ビッグデータを活用した取り組み

IIJ Techweek2014

2014/11/26

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

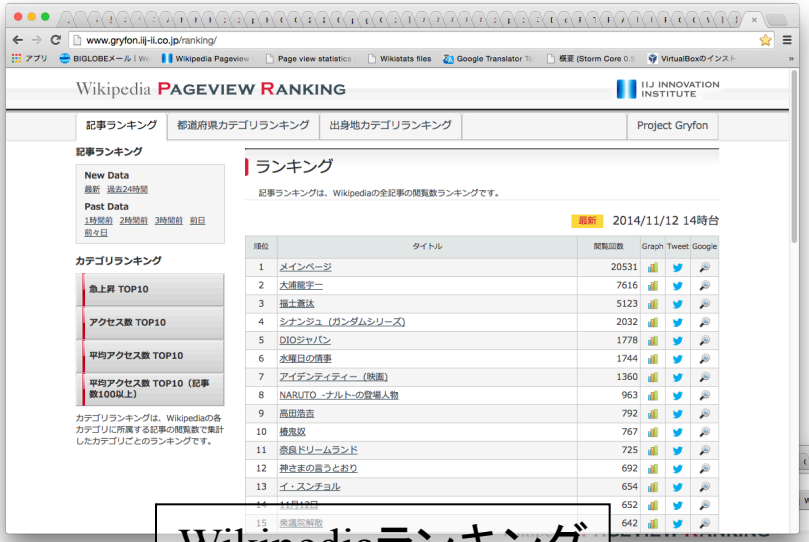
はじめに　ー　Project Gryfonの取り組み

- Project Gryfon
  - 当初：　クラウド基盤技術の研究開発を手がける
  - 現在：　研究開発の軸足を「基盤技術」から「ビッグデータ」に移行中

- ランキングによる（ビッグ）データの可視化
  - Wikipediaランキング
    - Wikipedia Pageview Count による記事ランキング
  - 統計データランキング
    - 政府公開データを使用したランキング
    - 作物統計調査 and 消費実態調査
  - Wikipedia カテゴリを利用したランキング
    - 都道府県 and 出身地
  - 感染症動向調査データランキング
    - 国立感染症研究所が公開する調査データによるランキング

Wikipedia ランキング

http://www.gryfon.iij-ii.co.jp/ranking/



Wikipediaランキング

出身地カテゴリランキング

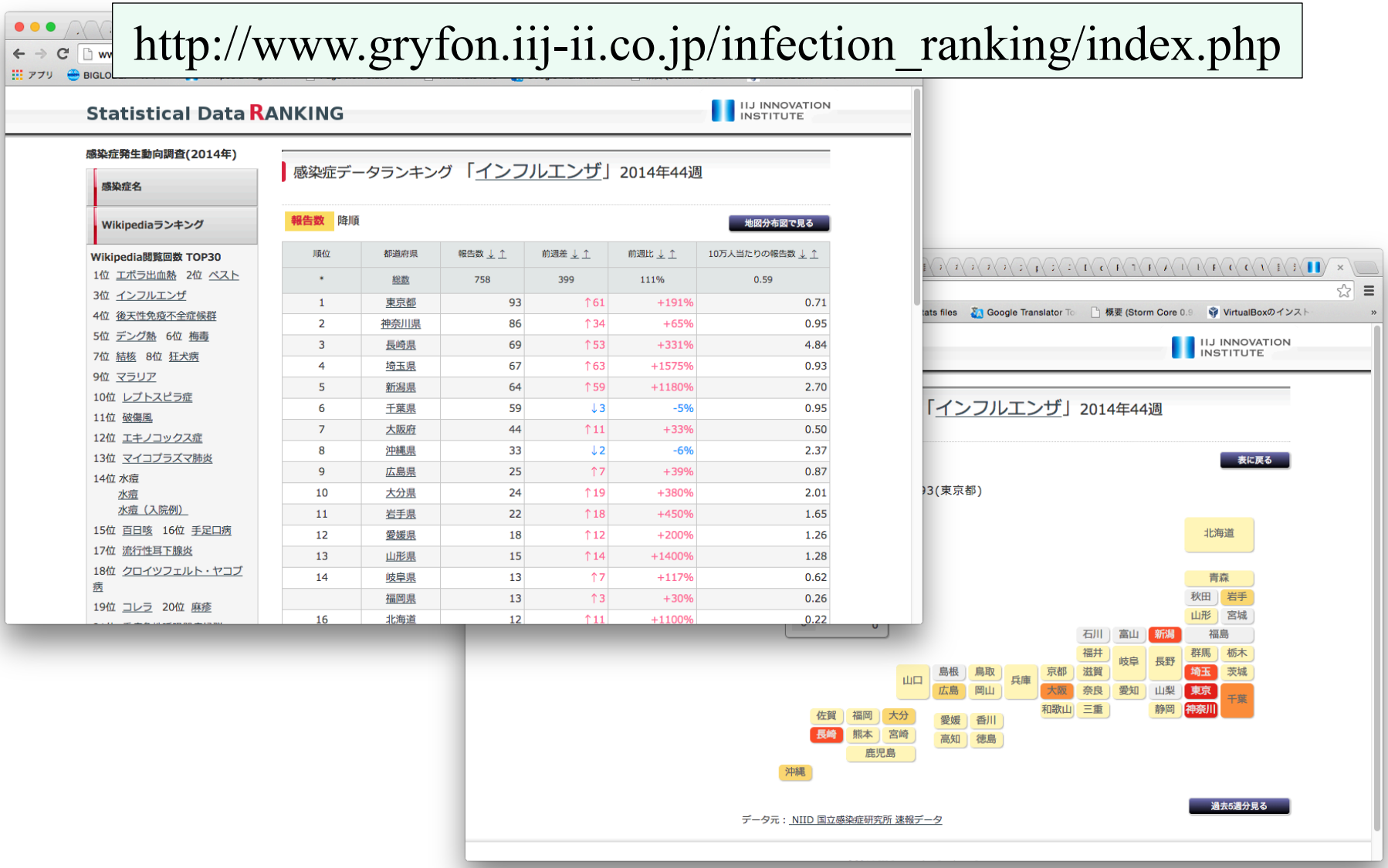
都道府県カテゴリランキング

Jタウンネットとのサービス連携

http://http://j-town.net/

感染症データランキング

http://www.gryfon.iij-ii.co.jp/infection_ranking/index.php



Statistical Data RANKING

IIJ INNOVATION INSTITUTE

感染症発生動向調査(2014年)

感染症名

Wikipediaランキング

Wikipedia閲覧回数 TOP30

1位 エボラ出血熱 2位 ペスト
3位 インフルエンザ
4位 後天性免疫不全症候群
5位 デング熱 6位 梅毒
7位 結核 8位 狂犬病
9位 マラリア
10位 レプトスピラ症
11位 破傷風
12位 エキノコックス症
13位 マイコプラズマ肺炎
14位 水痘
水痘
水痘（入院例）
15位 百日咳 16位 手足口病
17位 流行性耳下腺炎
18位 クロイツフェルト・ヤコブ病
19位 コレラ 20位 麻疹

感染症データランキング 「インフルエンザ」2014年44週

報告数 降順

地図分布図で見る

順位	都道府県	報告数	前週差	前週比	10万人当たりの報告数
*	総数	758	399	111%	0.59
1	東京都	93	↑61	+191%	0.71
2	神奈川県	86	↑34	+65%	0.95
3	長崎県	69	↑53	+331%	4.84
4	埼玉県	67	↑63	+1575%	0.93
5	新潟県	64	↑59	+1180%	2.70
6	千葉県	59	↓3	-5%	0.95
7	大阪府	44	↑11	+33%	0.50
8	沖縄県	33	↓2	-6%	2.37
9	広島県	25	↑7	+39%	0.87
10	大分県	24	↑19	+380%	2.01
11	岩手県	22	↑18	+450%	1.65
12	愛媛県	18	↑12	+200%	1.26
13	山形県	15	↑14	+1400%	1.28
14	岐阜県	13	↑7	+117%	0.62
	福岡県	13	↑3	+30%	0.26
16	北海道	12	↑11	+1100%	0.22

IIJ INNOVATION INSTITUTE

「インフルエンザ」2014年44週

表に戻る

93(東京都)

北海道 青森 秋田 岩手 山形 宮城 石川 富山 新潟 福島 福井 岐阜 長野 群馬 栃木 山口 島根 鳥取 兵庫 京都 滋賀 広島 岡山 大阪 奈良 愛知 山梨 埼玉 茨城 和歌山 三重 東京 千葉 静岡 神奈川 佐賀 福岡 大分 愛媛 香川 長崎 熊本 宮崎 高知 徳島 鹿児島 沖縄

過去5週分見る

データ元： NIID 国立感染症研究所 速報データ

データ活用ニーズを持つユーザとの協業の実際

- 目標は明確である必要はあるが・・・
  - 具体的過ぎるとすぐに行き詰る
  - 最初の成果を披露するまでの半年間の大半はアイデア出しに終始
  - どこかで踏ん切りをつけないと際限なく「アイデア出し」

- 1度成果を公開すると・・・
  - その後のアプローチの軸になる
  - 建設的な改善案が比較的スムーズに提案される
  - が、どこかで収束を迎える：ソースデータの利用アイデアが出尽くす

- ビッグデータはデータ活用の一部しか担えない
  - ビッグデータありきのデータ活用はエビデンスの問題で破綻する
  - 最初はビッグでないデータを活用して目標に直結する解析結果を得る
  - 初期の解析結果を「詳細化・補強する情報」をビッグデータから得る

