清華大學统计与数据科学系
Department of Statistics and Data Science, Tsinghua University

清华统计这两年

清华大学统计与数据科学系 双年报

Department of Statistics and Data Science, Tsinghua University **Biennial Report**

2023.7-2025.12

北京 · 清华园

“我们要坚持教育优先发展、科技自立自强、人才引领驱动，加快建设教育强国、科技强国、人才强国，坚持为党育人、为国育才，全面提高人才自主培养质量，着力造就拔尖创新人才，聚天下英才而用之。”

—— 习近平总书记在中国共产党第二十次全国代表大会上的报告

清华大学统计与数据科学系

Department of Statistics and Data Science, Tsinghua University

让数据成为推动社会进步的强大动力

Let data become a powerful driving force for social progress

目 录

CONTENTS

TSINGHUA UNIVERSITY
DEPARTMENT OF
STATISTICS AND DATA SCIENCE

第一部分

清华统计大事记

01 清华大学 成立统计与数据科学系

2024 年 7 月 10 日上午，清华大学在主楼接待厅举行统计与数据科学系成立大会。这是清华大学优化学科布局、服务国家战略的重要举措。

清华大学校长、中国科学院院士李路明，国家统计局党组成员、副局长蔺涛，清华大学丘成桐数学科学中心主任、求真书院院长、中国科学院外籍院士丘成桐，中国数学会理事长、上海科技大学副校长、中国科学院院士席南华，哈佛大学统计系和生物统计系教授、美国国家科学院院士、美国国家医学院院士林希虹，宾夕法尼亚大学沃顿商学院讲席教授、国际数理统计学会主席蔡天文，哈佛大学教授、清华大学统计学研究中心名誉主任刘军，中国科学院院士陈松蹊出席大会并致辞。清华大学副校长、中国科学院院士姜培学主持大会并宣读统计与数据科学系成立决定。中国科学院数学与系统科学研究院研究员、中国科学院院士严加安，琶洲实验室（黄埔）主任、西安交通大学原副校长、中国科学院院士徐宗本，相关领域专家学者、学会代表和学校相关单位负责人等百余人参加成立大会。

与会嘉宾共同揭牌

会上，李路明，蔺涛，姜培学，严加安，刘军，徐宗本，陈松蹊，林希虹共同为统计与数据科学系揭牌。

李路明讲话

李路明向与会来宾表示欢迎，向长期关心支持学校发展的社会各界表示感谢。他说，数据已经成为国家基础性战略资源。在全球科技竞争中，只有下好数据这个“先手棋”，才能占据优势地位、牢牢把握未来发展的主动权。清华大学成立统计与数据科学系，就是为了加强统计学与数据科学基础研究，积极促进学科交叉融合，把发展科技第一生产力、培养人才第一资源、增强创新第一动力更好结合起来，主动服务国家大数据战略实施和数字中国建设。

李路明回顾了学校在统计学与数据科学相关领域的发展历程并表示，当前，新一轮科技革命和产业变革深入发展，人类社会正在加速进入数字文明时代。面对崭新的发展机遇，一流大学作为国家战略科技力量的重要组成部分，理应有所作为，也必须有所作为。希望统计与数据科学系肩负历史使命、抓住历史机遇，落实立德树人根本任务，加强培养模式创新和课程体系建设；坚持目标导向和自由探索相结合，凝练基础研究关键科学问题，开展多学科交叉研究；充分激发创新创造活力，加强关键核心技术协同攻关，更好发挥数据要素作用，支撑发展新质生产力；深化高水平科技开放合作，搭建国际化交流平台，主动融入全球创新网络，切实肩负起构建人类命运共同体的科技创新使命。

01 清华大学
成立统计与数据科学系

蔺涛、丘成桐、席南华、林希虹、蔡天文、刘军、陈松蹊致辞（从左至右，从上至下）

蔺涛向清华大学统计与数据科学系成立表示祝贺，希望以此为契机进一步加强国家统计局和清华大学的合作，打造一流学术研究平台，形成一批具有原创性和影响力的研究成果，共同为以中国式现代化全面推进中华民族伟大复兴作出新的更大贡献。

丘成桐回顾了与统计学、数据科学相关领域顶尖学者的交流经历并表示，希望统计与数据科学系与数学中心以及求真书院展开密切的合作，共同推动基础科学领域取得新的更大进步。

席南华代表中国数学会祝贺清华大学统计与数据科学系成立并表示，期待未来清华的智力资源在统计与数据科学领域大放异彩，有力促进统计与数据科学的长远发展。

林希虹说，清华统计与数据科学系的成立，既能极大促进清华乃至中国统计与数据科学的发展，又能更有效地促进多个科学领域的交叉研究。

蔡天文表示，当前统计与数据科学学科的发展尤为重要，一方面能够吸引顶尖专家学者研究统计学与其他科学领域交叉的前沿问题，另一方面也能培养造就大批优秀人才。

刘军表示，统计与数据科学系将致力于发展有较大社会影响力的统计思想和方法；发挥清华在工科、商科、医科、生命科学、社会科学等方面的有利条件，在这些领域开展密切合作；发展互联网技术、大数据分析、人工智能等领域中的统计方法；培养一流统计与数据科学人才。

陈松蹊表示，统计学是以数据为研究对象的基础学科。未来统计学可以在统计与数据科学基础理论、人工智能和机器学习的统计基础理论以及其他学科交叉的重点领域发力，以数据为其他学科赋能。

统计与数据科学发展研讨会同期举行。徐宗本，普林斯顿大学讲席教授、比利时皇家科学院外籍院士范剑青分别以“大模型的极限理论”和“统计数据科学与经济社会”为题作主题报告。

清华大学在统计学与数据科学相关领域具有深厚的积累。我国概率统计学科的奠基人许宝騄 1930 年转入清华大学改学数学，1933 年从算学系毕业，他是中国早期从事概率论和数理统计学研究并达到世界先进水平的一位杰出学者。1979 年，清华大学重建数学系，并布局概率统计等方向，培养出以林希虹院士为代表的一批杰出统计学家。自 2000 年以来，数学科学系教授林元烈、杨瑛始终致力于推动统计学科建设，并于 2008 年促成“统计讲席教授团”的设立，举办统计学讲座、开设统计学课程，极大地促进了统计学科的发展。2011 年学校获批统计学一级学科博士学位授权点，在时任校长陈吉宁的推动及刘军教授的领导下，2015 年成立统计学研究中心（现统计与数据科学系）。清华大学工业工程系也为统计学发展提供了诸多支持。经过各方数十年的不懈努力，清华大学统计学在学术研究、学科建设、人才培养、社会服务等方面取得了长足进步，在数理统计、生物健康统计、统计机器学习及应用、经济与金融统计、工业统计与运筹学、交叉数据科学等重点应用方向形成了特色优势。

未来清华大学统计与数据科学系将紧密围绕国家大数据战略、人工智能行动和《数字中国建设整体布局规划》，立足“四个面向”的战略导向，针对国家重大需求、重大战略、重要部门，培养国际一流的统计学与数据科学领域综合性、创新型高层次人才，以全球视野对标世界一流，努力将统计与数据科学系建设成为国内外知名的产学研一体化学术重镇。

活动现场

02 教育部基础学科系列统计学“101 计划”工作启动会在清华大学举行

2024 年 11 月 19 日，由教育部高等教育司指导，清华大学主办，统计与数据科学系承办的教育部基础学科系列统计学“ 101 计划” 工作启动会在主楼接待厅举行。教育部高等教育司一级调研员侯永峰、清华大学副校长杨斌等出席会议。

