中国R会
The China-R Conference

会议手册

2022

第15届中国R会（北京）

The 15th 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：2022.11.19-2022.11.25

欢迎辞

经过十五年的磨砺，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

2022 年 11 月 23 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

Posit

Posit 公司，原名 RStudio，成立于 2008 年，创始人为 JJ Allaire，R 社区领军人物 Hadley Wickham 为首席科学家。该公司旨在为 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。Posit 公司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，Posit 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，该公司倡议与微软、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。2022 年 7 月 RStudio 公司更名为 Posit，公司章程将使命定义为创建用于数据科学、科学研究和技术交流的免费开源软件，这一使命意在超越”R for Data Science” ——沿用 R 语言成功的方法，并将其应用得更广泛。

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：目前已开展到第十五届，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 R 会官网上找到：`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》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。

第十五届中国 R 会筹备委员会

主　席：宋文轩
副主席：李诗瑶　郝嘉欣
秘书团：戴诗桐　闫涵超　王祎帆　任　焱　向　悦　孔令仁

分会场主席：

- 主会场：委员会联合筹办
- 人大专场 统计推断理论与应用：吕晓玲
- 机器学习：常象宇
- 区块链保险：苏锦华
- 人大专场 统计计算与深度学习：吕晓玲
- 统计软件分会场：谢益辉 黄湘云
- 统计计算：邱怡轩
- 社交媒体：王小宁
- 可视化：陈思明
- 生态环境：罗立辉
- 工业大数据：田春华
- 生物统计：边蓓蕾 王健桥
- 医疗卫生与健康：李璇

日程表

主会场（11.19 上午）

腾讯会议号：696226349；明德主楼 1031

会议链接：https://meeting.tencent.com/dm/UcyLhAME9aTD

本会场邀请到学界知名人士介绍统计学、数据科学与人工智能的融合发展，展示当前理论和应用领域引起关注的核心问题，希望能够带给听众多元的思维碰撞。

统计推断理论与应用（11.19 下午）

腾讯会议号：581200034；明德主楼 1031

会议链接：https://meeting.tencent.com/dm/g6FZSd4HQNhM

本会场五位报告人均为中国人民大学统计学院青年教师及博士生。报告主题涉及统计推断理论及其在各领域的应用。包括贝叶斯时空点过程模型、广义可加部分线性模型理论，多模态脑网络数据、纵向图像数据研究，以及图数据的条件相依性学习等。

机器学习（11.20 上午）

腾讯会议号：191863973

会议链接：https://meeting.tencent.com/dm/aynfqluCaId2

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，包括深度神经网络、马尔可夫决策过程、迁移学习、网络数据降维等主题。

区块链保险（11.20 上午）

腾讯会议号：519896259；明德主楼 1037

会议链接：https://meeting.tencent.com/dm/awPk0yCak9lN

区块链保险分会场围绕保险科技与前沿领域，分享去中心化保险、区块链保险、精准医疗与保险等话题，覆盖科普分享、论文汇报、行业产品横评、研究汇报等形式。

统计计算与深度学习（11.20 下午）

腾讯会议号：403996694；明德主楼 1037

会议链接：https://meeting.tencent.com/dm/sahiI1GOKmHL

本会场五位报告人均为中国人民大学统计学院青年教师及博士生。报告主题涉及统计计算与深度学习各领域的理论与应用。包括客户流失预测模型的序贯一步估计算法，异质数据个性优化的联邦学习方法、平均治疗效果的半参估计方法、空间自回归模型以及轨迹数据挖掘的深度学习模型等。

软件工具-1（11.20 晚）

腾讯会议号：674557544

会议链接：https://meeting.tencent.com/dm/4Edq5qJRMwIc

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工具的支持。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合作开发的工具、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，有效帮助到广大研究者们的研究工作。

统计计算（11.21 晚）

腾讯会议号：158759599

会议链接：https://meeting.tencent.com/dm/MU8gJ4GxcSn4

本会场专注于统计计算与科学计算问题，力图从硬件、软件、语言框架、算法和应用等各个层次探讨当前数据科学中常见的计算问题，预期向观众介绍高效利用计算资源的方法和技巧，以应对大型数据和复杂模型带来的挑战。

社交媒体（11.22 晚）

腾讯会议号：522177187

会议链接：https://meeting.tencent.com/dm/ORj8aHWGg8MZ

社交媒体专场聚焦社交媒体相关数据的挖掘，主要分享计算社会科学、图学习、政治参与、中文意见表达等主题的内容。

可视化（11.22 晚）

腾讯会议号：520839149

会议链接：https://meeting.tencent.com/dm/bo6rCgSundiR

可视化与可视分析是研究信息的视觉呈现、交互方法和技术的学科，有效地结合了人类认知能力与机器强大的计算能力，是分析理解复杂数据的有效手段。本会场邀请到六位可视化领域研究人员，报告主题涉及社交媒体文本可视分析、历史学群体可视分析研究、数据整理脚本的语义理解、可视化的智能自然语言交互以及基于交互式机器学习的数据重建与数据标注等。

软件工具-2（11.23 晚）

腾讯会议号：382794995

会议链接：https://meeting.tencent.com/dm/l2jC3WImqx95

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工具的支持。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合作开发的工具、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，有效帮助到广大研究者们的研究工作。

生态环境（11.23 晚）

腾讯会议号：579452861

会议链接：https://meeting.tencent.com/dm/Z2YTEWkK0vDn

在全球变暖背景下，冰冻圈的快速变化对全球和区域气候、水文、生态和经济社会发展产生了重要影响，本会场聚焦于植被物候、地质灾害、水资源、环境污染等的变化及其对气候的响应。

工业大数据（11.24 晚）

腾讯会议号：343593261

会议链接：https://meeting.tencent.com/dm/7964IGP6IzYH

工业是国民经济的核心要素和国家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，收到了学界和业界的普遍关注。本专题会场将结合生产调度、设备运营、绩效评估、运作效率优化等主题，探讨工业背景下数据分析的问题、挑战、算法技术等。

生物统计（11.24 晚）

腾讯会议号：372398290

会议链接：https://meeting.tencent.com/dm/fTEYn0CJfVix

近年来，大健康产业迅速崛起，其中基因研究得到广泛关注。本会场邀请到学界海内外青年学者共同探讨统计遗传学，包括统计方法及软件流程的开发及应用，还将介绍这些新工具带来的新的科学发现。

软件工具-3（11.25 晚）

腾讯会议号：799263224

会议链接：https://meeting.tencent.com/dm/t3BrN8gJcljI

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工具的支持。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合作开发的工具、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，有效帮助到广大研究者们的研究工作。

医疗卫生与健康（11.25 晚）

腾讯会议号：273990838

会议链接：https://meeting.tencent.com/dm/w5zkULTKi5o5

统计与数据科学的方法在医疗卫生与健康领域的应用日益成熟，成效显著。本专题会场从不同的角度探讨医疗健康与数据科学的相关议题，报告内容涵盖人工智能技术在医疗器械及制药领域的相关应用、真实世界的临床研究与产业实践，环境健康与人体健康的关联例证，贝叶斯 Lasso 并行多重中介模型，针对特定关联性研究的模型比较等，最后还会介绍一个药物研发的开源整合框架，希望借此机会增进医疗健康领域内不同方向的沟通与交流。

会议摘要

主会场（11.19 上午）

用统计学方法破解自然语言密码

邓柯 （清华大学统计学研究中心） ***8:30-9:05***

简介：邓柯，清华大学统计学研究中心长聘副教授。主要从事 Bayes 统计和统计计算方面的研究，并致力推动统计学与生物医学、人工智能、人文社科等领域的前沿交叉。北京大学应用数学学士（2003）、统计学博士（2008），哈佛大学博士后、副研究员（2008-2013）。2013 年进入清华大学工作，2015 年当选中国人工智能学会智慧医疗专业委员会副主任委员，2016 年获“科学中国人年度人物”的荣誉称号，2017 年当选中国现场统计研究会计算统计分会理事长、2018 年当选中国青年统计学家协会副会长，2019 年受聘担任北京“智源人工智能研究院”研究员，2020 年荣获“世界华人数学家国际联盟”最佳论文奖和“中国数字人文大会”最佳论文奖。主持多项国家级项目，担任多个国内外知名学术期刊的副主编和编委。

摘要：理解自然语言是当代人工智能研究的一个关键问题。近年来，基于深度学习，特别是大规模预训练模型，的自然语言处理方法取得了重要进展，在机器翻译、文本生成等预测性任务上取得了良好的表现。但是，目前来看，这种完全依赖大数据和大算力堆积起来的“黑盒模型”还不能真正理解自然语言并对语言现象背后的本质机理进行揭示。我们另辟蹊径，从统计学习的角度出发，提出了一系列贝叶斯统计模型对自然语言的局部结构进行建模和解析。大量实际数据分析表明，这些方法可以在小数据、弱监督场景下，以可解释的方式有效揭示中文自然语言的局部语法语义结构，并在中文分词、命名实体识别、关系提取等重要任务上取得良好的效果。

Learning an Explicit Sample Weighting Mapping for Robust Deep Learning

孟德宇 （西安交通大学） ***9:05-9:40***

简介：孟德宇，西安交通大学教授，博士生导师，任大数据算法与分析技术国家工程实验室统计与大数据中心副主任。发表论文百余篇，其中 IEEE 汇刊论文 60 余篇，计算机学会 A 类会议 40 篇，谷歌学术引用超过 18000 次。现任 IEEE Trans. PAMI，Science China: Information Sciences 等 7 个国内外期刊编委。目前主要研究聚焦于元学习、概率机器学习、可解释性神经网络等机器学习基础研究问题。

摘要：现有深度学习方法的有效性依赖于对训练数据集的高质量要求，当训练集呈现蕴含复杂标记噪声、类别不均衡等数据偏差问题时，其有效性往往不能得以保证，这被称之为深度学习的鲁棒性学习问题。这一问题已经严重制约了深度学习在现实场景中的有效应用，是领域亟需面对的瓶颈问题。本报告将特别针对样本加权这一类典型的处理数据偏差的方法论展开讨论，介绍该方法论从针对少量数据偏差类型的传统手工赋权设定方法，如何演进到更为前沿的针对更多数据偏差类型的自动化赋权方法。特别地，将深入讨论在该方法

论背后蕴含的元学习思想本质，挖掘其有效性理论内涵，从而揭示其可能对现实场景中复杂鲁棒深度学习问题的潜在泛化可用性。

谱域图神经网络理论基础

魏哲巍（中国人民大学高瓴人工智能学院） ***9:55-10:30***

简介：魏哲巍，教授，博导，入选国家高层次青年人才计划。研究方向为大数据算法理论、图机器学习。2008 年本科毕业于北京大学数学科学学院，2012 年博士毕业于香港科技大学计算机系；2012 年至 2014 年于奥胡斯大学海量数据算法研究中心担任博士后研究员，2014 年 9 月加入中国人民大学信息学院担任副教授，2019 年 8 月起任教授。2020 年 4 月加入高瓴人工智能学院。在数据库、理论计算机、数据挖掘、机器学习等领域的顶级会议及期刊上（如 SIGMOD、VLDB、ICML、NeurIPS、KDD、SODA 等）发表论文 50 余篇，并获得数据库理论顶级会议 PODS2022 十年最佳论文奖。担任 PODS、ICDT 等大数据理论会议论文集主席以及 VLDB、KDD、ICDE、ICML、NeurIPS 等国际会议程序委员会委员。主持多项自然科学基金青年项目、面上项目及重点项目子课题。担任鹏城实验室广州基地青年科学家。培养博士生获 2021 年百度奖学金（全球 10 人）。

摘要：近年来，由于图结构数据的强大表达能力，用机器学习方法分析和挖掘图数据的研究越来越受到重视。图神经网络（Graph Neural Networks）是一类基于深度学习的处理图数据的方法，在众多领域展现出了卓越的性能，其已经成为一种广泛应用的图分析方法。谱域图神经网络是图神经网络研究中一类重要的方法，它们在拉普拉斯谱域中设计和学习不同的图卷积，具有良好的理论保证和可解释性。本报告拟先介绍图神经网络的任务和一些前沿应用，然后从图傅里叶变换、图卷积的设计和图谱滤波器的多项式近似等方面探讨谱域图神经网络的理论基础，最后将讨论我们在谱域图神经网络所做的一些工作和对未来工作的展望。

