★2014北京站★

CHINA R

第七届中国R语言会议

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
· 北京大学商务智能研究中心
· 统计之都

欢迎辞

豫章故郡，洪都新府。公元 675 年，滕王阁中高朋满座，胜友如云，王子安脍炙人口的《滕王阁序》也因之而诞生。一想到“豫章”，就想到了“豫章，大木也，生七年乃可知也”。R 语言会议从 08 年的第一届开始，到现在已经第七届，算上筹备的时间，刚好过了七年。

七年的时间，周文王在牢里写完了周易，光武帝已经初定中原。七年来，地球的人口增长了大约 10 亿，就连阿森纳都成功杀死了吴冠小朋友，当然，莱昂纳多还是没有拿到奥斯卡。子曰：“善人教民七年，亦可以即戎矣”。

R 语言会议办了七年，在统计之都各位同仁的辛苦努力下，这次会议相比往届有了更大的突破。本届会议的报名人数已经突破 1800，参会单位超过 600，包括数十位从新疆、西藏、香港、澳门、台湾，甚至欧洲、美洲、澳洲远道而来的朋友。业界和学界的单位数量比率约为 2:1，人数比率为 1:1。报名者中，互联网占据了超过三成的席位，IT 行业也贡献了四分之一，改变了之前 R 语言会议总是统计圈占大头的局面。这与 R 语言的发展轨迹是完全一致的。上个世纪 R 语言主要应用于学界；2005 年 R 在欧美爆发；2008 年国内燃起火焰，第一届中国 R 语言会议也应运而生；2013 年借助大数据的浪潮走出了统计圈；2014 年开始，R 已经全面进入业界的工程应用。

回顾历届 R 语言会议，2008 年在北京召开第一届会议，大多数人甚至没有上台演讲的经验，靠着一股血勇搞了个在当时被认为是自娱自乐的活动。2009 年开始渐成气候，在鸟兄张翔的努力下，上海也形成了固定的分会场。2010 年是艰难的一年，好在所有的统计之都小伙伴没有放弃，当作使命坚持了下来。2011 年迎来了业界的广泛关注，改变了之前学界为主的局面。2012 年出现会议爆满的情况，无论北京还是上海都是一座难求。2013 年终于可以轻松下来，不再发愁参会者和赞助的问题。2014 年自然要寻求新的突破，于是汇集了全球领军的 R 语言公司、大数据时代顶尖的学者、产业界的大咖，开始重新定义数据科学的含义。

正所谓风虎云龙、风云际会，再说下去就是物华天宝、人杰地灵了。让我们回到本次 R 语言会议，主题是数据科学；这是个融合了数学和统计模型、IT 技术、业务知识的全新领域，在大数据的时代真正地实现了数据分析的价值。R 的初心就是统计，纵然被语言声名所累，这些年在 IT 界也积蓄了大量成功经验。当前，上千个优秀的 R 包绝大多数都是来自于具体的行业和领域，从这个角度来看，R 已是数据科学界当之无愧的弄潮儿。让我们一起怀着一期一会的心念，给这个数据的时代做一个不平凡的记号吧。

此时此刻，你我共聚中国人民大学。仰观数据之大，俯察品类之盛，所以游目骋怀，足以极视听之娱，信可乐也。

目录

R 语言及会议简介

R 语言介绍

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的统计及相关院系。除了这些作者之外，R 还拥有一大批贡献者，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 CRAN 上的程序包数目已经超过 5500 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 2013 年做的“本年度使用何种编程或统计类语言进行分析和数据挖掘”的调查中*，R 以 61%的得票率再次荣登榜首 (2012 年为 53%)，力压 Python(39%)、SQL(37%)、SAS(21%) 和 JAVA(17%)。此外 R 还击败了 Excel 和 Rapidminer(2010 和 2011 年排名第一)，在“过去十二个月中你在实际项目中使用的数据挖掘或分析工具”的调查中排名第一。

Rexer Analytics 5th 数据挖掘者调查报告指出：R 语言一直保持上升的势头，牢牢占据工具类的第一名。几乎有一半的调查对象 (47%) 声称使用 R 语言作为数据挖掘工具。(`http://t.cn/zWKsvEZ`)。

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。

中国 R 语言会议介绍

第七届中国 R 语言会议 (北京会场) 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

*http://www.kdnuggets.com/polls/2013/languages-analytics-data-mining-data-science.html

- 北京大学商务智能研究中心
- 统计之都

量邦科技是本次会议的协办方。

会议时间为 2014 年 5 月 24~25 日。其中，24 日在中国人民大学如论大讲堂，25 日在中国人民大学明德商学楼 102、202、302 三个分会场。

会务人员如下：

- 主席：冷静；
- 秘书长：霍志骥；
- 秘书团：何通、朱雪宁、张翼飞、王雪琪、曹志强、王梦钰、王贺、罗兰、敬冯时、陈梓衡、蔡笑炜、王昱雯、朱弘昊、刘辰昂、高腾、周震宇、陈雅慧、叶晓萌、李博、郑晔；
- 委员会：冷静、霍志骥、李妙竹、高涛、肖楠、谢益辉、邓一硕、陈堰平、陈昱、林祯舜、王剑、林荟、刘思喆、李舰、苏建冲、肖展航、陈森、魏太云、邱怡轩、陈钢、陈逸波、郝智恒、张翔、陈丽云等。

他们大多是中国人民大学的本科生和研究生，此外筹备人员的单位还包括清华大学、北京大学、北京师范大学、浙江大学、中山大学、中国传媒大学等多家大陆高校，耶鲁大学、普渡大学、爱荷华州立大学、密西根大学等海外高校，以及百度、京东、阿里巴巴、中国平安、中国移动、檬果咨询、SupStat 等企业。

中国人民大学应用统计科学研究中心简介

中国人民大学应用统计科学研究中心前身是成立于 1988 年的统计科学研究所。十几年来，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和风险管理与精算三个各具特色的研究方向。几年来，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。此外，我们本着创建和发展面向实际应用的研究中心的宗旨，创建了：竞争力与评价研究中心、数据挖掘中心、六西格玛质量管理研究中心、保险精算中心、统计资讯研究中心等子机构，在突出应用主题的研究中心下，本着联系实际和服务实际的思想，创建面向实际应用的网站，建立新型的学术交流、知识普及和与用户零距离连接的模式。它们是：

- 数据挖掘中心网站；
- 六西格玛质量管理研究中心；
- 保险精算中心；
- 数据库研究室。

随着我国经济体制的进一步改革，本中心积极适应市场经济的需要，面向全国开放，加强国际学术交流与合作，推动重大应用统计项目的研究。中心现有专兼职研究人员 35 人，学术委员会委员 19 人，其中既有统计科学领域国内外著名的学术带头人，如中科院院士严加安教授、彭实戈教授；又有一批国内外知名学者和业务骨干，如千人计划胡飞芳教授，长江学者朱力行教授、林共进教授，讲座教授马双鸽教授、周晓华教授，还有吴喜之教授、袁卫教授、耿直教授、赵彦云教授、金勇进教授等。中心研究队伍强大的教育背景、研究成果和学术声誉将使本中心成为全国一流并具有国际声誉和影响力的开放式应用统计研究机构。

自 2000 年 10 月以来，中心专兼职研究人员共出版著作、教材 176 部，其中著作 105 部，著作中，在国外出版 4 部，译著 14 部；教材 71 部，其中“十五”、“十一五”规划教材 15 部，以中心署名发表论文 670 余篇，其中：SCI 及 SSCI 检索论文 32 篇；会议论文 27 篇，其中 EI4 篇，ISTP1 篇；软件及专利成果 3 项；向各级部门提交研究咨询报告 136 份。中心研究人员作为第一主持人承担课题 153 项，经费总额达 1636.1 万元，其中国家级以上项目 45 项，教育部重大项目及重大攻关项目 19 项，多项研究成果获得国家级奖励。在学术交流方面，近年来本中心举办了 32 次国内外学术会议，还邀请了约 213 人次国外著名统计专家、学者来本中心讲学和合作研究。开设精品课程 5 门，培养硕士、博士及博士后约 825 人。中心丰富的资料、先进的设备为科学研究提供了良好的物质条件。本中心现拥有中文图书、资料近万册，外文图书、资料约 3000 册；中文

期刊 105 种，外文期刊 43 种，电脑 65 台，局域网终端 18 个，以及多媒体投影设备等；拥有中国社会经济统计数据库、全国工业企业统计数据库、IMD 数据库、OECD 国家投入产出数据库等。这些资料和设备为研究人员提高工作效率，快速获取和传递信息，加强科学研究和学术交流创造了方便的条件。

现代统计在经济学、管理学、社会学、心理学、新闻传播学、伦理学、宗教学、法学和人口学中都得到了广泛的应用。中国人民大学是一所以经济、管理、人文社会科学为主的综合性大学。本中心将依托中国人民大学的学科优势与各院系开展横向联合与合作，使统计学真正成为人文社会科学研究强有力的工具。在多年的研究中，我们与中国科学院、北京大学、中国科技大学、台湾辅仁大学和香港大学等国内外著名学术机构以及国家统计局、中国人民银行和一些市场调查公司，保险公司等社会实际部门建立了良好的学术合作与交流关系，这进一步促进了统计学在社会各方面的应用，也为将本中心建立成为应用统计研究与交流的平台提供了较为完善的外部环境。

中国人民大学统计学院简介

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。教育部人文社会科学重点研究基地" 应用统计科学研究中心" 设在统计学院。拥有统计学和风险管理与精算学两个博士点，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硕士点，统计学、风险管理与精算学两个本科专业方向，以及应用经济学下统计学博士后流动站。

