上海财经大学
SHANGHAI UNIVERSITY OF FINANCE AND ECONOMICS

数据挖掘与商务分析

第1讲 课程导论

主讲教师：肖升生

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 Economic

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

关于同学们

- 是否先修过高等代数，线性代数和概率论与数理统计?
- 是否先修或者自学过数据分析类的课程?
- 编程经验?
- 项目经验?
- 希望从本课程中学到什么?

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

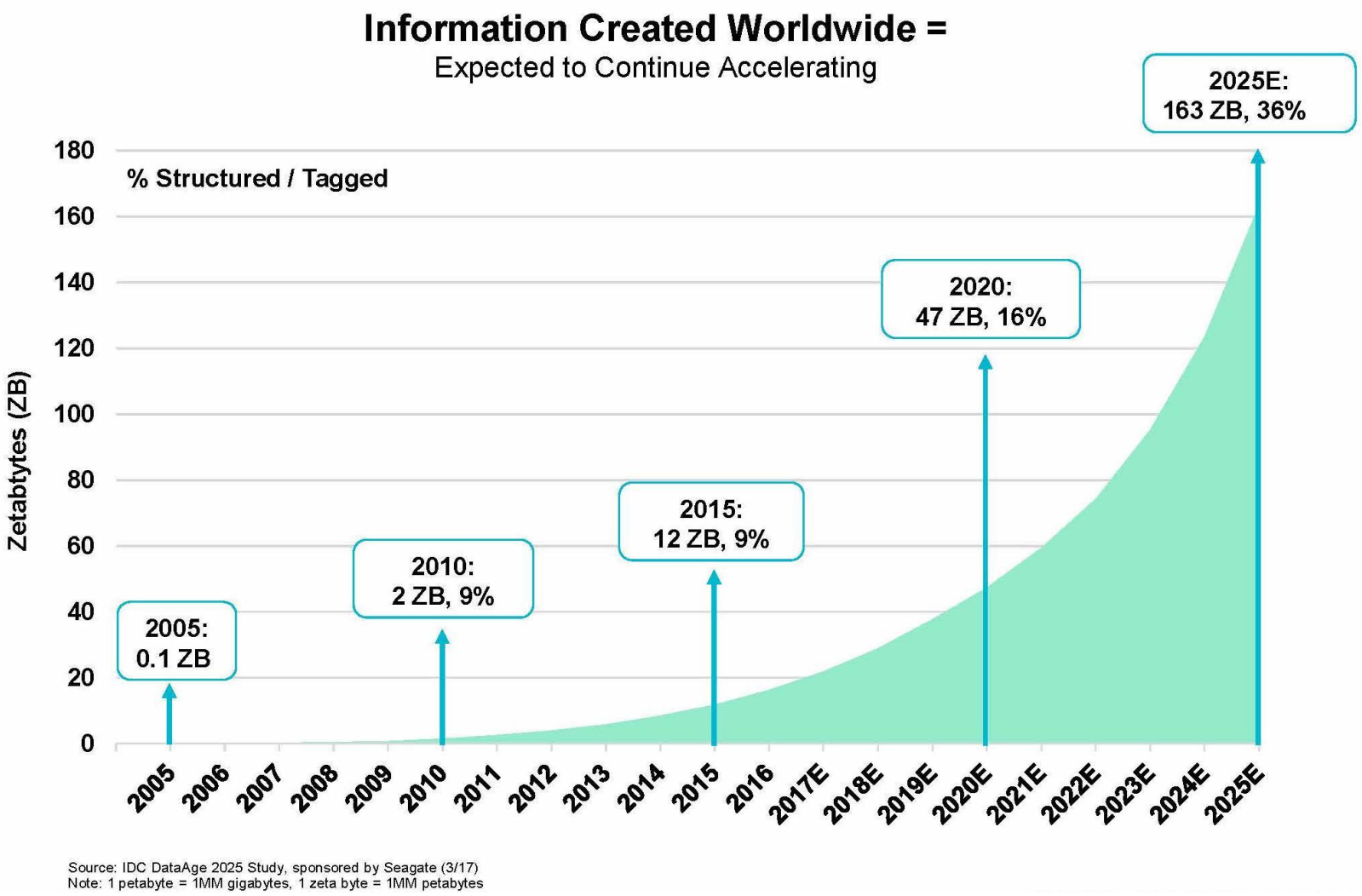
05 课程学习材料

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

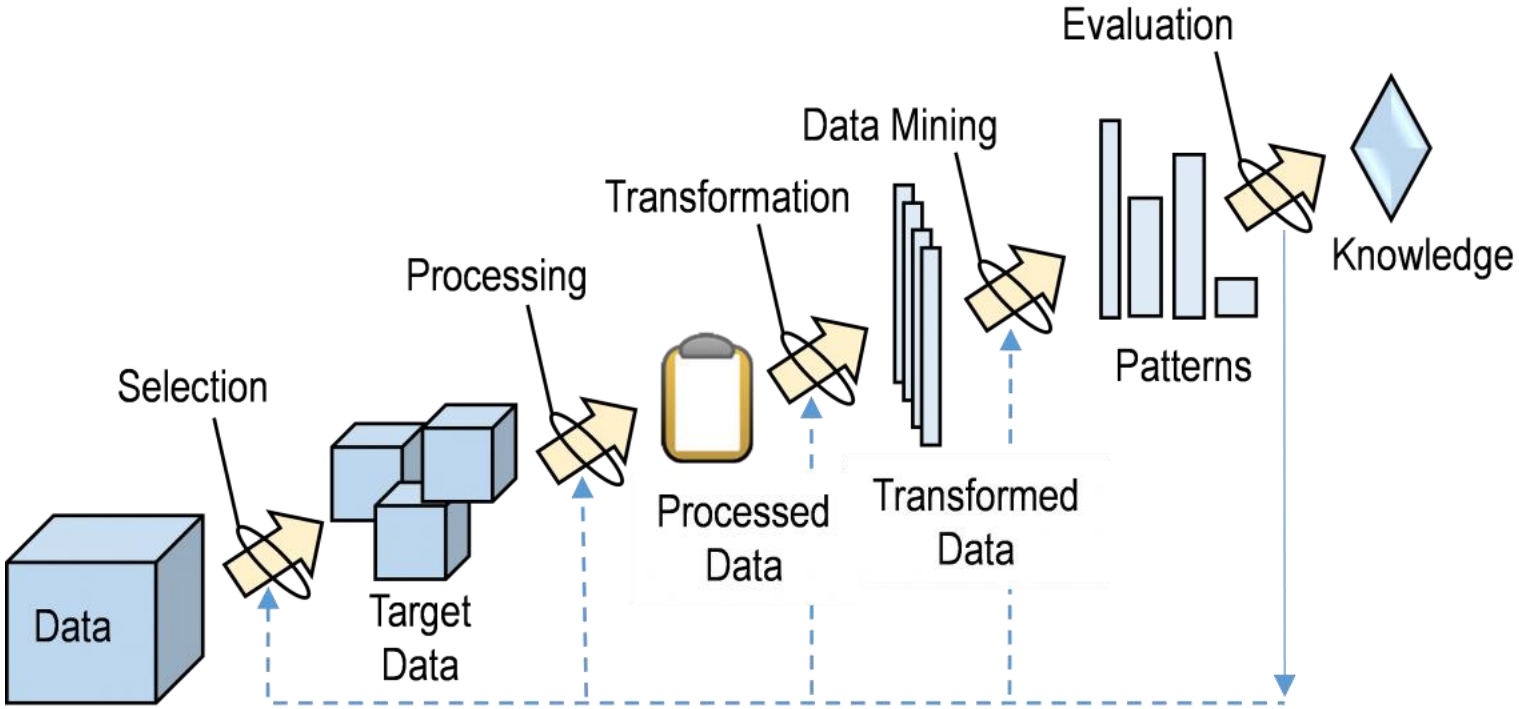
03 跨行业的数据挖掘流程

04 课程内容与设计

05 课程学习材料

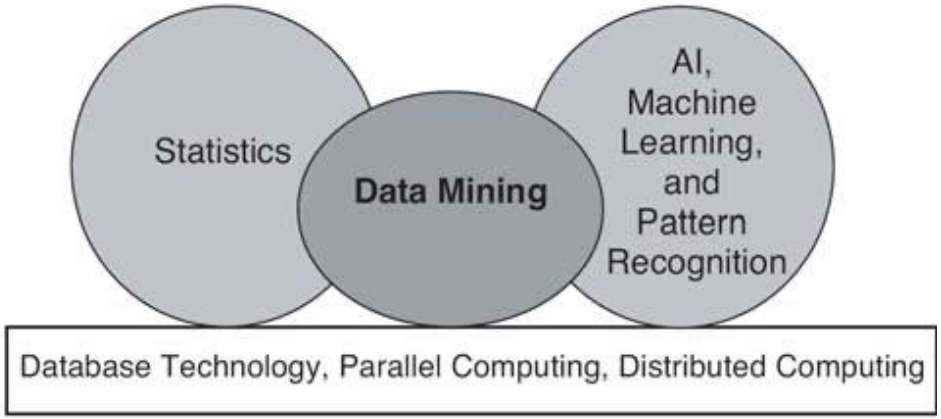
什么是数据挖掘

- 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



数据挖掘的起源

- 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
- 可借鉴的方法论来源
  - 统计学
  - 人工智能
  - 机器学习
  - 数据库技术
  - 分布式计算



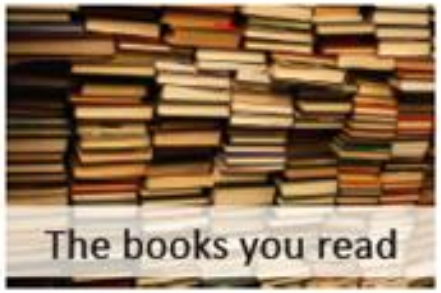
数据挖掘基本任务

- 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

- 数据挖掘的具体内容
  - 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

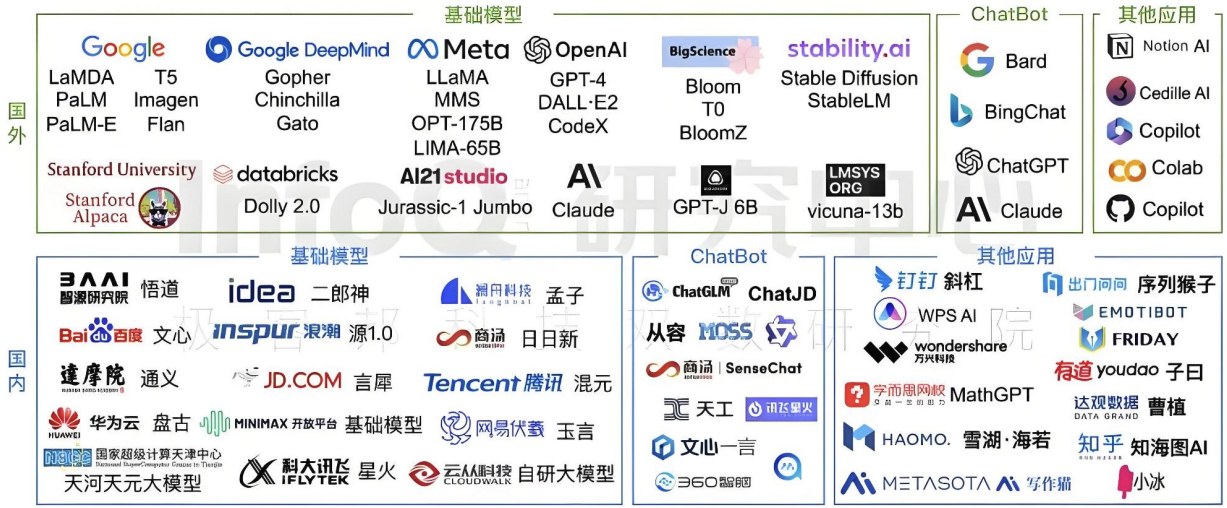

The way you're tracked


Your financial health?