Statistical Learning Methods for Neuroimaging Data Analysis with Applications

朱宏图（北卡罗来纳大学教堂山分校） ***10:30-11:05***

简介：朱宏图博士是北卡罗来纳大学教堂山分校生物统计学，计算机，和遗传学终身教授，曾任 MD 安德森癌症中心的诊断影像学 Bao-Shan Jing 讲席教授和生物统计学终身教授，滴滴出行首席统计学家。2000 年获得香港中文大学统计学博士学位。主要研究领域为统计学习、医疗图像处理、精准医疗、生物统计、人工智能和大数据分析。2011 年当选美国统计学会和数理统计学会会士。2016 年荣获德克萨斯州癌症预防与研究中心杰出研究奖。2019 年因强化学习在网约车出行中的应用荣获 Daniel Wagner 杰出应用奖。在多个大型医疗研究项目中担任统计分析师，并提供实验设计、数据分析和新方法开发。现有高水平期刊论文 290 多篇，包括 Nature，Science，Cell，Nature Genetics，Nature Communication，Nature Neuroscience，JAMA Psychiatry，PNAS，JMLR，AOS 以及 JRSSB；高水平会议论文 45 篇，包括 KDD，NIPS，ICDM，AAAI，MICCAI 以及 IPMI。担任多个国际顶级会议的区域主席，包括 Information Processing in Medical Imaging。担任（过）多个国际顶级期刊的编委，包括 Statistica Sinica，JRSSB，Biometrics，Annals of Statistics 和 Journal of American Statistical Association。

摘要： The aim of this talk is to provide a comprehensive review of statistical challenges in neuroimaging data analysis from neuroimaging techniques to large-scale neuroimaging studies to statistical learning methods. We briefly review eight popular neuroimaging techniques and their po-

tential applications in neuroscience research and clinical translation. We delineate the four common themes of neuroimaging data and review major image processing analysis methods for processing neuroimaging data at the individual level. We briefly review four large-scale neuroimaging-related studies and a consortium on imaging genomics and discuss four common themes of neuroimaging data analysis at the population level. We review nine major population-based statistical analysis methods and their associated statistical challenges and present recent progress in statistical methodology to address these challenges.

Optimal Integrating Learning for SQD-type Data

李扬 （中国人民大学统计学院） ***11:05-11:40***

简介：李扬，中国人民大学统计学院教授、博士生导师，统计咨询研究中心主任；国际统计学会推选会员、中国现场统计研究会常务理事、中国商业统计学会副会长、北京生物医学统计与数据管理研究会监事长；主要从事相关型数据分析、模型选择与不确定性评价、潜变量建模、临床试验设计等领域研究，承担国家自然科学基金面上项目、全国统计科学研究重大项目等科研项目二十余项，发表国内外期刊研究论文七十余篇。

摘要： In the era of data science, it is common to encounter data with different subsets of variables obtained for different cases. An example is the split questionnaire design (SQD), which is adopted to reduce respondent fatigue and improve response rates by assigning different subsets of the questionnaire to different sampled respondents. A general question then is how to estimate the regression function based on such block-wise observed data. Currently, this is often carried out with the aid of missing data methods, which may unfortunately suffer from intensive computational cost, high variability, and possible large modeling biases in real applications. In this article, we develop a novel approach for estimating the regression function for SQD-type data. We first construct a list of candidate models using available data-blocks separately, and then combine the estimates properly to make an efficient use of all the information. We show the resulting averaged model is asymptotically optimal in the sense that the squared loss and risk are asymptotically equivalent to those of the best but infeasible averaged estimator. Both simulated examples and an application to the SQD dataset from the European Social Survey show the promise of the proposed method.

统计推断理论与应用（11.19 下午）

Statistical Inference for Mean Function of Longitudinal Imaging Data over Complicated Domains

李杰 （中国人民大学统计学院） ***14:00-14:30***

简介： 李杰，中国人民大学统计学院师资博士后。2022 年毕业于清华大学，获得统计学博士学位。主要研究方向为函数型数据分析、时间序列和非参数统计。曾获国际统计学会 2021 年简·丁伯根奖一等奖，国际数理统计协会 2020 年 Hannan Graduate Student Travel Award，并在 Statistica Sinica 等期刊发表论文多篇。

摘要： Motivated by longitudinal imaging data possessing inherent spatial and temporal correlation, we propose a novel procedure to estimate its mean function. Functional moving average is applied to depict the dependence among temporally ordered images and flexible bivariate splines over triangulations are utilized to handle the irregular domain of images which is common in imaging studies. Both global and local asymptotic properties of the bivariate spline estimator for mean function are established with simultaneous confidence corridors (SCCs) as a theoretical byproduct. Under some mild conditions, the proposed estimator and its accompanying SCCs are shown to be consistent and oracle efficient as if all images were entirely observed without errors. The finite sample performance of the proposed method through Monte Carlo simulation experiments strongly corroborates the asymptotic theory. The proposed method is further illustrated by analyzing two sea water potential temperature data sets.

Generalized Bayesian Spatio-Temporal Point Process Model and Its Application

周峰 （中国人民大学统计学院） ***14:30-15:00***

简介： 周峰，中国人民大学统计学院讲师，中国人民大学杰出青年学者。主持国家自然科学基金青年项目，中国博士后基金特别资助、面上资助，入选博士后国际交流计划引进项目。主要研究方向包括统计机器学习、贝叶斯方法、随机过程、神经脉冲序列等。主要研究论文发表于 Journal of Machine Learning Research, Statistics and Computing, International Conference on Learning Representations (ICLR), Conference on Neural Information Processing Systems (NeurIPS) 等期刊、会议上。

摘要： The spatio-temporal point process is a common stochastic process model which is used to model the pattern of events occurring in time or space. Its application covers a wide range of domains including seismology, epidemics, neuroscience and high-frequency financial engineering. The traditional spatio-temporal point process model has limitations on flexibility, time-variability, multi-taskability, uncertainty and efficiency. To relieve the aforementioned limitations, in the first part of our work we propose the flexible time-varying nonlinear Hawkes process to extend the traditional Hawkes process in terms of both flexibility and time-variability; in the second part of our work we propose the heterogeneous multi-task nonparametric Cox process to extend the traditional nonhomogeneous Poisson process in terms of both flexibility and multi-taskability. In the meantime, for each model, we convert the non-conjugate problem to a conditional conjugate one by using the data augmentation technique, so as to derive efficient inference algorithms with analytical expressions. This work lays a solid foundation for the application of Bayesian spatio-temporal point processes in the big data scenario.

Blind source separation for multimodal brain networks

吴奔 （中国人民大学统计学院） ***15:00-15:30***

简介：吴奔，中国人民大学统计学院讲师，曾经在 Emory 大学生物统计与生物信息系、Michigan 大学生物统计系从事博士后研究工作。主要研究兴趣为贝叶斯统计、独立成分分析、神经影像数据分析等。在 JASA、Biometrics、中国科学（数学）、统计研究、系统工程理论与实践等期刊上发表过论文，正在加油尝试更多的期刊来延长个人简介。

摘要： There is a strong interest in analyzing multimodal brain networks in recent years. Integrating information from multimodal connections can potentially help better understand the formation and alteration in brain connectors due to neurodevelopment and disease progression. Investigating the interplay among multimodal brain networks is challenging due to several reasons such as the high noise of the imaging data, the different measures of connectivity across modalities, etc. In this talk, we will introduce a new blind source separation method that can be applied to decompose discrete representations of brain networks and achieve joint analysis of multimodal connections. We demonstrate our method with comprehensive simulations and present our findings on functional and structural brain connectivity from a real data study.

Multifold Cross-Validation Model Averaging for Generalized Additive Partial Linear Models

陈泽 （中国人民大学统计学院） ***15:45-16:15***

简介：陈泽，中国人民大学统计学院在读博士生，主要研究方向为变量重要性，模型平均等。

摘要： Generalized additive partial linear models (GAPLMs) are appealing for model interpretation and prediction. However, for GAPLMs, the covariates and the degree of smoothing in the nonparametric parts are often difficult to determine in practice. To address this model selection uncertainty issue, we develop a computationally feasible model averaging (MA) procedure.The model weights are data-driven and selected based on multifold cross-validation (CV) (instead of leave-one-out) for computational saving. When all the candidate models are misspecified, we show that the proposed MA estimator for GAPLMs is asymptotically optimal in the sense of achieving the lowest possible Kullback-Leibler loss. In the other scenario where the candidate model set contains at least one quasi-correct model, the weights chosen by the multifold CV are asymptotically concentrated on the quasi-correct models. As a by-product, we propose a variable importance measure to quantify the importances of the predictors in GAPLMs based on the MA weights. It is shown to be able to asymptotically identify the variables in the true model.Moreover, when the number of candidate models is very large, a model screening method is provided. Numerical experiments show the superiority of the proposed MA method over some existing model averaging and selection methods.

Learning conditional dependence graph for concepts via matrix normal graphical model

赖基正 （中国人民大学统计学院） ***16:15-16:45***

简介：赖基正，中国人民大学统计学院在读博士生，主要研究方向为文本挖掘，概念图模型等。

摘要： Conditional dependence relationships for random vectors is extensively studied and broadly applied. But it is not very clear how to construct the dependence graph for unstructured data like concept words or phrases in text corpus, where the variables(concepts) are not jointly observed with i.i.d. assumption. We assume that all the concept vectors learned from GloVe jointly follow a matrix normal distribution with sparse precision matrices. Different from knowledge graph methods, the conditional dependence graph describes the conditional dependence structure between concepts given all other concepts, which means that the concepts(nodes) linked by edges cannot be separated by other concepts. It represents an essential semantic relationship. A penalized matrix normal graphical model(MNGM) is then employed to learn the conditional dependence graph for both the concepts and the embedding 'dimensions'. Since the concept words are nodes in our graph with huge dimensions, we employ the MDMC optimization method to speed up the glasso algorithm. On the other hand, we propose a sentence granularity bootstrap to get 'independent' repeats of samples to enhance the penalized MNGM algorithm. We name the proposed method as Matrix-GloVe. In simulation studies, we check that the graph learned by Matrix-GloVe is more suitable for Graph Convolutional Networks(GCN) than a correlation graph. We employ the proposed method in two scenarios from real data and get good results.

机器学习（11.20 上午）

深度神经网络的学习理论

林绍波 （西安交通大学管理学院） ***8:30-9:05***

简介： 西安交通大学管理学院，教授、博士生导师。研究方向为函数逼近论、分布式学习理论、深度学习理论及强化学习理论。在应用数学顶级期刊 ACHA、SINUM、CA 及机器学习顶级期刊 JMLR，TPAMI，TIT 等发表论文 70 余篇。主持或以核心骨干参与国家级课题 11 项。

摘要： 深度学习在诸如图像处理、自然语言处理、运筹、博弈等领域取得了巨大的成功。但其成功的原因依然缺乏严格的理论解释与验证。在这种未知性下，学术界与业界掀起了深度学习浪潮，试图用深度神经网络去处理所有学习问题。很显然，在某些应用上，效果不如预期。该报告将从数学上（统计学习的角度）揭露深度神经网络的学习能力并在一定程度阐明深度学习的适用范围。特别地，该报告聚焦如下四个基本问题：1. 深度网是否一定比单层网好？2. 在什么情况下用深度学习会更有效？3. 为什么深度网在大数据时代取得这么大成功？4. 过参数化深度神经网络为何可规避过拟合现象？

Statistical Properties of Robust Markov Decision Processes

杨文昊 （北京大学前沿交叉学科研究院） ***9:05-9:40***

简介： 杨文昊，北京大学前沿交叉学科研究院数据科学（统计学）专业的博士研究生。其于 2018 年获得北京大学统计学学士学位。主要的研究兴趣包括统计学习和机器学习理论，目前集中在强化学习的理论研究上。其研究成果发表在 NeurIPS, ICLR, AISTATS 等国际会议和 Annals of Statistics 国际杂志上。

摘要： Robust MDPs are proposed to handle the sensitive estimation errors in value estimation of MDPs, where the transition probability is allowed to take values in an uncertainty set. In recent years, many works have proposed computationally efficient learning algorithms to solve robust MDPs and obtained the near-optimal robust policy and value function. However, the statistical performances of the optimal robust policy and value function are less studied. In this talk, we will introduce the basic theories and algorithms of robust MDPs and figure out two questions: (a) How many samples are sufficient to guarantee the accuracy of the robust estimators; (b) whether it is possible to make statistical inferences from the robust estimators. We will answer these questions from both non-asymptotic and asymptotic viewpoints.

