



/THEORY//IN//PRACTICE

Beautiful Data

The Stories Behind Elegant Data Solutions

O'REILLY®

Edited by Toby Segaran
& Jeff Hammerbacher

Beautiful Data

"Data indeed proves to be the 'Intel inside' of the next generation of computer applications. Inside this book, industry leaders describe how their projects are harnessing the power of data in new ways. A must-read for anyone interested in the future of data and problem-solving."

—Tim O'Reilly, founder and CEO, O'Reilly Media, Inc.

Discover just how wide-ranging—and beautiful—working with data can be. Through this collection of personal stories, 39 of the best data practitioners in the field explain how they developed simple and elegant solutions for a variety of projects, ranging from the Mars lander to a Radiohead video, and much more. With this book, you will:

- Explore the opportunities and challenges inherent in vast online datasets
- Learn how to visualize trends in urban crime using maps and data mashups
- Discover how crowdsourcing and transparency have advanced the state of drug research
- Understand how new data can alert users when it overlaps preexisting data
- Learn about the massive infrastructure required to process DNA data

***Beautiful Data* includes contributions from:**

Nathan Yau

Jonathan Follett and Matthew Holm

J.M. Hughes

Brian F. Cooper, Raghu Ramakrishnan,
and Utkarsh Srivastava

Jeff Hammerbacher

Jason Dykes and Jo Wood

Jeff Jonas and Lisa Sokol

Jud Valeski

Alon Halevy and Jayant Madhavan

Aaron Koblin with Valdean Klump

Michal Migurski

Jeffrey Heer

Coco Krumme

Peter Norvig

Matt Wood and Ben Blackburne

Jean-Claude Bradley, Rajarshi Guha,
Andrew Lang, Pierre Lindenbaum,
Cameron Neylon, Antony Williams,
and Egon Willighagen

Brendan O'Connor and Lukas Biewald
Hadley Wickham, Deborah F. Swayne,
and David Poole

Andrew Gelman, Jonathan P. Kastellec,
and Yair Ghitza

Toby Segaran

All author royalties will be donated to the Sunlight Foundation and Creative Commons.

US \$44.99

CAN \$56.99

ISBN: 978-0-596-15711-1



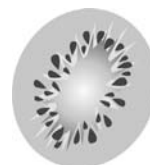
Safari
Books Online

Free online edition

for 45 days with purchase of
this book. Details on last page.

O'REILLY[®] oreilly.com

Beautiful Data



Edited by Toby Segaran and Jeff Hammerbacher

O'REILLY®

Beijing • Cambridge • Farnham • Köln • Sebastopol • Taipei • Tokyo

Beautiful Data

Edited by Toby Segaran and Jeff Hammerbacher

Copyright © 2009 O'Reilly Media, Inc. All rights reserved. Printed in Canada.

Published by O'Reilly Media, Inc. 1005 Gravenstein Highway North, Sebastopol, CA 95472

O'Reilly books may be purchased for educational, business, or sales promotional use. Online editions are also available for most titles (<http://my.safaribooksonline.com>). For more information, contact our corporate/institutional sales department: (800) 998-9938 or corporate@oreilly.com.

Editor: Julie Steele

Proofreader: Rachel Monaghan

Production Editor: Rachel Monaghan

Cover Designer: Mark Paglietti

Copyeditor: Genevieve d'Entremont

Interior Designer: Marcia Friedman

Indexer: Angela Howard

Illustrator: Robert Romano

Printing History:

July 2009: First Edition.

The O'Reilly logo is a registered trademark of O'Reilly Media, Inc. *Beautiful Data*, the cover image, and related trade dress are trademarks of O'Reilly Media, Inc. Many of the designations used by manufacturers and sellers to distinguish their products are claimed as trademarks. Where those designations appear in this book, and O'Reilly Media, Inc. was aware of a trademark claim, the designations have been printed in caps or initial caps.

While every precaution has been taken in the preparation of this book, the publisher and authors assume no responsibility for errors or omissions, or for damages resulting from the use of the information contained herein.

ISBN: 978-0-596-15711-1

[F]

All royalties from this book will be donated to Creative Commons and the Sunlight Foundation.

CONTENTS

	PREFACE	xi
1	SEEING YOUR LIFE IN DATA <i>by Nathan Yau</i>	1
	Personal Environmental Impact Report (PEIR)	2
	your.floodingata (YFD)	3
	Personal Data Collection	3
	Data Storage	5
	Data Processing	6
	Data Visualization	7
	The Point	14
	How to Participate	15
2	THE BEAUTIFUL PEOPLE: KEEPING USERS IN MIND WHEN DESIGNING DATA COLLECTION METHODS <i>by Jonathan Follett and Matthew Holm</i>	17
	Introduction: User Empathy Is the New Black	17
	The Project: Surveying Customers About a New Luxury Product	19
	Specific Challenges to Data Collection	19
	Designing Our Solution	21
	Results and Reflection	31
3	EMBEDDED IMAGE DATA PROCESSING ON MARS <i>by J. M. Hughes</i>	35
	Abstract	35
	Introduction	35
	Some Background	37
	To Pack or Not to Pack	40
	The Three Tasks	42
	Slotting the Images	43
	Passing the Image: Communication Among the Three Tasks	46
	Getting the Picture: Image Download and Processing	48
	Image Compression	50
	Downlink, or, It's All Downhill from Here	52
	Conclusion	52

