

ビューティフルデータ

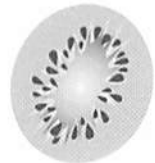
Toby Segaran 編
Jeff Hammerbacher

堀内孝彦、真鍋 加奈子、苅谷 潤 訳
小俣 仁美、篠崎 誠

O'REILLY®
オライリー・ジャパン

本書で使用するシステム名は、製品名は、それぞれ各社の商標、または登録商標です。
なお、本文中では™、®、©マークは省略しています。

Beautiful Data



Edited by Toby Segaran and Jeff Hammerbacher

O'REILLY®

Beijing • Cambridge • Farnham • Köln • Sebastopol • Tokyo

© 2011 O'Reilly Japan, Inc. Authorized translation of the English edition of Beautiful Data © 2009 O'Reilly Media, Inc. This translation is published and sold by permission of O'Reilly Media, Inc., the owner of all rights to publish and sell the same.

本書は、株式会社オライリー・ジャパンがO'Reilly Media, Inc.との許諾に基づき翻訳したものです。日本語版についての権利は、株式会社オライリー・ジャパンが保有します。

日本語版の内容について、株式会社オライリー・ジャパンは、最大限の努力をもって正確を期していますが、本書の内容に基づく運用結果についての責任は負いかねますので、ご了承ください。

はじめに

「ビューティフルコード」に続く本をデータを題材にして作らないかと提案されたとき、それは素晴らしい、と乗り気になった。とても野心的な試みだ。いまやあらゆる専門分野や日常生活の様々なところで、データの収集、可視化、処理が行われている。その幅広さには目を見張る。そこで我々は、賞賛に値する仕事をしている、非常に広範な領域の人々に声を掛けた。予想以上に多くの人々が協力してくれるというので、驚いた。

結果としてできあがったのが本書である。データを扱う仕事はあらゆる分野に存在し、しかもそれは美しいものだということを読者に伝えたい。題材は、役所との戦い、火星探査機、統計プログラムの利用方法、可視化、Radioheadのビデオ編集、地図、DNA、そして“データ哲学”としか言いようのない何かまで、多岐にわたる。

本書（原書）の印税は、データを自由に解放つことで世界をよりよい場所にしようとする2つの団体、クリエイティブコモンズとSunlight Foundationに寄付される。データとの出会いで世界をどう変えることができるのか、読者自身で考えてもらえれば幸いだ。

本書の構成

各章は、大まかにデータの収集、保存、編成、検索、可視化、分析の順番で並んでいる。

1章「データの中に生活をみる」はNathan Yauによるもので、パーソナルデータ収集という新分野の2つのプロジェクトにおける動機と挑戦について述べる。

2章「ビューティフル・ピープル」はJonathan Follett、Matthew Holmによるもので、ウェブを通して人々からデータを集める際には信頼、説得、テストが大事であると説く。

3章「火星上での組み込み画像処理」はJ. M. Hughesによるもので、宇宙探査という制約下で動作するデータ処理システムの設計における挑戦について議論する。

4章「PNUTShellにおけるクラウドストレージの設計」はBrian F. Cooper、Raghu Ramakrishnan、Utkarsh Srivastavaによるもので、現代的なウェブアプリケーションを支えるために、世界中に分散したデータセンターを一つの共通ストレージプラットフォームへと統合する目的でYahoo!が設計したソフトウェアについて述べる。

5章「情報プラットフォームとデータサイエンティストの登場」はJeff Hammerbacherによるもので、Facebookにおけるデータチームの経験を元に、情報処理ツールの進化とそれを活用する人間を追う。

6章「写真アーカイブの地理的な美」はJason DykesとJo Woodによるもので、人々が無償で提供するありふれたデータが、空間的データとしてカラフルに視覚化したときに持つ力について語る。

7章「データの自己発見」はJeff JonasとLisa Sokolによるもので、多くの人が必要とするであろう、データを全体として取り扱うための新しい考え方を説明する。

