HZ BOOKS
华章IT

数据分析与决策技术丛书

[PACKT]
PUBLISHING

Mastering Python Data Visualization

Python数据可视化

[印度] 科斯·拉曼（Kirthi Raman） 著

程豪 译

全面讲解Python在不同应用领域的可视化方法

涵盖Python的各种绘图选项，包含大量实际应用案例

机械工业出版社
China Machine Press

数据分析与决策
技术丛书

Mastering Python Data Visualization

Python数据可视化

[印度] 科斯·拉曼（Kirthi Raman） 著

程豪 译

机械工业出版社
China Machine Press

图书在版编目（CIP）数据

Python 数据可视化 /（印度）科斯·拉曼（Kirthi Raman）著；程豪译 . —北京：机械工业出版社，2017.3
（数据分析与决策技术丛书）
书名原文：Mastering Python Data Visualization

ISBN 978-7-111-56090-6

I. P… II. ① 科… ② 程… III. 软件工具 – 程序设计 IV. TP311.56

中国版本图书馆 CIP 数据核字（2017）第 032016 号

本书版权登记号：图字：01-2016-1889

Python 数据可视化

出版发行：机械工业出版社（北京市西城区百万庄大街 22 号　邮政编码：100037）
责任编辑：缪　杰　　责任校对：李秋荣
印　　刷：三河市宏图印务有限公司　　版　　次：2017 年 3 月第 1 版第 1 次印刷
开　　本：186mm × 240mm　1/16　　印　　张：17.75
书　　号：ISBN 978-7-111-56090-6　　定　　价：69.00 元

凡购本书，如有缺页、倒页、脱页，由本社发行部调换
客服热线：（010）88379426　88361066　　投稿热线：（010）88379604
购书热线：（010）68326294　88379649　68995259　　读者信箱：hzit@hzbook.com

The Translator's Words 译者序

海量信息的不断增长，不断刺激着读者对数据可视化的渴望与诉求。作为一种功能强大的开源编程语言，Python 包含了丰富的软件包和绘图技术，从而帮助用户完成数据分析、构建统计模型并展现研究结果。

本书尤其关注 Python 在众多应用领域中的可视化功能，全面覆盖 Python 的各种绘图选项，配合丰富的实际案例，为 Python 初学者和资深人士提供了一本实用指南。对于 Python，我不敢自称有丰富的实战经验，但却有过自学和运用的经历。在承担本书翻译工作的同时，我自己也重温了一次 Python 可视化之旅，收益颇多。故劝荐诸位，不妨深读此书，系统体验 Python 在数据可视化方面的贡献。与音乐一样，知识的传播没有国界。因此，翻译不仅是知识表达语言的转换，更是一次学习和交流的机会。与原作者对话，高山仰止，受益匪浅；与读者对话，高山流水，闻过则喜。

在此，感谢我的朋友钟琰在整个翻译过程中提供的帮助。感谢我的至爱刘钰洁在译稿校对阶段给出的建议。我要感谢我的博士生导师——中国人民大学的易丹辉教授。感谢我在美国联合培养期间的导师——美国哥伦比亚大学的韦颖副教授。特别感谢我父母和家人，是他们给予我前行的动力和勇气。最后，非常感谢机械工业出版社华章公司的编辑让我接触到这本书，并给予中肯建议。感谢身边所有的良师益友。

鉴于个人时间与水平有限，如有纰漏，还望各位读者予以反馈，不吝赐教！

程豪

2016 年 12 月 15 日

前　言 *Preface*

数据可视化旨在清楚明了地提供信息，帮助读者定性理解这些信息。俗话说，一图胜千字（百闻不如一见）。这里，可以换个说法，“一幅图讲述了一个故事，如同万语千言。”因此，可视化是一个宝贵的工具，有助于读者快速理解相应的概念。然而，与其说数据可视化是一种技能，还不如说它是一门艺术。这是因为，如过度使用数据可视化会适得其反。

