



中国 R 会议
The China-R Conference

第十二届 中国R会议(北京) 会议手册



The 12th China-R Conference (Beijing)

-  2019.5.24-2019.5.26
-  北京·中国农业大学

会议组织



主办方



协办方



狗熊会
CluBear



中国人民大学应用统计科学研究中心
Center for Applied Statistics of Renmin University of China



金牌赞助



独家视频支持



欢迎辞

蓄力一纪，可以远矣。经过十二年的磨砺，中国 R 语言会议又踏上了新的征程。每当这个时候，各位志同道合的朋友以 R 为相聚的理由，从数据科学的各类学术领域而来、从大数据的各种应用行业而来、从天南海北的各条奋斗战线而来，欢聚一堂，共襄盛举。这是 R 的独特魅力。R 的一个核心设计理念是“人的时间永远比机器的时间宝贵”，具有深厚的人文精神，其工程化应用又秉承了“总是有多种方法来做同一件事”的思想，极具包容性。它专注于数据科学和统计建模，保持自己的勃勃生机，又主动和其他的优秀工具融合，让大数据时代的舞台群芳竞艳。这也正如统计学，最大的好处是“可以在所有学科的后院玩耍”。参加会议的朋友们都热爱 R，但不执着 R，甚至不用 R，大有“圣人不凝滞于物”的境界。



十二年一个轮回，中国 R 会议又回到了最初开始的地方，来到中国人民大学。这么多年来，数据领域的各种热门词汇层出不穷，和 R 比较的工具也换了好几轮，但 R 和这个 R 会议一直在这里，这里没有人想一统天下，只想解决现实问题，因为我们知道“所有模型都是错误的，但有些是有用的”。伴随着国家产业升级的历史进程和大数据时代的热潮，本次会议的嘉宾人数、分会场数目又达到了一个新的高峰，演讲主题涵盖了医疗健康、生物信息、心理学、量化金融、工业工程、智能制造、人工智能、概率统计、机器学习、自然语言、智慧教育、天文海洋、新闻传播、数据保护、商业统计、数据可视化等诸多领域。我们真诚地欢迎您到来，一同感受数据科学为这个时代带来的惊喜与挑战。

统计之都敬上

人之大者

为中国人民大学而作

Moderato ♩ = 90

项海波 词/曲

The musical score consists of six staves of music for voice. The key signature is one flat, and the time signature is common time (indicated by '4'). The vocal range is soprano. The lyrics are written below each staff. Measure numbers 1 through 22 are indicated on the left.

1-4: Moderato ♩ = 90. Dynamics: *mf*, *p*. Lyric: 人 大 人 大 巍 巍 气 魄 煦 煦 文 化.

5-8: Dynamics: *f*. Lyric: 古 今 中 外 燥 河 汉 于 此 为 桀 至 真 至 善.

9-12: Dynamics: *p*. Lyric: 文 章 有 炳 寸 心 无 价 明 德 亲 民 扬 彼 大 道.

13-16: Dynamics: *f*, *p*. Lyric: 囊 我 中 华 至 真 至 善 文 章 有 炳 寸 心 无 价.

17-20: Dynamics: *f*. Lyric: 明 德 亲 民 扬 彼 大 道 囊 我 中 华.

目录

欢迎辞	1
会议介绍	1
第十二届中国 R 会议介绍	1
主办机构	2
赞助商介绍	3
第十二届中国 R 会议筹备委员会	4
统计之都简介及活动回顾	5
中国人民大学地图	5
主会场 & 专题会专场日程	6
Keynote(24 日, 八百人大教室)	16
易丹辉: 用数据诠释实际问题	16
俞声: 数据科学在医疗领域的机遇与挑战	16
李大庆: 城市交通的渗流模式	17
朱廷劭: 基于人工智能的在线自杀主动预防	17
吴喜之: 从模型驱动的集体推断到数据驱动的个体预测	18
张建民: 个人信息保护与企业数据产权	18
王朝东: 神经系统疾病诊疗辅助决策与多层次会诊系统建设	19
R01: 狗熊会专场 (25 日上午, 第三教学楼 3102, 主席: 周静)	20
潘蕊: 狗熊会数据科学精品案例库简介	20
李季: 智能热表与供热效率提升	20
黄丹阳: 小微企业支付交易数据的标签和模型方法	21
王菲菲: 智慧零售场景中的标题体系设计及应用	21
吴睿: 环保大数据应用场景及商业价值	22
R02: 隐私与数据保护 大成专场 (25 日上午, 第三教学楼 3103, 主席: 张建民)	23
吴沈括: 数据治理国际动向与前瞻	23
刘熙君: 京东数科个人信息保护实践	23
李玲: 互联网平台个人信息保护观察	24
陈昶屹: 侵犯网络隐私及个人信息案件的审理难点与审判趋势	24
赵中星: 并购融资过程中的数据实务分享	25
夏虞斌: 打通数据孤岛——基于可信硬件的安全数据处理	25
圆桌讨论: 圆桌讨论	25
R03: 可视化 (25 日上午, 第三教学楼 3104, 主席: 张杰)	26
刘永鑫: R 语言在宏基因组数据分析和可视化中的应用	26
余政彦: 从学术到业界的 3 种可视化设计探索与实践	26
钟永剑: D3 财务可视化: 财报呈现一目了然	26
胡悦: 数据可视化与反 p-hacking: 以政治科学研究为例	26
吕妍: 新媒体是如何做可视化报道的?	27
张杰: 数据可视化的色彩运用原理与实践	28

R04: 心理学 1(25 日上午, 第三教学楼 3105, 主席: 夏骁凯)	29
李英武: R 语言在面试考官多级评分中的应用	29
刘彦楼: 认知诊断模型信息矩阵估计软件包 dcminfo 的开发与应用	29
郭少阳: 基于潜混合模型的 4PLM 参数估计 BE3M 算法及其 R 实现	30
薛明锋: 心理测量理论在 R 语言上的实现	30
张雪儿: 基于项目反映理论在考试中的测验校准	30
张沥今: 基于贝叶斯估计的结构方程模型介绍	31
R05: 人大统计 (25 日上午, 第三教学楼 3204, 主席: 李扬)	32
杨翰方: Two-Way Partial AUC and Its Properties	32
孙怡帆: An integrative sparse boosting analysis of cancer genomic commonality and difference	32
马维: Properties of Covariate-Adaptive Randomization with Misclassification in Covariate	32
高光远: Evaluation of driving risk at different speeds	33
R06: 物联网 (25 日上午, 第三教学楼 3206, 主席: 徐浩)	34
熊志敏: 物联网数据安全一致性与分布式账本技术	34
李旭: 自动驾驶与车联网数据应用	34
刘鹏: 新能源汽车大数据与售后服务市场的融合应用	35
武坚利: 室内定位技术在航空制造业车间的落地应用	35
R07: 工业工程 (25 日上午, 第三教学楼 3308, 主席: 梁巧)	36
张晨: Modeling Tunnel Profile in Presence of Coordinate Errors A Gaussian Process Based Approach	36
唐昕迪: Dynamic Management of Autonomous Electric Taxis using Reinforcement Learning Method	36
董航: Modeling and Change Detection for Count-weighted Multi-layer Networks	36
刘心广: R 在某工厂自动化制造产线质量监控中的应用	37
梁巧: 基于电商评论数据的产品及服务质量在线监控	38
R08: 医疗健康 1(25 日上午, 第三教学楼 3310, 主席: 李昂、安澜)	39
刘跃伟: 基于简单空间插值和病例 - 交叉设计探讨大气污染对哮喘死亡的急性影响	39
郑德强: 基于高维 miRNAs 表达谱数据的疾病诊断标记物筛选研究	39
张兵: An R package to explore the effects of environmental factors on infectious diseases	39
夏昌发: 传染病动态传播模型的 R 实现	40
李昂: 基于混合效应模型的人群代谢组学差异性标志物筛选研究	40
R09: 人工智能 (25 日下午, 第三教学楼 3102, 主席: 常象宇、张源源)	42
何靖宇: XBART: 更快更准的提升树 (Boosting Tree) 算法	42
陈昱: 训练拳皇 97 AI	42
吴兴龙: 面向移动端的计算机视觉技术简介	42
骆颇: 工业推荐系统简介	43
熊熹: 深度召回在京东搜索中的应用	43
R10: 北大光华 BA 商业分析 (25 日下午, 第三教学楼 3103, 主席: 王汉生)	44
王汉生: 洞察数据 商业价值——北大光华商业分析硕士项目	44
王翀: 社交大数据	44
厉行: Leaderboard Effect: Who, When, and How?	45
R11: 北大光华 BA 营销大数据 (25 日下午, 第三教学楼 3104, 主席: 沈俏蔚)	46
楚燕来: The Unintended Consequences of Tariff Retaliation: Evidence from the Chinese Automobile Market	46

马雪静: Highest Contributions from Others: A Dampening Effect on PWYW Payment	46
吴少辉: How Is Mobile User Behavior Different?—A Hidden Markov Model of Mobile Application Usage Dynamics	46
张晗: What's Your Risk Attitude On A Grey Day: The Case of Air Quality and Financial Product Choice	47
R12: 医疗大数据 (25 日下午, 第三教学楼 3105, 主席: 陈显扬)	48
王朝东: 数据驱动的分子会诊与精准诊疗	48
马超超: 真实世界研究中的临床数据处理问题	48
段欣岑: 大数据技术在检验医学中的应用	49
章林 & 陈显扬: 图像识别自闭症的算法基础与应用	49
姜楠 & 陈显扬: 不同标志物对乳腺癌的早筛效率分析	49
杜智勇: 基因和代谢组的多层次组学联合分析	50
李奇斌: 基因大数据分析平台的开发和临床应用	50
R13: 软件工具 (25 日下午, 第三教学楼 3204, 主席: 覃文锋)	51
谢士晨: 财经数据分析之 pedquant 包	51
卢宾宾: R 语言空间数据处理与分析	51
杨健: 使用 Shiny 开发中文自然语言处理 Web 应用	51
李宇轩: 基于 knitr、rmarkdown 的 R+HTML 生态应用	51
覃文锋: R 社群的组织与参与	52
R14: 新闻传播 (25 日下午, 第三教学楼 3206, 主席: 王小宁)	53
塔娜: 社交网络上议题社群的公共焦虑研究	53
向安玲: 媒介数据挖掘与指数构建	53
姜柳: 不是科学家, 媒体怎么做“数据可视化”?	53
王妍: 影视大数据用户行为分析	54
李波: 电影《摇滚藏獒》营销案例分析	55
R15: 研究生专场 (25 日下午, 第三教学楼 3308, 主席: 张心雨)	56
Li Xixi: Forecasting with time series imaging	56
李杰: Kolmogorov-Smirnov simultaneous confidence bands for time series distribution function	56
赵一懋: 统计学学生在金融行业中求职的方向	57
黄涛: Large-scale Regression with Two-stage Best-score Random Forest	57
林毓聪: Long Distance Relation Extraction with Article Structure Embedding and Applied to Mining Medical Knowledge	58
黄湘云: 统计之都在线投稿系统	58
R16: 医疗健康 2(25 日下午, 第三教学楼 3310, 主席: 李昂、安澜)	59
李国星: 大气污染与人群健康的关系	59
陈善恩: 利用改进的高斯过程模型预测季节性流感的传播	59
王钒: Selection of mixed copulas for data with ties via penalized likelihood	60
安澜: R 语言在人群为基础的癌症登记数据的生存分析中的应用	60
林华亮: R 语言在大气颗粒污染物健康影响流行病学研究中的应用	61
R17: 智慧教育 (26 日上午, 第三教学楼 3102, 主席: 冯俊晨)	62
任万凤: 敏捷数据科学家如何玩转教学闭环产品?	62
Dan Bindman: A New Model of Knowledge Assessment and AI Adaptive Learning	62

饶丰: AI 在 K12 场景下的应用实践	62
何明: 基于电脑使用日志剖析和评估用户拖延行为	63
R18: 机器学习应用 (26 日上午, 第三教学楼 3103, 主席: 何珂骏)	64
涂富艺: 基于 Unet 的直肠肿瘤识别	64
叶小清: 肿瘤影像特征提取分析	64
刘晓玉: 肿瘤影像特征与淋巴结转移的相关性验证	64
苏蔚: 函数型数据变系数模型的估计 (Estimation of varying coefficient model for functional data)	64
夏强: 高维时间序列数据的降维处理——因子个数的确定研究	65
周海鹏: 机器学习在 LBS 中的应用	66
R19: 量化金融 (26 日上午, 第三教学楼 3104, 主席: 赵阳)	67
赵然: “金融科技”的春天	67
郭彪: 融资融券与 A 股收益率预测性	67
李孟育: 风险平价资产配置 -以商品期货、可转债为例	68
汪昊: 金融科技公司如何利用人工智能技术进行风控	68
张丹: 凯利公式 -用胜率和赔率量化你的投资	69
R20: 大数据应用 (26 日上午, 第三教学楼 3105, 主席: 李舰)	70
张耀峰: 基于警务大数据的犯罪事件智能预测	70
张忠元: 聚类方法的评价研究	70
蔡锐: Online learning 在大规模机器学习中的理论与应用	70
曾梓龙: 神经影像大数据中机器学习的应用	70
R21: 本科生专场 (26 日上午, 第三教学楼 3204, 主席: 冉佳鹭)	72
李家郡: 使用 Rstudio 结合 RcppArmadillo 制作可以快速随机计算稀疏矩阵奇异值分解的包	72
金滢: 伊辛图模型组合结构推断问题的计算 -统计权衡	72
康越: 基于岭比收缩的高光谱图像端元个数估计方法	72
刘秋华: 基于前列腺分割任务的 DDSP 网络设计和损失函数探讨	73
林枫: 从零开始的 COSplay	74
R22: 心理学 2(26 日上午, 第三教学楼 3206, 主席: 夏骁凯)	75
吕杰好: R 在心理学中的应用	75
高树青: “乱世重才, 治世重德”——经济不确定与德 - 才偏好	75
黄文昊: 人们对社会与金钱奖赏的预期共享神经环路: 结合多种分析方法的脑成像元分析研究	75
夏晓磊: 心理学脑电研究中数据的基本概念与分析实践	76
赵加伟: 基于 OSF 增强研究中的开放科学	76
R23: 生物信息 (26 日上午, 第三教学楼 3308, 主席: 伊现富)	77
连明: 机器学习在生物信息中的应用	77
张韬: Platform-independent approach for cancer detection from gene expression profiles of peripheral blood cells	77
李发金: Analysis for Ribosome Profiling Data	78
侯春宇: 利用蛋白质组学和生物信息学研究 PKC ζ 相互作用蛋白网络	78
伊现富: Genome-wide and cell type-specific pattern of transcriptional regulators cooperation in 3D chromatin	79
R24: 智能制造 (26 日下午, 第三教学楼 3102, 主席: 崔鹏飞)	80
崔鹏飞: 风电大数据落地应用实践	80

董兆宇：数据分析中的并行计算浅谈	80
于亚杰：数据分析在家电行业内的应用	80
陈肇江：智能制造在航空工业的探索	81
R25: 时间序列预测 (26 日下午, 第三教学楼 3103, 主席: 康雁飞)	82
陈艺天: 大数据时代的需求预测	82
Kang Yanfei: Feature-based time series forecasting	82
Wang Xiaoqian: Probabilistic forecasting based on time series features	82
王孟樵: 用 R 玩转中国生育率分析	83
白云: Text based crude oil price forecasting	84
R26: 自然语言处理 1(26 日下午, 第三教学楼 3104, 主席: 崔子璇、鲍莞倩)	85
张家兴: 金融对话机器人	85
段清华: 对话系统的历史与未来	85
敖翔: NLP+ 金融: 场景及技术趋势	85
沈泽希: 美人如花隔云端——对话机器人	85
R27: 自然语言处理 2(26 日下午, 第三教学楼 3105, 主席: 崔子璇、鲍莞倩)	87
戴明峰: 基于电商平台手机评论数据的文本挖掘	87
陈功: 面向中国学生的英语书面语动词形式错误自动检查——基于链语法的研究	87
孙亚: 基于 Wmatrix 语义赋码的商务话语概念隐喻分析	88
焦鲁: 混合效应模型 (Mixed-Effects Models) 在二语研究中的应用	88

第十二届中国 R 会议介绍

中国 R 会议 (The China-R Conference) 始于 2008 年，由统计之都 (Capital of Statistics, COS) 发起，联合各地高校、企业共同举办。会议旨在提供一个高质量的分享平台，让更多人了解、使用、推广、发展统计学方法及其在各领域的应用。R 会议起始于 R 语言的讨论，后来兼容并包，积极走向更广义的数据科学领域，聚各领域的学术专家、业界精英、技术大咖、莘莘学子于一堂，使各界参会者都得到充分的交流。作为国内最大的数据科学会议，R 会议已服务数万参会人员。

截至目前，R 会议已经在中国人民大学、北京大学、清华大学、华东师范大学、上海财经大学、中山大学、西安欧亚学院、厦门大学、江西财经大学、浙江财经大学、杭州师范大学、中南财经政法大学、湖北经济学院、西南财经大学、贵州大学、兰州财经大学、中国科学技术大学等多个高校举办。2018 年，第十一届中国 R 会议在北京、上海、广州分别举办，其中北京会场吸引了来自全国各地的 1300 余位参会者，在三天的会议中，各界人士汇聚一堂，进行思维的碰撞。今年将迎来第十二届中国 R 会议。

本届 R 会议由统计之都主办，由中国人民大学统计学院、中国人民大学应用统计科学研究中心和狗熊会协办，将于 5 月 24-26 日在北京举办。本届会议覆盖统计学、大数据、人工智能相关理论及其在各行各业的具体应用，包括医疗健康、生物信息、医疗大数据、心理学、量化金融、工业工程、智能制造、软件工具、计算平台、概率统计、统计理论、机器学习、人工智能、大数据应用、自然语言、新闻传播、社交网络、商务统计、人文科学等数据科学话题，我们欢迎您的到来！

24 日主会场地点为中国人民大学八百人大教室，25-26 日分会场地点为中国人民大学公共教学三楼，请您事先查阅好感兴趣的会场，并提前熟悉校园环境和路线，以便更加高效地参加会议。

主办机构

统计之都

统计之都 (Capital of Statistics, 简称 COS, 网址 <http://cosx.org/>)，成立于 2006 年 5 月，是一家旨在推广与应用统计学知识的网站和社区，其口号是“中国统计学门户网站，免费统计学服务平台”。统计之都发源于中国人民大学统计学院，由谢益辉创建，现由世界各地的众多志愿者共同管理维护，理事会现任主席为冯凌秉。统计之都致力于搭建一个开放的平台，使得科研人员、数据分析人员和统计学爱好者能互相交流合作，一方面促进彼此专业知识技能的增长，另一方面为国内统计学和数据科学的发展贡献自己的力量。

协办方

中国人民大学统计学院

中国人民大学统计学科始建于 1950 年，两年后成立统计学系，是新中国经济学科中最早设立的统计学系，2003 年 7 月，成立中国人民大学统计学院。多年来，本学科一直强调统计理论和统计应用的结合，不断拓宽统计教学和研究领域，成为统计学全国重点学科，在 2012 年、2017 年教育部全国统计学一级学科评估中排名第一。学院拥有统计学一级学科博士点和博士后流动站，拥有经济统计学和风险管理与精算学两个二级学科博士点，拥有预防医学与公共卫生一级学科硕士授权点，统计学、概率论与数理统计、风险管理与精算学、流行病与卫生统计学四个学术型硕士点，应用统计学专业学位硕士点，统计学、经济统计学、应用统计学（风险管理与精算）、数据科学与大数据技术四个本科专业，是全国拥有理学、经济学、医学三大门类统计学专业最齐全的统计学院。

狗熊会

狗熊会是一个致力于数据产业的高端智库。狗熊会帮助合作伙伴制定数据战略，培养数据人才，研究数据业务，发现数据价值，推动产业进步！狗熊会使命：“聚数据英才，助产业振兴”！

中国人民大学应用统计科学研究中心

中国人民大学应用统计科学研究中心是中华人民共和国教育部所属百所人文社会科学重点研究基地之一，它成立于 2000 年 9 月，其前身是 1988 年成立的中国人民大学统计科学研究所。研究中心积极培育中青年学术骨干，逐渐发展并形成了经济与社会统计、统计调查与数据分析、风险管理与精算、生物卫生统计，四个各具特色的研究方向。中心建设的重点研究平台是：1. 重大发展问题的统计技术创新研究。2. 现代统计技术与方法的应用性研究。3. 精算技术的创新与应用。4. 生物医学统计技术发展与应用。研究中心拥有国内一流的研究人员，承担多项国家及教育部项目，获得丰硕的研究成果。应用统计科学研究中心，始终将建立和发展应用统计学科基地作为战略定位，着重从制定应用统计研究的科学规划、密切联系实际选准科研攻关方向、注重研究工作的长期积累、加强重点研究平台建设等方面开展工作。

赞助商介绍

金牌赞助

图灵教育

北京图灵文化发展有限公司，始终以策划出版高质量的科技图书为核心业务，自成立以来累计销售图书已超 1000 万册，影响了数百万读者。旗下图灵教育品牌是国内计算机图书领域的高端品牌之一。图灵社区是图灵公司打造的综合性服务平台，集图书内容生产、作译者服务、电子书销售、技术人士交流于一体。

中国人民大学出版社

中国人民大学出版社成立于 1955 年，是新中国成立后的第一家大学出版社。1982 年被教育部确定为全国高等学校文科教材出版中心，2007 年获首届中国出版政府奖先进出版单位奖，2009 年获首届全国百佳图书出版单位荣誉称号，是中国最重要的高校教材和学术著作出版基地之一。我社统计学出版坚持精品战略，汇集了中国人民大学、北京大学、厦门大学、中央财经大学等国内众多知名高校的统计学教授的代表性教材和著作，受到了国内统计学老师的普遍认可。同时，紧跟学科发展前沿，率先出版了大数据系列教材和《数据科学概论》等。

RStudio

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（RConsortium），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。

人民邮电出版社

人民邮电出版社成立于 1953 年 10 月，隶属于工信部，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 20 亿元。

独家视频支持

IT 大咖说

IT 大咖说，IT 垂直领域的大咖知识分享平台，践行“开源是一种态度”，通过线上线下开放模式分享行业 TOP 大咖干货，技术大会在线直播点播，在线直播知识分享平台。200+ 合作社区，每周 30+ 场技术大会精彩分享，4000+ 业内大咖资源。让程序猿、攻城狮不再遗憾，随时随地，想看就看，让智慧流动起来！

第十二届中国 R 会议筹备委员会

主席：任怡萌

秘书长：王袆帆

秘书团：操懿、任焱、苏锦华、雷博文、李宇轩、顾小涵、王小宁、杨舒仪、黄湘云、夏骁凯

志愿者：白迎辰，蔡燕，柴树文，陈卉，陈美昆，陈一卿，程铁鹏，池义淳，董钦源，方焯，冯春进，侯松阳，高钰婷，顾琳，顾雨婷，何贤文，胡加曼，黄沁雪，黄依诺，黄允祇，霍霁雨，贾开文，姜倩云，姜威，解圆圆，李卜诺，李聪玥，李浩源，李璇，李雨桐，李昀芮，李子贤，梁湜，梁雯丽，梁小泽，梁旖韵，林嘉琦，林锦锋，刘姝畅，刘雨沙，刘雨馨，麻世钰，梅亚园，莫慧霖，尼玛，倪天辰，聂仪珂，任完美，邵蕾，师晓泉，宋逸楠，王博宁，王晨阳，王文丽，王志颖，魏晓容，吴爱娟，吴冕，吴柱容，向悦，谢雨淞，谢泽君，熊多多，徐嘉蔚，徐依格，杨春白雪，杨光，杨彤，叶中盛，于金朝，于子奇，虞汉婧，袁梦真，翟禹佳，张慧玲，张佩珊，张士琦，张一霖，张宇珠，张雨馨，赵佳欣，赵亦盈，赵颖，赵增辉，赵紫菡，郑敏行，周昕仪，周紫萱

统计之都简介及活动回顾

“统计之都”(Capital of Statistics, 简称 COS)网站成立于 2006 年 5 月 19 日，其主旨为传播统计学知识并将其应用于实际领域。纵观现今国内统计学理论和应用的发展，一方面我们不难发现统计学在应用领域的巨大潜力——现代管理、咨询、商业、经济、金融、医药、生物等等，无不需要数据的力量，而另一方面我们也不得不承认，国内统计学的应用很大程度上受理论的制约——无论是应用界的人们对统计学基础理论知识的欠缺，还是学术界所研究的理论对应用领域问题的轻视。“统计之都”网站便是基于这样的认识而创建的。我们希望，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。“统计之都”为非赢利性质网站，但大力欢迎所有商界和研究领域的朋友与我们在实际应用问题上合作。我们的口号是：

中国统计学门户网站，免费统计学服务平台

我们怀着“十年磨一剑”的决心，要将“统计之都”创建成中国的统计学“正直、人本、专业”的社区；我们抱着“己欲立而立人、己欲达而达人”的信条，要将“统计之都”以免费统计学服务平台的形式坚持办下去。我们希望“统计之都”在专业知识体系上有真正的王者风范，在面对用户需求时却又以谦恭的态度为大家服务。统计之都（下文简称 COS）目前由线上与线下两部分构成。其中，线上内容主要包括主站（<http://cosx.org/>）以及微信公众号（CapStat）；随着越来越多喜爱数据科学的朋友们加入，大家对于线下活动和书稿撰写翻译等等的需求也越来越旺。COS 线下活动总结：COS 线下活动总结：

