



# 大数据导论

# Introduction to Big Data



## 第1讲：绪论

叶允明  
计算机科学与技术学院  
哈尔滨工业大学（深圳）

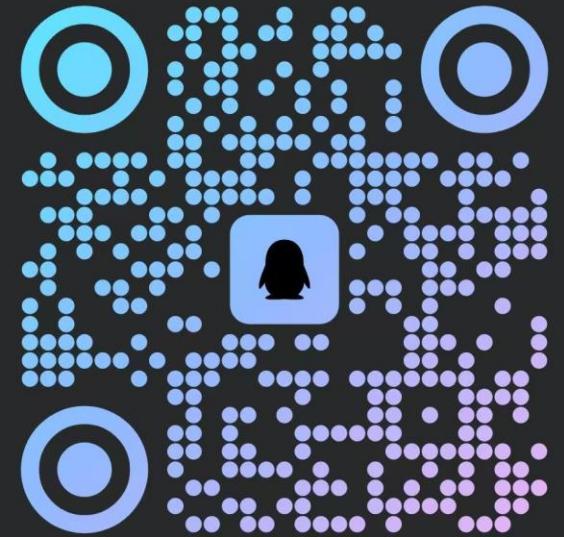
# 课程交流群

- QQ课程群
- Group number : 1059721420
- Group name : 2025大数据导论



2025大数据导论

群号: 1059721420



扫一扫二维码，加入群聊



QQ

# 助教



- 于哲浩
- Email:  
[25s151060@stu.hit.edu.cn](mailto:25s151060@stu.hit.edu.cn)
- Tel: 13292856828



- 周奇凤
- Email:  
[25S051012@stu.hit.edu.cn](mailto:25S051012@stu.hit.edu.cn)
- Tel: 13618009410

# 关于这门课程的学习目标

# 课程参考资料

- 教案与论文
- 梅宏. 大数据导论. 高等教育出版社, 2018.11.
- 林子雨. 《大数据技术原理与应用(第2版)》. 人民邮电出版社, 2017.
- Pang-Ning Tan, Michael Steinbach, Anuj Karpatne , Vipin Kumar著; 段磊, 张天庆等译. 数据挖掘导论 (原书第2版) . 机械工业出版社, ISBN: 9787111631620, 2019-07-29.
- Jiawei Han, Micheline Kamber, Jian Pei著; 范明, 孟小峰等译. 数据挖掘: 概念与技术. 机械工业出版社, ISBN: 9787111391401, 2012.

# 课程内容

- 大数据存储与处理框架（Hadoop）
- 数据治理方法：数据理解与预处理方法
- 大数据的关联规则挖掘及其应用
- 大数据的分类与预测算法
- 大数据的聚类与离群点检测算法
- 图数据及其典型大数据分析算法

# 从该课程你能学到什么？

- 如何在实际应用中设计和实现大数据项目
  - 大数据项目作为一个过程或工作流的思想 (process or workflow)
- 经典大数据算法
  - 例如常用的Map-reduce算法、数据挖掘算法
- 大数据软件工具
  - 开源工具、软件产品

# 课程形式和要求

- 先修课程：高等数学、代数与几何、概率论与数理统计、高级语言程序设计（Java, Python基础）
- 授课 & 实验
- 最终成绩：
  - 30% 小作业
  - 30% 实验
  - 40% 大作业

# 第一讲：绪论

- 大数据的历史背景
- 大数据的应用领域
- 大数据技术概况
- 大数据领域的学习资源

# 大数据的历史与背景

# 大数据现象

- 人类社会数字化、信息化和网络化进程的快速发展
  - 带来了各行各业数据的爆炸性增长!



我国网民数量居世界之首，每天产生的数据量也位于世界前列。

- 淘宝网站
  - ◆ 单日数据产生量超过5万GB
  - ◆ 存储量4000万GB
- 百度公司
  - ◆ 目前数据总量10亿GB
  - ◆ 存储网页1万亿页
  - ◆ 每天大约要处理60亿次搜索请求
- 一个8Mbps的摄像头
  - ◆ 一小时能产生3.6GB的数据
  - ◆ 一个城市每月产生的数据达上千万GB
- 医院
  - ◆ 一个病人的CT影像数据量达几十GB
  - ◆ 全国每年需保存的数据达上百亿GB

# 大数据是什么

“3V” 定义



规模性 (Volume)



多样性 (Variety)



高速性 (Velocity)

维基百科给出的定义：

大数据是指利用常用软件工具捕获、管理和处理数据所耗时间超过可容忍时间的数据集。

“4V” 定义



价值性 (Value) (IDC)



真实性 (Veracity) (IBM)

# 大数据领域的发展历程

- 大规模数据的处理与分析技术已发展多年，一直是研究热点。但量变会引起质变！
- 2007年1月，图灵奖得主Jim Gray指出：科学的发展正在进入“**数据密集型科学发现范式**”——科学史上的**“第四范式”**
- 《自然》杂志2008年9月出版一个关于大数据的专刊。



# 大数据与人工智能

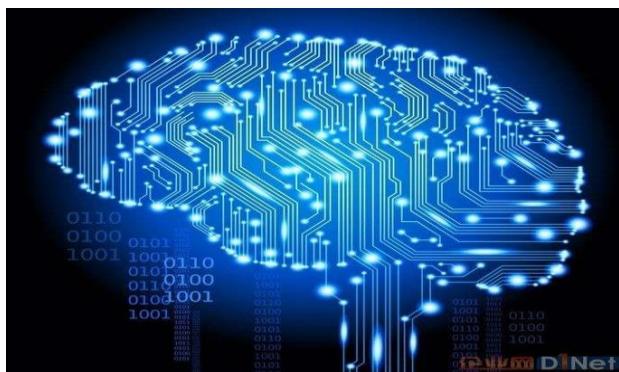
- 国务院：《新一代人工智能发展规划》，国发[2017]35号

## 专栏1 基础理论

1. 大数据智能理论。研究数据驱动与知识引导相结合的人工智能新方法、以自然语言理解和图像图形为核心的认知计算理论和方法、综合深度推理与创意人工智能理论与方法、非完全信息下智能决策基础理论与框架、数据驱动的通用人工智能数学模型与理论等。

