9 第九届 中国R语言会议（北京） 会议手册

中国人民大学

5月27日 — 5月29日

大统计与数据科学联合会议

联合主办

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

北京大学统计科学中心

伦敦政治经济学院统计学系

统计之都

百分点集团

共同协办

International Chinese Statistical Association 泛華統計協會

中国统计学会

欢迎辞

十年的时间很长，苏轼已两鬓成霜，木兰也百战回乡；十年的时间很短，罗隐还未成名，苏武仍在牧羊。自其变者而观之，觉宇宙之无穷，如级数般无穷；自其不变者而观之，识盈虚之有数，似统计般有数。这十年来，统计之都遇到的可能性可谓无穷，R 语言会议的滥觞却是有数。十年间，统计成了显学，大数据持续火热，也有很多新名词突然诞生然后烟消云散。还好我们坚持下来了，如今的 R 语言会议早已超越 R 语言之技，但又秉持着最初那群人用 R 砸向理论与实践间壁垒的发心。如今的统计之都仍然和当年一样坚持最初的理想，但今天主持大局的年轻人已是远胜当初。当然，当年的年轻人今天也还在，趁此良宵，携酒相邀，豪兴不浅，恭迎盛典。

九者阳之极也，R 语言会议经过这么多年的奇幻生长，终于迎来了第九届，也开始进入稳定传承和有序组织的节奏。十年利剑可成，统计之都磨砺了许久，在十年生日之前终于实现了正规化的运营，也开始探索新的目标，为中国的统计事业贡献自己的力量。5 月 27—29 日，“第九届中国 R 语言会议（北京）”将与“第七届中国人民大学国际统计论坛”和“2016 百分点数据与价值国际论坛”联合举办，共同打造迄今为止中国最大的统计盛会“大统计与数据科学联合会议”。欢迎您的莅临，与我们一同分享这场数据的盛筵。此外，欢迎各位老朋友前来参加统计之都成立十周年的庆典，共听清风，同赏明月，把酒言欢，共图一醉，岂不快哉！

统计之都敬上

2016 年 5 月 19 日

目录

大统计与数据科学联合会议介绍

对统计学而言，这或许是最好的时代。

信息技术的蓬勃发展让海量数据触手可及。大数据时代已然到来。在过去的许多年里，我们经历了大数据相关产业的高速发展，诸如生物医药、金融、移动互联网、车载信息技术等。毫无疑问，数据科学的巨大需求正在悄然兴起，统计学正大有可为。

对统计学而言，这或许是最坏的时代。

大数据相关科学与产业的发展太过迅速，随之而来的是具有复杂结构的海量的庞大数据集。在许多情况下，现有的统计模型无法拟合非结构化数据。通常，我们也无法算出大规模数据下的最大似然估计。在方兴未艾的大数据相关产业与科学面前，我们的领域知识显得如此苍白无力。放眼世界或立足中国，统计学都正面临着严峻挑战。

因此，我们需要一个交流观点、讨论想法的平台。为此，中国人民大学、统计之都、北京大学以及百分点集团倾力合作，联合举办迄今为止中国最大的统计大会。本次大会是由下述三个部分组成的联合会议：

- 国际统计论坛（The International Forum on Statistics, IFS）始于 2004 年。它是由中国人民大学统计学院与伦敦政治经济学院统计系共同赞助的两年一次的国际会议。该会议旨在联合国内外活跃学者，已经成为中国最有影响力的国际统计学会议之一。今年将迎来第七届国际统计论坛。
- 中国 R 语言会议（The China-R Conference）始于 2008 年，由统计之都（Capital of Statistics, COS）主办。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 会议起始于 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、技术大咖于一堂，使各界参会者都得到了充分的交流。作为国内最大的数据科学会议，R 会议已免费服务数万参会人员。2015 年，第 8 届中国 R 语言会议在北京、上海、广州、南昌、西安、武汉六个城市分别举办，其中北京大学举办的北京会场参会者逾 4000 人。今年将迎来第九届中国 R 语言会议。
- 数据与价值国际论坛（The Data and Value Forum, DVF）始于 2014 年。它是由北京大学商务智能研究中心（Business Intelligence Research Center, BIRC）发起的年会。该会议旨在推进统计学在生产中，特别是移动互联网、车载信息技术以及量化投资等方面的应用。今年，中国大数据产业坚定的践行者与领军者——百分点集团将与北京大学商务智能研究中心携手, 举办 2016 百分点数据与价值国际论坛。

本次联合会议的主题是：大统计与数据科学。会议的目标是多方面的：首先，我们希望建立国内外活跃研究人员的联系，以此促进研究观点的交流与统计理论的进步。其次，我们希望建立统计学研究人员与从业人员的联系，尤其关注于中国的大数据相关产业，从而让统计理论与产业实践相互促进、相得益彰。为此我们有幸得到百分点、量邦科技、以及考拉征信等多家企业合作伙伴的鼎力支持！

我们希望以本次会议为契机，推动统计学在数据科学领域的贡献，让统计学在数据科学领域发出强有力的声音。会议主题“大统计与数据科学”也正是由此而生。本次会议所探讨的话题将是广泛的，我们欢迎任何与数据分析相关的问题——无论是理论研究或是实际应用。为促进统计学在大数据产业中的应用，本次会议尤其欢迎源于大数据相关产业实际问题的研究。

主办机构

中国人民大学统计学院

人大统计学科始建于 1950 年，2003 年建院。全国重点学科，2007 年教育部二级学科评估排名全国第一，2012 年教育部统计学一级学科评估排名全国第一。拥有统计学一级学科博士点和博士后流动站，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，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）三个本科专业，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。

北京大学商务智能研究中心

北京大学商务智能研究中心依托北京大学光华管理学院，关注基于互联网的大数据研究与应用。中心尤其关注中文文本、网络结构、以及位置数据相关的科研课题。中心为学者提供相关数据资源，为企业提供相关分析方法，为学者和企业合作搭建一个有效的平台。

统计之都

统计之都（Capital of Statistics，简称 COS，网址 http://cos.name/）成立于 2006 年 5 月，是一个旨在推广与应用统计学知识的网站和社区。统计之都发源于中国人民大学统计学院，现由世界各地的众多志愿者共同管理维护，旨在搭建一个开放的平台，使得科研人员、企业数据分析人员和统计学爱好者能互相交流合作。统计之都的治站格言是“专业、人本和正直”，力图在此格言指导下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。

百分点集团

百分点集团成立于 2009 年，是中国领先的大数据技术与应用服务商。作为企业级大数据技术与应用的践行者，百分点拥有业界顶尖的研发团队、完善的研发体系以及成熟的商业实践。经过长期的技术沉淀和实践积累，百分点已经完成了大数据技术、管理、应用三位一体的大数据解决方案体系的构建。2015 年 9 月，百分点正式发布了全球首款企业级大数据操作系统 BD-OS。在技术、应用以及数据等方面的全面优势，使得百分点获得了超过 2000 家互联网及实体企业的广泛认可，涵盖了制造、金融、汽车、零售、快消、电商、媒体、政府等行业龙头企业。

赞助商介绍

白金赞助

量邦科技

北京量邦信息科技股份有限公司（简称量邦科技）是一家专注于大数据应用和科学云计算的新三板挂牌企业（股票代码 835352）。公司产品包括完备的经济金融数据库、数据云分析平台（开矿网）、投资策略开发平台（大宽网）、数据交易平台（淘数网）、投资软件产品线、投资机构服务和零售服务、高校创新实验室等。公司曾获“2015 金融服务行业 TOP10“等荣誉，是国家高新技术企业、中关村金种子工程企业、山东青岛财富管理基金业协会会员。

考拉征信

考拉征信——信用创造价值。

考拉征信是互联网大数据征信业的先行者、领导者和践行者！作为独立的第三方信用评估及信用管理机构，考拉征信是国内仅有的几家同时获得央行许可开展企业征信和筹备开展个人征信业务的征信机构之一，拥有国内首个专注于大数据征信模型研究的专业实验室，是国内第一家征信产品被银行接入的征信机构，并成为国内首个率先推出职业雇佣征信平台的征信机构。目前，考拉个人信用分、商户信用分等征信产品已成为金融信贷、职场招聘、租车、租房等领域信用评估的重要参考，受到光大银行、钱隆贷、保驾等合作伙伴的高度认可。其中，商户信用分更是获得了由国家院士、征信专家学者、金融行业风控专家等 11 位专家组成的专家评审组一致好评，具有重要的实用价值与社会意义，可应用于更多的信贷机构使用。

金牌赞助

懒投资

懒投资隶属于北京大家玩科技有限公司，2014 年 9 月上线运营，A 轮融 2100 万美元，来自策源创投、源码资本、福布斯富豪夏佐全先生。累计交易金额超 100 亿，为用户赚取 3.1 亿收益，注册用户超百万，无一例逾期。懒投资主要对接应收账款保理、融资租赁和消费金融等优质债权资产。2015 年 12 月，国资参股背景的大型担保机构中盈盛达在香港上市，懒投资作为其基石投资者受邀现场敲钟。这是国内首例互联网金融公司以基石投资者身份亮相国际资本市场。

Tableau

Tableau Software（纽交所代码：DATA）致力于帮助人们查看并理解数据。Tableau 帮助任何人快速分析、可视化并分享信息。超过 39,000 家客户通过使用 Tableau 在办公室或随时随地快速获得结果。数以万计的用户使用 Tableau Public 在博客与网站中分享数据。登录 http://www.tableau.com/zh-cn/products/trial 下载免费试用版，了解 Tableau 能够给您带来哪些帮助。

微软

微软是美国一家跨国电脑科技公司，以研发、制造、授权和提供广泛的电脑软件服务为主。总部位于美国华盛顿州的雷德蒙德，最为著名和畅销的产品为 Microsoft Windows 操作系统和 Microsoft Office 办公室软件，以及 XBOX 的游戏业务。微软是美国《财富》杂志 2015 年评选的全球最大 500 家公司的排行榜中的第 95 名。公司于 1975 年由比尔・盖茨和保罗・艾伦创立。初期主要为 Altair 8800 发展和销售 BASIC 直译器，在 1980 年代中期凭借 MS-DOS 在家用电脑作业系统市场上取得长足进步，后来出现的 Windows 使得微软逐渐统治了家用桌面电脑作业系统市场。同时微软也开始扩张业务，进军其他行业和市场，建立了 MSN 网站，在计算机

硬件市场上，微软商标及 Xbox 游戏机、Zune 和 MSN TV 家庭娱乐设备也在不同的年份出现在市场上。微软于 1986 年首次公开募股，此后不断走高的股价为微软缔造了四位亿万富翁和 12,000 位百万富翁。

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（R Consortium），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。

银牌赞助

纽约数据科学学院 (NYC Data Science Academy)

纽约数据科学学院 (NYC Data Science Academy) 是行业领先的教学教研机构，致力于推进全球数据科学和大数据应用进程、以及向企业界输送数据分析人才。学院下设三大业务板块: 商业咨询培训；数据科学社群以及数据科学家训练营。纽约数据科学学院的创始人 Vivian(张尚轩) 女士是业界知名数据科学家。她带领的团队由数据科学家与数据工程师组成。团队将前沿理论、实践经验与行业人脉等资源进行重新整合，以纽约数据科学学院为平台推出数据科学家训练营 – 全方位的帮助申请人在 12 周内迅速完成一系列数据科学培训、成为数据分析师/科学家。

第九届中国 R 语言会议筹备委员会

主　席　：张心雨
副主席　：冯璟烁
秘书长　：邓金涛
秘书团　：于嘉傲、杨舒仪、王小宁、王健桥、闫晗、彭晨昱
志愿者　：Rachel Tao、曹阳、查蓓蓓、常勤缘、车明佳、陈冠州、陈南、程大曦、代政、邓田娟、丁建军、董峰池、董峻华、范超、高昱竹、郭杰、郭肖晗、顾家齐、何谐、黄衍楠、贾燕飞、兰程、雷赛、李海蓓、李家郡、李明星、李沛佳、李远杰、李智凡、梁学文、刘贝、刘畅、刘慧婷、刘芸、毛璐、聂宇威、秦宇婷、齐立斐、邱雅娟、石美、宋文静、孙文昭、陶逸飞、王道兴、王高斌、王淑羽、王雪晴、王雨婷、王钰苑、王哲、魏凌云、吴豪、习淑婷、辛茹月、辛思、徐琳、薛娜、杨子涵、要卓、易安琪、余跃、袁帅、张碧怡、张昱旻、赵亮、郑晴朗、周生彬、周扬、周艺、周雨菡、周震宇、朱梓睿、邹梦文、朱鹏

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）虽以网站和论坛起家繁荣，但是随着越来越多喜爱统计的朋友们加入，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。目前，COS 的线下活动从一年两次的 R 会议，逐渐发展到沙龙、交流会、竞赛、讲座、培训等等。我们希望更多的新鲜血液可以就近加入 COS 的线下活动中。

COS 线下活动总结：

1. 中国 R 语言会议：目前已开展到第九届，分别在北京、上海、广州、杭州、西安、南昌、武汉等地举办。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 37 期，目前主要在北京，每月举办，详情参见统计之都微信公众号。
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 18 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，《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、张晔翻译，《R 绘图系统》呼思乐、张晔、蔡俊翻译，《R 语言编程实战（暂定）》（待出版）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、郑志勇等翻译等等。

联合会议主会场日程 & 人大地图

5 月 27 日（周五）联合会议主会场（世纪馆）

演讲嘉宾	主题	时间
	参会者入场	7:30-8:20
	致辞	8:30-8:45
李润泽	Computational Issues Related to Big Data Analysis	8:45-9:30
苏萌	大数据与商业价值	9:30-10:15
	自由讨论、休息	10:15-10:45
张志华	大数据分析中的统计学习方法	10:45-11:30
葛伟平	互联网征信数据处理和建模实践	11:30-12:15
	午餐	12:15-14:00
陈为	可视化是分析的一种手段：以城市数据为例	14:00-14:45
冯永昌	量化选股基础：三类因子模型的逻辑和实证	14:45-15:30
	自由讨论、休息	15:30-16:00
陈宇新	基于中国数据的商学研究现状	16:00-16:45
王汉生	数据，价值，回归	16:45-17:30

注：图上标记的餐厅都可以现金消费。

国际统计论坛会场日程

5 月 28 日（周六）国际统计论坛会场（八百人大教室）

主持人	特邀嘉宾	主题	时间
		报到	07:30-08:15
艾春荣	赵彦云	大会致辞	08:15-08:30
	Zongwu Cai	Inferences for Varying-Coefficient Panel Data Models with Cross-Sectional Dependence	8:30-9:50
	Qiwei Yao	Estimation of Extreme Quantiles for Functions of Dependent Random Variables	
		休息	9:50-10:10
Hailiang Yang	Pierre Mohnen	The econometrics of innovation: achievements and challenges	10:10-11:30
	Min-ge Xie	Statistical inferences and fusion learning in the era of data science	
		午餐	11:30-13:45
Jun Yan	Jun Yan	大会致辞	13:45-14:00
	Jason Fine	Dependent Censoring and Competing Risks: Confusion and Controversy	14:00-15:20
	Dianne Cook	Statistics on Street Corners	
		休息	15:20-15:40
Min-ge Xie	赵彦云	Study on internet big data statistics	15:40-17:40
	Hailiang Yang	Valuing Embedded Options in Insurance	
	Per Mykland	Assessment of Uncertainty in High Frequency Data: The Observed Asymptotic Variance	

第九届中国 R 语言会议北京会场日程

会场	28 日上午	28 日下午	29 日上午	29 日下午
国学馆报告厅	R01 互联网征信(考拉征信冠名) 主席：葛伟平	R07 量化分析(量邦科技冠名) 主席：冯永昌	R13 软件工具 主席：邱怡轩	R18 软件工具 主席：覃文峰
第一教学楼 1101	R00 COS play R 主席：冯凌秉		无	
第一教学楼 1204	R02 汽车联网 主席：周扬	R08 智能制造 主席：张玺		
第一教学楼 1205	R03 自然语言 主席：黄浩军	R09 计算平台 主席：高涛		
逸夫第 1 报告厅	2016 百分点数据与价值国际论坛 注:非 R 会议会场，需要单独门票。		R14 经济金融(懒投资冠名) 主席：邓一硕	R19 计算广告 主席：熊熹
逸夫第 1 会议厅	R04 概率统计 主席：吕翔	R10 生物医疗 主席：陈钢	R15 生物信息 主席：谢益辉	R20 社交网络 主席：朱雪宁
逸夫第 2 报告厅	R05 可视分析(Tableau 冠名) 主席：刘琳珂	R11 商务分析 主席：李宜熹	R16 机器学习 主席：常象宇	R21 机器学习 主席：常象宇
逸夫第 2 会议厅	R06 医疗健康 主席：黄帅	R12 生物统计 主席：杨灿	R17 智慧城市 主席：吴海山	R22 时空数据 主席：吴海山

5 月 28 日（周六）COS play R (第一教学楼 1101）

讨论主持人	主题	时间
谢益辉	R 中的那些好玩的东西 & 讨论	08:30-10:00
	自由讨论、休息	10:00-10:30
谢益辉	R 中的那些好玩的东西 & 讨论	10:30-12:00
	午餐	12:00-14:00
李舰、刘思喆	R 在产业界中的应用 & 讨论	14:00-15:30
	自由讨论、休息	15:30-16:00
李舰、刘思喆	R 在产业界中的应用 & 讨论	16:00-17:30

5 月 28 日（周六）上午分会场

分会场	演讲嘉宾	主题	时间
互联网征信 （考拉征信冠名专场） 国学馆报告厅 主席：葛伟平	黄丹阳	互联网征信中的信用评分模型	8:30-9:00
	曹斐	助力小微金融：考拉小微商户信用评分模型的开发与实践	9:00-9:30
	程其江	征信业互联网数据处理架构	9:30-10:00
		休息	10:00-10:30
		来自北京大学、人民大学、中国科学院大学、考拉征信、前隆金融、91 金融等统计、征信、互联网金融领域专家和学者，共同就互联网征信、数据价值和共享、信贷风控模型等问题开展深度讨论，一起探讨互联网征信理论研究、应用实践和未来发展。同时观众们还将有机会与行业大咖现场交流。	10:30-11:30
汽车联网 第一教学楼 1204 主席：周扬	王亮	基于车征数据的 UBI 创新产品	8:30-9:00
	李旭	车联网大数据 --- 数据实践驱动行业发展	9:00-9:30
	游皓麟	电力行业短期日负荷曲线预测	9:30-10:00
		休息	10:00-10:30
	潘蕊	车联网数据与商业价值	10:30-11:00
	胡晓伟	手机传感器在车联网领域中的几点应用	11:00-11:30

自然语言 第一教学楼 1205 主席：黄浩军	李嫣然	自然语言生成的现状与展望	8:30-9:00
	牟力立	Tree-Based Convolution and its Applications	9:00-9:30
	黄伟	基于深度学习的中文语义分析	9:30-10:00
		休息	10:00-10:30
	刘知远	深度学习与自然语言处理	10:30-11:00
	唐建	PTE: Predictive Text Embedding through Large-scale Heterogeneous Text Networks	11:00-11:30
概率统计 逸夫第 1 会议厅 主席：吕翔	刘乐平	关于“P 值”的那些事	8:30-9:00
	吕翔	一点关于 Large-Scale Test 的研究	9:00-9:30
	刘路	带图结构的大偏差理论	9:30-10:00
		休息	10:00-10:30
	陈堰平	动态线性模型的商业应用	10:30-11:00
	兰弘	Comparing Pruning Methods in Perturbation DSGE Models	11:00-11:30
	胡睿	基于 Copula 模型探究新闻情绪和股票收益的相关性	11:30-12:00
可视分析 (Tableau 冠名专场) 逸夫第 2 报告厅 主席： 刘琳珂	刘琳珂	R and Tableau: Smart Meets Fast	8:30-9:00
	Di Cook	Really? Using the nullabor package to learn if what we see is really there?	9:00-9:30
	陈思明	微博轨迹可视化	9:30-10:00
		休息	10:00-10:30
	沈毅	ECHARTS NEXT - 数据 · 视觉编码 · 交互	10:30-11:00
	闻啸	阿里云数据大屏探索	11:00-11:30
	谢佳标	利用 R 语言进行交互数据可视化	11:30-12:00
医疗健康 逸夫第 2 会议厅 主席：黄帅	霍剑	An New Confidence Interval for the Population Proportion in Binary Clustered Data	8:30-9:00
	万小红	Encoding and Decoding of Minds from Neural Activities	9:00-9:30
	李响	认知医疗与健康大数据分析	9:30-10:00
		休息	10:00-10:30
	黄帅	医疗问题中复杂系统的建模，监测，优化，以及控制问题	10:30-11:00
	徐增林	Association Discovery and Diagnosis of Alzheimer's Disease	11:00-11:30

5 月 28 日（周六）下午分会场

分会场	演讲嘉宾	主题	时间
量化分析 （量邦科技冠名专场） 国学馆报告厅 主席：冯永昌	冯永昌	CTA 策略研究方法和寻优中的统计学处理	14:00-14:30
	金戈	建立基于 R 语言的后验系统	14:30-15:00
	任坤	解密高频交易	15:00-15:30
		休息	15:30-16:00
	林伟林	让投资研究更简单-R 与投资研究	16:00-16:30
	李翛然	用 R 语言进行量化风控	16:30-17:00
	張家齊	樂透彩卷的投資策略與回測效果	17:00-17:30
智能制造 第一教学楼 1204 主席：张玺	张建锋	电信网络中的 KQI 和 KPI 的异常检测	14:00-14:30
	田春华	工业大数据分析实践分享	14:30-15:00
	王好	空间统计模型在半导体制造质量研究中的应用	15:00-15:30
		休息	15:30-16:00
	张玺	制造系统中利用传感数据对生产过程的监测与诊断	16:00-16:30
	彭皓	Reliability Optimization for Series Systems under Uncertain Component Reliabilities in the Design Phase	16:30-17:00
	谢帅	个性化制造让定制不再奢侈	17:00-17:30
计算平台 第一教学楼 1205 主席：高涛	孙锐	SparkR 的最新进展和趋势	14:00-14:30
	颜深根	基于 GPU 异构集群的大规模分布式深度学习算法优化	14:30-15:00
	Qu Zheng	Stochastic Dual Coordinate Ascent with Adaptive Probabilities	15:00-15:30
		休息	15:30-16:00
	周俊	大规模机器学习及其应用	16:00-16:30
	骆颇	设计模式选讲:以 caffe 为例	16:30-17:00