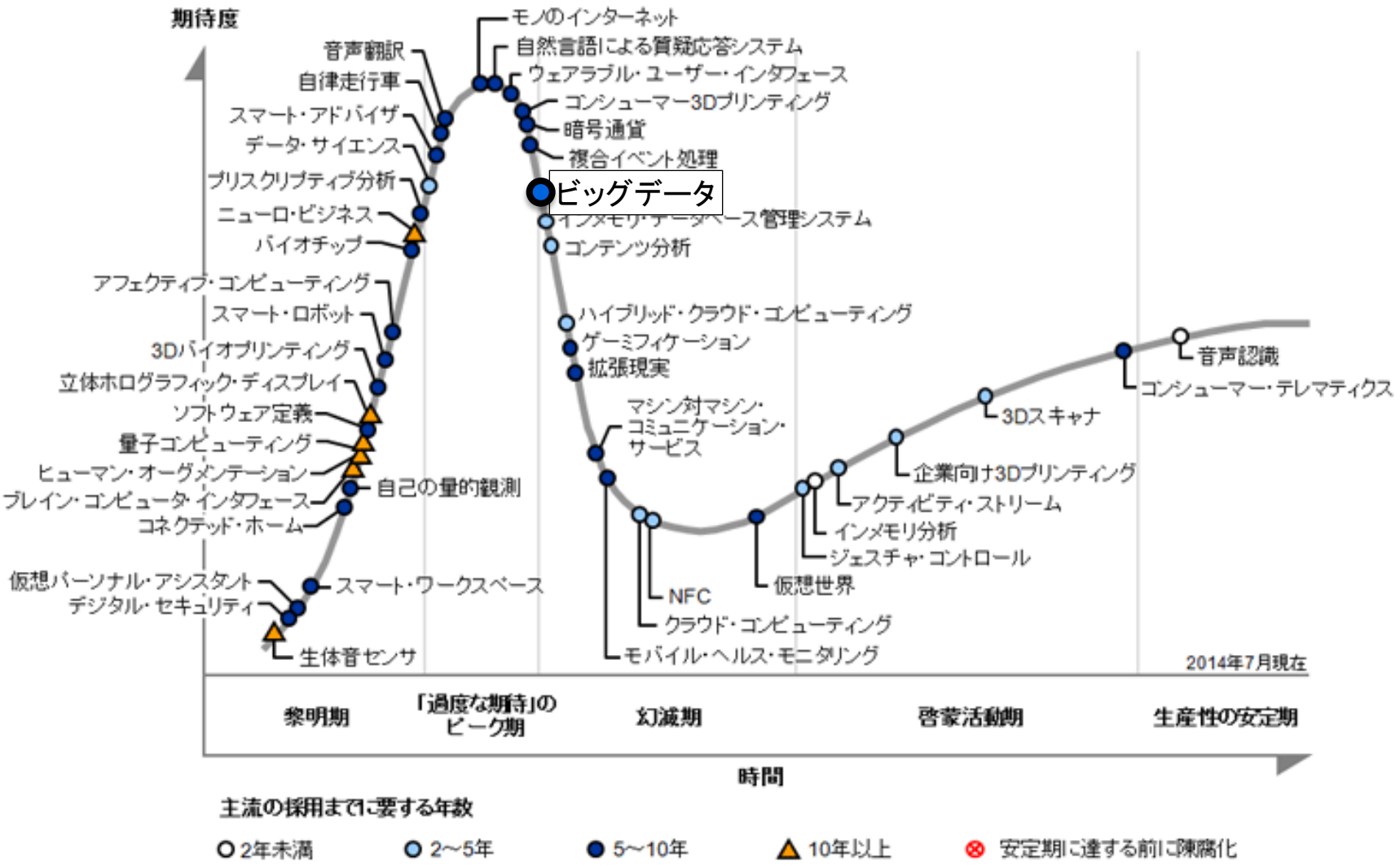
分かりやすいと評判のビッグデータの説明・・・ふたたび

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

技術動向： ガードナーのハイプ・サイクル（1）

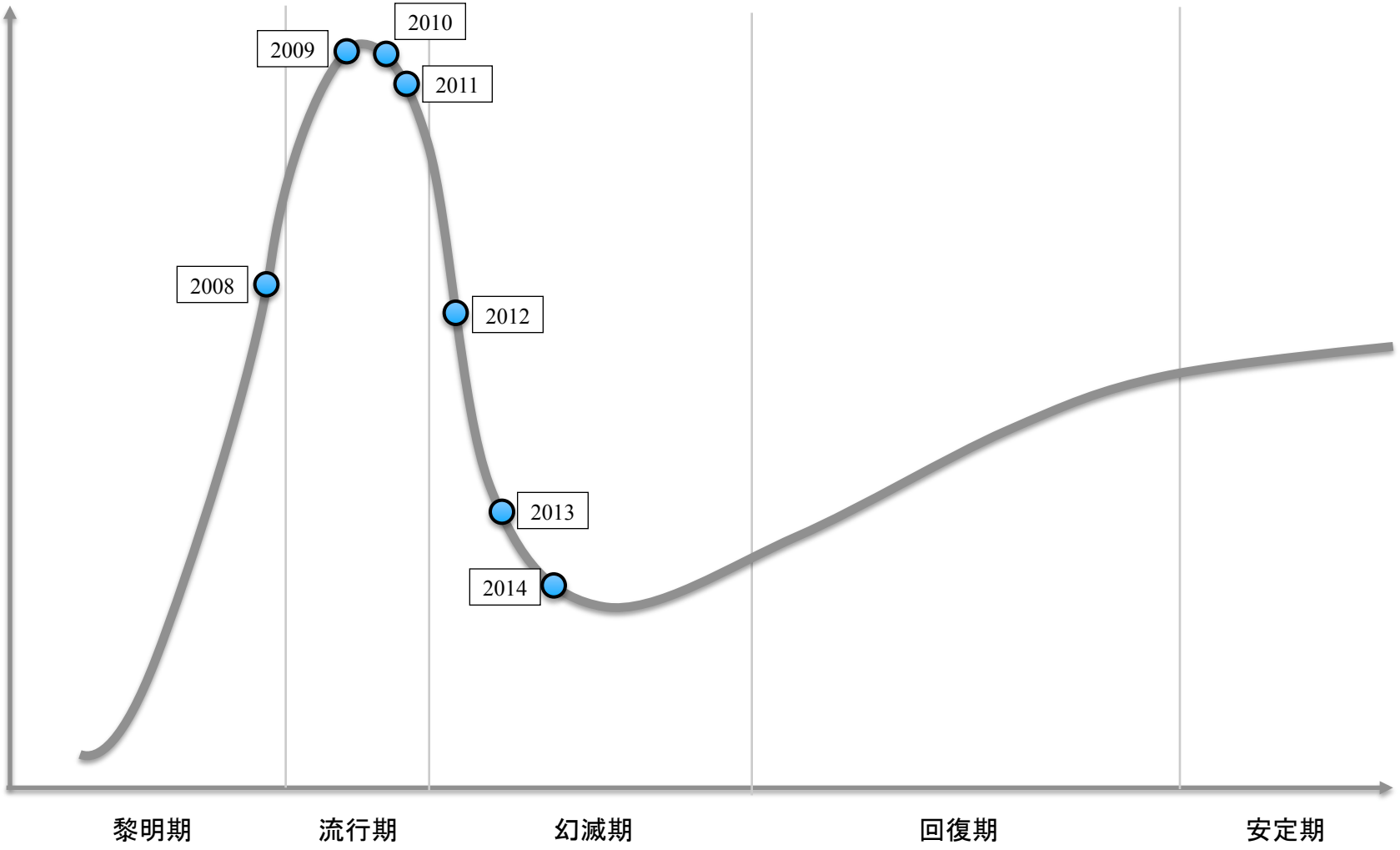


技術動向： ガードナーのハイプ・サイクル（2）

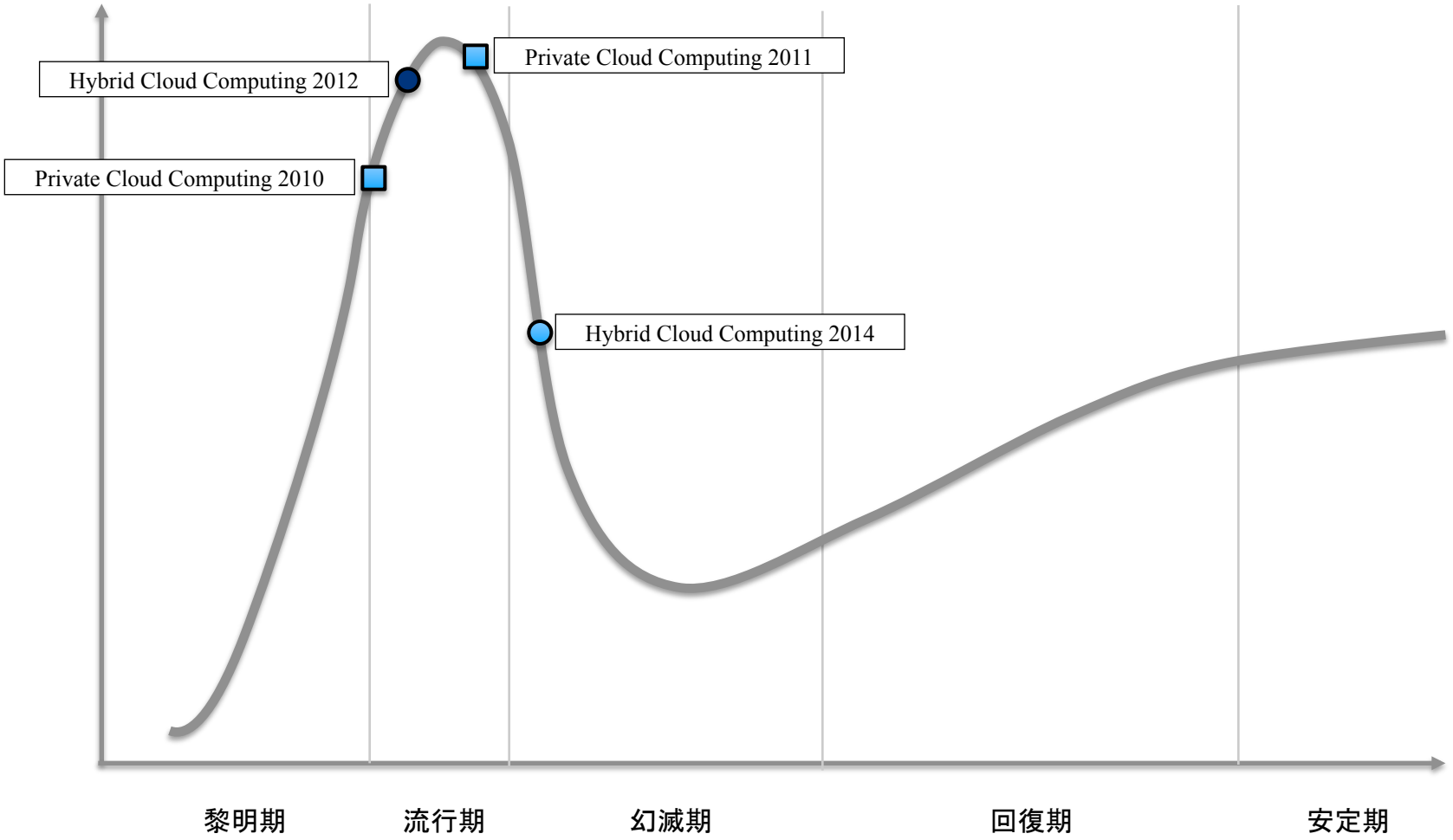
- ハイプ・サイクル： 英ガードナーが案出したグラフ図
  - 特定の技術の成熟度、採用度、社会への適用度を示す
  - 毎年8月頃に公表される

