工欲善其事，必先利其器

——创新项目统计分析方法及报告撰写

报告人　吴纯杰 教授

汇报人介绍：吴纯杰

上海财经大学
SHANGHAI UNIVERSITY OF FINANCE AND ECONOMICS

- **职务**
  - ●上海财经大学统计与管理学院副院长，教授，博士生导师。
- **研究成果**
  - ●主持2项国家自科基金项目和1项国家统计局重大项目
  - ●省部级项目3项，省部级奖励8项
- **教学建设和人才培养**
  - ●市级教学成果奖2项，市级教学项目3项，校级教改项目5项；
  - ●《数理统计》市精品课程，**国家一流本科课程**；
  - ●国家级**一流统计学本科专业**建设点负责人；
  - ●《统计软件》校重点、实验课程；
  - ●《数据世界探秘》校通识核心课程。
  - ●指导学生：

全国统计建模大赛**二等奖1项，**三等奖2项；

SAS高校数据分析大赛**冠军2届**，季军3届，四强1届；

“挑战杯”上海市二等奖、**特等奖，**国赛**一等奖，**“知行杯”等比赛省部级和国家级奖10余项。

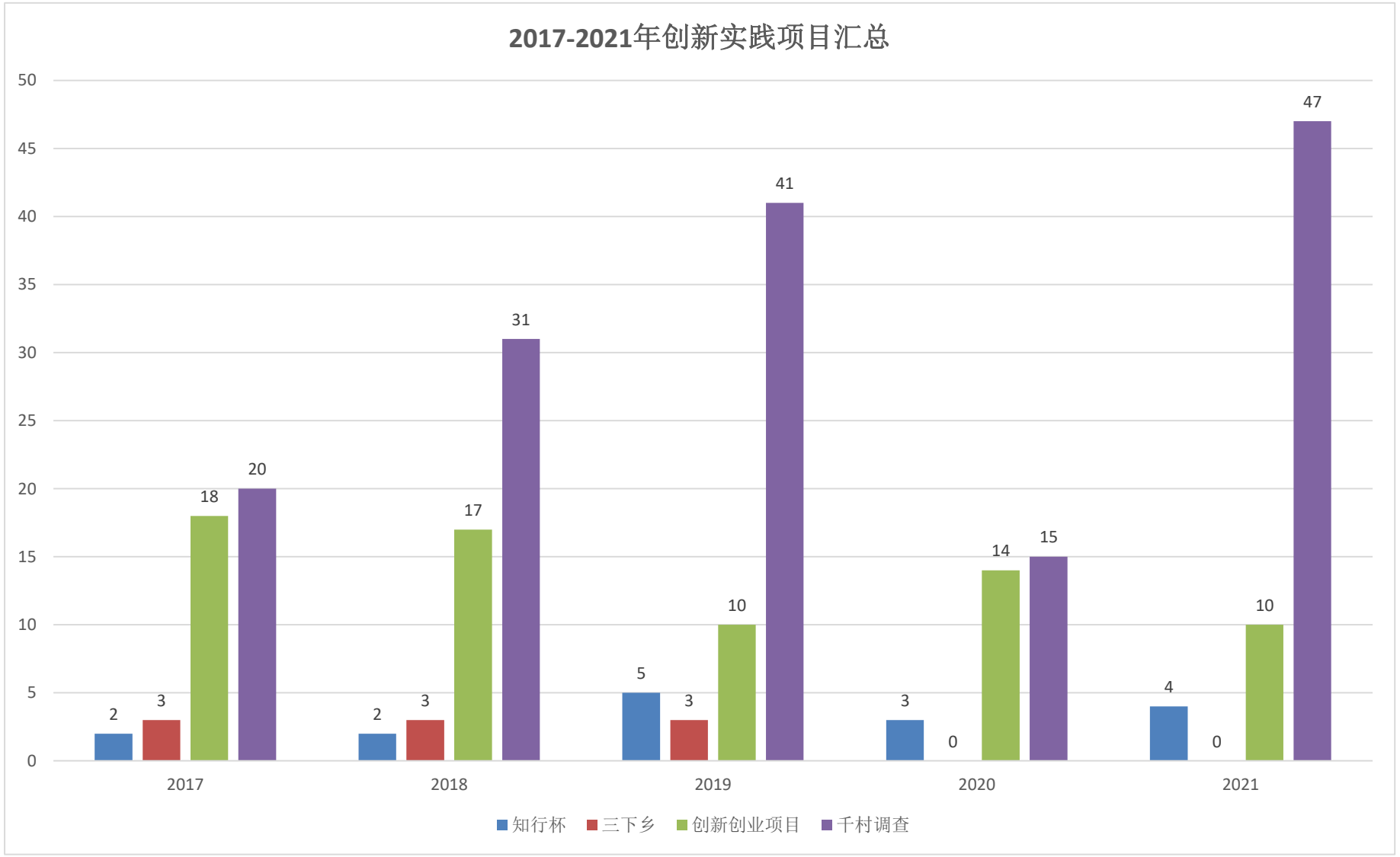
目录

- 大创项目分析
- 统计建模技术
- 统计建模工具
- 项目报告撰写

PART
一

大创项目分析

统管学院创新实践项目现状



上海财经大学
SHANGHAI UNIVERSITY OF FINANCE AND ECONOMICS

统管学院创新实践项目现状(近5年)

年度	项目名称	获奖等级
2017	大数据时代背景下的大数据文献计量	教育部
	基于数据挖掘对网络购物虚假评论的研究	教育部
	关于国内影院线上定价机制的优化分析	教育部
	创新实践的发展——个性化实用化校园纪念品的市场推广	教育部
	上海、北京两地雾霾形成影响因子分析及对比	教育部
	国内消费者对纯电动汽车购买行为因素探究	教育部
	基于网络直播弹幕的社会舆情传播模型研究	教育部
	基于机器学习理论的多种房价预测模型与传统模型的对比研究	上海市
	小红伞出租项目（上海财经大学校园版）	上海市
	基于层次分析法和ANP理论的互联网金融模式综合评价体系及行业前景预测	上海市
	针对OFO的研究分析及创新优化(以上海高校为例）	上海市
	探索不同促销侧率对广告竞价消费的激励效果——基于搜索引擎公司视角	上海市
	文具酱之校园文具配送	上海市
	上海市杨浦区养老服务的现状及其发展分析	上海市
	基于人事数据的大企业优秀员工“跳槽”原因分析及预测研究	上海市
	基于c++的扫雷分析与模型建立	上海市
	通过调查课表缓解食堂就餐拥挤问题	上海市
	制造企业经济资本实证研究	上海市
	大学生挑战杯，《京津冀冬季雾霾真的无药可救了吗？》	“挑战杯”全国大学生课外学术科技作品竞赛上海市二等奖
	京津冀冬季雾霾真的无药可救了吗？——基于农村贫富差距与城乡二元结构的经济视角	第三届中国“互联网+”大学生创新创业大赛全国铜奖
	青租界	第三届中国“互联网+”大学生创新创业大赛上海赛区特等奖
	青租界	第二届“汇创青春”——大学生文化创意作品展示活动上海市一等奖
	指尖非遗	第七届全国大学生电子商务“创新、创意及创业”挑战赛上海市三等奖
	票卜类证券文娱交易平台	第七届全国大学生电子商务“创新、创意及创业”挑战赛上海市二等奖
	电子食堂管理平台	“创青春”全国大学生创业大赛上海市铜奖

统管学院创新实践项目现状

年度	项目名称	获奖等级
2018	江南古镇在商业化发展模式下的现状与对策	校级优秀
	探究上海高校官方微信公众号文章点击率的影响力要素	校级优秀
	利用深度学习处理图像识别问题	校级优秀
	基于强化学习的上海市信号灯联动系统	校级优秀
	公交车班次动态化建议研究	通过
	寿险销售人员业务风险评估模型研究	通过
	集成学习框架下的个性化推荐算法研究	通过
	基于以共享单车为例的“共享经济”研究的“Prent”打印机租赁服务	通过
	“三尺店”——基于O2O的校内及校间二手买卖平台	通过
	以上海制药企业为切入点探究两票制影响	学院经费支持，通过
	基于蚂蚁花呗等消费贷对当代大学生影响的研究及花呗优化模型设计	学院经费支持，通过
	青年人社区治理与社区参与情况及其对基层群众自治推动性的分析--以上海市为窗口	学院经费支持，通过
	基于自然语言处理与统计方法对公众号文章标题与阅读量关系的研究	学院经费支持，通过
	研究网络直播对大学生的影响及如何“趋利避害”	学院经费支持，通过
	失能老人养老现状及问题研究-以上海杨浦区为例	学院经费支持，通过
	中小城市废弃物资源的回收与利用	学院经费支持，通过
	SUFE拿了么——学生代拿快递送货上门服务	学院经费支持，通过
2019	探究上海高校官方微信公众号文章点击率的影响力构成要素	挑战杯上海市二等奖
	垃圾普及化与常态化发展机制探究	挑战杯上海市二等奖
	依托旅游资源禀赋，推动乡村生态旅游发展	挑战杯上海市二等奖
	江南古镇在商业化发展模式下的现状与对策	挑战杯上海市二等奖
	互联网教育平台的现状、痛点与发展情景探究——以B2C平台为例	通过
	大数据人才供需匹配度研究	通过
	新能源汽车购买驱动因素研究——以上海市为例	校级优秀
	旅游古镇同质化趋势研究	校级优秀
	金融科技转型下上海金融市场人才需求探索—基于大数据建模分析	通过
	“爸爸妈妈去哪儿了”——以家长学校为例探讨提高家长陪伴质量的办法	通过
	微博流量真实度检测方法和影响分析	通过
	寻找长三角民宿经济的蓝海——对长三角民宿行业的分析（以上海为例）	校级优秀
	基于多信息源与文本挖掘的智能手机销量影响因素研究与决策建议	通过
	微博平台下的水军识别和面向用户的水军提示阅读器	通过
	垃圾分类普及化与常态化机制探索——以上海市奉贤区为例	校级优秀

统管学院创新实践项目现状

年度	项目名称	项目级别	获奖等级
2020	为什么双胞胎不一样？基于家庭资源配置动态变化的视角	校级	合格
	基于空间面板模型的气象因素与经济增长相关性研究	校级	合格
	基于网络搜索数据上海地区手足口病发病率组合预测模型的比较研究	校级	合格
	无人配送在城市物流领域的应用评估和市场优化	校级	合格
	移动游戏市场现存问题的探究	校级	合格
	新环境下教育行业趋势分析与教育企业破局策略	校级	合格
	基于分级诊疗制度的医疗机构推荐系统	校级	优秀
	我国线上老年教育可行性方案探究	校级	合格
	对互联网时代下网红经济发展现状和未来趋势的探究	校级	合格
	互联网医疗平台在上海市推广的现状、瓶颈与对策建议	校级	优秀
	老有所乐:基于心理维度的老年人健康预期寿命的测算研究及实证分析——以上海市为例	市级	合格
	区块链在物流行业应用场景的价值分析与技术实现	市级	合格
	区域软实力对区域经济发展水平的影响	市级	合格
	关于上海市15分钟社区医疗圈建设的分析评估以及优化建议	市级	优秀
2021	疫情冲击下惠企纾困政策分析与成效调研		挑战杯上海市特等奖
	老有所依，漂有所乐”--基于上海市数据的老漂族生活质量调研及对策研究		挑战杯上海市特等奖
	基于分级诊疗制度的医疗机构推荐系统		挑战杯上海市三等奖
	“禁塑令”落实成效与发展影响探索——以上海市“进一步加强塑料治理方案”为例	市级	通过
	当代青年养生图鉴——年轻群体保健品市场及用户画像探究	市级	优秀
	大学生低年级实习的影疫情下上海财经大学毕业生择业变化及相关企业招聘的创新措施响分析——以上海财经大学为例	校级	通过
	基于POI的上海市杨浦区“15分钟社区生活圈”适老化设施发展水平研究	国家级	优秀
	突发事件下全球贸易网络变化的趋势	校级	通过
	基于大数据方法预测社交网络信息传播趋势	国家级	通过
	统计学领域热点的发展分析与预测——基于对QS五十所统计强校的文本挖掘和社群网络分析	市级	通过
	影响直播带货经济效益因素的探究	校级	优秀
	农村家庭燃料使用变迁及其对健康的影响研究	市级	通过
	一线城市大学生主动返乡就业因素分析及建议——以上海为例	市级	通过

2014-2021年“知行杯”暑期实践统计表

项目名称	时间	带队学生	指导老师	获奖情况
有关绿色食品产业发展的研究—以上海市崇明县为例	2014	谯雅静	吴纯杰	“知行杯”大学生社会实践大赛优秀奖
大学生“手机人”族群化现状研究	2014	杨澜	宋达飞	“知行杯”大学生社会实践大赛一等奖
上海市电子垃圾产业化现状研究	2015	吴天昊	吴纯杰	“知行杯”大学生社会实践大赛优胜奖
对上海市城乡社区卫生服务体系资源配置不均问题的研究	2015	赵菁菁	李涛	“知行杯”大学生社会实践大赛三等奖
“碳测”PM2.5软件开发	2015	艾自卷		
社会公共服务短板谁来补--上海市城乡社区服务类民办非企业单位的调查报告	2016	王雪琳	吴纯杰	“知行杯”大学生社会实践大赛三等奖
休闲农业对新一轮农业供给侧结构性改革的价值探讨	2017	王艺博	吴纯杰	“知行杯”大学生社会实践大赛二等奖
上海市众创空间发展现状及对策建议	2017	王啸	柏杨	“知行杯”大学生社会实践大赛三等奖

2014-2021年“知行杯”暑期实践统计表（续）

项目名称	时间	带队学生	指导老师	获奖情况
DT 时代统计人才社会需求及毕业生契合度调查	2018	姚航	张鸣芳	“知行杯”上海市大学生社会实践项目大赛二等奖
大数据人才培养你知多少？——数据科学与大数据技术专业培养方案及学生认知度调查	2019	冉奚鸣	张鸣芳	“知行杯”上海市大学生暑期社会实践大赛三等奖
“旧巷新风，城市蝶变”——上海市城中村改造的社会影响调查	2019	邹佳旺	李涛	“知行杯”上海市大学生社会实践项目大赛二等奖
“老有所依，漂有所乐”——基于上海市数据的“老漂族”生活质量调研及对策研究	2020	门嘉齐	李涛	“知行杯”上海市大学生社会实践项目大赛一等奖
“珍馐盈门户 烟火气复来”——基于上海市典型区域食品地摊经济的消费者满意度影响因素调研	2020	王诚嘉	朱倩倩	“知行杯”上海市大学生暑期社会实践大赛三等奖
“扶贫旧貌换新颜”——助农直播打响脱贫攻坚收官战	2020	邬春妮	李涛	“知行杯”上海市大学生暑期社会实践大赛三等奖

项目选题原则

1. 关注“十四五”规划和2035中长期规划，政府工作报告
2. 经济社会民生热点、国际焦点
3. 选题切记大而空，聚焦一点来做
4. 有的题目虽好，但不具备可行性（数据获取）
5. 关注政府和学术公众号，比如“上海发布”、“人民日报”、国家统计局等
6. 研究对象范围不能太小，（获奖级别与研究区域范围对应）
7. 多读高质量学术论文（科研处、图书馆，权威A、B和核心期刊）

七、优先发展农业农村，全面推进乡村振兴

九、繁荣发展文化事业和文化产业，提高国家文化软实力

十二、改善人民生活品质，提高社会建设水平

http://www.gov.cn/zhengce/2020-11/03/content_5556991.htm

中共中央 国务院关于加强新时代老龄工作的意见

2021-11-24 19:40 来源： 新华社 【字体：大 中 小】 打印

新华社北京11月24日电

中共中央 国务院
关于加强新时代老龄工作的意见
（2021年11月18日）

有效应对我国人口老龄化，事关国家发展全局，事关亿万百姓福祉，事关社会和谐稳定，对于全面建设社会主
家具有重要意义。为实施积极应对人口老龄化国家战略，加强新时代老龄工作，提升广大老年人的获得感、幸福感
现提出如下意见。

- 健全养老服务体系
- 完善老年人健康支撑体系
- 促进老年人社会参与
- 着力构建老年友好型社会
- 积极培育银发经济
- 强化老龄工作保障

http://www.gov.cn/zhengce/2021-11/24/content_5653181.htm

大创项目选题分析

大创项目主题词云图

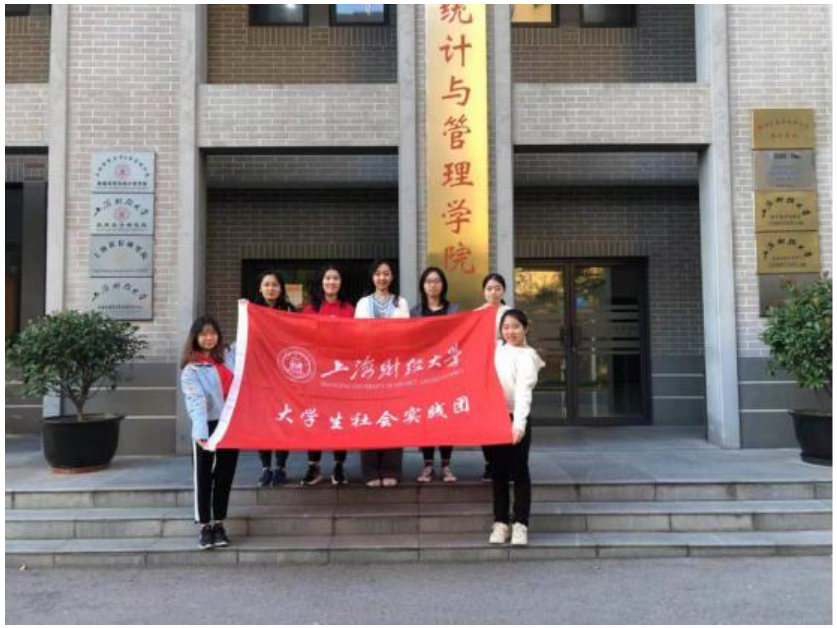
优秀作品主题词云图

优秀作品介绍

吴纯杰老师指导的王昱涵团队作品《**疫情冲击下上海帮扶企业纾困政策之落实、成效和再思考**》则通过横向对比上海各区帮扶政策的异同与个性化特色，对上海市各区县进行采访、考察与调研。小组成员实地走访部分企业并与国家统计局浦东调查队进行合作，得到政府部门相关支持，对上海市各区企业进行问卷调研，调研分析疫情期间企业纾困政策在不同方面的成效与不足和企业对政策的满意程度与意见，最后结合政策落实与企业需求，提出具有借鉴意义与创新思维的建议措施，以促进帮扶政策的有效实施与改进。该项目在2021年第十七届“挑战杯”上海市大学生课外学术科技作品竞赛中荣获**特等奖**，国赛一等奖。

优秀作品介绍

李涛老师指导的门嘉齐团队作品《“老有所依，漂有所乐”——基于上海市数据的老漂族生活质量调研及对策研究》以上海“老漂族”为研究对象，旨在有针对性的提高老漂族生活质量，为构建更为完善的公共服务体系、社区人文关怀提供有价值的参考建议。项目组通过建立生活质量评价指标体系，利用线上问卷、深度访谈、社区采访等调研手段，获取“老漂族”在上海和在其原住地的生活质量指标数据，计算生活质量评分。再通过配对t检验对比分析“老漂族”两地生活质量并建立逐步回归模型，获得影响“老漂族”生活质量的显著因素，针对性地向政府、社会、社区三个服务提供主体提出了改善“老漂族”生活质量的建议。本项目在2020年“知行杯”社会实践比赛中获得上海市一等奖，在2021年第十七届“挑战杯”上海市大学生课外学术科技作品竞赛中荣获**特等奖**，国赛三等奖。



上海财经大学
SHANGHAI UNIVERSITY OF FINANCE AND ECONOMICS

统管学院创新实践项目现状

刘旭老师指导的郭倩团队作品《基于分级诊疗制度的医疗机构推荐系统》基于当下医疗中普遍存在的“看病难”，医疗资源稀缺与浪费等问题，结合目前新医改的关注焦点分级诊疗制度引入“互联网+分级诊疗”，试图建立一个“等级推荐——个性择医——信息查询——反馈评价”的一体式分级诊疗系统，根据病人自己给出的病情描述，嵌入随机森林和XGBoost等机器学习方法进行高精度文本分类，得出一个病情严重等级，并由该等级为病人推荐合适的医疗机构、定制个性化的就诊方案。本文引入“互联网+分级诊疗”的基本框架，希望通过优质的用户体验让群众主动了解并接受分级诊疗的概念，创造新的发展机遇。该项目在2021年第十七届“挑战杯”上海市大学生课外学术科技作品竞赛中荣获三等奖。

统管学院创新实践项目现状

吴纯杰老师指导的徐音团队作品《垃圾分类普及化与常态化发展机制探索》，通过与政府部门、相关单位展开对接合作，考察垃圾分类减量的运行机制，寻求建立创新型垃圾分类引导监管模式和评价体系。项目组先后进行了近10场深度访谈与座谈会，获取了800余份居民问卷，建立了了解度和满意度评价模型，从而检测机制的执行效率和影响程度。项目还利用双界二分式条件意愿评估法建模分析了居民对垃圾分类的支付意愿，进一步支撑了实证调研的结论，也为未来上海市可能实行的收费制度提供了理论依据，对垃圾分类未来进入强制化时代与常态化轨道做出了合理的建议和展望。该项目在2019年第十六届“挑战杯”上海市大学生课外学术科技作品竞赛中荣获二等奖。

统管学院创新实践项目现状

黄涛老师指导的刘勉团队作品《江南古镇在商业化发展模式下的现状与对策》，则将关注点放在经济发展与传统文化保护之间的关系上。项目组前后历时近2年，发放超过1300份问卷，足迹遍布江浙沪地区38家古镇，爬取超过40000条商铺数据，旨在从供给侧和需求侧出发，建立逻辑回归模型，刻画人口旅游特征对旅游满意度的影响。项目建立自限增长模型，识别出古镇商业化进程中的5个阶段和4个特征点，提出针对性措施，并结合文化与创新这两个增长点，以石浦渔港古城和西塘为例，论证了适度的商业化发展有利于地方传统文化保护。该项目在2019年第十六届“挑战杯”上海市大学生课外学术科技作品竞赛中荣获二等奖。

统管学院创新实践项目现状

陈颖老师指导的夏璇团队作品《上海市高校官方微信号文章点击率的影响力构成要素》以促进高校官方微信公众号最大化发挥内容输出和多方互动作用为目标，利用python爬取上海8所高校官微文章发布的客观数据信息，构建多因素下的CART决策树模型、建立关键词分析指标，并利用主观问卷进行佐证，探究高校官微文章点击率的具体影响因素。项目组成员运用统计学科知识，综合宏观微观、定性定量、主客观等多方面考量，提出对现有高校官微的针对性改善建议，从而更广而精地展示和传播优秀校园文化。该项目在2019年第十六届“挑战杯”上海市大学生课外学术科技作品竞赛中荣获二等奖。

统管学院创新实践项目现状

李涛老师指导的陈睿团队《对易地扶贫就业政策满意度的评估及影响因素分析——以贵州省铜仁市万山区为例》实践项目着眼于“精准扶贫”的社会热点。团队前往贵州省铜仁市万山区，对地区当前大数据精准扶贫易地搬迁就业政策的实施情况开展调研活动。团队参与了万山区人社局政府座谈会，走访了4家向搬迁户提供就业岗位的企业，采访了数名移民搬迁户并回收精准扶贫就业政策满意度调查问卷300余份，面向当地移民搬迁户举办了2场“大数据精准扶贫”政策宣讲会，派发“大数据精准扶贫”主题宣传手册800余份。利用CHAID决策树模型和Logistic多元有序回归模型等对问卷数据进行处理，首先通过4类10项子满意度生成了总体满意度，之后通过模型拟合从26个解释变量中找到了满意度的4个关键影响因素，在此基础上，对当地“精准扶贫”易地搬迁政策的后续实施及有相同情况的地区的政策实施提出了意见及建议。该项目在2019年“知行杯”上海市大学生社会实践大赛中荣获二等奖。