Transferred Q-learning

李赛 （中国人民大学统计与大数据研究院） ***9:45-10:20***

简介： 李赛，中国人民大学统计与大数据研究院准聘副教授，博士生导师。2018 年毕业于罗格斯新泽西州立大学，获得统计博士学位，后于宾夕法尼亚大学生物统计系和统计系进行博士后研究，目前的研究方向包括高维数据分析、迁移学习、因果推断的统计方法及理论和在遗传学、流行病学和机器学习中的应用。

摘要： We consider Q-learning with knowledge transfer, using samples from a target reinforcement learning (RL) task as well as source samples from different but related RL tasks. We propose

transfer learning algorithms for both batch and online Q-learning with offline source studies. The proposed transferred Q-learning algorithm contains a novel re-targeting step that enables vertical information-cascading along multiple steps in an RL task, besides the usual horizontal information-gathering as transfer learning (TL) for supervised learning. We establish the first theoretical justifications of TL in RL tasks by showing a faster rate of convergence of the Q-function estimation in the offline RL transfer, and a lower regret bound in the offline-to-online RL transfer under certain similarity assumptions. Empirical evidences from both synthetic and real datasets are presented to back up the proposed algorithm and our theoretical results.

Dimension reduction for covariates in network data

赵俊龙 （北京师范大学统计学院） ***10:20-10:55***

简介：赵俊龙，北京师范大学统计学院教授。主要从事统计学和机器学习相关研究，包括：高维数据分析、统计机器学习、稳健统计等。在统计学各类期刊发表 SCI 论文四十余篇，部分结果发表在统计学国际顶级期刊 JRSSB，AOS、JASA，Biometrika 等。主持多项国家自然科学基金项目，参与国家自然科学基金重点项目。任中国现场统计学会高维数据分会理事，北京应用统计学会理事、北京大数据学会常务理事等。

摘要： A problem of major interest in network data analysis is to explain the strength of connections using context information. To achieve this, we introduce a novel approach, called network supervised dimension reduction, in which covariates are projected onto low-dimensional spaces to reveal the linkage pattern without assuming a model. We propose a new loss function for estimating the parameters in the resulting linear projection, based on the notion that closer proximity in the low-dimension projection corresponds to stronger connections. Interestingly, the convergence rate of our estimator is found to depend on a network effect factor, which is the smallest number that can partition a graph in a manner similar to the graph colouring problem. Our method has interesting connections to principal component analysis and linear discriminant analysis, which we exploit for clustering and community detection. The proposed approach is further illustrated by numerical experiments and analysis of a pulsar candidates dataset from astronomy.

区块链保险（11.20 上午）

区块链 + 保险科普

王叶 （香港中文大学信息技术管理硕士） ***9:00-9:30***

简介：王叶，香港中文大学信息技术管理硕士，在读期间曾参与若干区块链创业项目，曾任区块链基金的投资研究员，发布多篇区块链基础设施相关研报。

摘要： 2008 年，区块链技术诞生，并改善了转账体系。2015 年，以太坊与智能合约的诞生，从此，代码取代了各种金融中介的托管职能与信任成本，区块链 + 金融开始繁荣。当前，区块链与银行、证券、衍生品、保险的结合应用都有了较好的发展

Optimal Risk Pooling of Peer-to-Peer Insurance

陈泽 （中国人民大学财政金融学院） ***9:30-10:00***

简介： 陈泽，中国人民大学财政金融学院保险系助理教授，中国保险研究所研究员，中国人民大学“杰出青年”学者 B 岗。博士毕业于比利时鲁汶大学（KU Leuven）和清华大学。研究方向为保险精算与风险管理、数字经济、未来保险科技与元宇宙等话题。他目前的研究从经济学理论上关注了去国内外新兴的去中心化保险形式，并曾联合美国伊利诺伊大学（UICU）共同发布过研究报告白皮书，共同撰写过多篇关于去中心化理论下的保险理论及风险分担的研究，并被多个国际重要会议接收和报告。目前，他在国内外保险精算和金融学术期刊发表论文 10 余篇，如 Insurance：Mathematics and Economics, Scandinavian Actuarial Journal，European Financial Management，Methodology and Computing in Applied Probability 等；并主持和多项横纵向科研课题，如自然科学基金青年基金，教育部哲学社科基金一般项目以及北美精算师协会委托课题等。

摘要： Peer-to-peer insurance models, that jointly incorporate the forms of centralized insurer’s underwriting and decentralized peers’ risk sharing, are emerging. Un?der these innovative risk sharing forms, the risk is separated into two layers: the first below-deductible part is shared within a community, and the second above?deductible loss, exceeding the community’s risk-bearing capacity, is covered by an insurer. In this paper, we mathematically formalize two existing peer-to-peer in?surance models: the individual- and group-covered models. From the perspective of risk-averse participants, we investigate the existence, closed-form expression, and properties of optimal deductible, the primary feature of peer-to-peer insurance.

Web3.0 下的保险可能是什么样?

祁晨瑞 （中国人民大学统计学院） ***10:00-10:30***

简介：祁晨瑞，本科就读于中央财经大学保险学院，研究生现就读于中国人民大学统计学院。曾主持北京市级大学生创新创业训练比赛《我国巨灾保险模式研究》，负责人大 85 周年校庆数字藏品头像的智能合约开发，兼任司马数慧算法开发，量化比赛 worldquant 中国区第三名。

摘要：保险的初衷是进行风险的共担。但如今，保险业主要以大型保险公司为主导，在这样较为传统的保险业务的流程下，存在着高成本、低效率的问题，同时很多特殊风险难以

承保。区块链技术的应用，如：智能合约、Oracle 技术、零知识证明等，为传统保险业务流程中存有的委托代理问题、信息不对称问题提供了一种解决思路。合理地运用区块链技术，将削减大量人力流转所带来的成本，同时也拓宽了可承保的风险范围。近年来，国外涌现出了一批基于区块链的新兴保险平台，如：Nexus Mutual、Etherisc、Bridge Mutual，它们各自有着自己的生态体系与模式设计。通过分析它们之间的共性与差异，对其模式设计、流程、生态进行横向比较，我们认为保险科技不是与传统保险正面对立与竞争，而是对传统保险的补充和升级，开辟了新的保险需求或满足了传统机构并未覆盖到的风险需求。

长护险的精准定价

夏贺彦 （中国人民大学统计学院） ***10:30-11:00***

简介：夏贺彦，中国人民大学统计学院风险管理与精算系研究生，中国人民大学精算学会会长。本科就读于南开大学金融学院保险学系。研究方向为老年健康、长期护理保险的精算定价、风险管理与保险科技等。研究报告《商业医疗险的通胀识别及主因分析》获“寿再青骏杯”第二名，中国人寿再保险“青骏计划”实习生夏令营营员。

摘要：世界各国出生率下降、老龄化加剧已经成为全球性的问题，随之而来的是将至少数十年的养老成本和医疗成本等社会问题，因此开展老年健康险的研究具有重要的现实意义。尽管不少学者测算了未来的家庭和社会护理成本，长护险存在着很大的需求体量，但目前长期护理保险的定价和运行都存在着不同的问题。我们对这些问题进行介绍，在长护定价方面，由于对多状态之间的转移仅采用 Markov 过程描述并不符合现实，CHARLS 和 CLHLS 等微观调查数据库的现有数据也存在着严重的删失问题，这会导致对转移概率的测算存在偏差，风险异质性的存在也使得精准定价难以进行。运行方面，商业长护险由于较高的赔付率难以为继，很难扩大体量。针对以上两方面问题，本文将介绍生存分析技术在长护险定价模型中的应用，及目前商业公司在长护险运行中可采取的保险形式，和对个性化数据的采取及应用。并针对智能穿戴设备在精准医疗上的发展，以及区块链技术对病历联网协同的推进，做探讨与展望。

统计计算与深度学习（11.20 下午）

Nonparametric inference about mean functionals of nonignorable nonresponse data without identifying the joint distribution

李伟 （中国人民大学统计学院） *14:00-14:30*

简介：李伟，中国人民大学统计学院副教授。主要研究领域为因果推断、缺失数据、高维统计等。目前已在包括 Biometrika, Journal of Econometrics, Biometrics 等期刊上发表多篇学术论文，主持国家自然科学青年基金项目和全国统计科学研究重点项目各一项

摘要： We consider identification and inference about mean functionals of observed covariates and an outcome variable subject to nonignorable missingness. By leveraging a shadow variable, we establish a necessary and sufficient condition for identification of the mean functional even if the full data distribution is not identified. We further characterize a necessary condition for root n-estimability of the mean functional. This condition naturally strengthens the identifying condition, and it requires the existence of a function as a solution to a representer equation that connects the shadow variable to the mean functional. Solutions to the representer equation may not be unique, which presents substantial challenges for nonparametric estimation and standard theories for nonparametric sieve estimators are not applicable here. We construct a consistent estimator for the solution set and then adapt the theory of extremum estimators to find from the estimated set a consistent estimator for an appropriately chosen solution. The estimator is asymptotically normal, locally efficient and attains the semiparametric efficiency bound under certain regularity conditions. We illustrate the proposed approach via simulations and a real data application on home pricing.

Factor-Assisted Federated Learning for Personalized Optimization with Heterogeneous Data

王菲菲 （中国人民大学统计学院） *14:30-15:00*

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 Econometrics, 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： Federated learning is an emerging distributed machine learning approach, which can simultaneously train a global model from decentralized datasets while preserve data privacy. However, data heterogeneity is one of the core challenges in federated learning. The heterogeneity issue may severely degrade the convergence rate and prediction performance of the model trained in federated learning. To address this issue, we develop a novel personalized federated learning method for heterogeneous data, which is called FedFac. The proposed method is motivated by a common finding that, data in different clients contain both common knowledge and personalized knowledge. Therefore, the two types of knowledge should be decomposed and taken advantages of separately. We introduce the idea of factor analysis to distinguish the client-shared information and client-specific information. With this decomposition, a new objective function is established and optimized. Both theoretical and empirical analysis demonstrate that FedFac has higher com-

putational efficiency against the classical federated learning approaches. The superior prediction performance of FedFac is also verified empirically by comparison with various state-of-the-art federated learning methods on several real datasets.

Grouped spatial autoregressive model

胡威 （中国人民大学统计学院） ***15:15-15:45***

简介：胡威，中国人民大学在读博士生，研究兴趣为网络数据分析、网络数据采样方法、空间自回归模型、超高维数据分析等，研究论文发表于 Computational Statistics & Data Analysis、Electronic Journal of Statistics 等期刊上。

摘要： With the development of the internet, network data with replications can be collected at different time points. The spatial autoregressive panel (SARP) model is a useful tool for analyzing such network data. However, in the traditional SARP model, all individuals are assumed to be homogeneous in their network autocorrelation coefficients, while in practice, correlations could differ for the nodes in different groups. Here, a grouped spatial autoregressive (GSAR) model based on the SARP model is proposed to permit network autocorrelation heterogeneity among individuals, while analyzing network data with independent replications across different time points and strong spatial effects. Each individual in the network belongs to a latent specific group, which is characterized by a set of parameters. Two estimation methods are studied: two-step naive least-squares estimator, and two-step conditional least-squares estimator. Furthermore, their corresponding asymptotic properties and technical conditions are investigated. To demonstrate the performance of the proposed GSAR model and its corresponding estimation methods, numerical analysis was performed on simulated and real data.