- 5つの段階：
  - 黎明期（技術の引き金、Technology Trigger）
    - 「技術の引き金」またはブレークスルー（飛躍的前進）から始まる
  - 流行期（過剰期待の頂、Peak of Inflated Expectations）
    - 世間の注目が大きくなり、過度の興奮と非現実的な期待が生じる
  - 幻滅期（幻滅のくぼ地、Trough of Disillusionment）
    - 技術は過度な期待に応えられず急速に関心が失われる
  - 回復期（啓蒙の坂、Slope of Enlightenment）
    - いくつかの事業は、その利点と適用方法を理解するようになる
  - 安定期（生産性の台地、Plateau of Productivity）
    - 広範に宣伝され受け入れられるようになる

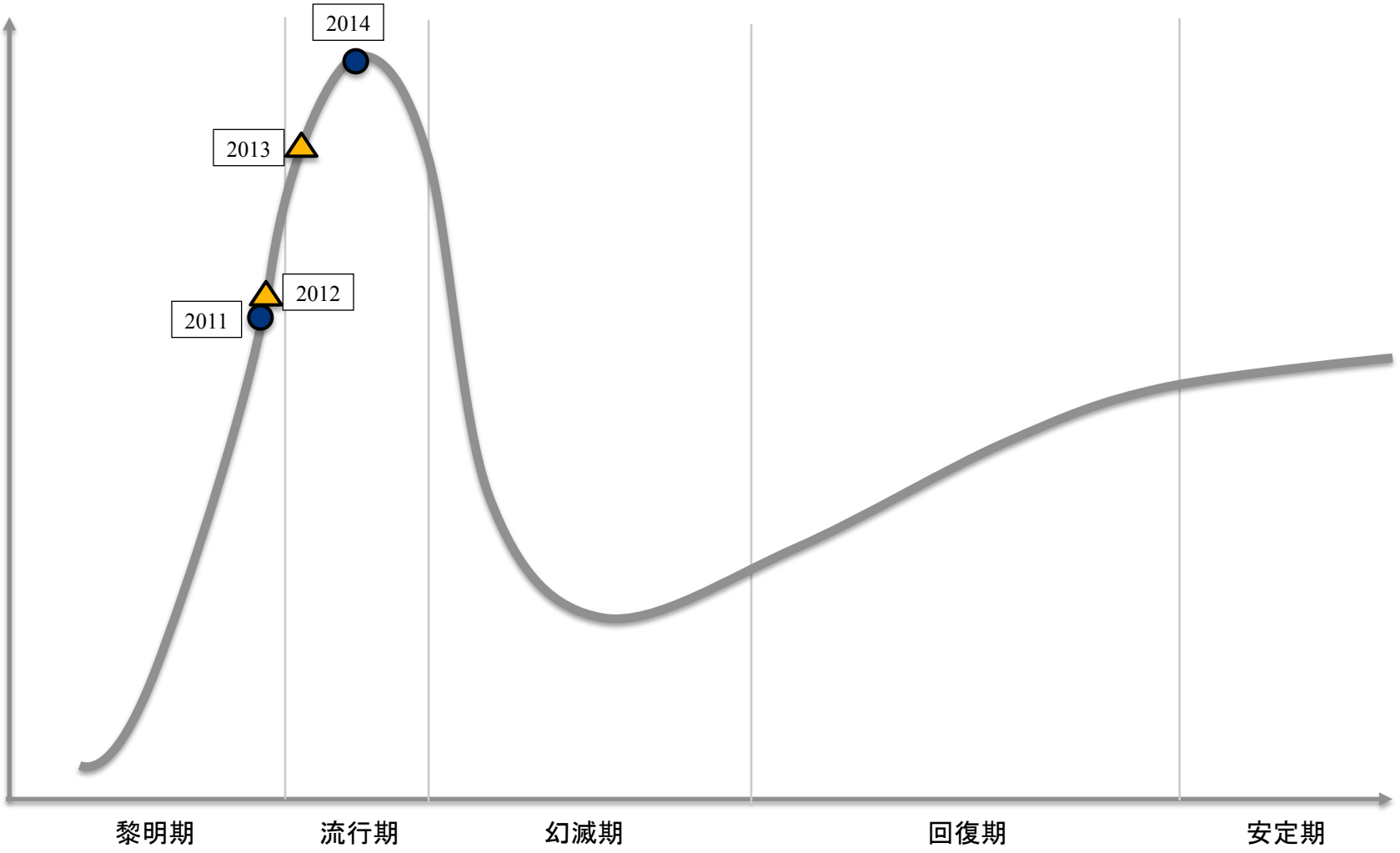
クラウドコンピューティング



クラウドプラットホーム



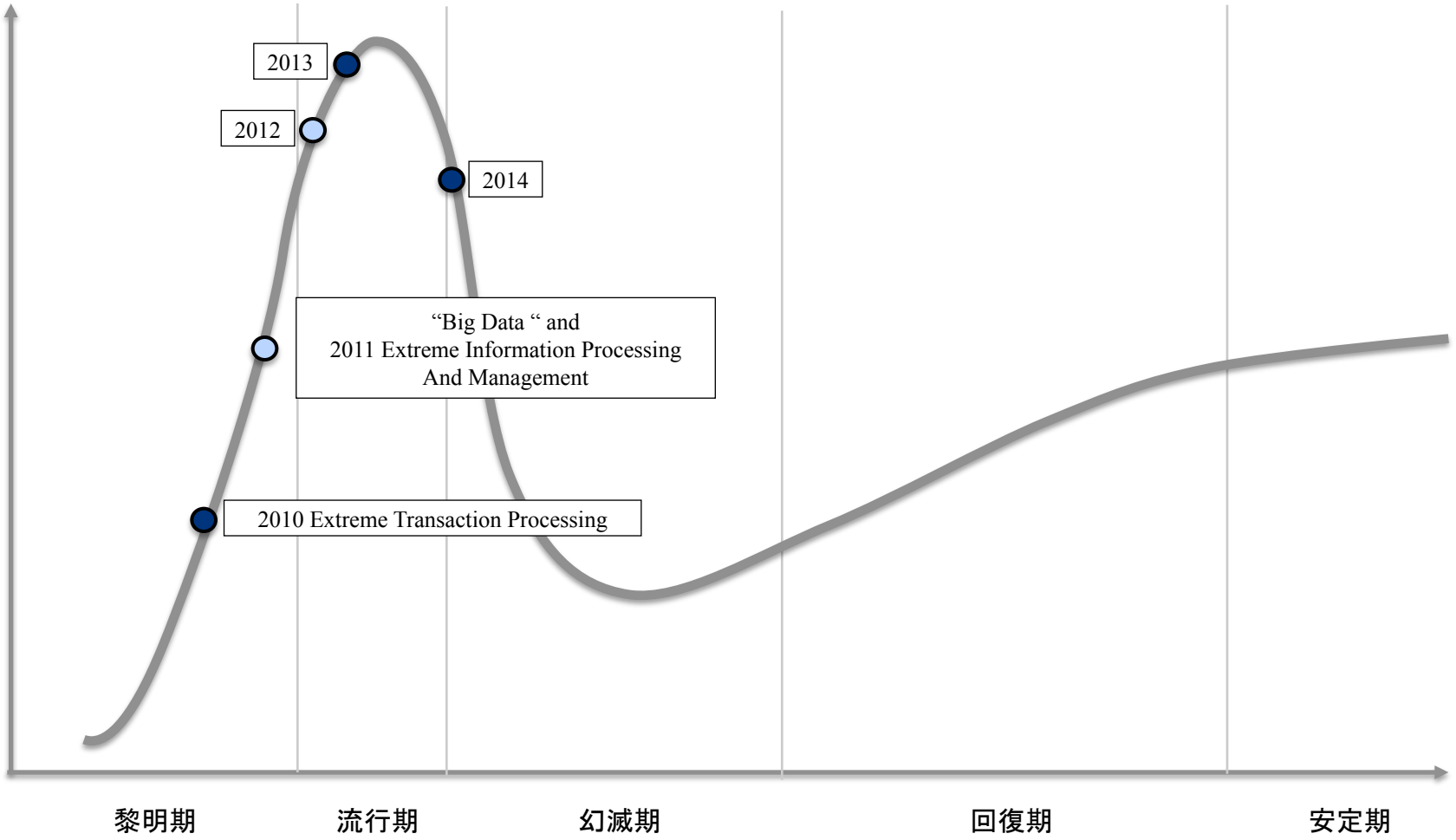
Internet of Thing（IoT）



各年のハイプ・サイクルの変化の追跡（1）

- 下記のトピックについて2008年～2014年の位置を追跡
  - 「クラウドコンピューティング」
    - 2008年に黎明期で登場 → バズワードとして一般に認知された
    - 2009年～2011年は流行期に存在 → 3年間は非常に長い
    - 2012年～2014年は幻滅期に存在 → 2014年は幻滅期の底
  - クラウドプラットホーム
    - 2010年～2011年は「プライベートクラウド」が流行期に存在
    - 2012年は「ハイブリッドクラウド」が流行期に登場 → 2013年には消滅
    - 2014年は「ハイブリッドクラウド」が幻滅期に再登場
  - 「Internet of Thing（IoT）」
    - 2011年に黎明期に登場 → 主流に定着まで5～10年
    - 2012年も黎明期に居座り → 主流に定着まで10年以上に変わる
    - 2013年は流行期に移動 → 主流に定着まで10年以上のまま
    - 2014年は流行期の頂点 → 主流に定着まで5～10年に戻る