- 目前最成功的人工智能应用领域：**大数据智能、大数据机器学习！**

➤ 深度学习需要大数据支撑！



# 大数据与数据挖掘

- 数据挖掘：从海量数据中发现“有趣的”的模式或知识  
( non-trivial, implicit, previously unknown and potentially useful )
- 1989 IJCAI Workshop on Knowledge Discovery in Databases
- 1991-1994 Workshops on Knowledge Discovery in Databases
- 1995-1998 International Conferences on Knowledge Discovery in Databases and Data Mining (KDD' 95-98)
  - Journal of Data Mining and Knowledge Discovery (1997)
- ACM SIGKDD conferences since 1998 and SIGKDD Explorations
- ACM Transactions on KDD starting in 2007
- 2008：“大数据”新的术语

# 大数据的应用领域

# 商业智能应用：决策支持

- 数据分析与决策支持
  - 市场分析与管理
    - ✓ 精准营销、客户关系管理(CRM)、购物篮分析、交叉销售、市场细分
  - 风险分析与管理
    - ✓ 预测（人、财、物）、客户维系、质量控制、竞争分析
  - 诈骗检测与异常模式发现

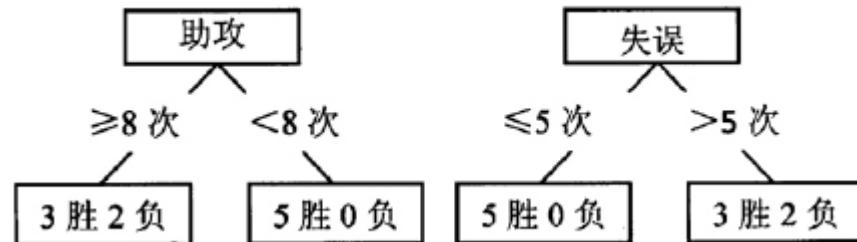
# 商业智能应用：推荐系统

- 应用领域：电商、信息推荐、电影、音乐等
- 目的：预测用户对商品是否喜欢、喜欢程度、个性化服务



# 体育应用：篮球针对性训练

- 对运动员成长轨迹进行深度挖掘、建模
- 找出运动员的“短板”与“长版”
- 加强对特长点和薄弱点的训练



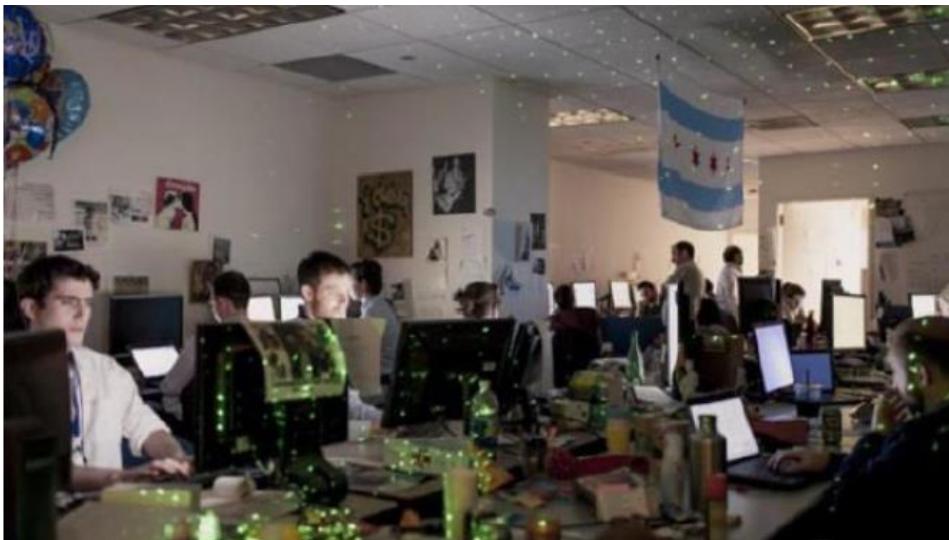
➤ 通过训练加强运动员助攻次数和减少失误率

避免此类情况发生



# 政治应用：美国总统大选

- 在总统候选人的第一次辩论之后，他们分析出哪些选民将倒戈，为每位选民找出一个最能说服他的理由
- 通过一些复杂的模型来精准定位不同选民，购买了一些冷门节目的广告时段，而没有采用在本地新闻时段购买广告的传统做法，广告效率相比2008年提高了14%
- 向奥巴马推荐，竞选后期应当在什么地方展开活动——那里有很多争取对象
- 借助模型帮助奥巴马筹集到创纪录的10亿美元



大数据团队

U.S. Edition ▾  
News Video TV Opinions More...  
U.S. World Politics Tech Health Entertainment Living Travel Money Sports  
New York City, NY 30° Sign In  
Search CNN

## How Obama's data crunchers helped him win

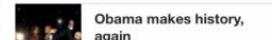
By Michael Scherer  
① Updated 11:45 AM ET, Thu November 8, 2012



President Obama's campaign manager hired an analytics department five times as large as that of the 2008 operation.

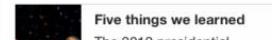


Full coverage: CNN's Election Center  
Get all the latest news at CNN's Election Center. There are race updates, a delegate counter and much more.