生物医疗 逸夫第 1 会议厅 主席： 陈钢	沈侠	What Does That P-value Mean?	14:00-14:30
	李舰	R 语言在医疗人工智能的应用	14:30-15:00
	陈钢	用数据撰写每个家族的传奇	15:00-15:30
		休息	15:30-16:00
	颜林林	癌症液体活检简介	16:00-16:30
	屈武斌	DNA 精准捕获	16:30-17:00
商务分析 逸夫第 2 报告厅 主席：李宜熹	杜晓梦	挖掘数据商业价值，助力企业精准决策	14:00-14:30
	毕然	大数据分析的道与术	14:30-15:00
	陈浪仙	大数据环境下的信用风控技术实践	15:00-15:30
		休息	15:30-16:00
	李宜熹	市场风险管理系统建置与开发	16:00-16:30
	漆晨曦	从大数据到智慧数据——电信企业大数据营销价值发掘和应用	16:30-17:00
	姜天英	互联网个人信用评估研究——基于不平衡样本视角	17:00-17:30
生物统计 逸夫第 2 会议厅 主席：杨灿	张淑芹	Drug-target Interaction Prediction by Integrating Multiview Network Data	14:00-14:30
	李更新	A Weighted Empirical Bayes Risk Prediction Model using Multivariate Traits for Sequencing data	14:30-15:00
	侯琳	Incorporating Network Information to Prioritize Results in Genome Wide Association Studies	15:00-15:30
		休息	15:30-16:00
	魏颖颖	Are All Transcription Factors Interacting with Each Other?	16:00-16:30
	王涛	A Dirichlet-tree Multinomial Regression Model for Associating Dietary Nutrients with Gut Microorganisms	16:30-17:00
	杨灿	IPAC: A Flexible Statistical Approach to Integrating Pleitoropy and Annotation for Characterizing Functional Roles of Genetic Variants that Underlie Human Complex Phenotypes	17:00-17:30

5 月 29 日（周日）上午分会场

分会场	演讲嘉宾	主题	时间
软件工具 国学馆报告厅 主席：邱怡轩	Mark Chen	Deep dive the In-database R	8:30-9:30
	宫雨	Unified Term for R and Julia	9:30-10:00
		休息	10:00-10:30
	Louise Wong	Microsoft R Server Overview	10:30-11:30
	蒋卓	“R” vs "Spark MLlib"在信用评分技术中的应用	11:30-12:00
经济金融 （懒投资冠名专场） 逸夫第 1 报告厅 主席：邓一硕	邓一硕	监管风暴中的互联网金融	8:30-9:00
	汪洋	中国金融运行面临转折性挑战	9:00-9:30
	张云松	R在Online lending中的数据化决策应用--模型工具、数据产品、实时决策	9:30-10:00
		休息	10:00-10:30
	张昊	同盾大数据反欺诈的实践与应用	10:30-11:00
	李慧	大小数据融合的金融营销建模	11:00-11:30
	趙致平	NBA 運動彩卷分析	11:30-12:00
生物信息 逸夫第 1 会议厅 主席：谢益辉	Martin Morgan	An Overview of Genomic Data Analysis in Bioconductor	8:30-9:00
	Lihua Julie Zhu	CRISPRseek and GUIDEseq Packages for Designing Effective and Target-specific gRNAs and Assessing the Precision of Engineered CRISPR-Cas9 Genome Editing System	9:00-9:30
	Charity Law	RNA-seq Analysis in Bioconductor	9:30-10:00
		休息	10:00-10:30
	Jovana Maksimovic	DNA methylation array analysis with missMethyl & other Bioconductor packages	10:30-11:00
	殷腾飞	Wrapping Your R Tools to Analyze National-Scale Cancer Genomics in the Cloud	11:00-11:30
	余光创	ggtree for Visualization and Annotation of Phylogenetic Trees	11:30-12:00

机器学习 逸夫第 2 报告厅 主席：常象宇	徐增林	张量大数据分析及其应用	8:30-9:00
	王尧	Recovering High-order Tensors from Highly Incomplete Observations: Models and Applications	9:00-9:30
	苏海波	推荐系统中的机器学习实践	9:30-10:00
		休息	10:00-10:30
	杨滔	信用风险预测模型	10:30-11:00
	刘汉中	Lasso Adjustments of Treatment Effect Estimates in Randomized Experiments	11:00-11:30
	陈嘉葳	机器学习在异常用户侦测上的应用	11:30-12:00
智慧城市 逸夫第 2 会议厅 主席：吴海山	王静远	智慧城市与城市大数据	8:30-9:00
	张忠元	北京出行便利性研究	9:00-9:30
	周景博	百度时空大数据上的智慧城市应用研究	9:30-10:00
		休息	10:00-10:30
	祝恒书	普适商务环境下的大数据分析	10:30-11:00
	罗应琏	你吐槽过的天气预报原来可以这样玩	11:00-11:30
	李伯楠	多维地理信息交互可视化与特征分区自动提取	11:30-12:00

5 月 29 日（周日）下午分会场

分会场	演讲嘉宾	主题	时间
软件工具 国学馆报告厅 主席：覃文峰	谢益辉	用 R Markdown 愉快地写作是怎样一种体验	14:00-14:30
	刘应耀	基于 R 的数据分析平台	14:30-15:00
	覃文锋	Rust and R Integration	15:00-15:30
		休息	15:30-16:00
	尹志	大数据时代的柳叶刀 --- data.table 使用体验	16:00-16:30
	肖凯	python 中的数据工具箱	16:30-17:00
	邱怡轩	SupR：让 R 语言走向多线程并行计算	17:00-17:30
推荐广告 逸夫第 1 报告厅 主席：熊熹	熊熹	大规模商品推荐系统——从原理到实践	14:00-14:30
	陈开江	认识另一种推荐系统：兴趣 feed	14:30-15:00
	冯扬	微博中的用户建模	15:00-15:30
		休息	15:30-16:00
	胡为松	基于大数据的广告营销	16:00-16:30
	郭贵冰	推荐系统中数据稀疏与冷启动问题的研究	16:30-17:00
	单艺	大数据驱动的智能招聘推荐系统	17:00-17:30
社交网络 逸夫第 1 会议厅 主席：朱雪宁	靳志辉	广聚人群，点通价值 --- 腾讯广点通广告受众定向的探索与实践	14:00-14:30
	刘跃文	基于学生就餐数据的社交网络挖掘及成绩预测	14:30-15:00
	兰伟	Network Autoregressive Factor Model	15:00-15:30
		休息	15:30-16:00
	毛仁歆	“化繁为简” --- 复杂网络在 DT 时代的应用	16:00-16:30
	周静	Tweet Or Retweet? Interaction Utility Derived From User-generated Content In Social Media	16:30-17:00
	陈弢	社交风控的思路和实践	17:00-17:30

机器学习 逸夫第 2 报告厅 主席：常象宇	王乃岩	Revisiting some Basic Components in Deep Neural Networks	14:00-14:30
	王树森	高效的随机矩阵计算	14:30-15:00
	钟琰	Unified Low-rank Matrix Estimation via Penalized Matrix Least Squares Approximation	15:00-15:30
		休息	15:30-16:00
	李文哲	Scalable MCMC method for Bayesian Models	16:00-16:30
	汪张扬	Task-Specific and Interpretable Feature Learning	16:30-17:00
	孟德宇	What's the Insight of Self-paced Learning	17:00-17:30
时空数据 逸夫第 2 会议厅 主席：吴海山	阎军	Detection and Attribution of Changes in Climate Extremes	14:00-14:30
	刘瑜	时空大数据支持下的空间交互研究及应用	14:30-15:00
	韩战钢	生物集群行为的系统研究	15:00-15:30
		休息	15:30-16:00
	王江浩	大规模时空数据分析与可视化：R 应用与实践	16:00-16:30
	李栋	基于迁徙数据的中国次区域识别	16:30-17:00
	周鹏	Application of R in Study on Deep-sea Polymetallic Nodules from the Pacific Ocean and Indian Ocean	17:00-17:30

Computational Issues Related to Big Data Analysis

李润泽（宾州州立大学）
时间：8:45~9:30　邮箱：*rli@stat.psu.edu*

简介： Runze Li is Verne M.Willaman Professor of Statistics, The Pennsylvania State University. He is a fellow of IMS and ASA. He was the co-editor of the Annals of Statistics from 2013-2015 and is an associate editor of Journal of American Statistical Association since 2006.His current researches concentrate on developing effective statistical procedures for high-dimensional data analysis, including variable selection, feature screening and hypothesis testing. He is also interested in applying these statistical procedures for analyzing real-life high-dimensional data such as genetic data analysis and functional MRI data analysis. His other research interests include non- and semi-parametric modeling and statistical applications to scientific research in social behavioral science and engineering.

摘要： This talk will be concerned with computational issues related to big data analysis. I will focus on two major issues: large size and large dimension. For large size data, I will discuss under what situations the commonly-used algorithms are valid and under what settings data analyst should be cautious in the use of the commonly-used algorithms. For large dimension problems, I will present algorithm to find global solutions of several useful regularization problems.

大数据与商业价值

苏萌（百分点集团）
时间：9:30~10:15　邮箱：*meng.su@baifendian.com*

简介： 百分点集团董事长兼首席执行官，美国康奈尔大学营销模型专业博士，国家“千人计划”入选者，美国营销科学院会员，曾担任北京大学光华管理学院博士生导师、副系主任，北京大学新媒体营销研究中心执行主任。他研究的领域包括营销模型、大数据分析、个性化营销、推荐引擎、联合分析、客户终身价值、消费者行为预测等，在美国康奈尔大学师从营销模型领域大师 Vithal R. Rao 教授，曾有多篇论文发表在国内外权威学术刊物上。苏萌博士一直致力于推动大数据在营销学、统计学、计算机科学等多个学科领域的交叉发展，倡导大数据商业应用的不断创新，2012 年出版的著作《个性化：商业的未来》是国内第一本专注于个性化技术与商业应用的书籍。2013 年，苏萌博士放弃了北大全职教授与副系主任等职务，离开了北大并全身投入到百分点的技术创新与管理工作中。2014 年，苏萌博士带领团队成功研发并推出其面向企业级应用的新一代大数据平台产品“百分点数据管家”。该产品浓缩了百分点在大数据领域深耕所积累的核心技术和算法模型，完美支持 PB 级海量数据的采集、存储、整合和挖掘。2015 年，苏萌博士带领百分点技术团队成功研发了全球首款大数据操作系统（BD-OS）。该数据操作系统可以实现海量数据的接入、加工、处理、消费等一整套流程的可视化、智能化、系统化处理，最大化的发现、分析企业内外部核心业务数据价值、辅助挖掘现有业务和应用系统的潜在商机，实现数据应用的完整闭环。

摘要： 大数据技术正在不断向各行各业进行渗透。深度学习、实时数据分析和预测、人工智能等大数据技术逐渐改变着原有的商业模式，推动互联网和传统行业发生着日新月异地变化。传统企业争先恐后地拥抱大数据，而却忽略了一个问题：大数据究竟如何为企业带来商业价值？本次演讲，苏萌博士将结合大数据生态、数据科学技术的发展与百分点的行业商业实践为大家剖析：如何利用大数据实现不同行业中企业的商业价值。孤立的技术并不能解决企业在大数据时代所面临的挑战，只有深耕于行业，深入探索行业痛点，结合大数据技术，构建一整套集底层数据平台、中层管理平台与顶层应用为一体的完整解决方案，才能满足不同行业中企业

的需求，帮助企业实现自身的商业价值。苏萌博士将结合百分点在金融、制造业与泛健康行业中的真实案例为大家娓娓道来，只有深耕于行业，才能发挥大数据的商业价值。

大数据分析中的统计学习方法

张志华（上海交通大学）
时间：10:45~11:30 邮箱：*zhang-ch@cs.sjtu.edu.cn*

简介：张志华，博士，上海交通大学计算机科学与工程系教授，上海交通大学数据科学研究中心兼职教授。在加入上海交通大学之前，是浙江大学计算机学院教授和浙江大学统计科学中心兼职教授。曾经获得 Google 公司全球 Visiting Faculty 计划的资助，并在 Google 北京研究院从事大规模机器学习算法的研发工作 1 年。目前主要从事人工智能、机器学习与应用统计学领域的教学与研究。是美国“数学评论”和 ACMcomputing 的特邀评论员，国际机器学习刊物 Journal of Machine Learning Research 的执行编委，作为程序委员曾服务于许多人工智能与机器学习领域国际会议，比如 IJCAI、AAAI、ICML、NIPS、CVPR 等。其公开课《机器学习导论》和《统计机器学习》受到广泛关注，迄今访问量达 10 余万次。

摘要：在当今的“大数据”时代，科学和工程技术领域源源不断地生成新的数据，且以亿万计的规模迅猛增长。本质上，知识不是直接呈现于数据中，而是需要通过建模、计算或推理等过程把数据变为知识。机器学习是连接统计与计算的桥梁，因此它在大数据分析中居于核心的地位。这个报告将主要讨论大数据分析中的统计学习方法以及潜在研究问题。具体地，报告将通过几个实例阐述贝叶斯机器学习的建模与计算、概率随机技术在大数据计算的应用、基于图结构的架构实现技术等问题。

互联网征信数据处理和建模实践

葛伟平（考拉征信）
时间：11:30~12:15 邮箱：*geweiping@kaolazhengxin.com*

简介：葛伟平，考拉征信服务有限公司联合创始人兼首席技术官，2005 年复旦大学计算机软件博士毕业，2012 年加盟拉卡拉，任集团副总裁，负责收单研发、系统运行、大数据平台体系架构建设和管理。2014 年作为股东代表，参与组建考拉征信服务有限公司，负责数据平台、评分模型、征信系统搭建工作，带领团队先后推出了多个企业和个人信用分产品，同时兼任中国科学院大学——考拉征信模型实验室主任。

摘要：谈谈在互联网征信如何应用 ElasticSearch、NLP 技术；以及在互联网征信模型建设过程中，数据处理、变量组合、特征选择，经典统计学方法与机器学习算法方面的一些体会，和如何以信贷类数据为主，加上消费类、公缴类、通讯类、用户行为等数据，整合不同类型的数据进行数据融合建模，以真实、客观反映信息主体的信用状况。

可视化是分析的一种手段：以城市数据为例

陈为（浙江大学）
时间：14:00~14:45 邮箱：*chenwei@cad.zju.edu.cn*

简介：陈为，1976 年生，浙江大学教授。研究兴趣是数据可视化。发表 70 余篇国际一流学术论文，出版可视化教材 2 部。担任国内外期刊编委和国际学术会议主席多次。承担国基重点、优青等 10 余项。合作研发了多个系统，如: 全球三维大气数据可视化平台、千万量级大图的可视分析系统。详见:http://www.cad.zju.edu.cn/home/chenwei。

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。本次报告将介绍可视化可以帮助用户解决的分析的任务，如展现、理解、推理等。以城市数据（手机基站位置、手机通话、微博、出租车轨迹、POI、房价等）的分析为例展示可视化的关键价值。

量化选股基础: 三类因子模型的逻辑和实证

冯永昌（量邦科技）
时间：14:45~15:30　邮箱：*fengyc@quanttech.cn*

简介：冯永昌，央行互联网金融博士后，北京大学对冲基金实验室联合创始人，中国期货业协会互联网金融委员会专家委员，上海期货交易所博士后导师，对冲基金人才协会资深专家会员，北京大学、清华大学 EDP、FMBA 讲师。北大光华统计学博士，人大统计学学士，美国芝加哥大学访问学者。发起创办了微量网、量邦科技、量客投资等多家公司，目前担任北京量邦信息科技股份有限公司（835352），微量网公司，量客投资公司董事长。

摘要：通过选择合适的天气相关变量与股票收益率进行关联分析来挖掘阿尔法源，并且构建股票组合后进行收益归因分析。所有的分析过程都在量化投资策略研究平台（大宽网）和数据云分析平台（开矿网）上进行，充分地利用了云平台的数据资源和计算能力。

基于中国数据的商学研究现状

陈宇新（上海纽约大学）
时间：16:00~16:45　邮箱：*yc1718@gmail.com*

简介：陈宇新教授现为国家“千人计划”专家，上海纽约大学商学部主任，杰出全球商学讲席教授。并曾任美国西北大学凯洛格商学院市场营销终身讲席教授，纽约大学斯特恩商学院终身教授。陈宇新教授于 1992 毕业于复旦大学物理学系，获理学学士；并分别于 1997 和 1999 在美国圣路易斯华盛顿大学获得市场营销学硕士和博士学位。在 1992 — 1994 期间，陈宇新教授曾为硕士学位研究生就读于浙江大学计算机科学系。陈宇新教授还应邀担任百分点集团首席模型科学家。在银行，电信，汽车，电商，旅游，保险，零售，社交媒体，广告，医疗等领域从事了一系列基于数据建模与分析的咨询和研究工作。陈宇新教授正积极推进大数据营销应用领域的相关研究工作，应邀参与了大数据相关领域国家自然科学基金重点和重大项目的立项，评审和顾问工作。陈宇新教授曾荣获 Frank M. Bass 最佳营销学博士论文奖、John D.C. Little 最佳营销学论文奖，INFORMS 营销协会长期影响提名奖，Paul E Green 最佳营销学论文奖等国际学术荣誉。陈宇新教授的研究领域主要涉及数据驱动营销, 互联网营销、竞争战略，零售、定价、广告、结构实证模型、贝叶斯计量经济学及行为经济学等。陈宇新教授现为国际营销科学顶级刊物《营销科学》4 名高级主编之一，并曾担任国际营销学及管理学顶级刊物《营销学研究期刊》,《营销科学》、《管理科学》和《定量营销和经济学》的副主编及《生产与运营管理期刊》,《客户需求与解决方案》的高级编委。陈宇新博士现还担任 INFORMS 营销协会顾问委员会理事。同时，陈宇新教授还受邀担任了任期五年的香港研究资助局商学部评审委员。

摘要：近年来，已有越来越多的海内外商学研究者投入到基于中国数据源的研究。本报告从最近 5 年来发表于 UTD-24 和 FT-45 两个国际一流学术刊物排名中 47 个期刊上的全部基于中国数据源的学术论文出发，结合数据挖掘和数据可视化方法，探讨了商学领域基于中国大数据的研究现状。分析聚焦于研究领域、研究人员和研究关键词三个维度，从研究趋势和学术合作两个方向为未来的中国商学数据研究提供了有益参考。

数据，价值，回归

王汉生（北京大学）
时间：16:45~17:30 邮箱：*hansheng@pku.edu.cn*

简介：王汉生，北京大学光华管理学院商务统计与经济计量系，嘉茂荣聘讲席教授，博导；北京大学商务智能研究中心主任；光华管理学院 MBA，EMBA，ExEd，本硕博教学指导委员会成员；美国统计学会（American Statistical Association）会士（Fellow, 2014）。1998 年北京大学数学学院概率统计系本科毕业，2001 年美国威斯康星大学麦迪逊分校统计系博士毕业。2003 年加入光华至今。国内外各种专业杂志上发表文章逾 80 篇，并合著英文专著 1 本，中文教材 2 本。国际统计协会（International Statistical Institute）、英国皇家统计协会（Royal Statistical Society）、美国统计协会（American Statistical Association）、美国数理统计协会（Institute of Mathematical Statistics）、泛华国际统计协会（International Chinese Statistical Association）的会员。先后历任以下国际学术刊物副主编（Associate Editor）:The Annals of Statistics (2008—2009), Computational Statistics & Data Analysis (2008—现在), Statistics and its Interface (2010—现在), Journal of the American Statistical Association (2011—现在), Statistica Sinica (2011—现在), Journal of Business and Economics Statistics (2012–现在), 中国科学数学（2013—现在）。在理论研究方面，关注高维数据分析。具体内容有: 变量选择、收缩估计、数据降维等。在应用方面，关注统计学方法在电子商务领域的应用，尤其关注中文文本分析、社会关系网络以及位置轨迹数据。

摘要：我们都说这是一个大数据时代，但是: 数据到底是什么？能否给数据一个朴素的定义？这个定义背后的时代特征又是什么？如果说，我们对数据的痴迷执着是因为: 数据可以产生价值。那么请问: 价值又是什么？价值会体现在商业实践的那几个方面？在怎样的场景环境下，价值才能够被客户感知？在这个数据爆炸，价值却不清晰的时代，如何实现从数据到价值的回归？背后有没有一般化的方法论？想听听熊大的看法吗？咱们人大世纪馆见！

COS play R

谢益辉、刘思喆、李舰

时间：8:30~17:30

摘要：COS 论坛上曾经有一段时间比较流行用 R “不务正业”，其大概意思也就是用 R 做一些并非与统计直接相关但好玩的事情，这个 COSPlay R 分会场的主旨便是大家一起分享一些自己觉得好玩或实用的 R 技法。上午会场由谢益辉主持，漫谈一些统计之都及 RStudio 出品的 R 包的基本功能、设计细节以及历史八卦等；下午会场分别由刘思喆和李舰主持，谈谈 R 的商业应用、行业案例、系统架构等。我们欢迎参加这个会场的听众也积极上台分享自己使用 R 的经验、乐趣、或困惑，每一位听众大约有五到十分钟时间，请有意参与的听众在会前事先做好准备。

互联网征信中的信用评分模型

黄丹阳（中国人民大学统计学院）
时间：8:30~9:00　邮箱：dyhuang89@126.com

简介：黄丹阳，2011 年于中国人民大学统计学院取得经济学学士学位，主修统计学专业，副修金融学专业。2015 年于北京大学光华管理学院取得经济学博士学位，统计学专业。同年毕业回到中国人民大学统计学院任教。研究方向包括搜索引擎营销背景下的超高维变量选择问题，社交网络建模。

摘要：面向小微商户以及个人消费的小微信贷是当前互联网金融的重要发展方向，并且正在经历爆发式增长。在这个增长过程中，如何在没有实物抵押的情况下，通过互联网大数据分析，实现快速准确征信是一个非常重要的问题。为此，不同的数据都可以做出一定的贡献。本研究一方面通过追踪用户历史行为数据，建立互联网征信的信用评分模型，另一方面通过跨平台的用户简历数据融合，进一步改善了预测精度。研究表明，用户历史行为对于用户信用评估具有重大作用，且跨平台数据融合将对于预测用户信用评估有进一步的帮助。

助力小微金融：考拉小微商户信用评分模型的开发与实践

曹斐（考拉征信服务有限公司）
时间：9:00~9:30　邮箱：caofei@kaolazhengxin.com

简介：曹斐，考拉征信高级数据分析师，北京大学软件与微电子学院硕士，曾在金融、保险、电信、生物等相关领域从事多年数据分析工作。擅长数据集成与治理、用户行为研究、金融数据建模，最容易被真正落到实处的数据产品所感动。

摘要：考拉小微商户信用分是国内首款针对小微商户领域推出的征信产品，通过采集这些小微商户的基本属性、每日经营交易流水数据、工商信息，以及外部的互联网公开信息，利用机器学习算法，从商户属性、信用记录、履约能力、成长能力、经营稳定、交易行为等维度评估小微商户信用，帮助他们凭借信用申请到贷款，缓解小微企业融资难问题。本次分享将贯穿一个完整的信用评分模型构建流程，从分析、理解小微商户的信用特点出发，介绍使用 R 语言建立信用评分模型的主要过程以及模型在市场中的应用实践情况。

征信业互联网数据处理架构

程其江（考拉征信服务有限公司）
时间：9:30~10:00　邮箱：chengqijiang@kaolazhengxin.com

简介：程其江，中科院研究生院硕士，考拉征信数据处理总监，熟悉 Hadoop 生态圈、Spark 生态圈、多年大数据平台建设经验，曾担任联想研究院高级研究员、Mop 网技术经理等。