8章「リアルタイムのポータブルデータ」はJud Valeskiによるもので、ウェブを通じてソーシャルデータ、ロケーションデータを配布するときに起きている限界に切り込み、この問題の解となりうる方式について議論する。

9章「ディープウェブを活用する」はAlon HalevyとJayant Madhavanによるもので、今はまだウェブのフォームの裏側にとらわれているデータを検索可能にするためにGoogleが開発したツールについて述べる。

10章「Radiohead [House of Cards] のプロモーションビデオができるまで」はAaron KoblinとValdean Klumpによるもので、賞をとったミュージックビデオ制作の裏側における、レーザー、プログラミング、撮影現場についての冒険物語だ。

11章「都市データの視覚化」はMichal Migurskiによるもので、我々の身近な世界で最も重要なあるデータを、オープンにして美しく加工するまでの経緯を詳しく説明する。

12章「sense.usの設計」はJeffrey Heerによるもので、データの可視化をソーシャルな活動として捉え直し、この新しい視点を元に150年におよぶアメリカの国勢調査データを探求する。

13章「データでできないこと」はCoco Krummeによるもので、人々が犯しやすいデータについての誤解と誤使用をわかりやすく示した。

14章「自然言語のコバスターデータ」はPeter Norvigによるもので、ウェブから収集した1兆語の自然言語コバスターデータを用いて、いくつかの刺激的な例題を読者に案内する。

15章「データの中の生命：DNA物語」はMatt WoodとBen Blackburneによるもので、DNAの中のデータの美しさと共に、そのデータを生成、収集、加工する巨大なシステム基盤について述べる。

16章「実世界のデータをビューティフルにする」はJean-Claude Bradley、Rajarshi Guha、Andrew Lang、Pierre Lindenbaum、Cameron Neylon、Antony Williams、Egon Willighagenによるもので、クラウドソース化と思い切った透明性の組み合わせにより、新薬発見研究の世界を前進させる試みについて紹介する。

17章「外見のデータ解析：数百万人の社会的ステロタイプ調査」はBrendan O'ConnorとLukas Biewaldによるもので、他人の顔写真を人々に匿名で評価してもらった際に現れる相関とパターンを紹介する。

18章「ベイエリア・ブルース：住宅市場崩壊の影響」はHadley Wickham、Deborah F. Swayne、David Pooleによるもので、ベイエリアにおける近年の住宅市場崩壊をオープンソースソフトウェアと公開データに基づき詳しく追った。

19章「政治に関するビューティフルデータ」はAndrew Gelman、Jonathan P. Kastellec、Yair Ghitzaによるもので、統計というツールとデータの可視化により、社会を組織する政治プロセスについての洞察を得た事例を紹介する。

20章「データをつなぐ」はToby Segaranによるもので、ウェブから得られる多くのデータセットをつなぎ合わせる難しさとその可能性について探求する。

本書の表記

本書では、以下の表記を使用しています。

等幅 (Constant Width)

サンプルコードを示す。

サンプルコードの使用について

本書の目的は、読者の仕事の手助けをすることです。一般に、本書に掲載しているコードは各自のプログラムやドキュメントに使用してかまいません。コードの大部分を転載する場合を除き、出版社に許可を求める必要はありません。たとえば、本書のコードブロックをいくつか使用するプログラムを作成するために、許可を求める必要はありません。なお、O'Reillyから出版されている書籍のサンプルコードをCD-ROMとして販売したり配布したりする場合には、そのための許可が必要です。本書や本書のサンプルコードを引用して問題に答える場合、許可を求める必要はありません。ただし、本書のサンプルコードのかなりの部分を製品マニュアルに転載するような場合には、そのための許可が必要です。