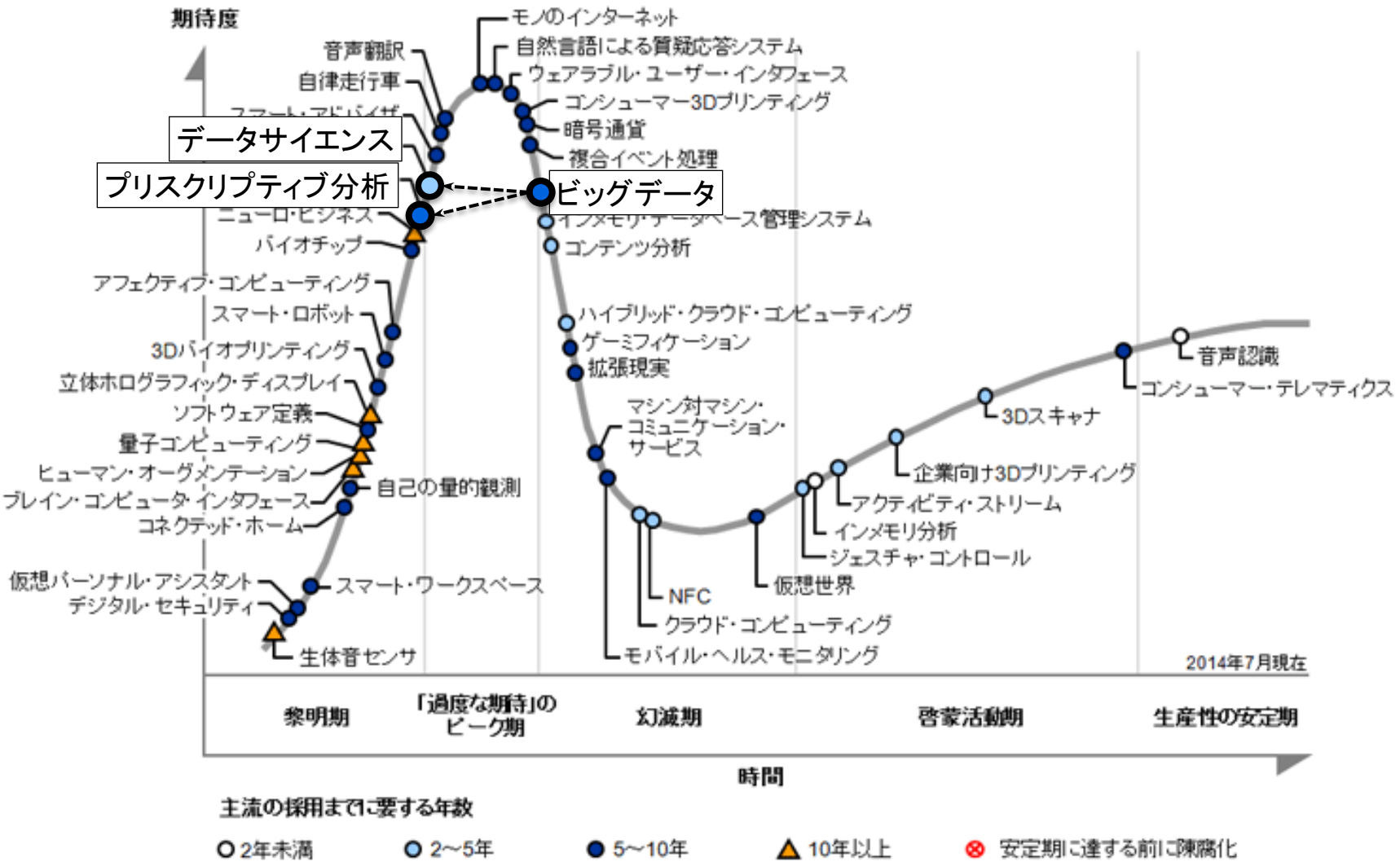
ビッグデータ


2013
2012
2014
“Big Data “ and
2011 Extreme Information Processing
And Management
2010 Extreme Transaction Processing
黎明期
流行期
幻滅期
回復期
安定期

各年のハイプ・サイクルの変化の追跡（2）

- 下記のトピックについて2008年～2014年の位置を追跡
  - 「ビッグデータ」
    - 2010年に「大規模トランザクション処理」として黎明期に登場
    - 2011年は「ビッグデータと大規模情報処理」として黎明期に存在
    - 2012年～2013年は「ビッグデータ」として流行期に移動
    - 2014年は幻滅期に移動
      - 技術は過度な期待に応えられず急速に関心が失われる？
- 再度、2014年のハイプサイクルのよく見てみると
  - 黎明期・流行期にあるキーワードには・・・
    - 解析結果の利用が想定される「スマートXXX」との名前のドットがちらほら
    - データ分析の側面をさらに指向する技術（キーワード）も出現
- 「ビッグデータ」動向の（少々強引な）解釈
    - 当初は「大規模データ処理」から派生したコンセプト
    - 2012年～2013年には「過度の興奮と非現実的な期待が生じた」
    - 現在はデータ分析とその技術に関心が移りつつある？

ビッグデータの現在（2014年）



ガードナーのハイプ・サイクル2014の考察

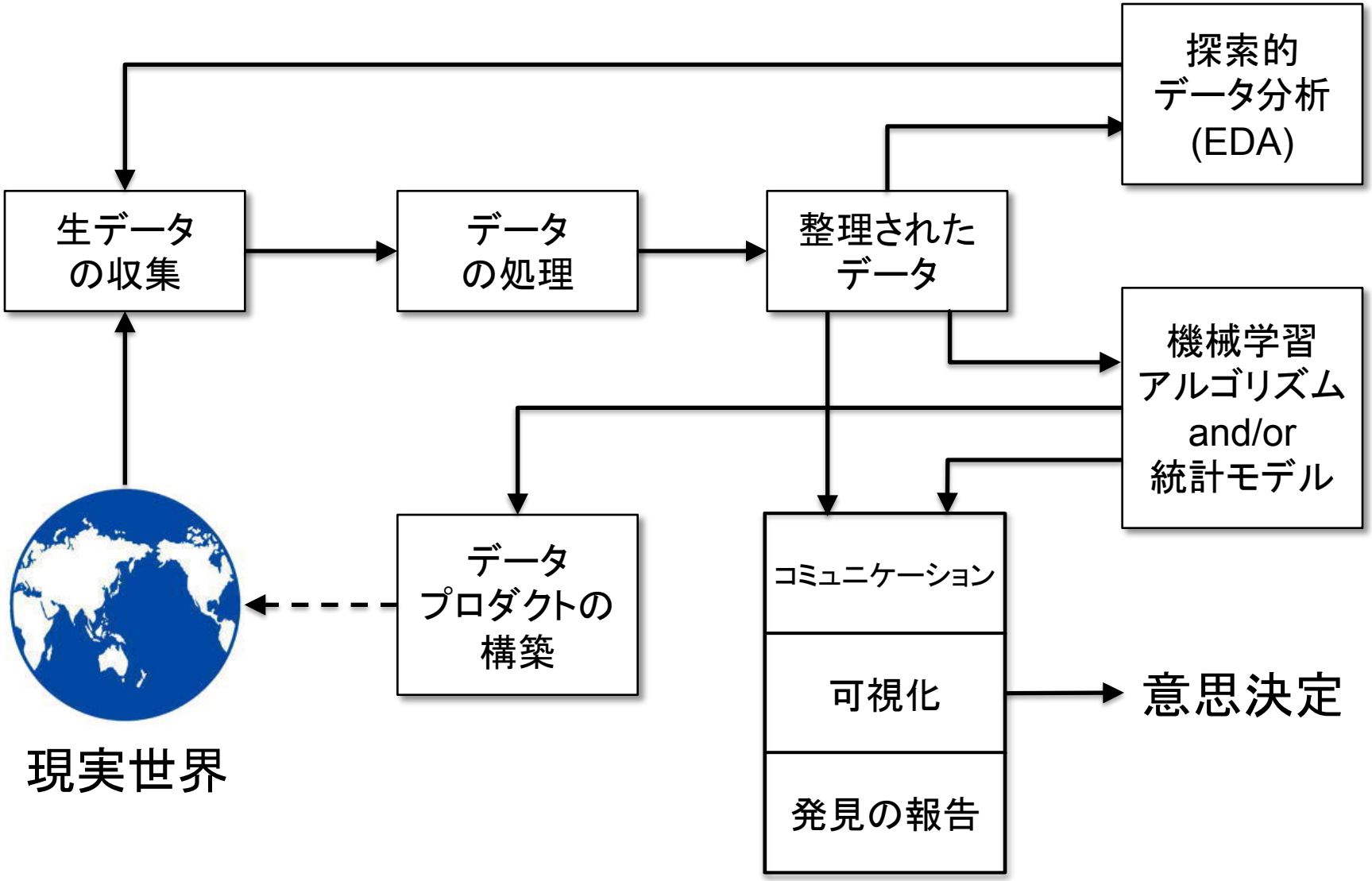
- ハイプサイクルにおける「ビッグデータ」の動向
  - 2011年に「大規模データ処理」から派生したコンセプト
  - 2012年～2013年には「過度の興奮と非現実的な期待が生じた」
  - 2014年には幻滅期に突入
    - が、コンセプチュアルなキーワードとしてハイプサイクルに残り続ける？
- 世間の関心はデータ分析とその技術に移りつつある
  - データサイエンス： より分析側面にフォーカスしたキーワード
  - プリスクリプティブ分析： ビッグデータ活用の具体化を示すキーワード
- 得られる知見： 経営学におけるビッグデータ活用の事例
  - ビジネス分析での分析フェーズの定義は共通フレームワークになる？
    - 記述的： 現在の状態を分析
    - 予測的： 将来の予測を分析
    - 規範的： 課題解決手段を分析 → 分析目的により変わってくる

今後注目を集めるキーワード(1)

- データサイエンス： 2014年に流行期で登場
  - 定義： データから知識を抽出する“科学”
    - 数学、統計学、情報工学にまたがる領域で研究された技術と理論を応用
      - 信号処理、確率モデル、機械学習、統計学習、コンピュータ・プログラミング、データ・エンジニアリング、パターン認識、ビジュアライゼーション、不確実性モデリング、データウェアハウス、ハイパフォーマンス・コンピューティング
    - ビッグデータのみを対象としているわけではない
      - 大規模データを容易に扱えることがデータサイエンスの特徴ではあるが

- コメント
  - 用語としては1960年代に登場していた
    - クラウドコンピューティングやビッグデータへの関心から再注目を集めた
  - 極めてコンセプチュアルなキーワード
    - 「クラウドコンピューティング」や「ビッグデータ」と同様に
    - 既にバズワードとして広く認知されている
    - ハイプサイクルに長期間存在し続けると予測される

データサイエンスのプロセス



Schtt, O'neil, "Doing Data Science", 「データサイエンス講義」 P.45

探索的データ分析（EDA: Exploratory Data Analysis）

- 探索的データ解析（Exploratory Data Analysis）
  - 統計学者 J.W.Tukey によって提唱された
  - データを視覚的に捉え、データ自身に仮説を語らせるアプローチ
  - 確証的データ解析（Confirmatory Data Analysis ）と対立する

- EDAのプロセス
  - データの理解： 分散の比較（Box Plot）
  - サニティーチェック： 欠損値や外れ値の確認
  - データの要約

- コメント
  - 昨年度講演の「発見的な探索」と同義？
  - 現在の “Wikipediaランキング” はこの分析の範疇内に閉じている？

今後注目を集めるキーワード(2)

- プリスクリプティブ分析:　2013年から黎明期で登場
  - 定義:　課題解決のための手段の分析
  - 記述的分析、予測的分析を前提として、予測を達成するために必要な「意思決定」とその意思決定に関連する「影響」を分析する
- コメント
  - 経営学(経営工学)の研究領域での用語
    - この領域でのビッグデータ活用を示すバズワードに見える
  - IBMが提唱するビジネス分析(BA)の3つのフェーズの最終フェーズ
    - 記述的分析(Descriptive Analytics)
    - 予測的分析(Predictive Analytics)　プリディテクティブ分析
    - 規範的分析(Prescriptive Analytics) プリスクリティブ分析
  - プリディテクティブ分析は既に2012年に安定期に到達
    - プリスクリティブ分析はそれを超える分析技術として注目されつつある
  - ベンチャー企業の Ayata がこのキーワードの商標権を持つ
    - 過去の研究をベースにした事業化が活発化しつつある

