
# 《数据挖掘》

大数据方向课程教学方法探讨

哈尔滨工业大学 计算机科学与技术学院

石胜飞

# 课程体系中的先修课

基础课程
- 必修：机器学习/人工智能

专业核心
- 必修：大数据计算基础、大数据分析

本课程
- 数据挖掘：教什么？如何教？

# 课程内容的选择依据

- 根据大学分系列课程的整体目标
  - 目标：“典型大数据系统的设计、实现与分析”
  - 在系列课程Ⅰ《大数据计算基础》中，学生掌握了大数据计算系统的概念与原理、具有一定的设计与开发的能力；
  - 在系列课程Ⅱ《大数据分析》中，学生能够掌握大数据分析基础知识，使用统计分析知识处理大数据分析的问题；
  - 本课程，将以数据挖掘实践为驱动，选择并整合各阶段所需核心知识，拓展学生解决数据挖掘问题的技能，并训练学生对算法进行灵活选择、改进以及创新的能力。
- 根据本领域的知识体系
  - 结合国内外数据挖掘研究以及工业应用领域，对于相关基础知识的需求
  - 适当引入近期国内外的研究成果（KDD、CIKM），以及工业界成功的框架

# IEEE-Computer Society Computer Science Curricula 2013

- ***Topics:***
  - The usefulness of data mining
  - Associative and sequential patterns
  - Data clustering
  - Market basket analysis
  - Data cleaning
  - Data visualization

# Computer Science Curricula 2013

- ***Learning Outcomes***:
  - Compare and contrast different conceptions of data mining as evidenced in both research and application
  - Explain the role of finding associations in commercial market basket data
  - Characterize the kinds of patterns that can be discovered by association rule mining
  - Evaluate methodological issues underlying the effective application of data mining
  - Identify and characterize sources of noise, redundancy, and outliers in presented data
  - Identify mechanisms (on-line aggregation, anytime behavior, interactive visualization) to close the loop in the data mining process
  - Describe why the various close-the-loop processes improve the effectiveness of data mining

# 课程总体思路

- 系统性地学习数据挖掘的基础理论、结合大数据环境下的实践能力培养
- 以课程大纲为主线，运用不同的实例和开放数据集，贯穿数据挖掘过程的各个环节教学
- 设立阶段性的任务（平时作业），强化对各个数据挖掘功能的理解和实践能力
- 设立若干实验题目，完成比较完整的数据挖掘任务，综合性较强

# 如何讲这门课?

- 从“教什么，就学什么”，转变到“我需要学什么？”
  - 每个章节都给定一个小的实践任务和不同规模的数据集，在相关的课程中教师讲解可能会用到的方法（课堂授课）。学生可以学以致用，或者根据参考文献找到更好的方法（课下自学）。
- 从各个孤立的知识点学习，转变到“问题驱动的系统性学习”
  - 从实际案例出发，以问题作为驱动，引出各个数据挖掘功能的介绍。
  - 树立整体意识，懂得各个独立的功能，在实际应用中，是相辅相成，互相渗透的
- 清楚：“学了那么多算法，我要用哪个更好？”
  - 每个问题都会讲解若干个算法，要引导学生掌握各个经典算法的适用性，以及影响算法性能的各个参数。突出大数据环境下如何提高系统的性能。
- “以赛代练”，强调实战能力，调动学生的积极性
  - 阶段性任务以及实验项目，设定各种指标进行排名

# 课程内容安排（大纲与教学设计）

- 本课程的教学方式从实践出发，结合适量的数据挖掘工程案例以及教学经验
- 以研究领域的公开数据集和相应研究问题为主线，深入浅出地讲授数据挖掘过程中的有关任务：

# 第1课 “Hello World”
# 从一个经典的数据集开始
# Titanic: Machine Learning from Disaster

- 第一节课，不讲大数据，不讲数据挖掘的故事，直接上数据。
- 数据集规模小，可用的模型很多，容易入手。
- 教学目的：自然引入数据挖掘的全过程概念、了解学生的知识结构与水平。
- 教学形式：教师演示的过程中，引导学生思考可能的方案。

# 第1课（2）：特征工程，真的精通这门“艺术”吗?

- 数据预处理以及特征工程，学生掌握的参差不齐，通过本课实例，系统介绍有关知识；
- 引导学生提出更多的特征，理解特征对于算法的重要性
- 展示典型的数据分析(EDA)和特征工程方法