1. 中国 R 会议：目前已开展到第十一届，分别在北京、上海、广州、杭州、西安、武汉、成都、贵阳、南昌、厦门、合肥、太原等地举办。历届会议纪要和幻灯片共享都可以在 COS 主站上找到：<http://china-r.org/>
2. 线下沙龙：目前我们在北京、上海和广州深圳开展线下沙龙活动。不同于规模庞大的 R 语言会议，沙龙形式更为轻巧，注重讨论交流。目前已经举办过 50 期，目前主要在北京、上海每月举办，详情参见统计之都主站及微信公众号。
3. 海外在线视频沙龙：我们在 Google Hangouts 举办在线沙龙，主要由海外嘉宾来分享学术、生活中的点点滴滴。目前已经举办 23 期：<http://meetup.cos.name/>.
4. 书籍出版，包括写作和翻译。如《Dynamic Documents with R and knitr》(2nd edition) 谢益辉著，《Implementing Reproducible Research》谢益辉等著，《bookdown: Authoring Books and Technical Documents with R Markdown》谢益辉著，《数据科学中的 R 语言》李舰、肖凯著，《R 语言实战》高涛、肖楠、陈钢翻译，《ggplot2: 数据分析与图形艺术》统计之都翻译，《R 语言核心技术手册》刘思喆、李舰、陈钢、邓一硕翻译，《R 语言编程艺术》陈堰平、邱怡轩、潘岚锋等翻译，《R 数据可视化手册》肖楠、邓一硕、魏太云翻译，《R 语言统计入门》邓一硕、郝智恒、何通翻译，《数据科学实战》冯凌秉、王群锋翻译，《R 语言实战》(第 2 版) 王小宁、刘撷芯、黄俊文翻译，《Rcpp: R 与 C++ 的无缝结合》寇强、张晔翻译，《R 绘图系统》呼思乐、张晔、蔡俊翻译，《R 语言编程实战》冯凌秉翻译，《量化投资与 R》(待出版) 邓一硕、冯凌秉、杨环翻译，《金融风险建模与投资组合优化》(待出版) 邓一硕、郑志勇等翻译、《ggplot2: 数据分析与图形艺术 (第 2 版)》黄俊文、王小宁、于嘉傲、冯璟烁，《统计之美：人工智能时代的科学思维》李舰，海恩著等等。

中国人民大学地图



注：图上标记的餐厅都可以支付宝消费。

5月24日（周五）主会场日程

主会场	演讲嘉宾	主题	时间
Keynote (八百人大教室)		参会者入场	08:00~09:00
		致辞	09:00~09:15
	易丹辉	用数据诠释实际问题	09:15~10:00
	俞声	数据科学在医疗领域的机遇与挑战	10:00~10:45
		自由讨论、休息	10:45~11:15
	李大庆	城市交通的渗流模式	11:15~12:00
Keynote (八百人大教室)	朱廷劭	基于人工智能的在线自杀主动预防	14:00~14:45
	吴喜之	从模型驱动的集体推断到数据驱动的个体预测	14:45~15:30
		自由讨论、休息	15:30~16:00
	张建民	个人信息保护与企业数据产权	16:00~16:45
	王朝东	神经系统疾病诊疗辅助决策与多层次会诊系统建设	16:45~17:30

第十二届中国R会议（北京）分会场日程

分会场	25日上午	25日下午	26日上午	26日下午
3102	R01 狗熊会专场 主席：周静	R09 人工智能 主席：常象宇、张源源	R17 智慧教育 主席：冯俊晨	R24 智能制造 主席：崔鹏飞
3103	R02 隐私与数据保护 (大成专场) 主席：张建民	R10 北大光华 BA 商业分析专场 主席：王汉生	R18 机器学习应用 主席：何珂骏	R25 时间序列预测 主席：康雁飞
3104	R03 可视化专场 主席：张杰	R11 北大光华 BA 营销大数据专场 主席：沈俏蔚	R19 量化金融 主席：赵阳	R26 自然语言处理 1 主席：崔子璇、鲍莞倩
3105	R04 心理学 1 主席：夏骁凯	R12 医疗大数据 主席：陈显扬	R20 大数据应用 主席：李舰	R27 自然语言处理 2 主席：崔子璇、鲍莞倩
3204	R05 人大统计 主席：李扬	R13 软件工具 主席：覃文锋	R21 本科生专场 主席：冉佳鹭	
3206	R06 物联网 主席：徐浩	R14 新闻传播 主席：王小宁	R22 心理学 2 主席：夏骁凯	
3308	R07 工业工程 主席：梁巧	R15 研究生专场 主席：张心雨	R23 生物信息 主席：伊现富	
3310	R08 医疗健康 1 主席：李昂、安澜	R16 医疗健康 2 主席：李昂、安澜		

注：分会场在中国人民大学公共教学三楼的各个教室，如3102表示公共教学三楼一层3102教室。

分会场	演讲嘉宾	主题	时间
25日上午			
狗熊会专场 第三教学楼 3102 主席：周静	潘蕊	狗熊会数据科学精品案例库简介	9:00-9:30
	李季	智能热表与供热效率提升	9:30-10:00
		自由讨论、休息	10:00-10:30
	黄丹阳	小微企业支付交易数据的标签和模型方法	10:30-11:00
	王菲菲	智慧零售场景中的标题体系设计及应用	11:00-11:30
	吴睿	环保大数据应用场景及商业价值	11:30-12:00
隐私与数据保护 (大成专场) 第三教学楼 3103 主席：张建民	吴沈括	数据治理国际动向与前瞻	8:30-8:55
	刘熙君	京东数科个人信息保护实践	8:55-9:20
	李玲	互联网平台个人信息保护观察	9:20-9:45
		自由讨论、休息	9:45-10:00
	陈祖屹	侵犯网络隐私及个人信息案件的审理难点与审判趋势	10:00-10:25
	赵中星	并购融资过程中的数据实务分享	10:25-10:50
	夏虞斌	打通数据孤岛——基于可信硬件的安全数据处理	10:50-11:15
		圆桌讨论	11:15-12:00
可视化专场 第三教学楼 3104 主席：张杰	刘永鑫	R 语言在宏基因组数据分析和可视化中的应用	8:30-9:00
	余政彦	从学术到业界的 3 种可视化设计探索与实践	9:00-9:30
	钟永剑	D3 财务可视化：财报呈现一目了然	9:30-10:00
		自由讨论、休息	10:00-10:30
	胡悦	数据可视化与反 p-hacking：以政治科学研究为例	10:30-11:00
	吕妍	新媒体是如何做可视化报道的？	11:00-11:30
	张杰	数据可视化的色彩运用原理与实践	11:30-12:00
R04 心理学 1 第三教学楼 3105 主席：夏晓凯	李英武	R 语言在面试考官多级评分中的应用	8:30-9:00
	刘彦楼	认知诊断模型信息矩阵估计软件包 dcminfo 的开发与应用	9:00-9:30
	郭少阳	基于潜混合模型的 4PLM 参数估计 BE3M 算法及其 R 实现	9:30-10:00
		自由讨论、休息	10:00-10:30
	薛明锋	心理测量理论在 R 语言上的实现	10:30-11:00
	张雪儿	基于项目反映理论在考试中的测验校准	11:00-11:30
	张沥今	基于贝叶斯估计的结构方程模型介绍	11:30-12:00
R05 人大统计 第三教学楼 3204 主席：李扬	杨翰方	Two-Way Partial AUC and Its Properties	9:00-9:30
	孙怡帆	An integrative sparse boosting analysis of cancer genomic commonality and difference	9:30-10:00
		自由讨论、休息	10:00-10:30
	马维	Properties of Covariate-Adaptive Randomization with Misclassification in Covariate	10:30-11:00
	高光远	Evaluation of driving risk at different speeds	11:00-11:30

分会场	演讲嘉宾	主题	时间
R06 物联网 第三教学楼 3206 主席：徐浩	熊志敏	物联网数据安全一致性与分布式账本技术	9:00-9:30
	李旭	自动驾驶与车联网数据应用	9:30-10:00
		自由讨论、休息	10:00-10:30
	刘鹏	新能源汽车大数据与售后服务市场的融合应用	10:30-11:00
	武坚利	室内定位技术在航空制造业车间的落地应用	11:00-11:30
R07 工业工程 第三教学楼 3308 主席：梁巧	张晨	Modeling Tunnel Profile in Presence of Coordinate Errors A Gaussian Process Based Approach	9:00-9:30
	唐昕迪	Dynamic Management of Autonomous Electric Taxis using Reinforcement Learning Method	9:30-10:00
		自由讨论、休息	10:00-10:30
	董航	Modeling and Change Detection for Count-weighted Multi-layer Networks	10:30-11:00
	刘心广	R 在某工厂自动化制造产线质量监控中的应用	11:00-11:30
R08 医疗健康 1 第三教学楼 3310 主席：李昂、安澜	梁巧	基于电商评论数据的产品及服务质量在线监控	11:30-12:00
	刘跃伟	基于简单空间插值和病例-交叉设计探讨大气污染对哮喘死亡的急性影响	9:00-9:30
	郑德强	基于高维 miRNAs 表达谱数据的疾病诊断标记物筛选研究	9:30-10:00
		自由讨论、休息	10:00-10:30
	张兵	An R package to explore the effects of environmental factors on infectious diseases	10:30-11:00
夏昌发			
	李昂	传染病动态传播模型的 R 实现	11:00-11:30
	李昂	基于混合效应模型的人群代谢组学差异性标志物筛选研究	11:30-12:00

分会场	演讲嘉宾	主题	时间
25日下午			
R09 人工智能 第三教学楼 3102 主席： 常象宇、张源源	何靖宇	XBART：更快更准的提升树（Boosting Tree）算法	14:00-14:30
	陈昱	训练拳皇 97 AI	14:30-15:00
		自由讨论、休息	15:00-15:30
	吴兴龙	面向移动端的计算机视觉技术简介	15:30-16:00
	骆颇	工业推荐系统简介	16:00-16:30
	熊熹	深度召回在京东搜索中的应用	16:30-17:00
R10 北大光华 BA 商业分析 第三教学楼 3103 主席：王汉生	王汉生	洞察数据 商业价值——北大光华商业分析硕士项目	14:00-14:45
	王翀	社交大数据	14:45-15:30
		自由讨论、休息	15:30-16:00
	厉行	Leaderboard Effect: Who, When, and How?	16:00-16:45
R11 北大光华 BA 营销大数据 第三教学楼 3104 主席：沈俏蔚	楚燕来	The Unintended Consequences of Tariff Retaliation: Evidence from the Chinese Automobile Market	14:00-14:30
	马雪静	Highest Contributions from Others: A Dampening Effect on PWYW Payment	14:30-15:00
		自由讨论、休息	15:00-15:30
	吴少辉	How Is Mobile User Behavior Different? — A Hidden Markov Model of Mobile Application Usage Dynamics	15:30-16:00
	张晗	What's Your Risk Attitude On A Grey Day: The Case of Air Quality and Financial Product Choice	16:00-16:30
R12 医疗大数据 第三教学楼 3105 主席：陈显扬	王朝东	数据驱动的分子会诊与精准诊疗	14:30-14:55
	马超超	真实世界研究中的临床数据处理问题	14:55-15:20
	段欣岑	大数据技术在检验医学中的应用	15:20-15:45
		自由讨论、休息	15:45-16:10
	章林 & 陈显扬	图像识别自闭症的算法基础与应用	16:10-16:35
	姜楠 & 陈显扬	不同标志物对乳腺癌的早筛效率分析	16:35-17:00
	杜智勇	基因和代谢组的多层次组学联合分析	17:00-17:25
R13 软件工具 第三教学楼 3204 主席：覃文锋	李奇斌	基因大数据分析平台的开发和临床应用	17:25-17:50
	谢士晨	财经数据分析之 pedquant 包	14:00-14:30
	卢宾宾	R 语言空间数据处理与分析	14:30-15:00
	杨健	使用 Shiny 开发中文自然语言处理 Web 应用	15:00-15:30
		自由讨论、休息	15:30-16:00
	李宇轩	基于 knitr、rmarkdown 的 R+HTML 生态应用	16:00-16:30
	覃文锋	R 社群的组织与参与	16:30-17:00

分会场	演讲嘉宾	主题	时间
R14 新闻传播 第三教学楼 3206 主席：王小宁	塔娜	社交网络上议题社群的公共焦虑研究	14:00-14:30
	向安玲	媒介数据挖掘与指数构建	14:30-15:00
	姜柳	不是科学家，媒体怎么做“数据可视化”？	15:00-15:30
		自由讨论、休息	15:30-16:00
	王妍	影视大数据用户行为分析	16:00-16:30
	李波	电影《摇滚藏獒》营销案例分析	16:30-17:00
R15 研究生专场 第三教学楼 3308 主席：张心雨	Li Xixi	Forecasting with time series imaging	14:00-14:30
	李杰	Kolmogorov-Smirnov simultaneous confidence bands for time series distribution function	14:30-15:00
	赵一懋	统计学学生在金融行业中求职的方向	15:00-15:30
		自由讨论、休息	15:30-16:00
	黄涛	Large-scale Regression with Two-stage Best-score Random Forest	16:00-16:30
	林毓聪	Long Distance Relation Extraction with Article Structure Embedding and Applied to Mining Medical Knowledge	16:30-17:00
R16 医疗健康 2 第三教学楼 3310 主席：李昂、安澜	黄湘云	统计之都在线投稿系统	17:00-17:30
	李国星	大气污染与人群健康的关系	14:00-14:30
	陈善恩	利用改进的高斯过程模型预测季节性流感的传播	14:30-15:00
	王钒	Selection of mixed copulas for data with ties via penalized likelihood	15:00-15:30
		自由讨论、休息	15:30-16:00
	安澜	R 语言在人群为基础的癌症登记数据的生存分析中的应用	16:00-16:30
	林华亮	R 语言在大气颗粒污染物健康影响流行病学研究中的应用	16:30-17:00

分会场	演讲嘉宾	主题	时间
26日上午			
R17 智慧教育 第三教学楼 3102 主席：冯俊晨	任万凤	敏捷数据科学家如何玩转教学闭环产品？	9:00-9:30
	Dan Bindman	A New Model of Knowledge Assessment and AI Adaptive Learning	9:30-10:00
		自由讨论、休息	10:00-10:30
	饶丰	AI 在 K12 场景下的应用实践	10:30-11:00
	何明	基于电脑使用日志剖析和评估用户拖延行为	11:00-11:30
R18 机器学习应用 第三教学楼 3103 主席：何珂骏	涂富艺	基于 Unet 的直肠肿瘤识别	8:30-9:00
	叶小清	肿瘤影像特征提取分析	9:00-9:30
	刘晓玉	肿瘤影像特征与淋巴结转移的相关性验证	9:30-10:00
		自由讨论、休息	10:00-10:30
	苏蔚	函数型数据变系数模型的估计 (Estimation of varying coefficient model for functional data)	10:30-11:00
	夏强	高维时间序列数据的降维处理——因子个数的确定研究	11:00-11:30
	周海鹏	机器学习在 LBS 中的应用	11:30-12:00
R19 量化金融 第三教学楼 3104 主席：赵阳	赵然	“金融科技”的春天	9:00-9:30
	郭彪	融资融券与 A 股收益率预测性	9:30-10:00
		自由讨论、休息	10:00-10:30
	李孟育	风险平价资产配置-以商品期货、可转债为例	10:30-11:00
	汪昊	金融科技公司如何利用人工智能技术进行风控	11:00-11:30
	张丹	凯利公式-用胜率和赔率量化你的投资	11:30-12:00
R20 大数据应用 第三教学楼 3105 主席：李舰	张耀峰	基于警务大数据的犯罪事件智能预测	9:00-9:30
	张忠元	聚类方法的评价研究	9:30-10:00
		自由讨论、休息	10:00-10:30
	蔡锐	Online learning 在大规模机器学习中的理论与应用	10:30-11:00
	曾梓龙	神经影像大数据中机器学习的应用	11:00-11:30
R21 本科生专场 第三教学楼 3204 主席：冉佳鹭	李家郡	使用 Rstudio 结合 RcppArmadillo 制作可以快速随机计算稀疏矩阵奇异值分解的包	9:00-9:30
	金滢	伊辛图模型组合结构推断问题的计算-统计权衡	9:30-10:00
		自由讨论、休息	10:00-10:30
	康越	基于岭比收缩的高光谱图像端元个数估计方法	10:30-11:00
	刘秋华	基于前列腺分割任务的 DDSP 网络设计和损失函数探讨	11:00-11:30
	林枫	从零开始的 COSplay	11:30-12:00

分会场	演讲嘉宾	主题	时间
R22 心理学 2 第三教学楼 3206 主席：夏晓凯	吕杰妤	R 在心理学中的应用	9:00-9:30
	高树青	“乱世重才，治世重德”——经济不确定与德-才偏好	9:30-10:00
		自由讨论、休息	10:00-10:30
	黄文昊	人们对社会与金钱奖赏的预期共享神经环路：结合多种分析方法的脑成像元分析研究	10:30-11:00
	夏晓磊	心理学脑电研究中数据的基本概念与分析实践	11:00-11:30
	赵加伟	基于 OSF 增强研究中的开放科学	11:30-12:00
R23 生物信息 第三教学楼 3308 主席：伊现富	连明	机器学习在生物信息中的应用	9:00-9:30
	张韬	Platform-independent approach for cancer detection from gene expression profiles of peripheral blood cells	9:30-10:00
		自由讨论、休息	10:00-10:30
	李发金	Analysis for Ribosome Profiling Data	10:30-11:00
	侯春宇	利用蛋白质组学和生物信息学研究 PKCζ相互作用蛋白网络	11:00-11:30
	伊现富	Genome-wide and cell type-specific pattern of transcriptional regulators cooperation in 3D chromatin	11:30-12:00

分会场	演讲嘉宾	主题	时间
26 日下午			
R24 智能制造 第三教学楼 3102 主席：崔鹏飞	崔鹏飞	风电大数据落地应用实践	14:00-14:30
	董兆宇	数据分析中的并行计算浅谈	14:30-15:00
		自由讨论、休息	15:00-15:30
	于亚杰	数据分析在家电行业内的应用	15:30-16:00
	陈肇江	智能制造在航空工业的探索	16:00-16:30
R25 时间序列预测 第三教学楼 3103 主席：康雁飞	陈艺天	大数据时代的需求预测	14:00-14:30
	Kang Yanfei	Feature-based time series forecasting	14:30-15:00
	Wang Xiaoqian	Probabilistic forecasting based on time series features	15:00-15:30
		自由讨论、休息	15:30-16:00
	王孟樵	用 R 玩转中国生育率分析	16:00-16:30
R26 自然语言处理 1 第三教学楼 3104 主席： 崔子璇、鲍莞倩	白云	Text based crude oil price forecasting	16:30-17:00
	张家兴	金融对话机器人	14:00-14:40
	段清华	对话系统的历史与未来	14:40-15:20
		自由讨论、休息	15:20-15:50
	敖翔	NLP+金融：场景及技术趋势	15:50-16:30
R27 自然语言处理 2 第三教学楼 3105 主席： 崔子璇、鲍莞倩	沈泽希	美人如花隔云端——对话机器人	16:30-17:10
	戴明锋	基于电商平台手机评论数据的文本挖掘	14:00-14:40
	陈功	面向中国学生的英语书面语动词形式错误自动检查——基于链语法的研究	14:40-15:20
		自由讨论、休息	15:20-15:50
	孙亚	基于 Wmatrix 语义赋码的商务话语概念隐喻分析	15:50-16:30
	焦鲁	混合效应模型 (Mixed-Effects Models) 在二语研究中的应用	16:30-17:10

用数据诠释实际问题

易丹辉（中国人民大学）

时间：09:15-10:00

简介：易丹辉，中国人民大学统计学院教授、博士生导师。主要从事统计方法在经济、金融、保险、管理、医疗等领域应用的研究。研究方向：风险管理与保险、预测与决策；出版专著：《结构方程模型：方法与应用》、《时间序列分析：方法与应用》、《非参数统计—方法与应用》、《统计预测—方法与应用》、《经济预测与决策》等，主编：《数据分析与 EViews 应用》等；发表学术论文百篇；主持承担国家自然科学基金、国家社科基金、教育部博士点基金、北京市哲学社会科学“十五”规划项目等 10 多项，主持承担“亚健康人群中医基本证候流行病学调查统计分析”、“晚期结直肠癌中医优势人群特征研究”、“中国心律失常注册研究”等委托项目近百项。

摘要：通常讲用数据说话，也就是要用数据诠释实际问题。无论数据类型怎样、数据量多少，只要能够客观反映实际现象的特征、变化，都可以在一定程度上解释说明实际问题。通过实际示例，说明数据质量是用数据说话的前提；选择合理科学的方法，是用数据客观真实诠释实际问题的保障；多种统计方法结合运用，层层递进、不断深入分析数据，才能更全面揭示事物的特征、变化规律，才能更好地诠释实际问题。

数据科学在医疗领域的机遇与挑战

俞声（清华大学）

时间：10:00-10:45

简介：俞声博士的研究方向是医学信息学，主要研究内容包括自动术语识别、关系提取、表示学习等自然语言处理问题，以及大规模医学知识图谱构建、表型提取、临床决策支持等问题。俞声博士现任职清华大学统计学研究中心副教授、数据科学研究院 RONG 教授，归国前是电子病历驱动的基因组学研究先驱 –i2b2 美国国家生物医学计算中心的成员之一，在多项精准医学重点项目中承担研发任务。俞声博士将统计学与人工智能技术应用于医学信息领域，在电子病历文本数据分析领域取得了一系列突破性成果，每年在医学信息学顶刊 JAMIA 上发表论文，并获选 Editor’s Choice。他所开发的无监督学习技术使疾病表型识别算法开发速度从每年 1-2 个提高到每年超过 1000 个，并应用于 Partners HealthCare Biobank、Veteran Affairs “Million Veteran Program” 等美国国家级精准医学研究项目。

摘要：医疗服务关系着每个人的生老病死，是人类生存的基本需求。为每一个人的健康服务，使医疗大数据自然成为了最崇高的一种大数据应用。当前，以电子病历为代表的各种医疗数据的电子化让统计学方法与机器学习等数据科学技术在医疗领域的大规模使用迎来了一个黄金时期，而医学信息学作为一个专门研究医疗大数据技术的交叉学科，也如雨后春笋般在哈佛大学等顶尖学府落地成立院系。目前，除了 IBM Watson、基因分析和医学影像识别等比较广为人知的商业化应用外，数据科学技术的运用已经深入到临床科研、医疗政策与保险政策制定、流行病学研究、新药研发、临床辅助决策与预后等医疗领域的各个方面。同时，医疗数据的可获得性、医学信息基础设施建设的缺失也是医疗数据科学未来发展所必须解决的紧迫问题。

城市交通的渗流模式

李大庆 (北京航空航天大学)

时间：11:15-12:00

简介：李大庆，以色列巴伊兰大学博士（最高荣誉毕业）。1982 年生，北京航空航天大学可靠性与系统工程学院研究员、博士生导师、北航首届校务委员会委员。国家优秀青年科学基金获得者。担任了中国系统工程学会系统可靠性专委会副秘书长、共同发起了中国管理科学与工程学会质量与可靠性管理研究会（筹），并担任中国优选法统筹法与经济数学研究会工业工程分会常务理事。

近年来，围绕复杂系统的可靠性管理，从系统故障的形成机理、系统故障的演化规律和系统故障的消解方法三方面研究复杂网络系统的故障规律和可靠性管理，以第一作者或通讯作者在 PNAS(3 篇)、Nature Physics、Nature Communications、RESS、Phys. Rev. Lett. 等国际著名期刊上发表研究成果；主持了包括国家自然科学基金，预研项目和预研重点基金等项目。指导研究生获得了北京市优秀毕业生、北航优秀毕业生等；研究生毕业后赴 MIT 等著名学府读博。

摘要：如果把城市看成一个生命体，交通网络就是其生命线。交通网络的健康运行是交通可靠性管理的核心问题。

同时，城市交通相关的位置服务是典型的大数据场景，目前的位置大数据计算依然面临着数据缺失、抽样偏差、计算复杂等瓶颈问题，导致较难出现达到应用级别的城市交通评估、预测和调控方法，成为目前智慧城市实施的主要困难之一。面对这些挑战，就需要基于现有积累的海量数据，挖掘相关的系统语义特征和演化行为，支撑更深层次的算法设计。近年来我们基于渗流理论，结合交通路况数据，对交通拥堵从产生、演化到恢复的全寿命周期进行分析，挖掘系统的弹性规律，希望可以为城市交通的可靠性管理提供新途径。

基于人工智能的在线自杀主动预防

朱廷劭 (中国科学院心理研究所)

时间：14:00-14:45

简介：朱廷劭，男，1971 年生。研究员，博士生导师，中国科学院“百人计划”学者。1993 年毕业于南京航空航天大学，分别于 1999 年中国科学院计算技术研究所和 2005 年加拿大 Alberta 大学获博士学位。

朱廷劭及其团队通过心理与信息科学的融合开展大数据心理学的交叉研究，实现了对用户心理特征的及时有效的识别，为心理学研究提供了新的思路。先后主持承担国家自然科学基金面上项目、科技部 973 和 863、国家社科基金重点、中科院 A 类先导专项等多项研究课题，发表论文 60 余篇。

摘要：自杀是一个比较严重的社会问题，目前针对自杀的干预或预防由于被动以及针对性不强，效果受到很大影响。我们提出利用社交媒体数据，基于人工智能技术，实现对自杀意念微博的自动识别，向用户发出帮助私信。在确保隐私保护的前提下，向用户提供多种的服务。自 2017 年 7 月正式上线值班以来，我们向 1.2 万余用户推送了帮助信息，得到了比较好的效果。通过机器学习方法，主动找到有需求的用户进行干预，提高了自杀干预的效率。