Obama makes history, again

A black man is returning to the White House. Four years ago, it was a first, the breaking of a racial barrier. Tuesday night, it was history redux. And more.



Five things we learned  
The 2012 presidential election shattered spending records, further polarized a

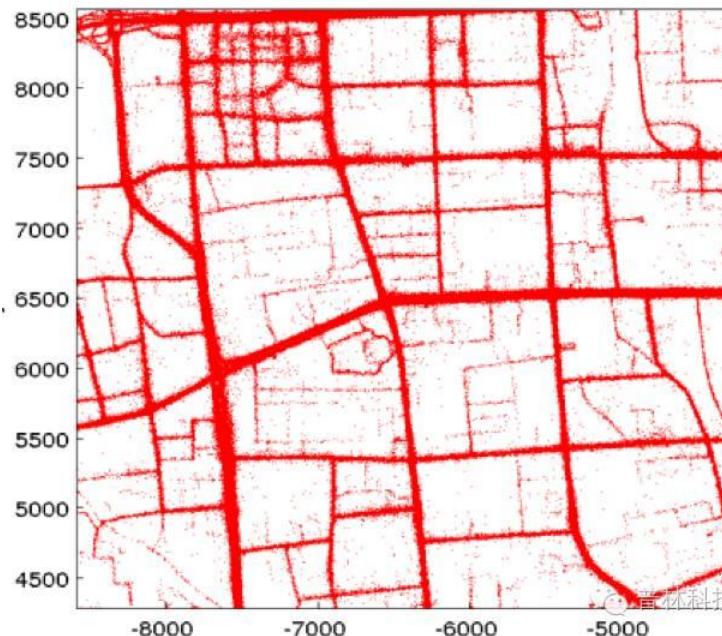
# 交通应用：拥堵预测

## ● 建立历史交通数据库

- 道路信息：从GIS数据库中导出北京市路网数据，包括道路的起点终点，中轴线经纬度，道路等级，车道数目等等。
- 车辆信息：数据来源是北京市6万辆出租车每天的GPS数据，出租车每50s生成一条GPS信息。

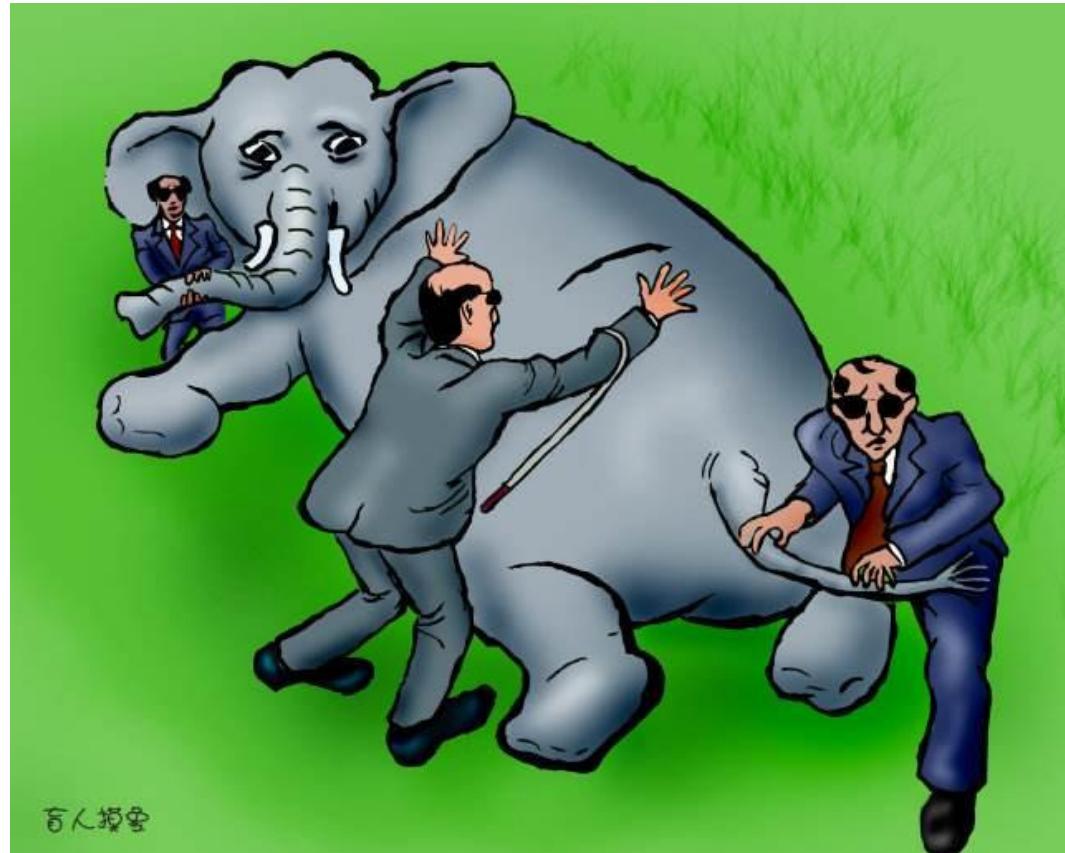
## ● 未来时间段车速预测

- 影响交通因素：天气状况，车辆数量，交通事故等，但车速可以包含以上信息。
- 根据历史数据中最相似的情况，从而进行预测，最相似的车速曲线，未来时刻的变化也可能相似。



# 大数据的核心问题

- 核心挑战：具有**多源、异构、信息碎片化、不确定性的特征**
- “关联”：发现**多源、异构的碎片化信息之间的关联关系**



图片引自[http://blog.sina.com.cn/s/blog\\_6773d7b90100jnsd.html](http://blog.sina.com.cn/s/blog_6773d7b90100jnsd.html)