清华大学（牵头高校）、北京大学、中国人民大学、华东师范大学、厦门大学、北京师范大学、南开大学、东北师范大学、复旦大学、上海财经大学、中国科学技术大学、上海交通大学、西南财经大学、云南大学、江西财经大学、南方科技大学等 16 所统计学“ 101 计划” 参与高校代表，70 余位海内外知名统计学者参会。

侯永峰表示，作为推动数字经济、人工智能发展的关键学科之一，统计学是我国基础学科的重要组成部分，在诸多领域发挥着越来越重要的作用。全国统计学者要凝聚共识，凝心聚力，聚焦培养统计学领域的拔尖创新人才，做好统计学“ 101 计划” 的探索实践，要明确一流核心课程、一流核心教材、一流核心教师团队和一流核心实践项目这四个基础核心要素的建设任务，全方位深化统计学本科教育教学改革，切实服务国家重大战略需求。

杨斌表示，作为统计学“ 101 计划” 的牵头单位，清华大学深感责任重大，使命光荣。2024 年 7 月，清华大学组建统计与数据科学系，旨在加强统计学与数据科学基础学科建设，积极促进学科交叉融合，主动服务国家数据科技产业发展和数字中国建设。学校将全力支持统计学“ 101 计划”相关工作，持续加大对统计学本科教育教学建设的投入力度，做好统筹和保障工作。

北京大学客座讲席教授、中国科学院外籍院士约翰 · 霍普克罗夫特（John Hopcroft）强调了统计学“ 101 计划” 的重要意义，建议教材建设针对教学对象和教学目标设置差异化难易程度，以提高教学的针对性。

02 教育部基础学科系列 统计学“101 计划”工作启动会在清华大学举行


陈松蹊致辞

统计学“ 101 计划”牵头专家、清华大学讲席教授、中国科学院院士陈松蹊介绍了计划的工作方案，并表示，将通过团队攻关的方式切实推动统计学“ 101 计划”落地生根，提升统计学基础学科人才培养能力，为我国现代化建设培养更多具有国际竞争力的高水平统计与数据科学拔尖创新人才。


阳化冰致辞

高等教育出版社副总编辑阳化冰介绍了支撑服务统计学“ 101 计划”核心教材的建设情况，并表示，高教社在教材建设、师资培训和教材国际化传播等方面有长期的积累，希望发挥平台优势，保障计划顺利推进。

杨斌向陈松蹊颁发聘书

陈松蹊向专家组成员颁发聘书

启动会上组建了统计学“ 101 计划”专家组。杨斌向统计学“ 101 计划”牵头专家陈松蹊颁发聘书。陈松蹊向 16 位专家颁发聘书。

统计学“ 101 计划”拟建设 13 门统计学核心课程。各核心课程教材建设牵头人介绍了建设计划，与专家学者就教材的编写思路、目标定位和预期成果进行了交流讨论。

基础学科系列“ 101 计划”是我国拔尖创新人才培养的一项筑基性工程，重点任务是建设一批有高阶性、创新性和挑战度的一流核心课程，一批反映国际学术前沿、具有中国特色的一流核心教材，一支“ 大先生”领衔的一流教师团队和一批科教融汇、产教融合的一流实践项目。以课程、教材、教师、实践项目等基础要素为“ 小切口”，牵引解决人才培养“ 大问题”，带动实现高等教育改革“ 强突破”。

教育部于 2021 年 12 月在计算机领域率先启动本科教育教学改革试点工作计划（简称“ 101 计划”），于 2023 年 4 月启动数学、物理学、化学、生物科学、基础医学、中药学、经济学、哲学等领域的基础学科系列“ 101 计划”。统计学“ 101 计划”工作启动会的举行、标志着统计学正式成为我国基础学科系列“ 101 计划”的成员。

03 清华大学统计与数据科学系新办公空间启用仪式

2025年3月17日，清华大学统计与数据科学系迎来发展历程中的重要里程碑—集教学、科研、交叉研究等功能于一体的新办公空间正式启用。值此之际，清华大学举办学术研讨会，伦敦政治经济学院姚琦伟教授作特邀学术报告，百余位在京高校统计学者共聚清华园，见证这一重要时刻。

空间新启｜学科发展新愿景

本次活动由统计与数据科学系邓柯副教授主持。他回顾了新办公空间从学校拨付到焕新启用的改造历程，并强调统计系建设工作始终以服务学科发展为核心理念。邓柯表示，近年来清华统计学科学术成果丰硕，学术声誉显著提升，新空间将为师生提供更优质的教学科研环境。

学界共贺｜学科建设新起点

统计与数据科学系刘军教授在致辞中指出，新空间标志着清华统计学科硬件设施跻身国际一流水平。他特别感谢学校支持与团队努力，并向全球统计学者发出加入清华的诚挚邀请。

跨界对话｜学科前沿与交叉

清华大学人工智能研究院常务副院长、计算机系孙茂松教授，人文学院院长刘石教授，中国科学院数学与系统科学研究院李启寨研究员，老一辈清华统计人代表、数学系荣休教授林元烈先生，北京大学 / 北京工商大学耿直教授、伦敦政治经济学院姚琦伟教授分别发言，从人工智能、人文社科、基础研究等视角探讨统计学科的交叉价值，并对清华统计系成立以来取得的成绩表示高度肯定，对清华统计学科在学术引领、促进跨学科交叉合作和高端人才培养的前沿探索提出建议，对清华统计学科以及在其引领下的全国统计学科的未来充满期待！

展望未来｜学科发展新坐标

陈松蹊院士总结强调，统计学人才培养应该改变传统范式，不应仅专注于写文章，更要注重理论的落地、学科交叉和高端人才的引进和培养工作。他特别提及教育部基础学科系列统计学“101 计划”落地清华，以及即将成立的“数据科学交叉研究院”。此外，陈院士介绍，清华即将启动统计学本科招生工作，将持续深化跨学科合作和人才培养工作，推动中国统计与数据科学学科迈向世界一流。

首都师范大学崔恒建教授赠送清华统计系墨宝

04 刘军教授
当选美国国家科学院院士

2025 年 4 月 30 日（当地时间 4 月 29 日），美国国家科学院公布了新一届美国国家科学院院士与外籍院士名单，本届共有 120 名院士以及 30 名外籍院士当选。清华大学统计与数据科学系刘军教授当选美国国家科学院院士。

刘军 教授

刘军 1985 年于北京大学获数学学士学位；1991 年于美国芝加哥大学获统计学博士学位。自 2000 年至今，刘军担任美国哈佛大学统计系终身教授，并于 2003–2015 年兼任哈佛生物统计系教授。他曾任哈佛统计系助理教授（1991–1994）；斯坦福大学统计系助理教授、副教授、终身教授（1994–2004）。他于 2015 年领导创建清华大学统计学研究中心（现统计与数据科学系），并任名誉主任至 2024 年。2024 年 7 月他以筹建发展委员会主任身份领导创建清华大学统计与数据科学系。