当前，有太多数据需要处理。这些数据包含着许多见解，这些见解是成功的关键。能够发现数据、清洗数据，并使用正确的工具实现可视化至关重要。本书讲解了用 Python 软件包实现数据可视化的不同方法，并给出很多不同领域的案例，比如，数值计算、金融模型、统计和机器学习，以及遗传学与网络。

本书提供在 Mac OS X 10.10.5 系统上运行的案例程序，具体用到 Python 2.7、IPython 0.13.2、matplotlib 1.4.3、NumPy 1.9.2、SciPy 0.16.0 和 conda 构建 1.14.1 版本。

本书主要内容

第 1 章阐述了数据可视化确实应该被称为“用于知识推断的数据可视化”。本章包含框架，讲解数据 / 信息如何转换为知识，以及有意义的呈现方式（通过取对数、颜色映射、散点图、相关性以及其他）如何能够帮助我们更容易地掌握知识。

第 2 章讲述可视化的重要性，展示可视化过程中的一些步骤，包括可选择的几种工具选项。可视化方法由来已久，很早之前我们就接触过这些方法；比如，连年幼的小孩都能解释条形图。交互式可视化有很多优点，本章将举例说明。

第 3 章解释了从 Continuum Analytics 使用 Anaconda 时，不必安装每个 Python 库的原因。Anaconda 有简化的打包和部署方法，这些方法使得 IPython notebook 与其他库的并行运算变得更加容易。

第 4 章包括交互式绘图方法及在计算物理和应用数学中的实践案例。一些著名的案例

包括用 SciPy 实现插值方法、近似、聚类、抽样、相关关系和凸优化。

第 5 章探索金融工程，该领域有很多数值计算和图表绘制的方法，是探索 Python 的一个有趣的案例。本章通过举例讲述股票报价、回归分析、蒙特卡洛算法和模拟方法。

第 6 章包含了用 NumPy、SciPy、matplotlib 和 scikit-learn 等工具进行处理的统计方法，比如，线性、非线性回归、聚类和分类。

第 7 章包含了有趣的案例，比如社交网络以及现实生活中的有向图举例，适用于这些问题的数据结构，以及网络分析。本章会用到一些具体的库，比如 graph-tool、NetworkX、matplotlib、scipy 和 numpy。

第 8 章包含模拟方法和信号处理案例，用以展示一些可视化方法。这里，我们也给出了其他高级工具的对比，比如 Julia 和 D3.js。

附录给出了 conda 概述，并列出多种 Python 库。

学习本书的准备工作

本书要求用户在操作系统上安装 2.7.6 或以上版本的 Python。对于书中的案例，可以使用 Mac OS X 10.10.5 的 Python 默认版本（2.7.6）来实现。其他会用到的软件包是 IPython——一个交互式 Python 环境。新版的 IPython 叫 Jupyter，该版本现在有 50 种不同语言的内核函数。

安装提前打包好的用于科学计算的 Python 发行版，如果可能的话，可以从 Continuum 安装 Anaconda，或安装 Enthought Python Distribution。Anaconda 一般自带 300 多个 Python 软件包。你可以用 pip 或 conda 安装不在自带软件包列表中的 Python 软件包。有一些案例可见附录。

本书适用对象

目前已有很多 Python 和数据可视化方面的书。然而，对于有一定 Python 知识储备的人来说，几乎很少有把两者内容结合在一起的书值得推荐。有关简化代码、重复使用的小生境（niche）技术的讨论更是少之又少。对于有强烈学习兴趣的 Python 开发人员，本书将提供一系列获得分析结果和产生惊人可视化效果的方法。

本书提供了解决实际问题的一系列分析方法。虽然本书并不是面向初学者的，但是如果有需要，你可以搜索书中推荐阅读的文献资料。如果这是你初次体验 Python 编程或数据可视化，提前阅读一些入门教材会有很大帮助。我最喜欢的书有 John Guttag 教授的《Introduction to Computer Science and Programming》（可从 MIT OpenCourseWare 上免费下载）和来自 UCLA 的 Nathan Yau 的《Visualize This》。

