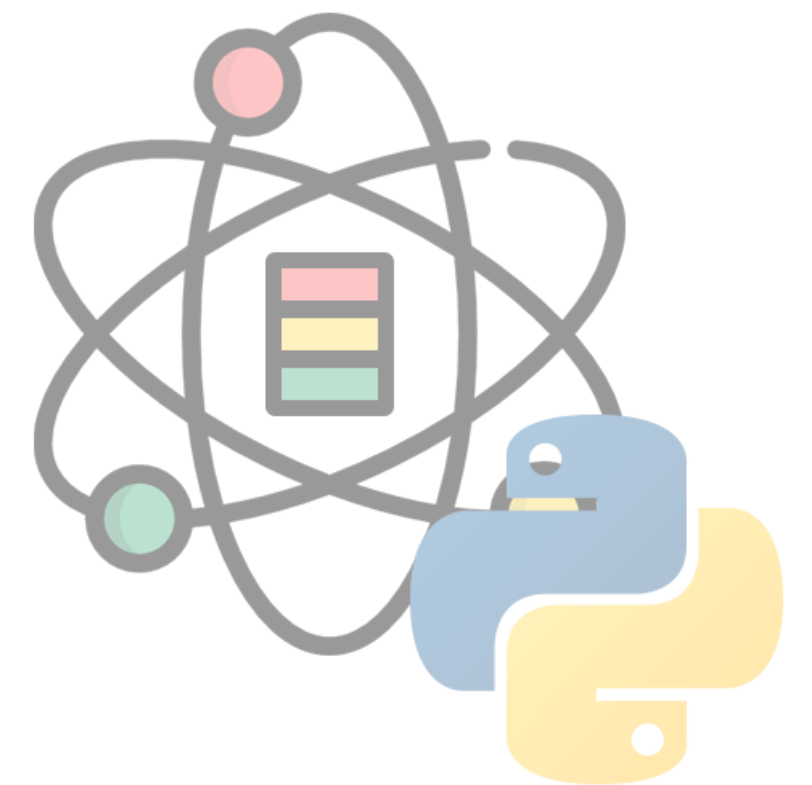


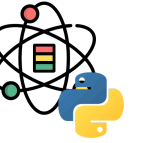
数据科学简介

Introduction of Data Science

范叶亮 Leo Van



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

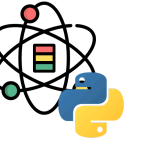


1974 年，Peter Naur 出版了“计算方法的简介调查”¹ 一书。该书中“数据科学” (**Data Science**) 一词被大量使用，同时对其作出定义：“数据科学是一门专门处理数据的科学。它被授权处置与其他科学领域中有关数据的表现与关联”。定义中强调了数据同其他科学领域之间存在的关系。

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.

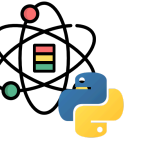


2001 年，William S. Cleveland 发表“**数据科学：为扩大统计技术领域的行动计划**”¹。文章计划扩大统计领域相关的技术工作范围，正是由于范围的扩张，作者将这一改变的领域称之为“数据科学”。计划中划分了 6 大技术范围，其具体内容和占比如下：

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. Cleveland, W. S. (2001). Data science: an action plan for expanding the technical areas of the field of statistics. *International statistical review*, 69(1), 21-26.

数据科学



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。

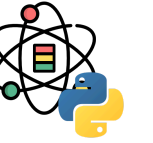
2005 年，美国国家科学委员会发布了“**长期数字数据收集促成二十一世纪的研究与教育**”¹。报告中将数据科学家 (Data scientists) 定义为信息和计算机科学家，数据库和软件工程师，程序员等那些对于成功管理信息数据至关重要的人们。

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: The Sexiest Job of the 21st Century,” Harvard Business Review Magazine: <https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century>, 2012.

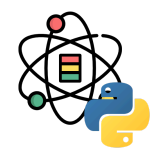
数据产品



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。

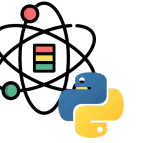
1. D. Patil, “Data jujitsu: the art of turning data into product,” tech. rep., *O’Reilly Media, Inc.*, 2012.



