第八届中国 R 语言会议（上海会场）

自由的统计语言

主办：

华东师范大学经济与管理学部

华东师范大学统计学院

统计之都

赞助：

辰智咨询

华院数据

永洪科技

优酷土豆集团

2015 年 11 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（`http://www.kdnuggets.com/polls/2015/analytics-data-mining-data-science-software-used.html` ）。

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

华东师范大学统计学院统计学一级学科博士点的前身— 概率论与数理统计博士点于1986 年批准建立，1987年被原国家教委设立为全国高等学校重点学科，是全国最早的3个概率论与数理统计国家重点学科之一。借由我国首批欧美留学归国的统计学家魏宗舒先生的努力推动和建设，在茆诗松、何声武、王静龙和郑伟安等教授的带领下，本学科茁壮成长。2011年获批一级学科博士学位授权点；2012年设立统计学博士后科研流动站；2012年教育部学科评估中，统计学位列全国第四位（并列），居88所参与统计学科评估高校中的前5%；2013年统计学被上海市教委列入高校一流学科（A类）建设计划；2013年，统计学“应用统计与理论研究创新引智基地”入选教育部和国家外国专家局的“高等学校学科创新引智计划”（又称“111计划”）；2013 年，由统计学牵头承担的“协同创新现代统计方法与理论”获得上海高校创新能力提升计划竞争性引导项目立项；2014年，本学科“统计应用与理论研究”进入“上海高校重点实验室建设计划”。

目前，本学科已建成由国家千人计划教授带领、学缘和年龄结构合理且国际化程度高的学术梯队，在2012年教育部学科评估中“专家团队情况”位列全国第一。本学科现有校级荣誉教授2人（美国国家科学院院士James Berger教授和美国国家工程院院士Jeff Wu教授），国家千人计划教授3人，紫江讲座教授3人，教育部统计学科评议组成员1人，教育部长江特聘教授1 人，“教育部新世纪优秀人才”1人，上海市曙光人才1人，“上海市优秀学术带头人计划”1 人，“上海市浦江人才计划”3人。目前，本学科专职教师35人，其中教授11人，副教授22 人，9人具有海外高水平大学的博士学位，91%教师拥有海外研修经历。近三年主持国家社会科学基金重大项目1项，国家社会科学基金重点项目1项，省部级重点基础研究项目1项，国家自然科学基金及国家社会科学基金18项，省部级项目23项，横向项目11项；近三年发表SCIE和SSCI论文116篇，其中高水平论文（SCI一区）4篇；获得省部级科研成果奖5项，其中“军队科技进步一等奖”1项；获得上海市教学成果奖1项；获选国家级规划教材2套。

本学科在培养高层次人才方面积累了丰富的经验，已建立起一个较为完善的培养本科、硕士、博士和博士后的培养体系，是我国统计学人才培养和科学研究的重要基地之一。在2012 年教育部学科评估中本学科“授予学位数”位列全国第一，2012年获得“上

海市研究生教育创新计划”和“上海市本科生科研创新实践基地”。毕业生中有许多知名高校教授和科研机构学者，如：国家千人计划教授（郑伟安、邵军、孙东初）、教育部“长江学者奖励计划”讲座教授（陈振庆，北京理工大学数理学院院长）、美国国家过敏和传染病研究所(NIAID) 统计学专家（秦靖，其1994年与别人合作的论文Empirical likelihood and general estimating equations 已经获得超过1000次的引用）、教育部长江特聘教授（王兆军，南开大学统计研究院副院长）和国家杰出青年基金获得者（王启华，中科院系统所）等；100 多名高级精算人才，包括20余位北美、英国和中国正精算师，30余位准精算师。许多毕业生成为社会中坚力量，如：1994届博士庄东辰于2003年获全国五一劳动奖章，1994届硕士皮六一（现为上海证券交易所交易管理部总监）2007年获评首届“上海金融十大杰出青年”，2008 年获全国证券期货监管系统劳动奖章。本学院与辉瑞、诺华、罗氏、默克、强生等多家制药企业联合培养了大量专业人才；与上海市统计局、江苏省昆山市统计局、太仓市统计局等签署全面合作协议，共同培养创新型应用统计人才。

统计之都简介*

“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[1]：

*统计之都网址：http://cos.name/

[1]微信号：CapStat

辰智咨询简介

辰智http://www.chenzhimr.com 是一家专业的大数据分析服务和商业数据应用提供商。公司以上海总部为中心，成立本地化网络，网点覆盖华东、华北、华南、华中、西南、东北。在全国范围内为各类客户提供专业的市场调查、商业数据、研究分析、策略咨询以及渠道大数据的挖掘分析服务。辰智主要面向商超、连锁、卖场、商业地产等有门店经营的企业提供顾客、商圈、品牌、行业的市场调研和咨询服务。

基于规范精准的数据分析以及创新的“互联网+”调研方式，辰智帮助雇主更加准确深入的了解他们的顾客，优化他们的门店网点分布，并开发更多合适的潜在顾客。同时，辰智专业的咨询团队还能帮助雇主进行准确的市场定位，提升品牌价值，优化和改进产品，建立起更加成功的营销和市场沟通策略。

另外，辰智提供各类专业的大数据应用工具。降低如GIS、R语言等大数据技术的使用门槛，让客户自己能够通过简单的操作，随心所欲地展示出详细的可视化分析结果，帮助分析自己想要了解的领域。

如今，凭借多年行业服务经验和对市场的准确理解与洞察，辰智已发展成为了业界首屈一指大数据商业服务机构。

华院数据简介

华院数据成立于2002年，是国内领先的大数据应用服务企业。在大数据分析、挖掘及应用领域深耕十几年，已经帮助电信、电商、零售、金融、制造等领域的客户利用大数据提升效率、降低成本和增加营收。目前华院数据的核心业务为：**产业大数据应用、深度孵化（数据实验室）、数据互联和智能引擎**。为了更高效、快速的参与和支持更多元化的大数据业务，华院正在以更大的决心和力度孵化并扶植一系列与大数据分析相关的垂直行业团队，目前已经成功孵化了数云（为零售企业提供数据化营销解决方案）、数创（O2O移动社会化营销服务平台）、华院分析（运营商数据相关的业务合作运营）、数尊（信用风险的评分服务）、析远（与海尔合资，专注于制造业大数据应用）、塔美（与凤凰文化合资，专注于商业地产大数据应用）、NewA Tech（在美国成立，专注于个性化、智能化数据挖掘工具开发）、数加（专注于在线教育）、小宝在线（专注于供应链服务的互联网金融平台）、集买（专注于商业地产O2O）等等。