统计学院现有教师 38 人，其中教授 14 人，副教授 15 人，讲师 9 人。党政教辅人员 9 人。兼职教授、讲座教授、客座教授共 17 人。50 多年来，共培养不同层次人才 5000 多人。2013 年 3 月，在校学生共 617 人，其中本科生 344 人，硕士 199 人，博士 74 人。大多数毕业生在金融、保险、证券、基金、信息等领域从事数据采集和分析工作。

北京大学商务智能研究中心简介

北京大学商务智能研究中心是一个面向互联网大数据的科研平台。中心尤其关注具备以下三种特征的互联网大数据：

1. 中文文本数据；
2. 网络结构数据；
3. 地理位置数据。

为此，中心依托北京大学光华管理，联合众多互联网企业、科研机构和社区。中心现有合作学者十余人，横跨统计学、营销、管理科学、计算机等众多学科。合作学者来自海内外知名大学。例如：北京大学、人民大学、中央财经、四川大学、西安欧亚学院、俄亥俄州立大学等。中心现有合作企业多个，覆盖互联网众多细分行业。例如：博雅立方，百度，新浪等。中心在研项目包含但不局限于：

1. 基于微博数据理解企业竞争态势；
2. 基于搜索 Cookie 数据了解消费者特征；
3. 基于 URL 首页文本数据萃取企业行业特征；
4. 基于社交媒体评论热度，理解股市走向等。

中心诚挚邀请有共同兴趣的学术机构、企业伙伴以及个人相互切磋，共同学习，互相提高！

统计之都简介

目前，统计之都网站主要由主站和论坛两部分构成。主站是由众多撰稿者定期更新的统计学博客，内容涵盖统计学理论、数据分析技巧、统计模型、统计计算和软件应用等广泛的内容，间或发表书评、文评以及其它统计学网站或博客导读；主站文章来源于本站撰稿人的贡献以及用户推荐，主站对这些文章有一定的审稿机制以保证文章质量。录用稿件的标准为：重实用、有新意、无抄袭、可重现。

论坛则是用户进行各类统计学问题讨论的互动平台，论坛结构全面，注册人数众多(六万人)；目前由数理统计、应用统计、软件应用、数据挖掘、生物统计、金融统计等板块构成，各板块均由相应领域具有坚实专业背景的专家担任版主，负责审核管理工作，保证论坛的高质量运转。其中的 R 语言板块是国内最活跃的 R 语言社区。

线下活动

除了线上的交流之外，统计之都社区还定期开展了一系列的线下活动，其中最有影响力活动的是自 2008 年以来定期举行的中国 R 语言会议和 2012 年以来定期开展的 COS 数据分析沙龙。中国 R 语言会议于每年的 5 月和 11 月分别于北京和上海召开，至今已累计举办 6 届，共吸引 6000 人次到场参会，该会议的成功举办对于国内 R 语言和数据分析的普及起了极大的推动作用；COS 数据分析沙龙以每 1-2 月一次的频次分别在北京、上海和深圳三个城市举行，每期都会邀请 1 至 2 位数据可视化、数据挖掘、电商数据分析、海量数据处理等前沿的统计学领域的嘉宾为大家分享经验，促进数据分析从业者面对面交流，分享经验，共享人脉。

社区愿景

建站至今，统计之都始终遵循互联网开源精神，网站的维护和管理均由志愿者完成。目前，负责管理维护统计之都网站运营的志愿者组织是 COS 理事会，但任何人都可以通过向管理员申请的方式成为主站作者或网站的维护者。

统计之都的治站格言是「专业、人本和正直」，在此格言指导下力图通过专业的知识和团队、人本的交流与传播、正直的态度和审视，来更好地推动统计学在中国的发展与传播。

目标

- 主站：传播统计学前沿动态，普及统计学基础知识，提供统计学应用实例。多角度多层次的为具有不同统计背景的人，从业余统计爱好者到统计专家，提供全方位统计学信息，力争每一篇文章都能让读者真正受益。
- 论坛：做广告最少、“讨论问题帖/下载帖”比率最大、单帖平均字数最多、为统计爱好者解疑答惑、学习氛围融洽的高质量统计学问答论坛。
- 沙龙：秉承交流 (Chat)、开放 (Open) 和分享 (Share) 的理念，为广大数据分析和数据挖掘的同仁提供了一个分享的平台。
- R 语言会议：促进 R 语言在中国的推广和发展，为各领域 R 语言使用者提供交流和互动的平台。

联系我们

如果有兴趣加入我们，请联系我们：

- 主页：`http://cos.name`；
- 邮箱：`admin@cos.name`；
- 微博：统计之都；
- 人人：统计之都；
- 微信：统计之都；扫描本手册封底二维码即可关注。

统计之都活动回顾

除了线上的交流之外，统计之都社区还定期开展了一系列的线下活动，线下活动总结：

1. R 语言是在统计和数据挖掘学界广泛应用的编程语言和开发环境，其免费、开源、灵活的特点与统计之都的文化不谋而合。2008 年起，统计之都在人民大学一间教室，举办了第一届中国 R 语言会议，自此，中国 R 语言会议与统计之都一同成长，规模越来越大，至今已成功举办了 6 届。如今 R 会议已经成为国内影响力最大的 R 语言社区盛会，聚学术专家、业界精英、技术大咖于一堂，使各届 R 使用者得到了充分的交流。应广大支持者的需求，从 2009 年第二届中国 R 语言会议起，在北京和上海设两个分会场，分别于 5 月下旬和 11 中旬举办。第七届 R 语言会议将于 5 月 24-25 日在中国人民大学如论讲堂举办。

   历届会议纪要：

   - 第六届中国 R 语言会议：http://cos.name/chinar/chinar-2013/
   - 第五届中国 R 语言会议：http://cos.name/chinar/chinar-2012/
   - 第四届中国 R 语言会议：http://cos.name/chinar/chinar-2011/
   - 第三届中国 R 语言会议：http://cos.name/chinar/chinar-2010/
   - 第二届中国 R 语言会议：http://cos.name/chinar/chinar-2009/
   - 第一届中国 R 语言会议：http://cos.name/chinar/chinar-2008/

2. COS 沙龙是依托于统计之都而成立的线下活动组织。其宗旨在于凝聚国内外统计领域、数据挖掘及数据分析领域相关人士，打造一个“开放、共享”的线下活动社区。口号是“共享知识、网络人脉”。目前，COS 沙龙以约每月 1 次的频率在北京、上海、广州等城市举行。自成立之初，累计举行 20 多场次，邀请了国内外众多学界、业界精英担当沙龙分享嘉宾，吸引了多达 1500 人次的与会人员。

   - 参与方式：每期沙龙开始前将在微博和论坛上进行通知，同时我们将建立参与过沙龙的朋友们的邮件列表。
   - 新浪微博：@ 统计之都　或者 `http://weibo.com/cosname`
   - COS 论坛：`http://cos.name/cn/`
   - 沙龙邮箱：`salon@cos.name`

3. 比赛、交流会、讲座：除了定期的活动之外，我们还将不定期举行或者和其他组织

共同举办一些交流会和讲座。北京的交流会主要以中国人民大学为中心，而上海的则更偏重于沟通合作。目前，我们主办或协办过的活动包括：

- 数据挖掘竞赛；
- 科普讲座：与科学松鼠会合作，分别在北京、上海、杭州联合举办过统计科普讲座——别让数字吓到你；
- 校内交流会：不定期在学校内举办统计建模、统计实际应用、统计学出国等经验交流会，包括中国人民大学、北京大学、北京师范大学、南开大学等。

4. 出版，包括电子手册、书籍撰写和翻译：

- **电子手册：**
  - 陈堰平：《shiny 官方教程中文版》；
  - 邱怡轩：《parallel 包中文文档》；
  - 陈钢：《R 入门 25 招》；
  - 邓一硕：《quantmod 学习笔记》；
  - 刘思喆：《R 语言环境下的文本挖掘》；
  - 刘思喆：《153 分钟学会 R》；
  - 刘思喆：《R 参考卡片》；
  - 陈丽云：《在 R 中玩转计量》；
  - 魏太云：《R 与最优化》；
  - 谢益辉：《R 语言忍者秘笈》(正在编写)。
- **书籍撰写：**
  - 李舰、肖凯：《数据科学中的 R 语言》(即将由西安交通大学出版社出版)；
  - 谢益辉：《现代统计图形》(正在编写)；
  - 李舰、肖楠、周扬、赵扬、魏太云等：《从数据到报告》(正在编写)。
- **书籍翻译：**
  - R in Action(《R 语言实战》，译者高涛、肖楠、陈钢，已经由人民邮电出版社出版)；
  - ggplot 2: Element Graphics for Data Analysis (《ggplot2：数据分析与图形艺术》，译者邱怡轩、魏太云、主伟呈、高涛、肖楠、潘岚锋，审校殷腾飞，已由西安交大出版社出版)；

- The Art of R programming(《R 语言编程艺术》，译者陈堰平、潘岚锋、邱怡轩，已由机械工业出版社出版)；
- R Graphics Cookbook(《R 数据可视化手册》，译者肖楠、邓一硕、魏太云，审校邱怡轩，已由人民邮电出版社出版)；
- Introductory statistics with R(《R 统计入门》，译者郝智恒、何通、刘旭华、邓一硕，已由人民邮电出版社出版)；
- Introductory Statistics with R(《R 语言统计入门》，译者郝智恒、何通、邓一硕、刘旭华，已由人民邮电出版社出版)；
- R in a Nutshell(《R 核心技术手册》，译者刘思喆、李舰、陈钢、邓一硕等，即将由电子工业出版社出版)；
- Financial Risk Modelling and Portfolio Optimization with R(译者邓一硕、郑志勇等，即将由机械工业出版社出版)；
- Applied Predictive Modeling (译者邱怡轩、肖楠、林荟、马恩驰，即将由机械工业出版社出版)；
- Seamless R and C++ Integration with Rcpp(译者寇强、张烨，即将由西安交大出版社出版)；
- The R Book(译者陈逸波、郝智恒、郑欣等，即将由机械工业出版社出版)。

