湖南科技大学 计算机科学与工程学院

数据分析技术

主讲人：陶 洁

邮 箱：553093468@qq.com

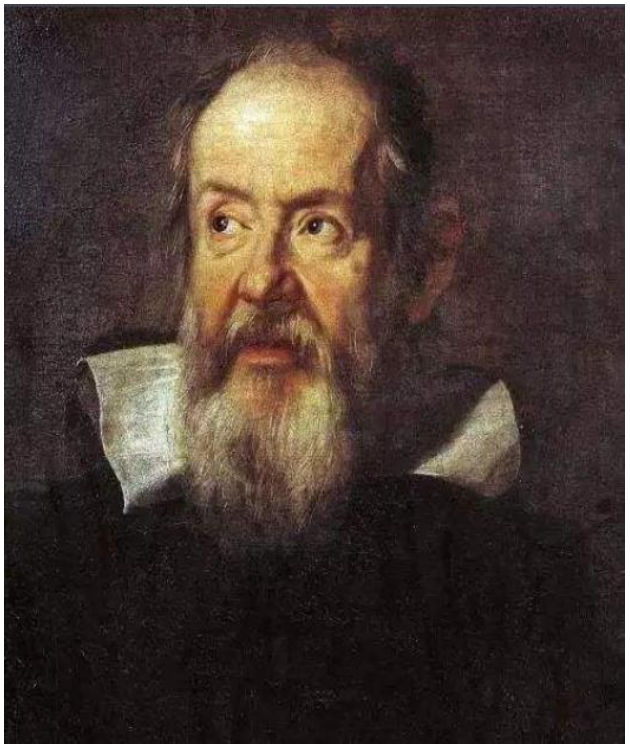
办公室：逸夫楼410

课程简介

- **数据分析技术：**
  用适当的数据挖掘和知识发现技术对收集来的大量数据进行分析，提取有用信息和形成结论而对数据加以详细研究和概括总结的过程。
- **课程概况：**
  3.5学分，48理论课时，8实践课时
- **教材信息：**
  《数据挖掘原理与算法》(第三版)
  By 毛国君，段立娟，王石，石云
  Pub. 清华大学出版社，2016

