会 议 及 论 坛 议 程

10 月 24 日主会场

地点：江西财经大学学术报告厅

时　间	议　　程	主 持 人
8:30 – 9:00	开 幕 式	严　武
9:00 – 9:45	Actuarial Science with R Terry O'Neill （邦德大学 精算学院院长，金融大数据分析中心主任）	朱雪宁
9:45 – 10:15	茶　歇	
10:15 – 11:00	Forecasting Big Time Series Data Using R Rob Hyndman （莫纳什大学统计学教授，商务与经济预测中心主任）	朱雪宁
11:00 – 11:45	A Dynamic Approach to Sparse Recovery 姚远（北京大学教授 博士生导师）	
12:00 – 14:00	午餐及午休	
14:00 – 14:45	High-Frequency Analysis of Lead-Lag Relations in Exchange Traded Volatility Market using R Michael O'Neill （Investors Mutual Ltd 分析师）	任万凤
14:45 – 15:30	The Effect of Objective Formulation On Retirement Decision Making Garry Khemka（邦德大学 副教授）	任万凤
15:30 – 16:00	茶　歇	
16:00 – 16:45	R 与社会网络分析 李舰（堡力山集团 副总裁）	任万凤

10 月 25 日金融大数据专场 （懒投资冠名）

地点：江西财经大学学术报告厅

时 间	演 讲 嘉 宾	题 目
8:30 – 9:00	李丰（中央财经大学 讲师，院长助理）	Efficient High-Dimensional Dynamic Tail-Dependence Modeling with Copulas
9:00 – 9:30	邓一硕（懒投资 CFO，副总裁）	互联网金融产品创新及经营活动中的挑战
9:30 – 10:00	谢军 （雅捷 首席科学家）	大规模并行数据库& R 技术
10:00 – 10:30	休息与讨论	
10:30 – 11:00	史彦琳（澳大利亚国立大学 讲师）	A Discussion on the Innovation Distribution of Markov Regime-Switching GARCH Model
11:00 – 11:30	任坤（凌云至善 量化私募基金研发合伙人）	R 语言在量化交易中的应用
12:00 – 14:00	午餐及午休	
14:00 – 14:30	夏家莉（江西财经大学 教授 博士生导师）	面向网络的财政数据情感分析
14:30 – 15:00	范青亮（厦门大学 助理教授）	Big Data: The Applications and Challenges in Economics
15:30 – 15:45	休息与讨论	
15: 45 – 16:15	高涛（阿里巴巴 工程师）	时间预测的基本方法简介
16:15– 16:45	杨炳铎（江西财经大学 讲师）	High Dimensional Models in Finance with R

10 月 25 日 应用与可视化专场（诸葛 IO 冠名）

地点：财税大楼二楼报告厅

时　间	演　讲　嘉　宾	题　　目
8:30 – 9:00	张源源 （乐动力 工程师）	传感器数据中的机器学习实践
9:00 – 9:30	肖凯（一号店 工程师）	当 R 遇到 Python
9:30 – 10:00	任万凤（诸葛 IO 数据科学家）	APP 用户行为路径挖掘助力产品优化
10:00 – 10:30	休息与讨论	
10: 30– 11:00	郎大为（SupStat 数据科学家）	重新定义你的地图：REmap
11:00 – 11:30	周扬（J. D. Power 数据分析师）	数据可视化爱好者的工具箱
12:00 – 14:00	午餐及午休	
14:00 – 14:30	李悦（7Park Data 数据分析师）	R 语言存活分析在商业中的应用
14:30 – 15:00	郝智恒（阿里巴巴 数据挖掘工程师）	R 语言在业界数据分析的应用
14:30 – 15:00	杨环（Mango Solutions 数据分析师）	Shiny，可视化玩具？

10 月 25 日（上午）视频专场

地点：财税大楼四楼报告厅

时　间	演　讲　嘉　宾	题　　目
8:30 – 9:15	Wush Chi-Hsuan Wu（台湾大学 博士）	Introduction to FeatureHashing
9:15 – 10:00	尤晓斌 （新加坡国立医疗集团数据分析师）	用数据科学优化人口健康模式
10:00 – 10:15	休息与讨论	
10: 15– 11:00	邱怡轩 （普渡大学 博士）	（统计模型 + 最优化）×（ADMM 算法 + 并行计算）= ？
11:00 – 11:45	谢益辉 （RStudio 工程师）	论 R 码农的自我修养

10 月 25 日（下午）统计与机器学习专场

地点：财税大楼四楼报告厅

时 间	演 讲 嘉 宾	题 目
14:00 – 14:30	刘路 （中南大学 研究员）	Sparse Graph Coloring Processes A Weak Convergence Result
14:30 – 15:00	朱雪宁 （北京大学 博士）	Network Vector Autoregression
15:00 – 15 :30	罗立辉（中科院寒旱所 博士，副研究员）	拆分抑或耦合：地学集成建模
15:30 – 15:45	休息与讨论	
15: 45– 16:15	曾锦山 （江西师范大学 助理教授）	非凸阈值迭代算法的收敛性及其在 SAR 成像中的应用
16:15– 16:45	汤耀华（香港大学 博士）	Learning to Trade via Recurrent Neural Network and Direct Reinforcement
16:45 – 17:15	曾若辰（香港大学 博士）	Quantile Hysteretic Autoregressive Models

Keynotes

Actuarial Science with R

【嘉宾介绍】

Terry O'Neill is Director of Centre for Actuarial and Financial Big Data Analytics at Bond University. He is also head of Actuarial Science. Terry has a PhD from Stanford University. He is a Fellow of the Institute of Mathematical Statistics, Fellow of the American Statistical Association and elected member of the International Statistical Institute. He was previously Director of the Research School of Finance, Actuarial Science and Applied Statistics at the Australian National University.

