数据挖掘与商务分析

课程导论

主讲教师：肖升生
xiao.shengsheng@shufe.edu.cn

关于授课教师

- **Instructor:** 肖升生
  - Email: xiao.shengsheng@shufe.edu.cn
  - Tel: 021-65904410-837
  - Office room: #837

- **Research areas:**
  (1) BA & Data Mining
  (2) Digital Economics

- **Faculty website:**
  https://de.sufe.edu.cn/18/4c/c12089a202828/page.htm

讲授提纲

01 数据类型与价值使用

02 数据挖掘、AI大模型与商务智能

03 跨行业的数据挖掘流程

04 课程内容与设计

05 课程学习材料

讲授提纲

01 数据类型与价值使用

02 数据挖掘、AI大模型与商务智能

03 跨行业的数据挖掘流程

04 课程内容与设计

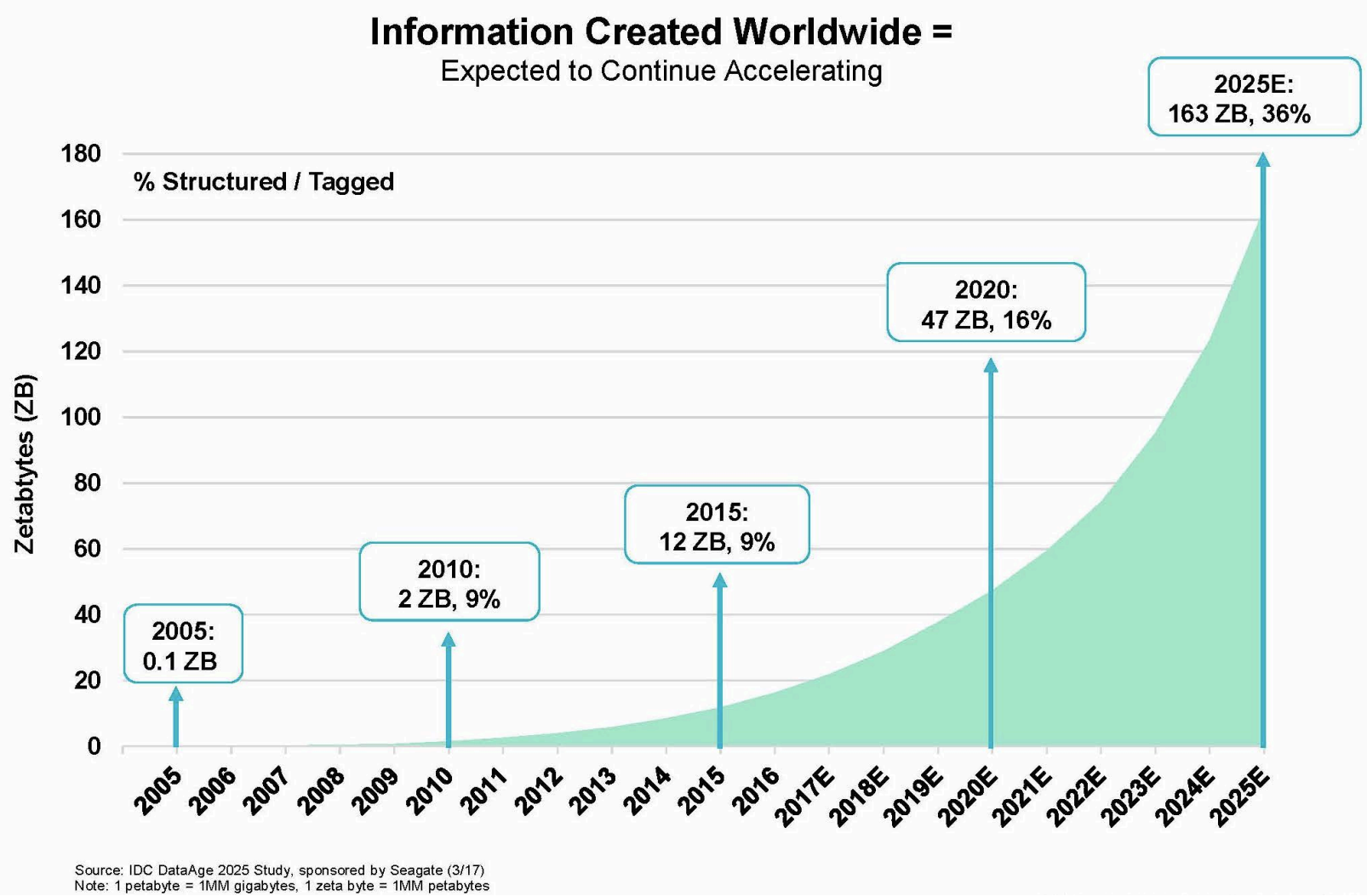
05 课程学习材料

数据类型与量级

- 数据类型：
  - 数值
  - 文本
  - 位置
  - 声音
  - 视频
  - …

资料来源: https://www.domo.com/

数据的快速增长



Bit、Byte、KB、MB、GB、TB、PB、EB、ZB、YB

数据的利用率低

- 数据被称为了新“石油”和新资产
- 但跟石油类似，数据的价值需要提炼
- 现状：“Data Rich but Information Poor”
  - 大量的信息隐藏在海量的数据背后
  - 绝大部分的数据都没有被分析和使用
  - 有用信息的挖掘需要耗费大量人力和物力

