中国R会
The China-R Conference

2020 会议手册

第13届中国R会（北京）

The 13th 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：2020.12.19-2020.12.20

欢迎辞

经过十三年的磨砺，中国 R 语言会议又踏上了新的征程。每当这个时候，各位志同道合的朋友以 R 为相聚的理由，从数据科学的各类学术领域而来、从大数据的各种应用行业而来、从天南海北的各条奋斗战线而来，欢聚一堂，共襄盛举。这是 R 的独特魅力。R 的一个核心设计理念是“人的时间永远比机器的时间宝贵”，具有深厚的人文精神，其工程化应用又秉承了“总是有多种方法来做同一件事”的思想，极具包容性。它专注于数据科学和统计建模，保持自己的勃勃生机，又主动和其他的优秀工具融合，让大数据时代的舞台群芳竞艳。这也正如统计学，最大的好处是“可以在所有学科的后院玩耍”。参加会议的朋友们都热爱 R，但不执着 R，甚至不用 R，大有“圣人不凝滞于物”的境界。

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。

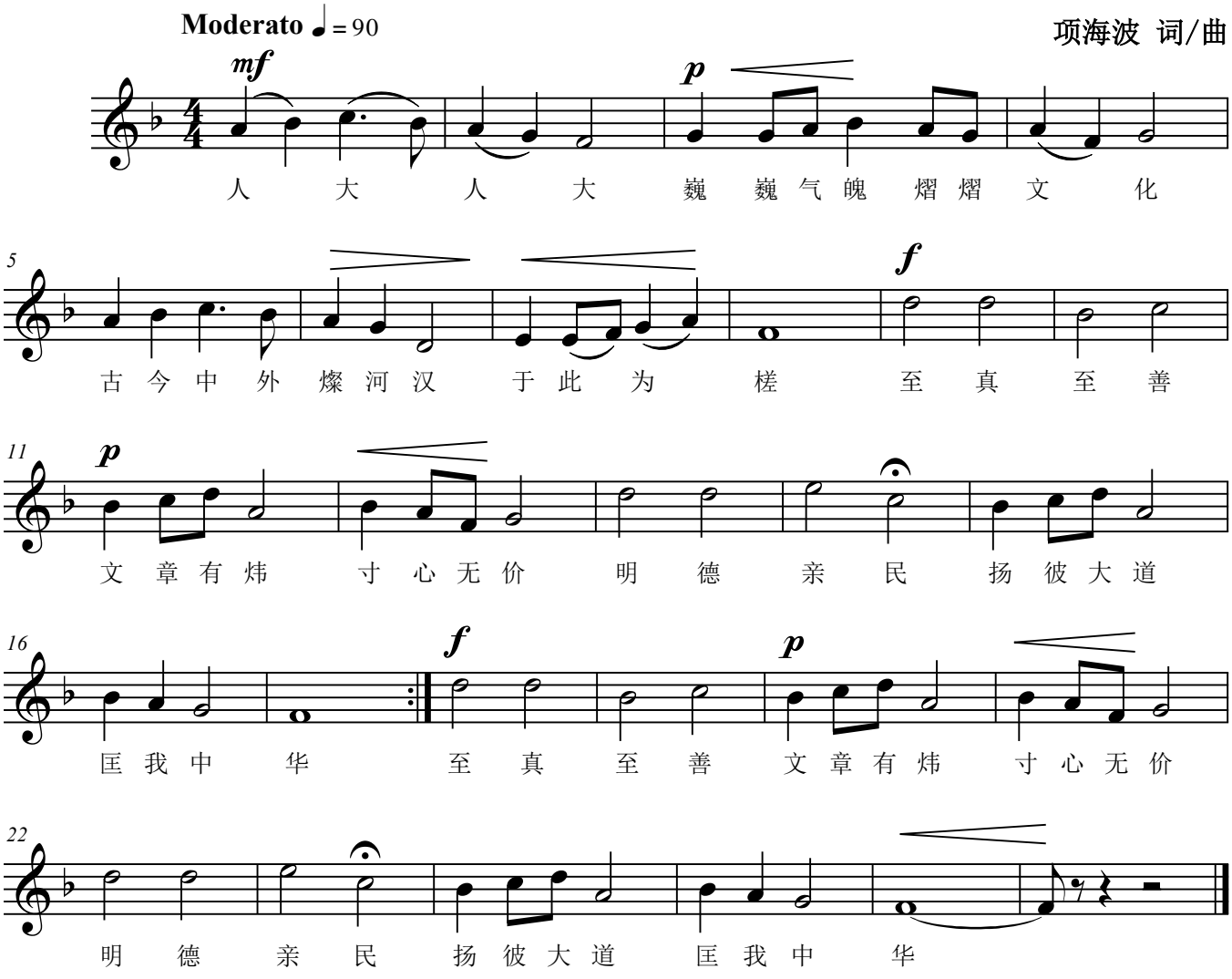
统计之都敬上

2020 年 11 月 30 日

人 之 大 者

为中国人民大学而作

项海波 词/曲



目录

第十三届中国 R 会介绍

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

截止目前，R 会已经在中国人民大学、北京大学、清华大学、华东师范大学、上海财经大学、中山大学、西安欧亚学院、厦门大学、江西财经大学、浙江财经大学、杭州师范大学、中南财经政法大学、湖北经济学院、西南财经大学、贵州大学、兰州财经大学、中国科学技术大学等多个高校举办。2019 年，第十二届中国 R 会议在北京、上海、哈尔滨分别举办，其中北京会场吸引了来自全国各地的 1300 余位参会者，在两天的会议中，各界人士汇聚一堂，进行思维的碰撞。今年将迎来第十三届中国 R 会。

本届 R 会由统计之都、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主办，由中国人民大学统计学院数据科学与大数据统计系承办，将于 12 月 19-20 日在北京举办。此次主题包含但不限于：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，我们欢迎您的到来！

为保障疫情期间会议参与者的身体健康，减少人员聚集及流动，本次会议将结合线上与线下会议形式，各会场已邀请到来自各个领域研究统计学与数据科学的高水平演讲嘉宾。中国人民大学在校师生可以选择线下或线上的形式参会，线下会场人数原则上不超过 50 人。校外参会者可线上参加会议。