【报告摘要】

R is a well documented programming environment with a strong academic following in many countries. It started with a core user group of statisticians, but now has a strong user base spread across a diverse range of disciplines. The open source nature of the product and the fact that it is used across disciplines including statistics, demography, economics and finance make it an ideal product for actuaries. It has already been embraced by many of the top actuarial industry bodies, research and teaching institutions. This talk will focus on the use of R at the forefront of data science in the field of Actuarial Science.

Forecasting Big Time Series Data using R

【嘉宾介绍】

Rob J Hyndman is Professor of Statistics in the Department of Econometrics and Business Statistics at Monash University and Director of the Monash University Business & Economic Forecasting Unit. Since 2005, he has been Editor-in-Chief of the International Journal of Forecasting and a Director of the International Institute of Forecasters.

He completed a Science honours degree at the University of Melbourne in 1988 and a PhD on nonlinear time series modelling at the same university in 1992. After lecturing at his alma mater for two years, he moved to Monash University as a lecturer in 1995 where he has been ever since. He was promoted to a personal chair in 2003.

Rob is the author of over 100 research papers in statistical science. In 2007, he received the Moran medal from the Australian Academy of Science for his contributions to statistical research, especially in the area of statistical forecasting. He is best known for his research on exponential smoothing, hierarchical forecasting, demographic forecasting and nonparametric smoothing. He has an h-index of 42 with 20 publications receiving over 100 citations each.

Rob is the coauthor of "Forecasting: methods and applications" (Wiley, 1998) with Makridakis and Wheelwright, and more recently of a free online textbook with George Athanasopoulos. He is also the founder of OTexts --- an online textbook publishing platform, and Cross-validated --- an online question and answer service in statistics and machine learning.

For 30 years, Rob has maintained an active consulting practice, assisting hundreds of companies and organizations. His recent consulting work has involved forecasting electricity demand, tourism demand, the Australian government health budget and case volume at a US call centre.

【报告摘要】

It is becoming increasingly common for organizations to regularly forecast many thousands or even millions of time series. For example, manufacturing companies often require weekly forecasts of demand for thousands of products at dozens of locations in order to plan distribution and maintain suitable inventory stocks. I will describe the best available algorithms for automatically forecasting large collections of univariate time series. These are implemented in the forecast package for R.

In many applications, there are also aggregation constraints that must be imposed. For example, a manufacturing company can disaggregate total demand for their products by country of sale, retail outlet, product type, package size, and so on. As a result, there can be millions of individual time series to forecast at the most disaggregated level, plus additional series to forecast at higher levels of aggregation. The disaggregated forecasts should add up to the forecasts of the aggregated data; this is known as "forecast reconciliation".

The optimal reconciliation method involves fitting an ill-conditioned linear regression model that is impossible to estimate using standard regression methods. I will demonstrate how this problem has been solved in R, making it possible for large scale forecasting to be implemented in practice.

A Dynamic Approach to Sparse Recovery

【嘉宾介绍】

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

【报告摘要】

In this talk we aim to solve an open problem raised by Jianqing Fan et al. in 2001, where convex l1-regularization (LASSO) causes bias in linear regression and nonconvex regularization is thus introduced for debias which however suffers from NP-hardness in finding global optimizers. We show a novel approach utilizing a technique from dynamics. Instead of optimizing a potential (objective) function, we evolve ODEs formed by gradient descent in dual space of l_1-norm. Equipped with early stopping regularization, it simultaneously achieves variable selection consistency and unbiased estimator, which is thus better than LASSO estimator. The dynamics leads to a simple discretization as linearized Bregman iteration algorithm, which has been widely used in image processing, matrix completion, as well as robust ranking, etc. R and Matlab packages are being developed in applications.

High Frequency Analysis of Lead-Lag Relations in Exchange Traded Volatility Markets using R

【嘉宾介绍】

Michael O'Neill BActS Hons/LLB, PhD, FIAA: Michael was awarded a PhD in Finance at the University of Queensland, Australia, in 2014. Previously he completed a combined degree at the Australian National University in 2005, graduating with an Actuarial Studies with First Class Honours and a University Medal, and a Bachelor of Laws. Michael qualified as a Fellow of the Actuaries Institute in 2007, and has been a director on the Institute's Board since 2008. Michael's research interests lie in volatility modelling and forecasting, state-price density analysis and analysis of supply and demand for volatility. Michael has worked as an equities analyst and portfolio manager at Investors Mutual Ltd since 2008. The company has received several awards as Australian fund manager of the year, most recently by Morningstar in 2015.

【报告摘要】

Exchange traded volatility markets have grown substantially since VIX futures were launched by the CBOE in 2004, followed by VIX options in 2006 and VIX ETPs in 2009. Daily open interest for volatility products is now in the tens of billions of dollars. Motivated by the wide range of options to trade stock market volatility on the exchange, Bollen, O'Neill and Whaley (2015) analyse the supply and demand for volatility. The aim of the study is to use high frequency data to determine where price discovery occurs, and whether the lead/lag relations between prices of exchange traded volatility have changed over time. Data are recorded at high frequency, and trading intensity has increased in these markets. In the analysis of lead/lag relations, the study calculates the Hiyashi

and Yoshida (2005) covariance matrix, in order to avoid any distortions and spurious correlations that might otherwise be caused by non-synchronous trading at high frequency. The calculation of this covariance matrix in R is computationally intensive; the estimator makes use of all data, regardless of the time intervals between samples. From this covariance matrix the authors draw conclusions for price discovery in markets for volatility and the length of leads/lags between markets, particularly where cost effective arbitrage is not possible.

The Effect of Obiective Formulation on Retirement Decision Making

【嘉宾介绍】

Dr. Garry Khemka did a Bachelor of Economics (Hons.) from Jadavpur University (Kolkata) before moving to Australia to pursue his education in Actuarial Studies. He completed a Masters of Actuarial Studies (with distinction) at the Australian National University and stayed on to complete his PhD in 2013. He is also a Fellow of the Institute of Actuaries of Australia. He was a lecturer at the Australian National University before joining Bond University in 2015 as one of the founding members of the Actuarial Science department.

