# 数据挖掘
# 第一讲· 概 述

授课教师：别荣芳 教授、博导

课程助教：周也、王舒扬

# Outline

# 7. 数据挖掘的应用

## 商务智能

提供商务运作的历史、现状和预测视图

- 联机分析处理
- 商务业绩管理
- 竞争情报
- 标杆管理
- 预测分析

## Web搜索引擎

解决数据爬取、索引和搜索所面临的挑战

- 处理不断增加的数据
- 处理在线数据
- 在快速增长的数据流上维护更新模型
- 处理个性化查询

## 科技数据分析

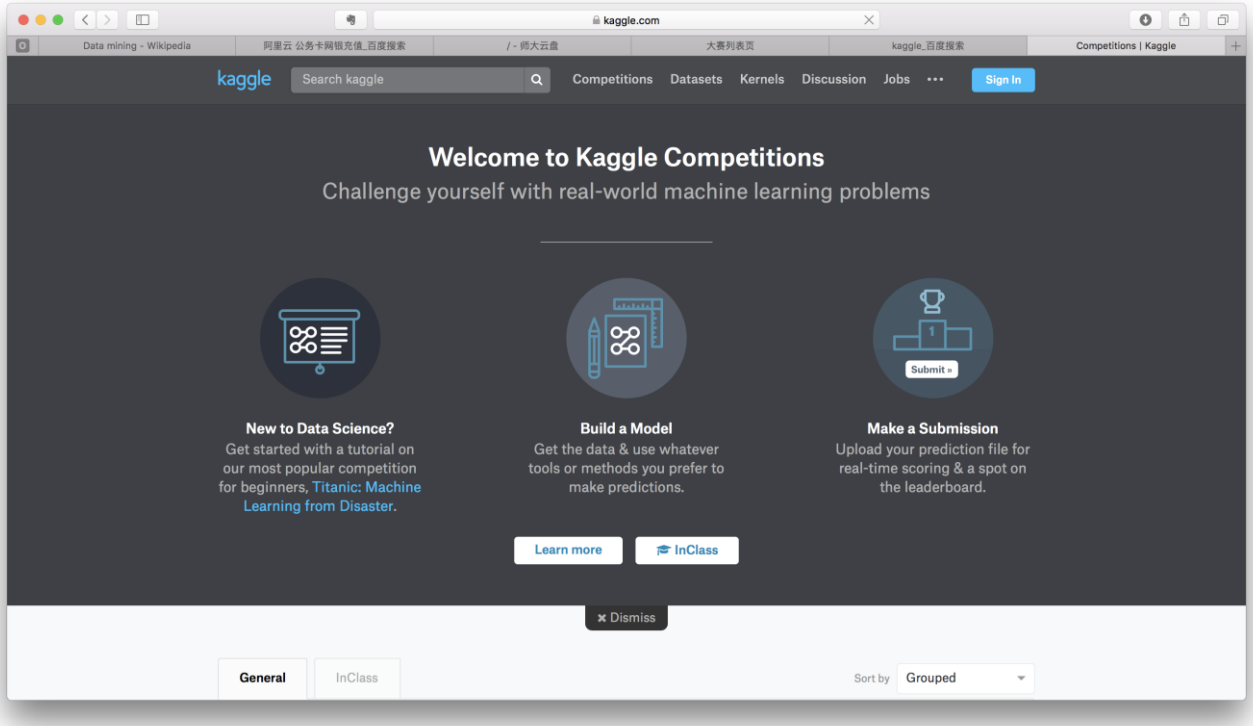
处理时空信息高维数据，流数据和异构数据

- 建立科学数据仓库
- 挖掘复杂数据类型
- 开发高级图形用户界面
- 研制可视化的工具
- …

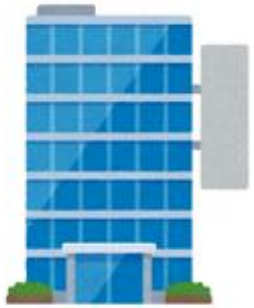
# 从竞赛平台看数据挖掘的应用

## 天池


## Kaggle



# Kaggle平台的工作形式



需求方通过平台发布数据及需要解决的问题

平台自动对结果打分

参赛者从平台获取数据，提交模型及结果

参赛者改进模型，不断提交更优的结果


赛期结束，评出本次竞赛的获奖者


平台基于参赛者在各项赛事中的表现，建立数据科学家排行榜

# Kaggle

kaggle

Search

Sign In | Register

- Home
- Compete
- Data
- Notebooks
- Discuss
- Courses
- More

## Competitions

Grow your data science skills by competing in our exciting competitions. Find help in the Documentation or learn about InClass competitions.

