数据分析与数据挖掘

第1课 概述

兰曼

计算机系

华东师范大学

参考书目

- **Python数据分析与挖掘实战**
  张良均等　机械工业出版社
- **利用Python进行数据分析(Python for Data Analysis)**
  唐学韬 等译　机械工业出版社
- **数据挖掘：概念与技术（第三版）**
  范明 孟小峰 等译　机械工业出版社
- **Data Mining : Concepts and Techniques (3rd Edition)**
  Jiawei Han, Micheline Kamber, Jian Pei

其他参考书目

- 机器学习实战 **(Machine Learning in Action)** 人民邮电出版社
- Python自然语言处理 (Natural Language Processing with Python)
- 社交网站的数据挖掘与分析 (Mining the Social Web)
- **Python** 金融大数据分析 (Python for Finance: Analyze Big Financial Data)

- 其他相关书本，文献，网络资源等

课程的主要内容

1. 概述和环境配置
2. 了解数据的基本概念
3. 数据分析的技术
4. 数据挖掘的算法和应用
5. 多领域的实例应用

1. 概述

- 为什么数据分析和数据挖掘？
- 为什么Python？
- 什么类型的数据可以被分析和挖掘？
- 功能
- 应用
- 关于这门课程

为什么数据分析和数据挖掘？

- “**Necessity is the mother of invention**” — Plato
- 数据的爆炸增长: 从TB 增长到PB
  - 数据收集和获得
    - 数据自动收集工具，数据库系统，网络，社会计算
  - 大量数据的主要来源
    - 商业：网络，电子商务，交易，股票…
    - 科学: 远程遥感，生物信息，科学模拟 …
    - 社会和个人: 新闻，数码照片，Twitter, YouTube, …

为什么数据分析和数据挖掘？

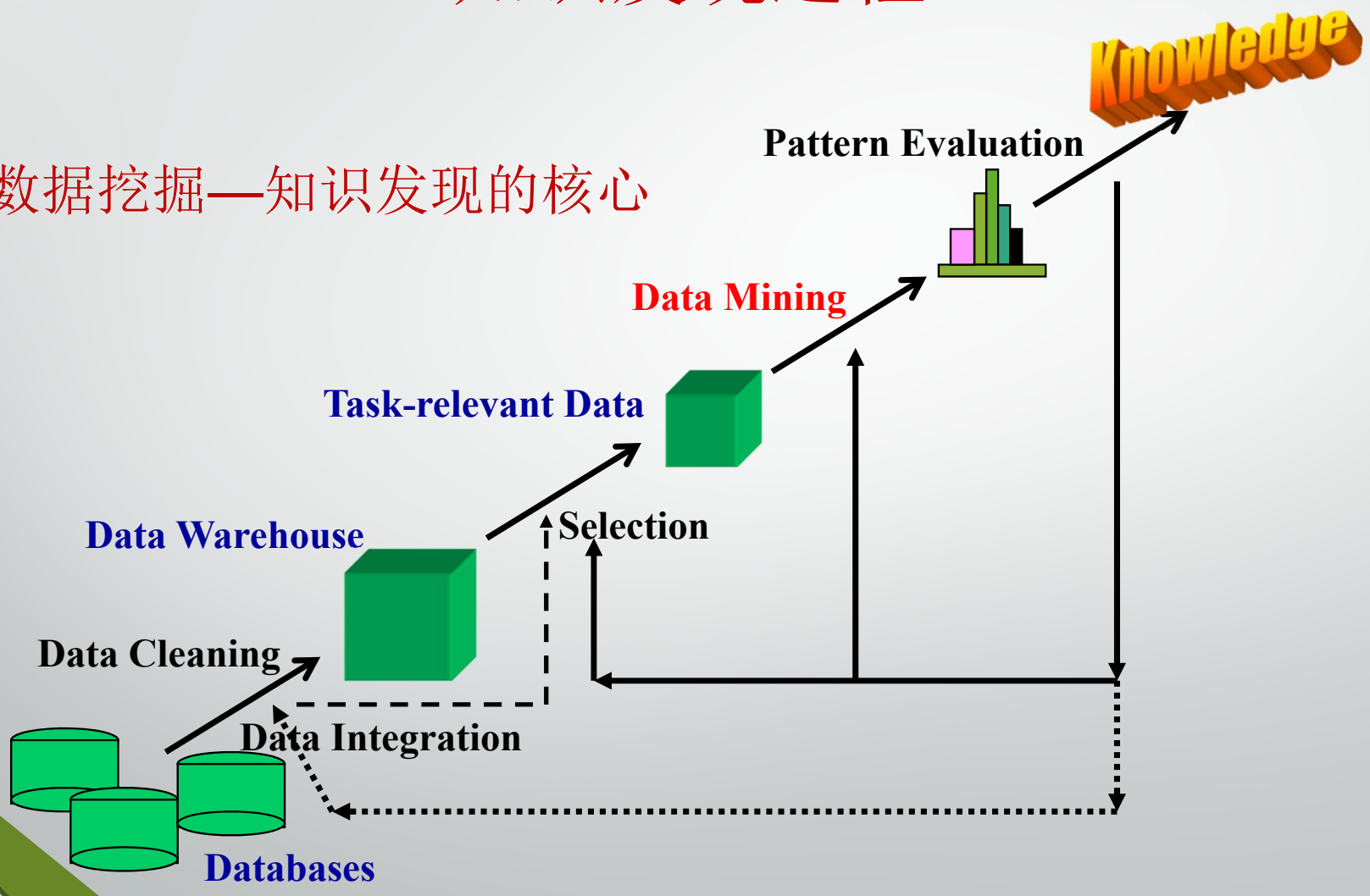
- 我们湮没在数据中，却渴望知识！
- 解决途径：数据分析和数据挖掘
- 数据分析 – 知己知彼
  - 了解数据，深入理解数据
  - 手段：了解数据分析的概念方法，掌握Python对多类型数据的分析技术
- 数据挖掘 – 运筹帷幄
  - 在了解数据的基础上，运用数据，预测未知未来
  - 手段：理解数据挖掘的概念原理，掌握数据挖掘在多领域中的应用实践