Dr. Khemka is a budding early career academic with publications in some of the top Actuarial journals, along with a wide number of conference presentations and keynote speeches. His current research interests span retirement planning and superannuation, long term care insurance products, disability income insurance and personal finance. He has also received various research grants, as a testament of his growing skill and expertise.

【报告摘要】

For a retiree who must maintain both investment and longevity risks, we consider the impact on decision making of focusing on an objective relating to the terminal wealth at retirement, instead of a more correct objective relating to a retirement income. Both a shortfall and a utility objective are considered; we argue that shortfall objectives may be inappropriate due to distortion in results with non-monotonically correlated economic factors. The modelling undertaken uses a dynamic programming approach in conjunction with Monte-Carlo simulations of future experience of an individual to make optimal choices. We find that the type of objective targeted can have a significant impact on the optimal choices made, with optimal equity allocations being up to 30% higher and contribution amounts also being significantly higher under a retirement income objective as compared to a terminal wealth objective. The result of these differences can have a significant impact on retirement outcomes.

R 与社交网络分析

【嘉宾介绍】

李舰，毕业于中国人民大学统计学院（本科）和北京大学软件与微电子学院（研究生），现就职于堡力山集团，担任副总。是 Rweibo、Rwordseg、tmcn 等 R 包的作者，《数据科学中的 R 语言》的作者，还参与翻译了《R 语言核心技术手册》和《机器学习与 R 语言》。

个人主页：http://jianl.org。

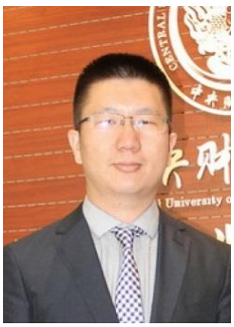
【报告摘要】

在大数据时代里，人们处理的数据早已超出了传统的数值和文本，尤其在移动互联网的时代下，社交网络数据成了非常关键的数据来源。社交网络分析（Social Network Analysis，SNA）是在传统的图与网络的理论之上对社交网络数据进行分析的方法，如今已经成了大数据分析不可或缺的一部分。本次报告将会介绍 SNA 的相关知识以及 R 中的实现方式，并结合业界常用的 Gephi 软件进行图形化的操作。此外，还会通过案例分享来说明社交网络分析的方法在业界的应用。

『金融大数据专场』

Efficient High-Dimensional Dynamic Tail-Dependence Modeling with Copulas

【嘉宾介绍】



李丰，中央财经大学统计与数学学院，讲师，硕士研究生导师，院长助理。

教育背景：

2008 — 2013 统计学博士，瑞典斯德哥尔摩大学统计学系

2003 — 2007 统计学本科，中国人民大学统计学院

E-mail：feng.li@cufe.edu.cn

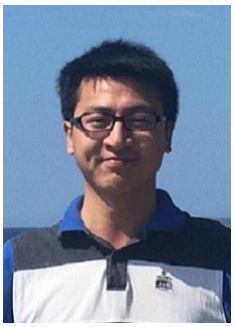
Web：http://feng.li

【报告摘要】

Copula tail-dependence modeling with flexible marginal distributions is widely used in financial time series. Most of the available Copula approaches for estimating tail-dependence are restricted within certain types of bivariate copulas due to computational complexity. We propose a general Bayesian approach for jointly modeling high-dimensional tail-dependence with general copulas functionals. Our method allows for variable selection among the covariates in the copula tail-dependence parameters. We apply a novel sampling technique into the posterior inference where the likelihood function is estimated from a random subset of the data, resulting insubstantially fewer density MCMC evaluations.

互联网金融产品创新及经营活动中的挑战

【嘉宾介绍】



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。

【报告摘要】

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

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

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

所有这些都需要根据数据和模型进行解决。

大规模并行数据库&R 技术

【嘉宾介绍】

28 年数据科学研究和工业实践经历，90 年代中期之前主要从事分析研究。1996 年后开始数据挖掘工业实践。中国移动、电信数据仓库第一案例实践者（广东邮电山东邮电）；2000 年创建中国银行数据仓库第一案例(工行浙江分行信贷分析系统)；创建中国教育数据挖掘第一案例(浦东教育质量评估)。此后完成江苏农行、上海农行等数据仓库工程和咨询项目 40 余例，典型工作：工总行法人客户贡献模型和风险转移模型。某些项目的水平，历 10 余年至今无人超出（某大型银行行长评价）

2005 年-2013 年重点研究教育数据挖掘，和其他教育专家一起在上海浦东和上海闵行进行了基于大规模考试的教育数据挖掘工程实践，历经 8 年。在 IRT 建模，教育数据挖掘和教育质量分析以及学生学业质量成因以及学生学科素养潜能分析上做出了领先的工作。

虽然离开学术界后奖项不再是努力方向，但还是获得人总行科技进步二等奖（项目奖）；农总行科技进步二等奖（个人奖），银监会科技进步奖（项目奖）

从业履历

2013.05 雅捷信息技术服务（上海）有限公司 副总经理 首席数据科学家

2006-2013Laurelway AB Stockholm 合伙人 兼技术总监

2004.05-2006.06 IBM 咨询业务部大中华区 CRM 签约首席咨询师

2002-2005 斯凯文软件技术（广东）有限公司 执行副总裁兼北京研发中心总监

1998.07-2001 创智软件 电信 BI 事业部总经理，金融业务并入 BI 事业部，任总经理；兼电子商务部总经理。

1997.07-1999 泰克软件项目经理、软件部副总经理

1993-1997 Daybreak Nuclear and Medical System, USA, 软件开发工程师

1995-1997 中科院西安分院 实习研究员

1991-1993 中科院西安分院 助理研究员

教育背景

1981-1985　上海复旦大学物理系 激光物理专业 学士

1987-1991　牛津大学 应用统计 博士

【报告摘要】

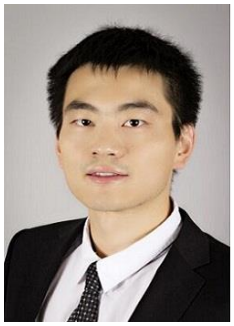
R 可以应用到金融，并可应用于金融大数据落地。金融大数据是一个非常大的领域。典型的大型省级银行拥有 5000 万客户，9000 万账户，100 个产品，10 个渠道，100 项客户属性。