摘要：互联网数据对征信业越来越重要，已成为个人和企业征信数据的重要来源。本报告从考拉征信互联网数据处理架构实践出发，详述数据采集、数据集成、数据存储、全文索引、自然语言处理以及数据服务的整个数据平台体系。着重介绍在各开源项目上改进工作，分享自然语言处理在征信数据处理中的实践经验，特别介绍文本分类、知识图谱在数据处理中应用。

基于车征数据的 UBI 创新产品

王亮（上海车征网络科技有限公司）
时间：8:30~9:00 邮箱：liang.wang@ichezheng.com

简介：2015 年创办车征，车征联合创始人兼 CTO。2013 年 -2015 年，宝尊电子商务金融事业部负责人，承揽了天猫保险大部分的保险业务。数十年保险行业经验，服务于多家保险公司及中介公司。

摘要：介绍各类车征的 UBI 创新产品。

车联网大数据——数据实践驱动行业发展

李旭（北京车网互联科技有限公司）
时间：9:00~9:30 邮箱：13810497045@139.com

简介：李旭，北京车网互联科技有限公司监事、行业总经理，北京大学光华管理学院 MBA，10 年车联网工作经验，熟悉车联网、汽车售后服务、物联网大数据和 UBI 保险等领域，作为主要发明人，已获车联网领域多项发明专利，曾牵头组织中国移动、广汽本田、东风本田、人保财险等多项车联网项目，参与完成《基于车联网的多维大数据综合运营服务系统》项目设计规划，是中英合作项目《车联网大数据联合运营》主要参与人。

摘要：车联网作为物联网的分支，较之互联网大数据，物联网领域的大数据有其独到的特点，相对规模较小、单位数据成本更高、紧密联系实际场景、商业价值转化路径更短。车联网作为物联网中最成熟的应用，充分体现出上述特点，也因此，可以广泛应用于主机厂、保险公司、汽车经销商、汽车租赁企业、专车公司等各类型的场景中。对于车联网大数据应用的探索，渐渐从理论研究过渡到实际应用，伴随数据规模的扩大，不断衍生出新的价值，带动行业更加快速的发展。

电力行业短期日负荷曲线预测

游皓麟（深圳市数据能源科技有限公司）
时间：9:30~10:00 邮箱：cador@sina.com

简介：游皓麟，高级数据分析师，专注于数据分析、挖掘、大数据领域，在互联网/电信/电力方面具有丰富的数据分析与挖掘建模经验，目前研究 NLP、知识图谱等内容。曾服务于华为技术软件有限公司、深圳市康拓普信息技术有限公司、深圳市数聚能源科技有限公司等企业，期间曾在小象学院兼职 R 语言数据挖掘讲师，参与过《R 语言与 Hadoop 大数据分析实战》书籍的翻译工作，著有《R 语言预测实战》，今年可出版。

摘要：电力负荷预测在电力系统计划与运行管理中起着重要作用，随着电网系统的建设更加完善，数据质量更有保障，以及外围数据的接入，使得负荷预测工作面临着全新的局面。本次演讲从变压器的日负荷曲线预测入手，介绍电力行业背景以及开展负荷预测的必要性，另外从算法角度，提出了基于特征学习的预测算法，希望与各界朋友交流讨论。

车联网数据与商业价值

潘蕊（中央财经大学）
时间：10:30~11:00 邮箱：*panrui_cufe@126.com*

简介：研究兴趣：高维数据变量选择；网络结构数据建模；地理位置数据（车联网数据）统计分析

摘要：随着互联网的发展和完善，海量数据正不断形成，车联网数据便是其中之一。车载设备提供的车联网大数据为分析用户驾驶行为提供了数据基础，该设备记录了用户驾驶过程中的车辆硬件数据、地理信息数据以及司机行为数据。车辆网大数据的出现带来了新的商机，基于驾驶人行为的保险（UBI）便是其中之一。本报告将通过车联网数据，探究影响车辆出险的重要因素，并据此构建相应的指标体系，对车辆未来出险情况进行预测，从而为车险公司提供合理的评判用户出险率的依据。

手机传感器在车联网领域中的几点应用

胡晓伟（浙江从泰网络科技有限公司）
时间：11:00~11:30 邮箱：*pengpeng@icongtai.com*

简介：胡晓伟, 男,1991 年 3 月生, 安徽六安人, 南京大学情报学硕士, 毕业后加入阿里巴巴集团天猫事业部商业智能部, 入职 10 个月破格晋升为资深数据分析师，随后离职创业。现任斑马行车算法专家，研究方向为手机内置传感器在车联网中的应用和 UBI 车险精算。

摘要：智能手机的高度普及让手机收集用户驾驶行为数据从而个性化精算车险成为可能，但另一方面，手机的电池容量低影响了用户使用此类 APP 的意愿，手机传感器精密性不高又影响了数据源本身的质量。本次演讲分享了本人通过 R 语言建模和训练，将手机传感器数据应用在斑马行车上以解决上述问题的几个成功案例，包括: (1) 斑马行车的业务和数据架构介绍 (2) 运用 R 语言构建低通滤波器并实现一种 ios 和 android 手机通用的加框自拟合计步器算法 (3) 运用 C5.0 算法设计手机传感器智能判断用户行为状态的分类器 (4) 介绍斑马行车对手机传感器的其他几点应用，包括基于加速计实现一种高精度的步态方向罗盘、运用 R 语言实现 GPS 与传感器融合导航等。

自然语言生成的现状与展望

李嫣然（香港理工大学）
时间：8:30~9:00 邮箱：*yanranli.summer@gmail.com*

简介：李嫣然，毕业于北京大学智能科学专业。现任香港理工大学研究助理，研究方向为自然语言处理中的语义表达和语言生成。

摘要：自然语言处理的研究包含着让机器理解自然语言和让机器产生自然语言。近年来，研究者们已经成功让机器自动生成诗歌或对联等富有韵律的文本。然而，自由文本的生成还不尽如人意。一方面，蓬勃发展的神经网络技术使得自然语言生成的研究取得了一定进展。但单纯依赖于神经网络技术的生成方法还存在诸多问题。另一方面，基于模板和规则的方法仍然发挥着作用。自然语言生成中的难点都有哪些？两类方法都适合解决哪些难点？两类方法能否有效结合？这些都是自然语言生成中亟待解决的问题。

Tree-Based Convolution and its Applications

牟力立（*Peking University*）
时间：9:00~9:30 邮箱：*doublepower.mou@gmail.com*

简介：Lili Mou received his Bachelor's degree in computer science from Peking University in 2012. He is now a Ph.D. student, supervised by Profs. Zhi Jin, Ge Li, and Lu Zhang. His recent research interests include deep learning applied to natural language processing as well as programming language processing.

摘要：Neural networks have wide applications in NLP, e.g., POS tagging, parsing, machine translation, etc. Several prevailing neural models include convolutional neural networks, recurrent neural networks, and recursive networks. In my talk, I will briefly review these models, and then introduce a novel tree-based convolutional neural network (TBCNN), which can capture structural information effectively. I have applied TBCNN to consitutency trees and dependency trees of natural language, as well as abstract syntax trees of programming language. Finally, I will discuss the advantages and disadvantages of different neural models (including the attention mechanism) in information processing.

基于深度学习的中文语义分析

黄伟（百分点集团）
时间：9:30~10:00 邮箱：*wei.huang@baifendian.com*

简介：毕业于上海交通大学计算机专业，曾在汤森路透等企业从事机器学习和自然语言处理方面的工作；现就职于百分点集团，负责机器学习和非结构化数据挖掘工作，特别是基于中文语义分析的商品自动分类、商品画像、情感分析和口碑分析等应用。同时一直在研究如何利用非结构化数据进行量化投资。

摘要：(1) 语义分析中的必要性，以及传统方法的局限性。大数据的核心目标是让应用变得“智能”，这就迫使我们“教会”计算机去理解自然语言的语义；传统的语义分析经历了形式化规则和机器学习两个阶段，虽然它们取得了很高的成就，但无论是在适用场景还是效果上都遇到了瓶颈，很难进一步提升。

(2) 利用深度学习进行语义分析。深度学习为语义分析提供了新思路和工具，这里介绍深度学习的基本原理和常用技术，以及在语义分析上的应用思路。

(3) 基于深度学习的情感分析实例。介绍百分点集团内部在情感分析上的应用和实现，以及取得的效果。

(4) 再评深度学习。介绍深度学习与传统机器学习的异同，以及在实践中如何合理选择。

深度学习与自然语言处理

刘知远（清华大学）
时间：10:30~11:00 邮箱：*liuzy@tsinghua.edu.cn*

简介：刘知远，清华大学计算机系助理研究员，主要研究方向为语义分析和社会计算。2011 年获得清华大学博士学位。已在自然语言处理等领域的著名国际期刊和会议发表相关论文十余篇。曾获清华大学优秀博士学位论文、中国人工智能学会优秀博士学位论文、清华大学优秀博士后等称号。

摘要：表示学习是机器学习的重要环节，在自然语言处理任务中扮演着重要角色。深度学习则是表示学习的重要技术之一，是最近的研究热点之一。报告将从词汇、短语、文档和知识图谱等几个层面，介绍以深度学习为代表的表示学习技术在自然语言处理领域的最新研究进展与前景。

PTE: Predictive Text Embedding through Large-scale Heterogeneous Text Networks

唐建（微软亚洲研究院）
时间：11:00~11:30 邮箱：*Jiatang@microsoft.com*

简介：唐建博士毕业于北京大学，目前为微软亚洲研究院机器学习组的副研究员。他的主要研究方向包括深度学习，统计主题模型以及这些方法在自然语言理解、网络分析、用户行为分析等领域的应用。他的主要论文都发表在机器学习和数据挖掘领域的国际顶级会议上包括 ICML、KDD、WWW、AAAI 以及 CIKM 等。他是机器学习领域国际会议 ICML2014 的最佳论文获得者以及多个国际会议的程序委员会成员包括 WWW、IJCAI、ACL、EMNLP。

摘要：Unsupervised text embedding methods, such as Skip-gram and Paragraph Vector, have been attracting increasing attention due to their simplicity, scalability, and effectiveness. However, comparing to sophisticated deep learning architectures such as convolutional neural networks, these methods usually yield inferior results when applied to particular machine learning tasks. One possible reason is that these text embedding methods learn the representation of text in a fully unsupervised way, without leveraging the labeled information available for the task. Although the low dimensional representations learned are applicable to many different tasks, they are not particularly tuned for any task. In this paper, we fill this gap by proposing a semi-supervised representation learning method for text data, which we call the textitpredictive text embedding (PTE). Predictive text embedding utilizes both labeled and unlabeled data to learn the embedding of text. The labeled information and different levels of word co-occurrence information are first represented as a large-scale heterogeneous text network, which is then embedded into a low dimensional space through a principled and efficient algorithm. This low dimensional embedding not only preserves the semantic closeness of words and documents, but also has a strong predictive power for the particular task. Compared to recent supervised

approaches based on convolutional neural networks, predictive text embedding is comparable or more effective, much more efficient, and has fewer parameters to tune.

关于“P 值”的那些事

刘乐平（天津财经大学）
时间：8:30~9:00 邮箱：*liulp66@163.com*

简介：1983.9 — 1987.7 毕业于江西大学数学系并获理学学士学位；
1995.9 — 1998.7 毕业于华东师范大学数理统计系并获理学硕士学位；
2000.9 — 2003.7 毕业于中国人民大学统计学系并获经济学博士学位；
2004-至今，天津财经大学统计学与金融学，教授、博导。

摘要：针对 P 值的误解与滥用，美国统计学会（ASA）2016 年 3 月正式发表了“关于统计显著性与 P-值”的官方声明。通过回顾 20 世纪 30 年代假设检验理论的起源与发展，比较 Fisher 显著性检验与 Neyman-Pearson 有效检验理论的差异，分析频率统计与贝叶斯分析的区别，强调统计思维在大数据时代的作用。结合国内统计学教学的现状，基于美国统计学会的官方声明，提出改进统计学教学的建议。

一点关于 Large-Scale Test 的研究

吕翔（中国人民大学）
时间：9:00~9:30 邮箱：*alanlvruc@gmail.com*

简介：中国人民大学大四学生。研究方向高维统计和并行计算。

摘要：Massive data is a common phenomenon in modern statistical problems. Due to its unique characteristics like heterogeneity, pre-established statistical learning methods become invalid. In this talk, we briefly review a new method to conduct inference for heterogenous massive data.

带图结构的大偏差理论

刘路（中南大学）
时间：9:30~10:00 邮箱：*g.jiayi.liu@gmail.com*

简介：现在主要研究方向为数理统计。在 Tran.Amer.Math 等杂志上发表论文若干。

摘要：我们介绍稀疏图上的一些过程的弱收敛定理和局部染色策略逼近最优策略的结果。我们说明这些结果与带图结构的大偏差理论的关系。

动态线性模型的商业应用

陈堰平（雪晴数据网）
时间：10:30~11:00 邮箱：*yanping.chen@xueqing.tv*

简介：雪晴数据网（www.xueqing.tv）创始人，主要从事统计咨询、数据分析、开发基于 R 语言的定制化统计软件，曾给惠普中国研发中心、花旗银行、东方航空、中国电信做过培训和咨询。现在同时也是统计之都理事会成员、中国 R 语言会议理事会成员，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。

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

(2) 介绍 DLM 的其他应用场景：百度旅游预测等

(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。

Comparing Pruning Methods in Perturbation DSGE Models

兰弘（对外经济贸易大学）

时间：11:00~11:30　邮箱：*lanhongken@outlook.com*

简介：对外经贸大学助理教授。博士毕业于德国洪堡大学，研究兴趣为随机动态模型的数据方法、宏观金融学。

摘要：We study the rationale and performance of DSGE perturbations that are pruned to guarantee stable simulations. We show that the moving average representation of the policy function is naturally pruned and express the nonlinear moving average recursively. This recursive algorithm differs from pruning algorithms and the rationale provide by series expansions in that it evaluates risk at the stochastic instead of the deterministic steady state. We compare seven different pruning algorithms at second and third order, documenting the differences between these algorithms and standard (non pruned) state space perturbations at <U+FB01>rst, second, and third order in a uni<U+FB01>ed notation. The nonlinear moving average is the most accurate and the series expansion the second most accurate; yet the two algorithms perform comparably, suggesting that this choice is unlikely to be a potential source of error. Alternative ad hoc algorithms from the literature suffer a loss of accuracy to varying degrees as they include terms inconsistent with or neglect terms consistent with the order of approximation.

基于 Copula 模型探究新闻情绪和股票收益的相关性

胡睿（中央财经大学）

时间：11:30~12:00　邮箱：*ruihu.cufe@outlook.com*

简介：中央财经大学统计学专业 12 级本科生，一个正在努力成为数据科学家的数据爱好者

摘要：应用多元 t-Copula 函数构建多维联合分布，对不同情绪的新闻文本数量和股票收益的相关性结构和尾部相关性进行了研究，发现正面新闻情绪和股票收益相关性较强而负面新闻情绪有一定的滞后性。结果表明, 在探究多个变量的非线性相关性问题上，Copula 函数更为灵活和准确，在金融领域有广泛的应用前景。本演讲将以 R 语言环境为背景，从新闻文本的爬取和分词入手，介绍如何得到关于新闻文本和股票收益的边际分布函数并通过多元 t-Copula 函数构建它们的联合分布，从而达到探究新闻情绪和股票收益相关性的目的。

R and Tableau：Smart Meets Fast

刘琳珂（*Tableau*）
时间：8:30~9:00 邮箱：*lliu@tableau.com*

简介：刘琳珂现任 Tableau 大中国区首席产品顾问，负责管理大中华区的咨询团队，帮助用户快速分析、探索、可视化数据的价值。刘先生在商务智能和数据仓库领域从业 15 年，曾任职于 Sybase，BusinessObjects，SAP，Qlikview，Oracle。从事商务智能和数据仓库解决方案和技术架构服务。其专注的领域包括: 商务智能、数据库、数据仓库、数据云架构、可视化分析等。

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.

Really? Using the nullabor Package to Learn if What We See is Really Tthere?

Di Cook（*Monash University*）
时间：9:00~9:30 邮箱：*dicook@monash.edu*

简介：莫纳什大学经济与商学院教授。

摘要：Plots of data often provoke the response ”is what we see really there”. In this talk we will discuss the use of the nullabor package to assess the significance of structure discovered by exploring data visually. Classically, quantifying significance with p-values required a rigorous protocol involving several steps: hypothesis formation, data collection, test statistic calculation, and comparison with a reference distribution requiring strict assumptions. The nullabor package implements the lineup protocol, which compares a plot of data with plots of null data. The lineup protocol is named after the ”lineup”, popular from criminal legal procedures. The nullabor package has several methods for generating null data, randomises and encodes the position of the data plot, and all the power of ggplot2 for making data plots. This package enables the data analyst to quantify their findings as different, or not, from spurious patterns. Joint work with Hadley Wickham and Heike Hofmann.

微博轨迹可视化

陈思明（北京大学）
时间：9:30~10:00 邮箱：*simingchen3@gmail.com*

简介：北京大学博士研究生，来自北京大学可视化与可视分析实验室。热爱并致力于发展可视分析技术，相信用技术与设计可以让数据生动鲜活。研究方向包括时空数据、社交媒体与网络安全的可视分析技术研究。研究成果发表于 IEEE VAST (TVCG)、VizSec 等国际知名会议与期刊上，参与 IEEE VAST Challenge 可视分析竞赛并获得多项一等奖。工作之余，兴趣在于旅行、对联与诗词。

摘要：我们以人们发布的带有地理信息的社交媒体数据为切入点，来观察、探索与分析个人的行为轨迹，乃至群体行为特征。以新浪微博为例，将用户带有地理信息的微博按照时间顺序连接起来，就可以构造出他们在实际物理空间中的稀疏轨迹。通过合理的可视化设计方案，可以构造出每个社交媒体用户带有明显个人特征的轨迹，例如旅行爱好者、商务白领、学者等，每个人的轨迹不尽相同。这些轨迹的每个采样往往都含有时间、文本、图片等丰富的信息，可以讲述一个个精彩的故事。我们的可视化系统允许用户探索自己以及好友的“微博足迹”，每个人都可以参与进来，共同分析轨迹特征。进一步地，我们实现了一个社会人群移动的群体行为的可视分析系统，让用户交互地发现群体行为中的空间、时间以及多维属性上的规律。

ECHARTS NEXT ——数据・视觉编码・交互

沈毅（百度）
时间：10:30~11:00　邮箱：*shenyi.914@gmail.com*

简介：百度资深前端工程师，目前主要负责维护 ECharts

摘要：介绍 ECharts 3 与可视化，包括可视化中的不同类别数据的常见可视化方式，数据的分类，颜色，glyph，尺寸等可视编码手段，还有可视化中常见的动画，交互等。

阿里云数据大屏探索

闻啸（阿里云）
时间：11:00~11:30　邮箱：*ninglang.wx@alibaba-inc.com*

简介：目前负责阿里云数据引擎数据可视化团队。从最早跟随团队将数据可视化概念引入公司，让公司对外数据展示项目升级换代，到抽象需求建立通用的 datav.js 数据可视化前端 js 组件库，并一步步将数据可视化真正落地到产品，帮助数据可视化在阿里巴巴内部从一个光鲜炫酷的新兴概念，扎根成为了帮助数据分析，简化数据理解的本质需求。

摘要：数据可视化的核心价值在于把多样的信息融合在一个界面中友好的体现，让人能更容易的把握整体业务全景，降低数据理解分析门槛，创建数据共享平台，实现可视化分析。

利用 R 语言进行交互数据可视化

谢佳标（乐逗游戏）
时间：11:30~12:00　邮箱：*jiabiao1602@163.com*

简介：多届中国 R 语言大会演讲嘉宾，目前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。

摘要：数据可视化可以是静态的或交互的。几个世纪以来，人们一直在使用静态数据可视化，如图表和地图。交互式的数据可视化则相对更为先进: 人们能够使用电脑和移动设备深入到这些图表和图形的具体细节，然后用交互的方式改变他们看到的数据及数据的处理方式。本演讲会带领大家一起了解如何用 R 语言绘制交互的柱状图、气泡图、时序图、社会网络图、股价图、文氏图、treemap 图、平行图等，并用游戏数据演示如何将这些交互图应用到实际生产环境中。令参会者们迅速掌握利用 R 绘制不同交互图。

An New Confidence Interval for the Population Proportion in Binary Clustered Data

霍剑（中国人民大学）
时间：8:30~9:00 邮箱：*huojian555@163.com*

简介：霍剑，中国人民大学统计学院博士生，研究方向：生物医学统计，曾参与医药行业、保险行业等国家级项目四项，获 2015 年中国卫生统计学术研讨会优秀论文奖等。

摘要：Binary clustered data are common in biomedical studies. In this paper we construct a new confidence interval of the response proportion in clustered data. The idea of our construction came from Wilson interval for a binomial proportion. The coverage probability of the existing confidence intervals are poor when the true portion is small or large. Our proposed confidence interval obviously improve the performance in that case. With regard to the criterions of coverage probability and expected length, the new confidence interval is better than the other existing methods in simulation studies. A real data example is also presented to show the application of our method.

Encoding and Decoding of Minds from Neural Activities

万小红（北京师范大学）
时间：9:00~9:30 邮箱：*xhwan@bnu.edu.cn*

简介：2013 年青年千人计划引进人才，认知神经科学与学习国家重点实验室 PI，麦戈文脑科学研究中心 PI。

摘要：Human brain constitutes of billions of neurons working together to represent moment-to-moment minds. One tale of neuroscience is to understand how the information is encoded in the brain, and the other tale is to decode the dynamic minds from the tons of recoded neural activities. In this talk, I will briefly introduce the current frontiers about how far we understand the encoding mechanisms and how far we can decode the minds from the recorded neural activities in neuroscience.