👋 New to Kaggle? Start here!

Our Titanic Competition is a great first challenge to get started.

Titanic: Machine Learning from Disaster
Start here! Predict survival on the Titanic and get familiar with ML basics
Getting Started • Ongoing • 16245 Teams — Knowledge

### All Competitions

Active | Completed | InClass | All Categories | Reward

Deepfake Detection Challenge
Identify videos with facial or voice manipulations
Featured • a month to go • Code Competition • 1819 Teams — $1,000,000

Google Cloud & NCAA® ML Competition 2020-NCAAM
Apply Machine Learning to NCAA® March Madness®
Featured • a month to go • 196 Teams — $25,000

Google Cloud & NCAA® ML Competition 2020-NCAAW
Apply Machine Learning to NCAA® March Madness®
Featured • a month to go • 120 Teams — $25,000

DS4G: Environmental Insights Explorer
Exploring alternatives for emissions factor calculations
Analytics • a month to go — $25,000

Google Cloud & NCAA® March Madness Analytics
Uncover the madness of March Madness®
Analytics • 2 months to go — $25,000

Abstraction and Reasoning Challenge
Create an AI capable of solving reasoning tasks it has never seen before
Research • 3 months to go • Code Competition • 197 Teams — $20,000

Bengali.AI Handwritten Grapheme Classification
Classify the components of handwritten Bengali
Research • a month to go • Code Competition • 1482 Teams — $10,000

Real or Not? NLP with Disaster Tweets
Predict which Tweets are about real disasters and which ones are not
Getting Started • a month to go • 3175 Teams — $10,000

Digit Recognizer
Learn computer vision fundamentals with the famous MNIST data
Getting Started • Ongoing • 2399 Teams — Knowledge

Titanic: Machine Learning from Disaster
Start here! Predict survival on the Titanic and get familiar with ML basics
Getting Started • Ongoing • 16245 Teams — Knowledge

House Prices: Advanced Regression Techniques
Predict sales prices and practice feature engineering, RFs, and gradient boosting
Getting Started • Ongoing • 4759 Teams — Knowledge

ImageNet Object Localization Challenge
Identify the objects in images
Research • 10 years to go • 65 Teams — Knowledge

Predict Future Sales
Final project for "How to win a data science competition" Coursera course
Playground • 10 months to go • 5750 Teams — Kudos

Categorical Feature Encoding Challenge II
Binary classification, with every feature a categorical (and interactions!)
Playground • a month to go • 672 Teams — Swag

Flower Classification with TPUs
Use TPUs to classify 104 types of flowers
Playground • 3 months to go • Code Competition • 199 Teams — Kudos

Connect X
Connect your checkers in a row before your opponent!
Getting Started • Ongoing • Simulation Competition • 457 Teams — Knowledge

# Kaggle经典案例

- **改进航班运营效率**
  - 2012年11月和2013年8月，通用公司在Kaggle上陆续发起了2组旨在改进航空公司航班效率的竞赛。从机场登机口分配到空中交通管理，许多环节都对航空公司航班的状态产生重大影响，进而引发航班延误，给航空公司带来损失。通用公司公布了由FlighStats公司提供的航空数据，希望参赛者开发出能够实时控制航班飞行路线、速度、高度和管制空域，从而优化航班的总体运行效率。
  - 第一组竞赛于2013年4月结束，决出的6个获胜团队，分享了25万美元奖金。
  - 第二组竞赛于2014年2月结束，5个获胜团队分享了另外25万美元奖金。
  - 在第二组竞赛中，由José Fonollosa提交的模型可以将现有航班效率提高12%以上，这将给航空公司节省巨额的费用。

# Kaggle经典案例

- **预测保险索赔情况**
  - 好事达保险公司（Allstate）希望能更准确地预测汽车的伤害索赔，以便优化保险的定价方案。于是该公司向竞争者们提供了2005年到2007年的保险数据，包括具体的汽车情况、以及每辆车相关的赔偿支出次数和数量，悬赏1万美金寻求解决方案。
  - 澳大利亚悉尼的保险精算顾问卡尔（Matthew Carle）使用基于决策树算法的模型夺得第一，他的模型能够获得比好事达保险公司的现有模型高出34%的精确度。