AI大模型

AI大模型：指拥有超大规模参数（通常在十亿个以上）、复杂计算结构的机器学习模型。它通常能够处理海量数据，完成各种复杂任务，如自然语言处理、图像识别等。

定义



典型代表：大语言模型

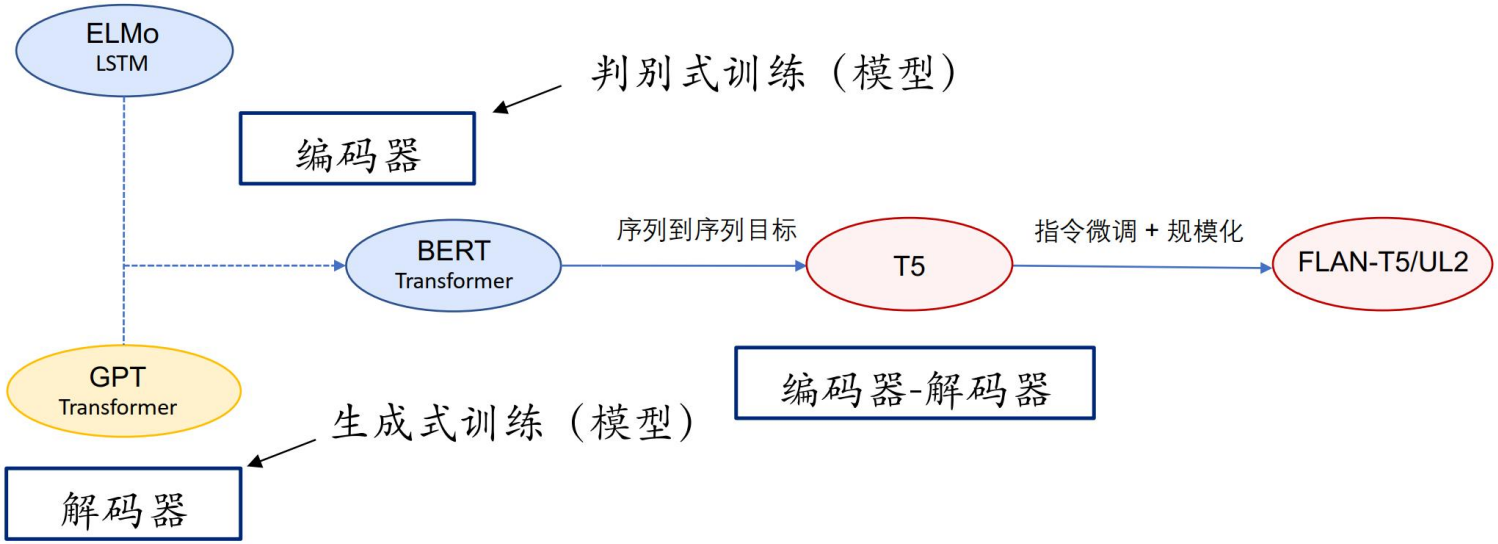
语言模型

- 语言模型：
  - 语言模型本质是在回答一个问题，即出现的语句适合合理（make sense）

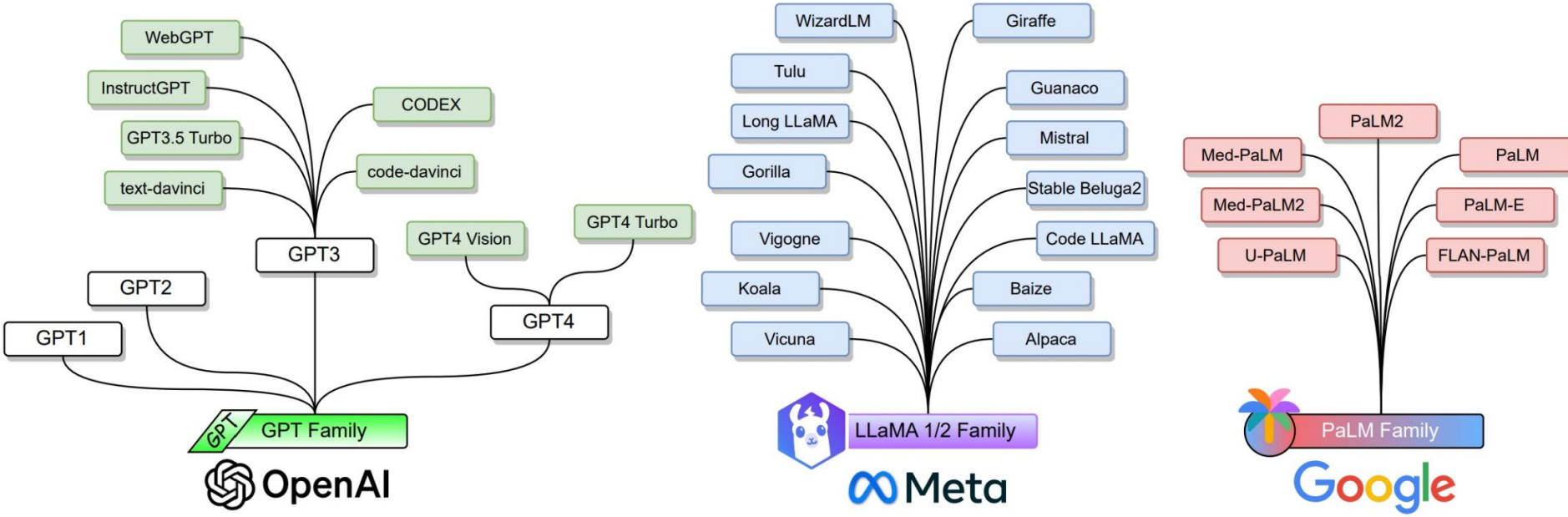
- 发展历程
  - 专家语法规则模型（至 80年代）
  - 统计语言模型（至 00年）
  - 神经网络语言模型（至今）

预训练语言模型

- **预训练语言模型**
  - 在大规模数据上**自监督训练**，经**微调或提示**后适配各类任务
- **主要技术架构**
  - 编码器：BERT、ALBERT
  - 解码器：GPT、Llama
  - 编码器-解码器：T5、BART

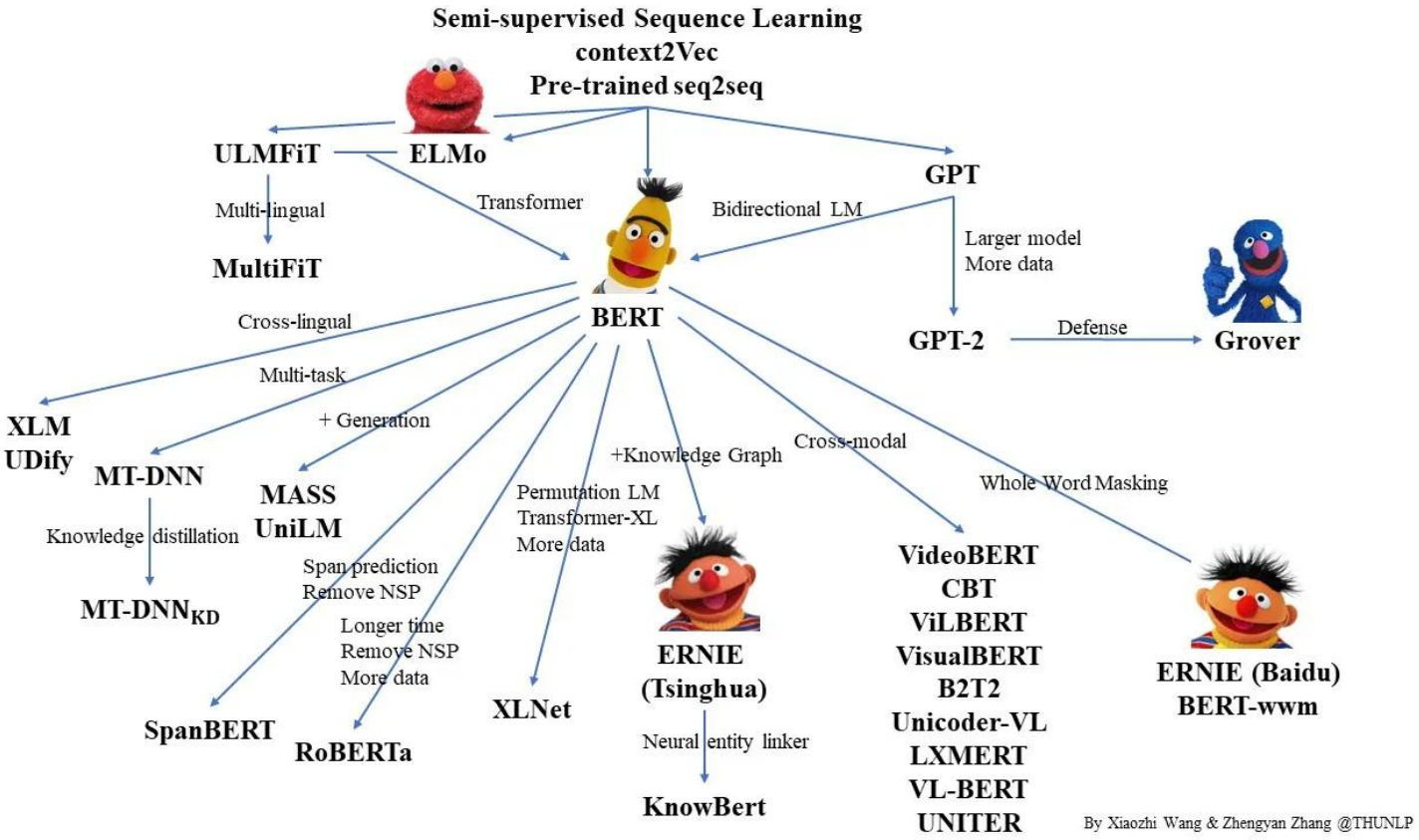


预训练语言模型发展



ShervinMinaee, TomasMikolov,NarjesNikzad,MeysamChenaghlu, RichardSocher, XavierAmatriain,JianfengGao. Large Language Models: A Survey. https://arxiv.org/pdf/2402.06196.pdf