创始人宣晓华是美国加州大学伯克利分校数学博士（师从美国科学院院士斯梅尔教授），浙江大学计算数学硕士。宣晓华也是中国工业和应用数学协会理事，上海分会副理事长，互联网金融千人会创始会员。

华院数据拥有：

- 6个区域服务中心，总部设在上海，北京、广州、西安、成都、乌鲁木齐设有分公司/办事处；
- 200多位模型工程师、软件工程师及资深顾问，团队云集了麻省理工学院、北京大学、复旦大学、浙江大学等著名学府的教授、学者，同时还有兼具多年管理、营销及数据挖掘实践经验的专家；
- 600多个不同行业的数据挖掘和分析的咨询项目；
- 100多个经过实际应用检验的数据挖掘模型；
- 30多个开发完善的专项应用软件。

华院定位—以数据分析挖掘为核心，行业垂直化的大数据应用的深度孵化器
华院使命—通过“数据分析”推动“科学决策”和“管理优化”

永洪科技简介

永洪科技是国内领先的数据可视化分析解决方案提供商，专注于为百亿级数据量的大型企业和各个垂直行业的中小企业提供灵活易用的数据分析解决方案。永洪科技的目标是让企业用户实现敏捷的数据化运营，实时洞察业务状况，支持战略决策。

永洪科技的管理团队拥有世界500强企业10年以上从业经历，曾多次获得国际大奖，包括Java One大奖、软件界的奥斯卡大奖JDJ读者奖等。

永洪科技的数据可视化分析软件永洪BI拥有分布式计算、分布式存储、分布式通信、云计算、数据处理、数据展现等多项技术专利。

目前，永洪科技已完成经纬创投的A轮和A+轮融资，业绩保持300%-500%的年增长率，软件版本从1.0迭代到5.5，员工人数从10人增长到近百人。2015年，永洪科技逐步完善了企业内部机制和国内市场的战略布局，除了北京总部以外还成立了上海和深圳分公司。

现在，永洪科技已经拥有大小客户达到100多家，主要涵盖政府、电信业、能源&电力业、金融业、制造业、零售&地产业、咨询&服务业、IT业、互联网等。包括中国移动、中国电信、中信银行、中国风电、艾瑞咨询、宝宝树、百程旅行网、积木盒子、途家网、Admaster、富国基金、北京市食品安全监控和风险评估中心、航天三院、北航等众多知名企业和学术机构都已经与永洪科技建立了战略合作关系。

优酷土豆集团简介

优酷土豆集团（NYSE:YOKU），专注于视频领域，是中国第一大视频应用和网络平台，旗下拥有中国排名第一和第二的视频网站优酷（www.youku.com ）和土豆（www.tudou.com ），以及合一影业。

优酷于2006年12月21日正式上线。一直以来，优酷以“世界都在看”为口号，坚持打造“阳光、真实、主流、有梦想”的品牌形象，现已成为中国互联网领域最具影响力、最受用户喜爱的视频媒体品牌。

土豆于2005年4月15日上线。多年来，土豆通过打造“青春、个性、自主、有趣”的品牌性格，号召“每个人都是生活的导演”，让用户能够充分发挥创造力和互动性，在土豆自由地创作、观看和分享视频节目。

优酷土豆集团拥有庞大的用户群、多元化的内容资源及强大的技术平台优势，为用户群提供最全、最多样的内容，帮助用户多终端、更便捷地观赏高品质视频，充分满足用户日益增长的互动需求及多元化视频体验。

目前，优酷、土豆的应用支持PC、手机、平板电脑、电视等多个终端，兼具UGC、PGC、版权、自制、电影五大内容形态，贯通视频内容制作、播出、发行三大环节，正在打造多屏文化娱乐生态系统，立志于成为全球华人最主要的视频来源，并分享快乐、智慧和感动，成就别人，实现梦想，传递正能量。

第八届中国 R 语言会议（上海会场）

会议指南

1. 日程安排

11 月 21 日	注册和主会场报告	华东师大中山北路校区，大礼堂
11 月 22 日	分会场报告	华东师大中山北路校区，科学会堂报告厅
11 月 22 日	分会场报告	华东师大中山北路校区，逸夫楼一楼报告厅

2. 会议议程

11 月 21 日，大礼堂

时间	内容	嘉宾	主持人
09:00-09:10	领导致辞	张日权	林祯舜
09:10-09:20	主席致辞	练勇强	
09:20-09:55	数据科学家的机遇、成长和创新创业	宣晓华	
09:55-10:30	Libra-an R package as Linearized BRegman Algorithm for High Dimensional Statistics	姚远	
10:30-10:50	**Break**		
10:50-11:25	大数据时代的可视化机遇	陈为	汤银才
11:25-12:00	如何在一个 BI 平台上实现数据准备、探索式分析和深度分析	王桐	
午餐			
14:00-14:30	互联网变现与计算广告	刘鹏	李舰
14:30-15:00	利害数据与关键分析技术	邹庆士	
15:00-15:30	当 R 真的遇到大数据：金融和学生学业质量溯因	谢军	
15:30-15:50	**Break**		
15:50-16:20	秩序的作用：商品陈列整齐是否总是比凌乱好？	叶巍岭	魏太云
16:20-16:50	R + Spark = 大数据时代的 R：SparkR 介绍	孙锐	
16:50-17:20	R 在开放数据的应用	谢宗震	

11 月 22 日，科学会堂报告厅

金融大数据会场			
时间	**内容**	**嘉宾**	**主持人**
09:00-09:30	二级市场、数据、趋势	刘道明	刘钟毓
09:30-10:00	互联网金融产品创新及经营活动中的挑战	邓一硕	
10:00-10:30	大数据反欺诈的实践与应用	张昊	
10:30-10:50	**Break**		
10:50-11:20	当金融工程遇到 R	任坤	李浩
11:20-11:50	影响台股指数涨跌的关键变量之分析：递归分类模型之运用	李孟育	
午餐			
工具及可视化会场			
14:00-14:30	商业大数据时代，GIS 和 R 更配	何宇兵	龚航俊
14:30-15:00	R and Tableau: Smart Meets Fast	胡羡祺	
15:00-15:30	slidify+rCharts+ECharts 制作炫酷 HTML5 报告	严紫丹	
15:30-15:50	**Break**		
15:50-16:20	借助 API 快速搭建自然语言处理平台	邢代涛	王昱栋
16:20-16:50	从用 R 读琅琊榜小说讲讲用 R 读书的一些事	张云雁	
16:50-17:20	数据科学的博客：从 knitr 到 jekyll	郎大为	

