

# 11

第十一届  
中国R会议 (北京)  
会议手册



中国人民大学



5月25日—5月27日



# 合作伙伴



战略合作伙伴



金牌赞助



银牌赞助



视频支持



# 欢迎辞

高朋远来千万里，盛会不觉十一年。R会议开到第十一届了，早就成了一种习惯。十载歌声犹萦在耳，一纪期盼不溢于面。每一年都是新征程，今年格外与众不同。世界潮流、浩浩荡荡，千年大计、百年变局，一时登览。我辈数据中人幸立潮头，躬逢盛世，正是乘风破浪的大好时机。大数据是一种原材料，凿开混沌得乌金，蓄藏阳和意最深，这样的好东西要凿开后才知道。数据科学是一种生产力，千锤万凿出深山，烈火焚烧若等闲，这种利器要不断试炼才管用。大数据不是空中楼阁，数据科学也不是屠龙之技，关键是要用。数据如果用得好，大则万物互联、小则数字化管理，升则人工智能，隐则风控于无形。尤其是要把握时代的机会，助力产业振兴，同时秉承专业、人本、正直的理念。



本次中国 R 会议会场共有 1 个主会场和 20 个分会场，汇聚了 115 位演讲嘉宾。会议名虽为 R，但 R 只是最初大家聚在一起的契机。君子不器、圣人不凝滞于物，这次所有报告主题的高频词是“数据”、“应用”、“研究”、“分析”，正好也概括了会议的初衷：融合学界和业界、探讨数据科学的应用。R 是为数据而生的编程语言，经过这么多年的进化 (evolution) 后，伴随着数据科学给各行各业带来了革命 (Revolution) 性的进展。参会者都来自于数据科学的各个领域，很多人都在数据应用的第一线奋战。难得有机会聚在一起，希望能跨领域、跨专业地进行交流和碰撞，相互借鉴、共同学习，把数据技术深入地应用到业务中，为产业升级和现代化的新征程贡献自己的力量。

统计之都敬上  
2018年5月19日

# 扬帆起航

Allegro  $\text{♩} = 117$

中国北京大学毕业生誓词  
项海波 作曲

The musical score consists of eight staves of music in G major (two sharps) and common time. The vocal line is in soprano range. The lyrics are integrated into the musical notes, with one note per syllable. The score includes measure numbers 6, 11, 16, 21, 27, 32, and 37.

**6**  
 明德博学求是笃行勤勉

**11**  
 为新朴实友爱志存高远争做

**16**  
 国民表率扬帆起航啊勇做社会栋

**21**  
 梁弘弘扬人大精神感念浓浓母校情肩负时代

**27**  
 使命共筑巍巍中国梦弘扬人大精神感念

**32**  
 浓浓母校情肩负时代使命共筑巍巍中国

**37**  
 梦

# 目录

欢迎辞	1
<b>会议介绍</b>	<b>1</b>
第十一届中国 R 会议介绍 . . . . .	1
主办机构 . . . . .	2
赞助商介绍 . . . . .	3
第十一届中国 R 会议筹备委员会 . . . . .	5
统计之都简介及活动回顾 . . . . .	6
中国人民大学地图 . . . . .	6
主会场 & 专题会专场日程 . . . . .	7
<b>Keynote(25 日, 北京香格里拉饭店)</b>	<b>15</b>
范剑青: Farming Big Data in R Environment . . . . .	15
山世光: 摸眼盲人管窥 AI 之 ABCDE . . . . .	15
邓柯: 统计学与健康中国 . . . . .	16
陈松蹊: 空气质量评估的气象调整方法 . . . . .	16
刘霁: 大规模深度学习的并行优化以及加速策略 . . . . .	17
王汉生: 数据交易与治理 . . . . .	17
<b>R01: 知识图谱与金融大数据应用 (26 日上午, 第一教学楼 1402, 主席: 王翀, 吴岚)</b>	<b>19</b>
鲍捷: NLP 和知识图谱技术在金融领域的应用 . . . . .	19
任亮: 基于产业知识图谱的公司金融创新 . . . . .	19
王守昆: 对话式交互的现状与未来 . . . . .	19
李文哲: 知识图谱技术以及它在不同垂直领域中的应用 . . . . .	20
孟嘉: 知识图谱面向金融行业的落地实践 . . . . .	20
<b>R02: 机器学习 (26 日上午, 第一教学楼 1404, 主席: 常象宇)</b>	<b>21</b>
孙剑: Model-driven Deep Learning . . . . .	21
徐增林: On Tensor Networks and Neural Networks . . . . .	21
储晨: 电商对抗智能 Adversarial Intelligence in E-commerce . . . . .	21
王乃岩: Towards Practical Deep Learning Model Compression and Acceleration . . . . .	22
黄庆昌: 在生产环境中使用 R 语言构建机器学习项目 . . . . .	22
<b>R03: 心理学 (26 日上午, 第一教学楼 1505, 主席: 夏骁凯)</b>	<b>23</b>
杨志: 脑健康研究中的数据挖掘 . . . . .	23
李燕: 基于项目反应理论的汉字识别在线测验的编制与应用 . . . . .	23
殷继兴: 心理学研究中的可重复性 –以神经成像偏见效应为例 . . . . .	23
麦子峰: 日间节律对睡眠缺失效应的调节机制——项探索性研究 . . . . .	24
甘怡群: 心理学研究统计的新趋势 . . . . .	24
夏骁凯: 统计之都 COStudy 心理学项目介绍 . . . . .	24
<b>R04: 医疗大数据 (26 日上午, 第三教学楼 3204, 主席: 李帆)</b>	<b>26</b>
陈显扬: 基于 R 开发的针对慢性病多中心项目研究一站式平台 . . . . .	26
何松: 面向药物重定位的多组学数据整合分析框架 . . . . .	26

---

李舰：一个肺癌大数据分析平台的建制 . . . . .	26
李嵘：半参数强分层整合分析的变量选择研究 . . . . .	27
<b>R05: 数据科学实践 (26 日上午, 第三教学楼 3310, 主席: 雷博文)</b>	<b>28</b>
石茂林: 数据挖掘在隧道掘进机设计及分析的应用 . . . . .	28
刘欣益: 基于小波变换与循环神经网络的在线手写签名识别方法 . . . . .	28
任乔牧: AI in Agriculture . . . . .	28
游皓麟: AI 技术与智能投放机器人 . . . . .	29
杨晨: 惧买与惜售——基于幂律的羊群效应检测方法及其量化策略 . . . . .	29
<b>R06: 数据可视化 (26 日上午, 第三教学楼 3102, 主席: 郎大为)</b>	<b>30</b>
曾勇: 开源的可视化数据分析平台 Kibana 从 0 到 100 . . . . .	30
王修坤: R 语言的可视分析应用 . . . . .	30
郎大为: 可视化中的静与动 . . . . .	30
周宁奕: 基于比特币交易的可视化分析 . . . . .	31
张杰: R 语言 ggplot2 之地理信息可视化 . . . . .	31
<b>R07: 统计理论 (26 日上午, 第三教学楼 3308, 主席: 周茂袁)</b>	<b>32</b>
张振: A New Constrained L1 Minimization Approach to the High-dimensional Markowitz Portfolio Optimization Problem . . . . .	32
刘杰: 不平衡数据下故障诊断算法综述 . . . . .	32
徐铣明: 关于微生物群落的 SDE 动态建模 . . . . .	32
黎磊: An ARIMA Model with Adaptive Orders for Predicting Blood Glucose Levels and Hypoglycemia . . . . .	33
周茂袁: Likelihood ratio-based distribution-free sequential change-point detection . . . . .	33
<b>R08: 智能对话 (26 日下午, 第三教学楼 3308, 主席: 李嫣然)</b>	<b>35</b>
吴俣: Deep Chit-chat: 教机器人如何唠嗑 . . . . .	35
翁嘉硕: 人工智能的情感交互与行业融合 . . . . .	35
史树明: 腾讯 AI Lab 的开放域智能对话研究 . . . . .	35
马宇驰: 语义识别的商业破局 . . . . .	35
吴金龙: 保险行业的对话机器人技术 . . . . .	36
<b>R09: 狗熊会专场 (26 日下午, 第三教学楼 3101, 主席: 潘蕊)</b>	<b>37</b>
潘蕊: 狗熊会人才培养与数据科学实践教学 . . . . .	37
庄池杰: 面向电力系统的数据分析与应用 . . . . .	37
杨慧: 公司数据能力的人才构建以及公司垂直数据人才社群运营 . . . . .	37
张维: 用数据的智慧改变建筑行业 . . . . .	37
金雄男: 大数据在足球运动中的应用及潜力 . . . . .	38
常莹: 知己知彼, 百战不殆——根据公开的广告数据优化 SEM 推广策略 . . . . .	38
<b>R10: 智能营销 (26 日下午, 第三教学楼 3310, 主席: 沈俏蔚)</b>	<b>39</b>
姚凯: 基于弹幕的在线社交对视频扩散的影响 . . . . .	39
张晗: 基于文本挖掘的个性化推荐 . . . . .	39
姜舒文: The Value of Seller Community: Evidence from Live Webcasting Platform . . . . .	39
梁屹天: 数字广告作假的实证分析 . . . . .	40
郭麦菊: The Value of Time: A Study of Pricing Strategy on a Ride-Sharing Platform . . . . .	40
申桐遥: Star popularity or star acting skill? The impact of star power on movie box office . . . . .	41
<b>R11: 量化金融 (26 日下午, 第一教学楼 1404, 主席: 叶征)</b>	<b>42</b>

刘正瑶：非零售信用风险内部评级法——客户评级建模 . . . . .	42
张佳：保险精算领域的大数据应用 . . . . .	42
王璟：区块链技术与商用特性介绍 . . . . .	42
周宇光：期权波动率的多彩空间 . . . . .	42
任坤：量化交易中的 R 高性能计算 . . . . .	43
谢士晨：信贷评分卡模型的开发与应用简介 . . . . .	43
<b>R12: 智能工程系统 (26 日下午, 第三教学楼 3102, 主席: 张玺, 王凯波)</b>	<b>44</b>
黄毅：制造业转型中的数字化精益 . . . . .	44
周杰, 崔鹏飞: 风电大数据分析: 机遇与挑战 . . . . .	44
李三华: 振动分析在高速旋转设备故障诊断中的应用 . . . . .	44
宁永铎: 基于大数据的硅片形状诊断和预报 . . . . .	44
王迪: 利用迁移学习实现储备粮关键品质指标的全程估计 . . . . .	45
<b>R13: 统计学在经济管理领域的应用 (26 日下午, 第三教学楼 3204, 主席: 张俊妮)</b>	<b>46</b>
徐敏亚: 组织情境下员工工作行为大数据研究 . . . . .	46
王菲菲: Analysis of Consumer Reviews Using Sequential Phrase Selection . . . . .	46
郇钰: Bayesian Estimation of the General Probability of Informed Trading Model . . . . .	46
李晨煦: 随机波动率模型下的闭形式隐含波动率曲面 . . . . .	47
宋晓军: Validating Power Laws in Economics and Finance . . . . .	47
张俊妮: Bayesian Estimation of Demographic Systems . . . . .	47
<b>R14: 医疗健康管理 (27 日上午, 第三教学楼 3308, 主席: 常象宇)</b>	<b>49</b>
鄂尔江: 患者住院数据的探索性分析 . . . . .	49
宁温馨: Semantics-driven Feature Extraction for High-throughput Phenotyping . . . . .	49
崔力文: 医疗资源消耗预测与预测任务导向的医疗编码表示学习 . . . . .	49
林毓聪: 引入文本章节结构进行远距离关系提取并应用于医学知识提取 . . . . .	50
文天才: 69863 例脑卒中患者合并疾病研究 . . . . .	50
<b>R15: 公共卫生 (27 日上午, 第三教学楼 3204, 主席: 蔡俊)</b>	<b>52</b>
杜向军: 基于 R 的流行病动力学模型测试 . . . . .	52
王锡玲: Epidemiology of human infections with influenza A(H7N9) virus and pandemic risk assessment	52
徐波: 流感疫情双峰现象的潜在机制 . . . . .	53
林华亮: R 在环境流行病学研究中的应用 . . . . .	53
李国星: 全球变暖的疾病负担研究 . . . . .	53
蔡俊: Non-inheritable risk factors during pregnancy for congenital heart defects in offspring: a matched case-control study . . . . .	54
<b>R16: 软件工具 (27 日上午, 第三教学楼 3201, 主席: 杜亚磊)</b>	<b>55</b>
杜亚磊: 初窥爬虫门径 . . . . .	55
李朋飞: Shiny: 从零到一搭建可视化 BI 平台 . . . . .	55
李智凡: Tensorflow 在 R 上的部署及使用 . . . . .	55
吕剑航: 基于 LSTM 的自动文本生成模型构建 . . . . .	55
周震宇: R 环境模型部署实践 . . . . .	56
黄湘云: R Markdown 应用之学位论文排版 . . . . .	56
<b>R17: 城市大数据 (27 日上午, 第三教学楼 3102, 主席: 李栋)</b>	<b>57</b>
戴劭勍: 地理要素的尺度效应、可变面积单元问题与空间统计的挑战 . . . . .	57

蔡纪烜：多源大数据探测城市多中心结构 . . . . .	57
李颖：多源数据融合辅助人口分析与政府管理 . . . . .	57
冯娟：城市产业结构及其经济复杂度研究 . . . . .	58
魏贺：数据 - 证据 - 决策：交通规划的例证与思考 . . . . .	58
王扬：综合能源数据分析平台构建及其在智慧城市中的应用 . . . . .	58
 <b>R18: 金融数据 (27 日上午, 第三教学楼 3101, 主席: 冯凌秉)</b>	 <b>60</b>
林伟林：数据分析在资产管理行业的实践 . . . . .	60
霍志骥：收益中的 Alpha 与 Beta . . . . .	60
赵然：量化基本面投资与大类资产配置 . . . . .	60
张云松：互联网征信的探索与实践 . . . . .	60
谢军：风险? 推荐?: 真实银行数据分析工作实践分享 . . . . .	61
罗小勇：量化风险管理与 R 语言一键式建模探索 . . . . .	61
 <b>R19: 车联网 (27 日上午, 第三教学楼 3310, 主席: 周扬)</b>	 <b>62</b>
陶建辉：超融合、超高性能的车联网大数据平台 . . . . .	62
张翔：区块链管理车联网数据 . . . . .	62
陈宸：大数据为二手车行业赋能 . . . . .	62
盛超：车联网数据与车辆可靠性研究实践 . . . . .	62
朱俊辉：LBS 数据科学实践 . . . . .	63
耿文童：车联网大数据的应用 . . . . .	63
 <b>R20: 西安欧亚学院专场 (27 日上午, 第三教学楼 3303, 主席: 张俊丽)</b>	 <b>64</b>
王艳：数据科学与大数据技术人才培养体系 . . . . .	64
吴睿：证券分析师的价值分析 . . . . .	64
张俊丽：物流车辆风险评估 . . . . .	64
贾蓓：西安市名牌战略实施效果调研 . . . . .	65
孙旭：文学书籍的市场发展探究 . . . . .	65

## 第十一届中国 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、兰州财经大学、中国科学技术大学等多个城市的高校举办。2017 年，第十届中国 R 语言会议在北京、上海、合肥、兰州、武汉、太原、西安等城市分别举办，其中清华大学举办的北京会场参会者逾 2000 人。今年将迎来第十一届中国 R 会议。

本届 R 会议由中国人民大学统计学院、北京大学光华管理学院、统计之都主办，狗熊会和中国人民大学应用统计科学研究中心协办，将于 5 月 25-27 日在北京举办。本届会议覆盖知识图谱与金融大数据应用、机器学习、量化金融、心理学、医疗大数据、统计应用、公共卫生、数据科学实战、智能营销、车联网、数据可视化、智能工程、城市大数据、统计理论、智能对话、医疗健康管理、金融数据和软件工具等数据科学话题，我们欢迎您的到来！

25 日主会场地点为海淀区紫竹院路 29 号香格里拉酒店 2 层会议厅，26-27 日分会场地点为中国人民大学公共教学一楼、公共教学三楼，请您事先查阅好感兴趣的会场，并提前熟悉校园环境和路线，以便更加高效地参加会议。

## 主办机构

### 中国人民大学统计学院

中国人民大学统计学科始建于 1950 年，2003 年建院。全国重点学科，2007 年教育部二级学科评估排名全国第一，2012 年教育部统计学一级学科评估排名全国第一。拥有统计学一级学科博士点和博士后流动站，经济统计学和风险管理与精算学两个二级学科博士点，拥有预防医学与公共卫生一级学科硕士授权点，应用统计学专业学位硕士点，统计学、经济统计学、应用统计学（风险管理与精算）三个本科专业，是全国拥有理学、经济学、医学三大门类统计学专业最齐全的统计学院。

### 北京大学光华管理学院

秉承北大百年风骨，以“创造管理知识，培养商界领袖，推动社会进步”为使命，作为北大工商管理教育的主体，北京大学光华管理学院是亚太地区最优秀的商学院之一。历经三十三年的发展，北大光华已形成完整的学科结构、一流的师资队伍、丰富的教学体系，吸引着有理想、有担当、有情怀、有责任感的有志之士砥砺前行，取得了令人瞩目的成绩，是中国商学教育的一面旗帜，并形成了植根于燕园独立精神与自由思想的“因思想，而光华”的独特精神气质。

### 统计之都

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

### 狗熊会

狗熊会，数据产业的高端智库，并以“聚数据英才、助产业振兴”为己任。通过精品案例，让更多的朋友享受数据分析的快乐，并助力其终身的职业幸福与成长。通过企业联合研究，陪伴中国的数据产业一起成长，共同见证他们的辉煌！

### 中国人民大学应用统计科学研究中心

中国人民大学应用统计科学研究中心是中华人民共和国教育部所属 100 所人文社会科学重点研究基地之一，它成立于 2000 年 9 月，其前身是 1988 年成立的中国人民大学统计科学研究所。研究中心积极培育中青年学术骨干，逐渐发展并形成了经济与社会统计、统计调查与数据分析、风险管理与精算、生物卫生统计，四个各具特色的研究方向。中心建设的重点研究平台是：1. 重大发展问题的统计技术创新研究。2. 现代统计技术与方法的应用性研究。3. 精算技术的创新与应用。4. 生物医学统计技术发展与应用。研究中心拥有国内一流的研究人员，承担多项国家及教育部项目，获得丰硕的研究成果。应用统计科学研究中心，始终将建立和发展应用统计学科基地作为战略定位，着重从制定应用统计研究的科学规划、密切联系实际选准科研攻关方向、注重研究工作的长期积累、加强重点研究平台建设等方面开展工作。

## 赞助商介绍

### 战略合作伙伴

#### 派生集团

派生科技集团有限公司（以下简称“派生集团”）于 2011 年在东莞成立，注册资金 10 亿元。聚焦实业、科技、金融三大战略投资方向，致力于成为一家以“大数据、人工智能、互联网科技”等技术力量驱动产融结合、提升产业运营效率的投资服务集团。目前集团主要有实业科技、环保科技、金融科技等核心业务板块，集团员工近 20000 人。2017 年，派生集团核心业务板块在东莞合计纳税超过 2.86 亿元。

### 金牌赞助

#### 星通联华

北京星通联华科技发展股份有限公司是一家以地质灾害、地下水监测及预警预报；公路运营管理及巡检养护信息化；道路、桥梁、隧道智能检测与养护管理等相关前沿科技产品的研发、应用推广和技术服务为方向的高新技术企业。公司自成立以来始终坚持以科技创新、服务客户为宗旨，在地灾、地下水、交通、建筑、环保、物流等领域提供监测、检测、系统开发及相关产品销售和服务等全方位的解决方案。在地灾、地下水领域提供业内领先的传感器技术及完整的整体解决方案，现已开发出种类齐全、技术先进、性能稳定可靠的传感器系列和多种应用软件；在交通领域，结合国内公路行业现状及发展趋势，目前已开发出用于公路行业建设、养护检测和管理的产品有：多功能道路综合检测车、激光弯沉检测车、公路探地雷达系统、智能手机路况数据采集系统、桥梁健康实时无线监测系统、公路质量检评系统、公路地理信息（GIS）管理系统、施工过程质量监管系统等。星通联华拥有一支高素质的优秀管理团队和研发团队，具备良好教育背景和丰富的专业经验，团队成员多年来持续专注行业内解决方案及服务，并积极推动商业模式创新，对新商业模式的发展具有坚定的信念。同时，公司以信任心使用人才，以事业感激激发人才，以诚信取信人才，不断吸引业内高端人才加盟，构筑专注、快速发展通道，最大限度的发展人才价值。星通联华作为国家智能交通标准化委员单位，参与多项标准的制定，承担各项交通运输部、科技部的科研项目，具有强大的行业影响力。展望未来，公司将抓住机遇，坚持以人为本，加大自主产品创新和推广，以客户导向为指引，不断挖掘客户需求，继续保持公司的产品和技术在市场上的领先地位。

#### 中国人民大学出版社

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

#### 天启智创

北京天启智创信息技术有限公司注册于 2016 年 10 月，2017 年 3 月正式运营。天启智创以人工智能算法、风控数据服务、决策引擎系统为核心技术能力，目前主要专注于服务金融信贷和保险行业客户，为客户提供营销、智能化决策系统、风险模型策略和数据服务。天启智创未来定位于互金、银行、保险等行业的金融科技能力构建，产品方向为数据模型服务、风控和决策系统产品、用户信用风险评估等。

## RStudio

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

## Elastic

Elastic 致力于构建大规模实时数据处理软件，场景主要涵盖搜索、日志、安全与数据分析等领域。公司成立于 2012 年，旗下拥有产品包括开源的 Elastic Stack（Elasticsearch、Kibana、Beats 和 Logstash）、X-Pack（商业特性）和 Elastic Cloud（一种托管服务）。迄今为止，这些产品的累积下载次数已超过 2.25 亿。Elastic 由 Benchmark Capital、Index Ventures 及 NEA 投资，投资额超过 1 亿美金。Elastic 拥有超过 800 位员工，分布于世界上 30 多个国家和地区。欲了解详情请访问：[elastic.co](http://elastic.co)。

## 银牌赞助

### 图灵教育

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

### 美库尔

Merkle Inc，世界顶级的大数据营销咨询公司。创立于 1971 年，总部设在美国首都大华府地区，全球 3800 多名员工分布在 21 座城市，为超过 200 家的国际大型品牌提供咨询服务，涉及多个行业和领域。公司于 2009 年初起，拓展部分核心业务到中国，为全球和本土企业提供客户关系市场营销服务，主要业务包括：市场营销策划、客户关系管理、统计建模和分析，数据处理分析、数据库开发、管理和服务。中国的办公地点为上海和南京。