4	CLOUD STORAGE DESIGN IN A PNUTSHELL	55
	<i>by Brian F. Cooper, Raghu Ramakrishnan, and Utkarsh Srivastava</i>	
	Introduction	55
	Updating Data	57
	Complex Queries	64
	Comparison with Other Systems	68
	Conclusion	71
5	INFORMATION PLATFORMS AND THE RISE OF THE DATA SCIENTIST	73
	<i>by Jeff Hammerbacher</i>	
	Libraries and Brains	73
	Facebook Becomes Self-Aware	74
	A Business Intelligence System	75
	The Death and Rebirth of a Data Warehouse	77
	Beyond the Data Warehouse	78
	The Cheetah and the Elephant	79
	The Unreasonable Effectiveness of Data	80
	New Tools and Applied Research	81
	MAD Skills and Cosmos	82
	Information Platforms As Dataspaces	83
	The Data Scientist	83
	Conclusion	84
6	THE GEOGRAPHIC BEAUTY OF A PHOTOGRAPHIC ARCHIVE	85
	<i>by Jason Dykes and Jo Wood</i>	
	Beauty in Data: Geograph	86
	Visualization, Beauty, and Treemaps	89
	A Geographic Perspective on Geograph Term Use	91
	Beauty in Discovery	98
	Reflection and Conclusion	101
7	DATA FINDS DATA	105
	<i>by Jeff Jonas and Lisa Sokol</i>	
	Introduction	105
	The Benefits of Just-in-Time Discovery	106
	Corruption at the Roulette Wheel	107
	Enterprise Discoverability	111
	Federated Search Ain't All That	111
	Directories: Priceless	113
	Relevance: What Matters and to Whom?	115
	Components and Special Considerations	115
	Privacy Considerations	118
	Conclusion	118

8	PORTABLE DATA IN REAL TIME	119
	<i>by Jud Valeski</i>	
	Introduction	119
	The State of the Art	120
	Social Data Normalization	128
	Conclusion: Mediation via Gnip	131
9	SURFACING THE DEEP WEB	133
	<i>by Alon Halevy and Jayant Madhavan</i>	
	What Is the Deep Web?	133
	Alternatives to Offering Deep-Web Access	135
	Conclusion and Future Work	147
10	BUILDING RADIOHEAD'S HOUSE OF CARDS	149
	<i>by Aaron Koblin with Valdean Klump</i>	
	How It All Started	149
	The Data Capture Equipment	150
	The Advantages of Two Data Capture Systems	154
	The Data	154
	Capturing the Data, aka "The Shoot"	155
	Processing the Data	160
	Post-Processing the Data	160
	Launching the Video	161
	Conclusion	164
11	VISUALIZING URBAN DATA	167
	<i>by Michal Migurski</i>	
	Introduction	167
	Background	168
	Cracking the Nut	169
	Making It Public	174
	Revisiting	178
	Conclusion	181
12	THE DESIGN OF SENSE.US	183
	<i>by Jeffrey Heer</i>	
	Visualization and Social Data Analysis	184
	Data	186
	Visualization	188
	Collaboration	194
	Voyagers and Voyeurs	199
	Conclusion	203

13	WHAT DATA DOESN'T DO <i>by Coco Krumme</i>	205
	When Doesn't Data Drive?	208
	Conclusion	217
14	NATURAL LANGUAGE CORPUS DATA <i>by Peter Norvig</i>	219
	Word Segmentation	221
	Secret Codes	228
	Spelling Correction	234
	Other Tasks	239
	Discussion and Conclusion	240
15	LIFE IN DATA: THE STORY OF DNA <i>by Matt Wood and Ben Blackburne</i>	243
	DNA As a Data Store	243
	DNA As a Data Source	250
	Fighting the Data Deluge	253
	The Future of DNA	257
16	BEAUTIFYING DATA IN THE REAL WORLD <i>by Jean-Claude Bradley, Rajarshi Guha, Andrew Lang, Pierre Lindenbaum, Cameron Neylon, Antony Williams, and Egon Willighagen</i>	259
	The Problem with Real Data	259
	Providing the Raw Data Back to the Notebook	260
	Validating Crowdsourced Data	262
	Representing the Data Online	263
	Closing the Loop: Visualizations to Suggest New Experiments	271
	Building a Data Web from Open Data and Free Services	274
17	SUPERFICIAL DATA ANALYSIS: EXPLORING MILLIONS OF SOCIAL STEREOTYPES <i>by Brendan O'Connor and Lukas Biewald</i>	279
	Introduction	279
	Preprocessing the Data	280
	Exploring the Data	282
	Age, Attractiveness, and Gender	285
	Looking at Tags	290
	Which Words Are Gendered?	294
	Clustering	295
	Conclusion	300