出展を明記する必要はありませんが、そうしていただければ感謝します。出展を明記する際には、Toby Segaran、Jeff Hammerbacher 編「ビューティフルデータ」(O'Reilly Media, Inc.) のように、タイトル、著者、出版社、ISBNなどを盛り込んでください。サンプルコードの使用について、正規の使用の枠を超える、またはここで許可している範囲を超えると感じる場合は、permissions@oreilly.com までご連絡ください。

意見と質問

本書(日本語翻訳版)に関するコメントや質問は以下に送付してください。

株式会社オライリー・ジャパン

〒160-0002 東京都新宿区坂町26番地27 インテリジェントプラザビル1F

電話 03-3356-5227

FAX 03-3356-5261

電子メール japan@oreilly.co.jp

本書に関する技術的な質問や意見につきましては、次の宛先に電子メールを送ってください。

bookquestions@oreilly.com (英語)

japan@oreilly.co.jp (日本語)

本書のWebページには、正誤表、サンプルコード、追加情報が掲載されています。以下のアドレスでアクセスできます。

<http://www.oreilly.com/catalog/9780596157128/>

<http://www.oreilly.co.jp/books/4873114897/>

オライリーに関するその他の情報(文献、会議、リソースセンター、O'Reilly Network)については、

次のオライリーのWebサイトを参照してください。

<http://www.oreilly.com>

<http://www.oreilly.co.jp>

目次

はじめに	v
1章 データの中に生活を見る (Nathan You)	1
PEIR (Personal Environmental Impact Report)	2
YFD (your.flowingdata)	2
パーソナルデータの収集	3
データの蓄積	4
データ処理	5
データの可視化	6
まとめ	13
参加するには	13
2章 ビューティフル・ピープル ——ユーザの存在を忘れることなくデータ収集の手段をデザインする (Jonathan Follett, Matthew Holm)	15
はじめに：ユーザへの共感がこれからのデザインの基本原則	15
プロジェクト：あるラグジュアリー新製品のための顧客調査	16
データ収集に伴う特有の課題	17
我々のデザインについて	19
結果と考察	30
3章 火星上での組み込み画像処理 (J. M. Hughes)	33
概要	33
はじめに	33
背景知識	35

詰めるか、詰めまいか.....	38
3つのタスク.....	39
スロットへの画像格納.....	41
画像の受け渡し：3つのタスク間でのやりとり.....	43
写真を取得する：画像のダウンロードと処理.....	46
画像圧縮.....	47
ダウンリンク、あとは地球へまっしぐら.....	48
おわりに.....	49

4章 PNUTShellにおけるクラウドストレージの設計

(Brian F. Cooper, Raghu Ramakrishnan, Utkarsh Srivastava).....	51
はじめに.....	51
データの更新.....	53
複雑なクエリ.....	60
他のシステムとの比較.....	63
まとめ.....	66

5章 情報プラットフォームとデータサイエンティストの登場

(Jeff Hammerbacher).....	69
図書館と脳.....	69
自己認識能力を持つようになったFacebook.....	70
ビジネスインテリジェンスシステム.....	71
データウェアハウスの停止と復活.....	72
データウェアハウスを超えて.....	73
チーターとゾウ.....	74
データの不合理な有効性.....	75
新しいツールと応用研究.....	76
MADスキルとCosmos.....	77
データスペースとしての情報プラットフォーム.....	78
データサイエンティスト.....	78
まとめ.....	79

6章 写真アーカイブの地理学的な美 (Jason Dykes, Jo Wood).....