主办方

统计之都

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

中国人民大学统计学院

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

诚聘英才

为满足学院事业发展的需要，中国人民大学统计学院现面向校外公开招聘教师/师资博士后：

招聘方向

经济与社会统计、风险管理与精算、概率论与数理统计、生物统计与流行病学、数据科学等方向。

招聘条件

1. 学风端正、治学严谨、道德高尚、为人师表、身心健康，热爱教育事业，遵守职业道德规范，具有强烈的事业心和协作精神。

2. 具有较为扎实的基础理论功底，较强的教学科研能力和良好的专业素养，并能履行相应岗位职责。

3. 具有博士学位。

岗位类别及待遇

提供在北京地区具有竞争力的住房、薪酬条件，保障子女享受国内一流教育条件。

报名方式

如您有意加盟中国人民大学统计学院，请您随时将您个人简历发送至 tongjihr@ruc.edu.cn，我们将通过电子邮件与您联系。

联系方式

中国人民大学统计学院办公室
张老师　陈老师　+86-10-62511318

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
Center for Applied Statistics of Renmin University of China

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。研究中心积极培育中青年学术骨干，逐渐发展并形成了经济与社会统计、统计调查与数据分析、风险管理与精算、生物卫生统计、数据科学与大数据统计，五个各具特色的研究方向。中心建设的重点研究平台是：1．重大发展问题的统计技术创新研究。2．现代统计技术与方法的应用性研究。3．精算技术的创新与应用。4．生物医学统计技术发展与应用。5．数据科学理论研究与大数据技术应用。研究中心拥有国内一流的研究人员，承担多项国家及教育部项目，获得丰硕的研究成果。应用统计科学研究中心，始终将建立和发展应用统计学科基地作为战略定位，着重从制定应用统计研究的科学规划、密切联系实际选准科研攻关方向、注重研究工作的长期积累、加强重点研究平台建设等方面开展工作。

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。

诚聘英才

中国人民大学统计学院数据科学与大数据统计系现面向校外公开招聘教师/师资博士后：

招聘方向

数据科学方向。

报名方式

如您有意加盟数据科学与大数据统计系，请您随时将您个人简历发送至 tongjihr@ruc.edu.cn，我们将通过电子邮件与您联系。

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

第十三届中国 R 会筹备委员会

主　席：任焱
副主席：操懿、向悦
秘书长：李璇
秘书团：任怡萌、王祎帆、苏锦华、孔令仁、刘中渊、陈卉、王小宁

统计之都简介及活动回顾

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

中国统计学门户网站，免费统计学服务平台

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

1. 中国 R 会：目前已开展到第十三届，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、上海每月举办，详情参见统计之都主站及微信公众号。
3. 海外在线视频沙龙：我们在 Google Hangouts 举办在线沙龙，主要由海外嘉宾来分享学术、生活中的点点滴滴。目前已经举办 23 期：http://meetup.cos.name/.
4. 书籍出版，包括写作和翻译。如《Dynamic Documents with R and knitr》(2nd edition) 谢益辉著，《Implementing Reproducible Research》谢益辉等著，《bookdown: Authoring Books and Technical Documents with R Markdown》谢益辉著，《数据科学中的 R 语言》李舰、肖凯著，《R 语言实战》高涛、肖楠、陈钢翻译,《ggplot2：数据分析与图形艺术》统计之都翻译，《R 语言核心技术手册》刘思喆、李舰、陈钢、邓一硕翻译，《R 语言编程艺术》陈堰平、邱怡轩、潘岚锋等翻译，《R 数据可视化手册》肖楠、邓一硕、魏太云翻译，《R 语言统计入门》邓一硕、郝智恒、何通翻译，《数据科学实战》冯凌秉、王群锋翻译,《R 语言实战》(第 2 版) 王小宁、刘撷芯、黄俊文翻译，《Rcpp: R 与 C++ 的无缝结合》寇强、张哗翻译，《R 绘图系统》呼思乐、张哗、蔡俊翻译，《R 语言编程实战》冯凌秉翻译，《量化投资与 R》（待出版）邓一硕、冯凌秉、杨环翻译，《金融风险建模与投资组合优化》（待出版）邓一硕、郑志勇等翻译、《ggplot2: 数据分析与图形艺术（第 2 版）》黄俊文、王小宁、于嘉傲、冯璟烁，《统计之美：人工智能时代的科学思维》李舰，海恩著等等。

专场日程

	线下会场	线上会场
12 月 19 日上午	数据科学专场 明德主楼 1030 腾讯会议 ID：151 375 534	灾害风险专场 腾讯会议 ID：168 820 803
12 月 19 日下午	应用实践专场 明德主楼 1030 腾讯会议 ID：956 896 432	机器学习专场 腾讯会议 ID：725 674 446
12 月 20 日上午	医疗健康专场 明德主楼 1030 腾讯会议 ID：957 482 185	统计软件专场 腾讯会议 ID：781 464 767