讲授提纲

01 数据类型与价值使用

02 **数据挖掘、AI大模型与商务智能**

03 跨行业的数据挖掘流程

04 课程内容与设计

05 课程学习材料

什么是商务智能

商务智能是利用数字智能技术从大量数据中提取信息，转化为可指导决策的知识和洞察力。

零售业库存优化

某大型零售商利用BI分析销售数据，精准预测需求，减少库存积压，提高资金周转率。

银行业风险控制

银行通过BI实时监控交易模式，有效识别欺诈行为，降低信贷风险，提升客户信任度。

制造业生产效率

制造企业运用BI分析生产线数据，优化资源配置，减少浪费，显著提升生产效率和产品质量。

商务智能的几个发展阶段

1. 传统报表阶段

依赖手工或批处理方式收集数据，借助简单的汇总统计，生成固定格式的周报、月报，回顾历史业绩，帮助管理者掌握基本业务状况。

2. 数据仓库与OLAP阶段

通过数据仓库以整合多源数据，并借助OLAP技术，使用多维数据模型，用户可灵活进行“切片、切块、钻取、旋转”操作分析业务表现。

3. 数据分析与智能化阶段

融合自助分析、大数据处理与AI技术，支持实时洞察、预测分析和自然语言交互，实现人人可用、主动发现的敏捷决策新范式。

示例: 汽车工业发展的三个阶段

机械化汽车

信息技术向汽车设计、生产制造等环节渗透
提高了生产效率

汽车雏形 ➡ 单件少量生产 ➡ 大规模生产

美国
T型车+流水线
1903-1927

美国、欧洲
自动化生产线+精益生产
1947-1980s

英国、法国、美国
蒸汽汽车
1705-1834

德国
内燃机四轮车
1876-1886

日本
多样化+准时化+精益生产
1970-1976

来源：阿里研究院

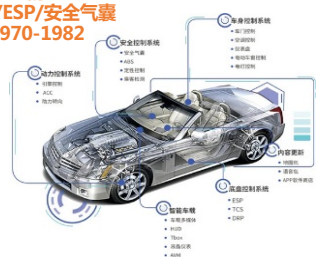
机电化汽车

数字控制和互联网技术向汽车产品和服务渗透
提升了汽车性能和舒适度，创造更高产品价值

➡ 机电一体 ➡ 网联车

电子控制式喇叭、微处理器控制的
ABS/ESP/安全气囊
1970-1982

车载无线电对讲
1990s



微电脑控制的
车辆集中电控、
GPS定位/离线导航/移动出行服务
1982-2000s

车机互联
Carplay、Anroid Auto
2013以来

智能化汽车

云计算、人工智能技术与汽车产业深度融合
重新定义汽车，重塑汽车产业

➡ 云网端AI一体

自动驾驶：自动泊车、智能巡航、自主驾驶
2018以来

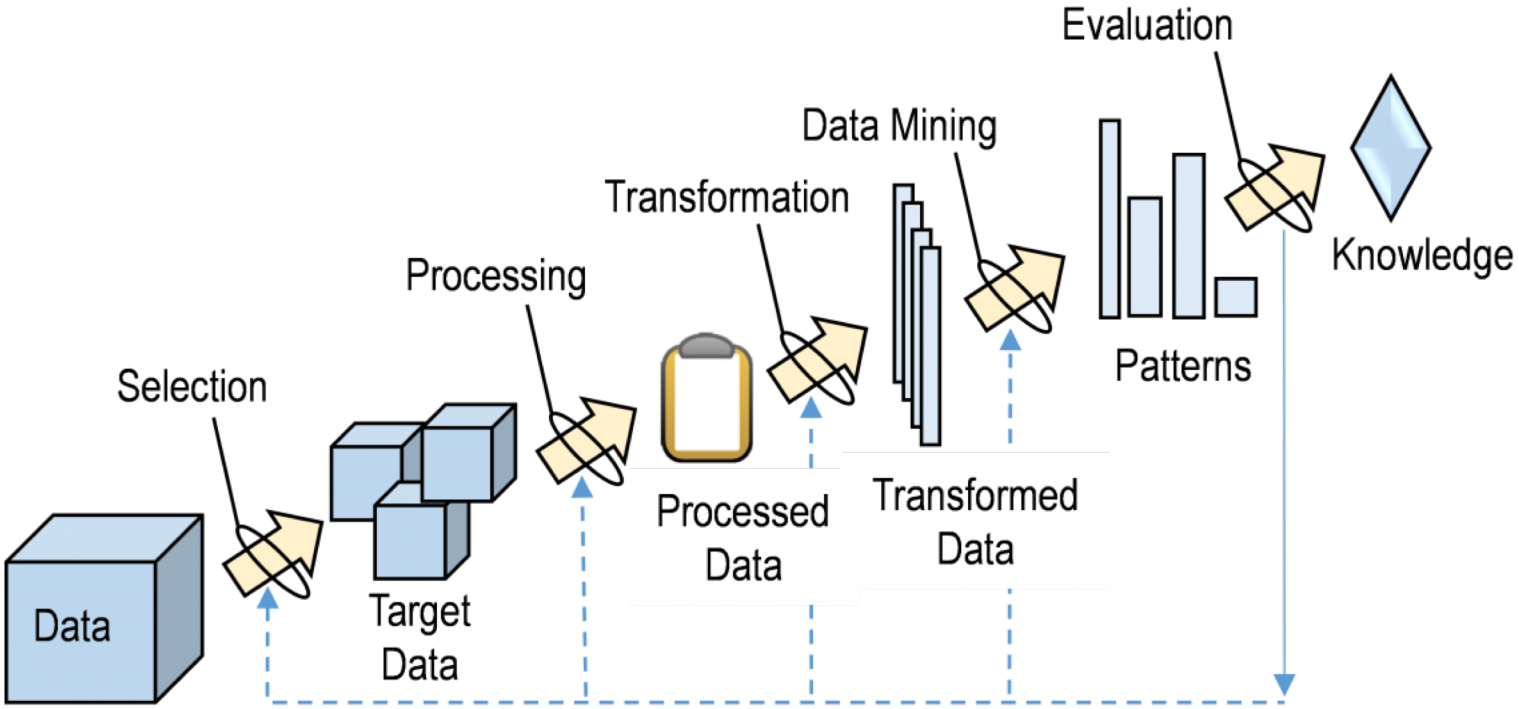
智能座舱：智能交互、智能仪表、360°影像
2018

订阅式软件服务：信息娱乐、系统升级
2020以来

智能导航：路况实时交互、路线动态优化
2020以来

什么是数据挖掘