什么是数据挖掘？

- 数据挖掘 (知识发现)
  - 从大量的数据中发现知识 (非零碎的, 隐晦的, 事先不知道的和潜在有用的)
  - 知识（规则、规律、约束、模式等）(rules, regularities, patterns, constraints)
  - 数据挖掘: 一个不恰当的名字？
- 别名
  - 知识发现 (KDD), 知识抽取，数据/模式分析,数据考古, 信息获取，商业智能，等.
- 注意: 所有的东西都可以“数据挖掘”？
  - (推论) 简单的搜索和查询过程
  - 专业系统或者小型的机器学习/统计程序

知识发现过程

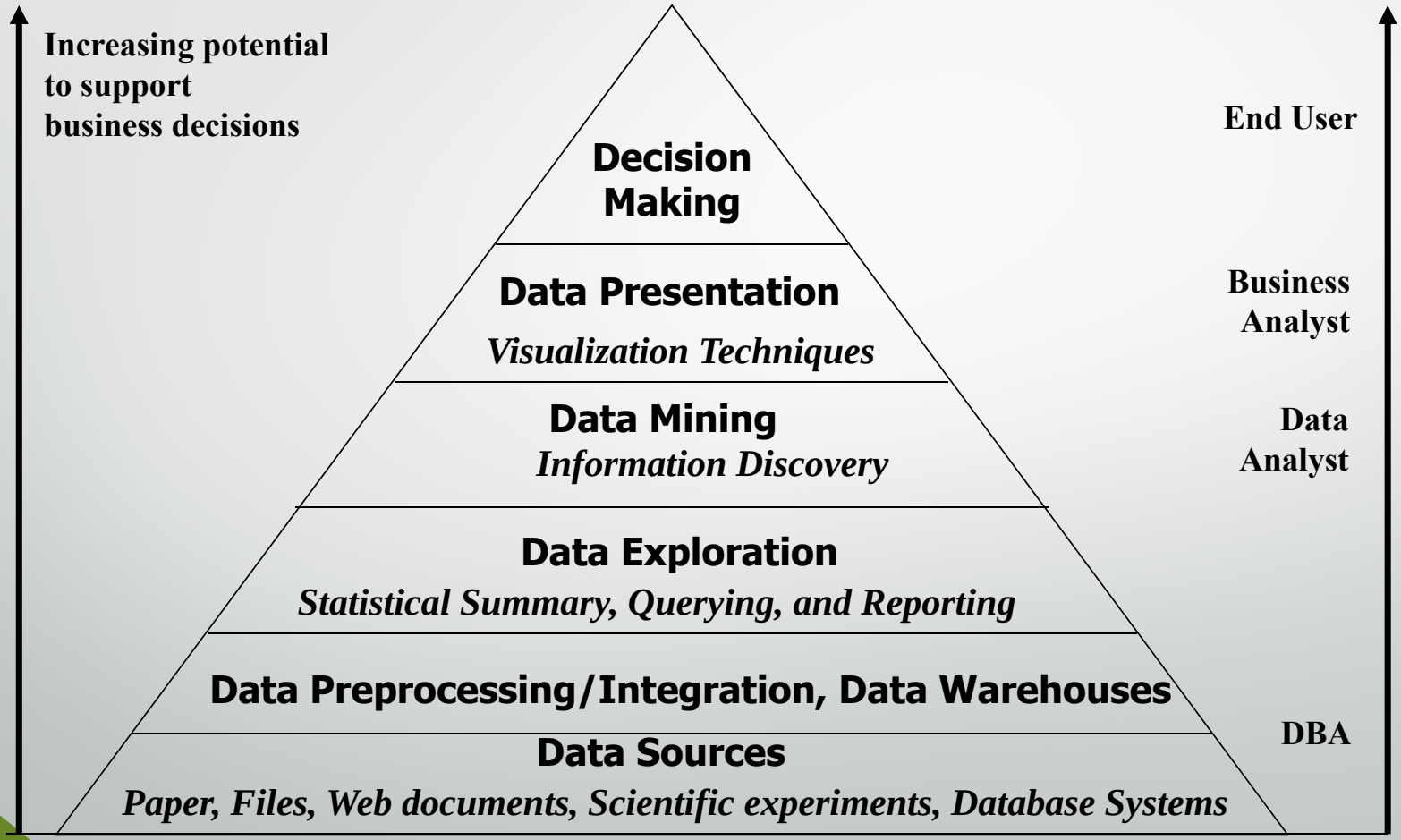
数据挖掘—知识发现的核心



知识发现过程中的几个关键步骤

- 学习应用领域：相关的先验知识和应用目标
- 创建目标数据集：数据选择
- 数据清理和预处理：（可能需要60％以上的精力！）
- 数据简化和转换
  - 查找有用的功能，缩减维度/变量，不变的表示
- 选择数据挖掘的功能
  - 归纳，分类，回归，关联，聚类
- 选择挖掘算法
- 数据挖掘：搜索感兴趣的模式
- 模式评估和知识演示
  - 可视化，转换，删除冗余模式等
- 使用发现的知识

