中国R会
The China-R Conference

第14届中国R会

The 14th China-R Conference (Beijing)

地点：

线下会场：北京-中国人民大学

线上会场：腾讯会议

时间：2021.11.20--2021.11.21

欢迎辞

经过十四年的磨砺，中国 R 会议又踏上了新的征程。每当这个时候，各位志同道合的朋友以 R 为相聚的理由，从数据科学的各类学术领域而来、从大数据的各种应用行业而来、从天南海北的各条奋斗战线而来，欢聚一堂，共襄盛举。这是 R 的独特魅力。R 的一个核心设计理念是“人的时间永远比机器的时间宝贵”，具有深厚的人文精神，其工程化应用又秉承了“总是有多种方法来做同一件事”的思想，极具包容性。它专注于数据科学和统计建模，保持自己的勃勃生机，又主动和其他的优秀工具融合，让大数据时代的舞台群芳竞艳。这也正如统计学，最大的好处是“可以在所有学科的后院玩耍”。参加会议的朋友们都热爱 R，但不执着 R，甚至不用 R，大有“圣人不凝滞于物”的境界。

这么多年来，数据领域的各种热门词汇层出不穷，和 R 比较的工具也换了好几轮，但 R 和 R 会一直在这里，这里没有人想一统天下，只想解决现实问题，因为我们知道“所有模型都是错误的，但有些是有用的”。迎着国家产业升级的历史进程和大数据时代的热潮，此次 R 会的主题包含但不限于：数理统计学、数据科学与大数据、人工智能的相关理论及其在各行各业的具体应用，包括机器学习、医疗健康、金融经济、软件工具、天文地理、社交网络等诸多话题。我们真诚地欢迎您的到来，一同感受数据科学为这个时代带来的惊喜与挑战。

统计之都敬上

2021 年 11 月 20 日

目录

会议相关单位

主办方

统计之都

CAPITAL OF STATISTICS
PROFESSION, HUMANITY & INTEGRITY

统计之都（Capital of Statistics，简称 COS，网址 https://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，理事会现任主席为常象宇。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。

中国人民大学统计学院

中国人民大学 RENMIN UNIVERSITY OF CHINA | 统计学院 SCHOOL OF STATISTICS

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

中国人民大学应用统计科学研究中心

中国人民大学应用统计科学研究中心是中华人民共和国教育部所属百所人文社会科学重点研究基地之一，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。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。中心着力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、数据科学与大数据统计等五个各具特色的研究方向，围绕各个方向的统计理论创新与应用建设重点研究平台，获得丰硕的研究成果。“十四五”期间，中心将围绕经济社会的数字化转型展开科研攻关，继续为统计学科的发展提供支撑平台。

承办方

中国人民大学统计学院数据科学与大数据统计系

中国人民大学统计学院数据科学与大数据统计系成立于 2020 年，它起源于 2014 年发起的大数据分析五校联合硕士项目以及统计学院自 2017 年开始提供的数据科学与大数据技术本科生项目。数据科学与大数据统计系致力于为不同专业背景（包括但不限于商业分析、金融科技、健康信息学、工程、数学以及计算机）的学生提供扎实的数据科学知识。我们的使命是培养未来的数据科学家。院系成员主要科研方向有大数据挖掘与统计机器学习方法、文本挖掘、消费者行为大数据统计分析、深度学习、大数据分布式计算，时空大数据分析、稀疏弱信号提取理论，大规模知识图谱方法，大数据网络技术及应用、图模型、高维数据统计分析、生物统计、分位回归、分层模型、计算机密集计算、极值和重尾分布等内容。数据科学与大数据统计系的愿景是把握机遇和挑战，发展具有持久的区域和全球社会影响的世界一流的数据科学中心。

赞助商

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 会议。为了 R 语言能更持续稳定发展，RStudio 倡议与微软、Tibco、Google 等几家商业公司成立了 R 联合团体（RConsortium），每年为 R 社区的开源项目提供大量资助，召集优秀人才解决 R 语言现存的重要且有挑战性的问题。

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）目前由线下与线上两部分构成。COS 线下活动总结：

1. 中国 R 会：目前已开展到第十四届，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、哈尔滨等地举办。历届会议纪要和幻灯片共享都可以在 COS 主站上找到：`http://china-r.org/`；

2. 线下沙龙：目前我们在北京、上海和广州深圳开展线下沙龙活动。不同于规模庞大的 R 语言会议，沙龙形式更为轻巧，注重讨论交流。目前已经举办过 50 期，主要在北京、上海举办，详情参见统计之都主站及微信公众号；

3. 海外在线视频沙龙：我们在 Google Hangouts 举办在线沙龙，主要由海外嘉宾来分享学术、生活中的点点滴滴。目前已经举办 23 期：`http://meetup.cos.name/`；

4. 书籍出版，包括写作和翻译。如《Dynamic Documents with R and knitr》(2nd edition) 谢益辉著,《Implementing Reproducible Research》谢益辉等著,《bookdown: Authoring Books and Technical Documents with R Markdown》谢益辉著,《数据科学中的 R 语言》李舰、肖凯著,《R 语言实战》高涛、肖楠、陈钢翻译,《ggplot2：数据分析与图形艺术》统计之都翻译,《R 语言核心技术手册》刘思喆、李舰、陈钢、邓一硕翻译,《R 语言编程艺术》陈堰平、邱怡轩、潘岚锋等翻译,《R 数据可视化手册》肖楠、邓一硕、魏太云翻译,《R 语言统计入门》邓一硕、郝智恒、何通翻译,《数据科学实战》冯凌秉、王群锋翻译,《R 语言实战》(第 2 版) 王小宁、刘撷芯、黄俊文翻译,《Rcpp: R 与 C++ 的无缝结合》寇强、张眸翻译,《R 绘图系统》呼思乐、张眸、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》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、冯璟烁著,《统计之美：人工智能时代的科学思维》李舰，海恩著,《现代统计图形》赵鹏、谢益辉、黄湘云著等等。

5. 线上内容主要包括主站（`http://cosx.org/`）和微信公众号（“统计之都”或搜索“CapStat”）。疫情当前，线下活动开展多有不便，2021 年 10 月，统计之都正式推出

COStudy 数据科学讲座，借助腾讯会议平台，以提问和讲述相结合的方式对数据科学、教育学习等问题进行深入探讨。COStudy 第一讲由中国人民大学统计学院教授吴喜之主讲，录屏可在公众号和哔哩哔哩（id 均为“统计之都”）观看。欢迎各位热爱数据科学的朋友持续关注和积极参与 COStudy 后续活动。

6. 在 Breiman《统计建模：两种文化》发表 20 周年之际，统计之都发起了征文活动，探讨统计学、数据科学的历史与未来、机遇与挑战、思想与技术，以启迪思考、开拓创新。当前约稿的文章受到中国乃至全球统计学者广泛阅读和讨论，总阅读量近 10 万。欢迎各位学界、业界人士共同参与！请联系邮箱：editor@cosx.org 或添加微信号（COStudy）讨论。

第十四届中国 R 会筹备委员会

主　席：孔令仁
秘书长：赵昊蛟
副主席：刘中渊　黄昱翔　聂宇舟
秘书团：任怡萌　王祎帆　任　焱　向　悦　郝嘉欣　周瑾纯

日程表

20 日上午 主会场

本会场邀请了统计学界的大咖讲解统计学和数据科学的历史、现在与未来，既有前辈带我们重走统计学发展之路，也有年轻力量展示当代青年学者和业界工作者关注的重要议题，希望能够给听众带来多元的思维碰撞。

8:40-9:00	**开场致辞**		
9:00-9:45	线下	**袁卫** 中国人民大学	在伦敦大学学院应用统计系（UCL）的中国留学生（1926-1939）
9:45-10:30	线上	**张志华** 北京大学	机器学习的本质：预测、表示和计算
10:30-10:40	**自由讨论、休息**		
10:40-11:10	线下	**林毓聪** 北京理工大学	从海量医学文本与电子病历中挖掘医学知识
11:10-11:40	线下	**李舰** 九峰医疗	企业数字化管理中的 R 工具

20 日上午 机器学习专场

本会场主要介绍机器学习模型，特别是深度学习模型在理论与实践方面的最新进展，包括因果推断、联邦学习、强化学习和图神经网络等主题。

9:00-9:30	线上	**朱正丹** 滴滴	因果推断在网约车交易市场的应用实践
9:30-10:00	线上	**李翔** 北京大学	Statistical Estimation and Inference via Local SGD in Federated Learning
10:00-10:10	**自由讨论、休息**		
10:10-10:40	线上	**周帆** 上海财经大学	Non-crossing Distributional Reinforcement Learning
10:40-11:10	线上	**许斯泳** 北京邮电大学	用于链接预测的主题感知异质图神经网络

20 日上午 软件工具专场（一）

软件工具是开展数据科学研究的基石，数据科学领域近年来的巨大突破都离不开快捷高效的软件和插件的支持。本会场将由国内外数据科学领域的专家介绍他们独立或是参与开发

的工具、插件，包括 R、Python 等语言在内，希望可以借此机会增进数据科学领域内的交流合作。

9:00-9:30	线上	**袁凡**	echarts4r: 从入门到应用
9:30-10:00	线上	**谭显英** 安联保险资产管理有限公司	使用 ShinyProxy 部署 Shiny Apps
10:00-10:30	线上	**张敬信** 哈尔滨商业大学	Tidyverse 优雅编程：从向量化、泛函式到数据思维
10:30-10:40	**自由讨论、休息**		
10:40-11:10	线上	**张亦龙** 默沙东	R for Clinical Study Reports and Submission
11:10-11:40	线上	**李丰** 中央财经大学	海量数据驱动场景下的分布式统计计算

20 日上午 软件工具专场（二）

本会场为软件工具场的 R 包介绍专场，将由国内外数据科学从业者介绍他们熟悉或是开发的 R 包，希望可以增进大家对 R 语言的了解并有效帮助广大研究者的科研工作。

9:00-9:30	线上	**张丹** 北京青萌数海科技有限公司	用结构化数据的方式来管理文本
9:30-10:00	线上	**俞丽佳**	R 语言中制作图形动画的多种方法
10:00-10:30	线上	**古杰娜** 麦肯锡咨询公司	徒手开发零依赖的 Htmlwidget R 包
10:30-10:40	**自由讨论、休息**		
10:40-11:10	线上	**任坤** 上海明汯投资	Using R in VS Code
11:10-11:40	线上	**毛任飞**	用 R 包 gm 生成音乐

