第七届中国 R 语言会议（杭州会场）

主办：

统计之都

协办：

阿里巴巴数据技术与产品部

天猫技术部

杭州师范大学

阿里巴巴复杂科学研究中心

阿里巴巴商学院

杭州市电子商务与网络经济研究中心

赞助：

阿里巴巴数据技术与产品部

天猫技术部

2014 年 11 月 29 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6000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2014年做的“使用何种编程或统计类语言进行分析、数据挖掘及数据科学的工作”的调查中，R 以49.0%的得票率荣登榜首，力压SAS、Python和SQL（`www.kdnuggets.com/polls/2014/languages-analytics-data-mining-data-science.html`）。连续三年位居榜首。

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。

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。

*统计之都网址：http://cos.name/

会议议程

时间	内容	嘉宾	主持人
9:00-9:30	会议致辞	闵万里	沈羽
9:30-10:00	从大数据分析迈向认知计算的应用实践	曹恒	
10:02-10:32	Big data Machine Learning: Status and Challenges	林智仁	
10:35-10:50	**茶歇**		
10:50-11:20	医疗大数据的云应用	郑光甫	沈羽
11:22-11:55	R 语言与大数据云端运算	陈景祥	
上午会议结束			

Track A ：数据、R 和工业应用 （恕园二号楼 1203）			
13:30-14:15	R-web：大型多人在线数据分析协作平台	林祯舜	林祯舜
14:20-15:05	初探基于 R 语言的电信用户数据挖掘应用	林倩莹	
15:10-15:55	Learning To Rank in Tmall Search	王勇	
16:00-16:45	用算法驱动业务：天猫超市数据化运营	李博	
Track B：R 工程开发（恕园二号楼 1205）			
13:30-14:15	大规模 CTR 预测系统在推荐中的应用	朱剑锋	李舰
14:20-15:05	R 中的数据可视化	魏太云	
15:10-15:55	R 与工程开发实践	李舰	
16:00-16:45	母婴潜客挖掘的数据之法	许亮	
Track C：用 R 进行数据产品开发全体验（恕园二号楼 2203）			
13:30-14:15	分析师们用得着的 R 包	陈逸波 郝智恒	周扬
14:20-15:05	用 node.js 和 R 集成开发数据产品原型	周扬	
15:10-15:55	基于地图的数据可视展现	周宁奕	
16:00-16:45	地图可视化中的数据挖掘	郎大为	
下午会议结束			

主办/协办单位：

统计之都，阿里巴巴数据技术与产品部，天猫技术部，
杭州师范大学，阿里巴巴复杂科学研究中心，阿里巴巴商学院，
杭州市电子商务与网络经济研究中心

组委会成员：

沈羽，李舰，魏太云，郝智恒，陈逸波

杭州师范大学地图

从大数据分析迈向认知计算的应用实践

曹恒[1,*]

[1]IBM中国研究院

摘要

数据是当今社会发展的重要战略资源,从数据中提取高价值信息的大数据分析技术正在以前所未有的速度改变着商业与个人生活，将人工智能和大数据分析相结合的认知计算更是为下一代信息系统提供了自学习的能力。本次报告将介绍IBM研究院在运用大数据分析及认知技术应用于实际业务问题的一些创新尝试。

*IBM中国研究院上海分院院长，多年IBM开发和研究中心的工作经验。目前担任负责IBM全球研究实验室的数据分析业务，上海实验室执行官

Big-data Machine Learning: Status and Challenges

林智仁[1,*]

[1] 台湾国立大学

摘要

Big-data machine learning has emerged as an important research topic because data larger than a machine's capacity are now very common. However, there are many challenges in applying big-data machine learning. First, most traditional machine learning algorithms were designed to run on a single computer. Second, sub-sampling data to one machine for analysis is always an option, so when distributed machine learning is preferable is an issue. After an overview of these challenges, we discuss distributed machine learning methods. In particular, a Newton method for large-scale logistic regression is given as an example. We also briefly review other existing developments and show how they are used in practical applications. Finally, we argue that big-data machine learning involves many issues other than algorithms. Programming environments, systems, and the application workflow must be considered together for a successful big-data machine learning project.

