

第十一届中国 R 会议（上海）暨华东地区数据科学会议



All for Data Science

主办方：

华东师范大学统计学院
华东师范大学数据科学与工程学院
华东师范大学教育信息技术学系

统计之都

赞助商：

RStudio

卓铭保险

人民邮电出版社

2018 年 12 月 8 日 – 9 日

R 语言简介

R 是一个有着统计分析功能及强大作图功能的语言环境和软件系统，由新西兰奥克兰大学统计系的 Ross Ihaka 和 Robert Gentleman 共同创立。R 语言可以看作是由 AT&T 贝尔实验室所创的 S 语言发展出的一种方言。

R 是在 GNU 协议 General Public Licence 下免费发行的，它的开发及维护现在则由 R 开发核心小组 *R Development Core Team* 具体负责，这个团队的成员大部分来自大学机构（统计及相关院系），包括牛津大学、华盛顿大学、威斯康星大学、爱荷华大学、奥克兰大学等。除了这些作者之外，R 还拥有一大批贡献者（来自哈佛大学、加州大学洛杉矶分校、麻省理工大学等），他们为 R 编写代码、修正程序缺陷和撰写文档。

R 的功能很大程度上是通过程序包（Package）来实现的，迄今为止，R 语言官网上的程序包数目已经超过 7000 个，广泛地覆盖了数据分析应用到的各类行业和领域。各种统计前沿理论方法的相应计算机程序都会在短时间内以软件包的形式得以实现，这种速度是其它统计软件无法比拟的。

在 KDnuggets 于 2015 年 7 月做的“首选何种编程语言进行分析、数据挖掘及数据科学的工作”的调查中，R 以 51% 的得票率荣登榜首，力压 Python、SAS 和 MATLAB (<http://www.kdnuggets.com/polls/2015/r-vs-python.html>)，连续 4 年位居榜首。在 2015 年 5 月的另一项调查“首选何种数据分析、数据挖掘、数据科学软件或工具”中，R 超过了 2014 年的冠军 RapidMiner，同样位列第一 (<http://www.kdnuggets.com/polls/2015/analytics-data-mining-data-science-software-used.html>)。

目前，几乎所有的西方大学与研究机构、以及越来越多的金融机构、制药公司、高科技企业都使用 R。R 的灵活性、开放性以及业界最广泛的支持是其不断完善和发展的根本原因，随着 R 越来越被学术界及业界认可，它也将在数据分析和统计建模中发挥越来越大的作用。

华东师范大学统计学院简介

华东师范大学是中国获批最早设立概率论与数理统计专业的大学之一。1986年，概率论与数理统计获原国家教委批准设立博士点，1987年成为全国高等学校重点学科，是全国最早的三个概率论与数理统计国家重点学科之一。

学院拥有统计学博士后流动站、统计学一级学科博士点、2个专业学位硕士点（应用统计硕士、保险硕士）、4个本科专业（统计学、经济统计学、金融工程、保险学）。另外，统计学院参与创办并自1985年起一直承办中国概率统计学会期刊《应用概率统计》。2017年与中国现场统计学会合作创建了统计学学术期刊《Statistical Theory and Related Fields》。

自2012年以来统计学院在*Annals of Statistics*, *Journal of the American Statistical Association*, *Biometrika*, *Annals of Probability*, *Insurance Mathematics & Economics*等国际顶尖SCI期刊上以第一作者或通讯作者共发表论文近180余篇。在数据科学的顶尖会议及期刊上发表论文26篇。2017年，成功加入北美精算师协会（SOA）的UCAP（Universities & Colleges with Actuarial Programs）高校计划。2017年，华东师范大学统计学院获批成立了统计与数据科学前沿理论及应用教育部重点实验室。同年，华东师范大学统计学在教育部学科评估中位列全国第三位（并列），并入选教育部“双一流”学科建设行列。

华东师范大学数据科学与工程学院简介

华东师范大学数据科学与工程学院（简称：数据学院）于2016年9月成立。数据学院的创建是华东师范大学五年来秉持“因势而谋，应势而动，顺势而为”基本理念，坚持求变求新，抓住历史机遇的结果。

数据学院是一所举“科学”与“工程”并重、既强调理想又讲究务实的学院。学院兼具人才培养、科学研究和社会服务三大功能。对于人才培养，学院力求培养符合新时代需求的综合型数据人才，使其兼具“数据科学家”的思考能力和“系统架构师”的务实才干。围绕此目的，学院将利用三年时间完善数据专业和学科的培养计划，编写一套教材，形成系统的从本科到博士的完整培养体系。对于科学研究和社会服务，学院认为二者是一枚硬币的两面。出于“做真的研究，做有用的研究”的初衷，学院将以实际问题和社会痛点为出发点，用创新思路和开源技术解决问题，形成应用创新和技术创新相互促进的良性循环。

为此，学院将坚持“应用驱动创新”的基本理念，积极发展与企业的合作伙伴关系，努力将学院打造成创新型企业的研发基地和智慧“外脑”，同时也将校企合作作为培养学生创新精神和创新能力的重要力量，培养能够实施“大众创业万众创新”战略的合格人才。

华东师范大学教育信息技术学系简介

华东师范大学教育信息技术学系于1978年成立，是全国创建最早、研究信息技术在教育中应用的文理交叉学科。建系近三十年来，经过师生共同努力，在学科建设方面达到了国内领先的水平，同时也树立了以计算机教育应用为特色的专业品牌。

教师主要研究聚焦于教育信息化理论与系统规划、学习科学与技术设计、数字媒体与数字出版研究、教育信息化装备、环境及技术标准研究以及计算机支持教育测量、评价及管理等领域。目前承担多项国家级课题并积极参与国际合作交流，与美国、英国、荷兰、日本、新加坡等国，以及香港、台湾等地区高等院校有长期的合作、交流关系。

教育科学和技术飞速发展的今天，教育信息技术学系努力寻找新的突破点和增长点。如先后与苏州工业园区、广州TCL集团共建了全国最早的两个教育技术学博士后科研工作站。又如面向全国进行教育技术应用（示范性）实验区建设，首批实验区建设已在江苏省、浙江省、山东省等所属的几个县、市级区域全面展开，协助这些地区教育信息化的整体推进。还如正在积极修订本科生和研究生的教学计划，并对专业实验室进行全面的改造。这些表明，华东师大教育信息技术学系正在展现它新的风貌。

统计之都简介*

统计之都（Capital of Statistics，简称COS）成立于2006年5月，是一个旨在推广与应用统计学知识的网站和社区，今年是其成立的第十二周年。其主要依托主站（<http://cosx.org/>）、微信公众号（CapStat）、新浪微博（统计之都）、中国R语言会议（<http://china-r.org>）、线下线上沙龙等平台活动推广与应用统计学知识，希望统计理论研究者能充分关注应用问题，而统计应用者也能正确把握统计学基本知识，将统计学这门应用学科真正的潜力开发出来。同时，统计之都也希望科研人员、数据分析人员和统计学爱好者能互相交流合作，一方面促进彼此知识和技能的增长，另一方面为国内的统计学和数据科学贡献自己的力量。

目前主站的原创文章量已近500篇，微信公众号粉丝数达57000余人，微博粉丝数达46000人，与多所院校合作举办R会议11届，会议累计合作的赞助商超过20家，2017年在5个省市举办R会累计报名人数近万人。目前已举办国内线下沙龙43期及海外线上沙龙23期，统计之都的核心会员累计创作、翻译相关书籍10余本。