客户行为描述往往是 5000 万行 5000 列的巨大表。这对所有现有的技术都是挑战。我们在这里报告一个 5700 万客户的银行是如何应用 R 于实际工作的。

本文报告一个大型银行使用 R，Hadoop 和 GPU 以获取 500 倍分析提速，是如何帮助营销定位客户的，是如何构建在一个架构中又如何发挥作用的。

A Discussion on the Innovation Distribution of Markov Regime-Switching GARCH Model

【嘉宾介绍】



史彦琳，于 2014 年 6 月在澳大利亚国立大学(ANU)取得统计学博士学位，目前在该校金融、精算与统计学院担任统计学讲师。研究方向包括时间序列分析，金融计量经济学和应用统计。曾在多个 SSCI 期刊上发表论文，并多次在国际学术会议如“International Congress on Modelling and Simulation”和“China Meeting of Econometric Society”上做过宣讲。

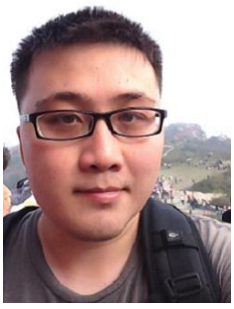
【报告摘要】

Markov Regime-Switching Generalized autoregressive conditional heteroskedastic (MRS-GARCH) model is a widely used approach to model the financial volatility with potential structural breaks. The original innovation of the MRS-GARCH model is assumed to follow the Normal distribution, which cannot accommodate fat-tailed properties commonly existing in financial time series. Many existing studies point out that this problem can lead to inconsistent estimates. To overcome it, the Student's t-distribution and General Error

Distribution (GED) are the two most popular alternatives. However, a recent study points out that the Student's t-distribution lacks stability. Instead, it incorporates the alpha-stable distribution in the GARCH-type model. The issue of the alpha-stable distribution is that its second moment does not exist. To solve this problem, the tempered stable distribution, which retains most characteristics of the alpha-stable distribution and has defined moments, is a natural candidate. In this paper, we conduct a series of simulation studies to demonstrate that MRS-GARCH model with tempered stable distribution consistently outperform that with Student's t-distribution and GED. The computational details of estimating the density function of the tempered stable distribution is also discussed. Further, our empirical study on the S&P 500 daily return volatility generates robust results. Therefore, we argue that the tempered stable distribution could be a widely useful tool for modelling the financial volatility in general contexts with a MRS-GARCH-type specification.

R 语言在量化交易中的应用

【嘉宾介绍】



任坤，毕业于厦门大学金融系和王亚南经济研究院，毕业后在深圳从事量化私募基金的策略研发和工具开发，是 pipeR、rlist、formattable 等扩展包的作者，在个人博客（http://renkun.me）中写了数十篇文章讨论数据分析相关工具、R 语言高级编程等主题。

【报告摘要】

量化交易的核心是策略，而好的策略离不开数据分析、交易执行和风险控制。R 语言作为数据分析、建模、可视化的重要工具，加上丰富的扩展包资源和快速成长的社区，可以显著提升效率，在量化研究方面起到重要作用。本报告将以量化策略的框架为基础，以具体案例为切入点，讨论量化策略从思想到回测、实盘以及分析环节中的方法和问题。

面向网络的财政数据情感分析

【嘉宾介绍】

夏家莉，教授、工学博士，博士生导师，江西省高校中青年学科带头人，江西财经大学财政大数据分析中心主任。

【报告摘要】

对网络上的相关财政信息进行采集、挖掘，进行情感分析，发现具有重大社会、政治、经济和民生影响的财经网络舆情热点话题，分析话题的传播路径以及挖掘关键结点，及时回应社会关切，正确引导社会舆论，为财政相关部门制定相关政策提供辅助决策信息，对于构建和谐的财政舆环境具有十分重要的意义。

Big data: The Applications and Challenges in Economics

【嘉宾介绍】

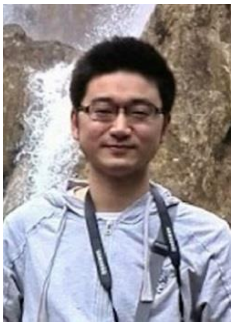
范青亮，经济学博士，毕业于美国北卡罗莱纳州立大学，研究方向为计量经济学、大数据分析、实证经济分析。曾在美国芝加哥布斯商学院、斯坦福大学、南加州大学等访问研究。其研究论文发表于Journal of Econometrics，《中国经济问题》等国际、国内刊物。

【报告摘要】

In this talk, I use two specific examples, first, the big data in NBA games, second, the big data in online computer games to illustrate the new applications and challenges of big data to economics. The first case is studied using FDA methods, while the second utilizes classic economic theory as well as data-mining methods. We believe that big data brings many new perspectives to social sciences including economics, and at the same time, new challenges both computationally and methodologically. We discuss the potential methods to solve these issues and the further directions to pursue.

