《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

出版社：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

《数据之美》

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

“ 数据被证实好比下一代计算机应用的‘ 因特尔内核 ’。在本书中，各业界领袖描述了他们的项目如何通过新的方式来驾驭数据的力量。对于任何对未来关于数据和问题解决感兴趣的读者来说，本书是必读的佳作。 ”
——Tim O ’ Reilly, O ’ Reilly Media公司创始人兼CEO
探索数据的范围可以多么广泛，其工作可以多么美丽！通过这部个人故事集合，在这个领域的39个最佳数据实践者阐释了他们如何为各种项目开发简单优雅的解决方案，包括从火星着陆探测器到Radiohead视频的制作……在本书中，你将：
探索海量在线数据集时面临的内在机遇和挑战
学习如何使用地图和数据“ 混搭 ”方式对都市犯罪趋势进行可视化
发现“ 众包 ”和透明如何改进药物研究现状
理解当新的数据和之前存在的数据交叠时如何向用户发送警告
学习处理DNA数据的大规模基础设施

《数据之美》

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

《数据之美》

书籍目录

第1章：在数据中观察生命
作者：Nathan Yau
第2章： 美丽的人们：设计数据收集方法时牢记用户
作者：Jonathan Follett, Matthew Holm
第3章：火星上的嵌入式图像数据处理
作者：J. M. Hughes
第4章：PNUTShell中的云存储设计
作者：Brian F. Cooper, Raghu Ramakrishnan, Utkarsh Srivastava
第5章：信息平台和数据科学家的兴起
作者：Jeff Hammerbacher
第6章：照片档案的地理之美
作者：Jason Dykes, Jo Wood
第7章：数据发现数据
作者：Jeff Jonas, Lisa Sokol
第8章：实时的可移动数据
作者：Jud Valeski
第9章：探寻Deep Web
作者：Alon Halevy, Jayant Madhavan
第10章：构建 Radiohead 的“ House of Cards ”
作者：Aaron Koblin, Valdean Klump
第11章：都市数据可视化
作者：Michal Migurski
第12章：Sense.us的设计
作者：Jeffrey Heer
第13章：数据所做不到的
作者：Coco Krumme
第14章：自然语言语料库数据
作者：Peter Norvi
第15章：数据中的生命：DNA漫谈
作者：Matt Wood, Ben Blackburne
第16章：美化真实世界中的数据
作者：Jean-Claude Bradley, Rajarshi Guha, Andrew Lang, Pierre Lindenbaum, Cameron Neylon, Antony Williams, Egon Willighagen
第17章：数据浅析：探索形形色色的社会定型
作者：Brendan O'Connor, Lukas Biewald
第18章：旧金山海湾之殇：次贷危机的影响
作者：Hadley Wickham, eborah F. Swayne, David Poole
第19章：美丽的政治数据
作者：Andrew Gelman, Jonathan P. Kastellec, Yair Ghitza
第20章：连接数据
作者：Toby Segaran