## 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。200+ 合作社区，每周 30+ 场技术大会精彩分享，4000+ 业内大咖资源。让程序猿、攻城狮不再遗憾，随时随地，想看就看，让智慧流动起来！

## 第十一届中国 R 会议筹备委员会

主席：杨舒仪

秘书长：董安澜

秘书团：毕季文，龚小艳，顾小涵，雷博文，李楠，李宇轩，任怡萌，王小宁，王祎帆，赵沫燕

志愿者：安澜、曹毛毛，曾千涵，崔思颖，常勤缘，董俊毅，房鸿宇，高钰婷，耿林圆，郭昊苏，何国星，何贤文，侯杰，黄嘉炜，姜姗，孔祥宜，李昂，李浩源，李杰桠，李俊杰，李璇，李泳欣，廖子宜，刘昊宇，刘馨宇，刘志恒，娄立威，鲁毅，马莉丽，马宁，聂仪珂，秦宇婷，邱雅娟，任焱，沈楠，石佳鑫，石晓辉，宋玉良，孙强强，孙瑄梓，田家赫，王梦一，王若彤，王哲，夏春秋，向悦，许亚昆，许智彤，闫晓雨，杨春白雪，杨珍珍，于金萍，于玉洁，张家玮，张文轩，张宵，张轩瑜，张媛媛，赵乘，钟厚岳，周昕仪

## 统计之都简介及活动回顾

“统计之都”(Capital of Statistics, 简称 COS)网站成立于 2006 年 5 月 19 日，其主旨为传播统计学知识并将其应用于实际领域。纵观现今国内统计学理论和应用的发展，一方面我们不难发现统计学在应用领域的巨大潜力——现代管理、咨询、商业、经济、金融、医药、生物等等，无不需要数据的力量，而另一方面我们也不得不承认，国内统计学的应用很大程度上受理论的制约——无论是应用界的人们对统计学基础理论知识的欠缺，还是学术界所研究的理论对应用领域问题的轻视。

“统计之都”网站便是基于这样的认识而创建的。我们希望，统计理论研究者能充分关注应用问题，而统计应用者也能正确把握统计学基本知识，将统计学这门应用学科真正的潜力开发出来。

“统计之都”为非赢利性质网站，但大力欢迎所有商界和研究领域的朋友与我们在实际应用问题上合作。我们的口号是：

中国统计学门户网站，免费统计学服务平台

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

统计之都(下文简称 COS)目前由线上与线下两部分构成。其中，线上内容主要包括主站(<http://cosx.org/>)以及微信公众号(CapStat)；随着越来越多喜爱数据科学的朋友们加入，大家对于线下活动和书稿撰写翻译等等的需求也越来越旺。COS 线下活动总结：

COS 线下活动总结：

1. 中国 R 会议：目前已开展到第十届，分别在北京、上海、广州、杭州、西安、武汉、成都、贵阳、南昌、厦门、合肥、太原等地举办。历届会议纪要和幻灯片共享都可以在 COS 主站上找到：<http://china-r.org/>
2. 线下沙龙：目前我们在北京、上海和广州深圳开展线下沙龙活动。不同于规模庞大的 R 语言会议，沙龙形式更为轻巧，注重讨论交流。目前已经举办过 40 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 23 期：<http://meetup.cos.name/>.
4. 书籍出版，包括写作和翻译。如《Dynamic Documents with R and knitr》(2nd edition) 谢益辉著，《Implementing Reproducible Research》谢益辉等著，《bookdown: Authoring Books and Technical Documents with R Markdown》谢益辉著，《数据科学中的 R 语言》李舰、肖凯著，《R 语言实战》高涛、肖楠、陈钢翻译，《ggplot2: 数据分析与图形艺术》统计之都翻译，《R 语言核心技术手册》刘思喆、李舰、陈钢、邓一硕翻译，《R 语言编程艺术》陈堰平、邱怡轩、潘岚峰等翻译，《R 数据可视化手册》肖楠、邓一硕、魏太云翻译，《R 语言统计入门》邓一硕、郝智恒、何通翻译，《数据科学实战》冯凌秉、王群锋翻译，《R 语言实战》(第 2 版) 王小宁、刘撷芯、黄俊文翻译，《Rcpp: R 与 C++ 的无缝结合》寇强、张晔翻译，《R 绘图系统》呼思乐、张晔、蔡俊翻译，《R 语言编程实战》冯凌秉翻译，《量化投资与 R》(待出版) 邓一硕、冯凌秉、杨环翻译，《金融风险建模与投资组合优化》(待出版) 邓一硕、郑志勇等翻译、《ggplot2: 数据分析与图形艺术 (第 2 版)》黄俊文、王小宁、于嘉傲、冯璟烁等等。

## 中国人民大学地图



注：图上标记的餐厅都可以现金消费。

## 5月25日（周五）主会场日程

主会场	演讲嘉宾	主题	时间
Keynote (北京香格里拉饭店)		致辞	09:00~09:15
	范剑青	Farming Big Data in R Environment	09:15~10:00
	山世光	摸眼盲人管窥 AI 之 ABCDE	10:00~10:45
		自由讨论、休息	10:45~11:15
	邓柯	统计学与健康中国	11:15~12:00
Keynote (北京香格里拉饭店)	陈松蹊	空气质量评估的气象调整方法	14:00~14:45
	刘霖	大规模深度学习的并行优化以及加速策略	14:45~15:30
	王汉生	数据交易与治理	15:30~16:15
		自由讨论、休息	16:15~16:35
		光华管理学院 BA 项目宣讲会	16:35~18:00

# 第十一届中国R会议北京分会场日程

分会场	26日上午	26日下午	27日上午
1402	R01 知识图谱与金融大数据应用 (光华专场) 主席：王翀 & 吴岚		
1404	R02 机器学习 主席：常象宇	R11 量化金融 (星通联华冠名) 主席：叶征	
1505	R03 心理学 主席：夏晓凯		
3204	R04 医疗大数据 主席：李舰	R13 统计学在经济管理领域的应用 (光华专场) 主席：张俊妮	R15 公共卫生 主席：蔡俊
3310	R05 数据科学实践 主席：雷博文	R10 智能营销 (光华专场) 主席：沈俏蔚	R19 车联网 主席：周扬
3102	R06 数据可视化 (Elastic 冠名) 主席：郎大为	R12 智能工程系统 (派生集团冠名) 主席：张玺 & 王凯波	R17 城市大数据 主席：李栋
3308	R07 统计理论 主席：周茂袁	R08 智能对话 主席：李嫣然	R14 医疗健康管理 主席：常象宇
3101		R09 狗熊会专场 主席：潘蕊	R18 金融数据 (天启智创冠名) 主席：冯凌秉
3201			R16 软件工具 主席：杜亚磊
3303			R20 西安欧亚学院专场 主席：张俊丽

注：分会场全部分布在中国人民大学教学一楼和教学三楼各教室，如1402表示教学一楼四层1402教室。

## 5月26日(周六) 上午分会场

分会场	演讲嘉宾	主题	时间
知识图谱与金融大数据应用 <b>(光华专场)</b> 第一教学楼 1402 主席: 王翀, 吴岚	鲍捷	NLP 和知识图谱技术在金融领域的应用	09:00-09:30
	任亮	基于产业知识图谱的公司金融创新	09:30-10:00
		自由讨论、休息	10:00-10:30
	王守崑	对话式交互的现状与未来	10:30-11:00
	李文哲	知识图谱技术以及它在不同垂直领域中的应用	11:00-11:30
	孟嘉	知识图谱面向金融行业的落地实践	11:30-12:00
机器学习 <b>第一教学楼 1404</b> 主席: 常象宇	孙剑	Model-driven Deep Learning	09:00-09:30
	徐增林	On Tensor Networks and Neural Networks	09:30-10:00
		自由讨论、休息	10:00-10:30
	储晨	电商对抗智能 Adversarial Intelligence in E-commerce	10:30-11:00
	王乃岩	Towards Practical Deep Learning Model Compression and Acceleration	11:00-11:30
	黄庆昌	在生产环境中使用 R 语言构建机器学习项目	11:30-12:00
心理学 <b>第一教学楼 1505</b> 主席: 夏骁凯	杨志	脑健康研究中的数据挖掘	08:30-09:00
	李燕	基于项目反应理论的汉字识别在线测验的编制与应用	09:00-09:30
	殷继兴	心理学研究中的可重复性 —— 以神经成像偏见效应为例	09:30-10:00
		自由讨论、休息	10:00-10:30
	麦子峰	日间节律对睡眠缺失效应的调节机制——一项探索性研究	10:30-11:00
	甘怡群	心理学研究统计的新趋势	11:00-11:30
医疗大数据 <b>第三教学楼 3204</b> 主席: 李舰	夏骁凯	统计之都 COStudy 心理学项目介绍	11:30-12:00
	陈显扬	基于 R 开发的针对慢性病多中心项目研究一站式平台	09:00-09:30
	何松	面向药物重定位的多组学数据整合分析框架	09:30-10:00
		自由讨论、休息	10:00-10:30
	李舰	一个肺癌大数据分析平台的建制	10:30-11:00
数据科学实践 <b>第三教学楼 3310</b> 主席: 雷博文	李嵘	半参数强分层整合分析的变量选择研究	11:00-11:30
	石茂林	数据挖掘在隧道掘进机设计及分析的应用	09:00-09:30
	刘欣益	基于小波变换与循环神经网络的在线手写签名识别方法	09:30-10:00
		自由讨论、休息	10:00-10:30
	任乔牧	AI in Agriculture	10:30-11:00
	游皓麟	AI 技术与智能投放机器人	11:00-11:30
	杨晟	惧买与惜售——基于幂律的羊群效应检测方法及其量化策略	11:30-12:00

分会场	演讲嘉宾	主题	时间
<b>数据可视化 (Elastic 冠名)</b> <b>第三教学楼 3102</b> 主席：郎大为	曾勇	开源的可视化数据分析平台 Kibana 从 0 到 100	09:00-09:30
	王修坤	R 语言的可视分析应用	09:30-10:00
		自由讨论、休息	10:00-10:30
	郎大为	可视化中的静与动	10:30-11:00
	周宁奕	基于比特币交易的可视化分析	11:00-11:30
	张杰	R 语言 ggplot2 之地理信息可视化	11:30-12:00
<b>统计理论</b> <b>第三教学楼 3308</b> 主席：周茂袁	张振	A New Constrained L1 Minimization Approach to the High-dimensional Markowitz Portfolio Optimization Problem	09:00-09:30
	刘杰	不平衡数据下故障诊断算法综述	09:30-10:00
		自由讨论、休息	10:00-10:30
	徐铣明	关于微生物群落的 SDE 动态建模	10:30-11:00
	黎磊	An ARIMA Model with Adaptive Orders for Predicting Blood Glucose Levels and Hypoglycemia	11:00-11:30
	周茂袁	Likelihood ratio-based distribution-free sequential change-point detection	11:30-12:00

## 5月26日(周六)下午分会场

分会场	演讲嘉宾	主题	时间
<b>智能对话</b> <b>第三教学楼 3308</b> 主席：李嫣然	吴俣	Deep Chit-chat: 教机器人如何唠嗑	14:00-14:30
	翁嘉颀	人工智能的情感交互与行业融合	14:30-15:00
	史树明	腾讯 AI Lab 的开放域智能对话研究	15:00-15:30
		自由讨论、休息	15:30-16:00
	马宇驰	语义识别的商业破局	16:00-16:30
	吴金龙	保险行业的对话机器人技术	16:30-17:00
<b>狗熊会专场</b> <b>第三教学楼 3101</b> 主席：潘蕊	潘蕊	狗熊会人才培养与数据科学实践教学	14:00-14:30
	庄池杰	面向电力系统的数据分析与应用	14:30-15:00
	杨慧	公司数据能力的人才构建以及公司垂直数据人才社群运营	15:00-15:30
		自由讨论、休息	15:30-16:00
	张维	用数据的智慧改变建筑行业	16:00-16:30
	金雄男	大数据在足球运动中的应用及潜力	16:30-17:00
<b>智能营销</b> <b>(光华专场)</b> <b>第三教学楼 3310</b>	常莹	知己知彼，百战不殆——根据公开的广告数据优化 SEM 推广策略	17:00-17:30
	姚凯	基于弹幕的在线社交对视频扩散的影响	14:00-14:30
	张晗	基于文本挖掘的个性化推荐	14:30-15:00
	姜舒文	The Value of Seller Community: Evidence	15:00-15:30

主席：沈俏蔚		from Live Webcasting Platform	
		自由讨论、休息	15:30-16:00
	梁屹天	数字广告作假的实证分析	16:00-16:30
	郭麦菊	The Value of Time: A Study of Pricing Strategy on a Ride-Sharing Platform	16:30-17:00
	申桐遥	Star popularity or star acting skill? The impact of star power on movie box office	17:00-17:30
量化金融 (星通联华冠名) 第一教学楼 1404 主席：叶征	刘正瑶	非零售信用风险内部评级法——客户评级建模	14:00-14:30
	张佳	保险精算领域的大数据应用	14:30-15:00
	王璟	区块链技术与商用特性介绍	15:00-15:30
		自由讨论、休息	15:30-16:00
	周宇光	期权波动率的多彩空间	16:00-16:30
	任坤	量化交易中的 R 高性能计算	16:30-17:00
	谢士晨	信贷评分卡模型的开发与应用简介	17:00-17:30
智能工程系统 (派生集团冠名) 第三教学楼 3102 主席：张玺，王凯波	黄毅	制造业转型中的数字化精益	14:00-14:30
	周杰，崔鹏飞	风电大数据分析：机遇与挑战	14:30-15:00
	李三华	振动分析在高速旋转设备故障诊断中的应用	15:00-15:30
		自由讨论、休息	15:30-16:00
	宁永铎	基于大数据的硅片形状诊断和预报	16:00-16:30
	王迪	利用迁移学习实现储备粮关键品质指标的全程估计	16:30-17:00
统计学在经济管理领域的应用 (光华专场) 第三教学楼 3204 主席：张俊妮	徐敏亚	组织情境下员工工作行为大数据研究	14:00-14:30
	王菲菲	Analysis of Consumer Reviews Using Sequential Phrase Selection	14:30-15:00
	郇钰	Bayesian Estimation of the General Probability of Informed Trading Model	15:00-15:30
		自由讨论、休息	15:30-16:00
	李晨煦	随机波动率模型下的闭形式隐含波动率曲面	16:00-16:30
	宋晓军	Validating Power Laws in Economics and Finance	16:30-17:00
	张俊妮	Bayesian Estimation of Demographic Systems	17:00-17:30

## 5月27日（周日）上午分会场

分会场	演讲嘉宾	主题	时间
医疗健康管理 第三教学楼 3308 主席：常象宇	鄂尔江	患者住院数据的探索性分析	09:00-09:30
	宁温馨	Semantics-driven Feature Extraction for High-throughput Phenotyping	09:30-10:00
		自由讨论、休息	10:00-10:30
	崔力文	医疗资源消耗预测与预测任务导向的医疗编码表示学习	10:30-11:00
	林毓聪	引入文本章节结构进行远距离关系提取并应用于医学知识提取	11:00-11:30
	文天才	69863例脑卒中患者合并疾病研究	11:30-12:00
公共卫生 第三教学楼 3204 主席：蔡俊	杜向军	基于 R 的流行病动力学模型测试	08:30-09:00
	王锡玲	Epidemiology of human infections with influenza A(H7N9) virus and pandemic risk assessment	09:00-09:30
	徐波	流感疫情双峰现象的潜在机制	09:30-10:00
		自由讨论、休息	10:00-10:30
	林华亮	R 在环境流行病学研究中的应用	10:30-11:00
	李国星	全球变暖的疾病负担研究	11:00-11:30
软件工具 第三教学楼 3201 主席：杜亚磊	蔡俊	Non-inheritable risk factors during pregnancy for congenital heart defects in offspring: a matched case-control study	11:30-12:00
	杜亚磊	初窥爬虫门径	08:30-09:00
	李朋飞	Shiny:从零到一搭建可视化 BI 平台	09:00-09:30
	李智凡	Tensorflow 在 R 上的部署及使用	09:30-10:00
		自由讨论、休息	10:00-10:30
	吕剑航	基于 LSTM 的自动文本生成模型构建	10:30-11:00
城市大数据 第三教学楼 3102 主席：李栋	周震宇	R 环境模型部署实践	11:00-11:30
	黄湘云	R Markdown 应用之学位论文排版	11:30-12:00
	戴劭勍	地理要素的尺度效应、可变面积单元问题与空间统计的挑战	08:30-09:00
	蔡纪烜	多源大数据探测城市多中心结构	09:00-09:30
	李颖	多源数据融合辅助人口分析与政府管理	09:30-10:00
		自由讨论、休息	10:00-10:30
	冯娟	城市产业结构及其经济复杂度研究	10:30-11:00
	魏贺	数据-证据-决策：交通规划的例证与思考	11:00-11:30
	王扬	综合能源数据分析平台构建及其在智慧城市中的应用	11:30-12:00

分会场	演讲嘉宾	主题	时间
金融数据 (天启智创冠名) 第三教学楼 3101 主席: 冯凌秉	林伟林	数据分析在资产管理行业的实践	08:30-09:00
	霍志骥	收益中的 Alpha 与 Beta	09:00-09:30
	赵然	量化基本面投资与大类资产配置	09:30-10:00
		自由讨论、休息	10:00-10:30
	张云松	互联网征信的探索与实践	10:30-11:00
	谢军	风险? 推荐? : 真实银行数据分析工作实践分享	11:00-11:30
	罗小勇	量化风险管理与 R 语言一键式建模探索	11:30-12:00
车联网 第三教学楼 3310 主席: 周扬	陶建辉	超融合、超高性能的车联网大数据平台	08:30-09:00
	张翔	区块链管理车联网数据	09:00-09:30
	陈宸	大数据为二手车行业赋能	09:30-10:00
		自由讨论、休息	10:00-10:30
	盛超	车联网数据与车辆可靠性研究实践	10:30-11:00
	朱俊辉	LBS 数据科学实践	11:00-11:30
	耿文童	车联网大数据的应用	11:30-12:00
西安欧亚学院专场 第三教学楼 3303 主席: 张俊丽	王艳	数据科学与大数据技术人才培养体系	09:00-09:30
	吴睿	证券分析师的价值分析	09:30-10:00
		自由讨论、休息	10:00-10:30
	张俊丽	物流车辆风险评估	10:30-11:00
	贾蓓	西安市名牌战略实施效果调研	11:00-11:30
	孙旭	文学书籍的市场发展探究	11:30-12:00

## Farming Big Data in R Environment

范剑青（普林斯顿大学）

时间：09:15-10:00

**简介：**范剑青 (Jianqing Fan)，现为复旦大学大数据学院教授、院长，以及普林斯顿大学 Frederick L. Moore'18 金融学讲座教授，2000 年荣获 COPSS 总统奖 (国际统计学领域最高奖项)，2006 年荣获洪堡基金会终身成就奖，2007 年荣获晨兴华人数学家大会应用数学金奖，2009 年荣获在美国文理与艺术界著名的 GUGGENHEIM 学者 (Fellow)，2012 年入选国家“千人计划”项目并当选台湾“中央研究院”院士，2013 年获泛华统计学会 (International Chinese Association) 的“许宝禄奖”，2014 年荣获英国皇家统计学会授予的“Guy Medal”银质奖章，现为国际统计学会 (International Statistical Institute) 会士、国际数理统计学会 (Institute of Mathematical Statistics) 会士、美国统计学会 (American Statistical Association) 会士、美国科学促进会 (American Association for the Advancement of Science) 会士、计量金融学会 (The Society for Financial Econometrics) 会士。主要研究领域为高维统计、机器学习、大数据科学、经济学、金融学、生物信息等。学术成果发表在 Annals of Statistics, Journal of American Statistical Association, Journal of Machine Learning Research, Econometrica, Journal of Econometrics, Journal of Financial Economics 等国际一流期刊上。目前为国际一流期刊 Journal of Econometrics 的联合主编，Journal of American Statistical Association 的副主编。

**摘要：**Correlated and heavy-tailed data arise frequently in a wide range of scientific and engineering problems: from genomics, medical imaging to neuroscience and finance. This talk introduce Factor-Adjusted Robust Multiple testing (FARM-test) and Factor-Adjusted Robust Model Selection (FARM-select). The former is introduced to control the false discovery proportion for large-scale simultaneous inference when variables are highly correlated and the latter deals the variable selection problems when covariates are highly correlated. We demonstrate that robust factor adjustments are extremely important in both improving the power of the tests and controlling FDP and improving model selection consistency rates. These will be demonstrated through newly developed R packages. We identify general conditions under which the proposed method produces a consistent estimate of the FDP. We also prove that factor adjustments significantly reduce the conditions needed for selection consistency. The results will be illustrated by numerical experiments.

