第十二届中国 R 会议（上海）

自由的统计语言

主办:

华东师范大学

统计之都

战略合作伙伴:

云筏科技

赞助:

图灵教育

西安交通大学出版社

2019 年 12 月 21 日–22 日

R 语言简介

R 是一个有着统计分析功能及强大作图功能的语言环境和软件系统，由新西兰奥克兰大学统计系的 Ross Ihaka 和 Robert Gentleman 共同创立。R 语言可以看作是由 AT&T 贝尔实验室所创的 S 语言发展出的一种方言。

R 是在 GNU 协议 General Public Licence 下免费发行的，它的开发及维护现在则由 R 开发核心小组 R Development Core Team 具体负责，这个团队的成员大部分来自大学机构（统计及相关院系），包括牛津大学、华盛顿大学、威斯康星大学、爱荷华大学、奥克兰大学等。除了这些作者之外，R 还拥有一大批贡献者（来自哈佛大学、加州大学洛杉矶分校、麻省理工大学等），他们为 R 编写代码、修正程序缺陷和撰写文档。

R 的功能很大程度上是通过程序包（Package）来实现的，迄今为止，R 语言官网上的程序包数目已经超过 7000 个，广泛地覆盖了数据分析应用到的各类行业和领域。各种统计前沿理论方法的相应计算机程序都会在短时间内以软件包的形式得以实现，这种速度是其它统计软件无法比拟的。

在 KDnuggets 于 2015 年 7 月做的“首选何种编程语言进行分析、数据挖掘及数据科学的工作”的调查中，R 以 51％的得票率荣登榜首，力压 Python、SAS 和 MATLAB（http://www.kdnuggets.com/polls/2015/r-vs-python.html），连续 4 年位居榜首。在 2015 年 5 月的另一项调查“首选何种数据分析、数据挖掘、数据科学软件或工具”中，R 超过了 2014 年的冠军 RapidMiner，同样位列第一。

目前，几乎所有的西方大学与研究机构、以及越来越多的金融机构、制药公司、高科技企业都使用 R。R 的灵活性、开放性以及业界最广泛的支持是其不断完善和发展的根本原因，随着 R 越来越被学术界及业界认可，它也将在数据分析和统计建模中发挥越来越大的作用。

华东师范大学统计学院简介

华东师范大学是中国获批最早设立概率论与数理统计专业的大学之一。1986 年，概率论与数理统计获原国家教委批准设立博士点，1987 年成为全国高等学校重点学科，是全国最早的三个概率论与数理统计国家重点学科之一。

学院拥有统计学博士后流动站、统计学一级学科博士点、2 个专业学位硕士点（应用统计硕士、保险硕士）、4 个本科专业（统计学、经济统计学、金融工程、保险学）。另外，统计学院参与创办并自 1985 年起一直承办中国概率统计学会期刊《应用概率统计》。2017 年与中国现场统计学会合作创建了统计学学术期刊《Statistical Theory and Related Fields》。

自 2012 年以来统计学院在 Annals of Statistics, Journal of the American Statistical Association, Biometrika, Annals of Probability, Insurance Mathematics & Economics 等国际顶尖 SCI 期刊上以第一作者或通讯作者共发表论文近 200 余篇。在数据科学的顶尖会议及期刊上发表论文 30 多篇。2017 年，成功加入北美精算师协会（SOA）的 UCAP（Universities & Colleges with Actuarial Programs）高校计划。2017 年，华东师范大学统计学院获批成立了统计与数据科学前沿理论及应用教育部重点实验室。同年，华东师范大学统计学在教育部学科评估中位列全国第三位（并列），并入选教育部“双一流”学科建设行列。2018 年，统计学被纳入上海高校高峰高原学科建设（II 类高峰）。

华东师范大学教育信息技术学系

华东师范大学教育信息技术学系于 1978 年成立，是全国创建最早、研究信息技术在教育中应用的文理交叉学科。建系近三十年来，经过师生共同努力,在学科建设方面达到了国内领先的水平,同时也树立了以计算机教育应用为特色的专业品牌。

教师主要研究聚焦于教育信息化理论与系统规划、学习科学与技术设计、数字媒体与数字出版研究、教育信息化装备、环境及技术标准研究以及计算机支持教育测量、评价及管理等领域。目前承担多项国家级课题并积极参与国际合作交流，与美国、英国、荷兰、日本、新加坡等国，以及香港、台湾等地区高等院校有长期的合作、交流关系。