刘军一直从事于贝叶斯统计理论、蒙特卡洛方法、统计机器学习、状态空间模型和时间序列、生物信息学、计算生物学等方向的研究，并做出杰出贡献，对大数据处理和机器学习领域有深远影响。他于 2002 年获得考普斯会长奖（COPSS Presidents' Award，公认的国际统计学界的最高荣誉）；2010 年获得世界华人应用数学最高荣誉晨兴应用数学金奖（三年一度，不超过 45 岁）；2014 年被 ISI 评为论文高频引用的数学家；2016 年获得泛华统计协会许宝騄奖（三年一度，不超过 51 岁）；2004、2005 年先后成为美国数理统计学会和美国统计学会会士（Fellow）；2022 年当选国际计算生物学会会士。刘军教授还曾任美国统计协会会刊（JASA）联席主编及多个国际一流统计杂志副编等职。截至 2025 年 5 月，他在各类国际顶尖学术杂志（如 Science，Nature，Cell，JASA，JMLR 等）及书刊上发表论文 300 余篇，出版一本专著，被引用 9 万余次（Google scholar）。他已经指导了 40 多位博士、30 多位博士后。

05 清华统计系聘任国际知名统计学家
Edoardo Maria Airoldi 教授为杰出访问教授

清华统计大事记

2025 年 7 月，清华大学统计与数据科学系再添国际学术力量，正式聘任美国天普大学（Temple University）Millard E. Gladfelter 讲席教授、国际知名统计与数据科学专家 Edoardo Maria Airoldi 教授为杰出访问教授，并举行聘任仪式。此次聘任是清华大学加强统计学与数据科学领域国际合作、推动学科前沿发展的重要举措，也是继汤家豪教授加盟后，统计与数据科学系国际化建设的又一里程碑。

Edoardo Maria Airoldi 教授是网络数据分析和统计机器学习等领域的权威学者，其研究成果在统计学、机器学习及交叉学科中具有广泛影响力。他曾担任哈佛大学统计系应用统计与数据科学实验室创始主任，并获多项国际学术奖项。此次聘任，Airoldi 教授将与清华统计与数据科学系建立深度合作，共同推进前沿研究、人才培养等工作，进一步提升清华大学在全球统计与数据科学领域的学术声誉。据悉，Airoldi 教授将积极参与清华大学的教学与科研工作，包括开设短期课程、举办学术讲座，以及推动清华统计系与国际知名院校的学术交流。统计与数据科学系亦将为 Airoldi 教授的访问提供全方位支持。

陈松蹊院士表示，Airoldi 教授的加盟是清华统计学科学术生态的重要补充。期待通过与国际顶尖团队的合作，清华大学统计与数据科学在理论研究与实际应用中取得突破性进展，培养更多具有全球视野的领军人才。

清华大学统计与数据科学系始终致力于国际一流学者的引进工作，为学科发展注入持续动力。此次 Airoldi 教授的加入，将进一步强化清华与国际统计学界的纽带，为清华统计学子和青年学者提供更广阔的学术平台。未来，清华大学统计与数据科学系将继续以开放姿态汇聚全球智慧，推动学科交叉融合，为世界统计学与数据科学发展贡献清华力量。

06 刘军教授加盟清华大学统计与数据科学系担任兴华卓越讲席教授

2025 年 8 月 30 日，清华大学在丙所会议室举行刘军教授“清华大学兴华卓越讲席教授”聘任仪式，国际著名统计学家刘军全职加盟清华。校党委书记邱勇、校长李路明出席聘任仪式。副校长姜培学主持仪式。

邱勇（左一）、李路明（右一）与刘军夫妇合影

李路明代表学校向刘军教授全职加盟清华表示热烈欢迎。他说，刘军教授与清华大学有着深厚渊源，多年来为学校统计学科的建设与发展作出了卓越贡献。清华大学高度重视教师队伍建设，坚持把人才强校作为学校发展的核心战略，持续加大高层次人才引进力度。期待刘军教授全职加盟清华后充分发挥自身广泛的学术影响力，抢抓以人工智能为引领的第四次工业革命机遇，推动清华统计学、数据科学等学科建设不断迈上新台阶，培养更多高水平拔尖创新人才，为学校加快迈向世界一流大学前列贡献力量。

刘军表示，自己在清华园出生、长大，对这所学校有着深厚感情。如今在人生一甲子之际选择回来，是对教育科研事业的热爱，也是家国情怀的召唤。统计与数据科学是跨学科研究的重要方法和思维方式，是人工智能发展的重要基石，具有广阔的发展前景。未来将依托清华在人工智能领域的综合学科优势，尽己所能推动统计学、数据科学等学科快速发展，助力统计与数据科学系建设成为国内外知名的产学研一体化学术重镇。

陈松蹊院士对刘军教授的全职加盟致以敬意，期待与刘军教授一同努力促进清华乃至中国统计与数据科学的高质量发展。

国家有关部委、北京市有关部门及兄弟高校负责人，校内相关院系和部处负责人等参加仪式。

刘军是享誉世界的著名统计学家，美国国家科学院院士、美国统计学会会士、国际数理统计学会会士和国际计算生物学会会士。他在贝叶斯推断、计算生物学、生物信息学等领域做出了一系列奠基性的工作，曾获国际统计学界最高荣誉考普斯会长奖、世界华人应用数学最高荣誉晨兴应用数学金奖。刘军从 2005 年开始担任清华大学客座教授，2015 年主导创建清华大学统计学研究中心（现统计与数据科学系）并担任名誉主任，积极吸纳多位海外优秀教师来校任职，显著提升了清华统计学科的国际影响力。2024 年，他作为筹建发展委员会主任，帮助清华大学建立了统计与数据科学系。

李路明为刘军颁发聘书

07 清华大学成立数据科学交叉研究院

2025 年 9 月 17 日，清华大学数据科学交叉研究院（Institute for Interdisciplinary Data Science, Tsinghua University，英文缩写：THU IIDS）正式成立。副校长姜培学院士，统计与数据科学系主任、讲席教授刘军院士，统计与数据科学系原主任、讲席教授陈松蹊院士等出席成立仪式。

揭牌仪式

清华大学数据科学交叉研究院管理委员会主任由副校长吴华强担任。陈松蹊院士担任研究院院长，统计与数据科学系副教授邓柯担任副院长、副教授俞声担任院长助理。科研院院长刘奕群代表学校宣读该机构成立的决议。刘军、陈松蹊分别发言。

清华大学数据科学交叉研究院依托院系为统计与数据科学系，工业工程系、计算机科学与技术系、地球系统科学系、人文学院、自动化系、医学院为共建院系。研究院将精准聚焦数据驱动的交叉研究、科技创新和应用落地，重点关注领域大数据的深度分析，着力推动统计学方法与人工智能技术的深度融合，积极推动先进数据科学技术的深度应用落地，引领前沿科技创新，服务国家重大战略，为国家在新一轮科技革命和产业变革中赢得主动、抢占先机提供强大智力支持。

在数据科学基础理论方面，研究院聚焦高维复杂数据建模、大数据与机理模型融合推断、大规模系统因果探寻等前沿基础问题，旨在产出具有国际影响力的理论成果，提升我国在数据科学基础研究领域的原始创新能力。

在交叉研究范式创新方面，研究院重点布局“人工智能 + 数据科学”“地球系统科学 + 数据科学”“健康科学 + 数据科学”“数字人文 + 数据科学”及“智慧政务 + 数据科学”等特色方向，旨在通过方法论创新，解决相关领域在宏观决策、精准治理、前沿探索中的重大数据分析瓶颈，形成具有清华特色的交叉研究新高地。