目　录 *Contents*

第 1 章 Chapter 1

数据可视化概念框架

当代，网络和社交媒体的兴起，产生了大量数据，而且数据量的增长已超乎想象。这种现象是怎么发生的？又是何时发生的？

十年前，一种处理问题的新方法演变为：跨企业的从数据源收集、整合大量数据，并进行运算的研究工作。他们这样做的目标是用海量数据改善决策过程。在此期间，促使 Amazon、Yahoo 和 Google 这样的公司在处理大量数据方面取得了显著进展。这些里程碑式的成就促使一些大数据分析技术的诞生。当然，我们不会追究大数据的细节问题，但是我们将尝试探索，为什么很多机构改变了他们以往的模式，用类似的想法获得更好的决策。

到底如何用这些海量数据做出更好的决策？这是我们的终极目标，但首先让我们理解数据、信息和知识间的差异，以及它们与数据可视化之间的关系。或许会有这样一个疑问，为什么要讨论数据、信息和知识。我们将就下面的脉络具体展开：怎样开始、用什么开始、这些内容如何有益于问题解决，以及可视化的作用。我们将通过简要回顾涉及的程序步骤，确定数据可视化所需的概念框架。

本章将包括以下主题：

- ❑ 数据、信息、知识和观点之间的差异
- ❑ 信息转化为知识，进而转化为观点
- ❑ 收集、处理和组织数据
- ❑ 数据可视化的历史
- ❑ 数据可视化如何帮助决策
- ❑ 可视化图像

1.1 数据、信息、知识和观点

数据、信息和知识被广泛用于计算机科学领域。通常，这些术语有很多种充满争议且不相一致的定义。在深入研究这些定义之前，我们先理解这些术语与可视化之间的关系。数据可视化的主要目标是从数据或信息中得出观点（隐含的真理）。本书有关数据、知识和观点的整个讨论属于计算机科学的范畴，而非心理学或认知科学。认知科学方面的文献请参见：https://www.ucsf.edu/news/2014/05/114321/converting-data-knowledge-insight-and-action。

1.1.1 数据

数据是得出结论的前提。尽管在一些特定的背景下，数据和信息看起来相关联。但实际上，数据是离散、客观事实的数字表示。作为后续工作的基础，数据会有不同的组织和安排形式，以方便得到回答实际问题的有用信息。

数据可以是非常简单却庞大冗杂的。离散数据本身不能用于决策。这是因为它没有意义，而且更重要的是，它们之间没有结构或关系。数据收集、转换和储存的过程因数据类型和储存方法的不同而有很多变化。数据有很多形式，一些常见形式如下：

- CSV 文件
- 数据库表格
- 文件格式（Excel、PDF、Word 等）
- HTML 文件
- JSON 文件
- 文本文件
- XML 文件

1.1.2 信息

信息是处理后的数据，为实际问题提供答案。当我们增加一种关系或一个关联时，数据就成为信息。这种关联通过提供数据背景来完成。这个背景有助于我们回答数据相关的问题。

比如，我们假定一名篮球员的数据包含身高、体重、位置、大学、出生日期、应招入队，选拔轮数，NBA- 登场和新成员排名。问题“哪位球员是首位应征入队、身高在 6 英尺[⊖]以上且担任控球后卫？”的回答是一条信息。