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

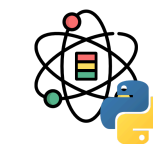
数据产品



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



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.

数据科学工具箱

数据科学常用工具

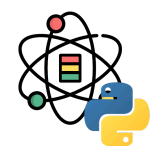


在数据科学领域，我们会用到多种多样的编程语言和工具。而编程语言和工具的选择取决于多种因素，例如：项目需要(目标，预算，时间等)；项目负责人和成员的专业背景和偏好，工具成本，功能性，可用性，学习曲线等等。

一般而言，这些编程语言和工具可以划分为如下 5 类：

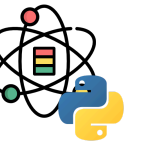
1. 统计编程语言：Python, R, SPSS, SAS
2. 数据挖掘和机器学习工具箱：scikit-learn (Python), mlr3 (R), Weka (Java)
3. 传统编程语言：C/C++, Java, Scala
4. 分析平台和框架：RapidMiner, KNIME, Hadoop, Spark, Hive
5. 其他：SQL, Excel, Tableau

KDnuggets 每年都会进行一项关于机器学习相关编程语言和工具使用的调研，在 2019 年，该项调查共有 1,800 个人参与，最终得票最高的 10 个编程语言和工具分别为：Python, RapidMiner, R, Excel, Anaconda, SQL, Tensorflow, Keras, scikit-learn, Tableau 和 Apache Spark。



编程语言和工具	2019 占有率	2018 占有率	2017 占有率	2016 占有率
Python	65.8%	65.6%	59.0%	45.8%
RapidMiner	51.2%	52.7%	31.9%	32.6%
R	46.6%	48.5%	56.6%	49%
Excel	34.8%	39.1%	31.5%	33.6%
Anaconda	33.9%	33.4%	24.3%	NA
SQL	32.8%	39.6%	39.2%	35.5%
Tensorflow	31.7%	29.9%	22.7%	6.8%
Keras	26.6%	22.2%	10.7%	NA
scikit-learn	25.5%	24.4%	21.9%	17.2%
Tableau	22.1%	21.5%	21.8%	18.5%

数据科学之战：Python 与 R



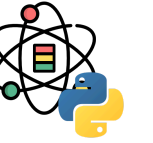
发展历史

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)。

R¹ 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。

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

数据科学之战：Python 与 R

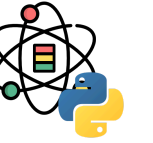


学习和使用

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。

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。

数据科学之战：Python 与 R

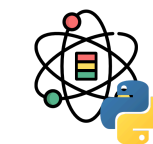


代码库

Python 提供一个代码库 **PyPi** (Python Package Index), 用户可以贡献自己的代码, 截止到 2019 年 10 月, PyPi 共有 200,539 个项目。除此之外, **Conda** 为不同操作系统提供了一个环境和包的管理平台, 除了能够管理 Python 以外, Conda 还能够管理 R, Ruby, Lua, Scala, Java, JavaScript, C/ C++, FORTRAN 等多种其他语言。

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 macOS 版本在该版尚未提供。截止到 2019 年 10 月, CRAN 已有 96 个镜像站点, 提供多达 15,121 个包。