11 月 22 日，逸夫楼一楼报告厅

互联网会场			
时间	**内容**	**嘉宾**	**主持人**
09:00-09:30	旅游 O2O 行内数据解析	张翔	胡优
09:30-10:00	里子和面子：R 语言及数据挖掘助力京东推荐系统	熊熹	
10:00-10:30	Growth hacking? App 增长分析新玩法	任万凤	
10:30-10:50	**Break**		
10:50-11:20	当游戏数据遇上 R 语言	谢佳标	练勇强
11:20-11:50	利用历史业务数据实现系统异常的实时监测	唐力	
午餐			
统计与机器学习会场			
14:00-14:30	缺失值处理与 R 语言	冯凌秉	王旭
14:30-15:00	古典概率的一些通用解法	杜传龙	
15:00-15:30	Introduction to Feature Hashing	吴齐轩	
15:30-15:50	**Break**		
15:50-16:20	贝叶斯动态线性模型的商业化应用	陈堰平	邹苗苗
16:20-16:50	如何攒一台深度学习服务器	肖凯	
16:50-17:20	旅游数据中的情感分析	毛苏晗	

3. 会议机构

主办单位：

华东师范大学经济与管理学部
华东师范大学统计学院
统计之都（http://cos.name/）

赞助单位：

辰智咨询
华院数据
永洪科技
优酷土豆集团

组委会老师：汤银才　林祯舜　李舰　葛建辉

会议主席：练勇强

副主席：刘钟毓

秘书长：龚航俊

组委会学生：胡优　王昱栋　牛青炎　邹苗苗　李浩　杨丹　王旭　耿晓满

会议详情请参见：http://cos.name/2015/11/2015shanghair

第八届中国 R 语言会议手册，上海， 2015 年 11 月 21 日 – 22 日
主办：华东师范大学统计学院；第 12 页

数据科学家的机遇、成长和创新创业

宣晓华[1,*]

[1] 华院数据

摘要

大数据在近年里受到国家，地方政府，企业，资本等的强劲关注，这也给数据科学家/工程师带来了前所未有的机遇。同时，中国大数据发展现状也非常需要大量数据科学家/工程师的产生和成长，以满足创新和创业的呼唤和落地。本次演讲将围绕这几个方面，阐述演讲者自己的观点，分享自己的经验和事例，介绍创新创业的合作模式。

*宣晓华是大数据技术和应用公司—华院数据技术（上海）有限公司创始人和董事长，也曾参与创办易保网络技术有限公司兼技术顾问；曾在美国加州惠普公司从事七年多的建模/仿真的算法研究和大型软件开发；宣晓华是美国加州大学伯克利分校数学博士，中国工业和应用数学协会理事，上海分会副理事长。

第八届中国 R 语言会议手册，上海， 2015 年 11 月 21 日 – 22 日
主办：华东师范大学统计学院；第 13 页

Libra-an R package as Linearized BRegman Algorithm for High Dimensional Statistics

姚远[1,*]

[1] 北京大学

摘要

In this first part of the two talk series, a new R package, Libra, will be introduced to solve variable selection and sparse recovery via a dynamic approach. In the heart of the package, lies an one-line iterative algorithm, namely Linearized Bregman Iteration which has been widely used in large scale image reconstruction since 2005. Here new application examples in high dimensional statistics will be demonstrated: linear regression, logistic regression, and discrete graphical models with sparsity constraint. The limit dynamics of such an algorithm, exhibits a simple gradient descent flow subject to differential inclusion constraint in favor of sparsity. Understanding such dynamics will provide us a deep insight on how model selection can be done with such a simple algorithm, which will be given in the second talk in the next day. This is a joint work with Stanley Osher (UCLA) and Wotao Yin (UCLA) on theory, as well as two brilliant students on both theory and the R package, Feng Ruan (Stanford and Peking University) and Jiechao Xiong (PKU).

[*] Professor Yuan Yao received his BS (1996, Harbin Inst. of Technology), MS (1998, Harbin Inst. of Technology), M.Phil (2002, City U of Hong Kong), and Ph.D. (2006, UC Berkeley). He did his postdoc at Stanford University during 2006 - 2009, and then he joined Peking University's School of Mathematical Sciences as a professor of statistics in the Hundred Talents Program. His current research interests include topological and geometric methods for high dimensional data analysis, statistical machine learning, as well as their applications in computational biology, computer vision, and information retrieval. He authored about 40 peer reviewed papers and one research monograph. He served as area or session chair in NIPS and ICIAM, as well as a reviewer of Foundation of Computational Mathematics, IEEE Trans. Information Theory, J. Machine Learning Research, and Neural Computation, etc.

第八届中国 R 语言会议手册，上海， 2015 年 11 月 21 日 – 22 日
主办：华东师范大学统计学院；第 14 页

大数据时代的可视化机遇

陈为[1,*]

[1]浙江大学

摘要

理解和利用数据是信息技术发展的迫切需求，数据可视化为人类洞察数据的内涵、理解数据蕴藏的规律提供了重要的手段和高效的人机界面，是和数据分析、数据挖掘等方法的有效补充，在一些重要场合将起到不可替代的作用。本次报告将介绍数据可视化的基本概念以及新媒体时代下的可视分析的内涵，阐述可视化学术界的近期关注重点，并展示面向大规模数值计算模拟、智慧地球、商业智能、数据新闻等应用的可视分析案例，如：空气污染传感器数据可视分析、出租车轨迹数据可视查询、多变量统计数据分布的定量可视分析、城市多维信息可视分析、国家气象局三维大气可视分析原型平台、基于可视化的透明式支持向量机可视化、NBA比赛新闻可视化等。欢迎对数据可视化有兴趣的同学来到现场进行任何形式的交流！

*陈为，1976年生，浙江大学计算机学院CAD&CG国家重点实验室，教授。研究兴趣是数据可视化和可视分析,承担多项国家自然科学基金项目。完成国际一流学术论文50余篇，包括22篇IEEE/ACM Transactions和IEEE Visualization会议论文。出版教材3部，包括2部数据可视化本科和研究生教材。任期刊Journal of Visualization、计算机辅助设计与图形学学报编委、IEEE TVCG和IEEE TITS客座编委。任VINCI国际会议大会主席(2010,2012)，IEEE Pacific Visualization 2013大会论文主席、IEEE Pacific Visualization 2015大会主席和ACM SIGGRAPH Asia Workshop on Visualization 2016大会论文主席。更多信息见：http://www.cad.zju.edu.cn/home/chenwei 和 http://www.cad.zju.edu.cn/home/vagblog

