数据挖掘的基本原理

Recommend
火龙果软件

报 告 内 容

- 数据挖掘的基本概念
- 数据挖掘的基本算法
- 数据挖掘实施方法论
- 总结与讨论

Recommend
火龙果软件

- ☞ 数据挖掘的基本概念
- ❖ 数据挖掘的基本算法
- ❖ 数据挖掘实施方法论
- ❖ 总结与讨论

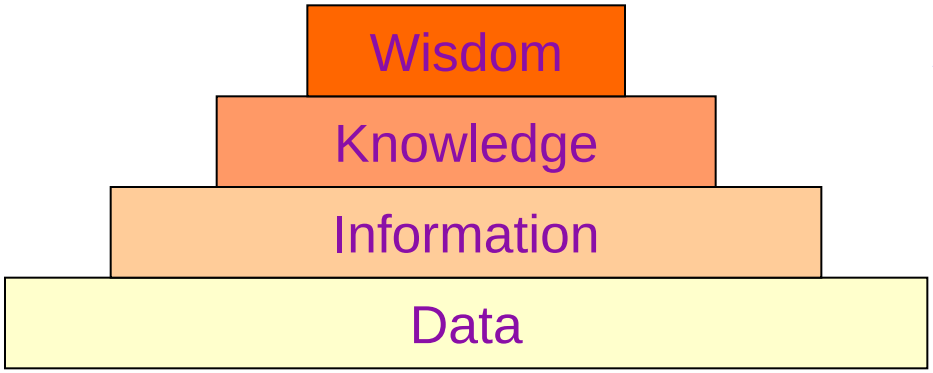
Recommend
火龙果软件

改变未来世界的十大新兴技术

- 《Technology Review》（麻省理工学院2002年1月出刊）
  - 机器与人脑的接口
  - 塑胶晶体管
  - 数据挖掘（Data Mining）
  - 数字权利管理
  - 生物测定学（Biometrics）
  - 语言识别处理
  - 微光学技术（Microphotonics）
  - 解开程序代码（Untangling Code）
  - 机器人设计
  - 微应用流体学（Microfluidics）

什么是数据挖掘？

- 存在太多数据挖掘的定义，但基本上有这样一种描述结构
  - To find / discover / extract / dredge / harvest 、、、
  - Interesting / novel / useful / implicit / actable / meaningful 、、、
  - Information / knowledge / patterns / trends / rules / anomalies 、、、
  - In massive data / large data set / large database / data warehouse 、、、



Knowledge + experience

Information + rules

Data + context

Recommend
火龙果软件

为什么会出现数据挖掘？

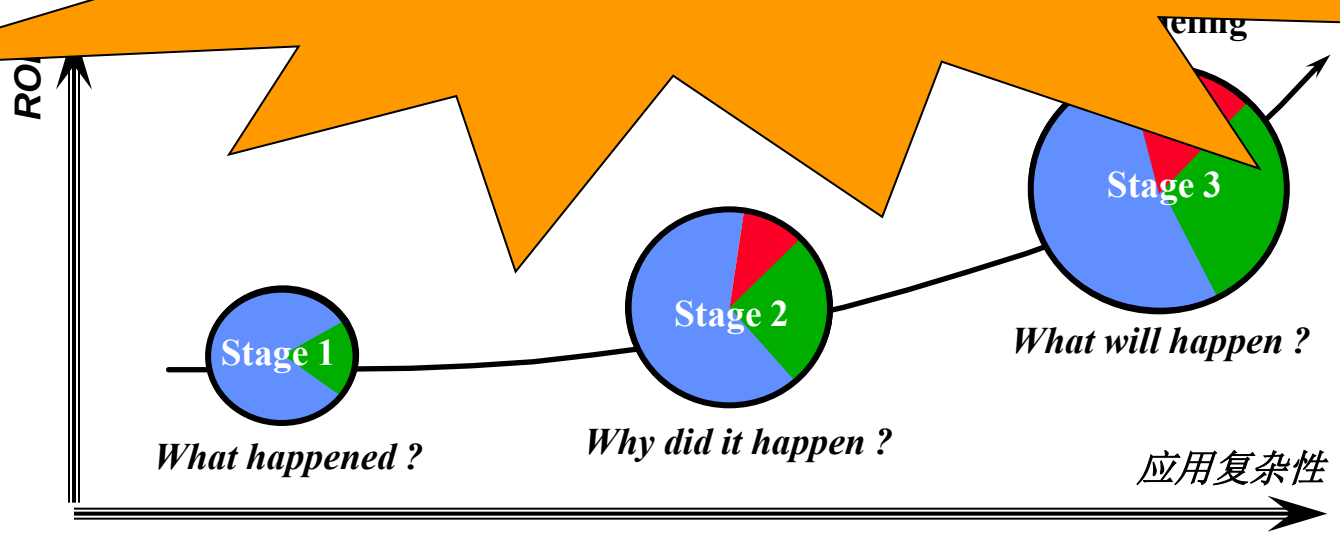
- 数据爆炸性增长是数据挖掘技术应运而生的根本原因。
  - 只见树木，不见森林（Drowning in data but starving for information）
  - 计算复杂度
  - 数据管理问题
  - 数据类型的多样性

处理大容量数据是数据挖掘技术区别于其他数据分析方法的唯一标志吗？

其他数据分析方法：商业智能

- E.F.Codd的数据分析模型
  - 绝对模型（Categorical Model）：依据预定义路径寻找原因，如查询
  - 解释模型（Exegetical Model）：依据多层[illegible]寻找原因，如多维分析
  - 思考模[illegible]析