Trajectory Representation Learning with Multilevel Attention for Driver Identification

李梦媛 （中国人民大学统计学院） ***15:45-16:15***

简介：李梦媛，中国人民大学统计学院在读博士生，主要研究方向轨迹数据挖据等。

摘要： Massive trajectory data have originated from the development of positioning technology. Learning GPS trajectory representation to characterize a driver's driving style is a challenging task with important applications in many areas, including autonomous driving, auto insurance, advanced driver assistance systems, urban computing, and the internet of things. Few studies have considered the interactions between different factors. In this study, we propose a novel trajectory representation method based on a multilevel attention mechanism (ATTraj2vec) and apply it to the task of driver identification. In addition to summarizing motion features from GPS trajectory data, we also extract spatial and temporal features. We use a multilevel attention mechanism to aggregate the interactions of motion features with temporal and spatial features progressively. Additionally, we adopt multi-loss to optimize our model simultaneously, which consists of a softmax loss for driver classification and Siamese loss for making trajectories from the same driver more similar. Classification experimental results on a real-world automobile trajectory dataset demonstrated that our proposed model significantly outperforms existing baselines. Meanwhile, the proposed method provides significant gains in the trajectory clustering of unseen drivers.

软件工具（11.20 晚上）

Switching from RStudio to VS Code

任坤 （明汯投资） ***19:00—19:30***

简介：任坤，就职于明汯投资，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： In this talk, I will introduce my motivation and experience of switching from RStudio to VS Code and how I implement new features for the R extension and R language server as I find potential improvements. The R development experience in VS Code has been vastly enhanced in recent two years and VS Code itself is evolving rapidly too. Besides the code editing features powered by the R language server, I will shed more light on the most exciting features such as remote development with SSH/WSL/Container, working with multiple terminals, and live collaboration.

R 语言中常见字符编码问题及其最佳实践

谭显英 （安联保险资产管理有限公司） ***19:40—20:10***

简介：谭显英是 R 语言的技术爱好者，也是 Github 的活跃用户，为很多 R 包做出过贡献，如 DT 和 data.table 等。他本职从事投资管理行业，长年在工作中使用 R 语言分析数据、搭建和部署模型、使用 shiny 和自动化报告（knitr, rmarkdown）展现成果等，对解决 R 语言在生产环境中面临的各种疑难杂症颇有心得。

摘要：如果字符串都用 UTF-8 编码该多好，然而 Windows 系统仍是许多 R 用户的工作环境，也是字符编码问题（“乱码”）的重灾区。本报告将会分析字符编码问题的由来，分享 R 语言本身在解决该问题的进展，并针对常见字符编码问题场景（源代码、数据文件和数据库等）给出解决方案或最佳实践。

与 R 互补的现代编程语言工具

覃文锋 （明汯投资） ***20:20—20:50***

简介：从事量化交易系统开发有关工作，曾参与开发和维护了 R Weekly 等开源项目。

摘要：主要讨论现代编程语言的特性，以及能够与 R 互补的一些常见工具。完成一个任务通常会用到多种工具来共同协作，本演讲会从编程语言原理，现代编程语言特性等角度，讨论异步编程、异构编程、即时编译、工具特性、性能优化等问题。

Using Python in R: An Example of Auto-encoder

俞丽佳 （无） ***21:00-21:30***

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： Many deep learning frameworks are built in Python. In order to improve the user numbers,wrapping the Python package into R package is essential for bioinformatics applications. In this talk, I'll use autoencoder as an example to show how to wrap a simple deep-learning Python package into a R package that can be run on GPU.

基于复杂网络的开源软件生态系统研究–以 R 软件为例

李传权 （江西财经大学统计学院） ***21:40-22:10***

简介：李传权，现就职于江西财经大学统计学院，讲师，硕士生导师，中南大学博士毕业，R 语言爱好者，统计之都十年粉丝。研究领域包括高维统计、网络数据分析等。

摘要： R 软件作为统计领域重要的开源软件，开发历史久远，生态系统较为成熟，对其系统架构和依赖关系进行深入研究，从而对国产开源软件及其生态系统的培养具有指导意义. 基于此，本文从复杂的有向网络角度出发，探讨 R 软件的发展，挖掘 R 软件包依赖关系中的社区，并研究社区的动态演变. 研究表明：R 软件迅速发展，功能多样；R 软件包间的依赖关系服从幂律分布和“小世界”现象；R 软件包的依赖网络中有" 统计模型"，" 高性能计算"，" 数据可视化"，" 网页技术"，" 数据预处理"，" 生物信息" 子社区. 综上，R 软件生态系统，作为一个成功的开源软件案例，其主导因素有：可满足整个数据分析全流程的需求，与时俱进地扩展，吸引了来自世界各地的开发者，并注重长期维护软件包的健康.

科研用开源数据分析平台的搭建与部署——以 xcmsrocker 为例

于淼 （杰克逊实验室） ***22:20-22:50***

简介：于淼，理学博士，杰克逊实验室科学家，研究方向为环境暴露组学，发表论文四十余篇，引用过千，《现代科研指北》作者，统计之都编辑部主编。

摘要：从简单的原始数据共享到完整的数据流程再现，目前科研中对研究结果的可再现性（reproducibility）不断提出更高的要求。影响研究结果可再现性的因素主要是软件的正常使用与模型的标准化构建与评价。前者经常受软件平台影响而后者则主要是缺乏标准化的脚本。以 Docker 为代表的容器化技术可以将软件正常运行所需要的所有依赖、集成开发环境乃至操作系统都打包为一个系统镜像，这样通过系统镜像的分发可以最大程度保障软件的可再现性。而以 Knitr、Jupyter Notebook 等为代表的文学化编程（literate programming）技术则可以很好的将代码运行与工作流（workflow）进行整合，这为再现机器学习模型的构建与评价过程提供了保障。基于容器化技术与文学化编程，我开发维护了一个基于 R 语言的开源数据处理平台项目 xcmsrocker，可用于基于代谢组学机器学习的环境研究。xcmsrocker 本质上是一个基于 Rocker 项目的系统镜像，后者是一个内置了 R 语言及其集成开发环境 RStudio 的 Linux 内核的系统镜像，可以跨平台安装部署到个人计算机或计算集群上并通过浏览器直接访问数据处理界面（RStudio），也支持 shiny 应用的部署。xcmsrocker 在这个镜像基础上做了两步开发，一步是集成了常见的代谢组学相关的生物信息学、化学信息学、机器学习等开源软件包，预装了相关的编译工具与依赖库；另一步是开发了 rmwf 包，为常见的代谢组学数据分析提供了数据处理模版与演示数据。同时，该镜像通过 API 可直接调用常见数据分享平台的数据接口，可实现在网络环境下下载原始数据并重现数据分析结果的全流程操作。此外，作为开源软件，研究人员也可以通过提交自己的工作流来方便其他研究人员再现自己的研究成果。

统计计算（11.21 晚上）

异构计算软件栈进展

张先轶 （澎峰科技有限公司） ***19:00-19:30***

简介：张先轶，本科和硕士毕业于北京理工大学，博士毕业于中国科学院大学，曾于中科院软件所工作，之后分别在 UT Austin 和 MIT 进行博士后研究工作。国际知名开源矩阵计算项目 OpenBLAS 发起人和主要维护者。中国计算机学会高性能计算专业委员会委员，ACM SIGHPC China 执行委员。2016 年，创办 PerfXLab 澎峰科技，提供异构计算软件栈与解决方案。2016 年获得中国计算机学会科学技术二等奖，2017 年获得中国科学院杰出科技成就奖，2020 年获美国 SIAM Activity Group on Supercomputing 最佳论文奖。

摘要：高性能的计算软件栈作为底层硬件和上层应用的桥梁，可以扩展芯片的应用范围，提升计算性能。国际主流芯片公司都投入大量资源建设异构计算软件栈，例如 Intel oneAPI，NVIDIA CUDA-X 等。本报告将介绍澎峰在异构计算软件栈的工作进展，包括底层计算库和框架的支持与优化，PerfXPy 以及面向新一代计算硬件的支持工作。

用 Taichi 在 Python 中书写高性能并行计算程序

赵亮 （太极图形公司） ***19:35-20:05***

简介：赵亮，2022 年初加入 Taichi，任 Taichi 编程语言产品经理、技术布道师。

摘要： Taichi 是一门嵌入在 Python 中的领域专用编程语言，具有书写方便、运行效率高、移植性好的优点。在这次报告中我将向大家介绍关于 Taichi 编程语言的基础知识，演示使用 Taichi 开发的一些精彩例子，与其它加速方案的比较，并和大家探讨 Taichi 在不同领域科学计算中的应用前景。

生成模型与快速实时的 AI 无损压缩技术

张世枫 & 康宁 （华为诺亚方舟实验室） ***20:10-20:40***

简介：张世枫，博士毕业于清华大学计算机系，现为华为诺亚方舟实验室主任工程师。张世枫的主要研究方向为生成模型与 AI 数据压缩，相关领域在国际知名会议发表多篇论文；康宁，博士毕业于香港大学计算机系，发表过 STOC, FOCS 等多篇理论计算机顶级会议，曾获 ACM/ICPC 香港赛区冠军。现任职于华为诺亚方舟实验室，从事 AI 压缩等方面的研究，在实时压缩方面有若干研究工作。

摘要：伴随深度生成模型技术的发展，生成模型用于 AI 无损压缩能显著提升数据的压缩率。然而，AI 压缩方法的吞吐率低，生成模型推理与动态熵编码的低吞吐率是两大性能瓶颈。本次报告将介绍基于流模型等多种类型生成模型的高效 AI 压缩方法及动态熵编码方法，这些方法在数据压缩率、吞吐率等性能均取得业界最优水平。

A fast and scalable statistical framework for genetic risk prediction of large-scale datasets

蔡铭轩 （香港城市大学） ***20:45-21:15***

简介：蔡铭轩，香港城市大学生物统计系助理教授，香港科技大学统计学博士。主要研究方向包括统计遗传学，统计计算，贝叶斯推断等，研究成果发表于 The American Journal of Human Genetics, Journal of Computational and Graphical Statistics, Bioinformatics 等国际期刊。

摘要： The rich resources of massive genetic data offer an unprecedented chance for individualized disease risk prediction. Through statistical modelling, the risk scores derived from genetic variants can effectively identify the individuals with higher disease risk from general population. However, multiple challenges arise when constructing risk prediction from massive data. First, the massive genetic data usually is comprised of hundreds of thousands of samples with millions of variants. Computational cost for standard statistical analysis becomes unfordable. Second, the individual-level genetic data are usually of restricted access due to privacy protection. Third, due to the large difference of genetic architectures between populations and the limited sample size from non-European populations, risk prediction has been less accurate for the non-European individuals. To improve the prediction accuracy in non-European populations, we propose a cross-population analysis framework for genetic risk prediction with both individual-level (XPA) and summary-level (XPASS) genetic data. By leveraging trans-ancestry genetic correlation, our methods can borrow information from the Biobank-scale European population data to improve risk prediction in the non-European populations. In a Chinese cohort, our methods achieved 7.3%-198.0% accuracy gain for height and 19.5%-313.3% accuracy gain for body mass index (BMI) in terms of predictive R2 compared to existing prediction approaches.