# 大数据技术概况

# 大数据技术体系

- 数据采集、数据存储、数据处理、统计分析、智能挖掘、可视化



# 大数据统计分析技术

- 条件查询
  - SQL语言查询（或类SQL）
- 聚合统计
  - 按地区汇总销售量
  - 按时间维度汇总
- 复杂报表
- 多维度、多层次统计分析：联机分析处理（OLAP）

	江苏	上海	北京	汇总
电器	940	450	340	1730
服装	830	350	270	1450
汇总	1770	800	610	3180

主要技术挑战：海量数据的检索性能！

# 从统计分析到智能挖掘

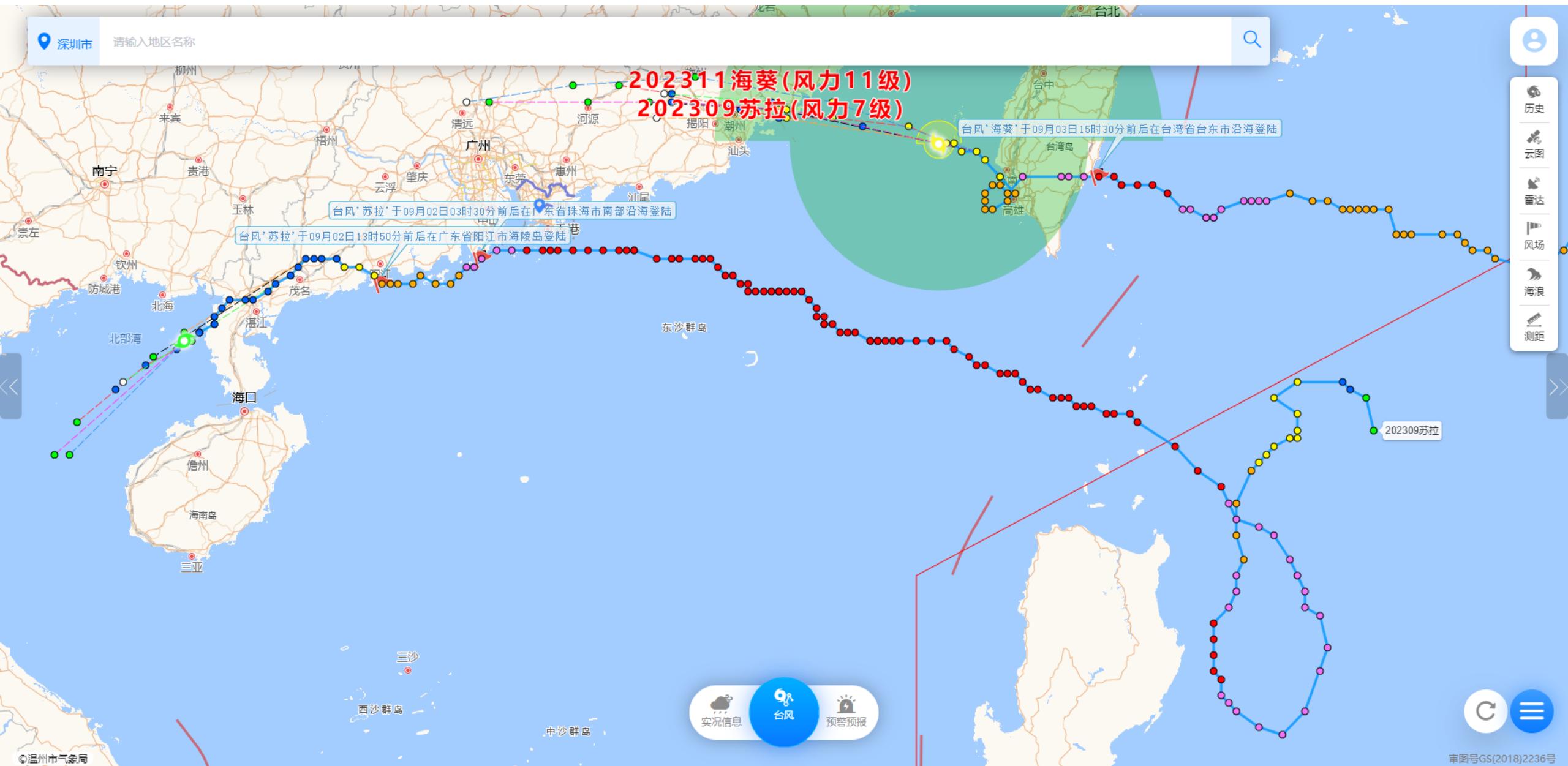
数据挖掘： Data Mining !