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。

《数据之美》

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

《数据之美》

精彩短评

1、众多的数据解决方案真实案例为我提供了很好的借鉴。
2、对于非专业的，很难。
3、我懂的看完依然懂，不懂的依然不懂，看了8章，弃……
4、数据,下一代的基本生产资料. 本书通过总结相关的信息收集系统向我们展示了信息收集中的各项问题与主要事项.
5、每章的内容都相互独立，没有个前后关系。内容也从数据库到可视化多而乱。
6、没办法，今年大数据太火，到处见广告，想不看都不行
7、说实话，感觉原版介绍的的内容还是不错的，举得例子涉及了很多领域，但是糟糕的是翻译是个外行，把数据之美重绘的一团糟，很多地方像在读google翻译的文字。
8、看不明白，可以么.....
9、不太懂，所以不知道好不好。
10、:
TP274/3232
11、在不同领域中数据获取，预处理，可视化分析的实践。
12、看目录可视化数据之类应该是很吸引我的，但是作者一直浮在表面讲案例，而且案例也没说出点普适的道理，几乎没什么干货，实在不值得一读。
13、在数据相关工作中浸泡过一段时间再来看，或许能收获更多
14、期待已久的好书，是《数据可视化之美》的姊妹篇。
15、这书可以说是数据可视化之美的前传，带有更多的案例分析和程序编写内容
16、可能水平还不够，有所收获但阅读过程挺痛苦，翻译得不行。
17、还可以吧，增广见闻不错
18、不大看得懂，感觉很专业。远不如《浪潮之巅》这种讲故事的来得吸引我
19、从采集来源、处理方式、展现形式等几个方面来介绍数据，值得一看
20、没怎么看懂 >_<
21、从技术的角度，不建议读，技术层面介绍的很少
从解决方案的角度，可以读一下，开阔视野。
另外，印刷质量一般
22、以后想做数据挖掘吧，但是那个啥，这本书有点范围广
23、内容一般
24、翻译很烂，电子版排版很烂，内容也过于零碎，挺让人失望的
25、介绍的东西很不错。了解了好的思路。
26、此本书既不讲解技术，也不讲解领域，通过对一些现实例子的分析来阐述数据带给我们的美。读完的感觉并不是技术上的收获，而是拓展了我们的视野。
27、一本数据科学家的散文集，对深入思考数据未来和解决实际问题的人会有帮助，不适合菜鸟。翻译生涩，可能是因为专业性太强。
28、所有括号里的英文都少了第二个字母，转换程序就算有BUG，难道就不手工查看吗？
29、涉及的领域相当广，对数据的表现方法和数据可视化以及用户交互这些方面都会有很好的启发。
30、刚刚收到书，打开发现合订的地方都歪了，裂了好几条纹，强烈要求换货
31、适合从事数据工作，并且有一定经验的人看，这本书可以用来开阔视野，里面有facebook和yahoo内部人讲的小故事，很不错！
32、快速看了几章，看不下去，觉得有点飘在空中，没有高瞻远瞩，也不是脚踏实地——总之，我不能领会
33、买了有两年了,最近研究相关内容,翻出来看看,内容组织特别混乱,个例之间毫无关联和铺垫可言,纯凑合骗傻子的书.
34、大数据分析和数字可视化可以渗透到各行各业，通过这些案例的解析，看自己所在的行业有哪些可以用来借鉴。推荐第17 18章
35、灰常好的一本书 不错啊