在科研基础设施与工具方面，研究院着力构建统计与数据科学知识图谱、研发数据分析与统计咨询 AI 智能体，旨在为全校乃至国内外交叉学科研究提供先进的公共算法平台与智能化方法支撑，显著提升科研效率。

清华大学数据科学交叉研究院首次管委会召开

2025 年 10 月 17 日，清华大学数据科学交叉研究院召开首次管委会会议，会议聚焦研究院建设发展，就机构共建、团队布局、统计咨询中心重启等议题达成共识：陈松蹊院长汇报研究院已与自动化系、医学院完成共建签约并新增两位管委会委员，正推进机构备案；首批设立数智医疗建模、时空大数据与政务治理、数字人文合作研究、地学数据科学四个团队；会议一致同意重启统计咨询中心，将其定位为服务全校师生的校级平台，建议配置两名固定教师、十名研究助理，给予教学工作量减免，并通过“北京清华前沿交叉创新研究院”灵活引进人才，同时恳请学校提供专项经费保障；吴华强主任在总结中强调应深化与崂山实验室及清华人工智能医院合作，在海洋治理与数智医疗方向实现创新突破。

后排左起：俞声、侯琳、杨瑛、张强、李国良、李飞跃、邓柯，
前排左起：吴华强、陈松蹊、刘奕群

08 清华大学自 2025 年秋季学期起设置统计学本科主修学位，迎来首届本科生

2025 年秋季学期，清华大学统计与数据科学系正式设置统计学本科主修学位，同步迎来首届该专业本科生（统 50 班）。此次本科主修学位的开设，正值清华大学统计与数据科学系成立一周年之际，标志着该系在人才培养体系建设上迈出关键一步，也进一步完善了清华大学在统计与数据科学领域的学科布局。

围绕统计学本科主修学位培养，该系结合统计与数据科学领域的核心需求，设立“统计学”与“数据分析与人工智能”两大培养方向，构建针对性课程与实践体系。在人才培养定位上，该系明确统计学作为从数据中提取信息、建立模型、进行推断的关键学科，是人工智能的底层逻辑与数据科学的基石，强调通过系统培养帮助学生形成科学能力与批判性思维。

同时，结合国家教育部基础学科系列统计学“101 计划”（清华大学为该计划统计学方向牵头单位），加强基础学科人才培养部署，鼓励首届本科生树立高远目标，在关注个人发展的同时培养家国情怀，未来投身解决“卡脖子”难题与重大科学问题，为社会、国家及科学领域发展贡献力量。

TSINGHUA UNIVERSITY
DEPARTMENT OF
STATISTICS AND DATA SCIENCE

第二部分

院士访谈

学科领航者的声音

展 望　PROSPECT

刘　军

兴华卓越讲席教授，美国科学院院士
统计与数据科学系主任

在过去的几年里，全球科技正经历一场前所未有的深刻变革。以大模型为代表的新一代人工智能技术，正广泛而迅速地重塑科学研究、产业体系乃至社会的各个层面。站在这一技术浪潮的前沿，统计与数据科学学科比以往任何时候都更加接近时代的核心议题，肩负着更为重要的责任。在这一历史节点上，我们更加迫切地需要回到统计学的根本逻辑，审视如何在新时代继续推动学科创新、服务国家战略、引领时代发展。

从天文观测到遗传学研究，从农业育种到医学试验，统计学的每一次进步都源自于实际需求。近年来，随着生物医学和多模态数据的爆发式增长，统计学的重要性愈发突出。它不仅是理解复杂系统和应对不确定性的核心工具，更是解决人类社会重大挑战的钥匙。面对层出不穷的新场景，我们作为学科的传承者，应紧抓这一机遇，以开放的姿态走向更广阔的世界，用新的方法回应新的需求，不断拓展统计学深度和广度。

正是在这一背景下，清华大学统计与数据科学系于 2024 年 7 月正式成立。在过去几年里，我们在理论研究和应用实践方面取得了显著突破。展望未来，我向大家提出三点期望。

第一，保持对数据的敬畏与洞察。

无论模型如何复杂、算力如何强大，如果忽视了数据背后的结构与偏差，最终都可能导致错误的结论。统计学的真正价值在于从数据中挖掘出更深层次的机制、规律和趋势，这种能力是我们不可放弃的核心竞争力。

第二，拓展学科的视野与担当。

统计学的发展从不局限于自身，未来也必然与医学、生命科学、社会治理、金融科技等战略领域深度融合。我希望大家继续保持开放的姿态，深刻融入 AI 新理论与方法的研究之中，积极参与跨学科的合作，共同推动创新，服务国家重大战略需求。

第三，追求学科的深度与温度。

在追求技术突破的同时，我们更应关注技术的实际应用与社会价值。统计学不仅要解决复杂的理论问题，更要切实改善人们的生活、增进人民福祉，发挥技术温度的同时体现人文关怀。

最后，我要特别感谢全系师生在过去两年的辛勤付出。是你们的智慧与努力，让清华统计与数据科学系始终走在学科前沿。未来，让我们继续携手，以统计之严谨、数据之洞察、AI 之创新，共同迎接一个更加智能、更加美好的世界。

寄 语　MESSAGE

陈松蹊

兴华讲席教授，中国科学院院士
数据科学交叉研究院院长

在全系老师的共同努力和学校领导的大力支持下，清华大学统计与数据科学系于 2024 年 7 月 10 日正式成立。一年来，我系经历了奠基与开拓并重、传承与创新并行的历程，各项建设工作扎实推进，在多个方面取得了显著成绩。

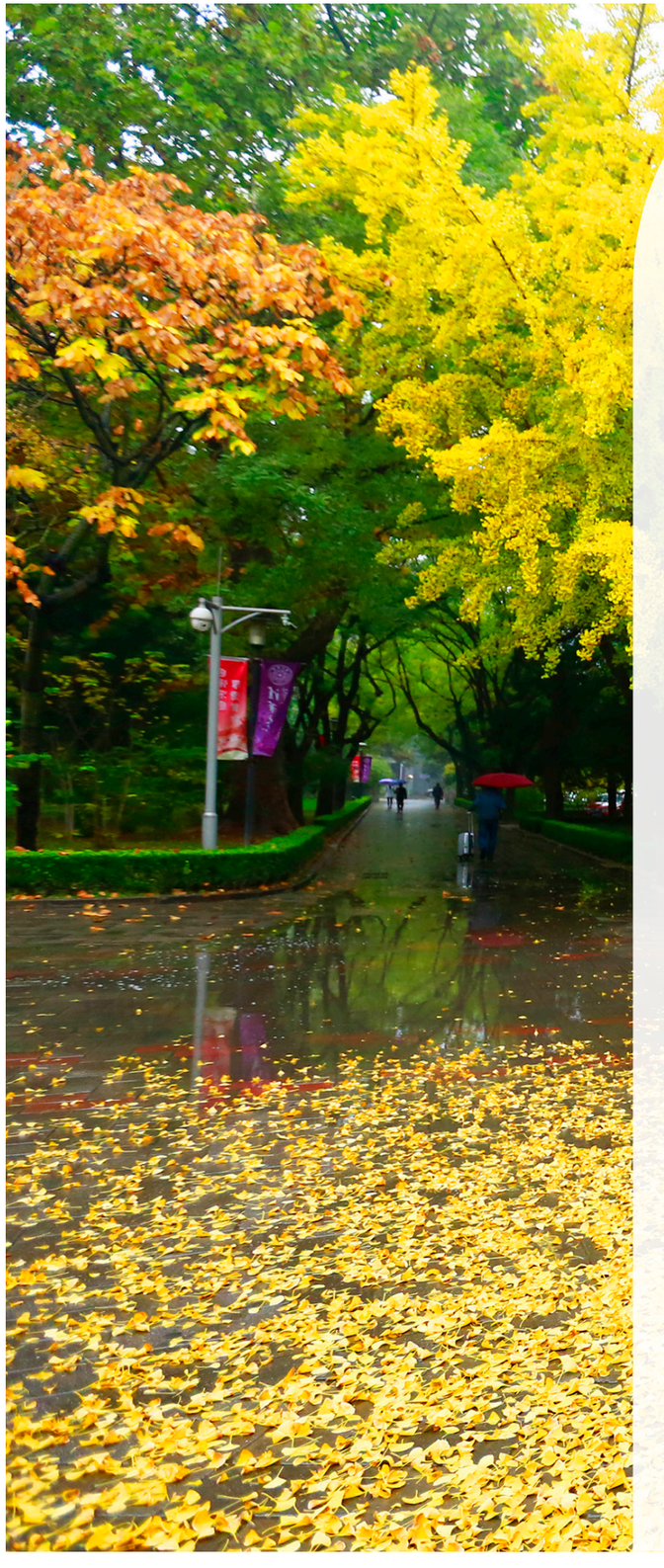
首先在人才培养方面开创新局，我们成功启动了本科生项目，首批至少 19 名本科生已顺利入学，开启了他们的统计与数据科学求索之路。对于一个系而言，拥有本科项目对于清华大学这样的顶尖学府至关重要，我们积极在各书院宣传统计与数据科学的魅力，努力吸引更多的清华本科生主修或辅修统计学。同时，非全日制数据分析师硕士项目也成功获批，目前正在招生进行中。我们力争把它打造成为数字中国建设培养优秀数据分析师的人才项目。