第八届中国 R 语言会议手册，上海， 2015 年 11 月 21 日 – 22 日
主办：华东师范大学统计学院；第 15 页

如何在一个BI平台上实现数据准备、探索式分析和深度分析

王桐[1,*]

[1] 永洪BI

摘要

面对复杂的海量数据，企业运营者该如何挖掘其中的价值？数据化运营如何开始？如何构建体系架构？永洪将分享多年积累的数据化运营最佳实践，让企业里面的每一个人都能轻松发掘大数据的价值，获取深度洞察力。

*王桐，北京永洪商智科技有限公司，副总裁。北京航空航天大学工学硕士，拥有8年商业智能领域的产品销售、市场营销经验，此前效力于甲骨文和IBM，均在咨询、销售岗位担任重要职位，曾成功推进多个大型项目的实施，在电商、政府、金融、互联网等行业积累了丰富经验。王桐目前主要负责产品销售和渠道拓展，已为上百家企业用户提供了完善的数据可视化分析解决方案，这些企业既有宝宝树等电商领域的明星公司，也有中国移动等传统巨头。

第八届中国 R 语言会议手册，上海， 2015 年 11 月 21 日 – 22 日
主办：华东师范大学统计学院；第 16 页

互联网变现与计算广告

刘鹏[1,*]

[1] 奇虎360

摘要

大量的互联网免费产品在获得了流量与数据以后，是如何规模化地创造收入的？在后向变现的过程中，广告的地位和作用如何？用到了什么样的计算技术？本讲座将以新书《计算广告》为背景，深入浅出地介绍计算广告的行业背景、基本问题、常见产品与技术方向，为大家了解互联网的商业模式，更好地从事互联网行业提供有价值的指导。

*刘鹏（@北冥乘海生），互联网商业变现专家，《计算广告》作者。现任360高级总监、商业产品首席架构师。

第八届中国 R 语言会议手册，上海， 2015 年 11 月 21 日 – 22 日
主办：华东师范大学统计学院；第 17 页

利害数据与关键分析技术

邹庆士[1,*]

[1] 中华R软体学会

摘要

Big Data是近年来热门的话题之一，本讲将从Big的新解谈起，以数据敏感度(data sensitive)与数据解析(data analytical)思维为根基，聚焦到关键的数据上，以解析出真正不同且重要的洞见。内容穿插各式数据分析实例，引出关键的解析技术，期能让数据发挥其利害的功用。

*现职：国立台北商业大学(商业技术学院 2004-2014)信息与决策科学研究所教授，台湾数据科学与商业应用协会理事长，中华R 软件学会理事长。经历：国立台北商业大学(商业技术学院 2004-2014)企业管理系副教授，世新大学信息管理学系副教授。中华大学(中华工学院 1996-1997)企业管理学系副教授，交通部运输研究所资料分析项目咨询顾问/授课教授，财团法人中兴工程顾问社数据分析项目咨询顾问，财团法人工业技术研究院资料分析项目咨询顾问/授课教授，中央气象局资料分析项目咨询顾问/授课教授，财团法人信息工业策进会数据分析项目咨询顾问/授课教授，故宫博物院资料分析项目咨询顾问，行政院主计总处数据分析授课教授，台北市计算机商业同业公会数据分析授课教授，财团法人金融研训院资料分析授课教授，新光人寿保险股份有限公司资料分析授课教授，中央警察大学鉴识科学学系资料分析授课教授，中华电信股份有限公司资料分析授课教授，南京理工、中央、中原、东吴、实践、空中等大学讲师/副教授/教授。

第八届中国 R 语言会议手册，上海， 2015 年 11 月 21 日 – 22 日
主办：华东师范大学统计学院；第 18 页

当R真的遇到大数据：金融和学生学业质量溯因

谢军[1,*]

[1] 上海雅捷信息

摘要

大数据该落地了。银行拥有大数据，典型的省级银行拥有5000万客户，9000万账号，其分析基础往往是5000行1000列巨大矩阵的运算。传统技术遇到了巨大挑战。本文报告以GPU为基础的大规模并行技术将数据库查询技术提速至少500倍。

教育传统上被认为是被排除在大数据之外，其实教育质量溯因也需要巨大规模的计算，本文也概要报告了上海闵行的大数据实践。

*本科毕业于复旦大学。1991年牛津大学应用统计博士。28年数据分析和数据挖掘服务经验，是中国电信行业、金融行业和教育行业数据挖掘的开拓者之一，若干大型银行风险和业务模型的创立者。央行科技进步二等奖、银监会科技进步二等奖，农总行科技进步二等奖获得者。目前共总集中在大规模并行计算和人工智能领域。

第八届中国 R 语言会议手册，上海， 2015 年 11 月 21 日 – 22 日
主办：华东师范大学统计学院；第 19 页

秩序的作用：商品陈列整齐是否总是比凌乱好？

叶巍岭[1,*]

[1] 上海财经大学

摘要

商品陈列的秩序是消费者进行商品评价的重要线索，为什么整齐的陈列（相较于凌乱的陈列）会更有利于消费者对商品给出更高的质量预估？我们提出了“画面效应”，即当商品陈列整齐（相对于凌乱）时，消费者对陈列画面的态度更加积极正面，继而导致消费者对产品质量的预估评价也显著更高。其中，消费者对陈列画面的态度在画面效应中起到完全中介作用。我们进一步指出画面效应的的调节变量：产品外观重要性。当消费者认为产品的外观不重要时，画面效应不发生，即凌乱与整齐的陈列不会引起消费者对陈列画面的态度差别，也不会引起消费者对产品质量预估评价的差异；而当产品的外观重要时，画面效应才会发生。演讲人综合已经有的产品污染和消费者污染效应，结合本研究的“画面效应”，以及研究团队正在进行的研究，对商品陈列整齐是不是总是比凌乱要好，作出全面的回答。

*叶巍岭，营销学博士，现任上海财经大学国际工商管理学院市场营销系副教授。上海财经大学“教书育人标兵”，及上海财经大学商学院杰出教学奖获得者。其研究团队的专业兴趣为零售消费者行为及广告效果。研究团队共主持国家自然科学基金2项，完成教育部课题1项，发表论文多篇。研究团队成果获2015年中国营销科学学术年会（JMS）优秀论文，第十四届（2015）“中国市场研究‘宝洁’论文奖”一等奖。

第八届中国 R 语言会议手册，上海， 2015 年 11 月 21 日 – 22 日
主办：华东师范大学统计学院；第 20 页

R+Spark=大数据时代的R：SparkR介绍

孙锐[1,*]