现象 ＝ 模型 ＋ 误差
数据挖掘寻找的是模型 ！



Human Discovery　　Machine-assisted Discovery

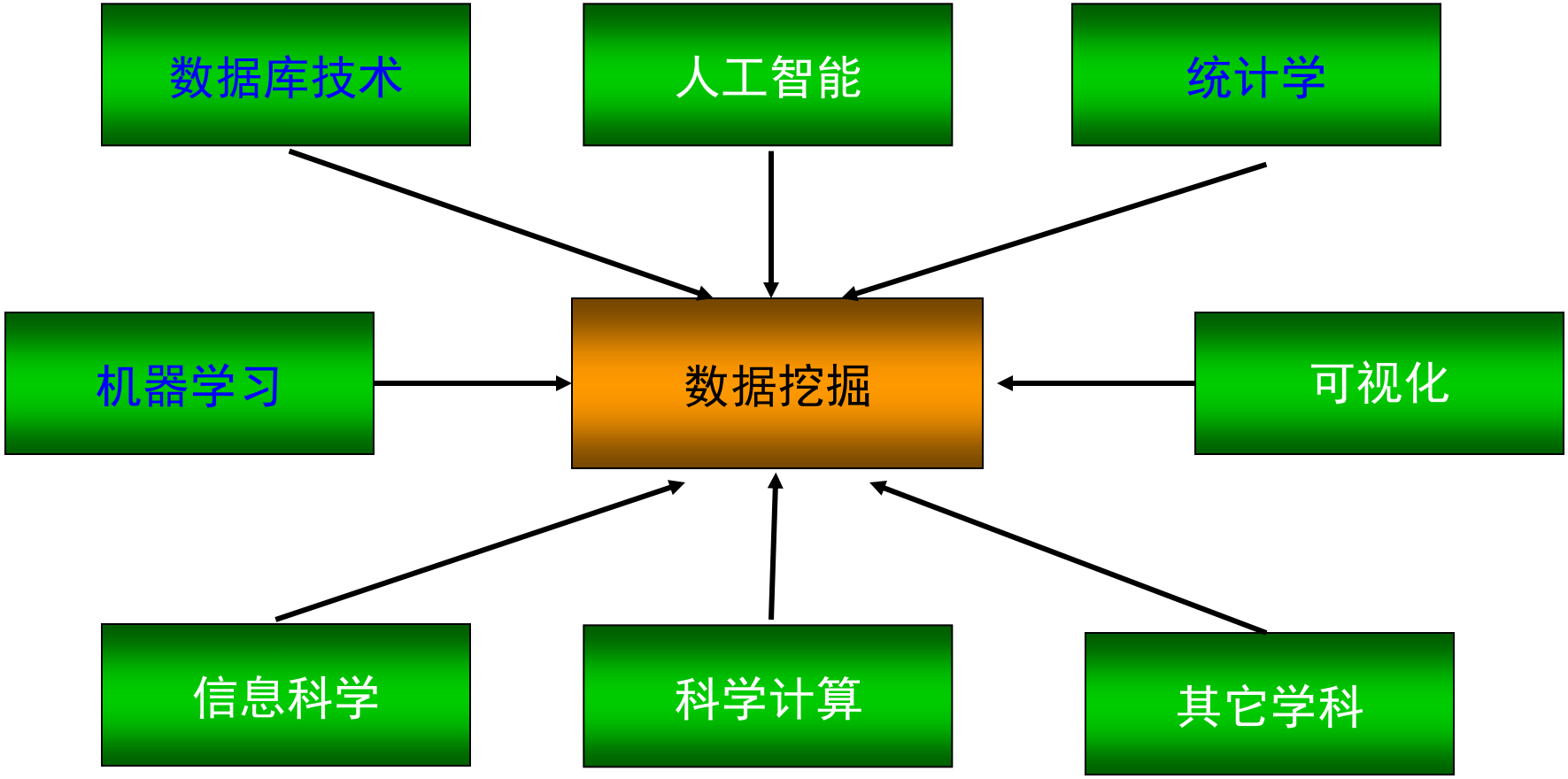
Recommend
火龙果软件

其他数据分析方法：统计学

- 从处理数据的角度看、、、
  - 数据规模不同
  - 数据来源不同：观测数据（Secondary Analysis）　　试验数据（Primary Analysis）
  - 数据　　同（结构化数据　　化数据　　据）
- 从分析目　　
  - 更关注　　alysis）
  - 更关注模型（Mo
- 但二者具有相当密切的
  - 从数据分析的角度，统计学现　是且仍将是数据挖掘最重要的技术支撑和思想源泉
  - 更加深入的渗透和交叉（如探索性数据分析，EDA）