# 第1课（3）数据分析

```
<class 'pandas.core.frame.DataFrame'>
Rang
Data
Pass
Surv
Pcla
Name
Sex
Age
SibS
Parc
Tick
Fare
Cabi
Emba
dtyp
memo
None
```

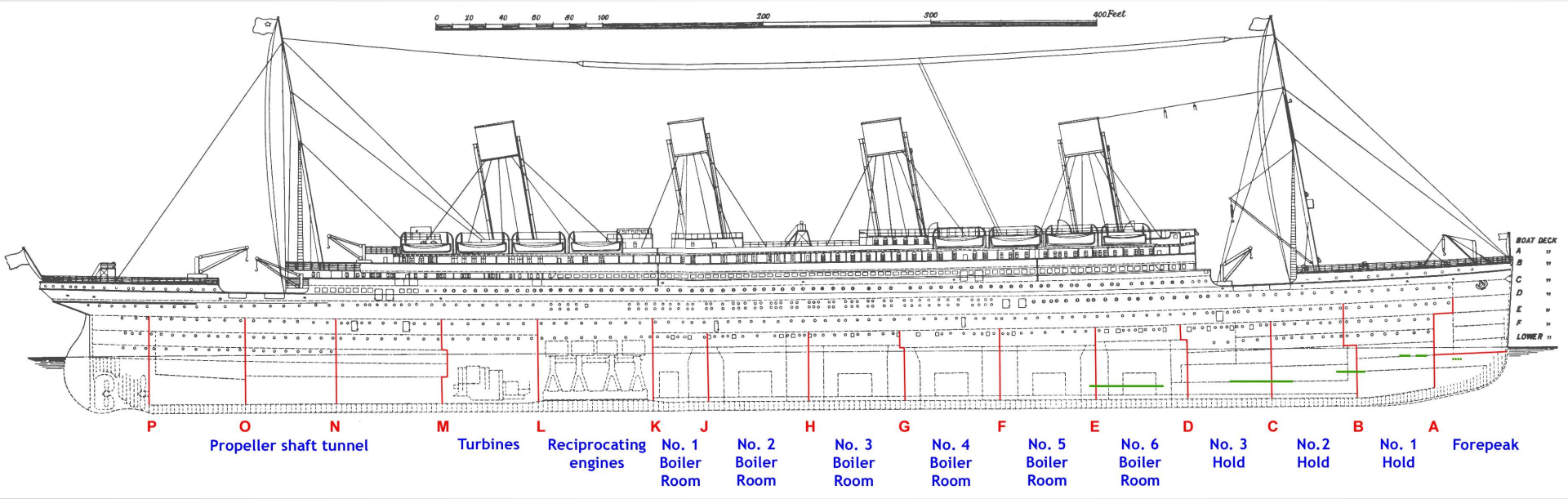
```
Training Set
Test Set
PassengerId column missing values: 0
Pclass column missing values: 0
Name column missing values: 0
Sex column missing values: 0
Age column missing values: 86
SibSp column missing values: 0
Parch column missing values: 0
Ticket column missing values: 0
Fare column missing values: 1
Cabin column missing values: 327
Embarked column missing values: 0
```

