数据科学简介

Introduction of Data Science

Mr. Black

目录

- 数据科学简介
- 数据科学工具箱
- 数据科学分工与流程

数据科学简介

数据科学 Data Science

1974年，Peter Naur出版了**“计算方法的简介调查”**[1]一书。该书中**“数据科学”（Data Science）**一词被大量使用，同时对其作出定义：“数据科学是一门专门处理数据的科学。它被授权处置与其他科学领域中有关数据的表现与关联”。定义中强调了数据同其他科学领域之间存在的关系。

1997年，Jeff Wu在**“统计=数据科学?”**[2]一文中重新探索了“统计（Statistics）”一词的含义，他认为统计工作应该是由数据收集，数据建模和分析以及决策制定三部分组成。同时他倡导将“统计”一词重命名为“数据科学”，将“统计学家（Statisticians）”一词重命名为**“数据科学家（Data Scientists）”**。

[1] P. Naur, Concise Survey of Computer Methods. Petrocelli Books, 1974.

[2] C. J. Wu, “Statistics = data science?,” 1997.

数据科学 Data Science

2001年，William S. Cleveland发表**“数据科学：为扩大统计技术领域的行动计划”**[1]。文章计划扩大统计领域相关的技术工作范围，正是由于范围的扩张，作者将这一改变的领域称之为“数据科学”。计划中划分了6大技术范围，其具体内容和占比如下：

1. **（25%）多学科调查**：包括在相关主题领域内的数据分析协作。
2. **（20%）处理数据的模型和方法**：包括统计模型；建模方法；等。
3. **（15%）数据计算**：包括硬件系统；软件系统；计算算法。
4. **（15%）教学方法**：包括小学，中学，大学，研究生，继续教育和企业培训的教学课程规划。
5. **（5%）工具评估**：包括实践中工具使用情况的调查，新工具需求的调查以及开发新工具的过程研究。
6. **（20%）理论**：包括数据科学的基础；模型方法，数据计算，教学和工具评估的基本方法；模型方法，数据计算，教学和评估的数学调查。

[1] W. S. Cleveland, “Data science: an action plan for expanding the technical areas of the feld of statistics,”International statistical review, vol. 69, no. 1, pp. 21–26, 2001.

数据科学 Data Science

2002年，国际科学理事会的科技数据委员会（CODATA）创立**Data Science Journal**杂志。2003年，**Journal of Data Science**创立。杂志为所有的数据工作者提供了一个很好的交流平台。

2005年，美国国家科学委员会发布了**“长期数字数据收集促成二十一世纪的研究与教育”**[1]。报告中将数据科学家（Data scientists）定义为信息和计算机科学家，数据库和软件工程师，程序员等那些对于成功管理信息数据至关重要的人们。

2012年，Tom Davenport和D.J. Patil在哈佛商业评论中发表**“数据科学家：21世纪最性感工作”**[2]。文章中将数据科学家评为21世纪最性感的职业。

[1] N. S. Board, “Long-lived digital data collections enabling research and education in the 21st century.”http://www.nsf.gov/pubs/2005/nsb0540/, 2005.

[2] T. H. Davenport and D. Patil, “Data scientist: Te sexiest job of the 21st century,” Harvard Business Review Magazine: http://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-Century/ar/1, 2012.

数据产品 Data Products

Patil在**“数据的柔术：将数据转化为产品的艺术”**[1]一文中解释说“**数据产品**是通过使用数据促进最终目标的产品”。因此可以说数据产品并不仅仅是指数据分析（Data Analysis）, 向高管提供的建议或是导致业务流程改善的洞察，而应该是一套完整有形的问题解决系统。

为了方便大家清楚理解数据产品的概念，我们比较两款产品：Excel和PYMK。Excel大家应该比较熟悉，是微软Office套件中用于数据处理、统计分析和辅助决策的表格处理软件。PYMK相对比较陌生，PYMK全称为People You May Know，是LinkedIn一套人物关系预测系统。

[1] D. Patil, “Data jujitsu: the art of turning data into product,” tech. rep., O’Reilly Media, Inc., 2012.