	嘉宾姓名	演讲题目	时间
线下会场			
数据科学专场 12 月 19 日上午 召集人：吕晓玲		致辞	8:50-9:00
	王菲菲	Distributed One-Step Upgraded Estimation for Non-Uniformly and Non-Randomly Distributed Data	9:00-9:30
	周静	Progressive Principle Component Analysis for Compressing Deep Convolutional Neural Networks	9:30-10:00
	吴奔	Scalar-on-Image Neural Networks with the Soft-Thresholded Gaussian Process Prior	10:00-10:30
		自由讨论、休息	10:30-10:40
	白琰冰	基于改进 BASNet 卷积神经网络的卫星影像洪水信息提取研究	10:40-11:10
	张源源	体育行业数字化实践	11:10-11:40
应用实践专场 12 月 19 日下午 召集人：邱怡轩	贺诗源	数据科学在天文学中的应用	14:00-14:30
	王祎帆	浅谈数据科学在国际金融中的应用	14:30-15:00
	李翛然	统计与 AI 在工业互联网的应用案例	15:00-15:30
		自由讨论、休息	15:30-15:40
	董峰池	机器学习在推荐系统中的应用	15:40-16:10
	肖一凡	自动机器学习的技术现状	16:10-16:40
医疗健康专场 **12 月 20 日上午** **召集人：任怡萌**	李舰	一个医疗数据建模和模拟平台的设计与实现	9:00-9:30
	王钒	Integrative Functional Linear Model for Genome-wide Association Studies with Multiple Traits	9:30-10:00
	李昂	低浓度臭氧暴露对老年人糖稳态影响的定组研究	10:00-10:30
		自由讨论、休息	10:30-10:40
	刘泓	Introduction to Transfer Learning: Theory and Algorithms	
	李璇	微博抑郁症患者的识别与分类	11:10-11:40
线上会场			
灾害风险专场 12 月 19 日上午 召集人：苏锦华		致辞（与线下会场同步）	8:50-9:00
	白琰冰	基于卫星影像和深度学习的巨灾房屋损毁评估研究	9:00-9:30
	唐辉	Debris Flow Hazard Assessment and Early-warning System Based on Machine Learning and Processes-based Model	9:30-10:00
	李政宵	Generalizing the Log-moyal Distribution and Regression Models for Heavy-tailed Loss Data	10:00-10:30
	熊政辉	中国地震巨灾模型的构建和行业应用	10:30-11:00
	杨熙	基于 MODIS GPP 产品的冬小麦保险费率厘定方法研究	11:00-11:30
机器学习专场 12 月 19 日下午 召集人：常象宇	宗福季	面向数字化转型的工业数据分析科研与教育	14:00-14:30
	杨涛	强化学习在电商场景的红包投放应用	14:30-15:00
	俞声	从电子病历到知识图谱	15:00-15:30
	鲁伟	深度学习语义分割理论与实战指南	15:30-16:00
	王梦佳	HR 数据智能	16:00-16:30
统计软件专场 12 月 20 日上午 召集人：谢益辉	谭显英	你只需要 library(data.table)	9:00-9:30
	苏玮	访问总量 1600 万+的疫情数据可视化应用的开发故事	9:30-10:00
	黄湘云	可重复性数据分析及其工业实践	10:00-10:30
	覃文锋	学习 R 的方法和 R Weekly	10:30-11:00
	宋骁	在 Kaggle 上分享你的数据分析工作	11:00-11:30

Distributed One-Step Upgraded Estimation for Non-Uniformly and Non-Randomly Distributed Data

王菲菲（中国人民大学）

时间：09:00-9:30

简介：王菲菲，中国人民大学统计学院助理教授，北京大学光华管理学院统计学博士。研究上关注文本挖掘及其商业应用、大数据建模、空间统计学、社交网络分析等，在 Journal of Econometrics、Journal of Business and Economic Statistics、Statistics in Medicine 等期刊上均有发表。

摘要：One-shot-type (or divide-and-conquer) estimators have been widely used for distributed statistical analysis. However, their outstanding statistical efficiency hinges on two critical conditions. The first is the uniformity condition, which requires that the sample sizes allocated to different Workers should be as comparable as possible. The second one is the randomness condition, which requires that the data should be distributed across Workers as randomly as possible. Considering that both conditions are often violated in practice, we prove both theoretically and empirically in this work that the violation of either condition can seriously degrade the statistical efficiency of one-shot estimators, or even make them inconsistent. To fix this problem, we propose a novel one-step upgraded pilot (OSUP) method. In the first step of the algorithm, a pilot estimate is computed based on randomly selected samples from different Workers. In the second step, we conduct one-step updating based on the pilot estimate by summarizing the derivative information on each Worker. We show theoretically that the resulting OSUP estimator can be as statistically efficient as the whole sample maximum likelihood estimator without any restrictive assumption about distribution uniformity and randomness. Extensive numerical studies are presented to demonstrate the finite sample performance of the OSUP estimator. Finally, by way of an illustration, an American Airlines dataset is analyzed on a Spark cluster.

Progressive Principle Component Analysis for Compressing Deep Convolutional Neural Networks

周静（中国人民大学）

时间：9:30-10:00

简介：周静，中国人民大学统计学院副教授、应用统计科学研究中心研究员，北京大学光华管理学院博士，研究方向为社交网络、空间计量、模型压缩等，在 Journal of Business & Economic Statistics，Statistic Sinica，Science China Mathematics，Electronic Commerce Research and Applications，管理科学，营销科学学报等国内外核心期刊发表论文十余篇，编著《深度学习：从入门到精通》教材一

本，主持国自科、北社科、统计局重点等多项省部级以上课题。担任人民邮电出版社数据科学与统计·商业分析系列教材编委会委员。

摘要：In this work, we propose a progressive principal component analysis (PPCA) method for compressing deep convolutional neural networks. The proposed method starts with a prespecified layer and progressively moves on to the final output layer. For each target layer, PPCA conducts kernel principal component analysis for the estimated kernel weights. This leads to a significant reduction in the number of kernels in the current layer. As a consequence, the channels used for the next layer are also reduced substantially. This is because the number of kernels used in the current layer determines the number of channels for the next layer. For convenience, we refer to this as a progressive effect. As a consequence, the entire model structure can be substantially compressed, and both the number of parameters and the inference costs can be substantially reduced. Meanwhile, the prediction accuracy remains very competitive with respect to that of the baseline model. The effectiveness of the proposed method is evaluated on a number of classical CNNs (AlexNet, VGGNet, ResNet and MobileNet) and benchmark datasets

Scalar-on-Image Neural Networks with the Soft-Thresholded Gaussian Process Prior

吴奔（中国人民大学）

时间：10:00-10:30

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、脑图像数据分析、金融高频数据分析等。

摘要：Deep neural networks have been adopted in the scalar-on-image regression which predicts the outcome variable using image predictors. However, training DNN often requires a large sample size to achieve a good prediction accuracy and the model fitting results can be difficult to interpret. In this work, we construct a novel single-layer Bayesian neural network(BNN) with spatially-varying coefficients (SVC) for the scalar-on-image regression. Our goal is to select interpretable image features and to achieve the high prediction accuracy with limited training samples. We assign the soft-thresholded Gaussian process (STGP) prior to the SVCs and develop an efficient posterior computation algorithm based on stochastic gradient Langevin Dynamics (SGLD). The BNN-STGP provides a large prior support for sparse, piecewise-smooth and continuous SVCs, enabling efficient posterior inference on image feature selection and automatically determining the network structures. We establish the posterior consistency of estimating the SVCs in the model and image feature selection consistency when the number of voxels/pixels grows much faster than the sample size. We compared our methods with state-of-the-art deep learning methods via extensive simulations and analyses of multiple real datasets including the task fMRI data from the ABCD study.