# 第1课（4）缺失值填充

- Age：简单用平均数或者中位数、根据相关性分析选择最相关的属性进行分组（聚类）后选择中位数等方法。
- 系统学习（复习）缺失值填充方法。

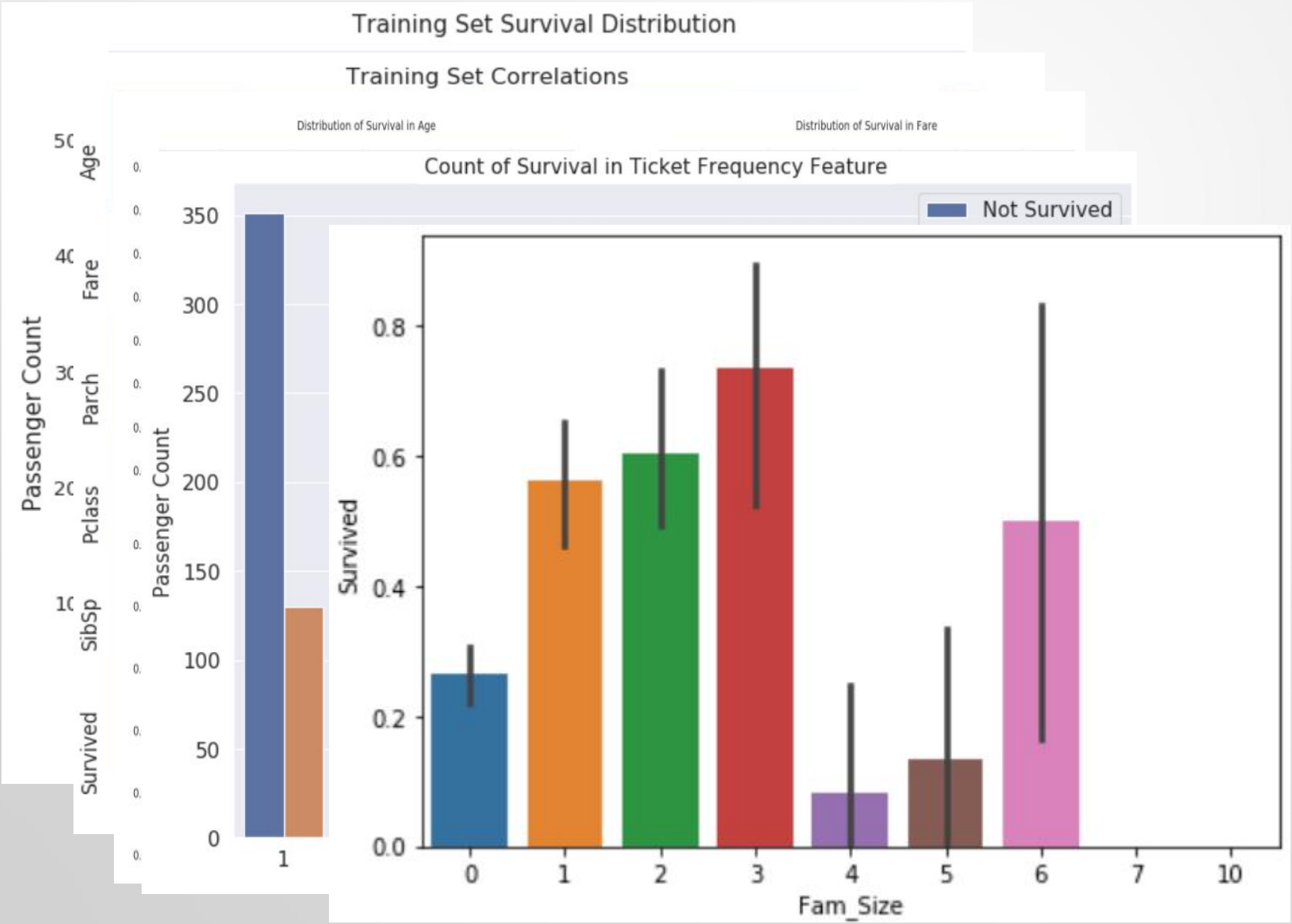
```
Median age of Pclass 1 females: 36.0
Median age of Pclass 1 males: 42.0
Median age of Pclass 2 females: 28.0
Median age of Pclass 2 males: 29.5
Median age of Pclass 3 females: 22.0
Median age of Pclass 3 males: 25.0
Median age of all passengers: 28.0
```