数据产品 Data Products

Excel和PYMK特性对比

特性	Excel	PYMK
系统	否（通用分析软件）	是（预测系统）
数据源	用户指定，无具体形式和内容要求	人员年龄，性别，工作等个人信息
数据理解	视用户操作而定	对数据有较充分理解
算法应用	视用户操作而定	使用相关智能算法
目标	无具体目标	寻找出可能认识的人
结果	不同操作产生不同结果	可能认识的人或人物关系网

数据产品 Data Products

在**“什么是数据科学?”**[1]一文中，Mike Loukides的第一句话就指出了**“未来是属于那些能将数据转化成产品的人和公司的”**，也就是说数据的真正价值只有在进行深度加工处理并形成产品之后才能够被体现出来。可以说有价值的数据是一个有待开发的金矿，需要人们利用“数据产品”这把利器去开采才能够得到金灿灿的黄金。同时，文章也指出了数据科学和数据产品之间的关系：数据科学使数据产品的创造成为可能，也就是数据科学在数据产品的创造开发过程中扮演着至关重要的角色。

[1] M. Loukides, “What is data science?,” tech. rep., O’Reilly Media, Inc., 2010.

跨界 Crossover

跨界（Crossover）一词在不同的领域有着各自具体的含义。**跨界音乐（Crossover Music）**[1]是指一个音乐作品被诠释成两种或更多的品味或流派。**跨界营销（Crossover Marketing）**[2]意味着打破传统的营销思维模式，实现多个品牌从不同角度诠释同一个用户特征，发挥不同类别品牌的协同效应。因此，跨界可以称得上是多种资源的一种融合创新。

开发数据产品同样也是一场跨界知识的融合。无论是组建一个数据产品开发团队还是成长为一个真正的数据科学家，都要对所涉及到的各种知识及其技能有所涉猎。当然“全”也并不意味着不“专”，正如开发数据产品的核心是数据科学的应用一样，数据科学家应掌握扎实的数据科学理论和应用能力。

[1] Wikipedia, “Crossover (music).” http://en.wikipedia.org/wiki/Crossover_music.

[2] 邓勇兵, “跨界营销: 体验的综合诠释,” 中国市场, 2007.

跨界 Crossover

知识类型	知识名称	和“开发数据产品”的关系	重要程度
领域知识	行业知识（管理，金融）	业务理解	★★☆☆☆
数学统计	基础数学（微积分，代数）	数据科学（基础）	★★★☆☆
数学统计	统计学	数据科学（统计分析）	★★★★☆
数学统计	应用数学（机器学习）	数据科学（建模分析）	★★★★☆
数学统计	统计编程（R，Python）	数据科学（模型计算）	★★★★☆
工程知识	数据库知识（MySQL，HIVE）	数据源获取	★★★☆☆
工程知识	软件工程（系统设计，Java）	系统开发（基础）	★★★☆☆
工程知识	计算框架（Hadoop，Spark）	系统开发（框架选择）	★★★☆☆
工程知识	前端技术（配色，HTML）	数据可视化	★★★☆☆

数据科学工具箱

数据科学常用工具

在数据科学领域，我们会用到多种多样的编程语言和工具。而编程语言和工具的选择取决于多种因素，例如：项目需要（目标，预算，时间等）；项目负责人和成员的专业背景和偏好，工具成本，功能性，可用性，学习曲线等等。

一般而言，这些编程语言和工具可以划分为如下5类：

1. 统计编程语言：SPSS，SAS，R，Python。
2. 数据挖掘和机器学习工具箱：Weka（Java），scikit-learn（Python）。
3. 传统编程语言：C/C++，Java，Scala。
4. 分析平台和框架：RapidMiner，KNIME，Hadoop，Spark，Hive。
5. 其他：SQL，Excel，Tableau。

KDnuggets每年都会进行一项关于“What Analytics, Big Data, Data Mining, Data Science software you used in the past 12 months for a real project?”（过去12个月中你在真实项目中所使用的数据分析，大数据，数据挖掘和数据科学软件是什么？）。在2016年，该项调查共有2895个人参与，最终得票最高的10个编程语言和工具分别为：R，Python，SQL，Excel，RapidMiner，Hadoop，Spark，Tableau，KNIME和scikit-learn。