从模型驱动的集体推断到数据驱动的个体预测

吴喜之（中国人民大学）

时间：14:45-15:30

简介：吴喜之，中国人民大学教授，本科毕业于北京大学数学力学系，美国北卡罗来纳大学统计博士。在国内外出版的专著和教材二三十部；研究涉及的领域或方向包括：序贯分析及最优停时、回归诊断、模型选择、贝叶斯统计、非参数统计、分类数据分析、纵向数据分析、偏最小二乘方法、结构方程模型、时间序列、数据挖掘及机器学习等等；曾在加州大学伯克利分校、加州大学戴维斯分校、北卡罗来纳大学教堂山分校、北卡罗来纳大学夏洛特分校、密歇根大学等二十几所国内外大学任教。

摘要：传统统计是模型驱动的，其核心是基于以样本均值为中心的统计量对假定的总体参数进行推断，以“统计显著性”为主要工具。而数据科学是数据驱动的，主要关注的是对个体的预测，主要原则是可预测性、可计算性、稳定性。我们必须完成从模型驱动的总体推断到数据驱动的个体预测的转换。

个人信息保护与企业数据产权

张建民（北京大成律师事务所）

时间：16:00-16:45

简介：张建民，北京大成律师事务所合伙人、律师，北京大学法律硕士，狗熊会联合创始人、数据合规研究中心主任，北京大学对冲基金实验室研究员，北大-UC 伯克利个人信息与数据保护国际项目学员。主要从事私募股权基金、私募证券投资及衍生的争议解决业务，被誉为“懂 K 线的律师”。张律师是数据资产调查（数据 FTA）法律服务产品的首创者，对数据合规、隐私保护、数据治理有着长期、深入的研究。

摘要：随着大数据、云计算、人工智能技术的普遍应用，大到国家，小到企业，数据愈发成为竞争的核心资源与资产形态。但是因为数据的天然属性，也伴生了巨大的权利冲突和现实张力。如何在尊重和保护个人数据权利，特别是隐私权的前提下促进数据的有效和合规利用，保护数据资产，成为数据从业者和数据行业普遍面临的重要议题。本报告将分别从数据主体（个人）与数据经营者（企业）两个角度，介绍个人信息保护与企业数据资产保护的立法脉络、典型案例，以及背后的规则与风险。

神经系统疾病诊疗辅助决策与多层次会诊系统建设

王朝东（首都医科大学宣武医院）

时间：16:45-17:30

简介：王朝东，教授、主任医师、博士研究生导师。现任首都医科大学宣武医院神经内科遗传代谢专业主任、国家老年疾病临床医学研究中心办公室主任，兼任国家重点研发计划项目“主动健康与老龄化科技应对”重点项目专家组成员、中国医师协会老年病学分会委员、中国优生科学协会理事、中国老年保健协会脑保健专业委员会副主任委员、北京市医学会帕金森病与运动障碍分会委员、北京市医学会遗传学分会委员。长期从事神经系统疾病的临床诊疗及分子机制、精准诊治和大数据应用等方面研究。主持国家科技部重点研发计划项目子课题 1 项、国家自然科学基金面上项目 4 项、省级课题 6 项。在国内外专业杂志发表论文 50 余篇，其中 SCI 收录 30 余篇。获中华医学科技二等奖、北京市科学技术一等奖、江西省科技进步三等奖各一项。

摘要：中国各级医院、医师诊疗水平参差不齐，优质资源集中在大医院和少数专家手里，基层医院和大夫诊疗水平提高的渠道少。临床大量的误诊误治病例，均由于缺乏精准的专业知识和标准化的流程。日益复杂和细化的临床诊疗要求，给传统的经验医学提出了前所未有的挑战。由于中国医生少、病人多，医生没有时间和精力掌握完整的系统知识，也不可能规范地按照标准流程开展诊治。因此，迫切需要一套简单、实用的工具系统，辅助神经专科和非专科大夫在短期内处理大量的医学信息和快速做出正确的决策。本报告将围绕神经系统疾病的精准诊疗辅助决策与多层次会诊系统的建设方案和临床应用场景进行介绍。

狗熊会数据科学精品案例库简介

潘蕊 (中央财经大学)

时间: 09:00-09:30

简介: 中央财经大学统计与数学学院副教授, 北京大学光华管理学院博士。研究方向为网络结构数据的统计分析。在 JASA、ANNALS 等统计学期刊有多篇论文发表。著有《数据思维实践》一书。狗熊会联合创始人(水妈), 精品案例负责人。

摘要: 狗熊会由北京大学光华管理学院王汉生教授(熊大)创办, 致力于“聚数据英才, 助产业振兴”, 关注数据科学的人才培养、案例教学、企业实训等方面。在数据科学案例教学方面, 狗熊会最为成熟的产品即精品案例库, 目前已有近 30 所高校使用。本次报告将详细介绍狗熊会数据科学精品案例库的主要内容, 以及 2019 年狗熊会在深度学习教学案例方向的规划。

智能热表与供热效率提升

李季 (中央财经大学)

时间: 09:30-10:00

简介: 中央财经大学商学院市场营销系教授, 大数据营销专业负责人, 大数据与营销创新研究团队负责人, 金融大数据营销研究中心主任, 狗熊会高级研究员。毕业于北京大学光华管理学院市场营销系, 获得管理学博士学位, 曾在哥伦比亚大学商学院进行访问研究。研究方向集中于市场营销模型, 客户关系管理, 新媒体营销, 数据库营销等方面。主持多项国家自然科学基金项目以及教育部人文社会科学基金项目, 获得北京市高等学校“青年英才计划”资助。在学术研究方面, 关注新产品扩散模型、生存模型、口碑推荐和社会网络分析方法等。在应用研究方面, 关注以大数据挖掘和分析为基础的客户关系管理, 以及企业在新媒体环境下的营销活动。

摘要: 在传统供热行业, 热网的水力平衡尤其是住宅区内的二次网水力平衡基本停留在人工手动调节的状态, 因此二次热网水力失衡的情况普遍存在。热网水力失衡导致部分用户的室温过冷或过热, 舒适性差。智能热表除了具有热量监测作用外, 还可以将热量数据上传云端并根据模型测算进行远程供热调节, 以达到二次热网的水力平衡。智能热表的应用, 既能降低水、电、热的单耗, 又可以节省人力成本, 提供热企业的整体效益。同时, 对热表数据、室外环境数据和用户住宅数据的深入分析, 可以帮助供热企业更加全面的了解供热效果, 为改善供热效率提供依据。

小微企业支付交易数据的标签和模型方法

黄丹阳 (中国人民大学)

时间: 10:30-11:00

简介: 黄丹阳, 中国人民大学统计学院副教授, 北京大学光华管理学院博士, 狗熊会联合创始人。研究兴趣为超高维数据分析, 社交网络数据建模, 互联网征信数据分析等。研究论文发表于在 Journal of Econometrics, Journal of Business and Economic Statistics, Electronic Journal of Statistics, Statistica Sinica 以及管理世界等国内外权威杂志。在互联网征信领域具有丰富的实践及研究经验, 主持多项相关纵向科研课题, 曾与考拉征信有多年合作经验, 现兼任百行征信研究中心外聘专家。

摘要: 小微企业在我国经济发展中作用重大。由于缺乏抵押物, 缺少电子化记录和财报不健全等原因, 对于小微企业, 一般的企业征信和企业风控模型很难有用武之地。本报告将通过互联网时代, 小微企业支付交易数据及拓展数据源的介绍, 讲解可以应用于小微企业画像的标签和模型方法。

智慧零售场景中的标题体系设计及应用

王菲菲 (中国人民大学)

时间: 11:00-11:30

简介: 王菲菲, 中国人民大学统计学院助理教授, 北京大学光华管理学院统计学博士。研究上关注文本挖掘及其商业应用、数据挖掘算法、社交网络建模、贝叶斯分析等, 研究论文发表于 Statistics in Medicine、The American Statistician、World Wide Web Journal 等国内外权威杂志上。在产业实践上, 在标签体系设计与应用、个性化推荐、大规模文本数据建模等领域具有丰富的实战经验。热衷案例创作, 是微信公众号狗熊会精品案例系列的案例组长。

摘要: 人工智能技术的飞速发展为各行各业带来了前所未有的变革与机遇, 那“人工智能”+“零售”能碰撞出怎样的火花呢? 本研究即着眼于人工智能时代下智慧零售的新液态。通过将人脸识别技术运用在连锁便利店中, 获得顾客的身份属性特征, 并进一步打通消费数据和门店数据, 构造人(顾客)、货(消费)、场(门店)三位一体的标签体系。以此为基础建立丰富的用户画像, 准确判断用户需求, 并进一步实现商品、服务、广告的个性化推荐, 从而助力智慧商超, 达到增加销售、培养忠诚顾客群体的目的。

环保大数据应用场景及商业价值

吴睿 (西安欧亚学院)

时间: 11:30-12:00

简介: 西安欧亚学院副教授, 欧亚·狗熊会数据科学研究院副院长, 西安欧亚学院金融学院院长助理, 西安交通大学博士生, 陕西省劳动模范, 主要研究方向统计建模、商业数据分析, 近期主持并参与了工业锅炉大气污染物不确定性分析、企业排污趋势及超标预警分析、证券分析师的价值分析等项目。

摘要: 由于工业废气排放所造成的严重污染, 给人类的生产生活和环境带来了许多的危害, 因此国家要求企业对废气排放的污染, 采取建立安全的监测预警系统, 能够对废气进行有效的监测和预警, 使其减少对人类的身体健康和环境的危害与污染。但部分排污企业为了削减治污成本, 擅自调整污染物监测设备, 甚至一些环保企业协助数据造假, 致使污染物在通过治理设施后仍然超标排放, 使得环境监管工作更加棘手。本报告通过对企业排污口在线监测的历史数据进行分析, 建立模型, 对异常数据进行有效识别, 并尝试建立企业超标趋势的预警模型, 为企业合理安排生产提供参考。

数据治理国际动向与前瞻

吴沈括 (北京师范大学)

时间: 8:30-8:55

简介: 北京师范大学刑事法律科学研究院暨法学院副教授、硕士生导师。中国互联网协会研究中心秘书长。联合国网络安全与网络犯罪问题高级顾问。意大利维罗纳大学法学博士、博士后、助理研究员、访问教授。意大利都灵大学合同研究员。最高人民法院咨询监督专家。中共北京市委政法委法学专家库专家。QS 集团世界高校学术声誉调查专家库专家、欧洲刑法学权威期刊《Diritto Penale XXI Secolo》编委。已在欧洲以意大利文、英文等出版专著 1 部, 合著 2 部, 外文核心期刊发文近 20 篇, 完成中国刑法, 刑事诉讼法和刑事司法解释全文意大利语译本。在《清华法学》等 CSSCI 刊物发文近 20 篇; 在《光明日报》等核心报刊发文逾 20 篇。主持国家社科基金、中央网信办、教育部留学基金委、北京市社科联以及北京师范大学等多项科研项目。

摘要: 随着数据日益成为国际竞争与博弈的战略资源, 全球数据治理呈现更为复杂的发展态势, 以美国、欧盟为代表的国外数据治理方案各自反映其特有的数据规制观, 我国数据治理建设需要满足中国产业发展的时代需求, 在宏观的制度设计与微观的风控合规两个层面做出有效的安排。

京东数科个人信息保护实践

刘熙君 (京东数字科技)

时间: 8:55-9:20

简介: 刘熙君女士, 现任京东数字科技法律合规部法律总监, 负责公司数据合规体系建设。刘熙君女士有十几年互联网和高科技公司法务工作经验, 曾任职奇虎 360 公司法律风控总监, TCL 互联网事业部法律总监, 擅长搭建企业产品法律风险评估体系、产品规则和制度等合规风控系统。

摘要: 大数据的互联互通, 要求数据具有一定的开放性和互通性, 而作为重要数据来源的个人信息, 则涉及隐私保护和个人信息保护的规则, 如何在数据使用中协调与个人信息保护的关系, 成为企业业务合规的重点。本次主题就数据资产和法律合规方面的落地实践做一些探讨。

互联网平台个人信息保护观察

李玲 (南都个人信息保护研究中心)

时间: 9:20-9:45

简介:南方都市报记者、南都个人信息保护研究中心研究员。专注于互联网隐私领域的报道多年，曾参与撰写“700元买同事个人信息”、“安卓应用市场乱象”、“借贷App过度索取乱象”、“精(惊)准推送”等多篇调查报道。自2017年3月起，测评过两千多家网站和App隐私政策，参与编著《互联网企业隐私政策研究与实例》一书，撰写过《千家网站和App隐私政策透明度测评报告》、《热门应用隐私设计观察报告》等多份互联网平台隐私保护报告。

摘要:自2016年年底推出重磅调查700元买同事个人信息后，南都持续在隐私领域深耕，专门成立隐私护卫队聚焦该领域的日常报道，同时推出多份网络平台信息安全测评报告。基于南都在隐私保护方面的长期观察和研究，本次分享将围绕千家App隐私政策透明度测评、奇葩App过度索取权限典型案例、以及App隐私设计等方面，讲解互联网平台的个人信息保护情况。

侵犯网络隐私及个人信息案件的审理难点与审判趋势

陈昶屹 (北京市海淀区人民法院)

时间: 10:00-10:25

简介:陈昶屹，北京市海淀区人民法院中关村人民法庭庭长，中国人民大学民商法博士、中国社会科学院宪法与行政法博士后、民商法副研究员，中国信息与网络研究会理事，全国模范法官、北京市审判业务专家。从事审判领域及研究兴趣为网络侵权责任、知识产权保护、个人信息保护等，曾主审全国首例“被遗忘权”案件，北京大学诉邹恒甫侵犯名誉权案、北京人人车旧机动车经纪有限公司诉车好多旧机动车经纪(北京)有限公司不正当竞争案等案件，在隐私及个人信息保护领域具有丰富的审判经验及研究经验，主持及参与多项相关国家级科研课题，曾出版《网络人格权侵权责任研究》等著作及发表个人信息保护相关论文。

摘要:虽然我国民法总则已经以民事基本法的方式正式宣示了个人信息受法律保护的基本原则，但是我国在个人信息及隐私保护的具体规则上，并没有像欧盟GDPR那样详细而体系化的规定，而且在个人信息及隐私保护与促进网络及大数据产业发展之间价值衡量与利益平衡的问题更为突出，这对此类案件的审理带来了前所未有的挑战，也为中国法官在网络治理中贡献中国规则与中国智慧提供了千载难逢的机遇。

并购融资过程中的数据实务分享

赵中星（北京大成律师事务所）

时间：10:25-10:50

简介：赵中星律师的执业领域为网络安全和数据合规、外商直接投资、境外投资和公司事务。赵中星律师专注于为科技、媒体和电信行业客户的日常运营、投融资、行业监管和数据合规等方面提供法律解决方案。赵中星律师曾为全球知名电商企业、数据分析公司、网络营销公司、互联网和手机安全产品公司、互联网金融公司、区块链、共享单车平台等商业模式所涉及的诸多与数据相关的前沿法律和监管问题提供咨询和建议。赵中星律师也是汤森路透 Practical Law 数据库的特约作者之一，曾就网络安全和数据合规、外商投资、境投融资、互联网行业监管多个专题撰写律师实务指引文章。赵中星律师曾就读于香港城市大学、香港中文大学、英国伦敦大学，分别取得法律、法律翻译和同声传译学位。

摘要：数据自由应用（FTA）是指通过对数据进行调查和分析，查明数据的收集、存储、处理、跨境传输以及商业应用等过程中是否存在可能侵犯数据主体权利和违反其他法律规定的情形。数据自由应用的尽职调查，旨在确保企业可对获取的数据按照既定目的进行各类型商业应用。在本次分享中，讲者将会就与数据自由应用相关的基本概念和内涵、数据资产的法律属性、数据自由应用尽职调查的要点、投融资过程中的数据风险及应对等关键问题展开探讨。此外，讲者也会结合投资并购交易中的真实案例，分享如何基于数据自由应用尽职调查的结果，并结合协议条款设置，有效预防和降低数据类资产的民事、行政、刑事风险。

打通数据孤岛——基于可信硬件的安全数据处理

夏虞斌（上海交通大学）

时间：10:50-11:15

简介：夏虞斌，上海交通大学副教授，主要研究方向为系统与数据安全，尤其是可信执行环境在数据安全计算的应用。在操作系统领域 USENIX ATC、EuroSys、FAST、MobiSys，体系结构领域 ISCA、HPCA，计算机安全领域 CCS、USENIX Security、NDSS 等著名国际学术会议发表论文。领导团队开发了基于 ARM TrustZone 的移动安全操作系统 T6，获得了第十四届“挑战杯”全国大学生课外学术科技作品竞赛特等奖，目前已部署在上亿部设备中。

摘要：随着大数据和人工智能等应用的不断增长，“数据产生价值”、“数据即资产”等理念已得到广泛共识。融合的数据比分散的数据产生更大的价值；然而，由于缺乏安全可信的通用多方数据处理技术，掌握用户数据的企业和机构出于自身利益和法律风险的考虑，往往严格控制其拥有的用户数据，极少与其他机构进行数据共享，在客观上导致“数据孤岛”问题。如何在数据不泄露的前提下，处理来自多个数据源的数据以产生更大的价值，是一个亟待解决的问题。本次报告将会介绍一种对数据的可信处理方法，利用体系结构扩展支撑上层应用新特性，对个人数据进行安全的处理与融合。

圆桌讨论

圆桌讨论

时间：11:15-12:00

R 语言在宏基因组数据分析和可视化中的应用

刘永鑫 (中国科学院遗传发育所)

时间: 08:30-09:00

简介: 中科院遗传发育所工程师, 生物信息学博士, 主要研究方向为宏基因组数据分析, 已在 Science、Nature Biotechnology、Plant Cell、Genomics Proteomics Bioinformatics、Science China Life Sciences 等期刊发表论文十余篇, 微信公众号“宏基因组”主创。

摘要: 近年来 R 语言在生物医学领域应用快速发展, 大多数高质量文章图表采用 R 语言进行可视化, 其中 Bioconductor 网站已发布 1649 个生物领域 R 包, 极大的推动了 R 语言在生物学中的应用。本人创立了“宏基因组”公众号, 推广 R 语言在宏基因组中统计和可视化的应用, 有超 4.1 万同行关注。

以 2019 年 4 月发表于《Nature Biotechnology》的一篇宏基因组学文章为例, 从统计分析方法和可视化的图表类型选择两个方面进行解读。报告的主要内容包括常用组间统计方法、主坐标分析、随机森林分类等分析方法; 可视化的方案有采用地图展示实验设计、箱线图 + 散点 + 统计展示多样性和实验数据、堆叠柱状图展示物种组成、物种树可视, 以及热图、维恩图等可视化的具体应用。

基于本示例, 带动同行进一步了解 R 语言在宏基因组数据统计分析、结果可视化过程中的套路和具体应用, 推动 R 语言在生物学中的应用和普及, 倡导科学领域研究过程方法透明共享和成果的可重复计算。

从学术到业界的 3 种可视化设计探索与实践

余政彦 (奇安信)

时间: 09:00-09:30

简介: 奇安信(原名 360 企业安全)可视化产品经理, 美国东北大学可视化设计与人民大学经济双硕士, 专注于数据可视化分析与可视化产品设计, 微信公众号“VisIt 有视没事”联合主创。

摘要: 在可视化设计中, 有哪些探索与实践的可能? 从个人经验出发, 分享可视化设计在学术与中美商业运用的几个例子。在学术中, 探寻可视化设计新的表现样式, 除了让数据更为美观, 更追求有效的跟用户传达信息; 毕业后的美国德勤咨询工作, 针对客户需求来设计后台可视化产品, 将可视化用于解决特殊业务问题; 回到中国后, 目前在奇安信负责图表规范的制定, 规范用于提升设计师与工程师在大屏的产出效率。从这三份看似毫无联系的经历, 串联出一个可视化设计的思路, 在可视化设计领域, 提供一点个人心得体会。

D3 财务可视化: 财报呈现一目了然

钟永剑 (北京数可视科技有限公司)

时间: 09:30-10:00

简介: 毕业于北京理工大学计算机科学与技术专业, 现担任北京数可视科技有限公司技术部总监, 负责数可视技术部研发工作, 研发的主要项目和产品包括 EVA 财务可视分析工具和 HANABI 数据可视化工具等。

摘要: 能够支持大数据集, 动态交互和动画效果。以专业的数据分析模型, 用图形视觉展现企业财报核心信息, 准确高效解读财报同时降低财报分析门槛。

数据可视化与反 p-hacking: 以政治科学为例

胡悦 (清华大学)

时间: 10:30-11:00

简介: 清华大学政治学系助理教授, 美国爱荷华大学 (University of Iowa) 政治学博士。主要研究领域为实证政治文化、语言政策和政治学方法论, 部分研究成果已经在 Journal of Politics、Chinese Sociological Review、Social Science Quarterly 等期刊上发表。方法研究涉及实验室和调查实验、大数据文本分析、网络分析、空间分析、数据可视化等, 并为 ‘interplot’、‘dotwhisker’ 等 R packages 的主要研发者和维护者, 下载量在全球范围内已达数万次。

摘要: 反 p-hacking 运动是近年政治科学乃至社会科学的热点议题。该运动旨在反对传统上以 $p = 0.05$ 作为评价社会科学研究结果质量与发表可能的“一刀切”标准。反 p-hacking 的一项重要措施是透明清晰地表达数据结果及其中不可确定性。dotwhisker 和 interplot 通过将分析结果总结 (summary)、不确定性模拟 (simulation) 和可视化 (visualization) 集成一站式方程, 将这一过程简化, 为社会科学常见分析结果提供简单有效的可视化工具, 避免如 p-hacking 等对分析结果的误读、误解。同时, 两软件均基于 ggplot2 预留足够的定制空间, 以满足不同情况下结果报告需要。本讲将系统介绍 dotwhisker 和 interplot 的使用流程和操作技巧, 并基于已发表学术文章进行案例演示

新媒体是如何做可视化报道的?

吕妍 (澎湃新闻)

时间: 11:00-11:30

简介: 澎湃新闻数据新闻主编, 负责澎湃新闻原创的美数课栏目和 PGC (全称: Professionally Generated Content) 为主的有数栏目; 负责策划的《海上鱼荒》、《汶川记忆地图》、《活在临界线上》、《数说相亲角》等项目, 曾获得美国新闻设计协会 (SND)、亚洲出版协会 (SOPA), 英国信息之美 (Information Is Beautiful)、腾讯传媒赏等奖项。

摘要: 上世纪五六十年代在美国开端的计算机辅助报道, 和历史更为悠久的报纸制图传统, 在互联网时代融合出数据新闻与可视化的这一新颖分支, 也成为媒体转型的重要抓手之一。在这一方面, 国内外媒体有哪些新鲜尝试? 是如何利用数据可视化讲故事的?

