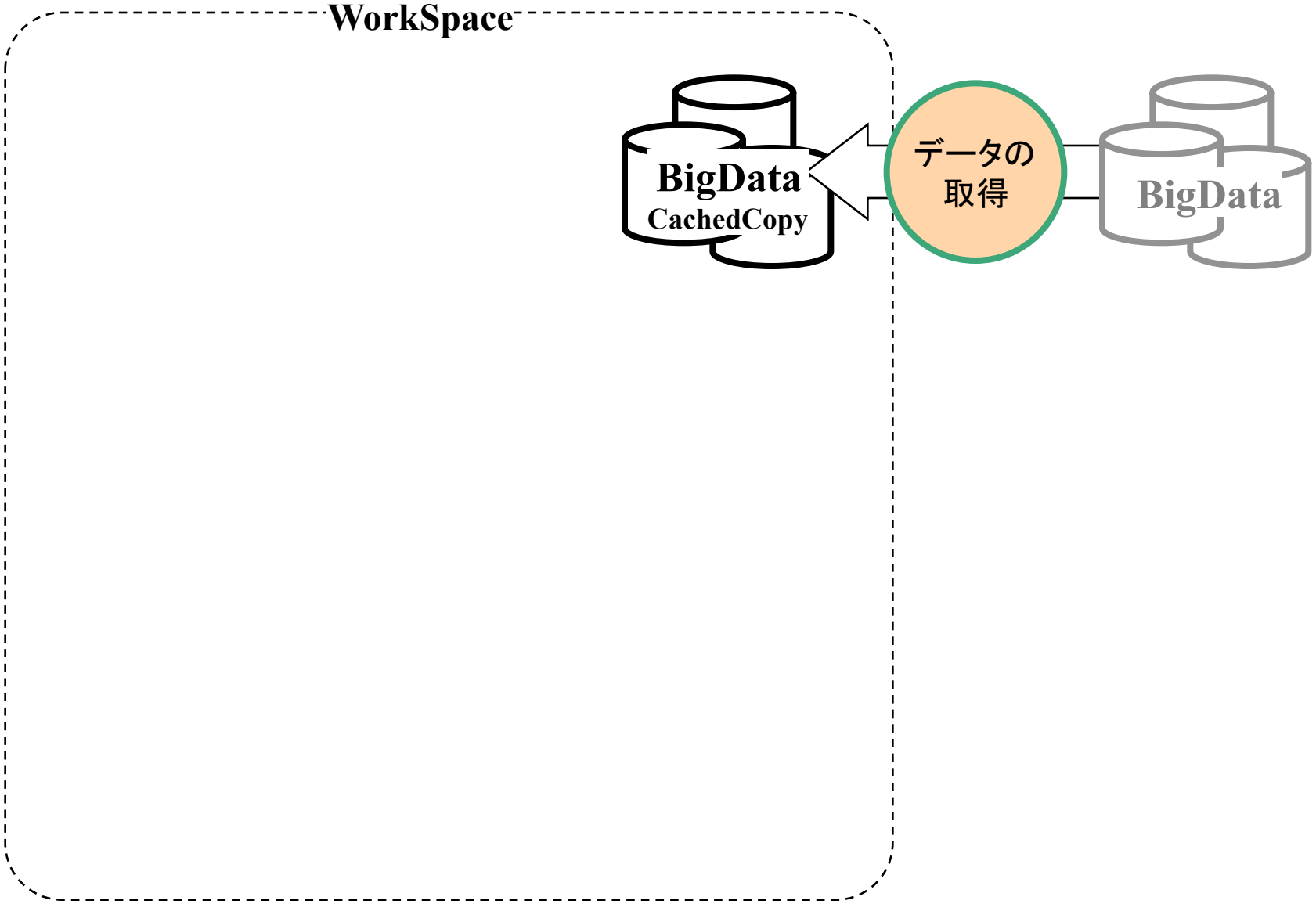
ビッグデータ解析のプロセス（2）



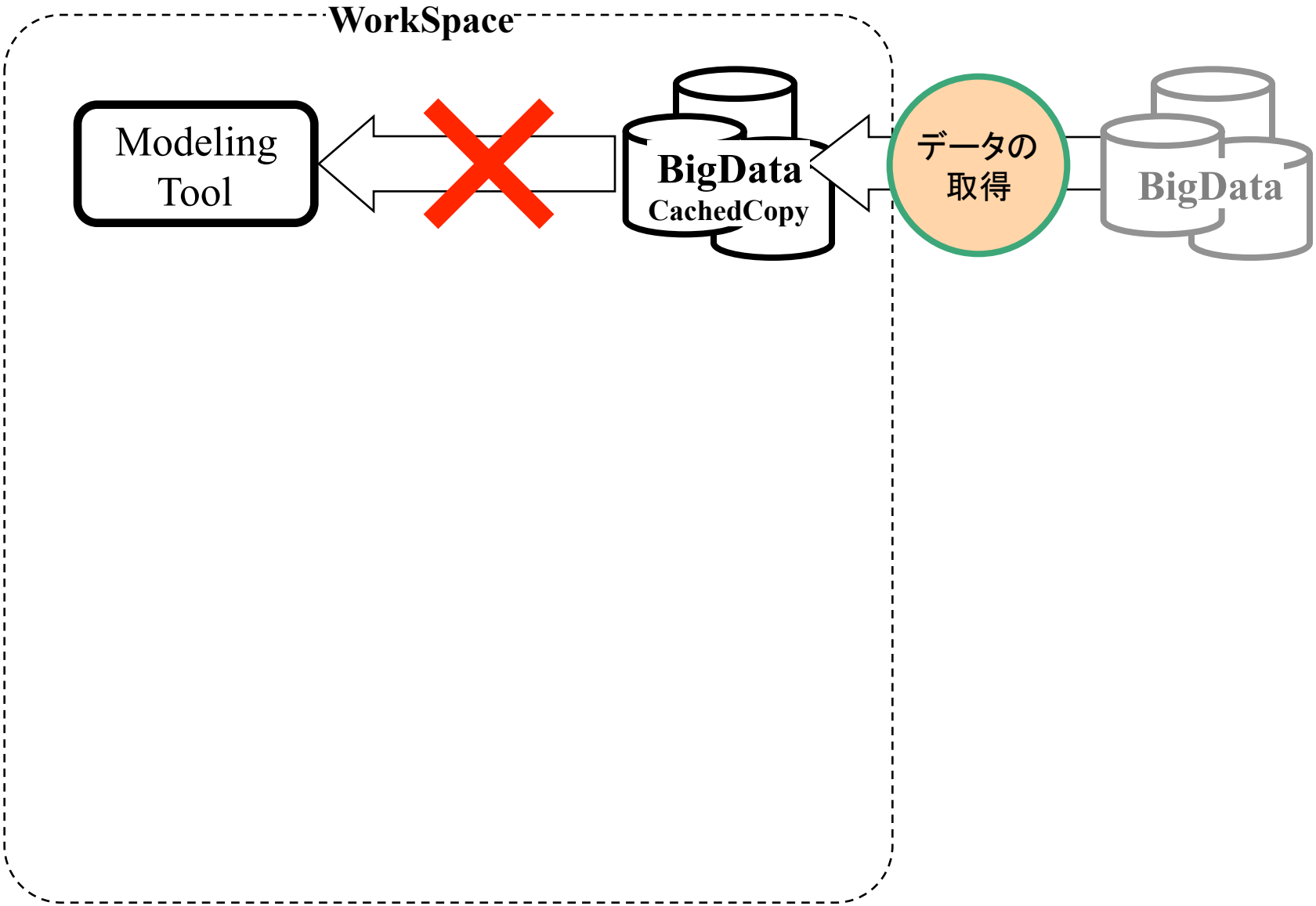
ビッグデータ解析のプロセス(3)

- データの取得(と理解)
  - データはネット経由で各所から入手できる
    - 自社データ、ソーシャルメディア、政府系データ、その他のリソース
  - 入手データの内容を理解する
    - 収録されるデータ、フォーマットなど・・・
- データの抽出・加工
  - モデリングに必要なデータを抽出する
    - テキストデータの場合は頻度等を求めて数値化
  - モデリングに適した形にデータを加工する
    - 単位やタイムゾーンなどを合わせる、個別の値を集計する
- モデリング
  - データの可視化と解析アルゴリズムの適用
    - 基本的に対話的なプロセス → データ解析アプリケーションの利用

