ソーシャル・ビッグデータ

3.1 ビッグデータの現状

「ビッグデータ」という言葉は「クラウド・コンピューティング」などと同じように非常にコンセプチュアルな用語です。定義が曖昧模糊としていて意味するところが時間の経過と共に変化していくので、非常に感覚的に使われることが多いです。それ故に長く生き残っていく用語だと思います。しかし、流行りすたりのあるバズワードとしての「ビッグデータ」は、2015年の現在、賞味期限切れを迎えつつあるように感じられます。この印象は、イギリスのシンクタンクであるガートナーが毎年発表しているハイプサイクルでも、2014年には「ビッグデータ」は幻滅期に突入していることから、それなりに裏付けのある話でしょう（図-1）。

もっとも日本国内に限定したハイプサイクルでは流行期の最後に位置するので、バズワードとしての「ビッグデータ」は実に微妙な立ち位置にあるのかもしれません（図-2）。

■ 日本におけるビッグデータの動向

ともあれ、ここ数年間の技術トレンドを象徴するバズワードであり続けた「ビッグデータ」も今ではやや古びてきて、今は「ビッグデータ」より少し具体化した他のバズワード、例えば「Internet of Things（IoT）」などが喧伝されることが多いように感じます。

「Internet of Things（IoT）」は、RFIDの国際標準化に貢献したKevin Ashtonが1999年に提唱した用語で、様々なオブジェクトにIDを付与することによりすべてをネットワーク化する、すなわち「モノのインターネット」を構築するコンセプトです。非常によく似た考えとして「Cyber-Physical System（CPS）」というコンセプトも知られており、こちらの方は「物理的な実体を制御するエレメントが情報を共有し協調して機能するシステム」と定義されています。いずれも「物理世界とサイバー世界を融合する」ところが共通する概念です[*1]。