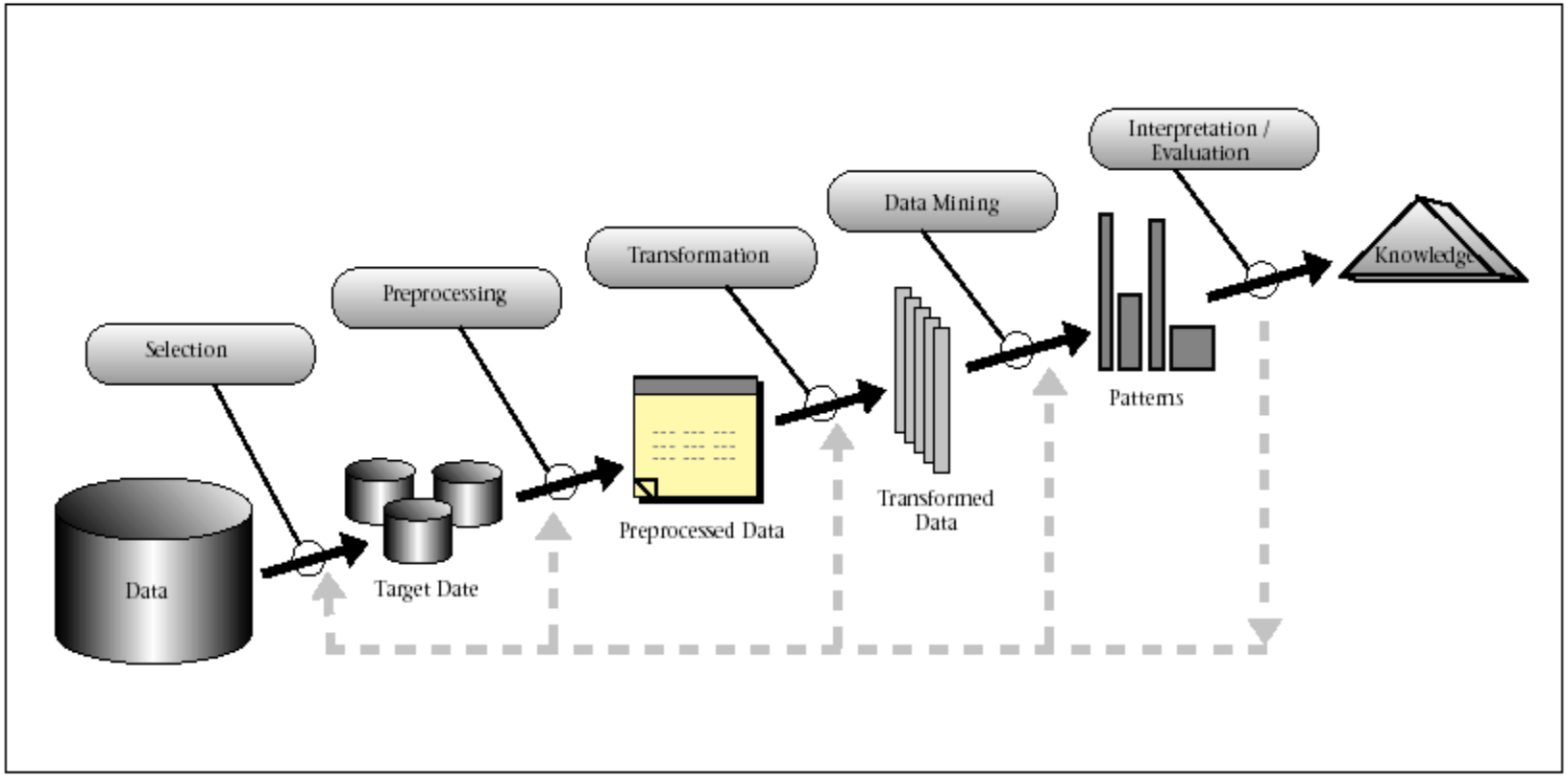
数据挖掘是数据驱动的探索性分析 ！

数据挖掘：多学科的汇合



Recommend
火龙果软件

数据挖掘是一个过程



An Overview of the Steps That Compose the KDD Process.

- *"from data mining to knowledge discovery in database". U. fayyad, G.P.Shapiro and P.Smyth (1996)*

数据挖掘过程中的数据预处理

- 数据清洗
  - 填充缺失值，修均噪声数据，识别或删除孤立点，并解决数据不一致问题
  - 主要分析方法：分箱（Binning）、聚类、回归
- 数据集成
  - 多个数据库、数据方或文件的集成
- 数据变换
  - 规范化与汇总
- 数据简化
  - 减少数据量的同时，还可以得到相同或相近的分析结果
  - 主要分析方法：抽样、主成分分析
- 数据离散化
  - 数据简化的一部分，但非常重要（尤其对于数值型数据来说）

数据挖掘过程中的数据探索

- 探索性数据分析（Exploratory Data Analysis, EDA）
  - 探索性地查看数据，概括数据集的结构和关系
  - 对数据集没有各种严格假定
  - “玩”数据
- 主要任务
  - 数据可视化（a picture is worth a thousand words）
  - 残差分析（数据＝拟合 + 残差）
  - 数据的重新表达（什么样的尺度－对数抑或平方跟－会简化分析？）
  - 方法的耐抗性（对数据局部不良的不敏感性，如中位数耐抗甚于均值）
- 常见方法
  - 统计量，如均值、方差、根方差、协方差、峰度、偏度、相关系数等
  - 统计图，如饼图、直方图、散点图、箱尾图等
  - 模型，如聚类

数据挖掘结果的评价

- 兴趣度度量：一个模式是有意义的，如果它易于被人理解，在某种程度上，对于新数据或者测试数据是有效的、潜在有用或者验证了用户渴望确认的某些假设。
  - 目前仍无很好的解决办法，很大程度上仍依靠人工
  - 不存在解决这个问题的简单技术，最终答案是不要把数据挖掘当作脱离数据内涵的简单技术来运用
  - 客观兴趣度：基于统计或模式的结构，如统计量、支持度、lift等
  - 主观兴趣度：基于用户对数据的确信程度，如意外程度、新奇程度或者可行动性等
  - 过度拟合（Over-fitting）问题

什么不是数据挖掘？

- 定量分析（Quantitative Analysis）的需要存在企业管理运行的各个侧面或环节，但并非所有的定量分析问题都可以归结到数据挖掘范畴的问题。
  - 简单的报表、图表及多维分析仍是日常分析工作的主要内容
  - 小样本数据的分析传统统计分析方法更成熟有效，如趋势预测
  - 某些特定业务问题无法用数据挖掘算法加以解决，例如
    - 资源最优配置问题是个运筹学问题
    - 某些物流管理问题或者供应链管理问题是个随机规划问题
    - 营销预演本质是个系统仿真问题

- 数据挖掘的基本概念
- ☞ **数据挖掘的基本算法**
- 数据挖掘实施方法论
- 总结与讨论

Recommend
火龙果软件

几个基本概念

- 模型（Model） vs 模式（Pattern）
  - 数据挖掘的根本目的就是把样本数据中隐含的结构泛化（Generalize）到总体（Population）上去
  - 模型：对数据集的一种全局性的整体特征的描述或概括，适用于数据空间中的所有点，例如聚类分析
  - 模式：对数据集的一种局部性的有限特征的描述或概括，适用于数据空间的一个子集，例如关联分析
- 算法（Algorithm）：一个定义完备（well-defined）的过程，它以数据作为输入并产生模型或模式形式的输出
- 描述型挖掘（Descriptive） vs 预测型挖掘（Predictive）
  - 描述型挖掘：对数据进行概括，以方便的形式呈现数据的重要特征
  - 预测型挖掘：根据观察到的对象特征值来预测它的其他特征值
  - 描述型挖掘可以是目的，也可以是手段

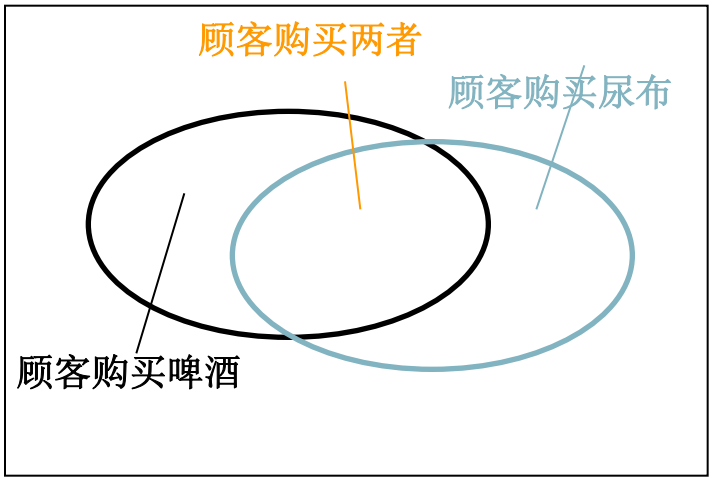
几类基本的挖掘算法

- 关联规则（模式、描述型）
  - 发现数据集中的频繁模式
  - 例如：buy(x,”diapers”) ⇒ buy(x,”beers”) [0.5%, 60%]
- 分类与预测（模型、预测型）
  - 发现能够区分或预测目标变量（唯一的）的规则或者函数
  - 分类的目标变量一般是范畴型的，而预测则是数量型的，并不必然带有任何时间延续型的暗示
  - 例如：股票市值的预测，病人病情的判断
- 聚类（模型、描述型）
  - 对数据分组以形成新类，类标记是未知的
  - 例如：市场细分
- 孤立点探测（Outlier Detection）（模式、预测型）
  - 分析异常或噪声数据的行为模式
  - 例如：欺诈检测

