Neusoft
Beyond Technology

数据科学与大数据人才培养建设及应用实践

东软集团股份有限公司

大数据产品总监 邹存璐 博士

2018年10月14日

Neusoft SaCa.RealRec

目录

CONTENTS

数据科学与大数据人才缺口

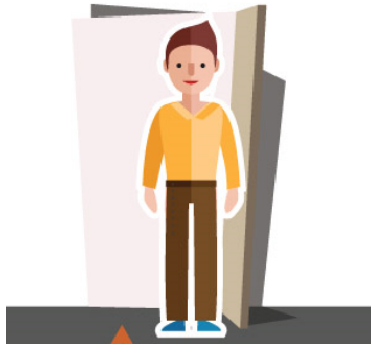
《2016年中国互联网最热职位人才报告》

数据分析师已成当下中国互联网行业需求最旺盛的六类人才职位之一，而且数据分析人才最为稀缺。报告表明，数据分析人才的供给指数仅为0.05，属于高度稀缺。此外，数据分析人才的跳槽速度也最快，平均跳槽速度为19.8个月。

《2016 Global Institute Report》

预计到2018年，大数据或者数据工作者的岗位需求将激增，其中大数据科学家的缺口在14万到19万之间，对于懂得如何利用大数据做决策的分析师和经理的岗位缺口则将达150万

企业基于角色的差异化能力要求



数据工程师

针对大数据存储查询需求，完成大数据系统架构设计以及研发工作。

1 程序开发(Python,Java...)
2 数据库设计(MySQL,MongoDB...)
3 分布式存储系统(HDFS...)
4 分布式计算系统(Spark...)
5 数据预处理(矩阵转置,缺失值填充...)
6 多维特征分析(关联性分析,主成份分析...)
7 熟悉数据软件和工具
8 基础算法（分类,聚类...）

数据分析师

根据明确业务分析预测目标完成数据采集、数据处理分析建模等技术工作。

1大数据底层原理
2 传统算法(分类，回归，聚类...)
3 时间序列分析
4 数据可视化
5 文本挖掘算法(分词,文本分类...)
6 数据分析、建模
7 业务模型理解
8 行业知识

数据科学家

根据客户目前业务特点以及数据现状，规划构建人工智能分析应用场景。

1 大数据底层原理
2 传统算法(分类，回归，聚类...)
3 时间序列分析
4 数据可视化
5 文本挖掘算法(word2vec,语法摘要)
6 深度学习(神经网络...)
7 商业模型理解
8 专业的行业知识

企业对数据分析人才需求问题

人才短缺、招聘困难

曾经创下全年面试几十人，只录用到合格人员2名的记录

理论基础技能强，缺乏应用实践技术结合能力

面试时基础理论技术能力很强，实际工作解决客户问题时缺乏技术业务灵活结合能力

人才技能培养周期长

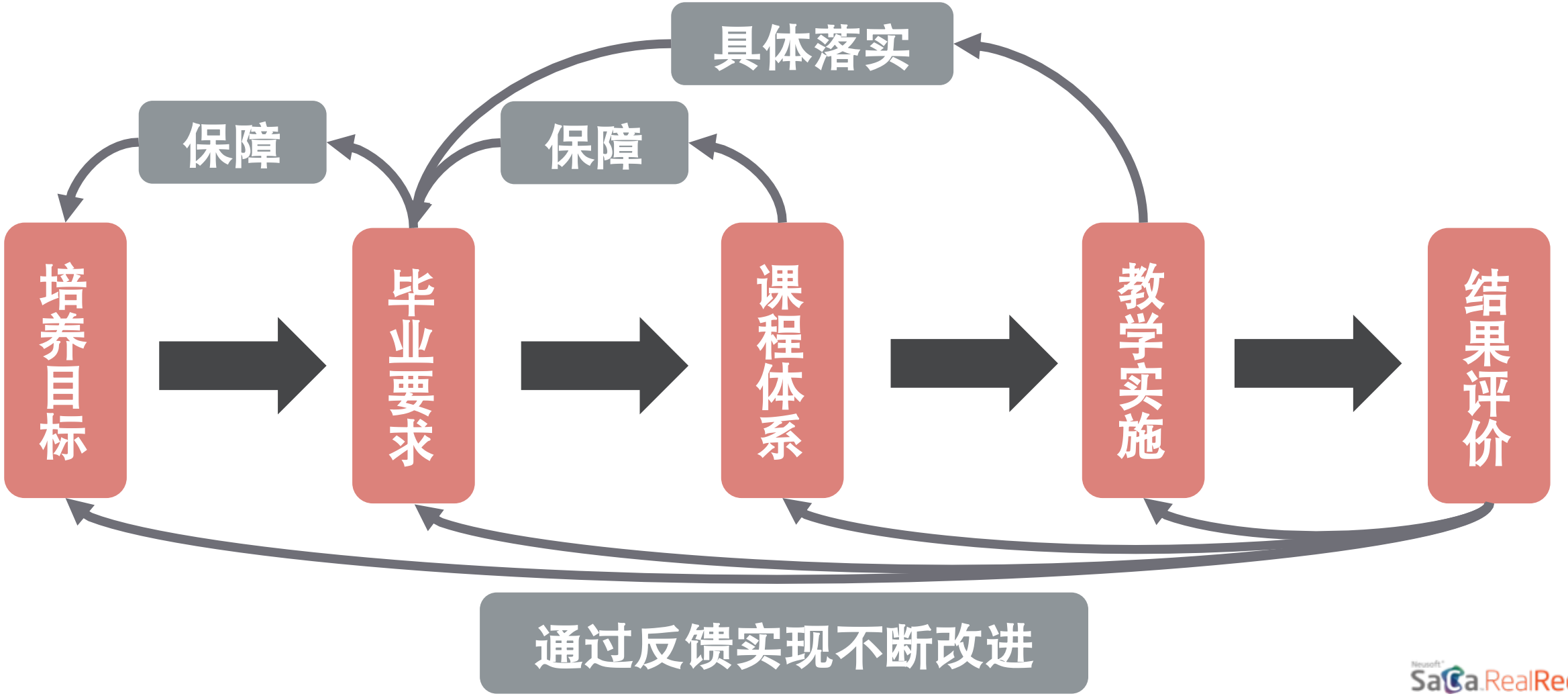
从新手到企业技术Leader平均培养周期大概在3年左右
事业部软件工程师众多，但对数据科学都是零基础，需要从头培训

市场人才竞争激烈，内部人才流失风险高

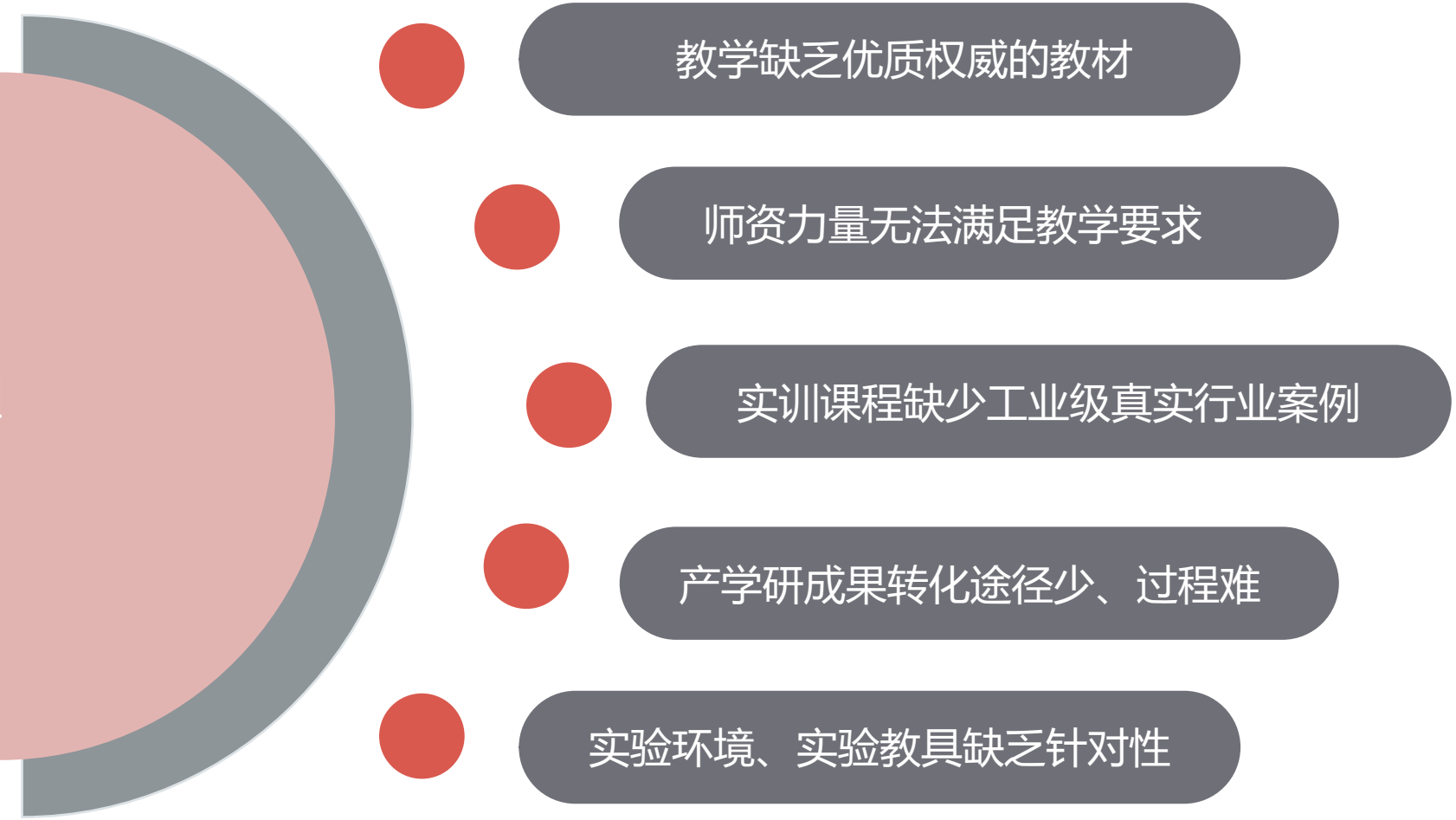
针对数据科学方面人才，包括工程师、算法分析师甚至是测试人员，
平均工作1年，市场工资翻倍