预训练语言模型发展



预训练语言模型的脉络

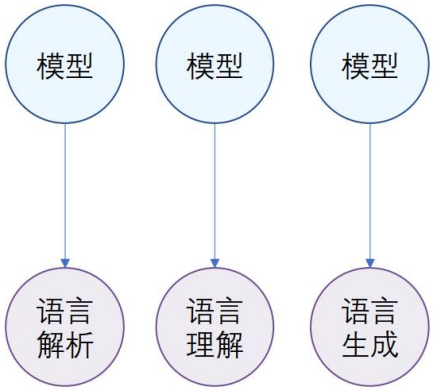
经典模型：ELMO，GPT，Bert……

大语言模型

- **大规模语言模型：** 通常指参数量超过 10B 的模型
  - 更多的计算量、推理开销更大
  - 泛化性能更强，出现涌现能力

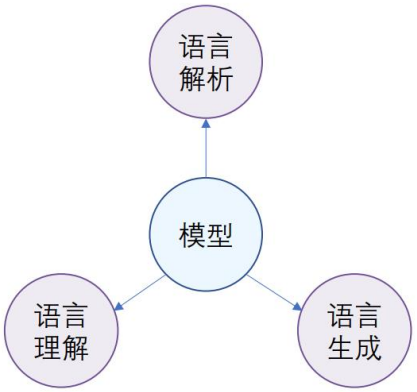
	预训练语言模型 （小模型、常规模型）	大规模生成式语言模型
典型模型	ELMo，BERT，GPT-2	GPT-3、ChatGPT、LLaMA
模型结构	BiLSTM，Transformer	Transformer
注意力机制	双向、单向	单向
训练方式	去噪自编码模型	自回归生成
擅长任务类型	理解、判断	生成
模型规模	1-10亿级参数	10-1000亿级参数
下游任务应用方式	微调	微调 & 提示学习
涌现能力	小数据领域迁移	上下文学习，思维链提示