20 日下午 数据科学专场

本会场五位报告人均为中国人民大学统计学院青年教师，报告主题涉及数据科学研究领域诸多理论方法与应用，包括文本主题建模识别多语料的主题引领与滞后关系、基于卫星遥感图像处理技术的经济发展水平量化研究、基于交易流水的信用卡套现交易及商户识别以及因果网络、贝叶斯方法等。

14:00-14:30	线下	**王菲菲** 中国人民大学	Jointly Dynamic Topic Model for Recognition of Lead-lag Relationship in Two Text Corpora
14:30-15:00	线下	**吴奔** 中国人民大学	Bayesian Spatial Blind Source Separation via the Thresholded Gaussian Process
15:00-15:30	线下	**刘越** 中国人民大学	基于局部因果网络学习的因果作用估计方法

15:30-15:40	自由讨论、休息		
15:40-16:10	线下	**白琰冰** 中国人民大学	AI+ 卫星遥感量化经济发展水平
16:10-16:40	线上	**黄丹阳** 中国人民大学	基于交易流水的信用卡套现交易及商户识别

20 日下午 气候变化专场

气候变暖是最近几十年的热门话题。人类活动引起的气候变暖尽管在地理和大气科学界存在着绝大多数共识，但这种共识是基于对陆面过程的理解上的，而地质科学和天文学界在联合国政府间气候变化专门委员会（Intergovernmental Panel on Climate Change，IPCC) 上却只有些微约的声音。科学的发展史就是一部人类对自然界的认识偏差的纠正史。基于此，本专题召集了人类活动、气候变化、极端灾害和生态环境四个报告，从不同的角度来探讨气候变暖及其背景下生态环境的变化。

14:00-14:35	线上	**韩旭军** 西南大学	西南干旱研究中的多源数据分析
14:35-15:10	线上	**宜树华** 南通大学	UAVEE-Net: 基于无人机的长期协作生态环境研究网络
15:10-15:20	自由讨论、休息		
15:20-15:55	线上	**谭亮成** 中科院地球环境研究所	中亚超级大旱与史前丝绸之路
15:55-16:30	线上	**罗立辉** 中科院西北生态环境资源研究院	气候变暖驱动的人类活动空间数据发展

20 日下午 工业大数据专场

工业是国民经济的核心要素，是国家核心竞争力的重要组成部分。工业大数据作为制造业数字化转型与智能化升级的关键技术，受到了学界和产业界的普遍关注。本专场结合设备运维、生产质量管理、运作效率优化等主题，探讨数据分析的问题、挑战和算法技术。

14:00-14:30	线上	**宋哲** 南京大学	工业数据预测模型的泛化和自适应能力提升探讨
14:30-15:00	线上	**杜娟** 香港科技大学（广州）	Knowledge-Infused Sparse Learning for Quality Improvements in Smart Manufacturing Systems
15:00-15:30	线上	**曾聿赟** 北京工业大数据创新中心	数据驱动的电动矿卡工作流程分析
15:30-15:40	自由讨论、休息		
15:40-16:10	线上	**姚树亮** 无锡合全医药有限公司	R 语言和统计学在药品研发和生产质量管理中的应用
16:10-16:40	线上	**蒋宗敏** 西北工业大学	复杂电力装备的数字孪生 OKDD 模型

20 日下午 软件工具专场（三）

14:00-14:30	线上	**林枫** 统计之都	统计之都编辑部投稿流程
14:30-15:00	线上	**张桐川** 广州微远基因	二代测序公司的 tidyverse 实战总结
15:00-15:30	线上	**阿力木·达依木** 剑桥大学	consort：临床研究流程图构建工具
15:30-15:40	**自由讨论、休息**		
15:40-16:10	线上	**赵鹏** 西交利物浦大学	mindr：R 语言制作思维导图
16:10-16:40	线上	**黄湘云** 美团	开发企业级 Shiny 应用的技术栈

21 日上午 灾害风险专场

灾害风险专场围绕气候变化，灾害预警以及风险管理分享遥感无人机技术应用与数据科学方法。

9:00-9:30	线上	**杨璞** 伦敦大学学院	气候变化灾害评估对国家碳减排政策的影响——从 2018 年诺贝尔经济学奖谈起
9:30-10:00	线上	**董智捷** 美国德州州立大学	Social media information sharing for natural disaster response
10:00-10:30	线上	**徐青松** 德国慕尼黑工业大学	The remote sensing image perception-cognition framework for the large-scale disasters: algorithms and applications
10:30-10:40	**自由讨论、休息**		
10:40-11:10	线下	**金泊翰** 中国人民大学	基于无人机等多源遥感数据的三维重建和智能评估技术
11:10-11:40	线上	**翁旭涛** 北京理工大学	基于最小化复发风险的最优消融时间预测模型

21 日上午 学生专场

本届专场将由来自国内外顶级高校的硕士博士学生分享他们在数据科学领域的学习以及成果，方向涵盖统计理论、统计软件、机器学习及其在推荐系统、医疗健康等领域的应用，将站在学生的视角探讨数据科学领域的进展与运用，并希望借此机会展现优秀同学们的理解和能力，增进不同专业之间的交流。

9:00-9:30	线上	**朱进** 中山大学	abess: 快速最优子集选取软件包
9:30-10:00	线下	**袁深** 中国人民大学	Self-Organized Hawkes Processes

10:00-10:30	线上	**牛子昂** 宾夕法尼亚大学	High-Dimensional Instrumental Variables Additive Model
10:30-10:40	**自由讨论、休息**		
10:40-11:10	线上	**李哲** 复旦大学	Distributed Community Detection for Large Scale Networks Using Stochastic Block Model
11:10-11:40	线下	**全国瑞、涂富艺** 中国人民大学	健康大数据分析共享平台介绍

21 日上午 数据科学企业应用专场

本会场主要介绍数据科学在企业方面的应用，包括数据科学技术的实践落地、数据在企业的运用准则和策略等，在变化中寻求不变。

9:00-9:30	线上	**李晓矛** Google	用户数据保护法规与应对策略
9:30-10:00	线上	**任万凤** 便利蜂	数据科学-在变化中寻求不变
10:00-10:30	线上	**熊熹** 京东	互联网业务中的因果推断应用
10:30-10:40	**自由讨论、休息**		
10:40-11:10	线上	**肖一凡** 小马智行	自动驾驶从零到一
11:10-11:40	线上	**张源源** 百姓车联	基于手机传感器数据的危险驾驶行为识别

线下线上地址

11 月 20 日上午	11 月 20 日下午	11 月 21 日上午
主会场 线下会场：明德主楼 1030 腾讯会议 ID：849677225	数据科学专场 线下会场：明德主楼 1030 腾讯会议 ID：601590474	灾害风险专场 线下会场：明德主楼 1031 腾讯会议 ID：176729652
机器学习专场 线下观看地址：明德主楼 1001 腾讯会议 ID：947208588	气候变化专场 线下会场：明德主楼 1031 腾讯会议 ID：652697869	学生专场 线下会场：明德主楼 1030 腾讯会议 ID：596401958
软件工具专场（一） 线下观看地址：明德主楼 1031 腾讯会议 ID：979877494	工业大数据专场 线下观看地址：明德主楼 1001 腾讯会议 ID：519869785	数据科学企业应用专场 线下观看地址：明德主楼 1016 腾讯会议 ID：400752859
软件工具专场（二） 线下观看地址：明德主楼 1016 腾讯会议 ID：472359905	统计软件专场（三） 线下观看地址：明德主楼 1016 腾讯会议 ID：437659984	

主会场（20 日上午）

在伦敦大学学院应用统计系（UCL）的中国留学生（1926-1939）

袁卫 （中国人民大学） ***09:00-09:45*** 线下

简介：袁卫，中国人民大学荣誉一级教授，国务院学位委员会学科发展战略咨询委员会委员，教育部社科委经济学部委员，国际统计学会（ISI）选举会员。获国家有突出贡献的中青年专家、全国优秀教师。曾任第四届国务院学位委员会委员，第五、第六届应用经济学评议组召集人，第七届统计学评议组召集人。

摘要：在 1926-1939 年 14 年间，同时也是上世纪前半叶世界统计中心 UCL 的鼎盛时期，陆续有 7 位中国留学生和访问学者，在这个不大的统计系中学习、研究，他们不仅将最前沿的统计理论和方法带回中国，使得中国的统计教育和研究紧跟国际前沿，而且他们中的杰出代表吴定良、许宝騄也为世界统计学科的发展做出了贡献。

机器学习的本质：预测、表示和计算
Prediction, Representation and Computation—The Nature of Machine Learning

张志华 （北京大学） ***9:45-10:30*** 线上

简介：张志华，北京大学数学科学学院教授。之前曾经先后任教于浙江大学和上海交通大学，任聘计算机科学教授。主要从事应用统计、机器学习与人工智能领域的研究和教学。是国际机器学习旗舰刊物 Journal of Machine Learning Research 的执行编委，并多次受邀担任国际机器学习和人工智能顶级学术会议的高级程序委员或领域主席。讲授有网络公开课《统计机器学习》、《机器学习导论》、《应用数学基础》和《强化学习》等。2021 年 9 月，发起成立了中国现场统计研究会机器学习分会。

摘要：机器学习的发展给统计学带来了深刻的影响。Leo Breiman 在他发表于 2001 年的著名论文“Statistical Modeling: The Two Cultures”中首次讨论了统计学和机器学习之间的文化差异，提出了统计学专注“Data Modeling Culture”，而定义机器学习为“Algorithmic Modeling Culture”。Bradley Efron 在其 2019 年 ISP(International Statistical Prize) lecture 和随后发表的论文“Prediction, Estimation, and Attribution”中再次发人深思地探讨了经典统计学和现代机器学习的分歧，他把机器学习定义为“Pure Prediction Algorithms”，而用“estimation”和“attribution”来刻画传统统计回归方法。这个报告中试图用“prediction, computation, and representation”三元素来阐述机器学习的本质。特别地，从“representation”角度来看待机器学习，表明它的发展贯穿着如何解决“curse of dimensionality”和利用“dimensionality blessing”。深度学习则完美诠释了这两者之间的权衡，它也是迄今为止把“Data Modeling