特别赞助单位介绍

光大证券

光大证券股份有限公司（以下简称“公司”）创建于 1996 年，系由中国光大（集团）总公司投资控股的全国性综合类股份制证券公司，是中国证监会批准的首批三家创新试点公司之一。2009 年 8 月 18 日在上海证券交易所挂牌上市，股票代码：601788。

公司成立十七年来，秉承“诚信专业卓越共享”的核心价值观和“合规稳健，创新发展”的经营理念，资本充足、内控严密、运营安全、服务优质、效益良好、创新能力和市场竞争能力突出。公司积极投身于国内外资本市场，各项业务迅速发展，业务规模及主要营业指标居国内证券公司前列，综合实力一直位居证券业第一集团军。

公司拥有一支高素质、专业化的员工队伍，在制度创新、产品开发、人力资源、风险管理、IT 平台和品牌管理方面综合优势明显。展望未来，公司将以交易为驱动、以增值服务为手段、以创造价值为目标，积极探索和大力培育资本中介服务业务，致力成为为客户投资、企业融资、资产定价和风险管理提供优质产品和高效服务的创造者和销售者，市场交易的组织者、流动性的提供者，力争把光大证券打造成具备全能型金融服务能力的国际一流品牌。

微量网

微量网 (www.wquant.com) 是国内顶尖的量化投资策略在线交易平台，策略提供者和理财投资者对接平台 (既包括针对理财者的量化投资策略评价和推荐，也包括针对策略提供者的策略生产和出售)。网站集策略研发、销售、交易为一体，投资者无需安装软件，通过网页或手机控制云端交易账户，尊享 7*24 小时无人值守交易。策略提供者可以通过量邦天语程序化交易系统研发投资策略，并通过微量网销售。微量网涵盖理财型和交易型策略，满足不同风险偏好理财投资者的需求，支持国内证券、期货、上海黄金以及渤海商品等八个交易所的所有品种交易，统一管理证券、期货、黄金和现货账户，交易方式包括网页端、移动 APP 端、移动微信端以及 PC 端等，资金不出个人法定交易账户，全程安全托管。微量网现面向全国招募量化达人成为策略提供者，有意者请扫描本段文字右上角二维码关注订阅号 (wquant)，或加入 QQ 群 188396749。

赞助单位介绍

SupStat

SupStat 是总部位于纽约的统计咨询公司, 在北京设立有分支机构，中文名为北京数博思达信息科技有限公司。公司为全球的企业和科研机构提供定制化的统计培训、咨询和软件。

Revolution Analytics

Revolution Analytics 是一家基于开源 R 项目基础上，提供软件与服务的领先商业提供商。公司为愿意用 R 这种世界上最强统计语言的企业带来高效能以及高生产率。

RStudio

RStudio 是一家致力于为 R 的统计计算环境提供软件、教育、服务的公司。公司团队成员积极地工作在增强 R 的一些开源项目中，包括 RStudio IDE(一种 R 的集成开发环境)、著名的 ggplot2,plyr 包，还有很多其他有名的 R 包。

VSNC

VSNC 的母公司 VSN International Ltd (VSNi) 是由英国洛桑试验站和 Numerical Algorithms Group (NAG) 共同成立的软件公司，公司主要产品有 GenStat，ASReml，CycdesigN，并提供相应的咨询培训服务。VSNC 秉承着母公司以用户为主，提供超世界一流的产品和服务的理念，全面负责 VSNi 产品和服务在中国以及亚洲市场的业务。

1degreenorth

1degreenorth 是一家为高性能计算 (HPC) 集群、公共和私有云基础设施和业务分析提供专业解决方案的公司。公司有一个充满激情的咨询与 HPC 架构师团队，从 1999 年以来一直致力于研究 HPC 集群与网格。公司目前为政府研究机构、跨国公司在新加坡和亚太地区在构建可伸缩的应用程序和基础设施的私人和公共云领域提供咨询和业务分析的解决方案。

合作出版社介绍

华章图书

北京华章图文信息有限公司是机械工业出版社与 (美国) 万国图文信息有限公司共同投资建立的合资企业，主要从事科技、经管、外语领域的图书出版服务业务，公司成立于 1995 年 10 月。几年以来，公司发展迅速。1995 年以计算机图书业务起步，1998 年奠定了国内计算机图书出版的四强格局；1998 年进入经管图书领域；2000 年创办互动出版网。全球采集内容，服务中国教育，成为华章的鲜明特色。公司出版的图书代表着计算机、经济管理、文化和英语教学领域的最新视野，能最好地满足专业人士、教师和学生不断变化的需求。

Springer

德国斯普林格 (Springer-Verlag) 出版社是世界上最大的科技出版社之一，它有着 150 多年发展历史，以出版学术性出版物而闻名于世，它也是最早将纸本期刊做成电子版发行的出版商。Springer Link 系统就是通过 www 发行的电子全文期刊检索系统，该系统目前包括 490 多种期刊的电子全文，其中 390 多种为英文期刊。根据期刊涉及的学科范围，LINK 将这些电子全文期刊划分成 11 个出色的《在线图书馆》，分别是：化学、计算机科学、经济学、工程学、环境科学、地理学、法学、生命科学、数学、医学、物理学与天文学。

中国人民大学出版社

中国人民大学出版社成立于 1955 年，是新中国建立之后成立的第一家大学出版社。1982 年人大出版社被教育部确定为全国高校文科教材出版中心，是中国高校教材、学术著作出版最重要的基地之一。2007 年荣获首届中国出版政府奖先进出版单位奖。2009 年荣获“2009 年度北京市新闻出版和版权工作先进集体”荣誉称号。2009 年 9 月在首次全国经营性图书出版单位等级评估中被评为一级出版社，被新闻出版总署授予“全国百佳图书出版单位”荣誉称号。

中国人民大学出版社秉承“出教材学术精品, 育人文社科英才”的理念，大力实施精品战略，以优秀的出版物传播先进文化。建社 50 多年来，累计出版图书一万余种，出版了一大批具有文化积累与文化传播价值的优秀教材和学术著作，涵盖了社会科学和人文科学各学科，包括哲学、经济学、政治学、法学、社会学、行政学、人口学、环境学、新闻学、档案学、财政学、金融学、管理学、会计学、商品学、历史学、语言文学、伦理学、心理学、美学、艺术以及新兴学科和边缘学科等。其中许多教材多次再版，一些

教材发行数量高达数十万册以至数百万册。2011 年人大出版社出书 2988 种，印制码洋达 8.3 亿元。经过长期的积累，中国人民大学出版社已发展成为具有图书、期刊、音像、电子和网络出版物等多媒体兼营的大型综合性出版社。

人民邮电出版社

人民邮电出版社成立于北京，现年出版通信、计算机、电子电工、工业技术、少儿、经管、摄影、旅游、交通、集邮等类新书及相关教材 2600 余种。其中，计算机、通信、摄影类图书零售市场占有率排名第一，电子电工、少儿、旅游类图书排名第二，具有较强的市场竞争力。期刊业务坚持以期刊为平台，走多元化发展之路。

第七届中国 R 语言会议北京会场日程

5 月 24 日白天(如论大讲堂)

环节	时间	内容
开幕式	08:45-09:00	会议主席冷静致辞 中国人民大学统计学院领导致辞
主场演讲 1 主持人：冷静	09:00-09:30	Hadley Wickham：R packages: principles and best practices
	09:30-10:00	David Smith：How the growth of R helps data-driven organizations succeed
	10:00-10:30	合影 • 讨论 • 休息
主场演讲 2 主持人：叶晓萌	10:30-11:00	余凯：Deep Learning Unfolds Big Data Era
	11:00-11:30	周明：计算机对联和诗词
	11:30-12:00	王汉生：A Statistical Model for Social Network Labeling
	12:00-13:30	中午休息
主场演讲 3 主持人：霍志骥	13:30-14:00	胡浩：云计算时代的量化投资
	14:00-14:30	靳志辉：广告定向中的用户分析
宣讲环节： 主持人：林祯舜	14:30-15:15	● 光大证券 ● 微量网 ● 中華 R 軟體研發暨應用協會 ● SupStat ● Rstudio ● Revolution ● 1degreenorth
	15:15-15:30	休息
主场演讲 4 主持人：罗兰	15:30-16:00	郑义：基金评选平台之建立
	16:00-16:30	廖逸竹：玩转三亿视频—数据分析在视频产业的应用
讨论环节： 主持人：林祯舜	16:30-17:30	讨论、答疑： 大数据/数据科学之产业&教育主题 嘉宾：吴喜之、王汉生、余凯、姚远、杜长嵘、胡浩

5 月 24 日晚上(明德商学楼楼负一层，泊星地咖啡厅)

时间	活动 • 数据之夜
18:00-21:00	● 嘉宾晚宴 ● 自由讨论 ● 穿插部分演讲

注：24 日晚上，我们已经包场泊星地咖啡厅供大家更加深入的交流，该咖啡厅也是嘉宾们晚宴和交流讨论的地点。欢迎广大参会人员一同前往晚宴、交流，普通参会者入场时收取场地费 100 元，餐费自理(30 元左右)，上限 100 人。