## 摸眼盲人管窥 AI 之 ABCDE

山世光（中科院智能信息处理重点实验室）

时间：10:00-10:45

**简介：**山世光，中科院计算所研究员、博导，现任中科院智能信息处理重点实验室常务副主任。他是第三批国家“万人计划”入选者，国家基金委优青，国家百千万人才工程入选者，中国计算机学会青年科学家奖获得者，科技部中青年科技创新领军人才。他的专业领域为计算机视觉和机器学习，在人脸识别等图像识别技术上有超过 20 年的研发经验，带领团队获得过十余次国内外学术竞赛冠亚军，所研发的人脸识别技术成功应用于公安部出入境管理局、十几省公安厅、华为手机等。已在国内外刊物和学术会议上发表论文 200 余篇，其中计算机学会认定的 A 类刊物和会议论文 70 余篇，论文被谷歌学术引用 13000 余次。曾应邀担任过 ICCV, ACCV, ICPR, FG, ICASSP 等 10+ 次领域主流国际会议的领域主席，现/曾任 IEEE TIP, CVIU, PRL, Neurocomputing, FCS 等国际学术刊物的编委 (AE)。研究成果获 2005 年度国家科技进步二等奖，2015 年度国家自然科学二等奖。

**摘要：**AI 热潮持续发酵，人类社会似乎正在快速进入所谓 AI 时代。那么，AI 时代真的指日可待了吗？本报告将在回顾 AI 领域近期部分重要进展的基础上，对此次 AI 热潮背后的最大推手——深度学习技术的源起和原理做介绍，然后将探讨深度学习给计算机视觉技术和系统的研发带来的方法论上的变迁，以及它最适用的领域和问题类型。最后，将对比人类智能，分析深度学习对全面实现 AI 时代的不足，以及未来需要继续努力的方向。

## 统计学与健康中国

邓柯（清华大学）

时间：11:15-12:00

**简介：**邓柯，清华大学统计学研究中心副教授、副主任，博士生导师。2008 年获北京大学统计学博士学位，同年进入哈佛大学统计系从事研究工作，历任博士后、副研究员，2013 年加入清华大学。2014 年入选“青年千人计划”并当选中国数学会概率统计学会第十届理事会理事，2015 年当选中国医疗保健国际交流促进会医学数据与医学计量分会常务委员，2017 年当选中国现场统计研究会计算统计分会首任理事长、中国现场统计研究会环境与资源分会常务理事，2018 年当选国际计算统计学会亚太地区分会理事。他还获得了“科学中国人（2016）年度人物”的荣誉称号。邓柯的研究兴趣包括统计建模、统计计算、生物信息、文本分析、医疗大数据分析、政府数据分析等领域。他的一系列研究成果发表在统计学和其他学科的顶级期刊上。

**摘要：**习近平主席在党的《十九大报告》中明确提出了“健康中国”战略，指出“要把人民健康放在优先发展的战略地位”，“切实解决影响人民群众健康的突出环境问题”、“加强食品安全监管”、“努力减少公共安全事件对人民生命健康的威胁”、“提供连续的健康管理服务和医疗服务”，从而“全方位、全周期保障人民健康，大幅提高健康水平，显著改善健康公平”。在这一重大系统工程的许多方面，都需要统计学发挥不可替代的关键作用。在本报告中，我将结合我们近年来在医疗大数据分析、食品安全监管、卫生技术评估等领域的研究和实践，探讨统计学在实施“健康中国”战略中的巨大机遇与挑战。我们也非常希望越来越多的统计届同仁关注相关的统计学研究与应用，与我们一道共同推动统计学在落实国家重大战略中发挥关键作用。

## 空气质量评估的气象调整方法

陈松蹊（北京大学）

时间：14:00-14:45

**简介：**陈松蹊，国家特聘专家，北京大学讲席教授，商务统计与经济计量系联合系主任、北京大学统计科学中心联席主任国家首批“千人计划”入选者，加盟北大后主要致力于商务统计与经济计量学学科建设及北大统计学研究队伍的建设工作。他是数理统计学会 (Institute of Mathematical Statistics) 资深会员 (fellow)，美国统计学会会士 (fellow)，国际统计学会 (International Statistics Institute) 当选会员 (elected member)，国际数理统计学会 (IMS) 理事会常务理事 (Council member)。他现在是 The Annals of Statistics(统计年鉴) 副主编 (自 2010 年)；Journal of Business and Economic Statistics 副主编 (自 2013 年)；曾任 Statistics and Its Interface 的联席主编 (2010-2013)。

**摘要：**Although air pollution is caused by emission of pollutants to the atmosphere, the observed pollution levels are largely affected by meteorological conditions which determine the dispersion condition of the pollutants. Effective air quality management requires statistical measures that are immune to the meteorological

confounding in order to evaluate spatial and temporal changes of the pollution concentration objectively. Motivated by a challenging task of assessing changes and trends in the underlying pollution concentration in a region near Beijing, we propose a spatial and temporal adjustment approach for the PM<sub>2.5</sub> and other five pollutants with respect to the meteorological conditions by constructing a spatial and temporal baseline weather condition based on historic data to remove the meteorological confounding. The adjusted mean pollution concentration is shown to be able to capture changes in the underlying emission while being able to control the meteorological variation. Estimation of the adjusted average is proposed together with asymptotic and numerical analyzes. We apply the approach to conduct assessments on six pollutants in the Beijing region from Year 2013 to Year 2016, which reveal some intriguing patterns and trends that are useful for the air quality management.

## 大规模深度学习的并行优化以及加速策略

刘霖 ( 罗切斯特大学 )

时间：14:45-15:30

**简介：**刘霖博士毕业于威斯康星大学麦迪逊分校计算机系，本科毕业于中国科大自动化系。当前是美国罗切斯特大学助理教授以及腾讯 AI lab 专家研究员。他的研究涉猎于诸多理论和应用方向，包括机器学习，平行算法，优化方法，强化学习，计算机视觉，游戏 AI 设计，生物信息学，多媒体，图形学等等。他在顶级的人工智能相关会议以及期刊上发表超过 40 篇论文，包括 NIPS, ICML, KDD, CVPR, ICCV, UAI, JMLR, PAMI 等。他曾获得 2010 数据挖掘顶级会议 KDD 最优论文提名奖和 2015 顶级人工智能会议 UAI Facebook 最优学生论文奖，他的论文曾入选 2017 机器学习顶级会议 NIPS 大会宣讲（比例 1%）。同时他还荣获 2017 IBM 最优教职员奖和 2017 MIT TR35 ( China )。

**摘要：**计算效率是大规模深度学习的重要瓶颈之一，同时也是人工智能技术落地的关键性因素之一。该报告将以个人的科研和实践为基础从多个层面和角度（比如，异步并行计算，区中心化的计算体系结构，硬件算法联合设计等等），介绍当前深度学习重要的并行框架，算法，和理论，并将展望未来的发展方向。另外还将探讨数据隐私保护和分布式学习算法结合。

## 数据交易与治理

王汉生 ( 北京大学 )

时间：15:30-16:15

**简介：**王汉生，北京大学光华管理学院商务统计与经济计量系，嘉茂荣聘讲席教授，博导；北京大学商务智能研究中心主任；光华管理学院 MBA, EMBA, ExEd, 本硕博教学指导委员会成员；美国统计学会 ( American Statistical Association ) 会士 ( Fellow, 2014 )。1998 年北京大学数学学院概率统计系本科毕业，2001 年美国威斯康星大学麦迪逊分校统计系博士毕业。2003 年加入光华至今。国内外各种专业杂志上发表文章逾 80 篇，并合著英文专著 1 本，中文教材 2 本。国际统计协会 ( International Statistical Institute )、英国皇家统计协会 ( Royal Statistical Society )、美国统计协会 ( American Statistical Association )、美国数理统计协会 ( Institute of Mathematical Statistics )、泛华国际统计协会 ( International Chinese Statistical Association ) 的会员。先后历任以下国际学术刊物副主编 ( Associate Editor )：The Annals of Statistics ( 2008—2009 ), Computational Statistics & Data Analysis ( 2008—现在 ), Statistics and its Interface ( 2010—现在 ), Journal of the American Statistical Association ( 2011—现在 ), Statistica Sinica ( 2011—现在 ), Journal of Business and

Economics Statistics (2012—现在), 中国科学数学 (2013—现在)。在理论研究方面, 关注高维数据分析。具体内容有: 变量选择、收缩估计、数据降维等。在应用方面, 关注统计学方法在电子商务领域的应用, 尤其关注中文文本分析、社会关系网络、以及位置轨迹数据。

**摘要:** 数据交易势不可挡。该趋势同任何组织或者个体, 之于数据交易的观点和态度无关, 这是一个基本的不可阻挡的趋势。就如同山洪一样, 它一定从高处往下流, 跟任何人的喜好无关。对山洪不管不顾, 或者刻意阻挡, 都有可能酿成洪灾。但是, 如果加以合理的规范治理, 它可以用于发电提供能源。对于数据交易一样, 无论是不负责任的放任自流, 还是不切实际的严苛管制, 都不是值得推崇的建设性方案。一个更具建设性的方案是, 集合更多的智慧, 为数据交易发展一套尽可能完备的理论框架。尽可能发挥数据交易的社会价值, 而极小化它所带来的伤害。该理论框架应该包括, 但不局限于, 数据治理 (含: 数据确权与合规), 交易标的与形态, 和数据定价理论。这是本次分享的核心内容。

## NLP 和知识图谱技术在金融领域的应用

鲍捷 (文因互联)

时间: 09:00-09:30

**简介:** 文因互联 CEO, 联合创始人。美国爱荷华州立大学 (Iowa State University) 博士。曾任伦斯勒理工学院 (RPI) 博士后, 麻省理工大学 (MIT) 访问研究员, 三星美国研发中心研究员, 三星问答系统 SVoice 第二代系统核心设计师。1998 年以来一直从事人工智能研究, 主要研究领域包括机器学习、神经网络、数据挖掘、自然语言处理、形式推理、语义网和本体工程, 发表了 70 多篇领域内相关论文。作为 W3C OWL(web 本体语言) 工作组成员, 在语义网的标准制定过程中起到了关键性作用。

**摘要:** 金融知识图谱是各种智能金融应用的关键技术之一。除了知识图谱领域普遍的技术挑战之外, 构建金融知识图谱还需要跨越低质量数据源与高精度图谱之间的鸿沟: 一方面, 多数金融领域数据的形态是充满噪音的 pdf 文档、ocr 扫描件和网页; 另一方面, 金融领域的许多应用又要求图谱中的数据非常精准。本报告结合文因互联在智能金融领域的实践, 介绍如何综合运用多种技术跨越这一鸿沟。其中, 语言模型、变体词发现、实体消歧等 NLP 技术发挥了重要的作用。在应用层面, 我们将介绍这些技术在自动化监管、自动化审计、咨询、投资等领域的案例。

## 基于产业知识图谱的公司金融创新

任亮 (知因智慧)

时间: 09:30-10:00

**简介:** 任亮, 北京知因智慧数据科技有限公司创始人 &CEO, 中科院大数据分析技术实验室副主任 (兼) 教授。知识图谱、机器学习、金融风险领域资深专家, 历任中国科学院教授、中国科学院大数据分析技术实验室副主任、中国科学院硕士研究生导师等职务。拥有近 20 年数据挖掘在客户风险管理、整合营销领域的项目实践及管理经验, 先后帮助中软、宇信科技等知名科技企业筹建金融风控部门并担任负责人; 任职 IBM 期间作为 GBS 全球企业咨询服务部金融大数据管理解决方案负责人亲自参与了一系列金融模型和解决方案的研发管理工作。此外, 在学术研究方面亦多有建树, 中国信用风险等相关论文受到产业界的高度评价, 是行业内为数不多的集产、学、研经验于一身的企业管理者。

**摘要:** 利用当前银行现实的数据和外部资源, 凭借知识图谱, 机器学习等前沿的人工智能技术, 在公司金融领域进行营销获客, 风险控制等业务创新; 分享近一年来在标杆金融机构的领先实践经验。

## 对话式交互的现状与未来

王守昆 (爱因互动)

时间: 10:30-11:00

**简介:** 王守昆先生现任爱因互动科技发展有限公司创始人、CEO。爱因互动成立于 2016 年 6 月, 致力于利用人工智能和自然语言处理技术向企业客户提供业界领先的对话服务, 帮助客户提升销售转化, 节省运营成本, 改善用户体验。在创办爱因互动之前, 王守昆先生曾经作为 CTO 和 CEO 参与了在线教育公司微学明日的创立和管理, 负责整体运营。在此之前, 王守昆先生在豆瓣网担任首席科学家和副总裁, 负责豆瓣网整体算

法架构设计和实施。王守昆先生毕业于清华大学自动化系, 分别于 1999 年和 2002 年获工学学士和工学硕士学位。

**摘要:** 近几年来, 随着人工智能浪潮的兴起, 对话式交互再次进入大众的视野。对话式交互 (CUI) 与图形界面交互 (GUI) 有着怎样的异同、对话式交互当前在个人和企业级市场有着怎样的应用、对话式交互能否真正重塑软件的使用和开发流程, 以及作为人工智能皇冠上的明珠的自然语言处理技术, 在对话式交互领域有着怎样的应用, 本文将从一个实践者的角度, 一一阐述作者对这些问题的思考。

## 知识图谱技术以及它在不同垂直领域中的应用

李文哲 (*Greedy Tech*)

时间: 11:00-11:30

**简介:** 美国人工智能公司 Greedy Tech 的创始人兼 CEO, 前凡普金科集团 (爱钱进) 的首席数据科学家、北京会牛科技的首席科学家兼投资总监、美国亚马逊和高盛的高级工程师。美国 USC 人工智能博士、先后在各类顶级会议上 (KDD、AAAI、AISTATS) 发表过 15 篇以上的论文, 其中三篇荣获了 Best Paper Award. 兼任多家中美创业公司和风投机构的技术顾问。

**摘要:** 主要讲解知识图谱技术 (推理技术、存储技术) 以及它在不同领域中的应用, 会结合工业界的实践案例。内容的大纲为: 1. 知识图谱的介绍; 2. 知识图谱解决的核心问题。; 3. 知识图谱技术在不同垂直领域 (金融科技、教育科技、电商、物联网) 中的应用以及实践。; 4. 结合规则以及深度学习的推理技术; 5. 未来技术展望以及挑战。

## 知识图谱面向金融行业的落地实践

孟嘉 (明略数据)

时间: 11:30-12:00

**简介:** 孟嘉, 明略数据技术合伙人, 大数据关系挖掘分析平台 SCOPA 的总架构师与负责人。2014 年底加入明略数据, 负责大数据关联分析平台 SCOPA 的研发与架构设计, 经历了 SCOPA 产品从 0 到 1 的过程, 见证一个新的产品如何一步步通过实际项目打开市场, 并帮助公共安全和金融行业客户解决实际问题。

**摘要:** 知识图谱技术在金融行业可以应用于诸多场景。明略数据凭借多年行业经验, 为金融机构构建金融风控和监管领域的金融知识图谱, 将海量数据治理成“企业、个人、机构、账户、交易、以及行为数据”等实体, 并基于知识图谱数据库进行高效存储和管理, 为金融客户发现隐藏在复杂网络之下的风险关系网络和资金异常流动情况, 全面提升风控专家在海量数据下, 精准甄别、避免监管套利, 提高合规审查的效率, 落地企业级 AI 服务, 切实支撑业务, 与金融机构客户并肩实现产业智能化升级。

## Model-driven Deep Learning

孙剑 (西安交通大学)

时间: 09:00-09:30

**简介:** 孙剑, 西安交通大学数学与统计学院信息科学系教授, 2009 年获得西安交通大学应用数学博士学位。主要关注视觉信息处理与分析中的数学模型与算法研究, 相关成果发表于 IJCV, IEEE TIP, CVPR, NIPS, MICCAI 等领域内著名国际期刊和会议。曾经在微软亚洲研究院 (2005-2008)、美国中佛罗里达大学 (2009-2010)、法国巴黎高等师范学院与法国国家信息与自动化研究院 (2012-2014) 做博士后或访问学者, 并在香港浸会大学、美国约翰霍普金斯大学、纽约大学等进行短期学术访问。担任 CVPR, MICCAI, ECCV 等高水平国际会议程序委员会委员, IJCAI-18 高级程序委员会委员。入选教育部新世纪优秀人才支持计划 (2012)、国家基金委优青项目 (2016), 研究成果获得中国工业与应用数学学会优秀青年学者奖。2017 年, 入选中组部万人计划“青年拔尖人才支持计划”。

**摘要:** 经典的深度学习方法将标准深度神经网络作为黑箱进行数据驱动的目标任务学习。我们提出模型驱动的深度学习思想, 将传统的基于领域知识或物理机制的建模方法与深度神经网络的数据驱动学习能力相结合, 构建模型驱动的深度学习方法。在该报告中, 将展示我们在模型驱动深度学习方法上的一些研究成果, 包括统计模型驱动的深度学习方法、ADMM 优化算法驱动的压缩传感深度神经网络、learning to learn 元学习算法等。并展示它们在图像处理与分析、深度神经网络优化中应用中的有效性。

## On Tensor Networks and Neural Networks

徐增林 (电子科技大学)

时间: 09:30-10:00

**简介:** Zenglin Xu (徐增林) is a Professor in School of Computer Science and Engineering at University of Electronic Science and Technology of China(UESTC). He is the founder and director of the Statistical Machine Intelligence and LEarning (SMILE) Lab. He is a recipient of China Thousand Talents(Youth) Program. He obtained his PhD in Computer Science and Engineering from the Chinese University of Hong Kong. His research interest includes machine learning and its applications on social network analysis, health informatics, and cyber security analytics. He has published over 70 papers in prestigious journals and conferences such as NIPS, ICML, IJCAI, AAAI, IEEE PAMI, IEEE TNN, etc. He is also the recipient of the APNNS young researcher award, and the best student paper honorable mention of AAAI 2015. Dr. Xu has been a PC member or reviewer to a number of top conferences such as NIPS, ICML, AAAI, IJCAI, etc. He regularly servers as a reviewer to IEEE TPAMI, JMLR, PR, IEEE TNN, IEEE TKDD, ACM TKDD, etc.

**摘要:** In the big data era, multiway data are almost everywhere, e.g., recommendation systems, face recognition, sensor networks, etc. Tensor factorization is an important approach to multiway data analysis. The speaker will first briefly introduce canonical methods as well as recent developments of tensor factorization. Then, the speaker will discuss the connections between tensor networks and deep neural networks, and especially how to compress deep neural networks with tensor networks.

## 电商对抗智能 Adversarial Intelligence in E-commerce

储晨 (阿里巴巴)

时间: 10:30-11:00

**简介:** 储晨, 博士, 毕业于中国科学技术大学少年班。现就职于阿里巴巴搜索事业部。研究方向包括, Anomaly Detection, Adversarial Machine Learning, Robust Recommender System 以及机器学习算法在图引擎上的加速等。

**摘要:** 搜索和推荐是电商两大重要的引导成交场景, 它们在诞生的第一天就受到各种攻击的威胁。而这种攻击不同于传统安全领域如 DDOS 等攻击, 是针对模型的攻击, 以攻击的方式使得目标商品获得更高的搜索排名和更多的流量。《电商对抗智能》旨在从系统层面——离线到实时, 从算法层面——图模型到深度学习以及从应用层面——识别到调控, 阐述阿里巴巴电商环境的 Adversarial Intelligence。

## Towards Practical Deep Learning Model Compression and Acceleration

王乃岩 (*TuSimple*)

时间: 11:00-11:30

**简介:** Naiyan Wang (王乃岩) is currently the principal scientist of TuSimple. He leads the algorithm research group in our Beijing branch. Before this, he got his PhD degree from CSE department, HongKong University of Science and Technology in 2015. His research interest focuses on applying statistical computational model to real problems in computer vision and data mining. Currently, he mainly works on the vision based perception and localization part of autonomous driving. Especially he integrates and improves the cutting-edge technologies in academia, and make them work properly in autonomous truck.

**摘要:** Deep neural networks have liberated its extraordinary power on various tasks. However, it is still very challenging to deploy state-of-the-art models into real-world applications due to their high computational complexity. In this talk, I will start with the background of deep model compression and acceleration, and discuss the practical aspect of this technique. Then I will introduce three recent works done in TuSimple by novel techniques in model distillation and sparse model structure selection. By combining these techniques, we can build a fully automatic pipeline for joint model training, performance boosting and model acceleration. These works all demonstrate superior performance in practice, and have been deployed in TuSimple's production.

## 在生产环境中使用 R 语言构建机器学习项目

黄庆昌 (派生集团)

时间: 11:30-12:00

**简介:** 派生集团数据中心数据挖掘部高级经理, 负责各业务线的营销模型、风控模型等建模工作。

**摘要:** 本次演讲主要介绍如何在生产环境中构建一个以 R 为主要开发语言的机器学习项目, 以及在实践过程中需要注意的一些关键步骤和要点。希望提供一些利用 R 语言解决实际生产环境问题的思路。

## 脑健康研究中的数据挖掘

杨志 (上海精神卫生中心)

时间: 08:30-09:00

**简介:** 杨志, 博士, 研究员, 上海市精神卫生中心心理健康与脑影像研究室 PI。曾任中国科学院心理研究所副研究员, 博士生导师, 中科院大学岗位教授。研究方向为脑功能影像数据挖掘及精神障碍的神经影像标志, 发表第一/通讯作者 SCI 论文 20 篇, 获国家发明专利一项, 主持国家自然科学基金三项, 研究成果获得教育部科技进步一等奖、北京市科技进步二等奖。

**摘要:** 脑的健康发展是心理健康和认知能力提升的物质基础。美国、欧洲、和我国先后启动的“脑研究计划”大力推动了对脑的工作机制和发展路径的研究。脑成像、脑电和行为研究是人类脑健康研究的最主要研究手段。就像“生物信息学”引爆了基因研究一样, 应用数据挖掘方法从以上脑数据中获得更丰富的信息将对脑研究领域做出重要贡献。演讲者将展示数据挖掘在脑研究中的作用, 并从脑科学的研究者角度提出对数据挖掘方法的需求。

## 基于项目反应理论的汉字识别在线测验的编制与应用

李燕 (北京师范大学)

时间: 09:00-09:30

**简介:** 李燕, 目前就职于北京师范大学心理学部的应用中心, 担任项目研发人员, 目前主要负责测验数据分析和文本分析。硕士毕业于英国伦敦大学学院, 有 3 年的 R 语言使用经验。

**摘要:** 传统心理学与教育评价研究依赖于纸笔测验, 如何大规模、准确、可重复地测量学生的素养和发展不容易。项目反应理论和计算机自适应测验技术的发展, 为评估学生情况提供了新的方法, 也带来了新的挑战。我们对小学生的汉字识别能力开展测查, 采用了项目反应理论、共同锚题设计和半自适应组卷的方法研发一套能够垂直等值和比较、可靠、稳定的汉字识别在线测验, 并建立北京市小学生识字量常模和一个可用于计算机自适应测验的项目测验库 (Item bank), 为今后的相关研究提供重要的研究工具。此外, 研究者还利用 R 语言中与计算机自适应测验相关的 catR 包模拟了计算机自适应测验, 取得了较好的结果。未来研究中, 本汉字识别在线测验不仅利于教师和家长跟踪和预测儿童的阅读能力发展, 还能作为区分发展性阅读障碍的重要指标, 利于教师提供有针对性的阅读干预。

## 心理学研究中的可重复性 –以神经成像偏见效应为例

殷继兴 (西北师范大学)

时间: 09:30-10:00

**简介:** 殷继兴, 现就读于西北师范大学心理学院, 目前主要关注的研究领域有神经成像偏见效应和心理学的可重复性危机, 并致力于学习如何使用 R 语言进行心理学研究数据的统计分析和提高研究的可重复性。

**摘要:** 可重复性是衡量研究效应是否存在的重要标准, 但近年来, 不少科学领域出现了“可重复性危机”——大量发表的研究无法重复。心理学对可重复性问题也越来越关注。在此背景下, 本研究对神经成像偏见效应的可重复性进行评估。神经成像偏见指的是神经成像可能会让外行解读出其结果以外的内容, 进而影响