# 第1课（5）好的特征工程要与领域知识结合



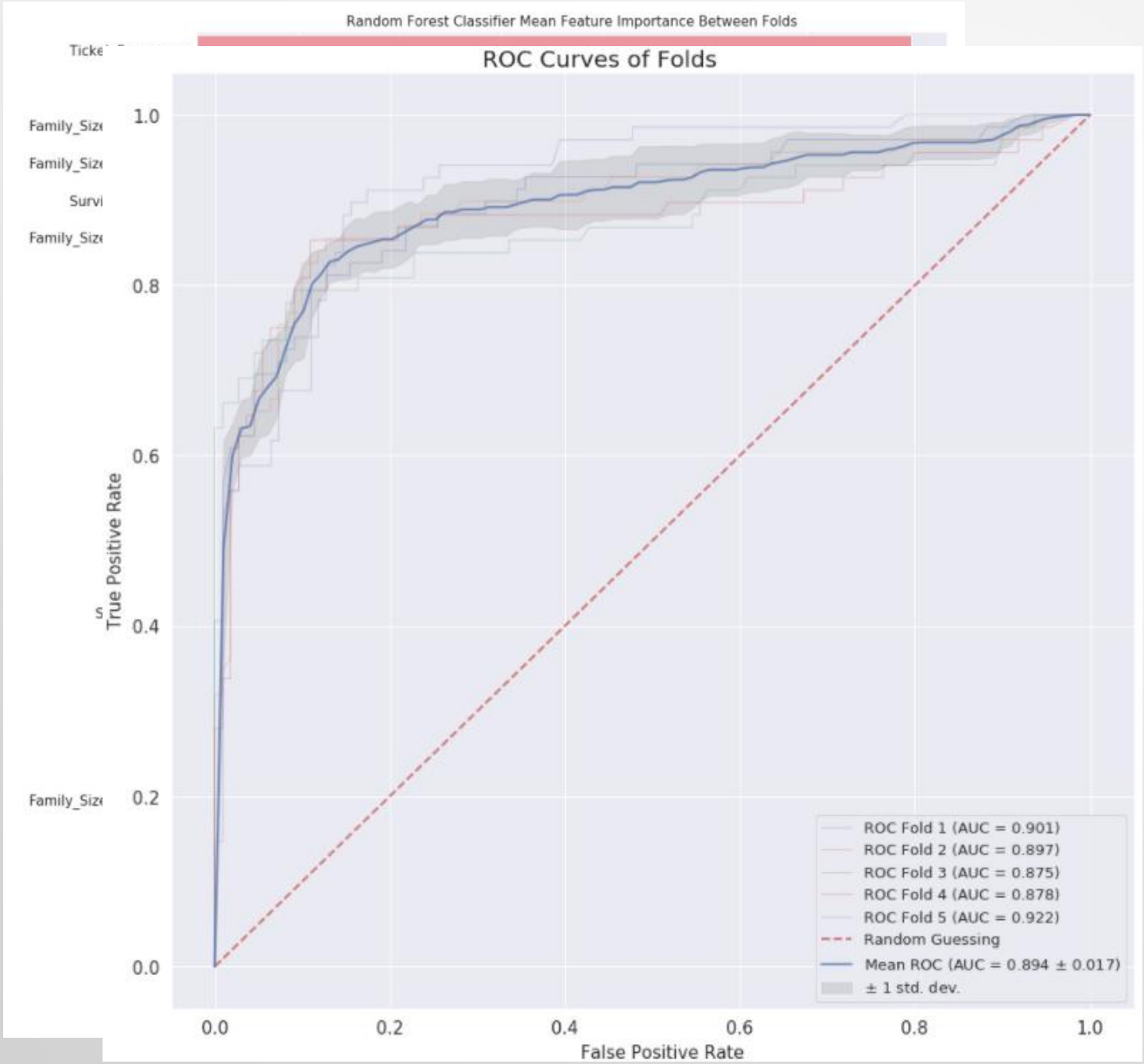

- Cabin属性与幸存下来的关系
- 大量缺失值如何填充？（每个人都会有不同的认识）

# 第1课（6）可视化分析的重要性



# 第1课（7）那么多模型，选哪个?

- 决策树、朴素贝叶斯、KNN、随机森林...
- 模型的评估：K折交叉验证（KFold，GroupKFold，StratifiedKFold...）
- 模型参数的选择与优化


Random Forest Classifier Mean Feature Importance Between Folds
ROC Curves of Folds
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC Fold 1 (AUC = 0.901)
ROC Fold 2 (AUC = 0.897)
ROC Fold 3 (AUC = 0.875)
ROC Fold 4 (AUC = 0.878)
ROC Fold 5 (AUC = 0.922)
Random Guessing
Mean ROC (AUC = 0.894 ± 0.017)
± 1 std. dev.

# 入门之后，学会自己获取数据

- 之前接触的训练数据都是给定的，在实际应用中，往往需要自行获取数据，并建立适当的特征
- 体育比赛结果预测，往往需要大数据的支持
- 秋季学期，NBA常规赛事恰好开始了...