新工科建设背景下的专业改革

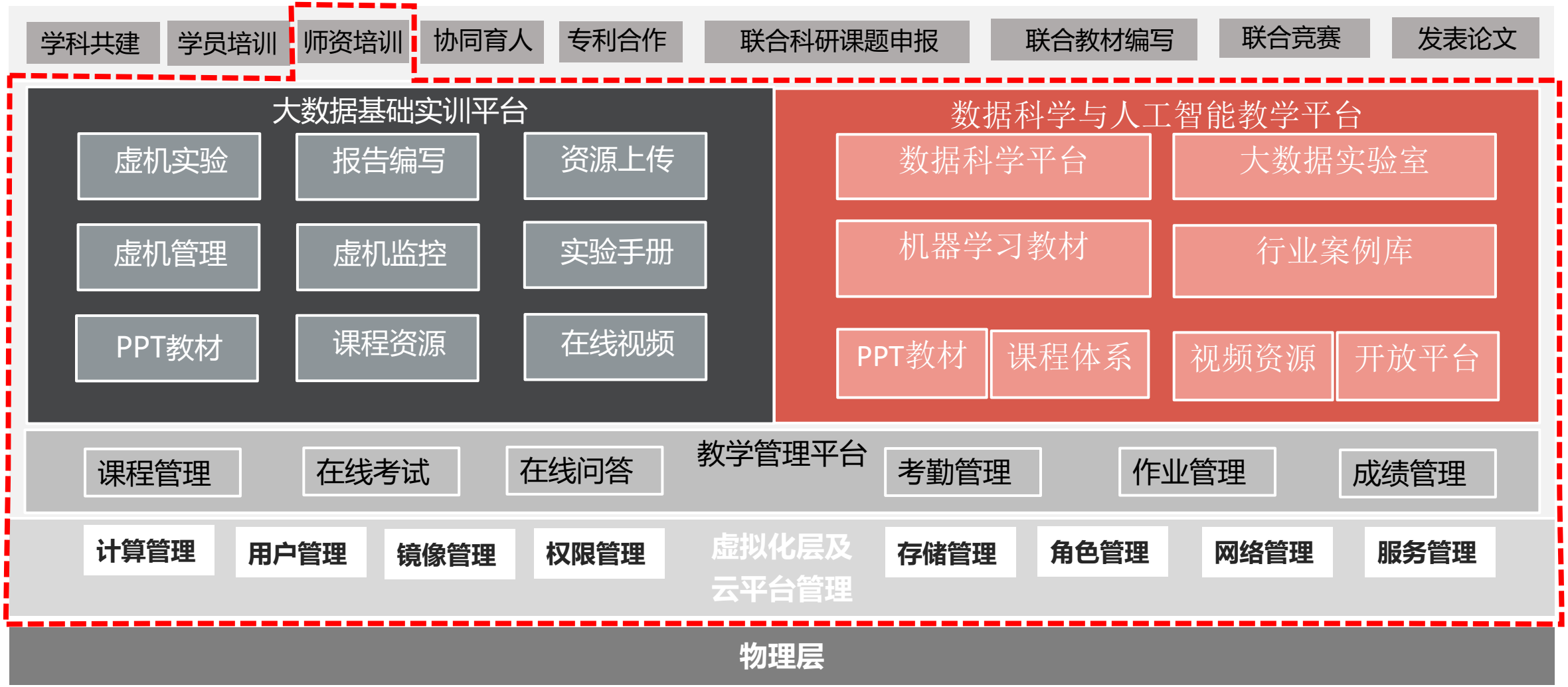
从知识传授转向能力培养，从学科导向转向产业需求导向



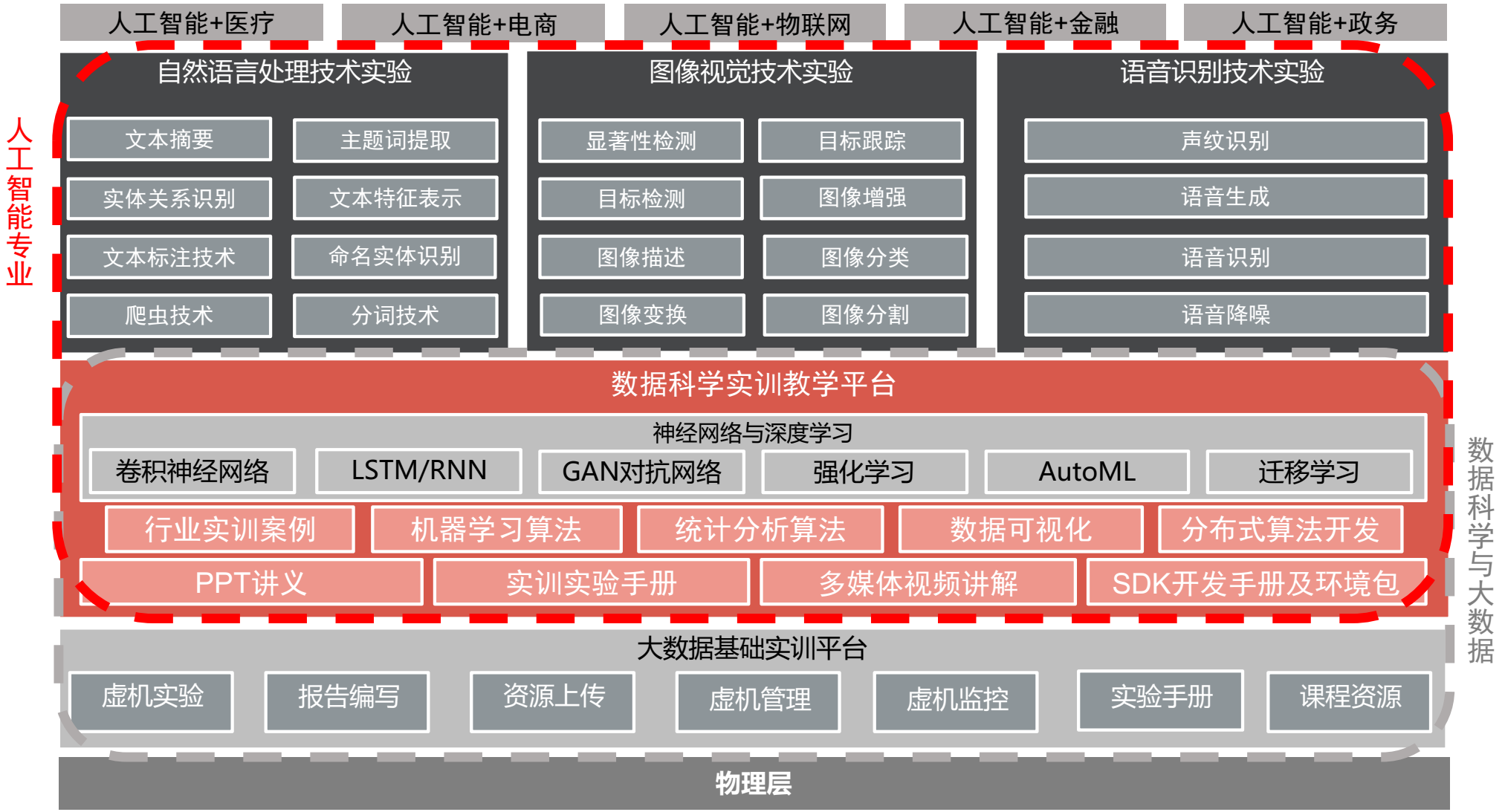
专业建设存在的问题与挑战



全面的培养方案


学科共建
学员培训
师资培训
协同育人
专利合作
联合科研课题申报
联合教材编写
联合竞赛
发表论文
大数据基础实训平台
虚机实验
报告编写
资源上传
虚机管理
虚机监控
实验手册
PPT教材
课程资源
在线视频
数据科学与人工智能教学平台
数据科学平台
大数据实验室
机器学习教材
行业案例库
PPT教材
课程体系
视频资源
开放平台
课程管理
在线考试
在线问答
教学管理平台
考勤管理
作业管理
成绩管理
计算管理
用户管理
镜像管理
权限管理
虚拟化层及
云平台管理
存储管理
角色管理
网络管理
服务管理
物理层