教育科学和技术飞速发展的今天，教育信息技术学系努力寻找新的突破点和增长点。如先后与苏州工业园区、广州 TCL 集团共建了全国最早的两个教育技术学博士后科研工作站。又如面向全国进行教育技术应用（示范性）实验区建设，首批实验区建设已在江苏省、浙江省、山东省等所属的几个县、市级区域全面展开，协助这些地区教育信息化的整体推进。还如正在积极修订本科生和研究生的教学计划，并对专业实验室进行全面的改造。这些表明，华东师大教育信息技术学系正在展现它新的风貌。

统计之都简介[1]

“统计之都”（Capital of Statistics，简称 COS）网站成立于 2006 年 5 月，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门户网站；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。

想要获取最新的咨询和活动的通知，请关注我们的微信[2]：

[1] 统计之都网址：http://cosx.org/

[2] 微信号：CapStat

云筏科技简介

云筏科技是一家拥有丰富经验的云计算解决方案提供商，我们致力于为用户提供优质和高性价比的定制化云服务，通过降低数据分析的学习和使用门槛，助力分析人员降本增效。公司主要产品包括多语言在线环境和云筏容器云平台，用户覆了国内顶尖高校及科研院所和各大教学平台。

图灵教育简介

北京图灵文化发展有限公司，始终以策划出版高质量的科技图书为核心业务。旗下图灵教育品牌是国内计算机图书领域的高端品牌之一。图灵社区是图灵公司打造的综合性服务平台，集图书内容生产、视频课程学习、作译者服务、电子书销售、技术人士交流于一体。

西安交通大学出版社简介

西安交通大学出版社 1983 年由文化部批准成立，1998 年西安交通大学出版社与西安交通大学音像出版社合并，成为具有图书、音像、电子、互联网出版权的综合性大学出版社。

2009 年我社被原国家新闻出版总署评为一级出版社，并获“全国百佳图书出版单位”荣誉称号，是西北地区唯一一家国家一级出版社。

西安交通大学出版社有限责任公司下设 15 个部门，其中编辑部门有理工分社、医学分社、青少分社、经管事业部、人文社科事业部。建社 36 年来，共出版新书 8000 余种，音像制品 1000 余种，电子出版物 300 余种。出版物中有 300 余种获得国家、省、部级奖励。

注重国际版权贸易工作，与施普林格公司、爱思唯尔公司等国际著名出版机构签署战略合作协议，共同打造高水平的英文版教材和学术专著。今后，将继续坚持“为教学科研服务”的办社宗旨，为我国出版事业的发展做出贡献。

第十二届中国 R 会议（上海）会议指南

1. 日程安排

2019/12/21 上午	注册和主会场报告	华东师范大学中山北路校区，科学会堂二楼报告厅
2019/12/21 下午	工业应用专场报告	
2019/12/22 上午	深度计算专场报告	
2019/12/22 下午	R 语言与数据思维专场报告	

2. 会议议程

12 月 21 日：

时间	内容	嘉宾	主持人
09:00-09:15	开场致辞		张翔
09:15-09:45	统计学习与机器学习的比较	汤银才	
09:45-10:15	大数据可视分析的现状与展望	成生辉	
10:15-10:30	休息		
10:30-11:00	知识图谱在本地生活新零售、餐饮场景下的构建和应用	李哲	张翔
11:00-11:30	当 R 牵手容器云	汤怡玮	
11:30-12:00	基于龙芯的数据科学工作环境	李舰	
午餐			
工业应用专场			
13:45-14:15	基于 R 的工业边缘 AI 应用-车体焊接质量智能预测 APP	刘心广	李舰
14:15-14:45	人工智能在医疗领域的商业化应用	李翛然	
14:45-15:15	现代数据科学标准：净土之路	朱俊辉	
15:15-15:30	休息		
15:30-16:00	人力资本场景下的 AI 实践之路	刘洋	李舰
16:00-16:30	基于 R 可靠性分析系统的构建	周世荣	
16:30-17:00	利用 AI 进行基于非结构化数据的高级搜索	赵明杰	