IBMによるデータ分析の定義

Understanding analytics

Definitions, sample applications and opportunities, and underlying technologies

	Descriptive	Predictive	Prescriptive
	What **HAS** happened?	What **COULD** happen?	What **SHOULD** happen?
What the user needs to **DO**	• **Increase** asset reliability • **Reduce** labor and inventory costs	• **Predict** infrastructure failures • **Forecast** facilities space demands	• **Increase** asset utilization • **Optimize** resource schedules
What the user needs to **KNOW**	• The **number and types** of asset failures • Why **maintenance costs** are high • The value of the **materials inventory**	• How to **anticipate failures** for specific asset types • When to **consolidate underutilized** facilities • How to **determine costs** to improve service levels	• How to **increase** asset production • Where to **optimally route** service technicians • Which strategic facilities plan provides the **highest long-term utilization**
How analytics gets **ANSWERS**	• **Standard reporting** - What happened? • **Query/drill down** - Where exactly is the problem? • **Ad hoc reporting** - How many, how often, where?	• **Predictive modeling** - What will happen next? • **Forecasting** - What if these trends continue? • **Simulation** - What could happen? • **Alerts** - What actions are needed?	• **Optimization** - What is the best possible outcome? • **Random variable optimization** - What is the best outcome given the variability in specified areas?
What makes this analysis **POSSIBLE**	• Alerts, reports, dashboards, **business intelligence**	• Predictive **models**, forecasts, statistical **analysis**, scoring	• Business rules, organization **models**, comparisons, **optimization**

Business value

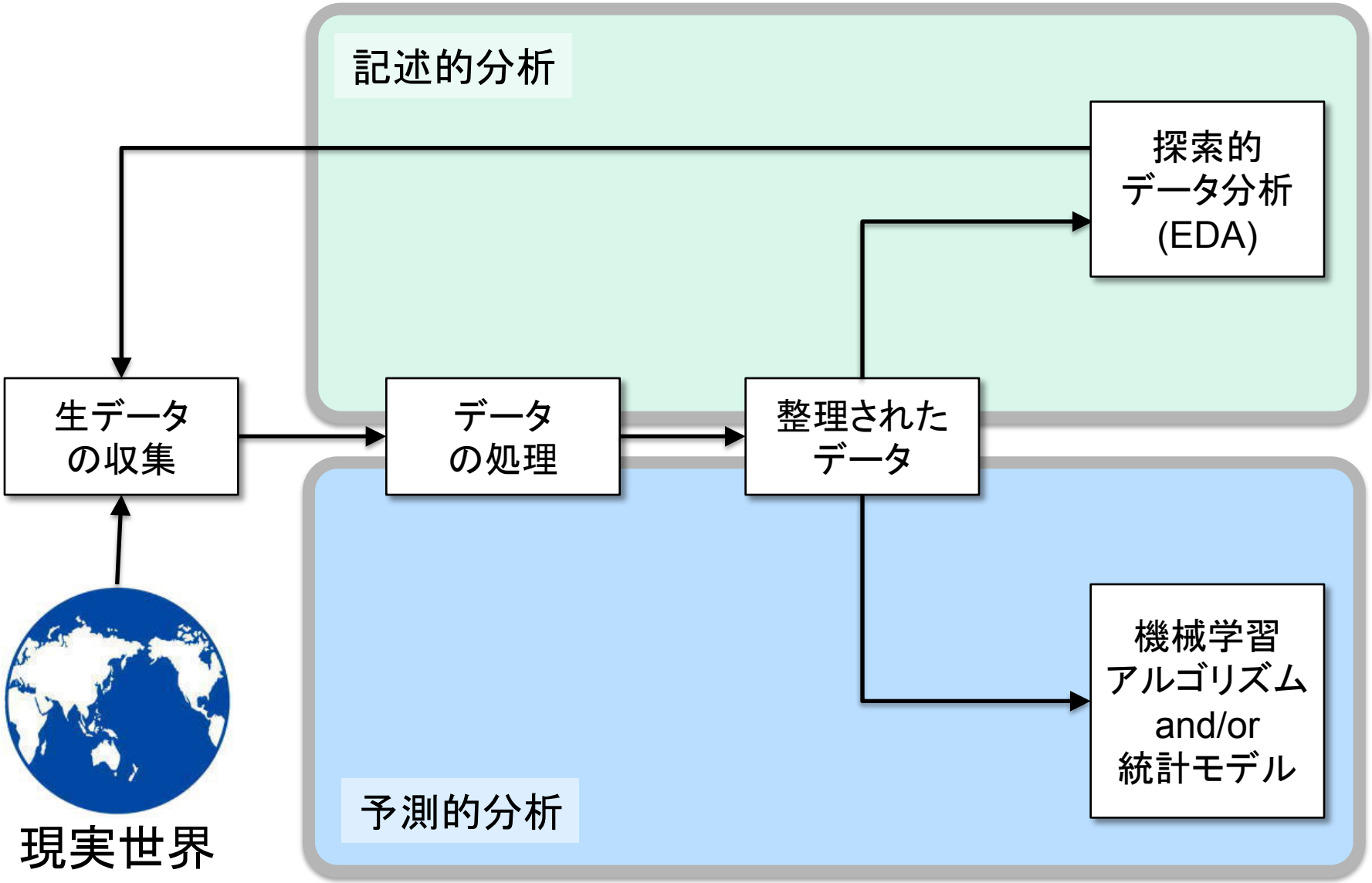
"Descriptive, predictive, prescriptive: Transforming asset and facilities management with analytics", http://docs.caba.org/documents/IS/IS-2014-49.pdf

IBMによるデータ分析の理解

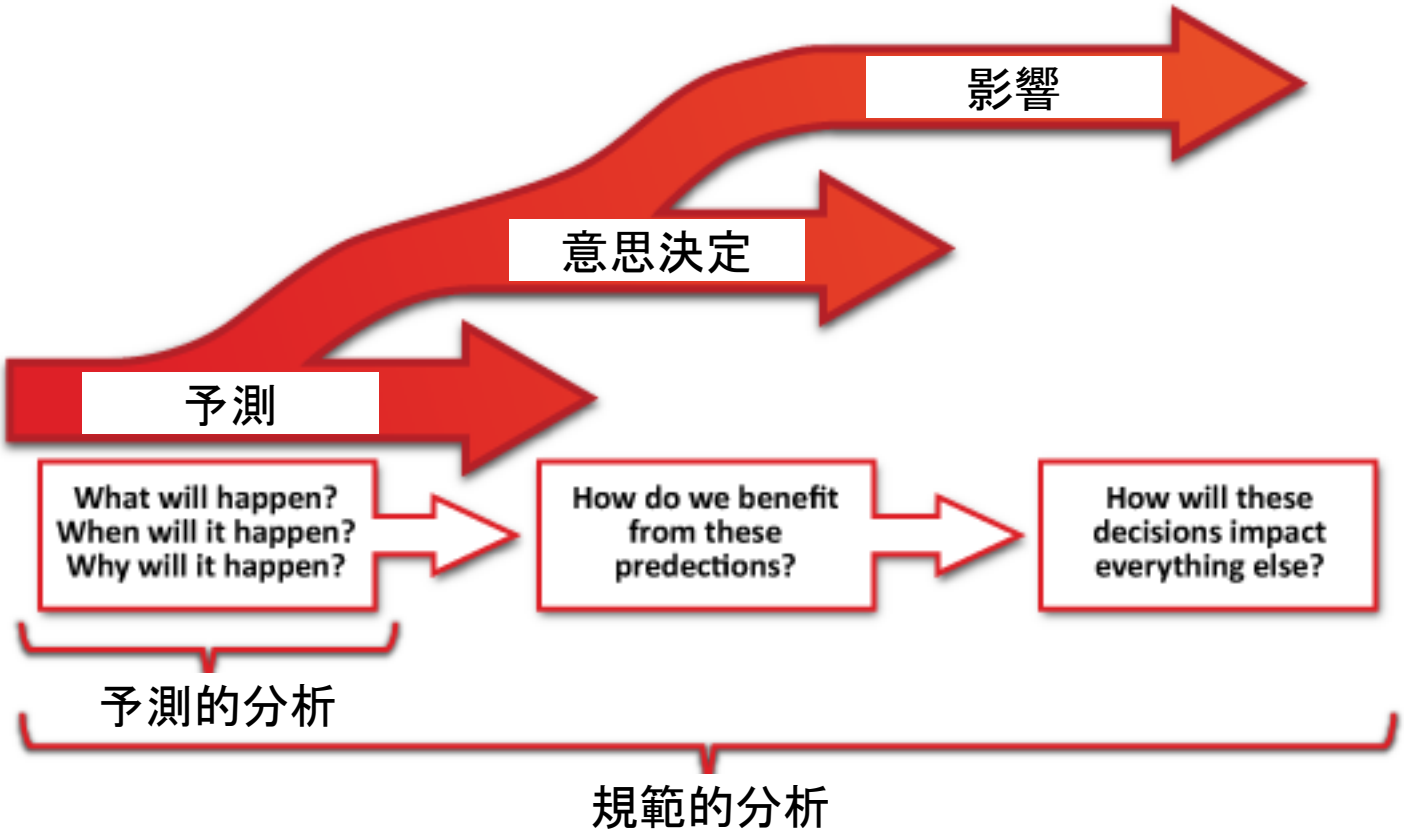
	記述的分析	予測的分析	規範的分析
	何が起こったのか？	何が起こり得るのか？	何が起こるべきなのか？
ユーザが行うたいこと	・資源の信頼性の拡大 ・労働力と在庫の圧縮	・インフラストラクチュア障害予測 ・設備の空間的な要求の予測	・資源利用の拡大 ・資源スケジューリングの最適化
ユーザーが知りたいこと	・資源の障害の数と種類 ・保守コスト増大の理由 ・在庫の価値	・特定の種類の資源障害を予想する方法 ・十分に利用されなかった設備を統合するべきタイミング ・サービスレベルを改善するコストを決定する方法	・資源の創出の増大する方法 ・サービス専門家を送るべき最適な場所 ・最も長期間活用できる戦略的な設備計画
答えを得るための分析方法	標準的なレポート－何が起こったのか？ クエリ/ドリルダウン－問題の正確な所在 アドホックなレポート－量、頻度、場所の詳細	・予測モデリング－次に何が起こるか？ ・予測－この傾向が続くとしたらどうするか？ ・シミュレーション－何が起こり得るのか？ ・警告－どんなアクションが必要か？	・最適化－起こりうる最良の結果は何か？ ・確率変数の最適化－特定の領域にて与えられた変動の範囲での起こり得る最良の結果は何か？
この分析を可能にする手段	・アラート、レポート、ダッシュボード、ビジネス・インテリジェンス	・予測モデリング、予測、統計的分析、スコアリング	・ビジネスルール、組織モデル、比較、最適化

“Descriptive, predictive, prescriptive: Transforming asset and facilities management with analytics”, http://docs.caba.org/documents/IS/IS-2014-49.pdf