古代数据分析技术的特点

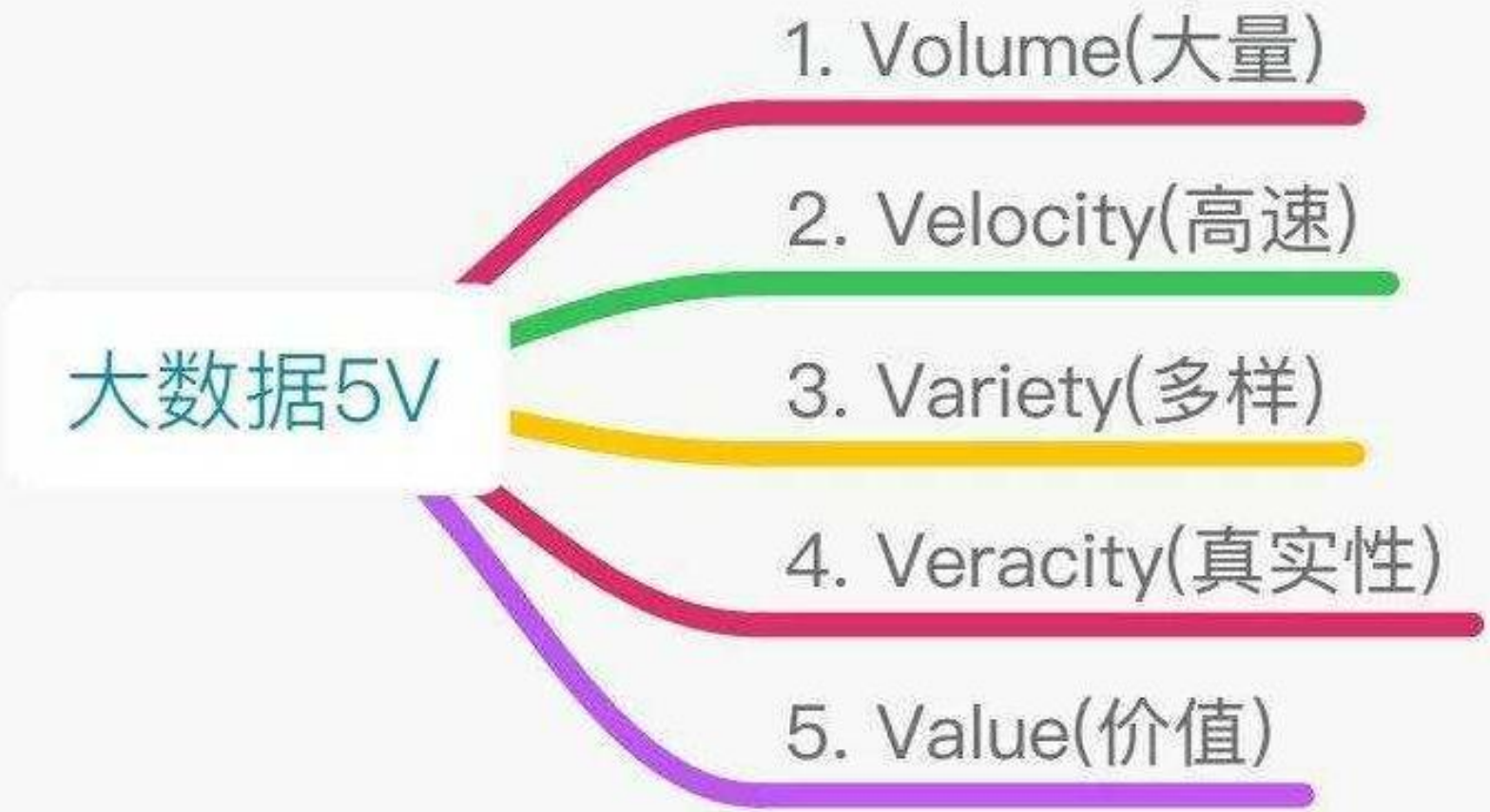
祖冲之
南北朝时期杰出的
数学家、天文学家



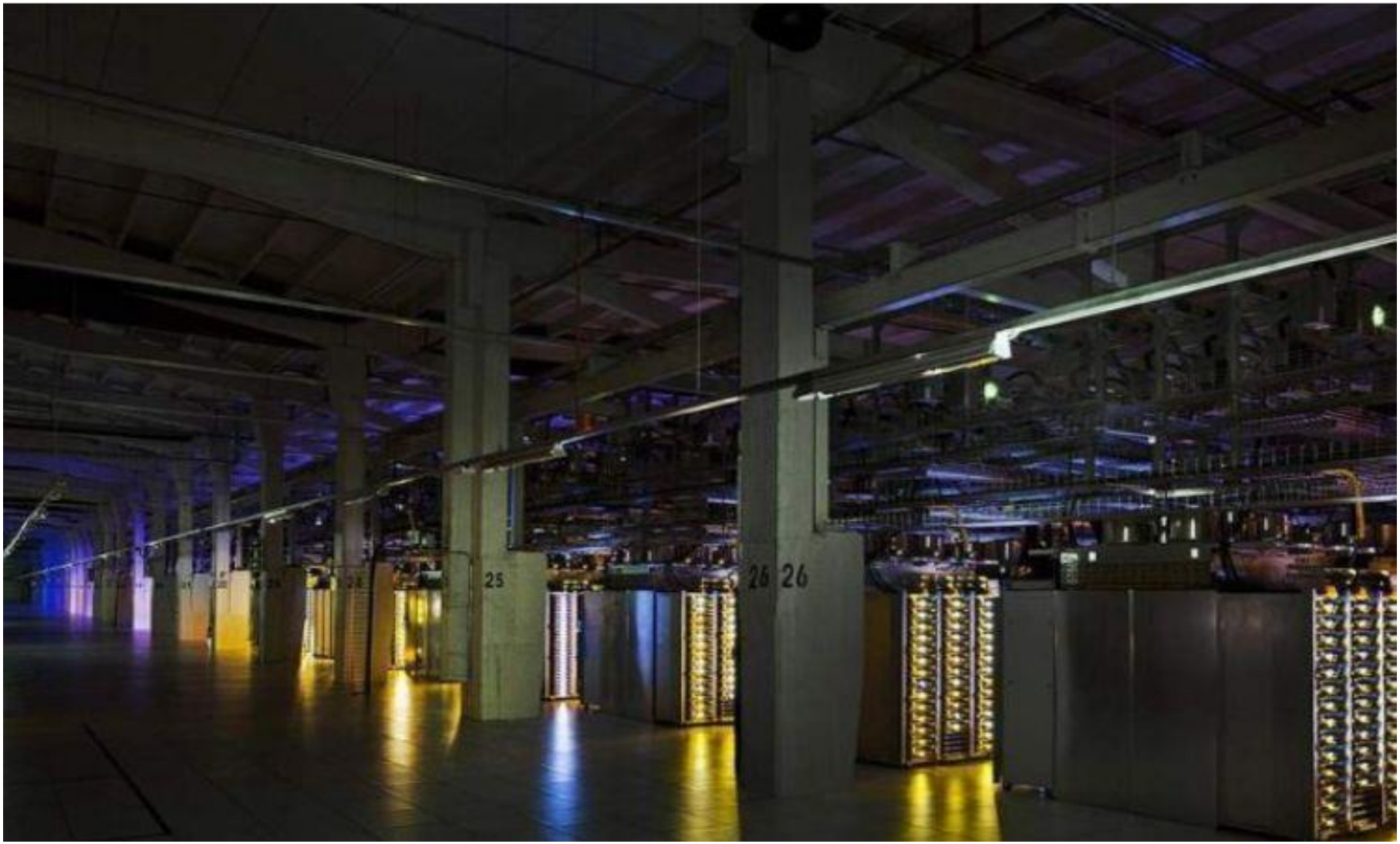
伽利略
近代科学实验奠
基人之一

竺可桢
中国近代地理学
气象学的奠基者

现代的数据分析技术的特点



数据分析技术的特点——谷歌数据中心

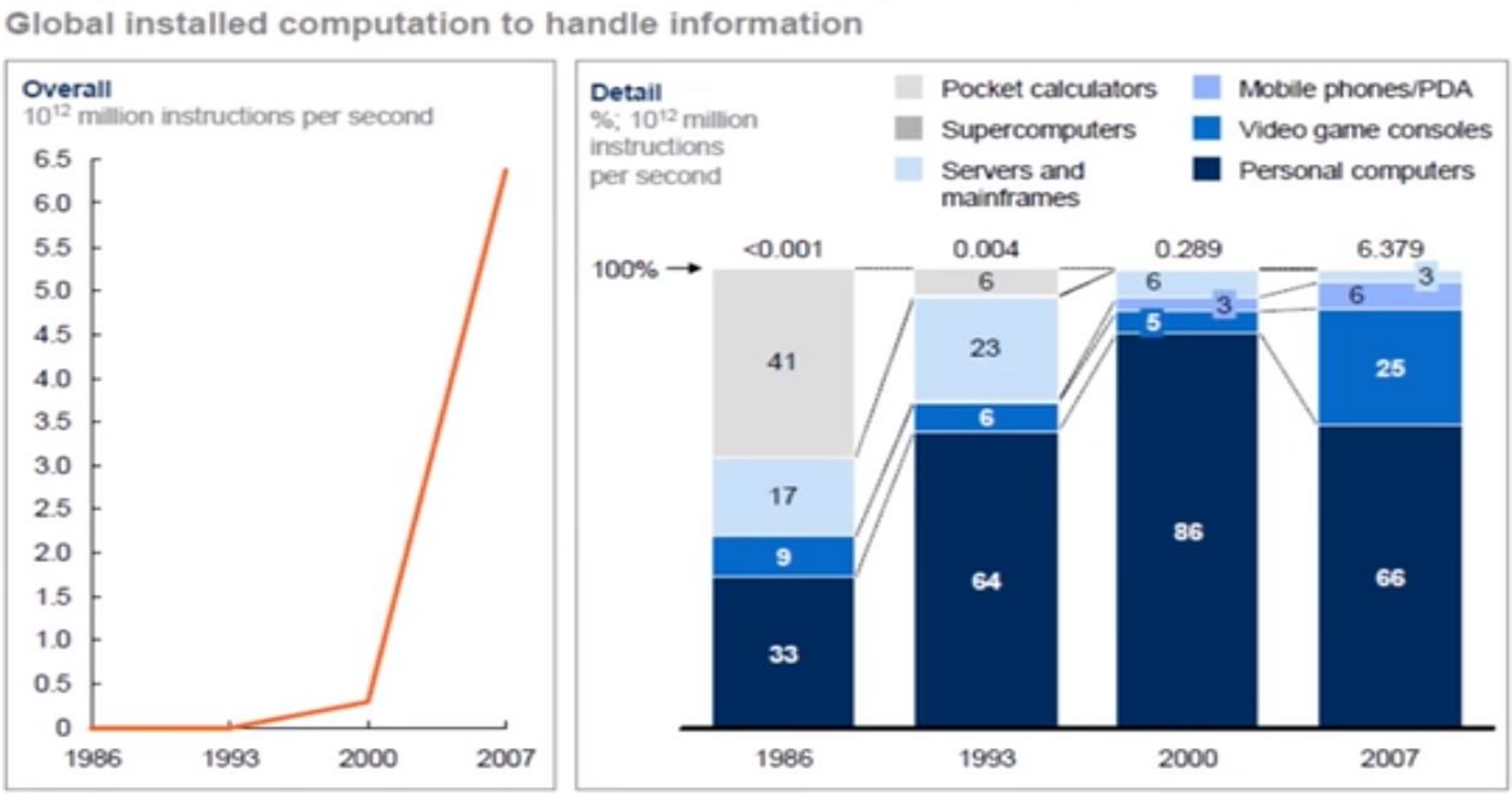


数据分析技术的产生与发展

耗电、散热？

大数据，低信息？

计算机——算力的提升



NOTE: Numbers may not sum due to rounding.

SOURCE: Hilbert and López, "The world's technological capacity to store, communicate, and compute information," *Science*, 2011