12 月 22 日：

深度计算专场			
时间	内容	嘉宾	主持人
09:15-09:45	在云上运行机器学习工作负载	高策	张翔
09:45-10:15	如何做设备端智能化	张先轶	
10:15-10:30	休息		
10:30-11:00	SQLFlow 在机器学习和数据运营上的应用	高朋	张翔
11:00-11:30	深度学习中三步加速梯度算法应用	练勇强	
11:30-12:00	产品经理眼中能灵活管理响应延迟的计算系统	张翔	
午餐			
R 语言与数据思维专场			
13:45-14:15	R 语言数据操纵之美	黄天元	徐浩
14:15-14:45	VSCode vs. RStudio	任坤	
14:45-15:00	智能电梯：故障预测建模分析	米汶权	
15:00-15:30	休息		
15:30-16:00	面向数据的思维模式和 R 语言数据项目开发	张丹	徐浩
16:00-16:30	如何让机器学习模型更有可解释性	陈堰平	

3. 会议机构

主办单位：

华东师范大学

统计学院

教育信息技术学系

统计之都（http://cosx.org/）

战略合作伙伴：

云筏科技

赞助单位：

图灵教育
西安交通大学出版社

组委会成员：

张翔，汤银才，李舰，徐浩，曾加

统计之都支持：

冯凌秉，任怡萌，向悦，任焱，孙腾飞

统计学习与机器学习的比较

汤银才[1]

华东师范大学统计学院

摘要

由深度学习推动的人工智能在图像与自然语言处理等领域的成功应用，给传统的统计学与计算机等学科带来了具体挑战，特别是对数据分析长期依赖的统计学科产生质疑，甚至否定。正如华为 CEO 任正非所指出的人工智能的发展离不数学与统计学(“国家若要强盛，数学是基础”；“人工智能就是统计学，我们要高度重视统计学”)。大数据分析离不开统计学，同样离不开计算机学科，二者协同完成海量数据的分析与应用。但是，在高等学校如何落到实处，尽快解决教什么、学什么和做什么样的创新性研究？这是我们师生共同关注的问题，也是企业数据从业人员需要高度重视的问题。

本报告将通过对当前一些人工智能相关话题的跟踪与大家一起探讨当前统计学习与机器学习的一些基本问题，内容包括:

1) 从统计学到数据科学
2) 统计建模与机器学习差异
3) 对深度学习的质疑
4) 贝叶斯深度学习

[1] 汤银才，华东师范大学教授，博士生导师，上海数萃大数据学院院长，觉云科技首席科学家；《应用概率统计》、《华东师范大学学报》和《Statsitical Theory and Related Fields》编委，中国数学会概率统计学会理事、中国运筹学会可靠性分会常务理事、中国现场统计研究会大数据分会分会常务理事兼副秘书长、中国现场统计研究会计算统计分会理事、上海市工业与应用数学会理事，已培养硕士毕业生 50 多名，博士毕业生 11 名，主持并完成国家自然科学基金 3 项，其他各类项目 20 多项，在国内外学术刊物上发表论文 100 多篇。先后获得上海市科学技术三等奖(2017)，上海市教育发展基金会申银万国奖(2009)，华东师范大学优秀任课教师奖（2008），上海市教学成果三等奖(1995)，上海市科技进步三等奖(2017,1996)，全国统计科学技术进步二等奖(1996)，上海市统计科学研究成果课题类一等奖(1995)等荣誉。著有《R 语言与统计分析》、《可靠性统计》、《贝叶斯统计》、《基于 WinBUGS 贝叶斯建模》(翻译, 2020)。

大数据可视分析的现状与展望

成生辉[1]

香港中文大学

摘要

大数据时代，数据的量不断上升，数据的复杂度也越来越高。挖掘并理解内部的信息是个挑战。可视分析主要将数据通过图形化的方式展示出来，更加直观的认识数据，并通过交互，来分析有价值的信息。此报告首先介绍一下大数据可视分析的产生与发展现状，之后介绍一下常见的可视化分析平台以及分析案例，最后对数据可视化的未来做一些分析和展望。

[1] 成生辉，大数据时代，数据的量不断上升，数据的复杂度也越来越高。挖掘并理解内部的信息是个挑战。可视分析主要将数据通过图形化的方式展示出来，更加直观的认识数据，并通过交互，来分析有价值的信息。此报告首先介绍一下大数据可视分析的产生与发展现状，之后介绍一下常见的可视化分析平台以及分析案例，最后对数据可视化的未来做一些分析和展望。

知识图谱在本地生活新零售、餐饮场景下的构建和应用

李哲[1]