人工智能人才培养方案



大数据基础实训

大数据开源技术部署安装、运维调优，支撑海量数据存储查询技术应用场景。

数据科学实训

支撑机器学习、模型训练的实训平台工具，包括传统机器学习以及神经网络深度学习技术。

人工智能技术实训

通用人工智能技术，包括自然语言处理实训、图像视觉实训、语音识别处理实训。

人工智能领域实训

面向人工智能+领域的综合实训平台，包括人工智能+医疗、电商、物联网、金融、政务等业务场景。

Neusoft SaCa RealRec

人工智能岗位需求

数据标注工程师

1 命名实体标注
2 实体关系标注
3 分类标注
4 标签数据管理
5 边框标注
6 描点标注
7 行业领域专业知识

人工智能测试工程师

1 机器学习算法评估指标
2 机器学习算法概论
3 数据分析处理
4 自动化测试脚本工具
5 标签数据集构建管理
6 数据可视化
7 熟悉数据软件和工具
8 分布式环境压力测试

机器学习工程师

1 机器学习算法原理
2 传统算法(分类，回归，聚类...)
3 神经网络算法应用
4 自然语言处理算法应用
5 图像处理算法应用
6 人工智能模型调优
7 应用开发（Python/Java/SQL）
8 业务场景实现

数据科学家

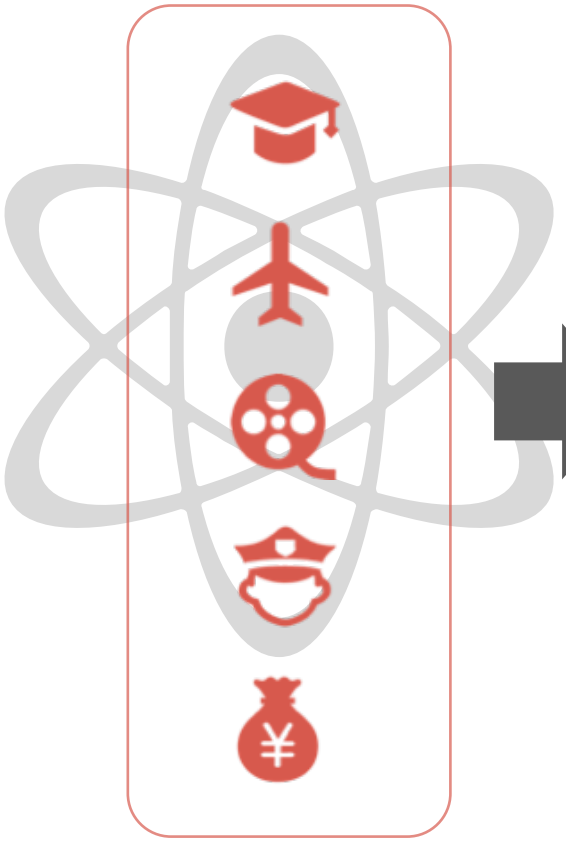
1 数学统计基础
2 器学习算法理论优化
3 神经网络架构设计
4 模型参数求解优化
5 损失函数定义
6 元学习（meta learning）方法
7 商业模型设计
8 专业的行业知识

目录
CONTENTS

基于多行业真实案例的大数据综合实训

大数据综合实训案例要贴近真实场景，通过案例背景介绍让学生熟悉行业背景，了解未来求职行业；使用真实数据，让学生从数据清洗开始就按照生产环境来学习数据操作的方法，对接工作需要；数据处理流程要开放，充分激发学生的创造性。

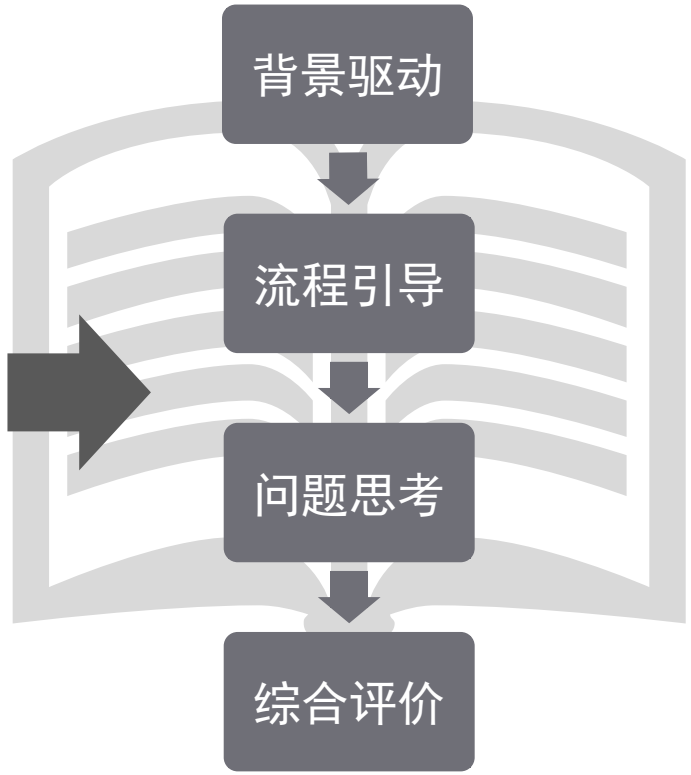
工业



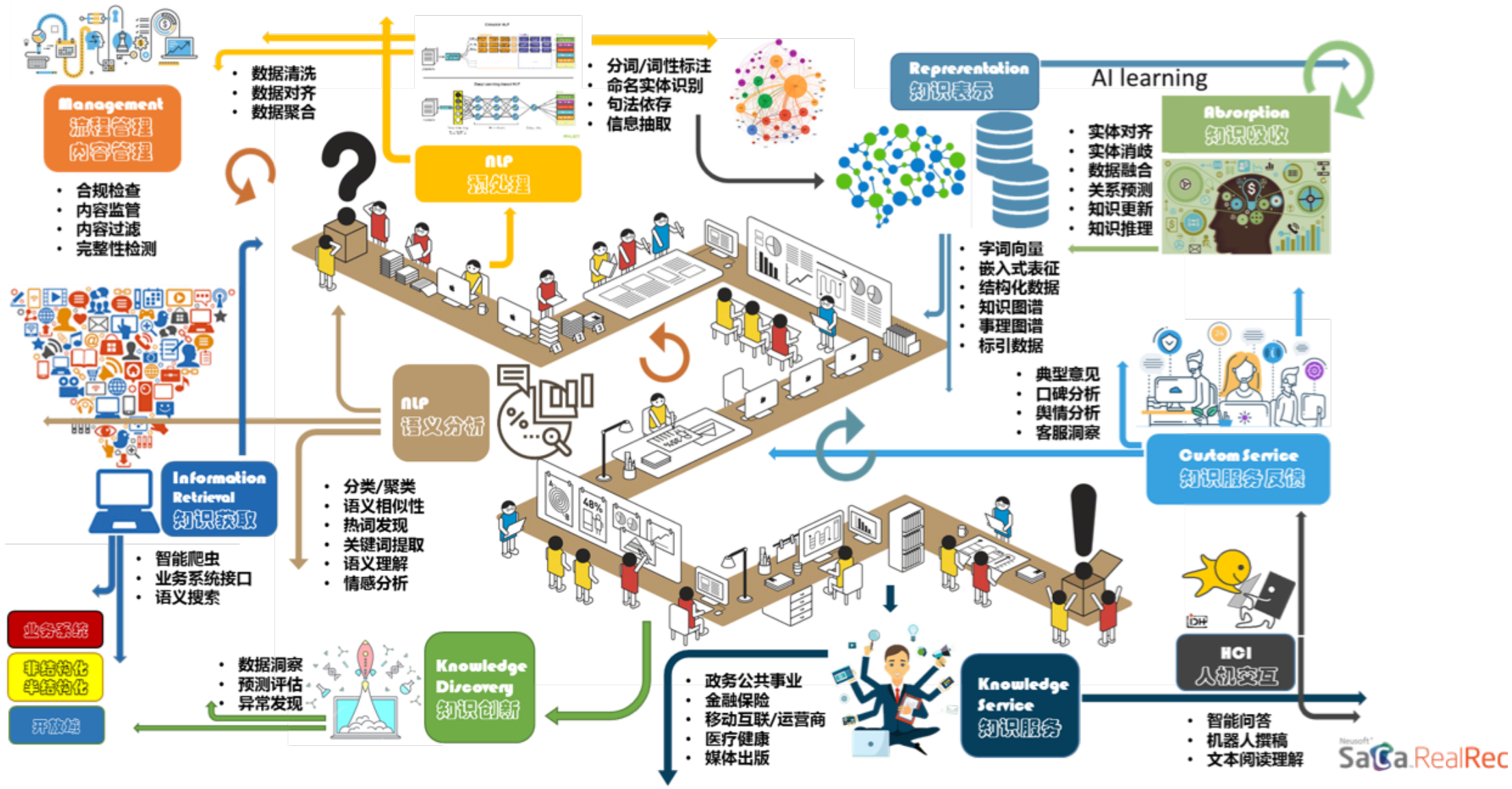
抽取

行业	数据	流程
• 金融 • 公安 • 教育 • 水产 • 工业 • 电商	• 真实度 • 数据字典 • 规模适度	• 数据采集 • 数据清洗 • 特征分析 • 模型训练 • 模型评估 • 预测服务