数据可视化的色彩运用原理与实践

张杰 (香港理工大学)

时间: 11:30-12:00

简介: 香港理工大学助理研究员, 著有《R 语言数据可视化之美》(印刷中)、《Excel 数据之美》以及 14 篇 SCI(E) 和 SSCI, 微信公众号“EasyCharts”联合主创。

摘要: 颜色对数据可视化的审美尤为重要。本次演讲将先介绍不同的颜色空间系统, 包括 RGB(红 (Red)、绿 (Green) 和蓝 (Blue))、HSL(色相 (Hue)、饱和度 (Saturation)、亮度 (Lightness))、HSV(色相 (Hue)、饱和度 (Saturation)、色调 (Value))、CIELUV、CIELCH、HSLuv 等颜色空间系统; 然后讲解了颜色主题方案的设计原理, 包括单色渐变系、双色渐变系、多色系三种类型; 接着介绍了 R 语言和 python 常用的颜色主题包; 最后介绍了几种不同的色彩应用案例。

R 语言在面试考官多级评分中的应用

李英武（中国人民大学）

时间：08:30-09:00

简介：李英武，中国人民大学心理学系副教授，目前主要从事高级心理测量学理论与 R 语言在心理测量中应用，个体认知能力增龄化过程影响及其心理健康影响机制，项目反应理论，大规模考试中的主观偏差（Bias in Personnel Selection），多水平理论（Multi-level theory research）等心理测量等方面研究。

摘要：本研究基于多面 Rasch 模型，使用 R 语言 immer 程序包对 4 组共 50 名面试者、20 名面试官在结构化面试中评定的成绩进行分析，追踪模型输出的异常值，并根据异常值剔除了由于面试官等具体测量情境因素引入的误差对原始分数的影响，得到面试者的能力估计值以及个体水平的评分者一致性信息。结果表明，多面 Rasch 模型的拟合度符合可接受的范围，面试结果是可被接受的；面试官的内部一致性水平总体较好；评分量尺的标准尚待修正；面试官的内部一致性水平和面试官彼此之间在严厉程度上的差异都可能对面试者的评定产生影响。本研究证实了用多面 Rasch 模型得到面试者能力估计值并以此作为决策的依据可以提高选拔的有效性、准确性和科学性。以后的研究中，还可以使用多面 Rasch 模型得到面试者个体层次的评分者一致性指标，不同侧面的偏差分析，如：面试官与面试者的偏差分析等，提升对面试误差来源的定位并给出详细的诊断信息。

认知诊断模型信息矩阵估计软件包 dcminfo 的开发与应用

刘彦楼（曲阜师范大学）

时间：09:00-09:30

简介：刘彦楼，博士，硕士生导师。曾发表多篇关于项目反应理论与认知诊断相关的论文，并进行认知诊断模型信息矩阵估计包 dcminfo 的开发工作。

摘要：认知诊断模型（Cognitive Diagnosis Model, CDM），又称诊断分类模型，是心理计量研究领域的热点之一。认知诊断模型参数估计值的信息矩阵在 CDM 研究中具有重要的理论及应用价值，如模型参数估计值标准误及置信区间计算、有限信息拟合统计量计算、项目水平上的模型比较，项目功能差异检验等。R 语言开源软件包 dcminfo (Liu & Xin, 2017) 开发的主要目的是估计 CDM 的期望信息矩阵、经验交叉相乘信息矩阵、观察信息矩阵以及三明治矩阵。将通过技术演示以及实证数据分析操作的方式展示如何通过 dcminfo 计算以上提及的四种矩阵，以及 dcminfo 在模型参数估计值标准误及置信区间 (Liu, Xin, Andersson, & Tian, 2019)、有限信息拟合统计量 (Liu, Tian, & Xin, 2016)、项目水平上的模型比较 (Liu, Andersson, Xin, Zhang, & Wang, 2018) 等研究中的具体应用。

基于潜混合模型的 4PLM 参数估计 BE3M 算法及其 R 实现

郭少阳 (华东师范大学)

时间: 09:30-10:00

简介: 郭少阳, 美国伊利诺伊大学香槟分校教育心理学硕士, 华东师范大学课程与教学系博士研究生。研究方向: 游戏化教育测评、项目反应理论、参数估计。

摘要: 在教育和心理测量领域, 四参数逻辑斯蒂克模型 (4-parameter logistic model, 4PLM) 能够平衡被试的随机作答 (或猜测) 偏差和社会赞许 (或失误) 偏差, 更为准确的反映被试的能力。然而, 由于 4PLM 项目参数估计的难度较大, 使其一直难以在中小样本量下使用。文章以被试的能力为潜类别变量, 使用贝叶斯三次期望最大算法 (Bayesian Expectation-Maximization-Maximization-Maximization, BE3M), 开发了基于 R 环境的 4PLM 项目参数估计程序 BE3M。模拟研究和实例数据的分析显示, (1) BE3M 算法兼具了 EM 算法和贝叶斯方法的优势, 能够以 EM 算法的执行时间得到与 MCMC 方法相当的估计精度; (2) 以传统 BEM 相比, BE3M 的估计结果受先验信息的限制和负面影响较少, 估计结果更为稳健。

心理测量理论在 R 语言上的实现

薛明锋 (北京师范大学)

时间: 10:30-11:00

简介: 薛明锋, 华南师范大学应用心理学学士, 目前就读于北京师范大学心理测量专业硕士, 发表 SCI 文章一篇, 若干其余文章, 有多年使用 R 语言的经历, 有多次科研项目经历。

摘要: 主要介绍经典测量理论、概化理论和项目反应理论的基本概念、假设和重要参数。在此基础之上, 展示如何在 R 上实现。分析 R 实现心理测量理论的优势以及劣势。

基于项目反映理论在考试中的测验校准

张雪儿 (中国人民大学)

时间: 11:00-11:30

简介: 张雪儿, 就读于中国人民大学心理学系, 跟随李英武副教授课题组进行心理测量以及工业组织心理学方向的研究, 参与编制《现代心理测量》等教材。

摘要: 测验校准是对考生实施测验, 并对考生对项目的反应进行二分法评分。然后, 将数学程序应用于项目反应数据, 以创建一个对特定的测验项目和应试者来说是唯一的能力分布, 传统的测验校准方法因迭代过程复杂而较难实现, 基于项目反映理论 (IRT) 的测验校准, 使用 R 软件可将不同阶段用最大似然法进行的迭代过程转化为程序实现, 从而便捷的产生有效结果, 为项目测验的结果提供参考框架, 本研究基于一个真实的考试数据集, 使用 R 软件进行 IRT 理论下的测验校准, 得出考生难度区分度和项目难度的精确分布。

基于贝叶斯估计的结构方程模型介绍

张沥今（中山大学）

时间：11:30-12:00

简介：张沥今，中山大学心理学系本科生，保送本系硕士研究生，师从潘俊豪副教授，研究领域为贝叶斯结构方程模型的统计分析及其应用。论文《贝叶斯结构方程模型及其研究现状》已被《心理科学进展》录用，曾在第 20 届全国心理学学术会议、第十三届海峡两岸心理测量与教育测验学术研讨会中以《含有序分类数据的贝叶斯 Lasso 因子分析模型》做分组口头报告，论文《渐近测量不变性中先验方差的选取》摘要已被 2019 年国际心理测量研讨会接收为分组口头报告。

摘要：在心理学研究中结构方程模型被广泛用于检验潜变量间的关系，其估计方法有频率学方法（如，极大似然估计）和贝叶斯方法两类。而传统的频率学派方法对模型施加的限制往往过于严格，这种限制在大样本情况下很容易拒绝实际上和数据拟合良好的模型。在传统方法中为了解决这种限制带来的问题，研究者通常会结合理论和修正指数的建议，在模型中增加交叉载荷或残差相关。但是这种基于修正指数的方法很容易受到研究者主观选择的影响，容易导致一类错误率的增大和模型的过拟合，削弱其泛化能力。贝叶斯结构方程模型通过结合先验信息可以较好地解决上述问题，此外，在模型识别和拟合、参数估计、处理复杂模型和小样本情况等方面贝叶斯方法都有着更好的表现，能够更好地满足应用研究者在实证研究中的需求，但其在国内心理学领域的应用不足。本次报告将详细介绍贝叶斯结构方程建模的原理和优势，并通过实例分析与传统估计方法进行深入对比，展示贝叶斯建模的分析步骤和评价标准，希望能够为大家带来新的结构方程建模思路，解决采用传统方法建模时难以克服的问题。

Two-Way Partial AUC and Its Properties

杨翰方 (中国人民大学)

时间: 09:00-09:30

简介: 杨翰方, 副教授, 博士生导师。现就职于中国人民大学统计学院, 隶属经济与社会统计教研室, 中国人民大学统计与大数据研究院师资团队成员。2007 年毕业于同济大学数学与应用系, 获得理学学士学位, 2012 年先后获得美国佐治亚州立大学风险管理硕士以及统计学博士学位。回国后在中国人民大学统计学院任讲师, 2016 年晋升副教授, 2017 年获得博士生导师资格。

摘要: Simultaneous control on true positive rate (TPR) and false positive rate (FPR) is of significant importance in the performance evaluation of diagnostic tests. Most of the established literature utilizes partial area under the receiver operating characteristic (ROC) curve with restrictions only on FPR, called FPR pAUC, as a performance measure. However, its indirect control on TPR is conceptually and practically misleading. In this paper, a novel and intuitive performance measure, named as two-way pAUC, is proposed, which directly quantifies partial area under the ROC curve with explicit restrictions on both TPR and FPR. To estimate two-way pAUC, we devise a nonparametric estimator. Based on the estimator, a bootstrap-assisted testing method for two-way pAUC comparison is established. Moreover, to evaluate possible covariate effects on two-way pAUC, a regression analysis framework is constructed. Asymptotic normalities of the methods are provided. Advantages of the proposed methods are illustrated by simulation and Wisconsin Breast Cancer Data. We encode the methods as a publicly available R package tpAUC.

An integrative sparse boosting analysis of cancer genomic commonality and difference

孙怡帆 (中国人民大学)

时间: 09:30-10:00

简介: 孙怡帆, 女, 中国人民大学统计学院副教授、博士生导师、概率论与数理统计系主任, 全国工业统计学教学研究会理事。主要从事机器学习理论与算法, 高维数据分析等领域研究。承担国家自然科学基金、教育部人文社科基金等科研项目五项。在 Statistics in Medicine, Physical Review X, Scientific Reports, Physical Review E, 统计研究等国内外期刊发表研究论文二十余篇。

摘要: In cancer research, high-throughput profiling has been extensively conducted. In recent studies, the integrative analysis of data on multiple cancer patient groups/subgroups has been conducted. Such analysis has the potential to reveal the genomic commonality as well as difference across groups/subgroups. However, in the existing literature, methods with a special attention to the genomic commonality and difference are very limited. In this study, a novel estimation and marker selection method based on the sparse boosting technique is developed to address the commonality/difference problem. In terms of technical innovation, a new penalty and computation of increments are introduced. The proposed method can also effectively accommodate the grouping structure of covariates. Simulation shows that it can outperform direct competitors under a wide spectrum of settings. The analysis of two TCGA (The Cancer Genome Atlas) datasets is conducted, showing that the proposed analysis can identify markers with important biological implications and have satisfactory prediction and stability.

Properties of Covariate-Adaptive Randomization with Misclassification in Covariate

马维 (中国人民大学)

时间: 10:30-11:00

简介: 2009 年毕业于浙江大学数学系, 2013 年毕业于美国弗吉尼亚大学并取得统计学博士学位。博士毕业后曾任职于世界知名制药企业研发部门担任资深生物统计学家。于 2017 年加入中国人民大学统计与大数据研究院并担任助理教授。研究兴趣包括生物统计、临床试验设计、健康医疗大数据等。

摘要: Covariate-adaptive randomization is extensively used in clinical trials to balance treatment allocation over covariates, which is often viewed as an essential component in ensuring valid treatment comparisons. Most randomization procedures and tests in clinical trials are based on the assumption that the covariates are measured accurately. However, in practice, measurement error is inevitable, and it will cause misclassification in covariate. Under covariate-adaptive randomization, the impact is tied to both randomization and analysis. It is unclear how the misclassification affects the treatment assignment and the corresponding statistical inference. In this talk, we study the impact of covariate misclassification on the properties of covariate-adaptive randomization. In particular, we show that, if there is misclassification, the within-stratum imbalance is no longer bounded in probability and the two sample t-test is still conservative. Numerical studies are also performed to assess the finite sample properties.

Evaluation of driving risk at different speeds

高光远 (中国人民大学)

时间: 11:00-11:30

简介: 高光远为中国人民大学统计学院讲师, 毕业于同济大学 (学士) 和澳洲国立大学 (博士)。目前的研究方向包括 UBI 车险定价, 贝叶斯准备金评估模型等。已在 ASTIN Bulletin, Scandinavia Actuarial Journal, Insurance: Mathematics and Economics, 保险研究等精算期刊上发表多篇论文, 并著有一本 Bayesian Claims Reserving Methods in Non-life Insurance with Stan: An Introduction (Springer)。

摘要: Telematics car driving data describes drivers' driving characteristics. This paper studies the driving characteristics at different speeds and their predictive power for claims frequency modeling. We first extract covariates from telematics car driving data using K -medoids clustering and principal components analysis. These telematics covariates are then used as explanatory variables for claims frequency modeling, in which we analyze their predictive power. Moreover, we use these telematics covariates to challenge the classical covariates usually used in practice.

物联网数据安全一致性与分布式账本技术

熊志敏 (*IOTA*)

时间: 09:00-09:30

简介: 熊志敏 (Jimmy Xiong), IOTA 中国社区负责人, 卓尘实验室创始人, IOTA 国际网络 IEN 成员。2016 年创办 IOTACHINA.COM 中国社区网站, 负责 IOTA 中国社区的布道推广工作。熊志敏先生于 2009 年博士研究生毕业于中科院, 目前从事分布式账本技术的研究, 投资及在物联网等行业的应用推广工作。

摘要: 我们将介绍在物联网机器经济时代海量数据的安全一致性, 数据存储, 数据交易, 数据应用等问题。物联网数据领域是分布式账本技术, 比如区块链技术的一个重要应用方向。IOTA 是新一代物联网机器经济分布式账本, 其设计为物联网数据进行转移交互和价值交易结算, 从而构建未来机器经济蓝图, 具有转账无需手续费, 良好的扩展性等特点。我们在本报告中将分享车联网, UBI 保险, 以及供应链等多个应用案例。

自动驾驶与车联网数据应用

李旭 (*Momenta*)

时间: 09:30-10:00

简介: 曾任北京车网互联科技有限公司常务副总裁, Momenta 产品总监; 毕业于北京大学光华管理学院, 北京大学商务智能研究中心行业专家, 已获车联网领域多项发明专利, 曾深度参与公司上市 (IPO) 及上市公司非公开发行项目的设计规划实施; 参与编著商业案例图书《找我》系列;

摘要: 我们日常工作中经常提到来自汽车的大数据, 我们的很多算法也依赖基于大数据的训练来提升性能, 伴随越来越多的车联网数据被生产和收集, 并在包括自动驾驶、保险、汽车制造、汽车金融、网约车等领域产生新的应用, 期待与各位一起分享与探讨, 我们如何发掘和转化浩繁车联网数据中的巨大价值。

新能源汽车大数据与售后服务市场的融合应用

刘鹏 (北京理工大学)

时间: 10:30-11:00

简介: 刘鹏, 工学博士, 北京理工大学副教授, 硕士生导师, 新能源汽车大数据联盟副秘书长。长期从事电动汽车动力电池系统应用理论, 运行管理与安全控制理论以及新能源汽车大数据分析理论等研究工作; 主持和参与国家 863 计划项目、北京市科技计划项目等 10 余项, 建立了国家和北京市新能源汽车监管平台; 获北京市科学技术奖三等奖、中国汽车工业科学技术一等奖各 1 项; 出版编著 5 部; 发表学术论文 26 篇, 其中 SCI 收录 5 篇, EI 收录 15 篇; 授权发明专利 5 项, 软件著作权 8 项, 申请发明专利 10 项; 主持参与国家标准 6 项、行业标准 2 项。

摘要: 随着国家大数据战略的推进, 大数据与新能源汽车产业的融合发展持续升温。目前, 大数据技术在新能源汽车安全监控、故障诊断、维护管理、回收利用、健康状态估计、残值评估等方面都具有极大的潜力, 在二手交易、UBI 保险等汽车后市场也将发挥极大的作用。数据创造价值, 而平台是数据的载体。本报告将结合演讲者多年来在新能源汽车国家监管平台以及国家大数据联盟的工作经验, 分享新能源汽车大数据发展现状, 汇报平台基本情况, 探讨新能源汽车大数据技术在安全预警、动力电池价值评估、梯次利用、售后服务等方面的应用与成果。

室内定位技术在航空制造业车间的落地应用

武坚利 (中国航空综合技术研究所)

时间: 11:00-11:30

简介: 就职于中国航空综合技术研究所, 现任软件产品事业部部长, 北京航空航天大学机械电子工程专业硕士。长期从事航空装备管理标准化、数字化、信息化、集成化工作, 获软件著作权 10 余项, 有丰富的装备制造业车间信息化总成经验。目前主要依托融融军民产业互联网平台, 面向高端装备制造业, 打造工业软件生态圈, 连接企业, 为企业赋能。

摘要: 面向航空装备制造业多品种、小批量、离散化特点, 分析航空装备制造企业的管理现状及痛点, 结合实际案例, 介绍航空制造业车间室内定位技术的应用情况, 以及基于室内定位带来的显著质效提升。

Modeling Tunnel Profile in Presence of Coordinate Errors A Gaussian Process Based Approach

张晨 (清华大学)

时间: 09:00-09:30

简介: Zhang Chen (张晨) is an Assistant Professor in Industrial Engineering, Tsinghua University. She received her B.Eng. degree in Electronic Science and Technology (Optics) from Tianjin University in 2012, and her Ph.D. degree in Industrial Systems Engineering from National University of Singapore in 2017. Her research interests include developing methodologies and algorithms for complex or large-scale systems with multivariate or high-dimensional data, including intelligent sampling and sensing for data collection, data mining and information extraction for system modeling, and on-line monitoring and efficient anomaly detection for streaming data.

摘要: This talk presents a Gaussian process (GP) based approach to model a tunnel's inner surface profile with point cloud data provided by Terrestrial Laser Scanner (TLS). We introduce a reading-surface profile which uniquely determines a three-dimensional tunnel in a Cartesian coordinate system. This reading surface transforms the cylindrical tunnel to a two-dimensional surface profile, hence allowing us to model the tunnel profile by GP. To account for coordinate errors induced by TLS, we take repeated measurements at designed coordinates. We apply Taylor approximation to extract mean and gradient estimations from the repeated measurements, and then fit the GP model with both estimations to obtain a more robust reconstruction of the tunnel profile. We present a case study to demonstrate that our method provides a more accurate result than the existing cylinder-fitting approach and has great potential for deformation monitoring in presence of coordinate errors.

Dynamic Management of Autonomous Electric Taxis using Reinforcement Learning Method

唐昕迪 (清华大学)

时间: 09:30-10:00

简介: Xindi Tang received her bachelor's degree from Department of Automation, Tsinghua University in 2012. She is now a 3rd year Ph.D. student in Department of Industrial Engineering, Tsinghua University. Her research interests lie on planning and operation strategies of urban electrical transportation system as well as intelligent transportation management.

摘要: Electrification and automatization are two tendencies of intelligent transportation. It is foreseeable that autonomous electric taxi will play an important role in the future. However, the travel pattern of autonomous vehicles are different from traditional ones, not mentioning there is extra charging need for electric vehicles. In this study, we propose a reinforcement learning method to dynamically assign vehicles to customers, which tackles the curse of dimensionality in traditional optimization model. We conduct numerical examples to verify effectiveness of proposed method.

Modeling and Change Detection for Count-weighted Multi-layer Networks

董航 (清华大学)

时间: 10:30-11:00

简介: 董航, 本科毕业于清华大学工业工程系, 现清华大学工业工程系博士四年级在读, 研究方向为网络数据的统计建模与监控。

摘要: In a typical network with a set of individuals, it is common to have multiple types of interactions between two individuals. In practice, these interactions are usually sparse and correlated, which is not sufficiently accounted for in the literature. We propose a multi-layer weighted stochastic block model (MZIP-SBM) based on a multivariate zero-inflated Poisson (MZIP) distribution to characterize the sparse and correlated multi-layer interactions of individuals. A variational-EM algorithm is developed in order to estimate the parameters in this model. We further propose a monitoring statistic based on the score test of MZIP-SBM model parameters for change detection in multi-layer networks. The proposed model and monitoring scheme are validated using extensive simulation studies and the case study from Enron email network.

R 在某工厂自动化制造产线质量监控中的应用

刘心广 (质瑞信息)

时间: 11:00-11:30

简介: 刘心广, 质瑞信息 CTO, 中国科学院博士, 副教授、高级工程师职称; 持有高校教师资格证, IEEE 会员, 美国质量协会 ASQ 高级会员。擅长机器学习和人工智能算法研究, 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。曾在杭州电子科技大学、聚光科技、艾维思通讯技术、OCLARO、II-VI 研发中心等从事电子信息领域数据分析、解决方案和智能软件产品研发, 目前负责工业领域大数据分析、智能算法和应用整合的数字化一体解决方案开发。

摘要: 某 3C 制造工厂在产品的自动化生产和组装过程中, 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, 产品质量相关数据分散在各个局部点位。本项目基于工厂业务应用场景需求, 通过数据的自动采集、互联互通和质量工程技术分析, 构建了三个层次的在线监控与报警系统, 对工厂制造质量进行实时数字化管理, 并在工厂内私有化部署上线, 是 R 语言在生产环境中的一个应用案例。

基于电商评论数据的产品及服务质量在线监控

梁巧 (清华大学)

时间: 11:30-12:00

简介: 梁巧, 清华大学工业工程系三年级博士生, 研究方向包括制造及服务过程中的统计建模和数据分析, 特别是基于文本类数据的统计过程控制。

摘要: 随着电子商务的兴起和快速发展, 人们逐渐习惯从网上购买商品和服务, 电商平台每天都会产生大量的用户评论, 这些评论直接反映了消费者的关注点和喜恶情况。与单一的打分形式相比, 文字评论涵盖的维度更广, 内容更丰富, 成为评估和监控网络商品和服务质量的有力工具。在评论文字中, 往往包含两种类型的信息: 一是“主题”, 它主要对应用户关注的商品或服务维度, 比如商品的各个质量特性; 二是“情感”, 反映了用户对商品或服务的好感度。这两部分信息结合起来, 则反映了用户感知到的商品和服务质量。已有的产品质量监控研究, 往往针对的是制造过程本身产生的波动和异常, 所用到的数据类型也大多是连续型、分类型数据, 本研究则从产品全生命周期的角度出发, 在售后阶段入手, 利用用户反馈的评论数据进行产品质量的监控, 为我们实现质量管理提供新的思路。

基于简单空间插值和病例 - 交叉设计探讨大气污染对哮喘死亡的急性影响

刘跃伟 (中山大学)

时间: 09:00-09:30

简介: 刘跃伟, 博士, 中山大学公共卫生学院流行病学系副教授。2001-2011 年间在华中科技大学获得预防医学本科、劳动卫生与环境卫生学硕士和博士学位。曾在湖北省疾病预防控制中心、美国疾病预防控制中心从事环境与健康相关科研和项目工作, 入选湖北省青年科技晨光计划、湖北省青年英才开发计划, 主持/参与十余项国家、省部级科研项目, 以第一/通讯作者在 Am J Respir Crit Care Med、Environ Sci Tech、Am J Epidemiol、Epidemiology、Hum Reprod、Environ Int 等杂志发表学术论文 20 余篇。目前是湖北省预防医学会环境卫生专业委员会副主任委员、BMC Public Health 副主编。

摘要: 我国是世界上大气污染最为严重的国家之一, 大气污染对健康的危害已受到政府、公众和学者的广泛关注。研究发现, 大气污染暴露可引发哮喘症状, 导致哮喘加重和就医行为增加, 但其是否增加因哮喘死亡的风险尚不清楚。本研究通过收集湖北省空气质量和死因监测数据, 采用病例 - 交叉研究设计, 基于研究对象 (N=4454) 家庭住址和简单空间插值进行暴露评估, 利用条件 Logistic 回归模型等定量评估 PM2.5、PM10、SO₂、NO₂、CO 和 O₃ 等大气污染物短期暴露对哮喘死亡的影响, 同时分析性别、年龄、季节等因素的效应修饰作用。数据处理和分析均通过 R 实现, 使用的软件包主要包括: data.table, doParallel, baidumap, geoChina, sp, raster, rgeos, survival, ggplot2 等。

基于高维 miRNAs 表达谱数据的疾病诊断标记物筛选研究

郑德强 (首都医科大学)

时间: 09:30-10:00

简介: 郑德强, 博士, 首都医科大学公共卫生学院流行病与卫生统计学系讲师, 研究方向包括医学高维变量筛选、交互作用分析、慢性病生存分析与风险评估、孟德尔随机化分析、统计学习。2016 年 7 月获北京大学数理统计专业博士学位, 获省部级科技进步二等奖 1 项, 发表学术论文 30 余篇, 其中在 JCEM、Environmental Pollution、JAHA 等期刊发表 SCI 论文 10 余篇, 现主持国家自然科学基金项目 1 项。

摘要: 食管癌是我国高发恶性肿瘤之一, 90% 以上病例为食管鳞癌。目前, 许多食管癌患者确诊时已进展至中晚期, 晚期食管癌总体 5 年生存率不足 15%, 食管癌的早期诊断尤为重要。内镜下食管活检为我国现阶段食管癌的有效筛查方法, 但是只有在患者有明显症状和组织有明显病变后诊断效能最高。新的无创性标记物筛选对于食管癌的早期诊断尤为重要。本研究基于一个食管鳞癌的病例对照高维 microRNAs 表达谱数据, 使用多种特征选取和统计学习的联合方法, 筛选出 3 个诊断性能最高的 miRNAs, 利用随机交叉验证方法对 3 个 miRNAs 的组合诊断性能进行了评价, AUC 和预测准确率超过 0.80 及 0.79, 对于不同期别食管癌与正常人的诊断 AUC 超过 0.76, 研究结果为确定食管鳞癌诊断的新型生物标记物提供了基础, 研究中的统计方法为其他癌症、疾病利用高维基因数据筛选生物标记物提供了参考。本研究的高维特征选取和统计学习主要基于 R 软件多个 Package 实现, 本报告将展示如何在 R 中实现相关统计分析和研究结果。

An R package to explore the effects of environmental factors on infectious diseases

张兵 (中山大学)

时间: 10:30-11:00

简介: 本科和研究生毕业于华中科技大学同济医学院, 现博士就读于中山大学公共卫生学院(深圳)。先后参加多次 R 语言大会并做报告。研究兴趣为传染病传播规律。个人主页为 www.spatial-r.com.