社交媒体（11.22 晚上）

媒介接触与媒介信任的关系及在政治参与中的作用

毛佳艺 （中国传媒大学） ***7:00-7:30***

简介：毛佳艺，中国传媒大学数据科学与大数据技术（传媒大数据方向）专业本科三年级在读，热衷于结合结构方程模型、文本分析等方法探索社交媒体领域的媒介使用与社会参与，目前正在进行小游戏广告安全、主流媒体国际传播能力相关的研究，参与北京市级大创、省部级科研项目。

摘要：伴随互联网技术的更新迭代，网络已成为中国公民政治参与的常见途径，而媒体在政治参与中的作用日益凸显。据此，本文旨在探讨网民媒介接触、媒介信任与政治参与之间的潜在关系与影响机制，帮助提升民众政治参与积极性。基于 2018 年“网民社会意识调查”数据，对所选取的 15 个指标进行相关分析，并通过因子分析提取出大众传播媒介信任、个人传播媒介信任、媒介接触、政治参与四类公因子，经聚类分析得到特征均不同的三大类网民群体，据此了解了当前中国网民的媒介接触、媒介信任、政治参与的具体情况。采用结构方程模型来分析政治参与的影响因素，结果显示媒介接触直接正向影响政治参与，个人传播媒介信任既直接正向影响又间接正向影响政治参与，大众传播媒介信任间接正向影响政治参与。根据研究结果我们给出以下建议：其一，应严格管理传播媒介内容生产，提高从业者综合素质；其二，政府机关部门可主动培育 KOL 向大众普及法治意识，鼓励大家积极参政议政；其三，可通过采用 VR、互动短视频等时下新潮元素来丰富政治信息传播的趣味性和互动性，让群众主动接触、乐于关注、积极参与政治生活。

中文网络舆情极化动态模拟——基于大数据和多主体建模的探索

张卓 （中南大学） ***7:40-8:10***

简介：张卓，中南大学社会学系博士，导师为中南大学自动化系 & 社会学系教授、教育部青年长江学者吕鹏，研究兴趣包括网络群体行为、舆情传播、群体性事件和安全管理等，致力于多智能体仿真模拟（ABM）、机器学习（ML）、自然语言处理（NLP）、复杂系统和社会科学研究主题耦合探索。相关工作发表于中科院 Top 期刊 KBS、CSF、CIS、《社会发展研究》等。最近的工作研究国际舆情和国际关系演变。

摘要：在移动互联网和大数据时代，线上互动和意见碰撞愈发激烈，大量研究开始关注网络舆论的形成过程和演化动态。基于多智能体 Ising 模型，本研究探索线上个体的行为方式和舆论极化的演化机制。我们选择了豆瓣平台作为数据源，并以网络唱片“内圆外方”为研究话题，进行舆情建模和动态周期模拟。模型迭代 10,000 次模拟来寻找可能的最优解，并检验其拟合度和稳健性。最优参数模拟可以反映网络舆情的全生命周期。从不同层次和指标来看，拟合度或匹配度达到了最高水平。在微观层面上，真实案例和模型模拟中的个体行为分布相似，呈现高斯正态分布；在中观层面上，匹配了真实案例和模型模拟中的支持和反对的极化舆论在离散分布和连续分布；在宏观层面上，匹配了网络舆情极化的时序相变点（爆发、上升、高峰、消失）和全生命周期。因此，该模型精准捕捉个体行为的核心机制，反演和刻画了网络舆情极化的演化动态。

不可知标签选择偏差下的去偏差图神经网络

范少华 （北京邮电大学） ***8:20-8:50***

简介：北京邮电大学 GAMMA Lab 博士生，Mila 联培博士，导师为石川教授。主要研究方向为因果图神经网络，因果机器学习等。目前已在 NeurIPS, KDD, WWW, TNNLS 等会议期刊发表一作论文。

摘要：多数用于节点分类的 GNN 方法没有考虑数据中的选择偏差，即训练和测试数据非同分布。同时在现实中，我们很难在训练时获得测试数据，从而导致测试数据变的不可知。在有选择偏差的节点上训练 GNN 会导致明显的参数估计偏差，从而严重的影响在测试节点上的泛化性能。在本文中，我们首先进行了实验性研究，验证了数据选择偏差会严重的影响到模型的泛化性能，并且从理论上证明了数据选择偏差将会导致 GNN 模型参数估计上的偏差。为了消除 GNN 估计的偏差，我们提出了具有差分去相关正则项的去偏差图神经网络。在多个有偏差数据集上数据集上验证了该方法可以有效的去除数据选择偏差所带来的不良影响。

基于极大噪声众包标注的中文意见表达式识别研究

张鑫 （天津大学） ***9:00-9:30***

简介：张鑫，天津大学三年级硕士研究生，指导教师为张梅山老师，研究兴趣包括自然语言处理 (NLP)、机器学习 (ML) 和多模态等领域，致力于探索和利用机器学习流程管线中的人类因素和人机交互，以及人在回路的机器学习方法。相关工作发表于自然语言处理国际顶级会议 ACL、EMNLP、COLING 上，最近的研究工作主要关注众包数据标注下的信息抽取、情感分析等任务。

摘要：最近的意见表达识别（OEI）工作在很大程度上依赖于人工构建的训练语料的质量和规模，这可能是非常难以满足的。众包是解决这个问题的一个实用方案，其目的是创建一个大规模但质量没有可靠保证的语料库。在本工作中，我们研究了具有极大噪声的众包标注者的中文 OEI 问题，以较低的成本构建了一个数据集。遵循我们先前的工作，通过将所有标注者视为众包标注者者的黄金标准来训练标注者者-适配器模型，并通过使用合成专家来测试该模型，该专家是所有标注者者的均值。由于这种用于测试的标注者均值在训练阶段从未被明确建模，我们提出通过 mixup 策略生成合成训练样本，使训练和测试高度一致。在我们构建的数据集上进行的模拟实验表明，众包对 OEI 来说是非常有前景的，而我们提出的标注者者混合可以进一步加强众包建模，超越了众多先前工作。

可视化（11.22 晚上）

基于地图隐喻的可视化构建方法及其应用

陈帅（北京大学） ***19:00-19:25***

简介：陈帅，北京大学智能学院博士研究生。主要研究方向为复杂异构数据的可视分析，尤其是针对社交、新闻媒体数据的分析，在 IEEE VIS、EuroVis 上发表多篇论文。

摘要：隐喻地图是一种利用地图作为隐喻来可视化非空间数据的方法，在文本、网络等数据可视化中具有广泛的应用。已有工作提出了不同的隐喻地图可视化形式，但是缺乏高效、统一的隐喻地图构建方法，普通用户在构建隐喻地图可视化上存在较大困难。本次报告将讨论隐喻地图的统一构建方法以及在不同类型数据、任务场景下的具体应用。

可视分析赋能历史学群体研究

黄锦畦（香港科技大学） ***19:25-19:50***

简介：黄锦畦，香港科技大学博士研究生，他的研究兴趣涵盖可视化分析对不同专业领域的应用，发表 CCF-A 类论文 4 篇。

摘要：群体，是解读历史的核心要素。历史学家通过研究历史人物的行为来探讨社会结构的变化和社会流动的趋势。例如，乔治-华盛顿——美国第一任总统，他和其他开国元勋之间还有什么不为人熟知的联系？朱熹——理学集大成者，他是如何一步步扩展他的传道版图，形成以他为核心的理学群体，进而使理学成为元、明、清三朝的官方哲学？数据科学的加入，让传统历史学群体研究有了新的视角与方法。我们用数据思维、计算机手段来思考和解答以上问题。本次报告将分享我们和历史学家合作的最新成果——CohortVA，一种交互式的可视分析方法，使历史学家能够将专业知识和洞察纳入迭代探索群体的过程中，极大地提高群体识别、人物筛选和假设验证的能力。

时变上下文语义序列的比较可视分析方法

赵宇恒（复旦大学） ***19:50-20:15***

简介：赵宇恒，复旦大学大数据学院博士研究生，复旦大学可视分析与智能决策实验室成员，导师为陈思明老师。目前研究方向为社交媒体可视分析，相关工作发表于 ACM CSCW 等会议与期刊。

摘要：社交媒体文本数据的可视分析旨在从文本中挖掘丰富的语义信息，结合可视化帮助人们快速理解舆情内容及其演变。已有研究提出采用语义序列的方法来总结社交媒体文本内容。然而现存的可视化方法难以支持同时比较时变信息和具有上下文信息的语义序列。此外，由于社交媒体事件往往存在不同的关键人物或焦点引导话题的演变，分析这些不同数据流中的语义序列也更加困难。本次报告将讨论如何创建新颖的翅膀隐喻可视化来支持语义序列的复杂比较分析。

数据整理脚本的语义理解

熊凯（浙江大学） ***20:15-20:40***

简介：熊凯，浙江大学 CAD&CG 国家重点实验室的博士研究生，导师为巫英才教授，是 ZJUIDG（https://zjuidg.org）科研小组的成员。主要研究方向是数据整理和可视分析，尤其关注在如何帮助数据工作者理解数据转换过程的语义。

摘要：数据整理（Data Wrangling）是一种通过清洗和转换操作将复杂凌乱的数据整理成理想数据格式的过程，是数据存取、数据建模和数据可视分析等任务的重要前置步骤。利用 R、Python 等编程语言来编写特定的脚本是完成数据整理工作的常用手段。在现实工作中，理解数据整理脚本的语义（即数据是如何发生变化的）是数据工作者的常见需求，如代码调试、程序复用等。然而，数据整理操作的类型及其代码的实现方式复杂多样，使得数据工作者在理解脚本语义时费时费力。为了帮助数据工作者高效地理解脚本的语义，我们提出了一系列研究工作：SOMNUS，利用基于图形图符的节点链接图可视化数据表格的变化过程；以及 COMANTICS，结合数据表格的差异及 CNN 模型自动推断数据整理脚本的语义。

可视化的智能自然语言交互

刘灿 （北京大学） ***20:40-21:05***

简介：刘灿，北京大学博士研究生，导师为袁晓如研究员。2018 年获北京大学理学士、经济学士。研究方向为深度学习驱动的可视化。近年来，在 IEEE TVCG, ACM CHI, IEEE PacificVis 等会议期刊发表近十篇论文。获 IEEE VIS 最佳海报提名奖，IEEE PacificVis 最佳海报奖、提名奖，ChinaVis 最佳综述奖、最佳论文提名奖。

摘要：可视化是数据分析的重要方法。自然语言是人类智能的结晶。两者的交叉研究从两方面提升用户对数据的理解。一方面，自然语言降低需求表达门槛，使普通大众更简便地构建和使用数据可视化；另一方面，自然语言作为可视化的一部分来辅助用户理解数据。

基于交互式机器学习的数据重建与数据标注

张宇 （华为中软基础软件创新实验室） ***21:05-21:30***

简介：张宇，华为中软基础软件创新实验室数据可视化技术专家。牛津大学计算机系博士，北京大学智能科学与技术专业本科。主要研究方向为交互式机器学习以及数字人文，相关论文发表于 ACM CHI, ACM TIIS 等会议与期刊。

摘要：数据可视化有很长的历史，历史上的可视化编码了有价值的历史数据集，但是记录了这些原始数据集的文档如今通常已经佚失。为了复原这些数据集，需要从历史上的可视化中进行数据重建。本报告主要介绍通过交互式机器学习从可视化中进行数据重建，通过图形化开发工具低代码地开发数据重建流程，以及这些工作对通用数据标注工具的启发。

软件工具（11.23 晚上）

R 机器学习：mlr3verse 核心工作流

张敬信 （哈尔滨商业大学） ***19:00—19:30***

简介：张敬信，博士毕业于哈尔滨工业大学基础数学，现为哈尔滨商业大学数学与应用数学系主任、副教授、应用统计硕导、数学建模主教练；主讲课程：高等数学、实变函数、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，参加国家自然科学基金项目 2 项；出版《R 语言编程：基于 tidyverse》（人民邮电）、《数学建模：算法与编程实现》（机械工业）。常驻知乎平台，关注 7.6 万。

摘要： mlr3verse 是最新、最先进的 R 机器学习框架，它基于 R6 面向对象语法和 data.table 数据底层，支持搭建" 图" 流学习器，理念非常先进、功能非常强大。本报告将围绕语法基础、图学习器、集成学习、特征工程、嵌套重抽样、超参数调参、特征选择、模型解释梳理用 mlr3verse 做机器学习的核心工作流程。

R 语言高效数据操作工具：tidyfst

黄天元 （中国科学院文献情报中心） ***19:40—20:10***

简介：黄天元，中国科学院文献情报中心特别研究助理，复旦大学理学博士，热爱数据科学与开源工具（R），致力于利用数据科学迅速积累行业经验优势和科学知识发现，在 CRAN 维护有 tidyfst、tidyft 和 akc 三个 R 包，著有《R 语言数据高效处理指南》、《文本数据挖掘——基于 R 语言》。知乎专栏：R 语言数据挖掘。

摘要： dplyr 和 data.table 是 R 开源社区优秀的数据操作包，两者可以完成很多类似的数据处理（如筛选、排序、分组汇总等），但是又有不同的特点。dplyr 的函数组织形式更加用户友好，而 data.table 则具有令人惊艳的计算性能。关于如何结合两者之间的特色构造更好的数据操作工具，R 社区有很多尝试，而 tidyfst 包就是其中之一。本报告分享了 tidyfst 包在开发过程的整个历程，并介绍 tidyfst 包作为高性能数据操作工具在使用上的便捷性。

A Metadata Approach for Analysis & Reporting in Clinical Trials

赵妤婕 （Merck） ***20:20—20:50***

简介： Yujie Zhao (赵妤婕), Ph.D. is a statistician from Merck. Yujie works on the methodology research in clinical trials with a focus on group sequential designs. She also works with a group of statisticians and programmers to demonstrate the capability of using R for data analysis in clinical trials. Yujie has published 5+ first-author papers on statistical computations, statistical process control, and tensor decomposition. Before joining Merck, she earned a Ph.D. degree in Industrial Engineering at Georgia Tech in 2021.