# Kaggle经典案例

- **预测医院住院病人流**
  - 位于加利福尼亚州的医疗保健机构HPN（Heritage Provider Network）在Kaggle上举办了一个为期近2年的竞赛，参赛者需根据36个月内的一系列医疗数据来预测哪些病人将会在一年内需要住院治疗。如果能准确预测病人入院需求，医疗机构可以预先进行预防和治疗，从而节省大量的医疗资金。
  - 该项竞赛设立了高达300万美元的奖金，成为Kaggle上奖金额最高的竞赛项目。该比赛已于2013年4月结束，共有1659个参赛队提交了超过35,000个模型。最终，第一名由一个名为POWERDOT的团队获得。

# Kaggle经典案例

- **Titanic生存预测**
  - Titanic: Machine Learning from Disaster，利用机器学习预测泰坦尼克乘客是否生还。迄今为止参赛的队伍已经达到17,065支，所有参与Kaggle的人都会拿这个题目练手。
  - 训练和测试数据是一些乘客的个人信息以及存活状况，要尝试根据它生成合适的模型并预测其他人的存活状况。
  - 题目虽简单，但取得好的名次也不容易。

# 阿里天池


PAKDD2021 第二届阿里云智能运维算法大赛 算法大赛

赛事简要：大规模内存故障预测是阿里巴巴进行智能化运维布局中的重要一环，课题难度大，价值高，通过大赛携手天池开发者共建智能运维生态圈。

举办方：Alibaba Cloud PAKDD 2021

奖金 $30000 | 团队 472 | 赛季1 2021-03-25 | 进行中

“AI Earth”人工智能创新挑战赛——AI助力精准气象和海洋预测 算法大赛

赛事简要：大赛聚焦全球大气海洋研究前沿方向，推进人工智能技术在气象和海洋领域的学术合作、人才培养、技术交流以及多学科交叉融合应用，探索利用人工智能技术突破行业关键性瓶颈，促进...

举办方：达摩院

奖金 ￥200000 | 团队 1912 | 赛季1 2021-03-24 | 进行中

“新内容 新交互”全球视频云创新挑战赛--算法挑战赛道 算法大赛

赛事简要：本届全球视频云创新挑战赛是由阿里云联手英特尔主办，与优酷战略技术合作，面向企业以及个人开发者的音视频领域的数据算法及创新应用类挑战。本届大赛包括两个赛道：“算法挑战...

举办方：阿里云 intel

奖金 ￥400000 | 团队 338 | 赛季1 2021-04-16 | 进行中

“新内容 新交互”全球视频云创新挑战赛 - 创新应用赛道 创新应用大赛

赛事简要：本届全球视频云创新挑战赛是由阿里云联手英特尔主办，与优酷战略技术合作，面向企业以及个人开发者的音视频领域的数据算法及创新应用类挑战。本届大赛包括两个赛道：“算法挑战...

举办方：阿里云 intel

奖金 ￥400000 | 团队 263 | 赛季1 2021-04-16 | 进行中

PAKDD2021 第二届阿里云智能运维算法大赛 算法大赛

赛事简要：大规模内存故障预测是阿里巴巴进行智能化运维布局中的重要一环，课题难度大，价值高，通过大赛携手天池开发者共建智能运维生态圈。

举办方：Alibaba Cloud PAKDD 2021

奖金 $30000 | 团队 472 | 赛季1 2021-03-25 | 进行中

“AI Earth”人工智能创新挑战赛——AI助力精准气象和海洋预测 算法大赛

赛事简要：大赛聚焦全球大气海洋研究前沿方向，推进人工智能技术在气象和海洋领域的学术合作、人才培养、技术交流以及多学科交叉融合应用，探索利用人工智能技术突破行业关键性瓶颈，促进...

举办方：达摩院

奖金 ￥200000 | 团队 1912 | 赛季1 2021-03-24 | 进行中

“新内容 新交互”全球视频云创新挑战赛--算法挑战赛道 算法大赛

赛事简要：本届全球视频云创新挑战赛是由阿里云联手英特尔主办，与优酷战略技术合作，面向企业以及个人开发者的音视频领域的数据算法及创新应用类挑战。本届大赛包括两个赛道：“算法挑战...

