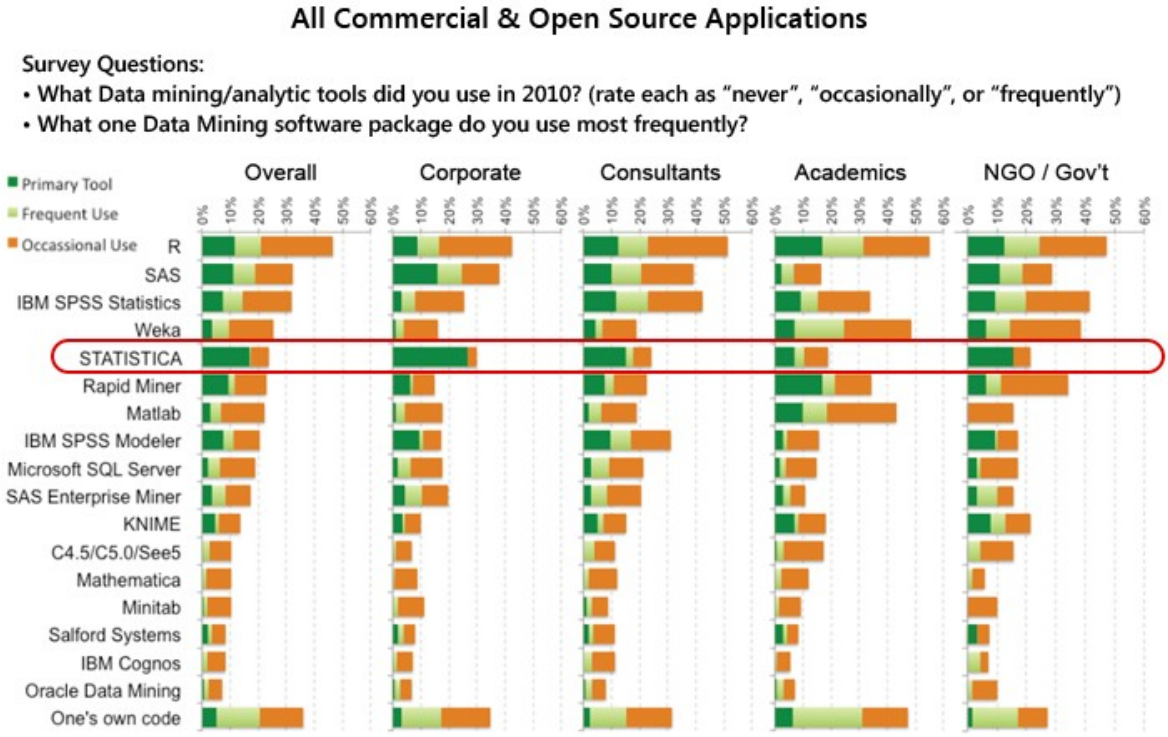
大数据环境下的 R 数据挖掘

近年来，由于存储设备的单位成本以惊人的速度下降（1G 硬盘空间的成本现在只需要几美分，这在过去难以想象），我们可以轻而易举地积累起大量的数据。电信 运营商，可以记录用户通话、短消息、无线上网产生的每一条信令，省级运营商一小时写入存储设备的数据量可以达到几百 G。电子商务网站，可以记录用户的每一 次交易，甚至每一次点击，可以复原用户的完整访问路径找出用户的兴趣点。城市监控体系，在各个重要路口，高速公路上的摄像头，每秒钟都在产生海量的视频数 据。在生命科学领域，对人体的 DNA 分析，一个个体就能产生几个 G 数据，可以想象如果一个生物信息数据库里包含了成千万的个体数据，信息量将会是怎样一个 规模，如此等等，不胜枚举。我们毫无疑问，正处于一个信息爆炸的时代。