記述的分析と予測的分析



規範的分析

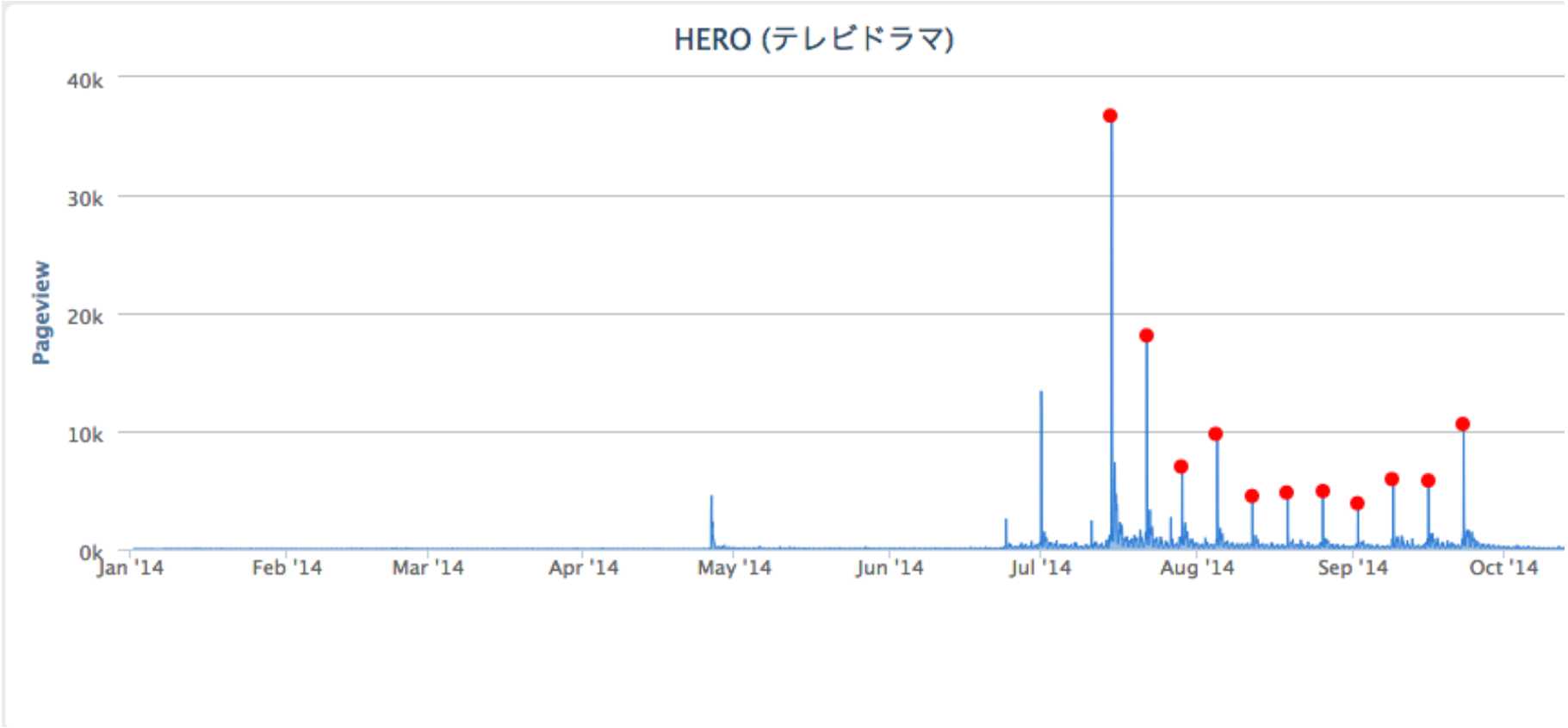

影響
意思決定
予測
What will happen?
When will it happen?
Why will it happen?
How do we benefit from these predections?
How will these decisions impact everything else?
予測的分析
規範的分析

Project Gryfon でのデータ分析への取り組み（1）

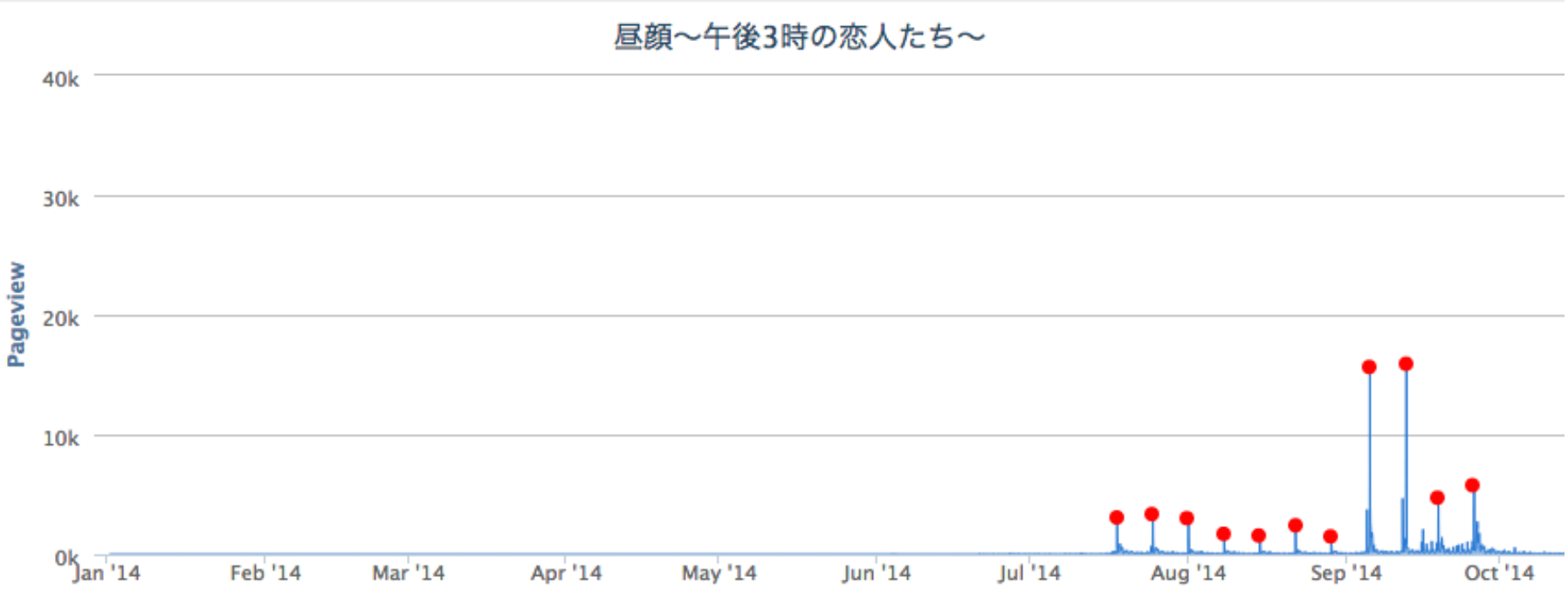
- Wikipedia PVC の時系列分析
  - 昨年度の「連続ドラマに着目した WikipediaPVC 解析」の拡大版
    - Wikipedia 日本語版全ページ（ドラマページだけでなく）を対象とする
    - その他のソーシャルメディア（Twitter）も取り込んだ要因分析
    - Apache Storm を利用したリアルタイム化と省力化・自動化

- Wikipedia PVC変動による分析手法の検討
  - 2014年第3クールの平均視聴率上位5件のドラマを使った試行
    - HERO 21.3%
    - 昼顔～平日午後3時の恋人たち～　13.9%
    - ST 赤と白の捜査ファイル　11.3%
    - 金田一少年の事件簿N　10.5%
    - ゼロの真実 監察医・松本真央　10.5%