案例



自然语言处理 全流程综合实训

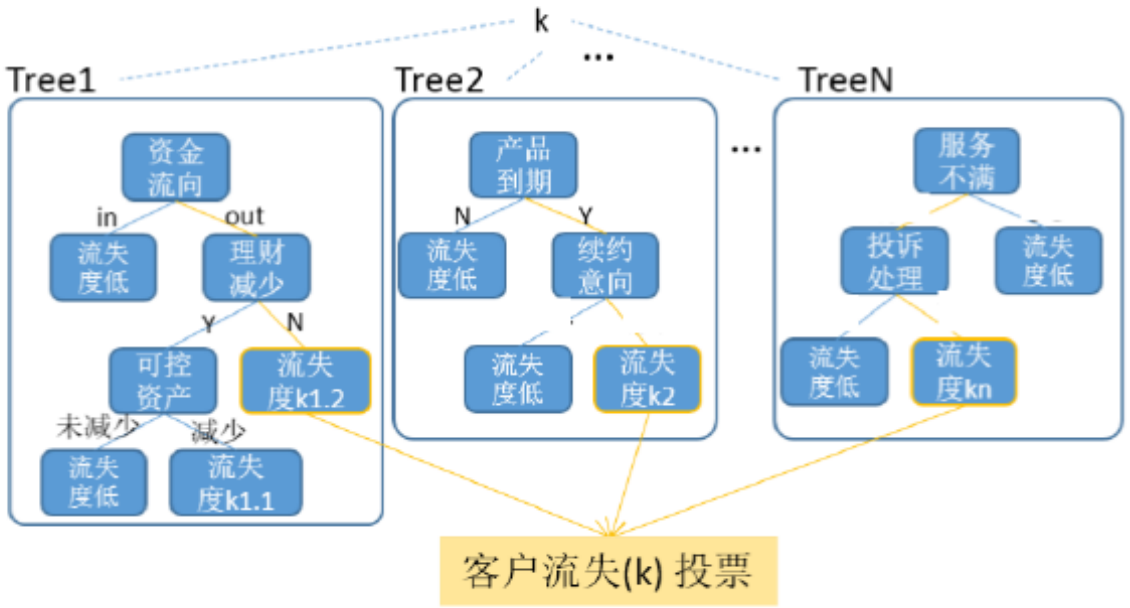


金融大数据

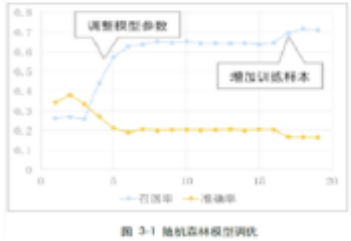
案例：银行高端客户流失预警

- 用户信息
- 存款
- 贷款
- 储蓄卡
- 信用卡
- 理财产品信息
- 客服记录信息

案例采用随机森林分析方法进行流失行为分析



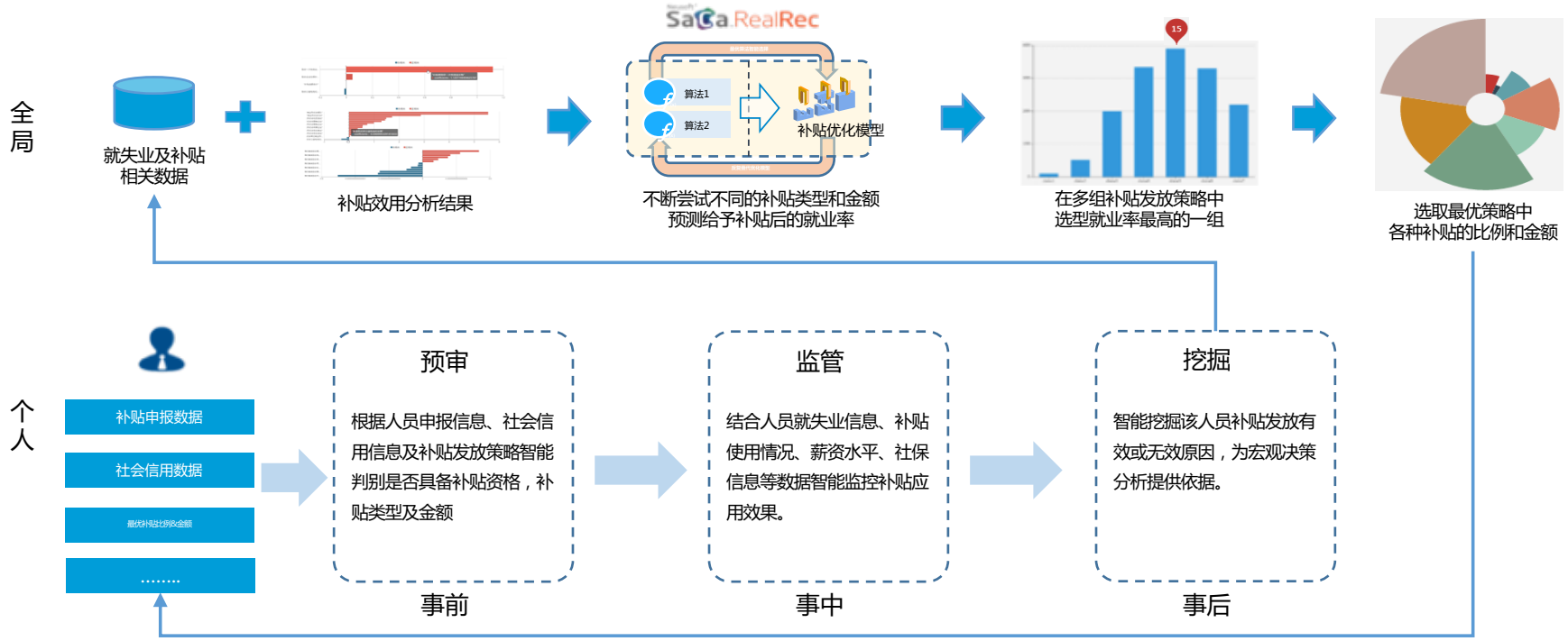
· 通过使用 某银行20万中高端用户的数据信息进行离线评估，进而实现客户流失预测71.2%的召回率。



· 银行方面对预测出的可能流失客户，制定有针对性的差异化产品、服务、营销活动。

政府大数据

案例：社保就业补助优化



- **背景需求**

湖北就业局每年失业补贴支出超过3亿，如何优化资金配置以最少的投入换取更高的就业率是核心问题。

- **解决方案**

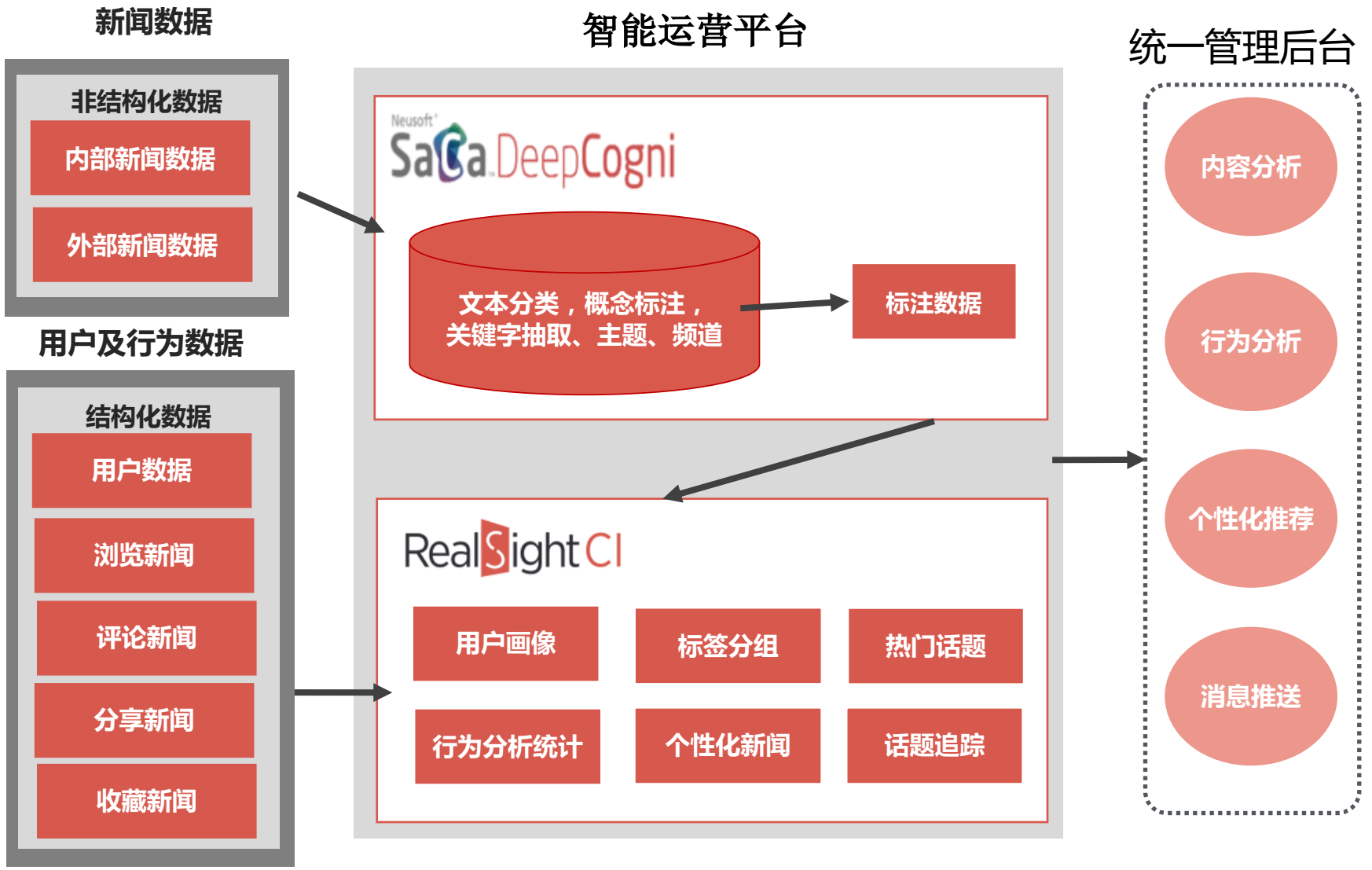
预测不同类别的人员在停发补贴后，不同时间段的就业变化趋势，挖掘在停发补贴后可以自主就业的人员特征，辅助决策不同时间段内可以停发或减少补贴的人员类别，优化补贴资源配置。