数据与新学习范式



过去

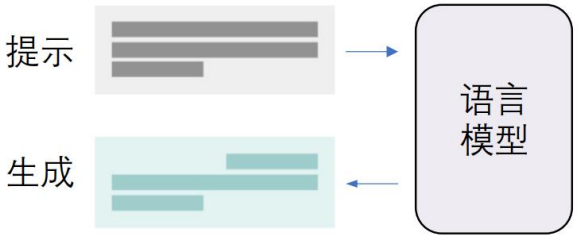
为每个任务训练独立的模型



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

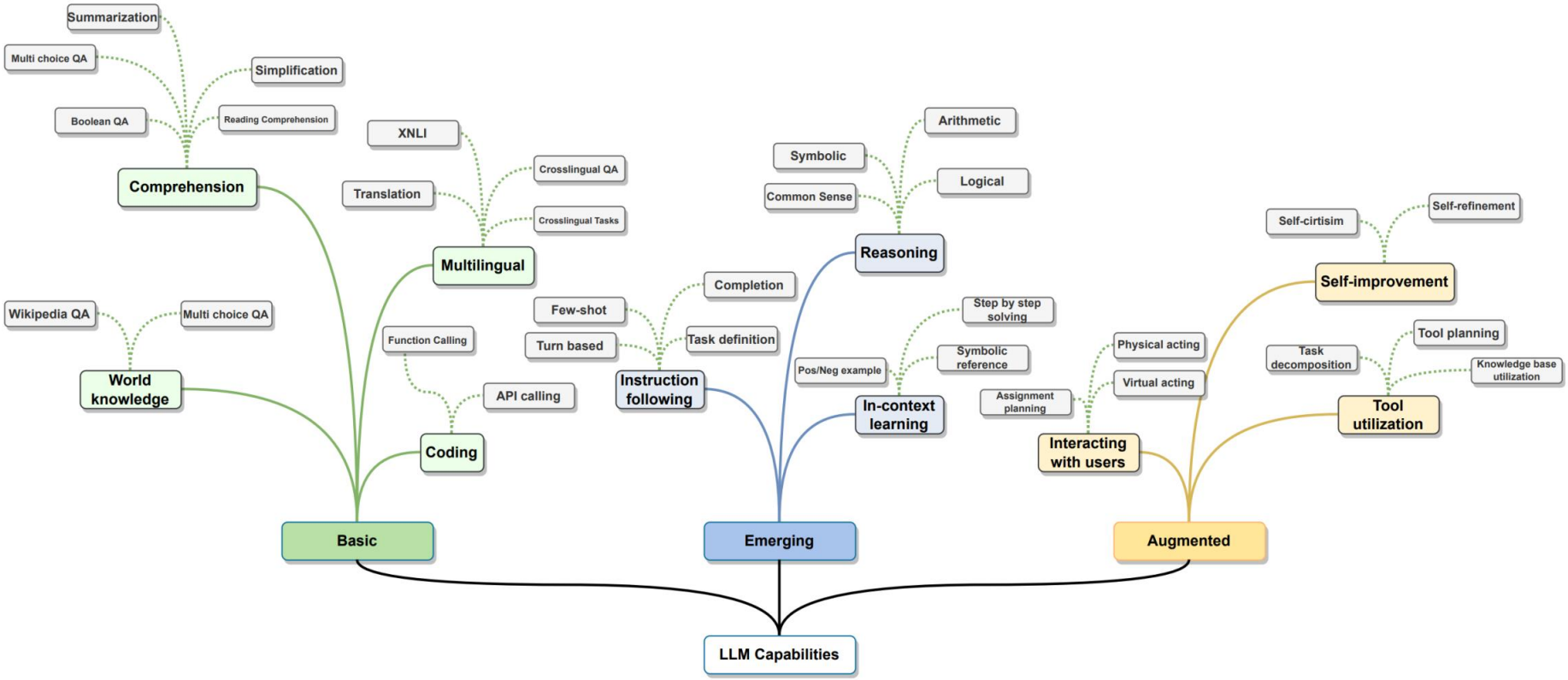
个体化训练 → 中心化训练 + 个体化微调



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



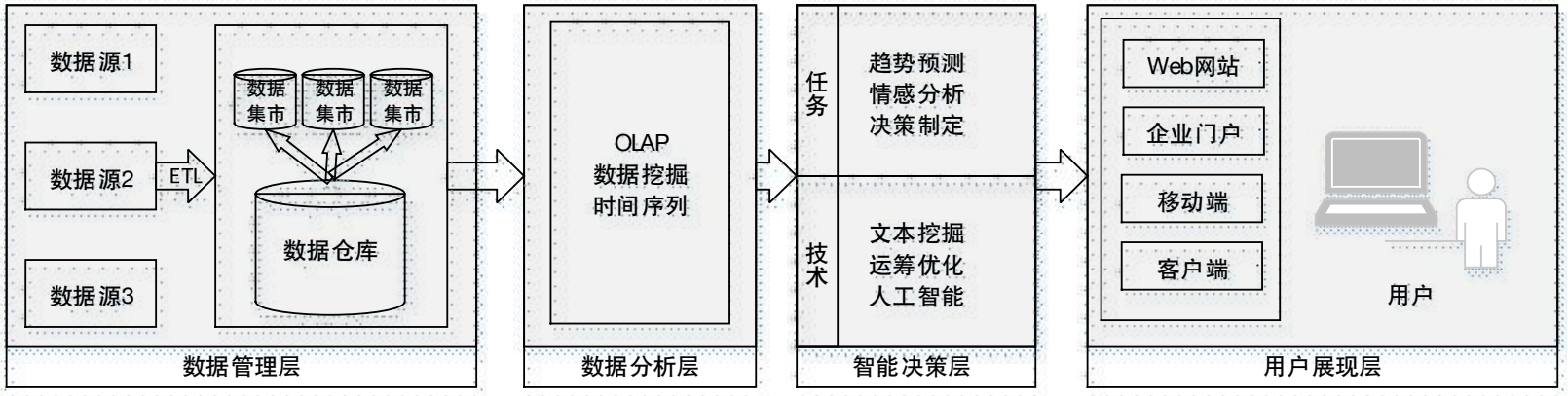
Shervin Minaee, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, Jianfeng Gao. Large Language Models: A Survey. https://arxiv.org/pdf/2402.06196.pdf

思考

- 数据挖掘与AI大模型有什么关系？
- 有了AI大模型，数据挖掘的学习还有必要吗?