36、挺能启发思路的，但是，Google是敏感词么？一律翻译成了G公司= =
37、作为一个地图应用的开发者，对这本书大量和地图数据结合的数据展现案例自然是非常有感
38、书还没有仔细看，大略翻了下，纸张不错，内容的话排版有些太密了。而且，封面太丑了。
39、没事时看着玩儿
40、快速翻了前两章，很杂乱，不知道说什么
41、就是一论文合集，作为资料看一看，没啥特别的感觉
42、此本书既不讲解技术，也不讲解领域，通过对一些现实例子的分析来阐述数据带给我们的美。读完的感觉并不是技术上的收获，而是拓展了我们的视野。
43、此书为你揭示数据另一面。为你展示了数据其实并不枯燥。
44、比较专业的书籍，可以开拓思维。
45、从不同地方扒拉过来拼在一起的一本书，部分章节还凑合，全书应快速翻阅。
46、太专业看不懂
47、经验是可以学习的，此书必读
48、还未读，翻了翻感觉还好
49、电子版做的比较差，很多单词缺字母
50、20个实际例子，由一线人员亲自讲解，涉及到了多个领域领域的数据分析处理的多个方面。各个章节问题领域不同，讲解内容侧重点各异，知识点较为零散且笼统。用于入门了解不错。
51、科普性质的书～
52、关于数据处理的一系列案例，只讲了大的方向和处理方法。
53、可以多方面了解下当前时代时髦的技术，整体来说扩宽视野，增加了解还是不错的
54、封面很漂亮，像个猕猴桃，呵呵
55、思想是最重要的。了解不同领域的人数据处理的想法也很重要。。这本书的好处在于简单，没有讨论技术细节，也是一本内行看门道，外行看热闹的书。
56、个人觉得翻译较差，译的很是晦涩难懂，很多句子很明显的成分都不够，不推荐！
57、看了4章 弃了
58、蛮好的，现在流行大数据，看了很有启发
59、大数据的国外应用实例，开拓眼界。
60、部门总监推荐的书，适合做数据分析商户分析的同学们学习。你能从一些简单的数据中发现别人发现不了的问题
61、通过很多例子来讲述数据使用的各种demo，很不错的书
62、不是一本实战型的书
63、这本书相对来说还算比较新，是从英文版翻译过来的。英文版2009年出版。中文版2010年10月出版。由20篇相互独立的文章组成。每篇讲一个数据处理相关的项目。不涉及具体的技术细节，仅仅是概括说明原理、思路、过程、结果。总体来说，阅读起来有点晦涩。感觉作者基本都明白英文版的意思，不过有些地方中文表达上不够通顺。这在IT业的翻译书中已经算不错的组合了，强过中文过关但是不懂技术的情况。其中讲数据可视化的文章有几篇。还都比较有意思。比如第六章“照片档案的地理之美”，说的是英国的一个名叫“Geograph”的项目，收集了大量的英国的照片及普通用户对照片的标签，作者分析这些标签，并且用图形化的方法把许多分析结果展现出来；第11章“都市数据可视化”，讲的是把警察局的犯罪发生的数据与地图结合起来，预测犯罪发生的地点与类型从而提早预防；第12章“Sense.us的设计”讲以可视化手段分析美国150年以来的人口数据，得出许多有趣的结论；第17章“数据浅析：探索形形色色的社会定型”说的是用图形化方法分析一个网站的大量用户相互之间的评论；第19章“美丽的政治数据”同样使用可视化手段分析选举相关数据。第4章“PNUTShell中的云存储设计”，说的是雅虎的一个云存储的项目PNUTShell的设计思路和优缺点。这个项目面对的应用主要是社交方...面的应用，数据一致性要求不高，可用性、扩展性要求很高。因此就对一致性做了一些牺牲，满足比较高的可用性和扩展性。数据只要最终按照操作顺序执行了相关的操作，最终一致就可以了。每一条数据都记录了版本号，好知道自己执行到那个步骤了。每一条记录还需要记录自己是不是主备份。写操作要先写主备份然后逐步同步到其他数据库上。如果系统发现用户比较频繁地写数据但是主备份所在服务器的物理距离与用户的物理距离比较远，就自动把主备份记录转移到距离用户更近的服务器上。如果主备份损坏，系统也会从剩下的数据中挑选最合适的一条做主备份。第9

章“ 探寻Deep Web ”说的是如何让搜索引擎自动搜索Form表单。Form表单可以有无穷个组合，这篇文章给出一些基本思路来让搜索引擎判断如何去选择下拉列表或者去填写文本框，目标是用尽量少的操作步骤尽量多地获取form表单后面的数据库中的内容。 阅读更多 ›
64、还没有仔细看，但内容不错
65、内容比较杂，但是有好多实践的经验是其他树立找不到的。
66、需要一定基础才能看懂。
67、这本书深入浅出，讲述了大数据应用的多个方面，有利于深入开发互联网资源，为企业升级服务。
68、这本书相对来说还算比较新，是从英文版翻译过来的。英文版2009年出版。中文版2010年10月出版。由20篇相互独立的文章组成。每篇讲一个数据处理相关的项目。不涉及具体的技术细节，仅仅是概括说明原理、思路、过程、结果。

总体来说，阅读起来有点晦涩。感觉作者基本都明白英文版的意思，不过有些地方中文表达上不够通顺。这在IT业的翻译书中已经算不错的组合了，强过中文过关但是不懂技术的情况。

其中讲数据可视化的文章有几篇。还都比较有意思。比如第六章“ 照片档案的地理之美 ”，说的是英国的一个名叫“ Geograph ”的项目，收集了大量的英国的照片及普通用户对照片的标签，作者分析这些标签，并且用图形化的方法把许多分析结果展现出来；第11章“ 都市数据可视化 ”，讲的是把警察局的犯罪发生的数据与地图结合起来，预测犯罪发生的地点与类型从而提早预防；第12章“ Sense.us的设计 ”讲以可视化手段分析美国150年以来的人口数据，得出许多有趣的结论；第17章“ 数据浅析：探索形形色色的社会定型 ”说的是用图形化方法分析一个网站的大量用户相互之间的评论；第19章“ 美丽的政治数据 ”同样使用可视化手段分析选举相关数据。

