



北京邮电大学

BEIJING UNIVERSITY OF POSTS AND TELECOMMUNICATIONS



Data Analytics with R

语言数据分析



所谓学习，归类而已

艾新波 / 2018 • 北京



# 课程体系

## R语言数据分析

### 上部：论道



- 第1章 气象万千、数以等观
- 第2章 所谓学习、归类而已
- 第3章 格言联璧话学习
- 第4章 源于数学、归于工程

### 中部：执具



- 第5章 工欲善其事必先利其器
- 第6章 基础编程
- 第7章 数据对象

第8章 人人都爱tidyverse

第9章 最美不过数据框

### 下部 博术



- 第10章 观数以形
- 第11章 相随相伴、谓之关联
- 第12章 既是世间法、自当有分别
- 第13章 方以类聚、物以群分
- 第14章 庐山烟雨浙江潮

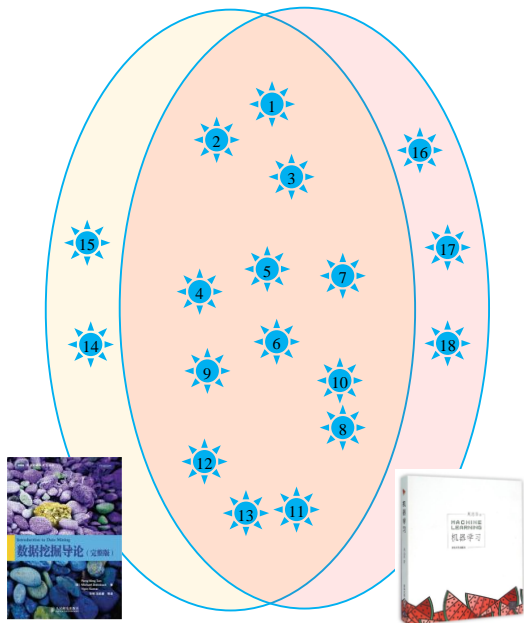
以机器学习为内核

数据分析  $\approx$  机器学习 / 数据挖掘

$\approx$  认识数据 + 关联 + 分类 + 聚类