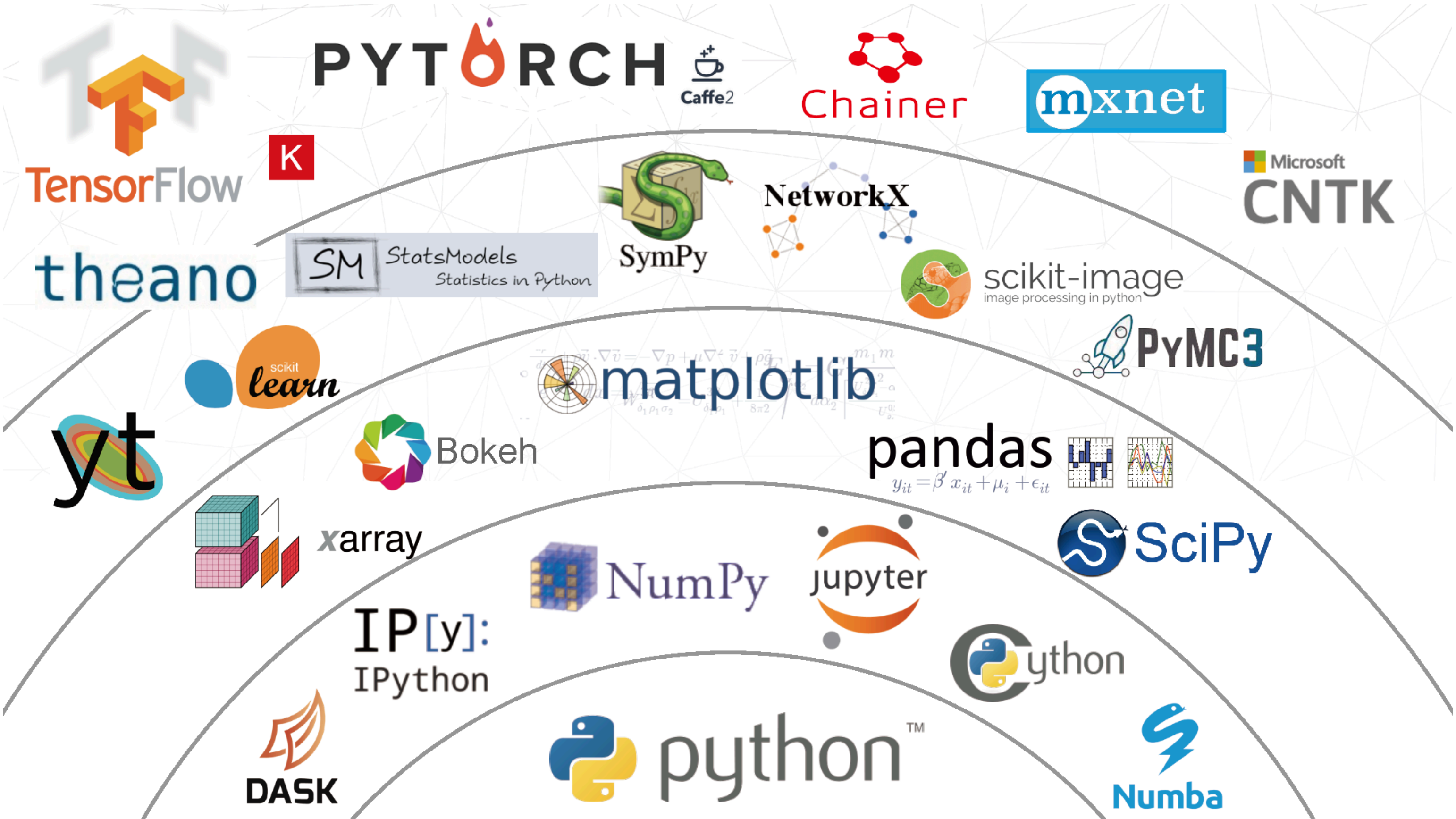
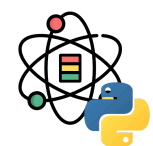
选择哪种语言



如何选择？

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？

Python 数据科学生态



数据科学分工与流程

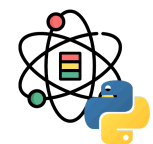
数据科学分工



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：

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. Donoho, D. (2017). 50 years of data science. *Journal of Computational and Graphical Statistics*, 26(4), 745-766.

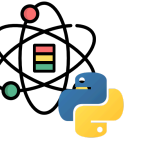


工业界数据分析和工作者采用的方法

年份/方法	CRISP-DM	My Own	SEMMA	KDD Process
2002 ¹	51%	23%	12%	NA
2004 ²	42%	28%	10%	NA
2007 ³	42%	19%	13%	7%
2014 ⁴	43%	27.5%	8.5%	7.5%