第4章“ PNUTShell中的云存储设计 ”，说的是雅虎的一个云存储的项目PNUTShell的设计思路和优缺点。这个项目面对的应用主要是社交方面的应用，数据一致性要求不高，可用性、扩展性要求很高。因此就对一致性做了一些牺牲，满足比较高的可用性和扩展性。数据只要最终按照操作顺序执行了相关的操作，最终一致就可以了。每一条数据都记录了版本号，好知道自己执行到那个步骤了。每一条记录还需要记录自己是不是主备份。写操作要先写主备份然后逐步同步到其他数据库上。如果系统发现用户比较频繁地写数据但是主备份所在服务器的物理距离与用户的物理距离比较远，就自动把主备份记录转移到距离用户更近的服务器上。如果主备份损坏，系统也会从剩下的数据中挑选最合适的一条做主备份。

第9章“ 探寻Deep Web ”说的是如何让搜索引擎自动搜索Form表单。Form表单可以有无穷个组合，这篇文章给出一些基本思路来让搜索引擎判断如何去选择下拉列表或者去填写文本框，目标是用尽量少的操作步骤尽量多地获取form表单后面的数据库中的内容。
69、说实话，这本书涉及的范围和深度比较复杂，好多章节看不太懂，大到天文，小到dna，各领域的数据分析的各个层面均有涉及，类似论文集的编排方法，也难以一以贯之，感受比较深的是语料分析的应用大开眼界，竟然想到了替换密码的破译，其他关于暗网的搜索，数据搜集表单设计，均有启发。可以翻阅，开拓视野。
70、比《大数据时代》在技术上更深入，又不是特别深入，对于我这种伪GEEK来说，刚刚好。一般俺不爱挑翻译的毛病，不耽误看就成。不过这本书，着实读着别扭，求店家送英文版……
71、每一个案例都是当前比较流行的网站数据架构方式，很值得一读，可以开阔眼界！
72、到写数据的部分就翻译的不错，一碰到大段论述就各种混乱。内容很有启发性，继续看吧。 大致翻完一遍，专业性非常强。。。所以，看不下去了。
73、考虑到这本书的出版年份，确实是当时的各种漂亮案例的集合。从数据的采集，处理，分析， 各方面的案件都有。尤其是最后几个比较全面的案例，看着不错， 令人感慨：数据有多美丽，在于想像有多美丽。
74、导师推荐的。但我读起来比较费劲。
75、纯英文，学术很强，完全看不进去，我想适合这个专业的博士或者研究人员阅读，不适合程序员

76、大致翻了一遍，有些章节值得仔细学习
77、书本质量还可以，印刷也不错。单篇文档内容还可以，但本书就值缺乏一个整体的思想，显得有些杂乱！建议内容挑选上做一些改进。
78、该给6星的书，，可惜了烂翻译……
79、对搞研究来说，并没有什么卵用
80、只能欣赏，上面的知识一时半伙还用不上，全当扩充知识面了
81、很有见地，文字轻巧却不乏教导意义,作者经验丰富，书中实例可操作性强
82、不是实操新型的书，各章节来源于美国各大顶尖技术公司
83、没想到还有这个的中文版，14章统计自然语言节，比《数学之美》详细多了！推荐
84、既没得到阅读的趣味，有没学到数据挖掘的知识。。
85、以前看过部分章节，觉得很好，这次活动顺便就买了
86、一本杂七杂八的书
87、太难了，完全看不懂在说什么，已GG
88、非常不错的数据分析入门书，这本书由很多篇article组成，讲诉了关于数据分析、数据可视化、链接数据等相关的若干项目，更多的是开阔了我的视野，在我研一期间给我一个很好的视野。
89、一是纸质、印刷都很棒，二是英文也不是太难，当然我只看了几篇的开头……
90、人造卫星中转站
91、可能偏向于技术层面吧，所以我感觉不出来作者对于数据美感的说明
92、对于我一个外行人来说，看完并没有感受到数据美的地方。
数据有不止一次提到数据和结果的因果性和相关性，这本书强调因果性。
书中的例子都是从项目的和要求入手的，来找数据。
再次验证弄明白诉求很重要