时序预测的基本方法简介

【嘉宾介绍】



高涛，统计之都成员，2015 年毕业于中国人民大学统计学院，现加入阿里巴巴 iDST 语音团队，曾在百度实习从事时序预测方面的工作，《R 语言实战》译者。

【报告摘要】

时序预测是一个古老但又实用的话题，不少人对时序预测又爱又恨，怒其不准，爱其简单。本演讲将结合笔者过去做时序预测的经验，介绍时序预测的基本思想和预测过程，结合 R 中的软件包介绍单时序和多时序的基本方法：(S)AR(I)MA(X)模型系列、ETS 模型系列、DLM 模型、复杂季节时序模型、函数型时序模型、分层时序、VAR 模型等模型。实际上，理解模型思路便是对数据模式的发现与理解的过程，本演讲希望通过对时序预测基本方法的阐述，和大家探讨实际数据预测的处理思路。

High Dimensional Models in Finance with R

【嘉宾介绍】

杨炳铎，博士毕业于美国北卡罗莱娜大学夏洛特校区数学与统计系，现任教于江西财经大学金融学院。博士毕业至今，一直从事金融计量经济学理论方法及其在经济金融领域的应用研究，论文在 Journal of Econometrics 和 Journal of Banking and Finance 等国际知名杂志上都有发表。

【报告摘要】

对经济金融中的高维数据进行建模是未来的热门研究方向。本次报告主要介绍高维条件下的向量自回归模型，方差协方差估计和投资组合选择以及它们在 R 软件中如何实现。

『应用与可视化专场』

传感器数据中的机器学习实践

【嘉宾介绍】

张源源，毕业于吉林大学信息与计算科学专业。先后在友盟、百度、乐动力等多家公司从事与数据和算法相关的工作。

【报告摘要】

相比 PC 电脑，手机因为有了更多更好的传感器，所以增加了很多比电脑好玩的特性。

作为业界最早的一批使用传感器数据做成产品的公司，乐动力在这个过程中进行了一些机器学习实践。

本次演讲主要讲述计步、自动识别运动、轨迹优化、自适应学习步长等模块背后的机器学习探索，以及在电量、内存、后台不能稳定运行等条件的限制下，如何取得尽量好的结果，最后剖析了乐动力的产品架构和发展历史，期望能抛砖引玉，启发大家对数据改变生活的认识。

当 R 遇到 Python

【嘉宾介绍】

肖凯，一个喜欢折腾数据的人，《数据科学中的 R 语言》作者之一，目前供职于 1 号店商务智能部。

【报告摘要】

介绍数据分析领域两种主流工具的特点，以及二者的协同配合。

APP 用户行为路径挖掘助力产品优化

【嘉宾介绍】

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

【报告摘要】

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

本次演讲主要通过国内某著名 APP 进行实际的用户行为路径挖掘，找到产品的核心优化点，定位流失人群显著属性，帮助 APP 更好的提升用户核心行为，从而带来 APP 下一个核心红利点。

重新定义你的地图：REmap

【嘉宾介绍】

郎大为, SupStat 高级数据科学家，雪晴数据网专职讲师。

博客地址：http://chiffon.gitcafe.io

【报告摘要】

GIS 地理信息可视化是可视化中难度较大，但适用范围较广的一个领域。REmap(http://lchiffon.github.io/REmap/)是基于 Echarts 的 R 包。不同于传统的静态地图，REmap 为广大的数据玩家提供了一种简便、动态、可交互的地理数据可视化工具。本报告集中展示了 REmap 的基本功能，以及运用 REmap 实现的天气预报等案例。

数据可视化爱好者的工具箱

【嘉宾介绍】

周扬，J.D.POWER 数据分析师，擅长数据可视化及利用 R 作为数据处理引擎建立数据应用级数据产品原型。熟悉 R, HTML5/CSS3, Python, JavaScript 等工程开发。曾在国际著名期 Bioinformatics(生物信息学) 上发表论文两篇，在 Nuclear Acid Research (核酸研究) 上发表论文一篇。

【报告摘要】

数据可视化作为数据的重要表现形式，在数据分析以及数据产出中都表现出极为重要的作用。而数据快速膨胀的今天，面向数据及数据可视化的工具和框架也迅速发展，如何在众多的工具和框架中选择适合自身数据需求的组合，成为每一个希望和正在从事数据工作的工作者需要思考的问题。演讲者作为数据可视化爱好者，希望通过自己在实现数据从端到端（即从原始数据到最终数据可视化产出物）的实际工作内容出发，以 workshop 的方式介绍：1）如何利用 R 语言作为数据处理、分析、建模引擎，提供数据可视化的有效支持；2）如何在不同数据分析需求条件下，选择数据可视化（数据呈现端）的工具；3）如何将数据处理、分析、建模的模块与数据可视化模块进行拼接和产物。

R 语言存活分析在商业中的应用

【嘉宾介绍】

李悦，纽约大学硕士毕业，专业金融传媒，现就职于位于纽约的卖方投资研究机构。

【报告摘要】

R 存活分析一直以来在医疗领域有着广泛的应用，特别是对于患者疾病控制的医学研究。近些年，这一历经实践检验并且有着数学理论基础的算法逐渐被其他领域借鉴，其中最为广泛的是商家对于顾客“存活”的分析，通过对顾客交易数据的分析，推算出顾客价值，从而预测营业额，并且更好理解客户，有针对性地投掷广告。目前在美国，这种借助 R 语言进行数据分析，实现有效经营的商业行为发展迅速，值得更广泛的跨领域跨国界交流。

R 语言在业界数据分析的应用

【嘉宾介绍】

郝智恒，阿里巴巴集团——数据技术产品部资深数据挖掘工程师。南开大学概率统计学硕士。《R 语言统计学入门》译者，5 年 R 语言编程经验。立志在经济，商业，数据中间需找立足点。

【报告摘要】

因为 R 对数据的处理特别灵活，可视化功能非常的强大，因此，它在淘宝数据分析工作中有着广泛应用。

本演讲分为三部分，首先会介绍下 R 语言在阿里零售平台的应用情况。这一部分我们介绍整个从数据读取，到分析，包括数据挖掘模型的建立，以及生产任务的部署的流程框架。同时也会介绍基于 shiny 的轻量级 demo 的制作。

第二部分，将会具体介绍 2-3 个具体案例，在这些案例中，我们会看到 R 语言强大的数据整合和处理能力，丰富的数据挖掘算法，强大的可视化展现以及基于 shiny 的 demo 应用。每个案例侧重介绍 R 的一方面特性。