统管学院创新实践项目现状

李涛老师指导的邹佳旺团队《“旧巷新风，城市蝶变”——上海市城中村改造的社会影响调查》实践项目着眼于特大型城市的精细化管理，选取上海市城中村为研究对象。项目组以寻找解决城中村改造的优化政策和方案的目标，对上海城中村的社会影响进行多层次多维度的考察，借此探究优化城中村改造的方法。以上海城中村为基点，对改造后的城中村居民进行问卷调查，考察他们在改造前、中、后三个时期的心理、经济与生活变化情况，同时也对周边人群及设施环境等数据以问卷调查、实地考察与数据爬虫的方式进行搜集。项目组走访了浦东唐镇和许浦村两地城中村，采访了当地居委会和承包改造的房地产公司，发放问卷300余份。在整理、筛选、清洗走访数据和问卷数据后，利用因子分析将两地村民的改造满意度进行对比分析，探究不同手法对城中村改造所产生的不同影响，从而为政府提供“因地制宜”的改造建议，最后项目组利用聚类分析，将城中村村民分为三类，以最大化改造满意度为目标，对每一类提出针对性政策，供政府参考。该项目在2019年“知行杯”上海市大学生社会实践大赛中荣获二等奖。

统管学院创新实践项目现状

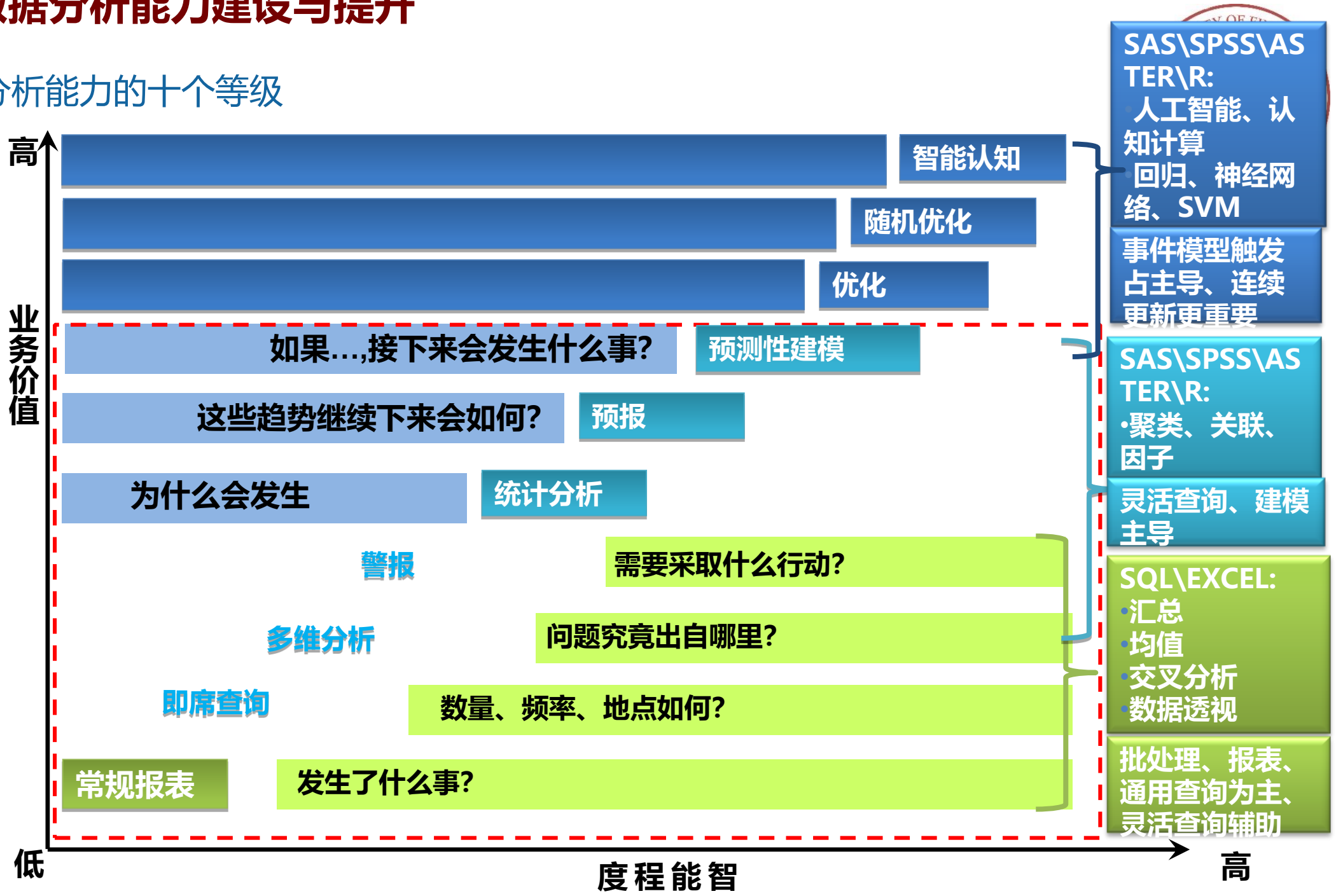
张鸣芳老师指导的冉奚鸣团队《数据科学与大数据技术专业培养方案及认知度调查》实践项目则以自身出发，对大学生的专业认知度进行调查，旨在为大学生选择大数据专业与大数据专业及就业提供有价值的参考意见，同时也希望为高校进行教育方式和专业引导提供建议。项目组在搜集了53所开设数据科学与大数据技术专业（本科）高校的培养方案，使用热力图、三维地图、雷达图来分析数据科学与大数据技术专业的开设情况，对数据科学与大数据技术专业的培养方案进行文本挖掘，进行词频分析，做出词云图。项目组同时设计了问卷进行了网络问卷采访，并前往北京、上海、浙江的13所高校进行实地调研，将所得数据通过建立结构方程模型，进行因子分析与主成分分析。双管齐下，从学校端和学生端着手探究对于数据科学与大数据技术专业认知度影响较大的因素并探究专业培养计划的改进方向。该项目在2019年“知行杯”上海市大学生社会实践大赛中荣获三等奖。

PART 一

统计建模技术

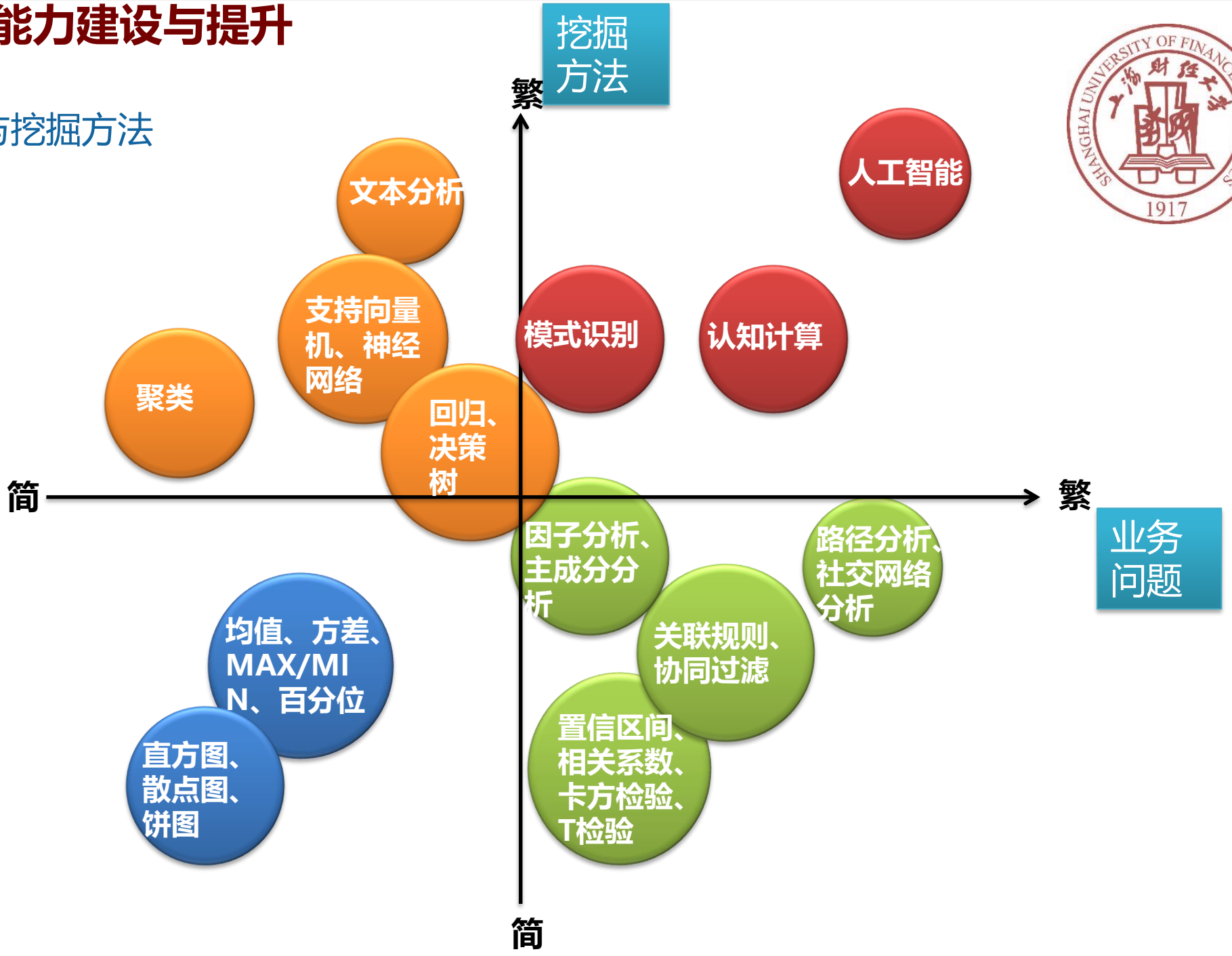
数据分析能力建设与提升

分析能力的十个等级



数据分析能力建设与提升

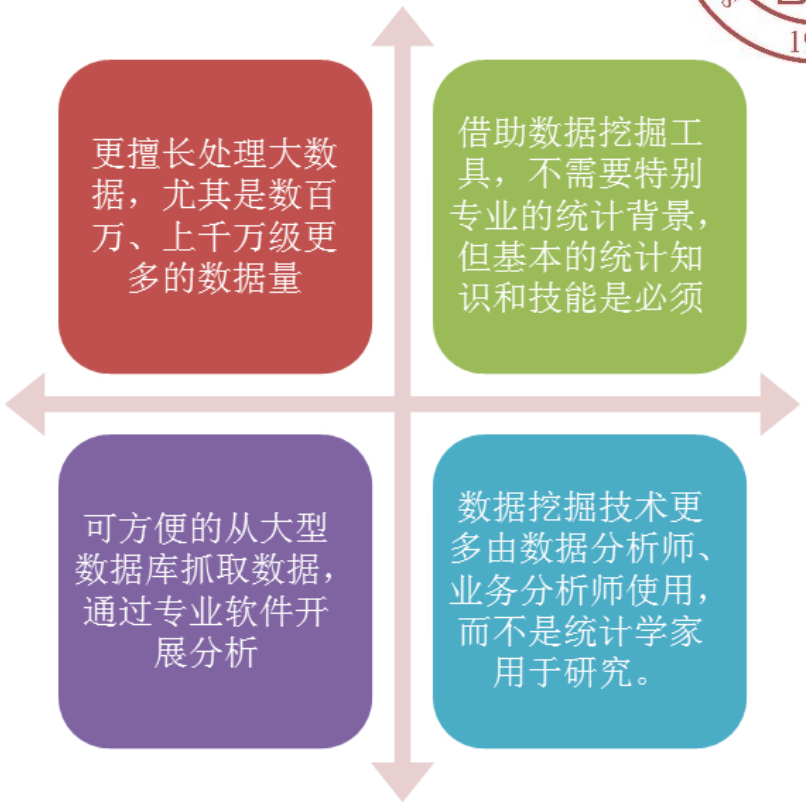
业务问题与挖掘方法



数据分析能力建设与提升

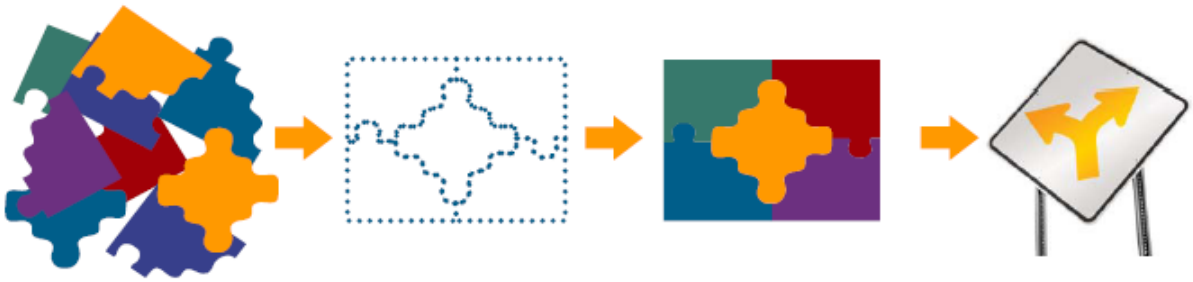
统计分析 vs 数据挖掘

统计分析	数据挖掘
基础是概率论，要对数据分布和变量间关系做假设，再用数据分析技术来验证。	不需要对数据的内在关系做过多的建设或判断，有挖掘工具中的算法自动寻找数据中隐藏的关系或规律。
预测中的应用常表现为一个或一组函数关系式	数据挖掘更注重预测的结果，有时并不会产生明确的函数关系式
“不管白猫还是黑猫，抓住老鼠才是好猫”	



数据分析能力建设与提升

数据挖掘



数据　　知识　　信息　　决策和行动

分析数据
建立模型
分析建模人员

模型
已发现的知识

把模型用在实际数据上
预测未来的行为
业务用户

发现:
找出隐藏在数据背后的模式，
这些模式能把数据转化为知识

部署:
应用已发现的知识达成实用的目的 –
例如：预测

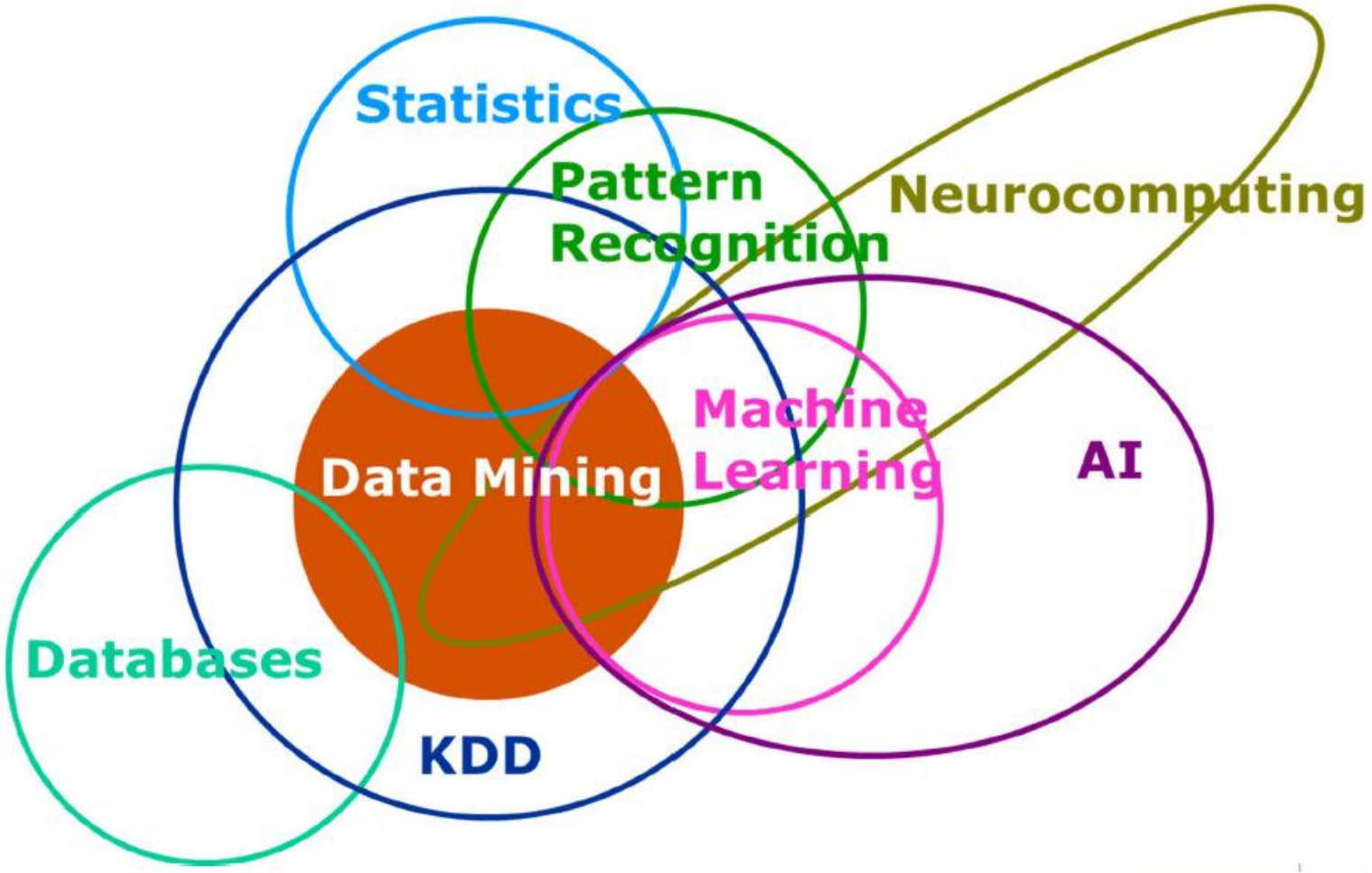
Data Mining is a process of discovering and interpreting patterns in data to solve problems

数据挖掘是一个发现和解释数据中的模式，并用于解决问题的过程

数据分析能力建设与提升

数据挖掘

数据挖掘融合了数据库、人工智能、机器学习、统计学、高性能计算、模式识别、神经网络、数据可视化、信息检索和空间数据分析等多个领域的和技术，是21世纪初期对人类产生重大影响的十大新兴技术之一。



数据分析能力建设与提升

数据挖掘技术介绍

基础学科：

数理统计学	• 线性回归 • Logistic回归
机器学习	• 决策树 • SVM
生物学	• 神经网络 • 遗传算法 • 群体智能

应用模式：

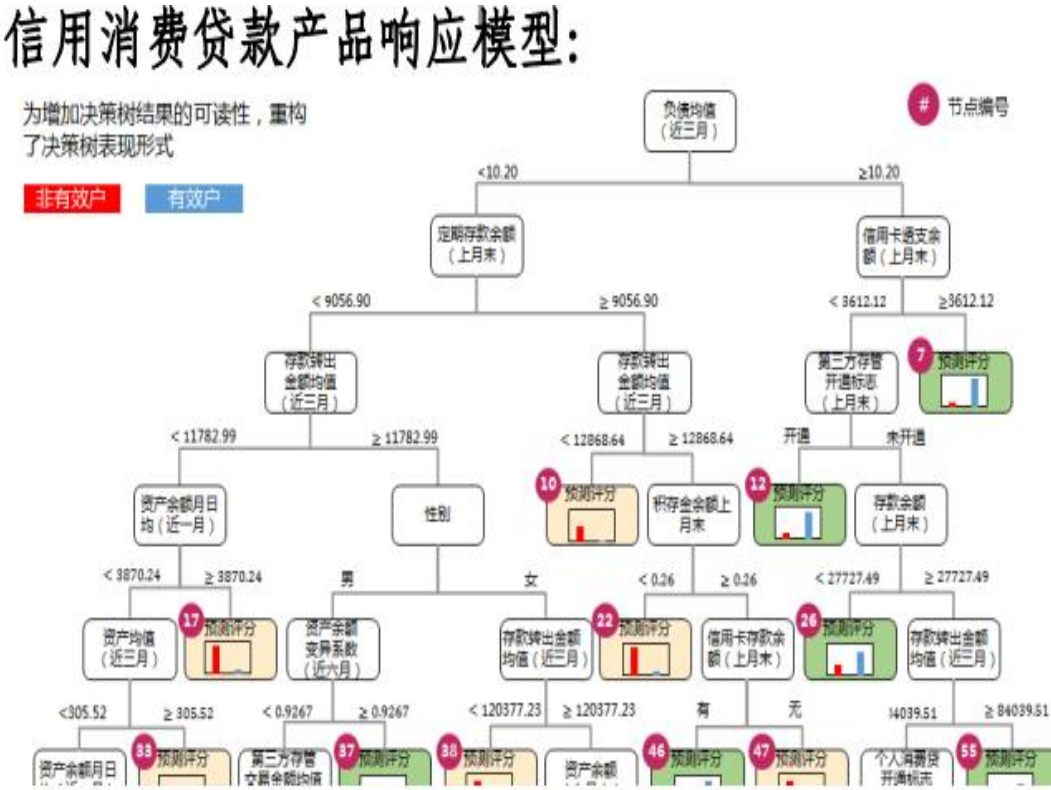
预测分析	决策树、Logistic回归、神经网络、判别分析、SVM、生存分析
描述分析	聚类、关联分析、因子分析、主成分分析、协同过滤、*序列规则（路径分析）*、社交网络分析

数据挖掘技术

决策树

•决策树是一种非常成熟的、普遍采用的数据挖掘技术。之所以称之为树，是因为其建模过程类似于一棵树的成长过程，从根部开始不断分叉最终形成一颗树状结构。其中所有样本数据集是树根，每个节点代表一个结论

1、模型结果直观，便于理解和部署； 2、搭建和应用速度快； 3、噪声数据有较高的承受能力； 4、可同时应对线性和非线性关系。	1、采用的贪心算法，做出在当前看来最好的选择，为从整体最优考虑； 2、目标变量适合分类变量； 3、如某些自变量类别较多时，注意。

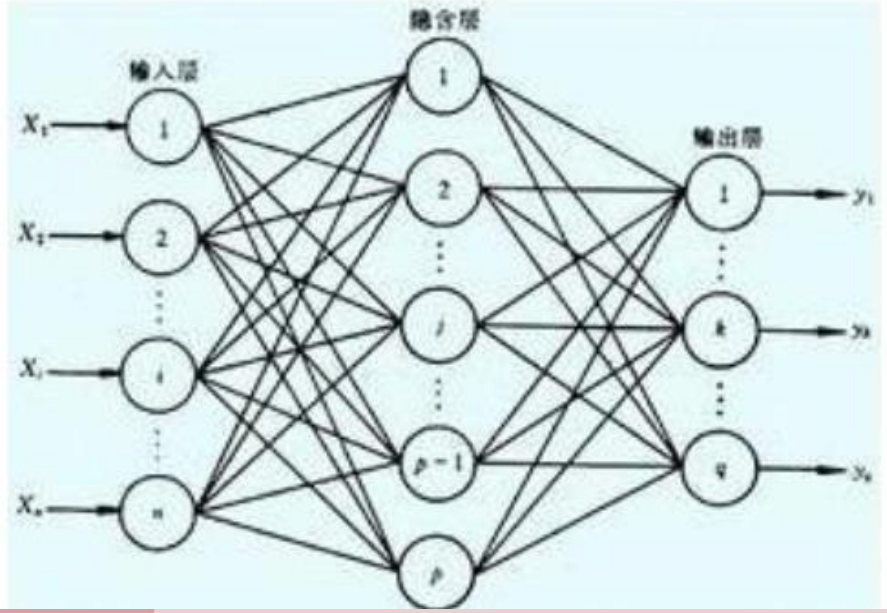


数据挖掘技术

神经网络

•神经网络是通过数学算法来模仿人脑思维，是数据挖掘中机器学习的典型代表。该模型由大量并行分布的人工神经元（微处理单元）组成的，该模型通过对多个非线性模型以及不同模型之间的加权互联，再将最终结果经过转换函数的转换输出为预测值。一个典型神经网络由输入层、隐含层和输出层三部分构成。作为分类、预测问题的重要技术，广泛应用于用户划分、行为预测和营销响应建模中

1、良好的自我学习能力； 2、适合处理非线性关系； 3、噪声数据有较高的承受能力。	1、“黑盒”，结论难以解释，信贷、医疗诊断等领域谨慎使用； 2、容易发生过拟合； 3、模型训练时间长。