# 第2课：NBA结果预测

October Schedule Share & more ▼

赛季数据

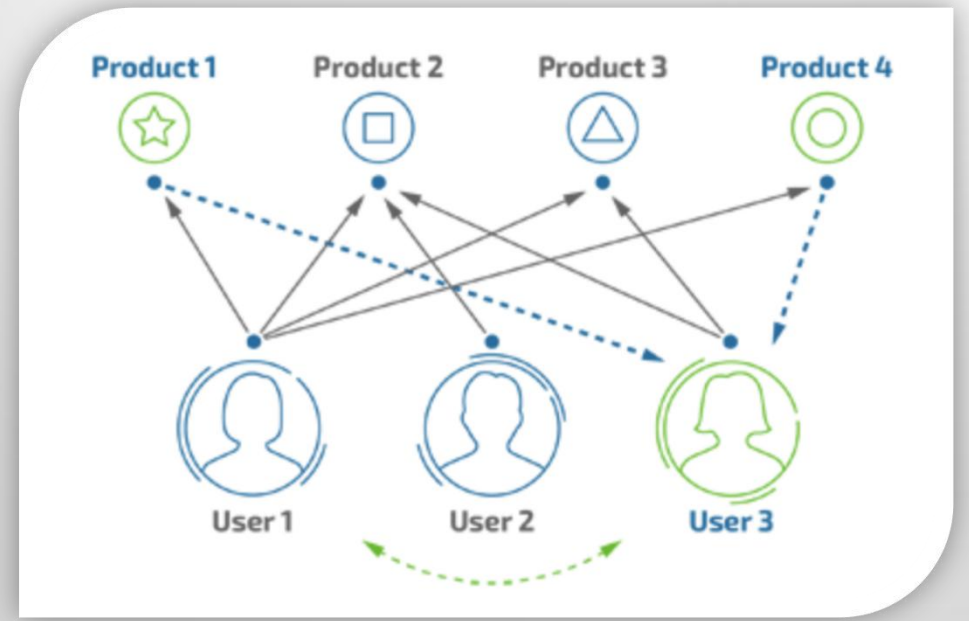
| 排名 | 球员 | 球队 | 场数 | 先发 | 场均篮板 | 场均助攻 | 分钟 | 效率 | % | 三分% | 罚球% | 进攻 | 防守 | 场均抢断 | 场均盖帽 | 失误 | 犯规 | 场均得分 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|

东部联盟

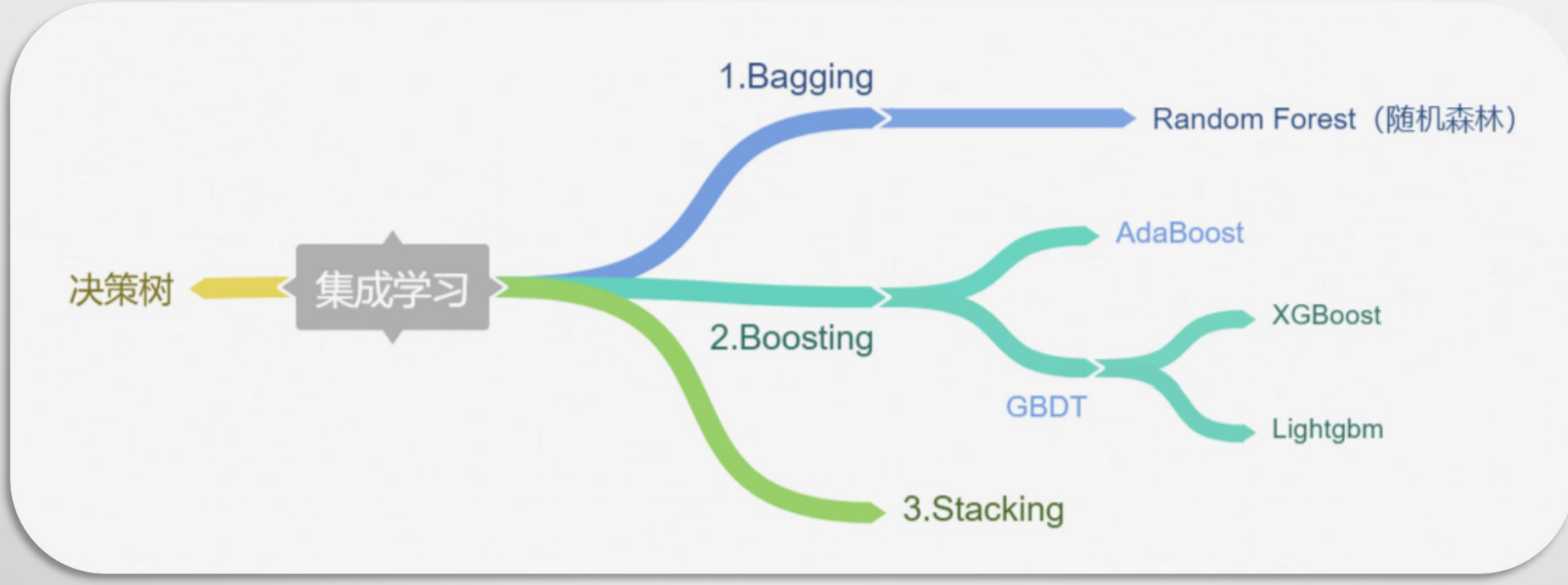
| 排名 | | 姓名 | 胜 | 负 | 连续 | 主场 | 客场 | 连胜 | 连败 | 最近10场 | 最近10个主场 | 最近10个客场 | 得分/场 | 场均失分 | 差距 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 01 | | e - 密尔沃基 雄鹿 | 60 | 22 | 负 1 | 负 1 | 胜 2 | 胜 7 | 负 2 | 7-3 | 7-3 | 5-5 | 118.1 | 109.3 | 8.8 |