统计之都大力欢迎所有应用和研究领域的朋友与我们在实际应用问题上合作！

*统计之都网址：<http://cosx.org/>

RStudio简介



RStudio公司成立于2008年，创始人为JJ Allaire，R社区领军人物Hadley Wickham 现任RStudio 首席科学家。RStudio旨在为R语言提供更便利的开发环境和数据分析工具，例如RStudio集成开发环境（IDE）、RStudio服务器、Shiny、Shiny服务器、ShinyApps.io、R Markdown、RStudio Connect等。RStudio坚定支持开源软件和社区，其产品多为免费开源软件，但同时RStudio也提供相应的企业级软件应用（如RStudio服务器专业版、Shiny服务器专业版等），以满足商业使用需求（如企业内部RStudio服务器管理、售后服务支持）。自2012年起，RStudio为世界各地的R会议提供了大量赞助和支持，包括官方R语言会议和中国R语言会议。为了R语言能更持续稳定发展，RStudio倡议与微软、Tibco、Google等几家商业公司成立了R联合团体（R Consortium），每年为R社区的开源项目提供大量资助，召集优秀人才解决R语言现存的重要且有挑战性的问题。

卓铭保险简介



卓铭保险（Charming Insurance）隶属润安国际保险经纪有限公司，国内领先的创新型保险科技平台，注册资本5000万人民币，拥有中国银保监会认可互联网保险销售资质。卓铭保险联合全球顶级的保险集团BUPA及国内知名保险公司众安在线、泰康在线、招商信诺、中意保险、复星联合健康等，致力于为当代精英人群提供全球保险定制服务。核心团队由BAT等知名互联网保险公司及PICC、招商信诺、众安在线等知名保险机构的资深人士组成，以保险科技InsurTech为驱动力，高度重视用户体验、服务及数据安全。

卓铭保险始终坚持用户至上，根据用户实际需求，与保险公司共同为用户定制针对性保险服务。用户不但可以享有国际保障权益，还可以享受保后一站式服务。

人民邮电出版社简介



人民邮电出版社
POSTS & TELECOM PRESS

人民邮电出版社成立于1953年10月，是工业和信息化部主管的大型专业出版社，隶属于中国工信出版传媒集团。人民邮电出版社是全国优秀出版社、全国百佳图书出版单位，荣获“中国出版政府奖先进出版单位”“全国文明单位”“中央国家机关文明单位标兵”等重要荣誉，出版领域涵盖科技出版、教育出版、大众出版，涉及信息技术、经济管理、摄影、心理学、少儿、大中专教材等十余个出版门类。年出版新书约3000种，再版图书超过5000种，年销售码洋超过18亿元。

First hack 数据骇客倡议书

“Hack（黑客）”常被大家视为一个有负面意义的名词。然而同样一个英文单词，如果将其翻译为“骇客”，似乎就显得中立了些。

我们办了10多年R语言会议，支持数据科学推广。现在大数据和人工智能的结合，使得统计专业和计算机越走越近，动手能力要求越来越高。不论是年度数据大会还是日常的科学沙龙，见到分享最多的主要是工业界的商业项目，和一些学术科研的成果。随着技术的成熟和科研方向的聚焦，很多话题某种程度上已经很窄或者很难看到新意。但与此同时，依然有很多有趣的非商业，非科研项目，比如纽约时报基于人口关系网络的时代迁移可视化等等¹。

只要数据的好奇心还在，对世界的探索欲望还在，相信还有很多值得把玩的数据课题。为了鼓励青年学生和刚毕业的开发者参与实操，我们倡议发起这个First hack数据骇客青年展。每年上海R会期间会挑选在校学生或者毕业2年以内的年轻数据开发者项目，给予表彰和奖励，欢迎大家报名。

2018年上海的R会上会有参展项目的展示和优秀项目的颁奖。奖品也是一个骇客好朋友赞助的，他的公司装备前线²专门给游戏和计算机爱好者定制键盘、鼠标、音响等极客设备。这次我们定制了数套千元级机械键盘，以R会的蓝白配色，和希腊字母的涂装，打造数据科学家趁手的神兵利器。

zFrontier装备前线，是一群中美两地热爱生活的技术宅，程序猿和发烧友在硅谷创立，为发烧友服务的极客装备平台。zFrontier用系统的方法发现世界上最好的装备，通过与设计师和生产商直接合作，使会员能以独家优惠的价格享受最好最新的产品。zFrontier坚信世界的各个方面都能被更优化。我们使用的产品决定了我们如何与世界互动，不同的互动方式使得我们的生活体验完全不同。所以，我们对选什么产品极端认真。我们的创始团队来自斯坦福，和上海交大，既有经验丰富的连续创业者，又有行业专家，刚成立就获得了硅谷和国内一线美元VC的投资。zFrontier已经获得众多发烧友的肯定和喜爱，你也来看看吧！³

¹<https://www.nytimes.com/interactive/2018/09/19/upshot/facebook-county-friendships.html>

²<https://www.zfrontier.com>

³<http://www.zfrontier.com/>

第十一届中国 R 会议（上海）暨华东地区 数据科学会议日程

1. 日程安排

12月8日	注册和主会场报告	华东师大闵行校区，体育馆
12月9日	分会场报告	华东师大中山北路校区，逸夫楼报告厅
12月9日	分会场报告	华东师大中山北路校区，科学会堂一楼报告厅
12月9日	分会场报告	华东师大中山北路校区，科学会堂二楼报告厅
12月9日	分会场报告	华东师大中山北路校区，文史楼 303

2. 会议议程

12月8日 主会场

时间	姓名	报告标题	主持人	
09:00-09:15	会议致辞			
09:15-09:50	张丹	R 语言商业实践-从金融市场到区块链	文茜	
09:50-10:25	刘思喆	数据科学如何助力在线教育革命		
10:25-10:40	Break			
10:40-11:15	夏虞斌	个人数据的加密运算		
11:15-11:50	谢军	Educational Big data and Machine Learning, Case Study from Shanghai		
午餐				
14:00-14:35	王昱舜	Data Analysis for Basketball Tactics	孙嘉怡	
14:35-15:10	谢宗震	智能制造导入实务		
15:10-15:25	Break			
15:25-16:25	Lightning Talk			
16:25-17:00	陈洁宁	AI on Device 精准脱贫		

12 月 9 日 逸夫楼报告厅

会场名称	时间	姓名	报告标题	主持人
自然语言 会场	09:00-09:30	钱亦欣	基于 R 语言的网络文学评论挖掘	夏晓凯
	09:30-10:00	Marco Li	文本挖掘在商业分析中的应用	
	Break			
	10:20-10:50	季雨清	数据如何助力人职匹配？	
	10:50-11:20	李翛然	如何做一个成功的商业对话机器人	
	11:20-11:50	吴子彧	AI 时代的智能服务	
午餐				
深度学习 会场	14:00-14:40	尹志	在浏览器里深度学习 – 使用 TensorFlow.js 构建人工智能应用	陈新宇
	14:40-15:20	付星宇	AlphaGomoku：一个基于 AlphaGo 的五子棋人工智能	
	Break			
	15:40-16:20	李翔	目标检测技术在携程图像智能化中的实践	
	16:20-17:00	陈新宇	MXNet Graph Optimization and Quantization based on Intel® MKL-DNN	

12 月 9 日 科学会堂一楼报告厅

会场名称	时间	姓名	报告标题	主持人
数据安全 会场	09:00-09:40	张恺	跨行业数据融合应用案例分享	张翔
	09:40-10:20	张翔	如何让车联网隐私数据安全的流通和使用	
	Break			
	10:40-11:20	余炀	基于区块链与 TEE 技术的数据隐私保护	
	11:20-12:00	郭健美	大规模数据中心的性能分析	
午餐				
数据平台 会场	14:00-14:35	韩俊仙	数据分析挖掘在制造领域的应用价值	程临峰
	14:35-15:10	孙繁荣	数据治理与数据资产管理	
	Break			
	15:30-16:05	刘心广	基于 R 构建某工厂质量在线监控平台	
	16:05-16:40	陈新河	“线上、线下课堂+数据竞赛”三位一体—数据科学家培养新模式	
	16:40-17:15	杨锐	保险大数据平台建设	

