设置“数据科学与大数据技术”专业的规划及设想

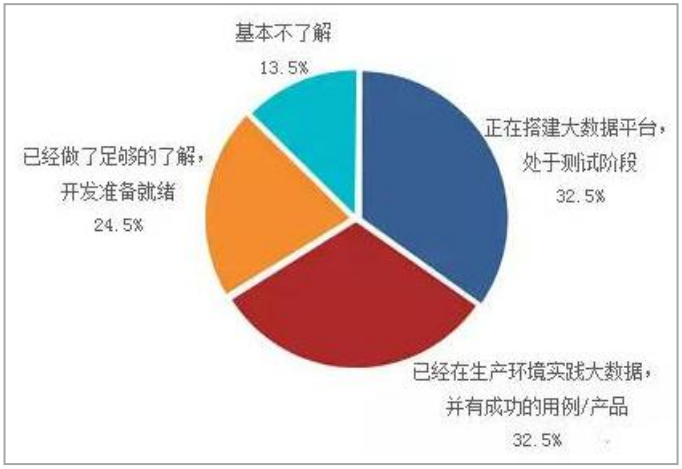
东北石油大学计算机与信息技术学院 文必龙
bilong_wen@126.com, 13206811009

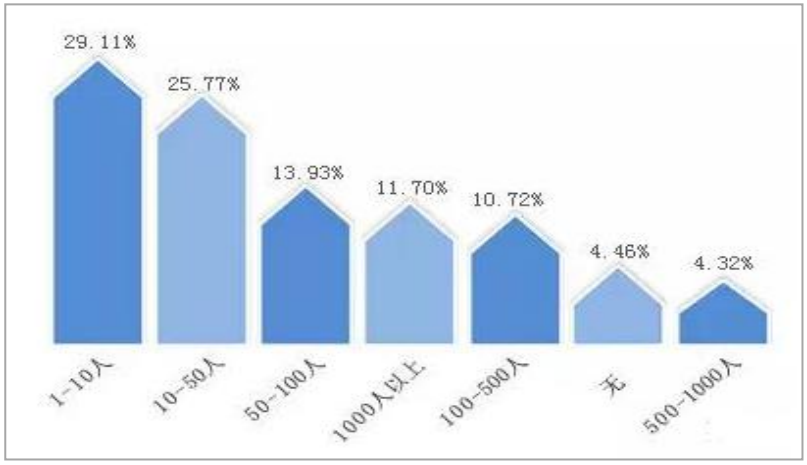
提纲

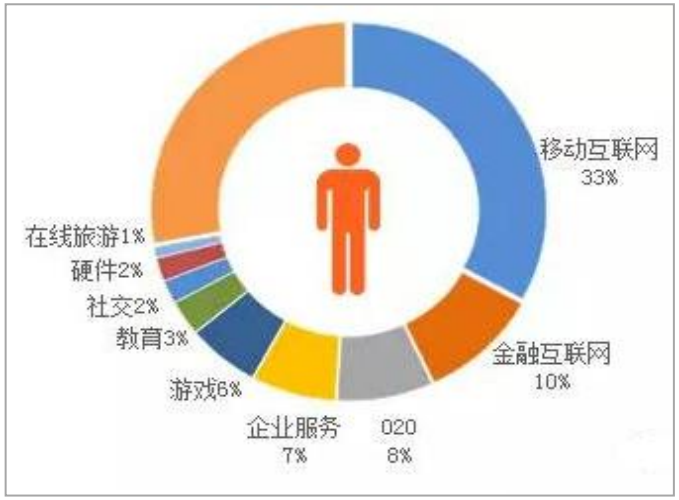
- 大数据人才培养现状
- 大数据技术基础
- 专业建设规划
- 开设大数据专业存在的问题

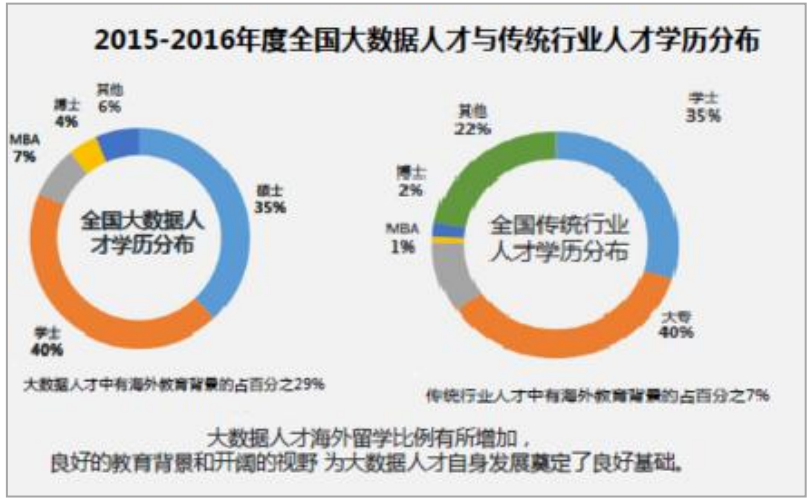
大数据人才需求调查

据中国数据分析行业网，对大数据平台有需求的公司的调查，目前全国的大数据人才只有46万，未来3-5年内大数据人才的缺口将高达150万多，大数据行业将面临全球性的人才荒。 领英发布的《2016年中国最热职位人才报告》基于领英平台上约50万的中国各个行业人才大数据的分析，报告表明，数据分析人才最为稀缺。









大数据人才类型

大数据最关键的部分的数据分析和挖掘数据价值，要获得这些，就需要大量的数据学科家，目前最为欠缺的数据人才主要有两种：数据处理人才和数据分析人才。

数据处理人才主要是从统计学、信息技术、软件工程领域诞生，主要负责数据处理的全过程，即数据的获取、存储、清洗、加工、建模、传输等。

数据分析人才主要诞生区域与前者相似，主要负责对大数据进行价值挖掘，包括对数据统计结果的甄别与分析，对数据分析结果的评估与展示，对用户数据需求的判断与反馈。

社会培训机构大数据技术培训内容

1. 数学基础和机器学习入门
2. Weka机器学习软件包
3. 深度学习基础：人工神经网络
4. 基于神经网络的词表示学习方法
5. 卷积神经网络基本原理
6. 循环神经网络（RNN）
7. 卷积神经网络文本情感分类、关系抽取
8. 循环神经网络分析语言模型、文本分类等

第一天
Hadoop与Spark大数据构架概述及案例简介
1 介绍Hadoop与Spark大数据层级构架
2 Hadoop与Spark区别与联系、定位
3 Spark生态系统概述以及版本演化
Hadoop、Spark安装部署

第二天
Spark程序设计实例
1 Scala语言基础
2 Spark程序设计方法
3 Spark程序设计实例
4 上机实验
1）日志统计
2）数据分析
Spark案例分析

第三天
Spark Streaming应用及案例分析
Spark SQL
MLlib
GraphX

学员对社会培训机构的选择要点

1. 个人基础情况

不是所有的人都适合花钱学大数据。个人技术背景比较好，自学能力强。接受新东西快，愿意动手。

2. 课程内容

每家机构都说自己的课程内容怎么好怎么全，培训大纲从hadoop到storm， 到spark， 到云计算，到openstack 到docker

3. 师资情况

大数据这一块目前还很少有公认的名师。毕竟大数据在国内发展起来没几年，有大数据经验的技术大咖都在一线搞技术，薪资高机遇好，没几个人愿意跑去做培训。老师的实际大数据经验并不多。

4. 教学服务

主要指答疑服务，不是指态度和待遇。高手点拨，手把手指导，除了老师之外还有一些在企业中做大数据工作的兼职人帮忙解决问题，QQ远程解决问题。

5. 学习环境

节约时间，不限定时间，集中时间。硬件不卡，软件不缺。系统全面，可承受破坏性操作。

6. 价格/性价比

学生对大数据的热情

1. 招生

专业简介中，提到大数据方向的，受到追捧。

2. 课外

本科生到实验室，都愿意学习与大数据相关的内容

课下咨询有关大数据内容的很多

假期花钱报名参加大数据技术培训班

3. 考研

踊跃选择大数据相关的导师

各高校积极申办大数据专业

缘于大数据时代催生的大量相关人才缺口，各大高校正紧锣密鼓启动大数据人才培养。2017年3月，教育部公布已有35所高校获批“数据科学与大数据技术”专业，站在互联网“风口”上的大数据，直接催热了大数据专业。