[*] Chih-Jen Lin is currently a distinguished professor at the Department of Computer Science, National Taiwan University. He obtained his B.S. degree from National Taiwan University in 1993 and Ph.D. degree from University of Michigan in 1998. His major research areas include machine learning, data mining, and numerical optimization. He is best known for his work on support vector machines (SVM) for data classification. His software LIBSVM is one of the most widely used and cited SVM packages. For his research work he has received many awards, including the ACM KDD 2010 and ACM RecSys 2013 best paper awards. He is an IEEE fellow, a AAAI fellow, and an ACM distinguished scientist for his contribution to machine learning algorithms and software design. More information about him can be found at http://www.csie.ntu.edu.tw/c̆jlin

医疗大数据的云应用

郑光甫[1,*]

[1] 台北医学大学

摘要

1.智慧型健康追踪系统

2.医疗云端资料库

3.健康数据是大数据

4.波特的大创意

5.如何解决波特的问题？

6.建立「个人健康评估模式」

7.建立「共病共药分析模式」

8.应用在制药产业的布局策略

*台湾中央大学前任副校长，现任台北医学大学生物统计中心主任，中华R软件开发与应用协会会长。专长包括生物统计，基因统计，财经预测，无母数回归方法。经历包括：美国佛罗里达州州立大学统计博士、美国纽约州立大学助教授、清华大学客座副教授、中央大学统计研究所特聘教授（现职）、美国哈佛大学公卫学院客座教授、中央大学教务处教务长、中央大学副校长、国家科学委员会自然科学咨议委员会委员、科技顾问组顾问、行政院教育部私立大学院校中程校务发展计划审查委员、教育部私立大学校院整体发展奖助审查委员、教育部科技大学评鉴委员、教育部学术审议委员会委员、行政院体育委员会委员、中国医药大学生物统计中心讲座教授兼主任、台北医学大学生物统计中心讲座教授兼主任。学术奖励包括：国立中央大学杰出教师、国立中央大学卓越教师、国科会杰出研究奖，优等奖、中国统计学社荣誉奖章、国际统计学院院士、教育部学术奖、泛华统计协会理事长、中华R软体研发暨推广协会理事长。

R语言与大数据云端运算

陈景祥[1,*]

[1] 淡江大学

摘要

除了商业属性的SAS 与SPSS 之外，R语言(R-Project) 已是各国统计专业人士最常使用的统计分析语言及软件。尽管如此，由于核心程序代码的限制，R软件仅能加载小于硬件内存大小的数据量。若再加上各类统计计算过程中所产生的临时变量所占的内存，若不做特殊处理，目前的R语言并不适合用于大数据运算分析。

幸运的是，除了像Revolution R等商业解决方案之外，R语言已经有不少专门的软件包可以解决大数据数据处理的问题。此外，R软件也有许多云端与平行化处理的包。我们将介绍R软件在大数据运算、平行运算、与云端计算的相关套件，以探讨R语言在实际学术与工商业应用的的可行性。

*台湾淡江大学统计系教授，中华R软件开发与应用协会秘书长。美国佛罗里达州立大学统计学博士，现为台湾台北淡江大学统计系专任副教授，专长为可靠度分析(Reliability Analysis)、统计计算(Statistical Computing)、数据挖掘(Data Mining)、与时间数列分析(Time Series)，此外陈教授还精通UNIX 操作系统、平行运算、大数据分析、与R 语言。陈教授是淡江大学NetStat 在线统计运算网站作者(http://netstat.stat.tku.edu.tw)，这个网站是中文第一个互联网上的统计分析平台。陈教授也是台湾第一位出版R语言专着的统计学者，他的著作「R 软件：应用统计方法」在台湾是学习R语言必备的参考书籍。除了学术研究，陈教授曾担任台湾锦华金融信息软件公司的顾问，将统计方法与软件工程结合，在金融信息方面有非常杰出得成果。此外在学科技能竞赛方面，陈教授曾担任2005 年芬兰国际技能竞赛(Word Skills Competition)信息与网络技术类国际裁判，2007 年日本国际技能竞赛(Word Skills Competition)信息与网络技术类教练，在提高学生计算器技能方面有很大的国际影响力。最近陈教授致力于R软件的应用及普及，目前已经开发了通用的数据分析与挖掘云平台R-web，陈教授是R-web平台技术研发团队的领军人物，实务应用及学术成果丰硕，是大数据应用方向的专家。

R-web：大型多人在线数据分析协作平台

林祯舜[1,*]

[1]先锋信息科技/辰智咨询

摘要

如果你是有经验的数据分析人员(或者数据科学家、数据挖掘工程师)，你会发现要将一个数据进行分析并得到有洞察力的建议，这是一个知识发现的过程，这个过程中有三个重要的组成部分: 数据、工具、人，数据在工具中经由人的探索及验证的过程逐渐积累形成知识，这个过程在未来会经由数据的交叉复用、分析方法(或算法)的镶嵌以及分析人员的协作与交流而加快知识的积累与生产，因此工具(或平台)的协作及弹性就越来越重要，这个报告要介绍第一个由华人统计学家基于R语言研发的数据分析云平台，阐述这个协作平台的设计理念及未来的愿景，希望这个平台在未来能让更多人能学习数据分析，理解分析思维，为培养未来的数据科学人才贡献一份心力。