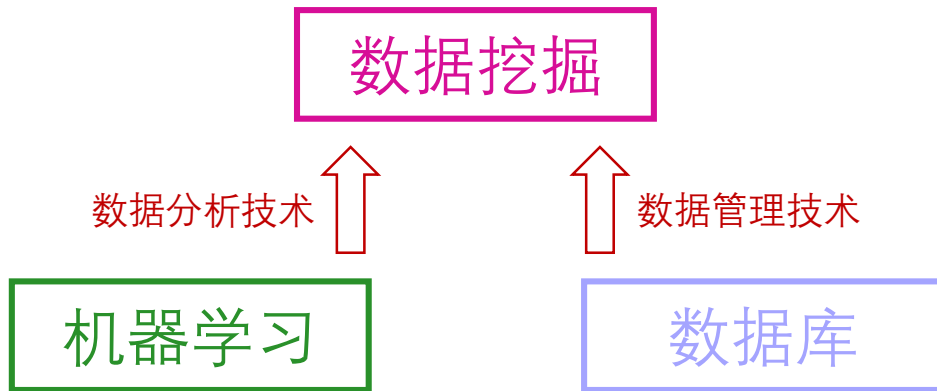
$\approx$  寻找关系结构（核心是归归类）

# 机器学习 vs 数据挖掘



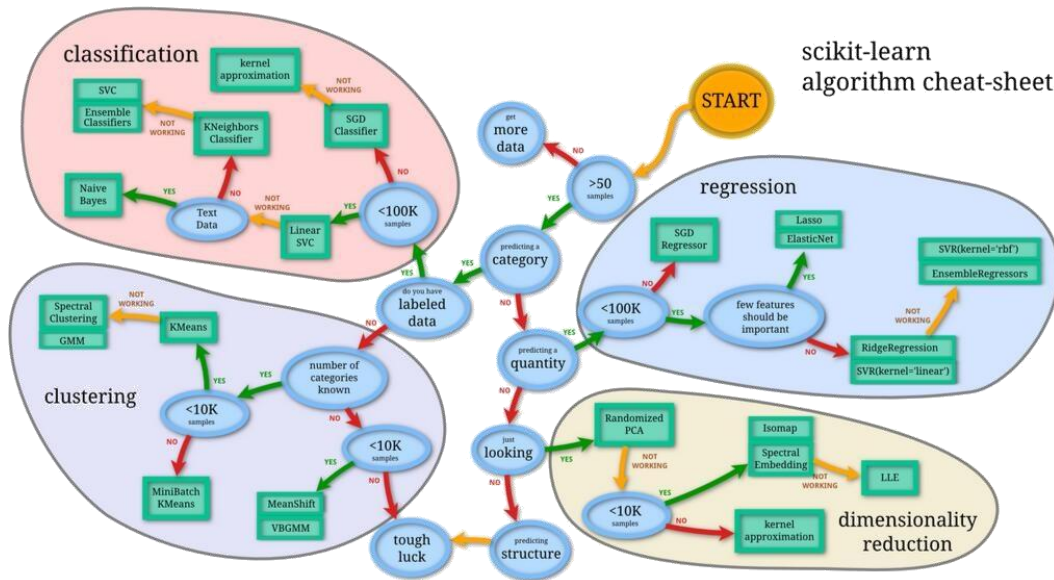
序号	数据挖掘导论	机器学习
相同	①决策树	
	②贝叶斯分类器	
	③人工神经网络	
	④支持向量机	
	⑤组合方法 (集成学习)	
	⑥关联规则 (规则学习)	
	⑦层次聚类	
	⑧原型聚类	
	⑨密度聚类	
	⑩维归约 (降维)	
	⑪线性代数 (矩阵)	
	⑫概率统计	
	⑬优化	
不同	⑭最近邻	⑯计算学习理论
	⑮异常检测	⑰半监督学习
		⑱强化学习

# 机器学习 vs 数据挖掘



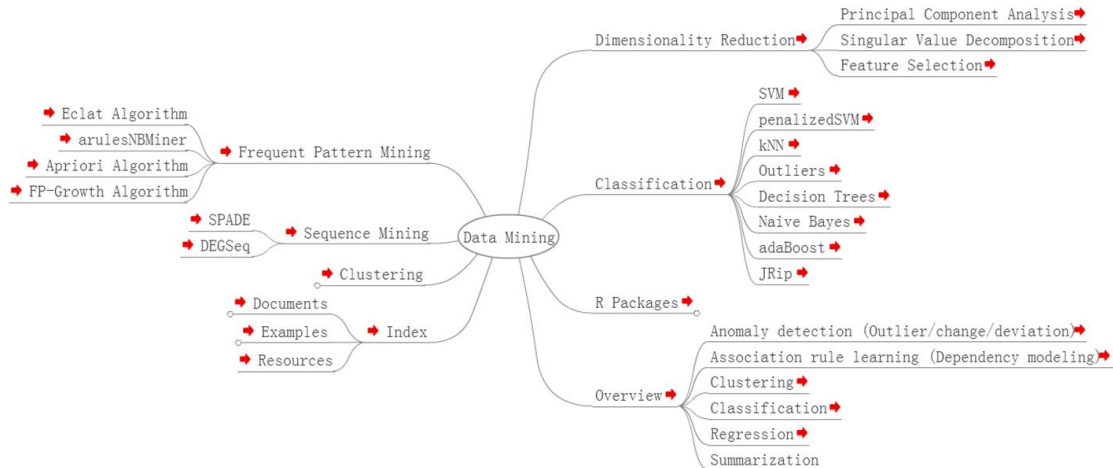
Source: 周志华·数据挖掘与机器学习

# 机器学习/数据挖掘脉络图



[http://scikit-learn.org/stable/tutorial/machine\\_learning\\_map/](http://scikit-learn.org/stable/tutorial/machine_learning_map/)