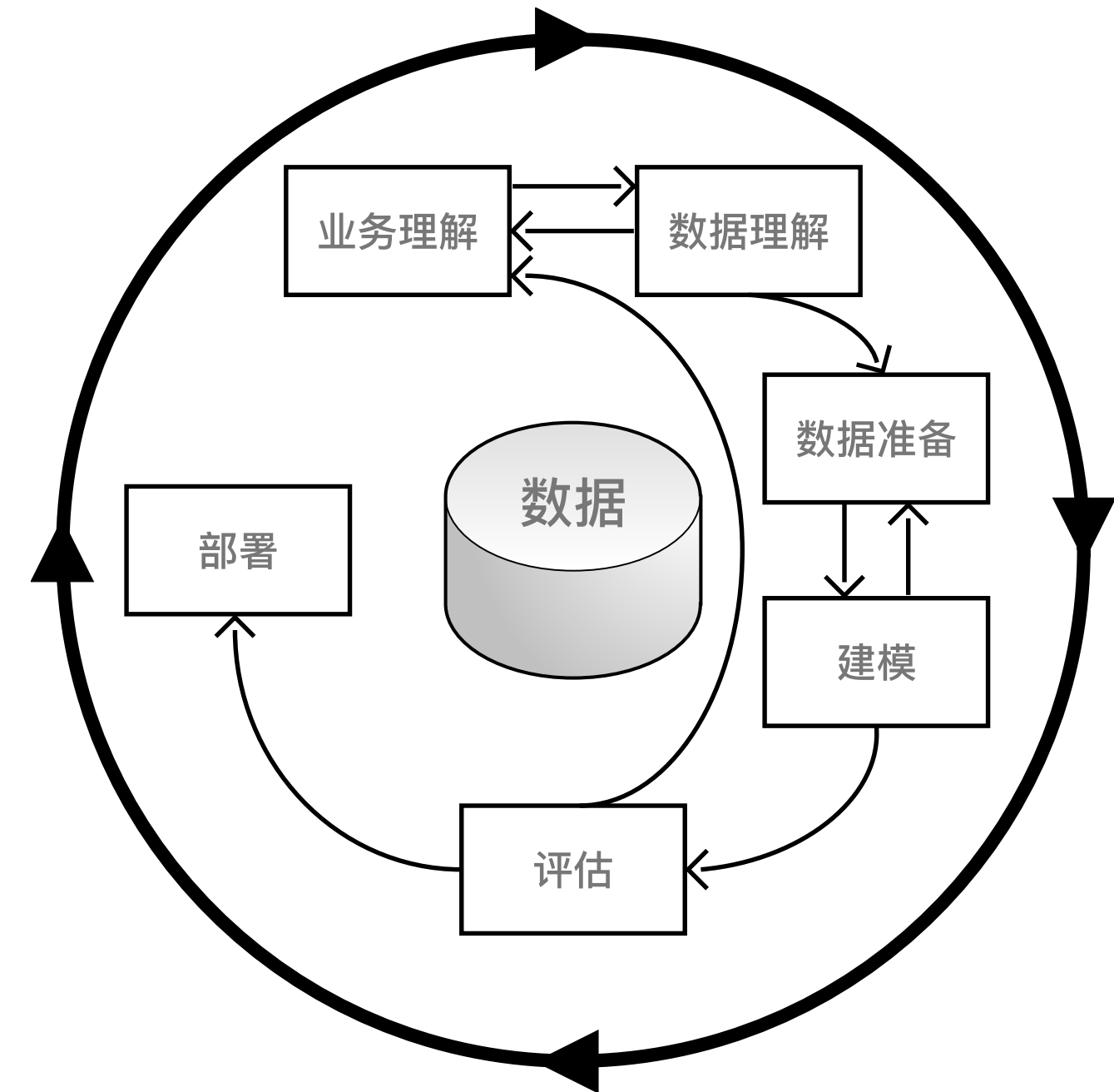
1. <http://www.kdnuggets.com/polls/2002/methodology.htm>
2. http://www.kdnuggets.com/polls/2004/data_mining_methodology.htm
3. http://www.kdnuggets.com/polls/2007/data_mining_methodology.htm
4. <https://www.kdnuggets.com/2014/10/crisp-dm-top-methodology-analytics-data-mining-data-science-projects.html>

