



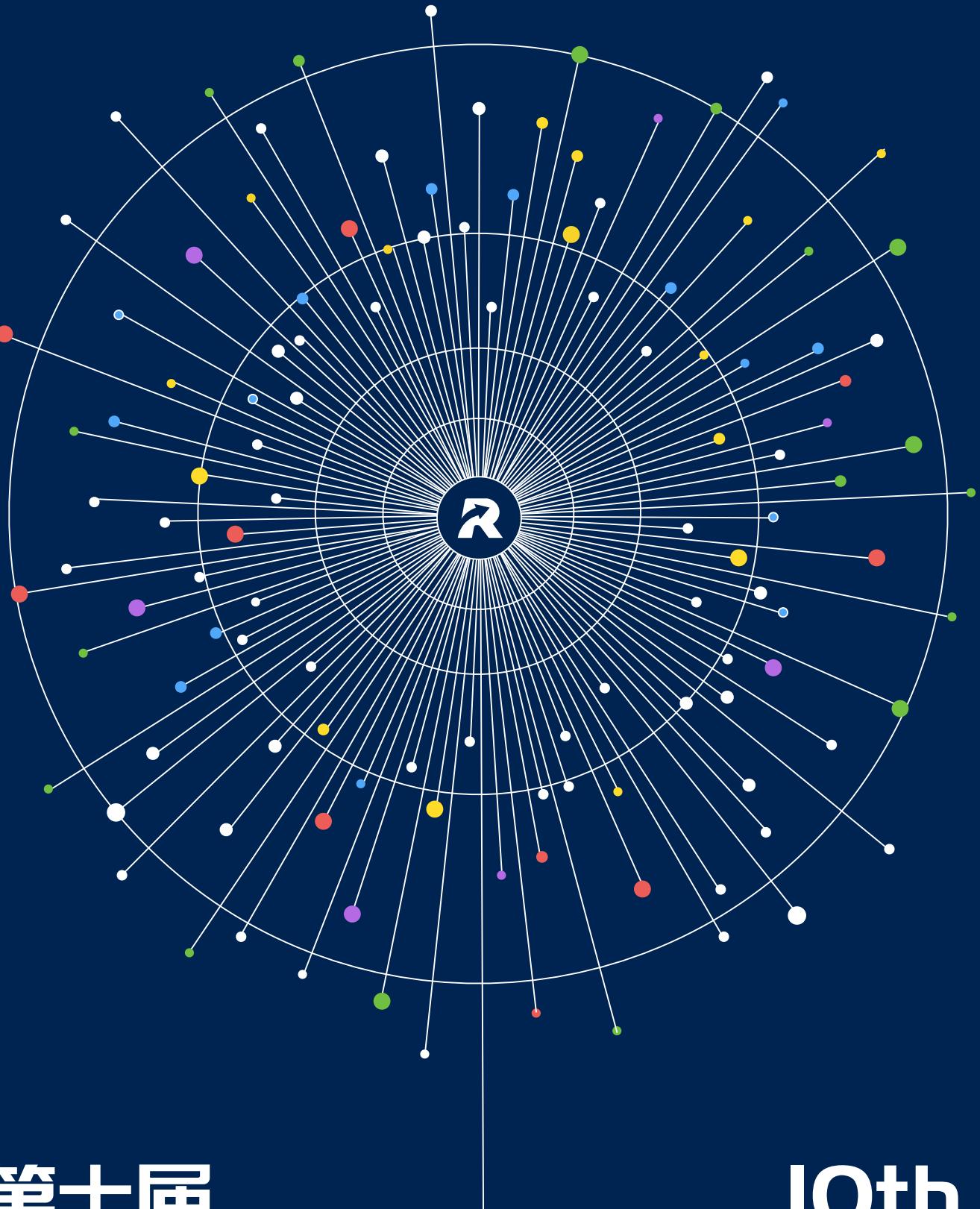
会议时间

5.19-5.21



会议地点

北京
清华大学



第十届
中国R会议 (北京)



10th
The China-R
Conference

合作伙伴



战略合作伙伴

懒投资
LANTOUZI.COM

卓銘保險
CHARMING INSURANCE



金牌赞助

elastic

R Studio®

同盾科技
www.tongdun.cn

KUANG-CHI



银牌赞助

华章科技

天善智能
TIANSHAN SOFT

中国人民大学出版社
China Renmin University Press

TURING
图灵教育



视频支持

IT大咖说
知识分享平台

欢迎辞

十载去，一剑初试锋。中国 R 会议已度过了十年的峥嵘时光，本届在清华大学统计学研究中心、北京大学商务智能研究中心、统计之都和狗熊会的携手努力下，第一次走进美丽的清华园。

十年的时间很长，苏轼已两鬓成霜，木兰也百战回乡；十年的时间很短，罗隐还未成名，苏武仍在牧羊。自其变者而观之，觉宇宙之无穷，如级数般无穷；自其不变者而观之，识盈虚之有数，似统计般有数。十年间，统计学成了显学，而很多新名词突然诞生然后烟消云散，但 R 会议依然屹立在这里。

桃李春风一杯酒，江湖夜雨十年灯。十年间，当初创办 R 会议的很多人经历的是官冗从、落尘笼、簿书丛，感叹的是隙中驹、石中火、梦中身。十年间，R 会议由一年一个城市到一年十多个城市筹办，由一年 120 人参会到上万人参会，由单个主题的小葱拌豆腐到本届覆盖医疗健康、公共卫生、生物信息、消费金融、量化投资、工业工程、智能制造、软件工具、计算平台、概率统计、机器学习、人工智能、自然语言、城市规划、社交网络、政务数据、商务统计、人文社科、心理学等 30 多个数据科学话题的满汉全席。

十年荏苒，R 会议变化了许多，也留下了许多。会议的嘉宾在变，会议的组织者在变；但 R 会议的名字不曾变，心不曾变，它代表着统计之都的“专业、人本、正直”不曾变。一岁一会，当思自强不息；半丝半缕，恒念厚德载物。在 R 会议十周年之际，我们庆幸未忘初心，也期待更新的未来。时光未老，理想仍在，白日放歌共纵酒，豪情作伴赴盛筵！这场 R 会议，敬请诸君共赏！

统计之都敬上
2017 年 5 月 19 日

江南好·十载去

贺中国R会议十周年庆

谢益辉词
项海波曲

Andante $\text{♩} = 69$

The musical score consists of eight staves of music for a single melodic line. The tempo is Andante with a tempo marking of $\text{♩} = 69$. The key signature is common time (indicated by '4'). The lyrics are integrated into the music, appearing below each staff. The lyrics are:

十载去一剑初试锋，筚路蓝缕计相统。
诗酒年华意自承，哦。
夜雨任几更，十载去一剑初试锋。
筚路蓝缕计相统，诗酒。
年华意自承，哦。
夜雨任几更，三代起十会再迎朋。
梅花南山月忆影，白鸟秋水虫鸣灯。
惊雷却无声。
三代起十会再迎朋，梅花南山月忆影，白鸟秋水。
虫鸣灯惊雷却无声。
惊雷却无声。

目录

欢迎辞	1
会议介绍	1
第十届中国 R 会议介绍	1
主办机构	2
赞助商介绍	3
第十届中国 R 会议筹备委员会	8
统计之都简介及活动回顾	9
清华大学地图	10
狗熊会专场地图	10
主会场 & 狗熊会专场日程	11
Keynote(19 日, 新清华学堂, 主席: 邓柯)	22
王永雄: Data Science, machine learning, precision medicine, and all that	22
刘军: Statistical learning with genomic big-data	22
李航: Building Better Connected World with Artificial Intelligence Technologies	22
郭建华: 大数据时代下的统计学思维—以文本挖掘为例	23
宗福季: 统计转移学习 (及其在统计过程控制的应用)	23
邓一硕: 中产阶级如何利用量化投资工具完成财富进阶	24
圆桌讨论	24
R00: 狗熊会专场 (19 日下午, 紫光国际会议中心, 主席: 狗熊会政委)	25
李广雨: 致辞	25
叶征: 物联网大数据分析技术在供应链金融保险和风控领域的应用?	25
苏永刚: 移动程序化广告	25
葛伟平: 数据融合与信用风险评估	25
赵锡刚: 证券分析师的价值分析	26
周扬: 基于车联网数据的商业价值探索	26
兰伟: 浅谈消费金融	27
R01: 资产管理 (懒投资冠名) (20 日上午, 6C101, 主席: 吴海山)	28
吴海山: Quantitative Venture Capital	28
董磊: 手机数据与经济活动测度	28
殷磊: 迁移学习在金融大数据风控中的应用	28
李翛然: 如何制造一次成功的投资	29
自由讨论	29
R02: 城市数据 (20 日上午, 6C102, 主席: 李栋)	30
吴梦荷: 基于区域关联视角的智慧城市发展	30
顾竹: 环境大数据的商业应用	30
张志成: 地理数据与商业网点选址实战	30
黄蔚欣: 基于室内定位数据 (IPS) 的时空行为分析	31
高楠: 不可或缺的优质地理大数据	31
朱雪宁: PM 2.5 数据的时空特征及统计建模	31

R03: 人文数据 (20 日上午, 6C201, 主席: 陈静 & 徐力恒)	33
陈静: 计算与人文: 作为新领域的“数字人文”	33
王成军: network diffusion: Simulate and Visualize Network Diffusion	33
郑文惠: 情感现象学与色彩政治学——唐诗色彩词的数字人文研究	33
王涛: 群像的描绘与类型的分析: 用数字工具挖掘《德意志人物志》	34
邱伟云: 词汇、概念、数字: 文本探勘技术于中国近代观念史研究中的应用与实践	35
自由讨论	35
R04: 数据科学与工业工程应用 (20 日上午, 6C202, 主席: 王凯波 & 朱宇)	36
王凯波: 卓越质量管理中的大数据分析	36
何曙光: 质保数据建模与分析	36
李彦夫: System reliability assessment and optimization	36
皋琴: Branding with social media: User gratifications, usage patterns, and brand message content strategies	37
姜海: 基于车辆 GPS 数据的交通大数据应用	37
自由讨论	38
R05: 生物信息 (20 日上午, 6A416, 主席: 侯琳 & 江瑞)	39
杜朴风: 生物序列分类中的特征快速生成与可视化	39
张淑芹: Hepatocellular carcinoma study based on HBV next generation sequencing	39
王涛: Prediction analysis for microbiome sequencing data	39
吴凌云: 条件随机场及其在生物信息学中的应用	39
杨灿: Adaptive False Discovery Rate regression with application in integrative analysis of large-scale genomic data	40
郭小波: Extending the adjusting-heritable-trait GWAS to bivariate analyse can help identify novel loci .	40
R06: 医学与基因组学 (20 日上午, 6A415, 主席: 韩思蒙 & 李程)	42
Harry Hua: R Usage in Pharmaceutical Industry	42
周健: 临床医生眼中的医疗大数据研究: 需求和挑战	42
吴健民: 消化道肿瘤基因组学研究进展	42
凌少平: Identifying tissue origin of cancer cells with somatic mutations and copy number alterations .	43
唐泽方: 癌症转录组大数据的可视化与再挖掘	43
江瑞: Identification of disease-causing single nucleotide variants in exome sequencing studies	44
R07: 汽车联网 (20 日上午, 6A414, 主席: 李旭)	45
侯志伟: 车联网时空数据挖掘与洞察	45
朱俊辉: 摩拜单车的数据科学实践	45
李晔彤: 互联网汽车数据服务分享	45
王犇: 机器学习在滴滴	45
张翔: 汽车消费的数字化决策	46
赵帅: 基于 R 语言的汽车驾驶行为数据分析	46
R08: 统计理论 A(20 日上午, 6A413, 主席: 杨立坚 & 李东)	47
马莹莹: Banded Spatio-Temporal Autoregressions with Application to Forecasting PM2.5	47
张兴发: On a vector double autoregressive model	47
顾莉洁: Prediction Interval for Autoregressive Time Series via Oracally Efficient Estimation of Multi-Step Ahead Innovation Distribution Function	47
蔡利: Simultaneous confidence bands for mean and variance function based on deterministic design .	48

张园园: A smooth simultaneous confidence band for correlation curve	48
自由讨论	48
R09: 人工智能与量化金融 (20 日下午, 6C101, 主席: 郭健)	50
郭健: 人工智能颠覆量化投资	50
丁磊: 数据驱动人工智能的实践	50
王鑫: 量化投资简介	50
张卓: 论机器学习在金融领域的应用	50
任坤: R 语言与量化投资实战	51
霍志骥: CTA 投资思路与常用 R 包	51
R10: 计算平台 (20 日下午, 6C102, 主席: 颜深根)	52
张先铁: 嵌入式上的深度学习初探	52
肖倾城: Exploring Heterogeneous Algorithms for Accelerating Deep Convolutional Neural Networks on FPGAs	52
杨军: Pluto: A Distributed Heterogeneous Deep Learning Framework	52
卢丽强: Evaluating Fast Algorithms for Convolutional Neural Networks on FPGAs	53
曾勇: Elastic Stack 与机器学习	53
R11: 软件工具 (20 日下午, 6C201, 主席: 肖楠)	54
谢佳标: 利用 css 对 shiny 页面优化及利用 htmlwidgets 包创建 HTML 控件	54
肖楠: Persistent Reproducible Reporting with Docker and R	54
任乾: Learning R Internals and C++ via Rcpp	54
覃文锋: 跟踪 R 社区动态 - R Weekly 的背后	55
自由讨论	55
R12: 工业制造 (20 日下午, 6C202, 主席: 田春华)	56
刘晨: 油气长输管道数据分析实践	56
张玺: 工程数据分析方法在半导体制造过程监测中的应用	56
张光磊: 工业大数据在风电行业的应用	56
王逢春: 电子制造业智能化的挑战与机遇	56
陈宸: 制造即服务, 数据即价值	57
田春华: 工业大数据分析: 实践与挑战	57
R13: Genomic Data Analysis in Bioconductor(20 日下午, 6A416, 主席: Charity Law)	58
Yang Liao: Rsubread: an efficient toolkit for mapping and counting short sequencing reads	58
Yunshun Chen: From reads to genes to pathways: differential expression analysis of RNA-Seq experiments in Bioconductor	58
Charity Law: Glimma: getting greater graphics for your genes	58
Alexandra Garnham: Deconvolving human and viral RNA in RNA sequencing data	59
自由讨论	59
R14: R 软件在社会科学中的应用 (20 日下午, 6A415, 主席: 苏毓淞)	61
吴江: 中文文本分析方便工具包 chinese.misc 介绍	61
陈华珊: ezdf: 用户友好的标签数据框	61
刘京辰: Latent Variable Modeling for Cognitive Assessment Through Second-Order Exponential Family	61
邵兴全: 法律的定量分析及其实践	62
李代: 再抽样法分析夫妻般配与家庭工资不平等	62

自由讨论	62
R15: 大数据人才培养 (20 日下午, 6A414, 主席: 王涛)	63
赵鹏: 微启的旋转门: 大数据教育界与工业界的生态进化	63
欧高炎: 大数据学科建设的关键因素	63
李扬: 数字金融 -实验室项目模拟系统——银行数据仓储, 数据测试, 数据安全三位一体的就业驱动 项目训练系统平台	63
刘乐平: 大数据历史长河中的统计思维与智慧	63
袁星星: 大数据教育平台的建设与探索	64
自由讨论	64
R16: 统计理论 B(20 日下午, 6A413, 主席: 杨立坚 & 李东)	65
王江艳: FACTOR AND RESIDUAL EMPIRICAL PROCESSES	65
王静: Free-knot spline for Generalized Regression Models	65
王冠男: Spatially Varying Coefficient Models	65
王文静: Quantile Regression Oultier Diagnostic: R package ‘quokar’	66
曹明: 哪种奇巧巧克力最好吃: Statistical ranking models 及其 R 实现	66
自由讨论	66
R17: 可视分析 (21 日上午, 6C101, 主席: 袁晓如)	67
萧庆: G2 - 面向统计的可视化语法	67
沈毅: WebGL 在前端数据可视化中的应用	67
陆曼: Interaction+: “让可视化动起来”的既有网页交互	67
黄伟: 运用 WebGL+GIS 开发网络安全应用	68
谷鸿秋: SAS 统计图表: 一键式的图表生成术	68
自由讨论	69
R18: 智能制造 (深圳光启高等理工研究院冠名) (21 日上午, 6C102, 主席: 邓柯)	70
季春霖: 工业大数据的应用	70
沈志勇: 数据智能实践 -从互联网到传统行业	70
陈宏: 大数据时代背景下设备安全管理与智能制造	70
田野: 数控机床大数据分析	71
自由讨论	71
R19: 商务统计 (21 日上午, 6C201, 主席: 黎波)	72
张耀武: 高维数据中的模型诊断及其在商务统计中的应用	72
徐旦: 从统计学生到互金数据科学家之路	72
吴岸城: 机器学习在营销管理中的应用	72
刘应耀: 人工智能颠覆客服行业的实践	72
陈卓: 新能源行业 R 语言数据分析实例	73
自由讨论	73
R20: 消费金融 (同盾科技冠名) (21 日上午, 6C202, 主席: 张云松)	74
叶伟: 人工智能助力线上消费金融的风险管理	74
叶梦舟: 以风险资本收益率驱动决策	74
张云松: 金融科技中的算法与可视化应用案例	74
肖勃飞: 消费金融反欺诈应用探索	74
自由讨论	75

R21: 医疗健康 (卓铭保险冠名) (21 日上午, 6A416, 主席: 李响)	76
李响: 患者表征学习方法与应用	76
俞声: 基于电子病历的高通量表型标记	76
黄正行: 医学临床中的人工智能技术	76
金博: “AI+ 慢性病管理”使精准医疗成为可能	77
自由讨论	77
R22: 文本挖掘 (21 日上午, 6A415, 主席: 张俊妮)	78
王厚峰: 新 AI 时代的智能问答	78
孙薇薇: 自然语言处理中的统计结构学习	78
王彦博: 商业银行“半监督”文本聚类技术应用	78
王菲菲: Bayesian Text Classification and Summarization via A Class-Specified Topic Model	79
张俊妮: 统计模型在关键词提取、文本分类和中文分词问题中的应用	79
自由讨论	80
R23: 深度学习应用 (21 日上午, 6A414, 主席: 陈昱)	81
陈昱: 增强学习打麻将	81
赵申剑: 字符级语言模型与机器翻译	81
薛少飞: 阿里巴巴语音识别声学模型的进化历程	81
张翔: 条件 GAN 用于车型设计和判别	81
郎大为: R 语言中的深度学习: 用 Mxnet 进行车型识别	82
自由讨论	82
R24: 心理科学 (21 日上午, 6A413, 主席: 夏骁凯)	83
朱廷劭: 基于社会媒体大数据的心理学研究	83
吕小康: 基于 R 与 Rstudio 的心理统计教学模式探索	83
胡传鹏: R 语言在加强心理学可重复性中的作用	83
蔡培林: 心理学研究规范化及在 R 语言的实现	84
余嘉元: 心理学在助老机器人研发中的应用	84
自由讨论	85
R25: 语言智能与产业应用 (21 日下午, 6A416, 主席: 刘知远)	86
吕正东: 从语言智能到法务智能	86
郑亚斌: 智能时代的量化资产管理	86
张超: 自然语言处理在医疗智能辅助中的应用	86
赵鑫: 面向社交媒体的商业大数据挖掘	87
吴珂皓: NLP 在金融报告自动化的实践	87
自由讨论	87
R26: 机器学习 (21 日下午, 6A415, 主席: 常象宇)	88
朱军: Triple Generative Adversarial Networks	88
熊熹: 大规模线上实验与机器学习	88
林绍波: Learning theory for deep nets	88
王流斌: 腾讯社交广告实践中智能出价新模式: oCPA	89
陈开江: bandit 算法与推荐系统	89
自由讨论	89
R27: 社交网络 (21 日下午, 6A414, 主席: 周静)	91

靳志辉: Building User Profiles from Online Social Behaviors, with Applications in Tencent Social Ads	91
高瀚: 微信中的社会传播课题与实践	91
周静: 从文本分析看小说中人物的复杂关系: 以琅琊榜为例	91
张忠元: On equivalence of likelihood maximization of stochastic block model and nonnegative matrix factorization, and beyond	92
陈成龙: Kaggle 数据挖掘比赛经验分享	92
R28: 公共卫生 (21 日下午, 6A413, 主席: 蔡俊)	94
张志杰: The relationship between meteorological factors and hand, foot, and mouth disease (HFMD): DLNMs-based time-series analysis	94
张兵: Assessment of the impact of climate on respiratory infectious disease via pomp package in R . .	94
蔡俊: R Epidemics Consortium and Using Its Packages to Analyze Influenza Data	94
李瑞云: 中国 H7N9 禽流感暴发模拟与预测	95
程渠: 基于 R 语言的登革热传播模型建立与参数化	95
贾鹏飞: 基于 R 语言环境下气候因素—登革热媒介蚊虫的动力学模型建立与研究	95

第十届中国 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、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、贵州大学等多个城市的高校举办。2016 年，第九届中国 R 语言会议在北京、上海、广州、杭州、西安、武汉、厦门、成都、贵阳等九个城市分别举办，其中中国人民大学举办的北京会场参会者逾 4000 人。今年将迎来第十届中国 R 会议。

2017 年，是中国 R 会议值得纪念的第 10 个年头。本届 R 会议由清华大学统计学研究中心、北京大学商务智能研究中心、统计之都、狗熊会共同携手筹办，将于 5 月 19-21 日在美丽的清华大学举办。在这样一个值得纪念的时刻，让我们相聚清华大学统计学研究中心，相聚 R 会议十周年庆典，也相聚这场数据与统计的盛宴！本届会议覆盖医疗健康、公共卫生、生物信息、消费金融、量化投资、工业工程、智能制造、软件工具、计算平台、概率统计、机器学习、人工智能、自然语言、城市规划、社交网络、政务数据、商务统计、人文社科、心理学等 30 个数据科学话题，我们欢迎您的到来！

19 日 keynote 主会场在新清华学堂，19 日下午的狗熊会专场在紫光会议中心二层，20 21 日各专题分会场在六教各教室，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

清华大学统计学研究中心

清华大学统计学研究中心依托清华大学在工科、商科、生命科学等方面的有利条件，深入开展统计基础理论、统计计算、生物及医学统计、工业统计和商业统计等领域的科研和教学工作。力争在理论和应用统计方面取得具有国际影响力的重要学术成果。

北京大学商务智能研究中心

北京大学商务智能研究中心依托北京大学光华管理学院，关注基于互联网的大数据研究与应用。中心尤其关注中文文本、网络结构、以及位置数据相关的科研课题。中心为学者提供相关数据资源，为企业提供相关分析方法，为学者和企业合作搭建一个有效的平台。

统计之都

统计之都（Capital of Statistics，简称 COS，网址 <http://cos.name/>）成立于 2006 年 5 月，是一个旨在推广与应用统计学知识的网站和社区。统计之都发源于中国人民大学统计学院，现由世界各地的众多志愿者共同管理维护，旨在搭建一个开放的平台，使得科研人员、企业数据分析人员和统计学爱好者能互相交流合作。统计之都的治站格言是“专业、人本和正直”，力图在此格言指导下通过专业的知识和团队、人本的交流与传播、正直的态度和审视，来更好地推动统计学在中国的发展与传播。

狗熊会

狗熊会致力于成为数据产业的高端智库，使命是：聚数据英才，助产业振兴！在知识普及方面，狗熊会为大众普及统计学和数据科学相关知识。公众号为大家提供包括“菜鸟专栏”、“R 语千寻”，“丑图百讲”等统计学基础知识学习；“熊大胡说”、“政委导读”、“狗熊文摘”等观点性文章。教育产业方面，关注以大学为代表的教育培训机构，提供以精品案例、教材、视频、音频为代表的相关教学知识产权产品，助力教师成长，进一步为合作伙伴数据科学教育的长期发展提供帮助。数据产业方面，狗熊会为合作企业提供联合研究，为数据化转型提供战略咨询，包括梳理业务逻辑，制定相应的数据战略，描绘可落地的实施路径，助力合作伙伴进入数据时代。目前合作过的企业横跨投资、互联网金融、电商、车联网、广告等众多行业。

赞助商介绍

战略合作伙伴

懒投资

懒投资隶属于北京大家玩科技有限公司，2014 年 9 月上线运营，A 轮融 2100 万美元，来自策源创投、源码资本、福布斯富豪夏佐全先生。累计交易金额超 266 亿，为用户赚取 7.8 亿收益，注册用户超百万，无一例逾期。懒投资主要对接应收账款保理、融资租赁和消费金融等优质债权资产。2015 年 12 月，国资参股背景的大型担保机构中盈盛达在香港上市，懒投资作为其基石投资者受邀现场敲钟。这是国内首例互联网金融公司以基石投资者身份亮相国际资本市场。

卓铭保险

卓铭保险（Charming Insurance）隶属润安国际保险经纪有限公司，注册资本 5000 万人民币，联合全球顶级的保险集团 BUPA、美国信诺保险及中国高端保险公司招商信诺、中意保险、永安保险等数 10 家国内专业的保险公司，致力于为当代精英人群提供全球高端保险服务。核心团队由 360、百度、阿里、腾讯等知名互联网公司及平安、PICC 等知名保险机构的资深人士组成，以 InsurTech 为驱动力，高度重视用户体验服务及数据安全。卓铭保险作为国内首个创新型科技保险公司，始终坚持用户至上，根据用户实际需求，与跨国保险公司共同为用户定制针对性保险服务。用户不但可以享有国际险种专有权益，还可以享受保后一站式服务。

金牌赞助

Elastic

Elastic 是一家世界领先的开源软件提供商，致力于结构化和非结构化数据的实时可用性，使用场景涵盖搜索、日志和数据分析等领域。公司成立于 2012 年，旗下拥有开源产品：分布式实时搜索与数据分析引擎：Elasticsearch；可视化展现与分析：Kibana；数据收集与处理中间件：Logstash；轻量级数据收集与网络层流量分析：Beats；除此之外还提供安全、预警、监控、图分析及机器学习的商业插件：X-Pack 以及托管的 Elasticsearch 云服务：Elastic Cloud；这些产品迄今累计已超过一亿次下载。Elastic 由 Benchmark Capital、Index Ventures 及 NEA 投资，总部位于荷兰阿姆斯特丹和美国加州山景城，公司员工及办事处遍布全球各地。

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（R Consortium），每年为 R 社区的开源项目提供大量资助，召集优秀人才解决 R 语言现存的重要且有挑战性的问题。

金牌赞助

同盾科技

同盾科技成立于 2013 年，总部位于浙江杭州，是国内专业的第三方大数据风控服务提供商。自创立以来，同盾始终坚持“跨行业联防联控”的理念，为非银行信贷、银行、保险、基金理财、三方支付、航旅、电商、O2O、游戏、社交平台等多个行业超过 6000 家客户提供基于大数据的风险控制、反欺诈及数据核验服务。作为互联网及金融大数据风控的领导者，同盾希望通过持续创新产品与技术，逐步累积海量正负向数据，不断提升服务可靠性，努力成为值得客户信赖的第三方风控服务提供商。

深圳光启高等理工研究院

光启是一家全球化的创新集团,2010 年由 5 位杜克大学、牛津大学博士归国创立，总部位于深圳。现已发展成为一个全球创新共同体。光启拥有核心自主知识产权和世界级的创新研发团队，掌握了颠覆式隐身技术、颠覆式新型空间技术和颠覆式无线互联技术。累计申请专利超过 4100 件，其中超过 2300 件已获授权。通过整合全球创新资源，设计未来、实现未来、分享未来进行跨代创新，推动时代变革。光启涉及航空航天工业、新型空间服务、智能装备、智慧城市、新型无线通信等产业。

银牌赞助

华章科技

北京华章图文信息有限公司（机械工业出版社华章公司）成立于 1995 年，是国内第一家中外合资的出版公司。20 年来出版了《算法导论》、《编译原理》、《Java 编程思想》、《深入理解计算机系统》、《代码之美》、《点石成金：访客至上的网页设计秘笈》、《PHP 和 MySQL Web 开发》、《深入理解 Java 虚拟机》等知名畅销书，累计出版图书 3000 多种，图书销售册数超过 2000 万册。

天善智能

天善智能 hellobi.com 致力于构建一个基于大数据领域的生态圈，链接一切与数据相关的资源，共同努力推动大数据、数据分析、商业智能 BI、数据挖掘、人工智能等领域在国内的普及和发展。社区包括技术问答、博客、活动、学院、招聘、读书频道等子版块，内容覆盖了与大数据、数据分析、数据挖掘和商业智能 BI、数据分析、数据挖掘和大数据相关的技术领域。2017 年初，社区组织编著的《数据实践之美》印刷出版，累计销量 4000+ 册。

图灵教育

北京图灵文化发展有限公司，始终以策划高质量的科技图书为核心业务，成立 11 年以来，累计销售图书已达 1000 多万册，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，是中华人民共和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。人大出版社依托中国人民大学的综合优势，始终高扬人文社会科学的旗帜，秉承“出教材学术精品，育人文社科英才”的出版理念，实施精品战略，以优秀的出版物传播先进文化，目前年出书 3000 余种，发行码洋近 10 亿。