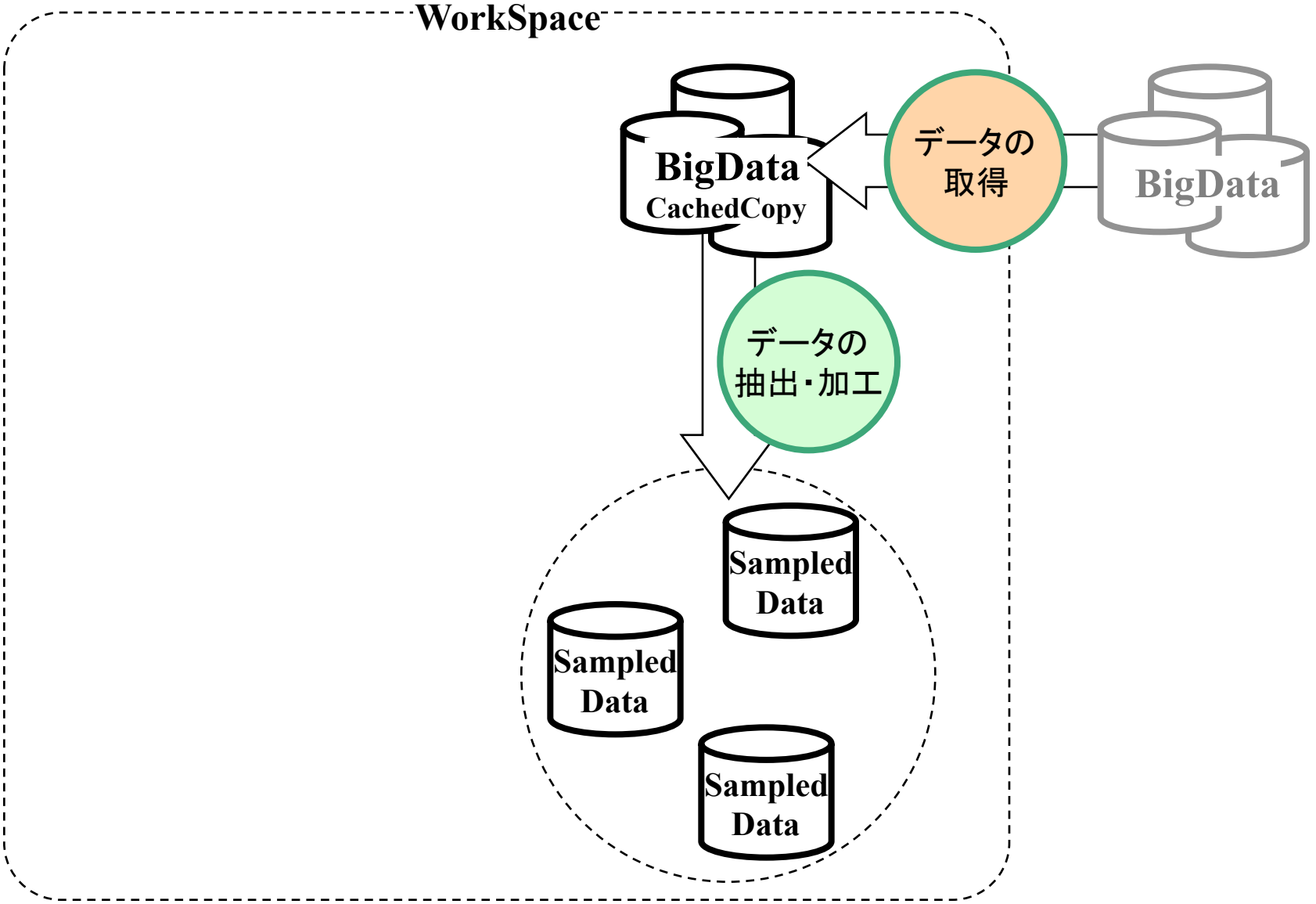
ビッグデータ解析の手順（1）



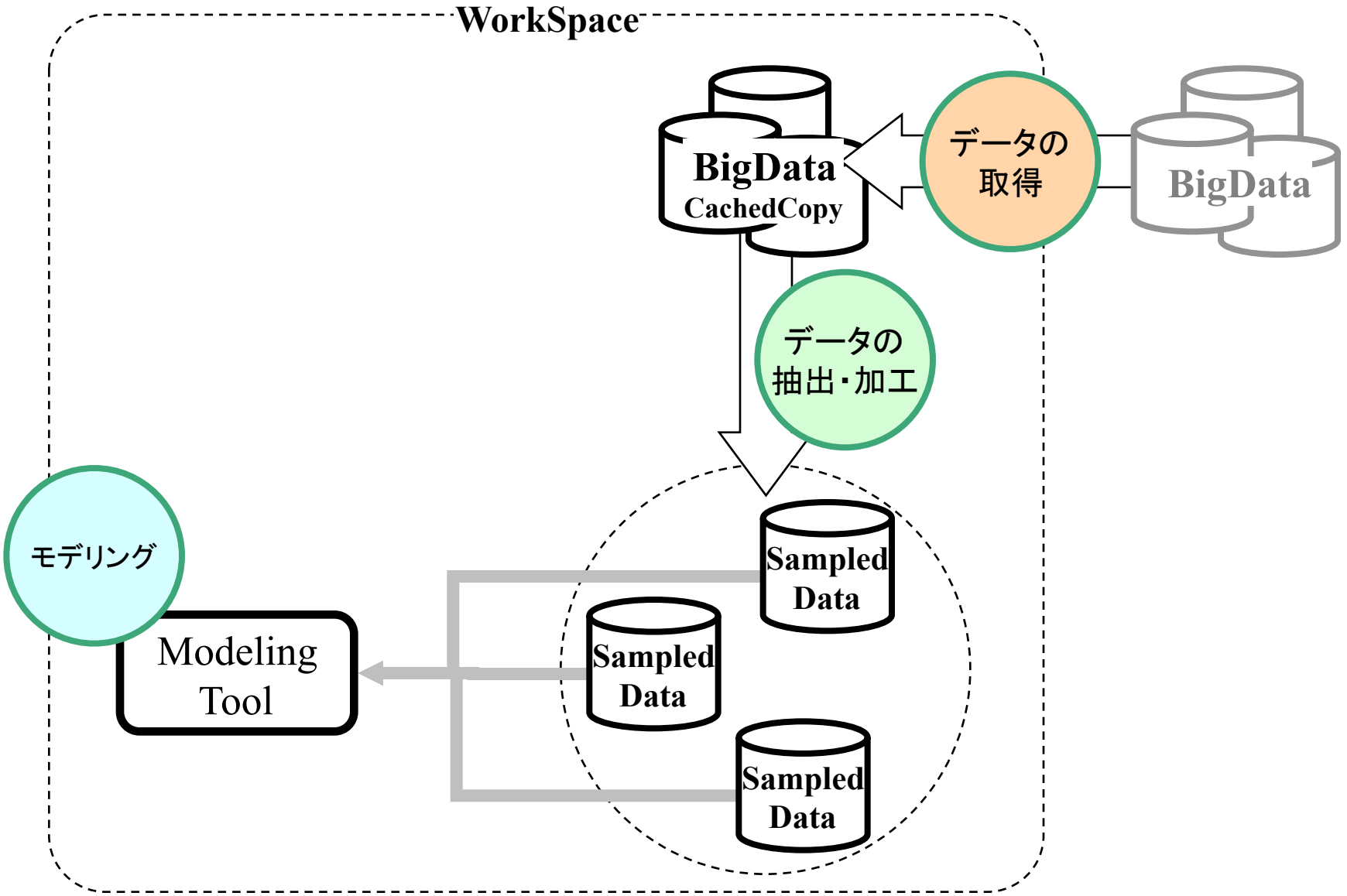
ビッグデータ解析の手順（2）



ビッグデータ解析の手順（3）



ビッグデータ解析の手順（4）

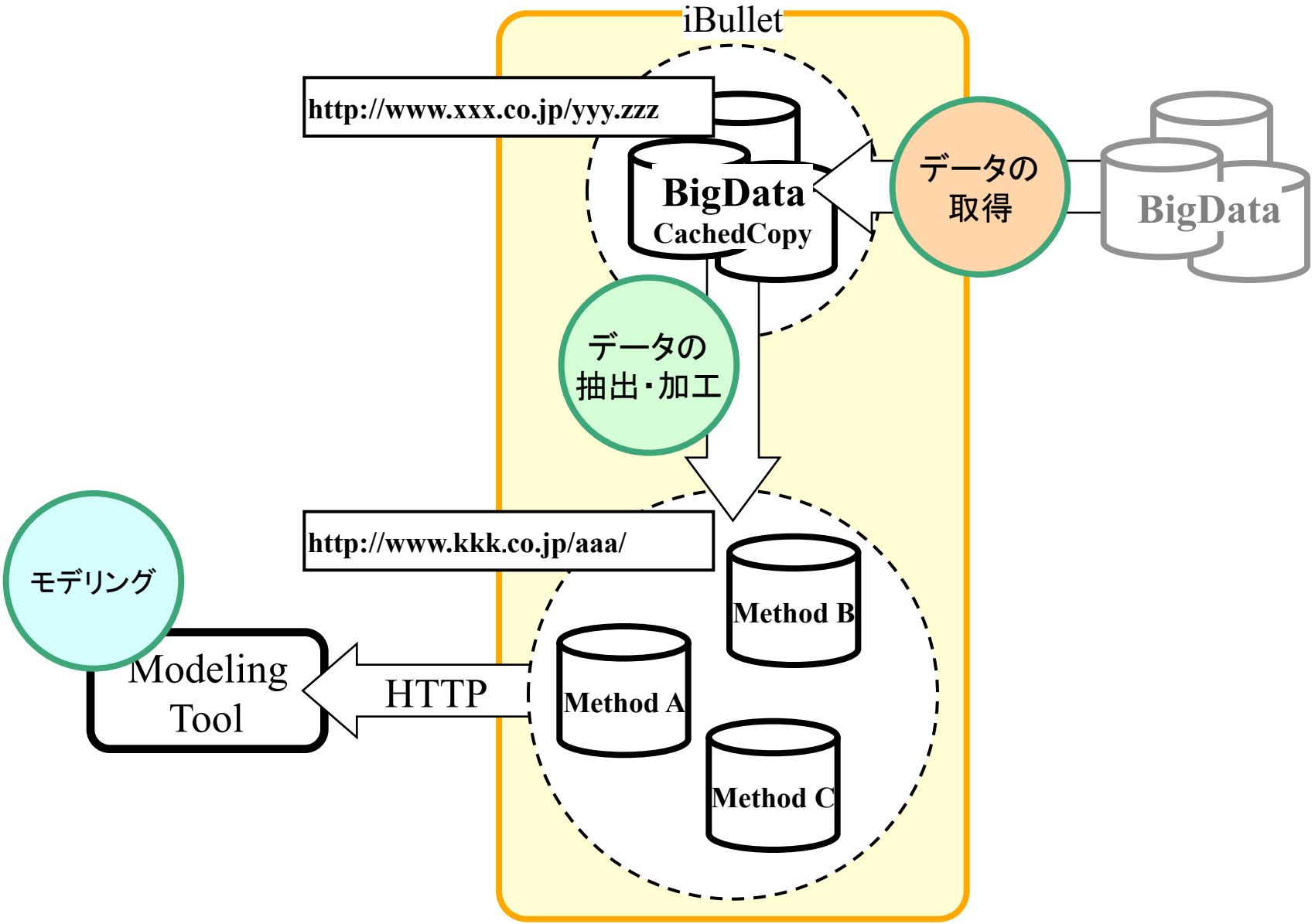


iBullet の導入（1）

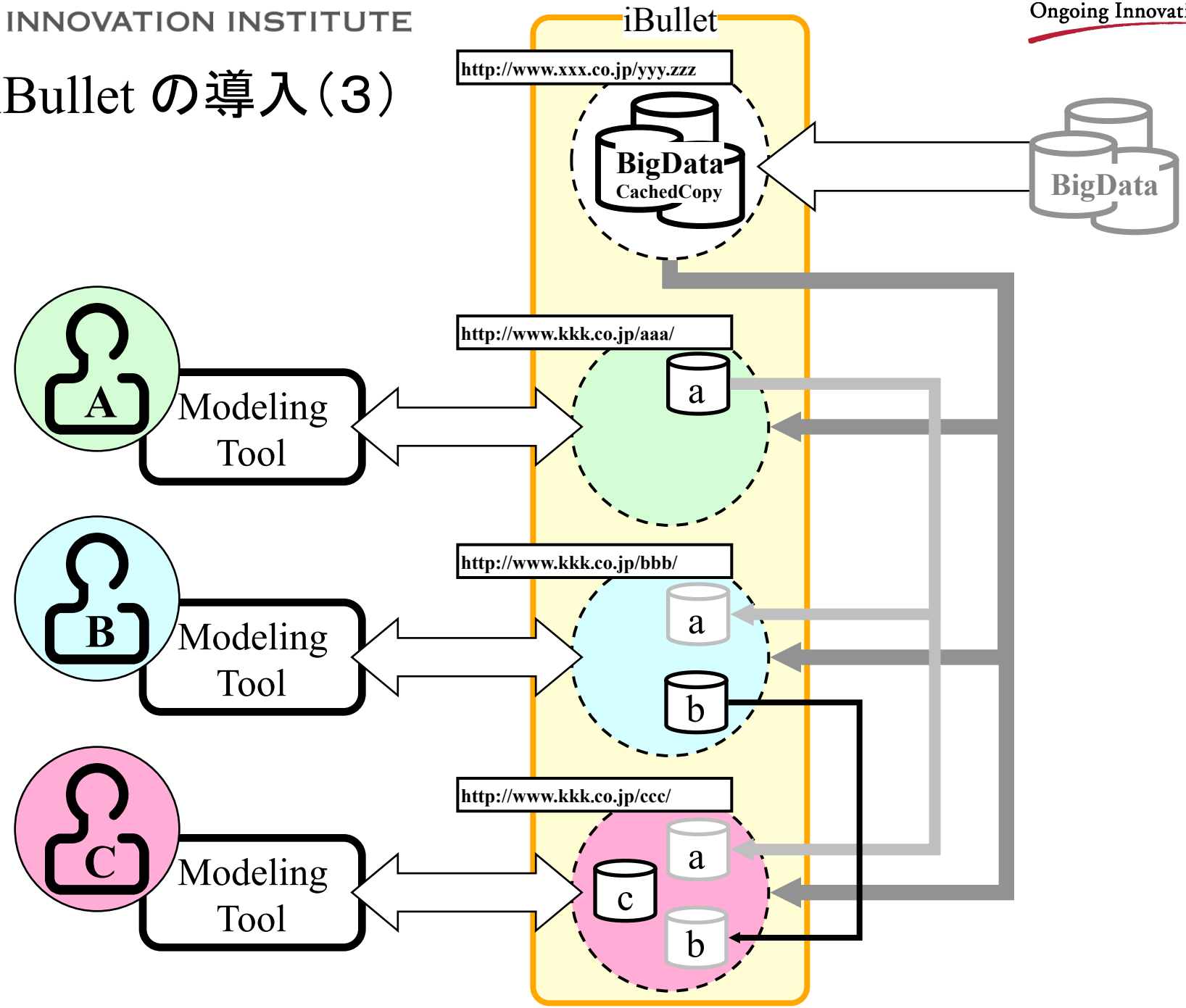
- キャッシュシステムをベースにした解析データ管理
  - ビッグデータ解析の作業過程で生成されたデータの一時保管
    - 『データの抽出・加工』工程に要する時間を節約できる
    - 『データの抽出・加工』工程の多段化により時間的効率がアップ
  - ビッグデータ解析の作業過程を外部より参照できる
    - 外部アプリケーションにはキャッシュドプロキシーとして動作
    - 『データの抽出・加工』工程で生成された中間データも任意に参照可能

- 狙い
  - ビッグデータ解析の手順（拡張モジュール）の保管
    - 解析作業の対象データと解析手順、解析結果をまとめて
    - 他の利用者によるビッグデータ解析作業の再現が容易になる
  - iBullet の分散化
    - 多段化された『データの抽出・加工』処理を任意のノードに配置