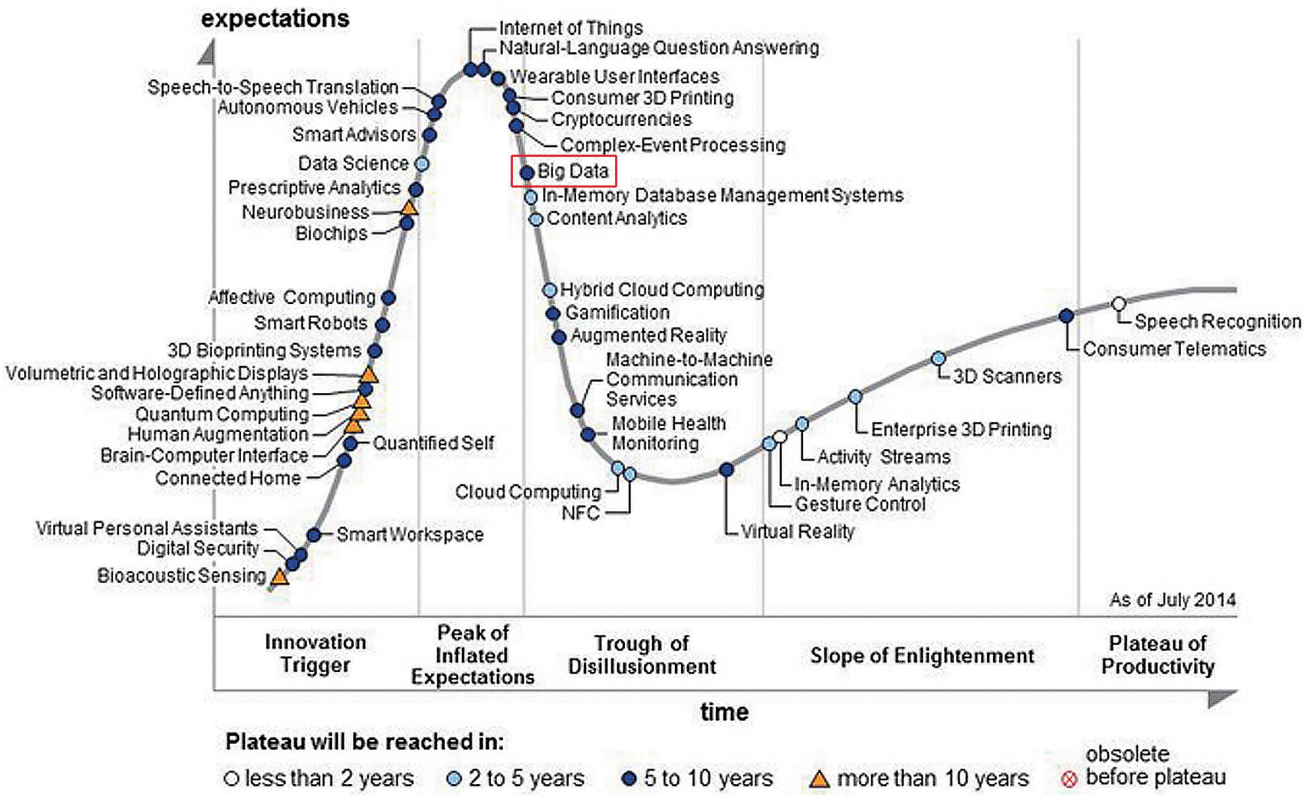


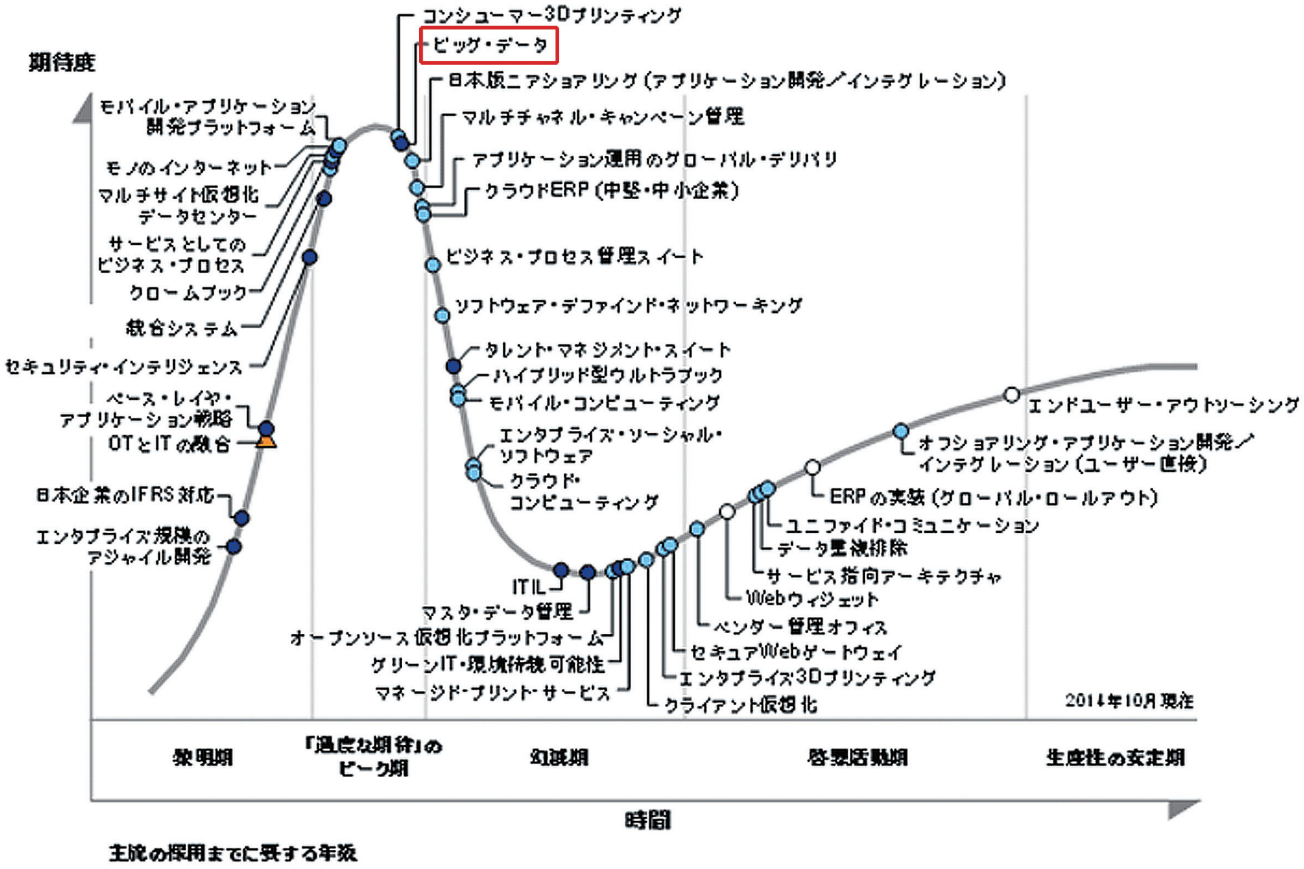
図-1 ハイプサイクル2014版

＊1 「サイバーフィジカルシステムとIoT（モノのインターネット）実世界と情報を結びつける」岩野 和生、高島 洋典、情報管理 2-15 vol.57 no.11（https://www.jstage.jst.go.jp/browse/johokanri/57/11/_contents/-char/ja/）。

今日、宅配便の荷物のタグや非接触型ICカードである定期券、あるいはスマートフォンなどには既にIDが付与されており、その移動などを追跡することが可能になっています。したがって「IDごとに生成され続ける位置情報」は文字どおりのビッグデータであり、これを分析することにより、様々な効率化・最適化を図ることができます。これがIoTやCPSが目指す「明るい未来」だとされています。例えば「スマートグリッド」や「スマートシティ」といった提案を目にしたことのある方は多いのではないでしょうか?

すなわちIoTやCPSは、本質的に社会システムの革新・刷新を目指すスケールの非常に大きなパラダイムです。その研究開発は社会生活を営む多くの人々に直接的な影響を与える可能性があります。例えば一般市民のプライバシーの問題などが挙げられます[*2]。筆者のようなビッグデータの分析手法に関心のある研究者にとって、社会的側面に配慮しなければならないという意味で、IoTやCPSに関わる研究は難しい課題を伴う研究対象であると考えます。

このような社会的問題が顕在化してくるのも、ビッグデータに関する研究開発が進展している証左だと思いますが、特にその研究の方向性を熟慮しなければならない踊り場状態にあるのが「2015年のビッグデータ研究の現状」ではないかと考えます。



出典：ガートナー（2014年10月）

図-2 日本におけるテクノロジのハイプ・サイクル：2014年