基于改进 BASNet 卷积神经网络的卫星影像洪水信息提取研究

白琰冰（中国人民大学）

时间：10:40-11:10

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。主要研究领域包括机器学习、深度学习、大数据分布式计算、卫星遥感、空间数据科学等。在 IEEE，Remote Sensing 等国际期刊发表多篇论文；主持美国 Microsoft 公司人工智能基金，参与日本学术振兴会 JSPS 基金项目；承担中国人民大学大数据硕士专业课程《大数据分布式计算》和本科生《统计学》的授课工作。担任北京大数据协会理事会理事。

摘要：高精度洪水信息提取能对灾后救助和恢复提供重要帮助，随着洪涝灾害在全球发生频率的提高和规模的增大，自动化的高精度洪水区域提取在气候变化背景下扮演着越发重要的角色。本研究主要基于 Sen1Floods11 多源卫星遥感影像数据集，利用卷积神经网络进行自动化洪水洪水区域提取。由于数据集存在严重的正负样本不平衡问题，即图像中大部分区域为负样本的陆地，水区域仅占极小比例，导致模型容易在负样本上表现较好，正样本预测效果较差。同时，由于数据量较少 (机器标注 4160 张图像，人工标注 446 张图像)，模型容易出现过拟合。因此，我们在 BASNet 模型的基础上，一方面，利用 focal loss 缓解样本不平衡问题，另一方面，加入随机翻转、随机旋转、通道标准化等数据增强方法，增强模型泛化能力。模型目前在 Sen1Flood11 数据集上，对地表水的预测效果超过已有模型，在 Test 数据集上 mIoU 达到 0.4097，BoliviaTest 数据上 mIoU 达到 0.4046。

体育行业数字化实践

张源源（北京百姓科服网络科技有限公司）

时间：11:10-11:40

简介：张源源，百姓科服算法总监，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。在此之前，曾在百度获得百度年度新人，百度最佳团队成员等荣誉，在乐动力期间，负责计步、跑步、运动识别等工作，是国内第一家入选 Appstore 年度精选应用的 App，对 GPS、IMU 数据有较多研究，有多项相关专利。

摘要：伴随着移动互联网的高速发展，点赞微信运动排行榜、朋友圈晒跑步轨迹、打卡马拉松赛事、吃完火锅做 Keep 等等已经成为我们很多人的日常行为，这些行为也见证了体育行业数字化的高速发展。

本次分享将以体育行业从业者视角，通过对体育行业数字化的背景、进展、未来的分析、陈述、展望，回顾移动互联网背景下的体育行业数字化整个进程，并通过走路、跑步、健身、球类运动等几个垂直类别的数字化方案剖析，分享一些 sensor、camera 数据的机器学习实践经验。

数据科学在天文学中的应用

贺诗源（中国人民大学）

时间：14:00-14:30

简介：贺诗源博士于 2017 年获 Texas A&M University 统计学博士学位，现任中国人民大学统计与大数据研究院助理教授，研究方向包括函数数据分析、统计计算、天文统计。

摘要：现代大规模巡天观测为数据科学提供了新的机遇与挑战。这些问题的系统解决，需要天文学、天体物理、宇宙学、计算机科学、统计学等学科的通力合作。本次报告选取部分角度，介绍数据科学如何在天文学中一展身手。我们的天文望远镜，只能观测到遥远恒星、宇宙深处传来的光线。但我们如何仅凭这些光线测量宇宙尺度、测量当前的宇宙膨胀速度、推测宇宙的过去与未来？如何从宇宙早期残留的影像（微波背景辐射）确定宇宙常数？大质量物质扭曲了我们观测的星体，我们如何检测引力透镜？地球这样的系外行星并不会发光（无法被望远观测），我们如何找到它们？如何知道三体世界中的比邻星附近，确实有行星存在？

浅谈数据科学在国际金融中的应用

王祎帆（中国人民大学）

时间：14:30-15:00

简介：王祎帆，中国人民大学统计学院经济统计学博士在读，研究方向为宏观经济、国际金融等，研究成果发表或即将发表在《数量经济技术经济研究》以及《国际金融研究》等期刊，曾参与《R Graphics Cookbook》的翻译工作，本科期间曾获国家奖学金以及优秀本科毕业论文一等奖。

摘要：随着信息时代的发展，全球化进程的加深，各类数据充斥在国际金融市场中，任何一点风吹草动都可能引发一场金融风暴。数据科学在其中可以发挥什么作用？我们应该如何利用这些数据？应该如何正确地利用这些数据？又应该如何从数据中发现机遇、发现风险，挖掘全球化下涌动的暗流？本次报告将对上述问题展开简要探讨，欢迎大家参会讨论。

统计与 AI 在工业互联网的应用案例

李翛然（深圳奇点信息技术有限公司、医道国际集团）

时间：15:00-15:30

简介：李翛然，就职于深圳奇点信息技术有限公司、医道国际集团，研究方向为大数据、AI、工业互联网。

摘要：近两年国家在大力推进工业互联网，而其核心定义、标准以及应用方式也正在被各行各业进行大力的探索。我们在一个工业互联网的项目中，通过整合 AI 技术与统计算法，使项目的业务效率得到了质的飞跃。本次演讲将会从实践案例出发，介绍其中的解决思路探索、技术演变、架构选型、测试、商业模式的落地，以及如何给客户带来真正的价值增长，向听众展示一个真实的工业互联网项目的实施过程。

机器学习在推荐系统中的应用

董峰池（北京字节跳动科技有限公司）

时间：15:40-16:10

简介：董峰池，18 年毕业于中国人民大学统计学院，后入职字节跳动做推荐算法工程师，目前所在团队负责西瓜视频推荐算法。工作之余运营着个人公众号“峰池”，致力于帮助在校同学更好更快的成长为一名推荐算法工程师。

摘要：推荐系统是当前机器学习算法应用比较成熟的一个分支，在抖音、今日头条等 App 上都取得了广泛的应用。可以说，工业界在推荐系统的实践上已经有一套较为成熟的方法论，而这一套理论在学校可能不易了解到。本次报告主要跟大家讲讲机器学习在工业界推荐系统的实践，介绍工业界推荐系统的整体框架，帮助大家了解手机 App 是怎么样从千万候选中找到你最感兴趣的那些视频。