摘要: 环境因素是影响传染病季节性特征最为重要的因素之一。相比于慢性疾病, 传染病的非独立性(t 时点感染病例数会影响 $t+1$ 时点人群的感染风险)以及隐性感染性(监测系统所观测到病例只是全部病例的一部分), 使得传统方法在探究环境因素与传染病发病关系时候存在众多悖论。环境因素可作用于病原体和影响宿主的易感性, 也会在一定程度上影响宿主感染后出现的临床表现, 进而影响其就诊行为。本研究以流行性腮腺炎为例, 通过传染病动力学方法, 分别将环境因素(温度、相对湿度和绝对湿度)作用靶点嵌套在传播过程(transmission process)和报告过程(reporting process), 并依托 iterated filtering 算法求解各自参数值及其可能得阈值范围。此方法可在一定程度上区分环境因素可能的作用靶点(传播过程还是报告过程)及解释环境因素对于传染病季节性影响存在区域异质性的原因。整个分析方法及结果展示都集成在 EFRID 程序包中(<https://github.com/Spatial-R/EFRID>)。

传染病动态传播模型的 R 实现

夏昌发 (北京协和医学院)

时间: 11:00-11:30

简介: 2015 年毕业于河北医科大学, 获医学学士学位; 2018 年毕业于北京协和医学院, 获流行病与卫生统计学硕士学位; 现于北京协和医学院攻读博士学位。先后获得北京市优秀毕业生、北京协和医学院优秀研究生等称号。发表学术论文 20 余篇, 其中作为第一作者发表论文 6 篇, 包括 Lancet Global Health、Cancer Letters、Tobacco Control 等国际知名杂志。曾为 Cancer、Cancer Medicine 等国际杂志的担任审稿专家。先后参与《中国癌症地图集》编制项目、肿瘤登记随访项目、上消化道癌筛查项目、宫颈癌卫生经济学项目等, 在项目中熟练使用 R 语言分析数据。

摘要: 通过数学模型模拟传染病的动态传播过程能有效地预测疾病的流行状态、并评价医疗卫生干预的效果。传染病动态传播模型是一种基于疾病传播动力学的模型体系, 常用于模拟传染病动态流行趋势, 并可扩展用于社会人口学特征的动态模拟等。该模型体系主要包括确定性房室模型、随机个体化模型和随机网络模型。确定性房室模型将人群划分为离散的疾病状态, 并进一步细分影响疾病传播的人口统计学、生物学和行为等特征, 通过在连续时间内求解传染病微分方程获得疾病在各个时点的流行状况。确定性模型中疾病和人口的转移参数在模拟过程中没有随机变异, 而随机模型则通过个体的离散化来估计模拟传播过程的潜在变异。随机个体化模型在个体测量水平上模拟人群中疾病传播的过程, 个体转移参数则通过在参数分布中随机抽样来获得。随机网络模型动态细化了传染源与易感个体的接触时间和次数, 因此该模型具有了网络模型的特征。借助 R 语言, 我们可以方便地刻画固定队列或动态人口各种疾病的传播类型, 如 SI、SIR 和 SIS 等, 并可根据需要自主扩展模型结构, 更为贴切地模拟任意复杂传染病的流行过程。

基于混合效应模型的人群代谢组学差异性标志物筛选研究

李昂 (北京协和医学院)

时间: 11:30-12:00

简介: 中国医学科学院, 北京协和医学院基础学院流行病与卫生统计学系硕博连读研究生, 目前的研究方向主要包括: 空气污染人群流行病学、代谢组学数据挖掘、证据权重分析等。参与编写《空气污染人群健康风险评估方法及应用》。曾获 2017 年“挑战杯”大学生课外学术科技作品竞赛省级特等奖、国家级三等奖。

摘要: 大量人群流行病学研究证实, 暴露于高水平的空气污染物可导致一系列的心血管和代谢结局, 但其潜在的生物学机制尚不明确。代谢组学技术可有效地从生物样品中定性和定量大量化合物, 从而鉴定受环境或疾病影响的代谢物和代谢通路。我们在传统代谢组学 PCA 分析、OPLS-DA 分析的基础上, 结合线性混合效应模型, 筛选出与细颗粒物暴露有关的差异性代谢产物。该方法能够考虑个体的多种混杂因素及其变化, 并能同时在模型中控制组间变异和组内变异对结局变量的影响。分析主要应用 R 软件 lme4、splines 包等。

XBART: 更快更准的提升树 (Boosting Tree) 算法

何靖宇 (芝加哥大学)

时间: 14:00-14:30

简介: 芝加哥大学布斯商学院三年级博士生, 毕业中科大。专注于研究机器学习算法、贝叶斯统计及在资产定价领域应用、量化投资。目前发表文章于 Journal of Business and Economic Statistics, Journal of Computational and Graphical Statistics, Bayesian Analysis, AISTATS 等杂志及会议。

摘要: XBART 是一个新型的提升树 (Boosting tree) 算法, 受到 BART (Bayesian additive regression trees) 模型启发, 在目前基于树的算法中有着极高的准确度和计算速度。模拟显示 XBART 取得比梯度提升树 (Gradient Boosting tree) 更高的准确性和更好的速度。

训练拳皇 97 AI

陈昱 (北京大学光华管理学院)

时间: 14:30-15:00

简介: 北京大学光华管理学院统计学博士生在读。研究方向为深度学习及强化学习, 关注强化学习在游戏领域的应用, 高性能强化学习工程实现等领域。

摘要: 我们介绍如何使用强化学习训练拳皇 97。包含 (1) 分布式 RL 训练系统及数据交互方案 (2) RL 环境设计 (3) 强化学习算法实现 (4) Multi-agent 对抗性以及 (5) 强化学习中的迁移问题。我们将展示训练结果, 并介绍我们开源的训练工具与教程。

面向移动端的计算机视觉技术简介

吴兴龙 (字节跳动)

时间: 15:00-15:30

简介: 2013 年硕士毕业于吉林大学数学学院; 毕业后加入三星中国研究院, 从事图像算法研究工作。2015 年加入字节跳动, 现为字节跳动 AILab 资深算法工程师, 主要研究方向为计算机视觉, 包括人脸, 手势, 分类, 分割等问题。

摘要: 分为字节跳动 & AILab 简介、抖音中典型的应用案例、移动端 CV 基础知识简介三个部分。

工业推荐系统简介

骆颇 (趣头条)

时间: 16:00-16:30

简介: 2017 年硕士毕业于复旦大学计算机科学与技术学院; 毕业后加入今日头条, 从事个性化推荐工作。2018 年加入趣头条, 现为趣头条算法 leader, 主要关注内容推荐。

摘要: 介绍个性化推荐系统的整体流程。主要内容为数据流, 用户画像, 协同召回, 向量化召回, 特征工程与排序模型在内容推荐中的实践。

深度召回在京东搜索中的应用

熊熹 (京东)

时间: 16:30-17:00

简介: 2009 年加入统计之都, 从此被改变人生轨迹 (某种程度上)。2015 年加入京东, 一直致力于机器学习算法在京东推荐与搜索业务中的应用, 目前主要负责京东搜索算法质量评估与用户体验优化。

摘要: 深度学习用于排序的应用比较多, 但是深度学习直接作用于一个几十亿索引级别的工业界搜索系统的召回阶段, 并且带来非常巨大的用户体验改善和直接商业价值的案例并不多。本报告主要介绍了京东最新的深度召回系统实践, 并系统介绍了基于 IVFPQ 的工业级 knn 的系统简洁与有效之处。

洞察数据 商业价值——北大光华商业分析硕士项目

王汉生 (北京大学光华管理学院)

时间: 14:00-14:45

简介: 王汉生教授现任北京大学光华管理学院商务统计与经济计量系系主任。1998 年北京大学数学科学学院, 概率统计系, 统计学本科, 2001 年美国威斯康星大学麦迪逊分校, 统计学博士。现为国际统计协会会员, 美国统计学会, 美国数理统计研究员, 英国皇家统计协会 (Royal Statistical Society), 以及泛华统计学会会员。

他发表英文学术论文五十余篇, 中文论文近二十篇。合著英文专著 1 本, 独立完成中文教材 2 本。先后担任多个学术刊物副主编 (Associate Editor)。现主要理论研究兴趣为: 高维数据分析、变量选择、数据降维、极值理论、以及半参数模型。主要应用研究兴趣为: 搜索引擎营销、社会关系网络。

摘要: 大数据、互联网、物联网等中国新型产业带来发展机遇和商业奇迹, 新时代下我国蓬勃壮大的信息产业、互联网产业、数据产业迅速崛起, 占据世界信息行业的一席之地。新的行业伴生出对新型人才需求的强烈呼声。为此, 自 2017 年起, 北京大学光华管理学院为适应和推动中国数据产业发展而重点打造了北大光华商业分析硕士项目, 成为我国第一个完全自主培养并密切联系中国国情的商业分析项目。培养精通数据商业价值的高级人才, 为中国数据产业培养技术与管理兼备的优秀人才, 推动中国数据产业的繁荣进步。王汉生教授讲述北大, 讲述光华, 讲述北大光华商业分析项目。

社交大数据

王翀 (北京大学光华管理学院)

时间: 14:45-15:30

简介: 王翀, 北京大学光华管理学院管理科学与信息系统学系副教授。2004 年毕业于北京大学数学科学学院应用数学专业, 2006 年获清华大学金融学硕士学位, 2012 年于香港科技大学商学院获得博士学位。2012 年至 2017 年任教于香港城市大学商学院。王翀教授关注现代信息技术, 如互联网、区块链, 人工智能等, 对社会、经济系统产生的颠覆性冲击。他的研究涉及社交媒体, 平台化商业模式, 群体智能与众包, 金融信息技术应用与监管等多个前沿领域。他的论文发表于 Information Systems Research, Journal of Management Information Systems, Decision Support Systems 等重要国际学术期刊, 并在 Information Systems Journal 担任 Associate Editor。

摘要: 线上社交已经成为了人们生活的重要组成部分, 每天有超过 10 亿用户在使用微信。随着互联网应用向移动端转移, 社交应用进一步成为整合各种服务的平台。与此同时, 各种新媒体社交平台 (如抖音、快手) 不断涌现。在给人们带来便利的同时, 基于互联网的社交服务实现了社会交往过程的数字化。互联网社交应用每天产生大量的用户行为和交互数据, 这些数据蕴含着广阔的研究机会和巨大的商业价值。那么, 复杂的网络社交后面的数据到底是什么形态? 如何实现对数据的分析? 数据分析能产生怎样的价值呢?

Leaderboard Effect: Who, When, and How?

厉行 (北京大学光华管理学院)

时间: 16:00-16:45

简介: 厉行, 北京大学光华管理学院市场营销系助理教授, 北京大学管理学学士、经济学硕士, 斯坦福大学经济学博士; 研究兴趣主要是数量营销、广告、实证产业组织、知识产权与创新。

摘要: We study the design of leaderboard for an online learning platform in order to help users finish their learning object, increase their activation and retention, and enlarge the user base of the platform. The online learning platform we study is an English-word-memorizing platform for Chinese learners, and it displays the learning hours of top ten users one day before. We show the existence of leaderboard effects during the learning process of users, i.e., users learn more when exposed to a longer leaderboard hour. Such effect exhibits a hump-shape as the user is moving towards the end of the book, with maximum value happened when users finish 60% of the process. The above reduced-form finding is consistent with the predictions in the goal literature studying the interplay of internal motivation and external motivation (e.g., Huang, 2018). Motivated by our reduced-form findings, we build a structural model which allows for rich heterogeneity in terms of base-line learning intensity, the position on the leaderboard referred, the leaderboard effects, and its variation along the learning process. Based on our model estimates and counterfactual analysis, we have the following advices: (1) identify users who are not positively responsive to the leaderboard, and shut down the leaderboard on their screen (who); (2) for each users, detect their specific period of learning process with positive leaderboard effect, and only show leaderboard during that period (when); (3) show the optimal length of the leaderboard based on their heterogeneous referral position (how). This paper shed light on the interface design of online learning websites, and it may also provide guidance for other websites with leaderboard.

The Unintended Consequences of Tariff Retaliation: Evidence from the Chinese Automobile Market

楚燕来 (中国人民大学)

时间: 14:00-14:30

简介: 楚燕来, 中国人民大学营销系教授

摘要: This paper adopts a novel perspective to assess the impact of tariff wars in the context of globalized production. Using automobile sales data in China, we find that China's retaliatory tariff against the U.S. increased sales of U.S. imports, increased sales of U.S. brands in the import segment, and decreased sales of China-made U.S. brands in the domestic segment. The increased sales of imported U.S. brands was driven by the advertising effect of the tariff war, while the decreased sales of China-made U.S. brands was caused by consumer boycotts. The net profit impact on U.S. brands was positive due to the much higher margins for imported cars. These findings suggest that a tariff war under globalization can be counterproductive and have unintended consequences.

Highest Contributions from Others: A Dampening Effect on PWYW Payment

马雪静 (北京大学光华管理学院)

时间: 14:30-15:00

简介: 马雪静, 北京大学营销系博士研究生

摘要: As an innovative way to monetize online content, Pay-What-You-Want (PWYW) pricing has emerged on many online platforms by enabling users to pay to content contributors voluntarily. Unlike traditional PWYW pricing which is usually implemented within private context, individuals' payment behavior can always be observed publicly in social context on these online platforms. In this paper, we analyze how others' payment behavior influence one's own payment behavior under such type of PWYW pricing. Specifically, we focus on the effect of the highest contribution from others and examine three potential explanations: anchoring, signaling, and crowd-out. We develop a set of theoretical predictions on individual payment incidence and payment amount based on each explanation. We then conduct an empirical analysis using a large-scale, individual-level dataset with 2.7 million observations from a live streaming platform with PWYW pricing. We uncover a robust, yet striking finding: namely, the highest payment from others has a negative effect on both individual payment incidence and payment amount, suggesting that the crowd-out effect is the primary driving force. Furthermore, we find that the crowd-out effect is attenuated by individuals' experience with live streaming viewing and gift payments on the platform. Evidence from a cohort analysis suggests that there exists individual heterogeneity. Our findings have important managerial implications for firms that adopt PWYW pricing in a social context. While highlighting large payments from others is commonly used by many online platforms to induce more and higher payments, it could in fact do the opposite by hurting the revenue of the platform, as the information of the highest payment from others reduces individuals' payment incentives among the majority of viewers.

How Is Mobile User Behavior Different?—A Hidden Markov Model of Mobile Application Usage Dynamics

吴少辉 (清华大学)

时间: 15:30-16:00

简介: 吴少辉, 清华大学营销系博士研究生

摘要: Mobile application usage is becoming an essential activity in many people's daily lives. Compared with PC internet, mobile internet usage is ubiquitous, temporally fragmented, and more context-dependent. Thus far research is limited on the mechanism of mobile app usage and the effects of context, even as usage continues to grow. In this paper, we aim to develop a framework to capture the underlying mechanism of mobile app usage, taking into account its unique time-fragmented feature and the possible impacts of contextual factors. To this end, we propose a hidden Markov model (HMM) and calibrate it using a consumer panel that contains real-time app usage information. We find three hidden states driving mobile app usage: utilitarian, social, and hedonic. Consumers have multiple intentions in the utilitarian state but are relatively single-minded in either the social or hedonic state. The state dynamic is volatile: Chains of continuous utilitarian states and hedonic states intercommunicate frequently with densely intermittent social states. Such a volatile state dynamic provides an antecedent for the fragmented-time mobile usage phenomena. Contextual factors—in particular, location and time of day—fluence the state dynamic, e.g., its volatility varies across different locations and times of day. In sum, our analysis depicts the following picture: In the mobile internet era, consumers take advantage of relatively long free-time windows for functional and entertainment activities, while exploiting short micro-moments for social activity. Furthermore, with the help of mobile internet technology, consumers seem to (1) utilize more micro-moments outside the office or in the morning as a complement to work, and (2) reserve more micro-moments in the evening for relaxation and entertainment, all at the expense of social activities.

What's Your Risk Attitude On A Grey Day: The Case of Air Quality and Financial Product Choice

张晗 (北京大学光华管理学院)

时间: 16:00-16:30

简介: 张晗, 北京大学营销系博士研究生

摘要: Using investment transaction data from an online financial product platform in China, we find that air pollution does not affect total investment amount, but it changes the distribution of investment on financial products with different risk levels. In particular, higher level of air pollution shifts consumers' financial investments towards products with higher risk. In addition to the distribution-shift effect, there also exists an adaption effect: If the severe air pollution has lasted for multiple days, the distribution-shift effect is weakened. Our results indicate that air pollution may change people's risk attitude in financial product consumption, and this effect becomes smaller if people get used to bad air quality.

数据驱动的分子会诊与精准诊疗

王朝东 (首都医科大学)

时间: 14:30-14:55

简介: 教授、主任医师、博士研究生导师。现任首都医科大学宣武医院神经内科遗传代谢专业主任、国家老年疾病临床医学研究中心办公室主任, 兼任国家重点研发计划项目“主动健康与老龄化科技应对”重点项目专家组成员、中国医师协会老年病学分会委员、中国优生科学协会理事、中国老年保健协会脑保健专业委员会副主任委员、北京市医学会帕金森病与运动障碍分会委员、北京市医学会遗传学分会委员。长期从事神经系统疾病的临床诊疗及分子机制、精准诊治和大数据应用等方面研究。

摘要: 中国各级医院、医师诊疗水平参差不齐, 优质资源集中在大医院和少数专家手里, 基层医院和大夫诊疗水平提高的渠道少。临床大量的误诊误治病例, 均由于缺乏精准的专业知识和标准化的流程。日益复杂和细化的临床诊疗要求, 给传统的经验医学提出了前所未有的挑战。由于中国医生少、病人多, 医生没有时间和精力掌握完整的系统知识, 也不可能规范地按照标准流程开展诊治。因此, 迫切需要一套简单、实用的工具系统, 辅助神经专科和非专科大夫在短期内处理大量的医学信息和快速做出正确的决策。本报告将围绕神经系统疾病的精准诊疗辅助决策与多层次会诊系统的建设方案和临床应用场景进行介绍。

真实世界研究中的临床数据处理问题

马超超 (北京协和医院)

时间: 14:55-15:20

简介: 北京协和医院检验科研究生, 师从北京协和医院检验科邱玲教授。从事临床实验室数据挖掘规范研究, 在协和实习及研究生期间一直致力于临床实验室数据从数据挖掘到人工智能转换的研究。目前以第一作者、共同第一作者及参与作者发表 SCI 5 篇, 中文核心 4 篇。曾执笔的临床实验室真实世界研究规范获得中国老年医学会团体标准立项。

摘要: 围绕临床实验室真实世界数据挖掘方法和人工智能初探这一主题展开, 进行了以下两方面的论述: 结合自身研究实例, 谈论临床实验室数据挖掘的流程规范; 论述临床实验室真实世界数据挖掘到人工智能的初探, 助力实验室管理。基于各指标大数据的分布情况, 借鉴工业质量控制方法, 建立可实时监测异常情况的个性化的算法和模型, 助力实验室质量管理, 为构建统一化、智能化实验室奠基。

大数据技术在检验医学中的应用

段欣岑 (复旦大学附属中山医院)

时间: 15:20-15:45

简介: 复旦大学附属中山医院检验科、生物统计室数据分析师。2018 年毕业于美国德保罗大学数据科学专业, 随即加入中山医院。从研究生阶段一直致力于卫生健康数据, 尤其是检验科数据分析方法和架构的探索。

摘要: 大数据与人工智能作为近几年的热词, 已经逐渐深入日常生活。随着检验技术的不断发展, 将大数据与人工智能技术应用于检验医学也变得极为迫切。报告将联系大数据的基本概念与目前检验科在实际工作中面临的困难, 参考大数据与人工智能在其它领域中日趋成熟的应用, 提出一系列如何将这些技术运用到检验医学与日常工作中的方法, 并分享部分中山医院检验科在这方面的初探成果。最后对今后大数据在检验医学方面的发展提出展望。

图像识别自闭症的算法基础与应用

章林 (杭州骐云软件有限公司) & 陈显扬

时间: 16:10-16:35

简介: 章林, 原微软企业咨询部高级咨询顾问, 曾担任项目经理及架构师, 善于跨学科技术应用及构建异种平台。目前主要研究方向是计算机前沿技术在医疗领域中的应用。

摘要: 模拟自闭症诊断流程, 建立我国人群的刺激 - 反应训练集。利用图像识别技术, 来分析情绪和行为的变化, 建立情绪和行为变化与自闭症诊断之间的模型联系。未来可以通过该系统, 早筛查幼儿是否患有自闭症和心理健康方面的问题。实现在家也能对儿童进行筛查, 从而提前诊断和治疗。

不同标志物对乳腺癌的早筛效率分析

姜楠 (清华大学第一附属医院) & 陈显扬

时间: 16:35-17:00

简介: 姜楠, 现为北京华信医院 (清华大学第一附属医院) 普外科副主任医师。任“中国抗癌协会康复会乳腺、甲状腺肿瘤分会”青年委员, “中国乳腺微创与腔镜手术联盟”成员, “乳腺癌人工智能与转化联盟”成员。

摘要: 甲状腺癌是头颈部恶性肿瘤中最常见的内分泌恶性肿瘤, 并且呈逐年上升趋势, 甲状腺乳头状癌 (PTC) 是最常见的类型。目前, 虽然超声引导细针穿刺活检被认为是鉴别甲状腺良恶性结节最有效的检查方法, 但这是一种侵入性手术。因此, 寻找无创、有效、可靠的血清标志物对于甲状腺癌的早期诊断是必要的。目前针对恶性肿瘤的分子标志物的研究主要集中在基因和蛋白领域, 例如, 纤维连接蛋白 -1、细胞角蛋白 -19 和 TPO 等 [1-3]。但是这些标志物的特异性和准确率都不高。我们应用非靶向代谢组学方案, 用前瞻性研究策略, 利用高分辨液相质谱技术对甲状腺乳头状癌的患者和正常人的血清进行检测, 进行非靶向代谢物的筛选, 通过多元统计分析以及机器学习方案, 确定能准确诊断和预测甲状腺癌的潜在肿瘤标志物以及预测模型。

基因和代谢组的多层次组学联合分析

杜智勇 (北京大学)

时间: 17:00-17:25

简介: 北京大学药学院生药学专业博士, 博士后, 有十年以上的药物研发经验, 并在多组学整合的技术研究上, 具有独特的创新性。

摘要: 通过代谢组学、脂质代谢组学、肠道微生物组、蛋白质组学、转录组学等多组学联用技术, 结合经典统计学以及多元统计学的分析手段, 建立基因 - 蛋白 - 代谢的分子联系, 从而揭示中药复杂体系对心肌肥厚、心肌缺血、心力衰竭的影响及潜在作用机制。

基因大数据分析平台的开发和临床应用

李奇斌 (云峰生物)

时间: 17:25-17:50

简介: 云峰生物创始人和 CEO, 2010 年毕业于中国科学院基因组研究所, 获生物信息学博士学位, 曾担任华大基因高级科学家和产品研发总监等职位。长期致力于与人类疾病的基因组学和数据分析算法开发, 在 Nature 和 Nature Genetics 等顶级学术期刊发表论文二十余篇。

摘要: 随着高通量测序技术的快速发展, 基因测序已经逐渐从实验室研究走向临床。目前基因数据分析和临床解读需要很多手工操作, 效率低下且容易犯错。我们采用了文献挖掘和人工校对等方法, 构建了临床级的基因解读知识库, 同时采用统计学习等算法, 构建了致病变异筛选算法, 可以快速从数万个基因变异中确定真实致病变异, 实现了遗传病的精准快速诊断。

财经数据分析之 pedquant 包

谢士晨 (中银富登)

时间: 14:00-14:30

简介: 名古屋大学环境经济学方向博士。目前就职于中银富登, 从事信用风险量化分析管理相关工作。曾经开发过 scorecard 包 (python 版本 scorecardpy), 广泛应用于信用风险量化模型评分卡的开发。

摘要: pedquant (Public Economic Data and Quantitative Analysis) 提供了接口获取 NBS, FRED, Yahoo, 网易财经, 新浪财经等财经数据, 以及可视化分析与策略开发等基本功能。该包可用于替换经典的 quantmod 包, 且能够与 tidyquant 结合使用。

R 语言空间数据处理与分析

卢宾宾 (武汉大学)

时间: 14:30-15:00

简介: 任职于武汉大学遥感工程学院, 博士毕业于爱尔兰国家地理计算中心, 硕士就读于北京大学遥感与地理信息系统研究所, 个人研究领域包括空间统计、地理加权建模技术、空间数据分析和开源 GIS 开发等, 在 IJGIS、CEUS、Journal of Statistical Software、Spatial Statistics 等期刊发表论文 12 篇, 负责开发并维护了地理加权建模技术 R 函数包 GWmodel, 囊括了地理加权回归分析、地理加权汇总统计量、地理加权主成分分析和地理加权判别分析等地理加权建模技术。

摘要: R 是当前最流行的统计计算、数据分析和图形可视化的开源平台软件之一, 尤其在空间数据相关的统计与分析领域所发挥作用越来越大。本报告将介绍如何在 R 中实现空间数据处理和基础分析操作, 同时围绕空间数据统计分析和可视化技术进行相关技巧介绍与技术展示, 不仅面向地理信息科学、遥感科学与技术等直接相关领域的听众, 同时也为社会经济、生态研究等领域内的科研人员从业者系统介绍使用 R 语言进行空间数据处理与分析的基础。

使用 Shiny 开发中文自然语言处理 Web 应用

杨健 (安利 (中国) 研发中心有限公司)

时间: 15:00-15:30

简介: 杨健, 硕士研究生毕业于美国康奈尔大学统计科学系, 先后在诺华、礼来等世界 500 强药企做临床试验统计分析, 编程等方面工作, 目前就职于安利 (中国) 研发中心有限公司任职 Senior Data Scientist, 主要负责安利中国研发的数据处理分析及建模。

摘要: 介绍两个利用 R Shiny 开发的中文自然语言处理 Web 应用。这两个应用可以帮助无编程经验的使用者快速绘制中文词云, 以及实现中文文本分类等任务。

基于 knitr、rmarkdown 的 R+HTML 生态应用

李宇轩 (中国人民大学)

时间: 16:00-16:30

简介: 李宇轩, 中国人民大学统计学院大四在读学生, 热爱 R 语言, 对 R+HTML 应用及文本分析有很浓厚的兴趣。

摘要: R 主要用于数据分析的过程, 但其用于 R+HTML 的生态应用则提之较少, 本文则是基于 knitr, rmarkdown 等 R 包工具, 介绍其设计思想、发展历史及基于分析报告、幻灯片及网站等的具体开发和应用。

R 社群的组织与参与

覃文锋 (暖房直租)

时间: 16:30-17:00

简介: 毕业于厦门大学公共卫生学院, 王亚南经济研究院。曾开发 jiebaR, awesome-R 等开源项目, R Weekly 编辑部成员。任职暖房直租开发工程师, 暖房直租是一个个人房东直租平台。

摘要: R 社区中有各种形式的组织和活动, 例如 R Forward 旨在促进女性和少数族裔群体在 R 社区中的参与和发展; 分散在世界各地的 R User Groups 定期组织 R 用户本地聚会, 分享社区动态; R Consortium 通过资助 R User Groups, 资助 R 开源项目的方式, 来促进 R 语言的发展。

本演讲将介绍 R Consortium, rOpenSci, R Forward, R-Ladies, R Weekly, R User Groups, R-GSoC 等的几个热门社群以及有关公益活动。分享参与和组织 R 社群的经验, 以及申请 R Consortium 等项目资金赞助的方法。

社交网络上议题社群的公共焦虑研究

塔娜 (中国人民大学)

时间: 14:00-14:30

简介: 塔娜, 中国人民大学新闻学院讲师, 中国人民大学新闻与社会发展研究中心研究员。2017 年毕业于清华大学计算机系, 获计算机科学与技术专业博士学位。研究方向为计算传播学, 新闻大数据。任教课程:《计算传播学》、《大数据与新闻传播实务》、《数字传播技术应用》、《新媒体运营实务》、“新闻与传播学科核心与特色课程创新计划”课程《新闻传播程序设计基础》等。近年来以第一作者或通讯作者身份发表多篇 CCF(中国计算机学会)A 类及 SCI 索引论文。目前主持国家自然科学基金项目青年项目“时空感知的异构社交网络传播模型研究”、北京市社会科学基金项目青年项目“异构社交网络传播模型及其对北京影响力最大化问题研究”等科研课题, 曾参与 973 项目“大数据群体计算的基础理论与关键技术”、国家自然科学基金项目“基于位置的社交网络关键技术研究”等科研项目。

摘要: Although a number of researches on individual level anxiety evaluation have been proposed, there are few researches on evaluating the public anxiety of a social network community, which can benefit various social network analysis tasks. However, we can not simply average anxiety scales of all individuals to calculate the public anxiety score of a community, because: (1) individuals are influenced by their connections in a community, so impacts from interpersonal relations on individuals' anxiety scales should be considered, i.e., the Structural factor; (2) public anxiety always relates to certain topics, topical discussions also reflect a community's anxiety level, which should also be considered, i.e., the Topical factor. In this paper we initiate the study of evaluating the public anxiety of topic-based social network communities (TSNC). We propose an evaluation framework to project a TSNC's anxiety level into a score in the [0, 1] range, using both Structural and Topical factors. We devise a cascading model to dynamically compute the anxiety score using the Structural influence. We propose a stochastic model to measure anxiety score of social network messages using a generalized user, and design a tree structure (MC-Tree) to organize messages of a TSNC to effectively compute anxiety score from the Topical factor. For large communities, computing public anxiety in real-time can be expensive, we show how to use a small sample of the community to compute the public anxiety within given confidence interval. Our model exhibits more than 80% precision and 90% recall in an empirical study on real-world data sets from Weibo.