北京大学
对外经济贸易大学
中南大学

中国人民大学	宿州学院	贵州师范大学
北京邮电大学	福建工程学院	安顺学院
复旦大学	黄河科技学院	贵州商学院
华东师范大学	湖北经济学院	贵州理工学院
电子科技大学	佛山科学技术学院	昆明理工大学
北京信息科技大学	广东白云学院	云南师范大学
中北大学	北京师范大学-香港浸会大学联合国际学院	云南财经大学
晋中学院	广西科技大学	宁夏理工学院
长春理工大学	重庆理工大学	
上海工程技术大学	成都东软学院	
上海纽约大学	电子科技大学成都学院	
浙江财经大学	贵州大学	

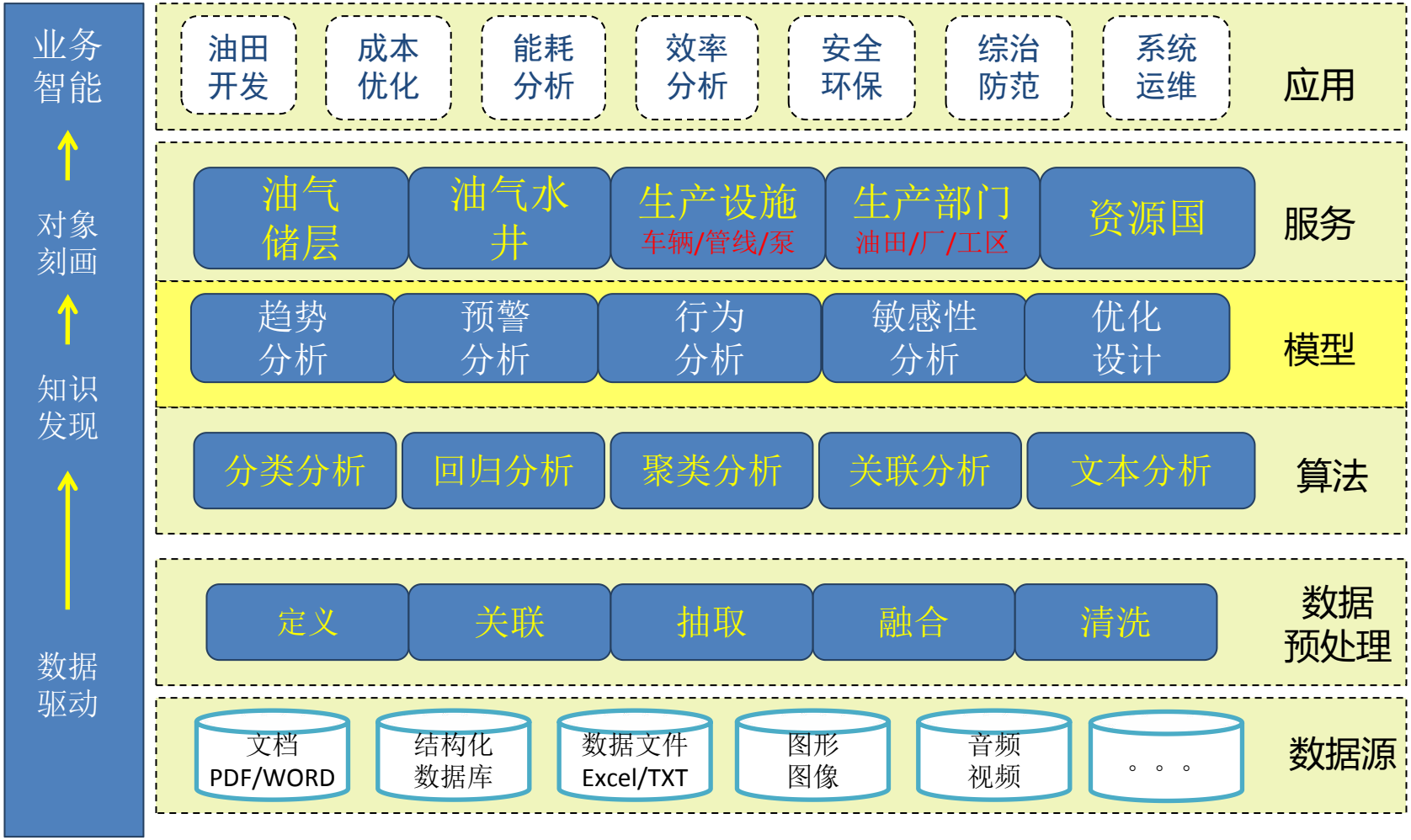
提纲

- 大数据人才培养现状
- 大数据技术基础
- 专业建设规划
- 开设大数据专业存在的问题

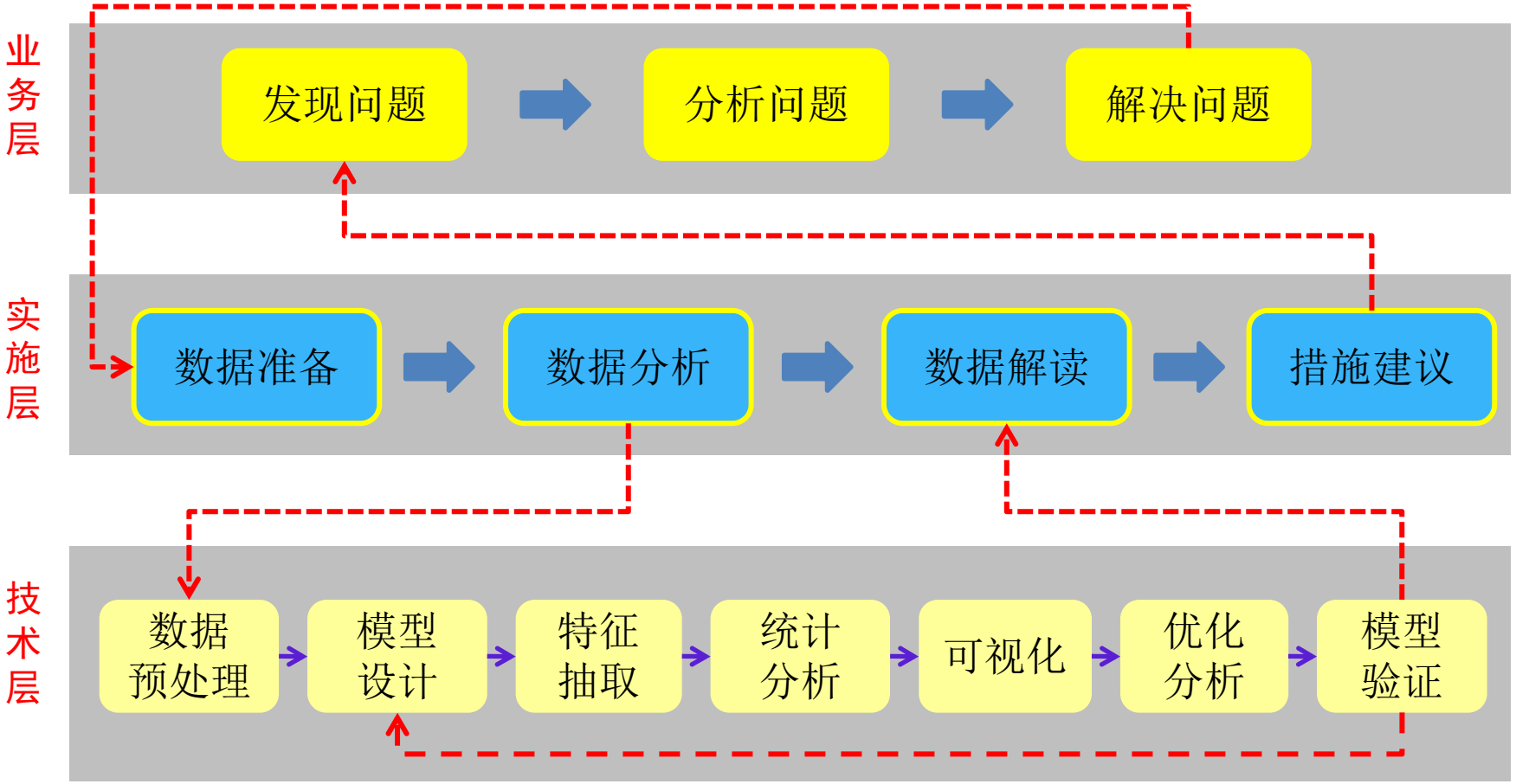
我院大数据相关的研究

应用研究	油田拉油 大数据分析	油田设备大 数据分析	基于大数据 的试井解释	油气集输 风险评估	智能油 田开发
方法研究	生物医学信息 抽取与分析	多媒体 语义标注	GIS空间 语义信息搜索	跨尺度多源 图像数据融合	
算法研究	智能分析 算法库	支持向量机 模型研究	神经网络 算法研究		
数据集成	油田信息 搜索引擎	课程知识 体系构建	油田数据 资源池	数据 标准化	
系统平台	基于Hadoop的实时 大数据平台	基于Oracle的 大数据平台	地震资料处理系统 海量数据管理平台		