12 月 9 日 科学会堂二楼报告厅

会场名称	时间	姓名	报告标题	主持人	
R 语言应用会场	09:00-09:30	谢佳标	利用 RMarkdown 快速实现定制化报表	黄俊文	
	09:30-10:00	张杰	R 语言之数据分布信息可视化		
	10:00-10:30	詹欣谕	R 语言使用者运用 Shiny 让服务更智能		
	Break				
	10:50-11:20	俞钟行	对在 AT&T/朗讯科技公司大数据环境下作统计分析工作的回顾		
	11:20-11:50	姚树亮	小项目取代日常重复工作(爬虫+COM 接口)		
午餐					
数据应用会场	14:00-14:40	金江	R 语言在制造行业与商业大数据平台集成应用案例	夏丰盛	
	14:40-15:20	黎建辉	OTA 酒店订单审核工作量预测		
	Break				
	15:40-16:10	曾加	基于 ros 实现无人驾驶小车		
	16:10-16:50	钱凯	OTA 违规酒店识别		

12 月 9 日 文史楼 303

会场名称	时间	姓名	报告标题	主持人	
统计科学会场	09:00-09:30	张四海	Classification of Regression Coefficients in Dynamic Panel Data Models	陈雨晴	
	09:30-10:00	车金星	随机信息度量下的变量选择集成方法		
	10:00-10:30	严晓东	Covariate-specified group structure recovery for high-dimensional regression		
	Break				
	10:50-11:20	成勤和	Conducting Meta-analysis under Confidence Distribution Framework Using gmeta in R		
	11:20-11:50	王庆勇	Data-driven analytics for video QoE management in the large scale mobile networks		
午餐					
生物信息与量化金融会场	14:00-14:35	郑小琪	基于 InfiniumPurify 包的肿瘤纯度估计和差异甲基化分析	沈佳瑾	
	14:35-15:10	李钧涛	基于自适应稀疏群 lasso 的生物信息挖掘		
	15:10-15:45	张伟伟	Accounting for tumor purity improves cancer subtype classification from DNA methylation data		
	Break				
	16:05-16:40	郭屹峰	基于强化学习的稳健对数最优策略理论研究		
	16:40-17:15	蔡艳丽	基于 HMM 文本挖掘的系统性金融风险度量研究		

3. 会议机构

主办单位：



华东师范大学统计学院



华东师范大学数据科学与工程学院



华东师范大学教育信息技术学系



统计之都 (<https://cosx.org>)

赞助单位：



RStudio



卓铭保险



人民邮电出版社

4. 会议组委会

组委会老师：林祯舜 汤银才 郎大为 高明 吴永和

主席：包亚杰

副主席：田雅慧

秘书长：付英男

组委会成员：沈佳瑾 毛悦 朱祺伟 程临峰 杨家麟 陈雨晴 文茜 孙嘉怡 刘嘉鑫
任怡萌 苏瑾 刘文丽 杨婉敏 程曼莉 李敏 刘思懿 陈子浩 陈佳恒 贾文鑫 赵彩云
况巧云 李浩淼 李蓉蓉 周世荣 李航 程歌星 王佳雯 陈雷慧 匡俊 李娜
陈远哲 朱仁煜 艾丽斯

志愿者：孟必莹 周龙羽曦 刘宇婷 罗冰莲 祁馨禾 蔡颖异 丁哲琪 张锦琳 余思敏
乃米热 吕紫瑄 尧紫琪 赖婷荷 时文怡 焦傲 章姝姝 邢译文 贺子益 徐瑾
谭静 詹晨 曾怡安 晋铭 杨雪柯 张子月 朱林染 付晓裕 黄婧 谭昕玥 何田田
唐嘉琪 张慧然 陈雨萌 何玉洁 何姣 雷美英

R语言商业实践-从金融市场到区块链

张丹^{1,*}

¹ 北京青萌数海科技

摘要

R 语言是一门统计语言，强大、易用，而且应用场景及其广泛。从金融市场量化投资到区块链交易，有很多的相似性的规则，相似的行为模式。在金融市场，我们现代金融学的理论基础，在框架下我们研究市场做数据分析，通过金融市场的金融资产交易行为，体现数据的价值。同样，在区块链的新型市场，我们通过也是通过分析市场的行为，发现交易价值。R 语言提供了非常强大的工具，只要我们能想到，就能动手就实践。通过场景落地，希望能给大家认识到R语言不仅是语言，还是我们探索世界的工具。

*张丹，资深R语言技术专家，是国内R语言技术社区的领军人物。拥有10年以上互联网应用架构经验，在R、Java、NodeJS、大数据、统计、数据挖掘算法等方面有深厚的积累。金融大数据专家，精通量化投资交易策略，熟悉中国金融二级市场、交易规则和投研体系。熟悉数据学科方法论，在外汇领域和区块链领域做落地的尝试。获得10项SUN及IBM技术认证，微软MVP。多次在互联网技术、数据科学相关技术大会中担任演讲嘉宾。著有《R的极客理想：量化投资篇》、《R的极客理想：工具篇》、《R的极客理想：高级开发篇》，英文版图书被CRC出版集团引进，在美国发行。个人博客：<http://fens.me> (Alexa全球排名70k。)

数据科学如何助力在线教育革命

刘思喆^{1,*}

¹51Talk

摘要

在2018上半年平均每天有0.81个教育项目获得融资，各类人才在资本催化下加速入场。各个公司的教研、技术、产品、数据等不同层面均有长足的变化。本次报告将以K12英语教育为例，探讨数据科学如何在商业目标和教育情怀间找到最优平衡，以及过程中涉及的业务形态和创新性数据技术。

*51Talk (COE) 首席数据科学家，负责流程算法优化变现、数据平台建设以及数据分析相关团队管理及技术指导工作。在加入51Talk前，京东推荐平台部高级经理，在《京东技术解密》一书中，被称为京东15位技术牛人之一。同时他还是中国人民大学大数据分析实验班、首经贸信息学院校外硕士生导师。国内R语言的早期推广者，15年的使用经验，《153分钟学会R》的作者，《R核心技术手册》的译者。

个人数据的加密运算

夏虞斌^{1,*}

¹ 上海交通大学

摘要

个人数据蕴含极大价值，然而目前还没有合理且合法的方式让普通开发者也有能力接触到大量用户数据产生价值。在这次报告中，我们提出了貔貅OS，允许用户对自己的数据进行全方位细粒度控制，同时支持第三方算法安全运行，从而允许任何人以较低成本合法获取大规模的用户真实数据。

* 上海交通大学副教授

Educational Big data and Machine Learning, Case Study from Shanghai

谢军^{1,*}

¹携隽数据

摘要

At Minhang at Shanghai, large scale educational examination data and survey have been collected and stored for more than 180,000 students from 204 schools, for more than 8 years. The data are integrated as an educational data warehouse: a big data set plays as fundamental infrastructure in the educational administration, research and service. The data warehouse is currently built on MS SQLServer, however, open sourced parallel SQL engine such as Impala on Hadoop would be the next generation infrastructure.

With the huge and highly dimensional data, we are able to go inside. For example, with IRT modeling, we are able to estimate the latent for students not only simple scores, include subject latent and latent for a specified intelligence for every student.

Machine learning such clustering, decision tree and neural network are applied. We model the scores or latent with factors collected from the joined survey by means of decision tree. Our aim to find out the key factors that influent the learning and teaching, which is key to improve further work. Multidimensional clustering helps us to link learning difficulty or problems with a group student.