会议视频服务独家合作伙伴

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。50+ 合作社区，每周 10+ 场技术大会精彩分享，2000+ 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！IT 大咖说，让智慧流动起来！

第十届中国 R 会议筹备委员会

主席：姜瑛恺

副主席：汪子栋

秘书长：于嘉傲

秘书团：邓金涛、冯璟烁、蒋斐宇、雷博文、李艺超、李宇轩、林毓聪、王健桥、王小宁、王毅然、徐崇元、徐嘉泽、闫施、杨舒仪、杨洋、张心雨、朱万闯

志愿者：毕嘉辉、蔡利、蔡振帝、常皓、常勤缘、车明佳、陈坤博、陈玲玲、储奕宇、戴庭萱、单娜阳、丁宗巨、董安澜、董航、杜国栋、杜泉莹、甘释宇、高彬、耿瑞轩、龚欢、郭瀚民、郭文魁、韩辰、胡贵平、胡涛、黄维佳、黄伟清、姬生翔、贾册、李浩然、李杰、李祺、李婷婷、李雨霏、梁骞、刘朝阳、刘璇、刘燕、卢剑秋、卢少强、卢心笛、吕学远、马宁、苗定豪、倪丹、欧阳志成、秦宇婷、邱雅娟、商丰瑞、汪洁、王尔实、王琦轩、王秋皓、王通、王欣薇、王雨洲、王昭、吴鹏、熊竹清、杨坤、姚启坤、于金萍、余丽珊、袁正、苑斌杰、翟頣、张翰宇、张家齐、张谦、张若帆、张思韫、张文轩、张晓宇、张怿良、张玉晨、张园园、张卓然、张紫嫣、赵永芳、郑浩天、周美孜、周璇、周艺、朱珂、朱淑怡、朱叶、朱之恺

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

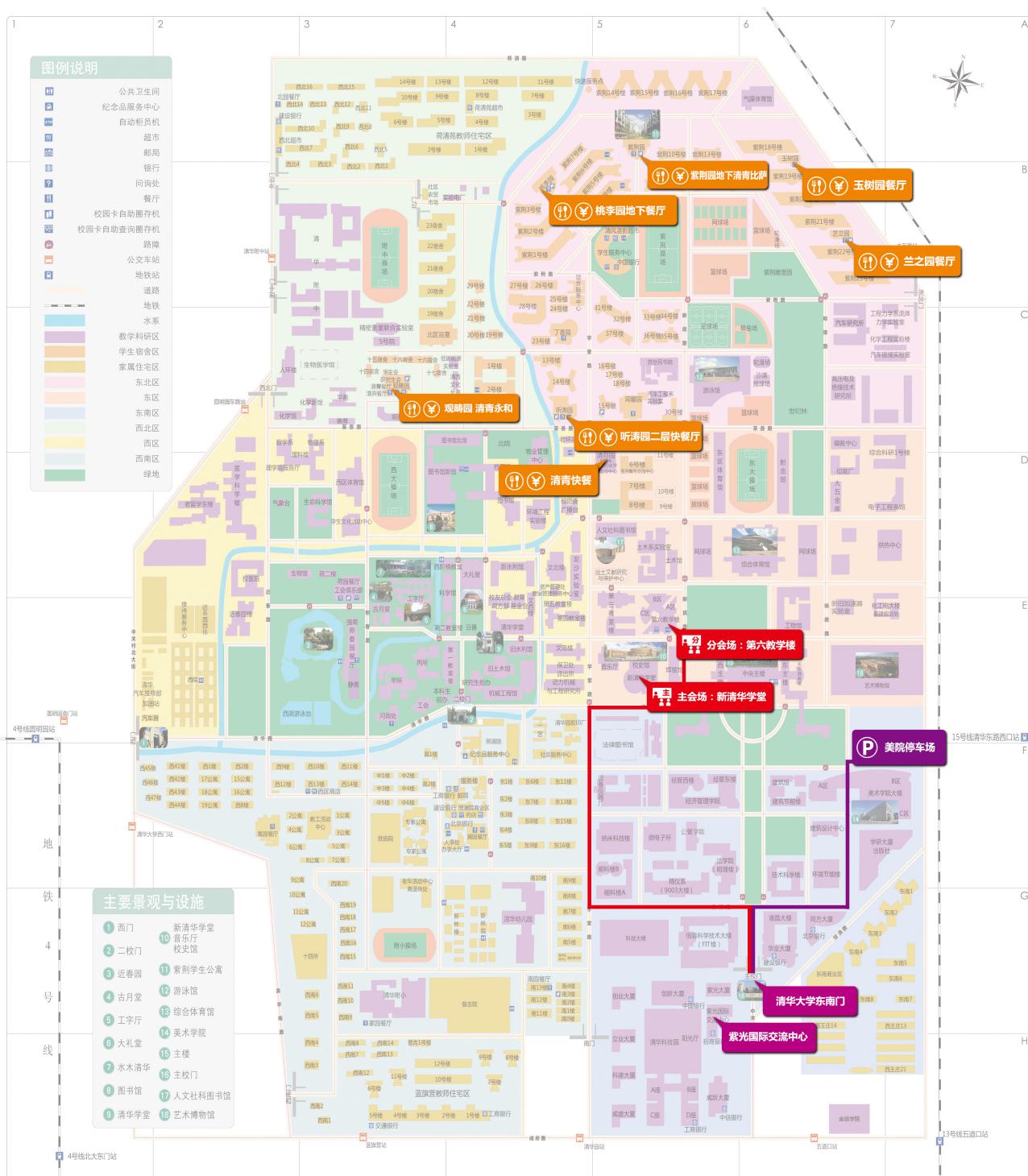
我们怀着“十年磨一剑”的决心，要将“统计之都”创建成中国的统计学“正直、人本、专业”的社区；我们抱着“己欲立而立人、己欲达而达人”的信条，要将“统计之都”以免费统计学服务平台的形式坚持办下去。我们希望“统计之都”在专业知识体系上有真正的王者风范，在面对用户需求时却又以谦恭的态度为大家服务。

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://cos.name/>)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：目前已开展到第九届，分别在北京、上海、广州、杭州、西安、武汉、成都、贵阳、南昌、厦门等地举办。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 37 期，目前主要在北京，每月举办，详情参见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 18 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、王群峰翻译，《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、郑志勇等翻译等等。

清华大学地图



注：该图表示前往会场、停车场的行车路线图，图上标记的餐厅都可以现金消费。

狗熊会专场地图



注：图上标记的位置是狗熊会专场的会场位置。

主会场 & 狗熊会专场日程

5月19日(周五)主会场(新清华学堂)

主会场	演讲嘉宾	主题	时间
主席： 邓柯		参会者入场	8:00~9:00
		致辞	9:00~9:15
	王永雄	Data Science, machine learning, precision medicine, and all that	9:15~10:00
	刘军	Statistical learning with genomic big-data	10:00~10:45
		自由讨论、休息	10:45~11:15
	李航	Building Better Connected World with Artificial Intelligence Technologies	11:15~12:00
		午餐	12:00~14:00
	郭建华	大数据时代下的统计学思维—以文本挖掘为例	14:00~14:45
	宗福季	统计转移学习(及其在统计过程控制的应用)	14:45~15:30
		自由讨论、休息	15:30~16:00
	邓一硕	中产阶级如何利用量化投资工具完成财富进阶	16:00~16:45
	圆桌讨论		16:45~17:30

5月19日(周五)下午狗熊会专场

(紫光国际会议中心二层宴会厅)

专场	演讲嘉宾	主题	时间
主席： 狗熊会政委	李广雨	致辞	13:55~14:00
	叶征	物联网大数据分析技术在供应链金融保险和风控领域的应用	14:00~14:30
	苏永刚	移动程序化广告	14:30~15:00
	葛伟平	数据融合与信用风险评估	15:00~15:30
	赵锡刚	证券分析师的价值	15:30~16:00
	周扬	基于车联网数据的商业价值探索	16:00~16:30
	兰伟	浅谈消费金融	16:30~17:00

第十届中国R会议北京会场日程

分会场	20日上午	20日下午	21日上午	21日下午
6C101	R01 资产管理 (懒投资冠名) 主席：吴海山	R09 人工智能与量化金融 主席：郭健	R17 可视分析 主席：袁晓如	
6C102	R02 城市数据 主席：李栋	R10 计算平台 主席：颜深根	R18 智能制造 (深圳光启高等理工研究 院冠名) 主席：邓柯	无
6C201	R03 人文数据 主席：陈静 & 徐力恒	R11 软件工具 主席：肖楠	R19 商务统计 主席：黎波	
6C202	R04 数据科学与工业工程应 用 主席：王凯波 & 朱宇	R12 工业制造 主席：田春华	R20 消费金融 (同盾科技冠名) 主席：张云松	
6A416	R05 生物信息 主席：侯琳 & 江瑞	R13 Genomic Data Analysis in Bioconductor 主席：Charity Law	R21 医疗健康 (卓铭保险冠名) 主席：李响	R25 语言智能与 产业应用 主席：刘知远
6A415	R06 医学与基因组学 主席：韩思蒙 & 李程	R14 R 软件在社会科学中的 应用 主席：苏毓淞	R22 文本挖掘 主席：张俊妮	R26 机器学习 主席：常象宇
6A414	R07 汽车联网 主席：李旭	R15 大数据人才培养 主席：王涛	R23 深度学习应用 主席：陈昱	R27 社交网络 主席：周静
6A413	R08 统计理论 A 主席：杨立坚 & 李东	R16 统计理论 B 主席：杨立坚 & 李东	R24 心理科学 主席：夏晓凯	R28 公共卫生 主席：蔡俊
6A411	自由讨论	自由讨论	自由讨论	自由讨论

注：分会场全部分布在清华大学第六教学楼各教室，如6C101表示六教C座101教室。

5月20日(周六)上午分会场

分会场	演讲嘉宾	主题	时间
资产管理 (懒投资冠名) 6C101 主席 : 吴海山	吴海山	Quantitative Venture Capital	8:30~9:00
	董磊	手机数据与经济活动测度	9:00~9:30
	殷磊	迁移学习在金融大数据风控中的应用	9:30~10:00
		自由讨论、休息	10:00~10:30
	李翛然	如何制造一次成功的投资	10:30~11:00
		自由讨论	11:00~11:30
城市数据 6C102 主席 : 李栋	吴梦荷	基于区域关联视角的智慧城市发展	8:30~9:00
	顾竹	环境大数据的商业应用	9:00~9:30
	张志成	地理数据与商业网点选址实战	9:30~10:00
		自由讨论、休息	10:00~10:30
	黄蔚欣	基于室内定位数据 (IPS) 的时空行为分析	10:30~11:00
	高楠	不可或缺的优质地理大数据	11:00~11:30
人文数据 6C201 主席 : 陈静&徐力恒	朱雪宁	PM 2.5 数据的时空特征及统计建模	11:30~12:00
	陈静	计算与人文 : 作为新领域的“数字人文”	8:30~9:00
	王成军	Network Diffusion: Simulate and Visualize Network Diffusion	9:00~9:30
	郑文惠	情感现象学与色彩政治学—唐诗色彩词的数字人文研究	9:30~10:00
		自由讨论、休息	10:00~10:30
	王涛	群像的描绘与类型的分析 : 用数字工具挖掘《德意志人物志》	10:30~11:00
	邱伟云	词汇、概念、数字 : 文本探勘技术于中国近代观念史研究中的应用与实践	11:00~11:30
数据科学与工业工程 应用 6C202 主席 : 王凯波 & 朱宇		自由讨论	11:30~12:00
	王凯波	卓越质量管理中的大数据分析	8:30~9:00
	何曙光	质保数据建模与分析	9:00~9:30
	李彦夫	System reliability assessment and optimization	9:30~10:00
		自由讨论、休息	10:00~10:30
	皋琴	Branding with social media: User gratifications, usage patterns, and brand message content strategies	10:30~11:00
	姜海	基于车辆 GPS 数据的交通大数据应用	11:00~11:30
		自由讨论	11:30~12:00

分会场	演讲嘉宾	主题	时间
生物信息 6A416 主席 : 侯琳 & 江瑞	杜朴风	生物序列分类中的特征快速生成与可视化	8:30~9:00
	张淑芹	Hepatocellular carcinoma study based on HBV next generation sequencing	9:00~9:30
	王涛	Prediction analysis for microbiome sequencing data	9:30~10:00
	自由讨论、休息		10:00~10:30
	吴凌云	条件随机场及其在生物信息学中的应用	10:30~11:00
	杨灿	Adaptive False Discovery Rate regression with application in integrative analysis of large-scale genomic data	11:00~11:30
	郭小波	Extending the adjusting-heritable-trait GWAS to bivariate analyse can help identify novel loci	11:30~12:00
	Harry Hua	R Usage in Pharmaceutical Industry	8:30~9:00
医学与基因组学 6A415 主席 : 韩思蒙 & 李程	周健	临床医生眼中的医疗大数据研究 : 需求和挑战	9:00~9:30
	吴健民	消化道肿瘤基因组学研究进展	9:30~10:00
	自由讨论、休息		10:00~10:30
	凌少平	Identifying tissue origin of cancer cells with somatic mutations and copy number alterations	10:30~11:00
	唐泽方	癌症转录组大数据的可视化与再挖掘	11:00~11:30
	江瑞	Identification of disease-causing single nucleotide variants in exome sequencing studies	11:30~12:00
	侯志伟	车联网时空数据挖掘与洞察	8:30~9:00
	朱俊辉	摩拜单车的数据科学实践	9:00~9:30
汽车联网 6A414 主席 : 李旭	李晔彤	互联网汽车数据服务分享	9:30~10:00
	自由讨论、休息		10:00~10:30
	王犇	机器学习在滴滴	10:30~11:00
	张翔	汽车消费的数字化决策	11:00~11:30
	赵帅	基于 R 语言的汽车驾驶行为数据分	11:30~12:00

分会场	演讲嘉宾	主题	时间
统计理论 A 6A413 主席：杨立坚 & 李东	马莹莹	Banded Spatio-Temporal Autoregressions	8:30~9:00
	张兴发	On a vector double autoregressive model	9:00~9:30
	顾莉洁	Prediction Interval for Autoregressive Time Series via Oracally Efficient Estimation of Multi-Step Ahead Innovation Distribution Function	9:30~10:00
		自由讨论、休息	10:00~10:30
	蔡利	Simultaneous confidence bands for mean and variance function based on deterministic design	10:30~11:00
	张园园	A smooth simultaneous confidence band for correlation curve	11:00~11:30
		自由讨论	11:30~12:00

5月20日(周六)下午分会场

分会场	演讲嘉宾	主题	时间
人工智能与量化金融 6C101 主席：郭健	郭健	人工智能颠覆量化投资	14:00~14:30
	丁磊	数据驱动人工智能的实践	14:30~15:00
	王鑫	量化投资简介	15:00~15:30
		自由讨论、休息	15:30~16:00
	张卓	论机器学习在金融领域的应用	16:00~16:30
	任坤	R 语言与量化投资实战	16:30~17:00
	霍志骥	CTA 投资思路与常用 R 包	17:00~17:30
计算平台 6C102 主席：颜深根	张先轶	嵌入式上的深度学习初探	14:00~14:30
	肖倾城	Exploring Heterogeneous Algorithms for Accelerating Deep Convolutional Neural Networks on FPGAs	14:30~15:00
	杨军	Pluto: A Distributed Heterogeneous Deep Learning Framework	15:00~15:30
		自由讨论、休息	15:30~16:00
	卢丽强	Evaluating Fast Algorithms for Convolutional Neural Networks on FPGAs	16:00~16:30
	曹勇	Elastic Stack 与机器学习	16:30~17:00
软件工具 6C201 主席：肖楠	谢佳标	利用 css 对 shiny 页面优化及利用 htmlwidgets 包创建 HTML 控件	14:00~14:30
	肖楠	Persistent Reproducible Reporting with Docker and R	14:30~15:00
	任乾	Learning R Internals and C++ via Rcpp	15:00~15:30
		自由讨论、休息	15:30~16:00
	覃文锋	跟踪 R 社区动态 - R Weekly 的背后	16:00~16:30
	自由讨论	自由讨论	16:30~17:00
工业制造 6C202 主席：田春华	刘晨	油气长输管道数据分析	14:00~14:30
	张玺	工程数据分析方法在半导体制造过程监测中的应用	14:30~15:00
	张光磊	工业大数据在风电行业中的应用	15:00~15:30
		自由讨论、休息	15:30~16:00
	王逢春	电子制造业智能化的挑战与机遇	16:00~16:30
	陈宸	制造即服务，数据即价值	16:30~17:00
	田春华	工业大数据分析：实践与挑战	17:00~17:30

分会场	演讲嘉宾	主题	时间
Genomic Data Analysis in Bioconductor 6A416 主席: Charity Law	Yang Liao	Rsubread: an efficient toolkit for mapping and counting short sequencing reads	14:00~14:30
	Yunshun Chen	From reads to genes to pathways: differential expression analysis of RNA-Seq experiments in Bioconductor	14:30~15:00
	Charity Law	Glimma: getting greater graphics for your genes	15:00~15:30
		自由讨论、休息	15:30~16:00
	Alexandra Garnham	Deconvolving human and viral RNA in RNA sequencing data	16:00~16:30
		自由讨论	16:30~17:00
R 软件在社会科学中的应用 6A415 主席 : 苏毓淞	吴江	中文文本分析方便工具包 chinese.misc 介绍	14:00~14:30
	陈华珊	ezdf: 用户友好的标签数据框	14:30~15:00
	刘京辰	Latent Variable Modeling for Cognitive Assessment Through Second-Order Exponential Family	15:00~15:30
		自由讨论、休息	15:30~16:00
	邵兴全	法律的定量分析及其实践	16:00~16:30
	李代	再抽样法分析夫妻般配与家庭工资不平等	16:30~17:00
		自由讨论	17:00~17:30
大数据人才培养 6A414 主席 : 王涛	赵鹏	微启的旋转门：大数据教育界与工业界的生态进化	14:00~14:30
	欧高炎	大数据学科建设的关键因素	14:30~15:00
	李扬	数字金融 - 实验室项目模拟系统——银行数据仓储，数据测试，数据安全三位一体的就业驱动项目训练系统平台	15:00~15:30
		自由讨论、休息	15:30~16:00
	刘乐平	大数据历史长河中的统计思维与智慧	16:00~16:30
	袁星星	大数据教育实训平台的建设与探索	16:30~17:00
		自由讨论	17:00~17:30
统计理论 B 6A413 主席 : 杨立坚&李东	王江艳	FACTOR AND RESIDUAL EMPIRICAL PROCESSES	14:00~14:30
	王静	Free-knot spline for Generalized Regression Models	14:30~15:00
	王冠男	Spatially Varying Coefficient Models	15:00~15:30
		自由讨论、休息	15:30~16:00
	王文静	Quantile Regression Oultier Diagnostic: R package ‘quokar’	16:00~16:30
	曹明	哪种奇巧巧克力最好吃 : Statistical ranking models 及其 R 实现	16:30~17:00
		自由讨论	17:00~17:30

5月21日(周日)上午分会场

分会场	演讲嘉宾	主题	时间
可视分析 6C101 主席：袁晓如	萧庆	G2 - 面向统计的可视化语法	8:30~9:00
	沈毅	WebGL 在前端数据可视化中的应用	9:00~9:30
	陆曼	Interaction+: “让可视化动起来”的既有网页交互	9:30~10:00
		自由讨论、休息	10:00~10:30
	黄伟	运用 WebGL+GIS 开发网络安全应用	10:30~11:00
	谷鸿秋	SAS 统计图表：一键式的图表生成术	11:00~11:30
		自由讨论	11:30~12:00
智能制造 (深圳光启高等理工研究院冠名) 6C102 主席：邓柯	季春霖	工业大数据的应用	8:30~9:00
	沈志勇	数据智能实践 – 从互联网到传统行业	9:00~9:30
	陈宏	大数据时代背景下设备安全管理与智能制造	9:30~10:00
		自由讨论、休息	10:00~10:30
	田野	数控机床大数据分析	10:30~11:00
		自由讨论	11:00~11:30
商务统计 6C201 主席：黎波	张耀武	高维数据中的模型诊断及其在商务统计中的应用	8:30~9:00
	徐旦	从统计学生到互金数据科学家之路	9:00~9:30
	吴岸城	机器学习在营销管理中的应用	9:30~10:00
		自由讨论、休息	10:00~10:30
	刘应耀	人工智能颠覆客服行业的实践	10:30~11:00
	陈卓	新能源行业 R 语言数据分析实例	11:00~11:30
		自由讨论	11:30~12:00
消费金融 (同盾科技冠名) 6C202 主席：张云松	叶伟	人工智能助力线上消费金融的风险管理	8:30~9:00
	叶梦舟	以风险资本收益率驱动决策	9:00~9:30
	张云松	金融科技中的算法与可视化应用案例	9:30~10:00
		自由讨论、休息	10:00~10:30
	肖勃飞	消费金融中反欺诈的应用研究	10:30~11:00
		自由讨论	11:00~11:30

分会场	演讲嘉宾	主题	时间
医疗健康 (卓铭保险冠名) 6A416 主席 : 李响	李响	患者表征学习方法与应用	8:30~9:00
	俞声	基于电子病历的高通量表型标记	9:00~9:30
	黄正行	医学临床中的人工智能技术	9:30~10:00
		自由讨论、休息	10:00~10:30
	金博	“AI+ 慢性病管理”使精准医疗成为可能	10:30~11:00
		自由讨论	11:00~11:30
文本挖掘 6A415 主席 : 张俊妮	王厚峰	新 AI 时代的智能问答	8:30~9:00
	孙薇薇	自然语言处理中的统计结构学习	9:00~9:30
	王彦博	商业银行“半监督”文本聚类技术应用	9:30~10:00
		自由讨论、休息	10:00~10:30
	王菲菲	Bayesian Text Classification and Summarization via A Class-Specified Topic Model	10:30~11:00
	张俊妮	统计模型在关键词提取、文本分类和中文分词问题中的应用	11:00~11:30
		自由讨论	11:30~12:00
深度学习应用 6A414 主席 : 陈昱	陈昱	增强学习打麻将	8:30~9:00
	赵申剑	字符级语言模型与机器翻译	9:00~9:30
	薛少飞	阿里巴巴语音识别声学模型的进化历程	9:30~10:00
		自由讨论、休息	10:00~10:30
	张翔	条件 GAN 用于车型设计和判别	10:30~11:00
	郎大为	R 语言中的深度学习: 用 Mxnet 进行车型识别	11:00~11:30
		自由讨论	11:30~12:00
心理科学 6A413 主席 : 夏骁凯	朱廷劭	基于社交媒体大数据的心理学研究	8:30~9:00
	吕小康	基于 R 与 Rstudio 的心理统计教学模式探索	9:00~9:30
	胡传鹏	R 语言在加强心理学可重复性中的作用	9:30~10:00
		自由讨论、休息	10:00~10:30
	蔡培林	心理学研究规范化及在 R 语言的实现	10:30~11:00
	余嘉元	心理学在助老机器人研发中的应用	11:00~11:30
		自由讨论	11:30~12:00

5月21日(周日)下午分会场

分会场	演讲嘉宾	主题	时间
语言智能与产业 应用 6A416 主席 : 刘知远	吕正东	从语言智能到法务智能	14:00~14:30
	郑亚斌	智能时代的量化资产管理	14:30~15:00
	张超	自然语言处理在医疗智能辅助中的应用	15:00~15:30
		自由讨论、休息	15:30~16:00
	赵鑫	面向社交媒体的商业大数据挖掘	16:00~16:30
	吴珂皓	NLP 在金融报告自动化的实践	16:30~17:00
		自由讨论	17:00~17:30
机器学习 6A415 主席 : 常象宇	朱军	Triple Generative Adversarial Networks	14:00~14:30
	熊熹	大规模线上实验与机器学习	14:30~15:00
	林绍波	Learning theory for deep nets	15:00~15:30
		自由讨论、休息	15:30~16:00
	王流斌	腾讯社交广告实践中智能出价新模式 : oCPA	16:00~16:30
	陈开江	bandit 算法与推荐系统	16:30~17:00
		自由讨论	17:00~17:30
社交网络 6A414 主席 : 周静	靳志辉	Building User Profiles from Online Social Behaviors, with Applications in Tencent Social Ads	14:00~14:30
	高瀚	微信中的社会传播课题与实践	14:30~15:00
	周静	从文本分析看小说中人物的复杂关系 :以琅琊榜为例	15:00~15:30
		自由讨论、休息	15:30~16:00
	张忠元	On equivalence of likelihood maximization of stochastic block model and nonnegative matrix factorization, and beyond	16:00~16:30
	陈成龙	Kaggle 数据挖掘比赛经验分享	16:30~17:00
		自由讨论	17:00~17:30
公共卫生 6A413 主席 : 蔡俊	张志杰	The relationship between meteorological factors and hand, foot, and mouth disease (HFMD): DLNMs-based time-series analysis	14:00~14:30
	张兵	Assessment of the impact of climate on respiratory infectious disease via pomp package in R	14:30~15:00
	蔡俊	R Epidemics Consortium and Using Its Packages to Analyze Influenza Data	15:00~15:30
		自由讨论、休息	15:30~16:00
	李瑞云	中国 H7N9 禽流感暴发模拟与预测	16:00~16:30
	程渠	基于 R 语言的登革热传播模型建立与参数化	16:30~17:00
	贾鹏飞	基于 R 语言环境下气候因素 - 登革热媒介蚊虫的动力学模型建立与研究	17:00~17:30

Data Science, machine learning, precision medicine, and all that

王永雄 (*Stanford University*)

时间: 9:15~10:00 邮箱: whwong@stanford.edu

简介: Professor Wong is a fellow of National Academy of Sciences in the United States and Academia Sinica (2010). He won the highest award in the field of Statistics COPSS Presidents' Award in 1993. Wong graduated from the University of California, Berkeley in 1976 with a Bachelor's degree. At the University of Wisconsin-Madison, he studied under renowned statistician Grace Wahba, and was awarded a PhD in Statistics in 1980. After graduation, he taught at the University of Chicago, served as an assistant professor, associate professor, and professor. In 1994 he joined the Chinese University of Hong Kong Department of Statistics. Since 1997, he taught and led his lab at the University of California, Los Angeles and Harvard University. In 2004, he was appointed Professor at Stanford University, and served as Head of the Department of Statistics at Stanford University in 2009.

摘要: Although many aspects in the healthcare industry have been impacted by information technology, the practice of medicine has not been disrupted a fundamental level. This is about to change because of the convergence of breakthrough advances in genomics, clinical informatics, and statistical learning methods. In this talk I will review recent developments in this direction. In particular, I will discuss the importance of large scale genomics and health record data, and the value of integrative analysis/modeling of heterogeneous data.

Statistical learning with genomic big-data

刘军 (清华大学; 哈佛大学)

时间: 10:00~10:45 邮箱: jliu@stat.harvard.edu

简介: Professor Liu is the director of Center for Statistical Science of Tsinghua University. He is also a professor in the Department of Statistics at Harvard University. Liu was an IMS Medallion Lecturer in 2002 and a Bernoulli Lecturer in 2004. He was elected a fellow of the Institute of Mathematical Statistics in 2004[3] and of the American Statistical Association in 2005. Liu received his B.Sc. from Peking University in 1985. He has a Ph.D. in math from Rutgers University in 1988, and a Ph.D. in statistics under the supervision of Wing Hung Wong from the University of Chicago in 1991.

摘要: The number of publicly available gene expression and genome sequence datasets has been growing dramatically. Various methods have been proposed to predict gene functions by integrating the publicly available datasets. I will use a few recent projects we carried out to illustrate the roles and importance of statistical modeling for extracting knowledge (i.e., learning) from genomic and genetic big-data and to predict treatment effects from genomic information. The take-home lessons are (a) statistical models are all “wrong” in certain technical aspects, but are extremely useful for synthesizing information (much more so than “techniques” -driven approaches); (b) statistical thinking is important for understanding predictions and observational findings; (c) Bayesian-like data integration and model averaging can provide more coherent and accurate answers to intricate problems. As examples, we show that our algorithm CLIC is capable of integrating over thousands of gene expression datasets to achieve much higher co-expression prediction accuracy compared to traditional co-expression methods. We also show that statistical learning model-based personalized chemotherapy selection results in significant higher survival rates compared with standard practice for breast cancer patients.