数据科学常用工具

编程语言和工具	2016年得票	2016年排名	2014占比	2015占比	2016占比
R	1419	1	38.5%	46.9%	49%
Python	1325	2	19.5%	30.3%	45.8%
SQL	1029	3	25.3%	30.9%	35.5%
Excel	972	4	25.8%	22.9%	33.6%
RapidMiner	944	5	44.2%	31.5%	32.6%
Hadoop	641	6	12.7%	18.4%	22.1%
Spark	624	7	2.6%	11.3%	21.6%
Tableau	536	8	9.1%	12.4%	18.5%
KNIME	521	9	15%	20%	18%
scikit-learn	497	10	na	8.3%	17.2%

数据科学之战：R与Python

发展历史

R语言[1]是一套用于统计编程和绘图的自由软件编程语言与操作环境。R语言是S语言的一种延伸和实现，由Ross Ihaka和Robert Gentleman于1995年设计开发的一种开源语言，因此称之为R语言。作为S语言的一种延伸，R语言主要利用C语言，Fortran和R语言开发完成

Python是由Guido Van Rossem于1991年创建的一门强调效率和代码可读性的编程语言。Python由Python软件基金会（PSF）负责其发展，其开发灵感主要来自于C语言和Modula-3，部分来自于ABC语言。Python的名字取自喜剧蒙提·派森的飞行马戏团（Monty Python's Flying Circus）。

[1] R. Project, “What is r?.” http://www.r-project.org/about.html.

数据科学之战：R与Python

学习和使用

R语言可以使用简短的几行代码完成一个统计模型。R语言也有其自己的代码样式表，但很少有人使用，不过保持一个良好的代码风格是一个还好的习惯。R语言可以使用不同点方式实现相同的功能，例如显式的循环（for）和隐式的循环（apply方法）等。在R语言中，可以还轻松的实现复杂的公式，同时一些常用的统计模型也是现成的方便使用。由于R语言的特点，开始学习时将会面临一个陡峭的学习曲线，不过一旦入门后就可以很容易的使用其高级特性。

Python是一个灵活的编程语言，由于其注重简便性和代码的易读性，Python的学习曲线相对平缓，可以很好的用于编写一些简短代码。不过由于Python缩进式的代码风格，对于类C语言的使用者多少会影响其学习和使用。由于Python是一门更加通用的编程语言，其更多的优势在于编写网站和其他应用脚本。由于Python看重可读性和易用性，使得它的学习曲线相对比较低并且平缓。除了可以用于数据分析外，还可以帮助使用者快速高效的完成其他工作。

数据科学之战：R与Python

代码库和社区支持

R语言有一个庞大的扩展包库（CRAN，The Comprehensive R Archive Network），用户可自行贡献开源的扩展包供其他人员使用。R语言提供最早的发布版本为0.49（1997年4月23日），当时CRAN仅有3个镜像站点，仅提供12个包，仅编译了少量类Unix平台版本，Windows和Mac OS版本在该版尚未提供。截止到2016年8月，CRAN已有101个镜像站点，提供多达6631个包。CRAN库中的扩展包包含了大量的R工具和数据集，这使得我们可以快速的获取和使用最新的技术和功能而不必一切都从头开发。由于R起源和发展的特点，研究人员、数据科学家、统计学家和数量分析专家对R提供了更多的支持。使用者可以从邮件列表(Mailing Lists)、用户贡献的文档、以及Stackoverflow网站等地方获取大量的社区支持。

Python也提供一个代码库（PyPi，Python Package Index），用户可以贡献自己的代码，不过相比CRAN而言，实践起来相对困难一些。不过单纯从统计分析角度而言，正如计算机科学教授Norm Matloff所言：Python并未建立起一个能与CRAN媲美的巨大的代码库，R在这方面领先巨大。

为什么选择R语言

我想如下问题可以帮助你进行选择：

1. 你要解决的问题是什么？
2. 学习一门新语言的成本是多少？
3. 在你的领域，常用的工具有哪些？
4. 其他常用的工具又有哪些？他们和常用的工具又有什么关系？

R：

- 优点：一图胜千言，生态系统，统计的通用语言
- 缺点：运行慢？陡峭的学习曲线

Python：

- 优点：IPython Notebook，通用语言，多用途语言
- 缺点：可视化，不成熟

数据科学分工与流程

数据科学分工

根据Donoho在**“数据科学50年”**[1]一文中的观点，将数据科学分为了6个部分，分别是：

1. 数据探索和准备（Data Exploration and Preparation）
2. 数据表示和转换（Data Representation and Transformation）
3. 数据加工计算（Computing with Data）
4. 数据建模（Data Modeling）
5. 数据可视化和展现（Data Visualization and Presentation）
6. 数据科学的科学性（Science about Data Science）

[1] D. Donoho, “50 years of data science,” 2015.