- 数据挖掘(Data Mining)：从大量的数据中使用智能化的方法自动地发现有用信息的过程

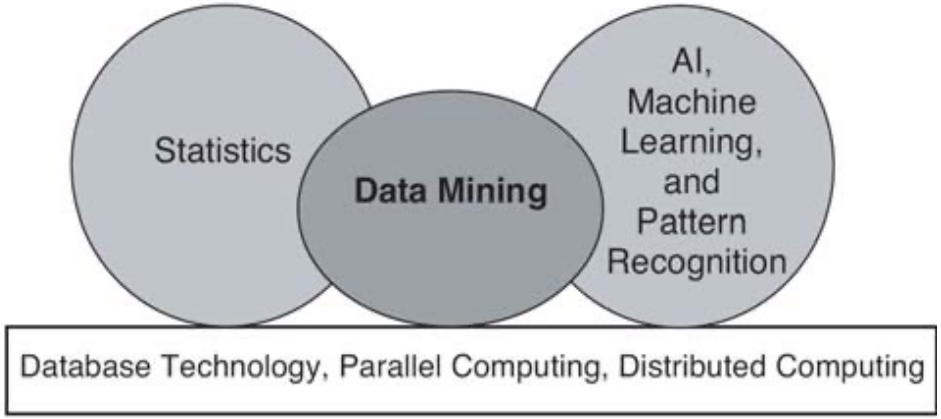


数据挖掘的关键环节

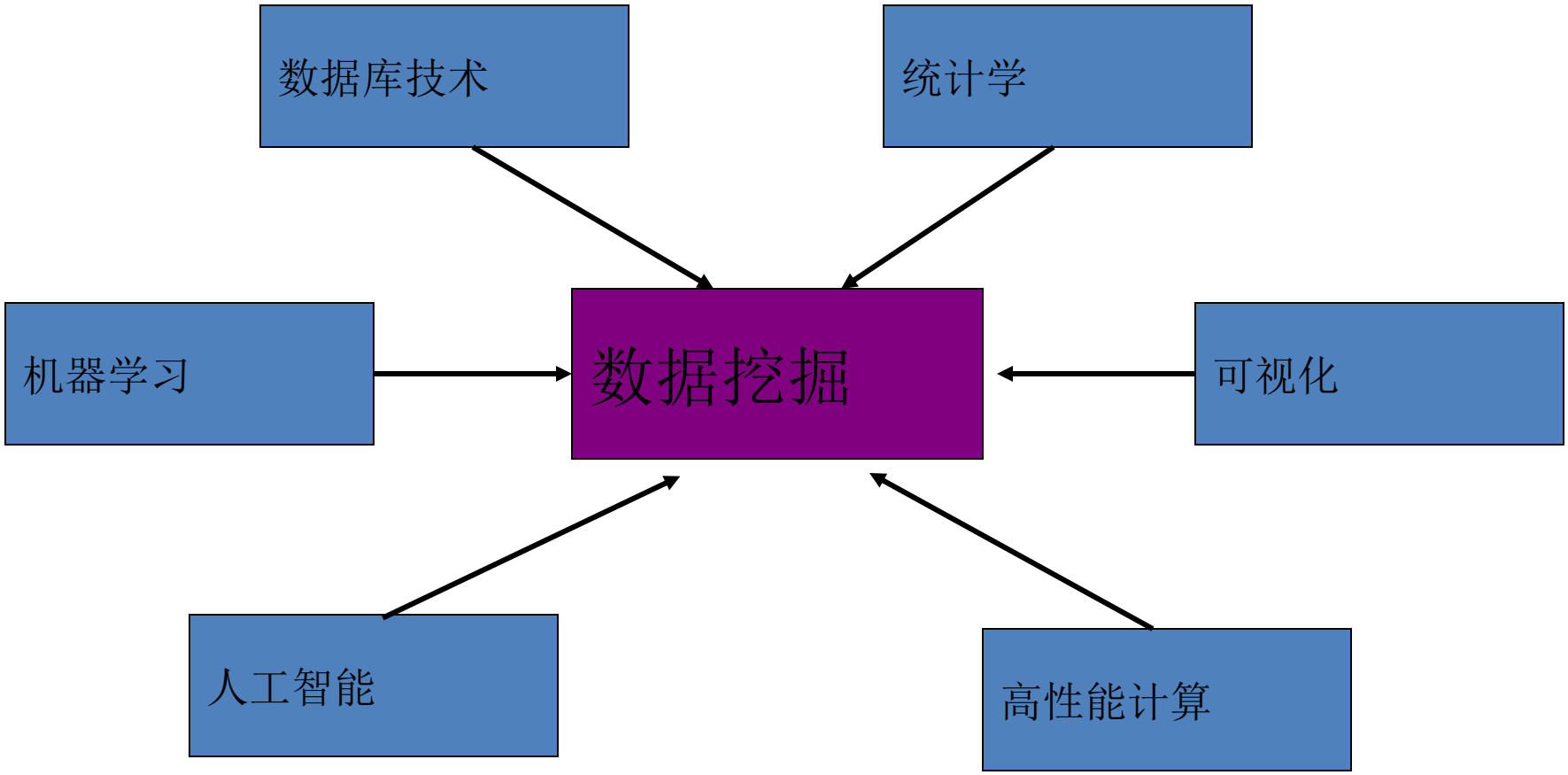

数据挖掘
数据预处理
数据清洗，去除重复值和异常值。
数据转换，将非数值型数据转化为数值型。
特征工程
特征选择，挑选对目标变量影响最大的特征。
特征构造，通过已有特征创建新的特征。
模型构建
分类算法，如决策树、随机森林等。
聚类算法，如K-means、DBSCAN等。
关联规则，如Apriori、FP-growth等。
模型评估
准确率，衡量模型预测正确的比例。
召回率，衡量模型找到所有正例的能力。
模型部署
生产环境集成，将模型嵌入业务系统。
持续监控，定期检查模型性能并进行调整。
应用场景
客户细分，根据消费行为将客户分为不同群体。
市场篮子分析，研究顾客购买商品之间的关联性。

数据挖掘要解决的问题

- 待解决的问题
  - 高维度
  - 异构性
  - 方法的可伸缩性
  - 分布式数据存储
- 借鉴的方法论来源
  - 统计学
  - 人工智能
  - 机器学习
  - 数据库技术
  - 分布式计算



数据挖掘是多学科融合产物



数据挖掘基本任务

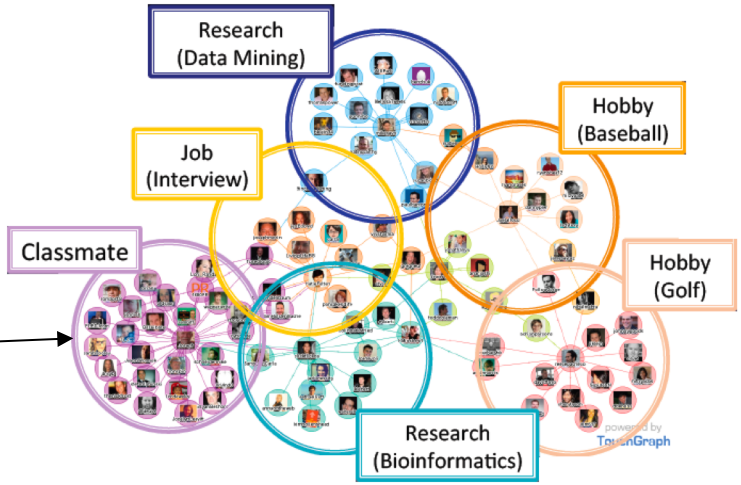
- 数据挖掘的基本任务：
  - 预测：根据已有属性值预测特定属性值
  - 描述：概括数据中潜在的关系模式

- 数据挖掘的基本内容：
  - 分类分析 [预测性]
  - 聚类分析 [描述性]
  - 关联规则分析 [描述性]

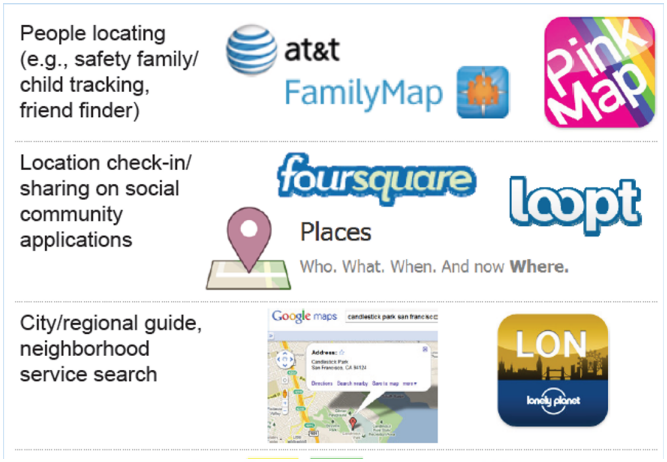
数据挖掘与商业应用

- 数据挖掘与商业应用

文本的大量使用



个体间联系的网络化



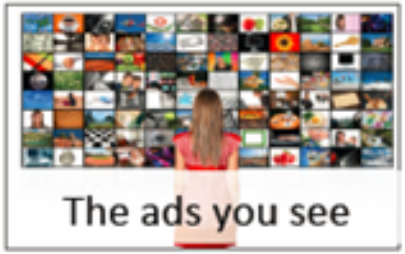
带有位置信息的智能终端化

数据挖掘的应用领域


The music you hear


The books you read


The media you follow


The ads you see


The things you buy


The people you date

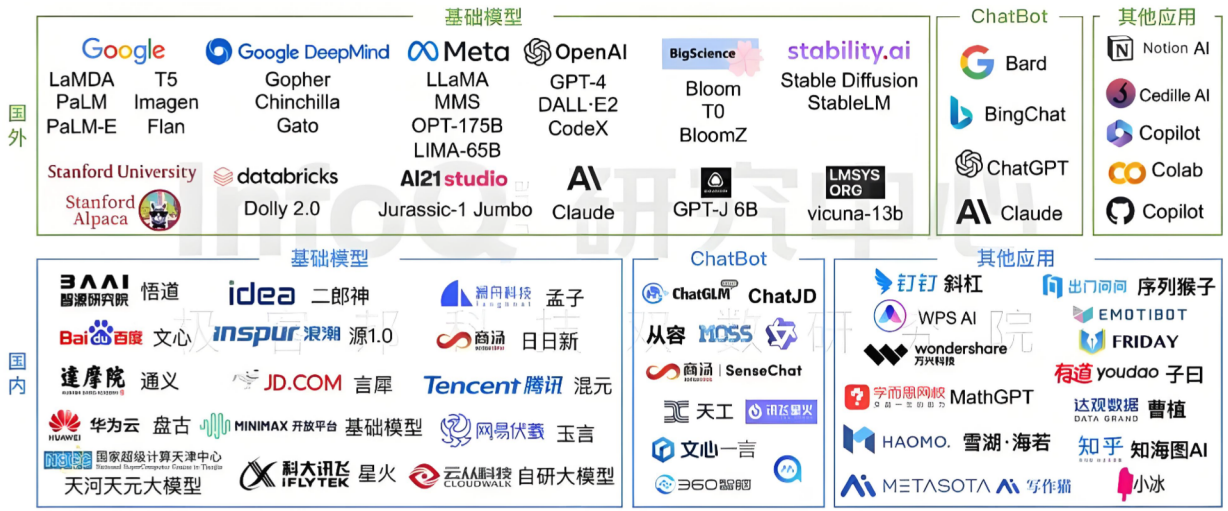

The next election


The way you're tracked

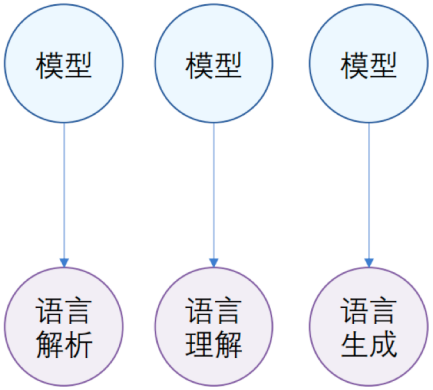

Your financial health?