Culture”和“Algorithmic Modeling Culture”融为一体的最佳技术途径。

从海量医学文本与电子病历中挖掘医学知识

林毓聪 （北京理工大学） ***10:40-11:10*** 线下

简介：林毓聪，北京理工大学医工融合研究院博士后，清华大学统计学研究中心博士，哈佛医学院蔡天西教授组访问学者。主要研究方向有医学信息学、医学知识挖掘与图谱构建、自然语言模型构建、神经网络建模等。主要工作发表在医学信息学一区期刊 Journal of Biomedical Informatics, Computer Methods and Programs in Biomedicine 与核心医学信息学会议 IEEE International Conference on Healthcare Informatics 中。

摘要：现如今，我们有着超过 2000 万篇医学大类论文以及数百 T 的医学电子病历数据，其中均蕴含着海量的医学知识。然而结构化好的医学知识——即医学知识图谱——才能够为智能医学诊疗提供机器可使用的医学知识支撑。然而，传统医学知识图谱构建的过程对专家标注的样本量的需求很大，为挖掘医学知识造成了主要瓶颈。在本讲座中，我将介绍 Hi-RES，一个用于高通量医学知识挖掘算法的框架。此外，我还将介绍一些正在进行的项目，包括 EHR 中的 EHR 知识挖掘和自动诊断框架，以展示医疗知识挖掘的巨大潜力。

企业数字化管理中的 R 工具

李舰 （九峰医疗） ***11:10-11:40*** 线下

简介：李舰，九峰医疗 CTO，“统计之都”核心成员之一。一直专注于数据科学在行业里的应用，参与编著了《统计之美》《数据科学概论》《数据科学中的 R 语言》，参与翻译了《R 语言核心技术手册（第 2 版）》《机器学习与 R 语言》。在 R 语言社区发布了 Rwordseg、tmcn 等包。

摘要：企业数字化转型是长期以来的热点，无论是传统行业还是高科技企业，在数字化管理方面都存在很大的提升空间，尤其《数据安全法》施行以来，对数字化管理提出了新的需求和挑战。演讲者结合自己的工作经验，探讨了 R 语言工具在数据安全、数据运营、精细化管理等方面的应用。

机器学习专场（20 日上午）

分会场主席：邱怡轩，普渡大学统计系博士，现为上海财经大学统计与管理学院副教授。研究方向包括统计计算、贝叶斯计算与推断、深度学习等，参与翻译了《应用预测建模》《R 语言编程艺术》《ggplot2：数据分析与图形艺术》等统计建模与数据分析方面的经典书籍，是 RSpectra、showtext、prettydoc、recosystem 等流行 R 软件包的作者。

因果推断在网约车交易市场的应用实践

朱正丹 （滴滴） ***9:00-9:30*** 线上

简介：朱正丹，滴滴专家算法工程师，花小猪价格策略负责人。

摘要：定价补贴是建立网约车交易市场良性生态和用户增长以及成本控制的重要抓手，运用因果推断来进行精细化定价补贴是滴滴的长期研究课题。本文将从数据、特征、构建模型以及评估等方面详解因果推断在滴滴的应用实践。

Statistical Estimation and Inference via Local SGD in Federated Learning

李翔 （北京大学） *9:30-10:00* 线上

简介：李翔是北京大学数学科学学院统计专业的博士研究生。其于 2018 年获得北京大学统计学、经济学 (双学位) 学士学位。主要的研究兴趣包括联邦学习、强化学习。其研究成果发于在 NeurIPS, ICLR, ICML 等国际会议。

摘要： Federated Learning (FL) makes a large amount of edge computing devices (e.g., mobile phones) jointly learn a global model without data sharing. In FL, data are generated in a decentralized manner with high heterogeneity. This paper studies how to perform statistical estimation and inference in the federated setting. We analyze the so-called Local SGD, a multi-round estimation procedure that uses intermittent communication to improve communication efficiency. We first establish a functional central limit theorem that shows the averaged iterates of Local SGD weakly converge to a rescaled Brownian motion. We next provide two iterative inference methods: the plug-in and the random scaling. Random scaling constructs an asymptotically pivotal statistic for inference by using the information along the whole Local SGD path. Both the methods are communication efficient and applicable to online data. Our theoretical and empirical results show that Local SGD simultaneously achieves both statistical efficiency and communication efficiency.

Non-crossing Distributional Reinforcement Learning

周帆 （上海财经大学） *10:10-10:40* 线上

简介：周帆，上海财经大学统计与管理学院副教授，美国北卡罗 莱纳大学教堂山分校生物统计学博士。主要研究方向包括深度学习，强 化学习，图网络，因果推断，多项研究成果发表于 JASA，Biometrics, Nature Genetics 等国际统计期刊和 NeurIPS, IJCAI，ICDM 等国际人工 智能会议，获得了泛华统计协会 New Researcher Award, 北卡生统系 Barry H. Margolin award 等奖项。

摘要： Although distributional reinforcement learning (DRL) has been widely examined in the past few years, there are two open questions people are still trying to solve. One is how to ensure the validity of the learned quantile function, the other is how to efficiently utilize the distribution information. To address these two issues, we first propose a non-decreasing quantile function architecture to guarantee the monotonicity of the obtained quantile estimates and then design a general exploration framework for DRL which utilizes the entire distribution of the quantile function. By comparing with some competitors, we are able to show that our method can achieve better performance on Atari 2600 Games especially in some hard-explored games.

用于链接预测的主题感知异质图神经网络

许斯泳 （北京邮电大学） ***10:40-11:10*** **线上**

简介：北京邮电大学石川教授团队三年级硕士生，研究方向为图神经网络，数据挖掘和机器学习。目前已在 ACL、CIKM 等国际会议上发表论文。

摘要：异质图神经网络（HGNNs）是一种可以聚合异质结构和属性信息的图表示学习方法。尽管 HGNNs 其捕获丰富语义的能力可以揭示节点不同方面，但它们仍然停留在简单地利用结构（例如元路径）的粗粒度级别。事实上，节点所包含丰富的非结构化文本内容，承载着由多方面主题感知因子所产生的潜在更细粒度的语义，这从根本上揭示了不同类型的节点会进行链接并形成特定的异质结构的原因。因此，本文提出了一个用于链接预测的主题感知异质图神经网络 THGNN，来层次性地挖掘主题感知语义并用于学习 HGs 中链接预测的多方面节点表示。在三个真实 HGs 上的实验结果表明，我们的方法在链接预测任务中优于最先进的方法，并体现了所学多方面主题感知表示的潜在可解释性。

软件工具专场（一）（20 日上午）

分会场主席：谢益辉，爱荷华州立大学统计学博士，现为 RStudio 软件工程师，曾负责 Shiny 包相关开发工作，后转入 R Markdown 相关扩展包的开发，包括 bookdown 和 blogdown。对统计计算、可视化、以及各类网页相关技术感兴趣，有志于对技术写作工具做减法工作。个人主页：`https://yihui.org`

echarts4r: 从入门到应用

袁凡 ***9:00-9:30*** **线上**

简介：东北财经大学统计学硕士，现从事数据分析类工作，R 语言新手。

摘要：本报告将对交互式绘图工具——echarts4r 包进行介绍。

使用 ShinyProxy 部署 Shiny Apps

谭显英 （安联保险资产管理有限公司） ***9:30-10:00*** **线上**

简介：CFA, 南开大学精算学硕士, 热爱编程, 对可重复化研究和报告充满热情, 是 data.table 项目成员及 DT 包共同作者，曾解决过许多 R 包中的字符编码问题，现任职安联保险资产管理公司组合及量化管理部负责人。

摘要： ShinyProxy 是 Open Analytics 开发的一个基于 Java 的开源软件，主要用于企业级地部署 Shiny Apps。它提供了统一安全的用户管理和登录验证等功能，支持并发，并且通过 Docker 技术来管理 App 环境。本报告将会为观众讲解 ShinyProxy 的工作模型，Docker 的基本概念，并通过示例展示如何个性化扩展 ShinyProxy，最后会与 RStudio Connect (Shiny Server) 进行简单比较。

Tidyverse 优雅编程：从向量化、泛函式到数据思维

张敬信 （哈尔滨商业大学） ***10:00-10:30*** 线上

简介：张敬信，博士，副教授，哈尔滨商业大学数学与应用数学系主任，数学建模主教练，主讲课程：高等数学、实变函数、泛函分析、数学建模、R 语言、数据挖掘等。发表 SCI 论文 4 篇，主持黑龙江省哲学社科项目 1 项，省教育厅科技项目 1 项，参加国家青年自然科学基金项目 1 项；即将出版《R 语言编程：基于 tidyverse》（人民邮电出版社）、《数学建模：算法与编程实现》（机械工业出版社）。常驻知乎平台，关注 7 万 +。

摘要： Tidyverse 代码整洁流畅、像文字叙述一样自然，是内在的编程思维在起作用。本报告将从向量化、泛函式编程谈起，再到数据思维、分解思维，梳理内在的编程思维脉络，并结合若干实例探讨如何用 tidyverse 优雅编程。

R for Clinical Study Reports and Submission

张亦龙 （默沙东） ***10:40-11:10*** 线上

简介： Yilong Zhang, Ph.D. is a statistician from Merck. Yilong works on late-stage clinical trial development in diabetes, cardiovascular, and oncology. He also works with a group of statisticians and programmers to demonstrate the capability of using R for regulatory submission. Yilong has published 20+ peer-reviewed papers including statistical methods in study design, missing data, and survival analysis. Before joining Merck, he earned Ph.D. degree in Biostatistics at New York University.

摘要： The use of open-source R is evolving in drug discovery, research and development for study design, data analysis, visualization, and report generation in the pharmaceutical industry. The ability to produce tables, listings and figures (TLFs) in customized rich text format (RTF) using R is crucial to enhance the workflow of using Microsoft Word to assemble analysis results. We developed an R package, r2rtf, that standardizes the approach to generate highly customized TLFs in RTF format. Code examples are provided to create customized RTF tables and figures with highlighted features. Based on the TLFs generated by r2rtf package, we further discuss a proposed process to submit the works to regulatory agency using the pkglite R package. The work is available in https://r4csr.org/.

