# <<数据之美>>

## 图书基本信息

书名：<<数据之美>>

13位ISBN编号：9787111315124

10位ISBN编号：711131512X

出版时间：2010年10月

出版时间：机械工业出版社

作者：Toby Segaran,Jeff Hammerbacher

页数：354

译者：祝洪凯,李妹芳,段炼

## 前言

我一直对数据挖掘很感兴趣，尤其是通过对海量、抽象甚至枯燥的数据进行挖掘分析后，利用数据可视化工具展现出来的那种绚丽多彩、富含意蕴的数据之美更是令我痴迷、叹为观止。
本书涉及领域很广，各领域的精英们向我们娓娓道来相关领域的数据信息系统的架构的设计，包括Yahoo！
的云存储架构、Deep Web数据抓取、Face book的信息平台、自然语言处理、“ 凤凰号 ”火星探测器的图像数据处理、探索数据生命的DNA漫谈，甚至是Radio head视频的制作、旧金山的次贷危机等。
阅读完本书之后，我自己的一个很大的收获是对于自己比较了解的领域，如云存储、Deep Web、NLP等有了进一步的理解和实践指导，而对于那些完全不熟悉的领域，如探索数据生命、火星探测器、制作Radio head视频等则更是开阔了视野，不但对数据有了新的认识，而且激发了思考问题的一些新的思维方式。
这本书令我很感怀的另一方面是，我发现这些“ 数据科学家 ”在兢兢业业构建平台处理数据的过程中，虽然遇到了很多困难和挑战，但是却依然如此坚持、执着地探索数据之美。
在翻译本书过程中，这种激情不仅激励着我完成这本书的翻译，同时也激励着我在生活、工作中要有毅力和恒心。
而纵观我身边的阿里巴巴云计算的同事们——这些“ 阿里数据科学家 ”们，也无一不是那种永远充满着激情致力于我们的“ 飞天 ”梦想！
这是我翻译的第一本书，很感激机械工业出版社华章公司编辑陈冀康先生慷慨地引我入门，并且对因为我前段时期项目开发非常紧张而导致翻译进度几乎停滞的宽容和理解表示深深感激。
感谢所有其他为本书付出努力的人们。
由于时间和精力有限，本书的疏漏、错误之处在所难免，还望各位读者不吝批评指正。

# <<数据之美>>

## 内容概要

本书揭示了数据发现可以是多么广泛和美丽!在本书中，39位业内最佳数据实践者揭秘了他们如何为各种项目开发简单优雅的解决方案，例如火星着陆探测器、Radiohead视频的制作等。
在本书中，你将：

- 探索在海量的在线数据集中所固有的机遇和挑战
- 学习如何使用地图和数据“混搭”(mashup)来对都市犯罪趋势进行可视化
- 发现“开放来源”(crowdsourcing)和透明化如何改善药物研究的现状
- 理解新的数据可能会覆盖已有数据时，如何向用户报警
- 了解DNA数据处理所需要的大规模的基础设施

## 作者简介

译者：祝洪凯 李妹芳 段炼 编者：（美国）托比（Toby Segaran） （美国）Jeff Hammerbacher

## 书籍目录

# <<数据之美>>

## 章节摘录

插图：正如由机器人完成的任务生成的数据非常宝贵，需要返回这些数据的通信带宽也是非常宝贵的。
对于较小的图像，比如那些通过子图定位或者抽样操作，图片大小已经减少了，因此直接执行“下行”操作而不做压缩处理是可行的。
更大的图像，比如全尺寸大小的ssI图像，“下行”操作会消耗很多带宽，因此在这种情况下，通常采用压缩方法来解决。
ICS采用像素映射和扩展，提供了两种压缩和减少图像大小的方式。
对于某个特定的图片，采用哪种压缩或减少图像大小方式，主要依赖于图像需要达到的保真程度，高保真被认为是图像的一个必要方面。
在一些情况下，每个像素8位就足够了；而在其他一些情况下，JPEG压缩本身造成的图像保真损失是可以接受的；而对于一些情况，图像需要保持尽可能高的保真，则可以采用无损压缩的方式。
在ICS内部，一台JPEG压缩器采用所有的整数算术计算和就地操作，提供所谓的“有损”压缩方式。
JPEG被认为是有损的，因为其压缩过程丢失了部分图像数据。
JPEG可以通过命令，对图像数据实现不同程度的压缩。
最终代码是松散式地基于Mars’98使命的JPEG压缩器；虽然凤凰号火星着陆探测器的ICS的实现只采用了其部分原始代码。
原始的JPEG压缩器使用的是浮点数乘以全尺寸大小的图像数组作为缓存，并采用动态内存分配方式。
对于这种方式如何在飞行软件上正常工作，我仍然感到很困惑，不过它确实能够正常工作。
在压缩代码中使用浮点数来表示像素数据，这也意味着对于每个图像，比起16位整数的原始图像表示方式，浮点数占用了其四倍的内存空间。
第二种压缩方式，也称为Rice无损压缩（Rice Lossless）或者Rice压缩，采用了由Jet Propulsion实验室的Robert Rice开发的一种算法。
该Rice算法可以对图像数据实现几乎2：1的压缩效果，且没有数据损失。
而JPEG算法在压缩过程中丢失了部分数据。
Rice压缩方法也是在图像槽中就地对图像进行压缩。
两种无压缩的缩小图像大小技术或者采用查询表，把12位的像素值映射到8位的像素值，或者采用位缩小技术，对像素数据向右移动4位，生成一个每个像素8位的图像。
JPEG和Rice压缩函数都接受12位或者8位的图像数据。

## 媒体关注与评论

“ 数据实际上已经是下一代计算机应用的真正核心。
本书中，各位业界精英描述了在他们的项目中如何以全新的方式来驾驭数据的力量。
对于任何对数据的未来和问题的解决感兴趣的读者来说，本书都是一部必读之作。
” ——Tim O'Reilly，O'Reilly Media公司创始人兼CEO

## 版权说明