AI大模型

AI大模型：指基于海量数据和强大算力训练出的、拥有数十亿甚至万亿参数的深度学习模型（如统DeepSeek，Chat-GPT等）。它们具备强大的模式识别、自然语言处理、生成和推理能力。

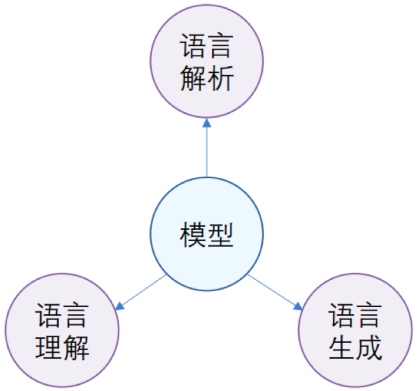


大模型与新学习范式



过去

为每个任务训练独立的模型



不久之前

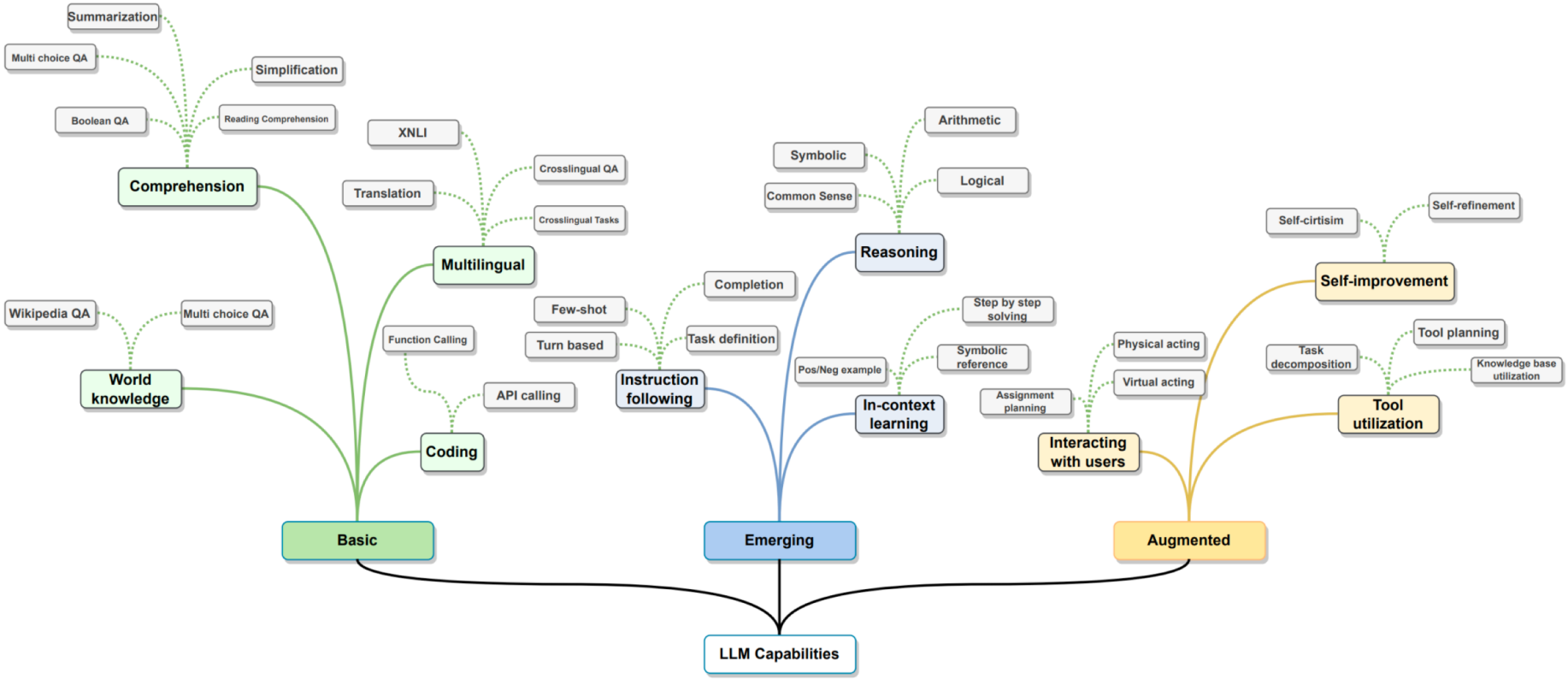
中心节点完成预训练，用户在此基础上面向任务微调

个体化训练 → 中心化训练 + 个体化微调

现在（大规模语言模型）

- 提示学习
  - 上下文学习
  - 思维链提示
- 轻量化微调

AI大模型的能力版图



Shervin Minaee, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, Jianfeng Gao. Large Language Models: A Survey. https://arxiv.org/pdf/2402.06196.pdf

思考

- 有了AI大模型，数据挖掘的学习还有必要吗？

讲授提纲

01 数据类型与价值使用

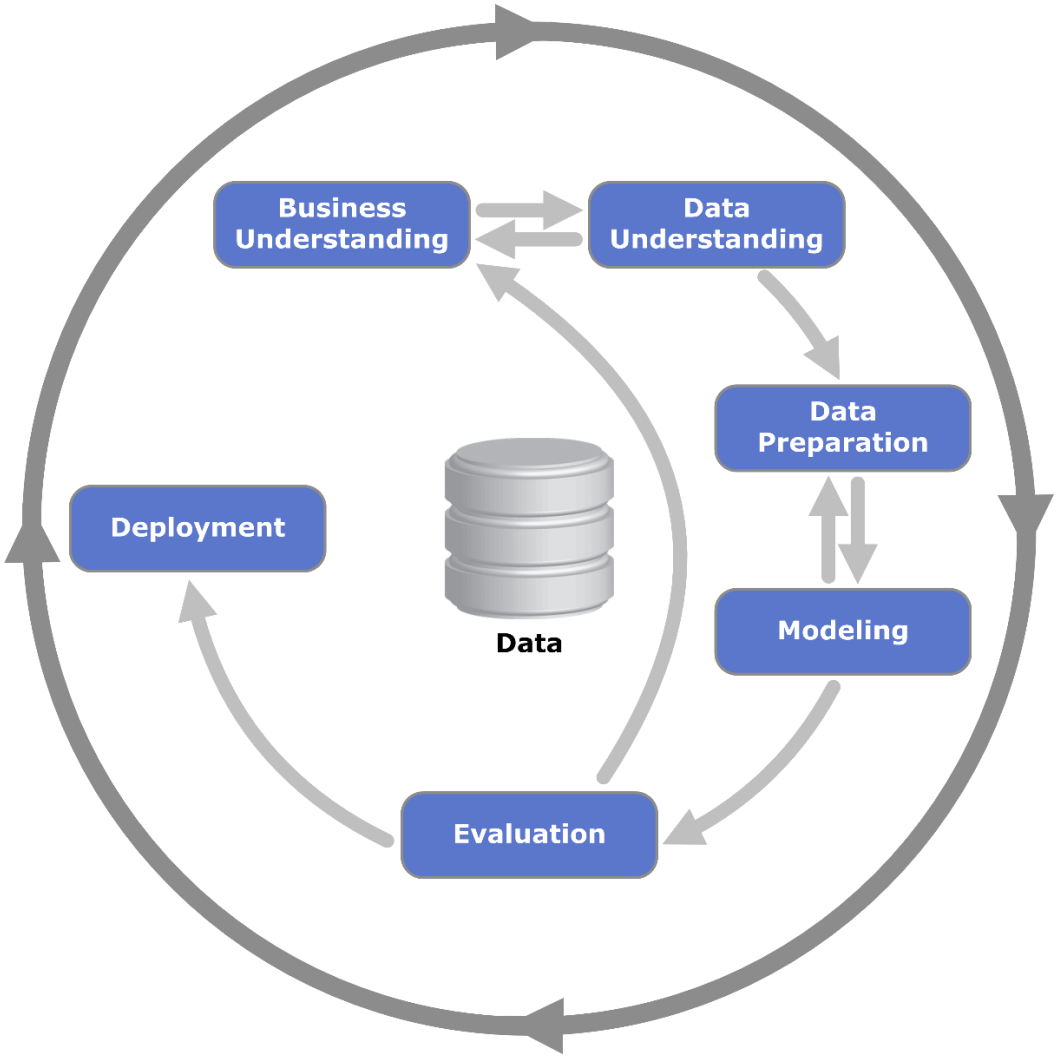
02 什么是数据挖掘

03 **跨行业的数据挖掘流程**

04 课程内容与设计

05 课程学习材料

跨行业的数据挖掘流程



Cross-industry standard process for data mining (**CRISP-DM**)