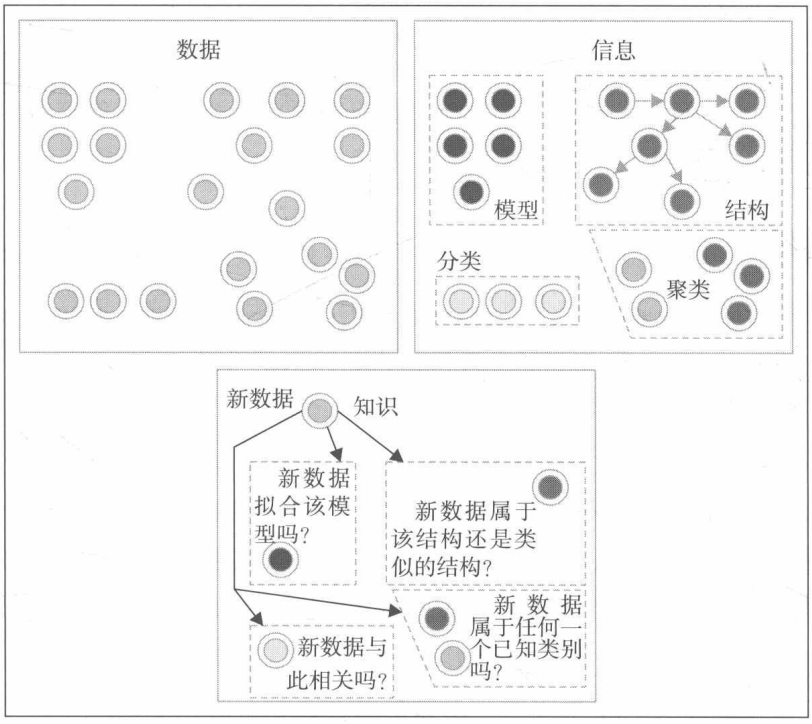
类似地，每个球员的得分也是一条数据。问题“今年每次比赛得分最高的选手是谁？分数是多少？”的回答“LeBron James，27.47”同样也是一条信息。

⊖ 1 英尺 =0.3048 米。

1.1.3　知识

当人类解释和组织信息，并用以决策时，知识便应运而生。知识是数据、信息和通过经验获得的技能。知识包括做出适当决策的能力和执行时所需的技能。

作为必不可少的部分（连接数据）允许我们理解每条信息的相对重要性。通过比较过去的结果和识别模式，我们不必从头开始寻找问题的解决方法。下图总结了数据、信息和知识的概念。



知识以不断增长的方式发生变化，特别是当信息被重新安排或被重新组织，或在一些计算算法发生变化时。知识像箭一样直击算法的结果，该算法与来自数据的过去信息有关。在许多情况下，可以通过与结果的视觉交互获得知识。另一方面，观点开启了通向未来的途径。

1.1.4　数据分析和观点

在我们深入研究观点的定义及其如何与实际问题相关联之前，我们不妨先看看如何获取观点。十年间，组织机构已尽力弄懂他们拥有的所有数据和信息，特别是探索数据量的大小。为了基于已有数据信息得到最佳或现实的决策，他们发现了数据分析的重要性（也就是**数据分析学**或**分析学**）。

分析学依赖数学算法来确定产生观点的数据间的关系。一种简单的方式是通过打比方来理解观点：当数据没有结构且与实际问题相对应时，通过将数据结构化，使其更接近实际目标，这有助于人们更清晰、更深刻地理解数据。观点是“我发现了”的那个时刻，得到突破性的结果。一个人不应该困惑于术语分析学和商务智能。当商务智能提供基于历史数据的分析结果时，分析学就具备了预测能力。

分析学通常用于更广泛的数据，为此，数据内外之间的协作时常发生。在一些实际问题的范式中，这种协作仅发生在海量数据的内部，但在大多数情况下，加入外界信息有助于链接点或完成拼图。最常见的两个外部数据链接源是社交媒体和用户群体。