举办方：阿里云 intel

奖金 ￥400000 | 团队 338 | 赛季1 2021-04-16 | 进行中

全球人工智能技术创新大赛【赛道三】 算法大赛

赛事简要：赛场三以“小布助手对话短文本语义匹配”为课题，要求参赛队伍根据脱敏后的短文本query-pair，预测它们是否属于同一语义。

举办方：中国人工智能学会

奖金 ￥1500000 | 团队 1929 | 赛季1 2021-04-09 | 进行中

全球人工智能技术创新大赛【赛道二】 算法大赛

赛事简要：赛场二以“PANDA大场景多对象检测跟踪”为课题，包含大场景多目标检测、追踪等视觉任务，要求参赛队伍关注动态大场景多对象数据处理算法的研究。

举办方：中国人工智能学会

奖金 ￥1500000 | 团队 1474 | 赛季1 2021-04-09 | 进行中

# 生活大实惠：O2O优惠券使用预测

## 赛题内容

| | 状态 | 举办方 | 赛季1 | 奖金 | 参赛队伍 |
|---|---|---|---|---|---|
| 天池新人实战赛o2o优惠券使用预测 | 进行中 | 阿里云 | 2022-02-01 | ￥0 | 19298 |

报名参赛

- O2O行业天然关联数亿消费者，各类APP每天记录了超过百亿条用户行为和位置记录，因而成为大数据科研和商业化运营的最佳结合点之一。 以优惠券盘活老用户或吸引新客户进店消费是O2O的一种重要营销方式。然而随机投放的优惠券对多数用户造成无意义的干扰。对商家而言，滥发的优惠券可能降低品牌声誉，同时难以估算营销成本。 个性化投放是提高优惠券核销率的重要技术，它可以让具有一定偏好的消费者得到真正的实惠，同时赋予商家更强的营销能力。本次大赛为参赛选手提供了O2O场景相关的丰富数据，希望参赛选手通过分析建模，精准预测用户是否会在规定时间内使用相应优惠券。

# 生活大实惠：O2O优惠券使用预测

- **相关数据集：**2016.01.01至2016.06.30，用户线下消费和优惠券领取使用行为的记录表，用户线上点击/消费和优惠券领取使用行为的纪录表。需要预测的是2016年7月份用户领取优惠券后是否使用。
- **解决方案概述：**根据这两份数据表，首先对数据集进行划分，然后提取了用户相关的特征、商家相关的特征，优惠券相关的特征，用户与商家之间的交互特征，以及利用本赛题的leakage得到的其它特征（这部分特征在实际业务中是不可能获取到的）。最后训练了XGBoost，GBDT，RandomForest进行模型融合。

https://github.com/wepe/O2O-Coupon-Usage-Forecast

## 数据集划分

可以采用滑窗的方法得到多份训练数据集，特征区间越小，得到的训练数据集越多。以下是一种划分方式：

| | 预测区间（提取label） | 特征区间（提取feature） |
|---|---|---|
| 测试集 | 20160701~20160731 | 20160315~20160630 |
| 训练集1 | 20160515~20160615 | 20160201~20160514 |
| 训练集2 | 20160414~20160514 | 20160101~20160413 |

划取多份训练集，一方面可以增加训练样本，另一方面可以做交叉验证实验，方便调参。

https://github.com/wepe/O2O-Coupon-Usage-Forecast

# 生活大实惠：O2O优惠券使用预测

**特征工程**

- 赛题提供了online和offline两份数据集，online数据集可以提取到与用户相关的特征，offline数据集可以提取到更加丰富的特征：用户相关的特征，商家相关的特征，优惠券相关的特征，用户-商家交互特征。
- 另外，赛题提供的预测集中，包含了同一个用户在整个7月份里的优惠券领取情况，这实际上是一种leakage，比如存在这种情况：某一个用户在7月10日领取了某优惠券，然后在7月12日和7月15日又领取了相同的优惠券，那么7月10日领取的优惠券被使用的可能性就很大了。在做特征工程时注意到这一点，提取了一些相关的特征。加入这部分特征后，AUC提升了10个百分点以上，这些特征在实际业务中是无法获取到的。
- **模型设计与模型融合：。。。**XGBoost，GBDT，RandomForest三种等加权融合，。。。

https://github.com/wepe/O2O-Coupon-Usage-Forecast

# 生活大实惠：O2O优惠券使用预测

- **用户线上相关的特征**
  - 用户线上操作次数
  - 用户线上点击率
  - 用户线上购买率
  - 用户线上领取率
  - 用户线上不消费次数
  - 用户线上优惠券核销次数
  - 用户线上优惠券核销率
  - 用户线下不消费次数占线上线下总的不消费次数的比重
  - 用户线下的优惠券核销次数占线上线下总的优惠券核销次
  - 用户线下领取的记录数量占总的记录数量的比重