| 02 | | a - 多伦多 猛龙 | 58 | 24 | 胜 2 | 胜 3 | 胜 1 | 胜 8 | 负 3 | 7-3 | 7-3 | 7-3 | 114.4 | 108.4 | 6.0 |
| 03 | | x - 费城 76人 | 51 | 31 | 胜 1 | 胜 1 | 负 1 | 胜 6 | 负 3 | 4-6 | 7-3 | 4-6 | 115.2 | 112.5 | 2.7 |
| 04 | | x - 波士顿 凯尔特人 | 49 | 33 | 胜 1 | 负 1 | 胜 3 | 胜 8 | 负 4 | 6-4 | 5-5 | 6-4 | 112.4 | 108.0 | 4.4 |
| 05 | | x - 印第安纳 步行者 | 48 | 34 | 胜 1 | 负 2 | 胜 2 | 胜 7 | 负 4 | 4-6 | 6-4 | 2-8 | 108.0 | 104.7 | 3.3 |
| 06 | | x - 布鲁克林 篮网 | 42 | 40 | 胜 3 | 胜 1 | 胜 2 | 胜 7 | 负 8 | 6-4 | 6-4 | 5-5 | 112.2 | 112.3 | -0.1 |
| 07 | | se - 奥兰多 魔术 | 42 | 40 | 胜 4 | 胜 9 | 胜 2 | 胜 6 | 负 4 | 8-2 | 9-1 | 4-6 | 107.3 | 106.6 | 0.7 |