Building Better Connected World with Artificial Intelligence Technologies

李航 (华为技术有限公司)

时间: 11:15~12:00 邮箱: *HangLi.HL@huawei.com*

简介: 李航博士的研究方向包括信息检索, 自然语言处理, 统计机器学习, 及数据挖掘。李航 1988 年日本京都大学电气工程系毕业, 1998 年获得日本东京大学计算机科学博士。他 1990 年至 2001 年就职于日本 NEC 公司中央研究所, 任研究员, 2001 年至 2012 年就职于微软亚洲研究院, 任高级研究员与主任研究员。李航一直活跃在相关学术领域, 曾出版过三部学术专著, 并在顶级国际学术会议和顶级国际学术期刊上发表过 120 多篇学术论文, 包括 SIGIR, WWW, WSDM, ACL, EMNLP, ICML, NIPS, SIGKDD, AAAI, IJCAI, 以及 CL, NLE, JMLR, TOIS, IRJ, IPM, TKDE, TWEB, TIST。他和同事的论文获得了 KDD2008 最佳应用论文奖, 他指导的学生获得了 SIGIR2008, ACL2012 最佳学生论文奖。李航参与了多项产品开发, 包括 Microsoft SQL Server 2005, Microsoft Office 2007, Microsoft Live Search 2008, Microsoft Bing 2009, Bing 2010, Office 2010, Office 2012。他拥有 42 项授权美国专利。李航还在顶级国际学术会议和顶级国际学术期刊担任许多重要工作, 如大会程序委员会主席, 资深委员, 及委员, 期刊编委, 包括 SIGIR, WWW, WSDM, ACL, NAACL, EMNLP, NIPS, SIGKDD, ICDM, ACML, IJCAI, 以及 CL, IRJ, TIST, JASIST, JCST。

摘要: We envision that with artificial intelligence technologies the telecommunication, enterprise, and consumer industries, in which Huawei has its main business, will enter a completely new horizon; specifically, all the products and services will be revolutionized to become more intelligent. Huawei is indeed pushing the frontier of research and development of technologies in those fields and has accomplished significant achievements. In this talk, I will introduce some of the best practices as well as the technology breakthroughs made in Huawei, with regard to building better telecommunication networks, better enterprise management, and better mobile devices. Specifically, I will describe the accomplishments made at research projects of Huawei Noah's Ark Lab. Finally, I will summarize the challenges and important research directions in artificial intelligence, for which more research, particularly fundamental research, is needed.

大数据时代下的统计学思维—以文本挖掘为例

郭建华 (东北师范大学)

时间: 14:00~14:45 邮箱: *jhguo@nenu.edu.cn*

简介: 郭建华, 东北师范大学教授, 博士生导师, 副校长。国务院学位委员会学科评议组统计学科召集人, 国家杰出青年科学基金获得者, 教育部“长江学者奖励计划”特聘教授, “新世纪百千万人才工程”国家级人选, 国务院政府特殊津贴获得者, 国家自然科学基金数学天元基金学术领导小组成员。

摘要: 从现实的世界出发去探知真实的世界, 是几乎一切科学的目的。为此, 人们搭建了一个想象的世界, 称之为模型。自然的, 模型既应与现实世界相吻合, 又应与我们心目中的真实世界相匹配。随着大数据时代的来临, 我们心中的“现实世界”变得越来越庞大, 模型的建立就变得越来越有挑战性。作为大数据建模的最重要工具之一, 统计学模型又是如何建立的呢? 本报告将以此为题, 逐步讨论统计学的思维方式, 提出了一种可称之为“结构降维”的建模思路, 并应用在文本挖掘领域。本报告将概述我们的基本思想和实际做法。

统计转移学习（及其在统计过程控制的应用）

宗福季（香港科技大学）

时间：14:45~15:30 邮箱：season@ust.hk

简介：宗福季教授现任香港科技大学工业工程与物流管理系教授，前系主任，及质量实验室主任，国际质量科学院 (IAQ) 院士，美国工业工程师学会 (IIE) 会士，美国质量学会 (ASQ) 会士，国际统计协会 (ISI) 当选会员，香港工程师学会 (HKIE) 会士。任职科大后，他积极参与有关质量改善和管理的教育及研究工作。他是大中华地区首名荣获美国质量学会 (ASQ) 六西格玛黑带的权威，亦是该学会特许的六西格玛黑带大师之一。宗教授目前是美国质量学会旗舰期刊 *Journal of Quality Technology* (JQT) 的主编，工业工程学会期刊 *IIE Transactions* 及 *Technometrics* 的副编辑。宗教授于国立台湾大学取得机械工程学士学位，其后于美国密歇根大学获工业工程硕士及博士学位。

摘要：随着信息技术与数据采集技术的迅速发展，在许多应用中人们越来越需要综合地利用多个数据源和多个领域的信息。近年来，迁移学习 (transfer learning) 提供了结合各个领域信息的有效框架。通过迁移来自源领域 (source domains) 的已有的知识，目标领域 (target domains) 里相似的问题可以得到更有效的解决。在迁移学习的框架下，统计模型与统计方法起到了很重要的作用，然而现有的迁移学习的综述多集中在机器学习领域，并没有强调统计模型与方法的应用。这次讲座将综述统计迁移学习 (statistical transfer learning)。通过总结迁移学习文献中的统计模型和统计方法，我将展示统计研究如何更好地帮助解决迁移学习问题。另外，我将讨论统计迁移学习在现实中有关统计过程控制，过程监控及质量控制的应用。

中产阶级如何利用量化投资工具完成财富进阶

邓一硕（懒投资）

时间：16:00~16:45 邮箱：dengyishuo@zuinianqing.com

简介：邓一硕，毕业于中央财经大学，北京大家玩财务总监、副总裁。曾参与翻译《R 核心技术手册》，《R 图形可视化手册》，《量化投资分析与 R 语言》。

摘要：随着经济的发展和人均收入的不断提高，拥有储蓄和投资能力的中产阶级人数大量增加。与此同时，房价的上涨，健康的投入又极大的消耗中产阶级的积蓄。如何利用量化投资工具，辅助进行资产配置决策，从而使得自身财富实现保值增值？本演讲将结合典型案例来分享量化投资工具在个人资产配置中的作用。

圆桌讨论

时间：16:45~17:30

摘要：圆桌讨论环节，各特邀嘉宾将就大数据时代的机遇与挑战、数据科学在各行业中的应用与前景、如何推动数据科学在国内的发展、如何促进学界与业界的配合等议题发表意见并展开讨论。

致辞

李广雨 (狗熊会)

时间: 13:55~14:00

物联网大数据分析技术在供应链金融保险和风控领域的应用?

叶征 (天津星通天安科技有限公司)

时间: 14:00~14:30 邮箱: davideye2000@126.com

简介: 毕业于北京大学光华管理学院, 师从著名金融学家曹凤岐教授。目前任星通天安科技有限公司副总裁, 主要负责物联网和车联网保险业务以及物联网金融风险管理等工作, 星通天安是一家专注于物联网大数据金融服务的公司。加入星通天安管理层前, 叶先生曾在中国人寿保险(集团)公司任职六年, 主管保险集团资产负债管理(ALM)、资本规划与经济资本(EC)管理等工作, 任职期间曾被中国人寿外派澳大利亚悉尼工作, 多篇论文在《保险研究》等核心期刊发表。除此之外, 叶先生还先后在国际著名投资银行、国际知名咨询机构和国际顶级金融保险集团实习、工作或任职, 参与了多家大型金融机构的全面风险管理(ERM)、内部评级法(IRB)建设中评级模型的开发、验证和优化, 以及经济资本管理的研究与应用等工作。

摘要: 虽然“互联网+”在时下极其引人注目, 但未来面临的重要发展趋势是“物联网+”, 因为服务于产品流通贸易环节的电子商务只是智能生产的一部分, 而物联网将囊括生产、贸易和使用所有环节。另外, 传统互联网是用户有意识的与网站发生交互留下行为信息, 而物联网却能在用户尚未意识到的情况下完成信息的搜集, 使物完全脱离人的状态获得感知与被感知的能力。由此, 物理世界与网络虚拟世界被打通形成互动反馈, 这样风控、金融和保险服务才能真正做到客观、实时、动态、前瞻。本演讲将报告物联网大数据分析技术在金融保险和风控领域的商业实践探索。

移动程序化广告

苏永刚 (蓬景数字)

时间: 14:30~15:00 邮箱: NA

简介: 北京大学计算机体系结构专业理学硕士学位。蓬景数字技术中心总经理, 负责蓬景数字广告发布平台与数据分析平台的产品研发、运营工作。在计算机系统结构、云计算与虚拟化技术、海量数据处理与高并发业务处理等方面有超过 10 年的深入研究和丰富经验。

摘要: 本报告介绍移动端的程序化广告, 即在移动设备上为广告主的精准营销需求提供全方位的服务。包括以多种多样的数据为基础的、建立在传统互联网广告业务的基础上的, 结合移动互联网的特点及优势的多种面向效果的解决方案。本报告还将介绍蓬景数字以及狗熊会联合研究组在针对不同的场景以及数据的相关研究工作, 包括基于数据的方法自动对广告出价、定向、投放、反馈、优化等各个环节。

数据融合与信用风险评估

葛伟平 (考拉征信)

时间: 15:00~15:30 邮箱: NA

简介: 葛伟平, 考拉征信服务有限公司执行总裁兼首席技术官, 2005 年复旦大学计算机软件博士毕业, 带领团队先后推出了企业和个人信用分模型、互联网金融征信服务平台、考拉云智风控引擎平台, 树立了考拉征信在征信行业品牌, 同时兼任中国科学院大学 <U+FF65> 考拉征信模型实验室主任。

摘要: 本次主要分享考拉征信依托海量数据为银行等提供信用卡申请评分模型构建服务的一般方法, 并介绍模型设计和数据处理的一般流程。

证券分析师的价值分析

赵锡刚 (对标科技)

时间: 15:30~16:00 邮箱: NA

简介: 毕业于同济大学运输管理工程专业, 99 年加入中国惠普有限公司任电信行业销售代表, 06 年成为惠普最年轻的行业销售总监。08 年加入安永会计师事务所全球电信中心负责中国区电信行业咨询业务, 当中结识俏江南张兰, 11 年进入俏江南全面负责俏江南上市业务和集团对外的整体业务, 经过 4 年的不断努力, 经历 A 股转 H 股, 从独立上市到卖给 CVC 的全过程, 2015 年离开俏江南, 2016 年 3 月成立对标科技。

摘要: 证券分析师就是给市场提供投资建议的人, 投资人听了他的建议交易股票, 分析师赚取交易的佣金提成。这种人可不容易的; 首先严格的准入条件, 要去考试拿资格证书, 其次, 严格的语言用词限制, 发出之前必须过内审, 必须实事求是, 不能引起市场恐慌, 再次严格的监管个人的语言和行为, 以防各种内部交易。但是国内分析师的评级可信吗? 分析师有用吗? 肯定有用, 为什么? 各大券商都花重金建立自己的分析团队, 要是没用早就都开除了。那怎么知道谁有用谁没用, 尤其是中小机构和个人投资者, 自己没有什么判断能力, 市场上有 4000 多位分析师, 每天发出将近 500 份研报, 平均都在千字左右, 怎么看啊, 看谁的啊! 这是一个甄别分析师分析能力的问题, 分析师分析的准不准主要是看他推的股票涨没涨, 涨了多少, 是不是跑赢了行业, 他是最初价值的发现者, 还是人云亦云的跟风者, 亦或是发了报告股票就下跌的悲催者。对标科技通过汇总所有历史上的分析师行为数据进行挖掘, 以收益率为核心通过统计分析, 为每一位分析师建立数据模型, 鉴别分析师的分析能力, 并将不同行为特点的分析师归类, 为中小投资者提供精准化的分析师群体行为的实时动态跟踪、关键信息的实时提醒并积累投资人行为; 最终制定自己的投资策略。

基于车联网数据的商业价值探索

周扬 (彩虹无线)

时间: 16:00~16:30 邮箱: zhouyanga9@gmail.com

简介: 周扬, 彩虹无线数据科学家, 数据科学部总监, 浙江大学客座讲师, 四川大学生物信息/生物统计专业硕士, 拥有国家发明专利一项, 先后在 NAR、Bioinformatics 发表论文三篇, 累计影响因子超过 18。多年来致力于车联网数据与汽车行业数据的价值研究, 为汽车智能制造、车辆工况研究、创新车险等方面提供数据赋能。

摘要: 当前,汽车行业整体处于数据来源一致性低、数据质量一般且可用性不强的基本状态。车联网数据作为采集频度高、数据质量好、来源稳定的数据源,成为了汽车主机厂商及周边行业的关注热点。其中包括的方向有:车险 UBI、无人驾驶、汽车营销、汽车后市场、车辆的生产制造及工况研究等核心方向。此次演讲,将基于彩虹无线多年来在车联网大数据行业的深耕,围绕实际商业应用场景,结合统计模型及算法,分享基于车联网数据商业应用的实践案例。

浅谈消费金融

兰伟 (西南财经大学 & 柠檬科技)

时间: 16:30~17:00 邮箱: lanwei@swufe.edu.cn

简介: 西南财经大学统计学副教授, 柠檬科技创始人。

摘要: 介绍目前消费金融的线上线下运营模式, 流量获取方式和风控模式, 以及目前网络图谱在反欺诈中的应用和进展。

Quantitative Venture Capital

吴海山 (合一创投)

时间: 8:30~9:00 邮箱: hswu85@gmail.com

简介: 吴海山, 合一创投 (Heyi Ventures) 首席数据科学家。2011 年从复旦大学博士学位, 毕业后加入 IBM 中国研究院。2012 年底加入美国普林斯顿大学进行博士后研究。2014 年 9 月至 2017 年 4 月任职于百度研究院大数据实验室, 担任百度时空大数据研究负责人。先后负责了百度经济测量、百度人群预警系统、百度商业地产选址系统等多个课题。研究成果获得了国内外知名媒体的广泛报道 (如 the Wall Street Journal, Bloomberg, the economist, Forbes, CNBC, CNN Money, MIT Technology Review, New Scientist, NPR, Washington Post, China Daily 等), 研发的百度经济指数每月 5 号会在彭博终端上更新。

摘要: 投资是一个艺术与科学的结合。对于二级市场投资来讲, 基于数据的量化投资策略已经取得了瞩目的成绩。但是对于一级市场的风险投资而言, 传统风险投资家在进行决策时, 更多倾向于通过自己的经验和直觉来进行决策, 数据和机器学习算法还未成为主流的方式。近年来随着互联网数据、可穿戴设备、小型卫星、物联网等多种传感器设备的普及, 我们越来越方便的可以对金融市场进行有效分析, 类似 Google Ventures, Correlation Ventures, KPCB 等多家 VC 公司也开始进行了基于数据的风险投资决策。这次讲演将主要介绍数据、机器学习是如何影响目前的风险投资市场, 以及将来的发展趋势。

手机数据与经济活动测度

董磊 (百度研究院)

时间: 9:00~9:30 邮箱: arch.dongl@gmail.com

简介: 清华大学建筑学学士、经济学学士、工学博士, 主要研究方向为时空数据分析。曾工作于百度 Big Data Lab, 从事基于移动端数据的分析与建模工作。研究论文发表于多个知名学术期刊, 并被 Economist, Bloomberg, New Scientist, MIT Tech Review 等专业媒体广泛报道。

摘要: Emerging trends in smartphones, online maps, social media, and the resulting geo-located data, provide opportunities to collect traces of people's socio-economical activities in a much more granular and direct fashion, triggering a revolution in empirical research. These vast mobile data offer new perspectives and approaches for measurements of economic dynamics and are broadening the research fields of social science and economics. In this paper, we explore the potential of using mobile big data for measuring economic activities of China from a bottom-up view. Firstly, We build indices for gauging employment and consumer trends based on billions of geo-positioning data. Secondly, we advance the estimation of store offline foot traffic via location search data derived from Baidu Maps, which is then applied to predict revenues of Apple in China and detect box-office fraud accurately. Thirdly, we construct consumption indicators to track the trends of various industries in service sector, which are verified by several existing indicators. To the best of our knowledge, we are the first to measure the second largest economy by mining such unprecedentedly large scale and fine granular spatial-temporal data. Our research provides new approaches and insights on measuring economic activities.

迁移学习在金融大数据风控中的应用

殷磊 (融 360)

时间: 9:30~10:00 邮箱: ylfego@163.com [13811747306](tel:13811747306)

简介: 现任融 360 天机风控 tech leader。曾任去哪儿技术总监, 百度资深架构师。北京理工大学计算机科学技术博士。专注大数据与人工智能方向研究。

摘要: 风控是金融领域研究的核心之一, 也是大数据应用的经典场景。金融产品丰富多样, 受众群体分布广泛, 不可能为其设计一个大而全且通用的风控模型。因此, 为不同的金融产品设计针对性的风控模型是非常必要的, 这正是迁移学习的用武之地。迁移学习不仅可以解决单一产品数据过少, 特征单一的问题, 还可以捕捉产品间相关性, 有效的识别个性化风险与系统化风险。

如何制造一次成功的投资

李翛然 (北京奇点创世信息技术有限公司)

时间: 10:30~11:00 邮箱: xrli_office@foxmail.com

简介: 李翛然, 北京奇点创世信息技术有限公司创始人。先后从事过寿险精算, 投资银行工作。于 2014 年创办北京奇点创世信息技术有限公司, 主要业务领域为二级市场金融风险管理。现已有 10 余家金融机构、私募基金采用该系统为客户和自营交易提供风险管理及投资顾问服务。其主要工作经历覆盖了一级市场的发行, 尽职调查, 搭建企业信用分析系统, 二级市场的量化分析, 风险管理 saas 系统。

摘要: 量化投资是近年来金融行业最火热的话题之一, 从高频, 套利交易, Alpha, 事件驱动, FOF 投资, 新的投资方法层出不穷, 那么, 到底一个投资者该如何选择策略? 这些策略的背后逻辑到底有哪些优点缺点? 在传统的金融学术和量化投资实战之间的巨大鸿沟有没有一些通用的方法论可以让一个新人成长? 这次简短的分享可以让大家对传统投资到量化投资有一个清晰而理性的认识, 同时可以对自己将来的投资生涯提供很多有意义的指导和帮助。

自由讨论

时间: 11:00~11:30

基于区域关联视角的智慧城市发展

吴梦荷 (北京清华同衡规划设计研究院有限公司)

时间: 8:30~9:00 邮箱: *viola_wumenghe@163.com*

简介: 城市与区域规划专业理学硕士, 现任职于清华同衡规划设计研究院技术创新中心, 从事城市规划相关的数据分析与数据分析产品研发。

摘要: 智慧城市旨在以新的科学技术手段优化城市发展路径, 这一范畴则包含了从微观到宏观的一系列尺度上的问题, 区域问题也是智慧城市发展的重要部分。如今网络化的城市关系正在形成, 城市的良性发展与区域关系密不可分, 因此以区域的视角解析智慧城市发展路径十分重要。研究基于智慧城市的理念和内涵, 探索基于城市间关联的区域分析框架, 采用新的数据源和技术方法, 把握区域发展格局、统筹城市间发展关系, 透视交通、人口、资本等特征, 从而为宏观区域发展提供智慧的解决之道。

环境大数据的商业应用

顾竹 (北京佳格天地科技有限公司)

时间: 9:00~9:30 邮箱: *guzhu@gagogroup.com*

简介: 北京佳格天地科技有限公司产品副总裁。南京师范大学本科、硕士, 美国纽约州立大学博士, 人工智能和大数据专家, 前 NASA 深度学习研究员。在美国纽约州立大学攻读博士期间, 就被 NASA 邀请参与遥感影像的重要项目。多年来专注遥感影像的深度学习。佳格是世界上首先采用深度学习来分析图像获取农业信息的公司。

摘要: 基于卫星遥感、GIS 等数据, 采用统计分析和机器学习技术, 可以挖掘出更为丰富的信息, 为社会生产、发展等各个领域应用。环境大数据智能共享云平台技术, 包含了针对空间环境数据特定优化的数据采集、分析、发布与可视化多个模块, 能够实现对气象, 环境, 地质等多类型环境数据的实时监测和关键环境变量的预报预测。其作为颠覆整个环境大数据行业的历史性突破技术, 获得国内外专家的广泛认可, 市场价值极其可观。

地理数据与商业网点选址实战

张志成 (NA)

时间: 9:30~10:00 邮箱: *94117106@qq.com*

简介: 《数据实践之美》合作者, 10 余年的商业网点选址分析与开店工作经验。服务过百胜餐饮、沃尔玛中国等公司, 曾作为外部顾问参与 IBM 农行网点优化。项目经验涵盖零售、餐饮、金融与服务、体验等商业业态。基于行业经验, 能够从业务角度正确解读数据。掌握主流的数据工具与简单的编程能力, 能够通过数据分析洞察业务机会。对数据驱动商业选址有一手的实战经验。

摘要: 电商与数据时代, 实体店作为重要的消费场景, 各种 app 推广主力渠道等, 在各种渠道中地位将会被继续强化, 新零售也开始通过数据来武装实体渠道, 从而帮助实体渠道能够实现科学选址、精细化运营等, 然而数据的应用应该首先以行业知识为基础和导向, 本次将会分享开店实战中是如何利用数据科学选址的, 从

中可以看到数据选取、方法与业务融合、执行落地缺一不可，也将会了解到数据时代实体渠道的更强生命力与机会点在哪里。

基于室内定位数据 (IPS) 的时空行为分析

黄蔚欣 (清华大学)

时间: 10:30~11:00 邮箱: huangwx@tsinghua.edu.cn

简介: 黄蔚欣, 清华大学建筑学院副教授, 日本京都大学博士, 数字建筑技术教学工作委员会副主任委员, 亚洲计算机辅助建筑学会 (CAADRIA) 委员, 中国建筑学会建筑师分会数字建筑设计专业委员会 (DADA) 联发起人, 清华大数据产业联合会会员。主要研究领域: 数字建筑设计、大数据行为分析, 设计认知等。

摘要: 时空位置信息对理解人群的环境行为具有重要的意义。传统的环境行为学研究方法使用拍照、绘图、跟踪、问卷等方式调查人们的行为, 可以较准确的记录人们的时空位置和活动的内容, 为分析少量个体在特定时段的行为提供了较为全面和准确的信息。然而, 这样的调研方式也存在样本数量少, 覆盖时间短、空间有限的不足。使用室内定位系统 (IPS) 的时空位置数据, 能够分析人群在大型公共建筑室内外空间、社区公共空间、居住空间等的行为, 总结行为模式, 比较不同人群、功能空间的特点, 为建筑设计、商业运营、公共安全管理提供动态依据。

不可或缺的优质地理大数据

高楠 (GeoHey)

时间: 11:00~11:30 邮箱: gaon@geohey.com

简介: 北京极海纵横信息技术有限公司 (GeoHey) 数据总监, 在地理数据治理、地理数据分析及可视化、地理信息商业咨询服务等专业领域积累八年经验, 曾为沃尔玛中国、万科等企业提供大数据服务, 专注于研究海量地理数据挖掘解决方案。

摘要: 在过去的工程实践中, 数据是一种比较稀缺的资源, 很多时候手握最好的硬件、软件、算法但苦于没有数据使得这些优质资源无用武之地。随着时间的推移, 数据的稀缺性渐渐降低, 数据甚至呈现出井喷的趋势, 越来越多的数据随处可见, 但数据质量参差不齐。尤其在数据挖掘、人工智能快速发展的时代, 人人都能手握最好的工具、模型、算法, 却难以准备出一份合格可用的数据供分析使用。如何高效的研发高质量的地理位置数据是我们重点开展的工作之一。

PM 2.5 数据的时空特征及统计建模

朱雪宁 (北京大学)

时间: 11:30~12:00 邮箱: xueningzhu@qq.com

简介: 光华管理学院商务统计系四年级博士生。研究上关注社交网络、高维数据、环境数据分析等; “狗熊会”公众号专栏作者。曾在 Annals of Statistics 以及 Statistics and Its Interface 有所发表。

摘要: 如今, PM 2.5 已经成为生活中经常谈论的高频词, 同时, 许多研究、报道也表明 PM 2.5 给呼吸系统、心肺功能带来不良影响, 危害健康。为了了解这一雾霾指标的时空分布规律, 本演讲从数据分析及统计建模的层面对 PM 2.5 数据进行研究。具体地, 本演讲将讨论 PM 2.5 相关的动态及空间相关特征。从统计建模上, 利用带有空间相关性的自回归模型对时空特征进行了建模。最后, 本研究给出空间中未知地点的预测插值方法。

计算与人文：作为新领域的“数字人文”

陈静（南京大学）

时间：8:30~9:00 邮箱：cjchen@nju.edu.cn

简介：陈静，南京大学艺术研究院副教授。南京大学博士，Rice University 博士后。主要研究兴趣为：文化与媒介研究、数字人文、新媒体艺术。

摘要：数字人文 (digital humanities)，源于“计算人文” (computing humanities)，是近 20 年来兴起的一个多学科交叉领域，研究主题从中世纪手稿的文本分析、历史文献主题挖掘、元数据框架、诗歌分析到计算机游戏、艺术品风格分析等，其参与主体包括艺术家、人文学者、社会科学家、统计学家、计算机科学家、地理专家、软件工程师等。数字人文主要关注的是在当今信息社会语境下，在知识生产方式及知识进行转型的重要时期，如何解决新出现的人类社会文化问题，或者通过新的研究方法、路径和工具对人文学科的进行再认识、再研究和再挖掘。

本发言将主要从“计算与人文”的关系对“数字人文”的发展脉络进行简要说明，并在此历史框架内，结合具体个案对统计学方法、自然语言分析、主题模型、社交网络、HGIS、Pyton、R 语言等数字人文学者常用的分析方法或者工具对人文研究的作用和影响进行说明。

network diffusion: Simulate and Visualize Network Diffusion

王成军（南京大学）

时间：9:00~9:30 邮箱：wangchj04@126.com

简介：Cheng-Jun Wang is currently an assistant research fellow in the School of Journalism and Communication, Nanjing University. He is the director of Ogilvy Data Science Lab, and also a research member of Computational Communication Collaboratory and Web Mining Lab. His research on computational communication appears in both SSCI and SCI indexed journals, such as Scientific Reports, PloS ONE, Physica A, Cyberpsychology.