认知医疗与健康大数据分析

李响（*IBM* 中国研究院）
时间：9:30~10:00 邮箱：*lixiang@cn.ibm.com*

简介：李响，2011 年毕业于浙江大学计算机学院并获得博士学位，目前是 IBM 中国研究院认知医疗部的研究员。研究兴趣包括医疗信息学、机器学习和数据挖掘，在疾病风险预测、治疗路径挖掘、病历信息抽取等方面进行了大量研究工作。

摘要：中国的医疗行业面临着日益严峻的挑战。一方面患者数量（特别是慢性病患者）非常庞大，另一方面医疗资源稀缺且分布严重不平衡。通过认知计算和大数据分析技术，能够优化医生的诊疗过程并提高患者的自我管理水平，是改善目前医疗状况的有效手段。一方面，通过自然语言处理与知识推理等技术，可以从大量

医学文献及临床指南中获取治疗方案建议和医学证据，为医生提供专业可靠的临床决策支持。另一方面，利用机器学习与数据挖掘技术，能够在临床研究数据、电子病历、医保数据、医学影像等健康数据的基础上，实现更精准的疾病风险预测、患者精细分群和治疗路径挖掘，不仅能为医生提供精准化的治疗路径建议，并且能为患者提供个性化的自我管理工具。目前，通过与专业医院及区域卫生机构合作，认知计算与大数据分析技术在医学研究和临床实践上已取得突破，未来将产生更深远的影响。

医疗问题中复杂系统的建模，监测，优化，以及控制问题

黄帅（*University of Washington at Seattle*）
时间：10:30~11:00　邮箱：*shuaih@uw.edu*

简介：黄帅，现任职美国华盛顿大学 -西雅图分校的工业工程系助理教授。黄帅于 2007 年在中国科技大学少年班获得统计学位，于 2012 年在美国亚利桑那州立大学工业工程系获得博士学位。其主要研究方向是结合统计、机器学习、运筹方法，研究一些医疗管理以及工程领域里面的复杂决策问题。具体应用比如老年痴呆、青少年糖尿病、手术感染等问题的监测和预防等等。从医疗问题出发，这些研究成果可以被广泛的应用在其他各类复杂系统之上，比如制造业或者供应链管理。他的研究获得了美国自然科学基金（National Science Foundation），Juvenile Diabetes Research Foundation，以及其他一些医学基金会以及医学机构的资助。有关他的具体研究工作，可以在他的主页上了解更多:https://sites.google.com/site/shuaihuang28/。

摘要：信息科技的发展提供了很多前所未有的机会去解决一些复杂的医疗问题。这些机会包括: 新的数据收集方法、新的监测手段、新的交流方式等。“解决”在这里比治愈更加广义，其意味着更好更有效的管理。比如在老年痴呆的疾病研究中，有一个被广泛接受的观点是，只要疾病能在发病的过程中被及时发现，有效的预防或者其他医疗方案就能被及时使用、去减缓发病过程或者病症，维持病人的大脑健康以及生活能力等等。又如在青少年糖尿病的管理之中，及时发现诱发糖尿病的因素能帮助携带发病基因的人群“对症下药”去改变自己的环境，改变自己的生活习惯和饮食结构等等。类似的例子还能在其他很多医疗问题中找到，比如抑郁症，或者美国老兵医院最近开始实行的个性化医疗计划和管理。

因此，本次演讲的目的是介绍这些医疗问题中的统计以及管理问题以及我在这些问题上的研究工作。事实上，在寻求这些问题的有效解决方案的过程中，我逐渐意识到这些问题之所以难解，是因为它们牵涉到一个个动态的复杂系统。这超出了统计中常有的一些概念比如总体或者样本的范畴。对于这些医疗问题中的复杂系统，怎么利用统计的方法去建模，怎么结合统计、运筹学以及其他管理科学去监测、优化、控制并且怎么把这些方法通过什么样的决策框架下真正在医疗决策问题中产生实效，是我的研究的主要内容。

Association Discovery and Diagnosis of Alzheimer's Disease

徐增林（电子科技大学）
时间：11:00~11:30　邮箱：*zlxu@uestc.edu.cn*

简介：徐增林，电子科技大学教授、博士生导师，中组部“青年千人计划”入选者，现任电子科技大学大数据研究中心数据挖掘与推理研究所轮值所长，并创建统计机器智能与学习实验室 (Statistical Machine Intelligence and LEarning, SMILE, http://bigdatalab.weebly.com/)。徐增林教授主要研究兴趣为机器学习及其在社会网络分析、互联网、计算生物学、信息安全等方面的应用。他在包括 IEEE TPAMI, IEEE TNN，NIPS, ICML, IJCAI, AAAI 等顶级会议和刊物发表论文近 30 篇，引用近千次，发表专著 2 部，书籍章节 2 篇，并于 2015 年的 AAAI 大会获得最佳学生论文奖提名。徐增林于 2012 年在多伦多召开的国际人工智能大会 (AAAI)

上做教学报告。徐增林教授是 JMLR，IEEE TPAMI 等机器学习与人工智能领域主要期刊的审稿人和香港教育资助局的基金评审人；多次担任人工智能领域的主要国际会议如 AAAI/IJCAI 等会议的程序委员会成员；多次担任机器学习和大数据研究方面的研讨会的组织委员会主席。

摘要： In biological and biomedical research, the analysis and diagnosis of many complex diseases, e.g., Alzheimer’s disease, can be based on a number of data sources or views, such as genetic variations and the phenotypic traits. This brings a new machine learning setting where the objectives are of two folds – to make diagnosis and to study the association between the genetic variations and the phenotypic traits. In this talk, we discuss a new sparse Bayesian approach for joint association study and disease diagnosis. In this approach, common latent features are extracted from different data sources based on sparse projection matrices and used to predict multiple disease severity levels; in return, the disease status can guide the discovery of relationships between data sources. I will also discuss how to take advantage of the linkage disequilibrium (LD) measuring the non-random association of alleles to guide the selection of genes. Finally, I show analysis on imaging genetics datasets for the study of Alzheimer’s Disease.

CTA 策略研究方法和寻优中的统计学处理

冯永昌（量邦科技）
时间：14:00~14:30 邮箱：*fengyc@quanttech.cn*

简介：冯永昌，央行互联网金融博士后，北京大学对冲基金实验室联合创始人，中国期货业协会互联网金融委员会专家委员，上海期货交易所博士后导师，对冲基金人才协会资深专家会员，北京大学、清华大学 EDP、FMBA 讲师。北大光华统计学博士，人大统计学学士，美国芝加哥大学访问学者。发起创办了微量网、量邦科技、量客投资等多家公司，目前担任北京量邦信息科技股份有限公司（835352），微量网公司，量客投资公司董事长。

摘要：介绍常见期货程序化交易策略的开发原理和几个常见策略，以及策略参数估计中常见的问题和统计学处理，抽样可以有效降低计算成本，而局部线性回归可以处理寻优后的参数曲面，更好的找到光滑区域，确定较好的参数。

建立基于 R 语言的后验系统

金戈（念空科技）
时间：14:30~15:00 邮箱：*jinge@gokudata.com*

简介：北京大学物理学学士，弗吉尼亚大学物理学博士。现任念空科技基金经理。曾任职千禧基金分析师、鸣石投资基金经理。

摘要：近 5 年国际国内市场投资经历，精通各种市场中性量化选股模型。对各种市场因子有深入独到的运用，并且利用国内外先进的风控系统严格控制并且对冲市场风险。善于使用数量模型发掘市场规律，探查市场风险，合理稳健的操作投资组合。使用 R 语言对股票，股指期货等各种投资标的进行数据处理，回测，交易，清洗等量化投研和交易工作。1）使用 R 进行平行计算，用以测算多因子绩效；2）调用择时策略所需要的技术指标；3）股票多空策略中的有约束条件的多因子优化；4）使用 R 进行数据库管理

解密高频交易

任坤（深圳凌云至善科技有限公司）
时间：15:00~15:30 邮箱：*renkun@outlook.com*

简介：毕业于厦门大学金融系和王亚南经济研究院，目前在深圳从事量化策略研发和工具开发的工作。

摘要：在金融交易领域中，高频交易经常备受争议，被许多人认为是扰乱和操纵市场的重要源头。该报告介绍了简单高频交易策略的金融原理、技术框架以及存在的一些难点和相关问题。

让投资研究更简单——R 与投资研究

林伟林（况客科技（北京）有限公司）
时间：16:00~16:30 邮箱：*linwl@hdinvesting.cn*

简介：况客科技联合创始人、汇迪投资管理 CEO，曾就职于 JT Capital 和博时基金固定收益部。

摘要：金融科技近年来成为最热的方向之一。随着云计算、可视化、去中心化等计算机相关技术的发展，计算机技术对金融的渗透已经从以“余额宝”为代表的销售渠道，逐步深入到金融机构日常的投资研究当中。况客科技是成立于 2015 年的一家金融科技创业公司，公司致力于通过互联网、大数据和人工智能的前沿技术提高资产管理行业的投资研究水平，让投资研究变得更加简单。作为一门简单实用的科学计算语言，本次演讲我将会介绍 R 在况客一体化投研平台上所扮演的重要角色。

用 R 语言进行量化风控

李翛然（北京奇点创世信息技术有限公司）
时间：16:30~17:00 邮箱：*lixiaoran@gingularity.com*

简介：干过精算，跑过投行。现在潜心为广大人民群众出谋划策，管好钱袋子。少赔点，多赚点。

摘要：(1) 量化风控初步介绍，个人及机构进行量化风控的意义何在
(2) 为什么选择 R 进行量化风控
(3) 实例展示如何进行风险量化及金融参数计算
(4) 我国二级市场的特点及描述性分析
(5) 如何将分析型工具 R，拓展成服务器架构性的风控系统

乐透彩卷的投资策略与回测效果

张家齐（木刻思股份有限公司）
时间：17:00~17:30 邮箱：*c3h3.tw@gmail.com*

简介：Founder of Taiwan R User Group & MLDM Monday

摘要：在这个演讲中，我们将详细探索
(1) 投资策略的各种重要参数
(2) 乐透彩卷的各种时间 pattern 与空间 pattern
(3) 如何运用乐透彩倦的 pattern 制定投资策略
(4) Trend Following Versus Mean Reversion
(5) 如何控管整体” 疯险” 与资金部位

电信网络中的 KQI 和 KPI 的异常检测

张建锋（华为技术有限公司）
时间：14:00~14:30 邮箱：zhangjianfeng3@huawei.com

简介：张建锋，2010 年毕业于四川大学数学系，获得统计学学士学位；2014 年加入华为中央研究院诺亚方舟实验室，从事电信领域大数据的研究和应用工作。

摘要：在现代生活中，电信网络已经成为和自来水系统等一样重要的基础设施，方便了每个人的生活。在网络的运维过程中，为了了解一个小区的网络运行状态，电信网络定义了很多统计指标（KQI 和 KPI）。当某个小区的网络出现异常时，这些指标相应地也会表现出一些异常。通常，网络工程师通过分析这些指标来管理整个网络。由于通信网络的规模日益庞大，一个中等规模的城市都拥有数万个小区。我们提出了一个网络 KQI、KPI 的异常检测算法，用以辅助网络工程师对全网小区的数据进行快速的异常检测，寻找 KQI 和 KPI 之间的异常关系，帮助网络工程师快速定位问题，提升网络运维效率。

工业大数据分析实践分享

田春华（*K2Data*）
时间：14:30~15:00 邮箱：tianchunhua@k2data.com

简介：昆仑智汇数据科技（北京）有限公司首席数据科学家，2004 年 1 月清华大学自动化系博士毕业。2004 年——2015 年在 IBM 中国研究院工作，负责数据挖掘研究和产品工作，分析应用成果在美国西南航空、香港水务署、韩国能源、和记黄埔等国际领先企业实施应用，发表学术论文（长文）82 篇（其中第一作者 42 篇），拥有 36 项专利申请（10 项已授权）。研究兴趣是数据挖掘算法与应用。

摘要：随着物联化和智能制造的推进，工业大数据成为互联网、计算机（如电信）大数据之外的一种重要的大数据应用类型。本报告将基于重型机械、风电、石油石化的 6 个客户应用案例，讨论工业大数据与商务大数据的差异，总结其关键时序或时空模式挖掘算法，以及对大数据平台在时序压缩、时空模式查询、分析并行化等方面的需求。最后，介绍我们在时间序列特征提取的 R package、基于 map-reduce 的 R 分析任务并行化引擎等方面的一些工作。

空间统计模型在半导体制造质量研究中的应用

王好（清华大学）
时间：15:00~15:30 邮箱：wh14@mails.tsinghua.edu.cn

简介：本科就读于天津大学工业工程系，现就读于清华大学工业工程系，博士二年级在读，研究方向为半导体制造和 3D 打印过程的统计质量模型

摘要：集成电路（芯片）是一种被广泛应用于工业生产和日常生活中电子元件。由于其结构紧凑，功耗较低，工作效率高，便于自动化控制等优点，集成电路产业在近几十年中得到了迅速的发展。芯片通常以半导体硅片 (也称晶圆) 为基础逐层加工，基于一定的生产工艺，元件层状分布在硅片表面上以满足一定的电路需求。在实际生产中，硅片表面可划分为若干（数十至数百个）区域，每个区域中包含一个集成电路芯片。集成电路的生产过程，涉及到数百步骤，需持续若干星期。随着制造技术的发展，单一硅片表面可生产的芯片数量规模

也逐渐增加。而由于空间位置的邻近，芯片之间必然存在着强烈的空间相关性。空间相关性的存在打破了传统统计模型各采样点“独立”的假设。这种新的数据特征对半导体制造行业的统计质量模型提出了挑战。空间相关性模型在疾病统计、气候研究以及生物分布等领域已经有了较广泛的应用，而在半导体制造行业中的研究还并不充分。以硅片表面的芯片缺陷为研究对象，结合空间统计学知识，可以建立包含空间信息的良品率模型。该模型可以并通过 R 语言进行建模和求解，并与现有模型对比。

制造系统中利用传感数据对生产过程的监测与诊断

张玺（北京大学）
时间：16:00~16:30 邮箱：*cristian1120@163.com*

简介：从事对复杂系统的工程数据分析和相关建模工作，达到对系统的有效监控、诊断和优化。研究成果在先进制造系统、公共医疗系统等领域得到良好应用。

摘要：随着先进传感技术和大规模数据采集技术的快速发展，在先进制造系统中采集各环节的生产数据已经成为可能，而利用这些传感数据实现生产过程的在线监控与诊断，已经成为先进制造领域的研究热点之一。本报告主要针对生产过程中产生的各类传感数据，结合工程物理知识，对生产工况进行有效识别和监测。

Reliability Optimization for Series Systems under Uncertain Component Reliabilities in the Design Phase

彭皓（中科院数学与系统科学研究院）
时间：16:30~17:00 邮箱：*penghao@amss.ac.cn*

简介：Hao Peng is an Assistant Professor in the Academy of Mathematics and Systems Science, Chinese Academy of Sciences. She received the Ph.D degree in Industrial Engineering from the University of Houston, Houston, TX in 2010. She received her Bachelor degree in Industrial Engineering from Tsinghua University, Beijing, China (2006). Her research interests are optimization for condition-based maintenance, quality and reliability engineering for evolving technologies. She was awarded the Marie Curie career integration grant from European Commission in 2012. She is a member of INFORMS and IIE.

摘要：We develop an optimization model to determine the reliability design of critical components in a serial system. The system is under a service contract, and a penalty cost has to be paid by the OEM when the total system down time exceeds a predetermined level, which complicates the evaluation of the expected cost under a given reliability design. Furthermore, in the design phase for each critical component, all possible designs are subject to uncertain component reliability. We propose three evaluation methods which take different types of uncertainty into account. Numerical results show that the full uncertainty method which includes the randomness of the number of failures as well as the randomness of the failure rates performs very well. We also show that ignoring the two types of uncertainty results in bad design decisions.

个性化制造让定制不再奢侈

谢帅（百分点集团）
时间：17:00~17:30　邮箱：*shuai.xie@baifendian.com*

简介：百分点集团咨询顾问，中国科学院大学 MBA。多年制造过程服务优化、供应链管理咨询经验。精益六西格玛黑带认证、TQM 认证、ITIL v3 Expert。目前从事大数据应用与智能制造相关咨询工作。

摘要：作为人类社会的支柱产业，制造业正受到大数据时代的巨大冲击。无论是德国提出的工业 4.0，还是《中国制造 2025》，都明确指出大数据已成为下一次工业革命中的关键技术。然而如何利用大数据等新一代信息技术，推动工业化和信息化的两化融合，是目前亟待解决的重要议题。随着互联网时代的到来，要求企业从以自身为中心转变为以用户为中心，从大规模制造转变为个性化定制，大数据将在这样的转变中起到关键性作用。本次演讲围绕制造产业链——从研发设计到生产运营，如何应用画像技术和数字协同等大数据手段实现个性化定制，并初步达成融合工业化生产和信息化协同的管理目标。

SparkR 的最新进展和趋势

孙锐（*intel*）
时间：14:00~14:30 邮箱：*rui.sun@intel.com*

简介：孙锐，英特尔上海大数据架构师。Hive，Spark 开源项目贡献者，SparkR 主力贡献者之一。

摘要：Apache Spark 从 1.4 版本加入 R 语言 API，为 R 社区提供了分析处理大数据的新手段。历经 1.5，1.6 以及即将到来的重要的 2.0 版本，Spark 社区一直在为 SparkR 贡献新的特性，提高 SparkR 的易用性和性能，同时 SparkR 的生态系统也有了一些发展。本演讲将介绍 SparkR 最新的特性和状态，重点是 UDF（在 DataFrame 上应用用户定义的 R 代码）和机器学习的算法，同时也将探讨它的发展趋势。

基于 GPU 异构集群的大规模分布式深度学习算法优化

颜深根（香港中文大学）
时间：14:30~15:00 邮箱：*yanshengen@163.com*

简介：颜深根博士毕业于中国科学院大学，是香港中文大学博士后，曾就职于百度研究院，现任 SenseTime 总监级主任研究员。研究兴趣包括大规模异构并行，深度学习，图像识别等。曾于 2013 年 6 月至 2014 年 2 月在美国北卡罗来纳州立大学访问交流。博士期间发表的两篇论文被并行计算领域顶级会议 PPoPP 13 和 PPoPP 14 分别录用。在百度期间主要负责大规模深度学习训练系统建设，大规模深度学习算法优化，另外在博士期间参与了《OpenCL 异构计算》一书的翻译及作为核心成员参与了 OpenCL 版本 OpenCV 的开发。

摘要：深度学习的出现极大的促进了机器学习相关领域的发展，其在视觉，语音，自然语言处理等诸多领域的成功应用，掀起了新一轮的人工智能热潮。相较于之前的机器学习算法，深度学习的一个重要特点在于通过大量的训练数据，来自动的提取特征，因而在模型训练阶段需要消耗大量的计算资源，训练时间往往长达数星期、甚至几个月。异构集群的特点在于使用更少的节点数，更低的能耗来提供更强的计算能力，因此非常适合于深度学习领域。由于异构集群的复杂性，如何将深度学习算法高效的映射到硬件上，是一个非常困难的问题。我们的工作采用基于 GPU 的异构集群，通过一系列的优化手段，最终获得了 32 个 GPU 相对于单 GPU 接近 25 倍的收敛加速。

Stochastic Dual Coordinate Ascent with Adaptive Probabilities

Qu Zheng（*The University of Hong Kong*）
时间：15:00~15:30 邮箱：*zhengqu@maths.hku.hk*

简介：My research focuses on developing and analyzing novel optimization algorithms capable of solving big data problems. Algorithms aspiring to achieve this goal must be highly granular and parallel / distributed in nature so as to exploit the power of modern high performance computer systems. Modern optimization methods need to address novel challenges brought up by the big data nature of the problems and need to rely on elements such as acceleration techniques, randomization, asynchronicity and communication avoiding strategies.

摘要：In this talk we present an adaptive variant of stochastic dual coordinate ascent (SDCA) for solving the regularized empirical risk minimization problems. Our modification consists in allowing the method to

adaptively change the probability distribution over the dual variables throughout the iterative process. Our method achieves provably better complexity bound than SDCA with the best fixed probability distribution, known as importance sampling. However, it is of a theoretical character as it is expensive to implement. We also propose a practical variant which in our experiments outperforms existing non-adaptive methods.

大规模机器学习及其应用

周俊（蚂蚁金服）
时间：16:00~16:30　邮箱：*jun.zhoujun@alibaba-inc.com*

简介：从事大规模机器学习的相关工作。

摘要：大数据给机器学习带来了很大的机遇和挑战。面向大数据量的机器学习，通常需要设计分布式系统跟算法来处理上百亿特征和数据。本报告将分享大规模机器学习的技术与过程，介绍大规模机器学习面临的问题以及在阿里的应用。

设计模式选讲：以 caffe 为例

骆颇（复旦大学）
时间：16:30~17:00　邮箱：*12110240007@fudan.edu.cn*

简介：复旦计算机学院计算机视觉方向研究生。

摘要：当架构一个领域专用框架的时候，我们既要考虑要当前的需要，也要为未来一定时间的需求变化留出空间。赋予系统拥抱变化的能力是否有一些实践经验可循呢？本报告将会以 caffe 为例介绍部分设计原则和设计模式，包括但不限于 SOLID 原则和 responsibility chain，composite，builder 等模式

What Does That P-value Mean?

沈侠（*University of Edinburgh*）
时间：14:00~14:30 邮箱：*xia.shen@ki.se*

简介：Dr Xia Shen is a statistical geneticist who received his PhD from Uppsala University. He is recently appointed as a Chancellor's Fellow (assistant professor) at the University of Edinburgh and also works part-time as a PI at Karolinska Institutet. Dr Shen has developed various statistical tools and conducted novel analyses in genetic studies, e.g. the hglm and bigRR packages for random effects modeling with applications in high-throughput genomic data analysis; genome-wide association analysis of genetic variance heterogeneity; multivariate methods in genome-wide association analysis. Dr Shen is also a developer of the GenABEL project for statistical genomics.

摘要：Scientific discovery via data analysis is of central importance in applied statistics. Modern big data science requires even more such effort, thus multiplicity becomes a ubiquitous issue in statistical inference using high-dimensional data. This presentation investigates the topic of multiple testing and points out different key perspectives of the use or misuse of inferential statistics such as p-values. Multiple empirical examples are given, including the applications in traditional epidemiology and high-throughput genomic data analysis for gene discovery. Highlighted here is the gap between statistical inference in science and conventional mathematical statistics, from which we should clearly emphasize the importance of focusing on data-driven rather than math-oriented statistics in education.