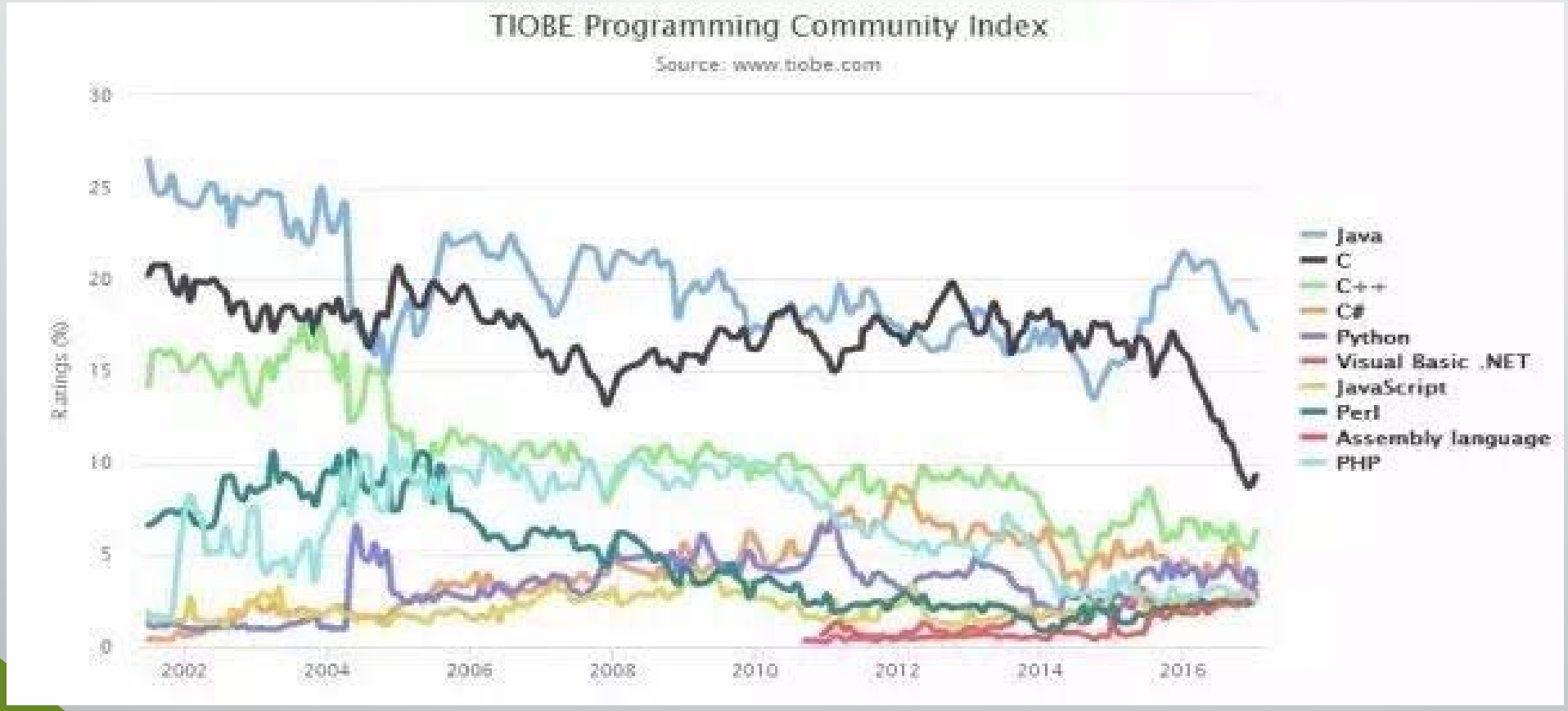
商业智能中的数据挖掘


Increasing potential
to support
business decisions
Decision
Making
Data Presentation
Visualization Techniques
Data Mining
Information Discovery
Data Exploration
Statistical Summary, Querying, and Reporting
Data Preprocessing/Integration, Data Warehouses
Data Sources
Paper, Files, Web documents, Scientific experiments, Database Systems
End User
Business
Analyst
Data
Analyst
DBA

1. 概述

- 为什么数据分析和数据挖掘？
- 为什么Python？
- 什么类型的数据可以被分析和挖掘？
- 功能
- 应用
- 关于这门课程

TIOBE Programming Trend (Jan. 2017)



为什么Python？

- 简单优雅
- 完善的基础代码库
- 丰富的第三方库
- 网络应用（网站、后台服务等）；日常需要的工具（包括管理员需要的脚本任务等）；把其他语言开发的程序再包装起来方便使用等
- *然而…*
- 运行速度慢
- 代码不能加密

1. 概述

- 为什么数据分析和数据挖掘？
- 为什么Python？
- 什么类型的数据可以被分析和挖掘？
- 功能
- 应用
- 关于这门课程

什么类型的数据?