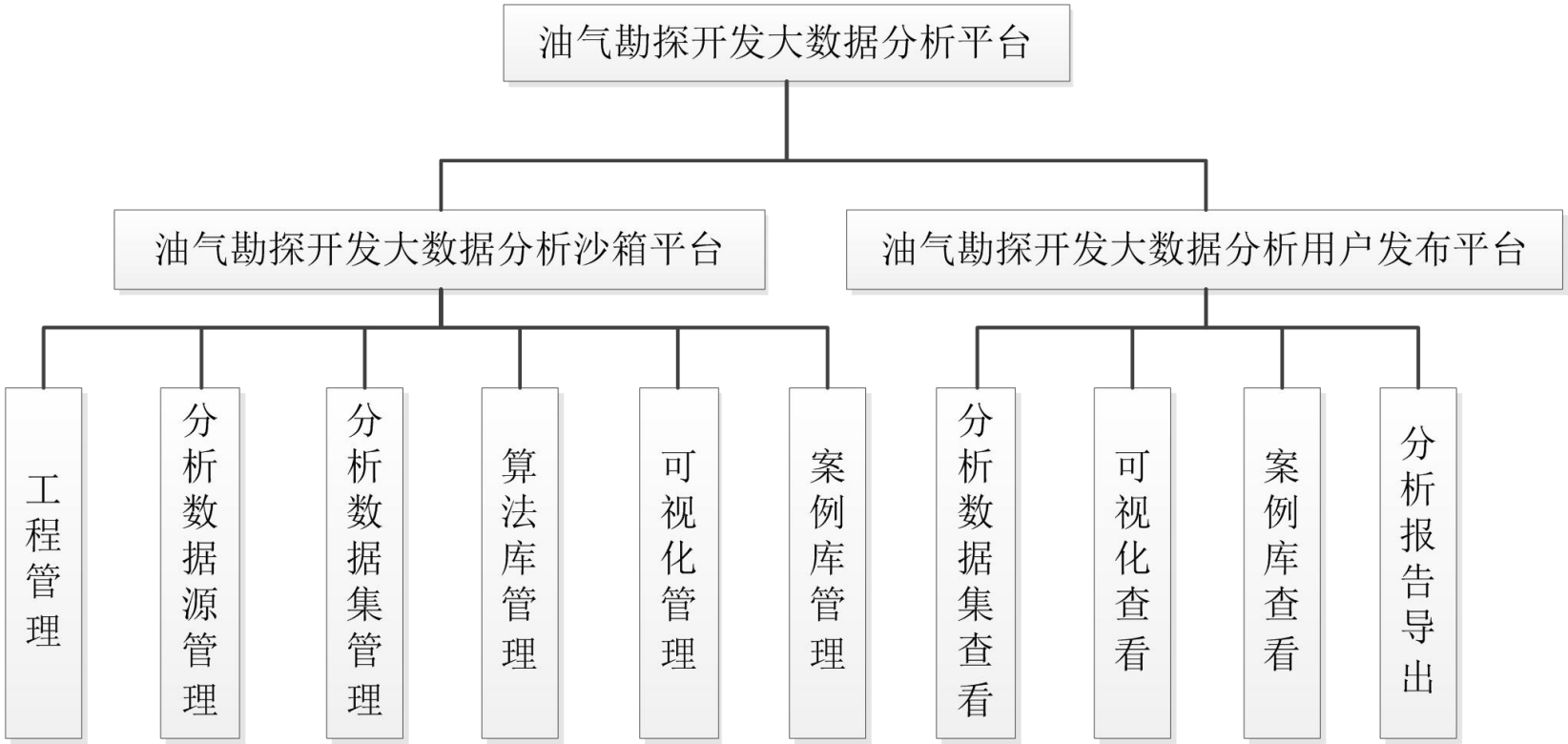
油田大数据应用体系结构



油田大数据分析模式

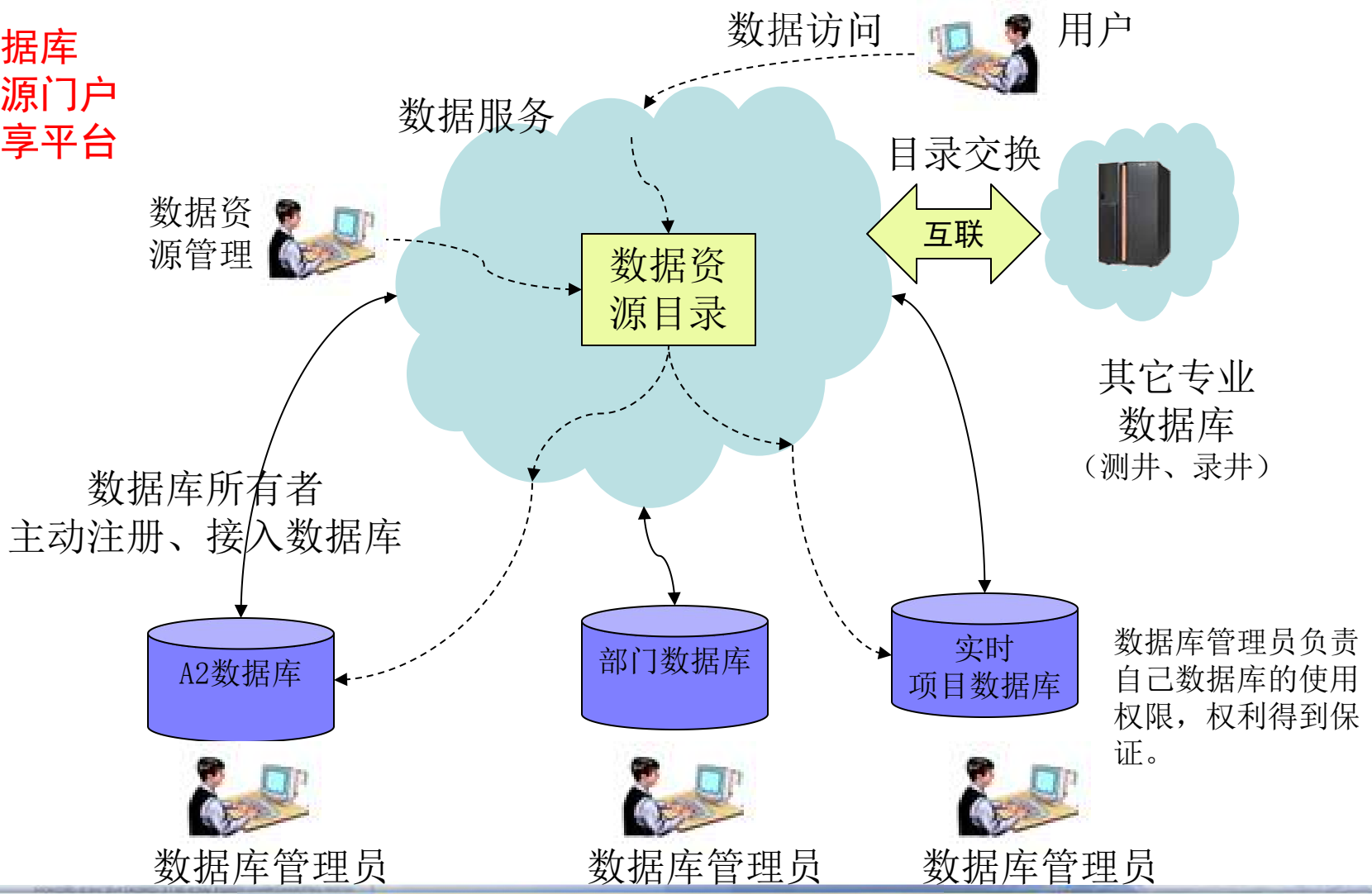

业务层
发现问题
分析问题
解决问题
实施层
数据准备
数据分析
数据解读
措施建议
技术层
数据预处理
模型设计
特征抽取
统计分析
可视化
优化分析
模型验证