在师资队伍建设方面，刘军院士于 2024 年 7 月全职加入统计系，这是我系发展历程中里程碑式的事件。我与刘军教授在 9 月份顺利完成了系主任的交接。自建系以来，我们引进了 4 位教研系列、1 位教学系列优秀教师，为我系注入了有生力量。此外，房祥忠教授和杨瑛教授也加入我系，担任统计咨询中心联席主任。

2024 年 9 月，我们成立了校级研究机构：数据科学交叉研究院，它是一个以数据科学为抓手、与其他学科进行交叉应用的一个平台，另外统计咨询中心也恢复运行。新系馆的启用，更为我们创造了优良的教学科研环境。

清华统计系取得的进步，离不开国内统计学界同仁们的鼎力支持，非常感谢大家的支持，也希望在今后得到全国统计与数据科学同仁的支持！

站在新的起点上，我们将继续携手并肩，以明德为先、以笃学为本，继续砥砺前行，推动清华乃至全国的统计学与数据科学事业再上新台阶。



TSINGHUA UNIVERSITY
DEPARTMENT OF
STATISTICS AND DATA SCIENCE

第三部分

师生共筑学术共同体 团队风采

师资力量

20 人
教师

4 人
正教授

11 人
副教授

4 人
助理教授

1 人
美国国家科学院院士

1 人
中国科学院院士

3 人
国家级人才项目入选者

7 人
国家级青年人才项目入选者

刘　军　讲席教授
统计与数据科学系系主任
美国国家科学院院士

研究方向：贝叶斯统计理论，蒙特卡洛方法，统计机器学习，状态空间模型和时间序列，生物信息学，计算生物学。

陈松蹊　讲席教授
数据科学交叉研究院院长
中国科学院院士

研究方向：超高维数据统计推断，高分辨数据同化，数学地球物理，气候变化统计学方法，计量经济，人口统计。

杨立坚　教授

研究方向：时间序列，函数型及高维数据的统计推断，以及统计学对经济学、金融学、农学、食品科学、地理学、遗传学、神经科学和管理科学的应用。

杨　瑛　教授

研究方向：统计学中的非 / 半参数估计，纵向数据分析，生存分析，生物统计，空间统计等的理论。

邓　柯　副教授

研究方向：贝叶斯统计理论，统计计算方法，交叉数据科学，人工智能方法。

李　东　副教授

研究方向：复杂时间序列的统计分析，非欧数据分析，空间统计，网络数据分析，机器学习，金融计量学。

侯　琳　副教授

研究方向：统计遗传学，生物信息学，生物统计。

俞　声　副教授

研究方向：文本类医学人工智能，包括大型语言模型、搜索引擎、自动问答、知识图谱等。

刘汉中　副教授

研究方向：因果推断，高维统计，大数据分析，机器学习。

林　乾　副教授

研究方向：数理统计，机器学习，深度学习。

吴未迟　副教授

研究方向：时间序列，变点推断，M 估计，网络数据。

杨朋昆　副教授

研究方向：高维统计理论，机器学习，算法及优化。

李　赛　副教授

研究方向：高维统计方法、机器学习与人工智能的统计学基础、因果推断。

王健桥　助理教授

研究方向：高维统计学，稳健估计，高维遗传和多组学数据分析，生物医学大数据分析。

从　鑫　助理教授

研究方向：大语言模型，自主智能体，数据分析自动化。

金华清　助理教授

研究方向：共形预测，临床试验设计，神经影像数据分析。

Henzi Alexander　助理教授

研究方向：概率预测，非参数回归，序贯假设检验。

周在莹　副教授

全国大学生数学建模竞赛优秀指导老师（全国一等奖），清华大学青教奖、清韵烛光奖、年度教学优秀奖等获得者。所授课程均在教评中位于前 5%。

邓婉璐　副教授

北京高等学校优秀专业课主讲教师，清华大学青教奖、青教赛一等奖、清韵烛光奖、宝钢优秀教师奖等获得者。

郑翔宇　讲师

人才引进与晋升

引进美国国家科学院院士刘军，兴华卓越讲席教授
引进时间：2025 年 7 月
研究方向：贝叶斯统计理论，蒙特卡洛方法，统计机器学习，状态空间模型和时间序列，生物信息学，计算生物学。

引进中国科学院院士陈松蹊教授，兴华讲席教授
引进时间：2024 年 7 月
研究方向：超高维数据统计推断，高分辨数据同化，数学地球物理，气候变化统计学方法，计量经济，人口统计。