项目案例

企业画像构建

- 财务指标维度
- 融资风险维度
- 经营风险维度
- 法律诉讼维度

02

多维度数据收集和处理

01

- 企业经营数据
- 舆情数据：抓取与企业相关的舆论情绪、社会评价
- 司法诉讼数据：企业法律纠纷情况

融资利率确定

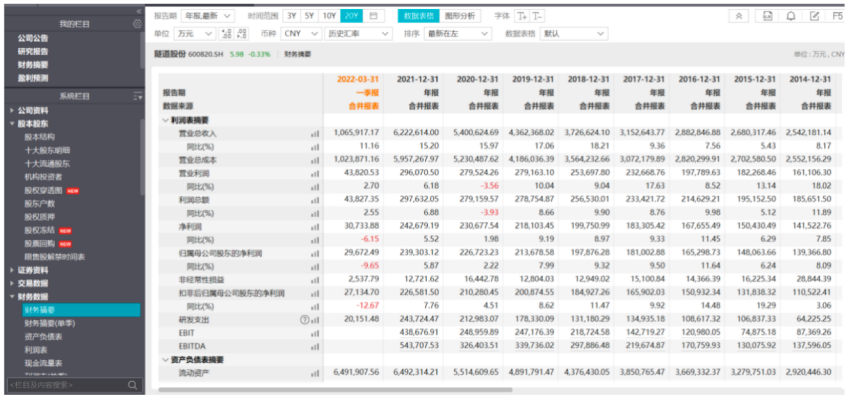
03

- 基于成本加成的融资利率确定模型
- 融资利率模型开发与落地部署

供应商企业风险评级及供应链金融利率确定

项目案例：业务理解与数据处理

结构化数据—财务信息



报告期	2022-03-31 一季报 合并报表	2021-12-31 年报 合并报表	2020-12-31 年报 合并报表	2019-12-31 年报 合并报表	2018-12-31 年报 合并报表	2017-12-31 年报 合并报表	2016-12-31 年报 合并报表	2015-12-31 年报 合并报表	2014-12-31 年报 合并报表
利润表摘要									
营业总收入	1,065,917.17	6,222,614.00	5,400,624.69	4,362,368.02	3,726,624.10	3,152,643.77	2,882,846.88	2,680,317.46	2,542,181.14
同比(%)	11.16	15.20	15.97	17.06	18.21	9.36	7.56	5.43	8.17
营业总成本	1,023,871.16	5,957,267.97	5,230,487.62	4,186,036.39	3,564,232.66	3,072,179.89	2,820,299.91	2,702,580.50	2,552,156.29
营业利润	43,820.53	296,070.50	279,524.26	279,163.10	253,697.80	232,668.76	197,789.63	182,268.46	161,106.30
同比(%)	2.70	6.18	-3.56	10.04	9.04	17.63	8.52	13.14	18.02
利润总额	43,827.35	297,632.05	279,159.57	278,754.87	256,530.01	233,421.72	214,629.21	195,152.50	185,651.50
同比(%)	2.55	6.88	-3.93	8.66	9.90	8.76	9.98	5.12	11.89
净利润	30,733.88	242,679.19	230,677.54	218,103.45	199,750.99	183,305.42	167,655.49	150,430.49	141,522.76
同比(%)	-6.15	5.52	1.98	9.19	8.97	9.33	11.45	6.29	7.85
归属母公司股东的净利润	29,672.49	239,303.12	226,723.23	213,678.58	197,876.28	181,002.88	165,298.73	148,063.66	139,366.80
同比(%)	-9.65	5.87	2.22	7.99	9.32	9.50	11.64	6.24	8.09
非经常性损益	2,537.79	12,721.62	16,442.78	12,804.03	12,949.02	15,100.84	14,366.39	16,225.34	28,844.39
扣非后归属母公司股东的净利润	27,134.70	226,581.50	210,280.45	200,874.55	184,927.26	165,902.03	150,932.34	131,838.32	110,522.41
同比(%)	-12.67	7.76	4.51	8.62	11.47	9.92	14.48	19.29	3.06
研发支出	20,151.48	243,724.47	212,983.07	178,330.09	131,180.29	134,935.18	108,617.32	106,837.33	64,225.25
EBIT		438,676.91	248,959.89	247,176.39	218,724.58	142,719.27	120,980.05	74,875.18	87,369.26
EBITDA		543,707.53	326,403.51	339,736.02	297,886.48	219,674.87	170,759.93	130,075.92	137,596.05
资产负债表摘要									
流动资产	6,491,907.56	6,492,314.21	5,514,609.65	4,891,791.47	4,376,430.05	3,850,765.47	3,669,332.37	3,279,751.03	2,920,446.30

非结构化数据—法律纠纷

案号	(2024) 辽0106民诉前调3094号	案由	租赁合同纠纷
案件类型	民事	当事人	原告：沈阳双晋交通科技有限公司 被告： 1. 李* 2. 黄** 3. 陈** 第三人：支付宝（中国）网络技术有限公司
法院	沈阳市铁西区人民法院	承办部门	-
承办法官	-	法官助理	-
案件状态	结案	立案日期	2024-02-27
开庭时间	-	结束时间	-

结构化数据—财务信息

上海润达医疗科技股份有限公司

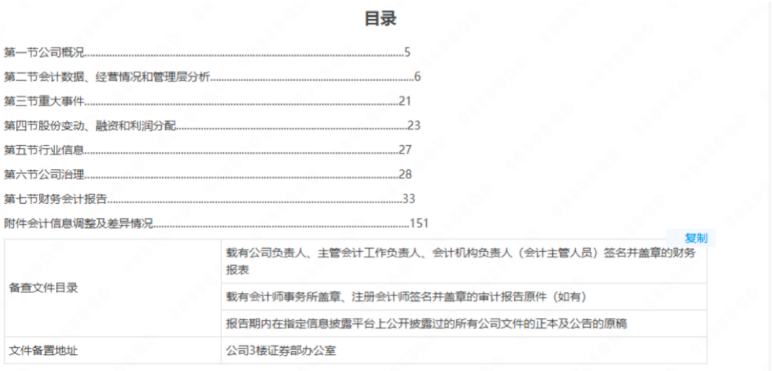
行政处罚

序号	披露日期	处罚日期	处罚类型	违规事由	文号	处理人	法律依据	原文
1	2022-05-11	-	监管关注	其他	-	上海证券交易所	-	
2	2016-05-18	-	整改通知	未依法履行职责	金地税限改〔2016〕5号	上海市地方税务局金山区分局	《中华人民共和国税收征收管理法》	

关联人处罚

序号	披露日期	当事人	处罚类型	事件标题	文号	处理人	原文
1	2022-04-09	何和平	罚款,警告(拟)	*ST实达:福建实达集团股份有限公司关于收到中国证券监督管理委员会《福建监管局行政处罚事先告知书》的公告	闽政监函[2022]134号	福建证监局	
2	2018-12-24	何和平	公开谴责	关于对上海普天邮通科技股份有限公司及有关责任人予以纪律处分的决定	纪律处分决定书[2018]76号	上海证券交易所	