第三部分，我们会交流一下我们在商业界数据分析和挖掘工作的积累的一些心得和体会。

整个演讲比较偏重于应用层面的工作和案例的介绍，希望能够给大家展现业界 R 语言应用的面貌，推动 R 语言在工业界更为广泛的应用。

Shiny，可视化玩具？

【嘉宾介绍】

杨环，现就职于 Mango Solution（檬果商务咨询），负责项目咨询和 R 语言开发。

【报告摘要】

shiny 包，用于 R 语言快速搭建网页交互应用，给统计计算分析代码裹上一层点击拖拽的外衣。随着 shiny 生态链上，大包小包落玉盘，shiny app 逐渐成长为商业应用中最后一环——可视化的实操工具。主讲人将以业界开发 shiny app 为例，对目前 shiny 相关技术及前景作一分享。后端到前端，从此全栈人；玩具亦产品，人生不赢乎。

『视频演讲专场』

Introduction to FeatureHashing

【嘉宾介绍】

Wush Chi-Hsuan Wu is a PhD student from the Institute of Electrical Engineering, National Taiwan University, where he is studying online advertising. He has contributed to several R packages, including digest, RcppCNPy, knitr and ckanr.
Wush is a co-founder of Taiwan R User Group. (http://www.meetup.com/taiwan-R).

【报告摘要】

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

A good approach for pre-processing smaller categorical data is `stats::model.matrix`. However, this approach is infeasible with recommender system data due to programming inefficiencies and pre-processing requirements(all data must be read in and synchronized). In this scenario, streaming algorithms are likely to fail while parallel algorithms become too complicated.

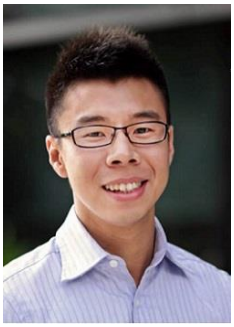
However a solution has been found - feature hashing (also known as the hashing trick). And in 2015, many analysts use this approach to encode such data. The speaker has developed an R package that efficiently processes vast amounts of categorical data.

This package, FeatureHashing(http://cran.r-project.org/web/packages/FeatureHashing/index.html)enables R users to easily apply the feature hashing with an interface similar to `stats::model.matrix`.

Attendees will learn the tips of using the feature hashing, exchange the experience of extending formula interface in R, and hear how some R users have successfully combined FeatureHashing with `xgboost` to do text mining.

用数据科学优化人口健康模式

【嘉宾介绍】



尤晓斌 ,现任新加坡国立医疗集团(National Healthcare Group)数据分析员，曾就读于新加坡国立大学统计系和厦门大学统计系。有 5 年的 R 使用经历。兴趣领域为：统计学习，数据科学，可视化以及人口医疗相关分析。

Email：alex.xbyou@gmail.com，Xiaobin.you@nus.edu.sg

【报告摘要】

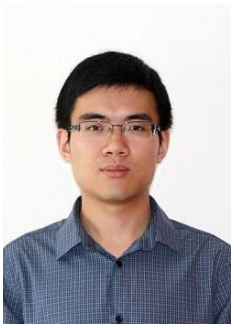
新加坡医疗系统在 2014 年 Bloomberg 医疗体系效率排名中位列第一。哈佛大学医学院教授哈兹尔廷用“价廉质优”一词，形容新加坡用 4%的国内生产总值交出了一流成绩单——全民医疗覆盖，低婴儿死亡率和高预期寿命。

新加坡医疗系统同样面临人口老龄化的挑战。据估计，至 2030 年新加坡 65 岁以上的人口将超总过人口的 20%。为了优化医疗系统以迎接人口老龄化的挑战，新加坡积极探索区域医疗模式，纵向整合综合医院，联合诊所，社区医院及疗养院等医疗资源，形成六大区域医疗协同合作的局面。

数据科学在探索区域医疗的过程中注入了新的科技活力。数据仓库的管理整合医疗机构运营数据以及 SNOMED，ICD 和 WHO drug dictionary 等标准化编码系统；统计学习建模能综合多方面信息协助决策；可视化及 GIS 有助于分析成果的阐释和理解。从数据预处理，建模到最终成果展示这一流程中，R 语言都扮演着重要的角色。

(统计模型+最优化)×(ADMM 算法+并行计算)=？

【嘉宾介绍】



邱怡轩，普渡大学统计系在读博士，统计之都理事会成员，感兴趣的领域包括统计计算与建模，R 语言相关技术等。曾参与翻译《R 语言编程艺术》《R 数据可视化手册》《ggplot2：数据分析与图形艺术》等书籍，是 showtext，rARPACK，recosystem 等 R 软件包的作者。个人主页 http://yixuan.cos.name/cn。

【报告摘要】

最近的几年里，“大数据”的观念越来越深入人心，但回归到问题的本源，如何在大规模的数据上进行统计建模并计算求解，依然是一个极具挑战性的问题。即使是最基本的统计模型，在面临很大的数据量时，要在可接受的时间内完成计算也并非易事。这其中至少包含了两方面的原因：(1)许多模型的设定很简单，但并没有显式的求解公式，例如 Lasso 及其他众多的带惩罚项的统计模型；(2)计算机硬件的发展使得并行计算已经相当普及，但很多模型的求解算法并没有利用这一便利条件。

为了克服这两方面的困难 最近几年 ADMM 算法(Alternating Direction Method of Multipliers) 开始受到越来越多的来自统计学和机器学习领域的关注。ADMM 是一种最优化算法，它主要针对带约束的凸优化问题。由于很大一部分的统计模型求解都可以归结为这一类优化问题，所以 ADMM 在统计学习里的应用非常广，典型的例子包括 Lasso 及其扩展，带正则项的广义线性模型，SVM，分位数回归，压缩感知等等。ADMM 具有几方面的优势：(1)是一种迭代算法，可以根据需要的精度来设定算法终止的时机；这在大规模数据处理中非常关键，因为实际中往往可以允许一定的精度误差，但要求算法在规定的时间内完成；(2)对于许多模型具有显式的迭代公式，实施简单；(3)提供了进行分布式计算的框架，充分利用硬件资源。