摘要：network diffusion, a R package which can help simulate and visualize the network diffusion. <https://github.com/chengjun/networkdiffusion> Network diffusion research focuses on how network structure exerts its impact on the diffusion process. The networkdiffusion package would help you simulate and visualize the most simple network diffusion with R. The algorithm is quite simple:

Generate a network g: g(V, E). Randomly select one or n nodes as seeds. Each infected node influences its neighbors with probability p (transmission rate,). Slides: <http://chengjun.github.io/network-diffusion>

情感现象学与色彩政治学——唐诗色彩词的数字人文研究

郑文惠（台湾政治大学）

时间：9:30~10:00 邮箱：wenhuei_cheng@yahoo.com.tw

简介：郑文惠，台湾政治大学文学博士，现任台湾政治大学中文系教授。

主编中国近现代思想及文学史专业数据库 (1830-1930)、《东亚观念史集刊》、《革命·启蒙·抒情——中国近现代文学与文化研究学思录》等。著有《文学与图像的文化美学——想像共同体的乐园论述》、《诗情画意——明代题画诗的诗画对应关系》、《钱选》、《王绂》、《中国书画传习汇编》等书, 及古典诗歌、文学与图像、遗民诗画、汉画、晚明版画、近代画报、近代小说、文学地景与记忆认同、世变与乐园、观念史、数字人文学等论文。

现主持「世变与文心/画像/书体——东汉末期价值逆反与文化再现」与观念话语行动: 数位视野下中国/台湾多元现代性研究」、「新/旧」的激变与交锋: 中国现代性形成的数位人文研究观念科技部个人型计划与整合型计划, 及「中国认同与现代国家的形成」、「中国近现代思想及文学史专业数据库 (1830-1930)」教育部迈向顶尖大学计划。

曾任哈佛大学、莱斯大学、海德堡大学、捷克国家科学研究院、日本关西大学、国际日本文化研究院、韩国江原大学、韩国翰林大学、复旦大学、中国美术学院、福建师范大学、江苏师范大学、新加坡南洋理工大学、香港教育学院等有关叙事文学、书法文化美学、文学与图像、遗民诗画、从遗民到后遗民的时间地理政治学、桃花源历史地理政治学、视觉文化与中国近现代画报、中国近现代报刊与文化研究、观念史、数字人文学等讲座或演讲。

摘要: 作者: 郑文惠 *、余清祥 **、颜静馨 ***、刘昭麟 ****、邱伟云 *****

摘要: 本次演讲主要以台湾政治大学历史与思想数字人文实验室团队近年来以数字技术进行文学文本研究之重要成果与未来开展为内容。以古典诗歌作为数字人文方法实践之场域, 主因于中国古典诗歌大多以具体可感的形象描摹抽象的心理情感, 亦即诗歌中一个个词组, 几乎是传达诗人心理情感的一个个意象, 承载了象征诗人心理情感与思想观念的意义系统。而诗人在独特的身体感知中, 以诗歌的修辞技术, 标记出本己的思想情感, 呈显为独特的诗歌风格, 从而蔚为一代的记忆表征, 也积淀了世代间不同的思想价值与文化风俗。2015年, 我们借用高分子化学的“分子链”概念, 施作于诗歌的意象丛及主题研究等, 借由数字技术尝试从“句链”中勾勒出色彩词在诗歌中的构词, 及其出现位置与对仗词、搭配词, 考察其中所透显出的情感现象学与色彩政治学。2016年, 我们纳入颜色词的同义字, 在原有技术基础上, 结合 R 进行文本探勘, 运用统计理论模型, 更为全面且深入地研究唐诗颜色光谱学。大体而言, 唐诗颜色光谱学除与诗人个人独特的联觉通感、视觉想像、心理情感与感觉结构息息相关外, 还涉及佛道宗教信仰、经世与隐逸思想、园林文化、祭典仪式、身分地位、染织技术、彩绘技术、化妆术…等等, 从中不仅可掌握诗人独特的颜色修辞与诗歌主题风格的关系及其深层的颜色心灵光谱, 也可深入理解透过各期唐诗多重性的颜色光谱所开展的隐喻系统, 及所表征不同时期宗教、思想、技术、政治、经济、阶级等社会文化的变革。2017年我们将尝试拓展可纳入句链结构的元素, 以“色彩词”为对象, 将古典诗歌重要的声、律与前述技术成果结合, 探索声音在诗歌中如何与情感、意象互动, 而数字技术又能为深具传统的古典诗歌研究带来怎样的风貌, 此为本团队系列研究未来开展的方向。

* 台湾政治大学中国文学系教授。(通讯作者)

** 台湾政治大学统计学系教授。

*** 台湾中正大学中国文学系博士生。

**** 台湾政治大学资讯科学系特聘教授。

***** 山东大学历史文化学院副研究员。

群像的描绘与类型的分析: 用数字工具挖掘《德意志人物志》

王涛 (南京大学)

时间: 10:30~11:00 邮箱: t.wang@nju.edu.cn

简介: 会编程的历史学家

摘要: “历史学的数位转型”是大势所趋, 它将在宏观层面影响历史学的整体面貌, 在微观层面改变个体史学研究者的工作方式。在中文学术圈数字人文方兴未艾, 但这种思路与方法主要被用来研究中国问题。中文

学界从事世界史研究的学者鲜有涉猎数字人文的佳作。本课题是运用数字人文工具研讨世界史问题的一次有益尝试：以德意志学界重要的人物传记参考书为蓝本，对历史人物进行了群体与类型的研究。传统的人物研究也以个体传记为主，本课题开创性以德意志群体人物为研究对象，并且主动运用数字史学的观念与方法，力图在德意志人物传记的研究中发现隐含的问题。本课题的具体应用，将拓宽我们对德意志历史的认识，加深我们对欧洲文明的理解；同时，我们在新工具与新思维的具体运用中结合历史问题的分析，不仅能够对传统结论提出改进意见，也能够在学术实践中对数字史学的技术进行评判，从而推动数字人文的发展。

词汇、概念、数字：文本探勘技术于中国近代观念史研究中的应用与实践

邱伟云（山东大学）

时间：11:00~11:30 邮箱：brianacwu@163.com

简介：邱伟云，博士，山东大学历史文化学院副研究员，台湾政治大学历史与思想数位人文实验室成员。

摘要：关键词、观念史与概念史研究法，皆以辞汇为研究对象，关注辞汇自身及其在修辞结构乃至话语系统中的变化状况，为人文学领域中着重辞汇研究的一套人文研究法。自然语言处理与文本探勘方法，亦以辞汇为研究对象，着重于辞汇撷取技术以及辞汇在文本脉络中的视觉化呈现，及核心词汇与其他关键词共现互动现象，为资讯科学领域中着重辞汇研究的一套数字研究法。从上述两套分属人文与资科领域之研究法说明可知，两套方法都共同关注辞汇，因此产生了跨领域协作研究的可能，也使数字技术得以与人文研究产生对话空间，因此关键词/观念史/概念史研究法的数字化转向，可说是百花齐放的数字人文学发展中一道不可忽视的风景。本次演讲正欲以过去结合文本探勘技术与中国近代观念史研究的诸多案例，说明数字技术如何协助人文研究？人文思维又可提供数字技术哪些思考方向？数字与人文该如何搭配才能进行协同研究？报告将以几种已运用于中国近代观念史研究上的数字人文方法为例，说明这些方法的操作过程及其优点，以及人文学者在数字人文协作研究过程中怀有哪些疑问？遭遇哪些难题？还希望有什么突破？对于未来发展前景有何期待？以上即是本次演讲的主要内容所在。

自由讨论

时间：11:30~12:30

卓越质量管理中的大数据分析

王凯波 (清华大学)

时间: 8:30~9:00 邮箱: kbwang@163.com

简介: 王凯波博士是清华大学工业工程系教授、系副主任。他在香港科技大学获得工业工程与工程管理学博士学位。王凯波的研究主要关注复杂系统的质量建模、监视与控制。他是多个自然科学基金与企业资助科研项目的负责人，在质量控制领域 SCI 索引的国际期刊发表了 30 余篇论文，其中包括 Journal of Quality Technology, IIE Transactions, IEEE Transactions of Automation Science and Engineering, Quality and Reliability Engineering International 等。王凯波博士现为 INFORMS 质量、统计与可靠性分会 (QSR) 主席，是美国质量协会 (ASQ) 资深会员，IIE、INFORMS、IEEE 会员。更多信息，请访问 <http://www.ie.tsinghua.edu.cn/kbwang/>。

摘要: 质量管理就是走独木桥，而大数据为拓宽质量管理的道路提供了新的广阔支撑。大数据除了在网络和社交媒体领域存在之外，在各类工程系统中同样广泛存在，而且价值巨大。本报告将以工程质量改善项目为案例，介绍各类制造数据在质量管理和提升项目中的应用。案例包括劳动力密集型企业质量改善、半导体生产过程质量改善、太阳能生产质量改善等。

质保数据建模与分析

何曙光 (天津大学管理与经济学部)

时间: 9:00~9:30 邮箱: shuguanghe@tju.edu.cn

简介: 何曙光博士，天津大学管理与经济学部教室。他在天津大学管理学院获管理科学与工程博士学位。何曙光的研究主要关注基于数据分析的质量改进、过程监控和基于质保数据的产品可靠性评估等。近年来在学术期刊发表论文 30 余篇，包括 Reliability engineering and system safety, Journal of quality technology, Annals of operations research 等。

摘要: 当前几乎所有的耐用品都提供质量保证，在规定的质保期有生产方或销售方对失效产品进行免费维修或更换。在质保服务过程中，会累计海量的数据。本报告某汽车制造企业质保数据为例，介绍二维质保条件下的产品可靠性评估、质保索赔预测等方面的研究进展和应用。

System reliability assessment and optimization

李彦夫 (清华大学)

时间: 9:30~10:00 邮箱: liyanfu@tsinghua.edu.cn

简介: 李彦夫，博士，博士生导师，现任工业工程系教授，入选 2016 年国家青年千人计划。长期致力于系统可靠性评估与优化方法的研究，以及将其应用于可再生能源系统，核能和计算机软件系统，并取得了一系列原创性学术成果。在可靠性，电力以及软件工程知名期刊发表多篇论文。主持或参与多项企业委托项目，合作方包括法国电力公司，阿尔斯通等公司。IEEE 高级会员，可靠性工程顶级期刊 IEEE Transactions on Reliability 副主编，中国航空学报青年编委。

摘要: Reliability is a fundamental attribute for the safe operation of any modern technological system. The demands from various industry sectors for the quantification of system reliability date back to the early 20th century and steadily grow till our times. Furthermore, the search for optimal system design, operation and maintenance strategies that minimize expense and maximize reliability has become an increasingly relevant task since the 1960s. These tendencies render the system reliability assessment and optimization two important topics in academic research and two necessary tasks in industrial applications. Consequently, a number of models and methods have been developed. Yet, new challenges emerge from the latest technological systems or the ongoing projects, such as the smart grids, mainly characterized by the complex and possibly intelligent behaviors of the components and the hybrid uncertainties embedded in the available modeling information.

Developing new methods to confront these challenges is the goal of my research. The research works are grouped under the two main axes: 1) reliability assessment of components and systems; 2) optimization.

Branding with social media: User gratifications, usage patterns, and brand message content strategies

皋琴 (清华大学)

时间: 10:30~11:00 邮箱: gaoqin@tsinghua.edu.cn

简介: 皋琴, 副教授, 工学博士(清华大学), 现任清华大学人因与工效研究所所长, (美国)人因与工效学会(HFES)中国分部主席, International Journal of Human-Computer Interaction 期刊编委, 2013 年入选北京市高校青年英才计划。研究方向: 复杂系统中的人机交互、社会化计算与用户体验、通用设计、跨文化研究和服务设计等。

摘要: The emergence of social media provides a new platform for developing brand-consumer relationships. The aim of the current study is to examine the differences in Chinese users' gratifications of different social media and the impact of brand content strategies on the quality of brand-consumer communication via social media. In the first study, 209 SNS and 161 microblog users were surveyed. Five dimensions of social media gratifications emerged from the factor analysis. Significant differences in the strengths of gratifications were found between SNS and microblog users. Usage patterns of SNS and microblog are analyzed and compared. In the second study, we examined the impact of users' gratification and the type of social media on the effectiveness of different brand content strategies through a two-week experiment involving 60 SNS users and 61 microblog users. Implications for developing branding strategies on different social media platforms are discussed.

基于车辆 GPS 数据的交通大数据应用

姜海 (清华大学)

时间: 11:00~11:30 邮箱: haijiang@tsinghua.edu.cn

简介: 姜海博士现任清华大学工业工程系副教授、博士生导师, 运筹与统计研究所副所长, 2016 年获国家自然科学基金 - 优秀青年科学基金(“优青”)资助。现任中国运筹学会 - 行为运作管理分会秘书长, 中国运筹学会 - 随机服务与运作管理分会理事, Computers & Industrial Engineering (IF=2.086, 工业工程领域 SCI 期

刊排名 9/44) 交通方向的分区编辑 (Area Editor)。他擅长将消费者行为模型、数据挖掘技术和大规模优化方法三者结合, 从系统的角度分析问题, 为政府、企业和个人提供以定量模型为基础的解决方案和决策工具。

摘要: 我们将围绕车辆 GPS 数据介绍若干大数据应用, 包括:

1. 城市路网的自动识别
2. 驾驶员驾驶风险的评判

我们将基于车辆 GPS 数据构建优化模型, 并在实际问题中对模型的性能进行检验。

自由讨论

时间: 11:30~12:00

生物序列分类中的特征快速生成与可视化

杜朴风 (天津大学)

时间: 8:30~9:00 邮箱: pufengdu@gmail.com

简介: 天津大学教师, 从事生物信息学研究, 主营生物序列分类业务。

摘要: 在生物序列分类过程中, 我们需要快速的生成特征, 也需要通过可视化来帮助进行分类算法的设计和选择。在这个报告里, 我们将讨论一些常用的特征生成技术, 以及利用 R 所进行的特征可视化。

Hepatocellular carcinoma study based on HBV next generation sequencing

张淑芹 (复旦大学)

时间: 9:00~9:30 邮箱: zhangs@fudan.edu.cn

简介: 博士毕业于香港大学数学系, 目前为复旦大学数学学院副教授。主要研究方向为计算数学、统计学、最优化方法在生物及医学数据中的建模、计算及相关分析, 尤其网络数据的建模及分析。

摘要: Hepatocellular carcinoma (HCC) is one of the most common type of cancer in our country. There have been many studies on it. In this talk, we will introduce our recent work on HCC classification based on HBV next generation sequencing data. The clinical phenotype data are also analyzed, and their relations with HBV are studied.

Prediction analysis for microbiome sequencing data

王涛 (上海交通大学)

时间: 9:30~10:00 邮箱: neowangtao@sjtu.edu.cn

简介: 王涛: 2007 年东南大学数学系学士, 2010 年华东师范大学金融与统计学院硕士, 2013 年获香港浸会大学数学系统计学专业哲学博士学位。2014 年赴美国耶鲁大学公共卫生学院生物统计系从事博士后研究工作, 2016 年 1 月回国任上海交通大学特别研究员。

主要致力于研究高维复杂数据的统计降维技术和变量选择技术, 以及研究人类微生物组数据等生物医学数据的统计分析方法。近年来分别在 Journal of the American Statistical Association、Journal of the Royal Statistical Society: Series B、Biometrika、Biometrics、Bernoulli、Statistica Sinica、BMC Systems Biology 等知名学术期刊上发表 SCI 论文二十余篇。

摘要: One primary goal of human microbiome studies is to predict host traits based on human microbiota. However, microbial community sequencing data present significant challenges to the development of statistical methods. In particular, the samples have different library sizes, the data contain many zeros and are often over-dispersed. To address these challenges, we introduce a new statistical framework, called predictive analysis in metagenomics via inverse regression (PAMIR). We demonstrate the advantages of PAMIR through numerical studies.

条件随机场及其在生物信息学中的应用

吴凌云 (中科院)

时间: 10:30~11:00 邮箱: lywu@amss.ac.cn

简介: 吴凌云, 中国科学院数学与系统科学研究院研究员, 博士生导师, 应用数学所运筹学研究室主任, 生物信息学研究中心主任. 中国运筹学会常务理事, 科普工作委员会副主任, 计算系统生物学分会副理事长. 2002 年于中国科学院获运筹学与控制论专业理学博士学位. 曾在香港科技大学和美国康奈尔大学 Weill 医学院从事博士后研究工作. 目前的研究兴趣是运筹学与生物信息学, 特别是运筹学方法在生物信息学与系统生物学中的应用. 主要工作包括: 测序算法, 单体型推断, 蛋白质结构预测与比对, 蛋白相互作用预测, 蛋白质修饰位点预测, 分子生物网络分析比较, 复杂疾病生物标记物建模等. 主持过青年基金, 面上基金, 重大研究计划培育项目等多项国家自然科学基金. 2014 年获中国运筹学会青年科技奖.

摘要: 海量分子生物学数据和复杂数据结构对现有的生物信息学模型和算法提出了巨大的挑战。条件随机场是一类重要的概率图模型, 是隐马尔可夫模型的推广, 具有更广的适用范围和更好的效果, 在语言识别和图像处理等领域已经有非常广泛的应用。本报告将介绍条件随机场的模型、算法和我们开发的 R 软件包 CRF, 以及条件随机场在生物信息学领域的应用。

Adaptive False Discovery Rate regression with application in integrative analysis of large-scale genomic data

杨灿 (Hong Kong Baptist University)

时间: 11:00~11:30 邮箱: eeyang@hkbu.edu.hk

简介: Dr. Yang Can's research interests include statistical genomics, bioinformatics, and machine learning. He is particularly interested in developing computationally efficient and statistically rigorous methods to address the challenging problems in the areas of statistical genomics, machine learning and etc. He has made contributions in development of statistical theory, methodology and algorithm, as well as scientific discovery. His research papers have appeared in a number of high-impact journals, including American Journal of Human Genetics, Annals of Statistics, Bioinformatics, IEEE Transactions on Pattern Analysis and Machine Intelligence, PLoS Genetics, and Proceedings of the National Academy of Sciences.

摘要: To address scientific questions, we often design experiments and collect data from experiments. Conventionally, we often focus on the data set at hand and improve analysis results by refining models. The rising of Big Data may change the way of doing research – What if combining our data at hand with other existing information that hides in the Big Data Mountain?

Extending the adjusting-heritable-trait GWAS to bivariate analyse can help identify novel loci

郭小波 (中山大学)

时间: 11:30~12:00 邮箱: mc03gxb@126.com

简介: 郭小波, 副教授, 硕士生导师, 2012 年毕业于中山大学统计科学系, 获理学博士, 2013 年 5 月受聘于中山大学讲师职位, 2017 年 1 月受聘中山大学副教授职位, 2016 年 4 月获聘澳大利亚墨尔本大学荣誉研究员 (Honorary Fellow)。曾于 2011-2012 年在美国耶鲁大学留学, 2013.8, 2015.4 分别在新加坡基因研究所、新加坡眼科研究所访问。主要从事组学数据、双生子数据、复杂医学数据、生物数据的整合与分析。目前已在统计专业著名杂志 Biometrics, Genetics Epidemiology, Statistics in Medicine, 综合性著名杂志 Nature Communications, British Journal of Cancer, Scientific Reports, Oncotarget, Plos One 等发表了学术论文近二十篇, 参与出版了《中华医学统计百科全书 - 遗传统计分册》, 2011 年获广东省统计科研优秀成果一等奖 (排名第二)。主持一项国家自然科学基金青年基金、一项中山大学青年教师培育项目、共同主持一项国际多中心合作项目。参与两项在研的国家自然科学基金重点项目、一项国家自然科学基金重大研究项目。

摘要: In this talk, we consider a large-scale testing problem in genomic data analysis. Recent international projects, such as the Encyclopedia of DNA Elements (ENCODE) project, the Roadmap project and the Genotype-Tissue Expression (GTEx) project, have generated vast amounts of genomic annotation data, e.g., epigenome and transcriptome. There is great demanding of effective statistical approaches to integrate genomic annotations with the results from genome-wide association studies (GWAS). To explore genetic architecture of human complex phenotypes, rather than only relying on GWAS, we introduce Adaptive False Discovery Rate (AdaFDR) regression to integrate genomic annotations with GWAS. For a given phenotype, not only AdaFDR increase the power of mapping its risk variants, but also adaptively incorporates relevant annotations for prioritization of genetic risk variants, allowing nonlinear effects among these annotations, such as interaction effects between genomic features. The developed algorithm is scalable to genome-wide analysis. Using AdaFDR, we performed integrative analysis of genome-wide association studies on human complex phenotypes and genome-wide annotation resources, e.g., Roadmap epigenome. The analysis results revealed interesting regulatory patterns of risk variants, offering new biological insights on genetic architectures of complex phenotypes.

R Usage in Pharmaceutical Industry

Harry Hua (Boehringer Ingelheim)

时间: 8:30~9:00 邮箱: harry.hua@boehringer-ingelheim.com

简介: Dr. Hairui (Harry) Hua is currently Senior Statistician in Boehringer-Ingelheim. He received bachelor degree in Statistics from Fudan University and then went to UK to pursue his master degree and PhD degree in Statistics in University of Bristol and University of Birmingham. In 2015, He worked in Roche UK for more than 1 year and joined BI China since last July till now. He mainly worked on the Phase I to III trials in oncology. His research area includes semiparametric modeling in survival analysis & individual patient meta-analysis.

摘要: While SAS remains an important tool in the pharmaceutical industry, more and more pharma companies are starting to use R as a complimentary tool to streamline their analytic processes. In the drug development stage, R is becoming a popular tool in daily work, for example, statisticians often use R to do scenario simulations for trial design. However, R is still rarely utilized in formal regulatory submissions although no agencies prohibit its use for statistical analysis. A key feature of R is the very large number of user contributed free code packages, however few of these have been fully validated. In the conservative pharma reporting environment, the applicability of user contributed R functions is therefore limited so far. My presentation will address two aspects. First I will introduce an RShiny App for sample size calculation developed by BI statisticians. This internal App supports project teams to determine Go/No-go criteria, i.e. to determine whether they should start Phase III trials based on the Phase Ib/II data. Then I will talk about the process that BI is now working on to validate external R packages (those in addition to base R and its default, recommended, packages). The aim of this process is to identify a group of high quality R packages which could be used for formal clinical reporting.

临床医生眼中的医疗大数据研究：需求和挑战

周健 (北京大学人民医院)

时间: 9:00~9:30 邮箱: zhoujian@bjmu.edu.cn

简介: 周健博士，2012 年毕业于北京大学医学部，获医学博士学位，2012 年至今于北京大学人民医院历任住院医师、主治医师，曾参与国家、企业、医院等多项临床数据库的设计及优化。主要研究方向包括：肺癌流行病学研究及胸外科创新性手术技术开发。

摘要: 一直以来，随机对照试验 (Randomized controlled trial, RCT) 被认为是治疗性研究的金标准。而基于医疗大数据的真实世界研究 (Real world research, RWR) 也受到了广泛的关注，基于医疗大数据的真实世界研究反映了现实医疗的情况，代表着广泛人群的治疗情况，可以代表疾病人群的全貌。近些年来，基于医疗大数据的研究层出不穷，不少企业及医院也致力于医疗大数据的开发应用。从临床医生的角度来看，其对于医疗大数据的需求主要集中在诊断、治疗、随访、科研等方面。而目前各种医疗大数据解决方案仍存在多种挑战：如何建立标准通用的结构化术语集，如何实现非结构化病例的高效结构化，如何打破不同中心、不同数据库之间的信息孤岛、如何轻松实现医生想要的信息检索功能等。

消化道肿瘤基因组学研究进展

吴健民 (北京大学肿瘤医院)

时间: 9:30~10:00 邮箱: wujm@bjmu.edu.cn

简介: 吴健民, 研究员、博士研究生导师, 北京大学肿瘤医院肿瘤生物信息中心主任, 兼信息部副主任。入选 2016 年第十二批北京市“海聚工程”计划。曾担任澳大利亚悉尼 Gravan 医学研究所 PI, 国际肿瘤基因组协作联盟 (ICGC) 胰腺癌项目多组学数据分析负责人, 在大队列的癌症基因组、蛋白组及相关生物信息学研究上有丰富经验。共发表 SCI 论文 35 篇, 他引 1953 次, 单篇最高引用 775 次。作为通讯作者先后在 Nat Methods 和 Nat Rev Cancer (Analysis Article) 等杂志发表研究成果; 合作研究多次在 Nature (4 次) 和 Cell (2014) 等杂志发表。目前致力于综合计算和实验手段, 整合多组学和临床数据深入研究国内高发癌种的发病机制、精准分子分型和个体化治疗。

摘要: 吴健民博士曾担任国际肿瘤基因组协作联盟 (International Cancer Genome Consortium, ICGC) 胰腺癌项目多组学数据分析负责人 (2010-2015)。这里将介绍 ICGC 胰腺癌项目的最新研究进展, 以及 2016 年回到北京大学肿瘤医院后在国内高发癌症之一的胃癌方面的多组学研究情况。

Identifying tissue origin of cancer cells with somatic mutations and copy number alterations

凌少平 (志诺维思)

时间: 10:30~11:00 邮箱: frank.ling@genowis.com

简介: 凌少平博士, 现任志诺维思基因科技有限公司 CEO 兼首席科学家, 自动化学士、信号与信息处理硕士、基因组学博士, 师从著名华人进化遗传学家吴仲义院士。凌少平博士曾任中科院北京基因组研究所生物信息技术主管、计算肿瘤基因组研究组组长, 在肿瘤异质性、肿瘤演化基因组和生物信息学方面具有较深的研究基础, 曾在 Nature Genetics、PNAS、Annual Review of Genetics, Molecular Biology & Evolution 等权威杂志上发表多篇文章。他主导设计的算法已经应用于肝癌 (HCC)、急性白血病 (AML)、侵袭性 NK 细胞白血病 (ANKL)、结直肠癌 (CRC)、垂体瘤、宫颈癌等诸多肿瘤基因组研究工作中。凌少平博士 2015 年曾代表中科院参与“国际肿瘤基因组分析金标准”大赛 (ICGC-TCGA Dream Somatic Mutation Challenge) 并获得点突变分项冠军和结构变异分项亚军。2016 年作为志诺维思首席科学家率公司团队再次参赛, 并获得结构变异分项亚军和点突变分项季军。2016 年领导志诺维思推出“抗癌登月”大数据平台和个人基因组云系统受到张高丽副总理的关注!

摘要: A substantial proportion of cancer cases present with a metastatic tumor and require further testing to determine the primary site; many of these are never fully diagnosed and remain cancer of unknown primary origin (CUP). It has been previously demonstrated that epigenomic variations detected in whole-genome bisulfite sequencing data of plasma cell-free DNA (1-3) can be used to identify its site of origin with limited accuracy. Recently, tissue-specific mutation accumulation pattern were found (4-6). We hypothesized that tissue origin of cancer cells can be identified by genomic variations detected from whole genome/exome sequencing data of tumor cells even plasma cell-free DNA. We presented a kernel machine to identify tissue origin based on somatic single nucleotide variations, copy number alterations and mutational signature from whole genome/exome sequencing 5610 cases across 24 cancer types from TCGA. The model achieved 80% of accuracy (79% of the F1 score) and the 88% of top2 accuracy (88% of the top2 F1 score) with 100 replicates of 5-fold cross-validation.

癌症转录组大数据的可视化与再挖掘

唐泽方 (北京大学)

时间: 11:00~11:30 邮箱: tangzefang@pku.edu.cn

简介: 唐泽方。北京大学生命科学学院 BIOPIC 张泽民组博士三年级研究生。2014 年加入北京大学生命科学学院 BIOPIC 张泽民实验组攻读博士学位, 研究 TCGA 癌症组织大数据与 GTEx 正常组织大数据的整合与数据挖掘。以通讯作者和一作身份在 Bioinformatics 杂志上发表癌症大数据可视化手机 APP GE-mini (gemini.cancer-pku.cn), 以一作身份在 Nucleic Acids Research 杂志上发表癌症大数据分析网站 GEPPIA (gepia.cancer-pku.cn)。目前研究兴趣在于利用 TCGA 、GTEx 大数据进行数据再挖掘。

摘要: 大型的国际项目如 TCGA, GTEx 创造出了大量的转录组数据, 为人们提供了数据挖掘、理解基因功能的机会。而如何能快速获取到这些生物大数据, 从其中能够得到什么有价值的信息, 是人们一直在探索的命题。为了让没有生物信息学背景的研究人员也能够轻易获取、分析生物大数据, 我们通过 R 、Perl 等语言对数据进行处理、可视化, 设计了癌症大数据可视化手机 APP GE-mini (gemini.cancer-pku.cn) 以及癌症大数据分析网站 GEPPIA (gepia.cancer-pku.cn)。研究人员能够通过 GE-mini 和 GEPPIA 来提出问题或是验证假设。我将在报告中介绍它们。

Identification of disease-causing single nucleotide variants in exome sequencing studies

江瑞 (清华大学)

时间: 11:30~12:00 邮箱: ruijiang@tsinghua.edu.cn

简介: 江瑞, 副教授, 博士生导师, 2002 年毕业于清华大学自动化系, 获得工学博士学位。目前任清华大学数据科学研究院医疗健康大数据研究中心副主任。主要研究兴趣包括: 1. 医学影像智能信息处理; 2. 电子病历智能信息处理; 3. 基因组学研究: 非编码调控元件的识别及其目标基因的预测; 4. 遗传学研究: 全基因组遗传变异对特定疾病的影响预测; 5. 多组学研究: 候选基因对特定疾病的影响预测。

摘要: Exome sequencing has been widely used in detecting pathogenic nonsynonymous single nucleotide variants (SNVs) for human inherited diseases. However, traditional statistical genetics methods are ineffective in analyzing exome sequencing data, due to such facts as the large number of sequenced variants, the presence of non-negligible fraction of pathogenic rare variants or de novo mutations, and the limited size of affected and normal populations. Here, we propose bioinformatics approaches, SPRING, snvForest and GLINTS, for identifying pathogenic nonsynonymous SNVs for a given query disease. SPRING integrates six functional effect scores calculated by existing methods and five association scores derived from a variety of genomic data sources to calculate the statistical significance that an SNV is causative for a query disease. snvForest adopts an ensemble learning method to assign prediction scores to candidate SNVs. These methods are designed to use with a set of seed genes known as associated with the disease of interest, and thus is suitable for studies on diseases with some prior knowledge. GLINTS further incorporates three disease phenotype similarity data to facilitate the detection of causative SNVs without any knowledge of seed genes for a query disease. This method is therefore suitable for research on diseases whose genetic bases are completely unknown. With a series of comprehensive validation experiments, we demonstrate the effectiveness of these methods, not only in simulation studies, but also in detecting causative de novo mutations for autism, epileptic encephalopathies and intellectual disability.