[1]Intel

摘要

Spark 1.4.0版本在Scala, Java和Python语言之外正式引入了R语言API（即SparkR）。SparkR为熟悉R语言的数据科学家提供了一种新工具，使得他们能够基于Spark大数据平台的分布式引擎在R中处理大数据。

本报告将概要地介绍SparkR的背景，历史，架构，API和状态。本报告将帮助R社区熟悉SparkR，并希望能吸引R社区参与到SparkR的讨论和开发中。

*孙锐，英特尔上海大数据团队架构师。HIVE/Shark/Spark贡献者，SparkR主力贡献者之一。

第八届中国 R 语言会议手册，上海， 2015 年 11 月 21 日 – 22 日
主办：华东师范大学统计学院；第 21 页

R在开放数据的应用

谢宗震[1,*]

[1]DSP智库驱动

摘要

开放数据是一个尚未被大量开发的巨大资源。政府、企业、非营利组织为了要完成他们的工作或是任务而收集了大量各式各样不同的数据。有越来越多的案例显示，成功的关键在于“善用数据、跨域合作”，从掌握现况、洞悉趋势、服务创新到政策研究，数据的价值和应用层面相当广泛。这场演讲将跟各位分享透过R语言在政府、企业、非营利组织等开放数据进行加值应用的真实案例。

*清华统计博士，擅长与跨领域专家合作，开发的R包iNEXT被应用于生物、遗传、新闻、文学、电竞等领域。现职为DSP智库驱动知识长，辅导超过300位企业人士成为数据分析人才。

第八届中国 R 语言会议手册，上海， 2015 年 11 月 21 日 – 22 日
主办：华东师范大学统计学院；第 22 页

二级市场、数据、趋势

刘道明[1,*]

[1] 光大云付互联网

摘要

1、互联网数据在股票市场中的应用：情绪指标

2、杠杆交易数据分析：股灾前后的交易客户画像

3、从P2P到互联网财富管理：互联网金融的趋势

*刘道明，光大云付互联网股份公司常务副总裁。多年证券行业研究、证券业务管理经验。历任光大证券研究所金融工程部总经理、光大证券信用业务管理总部总经理助理，2015年中开始参与筹建光大集团旗下互联网金融平台—光大云付。

第八届中国 R 语言会议手册，上海， 2015 年 11 月 21 日 – 22 日
主办：华东师范大学统计学院；第 23 页

互联网金融产品创新及经营活动中的挑战

邓一硕[1,*]

[1]懒投资

摘要

2014年以来，随着互联网金融企业数量的持续增加以及股票市场的繁荣，行业竞争骤然加剧，为了保持竞争力，互联网金融企业必须不断做出产品创新。在此背景下，可灵活存取的活期类投资产品、挂钩股市的结构化产品纷纷面市。

产品创新在为企业带来竞争优势的同时，也为企业带来了运营管理和风险控制上的挑战：像活期类产品就要求企业能够对产品的流动性做出较为精准的预测和管理，以满足投资者的赎回要求，因而，需要企业去不断发掘投资者的申购赎回规律；像债权转让类产品，为了引导投资者理性转让，增加市场流动性，需要对转让价格进行引导，此时需要动态告知投资者项目转让成功的概率。

此外，为了大量获客，企业常常推出力度较大的推广活动，如注册返现、投资返现等，这类活动往往会吸引众多职业羊毛党来薅羊毛，其中不乏伪造信息的“黑羊毛党”，“黑羊毛党”的存在会造成推广成本的剧增，从而降低推广质量，因而，如何甄别“黑羊毛党”也是一个很有意思的挑战。所有这些都需要根据数据和模型进行解决。

*邓一硕，北京大家玩科技有限公司（懒投资）CFO、副总裁，风险控制委员会委员；毕业于中央财经大学统计与数学学院，毕业后曾效力于首钢集团计财部，2014年起加入北京大家玩科技有限公司（懒投资），历任金融项目部总监、财务总监。统计之都理事会理事，曾翻译《R语言核心技术手册》等书籍。

第八届中国 R 语言会议手册，上海， 2015 年 11 月 21 日 – 22 日
主办：华东师范大学统计学院；第 24 页

大数据反欺诈的实践与应用

张昊[1,*]

[1] 同盾科技

摘要

进入“互联网+”时代后，随着大数据技术的不断更新迭代，互联网金融行业已经打破了传统金融寡头的垄断格局，第三方支付、P2P、众筹平台、大数据金融等互联网金融模式层出不穷，O2O、电商、支付行业不断创新。各行业蓬勃发展的同时也出现了各种各样的欺诈问题，如刷单、盗卡、盗号、身份冒用、团伙作案等。本次演讲通过简要介绍同盾科技“跨行业联防联控”的风控实践经验，分析了互联网产品创新所面临的各种欺诈风险问题和特点，介绍了大数据风控技术的反欺诈效果。最后简要探讨了大数据技术在风险控制领域未来的发展方向。

*张昊，同盾科技联合创始人，风控总监。本科毕业于南开大学软件学院，研究生毕业于复旦大学计算机学院。加入同盾科技前，曾在PayPal GRS部门从事支付反欺诈模型相关的工作，包括数据准备、线上/线下模型验证、模型性能监控、变量分析等。加入同盾科技后，主要负责实时风险决策引擎产品的研发、公司产品规划，以及行业风控解决方案等工作。对建模、风险决策分析（如R/Python/SQL等分析工具）有一定实践经验，具备较丰富的软件开发和项目管理经验，长期关注互联网金融、支付、电商、O2O等行业的风险问题。

第八届中国 R 语言会议手册，上海， 2015 年 11 月 21 日 – 22 日
主办：华东师范大学统计学院；第 25 页

当金融工程遇到R

任坤[1,*]

[1]凌云至善量化私募基金

摘要

金融工程是一个高度专业化、科学化的领域，同时也是与实际市场中各种风险管理需求紧密相连的学科。该演讲从金融衍生品的基本研究出发，展示R语言灵活的特性和丰富的扩展包如何在金融数据和量化交易策略的分析、可视化、报告等几乎各个方面满足分析者的实际需求，并讨论了在交易策略的研发和实际执行过程中出现的鸿沟、问题以及相关的思考。

*凌云至善量化私募基金研发合伙人，毕业于厦门大学金融系和王亚南经济研究院，现于深圳从事金融衍生品交易策略研发和平台开发，是pipeR, rlist, formattable 等R扩展包的开发者。