*谢军博士，本科复旦、博士牛津大学。一生致力于数据相关事业。中国第一个电信数据仓库（1996年山东邮电局）第一个银行数据仓库（2001年工商银行浙江分行），中国移动第一个CRM（1998 延吉移动）、中国电信第一个CRM（1997 鞍山电信）的主设计师、架构师。中国工商银行法人信贷风险模型的主师。IBM大中华区CRM首席咨询师，若干个上市公司 C T O。近10年来致力于数据驱动的现代教育学探索，44界全球教育评估联合会分会场主席。几十年来始终在工作一线，编码高手，至今编码并乐此不疲，目前技术兴趣是云计算，业务兴趣是现代教育学。

Data Analysis for Basketball Tactics

王昱舜^{1,*}

¹台湾交通大学

摘要

Analyzing players' performance and behaviors based on statistical and historical data is becoming an effective way for coaches in National Basketball Association (NBA) to develop winning strategies. For example, the analysis can help determine the line-up as well as match-ups in a game and choose the offensive and defensive strategies against the opposing team. It also can help players identify their weaknesses so as to polish their skills. In this talk, I will share some experience in analyzing basketball data and the findings from the data.

*Yu-Shuen Wang (王昱舜) is an associate professor of the Department of Computer Science at National Chiao-Tung University. He received his PhD degree from Visual System Laboratory, National Cheng Kung University, Tainan, Taiwan, ROC, in 2010. Currently, he leads the Computer Graphics and Visualization Lab at the Institute of Multimedia Engineering. Prof. Wang's research interests include Computer Graphics, Data Visualization, and Human Computer Interface.

智能制造导入实务

谢宗震^{1,*}

¹DSP智库驱动

摘要

化工产业同其他制造业，早已在上个世纪70年代导入DCS系统，实现生产即时监控与 rule-based 自动控制，作为发展百年的工程学科，知识体系也相对完整。然而，在多耦合复杂性制程系统中，传统的「理论模型」难以解析，需要依靠AI方法建置「数据模型」以达成节能减碳、降耗增效之目标。本次演讲将分享智库驱动如何基于AI精神发展出关键技术与系统，实现价值转兑的解决方案。

*Johnson Hsieh，DSP智库驱动知识长、清华统计博士、行政院青年咨询委员。提供企业策略与资料分析顾问服务，客户来自多种产业，包括制造、能源、电信、金融、媒体内容以及学研法人机构等。在工作之余主持《D4SG 资料英雄计画》，运用科技和资料的力量改造社会。

AI on Device 精准脱贫

陈洁宁^{1,*}

¹R-Laides Taipei

摘要

以「资料力做公益」j计划在社会救助网及脱贫议题下，如何运用R来处理相关议题，并教育社会救助网相关工作者使用资料及运用资料。内容涵盖R Shiny以及其他尝试过的工具，Azure Machine Learning & Power BI。

*Ning Chen是台北R-Laides的创办人及营运人，她擅长网站、产业分析及社群经营，并相信资料可以让世界变得不一样。她尝试将时间及精力花在创造一个对于初学者及专家都友善的协作环境。

基于R语言的网络文学评论挖掘

钱亦欣^{1,*}

¹ 上海长江时代众创空间数字技术有限公司

摘要

评论挖掘是文本分析的重要应用之一，常见的评论挖掘主要集中在电商等行业。而网络文学的评论由于数据生成场景有一定特殊性分析方式与商品评论有所不同。通过对网络文学评论展开挖掘，能了解读者的观点，对于作者创作，平台运营都有所帮助。

*上海大学经济学院统计学硕士，上海长江众创一鱼数据项目数据科学家，Hadley Wickham的忠实信徒。自2013年起开始使用R语言进行统计分析与数据可视化等工作，研究方向为文本挖掘、贝叶斯分析等，参与过中国房地产司法拍卖指数编制等项目。于图灵社区、雪晴数据网等社区翻译并创作R语言、数据挖掘等相关文章数十篇。开设有个人知乎专栏《数据科学译文系列》

文本挖掘在商业分析中的应用

Marco Li^{1,*}

¹ 安索帕集团

摘要

在社交媒体中，每时每刻都有消费者讨论内容产生，这些数据对于品牌主了解消费者关注的话题，对品牌的态度，对产品的使用体验都至关重要，本次分享讲述了自然语言处理如何在商业分析领域进行文本数据的挖掘应用，使其洞察结果转化为能够支持品牌传播策略的重要依据。

*Marco Li长期从事于市场研究、商业咨询领域，是安索帕中国集团的资深数据总监，带领品牌商务技术部门，从事数据挖掘与建模，数字生态系统解决方案以及人工智能的商业应用方面的工作，Marco带领的团队服务了涵盖母婴、汽车、快消、金融等行业的30多个客户。

数据如何助力人职匹配？

季雨清^{1,*}

¹Seedlink Technology

摘要

将以一个实际应用为例，探讨在解决人职、人企匹配问题上，能够怎样使用数据，效果如何，以及有哪些有待解决的问题。

*学过一些心理学，玩过一些数据，能写几行代码。

如何做一个成功的商业对话机器人

李翛然^{1,*}

¹深圳奇点信息技术有限公司

摘要

机器人行业当中，对话系统作为人机交互的必备入口已经成为业内的普遍共识。但是，随着近两年的AI投资及市场情况冷却，如何在机器人行业中寻求技术与客户发展的平衡点已经成为各AI公司必须解决的现实问题。本次演讲将会从聊天机器人的核心技术应用及软件架构入手，为听众带来一次另一个角度的机器人工业及应用方式解读。用户画像等，还能为企业带来更大的营销价值及销售转化机会等。”

*李翛然，于利兹大学金融数学毕业。先后从事过保险精算，投资银行工作。于2014年创办奇点信息技术有限公司，为各大机构提供智能化管理系统与机器人业务。现已有10余家金融机构、医院、教育系统机构采用其提供的智能服务为行业助力。其股票投资机器人股神SAI于2017年获得KPMG中国金融科技双创大赛TOP30奖项。

AI时代的智能客服

Evan lai^{1,*}

¹深圳追一科技有限公司

摘要

为了帮助企业真正构建智能服务平台，追一科技基于深度学习和自然语言处理等前沿AI技术，从智能化交互切入，并与企业业务系统融合，从而形成一体化智能服务解决方案，覆盖服务、营销和数据挖掘等应用场景。在AIForce的平台架构下，对话机器人、知识库运营、坐席辅助、智能分析等产品及服务能够形成有效的闭环，不断驱动智能化水平提升，形成数字化的企业大脑。经过充分训练和运营，AI意图识别准确率可以达到95%以上，有效辅助人工提升服务质量与效率，创新用户体验，帮助企业实现降本提效。更为重要的是，通过智能化数据分析并结合用户画像等，还能为企业带来更大的营销价值及销售转化机会等。

*Evan Lai 赖贊，追一科技AI商业与战略总监，专注金融科技领域。Evan是追一科技AI商业与战略总监，负责追一科技人工智能技术在垂直行业领域的战略规划，咨询，拓展和应用，目前专注金融科技领域。在加入追一科技之前，Evan在金融行业和高科技领域拥有10余年的丰富实操经验，曾先后在全球知名软件公司EMC, Oracle担任中国区研发项目负责人。在加入追一科技之前，Evan也曾任职于点融网和宽资本，主要负责兼并，收购，投资和资产出售，领导并推动了近亿元项目的实施。早前，Evan还曾任职于国内某知名AI图像识别领域独角兽企业，为金融科技事业部负责人，先后主导并推动农行，建行，人民银行等多家国内知名银行人脸识别项目上线推广；毕业于北京航空航天大学，拥有软件工程学位，并在英国爱丁堡大学获得计算机硕士学位，同时拥有金融领域证券、基金及人力资源，项目管理等多个认证和资质。

在浏览器里深度学习 - 使用TensorFlow.js构建人工智能应用

尹志^{1,*}

¹宁波工程学院

摘要

深度学习技术已然在各类场景大放异彩。而深度学习工程师们的日常却是面对海量的数据、恼人的训练时间。那么，在守候模型训练时你有没有想过，利用你的模型构建一个Web应用呢？对，就让你的深度学习模型跑在浏览器里！本报告将简要介绍TensorFlow.js框架。内容包括TensorFlow.js的核心概念、如何利用浏览器进行模型训练、常规的深度学习模型如何导入TensorFlow.js、Mobilenet模型等。我们会演示如何利用TensorFlow.js实现简单的基于Web的人工智能应用。