*2 NICT（2013年11月25日、プレスリリース）、「大規模複合施設におけるICT技術の利用実証実験を大阪ステーションシティで実施」（http://www.nict.go.jp/press/2013/11/25-1.html）。日経コンピュータ（2014年3月11日、清嶋 直樹）、「JR大阪駅ビルの「顔識別」実証実験、プライバシー侵害の懸念から延期」（http://itpro.nikkeibp.co.jp/article/NEWS/20140311/542723/）。日経コンピュータ（2014年11月7日、清嶋 直樹）、「JR大阪駅ビルの監視カメラを使った顔識別実験、範囲限定で再開へ」（http://itpro.nikkeibp.co.jp/atcl/news/14/110701801/）。

実際に、IoTやCPSが指向する「実世界とサイバー空間を結びつける」アプローチは、「事実をデータ化する」という意味において、ビッグデータ分析に新たな可能性を提供してくれると思います。例えば「ある人間がある時間にある場所にいた」や「あるプラントのある箇所がある時間にある温度に達していた」といった事実を示すデータが分析に活用できるのであれば、少なくともその事実については（論理的に揺らぎのない）確定情報として扱えるので、分析結果の精度向上に寄与するであろうと推測できます。また、そのような確かなデータに基づいて分析・予測しなければならない社会的な需要もあります。例えば「原子力発電所の運転状態監視」であるとか「犯罪者の行動追跡」といった緊急性が高く、確定性が重視される分析では「事実を示すデータ」の価値は絶大です。

3.2 ソーシャル・ビッグデータ

IoTによりもたらされる「事実を示すデータ」を、ここでは便宜的に「ファクト・ビッグデータ」と呼ぶことにします。それとは対照的な特性を持つデータ、例えば「暮らしを豊かにする…」とか「あなたのお好みの…」といった個人の感性や嗜好性を考慮する分析で使用するデータは「ソーシャル・ビッグデータ」と呼べるのではないかと思います。TwitterやFacebookといったのソーシャル・メディアで流通するメッセージなどがその典型的な事例と言えるでしょう。