自动机器学习的技术现状

肖一凡（华为（北京）研究院）

时间：16:10-16:40

简介：肖一凡，2017 年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。先工作于华为手机产品线，负责华为手机语音助手的研发与交付任务。后在华为北京研究院，负责自动机器学习算法的研究与商业落地。研究方向主要集中在自然语言处理、自动机器学习算法，同时对算法的工程交付、线上运营有丰富经验。

摘要：近十年来，机器学习技术的发展推动了巨大的商业变革，也涌现出了一大批以人工智能技术为核心竞争力的创业公司。然而设计一个可靠的机器学习算法，需要投入巨大的人力物力，甚至其复杂程度已经超越了人力所能处理的范围。自动机器学习，就是随着机器学习发展到成熟阶段后，诞生出来的一种新兴技术。自动机器学习不仅可以大幅减轻算法设计带来的人力负担，而且也创造了一系列超越人类设计水平的精度记录。那么自动机器学习究竟是不是万能的？它能做哪些事情、擅长哪些事情、又不能做哪些事情呢？本报告会从自动机器学习的起源讲起，循序渐进地介绍当前自动机器学习的主流技术，即使非本领域的听众也可以理解其主要原理。同时结合本人在工业界数年的开发与商业交付经验，向大家介绍商业实践中更倾向于使用什么样的算法，以及如何跨过算法到工程交付之间的鸿沟。

一个医疗数据建模和模拟平台的设计与实现

李舰（九峰医疗）

时间：9:00-9:30

简介：李舰，九峰医疗首席数据科学家，“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，著有《统计之美》《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：医疗大数据的分析和建模是当前健康领域的热点问题，在真实的场景中，不同医疗机构的信息系统种类繁多、非常复杂，如何将各个系统的数据整合到一起进行分析是很多应用成功的关键。此外，如果可以在数据还未清洗完成时通过蒙特卡洛方法进行模拟，可以加快科研的进度和促进同行的交流。本次报告介绍了一个基于医院信息系统数据的科研平台的设计和实现方式，并分享开发中的相关经验。

Integrative Functional Linear Model for Genome-wide Association Studies with Multiple Traits

王钒（中国人民大学）

时间：9:30-10:00

简介：王钒，中国人民大学统计学博士生、美国密歇根大学硕士，8 年临床医学、流行病学统计建模与数据分析经验，4 年制药业生物统计分析经验。研究兴趣为高维数据分析，函数型数据分析，临床试验与研究。论文发表在 Biostatistics, Statistical Methods in Medical Research, JAMA Internal Medicine 等期刊。

摘要：In recent biomedical research, genome-wide association studies (GWAS) have demonstrated great success in investigating the genetic architecture of human diseases. For many complex diseases, multiple correlated traits have been collected. However, most of the existing GWAS are still limited because they analyze each trait separately without considering their correlations and suffer from a lack of sufficient information. Moreover, the high dimensionality of single nucleotide polymorphism (SNP) data still poses tremendous challenges to statistical methods, in both theoretical and practical aspects. In this article, we innovatively propose an integrative functional linear model for GWAS with multiple traits. This study is the first to approximate SNPs as functional objects in a joint model of multiple traits with penaliza- tion techniques. It effectively accommodates the high dimensionality of SNPs and correlations among multiple traits to facilitate information borrowing. Our extensive simulation studies demonstrate the sat- isfactory performance of the proposed method in the identification and estimation of disease-associated genetic variants, compared to four alternatives. The analysis of type

2 diabetes data leads to biologically meaningful findings with good prediction accuracy and selection stability.

低浓度臭氧暴露对老年人糖稳态影响的定组研究

李昂（中国医学科学院、北京协和医学院）

时间：10:00-10:30

简介：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博士研究生，研究方向为环境流行病学和分子流行病学。博士期间的主要研究内容为空气污染暴露对老年人群的健康效应研究，研究论文发表在 Environment International, Science of the total environment 等杂志。

摘要：我国面临着非常严重的空气污染问题，自 2013 年国家开始实施《大气污染防治行动计划》以来，颗粒物浓度逐年递减，但臭氧浓度不降反升。已有研究发现，臭氧暴露与心血管疾病死亡存在关联性，且可通过氧化应激损伤和炎症反应的作用机制诱导心血管疾病的发生。但是，既往研究多集中在较高浓度的臭氧浓度暴露且研究人群多为一般人群。本研究旨在探讨在低浓度臭氧浓度暴露下，对非糖尿病老年人群糖稳态相关标志物的影响。本研究为明确易感人群，以及国家修订臭氧浓度限值提供相关研究证据。

Introduction to Transfer Learning: Theory and Algorithms

刘泓（清华大学）

时间：10:40-11:10

简介：刘泓，清华大学电子系本科生，研究方向为迁移学习，曾在 ICML, NeurIPS, CVPR 发表论文。

摘要：Transfer learning—transferring knowledge learned from a large-scale source dataset to a small target dataset—is an important paradigm in machine learning with wide applications. Models equipped with transferability learn efficiently from their past experience, lowering computation and energy cost, and are robust to changes in underlying distributions, enabling fairness amongst different ethnic groups and genders. I will start with mainstream algorithms of transfer learning in practice, and then focus on the recent advances in their mechanism and theoretical insights.

微博抑郁症患者的识别与分类

李璇（中国人民大学）

时间：11:10-11:40

简介：李璇，中国人民大学统计学院本科大四在读学生，研究生将就读于清华大学万科公共卫生与健康学院。感兴趣的研究内容为生物统计、文本分析和健康大数据，本次报告相关内容已被《统计研究》编辑部等组织的“第十八次全国中青年统计科学研讨会”收录。

摘要：抑郁症一直以来是社会关注的焦点问题，在我国，抑郁症患者的数量庞大而就诊率低。随着互联网的发展，近年来，线上社交网络成为患者表达抑郁情绪的重要途径之一，这对于抑郁症患者的早期识别提供了一种新的可能。本研究通过对社交平台“微博”用户的文本信息提取，利用卷积神经网络模型，实现了对微博文本的抑郁分类。同时通过与其他三种机器学习算法的对比，验证了卷积神经网络模型的有效性。最后，通过聚类方法对抑郁文本的特征进行了分析。研究结果表明，基于卷积神经网络模型可以更有效地对微博文本的抑郁情况进行分类，而抑郁症患者在微博的表达也具有明显的类别特征。因此，本研究可以为情绪障碍类疾病患者的精准干预与治疗工作、抑郁症患者语料库的构建工作等献力，从而使抑郁症患者受到更多的群体社会关怀。