判断和决策。早期的研究发现神经成像证据的呈现会过高地影响外行的决策, 这些研究被广泛报道并影响到一些现实的案件。但后续的研究则大多未能重复出这一结果, 或者发现效应只在某些特定的情况下出现。那么该效应是否存在? 本研究对该效应的相关文献首先进行系统回顾, 然后采用元分析方法对其效应量进行量化的估计, 并通过 p-curve 来探索文献中的 p-hacking 现象。此外, 我们还通过一个预注册的研究来重复该效应。本研究的结果表明, 神经成像偏见很可能不存在或者效应量小到可以忽略。通过证否一个被广泛接受的效应, 本研究表明, 可以通过多种方法来判断一个效应的可重复性。同时, R 语言的使用能够促进本研究的可重复性。

## 日间节律对睡眠缺失效应的调节机制——一项探索性研究

麦子峰 (华南师范大学)

时间: 10:30-11:00

**简介:** 麦子峰, 华南师范大学心理学院本科基地班大三学生, 跟随马宁教授进行睡眠方面的研究, 主要研究内容包括急性睡眠剥夺对个体认知与情绪功能的影响, 以及节律对睡眠缺失的调节机制。

**摘要:** 节律, 指个体生命活动随时间的推移而呈现出规律的变化, 前人研究发现睡眠缺失效应容易受到日间节律的调节, 但对于具体的调节机制, 尤其是中国人群中的规律, 尚未有研究涉及。因此, 本研究通过严密的实验室控制, 对比正常睡眠和急性睡眠剥夺两种状态下, 个体在不同时间点上主观疲劳感、客观行为表现、以及体温等心理与生理指标的差异。结果发现, 急性睡眠剥夺后个体的主观疲劳评分在下午一直维持在较高水平; 而个体在 Go/No go 任务中的辨别力则随着时间的推移而提升。研究揭示了节律对睡眠缺失后的个体在主观评定和客观行为表现上不同的调节作用, 即个体对自身状态的主观感知并没有明显变化, 但其认知表现会受节律调节从而呈现一定程度的恢复。

## 心理学研究统计的新趋势

甘怡群 (北京大学)

时间: 11:00-11:30

**简介:** 甘怡群, 北京大学心理与认知科学学院教授。1998 香港中文大学心理学系获得博士学位。发表了 90 余篇一作或通讯作者的研究论文, 论文发表于 SSCI 一区期刊 “Journal of Personality” 和 “Health Psychology”。主持多个国家级科研项目。目前任两个国际 SSCI 期刊 European Journal of Cancer Care 和 Journal of Pacific Rim Psychology 的副主编, 国际期刊 Applied Psychology: Health and Well-being, Stress and Health, PsyCH Journal 编委, 国内期刊《心理科学进展》, 《中国心理卫生杂志》, 《中国临床心理学杂志》编委, 同时也是 30 个国际性学术期刊的审稿人。研究领域集中在应激, 应对和心理健康方面。2016 年被中国心理学会认定为 “心理学家”。

**摘要:** 心理学研究的可重复性危机的部分原因来自零假设统计检验的运用。因此, 美国心理学会 (APS) 建议弱化零假设统计检验, 而推荐 3 种统计方法: 效应量, 置信区间和元分析。我们用实例说明了用置信区间和元分析的思维, 相比零假设统计检验的二分法思维从理解统计结果角度的优越性。这个报告介绍了如何用 SPSS 和 Excel 的宏程序计算效应量, 置信区间, 以及如何用 CMA 进行元分析。最后, 报告提出了心理学研究可重复性在统计学角度的新观点, 以及为避免低功效的统计, 事先估计样本量提出了建议和实际操作的方法。

## 统计之都 COStudy 心理学项目介绍

夏骁凯 (华南师范大学)

时间: 11:30-12:00

**简介:** 夏骁凯, 统计之都 COStudy 心理学项目临时负责人。现就读于华南师范大学心理学院, 攻读认知心理学博士学位, 主要研究方向为认知控制功能及其计算模型。致力于推广数据科学相关技术与 R 编程等在心理学领域中的使用。

**摘要:** COStudy 心理学作为 COStudy 的首期项目, 旨在促进数据科学与心理学的学科交流。由于心理学善于与多学科交叉的特征, 以及伴随着近些年来认知心理学、心理测量学等的快速发展, 心理学科内部越来越需要融合了计算机科学与统计学等的数据科学的力量。然而目前国内两个学科的合作仍然存在不小的隔阂。心理学学科融汇社会科学与自然科学的特性, 从而与数理专业沟通存在障碍。而其他学科对于目前心理学的发展现状往往知之甚少, 数据科学的专业人士对于与心理学合作的大量切入点尚未可知。我们希望通过该项目的工作, 针对心理学的学科特性, 将数据科学的相关知识介绍到心理学界, 并降低心理学专业人员使用相关技术的门槛。我们同时希望将心理学科的研究进展和方向, 更重要的是对数据科学的需求介绍到数据科学界, 以促成数据科学界与心理学界的合作。

## 基于 R 开发的针对慢性病多中心项目研究一站式平台

陈显扬 (北京骐骥生物技术有限公司)

时间: 09:00-09:30

**简介:** 陈显扬, 中国科学院生物学博士, 公司 CTO, 十几年生物学研发经验。发表 17 篇 SCI 文章和 6 项发明专利, 在生物信息学以及代谢组学研究、代谢病智能诊断方面颇有建树, 曾获罗氏青年科学家论坛金奖。

**摘要:** 慢性病多中心研究是循证医学一项重要内容, 一般分为回顾性研究和探索性研究。无论是哪种研究方式, 数据的处理是对临床医生极大的挑战。本平台第一次将病例回顾性研究, 以及代谢组学探索性研究相结合, 基于 R 开发的针对慢性病多中心项目研究一站式平台。针对病例数据, 我们利用 R server 进行的交互式的功能开发。可以通过平台导入数据库筛选的数据, 然后进行数据清洗, 并解决倾向性问题。随后, 可以根据需要开展 p 值比较, 描述性分析; 同时, 平台也提供病例研究的常用模型和功能分析: 生存分析, 功效分析, 逻辑回归, 随机森林等等, 并提供 ROC 和模型评价功能。特别的, 平台还具有适合于探索性代谢组学研究的数据处理系统, 并将回顾性数据和探索性数据进行结合, 从而得到更加准确的模型和预测结果。同时, 我们也给出两个案例, 利用 R 开发的一站式平台, 高效分析阿尔兹海默症新型分子标志物, 以及糖尿病并发症的预测模型。

## 面向药物重定位的多组学数据整合分析框架

何松 (军事科学院)

时间: 09:30-10:00

**简介:** 何松, 2013-至今军事科学院军事医学研究院博士生生物信息学专业。2009-2013 清华大学自动化系本科生。

**摘要:** 多组学数据融合算法为药物重定位也提供了新的机遇。在以往的研究中, 研究者往往只关注药物的某一方面的属性, 如临床医生更关注药物的副作用属性 (即临床属性), 药理学家更关注药物的药靶关联属性 (即药理学属性), 而化学家则更关注药物的结构属性 (即化学属性)。但药物并不因为研究者的视角而改变, 站在单一角度看待药物属性不可避免的会带来盲目摸象式的问题和噪声。将多组学数据融合算法应用于整合药物的多层属性也将指向更为准确的药物重定位。另一方面, 以往的药物重定位研究都是将药 - 靶, 药 - 病, 甚至是靶 - 病关联关系 “分而治之” 地割裂看待, 这样会带来噪声叠加的问题。例如, 预测的药 - 靶关联关系和靶 - 病关联关系都有一定程度的假阳性, 如果以靶标为中介, 关联药物和疾病, 那么在 “药 - 靶 - 病” 关系中, 只要有一条关联边是假阳性, 预测的药 - 病关联就是假阳性的。在本研究中, 我们提出了基于多组学数据融合的药 - 靶 - 病三元关联关系预测计算框架 PAMDF。通过整合药物四方面的属性、靶标四方面的属性以及疾病三方面的属性, 分别构建融合的药物相似性网络、靶标相似性网络和疾病相似性网络, 提出并利用三元异质网络上的重启随机游走算法预测 “药 - 靶 - 病” 关联关系。

## 一个肺癌大数据分析平台的建制

李舰 (统计之都)

时间: 10:30-11:00

**简介:** 李舰, 辅仁大学博士生, 统计之都的核心成员之一, 某医疗类创业公司合伙人, 兼任华东师范大学硕士生导师, 是 R 语言社区的活跃用户, 贡献了 tmcn、Rwordseg 等包, 著有《数据科学中的 R 语言》, 参与翻译了《R 语言核心技术手册》、《机器学习与 R 语言》。

**摘要:** 以国内某大型医院的真实项目为例, 介绍一个肺癌大数据分析的平台。包含了医院信息系统的整合、医疗数据的清洗、数据仓库的设计、分布式平台的建设、分析模块的开发、可视化的展现、业务上的应用等。尤其是分析模块层面, 重点介绍统计学、机器学习、商业智能在医疗数据分析中的应用, 以及深度学习和图像技术在医学影像诊断方面的应用。

## 半参数强分层整合分析的变量选择研究

李嵘 (中国人民大学)

时间: 11:00-11:30

**简介:** 李嵘, 本科毕业于中央民族大学理学院统计系, 2017 年起就读于中国人民大学统计学院, 目前是生物医学统计方向一年级研究生, 并逐渐跟随老师从事高维基因数据的变量选择问题。

**摘要:** 在许多医学相关研究中, 通常采用半参数模型来刻画基因与环境对疾病的影响。而基因对疾病的影响可能不仅仅是单个基因的存在带来的影响, 有时一些基因的共同存在会对疾病的患病风险造成更严重的影响, 故在疾病的研究中纳入交互效应将使研究的讨论更准确。在如此众多的危险因素中间, 如何识别、筛选出真正重要的因素具有重大意义。若针对同一个问题, 有许多不同来源的研究, 这些研究之间有一定的相关性, 同样也有来自不同来源的异质性。整合分析就是一个综合考虑多方面影响, 分析多个数据集的方法。为了解决半参数交互效应整合分析模型的变量选择问题, 我们提出了两种解决思路, 一种是基于罚函数的算法构建, 另一种为改进的 TGDR 方法。这两种解决方法均通过参数改写以及成组变量选择的方式、保证交互效应的分层结构以及同一变量在不同数据集中的一致性。罚函数因其良好的理论性质, 在准确性上占有一定的优势, 但是, TGDR 方法简单的操作思路大大缩短了其所用的时间, 在实际数据分析中是一种比较受欢迎的方法。我们通过设置不同的模拟场景, 讨论在何种场景下本模型设定优于各数据集分开建模的模型以及将各数据集一起整合的模型, 并比较在不同协变量结构、不同参数真值下, 不同模型设定的预测、估计和选择效果。除此之外, 还可比较在同样模拟场景的设置中, 两种不同解决方法的预测、估计和选择效果, 为实证分析选择解决方法提供一些参考和建议。最后, 通过皮肤黑素瘤数据建模分析, 得到了一些实际有用结论。

## 数据挖掘在隧道掘进机设计及分析的应用

石茂林 (大连理工大学)

时间: 09:00-09:30

**简介:** 石茂林, 大连理工大学博士在读生, 机械设计及理论专业。主要从事专业为工业大数据挖掘, 数据处理, 数据建模, 函数分析, 统计学习等。

**摘要:** 本演讲主要介绍了大连理工大学机械工程学院复杂机电系统创新设计及重大装备团队在工业大数据近两年的研究工作。内容分为两个方面, 一是基于工业大数据的装备载荷预测, 主要解决了非结构数据与结构数据的有机融合问题, 大数据的快速建模问题等; 二是提出了针对工业大数据聚类分析方法, 通过将属性相关先验信息引入聚类目标函数提升分类准确度, 以帮助后续的装备设计及分析。以上两个算法均在隧道掘进机上取得了成功应用, 是数据分析方法在工业领域的一个成功应用。

## 基于小波变换与循环神经网络的在线手写签名识别方法

刘欣益 (中国人民大学)

时间: 09:30-10:00

**简介:** 本人刘欣益, 来自中国人民大学统计学院应用统计专业一年级研究生。本科毕业于南开大学数学学院统计系。现与导师合作完成一篇利用循环神经网络的在线手写签名识别研究项目。

**摘要:** 在线手写签名由于其个体的独一无二以及难以模仿, 仍然被广泛地使用在日常生活与工作之中。本文受启发与卷积神经网络在序列建模中的成功应用以及小波变换对于序列特征提取的有效性, 构建了基于小波变换提取签名特征然后使用卷积神经网络对在线手写签名进行度量学习的模型。卷积神经网络学习的目的就是为了减少真签名之间的距离, 同时使得真签名与假签名的距离大于一个给定的阈值。本文对不同神经网络的结构, 特征的个数以及小波变换的基函数以及小波分解的层数均进行了实验。实验表明, 小波变换对于提取签名特征具有非常高的有效性同时也具有一定的稳定性。通过小波变换不仅能有效的提取复杂签名中的特征, 同时也能将样本的维度减少到原始的 25%-50%, 加速模型训练速度以及模型的收敛速度, 并且提升了模型的识别性能。不同的小波基以及小波分解的层数均会对最终模型识别的精度产生影响。另一方面由于神经网络三元组输入的设计结构, 能够在原始样本的基础上提升训练样本的个数, 一定程度上缓解了签名样本获取困难以及样本个数较少的问题。

## AI in Agriculture

任乔牧 (南京农业大学)

时间: 10:30-11:00

**简介:** 任乔牧, 南京农业大学工学院三年级本科生, 中国计算机学会、中国人工智能学会、中国中文信息学会会员, 江苏省智能化农业装备重点实验室成员, 南京农业大学工学院大数据实验室机器学习组组长, 南京农业大学工学院嵌入式物联网创新工作室技术总监。曾在东南大学认知智能研究所从事知识图谱、问答系统、自然语言处理等方面的研究, 目前在东南大学从事计算机视觉、计算摄像学方面的研究, 同时在南京农业大学工学院大数据实验室从事机器学习、计算机视觉和农业工程的交叉研究。累计参与各类科研项目十余

项, 累计在国际级、国家级等各类科技竞赛中获奖十余项, 曾作为 NAU-China 软件队负责人前往 Boston 参加 International Genetically Engineered Machine Competition 的汇报展示并获得 Gold Medal, 另有多篇论文在投。感兴趣方向包括但不限于机器学习、自然语言处理、计算机视觉及其在农业、医疗等领域的应用。

**摘要:** 随着机器学习、计算机视觉、自然语言处理等领域的发展, 人工智能技术被广泛地应用到各行各业中, 目前人工智能已经在医疗、交通等领域发挥了巨大作用。与此同时, 全球人口快速增长、自然资源不断被消耗以及城市化的快速发展, 都对农业提出新的需求。将机器学习、计算机视觉、自然语言处理等应用在农作物病害检测、农业问答系统等场景, 使得更高效、更精准的农业成为可能。本报告结合目前机器学习、计算机视觉、自然语言处理、植物表型等领域的研究成果和南京农业大学有关实验室的最新工作, 介绍机器学习、计算机视觉等的发展以及它们在农业领域的应用, 展望人工智能如何更好地推动农业的发展。

## AI 技术与智能投放机器人

游皓麟 (*Tap4Fun*)

时间: 11:00-11:30

**简介:** 游皓麟, 数据挖掘专家, 目前就职于 tap4fun, 专注人工智能算法研究以及智能投放机器人的研发落地。曾服务于华为技术软件有限公司等企业, 多次出席 R 语言会议并作为重要嘉宾发表演讲, 在小象学院担任过 R 语言数据挖掘和机器学习讲师, 著有《R 语言预测实战》。

**摘要:** 随着 AI 技术日新月异的发展, 广告投放领域也迎来了新的机会。AI 技术除了在图像、语音、自动驾驶等领域风生水起之外, 在广告投放领域也有其很大的发挥空间。与传统投放方式有所不同, 在 AI 技术的加持下, 广告投放正变得更加智能、更加有趣, 基于 AI 技术实现的智能投放机器人, 可以在一定程度上替代人工, 并且具有更好投放效果。本次分享, 主要聚焦 AI 技术的新进展, 新趋势和新成果, 同时分享广告投放领域的智能投放机器人实际案例, 希望与业界的朋友, 多交流, 多讨论, 多碰撞。

## 慎买与惜售——基于幂律的羊群效应检测方法及其量化策略

杨晟 (江西财经大学)

时间: 11:30-12:00

**简介:** 杨晟同学本科毕业于中央财经大学, 研究生现就读于江西财经大学金融学院。曾参加过“第十六届(2017年)中国金融工程学年会暨金融创新与风险管理(国际)论坛”和“2017 厦门大学金融工程与量化金融学术会议”, 并在会议上宣读了论文。曾获全国统计建模类研究生组二等奖, 首届河南省高校量化投资模拟大赛二等奖。

**摘要:** 本文通过研究慎买与惜售的羊群效应, 推导出了使用幂指数来度量慎买惜售的指标, 并使用中国股市 15 分钟数据进行了实证检验。实证表明, 中国股市整体上存在慎买惜售效应, 慎买效应显著大于惜售效应, 且这种效应在股市的不同时期也存在显著的不同。最后, 我们根据慎买惜售效应对方正证券金融工程部设计的“聪明钱策略”进行了改进, 改进后的交易信号准确度和策略收益率都得到了显著的提升。

## 开源的可视化数据分析平台 Kibana 从 0 到 100

曾勇 (*Elastic*)

时间: 09:00-09:30

**简介:** 曾勇 (Medcl), Elastic 技术布道师, 在分布式搜索、高性能、高可用架构、自动化运维等方面积累多年的经验。Elastic 开源社区负责人。Elastic 在中国的第一位员工。阿里云 MVP。开源软件爱好者。

**摘要:** Kibana 是时下非常新颖的一个开源的数据分析和可视化平台, 可以提供各种便利的分析能力和可视化展现效果, 本次分享将主要介绍 Kibana 的发展历史和具体的使用方法。演讲主题包括以下几个方面: 1. Kibana 项目历史; 2. Kibana 基本概念; 3. Demo 演示; 4. Kibana Canvas; 5. Kibana Geo 地理位置分析。

## R 语言的可视分析应用

王修坤 (阿里巴巴)

时间: 09:30-10:00

**简介:** 王修坤, 毕业后就职于阿里巴巴, 工作三年, 高级数据工程师。主要工作方向是基于机器学习算法的风控领域研究, 业余工作是基于可视分析的数据挖掘方法研究, 主要基于 shiny+d3, plotly, highcharts 等工具, 目前完成过《杭州市房价分析》、《基于可视化的论坛文本聚类》、《文本挖掘平台开发》等案例。

**摘要:** 可视分析相比较可视化存在差异, 可视分析的对象是分析师, 它通过可视组件和交互等元素来挖掘复杂的数据形态, 包括结构化和非结构, 所以我们的可视分析研究主要旨在探索利用可视化角度的数据挖掘方法。

第一部分: 介绍几个经典的可视分析案例; 第二部分: 常用的分析方法及组件选择, 从简单的 linePlot、barPlot 到负责的。apPlot、networkPlot; 第三部分: 工具介绍及案例: 《杭州市房价分析》、《基于可视化的论坛文本聚类》。

## 可视化中的静与动

郎大为 (*J.D. Power*)

时间: 10:30-11:00

**简介:** J.D. Power 资深数据分析师, 主要方向为汽车行业的数据咨询。浙江大学软件学院, 华东师范大学校外导师, 统计之都编辑部成员, 一个被可视化耽误的分析师, wordcloud2, REmap, leafletCN 等包的作者, recharts, RWeixin 等包的维护者。

**摘要:** R 语言的图形展示功能已经广泛得到了学界和业界的一致认可。可视化, 一个为了更好的理解数据, 展示数据的技能, 也逐渐被数据相关从业者列为一项必备技能。然而, 习惯了传统的静态可视化之后, 各类数据玩家又被更炫酷的动态可视化所吸引, 但复杂的概念, 众多的专有名词和计算机背景的知识更让学习动态可视化这个过程难上加难。正如统计之都某位远古大神介绍所说: “在可交互图形的实现机制上, 基于浏览器作为图形展示平台, 利用 javascript 作为图形绘制和交互基础的机制成为主流方向。htmlwidget 作为 RStudio 的最新发布的 R 包, 致力于将优秀的基于 javascript 可视化包, 结合 R 语言语法和 R 用户的习惯, 形成 R 层面的代码封装……”对于分析师, 可以选择像 PowerBI, tableau 这样的商业软件。但除此之外, R 语

言中的如 recharts, REmap 等包, 同样有一条动态可视化的速成之路。哪怕是段简单 R 的代码, 也可以开启一段动态可视化的旅程。在这条路上, 可视化的静与动, 其实并没有太多的障碍。

## 基于比特币交易的可视化分析

周宁奕 (众安保险)

时间: 11:00-11:30

**简介:** 周宁奕前建筑设计师, 前阿里巴巴 datav 成员, 主攻 WebGL、WebGIS, 现众安保险数据科学实验室可视化方向负责人, 独立应用糊涂作者。

**摘要:** 相比传统金融数据, 比特币的交易和区块链的数据更公开, 某些方面有更多的来源和更多的细节。本次讲座基于抓取的比特币数据, 通过众安保险的可视化产品 zatlas, 和用户现场演示数据的结构和分析的方法

- a. 数据获取 1. 虚拟币的运行机制 2. 交易所数据的获取 3. 区块链数据的获取
- b. 数据分析 1. 虚拟币的版块 2. 流动性分析 3. 交易网络分析 4. 相似度分析 5. 竞争分析 6. 价差分析
- c. zatlas 可视分析产品 1. 数据源导入 2. 图表组件 3. 发布分享 4. 更多的功能

## R 语言 ggplot2 之地理信息可视化

张杰 (香港理工大学)

时间: 11:30-12:00

**简介:** 香港理工大学 Research Assistant, Excel 教程《Excel 数据之美》作者; Excel 图表插件 EasyCharts 开发者; 十余篇 SCI 论文的水货达人; 微信公众号 EasyCharts 联合创始人; 预计 2018 年下半年出版《R 语言数据可视化之美》

**摘要:** 本次演讲重点讲解 R 语言基于 ggplot2 包的地理信息可视化, 先介绍不同的地图投影模式, 讲解世界地图、美国和英国等世界各国、中国(包括省级、市级到县级不同的行政单位)、局部地图等, 地图数据的获取与绘制, 特别会讲解标准中国和美国地图的绘制; 再接着讲解不同的地图类型, 包括等值区间地图、带散点、气泡、柱形、饼图和连接线的地图、等位地图、地铁线路图等。

## A New Constrained L1 Minimization Approach to the High-dimensional Markowitz Portfolio Optimization Problem

张振 (南方科技大学)

时间: 09:00-09:30

**简介:** 张振, 南方科技大学数学系副教授, 研究兴趣包括偏微分方程数值解, 多尺度建模, 和高维数据分析。2007 年本科毕业于中国科学技术大学数学系, 2013 年毕业于香港科技大学数学系, 获博士学位, 研究方向为计算数学, 博士毕业论文获得香港数学学会最佳博士论文奖。2013-2015 年在新加坡国立大学数学系从事博士后研究。2015 年至今工作于南方科技大学数学系, 2017 年入选第十三批中组部“千人计划”青年项目。目前是广东省工业与应用数学学会常务理事。在高维数据分析特别是 L1 正则化方面有一定研究, 与郭建华和荆秉义教授共同合作针对一类高维生物信息数据进行稀疏正则化建模和分析, 工作发表在 Journal of the American Statistical Association 上。近期研究主要在高维数据线性判别分类的降维以及其在 Markowitz 资产投资优化问题上的应用。

**摘要:** Covariance matrix and mean vector of asset returns are two important ingredients in portfolio optimization under the theory by Markowitz. The most traditional way to estimate the optimal portfolio weights is to plug in the sample mean and sample covariance matrix. However, the out-of-sample performance of such estimator is bad, especially in the case where the number of stocks is large compared to the sample size. Recently, we propose an estimator of the mean-variance optimal portfolio, which we call the Linear Programming Optimal (LPO) portfolio. We show that such portfolio can asymptotically yields the maximum expected return and meanwhile satisfies the risk constraint. Moreover, the LPO problem could be solved easily using DASSO algorithm, which shares similar properties as LARS algorithm and was initially proposed for solving Dantzig selector. Numerical results by R show that our method produces better performance compared to other commonly used approaches.