R 语言在医疗人工智能的应用

李舰（统计之都）
时间：14:30~15:00 邮箱：*lijian.pku@gmail.com*

简介：李舰，统计之都的核心成员之一，R 语言社区的活跃用户，开发了 Rweibo、Rwordseg、tmcn、Rofficetool 等包，也是中国 R 语言会议的组织者之一。撰写了《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）》和《机器学习与 R 语言》。专注于数据科学在行业里的应用，在制药医疗、零售快消、工业制造等领域有丰富的实践经验。

摘要：近年来，大数据医疗越来越热，人工智能也如火如荼，深度学习、GPU 计算等先进技术也进入了寻常百姓家。这些领域的结合是如今大数据医疗的方向。本次演讲将会根据真实的应用场景举例，介绍 R 语言和这些前沿技术的融合及在医疗健康领域发挥越来越重要的作用。

用数据撰写每个家族的传奇

陈钢（*WeGene*）
时间：15:00~15:30 邮箱：*cg@wegene.com*

简介：陈钢，WeGene 联合创始人兼 CTO，中国计算机学会生物信息学专业组创始委员。2006 年至 2012 年于中南大学获计算机博士学位。2009 年至 2010 年在德克萨斯大学医学部做访问学者。2012 年加入深圳华

大基因，历任华大基因研究院研究员、华大科技云平台部门副总监，以及负责云计算及互联网业务的副总裁。2013 年起担任香港中文大学生物信息学兼职助理教授。2015 年 4 月以联合创始人身份加入 WeGene，担任首席技术官。2016 年起担任香港医管局大数据顾问。在工作学习期间，翻译并出版《R 语言实战》、《Web 智能算法》等技术书籍十余本，在国际学术会议和期刊上发表生物医学数据分析相关论文十余篇。

摘要：2016 年 4 月份在 Nature Genetics 上发表的论文通过对 1244 个男性的 Y 染色体测序数据，推算出现代人类的祖先“亚当”诞生于 19 万年前；牛津大学关于成吉思汗 Y 染色体的研究表明，现在全球约有 1600 万成吉思汗的后代；全球大部分现代人类的基因组中都有另一个人科物种尼安德特人的基因组片段，并给我们带来了抑郁，尼古丁成瘾……

古人类和现代人类的基因组数据让我们得以从分子的层次上去窥探每一个现代智人的历史。DNA 在一代一代人之间的传递，也记录下了每一个家族的传奇。

大规模基因组数据的积累，以及数据存储和分析技术的进步，使得我们可以挖掘出每一份基因组数据中所蕴藏的家族传奇。我将跟大家分享自己的基因组中的故事，所涉及的数据分析工具，算法和模型，以及这个事情的价值和意义。

癌症液体活检简介

颜林林（北京大学生命科学学院）
时间：16:00~16:30　邮箱：*yanll@pku.edu.cn*

简介：颜林林，本科毕业于厦门大学生命科学学院生物专业，现为北京大学生命科学学院生物信息专业博士研究生，研究方向为人类基因组学中的高通量测序技术应用及数据分析。拥有十多年软件开发项目实战经验，涉足视频点播、图像处理、数据库信息系统、网络安防等多个领域，技术上擅长 C/C++ 语言编程。崇尚开源，故对 Linux、R 语言等都非常热衷。

摘要：液体活检，是从血液等体液中提取癌细胞 DNA，用以获得癌症相关信息的检测技术。由于其无创伤和高灵敏等特点，近年来被众多公司和机构热捧，被誉为“癌症诊断领域的颠覆性技术”。本次报告将会介绍液体活检的基本原理及当前进展，并从生物信息学和统计学的角度，阐述该领域现在面临的诸多问题，以及作为非医学专业人士，如何利用大数据技术，共同参与到这场人类与癌症持续了几十年的战争中来。

DNA 精准捕获

屈武斌（艾吉泰康生物科技（北京）有限公司）
时间：16:30~17:00　邮箱：*quwubin@gmail.com*

简介：艾吉泰康生物科技（北京）有限公司，合伙人，生物信息负责人。2009-2015 年在军事医学科学院从事 PCR 引物、探针方面的研究，发表相关学术论文 IF $>$ 40，参与编著《PCR Primer Design》，主持国家课题 2 项。

摘要：和全基因组测序不同，靶向捕获测序技术是一种可以对基因组中感兴趣的特定区域进行有效测序的方法，具有快速、经济、测序深度高且容易分析等优势。本次报告将重点介绍基因组精准捕获技术中的核心技术，即基于液相探针杂交的捕获和基于多重 PCR 扩增的捕获技术；同时介绍我们如何借助云计算实现一键完成超过 2000 个样品的数据分析。

挖掘数据商业价值，助力企业精准决策

杜晓梦（百分点集团）
时间：14:00~14:30 邮箱：*xiaomeng.du@baifendian.com*

简介：百分点集团数据科学总监，北京大学营销模型专业博士。专长于营销模型、消费者行为预测、互联网广告、社会媒体营销；擅长大数据统计建模及数据挖掘，精于归因模型、流失预警模型、社会网络分析等大数据商业模型，于 INFORMS Marketing Science Conference 等国际顶级学术会议上发表研究报告。

摘要：本次分享结合企业产品、销售、营销、服务等业务部门痛点，阐述如何运用大而灵活的数据整合分析能力，赋予企业新的核心市场竞争力。另外，将介绍如何通过对海量多源异构数据的整合、挖掘和应用，充分提升大数据商业应用价值，即以数据驱动产品研发设计、精准营销、提升消费者服务体验等等。同时也将分享大数据指导企业决策的应用场景和实际案例。

大数据分析的道与术

毕然（百度）
时间：14:30~15:00 邮箱：*biran1983@gmail.com*

简介：百度资深数据技术专家，在检索系统、在线广告、商业营销等领域有丰富的大数据分析和建模经验。曾因对百度的杰出贡献，获得首届百度百万美金最高奖，并多次获得技术创新奖。专注于理论与实践的结合，涉猎大数据技术、经济与商业机制、互联网产品战略、营销策略等多个领域，深究其根源并擅长跨界思考。乐于分享，百度技术学院的明星讲师，开设课程《大数据分析的道与术》、《经济学与互联网商业产品设计》和《机器学习的设计故事》等。著有《大数据分析的道与术》一书。

摘要：(1) 做好数据分析的四个关键
(2) 我们能相信统计吗?
(3) 大数据中“大”的价值
(4) 数据分析的实用方法

大数据环境下的信用风控技术实践

陈浪仙（百融 (北京) 金融信息服务股份有限公司）
时间：15:00~15:30 邮箱：*langxian.chen@100credit.com*

简介：10 年数据与信息挖掘相关的工作经验，持续跟踪数据存储处理和信息挖掘的技术进展。一直在进行技术和数据相结合以解决业务问题的工作，以将数据中的信息以更友好的方式和更清晰的表达通过产品和系统交付给用户。

摘要：新兴互联网金融如何利用数据和机器学习技术提高效率降低风险，泛征信数据如何在国内的实际产业环境中得到应用。

市场风险管理系统建置与开发

李宜熹（台湾高雄第一科技大学）
时间：16:00~16:30 邮箱：*eclee@nkfust.edu.tw*

简介：李宜熹，台湾中山大学财务管理博士、中南大学管理科学博士生，主要专长为金融风险管理、金融资讯探勘与金融资讯系统开发。目前就职于台湾高雄第一科技大学金融系，在进入学术界之前，曾协助台湾的金融业，开发金融风险管理与投资组合管理系统。

摘要：本讲题主要借由讲者过去协助金融业建置市场风险管理资讯系统的经验，说明市场风险管理模型导入金融产业的步骤与验证程序，并说明及展演系统雏形架构与模组功能。内容涉及金融资料源的 ETL、市场风险模型 (历史模拟法、时间加权法、波动率调整法、极值理论等之组合应用、以及经济情境产生器 ESG 的开展与应用价值)。本讲题的内容强调统合金融学 (知识领域)、统计与数学及信息工程等三个领域的重要性与发展性。

从大数据到智慧数据——电信企业大数据营销价值发掘和应用

漆晨曦（中国电信股份有限公司广东研究院）
时间：16:30~17:00 邮箱：*qicx@gsta.com*

简介：女，中国电信股份有限公司广东研究院市场运营研究所副所长，高级经济师，统计学硕士，从事电信数据分析、数据挖掘、精确营销、BI 架构及规范等专业研究 19 年。是国内通信行业第一个数据仓库广东电信市场经营分析系统的架构和规范设计者；国内通信行业第一本市场经营分析著作《电信市场经营分析方法与案例》的第一作者；牵头翻译（第一作者或独立作者）社交网络分析、大数据及相应大数据支撑的营销应用方面专业书籍 7 本，均由人民邮电出版社出版。

摘要：报告内容分成两部分，第一部分先跟大家分享我对大数据是什么，大数据价值何在的理解，在此基础上回到电信企业本身，追根溯源看看电信行业数据分析从最早报表分析再到支撑精确营销的数据挖掘高级分析及其分析平台的发展历史，然后看看大数据为电信企业带来的新机会；第二部分，先跟大家回顾电信企业精确营销概念，以及支撑精确营销理念落地过程中相应 IT 系统的发展演变，最后回到大数据支撑客户连接营销转型的营销战略新趋势。

互联网个人信用评估研究——基于不平衡样本视角

姜天英（山西财经大学）
时间：17:00~17:30 邮箱：*1224627533@qq.com*

简介：姜天英，山西财经大学统计学院 2014 级研究生，研究方向: 应用统计学和大数据指数分析。

摘要：消除信用信息的不对称性，搭建有效的企业、个人信用体系，是互联网征信可持续发展的基础和保障。本研究分别运用随机过抽样、SMOTE 方法、LLE+isomap 方法以及随机过抽样 +LLE+isomap 的方法进行重抽样，而后建立决策树、支持向量机和随机森林模型，对互联网个人信用进行评估。研究表明互联网大数据背景下的个人信用评估研究具有可行性且过抽样方法较好地提高了模型的分类性能以及随机森林分类效果最

优，并总结出信用等级好的用户的一般特征；同时基于变量重要性的探索，反驳了“变量越全面结果越准确”的说法。

Drug-Target Interaction Prediction by Integrating Multiview Network Data

张淑芹（复旦大学）
时间：14:00~14:30 邮箱：*zhangs@fudan.edu.cn*

简介：复旦大学数学科学学院副教授。主要研究兴趣是统计学、计算数学、最优化方法及其在生物信息学、分子生物学等方面的应用。

摘要：Drug-Target interaction prediction is a key step in further drug repositioning, drug discovery and drug design. Many mathematical models and statistical computation algorithms have been developed to identify potential drug-target pairs. In this work, we proposed a novel method to predict drug-target interactions by integrating different types of data. We applied our algorithms to LINCS L1000 database and DrugBank 3.0, and compared our algorithms with other algorithms. The evaluation results show our method can produce high prediction accuracy within short time. Finally, we predicted 54 possible drug-target interactions.

A weighted Empirical Bayes Risk Prediction Model using Multivariate Traits for Sequencing Data

李更新（*Wright State University*）
时间：14:30~15:00 邮箱：*gengxin.li@wright.edu*

简介：Gengxin Li has completed her Ph.D from Michigan State University and postdoctoral studies from Yale School of Public Health. She is an Assistant Professor in the Department of Mathematics and Statistics at Wright State University. Her research interests: statistical genetics and bioinformatics.

摘要：Empirical Bayes classification method is a powerful risk prediction approach for microarray data, but it is challenging to apply this method to risk prediction area using the exome sequencing data. A major advantage of using this method is that the effect size distribution for the set of possible features is empirically estimated and that all subsequent parameter estimation and risk prediction is guided by this distribution. Here, we generalize Efron's method to allow for some of the peculiarities of the exome sequence data. In particular, we incorporate quantitative trait information to binary trait prediction model, and a new model, named Weighted Empirical Bayes Multiple Traits Model, is proposed and we further allow this model to properly incorporate the annotation information of single nucleotide polymorphisms (SNPs). In the course of our analysis, we examine several aspects of the possible simulation model, including the identity of the most important genes, the differing effects of synonymous and non-synonymous SNPs, and the relative roles of covariates and genes in conferring disease risk. Finally, we compare the proposed methods to other classifiers.

Incorporating network information to prioritize results in genome wide association studies

侯琳（清华大学）
时间：15:00~15:30 邮箱：*lin_hou@163.com*

简介：侯琳博士于 2011 年获得北京大学统计学博士学位，2012 年进入耶鲁大学生物统计系从事研究工作，历任博士后、副研究员。2015 年 9 月加入清华大学统计学研究中心任助理教授。侯琳博士的研究兴趣包括统计学以及统计学在生物大数据和个体化医疗中的应用，包括统计遗传学中全基因组关联分析的研究；下一代测序数据的建模和分析；癌症基因组学；大规模生物相互作用网络的研究等。

摘要：Although Genome Wide Association Studies (GWAS) have identified many susceptibility loci for common diseases, they only explain a small portion of heritability. It is challenging to identify the remaining disease loci because their association signals are likely weak and difficult to identify among millions of candidates. One potentially useful direction to increase statistical power is to integrate other information, such as functional genomics and gene pleiotropy, to prioritize GWAS signals. In this talk, we will discuss the methods we developed recently for post-GWAS prioritization. We use Markov random field framework to incorporate biological networks. Applications to real data demonstrated that our method can identify more replicable genes, and the prioritized genes are enriched in disease related pathways.

Are all Transcription Factors Interacting with Each Other?

魏颖颖（香港中文大学）
时间：16:00~16:30 邮箱：*ywei@sta.cuhk.edu.hk*

简介：魏颖颖于 2009 年毕业于清华大学数学系，2014 年在约翰霍普金斯大学获得计算机科学硕士学位及生物统计博士学位，同年起在香港中文大学统计系任助理教授。

摘要：Understanding the interactions of different transcription factors (TF)s is a crucial first step toward deciphering gene regulatory mechanism. With advances of high-throughput sequencing technology, the genome-wide binding sites of many TFs have been profiled under different biological contexts. It is of great interest to quantify the interactions among different TFs. Analyses of the overlapping patterns of binding sites have been widely performed, mostly based on ad hoc methods. Due to the heterogeneity and the tremendous size of the genome, such methods often lead to biased even erroneous results. In this talk, by surveying the huge amount of sequencing data accumulated in public databases, we discover a Simpson's paradox type of phenomenon in assessing the genome-wide spatial correlation of TF binding sites. Built upon such observations, we propose a testing procedure for evaluating the significance of overlapping from a pair of proteins, which accounts for background artifacts and genome heterogeneity. Furthermore, to characterize the co-activation patterns among TFs across the genome and under diverse biological conditions, we propose a dynamic Poisson graphic model, which can be applied to a large class of multivariate counts data.

A Dirichlet-tree Multinomial Regression Model for Associating Dietary Nutrients with Gut Microorganisms

王涛（上海交通大学）
时间：16:30~17:00 邮箱：*neowangtao@sjtu.edu.cn*

简介：2007 年东南大学数学系学士，2010 年华东师范大学金融与统计学院硕士，2013 年获香港浸会大学数学系统计学专业哲学博士学位。2014 年赴美国耶鲁大学公共卫生学院生物统计系从事博士后研究工作，2016 年 1 月回国任上海交通大学特别研究员。

主要致力于研究高维复杂数据的统计降维技术和变量选择技术，以及研究生物医学数据的统计分析方法。近年来分别在 Journal of the Royal Statistical Society: Series B、Biometrika、Biometrics、Bernoulli、Statistics and Computing、Statistica Sinica 等知名学术期刊上发表 SCI 论文近二十篇。

摘要：Understanding the factors that alter the composition of the human microbiota may help personalized healthcare strategies and therapeutic drug targets. In many sequencing studies, microbial communities are characterized by a list of taxa, their counts, and their evolutionary relationships represented by a phylogenetic tree. In this paper, we consider an extension of the Dirichlet multinomial distribution, called the Dirichlet-tree multinomial distribution, for multivariate, over-dispersed, and tree-structured count data. To address the relationships between these counts and a set of covariates, we propose the Dirichlet-tree multinomial regression model for which we develop a penalized likelihood method for estimating parameters and selecting covariates. For efficient optimization, we adopt the accelerated proximal gradient approach. Simulation studies are presented to demonstrate the good performance of the proposed procedure. An analysis of a data set relating dietary nutrients with bacterial counts is used to show that the incorporation of the tree structure into the model helps increase the prediction power.

IPAC: A Flexible Statistical Approach to Integrating Pleitoropy and Annotation for Characterizing Functional Roles of Genetic Variants that Underlie Human Complex Phenotypes

杨灿（*Hong Kong Baptist University*）
时间：17:00~17:30 邮箱：*eeyang@hkbu.edu.hk*

简介：Can Yang received the bachelor’s and master’s degrees in Automatics from Zhejiang University, China, in 2003 and 2006, respectively. He received the PhD degree in electronic and computer engineering from the Hong Kong University of Science and Technology in 2011. He worked as a postdoctoral associate (2011-2012) and an associate research scientist (2012-2014) at Yale University, New Haven, Connecticut. Now he is working as an assistant professor at department of mathematics, Hong Kong Baptist University. His research interests include statistical learning and bioinformatics. He was the Winner of the 2012 Young Scientist Award in Engineering Science.

摘要：Recent international projects, such as the Encyclopedia of DNA Elements (ENCODE) project, the Roadmap project and the Genotype-Tissue Expression (GTEx) project, have generated vast amounts of genomic annotation data measured at the multiple layers, e.g., epigenome and transcriptome. On the other hand, increasing evidence suggests that seemly unrelated phenotypes can share common genetic factors, which is known

as pleiotropy. A big challenge in integrative analysis is how to put pleiotropy and annotation into a unified model and automatically select most relevant genomic features from a potentially huge set of genomic features. In this talk, we introduce a flexible statistical approach, named IPAC, to integrating pleiotropy and annotation for characterizing functional roles of genetic variants that underlie human complex phenotypes. IPAC enabled us to automatically perform feature selection from a large number of annotated genomic features and naturally incorporate the selected features for prioritization of genetic risk variants. IPAC not only demonstrated a remarkably computational efficiency (e.g., it took about 2 3 minutes to handle millions of genetic variants and thousands of functional annotations), but also allowed rigorous statistical inference of the model parameters and false discovery rate control in risk variant prioritization. With the IPAC approach, we performed integrative analysis of genome-wide association studies on multiple complex human traits and genome-wide annotation resources, e.g., Roadmap epigenome. The analysis results revealed interesting regulatory patterns of risk variants. These findings undoubtedly deepen our understanding of genetic architectures of complex traits.

Deep Dive the In-database R

Mark Chen（微软）
时间：8:30~9:30 邮箱：*markchen@microsoft.com*

简介： Mark is an Advanced Analytics Tech Lead at Microsoft focused on Azure Machine Learning, Cortana Intelligence, Cognitive Services, Big Data, and Data Science tools and education.

He also works on enabling BI analysts to go from hindsight to foresight by integrating Excel, Power BI, and other tools with Azure ML web services and Microsoft R Server.

摘要： In this sessions, you'll learn an end-to-end solution for predictive modeling using SQL Server R Services.

This sessions is based on a well-known public data set, the New York City taxi dataset. You will see a combination of R code, SQL Server data, and custom SQL functions to build a classification model that indicates the probability that the driver will get a tip on a particular taxi trip. You'll also learn your R model to SQL Server and use server data to generate scores based on the model.

Unified Term for R and Julia

宫雨（中国石油大学（北京））
时间：9:30~10:00 邮箱：*armgong@yahoo.com*

简介： 博士，副教授，在中国石油大学（北京）商学院从事管理信息系统和数据挖掘的教学和研究工作，业余时间喜欢编写与统计计算和数据分析相关的程序。

摘要： 传统上集成 R 和其它语言都是通过扩展包完成的，这种方式主要用来实现 R 和其它语言的交互。但在 R 中编写被嵌入的语言代码时，缺失了 REPL 的很多功能，如语法完成、结果输出等。

本次演讲讨论修改 R 的源代码，实现在 R term 中嵌入 julia 的 REPL，实现了两种语言编程环境的统一，通过这种方式使用者可以在一个 term 中 R 和 julia 的环境中切换，方便编程和调试，同时还支持 R 和 julia 变量和对象的交换。

具体的技术细节包括: (1) 截获 R term 的输入，实现 julia 的 REPL，完成 R 和 Julia 的 REPL 相互切换 (2) 在修改后 R term 中实现 julia 的语法完成、输出等功能 (3) 在不依赖扩展包的情况下，R 和 julia 之间的变量和对象交换

Microsoft R Server Overview

Louise Wong（微软）
时间：10:30~11:30 邮箱：*Louise.Wong@microsoft.com*

简介： 黄河燕是微软在大中华地区负责高级的分析和大数据解决方案的黑带团队。在此之前，在马来西亚负责微软 SQL Server 业务营销，有超过 20 年解决方案的经验。在加入微软之前，他曾于 IBM 和 Oracle 公司担任企业解决方案职务，并拥有电子工程学士学位以及工商管理硕士学位。

摘要： Microsoft R Server is the most broadly deployable enterprise-class analytics platform for R available today. Supporting a variety of big data statistics, predictive modeling and machine learning capabilities, R Server supports the full range of analytics – exploration, analysis, visualization and modeling based on open source R.

Welcome to start learn about Microsoft R Server today!

”R” vs ”Spark MLlib” 在信用评分技术中的应用

蒋卓（考拉征信）

时间：11:30~12:00 邮箱：*jiangzhuo@kaolazhengxin.com*

简介： 考拉征信数据分析师、统计学硕士。在金融领域和基因数据分析领域积累了丰富的经验。

摘要： 友好的对象化编程及活跃的开源环境使得 R 成为使用最受欢迎的数据分析语言。随着数据量的增大和变量的增多，为征信技术带来了新的挑战。Spark 凭借其分布式处理的通用性、高效迭代、容错能力等优势提出了新的解决方案。本报告首先介绍常用的信用评分技术，然后基于征信数据对 R 和 Spark 的建模效果进行对比。

监管风暴中的互联网金融

邓一硕（懒投资）
时间：8:30~9:00 邮箱：*dengyishuo@zuinianqing.com*

简介：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。

摘要：2015 年下半年开始，互联网金融行业负面新闻迭出，一时间曾被寄予改革和金融创新希望的互联网金融被媒体妖魔化，投资者人心浮动。在此背景下，监管部门着手制定监管意见，并自 2016 年开始对互联网金融行业进行整改，监管、从业、媒体等相关人士皆对互联网金融行业持观望态度。那么在判断其发展前景之前，应当首先厘清互联网金融到底是什么？它因何而生，如何蔚然成风？是风险源头，还是创新萌芽？未来互联网金融会怎么存在？在行至半途之时，回顾初衷也许非常重要。

中国金融运行面临转折性挑战

汪洋（江西财经大学）
时间：9:00~9:30 邮箱：*wangyang@jxufe.edu.cn*

简介：江西财经大学金融学院教授，博士生导师，研究方向为国际金融理论与政策。

摘要：本报告分析了 2014 年下半年以来中国货币政策面临的周期性转折，尤其是 2015 年 8.11 人民币汇率改革之后，中国外汇储备快速下跌之后，基础货币出现的根本性变化。对未来中国货币政策的框架选择进行了展望。

R 在 Online lending 中的数据化决策应用 –模型工具、数据产品、实时决策

张云松（融 *360*）
时间：9:30~10:00 邮箱：*stevenzys@hotmail.com*

简介：张云松，融 360 风控决策总监，多年金融风控决策分析、大数据征信产品、消费金融产品设计等行业经验，曾就职于 Experian、德勤、京东等企业。目前专注于将机器学习与数据挖掘模型应用于在线贷款类产品的获客、转化、反欺诈、授信决策、贷后和账户管理。

摘要：互联网金融中在线授信产品最近层出不穷，对这些小额信贷产品如何进行数据化风控决策是产品成败的核心，本次报告将介绍如何运用 R 构建模型开发评测体系，自动化建模提升数据模型的效率。同时，将分享大量 R、shiny、opencpu 等在 online lending 业务中模型策略、反欺诈模型、数据产品等应用中的实际案例。

同盾大数据反欺诈的实践与应用

张昊（杭州同盾科技有限公司）
时间：10:30~11:00 邮箱：*bo.ding@tongdun.cn*

简介：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。

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。本报告包括:

(1) 互联网反欺诈的特点及相关的模型演示
(2) 互联网产品创新所面临的各种欺诈风险问题和特点
(3) 大数据风控技术的反欺诈效果
(4) 大数据在反欺诈行业的应用前景

大小数据融合的金融营销建模

李慧（百分点集团）
时间：11:00~11:30 邮箱：*hui.li@baifendian.com*

简介：百分点集团数据建模分析师，本科时期主修统计学专业辅修金融，而后赴法国深造学习，并于法国雷恩一大、法国国立信息统计与分析学校（ENSAI）获得双校统计与计量经济学硕士学位。回国后在西门子中国研究院、埃森哲等公司实习，并在 15 年加入百分点供职至今。主要工作方向为金融领域的大数据建模。

摘要：金融行业拥有大量而高质的数据，但在国内，对金融数据的挖掘和应用并不充分。该主题主要结合金融行业数据与互联网多维数据一同解决金融领域营销问题，从交叉销售、潜客挖掘等角度阐述在大数据时代，基于大数据对客户进行细致分析和研究的意义。

NBA 运动彩卷分析

赵致平（木刻思股份有限公司）
时间：11:30~12:00 邮箱：*whizzalan@gmail.com*

简介：数学这种高深莫测的武功就是需要程式将具现化，没有尝试用资料验证过的事情不太相信，基本上我连自己也不太信。期望能在资料的世界中，找到我活著的意义。

摘要：简单招募大家来玩 NBA 的运动彩卷，常被认定邪恶庄家盘口介绍，会被骗钱的各种游戏玩法，以及认识生态圈后看要当食物链的那块物种利用爬虫行为爬取相关数据，并作一些分析提供给自己作下注的决策支援资讯。

An Overview of Genomic Data Analysis in Bioconductor

Martin Morgan（*Roswell Park Cancer Center*）
时间：8:30~9:00 邮箱：*martin.morgan@roswellpark.org*

简介： Dr. Morgan earned his undergraduate and Master's degrees in Botany at the University of Toronto. Dr. Morgan's PhD studies at the University of Chicago involved the evolutionary consequences of frequency-dependent selection, and of multilocus deleterious mutation.