93、个人感觉更偏向于做数据展示。给处理数据的人提供了一些指引和思路
94、应该是正品，有防伪标志，东西不错 物流太给力了，第二天就收到了
95、我看不太懂。但是很解闷。
96、需要过段时间看才能感觉
97、Kindle版恐怕是校对得不认真，出现很多不应该的错漏，影响阅读。
98、可以通过这本书了解下数据挖掘、数据分析领域相关的研究方法
99、太晦涩了，看不懂。
100、看了两章，内容不是想象中的讲述机器学习及数据挖掘的书籍。讲的是数据在不同领域的应用，感觉没什么收获。
101、我自己的错，现在已经没有耐心读进纯技术的书籍了.. 唉

《数据之美》

精彩书评

1、一直认为o'really出的书都带有很重的哲学色彩，适合菜鸟和大神阅读，这本“ 菊花 ”版的也不例外。诚如副标所题“ 背后的故事 ”，该书根据数据的 ”提取-处理-可视化“ 松散的排列思路，选取了20个 ”优雅的数据解决方案“ 。作为数据挖掘的新生信徒，关注该书的初衷来源于对个人数据的整理，亦即该书第一章，内容展现的是方案设计过程中的一些思想和实用的原则，以及一些关键的经验和技巧，一如该书其他章节。虽是跑马观花地浏览，但从中了解了大神在对待大数据时的好奇和激情，对问题分析的庖丁解牛和对技术运用的游刃有余，并加深了对数据处理的目的性和实验性认识，随之而来的便是对数据进行探索的一种冲动，希望再读时，是同样的热情。
2、感觉买亏了，看到评价还行，但是没看人数，这是致命的啊！看了一下，感觉没什么意思，弱弱的。各行业的术语都出来了，各种无法理解啊，有些还有点像流水账。但也有可能是我自己的原因吧？覆盖行业比较广泛，可以挑自己熟悉的行业看看就行，没必要全读。感觉更像一本大杂烩。
3、这本书相对来说还算比较新，是从英文版翻译过来的。英文版2009年出版。中文版2010年10月出版。由20篇相互独立的文章组成。每篇讲一个数据处理相关的项目。不涉及具体的技术细节，仅仅是概括说明原理、思路、过程、结果。总体来说，阅读起来有点晦涩。感觉作者基本都明白英文版的意思，不过有些地方中文表达上不够通顺。这在IT业的翻译书中已经算不错的组合了，强过中文过关但是不懂技术的情况。其中讲数据可视化的文章有几篇。还都比较有意思。比如第六章“ 照片档案的地理之美 ”，说的是英国的一个名叫“ Geograph ”的项目，收集了大量的英国的照片及普通用户对照片的标签，作者分析这些标签，并且用图形化的方法把许多分析结果展现出来；第11章“ 都市数据可视化 ”，讲的是把警察局的犯罪发生的数据与地图结合起来，预测犯罪发生的地点与类型从而提早预防；第12章“ Sense.us的设计 ”讲以可视化手段分析美国150年以来的人口数据，得出许多有趣的结论；第17章“ 数据浅析：探索形形色色的社会定型 ”说的是用图形化方法分析一个网站的大量用户相互之间的评论；第19章“ 美丽的政治数据 ”同样使用可视化手段分析选举相关数据。第4章“ PNUTShell中的云存储设计 ”，说的是雅虎的一个云存储的项目PNUTShell的设计思路和优缺点。这个项目面对的应用主要是社交方面的应用，数据一致性要求不高，可用性、扩展性要求很高。因此就对一致性做了一些牺牲，满足比较高的可用性和扩展性。数据只要最终按照操作顺序执行了相关的操作，最终一致就可以了。每一条数据都记录了版本号，好知道自己执行到那个步骤了。每一条记录还需要记录自己是不是主备份。写操作要先写主备份然后逐步同步到其他数据库上。如果系统发现用户比较频繁地写数据但是主备份所在服务器的物理距离与用户的物理距离比较远，就自动把主备份记录转移到距离用户更近的服务器上。如果主备份损坏，系统也会从剩下的数据中挑选最合适的一条做主备份。第9章“ 探寻Deep Web ”说的是如何让搜索引擎自动搜索Form表单。Form表单可以有无穷个组合，这篇文章给出一些基本思路来让搜索引擎判断如何去选择下拉列表或者去填写文本框，目标是用尽量少的操作步骤尽量多地获取form表单后面的数据库中的内容。