*林祯舜博士是数据科学及营销科学方面的专家，毕业于人民大学统计学院并获得博士学位，在企业界，目前担任信息技术咨询公司的总经理，在学术界，目前是兰州商学院及吉林大学的兼职教授。林博士学术领域的研究方向包括数据挖掘，机器学习，统计计算，网站效果测量与点击流数据分析。林博士的学术论文是关于互联网点击流的模型应用，这个模型是第一个应用在互联网媒体规划的基础模型，主要的贡献是打破互联网媒体规划和传统电视媒体规划使用相同模型的迷思，对互联网媒体规划的理论与应用找出一个突破口，目前全球互联网监测公司所使用的媒体规划模型，都是在林博士发表的论文基础上加以改进并产品化。这篇论文在2010年五月被美国网站分析协会(Web Analytics association)选为网站分析领域最需要阅读的14篇论文之一。国际顶尖广告研究期刊，广告研究学报(Journal of Advertising Research)在2011创刊50周年的特刊上，针对互动网络的专文中，特别提到林博士论文在互动媒体规划方面的贡献，这是该领域被提到的四篇文章之一。林博士相关的学术研究论文发表在Journal of Advertising(JA)，Journal of the American Society for Information Science and Technology (JASIS&T)，Information Research，营销科学学报等期刊。

基于R语言的电信用户数据挖掘应用

林倩莹[1,*]

[1]SupStat

摘要

2013年我们移动通信用户总数超过10亿大关，而预计今年用户总数将会增加至20亿。如此庞大的用户量必然产生了巨大的数量。作为数据挖掘应用的重点行业，电信业如何利用这个庞大的原始数据，针对不同的客户采取不同的营销策略，从而为公司增加收益呢？

对于电信公司来说，用户新增入网之后，将会经过三个阶段，分别是新增入网时的成本投入阶段，之后到成熟稳定的价值贡献阶段，最后用户就会逐渐衰退流失，为零负收益阶段。显然，用户在稳定期时收益贡献率最高，那如何判别新增入网的用户之后是否会成为这个时期的稳定用户呢。本报告通过观察，先定义稳定用户为在网时长24个月以上用户，作为用户的稳定性标识，从800万数据中抽取1

[*]SupStat数据科学家

Learning To Rank in Tmall Search

王勇[1,*]

[1] 天猫技术部

摘要

在搜索引擎服务中，对返回结果的排序是非常重要的一个环节。早起的排序规则通常是基于对业务的了解和不断尝试而手动定制的。随着机器学习领域的不断发展，最近几年排序学习（Learning to Rank）越来越受人关注。排序学习致力于通过机器学习手段自动的学习排序规则，相比手动制定排序规则，自动的排序学习可适应性更强，在目标设定合理的情况下，能保证学习到的规则尽量靠近理想目标。本次分享我们将简要介绍在天猫商品搜索上实施的排序学习框架。

*天猫技术部，搜索研发专家

用算法驱动业务：天猫超市数据化运营初探

李博[1,*]

[1] 天猫技术部

摘要

在天猫超市这个业务背景下，初步探索了如何利用算法来解决业务难点并驱动业务模式升级，内容包括：算法选品、销量预测、流量匹配等。

*天猫技术部，数据挖掘团队

大规模CTR预测系统在推荐中的应用

朱剑锋[1,*]

[1] 京东

摘要

CTR（点击率）预测是推荐算法中的重要环节，也是个性化推荐引擎的核心组件，对召回商品CTR的准确预测会直接提升用户体验和网站收益。与其他语言/平台相比，R语言提供了强大的数据预处理、模型构建及评测能力，完全可以支持工业界级别的数据规模，并具有灵活的扩展性。

[*]朱剑锋，2010年毕业于北京大学元培学院统计学专业。曾就职于深圳华大基因研究院，担任核心算法研究单元负责人，主要研究方向包括宏基因组、群体遗传学、序列比对、产前遗传筛查等，具有丰富的统计建模，机器学习经验，研究结果发表于Nature等期刊。现就职于京东商城推荐搜索组，主要工作为推荐系统中的算法设计、实现及优化，使用R语言5年。

R中的数据可视化

魏太云[1,*]

[1] 统计之都

摘要

R 官方网站的第一句话是这样介绍R 语言的：“R 是一个用于统计计算和绘图的自由软件环境。”这句话正好突出了R的两大特色：数据分析和数据可视化。经过长年的开发和完善，目前R主要支持了四套图形系统：基础图形（base）、网格图形（grid）、lattice 图形和ggplot2。除了简要介绍R图形技术之外，本演讲还会注重展示一些整合了数据分析和可视化的综合案例。