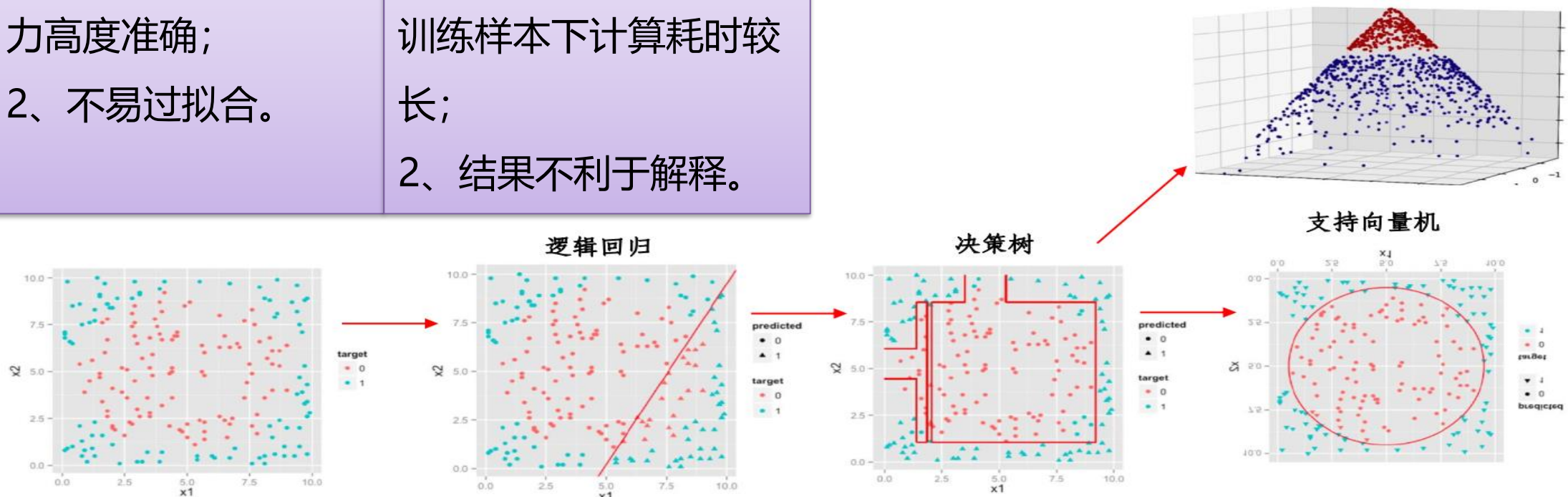


数据挖掘技术

➢ 支持向量机（SVM）

- 支持向量机（Support Vector Machine）是Vapnik等人于1995年率先提出的，是近年来机器学习研究的一个重大成果，其作为一种分类算法，将数据映射到较高的维上，寻找一个最优分类超平面。

1、对于复杂的非线性决策边界的建模能力高度准确； 2、不易过拟合。	1、需要一定规模的训练样本，但其在大规模训练样本下计算耗时较长； 2、结果不利于解释。

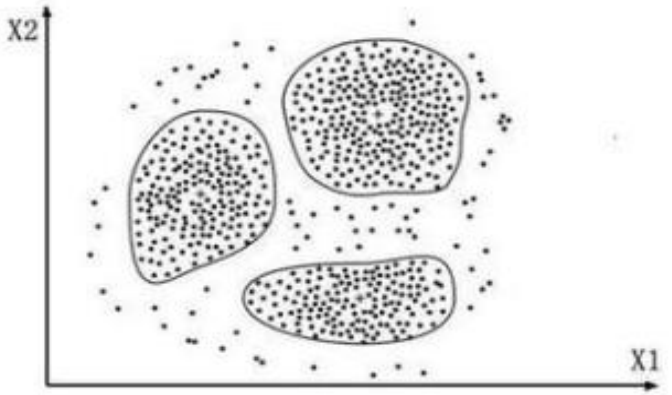


数据挖掘技术

聚类（Clustering）

- 通俗的比喻，“物以类聚，人以群分”。聚类算法用几个特定的业务指标，将观察对象按照相似性和相异性进行不同群组划分。经过划分后，每个群组内各对象相似度会更高，而群组间的相异性更高。其中k-means算法是目前最主流的算法和应用也最广泛

1、技术成熟，算法可靠，非常适合群体细分的工具； 2、结果易于业务理解，逻辑可以解释； 3、简洁、高效，其算法复杂度和数据集大小是线性相关的。	1、 k-means中的k事先未知； 2、对数据噪声和异常值敏感。



数据挖掘技术

关联规则

- 关联规则的主要目的是要找出数据集中的频繁模式，即多次重复出现的模式和并发关系。应用关联规则最经典的案列就是购物篮分析，通过分析客户购物篮中常见商品组合，挖掘客户的购物习惯。

婴儿尿不湿→啤酒[支持度=10%，置信度=70%]

在所有顾客中，有　　的顾客同时购买了婴儿尿不湿和啤酒，而在所有购买了婴儿尿不湿的顾客中，占70%的人同时还购买了啤酒。

1、广泛用于商品推荐模型； 2、结果清晰可用； 3、算法简单，方便部署。	1、需要频繁扫描交易数据集，对大型数据集来说，效率较低； 2、不能推断出因果关系。

数据挖掘技术

协同过滤（Collaborating filtering）

•协同过滤技术是迄今为止最成功的电商推荐系统技术，通过收集群体用户的偏好信息，预测个体用户可能感兴趣的内容。其基于如下基本假设：如果一个人A在一个问题上和另一个人B持相同观点，那么对于另外一个问题，比起随机选择一个路人C，A更有可能同B持有相同观点。

基于用户的协同过滤

- 首先根据用户的历史行为信息寻找与新用户相似的其他用户
- 再根据这些相似

基于商品的协同过滤

- 预先计算好商品的相似度矩阵。以

用户	商品1	商品2	商品3	商品4
用户A	4	?	3	5
用户B	?	5	4	?
用户C	5	4	2	?
用户D	2	4	?	3
用户E	3	4	5	?

$$s(i,j)=\frac{(i\cap j)}{freq(i)\cdot freq(j)}$$

$$s(C,A)=\frac{(5-3.667)(4-4)+(2-3.667)(3-4)}{\sqrt{(5-3.667)^2+(2-3.667)^2}\times\sqrt{(4-4)^2+(3-4)^2}}=0.781$$

$$p_{C,4}=3.667+\frac{0.781\times(5-4)+(-0.515)\times(3-3)}{0.781+0.515}=4.269$$

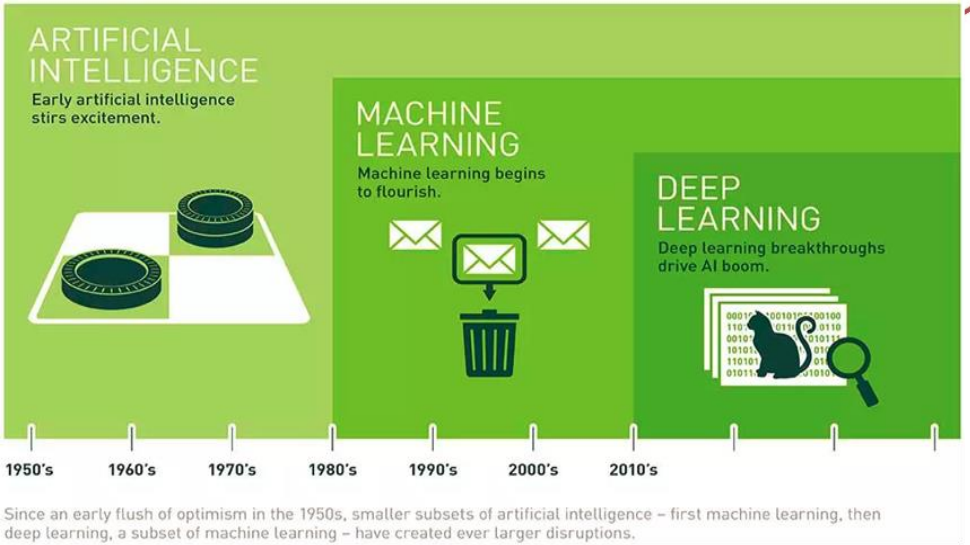
数据分析能力建设与提升

人工智能

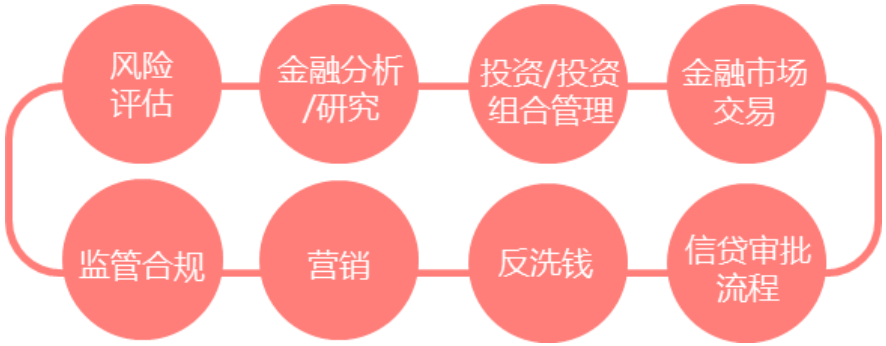
人工智能——人工智能就是要让机器的行为看起来像人所表现出的智能行为。（人工智能概念提出者约翰·麦卡锡）

机器学习——是实现人工智能的一种算法，是从数据中自动分析获得规律，并利用规律对未知数据进行预测的算法。（摘录于中文维基百科）

深度学习——是一种机器学习的模型，使用包含复杂结构或由多重非线性变换构成的多个处理层的神经网络对数据进行预测。（摘录于中文维基百科）



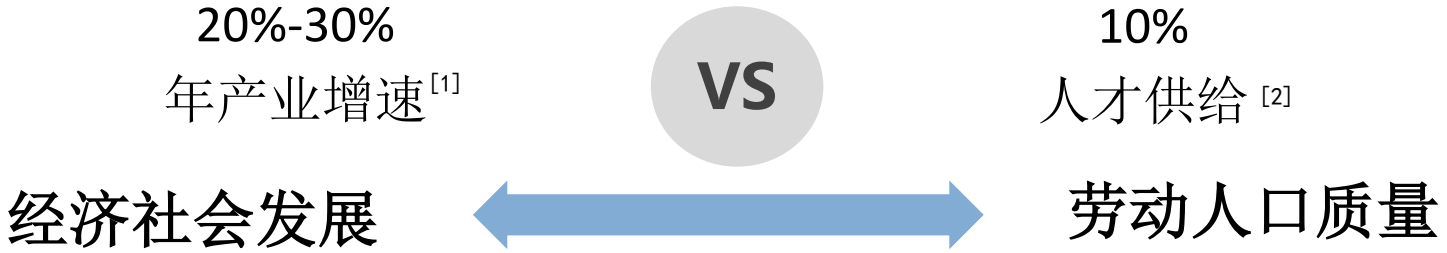
认知计算——一种能够规模化学习、有目的推理，并与人类自然交互的系统。是近年来由一家国际IT服务商提出的概念，其对人工智能在商业应用中理解是实现”认知“而非”智能“，认为当前人工智能要做的是“认知”数据的价值。



案例分享

基于简历大数据的数字经济人才发展与培养分析

- **数字经济**已经成为经济增长的助推器
- **数字经济人才**:具有一定数字能力的数字经济从业者

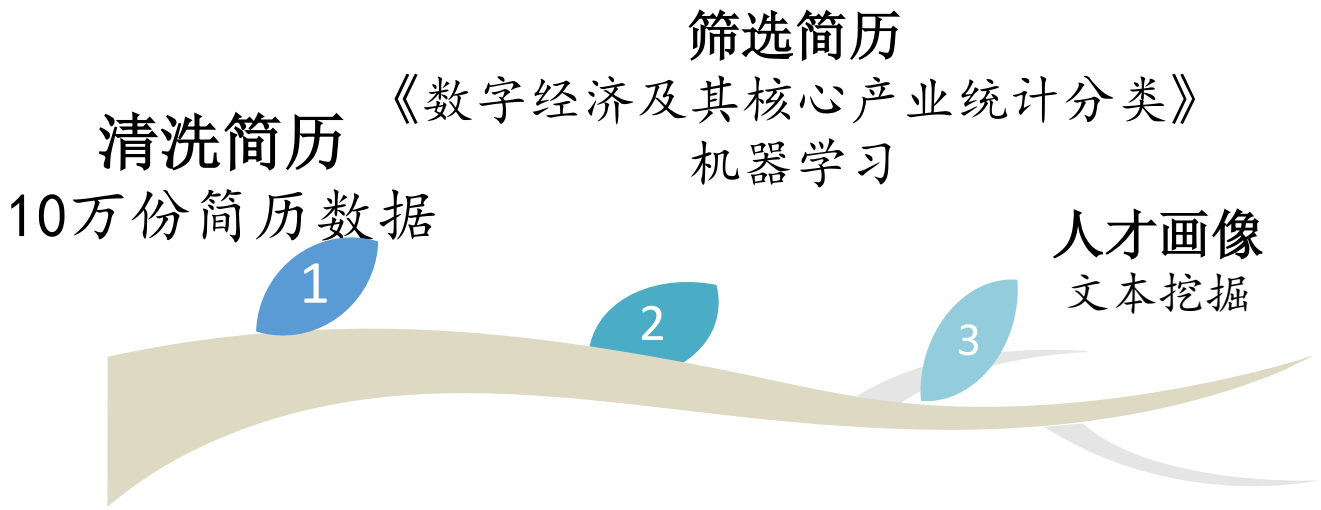


兼具专业素养和数字能力的复合型人才是企业转型的核心竞争力。

[1]中国电子信息产业发展研究院.《2021中国数字经济发展形势报告》(2021年).
[2]中国信通院.《2020-2021年数字化新就业新职业新岗位研究报告》(2021年).

案例分享

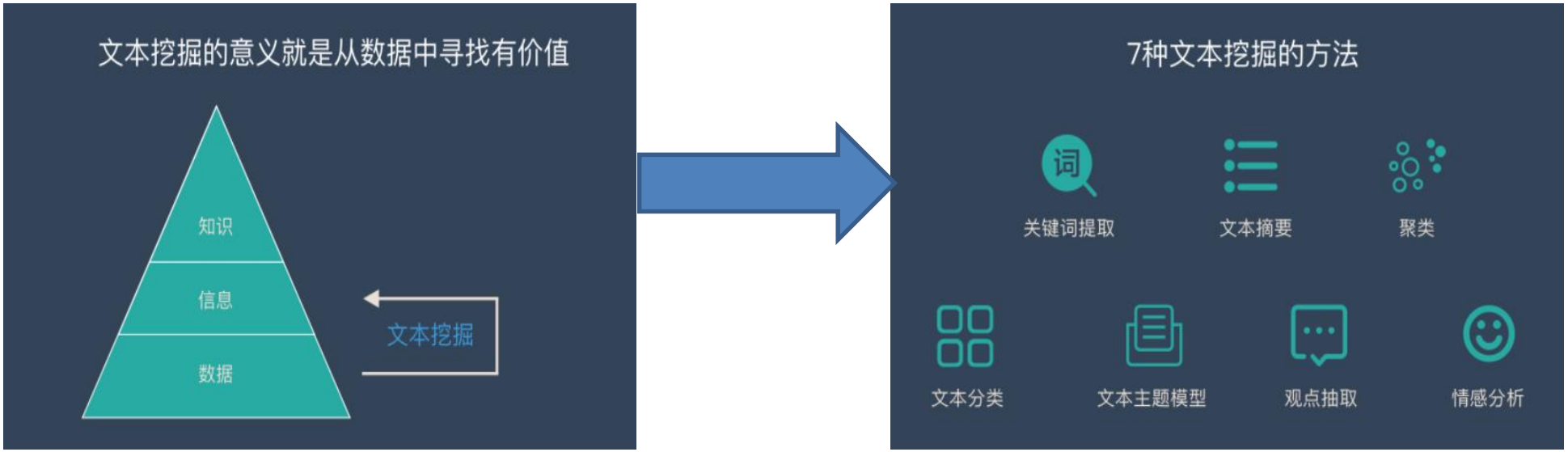
基于简历大数据的数字经济人才发展与培养分析



黄枫指导，上海财经大学2021年优秀大创作品 《数字经济人才发展与培养——基于简历大数据的分析》

案例分享

基于文本挖掘的人才画像



案例分享

基于文本挖掘的人才画像

- 互联网技术的发展为人才招聘市场带来了海量的数字简历。
- 传统的人工识别、 筛选、分类在如此庞大的数据量面前显得力不从心。
- 关键词分类是常用的机器文本分类方法，对于某些特定的领域分类效果较好。
- 所涉及的数字经济涉及面广，与各个行业均有交叉，因此很难构造出特定的关键词实现对数字经济人才的有效提取，导致单纯的关键词方法效果不佳。
- 引入深度学习模型进行简历筛选。最终采用 BERT 预训练+精调模型进行分类，在测试数据集上达到了 90%的分类准确率。

黄枫指导，上海财经大学2021年优秀大创作品 《数字经济人才发展与培养——基于简历大数据的分析》

案例分享

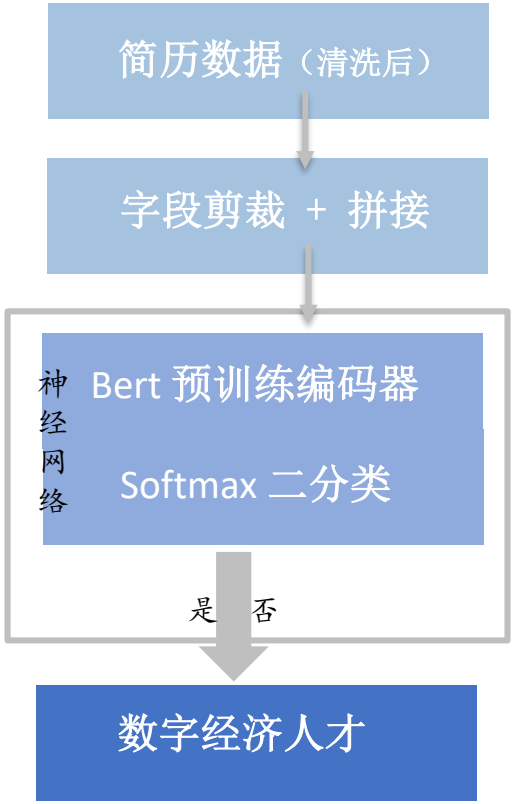
基于简历文本挖掘的人才画像

- 用户画像研究常见于图书情报学、计算机科学等领域，通过已有数据对用户行为、偏好进行分析，在用户行为预测、个性化推送、精准营销等场景中多有应用。
- 依托自然语言处理技术，深入挖掘十万份简历里数字经济人才的文本信息，构造数字能力这一特征变量（下设编程水平、数字意识等多个次级指标），基于教育经历、工作经历、项目经历，绘制人才发展路径，连接人才成长网络。

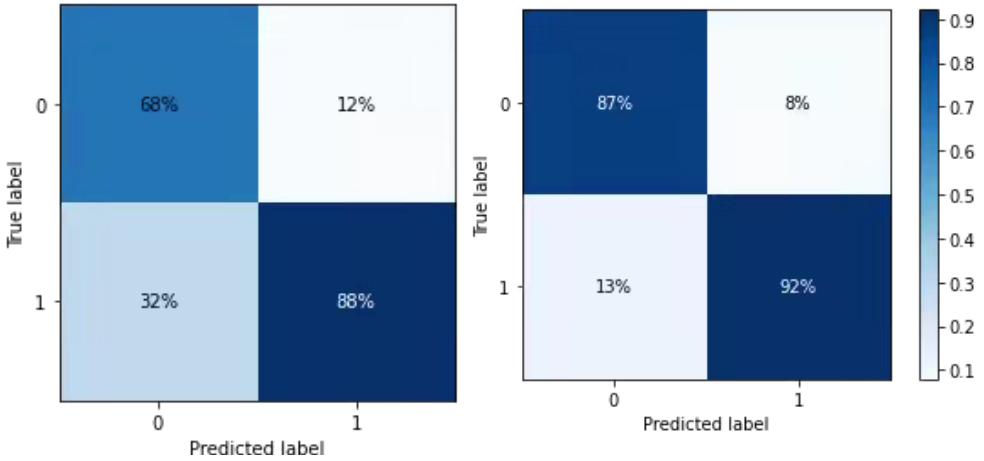
黄枫指导，上海财经大学2021年优秀大创作品 《数字经济人才发展与培养——基于简历大数据的分析》

案例分享

基于简历文本挖掘的人才画像



关键词 ⟷ 深度学习



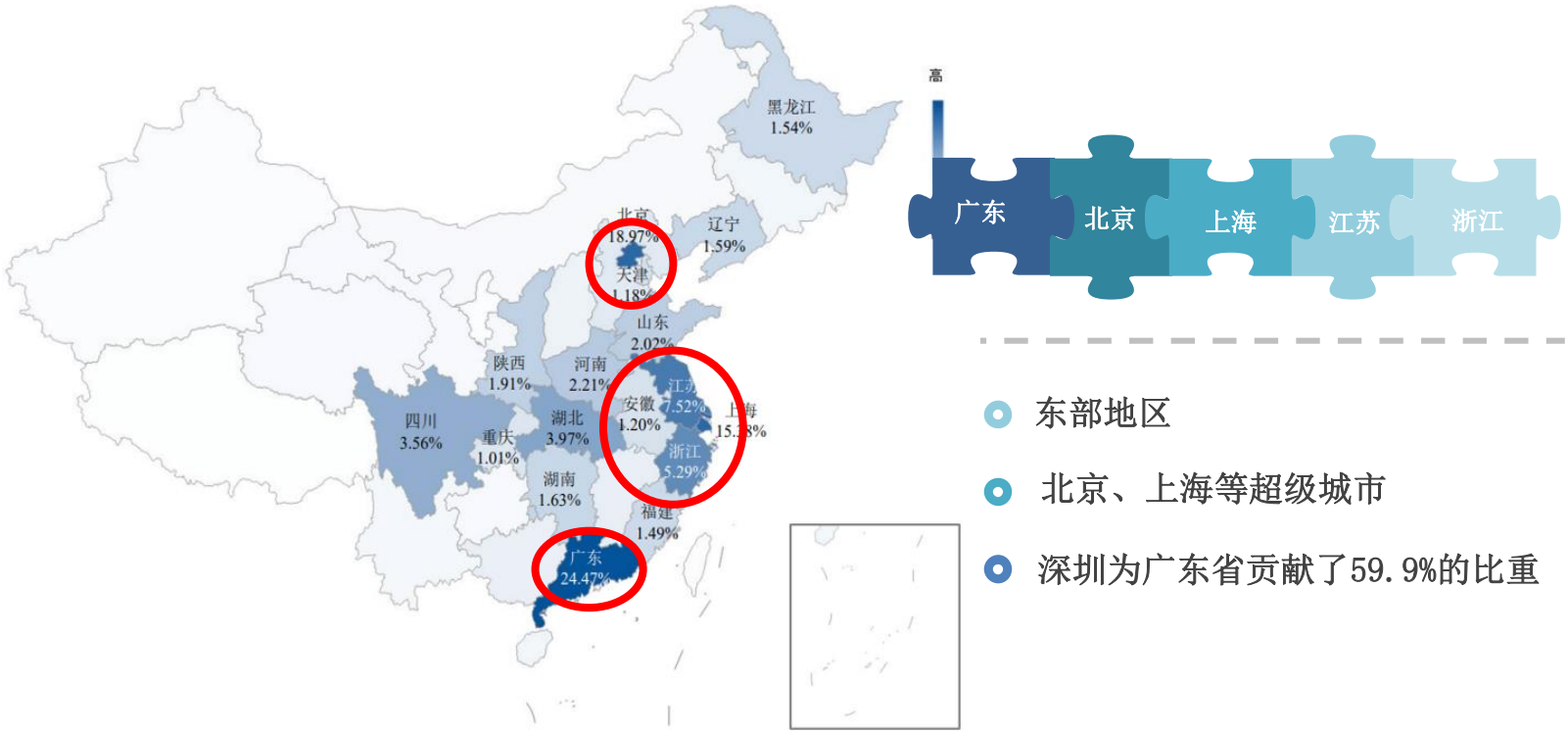
	精确率	召回率	f 1
关键词分类	0.881	0.726	0.795
深度学习分类	0.921	0.91	0.915