海量数据驱动场景下的分布式统计计算

李丰 （中央财经大学） ***11:10-11:40*** 线上

简介：李丰博士现任中央财经大学统计与数学学院副院长、副教授。博士毕业于瑞典斯德哥尔摩大学，研究领域包括贝叶斯统计学，预测方法，大数据分布式学习等。曾获瑞典皇家统计学会 Cramér 奖，主持和参与多项国家自然科学基金项目。李丰博士最新研究成果发表在统计计算和运筹期刊 Journal of Computational and Graphical Statistics，European Journal of Operational Research. 经济与管理学期刊 Journal of Business and Economic Statistics,

International Journal of Forecasting 等, 同时著有 Bayesian Modeling of Conditional Densities,《大数据分布式计算与案例》和《统计计算》。李丰博士在世界贝叶斯大会，国际预测大会等作过邀请报告。

摘要：随着海量数据的涌现，全新的商业管理和决策场景依赖灵活高效的数据科学方法。特别是在分布式存储和计算的新常态下，依赖分布式平台上的实时数据科学方法可以提供最佳商业决策。这对计算能力有了前所未有的挑战。如何快速在数据浪潮中迅速提供分布式统计模型的解决方案显得至关重要。本报告通过介绍海量数据驱动场景下 Spark 分布式平台的统计模型算法开发，为将来志在数据驱动的新兴领域的同学知识储备和能力提升提供一些启发。

软件工具专场（二）（20 日上午）

分会场主席：谢益辉

用结构化数据的方式来管理文本

张丹 （北京青萌数海科技有限公司） ***9:00-9:30*** **线上**

简介：” 张丹，R 语言实践者，北京青萌数海科技有限公司 CTO，微软 MVP，10 年以上互联网应用架构经验，在 R、Java、NodeJS、大数据、数据挖掘等方面有深厚的积累。精通量化投资交易策略，熟悉中国金融二级市场、交易规则和投研体系。熟悉统计学和数据学科方法论，在海关、外汇、金融等监管领域，都有成功落地的模型应用。著有《R 的极客理想：量化投资篇》、《R 的极客理想：工具篇》、《R 的极客理想：高级开发篇》，英文版图书被 CRC 出版集团引进，在美国发行。个人博客：http://fens.me 。”

摘要：在互联网的今天，我们每天都会生产和消费大量的文本信息，如报告、文档、新闻、聊天、图书、小说、语音转化的文字等。海量的文本信息，不仅提供扩宽的研究对象和研究领域，也为商业使用带来了巨大的机会。量化文本分析（Quantitative Analysis of Textual Data），一种新的方式，用结构化数据的方式来管理文本。quanteda 包，提出以语料库的形式管理文本，语料库被定义为文本的集合，其中包括特定每个文本的文档级变量，和整个集合的元数据。用户可以轻松地按单词、段落、句子甚至用户提供的分隔符分割文本和标签，按文档级变量将它们分组为更大的文档，形成基于逻辑条件的变量组合。

R 语言中制作图形动画的多种方法

俞丽佳 ***9:30-10:00*** **线上**

简介：临床检测诊断从业者，从事临床分子检测质量评价和计算生物学研究。

摘要：本报告将回顾 R 语言中制作图形动画的工具包和下游应用，包括 animation, gganimation, plotly, moveVis 等内容。

徒手开发零依赖的 Htmlwidget R 包

古杰娜 （麦肯锡咨询公司） ***10:00-10:30*** **线上**

简介：古杰娜，目前在麦肯锡咨询公司担任软件工程师，活跃于开源社区，热衷于用业余时间开发开源软件包。个人网站：https://www.jienamclellan.com/

摘要：Htmlwidgets 提供了一个可以创建 R 链接 JavaScript 包的框架，使得在 R 环境下可以使用 JavaScript 可视化库，极大的方便和丰富了 R Markdown 文档和 Shiny Web 应用程序的开发和共享分析成果。本报告将介绍近期开发的 Htmlwidget R 包（faq：https://github.com/jienagu/faq 和 flashCard：https://github.com/jienagu/flashCard ）以及其从前端（JavaScript）到后端（R）开发流程。

Using R in VS Code

任坤 （上海明汯投资） ***10:40-11:10*** **线上**

简介：任坤，上海明汯投资资深投资经理，微软最有价值专家（MVP），R 语言开源社区的活跃贡献者，是 VS Code R 语言扩展以及 R Language Server 的主要开发者和维护者，也贡献于许多其他的 R 扩展包，例如 data.table, lintr 等。2016 年底出版了 Learning R Programming，中文版为《R 语言编程指南》。

摘要：As Visual Studio Code was ranked the most popular development environment in the Stack Overflow 2021 Developer survey, we have been constantly improving the R support in VS Code in the recent two years. In this talk, I will introduce the powerful code editor specifically for R users and focus on the vscode-R extension with its development and demonstrate its powerful code editing features based on the R language server and flexible interactivity with one or multiple R sessions.

用 R 包 gm 生成音乐

毛任飞 ***11:10-11:40*** **线上**

简介：毛任飞，基于科学的课程设计师 (science-based course designer) 和教师，有七年的教育经验，学生从学龄前至初中。课程主题包括创造力和非智力因素（non-cognitive factors）。独立 R 语言开发，作品有 R 包 gm，用来生成乐谱和音频。独立音乐人，作品专辑《夜》正在开发中，见`https://flujoo.github.io/en/my-music-album-night/`。擅长自学，所有工作和兴趣所需要的知识和技能均靠自学获得。

摘要：本报告将简单介绍 R 包 gm，包括如何用它生成音乐，如何在 RMarkdown 等环境中使用，以及如何用它来算法作曲。

数据科学专场（20 日下午）

分会场主席：吕晓玲，中国人民大学统计学院教授，数据科学与大数据统计系系主任，博士生导师，中国人民大学数据挖掘中心主任。本科与硕士毕业于南开大学数学系概率统计专业，博士毕业于香港城市大学管理科学系。曾经是奥地利约翰开普勒大学应用统计系以及美国加州大学伯克利分校统计系访问学者。一直从事统计机器学习、数据科学领域的研究。主持教育部人文社会科学研究项目以及中国国家自然科学基金项目。学术论文在 Journal of American Statistical Association, The Canadian Journal of Statistics、Statistics and Probability Letters、Knowledge-Based Systems、Statistics and Its Interface、Journal of Electronic Commerce Research、Public Personnel Management、Journal of Advertising Research 等 SSCI/SCI 检索的国际学术期刊发表。

Jointly Dynamic Topic Model for Recognition of Lead-lag Relationship in Two Text Corpora

王菲菲 （中国人民大学） ***14:00-14:30*** **线下**

简介：王菲菲，中国人民大学统计学院副教授。研究上关注文本挖掘及其商业应用、社交网络分析、大数据建模等，研究论文发表于 Journal of Econometric, Journal of Business and Econometric Statistics, Journal of Machine Learning Research, 中国科学（数学）等国内外高水平期刊上。主持并参与了国家自科基金项目、教育部社科重大项目、国家重点研发项目等多个课题。

摘要： Topic evolution modeling has received significant attentions in recent decades. Although various topic evolution models have been proposed, most studies focus on the single document corpus. However in practice, we can easily access data from multiple sources and also observe relationships between them. Then it is of great interest to recognize the relationship between multiple text corpora and further utilize this relationship to improve topic modeling. In this work, we focus on a special type of relationship between two text corpora, which we define as the ”lead-lag relationship”. This relationship characterizes the phenomenon that one text corpus would influence the topics to be discussed in the other text corpus in the future. To discover the lead-lag relationship, we propose a jointly dynamic topic model and also develop an embedding extension to address the modeling problem of large-scale text corpus. With the recognized lead-lag relationship, the similarities of the two text corpora can be figured out and the quality of topic learning in both corpora can be improved. We numerically investigate the performance of the jointly dynamic topic modeling approach using synthetic data. Finally, we apply the proposed model on two text corpora consisting of statistical papers and the graduation theses. Results show the proposed model can well recognize the lead-lag relationship between the two corpora, and the specific and shared topic patterns in the two corpora are also discovered.

Bayesian Spatial Blind Source Separation via the Thresholded Gaussian Process

吴奔 （中国人民大学） ***14:30-15:00*** **线下**

简介：吴奔，中国人民大学统计学院讲师，Emory 大学生物统计与生物信息系博士后，

Michigan 大学生物统计系博士后。主要研究方向为贝叶斯统计、独立成分分析、神经影像数据分析、金融高频数据分析等。

摘要： Blind source separation (BSS) aims to separate latent source signals from their mixtures. For spatially dependent signals in high dimensional and large-scale data, such as neuroimaging, most existing BSS methods do not take into account the spatial dependence and the sparsity of the latent source signals. To address these major limitations, we propose a Bayesian spatial blind source separation (BSP-BSS) approach for neuroimaging data analysis. We assume the expectation of the observed images as a linear mixture of multiple sparse and piece-wise smooth latent source signals, for which we construct a new class of Bayesian nonparametric prior models by thresholding Gaussian processes. We assign the von Mises-Fisher priors to mixing coefficients in the model. Under some regularity conditions, we show that the proposed method has several desirable theoretical properties including the large support for the priors, the consistency of joint posterior distribution of the latent source intensity functions and the mixing coefficients, and the selection consistency on the number of latent sources. We use extensive simulation studies and an analysis of the resting-state fMRI data in the Autism Brain Imaging Data Exchange (ABIDE) study to demonstrate that BSP-BSS outperforms the existing alternatives for separating latent brain networks and detecting activated brain activation in the latent sources.