基于卫星影像和深度学习的巨灾房屋损毁评估研究

白琰冰（中国人民大学）

时间：9:00-9:30

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。主要研究领域包括深度学习、大数据分布式计算、卫星遥感、巨灾损毁评估等。在 IEEE，Remote Sensing 等国际期刊发表多篇论文；主持美国 Microsoft 公司人工智能基金，担任北京大数据协会理事会理事。

摘要：近年来随着中国城乡居民住宅地震、台风和洪水巨灾保险制度的实施和产品落地，中国未来在量化房屋资产及损失智能保险领域有了巨大的需求。本研究依托来自世界范围的 85 万个受灾房屋相关的卫星影像研发了基于深度学习算法的针对房屋损失评估量化分析云平台，该平台提供云端交互式房屋损毁量化分析服务，未来可以提供给保险公司用于风险量化和理赔服务。

Debris Flow Hazard Assessment and Early-warning System Based on Machine Learning and Processes-based Model

唐辉（德国地球科学中心 *German Research Centre for Geosciences (GFZ)*）

时间：9:30-10:00

简介：唐辉，德国亥姆霍兹地球科学研究中心 (GFZ) 研究员。主要从事机器学习，深度学习以及数据同化在地球科学领域的应用与研究。主持多项美国自然科学基金 (NSF)，德国科学基金会 (DFG)，以及中德合作中心合作项目 (DFG-NSFC)，目前为地表过程与自然灾害研究组组长。在 Journal of Geophysical Research, Geophysical Research Letter, Earth Science Review 等国际顶级期刊发表十余篇论文。多次担任 Nature, Nature Geoscience, Journal of Geophysical Research, Geophysical Research Letter 以及德国科学基金会，美国自然科学基金和瑞士国家科学基金会 (SNSF) 审稿人。

摘要：Debris flows threaten life and infrastructure in areas close to steep mountain fronts. Rainfall intensity-duration (ID) thresholds are commonly used to assess the likelihood of debris flows in distinct regions with similar climate, soils, and topography. Currently employed ID thresholds are empirical and developed with historical data, and therefore most applicable to those settings where debris flows have been recorded in the past. We propose a method that combines process-based numerical modeling and machine learning to derive thresholds for runoff-generated debris flows in a variety of settings to build an early warning system. By using a support vector machine method, we train logistic regression functions using a combination of monitoring data and model results. Our training dataset includes post-wildfire debris flows in the San Gabriel Mountains, California, USA, runoff-generated debris flows in Chalk Cliffs, Colorado, USA, and runoff events in the Venetian Dolomites, Italy. Our proposed approach is based on rainfall thresholds that can be used in these areas with no historical data on

runoff-generated debris flow occurrence. This result is consistent with previously derived rainfall ID thresholds for post-fire debris flows in southern California. Data from two other dolomitic sites in the northeastern Italian Alps (Acquabona and Cancia) provide data for further independent testing of our thresholds using at least four previously observed debris flow events.

Generalizing the Log-moyal Distribution and Regression Models for Heavy-tailed Loss Data

李政宵（对外经济贸易大学）

时间：10:00-10:30

简介：李政宵，毕业于中国人民大学统计学院风险管理与精算学专业，获经济学博士学位，现为对外经济贸易大学保险学院统计与精算系副教授，主要研究领域为精算模型，巨灾风险管理和相依风险度量。近年来在《ASTIN Bulletin: The Journal of the IAA》、《统计研究》、《数理统计与管理》、《系统工程理论与实践》、《保险研究》等国内外核心期刊发表多篇学术论文，主持国家自然科学基金青年项目一项《复杂数据结构下的巨灾保险定价模型及其应用》，先后参与国家社会科学基金重大项目、国家自然科学基金一般项目、教育部人文社会科学重点研究基地重大项目等多项课题研究。

摘要：Catastrophic loss data are known to be heavy-tailed. Practitioners then need models that are able to capture both tail and modal parts of claim data. To this purpose, a new parametric family of loss distributions is proposed as a gamma mixture of the generalized log-Moyal distribution from Bhati and Ravi (2018), termed the generalized log-Moyal gamma (GLMGA) distribution. While the GLMGA distribution is a special case of the GB2 distribution, we show that this simpler model is effective in regression modeling of large and modal loss data. Regression modeling and applications to risk measurement are illustrated using a detailed analysis of a Chinese earthquake loss data set, comparing with the results of competing models from the literature. To this end, we discuss the probabilistic characteristics of the GLMGA and statistical estimation of the parameters through maximum likelihood. Further illustrations of the applicability of the new class of distributions are provided with the fire claim data set reported in Cummins et al. (1990) and a Norwegian fire losses data set discussed recently in Bhati and Ravi (2018).

中国地震巨灾模型的构建和行业应用

熊政辉（中再巨灾风险管理股份有限公司）

时间：10:40-11:10

简介：熊政辉，中国再保险（集团）股份有限公司博士后，博士毕业于中国地震局地球物理研究所，长期从事地震巨灾模型研究和应用工作，先后参与多个国家重点研发计划重点专项，目前申请、获批巨灾领域发明专利共 7 个，软件著作权 4 项。

摘要：破坏性地震是典型的低频率、高损失、难预测的“黑天鹅事件”。国内外的巨灾保险实践表明，基于物理机制和计算机技术的地震巨灾模型，可以更科学合理地对地震风险进行损失评估和风险定价。基于中再集团近些年在巨灾保险领域的实践探索和研究成果，本报告将介绍地震巨灾模型的构建技术方法和平台特点优势，以及模型在（再）保险行业的应用实践。并阐述中再集团的巨灾风险管理理念：以巨灾模型为契机，聚合各方资源，打造开放合作的巨灾风险管理“新生态”，通过巨灾风险管理服务国家治理。