      - 処理内容に応じたスペックのノードに割り振る

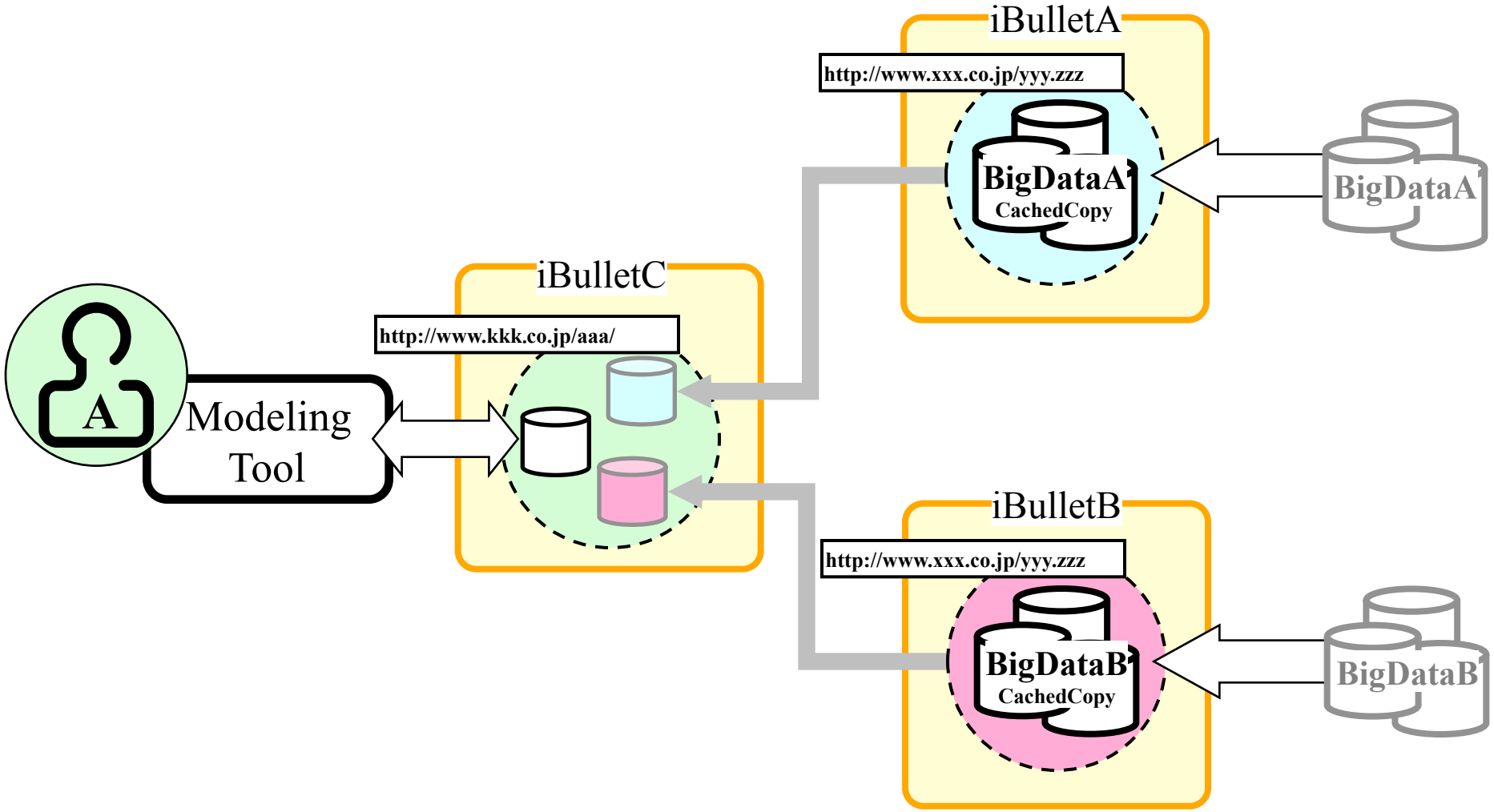
iBullet の導入（2）



iBullet の導入（3）



iBullet の導入（4）



事例：連続ドラマに着目したWikipediaPVC解析

- 目的： iBullet が想定する解析プロセスを具体的に検証
  - 機能紹介のためのデモとしても活用できる
- 動機： Wikipedia ランキングの挙動
  - Wikipediaのドラマページが放映時に顕著な反応する
    - 視聴率の高いドラマはランキング50位内に頻繁に登場
    - Wikipedia のページビューと視聴率には何らかの相関がある？
- 方法： 前述のビッグデータ解析プロセスに基づく
  - 対象データは Wikipedia から入手できるもののみとする
    - ページビューデータ： Wikipedia の辞書ページの参照カウント
    - ページデータ： Wikipedia の辞書ページ（Mediawikiフォーマット）
  - ページビュー情報からドラマ関連情報抽出して解析を行う
    - 各ドラマ毎、放映された四半期のドラマのページビュー情報を抽出
    - 抽出処理にはドラマページの情報を活用
  - モデリングには Excel と R を活用