Dr. Morgan spent 10 years as an Assistant and then Associate Professor at Washington State University, before joining the Fred Hutchinson Cancer Research Center in 2005. At the Hutch, Dr. Morgan worked on the Bioconductor project for the analysis and comprehension of high-throughput genomic data; he has led Bioconductor since 2008. Dr. Morgan recently moved to Roswell Park Cancer Institute in Buffalo, NY, where the Bioconductor project is now based.

摘要： Bioconductor is a collection of more than 1,100 individually code-reviewed software packages, hundreds more annotation and experiment data packages, and specialized data structures for high-throughput genomic (especially sequence) analysis. This talk will overview Bioconductor software and principles across several domains, especially sequence analysis. We will mention common work flows, statistical challenges, working with large data, and placing results in biological context. The principles apply to other areas where Bioconductor helps analysis, including microarrays (expression, methylation, copy number, SNP), flow cytometry, proteomics, and image analysis.

CRISPRseek and GUIDEseq packages for Designing Effective and Target-specific gRNAs and Assessing the Precision of Engineered CRISPR-Cas9 Genome Editing System

Lihua Julie Zhu（*University of Massachusetts Medical School, Worcester, MA, USA*）
时间：9:00~9:30 邮箱：*Julie.Zhu@umassmed.edu*

简介： Lihua Julie Zhu obtained her Ph.D. in Nutritional Sciences from the University of Wisconsin-Madison in 1999, and her M.S. in Computer Science from DePaul University in Chicago in 2001. She joined the Bioinformatics Core of the Robert H. Lurie Comprehensive Cancer Center (RHLCCC) of Northwestern University in 2001, where she was the Director of Bioinformatics Consulting Core from 2003 to 2005 and the Director of Clinical Informatics Group from 2005 to 2007. She joined the University of Massachusetts Medical School (UMMS) as a Research Assistant Professor in 2007. Since 2015, Dr. Zhu is a Research Professor, Head of Bioinformatics Core of the Department of Molecular, Cell and Cancer Biology (MCCB) of UMMS.

摘要： The most recently developed genome editing system, CRISPR-Cas9 has greater inherent flexibility than prior programmable nuclease platforms because sequence-specific recognition resides primarily within the associated sgRNA, which permits a simple alteration of its recognition sequence. The short Protospacer Adjacent Motif (PAM), which is recognized by Cas9, is the chief constraint on the target site design density. Because of its simplicity and efficacy, this technology is revolutionizing biological studies and holds tremendous promise for therapeutic applications. At least three companies have been founded to leverage this technology for therapeutic uses. However, imperfect cleavage specificity of CRISPR/Cas9 nuclease within the genome is a cause for concern

for its therapeutic application. To facilitate the adoption and improvement of this technology, we have developed CRISPRseek for designing target specific gRNAs, and GUIDEseq for identifying genome-wide offtarget sites from GUIDE-seq experiments to assess the precision of engineered CRISPR-Cas9 nucleases.

RNA-seq Analysis in Bioconductor

Charity Law（*The Walter and Eliza Hall Institute of Medical Research, Melbourne, Australia*）
时间：9:30~10:00 邮箱：*law@wehi.edu.au*

简介： Molecular Medicine Division, The Walter and Eliza Hall Institute of Medical Research, Melbourne, Australia.

摘要： Transcriptome sequencing is a popular application in functional genomics research, and the Bioconductor project hosts a wide collection of tools that are capable of performing a complete analysis, from read mapping, normalization and exploratory data analysis through to differential expression and pathway analysis. This talk will provide an overview of some of the most popular packages and showcase complete workflows for RNA-seq analysis (from raw data through to biologically relevant gene lists and pathways) as well as new developments in this area.

DNA methylation Array Analysis with MissMethyl & other Bioconductor Packages

Jovana Maksimovic（*Murdoch Childrens Research Institute, Melbourne, Australia*）
时间：10:30~11:00 邮箱：*jovana.maksimovic@mcri.edu.au*

简介： After completing a Bachelor of Science (Honours) / Bachelor of Bioinformatics at La Trobe University, majoring in biochemistry, genetics and computer science, Dr Maksimovic worked at the Department of Primary Industries for two years as part of their graduate program. During this time she worked on many diverse research projects and developed an interest in the biology of lactation.

She began her PhD at Monash in 2007 on a Department of Primary Industries project investigating the expression and regulation of a gene family involved in the production of a subset of milk oligosaccharides that are of particular interest in infant nutrition. She currently works as a Postdoctoral Scientist in the bioinformatics group at Murdoch Childrens Research Institute.

摘要： DNA methylation is one of the most widely studied epigenetic modifications due to its role in both development and disease. Since its release in 2011, the Illumina HumanMethylation450 (450k) array has been enthusiastically embraced by the epigenetics community as a cost-effective way to measure methylation at >450,000 sites across the human genome. This year, Illumina has increased the genomic coverage of the platform to >850,000 sites with the release of their EPIC array. In this talk, I will describe the general approach to methylation array analysis using popular Bioconductor packages to get from raw data to identifying biologically interesting methylation sites. I will focus particularly on our package, missMethyl, a suite of tools developed to address some of the challenges and biases inherent in the analysis of these arrays. Specifically, missMethyl has functions tailored for the arrays that range from normalisation and statistical testing, to a more whole systems

approach in the form of gene set testing. The key ideas will be described using practical examples from case studies.

Wrapping Your R tools to Analyze National-Scale Cancer Genomics in the Cloud

殷腾飞（*Seven Bridges Genomics*）
时间：11:00~11:30 邮箱：*yintengfei@gmail.com*

简介：爱荷华州立大学计算生物学博士，Seven Bridges Genomics 公司产品经理，个人主页 http://tengfei.name。

摘要：The Cancer Genomics Cloud (CGC), built by Seven Bridges and funded by the National Cancer Institute hosts The Cancer Genome Atlas (TCGA), that is one of the world's largest cancer genomics data collections. Computational resources and optimized, portable bioinformatics tools are provided to analyze the cancer data at any scale immediately, collaboratively, and reproducibly. Seven Bridges platform is not only available on AWS but also available on google cloud as well. With Docker and Common Workflow Language open standard, wrapping a tool in any programming language into the cloud and compute on petabyte of data has never been so easy.

The open source Bioconductor package "sevenbridges" is developed to provide full API support to Seven Bridges Platforms including CGC, supporting flexible operations on project, task, file, billing, apps etc., users could easily develop fully automatic workflow within R to do an end-to-end data analysis in the cloud, from raw data to report. The "sevenbridges" package also provides interface to describe your tools and workflow in R and make it portable to CWL format in JSON and YAML, that you can share easily with collaborators, execute it in different environment locally or in the cloud, everything is fully reproducible. Combined with the R API client functionality, users will be able to create a CWL tool in R and execute it in the Cancer Genomics Cloud to analyze the huge amount of cancer data at scale. This package is the hub to connect docker technology, CWL, R development environment, cloud execution and reproducible research.

ggtree for Visualization and Annotation of Phylogenetic Trees

余光创（香港大学）
时间：11:30~12:00 邮箱：*guangchuangyu@gmail.com*

简介：余光创，香港大学公共卫生学院博士生，研究方向为甲型流感病毒的进化。

摘要：Trees are commonly used to present the evolutionary relationships of species. As phylogenetic trees become more widely used in multidisciplinary studies, there is an increasing need to incorporate various type of information in the tree visualization. Users then require programmable software to allow high levels of customization and integration, rather than standalone applications that focus on specific analyses and data types. To fill this gap, the speaker has developed a Bioconductor package, ggtree. The ggtree package inherits versatile properties of ggplot2 and thus allows constructing complex tree views by freely combining multiple layers of annotations. This talk will cover the design of ggtree (why it is special and different from other packages based on ggplot2), getting trees (especially non-standard formats) into R, tree visualization and annotation with different types of data including the user's own experimental data.

张量大数据分析及其应用

徐增林（电子科技大学）
时间：8:30~9:00 邮箱：*zlxu@uestc.edu.cn*

简介：徐增林，电子科技大学教授、博士生导师，中组部“青年千人计划”入选者，现任电子科技大学大数据研究中心数据挖掘与推理研究所轮值所长，并创建统计机器智能与学习实验室 (Statistical Machine Intelligence and LEarning, SMILE, http://bigdatalab.weebly.com/)。主要研究兴趣为机器学习及其在社会网络分析、互联网、计算生物学、信息安全等方面的应用。他在包括 IEEE TPAMI, IEEE TNN，NIPS, ICML, IJCAI, AAAI 等顶级会议和刊物发表论文近 30 篇，引用近千次，发表专著 2 部，书籍章节 2 篇并于 2015 年的 AAAI 大会获得最佳学生论文奖提名。于 2012 年在多伦多召开的国际人工智能大会 (AAAI) 上做教学报告。徐增林教授是 JMLR，IEEE TPAMI 等机器学习与人工智能领域主要期刊的审稿人和香港教育资助局的基金评审人；多次担任人工智能领域的主要国际会议如 AAAI/IJCAI 等会议的程序委员会成员；多次担任机器学习和大数据研究方面的研讨会的组织委员会主席。

摘要：张量（矩阵向三维或更高维的扩展）广泛存在于许多现实系统中，例如社会网络中的 < 用户 -图片 -图片标记 -地理位置 > 和电子商务中的 < 用户 -商品 -交易 -时间 >。比起向量或矩阵表示，张量可以有效包含不同模之间的关联关系并有助于大大提高预测精度。本报告将介绍张量的基本概念、张量分解方法及其应用。重点探讨张量分解的贝叶斯方法，并探讨当前存在的主要问题及解决方案。

Recovering High-order Tensors from Highly Incomplete Observations: Models and Applications

王尧（西安交通大学）
时间：9:00~9:30 邮箱：*yao.s.wang@gmail.com*

简介：王尧，于 2014 年获西安交通大学应用数学博士, 现为西安交通大学数学与统计学院助理教授 (讲师)。博士期间获国家留学基金委资助，在乔治亚健康科学大学神经科学系与乔治亚理工大学工业与系统工程系各进行了 12 个月的学术访问。现主要研究方向为稀疏信息处理、高维统计推理、计算生物学等，相关研究成果分别发表于 IEEE TIP、IEEE TSP、中国科学等国内外知名杂志。

摘要：Various applications, e.g., video surveillance, hyperspectral imaging and dynamic MR image reconstruction, can be formulated as recovering a high-order tensor from highly incomplete observations. Despite an increasingly common interest, high-order tensor recovery remains a challenging problem because of the underlying complex structures of tensors. The existing approaches are developed through unfolding the tensor into different matrix forms and then using conventional matrix recovery techniques. Such matricization fails to effectively exploit the tensor structure and may lead to suboptimal procedure. The focus of the talk is on introducing some novel models to remedy this issue for three tensor related applications, i.e., background subtraction from compressive measurements, hyperspectral compressive sensing, and hyperspectral/multispectral image super-resolution. As compared with the existing models, our models are capable of characterizing extensive structures that underlie the high-order tensors, yielding better quality of recovery from generally fewer observations.

推荐系统中的机器学习实践

苏海波（百分点集团）
时间：9:30~10:00 邮箱：haibo.su@baifendian.com

简介：清华电子系博士毕业，他擅长文本分析、机器学习、个性化推荐以及计算广告学，多篇论文发表于国内外顶尖学术会议和期刊，曾经负责微博的信息流效果广告系统。

摘要：随着互联网信息的不断爆炸，个性化推荐已经成为解决信息爆炸的重要手段，推荐系统是目前互联网世界最常见的智能产品形式，其中机器学习在推荐系统中扮演着重要的角色，对于推荐的效果至关重要，本次分享将介绍百分点推荐系统中机器学习的发展历程，从单机协同过滤、实时分布式协同过滤到点击率模型等等，和听众分享百分点在这些方面的实践经验。

信用风险预测模型

杨滔（桃树（天津）科技有限公司）
时间：10:30~11:00 邮箱：847522835@qq.com

简介：杨滔，奥克兰大学机器学习博士，悉尼科技大学博士后，ALH 算法创始人，第十四届亚太数据挖掘会议最佳论文亚军奖得主。曾任阿里巴巴集团数据科学家和 F 团首席科学家。

摘要：基于 R 语言研发自动化信用风险预测建模工具。基于此工具，在多家银行成功落地。

Lasso Adjustments of Treatment Effect Estimates in Randomized Experiments

刘汉中（*University of California, Berkeley*）
时间：11:00~11:30 邮箱：lhz1985@berkeley.edu

简介：I'm a postdoctoral scholar working with Professor Bin Yu in the Department of Statistics at UC Berkeley. Prior to this, I visited UC Berkeley as a visiting student from 2012 to 2014 and I received my PhD in statistics from Peking University in 2014, advised by Professor Bin Yu and Jinzhu Jia. My research interests lie in statistical theory and methodologies for solving high-dimensional data problems and causal inference. Currently, I'm working on combing bootstrap and sparse modeling method (e.g., Lasso) to construct confidence intervals for parameters in high-dimensional sparse linear model, as well as on estimating causal effect of treatment using machine learning methods under Neyman-Rubin causal model.

摘要：We provide a principled way for investigators to analyze randomized experiments when the number of covariates is large. Investigators often use linear multivariate regression to analyze randomized experiments instead of simply reporting the difference of means between treatment and control groups. Their aim is to reduce the variance of the estimated treatment effect by adjusting for covariates. If there are a large number of covariates relative to the number of observations, regression may perform poorly because of overfitting. In such cases, the Lasso may be helpful. We study the resulting Lasso-based treatment effect estimator under

the Neyman-Rubin model of randomized experiments. We present theoretical conditions that guarantee that the estimator is more efficient than the simple difference-of-means estimator, and we provide a conservative estimator of the asymptotic variance, which can yield tighter confidence intervals than the difference-of-means estimator. Simulation and data examples show that Lasso-based adjustment can be advantageous even when the number of covariates is less than the number of observations. Specifically, a variant using Lasso for selection and OLS for estimation performs particularly well, and it chooses a smoothing parameter based on combined performance of Lasso and OLS.

机器学习在异常用户侦测上的应用

陈嘉葳（*Taiwan R User Group*）
时间：11:30~12:00　邮箱：*sk413025@gmail.com*

简介：Taiwan R User Group 共同组织人，是社群的长期志工，主要协助举办资料分析与机器学习相关议题的聚会，同时也经常在台湾的技术交流会议上分享机器学习相关技术。

摘要：异常用户侦测是机器学习在实际工作中很重要的应用。当我们在建模的时候，经常遇到许多问题，例如: 异常用户只占极少数，用户检举与专家提供的标记只有单一类别，或是用户恶意检举与人为疏失导致标记不可信任等等。本次演讲将会介绍讲者面对这些问题的一些经验。

智慧城市与城市大数据

王静远（北京航空航天大学）
时间：8:30~9:00 邮箱：*jywang@buaa.edu.cn*

简介：王静远，2011 年 7 月毕业于清华大学计算机系，获工学博士学位。现任北京航空航天大学计算机学院讲师。研究兴趣包括数据挖掘、城市计算、数据中心网络拥塞控制等智慧城市关键技术。主持国家自然科学青年基金一项、面上项目一项、重点项目子课题一项。参与国家自然科学基金、973 项目、863 项目、支撑计划项目多项。发表学术论文 20 余篇。申请中国专利 9 项，美国专利 2 项，AVS 标准提案一项。担任 SCI 期刊 Frontier of Computer Science 的 Administrator Editor。CCF 大数据专委会通讯委员。

摘要：本报告介绍了北航计算机学院智慧城市研究小组基于浮动车 GPS 轨迹数据、手机定位数据、公交一卡通数据、城市公共卫生病例数据等多源、异构的城市数据，进行的关于城市结构分析、城市交通规划、城市职住分析等研究工作。报告的内容涉及计算机科学同城市规划学科、城市地理学科、流行病学等不同学科之间的交叉融合，是大数据分析与挖掘技术在跨学科领域应用的一个很具有代表性的案例。

北京出行便利性研究

张忠元（中央财经大学）
时间：9:00~9:30 邮箱：*zhyuanzh@gmail.com*

简介：张忠元, 理学博士, 中央财经大学教授, 博士生导师, 中国计算机学会高级会员, 果壳网科学顾问。主要研究兴趣在复杂网络分析和数据挖掘。在 Data Mining and Knowledge Discovery, Physical Review E, EPL, Knowledge and Information Systems, Scientific Reports, 中国科学等国内外著名期刊上发表学术论文十余篇。爱思唯尔杰出审稿人, 担任 Data Mining and Knowledge Discovery, Physica A, Management Science 等著名期刊的匿名审稿人。

摘要：利用出租车数据分析了北京地区出行便利性，结果表明北京交通路网建设有可提高之处。

百度时空大数据上的智慧城市应用研究

周景博（百度）
时间：9:30~10:00 邮箱：*zhoujingbo@baidu.com*

简介：周景博于 2014 年从新加坡国立大学博士毕业，主要研究方向为时空数据挖掘，包括轨迹预测，交通流量预测等。周景博于 2015 年底加入百度研究院大数据实验室任数据科学家，从事百度时空数据和用户行为预测相关的研究工作。

摘要：以手机百度和百度地图为代表的百度移动应用每天会收到数十亿次的搜索和定位请求。本次演讲将介绍如何利用数据挖掘技术来探究百度时空大数据里所隐藏的信息，来帮助我们洞察变化中的真实世界。在此基础上，我们会进一步介绍若干基于百度时空大数据的智慧城市相关项目上的应用，包括智能选址，群体行为研究和城市空间量化分析等。

普适商务环境下的大数据分析

祝恒书（百度）
时间：10:30~11:00 邮箱：*zhuhengshu@baidu.com*

简介：祝恒书博士现任百度研究院大数据实验室数据科学家，负责百度商务智能大数据领域的科研工作。他分别于 2009 年和 2014 年在中国科学技术大学获得了计算机应用与技术专业学士和博士学位，期间他曾作为国家公派访问学者前往美国新泽西州立 -罗格斯大学商学院进学术访问。他曾获得了中国科学院院长特别奖（2014）、教育部学术新人奖（2012）等荣誉以及包括著名国际会议 ICDM-2014 等在内的 4 项最佳论文奖或提名奖。

他的研究方向是数据挖掘及知识发现中的关键技术和应用，主要包括普适商务环境下的大数据分析、移动智能计算以及计算社会学等方面。近年来，他在国际顶级学术期刊（如 IEEE TKDE、IEEE TMC、IEEE TC、ACM TIST）和会议（如 KDD、IJCAI、ICDM、CIKM）上发表学术论文 30 余篇，并作为发明人申请了超过 20 项国内外专利。他是 IEEE、ACM、CCF 会员，以及中国计算机学会大数据专家委员会首批通讯委员，并多次受邀担任了 KDD、IJCAI、ICDM 等大数据领域顶级国际会议的程序委员会委员和包括 IEEE TKDE 等在内的 10 余个国际著名学术期刊的审稿人。

摘要：大数据为深度理解用户、开发新型商务智能应用与服务带来了全新的机遇和挑战。本次报告将首先介绍大数据时代的相关背景知识，然后从理论与实践的角度详细的阐述了数据挖掘相关技术在大数据时代所扮演的重要角色，并结合报告人近年来在商务智能、智慧城市计算等领域的研究成果以及百度大数据实验室的部分应用实践进行了案例介绍，最后对数据挖掘技术在普适商务情境下的应用前景进行了展望。

你吐槽过的天气预报原来可以这样玩

罗应琏（北京维艾思气象信息科技有限公司）
时间：11:00~11:30 邮箱：*ALuo@WeatherData.com.cn*

简介：目前服务于北京维艾思气象信息科技有限公司，隶属中国气象局公众气象服务中心，负责气象大数据以及天气保险业务。过去在德勤 (Deloitte) 管理咨询与凯捷 (Capgemini) 管理咨询担任高级经理 (Senior Manager) 职务参与世界五百强的联想集团、金士顿科技。

摘要：天气的变幻莫测会为很多行业带来风险，但在大数据时代下，目前气象数据不仅可以让人们知道分钟级的天气预报、了解气候，也为城市规划、防灾减灾、工程设计和环境评估等决策提供了数据的支撑，还能帮助保险、旅游等行业减少损失、提高盈利。此外，互联网企业和创客对于气象数据价值的进一步挖掘正慢慢带给他们更高的回报。分享与讨论方向会从天气切入: 1、海量天气数据与其他公共数据 2、大数据分析: 行业数据 + 天气数据 3、天气大数据应用 -保险 4、企业与初学者发展的数据的思考分享。

多维地理信息交互可视化与特征分区自动提取

李伯楠（深圳市位和科技有限责任公司）
时间：11:30~12:00 邮箱：*info@wayhe.com*

简介：位和科技市场总监，毕业于中科院地理所与宾夕法尼亚州立大学。

摘要：介绍多维地理信息交互可视化与特征分区自动提取技术，并讨论基于 Twitter 数据的美国区域语言变化分析、全球地表温度变化和异常模式分析、基于 30 年 24 小时降水极值的气候特征分析等案例及分析方法。