基于局部因果网络学习的因果作用估计方法

刘越 （中国人民大学） ***15:00-15:30*** **线下**

简介：刘越，中国人民大学统计学院讲师，主要从事于因果网络，因果推断，可信机器学习等方向的研究。多篇文章发表于 JMLR, TIST,UAI 等机器学习与统计学期刊及会议。

摘要：从观测数据中获得变量之间的因果作用和因果关系是数据科学的中心目标之一，近些年，基于贝叶斯网络的因果推断方法受到越来越多的关注。但是因果网络学习计算耗时严重。很难推广到大规模的情景。其次，通过观测数据我们只能学习到因果网络的等价类，而在这个等价类中我们往往无法获得唯一的因果作用，有的因果关系我们无法通过观测数据获得。鉴于此，本报告中，讲者将介绍将介绍因果图模型的局部学习方法，以及如何将局部学习方法应用于因果作用估计。

AI+ 卫星遥感量化经济发展水平

白琰冰 （中国人民大学） ***15:40-16:10*** **线下**

简介：白琰冰，中国人民大学统计学院数据科学与大数据统计系讲师/硕士研究生导师，中国人民大学杰出学者，博士毕业于日本东北大学，曾任加利福尼亚大学欧文分校计算机学院访问学者。主要研究方向为数据科学（深度学习、时空大数据分析、计算机视觉）及其在经济社会统计、环境统计及巨灾风险管理领域的应用研究。主要讲授课程包括统计学、并行计算和软件设计及研究生大数据分布式计算，作为主编编写中国人民大学出版社《数据科学并行计算》教材，曾获得 2020 年王宽诚教育基金会资助项目。

摘要：卫星遥感影像是一种重要的时空数据来源，当我们利用人工智能技术挖掘卫星遥感大数据，并进行系统地量化、分析和预测，那我们的社会经济就变成了可感知的活体。本

次报告内容关注当前国内精准扶贫这一热点问题大背景，针对当前国内 GDP 多以省级单位为单元、统计口径较为粗糙，而各级政府又特别关心本域经济发展的问题，首次系统的提出了基于卫星遥感影像数据和深度学习算法的中国大陆地区县级尺度 GDP 预测方法。本研究建立了一个迁移学习框架，将夜间灯光强度作为经济活动程度的一个代理变量，通过注意力机制卷积神经网络提取白天卫星遥感影像中的特征，将特定层的输出特征进行降维和统计计算，最后通过回归器来实现县级行政单元 GDP 的预测。本研究使用用 2017 年的卫星遥感影像数据对模型进行训练，利用训练的模型对 2018 年的卫星遥感影像数据进行 GDP 预测，预测结果值达到 0.70 的 R2。

基于交易流水的信用卡套现交易及商户识别

黄丹阳 （中国人民大学） ***16:10-16:40*** 线上

简介：黄丹阳，现任中国人民大学统计学院副教授，博士生导师，中国人民大学杰出青年学者。主持国家自然科学基金面上项目，青年项目，北京市社会科学基金青年项目等多项科研课题，曾获北京市优秀人才培养资助。长期从事复杂网络建模、大型网络计算、超高维数据分析等方向的理论研究工作。研究论文发表于国内外权威期刊包括 Journal of Econometrics, Journal of the American Statistical Association, Journal of Business and Economic Statistics, 以及《统计研究》等。

摘要：信用卡套现是一种威胁正常金融秩序的风险行为。有效识别具有套现风险的商户及其风险交易，对信用卡风控具有重要意义。传统的信用卡风险识别方法需要先积累大量的标注数据，对持卡人拥有充分的先验信息。而信用卡标注数据的稀缺大大限制了传统方法的应用。本文充分挖掘海量交易流水数据，提出基于无监督学习的套现交易及风险商户识别方法。一方面，该方法无需关于信用卡的标注数据或先验信息，能够以数据驱动的方式过滤行为异常的套现交易及风险商户，具有更广泛的应用前景。另一方面，该方法综合商户的交易金额属性，及商户与消费者之间的关联关系，构建出一系列可解释性强的套现风险指标，为风控管理提供直观的指导参考。基于某第三方支付平台实际数据的实证分析表明，本文方法能够有效区分具有不同行为表现、不同风险等级的商户群体，为实际的套现交易识别提供可靠的决策支持。

气候变化专场（20 日下午）

分会场主席：罗立辉，博士、研究员、博士生导师。主持了国家自然科学青年基金、中科院青年人才成长基金、中国博士后面上基金、中国博士后特别资助、国家自然科学面上基金、西部之光、揭榜挂帅等项目，承担了 973 计划、中国科学院 A 类先导专项、国家重点研发计划等子课题和专题。10 多年来长期在青藏高原、祁连山、黑河流域、东北大兴安岭等寒区旱区从事卫星遥感、无人机低空遥感与地面监测、模型开发等野外实验与应用研究工作。在国内外期刊发表论文 40 多篇，授权发明专利 7 项。

西南干旱研究中的多源数据分析

韩旭军 （西南大学） ***14:00-14:35*** 线上

简介：韩旭军，男，1980 年生，博士，西南大学地理科学学院教授，博士生导师，遥感大数据应用重庆市工程研究中心副主任。主要多源遥感陆面数据同化、模型集成与遥感大数据研究。先后承担国家自然科学基金面上项目 2 项、国家自然科学基金青年项目 1 项、科技部 973 项目专题 1 项和中科院重要方向项目等课题，参与科技部 863 计划等科研项目。在数据同化理论方法及应用领域开展了系统性研究开发了多源遥感陆面数据同化系统框架 DasPy，并开放源代码（`https://github.com/DasPy`）。

摘要：我国西南地区极端干旱事件频发，本研究利用多源遥感数据和陆面过程模型，获取地表水和土壤水的时空变化趋势，在此基础上对我国西南地区的历史干旱事件开展大数据分析，为西南地区干旱的预测预警机制研究提供思路。

UAVEE-Net: 基于无人机的长期协作生态环境研究网络

宜树华 （南通大学） ***14:35-15:10*** **线上**

简介：宜树华，南通大学教授，博士生导师，2006 年获得加拿大麦克马斯特大学博士学位，2006-2008 在美国阿拉斯加大学费尔班克斯分校开展博士后工作，于 2009 年和 2014 年分别获中国科学院百人计划项目和国家基金委优秀青年项目资助。长期从事寒区生态系统的模式模拟研究；初步建立了基于无人机长期-协同生态环境监测网络，在中国生态脆弱区设置了 3500 多个工作点，飞行 1.5 万余次，获得固定航点照片 20 余万张，开展了植被盖度、生物量、植物物种、高原鼠兔等研究。以第一作者或者通讯作者在 PNAS，GRL 等杂志上发表 SCI 论文 38 篇，被 Nature Geoscience、Nature Climate Change、IPCC 等引用。获得青藏高原青年科技奖（2015），大疆创新无人机开发者大赛第 3 名（2015），四川省科学技术进步一等奖（排名第三，2018），中国草学会 2019-2020 草业科技奖二等奖（排名第一）。

摘要：轻小型无人机在生态环境监测研究中发挥着越来越重要的作用，然而，各自为阵、缺乏长期定位监测研究的现状极大的限制了资源共享和监测研究的规范性。当前软硬件条件以及政策都非常有利于建立轻小型无人机的研究网络。本文介绍了无人机生态环境研究网络（UAVEE-Net）。UAVEE-Net 起源于 2015 年，是一个开放性平台，现有来自 18 个大学和研究所加入这一协作网络。通过协同观测已经在中国脆弱生态区设置了 3669 个工作点，执行飞行 1.5 万余次，在固定航点上拍摄了 1.87 万张照片。协作不仅仅存在于野外观测之间，也存在于野外以及室内分析之间，同时航拍公民科学家的启动也有助于更多的人为 UAVEE-Net 贡献力量。通过协作分析航拍照片而获得的信息已经初步应用于高原鼠兔、牦牛、羊道、植被盖度、地上生物量和物种等研究。UAVEE-Net 将继续开展协同观测以覆盖更多的区域获取更长的时间序列，并且在轻小型无人机搭载更多探头后监测更多的变量。网络将注重于（但不限于）从不同的角度解决如下的科学问题：物种丰富度和生产力的关系、生态系统突变实证以及物种空间分布。UAVEE-Net 将是现有长期生态环境联网观测、全球变化控制实验以及卫星遥感研究的有益补充。

中亚超级大旱与史前丝绸之路

谭亮成 （中国科学院地球环境研究所） ***15:20-15:55*** **线上**

简介：谭亮成，男，1980 年生，博士，中国科学院地球环境研究所研究员，博士生导师，地球环境学报执行主编。兼任西安交通大学和长安大学客座教授，Science Bulletin 特邀编委和 Scientific Reports 编委，是中国地理学会环境变化和环境考古专业委员会副主任委员、中

国第四纪科学研究会喀斯特与环境专业委员会委员、中国青藏高原研究会高寒环境与人类适应过程专业委员会委员。获陕西省青年科技新星（2015）、中科院青年创新促进会优秀会员（2016），陕西省首届杰出青年基金（2018）、第六届刘东生杰出青年奖（2019）。长期从事石笋与古气候研究，围绕全新世突变事件的时空特征、机制、影响和百年-年代际气候变率，在我国季风区、热带东南亚以及中亚干旱区开展大空间范围、高精度测年石笋的高分辨同位素和元素地球化学研究。发表论文 80 余篇，包括在 PNAS、Sci. Bull.、EPSL、GRL 等第一作者/通讯作者 SCI 论文 40 余篇。研究成果被包括 IPCC AR5、Science、Nature、PNAS、NG、NCC 等国际顶级期刊论文等引用超过 3000 次。部分成果被选为 Nature 出版集团的亮点文章、Science Bulletin 和 Quaternary Research 封面文章。先后主持国家自然科学基金青年、面上和重大项目课题，以及中国科学院青促会优秀会员和“西部之光”人才培养项目，参与国家重点基础研究规划（973）、国家重大科学研究计划、国家重点研发计划及中国科学院战略先导专项等。

摘要：中亚干旱区是丝绸之路的核心区，也是史前人类扩散和东西方文化交流的重要通道。该区气候干旱，植被和生态群落受制于水文因素。由于有精确年代控制的高分辨率古气候记录的缺乏，中亚地区全新世水文变化及极端干旱事件在史前人口迁移和跨欧亚大陆文化交流中扮演的角色还不清楚。报告基于来自吉尔吉斯斯坦的石笋多指标记录，重建了中亚干旱区目前年代最精确（测年误差 6‰）、分辨率最高（ 3 年）、涵盖中晚全新世的降水变化序列。发现在 5820-5180 a BP 期间，中亚存在一次持续时间为 640 年的超级干旱事件，主要受西风带北移的影响。这次干旱事件对中亚史前文化的发展产生了重要影响。

气候变暖驱动的人类活动空间数据发展

罗立辉 （中国科学院西北生态环境资源研究院） ***15:55-16:30*** **线上**

简介：罗立辉，博士、研究员、博士生导师、中共甘肃省委十三五规划专家。本硕博先后于兰州大学、中国科学院大学、德国马普生物地球化学研究所就读。先后主持了国家自然科学青年基金、中科院青年人才成长基金、中国博士后面上基金、中国博士后特别资助、国家自然科学面上基金、西部之光、揭榜挂帅等项目，主持了 973 计划、中国科学院 A 类先导专项、国家重点研发计划等子课题和专题。10 多年来长期在青藏高原、祁连山、黑河流域、东北大兴安岭等寒区旱区从事卫星遥感、无人机低空遥感与地面监测、模型开发等野外实验与应用研究工作。在国内外期刊杂志发表论文 40 多篇，授权发明专利 7 项。