在本章，我们应用分析法理论得出观点、驱动商业价值，以及改善决策和更好地理解用户，我们得出真实生活故事中有价值的结论。

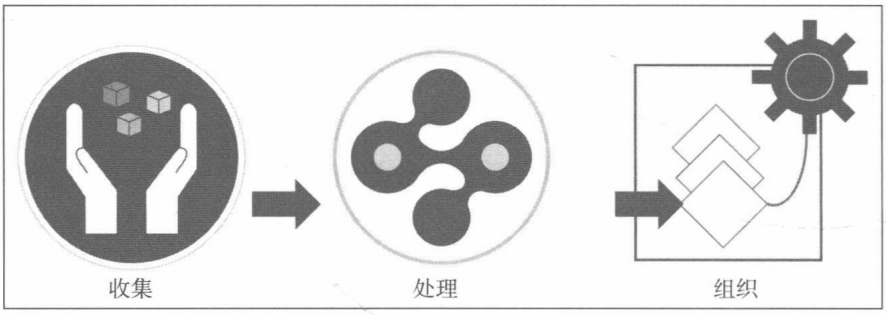
1.2 数据转换

现在，我们了解了数据的定义，但问题是：为什么要收集数据？数据对于描述物质或社会现象以及进一步回答这些问题非常有用。出于这个原因，确保数据的无误、精确和完整是很重要的；否则，错误、不精确和不完整的数据将导致响应结果的不精确或不完整。

数据有不同种类，其中包括*过去表现数据*、*实验数据*和*基准数据*。过去表现数据和实验数据当然很容易理解。另一方面，基准数据是用一个测度标准来比较两种不同项目或产品的特征。数据被转换为信息，得到进一步处理，然后用来解答问题。因此，很明显下一步就是转换的实现。

1.2.1 数据转换为信息

根据数据的内容和重要性，数据收集和储存有一些不同的方式。例如，如果数据是关于篮球季后赛的，那么这些数据将储存为文本和视频格式。另一个例子是一个国家所有城市的温度记录，这些数据通过不同形式收集得到。从数据转换为信息包含数据的收集、处理和组织，如下图所示：



收集来的数据需要处理和组织过程，这些数据后续可能没有结构、没有模型或没有模式。然而，该处理过程至少给我们一种从数据中发现问题答案的组织方式。这种处理可以是一种基于篮球员总得分的简单分类，或者根据城市和州名的分类。

从数据到信息的转换也可以不仅仅是分类，比如统计建模或计算算法。将数据转换为信息确实很重要，这样数据可以被查询、访问和操作。海量数据的转换可能包括这样几种处理方法：过滤、聚集、应用相关性、归一化和分类。

1. 数据收集

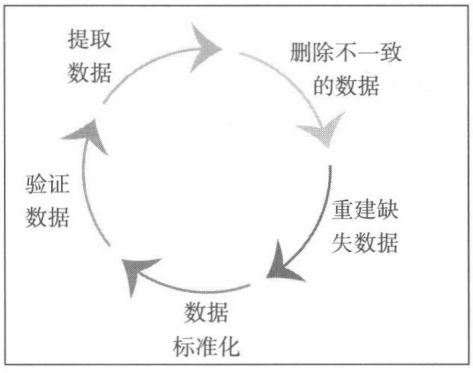
数据收集是一个耗时的过程。因此，人们正在寻找更好的自动数据采集方法。然而，人工数据收集仍然很常见。如今，数据的自动收集过程用到输入设备，比如传感器。例如，通过传感器检测水下珊瑚礁；农业上用传感器检测土壤性质、控制灌溉和施肥方法是另一个应用领域。

另一种自动收集数据的方法是通过扫描文档和日志文件，这也是一种服务器端数据收集的形式。人工处理包括基于网络且储存于数据库的数据收集方法，这些数据可以转换为信息。现在，基于网络的协作环境正受益于交流改善和数据分享。

传统的可视化和可视化分析工具专门为单个用户、单机可视化应用而设计。将这些工具的功能拓展到支持协作的层面需要一个漫长的过程，才能扩大真实世界中可视化的适用范围和应用领域。

2. 数据预处理

如今，基于数据量、数据来源的多重异质性和数据类型的不同，数据很容易受到噪音和不一致的影响。现有一些数据预处理技术，比如*数据清洗*、*数据集成*、*数据压缩*和*数据转换*。数据清洗用于数据中的噪音清理和矛盾修正。数据集成将多个数据源的数据合并起来，通常被称为数据仓库。数据压缩可以通过诸如合并、聚集和消除冗余特征等方法减少数据量。数据转换将数据缩放到一个较小的区间，从而提高处理和可视化的精确性和效率。数据的转换周期如下图所示：