- 面向数据库的数据集以及应用
  - 关联数据库, 数据仓库, 事务型数据库
- 高级数据集和高级的应用程序
  - 数据流和数据传感器
  - 时间序列数据，时间数据，序列数据, (包括生物序列)
  - 结构化数据,图,社交网络和多连接数据
  - 面向对象数据库
  - 异构数据库和遗留数据库
  - 空间数据和时空数据
  - 多媒体数据库
  - 文本数据库
  - 万维网

1. 概述

- 为什么数据分析和数据挖掘？
- 为什么Python？
- 什么类型的数据可以被分析和挖掘？
- 功能
- 应用
- 关于这门课程

(1) 关联和相关

- 频繁模式(或频繁项集)
  - 在超市购物中，哪些商品是经常被一起购买的？
- 关联规则挖掘:
  - 在事务数据库，关系数据库和其他信息库中或对象集合中寻找频繁模式，关联，相关性或因果结构。
  - 关联规则形式：前件 → 后件 [支持度，可信度]
    - 尿片→ 啤酒 [0.5%, 75%]
    - 买(x, “尿片”) → 买(x, “啤酒”) [0.5%, 60%]
    - 专业(x, “CS”) ^ 选修(x, “DB”) → 积点分(x, “A”) [1%, 75%]

(1) 关联和相关分析

- 关联，相关与因果关系比较:
  - 强关联的两个项是否也强相关？
  - 如何在大型数据集中挖掘这样的模式和规则？
  - 如何在分类，聚类或者其他应用使用这类模式？
- 应用:
  - 篮子数据分析，交叉营销，目录设计，亏损诱导策略，聚类，分类等

(2) 分类

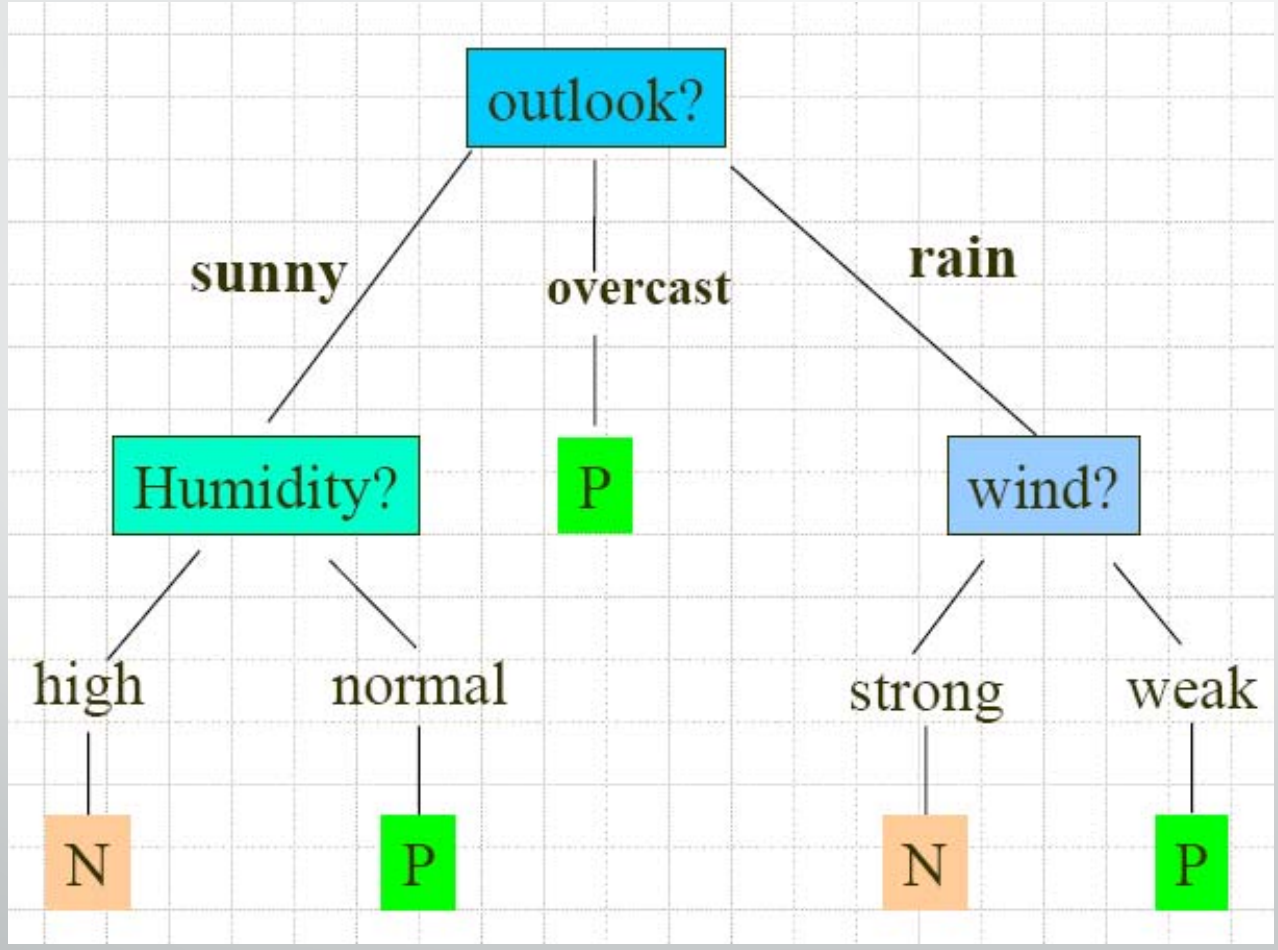
- 分类以及类标预测：
  - 根据训练样本构建模型
  - 描述和区分类别以便将来预测
    - 例如：基于兴趣将客户进行分类
  - 预测未知样本的类标
- 算法：
  - 决策树，朴素贝叶斯，k-最近邻，支持向量机，神经网络…
- 应用:
  - 信用卡欺诈检测，客户分类，网页分类，垃圾邮件过滤，图像识别 …