*尹志，浙江大学物理学博士，现就职于宁波工程学院理学院。云朵网络首席数据科学家。水过机器学习论文，做过数据挖掘项目，打过数据科学比赛。研究方向集中在推荐系统、文本挖掘、医学影像等领域，对解决各类数据科学相关的实际问题尤感兴趣。

AlphaGomoku：一个基于AlphaGo的五子棋人工 智能

付星宇^{1,*}

¹ 广州似然科技有限公司

摘要

- AlphaGo的发展历史。
- AlphaGo的inference structure。
- 我们是如何将AlphaGo算法应用到五子棋游戏上的（对AlphaGo的创新）。
- 和现场观众人机对战

*职业经历：似然实验室联合创始人。负责：独立研究项目负责人，实习生项目管理，机器学习研修班负责人。公司官网：<http://www.maxlikelihood.cn/> 某初创量化投资科技公司联合创始人。负责：CTA策略开发，机器学习策略开发，交易系统的设计和实现。教育背景：中山大学数学与应用数学大四在读。GPA：3.9/4.0；连续三年获得校级一等优秀奖学金。本科期间前往加州大学伯克利分校交换，全部课程满分。GPA：4.0/4.0。2017国际基因工程机器大赛(IGEM)，全球软件组第一。科研项目经历：- AlphaGomoku: an AlphaGo-based Gomoku artificial intelligence using curriculum learning - A Machine learning framework for stock selection - Robust log optimal strategy with reinforcement learning

目标检测技术在携程图像智能化中的实践

李翔^{1,*}

¹ 携程

摘要

携程作为OTA行业的领跑者，拥有全球百万家酒店数以亿计的酒店图像。当前携程的图像智能化系统，已实现海量图像的智能化审核、处理、识别和应用，从而降低人工干预、提升用户体验、增加订单售卖。本次报告将围绕其中的图像目标检测技术展开，介绍一系列基于深度学习的目标检测算法，并分享其在携程酒店图像智能化中的多个具体应用和实践经验。

*为识别、度量学习、迁移学习和深度学习。在ICCV和CVPR等国际顶级学术会议及国际权威期刊上发表多篇论文。

MXNet Graph Optimization and Quantization based on Intel MKL-DNN

陈新宇^{1,*}

¹华东师范大学

摘要

This presentation mainly talks about the efficient graph optimization techniques and 8-bit low precision inference of Apache MXNet (incubating) based on Intel Math Kernel Library for Deep Neural Networks (Intel MKL-DNN). While convolutional neural networks (CNN) shows state-of-the-art accuracy for wide range of computer vision tasks, it still faces challenges during industrial deployment due to its high computational complexity of inference. Low precision is one of the key techniques being actively studied recently to conquer the problem. With hardware acceleration support, low precision inference can execute more operations per second, reduce the memory bottlenecks, permit better cache usage, and deliver higher throughput and lower latency for workload.

*I am a graduate student at ECNU pursing the major of Statistics. Specialties: high performance computing, high level algorithmic optimization and low level hardware-specific hot spots tuning, vectorization, code parallelization, deep learning frameworks optimizations (Caffe, Keras, Theano, PyTorch, TensorFlow, MXNet, R).

跨行业数据融合应用案例分享

张恺^{1,*}

¹ ucloud

摘要

跨行业数据融合应用案例分享随着大数据在各个行业的应用趋于成熟，基于企业内部用户日志数据或业务系统数据产生的用户画像，个性化推荐，精准营销等数据产品已成为中大型企业的“标配”。如何让用户画像维度更多？如何使个性化推荐效果更好？如何把精准营销做得更精准？一个可行的方案就是走出企业，仰仗同行业或跨行业的各类数据进行融合，弥补企业自身数据的局限性。在分享中，会介绍到诸如金融风控，跨行业交叉营销等需要多种数据融合应用的典型场景案例。

*现任ucloud数据平台部产品和数据运营负责人。毕业于上海交通大学，获项目管理硕士学位。曾在金融、房地产、旅游、汽车、电商等多个行业龙头企业中从事数据相关工作十年以上。从数据采集到数据存储，再到数据建模和应用都有丰富的经验。

如何让车联网隐私数据安全的流通和使用

张翔^{1,*}

¹ 车轮互联

摘要

大数据行业发展良久，已经从开放的匿名数据，渗透到敏感的隐私数据，从单一数据源走向多源数据融合。伴随着数据保护法规的完善，如何安全的让数据流通起来，产生更多的价值，同时又保护数据所有者的隐私和分享利益的权力？我们在汽车行业开源了carro.io项目，通过加密技术保护隐私，通过区块链认证权益，通过零知识证明验证交易，通过IPFS实现完全去中心化的存储，最后通过安全计算平台实现跨数据源的交叉建模。类似方案其实不止在汽车领域，在基因，医疗，人工智能等等多个领域都有尝试，2018年可以说是数据安全流通领域的元年。

*车轮互联副总裁，10年COS老水友，也是集信+ jx.plus 信用区块链创始人，《重构区块链》bcrb.io 的作者

基于区块链与TEE技术的数据隐私保护

余炀^{1,*}

¹TEEX

摘要

随着大数据技术的蓬勃发展，数据本身的价值得到了越来越多的关注。一方面，企业或个人用户拥有大量的数据交易与线上数据分析的需求；另一方面，数据流通过程中的隐私性变得至关重要。本次报告将介绍如何通过区块链与TEE（可信执行环境）技术保护数据流通过程中的隐私安全。

*余炀，复旦大学计算机博士，TEEX联合创始人，在体系结构、系统安全及虚拟化领域具有深厚的积累。TEEX致力于将区块链与TEE（可信执行环境）技术相结合，打造一个可信计算平台，保护数据和执行过程的隐私安全。

大规模数据中心的性能分析

郭健美^{1,*}

¹阿里巴巴集团

摘要

数据中心已成为支撑大规模互联网服务的标准基础设施。随着数据中心的规模越来越大，数据中心里每一次软件或硬件的升级改造都会带来高昂的成本。合理的性能分析有助于数据中心的优化升级和成本节约，而错误的分析可能误导决策、甚至造成巨大的成本损耗。本报告介绍大规模数据中心性能监控与分析的挑战与实践。

*郭健美，阿里巴巴集团系统软件事业部高级技术专家，目前主要从事大规模数据中心的性能分析和软硬件结合的性能优化。中国计算机学会软件工程专委会委员。曾主持国家自然科学基金面上项目、入选上海市浦江人才计划（A类）、获得ACM SIGSOFT “杰出论文奖”。担任ICSE’18 NIER、ASE’18、FSE’19等重要会议程序委员会委员。

数据分析挖掘在制造领域的应用价值

韩俊仙^{1,*}

¹ 北京桑兰特科技有限公司

摘要

工业互联网及智能制造是国家战略，如何让数据分析挖掘在制造企业发挥作用？如何从机理模型到数学模型，帮助企业创造价值？应该是数据科学家思考的问题。

*高级工程师、六西格玛黑带大师，第一届全国六西格玛管理推进委员会专家委员。日本质量工程（田口方法）研究会国际会员、北京大学数学科学学院兼职教授、硕导，北京工业大学数理学院兼职教授、硕导，阿里云全球MVP（最有价值专家）。韩俊仙在企业从事质量管理、现场质量控制、优化、改进等工作多年，具有丰富的制造业现场工作和指导经验，在DOE、参数设计、SPC等方面具有很深的造诣。为华为、京东方、航天科工、上海电气、金风科技、美的、格力、方太、老板、苏泊尔等六十余家企业提供技术服务。