异常值检测是非常规数据的识别，这些数据可能不会落入收集数据的预期行为或模式。异常值也称为离群点或噪音；比如信号数据，一个非常规的特别信号被视为噪音。交易数据中的一个离群点是欺诈交易。准确的数据收集对于保持数据完整性必不可少。然而，从另一角度考虑，异常值也非常重要，比如寻找诈骗保险理赔。

3. 数据处理

数据处理是转换过程中的重要一步。当务之急是关注数据质量。依存模型和聚类有助于准备分析数据和更好地理解处理步骤。虽然也有其他处理技术，但是我们在这不做过多赘述，仅以两种最受欢迎的处理方法为例。

依存模型是建模数据以确定表现方式性质和结构的基本原则。该过程寻找数据元素间的关系；比如，百货公司可能收集顾客购买习惯的数据。该过程有助于百货公司减掉频繁购买的信息。

聚类是在数据中发现群组，从某种方式上看，“相似性模式”没有用数据中已知的结构。

4. 组织数据

数据库管理系统允许用户以结构化的形式存储数据。然而，数据库太大而不能存入内存。有以下两种结构化数据的方法：

- 以结构化的形式将大量数据储存到磁盘中，比如，表、树或图表
- 为了快速访问，以结构化的形式将数据储存到内存中

数据结构由将数据结构化为可被储存和访问的一系列不同格式构成。常用的数据结构类型有数组、文件、表、数、列表、映射等。任何数据结构都是为特定目的而设计的，通过组织数据来进行数据储存、访问和操作。一种数据结构可能被选择或设计来储存数据，以实现用不同算法更快访问的目的。

经过高效收集、处理和组织所存储的数据，使数据更容易被理解，这也有助于更好地理解数据中蕴含的信息。

5. 获取数据集

针对接触不到组织数据的读者，下面列举出一些丰富的数据集资源：

- http://grouplens.org（来自明尼苏达大学）
- http://ichart.finance.yahoo.com/table.csv?s=YHOO&c=1962
- http://datawrangling.com/some-datasets-available-on-the-web
- http://weather-warehouse.com（天气数据）
- http://www.bjs.gov/developer/ncvs/（Justice 统计局）
- http://census.ire.org/data/bulkdata.html（人口普查数据）

- http://www.pro-football-reference.com（足球参考）
- http://www.basketball-reference.com（篮球参考）
- http://www.baseball-reference.com（棒球参考）
- http://archive.ics.uci.edu/ml/datasets.html（机器学习）
- http://www.pewresearch.org/data/download-datasets/
- http://archive.ics.uci.edu/ml/datasets/Heart+Disease（心脏病）

1.2.2　信息转换为知识

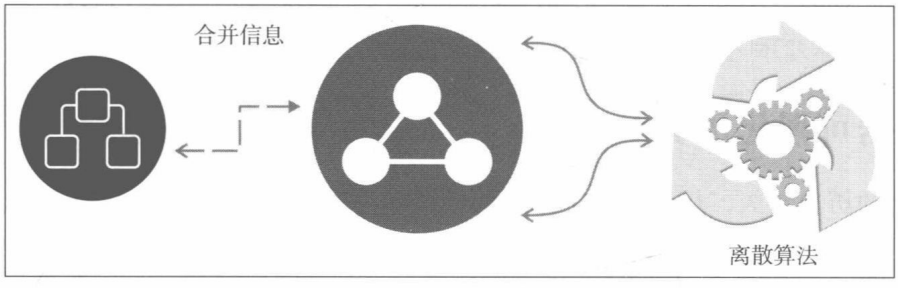
信息是可量化的、可测度的、有形式的，可以被访问、生成、存储、分发、搜索、压缩和复制。信息可以通过数量或信息量进行量化。