摘要： In clinical trials, there is a growing trend to get reproducible analysis and & reporting. In this presentation, we will present an end-to-end automation framework to construct clinical datasets into metadata. Additionally, we will demonstrate the generation of analysis reports by metadata. A nice feature of this metadata approach is its automation. For example, users can update the analysis by simply updating operations and all deliverables can be automatically updated based on

upstream metadata changes. The work is available at https://github.com/Merck/metalite and https://github.com/Merck/metalite.ae.

用 R 语言解读传染病模型

张丹 （北京青萌数海科技有限公司） ***21:00-21:30***

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、大数据、数据分析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。熟悉数据学科方法论，在海关、外汇等监管科技领域均有落地项目。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： 新冠疫情几次变异，极大地影响着我们的正常生活和工作。特别是 2022 年 2 月以来的 Delta 变异株感染，在上海和北京这种人口超大型城市中，有着超强的传染力。在流行病学领域，有几种不同传染病的传播模型，可以模拟病毒的传播过程。本次分享将使用 R 语言，来给大家演示病毒传播的过程。了解了病毒传播的逻辑，能让我们更加坚定战胜病毒的决心。本次分享的传染病模型，涉及到 2 个包 EpiModel（数学模型），nCov2019（下载数据和可视化）。

用 dataMojo R 包开发高效数据分析应用

古杰娜 （麦肯锡咨询公司） ***21:40-22:10***

简介： 古杰娜，目前在麦肯锡咨询公司担任软件架构师，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： 本报告将介绍近期开发的 dataMojo R 包（https://github.com/jienagu/dataMojo）以及用其开发的语法简洁且高效数据分析应用（https://github.com/jienagu/demo_mojo_app）。dataMojo R 包是基于 data.table 为框架的数据分析扩展包，能够覆盖很多数据处理工作中的场景。本报告将通过一系列实例展示此 R 包的独特优势。

生态环境（11.23 晚上）

青藏高原植物物候变化及驱动机制研究进展

沈妙根 （北京师范大学） ***19:00-19:45***

简介：沈妙根，北京师范大学教授，博士生导师，中组部万人计划青年拔尖人才，以第一或通讯作者在 Nature Reviews Earth & Environment、美国国家科学院院刊 (PNAS)、Global Change Biology 等 SCI 期刊发表论文 30 余篇，篇均 SCI 引用 60 余 次，其中 4 篇长期入选 ESI 前 1% 高被引论文。曾获青藏高原青年科技奖、西藏自治区科学技术一等奖和农业部神农中华农业科技奖等。入选全球前 2% 顶尖科学家榜单。

摘要：物候变化是高寒地区陆地生态系统响应气候变化的敏感指标。在全球气候变化背景下，青藏高原植物物候变化引起的生态系统变化改变陆面和大气过程，进而影响高原和周边地区的天气和气候。本报告阐述青藏高原植物物候时空变化和驱动机制等。

室内固体燃料排放、空气污染及其暴露风险

沈国锋 （北京大学） ***19:45-20:30***

简介：沈国锋博士现为北京大学城市与环境学院的新体制研究员，研究生导师。研究方向是环境污染化学和区域环境过程，重点研究有毒有害污染物（包括传统污染物和新污染物）的来源、环境归趋、暴露、健康风险和控制政策等。近 5 年主持国家自然科学基金委优秀青年基金、面上基金，科技部“第二次青藏高原综合科学考察”专题，中科院 A 类先导专项子课题等国家和省部级项目，作为主要成员参与国家基金委重大项目、重点基金、国际合作等项目。在 Nature 子刊，Sci. Adv., PNAS, NASR，ES&T,《科学通报》等领域的知名期刊上合作发表 SCI 论文 150 多篇，其中独立通讯或第一作者 80 余篇。参编英文专著 2 部，作为主要技术人员参与制定国际测试标准 2 个，国家测试标准 2 个。目前担任国际杂志 AECT, Sustainable Horizons 副主编，CREST，B&B，ESE,《颗粒学报》、《环境科学》、《生态环境学报》等编委或青年编委。兼任中国地理学会环境地理专委会副主任委员。

摘要：清洁可负担的现代能源是可持续发展的重要指标，同时也与健康、气候应对、土地利用、不平等性等相关。居民生活中的固体燃料使用与室内环境、区域空气控制等直接相关，进而影响居民健康和气候变化。本报告基于多年实验和模型研究得到的基础数据，介绍了居民生活能源利用、污染物排放、空气质量影响贡献及清洁干预效果。

祁连山的地貌过程与数值模拟

耿豪鹏 （兰州大学） ***20:30-21:15***

简介：耿豪鹏，兰州大学资源环境学院地貌与第四纪地质研究所，副教授, 博导，副院长。2014 年获得兰州大学自然地理学博士学位，2011 年-2013 年于美国加州大学伯克利分校联合培养，研究方向为地貌演化与数值模拟，主要从事西北活动造山带地貌演化的量化理论研究，以及地表风化侵蚀过程的模型研发工作。在包括 Sci. Bull.、EPSL、Catena、ESPL、中国科学等在内的国内外知名期刊发表学术论文 30 余篇。主持国家自科基金青年项目与面上项目，参与国家自科基金重点项目 2 项。服务于“地貌学”国家级精品课程授课团队，获批甘肃省线下一流课程（地貌学），发表教学论文 2 篇，管理论文 3 篇，2018 年获兰州大学师德师风建设“先进个人”，2018 年获“甘肃省技术标兵”称号。

摘要：祁连山位于青藏高原东北缘，是高原最为年轻的组成部分，坐落在西北干旱区的边缘，是我国重要的生态安全屏障。祁连山的地貌演化过程对理解高原周缘的生长及其环境效应具有重要的意义。本研究将介绍祁连山地貌过程的基本特征、多尺度侵蚀速率的控制因素，以及风化限制区滑坡侵蚀对过去气候变化的响应机理，为干旱区活动造山带演化提供了新的理论认识。

祁连山冰冻圈变化及其影响

杜文涛 （中国科学院西北生态环境资源研究院） ***21:15-22:00***

简介：杜文涛，中国科学院西北生态环境资源研究院正高级工程师，博士生导师，中国科学院西部之光青年学者、关键技术人才入选者，甘肃省生态环境标准化委员会委员，冰冻圈科学国家重点实验室副主任。主要从事高山气候、雪冰变化归因及其水文、环境效应研究及技术革新，曾获甘肃省科技进步三等奖、甘肃省学术大会优秀论文奖及兰州市优秀科技工作者。

摘要：祁连山是气候环境变化和人文交流的重要区域，也是我国冰冻圈监测研究的萌生地。全球变暖背景下，祁连山气候曲线和极端事件均反映出暖湿化特征，本报告详细阐述祁连山冰冻圈变化过程、原因及其影响。

工业大数据（11.24 晚上）

数据驱动的电动汽车充电过程监控

宋哲 （南京大学商学院） ***19:00-19:30***

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。主要研究方向为工业大数据驱动的预测建模、复杂网络建模与仿真、智慧能源管理。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 3000 多次, 获中国和美国发明专利 13 项。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。担任十多个国际一流期刊的审稿人, 如 IEEE Trans. Industrial Informatics， European Journal of Operational Research、IEEE Trans. Industrial Electronics、IEEE Trans. Systems, Man, and Cybernetics 等；INFORMS，IISE，IEEE 协会会员。

摘要：据公安部统计，截至 2022 年 9 月底，全国新能源汽车保有量达 1149 万辆，占汽车保有量的 3.65%。其中，纯电动汽车保有量 926 万辆，占新能源汽车总量的 80%。由于车质量参差不齐，用户驾驶习惯和充电安全意识不够高，近年来电动汽车自燃和充电过程中起火事故常有发生，给人民群众生命财产带来极大的安全隐患。电动汽车在充电过程中，车辆 BMS（Battery Management System）系统和充电桩之间进行通讯，会产生大量的有价值实时数据，比如电流、电压、SOC、温度等等。将这些数据“养殖”起来，结合动力电池故障机理和残差方法，经典的统计控制图和机器学习算法就能起到事半功倍的故障监测和预警效果。本次报告通过几个案例展示了如何将经典的统计学和机器学习算法应用到电动汽车充电过程安全健康评估领域。

针对生产设备维护和调度联合优化的建模方法

张玺 （北京大学工业工程系） ***19:30-20:00***

简介：报告人张玺目前是北京大学工业工程与管理系副教授，研究领域主要面向工业系统以及高端制造过程的实时监测、诊断、控制、优化和运维管理。在数据科学、工业工程、质量与可靠性工程、制造工程等领域内知名期刊诸如 JMLR、Technometrics、IISE Transactions、JQT 等发表多篇学术论文，已授权 13 项国家发明专利及 3 项软件著作权。

摘要：在多工序制造系统中，生产调度和设备维护是两个不可分割的且需要同时决策的任务集。在对生产现场进行排产和设备维修决策时，现有的研究往往简化生产设备退化造成的不确定性影响来达到同时优化的目的，极易忽略这两者之间的相互作用，从而降低了整个制造系统的生产效益。本次报告重点围绕这一问题展开，提出了针对生产调度和设备维护之间的交互关系的建模思想，并依此建立了生产调度和设备维护的联合优化方法。

Anomaly Detection for Fabricated Artifact by Using 3D Point Cloud Data

杜娟

（香港科技大学（广州）） ***20:00-20:30***

简介：现任香港科技大学（广州）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联属助理教授。主要从事数据驱动的智能制造系统质量改善研究工作，在工业数据分析、质量控制和先进制造领域发表多篇高水平期刊论文，并获多项国家发明专利和软件著作版权。多次应邀在国际、国内权威学术会议作报告，现为 IISE 的高级会员。

摘要： Recently, various advanced 3D scanners have been widely used in manufacturing industries to collect 3D point cloud data of fabricated artifacts. The extra dimension of 3D point cloud data can provide more detailed descriptions about anomalies in artifact surfaces than 2D image data. 3D point cloud data can be categorized into structured and unstructured point clouds. Compared with structured 3D point cloud data, unstructured point cloud data can capture the surface geometry more completely. However, anomaly detection by using unstructured 3D point cloud data are more challenging due to unstructured data representation, inconsistent point sizes, and high dimensionality. To deal with these challenges, this talk will present some recent advances in anomaly detection by using unstructured 3D point cloud data. The accuracy and robustness of the proposed method are validated by simulation studies and case studies.