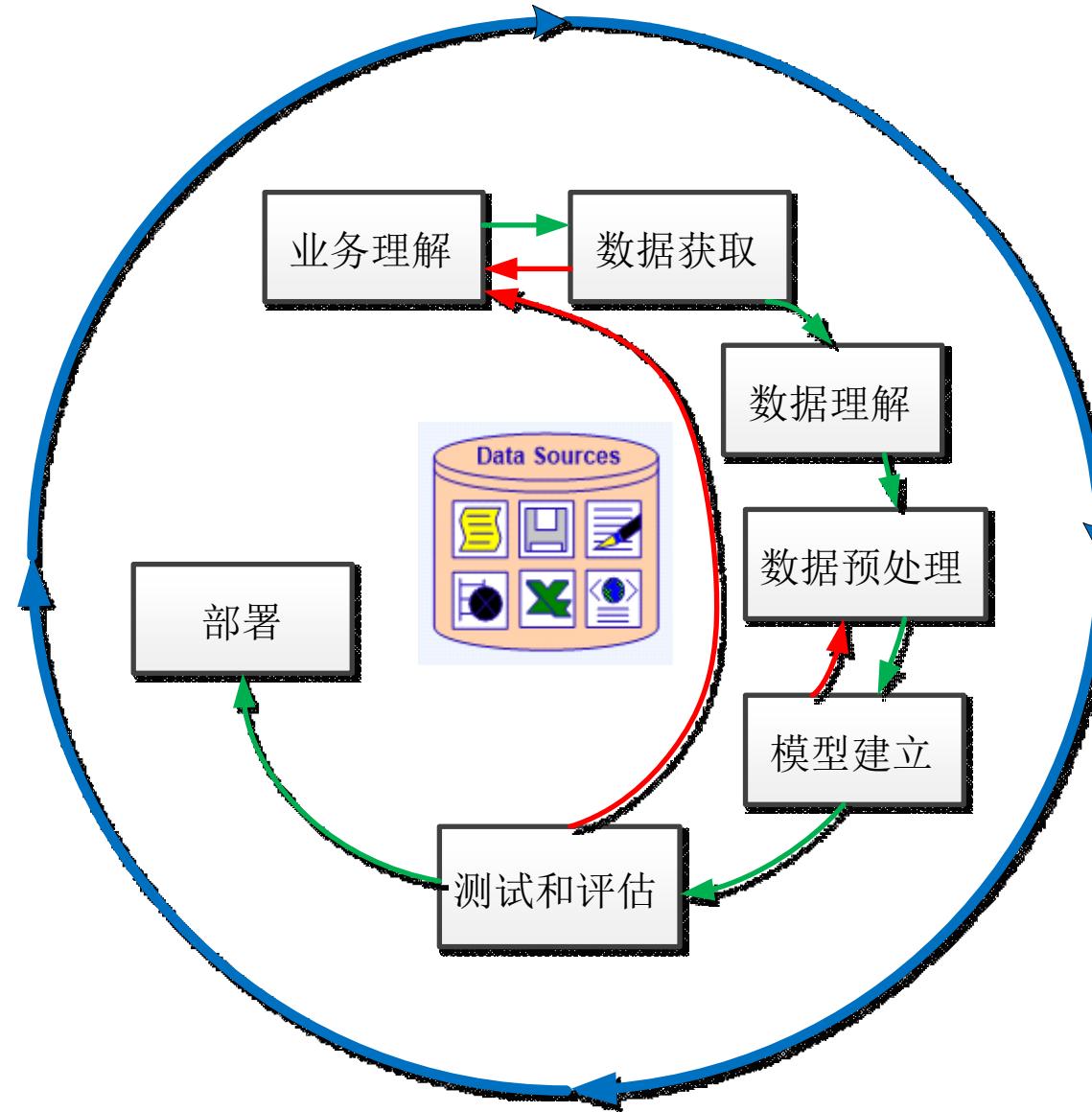
# 数据挖掘是什么？

- 数据挖掘（从数据中发现知识）
  - 从大量数据中提取有趣的（非平凡的，隐含的，以前未知的和潜在有用的）模式（pattern）或知识
- 替代名称
  - 数据库中的知识发现（Knowledge discovery in Databases, KDD）
  - 知识抽取（knowledge extraction）、模式挖掘（pattern mining）等
- 哪些数据处理和分析任务不是“数据挖掘”
  - 查询处理
  - 专家系统或小型ML /统计程序