Wikipedia Pageview Count（1）

- “Page view statistics for Wikimedia projects”
  - http://dumps.wikimedia.org/other/pagecounts-raw/
    - Wikimedia プロジェクトの各ページのページビュー数を集計
    - 2013年1月より公開開始
    - 2008年以降～現在までのページビューデータを毎時追加

- データフォーマット:
  - テキストファイル：1行ごとにスペース区切りで下記の情報を記録
    - <Project>　　プロジェクト種別（言語＋プロジェクト）
    - <PageTitle>　　ページタイトル（HTTPエンコード）
    - <Pageview>　　ページビュー数
    - <PageSize>　　ページサイズ

- 欠損データなどの詳細は下記のページで紹介してます
  - http://www.gryfon.iij-ii.co.jp/ranking.html

Wikipedia Pageview Count（2）

データサイズ（2007/12/09 18:00から2013/11/16 23:00まで）

Project: ja
Namespace: 0

	圧縮	伸張	日本語
2007	13,398,810,576	57,129,124,521	5,240,415,228
2008	374,872,455,581	1,398,699,752,241	104,537,738,517
2009	502,812,348,159	1,845,979,467,189	118,126,627,592
2010	554,016,537,086	1,983,856,917,276	131,154,527,976
2011	667,206,244,576	2,401,022,366,436	148,014,598,499
2012	779,103,332,802	2,813,104,716,667	157,935,218,023
2013	663,407,267,632	2,430,725,389,233	119,934,028,993
合計	3,554,816,996,412	12,930,517,733,563	784,943,154,828
GB	3310.68	12042.48	731.04

6.07%

Wikipedia Page Data

- “Wikimedia Downloads -- Database dump progress”
  - http://dumps.wikimedia.org/backup-index.html
  - http://dumps.wikimedia.org/jawiki/（日本語版）
    - Wikipedia のダンププロセスは常時稼動している
    - 各言語ごとに巡回し、概ね1ヶ月おきに新しいダンプができる
  - データフォーマット
    - 基本的には XML フォーマット
    - 辞書本文は Mediawiki フォーマット（<TEXT>でタグ付け）
    - 様々な収録データの組み合わせでファイルを公開している
      - 我々が使っているのは jawiki-<date>-pages-articles.xml.bz2

- jawiki-20131005（2013/10/05のスナップショット）
  - 全ページ: 1,752,890 ページ
  - 辞書ページ: 1,411,191 ページ （80.5%）
  - リダイレクトを除く: 883,537 ページ （50.4%）

Wikipedia 日本語版のドラマ関連ページ

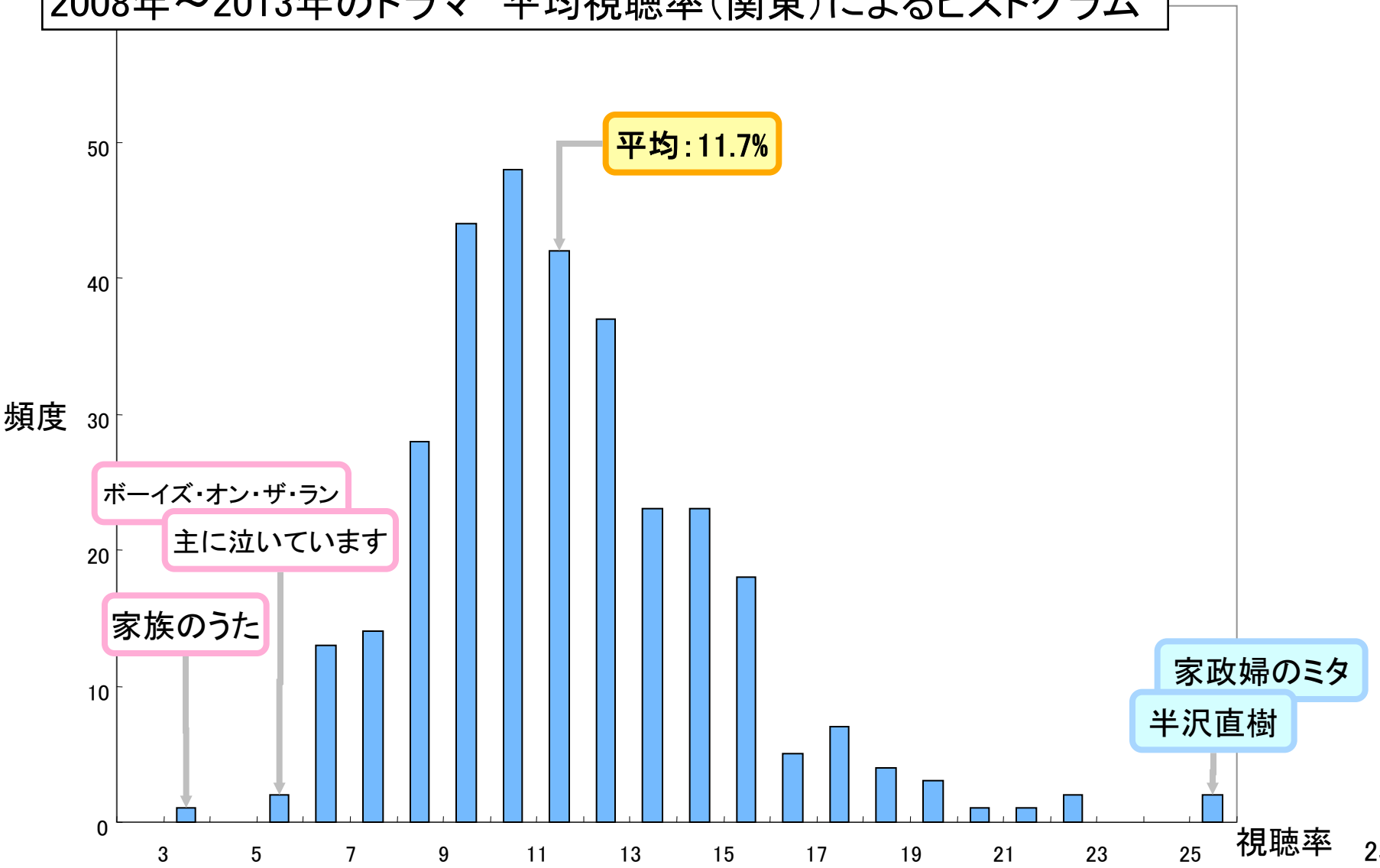
- Wikipedia 日本語版ドラマページは次の3つのパターン
  - 原作のページに「テレビドラマ」のセクションがある
    - 最初はこのパターンが多い
  - テレビドラマ単独のページ
    - オリジナルドラマの場合
    - 上記のパターンから独立した（「半沢直樹」はこのパターン）
  - シリーズ化されたドラマのページ
    - 各シーズンごとにセクションがある
    - 「登場人物」等が別ページに独立している場合もある
      - 解析者にとって「相棒」は最悪のページ
- 今回の解析で着目した各ページの情報
  - 視聴率情報
    - 関東圏の全体視聴率、各回の平均（瞬間）視聴率
  - キャスト、スタッフのページへのリンク
    - 原作ものの場合は原作者も・・・

Wikipedia 日本語版ドラマページの解析（1）

- 2008年以降2013年第3四半期までの317件を選択
  - 四半期の期間で全8回～11回のドラマをピックアップ
    - 四半期枠からはみ出てしまう NHK のドラマは除外
    - 特番扱いの民放スペシャルドラマも除外
- ドラマの視聴率について
  - ドラマ視聴率は制作者の通信簿
    - 20%～:　超優良（6件）
      - 「半沢直樹」「家政婦のミタ」「ごくせん」「CHANGE」「JIN」「相棒」
    - 15～20%:　優秀
      - ここに入ればスポンサーに対し強気に出れる
      - 1回延長 or 放映時間延長
    - 10～15%:　良
      - 最初はここを目指す
    - 5～10%:　可
      - スポンサーに怒られる
      - 途中で打ち切られる（10回 or 8回）
    - ～5%: 不可
      - 2008年以降では「家族のうた」のみ