饿了么

摘要

知识图谱在本地生活新零售、餐饮场景下的构建和应用简介：饿了么有数亿的用户、数百万商家和数亿的商品。知识图谱如何将这些数据链接起来，智能化的助力饿了么新零售、餐饮的发展？本次将详叙饿了么在知识图谱的构建与落地应用的详细历程，并分享知识图谱对饿了么新零售、餐饮的赋能成果。

[1] 李哲，饿了么资深数据专家，管理科学与工程博士，负责饿了么标签画像系统、数据挖掘以及营销算法，在机器学习赋能智能营销的领域有丰富的经验。之前先后任职于 Google、eBay。

当 R 牵手容器云

汤怡玮[1]

云筏科技

摘要

介绍容器云和微服务在科研和数据分析领域的前景；利用两个场景实例介绍如何将 R 与容器云结合，其中包括：1.使用容器云 paas 部署高可用的 rshiny 应用集群；2.使用 docker 实现应用框架和业务服务分离。

[1]汤怡玮，云筏科技联合创始人&CTO，负责云筏科技技术架构的整体开发和维护，拥有多年大数据分析及高通量组学分析实战经验。

基于龙芯的数据科学工作环境

李舰[1]

统计之都

摘要

自主可信计算在新时代下有着重要的意义，基于国产龙芯平台的通用 CPU 以其良好的兼容性和飞速提升的计算性能，在很多关键领域得到了大规模的产业化应用，对维护个人和国家信息安全发挥了重要作用。随着 3A4000 平台的推出，未来必将在个人电脑领域也占据重要的地位。本次演讲以 R 语言为基础，介绍龙芯平台下的常用数据科学工具，可以作为把工作环境迁移到龙芯系统的参考资料。

[1] 李舰，“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：人工智能时代的科学思维》《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。

基于 R 的工业边缘 AI 应用-车体焊接质量智能预测 APP

刘心广[1]

西门子（中国）有限公司

摘要

边缘计算和人工智能是西门子关于未来工业智能制造远景中的核心技术之一，也是工厂自动化向数字化智能化转型升级的重要环节，赋予自动化元件具备感知、执行和反馈的综合能力，形成一个个小的智能单元，组合构建工厂智能体，迎接未来的工业升级需求和挑战，而其中的边缘应用是核心组件。本材料即基于汽车行业白车身焊接质量的业务需求，介绍 R 语言在边缘 AI 项目中的建模，以及 R 语言与 Docker 和 Kubernetes 组合的微服务应用管理，在某上线 APP 中的应用概况。

[1] 刘心广，中国科学院博士，高级工程师，西门子数字化工业集团工业 AI 与边缘计算首席数据科学家，从事工业领域大数据分析、智能算法和应用整合的一体解决方案研究，负责商业模型的算法架构和产品开发，尤其擅长产品质量相关的分析/预测/优化/提升等，现在西门子主导汽车/电池/电子等行业的工业人工智能的深化应用。

人工智能在医疗领域的商业化应用

李翛然[1]

奇点信息技术

摘要

在人工智能的系统建设中，NLP 系统由于其结果判断没有统一标准，产品体验者背景不同，客户预期也处于几乎全部处于概念验证阶段。因此，绝大部分的 NLP 项目都处于实验室“看上去”很美的状态。那么，一家商业性的 AI 公司，如何在其中寻找平衡点，并形成一套范式的商业化模式，变成为其在经济增速放缓的大环境时期的必备技能。本次演讲将会从实践案例出发，穿插其中的技术演变，架构选型，测试以及商业模式的落地，如何给客户带来真正的价值增长，向听众展示一个真实的人工智能（NLP）项目的实施过程。

[1] 李翛然，于利兹大学金融数学毕业。 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，为各大机构提供智能化管理系统与机器人业务。现已有 10 余家金融机构、医院、教育系统机构采用其提供的智能服务为行业助力。其证券类产品 2017 年获得 KPMG 中国金融科技双创大赛 TOP30 奖项。2018 年为全中国（也去也是全球最大）最大的肺癌数据库完成知识结构化及自动化服务。其工业物联网及自动化系统获得 2019 年微众银行区块链应用大赛一等奖。

现代数据科学标准：净土之路

朱俊辉[1]