本报告将介绍 ADMM 的基本原理和算法，其分布式计算框架，以及若干常用的统计和机器学习模型在 ADMM 下的实现。演讲者还将介绍其编写的 R 软件包 ADMM，展示这些模型在 R 中的实际用法。我们将比较 ADMM 软件包与其他已有软件包（如 glmnet 的 Lasso，quantreg 的中位数回归等）的性能，突出 ADMM 的计算优势所在。

论 R 码农的自我修养

【嘉宾介绍】

谢益辉，爱荷华州立大学统计学博士，现为 RStudio 码农一枚。

【报告摘要】

简单谈谈作为一个 R 码农的一些个人经验，以 R 包的开发为主，包括函数的模块化、文档的自动化（ roxygen2 ）、测试云端化（ Travis CI ）、文档人性化、开发社区化（ Github ）、应用网页化（ JavaScript/htmlwidgets/Shiny/R Markdown ）。

『统计与机器学习专场』

Sparse Graph Coloring Processes A Weak Convergence Result

【嘉宾介绍】

刘路，中南大学数学与统计学院，研究员。在 Journal of symbolic logic 和 Transaction of AMS 上发表论文若干。目前研究方向：过程统计、大偏差理论。http://r.m.baidu.com/4c1arq7

【报告摘要】

For a given sparse graph sequence G^N, we give definition of convergence of graph structure. We study a colouring process. At each time t, one uncoloured node, $\pi(t)$, is coloured by a state in S. The colouring state is random with distribution $v_t\left(\cdot\middle|G^{N,S,t}_{\pi(t),d}\right)\in \mathrm{P}(S)$. Where $v_t\left(\cdot\middle|\pi(t),G^{N,S,t}_{\pi(t),d}\right)$ depends on time t, the subgraph (with colour) within distance d to node $\pi(t)$, denoted by $G^{N,S,t}_{\pi(t),d}$. We prove that for a convergent sequence of sparse graph, any such colouring strategy, $v_t\left(\cdot\middle|G^{N,S,t}_{\pi(t),d}\right)$, which is bounded away from 0 and 1, induces a convergent sequence of stochastic process on coloured graph structure. Actually, what we really prove is somewhat stronger.

We point out the possible applications on large deviation theorems and control problems.

Network Vector Autoregression

【嘉宾介绍】

朱雪宁，本科毕业于中山大学数学与应用数学专业，现为北京大学光华管理学院商务统计系在读博士生。研究方向为社交网络分析，搜索引擎营销等。

【报告摘要】

We consider here a large-scale social network with a continuous response observed for each node at equally spaced time points. The responses from different nodes constitute an ultra high dimensional vector, whose time series dynamics is to be investigated. In the meanwhile, the network structure needs to be taken into consideration. To this end, we propose a network vector auto regressive (NAR) model. NAR models each node's response at a given time point as a linear combination of(a) its previous value, (b) the average of its connected neighbors, (c) a set of node-specific covariates, and (d) an independent noise. The corresponding coefficients are referred to as the momentum effect, the network effect, and the nodal effect respectively. The second-order stationary conditions for the NAR model are derived, and it is found that the network structure plays an important role. In order to estimate the NAR model, an ordinary least squares type estimator is developed, and its asymptotic properties are investigated. We further illustrate the usefulness of the NAR model through a number of interesting potential applications. Simulation studies and an empirical example are presented to demonstrate the performance of the newly proposed methodology.(Joint work with Rui Pan, Guodong Li, and Hansheng Wang.)

拆分抑或耦合：地学集成建模

【嘉宾介绍】

罗立辉，博士，中国科学院寒区旱区环境与工程研究所寒旱区科学大数据中心副研究员。主要研究方向包括冻土分布与灾害模型、寒区生态—水文建模、寒区灾害预警与决策支持。

【报告摘要】

地学建模将地球系统的五大圈层：大气圈、水圈、冰冻圈、岩石圈和生物圈作为一个相互作用的整体来考虑，聚焦于地球系统在时间和空间上的动态变化，而经济学模型则要揭示大量不同性质的经济活动之间的相互关系。随着地学模型的发展，需要从模型集成的角度来构建更为复杂的相互作用模型。地学模型的集成一般有两种方法：（1）模型直接耦合。将多个代表不同物理过程的地学模型进行耦合，以形成新的更为庞大的、复杂的地学模型。（2）模型拆分再耦合。将多个地学模型拆分成大量的独立模块，然后从已拆分的众多模块中优选多个模块进行耦合，以形成一个新的地学模型。本报告从地学模型的机理出发，采用 R 语言来制备地学时空数据，开发了地学模型中的冻土变化模型 EQTEC，并进行模拟计算与可视化分析；拆分和优选不同地学模型以组合成新模型，由此创建的地学建模环境（平台）将为地学模型集成创造出各种可能，由此我们开发了地学集成建模框架 HOME，并在 HOME 平台中将已开发的 EQTEC 与生态-水文模型进行了集成应用研究。

非凸阈值迭代算法的收敛性及其在 SAR 成像中的应用

【嘉宾介绍】

曾锦山，男，江西师范大学计算机信息工程学院副教授。曾锦山于 2008 年 7 月获得西安交通大学理学学士学位，同年 9 月进入该校应用数学系攻读研究生，并于 2015 年 6 月获得西安交通大学理学博士学位。自 2015 年 9 月起在江西师范大学计算机信息工程学院任职。从 2013 年至 2014 年在美国加州大学洛杉矶分校(UCLA)访问。研究兴趣主要包括机器学习、分布式优化、稀疏优化、及其信号处理。现已发表 SCI 论文 11 篇 ,国际会议论文 1 篇 ,授权专利 1 项。目前担任 SIAM Journal on Scientific Computing, IEEE Trans. Signal Processing/ Image Processing， IET Signal Processing 和 Science China Information Science 等多个学术期刊的审稿人。