关联规则的基本概念

- 基本定义
  - 给定（1）事务数据集（2）每个事务是数据项的集合，试图发现项集中的频繁模式或关联关系
  - 所谓频繁模式或者关联规则就是一个具有“A ⇒ B”形式的逻辑蕴涵式
  - 频繁模式并不必然蕴涵着因果关系或相关关系！
  - 算法实现基本上基于APRIORI法则：频繁项集的所有非空子集一定也是频繁（Frequent）的
- 基本分类
  - 布尔关联规则 vs 定量关联规则
    - buy(x,”diapers”) ⇒ buy(x,”beers”)
    - Age(x,”30..39”) ^ income(x,”42k..48k”) ⇒ buy(x,”PC”)
  - 单维关联规则 vs 多维关联规则
  - 单层关联规则 vs 多层关联规则
    - Age(x,”30..39”) ^ income(x,”42k..48k”) ⇒ buy(x,”IBM PC”)
  - 序列模式（Sequence Pattern）
    - 数据项是一个包含时间标签的序偶[item(i),t]

关联规则的度量



- 发现具有最小置信度和支持度的全部规则 $X \wedge Y \Rightarrow Z$
  - 支持度(*support*), *s*, 事务中包含{X & Y & Z}的概率
  - 置信度(*confidence*), *c*, 事务中包含{X & Y}的条件下, 包含*Z*的条件概率

Transaction ID	Items Bought
2000	A,B,C
1000	A,C
4000	A,D
5000	B,E,F

- *令最小支持度为50%, 最小置信度为50%, 则有*
  - $A \Rightarrow C$ (50%, 66.6%)
  - $C \Rightarrow A$ (50%, 100%)

对支持度与置信度的批判

- 示例
  - 总共5000名学生，其中
    - 3000人玩篮球
    - 3750人吃谷类食品
    - 2000人既玩篮球又吃谷类食品
  - *play basketball* ⇒ *eat cereal* [40%, 66.7%] 是一个误导规则，因为吃谷类食品的学生占学生总数的75%，比66.7%更高
  - *play basketball* ⇒ *not eat cereal* [20%, 33.3%] 其实是一个更精确的规则，尽管它的支持度和置信度都比较低

	basketball	not basketball	sum(row)
cereal	2000	1750	3750
not cereal	1000	250	1250
sum(col.)	3000	2000	5000

Recommend
火龙果软件

关联规则的应用

- 市场购物篮分析（Market Basket Analysis）
  - 例如一个事务是客户的一个购物清单，同一客户的两份清单被认为是两个不同的事务
  - 数据项是所有可能陈列货物的全集
  - 目标是发现同时出现的货品组合间的关联模式
  - 应用：商品货价设计、仓储规划、网页布局、产品目录设计等等
- 交叉销售（Cross Selling）
  - 客户依次购买不同产品的序列
  - 目标是发现在购买某一产品组合之后客户可能购买的另一产品或服务
  - 应用：网络故障分析、网站门户设计等

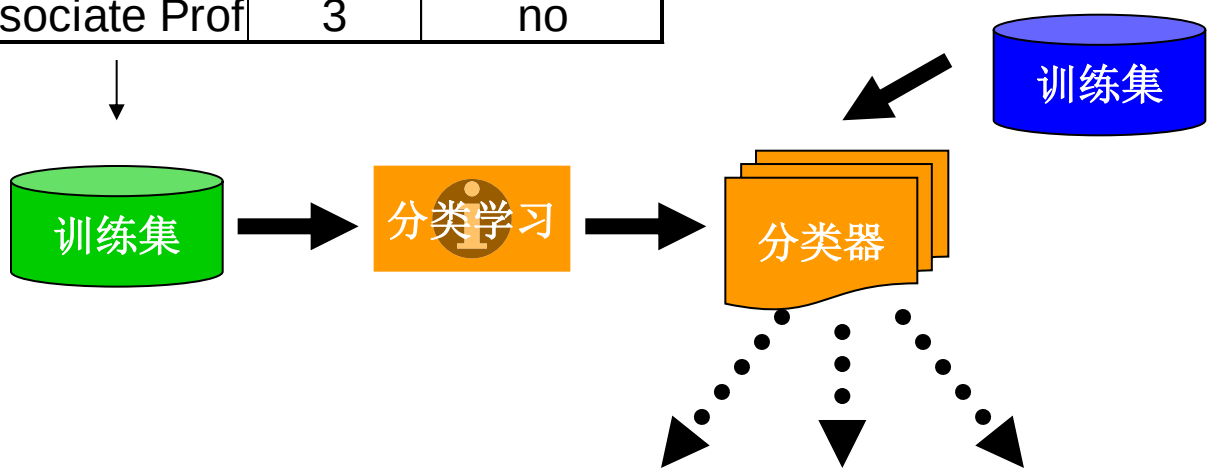
分类问题的基本定义

- 给定一数据集合（训练集）
  - 数据记录由一系列变量组成
  - 其中有一个变量是目标分类标签
- 寻找一模型，使目标分类变量值是其他变量值的一个函数
- 利用上述函数，一未知分类变量值的数据记录能够尽可能准确地被判定到某一类别中去
  - 一般会有另一独立地数据集（测试集）用以验证所构建分类函数的准确性，避免过度拟合

分类过程示意

NAME	RANK	YEARS	TENURED
Mike	Assistant Prof	3	no
Mary	Assistant Prof	7	yes
Bill	Professor	2	yes
Jim	Associate Prof	7	yes
Dave	Assistant Prof	6	no
Anne	Associate Prof	3	no

NAME	RANK	YEARS	TENURED
Tom	Assistant Prof	2	no
Merlisa	Associate Prof	7	no
George	Professor	5	yes
Joseph	Assistant Prof	7	yes



NAME	RANK	YEARS	TENURED
Jef	Professor	4	?

IF rank = ‘professor’
OR years > 6
THEN tenured = ‘yes’

Jef is YES!