(2) 分类（训练数据集）

This follows an example from Quinlan's ID3

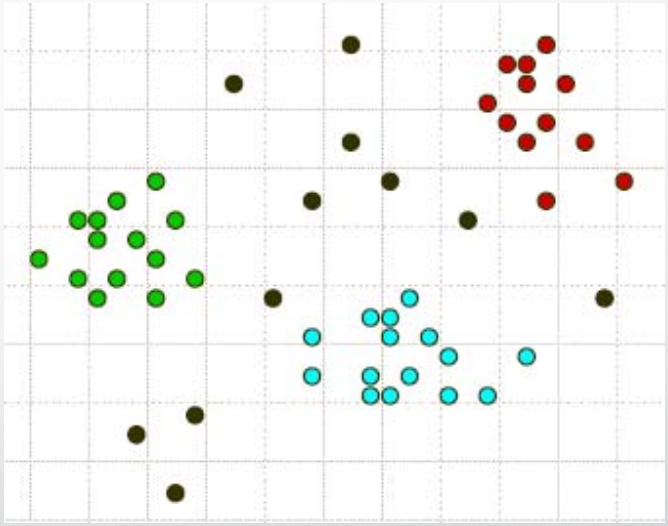
Outlook	Temperature	Humidity	Windy	Class
sunny	hot	high	false	N
sunny	hot	high	true	N
overcast	hot	high	false	P
rain	mild	high	false	P
rain	cool	normal	false	P
rain	cool	normal	true	N
overcast	cool	normal	true	P
sunny	mild	high	false	N
sunny	cool	normal	false	P
rain	mild	normal	false	P
sunny	mild	normal	true	P
overcast	mild	high	true	P
overcast	hot	normal	false	P
rain	mild	high	true	N

(2) 分类（决策树模型）


outlook?
sunny
overcast
rain
Humidity?
P
wind?
high
normal
strong
weak
N
P
N
P

(3) 聚类分析

- 聚类: 数据对象的集合
  - 在同一个集群中的对象彼此相似
  - 不同集群中的对象不同
- 聚类分析
  - 将一组数据对象分组到集群中

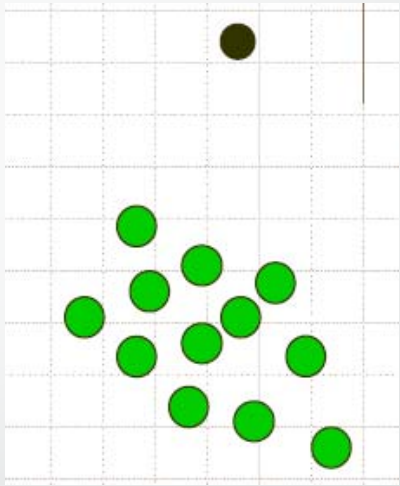


(3) 聚类分析

- 无监督学习（即，类标未知）
- 将数据分组形成新的类别（即，聚类）
- 原则：最大化类内相似度和最小化类间相似性
- 许多方法以及应用

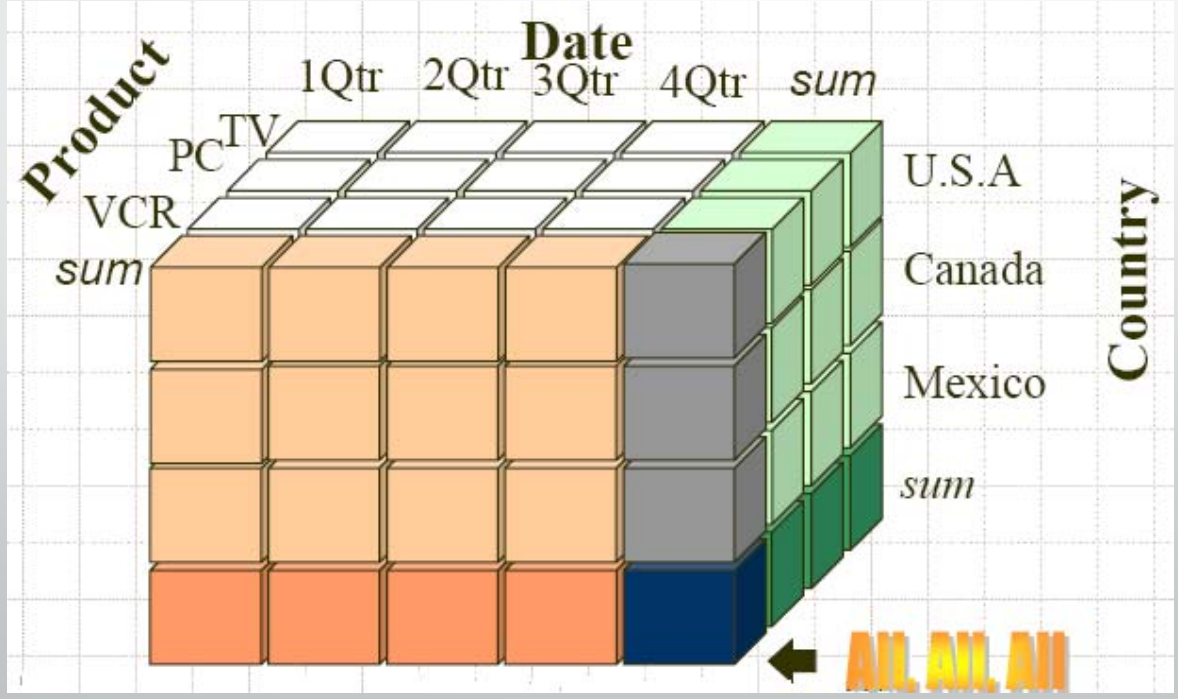
(4) 异常值检测

- 什么是异常值？
  - 不符合数据的一般行为的数据对象
  - 例如：运动：Michael Jordon, ...
  - 噪音还是异常？— One person's garbage could be another person's treasure
- 方法：远离其他集群的集群和对象是异常值；通过聚类或回归分析的乘积来判断…
- 寻找没有聚类的异常值是一个挑战
- 可以用于欺诈检测，罕见事件分析