美团点评

摘要

近年来，随着机器算力的不断提升，人工智能不断打破应用场景的天花板，数据科学工具箱日趋完善，生产效率大幅提升。本次演讲，将着重介绍以 Tidyverse 为中心的现代数据科学标准工作流，通过若干案例演示在大数据、人工智能时代，现代数据科学工作流的新趋势。

[1] 朱俊辉，来自美团点评技术团队，R 语言脑残粉，geosparkR 包开发者，专注于探索智能交通和金融科技领域的 AI 应用。

人力资本场景下的 AI 实践之路

刘洋[1]

e 成科技

摘要

分享 e 成科技在人工智能技术与人力资源领域结合的工作，包括招聘和员工服务场景的对话机器人、咨询访谈机器人，AI 面试机器人以及人力资源行业的人才岗位画像、知识图谱、人岗匹配等。重点探讨如何将不同 AI 技术与人力资本的特定业务场景结合，解决领域内的难题。此外演讲还会探讨 AI 在 toC 和 toB 场景下应用的差异及一些个人体会。

[1] 刘洋，e 成科技 AI 算法负责人，在 e 成科技负责 NLP、对话机器人及机器学习算法研发。具有丰富的开发和管理经验，曾先后就职于腾讯社交网络事业群及阿里巴巴创新业务事业群，负责用户画像、深度学习、搜索排序及问答系统等项目。e 成科技（https://ai.ifchange.com/）作为人力资本数字化平台，是数字化人才战略领先者，开创性地将 AI 技术与人才战略升级场景深度结合，形成数字化招聘、数字化员工服务、数字化人才咨询等支柱产品线，为企业招对人，用好人，助力人才战略成功创造价值。

基于 R 可靠性分析系统的构建

周世荣[1]

华东师范大学

摘要

可靠性分析是产品研发、测试与交付的重要一环，华为作为全球最大的通信设备研发和制造企业，其可靠性工程师每天都要分析大量的寿命数据，进而评估产品的可靠度、试验所需的样品量和试验时间。在美国制裁与限制华为在大前提下，我们与华为各部门的可靠性工程师开始进行密切合作，提出了数据与业务驱动的可靠性生态构建方案，并基于 shiny 为华为打造了一个可靠性分析平台。本次报告主要对该平台的功能进行介绍与演示。

[1] 周世荣，华东师范大学统计学院博士生，上海数萃大数据学院金牌讲师，第十一届中国 R 语言会议(上海会场)组委会成员，R 与 Python 语言重度用户，作为核心成员参与并完成多项可靠性横向课题。研究方向包括可靠性统计、贝叶斯统计、机器学习前沿算法等。

利用 AI 进行基于非结构化数据的高级搜索

赵明杰[1]

微软中国

摘要

利用 Azure 高级搜索服务基于各类非结构化数据，图片表单，建立索引，以实现各类高级搜索，并生成统一的结构化数据与报表。

[1] 赵明杰，微软全渠道事业部 云计算解决方案架构师 负责支持微软全球合作伙伴数字化转型的架构设计 有着丰富的人工智能，机器学习，物联网与 DevOps 经验。

在云上运行机器学习工作负载

高策[1]

杭州才云科技有限公司

摘要

机器学习在近几年逐渐走出了高校实验室，开始在工业界落地。随着大规模的应用，如何更好，更具扩展性地管理和维护机器学习应用，成为了一个热点话题。本次演讲介绍了在 Kubernetes 的帮助下，在数据准备到模型训练再到模型服务上线过程中遇到的问题是如何被一一解决的。其中包括对显卡等硬件加速器资源的统一管理，大规模分布式训练的支持，集群利用率的优化等话题。

[1] 高策，杭州才云科技的机器学习系统工程师。他目前是机器学习基础设施开源项目 Kubeflow 的子项目：自动机器学习系统 Katib，TensorFlow 和 PyTorch Operator 的维护者。他的研究兴趣方向是分布式系统和可伸缩的机器学习场景的基础设施。他曾是 KubeCon China 2018，第 10 届中国 R 语言会议（上海）的讲者。

如何做设备端智能化

张先轶[1]

PerfXLab

摘要

深度学习已经成为当前技术发展主要趋势之一，除了在服务器端训练深度神经网络模型之外，如何高效率的在嵌入式端执行模型推理，也成为了当前业界关注的问题之一。本报告将分享我们在嵌入式人工智能落地实践过程中，成功的经验和踩过的坑。

[1] 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。2016 年，创办 PerfXLab 澎峰科技，提供嵌入式 AI 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（研究集体主要完成者）。