精确率 = 预测为正类且正确的数量/预测为正类的数量
召回率 = 预测为正类且正确的数量/样本中正类的数量
F1分数为精确率和召回率的调和平均数，用来反映模型的综合性能

案例分享

结果展示

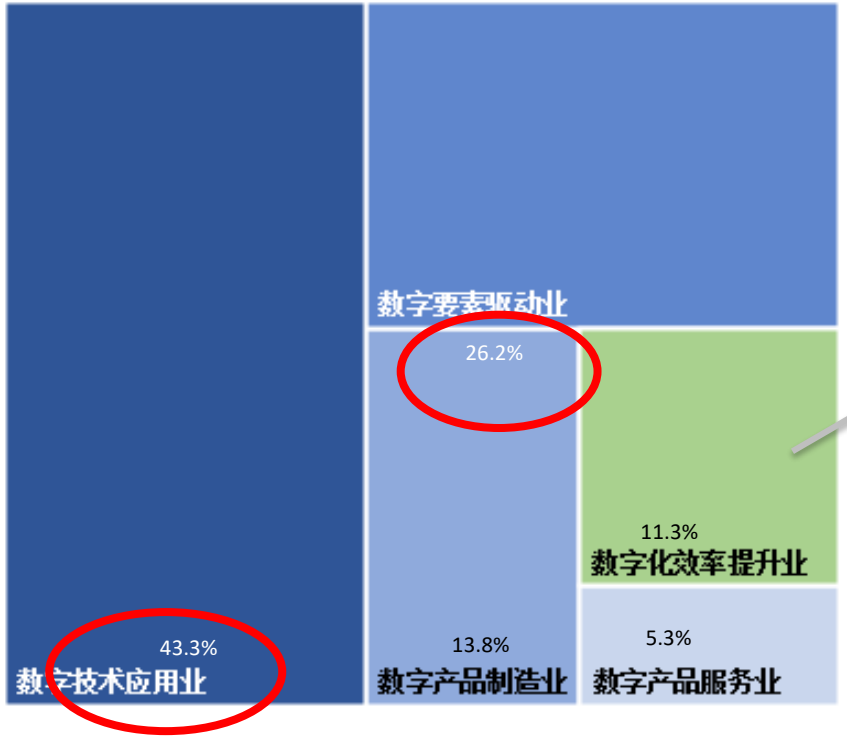
数字经济人才公司地域分布

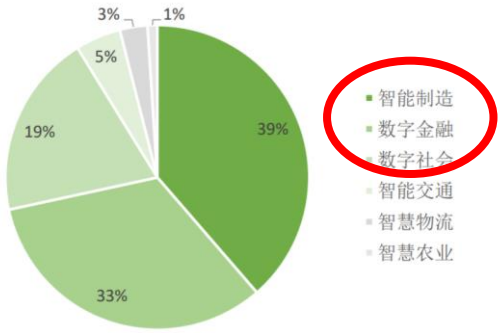


案例分享

结果展示

数字经济人才行业大类分布





数字技术应用业、数字要素驱动业占比近**3/4**。

产业数字化人才只占**1/10**。

黄枫指导，上海财经大学2021年优秀大创作品 《数字经济人才发展与培养——基于简历大数据的分析》

数据分析项目方法论

常见数据挖掘项目类型分类

阿里分类

- 目标客户的特征分析
- 目标客户预测（响应、分类）
- 运营群体活跃度定义
- 用户路径分析
- 交叉销售模型
- 信息质量模型
- 服务保障模型
- 客户评价模型
- 信用分析模型
- 商品推荐模型
- 决策支持

SAS分类

- 市场营销
- 风险管理
- 政府行政管理
- WEB网站运营
- 物流行业
- 其他

相关性、重要性和影响力分析，在当今比以往任何时候都重要，因为可以收集到的数据越来越多，发现隐藏在数据背后的知识具有战略价值，各种类型的分析应用将层出不穷、蓬勃发展，很难穷尽描述所有的分析应用类型。

千举万变，其道一也。

—《荀子·儒效》

市场营销	风险管理	政府行政管理	Web网站运营	物流行业	其他
营销活动响应分析建模	客户信用风险评分	避税分析	网站分析	需求预测	文本分析
净提升度分析建模	市场风险评分建模	社保欺诈侦测	社交媒体分析	供应链分析	业务流程分析
客户挽留分析	运营风险评分建模	洗钱分析侦测	访问量测试（A/B测试）		
购物篮分析	欺诈侦测	恐怖主义侦测			
自动推荐系统					
客户细分					

数据分析项目方法论

项目分类之目标客户特征分析

目标客户特征分析几乎是现代企业数据化运营实践中最普遍、频率最高的业务分析需求之一，原因在于数据化运营的第一步就是找准你的目标客户、目标受众，然后才是相应的运营方案、个性化的产品与服务等。

- **寻找目标客户**
  虚拟目标客户特征分析vs真实目标客户特征分析
- **寻找运营的抓手**
- **用户群体细分的依据**
- **新品开发的线索和依据**

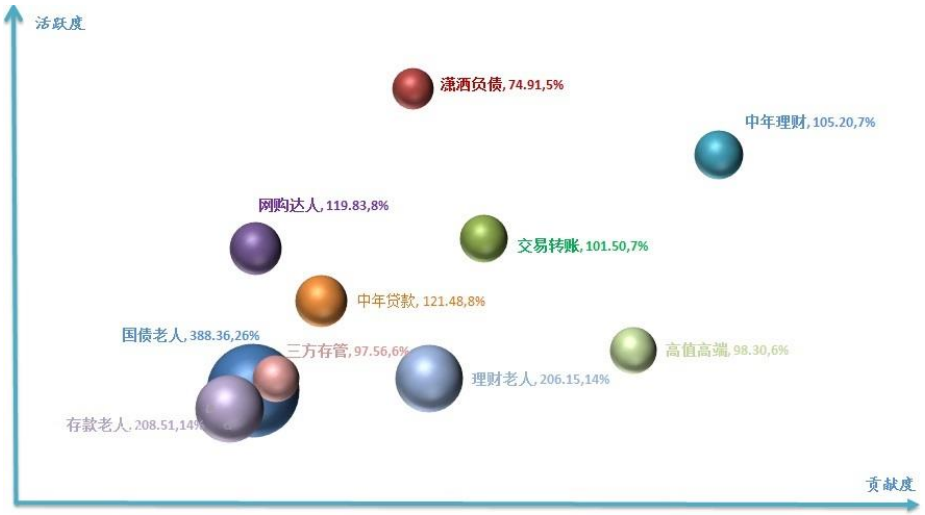
分析技术：
- ✓RFM
- ✓聚类
- ✓决策树
- ✓假设检验
- ✓Excel透视表

1、基于预先定义的划分

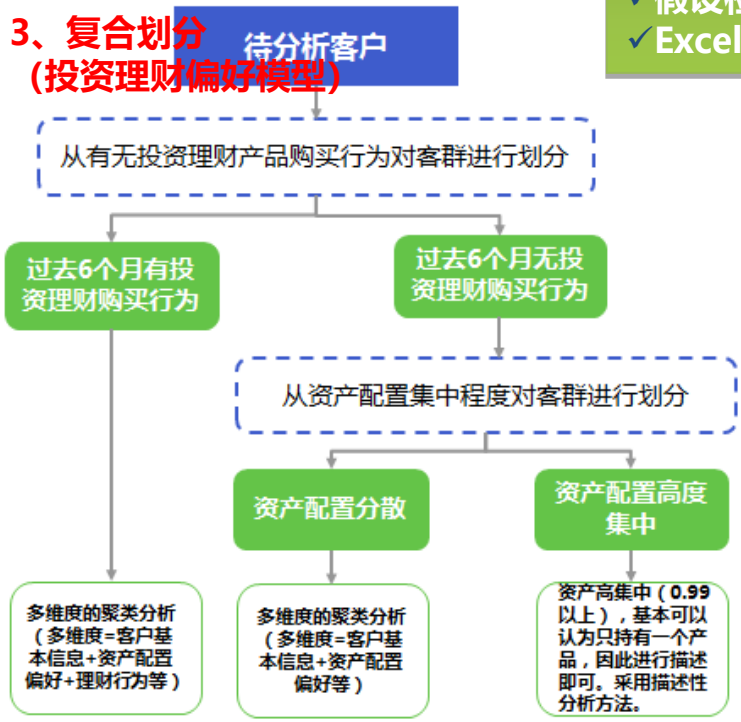
表1：全行个人客户五级分类结果

五级分类	2014[2]		2013		变化	
	客户数量（户）	占比	客户数量（户）	占比	客户数量	占比
A类	248,248	0.05%	502,800	0.11%	-50.63% ↓	-0.06% ↓
B类	12,324,828	2.65%	11,086,637	2.53%	11.17% ↑	0.12% ↑
C类	84,284,461	18.13%	79,168,558	18.09%	6.46% ↑	0.03% ↑
D类	176,606,341	37.98%	186,513,627	42.63%	-5.31% ↓	-4.64% ↓
E类	191,519,382	41.19%	160,288,585	36.63%	19.48% ↑	4.56% ↑
合计	464,983,260	100.00%	437,560,207	100.00%	6.27% ↑	-

2、基于数据分析的划分（中高端客户行为细分）



3、复合划分（投资理财偏好模型）



数据分析项目方法论

项目分类之目标客户的预测（响应、分类）模型

预测模型是数据挖掘中最常用的一种模型类型，几乎成了数据挖掘技术应用的一个主要代名词。这里的预测模型包括产品响应模型、流失预警模型等。这类模型的核心就是响应概率（Probability）。

个人信用消费贷款产品响应模型

1 客户人口统计学特征	具体包括性别、年龄、婚姻状态、行业、职业等人口统计学特征
2 客户银行关系属性	包括客户行龄、是否为员工、开户分行、是否为财富客户等
3 客户价值属性	包括星级、各业务的贡献、成本等
4 客户资产负债属性	例如客户资产、负债规模、波动、资产负债比、客户资金能力等
5 客户产品特征	包括相关客户产品持有持有情况
6 客户行为特征	包括存款、贷款、信用卡、中间业务等业务的客户交易行为特征
7 客户营销活动数据	客户参加营销活动及反馈的历史数据

模型名称	验证集:误分类率	训练集:误分类率	最终变量数	决策树叶节点数	综合评价
决策树-3-6	20.12%	19.61%	22	141	该模型在验证集上拥有最低的误分类率，但决策树节点相对较多
决策树-3-5	20.61%	20.37%	15	69	分类效果较好，采用变量数较少，但决策树节点较多
决策树_信用卡	20.68%	19.35%	31	359	该模型依据“是否持有信用卡”先行粗分，再自动运行，为达到同等分类效果，采用了过多变量和节点
神经网络	20.87%	21.08%	15	-	模型效果不错，但缺乏业务解释性
决策树-2-6 ✓	20.95%	20.53%	15	20	模型效果不错，且用了最少的变量和节点，推荐采纳
回归	21.27%	21.42%	15	-	较之其他5个模型，分类效果相对较弱

分析**技术**：
- ✓逻辑回归
- ✓决策树
- ✓神经网络
- ✓支持向量机

提取客户在成为有效户前六个月观察期内的多维度数据，构成预测变量

观察期：共6个月

观察期**期末时点**数据，例如资产余额、某产品开通情况、行龄 — 期末

例如最近一月的存款余额变动净值 — 近1月

例如近三月资产均值、近三月最大资产余额 — 近3月

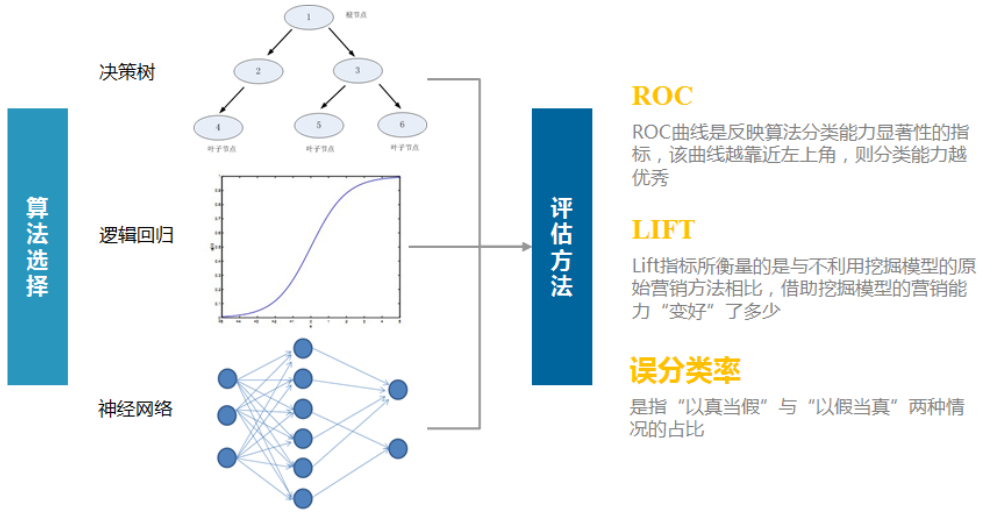
近6月

例如近六月各月末资产变异系数、近六月各月末资产最大值与最小值差额

提取有效户初次贷款行为所在月份

表现期：1个月

- 本模型认为客户前六个月观察期内的综合表现，将“酝酿”并最终决定其之后一个月表现期内的贷款行为



数据分析项目方法论

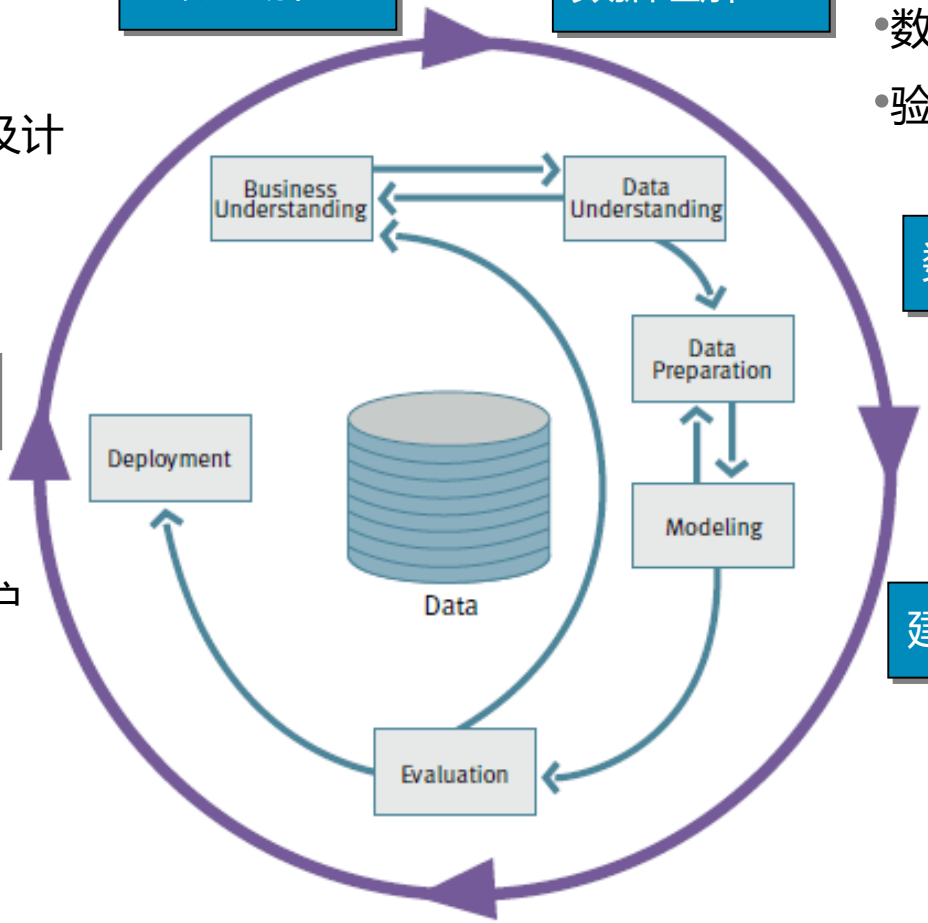
数据挖掘项目经典流程—CRISP-DM(IBM)

业务理解

- 确定业务目标
- 可行性研究
- 制定建模思路及计划

数据理解

- 收集原始数据
- 数据探索
- 验证数据质量



数据准备

- 选择数据
- 清洗数据
- 加工数据

建模

- 选择建模技术
- 构建模型
- 验证模型

评估

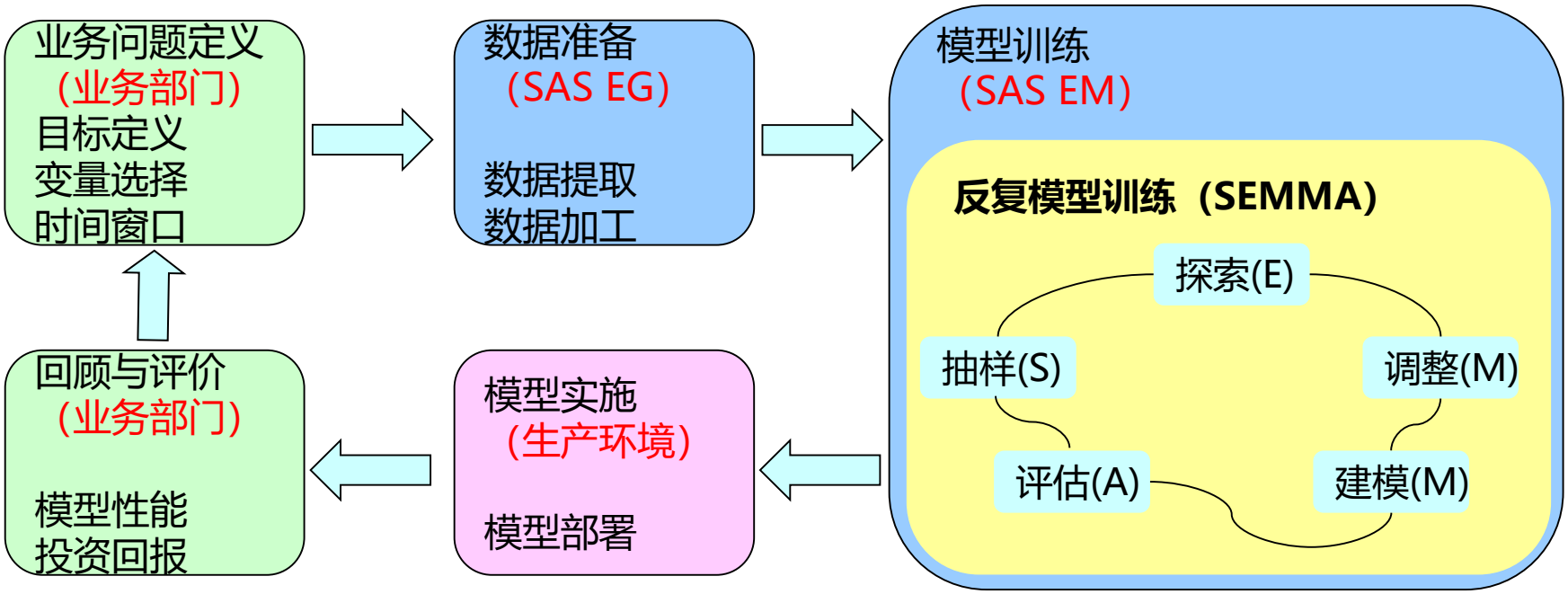
- 业务结果评估
- 建模过程评估
- 确定后续行动

部署

- 制定部署计划
- 模型监控和维护
- 模型资产总结

数据分析项目方法论

数据挖掘项目经典流程—SEMMA(SAS)



研究目标		数据分析方法
看现状（分布）	→	描述统计
探关系	→	描述统计、相关性分析
找影响因素	→	显著性检验、模型
做预测	→	模型

华东师范大学李艳副教授，《市场调查与分析大赛参赛准备与技巧》（2021年）

看现状（分布）：以图形完整展示分布，以统计量探索分布特征

图形描述

图形	功能	适用数据类型 名义	有序	定量
柱状图	(1)展示定性变量分布 (2)分类展示指标值	√	√	√
饼图	展示构成结构	√	√	√
环形图	比较多个总体结构	√	√	√
直方图	展示定量变量分布			√
折线图	展示定量变量分布			√
箱线图	展示定量变量分布			√

华东师范大学李艳副教授，《市场调查与分析大赛参赛准备与技巧》（2021年）

图形描述

图形	功能	适用数据类型 名义	有序	定量
线图	展示指标值随时间的变化			√
散点图	展示定量变量间关系			√
气泡图	展示变量间关系			√
雷达图	展示多个指标值（及其关系）			√

华东师范大学李艳副教授，《市场调查与分析大赛参赛准备与技巧》（2021年）

集中趋势描述统计量

统计量	定义/计算	适用数据类型 名义	有序	定量
众数	出现次数最多的值	√	√	√
算术平均$\bar{X}$	$\frac{1}{n}\sum_{i=1}^{n}X_i$			√
截尾平均	去掉两边极端值后的算术平均			√
中位数	$\frac{n+1}{2}$位置的值		√	√
下四分位数Q_L	$\frac{n+1}{4}$位置的值		√	√
上四分位数Q_U	$\frac{3(n+1)}{4}$位置的值		√	√

华东师范大学李艳副教授，《市场调查与分析大赛参赛准备与技巧》（2021年）

离中趋势描述统计量

统计量	定义/计算	适用数据类型 名义	有序	定量
异众比率	非众数数据所占比例	√	√	√
极差	最大值减最小值			√
四分位差IQR	$Q_U - Q_L$			√
标准差S	$\sqrt{\frac{1}{n-1}\Sigma_{i=1}^{n}(X_i - \bar{X})^2}$			√
变异系数	$S/\bar{X}$			√

华东师范大学李艳副教授，《市场调查与分析大赛参赛准备与技巧》（2021年）

看关系：对两个变量进行相关性分析

变量类型

X1	X2	两个变量相关性分析方法
定性	定性	列联表检验（交叉表） 对应分析
定性	定量	Kruskal-Wallis检验 方差分析（ANOVA） 多重比较
定量	定量	Pearson线性相关系数 Spearman等级相关系数 Kendall τ相关系数