商务智能

商务智能是利用计算机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。



商务智能发展的几个阶段

阶段一：数据集成

整合企业多源数据，实现数据的实时更新和统一视图。

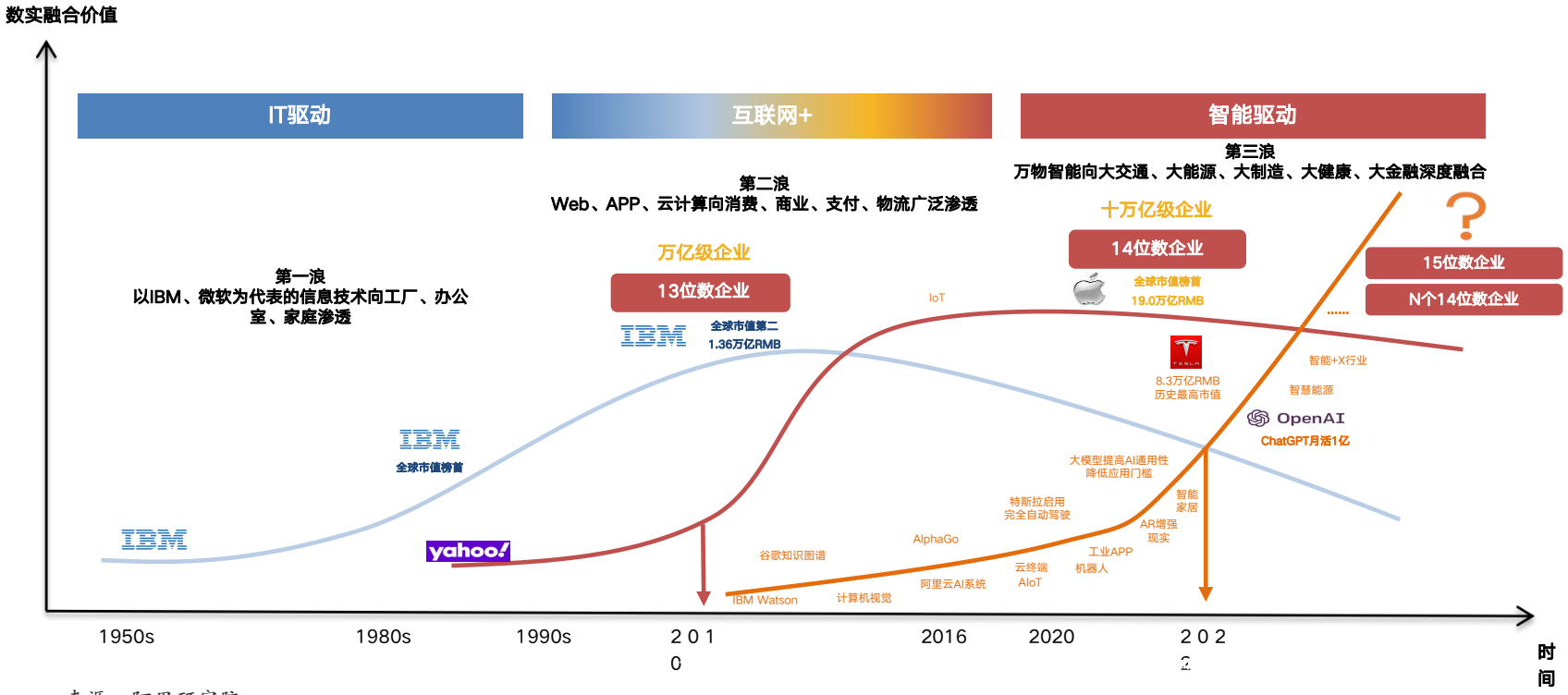
阶段二：数据挖掘

运用大数据分析和机器学习技术，实时挖掘数据中的潜在价值。

阶段三：智能决策

通过AI算法，系统能够自动提出决策建议，辅助管理者快速响应市场变化。

商务智能：数实融合的三次浪潮



来源：阿里研究院

示例: 汽车工业的三个阶段

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

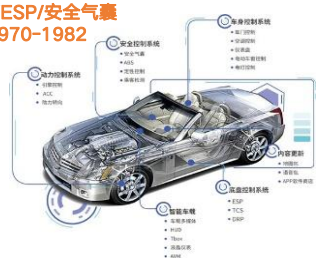
来源：阿里研究院

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

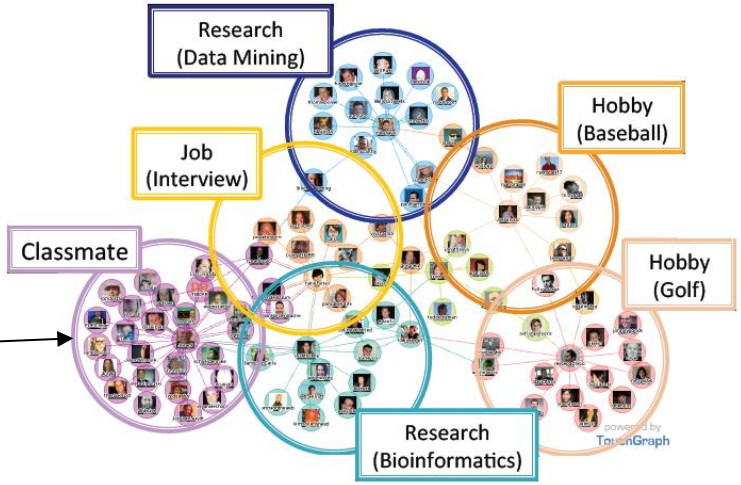
订阅式软件服务：信息娱乐、系统升级
2020以来

智能导航：路况实时交互、路线动态优化
2020以来

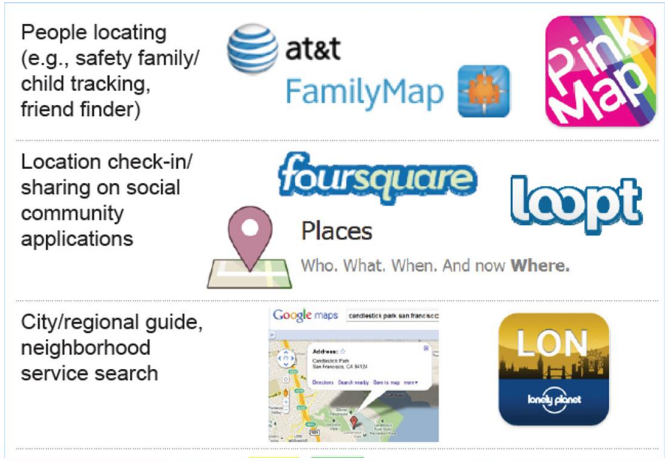
数据挖掘与商业应用

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

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

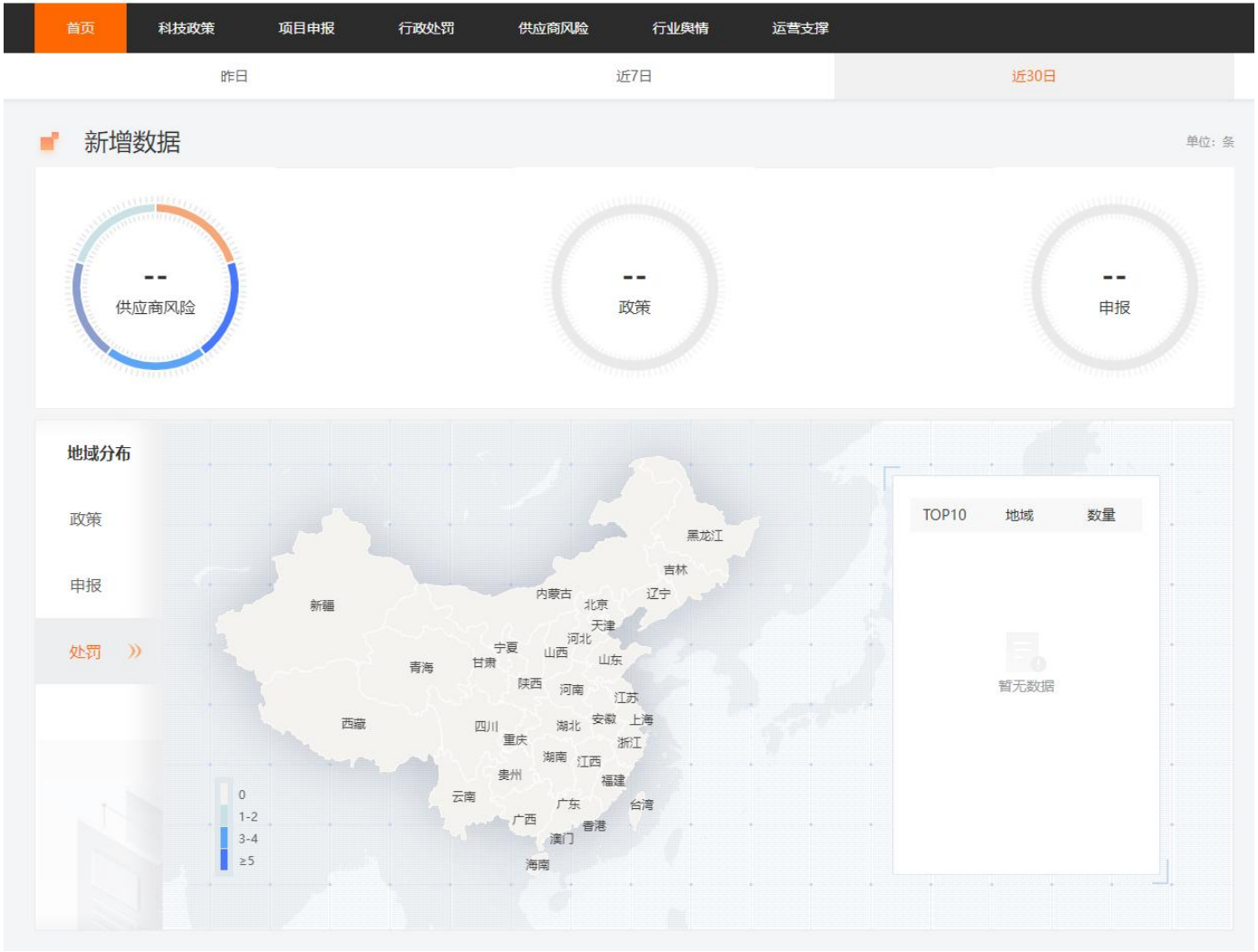
04 课程内容与设计

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


首页
科技政策
项目申报
行政处罚
供应商风险
行业舆情
运营支撑
昨日
近7日
近30日
新增数据
单位：条
--
供应商风险
--
政策
--
申报
地域分布
政策
申报
处罚
TOP10
地域
数量
暂无数据
0
1-2
3-4
≥5
黑龙江
吉林
辽宁
内蒙古
北京
新疆
天津
河北
宁夏
山西
山东
青海
甘肃
陕西
河南
江苏
西藏
四川
湖北
安徽
上海
重庆
浙江
湖南
江西
贵州
福建
云南
广东
台湾
广西
香港
澳门
海南

项目案例


科技政策
政策技术领域趋势图
2021-08
2021-09
2021-10
2021-11
2021-12
2022-01
2022-02
人工智能
区块链
云计算
大数据
信息安全
数据安全
生物特征识别
5G
运营
新增政策中央部委分布
央行
银保监会
证监会
工信部
科技部
发改委
网信办
0
1
2
3
4
5
6
7

项目案例

闹了半天，腾讯依然是国内第一大市值企业但茅台追至第二。1、3月3日消息，东方财富Choice数据，在中国所有上市公司中，腾讯控股以35885亿元的市值蝉联第一，但茅台却超过了阿里巴巴和中国工商银行，成为中国第二大市值企业，阿里巴巴近年来罕见地成为第三。工商银行排第四。 3、 腾讯股票2021年其实也