勘探开发大数据分析开台EPBDAP

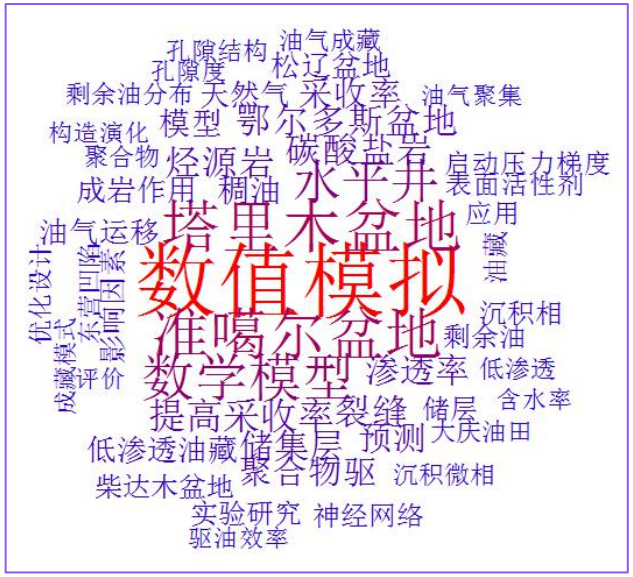


基于数据资源目录的大数据资源池

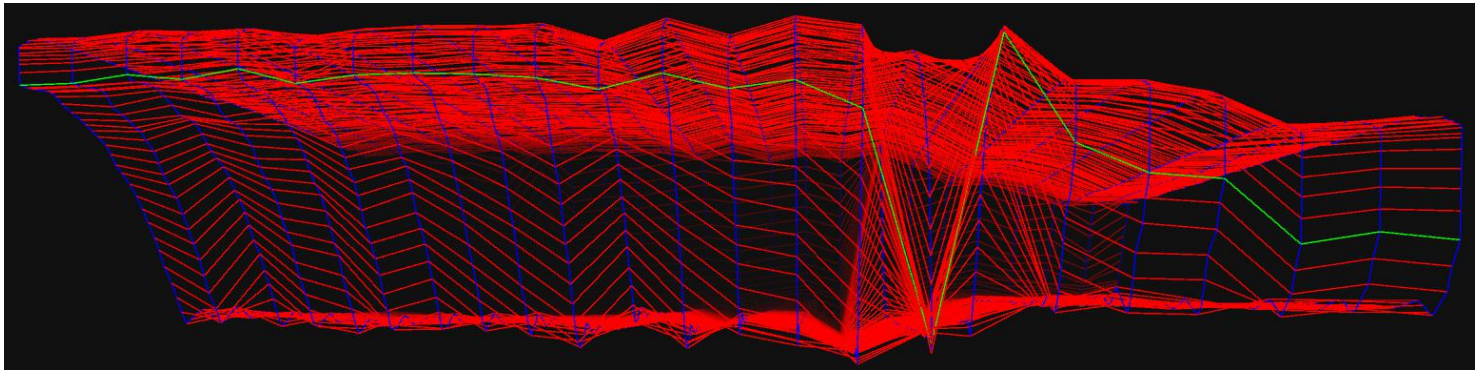
虚拟数据库
数据资源门户
数据共享平台



数据可视化技术

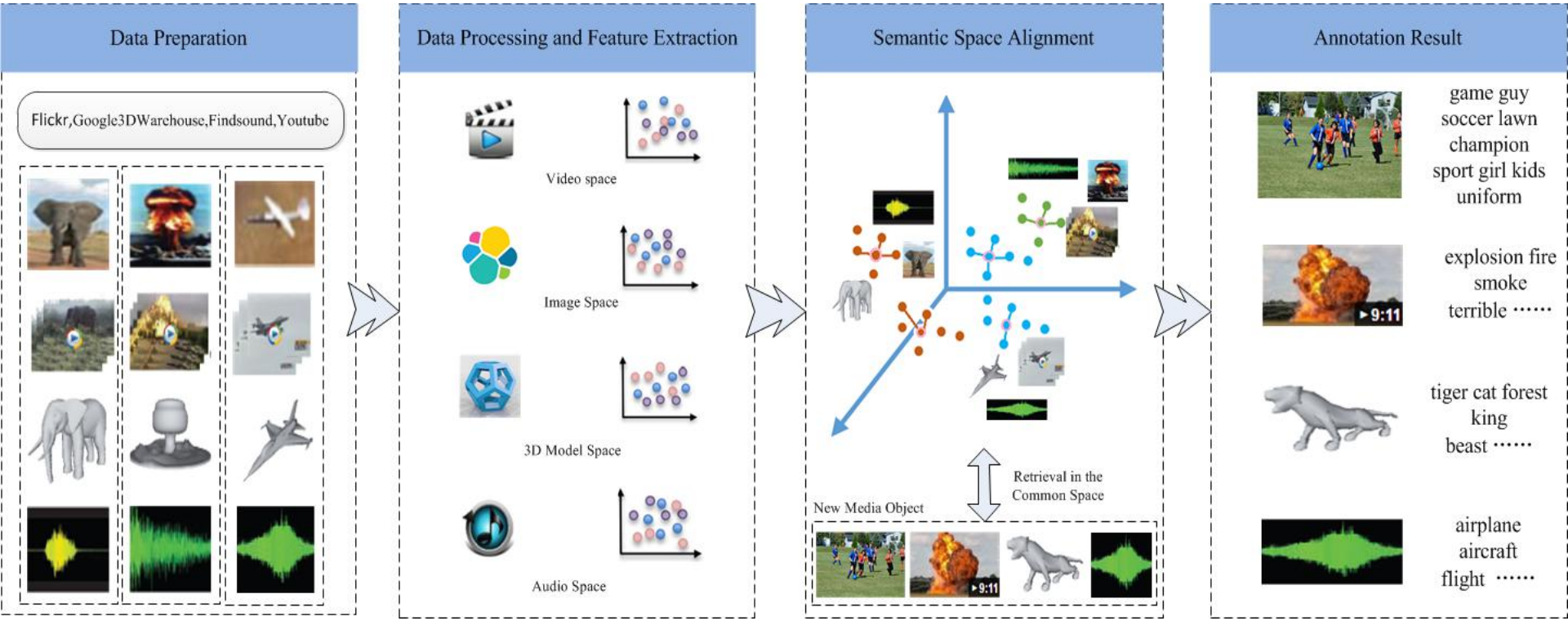

油气成藏
孔隙结构
松辽盆地
孔隙度
剩余油分布
天然气
采收率
油气聚集
构造演化
模型
鄂尔多斯盆地
聚合物
烃源岩
碳酸盐岩
启动压力梯度
成岩作用
稠油
水平井
表面活性剂
塔里木盆地
应用
油气运移
油藏
数值模拟
优化设计
沉积相
准噶尔盆地
剩余油
数学模型
渗透率
低渗透
提高采收率
裂缝
储层
含水率
低渗透油藏
储集层
预测
大庆油田
聚合物驱
沉积微相
柴达木盆地
实验研究
神经网络
驱油效率


沙河村
肃宁县官厅中学
官亭村
柳家庄村
崔家庄村小闾庄村
李庄村
郑庄村
大王庄村
大闾庄村
大迁民庄村
东河庄村
索庄村
吕家庄村
南于庄村
岳家庄村
东风庄村
镇海寺
泗水岸村
河头店村
魏娄村
小张庄村
豆闾庄村
留名佛村
闾郝村
杜王郭村
方周村
晁固村
贾刘店村
李举村
张村
山秋村