地理データの中の美：Geograph.....	82
視覚化と美、そしてツリーマップ.....	84

Geographの単語頻度における地理学的視点	87
発見の中の美	94
考察とまとめ	96
7章 データの自己発見 (Jeff Jonas、Lisa Sokol)	99
はじめに	99
ジャスト・イン・タイム発見の利点	100
ルーレットでイカサマをする	101
企業内発見能力	104
横断検索ですべては解決しない	104
ディレクトリ：プライスレス	105
関連性：問題は何か？ それは誰にとって問題なのか？	107
構成要素と注意点	108
プライバシーについての考察	110
まとめ	110
8章 リアルタイムのポータブルデータ (Jud Valeski)	113
はじめに	113
最新技術	114
ソーシャルデータの標準化	121
まとめ：Gnip経由での仲介	124
9章 ディープウェブを活用する (Alon Halevy、Jayant Madhavan)	127
ディープウェブとは	127
ディープウェブへのアクセスを提供するための2つの手法	129
まとめと今後について	139
10章 Radiohead「House of Cards」の プロモーションビデオができるまで (Aaron Koblin、Valdean Klump)	141
ビデオ作成のきっかけ	141
データキャプチャ機器	142
2つのデータキャプチャシステムを使う利点	146
データのキャプチャ「撮影」	147
データ処理	151
データの後処理	152

ビデオの公開	153
まとめ	156
11章 都市データの視覚化 (Michal Migurski)	159
はじめに	159
背景	160
難問の解決	161
データサービスの公開	165
再検討	170
まとめ	173
12章 sense.us の設計 (Jeffrey Heer)	175
データの視覚化と社会調査のためのデータ分析	176
データ	177
データの視覚化	179
コラボレーション機能	185
探索者と観察者	190
まとめ	193
カラー図版	195
13章 データでできないこと (Coco Krumme)	227
データの導きを得られないとき	230
まとめ	239
14章 自然言語のコーパスデータ (Peter Norvig)	241
ワードセグメンテーション	242
暗号	251
スペル訂正	259
他の適用分野	265
議論とまとめ	266
15章 データの中の生命 : DNA 物語 (Matt Wood, Ben Blackburne)	269
データ格納庫としてのDNA	269
デジタルストレージとしてのDNA	275

データソースとしてのDNA.....	276
DNAの未来.....	283
16章 実世界のデータをビューティフルにする (Jean-Claude Bradley, Rajarshi Guha, Andrew Lang, Pierre Lindenbaum Cameron Neylon, Antony Williams, Egon Willighagen).....	285
現実のデータが持つ問題.....	285
実験ノートの実データを公開する.....	286
クラウドソースされたデータを検証する.....	288
ネット上でのデータの表現.....	289
改善サイクルを回す：可視化して新しい実験を見つける.....	296
オープンデータとフリーサービスでデータウェブを構築する.....	299
17章 外見のデータ解析：数百万人の社会的ステロタイプ調査 (Brendan O'Connor, Lukas Blewald).....	303
はじめに.....	303
データの前処理.....	304
データの調査.....	306
年齢、魅力、性別.....	309
タグを見る.....	314
性差のある語はどれ？.....	318
クラスタリンク.....	320
まとめ.....	324
18章 ベイエリア・ブルース：住宅市場崩壊の影響 (Hadley Wickham, Deborah F. Swayne, David Poole).....	327
はじめに.....	327
データをどのように得たのか.....	327
ジオコーディング (Geocoding).....	328
データのチェック.....	329
分析.....	329
インフレの影響.....	331
金持ちはより金持ちに、貧乏人はますます貧乏に.....	332
地理的差異.....	333
国勢調査の情報.....	337

サンフランシスコの調査.....	341
まとめ.....	344
19章 政治に関するビューティフルデータ	
(Andrew Gelman, Jonathan P. Kastellec, Yair Ghitza).....	345
事例1：選挙区画の再編成と政党の偏り.....	346
事例2：推定値の時系列.....	347
事例3：年齢と票.....	349
事例4：最高裁判事指名に関する世論と上院での投票.....	350
事例5：ペンシルバニア州における党派の局所的な偏り.....	351
まとめ.....	353
20章 データをつなぐ (Toby Segaran).....	355
パブリックなデータには、実際どんなものがあるか.....	355
データの結合が可能にすること.....	356
企業の中で.....	358
データ結合の障害.....	358
解決策.....	363
まとめ.....	367
訳者あとがき.....	369
索引.....	371