编辑 | 关注 | 转发 | 原文阅读

关键词：罚款 反垄断

资讯来源： 今日头条　　相关报道数： 0

发布时间： 2022-01-03 21:12:57　　涉及对象： 腾讯

风险分类： 业务风险　　风险等级： 显著风险

近30天的风险数

腾讯 0

正文

闹了半天，腾讯依然是国内第一大市值企业但茅台追至第二。1、3月3日消息，东方财富Choice数据，在中国所有上市公司中，腾讯控股以35885亿元的市值蝉联第一，但茅台却超过了阿里巴巴和中国工商银行，成为中国第二大市值企业，阿里巴巴近年来罕见地成为第三。工商银行排第四。 3、 腾讯股票2021年其实也康钊运营商财经网1641209927 { “rich_content” :{ “text” : “闹了半天，腾讯依然是国内第一大市值企业但茅台追至第二。1、3月3日消息，东方财富Choice数据，在中国所有上市公司中，腾讯控股以35885亿元的市值蝉联第一，但茅台却超过了阿里巴巴和中国工商银行，成为中国第二大市值企业，阿里巴巴近年来罕见地成为第三。工商银行排第四。

3、 腾讯股票2021年其实也不断下滑，打破了神话，从年内高点773.9港元一路下滑453港元。

腾讯市值从2017年一跃成为中国第一，最高曾超5万亿港元，最高排名全球第七。不过2018年腾讯股票却出现了反转，曾在3个月下跌了18.9%，市值蒸发了8000多亿港元，折合约6500亿元人民币。致使马化腾让出中国首富宝座。 当时是因为腾讯第一大股东南非公司Naspers表示，将出售2%的腾讯股份。