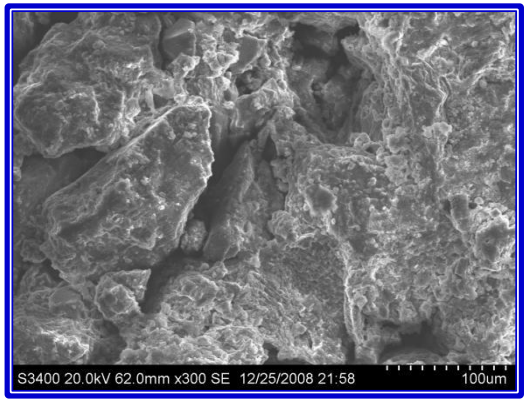


多媒体语义标注

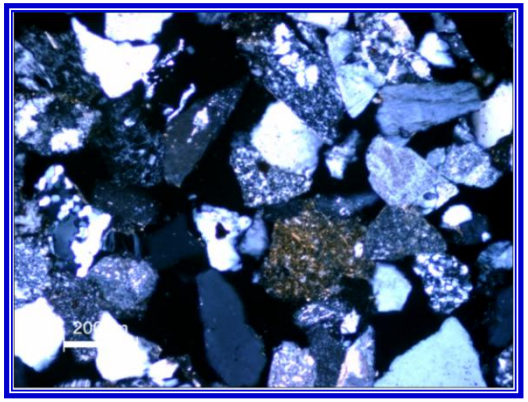
通过对图像、视频、音频、三维模型等多媒体数据的内容识别，实现其文本语义标签的自动生成。