} 也可用于有序数据

华东师范大学李艳副教授，《市场调查与分析大赛参赛准备与技巧》（2021年）

模型

1、降维/信息提取/排序： 因子分析、主成分分析等

2、分析影响机制/做预测： 回归分析、机器学习方法等

3、分析影响路径与多变量间作用关系： 贝叶斯网络（含结构方程模型）等

Q：方法越炫酷越好吗？

A：No！方法恰当且简单有效是最高境界！高级分析方法往往要求更为复杂的数据！

Q：除了统计学方法，可以用其他分析方法吗？

A：Yes！只要是有利于深入分析的方法都可以使用。

华东师范大学李艳副教授，《市场调查与分析大赛参赛准备与技巧》（2021年）

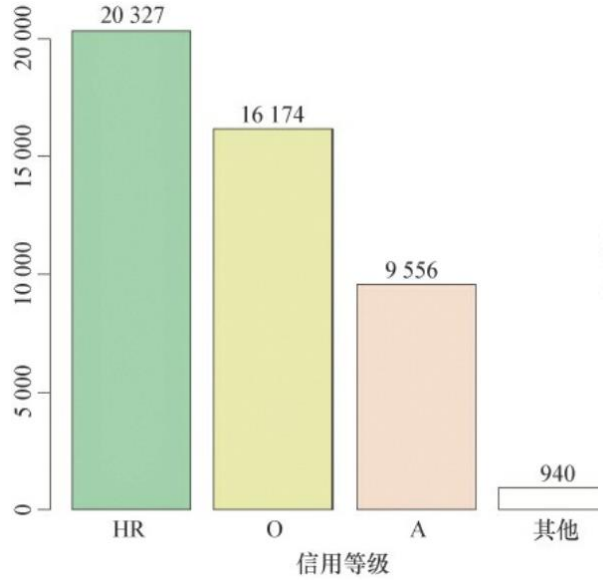
描述统计——数据可视化

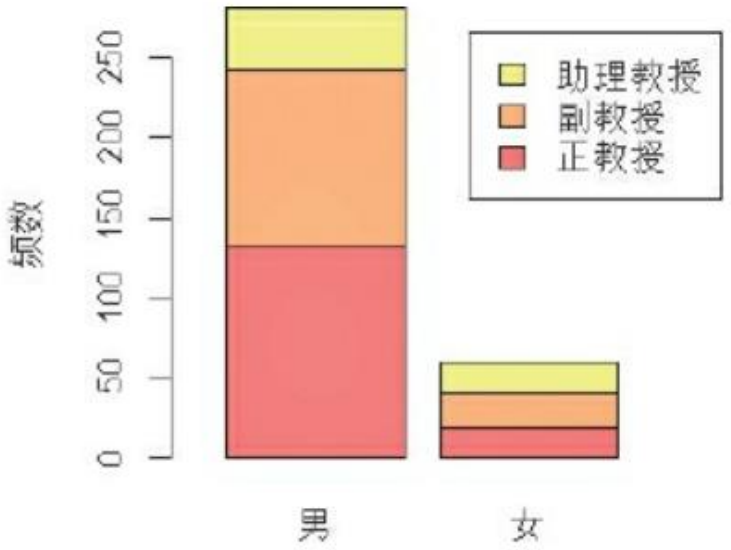
柱状图

- 柱状图是针对分类数据所作的统计图。 每根柱子代表一个类别，柱子的高度是这个类别的频数，有时也是百分比。

堆积柱状图

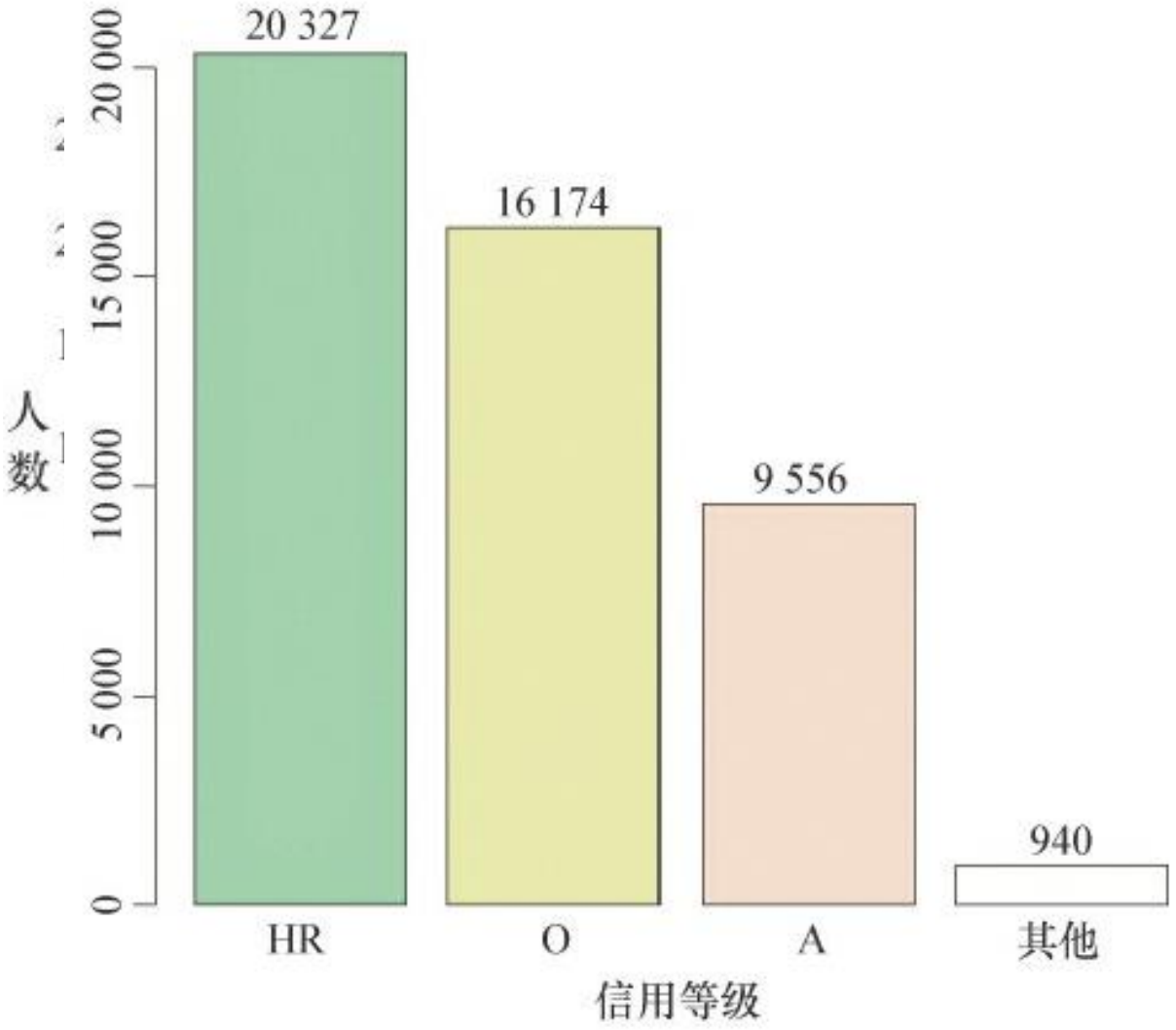
- 堆积柱状图和柱状图的本质一样，都是在展示频数。 只不涉及两个离散型分类变量。





描述统计——数据可视化

柱状图之错误使用



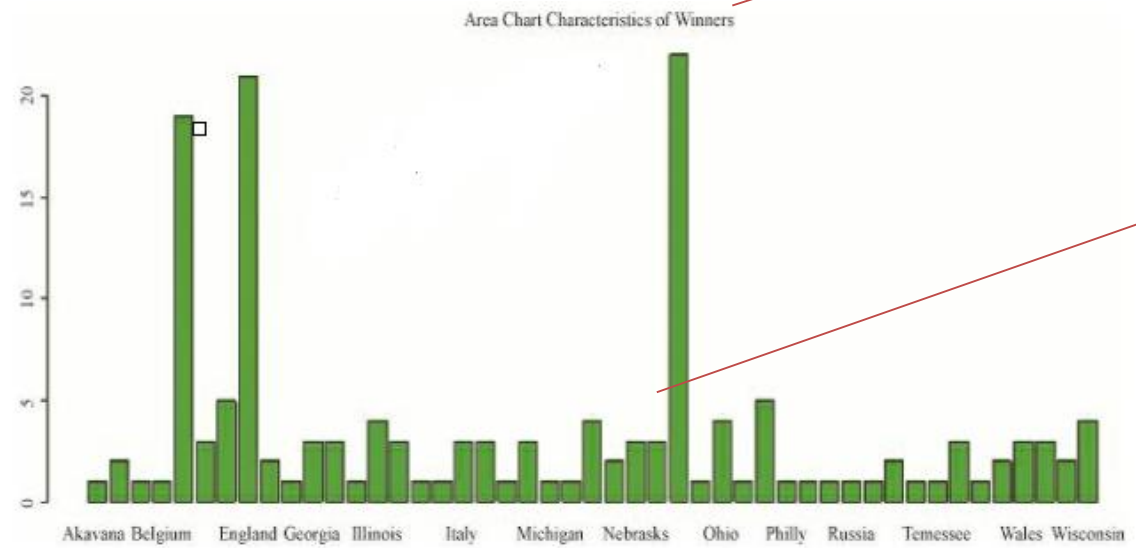
点评1：这不是在画统计图， 而是在画诗， 这幅图画的是《题西林壁》中的“远近高低各不同”。 最高的柱子高2万多， 最矮的柱子才60。

点评2：美观问题。 人都说距离产生美， 柱子之间需要留出空隙，让人喘口气。 横坐标“信用等级”也体现了自己无处安放的青春， 非要跟频数60挤在一起才有安全感吗?

点评3：是图的标题。 这个图的大名叫“柱状图”， 你却起个绰号叫“分布图”。?

描述统计——数据可视化

柱状图之错误使用



获奖者地区分布频数图

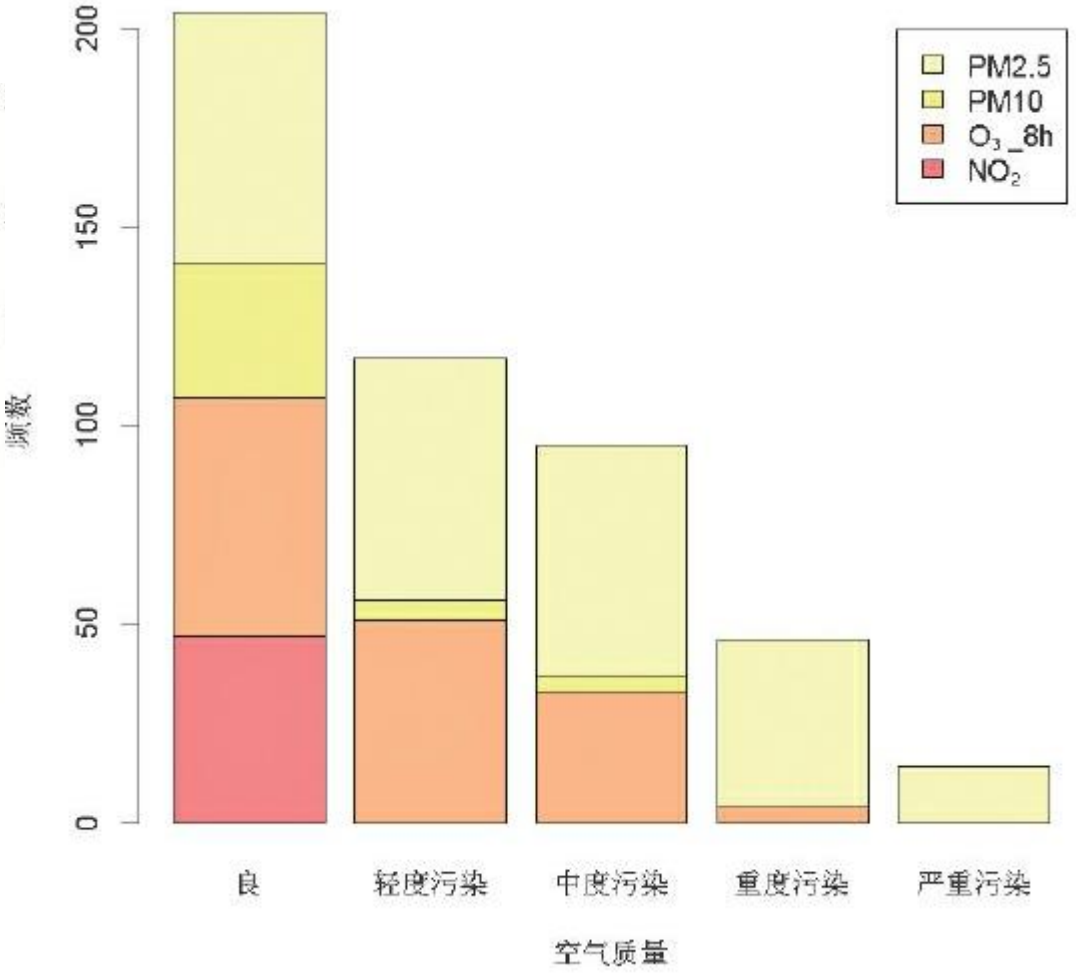
点评2：图的标题出现了两次，这是分析报告里经常看到的。图的上方，标注了一次标题(更多时候是统计软件默认的标题，而作者没有修改或者去掉)，然后图的下方又写了一遍。

点评1：洋洋洒洒几十根柱子，精心排列得奇丑无比。而且由于柱子数太多，很多标签无法显示，根本无法知道每根柱子对应哪个地区，相当于这个柱状图没有传递任何信息！

点评3：图的标题和纵轴标题大名叫“柱状图”，就不要再给起个“频数图”或者“分布图”这种名字了。另外，这个图缺少纵轴标题，可以标注“频数”或者“人数”。

描述统计——数据可视化

堆积柱状图之错误使用

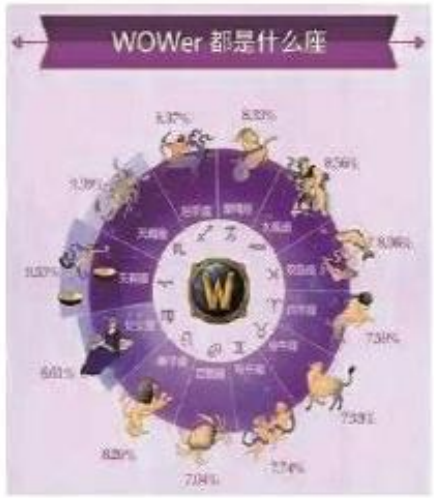


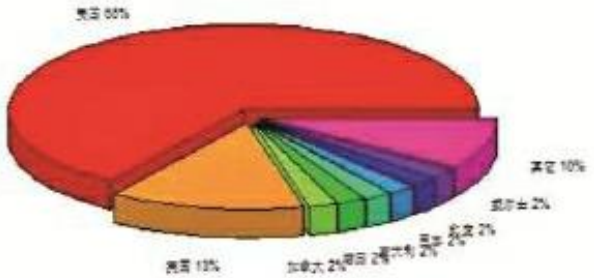
点评2：这些柱子上面最多出现了4种颜色，然而标签却显示出7种物质。看原始数据才发现，CO或者O3频数太低，根本显示不出来。

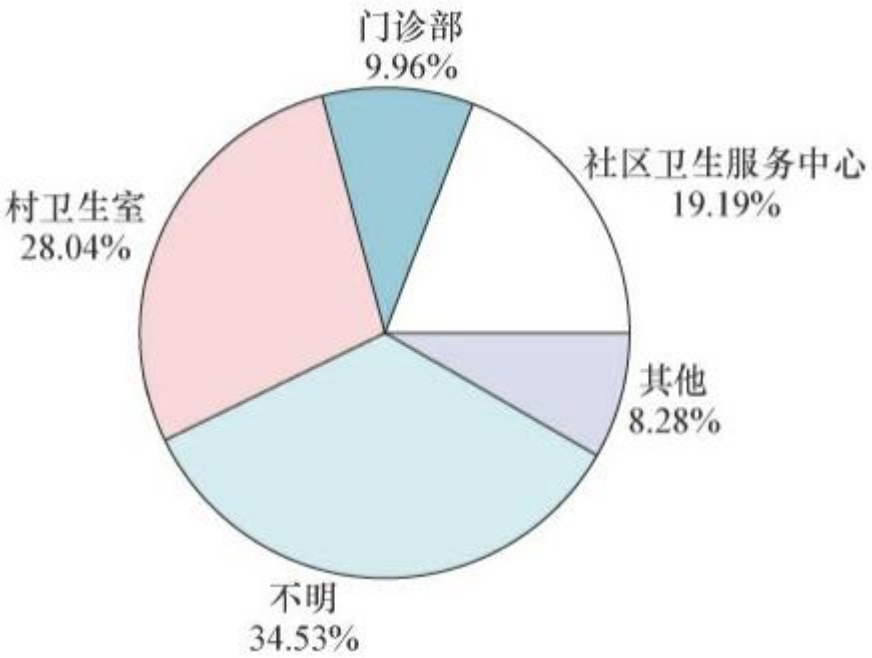
点评1：这是在对读者进行色弱测试吗？很难看出，哪段是PM2.5，哪段是PM10。注意，但凡类别较多，需要画堆积柱状图的时候，应选择区分度比较强的配色，让人能识别出每段柱子都是哪个类别。

描述统计——数据可视化

饼图之错误使用







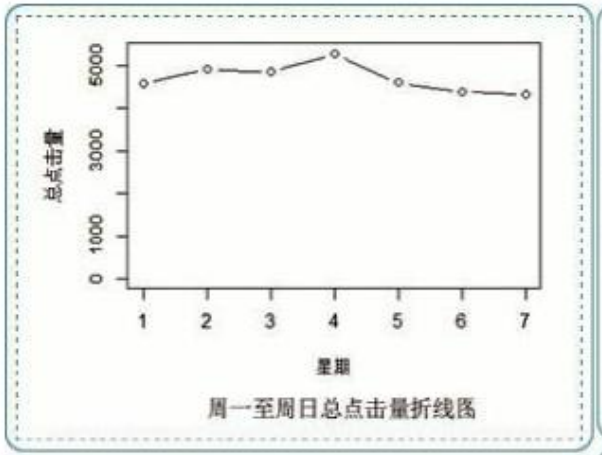
点评1：饼的块数过多(如果只有两类也不适合画饼图)。

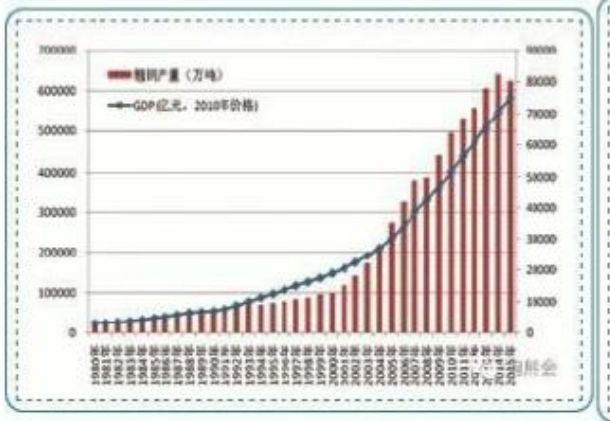
点评2：饼的标签单独打在旁边的时候，对应起来很费劲，比如右下角的饼图： 这个饼分了9块，右侧的 只有8个。另外一个34.53%的 立的标签呢?

：饼的标签，一般只标注百分 很少标注频数或者两者都标注。 角的饼图就同时标注了频数和 化，异常混乱。

描述统计——数据可视化

折线图之错误使用





点评1：左上图：一根线飘在空中，让人不明所以。不妨对纵轴展示范围进行调整。

点评2：右上图：三根折线两个纵轴，让人难以比较

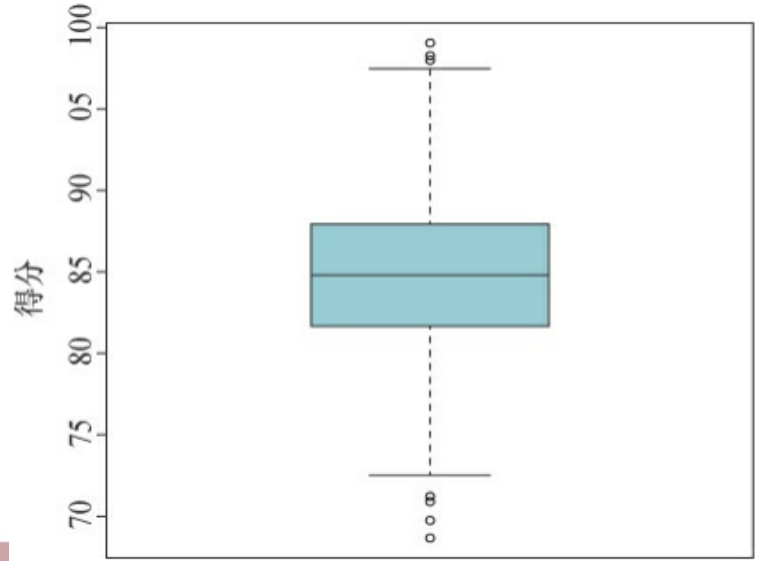
点评3：左下图：少了纵轴标题，横轴标签过于密集。

点评4：右下图：只能用一个词来表达：一团乱麻。 如果有太多的信息想要表达，而且非要在一个图中，就是这个效果

描述统计——数据可视化

箱线图（盒须图）

- 箱线图（boxplot） 是一种针对连续型变量的统计图。
- 箱子的中间一条线是数据的中位数，代表了样本数据的平均水平。
- 箱子的上下限， 分别是数据的上四分位数和下四分位数，意味着箱子包含50%的数据。 箱子的高度在一定程度上反映了数据的波动程度。
- 在箱子的上方和下方， 又各有一条线。 如果有点冒出去， 应理解为“异常值”。



描述统计——数据可视化

茎叶图

- 茎叶图可以同时展示原始数据和分布的形状。
- 图形由“茎”和“叶”两部分组成。 通常以数据的高位数字作为树茎，低位数字作为树叶。

	A	B
1	进球时间	时间段
2	57	下半场
3	65	下半场
4	89	下半场
5	5	上半场
6	10	上半场
7	61	下半场
8	81	下半场
9	73	下半场
10	92	伤停补时
11	41	上半场

部分进球时间

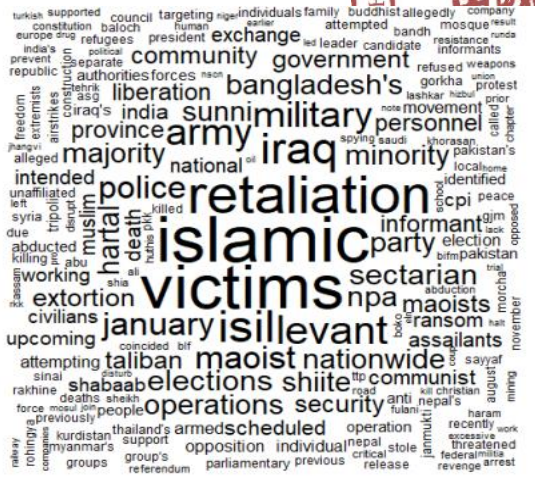
```
0 | 5
1 | 089
2 |
3 | 122477
4 | 125889
5 | 016779
6 | 125
7 | 135
8 | 017789
9 | 00122266
```

甲		乙
97	0	78
6331	1	0579
83	2	13

甲乙两名球员八场比赛得分对比

描述统计——数据可视化

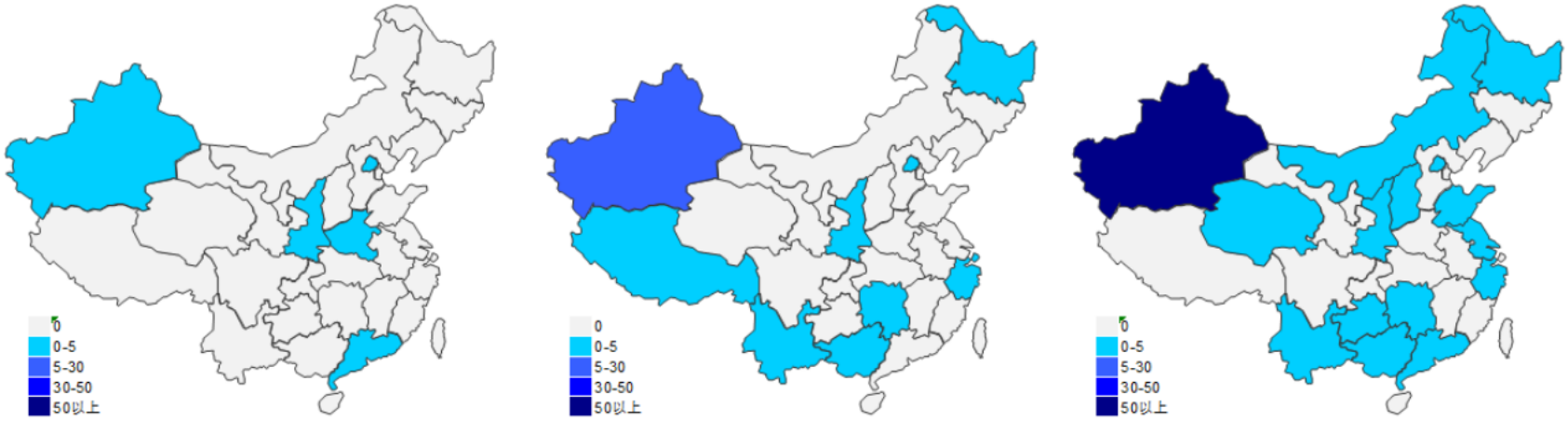
词云图



2015 2016 2017

2015-2017恐怖袭击动机

地理图



中国恐怖袭击事件地域分布图

例：国内高校相关课程及公司技能需求核心词云图对比

图表备注："数据"为国内高校课程和公司技能需求均最高频的常规词，未被纳入词云呈现。

例：全国高校大数据专业分布地图

数据分析可视化工具

- Excel：少量数据，作图不频繁；缺点：图表类型不多，不完美，需要研究小技巧，弥补工具缺陷。
- Tableau：分析大数据，很强用户交互性，商业智能和数据可视化分析领域最强软件，操作简单，图表精美。
- D3、Python、R等编程工具：编程实现个性化，D3专注于数据可视化，Python和R都有可视化包，方便实现各种类型。