- **收益**

就失业预测准确率高于75%，并为补贴精准发放及补贴申请审核提供可靠依据。

媒体大数据

案例：央视新闻客户端数据分析平台



项目背景

央视新闻从2011年启动新媒体战略，需要符合互联网传播特征、利于用户互动的产品组合，带动新闻生产由电视媒介为主向多媒体平台的变革，使“央视新闻”成为新媒体领域具备强大舆论引导力和话语权的一流主流媒体。

解决方案

实施了客户智能行为分析、个性化推荐、精准营销，整合用户关注信息，进行精准化推荐；为用户提供个性化资讯内容，以提高用户使用的黏性。

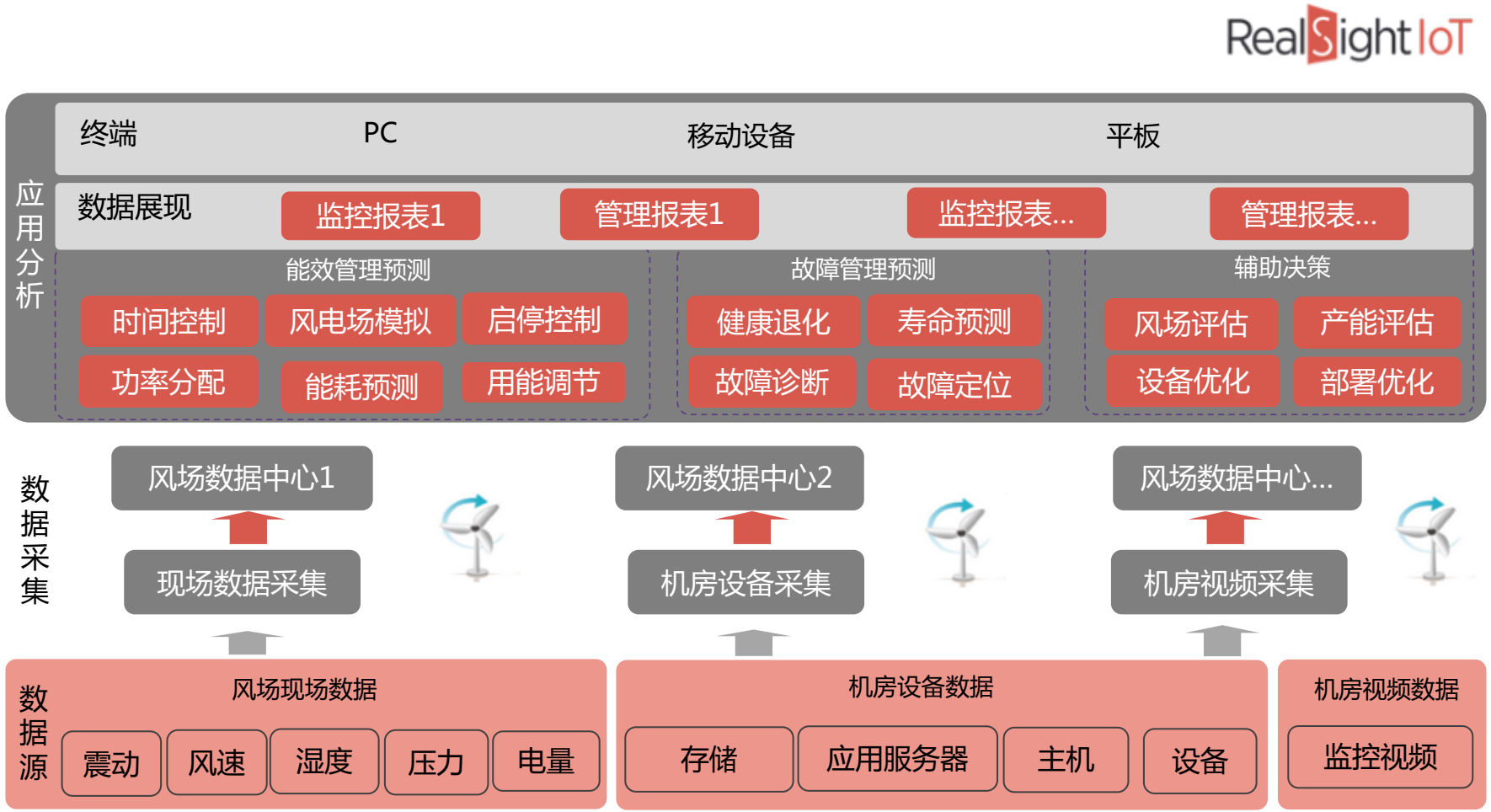
项目价值

媒体终端访问量提升20%以上。

Neusoft SaCa RealRec

工业大数据

案例：风力发电大数据平台



- **背景需求**

海量风机全生命周期运维及动态优化，提升风机发电量和节维风场运维成本。

- **解决方案**

风机能效预测、故障诊断与定位、风场能效优化。

- **收益**

为风场减少约10%的维保费用，年发电量提升5%-7%。

目录

CONTENTS

实训平台主要内容

数据挖掘分析平台

提供一站式数据挖掘模型构建与预测服务的高校教学平台。帮助高校提升构建智能应用的能力，降低复杂机器学习算法的学习成本。

实训案例库

基于东软20多年的垂直行业经验，提供丰富的案例资源，实现理论教学与现实应用的完美结合。

机器学习教材

提供面向大数据分析挖掘的丰富教材内容以及完整知识体系，配合丰富的教学形式，形成机器学习理论知识的最佳学习路径，帮助学生更快掌握并提升理论知识水平。

智能模型应用商店

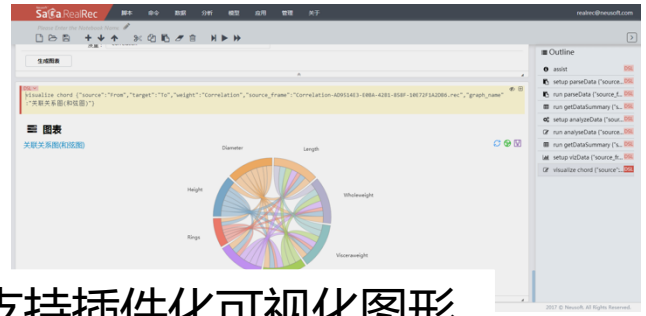
提供易掌握、易开发、易销售的创新创业平台。打造以运营管理商、第三方软件提供者和企业用户共同参与的模型生态系统。为企业和开发者提供交流与交易平台，帮助高校创业科研团队将学习成果直接应用到商业环境中。