《数据之美》

章节试读

1、《数据之美》的笔记-第215页

大规模地搜索没有因果关系的相关性属于偶然计算，不是科学。即使是所谓的大数据，科学依然是一个很强的由假设驱动的过程。

2、《数据之美》的笔记-第7页

your.flowingdata(YFD)

个人数据收集从用户角度，我们需要使得数据收集变得尽可能简单。它应该是无干扰的、直观的且易于访问的，这样数据收集才更有可能成为日常生活的一部分。

3、《数据之美》的笔记-第302页

第17章 数据浅析：探索形形色色的社会定型
通过日常生活的数据感知年龄、性别、智商和魅力每天我们都通过自己购买的东西、使用的Web站点、搜索的查询、发送的消息和去过的地方 方方面面地展现自己。
我们的日常生活无时无刻在产生大量混乱、无序、碎片化的信息，这些集合隐藏着关于我们的某种模式。

4、《数据之美》的笔记-第109页

一个组织最多只和它的所有成员的洞察力总和一样有智慧。
这句话类似于人不可能靠自己把自己提起来。人做的系统归根到底得人能理解才行，所以神经网络失败了。

5、《数据之美》的笔记-第10页

数据存储
PEIR采用了PostGIS，为PostgreSQL数据库增加地理对象的支持。
YFD 采用了Django，基于Python语言的MVC模式，支持敏捷高效开发。
PEIR：http://peir.cens.ucla.edu

6、《数据之美》的笔记-第26页

表单布局设计：

Web表单排版和可访问性；
给人们一些空间；
适应于不同的浏览器，并测试兼容性；
交互设计考虑：动态表单长度；
设计信任；
为精准的数据收集而设计；
动机；
报告即时数据结果；

7、《数据之美》的笔记-第12页

Mass Observation

8、《数据之美》的笔记-第21页

第2章 设计数据收集方法时必须始终牢记受众的期望和需求
数据收集面临的挑战：可访问性、信任和用户动机

用户体验User Experience直接关乎主动数据收集的质量，多花点心思设计一个表单是值得的！考虑表单排版、兼容性、动态的交互和反馈、建立目标用户的心理信任和配合的兴趣。

9、《数据之美》的笔记-第9页

记下

10、《数据之美》的笔记-第128页

系统的实时处理能力意味着避免不必要的整体轮循，而只触发少数几个频繁更新的用户吗？

11、《数据之美》的笔记-第6页

“ 虽然搜集和返回的数据类型可能已经随时间变化，但是每个人的需求是不变的。也就是说，那些收集关于自己和他们周围数据的个人，他们还是会收集这些数据，以获得对流动的数据的信息更好的理解。绝大多数时候，我们不是追求数据本身；我们感兴趣的是数据的真正含义。这是个微小的差别，却是非常重要的一点。这个需求要求系统能够处理个人数据流，高效准确的处理这些数据，把这些信息通过易于理解且有用的方式发给非专业人员。我们想要的远远不只是一个电子表格的数据，我们想要的是隐含在这些数据中的故事。 ”

翻译都没有人是说中国话的么？理解这么一段话都要累死了有木有，这是逼得人非得看原文不可。。。。。

12、《数据之美》的笔记-第27页

制作数据地图
工具：Modest Maps + Open Street Map
1、贴图数据
2、选择颜色机制红色通常表示停止或者前方有危险，而绿色则意味着进展或者增长，尤其是站在环境的立场上看。3、考虑交互性
4、呈现
5、分享