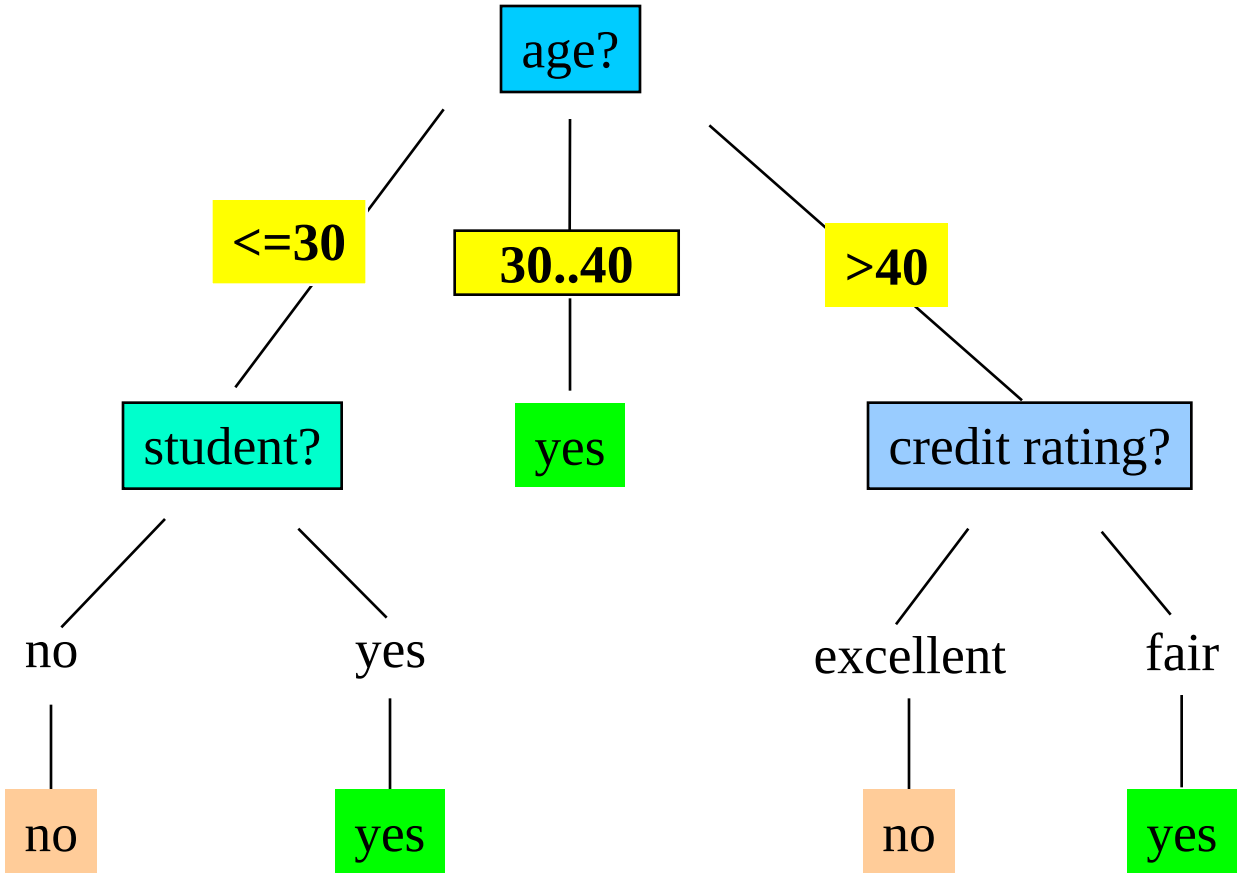
Recommend
火龙果软件

分类中的决策树（Decision Tree）归纳

- 决策树
  - 类似于流程图的树型结构
  - 内部节点代表对某个属性的一次测试
  - 分支代表测试的输出结果
  - 叶节点代表分类标签或分布
- 决策树的生成包括两个阶段
  - 树的创建
    - ✓ 首先，所有训练样本都位于根节点
    - ✓ 递归地基于选择属性来划分样本集
  - 树的修剪
    - ✓ 识别并删除那些反映噪声或孤立点的分支
- 应用决策树：对未知样本进行分类
  - 在决策树上测试样本的各个属性值

决策树示意

- 示例：是否购买计算机？



Recommend
火龙果软件

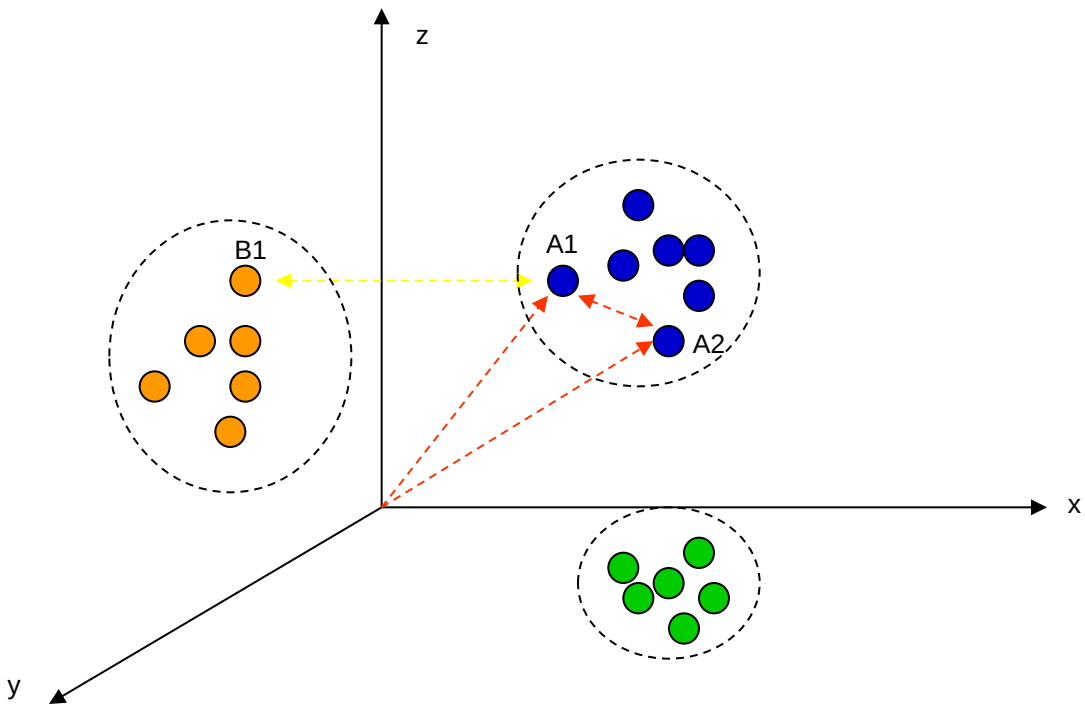
聚类的基本概念

- 基本定义
  - 将数据对象集划分成事先未知的分组或类别
  - 聚类的原则：类内相似度高，类间相似度低
  - 相似度一般为某种距离函数D(i,j)
  - 聚类既可以作为独立分析工具考察数据分布结构，也可以作为其他分析方法的预处理步骤
  - 很不幸，对聚类结果的评价一般都是主观的
- 基本分类
  - 将数据对象集划分成事先未知的分组或类别