开放平台

开放开发算法接口，开发可视化图形接口和相应的SDK，赋予学生创新创业的方法和能力，增强学生的动手能力和实践能力。

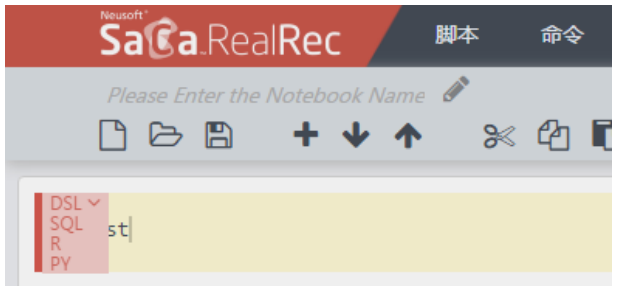
Neusoft SaCa RealRec

数据挖掘分析平台



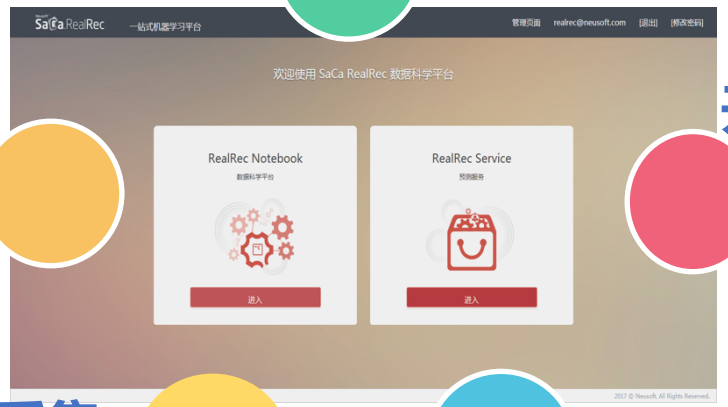
支持插件化可视化图形开发，增强前端开发实践能力

支持DSL，Python，R，SQL语言



多语言支持

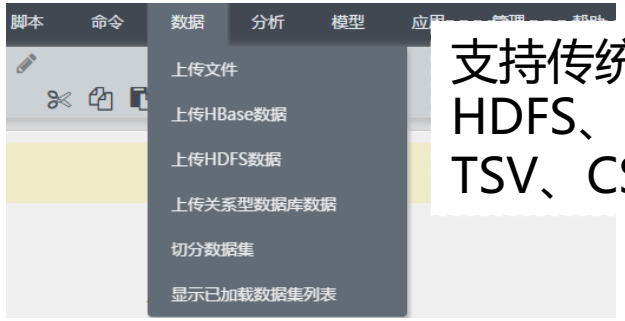
可视化



开放平台

支持算法二次开发，优化已有算法，扩展和丰富算法包

数据采集



支持传统的关系型数据库、HDFS、Hive、Hbase、TSV、CSV、Excel等数据源

创新大赛

平台支持H3C举办第四届全国高校云计算应用创新应用大赛

案例实训手册

- 水产品鲍鱼产量预测
- 贫困生识别
- 就业局就失业分析
- 客户流失预测
- 算法SDK

- 银行欺诈预测
- 预测电影票房
- 航食配餐
- 个性化推荐
- 风机预测性维护

01 案例均都来自**行业真实**的客户案例

02 案例分三级，**由易到难**逐步深入

03 案例精心设置引导性和课后**思考问题**

04 可视化、算法开发增强学生**实践能力**

Neusoft SaCa RealRec

体验丰富的机器学习算法

分类算法

Decision Tree
Random Forest
Linear Regression
Naïve Bayesian
SVM
Logistic Regression

聚类算法

K-Means
Streaming K-Means
Bisecting K-means
power iteration
Gaussian Mixture

深度学习

FeedForward Classification
FeedForward Regression
LSTM RNN
CNN
Deep Belief
RBM

文本分析

LDA
TFIDF
Word2Vec
Word Segmentation
Sentiment Analysis

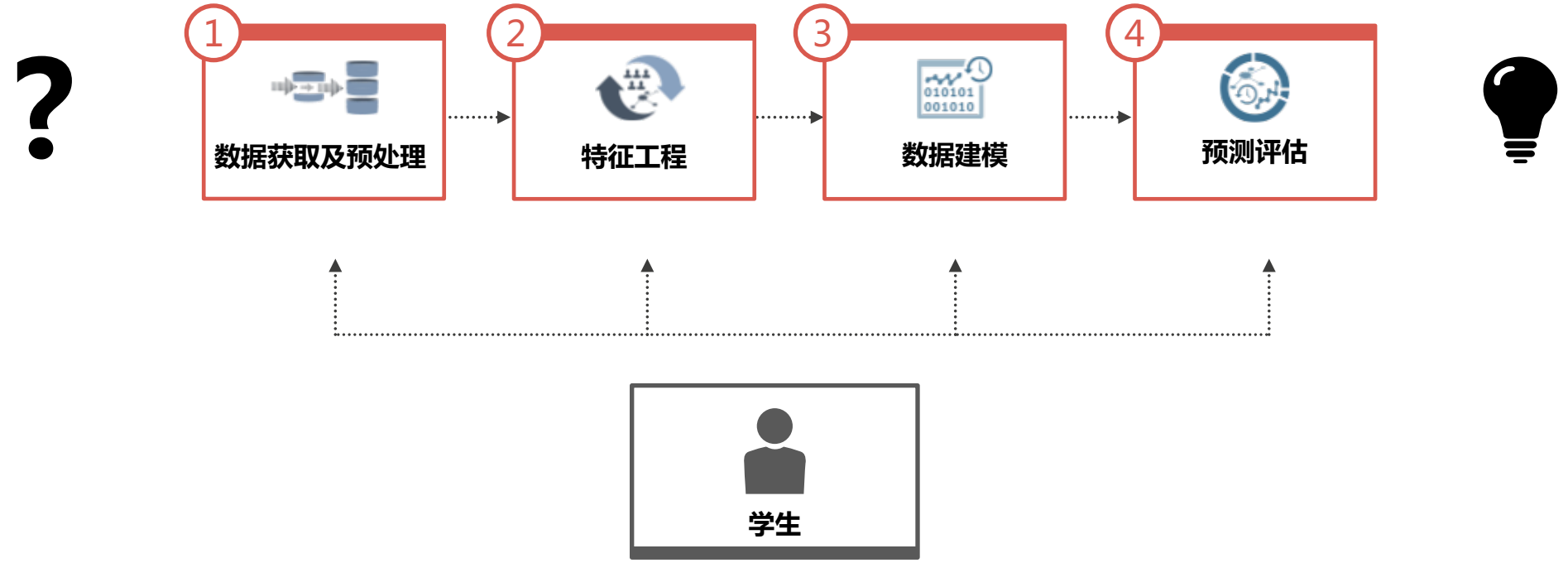
推荐算法

Factorization Based
Neighborhood Based
Popularity Based
CF
SlopeOne

图挖掘算法

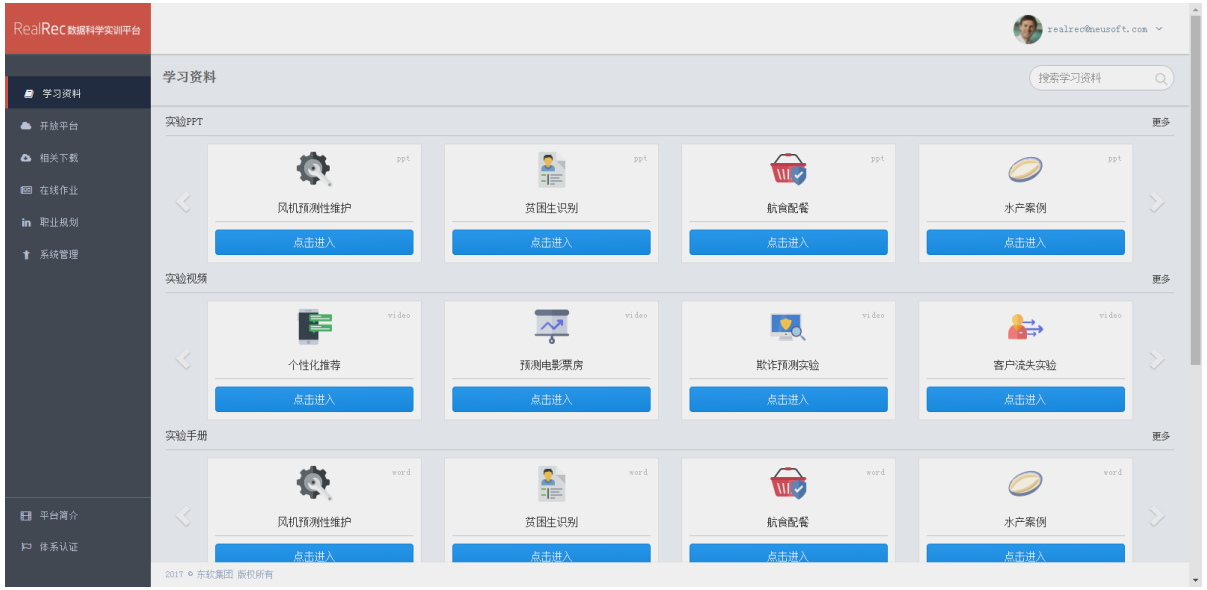
Page Ranking
Label Propagation
Connected Component
Triangle Count