(5) 数据立方体

- 提供数据的多维视图，以便于数据分析
- 例如：销售量是产品(product)，时间(Date), 和国家(Country)的函数

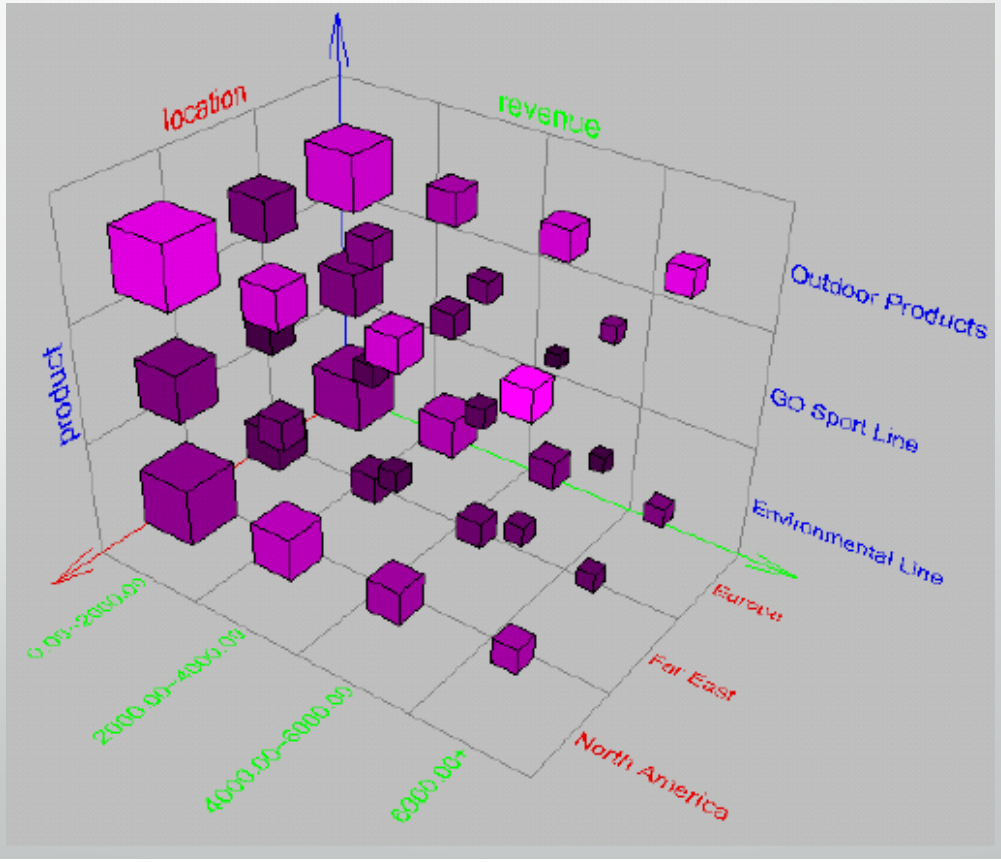


(6) 可视化

一张图片胜过千言万语

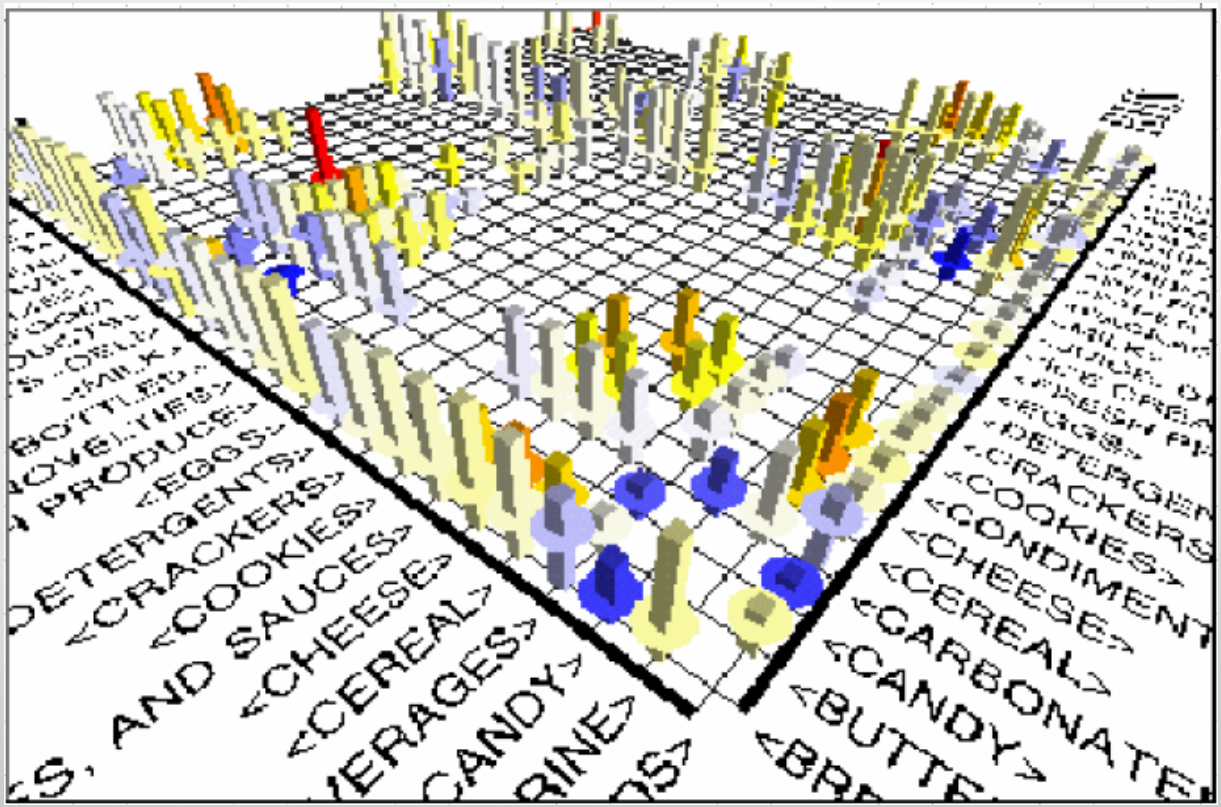
One picture may worth 1000 words

数据方体的
可视化



(6) 可视化

关联规则的
可视化



1. 概述

- 为什么数据分析和数据挖掘？
- 为什么Python？
- 什么类型的数据可以被分析和挖掘？
- 功能
- 应用
- 关于这门课程

客户关系管理

- 客户关系管理是新型经济的一个重要的方面
- 给定一个客户，我们是否可以不询问他而直接推荐他感兴趣的产品？
- 我们是否能满足他/她的需求？
- 我们可以为每个客户提供个性化的折扣方案吗？

Web分析

- 网页内容：
  - 搜索：网络搜索引擎 Google
  - 网页聚类：从网络中寻找相似的网页
- 网络日志
  - 如果客户访问页面A和页面B，则他/她很可能转到页面C，然后购买产品E
  - 序列聚类，找到具有非常相似的页访问序列的客户组

安全领域

- 网络安全
  - 用于检测网络入侵
  - 分析所发布的命令
  - 分析网络流量
- 监视器
  - 分析房间或飞机上的异常运动
  - 在X射线扫描期间快速检测武器

地理数据

- 在地图上显示的数据
- 天气预报
- 寻找污染源
- 分析犯罪模式
- 位置规划
- 流量分析

性能优化

- 缓存预取
- 使用数据挖掘的语义压缩
- 通过聚类更好的索引
- 供应链管理

生物信息学

- DNA序列分析：
  - 索引
  - 聚类
  - 压缩
- 基因表达分析
  - 函数预测
  - 可视化
  - 聚类

其他应用

- 体育
  - 分析球员和球队的战略
- 天文学
  - JPL和Palomar天文台在数据挖掘的帮助下发现了22个类星体
- 互联网使用援助
  - IBM Surf-Aid将数据挖掘算法应用于市场相关页面的Web访问日志，以发现客户偏好和行为页面，分析Web营销的有效性，改进网站组织等。