摘要：自第一次工业革命以来，人类活动已经深刻影响了地球各圈层，且这种影响还将持续扩大和增强。青藏高原作为一个具有全球意义的生态系统单元，同时也是我国重要的生态安全屏障，在水土保持、生物多样性保护、水源涵养和碳收支平衡等诸多方面发挥着至关重要的作用。但近 30 年来，随着青藏高原人类活动范围的扩大和强度的快速增长，人类活动所造成的各种生态环境问题也日益突出，并严重影响着青藏高原生态功能的发挥。青藏高原人类活动强度空间数据的研究与制备，将有助于深入理解该地区人类活动的影响强度和范围，揭示气候变暖背景下人类活动的变化规律，对于进一步量化辨识人类活动与气候变化对生态系统的影响，以及促进该区域的可持续发展都具有重要意义。

工业大数据专场（20 日下午）

分会场主席：田春华博士，北京工业大数据创新中心/昆仑数据首席数据科学家，毕业于清华大学自动化系，此前就职于 IBM 中国研究院，屡次获得 IBM 全球研究部杰出成就奖和主

要技术负责人奖。2017 PHM Data Challenge 冠军队导师。连续三年担任中国工业大数据竞赛评委，宝洁全球大数据黑客马拉松评委。发表学术论文（长文）92 篇，拥有 79 项专利申请（40 项有效授权），曾担任 IEEE、INFORMS、ACM 等学术组织和国际学术会议分会主席、执行委员、国际学术期刊审稿人。在高端装备制造、石油石化、新能源、航空与港口等行业，帮助国内外领先企业，成功实施资产管理、运营优化等数据分析项目，参与工业大数据、工业互联网白皮书的编写，著有《工业大数据分析实践》等书籍。

工业数据预测模型的泛化和自适应能力提升探讨

宋哲 （南京大学） ***14:00-14:30*** **线上**

简介： 南京大学商学院教授，博士生导师，美国爱荷华大学（University of Iowa）工业工程博士、博士后。国际知名期刊 IEEE Transactions on Sustainable Energy 副主编；Journal of Intelligent Manufacturing 副主编。IEEE Power Engineering Society Letters, Industrial Engineering & Management 编委成员。在大数据分析建模和管理决策优化方向已经发表高影响因子国际期刊论文 30 多篇，被引用 2800 多次, 获中国和美国发明专利 13 项。

摘要： 数据预测模型是工业智能的关键技术，通过机器学习算法从历史数据“学”出来，然后部署到现场对工业系统未来状态进行预测。预测模型往往在训练时表现出良好的准确率，然而在工业领域，由于系统 A 的历史生产运行数据并不能够覆盖未来所有可能出现的状态，基于 A 历史数据“学习”的预测模型在部署到同类型 B 系统，或者遇到未曾“见过”的新场景时往往会表现出较差的预测性能，存在“一机一模型”现象，模型泛化能力弱；同时任何系统会存在“老化”等现象，从而造成预测准确率逐渐退化，模型缺乏自适应能力。当前研究偏重于“学习”算法创新，机器学习业务流程研究匮乏。如果把预测模型看成一个工业产品，那么生成这个产品的工艺流程以及后续运行维护保养则显得至关重要。学习算法只是这个生产过程中的一环，并不能决定最终产品质量和现场使用性能。因此我们有必要另辟蹊径，从业务流程视角来审视和优化机器学习，提高模型的泛化和自适应能力。

Knowledge-Infused Sparse Learning for Quality Improvements in Smart Manufacturing Systems

杜娟 （香港科技大学（广州）） ***14:30-15:00*** **线上**

简介： 杜娟，现任香港科技大学（广州）（筹）系统枢纽智能制造学域助理教授，广州市香港科大霍英东研究院副研究员，香港科技大学机械与航空工程系附属助理教授。于 2014 年获得哈尔滨工业大学英才学院学士学位，2019 年获得北京大学博士学位，2017-2019 年佐治亚理工大学联合培养。博士论文获得中国管理科学与工程学会优秀博士论文，北京大学优秀博士论文。主要从事数据驱动的智能制造系统质量改善研究工作。发表 13 篇高水平期刊论文，其中 3 篇论文获得国际运筹与管理科学学会 (INFORMS) 质量统计可靠性分会或数据挖掘分会最佳论文（提名）奖，并获得 6 项国家发明专利和 1 项软件著作版权，曾于 2018 年获得美国统计学会 (ASA) 质量与生产分会会议旅行奖。多次应邀在国际、国内权威学术会议作报告。研究获得中国国家自然科学基金等项目支持，2020 年入选上海市青年科技英才扬帆人才计划。现为 IEEE, ASME, IISE 和 INFORMS 的会员。更多信息可以参看 https://sites.google.com/view/juandu/home.

摘要： The rapid development of sensing and computing technologies has resulted in unprece-

dented data-rich environments in smart manufacturing systems, which brings significant challenges and great opportunities for quality improvements in smart manufacturing systems. With massive data readily available and manufacturing domain knowledge, it is desired to develop new computational methodologies for quality improvement that will realize process monitoring, prognostics, diagnosis, control, and intelligent decision making. This presentation will discuss research opportunities, challenges, and advancements in knowledge-infused sparse learning for quality improvement in smart manufacturing systems. Emphasis is given on the recent publications on (i) optimal shape control strategy for compliant part assembly via sparse learning and (ii) ranking features to promote diversity for fault diagnosis in manufacturing systems. Real case studies will be used to illustrate the developed new methodologies.

数据驱动的电动矿卡工作流程分析

曾聿赟 （北京工业大数据创新中心） ***15:00-15:30*** **线上**

简介：曾聿赟，男，1991 年 3 月出生。2018 年毕业于清华大学核科学与技术专业，获工学博士学位，同年进入清华大学博士后流动站，从事工业互联网与设备健康分析的研究工作，2020 年出站。2020 年起进入北京工业大数据创新中心，担任工业数据分析师，从事工业互联网相关的数据分析和挖掘工作。

摘要：电动矿卡的工作流程分析问题为基于电动矿卡工作过程中上传的状态监测数据和卫星定位数据等，识别矿卡工作中的各个工作环节，如装载、卸载、充电、检修等，并对其基本工作循环进行切分，是进行数字化矿山运营调度优化的基础工作。考虑到矿卡工作过程中工作区位置易发生改变的问题，矿卡的流程分析以工作区的智能识别为基础。分析过程中，将矿场离散化为一系列地块，首先通过统计各个地块内矿卡的状态数据得到特征，建立工作区识别模型，识别出各个工作区，然后根据矿卡路径上途径工作区的情况及其在各个工作区内的运行状态得到矿卡工作流程的分析结果。

R 语言和统计学在药品研发和生产质量管理中的应用

姚树亮 （无锡合全医药有限公司） ***15:40-16:10*** **线上**

简介：姚树亮，六西格玛黑带，曾就职于世界 500 强欧美跨国制药企业，国内某大型上市公司及前沿生物药研发公司。期间负责新项目的质量体系建立和维护，有多种剂型（片剂/胶囊/软胶囊/颗粒/微丸/干混悬剂，小容量注射剂/冻干/脂质体）的质量管理经验。且在体系维护/变更管理，偏差调查中有丰富的经验。生物药公司任职期间，负责公司的所有报告的数据分析，数据可视化，产品业务数据的数据可视化电子系统和服务器搭建。支持研发和生产业务。辅助产品开发和决策。

摘要：本报告介绍 R 语言和统计学在制药行业药品研发与生产质量管理中的应用。包括药品处方筛选与工艺研发、工艺转移评价及商业化生产的质量管理中的常规应用和 case by case study、trouble shooting 应用。

复杂电力装备的数字孪生 OKDD 模型

蒋宗敏 （西北工业大学） ***16:10-16:40*** 线上

简介：蒋宗敏，高级工程师，西北工业大学在读博士，中国西电集团科技带头人，长期从事电力能源领域物联网、设备智能化和数字系统研发工作。

摘要：电力装备作为电能传输、控制的载体，其可靠、高效和经济运行得到广泛关注。最近几年，电力行业大力开展数字孪生 + 不同场景的应用探索。提出了数字孪生体的 OKDD 模型以及实现方法。OKDD 模型使数字孪生体具象化，便于操作和扩展。通过该模型可以方便的构建从设备单元级-系统级-系统之系统级的多层级架构的复杂电网系统。OKDD 模型能够促进多业务系统信息集成、专业知识管理、沉淀与复用，利于算法、业务模型持续的迭代优化。它能够助力电力资产的精益运营，复制性强，为当前电力装备数字孪生体的统一信息框架提供了参考。

软件工具专场（三）（20 日下午）

分会场主席：谢益辉

统计之都编辑部投稿流程

林枫 （统计之都） ***14:00-14:30*** 线上

简介：林枫，美国华盛顿大学在读博士。本硕毕业于中国科学技术大学。研究兴趣是应用驱动的学习与决策问题。现任 COS 执行主编。

摘要：围绕“如何向统计之都投稿”这个问题，介绍三种投稿方式并实际演示一些相对复杂的步骤。

二代测序公司的 tidyverse 实战总结

张桐川 （澳洲格里菲斯大学） ***14:30-15:00*** 线上

简介：张桐川，广州微远基因生信科学家，R 语言爱好者，cosx 三年水友，中国科学技术大学本硕，澳洲格里菲斯大学博士，统计之都 id tctcab。

摘要：进入二代测序病原检测领域后，惊奇地发现部门里好多人都挺熟悉 r 和 tidyverse 并运用在生产管线里。这次报告的主题是分享一些有意思的观察以及个人一点微小工作。

consort：临床研究流程图构建工具

阿力木·达依木 （剑桥大学肿瘤学系） ***15:00-15:30*** 线上

简介：阿力木·达依木博士于 2014 年在北京大学取得预防医学学士学位，2014-2020 年间在

山东大学生物统计学系师从薛付忠教授获得博士学位。毕业后加入默沙东研发统计部门担任高级统计师，后于 2021 年进入剑桥大学肿瘤学系临床试验中心从事博士后研究，主要研究领域包括临床实验设计、纵向数据分析等方面的研究。

摘要：临床试验报告统一标准（CONSORT）声明中的提出的流程图已被广泛应用于报告临床试验研究中人群脱落和排除情况，此外它同时被其他优秀的期刊和组织应用在其他医学研究领域，以提高临床研究报告的质量。但是目前通常的做法是收集到数据以后手工计算并填入事先准备的 Word 文档，在数据逐渐累加的过程中不断进行更新调整，但是该方法低效且易出错。尽管已有 R 包可以绘制流程图，但是存在复杂度高或无法生成高质量流程图的问题。consort 包在兼顾易用性的同时提高灵活性，为医学研究者生成高质量流程图提供可靠的工具。

mindr：R 语言制作思维导图

赵鹏 （西交利物浦大学） ***15:40-16:10*** **线上**

简介：赵鹏，西交利物浦大学健康与环境科学系助理教授，博士生导师，英国利物浦大学荣誉学术成员。“统计之都”成员，mindr、rmd、bookdownplus、beginr 等 R 扩展包的作者。著有《学 R：零基础学习 R 语言》（赵鹏，李怡，2018）、《现代统计图形》（赵鹏，谢益辉，黄湘云，2021）。