基于 Modern Data Stack 的时空数据平台

张源源 （百姓车联） ***20:30-21:00***

简介：张源源，在百度、乐动力、阿里、百姓车联等多家赛道内头部公司有过行业内开创性的工作，在传感器数据、手机信令数据、轨迹数据等领域的数据科学、数据平台工作有近 10 年积累。

在乐动力期间，独立负责国内第一家入选苹果 Appstore 年度精选 App 的数据科学工作。

在阿里期间，开发的 AI 运动是业内第一个在手机端实时进行健身动作计数的应用，在疫情期间帮助数十万高校师生顺利开展体育课教学，并推广给数千万的中小学生；负责的走路、跑步运动商业化算法工作，年创收近亿元。

在百姓车联期间，带领团队开发了业内领先的危险驾驶行为识别 SDK 和 UBI 解决方案，申请了 7 项专利、发表了 2 篇 CCF-A 类 paper，并成为保险行业第一个入选中国人民银行金融科技创新应用监管沙盒的项目；正在带领团队开发业内首个可扩展、高性能、云原生、一站式的时空数据平台，旨在解决时空数据领域 Data/AI Infra 缺失的问题。

摘要：作为一种特殊数据类型，时空数据可能是最被低估的数据金矿，一方面，现实世界中的数据超过 80% 与地理位置有关，并继续高速增长着，另一方面，大量的时空数据躺在那里，等待它的戈多。与此同时，时空数据内部结构也发生着巨大变化，近 10 年，随着移动互联网、车联网的发展，时空数据的主要载体已经从静态的遥感数据逐步转变为轨迹数据、手机信令数据，但存储、索引、计算、可视化等各个环节的工具都没有跟上趋势。

毫无疑问，落后的生产力工具抑制了数据使用需求，但疫情防控、交通物流、O2O、车联网、保险等行业又有海量的需求在爆炸性增长着，行业急需可扩展、高性能、云原生、一站式的时空数据平台。

本次分享将从存储、索引、计算、可视化等各个环节出发，讲述我们从 0 到 1 建设时空数据平台的思考和抉择。

知识-技术-实践的知识传播-图书出版人的视角

吕潇 （机械工业出版社） ***21:00-21:30***

简介：机械工业出版社电工电子分社策划编辑/副编审，本科毕业于西安电子科技大学微电子学院微电子学专业，专职从事科技图书出版工作 14 年，选题方向为电子技术、数据科学、建筑电气领域，兼主管部门数字化产品开发与策划。主导或参与多个国家自然科学基金、社会科学基金、国家出版基金项目，策划工程技术类图书 90 余种，编辑加工书稿 6000 余万字；曾获中国编辑学会学术论文二等奖，参与编写图书 3 本，所策划及担任责任编辑的图书曾获中国出版传媒集团、中国通信学会、机械工业信息研究院等机构颁发的奖项。

摘要：碎片化时代的系统化知识价值，广义上图书作为知识载体的的意义；不是讲技术的书都是学校教材，科技出版是服务岗位、服务垂直应用的——以《工业大数据分析算法实战》为例简单说明；数据科学题材的出版需求，以及出版社的枢纽作用

制药工业的数字化与统计/机器学习智能化应用

姚树亮 （苏州金玉问道数据科技有限公司） ***21:30-22:00***

简介：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。曾负责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。负责公司研发及生产数据的统计学支持，及部分机器学习算法在制药工业生产与产品工艺优化的应用。

摘要： 1. 计算机存储与大数据存储、计算的发展 2. 人类基因组计划与制药工业发展 3. 制药行业的数字化转型热潮 4. 统计统计和机器学习在制药行业的数字化/智能化应用。

基于设备机台数据分析的生产人员绩效评估

解光耀 （昆仑数据） ***22:00-22:30***

简介：昆仑数据高级数据分析师，清华大学工程物理系本科、博士，从事于工业数字化场景建设的场景探索、模型研发、数据架构设计等工作。长期从事工业设备的状态监测与异常识别、智能运维、故障诊断等方面的研发工作，在工程机械、装备制造、电子制造等行业拥有多年丰富的工作经验。

摘要：1. 工厂运营人员管理的痛点 2. 如何用数据驱动的方法改变现有的人员管理模式 3. 设备数据在人员绩效业务上的表达 4. 绩效评估模型应用效果

生物统计（11.24 晚上）

Neural Network on Interval Censored Data with Application to the Prediction of Alzheimer’ s Disease

孙韬 （中国人民大学） ***19:00-19:45***

简介：孙韬，中国人民大学统计学院讲师，博士毕业于匹兹堡大学生物统计系，研究方向为复杂生存数据模型，老年慢性病预防与管理。主持国自然与国家统计局项目各一项，论文发表于 Science, Biostatistics, Biometrics 等期刊。

摘要： Alzheimer’s disease (AD) is a progressive and polygenic disorder that affects millions of individuals each year. Given that there have been few effective treatments yet for AD, it is highly desirable to develop an accurate model to predict the full disease progression profile based on an individual’s genetic characteristics for early prevention and clinical management. This work uses data composed of all four phases of the Alzheimer’s Disease Neuroimaging Initiative (ADNI) study, including 1740 individuals with 8 million genetic variants. We tackle several challenges in this data, characterized by large-scale genetic data, interval-censored outcome due to intermittent assessments, and left truncation in one study phase (ADNIGO). Specifically, we first develop a semiparametric transformation model on interval-censored and left-truncated data and estimate parameters through a sieve approach. Then we propose a computationally efficient generalized score test to identify variants associated with AD progression. Next, we implement a novel neural network on interval-censored data (NN-IC) to construct a prediction model using top variants identified from the genome-wide test. Comprehensive simulation studies show that the NN-IC outperforms several existing methods in terms of prediction accuracy. Finally, we apply the NN-IC to the full ADNI data and successfully identify subgroups with differential progression risk profiles.

可变剪接的遗传调控及其在复杂性状和疾病中的独特的重要作用

祁婷 （西湖大学） ***19:45-20:30***

简介：祁婷博士，西湖大学副研究员。主要研究方向是统计遗传学，通过整合多组学数据解析人类复杂性状和常见疾病的遗传机制。相关工作已经发表在 Nature Genetics，Nature Communications 等杂志。

摘要：生物学中心法则描述了遗传信息的传递过程，包括基因的转录、RNA 的剪接、修饰和翻译。在此过程中，RNA 聚合酶以 DNA 为模板合成前体信使 RNA，开启遗传信息的传递。前体信使 RNA 通过不同的剪接方式（即：在不同的剪 接位点剪掉内含子，连接外显子），生成多样化的成熟信使 RNA，这个过程 称之为可变剪接（或选择性剪接）。据统计，约 95% 的人类基因存在可变剪接，有些基因的剪接方式多达数百种，这是基因表达调控和蛋白质组多样性形成的 重要机制。可变剪接的异常会导致生理状态的失衡和疾病的发生。因此，建立 全面的可变剪接遗传调控图谱及其与常见疾病之间的关联图谱对可变剪切的分子 机制研究以及疾病治疗新靶点的发掘有着重要意义。

研究团队开发了一款高效的 RNA 可变剪接遗传调控位点（splicing QTL 或 sQTL）定位新方法，将其命名为 THISTLE；利用该方法系统地分析了 2865 个脑组织的 转录组和遗传学数据，

绘制了迄今为止最全面的可变剪接遗传调控图谱；通过将该 sQTL 图谱映射到精神分裂、阿尔兹海默症、帕金森氏症等大脑相关性状 和疾病的全基因组关联分析数据中，鉴定出 244 个易感基因，其中 61% 基因的机制 无法被基因转录水平的遗传调控所解释，揭示了 RNA 可变剪接在复杂性状和疾病 遗传机制中独特的重要作用。

STAARpipeline: A comprehensive framework for flexible and scalable rare variant association analysis using whole-genome sequencing data and annotation information

李子林 （Indiana University） ***20:30-21:15***

简介：李子林，印第安纳大学医学院生物统计与健康数据科学系助理教授。历任哈佛大学陈曾熙公共卫生学院生物统计系研究员、副研究员和博士后，本科与博士毕业于清华大学数学科学系，师从林希虹院士。主要研究方向为高维数据中的统计方法理论和遗传统计学。相关研究成果在 Journal of American Statistical Association、Nature Methods、Nature Genetics、The American Journal of Human Genetics 等国际学术期刊发表。入选首批美国国家心肺血液研究所生物数据云计算平台研究员（National Heart, Lung and Blood Institute BioData Catalyst Cohort I Fellow），获得国际数理统计协会颁发的 2021 年度 New Researcher Travel Award。

摘要： Large-scale whole-genome sequencing (WGS) studies have enabled the analysis of rare variant associations with complex human diseases and traits. Variant set analysis is a powerful approach to studying rare variant associations. However, existing methods have limited ability to define the variant set in the genome, especially for the noncoding genome. We propose a computationally efficient and robust rare variant association-detection framework, STAARpipeline, to automatically annotate a WGS study and perform flexible rare variant association analysis, including gene-centric analysis and fixed-window and dynamic-window-based non-gene-centric analysis by incorporating variant functional annotations. In gene-centric analysis, STAARpipeline groups coding and noncoding variants based on functional categories of genes and incorporate multiple functional annotations. In non-gene-centric analysis, in addition to fixed-size sliding window analysis, STAARpipeline provides a data-adaptive-size dynamic window analysis. All these variant sets could be automatically defined and selected in STAARpipeline. STAARpipeline also provides analytical follow-up of dissecting association signals independent of known variants via conditional analysis. We applied the STAARpipeline to analyze the total cholesterol in 30,138 samples from the NHLBI Trans-Omics for Precision Medicine program. All analyses scale well in computation time and memory. We discover several potentially new significant associations with lipids, including a finding of rare variants in an intergenic region near JKAMPP1 associated with total cholesterol. In summary, the STAARpipeline is a powerful and resource-efficient tool for association analysis of biobank-scale WGS studies.

Powerful, Scalable and Resource-Efficient Rare Variant Meta-Analysis of Whole-Genome Sequencing Studies Using Summary Statistics and Functional Annotations

厉希豪 （Harvard T.H. Chan School of Public Health） ***21:15-22:00***

简介：Xihao Li is a postdoctoral research fellow in the Department of Biostatistics at Harvard T.H. Chan School of Public Health, mentored by Professor Xihong Lin. Prior to this, he received his Ph.D. in Biostatistics at Harvard University. Dr. Li's research interests lie in developing novel statistical

methodologies that enable scalable and integrative analysis of large-scale whole-genome/whole-exome sequencing data and multi-omics data. He has also worked on methodological projects to develop statistical approaches for rare disease clinical trials and real-world evidence studies.

摘要： Large-scale whole-genome/exome sequencing (WGS/WES) studies have enabled the analysis of rare variants (RVs) associated with complex human traits and diseases. Existing RV meta-analysis approaches are not scalable when applied to WGS/WES data. We propose MetaSTAAR, a powerful and resource-efficient RV meta-analysis framework, for large-scale WGS association studies. MetaSTAAR accounts for population structure and relatedness for both continuous and dichotomous traits. By storing LD information of RVs in a new sparse matrix format, the proposed framework is highly storage efficient and computationally scalable for analyzing large-scale WGS/WES data without information loss. Furthermore, MetaSTAAR dynamically incorporates multiple functional annotations to empower RV association analysis, and enables conditional analyses to identify RV-set signals independent of nearby common variants. We applied MetaSTAAR to identify RV-sets associated with four quantitative lipid traits in 30,138 related samples from the NHLBI TOPMed Program Freeze 5 data, consisting of 14 ancestrally diverse studies and 255 million variants in total, as well as the UK Biobank WES data of 200,000 related samples.