- 随着信息技术的高速发展，人们希望能够提供更高层次的数据处理功能。新的需求推动数据分析技术的诞生。

“顾客买了哪些商品？”

“商品之间有什么联系？”

“用这些联系来促进销售”

数据分析技术应用实例


数据挖掘原理与算法
（第3版）
清华大学出版社


经常一起购买的商品
总价：￥204.00
全部加入购物车
本商品：计算机科学丛书：数据挖掘：概念与技术(原书第3版) 平装 ￥61.50
数据挖掘导论(完整版) - 陈封能 (Pang-Ning Tan) 平装 ￥57.00
利用Python进行数据分析 - 麦金尼 (Wes McKinney) 平装 ￥85.50


购买此商品的顾客也同时购买
¥57.00 prime
¥54.50 prime
¥85.50
¥62.10 prime

课程简介

- 第1章 绪论
- 第2章 知识发现过程与应用结构
- 第3章 关联规则挖掘理论和算法
- 第4章 分类方法
- 第5章 聚类方法
- 第6章 时间序列和序列模式挖掘
- 第7章 Web挖掘技术
- 第8章 空间挖掘

数据挖掘技术的产生与发展

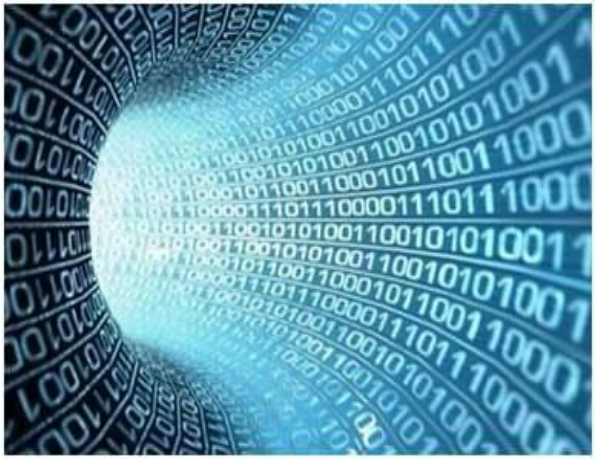
大量的数据、强大的算力、实际需求

data → information ↔ knowledge

- 数据挖掘技术的产生
- 数据挖掘概念
- 数据挖掘技术的分类问题
- 数据挖掘常用的知识表示模式与方法
- 不同数据存储形式下的数据挖掘问题
- 粗糙集方法及其在数据挖掘中的应用
- 数据挖掘的应用分析

数据分析技术的定义

- 数据分析技术是从大量的、不完全的、有噪声的、模糊的、随机的数据集中识别有效的、新颖的、潜在有用的，以及最终可理解的模式的非平凡过程。它是一门涉及面很广的交叉学科，包括机器学习、数理统计、神经网络、数据库、模式识别、粗糙集等相关技术。



中彩票，可以用数据挖掘技术吗?