# 推荐算法：温故知新

- 简单直观的：基于关联规则推荐的（复习相关算法）
- 进阶：协同过滤
- 更进一步：**ALS矩阵分解算法并行化**（Spark实例）

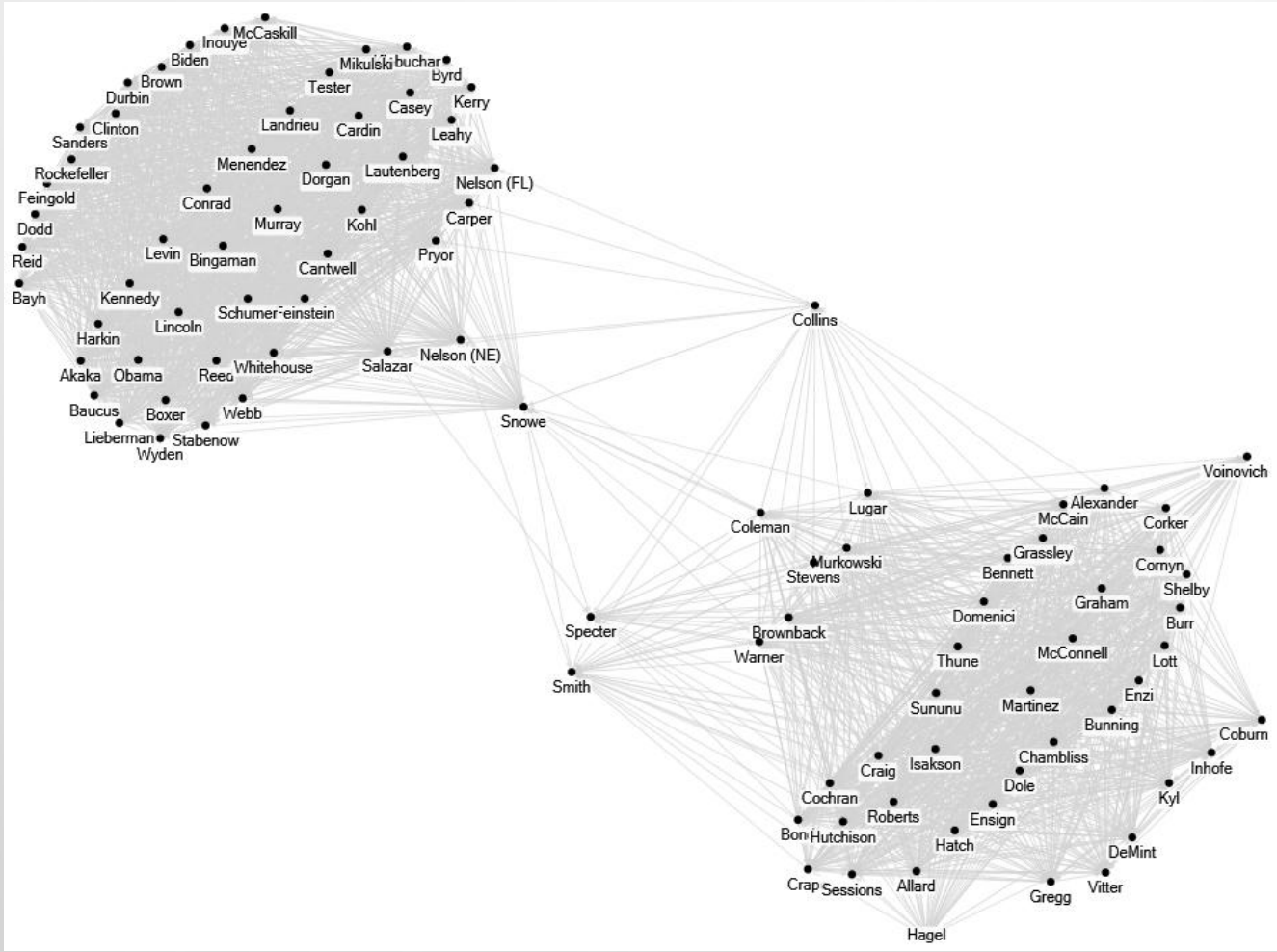


# 流行框架：集成学习的深入理解

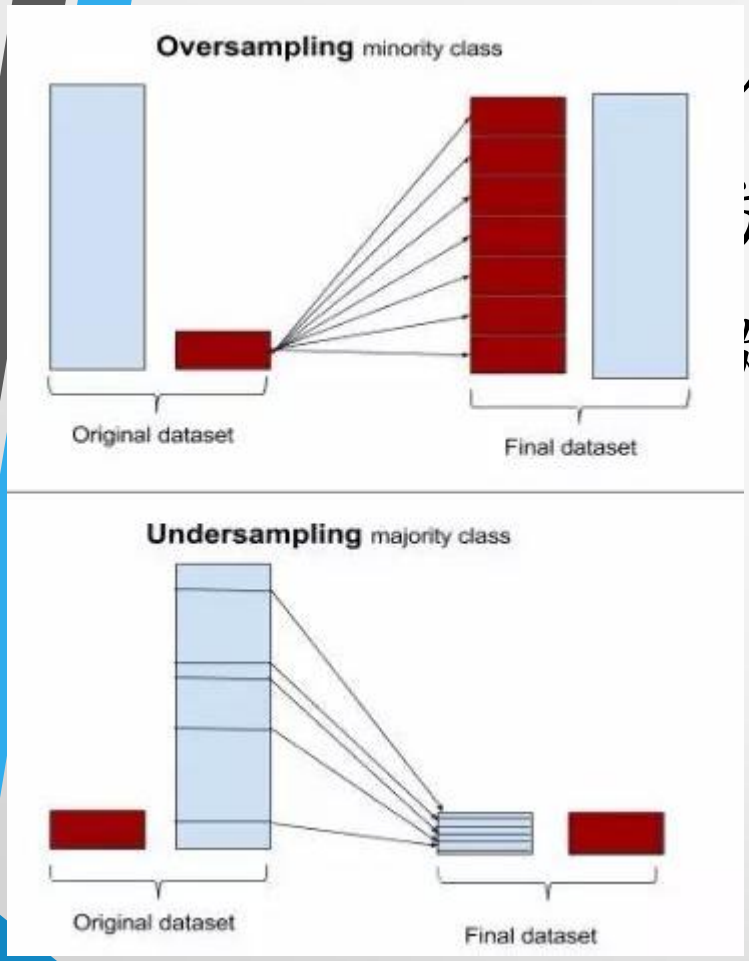


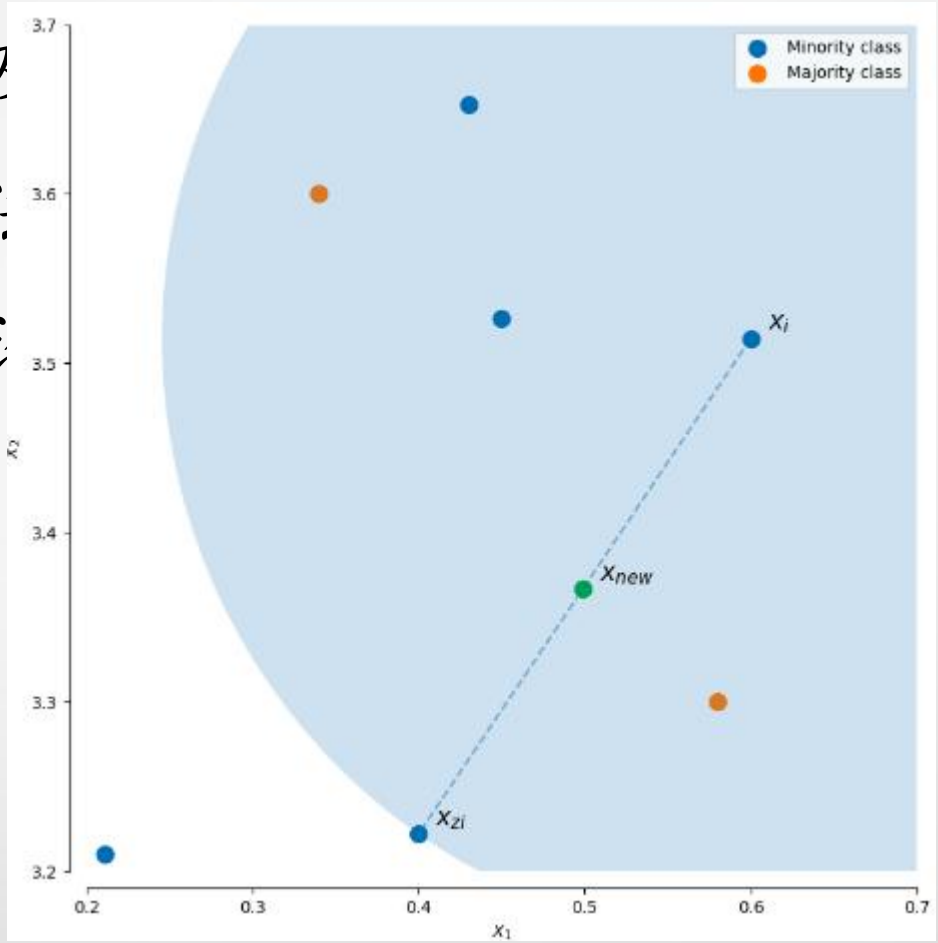

- 实验：医疗数据预测

# 图挖掘：社会网络分析


McCaskill
Inouye
Biden
Brown
Durbin
Clinton
Sanders
Rockefeller
Feingold
Dodd
Reid
Bayh
Harkin
Akaka
Obama
Baucus
Boxer
Lieberman
Wyden
Stabenow
Webb
Reed
Whitehouse
Kennedy
Lincoln
Levin
Bingaman
Conrad
Menendez
Landrieu
Tester
Mikulski
Cardin
Dorgan
Cantwell
Murray
Kohl
Lautenberg
Casey
Byrd
Kerry
Leahy
Nelson (FL)
Carper
Pryor
Schumer
Feinstein
Salazar
Nelson (NE)
Snowe
Collins
Specter
Smith
Coleman
Lugar
Murkowski
Stevens
Brownback
Warner
Alexander
McCain