5 月 25 日 A 场(明德商学楼 102) 会场秘书：肖展航

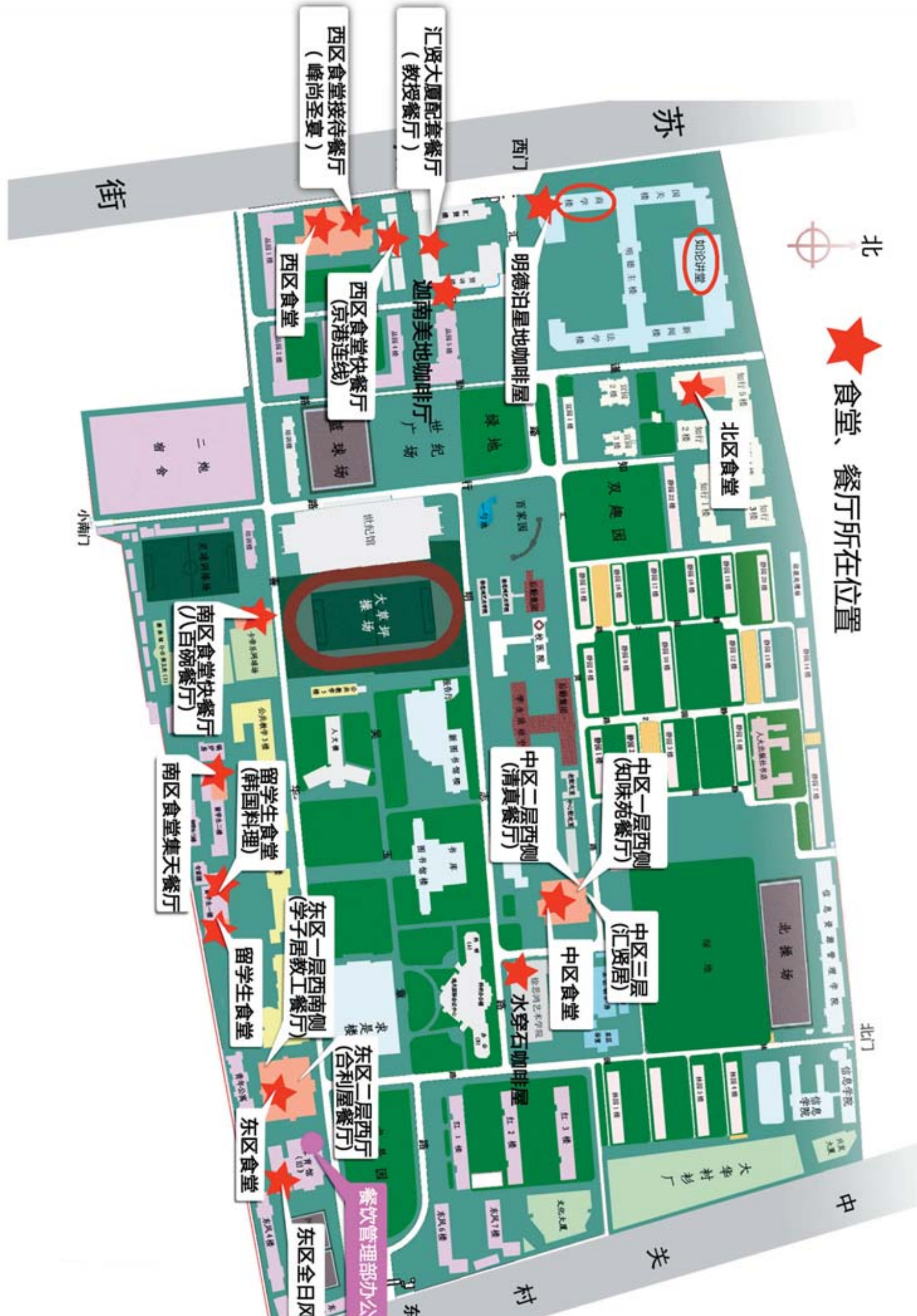
环节	时间	内容
专题 1 数据模型 主持人：李毅	09:00-09:30	张家齐：Hacking Models with R
	09:30-10:00	James Wicker：Multi-Cluster Detection
	10:00-10:30	讨论 • 休息
专题 2 量化投资 主持人：郑晔	10:30-11:00	景亮：基于 R 的程序化交易
	11:00-11:30	牟官讯：开发的血和泪，交易的冰与火
	11:30-12:00	张丹：R 语言与金融大数据应用
	12:00-14:00	中午休息
专题 3 R 数据可视化 主持人：张翼飞	14:00-14:30	王亮博：Interactive Visualization with R
	14:30-15:00	周扬：它山之石可以攻玉：recharts 图形包
	15:00-15:30	Hadley Wickham: ggvis
	15:30-16:00	讨论 • 休息

5 月 25 日 B 场(明德商学楼 202) 会场秘书：陈森

环节	时间	内容
专题 4 大数据产业 主持人：何通	09:00-09:30	陈堰平：大数据的新方向：公开同享趋势下的新数据产业
	09:30-10:00	陈景祥：R-Web：大数据分析及导引云平台
	10:00-10:30	讨论 • 休息
专题 5 R 中的大数据 主持人：高腾	10:30-11:00	寇强：突破 R 内存瓶颈的若干技术
	11:00-11:30	吴齐轩：Large Scale Learning with R
	11:30-12:00	丘右玮：Big Data Analysis with RHadoop
	12:00-14:00	中午休息
专题 6 R 高级技术 主持人：敬冯时	14:00-14:30	任坤：构建高效率的数据流水线：在 R 中使用管道操作
	14:30-15:00	张晔：科研角度下的 R 包开发
	15:00-15:30	邱怡轩：R 中大规模矩阵的特征分解、svd、nmf 等
	15:30-16:00	讨论 • 休息

5 月 25 日 C 场(明德商学楼 302) 会场秘书：苏建冲

环节	时间	内容
专题 7 R 与其他工具 主持人：王梦珏	09:00-09:30	郭韦廷：Data Analysis with R and Python
	09:30-10:00	李舰：R 与 Office 的整合
	10:00-10:30	讨论 • 休息
专题 8 数据分析实战 主持人：黄俊文	10:30-11:00	林荟：数据分析在传统行业商业决策中的应用
	11:00-11:30	欧阳鹤：小而美的数据产品
	11:30-12:00	刘思喆：R 与企业级数据挖掘
	12:00-14:00	午休
专题 9 生物、心理学 主持人：陈纲	14:00-14:30	肖楠: Integrated Pipeline for Systems Pharmacology in R/Bioconductor
	14:30-15:00	杨环：R 在新药研发中的应用
	15:00-15:30	江歌：Combining R with Psychology——An illustration with SEM
	15:30-16:00	讨论 • 休息


食堂、餐厅所在位置
北
苏
街
中
关
村
街
西门
东门
北门
小南门
如论讲堂
世纪馆
大草坪操场
世纪广场
绿地
双趣园
北操场
西区食堂接待餐厅（峰尚至宴）
汇贤大厦配套餐厅（教授餐厅）
西区食堂
西区食堂快餐厅(京港连线)
迦南美地咖啡厅
明德泊星地咖啡屋
北区食堂
南区食堂快餐厅(八百碗餐厅)
留学生食堂(韩国料理)
南区食堂集天餐厅
留学生食堂
东区一层西南侧(学子居教工餐厅)
东区二层西厅(合利屋餐厅)
东区食堂
餐饮管理部办公地点
东区全日风味餐厅
中区一层西侧(知味苑餐厅)
中区二层西侧(清真餐厅)
中区三层(汇贤居)
中区食堂
水穿石咖啡屋

R packages: principles and best practices

Hadley Wickham(RStudio)
h.wickham@gmail.com

嘉宾介绍

Hadley Wickham 是 RStudio 公司的首席科学家，同时也是美国 Rice 大学的助理教授。他开发了著名的 ggplot2, plyr, dplyr, shiny, devtools, stringr, ggvis, httr 等包，在 R 社区中很受欢迎。

演讲摘要

R packages have a reputation for being complex, unwieldly beasts that need decades of study to master. In this talk, I'll show you that when you have the right tools, R packages are easy; so easy, in fact, that they should be your default whenever you combine code, data or documentation. Packages are great just for yourself, and they're also great if you want to share. Sharing a package requires a little more work so that it works everywhere (not just on your computer), but the right mindset and the right tools make easy and a small additional time investment makes it possible for you to share your work with the world through github or CRAN.

You might think that you'll never need a package because you only use R to do data analysis. But many data analysis are complex, and can't be solved with a few lines of R code. Instead, you need to write functions to capture common solutions to repeated problems. As soon as you start writing functions, it's a good idea to learn a little bit about packages so that you can make functions that are well-documented and well-tested.

How the growth of R helps data-driven organizations succeed

David Smith(Revolution Analytics)
david@revolutionanalytics.com

嘉宾介绍

David Smith 是 Revolution Analytics 公司的 Chief Community Officer。他领导着该公司的开源解决方案团队。借助他的数据科学背景，他每天都在 Revolution 的博客网站上撰写 R 语言在预测性分析中应用的文章。他被福布斯杂志评为“大数据”主题中十大最有影响力人物之一。他是 R 语言的培训手册“An Introduction to R”的作者之一，并且是 ESS 项目 (Emacs Speaks Statistics，Emacs 与 R 相互的插件) 的最初开发者之一。在加入 Revolution Analytics 之前，David 是 Insightful 公司负责 S-PLUS 产品管理的董事之一。他的 twitter 账号是 @revodavid。

演讲摘要

Adoption of the R language has grown rapidly in the last few years, and is ranked as the number-one data science language in several surveys. This accelerating R adoption curve has been driven by the Big Data revolution, and the fact that so many data scientists —having learned R at university —are actively unlocking the secrets hidden in these new, vast data troves.