非结构化数据—年度报告



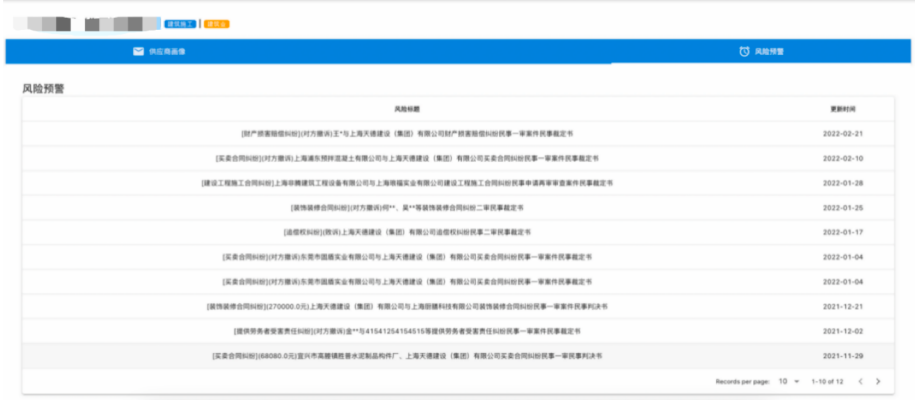
目录

复制

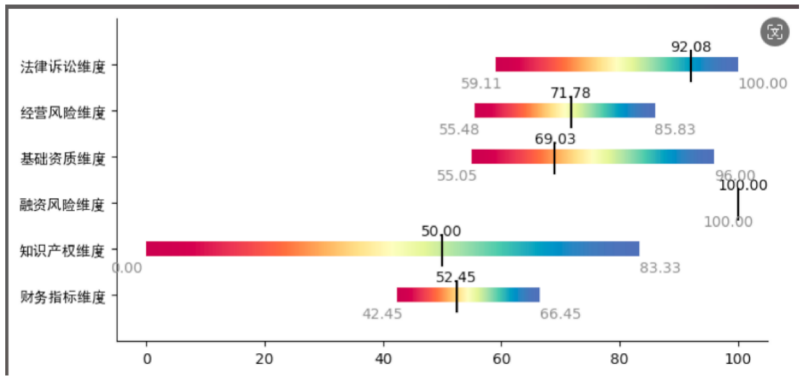
备查文件目录	载有公司负责人、主管会计工作负责人、会计机构负责人（会计主管人员）签名并盖章的财务报表
	载有会计师事务所盖章、注册会计师签名并盖章的审计报告原件（如有）
	报告期内在指定信息披露平台上公开披露过的所有公司文件的正本及公告的原稿
文件备置地址	公司3楼证券部办公室