用 R Markdown 愉快地写作是怎样一种体验

谢益辉（*RStudio*）
时间：14:00~14:30　邮箱：*xie@yihui.name*

简介：统计之都创始人，爱荷华州立大学统计学博士，RStudio 软件工程师。

摘要：我第一次萌生写书的想法是 2007 年，那时候我对 LaTeX 兴趣盎然，毕竟才被折磨两三年而已。尽管捣鼓 CTEX、WinEdt 之类的东西很费神，看到整齐利落的 PDF 输出时，觉得一切都值了。过了一年，我开始捣鼓 Sweave，它让我可以把 R 代码直接嵌入 LaTeX 文档，而不必再复制粘贴结果，所以更新结果很方便。尽管捣鼓 Sweave 很费神，看到 R 代码直接输出图片到书中，觉得一切都值了。再过了一年，我发现了 LyX，原来我在反斜杠世界中浪费了那么多时间！于是我把 Sweave 移植到 LyX 中，虽然捣鼓 LyX 很费神，但看到这反斜杠再也遮不住我眼，觉得一切都值了。这样过了两年，我意识到 R 核心团队成员也是肉身凡体，并不是神，Sweave 虽然有一个很好的想法，但其实现太多局限性，于是我开始造一个新轮子，即 knitr。此时，十年来 Sweave 没能完成的功能支持在几个月内都通过 knitr 完成了。然而，Markdown 开始盛行了。LaTeX 用户中常见斯德哥尔摩综合症，幸好我不是患者，此时我开始想，为什么我要折腾那么多反斜杠的玩意。写作不就是章节段落加粗斜体图表引用这么点事吗，为什么会弄成 LaTeX 那样在各种命令中万劫不复。于是，knitr 开始支持 Markdown。但 Markdown 始终太简单太弱了，世人多瞧不上。此时，Pandoc 出场了，给我一个 Markdown 文档，我撬起地球给你看。于是真正的 R Markdown 出世了。然而 Pandoc 还是缺乏一些学术写作的功能。从我萌生写书的想法过了十年，我终于觉得我造出了适合写书的工具:bookdown。

基于 R 的数据分析平台

刘应耀（阿里巴巴）
时间：14:30~15:00　邮箱：*yingyao.lyy@alibaba-inc.com*

简介：阿里巴巴集团的阿外，高级技术专家，负责数据驱动业务创新以及智能机器人研发。

摘要：我的团队日常的算法原型、数据分析、个性化报告都是使用 R，为提高效率，还搭建了一个 R 的数据分析平台，用于实时产出个性化报告。整合了 rstudio server（数据分析、算法及 Rmd 报告开发)+ shiny server(交互式报告) + Grails(实时数据处理及报告展现）。

SupR：让 R 语言走向多线程并行计算

邱怡轩（普渡大学）
时间：15:00~15:30　邮箱：*yixuan.qiu@cos.name*

简介：普渡大学统计系博士生，统计之都理事会成员，感兴趣的方向包括复杂数据的统计推断，大规模统计计算和统计学习等。参与翻译了《ggplot2: 数据分析与图形艺术》、《R 语言编程艺术》、《应用预测模型》等书籍，是 RSpectra、showtext、recosystem 等 R 软件包的开发者。

摘要：R 语言是一个深受众多数据科学家喜爱的数据分析软件和平台，然而随着数据规模的增大，它的一些弊端也逐渐显露。例如，R 对并行计算和分布式数据存储的支持不太理想，这使得它在大规模数据的分析上

有所欠缺。即使目前有若干支持 R 并行计算的扩展包，但它们大都是基于进程级别的并行，其劣势是内存占用大，通信成本高。

对于数据分析平台，理想的并行模式是在单机上进行多线程的并行（例如 C/C++ 的 OpenMP 和 Java 的 Thread 类），集群上进行多机之间的通信，一个典型的例子就是目前非常流行的 Apache Spark。至今为止 R 语言的官方版本尚不能很好地支持多线程并行，原因是 R 的解释器和内存调度不是线程安全的。但考虑到 R 语言长久的社区支持和软件包积累，如果能让 R 语言实现这样的并行机制，将能极大地节省数据分析者的开发成本。

为了解决这一难题，普渡大学的刘传海教授基于 R 的官方版本试验开发了一款同时支持多线程和分布式计算的修改版 R——SupR。SupR 对官方 R 在源代码级别上进行了修改和补充，同时借鉴了 Spark 平台的诸多特性，实现了如下适合大规模数据分析的特性:

(1) 保持 R 的语法和内部数据结构不变

(2) 提供类似于 Java 的多线程计算

(3) 提供类似于 Spark 的集群运算

(4) 内置的分布式文件系统支持

SupR 目前处于开发阶段，正式版将以开源软件的形式发布。当前的开发信息可以在 http://www.stat.purdue.edu/ chuanhai/SupR/internal/index.html 上获取。我们希望 SupR 能吸引更多的开发者和使用者，将 R 语言真正带向大规模数据分析的世界。

大数据时代的柳叶刀——data.table 使用体验

尹志（宁波工程学院）
时间：16:00~16:30 邮箱：*ronaldo_yin@hotmail.com*

简介：尹志，浙江大学物理学博士，现就职于宁波工程学院理学院。水过机器学习论文，做过数据挖掘项目，打过数据科学比赛。研究方向集中在推荐系统、文本挖掘等机器学习领域，对解决各类数据科学相关的实际问题尤感兴趣。

摘要：“跑一下你的 Hadoop，我有一堆大数据需要你处理。”他递给我一个容量为 2G 的 U 盘，淡淡地说道。

在大数据的浪潮下，许多从业人员开始将 Hadoop，Spark 等工具奉为圭臬。然而，在各类数据挖掘场景下，我们更多面对的是 GB 量级的“中型数据”。撇开庞大臃肿的 Hadoop，快速拥抱数据、挖掘有效信息才是我们真正的目标。本报告将介绍 R 语言 data.table 包。该工具堪称大数据时代的柳叶刀 –快速、高效、自成一体。报告将围绕 data.table 的历史、特点、使用及技巧展开。

python 中的数据工具箱

肖凯（蚂蚁金服）
时间：16:30~17:00 邮箱：*xccds@yeah.net*

简介：一个爱折腾数据的人，《数据科学中的 R 语言》作者之一。

摘要：将介绍 python 中用于数据分析挖掘的包。例如 numpy，scipy，pandas，scikit-learn，theano 等等。会谈及使用这些包的应用场景，它们之间的联系，以及和 R 的区别。

Rust and R Integration

覃文锋（厦门大学）
时间：17:00~17:30 邮箱：*mail@qinwf.com*

简介：厦门大学公共卫生学院学生。

摘要：Rust is a systems programming language that runs blazingly fast, prevents segfaults, and guarantees thread safety.

The rustr Rust library and rustinr R package simplifies integrating Rust code with R. It provides Rust interfaces that map many types of R objects (vectors, functions, environments, ...) to Rust type. Object interchange between R and Rust is easy. Rust code can be compiled, and loaded on the fly, or added to a package.

大规模商品推荐系统——从原理到实践

熊熹（京东商城）
时间：14:00~14:30 邮箱：xiongxi@jd.com

简介：人大统计本科毕业，明大生统博士辍学，京东推荐一年晃悠，统计之都长期打杂。

摘要：以京东商城的推荐及其他个性化系统为例，以主流商品推荐系统算法和实践为脉络，介绍一下如何通过数据分析和算法切实提高用户体验和各项转化指标的一些应用。

认识另一种推荐系统: 兴趣 feed

陈开江（边聊边逛）
时间：14:30~15:00 邮箱：kaijiangchen@gmail.com

简介：陈开江 @ 刑无刀 2013 年之前在新浪微博搜索部和商业产品部任资深算法工程师，先后负责过微博反垃圾、基础数据挖掘、智能客服平台、个性化推荐等产品的后端算法。2012 年至 2013 年领导翻译了《机器学习: 实用案例解析》一书。2013 年末加入传统媒体公司车语传媒，任算法主管，负责从零打造公司转型产品考拉 FM 的个性化推荐系统，如今个性化推荐已成为考拉 FM 与其他 FM 之间最大差异化特性。2015 年初，离职创业，公司拿到 IDG 和晨兴资本的天使投资，产品几经调整，如今专注在用视频分享购物经验，App 名称: 边逛边聊。

摘要：Feed 即信息流，是社交网络中信息流动的总线系统，纵观国内外的社交平台，feed 无一例外都是粘住用户的核心。将内容按照时间顺序呈现给用户，一直是信息流的标准做法，且用户接受很快。但是近年来移动互联网和智能手机的发展，使得创造内容的门槛空前降低，以至于时间顺序排列的 feed 流开始出现明显的信息过载，据 Facebook 数据显示，80% 的新鲜事是不会被用户读到的，Instagram 也表示 70% 的内容被用户错过。因此，让用户看到最想看的内容就成为 feed 发展的必然趋势，即“兴趣 feed”替代“时间线 feed”。已经有一些大的社交平台完成了兴趣 feed 的转型，而且越来越多的社交平台开始向兴趣 feed 切换，这既是互联网产品发展的必然，也有商业利益的驱使，但是关键还是得益于大数据技术的发展。如何为你的社交产品打造一个个性化的兴趣 feed，将会成为新的大数据技术发展方向。

微博中的用户建模

冯扬（新浪微博）
时间：15:00~15:30 邮箱：fengyoung82@sina.com

简介：2010 年毕业于北京理工大学，获得信息科学博士学位；曾先后就职于新浪微博、腾讯 SOSO、搜狗等互联网公司，主要从事推荐算法研究及推荐系统的研发工作；2014 年 9 月加入新浪微博商业平台及产品部，任推荐技术专家，从事微博推荐广告策略平台的设计及研发。

摘要：如何认识和刻画每一个用户，是所有社交类型的平台都需要面对的问题。本次分享将给大家带来新浪微博中的用户建模，包括:

(1) 微博中的用户画像的构建

用户画像是对用户的信息进行标签化。一方面结构化的画像信息方便计算机的识别和处理；另一方面，标签本身也具有准确性和非二义性，利于人工的整理、分析和统计。分享内容中包含了微博中如何提取用户兴趣标签，并扩展到提取用户的能力标签，此外考虑标签的时效性，提取用户的长短期兴趣。

(2) 微博中的社交关系模型

不同于传统互联网媒体，微博作为社交媒体最大的优势在于引入了非对等的用户关系，形成了关系网络，从而仅令传播更加高效。在微博中构建用户关系模型，就是针对这种关系网络中的节点（代表了用户）、边（代表了关系和方向）、关系圈（代表了由关系聚合而成的群体）进行分析，全面地描述和刻画社交媒体中的关系网络。

(3) 微博中的群体模型

经过多年的经营，微博中形成了各式各样的群体，这些群体本身具有一定的共性，而群体之间又存在着差异，体现在群体的特征、受众（粉丝）、行为模式上；针对用户群体的建模能有效帮助平台以及独立账号进行成长。

除了上述三个方面以外，分享的末尾会针对微博用户模型在微博推荐、微博商业中的应用进行简要介绍。

基于大数据的广告营销

胡为松（百分点集团）
时间：16:00~16:30　邮箱：*weisong.hu@baifendian.com*

简介：百分点集团研发经理，北京邮电大学硕士，长期从事计算广告相关领域的工作和研究，现负责百分点营销管家的技术和研发。

摘要：互联网广告经过短短十几年的发展，从合约广告，到竞价售卖，最后到实时竞价，逐渐形成了以人群为投放目标的技术型投放模式。计算广告的核心问题，是利用数据和算法，为一系列用户与环境的组合，找到最合适的广告投放策略，以优化整体广告活动的利润。数据和算法在互联网广告领域起着越来越重要的作用。尤其对于实时竞价产品，用定制化标签指导广告投放是实时竞价的关键产品目标。本次演讲主要介绍计算广告的发展和现状，以及如何用标签和模型指导互联网广告营销。

推荐系统中数据稀疏与冷启动问题的研究

郭贵冰（东北大学）
时间：16:30~17:00　邮箱：*guogb@swc.neu.edu.cn*

简介：2015 年 11 月，以引进人才的方式加入东北大学软件学院，任职副教授。研究兴趣包括推荐系统，信任计算，社交网络分析，数据挖掘，现已发表 30 余篇国际学术会议和期刊文章。师从新加坡南洋理工大学 (NTU) 的 Assoc. Prof. Jie Zhang 和 Prof. Daniel Thalmann，于 2015 年 7 月获得博士学位；2015 年 1 月至 10 月，在新加坡管理大学 (SMU) 朱飞达教授的实验室担任研究员。

积极活跃于学术社区，策划举办第一届 IFUP 。2016 国际研讨会在 ACM UMAP 2016 上，受邀成为多个重要国际学术会议的程序委员会会员 (PC member)，包括 AAAI 2014、WI 2015、RecSys 2016、IJCAI 2016 等，受邀成为多个重要国际学术会议和期刊的审稿人，包括 AAAI 2015、WWW 2014-2015、RecSys 2014、TKDE、KAIS、KBS、ECRA、WIAS 等。此外，设计开发了开源 Java 推荐工具库 LibRec，吸引了推荐系统领域众多学术界和工业界人士的关注，在开源代码仓库 GitHub 的推荐系统领域中排名前列。

摘要：数据稀疏和冷启动是推荐系统面临的两个重要挑战。本演讲将从三个研究方向上讨论如何进一步解决这两个关键问题。第一个方向是仅基于评分数据；第二个方向是融合社交信任数据；第三个方面是融合上下文情景数据的推荐系统。每一个方向都将讨论两个我们的相关研究工作。此外，还将摘要介绍基于 Java 的推荐系统工具库 LibRec，以增强推荐算法的可重复性。

大数据驱动的智能招聘推荐系统

单艺（猎聘网）
时间：17:00~17:30 邮箱：*shanyi@liepin.com*

简介：单艺先生目前担任猎聘网首席数据官，负责机器学习技术和产品研发、商业数据分析以及大数据基础设施建设。他的主要兴趣在于数据挖掘和商业分析。他具有 16 年的数据挖掘和系统研发经验。之前，他担任 Omni-Dimension Inc 和 WPP Group/ 奥美 ITOP 24/7 Networks 的 CTO 职务，负责数据驱动的互联网广告优化技术和精准广告网络的研发；还曾经担任空中网悟空搜索副总裁和美国 Yahoo！网页搜索资深工程师，从事大规模搜索技术和文本挖掘技术的研发。

单艺毕业于清华大学和美国 University of Arizona，获得了管理信息系统专业的学士和硕士学位。

摘要：在传统的招聘过程中，求职者需要从成千上万的职位中寻找适合自己的公司和职位，而招聘经理则需要从海量简历中排除掉大批的不合格候选人。这对于双方来说，不仅费时费力，而且由于严重的信息不对称导致整个过程效果和效率不佳。

为了解决这个问题，猎聘大数据研究院运用了先进的 Hadoop、文本挖掘、协同推荐和机器学习等技术，开发了大数据驱动的智能招聘系统，机器伯乐。我们将在本次分享中介绍机器伯乐的总体设计、关键技术和开发心得。内容将会包括:

(1) C 端推荐总体设计
(2) B 端推荐总体设计
(3) 协同推荐算法应用
(4) 大规模矩阵分解算法应用
(5) 增强学习算法的应用

广聚人群，点通价值——腾讯广点通广告受众定向的探索与实践

靳志辉（腾讯）
时间：14:00~14:30 邮箱：*rickyjin@qq.com*

简介：北京大学计算机系计算语言所硕士，日本东京大学情报理工学院统计自然语言处理方向博士。2008 年加入腾讯，主要工作内容涉及统计自然语言处理和大规模并行机器学习工具的研发工作。目前为腾讯社交与效果广告部质量研发中心总监，主要负责腾讯用户数据挖掘、精准广告定向、广告语义特征挖掘、广告转化率预估等工作。

摘要：作为腾讯社交与效果广告的承载平台，广点通以“广聚人群，点通价值”为理念，汇聚腾讯内、外部超精品流量，致力于为广告主提供智能化的数字营销服务，提升内、外部流量的变现能力，在此过程中为用户传递最感兴趣的资讯信息。

广点通广告系统设计理念之一就是利用前沿的数据挖掘技术深度挖掘腾讯用户数据，实现精准广告定向。广点通技术团队，通过将数据挖掘技术和广告平台的整合，建设了效果广告的个性化定向体系。我们将向您介绍腾讯广告受众定向的挑战、发展，以及面向未来的思考。

基于学生就餐数据的社交网络挖掘及成绩预测

刘跃文（西安交通大学）
时间：14:30~15:00 邮箱：*liu.yuewen@qq.com*

简介：刘跃文是香港城市大学及中国科学技术大学博士、新加坡国立大学访问学者、西安交通大学博士后。曾任香港城市大学研究员、腾讯科技（深圳）有限公司高级工程师，现任西安交通大学管理学院信息管理与电子商务系助理教授。

刘跃文主讲课程包括新一代信息技术与管理、数据库系统及应用、数据挖掘与知识发现等。研究领域包括社交网络用户行为研究、创新扩散研究、电子商务研究等，其研究成果发表在 Decision Support Systems 等国际期刊上，并获得国家自然科学基金青年项目、中国博士后科学基金特别资助等多个项目资助，获批十余项国家专利及数项软件著作权。

摘要：大学生的学习成绩是影响其出国、深造、就业等未来发展的重要因素。部分大学生，由于不及格科目过多，甚至无法拿到毕业证及学位证。因此，管理大学生的学业，督促其学习，是大学教育工作的一个重要任务。目前，大学的学业管理模式主要是事后干预，即在成绩发布之后与学生交流其学习过程中存在的问题。本研究尝试使用学生的就餐数据来预测其学习成绩，目标是及早发现学业有潜在问题的学生，以达到事前发现、事前干预的目标。

Network Autoregressive Factor Model

兰伟（西南财经大学 & 柠檬科技）
时间：15:00~15:30 邮箱：*lanwei@swufe.edu.cn*

简介：高校青椒，闯荡消费金融。

摘要： This article introduces a network autoregressive factor model for an n-individual response vector for large scale network data. The network autoregressive model has been widely used to model the dependency of individuals in some social networks, e.g., the social network site (SNS). However, the dependency can also be generated by some common factors, for instance the overall interactive degree within the website.

The proposed method explores the dependent relationship between the n individuals from both the social network and common factors. We propose an EM algorithm to obtain the maximum likelihood estimators. Then, we show the large sample properties of the estimators corresponding to the network effect, factor loading and common factors. Simulation experiments are presented to demonstrate the finite sample performance. An example is analyzed from the Chinese Weibo to illustrate the usefulness of the proposed network autoregressive factor model.

“化繁为简”——复杂网络在 DT 时代的应用

毛仁歆（蚂蚁金服）
时间：16:00~16:30 邮箱：*maorenxin@gmail.com*

简介： 毕业于南京大学（学士）及香港中文大学（硕士），现就职于蚂蚁金服，主要研究兴趣为数据挖掘、机器学习以及复杂网络。求学及在职期间在国际 SCI 权威期刊发表研究型论文多篇，影响因子累计达 35.51，并同时发表专利十余篇。

摘要： 小世界、六度分割、无标度，这些名词汇聚在一起成为我们熟知的复杂网络。当我们数据很“小”的时候，繁杂的结构已经让我们无所适从。在 DT 时代，当数据量级达到百亿、千亿的时候，复杂网络的问题仍然困扰着我们: 小世界、无标度这些性质是否仍然存在? 拓扑空间是否仍然能够表达欧式空间的隐藏结构? 如此复杂的结构是否存在至简的方法来解决我们的实际问题?

这次分享我将为大家带来 DT 时代的复杂网络介绍，以及其在蚂蚁金服的风控、营销等场景的真实案例，备以分布式图计算框架介绍，脑洞即将大开。

Tweet Or Retweet? Interaction Utility Derived From User-generated Content In Social Media

周静（北京大学）
时间：16:30~17:00 邮箱：*jing.zhou@pku.edu.cn*

简介： 周静，北京大学光华管理学院博士研究生，研究方向为营销模型、空间模型、社会化网络营销，热衷于网络数据的建模与分析。

摘要： With the flourish explosion of online social networks (e.g. Facebook, Twitter or Weibo), user-generated content (UGC) is becoming a dominating way for people to communicate with each other on social platforms. For platform providers, a good UGC performance can attract more advertisers and directly influence their revenue. Therefore, how to encourage people to contribute content becomes a problem of interest. However, the underlying motivation on user generated content is still not well understood. Although previous literatures suggest the motivation can be driven by intrinsic utility and image-related utility, they ignore the interaction perspective of social media. Therefore, we propose in this paper a more comprehensive utility framework.

Combining with previous utility theory, we propose an interaction utility concept which is used to describe the utility that users derive from interacting with their friends. Then incorporating interaction utility is our first contribution in this paper. As our second contribution, we exert a time budget constraint in our utility framework and this enables us to evaluate the tradeoff between generating and consuming content. Finally, we differentiate the impacts from different social neighbors on users' tweet and retweet. Both analytical model and empirical approach are proposed in this paper.

社交风控的思路和实践

陈弢（蚂蚁金服）
时间：17:00~17:30　邮箱：*50538840@qq.com*

简介：陈弢，蚂蚁金服高级数据分析专家。2008 年毕业于香港科技大学，获计算机专业博士学位。主要研究方向为人工智能，数据挖掘和机器学习。在国际期刊和会议上发表近 30 篇学术论文，其中关于多维聚类的系统研究发表于《人工智能杂志》。担任 IJCAI、UAI、以及 PAKDD 的程序委员会成员。曾为国际一流期刊如 Artificial Intelligence Journal、Journal of Machine Learning Research 审稿。参与审稿的国际会议包括 UAI、IJCAI、AAAI、ICML、ECSQARU、PAKDD 和 PGM。

摘要：一个社交应用的成熟需要经过社交关系的沉淀、维系和活跃几个阶段。在不同的阶段，安全风控也面临着不同的风险。本次分享主要针对社交风险中典型的赌博社群、垃圾内容和垃圾关系等来阐述如何基于社交大数据从内容、行为和用户画像的维度进行立体的安全防控。同时，我们也脑暴下基于一个实人的、健康的社交环境可以衍生出的社交功能。

Revisiting some Basic Components in Deep Neural Networks

王乃岩（图森）
时间：14:00~14:30 邮箱：*winsty@gmail.com*

简介：王乃岩，现为北京图森互联科技有限公司首席科学家，负责算法研发业务。在这之前，他于 2015 年毕业于香港科技大学计算机科学与工程系。他的主要研究方向为计算机视觉与数据挖掘，特别在于将统计计算模型应用到这两者的实际问题中去。

摘要：Deep neural networks have significantly advanced the state-of-the-art performance in various area, such as computer vision and speech recognition. In this talk, I will review some basic components of deep neural networks, and provide some new insights into it. In the first part, I will focus on the meta-operations in DNN including convolution layer and fully connected layer. These layers are both based on the well-known generalized linear model. Though multiple layer stacking and non-linear activation functions could mitigate this issue, their representation power are still limited. Thus, we propose an efficient and universal factorized bilinear operation to enhance them. In the second part, I will focus on the batch normalization layer, which has become a de facto component in state-of-the-art DNN architecture to reduce gradient explosion and vanishing. Nevertheless, our observations strongly suggest that besides its original use, it has an interesting side effect: the domain information of data is stored in the statistics of batch normalization layer. Based on that, we propose to modify batch normalization layer for domain adaptation task. The proposed method is simple yet effective. It can archive state-of-the-art performance in various vision recognition tasks with only several lines of codes.