数据分析和挖掘流程

数据分析和挖掘是一个复杂的过程，在进行数据分析和挖掘工作的过程中，我们需要一个过程模型指导每一步工作。迄今为止，很多专家和学者提出了多种数据分析和挖掘的过程模型，下表显示了工业界数据分析和挖掘工作者近年来所采用的方法。

工业界数据分析和工作者采用的方法

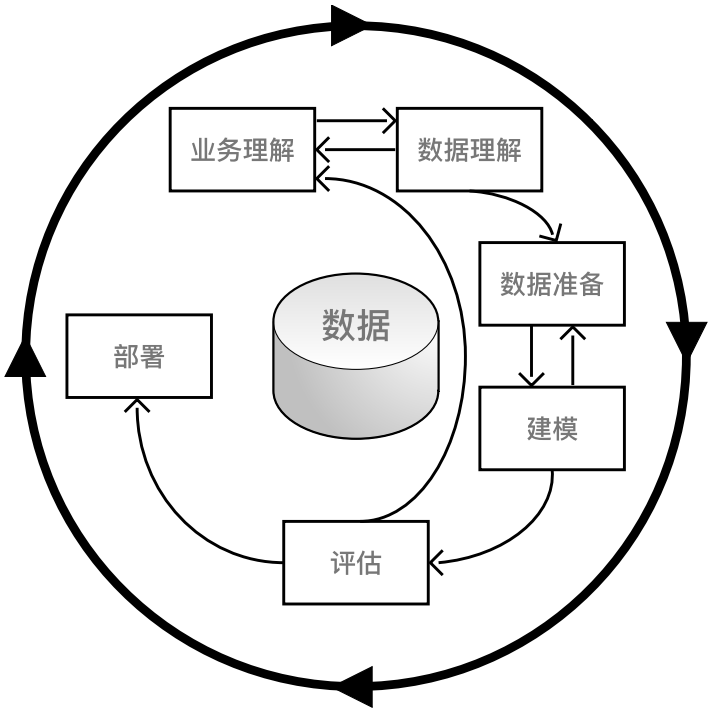
年份/方法	CRISP-DM	SEMMA	KDD Process	Other	None
2002[1]	51%	12%	-	34%	4%
2004[2]	42%	10%	-	40%	7%
2007[3]	41%	13%	7%	33%	5%

[1] http://www.kdnuggets.com/polls/2002/methodology.htm
[2] http://www.kdnuggets.com/polls/2004/data_mining_methodology.htm
[3] http://www.kdnuggets.com/polls/2007/data_mining_methodology.htm