一般に「雑多な情報の寄せ集め」と理解されがちなソーシャル・ビッグデータですが、Amazonのチーフ・サイエンティストであるAndreas Weigendによれば、ソーシャル・ビッグデータは急激に増大しており、2009年1年間に生成されたデータ総量は、有史以来2008年までに生成されたデータ総量を超えており、これが顧客の商品情報の探索や収集に革命的な変革をもたらしたと報告しています[*3]。

Weigendは、ソーシャル・メディアのユーザが、相互にメーカ・サイトや自身やその友人の感想、あるいは購入した商品の周辺について明確な情報をポストする事例が増えており、メーカが広報宣伝の一環として提供する情報より役立つ傾向にあること、また一部のメーカは、ユーザの率直なデータ提供を奨励し、有益なデータには褒賞する系統立った方法を用意していることを指摘しています。

これらの事実を背景に、今日のオンライン・マーケティングの世界が、ユーザ間の共同作業により第三者にとって有益で明確なデータを作成するモデルに移行（e-businessからme-businessへの移行）しつつあることをWeigendは「ソーシャルデータ革命」と呼んでいます。ソーシャル・ビッグデータの分析は「顧客の期待が何を引き起こすのか?またそのような期待に直面する企業には何ができるのか?」との問いに対する答えを提示できるのかもしれません。インターネットで流通するビッグデータがますます増加していくことは自明ですが、今日のIoTの普及やソーシャルメディアの台頭が、流通するビッグデータの特性の二局化を促す状況を加速させているように考えています。

3.2.1 ソーシャル・ビッグデータの性質

ビッグデータ分析という観点で、ソーシャル・ビッグデータをファクト・ビッグデータとの対比で考えると、「機械的な生成が困難」「含意性が多元的」「意味的な曖昧さが許容される」「推測・憶測が許容される（誤りが許容される）」といった性質を持っていることが挙げられます。少々感覚的な表現を使えば、ファクト・ビッグデータが「機械が生成する硬いデータ」であるのに対し、ソーシャル・ビッグデータが「人間が生成する柔らかいデータ」と言えるのでないかと思います。

通常、データ生成に人間が介在すると、そこから得たデータの分析はその分難しくなります。例えば、大きな公園において一番過ごしやすいポイントを調査することを考えてみてください。「過ごしやすい」を「適正な気温」と仮定し、公園内に温度センサーのついたデバイスを1万個配置し、24時間データを収集し続けるシステムを構築して蓄積されたファクト・ビッグデータを分析する場合は非常に明快な結果が得られます。デバイスは公園内にくまなく設置できますし、設置さえすれば確実にその場所の気温データを得ることができます。1万個もあれ

*3　The Social Data Revolution(s) Andreas Weigend MAY 20, 2009（https://hbr.org/2009/05/the-social-data-revolution.html）。

ば、そのうちの幾つかは故障するでしょうが、故障が確認できれば直ちに交換すれば良いのです。人間が「過ごしやすい」と感じる気温の地点をくまなく調べ上げることができるでしょう。

これに対して、調査員を1万人雇って「各ポイントの温度を調べてください」と依頼するのがソーシャル・ビッグデータでの調査・分析の方法です。生真面目な調査員であれば温度計を持参し「ポイントXXはYY度でした」と温度センサー付きデバイスに迫る正確な報告をしてくれるかもしれませんが、温度計など持たずに「池の周辺は寒いです」と非常に曖昧な報告をする調査員もいるでしょう。華氏の温度計を見て「72度です」と答えたり、ポイントまで出向かずに適当に嘘の報告をしたり、中には「森林地帯は過ごしやすいです」と見当違いな報告をしてくる調査員もいるかもしれません。これでは公園内の「適正な温度」のポイントを見つけるのは大変です。

ですが、調査員にお願いするのが「過ごしやすいポイントを探してください」という依頼だとしたら、調査結果はガラッと変わってくるでしょう。調査員は、まず過ごしやすいポイントに向かって移動するでしょう。その途上で知り合った他の調査員と意見交換したり、一緒にポイントを探し始めたり、意見が一致するポイントが見つかれば「ここが一番いいよね」といったやり取りをするのかもしれません。その結果、時間の経過と共に調査員が公園内の幾つかのポイントに集まるようになることが期待できます。そのタイミングを見計らって個々の調査員にヒアリングを行います。「今どこにいるのか?」「そこは過ごしやすいのか?」「そう思う理由は?」といった質問をすると、公園内の過ごしやすいポイントについて第三者が共感できる情報が得られるわけです。