车联网时空数据挖掘与洞察

侯志伟 (北京车网互联科技有限公司)

时间: 8:30~9:00 邮箱: houzhiwei@che08.com

简介: 侯志伟, 数据分析师, 专注于车联网时空数据分析、挖掘及其可视化。曾多次获得数学建模国家一等奖, 且均为交通方向。已获得专业领域内发明专利 3 项, 发表中文核心期刊论文 1 篇。擅长领域: 时空数据挖掘、用户画像系统、智能优化算法, Spark 高性能计算等。

摘要: 车联网作为物联网的先行者、自动驾驶的必由之路, 业已开始步入蓬勃发展期, 海量的多源异构的数据随之而生, 这其中尤以时空轨迹数据为盛。如何挖掘如此大规模的数据金矿, 并洞察背后的价值, 这一问题在如今的数据时代显得极为迫切。本次演讲主要分享车网互联在车联网领域数据的认知和经验, 围绕以下三个业务核心进行介绍: 事件识别, 行为评价, 用户画像。

摩拜单车的数据科学实践

朱俊辉 (摩拜单车)

时间: 9:00~9:30 邮箱: harryzhu@mobike.com

简介: 朱俊辉, 摩拜单车算法工程师, 熟悉 R 语言和 Python, 专注于供应链量化和可重复性研究。

摘要: 摩拜单车在最近的一年里发展飞速, 许多实际问题亟待通过数据驱动的方法去解决。本次演讲将主要从供应链优化的角度, 谈一谈在运营效率的提升方面, 摩拜数据科学应用的现状以及对策。

互联网汽车数据服务分享

李晔彤 (斑马网络)

时间: 9:30~10:00 邮箱: liyetong@saicmotor.com

简介: 李晔彤, 斑马网络数据挖掘工程师, 从事车辆轨迹, 硬件, 车主数据分析与挖掘。毕业于西安交通大学和伦敦政治经济学院, 应用数学专业。

摘要: 介绍斑马互联网汽车的数据应用, 包括轨迹 poi 分析, 驾驶行为分析, 硬件使用分析和用户使用分析。通过介绍专车识别, 驾驶评分, 油耗预测等业务模型, 分享建模工作中的心得。

机器学习在滴滴

王犇 (滴滴出行)

时间: 10:30~11:00 邮箱: benwang177@gmail.com

简介: 现任滴滴大数据 -顺风车策略团队负责人, 负责顺风车分单调度、拼车、信任值、定价、画像、智能补贴等相关算法策略的迭代优化; 曾任腾讯微博 & 腾讯新闻数据挖掘 & 推荐系统负责人; 曾任 58 集团 - 数据智能部负责人; 个人兴趣在于利用大数据 & 大规模机器学习方法持续改进业务和产品体验。

摘要: 每天滴滴出行平台产生海量出行数据, 而滴滴正利用这些数据不断建立各种机器学习模型来优化线上产品体验, 从分单到定价, 滴滴的机器学习和传统互联网公司的推荐广告算法的差异很大, 这次分享会介绍滴滴平台典型的机器学习应用, 进一步会介绍在顺风车场景如何利用机器学习来构建更加智能理性的大数据运营引擎.

汽车消费的数字化决策

张翔 (车轮互联)

时间: 11:00~11:30 邮箱: birdzhangxiang@gmail.com

简介: 汽车盒子数据科学家, 车轮互联数据副总裁。

摘要: 在移动互联网时代, 多屏媒体, O2O 多维互动, 给消费者购物带来了更多信息和更多选择。也给了企业更丰富, 更有挑战的营销环境。在众多影响决策的微时刻 (micro-moment) 和关键时刻 (moment of truth) 中, 汽车消费者的思维已经不自觉的进入了“车型鄙视链”的精神世界和换车魔力象限的领域。利用车轮查违章, 车轮社区 (覆盖 2 亿真实车主的 APP 应用) 中用户对车型 PK 投票的数据, 我们真实再现了这个车型鄙视链, 从中会发现每一款车, 你都可以找到选择他的理由。这为更加细分, 更加个性化的汽车市场提供了理论支撑。以此报告希望能够协助用户选到最适合自己的车, 也协助车厂在细分市场更加精准的定位, 甚至可以预测未来的汽车销量。

基于 R 语言的汽车驾驶行为数据分析

赵帅 (中国汽车技术研究中心数据资源中心)

时间: 11:30~12:00 邮箱: zhaoshuai@catarc.ac.cn

简介: 吉林大学车辆工程专业硕士学位, 中国汽车工程学会 (SAE-China) 会员、中国计算机学会 (CCF) 会员。曾任汽车仿真与控制国家重点实验室研究员, 现任中国汽车技术研究中心数据资源中心数据技术部部长助理, 全面负责数据建设及挖掘工作。

从事研究领域包括车辆数据集成、机器学习、深度学习等, 擅长基于 R 语言、MATLAB 的算法模型开发。个人曾获北美大学生数据建模竞赛一等奖、全国研究生数学建模竞赛一等奖, 并多次在天池大数据算法大赛中获奖。

摘要: 报告主要介绍了汽车驾驶员驾驶行为数据分析的思路和结果。本例的驾驶行为数据主要采集自车辆的 CAN 总线与陀螺仪数据, 报告首先介绍了数据的预处理方案, 包括数据的滤波方法及坐标转换方法, 然后介绍了常规类驾驶行为与特殊驾驶行为的识别算法, 最后介绍了驾驶行为的统计结果以及对车辆性能的预估。本报告所涉及的数据处理、数据分析基本都使用 R 语言进行, 相关的 R 包括 ggplot2、dtw、corrplot、sqldf 等。

Banded Spatio-Temporal Autoregressions with Application to Forecasting PM2.5

马莹莹 (北航)

时间: 8:30~9:00 邮箱: mayingying_11@163.com

简介: 北京航空航天大学经管学院助理教授, 研究方法为社交网络数据分析, 高维数据分析, 付费搜索广告营销。

摘要: We propose a new class of spatio-temporal models with unknown and banded autoregressive coefficient matrices. The setting represents a sparse structure for high dimensional spatial panel dynamic models when panel members represent economic (or other type) individuals at many different locations. The structure is practically meaningful when the order of panel members is arranged appropriately. Note that the implied autocovariance metrics are unlikely to be banded, and therefore, the proposal is radically different from the existing literature on the inference for high-dimensional banded covariance matrices. Due to the innate endogeneity, we apply the least squares method based on a Yule-Walker equation to estimating autoregressive matrices. A ratio-based method for determining the bandwidth of autoregressive matrices is also proposed. Some asymptotic properties of the inference methods are established. The proposed methodology is further illustrated using both simulated and real data sets.

On a vector double autoregressive model

张兴发 (广州大学)

时间: 9:00~9:30 邮箱: xingfazhang@hotmail.com

简介: 广州大学经济与统计学院统计系副教授, 副系主任。研究方向: 时间序列分析。

摘要: Motivated by the double autoregressive (DAR) model, in this talk, we study a vector double autoregressive model (VDAR). The model is a straightforward extension from univariate case to multivariate case. Sufficient ergodicity conditions are given for the model. Without existence of second moment conditions for observed time series, the quasi maximum likelihood estimator (QMLE) of the parameter in the model is shown to be asymptotically normal, which does not hold for classic vector autoregressive (VAR) model with i.i.d errors. Simulation results confirm that our estimators perform well. A given empirical study implies the proposed model has potential applications in practice. Keywords: Vector double autoregressive model, quasi maximum likelihood estimation

Prediction Interval for Autoregressive Time Series via Oracally Efficient Estimation of Multi-Step Ahead Innovation Distribution Function

顾莉洁 (苏州大学)

时间: 9:30~10:00 邮箱: gulijie@suda.edu.cn

简介: 我是苏州大学数学科学学院的一名教师, 主要研究方向是非参数与半参数统计方法, 主要研究兴趣是抽样调查、时间序列及函数型数据的统计推断。

摘要: Kernel distribution estimator (KDE) is proposed for multi-step ahead prediction error distribution of autoregressive time series, based on prediction residuals. Under general assumptions, the KDE is proved to be oracally efficient as the infeasible KDE and the empirical cdf based on unobserved prediction errors. Quantile estimator is obtained from the oracally efficient KDE and prediction interval for multi-step ahead future observation is constructed using the estimated quantiles and shown to achieve asymptotically the nominal confidence levels. Simulation examples corroborate the asymptotic theory.”

Simultaneous conficence bands for mean and variance function based on deterministic design

蔡利 (苏州大学)

时间: 10:30~11:00 邮箱: caili16@126.com

简介: 苏州大学数学与科学学院在读博士二年级学生。

摘要: Asymptotically correct simultaneous confidence bands (SCBs) are proposed for the mean and variance functions of nonparametric regression model based on deterministic designs. The variance estimation is as efficient up to order $n^{-1/2}$ as an infeasible estimator if the mean function were known. Simulation experiments provide strong evidence that corroborates the asymptotic theory. The proposed SCBs are used to analyze two sets of strata pressure from the Bullianta Coal Mine in Erdos City, Inner Mongolia, China.

A smooth simultaneous confidence band for correlation curve

张园园 (苏州大学)

时间: 11:00~11:30 邮箱: zhangyuany2014@163.com

简介: 苏州大学 2014 级研究生, 主要研究方向为非参数与半参数统计推断, 函数型数据分析, 时间序列分析, 并对大规模机器学习研究感兴趣。

摘要: A smooth simultaneous confidence band (SCB) is proposed for a local measure of variance explained by regression, termed correlation curve in Doksum et al. (1994), based on local quadratic estimation. The proposed estimator of correlation curve is oracally efficient in the sense that it is as efficient as an infeasible correlation estimator with the variance function known. Simulated and real-data examples are provided to illustrate the usefulness of the proposed oracle SCB.

自由讨论

时间: 11:30~12:00

人工智能颠覆量化投资

郭健 (深度资产管理有限公司)

时间: 14:00~14:30 邮箱: bayesso@163.com

简介: 郭健博士是深度资产管理有限公司创始人, 致力于打造世界领先的人工智能对冲基金。郭健教授曾在美国哈佛大学任教, 从事机器学习、大数据挖掘、复杂网络分析、高维统计学等领域的研究, 并担任微软、谷歌、雅虎等公司在人工智能方面的技术顾问。郭健还担任一系列国际一流统计学和机器学习期刊的审稿人。郭健博士本科毕业于清华大学数学科学系, 之后获得美国密歇根大学统计学博士学位, 其博士期间的研究工作多次获得美国统计学会、国际运筹与管理学会、国际生物统计学会颁发的最佳学生论文奖。

摘要: 人工智能和大数据的发展, 正在颠覆传统金融投资和交易行业。本报告系统介绍人工智能、机器学习和统计学模型如何改变传统量化投资。我们将通过对谷歌 AlphaGo 人工智能系统的深入剖析, 展开对人工智能金融交易的介绍, 并对其发展前景进行预测。我们还讨论如何将人工智能技术与传统量化投资模型相结合, 以提升模型收益, 降低交易风险。

数据驱动人工智能的实践

丁磊 (百度公司)

时间: 14:30~15:00 邮箱: ding@shuju.io

简介: 丁磊博士是百度金融首席数据科学家, 曾任职 PayPal 全球消费者数据科学部负责人, 通过一系列的人工智能和个性化产品大幅度提升了全球电商和支付用户的消费体验。丁博士曾在哥伦比亚大学和 IBM Watson 研究院工作, 在人工智能和大规模机器学习等领域有丰富的成果。丁博士曾在斯坦福大学学习管理。

摘要: 如果说数据是原油, 那么人工智能就是从原油中提炼各种高价值产品的加工厂。丁博士将结合十多年在零售、金融、广告等行业开发人工智能产品的实践, 分享他关于人工智能技术在商业领域的深度思考。

量化投资简介

王鑫 (NA)

时间: 15:00~15:30 邮箱: uxeverest@qq.com

简介: 清华大学物理系学士、中国科学院理论物理博士, 美国莱斯大学物理与天文系博士后, 历任杭州某量化对冲基金资深基金经理、基金管理部总监、量化分析师。在数学建模相关领域具有相当造诣, 擅长结合物理实验建模, 精通分子动力学模拟, 擅长各类算法及海量数据存储技术。多年大型高性能分布式并行计算处理经验, 擅长复杂数据分析和大数据挖掘。

摘要: 讲解量化投资特点、分类等方面的基本概念, 并对主要投资策略的理论基础、适用条件、应用特点等角度做进一步阐述, 在此基础上介绍量化投资中常用的风险衡量、业绩评估指标等。

论机器学习在金融领域的应用

张卓 (卓识投资有限公司)

时间: 16:00~16:30 邮箱: zhangzhuo Jason@163.com

简介: 清华大学电子工程系本科生, 曾获得清华大学最高荣誉特等奖学金。博士就读于普林斯顿大学电子工程系, 从事机器学习人工智能方面的研究。博士毕业后曾就职于华尔街最大的做市商骑士资本, 其独立开发的策略日均交易额达 10 亿美金。现回国创立卓识投资, 任总经理, 负责开发期货 CTA 和股票阿尔法策略。

摘要: 近年来, 量化交易已经在国内二级市场得到了充分发展, 人们越来越认识到量化模型对于风险控制的重要性。而机器学习也慢慢渗透进入这个领域, 顶尖的技术人才开始试图用更复杂的非线性模型来解释略显神秘的金融市场的不确定性。

R 语言与量化投资实战

任坤 (上海明法投资)

时间: 16:30~17:00 邮箱: renkun@outlook.com

简介: 上海明法投资资深基金经理, 主要从事股票量化对冲、期货量化策略的研发。编写了 `formattable`、`rlist` 等扩展包, 是《Learning R Programming》的作者。

摘要: 量化投资是用数量化的方法, 基于历史数据, 发现、分析和验证投资逻辑, 而 R 语言则是量化研究的重要利器。从股票到期货和其他金融衍生品, 丰富的数据一方面为多样化的投资逻辑提供了可能, 另一方面也为投研人员带来了一些挑战。该演讲从量化投资所涉及的数据处理方面的挑战入手, 从数据操作、高性能计算等方面介绍 R 语言和相关扩展包如何提升量化投资研究的生产力。

CTA 投资思路与常用 R 包

霍志骥 (私募)

时间: 17:00~17:30 邮箱: huo.zhiji@qq.com

简介: 中国人民大学统计学 2012 级本科生。量化研究员, 主要从事期货量化策略的研发, 四年衍生品投资经历。曾任私募主观交易员, 从事期货交易, 后转入量化研究的领域。有一系列错误与正确的投资经验。

摘要: 量化投资是在传统的投资思路上, 运用了数据验证与量化的工具, 能够极大的增强策略的可靠性, 可复制性, 可解释性, 并提高开发的效率, 降低研发的成本。而 CTA 类型的策略, 则是量化被广泛运用的领域之一。该演讲将结合个人从主观交易员转向量化研究的经历, 介绍 R 中常用的包与 CTA 策略构建的一些经验。

嵌入式上的深度学习初探

张先轶 (澄峰科技)

时间: 14:00~14:30 邮箱: xianyi@perfxbal.com

简介: 张先轶, PerfXLab 澄峰科技创始人, 中科院博士, 曾先后于美国得州大学奥斯汀分校, 麻省理工学院进行博士后研究工作, 主要研究方向为矩阵计算, 高性能计算, 性能优化等。全球领先的开源矩阵计算项目 OpenBLAS 发起人与维护者, 获得 2016 年中国计算机学会科学技术二等奖。

摘要: 嵌入式系统的深度学习已经成为主要趋势之一。将模型的 Inference 直接在嵌入式设备本地运行, 除了本身模型不能过于复杂外, 还需要深度学习框架与底层优化库的配合。本报告讲介绍我们团队在这方面的工作, 包括底层库的优化, 框架精简, 以及模型压缩等。

Exploring Heterogeneous Algorithms for Accelerating Deep Convolutional Neural Networks on FPGAs

肖倾城 (商汤)

时间: 14:30~15:00 邮箱: walkershaw@foxmail.com

简介: 北京大学高能效计算与应用中心研究生, 商汤集团 FPGA 研发实习生。

摘要: Convolutional neural network (CNN) finds applications in a variety of computer vision applications ranging from object recognition and detection to scene understanding owing to its exceptional accuracy. There exist different algorithms for CNNs computation. In this patent, we explore conventional convolution algorithm with a faster algorithm using Winograd's minimal filtering theory for efficient FPGA implementation. Distinct from the conventional convolution algorithm, Winograd algorithm uses less computing resources but puts more pressure on the memory bandwidth. We first propose a fusion architecture that can fuse multiple layers naturally in CNNs, reusing the intermediate data. Based on this fusion architecture, we explore heterogeneous algorithms to maximize the throughput of a CNN. We design an optimal algorithm to determine the fusion and algorithm strategy for each layer. We also develop an automated toolchain to ease the mapping from Caffe model to FPGA bitstream using Vivado HLS.

Pluto: A Distributed Heterogeneous Deep Learning Framework

杨军 (阿里巴巴)

时间: 15:00~15:30 邮箱: yangjunpro@gmail.com

简介: 目前在阿里云 iDST 大规模算法团队负责大规模深度学习基础设施相关建设工作, 对大规模分布式机器学习的开发、建设以及在不同业务场景中的落地应用有较为深入的理解和认识。之前先后在奇虎 360 担当广告技术部门架构师, Yahoo! 北京研发中心担当效果广告系统技术负责人。

摘要: 本分享会介绍阿里云 iDST PAI 团队研发的一款分布式深度学习框架 Pluto。在 Pluto 里, 阿里云 PAI 团队基于 Caffe 和 TensorFlow 这两款开源框架进行了分布式性能的深度优化定制, 相较于优化前取得了

显著的性能提升, 在一些场景下取得了 10X 的收敛加速比提升。并成功应用到了集团安全、金融风险建模、证件类图片识别、客服问答、机器翻译等集团核心业务建模场景里, 显著提升了建模迭代效率。

Evaluating Fast Algorithms for Convolutional Neural Networks on FPGAs

卢丽强 (*sensetime*)

时间: 16:00~16:30 邮箱: *liqianglu@pku.edu.cn*

简介: 北京大学信科 13 级本科生高能效计算与应用中心 sensetime FPGA 研发实习生

摘要: In recent years, Convolutional Neural Networks (CNNs) have become widely adopted for computer vision tasks. FPGAs have been adequately explored as a promising hardware accelerator for CNNs due to its high performance, energy efficiency, and reconfigurability. However, prior FPGA solutions based on the conventional convolutional algorithm is often bounded by the computational capability of FPGAs (e.g., the number of DSPs). In this paper, we demonstrate that fast Winograd algorithm can dramatically reduce the arithmetic complexity, and improve the performance of CNNs on FPGAs. We first propose a novel architecture for implementing Winograd algorithm on FPGAs. Our design employs line buffer to effectively reuse the feature map data among different tiles. We also effectively pipeline the Winograd PE engine and initiate multiple PEs through parallelization. Meanwhile, there exists a complex design space to explore. We propose an analytical model to predict the resource usage and reason about the performance. Then, we use the model to guide a fast design space exploration. Experiments using the state-of-the-art CNNs demonstrate the best performance and energy efficiency on FPGAs. We achieve an average 785.1 GOP/s for the convolutional layers and 749.4 GOP/s for the overall AlexNet and an average 2653.4 GOP/s for the convolutional layers and 2272.6 GOP/s for the overall VGG16 on Xilinx ZCU102 platform.

Elastic Stack 与机器学习

曾勇 (*Elastic*)

时间: 16:30~17:00 邮箱: *medcl@elastic.co*

简介: 曾勇 (Medcl), Elastic 工程师与布道师, 2015 年加入 Elastic 公司, 在加入 Elastic 之前, 在搜索和运维等方面积累了超过七年的经验。Elasticsearch 国内首批用户, 自 2010 年起就开始接触 Elasticsearch, 是 Elasticsearch 中文社区的发起人, 同时也是 Elastic 在中国的首位员工。

摘要: 目前常规的分析手段往往只关注到了总体的趋势, 而忽略了异常的个体行为, 比如在海量的信用卡订单中, 我们可以通过统计可以知道总体的交易量、每笔交易、平均交易时间等等, 但是如何快速识别和定位其中存在盗刷可能的交易同样重要, 而通过机器学习, 您就可以在海量的订单数据中发现异常的数据, 定位异常的单笔交易行为。另外传统的机器学习往往需要经历较长的离线处理, 如果能够实时的对数据进行基于机器学习的分析将大大改善提升数据分析的能力和改善客户体验, 本次分享将主要介绍如何基于 ElasticStack 及 X-Pack 提供的机器学习能力来实现的实时行为分析。

利用 css 对 shiny 页面优化及利用 htmlwidgets 包创建 HTML 控件

谢佳标 (乐逗游戏)

时间: 14:00~14:30 邮箱: jiabiao1602@163.com

简介: 乐逗游戏高级数据分析师, 负责大数据挖掘及可视化。资深 R 语言用户, 有九年以上数据挖掘工作实战经验, 多次在中国 R 语言大会上作主题演讲。合著过《R 语言与数据挖掘》、《数据实践之美》, 新书《R 语言游戏数据分析与挖掘》也即将上市。

摘要: 本演讲将介绍如何利用 CSS 对 shiny 页面进行个性化设计及在网页中嵌入视频; 并通过一个详细案例介绍了利用 htmlwidgets 包开发 HTML 控件, 基于 D3.JS 库创建简单的交互桑基图, 包括控件创建、函数修改、数据调用及与 shiny 结合的演示。

Persistent Reproducible Reporting with Docker and R

肖楠 (Seven Bridges Genomics)

时间: 14:30~15:00 邮箱: me@nanx.me

简介: Nan is a Genomic Data Scientist at Seven Bridges, where he and his colleagues build innovative data-driven products for petabyte-scale biomedical data analysis, accelerating breakthroughs in genomics research for cancer, drug development, and precision medicine. With the help of Docker, their scalable, cloud-based Seven Bridges Platform empowers rapid, collaborative analysis of millions of genomes in concert with other forms of biomedical data. As an active contributor to the R community, Nan is the author of 10+ R/Bioconductor packages covering topics of machine learning, reproducible research, and data visualization.

摘要: Automatic report generation has a massive number of use cases for reproducible research and commercial applications. Fortunately, most of the problems involved in this topic have been elegantly solved by knitr and the R Markdown specification for the R community. However, the issues on data persistence and operating system-level reproducibility were rarely considered in the context of reproducible report generation. Today, such issues have become a major concern in the current software implementations. In this talk, we will discuss potential approaches to tackle such problems, particularly with the help of modern containerization technologies. We will also demonstrate how to compose a persistent and reproducible R Markdown report with the help of the two R packages we developed: docker-r and liftr. Specifically, you will learn to dockerize your existing R Markdown documents, how to apply it to the analysis of petabyte-scale cancer genomics data on the Cancer Genomics Cloud, and how to distribute or reuse such containerized reports.

Learning R Internals and C++ via Rcpp

任乾 (深圳谷雨科技)

时间: 15:00~15:30 邮箱: enqian@outlook.com

简介: 量化工程师, 主要方向为股票、期货策略。

摘要: In the realm of high performance computing with R, users might take a learning path from R, Rcpp to some R internals. However, each one of the three parts can be challenging without a proper understanding of the other two. This lecture attempts to share my experience and viewpoint with those who have similar interests in gaining better understanding of how R works behind the scene while advancing their C++ skills.

跟踪 R 社区动态 - R Weekly 的背后

覃文峰 (*R Weekly*)

时间: 16:00~16:30 邮箱: *wenfeng.qin@qq.com*

简介: R Weekly 创始人之一

摘要: RWeekly.org 搭建了一个一站式的信息平台, 通过网站, 邮件, 新浪微博 @rweekly 等渠道, 实时地向来自 140 多个国家的读者推送社区的最新动态。每周的资讯速递帮助 R 用户快速地掌握社区一周内的最新进展。近年来, R 社区发展迅速, CRAN 现在已有 10000+ 的程序包。学会发现, 学习和使用现有的基础资源, 掌握社区的最佳实践, 可以节省时间、减少重复的轮子。这个讲座将会介绍 R Weekly 的一些有趣的发现以及背后的故事。

自由讨论

时间: 16:30~17:00

油气长输管道数据分析实践

刘晨 (中国石油规划总院)

时间: 14:00~14:30 邮箱: 390900285@qq.com

简介: 中国石油规划总院管道信息部副主任, 从事大型企业数据分析、ERP、信息规划 10 余年经历。

摘要: 经过十余年的信息化建设, 中国石油已形成以 ERP 为核心, 以生产系统为支撑, 以传感器数据为基础的信息化架构。依托生产与管理数据, 对压缩机生产与能耗情况进行分析预测, 实时掌握核心生产设备运行情况, 并对工况进行预测, 寻找最佳运行模式, 降低运行成本。

工程数据分析方法在半导体制造过程监测中的应用

张玺 (北京大学)

时间: 14:30~15:00 邮箱: xi.zhang@pku.edu.cn

简介: 从事数据融合和质量工程的科研与教学, 研究兴趣集中在对复杂工程和服务系统的过程监测、诊断、控制与优化。

摘要: 随着半导体器件关键尺寸的不断减小、集成度的不断提高和晶圆直径的不断增大, 半导体制造过程变得越来越复杂, 对半导体制造装备及其自动化水平要求越来越高。各种传感器技术的发展也同时给制造过程监测带来了前所未有的契机。本次演讲主要围绕过程中各种传感器数据, 利用一些工程数据分析方法来实现对生产过程的有效监测, 从而最终达到质量提升的目的。

工业大数据在风电行业的应用

张光磊 (金风科技)

时间: 15:00~15:30 邮箱: zhangguanglei@goldwind.com.cn

简介: 清华大学自动化系控制理论方向博士, 曾就职于理光软件研究所, 目前在金风科技数据分析部门从事工业大数据分析方面的研发工作。

摘要: 风电行业作为清洁能源的首选已经有几十年的发展历史, 积累了大量的数据和经验, 如何利用大数据分析在其他行业的成功经验应用于风电行业将是十分有意义的事情。

电子制造业智能化的挑战与机遇

王逢春 (台达电子)

时间: 16:00~16:30 邮箱: fengchun.wang@deltauw.com

简介: 王逢春博士于 2014 年 12 月加入台达电子, 目前是台达电子技术长办公室 (CTO Office) 解决方案总监, 负责智能制造, 电动车充电桩运营等工业物联网领域整体解决方案的设计、开发和实施。在此之前, 王

博士在 IBM 中国研究院工作，在行业知识资产化，业务转型方法与技术，服务创新，智慧城市战略规划，食品安全及风险分析解决方案等领域担任高级研究员工作，并于 2006 至 2007 年度担任 IBM 中国研究院院长助理的工作。王博士于 2004 年毕业于重庆大学并拥有工业工程博士学位。

摘要：制造业已经走过或者正在经历自动化，信息化的浪潮，而制造业的未来属于智能制造。如何才能做到智能化，工业大数据分析是其中的关键要素。本演讲以电子制造业龙头企业台达电子自身对智能制造的设计，对智能制造的布局出发，分享了工业企业大数据分析提升良率上的探索，并提出了对大数据分析技术的问题和挑战。

制造即服务，数据即价值

陈宸（三一集团）

时间：16:30~17:00 邮箱：sanygroup@foxmail.com

简介：本科数学，硕博模式识别，现任三一集团数据科学家。

摘要：制造即服务，数据即价值。

工业大数据分析：实践与挑战

田春华（NA）

时间：17:00~17:30 邮箱：tianchunhua@k2data.com.cn

简介：昆仑数据首席数据科学家。2004 年 1 月清华大学自动化系博士毕业。2004 年 -2015 年在 IBM 中国研究院，负责数据挖掘算法研究和产品工作，在高端装备制造、石油石化、新能源、航空与港口等行业，帮助中国、亚太、欧美领先企业，成功实施资产管理、运营优化、营销洞察等各类数据分析项目。发表学术论文（长文）82 篇（第一作者 42 篇），拥有 36 项专利申请（10 项已授权），研究兴趣是数据挖掘算法与应用。

摘要：通过 9 个行业案例分析，归纳出工业大数据分析与经典商业数据分析的区别，并尝试归纳总结出工业大数据分析 3 段建设方法论，和 6 种分析范式，最后展望工业大数据为分析算法研究带来的机遇。

Rsubread: an efficient toolkit for mapping and counting short sequencing reads

Yang Liao (the Walter and Eliza Hall Institute)

时间: 14:00~14:30 邮箱: liao@wehi.edu.au

简介: Yang Liao is a postdoctoral researcher in the Bioinformatics division of the Walter and Eliza Hall Institute of Medical Research (WEHI). He is the co-author of the Subread and Rsubread package for genomic analysis. With his computer science background, Yang Liao's research interests focus on high performance computing in Bioinformatics, including highly efficient read mapping, quantification and downstream analysis.

摘要: Read mapping and quantification tools play a critical role in many genetic analysis pipelines that take high-throughput sequencing data as input. The accuracy and sensitivity of the read mapping tool directly determine the validity and quality of the outcomes from downstream analysis. More importantly, the very large (and continuously increasing) amount of data generated in high-throughput sequencing brings on the needs to highly efficient tools for read mapping and quantification.

From reads to genes to pathways: differential expression analysis of RNA-Seq experiments in Bioconductor

Yunshun Chen (the Walter and Eliza Hall Institute)

时间: 14:30~15:00 邮箱: yuchen@wehi.edu.au

简介: Yunshun (Andy) Chen is a Postdoctoral Research Fellow in the Bioinformatics Division at the Walter and Eliza Hall Institute (WEHI) of Medical Research. His research mainly focuses on differential gene expression of the next-generation sequencing data. He is one of the authors and the main maintainer of the edgeR package - the arguably world's most popular R package specifically designed for count-based sequencing data. His other research interests include DNA methylation, alternative splicing, microRNA and single cell RNA-Seq data.

摘要: In recent years, RNA sequencing (RNA-seq) has become a very widely used technology for profiling gene expression. One of the most common aims of RNA-seq profiling is to identify genes or molecular pathways that are differentially expressed (DE) between two or more biological conditions. Changes in expression can then be associated with differences in biology, providing avenues for further investigation into potential mechanisms of action.