- **用户线下相关的特征**
  - 用户领取优惠券次数
  - 用户获得优惠券但没有消费的次数
  - 用户获得优惠券并核销次数
  - 用户领取优惠券后进行核销率
  - 用户满0~~50/50~~200/200~500 减的优惠券核销率
  - 用户核销满0~~50/50~~200/200~500减的优惠券占所有核销优惠券的比重
  - 用户核销优惠券的平均/最低/最高消费折率
  - 用户核销过优惠券的不同商家数量，及其占所有不同商家的比重
  - 用户核销过的不同优惠券数量，及其占所有不同优惠券的比重
  - 用户平均核销每个商家多少张优惠券
  - 用户核销优惠券中的平均/最大/最小用户-商家距离

https://github.com/wepe/O2O-Coupon-Usage-Forecast

# 生活大实惠：O2O优惠券使用预测

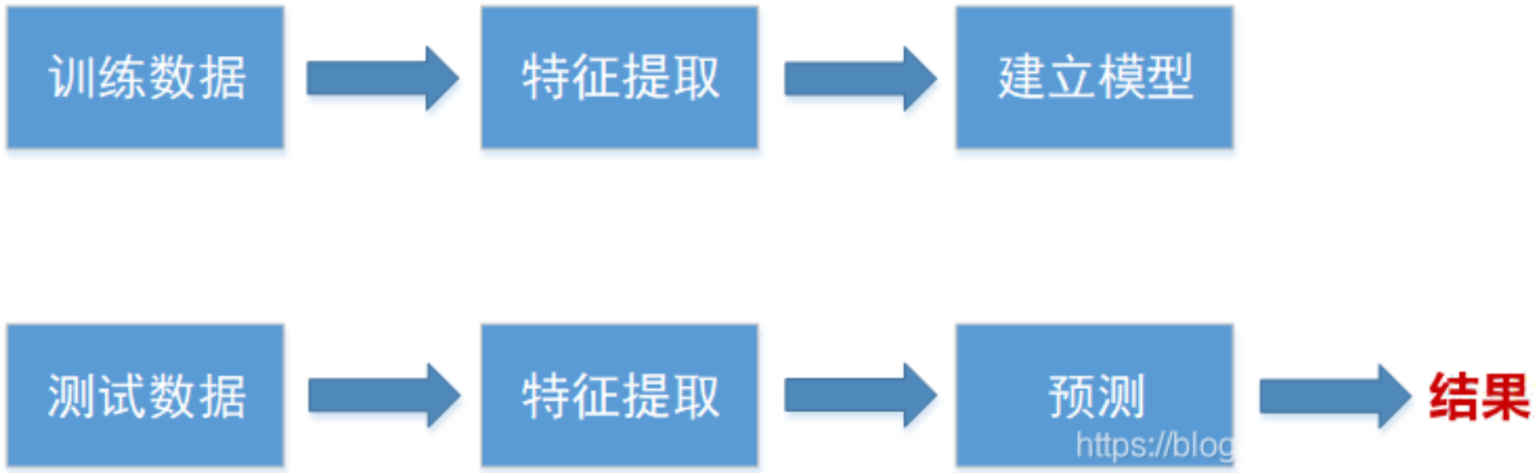
- **用户-商家交互特征**
  - 用户领取商家的优惠券次数
  - 用户领取商家的优惠券后不核销次数
  - 用户领取商家的优惠券后核销次数
  - 用户领取商家的优惠券后核销率
  - 用户对每个商家的不核销次数占用户总的不核销次数的比重
  - 用户对每个商家的优惠券核销次数占用户总的核销次数的比重
  - 用户对每个商家的不核销次数占商家总的不核销次数的比重
  - 用户对每个商家的优惠券核销次数占商家总的核销次数的比重