SQLFlow 在机器学习和数据运营上的应用

高朋[1]

和鲸科技

摘要

这次演讲是一个 SQLFLow 的介绍性演讲，SQLFlow 是一个将 SQL 语句转换成机器学习训练和推断的框架，能够减少日常机器学习项目的工程复杂度，也可以帮助运营专家降低机器学习的应用门槛，主要会介绍 SQLFlow 的整体设计和架构，以及我们在日常的数据运营的过程中如何将 SQLFlow 应用到我们的数据工作流当中。

[1] 高朋，科赛网的研发工程师，主要负责数据科学竞赛和数据科学协同平台的研发工作，承办每年全国性的高校大数据挑战赛和全国人工智能大赛，关注数据治理和应用加机器学习的结合以及工作流的整合，对大规模系统软件有长期的实践。

深度学习中三步加速梯度算法应用

练勇强[1]

华东师范大学

摘要

在训练深度学习的模型时，梯度下降算法是一个用得非常广泛的优化方法。然而梯度下降算法收敛速度非常慢，因此许许多多的加速算法被提出用来加快收敛速度。结合二次函数收敛性质，Nesterov 加速梯度算法迭代步骤中的姐妹序列，以及神经网络中的平行切线法，我们提出了三步加速梯度算法，并将三步加速梯度算法融入到深度学习反向传播算法和随机梯度算法中，重写了 R 软件包 neuralnet，命名为 supneuralnet。利用 supneuralnet 包可以计算本报告中所有有关深度学习的算法。最后利用四个案例展示出我们的算法比其它算法更优。

[1] 练勇强，华东师范大学统计学院博士，2016 年国家公派美国普渡大学统计系联合培养博士一年，上海数萃大数据科技有限公司核心成员。关注于机器学习算法设计、贝叶斯计算、随机模拟方面的科学研究，第八届中国 R 语言会议(上海会场)组委会主席，2014 年和 2017 年中国 R 语言会议(上海会场)演讲嘉宾。

产品经理眼中能灵活管理响应延迟的计算系统

张翔[1]

车轮互联

摘要

深度学习和并行计算系统的融合带来了行业应用的大升级。随着模型复杂度的提高，带来了效果的提升，也对基础设施算力有了更高的要求。但大部分应用厂商没有那么扎实的系统架构能力，所以产品经理需要在效果收益、实施成本，时间延迟等要素间平衡。模型效果在前期可以确定，但是延迟和实施成本会随着容量的变化产生不确定性。这里探讨一种基于响应时间的深度模型部署和管理思路：Time based model serving。

[1] 张翔，车轮互联副总裁，10 多年互联网和大数据从业者，创建了艾瑞咨询数据挖掘部门，服务 BAT 和一线零售快消企业，在旅游和汽车行业创业。同时也是学术爱好者，10 多年 COS 老水友，筹建了上海 R 会组织。写了一本电子书《重构区块链》 bcrb.io。

R 语言数据操纵之美

黄天元[1]

复旦大学

摘要

数据框（data.frame）是 R 语言最重要的数据结构之一，它的本质是一张二维表，每行代表一个实例，每列代表一个属性。对二维表进行创建、插入、删除、排序、分组等操作，是数据处理基础中的基础。尽管在基本包中已经能够对数据框进行很多基本操作，但在 R 语言社区里，为了使这些过程变得高效便捷，有众多的开发者进行了探索尝试。其中，最突出的两个包就是 dplyr 和 data.table。前者的语法表达极其简练而符合人类逻辑，给人以美的感受；后者则具备惊人的处理速度，让所有需要处理海量数据的用户无法抗拒。本演讲将会对两者进行一个比较，并对今后的发展趋势进行介绍。

[1] 复旦大学博士在读，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，涉猎内容包括但不限于信息计量、机器学习、数据可视化、应用统计建模、知识图谱等，著有《R 语言数据高效处理指南》一书。知乎专栏：R 语言数据挖掘。

VSCode vs. RStudio

任坤[1]