## 不平衡数据下故障诊断算法综述

刘杰 (北京航空航天大学)

时间: 09:30-10:00

**简介:** 刘杰, 北京航空航天大学卓越百人计划副教授, 博士。2015 年 2 月 -2017 年 6 月在巴黎萨克雷大学以博士后身份进行研究工作。2017 年 9 月加入北京航空航天大学可靠性与系统工程学院。目前已发表学术论文二十余篇。其主要研究方向为基于机器学习方法的故障诊断、故障预测和健康管理。代表性成果包括故障漂移模型下高效在线学习模型、基于局部数据特征的不平衡数据下故障诊断算法等。

**摘要:** 不平衡数据指的是某些类样本数量远远小于其他类。具有少量样本的类称为少数类, 而具有大量样本的类称为多数类。许多实际的工程应用中都存在不平衡数据集, 比如欺骗信用卡监测, 医疗诊断, 信息检索, 文本分类等, 其中对于少数类的分类准确率更为重要。不平衡数据严重影响传统数据驱动模型的故障诊断准确度。目前针对不平衡数据的故障诊断算法主要分为三类: 一是采样方法, 即通过对原始数据的采样平衡不同类数据之间的差异; 二是改进现有算法, 最常见的方法是通过代价敏感模型提高少数类样本的错分代价, 进而提高故障诊断准确率; 三是集成方法。本报告将对不平衡数据的产生原因以及其对传统数据驱动模型的影响进行介绍, 并详细阐述针对不平衡数据的解决方法和目前面临的挑战。

## 关于微生物群落的 SDE 动态建模

徐铣明 (南开大学)

时间: 10:30-11:00

**简介:** 徐铣明, 南开大学统计研究院助理教授, 南开大学百名青年学科带头人。2012 年博士毕业于多伦多大学统计学系。主要研究方向为稳健统计学以及其在生物信息学中的应用, 具体包括复合似然函数, 肠道微生物组与疾病的关系, 微生物群落的动态结构, 基于多序列的进化树构建等。主持国家自然科学基金青年项目一项。担任 Bernoulli , Annals of Applied Probability 等国际期刊的审稿人。

**摘要:** 关于肠道菌群 (gut microbiome) 及其与人体健康之间的研究是近几年生物和医学领域的热点, 研究表明肠道菌群与肥胖以及包括肠炎, 糖尿病, 直肠癌在内的多种疾病有着紧密的关系。但是, 因为数据获取的难度和数据本身的复杂度, 有关肠道菌群结构随着时间和外部因素 (treatment) 改变而发生动态变化的研究还比较少。本研究采用随机偏微分方程 (SDE) 对菌群的动态结构进行建模, 并使用 R 语言对所提出的方法进行了随机模拟试验和真实数据的分析。

## An ARIMA Model with Adaptive Orders for Predicting Blood Glucose Levels and Hypoglycemia

黎磊 (北京航空航天大学)

时间: 11:00-11:30

**简介:** 黎磊, 北京航空航天大学可靠性与系统工程学院系统工程专业博士在读, 研究兴趣包括应用统计、信号处理以及时间序列分析。博士课题主要关注基于实时血糖监测数据的糖尿病患者血糖控制。

**摘要:** The Continuous Glucose Monitoring System (CGMS) is an effective tool which enables the users to monitor their blood glucose (BG) levels. Based on the CGM data, we aim at predicting future BG levels so that appropriate actions can be taken in advance to prevent hyperglycemia or hypoglycemia. Due to the time-varying non-stationarity of CGM data, verified by Augmented Dickey–Fuller (ADF) test and Analysis of Variance (ANOVA), an Autoregressive Integrated Moving Average (ARIMA) model with an adaptive identification algorithm of model orders is proposed in the prediction framework. Such identification algorithm adaptively determines the model orders and simultaneously estimates the corresponding parameters using Akaike Information Criterion (AIC) and least square estimation (LSE). A case study is conducted with the CGM data of diabetics under daily living conditions to analyze the prediction performance of the proposed model together with the early hypoglycemic alarms. Results show that the proposed model outperforms the adaptive univariate model and ARIMA model.

## Likelihood ratio-based distribution-free sequential change-point detection

周茂袁 (中国民航大学)

时间: 11:30-12:00

**简介:** 周茂袁 (中国民航大学), 蓝天青年学者, 2013 年获南开大学统计学博士学位, 硕士生导师, 现任统计学系副教授。研究兴趣: 流数据的实时监控和在线分类。多次受邀作国际会议邀请报告。主持或参与国家自然科学基金多项, 发表多篇 SCI 期刊论文, 其中包括 *Journal of Statistical Computation and Simulation*、*Quality and Reliability Engineering International*、*OPERATIONAL RESEARCH* 等。为 “*Journal of Statistical Computation and Simulation*”, “*Operational Research*”, “*Computers & Industrial Engineering*”, “系统科学与数学” 等 SCI 或北大核心期刊审稿人。

**摘要:** Most existing control charts are for monitoring location or scale parameters, rather than any change in process distribution such as shift in shape. Goodness-of-fit (GOF) test can detect any change in distribution. This paper develops a new distribution-free control chart by integrating a powerful two-sample nonparametric likelihood ratio GOF test into the effective change-point model. Our proposed chart is easy in computation, convenient to use, and very efficient in detecting any change in process distribution, including shifts in location, scale, and shape. It is also robust in detecting various magnitudes of shifts and especially powerful in monitoring any distributional change involving a decrease in scale.

## Deep Chit-chat: 教机器人如何唠嗑

吴俣 (北京航空航天大学)

时间: 14:00-14:30

**简介:** 吴俣, 北航 -微软亚洲研究院联合培养博士生。参与了微软多款聊天机器人相关产品的开发, 并在近两年来在 ACL、AAAI 等国际顶尖学术会议和期刊发表关于对话机器人的论文近 10 篇, 并担任 ACL, COLING 等会议对话系统和人机交互领域审稿人。

**摘要:** 本报告主要介绍闲聊导向对话机器人技术前沿进展。首先, 本报告介绍检索模型和生成模型的聊天机器人算法发展, 以及各大聊天机器人比赛结果和相关经验。之后, 本报告讨论如何构建聊天机器人的技术护城河, 深耕聊天机器人的相关技术, 让公司在白热化的聊天机器人竞争中在技术领域取得业界领先地位。

## 人工智能的情感交互与行业融合

翁嘉硕 (竹间智能)

时间: 14:30-15:00

**简介:** 纽约州立大学计算机硕士毕业, 熟悉算法、编程语言、搜索引擎、网络安全以及邮件安全, 使用过的语言超过 35 种。作为 AI 领域的技术专家, 他带领团队负责竹间在 AI 领域产品研发与技术规划, 领域主要涵盖对话机器人、计算机视觉、金融科技等领域。此前, 翁嘉硕在中国大陆及台湾的多个科技类创新企业担任 CTO、首席架构师等职位, 带领团队进行 AI 及大数据领域的研究开发。

**摘要:** 自然语言理解与情感识别技术的结合, 使对话机器人能提供拟人化服务与对话体验, 并在商业与行业应用中找到越来越多的落地场景。

## 腾讯 AI Lab 的开放域智能对话研究

史树明 (腾讯 AI Lab)

时间: 15:00-15:30

**简介:** 史树明博士 2016 年 10 月加入腾讯, 任人工智能实验室 (AI Lab) 自然语言处理中心研究主管, 主要研究方向为语义理解和智能人机交互。他在 ACL、EMNLP、WWW、SIGIR、CIKM、AAAI 等国际会议上发表论文 30 多篇, 曾多次担任 ACL、EMNLP、WWW、AAAI 等会议的程序委员会委员以及 TOIS、TKDE 等期刊的审稿人。他毕业于清华大学计算机科学与技术系, 加入腾讯之前曾任职于微软亚洲研究院 (主管研究员) 和阿里巴巴集团 (资深算法专家)。

**摘要:** 本报告主要介绍腾讯 AI Lab 在开放域人机对话方向上的一些思考和实践。首先简要分析开放域人机对话的难度与挑战, 接着重点介绍腾讯 AI Lab 在文本理解和文本生成方面的研究进展, 以及如何利用文本理解和文本生成技术来提升对话系统。

## 语义识别的商业破局

马宇驰 (三角兽)

时间: 16:00-16:30

**简介:** 马宇驰, 三角兽科技三位创始人之一, 董事长 &COO, 融资、市场、品牌管理专家, 企业运营经验丰富。连续创业者, 曾建立 2 家公司, 公司运营经验丰富。2010 年创建的品牌营销公司, 曾为 24 券、优众网、沃尔沃、恒信钻石、凯迪拉克等提供品牌策略和品牌营销方案。2015 年参与创建 O2O 公司, 获得徐小平投资。企业品牌和市场专家, 在国际一线企业近 10 年经验。曾在 Viacom、奥美公关、Amway China 等国际巨头公司, 负责广告、公关和企业品牌工作。为可口可乐、统一等客户提供多年广告投放全案策划。在奥美公关作为 Intel 笔记本处理器公关负责人, 主推笔记本处理器“酷睿”系列。在 Amway 负责企业品牌全国户外投放, 建设品牌视觉资料库, 和企业文化在店铺和展览馆的推广。

**摘要:** 2017 年人工智能高速发展, 成为国家战略, 自然语言理解和人机交互是其中充满想象力的重要领域, 语义技术在当下已经可以满足一些应用场景, 商业化落产生价值, 真实的用户交互数据也将更快促进技术的发展。三角兽在语义理解、开放域聊天、多轮对话和跨域中控四个方面领先国内, 将分享其自然语言理解和人机对话技术的研发和商业化落地。

## 保险行业的对话机器人技术

吴金龙 (爱因互动)

时间: 16:30-17:00

**简介:** 2010 年获得北京大学数学院计算数学专业博士学位, 期间研究方向为推荐系统中的协同过滤算法。毕业后加入阿里云, 主要从事 PC 和云手机的输入法开发。2011 年加入世纪佳缘, 负责世纪佳缘用户推荐系统的开发。作为世纪佳缘资深总监, 领导世纪佳缘技术部, 负责佳缘数据和 AI 相关的各项工作, 并负责开发了中文对话机器人 (bot) 创建平台『一个 AI』(<http://www.yige.ai>)。因为坚信对话交互是未来趋势, 2017 年初我离开世纪佳缘, 以技术合伙人身份加入爱因互动 (<https://einplus.cn>), 负责算法部门工作。个人微博和博客分别为 @breezedeus 和 <http://breezedeus.github.io>。

**摘要:** 爱因互动专注于为企业提供垂直领域的商用对话机器人, 用友好、自然的人机沟通方式提升用户体验, 促进销售转化。本次演讲将主要介绍爱因互动对保险行业的 AI 化思考, 以及如何把对话机器人技术引入到保险行业, 提升保险行业的售前效率。爱因互动的保险行业解决方案都包含了哪些黑科技? 请拭目以待 -。

## 狗熊会人才培养与数据科学实践教学

潘蕊 (狗熊会)

时间: 14:00-14:30

**简介:** 中央财经大学统计与数学学院副教授, 北京大学光华管理学院博士。研究兴趣: 高维数据、网络结构数据与车联网数据。在 JASA、Annals 等期刊上均有发表。狗熊会公众号丑图百讲和精品案例系列作者。

**摘要:** 狗熊会是数据产业的高端智库, 以“聚数据英才, 助产业振兴”为使命。本次报告分为两个部分。第一部分将分享狗熊会的人才培养理念, 并且介绍数据科学的在线学习平台【熊学堂】。第二部分将分享狗熊会的【教学平台】和【云实训平台】, 这两个平台将助力数据科学在教学和实战方面的学科建设。

## 面向电力系统的数据分析与应用

庄池杰 (清华大学)

时间: 14:30-15:00

**简介:** 清华大学电机工程与应用电子技术系副教授, 主要研究方向为光电传感、电气工程领域科学计算及数据分析、间隙放电。主持科技部国家重大研发计划课题、子课题各 1 项, 国家自然科学基金项目 (面上、青年) 2 项; 承担国家电网、南方电网等企业委托项目数十项。

**摘要:** 随着电力系统设备信息化与智能化程度的不断提高和配用电数据量的迅速增长, 研究适用于输变电行业数据分析算法并建立有效的知识发现模型, 对提升电力行业运行、管理水平以及创新业务模式具有重要意义。从系统控制、设备运维、营销管理三个维度介绍了数据分析在电力行业的应用以及涉及的数学模型。

## 公司数据能力的人才构建以及公司垂直数据人才社群运营

杨慧 (*TalkingData University*)

时间: 15:00-15:30

**简介:** 现任 TalkingData CEO 助理, TDU 执行校长。中国人民大学商学院企业管理系 2010 级管理学博士, 香港中文大学管理学系博士后, 研究方向为战略管理、公司治理。曾先后供职于德电咨询、埃森哲 (中国) 有限公司, 方向为 TMT 行业战略咨询。杨慧博士多年关注互联网领域, 曾参与国家科技部商业银行信息科技风险监管支撑计划。她在研究期间, 走访了包括腾讯、网易、易车网等多家互联网企业采访中高层人员。目前她在此领域的出版物有《互联网时代的新创客》, 《互联网时代 · 新战略全景 -New Strategic Landscape under Internet》以及《体验互联网新思维》。

**摘要:** TalkingData 简介; 大数据行业现状与趋势; 大数据人才教育现状与问题; TDU 模式与人才观。

## 用数据的智慧改变建筑行业

张维 (北京广联达平方科技有限公司)

时间: 16:00-16:30

**简介:** 北京广联达平方科技有限公司数据科学家大数据研究院院长

**摘要:** 建筑行业是一个资金密集型、人力密集型、技术密集型、数据密集型的超重型行业。今天分享北京广联达平方团队如何凭借大数据创新的力量在这样一个巨擘林立的传统行业里帮助行业客户轻盈起舞, 利用数据的智慧持续驱动行业客户的研究、生产、营销、经营效率不断升级。

## 大数据在足球运动中的应用及潜力

金雄男 (北京同道伟业科技公司)

时间: 16:30-17:00

**简介:** 北京同道伟业体育科技有限公司总经理。北京科技大学管理信息系统专业毕业, 曾就职于北京现代发展规划本部, 北京京裕华通科技有限公司。

**摘要:** 数据是枯燥的, 只有让数据变成生产力, 才能体现数据的魅力。做足球数据是寂寞的, 对中国足球没有足够的情怀, 是很难坚持下来的。当中国足球长期在低水平徘徊的时候, 惟愿同道伟业以及他们的同行们, 砥砺奋进, 用数据让中国足球更加科学地训练, 更加科学地踢球, 早日让中国足球心中有数!

## 知己知彼, 百战不殆——根据公开的广告数据优化 SEM 推广策略

常莹 (狗熊会)

时间: 17:00-17:30

**简介:** 北京大学光华管理学院商务统计学硕士毕业, 曾在雅虎、美丽联合集团等公司任十余年互联网数据分析师。现为狗熊会数据分析师, 西安欧亚学院兼职教师。

**摘要:** SEM(Search Engine Marketing) 广告是互联网效果类广告中的重要分支, 它符合效果类广告的基本业务逻辑, 相对入门容易、成本可控, 尤其对于中小企业主和互联网广告的入门商家是非常好的推广手段。进行 SEM 推广的广告主共性的诉求之一: 希望可以了解同行的推广策略和市场概况, 有针对性地调优自己的推广策略。但是搜索引擎发布的数据非常有限, 同行之间共享关键业务数据显然也难以达成; 这个问题成为了 SEM 这个具有丰富实践积累的领域里相对空白的一个领域。本次报告将演示一种使用搜索引擎上公开发布的广告数据来了解市场竞争形势、定位竞争对手、优化投放策略的方法, 为满足 SEM 广告知己知彼的诉求提供一种解决方案。这一解决问题的思路同样也适用于其他一些市场和竞品监控问题。

## 基于弹幕的在线社交对视频扩散的影响

姚凯 (中央财经大学)

时间: 14:00-14:30

**简介:** 姚凯, 本科毕业于北京师范大学计算机系, 并被保送至北京大学计算机系完成硕士学位, 主要的研究内容分别是图像处理和视频编码。博士毕业于北京大学光华学院企业管理专业市场营销方向, 美国宾夕法尼亚大学沃顿商学院联合培养博士, 现任中央财经大学商学院市场营销系助理教授。研究领域包括: 互联网营销、大数据营销、实地实验和金融大数据。通过对大数据进行分析, 提高大数据的流动性, 避免数据孤岛, 通过将不同电商的数据整合起来, 达到更好的预测效果。主要给本科生和MBA学生教授市场营销, 数据分析, 大数据编程, 大数据技术及应用等课程。

**摘要:** 弹幕是观众在观影过程中发表的评论信息, 并实时嵌入视频播放过程, 其他观众在观看过程中也能看见, 近年来在商业实践中得到了广泛的应用。该研究从消费者行为的角度出发, 探究消费者使用弹幕的动机, 以及弹幕如何影响视频的扩散。为了避免研究中的内生性问题, 该研究同时获取了两个视频网站中相同视频的多期信息, 以此分析消费者在观看视频过程中发送的弹幕如何影响视频的扩散。同时, 本研究利用文本挖掘方法对消费者海量弹幕信息进行建模分析, 深入探索消费者使用这类实时评论的动机及效果。

## 基于文本挖掘的个性化推荐

张晗 (北京大学光华管理学院)

时间: 14:30-15:00

**简介:** 张晗, 北京大学光华管理学院营销系博士生, 主要关注于使用各种各样的模型解决营销问题, 发现消费者行为的规律, 帮助企业经营。

**摘要:** 个性化推荐一直是营销关注的热点, 其中最常用的方法是协同过滤方法, 以结构化的评分为输入, 输出预测的用户评分。大部分现有研究集中在对于协同过滤方法中不用评论的权重以及不同用户的权重上。我们使用话题模型, 以用户非结构化的文本评价作为输入, 来预测消费者评分。目前, 鲜有研究使用中文评论作为输入, 来预测消费者评分。除此以外, 我们的方法能够非常容易作为现有个性化推荐方法的一个结构化输入, 来改善现有的个性化推荐系统, 并且非常适合用在资讯推荐, 以及微博内容, 微博用户的推荐上。研究结果表明, 基准模型的误差 (MAE) 是我们模型的 1.1 倍; 不同话题数量的选择, 对于预测精度影响不大; 模型在对重度用户预测的表现上要略比轻度用户好。

## The Value of Seller Community: Evidence from Live Webcasting Platform

姜舒文 (北京大学光华管理学院)

时间: 15:00-15:30

**简介:** 姜舒文, 本科毕业于北京大学元培学院市场营销方向。现就读于北京大学光华管理学院市场营销系, 研究方向为营销模型方向。研究领域包括: 社交媒体、互联网营销和社会影响

**摘要:** Online social commerce, which features individual sellers of products and service in online communities, is growing fast. An often-observed characteristic of such e-commerce is the social connections between the sellers (Stephen and Toubia, 2010). For example, Etsy sets up Etsy Teams where individual sellers can interact and connect with other members. This paper examines the economic value of joining a seller community, and focuses on two types of benefits: one is the accessibility of the broadened customer base through seller network and the other is the value from community learning or knowledge spillover. Our empirical context is a live webcasting platform. Live webcasting has become a multi-billion business with millions of users broadcasting performance, game, or simply the minutiae of their daily life to other online users. We obtain data from a leading live webcasting platform in China where the broadcasters act as microenterprises and attract viewers by providing entertaining performance in expectation of virtual gifts (revenue) from viewers. Broadcasters can choose to join webcaster communities on the platform. The context is particularly suitable for our research question as the social aspect is prominent in such business and the community involves broadcasters with differentiated levels of experience. These aspects help the identification of the two types of community value. Our empirical analysis reveals that the broadcaster community offers significant value to its members. On average, community increases webcaster revenue by 40%. Evidence suggests that the revenue increase results from both broadened viewership through community referral and better broadcasting quality from community learning. We also explore the heterogeneous effect for different types of webcasters and the role of community composition.