媒介数据挖掘与指数构建

向安玲 (清华大学)

时间: 14:30-15:00

简介: 向安玲, 清华大学新闻与传播学院博士研究生, 武汉大学信息管理学院管理学硕士(硕士研究方向: 数字出版与新媒体; 武汉大学信息管理学院编辑出版专业学士、经济管理学院经济学双学士), 清博研究院副院长, 清博指数体系创始研发者(含 WCI、BCI、TGI 等数十个指标体系, 十余部委内部新媒体官方评价标准)。研究方向: 新媒体、大数据、舆情、数字出版、互联网 + 文化产业。已发表相关论文十余篇, 主持参与近十个相关项目, 有较丰富的科研研究经历。

摘要: 媒介数据挖掘与指数构建; 摘要: 从大数据到小数据: 快速筛选高价值内容; 从历史到未来: 洞察规律、预判趋势; 从真实到虚拟: 断物识人、拟态仿真。基于百亿级网络公开数据、百万级新媒体账号标签、数百个指数模型, 实现对媒介数据的深度挖掘与分析, 通过泛指数体系的构建, 打造人文传播研究的大数据基础设施。

不是科学家，媒体怎么做“数据可视化”？

姜柳（北京数可视科技有限公司）

时间：15:00-15:30

简介：姜柳，北京数可视科技有限公司数据编辑，毕业于美国雪城大学计算机辅助新闻专业。和腾讯科技以及谷雨频道多次合作数据新闻制作，创作有《宝贝，我们等你回家过年——图解十年千位儿童的被拐路线》《扒完北京地铁 5 年的数据，我们发现了最“悲剧”的通勤线路》等作品，和腾讯科技合作的数据快讯的累计阅读量超过 800 万。乐于从数据中去发掘新闻故事，希望通过交互设计、信息图、漫画等设计去展现信息之美。

摘要：“数据可视化”听起来是一个非常宽泛、什么都可以往里面装的表现形式，越来越多的交互作品和平面信息图都开始用以“数据”为噱头进行包装。和专业的数据团队相比，媒体行业的“数据可视化”业务优势在哪里，什么才算得上是一个“有意义”的数据新闻作品，商业公司怎样才能把“数据可视化”的业务特色贯穿到所有客户订制作品里面？分享环节将和大家一起讲述商业公司制作可视化作品的特色和流程，以及如何在甲方需求和自身专业特色之间寻求平衡。

影视大数据用户行为分析

王妍（中国传媒大学）

时间：16:00-16:30

简介：王妍，中国传媒大学数据科学与智能媒体学院副教授，博士后。2010 年中国科学院数据科学与虚拟经济研究中心获得理学博士学位。研究领域：统计建模与传媒大数据分析、城乡投入产出分析与低碳经济。长期从事广电和影视大数据建模和分析工作。主要承担本科生和研究生概率论、数理统计、多元统计和数据分析等课程。

摘要：利用有线电视双向互动用户采集回传动态数据，建立影视节目的用户偏好和收视习惯画像模型，基于模型结果对每个用户进行偏好和习惯标签设置；在用户画像基础上，建立用户收视偏好和习惯的分群模型，通过最优分群结果进行群集特征描述；最后对用户个体画像和分群结果进行可视化展示研究。

电影《摇滚藏獒》营销案例分析

李波（中国传媒大学）

时间：16:30-17:00

简介：李波，女，中国传媒大学理学院统计系副教授。北京航空航天大学系统工程专业博士。主要承担研究生和本科生随机过程、数学建模与数学实验、时间序列分析、受众数据分析、数据挖掘等课程。主要从事影视文化产业领域优化、建模、大数据处理分析等研究工作。近几年来主持和参与国家自然科学基金项目、广电总局项目，北京市科委项目等各类科研项目 10 余项。在 SCI、CSSCI、EI 等各类学术期刊发表论文 40 余篇，出版学术专著 2 部。

摘要：通过从各大门户网站、搜索引擎、社交平台、视频网站、微博、微信、app 等主流媒体采集相关数据，从主创团队、制/发公司、前期作品、相关作品、类型/标签、物料元素、宣发营销效果、档期、排片、票房评估等方面对电影《摇滚藏獒》做全方位的分析，为电影的制作、宣传、发行、放映等相关从业者制定营销方案、监测影片市场动态、了解影片宣发物料投放效果、调整排片场次，掌握观影人群构成及其观影反馈、发掘物料投放目标，以及为广告商制定广告推广方案、挖掘潜在广告营销客户提供数量化的决策支持参考。

Forecasting with time series imaging

Li Xixi (北京航空航天大学)

时间: 14:00-14:30

简介: A master student from Beihang University. My research focuses on time series forecasting, machine learning and business intelligence.

摘要: Feature-based time series representation has attracted substantial attention in a wide range of time series analysis methods. Recently, the use of time series features for forecast model selection and model averaging has been an emerging research focus in the forecasting community. Nonetheless, most of the existing approaches depend on the manual choice of an appropriate set of features. Exploiting machine learning methods to automatically extract features from time series becomes crucially important in the state-of-the-art time series analysis. In this paper, we introduce an automated approach to extract time series features based on images. Time series are first transformed into recurrence images, from which local features can be extracted using computer vision algorithms. The extracted features are used for forecast model selection and model averaging. Our experiments show that forecasting based on automatically extracted features, with less human intervention and a more comprehensive view of the raw time series data, yields comparable performances with the top best methods proposed in the largest forecasting competition M4.

Kolmogorov-Smirnov simultaneous confidence bands for time series distribution function

李杰 (清华大学)

时间: 14:30-15:00

简介: 李杰, 清华大学统计学研究中心直博二年级学生, 师从杨立坚教授。主要研究方向为非参数统计、时间序列和函数型数据分析。

摘要: Claims about distribution functions of time series are more often folklores than substantiated conclusions, due to lack of hypotheses testing tools. In this work, Kolmogorov-Smirnov type simultaneous confidence bands (SCBs) are constructed based on a simple random sample (SRS) drawn from a realization of time series, together with smooth SCBs using kernel distribution estimator (KDE). All SCBs are shown to enjoy the same limiting distribution as the standard Kolmogorov-Smirnov SCB for i.i.d. sample. This theoretical fact has been validated in simulation experiments performed on various time series. Hypotheses testing based on these SCBs has led to the unexpected finding that with proper rescaling, Gaussian distribution and most student's t-distributions are all acceptable alternatives of the S&P 500 daily returns' stationary distribution. This discovery challenges the long held belief that daily financial returns' distribution is fat-tailed and leptokurtic.

统计学学生在金融行业中求职的方向

赵一懋 (北京大学)

时间: 15:00-15:30

简介: 赵一懋, 本科毕业于中国科学技术大学统计学, 研究生就读于北京大学数学科学学院应用统计金融统计方向。

摘要: 随着金融行业的飞速发展, 各大基金、券商等金融机构对统计学的人才需求日益增加。那么, 金融机构的哪些部门对统计学的人才更加偏好呢? 统计学的毕业生在金融行业的求职上有哪些方向呢? 作为一个刚经历秋招季的统计学毕业生, 本演讲准备讲述一名普通的统计学毕业生在秋招季面临的挣扎与选择, 详细介绍一些需要统计学毕业生的金融机构或者部门, 以及该部门的具体工作。

Large-scale Regression with Two-stage Best-score Random Forest

黄涛 (中国人民大学)

时间: 16:00-16:30

简介: 黄涛, 男, 本科毕业于中国人民大学统计学院, 现为中国人民大学统计与大数据研究院博士一年级研究生。

摘要: 我们提出了一种针对大规模回归问题的新算法, 即两阶段最佳得分随机森林 (TBRF)。“最佳得分”意味着从一定数量的纯随机回归树候选中选择一个具有最佳经验性能的回归树, 而“两阶段”意味着将原始随机树的切割过程分成两阶段: 第一阶段, 特征空间基于树结构被划分为非重叠的单元 (即叶节点); 第二阶段, 在上阶段得到的非重叠的每一个单元, 我们继续基于树结构对特征空间进行划分。该算法的优点可归纳如下: 首先, TBRF 中的纯随机性使得几乎最优的收敛速率能够达到, 并且这解决了长期困扰现有大规模回归算法的边界不连续性问题。其次, 两阶段过程为并行计算铺平了道路, 从而提高了计算效率。最后, 作为一个一般性的框架, TBRF 可以令各种回归器, 例如线性预测器和最小二乘支持向量机 (LS-SVM) 也嵌入到第一阶段的树的叶节点里, 具体取决于基础数据集的特征。在大规模回归数据集上的实验, TBRF 良好的预测精度和高计算效率得到验证。

Long Distance Relation Extraction with Article Structure Embedding and Applied to Mining Medical Knowledge

林毓聪 (清华大学)

时间: 16:30-17:00

简介: 林毓聪, 清华大学统计学博士研究生, 主要研究方向为医学大数据分析、医学知识图谱构建。

摘要: As a central work in medical knowledge graph construction, relation extraction has gained extensive attention in the fields of natural language processing and artificial intelligence. Conventional works on relation extraction share a common assumption: a sentence can express a relation of an entity pair only if both entities appear in this sentence. Under this assumption, plenty of informative sentences are precluded. In this paper, we break the assumption and propose a new relation extraction model that incorporates article structure information, which not only provide additional information, but also allows extracting long distance relations. We apply the model to online medical relation extraction and demonstrate its advantage over conventional models.

统计之都在线投稿系统

黄湘云 (中国矿业大学(北京))

时间: 17:00-17:30

简介: 统计之都副主编, 中国矿业大学(北京)统计学硕士, 感兴趣的领域有可重复数据分析, 数据可视化和混合效应模型。R 语言信仰粉, 活跃于 GitHub, 统计之都论坛等社区。

摘要: 统计之都副主编带你探秘最具透明性、时尚性和专业性的 blogdown 在线投稿系统, 以真实投稿案例分享全透明的专业严谨的审稿流程, 我们来稿必回! 欢迎大家踊跃给统计之都编辑部投稿, 只要文章符合专业、人本、正直的气质, 我们每一篇都会安排相关专业领域的大佬在线审稿, 快来和大佬在线互动吧! 希望大家在交流的过程中, 不断碰撞出知识的火花, 为普及统计学知识贡献自己的力量!

大气污染与人群健康的关系

李国星 (北京大学)

时间: 14:00-14:30

简介: 李国星, 博士, 北京大学公共卫生学院, 副教授。主要研究领域: 环境流行病学, 特别是大气污染和气候变化的健康影响和疾病负担。作为项目负责人, 先后主持国家自然科学基金面上项目 1 项、省部级课题子课题 1 项和中华医学会课题 1 项, 并参与多项国际自然科学基金和环保部公益项目课题。作为第一或通讯作者累计发表英文论文 24 篇, 包括在 *lancet planetary health, environmental international, stroke, environmental pollution, environmental research, science of the total environment* 等期刊的发表; 参编专著教材 4 部, 包括《现代环境卫生学》, 《空气颗粒物与健康》等。目前担任中华预防医学会卫生工程分会青年委员; 北京环境诱变剂学会青年专业委员会常务委员。

摘要: 近年来, 大气污染的不良效应已经引起了学术界和公众的广泛关注。本研究拟全方位的介绍大气污染对我国公众的健康影响。在对大气污染物不良效应评估的基础上, 利用我国大气污染防治行动计划推出的契机, 在国际上率先对该计划对大气污染进行干预所带来健康效应进行了评估, 该结果为评价和推动我国大气污染的治理提供了可靠证据, 另外, 结合本人使用 R 软件的经验, 分享心得。

利用改进的高斯过程模型预测季节性流感的传播

陈善恩 (北京大学)

时间: 14:30-15:00

简介: 北京大学工业工程与管理系 2016 级博士研究生, 主要从事基于医疗大数据的慢性病预诊、传染性疾病时空传播和复杂系统的可靠性建模。目前以第一作者身份发表 4 篇 SCI 论文, 申请国家发明专利 1 项、国家计算机软件著作权 2 项。曾获北京大学研究生国家奖学金、波音二等奖学金、三好学生标兵、三好学生等荣誉。

摘要: 季节性流感的传播预测对预防流感爆发、保护公众健康至关重要。目前的研究大多集中在流感的时空传播机理建模, 尚未有研究将气象因子纳入流感传播的预测建模中。本研究建立了一个基于改进高斯过程回归模型的非参数流感预测模型, 将气象因子纳入考虑, 以捕捉流感时间序列中隐藏的相关性。为了确定最具解释性的气象因子, 我们首先采用 L1 正则化方法识别最优气象因子子集。基于高斯过程回归模型, 我们设计了三种协方差函数来描述流感活动的非平稳和周期性, 利用设计的交叉协方差函数对流感与气象的相关性进行了建模。最后, 我们利用深圳 CDC 所采集的 2011-2015 年的流感数据对改进高斯过程回归模型进行了验证, 并与现有的流感预测多变量统计模型进行了比较。结果表明, 特定气象因子对流感传播有显著影响, 将气象因子纳入建模过程能显著提高流感预测的准确性。

Selection of mixed copulas for data with ties via penalized likelihood

王钒 (中国人民大学)

时间: 15:00-15:30

简介: 王钒, 中国人民大学博士研究生。本科毕业于加拿大滑铁卢大学统计学专业, 研究生毕业于美国密歇根大学安娜堡分校应用统计专业, 曾在美国制药业从事临床研究及健康数据分析 4 年。目前的研究兴趣主要集中在健康大数据, 基因数据与疾病的相关分析, 函数型回归模型等。

摘要: The link between Obesity and Hypertension is one of the most popular topics which have seen much discussion in recent decades but still difficult to be captured comprehensively and accurately. However, the distribution of BMI and blood pressure is usually fat tailed and severely tied. This paper adopts the ideas from Cai and Wang 1 by using data-driven copula selection approach with penalized likelihood to measure fat tailed correlation, and from Li et al.2 by employing Interval Censoring method to address tied data issue. Minimax Concave Penalty (MCP) is borrowed to perform the unbiased selection of Mixed copula model instead of Smoothly Clipped Absolute Deviation (SCAD), which was used in Cai and Wang1, for MCP is faster to get un-penalized solution. Interval Censoring method, inspired from survival analysis, is applied by considering ranks as intervals with upper and lower limits, and maximizing pseudo- likelihood to get point estimates. This paper describes the model and corresponding iteration algorithm. Also, a simulation to compare the proposed model (Mixed copula model via MCP with Interval Censoring method) versus existing model (Mixed copula model via SCAD with “Jitter” from R package applied to address tied data issue) in different conditions is presented. Additionally, the model is also applied to health data collected from China Health and Nutrition Survey (CHNS). Both numerical studies and real data analysis show positive results.

R 语言在人群为基础的癌症登记数据的生存分析中的应用

安澜 (北京协和医学院)

时间: 16:00-16:30

简介: 2018 年毕业于吉林大学白求恩医学部。现为国家癌症中心/中国医学科学院北京协和医学院肿瘤医院流行病与卫生统计学研究生。主要研究方向为: 人群为基础的高精度肿瘤监测体系构建、人群为基础的癌症生存分析研究。目前主要完成了 2015 年中国肝癌流行情况分析和 Elandt-Johnson 模型推算完全寿命表方法学原理及其在中国人群中的应用, 并在 NAACCR / IACR Combined Annual Conference 汇报。

摘要: 近几十年来, 随着社会的进步与经济的发展, 人类疾病谱发生了巨大变化, 癌症已成为严重威胁人类健康的世界性公共卫生问题。对人群为基础的癌症登记数据进行生存分析是评估癌症疾病负担、评估筛查效果和监测卫生服务绩效的基础, 其对于评估国家癌症治疗方案的疗效和公平性至关重要。本研究将以美国 seer 数据库中的肺癌为例对人群为基础的癌症登记数据的生存分析展开介绍, 分析主要应用 R 软件包 survival, dplyr, survminer 等。

R 语言在大气颗粒污染物健康影响流行病学研究中的应用

林华亮 (中山大学)

时间: 16:30-17:00

简介: 林华亮, 中山大学公共卫生学院, 副教授, 博导; 获得广东省杰出青年医学人才称号。2011 年毕业于香港中文大学公共卫生学院, 获博士学位。中华预防医学会环境卫生分会委员, 媒介生物学分会委员, 中国卫生信息学会卫生地理信息专业委员会会员; 华南预防医学杂志编委。主要研究方向为环境流行学, 对室内外空气污染、气候变化对人群健康的影响有多年的研究经验。发表 SCI 文章 100 余篇, 其中第一/通讯作者 60 余篇, 文章被引用 (Google Scholar) 超过 2900 次, H-Index 达到 34。获得国家发明专利授权 1 项。

摘要: 大气污染和气候变化是我们目前面临的重要的环境问题, 其对居民健康的短期和长期影响近年来引起了公众的广泛关注。但是前期的研究中, 较少考虑这些因素对健康影响的非线性、滞后和累积效应; 本研究利用时间序列分析、分布滞后非线性模型分析大气污染和气候变化对不同健康结局的暴露反应关系。分析主要应用 R 软件包 mgcv, dlnm, ggplot2, dplyr 等。

敏捷数据科学家如何玩转教学闭环产品?

任万凤 (51Talk 无忧英语)

时间: 09:00-09:30

简介: 任万凤, 毕业于北京大学数学学院应用统计硕士, 目前在 51Talk 担任算法专家, 主要研究方向为教学闭环信息化、少儿英语启蒙学习路径、老师推荐等, 业余时间也是教育自媒体大 V, 探索优质的内容在青少儿英语启蒙中应用。在 51Talk 之前是国内第一批 Growth hacker, 擅长用户增长、用户行为分析、精细化运营等, 曾助力多家企业实现增长。曾主要译作《Tableau 数据可视化实战》、《金融风险建模及投资组合优化—使用 R 语言》等书籍。

摘要: 在互联网行业想要一个人玩转全闭环的产品并获得良好的用户体验可能吗? 可能! 但这需要什么样的能力呢? 大部分数据从业人通过数据分析或挖掘助力各环节效率提升, 但在线英语教育行业仍处于低洼状态, 数据的采集和架构都存在较大的问题, 就更提数据应用, 而数据驱动的产品落地才是作为敏捷数据科学家的唯一归属。本次演讲主要通过如何架构底层数据使得小团队玩转整个闭环的教学产品成为可能。通过针对用户的个性化的约课攻略、对内部的魔镜系统和 CMS 教学内容管理系统案例来分解一个全栈数据科学家在设计、实现到推动全闭环教学产品的成功需要什么样的能力, 而这些能力又是在各个环节中起着什么样的作用, 对于一个数据人来说又应如何培养呢?

A New Model of Knowledge Assessment and AI Adaptive Learning

Dan Bindman (Yixue Squirrel AI)

时间: 09:30-10:00

简介: Dan Bindman received his Ph.D. from the Institute For Math Behavioral Sciences at UCI in 2002. He then spent the next 12 years at ALEKS, a pioneer in online adaptive learning focused on Math and Chemistry, where he eventually became Editorial Director and Chief Architect for the Math Products. He is now Chief Data Scientist for Yixue Squirrel AI, the first AI adaptive learning system in China, where the focus is on “after school” courses preparing K-12 students in China for the high stakes tests used there.

摘要: In this talk, Dr. Dan Bindman will describe a powerful new multi-dimensional model for knowledge assessment and learning that he is now working to implement with Yixue Squirrel AI for their adaptive learning products. This new model gives extremely accurate and high resolution predictions at the highest granularity level: after a student has answered only 20 to 25 questions in a given product, the system can accurately predict the student’s probability of answering each question in the product at the current time. And unlike systems such as ALEKS, the model does not require Knowledge Structures linking the topics, saving a huge amount of costs compared to adaptive learning systems that require this step. It also can be used “out of the box” with any mix of question formats (free response or multiple choice) from any mix of subject areas without any content tweaking. Finally, all results so far indicate each student’s knowledge (as represented by the model) can be accurately updated as the student learns and works on questions in the product without the need of periodic reassessments, eliminating a big “pain point” for students that occur with many adaptive learning systems. The results from a real-world large-scale application of the model will be given to show the model’s unprecedented accuracy and predictive power.

AI 在 K12 场景下的应用实践

饶丰 (一起教育科技)

时间: 10:30-11:00

简介: 饶丰, 毕业于北京邮电大学. 前微信语音技术负责人, 目前担任一起教育科技 AI 部门负责人. 负责 AI 技术的研发和应用。

摘要: 如何利用 AI 技术让学习成为美好体验, AI 技术在学习和教育过程中应该扮演一个什么样的角色。我们通过构建 AI 助教这么一个角色来解放老师, 同时赋能学生。我们通过语音评测技术实现了口语作业的自动批改和评测。通过图像识别技术实现了纸质作业的拍照批改, 从而解放老师, 使老师从繁杂的批改工作中解放出来。我们通过 AI 技术进一步分析数据, 从而更好地理解学生, 从而提供更具个性化的, 更高效的教学。本次演讲主要介绍 AI 技术是如何落地在 K12 场景, 以及 AI 技术在实际应用中所遇到的一些问题。

基于电脑使用日志剖析和评估用户拖延行为

何明 (上海交通大学)

时间: 11:00-11:30

简介: 何明, 重庆大学学士, 中国科学技术大学博士, 曾于美国北卡罗来纳大学夏洛特分校访问交流, 目前为上海交通大学电子科学与技术方向博士后研究人员、前 OPPO 研究院人工智能算法研究员。主要研究方向为深度强化学习、数据挖掘与知识发现机器学习方法及其应用, 倾重于移动端用户行为分析与建模。在 TIP、TWEB、DASFAA、IEEE Access 等重要学术会议和期刊共发表论文 10 余篇, 曾获数据挖掘领域国际会议 KSEM2018 的最佳论文奖。

摘要: 在互联网给我们生活和工作带来极大便利的同时, 也给我们的工作带来了各式各样的问题, 如拖延症。尤其是近些年来, 延症给我们的生活和工作产生了众多的负面影响, 如降低工作效率、带来消极情绪等。然而如何评估和量化拖延症, 是一个非常棘手的问题。传统的评估方法主要基于调查问卷, 但调查问卷本身存在准确度低、规模小、成本高等问题, 需要一种更为准确、更为便捷的评估方法。基于此, 我们通过将心理学理论和数据驱动方法相结合, 从电脑使用行为日志中抽取出若干拖延行为特征, 如意志力、专注程度等, 并根据抽取的特征设计了基于 GBDT 和 CLTree 的用户拖延症评估模型, 实验结果清晰表明了抽取特征的有效性和评估模型的准确性。

基于 Unet 的直肠肿瘤识别

涂富艺 (中国人民大学)

时间: 08:30-09:00

简介: 中国人民大学博士在读, 研究方向为实验设计。

摘要: 就肿瘤本身而言, 其可以出现在直肠的任何地方, 可能有任何形状、大小和对比度。这些原因促使我们探索一种机器学习解决方案, 利用一个灵活的, 高容量的, 同时是非常有效的深度神经网络 (DNN) 来实现全自动的直肠肿瘤分割。然而, DNN 的成功训练往往需要数千个带标签的培训样本, 受数据量的限制, 我们进一步采用基于 U-Net 的直肠肿瘤分割算法, 并通过抽取感兴趣区域 (ROI) 中的子图像进行神经网络的训练, 能更有效地利用带标签的样本, 并得到更高的准确率。此外, 针对我们在实验中遇到的具体问题, 我们进一步改进了 U-Net 的方法, 即不再将随机抽取的大量子图像作为模型的输入, 而是直接随机抽取每张图像的 ROI 作为模型输入, 我们验证了这种新方法更加高效和准确。最后我们在 U-Net 模型的基础上增加了边缘区域生长算法, 并对比了原始方法与改进方法的 Dice 系数。

肿瘤影像特征提取分析

叶小清 (中国人民大学)

时间: 09:00-09:30

简介: 中国人民大学博士二年级在读, 研究方向为肿瘤影像特征提取。

摘要: 基于上述对肿瘤 CT 图像的分割, 对肿瘤影像进行特征提取。特征提取的准确性和全面性直接影响淋巴转移分类模型的有效性, 但是, 过多地对特征进行提取, 又会影响淋巴转移分类模型的速度。于是, 我们提取了传统的二维和三维特征, 如: 表面积、体积、形状, 纹理, 小波系数等, 又提取了最大最小横截面积、中心位置等创新特征, 进而通过 PCA 进行降维。

肿瘤影像特征与淋巴结转移的相关性验证

刘晓玉 (中国人民大学)

时间: 09:30-10:00

简介: 中国人民大学博士一年级在读, 研究方向为 Causal Inference, Treatment Effect。

摘要: 在对前述报告所提取的图像特征和病人的临床数据 PCA 降维后, 构造大量弱分类器, 采用基于随机森林和 xgboost 的集成分类方法, 分析了直肠肿瘤区域影像特征与是否淋巴结转移之间的关系, 建立了基于 CT 影像和临床数据的全自动淋巴结转移分类模型并评估且对比了两种模型的表现。

函数型数据变系数模型的估计 (Estimation of varying coefficient model for functional data)

苏蔚 (中国人民大学)

时间: 10:30-11:00

简介: 中国人民大学统计学院硕士在读, 研究方向为 functional data analysis (函数型数据分析)。

摘要: 本文提出了函数型数据变系数模型及其估计和推断, 并将这一模型应用于研究在气象因素的影响下, 沈阳市空气污染源排放情况对于污染物浓度的影响。

高维时间序列数据的降维处理——因子个数的确定研究

夏强 (华南农业大学)

时间: 11:00-11:30

简介: 夏强, 博士, 华南农业大学数学与信息学院, 教授, 研究方向: 时间序列分析, 高维数据分析, 贝叶斯计算。

摘要: For dealing with high-dimensional stationary time series, the factor model is often used to reduce the dimension. In this talk, we suggest a method of determining the number of factors in factor modeling. When the factors are of different degree of strength, the eigenvalue-based ratio method of Lam and Yao needs a two-step procedure to estimate the number of factors. As a modification of the method, however, our method only needs a one-step procedure for the determination of the number of factors. The resulted estimator is obtained simply by minimizing the ratio of the contribution of two adjacent eigenvalues. The finite sample performance of the method is well examined and compared with some competitors in the existing literature by Monte Carlo simulations and a real data analysis.