项目案例：成果交付

企业画像各维度得分详情

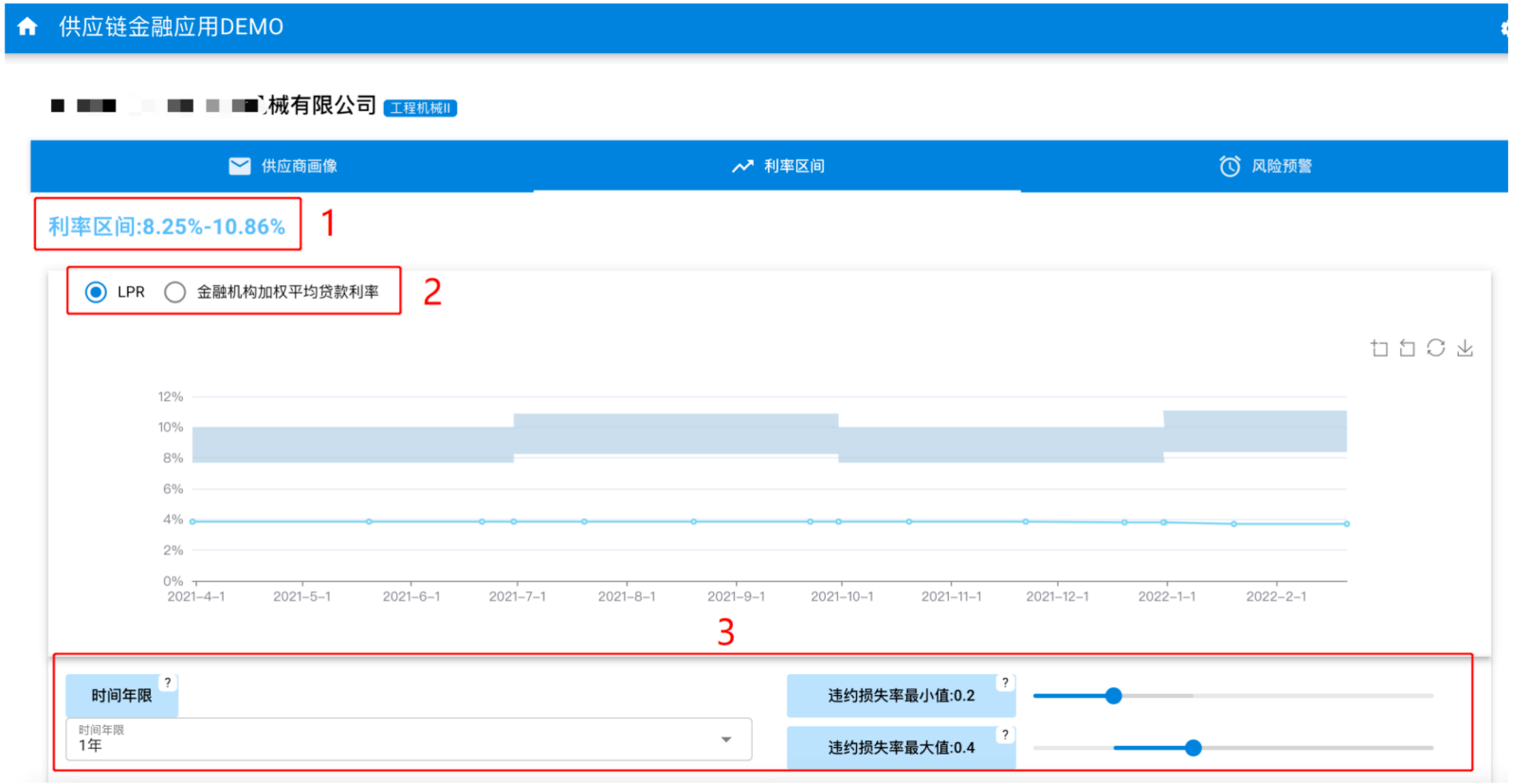


指标预警信息列表详情

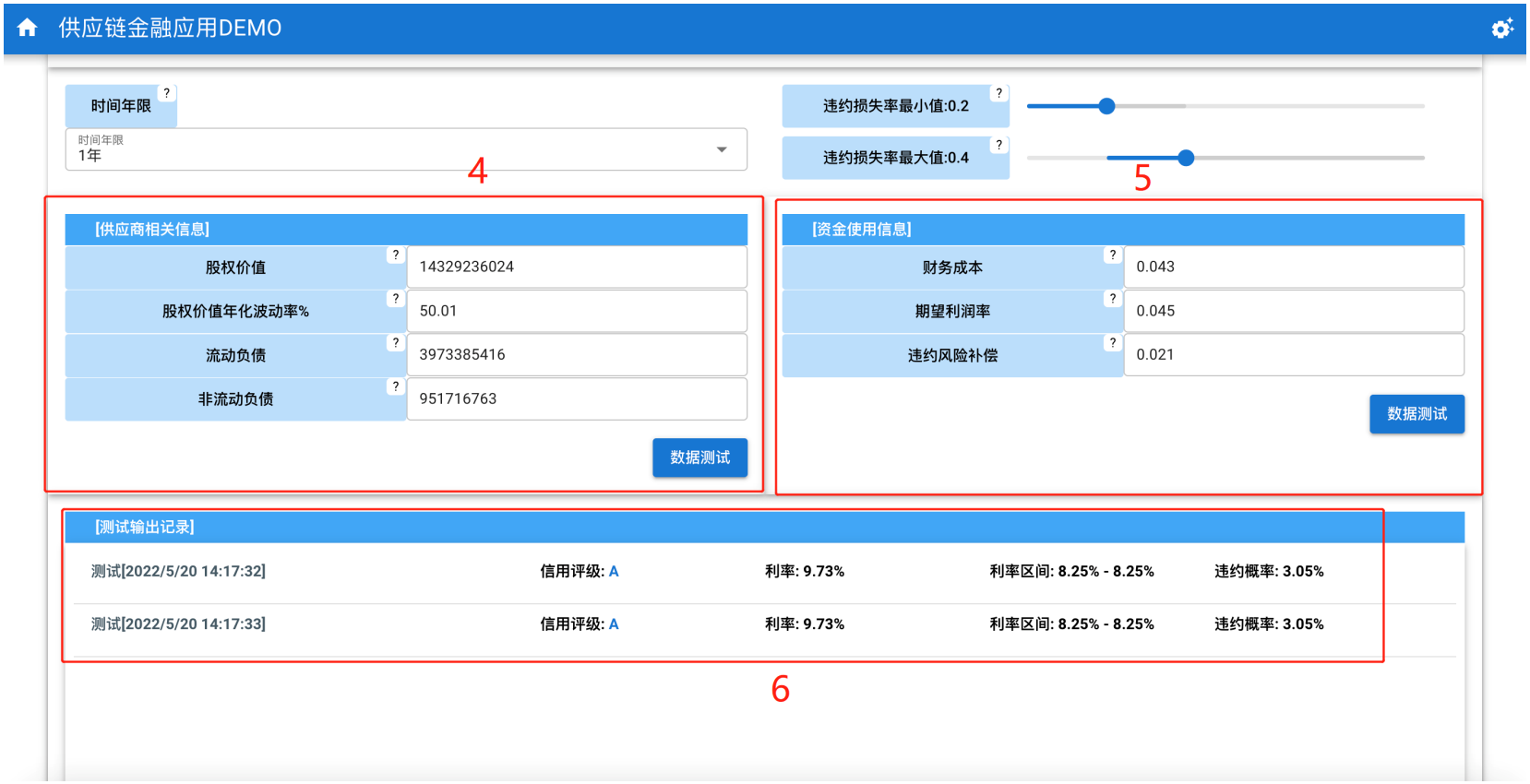


企业各维度得分情况在同行业中所处位置

项目案例：成果交付


供应链金融应用DEMO
工程机械II
供应商画像
利率区间
风险预警
利率区间:8.25%-10.86%
1
LPR
金融机构加权平均贷款利率
2
12%
10%
8%
6%
4%
2%
0%
2021-4-1
2021-5-1
2021-6-1
2021-7-1
2021-8-1
2021-9-1
2021-10-1
2021-11-1
2021-12-1
2022-1-1
2022-2-1
3
时间年限
时间年限
1年
违约损失率最小值:0.2
违约损失率最大值:0.4

项目案例：成果交付


供应链金融应用DEMO
时间年限
时间年限
1年
4
违约损失率最小值:0.2
违约损失率最大值:0.4
5
[供应商相关信息]
股权价值
14329236024
股权价值年化波动率%
50.01
流动负债
3973385416
非流动负债
951716763
数据测试
[资金使用信息]
财务成本
0.043
期望利润率
0.045
违约风险补偿
0.021
数据测试
[测试输出记录]
测试[2022/5/20 14:17:32]
信用评级: A
利率: 9.73%
利率区间: 8.25% - 8.25%
违约概率: 3.05%
测试[2022/5/20 14:17:33]
信用评级: A
利率: 9.73%
利率区间: 8.25% - 8.25%
违约概率: 3.05%
6