## 数字广告作假的实证分析

梁屹天 (清华大学经管学院)

时间: 16:00-16:30

**简介:** 梁屹天在 2017 年毕业于加拿大英属哥伦比亚大学尚德商学院的市场营销系。他的主要研究方向之一是基于实证产业经济学的营销定价, 涉及几个不同产业, 包括网络游戏、电影与政府补贴项目。此外, 他也在探索如何把营销数量模型和营销心理学进行结合。他这方面的一项研究是探讨睡眠不足对消费者行为产生的影响。

**摘要:** 数字广告在 2017 年的全球规模达到两千亿美元, 并仍在稳健增长。但是伴随而来的广告作假问题却为该行业的发展蒙上一曾阴影。虽然作假问题在业界引起了广泛的关注, 学术中针对它的研究多是从理论的角度出发, 而实证的研究大部分来自于计算机领域, 并以开发更好的检测方法为主导。不同于之前的文献, 此论文将从实证的角度去分析作假的经济机制。我们集中研究两种经济因素: 市场地位 (即企业的大小) 与市场状态 (即流量的波动), 对企业作假的倾向性与策略的影响。在此基础上, 我们进一步分析这两种经济因素是否会对渠道中的上游 (广告中介) 和下游 (媒体) 产生不同的影响。我们发现上下游的作假行为很不一样。总体来说, 上游的作假行为比下游更复杂。我们推测这种差异是由上下游面对的不同合同规则所带来的。

## The Value of Time: A Study of Pricing Strategy on a Ride-Sharing Platform

郭麦菊 (北京大学光华管理学院)

时间: 16:30-17:00

**简介:** 北京大学光华管理学院市场营销系博士二年级量化营销模型方向

**摘要:** 本研究主要是在滴滴排队场景下, 研究乘客在快车和优享之间是如何做选择的。研究主要采用二维纵向产品差异化这一理论框架, 研究乘客在排队场景下价格和等待时间(用排队人数来测量)对乘客选择的影响, 目标是得到乘客金钱价值和时间价值的关系然后得出快车和优享的相对价格调整方向。数据分析结果表明价格弹性系数和等待时间弹性系数都是负向显著且这两个系数在数值上非常接近。这表明排队模式和动态调价对需求的调节力度相同, 也就是对乘客来说, 排队人数多一个人和涨价一块钱对减少需求的作用是相同的。另外, 我们还发现对排队时间敏感的人对价格更不敏感, 对价格敏感的人对排队时间更不敏感, 所以我们认为乘客在进行打车选择时需要在金钱价格和时间价值之间进行权衡。

## Star popularity or star acting skill? The impact of star power on movie box office

申桐遥 (北京大学光华管理学院)

时间: 17:00-17:30

**简介:** 申桐遥, 北京大学光华管理学院企业管理系市场营销专业博士在读。主要方向: 营销模型, 娱乐产业研究

**摘要:** Movie star is undoubtedly a critical influencing factor on movie box office, though the exact impact is unclear. Almost all of the existing papers measure star power with acting skill-related measurement. However, an interesting phenomenon in China during recent few years has inspired us that star acting skill is not the only dimension of star power. The phenomenon is that many popular stars with terrible acting skill are invited in movies. To our surprise, those movie turns out to be very successful in terms of box office. This contradiction inspired us that star's popularity could also drive movie success.

## 非零售信用风险内部评级法——客户评级建模

刘正瑶 (中国工商银行)

时间: 14:00-14:30

**简介:** 注册金融风险管理师 (FRM), 北京大学光华管理学院 2014 届金融硕士研究生, 中国人民大学财政金融学院信用管理学士学位。现就职于中国工商银行总行风险管理部, 主要从事非零售信用风险计量工作, 负责非零售客户评级模型的开发、监测和相关系统开发等工作。

**摘要:** 内部评级法是巴塞尔协议第一支柱中用于计量信用风险资本的高级方法, 相比于标准法, 其对风险的计量具有更强的敏感性, 广泛运用于国际先进银行的风险管理实践中。初级内部评级法要求银行自行估计客户的违约概率, 本演讲将重点介绍客户评级的具体建模方法, 包括数据清洗、单变量分析、多变量分析、定性指标分析、模型校准等内容, 并对内部评级建模方法与量化结果应用的相关问题进行深入思考与探讨。

## 保险精算领域的数据应用

张佳 (安永 (中国) 企业咨询有限公司)

时间: 14:30-15:00

**简介:** 北京大学数学科学学院金融数学学士、硕士, 北美精算师协会会员 (FSA) 安永咨询精算与保险风险管理咨询总监, 有超过 12 年的保险业咨询经验, 专注于保险公司偿付能力体系与资本管理、资产负债管理、风险管理、准备金评估与财务预测等方面的研究。张女士还是中国保监会保险资产负债管理监管规则制定项目组主要成员, 自 2015 年起协助保监会开展行业资产负债管理和资产配置调研、保险公司资产负债管理能力评估标准和量化评估标准制定等工作。

**摘要:** 详细讲解保险业务端“大数据”、保险资产端“大数据”以及大数据方法在保险业的应用和案例。

## 区块链技术与商用特性介绍

王璟 (布比 (北京) 网络技术有限公司)

时间: 15:00-15:30

**简介:** 区块链领域资深技术专家, 加入布比公司之前, 他曾就职于华为等多家知名企业。数年来深入钻研区块链技术和产品, 他带领的布比技术团队拥有数十项核心专利技术, 已经开发了国内领先的商业级区块链基础设施平台, 在过去的三年多时间里与八百多家机构的高层或业务部门有过区块链技术科普和应用落地的交流, 拥有丰富的区块链商业落地经验。

**摘要:** 回顾区块链的商用发展史, 介绍区块链技术的本质——分布式共享记账技术; 以及什么样的场景一定要用区块链, 区块链的商用路径以及未来最可能爆发的应用场景。

## 期权波动率的多彩空间

周宇光 (上海汨原投资管理中心)

时间: 16:00-16:30

**简介:** 汨原投资合伙人, 本科毕业于北京大学光华管理学院, 研究生就读于纽约哥伦比亚大学统计系, 曾获得全国数学奥林匹克竞赛二等奖。出国前就职于普华永道会计师事务所和泰达荷银基金管理有限公司, 自2010 年起先后在美国纽约华尔街知名自营交易公司 LaBranche 和大型对冲基金 Ramius 任衍生品交易员、投资经理, 从事衍生品套利交易, 业务覆盖欧美及亚太主要市场。归国后创立汨原投资。

**摘要:** 介绍期权隐含波动率的概念、3D 波动率曲面、以及波动率曲面在实际交易中的应用。

## 量化交易中的 R 高性能计算

任坤 (上海明法投资)

时间: 16:30-17:00

**简介:** 上海明法投资资深投资经理, 从事股票量化模型、股票和期货中高频交易模型的研究和开发。主要基于多因子模型, 利用 R 和 C++ 对低频和高频金融数据进行处理、分析和建模, 以及组合优化、实盘系统的设计和开发。工作之余积极参与开源社区 (主要是 GitHub) 中许多 R 扩展包的问题和讨论, 并贡献了formattable、rlist 等扩展包。2016 年 11 月出版了 Learning R Programming 一书, 2017 年 10 月中文版和日文版上市。

**摘要:** 量化交易尤其是股票量化模型中涉及到许多模型计算, 在实际中由于数据的可用性和交易的时效性, 这些计算需要达到一定的性能要求。本演讲从具体的场景和例子出发, 介绍量化模型从回测到实盘的过程中主要面对的数据结构和计算场景, 并展示如何充分利用向量化、高性能扩展包、Rcpp、多线程等方式逐步优化代码, 上千倍地提升代码性能, 以及每一步性能提升的技术原理, 主要是 R 对象的工作方式和复制行为。了解这些原理和技术有助于写出简洁、高性能的 R 代码, 满足在同时追求模型和性能的计算场景下能够使用合适的技术来解决问题。

## 信贷评分卡模型的开发与应用简介

谢士晨 (小米金服)

时间: 17:00-17:30

**简介:** 谢士晨, 名古屋大学博士, R 包 scorecard(python 版本 scorecardpy) 作者。目前就职于小米金服, 从事风险数据分析相关工作。

**摘要:** 本次报告将介绍评分卡模型相关概念、开发过程与其应用, 并且结合 scorecard 包演示相关功能。该包提高了评分卡的开发效率与模型稳健水平, 是评分卡开发的开源解决方案。scorecard 提供的主要功能包括(1) 信息值计算; (2) 变量初步筛选; (3) 变量分箱与 woe 转换; (4) 评分卡刻度转换; (5) 模型效果评估等。

## 制造业转型中的数字化精益

黄毅 (精益汇智)

时间: 14:00-14:30

**简介:** 黄毅博士, 精益汇智联合创始人兼 CEO, 已领导团队为 30+ 大中型工业企业, 实现其 MES、智能运营分析、数字化诊断等精益运营管理数字化系统的构建, 擅长将物联网、云计算、高阶分析和模块化软硬件等数字化和智能化技术, 融入企业生产运营场景, 帮助企业打通信息孤岛、及时发现浪费、固化精益知识并迈向智能转型; 黄毅博士, 清华大学自动化学士学位, 清华大学工业工程博士学位, 麻省理工学院 MIT 人工智能实验室访问学者。

**摘要:** 在智能制造数字化转型中, 盲目地追求无人化、云端化、人工智能化对制造企业往往不是动力而是负担, 而精益运营管理是企业从数字化转型中能真正获得价值的关键方法, 制造企业在构建数字化和智能化工业场景时, 一手是传统精益方法, 一手是先进 IT/OT 技术, 如何有效融合将关系企业的数字化成败。

## 风电大数据分析: 机遇与挑战

周杰, 崔鹏飞 (金风科技, 昆仑数据)

时间: 14:30-15:00

**简介:** 周杰, 金风科技研发中心整机系统安全主任工程师, 十余年风力发电机组系统开发及整机系统安全设计经验, 在风力发电机组仿真设计、性能优化、诊断预警、现场解决方案等领域拥有数多项国内发明专利及 7 项海外专利, 近年专注于风电大数据在机组设计改善、可靠性提升、智能运维等方面的应用。崔鹏飞, 昆仑数据分析高级工程师, 毕业于哈尔滨工业大学控制科学专业, 目前主要从事于新能源领域的故障诊断, 图像处理等方面的数据分析工作, 拥有图像处理方面多项国内发明专利。

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

## 振动分析在高速旋转设备故障诊断中的应用

李三华 (昆仑数据)

时间: 15:00-15:30

**简介:** 数据科学家, 北京大学硕士研究生, 2017 年 PHM 冠军。现任职于北京工业大数据创新中心, 负责工业大数据建模、振动分析算法的研发等工作。

**摘要:** 介绍研发的 R 语言振动分析包和常规分析算法、图谱, 并介绍实际的应用场景和诊断案例。

## 基于大数据的硅片形状诊断和预报

宁永铎 (有研半导体材料有限公司)

时间: 16:00-16:30

**简介:** 四川大学微电子学学士学位, 北京有色金属研究总院材料科学与工程硕士学位。现就职于有研半导体材料有限公司, 担任技术研发部经理, 主要负责半导体硅材料工艺技术研发与产业化、品质分析与持续改进、技术类问题交流与沟通。

**摘要:** 以半导体硅片几何参数检测原始数据为数据源, 基于数据挖掘理论和机器学习技术实现硅片几何参数异常的自动化诊断。首先以半导体硅片生产制造环境中的几何参数异常产品为研究对象, 收集了 19 种异常、共 1033 片几何参数异常片作为实验品, 采用数据挖掘方法, 以差分分析、回归分析、方差分析为基本算法, 提取出异常硅片的特征统计量, 这些特征统计量共构造出 9 个特征值, 不同异常原因的硅片在这 9 个特征值上有显著的分离度; 随即分别提取了每枚硅片的特征值, 应用机器学习技术, 用提取出的特征值训练机器学习模型 (分类器), 再检验分类器的出错率。在机器学习阶段分别尝试了 linear discriminant analysis、adaboost、bagging、random forest、support vector machine 和 k-nearest neighbor 等 6 种分类器。实验结果表明, support vector machine 分类器的误判率最高 (3.58%), 而 random forest 分类器的误判率最低 (0.3%), 这些分类器的低误判率均可以满足硅片制造系统的要求, 证明结合数据挖掘方法和机器学习技术可以实现异常品的自动化诊断。

## 利用迁移学习实现储备粮关键品质指标的全程估计

王迪 (北京大学)

时间: 16:30-17:00

**简介:** 北京大学工学院工业工程与管理系 2015 级博士研究生。目前主要从事复杂工业系统过程监控、诊断与优化方面的研究。目前以第一作者身份完成 6 篇学术论文。已公开发表 2 篇 EI 论文, 其中一篇获得 2017 年 IEEE 系统集成与信息学 (IEEE & SICE System Integration) 国际会议最佳论文奖; 完成 4 篇 SCI 论文, 均投稿至本领域顶级期刊。申请国家发明专利 2 项; 获得国家计算机软件著作权 1 项。曾获北京大学五四奖学金、专项奖学金、三好学生、优秀科研奖等荣誉。

**摘要:** 粮食在仓储过程中会因为多种原因而产生变质, 导致粮食的损失。仓储中粮食的损失已成为全世界广泛关注的问题。因此, 对仓储中粮食品质的监控至关重要。本研究以国家储备粮库中的储备粮作为研究对象, 以粮食温度作为评价粮食品质的指标, 提出了一种利用迁移学习实现储备粮关键品质指标的全程估计方法。针对仓储中粮温传感器网络在数据采集过程中存在的数据稀疏性和数据缺失的问题, 本研究借助同一粮库下与目标粮仓具有相同仓储环境的其他粮仓的粮温传感数据, 采用一种将多任务学习模型和自相关模型相结合的时空动态温度场的迁移学习方法, 实现对目标粮仓温度场的准确估计。通过本研究能够解决目前在仓储中存在的由于传感器数据不足、数据缺失而无法获得粮食温度场全程信息的问题, 可以得到粮食温度场的全面而准确的信息, 为仓储中粮食品质的监控、决策等措施提供依据。

## 组织情境下员工工作行为大数据研究

徐敏亚 (北京大学光华管理学院)

时间: 14:00-14:30

**简介:** 徐敏亚现任北京大学光华管理学院商务统计与经济计量系副教授。她 2004 年毕业于中国科技大学统计系获学士学位, 后于 2008 年在美国罗格斯大学统计学获得博士学位。她专长于统计方法在管理中的应用。

**摘要:** 本演讲根据中国某 IT 公司的工作平台数据, 对员工的工作行为与绩效之间的关系进行了广泛的探索, 特别地探讨了好的销售人员的工作行为特征。同时利用公司工作软件记录的日志数据来识别销售人员与其它员工的工作社交网络, 研究他们的工作社交网络与销售业绩的关系。

## Analysis of Consumer Reviews Using Sequential Phrase Selection

王菲菲 (中国人民大学)

时间: 14:30-15:00

**简介:** 北京大学光华管理学院 2017 届统计学博士, 中国人民大学统计学院讲师。感兴趣的研究方向为文本挖掘、空间统计学以及贝叶斯分析等。

**摘要:** Text mining has attracted more and more attention with the accumulation of text documents in all fields. In this talk, we focus on the analysis of consumer reviews. The goal is to exploit the dependent relationship between textual information extracted from consumer reviews and the corresponding rating score. To handle the unstructured texts, one common practice is to structuralize the text documents via vector space models, which often lead to an extremely large term set and suffer from the curse of dimensionality. Under this context, we propose a novel sequential phrase selection method for vector space models under a linear regression setup. Results show that this selection technique can effectively detect the relevant phrases.

## Bayesian Estimation of the General Probability of Informed Trading Model

郇钰 (北京大学光华管理学院)

时间: 15:00-15:30

**简介:** 北京大学光华管理学院商务统计与经济计量系应届博士研究生, 本科毕业于山东大学数学学院统计学专业。博士研究方向为金融市场微观结构。

**摘要:** The probability of informed trading (PIN) is a widely used direct measure of market information asymmetry risk. Maximum Likelihood Estimation (MLE) of the PIN model often encounters numerical problems. In this paper, a Bayesian method is proposed to estimate the PIN model, which uses Gibbs sampling and the adaptive rejection sampling algorithm. Simulation studies reveal that the Bayesian method overcomes numerical problems, and also leads to more accurate estimates than the MLE methods. We apply our method to obtain annual PIN estimates for all stocks in Shanghai and Shenzhen Stock Exchanges in China over seven

years between 2009 and 2015. We find that in the Chinese stock market, PIN can explain observed differences in spreads, but does not affect future returns.

## 随机波动率模型下的闭形式隐含波动率曲面

李晨煦 (北京大学光华管理学院)

时间: 16:00-16:30

**简介:** 北京大学光华管理学院在读博士研究生, 南开大学数学科学学院理学学士学位。研究方向包括: 金融工程、金融计量和统计建模等。

**摘要:** 期权类产品是衍生品市场中交易最为活跃的衍生品之一。在对大多数的期权类产品进行报价时, 通常报告的并非价格, 而是该期权类产品通过 Black-Scholes 公式反解出的隐含波动率。隐含波动率和期权价格之间一一对应, 且为期权的价格及未来一段时间内的市场风险提供了一种更直观的度量, 因此在实际交易中有着广泛的应用。本文试图回答什么样的随机波动率模型能最好地拟合市场中观测到的隐含波动率的数据。对于任意的随机波动率模型, 我们得到了隐含波动率的任意几何形状特征 (如曲面水平、斜率、凸度) 的闭形式二元展开。这一闭形式展开使我们得以显式地分析随机波动率模型中的不同参数对隐含波动率曲面形状的影响。反过来, 利用这一闭形式展开公式, 我们可以构造一个“隐含随机波动率模型”, 以期直接拟合市场中发现的隐含波动率曲面的几何形状特征。

## Validating Power Laws in Economics and Finance

宋晓军 (北京大学光华管理学院)

时间: 16:30-17:00

**简介:** 北京大学光华管理学院商务统计与经济计量系助理教授, 西班牙马德里卡洛斯三世大学经济学博士。主要研究兴趣是理论计量经济学, 包括非参数, 半参数方法, 假设检验和自助法, 以及计量经济学的应用。

**摘要:** In the past decades, many economical and financial variables are identified as obeying the power law distribution, e.g. city size, firm size, macroeconomic disasters. In this talk, we propose a novel testing procedure to validate if the power law with an unknown exponent holds for the underlying variable. The test is asymptotically distribution free, and easy to apply. In our empirical study, we find strong evidence to reject several variables, which were believed to obey the power law. We also try to explore the potential reasons behind the failure of power law.

## Bayesian Estimation of Demographic Systems

张俊妮 (北京大学光华管理学院)

时间: 17:00-17:30

**简介:** 北京大学光华管理学院统计学副教授, 哈佛大学统计学博士。她的研究领域为贝叶斯分析、人口统计学、文本挖掘、因果推断。著有中文教材《数据挖掘与应用》, 即将出版英文专著《Bayesian Demographic Estimation and Forecasting》(与 John Bryant 合作)。

**摘要:** Many problems in applied demography consist of estimating demographic systems. We present a general approach to estimating demographic systems based on the idea of a demographic account. A demographic account is a set of linked tabulations of demographic series, such as fertility, mortality, migration, and population. The estimation methods are fully Bayesian, and tackle challenges such as disaggregation, measurement error, missing data, and combining multiple data sources. We simultaneously estimate (i) the true counts of events and populations, (ii) the demographic rates (e.g. fertility rate, mortality rate), and (iii) indicators of data quality. The methods also generate uncertainty measures for all estimated quantities. We illustrate the methods using examples.

## 患者住院数据的探索性分析

鄂尔江 (清华大学)

时间: 09:00-09:30

**简介:** 鄂尔江, 男, 清华大学工业工程系博士生, 研究方向为医疗知识管理和医疗数据分析。

**摘要:** 利用某医院 3 万多条患者住院记录, 对患者基本信息、诊断信息、手术信息、费用信息进行分析。一方面从患者的角度研究患者疾病风险, 另一方面从医院的角度研究医院运营管理策略。

## Semantics-driven Feature Extraction for High-throughput Phenotyping

宁温馨 (清华大学)

时间: 09:30-10:00

**简介:** 宁温馨, 25 岁, 清华大学工业工程系博士研究生, 喜欢通过数据科学方法解决实际问题。本科毕业于清华大学工业工程系并取得工学学士学位及计算机科学辅修学位, 博士阶段研究集中在利用计算语言学和机器学习等方法进行电子病历数据分析, 期间于哈佛大学生物统计系进行半年的访问学习, 在国际知名学术期刊及会议上发表多篇论文。业余对量化投资及交易同样具有浓厚的兴趣。

**摘要:** Phenotyping algorithms can efficiently and accurately identify patients with a specific disease phenotype and construct electronic health records (EHR)-based cohorts for subsequent clinical or genomic studies. Previous studies have introduced unsupervised EHR-based feature selection methods that yielded algorithms with high accuracy. However, those selection methods still require expert intervention to tweak the parameter settings according to the EHR data distribution for each phenotype. To further accelerate the development of phenotyping algorithms, we propose a fully automated and robust unsupervised feature selection method that leverages only existing medical knowledge sources, instead of EHR data. Methods: SEmantics-Driven Feature Extraction (SEDFE) collects medical concepts from knowledge sources as candidate features and gives them vector-form distributional semantic representations derived with neural word embedding and the Unified Medical Language System Metathesaurus. A number of semantically closest features that sufficiently characterize the target phenotype are determined by a linear decomposition criterion and are selected for the final classification algorithm. Results: SEDFE was compared with the EHR-based SAFE algorithm and domain experts on feature selection for classifying four phenotypes including coronary artery disease, rheumatoid arthritis, Crohn's disease, and ulcerative colitis, using both supervised and unsupervised approaches. Algorithms yielded by SEDFE achieved comparable accuracy to those yielded by SAFE and expert-curated features. SEDFE is also robust to the input semantic vectors. Conclusion: SEDFE attains unsupervised feature selection for EHR phenotyping with satisfying performance. Being fully automated and EHR-independent, this method promises both efficiency and accuracy in developing algorithms for high-throughput phenotyping.