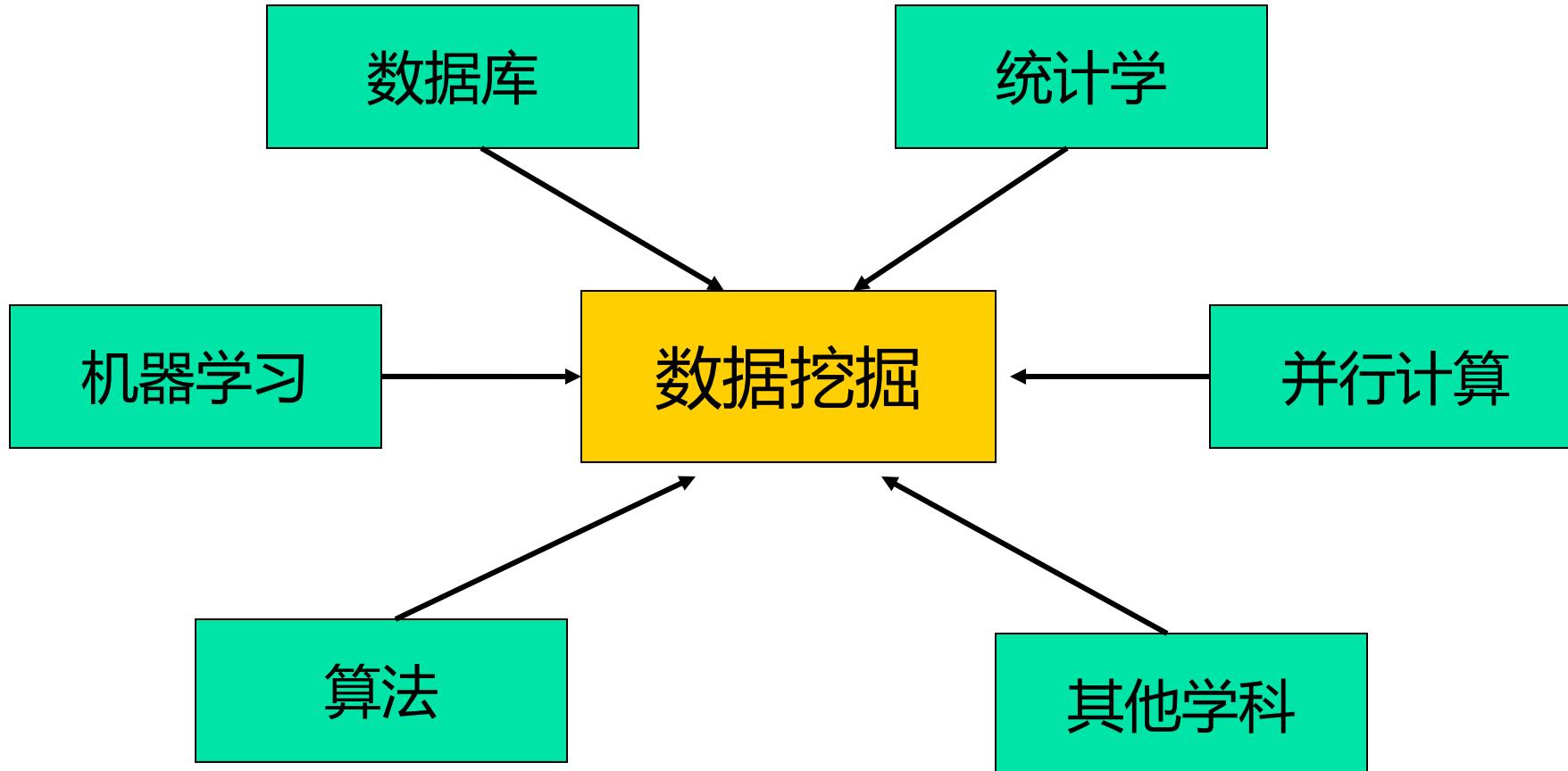
# 一个数据挖掘应用：台风预报



# 实际数据挖掘项目的过程模型



# 数据挖掘：多学科融合



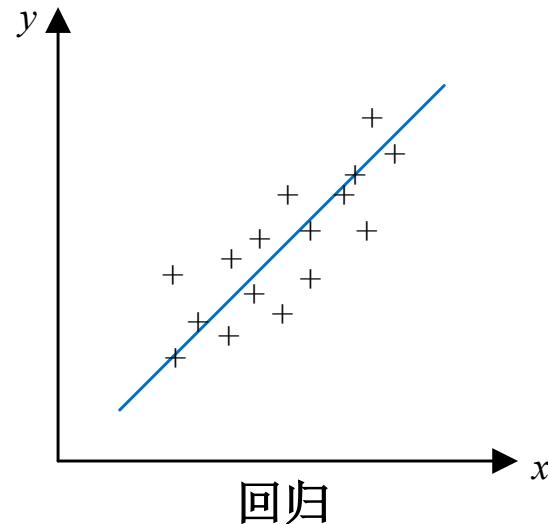
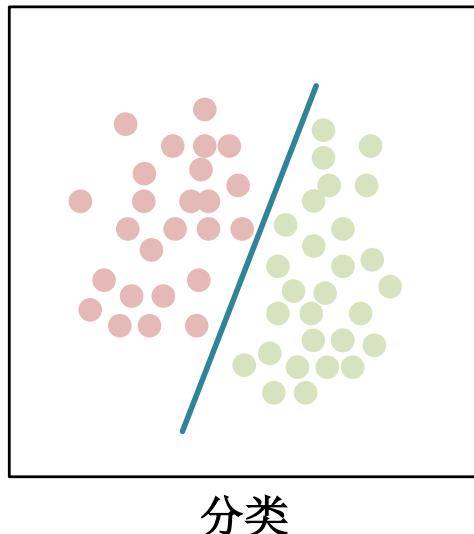
# 常见数据挖掘任务

- 多维概念描述：特征化概括和对比区分
- 关联规则挖掘
- 分类和回归预测
- 聚类分析与离群点检测
- 协同过滤推荐
- 趋势和演变分析
  - 子图模式挖掘、周期性分析

# 分类与回归

- 分类(classification): 预测给定数据对象的类别 (class, 离散值)
- 回归(regression): 预测给定数据对象对应的目标值 (连续值)

$$y = f(x), \quad \text{其中 } x \in D$$

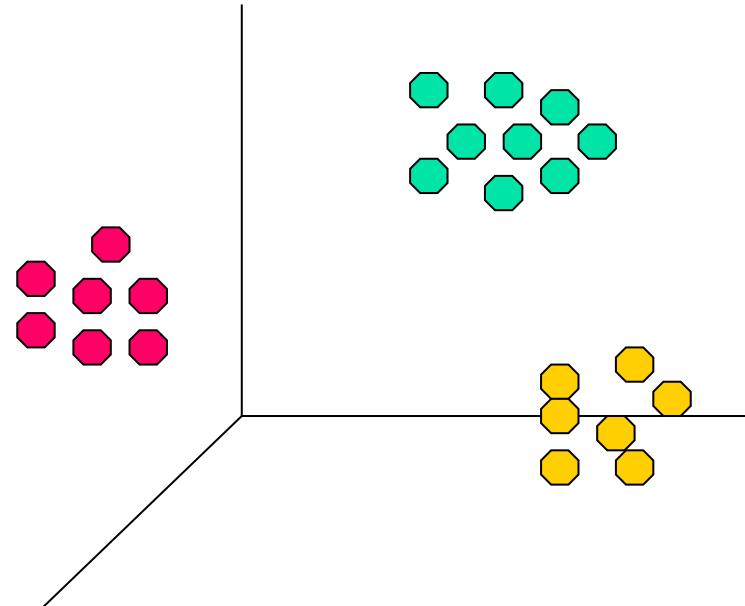


# 分类示例



# 聚类分析 (cluster analysis)

- 给定一个数据对象集合，以及数据对象之间的相似性度量，找到这样的一组簇 (cluster)：
  - 同一个簇中的数据点彼此更相似，不同簇中的数据对象彼此不太相似。



# 聚类分析示例

Baidu 百度

苹果

找到相关图片约6581张 版权 高清 最新 动图 1024x768 全部颜色

相关搜索：水果苹果 苹果壁纸高清 苹果壁纸 苹果Logo 苹果素描 苹果11图片 苹果手机 苹果12真实图片 苹果🍎 apple watch 苹果七 苹果八 苹果的照片 苹果所有型号手机 iPhone 8

The image shows a 5x8 grid of 40 photographs related to the search term 'apple'. The images include various types of apples (red, yellow, green), apple trees in different settings (outdoor orchards, indoor displays), and various Apple products (laptops, phones, tablets, desktops). A red box highlights the search bar at the top left of the Baidu search interface.