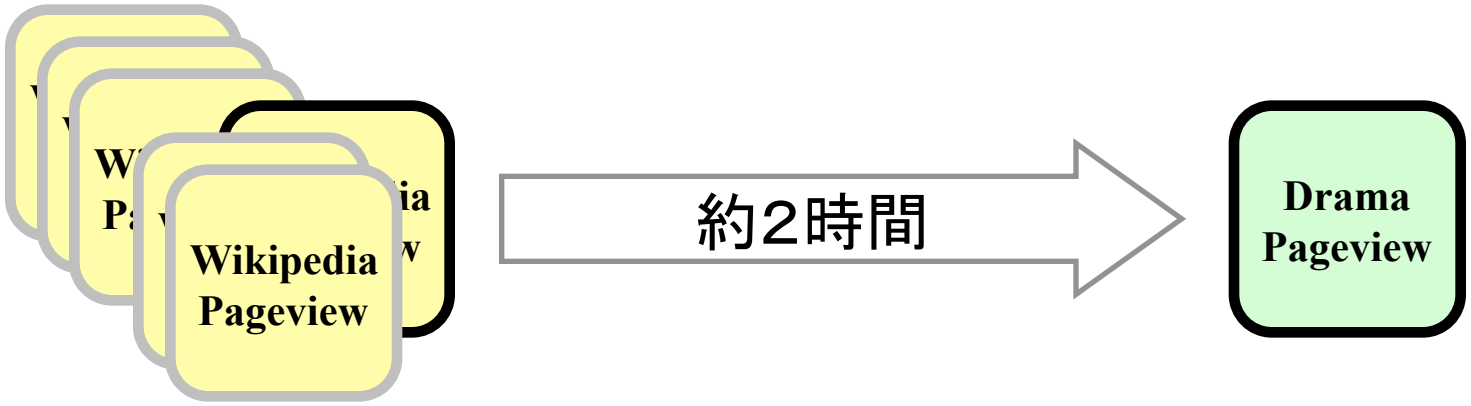
Wikipedia 日本語版ドラマページの解析（2）

2008年～2013年のドラマ　平均視聴率（関東）によるヒストグラム

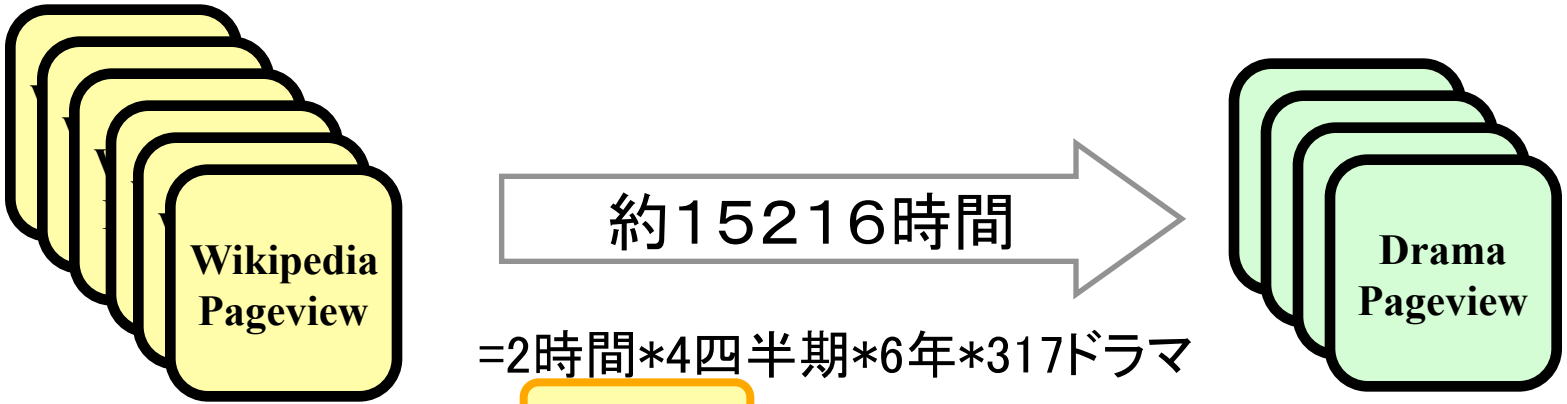


ドラマページ情報の抽出・加工（1）

1つのドラマが放映された4半期のページビュー情報を抽出する



全ドラマについて2008～2013年のページビュー情報を抽出すると・・・

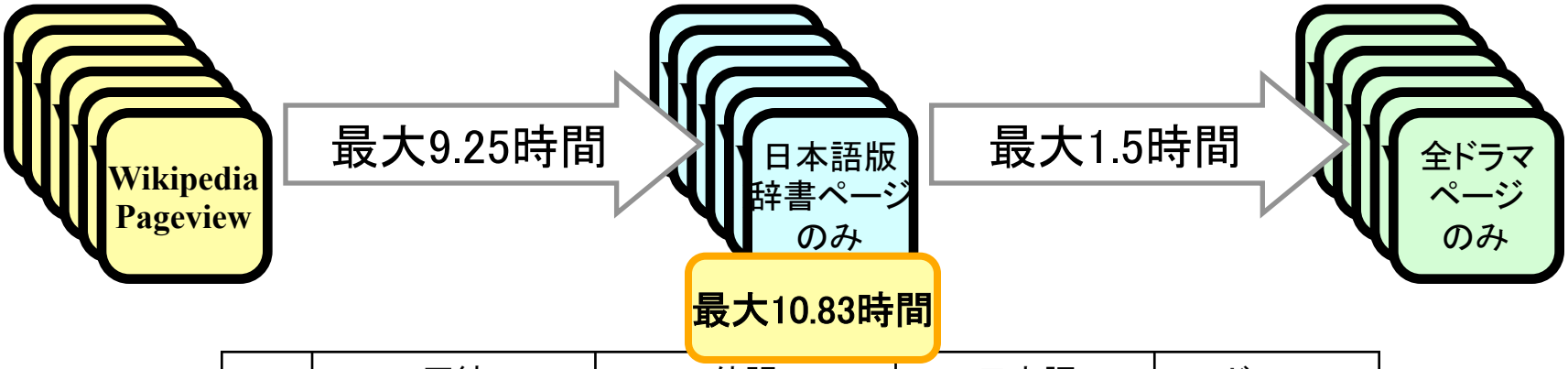


=2時間*4四半期*6年*317ドラマ

= 約21ヶ月

ドラマページ情報の抽出・加工（2）

ノード6台並列で段階的にデータ抽出を行う



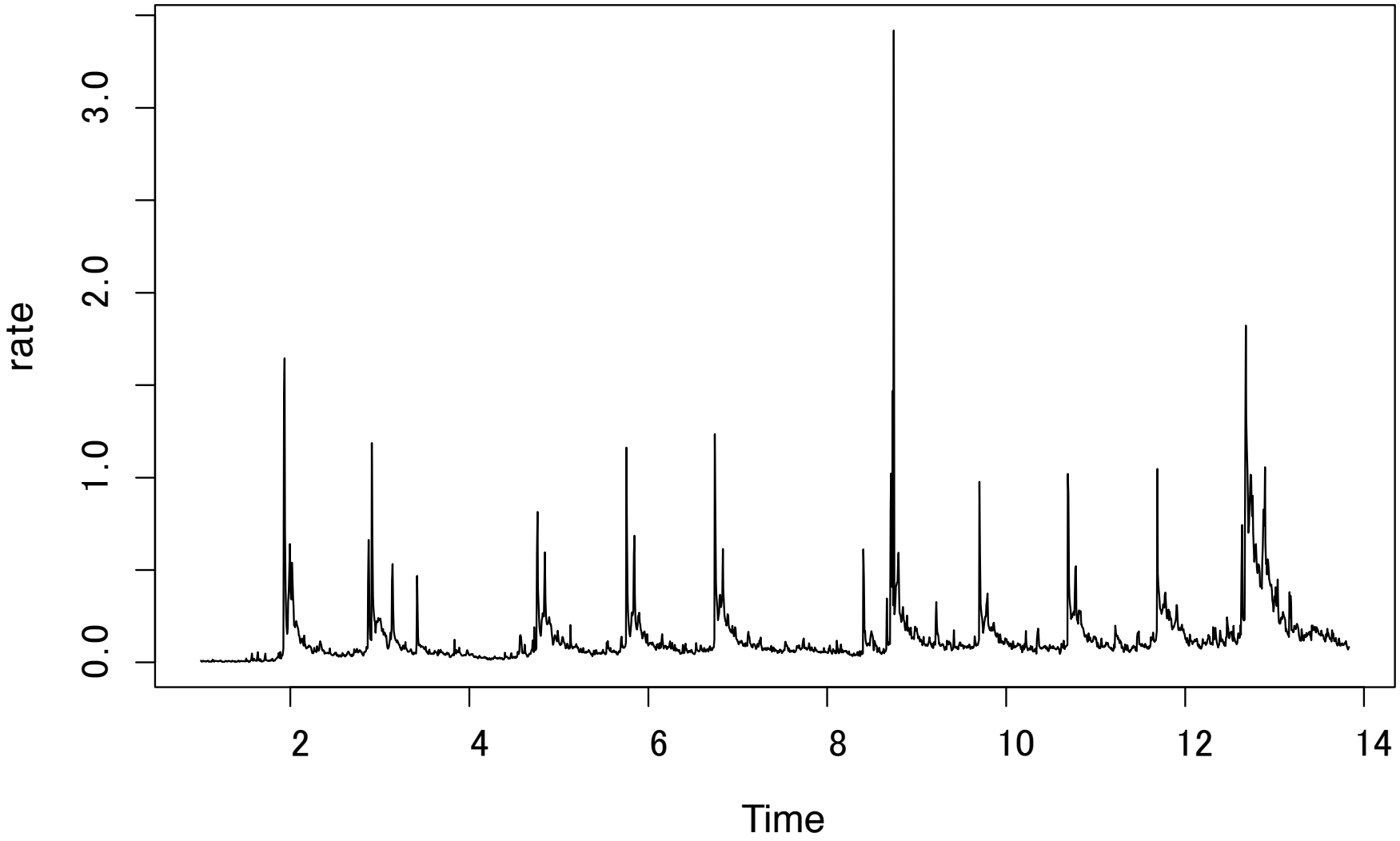
	圧縮	伸張	日本語	ドラマ
2007	13,398,810,576	57,129,124,521	5,240,415,228	17,181,933
2008	374,872,455,581	1,398,699,752,241	104,537,738,517	281,512,606
2009	502,812,348,159	1,845,979,467,189	118,126,627,592	274,634,787
2010	554,016,537,086	1,983,856,917,276	131,154,527,976	277,636,429
2011	667,206,244,576	2,401,022,366,436	148,014,598,499	271,058,710
2012	779,103,332,802	2,813,104,716,667	157,935,218,023	285,049,720
2013	663,407,267,632	2,430,725,389,233	119,934,028,993	221,514,914
合計	3,554,816,996,412	12,930,517,733,563	784,943,154,828	1,628,589,099
GB	3310.68	12042.48	731.04	1.52

Rを使ったデータ解析

- ドラマページのページビューを解析
  - ドラマ放送時にはページビューに顕著な変化が現れる
    - 各回の放送中の1時間
    - 放送終了後24時間
    - 放送終了後168時間（7日間）