## 医疗资源消耗预测与预测任务导向的医疗编码表示学习

崔力文 (清华大学)

时间: 10:30-11:00

**简介:** 2009 年 8 月进入清华大学物理系就读, 2013 年 7 月本科毕业并获得理学学士学位。2013 年 8 月免试进入清华大学工业工程系攻读管理科学与工程博士至今, 期间曾于 2015 年前往美国普渡大学克兰纳特管理学院, 进行为期 6 个月的合作研究。

**摘要:** 医疗资源稀缺是当今许多国家面临的大挑战。根据患者的具体情况准确预测其医疗资源消耗, 有利于医院提高医疗管理水平, 并辅助政府制定更合理的报销政策, 从而实现对医疗资源的有效利用。随着医疗信息技术的广泛应用, 研究人员可以获得更丰富的医疗数据, 从而得以利用机器学习技术从中挖掘有价值的信息, 在数据的支持下, 提高医疗资源利用效率及医疗服务水平。本研究使用的电子病历 (Electronic Health Record, EHR) 数据集, 采集自北京市五家三甲医院约 75 万份住院病案首页, 包含患者的性别、年龄、诊断、手术操作、医疗资源消耗情况等信息。其中, 诊断信息和手术操作信息使用医疗编码进行表示。以往关注医疗资源消耗预测的文献, 通常致力于对单一指标 (如医疗费用) 的预测。然而, 在实际应用中, 医疗资源消耗情况通常通过多个指标来共同衡量。本研究通过建立多输出机器学习模型, 来同时预测多个指标。这不仅能够得到更加符合实际需求的简单统一的预测规则, 同时还能节约计算资源, 并能够利用多个任务间的相关性来提高模型预测的准确度。将 EHR 数据集包含的医疗编码转换为特征向量, 是搭建机器学习模型的基础步骤。研究发现, 医疗编码的向量表示方式会对预测模型的表现产生很大影响。在以往的研究中, 通常不考虑具体的预测任务, 基于医学知识或无监督学习模型生成医疗编码向量。本研究使用预测任务来指导医疗编码向量的生成, 有效提升了医疗编码向量的预测能力。本研究关注两种医疗编码向量生成方式。第一种是利用自然语言处理 (Natural Language Processing, NLP) 领域的表示学习模型, 生成医疗编码的低维连续向量表示。这种医疗编码表示方法预测能力很强, 但可解释性较差, 在实际应用中推广的难度较大。第二种是对医疗编码进行分组, 然后构建基于组的独热向量 (one-hot vector) 来表示医疗编码。这种方法由于可解释性很强, 在相关文献及应用中较为常见, 但预测能力通常较差。本研究一方面在分组过程中融入表示学习模型, 提升了医疗编码分组的预测能力; 另一方面利用编码的树型分层结构, 使得分组结果贴近临床经验, 更容易被医务工作者所接受。本研究有效提升了医疗资源消耗预测的准确度, 使得模型能够更好地服务于实际需求; 对可解释性的充分探讨, 使得成果具有了更高的实际应用价值。

## 引入文本章节结构进行远距离关系提取并应用于医学知识提取

林毓聪 (清华大学)

时间: 11:00-11:30

**简介:** 林毓聪, 清华大学统计中心博士, 师从俞声教授。研究方向为自然语言处理、医学电子病历分析、关系提取、知识图谱构建等。

**摘要:** 作为构建知识库的核心工作, 关系抽取在人工智能时代得到了广泛的关注。传统的关系抽取作品都有一个很强的假设, 即只有两个实体都出现在这个句子中, 句子才能表达实体对关系的含义。这个假设排除了大量含有其他句法结构的句子。在本文中, 我们打破了原有假设, 提出了一种结合文章结构信息的新型关系抽取模型, 该模型可以处理长距离关系抽取, 使得它在文本语料库中提取关系时更加适用。我们将该模型应用于在线医学关系提取, 并在完整数据集中与其他关系抽取模型进行了比较。实验证明我们的模型达到了最高的精度。

## 69863 例脑卒中患者合并疾病研究

文天才（中国中医科学院中医药数据中心）

时间：11:30-12:00

**简介：**文天才，中国中医科学院中医药数据中心软件工程研究室主任，高级工程师，硕士生导师。主要从事医学数据分析、临床试验数据标准、临床试验数据管理技术与方法、医院绩效评价方法的研究。主持科技部重大专项课题 1 项，国家自然科学基金面上项目 1 项，局级课题 4 项，获得国家发明专利 4 项，发表论文 20 余篇。

**摘要：**本研究的目的是要评估中国脑卒中患者合并疾病类型和特点。从全国 439 家医院采集 2015 年 1 至 12 月份 69863 例首次脑卒中出院患者，包括 7904 例脑出血（11.31%），61079 例脑梗死（87.43%）和 880 例（1.26%）例脑血管闭塞患者。患者年龄范围从 18 岁到 104 岁，平均年龄 67 岁。依据 ICD-10 诊断 3 位类目对所有合并疾病进行归类，利用 R 3.4.4 软件 arules 和 ape 包进行合并疾病关联规则和层次聚类分析，利用 Gephi 0.9.2 进行合并疾病网络分析。最终获得疾病共计 966 类，脑卒中合并 I10 特发性（原发性）高血压、I25 慢性缺血性心脏病、E11 非胰岛素依赖型糖尿病最多且关联程度最高，心脑血管疾病、代谢紊乱及脑卒中引起的并发症或后遗症关系密切。上述核心疾病在所有合并疾病中具备高点度中心性、高接近中心性、高中介中心性和低群集系数，脑卒中合并疾病网络呈幂率分布具备无标度网络特征。

## 基于 R 的流行病动力学模型测试

杜向军 (中山大学)

时间: 08:30-09:00

**简介:** 杜向军, 中山大学公共卫生学院(深圳)教授, 博士生导师。2017 年入选中山大学百人计划, 2018 起加入中山大学公共卫生学院(深圳), 入选第十四批中组部“千人计划”青年项目。2005 年华中科技大学物理系本科毕业, 2010 年中科院生物物理所生物信息学博士毕业, 之后在美国国立卫生研究院生物信息学中心(NCBI)、密歇根大学以及芝加哥大学生态进化系从事计算系统生物学研究, 在运用学科交叉方法解决与公共卫生相关的传染病研究方面有着丰富的经验, 代表性成果包括基于序列信息的季节性流感快速准确的抗原监测、疫苗株推荐、动力学模拟以及流行强度提前预测等。实验室侧重通过计算系统生物学的方法, 综合运用数据分析以及理论建模等手段, 定量的研究多种相关因素与传染病的产生、演化、传播以及致病的关系, 揭示其背后的生物学机制, 指导传染病日常监测、防控与治疗。

**摘要:** R 在传统的流行病与卫生统计研究中已经起到举足轻重的作用。随着流行病动力学模型理论研究的不断推进, R 在基于模型测试来揭示传染病传播与流行中复杂的规律方面也起到越来越重要的作用。以流感为例, 通过提出不同假设, 比较不同模型, 我们将探索如何利用 R 来揭示传染病传播与流行规律背后的精细生物学机制, 用于指导更有针对性的传染病监测与防控。

## Epidemiology of human infections with influenza A(H7N9) virus and pandemic risk assessment

王锡玲 (复旦大学)

时间: 09:00-09:30

**简介:** 王锡玲, 复旦大学公共卫生学院副教授。2014 年博士毕业于香港大学公共卫生学院。主要研究方向为传染病流行病学与统计建模, 具体包括流感的季节性及其驱动因素研究、疾病负担估计; 禽流感 H7N9 的流行病学参数估计、大流行风险评估及干预措施有效性评价; 流感感染与发作性睡病的关联研究。以第一或通讯作者(含共同)发表 SCI 论文 14 篇。2017 年关于禽流感 H7N9 的论文以快速通道的形式发表在柳叶刀传染病学杂志 Lancet Infectious Diseases(IF=19.9, 传染病学领域排第一), 并同期配发了专家述评。主持国家自然科学基金青年项目一项。担任预防医学会生物统计学分会第一届青年委员会秘书长; 担任 Emerging Infectious Diseases 等国际期刊的审稿人。

**摘要:** Background: A surge in laboratory-confirmed human cases of A(H7N9) virus infection in 2016–17 has prompted concerns of an increasing pandemic threat. Our study aimed to describe the epidemiological characteristics, clinical severity of A(H7N9) case-patients and assess human-to-human transmissibility of A(H7N9) virus in the 2016–17 epidemic wave, compared with previous waves. Methods: We described the epidemiological characteristics and clinical severity profile of laboratory-confirmed human cases of A(H7N9) virus infection across 5 epidemic waves in mainland China. We estimated the bounds on the effective reproductive number ( $R_e$ ) of A(H7N9) virus by analyzing clusters of case-patients. Results: The 2016–17 A(H7N9) epidemic wave began earlier, spread to more counties in affected provinces and had more confirmed cases than previous epidemic waves. Proportions of cases in semi-urban and rural residents in the 2015–16 and 2016–17 epidemic waves (63% and 61%) were higher than those in the first three epidemic waves (38%, 56% and 55%). The clinical severity of hospitalized cases in 2016–17 was comparable to the previous epidemic waves. The upper limit of  $R_e$  for A(H7N9) virus was 0.12 (95% CI 0.10, 0.14), and was not significantly different across waves. Conclusion: Case

sources changed gradually across epidemic waves, while clinical severity and transmissibility has not changed substantially. Continued vigilance and sustained intensive control efforts are needed to minimize the risk of human infection with A(H7N9) virus.

## 流感疫情双峰现象的潜在机制

徐波 (清华大学)

时间: 09:30-10:00

**简介:** 徐波, 清华大学地球系统科学系生态学专业博士在读, 研究兴趣为传染病动力学模型和病毒进化动力学。

**摘要:** 流感大流行的多峰现象在历史上出现了多次, 例如 1918 年西班牙流感在英格兰引发了三波疫情, 在美国引发了两波疫情; 2009 年甲型 H1N1 流感在墨西哥和加拿大分别引发了三波和两波疫情。但是对于引起这一现象的机制还缺乏一个较为系统的总结。本研究着眼于单一流感季/年内出现两波疫情的现象, 提出了一个判定双波疫情的标准 (2-wave metric), 总结和提出了共 16 种可能导致这一现象的机制, 使用 R 语言中的 fitR 函数集, 将它们分别抽象化为不同的流行病学模型, 并探讨不同的干预措施可能造成的影响。

## R 在环境流行病学研究中的应用

林华亮 (中山大学)

时间: 10:30-11:00

**简介:** 林华亮, 中山大学公共卫生学院副教授, 博导; 获得广东省杰出青年医学人才称号。2011 年毕业于香港中文大学公共卫生学院, 获博士学位。中华预防医学会环境卫生分会委员、媒介生物学分会委员、中国卫生信息学会卫生地理信息专业委员会会员; 华南预防医学杂志编委。主要研究方向为环境流行学, 对室内外空气污染、气候变化对人群健康的影响有多年的研究经验。承担和参与了国家自然科学基金、973 课题、国家卫计委卫生行业专项课题、广东省自然基金、广州市产学研协同创新重大课题等多项课题。获国家发明专利和实用新型专利各 1 项。目前已发表 SCI 论文 90 余篇, 其中第一作者或通讯作者文章 50 余篇; 总影响因子 350 余分, 文章被引用超过 2000 余次, H-Index 达到 27。主要文章发表在多个国际学术期刊, 如 Hypertension, Stroke, Environment International, Proceedings of the National Academy of Sciences, Am J Respir Crit Care Med, Environmental Health Perspectives, Environmental Pollution, Atmospheric Environment 等。

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

## 全球变暖的疾病负担研究

李国星 (北京大学医学部)

时间: 11:00-11:30

**简介:** 李国星, 博士, 北京大学医学部公共卫生学院讲师。主要研究领域: 环境流行病学, 特别是大气污染和气候变化的健康影响和疾病负担。作为项目负责人, 先后主持国家自然科学基金面上项目 1 项、省部级课题子课题 1 项和中华医学会课题 1 项, 并参与多项国际自然科学基金和环保部公益项目课题。作为第一或通讯作者累计发表英文论文 21 篇, 包括在 Environmental International, Stroke, Environmental Pollution, Environmental Research, Science of the Total Environment 等期刊; 参编专著教材 4 部, 包括《现代环境卫生学》、《空气颗粒物与健康》等。目前担任中华预防医学会卫生工程分会青年委员、北京环境诱变剂学会青年专业委员会青年委员。

**摘要:** 全球变暖已经引起了学术界和公众的广泛关注。但其对我国相关的疾病负担研究开展尚少。本研究利用 R 软件 dlnm 包, 首先基于我国某城市多年的气象数据, 污染物数据和健康数据, 构建温度与健康之间的暴露反应关系模型; 其次利用 IPCC 提出的多种气候模式, 预测未来不同情景模式下可能的健康收益/损失, 为全方位的理解气候变化对我国公众的健康效应提供科学证据。

## Non-inheritable risk factors during pregnancy for congenital heart defects in offspring: a matched case-control study

蔡俊 (清华大学)

时间: 11:30-12:00

**简介:** 蔡俊, 清华大学地球系统科学系生态学专业博士生, 研究兴趣包括流感传播动态、传染病流行病学和环境健康。博士课题关注流感的时空传播动态, 曾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 6 年 R 语言编程和数据分析经验, 是 geoChina 和 humidity 包作者以及 animint、incidence 和 ncf 包贡献者, R Epidemics Consortium 成员, 是 R Graphics 中文版本《R 绘图系统》的译者之一。近期对 R 在传染病数据分析和建模中的应用感兴趣。

**摘要:** 先天性心脏病是先天畸形中最常见的一类, 其产生是环境和遗传等因素共同作用的结果。同时由于多种亚型先天性心脏病病因的异质性, 导致很难评价单个因素对先天性心脏病的影响。本演讲将介绍如何用 R 构建配对病例对照研究 (matched case-control study) 评价江苏和安徽两省母亲孕期的环境暴露与后代患先天性心脏病风险之间的可能关系, 涉及病例对照匹配过程、条件 logistic 回归分析 (conditional logistic regression analysis)、成对多重比较及修正 (pairwise multiple comparisons with correction) 和剂量反应分析 (dose-response analysis)。

## 初窥爬虫门径

杜亚磊 (无忧英语)

时间: 08:30-09:00

**简介:** 51talk 算法工程师

**摘要:** 统计是一门数据的搜集, 整理, 分析, 展示的科学。互联网时代, 爬虫是搜集数据的主要方式之一。这次分享主要介绍爬虫的基础知识 (HTML, CSS, JS) 和实用技巧 (Ajax, Proxy 等), 列出相关的学习资源。但并不过多涉及反爬虫策略的破解, 也没有深入的代码。最后提出一些有趣的想法和案例, 供大家探讨。

## Shiny: 从零到一搭建可视化 BI 平台

李朋飞 (无忧英语)

时间: 09:00-09:30

**简介:** 无忧英语算法工程师。

**摘要:** Shiny 是 R 语言的一个包, 它让你不需要任何的网站和网页前端知识, 仅仅通过 R 语言, 就可以搭建一个可视化 BI 平台。只要你对 R 语言够熟悉, 就可以将所有 R 语言的图形, 数据甚至模型以各种形态展现出来, 定制完全个性化的 BI 平台。Shiny 足够简单, 从零到搭建一个简单的可视化页面只需要很短的时间就能够完成, Shiny 功能足够强大, 各种交互组件都有定制, 可以让你通过搭积木的方式搭建一个自己的 BI 平台, 本次演讲将通过一个实例来讲述 shiny 的应用和可视化 BI 的搭建。

## Tensorflow 在 R 上的部署及使用

李智凡 (中国人民大学)

时间: 09:30-10:00

**简介:** 中国人民大学统计学院数理统计方向硕士一年级在读, 有一定经验的 R 语言使用者。主要兴趣方向: 概率图模型, 统计软件。

**摘要:** Tensorflow 是由 google 开发的人工智能学习系统, 被广泛用于语音识别或图像识别等多项机器学习和深度学习领域, 也是目前世界上最受欢迎的人工智能学习系统之一。目前, 绝大多数人工智能方向的学者、工程师会选择 python 作为 tensorflow 开发的语言和工具, 而事实上, Rstudio 也完成了 R 语言对于 tensorflow API 的接口, 意味着 R 的用户也能利用 tensorflow 进行深度学习模型的开发。本报告将对于 tensorflow 的基本概念做简要介绍, 利用 tensorflow 实现常用的 CNN 与 RNN 模型, 对于另一深度学习模型 keras 在 R 上的实现, 也会作简要介绍。

## 基于 LSTM 的自动文本生成模型构建

吕剑航 (中国人民大学)

时间: 10:30-11:00

**简介:** 中国人民大学统计学院 2014 级本科生, 主修经济统计学, 辅修数理统计学第二学位、互联网金融第二专业; 已推免保送至北京大学光华管理学院商务统计与经济计量系硕博连读项目。

**摘要:** 随着互联网技术的发展, 人工智能的发展在诸多领域都取得了显著性突破。深度学习也被应用到了自然语言处理、机器翻译和图像识别等诸多领域。本次展示将从深度学习中最经典的循环神经网络 (RNN) 和长短期记忆网络 (LSTM) 出发, 对其原理、结构和理论推导进行详尽的剖析; 在此基础上, 本次展示还将利用 TensorFlow 构建一个基于 Char\_RNN 的自动文本生成模型, 并分别以日文文集、五言古诗和网络小说作为训练集进行训练, 且对试验结果进行展示; 最后, 本次展示还将对深度学习的本质和人工智能未来的发展方向进行展望。

## R 环境模型部署实践

周震宇 (中国人民大学)

时间: 11:00-11:30

**简介:** 周震宇, 中国人民大学统计学院在读研究生一枚, 方向为数理统计, 兴趣为机器学习与文本挖掘, 爱码代码。最常用 R, python 次之。坚持没有工具无高低优劣之分, 所有工具皆服务于场景。

**摘要:** 一个成熟完整的统计建模分析流程中, 模型部署是不可或缺的一环, 依赖它才能运用现有模型对实时的流数据做评分与推断, 从而在生产环境中完成对原模型效果与性能的反馈, 实现建模流程的闭环。部署模型可利用的方式有很多: 移动终端、GPU、云环境……本报告将概述模型部署的基本概念与用途, 并且聚焦于如何在 R 的环境中用网络服务这种方式来部署上线一个模型, 介绍的工具主要包括: plumbeR、openCPU、fiery、httpuv……报告会从性能上对这些不同的工具框架做出对比。最后, 报告将会给出部署模型可能遇到的一些问题以及它们的解决方案。

## R Markdown 应用之学位论文排版

黄湘云 (中国矿业大学 (北京))

时间: 11:30-12:00

**简介:** 中国矿业大学 (北京) 2015 级统计学硕士, 研究兴趣包含统计计算、统计图形、混合模型 (mixed models), 曾在新浪实习, 目前就职于京东。

**摘要:** 我为什么入坑 LaTeX, 又扶着 R Markdown 这个梯子爬出来, 介绍如何将 R Markdown 用于学位论文的模板开发和 R Markdown 的其他应用, 如数据分析报告和个人博客的搭建

## 地理要素的尺度效应、可变面积单元问题与空间统计的挑战

戴劭勍 (中国科学院城市环境研究所)

时间: 08:30-09:00

**简介:** 戴劭勍, 硕士研究生, 中国科学院城市环境研究所, 研究方向空间统计、遥感与生态过程模型。

**摘要:** 地理要素存在着尺度效应, 地理要素的特征在不同的时空尺度上存在着差异, 由此使得在进行地理分析时难以避免可变面积单元问题。这两个问题至今仍是地理学的研究热点, 空间统计作为 GIS 中对于地理要素的时空分布描述的关键学科。尺度效应与可变面积单元问题给空间统计带来了如何的挑战。我们以地理探测器结果的不确定性进行讨论与分析。试图对这一方面进行深入的研究。

## 多源大数据探测城市多中心结构

蔡纪烜 (腾讯)

时间: 09:00-09:30

**简介:** 蔡纪烜, 香港中文大学博士毕业生, 现于腾讯位置大数据组从事空间数据挖掘。在地理大数据获取、管理以及数据挖掘有丰富经验, 关注利用新兴数据研究人口迁移、城市规划及人居环境。

**摘要:** 大数据时代的到来使得近几年的城市定量研究有了质的飞跃, 然而如何集成不同类型、不同口径的空间大数据, 发挥多源大数据各自的优点是当下该类研究的一大挑战。对此, 我们综合了空间统计的方法以及图像处理算法, 基于遥感影像、社交媒体以及 POI 等多源数据, 研究了北京、上海和重庆三个形态各异的多中心城市之结构。该研究发表于 Remote Sensing of Environment。

## 多源数据融合辅助人口分析与政府管理

李颖 (北京清华同衡规划设计研究院)

时间: 09:30-10:00

**简介:** 北京清华同衡规划设计研究院技术创新中心规划师

**摘要:** 人是城市运行的主体, 人口分布与活动特征对设施、产业、交通等相关规划与决策制定具有重要参考意义。传统的人口分析主要依赖各部门调查数据, 人工成本高、数据更新慢、时空粒度粗糙, 难以即时反映人口的真实动态, 导致城市管理措施的制定通常因为缺乏对“人”这一活动主体的充分认识而存在滞后性。

随着互联网和移动设备产生的新数据获取与分析技术的逐渐成熟, 精准识别人口位置和活动轨迹成为可能, 这使得人口研究中空间与时间尺度的精细化程度和量化水平都得到了提升。然而, 新数据环境中的多源数据口径不一, 目前无论在数据选取、指标选取还是模型构建上均在缺乏系统性。研究通过丰富的多源异构大数据应用经验, 积累了手机信令、互联网定位、交通出行、空间要素等具有精准时空属性的新兴数据源关联对比与指标设计经验, 并实现了新数据结合传统调查数据的交叉校验, 建立了更精准、更高效、更全面的人口研究指标体系与技术路线。研究以朝阳区动态人口监测为例, 主要实现(1)建立了基于总量、时序、空间的三维数据检验方法, 对各类数据进行口径、质量的适用性的评估, 为数据分析与融合提供可靠基础;(2)将多源数据的优势紧密结合动态人口刻画的需求, 挖掘人口数量、空间分布与通勤活动等规律, 并对人口与城市用地空间布局、产业结构调整进行关联分析, 辅助制定合理的管理措施, 为实现人口与资源环境协调发展规划提供量化

依据。(3) 引入机器学习算法进行了人口数量与空间分布的预测, 实现了区域人口发展趋势的提前预警; (4) 构建了通过敏感性验证的人口分析指标体系, 并将其整合到人口监测系统平台中, 为管理者掌握城市人口规模变动、摸清人口活动规律提供了更高效的途径, 促进城市功能与人口布局的协同优化, 提高城市发展质量, 也为政府进行相关规划决策提供了科学依据。

## 城市产业结构及其经济复杂度研究

冯娟 (量子数聚)

时间: 10:30-11:00

**简介:** 冯娟, 华中科技大学硕士, 量子数聚研究总监, 首席分析师, 长期从事政府及企业大数据分析挖掘工作, 拥有十多年的政府大数据项目经验, 曾在国家工商总局, 北京财政局, 北京地税局等多个政府部门担任过决策支持及数据挖掘项目总分析师。

**摘要:** 本次报告着重探讨了如何基于城市企业大数据进行城市产业结构及产业经济研究。报告内容主要包括以下几部分: 1 城市企业大数据介绍 2 企业大数据研究城市产业结构的方法论 3 城市产业经济模型介绍 (产业偏好及竞争力优势模型, 产业相似度模型, 产业生命周期模型, 产业经济复杂度模型等) 4 产业模型应用案例介绍