Glimma: getting greater graphics for your genes

Charity Law (the Walter and Eliza Hall Institute)

时间: 15:00~15:30 邮箱: law@wehi.edu.au

简介: Charity Law is a statistical bioinformatician whose work focuses predominantly on gene expression analyses of high-throughput data. The impact of her work is best illustrated by the popularity of voom , a method for RNA-seq gene expression analysis that she developed which has been cited 602 times since its publication in 2014 (Source: Google Scholar). She currently holds the position of senior research officer in the Molecular Medicine Division at Walter and Eliza Hall Institute of Medical Research, Australia. In addition to differential gene expression, her research interests include differential isoform usage, transcript expression, and histone modification analyses.

摘要: RNA-sequencing is a popular technology used by scientists to study changes in gene expression levels across tens of thousands of genes simultaneously. Representing gene expression levels, the counts in each sample are typically analysed by categorising samples into groups of interest, and obtaining gene-wise summary statistics in the form of log-fold changes, t-statistics, p-values, and the like. The data and its results can be explored by plotting one summary statistic against another and highlighting genes that are significant or of interest. The new Bioconductor package, Glimma, generates interactive graphics for plots typically found in the limma package with the enhanced feature of connecting many levels of information within the analysis on a single html page using d3.js. A Glimma-style mean-difference plot, or the more generic xy-plot, allows one to click on the points to bring up a new plot of sample-wise expression levels that is displayed alongside the original plot. This feature enables researchers to interrogate the data more intensely than ever before without the need to repeat the work for every gene under examination. The plots include options to search and select for genes of interest, and zoom in and out for better resolution. Unlike the traditional multi-dimensional scaling (MDS) plot, Glimma's MDS plot shows several dimensions and group combinations on the same page. The functions within Glimma are tailored to integrate smoothly with objects native to limma, edgeR and DESeq2, and can be extended for use with microarray, single-cell and methylation data analyses.

Deconvolving human and viral RNA in RNA sequencing data

Alexandra Garnham (the Walter and Eliza Hall Institute)

时间: 16:00~16:30 邮箱: garnham.a@wehi.edu.au

简介: Alexandra Garnham (PhD) is the head of the Bioinformatic Support Unit at the Walter and Eliza Hall Institute of Medical Research. Her work focuses on the analysis of high-throughput sequencing data as well as biostatistical analysis. Her research interests include gene expression and regulation, data visualisation and dimension reduction. She is also a member of the R Ladies organisation whose aim is to promote gender diversity in the R community.

摘要: It is estimated that 15-20% of all human cancers are associated with viral infections. Viruses can influence various stages of the oncogenic process, however discovering the biological significance of their contribution can be challenging. The prevalence of a virus with a particular cancer can range from 15-100%. An option in determining the abundance of viral presence in a tumour sample would be to perform RNA sequencing on the tumour. We have developed a pipeline utilizing the Rsubread Bioconductor package that enables us to deconvolve viral RNA from human, thereby allowing us to detect and quantify the presence of viruses. We demonstrate this pipeline using RNA sequencing data from human Head and Neck Squamous Cell Carcinomas (HNSC) acquired from The Cancer Genome Atlas.

自由讨论

时间: 16:30~17:00

中文文本分析方便工具包 `chinese.misc` 介绍

吴江 (清华大学)

时间: 14:00~14:30 邮箱: husserlhusserl@sina.com

简介: 清华大学社会科学学院博士后, 主要研究方向为社会科学方法论、量化分析、政治传播。

摘要: 尽管现在文本挖掘技术发展迅速, 各种新技术和新工具不断出现, 但用 R 语言进行中文文本分析的人, 特别是初学者, 还时常在如何读取文件并避免乱码、如何分词、如何统计词频这样的问题上遇到困难。`chinese.misc` 包尝试缓解这一问题。该 R 包的功能非常实用, 主要用于对中文文本进行数据清理工作, 此外还包含另外一些常用的处理和分析功能。在生成文档 -词语矩阵的功能上, 可以代替对中文不是太支持的 `tm` 包。此外, 在读取文件、去除停用词、描述性分析等方面, 该包在封装既有函数的基础上提供了更为方便和灵活的形式。

ezdf: 用户友好的标签数据框

陈华珊 (中国社科院社会发展战略研究院)

时间: 14:30~15:00 邮箱: chenhs@cass.org.cn

简介: 副研究员

摘要: ‘ezdf’ 包的目的是使 R 支持类似 SPSS 或 Stata 那样对用户友好的标签输出。‘ezdf’ 包并不是要定义一套新的制表函数, 而是控制相关制表函数 (如 ‘pander’) 在输出时, 能够自动带上对应的标签。除此之外, ‘ezdf’ 也封装了几个常用的制表方法。

众所周知, 在 R 的体系当中, 并无变量标签或者数值标签的定义。对于类别变量, 在 R 中使用 ‘factor’ 类型可起到部分标签的功能。对于变量标签, 在 ‘data.frame’ 中尽管可以直接使用标签来命名变量, 例如 ‘df\$ 年龄’, 但是实际使用中多有不便。

在 R 中导入 SPSS 或 Stata 等传统统计软件的数据格式可有多个包来实现, 例如 ‘foreign’、‘readStata13’、‘haven’、‘sas7bdat’ 等等。这些包在导入数据时, 都能保持原数据中所定义的标签。然而所有这些包目前来说各有优缺点, 即使对同一个格式也做不到支持各个版本的导入, 因此难以提供一揽子解决方案。更重要的, 各个包导入数据之后所定义的标签属性各不相同, 导致对标签的使用难以统一。更不用说, 在制作表格或者统计结果输出时, 能够让 R 做到标签友好。

Latent Variable Modeling for Cognitive Assessment Through Second-Order Exponential Family

刘京辰 (*Columbia University*)

时间: 15:00~15:30 邮箱: jcliu@stat.columbia.edu

简介: Jingchen Liu is Associate Professor in Statistics at Columbia University. He holds a Ph.D. in Statistics from Harvard University. He is the recipient of 2013 Tweedie New Researcher Award given by the Institute of Mathematical Statistics and a recipient of the 2009 Best Publication in Applied Probability Award

given by the INFORMS Applied Probability Society. He has research interests in statistics, applied probability, Monte Carlo methods, and psychometrics.

摘要: Latent variable models are popular in the analysis of marketing, e-commerce, social network, and many other fields where human behaviors are observed and are summarized to a few characteristics. In this talk, I discuss a framework for latent variable models through a low-rank second-order exponential family. In this framework, the computational overhead is substantially reduced, which is crucial especially for nonlinear models and big data analysis. It is also convenient to incorporate additional graphical structures and other covariates. An R package is developed. I will illustrate the model and the package through several real data examples.

法律的定量分析及其实践

邵兴全 (中豪律师集团)

时间: 16:00~16:30 邮箱: 316373595@qq.com

简介: 受过法律与经济学系统教育, 具有丰富的司法实践经验, 致力于法律的量化分析。

摘要: 一直以来, 法学被归入社会科学的范畴, 主要采用定性及案例分析的方法展开研究。但随着法律经济学在英美国家的兴起, 以统计为基础的研究方法, 越来越多被用于法学研究与司法实践。在我国, 司法判例被不断地公布, 对其进行定量分析已具备初步基础, 而今, 无论是理论界与司法实务部门, 都在积极采用大数据改进我们对司法系统的认识。本次演讲围绕法律的定量分析与隐私权保护展开, 结合民商事、刑事等案件, 展示如何对其进行定量分析, 并得出有意义的结论。另外, 本次演讲也会探讨大数据时代的隐私权保护问题。

再抽样法分析夫妻般配与家庭工资不平等

李代 (北京大学)

时间: 16:30~17:00 邮箱: lidaipku@163.com

简介: 北京大学社会学系博士研究生

摘要: 近年来, 关于同型婚配的研究在社会学界得到越来越多的关注。本文采用 LHSC1996 与 CGSS2012 两个截面数据, 首先用对数线性模型测量调查数据显示的夫妻教育匹配的同型程度, 用相关系数测量工资收入上的相似程度。然后使用模糊置换检验, 估量在控制年龄、城乡和地域之后教育程度上存在的匹配对家庭工资收入不平等指数泰尔指数 (Theil Index) 影响, 并通过比较两个截面数据考察其变化趋势。

自由讨论

时间: 17:00~17:30

微启的旋转门：大数据教育界与工业界的生态进化

赵鹏（看准数据招聘集团）

时间：14:00~14:30 邮箱：changmeng@kanzhun.com

简介：看准数据招聘集团创始人，资深人力资源和品牌营销专家。他创立于 2013 年 12 月的看准数据招聘集团旗下现拥有看准网、BOSS 直聘和店长直聘三个子品牌，总服务用户数超过 1 亿。其中，于 2014 年 7 月上线的 BOSS 直聘，在全球首创求职者与招聘方的“在线直聊模式”，并通过大数据技术实现人才与岗位的精准推荐，已成为中国移动互联网招聘领域单款最大 APP，截至 2017 年 4 月底，拥有注册求职者 1950 万，注册 Boss 403 万。赵鹏先生在人力资源科技领域有 12 年经验，曾担任智联招聘 CEO，期间带领连续亏损 13 年的公司扭亏为盈。赵鹏先生是中国大学生志愿者西部支教计划的缔造者之一，服务于青年志愿者事业十年。赵鹏先生 1994 年毕业于北京大学法律系。

摘要：大数据时代中，技术变革和人才稀缺成为两个核心话题。作为服务了数十万公司的移动互联网招聘平台，BOSS 直聘积累了海量人才大数据，清楚地看到当下企业应用大数据技术，与人才断层下的几个困境。构建自我进化的生态系统，或许能够成为解决问题的一个方法。

大数据学科建设的关键因素

欧高炎（北京大数据研究院）

时间：14:30~15:00 邮箱：gaoyano@boyabigdata.cn

简介：欧高炎，北京大学理学博士，博雅大数据学院院长。全球首家大数据教育、竞赛和服务平台“数据嗨客”创始人。中国人民银行征信中心《大数据新算法用于信用模型构建的效果评估》项目组负责人。

摘要：通过介绍博雅大数据学院在数据科学相关专业建设、大数据教育实训平台建设方面的经验，探讨大数据人才培养的模式，交流大数据教育和学科建设的经验。

数字金融 - 实验室项目模拟系统——银行数据仓储，数据测试，数据安全三位一体的就业驱动项目训练系统平台

李扬（文思海汇）

时间：15:00~15:30 邮箱：[yang.li6@pactera.com](mailto.yang.li6@pactera.com)

简介：目前就职于全球 IT 服务提供商 Pactera 文思海辉，担任技术总监。12 年研发经验，曾负责金融 & 电力系统 ERP 的架构设计和研发工作。8 年 IT 教育培训与校企合作经验，主持设计集团人才培养顶层架构，聚焦于机器学习技术研究与实训平台的研发工作。

摘要：教育部《关于“十三五”期间全面深入推荐教育信息化工作的指导意见》提出“信息化已成为国家战略，教育信息化正迎来重大历史发展机遇”。文思海辉以就业为驱动的金融训练系统平台，整合了集团数字金融解决方案的银行数据仓储工坊系统、数据测试 ATQ 管理系统与数据安全攻防产品系统，形成三位一体的实验室项目训练系统，并为校企合作专业共建和教学实验室建设提供了应用层、测试层、信息安全层，提供了一个完整的培养体系解决方案。

大数据历史长河中的统计思维与智慧

刘乐平 (天津财经大学)

时间: 16:00~16:30 邮箱: liulp66@163.com

简介: 2003 年博士毕业于中国人民大学统计系。愿与统计之都和狗熊会的小伙伴们一起共创中国统计新纪元。

摘要: 统计是动态的历史, 历史是静态的统计。如果大数据是海洋, 那么统计学定是汇入这海洋的主干河流之一。纵向梳理公元前至今统计历史长河中的年代大事, 横向比较数据统计分析的重要科学发现, 探究人类科学广场上雄伟的统计智慧殿堂。

大数据教育平台的建设与探索

袁星星 (北京大数据研究院)

时间: 16:30~17:00 邮箱: xingxingy@boyabigdata.cn

简介: 北京大数据研究院博雅大数据学院产品经理; 大数据教育实训平台数据嗨客产品负责人。

摘要: 当前在线教育市场方兴未艾, 作为细分领域的数据教育更是处于摸索发展阶段。《大数据教育实训平台的建设与探索》结合大数据教育的现状, 剖析行业痛点, 从教、学两大主题, 通过介绍北京大数据研究院博雅大数据学院产品数据嗨客, 交流在大数据教育实训平台建设上的实践经验。

自由讨论

时间: 17:00~17:30

FACTOR AND RESIDUAL EMPIRICAL PROCESSES

王江艳 (南京审计大学)

时间: 14:00~14:30 邮箱: wangjiangyan2007@126.com

简介: 王江艳, 理学博士, 2016 年毕业于苏州大学概率论与数理统计专业, 现为南京审计大学理学院统计科学与大数据研究院讲师。

摘要: The distributions of the factor return and specific error for an individual variable are important in forecasting and applications. However, they are not identified with low-dimensional time series observations. Using the recently developed theory for large-dimensional approximate factor model for large panel data, the factor return and specific error can be estimated consistently. Based on the estimated factor returns and residual errors, we construct the empirical processes for estimation of the distribution functions of the factor return and specific error, respectively. We prove that the two empirical processes are oracle

efficient when $p \log CT^{3/2}$ where p and T are the dimensionality and sample size, respectively. This demonstrates that the factor and residual empirical processes behave as well as the empirical processes pretending that the factor returns and specific errors for an individual variable are directly observable. Based on this oracle property, we construct the simultaneous confidence bands

for the distributions of the factor return and specific error. Extensive simulation studies check that the estimated bands have good coverage probabilities. Our real data analysis shows that the factor return distribution has a structural change during the crisis in 2008.

Free-knot spline for Generalized Regression Models

王静 (University of Illinois at Chicago)

时间: 14:30~15:00 邮箱: jiwang12@uic.edu

简介: Jing Wang is currently an associate professor in Statistics at University of Illinois at Chicago. Her main research area is in application and theory of kernel and spline smoothing methods in semi-parametric and non-parametric regression models.

摘要: A computational study of bootstrap confidence bands based on free-knot spline technique is explored for generalized regression models,

Spatially Varying Coefficient Models

王冠男 (College of William & Mary)

时间: 15:00~15:30 邮箱: gwang01@wm.edu

简介: I am GuanNan Wang. I graduated with a PhD in Statistics from the University of Georgia, U.S. in 2015. Since then, I joined the faculty group of department of Mathematics at College of William & Mary.

摘要: In this paper, we study the estimation of spatially varying coefficient models for data distributed over complex domains. We use bivariate splines over triangulations to represent the coefficient functions. A

convergence rate for the bivariate spline estimators is derived. The estimators of the coefficient functions are consistent, and we establish the rate of convergence of the proposed estimators. A penalized least squares method is proposed to estimate the model with a penalization term. We also propose hypothesis tests to examine if the coefficient function is really varying over space. The proposed method is computational expedient, thus usable for analyzing massive datasets. The performance of the estimators and the proposed tests are evaluated by several simulation examples and a real data analysis.

Quantile Regression Oultier Diagnostic: R package ‘quokar’

王文静 (中国人民大学)

时间: 16:00~16:30 邮箱: wenjingwang1990@ruc.edu.cn

简介: 本人是中国人民大学统计学院博士二年级学生, 从 2016 年 9 月到 2017 年 3 月在澳大利亚莫纳什大学联合培养, 期间师从 Dianne Cook 教授开发 R 语言包 quokar, 目的在于做分位回归中的异常值检验。

摘要: Extensive toolbox for estimation and inference about quantile regression has been developed in the past decades. Recently tools for quantile regression model diagnostic are studied by researchers. We built R package ‘quokar’ to implement outlier diagnostic methods in R language. This talk offers a brief tutorial introduction to this package. Package ‘quokar’ is open-source and can be freely downloaded from Github: <http://www.github.com/wenjingwang/quokar>. To move one step further, we also plot the diagnostic model into data space to observe how does the model performs using R package ‘rggobi’.

哪种奇巧巧克力最好吃: Statistical ranking models 及其 R 实现

曹明 (University of Texas School of Public Health)

时间: 16:30~17:00 邮箱: ming.cao@outlook.com

简介: 曹明即将于 2017 年夏从 University of Texas School of Public Health 生物统计系博士毕业, 主要研究方向是 (社交) 网络的统计模型和估计 (statistical network analysis), 统计软件开发 (statistical software development) 和专为概率设计的编程语言 (probabilistic programming)。本科是同济大学的软件工程, 去美国读博之前做过两年码工。

摘要: 排序 (ranking) 是一种普遍的需求, google 出来排在最前面的几个结果 (PageRank) 是否就是你想要的? 上赛季的金州勇士队常规赛创纪录的 73 胜却没有赢下最终的总冠军, 他们的“真实实力”到底是不是第一呢? 我们就从 sports analytics 里常用的 Bradley-Terry model 说起, 以最近 John Hopkins 一个十分有趣的项目: 哪种奇巧巧克力 (Kitkat) 最好吃为例, 谈谈 ranking 的统计模型, 以及相关的几个 R package。

自由讨论

时间: 17:00~17:30

G2 - 面向统计的可视化语法

萧庆 (蚂蚁金服)

时间: 8:30~9:00 邮箱: xiaoqing.dongxq@alipay.com

简介: 萧庆, 蚂蚁金服数据可视化团队技术专家, G2 的架构师和核心研发, 多年从事数据可视化研发, 对可视化相关图形, 统计, 图形语法有深入的思考和丰富的项目经验

摘要: G2 是一套基于 (The Grammar Of Graphics) 的图形语法, 以数据为驱动, 具有高度的易用性和扩展性, 内置常用的统计函数, 配备坐标系、度量、辅助元素等组件, 用户无需关注各种繁琐的实现细节, 一条语句即可构建出各种各样的可交互的统计图表。G2 始于图形语法, 打开数据可视化的无限可能。

WebGL 在前端数据可视化中的应用

沈毅 (百度)

时间: 9:00~9:30 邮箱: shenyi.914@gmail.com

简介: 2012 年浙江大学专业毕业后一直在百度做前端开发, 目前为百度资深研发工程师, 主要从事 ECharts 的研发。个人研究领域有二维, 三维的前端图形绘制, 数据可视化等。视觉系, 对游戏开发, 程序生成设计, 绘画等感兴趣

摘要: 我们在前端可视化库 ECharts 中选择了 Canvas 作为底层的绘图接口, ECharts 目前在 GitHub 上拥有 16k 的 star, 其拥有丰富绚丽的可视化效果, 深度全面的交互操作, 以及对大数据量稳定高效的展现等特性, 是 GitHub 上以及国内最热门的开源前端可视化库之一, 能够实现这些特性主要得益于 Canvas 的强大能力以及我们基于 Canvas 封装的二维图形库 ZRender 对图形操作的便捷性。

但是随着可视化形式的拓展, Canvas 在某些展现方式上也显得力不从心, 例如:

1. 用 Canvas 绘制几十万的图形依然有压力
2. 各种三维数据的展现需求, 以及大屏需要的一些酷炫的三维效果
3. 用 JS 计算布局存在的性能瓶颈, 需要通过一些新的思路去有进一步的提升

本次分享主要介绍我们是如何在现在的 ECharts 中集成 WebGL 去解决这些问题的, 包括:

1. ECharts 简介以及目前的情况
2. 用 WebGL 绘制地球, 三维的散点图, Surface 等三维图表
3. 用 WebGL 加速散点图等常见二维图表的绘制
4. 以及利用 GPGPU 进行一些布局运算的性能优化

Interaction+: “让可视化动起来”的既有网页交互

陆昊 (北京大学)

时间: 9:30~10:00 邮箱: lumin.vis@gmail.com

简介: 陆昊, 现就读于北京大学信息科学技术学院, 师从可视化与可视分析实验室袁晓如研究员。研究方向为时空数据的可视化与可视分析、人机交互, 有多篇关于城市轨迹数据可视分析工作发表于可视化领域顶级期刊与重要会议。

摘要: “让可视化动起来”的网页交互工具 Interaction+, 它能让你不写一行代码就与网页上的各种各样的可视化进行交互。这些可视化可以是像纽约时报上那样的数字媒体新闻、博客财报中的各式各样的统计图表, 也可以是 d3.js 编写的千奇百怪的可视化工作、艺术家制作的漂亮的信息图等。Interaction+ 的核心思想是在已有的可视化作品的制作流程之外, 将交互的对象从常规的数据转移到视觉图元, 支持用户在认知过程中对视觉图元的整理、查询、过滤等任务。具体而言, Interaction+ 从已有的网页可视化作品中获得其视觉图元及属性信息, 并将其作为数据驱动, 提供一套基础而完整的交互功能, 让用户能在原可视化中进行选择、过滤、查询、比较、打标签等交互。

运用 WebGL+GIS 开发网络安全应用

黄伟 (360)

时间: 10:30~11:00 邮箱: huangwei01@b.360.cn

简介: 黄伟, 现就职于 360 企业安全, 从事前端开发与可视化研发工作, 发表多篇可视化相关论文和专利, 近期研究领域为多维数据可视化与可视分析, 时空数据可视化, 可视化数据挖掘和 BI。

摘要: 随着互联网技术的发展, 网络几乎渗入到了人们工作和生活的各个方面, 在给人们生活带来方便的同时, 网络攻击和网络犯罪也随之产生。近年来, 网络攻击的数量越来越多, 规模越来越大, 攻击复杂度也越来越高, 传统的网络安全保障机制也越来越不足以应对。网络安全可视化应运而生, 并成为网络安全研究领域的一个热点。本次分享将围绕 360 企业安全天眼团队利用 GIS 和 WebGL 技术在网络安全可视化领域的实践经验, 具体包括: 1) APT 先知计划: 以“上帝视角”将攻击信息及 POI 信息在地图上显示出来, 让客户感知安全态势, 帮助企业和组织及时作出应对策略; 2) 伪基站追踪系统: 结合可视化和数据挖掘揭示伪基站出现在城市中出现的模式及发送短信的规律, 并能够实时显示伪基站的位置, 有效帮助执法机关打击不法分子; 3) 全国僵木蠕毒态势感知: 结合 GeoHash 技术和 Openlayers 以热力、蜂窝聚合显示全国僵木蠕毒数据; 4) 春运铁路网热度: 利用 WebGL 可视化春运时期全国铁路线路及各线路热度; 5) 大图可视化引擎: 针对海量数据可视化中布局及绘制的难点, 开发大规模网络关系布局算法库在服务端进行布局, 前端利用 WebGL 技术显示数百万点边图。”

SAS 统计图表: 一键式的图表生成术

谷鸿秋 (北京天坛医院)

时间: 11:00~11:30 邮箱: guhongqiu@yeah.net

简介: 著名医学院非著名毕业生; 土鳖博士, 野生码农, 科研搬砖工; 微信公众号「统技思维」出品人; 人大经济论坛/SAS 中文论坛卸任版主; 目前主要从事流行病学、公共卫生、临床试验、实效研究等临床研究领域的统计学设计和统计分析方法研究。

摘要: 一直以来, 统计表格的常规制作模式是: 1. 运行统计软件 (菜单/代码) 2. 设计统计表格 (Word/Excel) 3. 复制粘贴填充结果 (Ctl+C/V)。这种操作模式不仅低效, 而且容易出错, 也不利用重复性研究。在总结了大量医学研究学术期刊后, 笔者提炼出了学术期刊中最基本的 9 种统计表格, 借助 SAS 的宏程序和 ODS 系统, 开发了一套可以一键式制作统计表格的工具。。SAS 作为一款优秀的统计分析软件, 其统计绘图功能却一直被大众所诟病, 绘图语法也为大众所畏惧。在大众的印象中, SAS 的统计绘图功能太难学, 画出的图形太丑, 太死板。其实 SAS 公司一直在改进 SAS 的绘图功能, 自 SAS 9.2 引入 ODS Graphics System 后, SAS 绘图的语法变得更加简洁, 实现也更加方便, 终极绘图武器绘图模板语言 (Graph Template

Language, GTL) 更是让 SAS 的绘图功力大大增加。ODS Graph 设计器使得不会编程, 不懂 GTL 语言的人也能迅速画出 ODS Graph。分享本次分享将围绕 SAS 统计表格输出系统的开发过程、原理、构成以及使用, 以及 SAS 的绘图系统, ODS 绘图系统的构成, 各种统计图形的绘制举例, 统计图形的美化等内容展开, 并对 SAS sgplot 与 R ggplot2 的做简要对比。

自由讨论

时间: 11:30~12:00

工业大数据的应用

季春霖（深圳光启高等理工研究院）

时间：8:30~9:00 邮箱：chaofeng.kou@kuang-chi.com

简介：季春霖，深圳光启高等理工研究院联合创始人，副院长；深圳市统计学会副会长；哈佛大学博士后，杜克大学博士，剑桥大学硕士；广东省自然科学基金杰青项目获得者；发表包括 Science 在内的论文 60 余篇，授权专利 470 余项。

摘要：本文主要介绍了统计方法和计算在新材料开发特别是超材料开发中的应用。在超材料开发中，利用统计学、信息学方法，通过数据挖掘探寻材料结构与性能之间的关系模式，缩短材料开发周期，降低开发成本。通过仿真与实验的有机结合，建立超材料基因数据库，结合物理学、人工智能、大数据、材料学，通过统计方法预测超材料的复杂电磁响应，实现新材料定制化。另外，介绍了统计方法在发动机设计诊断等工业问题中的应用。

数据智能实践 –从互联网到传统行业

沈志勇（百度）

时间：9:00~9:30 邮箱：shenzhiyong@baidu.com

简介：沈志勇博士，百度云首席数据科学家。大数据分析技术国家工程实验室学术委员会成员，大数据流通与交易技术国家工程实验室专家委员会成员。本科毕业于北大数学学院概率统计专业，随后于中科院软件所获得博士学位。曾任百度大数据实验室副主任，惠普中国研究院研究员。

摘要：行业的发展的需求，正从信息化，慢慢的转为数据驱动，最近开始追求智能化。随着各行业相继完成信息化进程，数据在各行业内逐渐形成积累，数据驱动的决策与业务优化的需求越来越明确，在某些数据驱动应用充分的行业，如金融，又在此基础上开始追求业务的智能化。互联网行业较早的经历和完成了从数据驱动到智能化的演化，在这里介绍一些相关的案例。

大数据时代背景下设备安全管理与智能制造

陈宏（郑州恩普特科技股份有限公司）

时间：9:30~10:00 邮箱：chenhong@zotp.cn

简介：工学博士（后）、副教授，教授级高工，研究生导师，郑州大学振动工程研究所所长，郑州恩普特科技股份有限公司总工程师。中国振动工程学会故障诊断专业委员会理事，故障诊断专业委员会常务理事，河南省振动工程学会秘书长，河南省机械工程学会青年工作委员会副秘书长。主要研究方向为设备智能诊断与智能管理。主持或参与 6 项国家级科研项目、10 余项省市级科技攻关项目，以及多项企业横向合作项目的研究与开发工作。获得省部级科技奖励 3 项，已鉴定验收科研成果 8 项，授权发明专利 5 项，在国内外学术刊物及会议上发表研究论文 74 篇，其中 EI/SCI/STP 检索十余篇。

摘要：1. 工业 4.0 与中国制造 2025

2. 设备安全管理与智能制造

3. 当前存在的问题

-
- 4. 应对措施与解决方案
 - 5. 未来的发展需求与展望

数控机床大数据分析

田野（杭州数途科技信息有限公司）

时间：10:30~11:00 邮箱：Eric.tian@inrevo.io

简介：现任 Inrevo 杭州数途信息科技有限公司创始人。先后在纽约知名互联网企业服务企业 Register.com, NetworkSolutions, Web.com, Travelclick.com 等纳斯达克上市企业担任公首席架构师, 全球网络运营资深运营经理, 世界 500 强企业雅芳集团全球网络与存储云计算转型总监, 亚马逊全球网络实施高级经理等职务。现从事于工业物联网技术和工业大数据处理的研发, 所创 Inrevo 公司以 I-IoT 驱动精益生产和设备管理理念为核心, 基于大数据技术实现对工业生产过程中数据进行实时精准的分析处理, 对工厂运行的各个环节实现透明化数字化管理, 为持续改进, 业务决策提供关键的大数据技术支撑。

摘要：节能减排是缓解工业能源消耗的主要途径, 数途科技基于大数据技术, 重点研究工业智能制造中的 CNC 数控机床智能生产。以精益生产为核心, 对机床生产过程产生的实时数据开展分析。通过顶层设计, 实现智能机床数据的云端整合。通过设计高维分布式大数据分析算法, 实现智能机床生成的能量消耗实时预测。最终实现对工厂运行从排期到生成的全流程监控, 全过程优化。以最低能耗实现产品精准化生产, 真正达到工业生成的智能制造。该技术服务于数控机床零件加工企业, 在冲压、注塑, 过程制造, 离散化生产线等领域都拥有巨大的应用场景。

自由讨论

时间：11:00~11:30

高维数据中的模型诊断及其在商务统计中的应用

张耀武 (杉数科技)