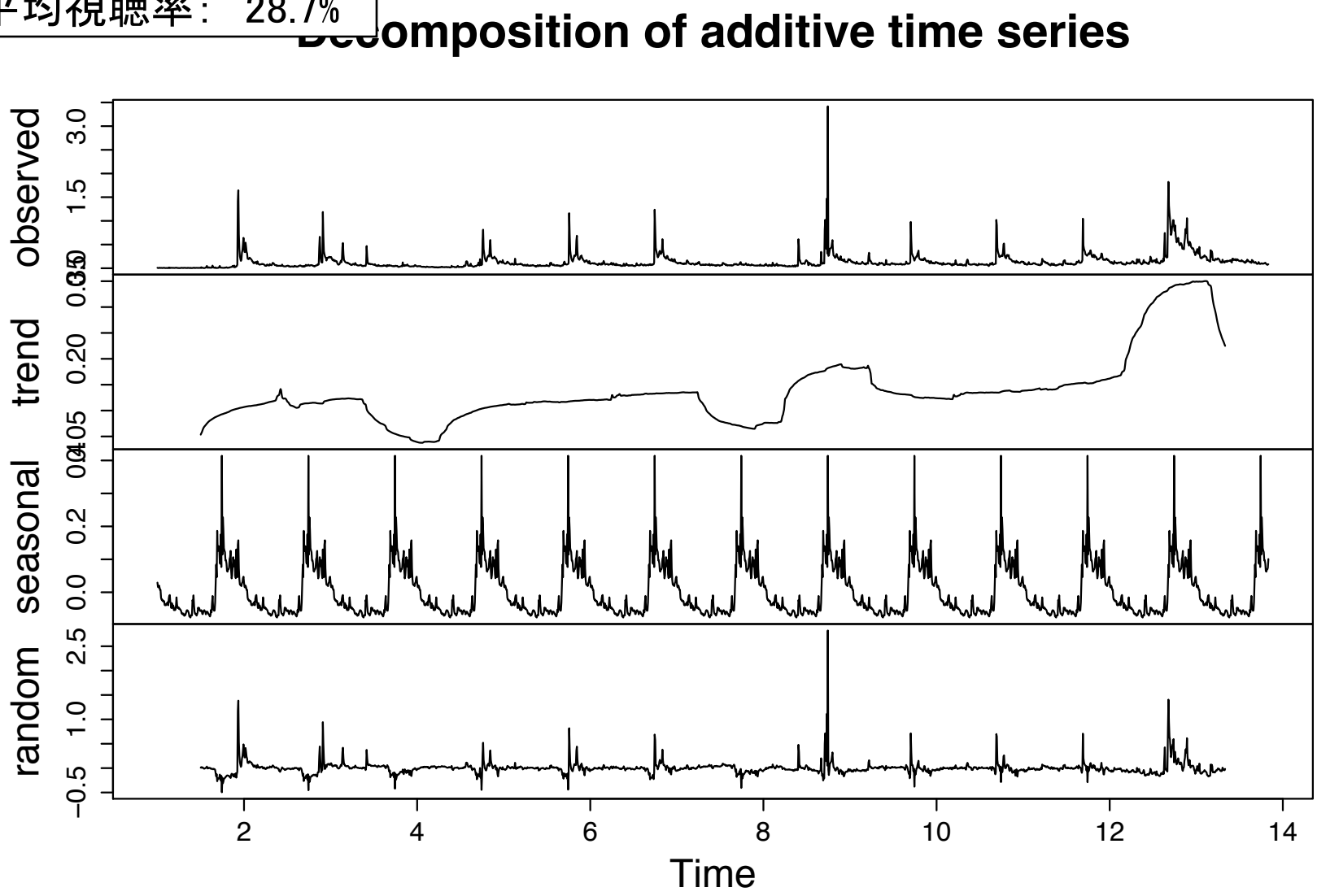
- Rの時系列解析を利用してページビュー変動を可視化
  - 関数 ts()：　　時系列データへの変換する
    - パラメータとして変動周期（24 or 168）を設定
  - 関数 decompose()：　時系列データを要因ごとに分解する
    - トレンド（Trend）、季節要因（seasonal）、ノイズ（random）
      - 移動平均法による　→　時系列解析としては古典的な手法
      - Rではデフォルトでビルドインされている
  - 結果
    - 変動周期 = 168 の場合にトレンドはドラマの評価が現れる