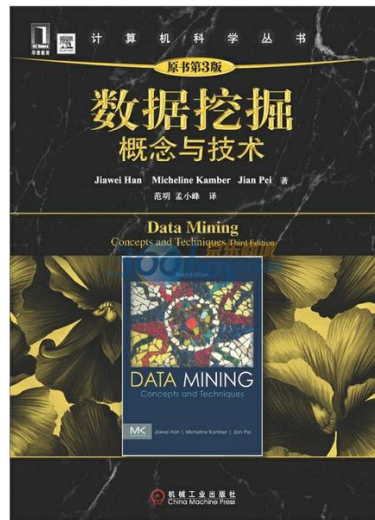
# 机器学习/数据挖掘脉络图



<http://www.healthcaresimulations.com/mindmaps/DataMining.html>

# 机器学习/数据挖掘脉络图

1. 认识数据/数据预处理
2. 挖掘频繁模式、关联和相关性
  - 频繁项集、频繁子序列（序列模式）、.....
3. 分类
  - 决策树归纳、贝叶斯分类方法、组合方法、神经网络、支持向量机、最邻近分类器、.....
4. 聚类分析
  - 层次方法、基于密度方法、基于网格方法
5. 离群点分析





**以机器学习为内核**

**数据分析  $\approx$  机器学习 / 数据挖掘**

**$\approx$  认识数据 + 关联 + 分类 + 聚类**

**$\approx$  寻找关系结构（核心是归归类）**

**以机器学习为内核**

**数据分析  $\approx$  机器学习 / 数据挖掘**

**$\approx$  认识数据 + 关联 + 分类 + 聚类**

**$\approx$  寻找关系结构 (核心是归归类)**

# 机器学习只是归归类

有监督学习——分类

无监督学习——聚类

构成了机器学习的主体部分

所以，机器学习不过是归一归类

# 机器学习的应用领域

Industries / Fields where you applied Analytics, Data Mining, Data Science in 2014? [221 voters]	
	<div> <div></div> 2014 % of voters                     <div></div> 2012 % of voters                 </div>
CRM/Consumer analytics (49)	<div> <div></div> 22.2%                     <div></div> 28.6%                 </div>
Banking (37)	<div> <div></div> 16.7%                     <div></div> 14.3%                 </div>
Health care (was Healthcare/HR) (36)	<div> <div></div> 16.3%                     <div></div> 16.3%                 </div>
Retail (30)	<div> <div></div> 13.6%                     <div></div> 14.8%                 </div>
Fraud Detection (30)	<div> <div></div> 13.6%                     <div></div> 12.8%                 </div>
Science (30)	<div> <div></div> 13.6%                     <div></div> 11.7%                 </div>
Other (30)	<div> <div></div> 13.6%                     <div></div> 10.2%                 </div>
Finance (24)	<div> <div></div> 10.9%                     <div></div> 10.2%                 </div>
Advertising (23)	<div> <div></div> 10.4%                     <div></div> 13.3%                 </div>
Oil / Gas / Energy (21)	<div> <div></div> 9.5%                     <div></div> na                 </div>
E-commerce (21)	<div> <div></div> 9.5%                     <div></div> 5.1%                 </div>
Manufacturing (20)	<div> <div></div> 9%                     <div></div> 7.10%                 </div>
Telecom / Cable (20)	<div> <div></div> 9%                     <div></div> 6.6%                 </div>
Social Media / Social Networks (19)	<div> <div></div> 8.6%                     <div></div> 12.2%                 </div>
Insurance (19)	<div> <div></div> 8.6%                     <div></div> 7.7%                 </div>
Credit Scoring (18)	<div> <div></div> 8.1%                     <div></div> 7.1%                 </div>
Education (17)	<div> <div></div> 7.7%                     <div></div> 14.3%                 </div>