时间: 8:30~9:00 邮箱: yaowu@shanshu.ai

简介: 上海财经大学统计学博士, 研究方向为大数据挖掘, 高维数据等。杉数科技高级算法工程师, 京东定价与库存项目和顺丰件量预测项目核心成员。

摘要: 通常大家认为, 模型预测准确性越高, 越有利于做决策。但现实中, 却存在精度高的模型导致错误决策的现象。因此, 盲目追求模型精度是不科学的。事实上, 作为一家决策型公司, 我们发现模型的正确性才是至关重要的, 即使预测精度不是很高。我们在这里介绍一种普遍使用的模型诊断方法, 利用提供的数据, 检验当前使用的模型 (可以是非参数或者半参数的) 是否适用。这是数据分析和决策中很关键的一步, 而恰恰容易被忽略。

从统计学生到互金数据科学家之路

徐旦 (读秒)

时间: 9:00~9:30 邮箱: dan.xu@idumiao.com

简介: 2011 届中山大学统计学士, 2013 届美国哥伦比亚大学精算硕士。毕业后先后就职于美国最大银行之一 Capital One, 美国新兴金融科技公司 ZestFinance。2016 年回国, 现就职于读秒 www.idumiao.com (原积木盒子零售事业部)。

摘要: 针对尚未走出校门的统计学生, 介绍数据科学家在实际工作中的工作方式、理念和工作内容, 包括一些常用的 R 包和函数。并将结合一个完整的从业务目标分析, 到接入外部数据源, 最后到模型部署上线的实际案例进行演示。

机器学习在营销管理中的应用

吴岸城 (菱歌科技)

时间: 9:30~10:00 邮箱: wuanch@gmail.com

简介: 吴岸城, 毕业于浙江大学计算机系, 目前研究方向在对话、视觉、推荐领域。有 13 年企业级软件服务与电信增值业务软件研发经验。曾在中兴、亚信担任研发管理人员, 现任菱歌科技首席数据科学家。

摘要: 本次报告主要关于 TURank 在影响力算法的进化和应用, 以及 CNN 在图像场景识别的应用。

人工智能颠覆客服行业的实践

刘应耀 (阿里巴巴集团)

时间: 10:30~11:00 邮箱: yingyao.lyy@alibaba-inc.com

简介: 刘应耀, 花名阿外, 阿里巴巴高级技术专家, 隶属智能创新中心。十多年数据产品及算法平台研发经验, 最近几年聚焦大数据和人工智能领域, 孵化智能机器人、人工智能辅助系统、知识图谱等系列产品, 开创阿里客服的智能时代。并通过智能硬件实验室等企业内创业模式, 试点前台机器人等智能硬件。

摘要: 阿里客服体系, 每天 5 万热线电话, 30 万人在线咨询问题, 服务成本高、服务效率低且用户体验差。我们基于语音识别、语义理解、个性化推荐、客户模型等技术, 构建一整套完整的智能服务解决方案。包括陪伴用户的智能助理式机器人、人工咨询的智能推荐系统、交易纠纷的智能决策系统等等, 用数据和智能提升用户体验, 降低服务成本。报告会先演示智能机器人产品系列功能, 包括语音识别、咨询服务、生活助理等, 然后介绍客服整体智能解决方案、技术架构、关键算法、以及数据驱动理念。

新能源行业 R 语言数据分析实例

陈卓 (CCF 中国计算机学会)

时间: 11:00~11:30 邮箱: 15210607169@163.com

简介: 陈卓, CCF 中国计算机学会会员, ACM 国际计算机协会会员, CIPS 中国中文信息学会会员, 曾任北京东润环能科技有限公司高级算法工程师、产品经理, 金风科技“金风慧能大数据平台”数据分析主管, 主要工作方向为新能源风电行业的数据分析与数据发掘。

张锐, 中科院大气物理所大气科学和地球流体力学数值模拟国家重点实验室博士, 北方大贤风电科技(北京)有限公司高级顾问, 长期从事新能源与气象行业的学术研究与工程化推广工作, 主要研究方向为新能源行业的气象预测、功率预测与数字化应用。

摘要: 随着可再生能源行业的快速发展, 越来越多的从业主体希望数据分析能够在实际业务中发挥更多的作用。R 语言作为一种易于上手的开源语言, 成为了许多从业者的首选分析工具。演讲将以新能源行业的基本特点入手, 通过环境特征分析、设备性能分析、运营管理分析、行业信息与舆情分析四个业务方向, 介绍新能源领域数据分析的重要价值与工作方法, 最后分享基于公有云 R 语言分析平台的建设经验, 展示 R 语言并行处理海量数据的思路与方法。

自由讨论

时间: 11:30~12:00

人工智能助力线上消费金融的风险管理

叶伟(同盾科技)

时间: 8:30~9:00 邮箱: *data_mine@163.com*

简介: 同盾科技数据部信贷建模总监, 大数据风控、信贷建模领域十多年工作经验。

摘要: 基于同盾大数据平台采用人工智能技术为消费金融商户提供全流程风险管理解决方案。

以风险资本收益率驱动决策

叶梦舟(融360)

时间: 9:00~9:30 邮箱: *yemengzhou@rong360.com*

简介: 2016年11月加入融360担任首席风控官; 在中国和美国商业银行有近二十年的风控领域专业经验, 先后服务于中国光大银行、美国摩根大通银行和花旗银行; 曾任花旗银行北美执行董事, 兼任花旗北美CRO的首席战略顾问; 南京大学经济学学士、美国印第安纳大学MBA。

摘要: 1. 巴塞尔监管资本概述 2. 如何用模型确定监管资本需求 3. 如何衡量产品/项目的资本收益率并服务于决策 4. 监管资本要求对银行业务的冲击和影响, 以及未来发展趋势。

金融科技中的算法与可视化应用案例

张云松(北京天启智创信息技术有限公司)

时间: 9:30~10:00 邮箱: *stevenzys@hotmail.com*

简介: 张云松毕业于中科院, 多年零售金融行业咨询和互联网公司从业经历, 专注数据算法、决策分析、风险管理及金融产品设计的工作。曾就职于互联网金融企业融360, 负责运营在线授信的小额现金贷产品, 目前在金融科技方向进行创业。一直致力于数据科学应用于零售金融的业务流程, 通过数据化决策等金融科技方式辅助金融业务中业务中获客、转化、反欺诈、风险定价、审批授信、贷后催收等。

摘要: 随着机器学习算法和AI的普及应用, 更多的算法和数据科学在金融业务领域得到实践并发挥了重要作用, 同时互联网在更快更深入的变革传统金融业务, 在金融业务中对金融科技的需求越来越迫切。本次分享结合目前互联网金融中的实际业务场景和需求, 介绍算法和各种数据科学技术如何在决策模型、反欺诈、审批授信、风险定价等实际业务中应用的案例。

消费金融反欺诈应用探索

肖勃飞(成都柠檬时光科技有限公司)

时间: 10:30~11:00 邮箱: *xiaobofei@163.com*

简介: 曾担任中国最大的金融 IT 服务商东南融通 BI 数据挖掘团队负责人, GE 智能医疗研发负责人, 京东风控决策支持部负责人, 四方伟业首席科学家和大数据产品部总经理; 在 CRM 系统、精准营销、信贷风控、智能医疗、电商交易风控、反欺诈行业应用方面具有丰富的经验, 同时在大数据和数据挖掘的产品化方向经验丰富。

摘要: 1, 反欺诈产品综述; 2, 介绍图查询引擎设计; 3, 介绍反欺诈模型; 4, 介绍反欺诈可视化产品

自由讨论

时间: 11:00~11:30

患者表征学习方法与应用

李响 (IBM)

时间: 8:30~9:00 邮箱: lixiang@cn.ibm.com

简介: 李响, IBM 中国研究院资深研究员, 毕业于浙江大学计算机学院并获得博士学位。现主要从事认知计算和医疗大数据分析的研究, 在疾病风险预测、治疗路径挖掘、病历信息抽取等方面进行了若干研究工作, 发表医学信息学和人工智能顶级论文 20 余篇。

摘要: 真实世界的电子病历 (EMR) 数据存在高维性、时序性和稀疏性的特点。从电子病历数据中提取患者的表征 (representation) 是进行疾病风险预测、患者表型分群等数据分析的重点和难点。传统的医疗数据分析通常采用基于向量的患者表征, 对高维、时序和稀疏数据的处理存在一定的问题。为此, 近年来研究者们提出了多种新的患者表征学习方法, 包括基于时序模式的表征、基于主题模型的表征、基于张量的表征、以及基于深度模型的表征等, 这些方法能够从高维、时序和稀疏的电子病历数据中提取出更为有效的患者表征。

基于电子病历的高通量表型标记

俞声 (清华大学)

时间: 9:00~9:30 邮箱: syu@tsinghua.edu.cn

简介: Dr. Yu Sheng is Assistant Professor at the Center for Statistical Science of Tsinghua University. He received his BS and MA degrees in statistics from Nankai University and the University of Michigan, and he received his PhD degree in systems engineering (operations research) from the George Washington University. He started his research in medical informatics since his research work at Harvard University. His current research interests include natural language processing and deep learning in medicine, and data analysis and knowledge extraction from EHR and online data.

摘要: 电子病历自诞生之日起就被视为潜在的医学数据挖掘和知识提取的宝库。电子病历中记录的诊疗细节包含了对患者各种表型的描述, 这些表型可以被提取出来, 丰富现有生物样本库中的表型信息, 或根据表型自动建立新的大型队列, 助推各种各样的生物医学研究。然而, 由于电子病历中语意的复杂性, 准确提取病历中的表型信息并不简单。通常, 需要使用机器学习模型, 综合诸多专家设计的特征, 拟合人工标注的金标准, 形成表型标记算法。这样的算法生成过程耗时长、消耗人力巨大, 无法满足大数据时代对电子病历挖掘的需求。2012 年底, 医学信息学界提出要实现“高通量”表型标记, 旨在去除生成算法过程中的一切人工因素。本报告介绍我们在高通量表型标记方向的一系列创新研究: AFEP、SAFE、以及世界上第一个高精度高通量表型标记算法生成技术 PheNorm。

医学临床中的人工智能技术

黄正行 (浙江大学)

时间: 9:30~10:00 邮箱: zhengxinghuang@zju.edu.cn

简介: 先后负责了国家自然科学基金面上项目一项、青年基金项目一项、博士后特别资助和一等资助各一项, 作为研究骨干参加国家 863 项目和工信部重大专项项目各 1 项, 参加国际合作项目 2 项。在国际、国内

重要学术期刊和国际医学信息学核心学术会议上, 以第一作者或通讯作者身份共发表学术论文 30 多篇。所发表的论文被国内外的其他研究组引用次数 500 多次。申报专利 2 项, 获得软件著作权 3 项。被国际学术期刊 Artificial Intelligence in Medicine 聘为编委, 担任 IEEE International Conference on Healthcare Informatics 等医学信息学领域国际核心学术会议的 Committee member。近年来, 致力于医学大数据挖掘与应用研究。努力创新, 注重将研究成果与临床实践相结合。前期工作中, 围绕着如何有效挖掘和利用医学大数据, 分析和优化心血管等典型复杂性疾病的临床诊疗实践这一医学和信息科学的交叉研究课题, 在(1)面向诊疗过程实践的临床路径分析与优化; 和(2)临床诊疗实践中的疾病风险评估与不良事件预测, 等研究方向上取得了若干的研究成果。

摘要: 医疗信息化的规模正以前所未有的速度增长, 医疗卫生领域已进入“人工智能时代”。医疗人工智能的研究与应用对提升医药卫生服务水平、促进医疗产业发展等方面发挥着至关重要的作用。本报告在对医学临床人工智能技术发展历史进行回顾的基础上, 重点阐述了医学临床人工智能的研究现状、应用领域、总结展望等。

“AI+ 慢性病管理”使精准医疗成为可能

金博 (大连理工大学)

时间: 10:30~11:00 邮箱: jinbo@dlut.edu.cn

简介: 金博, 大连理工大学副教授。致力于数据挖掘、大数据分析、创新管理、商务智能等领域的科学研究。主持了国家自然科学基金青年项目、辽宁省高校科研项目、国家重点实验室开放课题等课题, 参与科技部国家重点研发计划“精准医疗研究”项目、国家自然科学基金重大研究计划培育项目和面上项目、863 计划项目等国家级课题。在相关领域重要国际期刊及会议上发表论文 60 余篇, 近年来多篇论文在数据挖掘领域顶级期刊 (KDD、AAAI、ICDM、SDM、PAKDD 等) 收录, 担任数据挖掘领域三大顶级会议 KDD、ICDM、SDM 的程序委员, 是 ACM、IEEE 和 CCF 高级会员。

摘要: 调查显示, 慢性病及其并发症的急性发作已成为威胁我国老年患者健康的最主要因素。以帕金森症、阿兹海默症等神经系统慢性退行性疾病为研究对象, 针对临床医学研究中的慢性病并发症评估、药品不良反应预测、联合用药推荐等难题, 采用机器学习和医疗大数据分析的方法, 在前期积累的海量医疗数据基础上, 构建人工智能 + 慢性病管理的模式, 以数据为驱动, 使精准医疗成为可能, 为提高我国医疗信息服务水平、合理利用医疗资源、探索新的慢性病并发症个性化治疗模式提供理论与实践支撑。

自由讨论

时间: 11:00~11:30

新 AI 时代的智能问答

王厚峰 (北京大学)

时间: 8:30~9:00 邮箱: wanghf@pku.edu.cn

简介: 王厚峰, 北京大学计算语言学研究所所长。研究兴趣为自然语言处理, 近年来, 主要集中于问答系统和情感分析的研究。曾作为首席专家负责过 863 项目, 作为首席专家负责过国家社科基金重大项目。发表学术论文 70 余篇。

摘要: 问答系统是人工智能和自然语言处理中广受关注的问题, 是自动客服、人机对话、自动阅读理解、高级检索等应用中最核心的内容。随着新一轮人工智能热的兴起, 自动问答更是受到了前所未有的重视, 这不仅表现在学术界, 而且也体现在工业界。然而, 问答系统仍然存在大量尚未解决问题, 包括问题的理解和答案的形成。报告简要介绍了新一轮人工智能的兴起过程及问答系统的发展状况, 分析了问答系统中主要的难点以及我们在开展的相关研究工作。

自然语言处理中的统计结构学习

孙薇薇 (北京大学)

时间: 9:00~9:30 邮箱: ws@pku.edu.cn

简介: 孙薇薇, 女, 1983 年 2 月出生, 博士, 副教授, 北京大学计算机科学技术研究所语言计算与互联网挖掘研究组成员。研究方向为: 计算语言学/自然语言处理, 研究子领域为深层自然语言理解、句法分析、组合语义分析等。2002 年考入北京大学中国语言文学系, 2006 年 7 月获得“应用语言学”专业文学学士学位。自 2003 年 9 月起选修北京大学信息科学技术学院开设的“软件工程”双学位, 于 2006 年 7 月获得理学第二学士学位。2006 年 9 月考入北京大学信息科学技术学院, 于计算语言学研究所攻读硕士学位, 导师为穗志方教授。2009 年 7 月毕业, 获“计算机软件与理论”方向理学硕士学位。2009 年 10 月进入德国萨尔州大学 (Saarland University) 计算语言学系攻读博士学位, 师从 Hans Uszkoreit 教授, 同时在德国人工智能研究中心 (German Research Center for Artificial Intelligence) 语言技术实验室 (Language Technology Lab) 担任科研助理。2012 年 4 月通过博士论文 (题目为 LearningChinese?Language?Structures?with?Multiple?Views) 答辩, 获得萨尔州大学“数学与计算机科学”方向的工学博士学位。2012 年 5 月进入北京大学计算机科学技术研究所工作, 2015 年 8 月起任副教授。2016 年 11 至 12 月于香港城市大学翻译及语言学系任访问学者。

摘要: 统计机器学习技术是自然语言处理的基石之一, 和一般的应用场景相比, 自然语言的一大特点是其句法语义的结构性, 相应地结构化学习在自然语言处理中尤为重要。本报告主要讨论自然语言处理中的统计结构学习问题, 从语言本体研究的角度介绍句法语义结构表征方式, 从应用角度介绍这些结构和结构化学习之间的联系, 并以图结构学习为重点介绍自然语言领域近几年的重点研究工作。

商业银行“半监督”文本聚类技术应用

王彦博 (中国民生银行)

时间: 9:30~10:00 邮箱: wangyanbo@cmbc.com.cn

简介: 王彦博, 中国民生银行公司业务管理部数字化中心负责人。从事大数据挖掘科研及应用工作十余年, 具有丰富的智能化信息分析经验, 推行“智慧银行”大数据金融战略。入职民生银行以前, 曾任职英国国家文本挖掘研究中心副研究员; 兼任美国 IGI Global 出版社《知识社区与社会网络进展》系列丛书副主编, ACM 《智能系统与技术》、英国剑桥大学《知识工程回顾》、印度 DIVA 《数据挖掘与新兴技术》、德国 IBAI “业界数据挖掘”、IEEE “计算机应用与系统建模”等国际期刊、学术会议论文审稿专家。英国利物浦大学计算机科学博士、曼彻斯特大学计算机科学博士后; 发表著作 1 部, 著作章节 3 篇, 学术论文 50 余篇, 参与编写金融专业书刊 2 部, 获得国家专利 1 项、国家级奖励 1 项、省部级奖励 5 项。

摘要: 在商业银行日常经营管理过程中, 经常会产生大量非结构性文本数据。如何对这些文本数据进行分析挖掘, 从中提炼出有价值的信息并加以有效应用, 已经成为大数据时代商业银行需要解决的一项重要课题。通过构建“半监督”文本聚类技术, 对文本主题、类别、关键词和样本之间的关系进行学习, 从而实现对非结构性信息的结构化转换和提炼, 相关应用对商业银行经营管理提升起到积极推动作用。

Bayesian Text Classification and Summarization via A Class-Specified Topic Model

王菲菲 (北京大学)

时间: 10:30~11:00 邮箱: *ff11161224@126.com*

简介: 王菲菲, 北京大学光华管理学院商务统计与经济计量系博士研究生, 2012 年毕业于中国人民大学统计学院, 获经济学学士学位。感兴趣的研究领域有: 文本挖掘, 贝叶斯分析等。目前的研究课题集中在文本挖掘领域, 尤其是主题模型在营销和社交网络方面的应用, 近期也开始涉猎空间统计学在疾病分布方面的研究以及人口学领域的研究工作等。

摘要: We propose the Class Specified Topic Model (CSTM), an extension of the Latent Dirichlet Allocation (LDA) model, to address the problems of text classification and summarization of texts within classes. We assume that each document has a probability distribution over a set of class-specific topics and a set of common topics shared across classes. Each class-specific or shared topic has its own probability distribution over a dictionary of words or phrases. Bayesian inference of the CSTM in semi-supervised scenario is developed, with supervised scenario as a special case. We analyze the 20 Newsgroup dataset, a benchmark dataset for text classification, and demonstrate that the CSTM has better performance in text classification and summarization than a two-stage approach based on LDA and a L1 penalized logistic regression.

统计模型在关键词提取、文本分类和中文分词问题中的应用

张俊妮 (北京大学)

时间: 11:00~11:30 邮箱: *junnizhang@163.com*

简介: 张俊妮为北京大学光华管理学院商务统计与经济计量系副教授, 获美国哈佛大学统计学博士学位, 任北京大学商务智能研究中心副主任、北京大学光华管理学院责任与社会价值中心副主任。主要研究领域为贝叶斯分析、人口统计学、文本挖掘。

摘要: 本报告主要讨论文本挖掘中的三个问题: 提取关键词、文本分类以及分词。针对分类别的关键词提取问题, 我们提出了基于假设检验的三种统计方法。针对文本分类问题, 我们提出了一种将分类别的关键词引

入主题模型的方法, DWTM (The Discriminative Words Topic Model)。我们接着提出了一种结合主题信息的分词模型 WSTM (The Word Segmentation Topic Model), 该模型能够同时进行分词并使用分类别的关键词进行分类。我们使用一些英文和中文数据集, 比较了这些方法和现有的其他方法。

自由讨论

时间: 11:30~12:00

增强学习打麻将

陈昱 (北京大学)

时间: 8:30~9:00 邮箱: yu.chen@pku.edu.cn

简介: 北京大学光华管理学院统计系博士生。研究兴趣是时空统计以及深度学习。

摘要: 本报告和大家分享如何使用深度学习技术教会计算机打麻将，并且提高电脑 AI 的水平。使用到的技术包括卷积神经网络，增强学习，以及一些实现技巧。本文使用的技术很大程度上受到 AlphaGo 的启发，不同之处在于，1) 对麻将随机性的处理上，2) 使用最近一年来更为强大的增强学习算法，3) 更高效地利用稀有的训练数据。除了纯粹技术细节，我们还会分享从无到有地用深度学习解决问题的关键步骤，以及一些良好习惯，希望对大家有所帮助和启发。

字符级语言模型与机器翻译

赵申剑 (上海交通大学)

时间: 9:00~9:30 邮箱: sword.york@gmail.com

简介: 我来自上海交通大学计算机系，目前正在从事机器翻译相关应用和研究。

摘要: 局限于模型和计算能力，词级语言模型和机器翻译是一直以来的标准。但随着网络结构的发展和计算能力的提高，词级模型的缺点似乎可以通过字符级模型来解决。本次演讲主要探讨词级和字符级模型各自优缺点，并讨论目前字符级模型的发展情况，如语言模型和机器翻译中的进展。

阿里巴巴语音识别声学模型的进化历程

薛少飞 (阿里巴巴)

时间: 9:30~10:00 邮箱: shaofei.xsf@alibaba-inc.com

简介: 薛少飞，阿里巴巴 iDST 语音识别专家，中国科学技术大学博士。现负责阿里声学模型研究与应用：包括语音识别声学建模和深度学习在业务场景中的应用。博士期间的研究方向为语音识别说话人自适应，提出基于 Speaker Code 的模型域自适应方法，在语音相关的会议和期刊上发表论文十余篇。

摘要: 近年来，随着技术的发展，基于深度学习的语音识别已经成为业界主流的方法。本次演讲将首先带着大家梳理基于深度学习的语音识别声学模型发展历程，之后将分享阿里巴巴在语音识别声学建模上所做的技术突破，并展示我们在语音识别应用上的一些案例。

条件 GAN 用于车型设计和判别

张翔 (车轮互联)

时间: 10:30~11:00 邮箱: birdzhangxiang@gmail.com

简介: 10 年的 COS 水友, 车轮互联数据副总裁

摘要: 在移动互联网时代, 多屏媒体, O2O 多维互动, 给消费者购物带来了更多信息和更多选择。也给了企业更丰富, 更有挑战的营销环境。在众多影响决策的微时刻 (micro-moment) 和关键时刻 (moment of truth) 中, 汽车消费者的思维已经不自觉的进入了“车型鄙视链”的精神世界和换车魔力象限的领域。利用车轮查违章, 车轮社区 (覆盖 2 亿真实车主的 APP 应用) 中用户对车型 PK 投票的数据, 我们真实再现了这个车型鄙视链, 从中会发现每一款车, 你都可以找到选择他的理由。这为更加细分, 更加个性化的汽车市场提供了理论支撑。以此报告希望能够协助用户选到最适合自己的车, 也协助车厂在细分市场更加精准的定位, 甚至可以预测未来的汽车销量

R 语言中的深度学习: 用 Mxnet 进行车型识别

郎大为 (*J.D. Power*)

时间: 11:00~11:30 邮箱: *chiffonlang@icloud.com*

简介: J.D. Power 数据分析师, 致力于汽车行业的数据咨询, 曾任职于 Supstat, Ctrip, 浙江大学毕业导师, REmap, wordcloud2, leafletCN 等包的作者。

摘要: 深度学习发展到今天, 慢慢与一些传统的概念开始交叉, 迁移学习就是其中之一, 迁移学习可以通过已有模型的基础上进一步调整, 训练, 以大幅减少建模与训练的时间。本文将会以车型分类为例, 介绍如何使用一个预先训练好的模型, 在 R 语言中使用 mxnet 进行模型的微调 (fine-tuned), 并介绍入门深度学习的一些经验。

自由讨论

时间: 11:30~12:00

基于社交媒体大数据的心理学研究

朱廷劭 (中国科学院)

时间: 8:30~9:00 邮箱: tszhu@psych.ac.cn

简介: 中国科学院心理研究所研究员, 博士生导师, 入选中国科学院“百人计划”。获得中国科学院计算技术研究所硕士学位和博士学位, 于 2005 年获得加拿大 University of Alberta 博士学位。朱廷劭研究员的工作涉及机器学习、汉语转换以及网络行为心理研究等多个领域, 并取得创新性成果。他开展的网络行为心理研究, 从网络行为的分析实现对用户人格、心理健康以及社会态度的感知, 并在此基础上实现群体心理的预警预报和有效干预。

摘要: 互联网时代的到来, 能够将普通人的日常行为以空前的规模和精细程度进行记录, 形成网络行为大数据, 为个性心理研究提供了前所未有的机遇, 也提出了新的理论与技术问题。我们对反映个性特征的社会媒体大数据开展研究, 运用机器学习方法构建利用社交媒体数据预测用户个性心理特征的计算模型, 并开始尝试将模型预测作为测量手段运用于个性心理学研究。这些初步工作, 为网络技术支持下个性心理研究的纵深化、精细化发展做了铺垫, 并开始显露出巨大的应用潜力。

基于 R 与 Rstudio 的心理统计教学模式探索

吕小康 (南开大学)

时间: 9:00~9:30 邮箱: xkdog@126.com

简介: 南开大学副教授。主要研究方向包括: 文化与社会心理学, 尤其是对医学现象的社会学、心理学、人类学交叉视角研究及本土化阐释; 统计方法及统计社会学, 尤其是基于 R 语言的数据分析与可视化实践。出版了《R 语言统计学基础》、《AP 微积分基础教程》、《AP 统计学基础教程》等多篇专著, 在心理学报、心理科学进展、心理科学, 等心理学核心刊物上发表多篇文章。

摘要: 传统心理统计的教学模式较为依赖于纸笔运算, 所倾向使用的统计软件多为 SPSS, 教学过程中统计知识与软件的结合相对分离。要促进心理统计知识与技能的学习, 统计计算软件的深度参与是必不可少的, 但这种软件自身知识的学习同时可能加重学生的学习负担, 从而导致其学习动机和可持续性不足。本演讲将基于作者在南开大学周恩来政府管理学院的多年本科及硕士的统计类课程教学实践, 说明和演示如何在教学过程中贯穿“用统计软件为统计学习服务”的基本理念, 如何利用 R 与 Rstudio 简化统计教学的流程, 使学生更为便利地接受 R 语言这一统计计算工具, 同时利用 Rstudio 进行作业布置与管理。统计工具的大众化需要各学科内从事具体教学科研的工作人员不断提供的尝试方式, 最大化地体现新工具较之传统工具的优势, 如此才能形成一种良性的教学文化, 使得一种统计工具真正能够在学科领域扎根并流行。

R 语言在加强心理学可重复性中的作用

胡传鹏 (清华大学)

时间: 9:30~10:00 邮箱: hcp4715@163.com

简介: 清华大学博士五年级, 研究兴趣为社会认知神经科学。近年来, 由于心理学中的可重复危机, 开始关注如何加强心理学研究的可重复性问题。在《心理科学进展》上发表《心理学研究的可重复性问题: 从危机

到契机》，专门分析心理学研究中的可重复性问题。2016 年 10 月第 19 届全国心理学大会期间，举办《加强心理学研究的可重复性》工作坊。

摘要: 自 2011 年, 由于一系列的重复失败事件, 可重复性问题成为了心理学界一个持续的热点问题。统计方法严谨性(如过度依赖于 p 值)和开放性(大量的可疑研究操作)不足是导致心理学研究可重复率过低的重要原因。为了应对心理学中的可重复危机, 研究者们倡导使用更加多样的统计方法以及公开透明的研究实践。由于 R 语言中分析方法的灵活性与开放性, R 语言在加强心理学研究的可重复性上可以起到重要作用。首先, R 语言能够加强心理学研究中统计的严谨性和统计方法的多样性。由于众多心理学相关的软件包(psych, MBESS, lavaan, BootES, BayesFactor, Metafor 等), 使用 R 语言, 研究者可以进行多样的统计, 避免过度依赖于 p 值。例如, 使用 BootES, 研究者可以快速地对效应量的置信区间进行估计, 使用 BayesFactor, 研究者可以使用贝叶斯因子来进行统计。其次, R 语言能够加强心理学研究中数据分析的公开与透明的程度。与心理学中最常用的 SPSS 软件相比, R 语言代码而非鼠标点击进行数据的预处理以及分析, 能够精确地记录数据分析的过程。这对于将数据分析过程透明化具有重大意义。最近, 合理使用 Rmarkdown 的强大功能, 能够将数据处理与结果报告结合起来, 能够让研究者完全地数据分析与论文撰写无缝结合, 例如 papaja 工具包的出现, 可以让研究者直接使用 Rmarkdown 完成 APA 格式的文稿写作。正是由于 R 语言的这些优势, 在最近的重复研究中, R 语言被广泛地使用。例如, 2015, 发表在 Science 上的大规模重复实验中, 其数据分析用 R 完成。