而2021年腾讯则是因为互联网企业的反垄断。

舆情热度 ★★☆☆☆

- 二级：舆情热度>5000

媒体分类 ★★☆☆☆

- 二级：自媒体大V

供应商相关 ★★★★★

- 五级：文章标题和首段都提到了：腾讯

风险词汇 ★★★★☆

- 四级：罚款、反垄断

讲授提纲

01 数据类型与价值使用

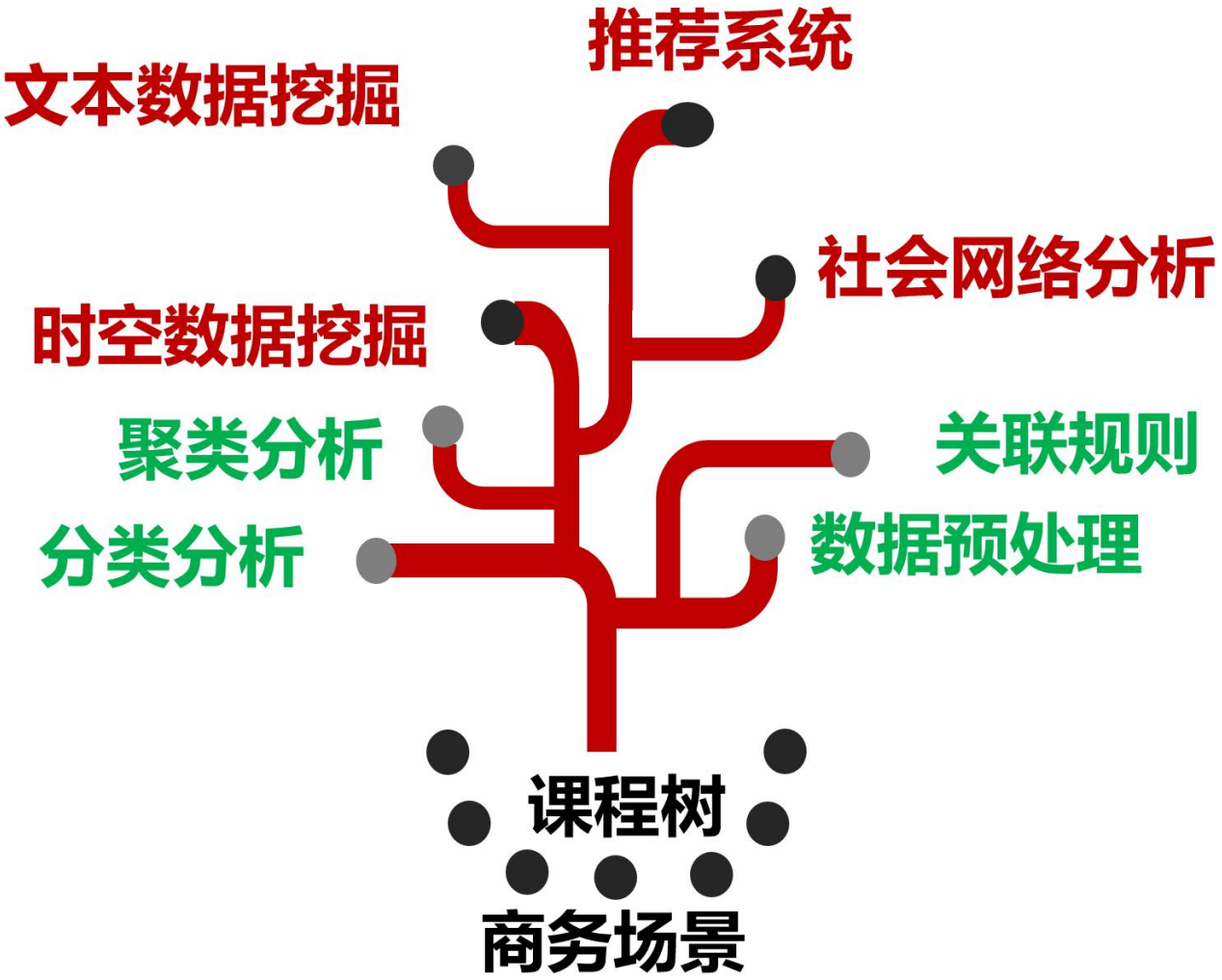
02 什么是数据挖掘

03 跨行业的数据挖掘流程

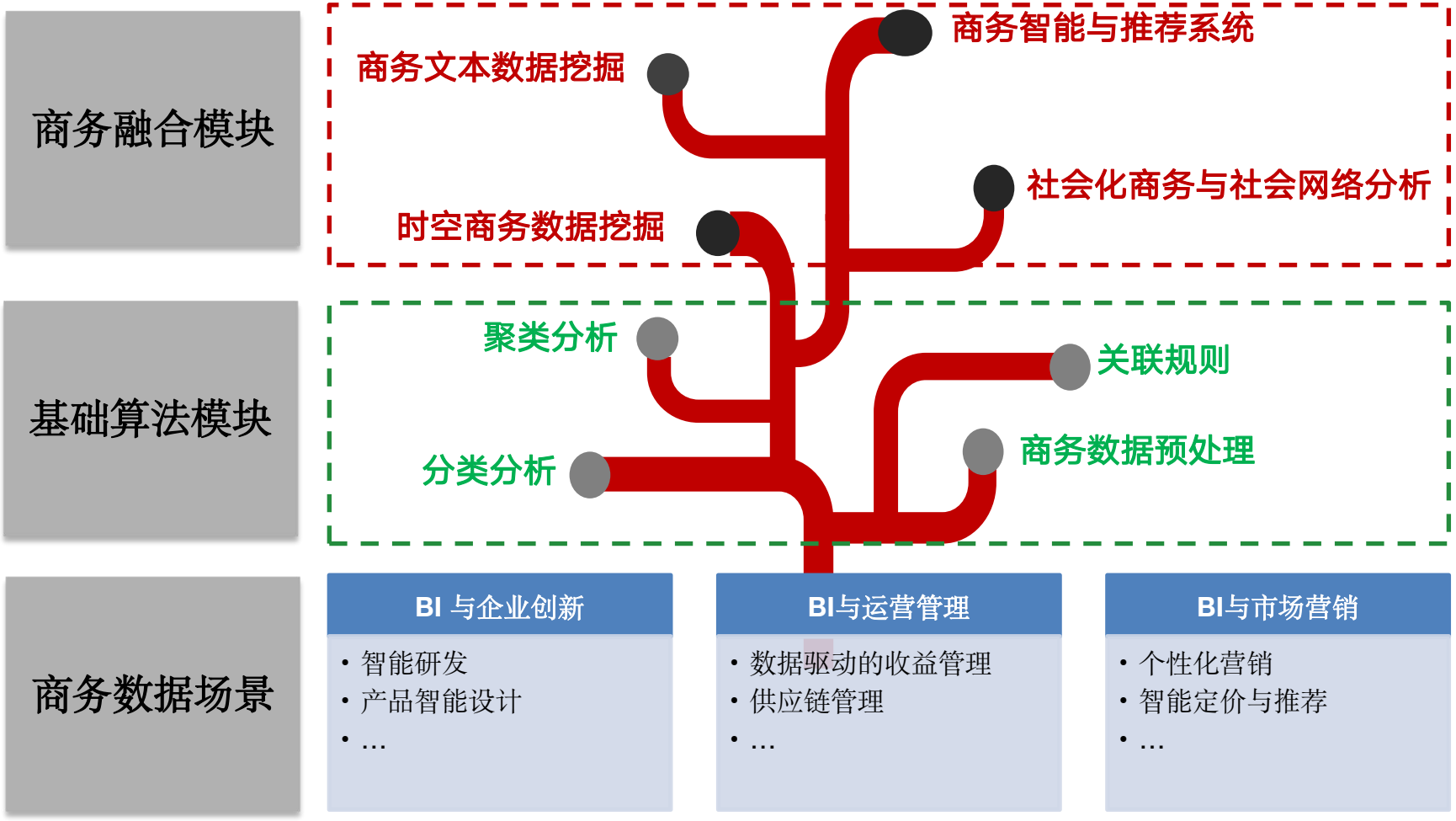
04 课程内容与设计

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
- 随堂测试 (15%)
- 个人作业(35%)
  - 数据分析实践
  - 数据分析和方法原理练习
- 期末Project (40%)
  - 个人/团队均可
  - 总人数<=3 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

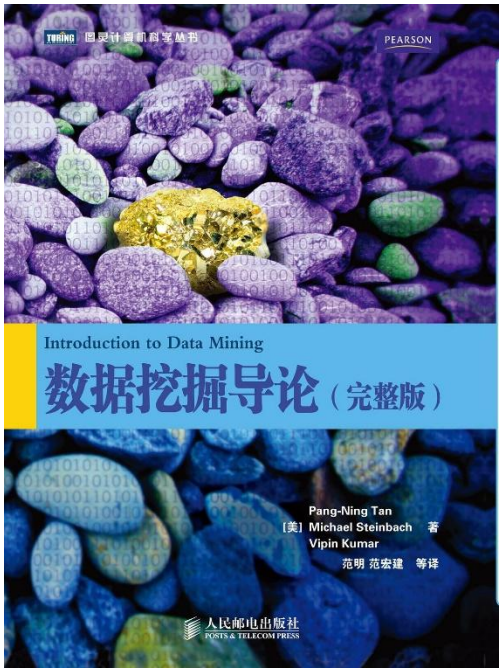