Direct Marketing/ Fundraising (16)	<div> <div></div> 7.2%                     <div></div> 9.7%                 </div>
Medical/ Pharma (16)	<div> <div></div> 7.2%                     <div></div> 6.6%                 </div>
Software (16)	<div> <div></div> 7.2%                     <div></div> 5.6%                 </div>
Biotech/Genomics (15)	<div> <div></div> 6.8%                     <div></div> 7.7%                 </div>
Search / Web content mining (14)	<div> <div></div> 6.3%                     <div></div> 8.2%                 </div>
Government/Military (14)	<div> <div></div> 6.3%                     <div></div> 5.1%                 </div>
Automotive (13)	<div> <div></div> 5.9%                     <div></div> na                 </div>
HR/workforce analytics (13)	<div> <div></div> 5.9%                     <div></div> na                 </div>
Web usage/Log mining (13)	<div> <div></div> 5.9%                     <div></div> 6.6%                 </div>
Investment / Stocks (11)	<div> <div></div> 5.0%                     <div></div> 4.1%                 </div>
Travel / Hospitality (7)	<div> <div></div> 3.2%                     <div></div> 3.1%                 </div>
Mobile apps (5)	<div> <div></div> 2.3%                     <div></div> na                 </div>
Security / Anti-terrorism (5)	<div> <div></div> 2.3%                     <div></div> 3.6%                 </div>
Games (4)	<div> <div></div> 1.8%                     <div></div> na                 </div>
Entertainment/ Music/ TV/Movies (4)	<div> <div></div> 1.8%                     <div></div> 4.6%                 </div>
Social Policy/Survey analysis (4)	<div> <div></div> 1.8%                     <div></div> 1.0%                 </div>
Junk email / Anti-spam (4)	<div> <div></div> 1.8%                     <div></div> 0.5%                 </div>
Social Good/Non-profit (3)	<div> <div></div> 1.4%                     <div></div> na                 </div>

Source: KDnuggets

## 机器学习的应用领域

- 垃圾邮件过滤
- 个性化推荐
- 搜索引擎
- DNA序列分类
- 欺诈侦测
- 医学诊断
- 经济与金融
- 计算机视觉
- 语音识别
- 自然语言处理
- .....

# 曾经以为的归归类

## 生物分类

纲、目、属、种

## 图书分类

经、史、子、集

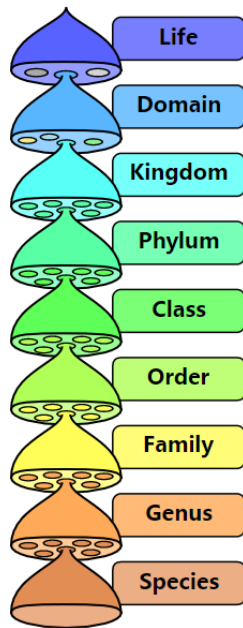
## 学科分类

01 哲学、02 经济学、.....08工学、.....

## 岗位分类

校长、副校长、财务处长、.....

## ABC分类法.....



85. 人类认知世界的基本方法，所有科学的核心问题是（ ）

- A. 分类
- B. 诊断
- C. 评估
- D. 心理测验

Source: 某大学生心理健康知识竞赛题库

## 归类，超乎常人的想象

现代认知心理学家认为：

人们认识事物时往往先把被认识的对象进行分类，以便寻找其中同与不同的特征，因而分类学是人们认识世界的基础科学，而**分类的方法则是人们认识世界的基本方法**。这种方法帮助我们有条理的认识这个纷繁复杂的世界。