很不幸的是，我们得到了大量的数据，而这些数据中的绝大部分，在它的生命周期里基本上都被闲置着，从来没有考虑过产生任何的价值，唯一的用途就是“保存备 查”。尽管“啤酒与尿布”的故事，已经写入教科书有 10 多年了，几乎每一个接受过专业教育的同仁都知道数据挖掘能产生的价值，但是直到今天，我们对数据的 处理依然停留在按预定指标进行统计这种很低的水平上。造成这种情况的原因有很多。一方面，由于业务人员和 IT 人员的工作鸿沟，使到即使能提出数据分析的需 求都成了一个很大的困难。在各公司里保管数据的大多是 IT 人员，他们对业务的了解可能并非很深入，而业务人员也鲜有对数据有深入认识者，他们通常都缺乏必 要的数学素质和知识基础去进行建模和深入的分析工作。另一方面，数据分析专家具有深厚的数学处理能力，善于建模和构筑算法，但是由于无法得到合适的需求， 他们的能力也无从施展。另外数学家、统计学家们很多并不熟悉现代的 IT 软硬件设备的特性，对于集群、分布式系统、大规模存储、云计算、数据库等认识几乎为 零，对于算法的实现可能还停留在对着 PC 写 C 语言程序的水平上，对于海量数据，无法利用现代化设备的能力，使到算法是否能真正实现变成生产力存有很大的疑 问。

现在这门《数据分析系列网络课程》正是要打破这种鸿沟。用新兴的互联网教育模式，把各应用领域的业务专家、数据分析专家、IT 专家推荐给学习者，向有志于 学习数据分析知识发挥数据价值的朋友能得到低成本交流的机会。我们的目标是在中国传播“技术成就梦想，数据产生价值”的观念，使学习者能快速提升其个人能 力，在新的挑战面前获取更多个人机会，企业能在保存的海量数据中炼出黄金。

R 是一套完整的数据处理、计算和制图软件系统。是一个免费的自由软件，它有 UNIX、LINUX、MacOS 和 WINDOWS 版本，都是可以免费下载和使 用的,在那儿可以下载到 R 的安装程序、各种外挂程序和文档。在 R 的安装程序中只包含了 8 个基础模块，其他外在模块可以通过 CRAN 获得。R 既是功能强大的 统计和分析软件，同时也是完美的数据可视化制作工具，丰富的图形函数和外置包，几乎无限的扩展能力，使到我们的想象空间永远都不会达到上限



2011 年统计的数据分析软件使用率情况，R 语言多项雄踞首位，是最热门的分析利器

《R 语言数据分析、展现与实例》课程介绍如下：

1 基础数据分析知识，包括一些概率统计里的概念、术语，和基本统计量的计算方法等。
2 一些常用的数据分析和数据挖掘算法，以及有关的各种领域里的实际应用案例分析
3 世界最流行的开源数据分析软件 R 及其编程方法
4 数据展现，介绍 R 及其强大的图表功能

课程大纲：

第 1 课 R 语言基础
- R 简介
- 数据类型介绍
- R 的数据可视化
- 常用 R 包介绍
- R 集成开发环境

第 2 课 数据整理
- 数据的读入输出
- 控制流
- 各种图表
- 常用统计量计算

第 3 课 数据展现 1
- 基本制图函数综述
- 理解关键制图参数

第 4 课 数据展现 2
散点图
线图与时间序列谱图
案例：股价走势可视化展现

第 5 课 数据展现 3
柱形图
点图
饼图
直方图
案例：销售数据可视化展现

第 6 课 数据展现 4
箱线图
热力图
等高线
地图
案例：Facebook 好友联系图

第 7 课 预知未来的回归模型 1
线性回归模型
案例：网页流量预测

第 8 课 预知未来的回归模型 2
logistic 回归
广义线性回归
非线性回归
案例：婚外情频率预测

第 9 课 预知未来的回归模型 3
回归检验与方差分析
案例：上两周周案例的进一步分析优化

第 10 课 挖掘关联和推荐技术
MINE 方法
apriori 购物篮分析
案例：超市购物篮分析

第 11 课 万事皆选择 1
分类算法（线性判别法，贝叶斯分类器，决策树，最近邻算法）
案例：汽车销量
走势预测，上涨还是下跌？

第 12 课 万事皆选择 2
聚类算法（层次聚类法，谱系图，k 平均值法，k 中心法）
案例：推荐系统

第 13 课 大道至简
降维技术
主成分分析和因子分析
案例：业绩综合指标设计

第 14 课 沿着时间轴前进
时间序列分析
案例：未来股价预测

第 15 课 R 数据挖掘实际场景综合案例分析及前沿技术选讲