13、《数据之美》的笔记-第26页

设计数据收集方案时要遵循的一些指南：
尊重用户
在整个设计过程中应该保持以人为本，需要了解考虑用户的情绪反应。
他们不是傻瓜，而是我们的潜在客户。
展现真是的人们
用角色替身来指导我们思考，包括年迈的父母、一些很了解其先前情况的商业伙伴。

14、《数据之美》的笔记-第5页

第一章，在数据中观察生活

定义 - 收集 - 存储 - 处理 - 可视化

15、《数据之美》的笔记-第6页

在数据中观察生活绝大多数时候，我们不是追求数据本身；我们感兴趣的是数据的真正含义。这个微小的区别，却是非常重要的一点。这个需求要求系统能够处理个人数据流，高效准确地处理这些数据，把这些信息通过易于理解且有用的方式分发给非专业人员。个人环境影响报告（Personal Environmental Impact Report, PEIR）我们重点利用日常的移动技术（如手机）来收集关于周围和自己的数据，因此人们可以对如何与身边的事物进行交互有更好的理解。例如：DietSense是一个在线服务，它允许人们自我监测饮食选择以及进一步向饮食专家咨询；Family Dynamics帮助家庭和生活教练记录一个家庭每日交互的关键特征，如户外驻扎和家庭聚餐；Walkability帮助居民和行人，提倡通过观察发表他们对于附近的步行适宜性和与公共交通的联系的看法。

16、《数据之美》的笔记-第1页

第一章的作者介绍了两个他以及他参与的项目，个人环境影响报告（PEIR）和your.flowingdat(YFD)，两个项目都是利用“移动”（智能手机，Twitter）大作文章。

PEIR似乎是测量个人的碳排放量，利用手机的GPS功能，每间隔几分钟定位来测算移动的速度，判断碳排放，还有和facebook上朋友分享以及比较碳排放的功能。可能个人比较愚钝，没太搞懂这个项目都要做什么，怎么做。

第二个项目YFD是作者自己的小玩意，利用twitter让用户发送自己的状态，比如吃的菜，在睡觉，开心的情绪，记录自己每天的生活。等到你想利用这些数据时，比如你想减肥，你的目标是“I want to fit into my pants - all of them”，你可以查看到你最近都吃了哪些菜，喝了些什么，你现在有多重？

总结来看，这两个项目都是对个人生活方式的一种数据化记录，并通过数据可视化技术来友好地把结果返回给用户，使其获益，以便于继续乐于提供自身数据信息来供人们研究生活方式。现代的个人数据收集方法要比Mass Observation方便多了！

17、《数据之美》的笔记-第122页

什么《数据发现数据》
这个垃圾文章
总是在说系统应该怎样、能够怎样
却对技术细节什么也不谈
这说明作者是个满嘴胡吹的家伙
另外
他所举例的用处不是为广大的用户带来利益
而是如何帮助赌场发现作弊行为
——这也tmd太恶心了

18、《数据之美》的笔记-第112页

并生成一条即时报警信息：５７６４号员工和４４００３２１号被逮捕人有相同的电话号码！这个场景很生动，经常在犯罪心理里看到，而且计算量也不大，我们的工作里也可以弄这么一套发现系统。

19、《数据之美》的笔记-第9页

异步数据收集
YFD 通过解析文本时间，解决不能及时上传数据的瓶
PEIR 通过缓存把数据存储在手机本地的内存，直到手机可以重新连上网络才上传数据，解决网络连接不可能100%存在的问题。

存在的问题是：期望人们在事件发生时收集数据是不合理的。人们会忘记或者不方便收集数据。因此，提供用户在后期也能够输入数据的功能是很重要的，这一点又反过来影响了数据流的下一步设计。

20、《数据之美》的笔记-第110页

数据发现数据

实时发现的好处
立足于“ 数据发现数据 ”的高级信息管理系统，不会依赖于用户向计算机凭空提出正确的、相关的和即时的问题。