数据挖掘全流程实践

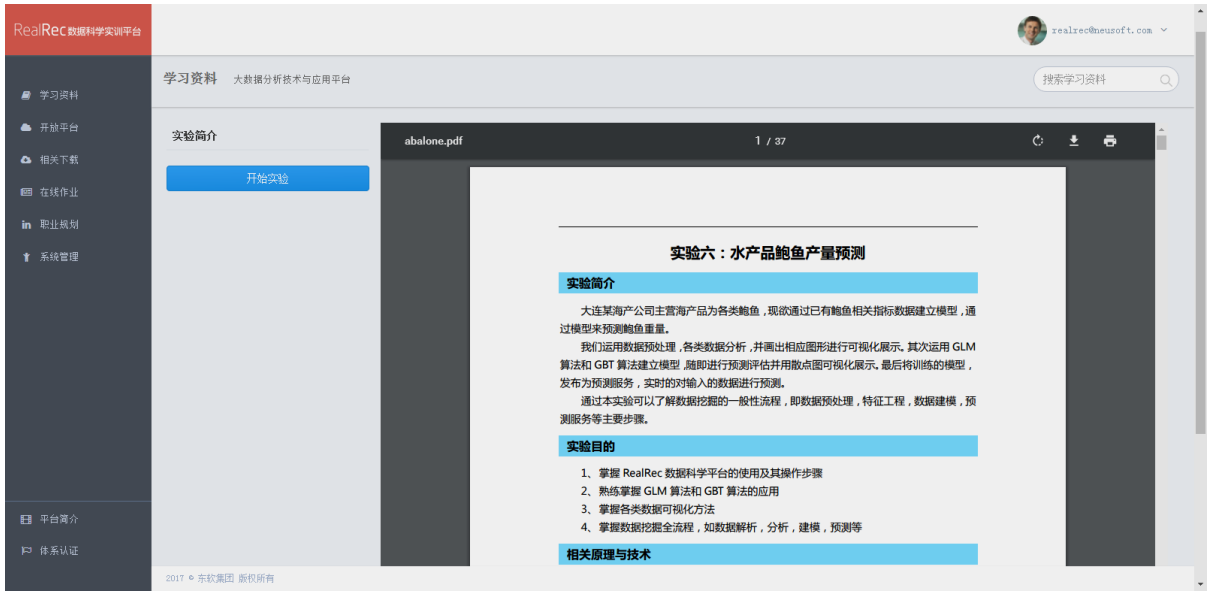


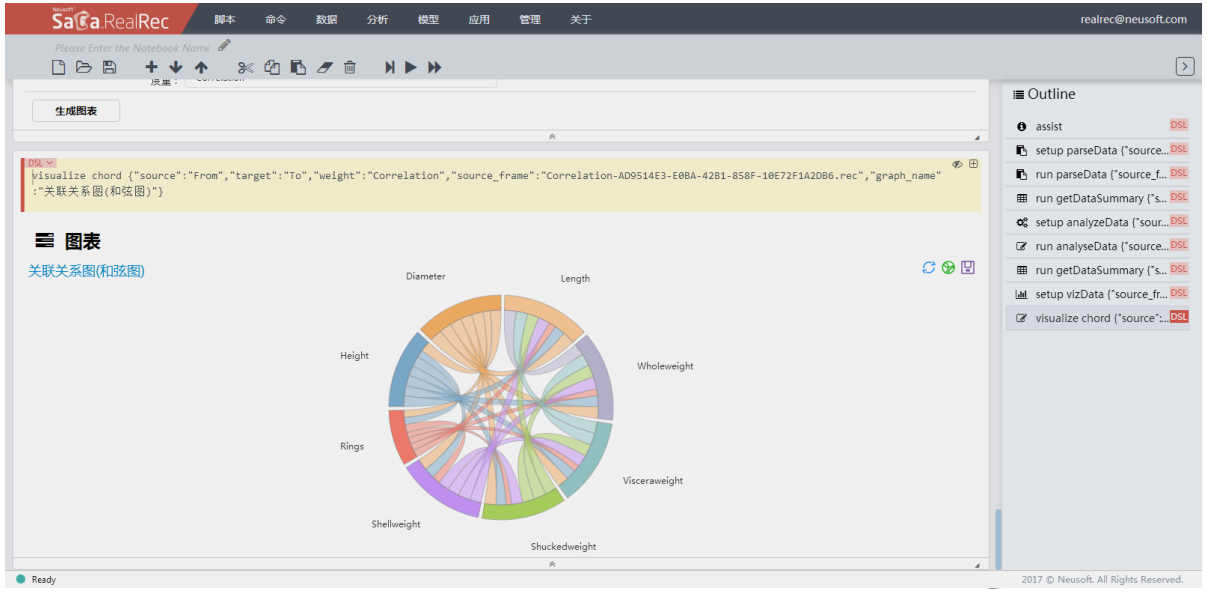
在一门课程当中，在一门实验当中，把大数据整个的从数据分析到数据挖掘，到数据的呈现这个流程都走一遍，这样对学生来讲，可以综合性的快速了解一个知识体系!

数据科学实训平台掠影

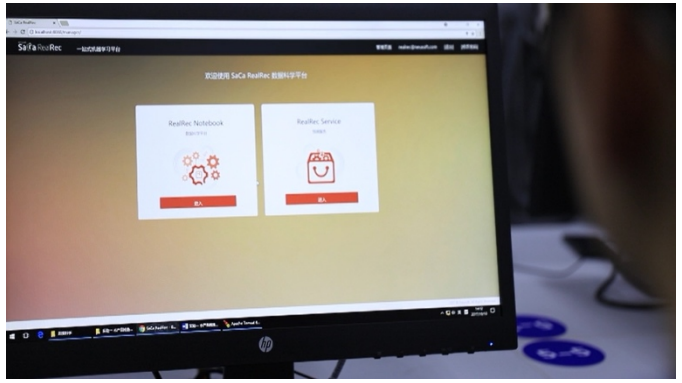

RealRec 数据科学实训平台
realrec@neusoft.com
学习资料
开放平台
相关下载
在线作业
职业规划
系统管理
平台简介
体系认证
学习资料
搜索学习资料
实验PPT
更多
风机预测性维护
贫困生识别
航食配餐
水产案例
点击进入
实验视频
个性化推荐
预测电影票房
欺诈预测实验
客户流失实验
实验手册
2017 © 东软集团 版权所有


SaCa.RealRec
一站式机器学习平台
管理页面
realrec@neusoft.com
[退出]
[修改密码]
欢迎使用 SaCa RealRec 数据科学平台
RealRec Notebook
数据科学平台
进入
RealRec Service
预测服务
进入
2017 © Neusoft. All Rights Reserved.


RealRec 数据科学实训平台
realrec@neusoft.com
学习资料
大数据分析技术与应用平台
搜索学习资料
实验简介
开始实验
abalone.pdf
1 / 37
实验六：水产品鲍鱼产量预测
实验简介
大连某海产公司主营海产品为各类鲍鱼，现欲通过已有鲍鱼相关指标数据建立模型，通过模型来预测鲍鱼重量。
我们运用数据预处理，各类数据分析，并画出相应图形进行可视化展示。其次运用 GLM 算法和 GBT 算法建立模型，随即进行预测评估并用散点图可视化展示。最后将训练的模型，发布为预测服务，实时的对输入的数据进行预测。
通过本实验可以了解数据挖掘的一般性流程，即数据预处理，特征工程，数据建模，预测服务等主要步骤。
实验目的
1、掌握 RealRec 数据科学平台的使用及其操作步骤
2、熟练掌握 GLM 算法和 GBT 算法的应用
3、掌握各类数据可视化方法
4、掌握数据挖掘全流程，如数据解析，分析，建模，预测等
相关原理与技术
2017 © 东软集团 版权所有


SaCa.RealRec
脚本
命令
数据
分析
模型
应用
管理
关于
realrec@neusoft.com
生成图表
Outline
assist
setup parseData {"source...
run parseData {"source_f...
run getDataSummary {"s...
setup analyzeData {"sour...
run analyseData {"source...
run getDataSummary {"s...
setup vizData {"source_fr...
visualize chord {"source":...
visualize chord {"source":"From","target":"To","weight":"Correlation","source_frame":"Correlation-AD9514E3-E0BA-42B1-858F-10E72F1A2DB6.rec","graph_name":"关联关系图(和弦图)"}
图表
关联关系图(和弦图)
Diameter
Length
Height
Wholeweight
Rings
Visceraweight
Shellweight
Shuckedweight
Ready
2017 © Neusoft. All Rights Reserved.