数据分析和挖掘流程

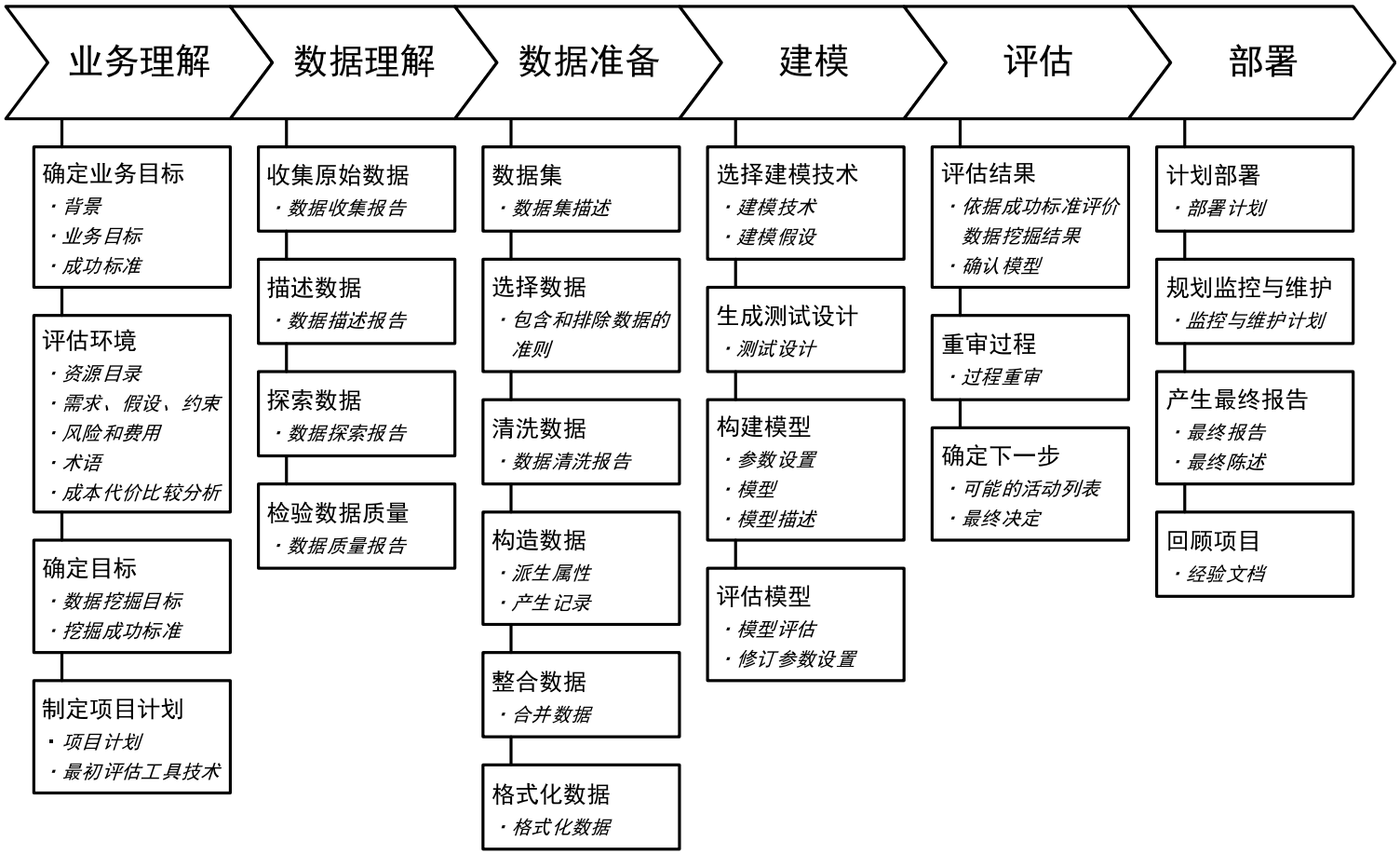
CRISP-DM[1]全称为跨行业数据挖掘标准流程（Cross Industry Standard Process for Data Mining）Shearer于2000年提出。CRISP-DM对一个数据分析和挖掘项目的生命周期提供一个总体的描述。

- 业务理解 (Business understanding)
- 数据理解 (Data understanding)
- 数据准备 (Data preparation)
- 建模 (Modeling)
- 评估 (Evaluation)
- 部署 (Deployment)



[1] C. Shearer, “Te crisp-dm model: the new blueprint for data mining,” Journal of data warehousing, vol. 5, no. 4, pp. 13–22, 2000

CRISP-DM


业务理解
确定业务目标
・背景
・业务目标
・成功标准
评估环境
・资源目录
・需求、假设、约束
・风险和费用
・术语
・成本代价比较分析
确定目标
・数据挖掘目标
・挖掘成功标准
制定项目计划
・项目计划
・最初评估工具技术
数据理解
收集原始数据
・数据收集报告
描述数据
・数据描述报告
探索数据
・数据探索报告
检验数据质量
・数据质量报告
数据准备
数据集
・数据集描述
选择数据
・包含和排除数据的准则
清洗数据
・数据清洗报告
构造数据
・派生属性
・产生记录
整合数据
・合并数据
格式化数据
・格式化数据
建模
选择建模技术
・建模技术
・建模假设
生成测试设计
・测试设计
构建模型
・参数设置
・模型
・模型描述
评估模型
・模型评估
・修订参数设置
评估
评估结果
・依据成功标准评价数据挖掘结果
・确认模型
重审过程
・过程重审
确定下一步
・可能的活动列表
・最终决定
部署
计划部署
・部署计划
规划监控与维护
・监控与维护计划
产生最终报告
・最终报告
・最终陈述
回顾项目
・经验文档

Thanks