数据治理与数据资产管理

孙繁荣^{1,*}

¹ 上海长江时代众创空间数字技术有限公司

摘要

从企业数据治理角度阐述在数据治理和数据资产管理过程中面临的问题和挑战，数据治理框架体系和数据资产管理相关的架构、流程和运营，以及在实施过程中的最佳实践。

* 上海长江时代众创空间数字技术有限公司CTO, 大数据创新实验室主任。大数据技术专家，曾任富士康、毕博GDC、惠普等知名企业研发经理、高级架构师、及产品经理，20年以上企业级关键信息系统建设经验，主导研发云计算SaaS应用、云存储产品、大型MPS（主生产计划系统），MES（生产执行系统）系统、金融行业解决方案、数据资产管理系统等。

基于R构建某工厂质量在线监控平台

刘心广^{1,*}

¹ 上海质瑞信息科技

摘要

某自动化工厂在产品生产和组装过程中，由于多工位多类型自动化设备的引入，数据实时产生，但类型多样、各自孤立。本项目即基于工厂内部定制化的分层级质量监控需求，通过数据的互联互通、关联分析和质量参数在线监测与报警系统，使用R语言与数据解析融合、API和web应用相结合，通过私有云部署，实现工厂内部自定义的质量监控和报表系统。

*刘心广，中国科学院博士，可靠性工程高级工程师职称，持有高校教师资格证，产品可靠性性能高级检测员，IEEE可靠性组会员，美国质量协会ASQ高级会员，ASQ认证可靠性工程师(CRE)、质量工程师(CQE)、六西格玛黑带(CSSBB)。熟悉产品质量可靠性工程技术和方法论在产品开发中的应用，擅长数据分析、模型构建和软件二次开发，通过大数据采集、清洗、存储、分析计算和应用的一体化技术，推动提升工业质量大数据的分析和应用，实现数据的信息化和价值化。目前负责工厂智能制造中质量数字化集成解决方案开发。

“线上、线下课堂+数据竞赛”三位一体—数据科学家培养新模式

陈新河^{1,*}

¹ 中关村大数据产业联盟

摘要

1. 工业界需要啥样的数据科学家；
2. 数据科学课程体系如何设立；
3. “线上、线下课堂+数据竞赛”三位一体的数据科学家培养模式；
4. 某股份银行采用三位一体的数据科学家培养模式复盘；
5. 三位一体的数据科学家培养模式不足和待改进的地方；
6. 对高校培养数据科学专业的建议。

*陈新河，中关村大数据产业联盟副秘书长、DT 大数据产业创新研究院（DTiii）院长、聚合数据独立董事；曾任工业和信息化部电子科技情报研究所副主任。在 IT 领域 20 多年的研究、观察和思考，同样的数据，不同的观点。参加国务院颁布《促进大数据发展行动纲要》文件编制，主持国家发展和改革委员会“十三五”规划前期研究重大课题-《“十三五”信息经济发展研究》，社科基金特别重大课题《大数据治国战略研究》核心成员，《关于全面实施“大数据治国”战略的建议》获多位党和国家领导人重要批示，成功竞标获得《北京市软件和信息服务业“十二五”发展规划》编制工作，2004 年主持课题《未来 5-15 年电子信息技术发展趋势分析》获部级奖励。IT 思想贡献：互联网是以人均 GDP 为基数的产业，移动互联网是以人口数为基数的产业。

保险大数据平台建设

杨锐^{1,*}

¹ 西安财经学院

摘要

大数据时代的到来，繁琐庞大的数据量对保险业大数据带来了极大的挑战。保险业现有的传统数据处理模式已经不能应对处理现有的数据量了，保险业的数据分结构化和非结构化，如何能在大数据环境下得到保险业想要的数据处理结果，建设保险业大数据平台显得尤为重要。保险业大数据平台建设是利用现代智能技术手段对保险数据进行系统分析，发现有价值的信息，并在数据的支持下做出正确的决策，进一步提高保险业的核心竞争力。

*我是一名全日制的研二在校学生，本科学习计算机科学与技术专业，硕士学习应用统计学。在校期间曾获得专业奖学金二、三等奖，优秀三好学生等荣誉。读研期间多次参加学术会议，论文被保险学会和全国商业统计学会录用。在学校积极参加活动，培养了我较强的组织能力和较强的责任心。在课余时间充实自我，喜欢游泳，完善各个方面的能力。

利用RMarkdown快速实现定制化报表

谢佳标^{1,*}

¹跨越速运

摘要

近年来R Markdown已经逐步演变成制作自动化报表和文档相对完整的生态系统。本演讲将分享如何快速定制化个性主题，添加GIF动图、HTML Widgets和shiny等技巧，实现报表交互；并详细介绍dashboard和Slidy的有趣功能，制作专属你的个性化报告。

*跨越速运数据挖掘专家。2017-2018 微软MVP 资深R语言用户，十余年数据挖掘工作实战经验，多次在中国R语言大会上作主题演讲。著作书籍有《R语言游戏数据分析与挖掘》、《R语言与数据挖掘》、《数据实践之美》。

R语言之数据分布信息可视化

张杰^{1,*}

¹香港理工大学

摘要 本次演讲主要介绍并对比十多种不同展示数据分布情况的图表，包括带统计直方图、核密度估计曲线图、误差线的柱形图和散点图、箱型图、小提琴图、瓶状图、豆状图、条带图、云雨图、海盗图等，尤其对箱型图的不同应用作详细介绍。同时对二维统计直方图和核密度估计图也作对比介绍。

*香港理工大学Research Assistant; Excel教程《Excel数据之美》作者; Excel图表插件EasyCharts开发者; 著有十余篇一作的SCI(E)论文; 微信公众号EasyCharts联合创始人; 2018年第十一届中国R会议（北京）数据可视化专场演讲嘉宾; 学术研究方向为颜色科学、机器视觉与深度学习; 预计2018年下半年出版《R语言数据可视化之美》。

R语言使用者运用Shiny让服务更智能

詹欣谕^{1,*}

¹R-Ladies Taipei

摘要

Shiny是R语言使用者轻松开发交互 Web 应用的R包，不需要用JavaScript用几行代码就可以构建 Web 应用程序，本次演讲将分享实际应用Shiny实现界面布置与搭建。内容包含 2018年台北R-Ladies Kaggle竞赛看板、智能推荐脱贫用户、Azure Machine Learning with shiny。

*Kristen Chan，台北R-Ladies的营运人、WaveIn伸波通讯资料科学家。擅长电商及通讯领域，喜欢运用资料的力量解决问题。

对在AT&T/朗讯科技公司大数据环境下作统计分析工作的回顾

俞钟行^{1,*}

¹ 上海思科统计质量咨询服务有限公司

摘要

对20多年前，在AT&T/朗讯科技公司大数据环境下作统计分析工作时，在当时无任何中文资料情况下，用S-PLUS软件自己编程作统计分析，解决诸多实际工作问题的回顾。

*俞钟行，上海思科统计质量咨询服务有限公司经理。

小项目取代日常重复工作（爬虫+COM接口）

姚树亮^{1,*}

¹礼来苏州制药

摘要

每一个完整功能的实现，都需要综合各方面的知识构成一个完整项目。本次演讲内容主要实现的作用是：自己定期会写一份报告，并需要批准，批准后，告诉一些指定的人。这个需求转化为R语言编程主要涉及：爬虫，数据处理和逻辑判定，通过COM调用本地Outlook发送邮件给指定的人，计划任务。将以上过程通过编程实现，使用计划任务，每天运行一次，检查有无新文件批准，如果已批准，则发邮件给指定的人，如果没批准，或批准了，但是发送过邮件了，则不发送。

*姚树亮，制药行业的小实验员一枚，爱好统计分析。

R语言在制造行业与商业大数据平台集成应用案例

金江^{1,*}

¹SAP

摘要

本环节将介绍制造行业的生产制造和售后交付场景中，利用R语言强大的数据科学能力，结合商用大数据平台实时计算能力，帮助提升产品良率和产品可靠性的案例，解决真实商业管理领域的痛点问题。