*统计之都现任理事会主席，感兴趣的领域是统计建模、量化投资、数据可视化，合作翻译出版了《ggplot2：数据分析与图形艺术》、《R数据可视化手册》等书籍。

R与工程开发的实践

李舰[1,*]

[1]Mango Solutions

摘要

R语言最初因为其矩阵运算和内置统计模型、作图引擎的优势兴起于学术界，经历了一段曲高和寡的日子。但本质上，R的风格并不学术，反而由于其深刻地权衡了开发时间与运行时间的原因，非常符合业界的习惯，经过多年大浪淘沙，各行各业都把R当成了数据科学的首选工具，其开发快、资源丰富、入门容易、功能强悍等特点在业界得到了共鸣，产生了非常多的成功案例。

但是另一方面，由于R的调试功能非常弱，所以很容易被认为不适合大规模的工程开发。实际上，R由于其平台的灵活性，可以借助很多成熟的工具和软件工程的思路，有着很多工程领域最佳实践的经验。演讲者结合自己多年来在各行各业中使用R的经验，尤其是大规模系统的开发经验，介绍R工程开发的实践，并结合具体案例进行演示和说明。

*李舰，毕业于中国人民大学统计学院（本科）和北京大学软件与微电子学院（研究生），现就职于Mango Solutions，担任中国区数据总监，负责数据分析相关的咨询项目及公司基于R语言的产品的开发。Rweibo、Rwordseg、tmcn等R包的作者。《数据科学中的R语言》的作者，《R语言核心技术手册》的译者。邮箱：lijian.pku@gmail.com，主页：http://www.jianl.org。

母婴潜客挖掘的数据之法

许亮[1,*]

[1] 天猫

摘要

2012年，美国一名男子闯入他家附近的一家零售连锁超市塔吉特内抗议：你们竟然给我17岁的女儿发婴儿尿片和童车优惠券。一个月后，这位父亲前来道歉，因为这时他知道自己的女儿的确怀孕了。想知道其中的奥秘吗？一起来和我们探索母婴用户数据的奥秘吧。让大数据解开这个谜团。

*天猫数据产品专家

分析师们用得着的R包

陈逸波[1,*]　　郝智恒[2,†]

[1]阿里巴巴数据技术产品部
[2]阿里巴巴数据技术产品部

摘要

在与分析师们的日常合作交流中，我们发现，除了制作各种统计报表之外，大量的工作都涉及到对统计指标进行不同维度不同角度的测算（以寻找数据变化的原因），并形成分析报告供业务方参考。探索数据的过程是有趣而充满挑战的，但是重复的鼠标操作又略显乏味，此时如果使用R作为分析工具，会带来不少的便利。以Rstudio为首的R社区已经开发了大量的R包来帮助进行可重复的分析研究工作，本次报告就是介绍这些工具的常规用法，以及我们在实际工作中的一两个例子。

*阿里巴巴数据技术产品部，资深数据挖掘工程师
†阿里巴巴数据技术产品部，资深数据挖掘工程师

利用R和NodeJS实现数据产品原型

周扬[1,*]

[1]JDPOWER

摘要

R作为数据分析及处理的重要工具，成为越来越多数据分析师的宠儿。本次分享将围绕R的服务器应用，描述R的服务器应用以及在线服务上的可选方案，并利用本人在实际工作中运用的利用Node.js作为服务器后端，并通过RIO（nodejs包）调用R代码，根据用户的操作生成所需要的结果文件（JSON），再通过前端的展现库d3js 和ECharts进行图形渲染，以实现简单的可视分析应用。

*JDPOWER 数据分析师,著名开源可视化工具recharts的作者。

基于地图的数据可视展现

周宁奕[1,*]

[1] 淘宝数据平台

摘要

3D地图有很大的商用和分析价值，以下几个问题可能是你都问过的。怎样画地图？哪些数据哪些服务，如何让地图变3d，如何在地图上画实时数据可视化、如何在地图里增加时间的维度，如何让地图增加隐藏的交互？这个演讲，将解答你的这些疑问。

*淘宝数据平台

地图可视化中的数据挖掘

郎大为[1,*]

[1]SupStat

摘要

对于地图数据的可视化一直是数据可视化的焦点之一，最基本的地图可视化就是为不同的区域填充深浅不同的颜色。在R中，我们可以使用一些数据挖掘的技术，思路(loess,knn)来更好的绘制这类基本的地图，做出连续的颜色变化，多选项时的混合颜色

[*]SupStat数据科学家