数据分析和挖掘流程

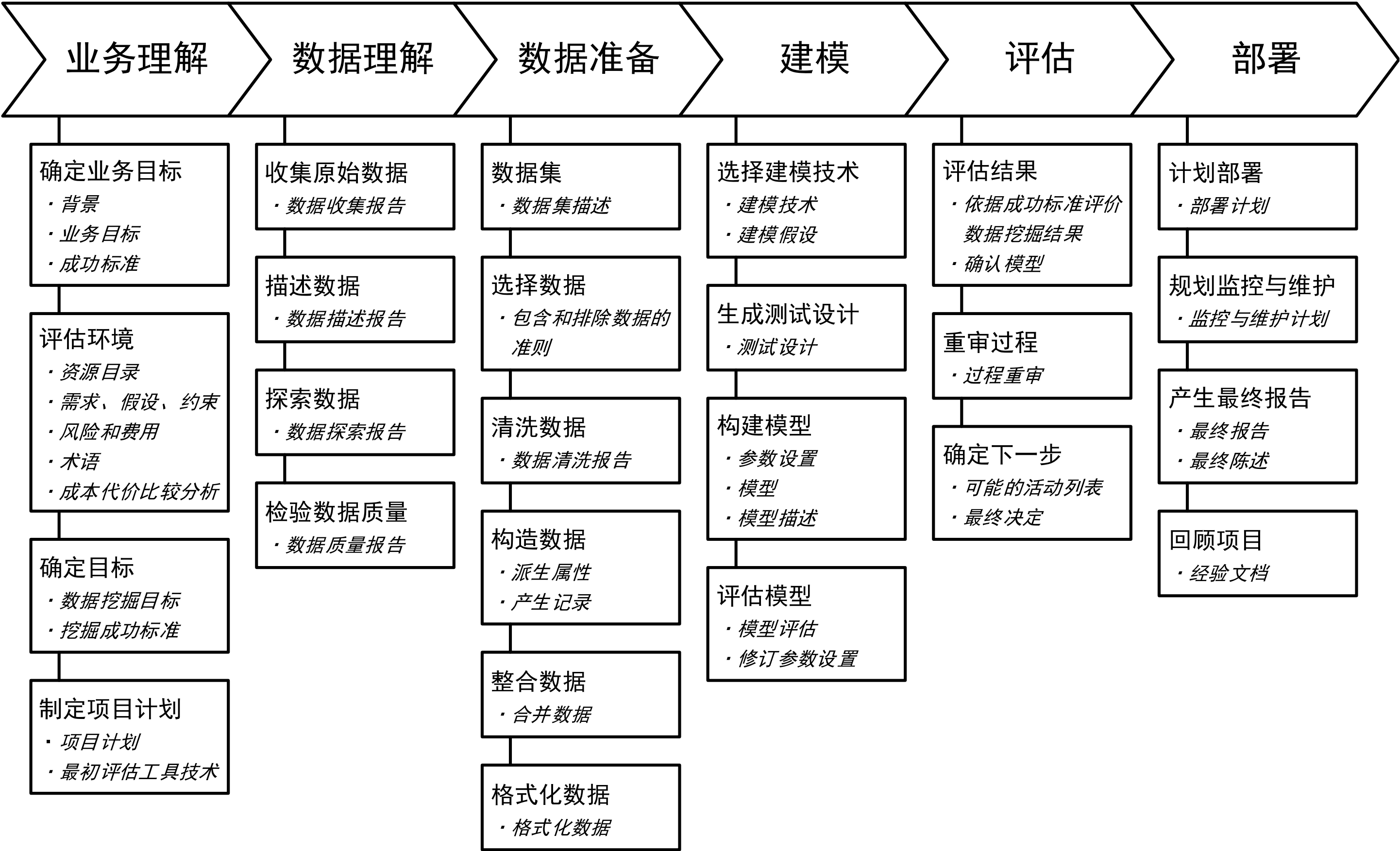
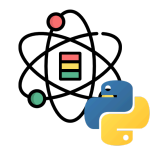


CRISP-DM¹ 全称为跨行业数据挖掘标准流程 (Cross Industry Standard Process for Data Mining) Shearer 于 2000 年提出。CRISP-DM 对一个数据分析和挖掘项目的生命周期提供一个总体的描述。

- 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. Shearer, C. (2000). The CRISP-DM model: the new blueprint for data mining. *Journal of data warehousing*, 5(4), 13-22.



感谢倾听

本作品采用  授权

版权所有 © 范叶亮 Leo Van