*金江目前在SAP大中华区的客户创新及企业平台团队担当大数据业务架构师，负责业务拓展工作，拥有超过16年企业信息化建设的咨询经验。目前负责离散制造与高科技行业的大数据业务扩展，致力于为中国客户提供SAP大数据及物联网等领域的先进技术，结合国内外最新的案例和实践，用信息化手段给中国企业带来更多降本增效的业务推进动力。

OTA酒店订单审核工作量预测

黎建辉^{1,*}

¹ 携程

摘要

OTA酒店订单审核工作量预测项目通过对审核人员所属三级组别的日、周、月时间维度的工作量回归预测，同时考虑节假日、星期等时间因素。将月维度的预测结果用于审核人员的人员招聘储备和审核工作量的趋势分析；日和周维度的预测结果，根据审核人员过去两周日均审核工作量，换算成所需审核人员数，用于审核工作量的动态合理排班，从而提升酒店订单的审核及时性，降低担保订单投诉比例，提升用户的酒店全流程体验和OTA企业的品牌形象。

*黎建辉自2015年6月加入携程旅行网，在大住宿事业部酒店数据智能团队担任高级数据分析经理。主要致力于用机器学习，文本挖掘，图像处理的方法解决携程酒店服务的业务问题。

基于ros实现无人驾驶小车

曾加^{1,*}

¹浙江大学

摘要

ROS(Robot Operating System) 是一个开源的机器人控制软件平台，随着机器人行业的发展而在研究人员、创业公司和大型公司中日益普及。而Turtlebot3是由ROS官方打造的一个优秀的软硬件学习平台。本次演讲将介绍ros用于无人驾驶研究的优势，以及如何利用Turtlebot3实现一辆具有地图绘制、障碍检测、定点导航等功能的小车。

*Tephra Lab是主要由几位浙江大学在读研究生组成的小团队，在课余时间热衷于研究一些创意小项目，以及参与统计之都线下沙龙、文章翻译等活动。

OTA违规酒店识别

钱凯^{1,*}

¹ 携程

摘要

OTA违规酒店识别项目通过计算每两家酒店之间的图片和酒店名称的相似度，找到疑似在OTA上一店多开的酒店。项目中，根据酒店的经纬度，对距离在1km范围内的酒店进行两两匹配，计算匹配到的每一对酒店之间的大堂、外观图片相似度，得到每一对酒店之间相似图片的张数，以及计算每一对酒店名称之间的相似度，综合考虑以上两点，识别出疑似一店多开的酒店，交给业务人员核查，对确实存在一店多开违规行为的酒店进行相应的惩罚。项目目的是为了减少酒店一店多开骗取OTA优质流量和影响客人入住体验的行为，同时维护OAT企业的品牌形象。

*携程旅行网，酒店数据智能部，高级数据分析师。构建机器学习模型，降低用户预订酒店过程中遇到的服务缺陷，提升订单产量，提升用户预订体验。

Classification of Regression Coefficients in Dynamic Panel Data Models

张四海^{1,*}

¹ 上海师范大学

摘要

This paper proposes a new method for classification of regression coefficients in dynamic panel data models. The regression coefficients are assumed to be heterogeneous across groups but homogenous within a group, however, neither the number of groups nor the members of each group are known. Some statistical procedures have been developed for classification of regression coefficients. Moreover, the resulted estimators for the number of groups and the classification can be proved to be consistent under some mild conditions. Monte Carlo simulation study shows that the new classification method has desired finite sample properties. A real data application is carried out for illustration.

*张四海，中共党员，硕士研究生，现就读于上海师范大学数理学院。

随机信息度量下的变量选择集成方法

车金星^{1,*}

¹南昌工程学院

摘要

大数据时代给我们带来了无论变量维数还是数据条数上都空前巨大的大数据集。这为统计机器学习领域提供了必需的数据保障。然而，面对这些大数据集，如何从中挖掘出关键的数据，如何有效地选择信息变量，如何有效地选择关键数据，如何利用选择的关键数据来推断事物的未来发展，成为了一个至关重要的研究课题。报告人近年来围绕变量选择和数据预测等问题通过模型建立、算法设计和理论分析等方面进行了系统性的研究，并将相关算法应用于数值模拟数据集和工程领域的一些真实数据集。在本报告中，报告人及合作者围绕变量选择问题，定义了三类变量，主要讨论如何就数据选择、模型选择、变量选择扰动下的集成学习问题。在理论上，给出了相关性度量定理、收敛性定理和三类变量选择性能定理。在数值模拟实验中，这一算法取得了很好的效果，并进行大样本和实际案例实验。通过模型对比试验，验证了该模型的优越性和有效性。这一工作为大样本学习的预处理提供了理论奠基。

*车金星，男，副教授，硕士生导师，南昌工程学院第五届“十佳青年教师”、“瑶湖杰青”特聘岗位，中国人工智能协会(CAAI)会员、中国国家科技专家库成员、国家自然科学基金项目评审专家、江西省科技奖（自然科学）三等奖第一完成人。研究方向为统计机器学习与数据挖掘，以及在网优、通讯定位，以及电力负荷、电价、风能等能源生产及需求的统计预测上的应用；预测理论与方法；特征选择。IEEE Transactions on Neural Networks and Learning Systems、Journal of Applied Statistics 等 20 余个国际著名期刊的审稿人；发表学术论文 30 余篇，含 SCI 一区 4 篇、二区论文 6 篇；主持、参与国家级、省级课题多项，主持完成国家青年基金、省青年基金（结题评定为优）各一项。指导学生参加全国数学建模竞赛、并获得国家级省级奖项多项，指导的“三下乡”团队 2 次获得省级优秀服务队，主持完成通信类数据、教育类数据等多个横向项目。

Covariate-specified group structure recovery for high-dimensional regression

严晓东^{1,*}

¹ 山东大学

摘要

This paper studies integrative analysis of multiple units in the context of high-dimensional linear regression. We consider the case where a fraction of the covariates pose different effects on the responses across various units, e.g., some covariate-specific coefficients are the same for all the units, while others have a grouping structure. We propose a double penalized least squares approach by combining quadratic loss function with a fusion penalty term to penalize the difference between any two units' coefficients of the same covariate for identifying latent grouping structure, as well as a sparsity penalty to detect nonzero effects. Without the need of knowing the grouping structure of every variable among the data units and the sparsity construction within the variables, the proposed double penalized procedure can automatically recover sorts of structures of covariate-specific effects such as heterogeneous, homogeneous and sparsity, and estimate the parameters simultaneously. We proceed the alternating direction method of multipliers algorithm (ADMM) through effectively utilizing the storage and reading of the datasets, and demonstrate convergence of the proposed procedure. We show that the proposed estimator enjoys the oracle property in recovering the underlying covariate-specific structure of heterogeneous, homogeneous and sparsity. Simulation studies demonstrate the good performance of the new method with finite samples, and a real data example is provided for illustration.

*严晓东(Xiaodong Yan)，山东大学经济学院副研究员(Associate Professor at School of Economics, Shandong University)。
研究方向(Research Interests): 计量经济(Econometrics)、计量金融(Quantitative Finance)、风险管理(Risk Management)、
大数据分析(Big Data Analysis)、学习经历(Studying Experience)。

Conducting Meta-analysis under Confidence Distribution Framework Using gmeta in R

Jerry Cheng^{1,*}

¹Rutgers University

摘要

A variety of meta-analysis methods have been developed under the framework of combining confidence distributions. Under this framework, traditional approaches, such as p-value combination, fixed-effects model, and random effects models, are subsumed. More importantly, innovative meta-analysis methods are developed. The examples are robust meta-analysis, exact meta-analysis for binary data with rare events in 2 by 2 table, meta-analysis for heterogeneous studies, non-parametric meta-analysis, etc. In this presentation, we discuss these new meta-analysis methods and demonstrate a gmeta R package with numeric examples and graphic output.