数据分类与处理

单变量数据

数据类型	集中趋势与离散趋势测量	数据可视化
分类变量	众数，中位数，分位数，异众比率	柱状图，饼图，频数统计表
连续型变量	中位数，分位数，平均数，极差，方差，偏态，峰态	直方图，折线图，散点图，箱线图，茎叶图

多变量数据

数据类型	数据可视化	统计建模分析
分类变量	堆积柱状图，列联表	列联表分析，逻辑回归
连续型变量	散点图，直方图，折线图	多元回归，时间序列模型
分类变量+连续型变量	箱线图，茎叶图，堆积柱状图	逻辑回归，回归+哑变量

资料收集与文献整理
——避免大段文字，适当加粗突出内容

政策类别	美国	日本	共同点
货币政策	回购政策 **降息**政策，降级基准利率 调整贴现窗口工具，取消准备金要求 **量化宽松** 重启非常规流动性工具（PDCF、CPFF）	扩大现金流供给，提供**无利息贷款**、利率下调 提供更为灵活的信贷手段 启动**量化宽松**政策 扶持其他担保贷款业务	信贷降息
财政政策	2万亿美元**财政刺激**方案 联邦税申报及纳税期限延迟 向以前年度结转营业亏损以获得税收返还 企业替代性**最低税**抵免	15万亿日元**财政刺激**计划 **税金减免**政策	纳税减缓
企业政策	发放其他形式企业**补贴**、奖励 扩大保险范围、力度	发放大量企业**补贴** 开设中小企业咨询台 推延保险金缴纳 提供远程雇员教育 扩大失业保险	补贴补助 社保减缓

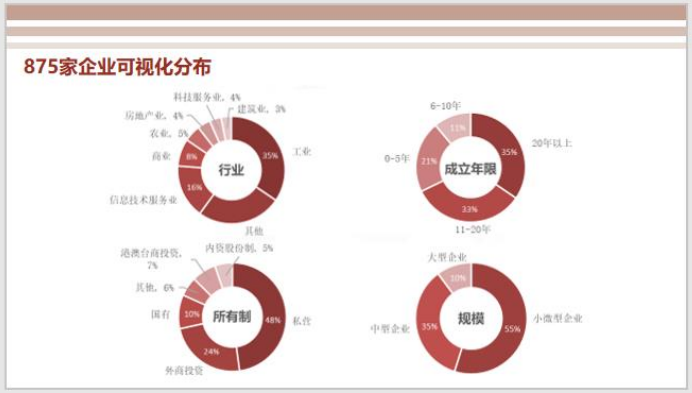
重点国家纾困政策归纳梳理

全国重点城市政策关键词词云图

上海及各区县政策关键词词云图

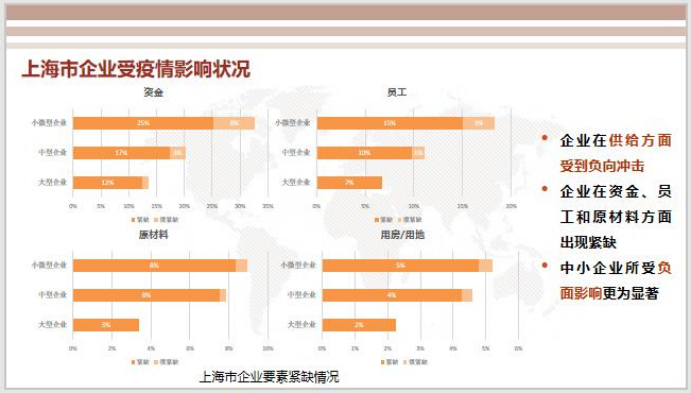
吴纯杰、黄枫指导，2021年“挑战杯”上海市特等奖，国赛一等奖作品《疫情冲击下惠企纾困政策分析与成效调研》

问卷数据可视化——注意配图色系统一，图片内容清晰大方

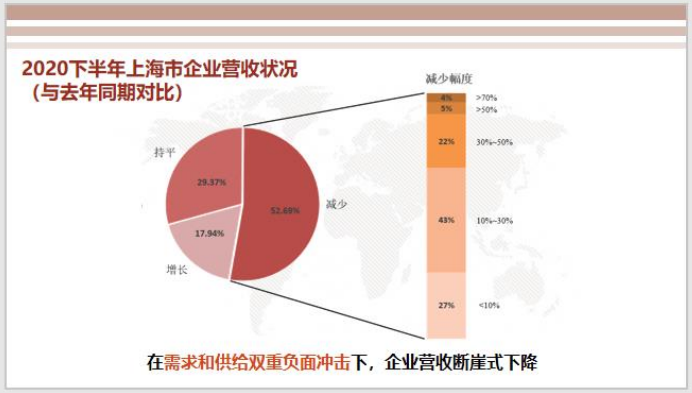


10

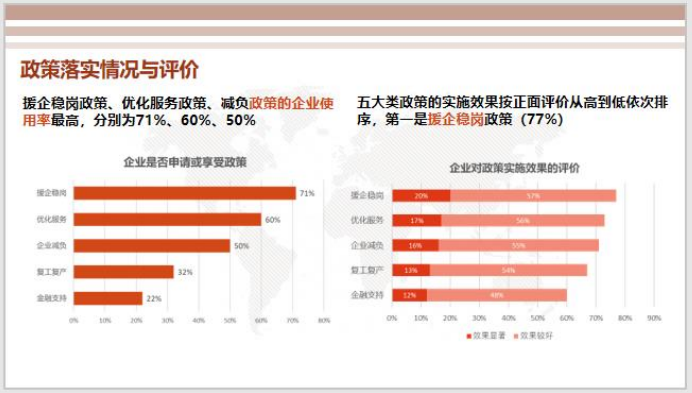
11



12



13



14

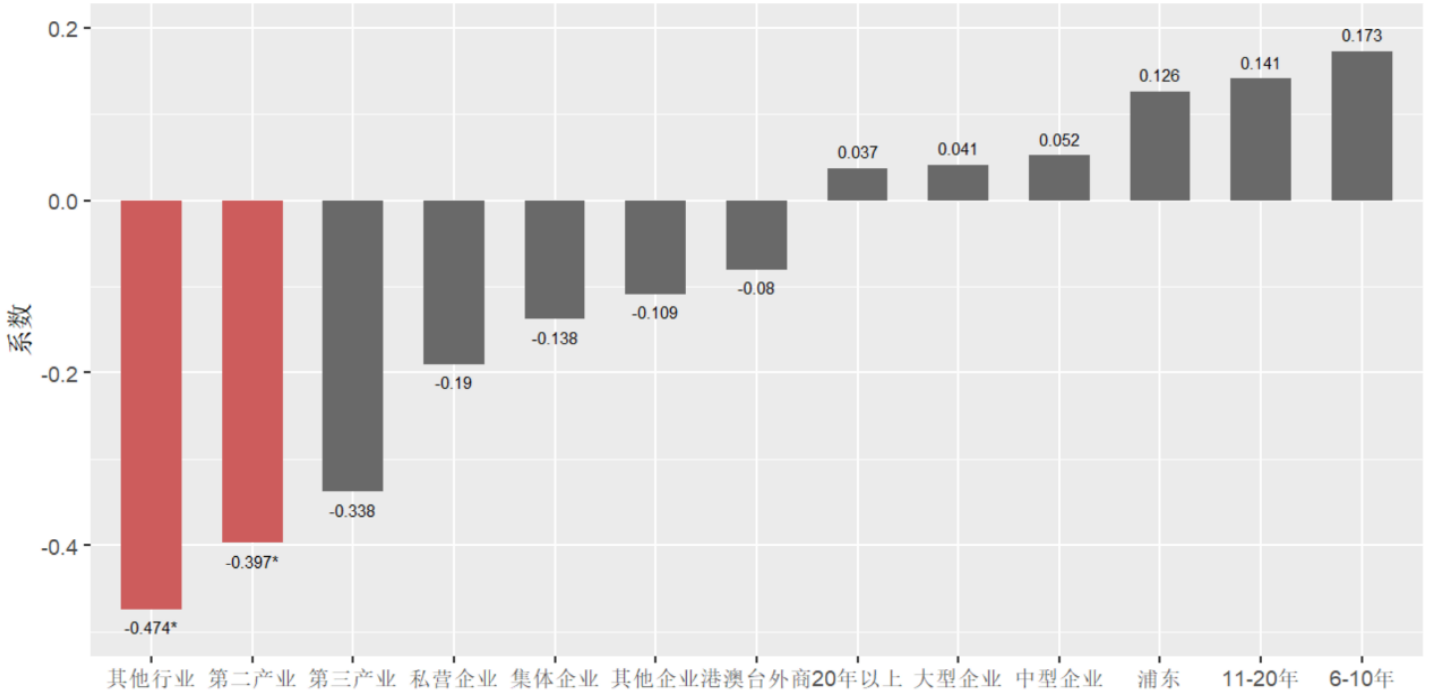
15

吴纯杰、黄枫指导，2021年“挑战杯”上海市特等奖，国赛一等奖作品《疫情冲击下惠企纾困政策分析与成效调研》

定量研究——定序回归下企业纾困政策效果对比

鉴于不同企业类型之间有顺序关系，借鉴逻辑回归的构造和参数估计方法，引入定序回归模型。

选择**非浦东地区、属于第一产业、成立0-5年、小微型、国有的企业**作为**参照水平。**



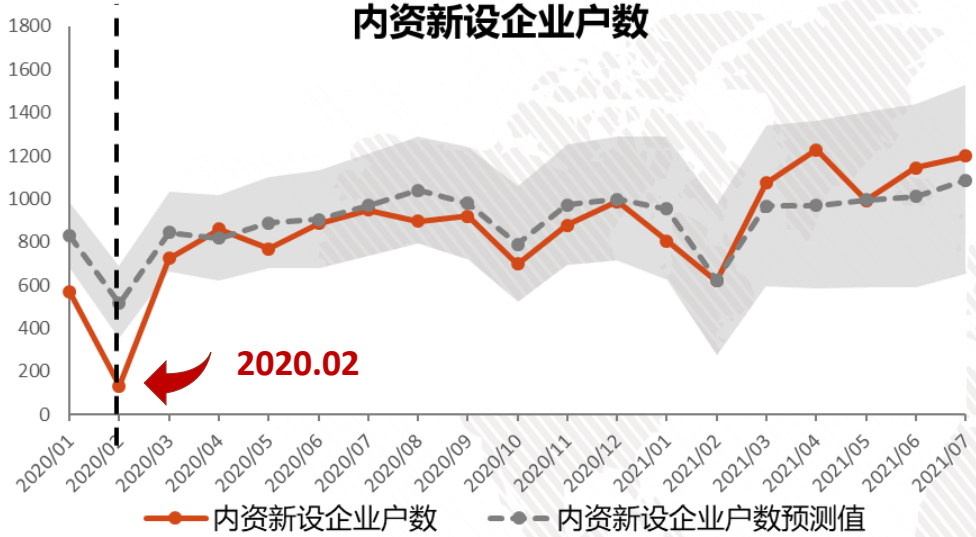
总体评价定序模型系数图

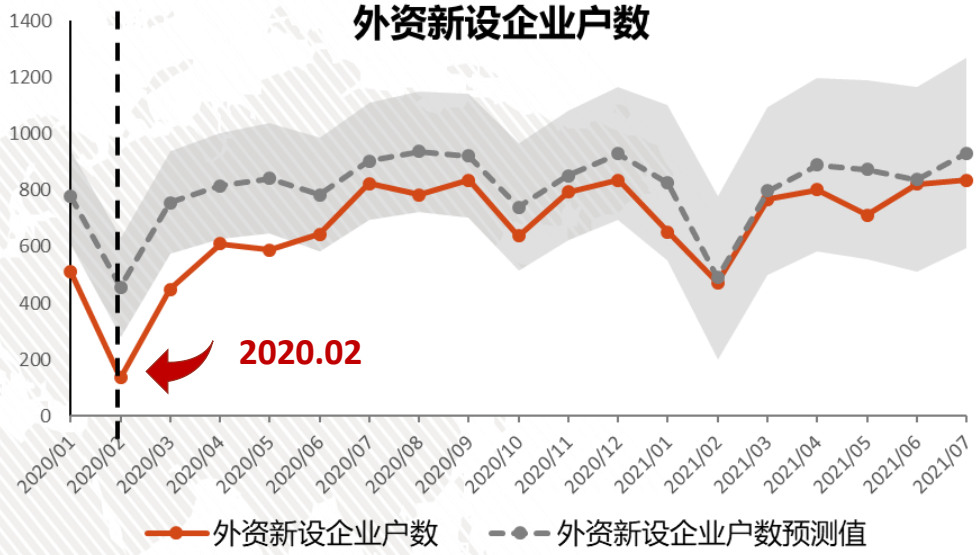
*解释变量如下：

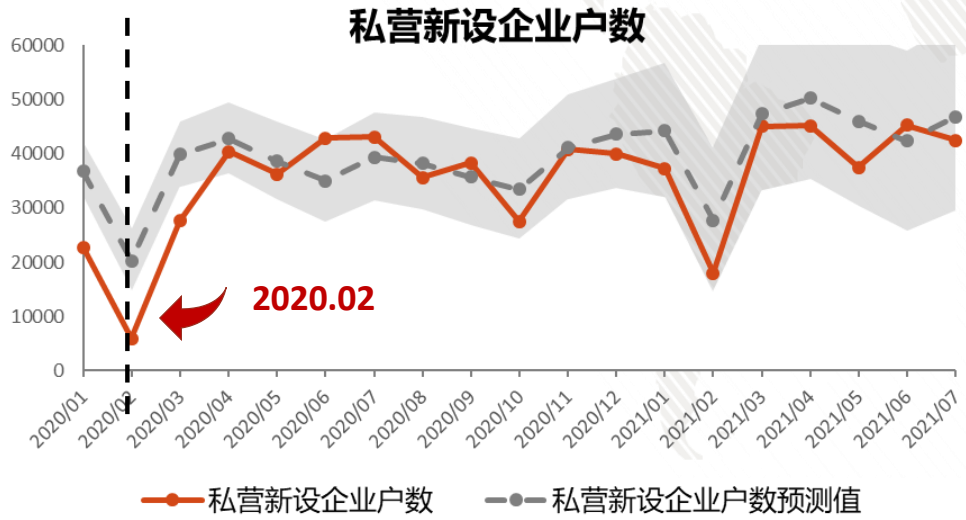
企业所属地区
企业所属行业
企业年限
企业规模
企业所有制

吴纯杰、黄枫指导，2021年“挑战杯”上海市特等奖，国赛一等奖作品《疫情冲击下惠企纾困政策分析与成效调研》

时间序列模型——构造若没有疫情发生的“反事实结果”







- 三类企业2020年2月新设企业实际值远低于预测值
- 随着2月上旬，上海市各区县企业纾困政策密集出台，3月新增企业数量开始回升
- **复苏速度：内资>私营>外资**
- **复苏幅度：私营>内资>外资**

吴纯杰、黄枫指导，2021年“挑战杯”上海市特等奖，国赛一等奖作品《疫情冲击下惠企纾困政策分析与成效调研》

数据分析方法论

条条大路通罗马

天下同归而殊途，一致而百虑。

——《周易·系辞》

- 不同的思路之间是普遍联系的
- 不同的算法之间是普遍联系的
- 数据本身是普遍联系的

PART 三

统计建模工具

统计建模工具使用情况

2.1 工具篇

Python、Excel、MySQL为数据从业者使用最多的数据分析工具

从数据分析结果来看，Python可以说是数据从业者中最受欢迎的编程语言，问卷调研的受访者中，超过 7 成在工作中需要使用 Python。

此外，Excel 和 MySQL的使用比重占到调研用户的半数左右。 Hive 、 Hadoop MapReduce 和 Spark 等数据分析常用工具紧随其后。

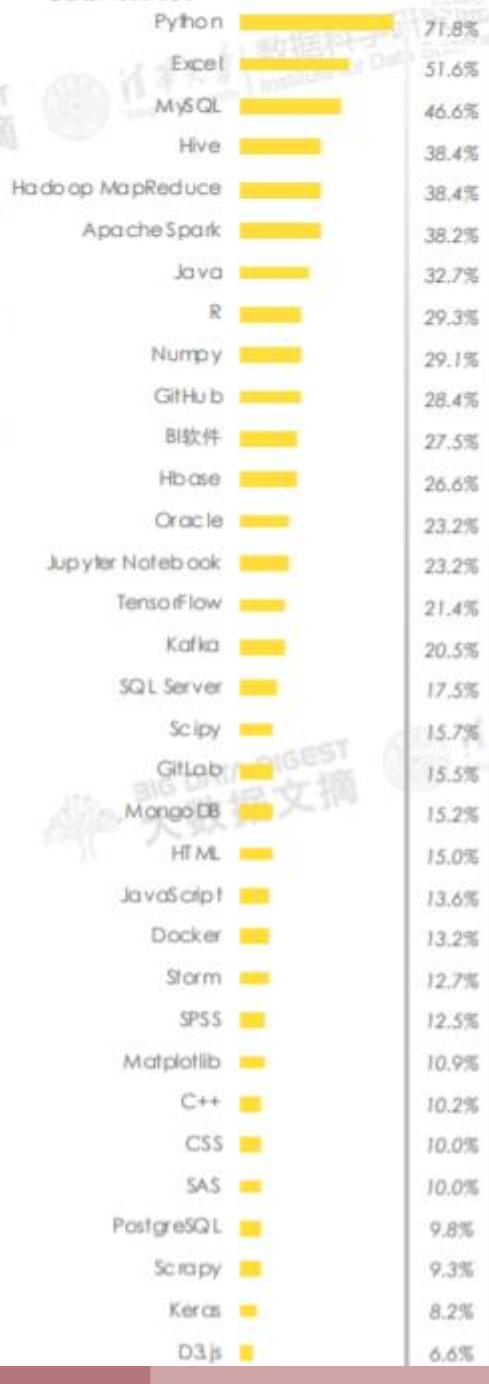
C、C++、JS 三者含金量最高

根据数据相关职位描述中要求的编程语言信息以及相对应的薪资水平，我们计算出了每种编程语言的"技能含金量指数"[4]。

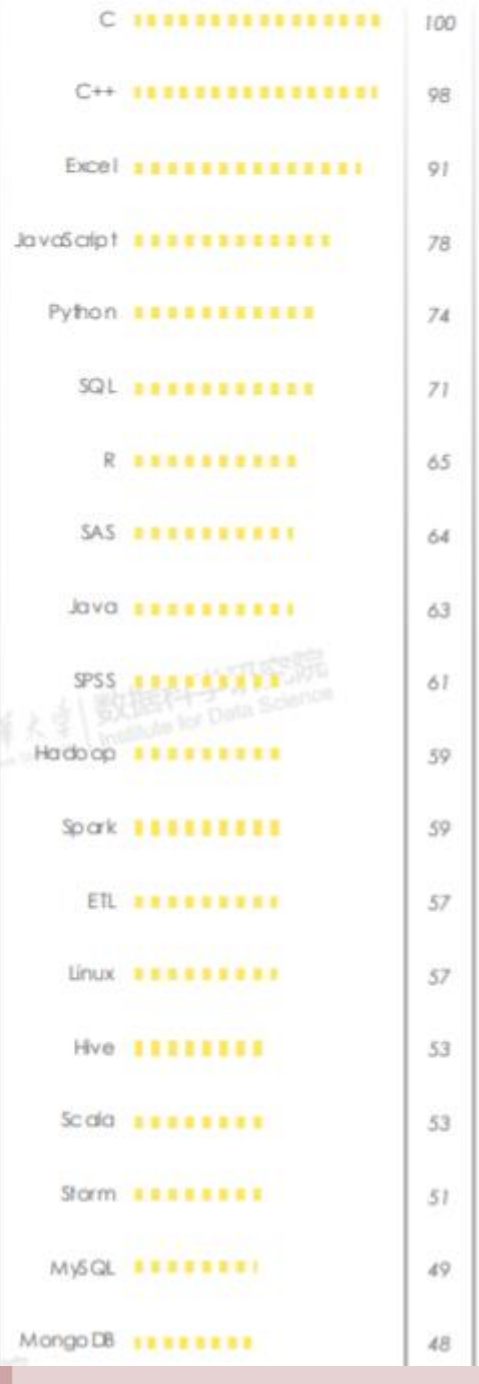
其中，C和C++ 两种编程语言含金量最高，其次则是 JavaScript、Python 和 SQL。

可以看出，在数据从业者中，Python 虽然最受欢迎，但大量的使用者也拉低了该语言的含金量。相反，使用人数占比相对较少的 C、C++ 和 JavaScript 在市场上更具薪资竞争力。

数据行业从业者各类数据分析工具使用人数比例



各类数据分析工具含金量指数

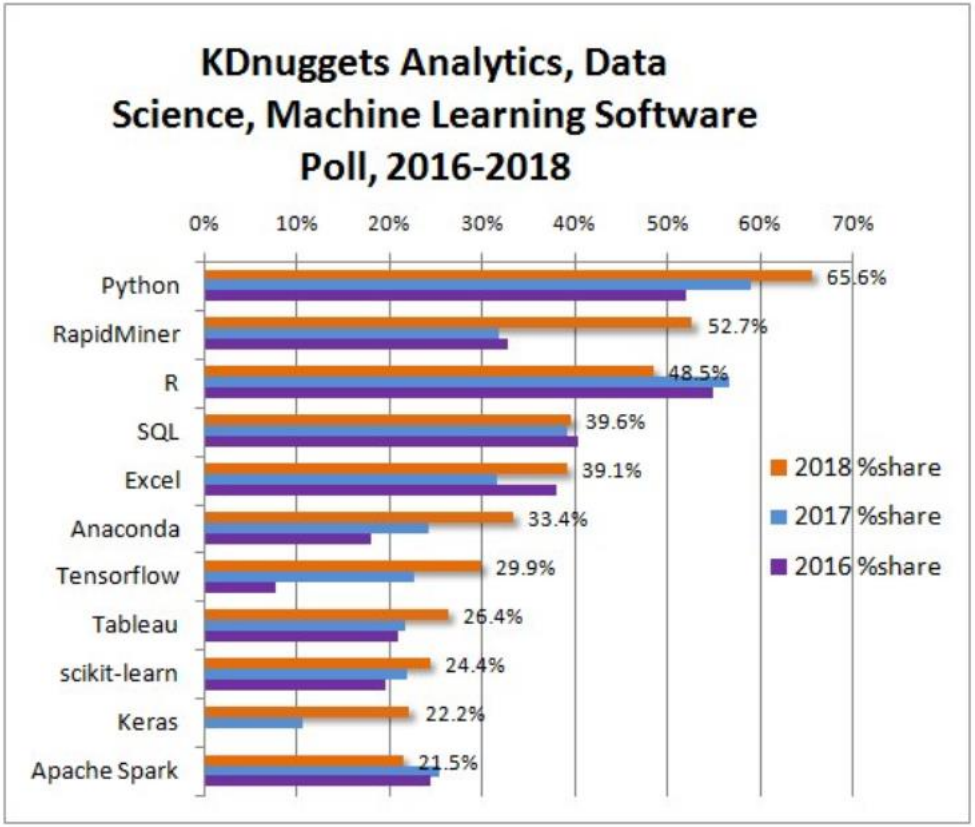


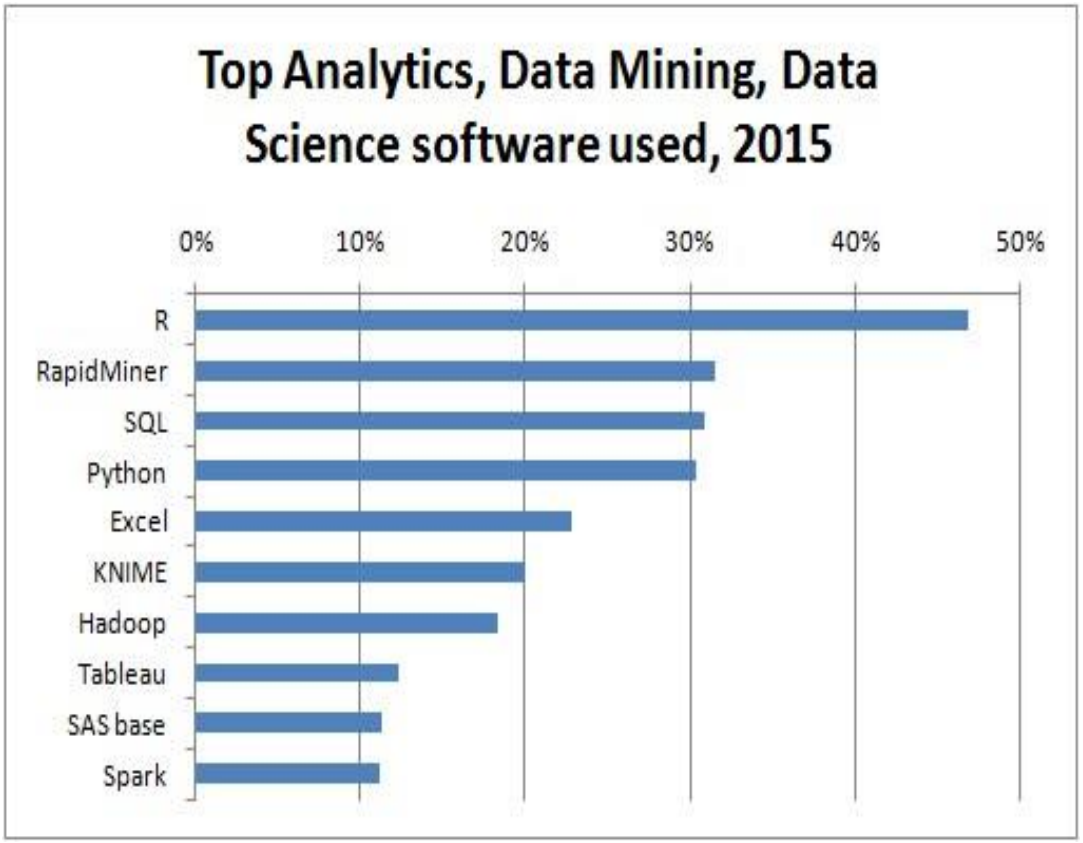
数据分析必备硬技能


理论知识
统计基础
机器学习
编程语言
R
Python
其他：SAS, Stata, SPSS
Excel
数据库
SQL
NoSQL
大数据工具
Hive
HBase
Spark
可视化工具
Tableau
Power BI