机器学习在 LBS 中的应用

周海鹏 (*TalkingData*)

时间: 11:30-12:00

简介: 周海鹏, 中科院硕士毕业, 一直从事云存储、云计算开发及架构工作, 专注于分布式存储、分布式计算、大数据分析等方向。长期从事 IT 技术工作, 历经 10 年发展, 从实践到理论均有所积累。现任 TalkingData 技术副总裁, 从事大数据计算平台工作, 对分布式存储和分布式计算、VLDB、大数据分析等有深刻实践, 主持研发实时流式 OLAP 计算框架, 分布式索引, 分布式查询系统。同时关注高可靠、高可用、高扩展、高性能系统服务, 以及 Hadoop/HBase/Storm/Spark 等离线、流式及实时分布式计算技术。参与多次大数据论坛, 在业内具有一定的影响力。

摘要: 位置服务在经济中起到越来越重要的作用, 但是传统 LBS 服务多数是基于纯地理数据支持商业分析的, TalkingData 基于人本数据, 结合机器学习等手段, 整合了数据、算法, 可以更实时、全景地观察现实世界, 提高了分析效率、增强了分析的客观性。

解决的核心问题是:

- 1、如何利用大数据, 解决线下世界人口、客流、画像等观测问题。这块主要是解决像零售、政府、金融等机构对现实世界进行调研的需求。
- 2、如何利用人工智能, 结合上述的现实世界人口、客流、画像, 结合传统的 POI、道路、交通等数据, 对现实世界, 产业布局进行分析的需求。
- 3、如果利用产业布局的分析结果, 在新兴城市、新兴市场中快速找到合适的渠道网络、定位人群、推广营销等。
- 4、风控数据模型提升应用
- 5、R 语言一键式建模探索

“金融科技”的春天

赵然 (中信建投研究发展部)

时间: 09:00-09:30

简介: 赵然, 中信建投研究发展部非银金融、前瞻研究组首席分析师。中国科学技术大学统计与金融系硕士。曾任中信建投金融工程分析师, 2018 年 wind 金牌分析师金融工程第 2 名团队成员。研究成果包括: 借鉴海外资管公司方法论, 结合国内资本市场实情, 利用机器学习等量化算法, 构建了自上而下, 从因子到资产, 从宏观到行业的全球大类资产“战略 + 战术”配置框架。目前专注于金融科技领域(资产配置平台、智能投研、智能投顾、金融信息服务、金融大数据等)的研究。

摘要: 金融科技伴随数字技术的发展应运而生, 越来越多的互联网巨头、综合金融集团和创业公司都投入到了这场新的战役。在未来数年或将影响传统的金融生态, 重塑金融业现存商业模式。对于金融机构来说, 如何拥抱变革, 重塑现有的盈利模式, 构建新的生态系统至关重要。跨界合作, 与不同行业进行融合, 通过科技手段进行全方位数字渠道的覆盖, 最终建立新金融商业模式。金融服务本身也是一种企业服务, 金融服务与产品的边界性将会大大的削弱, 如何基于 B 端机构和 C 端个人投资者的需求, 提供个性化的金融服务将是金融科技领域重要的用武之地。本报告将简介金融科技领域面临的机遇和挑战。

融资融券与 A 股收益率预测性

郭彪 (中国人民大学)

时间: 09:30-10:00

简介: 郭彪, 中国人民大学财政金融学院副教授, 应用金融系副系主任, 曾就读同济大学(经济学本科)、德国 Konstanz 大学(经济学硕士)、瑞士苏黎世联邦理工大学(金融学硕士)、英国诺丁汉大学(金融学博士)。曾在国内外著名对冲基金公司 Man Group、金融软件公司和量化咨询公司从事量化研究工作。主要研究领域为资产定价和风险管理。在 Journal of Futures Markets、Journal of Financial Research、Finance Research Letters 等核心学术期刊上发表十余篇学术论文, 并主持和参与多个国家自然科学基金、四川省金融局、进出口银行、大连商品交易所等项目。

摘要: 基于 A 股市场融资及融券余额的巨大差距, 我们拓展 Hong 等 (2015) 理论模型, 得出影响股票收益率的变量: 融券比率(融券余量/流通股数)和融资回补天数(融资比率/日均换手率)。使用 FM 回归及因子分析, 我们实证检验 A 股市场中融券比率与融资回补天数的解释和预测股票收益率的能力。实证结果表明, 在存在融券限制的条件下, 融券比率相比融券回补天数能更好的代表套利者对股票价格高估的看法; 而由于融资约束相对较低, 融资回补天数相比融资比率(LR)能更好的代表套利者对股票价格的低估看法。实证结果与存在融券数量限制下的理论模型相符。

风险平价资产配置 -以商品期货、可转债为例

李孟育 (南华期货股份有限公司)

时间: 10:30-11:00

简介: 专长: 统计计算、数据分析、衍生品定价、量化分析、科技管理。现任职于南华期货股份有限公司期货研究所量化投资组。曾任职台湾国立嘉义大学金融系助理教授、金融工程公司知识长等岗位。取得台湾国立交通大学资讯管理博士 (双辅修: 统计、应用数学)。曾经主持台湾国科会专题研究项目、学术论文发表于 SCI/SSCI 等国际期刊, 曾经获得多个研讨会最佳论文奖。

摘要: 本文以风险平价 (Risk Parity) 方法来进行两类投资组合的资产配置, 并且说明所使用的 R 包。第一类是商品期货。南华商品指数是中国第一个商品指数, 也是广为被国内投研机构所引用的标杆。本研究以波动率来进行风险预算, 进行权重配置, 并且跟原先商品指数作比较。第二类是可转债, 本研究利用可转债的两个风险指标: 标的股票收益率方差、可转债相对应的 delta 来进行资产配置。

金融科技公司如何利用人工智能技术进行风控

汪昊 (汉升链商)

时间: 11:00-11:30

简介: 汪昊, 区块链公司科学家, 前恒昌利通大数据部负责人, 美国犹他大学本科/硕士, 在百度, 新浪, 网易, 豆瓣等公司有多年的研究和技术管理经验, 擅长机器学习, 大数据, 推荐系统, 社交网络分析等技术。在 TVCG 和 ASONAM 等国际会议和期刊发表论文 10 篇。本科毕业论文获国际会议 IEEE SMI 2008 最佳论文奖。ACM/ICPC 北美落基山区域赛金牌第三名。

摘要: 金融科技大潮近年来汹涌澎湃, 催生了诸多的科技应用场景。风控作为金融信贷环节不可缺少的一环, 帮助金融科技排除了大量的欺诈风险。为风控服务的人工智能技术的应用由来已久, 包括逻辑回归, GBDT, 混合模型以及最新的深度学习技术越来越广的应用在风控的各个场景中。本次演讲将结合风控在金融科技公司落地的具体案例, 介绍如何利用大数据和人工智能帮助企业排除欺诈用户, 排除金融风险。

凯利公式 -用胜率和赔率量化你的投资

张丹 (青萌数海)

时间: 11:30-12:00

简介: 张丹, 青萌数海 CTO, 微软 MVP。资深 R 语言技术专家, 在国内 R 语言技术社区的领军人物。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)。

摘要: 职业做投机交易的人, 应该都听说过凯利公式, 这是一个通过计算胜率和赔率, 来选择最佳投注比例的公式, 目的是长期获得最高的盈利。只要找到长期看必胜的局, 接下来就是让时间帮我们赚钱了。

设游戏赢的概率是 80%, 输的概率是 20%, 赢时的净收益率是 100%, 输时的亏损率也是 100%。如果赢, 你每赌 1 元可以赢得 1 元; 如果输, 则每赌 1 元将会输掉 1 元。赌局可以进行无限次, 每次下的赌注可由你自己任意定。如果你的初始资金是 100 元, 那么怎么样下注, 才能使得长期收益最大?

让我们用 R 语言来实现凯利公式, 找到答案!

基于警务大数据的犯罪事件智能预测

张耀峰 (湖北经济学院)

时间: 09:00-09:30

简介: 张耀峰, 博士、教授, 现任湖北经济学院湖北数据与分析中心主任, 中国现场统计研究会大数据统计分会秘书长, 中国商业统计学会数据科学与商务智能分会副秘书长, 武汉烽火普天信息技术有限公司等多家企业合作伙伴和高级顾问, 研究方向为商业大数据、舆情大数据、警务大数据等。

摘要: 随着警务信息化的发展, 公安大数据平台建设日益完善, 为利用大数据技术开展智慧警务研究提供了基础。本报告利用 2015-2018 年武汉市近 70000 条入户盗窃的 110 报警数据, 分别利用深度学习方法对入户盗窃犯罪行为进行预测研究。通过数据提取、数据预处理、模型训练等过程, 预测下一时间段犯罪行为发生的具体地点, 预测准确率达 50% 以上。根据该研究结果, 已经申请了国家专利 1 项, 为武汉市公安局大数据实战应用中心开发可视化大屏一块, 2018 年 7 月大屏系统上线以来运行良好。

聚类方法的评价研究

张忠元 (中央财经大学)

时间: 09:30-10:00

简介: 张忠元, 中央财经大学统计与数学学院教授, 博士生导师。

摘要: Clustering analysis is critical towards understanding the hidden patterns behind the data. However, how to evaluate the quality of different clustering methods is still challenging and remains unsolved. The most widely used metric, normalized mutual information (NMI), was proved to have finite size effect, and its improved form relative normalized mutual information (rNMI) has reverse finite size effect. Corrected normalized mutual information (cNMI) was thus proposed and has neither finite size effect nor reverse finite size effect. However, in this paper we show that cNMI violates the so-called proportionality assumption. In addition, NMI-type metrics have the problem of ignoring importance of small clusters. Finally, they cannot be used to evaluate a single cluster of interest. In this paper, we map the computed cluster labels to the ground-truth ones through integer linear programming, then use kappa index and F-score to evaluate the clustering methods. Experimental results demonstrate the advantages of our method.

Online learning 在大规模机器学习中的理论与应用

蔡锐 (宾州州立大学)

时间: 10:30-11:00

简介: 蔡锐, 曾于统计之都打杂, 后远遁美国宾州苦寒之地读博, 现为宾州州立大学博士学生。

摘要: Online learning 是处理大数据时很常用的一种方法。与分布式计算不同, online learning 在有内存和存储限制的情况下依然可以实现对大数据的处理。具体来说, 当新的数据点或数据集到达的时候, 我们依据这些新的数据来更新模型中的参数, 从而实现对数据的建模。本次报告将回顾 online learning 发展历程, 主要关注点, 理论性质, 实际应用, 以及本人的一点微不足道的工作。

神经影像大数据中机器学习的应用

曾梓龙 (北京师范大学)

时间: 11:00-11:30

简介: 曾梓龙, 北京师范大学认知神经科学与国家重点实验室硕士生, 研究涉及脑连接组、神经影像数据处理和机器学习。

摘要: 随着神经影像数据的日益丰富, 越来越多的研究者采用机器学习算法或者深度学习算法进行神经科学的研究。本次报告将介绍机器学习算法如何应用到大脑皮层的划分任务以及婴幼儿自闭症的早期诊断任务。

使用 Rstudio 结合 RcppArmadillo 制作可以快速随机计算稀疏矩阵奇异值分解的包

李家郡 (中国人民大学)

时间: 9:00-9:30

简介: 李家郡, 中国人民大学统计学院 2015 级本科生, 主修统计学, 辅修计算机科学技术, 编程、羽毛球、足球、骑行爱好者。主要研究方向包括大数据流算法、矩阵计算, 有多年 C++、R、Python、Java 编程经验, 初步开始探索制作个人博客 <https://llijiajun.github.io/github-io/>。

摘要: 制作 R 包是件挺有意思的事, 但仅仅使用 R 语言制作的 R 包在计算一些结果时, 很难达到令人满意的效率。利用 Rcpp 手段制作高效运算包可以很大程度上提高工作效率。而矩阵的奇异值分解问题可以广泛应用于各个领域, 随着社交网络、自然语义识别等领域的发展, 这些问题往往面临着求解大规模稀疏矩阵奇异值分解的问题, R 语言现有的奇异值分解手段仍有优化空间。本演讲基于制作 R 包的过程展开, 以快速随机计算稀疏矩阵奇异值分解为例, 探究使用 Rcpp 结合现有矩阵代数库 Armadillo 实现 R 包中可能遇到的问题, 并借此倡导广大 R 语言和其他兼修多种语言使用者加入到扩充 R 包的队伍中, 以实现更多有趣的课题。

伊辛图模型组合结构推断问题的计算 -统计权衡

金滢 (清华大学)

时间: 9:30-10:00

简介: 金滢, 清华大学数学科学系 2015 级本科生, 主修概率与统计方向, 美国斯坦福大学统计博士 2019Fall 录取。2017 年 9 月加入清华统计学研究中心教授团队研究非平稳时间序列理论, 一作正撰稿一篇; 2018 年 3 月起在西北大学运筹系及哈佛大学生物统计系教授指导下研究图模型结构推断理论, 一作正撰稿一篇; 2018 年 7-9 月赴哈佛大学统计系进行网络分析模型算法相关暑期科研。现有研究方向主要为图模型和网络分析、高维统计、时间序列。

摘要: In various problems, data are represented and interpreted in the form of graphs, where conditional dependency relationships among nodes are illustrated by edges between pairs of nodes. Ising model is a popular kind of graphical model that is suitable for binary-valued nodes. Under the framework of oracle model, which captures the computational properties of a wide range of algorithms, we focus on inferring combinatorial structures of Ising model, for example, the existence of an s-clique in a given graph. We are interested in the detection limits both with or without polynomial computational limits, equivalently, the computational-statistical tradeoffs in this problem.

In our work, we establish the general computational lower bounds for inferring structural properties in simple zero-field Ising model, under which no polynomial-time queries can distinguish between two hypotheses. We found that the bound is related to certain structural property of the underlying graph which we define as vertex overlap ratio. Meanwhile, we propose specific query functions and test functions with polynomial computational budgets which attain the lower bounds for specific problems. We also discussed the relationship between information-theoretic limits and computational-efficient limits. We found that for clique detection and nearest neighbor graph detection problem, there is a gap between computational efficiency and statistical accuracy. However, for relatively sparse structure like perfect matching, there is no gap between computational efficiency and statistical accuracy.

基于岭比收缩的高光谱图像端元个数估计方法

康越 (西安交通大学)

时间: 10:30-11:00

简介: 康越, 西安交通大学数学与应用数学专业大四在读, 毕业后将前往加州大学戴维斯分校攻读统计学博士, 研究兴趣为高维数据分析及其在各个领域的应用。在 IEEE Transactions on Geoscience and Remote Sensing 杂志上投稿文章一篇, 运用统计学知识对高光谱图像端元个数的估计问题提出了更为高效的方法。

摘要: 高光谱图像端元个数的估计问题一直都是该领域的核心问题之一。经典的算法如 HFC、HySime 等对于很多实际数据的模拟效果很差, 这与他们对算法内部的参数以及噪音类型的高敏感度有很大关系, 并且本身的理论也普遍存在一定缺陷。本研究着眼于利用统计理论构造了一种克服上述缺点新方法 (thresholding ridge ratio criterion), 并且在模拟与实际数据中表现出了很大的优越性。

基于前列腺分割任务的 DDSP 网络设计和损失函数探讨

刘秋华 (中国人民大学)

时间: 11:00-11:30

简介: 刘秋华, 中国人民大学信息学院 2015 级本科生, 主修应用数学, 美国密歇根大学生物统计硕士 2019FALL 录取。于 2017-2018 年间获两次北美大学生数学建模竞赛一等奖、一次全国大学生数学建模竞赛国家二等奖; 于 2017 年 3 月加入中国人民大学数学科学研究院数学智能应用实验室, 期间参与三个科研项目, 目前学生二作 BMC Systems Biology 发表一篇, 一作投稿 SIAM Journal on Imaging Science 一篇; 于 2018 年 6-9 月到美国斯坦福大学参加暑期科研, 目前一作正撰稿一篇。研究方向主要为深度学习领域数学统计理论、医疗图像处理、机器学习建模等。

摘要: The high incidence rate of prostate disease poses a requirement in early detection for diagnosis. As one of the main imaging methods used for prostate cancer detection, Magnetic Resonance Imaging (MRI) has wide range of appearance and imbalance problems, making automated prostate segmentation fundamental but challenging. Here we propose a novel Densely Dilated Spatial Pooling Convolutional Network (DDSP ConNet) in encoder-decoder structure, which employs dense structure to combine dilated convolution and global pooling, thus supplying coarse segmentation results from encoder and decoder subnet and preserving more contextual information. Furtherly, to obtain richer hierarchical feature maps, residual long connection is adopted to fuse contexture features. Meanwhile, we adopt DSC loss and Jaccard loss functions to train our DDSP ConNet and we surprisingly found and proved that, in contrast to re-weighted cross entropy, DSC loss and Jaccard loss have a lot of benign properties in theory, including symmetry, continuity and differentiability about the parameters of network. To corroborate the effectiveness of our DDSP ConNet with DSC loss and Jaccard loss, extensive experiments have been done on the MICCAI PROMISE12 challenge dataset. In the test dataset , our method achieves a score of 85.78, outperforming most of other competitors.

从零开始的 COplay

林枫 (中国科学技术大学)

时间: 11:30-12:00

简介: 统计之都副主编, 中国科学技术大学在读硕士, 感兴趣的领域包括机器学习、优化理论等。

摘要: 统计之都 (COS) 一直致力于推广与应用统计学知识, 分享有趣的数据科学方法与观点, 为国内统计学和数据科学的发展贡献力量。我们也一直欢迎有志于将统计学和数据科学发扬光大的各界朋友加入到这个大家庭来。让我们一起探秘 COS 编辑部时尚时尚最时尚的在线投稿系统, 解锁 COS 丰富的“玩法”, 从零开始一场统计学与数据科学的精彩 PLAY!

R 在心理学中的应用

吕杰好 (中央财经大学)

时间: 09:00-09:30

简介: 现任中央财经大学心理学系讲师, 心理学博士, 本硕毕业于西南大学心理学部, 博士毕业于英国伦敦玛丽女王大学 (Queen Mary University of London) “动态学习与决策实验室”。美国判断与决策协会成员, 欧洲决策协会成员。研究方向为合作行为的机制及影响因素。迄今为止, 科研方面, 主持科研项目 4 项、发表英文学术论文 SCI/SSCI6 篇。教学方面, 主持教学项目 1 项, 此教学项目为 2018 年中央财经大学新开课程建设项目“基于 R 的心理统计及其可视化”, 参与教学项目 1 项, 主讲课程为 English Academic Writing, 决策心理学, 财税心理学和 R 在心理学中的应用。R 在心理学中的应用的课程开设面向本科生、硕士生。

摘要: 随着数据科学中的重要作用, 具有强大的统计计算功能和便捷的数据可视化系统的 R 统计在各个学科中得到了广泛的应用。基于实证研究中的心理学也日益显现了 R 在科学研究中的数据分析的重要性。R 的免费、开源、灵活性等特性也使依赖于数据分析的心理学实现更为深度的学习。同时开放科学的普及, 在科研研究论文在国际高水平期刊中发表时, 有些杂志要求上传原始数据和数据分析情况, 采用 R 软件能够更加促进知识交流。因此本次报告中第一部分数据统计分析, 将结合心理学常用的实验法中, 如何使用 R 计算出在心理学论文中所需要报告的描述统计、推论统计; 第二部分数据可视化, 以 ggplot2 介绍如何使用 R 来进行单因素实验设计、多因素实验设计的实验结果中的 bar chart 和 line chart 的图形的呈现。第三部分, 元分析, 简单介绍如何使用 R 来进行元分析研究。通过这三个部分的简单介绍, 希望帮助和引导心理学专业的本科生、研究生以及研究者开始使用 R 软件来进行数据分析, 从而达到使心理学研究者们更便捷地进行交流和学习。

“乱世重才，治世重德”——经济不确定与德 - 才偏好

高树青 (北京师范大学)

时间: 09:30-10:00

简介: 北京师范大学博士, 主要研究领域为社会心理学、人格心理学、进化心理学与网络心理学, 主要研究内容包括: 网络情绪表达的地区差异及其社会生态机制, 空气污染与风险决策、收入不平等与道德判断以及价值观关系等内容。擅长结合实验室研究与宏观网络数据, 探讨宏观社会生态变量对人们心理与行为的影响。

摘要: 近些年来, 经济全球化越发深入, 各国经济发展联系越发紧密, 与此同时, 全球经济的不确定因素也明显增加。越来越多的研究开始关注经济不稳定对人们心理和行为的消极影响。在社会价值观念中, 道德观念能够引人向善, 促进社会的和谐与稳定。人们对于道德与才能的偏好不仅体现了人们在自我建构时如何权衡发展, 同时反映了一个时期内社会道德观念是否处于最重要的位置。过分推崇才能势必导致道德观念的下滑。已有研究者指出, 在过去的几十年间, 美国正在经历道德的衰落, 那么中国的情况如何? 中国德 - 才观念的变化机制是什么成为重要的研究问题。本研究认为宏观经济的不稳定可能促进了人们德 - 才观念的变化。经济不稳定使得个体具有更强的控制感需求, 在德 - 才偏好中, 更注重能够自我获益的, 可控的才能品质, 降低了对道德品质的亲和。因此, 在经济不稳定的情况下, 个体更加偏好采用才能进行自我建构, 认为对自己而言, 才能比道德更加重要。本研究通过三个研究对这个问题进行了探讨, 研究一通过宏观层面的公开数据考察了经济政策不确定与道德和能力词汇搜索比率之间的关系。研究二通过新闻材料操纵被试对于中国经济环境不确定程度的认知, 同时, 采用道德和能力词汇追选的方式, 探究经济不确定对被试描述自我时选取道德和能力词汇偏好的作用。研究三更加聚焦个体层面经济不确定的作用, 通过自传体回忆范式启动被试的经济不确定, 进一步探究经济不确定对个体道德和能力评分的影响, 同时考察控制感在二者间的中介作用。三个研究证明了在宏观和个体水平, 经济不确定性增加了人们对于德 - 才中才能的偏好。

人们对社会与金钱奖赏的预期共享神经环路：结合多种分析方法的脑成像元分析研究

黄文昊（首都师范大学）

时间：10:30-11:00

简介：黄文昊，首都师范大学心理学院研究生，主要的研究兴趣为情绪和社会认知，通过结合行为，脑成像和药理学实验等技术手段探索社会行为的神经机制以及其与情绪的相互影响。

摘要：在日常生活中，作为各种目标导向行为的诱因，社会性奖赏与物质性奖赏扮演着同样重要的角色。近年来，大量的神经影像学研究在尝试回答这样一个问题，人类的大脑中是否存在特定的神经回路来表征社会性奖赏，或者是社会性奖赏与物质性奖赏以类似的形式在大脑中编码。本次报告将为大家介绍结合多种分析方法的脑成像元分析在解答这一问题上的尝试，此外，将以本研究为例介绍激活似然性估计方法（Activation Likelihood Estimation, ALE）的基本原理和操作，以及功能解码（functional decoding）和脑连通性元分析模型（meta-analytic connectivity modeling, MACM）等分析方法在元分析研究上的应用。

心理学脑电研究中数据的基本概念与分析实践

夏晓磊（中国科学院心理研究所）

时间：11:00-11:30

简介：夏晓磊，中国科学院心理研究所在读博士，擅长脑电实验设计和数据分析技术，使用脑电技术以第一作者/并列一作身份在疼痛领域顶尖杂志 Pain 上发表论文 2 篇（含封面文章 1 篇）、在神经成像领域优秀期刊 NeuroImage 上发表论文 1 篇。

摘要：脑电图（EEG）技术具有安全无创、价格便宜、容易操作的特点，在基础研究和临床上的应用越来越普及。本次报告将对脑电原理、实验和数据分析实践中涉及的一些问题进行探讨，对初学者经常犯的错误和有疑问的地方做重点讨论，帮助大家澄清一些基本问题、正确利用脑电技术服务好自己的科研事业。

基于 OSF 增强研究中的开放科学

赵加伟（天津师范大学）

时间：11:30-12:00

简介：赵加伟，天津师范大学应用心理专业硕士研究生；Open Science Club 成员。

摘要：心理学界重复性危机至少有 60 年的历史（Sterling, 1959），且这场危机仍在持续。自 2011 年以来，心理学研究者越来越意识到危机的严重性，Open Science Framework（OSF）是对此危机中搭建的用于开放科学的平台。这是一个免费的 web 应用程序，研究人员在 OSF 中可从头到尾管理他们的研究，如预注册（pre-register）、公开数据与材料、远程合作。预注册简单形式包含基本研究设计，也可以包括研究程序、结果和统计分析计划的详细预先说明。OSF 提供一个受认可的预注册平台，注册后可作为研究凭证。同时，OSF 上可以实现多人远程协作，免费存储数据且自带版本控制。此外，OSF 与其它平台能无缝结合，如 Amazon、Box、Google Drive、Github。

机器学习在生物信息中的应用

连明 (中国科学院北京基因组研究所)

时间: 09:00-09:30

简介: 本科毕业于天津医科大学, 生物信息学专业。目前就读于中科院北京基因组研究所, 基因组学方向, 硕士二年级, 研究涉及药物基因组, 宏基因组, 转录组和机器学习。

摘要: 简述目前机器学习在生物信息学领域的一些典型和成功的案例, 比如无监督聚类算法在宏基因组、单细胞测序数据中的应用, 传统机器学习分类算法在临床样本诊断中的应用, 等等。另外会结合本人在项目研究中遇到的一些问题, 谈谈自己的经验教训, 例如, 数据拿到手之后为什么不应该直接就开始建模, 而是应该先对数据的一些分布特点进行观察, 等等, 在此就不赘述了。

Platform-independent approach for cancer detection from gene expression profiles of peripheral blood cells

张韬 (中国科学院北京基因组研究所)

时间: 09:30-10:00

简介: 张韬, 男, 中国科学院北京基因组研究所, 生物信息学专业硕士研究生。

发表论文:

(1) Yadong Yang, Tao Zhang, Rudan Xiao, Xiaopeng Hao, Huiqiang Zhang, Hongzhu Qu, Bingbing Xie, Tao Wang and Xiangdong Fang: Platform-independent approach for cancer detection from gene expression profiles of peripheral blood cells. *Briefings in Bioinformatics* 2019, doi: 10.1093/bib/bbz027. (IF= 6.302)

(2) 张韬, 杨亚东, 方向东. 应用于精准医学研究的转录组可变剪接分析 [J]. 发育医学电子杂志, 2016, 4(2):78-84.