【报告摘要】

在近年来的稀疏建模研究中，由于非凸罚在诱导稀疏性方面往往比凸罚更有优势，从而非凸罚受到广泛关注。然而与凸罚方法相比，对应的非凸罚算法的收敛性分析通常更具挑战性。本报告将主要介绍一类用于求解稀疏性问题的非凸阈值迭代算法及其收敛性理论，并将所建立的算法与理论应用于典型的稀疏性问题—合成孔径雷达(SAR)成像。数值实验揭示了算法在 SAR 成像应用中的有效性。

Learning to Trade via Recurrent Neural Network and Direct Reinforcement

【嘉宾介绍】

汤耀华，香港大学统计与精算系在读博士，主要研究方向是深度学习在机器翻译和股票交易两个领域的应用。

【报告摘要】

In this research, we consider using recurrent neural network, especially Long-Short Term Memory(LSTM), to guide daily trading of stocks. This is based on the dynamic changing nature of stocks' prices. Unlike the previous researches that only capable of handling a single stock or index, we combine portfolio management and stock selection into a united model. The performance functions that we consider for reinforcement learning are profit or wealth. Experimental results show that our model outperforms common baseline strategies.

Quantile Hysteretic Autoregressive Models

【嘉宾介绍】

曾若辰，香港大学统计及精算系在读博士。

【报告摘要】

The traditional threshold autoregressive model only describes the dynamics of the conditional mean process, so they could be inadequate to capture nonlinearities in the entire quantile process, when different regimes that coincide in the mean process fail to coincide in quantile processes. To capture different behaviors across different quantiles, Galvaoet

al.(2011) proposed the threshold quantile autoregressive models, built on top of the quantile autoregressive model introduced by Koenker and Xiao (2006). Empirical study has found that the for US monthly unemployment growth series, different threshold autoregressive processes for different quantiles were proposed. On the other hand, hysteresis has been widely observed in many macroeconomic series. Li et al. (2012) proposed the hysteretic autoregressive model to incorporate hysteresis into the classical two-regime self-exciting threshold autoregressive model by introducing a more flexible regime switching mechanism. It thus motivates the proposal of a quantile hysteretic autoregressive model that can capture hysteresis in the entire process.

统计之都简介

统计之都（Capital of Statistics，简称 COS）成立于 2006 年 5 月，是一个旨在推广与应用统计学知识的非营利性组织。统计之都网站最初由谢益辉创办，现由世界各地的众多志愿者共同管理维护。目前统计之都的组织形式为线上和线下相结合，其愿景是成为国内杰出的统计学服务平台之一。

目前，统计之都的线上活动主要为主站（http://cos.name/）与中文论坛（http://cos.name/cn/）相结合。其中主站内容以统计学知识介绍和应用案例为主，并同时发表书评、文评以及其他数据科学相关内容导读（“每周精选”栏目）；主站文章来源于本站撰稿人的贡献以及用户推荐，并对这些文章有一定的审稿机制以保证文章质量。中文论坛覆盖数据挖掘、机器学习、生物医学统计、金融统计等各个专业板块，也包括出国留学、就业招聘、招生考研等信息交流板块。

除了线上的交流之外，统计之都社区还定期开展一系列的线下活动，其中最重要的是自 2008 年以来一直延续至今的中国 R 语言会议，迄今为止，R 语言会议的累积参会者超过两万人，累积参会单位有三千多家。除此之外，自 2012 年，统计之都开展了一系列的数据分析沙龙，其主题涵盖了可视化、生物信息、金融、空间统计、海量数据处理等众多前沿的统计学问题。 统计之都希望团结来自学术界、业界的力量，更好的为国内国外的统计学和数据科学工作者服务，传播知识，提供有效的交流和合作的平台。

统计之都活动回顾

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

1. R 语言是在统计和数据挖掘学界广泛应用的编程语言和开发环境，其免费、开源、灵活的特点与统计之都的文化不谋而合。2008 年起，统计之都在人民大学一间教室，举办了第一届中国 R 语言会议，自此，中国 R 语言会议与统计之都一同成长，规模越来越大，至今已成功举办了 7 届。如今 R 会议已经成为国内影响力最大的 R 语言社区盛会，聚学术专家、业界精英、技术大咖于一堂，使各届 R 使用者得到了充

分的交流。应广大支持者的需求，从 2009 年第二届中国 R 语言会议起，在北京和上海设两个分会场，分别于 5 月下旬和 11 中旬举办。第八届 R 语言会议今年 4 月在西安欧亚学院成功举办，5 月在中山大学举办了广州分会场。 6 月 6-7 日，北京大学举办了北京会场，报名人数超过了 4000 人，达到历史之最。今年除了在南昌举办外，还将在武汉、上海等城市举办分会场。关于中国 R 语言会议的历史纪要和最新进展，可以关注官方网站 http://china-r.org/。

2. COS 沙龙是依托于统计之都而成立的线下活动组织。 其宗旨在于凝聚国内外统计领域、数据挖掘及数据分析领域相关人士，打造一个“开放、共享”的线下活动社区。口号是“共享知识、网络人脉”。

目前，COS 沙龙以约每月 1 次的频率在北京、上海、广州等城市举行。自成立之初，累计举行 20 多场次，邀请了国内外众多学界、业界精英担当沙龙分享嘉宾，吸引了多达 1500 人次的与会人员。

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

3. 比赛、交流会、讲座：除了定期的活动之外，我们还将不定期举行或者和其他组织共同举办一些交流会和讲座。北京的交流会主要以中国人民大学为中心，而上海的则更偏重于沟通合作。目前，我们主办或协办过的活动包括：

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