# 关联规则挖掘

- 挖掘事物之间的关联关系
- 给定一组记录（数据对象），每个记录包含来自给定集合的一些项目（Item）
- 生成项集（itemset）之间的关联规则：

$X \rightarrow Y$

<i>TID</i>	<i>Items</i>
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

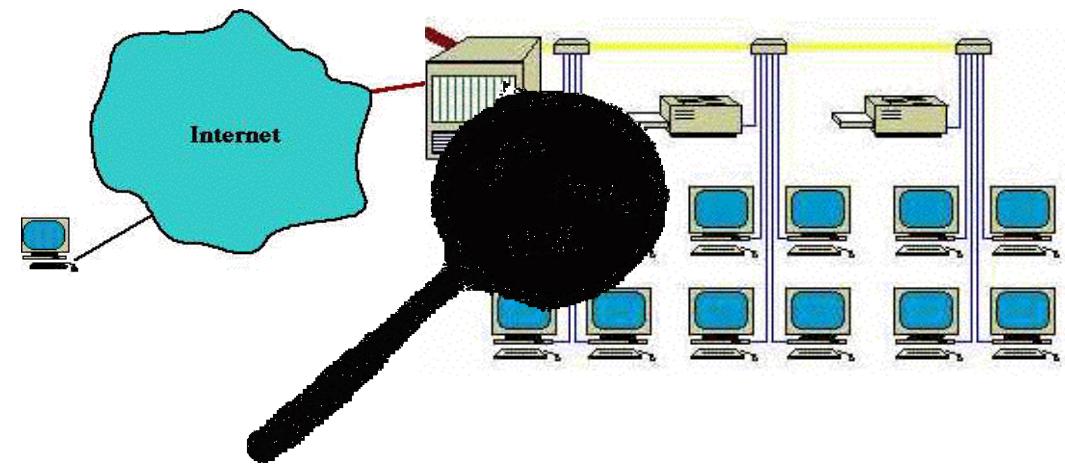
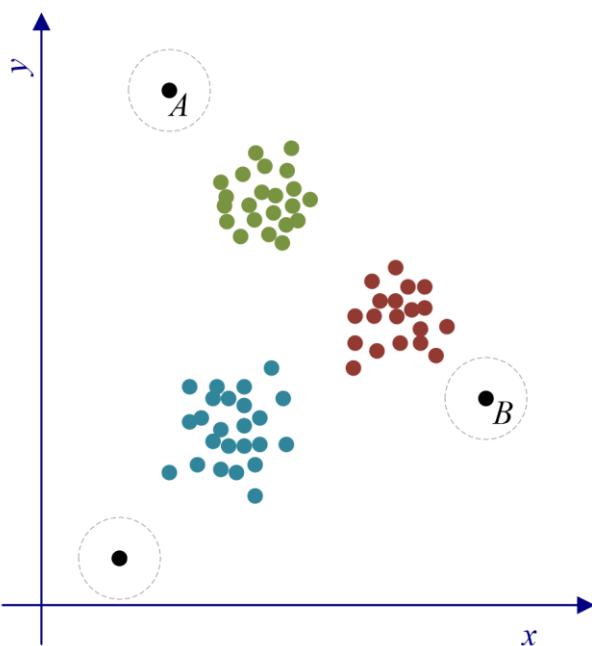
挖掘的关联规则：