王燕爽：分类能力与学习成绩，东北师范大学，2006年



## 我们所能做的，大部分不过是归归类而已

To perceive is to categorize,

**感知是归类**

to conceptualize is to categorize,

**概念化是归类**

to learn is to form categories,

**学习是归类**

to make decisions is to categorize

**决策还是归类**



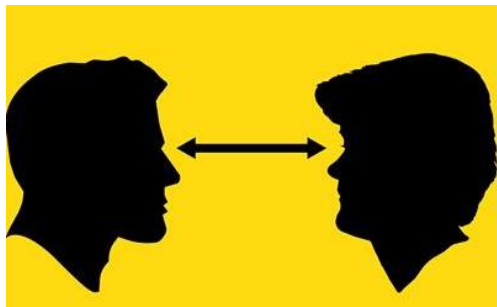
**Jerome Bruner**

1915 — 2016

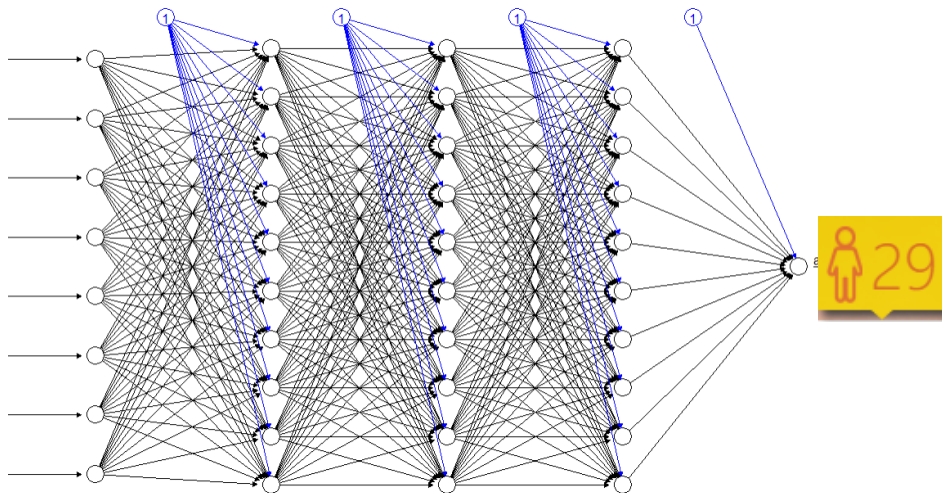
Theory of Categorization

<http://www.uwyo.edu/aded5050/5050unit8/bruner.asp>

我们所能做的，大部分不过是归归类而已



# 机器学习也只不过是归归类



类别={人类可能的年龄}

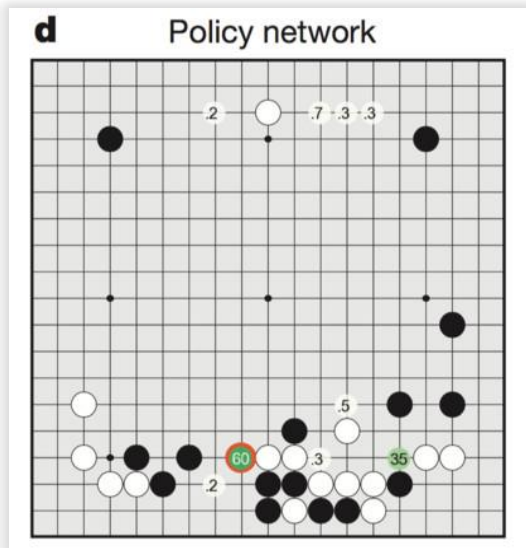
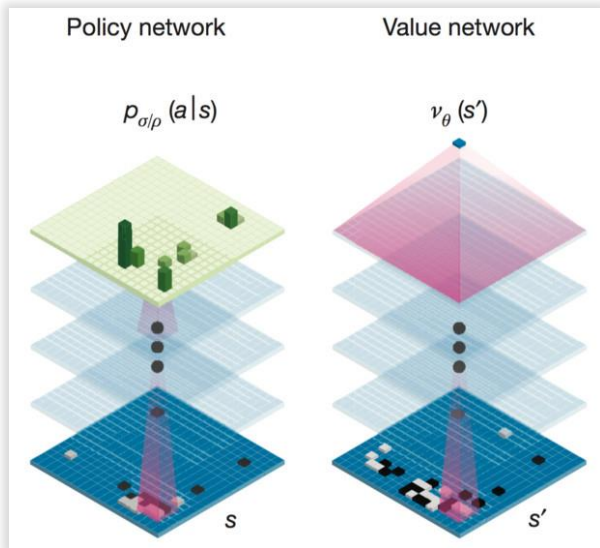
# 机器学习也只不过是归归类



类别={天空,建筑,道路,人行道,栅栏,植物,柱子,汽车,标识,行人,自行车}

Ros, G., Sellart, L., Materzynska, J., Vazquez, D., & Lopez, A. M. (2016). The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 3234-3243).