上海明汯投资

摘要

长期以来，RStudio 作为 R 语言的主要开发环境，为数据科学的实际使用和推广起到了好的作用。而近年来快速兴起的 VSCode 作为微软联合社区共同开发的综合型编辑器，其强大的功能和可扩展性受到了广大开发者的好评。我们是否能基于 VSCode API 以及 Langauge Server Protocal 为 R 语言打造更为强大的开发环境？该演讲分享了我们在 vscode-R 和 langauge server 中取得的最新进展，展示了诸如基于代码静态分析的自动补全、文档提示、代码跳转、定义显示、Session 定义、画图与 WebView 等等强大功能背后的实现，以及未来可以实现功能的展望，让 R 语言的用户和开发者有更多的选择，通过更强大的开发环境大幅提升生产力。

[1] 任坤，上海明汯投资资深投资经理，R 社区的活跃开发者。开发了静态表格可视化 formattable 扩展包，非结构化数据处理 rlist 扩展包，是高性能数据处理包 data.table 的贡献者、R Language Server 主要开发者之一，以及 vscode-R 的贡献者。2016 年出版 Learning R Programming 一书，并被翻译为中文（《R 语言编程指南》）和日语于 2017 年出版。

智能电梯：故障预测建模分析

米汶权[1]

复旦大学

摘要

如今城市快速发展，城市中的数据信息也呈爆炸式增长。电梯作为人们出行的一个重要环节，使得电梯安全运行将是保障城市生活正常运作的一个基本点。本报告将基于上海某智能电梯公司提供的大量数据，从中提取出相关特征，利用多种机器学习算法，对电梯数据进行深入挖掘，为电梯故障进行提前预警。

[1] 米汶权，复旦大学大数据学院应用统计在读研究生。2019 年参与复旦大学-新再灵联合实验室项目：“智能电梯：故障预测建模分析”。

面向数据的思维模式和 R 语言数据项目开发

张丹[1]

青萌数海

摘要

目前很多公司机构已经完成了数据的原始积累，如何让沉睡的数据发挥价值，是急需要功课的难关。本次分享的内容分为 2 个部分：面向数据的思维模式和 R 语言数据项目。 数据项目和软件项目、互联网项目都有非常大的不同，不确定性、跨学科知识点、工程落地，都是影响数据项目成功与失败的重要因素。掌握数据思维，科学的方法论，专业的团队，便利的工具，才能让数据项目走向成功。 数据分析师每天都有大量的数据需要处理，我们会根据业务的要求做各种复杂的报表，包括了聚合、分组、排序、筛选、转置、差分、填充、移动、清洗、回归、分布检验、高数计算 等等。有时为了计算一个业务指标，你的 SQL 怎么写都不会少于 10 行时。用 R 语言可以高效地、优雅地解决数据分析中的问题，统计建模、数据挖掘、原型开发，可以节约大量的时间，让我们专业于建模!

[1] 张丹，青萌数海 CTO，微软 MVP，数据科学家。10 年以上互联网应用架构经验，在 R、Java、NodeJS、大数据、数据挖掘等方面有深厚的积累。精通量化投资交易策略，熟悉中国金融二级市场、交易规则和投研体系。 熟悉数据学科方法论，在外汇、海关、区块链等领域均有落地的尝试。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 。

如何让机器学习模型更有可解释性

陈堰平[1]

微软中国

摘要

机器学习、深度学习往往给人一种黑盒的感觉，也就是它所表现出来的可解释性程度不高或者是很低。模型可解释性方面的研究，在近两年的科研会议上成为关注热点。在工业界中，数据科学或机器学习的主要焦点是解决复杂的现实世界的应用问题，而不是理论上有效地用数据来建立模型。解释模型如何对业务起作用总是会带来一系列挑战。有一些领域的行业，特别是在保险或银行等金融领域，数据科学家通常最终不得不使用更传统的机器学习模型（线性或基于树的），原因是模型可解释性对于企业解释模型所采取的每个决策非常重要。

微软这些年一直致力于研究如何创造出具有可理解性的人工智能，不但在 MIT 开源协议下开源了 InterpretML 软件工具包，在微软的云端 AI 平台 Azure Machine Learning Service 中还提供了众多模型可解释性的包。本演讲将介绍如何利用微软的技术解释模型、训练过程中的可解释性、模型推断时的可解释性。

[1] 陈堰平，微软(中国)有限公司 Data&AI 解决方案架构师，中国青年统计学家协会常务理事，微软认证讲师，2017~2018 年入选微软最有价值专家。在高级分析、数据挖掘、人工智能等领域有十年的经验，为企业级客户提供相关项目的架构设计、咨询培训、测试开发支持等服务，服务过金融、互联网、通信、咨询、航空、医疗等行业的客户。