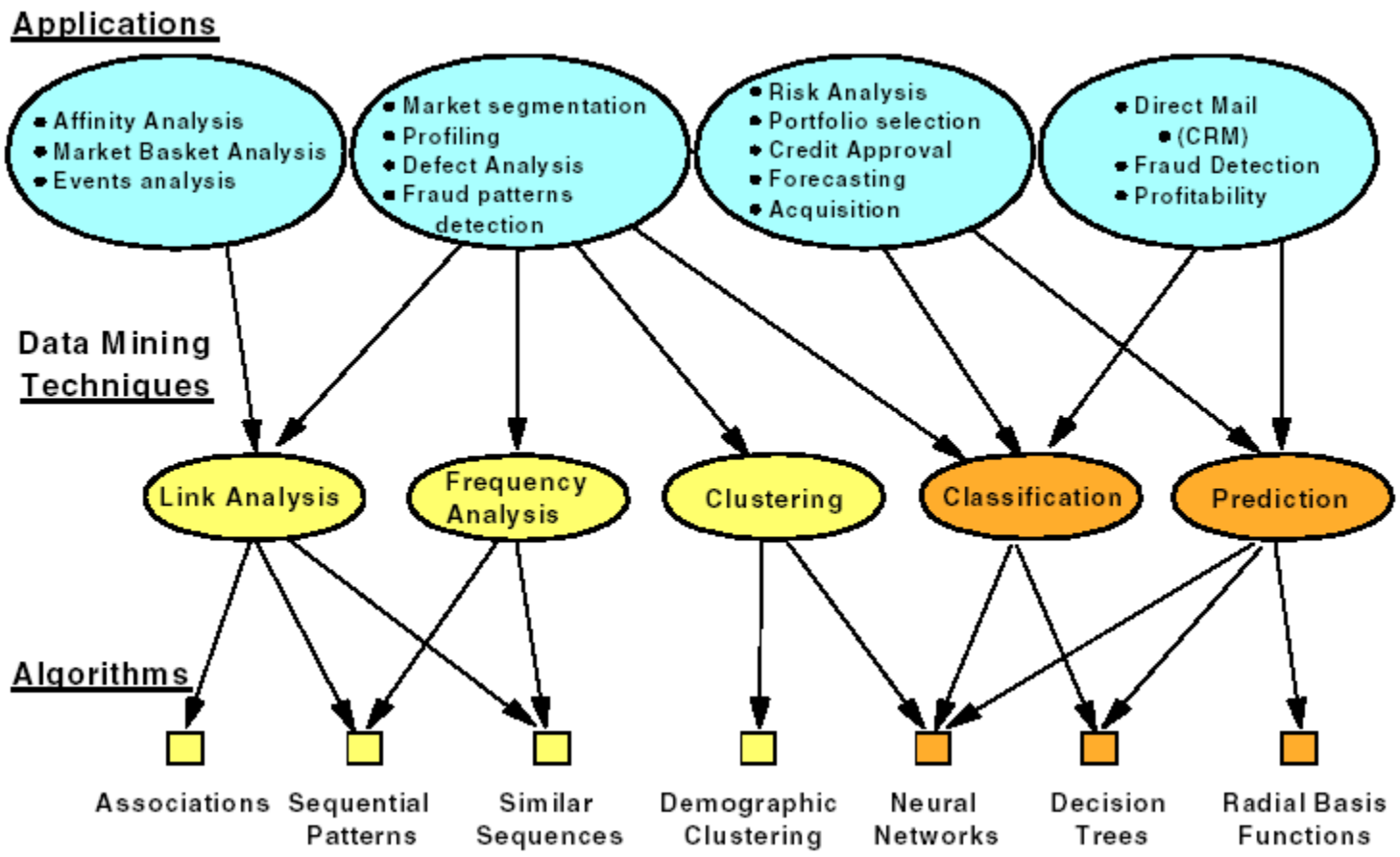
聚类示意

- 基于欧氏距离的三维空间中的聚类

$$d(i,j)=\sqrt{(|x_i-x_j|^2+|y_i-y_j|^2+|z_i-z_j|^2)}$$



从算法到应用



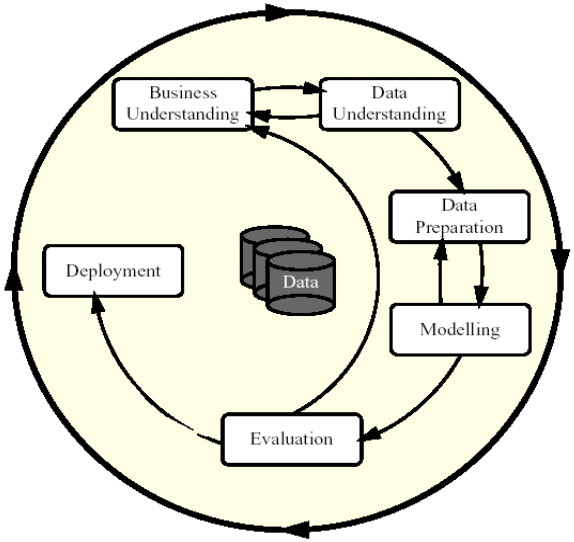
Recommend
火龙果软件

- 数据挖掘的基本概念
- 数据挖掘的基本算法
- **数据挖掘实施方法论**
- 总结与讨论

Recommend
火龙果软件

数据挖掘过程模型（DM Process Model）

- 用以管理并指导Data Miner有效、准确开展数据挖掘工作以期获得最佳挖掘成果的一系列工作步骤的规范标准。
  - 由厂商提出的
    - ✓ SPSS的5‘A（Assess, Access, Analysis, Act, Automat）
    - ✓ SAS的SEMMA（Sample, Explore, Modify, Model, Assess）
    - ✓ MICROSOFI的OLEDB for DM
  - 由行业组织提出的
    - CRISP-DM（CRoss Industry Standard Process for DM）