摘要： mindr 是一个用于制作思维导图的 R 扩展包。它可以将 R Markdown 文档中的章节标题、R 脚本中的注释文字和常见的 FreeMind 思维导图三者之间两两任意转换，并且可以将思维导图以 HTML Widget 的形式展示和保存。此外，它还兼有以思维导图的形式呈现计算机中指定目录的功能。这里我们介绍一下 mindr 的主要功能和开发历程。

开发企业级 Shiny 应用的技术栈

黄湘云 （美团） ***16:10-16:40*** **线上**

简介：黄湘云，现任统计之都编辑，毕业于中国矿业大学（北京），多年 R 语言爱好者，与赵鹏、谢益辉一起合著了《现代统计图形》，目前在美团搬砖。

摘要： R Markdown 和 R Shiny 分别是什么？R Markdown 和 R Shiny 有什么能力？适合在什么样的场景下使用？一个完整的 R Markdown 和 R Shiny 应用通常都包含哪些部分？如何去学习和使用 R Markdown 和 R Shiny 两个工具？

灾害风险专场（21 日上午）

分会场主席：苏锦华，中国人民大学统计学院硕博一年级，百姓车联驾驶行为算法炼丹员，司马数慧技术负责人，在 Remote Sensing，地理研究，Ubicomp workshop 有多篇发表，2021EMNLP 审稿人，数次 nlp 投稿未果，含泪参与了统计计算、医学影像的工作在投，关注保险科技，无人机技术，NLP 技术。home lab 爱好者，图吧垃圾佬，欢迎交流捡垃圾心得。

气候变化灾害评估对国家碳减排政策的影响-从 2018 年诺贝尔经济学奖谈起

杨璞 （伦敦大学学院） ***9:00-9:30*** 线上

简介：伦敦大学学院巴特莱特可持续建筑学院在读博士，气候变化经济学论坛首届执行委员会成员。研究方向为气候变化经济学、二氧化碳碳减排政策和综合评估模型等。目前的研究主要是通过综合评估模型核算社会碳成本并进行国家减排目标的评估。在 One Earth, Global Environmental Change, Journal of Cleaner Production, Journal of Environmental Management, Renewable Energy 等期刊上发表多篇论文。

摘要：近年来，气候变化导致的极端天气灾害频发，如何促进碳减排行为是应对全球气候变化的重要挑战。2018 年诺贝尔经济学奖授予 William Nordhaus 教授，表彰他将气候变化纳入宏观经济分析所作的贡献。他所开发的综合评估模型从定量角度构建了经济系统与地球系统之间的相互作用，将避免的灾害损失看作减排收益，测算了经济最优水平的排放和碳排放社会成本，为减排政策提供了重要参考。从经济最优视角指定谈减排政策的核心是平衡减排成本和减排收益，因此，气候变化灾害评估（减排收益）会显著影响估计的减排政策。本次报告我将分享我对全球社会碳成本和国家碳减排政策如何受到气候变化灾害评估影响的一些思考。

Social media information sharing for natural disaster response

董智捷 （美国德州州立大学） ***9:30-10:00*** 线上

简介：董智捷博士现任美国德州州立大学工业工程系 tenure-track 助理教授，曾任联邦快递高级运筹科学家。获南京大学、哥伦比亚大学和康奈尔大学的本科、硕士和博士学位。长期从事应急交通运输管理的研究，开辟了应急物流中供应商选择机制领域，率先建立共享交通下应急疏散理论，并建立应急管理的机器学习优化算法。在 Omega，TR Part B 等期刊发表了十多篇高水平论文；主持了含美国自然科学基金（NSF）在内的约 150 万美金项目。荣获了美国 NSF CISE CRII Award，获评 NSF 应急事件领域最有潜力的青年学者、NSF CMMI Panel Fellow、哈佛大学商学院最佳数据科学和决策统计青年学者，及 INFORMS 2020 年 Early Career Award，并连续三年获得 IISE 年会最佳论文奖。现任期刊 Sustainable Futures 副主编及自然期刊系列的期刊 Communication Engineering 编委。

摘要： Social media has become an essential channel for posting disaster-related information, which provides governments and relief agencies real-time data for better disaster management. However, research in this field has not received sufficient attention, and extracting useful information is still challenging. The work will be presented aims to improve disaster relief efficiency via mining and analyzing social media data like public attitudes toward disaster response and public demands for targeted relief supplies during different types of disasters using machine learning models. The change of public opinion during different natural disasters and the evolution of peoples' behavior of using social media for disaster relief in the face of the identical type of natural disasters as Twitter continues to evolve are also studied. The research results demonstrate the feasibility and validation of the proposed research approach and provide relief agencies with insights into better disaster management.

The remote sensing image perception-cognition framework for the large-scale disasters: algorithms and applications

徐青松 （德国慕尼黑工业大学） ***10:00-10:30*** 线上

简介：徐青松，德国慕尼黑工业大学博士生，西湖大学访问学生。2018 年在成都理工大学获土木工程学士学位，其后保送至中国科学院成都山地灾害与环境研究所，于 2021 年获得岩土工程硕士学位。同年，获得国内外多所名校博士入学资格。目前在西湖大学人工智能实验室访问。研究领域主要有：遥感图像算法研究，机器学习与气候水文灾害交叉领域等。GeoCV 框架主要推动者之一。本科期间，先后获得国家级、省校级奖项 20 余项，如“地质 +”大学生创新创业大赛金奖，“互联网 +”大学生创新创业大赛银奖，四川省优秀毕业生等。硕士期间，连续三届获得中国科学院大学优秀学生，国家奖学金等。已发表 SCI 期刊 9 篇（其中第一作者 6 篇），如 IEEE TGRS、Landslides 等，计算机国际顶级会议 1 篇。获得 4 项发明专利，多项软件著作权。同时担任 SCI 期刊 Pattern Recognition, IEEE TGRS，计算机顶会 AAAI 等审稿人。

摘要： The perception-cognition framework focus on regional/global disaster (climate change, floods, earthquakes) assessment and early warning . Compared with the current remote sensing image processing methods based on machine learning, the biggest difference of our proposed perception-cognition framework is that the dynamic properties of disaster are considered and an end-to-end mapping and fusion of perception and cognitive data is realized. Specifically, it consists of three main parts: 1) Where is the scene target/change target in remote sensing images? (Perception: Where), 2) How to dig deeper into the spatial and temporal geographic information of the target? (Cognition: with whom), and 3) How to realize the refined management and fusion of these information (perception and cognition)? (active cognition: where to go).

基于无人机等多源遥感数据的三维重建和智能评估技术

金泊翰 （中国人民大学） ***10:40-11:10*** 线下

简介：中国人民大学环境学院环境科学专业本科生，现任人大无人机协会会长，主要研究方向为无人机和卫星遥感，目前已报送北大光华管理科学与信息系统直博生。曾主持大学生创业训练计划国家级项目：机器学习和无人机遥感在农村非正规垃圾场地监测中的应用研究，在 Remote Sensing 上发表全球甲醛健康风险相关研究一篇。目前在司马数慧公司从事无人机遥感数据采集与处理工作。

摘要：台风、暴雨、洪水、山体滑坡等重大灾害发生后，保险公司和救援人员需要第一时间了解企业、建筑及工程等受灾标的灾前和灾后情况，精确、快速的评估标的损失，以迅速开展施救和定损工作，减少客户损失，助力灾后恢复重建。基于无人机、遥感卫星的多源多模态大数据和深度迁移学习技术，可实现对灾情现场的快速 3D 重建、灾前灾后对应位置的标注和对比。利用深度卷机网络实现对现场房屋的实时识别，实现受损标的长度、面积、土方体积等的快速测量和评估。选取典型灾害场景、典型区域开展试点研究，验证方法的适用性和准确性，探索保险应用模式。

基于最小化复发风险的最优消融时间预测模型

翁旭 （北京理工大学） ***11:10-11:40*** **线上**

简介：翁旭涛，北京理工大学计算机院在读博士研究生。主要研究方向为结构化医学数据挖掘及其在诊疗上的应用。主要工作发表在 Computer Methods and Programs in Biomedicine 等国际期刊上。

摘要：有监督学习对于分析结构化电子健康病例 (Electronic Health Records, EHRs) 而言是一种重要的方法。然而，因其数据中往往存在噪音，利用 EHRs 进行有监督学习时常存在一些困难。当监督信息中存在不可忽视的噪音时会给模型带来监督矛盾的情况。同时, 治疗方案通常由经验丰富的医生进行合理的方案规划。然而, 由于不同医生的经验不同以及病患的个体差异, 其最终治疗结果可能呈现出不同情况。但是不同经验的医生对于治疗方案的规划会统一记录在电子病历中，则若单纯以记录信息作为监督也可能带来监督矛盾的情况，并最终影响模型预测性能。我们针对上述问题提出了一种方法，该方法根据训练样本中其他关联的监督信息，通过最优化方法对存在噪音的弱监督信息进行修正，其修正的方向可以使关联监督预测值偏向于期望的分布，从而降低噪音对于有监督学习的不良影响。

学生专场（21 日上午）

分会场主席：聂宇舟，中国人民大学统计学院数据科学与大数据专业本科在读，热衷于探索计算机技术和深度学习领域中多模态学习、生成模型和图神经网络等知识，目前正在完成医疗大数据交叉相关工作。此外努力在公众平台分享自己在专业领域的学习和理解，希望可以和同样有志于数据科学领域的同学交流，同时也是 Unix 爱好者 (初学者)，欢迎同好交流心得。曾参与 WTD! 项目收集分析顶级统计学家履历成果，并负责北京市级大创、科研基金项目。

abess: 快速最优子集选择软件包

朱进 （中山大学） ***9:00-9:30*** **线上**

简介：中山大学数学学院统计学专业博士生。主要研究领域为统计推断，高维数据分析等。研究论文发表于 Proceedings of the National Academy of Sciences、Journal of the American Statistical Association、Journal of Statistical Software 等杂志。开源统计软件社区贡献者，发布 Ball，abess 等多个统计软件，累计下载量逾 5 万次。