高效的随机矩阵计算

王树森（浙江大学）
时间：14:30~15:00 邮箱：*wss@zju.edu.cn*

简介：王树森于 2011 年在浙江大学计算机学院获得本科学位并攻读博士学位，将在今年毕业后前往加州大学伯克利分校统计系任博士后研究员。研究方向主要是机器学习、随机数值计算、优化，在机器学习顶级期刊 JMLR 发表 3 篇文章，博士期间获得过微软学者、百度奖学金等多项荣誉和奖励。

摘要：大数据给矩阵计算和机器学习带来了巨大的挑战。许多经典的矩阵计算和机器学习方法由于时间复杂度、空间复杂度过高，故不适用于大数据问题。近年来矩阵数据的快速近似方法受到了广泛关注。矩阵快速近似方法采取用精度换时间、空间的策略，损失少量精度从而大大降低时间复杂度和空间复杂度。矩阵快速近似方法已经广泛用于加速特征分解、奇异值分解、矩阵求逆等矩阵计算操作，也广泛用于最小二乘法、矩阵恢复、高斯过程、核 PCA、谱聚类、流形学习等机器学习方法。本次报告系统讲解随机矩阵计算的一些基础知识和典型应用:

(1) 随机投影和随机列选择
(2) 近似回归问题
(3) 近似奇异值分解
(4) 半正定矩阵近似及在核方法的应用

Unified Low-rank Matrix Estimation via Penalized Matrix Least Squares Approximation

钟琰（中国人民大学）
时间：15:00~15:30 邮箱：*yanzhong07@foxmail.com*

简介：中国人民大学统计学院研究生。

摘要：Low-rank matrix estimation naturally arises in a number of statistical and machine learning tasks. For example, the coefficient matrix has been considered to have a low-rank structure in multivariate linear regression or multivariate quantile regression. In this paper, we propose a method called Penalized Matrix Least Squares Approximation (PMLSA) for unified yet simple low-rank matrix estimation. Our general theoretical framework includes all the aforementioned regression models and many others as special cases. Specifically, PMLSA can transfer many different types of low-rank matrix estimation problem into their asymptotically equivalent least-squares forms. Then the equivalent forms can be efficiently solved by an algorithm called matrix FISTA and easily derived an analytic form of the degrees of freedom. We construct a BIC-type criterion using the derived degrees of freedom for selecting tuning parameters. Furthermore, we justify using the BIC-type criterion to select the estimated rank is asymptotically consistent with the true rank under mild conditions. Extensive experimental studies confirm our theory.

Scalable MCMC method for Bayesian Models

李文哲（普惠金融（爱钱进））
时间：16:00~16:30 邮箱：*nadalwz1115@hotmail.com*

简介：普惠金融（爱钱进）的首席数据科学家，负责公司的人工智能、大数据技术以及创新产品的研发。在大数据、机器学习、深度学习、自然语言处理，图数据库等领域有丰富的研究和实践经验。在美期间，先后就职于亚马逊、高盛、Fiserv 等多家公司。南开大学本科，美国 Texas AM 大学人工智能硕士，美国南加州大学机器学习博士，荷兰阿姆斯特丹大学访问学者，主要的研究方向为图模型、贝叶斯优化、深度学习、知识表示，先后发表数篇论文在 AAAI、KDD、AISTATS、CHI 等国际顶级会议和期刊上。

摘要：Scaling up bayesian models is becoming one of the most popular research topics in machine learning community. In this talk, I will introduce stochastic gradient Markov chain Monte Carlo (SG-MCMC) algorithm for scalable inference in Bayesian Models, in particular, mixed-membership stochastic blockmodels (MMSB). Our algorithm is based on the stochastic gradient Riemannian Langevin sampler and achieves both faster speed and higher accuracy at every iteration than the current state-of-the-art algorithm based on stochastic variational inference. In addition, I will introduce a new approximation scheme that can handle models that entertain a very large number of communities. Finally, I will show how our algorithm can scale up to Friendster network with millions of nodes and billions of edges.

Task-Specific and Interpretable Feature Learning

汪张扬（*UIUC*）
时间：16:30~17:00 邮箱：*masterwant@gmail.com*

简介： Zhangyang (Atlas) Wang is joining the Computer Science and Engineering (CSE) Department, at the Texas A&M University (TAMU), as an assistant professor. During 2012-2016, he has been a Ph.D. student in the Electrical and Computer Engineering (ECE) Department, at the University of Illinois at Urbana-Champaign (UIUC), working with Professor Thomas S. Huang. Prior to that, he obtained the B.E. degree at the University of Science and Technology of China (USTC), in 2012.

His principal research interest has been addressing machine learning and visual data analytics problems using advanced feature learning techniques, with a recent focus on deep neural network models and theories. He has published over 25 papers in top-tier journals (IEEE TIP/TCSVT/TGRS) and conferences (CVPR/AAAI/IJCAI/ACM MM/SDM/NIPS/BMVC, etc.), and recently coauthored the book "Sparse Coding and Its Applications in Computer Vision". He received the prestigious UIUC Dissertation Completion Fellowship (2016), Huang Graduate Research Award (2016), CSC Fellowship (2016), Baidu Research Scholarship (2015), Cognitive Science/Artificial Intelligence Research Award (2015), among many others. The Adobe DeepFont system, to which he was the leading contributor, has led to high-impact technical products and attracted lots of media coverage. He also completed three successful summer internships in Microsoft Research (2015), Adobe Research (2014), and US Army Research (2013).

摘要： Deep learning models have had tremendous impacts over the recent years, in a variety of machine learning and artificial intelligence applications. Meanwhile, a questions has been raised by many: is deep learning just a triumph of empiricism? There has been emerging interests in reducing the gap, between the theoretical soundness and interpretability, and the empirical success of deep models. In this talk, I will introduce my research on bridging traditional learning models that emphasize problem-specific reasoning, and deep models that allow for larger learning capacity. The overall goal is to devise the next-generation deep architectures that are: 1) Task-specific, namely, being optimized for the specific task by fully exploiting available prior knowledge and problem structures, rather than applying generic data-driven models as "black-boxs"; and 2) Interpretable, namely, being able to learn a representation which consists of disentangled and semantically sensible latent variables, and to display more predictable behaviors. I will present a few concrete model examples, to reveal how the analytic tools in the classical optimization problems can be translated to guide the architecture design and performance analysis of deep models. As a result, those models demonstrate improved performance, intuitive interpretation, as well as efficient parameter initialization. I will then show how my developed feature learning models are applied to constructing and solving complicated, end-to-end optimization models efficiently. Finally, I will conclude this talk by discussing several interesting future directions, with a focus on several newly-emerging interdisciplinary applications such as medical and healthcare data analytics.

What's the Insight of Self-paced Learning

孟德宇（西安交通大学）
时间：17:00~17:30 邮箱：*dymeng@mail.xjtu.edu.cn*

简介： 孟德宇，博士，西安交通大学数学与统计学院副教授，博导。曾于 2006 年赴英国 Essex 大学进行学术访问，于 2009 年赴香港中文大学进行博士后研究，2012-2014 年赴卡内基梅隆大学进行学术合作。在 TIP,

TKDE, TSMCB, TNNLS, PR, Neural Computation 等国际期刊与 CVPR, ICCV, ECCV, AAAI, ICML, NIPS, ACM MM 等计算机顶级会议发表论文多篇。担任 ICML, CVPR, ICCV, ACM MM, AAAI 等 CCF A 类会议程序委员会委员，2016 年 AAAI 会议高级程序委员会委员。2010 年获陕西省青年科技奖，陕西省优秀博士论文奖。目前主要聚焦于机器学习、数据挖掘、计算机视觉、多媒体分析等方面的研究。

摘要： Self-paced learning (SPL) is a recently proposed learning regime inspired by the learning process of humans and animals that gradually incorporates easy to more complex samples into training. While several easy SPL implementation strategies have been proposed, it is still short of a general paradigm for guiding the construction of rational SPL learning regimes targeting specific applications. To resolve this problem, we provide an axiom for insightfully formulating the underlying principles of self-paced learning. This axiomatic understanding not only involves the previous SPL learning schemes as its special cases, but also can be utilized to extend a series of new SPL implementation regimes based on certain application aims. In the recent two years, we have constructed several SPL realizations, including SPaR, SPLD, SPCL, SPMF, based on this axiom, and achieved the best performance in several known benchmark datasets, e.g., Web Query, Hollywood2, and Olympic Sports. Especially, this paradigm has been integrated into the system developed by CMU Informedia team, and achieved the leading performance in challenging semantic query (SQ)/000Ex tasks of the TRECVID MED/MER competition organized by NIST in 2014.

In this talk, I'll introduce some of our recent developments on the insightful understanding under SPL regime. We will use these results to explain the intrinsic reason why SPL can work in applications with highly noisy scenarios

Detection and Attribution of Changes in Climate Extremes

阎军（*University of Connecticut*）
时间：14:00~14:30 邮箱：*jun.yan@uconn.edu*

简介： Jun Yan is a Professor at the Department of Statistics, University of Connecticut. He received his B.Econ and M.Econ from Renmin University of China, 1993 and 1996, respectively, MA in Economics from the University of Miami, 1998, and Ph.D. in Statistics from the University of Wisconsin, 2003. He was an assistant professor at the University of Iowa during 2003-2007 before he moved to the University of Connecticut. His methodological research interests include survival analysis, clustered data analysis, spatial extremes, estimating functions, statistical computing, and applications in public health and environmental sciences. He is committed to making his statistical methods available via open source software and has authored and is actively maintaining a collection of R packages in the public domain.

摘要： To detect changes in climate extremes, no fully satisfactory analog of the widely used optimal fingerprinting method for mean climate states has been available. The state-of-the-art method incorporates the signals into the location parameters of generalized extreme value (GEV) distributions. Since the coefficient of the signal is shared by all the grid boxes while other parameters are grid box specific, the estimation was done with an independent likelihood profiled on the shared parameter. The profile method is computing intensive combined with a bootstrap procedure for inferences, which is prohibitive for multiple signals.Further, it discards spatial dependence which may lead to low efficiency in estimation and, hence, low power in detection. We propose a combined score equation (CSE) method that combines the score equations of the GEV model at each grid box such that an approximate correlation function of the scores is used to improve the estimation efficiency of the signal effect. Under working independence, it reduces to the existing method, but the estimation is much faster with a coordinate descent algorithm. Unlike the pairwise likelihood (PL) method assuming max-stable processes the CSE method does not need full specification of spatial dependence. It provides a close analog to the optimal fingerprinting in detection and attribution of changes in climate extremes. The method is applied to extreme temperature in Australia under a perfect model setting and in Northern Europe with real data. In the latter application, anthropogenic impact is detected when the natural forcing is present in the model simultaneously, which has not been reported before due to methodological limitations. The favorable properties of the CSE method are further demonstrated in extensive simulation studies.

时空大数据支持下的空间交互研究及应用

刘瑜（北京大学）
时间：14:30~15:00 邮箱：*liuyu@urban.pku.edu.cn*

简介： 男，1971 年 10 月生，山东诸城人，北京大学遥感与地理信息系统研究所教授。目前主要研究方向为地理信息科学，参与或负责科研项目十余项，发表论文 100 余篇，其中 SCI/SSCI 收录 50 余篇。任 Computers, Environment and Urban Systems 副主编、Journal of Spatial Information Sciences、地理与地理信息科学编委。社会学术任职有中国地理学会青年工作委员会副主任、中国地理学会地图学与地理信息系统专业委员会委员、中国 GIS 协会理论与方法工作委员会委员等。

摘要： 在地理学研究中，空间交互 (Spatial interaction) 指的是两个场所之间的联系。这种联系通常可以基于人流、货流、资金流等量化。研究空间交互有助于理解一个区域内部的结构以及动态演化特征。空间大数

据中，个体的移动轨迹以及个体之间的社交关系，都可以在聚集层面量化两个场所之间的交互强度，前者如两个城市间的人流总量，后者如两个城市之间的互粉好友对数。本报告将针对城市间和城市内两个尺度，介绍时空大数据支持下的空间交互分析方法，以及其在相关领域的应用。

生物集群行为的系统研究

韩战钢（北师大）
时间：15:00~15:30 邮箱：*zhan@bnu.edu.cn*

简介：韩战钢，北京师范大学系统科学学院教授，副院长，系统分析与集成实验室主任，联合国教科文组织姊妹大学复杂系统数字校园副主席兼亚洲区主席。他拥有理论物理硕士学位和人工智能博士学位。他曾经在比利时布鲁塞尔自由大学 (ULB) Solvay 研究所，美国加州大学洛杉矶分校 (UCLA) 访问，进行复杂系统研究。曾主持和参加国家自然科学基金，863，国家科技部攻关等多项国家级科研项目，以及欧盟 FP7 项目和澳中合作项目。在国际和国内期刊发表多篇文章，并多次在国际会议上报告研究成果。他现在的科学研究集中于通过实验和模型方法研究蚁群、鱼群和机器人群体的集群行为（collective behavior）。以蚁群和鱼群为科学观测对象，通过机器人群体实现仿生控制。研究包括几类群体的集群行为中信息获取、传播和对集群行为的影响，对称破缺（symmetry breaking）的出现与机制分析，系统处于临界态的实验设计、观测与机制分析，系统追逃行为研究，机器人系统自组织行为实现，机器人系统处于临界态的工程实现。

摘要：生物群体与活性物质的集群行为（collective behavior）研究正成为国际科学研究的新热点。集群运动作为集群行为研究分支广泛存在于自然界生物系统中，从菌落的漂游，蚁群、蜂群的分工协作，蝗虫的集群侵袭，鱼群的集群游动，鸟群的迁移及躲避敌害，都表现出典型的自组织集群行为。本报告主要从蚁群、鱼群、鸟群及群体机器人控制来介绍集群行为研究的国际前沿进展，从复杂性研究的角度，探究众多系统通过简单的局部相互作用而涌现出复杂集群行为的普适机制，揭示系统中个体信息获取、传播和利用的规律，介绍生物与机器系统的融合，以及对称破缺（symmetry breaking）的出现与机制分析，系统处于临界态的观测与机制分析，系统追逃行为等研究。

大规模时空数据分析与可视化:R 应用与实践

王江浩（中国科学院地理科学与资源研究所）
时间：16:00~16:30 邮箱：*wangjh@lreis.ac.cn*

简介：王江浩，博士，中国科学院地理科学与资源研究所，资源与环境信息系统国家重点实验室助理研究员，北京城市实验室主要创始人之一。主要研究方向包括空间统计、时空数据挖掘、资源环境遥感与应用。近年来主要从事时空大数据分析和制图的新理论和新技术方法研究。现已积累了海量的手机定位数据和社交网络数据，建立了面向专题的时空数据库；在此基础上开展人群时空动态分布特征分析与制图，通过与普查数据对比，证明了地理微博数据在人口分布与流动研究的可行性；利用地理微博数据构建城际人口流动流入流出矩阵，开展人口流动的多尺度空间特征分析、挖掘与制图，并分析了影响人群流动的自然、社会、经济和文化因素。现已在 IJGIS、IEEE TGRS、Annals of AAG 等学术期刊上发表 SCI/SSCI 论文 20 余篇。

摘要：以动态、异构、多源为特征的时空大数据已经成为我们观察人类自身社会行为的“显微镜”和监测大自然的“仪表盘”。这些数据包括手机信令数据、移动位置服务数据、社交网络签到数据、兴趣点数据等。时空大数据可以构建全方位、跨领域、多角度、高时效性的全息立体信息网络，已深入应用到包括交通管理、环境治理、城市规划、城市计算、社会计算、智慧城市建设等领域。面对新型的时空大数据，如何进行定量化

分析、挖掘与制图表达是空间数据挖掘的重要挑战。本报告将重点介绍 R 语言下，开展大规模时空数据分析与可视化，介绍所采用的大数据计算策略，如分布式数据存储与计算、网格化数据处理、高性能计算算法构建等内容。其目的是用 R 语言对海量异构的时空数据进行深度分析与表达。

基于迁徙数据的中国次区域识别

李栋（北京清华同衡规划设计研究院有限公司）
时间：16:30~17:00 邮箱：*LDZGY@qq.com*

简介：李栋，博士，高级工程师，北京清华同衡规划设计研究院技术创新中心副主任，北京城市实验室（BCL）主要创始人之一。工作以来参与和负责多项国家和地区重大规划项目、国际学术课题研究等工作，多次获省部级以上奖励。近年来致力于在城市研究和规划中应用大数据解决现实问题，开展算法、模型与指标的研究，重点关注出租车轨迹、互联网签到、照片等新型数据的获取、处理、以及相关的时空分析和挖掘等。研究成果在中英文同行评议期刊、重要会议中多次发表。

摘要：基于互联网公开的互联网人口迁徙数据，应用网络分析方法和指标，对中国次区域进行了重新划分，并测算了其联络度、重要性等评价指标。本研究对中国的区域和城市研究与规划具有指导意义。

Application of R in Study on Deep-sea Polymetallic Nodules from the Pacific Ocean and Indian Ocean

周鹏（国家海洋局第二海洋研究所）
时间：17:00~17:30 邮箱：*hockeyextremophiles@yahoo.com*

简介：Peng Zhou, received B.Sc. in Biological Sciences from Shandong University and Ph.D. in Genetics from Zhejiang University. He has research experiences on microbiology in Lawrence Berkeley National Laboratory, University of California, and University of Massachusetts. Currently, he is working on marine biology at the Second Institute of Oceanography, SOA. His recent research focuses on data mining and visualization of the high-throughput sequencing data in marine ecology, comparative genomics and metagenomics.

摘要：Deep-sea polymetallic nodules are valuable for high abundance of metals, such as manganese, nickel, cobalt and copper. Microorganisms were known involved in the formation of nodules. To better understand what microorganisms may be involved in the formation of nodules, we investigated samples of 20 deep-sea polymetallic nodules and 9 surrounding sediments from the Pacific Ocean and Indian Ocean. The statistical analysis and data visualization was performed in R environment. The element analysis showed that manganese, cobalt, copper, and nickel are much more abundant in the nodules than those in the surrounding sediments. Copper and nickel have significant high correlation of 0.96 (p-value<0.001). Correlation-based network analysis was based on metal composition and relative abundance of taxa in the microbial communities, which helped obtain a comprehensive understanding of relationships between metals and microorganisms. The order Rhizobiales, to which some Mn(II)-oxidizing microorganisms belong, shows high correlation (r>0.6, p-value<0.001) with manganese, nickel and copper. The order Solirubrobacterales shows high correlation (r>0.6, p-value<0.001) with manganese and cobalt. The results suggest that the microorganisms belonging to these two orders may play important roles in the geochemical formation of nodules. This study shed light on further studies concerning the formation of deep-sea polymetallic nodules.

白金赞助

考拉征信

Quant Tech
量邦科技

金牌赞助

tableau

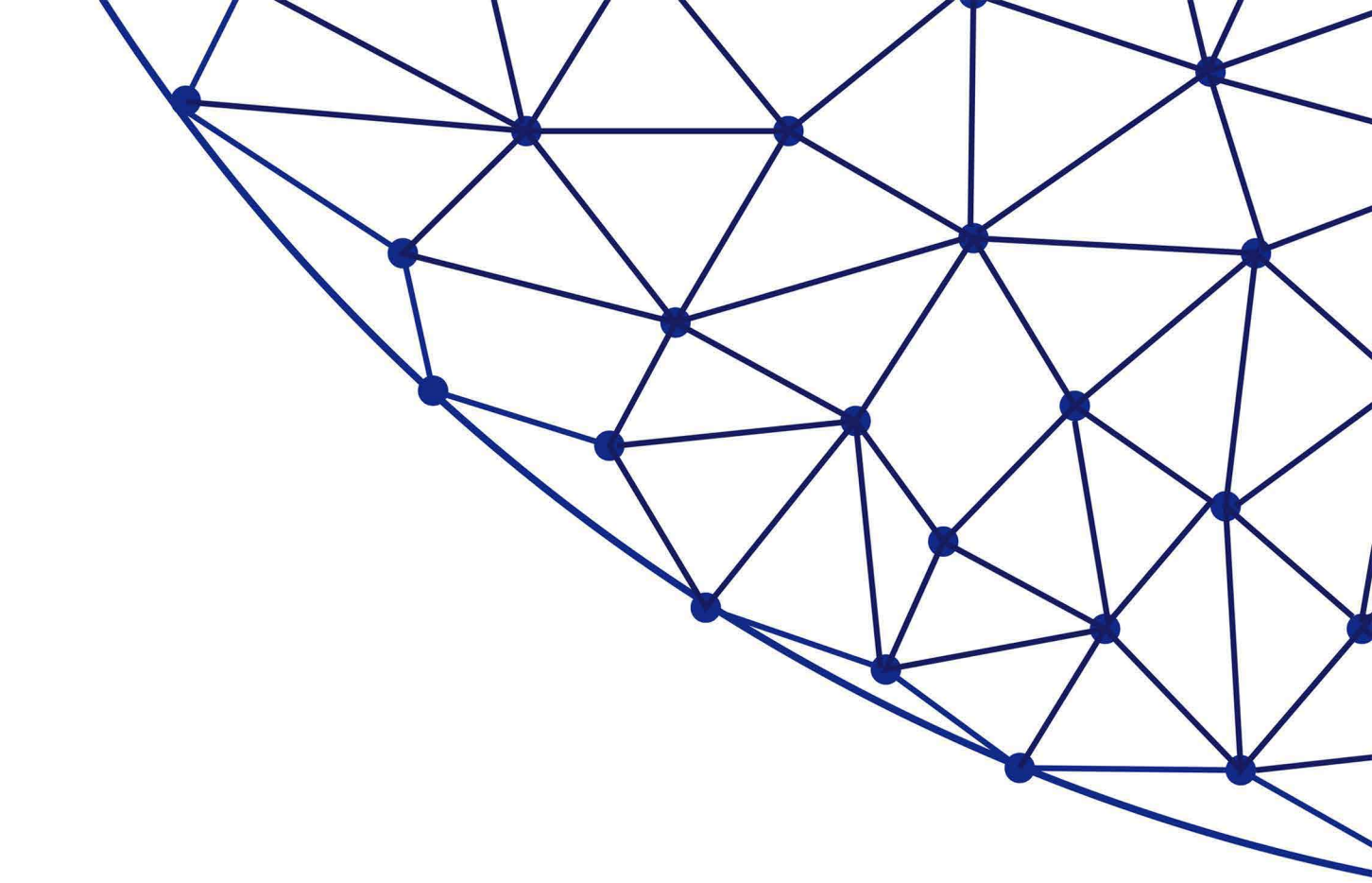
Microsoft

懒投资
LANTOUZI.COM

RStudio

银牌赞助

NYC DATA SCIENCE ACADEMY



专业 人本 正直

百分点

狗熊会

人大统计学院

统计之都