In more than 6 years of writing for the Revolutions blog, I’ ve discovered hundreds of applications of R in business, in government, and in the non-profit sector. Sometimes the use of R is obvious, and sometimes it takes a little bit of detective work to learn how R is operating behind the scenes. In this talk, I’ ll begin by presenting some recent statistics on the growth of R. Then I’ ll recount some of my favorite applications of R, and show how R is behind some amazing innovations in today’ s world.

Deep Learning Unfolds Big Data Era

余凯 (百度 IDL)
yukai@baidu.com

嘉宾介绍

余凯是百度深度学习研究院 (IDL) 常务副院长, 南京大学和北邮兼职教授，中科院计算所客座研究员，国家“千人计划”专家、中关村高端领军人才及北京市海外高层次人才。余凯先生先后毕业于南京大学和慕尼黑大学，毕业后曾在微软、西门子和 NEC 工作。曾任斯坦福大学计算机系 Adjunct Faculty。他至今发表数十篇论文，论文共计被引用 5000 余次，曾荣获 ICML-2013 的最佳论文奖银奖，并曾在 PASCAL VOC, ImageNet 等竞赛中获国际第一。2013 年，他领导的百度语音团队荣获“2013 百度最高奖”，其团队开发的基于图像技术的“百度魔图 PK 大咖”成为 2013 年最火爆的移动图片应用产品之一。近年来，他领导的团队使得深度学习在互联网广告业务和网页搜索排序获得突破性进展。

演讲摘要

Dr. Kai YU, Head of Institute of Deep Learning (IDL) at Baidu, will speak about Baidu’s recent efforts in developing cutting-edge technologies in the areas of deep learning and broader artificial intelligence. By leveraging big data aided by massive parallel computation, enthusiastic IDL researchers have led to significant improvements to Baidu’s core business, such as search, ads, speech recognition and computer vision. More importantly, these technologies have been shaping up a foundation for imaginative long-term innovation.

计算机对联和诗词

周明 (微软亚洲研究院)
mingzhou@microsoft.com

嘉宾介绍

周明是微软亚洲研究院自然语言计算组首席研究员，中国计算机学会通讯动态栏目主编。哈工大、南开大学等高校博导、清华 -微软联合实验室联合主任。周明先生于 1991 年在哈工大获得博士学位，1991 年到 1999 年在清华大学计算机系任博士后和副研究员。1999 年加盟微软亚洲研究院，随后开始担任计算语言计算组的负责人 (曾短期兼任过语音组的主任)。他是中国第一个中 -英机器翻译系统 CEMT—I、中 —日机器翻译系统 J–北京的发明人。此外，他领导团队开发了微软对联、微软中日文输入法、英语写作助手、微软聊天机器人、必应词典、英库问答、微博搜索、微软中 -英翻译等系统，其团队为必应搜索、Office、SQL、Windows 及微软语音翻译系统等产品做出重要贡献。其与中科院计算所合作的基于 Kinect 的手语的识别和翻译也声名不菲。他曾任首届亚洲信息检索大会 (AIRS2004) 程序委员会主席、中国计算机学会自然语言处理和中文计算大会 (NLP&CC) 程序委员会主席 (2012 年首届大会) 和大会主席 (2013 年)，并曾多次担任 ACL、SIGIR、EMNLP、IJCAI、COLING 等国际学术会议的领域主席。

演讲摘要

对联和诗词是中国的重要文化遗产。对联和诗词有严格的对仗、平仄、韵律等要求，并讲究意境的精妙。用计算机自动产生对联和诗词是人工智能的一项难题，而且在学术界的研究也不多见。本研究创造性地把对联和诗词生成看作是一种特殊的机器翻译过程。我们提出了一个基于短语的机器翻译的解码方法。对用户输入的上联，系统产生下联的多个候选。然后一组基于对联要求的语言学规则惩罚不符合对联要求的候选。最后，通过一个 Ranking 机制综合利用多属性进行重新排序。基于这个方法，我们开发了微软对联系统 (http://duilian.msra.cn)。

在对联研究基础上，我们进一步扩展到对诗词的自动生成研究，我们目前以绝句为例进行了初步研究。目前的场景是用户给出几个关键词，用于描述自己的意图，然后系统生成一首绝句。系统首先通过语言模型生成绝句的第一句。然后，采用统计机器翻译的方法逐句生成以下三句。在生成第 N 句的时候，考虑了以前生成的 N-1 句以避免词语的重复、遵循对仗和平仄，并保证意义的连贯。通过初步的实验验证了本方法的有效性。

A Statistical Model for Social Network Labeling

王汉生 (北大光华)
hansheng@gsm.pku.edu.cn

嘉宾介绍

王汉生是统计学博士，北京大学教授、博士生导师，现任北京大学商务智能研究中心主任和北京大学光华管理学院商务统计与经济计量系主任。先后毕业于北京大学数学科学学院概率统计系 (1998)，美国威斯康星大学麦迪逊分校 (2001)。现为国际统计研究员、美国统计学会、美国数理统计研究员；英国皇家统计协会以及泛华统计学会会员。同时也是 Computational Statistics & Data Analysis (2008—现在),Statistics and its Interface (2010—现在)，Journal of Business and Economic Statistics(2012 至今) 和 Journal of the American Statistical Association (2011 至今) 的副主编。至今已发表英文学术论文五十余篇，中文论文近二十篇。同时曾合著英文专著一本，独立完成中文教材一本。关注的理论研究领域包括：高维数据分析、变量选择、数据降维、极值理论以及半参数模型等；关注的应用研究为：搜索引擎营销和社会关系网络等。

演讲摘要

We consider here a social network from which one observes not only network structure (i.e., nodes and edges) but also a set of labels (or tags, keywords) for each node (or user). These labels are self-created and closely related to the user's career status, life style, personal interests, and many others. Thus, they are of great interest for online marketing. To model their joint behavior with network structure, a statistical model is developed. The model is based on the classical p1 model but allows the reciprocation parameter to be label dependent. For both dense and sparse networks, we obtain maximum likelihood estimators, which are statistically efficient but computationally expensive. To alleviate the computational cost, a novel conditional maximum likelihood estimator is proposed for large scaled sparse network. The asymptotic properties of these estimators are investigated. Simulation studies are conducted and a real Sina Weibo dataset is analyzed.

云计算时代的量化投资

胡浩 (微量网)
huh@quanttech.cn

嘉宾介绍

胡浩现任微量网络科技 CEO(互联网证券金融领域的国家高新技术企业)，毕业于中国人民大学，获统计学硕士和金融工程博士学位，曾担任中信证券首席金融工程分析师、多家大型资产管理机构量化投资负责人。胡浩博士长期从事数量金融研究，曾带领中信证券金融工程团队在《新财富》最佳分析师评选、中国证券业协会、深圳证券交易所征文大赛等活动中多次获奖。胡浩博士致力于以大数据分析为基础、结合金融理论和投资者行为分析解释资本市场现象，构建 A 股市场量化投资策略体系。他目前主导的“微量网”项目是互联网证券金融的领导品牌，搭建了投资策略提供者和策略使用者之间的“云交易”平台。

演讲摘要

随着资本市场的发展，量化投资逐渐为国人所熟悉，但看起来，似乎只有专业人士才能进行量化投资，其实不然，量化投资的核心在于你是否具有“模式”投资的思维，而数据存储、模型测算、IT 执行等在云计算时代不再成为一个难题。换句话说，如果你从量化的角度思考资本市场并且找到了某些规律性的东西，那么在云计算时代，在外部系统的帮助下你也可以成为一个高效的量化投资者。

广告定向中的用户分析

靳志辉（腾讯）
zhihuijin@gmail.com

嘉宾介绍

靳志辉先后毕业于北京大学计算机系计算语言所 (硕士)，日本东京大学 (统计自然语言方向博士)。目前，在腾讯科技北京有限公司工作，担任研究员。曾参与腾讯效果广告平台的研发工作，工作范畴主要涉及统计自然语言处理和大规模机器学习，以及把这些技术工具应用于腾讯海量的用户行为分析和广告定向中。

演讲摘要

腾讯拥有庞大的互联网用户和流量，如何挖掘这些海量的用户的行为数据以支持腾讯广告业务中的精准定向是腾讯互联网业务中的一个难题。在尝试精准广告定向的过程中，我们有几个任务需要解决：

- 如何使用高效的机器学习算法对海量的用户行为数据进行语义挖掘?
- 如何利用腾讯特有的社交行为数据挖掘用户的意图和兴趣?
- 直接产生兴趣数据的用户相对较少，而相似的用户可能会有相似的兴趣，能否通过相似用户计算，预测用户的兴趣?