*Jerry Cheng is an Assistant Professor of Biostatistics with RWJ Medical School at Rutgers University. His research interest is statistical computing, large scale data analysis, data mining, and survival analysis.

Data-driven analytics for video QoE management in the large scale mobile networks

王庆勇^{1,*}

¹ 国防科技大学

摘要

Video streaming is becoming one of the most popular services over mobile networks. However, it is difficult to ensure quality of user experience (QoE) of video streaming because QoE of video streaming is affected by multiple factors of mobile networks. Many previous research efforts have been made to improve Quality of Service (QoS) of video streaming over mobile networks while QoS improvement does not directly enhance QoE. Essentially, we need to assess QoE from user perspective and identify the relation between QoE and QoS so that we can improve QoE of video streaming. There are few studies on standardizing QoE assessments. One of recent proposals on standardizing QoE of video streaming is video Mean Opinion Score (vMOS), which can model QoE of video streaming in 5 discrete grades. However, there are few studies on quantifying vMOS and investigating the relationship between vMOS and other QoS parameters. In this paper, we address this concern by proposing a novel data analytical framework based on video streaming QoE data. In particular, our analytical model consisting of K-means clustering and logistic regression; this model integrates the benefits of both these two models. Moreover, we conduct extensive experiments on realistic dataset and verify the accuracy of our proposed model. The results show that our proposed framework outperforms other existing data analytical methods in terms of prediction accuracy. Moreover, our results also show that vMOS is essentially affected by many QoS parameters such as initial buffering latency, stalling ratio and stalling times. Our results offer a number of insights in improving QoE of video streaming over mobile networks.

*Qingyong Wang has authored and coauthored over ten journal and conferences, he also is a reviewer for journals and TPC member of conference. His research interests include machine Learning, Large-Scale Distributed Systems and System Performance Evaluation Theory in Big Data Analysis.

基于InfiniumPurify包的肿瘤纯度估计和差异甲基化分析

郑小琪^{1,*}

¹ 上海师范大学

摘要

In this talk, I will present a set of statistical methods for the analysis of DNA methylation microarray data, which account for tumor purity. These methods are an extension of our previously developed method for purity estimation; our updated method is flexible, efficient, and does not require data from reference samples or matched normal controls. We also present a method for incorporating purity information for differential methylation analysis. In addition, we propose a control-free differential methylation calling method when normal controls are not available. Extensive analyses of TCGA data demonstrate that our methods provide accurate results. All methods are implemented in InfiniumPurify.

*郑小琪，上海师范大学教授，博士生导师，主要从事生物统计和生物信息领域的研究。2009年毕业于大连理工大学应用数学系获博士学位，随后至上海师范大学工作，历任讲师、副教授、教授。2012至2014年及2015年7月，两次赴哈佛大学公共卫生学院“生物统计与计算生物学系”进行学术访问。2008年至今累计发表SCI论文50余篇，包括第一或通讯作者论文41篇，其中多篇发表在本领域顶级杂志上。近五年的代表性学术论文包括Genome Biology (IF = 11.908) 三篇, Bioinformatics (IF = 7.307) 两篇, PLoS Computational Biology (IF = 4.542) 一篇等。统计到2018年2月，所发表论文被国内外其他研究学者引用604次，他引论文分布在Cell、Nature、Nature Reviews Genetics 等20多种国际学术刊物上。

基于自适应稀疏群lasso的生物信息挖掘

李钧涛^{1,*}

¹河南师范大学

摘要

针对群lasso惩罚类统计学习方法处理二分类高维生物数据面临的提前变量分群，自适应的群内变量选择，生物可解释性等难题，致力于开展基于网络分析的变量分群策略和新型自适应惩罚机制研究，据此提出了融合网络分析和信息学理论方法的自适应稀疏群lasso。首先，将网络分析中的网络模块识别与群lasso中的变量分群有机联系起来，利用加权基因共表达网络分析方法辨识出具有良好生物交互关系的模块。其次，利用条件交互信息等信息论方法在每一个被划分的群内构建变量重要性的评价准则，据此构造具有生物可解释性的权重系数并将其添加到惩罚项的合适位置来自适应地进行变量选择。最后，借助于wgcna和sgl等R工具包，在四种高维癌症生物数据上验证了所提的自适应稀疏学习机能够有效地进行分类和群体基因选择。

*李钧涛现为中国人工智能学会智能空天系统专业委员会委员，中国自动化学会数据驱动控制、学习与优化专业委员会委员，长期从事统计学习、数据挖掘等方面的研究。近年来主持、参与国家级、省部级重点等项目16项，发表学术论文40余篇，其中SCI, EI检索论文36篇。先后获得“河南高校科技创新人才”、“河南省高校青年骨干教师”等称号。

Accounting for tumor purity improves cancer subtype classification from DNA methylation data

张伟伟^{1,*}

¹东华理工大学

摘要

Tumor sample classification has long been an important task in cancer research. Classifying tumors into different subtypes greatly benefits therapeutic development and facilitates application of precision medicine on patients. In practice, solid tumor tissue samples obtained from clinical settings are always mixtures of cancer and normal cells. Thus, the data obtained from these samples are mixed signals. The “tumor purity”, or the percentage of cancer cells in cancer tissue sample, will bias the clustering results if not properly accounted for. In this paper, we developed a model-based clustering method and an R function which uses DNA methylation microarray data to infer tumor subtypes with the consideration of tumor purity. Simulation studies and the analyses of The Cancer Genome Atlas (TCGA) data demonstrate improved results compared with existing methods.

*张伟伟，2001-2005就读于郑州大学数学系，2005-2007就读于大连理工大学基础数学专业，2014-2017就读于上海师范大学计算生物学专业，2007-至今工作于东华理工大学理学院。以第一作者身份发表论文两篇，主持省级课题三项，参与国家基金三项。

基于强化学习的稳健对数最优策略理论研究

郭屹峰^{1,*}

¹ 中山大学

摘要

通过二阶Taylor 展开逼近传统对数最优策略目标函数(GLOS)，提出新的稳健对数最优策略(RLOS)，成功避免了对分布函数的估计及其误差，极大地减少运算量，并通过不等式放缩证明在一定条件下稳健对数最优策略(RLOS)误差存在上界。再基于强化学习技术提出基于强化学习的稳健对数最优策略(RLOSRL)，利用bootstrap在沪深300 股指进行随机选股回测，在不同时间长度下与其他策略进行对比，成功验证RLOSRL 的收益性和稳健性均优于其他策略。更多细节可阅读研究论文：
<http://arxiv.org/abs/1805.00205>

*中山大学数学学院大四在校生，加州大学伯克利分校交换生，华南统计科学中心成员，曾于中国人寿、朝旭投资等公司实习从事投资组合开发。

基于HMM文本挖掘的系统性金融风险度量研究

蔡艳丽^{1,*}

¹ 中南财经政法大学

摘要

本文合成了系统性金融风险的传统金融指标综合指数，然后添加合成了由HMM文本挖掘获得的百度指数，更精细地刻画了金融风险，并比较了添加前后的效果。研究结果发现：传统金融指标合成的综合指数能很好的拟合各金融市场指数，具有一定预测性；添加由HMM文本挖掘的百度指数合成后，既保留了传统金融综合指数的优点，又修正了其存在的部分异常区间，整体上具有更好的拟合效果和先行预测性。

*蔡艳丽，2014年毕业于中南财经政法大学获统计学学士，现中南财经政法大学硕士研究生在读，研究方向为金融统计，在校期间参与国家社科课题《重大风险事件与股市流动性及波动性的关系研究》，主持校级课题《基于NSGAII与粒子群算法对比的共享汽车最优网点选址问题研究》，于学术杂志《商情》发表论文《“退欧”对英镑汇率的影响研究——基于GARCH族模型》