第八届中国 R 语言会议手册，上海， 2015 年 11 月 21 日 – 22 日
主办：华东师范大学统计学院；第 26 页

影响台股指数涨跌的关键变量之分析：递归分类模型之运用

李孟育[1,*]

[1] 台湾嘉义大学财经系

摘要

本文根据R语言的递归分类模型(Model based Recursive Partitioning, MOB)来萃取台股指数涨跌的关键知识规则，并且据此来设计交易策略，并且运用文字探勘方法来找出影响台股的关键变量。递归分类模型乃是结合分类树与回归式的方法，其目的乃是从众多自变量找出分类规则，及其分类节点的回归式。等距或比例尺度的自变量，在分类树可能被递归重复使用，使得关键的自变量会不断地重复呈现于知识规则内。据此，本研究以三大法人(外资、投信、自营商)于现货、期货与选择权交易部位的信息作为自变量。以未来不同天数之累积报酬率应变量。资料期间为2007年7月2日至2013年底，每次以8季为样本内数据、下一季则做为样本外数据，来萃取台股指数的知识规则，并据此来设计做多、做空的交易规则。利用文字探勘来分析这些所有非分类树，其“累计次数”与“不重复次数”的统计显示，使用次数最多的变量乃是“外资买卖超金额(千)”，而且出现在每一个知识规则。

*现职：(台湾)国立嘉义大学财务金融学系助理教授。最高学历：(台湾)国立交通大学信息管理研究所博士，双辅修：应用数学、统计。李孟育博士专长于数据分析、财务工程、实质选择权、科技管理、衍生性金融商品。自2008年起任职于(台湾)国立嘉义大学财务金融学系，担任助理教授。自2007-2008年，任职于(台湾)亚洲大学财务金融学系。自2004-2007年，则于民营机构担任金融工程师等职务。李博士曾经多次荣获教学、服务、导师奖项。拥有八年的教学经验与三年企业经验，其论文曾经荣获多个研讨会最佳论文奖。

第八届中国 R 语言会议手册，上海， 2015 年 11 月 21 日 – 22 日
主办：华东师范大学统计学院；第 27 页

商业大数据时代，GIS和R更配

何宇兵[1,*]

[1]辰智咨询

摘要

地理信息系统（GIS）在商业领域的价值日益体现，被越来越广泛应用在选址规划、市场拓展、营销管理、物流优化等方面。GIS在空间数据的采集、组织、管理、显示、分析、共享方面的技术优势，使其成为商业企业构建大数据分析基础平台的必需选项。R是一个能够自由有效地用于统计计算和绘图的语言和环境，已经位居数据挖掘领域所有语言之首，实际上已经成为专业数据分析领域的标准。R的扩展包数量还在不断地增长中，它能解决的问题还有无限可能。GIS+R的结合必将商业企业大数据分析平台的构建带来重要的影响。辰智在自主研发的面向行业的商业大数据分析解决方案商圈秀中对GIS+R进行了有益的尝试，事实证明商业大数据时代，GIS+R是一种趋势。

*上海辰智商务信息咨询有限公司GIS商业应用研究中心技术总监，从事GIS（地理信息系统）技术开发和商业应用研究工作近10年。曾就职Esri，担任高级架构师，两获Esri中国技术创新奖二等奖；后就职于麦当劳中国区担任GIS数据应用系统分析师；2014年加入辰智咨询就任辰智研究院GIS商业应用研究中心技术总监。具有丰富的GIS系统架构设计、分析和应用经验。主要关注GIS在选址规划、经营数据地理可视化、市场潜力分析、物流优化等领域的应用。服务的客户涉及政府、零售、餐饮、金融、保险、物流等行业，主要客户包括百胜、麦当劳、星巴克、阿迪达斯、乐购、沃尔玛、屈臣氏、交通银行、国元保险、德邦、大众等。

第八届中国 R 语言会议手册，上海， 2015 年 11 月 21 日 – 22 日
主办：华东师范大学统计学院；第 28 页

R and Tableau: Smart Meets Fast

胡羡祺[1,*]

[1] Tableau

摘要

Tableau is a visual reporting application that connects directly to R. It's designed for you, the domain expert who understands the data. Its drag-and-drop interface allows you effortlessly connect to libraries and packages, import saved models, or write new ones directly into calculations, visualizing them in seconds. Join us to see how you can use Tableau alongside R to speed up your data science projects and get them in front of more eyes, leading to smarter, data-driven business decisions.

*Sein Chyi is the Associate Sales Consultant based in Shanghai. She has been working with well-known companies and established MNC across various industry and countries in Asia Pacific, to integrate and adopt Tableau successfully into their organizations. With an expertise in Customer Insight Analytics for Retail and Banking industry, Sein Chyi provides effective solutions that help organization to gain insights and understand their data at the speed of thought. Sein Chyi holds a degree in Mathematics from Nanyang Technological University Singapore.

第八届中国 R 语言会议手册，上海， 2015 年 11 月 21 日 – 22 日
主办：华东师范大学统计学院；第 29 页

slidify+rCharts+ECharts 制作炫酷HTML5报告

严紫丹[1,*]

[1] 陆金所

摘要

slidify，可以创建HTML5 slides；rCharts，提供了各种可直接嵌入slidify的图表插件；ECharts，国内热门的开源作图工具。本文将介绍如何结合上述三大工具制作出不依赖网络，可交互，可重复的HTML5报告。

*严紫丹，互联网金融公司分析师

第八届中国 R 语言会议手册，上海， 2015 年 11 月 21 日 – 22 日
主办：华东师范大学统计学院；第 30 页

借助API快速搭建自然语言处理平台

邢代涛[1,*]

[1]SupStat

摘要

如今，很多实验室和机构开放了自然语言处理方面的平台。用户借助API，可以快速开发NLP方面的产品。NLPApiTools包将整合这些平台资源，提供R语言的接口和扩展功能。本例以哈工大语言云平台为例，介绍了借助NLPApiTools包，快速开发语义分析方面的应用，并在shiny上进行可视化的展示。

*邢代涛，supstat数据科学家，雪晴数据网的讲师。邮箱：daitao.xing@supstat.com.cn

第八届中国 R 语言会议手册，上海， 2015 年 11 月 21 日 – 22 日
主办：华东师范大学统计学院；第 31 页

从用R读琅琊榜小说讲讲用R读书的一些事

张云雁[1,*]