通过应用离散算法，信息可转换为知识，知识要比信息更可量化。在某些领域，知识持续经历了一个不断发展的周期。当数据发生实时变化时，这种演变过程随之发生。

知识就像是帮助你做面包的面粉和酵母成分的烹饪配方。另一个看待知识的方法是数据和信息的结合，并加入经验和专家意见，以帮助决策。知识不仅仅是过滤或算法的结果。

转换中包括哪些步骤？这种变化如何发生？当然，它本身是不能发生的。尽管信息这个词是基于定义的不同阐释，但是，我们将在计算的范围内进一步探索。

有一个简单的类比用以说明信息和知识之间的区别：一门特定课程的课程材料为你提供有关概念的重要信息，随后老师引导学生通过讨论来理解概念。这有助于学生获得课程知识。类似地，信息转换为知识也需要完成一些工作。下图展示了信息转换为知识的过程：



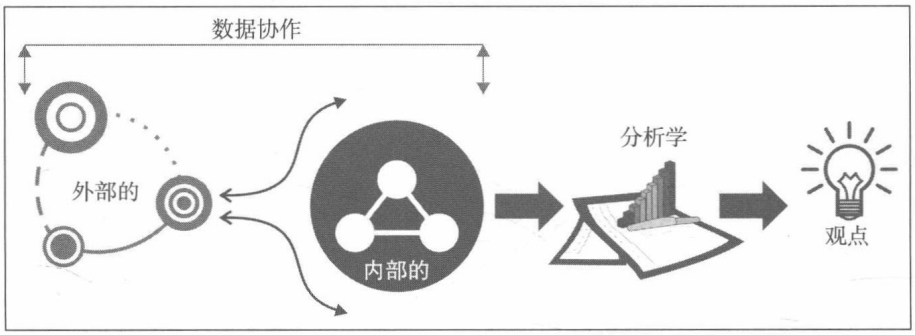
正如上图所示，信息通过一些离散算法进行合并和运行后，就能转换为知识。需要通过整合信息得到更多的知识。通过这种转换获得的知识有助于回答有关数据或信息的问题，比如，公司在哪个季度销售收益最高？广告拉动销售的贡献有多大？今年发布了多少新产品？

1.2.3　知识转换为观点

在传统的系统中，信息经处理、分析并形成报告。自因特网诞生以来，我们可以获取经过处理的信息，而且社交媒体融合成为一种处理实际问题的新方式。

一些组织机构已开始分析外部数据来获得观点。比如，通过 Twitter 上消费者的推文完成对用户情绪的测度，以此来追踪他们对产品品牌的意见。在某些情况下，较高比例的用户会在社交媒体上发布新产品的好评，比如一台 iPhone 或平板电脑。分析工具能够提供该情绪的数据化证据，这就是数据可视化扮演的重要角色。

下面是知识转化为观点的另一个例子。2009 年 Netflix 公司宣布了一场比赛，该比赛基于已有的电影分级，评选用来预测用户对电影评级的最佳协同过滤算法。比赛的获胜者用语用学理论，在预测用户分级方面提高 10.05% 的正确率，增加了 Netflix 公司的商业价值。



知识转换为观点是通过如上图所示的协作和分析来实现的。观点意味着看到解决方案，并发现需要做的事情。得到数据和信息很容易，一些组织机构已经知道获取方法，但是得到观点却很难。观点的得出需要新的创造性思维和连点成线的能力。除了应用创造性思维，数据分析和数据可视化在观点得出的过程中也发挥着很大作用。数据可视化被视为艺术和科学的结合。

1.3 数据可视化历史

可视化的历史悠久，最早用墙上的原始绘图和图像，表中的数字以及黏土上的图像来呈现信息。然而，它们并没有被称为可视化或数据的可视化。数据可视化是一个新术语；它传达出可视化不仅仅是以图表的形式展示数据。数据背后的信息应该用效果良好的图表直观揭示出来；图表本身应该帮助读者看到数据结构。

计算机出现前的可视化

在巴比伦时代早期，图片被绘制在黏土上，随后被渲染在纸草上。那些图的目标是给人们提供对信息的定性理解。众所周知，作为一种信息的可视化展示，我们对图片的理解是一种本能，因此理解过程非常轻松。本节只包括可视化历史的部分细节。关于精心设计的细节和例子，我们推荐两个有趣的资源：