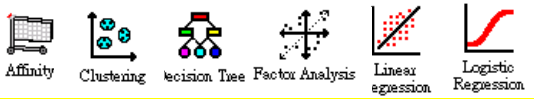


Recommend
火龙果软件

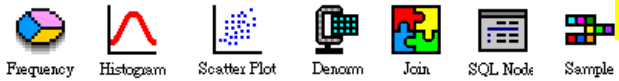
TERATATA的挖掘方法论

Scope Business Problem 度量数称胜

Deploy工欲善其事，必先利其器



Model 奇正之变，不可胜穷也



Explore Business Facts in DB (EDA) 望闻问切

Select & Sample 必也正名乎

Clean Data 谨防假做真时真亦假，无为有处有还无

Teradata Data Warehouse

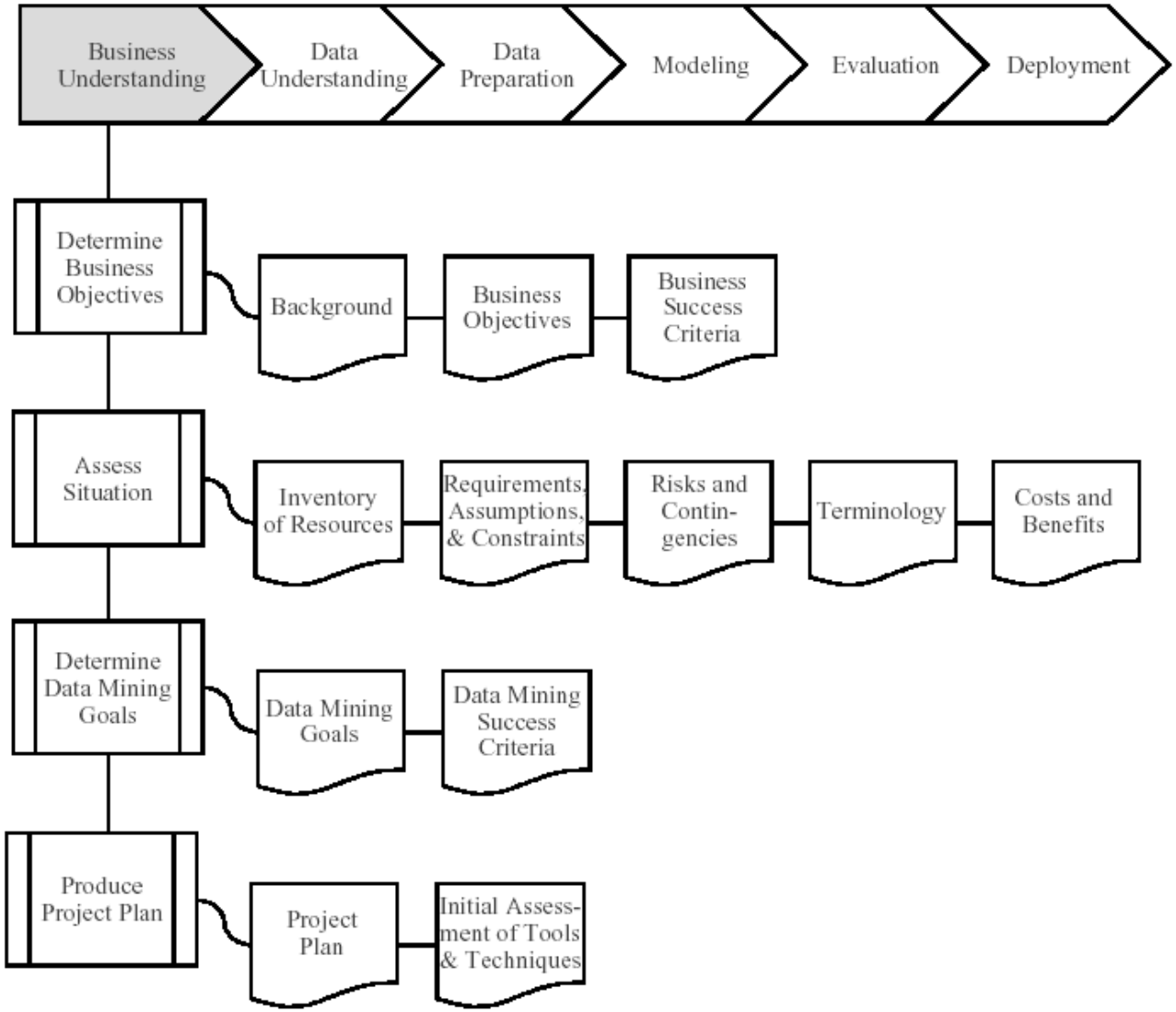
Recommend
火龙果软件

CRISP-DM中的元任务（Generic Tasks）

Business Understanding	Data Understanding	Data Preparation	Modeling	Evaluation	Deployment
Determine Business Objectives *Background* *Business Objectives* *Business Success Criteria* **Assess Situation** *Inventory of Resources* *Requirements, Assumptions, and Constraints* *Risks and Contingencies* *Terminology* *Costs and Benefits* **Determine Data Mining Goals** *Data Mining Goals* *Data Mining Success Criteria* **Produce Project Plan** *Project Plan* *Initial Assessment of Tools and Techniques*	**Collect Initial Data** *Initial Data Collection Report* **Describe Data** *Data Description Report* **Explore Data** *Data Exploration Report* **Verify Data Quality** *Data Quality Report*	*Data Set* *Data Set Description* **Select Data** *Rationale for Inclusion / Exclusion* **Clean Data** *Data Cleaning Report* **Construct Data** *Derived Attributes* *Generated Records* **Integrate Data** *Merged Data* **Format Data** *Reformatted Data*	**Select Modeling Technique** *Modeling Technique* *Modeling Assumptions* **Generate Test Design** *Test Design* **Build Model** *Parameter Settings* *Models* *Model Description* **Assess Model** *Model Assessment* *Revised Parameter Settings*	**Evaluate Results** *Assessment of Data Mining Results w.r.t. Business Success Criteria* *Approved Models* **Review Process** *Review of Process* **Determine Next Steps** *List of Possible Actions* *Decision*	**Plan Deployment** *Deployment Plan* **Plan Monitoring and Maintenance** *Monitoring and Maintenance Plan* **Produce Final Report** *Final Report* *Final Presentation* **Review Project** *Experience Documentation*