心理学研究规范化及在 R 语言的实现

蔡培林 (天津师范大学)

时间: 10:30~11:00 邮箱: NA

简介: 天津师范大学应用心理专业硕士。PsychoR 团队成员, 研究方向为科研规范化与可重复研究的实现。

摘要: 自 2015 年 RPP 项目的结果公布, 表明其中的 100 项心理学研究中只有 39 项得到重复以来, 心理学可重复性的危机已昭然若揭。不可重复背后主要涉及到各种不规范的研究操作, 研究中使用的统计方法和出版偏见。为应对这种危机, 新提出的 TOP 标准强调研究要提前注册, 完整公开, 开放数据与材料。运用 R 语言各种包和函数的强大功能, 能有效增强心理学研究的规范化, 促进研究的公开、透明和开放, 从而提升研究的可重复性。

心理学在助老机器人研发中的应用

余嘉元 (南京师范大学)

时间: 11:00~11:30 邮箱: yujiayuanwx@163.com

简介: 南京师范大学教授, 博士生导师, 享受国务院特殊津贴专家。目前担任中国心理学会理事, 心理测量专业委员会副主任、中国机器学习学会理事和《心理学报》编委等职务。研究内容主要包括心理测量和认知心理学, 出版了《教育和心理测量》等多部专著, 在国内外学术刊物上发表了 200 余篇论文。在心理测量方面, 主要是对项目反应理论的研究, 包括对项目反应模型和参数估计方法的研究。在认知心理学方面, 主要是对问题解决的策略、联结主义(又称人工神经网络)及其应用进行了研究。

摘要: 当前 70 岁以上老年人存在不同程度的孤独感和抑郁感, 影响了他们的心理健康水平。在研发助老机器人的过程中, 我们采用隐马尔可夫模型和神经网络对老年人的语音进行分析, 通过模糊模式识别方法对他

们的孤独和抑郁程度进行评定。然后在事先构建的心理辅导知识库中提取相应的专家知识，对老年人进行个别的干预，从而缓解其孤独感和抑郁感。

自由讨论

时间：11:30~12:00

从语言智能到法务智能

吕正东 (深度好奇 (北京) 科技有限公司)

时间: 14:00~14:30 邮箱: luz@deeplycurious.ai

简介: 吕正东, 俄勒冈健康与科学大学计算机科学博士。曾于德州大学奥斯汀分校师从国际大数据及人工智能国际权威之一。Inderjit Dhillon 教授 (ACM、IEEE、SIAM 院士), 之后曾任职于微软亚洲研究院、华为诺亚方舟实验室等著名研究机构。长期从事机器学习及人工智能的研究, 在深度学习、自然语言处理、多模态学习和半监督学习领域卓有建树, 是深度学习领域 (尤其是自然语言处理方向) 具有世界一流水平并享有国际声誉的科学家和技术专家。2016 年创立深度好奇 (北京) 科技有限公司并任 CTO。

摘要: 虽然法律服务一贯具有较强的技术免疫力, 但是倚重信息检索、文件整理和逻辑推理的法律事务确实是人工智能特别是语言智能 “发挥所长” 的绝佳领域。本报告将围绕语义解析这一自然语言处理的终极任务, 探索 NLP 技术将如何重塑以法律为首的传统行业。我们提出和发展了深度学习和符号智能结合的方法在法律领域语义解析上的应用, 这些方法能够系统性地利用领域知识, 并在弱监督信号下进行有效的学习。以高效准确的语言技术为基础, 我们可以构建行业专家的辅助系统, 为人工服务中标准化的部分带来优化与变革。

智能时代的量化资产管理

郑亚斌 (鸣熙资产管理有限公司)

时间: 14:30~15:00 邮箱: dianshi_investment@163.com

简介: 郑亚斌, 2006 年于清华大学计算机系获得学士学位, 2011 年于清华大学计算机系获得博士学位, 主要研究方向为自然语言处理、人工智能。2011 年 7 月至 2013 年 9 月就职于国信证券经济研究所, 任金融工程分析师, 研究兴趣涵盖量化择时、行业配置、选股及量化对冲策略。2013 年 9 月至 2016 年 2 月就职于青骓投资管理有限公司, 担任投资经理, 管理 “光大 - 青骓 CTA 二期” 产品。2016 年 3 月至今就职于鸣熙资产管理有限公司, 担任投资总监。

摘要: 信息爆炸的互联网时代背景下, 如何利用人工智能技术提供高效准确的资产管理服务成为金融行业关心的话题。日趋增长的差异化资产管理需求也为传统资产管理行业提出了新的挑战。该报告将介绍如何利用自然语言处理、机器学习等相关技术, 从模型预测、信息检索、智能投顾等角度辅助投资决策。智能时代下的资产管理需要最大化地结合机器快速准确的处理效率及投资专家丰富的投资经验。

自然语言处理在医疗智能辅助中的应用

张超 (康夫子科技有限公司)

时间: 15:00~15:30 邮箱: zhangchao@kangfuzi.cn

简介: 张超, 北京康夫子科技有限公司创始人。曾在新加坡国立大学从事人工智能方向的研究工作, 后担任百度自然语言处理部资深研发工程师、文本知识挖掘方向负责人, 是知识图谱、实体建模方面专家。

摘要: 医疗领域是人工智能重点应用领域, 本报告重点阐述康夫子公司将自然语言处理技术医疗智能化中的应用研究。在应用层面, 主要表现为针对医生行医过程中提供临床辅助 (如: 病历书写辅助、诊断辅助)、针

对医学科研提供的病历可视化服务以及针对患者提供的导诊服务等等。在技术层面, 本报告概述了自然语言处理技术在知识图谱构建、病历结构化、智能诊断、对话交互、语义理解等方面的应用。

面向社交媒体的商业大数据挖掘

赵鑫 (中国人民大学)

时间: 16:00~16:30 邮箱: batmanfly@qq.com

简介: 赵鑫, 现为中国人民大学信息学院教师。师从北京大学李晓明教授, 专注于研究社交用户的兴趣建模。近五年内在国内外著名学术期刊与会议上以第一作者或者第二作者身份发表论文 40 余篇, 其中包括信息检索领域顶级学术期刊 ACM TOIS 和学术会议 SIGIR、数据挖掘领域顶级学术期刊 IEEE TKDE 和学术会议 SIGKDD、自然语言处理顶级会议 ACL 和 EMNLP。所发表的学术论文取得了一定的关注度, 据 Google Scholar 统计, 已发表论文共计被引用近 1400 次, 其中以第一作者发表的《Comparing Twitter and Traditional Media Using Topic Models》单文被引用 640 次。担任多个重要的国际会议或者期刊评审、CCL 2016 和 AIRS 2016 出版主席、NLPCC 2017 和 SMP 2017 的领域主席等。

摘要: 最近几年, 随着互联网技术的快速发展, 社交媒体服务在用户的真实生活中发挥着越来越重要的作用, 得到了广泛使用。同一用户可能同时拥有多个社交媒体网站的账号, 分别对应着不同的网络社区身份。以这些社区身份为基础, 用户可以同时参与到多个社交媒体平台, 享受其中提供的应用服务。因此, 在打造电子商务服务时, 能否同时围绕用户的“真实身份”与“在线社交身份”, 是一个很重要的思维创新。同时利用电子商务平台上的数据以及社交媒体平台上的用户数据, 将能够解决一些之前电子商务平台网站很难解决的技术挑战, 如冷启动推荐问题等。本次报告将围绕用户画像构建、用户意图检测和用户需求推荐等方面来进行相关内容介绍。

NLP 在金融报告自动化的实践

吴珂皓 (NA)

时间: 16:30~17:00 邮箱: wukehao@memect.co

简介: 吴珂皓, 北京文因互联数据科学家, 曾在美国杜兰大学负责大规模数据分析、管理维护研究中心数据仓库和高性能计算集群, 发表多篇 SCI 论文, 现负责文因互联报告自动化项目。

摘要: 投行、咨询公司依赖着昂贵的人力撰写分析研究报告, 存在大量的重复劳动工作, 由自然语言理解、知识图谱和自然语言生成技术组成的报告自动化技术正在逐步帮助这些公司降低成本提高工作效率。依靠自然语言理解对 PDF 进行结构化处理, 在主要内容分析、篇章语义的结构化、表格数据的结构化、文本摘要上都有不错的发挥; 而知识图谱在知识推理和检索方面都有不错的表现; 依靠自然语言生成和数据可视化帮助分析师自动生成报告, 降低人力成本。

自由讨论

时间: 17:00~17:30

Triple Generative Adversarial Networks

朱军 (清华大学)

时间: 14:00~14:30 邮箱: dcszj@mail.tsinghua.edu.cn

简介: 朱军, 清华大学计算机系长聘副教授、卡内基梅隆大学兼职教授、智能技术与系统国家重点实验室副主任、深度学习技术与应用国家工程实验室副主任。2001 到 2009 年获清华大学计算机学士和博士学位, 之后在卡内基梅隆大学做博士后, 2011 年回清华任教。主要从事人工智能基础理论、高效算法及相关应用研究, 在国际重要期刊与会议发表学术论文近百篇。受邀担任人工智能顶级杂志 IEEE TPAMI 和 AI 的编委、《自动化学报》编委, 担任机器学习国际大会 ICML2014 地区联合主席, ICML (2014-2017)、NIPS (2013, 2015)、UAI (2014-2017)、IJCAI (2015, 2017)、AAAI (2016, 2017) 等国际会议的领域主席, 中国计算机学会 (CCF) 学术工委主任助理。获微软学者、CCF 优秀博士论文奖、CCF 青年科学家奖、国家优秀青年基金、中创软件人才奖等, 入选国家“万人计划”青年拔尖人才、IEEE Intelligent Systems 杂志评选的“AI’s 10 to Watch”(人工智能青年十杰)、及清华大学 221 基础研究人才计划。

摘要: Generative adversarial nets (GANs) are good at generating realistic images and have been extended for semi-supervised classification. However, under a two-player formulation, existing work shares competing roles of identifying fake samples and predicting labels via a single discriminator network, which can lead to undesirable incompatibility. In this talk, I will present triple generative adversarial net (Triple-GAN), a flexible game-theoretical framework for classification and class-conditional generation in semi-supervised learning. Triple-GAN consists of three players - a generator, a discriminator and a classifier, where the generator and classifier characterize the conditional distributions between images and labels, and the discriminator solely focuses on identifying fake image-label pairs. With designed utilities, the distributions characterized by the classifier and generator both concentrate to the data distribution under nonparametric assumptions. Our results on several datasets demonstrate the promise in semi-supervised learning, where Triple-GAN achieves comparable or superior performance than state-of-the-art classification results among DGMs; it is also able to disentangle the classes and styles and transfer smoothly on the data level via interpolation on the latent space class-conditionally.

大规模线上实验与机器学习

熊熹 (京东)

时间: 14:30~15:00 邮箱: xiongxi@jd.com

简介: 2015 年加入京东, 一直致力于机器学习算法在京东个性化与推荐业务中的应用, 目前主要负责个性化业务中大规模线上实验, 指标定义, 异常追踪和用户体验优化等。曾在国内外知名大公司和研究机构从事复杂实验设计的理论和实践工作, 并持续跟踪大规模线上实验与机器学习在其中应用的前沿研究。由于在利用人工智能技术提升个性化用户体验以及更全面科学地定义个性化对京东的贡献等工作上的突出贡献, 曾获得 2016 年度 CTO 特别奖。

摘要: 大规模线上实验在京东每一天都在发生, 大到一个全新的模块乃至平台上线, 小到一个 icon 颜色, 样式的更改, 主要以 AB 试验的形式进行。大多数试验遵循直觉, 数据收集和整理的工作冗长, 但是对需要测试的指标以及收集到的数据的验证工作比较简单, 容易造成区分度不足乃至和真实结论南辕北辙的情况。本次报告会详细介绍线上实验的基本科学原则, 实施细节, 容易犯的错误; 并结合 google, 微软, LinkedIn, Amazon 等公司的最新研究论文, 以及京东个性化推荐中的实践, 从案例中学习如何使用机器学习和人工智能技术来验证数据一致性, 降低误差等。

Learning theory for deep nets

林绍波 (温州大学)

时间: 15:00~15:30 邮箱: sblin1983@gmail.com

简介: 2014 年 10 月毕业于西安交通大学。2015 年 3-2016 年 3 月, 香港城市大学博士后。现工作于温州大学统计系。研究方向为分布式学习理论与深度学习理论。

摘要: Deep learning has attracted avid research activities in the past few years. Compared with comprehensive application studies, the theoretical verifications lag heavily behind. This talk aims at developing a learning theory for deep learning to illustrate the power of deep nets. We construct a deep net containing pre-training stage, learning stage and fine-tuning stage to embody the three features of deep learning: multi-layered neural networks, large-scale algorithms and fine-tuning. Our constructed deep net is proved to attain the optimal learning rate when the ambient space is a lower dimensional manifold. This optimal learning rate is better than the existing results for shallow nets and therefore, shows the outperformance of deep nets.

腾讯社交广告实践中智能出价新模式: oCPA

王流斌 (腾讯)

时间: 16:00~16:30 邮箱: ubiwang@tencent.com

简介: 王流斌, 2010 年硕士毕业于北京大学软件工程专业, 同年加入腾讯, 先后参与过搜索广告、情境广告、社交广告的系统研发和策略优化工作, 专注于大规模并行机器学习系统研发、特征选择、转化率预估及应用等技术方向。担任 Tech Lead 的 oCPA 项目获得腾讯 2016 年度公司级技术突破奖。

摘要: 长期以来因为数据和技术的限制, 业内的广告系统大多只将广告的效果优化止于展现和点击阶段。我们系统中是如何衡量和优化广告转化效果的呢? 此次分享首先从营销漏斗开始介绍什么是转化。接着以电商和 App 为例讲解转化归因和转化跟踪技术。然后讲解转化率预估的建模方法、挑战和技术实现。最后介绍转化率在广告出价排序阶段的应用。希望通过分享让大家对转化闭环生态体系中的相关技术应用有一个整体的了解和认识。

bandit 算法与推荐系统

陈开江 (深圳市天农科技有限公司)

时间: 16:30~17:00 邮箱: kaijiangchen@gmail.com

简介: 陈开江 @ 刑无刀, 天农科技 CTO, 曾任新浪微博资深算法工程师, 考拉 FM 算法主管, 个性化导购 APP “Wave” 和 “边逛边聊” 联合创始人, 多年推荐系统从业经历, 在算法、架构、产品方面均有 “些许” 实践经验。

摘要: 推荐系统里面有两个经典问题: EE 问题和冷启动问题。前者涉及到平衡准确和多样, 后者涉及到产品算法运营等一系列东西。bandit 算法是一种简单的在线学习算法, 常常用于尝试解决这两个问题, 本文为你介绍基础的 bandit 算法及一系列升级版, 以及对推荐系统这两个经典问题的思考。

自由讨论

时间: 17:00~17:30

Building User Profiles from Online Social Behaviors, with Applications in Tencent Social Ads

靳志辉 (腾讯)

时间: 14:00~14:30 邮箱: rickyjin@qq.com

简介: Rickjin(靳志辉) ; 北京大学计算机系计算语言所硕士, 日本东京大学情报理工学院统计自然语言处理方向博士。2008 年加入腾讯, 主要作品内容涉及统计自然语言处理和大规模并行机器学习工具的研发工作。目前担任腾讯社交与效果广告部质量中心研发总监, 主要负责腾讯用户数据挖掘、精准广告定向、广告用户体验优化、广告转化率预估等工作。

摘要: The QQ (800M monthly users) and Wechat (700M monthly users) are the two largest instant messaging / social networks in China. Tencent Social Ads is the advertising system for both Wechat and QQ, serving well over 10B page views per day, for hundred million daily users.

We strive to understand as much as possible on our users' multiple aspects, so as to serve the best personalized ads for them. The rich user behaviors on Tencent's many products lay a solid foundation in user profiling. We develop audience targeting on many dimensions, including demographics, interests, intents, transactions, physical locations, and access environment, etc.

In this presentation, we will share our experience in large-scale user data mining for audience targeting, and discuss the challenges we face and the solutions we have employed.

微信中的社会传播课题与实践

高瀚 (腾讯)

时间: 14:30~15:00 邮箱: 364493790@qq.com

简介: 高瀚, 2013 年毕业于中山大学, 获数学学士及应用统计硕士学位。毕业后加入腾讯, 主要从事社交网络、社会传播以及 LBS 等领域的研究。先后主导“腾讯 LBSN (基于地理位置的社交网络)”、“宜出行 (城市热力图)”等系统的研发, 其中“宜出行”成为微信城市服务中的亮点功能之一, 并获得 IEEE 大数据峰会 (深圳分会) 技术创新奖。目前正在尝试将传统的社交网络理论与机器学习相结合, 应用于在微信业务中, 并取得了一定的成果。

摘要: 俗话说“酒香不怕巷子深”, 表面上说的是酒香引人, 实际上是指好酒在街坊邻里间口耳相传, 酒借着口碑飘香千里, 毋须大张旗鼓的门面, 也自会有客似云来。这就是口碑营销, 其背后是社会传播在起作用。线上社交工具的兴起, 为传统的社会传播学带来了全新的研究视角, 也提供了广阔的应用场景。本次分享将简要介绍微信中的社会传播问题、研究以及应用。

从文本分析看小说中人物的复杂关系: 以琅琊榜为例

周静 (中国人民大学)

时间: 15:00~15:30 邮箱: zhouding_89@126.com

简介: 周静, 中国人民大学统计学院助理教授, 北京大学光华管理学院管理学博士, 研究上关注复杂网络数据建模、营销模型、消费者行为分析等, 研究论文发表于 Journal of business and economic Statistics、Science China Mathematics、营销科学学报等国内外权威杂志上。在产业实践上, 对客户流失预警模型、用户欺诈模型等相关模型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的作者之一。

摘要: 本报告通过对人气网络小说《琅琊榜》进行小说三要素的文本分析, 从人物形象、故事情节和典型环境三个方面进行剖析。在人物形象的分析中主要探索不同人物之间的关系、从他们的动作、语言等方面去探索他们不同的性格特征。在故事情节上, 主要对小说的开端、发展、高潮和结局做了相应的分析, 同时为了研究人物之间复杂的关系, 我们对角色之间的亲密度、出场密度和称谓的变化等进行了分析。最后选取了几个典型环境来分析故事情节的发展。

On equivalence of likelihood maximization of stochastic block model and nonnegative matrix factorization, and beyond

张忠元 (中央财经大学)

时间: 16:00~16:30 邮箱: zhyuanzh@gmail.com

简介: 张忠元目前为中央财经大学统计与数学学院教授, 博士生导师, 中国计算机学会高级会员和果壳网科学顾问。主要研究兴趣在机器学习和复杂网络分析。在中国科学、Data Mining and Knowledge Discovery、Physical Review E、EPL(Europhysics Letters)、Scientific Reports、Knowledge and Information Systems、BMC Bioinformatics 等期刊发表过论文。

摘要: Community structures detection in complex network is important for understanding not only the topological structures of the network, but also the functions of it. Stochastic block model and nonnegative matrix factorization are two widely used methods for community detection, which are proposed from different perspectives. The relations between them are studied in this talk. The logarithm of likelihood function for stochastic block model can be reformulated under the framework of nonnegative matrix factorization. Besides the model equivalence, the algorithms employed by the two methods are different.

Furthermore, we design new matrix factorization model for signed network, and its effectiveness is evaluated.

Kaggle 数据挖掘比赛经验分享

陈成龙 (腾讯科技(深圳)有限公司)

时间: 16:30~17:00 邮箱: c.chenglong@gmail.com

简介: 陈成龙, 2015 年博士毕业于中山大学, 研究图像篡改检测, 在图像领域顶级期刊 IEEE Transactions on Image Processing 上发表论文 2 篇, Kaggle CrowdFlower 和 HomeDepot 搜索相关性比赛分获第一和第三名, 曾在 Kaggle 数据科学家排行榜上排名全球第十。目前在腾讯社交与效果广告部任职数据挖掘工程师, 负责 Lookalike 相似人群扩展相关工作。

摘要: Kaggle 是一个全球范围内具有很高影响力的大数据比赛平台, 举办过很多有名的比赛, 如 KDD Cup。同时, 不少知名的公司 (如 Google, Facebook, Microsoft 等) 也在 Kaggle 上发布题目, 开放数据, 吸引全球上万名数据科学家共同来解决业界难题。此次分享会首先介绍 Kaggle 比赛的一些基本情况, 包括参赛方式, 比赛流程, 组队方式, 在线论坛和编程环境等。进一步, 我们会介绍 Kaggle 比赛项目类型, 以及相应

的常用机器学习技术和工具，涵盖图像分类，搜索相关性和 pCTR 等任务。最后会结合具体的比赛项目，分享特征工程，模型训练和模型集成等方面的一些经验。

The relationship between meteorological factors and hand, foot, and mouth disease (HFMD): DLNMs-based time-series analysis

张志杰 (复旦大学)

时间: 14:00~14:30 邮箱: epistat@gmail.com

简介: 复旦大学流行病学与卫生统计学, 副教授, 研究方向为空间流行病学、统计方法与模型。

摘要: 简单回顾一下时间序列分析以及 R 中的时间序列分析程序包, 然后聊一下以往时间序列分析的问题, 引入 DLNMs 方法。以手足口病为例, 重点介绍一下该模型在实际研究中的应用, 建模过程中的一些细节, 快速演示一下软件的操作, 以及结果的解释, 以期让听众能有效地掌握该技术方法。

Assessment of the impact of climate on respiratory infectious disease via pomp package in R

张兵 (广东省公共卫生研究院)

时间: 14:30~15:00 邮箱: zhangbing4502431@outlook.com

简介: 2014 年 6 月从华中科技大学劳动卫生与环境卫生学系毕业, 先就职于浙江省疾病预防控制中心, 现如今就职于广东省公共卫生研究院。支持开源软件, 喜欢折腾代码和数据, 对传染病动力学模型、时空数据分析和数据可视化感兴趣, 现今主要研究气象环境因素对传染病发病的影响。

摘要: 气象因素与传染病发病关系密切, 已有很多研究如广义相加模型、小波分析等一系列方法来探讨过气象因素是如何作用于传染病的, 但上述方法都不能解决哪些气象因素作用于病原体, 哪些气象因素作用于人体。本研究通过构建一个含阈值和滞后效应的传染病动力学模型, 并通过基于隐马尔科夫模型和粒子滤波算法从生态学角度探讨气象因素作用于传染病的可能机制。

R Epidemics Consortium and Using Its Packages to Analyze Influenza Data

蔡俊 (清华大学)

时间: 15:00~15:30 邮箱: cai-j12@mails.tsinghua.edu.cn

简介: 蔡俊, 清华大学地球系统科学系生态学专业 2012 级直博生, 研究兴趣包括流感传播动态、传染病流行病学和环境健康。博士期间主要从事中国内地 2009 年甲型 H1N1 流感时空传播动态研究, 并于美国国立卫生研究院 Fogarty 国际中心国际流行病学和人口研究司短期访学。同时是一名 R 语言爱好者, 拥有 5 年 R 语言编程和数据分析经验, 是 geoChina 和 humidity 包作者以及 animint 和 incidence 包贡献者, R Epidemics Consortium 成员。近期对 R 在传染病建模中的应用感兴趣。

摘要: R 流行病联盟 (R Epidemics Consortium, RECON) 聚集了一群传染病建模、公共卫生和软件开发方面的国际专家, 通过使用 R 软件创建下一代疾病暴发响应分析工具。RECON 目前包括使用最前沿的统计方法对疾病暴发数据进行处理、可视化以及分析的专门软件包, 以及更多针对疾病数据清理、版本控制和加密等

通用工具。本演讲将介绍 RECON 的创立背景、目标、成员以及拥有的 R 流行病方面的软件包项目和资源，最后以分析流感暴发数据为例，展示如何利用 RECON 的 incidence 和 EpiEstim 包快速绘制流行曲线并估计随时间变化再生数。

中国 H7N9 禽流感暴发模拟与预测

李瑞云 (北京师范大学)

时间: 16:00~16:30 邮箱: ruiyunli@mail.bnu.edu.cn

简介: 北京师范大学 2014 级博士研究生，研究方向：环境健康

摘要: H7N9 禽流感病毒的出现对中国以及世界公共健康构成了重大挑战。然而，对于其在家禽中的传播及扩散模式和家禽到人的跨宿主传播机制知之甚少。本文将流行病学模型和数据同化方法结合起来，并利用人感染 H7N9 禽流感病例来估计流行病学重要参数，并且对家禽和人感染禽流感做出了预测。研究结果表明，尽管 H7N9 禽流感病毒在家禽中造成了较大规模的感染 (33%)，但从家禽至人的跨宿主传播的可能性较低。此外，我们能较准确的预测出 H7N9 在人类中传播时的峰值时间和爆发强度。该研究结果说明，H7N9 禽流感病毒在禽类中的传播模式以及实时的跨宿主传播是可预测的。

基于 R 语言的登革热传播模型建立与参数化

程渠 (清华大学)

时间: 16:30~17:00 邮箱: chengtooto@126.com

简介: 清华大学地学系博士生

摘要: 登革热是世界上传播最快的蚊媒病毒传染病。2014 年广州市共报告 38036 例病例，占 1990 到 2015 年中国大陆报告病例数的 52%。数学模型可以被用于研究 2014 年广州市登革热暴发的决定因素。本演讲的主要内容包括利用 R 语言建立登革热传播数学模型；利用区域敏感性分析法 (regional sensitivity analysis) 来对模型进行参数化；构建不同情景来研究暴发的决定因素。

基于 R 语言环境下气候因素—登革热媒介蚊虫的动力学模型建立与研究

贾鹏飞 (北京师范大学)

时间: 17:00~17:30 邮箱: j_pengfei@yahoo.com

简介: 贾鹏飞，北京师范大学全球变化与地球系统科学研究院 2014 级博士生。博士期间主要从事登革热媒介白纹伊蚊种群的时空建模，以及全球变化与媒介种群波动的分析工作。该研究隶属 973 国家重大专项“气候变化对人类健康的影响”子课题。博士阶段科研工作突出，曾获“北京师范大学学术一等奖学金”，在国际上疾病媒介研究期刊 Parasites & Vectors (SCI 二区) 发表学术论文 2 篇，多次参加国内外学术会议并做专题报告。在该课题研究过程中，该种群模型中动力学方程的建立和求解主要通过 R 语言编程实现，后续的绘图工作多利用其中的 ggplot 工具完成，对 R 语言在公共卫生方面的研究有一定的心得体会。

摘要: 气候变化是一个典型的全球尺度环境问题, 其中全球变暖给我们带来的影响毋庸置疑。登革热作为一种蚊虫传播的病毒病在全球均有分布, 并广泛流行于全球热带及亚热带地区。白纹伊蚊作为该疾病的重要传播蚊虫媒介, 在全球大面积扩散造成了显著影响。从蚊虫角度出发, 白纹伊蚊幼虫生长发育和活动规律与气温、降水量、光周期等自然因素密切相关。该研究通过构造数学微分方程组的形式, 建立“气候因素—白纹伊蚊”种群动态模型, 并模拟中国大陆的白纹伊蚊生长和繁殖情况。其中方程组的求解过程以及模拟结果的展示均借助 R 语言强大的编程环境和绘图工具完成。该工作在一定程度上说明 R 语言在公共卫生领域有强的应用前景, 同时表明该机理模型对在未来气候情境下的种群预测和防治工作有重大的指导意义。

这是一条来自韬映资本的广告。你之所以会看这段文字，是因为此刻的你既没有去度假，也不在去度假的路上，但你想。你想让自己闲下来，大脑放空，就像这张白纸。陪陪家人，做点浪费时间的事，比如晒晒太阳。如果你真的想，韬映资本乐意为你实现这样的梦想。稳健可观的收益回报，提

前步入财务自

由的生活。

扫描下面的

二维码，给

自己一个机

会吧。毕竟，

有钱有闲的

生活，有谁

会嫌弃呢？



韬映资本，

致力于为当

代精英提供

卓越的财富

管理服务的

私募基金。以合

伙人思维、跨界理

念，汇聚国内外顶、

尖金融机构精英，

拥有高执行力、高

素质团队。专注私

募债权、私募股权

投资、上市公司并

购重组，并围绕新

文化教育、医疗健

康、影视传媒、互

联网+等领域上市

及拟上市公司整合

产业链。联手国泰

君安、财通基金，

推出债券基金悠系

列、臻系列，定增

基金致系列，历史

平均年化收益率

35%。让远方更

近，让未来

可期。



主办单位



清华大学
Tsinghua University

清华大学统计学研究中心



北京大学
PEKING UNIVERSITY

北京大学商务智能研究中心



CAPITAL OF STATISTICS
PROFESSION, HUMANITY & INTEGRITY

统计之都

协办单位



狗熊会
CluBear

聚数据英才、助产业振兴

狗熊会



清华大学统计学研究中心



统计之都



狗熊会

10th

The China-R Conference