本次演讲主要分享一下腾讯广点通广告定向团队在以上问题上做了一些积极的尝试所得到的一些初步成果。

基金评选平台之建立

郑义 (国立中山大学)
yihjeng@gmail.com

嘉宾介绍

郑义是美国爱荷华大学财务博士、CFA，专长为权益金融商品设计、投资组合理论与金融资讯系统开发，现任台湾中山大学副教授，曾任台湾期货交易所商品研发小组委员、保德信投信投资研究部副总经理、复华投信新金融商品部副总经理与资深谘询顾问、宝来证券新金融商品部专案谘询顾问等，具丰富的产官学经验。

演讲摘要

本团队运用 R 语言，将多个基金指标融合为单一综合指标，并藉此挑选较佳的基金产品，提供消费者简易且有效的基金评选平台，此外有鉴于退休规划之需求日益提升，本平台亦推荐数种严控风险的投资组合，做为长期投资之参考。

玩转三亿视频—数据分析在视频产业的应用

廖逸竹 (优酷土豆)
liaoyizhu@youku.com

嘉宾介绍

廖逸竹毕业于台湾大学工商管理系。现为优酷土豆集团数据分析部的高级经理, 负责以商业决策为导向的相关分析, 包括用户多屏行为、视频内容特点、用户与内容关连等相关议题。曾在台湾的管理咨询公司、商业银行、及雅虎台湾从事商业分析, 关注领域为客户区隔、测试设计、风险预测模型、及客户价值极大化分析。

演讲摘要

每日有 1.2 亿互联网用户与优酷土豆互动, 藉由对观众观看行为、影片搜索、评论互动等行为的解析, 得以了解不同类型影片的观众群、跨屏幕播放行为差异、影片关键情结点、内容偏好、UGC 播放的重要影响因子等议题, 进而将所获信息及知识转化为对公司、内容及产品运营的正面影响。数据分析对网络视频的影响正如火如荼发生, 玩转优酷土豆三亿个视频, 且听我们如何化数据为故事, 化故事为行动!

Hacking Models with R

张家齐 (Taiwan R User Group)
c3h3.tw@gmail.com

嘉宾介绍

家齐是一位热爱分析资料的工程师，热爱分析资料，建立模型，讨论数学。由于，早年喜欢作期货与选择权的程式交易，而纵观 Open Source 的软体中提供最多，跟投资策略分析相关资源的，大概就是 R 语言了。此外，在当时的台湾 Open Source 社群中，大多数也都集中在网站技术的讨论，鲜少有资料相关的社群与活动！因此，在 2012 年时，就找了高中学长 Wush，一起共同创办了 Taiwan R User Group 社群，以及相关的聚会！非常高兴能够有机会和社群的许多朋友们，一起组织聚会，一起讨论，一起成长。

演讲摘要

在这个「大资料」时代掘起的「掏资料潮」中，Data Mining 等相关的技术被应用的越来越广泛，也越来越深刻。不过，在真实的生活应用中，许多传统的 Modeling 技术，还是常常会遇到许多困难与挑战。因此，学会如何改写 Model 来因应环境的需求，就成为了资料分析人员很重要的技能之一。在这场演讲中，我将会介绍「R 中许多典型的资料模型」、「原始模型遇到的问题」、「模型背后的最佳化问题」、「如何改写模型并改写其 Solver」。

Multi-Cluster Detection

James Wicker(Chinese Academy of Sciences)
jewicker@gmail.com

嘉宾介绍

James Wicker graduated with a Bachelor’ s Degree in Physics from New College Florida in 1997. He went to graduate school at the University of Tennessee – Knoxville and earned a Master's Degree in Statistics in 2003 and a Ph.D. in Physics in 2006. His Ph.D. dissertation focused on developing new methods in regression and cluster analysis and applying them to analysis of physical systems. In 2007, he came to National Astronomical Observatories, Chinese Academy of Sciences in Beijing as a postdoctoral researcher. In 2009, he became an editor for the research journal Research in Astronomy and Astrophysics, which is also based at National Astronomical Observatories, Chinese Academy of Sciences. He is still doing research on developing new methods of statistical analysis, especially related to mixture modeling.

演讲摘要

A major challenge in mixture model analysis is determining the number of clusters present in a data set. I propose a new method to compute univariate mixture models that combines the advantages of both genetic algorithms and information scoring. Information scoring overcomes handicaps that are inherent in hypothesis testing, and as applied to mixture modeling, information scoring can overcome these ambiguities. I implement a restricted log-likelihood maximization procedure into a genetic algorithm that can accurately identify the number of clusters present in a univariate mixture model analysis situation. Repeated trials on simulated data sets demonstrate the accuracy and reliability of this method, and application to real data sets uncovers hidden structure in the underlying probability density functions.

基于 R 的程序化交易

景亮 (量邦科技)
jingl@quanttech.cn

嘉宾介绍

景亮现任量邦科技策略研发总监，毕业于中国科学技术大学 (物理学学士)，美国印第安纳大学布鲁明顿分校 (物理学硕士)，美国德克萨斯大学圣安东尼奥分校 (应用统计学博士)。曾任美国德克萨斯大学统计咨询中心高级分析师，具有多年统计学行业应用经验和丰富的量化金融投资研究实践经验。

演讲摘要

R 作为最流行的统计分析和数据可视化编程语言有其独特的优势和广泛的使用者基础，策略编写语言作为程序化交易策略开发中最核心的部分直接决定着开发的效率和策略的质量，如何把 R 融入程序化策略开发之中、充分挖掘其优势是一个值得深入研究的问题。量邦科技作为国内顶尖的量化投资平台开发商在这一领域做出了一定的尝试：

- 我们在程序化交易策略研发平台上，把 R 植入作为开发交易信号的编程语言；
- 上游无缝接入行情数据，下游对接信号汇总和策略表现分析模块。

如此一来，R 语言爱好者可以直接使用 R 语言开发程序化交易策略。

开发的血和泪，交易的冰与火

牟官讯（个人投资者）
lyxmoo@gmail.com

嘉宾介绍

牟官讯毕业于上海石油化工专科学校数据处理专业 (计算机应用方向)，多年的电信行业基础软件经验，过去曾从事电信级的应用和软件开发。现收集国内 A 股高频交易数据，从中进行用户行为的研究，投资开发了交易数据分析平台，从历史交易数据中发掘有价值的交易机会。

演讲摘要

开发高效率计算的代码技巧；如何提升算法代码的通用性；如何从历史交易数据中实时构建动态贝叶斯网络进行预测。

R 语言与金融大数据应用

张丹 (SupStat)
bsspirit@gmail.com

嘉宾介绍

张丹是 R 语言资深用户,《R 的极客理想》作者，系统架构师，曾开发多种不同类型的系统及应用，目前在量化投资领域创业中。张丹在其个人博客 (http://blog.fens.me) 原创了大量关于 R 语言和 Hadoop 大数据技术的文章。2013 年，他的 RHadoop 系列文章，在统计之都发表。他还是 Dataguru 培训讲师，教授课程《Hadoop 应用开发实战案例》、《Mahout 机器学习平台》。

演讲摘要

基于 Hadoop 存储证券的日内交易数据，通过 RHive 连接 R 语言与 Hive，建立相关性算法模型，在历史数据中回测，构建投资决策组合，并生成可视化结果用于展示。

Interactive Visualization with R

王亮博 (台大生医电资所)
ccwang002@gmail.com

嘉宾介绍

王亮博是台大生医电资所在学硕士。喜欢写 R、Python、统计与生物资讯。目前为 Taiwan R Users Group 工作人员及 Taipei.py 常客。

演讲摘要

近年来各种网路服务诞生，从要求画图好看，到要能与使用者互动。对于常见的图表而言，现在已有套件如 D3.js、ECharts 能提供解决方案。而 ggplot2 的强大功能已经为 R 使用者提供简洁又高质量的图表解决方案。如何将 ggplot2 的图表加入互动的元素，其中一个解决方案使用 gridSVG 作接口。本讲题将以 gridSVG 为出发点，介绍 grid 框架、SVG 互动语法，并示范如何于 R 中接合 D3.js 来实现互动图表。

它山之石可以攻玉：recharts 图形包

周扬 (AdMaster)
zhouyanga9@gmail.com

嘉宾介绍

周扬现就职于 AdMaster 数据研究院，主要负责数据分析、建模及其展示。R、Javascript 两栖码农，数据可视化爱好者，recharts 图形包重要参与者。

演讲摘要

数据可视化作为理解数据的重要媒介，让光秃秃的数据充满了活力和魅力。Echarts 是国内优秀数据可视化团队设计与实现的基于浏览器的图形库 (js 库)，已经获得广泛的使用和好评。然而 R 作为一个统计分析、数据建模和图形可视化的重要工具，由于其原生图形设备在动态可交互图形方面提供的支持有限，需要借助于浏览器作为数据展示平台实现图形的动态可交互。因此，recharts 基于将 Echarts 图形库引入 R 平台，为 R 用户群提供动态可交互图形的一个选择。并且通过与 knitr、Shiny、slidify 等优秀 R 包的连接实现了丰富和精彩的应用。

ggvis sneak peek

Hadley Wickham(RStudio)
h.wickham@gmail.com

嘉宾介绍

Hadley Wickham 是 RStudio 公司的首席科学家，同时也是美国 Rice 大学的助理教授。他开发了著名的 ggplot2, plyr, dplyr, shiny, devtools, stringr, ggvis, httr 等包，在 R 社区中很受欢迎。

演讲摘要

I'll give you a sneak peek at ggvis, the successor to ggplot2. Like ggplot2, ggvis allows you to describe visualisations declaratively. Unlike ggplot2, ggvis graphics are fundamentally of the web: they're built using html, js, and css. More importantly, ggvis graphics are fundamentally reactive. You can bind plot parameters to sliders and dropdowns, and visualise streaming data as it comes in.