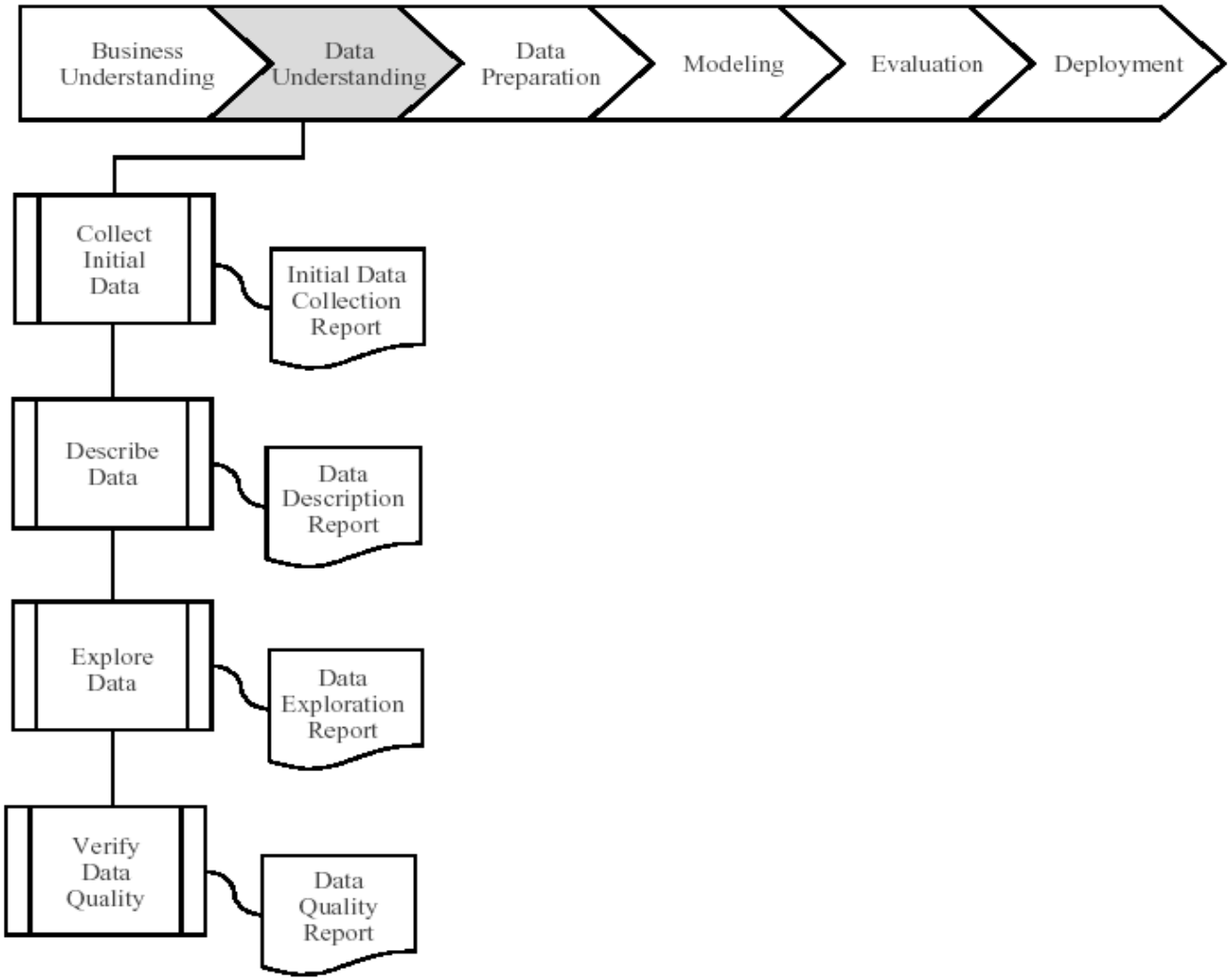
Recommend
火龙果软件

CRISP-DM中的业务理解



Recommend
火龙果软件

CRISP-DM中的数据理解



Recommend
火龙果软件

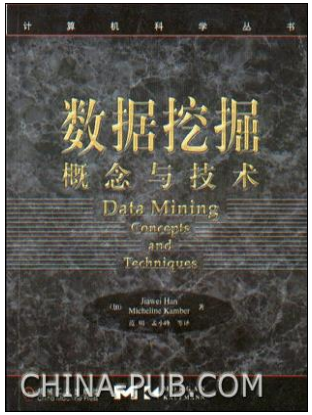
- 数据挖掘的基本概念
- 数据挖掘的基本算法
- 数据挖掘实施方法论
- 总结与讨论

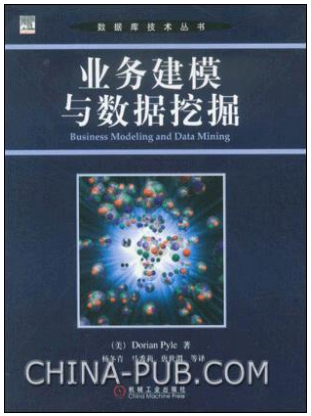
Recommend
火龙果软件

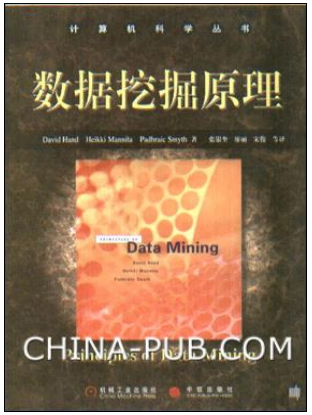
总结

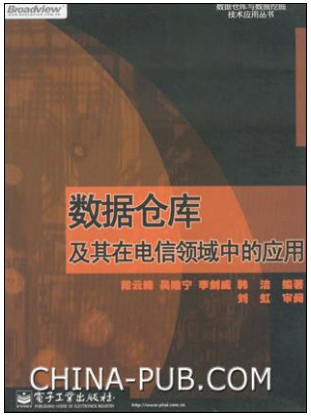
- 数据挖掘的思想也许比我们想象的更加深邃
- 数据挖掘的算法太多了，但用到一种，就要透彻把握一种
- 数据挖掘是一个循环的探索过程，各个环节同等重要
- 实施数据挖掘是一个战略性的举措
- 数据挖掘不是万能的，没有数据挖掘也未必是万万不能的

参考文献









Recommend
火龙果软件

Thank you !

Recommend
火龙果软件