「半沢直樹」のページビュー



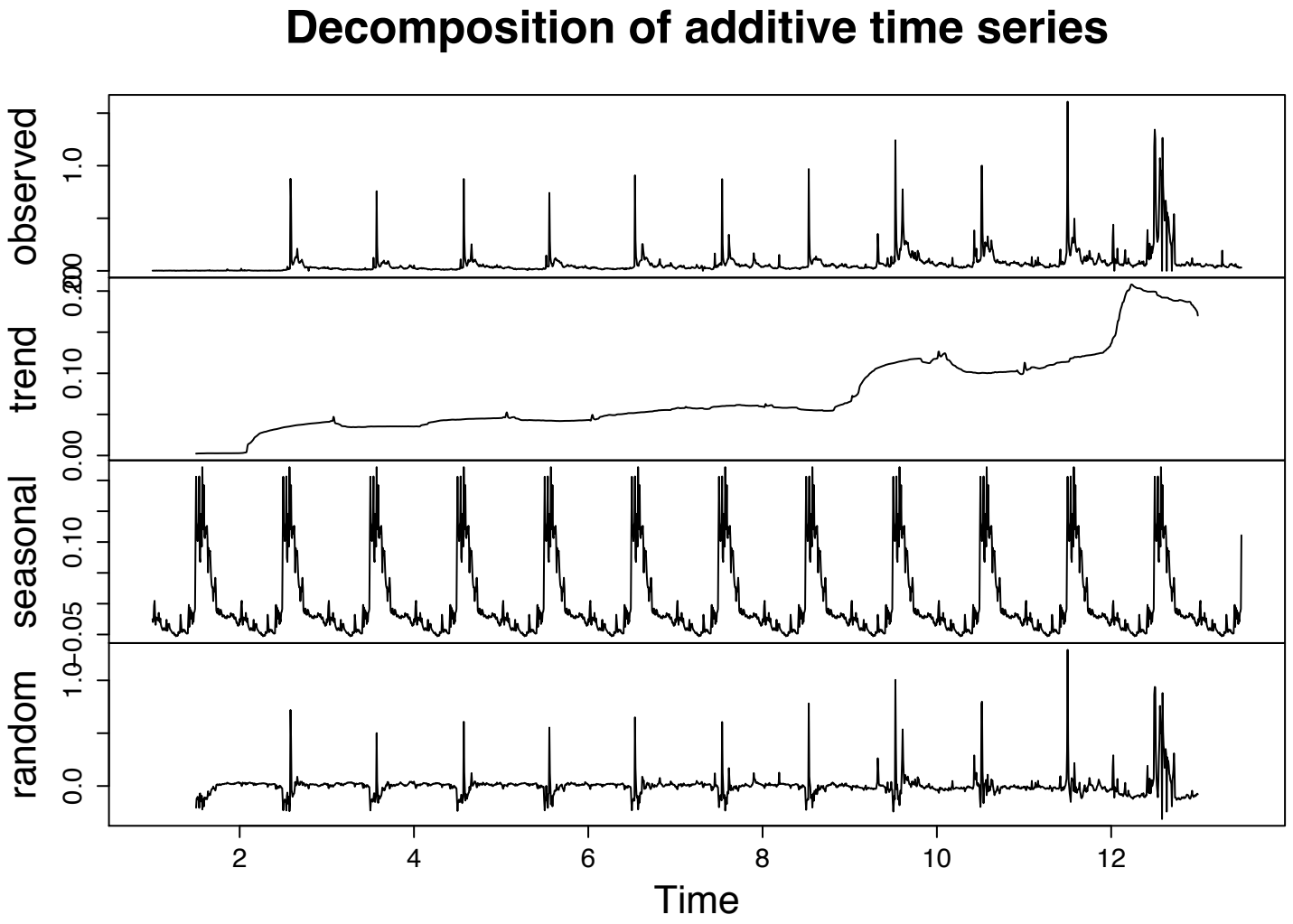
「半沢直樹」の時系列解析（1週間周期変動）

平均視聴率： 28.7%



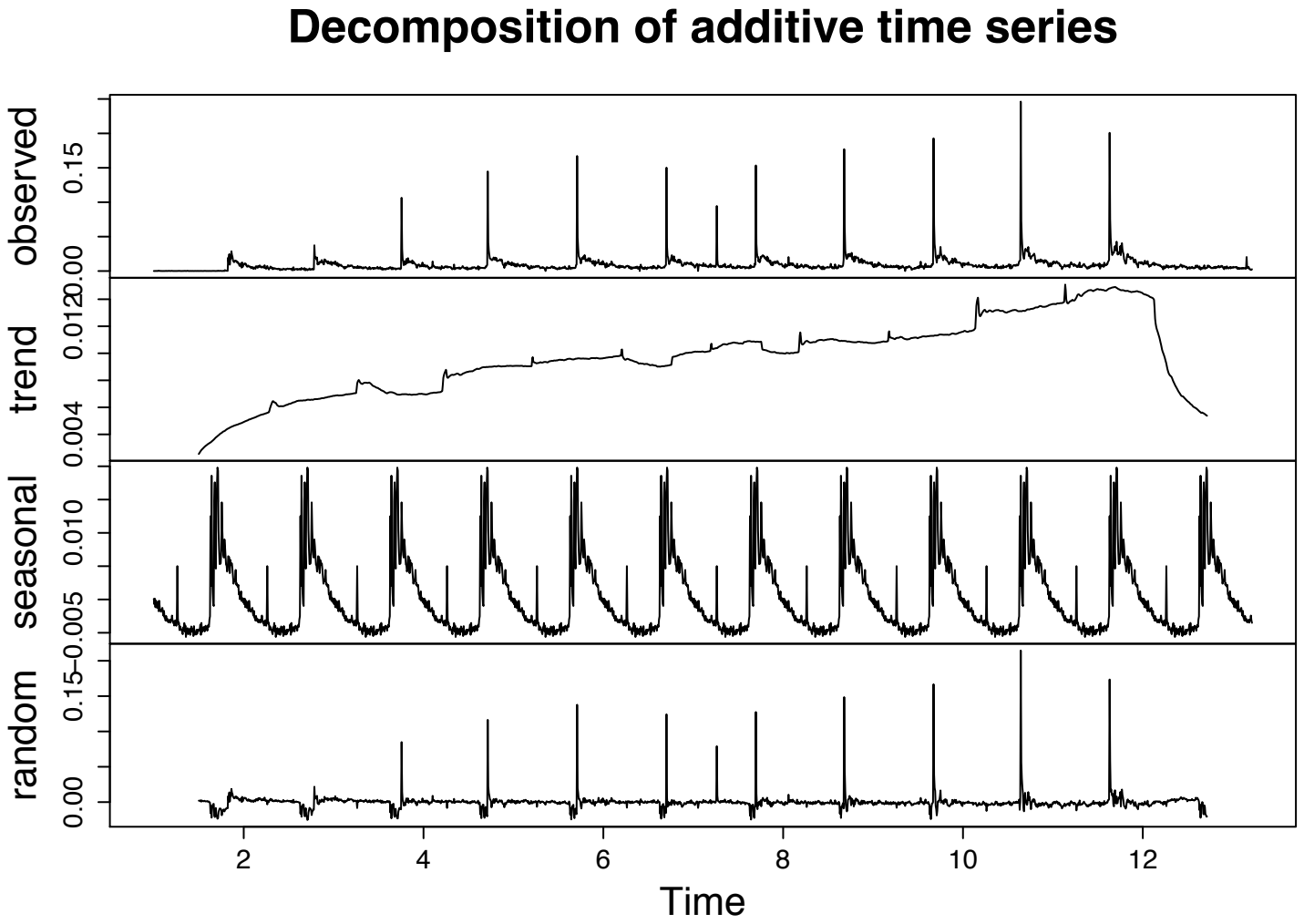
「家政婦のミタ」の時系列解析（1週間周期変動）

平均視聴率： 25.2%



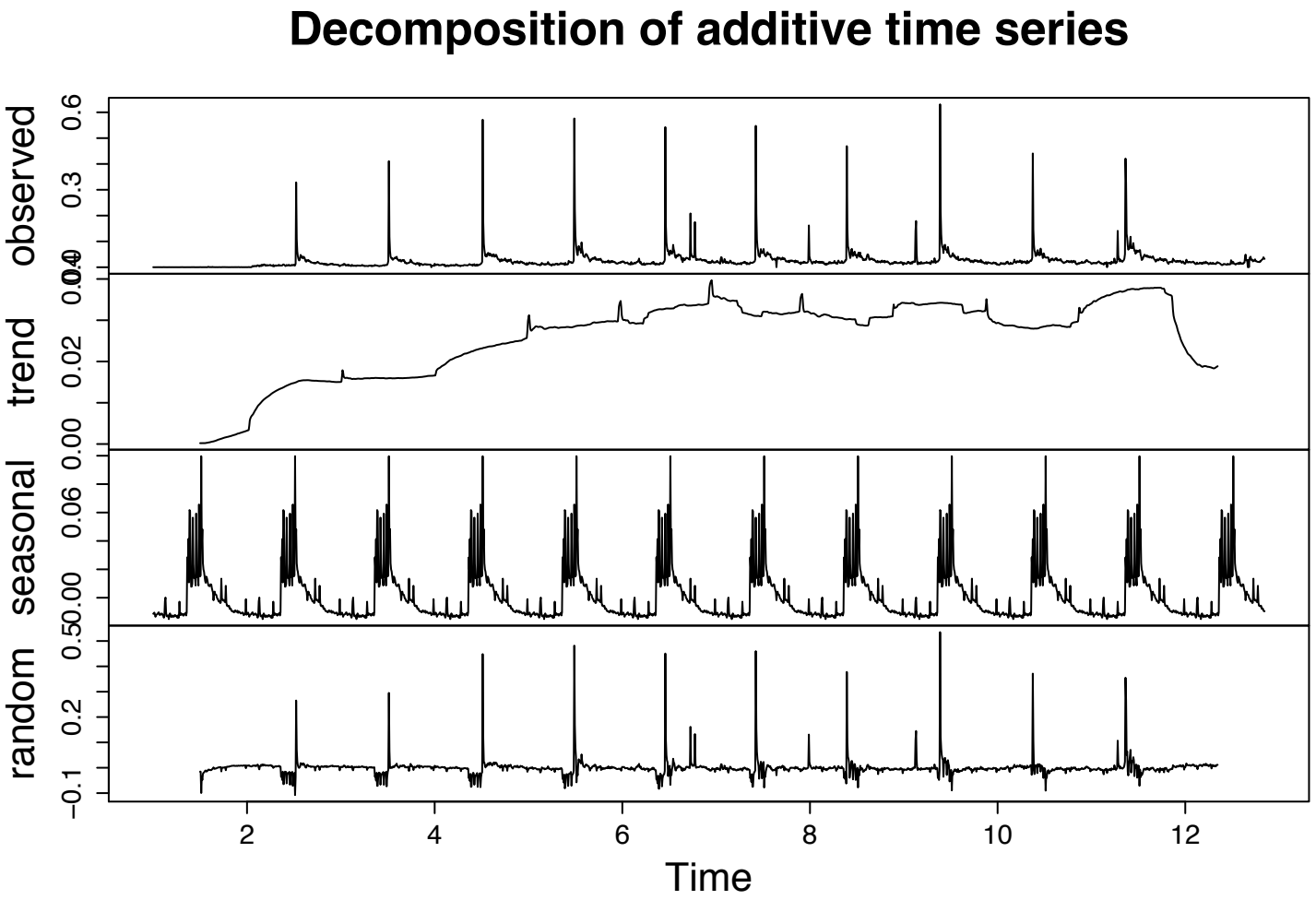
「結婚しない」の時系列解析（1週間周期変動）

平均視聴率： 11.8%



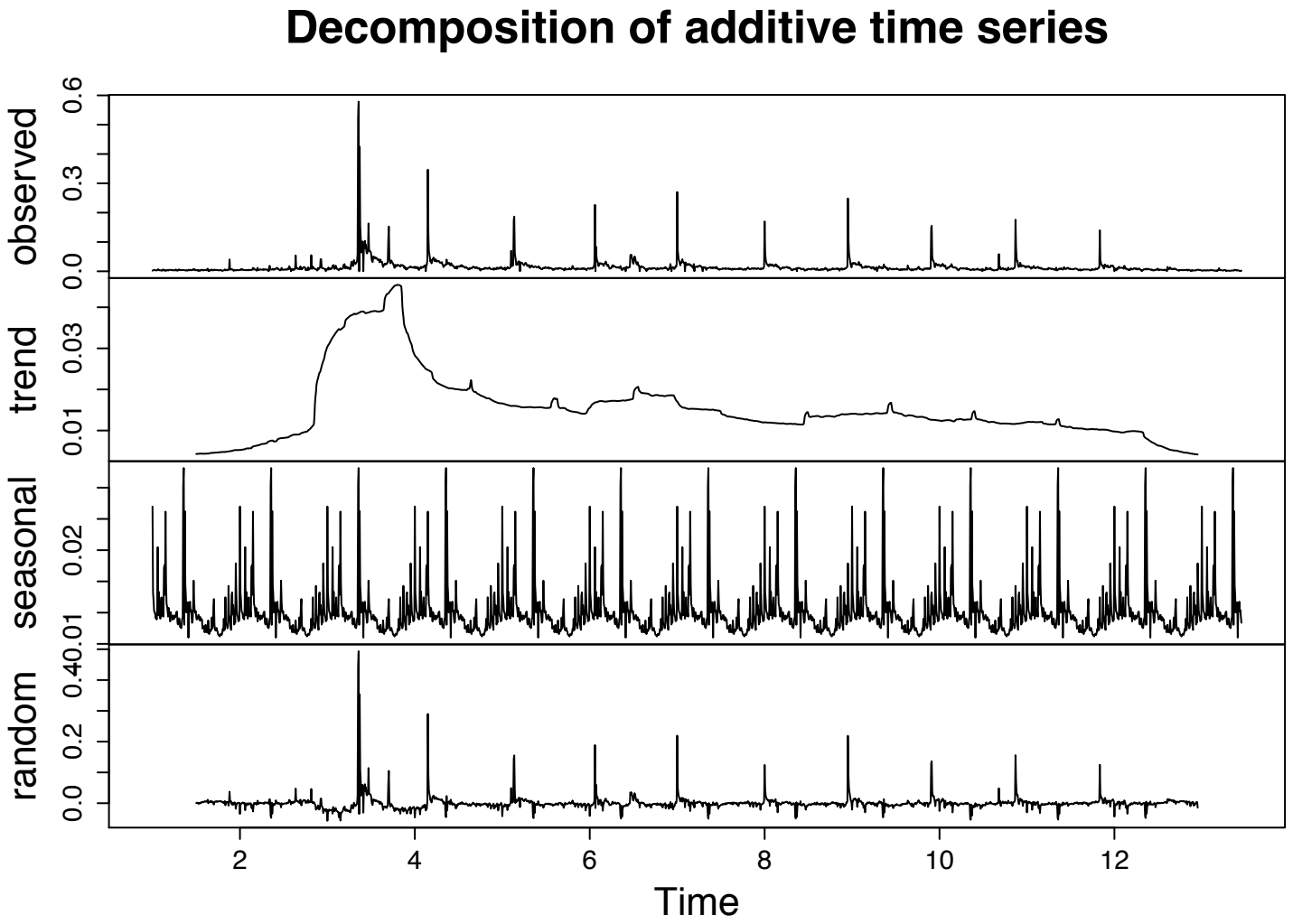
「最高の離婚」の時系列解析（1週間周期変動）

平均視聴率： 11.8%



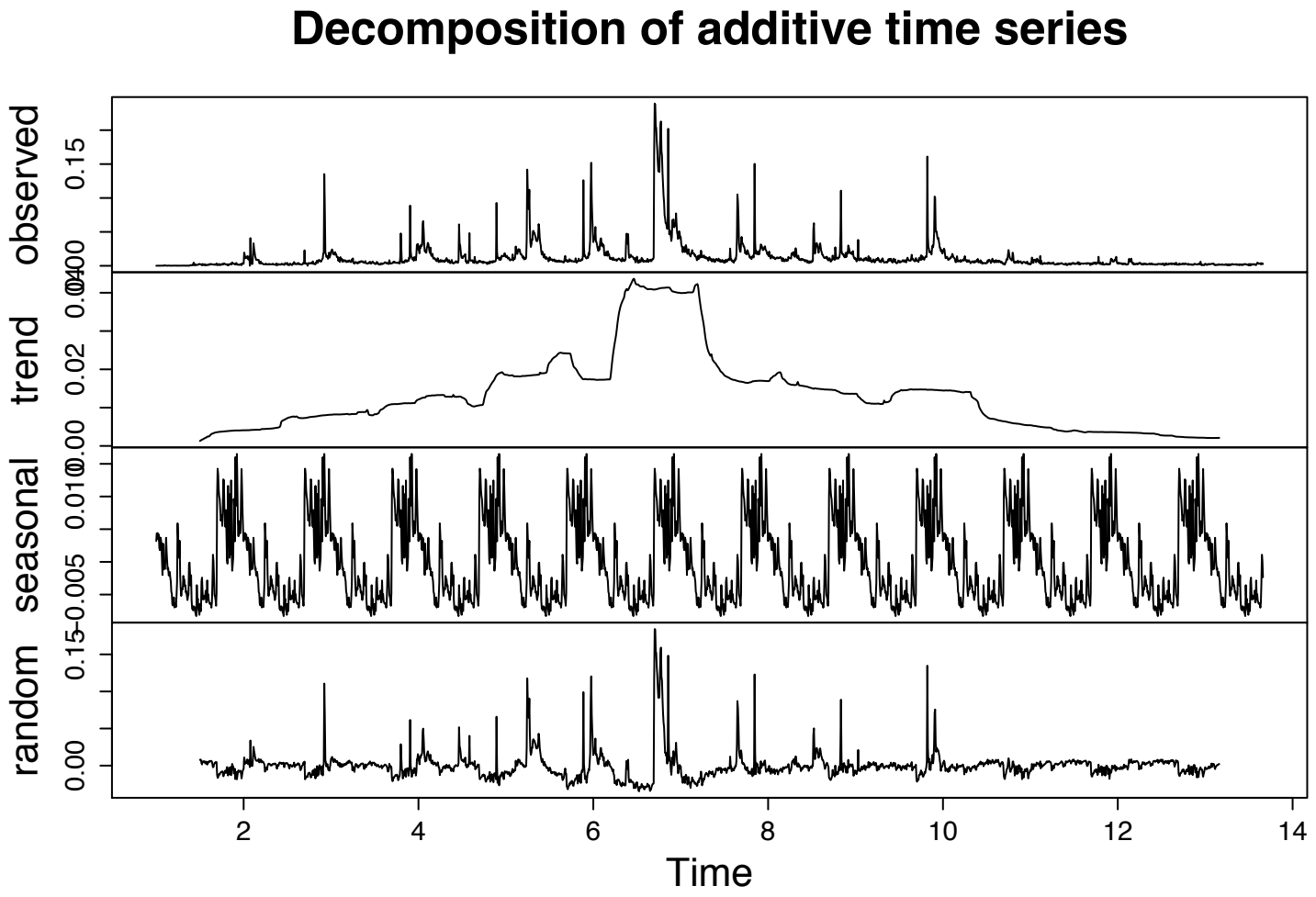
「ぴんとこな」の時系列解析（1週間周期変動）

平均視聴率： 7.3%

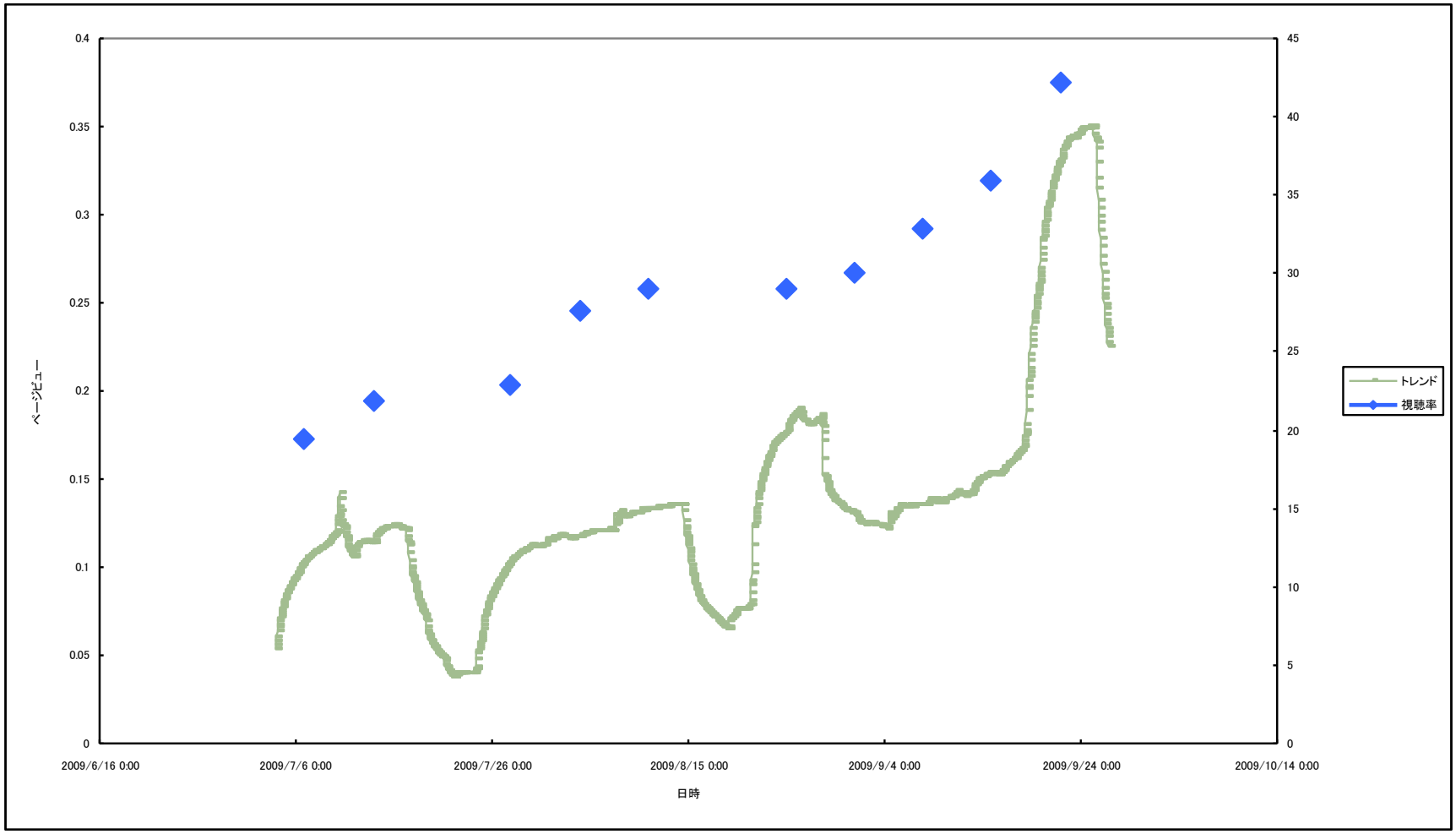


「家族のうた」の時系列解析（１週間周期変動）

平均視聴率： 3.9%



「半沢直樹」のトレンドと視聴率



まとめ（1）

- Indexer Bullet について
  - キャッシュシステムをベースにした解析データ管理
    - ビッグデータ解析の作業過程で生成されたデータの一時保管
    - ビッグデータ解析の作業過程を外部より参照できる
  - 得られる効果
    - 『データの抽出・加工』の処理を多段化できる
      - 時間的コストのかかる処理の実行頻度を抑制する
    - 頻繁な反復を繰り返す『モデリング』作業の効率化に寄与する