基于 MODIS GPP 产品的冬小麦保险费率厘定方法研究

杨熙（中国太平洋财产保险股份有限公司）

时间：11:10-11:40

简介：杨熙，中国人民大学地理信息系统专业硕士，研究方向为空间统计与遥感风险评估，目前就职于中国太平洋财产保险公司，主要负责农险产品开发工作。

摘要：农作物保险是国内外减少灾害造成的种植户经济损失，保障农民基本生产收入的重要手段。国内传统的农作物保险费率是基于行政单元的统计数据厘定的，忽略了行政单元内部灾害的空间风险差异，因此如何获得行政单元内部农户级农作物纯保险费率，成为精细化农作物保险的关键问题。针对农户级的冬小麦纯保险费率，以河南省周口市为实验区，利用 MODIS GPP 总初级生产力数据产品生成历年冬小麦生长季的 GPP 数据，同时利用 Landsat 数据计算历年公里级的冬小麦种植面积比。通过信度模型模型和经验费率法厘定得到实验区基于格网单元的冬小麦纯保险费率。研究表明：遥感数据可以为农作物保险空间精细费率厘定提供数据保障，利用遥感数据可以得到公里级格网单元的冬小麦纯保险费率。将利用遥感数据得到的农作物纯保险费率用于农作物保险中，提高了农作物保险的空间精细水平，可以进行基于地块的空间差异化农户投保，有利于推进农业保险精细化发展。

面向数字化转型的工业数据分析科研与教育

宗福季（香港科技大学）

时间：14:00-14:30

简介：宗福季教授现任香港科技大学讲座教授, 广州校区信息枢纽署理院長，工业工程与决策分析系前系主任, 及质量与大数據分析实验室主任，国际质量科学院 (IAQ) 院士，美国统计学会 (ASA) 会士, 美国工业工程师学会 (IISE) 会士，美国质量学会 (ASQ) 会士，国际统计协会 (ISI) 當選会员，香港工程师学会 (HKIE) 会士。宗教授在质量控制及工業大数據领域有着广泛而深入的研究，并积极推动有关质量大数据的研究及教育工作。宗教授是美國质量学会旗艦期刊 Journal of Quality Technology (JQT) 的前主編, 工业工程学会期刊 IISE Transactions 及 Technometrics 的副编辑。宗教授于密歇根大学获工业工程硕士及博士学位。

摘要：This talk will present and discuss the challenges and opportunities that data science and analytics face in the era of digital transformation, and the roles we play to drive such transformation. In particular, there is a big opportunity for industrial and business analytics, under the digital transformation paradigm, in order to further explore ways of creating value from data and big data. On research: I will update the recent progress in our Quality and Data Analytics Lab on change detection in heterogeneous data streams. On education: I will share the recent development of HKUST 2.0: a cross-disciplinary paradigm and a unique Information Hub in the Greater Bay Area.

强化学习在电商场景的红包投放应用

杨涛（阿里健康科技（中国）有限公司）

时间：14:30-15:00

简介：杨涛，阿里健康引擎算法初创团队成员，现就职于阿里巴巴创新业务事业群，负责阿里健康电商导购、流量营销分发、个性化推荐等模型建设，致力于提升商业产品变现能力和用户产品体验。在商品导购、搜索推荐、权益营销方面有丰富经验，深耕业务特点和生活场景，从医药服务共性需求出发，探索强化学习、迁移学习、可解释性推荐、异构建模等模型算法研究，为超过一百家医药自营商家的数字化营销提供一站式智能解决方案。

摘要：对于权益投放，一直是用户增长领域中一个老生常谈的话题，在各大互联网公司、各大业务场景中都屡见不鲜。但是，在淘宝这个强购物心智的场景中，解决给什么样的人发什么样的权益能达到什么样的效果，这都是一个很有挑战的问题。尤其是对于商家而言，如何充分利用手头的现金流资金进行权益促达，有效促进流量的优质转化，进行店铺级别的拉新与留存，这都是目前运营和业务同学的迫切痛点。目前阿里健康的业务场景中，用户来访的意图也呈现出越来越多元化的趋势，如何准确地筛选目标人群进行干预，给什么样的人投放多少钱的红包，以此来提高转化率，是权益发放面临的难题。考虑到最大化增量 ROI 的多目标的红包面额推荐问题，我们通过强化学习的思路去着力理解和解决目前营销活动中的权益投放问题。

从电子病历到知识图谱

俞声（清华大学）

时间：15:00-15:30

简介：俞声，博士，清华大学统计学研究中心副教授，清华大学数据科学研究院 RONG 教授，长期从事医学自然语言处理技术与电子病历分析技术研究。俞声独立开发的电子病历自然语言处理系统被美国哈佛医学院、麻省总医院、退伍军人医学中心等顶尖医学研究机构使用，至今已分析电子病历数十亿篇次。俞声发明的高通量表型提取技术使 i2b2 疾病表型识别算法开发速度从每年 1-2 个提高到每年超过 1000 个，并应用于 Veteran Affairs “Million Veteran Program”等美国国家级精准医学研究项目；该系列论文获评医学信息学顶刊 Journal of the American Medical Informatics Association 的编辑选择奖、国际医学信息学学会 2019 年年鉴最佳论文奖，并按标准化生物医学实验方法发表于 Nature Protocols。归国后，俞声带领团队围绕中文电子病历和智能诊疗发展了高通量知识图谱构建、无监督中文医学术语发现、医学机器翻译等一系列技术。

摘要：以知识图谱中的庞大知识量使计算机更加智能是后深度学习时代人工智能发展的一个重要目标。对于医学知识图谱的构建，电子病历中所蕴涵的丰富信息不能忽视，具有巨大的挖掘价值。本报告介绍利用电子病历提取知识图谱信息的几个近期成果：1、利用一种无监督多粒度分词技术实现医学术语发现，用以构造图谱的节点。2、一种知识决定的术语嵌入技术，用于术语正则化，将同义术语聚合。3、一种高通量关系提取技术，通过远监督方法自动生成百万级训练样本，并结合知识型文章和电子病历信息进行联合高精度关系提取。

深度学习语义分割理论与实战指南

鲁伟（杭州脉流科技有限公司）

时间：15:30-16:00

简介：鲁伟，贝叶斯统计方向硕士毕业。目前是一家医疗科技公司深度学习算法工程师，主要研究方向为医学图像处理和深度学习应用。著有《深度学习笔记》一书，公众号机器学习实验室主理人。