では、どちらの調査結果が適切でしょうか?それは調査目的によって変わってくるでしょう。公園の管理や環境保全の担当者であればファクト・ビッグデータだけを必要とするのではないかと思います。しかし、公園内の売店のオーナーであれば、むしろソーシャル・ビッグデータによる調査結果を欲するかもしれません。この事例での両者の決定的な違いは、ファクト・ビッグデータは調査方法が一様で結果が均質な事実に基づく客観的なデータで構成されているのに対し、ソーシャル・ビッグデータは調査員(人間)の主観に基づくデータの集積物だということです。両者は共にビッグデータですからエラーデータに対処する「信頼性の考慮」が必要ですが、更にソーシャル・ビッグデータに関しては「データは正しいのか?誤りなのか?どの程度信用するか?」といった信用性も考慮した分析手法が必要になります。

3.3 ビッグデータとしてのWikipedia

ソーシャル・ビッグデータと言えばTwitterで流通しているメッセージをイメージされる方が多いと思うのですが、筆者はWikipediaに着目しています。

最もポピュラーな電子辞書サービスであるWikipediaの概要をここで改めて紹介する必要はないと思いますが、ビッグデータとしてもWikipediaは注目すべき存在です。Wikipedia日本語版の「Wikipedia:統計」ページ[*4]、あるいはWikipedia英語版の"Wikipedia:Statistics"[*5]を見ると最新のWikipediaの記事総数などが掲載されています。

2015年7月20日時点でのWikipediaの総記事数は4,920,887件で総項目数(周辺情報も含める)36,748,410ページ、日本語に限定しても総記事数974,894件、総項目数2,785,007ページにもなります。このデータはCC-BY-SA 3.0[*6]というライセンス条件に基づいて改変、複製などの2次的利用をすることもできます。一般的なオープンソースソフトウェアと同様に使える「ライセンス的に最も安全なビッグデータ」と言えるかもしれません。

Wikipediaのデータは"Wikimedia Download"[*7]というページから辿ることができます。Wikipediaでは、常時データバックアップのタスクが動いており各言語で順次バックアップを行っています。筆者の知見では概ね1ヵ月単位で新たなバックアップが作成されます。

*4 https://ja.wikipedia.org/wiki/Wikipedia:%E7%B5%B1%E8%A8%88
*5 https://en.wikipedia.org/wiki/Wikipedia:Statistics
*6 「クリエーティブ・コモンズ」(https://ja.wikipedia.org/wiki/%E3%82%AF%E3%83%AA%E3%82%A8%E3%82%A4%E3%83%86%E3%82%A3%E3%83%96%E3%83%BB%E3%82%B3%E3%83%A2%E3%83%B3%E3%82%BA)。
*7 http://dumps.wikimedia.org/

このWikipediaデータは既に様々なところで活用されています。著名なところで言えば、例えばDBpedia[*8][*9]、これはWikipediaデータからLinked Open Data(LOD)を生成しデータベース化するプロジェクトです。このデータベースは自然言語処理やテキストマイニングなどに活用されるようです。

3.3.1 ソーシャル・ビッグデータとしてのWikipedia

ではWikiシステムを使った人海戦術で電子辞書としてのコンテンツを維持しているWikipediaはソーシャル・ビッグデータでしょうか?「辞書としての有用性を追求し、ライターが客観的な事実に基づく記述をするよう促すため、システムと運用の両面で数々の工夫を施しているWikipediaの記事データはファクト・ビッグデータに分類すべき」というのが筆者の意見です。もちろん、Wikiシステムを使って人手により記述・修正されているデータですので、機械的に生成されるデータに比べて、記述の誤りや事実誤認、記事相互での矛盾など、データとしての完全性に問題はありますが、その反面、事実が確認不能な事柄や定説が定まらない事柄なども網羅できる「間口の広いファクト・ビッグデータ」とも理解できます。

もっともWikipediaにはソーシャル・ビッグデータとしての顔もあります。2013年から公開されるようになった"Page view statistics for Wikimedia projects"[*10]がそれに該当します。これはWikimediaプロジェクトの全ページについて、1時間単位でページビューカウントを集計したデータで、ここから辿ると2008年(正確には2007年12月)以降のページビュー情報を入手することができます。