工欲善其事必先利其器

来源：https://www.kdnuggets.com/2018/05/poll-tools-analytics-data-science-machine-learning-results.html





SAS

- 全称：STATISTICAL ANALYSIS SYSTEM。
- 最早由北卡罗来纳大学的两位生物统计学研究生编制，并于1976年成立了SAS软件研究所，正式推出了SAS软件。
- 在数据处理和统计分析领域，SAS系统被誉为国际上的标准软件系统，并在96～97年度被评选为建立数据库的首选产品。
- 全世界120多个国家和地区的近三万家机构所采用，直接用户则超过三百万人，遍及金融、医药卫生、生产、运输、通讯、政府和教育科研等领域。
- 1997年，推出SAS 6.12版；2000年，推出SAS 8.0版；目前最新版本为9.4版。版本9.0+支持中文操作界面。
- SAS的特点
  - ●功能强大，统计方法齐，全，新。
  - ●使用简便，操作灵活。
  - ●提供联机帮助功能。

最新版本：9.5
官方地址：https://www.sas.com/zh_cn/home.html

R语言

- R语言由S语言演变而来，S语言于上世纪70年代诞生于AT&T贝尔实验室。
- 基于S语言开发的商业软件S-plus，在国外学术界应用很广。
- R语言由Auckland 大学统计系的Robert Gentleman和Ross Ihaka于1995年编写而成（R命名取其两人名字的首字母）。
- R很快得到广泛用户的欢迎，目前它是由R核心发展团队维持，它是一个由志愿者组成的工作努力的国际团队。

最新版本：4.2.0
下载地址：https://www.r-project.org/

Ross Ihaka & Robert Gentleman

R语言

- R的优势
  - R使用成本低。
  - R扩展性强。
  - R使用简单。

- R的劣势
  - R初始设计完全基于单线程和纯粹的内存计算，处理大数据受到限制。
  - R非“傻瓜”软件，需要一定的编程基础，需要足够的统计知识。
  - R的技能核定并没有官方或者机构标准，企业想招到R相关人才也不那么简单；
  - R的迁移成本高：对于大量工作已由其他软件实现（比如用SAS）的公司来讲，迁移成本很高。

R优秀扩展包弥补了性能问题！！（截至2022年3月30日，CRAN已经收录了各类包19033个。例如用于经济计量、财经分析、人文科学研究以及人工智能。）

中国R会

第八届中国R会议（北京大学，2015.5）

中国R会

第九届中国R会议（中国人民大学，2016.5）　第十届中国R会议（清华大学，2017.5）

PYTHON

- Python是一种面向对象的解释型计算机程序设计语言，由荷兰人Guido van Rossum于1989年发明，第一个公开发行版发行于1991年。
- 1989年圣诞节期间，在阿姆斯特丹，Guido为了打发圣诞节的无趣，决心开发一个新的脚本解释程序，做为ABC 语言的一种继承。之所以选中Python（大蟒蛇的意思）作为该编程语言的名字，是因为他是一个叫Monty Python的喜剧团体的爱好者。
- Python特点：
  - Python是一种面向对象、解释型计算机程序设计语言。
  - Python是纯粹的自由软件，它的语法简洁、易读以及有很强可扩展性。
  - Python具有丰富和强大的库，能够把用其他语言制作的各种模块（尤其是C/C++）很轻松地联结在一起。

最新版本：3.10.4
下载地址：https://www.python.org/

PYTHON

优点

1. 简单明了，学习曲线低，比很多编程语言都容易上手。
2. 开放源代码，拥有强大的社区和生态圈，尤其是在数据分析和机器学习领域。
3. 解释型语言，天生具有平台可移植性，代码可以工作于不同的操作系统。
4. 对两种主流的编程范式（面向对象编程和函数式编程）都提供了支持。
5. 代码规范程度高，可读性强，适合有代码洁癖和强迫症的人群。

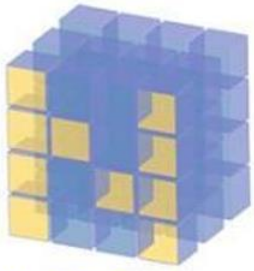
缺点

1. 执行效率稍低，对执行效率要求高的部分可以由其他语言（如：C、C++）编写。
2. 代码无法加密，但是现在很多公司都不销售卖软件而是销售服务，这个问题会被弱化。
3. 在开发时可以选择的框架太多（如Web框架就有100多个），有选择的地方就有错误。

数据采集：以Scrapy为代表的各类方式的爬虫。

数据链接：Python有大量各类数据库的第三方包，方便快速的实现增删改查。

数据清洗：Numpy、Pandas，结构化和非结构化的数据清洗及数据规整化的利器。



NumPy

Pandas

数据分析：Pandas、StatsModels、Scipy，统计分析，科学计算、建模等。

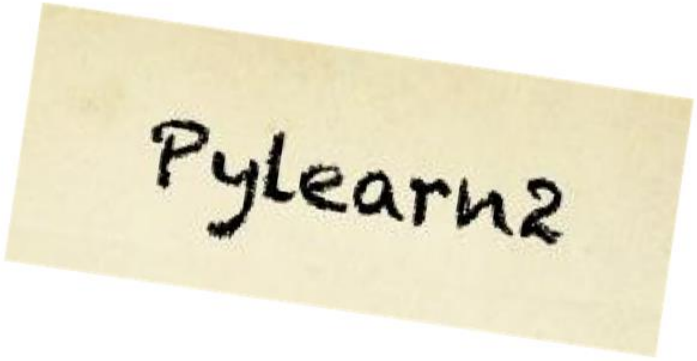
机器学习与深度学习


K
Keras


scikits
learn
machine learning in Python


theano


Caffe


Pylearn2


LASAGNE
Cutting through complexity

数据可视化：Matplotlib、Seaborn等等大量各类可视化的库。

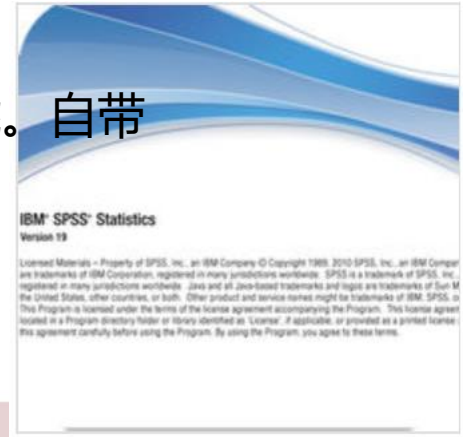
MATLAB

- MATLAB是美国MathWorks公司出品的商业数学软件，用于算法开发、数据可视化、数据分析以及数值计算的高级技术计算语言和交互式环境，主要包括MATLAB和Simulink两大部分。
- 20世纪70年代，美国新墨西哥大学计算机科学系主任Cleve Moler为了减轻学生编程的负担，用FORTRAN编写了最早的MATLAB。
- 1984年由Little、Moler、Steve Bangert合作成立了的MathWorks公司正式把MATLAB推向市场。到20世纪90年代，MATLAB已成为国际控制界的标准计算软件。
- Matlab特点：
  - 高效的数值计算及符号计算功能，能使用户从繁杂的数学运算分析中解脱出来；
  - 具有完备的图形处理功能，实现计算结果和编程的可视化；
  - 功能丰富的应用工具箱(如信号处理工具箱、通信工具箱等)，为用户提供了大量方便实用的处理工具。

最新版本：Matlab 2021
官方地址：https://www.mathworks.com/

SPSS

- SPSS是世界上最早的统计分析软件，由美国斯坦福大学的三位研究生Norman H. Nie、C. Hadlai (Tex) Hull 和 Dale H. Bent于1968年研究开发成功，同时成立了SPSS公司，并于1975年成立法人组织、在芝加哥组建了SPSS总部。
- 1984年SPSS总部首先推出了世界上第一个统计分析软件微机版本SPSS/PC+，开创了SPSS微机系列产品的开发方向，极大地扩充了它的应用范围，并使其能很快地应用于自然科学、技术科学、社会科学的各个领域。
- 2009年7月28日，IBM公司宣布将用12亿美元现金收购统计分析软件提供商SPSS公司。如今SPSS的最新版本为25，而且更名为IBM SPSS Statistics。
- SPSS的特点
  - ●操作简便：界面非常友好，大多数操作可通过鼠标拖曳来完成。
  - ●编程方便：具有第四代语言的特点。对于常见的统计方法，SPSS的命令语句、子命令及选择项的选择绝大部分由“对话框”的操作完成。
  - ●功能强大：具有完整的数据输入、编辑、统计分析、报表、图形制作等功能。自带11种类型136个函数。



最新版本：25.0

官方地址：https://www.ibm.com/analytics/spss-statistics-software

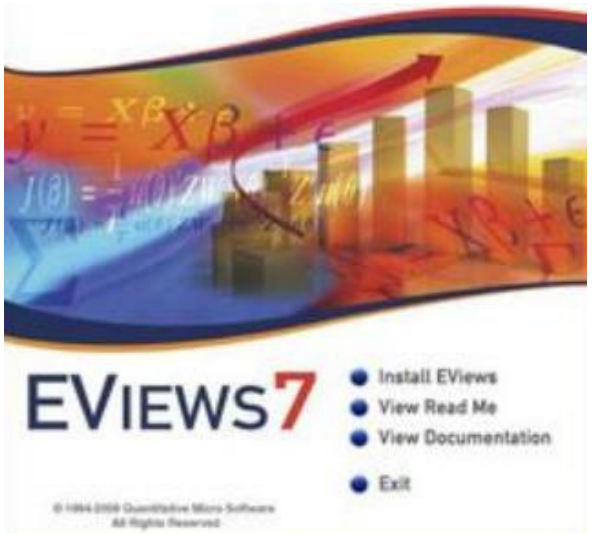
STATA

- Stata 是一套提供其使用者数据分析、数据管理以及绘制专业图表的完整及整合性统计软件。
- Stata的架构师是William Gould，诞生于1985年（确切说是1984年12月），是StataCorp的核心产品。
- 目前常用的Stata的版本为2009年7月推出的Stata 11.0。Stata 11包括四种版本：Small（小型版）、IC（标准版）、SE（特别版）和MP（多处理器版）。其中MP版本最为强大。MP版和SE版的功能完全相同，但MP版的运算速度比SE版的要快很多。
- Stata的特点
  - Stata的统计功能很强。
  - Stata的作图模块，可以满足绝大多数用户的统计作图要求。在有些非绘图命令中，也提供了专门绘制某种图形的功能，如在生存分析中，提供了绘制生存曲线图，回归分析中提供了残差图等。
  - Stata提供了多元统计分析中所需的矩阵运算。

最新版本：Stata 15
官方地址：https://www.stata.com/

EVIEWS

- Eviews是Econometrics Views的缩写，通常称为计量经济学软件包。是专门为大型机构开发的、用以处理时间序列数据的时间序列软件包的新版本。
- 1994年QMS（Quantitative Misro Software）公司在Micro TSP基础上直接开发成功Eviews并投入使用。
- Eviews的特点
  - Eviews处理的基本数据对象是时间序列
  - Eviews具有操作简便且可视化的操作风格，体现在从键盘或从键盘输入数据序列、依据已有序列生成新序列、显示和打印序列以及对序列之间存在的关系进行统计分析等方面。
  - Eviews具有现代Windows软件可视化操作的优良性。
  - Eviews还拥有强大的命令功能和批处理语言功能。



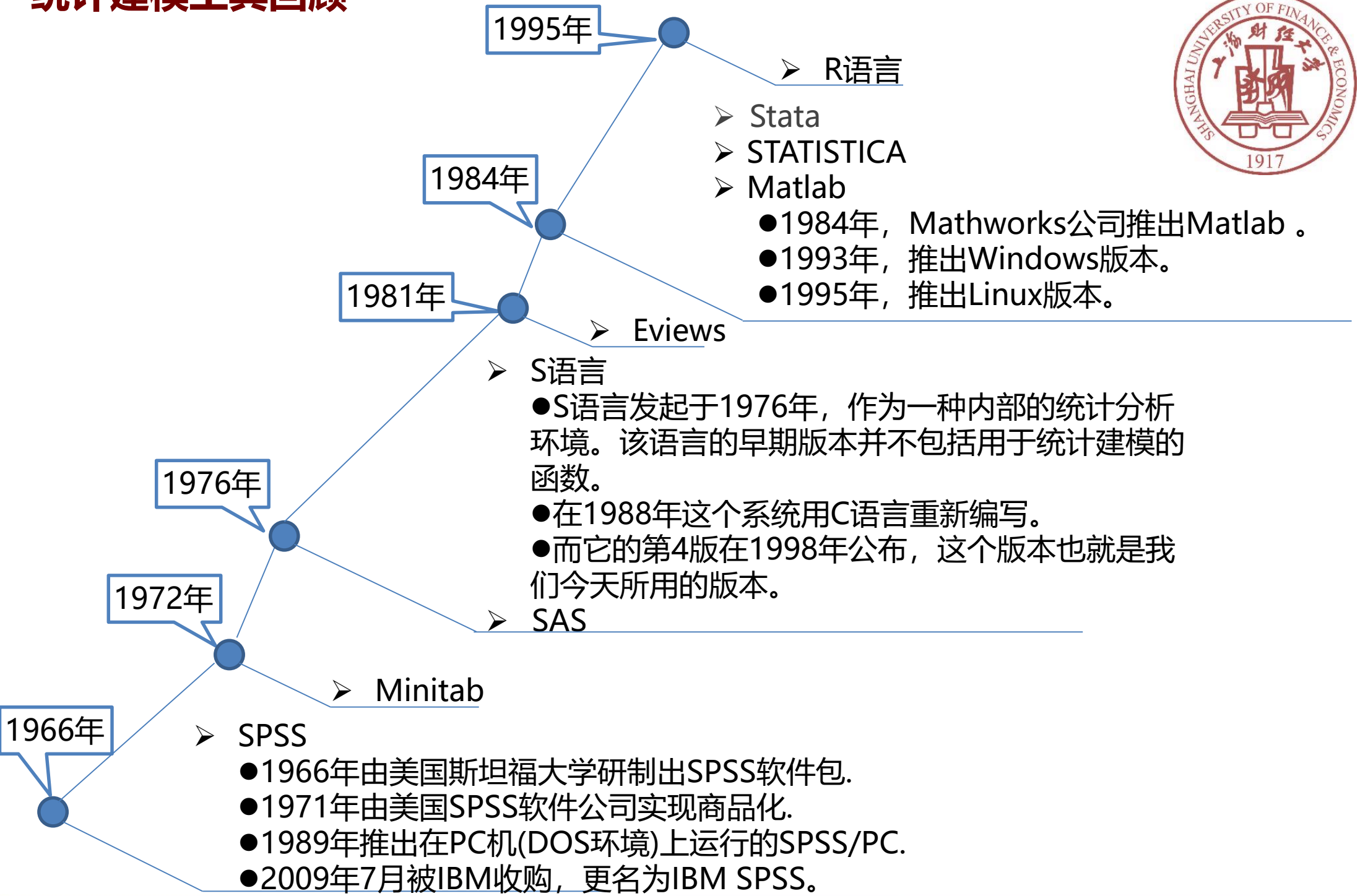
最新版本：Eviews 10
官方地址：http://www.eviews.com/home.html

统计建模工具回顾

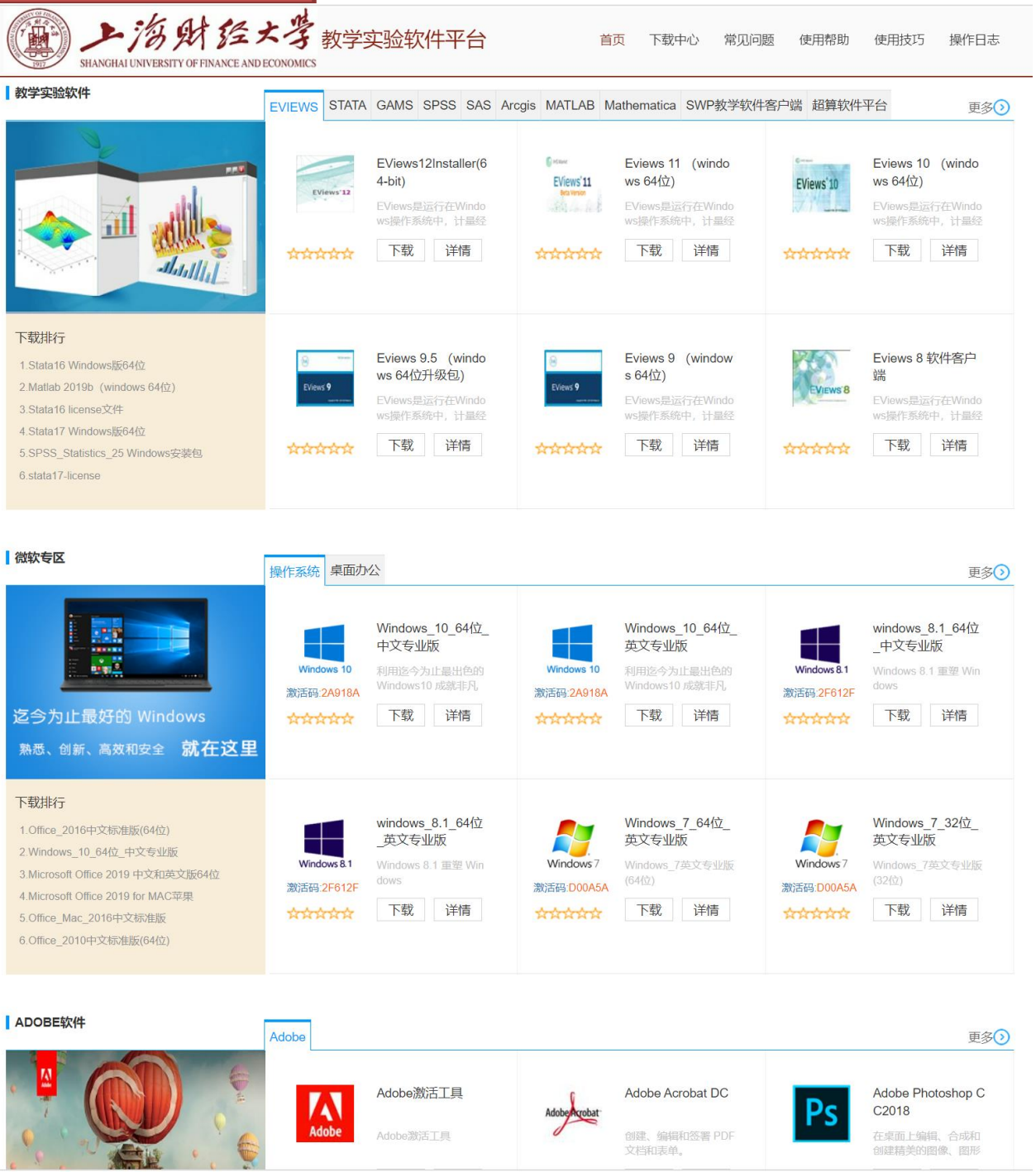
http://bbs.pinggu.org/

数据交流中心	调查问卷专版　数据求助
计量经济学与统计软件	Stata专版　stata上传下载区　EViews专版　Gauss专版　LISREL、AMOS等结构方程模型分析软件　IRT理论相关软件　winbugs及其他软件专版　HLM专版　LATEX论坛　统计从业与统计经管代码库　统计软件培训班VIP答疑区
商业数据分析	数据分析与数据挖掘　SPSS论坛　SAS专版　SAS上传下载区　R语言论坛　Excel　MATLAB等数学软件专版　Clementine&Mode　JMP论坛　数据分析师（CDA）专版　每天一个数据分析师
大数据技术	Hadoop论坛　python论坛　数据可视化　SQL及关系型数据库数据　Oracle数据库及大数据解决方案　mahout论坛　数据仓库技术　spark高速集群计算平台　nosql论坛　openstack云平台　storm实时数据分析平台　行业应用案例
机器学习技术	人工智能(自然语言处理/机器学习/智能设备与机器人)　人工智能论文版
IT基础	Scala及其他JVM语言　Linux操作系统　C与C++编程　JAVA语言开　比特币与区块链

统计建模工具回顾


1995年
R语言
Stata
STATISTICA
Matlab
●1984年，Mathworks公司推出Matlab 。
●1993年，推出Windows版本。
●1995年，推出Linux版本。
1984年
1981年
Eviews
S语言
●S语言发起于1976年，作为一种内部的统计分析环境。该语言的早期版本并不包括用于统计建模的函数。
●在1988年这个系统用C语言重新编写。
●而它的第4版在1998年公布，这个版本也就是我们今天所用的版本。
1976年
SAS
1972年
Minitab
1966年
SPSS
●1966年由美国斯坦福大学研制出SPSS软件包.
●1971年由美国SPSS软件公司实现商品化.
●1989年推出在PC机(DOS环境)上运行的SPSS/PC.
●2009年7月被IBM收购，更名为IBM SPSS。