刘汉中，晋升长聘副教授
晋升时间：2024 年 2 月
研究方向：因果推断，高维统计，大数据分析，机器学习。

林乾，晋升长聘副教授
晋升时间：2025 年 2 月
研究方向：数理统计，机器学习，深度学习。

引进王健桥助理教授
引进时间：2024 年 12 月
研究方向：高维统计学，稳健估计，高维遗传和多组学数据分析，生物医学大数据分析。

引进从鑫助理教授
引进时间：2025 年 8 月
研究方向：大语言模型，自主智能体，数据分析自动化。

引进金华清助理教授
引进时间：2025 年 9 月
研究方向：共形预测，临床试验设计，神经影像数据分析。

引进郑翔宇讲师
引进时间：2025 年 10 月

引进李赛副教授
引进时间：2025 年 12 月
研究方向：高维统计方法、机器学习与人工智能的统计学基础、因果推断。

引进 Henzi Alexander 助理教授
引进时间：2025 年 12 月
研究方向：概率预测，非参数回归，序贯假设检验。

访问教授

杰出访问教授

汤家豪　教授

伦敦政治经济学院 荣休教授，挪威科学与文学院外籍院士
聘任时间：2019 年 10 月
研究方向：非线性时间序列分析。

Edoardo Maria Airoldi　教授

美国天普大学 Millard E. Gladfelter 讲席教授
聘任时间：2025 年 7 月
研究方向：因果推断，网络数据分析，人工智能的统计学基础。

卓越访问教授

房祥忠　教授

北京大学 退休教授
聘任时间：2025 年 5 月
研究方向：可靠性，生存分析，生物统计，基尼系数，不确定性量化。

行政团队

综合办公室

田　园
综合办主任、人事

钱　倩
科研、学科、财务

肖　剑
财务、资产、后勤

王旻钰
外事、学术活动

教学、学生办公室

侯禹珊
党建、宣传、研究生学生工作

王　泽
研究生教务

张静怡
本科生教务、本科生学生工作

科研管理团队

宋希婷
统计咨询师
邓柯课题组科研助理

李　楠
数据科学交叉研究院执行秘书
陈松蹊课题组科研助理

冯　冉
刘汉中、林乾、吴未迟
杨朋昆课题组科研助理

博士生团队

2019 级及之前： 潘长在　任吉杨　宋　爽　孙　爽　陶宇心
王　掣　王海沣　吴方维　郑思捷　周墨钦

2020 级： 白露佳　冯永真　胡祺睿　李冬煜　卢伟灏　卢　鑫
王　达　徐曼芸　余　成　苑洪意　张卓婧

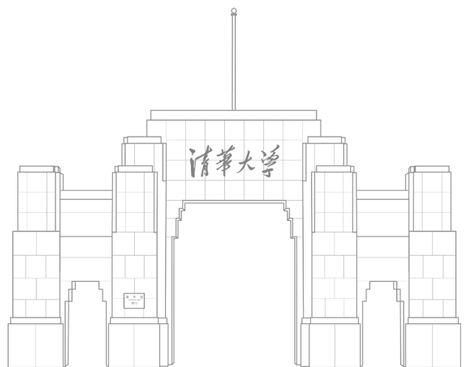
2021 级： 付子初　韩庭萱　李弘梓　陆　瑶　罗天派　马　沄
王羽超　王梓涵　易盈淮　于丁一　张皓博　赵政昀

2022 级： 蔡乐衡　范歆远　江柔蓝　李易诚　罗颖橙　吕逸晨
马昕桐　潘庆一　孙弘毅　应怀原　于浩洋　张灿睿

2023 级： 陈谷涵　陈新佑　付萬嘉　黄　栋　李不凡　盛梓萱
王　琛　王孜睿　徐墨姝　张译尹　周松池

2024 级： 陈诗睿　方　达　高子瞻　沈长乐　孙健博　张安磊
张泓昊　周　扬　祝尔康

2025 级： 陈　双　赖杰鹏　李维佳　林子开　刘　睿　谭圃蕙
王一竹　韦扬昊　谢雨露　张　博　张睿桐　朱梓芸



人才培养

统计学主修专业

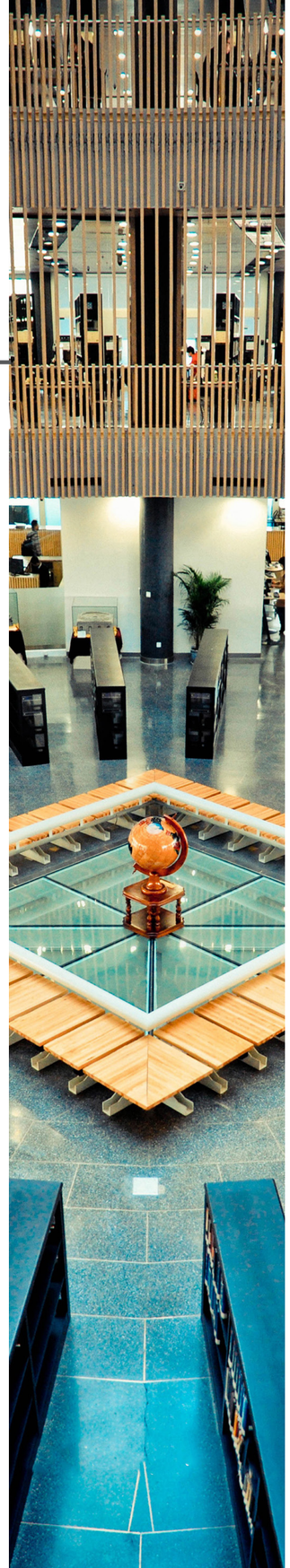
自 2025 年起，清华大学统计学本科专业计划每年通过高考招收约 30 名新生。

本科培养体系注重理论与实践的融合，形成了系统完整、面向未来的专业教育框架。

课程设置主要包含以下几个模块：

··· 数学与计算机基础课程 ···

··· 专业主修课程 ···



··· 专业选修课程 ···

统计系在近十年不断探索和积累辅修专业的培养经验，为本科教育体系的完善和课程建设的创新打下了坚实基础。

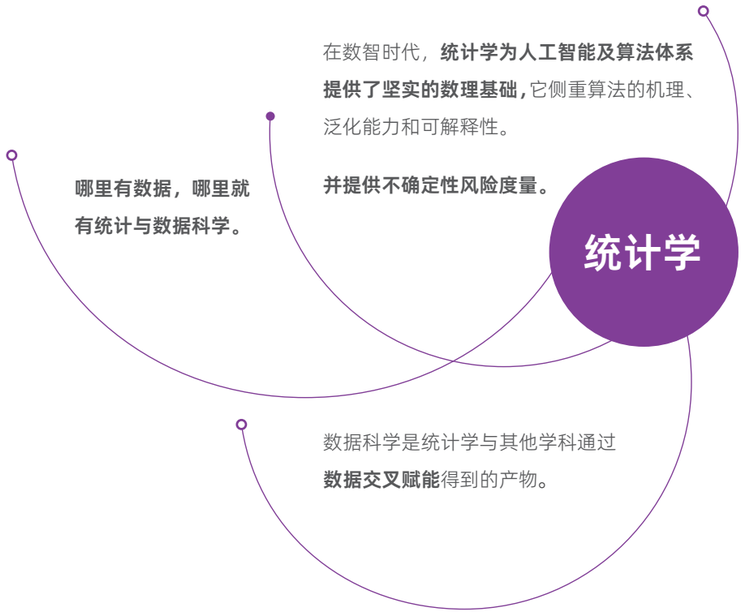
近年来，统计学本科专业课程质量备受认可。其中，邓婉璐和周在莹两位教师开设的课程多次在教学评估中名列前茅，并分别获得第九届和第十届“清韵烛光·我最喜爱的教师”称号，充分彰显了统计系在人才培养方面的卓越成就。

欢迎对统计学感兴趣的同学报考清华大学统计系！

更多课程详情，欢迎大家关注微信公众号：清华大学统计与数据科学系

统计学是关于数据的科学。

在后达尔文时代开宗立科，100 年来已成为理、工、农、医、经济、社会等实证分析和科学推断的方法论。



统计学辅修专业

培养目标

以统计学理论和应用并重为特色，致力于培养具备坚实统计学基础理论知识、统计分析技术、了解统计学研究的前沿动态，并能熟练应用统计学原理解决生产生活中实际问题的跨学科人才；为跨专业本科生攻读统计学相关专业研究生和在工业界、公共卫生领域、政府机关、研究机构等统计相关的岗位就业奠定基础。