Corker
Grassley
Bennett
Cornyn
Shelby
Graham
Domenici
Burr
Thune
McConnell
Lott
Sununu
Martinez
Enzi
Bunning
Coburn
Craig
Isakson
Chambliss
Inhofe
Dole
Kyl
Cochran
Roberts
Ensign
Hutchison
Hatch
DeMint
Crapo
Sessions
Allard
Gregg
Vitter
Hagel
Voinovich

# 异常检测：学以致用


Oversampling minority class
Original dataset
Final dataset
Undersampling majority class
Original dataset
Final dataset


Minority class
Majority class
$x_i$
$x_{new}$
$x_{zi}$
3.7
3.6
3.5
3.4
3.3
3.2
0.2
0.3
0.4
0.5
0.6
0.7
$x_1$
$x_2$

# 面向大数据：分布式数据挖掘算法

- 电商平台有上亿的用户和产品，每天产生百亿规模的用户反馈数据。
- 每天有100亿规模的用户行为数据。如此超大规模的训练数据，给分布式机器学习带来了巨大的挑战，也引入了有趣的研究问题。

- 数据与模型并行模式
- 同步与异步算法简介
- Spark ML经典算法简介

# 接触前沿：大数据挖掘新进展

- 近两年，KDD、SIGMOD、ICDE、CIKM等学术会议的相关研究成果