      - 解析対象データの取り出しコストを最小化する
  - 次の展開
    - ビッグデータ解析の手順（拡張モジュール）も一時保管
    - iBullet の分散化

まとめ（2）

- Wikipedia ページビュー情報について
  - 完全にパブリックなソーシャルデータ
    - 任意の解析について第3者の追試が可能
      - 第3者による誤りの指摘や修正が可能
      - 情報量は少ないが信頼性は期待できる
  - 時系列解析により比較的容易に社会的トレンドを把握できる
    - ページカウントは社会的イベントに敏感に反応する
      - 今、何が起きているのか大づかみに把握するには便利
      - ソーシャルメディアからの情報との組み合わせで詳細化は可能
  - 辞書情報が『データの抽出・加工』に役立つ
    - 調査対象ごとページが存在する
      - 各ページには調査対象と関連性のある情報が網羅されている

おまけ

- Wikipedia ドラマページの解析について
  - できれば現在クールのドラマの最終視聴率の予測がしたかった
  - ドラマ視聴率とWikipedia ページビュー
    - Wikipedia ページビューはネット上でのドラマの関心度
      - ドラマの視聴量とはなんらかの相関はあるが・・・それだけではない
      - 人気の高いドラマは視聴率との相関性は高そう
      - 人気の低いドラマは・・・
  - ドラマ視聴量の抽出はできないか？
    - 時系列解析の要因分解の手法を応用して
      - 社会的イベントは指数関数を取る
      - Wikipedia ページビューを対象にした要因分解手法は？