招生对象与条件

具有清华大学学籍的全日制在校本科生；
主修专业学习成绩优良；
已经完成微积分、线性代数等课程学习，并成绩优良；
已修课程中无不及格课程；
没有选修其他辅修或第二学位；
每年招收名额由主办单位决定，并报校教务处备案。

学习时间和证书

学习时间：本科主修专业毕业前完成，按照学分制管理机制，修满 28 学分，成绩合格并获得第一学位者，可获得清华大学统计学辅修专业证书。

课程设置与学分分布

统计学辅修专业总学分 28 学分，课程设置由必修课和选修课两部分组成。辅修专业课程应在主修期间完成。主修专业已达到毕业要求而未完成辅修课程的同学，不允许延长学习年限，但可申请毕业离校后继续修读剩余课程，在主修专业学制 +2 年内完成，可获得辅修专业证书，否则已修辅修课程按任选课记入成绩单。

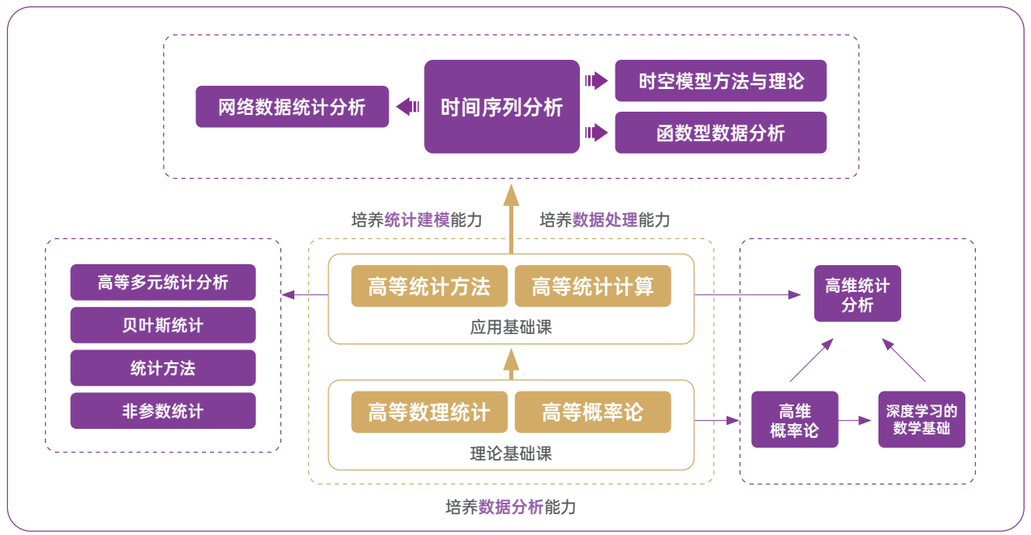
数据分析师硕士项目（非全时专业学位）

2025 年清华大学统计与数据科学系正式启动“数据分析师（非全日制）”硕士项目（Master of Data Analysis, MDA），面向社会公开招生。项目立足数字经济发展需求，旨在培养具备数据建模、商业洞察、智能决策能力的高层次数据分析人才。

本项目以职场人士能力提升与转型发展为主要目标，同时欢迎相关学科背景的应届本科毕业生报考。项目以“紧跟前沿、全栈实践、问题驱动、无缝衔接”为设计理念，课程涵盖统计学方法与思维、数据科学开发技术、人工智能与大模型应用、行业数据分析等核心内容，采用夜间 / 周末的灵活学习方式。毕业颁发硕士研究生毕业证和硕士学位证，学位类别：应用统计硕士（专业学位代码 0252）。

博士生培养

统计与数据科学系面向国际学术前沿和国家重大需求，构建了系统严谨又富有创新活力的人才培养体系。博士生在这里不仅接受扎实的理论训练和严格的科研锻炼，更在跨学科交流、教学实践和社会服务中全面成长。通过资格考试、科研训练、学术交流和助教实践的全链条培养机制，学生能够不断提升学术水平与创新能力，成长为兼具国际视野、专业素养和社会担当的高层次统计学人才。



博士生课程修读框架

教学成果展示

近两年统计系教学评估前 5% 课程

时间	教师	课程	备注
2023–2024 学年度秋季学期	邓婉璐	初等概率论	同规模同类型课堂全校排名第一
2023–2024 学年度秋季学期	邓婉璐	因果推断导论	
2023–2024 学年度春季学期	邓婉璐	多元统计分析	同规模同类型课堂全校排名第一
2023–2024 学年度春季学期	邓婉璐	贝叶斯统计导论	
2023–2024 学年度春季学期	周在莹	实验设计和分析	
2023–2024 学年度春季学期	周在莹	线性回归分析	
2024–2025 学年度秋季学期	邓婉璐	初等概率论	同规模同类型课堂全校排名第一
2024–2025 学年度秋季学期	邓婉璐	因果推断导论	
2024–2025 学年度秋季学期	周在莹	非参数统计导论	同规模同类型课堂全校排名第一
2024–2025 学年度秋季学期	周在莹	统计计算与软件	
2024–2025 学年度春季学期	邓婉璐	多元统计分析	
2024–2025 学年度春季学期	邓婉璐	贝叶斯统计导论	
2024–2025 学年度春季学期	周在莹	线性回归分析	同规模同类型课堂全校排名第一
2024–2025 学年度春季学期	周在莹	实验设计和分析	同规模同类型课堂全校排名第一

- 邓柯 -

教学奖项

《经典与前沿：数字人文的跨学科教学模式》荣获 2025 年清华大学教学成果一等奖

- 侯琳 -

教学奖项

《数智驱动、系统重构的工业工程交叉人才培养体系建设》荣获 2025 年清华大学教学成果一等奖

- 邓婉璐 -

教学奖项

- 入选 2025 届毕业生心目中的“好老师 · 好课程”
- 北京高校优质本科课程(重点)
- 宝钢优秀教师
- 清华大学 2023 年度青年教师教学优秀奖
- 清华大学 2024、2023 年度教学优秀奖

- 周在莹 -

教学奖项

- 国家级一流本科生线下课程
- 2024 年清华大学精品课程
- 第十届“清韵烛光·我最喜爱的教师”
- 清华大学 2024 年课程思政示范课程、示范教师
- 清华大学 2024、2023 年度教学优秀奖
- 全国大学生数学建模竞赛优秀指导老师
- 清华大学 2022 年度青年教师教学优秀奖

团队文化

校友风采

统计与数据科学系校友返校日侧记

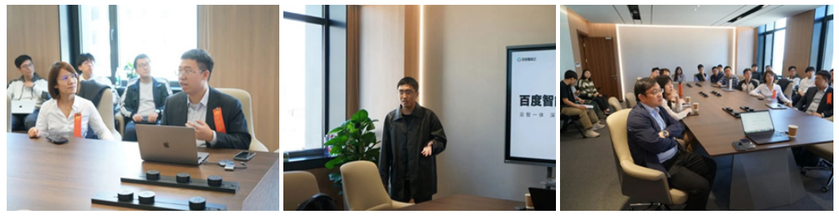
值清华大学 114 周年校庆之际，统计与数据科学系以“数聚时光”为主题，诚邀四海校友重返故园。一场穿越时空的数据诗行，在春日的清华园悄然铺展。