摘要：图像分类、目标检测和图像分割是基于深度学习的计算机视觉三大核心任务。三大任务之间明显存在着一种递进的层级关系，图像分类聚焦于整张图像，目标检测定位于图像具体区域，而图像分割则是细化到每一个像素。基于深度学习的图像分割具体包括语义分割、实例分割和全景分割。语义分割的目的是要给每个像素赋予一个语义标签。语义分割在自动驾驶、场景解析、卫星遥感图像和医学影像等领域都有着广泛的应用前景。本文作为基于 PyTorch 的语义分割技术指南，对语义分割的基本技术框架、主要网络模型和技术方法提供一个实战性指导和参考。

HR 数据智能

王梦佳（阿里巴巴）

时间：16:00-16:30

简介：王梦佳，阿里巴巴城市大脑创始团队成员，阿里云数据中台算法负责人，负责数十家企业的数字化转型及智能升级项目，涉及多行业从 0 到 1 的产业智能方案搭建，尤其在交通物流，新零售等领域。现就职于阿里巴巴企业智能事业部，立足阿里生态全域大数据，应用大数据、机器学习、统计分析和数据可视化等技术，和业务创新发展相结合，探索和发展基于数据智能的企业信息智能之路。

摘要：本报告将介绍数据智能和 AI 应用于人力资源管理（资源规划、招聘配置、培训发展、绩效管理）的探索与研究。

你只需要 library(data.table)

谭显英（中意资产管理有限责任公司）

时间：09:00-09:30

简介：谭显英，CFA，南开大学精算学硕士，现任职中意资产高级量化经理。他是 data.table 的项目成员，DT 包的共同作者。他曾与 R 中的字符编码问题进行过多次战斗，并取得了些许成绩。

摘要：data.table 是一个语法简洁、功能强大、运算高效的 R 包，能够覆盖数据处理工作中的大部分场景。本演讲将会通过一系列的实例，向大家分享 data.table 在数据处理方面的独特优势。

访问总量 1600 万 + 的疫情数据可视化应用的开发故事

苏玮（*Yahoo! Japan* 前端工程师）

时间：09:30-10:00

简介：2016 年从华中科技大学生物信息学专业毕业后，进入东京大学应用生命工学专业就读研究生，主要研究基于 R 语言的差异基因分析工具的开发。研究生期间自学 shiny 并独立完成差异基因分析工具 TCC-GUI 后进入日本雅虎，从事雅虎广告平台投放系统的前端开发工作。新冠疫情爆发后，就持续追踪日本疫情变化，再次利用 shiny 对其进行数据可视化，最终开发并创建了访问量破千万的 shiny 应用《新冠疫情速报》。

由于在《疫情速报》开发中对疫情数据的迅速且丰富的可视化上所做出的出色成果，在 3 月份时被日本厚生劳动省聚集性感染对策专家组邀请协助进行疫情数据可视化工作。本次演讲的内容也是以《疫情速报》的开发故事，结合一个从学术界到互联网前端开发的从业人员的经验，帮助 shiny 应用开发初学者减少一些可能会走的弯路。

摘要：在年初的新冠肺炎爆发之后，以霍普金斯大学疫情数据仪表盘应用为首，互联网上出现了大量用 JavaScript、Python、R 等各类语言开发的同类应用。数据分析师、程序员、各种传媒等以同一个题材进行数据可视化的热情空前高涨。作为一个生活在日本的海外游子和数据可视化的爱好者，《新冠肺炎疫情速报》在日本确诊患者仅有 9 名时立项开发，所有功能经过数次迭代后，仅靠口口相传的形式在 10 个月时间里获得了 1600 多万浏览量。其翔实的数据和通俗易懂的可视化甚至被日本专家学者列为参考网站之一。本次报告将面向 shiny 应用开发初学者，通过分享《新冠速报》开发背后的故事，谈谈开发中的坑和相应的处理办法，从而帮助新手在开发自己应用的过程中少走一些弯路。

可重复性数据分析及其工业实践

黄湘云（北京三快在线科技有限公司）

时间：10:00-10:30

简介：黄湘云，中国矿业大学（北京）统计学硕士，统计之都副主编，美团 AI 平台数据研发工程师。多次混迹 R 会蹭吃蹭喝，开源了多本笔记，和谢益辉、赵鹏合著的书《现代统计图形》即将出版。

摘要：文学编程自提出以来，不断有新的尝试去实现高德纳先生的想法，从 noweb、Sweave 到如今的 R Markdown、Jupyter Notebook，一路走来，新的技术和理念推动了数据分析的可重复性，提升了团队的协作效率，数据可重复、问题可重复、过程可重复、研究可重复、经验可沉淀，工具和技术必将赋能业务决策和平台建设。

学习 R 的方法和 R Weekly

覃文锋（*R Weekly*）

时间：10:30-11:00

简介：覃文锋，毕业于厦门大学公共卫生学院，王亚南经济研究院。R Weekly 团队创始人。活跃于 R 开源社区，开发维护了多个热门的 R 开源项目。目前在从事机器学习的有关工作，个人感兴趣的领域包括机器学习、优化理论等。

摘要：向初学者介绍学习 R 语言的方法，以及一些推荐的资源。介绍 R Weekly 项目的构成和如何参与到这个开放的团队。R Weekly 搭建了一个一站式的信息平台，通过网站，邮件，播客等方式，向来自世界多个国家的读者推送 R 社区的最新动态。每周的资讯速递帮助 R 用户快速地掌握社区一周内的最新进展，帮助用户发现，学习和使用现有的基础资源，掌握社区内的最佳实践。

在 Kaggle 上分享你的数据分析工作

宋骁（泛为科技）

时间：11:00-11:30

简介：宋骁，毕业于华东师范大学。目前在上海泛为科技担任数据分析师，负责互联网广告投放系统的 BI 报表开发工作。业余时间混迹于 Kaggle，Cos 等网站。

摘要：Kaggle 是谷歌旗下的集竞赛、代码分享、论坛于一体数据科学网站。本报告介绍 Kaggle 入门的使用技巧，如数据分享、代码版本控制、竞赛提交等功能。其次介绍 Kaggle 对 RMarkdown 等可重复性工具的支持，鼓励大家在 Kaggle 上提交公开代码。最后介绍了 Kaggle 提供的计算资源的使用方法以及它在数据科学教学及研究中能起到的作用。

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

RStudio