大数据的新方向：公开同享趋势下的新数据产业

张尚轩、陈堰平 (SupStat)
vivian.zhang@supstat.com, yanping.chen@supstat.com

嘉宾介绍

张尚轩是 SupStat Inc(分公司为北京数博思达信息技术有限公司) 首席技术官和联合创始人。她负责美国市场的业务拓展和多边合作，并将美国大数据的软硬件解决方案带入中国市场。她在美国获得计算机/统计学双硕士学位，曾在布朗大学统计研究中心、斯隆凯特琳癌症中心、纽约石溪大学医疗中心等机构工作，参与多个重要的研究课题，并在影响因子第一名的 JASA 统计学杂志发表最新学术文章。她创立了纽约公开数据 Meetup, 专注于利用公开数据教授一般民众和技术人员数据分析方法，为社会创造透明高效的运作秩序，为企业提供最优质最好的数据源来发展业务。在不到一年之内，她为技术和数据社区提供了 80 余场免费的教学讲座。她亦是纽约数据科学学院的创始人，在纽约曼哈顿地区提供大数据专题教学，涵盖大量流行的数据分析和可视化编程工具 (R, Python, Hadoop, D3.js, Processing, Location data query 等)，帮助企业培训优质的大数据人才。

演讲摘要

- 分享美国政府公开数据的进展情况，以纽约，芝加哥，旧金山等主要城市为例，以具体的例子来展示政府是如何与一般民众沟通信息，鼓励创新和监督。以纽约为例，分享各类数据公开之前的几个步骤和需要的条件。
- 分享美国公司公开数据的使用情况，以 Oscar Health Care, On Deck Capital, Engima.io 为例，企业是如何从公开数据中受益获利以及发展出与众不同的竞争力。
- 分享美国的公司之间又是如何通过分享数据，创造新的价值和便利。以医疗体系为例，医生之间，医院之间，医疗体系之间实现了快速电子医疗档案的传递，可携带设备公司与医疗体系之间便捷的数据传递同享。
- 最后分享公开数据一些有趣的小数据产品，例如利用云图精确预测下一分钟是否下雨的移动应用产品，例如气象报告来卖天气保险产品等。和公开数据的比赛，例如纽约 Big Apps 比赛，每年一次鼓励全世界的开发者来比赛，把公开数据的价值释放出来。

R-Web: 大数据分析及导引云平台

陈景祥 (淡江大学)
steve@home.com.tw

嘉宾介绍

陈景祥是中华 R 软件研发暨应用协会 (www.carra.org.tw) 秘书长，淡江大学副教授。

演讲摘要

R 语言项目 (R-Project) 经过多年的发展，目前已是各国统计专业人士最常使用的分析工具。近两年来，随着大数据观念的普及，R 语言在数据科学的应用上也逐渐受到各个应用领域专家的关注，并已经成为主要的分析工具，虽然 R 本身包含了完整的程序语言功能以及众多的包 (package)，但是数据分析与应用人员未必都具有 R 的编程能力，因此开发一个只需鼠标点选即可完成分析任务的用户图形接口 (GUI) 就扮演了相当重要的角色。在 R 中原本就已经有若干图形界面的包可供选用，例如 R-Commander、Rattle，以及可供制作图形接口的 JGR、PMG、gWidget 等等，但是这些套件都各有缺点，在中文接口与计算结果的呈现也未必理想。R-Web 是第一个针对中文所开发的大数据分析及导引云平台，用户仅需使用计算机或行动装置的浏览器即可进行数据分析。R-Web 除了数据处理与一般统计分析之外，另外还包含数据挖掘、时间数列、广义线性模式 (GLM)、存活分析、以及结构方程模式 (SEM) 等多样的分析方法，对于初学者或对分析方法不熟悉的使用者而言，R-Web 也提供了分析目标导引系统，让用户可以经由问与答方式来找到适用的统计分析方法，提高分析效率及增加分析知识。

突破 R 内存瓶颈的若干技术

寇强 (华南统计科学研究中心)
qkou@umail.iu.edu

嘉宾介绍

寇强：微博：@Gossip useR，华南统计科学研究中心成员，信息学博士在读，研究方向为串联质谱的数据分析和软件开发。

演讲摘要

R 的内存计算一直被人诟病，除去利用近年兴起的 Hadoop 之外，R 众多的扩展包为解决 R 的内存瓶颈提供了各种思路，包括 hashing、硬盘缓存、保存重复计算结果、利用数据库后台等等。这里整理比较一下各种相关技术，提供若干性能测试，并加上一些个人的使用体会。

Large Scale Learning with R

吴齐轩（台大电机所）
wush978@gmail.com

嘉宾介绍

Wush Wu 是台大电机所的博士生，并且和宇汇知识科技合作，研发网路广告的推荐引擎。R 是 Wush 最熟悉的工具，平时工作几乎都使用 R 来完成，包括利用 R 爬资料、跑实验、分析数据到撰写报告和论文。也由于对 R 的喜爱，所以和家齐于 2012 年创立 Taiwan R User Group。实务经验上，目前 Wush 利用 Open Source R、Rcpp 和 pbdMPI 建立了分散式的学习系统来建立推荐模型，目前正在商转中。在了解业界的环境和挑战之后，目前则尝试将整个分析的流程系统化及自动化，建立一套能够持续改善推荐模型的 SOP，更期望将所谓让数据说话的思维落实到企业决策中。

演讲摘要

在资料爆炸的时代，运用大数据挖掘与探索商机是现在相当热门的议题。但事实上要驾驭大数据却不是件容易的事情，尤其在建立模型的部份，若在工具上没能跨过门槛，就很难在有限时间产生资料的价值。这次 Wush 将分享运用 R、Rcpp 和 pbdMPI 所开发的高效能的大数据运算平台，包含完成对超过 1 亿笔资料，仅花费 1 小时的建模经验，以及跨过分析门槛与挖掘知识的过程。Wush 除了介绍影响 R 运算速度的问题，以及实际克服问题的过程，同时也会分享如何运用系统化的概念创造资料价值的故事。

Big Data Analysis with RHadoop

丘右玮 (硕源资讯)
tr.ywchiu@gmail.com

嘉宾介绍

丘右玮 (David Chiu) 是硕源资讯 (numerinfo.com) 共同创办人，TW.R Officer，也曾经是趋势科技的工程师。David 是一位致力于提供 Data as a Service 的创业者与资料科学家，熟悉使用 Hadoop 进行巨量资料处理，暨长时间专注使用各式 Data Mining 技术从事资料分析；为台湾 Python 及 R 社群的忠实听众，喜爱参与社团交流与分享，希望能多了解如何使用 Python & R 让资料分析更简单上手。目前正在替 Packt 撰写 Machine Learning With R Cookbook 及编评 Bioinformaics With R Cookbook。

演讲摘要

谈到海量资料，通常大家脑海中联想到的就是使用 Hadoop 的 MapReduce 和 HDFS，但是撰写 MapReduce，则就必须要学会撰写 Java 或透过 Thrift 接口才能撰写。但 R 是否有办法运行在 Hadoop 上呢? 而使用 R + Hadoop，是否就真的能结合 R 强大的分析功能，分析海量资料呢?

本次讲题将介绍如何撰写 R 的 MapReduce 程式，并实际示范如何使用 RHadoop 进行海量资料分析。更重要的是，此次将探讨使用 RHadoop 是否为海量资料分析找到一盏明灯? 或者只是另一套实作方法而已?

构建高效率的数据流水线：在 R 中使用管道操作

任坤（厦门大学）
renkun@outlook.com

嘉宾介绍

任坤是厦门大学王亚南经济研究院金融硕士生，研究兴趣为计算统计和金融量化交易。

演讲摘要

在数据驱动的统计计算和数据分析中，对数据使用一连串指令来做处理与可视化是很常见的情况。但是由于传统的函数写法导致后调用的函数需要先写出来，所以一连串指令常常是多层嵌套、很长的表达式，既难阅读也难以维护。讲者编写的 pipeR 扩展包借鉴了 F# 语言中的管道操作符背后的思想，定义了三种适合 R 中使用的管道操作，可以方便地构建流水线式的数据处理过程，可以和 dplyr 等扩展包一起使用，大幅简化数据操作过程，使之变得清晰、易读、可维护。

科研角度下的 R 包开发

张晔 (中山大学)
zhangyet@gmail.com

嘉宾介绍

张晔是中山大学数学与计算数学学院计算数学专业在读硕士，华南统计科学中心研究人员。合作翻译 Financial risk modelling and portfolio optimization with R, Data mining with Rattle and R, Rcpp: Seamless R and C++ Integration 等图书。研究方向为生物统计。近期研究方向为生物调控网络。关注的技术点为 Rcpp 和 R 语言下的并行计算。

演讲摘要

科研工作需要将层出不穷的想法付诸实践，并在实验中不断修正想法。对于统计科学的研究人员来说，R 语言灵活高效，贴近统计学家的思维，同时又是一门正在发展的编程语言。演讲者将会结合自身的研究工作，讨论一下科研工作中的 R 语言开发。一方面，统计方法的算法描述往往是简单明了的，另一方面，统计科研中的编程工作并不简单。这是因为从算法描述到代码实现之间充满了大量的细节。主要的开发困难在于数据结构和接口的设计。而这需要软件工程的思维。要求一个统计学研究人员掌握计算机专业的专业知识略显苛刻，我们更推崇一种“统计学家提供原型，程序员进行优化改造”的工作范式。但为了可重复的研究，编写良好的 R 程序依然是非常重要的技能。R 提供了简单实用的面向对象系统 (S3 和 S4) 和一个强大的 C++ 语言接口 (Rcpp)，为我们的研究提供了极大的便利。