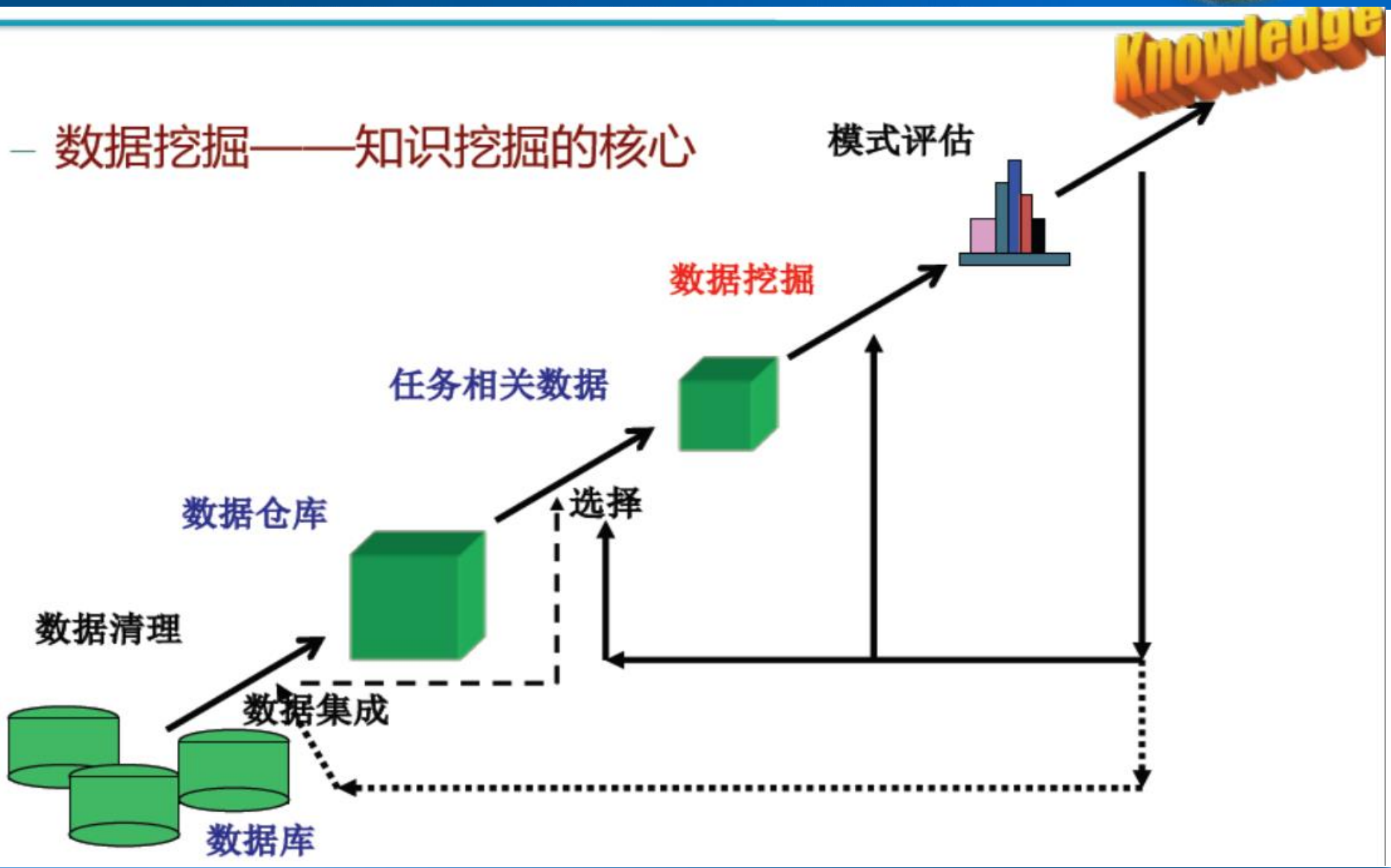
号码走势图 | 综合分析表

大乐透号码分布图 --请选择期号-- 至 --请选择期号-- 查看

期号	号码	中奖号码分布图																																														
		前区																																			后区											
		01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	01	02	03	04	05	06	07	08	09	10	11	12
09075	01 07 11 32 33 + 09 11	1						7				11																					32	33											9		11	
09076	03 09 11 18 24 + 04 05			3						9		11							18						24															4	5							
09077	07 20 28 32 34 + 06 08							7													20								28				32		34							6		8				
09078	11 23 26 31 34 + 05 11											11												23			26					31			34						5						11	
09079	01 02 07 13 34 + 04 10	1	2					7						13																					34					4						10		
09080	07 16 19 29 35 + 02 11							7									16			19										29						35		2									11	