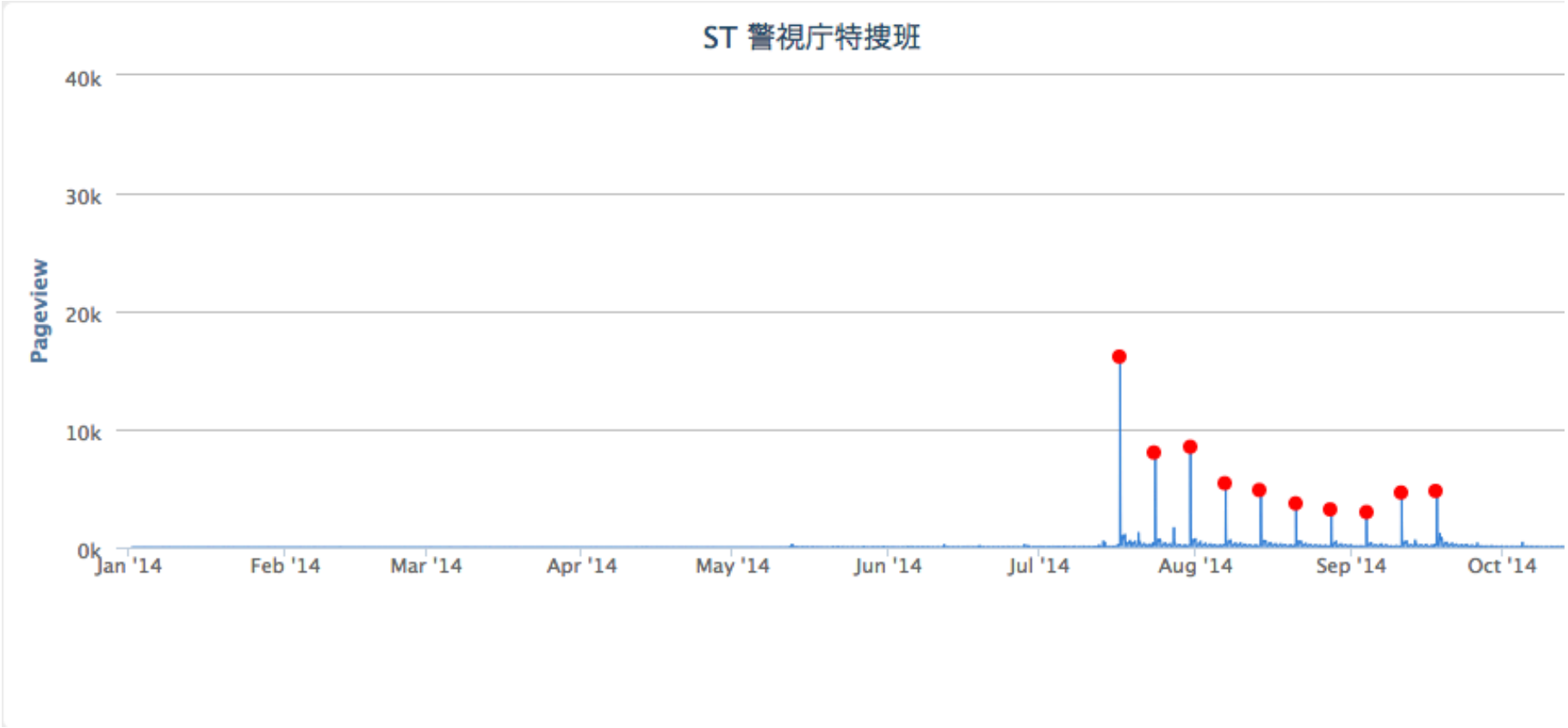
HERO　第2シーズン（21.3%）



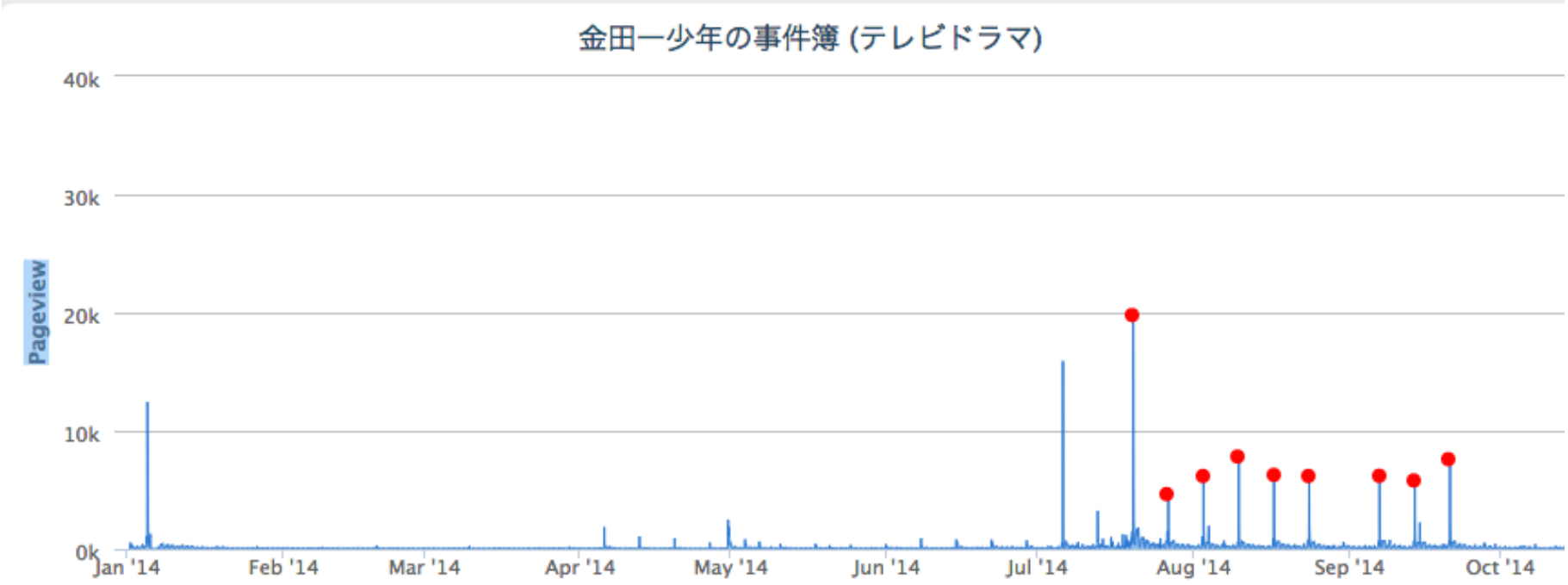
昼顔～平日午後3時の恋人たち～（13.9%）



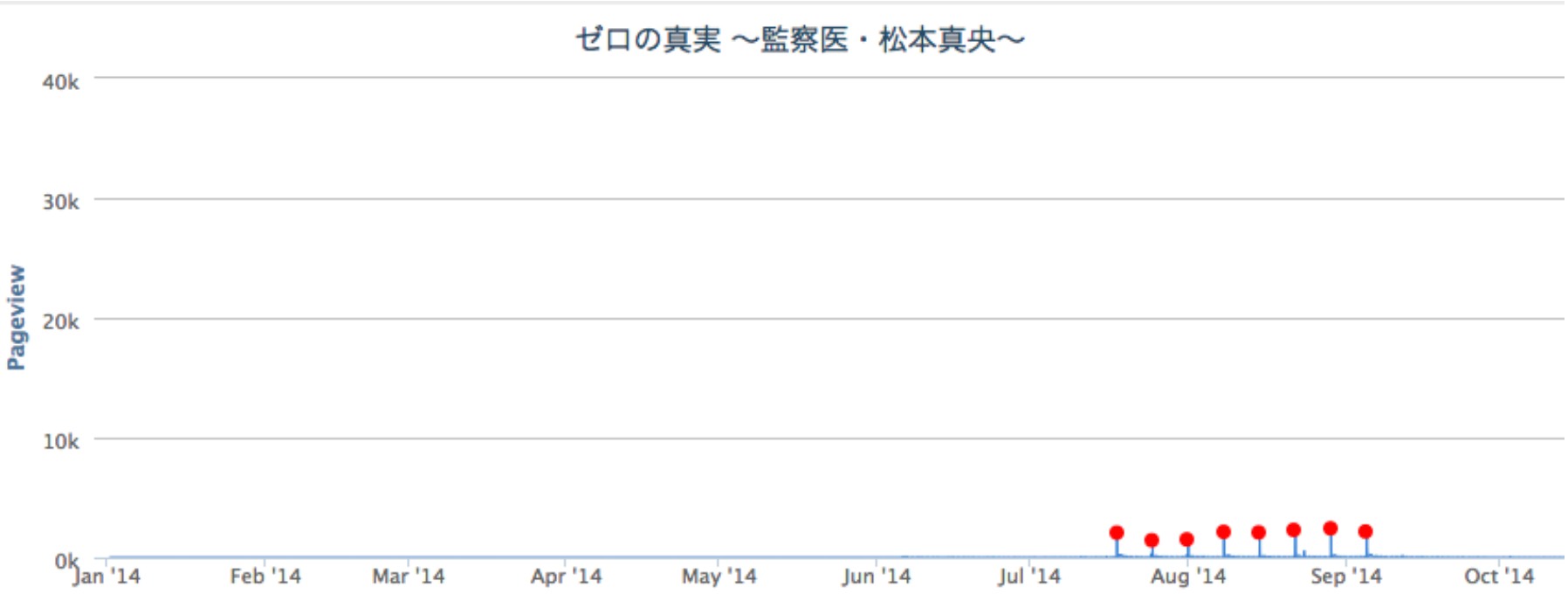
ST 赤と白の捜査ファイル（11．3%）



金田一少年の事件簿N（10.5%）



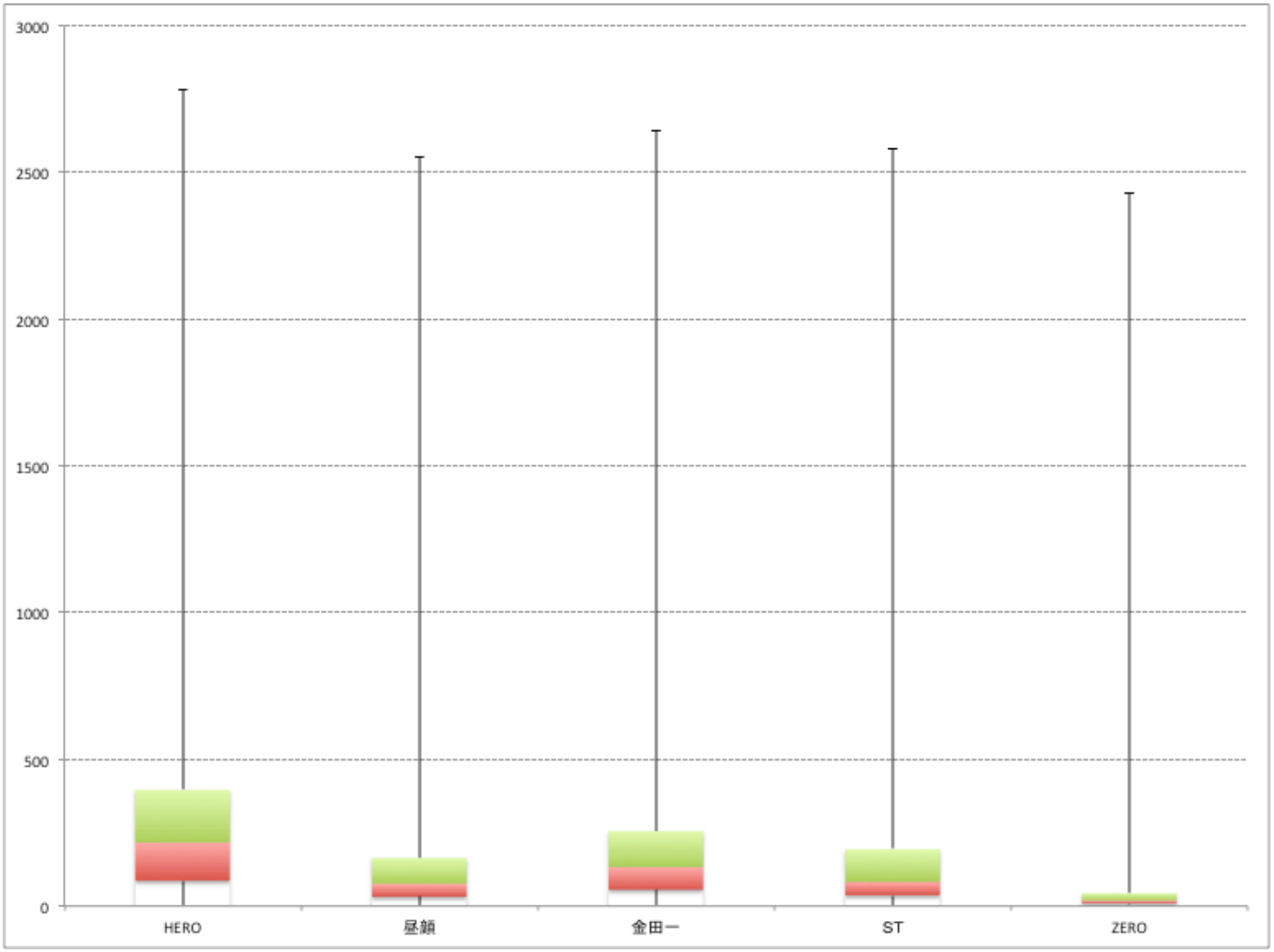
ゼロの真実 監察医・松本真央（10.5%）



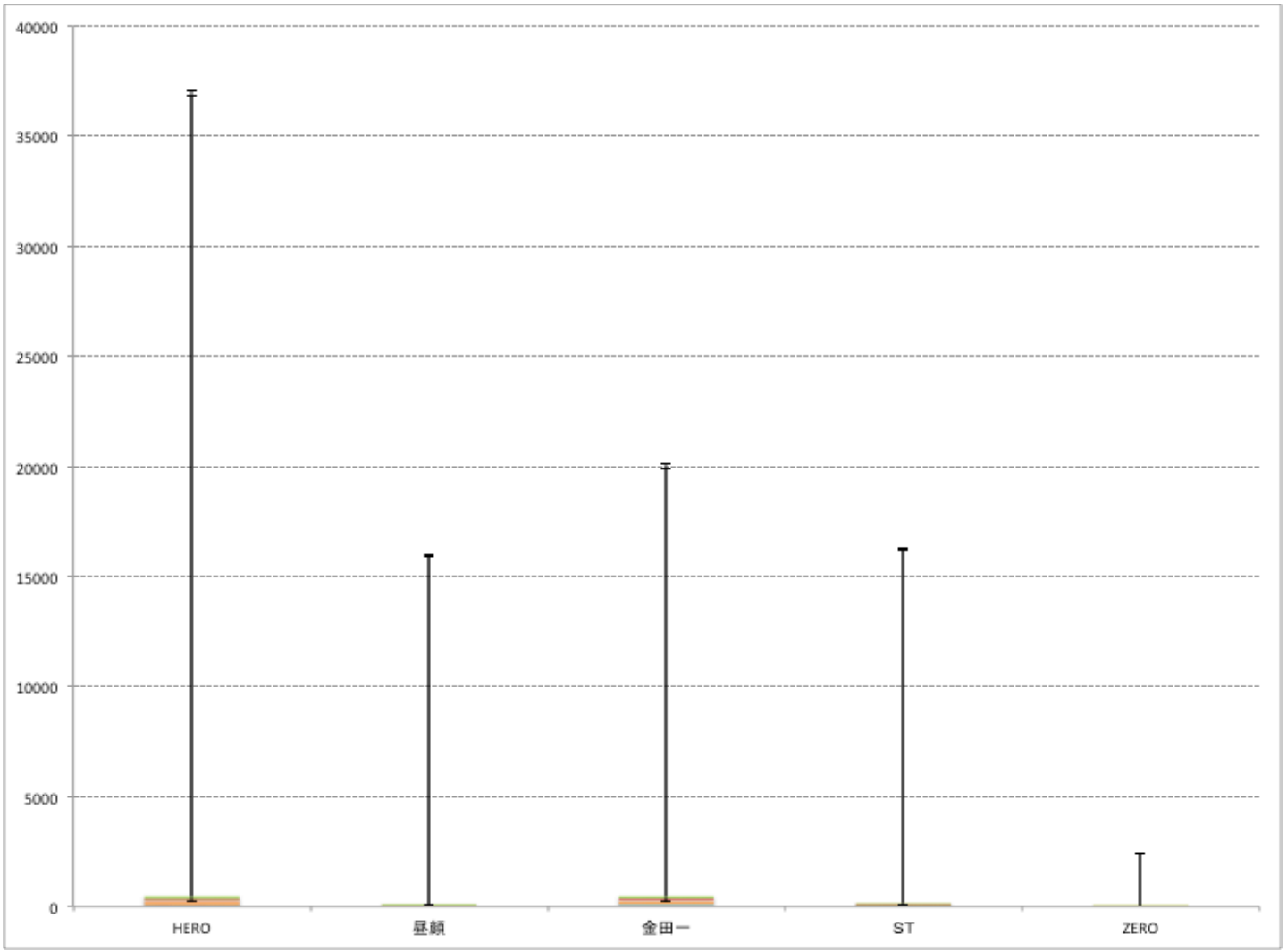
Project Gryfon でのデータ分析への取り組み（2）

- 探索的データ分析（記述的分析）
  - Wikipedia PVC変動の時系列グラフ
    - PVC変動はドラマ放映と連動している
    - ドラマ放映以外にもPVCのピークが発生している
      - 制作発表＆旧シリーズの再放送など
  - Box Plot によるPVC値の分布の把握
    - 全体の50%は比較的小さな値でまとまる
      - ドラマ放映時は瞬間的に最大値周辺まで跳ね上がるが持続しない
    - ドラマの放映クール（期間）以外は中央値は限りなく0に近づくケース有り
      - ドラマ放映期間しか関心を持たれない？
- Wikipedia PVC 分析に関する考察
  - PVC値の瞬間的変動が発生した期間を検出する
    - 社会的イベントの発生時期を把握
    - 一般的な異常値検出アルゴリズムが適用可能？
  - 任意の期間のPVC値の分布を把握する
    - 対象トピックの社会的認知度を示す指標になり得る？

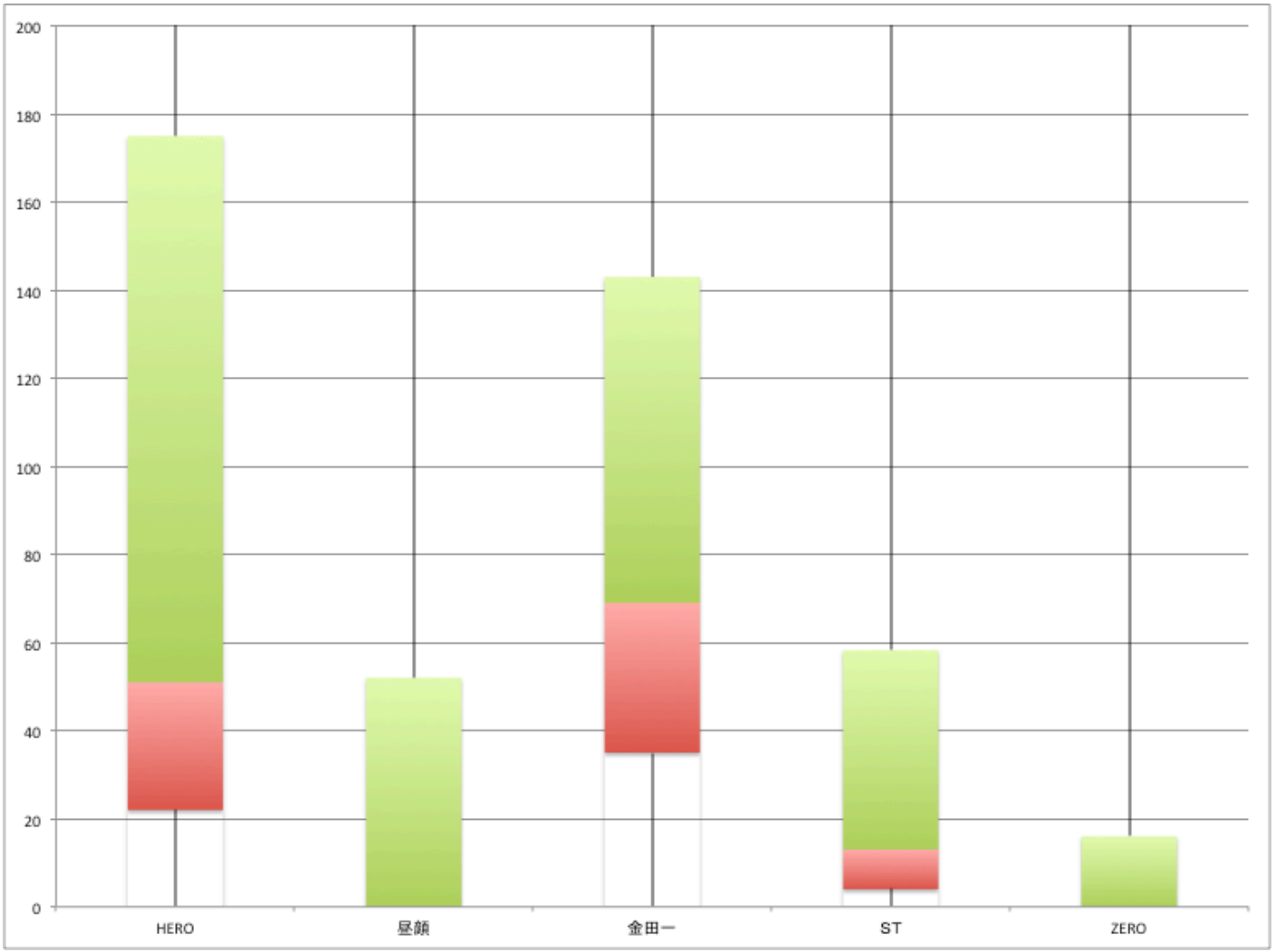
ドラマPVC Box Plot　2014年7月～9月



ドラマPVC Box Plot　2014年1月～10月（1）



ドラマPVC Box Plot　2014年1月～10月（2）



ドラマPVC Box Plot　データ

2014 07-09	HERO	昼顔	金田一	ST	ZERO
最小値	0.00	0.00	0.00	0.00	0.00
第1四分位	86.75	32.00	56.00	37.00	9.00
中央値	217.00	76.00	133.00	83.00	19.00