筆者の研究グループでは2013年6月にこのデータに基づくランキングサービス[*11]を立ち上げて以来、ランキングの変動の面白さ、すなわち「世間で話題になっているトピックがランキング上位にマークされる」振る舞いに注目していました(図-3)。

3.3.2 Tobias Preisの研究

ウォーリック大学の金融行動学の准教授であるTobias Preisによれば、インターネットを利用する個人の情報探索行動がこのような振る舞いを発生させていると説明しています。例えば、マスメディアなどで大々的に報道されているトピックに接すると、日常的にインターネットを利用する個人はサー

図-3 Wikipediaランキング

*8 http://wiki.dbpedia.org/
*9 http://ja.dbpedia.org/
*10 http://dumps.wikimedia.org/other/pagecounts-raw/
*11 http://www.gryfon.iij-ii.co.jp/ranking/

チエンジンで検索したり、Wikipediaを閲覧して、トピックに関する情報を収集する行動をとります。この行動の痕跡は検索エンジンのクエリーデータやWikipediaのページビューとして記録されると推測されます。この現象に注目したPreisは2010年の論文"Complex dynamics of our economic life on different scales : insights from search engine query data"[*12]で、サーチエンジンのクエリーデータと株式市場の変動が相関していることを突き止めました。

続く論文"Quantifying Trading Behavior in Financial Markets Using Google Trends"[*13]では、Google Trendsから得られるクエリーデータに含まれる金融関連の98の用語の検索量の増大が金融市場の大きな損失に先行する傾向があることを示唆しました。

更に論文"Quantifying Wikipedia Usage Patterns Before Stock Market Moves"[*14]では、Google Trendsでの分析の知見を使ってWikipediaの閲覧回数が株式市場の大規模な変動と相関することを発見しています。

これらの論文でのPreisの結論では、検索エンジンのクエリーデータやWikipediaのページビューといったオンラインデータから、意思決定を迫られている人々の情報収集活動について新たな知見が得られる可能性を指摘しています。例えば、株式市場の大暴落といった事象は個々の投資家の意思決定の結果ですが、オンラインデータに注目していれば、その兆候を早期に発見することができるわけです。これはソーシャル・ビッグデータを活用した典型的なソリューションの一例でしょう。

3.4 まとめ

本稿ではソーシャル・ビッグデータとその性質を中心に議論してきました。Andreas Weigendが主張する「ソーシャルデータ革命」は、Eコマースといった特定の分野では、商品購入における人間行動に関する膨大なデータが蓄積され始めており、そのデータの分析を前提とした新たなアプローチの必要性が主張されています。

ソーシャル・ビッグデータが「人間の行動に関するデータ」であることから、その分析方法の開発は社会科学や行動科学の領域の研究となるのでしょう。しかしながら、人間の行動に関してこれだけ広範囲かつ詳細なデータを入手できるような状況がこれまでなかったことを考えると、価値ある知識を抽出できる有効な手法を見出すには相応の時間が必要でしょう。筆者が知る限りでは、マイクロブログやSNSなどから得たデータを元に個人の行動を追跡、分析するアプローチは「なんらかの分析結果は得られるものの、そこから有意な知見を見出すことが難しい」という意味においてあまり成功していないと考えます。

一方、インターネット利用者の行動をマクロな現象と捉え、複雑系の分析手法を応用するTobias Preisのアプローチの方がむしろ有意な知見が得られやすいのではないかと考えます。この分析にはクエリーデータやWikipedia PVCなどが用いられていますが、Preisの論文は、独立した個人による集団行動の予兆を捉えることにより、短期的な予測が可能であることを示しているように思います。また、この分析によって得られた変化の著しい現象をより詳細に分析するには、ソーシャルメディアから得られたデータを使って個人の行動を追跡することにより、要因などの知見を得ることができるのではないかと、考えています。

執筆者:
藤田 昭人（ふじた あきと）
株式会社IIJイノベーションインスティテュート(IIJ-II)企画開発センター チーフアーキテクト。2008年IIJ入社。
構造化オーバーレイ研究の知見を活用したクラウドコンピューティング技術の研究開発に従事している。

*12 http://www.tobiaspreis.de/publications/prs_ptrsa_2010.pdf
*13 http://www.nature.com/srep/2013/130425/srep01684/full/srep01684.html
*14 http://www.nature.com/srep/2013/130508/srep01801/full/srep01801.html