上财教学实验软件平台
https://software.shufe.edu.cn



我与统计建模工具

求学

计算机语言：C、Fortran

计算机相关：数据库、数据结构、高级程序设计

统计软件：SAS、R、Matlab

证书：程序员

教学

2006-2010：统计软件S-Plus、R语言、C语言、Fortran语言

2012-2014：计算机编程（专硕）、统计软件和统计计算（博士、硕士）

2012-2022：统计软件（本科、硕士）

科研

早期：Fortran

现在：R语言

未来：R/Python

软件课程开设情况

学年	学期	课程名称	学历层次	课程类别	课时	人数
2018-2019	1	统计软件	大学本科	选修课	64	47
2018-2019	1	统计软件（SAS）	硕士研究生	学位基础课	48	54
2018-2019	1	统计软件	大学本科	必修课	64	65
2017-2018	1	统计软件	大学本科	必修课	64	70
2017-2018	1	统计软件（SAS）	硕士研究生	学位基础课	48	52
2016-2017	2	统计软件（SAS）	硕士研究生	学位基础课	48	22
2016-2017	1	统计软件	大学本科	必修课	64	63
2016-2017	1	统计软件（SAS）	硕士研究生	学位基础课	48	68
2015-2016	2	统计软件（SAS）	硕士研究生	学位基础课	48	61
2015-2016	1	统计软件	大学本科	专业必修课	64	70
2014-2015	1	统计软件（SAS）	硕士研究生	学位基础课	51	45
2013-2014	2	统计软件（SAS）	硕士研究生	公共选修课	36	23
2013-2014	2	统计软件	大学本科	任意选修课	34	36
2013-2014	1	统计软件（任选）	大学本科	任意选修课	34	46
2013-2014	1	统计建模与统计软件	硕士研究生	学位基础课	36	67
2012-2013	2	统计软件	硕士研究生	专业选修课	27	43
2012-2013	1	统计软件	大学本科	学科共同课	68	67
2012-2013	1	统计软件	大学本科	学科共同课	68	30
2012-2013	1	计算机编程	硕士研究生	专业必修课	36	34
2012-2013	小	应用统计软件（双）	大学本科	专业方向课	30	49
2011-2012	2	统计建模与统计软件	硕士研究生	学位基础课	36	45
2010-2011	1	统计计算与统计软件	硕士研究生	专业必修课	45	18
2010-2011	小学期	应用统计软件（双）	大学本科	专业方向课	30	42
2009-2010	2	计算机程序设计	大学本科	学科共同课	68	51
2009-2010	2	计算机程序设计	大学本科	学科共同课	68	47
2009-2010	1	统计计算	硕士研究生	专业必修课	36	6
2009-2010	小学期	应用统计软件（双）	大学本科	专业方向课	30	70
2008-2009	2	统计软件	硕士研究生	专业选修课	27	23
2008-2009	2	计算机程序设计	大学本科	学科共同课	52	49
2008-2009	2	计算机程序设计	大学本科	学科共同课	52	47
2007-2008	2	统计计算	硕士研究生	专业选修课	27	38
2007-2008	2	计算机程序设计	大学本科	学科共同课	51	44
2007-2008	2	计算机程序设计	大学本科	学科共同课	51	44
2007-2008	1	统计软件	大学本科	学科共同课	68	20
2006-2007	2	统计软件	大学本科	学科共同课	68	46
2006-2007	小学期	应用统计软件（双）	大学本科	专业方向课	30	38
2005-2006	2	统计软件	大学本科	学科共同课	68	40
2005-2006	2	统计软件	大学本科	学科共同课	68	44


41
42

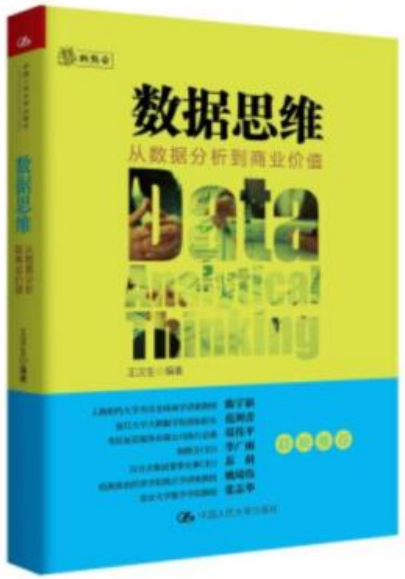

21
26
31
36


Thank you!

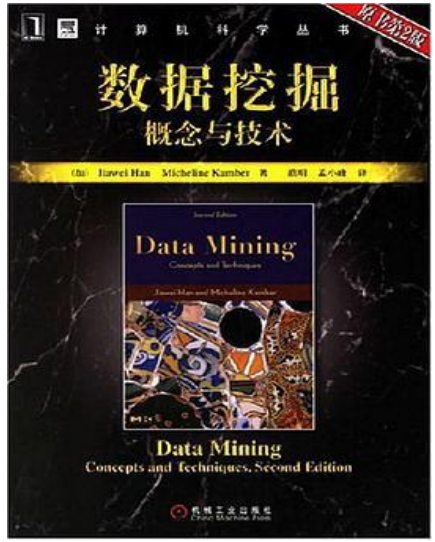

sas
THE POWER TO KNOW.
Data Mining Champion
2013 • 北京
上海财经大学
单位：赛仕软件（北京）有限公司
单位：中国科学软件网
sas

推荐书单

数据思维

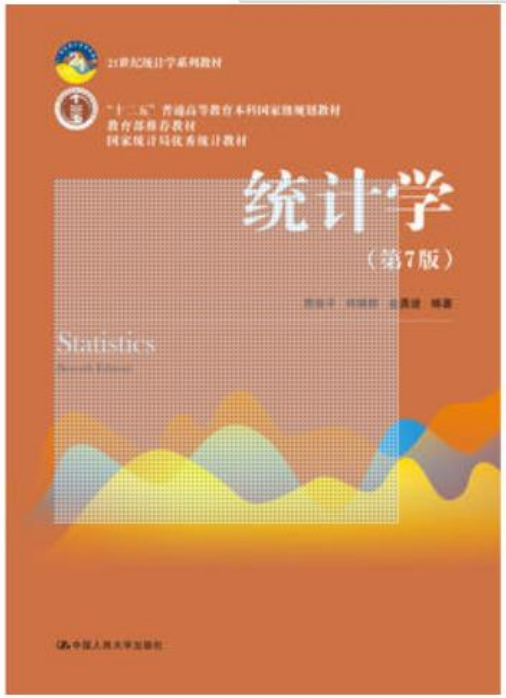


数据挖掘&机器学习

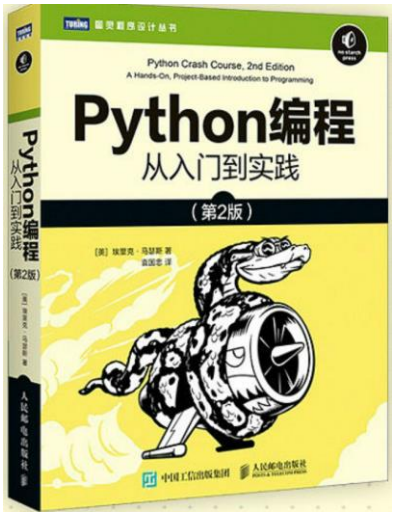


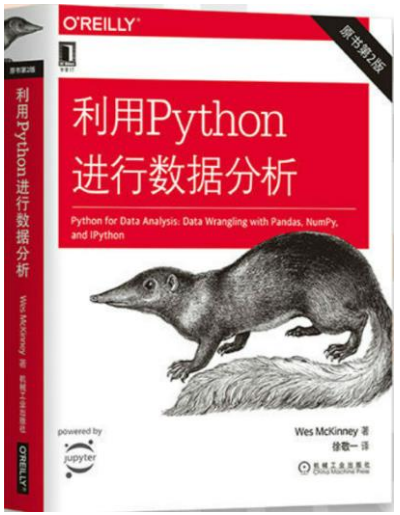
推荐书单

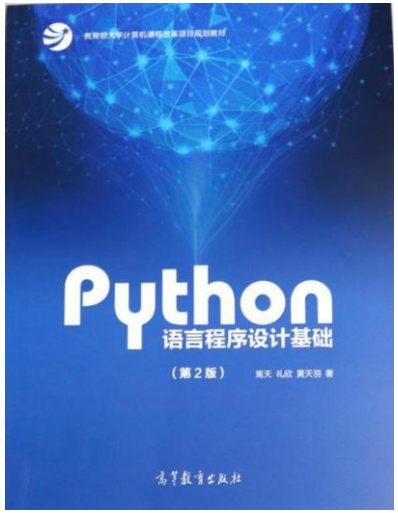
统计学



Python







PART 四

项目报告撰写

报告框架

1 封面（标题等）

2 目录 （论文目录、表格目录、插图目录）

3 摘要和关键词

4 论文主体

5 参考文献

6 附录（问卷、深度访谈和补充材料等）

华东师范大学李艳副教授，《市场调查与分析大赛参赛准备与技巧》（2021年）

上海财经大学
SHANGHAI UNIVERSITY OF FINANCE AND ECONOMICS

目录范例

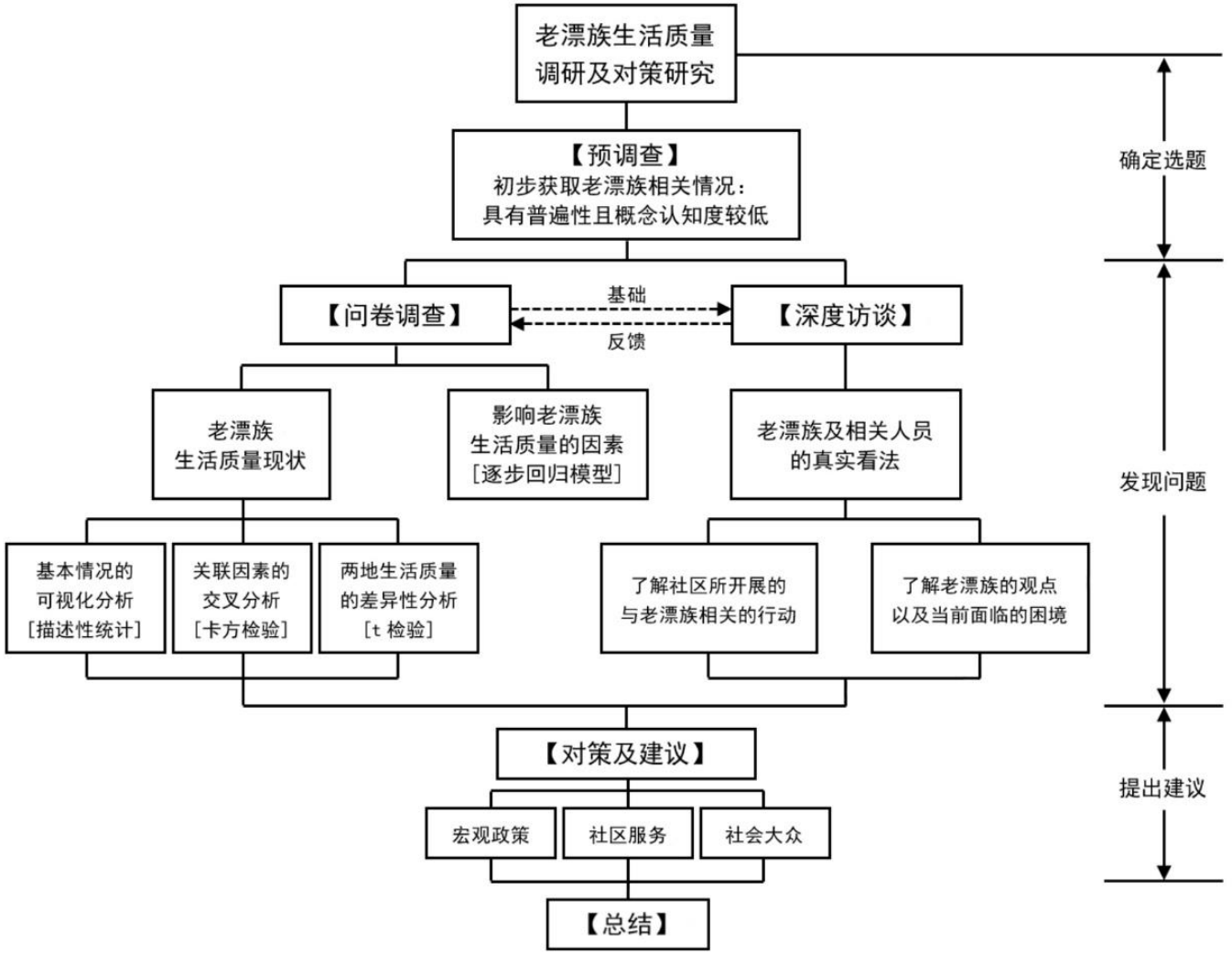
吴纯杰、黄枫指导，2021年挑战杯国赛一等奖作品《疫情冲击下惠企纾困政策分析与成效调研》

上海财经大学
SHANGHAI UNIVERSITY OF FINANCE AND ECONOMICS

图表目录示例

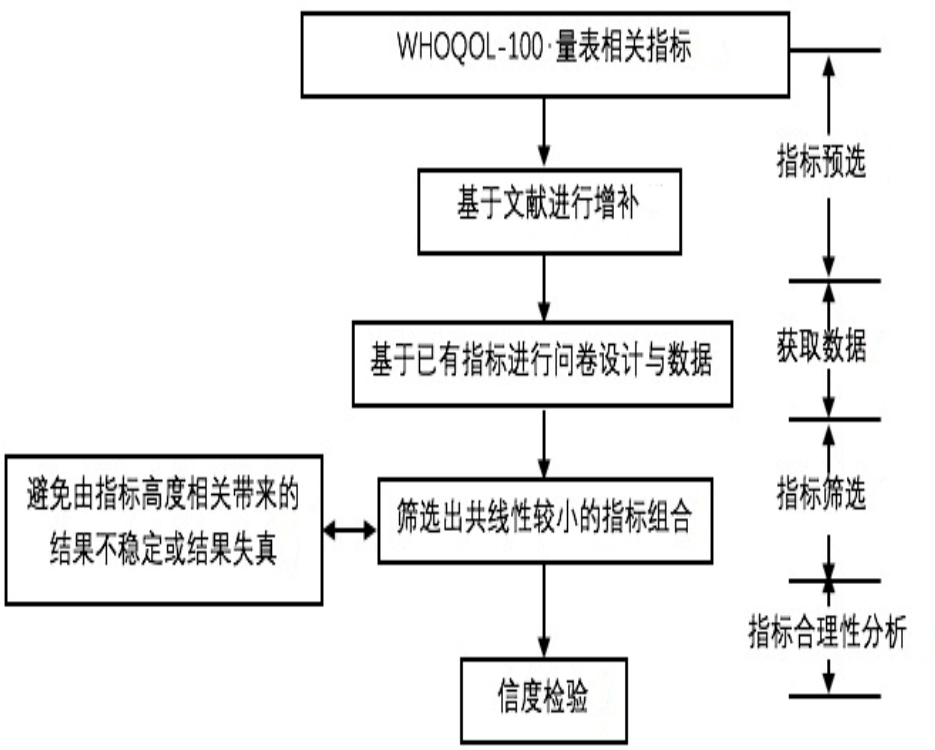
吴纯杰、黄枫指导，2021年“挑战杯”国赛一等奖作品《疫情冲击下惠企纾困政策分析与成效调研》

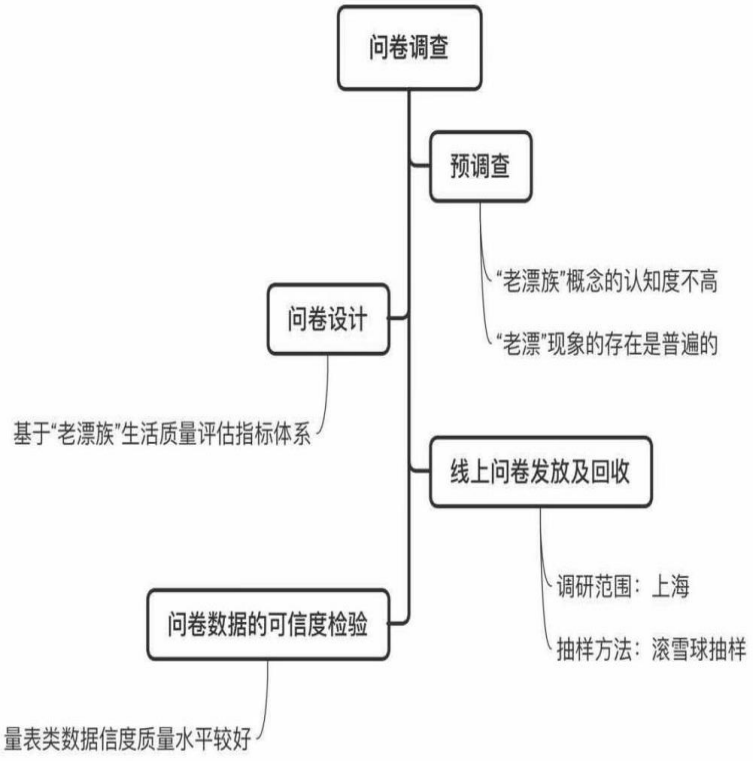
全文思维导图



门嘉齐，《2021年挑战杯国赛三等奖、上海市特等奖作品案例分享》

模块思维导图





门嘉齐，《2021年挑战杯国赛三等奖、上海市特等奖作品案例分享》

摘要

- 一定要将论文创新点、主要想法、做法、结果、分析结论表达清楚。
- 一篇完整的陈述性短文，具有自含性和独立性。
- 应有论文全文的主要信息，突出新见解或创新性，或突出创造性成果。
- 基本要素：（1）目的；（2）方法；（3）结果；（4）结论；（5）其他。

华东师范大学李艳副教授，《市场调查与分析大赛参赛准备与技巧》（2021年）

论文主体

1 引言（研究背景、文献综述、研究目标与思路）

2 调查设计与实施（调查方案设计、问卷调查安排与实施、样本情况分析、质量控制效果分析）

3 对问题的研究分析（按数据分析流程描述、推断、建模等开展）

4 结论、建议、不足与展望

华东师范大学李艳副教授，《市场调查与分析大赛参赛准备与技巧》（2021年）

调研流程——线下深度访谈

调研对象

符合条件的“老漂族”：整群抽样

已有老漂活动经验的有代表性的社区居委会

问题设计

基于线上问卷：
- 改编和增加开放性问题
- 合并小问题
- 将问题口语化

- “老漂族”访谈时所发现的问题
- “老漂族”人数统计
- 老漂活动开展情况
- 社区的福利政策

访谈流程

表明调查组身份 → 说明调查目的 → 询问受访意愿及被记录意愿 → 进行访谈
*注意对受访者个人信息的保密，以及控制访谈时间

门嘉齐，《2021年挑战杯国赛三等奖、上海市特等奖作品案例分享》

典型数据分析流程

数据获取

- 获取源数据
- 理解数据结构
- 了解数据收集机制

数据清理

（提升数据质量）

- 处理缺失值
- 数据类型转换
- 删除低价值数据 …

数据转换

（准备分析数据）

- 合并表格
- 转换表格形式
- 创建新变量 …

数据探索

（充分理解数据和信息）

- 数据可视化
- 各种表格转换
- 非监督学习 …

统计分析与建模

（得出分析结论）

- 推断统计
- 基本统计量计算
- 机器学习…

分析呈现

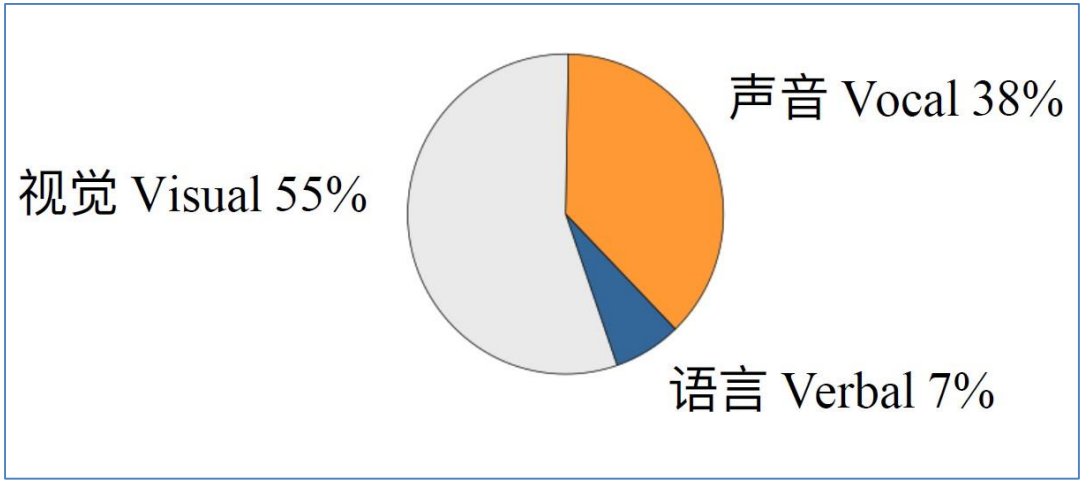
（报告结论/洞见）

- 分析报告
- 数据产品（数据看板）…

展示答辩

- 科学的研究设计
- 严谨的研究过程
- 高质量的数据
- 正确的分析方法
- 有价值的结论与建议

- 在有限的时间内，有理有据地展示出研究的精华
- 做好万全的准备
- 充分调动各种元素

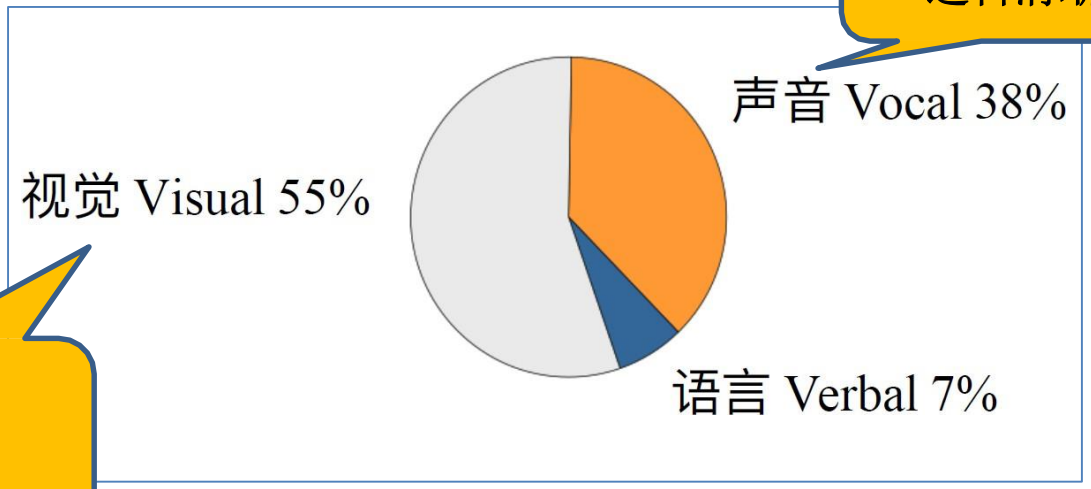


华东师范大学李艳副教授，《市场调查与分析大赛参赛准备与技巧》（2021年）

展示答辩

- 科学的研究设计
- 严谨的研究过程
- 高质量的数据
- 正确的分析方法
- 有价值的结论与建议

- 在有限的时间内，有理有据地展示出研究的精华
- 做好万全的准备
- 充分调动各种元素



- 声音自然
- 逻辑清晰

- 版面简洁
- 文字精炼
- 图表美观
- 动画流畅

华东师范大学李艳副教授，《市场调查与分析大赛参赛准备与技巧》（2021年）

如何组建一支优秀的队伍

- 有积极性、脚踏实地
- 交流、分工、配合
- 以成员特长的合理利用
- 跨院、跨专业、跨校（市级及以上）

如何在大创项目中发挥出色

- 提出问题、解决问题
- 结论具有推广性、有应用或理论价值
- 组长的魅力和领导力，组员的执行力和协作
- 知识储备，多读论文（先从文献综述训练）
- 发挥指导老师作用（方向把握、大局观、资源支持等）

项目经验分享

开题选题

1.结合时事政策，注重可行性

2.大方向需要规划

3.与指导老师积极交流

项目进行中

1.组长起到统领作用

2.各成员积极反馈进度

3.发挥每个人擅长的地方

项目总结

1.注重论文修改与图片美化

2.演讲稿逐字打磨

3.总结经验教训

陆佳雯，《2021年“挑战杯”国赛一等奖、上海市特等奖作品经验分享》

谢 谢!

Thank You