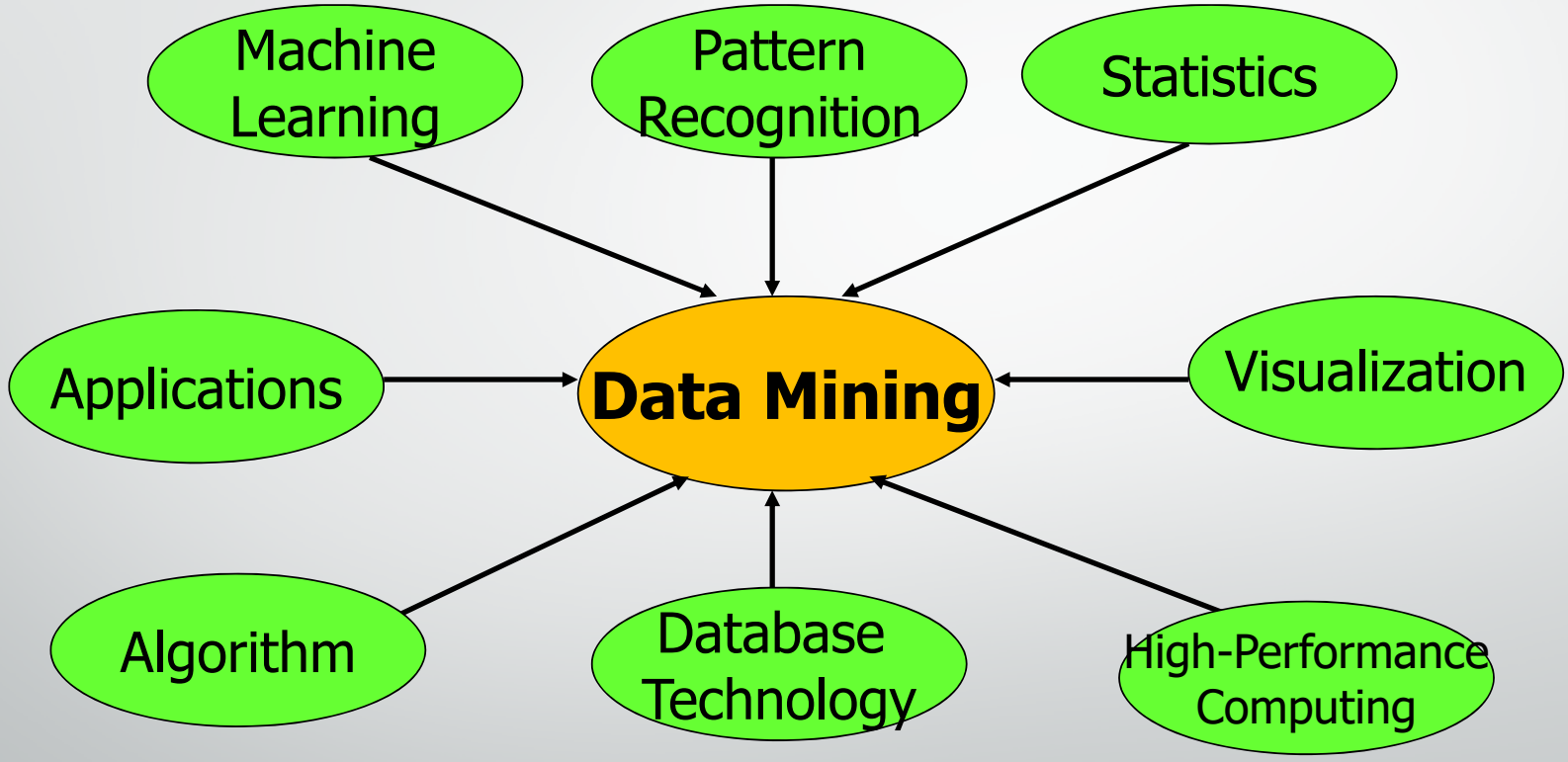
为什么不使用传统数据分析？

- 数据量巨大
  - 算法必须能够具有很好的扩展性以便处理T量级的数据
- 高维数据
  - 基因序列Micro-array具有十几甚至几十万维的维度
- 高复杂性的数据
  - 数据流，时间序列数据
  - 结构数据，图，社交网络，空间、时空、多媒体、文本和网络数据
  - 异构数据库，异源数据
  - 软件程序，科学实验数据
- 不断有新的和复杂的应用程序

1. 概述

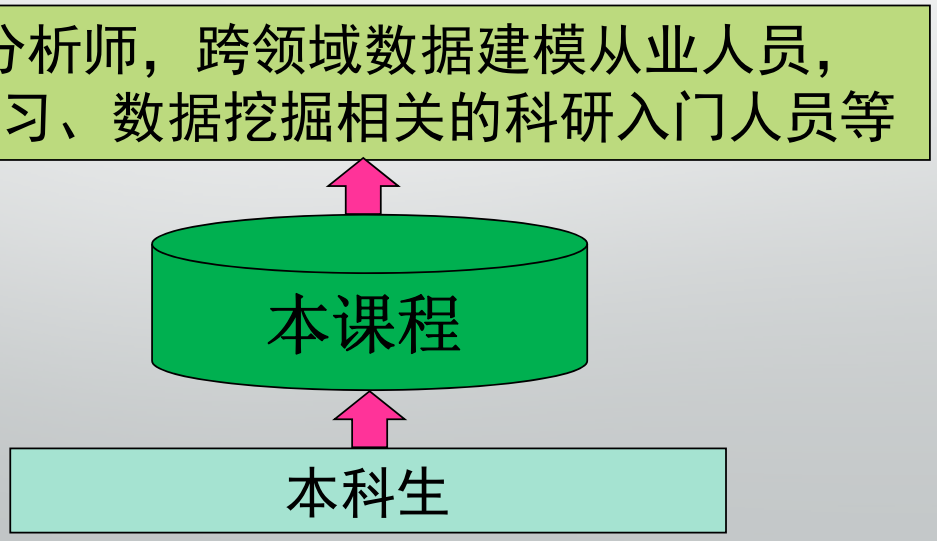
- 为什么数据分析和数据挖掘？
- 为什么Python？
- 什么类型的数据可以被分析和挖掘？
- 功能
- 应用
- 关于这门课程

多学科的交融


Machine Learning
Pattern Recognition
Statistics
Applications
Data Mining
Visualization
Algorithm
Database Technology
High-Performance Computing

适用人群

- 对数据分析和数据挖掘相关技术有兴趣
- 想具有跨领域/跨行业的数据分析和建模的从业人员
- 想转行/领域从事数据分析师行业的学习者
- 想使用Python实现机器学习和数据挖掘的学习者
- 尚不会使用Python的数据分析从业者



课程评估分布（1）

• 个人作业:	20 分 * 3	60 分
• 小组项目作业:	40 分	40 分
• **总共:**		**100 分**

课程评估分布（2）

- 小组项目作业：**40分**
  - 每个小组给出相同的训练数据，可以使用任何工具和软件及学习的技术构建模型
  - 每个小组给出不同的测试数据（符合同一分布）进行预测
  - 3 - 4 人/每组
  - 分数分布
    - 系统准确性: 20 (性能最好的系统获得满分20分)
    - 书面报告: 10 (pdf 或者doc格式均可)
    - 口头报告: 10 (口头汇报+PPT)
  - 2017年3月31日 – 发布任务和训练数据
  - 2017年5月7日 – 发布测试数据
  - 2017年5月21日 – 提交结果；2017年5月28日 – 提交书面报告
  - 2017年6月 – 小组口头汇报