- **优惠券相关的特征**
  - 优惠券类型(直接优惠为0, 满减为1)
  - 优惠券折率
  - 满减优惠券的最低消费
  - 历史出现次数
  - 历史核销次数
  - 历史核销率
  - 历史核销时间率
  - 领取优惠券是一周的第几天
  - 领取优惠券是一月的第几天
  - 历史上用户领取该优惠券次数
  - 历史上用户消费该优惠券次数
  - 历史上用户对该优惠券的核销率

https://github.com/wepe/O2O-Coupon-Usage-Forecast

# 生活大实惠：O2O优惠券使用预测

这是一个典型的分类问题，其过程就是根据已有训练集进行训练，得到的模型再对测试进行测试并分类。整个过程如下图所示：




https://github.com/wepe/O2O-Coupon-Usage-Forecast

# 2021数字中国创新大赛大数据赛道-城市管理大数据专题


## 早高峰共享单车潮汐点的群智优化

### 赛题说明

共享单车，延伸了城市公共交通脉络，解决了市民出行“最后一公里”问题。然而，随着共享经济模式被越来越多市民接受，成为出行习惯，潮汐现
上休息的人类活动规律的客观存在，加之上下班时间段的集中，导致早晚高峰“一车难寻”、“无地可停”的供需矛盾。本题希望通过对车辆数据的
阶段潮汐点进行有效定位，进一步设计高峰期群智优化方案，缓解潮汐点供需问题，以期为城市管理部门和共享单车运营方研究制定下一步优化措施

- https://data.xm.gov.cn/contest-series/digit-china-2021/#/3/contest_explain

# 2021数字中国创新大赛大数据赛道-城市管理大数据专题

## 赛题说明

- 通过对车辆数据的综合分析，对厦门岛内早高峰阶段潮汐点进行有效定位，进一步设计高峰期群智优化方案，缓解潮汐点供需问题，以期为城市管理部门和共享单车运营方研究制定下一步优化措施提供数据支撑。
- “潮汐”现象：城市公共自行车从业者将发生在早晚高峰时段共享单车“借不到、还不进”的问题称之为“潮汐”现象。本题涉及的“潮汐现象”聚焦“还不进”的问题，识别出早高峰共享单车最淤积的40个区域

## 赛题任务

- 任务一：为更好地掌握早高峰潮汐现象的变化规律与趋势，参赛者需基于主办方提供的数据进行数据分析和计算模型构建等工作，识别出工作日早高峰07:00-09:00潮汐现象最突出的40个区域，列出各区域所包含的共享单车停车点位编号名称，并提供计算方法说明及计算模型，为下一步优化措施提供辅助支撑。
- 任务二：参赛者根据任务一Top40区域计算结果进一步设计高峰期共享单车潮汐点优化方案，通过主动引导停车用户到邻近停车点位停车，进行削峰填谷，缓解潮汐点停车位（如地铁口）的拥堵问题。允许参赛者自带训练数据，但需在参赛作品中说明所自带数据的来源及使用方式，并保证其合法合规。

# 小结

- 数据挖掘是从海量数据中发现有趣模式的过程。作为知识发现过程，它通常包括数据清理、数据集成、数据选择、数据变换、模式发现、模式评估和知识表示。

- 数据挖掘功能用来指定数据挖掘任务发现的模式或知识类型，包括特征化和区分，频繁模式、关联和相关性挖掘，分类和回归，聚类分析和离群点检测。

- 数据挖掘融汇来自其他一些领域的技术。这些领域包括统计学、机器学习、数据库和数据仓库系统，以及信息检索。数据挖掘研究与开发的多学科特点大大促进了数据挖掘的成功和广泛应用。

- 数据挖掘有许多成功的应用，如商务智能、Web搜索、生物信息学、卫生保健信息学、金融、数字图书馆和数字政府。

# 第1章 学习要求与作业题

- 本章涵盖基础概念与宽泛应用，建议认真阅读教材内容
- 讨论：
  1. 上网搜集资料，阐述数据挖掘与数据库、数据挖掘与机器学习的关系
  2. 与挖掘少量数据相比，挖掘海量数据的主要挑战是什么？
  3. 给出1个现实生活中成功的数据挖掘应用的案例
-

# 我们在数据的海洋里，渴望知识的淡水