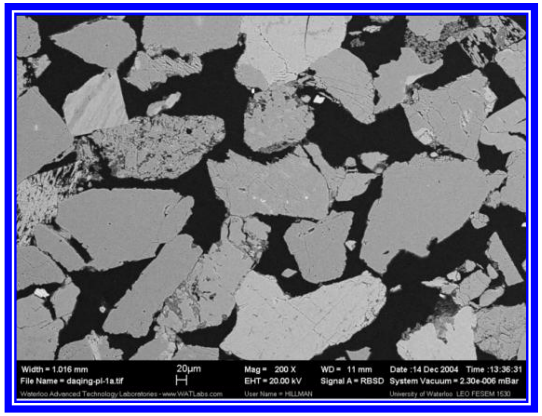
致密油储层孔隙结构跨尺度多源融合、重构、识别



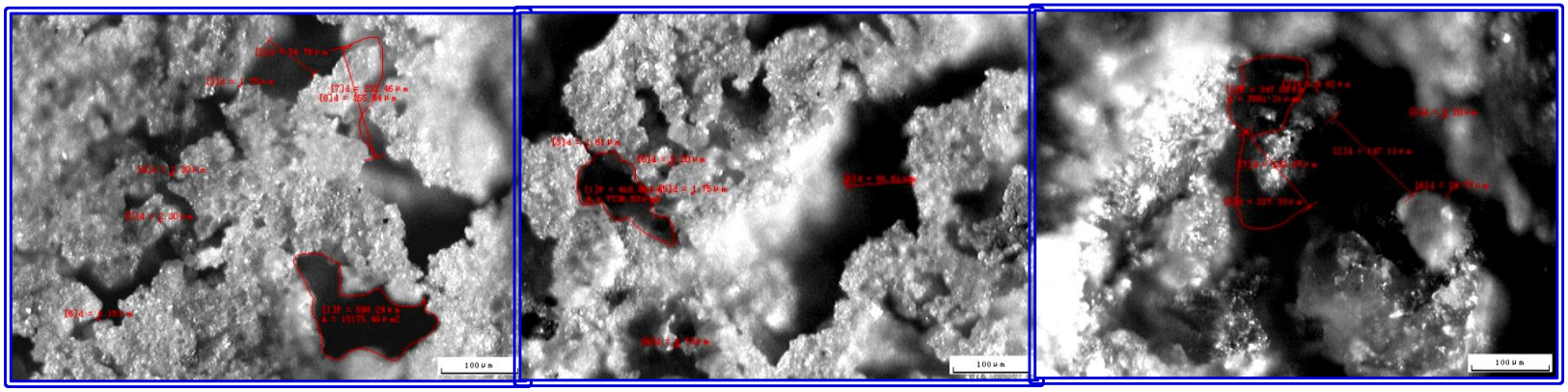
扫描电镜图



岩芯样品

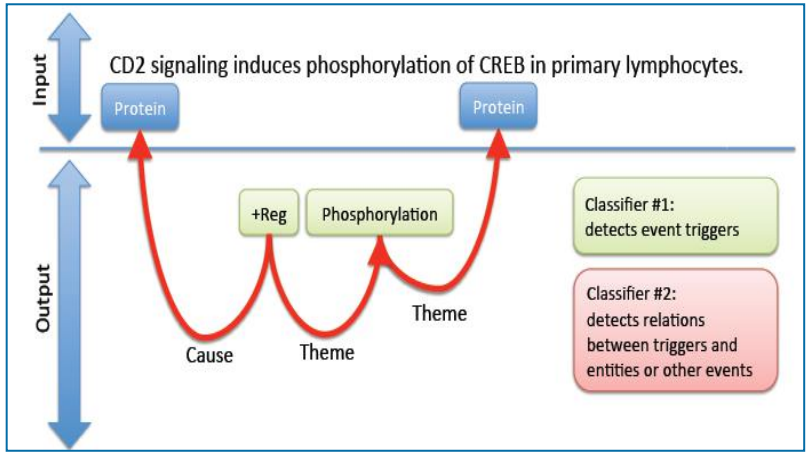


电子扫描BSEM图像

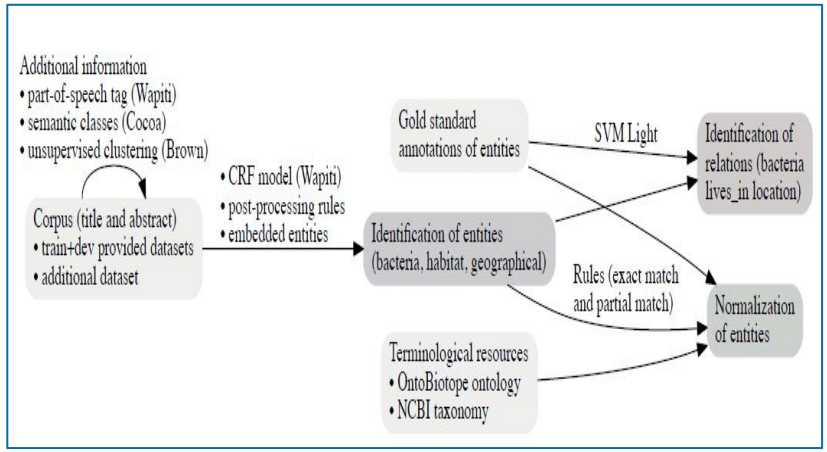


生物医学文本信息抽取

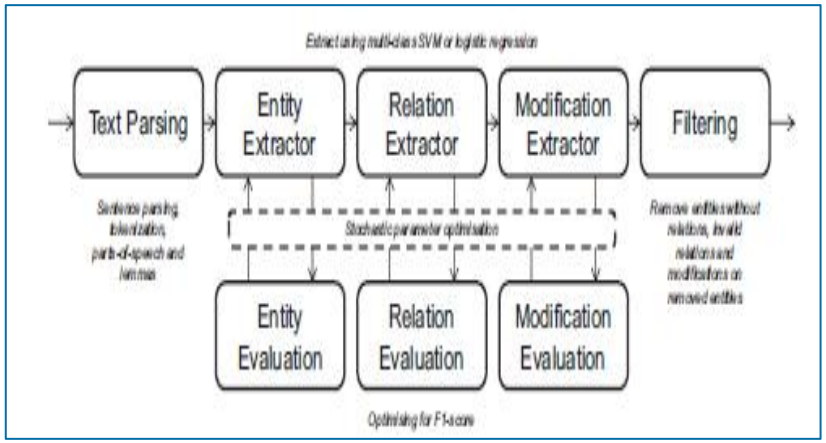
对生物医学文本中的实体、关系、事件自动识别



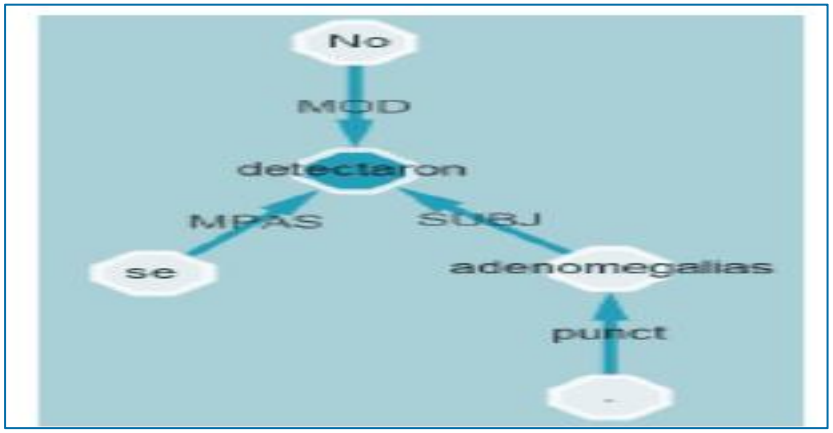
生物医学文本命名实体及关系的识别



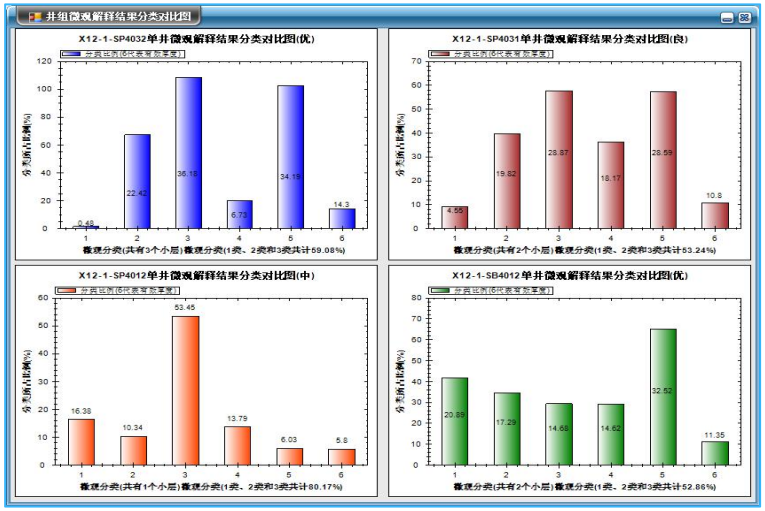
信息抽取和评价

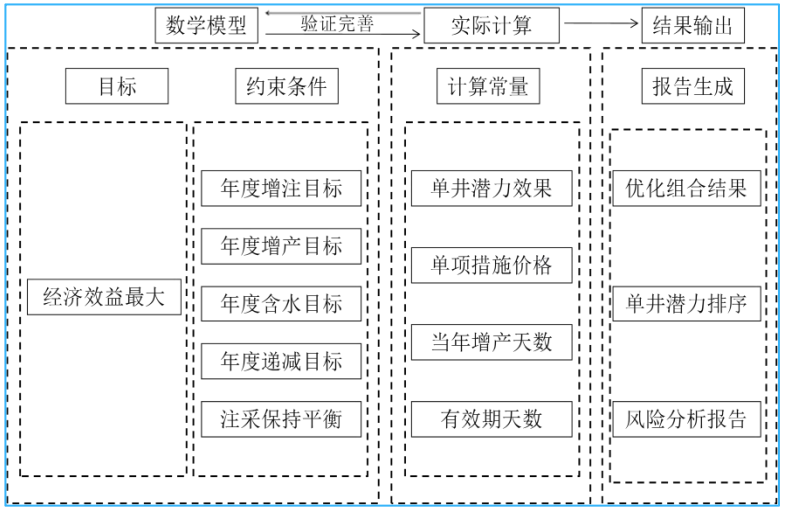


句法分析

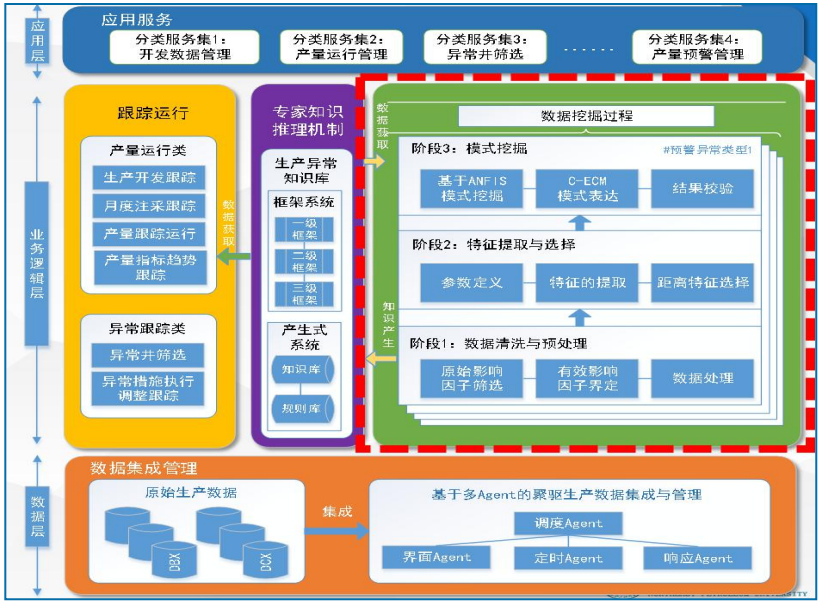


基于智能计算的油田开发调整规划优化

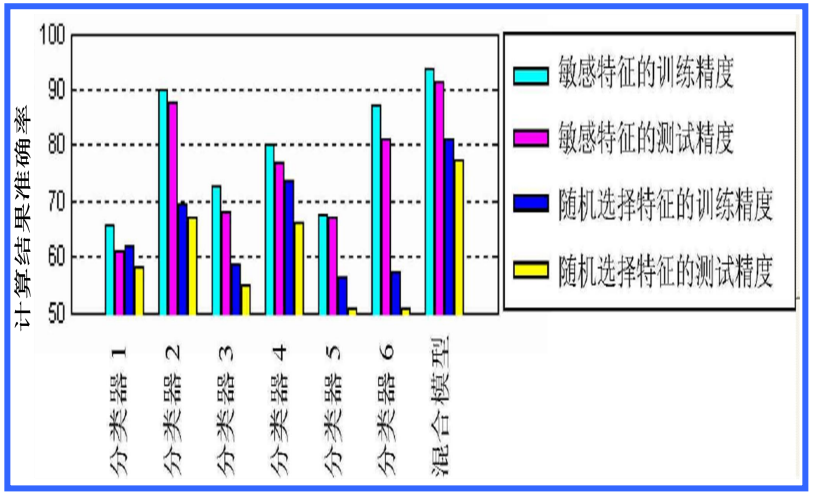




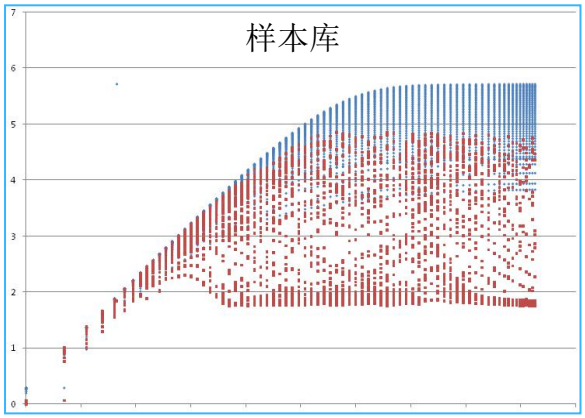
基于大数据的三次采油跟踪运行与预警系统

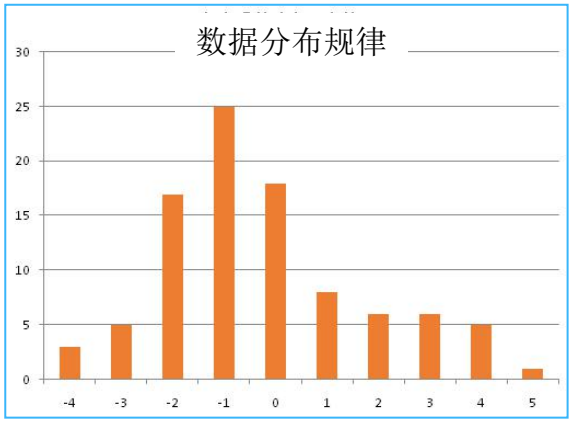


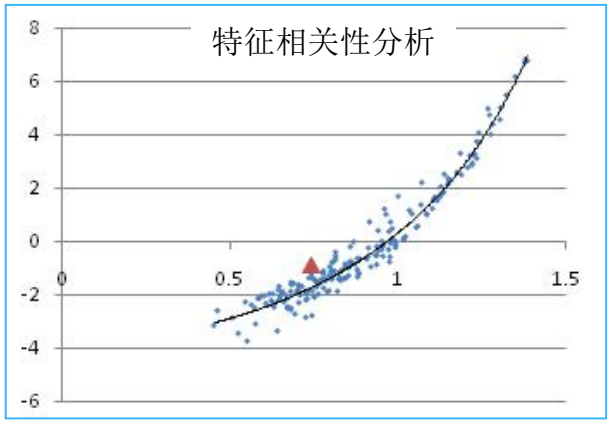
输入	分类器 1		分类器 2		分类器 3		分类器 4	
特征	训练	测试	训练	测试	训练	测试	训练	测试