[1]PayPal

摘要

前不久琅琊榜大热的时候，开了一个神奇的脑洞，用R读了一下琅琊榜小说，写了个博文。写完后意识到R除了用来玩玩数据爬爬虫，它还能给我们读书读代码等等带来各种便利，因此脑洞一发不可收拾，又给自己挖了几个坑。所以这里继续借琅琊榜的东风讲一下近期脑洞的内容与一些研究结果，旨在给大家展示一下R语言除了用来做个报告画个图表搞搞数据挖掘外，还可以玩的一些东西。

主要内容包括：

一、聊聊用R语言打开琅琊榜的正确姿势

二、谈谈用R语言改变阅读习惯的可能性，以读某本书代码为例

三、讲讲没做的一些脑洞，以及其他

*R user，非计算机非统计野路子出身，属于因为工作数据量大需求麻烦才从EXCEL转用R的，结果感受到了R的魅力，掉入R语言的大坑一发不可收拾。平时主要关注R处理数据和图表自动化方面，有时候会用R画画图表爬爬虫搜罗优惠信息。(博客园@尾巴AR)

第八届中国 R 语言会议手册，上海， 2015 年 11 月 21 日 – 22 日
主办：华东师范大学统计学院；第 32 页

数据科学的博客：从knitr到jekyll

郎大为[1,*]

[1] 雪晴数据网

摘要

数据科学的概念与大数据一起兴起，一系列的数据科学的博客也慢慢出现在了业界。本报告从R语言中的knitr包(用于制作自动化文档)开始，到使用jekyll搭建基于github的博客的简单演示，展示了一条值得花时间的学习路线。希望通过这个过程，激起听众对数据科学博客的热情，了解前沿技术，分享项目经验，并督促自己不断学习。最终经历这条从自动化文档到构建数据科学博客的学习路线：1)使用knitr撰写自动化报告；2)github代码分享与版本管理；3)RSS订阅他人的博客；4)构建自己的数据科学博客。

*SupStat数据科学家，雪晴网专职讲师。擅长数据挖掘与数据可视化等领域。Remap，APItools包的作者。常用的工具是R，python，同时是前端的爱好者。博客，七风阁：http://chiffon.gitcafe.io

第八届中国 R 语言会议手册，上海， 2015 年 11 月 21 日 – 22 日
主办：华东师范大学统计学院；第 33 页

旅游O2O行内数据解析

张翔[1,*]

[1] 淘在路上

摘要

移动互联网和O2O创业是近两年的热潮，在赶上风口的同时，很多互联网时代老操盘手也把历史的糟粕带了进来。站在投资商，公司管理者的角度，和大家分享，以旅游O2O为例如何正确的解读和使用数据，避开数据中的陷阱，让这个行业更加阳光。

*淘在路上数据副总裁，Growth Hacker。长期任职于互联网领域，曾于2009年负责创立艾瑞咨询的数据挖掘部门，2013年创立gdcoast伟大海岸旅行网站，2014年至今，致力于协助旅行，汽车等相关行业企业的数据建设和应用。

第八届中国 R 语言会议手册，上海， 2015 年 11 月 21 日 – 22 日
主办：华东师范大学统计学院；第 34 页

里子和面子：R语言及数据挖掘助力京东推荐系统

熊熹[1,*]

[1] 京东商城

摘要

正如《一代宗师》所言，人活在世上,有的活成了面子，有的活成了里子。大多数用R的人，做的都是面子，并且R也擅长于此。但是在海量数据，高并发和复杂应用场景中，R的灵活和丰富的机器学习资源，依然能帮助我们做好不少里子工作。京东的推荐系统，正是一个最好的实例。

[*] 京东推荐算法工程师

第八届中国 R 语言会议手册，上海， 2015 年 11 月 21 日 – 22 日
主办：华东师范大学统计学院；第 35 页

Growth hacking？App 增长分析新玩法

任万凤[1,*]

[1] 诸葛IO

摘要

互联网+ 时代的降临使得疯狂的创业者们笑开了颜，但不幸的是互联网寒冬也伴随而来，人口红利也逐渐消失，因此创业者更需要把有限的资本花在刀刃上，找到下一个产品红利点。若想在这样的大环境下实现价值业务增长、快速找到Growth hacking策略，就必须对产品本身和用户群体行为有深入的分析。但大部分APP却没能利用好数据价值驱动决策，仍然闭着眼睛做产品，其中用户数据的流动（用户路径）才是产品的立身之本，根据用户的核心行为路径，挖掘出最易导致用户流失的产品位置并加以优化，帮助产品快速迭代，更加贴近核心价值用户需求。

本次演讲主要通过国内知名App进行实际的用户行为路径挖掘，找到产品的核心优化点，定位流失人群显著属性，同时通过Growth hacking挖掘帮助APP找到有助于用户留存的价值行为，为APP找到下一个核心红利点。

*任万凤，毕业于北京大学数学学院应用统计硕士，研究方向为移动互联网用户分析、社交网络兴趣识别、精准营销等。曾就职于天猫BI部，参与双11商品流量调控及预测等相关项目。毕业后加入诸葛IO(zhugeio.com)，担任资深数据分析师，从事社交用户兴趣识别、精细化运营、用户行为路径等相关工作。曾主要翻译《Tableau数据可视化实战》等书籍。

第八届中国 R 语言会议手册，上海， 2015 年 11 月 21 日 – 22 日
主办：华东师范大学统计学院；第 36 页

当游戏数据遇上R语言

谢佳标[1,*]

[1]深圳创梦天地科技

摘要

分别利用excel和R语言进行LTV（游戏生命周期）和ROI（投入产出比）模型预测，通过对比体现R语言的简洁性和灵活性；并介绍如何利用对应分析研究游戏玩家的喜好，利用社会网络图体现游戏玩家中的社交性。演讲过程中结合现成源码的解读与演示，令参会者们迅速理解R语言在游戏行业的应用。

*在创梦天地担任高级数据分析师一职，作为创梦天地数据挖掘组的负责人，带领团队对游戏数据进行深度挖掘，主要利用R语言进行大数据的挖掘和可视化工作。本人从事数据挖掘建模工作已有8年，曾经从事过咨询、电商、电购、电力、游戏等行业，了解不同领域的数据特点。有丰富的利用R语言进行数据挖掘实战经验。

第八届中国 R 语言会议手册，上海， 2015 年 11 月 21 日 – 22 日
主办：华东师范大学统计学院；第 37 页

利用历史业务数据实现系统异常的实时监测

唐力[1,*]

[1] 携程旅行网

摘要

互联网企业的系统可用性与客户服务质量及企业经济收益密切相关，主要受网络异常和数据库异常等因素的影响。除对系统采取必要的防护措施外，我们还应该综合利用历史业务数据，快速有效地检测系统异常以尽可能的降低系统故障所带来的损失。本次演讲将着重介绍利用历史业务数据监测系统异常的实际背景，理论基础和具体实施步骤并加以案例分析。