R 中大规模矩阵的奇异值分解与矩阵补全

邱怡轩 (普渡大学)
yixuan.qiu@cos.name

嘉宾介绍

邱怡轩毕业于中国人民大学统计学院 (硕士)，目前为普渡大学统计系在读博士，统计之都理事会成员。感兴趣的领域包括统计建模与计算，R 语言相关技术等，参与翻译了《R 语言编程艺术》《ggplot2：数据分析与图形艺术》《R 数据可视化手册》等书籍，是 R2SWF，showtext，rARPACK 等 R 程序包的作者。个人主页 yixuan.cos.name/cn。

演讲摘要

奇异值分解 (SVD) 及与其相关的特征值分解是统计模型中重要的代数运算工具，在传统的统计方法，如回归分析、主成分分析中有广泛的使用。R 中提供了 svd() 和 eigen() 等函数来完成相应的运算，然而当矩阵的维度较大时，其计算量通常会变得难以承担。对于一些特定的问题，我们只需要求解一部分的特征值 (例如最大的 k 个)，这可以通过 rARPACK 软件包中的相关函数来实现。本演讲将首先介绍 rARPACK 软件包的基本用法，并提供它与 R 中其他工具的性能比较。演讲的第二部分是 SVD 的一项有趣的应用，称为矩阵补全 (Matrix completion)，它与推荐系统、图片修复等具有紧密的联系。演讲中将以一个恢复受损图片的例子来介绍矩阵补全的基本原理和实现过程。

Data Analysis with R and Python

郭韦廷 (Taiwan R User Group)
waitingkuo0527@gmail.com

嘉宾介绍

郭韦廷是 Pandas (Python 用来做 Data Analysis 的套件之一) 的源码贡献者之一。Stackover Flow 上 Pandas 的 Top Answer 之一。

演讲摘要

近几年 Python 发展出了许多 Data Analysis 的套件，越来越多人开始使用 Python 做 Data 相关的服务。相较专门用来做 Data Analysis 的语言，Python 更易整合各式各样的资源，介接 Database、做个简单的 Web Dashboard、开 API 跟其他程式介接…这个 Talk 会介绍如何用 Python 来做 Data Analysis，还有一些 R 和 Python 的比较。

R 与 Office 的整合

李舰 (Mango Solutions)
lijian.pku@gmail.com

嘉宾介绍

李舰现就职于 Mango Solutions (China)，担任首席顾问，负责数据分析相关的咨询项目及公司产品中分析模块的开发。开源社区中 Rweibo、Rwordseg、tmcn 等 R 包的作者。中国 R 语言会议 (上海会场) 的组织者。《数据科学中的 R 语言》一书的作者 (即将由西安交大出版社出版)。

演讲摘要

R 是最强大且便利的统计分析工具，Office 是最为人熟知而随处可得的办公软件，如果一个分析人员的工作电脑上只能装两个软件的话，相信很多人会选择 Office 和 R。关于 Office 与 R 的整合，网络上存在很多很好的资源，比如 RExcel、R2PPT、ReporteRs 等。这些工具到底有哪些妙用？他们的实现机制到底是什么？如何使用才是最有效率的方式？本次报告将会对这些问题进行解答。在行业中，大部分的分析报告都是基于 Office 产生，尤其是 PPT 的报告，在可重复研究日趋火爆的今天，关于 Office 的自动化报告的方案并不常见。在本次报告中，演讲者还将会介绍一个自己编写的 R 包，可以通过 DCOM 的方式对 Office 中的对象进行自如地操作，并能自动解析 PPT 的各模块，以一个自动化报告的需求为例，介绍基于模板自动生成报告的流程。

数据分析在传统行业商业决策中的应用

林荟 (杜邦先锋)
longqiman@gmail.com

嘉宾介绍

林荟先后毕业于北京师范大学数学科学学院 (本科)，美国爱荷华州立大学统计系 (博士)。2009-2013 年曾为爱荷华州立大学兽医学院和商学院提供统计咨询服务；2013 年 5 月起任杜邦先锋全球总部市场部统计师，主要工作是领导建立商业预测模型、分析消费者行为数据和提供统计咨询。

演讲摘要

在大数据成为热点、电商高度发展的今天，数据分析在传统行业 (如农业) 商业决策中扮演的角色变化似乎被遗忘在舞台清冷的角落。本次演讲不打算搅和大数据这杯混水，而是立足于小样本建模分析在传统商业决策中的应用。

当然，这是另外一杯机遇和挑战并存的混水。具体说来主要讨论如下几点：

- 商业数据分析在传统行业的和电商邻域扮演的角色有什么不同?
- 数据分析如何帮助商业决策?
- 几个需要注意的问题
- 模拟应用案例：用 Group Lasso 逻辑回归构建评分系统
- 机遇和挑战

小而美的数据产品

欧阳鹤（魅力惠）
iamoyh@163.com

嘉宾介绍

欧阳鹤毕业于复旦大学广告系。曾就职于路易威登零售、顾客零售营销部门。目前在奢侈品闪购网站魅力惠从事网站数据分析工作。谷歌分析认证网站分析师。受统计之都的影响于 2012 年开始自学 R 语言。参加过 2012 年与 2013 年的上海 R 语言会议。兴趣：信息图表设计，可重复性研究与自动化报告。

演讲摘要

读大学时我从电气工程与自动化转专业到了广告系。朋友说，是从 Hard 模式跳到了 Soft 模式。工作后，我从零售与营销转到了数据分析。有人说，Soft 调回了 Hard。其实，文理相长。数据分析工作可以是技术与艺术的完美结合。数据产品：以“产品经理”和用户的角度去思考。小而美：有效的信息沟通，”不要炫的，要有效的”。前辈的金玉良言，Edward Tufte, Stephen Few 应用案例: 魅力惠是以闪购活动的形式来组织销售。每个活动持续 1 到 2 周。又快又轻又好的活动销售报告是频繁而核心的业务需求。Ubuntu+R+Git 布环境, shiny 搭骨架, RMySQL 读写数据，plyr 与 reshape2 清理数据，ggplot2 与 ggmap 绘图，knitr 转换成报告网页，Google Analytics 监测应用访问及使用。

R 与企业级数据挖掘

刘思喆 (京东)
sunbjt@gmail.com

嘉宾介绍

刘思喆，京东数据部推荐团队算法组经理，主要负责商品推荐算法、用户行为预测等内容。在加入京东前，曾供职于亚信联创 BOC、神州数码思特奇 DSS，为电信运营商提供数据挖掘及业务咨询等顾问服务。自中国人民大学统计学院毕业 9 年来，一直追求为企业提供高效、完备的数据解决方案，尤其在统计分析、预测分析、数据可视化、机器学习、文本挖掘及社交网络等领域。博客：http://bjt.name。

演讲摘要

京东 (JD.com) 是中国最大的自营式电商企业，拥有超过 1 亿的注册用户，13 大类约 4,020 万 SKUs 的丰富商品，业务覆盖采销、平台、物流、售后、支付等全流程。在此复杂的场景下，京东生产了海量的结构化和非结构化数据，例如用户级别的购买、浏览、点击、搜索、评论行为数据，以及商品、商户、供应链等数据信息。R 语言作为方便快捷的可视化及建模工具，在一些特定问题上有效的支持了业务的需求。本文将从京东复杂的数据环境讲起，介绍以 R 语言和 Hadoop 为核心的数据建模的技术框架，并分享几个 R 语言实践的案例。

Integrated Pipeline for Systems Pharmacology in R/Bioconductor

肖楠 (中南大学)
road2stat@gmail.com

嘉宾介绍

肖楠是中南大学数学与统计学院统计学系在读博士，统计之都论坛 R 语言版版主。《R 语言实战》、《ggplot2：数据分析与图形艺术》、《R 数据可视化手册》等书籍译者；protr、Rcpi 等 R 包作者。关注领域为统计机器学习、化学信息学与生物信息学、定量与系统药理学。

演讲摘要

Multiscale molecular representation and modeling is a fundemental problem in systems pharmacology research. We developed R/Bioconductor packages and web apps emphyyyasizing the comprehensive integration of bioinformatics and chemoinformatics into a molecular informatics platform for drug discovery. We will share the experience and pitfalls during the package development process.

R 在新药研发中的应用

杨环 (Mango Solutions)
huan.a.young@gmail.com

嘉宾介绍

杨环现就职于 Mango Solutions (China)，担任咨询顾问。毕业于厦门大学和伦敦政治经济学院。

演讲摘要

一款新药的平均研发时间达到十年之久，耗资通常 10 亿美元之巨，整个研发过程中的任何决策都至关重要。尤其在最近几年，很多大药厂纷纷遭遇专利保护到期的困境，而新药研发的进度也越来越缓慢。在这样特殊的时期，在 FDA 的引导和各大药厂的实践下，新药研发中的建模和模拟成了药厂摆脱困境的良药，而这个领域最受欢迎的工具就是 R。演讲者将会结合 Mango Solutions 为各大药厂提供服务的经验，介绍新药研发尤其是建模和模拟的流程，展示各类统计模型和数学方法在新药研发中的应用以及系统和工具的实践，尤其是 R 在其中所起到的关键作用。

Combining R with Psychology——An illustration with SEM

江歌 (University of Notre Dame)
gjiang2@nd.edu

嘉宾介绍

Ge Jiang is currently a PhD student at University of Notre Dame, his major is Quantitative Psychology and minor is Applied Computational Mathematics and Statistics. His research interests lied in psychometrics and factor analysis and I kind of enjoy the pleasure of being a "ma nong" and want to apply these statistical methods more inside psychology field.

演讲摘要

R is an advanced software that has been adopted and created many disciplines, including biostatistics, econometrics, psychometrics, and social statistics. In quantitative psychology, it plays a crucial role in conducting simulation and testing hypotheses. This topic mainly presents how R is adopted in SEM to test model fit and developing new test statistics.