09081	02 15 20 25 33 + 03 06		2													15					20					25								33					3			6						
09082	05 11 26 29 32 + 08 11					5						11															26			29			32											8			11	
09083	02 05 28 34 35 + 01 07		2			5																							28						34	35	1						7					
09084	05 09 19 26 35 + 02 10					5				9										19							26									35		2								10		
期号	号码	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	01	02	03	04	05	06	07	08	09	10	11	12

数据分析技术

– 数据挖掘——知识挖掘的核心



数据分析技术在生活中的应用

数据分析技术相关的科学研究

图 4 2013-2018 KDD 研究性论文投稿与接收情况

参考资料


Java


python


Google
scholar

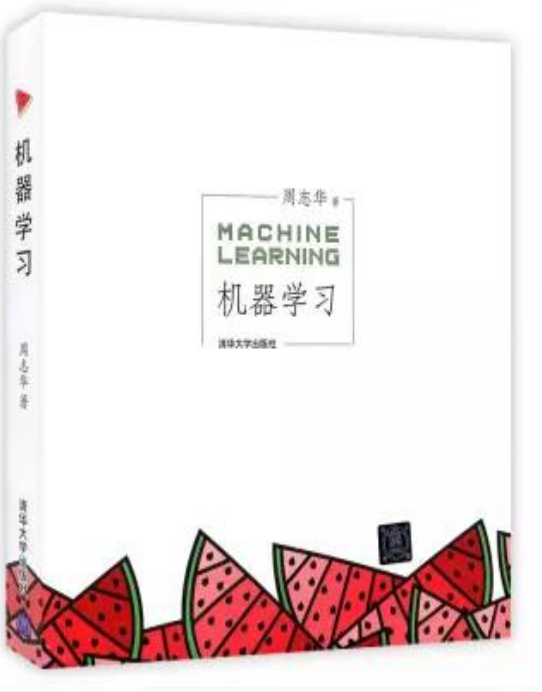

Weka GUI Chooser
Program Visualization Tools Help
WEKA
The University of Waikato
Applications
Explorer
Experimenter
KnowledgeFlow
Simple CLI
Waikato Environment for Knowledge Analysis
Version 3.6.2
(c) 1999 - 2010
The University of Waikato
Hamilton, New Zealand