*唐力，就职于携程旅行网，担任数据分析师，毕业于华东师范大学统计学专业，主要兴趣是数据挖掘及其应用。郑锦超，就职于携程旅行网，担任数据分析师，毕业于华东理工大学理学院，感兴趣的领域是互联网企业的数据采集、数据挖掘、数据可视化。

第八届中国 R 语言会议手册，上海， 2015 年 11 月 21 日 – 22 日
主办：华东师范大学统计学院；第 38 页

缺失值处理与R语言

冯凌秉[1,*]

[1]澳大利亚国立大学

摘要

有过数据分析经验的人对于缺失值问题应该都不会陌生，但是对于如何看待缺失值的存在、如何处理缺失值以及分析不同的缺失值处理方法对于统计推断的影响等问题，就少有人问津了。但其实，缺失值处理过程中的草率或欠知都有可能对数据分析结果的可信度和有效性产生显著的影响。本演讲将着力概述缺失值处理的主流思路和方法，并在R语言环境背景下简要介绍其处理流程。

*冯凌秉，澳大利亚国立大学统计学博士，江西财经大学金融管理国际研究院讲师，硕士研究生导师。

第八届中国 R 语言会议手册，上海， 2015 年 11 月 21 日 – 22 日
主办：华东师范大学统计学院；第 39 页

古典概率的一些通用解法

杜传龙[1,*]

[1]Iowa State University

摘要

There are many interesting puzzles related to probability. Many of the puzzles are usually easy to understand even for people without statistics background, yet very hard to solve. In this talk, I will discuss some general approaches for solving probability puzzles.

*统计博士，实用派Linux fan，编程爱好者，个人主页：http://www.legendu.net

第八届中国 R 语言会议手册，上海， 2015 年 11 月 21 日 – 22 日
主办：华东师范大学统计学院；第 40 页

Introduction to Feature Hashing

吴齐轩(Wush Wu)[1,*]

[1] 国立台湾大学

摘要

In the world of online advertising, recommendation systems produce vast amounts of categorical data. This data has many, many levels of behavioural and text data. Too many to conveniently recode in R!

A good approach for pre-processing smaller categorical data is 'stats::model.matrix'. However, this approach is infeasible with recommender system data due to programming inefficiencies and pre-processing requirements (all data must be read in and synchronized). In this scenario, streaming algorithms are likely to fail while parallel algorithms become too complicated. However a solution has been found – feature hashing (also known as the hashing trick). And in 2015, many analysts use this approach to encode such data. The speaker has developed an R package that efficiently processes vast amounts of categorical data. This package, Feature Hashing (`http://cran.r-project.org/web/packages/FeatureHashing/index.html`) enables R users to easily apply the feature hashing with an interface similar to 'stats::model.matrix'. Attendees will learn the tips of using the feature hashing, exchange the experience of extending formula interface in R, and hear how some R users have successfully combined Feature Hashing with 'xgboost' to do text mining. Often use ordinal variables to record interviewees' attitude in questionnaire survey.

*Wush Chi-Hsuan Wu (吴齐轩) is a PhD student from the Institute of Electrical Engineering, National Taiwan University, where he is studying online advertising. He has contributed to several R packages, including digest, RcppCNPy, knitr and ckanr. Wush is a co-founder of Taiwan R User Group which hosts Machine Learning and Data Mining related talks weekly on Mondays(`http://www.meetup.com/taiwan-R`).

第八届中国 R 语言会议手册，上海， 2015 年 11 月 21 日 – 22 日
主办：华东师范大学统计学院；第 41 页

贝叶斯动态线性模型的商业化应用

陈堰平[1,*]

[1]SupStat

摘要

动态线性模型（DLM）是一类应用广泛的时间序列模型，贝叶斯预测方法是这种模型的经典预测算法。贝叶斯预测方法不仅仅依赖于t时刻以往的历史数据和根据模型的知识进行预测，还可包括专家的经验信息以及主观的判断来进行预测，这对于预测突发事件特别有用，而历史数据以及预先规定的模型并不能完全反映它们。当发现模型性能不好时，可求助于专家的经验和信息，对模型进行改进。贝叶斯预测方法，相对于Box-Jenkins传统的时间序列方法而言，有它的优点，它不必假设Box-Jenkins方法所必须的平稳性假设。贝叶斯预测方法通过人的主观经验给出先验分布，使得对数据量的要求大大减少。

本演讲分三个部分：

（1）以多渠道营销的动态ROI评估为案例背景，介绍DLM的模型形式及贝叶斯预测方法的原理；

（2）介绍DLM的其他应用场景：销售预测、资本资产定价模型的扩展、国际汇率预测、网络安全监测等；

（3）介绍我们在实际项目中如何设计估计DLM模型的R包，如何将R包的分析功能通过API的方式整合到业务系统中。

*陈堰平，北京数博思达信息科技有限公司（SupStat Analytics）联合创始人，微软公司合作伙伴，有多年从事统计咨询、数据分析、定制开发基于R语言的分析工具的经验，曾给花旗银行、东方航空、中国电信、中国联通等公司做过培训和咨询。现在也是统计之都管理团队成员，中国R语言会议理事会成员，曾获CQF国际数量金融认证，译作有《R语言编程艺术》《实用数据分析》，目前还参加其他几本R语言图书的编写和翻译。

第八届中国 R 语言会议手册，上海， 2015 年 11 月 21 日 – 22 日
主办：华东师范大学统计学院；第 42 页

如何攒一台深度学习服务器

肖凯[1,*]

[1] 开智微播

摘要

首先介绍一点深度学习的背景，然后是搭建服务器的各项注意要点，最后是基于深度学习方法展示一个关于文本分类的例子。

*肖凯，一个喜欢折腾数据的人，《数据科学中的R语言》作者之一，开智微播科技公司工程师。

第八届中国 R 语言会议手册，上海， 2015 年 11 月 21 日 – 22 日
主办：华东师范大学统计学院；第 43 页

旅游数据中的情感分析

毛苏晗[1,*]

[1] 淘在路上

摘要

在旅游评论数据中，用CRF算法训练词性的分类器，并运用一些规则提取出评论数据中的关键性的评价，以此看出评论者的好恶。其他一些自然语言处理的工具在旅游数据情感分析中的小应用。

*数据架构师，12年毕业于南京大学本科，先后在eBay和淘在路上工作，主要从事数据处理数据挖掘方面的工作。