软件工具（11.25 晚上）

ModelOps 在数据科学平台中的实践与应用

殷自强（和鲸科技） ***19:00—19:30***

简介：殷自强，和鲸科技联合创始人，担任和鲸科技执行总裁兼首席产品官，毕业于上海交通大学信息安全专业，负责公司产品战略设计与规划，拥有丰富的数据科学产品、PLG 模式、数据驱动型组织客户旅程研究经历，主导了和鲸 ModelWhale 与和鲸社区从 0 到 1 的产品设计。

摘要：模型的全生命周期有许多自己的特点，传统 Devops 方法在模型管理中并不适用，同时相比于机器学习方法，延拓到更广义的决策模型在国内的研究与应用的适用性更广，本报告讨论了此类模型全生命周期的特点，如何更好的定义和管理过程中不同生产要素的版本、关系和交付，同时能够站在跨角色协同与工作流协同的角度对 ModelOps 进行设计

别再让你的 Shiny 代码源文件乱成一团了

苏玮（日本雅虎） ***19:40—20:10***

简介：苏玮，硕士毕业于东京大学应用生命工学专业，现任日本雅虎搜索广告前端工程师。本科开始接触 R 语言，持续对 R 语言和前端交互方面的发展保持关注。曾于日本新冠流行初期独立开发出访问量破千万级的 Shiny 应用。

摘要：对于倾向于研究的科研和分析人员来说，Shiny 框架的出现，方便了非软件开发人员也能够快速、独立完成带有交互性的数据展示页面应用。但也正是由于 Shiny 应用的大部分开发者更偏向研究人员的这一属性，在应用的可维护性方面往往会有所欠缺，导致项目难以获得良好的管理、扩展和协作。本报告主要基于个人的开发经验，面向 Shiny 的初中级开发者介绍一种简单易懂的 Shiny 应用代码文件的管理方法，降低 Shiny 应用的开发门槛和学习成本，从而快速搭建具有工程思维的 Shiny 应用。

R 语言中的并行计算，从 parallel 到 foreach 和 future

成超（先声药业） ***20:20—20:50***

简介：现任先声药业高级统计师。博士毕业于上海财经大学应用统计专业，研究方向为高维数据的降维、变量选择和高性能计算。开发了针对高维数据的稳健亚组识别与变量选择的 R 包 RSAVS。热爱开源、Linux、R 和电子游戏。

摘要：很多时候我们需要利用大量的模拟来验证所提出的统计方法的效果，随着计算资源的不断发展，并行计算成为了一个有效的提升计算速度的方法。在这次分享中，我将基于自己在博士阶段的经历，向大家介绍如何在 R 语言中进行并行计算。从利用 parallel 包自己准备并行环境，到利用 foreach 和 future 这样的包实现更简单易用的并行计算。

森林图绘制包：forestploter

阿力木·达依木（剑桥大学） ***21:00-21:30***

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。于 2021 年进入剑桥大学肿瘤学系临床试验中心从事临床试验设计、纵向数据分析等方面的研究。consort 和 forestploter 包的开发者和维护者。

摘要：森林图在临床研究中主要用于 meta 或回归分析中展示效应大小，现有的 R 包存在用途单一、过于复杂或不灵活等问题。forestploter 包将森林图内容、效应以及主题进行分开，在降低森林图绘制复杂性的同时提高灵活性。本报告将介绍 forestploter 包背后的理念，同时通过示例展示绘制简单及复杂的森林图，以及主题修改以及事后对森林图进行编辑。

DT：给交互式表格设置静态样式

袁凡（合众人寿保险股份有限公司） ***21:40-22:10***

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：本报告主要介绍在使用交互式表格包 DT 时，如何给表头、表格主体、按行、按列来设置（CSS）样式，以及往表格里插入 Unicode 字符、图片、超链接、字体图标（icon）、迷你图（sparkline）的基本方法。

医疗卫生与健康（11.25 晚上）

人工智能医疗器械漫谈

李舰（江西中科九峰智慧医疗科技有限公司） ***19:00-19:30***

简介：李舰，江西中科九峰智慧医疗科技有限公司 CTO，一直专注于数据科学在行业里的应用，编著了《统计之美》《数据科学概论》《数据科学中的 R 语言》等书，在 R 语言社区贡献了 MSToolkit、Rwordseg、tmcn 等包。在医疗健康领域发表了期刊 5 篇，主持省部级科技项目 3 项。

摘要：医疗器械直接关系着人类的生命健康，我国对其进行了严格的监管，根据安全性对医疗器械进行分类管理，其中第三类医疗器械是安全管理级别最高的类别，NMPA 明确了用于辅助决策的人工智能医用软件必须按照第三类医疗器械管理。演讲者曾经带领团队经历了一款人工智能三类证产品从研发到获批的漫长过程，将会在本次演讲中分享人工智能医疗器械的研发及注册的相关经验。

AI 与新药研发的真实世界的碰撞与机遇

李翛然（医图生科（苏州）生命科学技术有限公司） ***19:30-20:00***

简介：李翛然，毕业于利兹大学金融数学，现为医图生科（苏州）生命科学技术有限公司联合创始人/CEO。先后从事过保险精算，投资银行工作。于 2014 年创办奇点信息技术有限公司，为各大机构智能化管理系统与机器人业务。先后有 10 余家医院采用其 AI 辅助诊断系统，后公司被收购。现公司专业从事 AI 新药研发业务。

摘要：近年来 AI 与制药行业的碰撞在资本的加持下，逐渐进入了大众的视野。不过对于普通观众来说，到底这两个“最前沿”的人类科学技术领域到底是如何互相影响与协同，一直是一个未曾掀开的面纱。本次的主题为一家 AI 制药企业的负责人角度，看待一个创新药物的研发流程与 AI 的互动。

真实世界研究：从数据到证据

周梦戈（中国医学科学院基础医学研究所、北京协和医学院基础学院流行病与统计学系） ***20:00-20:30***

简介：周梦戈，女，流行病与卫生统计学博士，中国医学科学院基础医学研究所、北京协和医学院基础学院流行病与统计学系助理研究员。周梦戈博士毕业于首都医科大学，后加入清华大学万科公共卫生学院进行博士后研究。主要研究方向为慢性病（尤其是心血管疾病和糖尿病）的流行病学及临床研究，感染性疾病的预防研究等。以第一作者在 The Lancet Diabetes & Endocrinology、Journal of the American College of Cardiology、Mayo Clinic Proceedings 及 Diabetologia 等杂志发表多篇论文及摘要。参与多本书籍和指南的编写工作，包括《中国心血管病防治蓝皮书》、《高血压》、《中国成人血脂异常防治指南 2016》等。目前任《中国误诊学杂志》编辑委员会青年副主编，Cardiovascular Innovations and Applications 杂志青年编委等。

摘要：真实世界研究从 1992 年正式提出其概念，至今已近三十年。但近年来，大家才逐渐认识到它在医疗产品上市前的临床评估、上市后的安全监管、效果评估等方面的应用意义，

关注度日益增加。从政策层面看，美国食品药品监督管理局于 2017 年发布《采用真实世界证据支持医疗器械的法规决策》，2020 年，国家药监局发布《真实世界证据支持药物研发与审评的指导原则（试行）》。从医疗大环境看，医疗大数据的构建给真实世界研究提供了前所未有的便利。但真实世界研究所获得的数据往往存在异质性强、数据缺失多的问题，因此对统计方法提出了更高的要求。本报告将首先厘清什么是真实世界研究及其优势；进而介绍真实世界研究的数据预处理方法（主要是缺失数据的填补方法）以及目前常用的统计分析方法，并讨论这些方法的优势和劣势；最后，将基于中国心血管疾病医疗质量改善项目介绍我们前期做的一些真实世界研究以及展望。

贝叶斯 Lasso 并行中介模型

张沥今（斯坦福大学） ***20:30-21:00***

简介：张沥今，斯坦福大学发展与心理科学项目博士研究生，研究方向为心理测量，目前主要关注如何提升和评估结构方程模型的泛化推广能力。个人主页 https://lijinzhang.com

摘要：中介变量在帮助研究者理解变量间影响机制中发挥了重要的作用。而在行为研究中，变量间的影响机制通常存在多个潜在的中介变量。目前常用的中介效应检验方法包括 Sobel 法、百分位数 Bootstrap 法和偏差校正的 Bootstrap 法。但是当模型中存在多个中介变量时，这些方法难以有效地处理模型过拟合问题。为了在并行中介模型中提供更有效的变量选择工具，Serang 等人（2017）将频率学派下的 Lasso 方法应用到并行中介模型进行辩论选择，发现该方法能够更好地平衡模型复杂性和泛化能力，在小样本情况下也有着更好的表现。然而这种方法存在两个局限：（1）没有考虑到行为研究中对潜变量测量时测量误差的存在；（2）不能提供中介效应的区间估计结果。为了解决这两个局限，本研究将贝叶斯 Lasso 方法扩展到含有潜变量的并行中介模型中，并采用模拟研究与传统的中介变量选择方法进行对比。实证研究进一步展示了如何在实际数据中构建贝叶斯 Lasso 并行多重中介模型，并深入对比了该方法与传统方法的表现差异。

金属元素复合暴露与炎症因子关联性研究：基于三种统计模型比较

李昂（中国医学科学院基础医学研究所，北京协和医学院基础学院，流行病与卫生统计学系） ***21:00-21:30***

简介：李昂，医学博士，中国医学科学院基础医学研究所，北京协和医学院基础学院流行病与卫生统计学系博士后。研究方向为环境流行病学。主要关注空气污染、重金属、环境内分泌干扰物暴露的人群健康效应。目前担任 Frontiers in Public Health 和 Metabolites 客座编辑。研究论文发表在 Environment International、Science of the total environment、Environmental Pollution、Exposure and Health 等杂志。

摘要： Objective: We aimed to investigate whether urinary metal exposome mixtures are associated with the homeostasis of inflammatory mediators in middle-aged and older adults. Methods: A four-visit repeated-measures study was conducted with 98 middle-aged and older adults from five communities in Beijing, China, and 391 observations were included in the analysis. The urinary concentrations of 10 metals were measured at each visit using inductively coupled plasma mass spectrometry (ICP-MS), and the detection rates were all above 84%. Similarly, 14 serum inflammatory mediators in six categories reflecting inflammation regulatory homeostasis were measured using a Beckman Coulter analyzer and the Bio-Plex MAGPIX system. A linear mixed model (LMM), LMM with least absolute shrinkage and selection operator regularization (LMM-

LASSO), and Bayesian kernel machine regression (BKMR) were adopted to explore the effects of urinary metal mixture on inflammatory mediators. Results: In LMM, a two-fold increase in urinary cesium (Cs) and chromium (Cr) was statistically associated with -35.22% (95% confidence interval [CI]: -53.17, -10.40) changes in interleukin 6 (IL-6) and -11.13% (95%CI: -20.67, -0.44) in IL-8. Urinary copper (Cu) and selenium (Se) was statistically associated with IL-6 (88.10%, 95%CI: 34.92, 162.24) and tumor necrosis factor-alpha (TNF-α) (22.32%, 95%CI: 3.28, 44.12), respectively. Similar results were observed for the LMMLASSO and BKMR. Furthermore, Cr, Cs, Cu, and Se were significantly associated with other inflammatory regulatory network mediators. For example, urinary Cs was statistically associated with endothelin-1, and Cr was statistically associated with endothelin-1 and intercellular adhesion molecule 1 (ICAM-1). Finally, the interaction effects of Cu with various metals on inflammatory mediators were observed. Conclusion: Our findings suggest that Cr, Cs, Cu, and Se may disrupt the homeostasis of inflammatory mediators, providing insight into the potential pathophysiological mechanisms of metal mixtures and chronic diseases. Keywords: Bayesian kernel machine regression; Inflammatory mediator homeostasis; Middle-aged and older adults; Repeated measurement; Urinary metal exposome

safetyGraphics: 具有完全交互性的开源整合框架

徐永 （上海语擎信息科技有限公司） ***21:30-22:00***

简介：徐永毕业于西安交通大学电子工程专业，硕士学位。曾在全球排名第一的 IT 市场研究公司 Gartner 任中国首席数据分析师, 后参与创办上海语擎信息科技有限公司, 完成全定制化搜索引擎、社会化协作框架等产品. 最近三年进入医学临床数据分析及药物警戒数据合规系统的研发, 完成了从职业程序员向医学数据分析的跨界. 熟悉 Java, Python 和 R 语言, 在桌面应用、前后端编程、数据结构、统计学算法及可视化方面拥有丰富经验, 并能深刻理解数据结构与实际需求的匹配和相应的数据库搭建.

摘要： safetyGraphics 是一个在美国 ISG 工作组指导下新开发的一套系统性整合框架. 它可以轻易地把来自不同 R 包的图表组成相互关联的一套分析工具, 共享原始数据, 可随时针对不同的药物实验建立一套专门的图表顺序分析工作流程. 它可以让药物实验中的医师只需要鼠标点击就能完成所有的分析工作, 从而大大减轻了他们的工作压力. 它是开源的, 也是轻量级的, 特别适合中小规模的药物研发项目. 相对于商业软件的高门槛及个性化需求响应缓慢等问题, safetyGraphics 轻便灵活, 定制与上手的时间都很短, 成本低, 是 R 领域开源生态中必不可少的一个部分.

会议主办方

中国人民大学统计学院
中国人民大学应用统计科学研究中心
统计之都

会议承办方

中国人民大学统计学院 数据科学与大数据统计系

会议赞助商

Posit