KDnuggets™

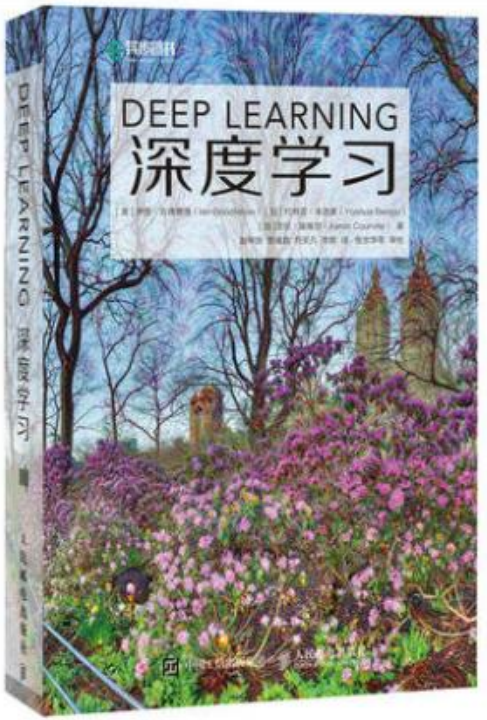

UCI
Machine Learning Repository

参考数目

昆明理工大学
袁梅宇

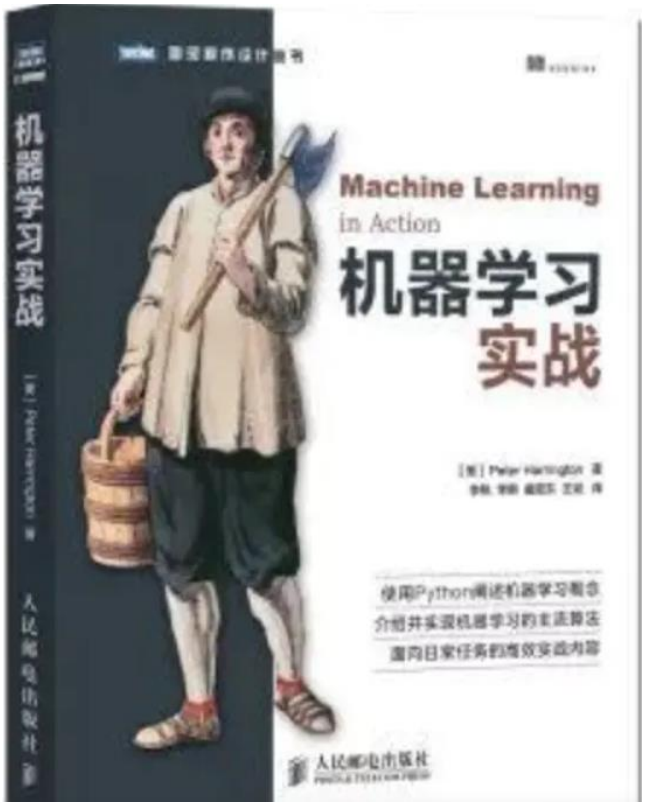


南京大学
周志华

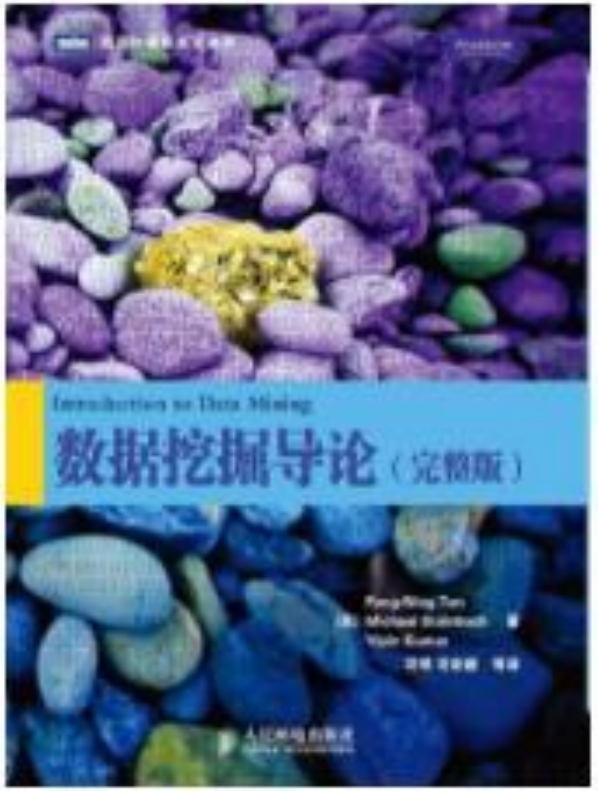


美国
伊恩·古德费洛

参考数目



使用python



理论与经典案例

数据挖掘技术举例——十大算法

一、 C4.5：分类决策树算法,其核心算法是ID3 算法

二、k-means：k-平均算法是解决聚类问题

三、SVM：支持向量机

四、Apriori：挖掘布尔关联规则频繁项集的算法

五、EM ：最大期望算法

六、PageRank：网页排名、搜索引擎

七、AdaBoost：自适应增强

八、KNN：k-近邻算法

九、Naive Baye：朴素贝叶斯分类器

十、CART：分类回归树

投 票(可多选)

- 网上购物时，用户推荐系统可能用到以下哪些方法？

A、决策树

B、k-means

C、SVM

D、Apriori

E、EM

F、KNN

G、Naive Baye

H、其它（填写）_____.

谢谢！