C3	65.67	61	90	87.86	72.67	68	80.33	77
C4	62	58.03	69.27	66.93	58.5	55.1	73.73	66.3
输入	分类器 5		分类器 6		6 分类器均值		混合模式	
特征	训练	测试	训练	测试	训练	测试	训练	测试
C3	67.67	67	87.33	81	77.28	73.61	93.67	91.33
C4	56.5	50.93	57.37	50.8	62.89	58.02	81.79	77.15

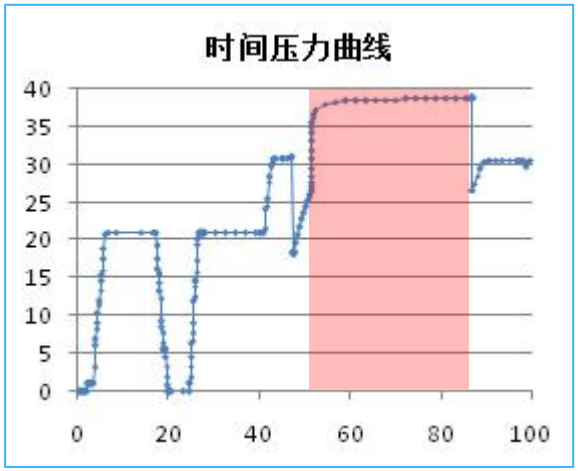


基于大数据的试井解释

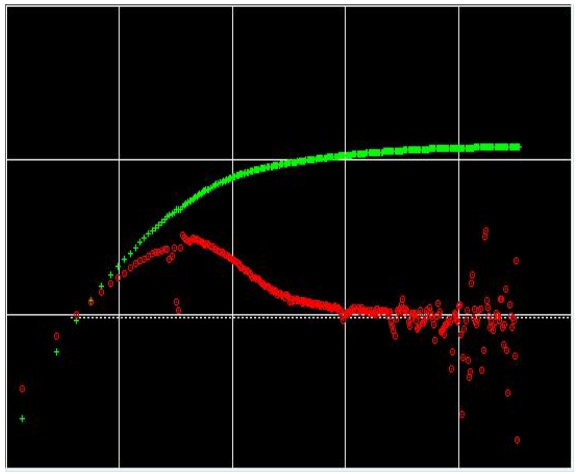




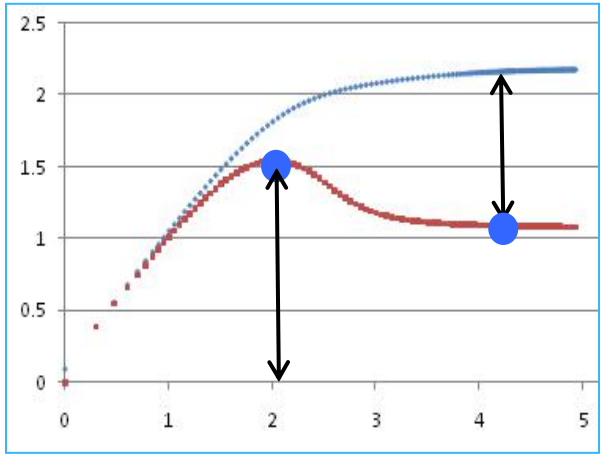




压力曲线预处理及流动段提取



双对数曲线预处理



曲线特征抽取

提纲

- 大数据人才培养现状
- 大数据技术基础
- 专业建设规划
- 开设大数据专业存在的问题

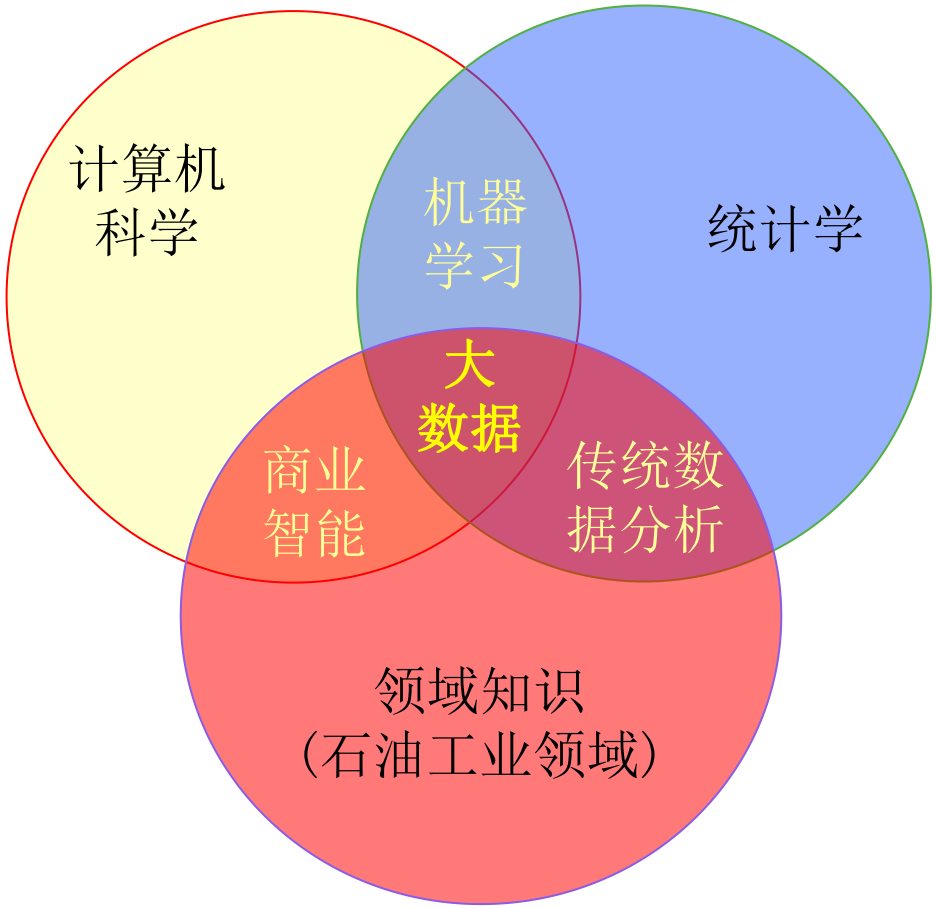
专业发展定位与方向定位

- **三方向的职位：**旨在培养大数据领域专业人才，如“大数据开发工程师、数据分析师、大数据管理工程师”等。

方向	对应岗位
Hadoop大数据开发	**大数据开发工程师：** 数据建模、数据集成、数据ETL处理、数据质量分析
数据挖掘、数据分析&机器学习	**数据分析师：** 结合业务对数据深入分析，为业务的策略、产品优化提供数据支持。 **数据挖掘工程师：** 信息检索、文本处理、机器学习、数据挖掘技术研究与应用。
大数据运维&云计算	**数据管理工程师** 分布式/并行数据环境、性能优化