摘要：最优子集选择的目的是选取一部分变量构建模型，使得模型具有最高的准确性或解释性。最优子集选择在科学研究和实际应用中都有巨大的价值。我们将介绍一个用于处理最优子集选择的软件包 abess，它利用剪接技术以解决多种机器学习问题，如线性回归、分类和主成分分析。特别地，对于线性回归，abess 可以在多项式时间内以高概率获得最优解。我们的高效实现使 abess 求解最佳子集选择问题的速度与其它知名的变量选择软件一样快，甚至是快 100 倍。abess 软件包目前发布在 PyPI 和 CRAN，源代码可以在 github 上获取：`https://github.com/abess-team/abess`。

Self-Organized Hawkes Processes

袁深 （中国人民大学） ***9:30-10:00*** 线下

简介：袁深，中国人民大学高瓴人工智能学院 2021 级博士生，导师为许洪腾副教授，目前的研究方向为机器学习及其在医疗中的应用。

摘要： We propose a novel self-organized Hawkes process (SOHP) to model complex event sequences based on extremely few observations. Motivated by the fact that the complicated global relations among events are often composed of simple local relations, we model the event sequences by a set of heterogeneous local Hawkes processes rather than a single Hawkes process. In the training phase, we learn the Hawkes processes with a self-organization mechanism, selecting training sequences adaptively for each Hawkes process by a bandit algorithm. The reward used in the algorithm is originally defined based on an optimal transport distance. Additionally, we leverage the superposition property of the Hawkes process to enhance the robustness of our algorithm to the data sparsity problem. We apply our SOHP method to sequential recommendation problems in the continuous-time domain and achieve encouraging performance in various datasets.

High-Dimensional Instrumental Variables Additive Model

牛子昂 （宾夕法尼亚大学） ***10:00-10:30*** 线上

简介：牛子昂同学本科毕业于中国人民大学统计学院，现就读于宾夕法尼亚大学应用数学与计算科学专业。他的研究兴趣主要集中于计算统计、高维统计推断、因果推断、理论机器学习等方向。

摘要： Instrumental Variables Regression (IVR) is a fundamental tool for handling unmeasured confoundedness in causal inference. If measurements of multiple covariates X and the response Y are confounded by some unobserved variables, the casual effect cannot be identified due to confounding bias. If an instrumental variable Z is available, which satisfies the following: (1) it influences X directly; (2) it is conditionally independent of Y given X; and (3) it is independent of the unmeasured confounders, the causal effect can be identified. The classic two-stage least squares algorithm simplifies the estimation problem by first regressing the endogenous variables on the instrumental variables and then regressing the response variable on the fitted values obtained from the first stage. In this paper, we consider a flexible two-stage instrumental variables regression for high-dimensional data. Our model allows for non-linear relationship between the instrumental variables and the covariates, and allows data in both stages to be high-dimensional. We provide non-asymptotic analysis for the estimation errors of the parameters of interest. Moreover, we employ a debiased procedure to establish valid inference for the parameters using the framework in *Confidence intervals and hypothesis testing for high-dimensional regression*. Extensive numerical experiments show that our method yields consistent estimation and has more flexibility than existing methods in the literature. We also apply our method to a real-world dataset and obtain intriguing result about the genetic effects on mouse obesity.

Distributed Community Detection for Large Scale Networks Using Stochastic Block Model

李哲 （复旦大学） *10:40-11:10* 线上

简介：李哲，本科毕业于复旦大学物理系，目前为复旦大学大数据学院统计学专业研究生二年级，研究方向为分布式计算、高维数据建模、机器学习及深度学习。

摘要： With rapid developments of information and technology, large scale network data are ubiquitous. In this work we develop a distributed spectral clustering algorithm for community detection in large scale networks. To handle the prob- lem, we distribute l pilot network nodes on the master server and the others on worker servers. A spectral clustering algorithm is first conducted on the master to select pseudo centers. The indexes of the pseudo centers are then broadcasted to workers to complete distributed community detection task using a SVD type algorithm. The proposed distributed algorithm has three merits. First, the communication cost is low since only the indexes of pseudo centers are communicated. Second, no further iteration algorithm is needed on workers and hence it does not suffer from problems as initialization and non-robustness. Third, both the computational complexity and the storage requirements are much lower compared to using the whole adjacency matrix.

健康大数据分析共享平台

全国瑞、涂富艺 （中国人民大学） *11:10-11:40* 线下

简介：全国瑞、涂富艺均为中国人民大学统计与大数据研究院研究生。

摘要：我们旨在建立具备强大数据存储功能及大规模计算和分析能力的健康大数据分析共享平台，整合处理调查问卷、健康体检、生物样本等多种来源的数据，实现对增龄过程中机体重要器官与系统功能和健康风险因素等动态数据进行存储、处理、整合及分析，并实现平台的共享机制。

数据科学企业应用专场（21 日上午）

分会场主席：任万凤，毕业于北京大学数学院应用统计硕士，目前在便利蜂担任商品策略算法负责人，置身于产业互联网中拥抱数据科学的变化，当前研究方向包括不限于商品竞价、定价/促销、推荐、用户营销、用户增长等；之前在 51Talk 负责算法部，连续三年产出多个行业首创的数据科学产品，带领公司实现了产业数据化变革，见证了整个教育行业的兴衰史。业余也是一个英语启蒙教育大 V, 帮助全网实现云养娃。

用户数据保护法规与应对策略

李晓矛 （Google） *9:00-9:30* 线上

简介：李晓矛，现为谷歌数据科学家。

摘要：报告探讨欧美中三地的用户数据保护法规的形成过程，并总结主要的行业应对措

施。

数据科学-在变化中寻求不变

任万凤 （便利蜂） ***9:30-10:00*** 线上

简介：任万凤，毕业于北京大学数学院应用统计硕士，目前在便利蜂担任商品策略算法负责人，置身于产业互联网中拥抱数据科学的变化，当前研究方向包括不限于商品竞价、定价/促销、推荐、用户营销、用户增长等；之前在 51Talk 负责算法部，连续三年产出多个行业首创的数据科学产品，带领公司实现了产业数据化变革，见证了整个教育行业的兴衰史。业余也是一个英语启蒙教育大 V, 帮助全网实现云养娃。

摘要：产业互联网和消费互联网看起来都是叫互联网，实则截然不同的底层逻辑，本演讲主要从一个产业互联网数据科学家的角度来看数据科学的变化和一些不变的最佳实践，剖析如何通过数据科学推动企业数字化革命。

互联网业务中的因果推断应用

熊熹 （京东） ***10:00-10:30*** 线上

简介：熊熹，京东数据科学家，2015 年加入京东，目前就职于在京东北美研究院；一直致力于机器学习算法和数据科学在京东搜索与推荐业务中的应用。曾在国内外知名大公司和研究机构从事复杂实验设计的理论和实践工作，并持续跟踪大规模线上实验与数据科学在互联网应用的前沿研究。

摘要：因果推断是数据科学一大重要应用场景；在互联网领域，因果推断在实验设计，策略评估和归因和长期目标指定等领域均具有不可或缺的作用。本报告主要归纳了目前比较主流的应用场景案例，原理及使用注意，也从“实用性”和“故事性”两个角度提出自己的思考。

自动驾驶从零到一

肖一凡 （小马智行） ***10:40-11:10*** 线上

简介：肖一凡，小马智行科技有限公司，自动驾驶算法工程师。硕士毕业于清华大学计算机系，毕业后加入华为技术有限公司。曾任华为手机产品线语音助手交付负责人、华为北京研究院 AutoML 高级算法工程师。2021 年加入小马智行，开始从事自动驾驶研发工作。研究方向主要集中在自然语言处理、AutoML 算法、感知算法等，同时对算法的工程交付、线上运营有丰富经验。

摘要：自动驾驶行业正在经历快速发展和变革，无论是传统车企、传统互联网公司，还是造车新势力、创业公司等都在从不同的角度加入这个行业的竞争。那么当前业界有哪些主流的技术方案？一个公司如何从零到一地开展自动驾驶研发工作？一个新的特性又是如何从零到一地加入到整个自动驾驶系统中？本报告会围绕这些问题展开讨论，欢迎大家参会一同探讨。

基于手机传感器数据的危险驾驶行为识别

张源源 （百姓车联） ***11:10-11:40*** **线上**

简介：张源源，负责百姓车联数据科学与平台团队，致力于用人工智能技术帮助司机安全驾驶，提升车险、车后行业效率。在一年内，带领团队研发了行业一流的危险驾驶行为识别系统，并通过持续的权益运营，显著降低司机危险驾驶行为次数，成为保险行业首个入选中国人民银行金融科技创新监管工具的应用，并获得第 5 届星斗奖数据驱动领军企业奖、领军个人奖，带领团队申请了多项相关专利，在领域内顶会 UbiComp 发表 paper 2 篇。曾负责阿里体育的数据算法团队，致力于用计算机视觉、传感器技术帮助体育运动数字化。在阿里期间开发的 AI 运动是业界第一个在手机端实时进行健身动作计数的应用，在疫情期间帮助 30 多所高校近 20 万名师生顺利开展体育课教学，从 0 到 1 为阿里体育开辟了新的业务方向，目前已经是阿里体育主要业务方向。

摘要：近几年，随着社会治安的逐渐变好，刑事案件越来越少，交通事故带来的不安全大有成为影响社会安全的头号危险因素的趋势。除了通过事后的交通安全处罚、交通事故影响车险定价、日常交通安全教育等等需要较高社会成本的方式去教育、引导大家安全驾驶，如果能对用户自己的行程进行打分乃至实时反馈开车好坏，并通过权益引导等方式激励大家持续安全驾驶，无疑对社会是很有意义的。对用户危险驾驶行为识别大致有三种技术方案，前装方案、后装方案、基于手机 APP 方案；其中前装方案要求车辆本身已经有丰富的传感器，但这种车辆市占率目前还是比较低的，后装方案需要在车辆上安装数百到上千元的硬件设备，有不小的硬件成本和安装成本，前装和后装方案的局限性极大了影响了技术方案的普及度。我们最终选了手机 APP 的技术方案，本次报告将主要带大家了解介绍这一工作，也分享一些手机传感器数据的处理经验。

会议主办方：

中国人民大学 RENMIN UNIVERSITY OF CHINA | 统计学院 SCHOOL OF STATISTICS

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

CAPITAL OF STATISTICS PROFESSION, HUMANITY & INTEGRITY

会议承办方：

中国人民大学数据科学与大数据统计系

赞助商：

RStudio