讲授提纲

01 数据类型与价值使用

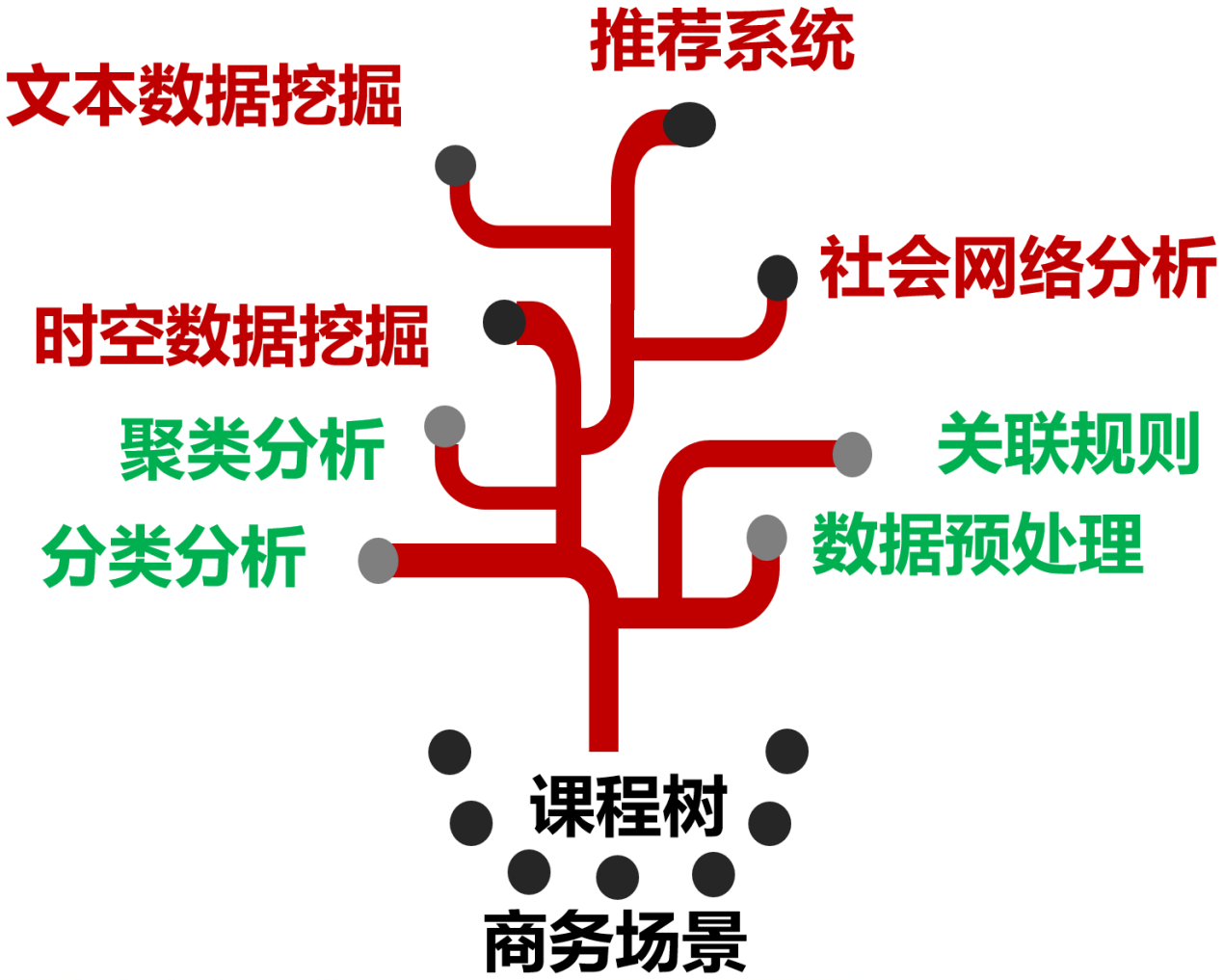
02 什么是数据挖掘

03 跨行业的数据挖掘流程

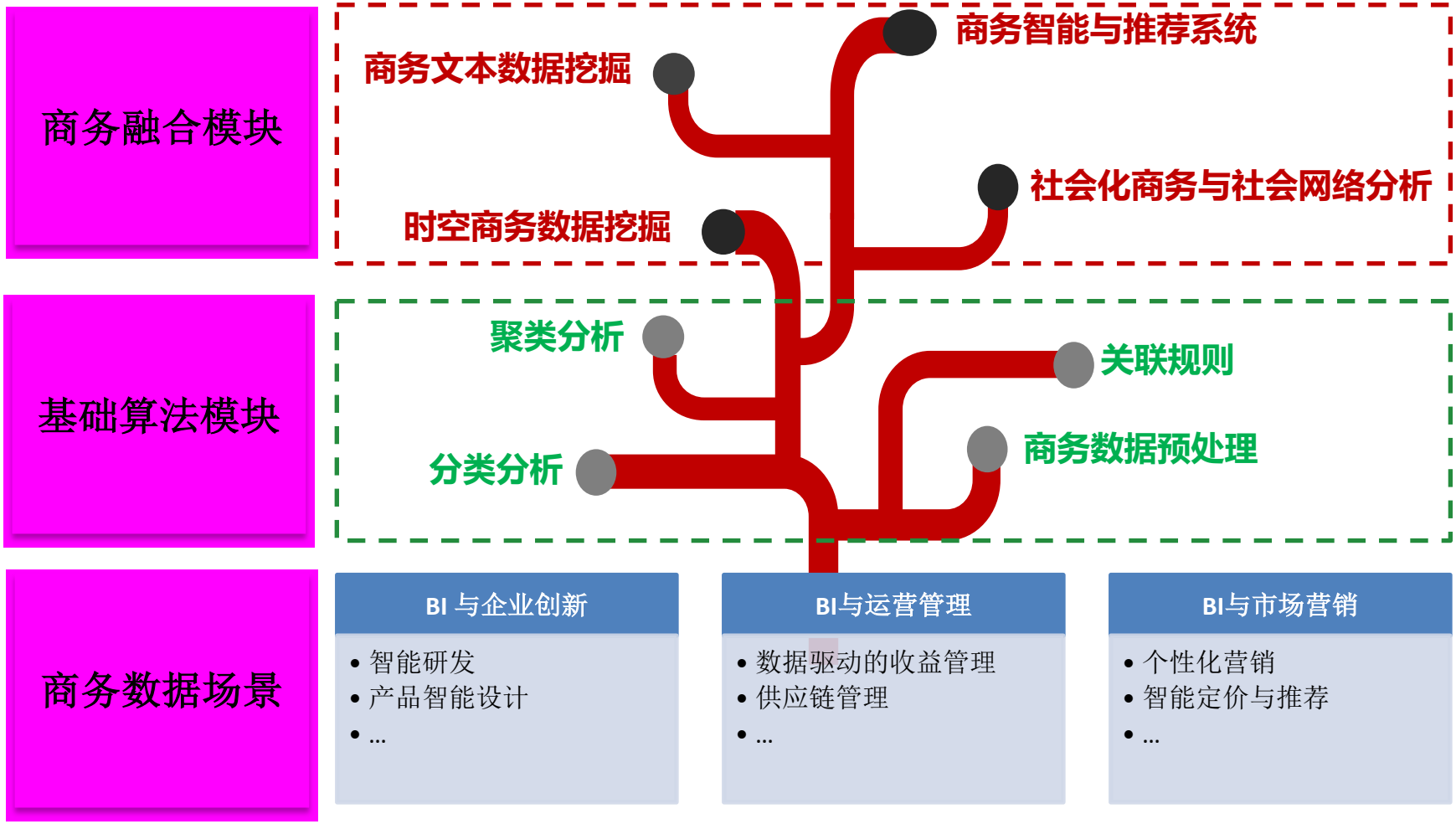
04 课程内容与设计

05 课程学习材料

课程内容



课程具体设计



https://www.xueyinonline.com/detail/215844957

先修课程要求

- 理论课程（建议）：
  - 概率论与数理统计
  - 高等数学
- 编程课程（建议）
  - 具有Python/R 编辑基础
  - 有意愿认真学习一门编程语言
- 不建议同时选修课程：
  - 《数据挖掘》课程
  - 《机器学习》课程

课程考核计划

- 考勤及课堂表现(10%)：
  - 随机点名
  - 课堂表现
- 随堂测测验（20%）
- 个人作业(30%)
  - 数据分析实践
  - 数据分析方法原理练习
- 期末Project （40%）
  - 个人/团队均可
  - 总人数<=5 人
  - 不能直接使用其他课程的期末项目

讲授提纲

01 数据类型与价值使用

02 什么是数据挖掘

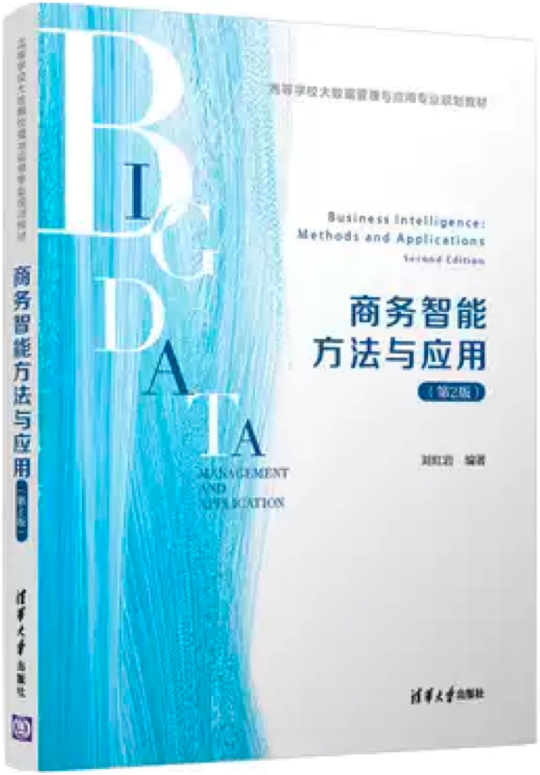
03 跨行业的数据挖掘流程

04 课程内容与设计

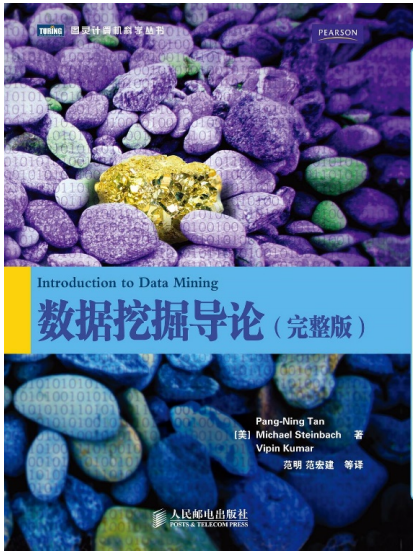
05 **课程学习材料**

课程参考教材

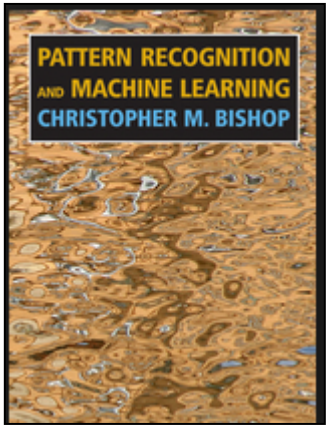
- 相关教材

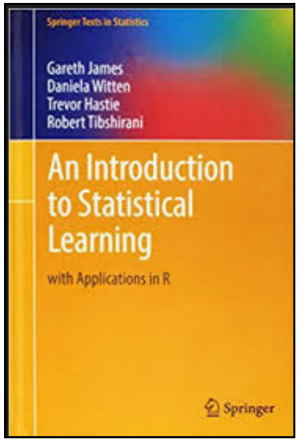

BIG DATA
Business Intelligence: Methods and Applications
Second Edition
商务智能方法与应用
（第2版）
刘红岩 编著
MANAGEMENT AND APPLICATION
清华大学出版社


数字经济专业系列教材
DIGITAL ECONOMY
商业智能
肖升生 | 主编
陈佳威 方慧 韩潇 李艳红 | 副主编

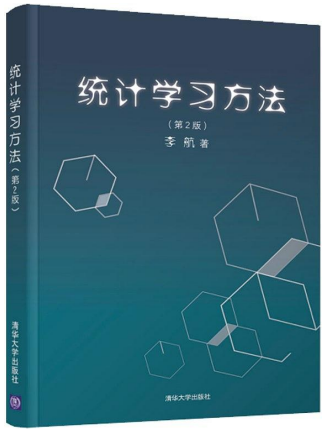

PEARSON
Introduction to Data Mining
数据挖掘导论（完整版）
Pang-Ning Tan
[美] Michael Steinbach 著
Vipin Kumar
范明 范宏建 等译
人民邮电出版社

课外阅读材料


PATTERN RECOGNITION
AND MACHINE LEARNING
CHRISTOPHER M. BISHOP


Springer Texts in Statistics
Gareth James
Daniela Witten
Trevor Hastie
Robert Tibshirani
An Introduction to Statistical Learning
with Applications in R
Springer


MACHINE LEARNING
机器学习


统计学习方法
（第2版）
李 航 著
清华大学出版社

更多学习资料

- 理论学习
  - ACM SIGKDD Conference on Knowledge Discovery and Data Mining
  - International Conference on Machine Learning
  - International Conference on Data Mining
  - IEEE Transactions on Knowledge and Data Engineering
- 实践学习
  - 天池大赛: https://tianchi.aliyun.com/
  - Kaggle: https://www.kaggle.com/

数据挖掘与商务分析

400年前发明了显微镜，改变了测量的标准，人类研究物体的细微程度从此不同。

大数据分析带来的变革，就像400年前的显微镜一样，我们能够掌握事件、行为的精细程度，也将从此进入全新的境界。

—— Erik Brynjolfsson