## 数据 - 证据 - 决策: 交通规划的例证与思考

魏贺 (北京市城市规划设计研究院)

时间: 11:00-11:30

**简介:** 魏贺, 北京市城市规划设计研究院, 研究方向: 交通政策、交通模型、交通规划。

**摘要:** 数据 - 证据 - 决策, 借助城市交通模型探索空间、产业、环境、设施和服务间规律与规则的关系, 基于多源数据的交叉校验分析因果逻辑, 借助逻辑认知和数据结论形成证据链条以支撑研判, 利用可视化和交互现实增强交通规划者与决策者对数据和证据的理解。由此计算、表征、挖掘和验证出的规律与规则具有“借助数据、挖掘证据; 发现规律、拟定规则; 模拟现状、验证规则; 推估未来、洞察规律”的特点。科学观应“实事求是, 因果推断”; 研判结论应形成“数据网、证据链、决策树”的递进逻辑; 输入输出应体现“透明化、平台化、开放化”的原则; 预测方法应挑战自我并审视认知, 积极应对未来不确定性。

## 综合能源数据分析平台构建及其在智慧城市中的应用

王扬 (国家电网)

时间: 11:30-12:00

**简介:** 王扬, 男, 1983 年 5 月生, 汉族, 研究生文化, 工学博士学位, 高级工程师职称。现任: 北京大学光华管理学院工商管理专业博士后, 国网天津市电力公司高级工程师。长期从事综合能源数据分析方向创新工作, 参与国家 863 计划项目, 主持天津市重点研发计划科技支撑重点项目和国家电网公司科技项目, 提出面

向智慧城市的综合能源数据分析方法，并在中新天津生态城智能电网等应用中取得良好效果。作为主要完成人完成的研究成果先后获得天津市科技进步奖、中国电力创新奖和国家电网公司科技进步奖等省部级科技奖励 7 项，为天津城市电网信息化和智能化水平的提升做出了重要贡献。

**摘要：**本成果属于信息科学学科领域，研究对象是综合能源数据。本成果针对综合能源数据来源广泛，结构复杂，且与用户、时间、空间信息关系紧密的特点，构建了高性能综合能源数据分析平台，提出了多尺度、细粒度的综合能源数据分析理论框架及方法，并将其应用于智慧城市建设。主要成果如下：（1）搭建了统一高效的综合能源数据存储与计算服务平台，实现天津地区城市电网数据共享和业务融合，业务应用从“搬数据”向“搬计算”的转变，电力业务由“非实时”向“准实时”的突破。（2）面向大规模综合能源数据，系统地提出了基于联合隐变量分解的多因素能源使用特性细粒度分析框架。综合考虑了用户用能信息和外部多源异构信息，对用户在多因素影响下的用能特性进行了细粒度分析与建模。（3）在细粒度用能特性分析框架基础上，构建了将时间维度与空间维度相结合的多尺度综合能源需求预测模型，提出了一种面向智慧城市的综合能源需求预测的方法，提升能源供应规划和营销策略的优化与决策支持。（4）打通电力数据与政府和其他行业数据通道，实现了面向智慧城市的综合能源信息应用服务场景，并利用 GIS 技术实现配电网分析和用户用能特性分析的可视化。本项目已建成一套完整的综合能源数据分析平台，支持 PB 级别数据存储，接入 18 个以上综合能源及经济社会类型数据，提供 12 个以上综合能源应用服务场景。大规模真实数据实验结果表明，本项目实现的多尺度、细粒度综合能源数据分析方法，在用户用能特性分析准确性方面提升 12%。研究成果已成功应用于包括中新天津生态城国家智慧城市试点建设项目在内的百余个智慧城市实际工程项目，总计产生直接经济效益 21733.07 万元，对推进我国智慧城市建设具有重要的经济和社会意义。依托本成果，出版专著 1 本，授权发明专利 6 项，实用新型专利 2 项，7 项发明专利进入实审，获软件著作权 4 项，在国内外重要刊物和国际会议上发表论文 32 篇，其中 SCI 收录 4 篇。鉴定意见认为研究成果达到了国际领先水平。

## 数据分析在资产管理行业的实践

林伟林 (况客科技)

时间: 08:30-09:00

**简介:** 林伟林, 况客科技联合创始人, 汇迪投资管理有限公司总经理, 厦门大学校外导师, 多年从事金融和数据分析领域, 致力于用数据分析提高资产管理行业的效率。

**摘要:** 本次演讲主要介绍在资产管理行业, 如何通过引入数据分析的方法和相应的 IT 系统化的解决方案提高行业的效率。本次演讲主要从以下几个方面探讨数据分析在资产管理行业发挥的作用。第一方面从客户服务这块, 通过数据分析的结果我们能更好的跟客户展示资产管理的过程; 第二方面, 我们在实际的投资决策中会大量的参考数据分析的结果; 第三方面, 我们通过对数据的分析, 可以帮我们做好风险控制以及客观的评价交易员的能力。

## 收益中的 Alpha 与 Beta

霍志骥 (博普资产)

时间: 09:00-09:30

**简介:** 霍志骥毕业于中国人民大学统计系, 本科期间在自营交易公司实习两年, 主要从事商品期货与外汇交易。毕业之后, 主要从事量化研究, 设计过一些套利和 CTA 策略。现就职于博普资产。

**摘要:** 一个资产组合的收益可以拆分成 Cash,Beta,Alpha,Lucky。做量化投资希冀于获得一个好的回报, 这个思路可以让我们理性的评估一个产品的收益, 从另外一个角度, 我们可以以此为标准来构建我们想要的投资组合并获得一个好的回报。

## 量化基本面投资与大类资产配置

赵然 (中信建投证券)

时间: 09:30-10:00

**简介:** 中国科学技术大学统计与金融系硕士, 2016 年加入中信建投证券研究所, 担任金融工程研究员。目前专注于大类资产配置及基本面量化相关研究, 研究成果包括大类资产配置框架, 宏观因子投资体系, 原油和黄金等资产的择时策略。

**摘要:** 在大多数认知里, 传统量化投资和价值投资似乎是两个不相容的概念, 价值投资关注经济逻辑, 量化模型笃信数学模型, 但我们认为两者并不矛盾, 经济逻辑是模型的支撑, 统计方法是有效的辅助工具, 特别是针对相对低频、数据量有限、数据质量不佳的基本面数据, 我们更需要建立经济逻辑与数量关系相互验证的策略。本报告以大类资产配置问题为例, 概述如何将一些统计学习的思想应用于投资实践中。

## 互联网征信的探索与实践

张云松 (天启智创)

时间: 10:30-11:00

**简介:** 张云松, 天启智创创始人, 毕业于中科院, 多年咨询公司和互联网公司从事数据算法、决策分析、风险管理的产品设计的工作。

**摘要:** 准确、快速地获得个人的征信类数据对互联网金融企业是至关重要的。传统征信数据覆盖度低、数据维度低、获取难度大, 很难满足行业的需求。互联网数据覆盖度高、数据维度高、获取容易, 自然就成为了互联网征信领域的关键数据。而怎么从各式各样的数据中准确地评估一个人的欺诈风险和信用水平是一个复杂的工作。本文就从数据、特征、模型角度来谈一下我们的探索和实践过程。

## 风险? 推荐?: 真实银行数据分析工作实践分享

谢军 (上海路瑞软件技术有限公司)

时间: 11:00-11:30

**简介:** 数据工作者, 牛津大学博士。30 年如一日, 6-12-6 工作方式。25 年金融领域经验。央行、银监会科技进步二等奖获得者, 农总行科技进步二等奖获得者。若干巨大银行和巨大证券公司重要模型设计者。目前的主要兴趣是计算几何, 自然语言理解, 业务是著书立传带徒弟。

**摘要:** 银行是一个面向市场的金融领域, 与咨询公司和学院的工作要求是不同的。在银行进行数据工程, 如数据整合与分析, 基于数据的机器学习, 你必须面临的挑战是拿出真正的经得起实践考核的产品。我将分享我的工作实践,, 一个投产的风控模型是什么样的, 它绝非搭建一个 TensorFlow 可以胜任; 我还将分享一行代码的推荐系统。两个模型一个复杂一个简单。这就是银行数据模型。希望能够帮助年轻同学建立严肃的数据科学姿态, 希望向业界传播反忽悠的数据工作态度。

## 量化风险管理 R 语言一键式建模探索

罗小勇 (快牛金科)

时间: 11:30-12:00

**简介:** 罗小勇, 毕业于武汉大学概率统计专业。专注金融保险行业数据分析、机器学习模型, 数据产品应用。现任快牛金科风控中心决策科学 Leader, 数据科学家, 负责集团风控模型, 智能创新, 数据产品研发和应用。

**摘要:** 1、量化风控管理介绍 2、风控数据模型提升应用 3、R 语言一键式建模探索

## 超融合、超高性能的车联网大数据平台

陶建辉 ( 涛思数据 )

时间: 08:30-09:00

**简介:** 陶建辉, 1994 年毕业于中国科大, 同年到美国印第安纳大学攻读天体物理博士, 曾在美国芝加哥 Motorola、3Com 等公司从事 2.5G、3G、WiFi 等无线互联网的研发工作, 国际顶尖无线数据专家。2008 年回到北京创办和信, 专注移动互联网 IP Push 和 IP 实时消息服务, 2010 年被台湾联发科收购。2013 年再度创业, 创办快乐妈咪, 专注母婴智能硬件和母婴健康服务, 2016 年初被太平洋网络收购。2017 年 5 月又再次走向战场, 创办涛思数据, 专注时序空间数据的实时高效的处理, 其产品 TDengine 比其他业内标杆数据库性能好 10 倍以上, 可广泛运用于物联网、车联网、工业大数据、金融等领域。

**摘要:** 随着车联网的兴起, 所采集的时序空间数据高速增长。一般的大数据解决方案都是 Kafka + Redis + NoSQL + Hadoop/Spark, 但这些套件都是用来处理通用的非结构化数据的, 因此在处理结构化的时序空间数据时, 效率就大打折扣。涛思数据的 TDengine 充分挖掘时序数据特点, 设计了独有的存储结构和时序数据处理模型, 将大数据平台所需要的数据库、消息队列、缓存、数据订阅等功能全部融合一起, 无论是数据插入、还是普通查询、流式计算, 速度都比现有方案快十倍以上, 而且大大降低了应用的开发难度和系统维护成本。并且, TDengine 的标签设计让大数据系统能轻松面对天天变化的业务分析需求。

## 区块链管理车联网数据

张翔 ( 车轮互联 )

时间: 09:00-09:30

**简介:** 车轮互联数据副总裁, 上海 R 组织筹建者, 11 年 COS 老水友。

**摘要:** 车轮互联开源项目 carro.io 使用区块链管理车辆资产以及数据。通过数据钱包进行数据加密, 保证用户对数据的绝对控制权。通过区块链智能合约记录数据的产生的授权关系, 保证公开透明不可篡改。通过分布式云存储对加密数据进行保存, 并与区块链的所有权关系保持映射, 避免区块链存储大文件的缺陷。最终基于该协议可以衍生出安全公平的车联网数据上报, 储存, 和交易使用平台。

## 大数据为二手车行业赋能

陈宸 ( 北京精真估信息技术有限公司 )

时间: 09:30-10:00

**简介:** 陈宸, 山东济南人, 本科毕业于山东大学数学系, 硕博毕业于北京科技大学模式识别专业, 从事机器学习 (CV 方向) 的研究, 博士课题涵盖图像处理、目标检测、识别、跟踪、三维重构等方向的算法研究, 曾就职于三星技术院、佳能研究院、电信科学技术研究院、三一集团, 现任北京精真估信息技术有限公司首席数据科学家, 负责公司机器学习 (数据挖掘、计算机视觉、自然语言处理) 相关算法的研究与产品化工作。

**摘要:** 1、中美二手车行业现状概述; 2、大数据时代的二手车行业; 3、一个例子: 二手车估值场景与算法; 4、机动时间: 如何成为一名数据科学家。

## 车联网数据与车辆可靠性研究实践

盛超 ( 彩虹无线 )

时间: 10:30-11:00

**简介:** 彩虹无线数据挖掘工程师, R 语言忠实粉丝, 拥有丰富的车联网数据应用实践能力, 服务过多个整车厂项目。

**摘要:** 车联网数据作为车辆实时状态评估的重要资源, 然而随着智能网联能力增强, 动力总成逐步迈向电动化, 车辆的即时通讯能力也大幅提升, 目前彩虹拥有的数据采集能力可以实现 10ms 级的 3000 个字段并发, 单车单日最高采集数据量达 10GB。因此, 围绕如何采集数据, 如何确认数据间的不同联系, 如何依赖于数据产品体系将车联网的数据进行高效利用, 并为车辆的实时可靠性研究提供辅助, 成为一个非常重要的话题。本演讲将围绕彩虹无线在车联网数据应用的丰富实践, 分享车辆数据与车辆可靠性的研究案例。

## LBS 数据科学实践

朱俊辉 ( Mobike )

时间: 11:00-11:30

**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:** 伴随着互联网下半场快速扩展的脚步, 在 LBS 领域, 数据驱动的方法论逐渐深入人心, 数据分析工具比如 Spark、ElasticSearch、Leaflet 等日渐成熟。本次演讲将主要从数据科学实战的角度, 谈一谈如何提升 LBS 数据分析效率。

## 车联网大数据的应用

耿文童 ( 车网互联 )

时间: 11:30-12:00

**简介:** 2014 年加入北京车网互联科技有限公司, 数据分析部总监。负责公司数据产品规划设计、行业应用建模和数据分析团队管理工作。基于车联网数据, 研发完成用户驾驶行为评价、用户出行分析等十余个可成熟商用的数据模型。作为主要发明人《车载加速传感器的三轴自校准方法及装置》、《基于车载数据的碰撞事件识别方法与装置》等数十项发明专利。

**摘要:** 基于车网互联积累的海量车联网实时数据, 围绕以车辆精细化管理, 智能网联基础研究, 驾驶行为分析等诸多车联网数据应用, 重点介绍车网互联在围绕车联网数据应用方面的实践案例。

## 数据科学与大数据技术人才培养体系

王艳 (欧亚学院)

时间: 09:00-09:30

**简介:** 西安交通大学在读博士, 西安欧亚校长助理, 金融学院院长, 陕西省统计学会副会长, 国家统计学会理事, 欧亚·狗熊会数据科学研究院院长。2016 年获得陕西省师德先进个人称号。所带领的数学教学团队获得了陕西省 2011 年“巾帼建功标兵岗”和 2015 年“优秀教学团队”的称号; 所带领的通识教育学院团队获得 2012 年陕西省“先进集体”和 2015 年陕西省“巾帼标兵岗”。

**摘要:** 在大数据时代, 社会各行各业都需要大量数据分析人才, 数据分析师已成当下中国互联网行业需求最旺盛的人才职位之一。企业大数据研发部门面临人才短缺、招聘困难、新入职人员缺乏应用实践技术结合能力、人才技能培训周期长及内部人才流失风险高等问题。为此, 部分高校正在积极探索数据分析人才的培养模式。高校也同样面临一些挑战: 教学缺乏优质权威的教材、师资无法满足教学要求、实训缺乏真实案例、产学研成果转化途径少、过程难等问题。通过对这些问题的思考, 对数据分析人才的应用场景、岗位需求分析的基础上, 结合数据科学与大数据技术专业的内涵和特征, 提出了一套应用型本科院校数据科学与大数据技术分析人才培养模式, 给出了明确的培养目标、课程体系的设置、实践教学及平台的建设意见。

## 证券分析师的价值分析

吴睿 (欧亚学院)

时间: 09:30-10:00

**简介:** 吴睿, 西安欧亚学院教师, 西安交通大学在读博士, 欧亚·狗熊会数据科学研究院副院长, 北京大学访问学者, R 语言论坛主讲嘉宾, 美国大学生数学建模竞赛及全国大学生数学建模竞赛指导教师, 2017 年陕西省劳动模范。

**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。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, 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。

## 物流车辆风险评估

张俊丽 (欧亚学院)

时间: 10:30-11:00

**简介:** 张俊丽, 西安欧亚学院教师, 西安交通大学在读博士, 欧亚·狗熊会数据科学研究院副院长, 专长数据分析与挖掘、数学建模, 2015 年“狗熊会”熊学院彩虹无线第一期优秀毕业生, 2016 年数据与价值论坛演讲嘉宾, 2017 年陕西省数学建模优秀指导教师。主持并参与了物流车辆数据分析、天然气用气量、电信用户流失分析、核电站继电器等多项校企合作项目。

**摘要:**车联网是指由车辆位置、速度和路线等信息构成的巨大交互网络。车联网主要依托移动通信与信息科学技术,通过无线通信技术、地理位置定位技术、汽车传感器技术以及行车记录仪技术等完成车辆行驶状态与周边环境采集、数据的传输与处理工作。基于车联网数据、车险数据为每辆车的驾驶行为进行多维度风险评级打分,对每辆车进行客户画像,给出每辆车驾驶行为的改进建议。通过对车联网数据的分析,会为保险公司、驾驶司机以及汽车制造商地带来商业价值。

## 西安市名牌战略实施效果调研

贾蓓 (欧亚学院)

时间: 11:00-11:30

**简介:**贾蓓,西安欧亚学院金融学院经济统计学专业负责人,专业研究方向为数据挖掘技术与软件应用。具有工信部数据分析师职业资格,负责并参与 30 余项数据挖掘与分析相关企业项目,完成“建行银行信用卡早期用卡与盈利关键行为分析”项目、“大明宫建材家居服务满意度研究”项目、“金龙鱼促销效果”评估项目、“对标科技股票分析师业务评测”项目等。

**摘要:**实施名牌战略,是西安市实施质量强市战略的重要举措。截止 2016 年,西安市已有 252 家企业申报了名牌产品,名牌产品数量达到了 330 个,政府部门为进一步提升名牌战略工作在实现“追赶超越”、建设“品质西安”中的作用,西安市政府部门委托西安欧亚学院金融学院对实施名牌战略在西安市经济社会发展的贡献情况进行调查并分析研究,为其更好的发展实施名牌战略提供参考建议。

## 文学书籍的市场发展探究

孙旭 (欧亚学院)

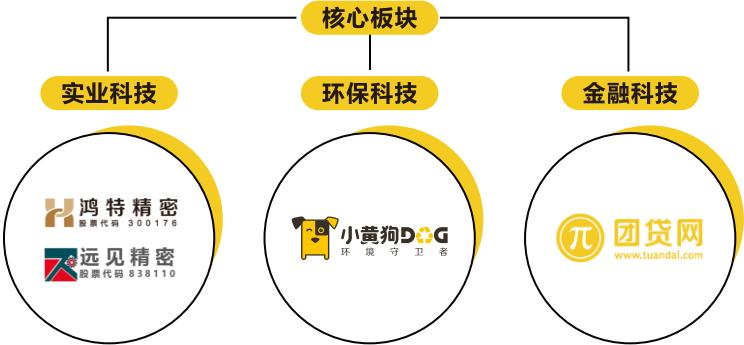
时间: 11:30-12:00

**简介:**孙旭,西安欧亚学院经济统计学专业 15 级学生,2017 年全国大学生数学建模竞赛国家二等奖获得者,第八届全国大学生高数竞赛三等奖,第三届“好贷杯”中国高校风险控制与管理能力挑战赛二等奖,2018 美国大学生数学建模竞赛三等奖。

**摘要:**文学是美的,高尔基说过“书是人类进步的阶梯”,阅读文学书籍能提高我们的审美能力,培养我们的生活情趣,净化我们的心灵。那么当今社会,文学书籍的阅读情况又是怎样的呢?通过在我国最大的图书销售网站当当网收集到文学类书籍的销售数据进行挖掘分析,以图书评论数来代表图书热度,建立了影响图书热度的数据模型,判断文学图书的受欢迎程度,并对有销售潜力的中小学类文学图书进行深度分析,可以对销售平台提供合理的营销策略,在提高平台销售量的同时又在青少年群体中推广了文学图书的阅读,提升了孩子们的文学修养。



派生科技集团有限公司（以下简称“派生集团”）于2011年在东莞成立，注册资本10亿元。聚焦实业、科技、金融三大战略投资方向，致力于成为一家以“**大数据、人工智能、互联网科技**”等技术力量驱动产融结合、提升产业运营效率的投资服务集团。目前集团主要有实业科技、环保科技、金融科技等核心业务板块，集团员工近20000人。2017年，派生集团核心业务板块在东莞合计纳税超过2.86亿元。



### ● 投资机构 ●

#### 海慧科技

实力雄厚投资机构，拥有丰富的产业资源

#### 盈生创新

多元化投资的经营理念，致力于投资高成长性项目

#### 民生资本

泛海资本全资控股，凭借杰出的创投表现倍受业界肯定

#### 北海宏泰

稳步发展的投资机构，已成功投资多家优秀企业

#### 巨人投资

商界传奇人物史玉柱先生创办建立，成功投资多家知名上市公司

## 新物种“小黄狗”

用智能化解决  
垃圾回收的最后一公里

小黄狗智能垃圾分类回收机，通过进驻城市居民社区、写字楼、酒店、闹市区及其他公共区域，以便捷、高效的有偿回收方式接收市民投放的废纸、塑料、金属、废旧纺织品、玻璃等废弃物。同时，公司推出“小黄狗”智能垃圾分类回收终端交易平台，并且建立完善的回收装运体系，打通线上线下回收行业生态圈，有效的将广大群众、废品回收商、再生资源产业、垃圾处理事业单位等有机整合，打造一套完整的废品回收生态链，实现用户投放现金返利、垃圾智能分类、回收商分区域托管收运、有害垃圾集中有效处理，响应国家政策导向，用科技改善垃圾分类现状，解决了垃圾回收最后一公里的难题。

目前小黄狗已经拥有完善的智能设备生产基地和近百人的专业技术开发运营团队，其智能垃圾分类回收终端+大数据云服务平台集人工智能、大数据、云服务、物联网、实时支付系统等科技前沿技术于一体，现已取得4项软件知识产权，外观专利顺利受理。同时公司也是中国循环经济协会垃圾资源化专委会、资源强制回收产业技术创新战略联盟理事单位。



“小黄狗”是派生集团旗下子公司  
小黄狗环保科技有限公司自主研发的  
**智能垃圾分类回收机**



关注我们

## 主办方



人大统计学院



光华管理学院  
Guanghua School of Management



统计之都

## 协办方



狗熊会  
CluBear

狗熊会



应用统计科学研究中心