大连理工大学综合实训成果



- 大连理工计算机学院针对学院的教学需求与特点，整合东软集团先进技术与产品资源推进教学改革，通过实验性引入SaCa RealRec数据科学平台为大数据人才的培养做好充足准备。
- 充分利用数据科学平台的交互性建模能力、配套教材、案例库、模型商店等，提供师生多样的教学内容，灵活的教学方式。
- 校方高度评价：“东软SaCa RealRec数据科学平台能满足大工**教学体系和科研体系**的实际需求，基于实际案例的操作实践对学生的学习和创新创业非常有价值！”

联合竞赛

03 智能检索和推荐

将相关模型发布线上服务，当求职者输入学历、专业等信息后，智能推荐相关职位

02 网络数据分析

解析分析数据，进行行业招聘态势分析，运用相关算法得到有价值的结论

01 网络爬虫

根据选定的招聘网站列表进行数据爬取，并将数据进行整理和存储

第四届全国高校云计算应用创新大赛报名正式开放

2017-09-09 大赛组委会 全国高校云计算应用创新大赛

第四届全国高校云计算应用创新大赛(https://cloud.seu.edu.cn/contest)的报名已经正式开放，本次大赛设置了丰富多样的命题，包括云上**创意、Spark编程、VR编程、容器技术和深度学习**等，相信广大选手都能在比赛过程中学习到云计算和相关领域内最新最流行的技术，找到展示自我，放飞创意的舞台！

本届大赛在前几届的基础上，听取了广大参赛选手的意见，进行了多项改革，也为参赛选手争取到了更多福利：

1. **决赛入围名额大幅提高！**未能入围决赛的选手也将获得分赛区的奖项
2. **总奖金超过10万元！**获奖队伍同时有机会获得50万元创业配套资金
3. **报名就送腾讯云（http://cloud.tencent.com）代金券！**体验真实的云计算平台和应用
4. **作品评审分组进行！**大赛评审时分研究生组和本/专科组分别进行，确保公平的获奖机会
5. **ACM大咖面对面！**获奖队伍将有机会受邀参加ACM中国理事会的各类高端学术活动

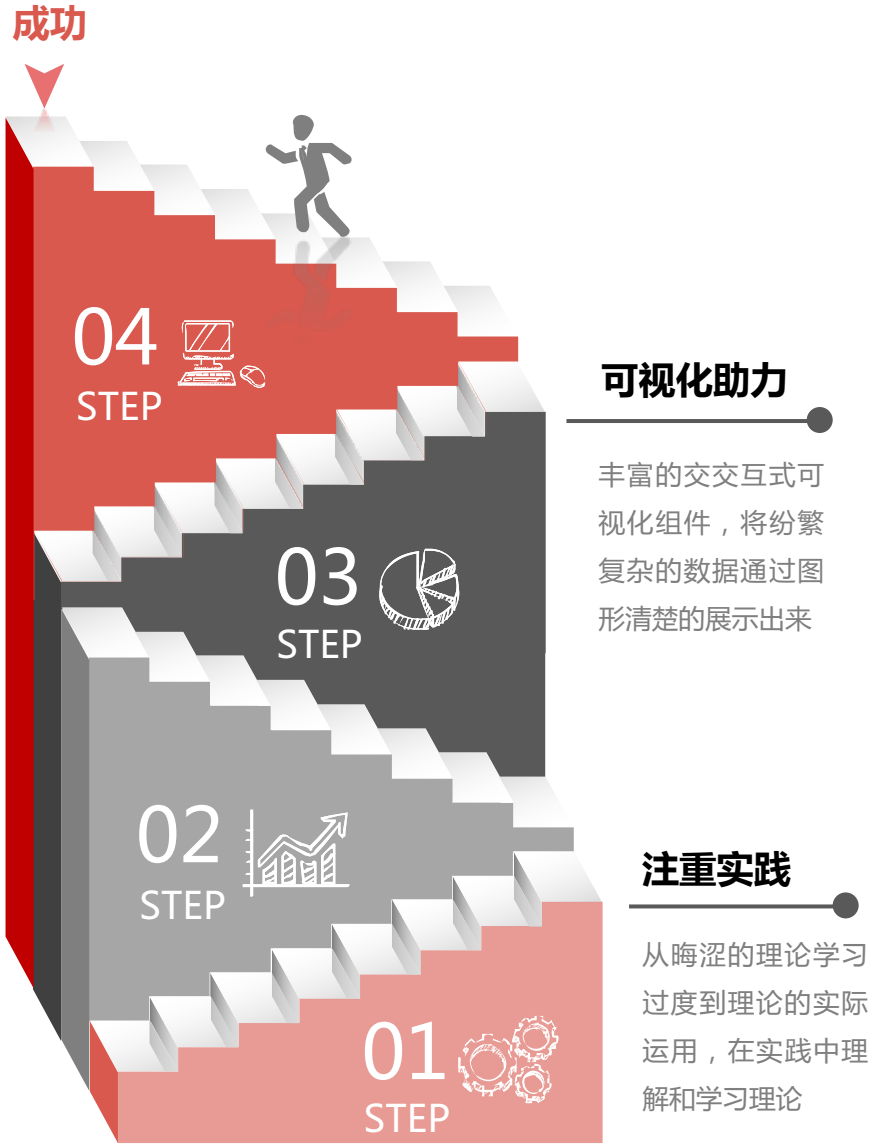
学生心得体会

掌握算法

通过丰富的案例实践，了解和学习丰富的各类算法，掌握不同业务场景下运用何种算法

引导思考

通过实践实验，理清思路，查找自身的不足，了解行业企业的实际应用场景

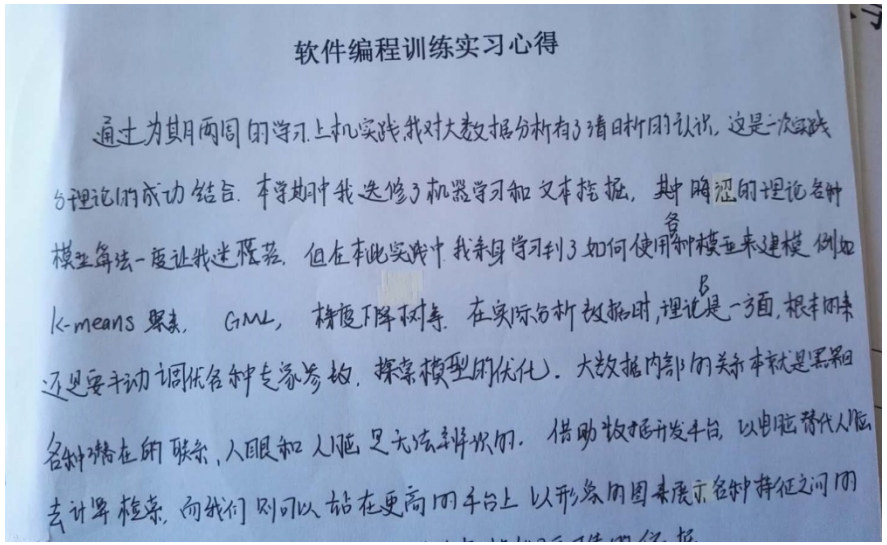


可视化助力

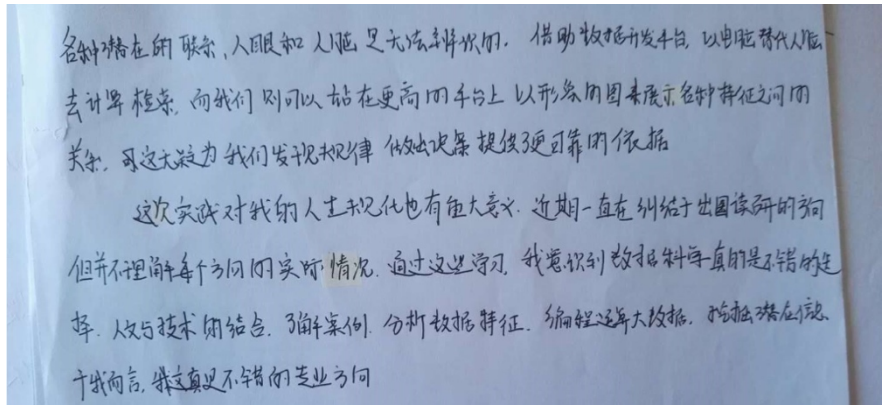
丰富的交交互式可视化组件，将纷繁复杂的数据通过图形清楚的展示出来

注重实践

从晦涩的理论学习过度到理论的实际运用，在实践中理解和学习理论


软件编程训练实习心得

通过为期两周的学习上机实践，我对大数据分析有了清晰的认识，这是一次实践与理论的成功结合。本学期中我选修了机器学习和文本挖掘，其中晦涩的理论各种模型算法一度让我迷茫。但在本此实践中我亲身学习到了如何使用各种模型来建模，例如k-means 聚类，GMM，特征降维对等。在实际分析数据时，理论是一方面，根本的还是要手动调优各种专家参数，探索模型的优化。大数据内部的关系本就是某种各种潜在的联系，人眼和人脑是无法辨识的。借助数据开发平台，以电脑替代人脑去计算探索，而我们则可以站在更高的平台上以形象的图来展示各种特征之间的关系，可这大量为我们发现规律，做出决策提供更可靠的依据


各种潜在的联系，人眼和人脑是无法辨识的。借助数据开发平台，以电脑替代人脑去计算探索，而我们则可以站在更高的平台上以形象的图来展示各种特征之间的关系，可这大量为我们发现规律做出决策提供更可靠的依据

这次实践对我的人生和文化也有重大意义。近期一直在纠结于出国读研的问但并不理解每个方向的实际情况。通过这次学习，我意识到数据科学一直的是不错的选择。人文与技术的结合，分析案例，分析数据特征，编程处理大数据，对我而言，我这真是不错的专业方向

Neusoft

Beyond Technology

Neusoft Saca.RealRec