03 跨行业的数据挖掘流程

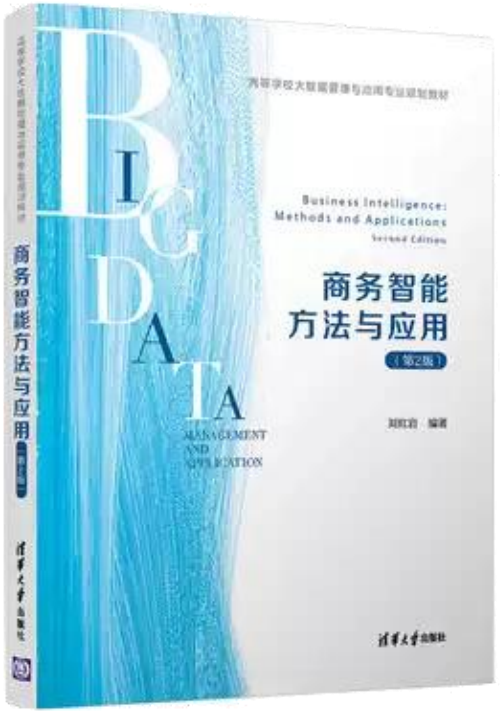
04 课程内容与设计

05 课程学习材料

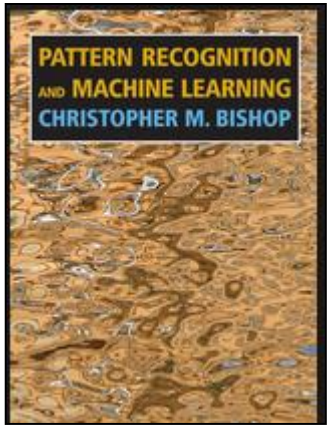
课程参考教材

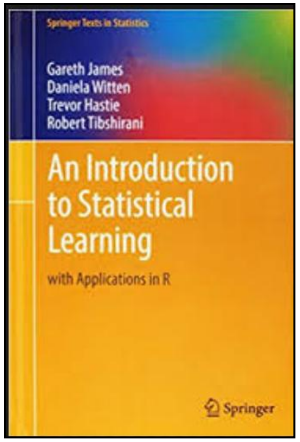
- 参考教材


TURING 图灵计算机科学丛书
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
POSTS & TELECOM PRESS

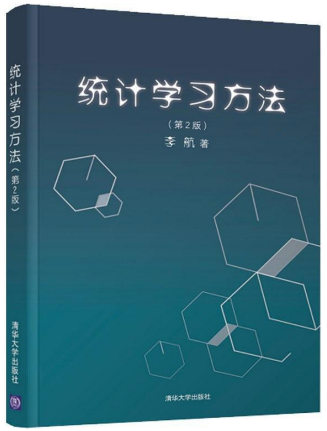

BIG DATA
Business Intelligence: Methods and Applications
Second Edition
商务智能方法与应用
（第2版）
MANAGEMENT AND APPLICATION
清华大学出版社

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