$\{Milk\} \rightarrow \{Coke\}$

$\{Diaper, Milk\} \rightarrow \{Beer\}$

# 离群点/异常检测

- 检测与正常行为之间存在的显著偏差
- 应用：
  - 信用卡欺诈检测、网络入侵检测等



# 推荐系统：协同过滤

- 给定用户偏好的数据库，预测新用户的偏好
- 示例：预测你喜欢的新电影，根据
  - 你过去的偏好
  - 其他有相同偏好的人，以及他们对新电影的偏好

	5			4	
		1	2	3	3
	4		4		
		3			
				2	1

# 推荐系统的成功应用案例：今日头条

- 根据浏览历史推荐新闻

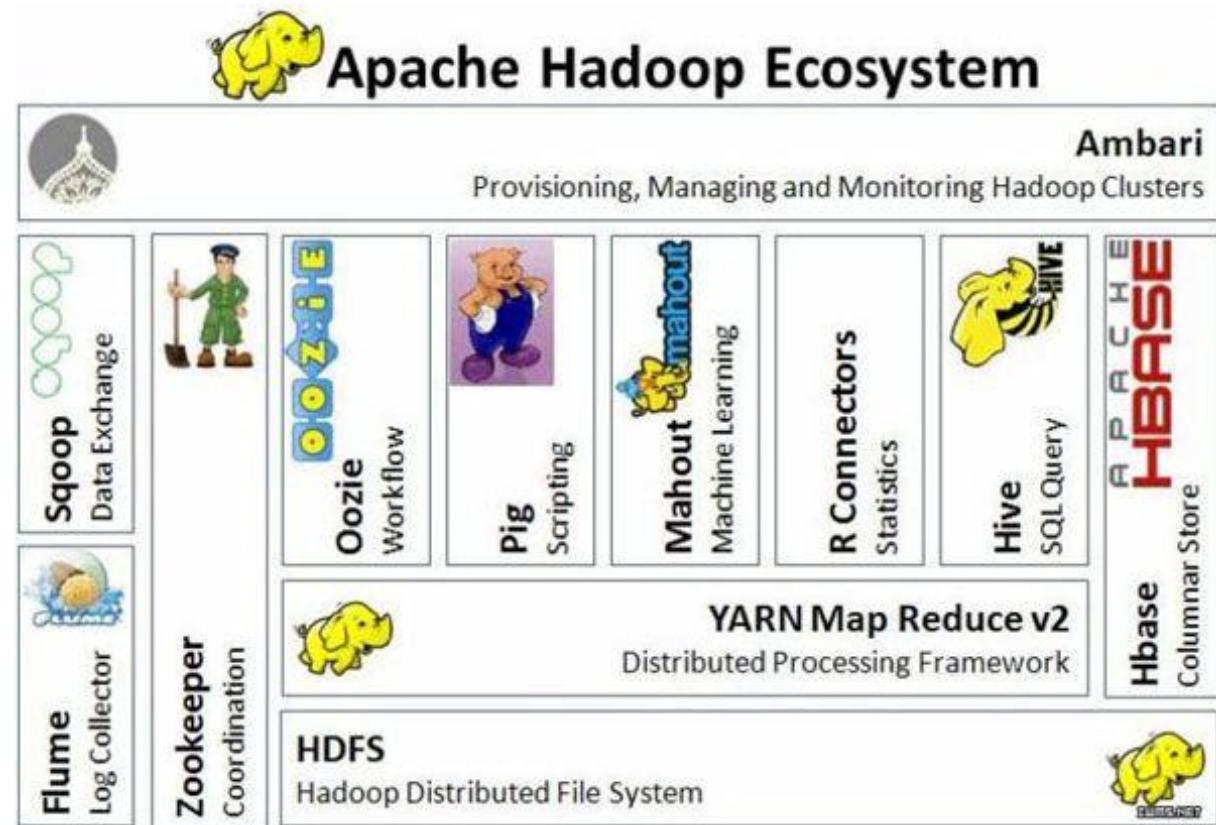
The screenshot illustrates the Toutiao news recommendation system. On the left, a sidebar menu lists categories such as '推荐' (Recommendation), '阳光宽频' (Broadband Sunlight), '热点' (Hot Topics), '图片' (Images), '科技' (Technology), '娱乐' (Entertainment), '游戏' (Games), '体育' (Sports), '汽车' (Cars), '财经' (Finance), '搞笑' (Funny), and '更多' (More). A red arrow points from the '推荐' button in the sidebar to the recommended news feed on the right. The main content area features a large image of a harbor with many fishing boats under a cloudy sky, with the caption '浙江舟山：千艘渔船归港备战台风“安比”'. Below this, a news item is highlighted with a yellow border: '习近平：欢迎塞内加尔成为第一个同中国签署“一带一路”合作文件的西非国家' (International, 人民网, 25 comments). To the right of the sidebar, a vertical navigation bar includes links for '要闻' (Top News), '社会' (Society), '娱乐' (Entertainment), '体育' (Sports), '军事' (Military), and '明星' (Famous Stars). The recommended news feed displays several articles with their respective thumbnails, titles, sources, and comment counts. A blue banner at the top of the feed reads '为您推荐了10篇文章' (Recommended 10 articles for you).

- 为您推荐了10篇文章
- 习近平：欢迎塞内加尔成为第一个同中国签署“一带一路”合作文件的西非国家  
国际 人民网 · 26评论 · 刚刚
- 习近平在南非媒体发表署名文章  
国际 新华网 · 1评论 · 刚刚
- 中国陆军首度军长大考，释放出什么信号？  
军事 上观新闻 · 20评论 · 刚刚
- 事业单位合并后，有职称的人员应该如何安置？  
社会 悟空问答 · 刚刚
- 面对美国颠倒黑白，华春莹的这些回应太精彩！  
国际 海外网 · 25评论 · 刚刚

# 大数据的学习资源

# 大数据存储与分析系统

- 关系数据库 (SQL) : MySQL, Oracle, .....
- 分布式文件系统: HDFS, CEPH, ...
- NoSQL: Not Only SQL
  - Key-value数据库
  - Redis、mongodb
  - 图数据库: Neo4J
  - .....
- 大数据分析处理
  - Spark, Storm, Flink, .....



# 传统数据挖掘系统

- 商业化系统
  - SAS Enterprise Miner
  - SPSS Clementine
  - Insightful Miner
  - Oracle/SQL Server提供的数据挖掘工具
  - .....
- 开源系统
  - Scikit-learn
  - Weka
  - Mahout
  - Spark .....

# 大数据领域的重要国际会议和期刊

## ● Conferences

- SIGMOD
- VLDB
- ICDE
- ACM SIGKDD Int. Conf. on Knowledge Discovery in Databases and Data Mining (**KDD**)
- SIAM Data Mining Conf. (**SDM**)
- (IEEE) Int. Conf. on Data Mining (**ICDM**)
- PAKDD, PKDD

## ■ Other related conferences

- WWW, SIGIR
- ICML, CVPR, NIPS, IJCAI

## ■ Journals

- IEEE Trans. On Knowledge and Data Eng. (TKDE)
- ACM Transactions on Information Systems
- ACM Transactions on Database Systems
- The VLDB Journal
- ACM Trans. on KDD

# Thank You for Your Attention

Contact me at: [yym@hit.edu.cn](mailto:yym@hit.edu.cn)

Tel: 26033008, 13760196623

Address: Rm.1402, H# Building

# 致谢

- 一小部分图表、文字来自互联网，仅供公益性的学习参考，在此表示感谢！如有版权要求请联系：[yym@hit.edu.cn](mailto:yym@hit.edu.cn)，谢谢！