第3四分位	396.25	164.25	255.25	194.00	44.00
最大値	36733.00	15917.00	19799.00	16163.00	2432.00

2014 01-10	HERO	昼顔	金田一	ST	ZERO
最小値	0.00	0.00	0.00	0.00	0.00
第1四分位	22.00	0.00	35.00	4.00	0.00
中央値	51.00	0.00	69.00	13.00	0.00
第3四分位	175.00	52.00	143.00	58.25	16.00
最大値	36733.00	15917.00	19799.00	16163.00	2432.00

Project Gryfon でのデータ分析への取り組み(3)

- 統計モデルによるデータ分析(予測的分析)
  - PVC値の瞬間的変動が発生した期間を検出する
- ChangeFinder
  - 自己相関モデル(ARモデル)により異常値を検出するアルゴリズム
    - 時系列で変化するデータを入力すると異常性を示すスコア値を算出する
  - HERO の PVC データを入力するも・・・
    - スコア値が発散し異常終了
  - Rの関数 ar を使って HERO の PVC データの自己相関性を確認
    - 2014 7月～9月の範囲では自己相関性は有意
    - 2014 1月～10月の範囲では自己相関の有意性は確認できず
  - 結論:時系列解析の各種モデルを単純に適用するのは無理
- Kleinberg のバースト検知アルゴリズム
  - 隠れマルコフモデルによるイベントの集中的な発生(バースト)を検出
    - ただいま実装中

おわりに

- データ活用ニーズを持つユーザとの協業
  - 目標は明確である必要はあるが具体的過ぎるとすぐに行き詰る
  - 成果を公開するとアプローチの軸になる
  - ビッグデータはデータ活用の一部しか担えない
- ハイプサイクルによる技術動向の把握
  - 世間の関心はビッグデータからデータ分析とその技術に移りつつある
  - 今後は様々な研究領域でのビッグデータ活用事例に注目が集まる
    - 自らのテーマと共通する課題や応用可能な手法をウォッチ
- データ分析への取り組み
  - 引き続き Wikipedia PVC への関心を持っている
  - 時系列分析で「社会的イベント」に関するより深い知見を得るのが目標