# 机器学习也只不过是归归类

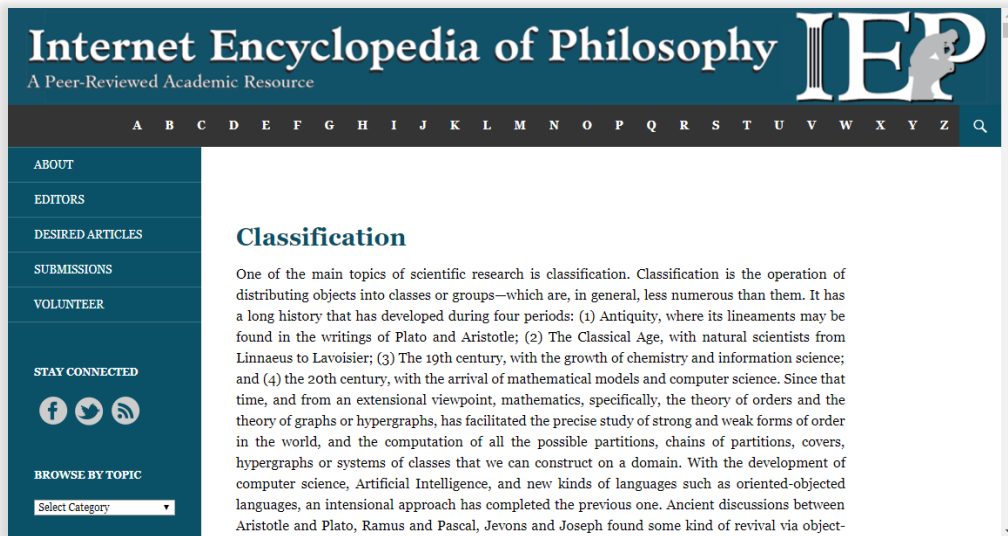


DAN MAAS:How AlphaGo Works

<http://www.d cine.com/2016/01/28/alphago/>

类别={落子位置}

# 对归类理论的哲学探讨



The screenshot shows the homepage of the Internet Encyclopedia of Philosophy (IEP). The header features the title 'Internet Encyclopedia of Philosophy' in a large, serif font, with the subtitle 'A Peer-Reviewed Academic Resource' below it. To the right is the IEP logo, which consists of the letters 'IEP' in a stylized, classical font. Below the header is a navigation bar with letters A through Z, and a search icon. On the left side, there is a sidebar with links to 'ABOUT', 'EDITORS', 'DESIRED ARTICLES', 'SUBMISSIONS', and 'VOLUNTEER'. Below these links are social media icons for Facebook, Twitter, and RSS. At the bottom of the sidebar is a 'BROWSE BY TOPIC' section with a dropdown menu labeled 'Select Category'. The main content area displays the title 'Classification' in a bold, serif font. Below the title is a paragraph of text discussing the history and importance of classification in scientific research.

## Internet Encyclopedia of Philosophy

A Peer-Reviewed Academic Resource

A B C D E F G H I J K L M N O P Q R S T U V W X Y Z

ABOUT  
EDITORS  
DESIRED ARTICLES  
SUBMISSIONS  
VOLUNTEER

STAY CONNECTED

f t r

BROWSE BY TOPIC

Select Category

### Classification

One of the main topics of scientific research is classification. Classification is the operation of distributing objects into classes or groups—which are, in general, less numerous than them. It has a long history that has developed during four periods: (1) Antiquity, where its lineaments may be found in the writings of Plato and Aristotle; (2) The Classical Age, with natural scientists from Linnaeus to Lavoisier; (3) The 19th century, with the growth of chemistry and information science; and (4) the 20th century, with the arrival of mathematical models and computer science. Since that time, and from an extensional viewpoint, mathematics, specifically, the theory of orders and the theory of graphs or hypergraphs, has facilitated the precise study of strong and weak forms of order in the world, and the computation of all the possible partitions, chains of partitions, covers, hypergraphs or systems of classes that we can construct on a domain. With the development of computer science, Artificial Intelligence, and new kinds of languages such as oriented-objected languages, an intensional approach has completed the previous one. Ancient discussions between Aristotle and Plato, Ramus and Pascal, Jevons and Joseph found some kind of revival via object-

*<http://www.iep.utm.edu/classifi/> (accessed on 20 April 2018)*

A decorative blue border with rounded corners frames the entire slide. Two thin blue crosshair-like lines are positioned diagonally, one in the upper right and one in the lower left, intersecting at the center of the text.

**谢谢聆听**  
**Thank you**

# 教师个人联系方式

艾新波

手机: 13641159546

QQ: 23127789

微信: 13641159546

E-mail: [13641159546@126.com](mailto:13641159546@126.com)

[axb@bupt.edu.cn](mailto:axb@bupt.edu.cn)

地址: 北京邮电大学科研楼917室

课程网址: <https://github.com/byaxb/RDataAnalytics>