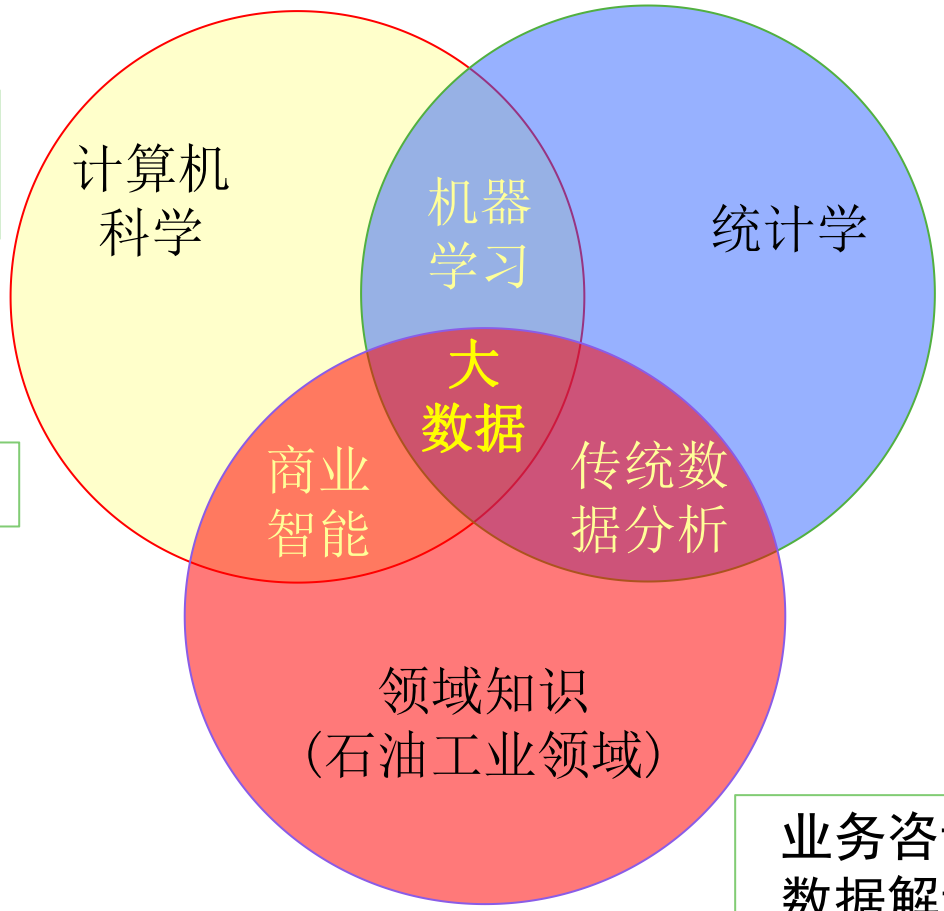
专业发展定位与方向定位

- **三学科结合：**采用多学科交叉的培养模式，融合计算机、统计学、石油工业技术等知识体系，形成了反映专业内涵的核心课程。



课程设置

- 三层次能力：在能力方面，形成计算机及数据库应用能力、大数据分析能力、面向领域的大数据应用能力。



课程设置

计算机基础

- 算法分析与设计
- 并行计算与分布式计算
- 数据结构
- 软件工程
- 数据库原理
- C++
- Java

大数据基础理论

- 大数据科学与技术导论
- 计算思维和数据科学

大数据核心技能

- 大数据与领域建模
- 数据可视化分析
- 商务智能及应用
- 数据挖掘与分析
- 大数据集成技术
- 概率论与数理统计

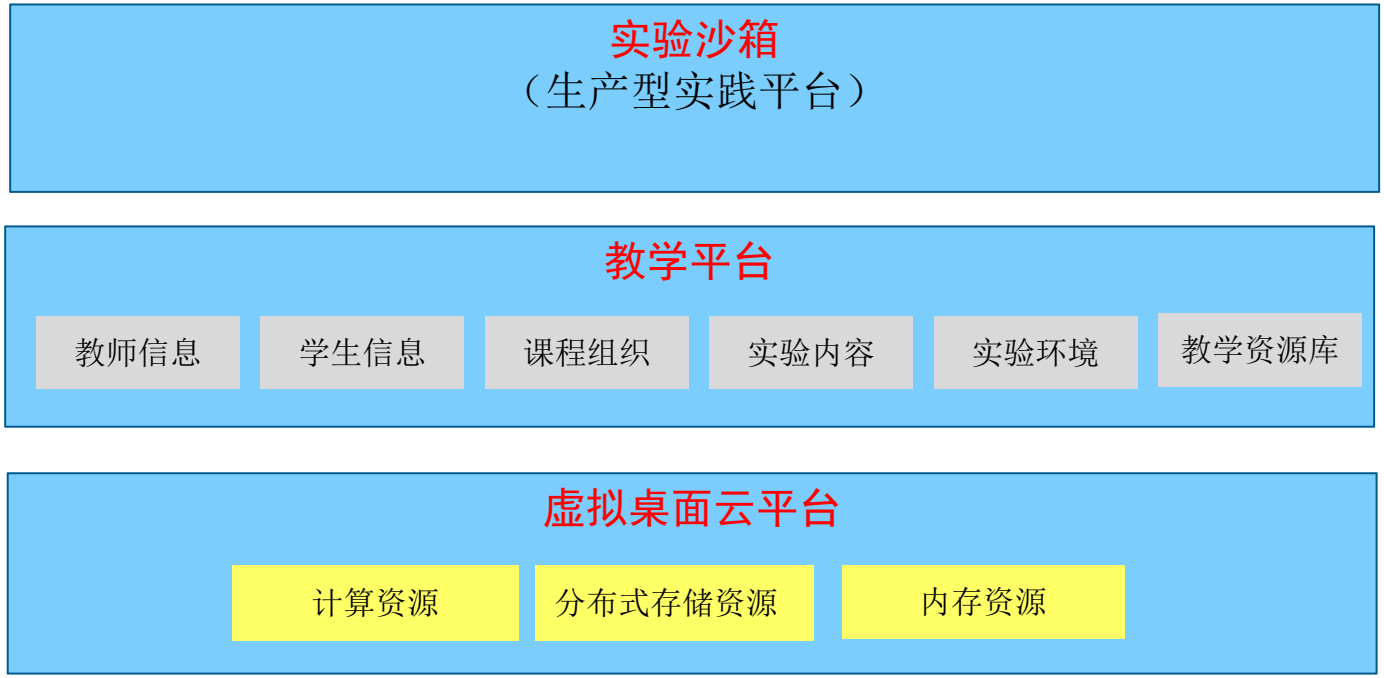
大数据应用技能

- Python
- R语言
- Matlab
- SPSS
- 大数据平台技术

领域大数据

- 石油勘探开发概论
- 石油地质统计学

实验室建设


实验沙箱
（生产型实践平台）
教学平台
教师信息
学生信息
课程组织
实验内容
实验环境
教学资源库
虚拟桌面云平台
计算资源
分布式存储资源
内存资源

培养模式

“产学研”相结合，开展数据人才培养实训基地

高等教育作为我国人才培养的主要基地，承担着大部分高素质人才产出的任务，在大数据时代也不例外。而且针对数据人才严重缺乏的现状，目前国内外高校也都采取了积极的应对措施。为应对大数据产业人才短缺的问题，当前切实有效的方法 是从技能培育阶段的高校入手，通过传统的人才培训体制和学科教育体系，将“数据”纳入教育范畴，并积极建立高校主导的数据科学研究中心，为数据人才的输出做好知识和技能储备。

有专家指出，由于无法提供真实的大数据环境，高校很难培养出市场真正需要的数据人才，于是“产学研”相结合的实践陆续推出。

通过建立实训基地，就可以为高校学生提供学校没有的数据环境和实战机会，使数据人才的业务应用呈综合实践能力得到培养

师资培养

尽管我院有一批从事大数据研究的青年教师。但要进行专业课教学，需要进行系统地学习和培训。这是一个系统工程。

引进　　培训　　合作

教学模式

- 课堂与实验室结合
- 线上/线下结合
- 科研与教学结合

提纲

- 大数据人才培养现状
- 大数据技术基础
- 专业建设规划
- 开设大数据专业存在的问题

开设大数据专业存在的问题

科研与教学的差异：

科研：

- ✓点--用多少学多少

教学：　系统性

- ✓线---一门课程
- ✓面--专业

开设大数据专业存在的问题

基础设施的欠缺：

硬件：

- ✓需要大量的投资

软件：

- ✓系统软件需要开发
- ✓教学资源需要积累
- ✓缺少系统化的权威性的专业教材

开设大数据专业存在的问题

师资培养：

- ✓配齐所有课程需要大量师资
- ✓培养成熟师资需要时间
- ✓课程范围涉及多个学科

一点建议

- 加强院校间大数据教育沟通与交流
- 在师资培养、教材建设加强合作
- 高校与油田、培训机构加强合作

谢 谢