授权专利:

(1) 2017109863490. 一种获取外周血基因模型训练数据的方法及装置。方向东, 杨亚东, 张韬. 2017.

摘要: 肿瘤是系统性疾病, 在肿瘤发生发展过程中, 除病灶位置外, 外周血中多种细胞的表达量也发生变化, 这使得我们有机会通过外周血在分子层次实现对肿瘤的追踪。过往研究发现血小板转录组在预测肿瘤中的作用, 但限于标准化方法不统一、数据集小等原因, 取得的模型很难扩展到独立的数据集中。我们一方面整合外周血全细胞转录组, 最大程度地降低了单一细胞类型受特定环境影响所导致的非特异性变化, 另一方面开发了创新性的秩归一化方式以屏蔽不同转录组检测平台、不同批次之间的噪音, 并通过人工智能方法实现高维数据特征筛选和模型构建, 大幅提升了可整合的数据量和肿瘤分类效果。这是首次通过大规模人群的外周血转录组数据来区分正常人和乳腺癌患者。

Analysis for Ribosome Profiling Data

李发金 (清华大学)

时间: 10:30-11:00

简介: 李发金, 清华大学生命学院二年级博士生, 主要研究方向为蛋白质的翻译调控。

摘要: Ribosome profiling is a technology used for sequencing the mRNA fragments protected by ribosomes during the process of decoding by translation, through which we could identify the potential translated regions on mRNA, measure the protein synthesis and find out some translation events ignored by other proteomic technologies such MS. In my speech, I will introduce some details about ribosome profiling technology and some methods our lab developed before used for analysis for ribosome profiling data.

利用蛋白质组学和生物信息学研究 $\text{PKC}\zeta$ 相互作用蛋白网络

侯春宇 (北京大学医学部)

时间: 11:00-11:30

简介: 侯春宇, 北京大学医学部第一临床医学院, 张宁教授实验室博士一年级学生, 攻读肿瘤学博士。硕士毕业于天津医科大学, 师从张宁教授。主要研究方向为肿瘤的转移。在硕士期间, 以第一作者发表 SCI 文章 2 篇, 获得 2017 年研究生国家奖学金, 天津医科大学优秀毕业生的称号。

摘要: 乳腺癌是一种威胁性极强的恶性肿瘤, 严重影响女性健康。对晚期的乳腺癌患者来说, 转移仍然是临幊上病人死亡的主要原因。蛋白激酶 C ζ (Protein kinase C ζ , PKC ζ) 是非典型蛋白激酶 C 的异构体, 是癌症中的关键调节因子。然而, PKC ζ 蛋白分子在调控肿瘤发生和转移的分子和细胞机制尚未完全了解。在这项研究中, 我们结合蛋白质组学和生物信息学的方法分析建立了 PKC ζ 的相互作用蛋白质 (Protein - protein interaction, PPI) 网络, 进一步了解 PKC ζ 在乳腺癌中的生物学作用。

Genome-wide and cell type-specific pattern of transcriptional regulators cooperation in 3D chromatin

伊现富 (天津医科大学)

时间: 11:30-12:00

简介: 伊现富, 天津医科大学教师。2012 年毕业于中国科学院上海生命科学研究院, 获遗传学博士学位, 2012 年至今于天津医科大学担任生物信息学专业教师。教学方面致力于生物信息学专业必备技能和组学数据分析的教授, 力图通过浅显易懂的讲解引领初学者入门并喜爱上生物信息学专业, 所有教学资料共享在 GitHub 中 (<https://github.com/Yixf-Education>)。科研方面主要从事组学数据的整合与分析、肿瘤等复杂疾病的研 究, 目前已在 Nucleic Acids Research 等杂志上发表多篇文章。

摘要: The intact cooperation of transcriptional regulators (TRs), including transcription factors, histone modifying enzymes and chromatin remodelers, precisely determine gene expression in the cell nucleus. Deciphering the relationship among these TRs in the context of 3D chromatin and specific cell type will facilitate the understanding of transcriptional regulation. In this study, we present a computational analysis to comprehensively investigate TR cooperations by integrating genome-wide Hi-C and 266 TRs ChIP-seq data in K562. We uncover lots of previously reported and unknown TR cooperations in 3D gene regulation. To generalize the analysis for cell types with limited TRs ChIP-seq data, we develop a novel strategy that incorporates Dnase I hypersensitive site, sequence motif and TR co-expression network to predict TR cooperations in particular cell type. Using Hi-C data from 7 human cell lines, we discover many shared and unique combinatory roles of TRs in sustaining different patterns of super enhancer, chromatin state and gene expression. Benefit from 3D genome data of ES, NPC and CN cell types, we evaluate our strategy and get several solid TR cooperations during mouse neural development. Our strategy can be used on any tissue or cell type, and the results will be an invaluable resource for transcriptional regulation research.

风电大数据落地应用实践

崔鹏飞 (昆仑数据)

时间: 14:00-14:30

简介: 昆仑数据数据分析主管, 毕业于哈尔滨工业大学控制科学专业, 目前主要从事于新能源领域的故障诊断, 图像处理等方面的数据分析工作, 拥有图像处理方面多项国内发明专利。

摘要: 从风电行业大数据分析从业者的角度, 结合风电行业大数据实际案例, 讲述风电大数据落地过程的机遇与挑战, 内容包括风电行业的数据情况、分析场景、现实中大数据分析面临挑战以及实际案例。

数据分析中的并行计算浅谈

董兆宇 (昆仑数据)

时间: 14:30-15:00

简介: 数据分析师, 曾任职金风科技, 现任职昆仑数据, 任职期间著有 30 余篇软件著作权和 3 篇发明专利, 目前主要专注于工业大数据的数据分析方向。

摘要: 介绍并行计算的基础知识, 概述主流的并行计算模式和框架, 内容中包括有实际问题的并行化过程。

数据分析在家电行业内的应用

于亚杰 (量奇科技 (北京) 有限公司)

时间: 15:00-15:30

简介: 量奇科技 (北京) 有限公司创始人兼总经理, 中国地质大学软件工程软件硕士, 有十余年行业经验, 工业大数据相关工作经验, 主要针对装备制造业、流程制造业等行业。2007 年任清华大学软件学院信息系统与工程研究所助理研究员; 2015 年在北京优医宝信息技术有限公司担任 CTO; 2016 年就职于昆仑智汇数据 (北京) 科技有限公司担任解决方案经理。2017 年成立量奇科技 (北京) 有限公司专注做工业大数据和企业信息化系统。

摘要: 从企业实际应用出发, 讲述数据分析在家电行业内的重要性和应用案例。

智能制造在航空工业的探索

陈肇江 (百分点公司)

时间: 16:00-16:30

简介: 数据建模师, 曾任职于航空工业 301 所、现任职于百分点公司, 主要从事航空装备管理信息化、航空智能制造标准研制、系统集成工作。目前主要专注于工业大数据分析工作。

摘要: 分析航空制造特点、航空智能制造标准体系框架、航空工业大数据分析探索与实践, 结合案例讲解飞机全生命周期从设计、制造、使用、维修涉及的主要业务、数据及数据分析建模的要点。

大数据时代的需求预测

陈艺天 (*Bigo*)

时间: 14:00-14:30

简介: 陈艺天, Bigo 算法专家, 前京东资深算法工程师。在京东期间, 主要从事电商统计建模与运筹优化相关的数据建模工作, 带领团队搭建着京东第一个动态定价系统, 该系统目前管理着京东自营约 50% 70% 中长尾商品的自动价格管理; 参与京东的销量预测系统的优化, 率先搭建起基于神经网络的预测模型框架, 该框架提升原京东物流单量约 50% 的准确度。多次受邀去 KDD 作技术报告。

摘要: 时序预测在诸多的商业决策中扮演着至关重要的角色; 在大数据爆炸的今天, 许多公司与机构面临对成千上万相关时序的预测问题: 如电商平台, 需要对未来一段时间所有商品销量做预测, 物流行业对库房每天的出单量作预测, 交通领域对一个城每条的街道的车流量作预测, 等等。在诸如此类的场景中, 应用经典的时序模型, 如基于 State space model 的 ARIMA, Exponential smoothing 模型, 又或是基于 Bayesian 方法的结构化时序模型, generalized additive model, 面临着诸多的挑战: 一方面大量的数据使得对单一时序的建模与调参变得不太现实; 另一方面, 在这些场景中, 其中大量的时序存着需要冷启动预测又没或数据稀疏的问题 (如电商平台每周都会有新产品上架), 而经典时序模型在这种场景下变得无能为力; 在本次的报告中, 我们将展现我们设计的基于时序因果卷积的深度学习模型, 并提出了基于该模型的两种概率预测的框架。中国最大的零售商的电商销售数据和物流数据的实验, 表明该模型可以大幅度提升预测准确度; 我们同时也在公开数据集和当前时序最新 state-of-the-art 的模型作了比较, 结果表明我们的框架在准确度和计算效率都大幅度优于其他模型。

Feature-based time series forecasting

Kang Yanfei (北京航空航天大学)

时间: 14:30-15:00

简介: Dr. Yanfei Kang is Associate Professor of Statistics at Beihang University in China. Prior to that, she was Senior R&D Engineer in Big Data Group of Baidu Inc. Yanfei obtained her Ph.D. degree at Monash University in 2014. She worked as a postdoctoral research fellow during 2014 and 2015 at Monash University. Her research interests include time series forecasting, time series visualization, statistical computing and machine learning.

摘要: Feature-based time series representation has attracted substantial attention in a wide range of time series analysis methods. Recently, the use of time series features for forecast model selection and model averaging has been an emerging research focus in the forecasting community. That calls for a more diverse time series benchmarks as the training data to model time series features and forecasting algorithm performances. We propose GeneRAting TIme Series with diverse and controllable characteristics, named GRATIS, with the use of mixture autoregressive (MAR) models. We generate sets of time series using MAR models and investigate the diversity and coverage of the generated time series in a time series feature space. Efficient Bayesian surface regression is used on the generated data to examine how time series features influence forecasting method performances, which enables us to predict the performances of the forecasting methods on test data. We illustrate the usefulness of our time series generation process and feature-based forecasting scheme with their applications on the M3 forecasting competition data.

Probabilistic forecasting based on time series features

Wang Xiaoqian (北京航空航天大学)

时间: 15:00-15:30

简介: 北京航空航天大学经济管理学院统计学专业在读博士。研究方向包括时间序列分析, 统计建模和机器学习。

摘要: The surge of time series data in the big data era leads to an explosive demand for time series forecasting methods. Compared with point forecasting, the literature on probabilistic forecasting, which can provide a comprehensive outlook of the expected future value and the future uncertainty, is highly limited. In this paper, we propose a general feature-based probabilistic forecasting framework, which is divided into “offline” and “online” parts. In the “offline” part, we explore how time series features affect the probabilistic forecasting accuracy of different forecasting methods by generalized additive models (GAM). Moreover, we introduce a threshold ratio for the selection of individual forecasting methods in the model averaging process. In the “online” part, we obtain the model average forecasts of new series by pre-trained GAM and optimal threshold ratio. We illustrate that our probabilistic forecasting framework outperforms all individual benchmark forecasting methods on M3 data, with improved computational efficiency. Another key advantage of our proposed framework is its interpretability of the effects of features on the probabilistic forecasting accuracy.

用 R 玩转中国生育率分析

王孟樵 (四川大学华西公共卫生学院)

时间: 16:00-16:30

简介: 王孟樵, 四川大学华西公共卫生学院流行病与卫生统计学系助理教授。

摘要: 本研究利用 R 语言及 ggplot2 包对中国的生育率进行回顾性纵向研究, 通过丰富多彩的静态和动态数据可视化呈现人口的不断老龄化和生育率的持续走低。时间序列研究基于霍尔特指数平滑模型恢复预测了删除的基于年龄组别和出生顺序的生育率数据, 从而对中国生育率自 2003 年到 2018 年进行了完整而全面的分析。总的来说, 生育率继续下降到一个相当低的水平, 而以东北三省份为突出代表显示出与低生育陷阱紧密相关的显著社会经济问题。调整生育限制和推动相关生育扶持政策正当其时。(Wang, Heliyon, 2019, <https://doi.org/10.1016/j.heliyon.2019.e01460>)

Text based crude oil price forecasting

白云 (北京航空航天大学)

时间: 16:30-17:00

简介: 白云, 北京航空航天大学本科生, 统计学专业, 已保研至北航管科专业继续攻读研究生, 研究方向: 时间序列分析和预测、文本挖掘、机器学习, 曾获 2018 年 ICIM 优秀论文奖, 参与中远海集团, 工商银行总行等多个大数据项目的算法分析与设计工作。

摘要: Crude oil price forecasting has attracted substantial attention in the field of forecasting. Recently, text based crude oil price forecasting has been an advanced research. Nonetheless, in order to achieve high performance, most of the existing approaches combine text and other factors related to crude oil price to forecast, which is computational and time consuming. Exploiting less factors and information to forecast the crude oil price becomes crucially important. In this paper, we only use news headlines related to the future price to forecast crude oil price. The two marketing index(sentiment index and topic intensity index) extracted from news is used to forecast. Specifically, for short news headlines, we use SeaNMF model to characterize the topic intensity of the market. Compared with other research, we take the time factor into consideration for the sentiment index. Our experiments show crude oil price forecasting based on text, with less factors related to crude oil price and a more accurate characterization of the market, yields better performance compared with others. Also, our text based forecasting method has been applied in other fields and yields good performance, which demonstrates the versatility and robustness of our approach.

金融对话机器人

张家兴 (蚂蚁金服)

时间: 14:00-14:40

简介: 张家兴博士, 蚂蚁金服人工智能部技术总监, 资深算法专家。带领算法团队探索深度学习、自然语言理解、对话机器人、智能客服、舆情分析等人工智能技术在金融领域的应用。张家兴 2006 年获得北京大学博士学位, 毕业后先后就职于百度、微软、阿里巴巴, 曾任微软亚洲研究院研究员。在人工智能、深度学习、分布式系统、算法等多个领域的顶级会议和期刊上 (NIPS、OSDI、CVPR、SIGMOD 等) 发表十多篇论文。

摘要: 传统金融领域存在大量触达用户的人工服务场景, 例如产品客服、贷款催收、产品销售、保险回访和续保等。目前, 对话机器人正在这些场景中帮助和代替人工, 极大的提升效率和降低成本, 给用户带来更好的体验, 推动普惠金融。蚂蚁金服在金融对话机器人中探索着各种人工智能技术, 在智能客服能领域树立了业界标杆, 也在理财、保险、微贷等金融业务中创造算法价值。本次报告会和大家一起探索如何用深度学习和自然语言理解等技术构建金融对话机器人。

对话系统的历史与未来

段清华 (金证优智科技有限公司)

时间: 14:40-15:20

简介: 人工智能领域资深工程师, 主要开发领域是人工智能、自然语言处理, 金融智能。带领团队完成过金融知识图谱构建, 金融非格式化数据挖掘与抽取, 金融问答系统, 量化分析平台, 基于机器学习的通用对话系统等。

摘要: 从早在 1977 年的 Genial Understanter System 的论文发布, 到今天日常的各种机器人, Siri、Echo 等已经进入我们的生活, 这个过程中对话系统经历了哪些, 对话系统未来的发展方向。

NLP+ 金融: 场景及技术趋势

敖翔 (中国科学院计算技术研究所)

时间: 15:50-16:30

简介: 敖翔, 博士, 中国科学院计算技术研究所副研究员, 硕士生导师, 中国科学院青年创新促进会会员。敖翔博士主要从事金融大数据挖掘与机器学习相关研究, 特别是金融相关文本挖掘和金融行为大数据挖掘等, 在 IEEE TKDE, ICDE, WWW, IJCAI, SIGIR 等知名期刊和会议上发表文章 20 余篇, 拥有 4 项授权专利, 先后主持纵、横向项目 10 余项。

摘要: 本次报告将通过介绍近年来自然语言处理服务金融领域的案例, 介绍 NLP+ 金融的部分场景、技术以及未来趋势, 重点将围绕对话系统、文本理解和文本生成在金融中的应用展开, 同时将展望未来 NLP+ 金融的发展趋势。

美人如花隔云端——对话机器人

沈泽希 (联想研究院)

时间: 16:30-17:10

简介: 就职于联想研究院人工智能实验室, 担任 AI Language Analyst 及 Algorithm Researcher; 毕业于 Universidad Complutense de Madrid, 获语言文学研究硕士学位。兼具理科思维的非典型文科生, 输出快乐、创意、趣味等人体必需的多种氨基酸维生素。“每个人都是一本书, 无论厚薄都渴望被人阅读; 快读的人只看封面, 速读的人只浏览目录, 而我却喜欢跳开装帧去细细地品读, 在抒情中发现艺术, 在叙事里追踪哲理, 图书馆就建在我的眼睛深处。”

摘要: 将自然语言应用于对话机器人, 其中存在哪些机遇与挑战

1. 人机交互“欢乐多”?

为什么即便当前对话机器人遍地开花, 而且其中不乏优秀样本, 但人们仍然更喜欢与人类进行交流
(鸡同鸭讲)

2. 人人交互也有“陷阱”?

有了合理的数据与完善的体系, 机器也能逆袭
(逻辑)

3. 人机交互“未来可期”?

多语言, 多渠道, 多模态, 痛点亦是空间
(应用范畴与国人思维)

基于电商平台手机评论数据的文本挖掘

戴明锋 (商务部研究院)

时间: 14:00-14:40

简介: 戴明锋, 2014 年中国人民大学统计学博士毕业, 现为商务部研究院高级统计师, 从事商务数据挖掘、贸易新业态新模式研究, 同时是多家大数据培训机构的兼职培训讲师。

摘要: 随着网购在国内越来越流行, 消费者对网上购物的需求和频率越来越高, 除了关心价格和质量外, 还有很多因素在影响消费者的购物决策, 了解消费者的需求影响因素至关重要, 本研究通过对电商平台上某款手机的评论做文本挖掘分析, 分析某一款手机的用户情感倾向, 从评论中找出该款手机的优点与不足, 提炼不同品牌手机的卖点。

面向中国学生的英语书面语动词形式错误自动检查——基于链语法的研究

陈功 (对外经济贸易大学)

时间: 14:40-15:20

简介: 陈功, 女, 博士, 副教授、硕士研究生导师。对外经济贸易大学英语学院专用英语系副主任。研究方向为语料库语言学、计算语言学。2014 年至 2015 年在英国谢菲尔德大学进行博士后研究。2013 年至 2014 年在教育部国际司工作, 主要从事会谈口译、中外教育合作、中英人文交流等方面的工作。2017 年获国家社科基金青年项目一项、2014 年获教育部人文社科青年项目一项 (现已结项)。参与国家社科基金项目, 北京市社科项目等若干项。在《外语教学与研究》、《外语学刊》等期刊上发表论文近十篇。向教育部提交研究报告四篇 (均得到采纳), 其中两篇研究报告得到部领导肯定性批示。出版专著一部。

摘要: 该研究以型式语法为理论基础, 通过链语法形式化语法体系对动词型式进行了形式化, 并对链语法动词词典进行了重构, 旨在构建一个更好的面向中国学生的英语书面语动词形式错误检查系统。测试结果显示, 重构后链语法词典的查错性能和句法分析能力得到提高。对错句检查的召回率比原词典提高了 4.5%, 准确率提高了 15.7%; 对本族者正确分析句子的准确率提高了 12.2%。研究表明, 该研究所基于的语言学理论 (动词型式语法) 和形式模型 (链语法) 可以较好地适用于中国学生书面英语动词形式错误检查系统的构建。

基于 Wmatrix 语义赋码的商务话语概念隐喻分析

孙亚 (对外经济贸易大学)

时间: 15:50-16:30

简介: 孙亚, 博士, 现任对外经济贸易大学英语学院副院长、教授、博士生导师、校学术委员会委员。于2003年获得复旦大学英语语言文学博士学位, 2010年7月—2011年7月加州大学伯克利分校语言学系访问学者。主要研究方向为语用学、认知语言学, 近年来研究兴趣为商务话语中的隐喻、商务语用学。目前主持国家社科基金项目1项、教育部人文社会科学研究青年项目等省部级项目2项, 已完成北京市社会科学理论著作出版基金资助项目1项。已出版专著3部和发表论文40余篇, 其中CSSCI期刊论文20余篇, SSCI期刊论文或书评5篇, 国外一般期刊论文2篇。主要代表作包括专著《隐喻与话语》(对外经济贸易大学出版社, 2013), 论文 Metaphor use in Chinese and US corporate mission statements(English for Specific Purposes, 2014) 等。

摘要: 本研究对使用语料库方法进行的隐喻研究进行简要回顾, 主要探讨基于网络的语义分析工具Wmatrix在商务话语隐喻研究中的应用, 说明该工具的语义域赋码功能在获取大量文本语料中的隐喻词目和分析隐喻规约性方面的作用。研究表明,Wmatrix的语义域赋码功能使研究者能大限度提取大规模语料中可能的隐喻词目; 隐喻词目的语义域赋码顺序为隐喻的规约性提供了有力证据。

混合效应模型 (Mixed-Effects Models) 在二语研究中的应用

焦鲁 (北京师范大学)

时间: 16:30-17:10

简介: 焦鲁, 北京师范大学心理学部, 博士研究生。在学术论文方面, 本人以第一作者身份发表了3篇SCI/SSCI论文, 并以合作者身份合作发表论文多篇。在课题方面, 本人参与导师负责的课题, 并独立负责一项校级课题, 并顺利结题。在科研经历方面, 本人多次参加国内外的学术会议, 并通过国家公派研究生项目赴美国联合培养1年。

摘要: 二语研究的统计分析始终面临着两个挑战。第研究结论的可推广性。例如, 基于某大学生群体得到的结论能否推广到其他语言背景的大学生群体之中, 或者基于某些词汇材料所得到的结论是否能够推广到该语言的所有词汇中。对此, 传统的方差分析需要分别进行被试分析和项目分析, 但往往得到不一致的结论。第二, 自变量的连续性变化。双语研究必须考虑被试的众多语言背景信息, 例如第二语言的熟练度, 习得年龄等。在传统的方差分析中, 研究者往往根据主观标准, 将这些连续变量划分为类别变量, 必然会丢失一部分数据信息。基于R和Rstudio, 混合效应模型可以很好地解决二语研究中的这两大挑战: 一方面, 混合效应模型可以同时考虑被试(subject)和项目(item)两方面的随机效应, 提高研究结论的可推广性, 另一方面, 混合效应模型可以同时将连续变量和类别变量作为自变量, 避免了因数据类型转换而导致数据信息丢失等问题。此次报告将包括两方面的内容, 一是简单地介绍混合效应模型及其如何利用R语言实现; 二是结合心理学领域中的具体研究, 介绍如何应用混合效应模型解决二语研究中的实际问题。



- 主 办 方**  CAPITAL OF STATISTICS
PROFESSION, HUMANITY & INTEGRITY
- 协 办 方**  中国人大大學
RENMIN UNIVERSITY OF CHINA |  统计学院
SCHOOL OF STATISTICS  中国人大应用统计科学研究中心
Center for Applied Statistics of Renmin University of China  狗熊会
CluBear
- 金牌赞助商**  R Studio®  人民邮电出版社
POSTS & TELECOM PRESS  TURING
图灵教育  中国人大出版社
China Renmin University Press
- 独家视频支持**  IT大咖说
知识共享平台