数学系 2015 届校友、现中国人民大学统计与大数据研究院助理教授孟澄作为今年返校优秀毕业生代表之一参加了校党委书记邱勇主持的校友座谈会。孟澄特别感谢了母校对统计学科建设的支持，向母校汇报了其多次带队斩获华为火花奖的事迹。孟澄在本科时即加入统计系（前统计学研究中心）邓柯副教授课题组，开启了科研之路。如今，清华统计学子已投身教学科研一线，并把科研和成果做在了祖国大地上。

2025 年 3 月，统计系启用新办公空间，校友们对统计系新家赞不绝口。统计系将教学启发、科研深耕、高效办公与多元功能有机整合，既承载学科历史积淀，更面向未来人才培养与技术创新的需求，为清华统计人打造了一座连接过去与未来的学术桥梁。

美国国家科学院院士、医学院院士、哈佛大学生物统计系林希虹教授作特邀报告。报告聚焦人工智能（AI）时代如何融合统计学、生成式机器学习以及基因组学和健康科学，推动科学探索。林教授指出，借助生成式机器学习的可扩展、稳健且增强的统计方法，对可信科学至关重要，尤其是在量化不确定性、提升可解释性和加速科学发现方面具有巨大潜力。要以开放心态拥抱 AI，利用合成数据技术增强统计推断，打造可解释、可扩展的分析体系，推动统计学、AI 与领域科学的深度融合，加速科学发现。

本次校庆，统计系特意向曾修读统计学辅修课程的校友们发出邀请。统计系副系主任邓柯副教授、邓婉璐副教授全程参与接待，与校友们共叙情谊。“虽然主修专业不同，但统计系的课程是我清华生涯最难忘的回忆之一。统计系也一直是把辅修学生当亲学生一样培养，让我们找到母系的归属感。”统辅校友们如是说。



2021 届统计学博士林毓聪（现北京理工大学生物医学工程领域助理教授）与 2022 届统计学本科辅修校友赵子健（现百度智能云 - 千帆大模型团队技术骨干），以“学术 + 产业”双视角为学弟学妹勾勒统计人才发展蓝图。

从顶尖学术到前沿产业，校友反哺印证了统计教育的通用价值。未来统计系将持续深化“学科交叉 + 实战赋能”培养体系，助力更多学子在数据浪潮中锚定征途。

就业数据

近两年统计系毕业生去向

姓名	年级	导师	毕业去向	毕业时间
潘长在	2018	邓　柯	中国电信	2024.6
王　掣	2018	邓　柯	中国海洋石油集团有限公司	2024.6
周墨钦	2019	邓　柯	某 IT 企业	2024.6
孙　爽	2019	杨立坚	宾夕法尼亚大学博士后	2024.6
郑思捷	2019	杨立坚	加州大学洛杉矶分校博士后	2024.6
王海沣	2019	邓　柯	某私募基金公司	2024.6
宋　爽	2019	刘　军	哈佛大学博士后	2024.6
陶宇心	2019	李　东	南方科技大学	2024.6
任吉杨	2019	刘汉中	阿斯利康公司	2024.6
白露佳	2020	吴未迟	鲁尔波鸿大学博士后	2024.10
卢伟灏	2020	林　乾	新加坡国立大学博士后	2024.12
卢　鑫	2020	刘汉中	美国圣路易斯华盛顿大学博士后	2025.6
苑洪意	2020	俞　声	某 IT 企业	2025.6
冯永真	2020	杨立坚	北京工商大学	2025.6
胡祺睿	2020	杨立坚	上海财经大学	2025.6
余　成	2020	李　东	芝加哥大学博士后	2025.6

卢鑫：探寻因果的清华统计人

卢鑫，男，汉族，共青团员，统计与数据科学系 2020 级博士生，师从刘汉中副教授，研究方向为因果推断，尤其致力于随机化试验的设计与分析；以第一作者身份在统计学领域顶刊 Biometrika 和 Journal of the American Statistical Association 发表文章；多次担任 Journal of Causal Inference（JCR 二区）期刊审稿人。曾获 2024 年国家奖学金与王大中奖学金，工业工程系未来教授奖学金，综合优秀奖学金（二等）。

1 谁的学术之路不迷茫

回顾本科阶段，卢鑫的综合排名并没有特别突出，仅处于院系中游。他还清楚地记得，第一次将本科毕业论文的手稿交给导师刘汉中老师时的场景，“读起来像枯燥的证明题”。导师在打印稿上密密麻麻写下的上百处批注让他深感挫败，甚至一度对自己的学术能力产生了怀疑。然而，他并没有因此气馁，而是选择直面问题，努力提升自己的科研能力。

进入研究生阶段，卢鑫为自己制定了严格的日程安排，保持着几乎从不懈怠的工作状态。在导师严谨治学风格的影响下，他逐渐养成了注重细节和追求卓越的科研习惯。从论文内容到报告细节，他总是反复推敲，不断优化。在组会上，他也常常对老师提出的问题“一时语塞”，但这并没有打击他的信心，反而激励他更加深入思考，寻找更清晰的表达方式和更完备的解决方案。

正是凭借不懈的坚持，他在学术道路上实现了蜕变，从最初的迷茫到如今在顶级学术期刊发表多篇高质量研究成果。他认为，成果固然重要，更重要的是对学术的热爱和追求。如今，作为组内的“大师兄”，他也在传递这种态度。在组会上，他认真点评学弟学妹们的报告，帮助他们发现不足，为他们的成长助力。

卢鑫在博士生论坛作报告

2 迎接挑战：勇敢开启学术征程

卢鑫的第一篇文章源自导师布置的课题。那时，导师刘汉中副教授在和加州大学伯克利分校的丁鹏副教授的学术交流讨论中提出了一个关于如何在群组随机化试验中应用重新随机化的问题（注：重新随机化是美国科学院院士 Donald B. Rubin 推崇的一种试验设计方法）。彼时的他首次参加组会，便自告奋勇接下了这个课题。

面对完全陌生的研究领域，零科研经验的他并没有退缩，而是全身心投入，认真钻研，通读了大半本经典教材 Cluster Randomised Trials。在他的努力和两位老师的悉心指导下，这项工作最终被统计学顶级期刊 Biometrika 接收。这一成果给丁鹏老师留下了深刻的印象。2024 年初，卢鑫受邀前往加州大学伯克利分校进行为期五个月的学术访问，并在访问期间与丁鹏老师合作解决了一个关于重新随机化与 P 值（p-value）操纵的问题。

从组会上第一次大胆接下挑战，到独自远赴异国开展合作，卢鑫的每一次勇敢尝试都为他带来了丰硕的成果，助力他在学术征途中不断攀登。

卢鑫在学术会议上作壁报展示

3 拓宽边界：开辟多元研究领域

卢鑫在完成每一个研究后，并没有“一招鲜、吃遍天”地开展同质化工作，而是尽量避免做相似的课题，这不仅有助于他保持对科研的热情，也拓宽了他的学术视野。

在完成前两项研究后，他注意到了清华大学交叉信息研究院的王禹皓助理教授发表的一篇文章。这篇文章探讨了在协变量发散情况下如何估计处理效应的问题，而卢鑫决定将这一研究推广到协变量维数和样本数同阶发散的情形。为了解决这一复杂问题，他投入了大量时间学习随机矩阵的理论和相关的中心极限定理，在此过程中，为了证明一个关于随机矩阵的性质，他主动联系了丘成桐数学科学中心的杨帆副教授，并得到了他的帮助。最终，这项工作获得了统计学顶级期刊 Annals of Statistics 的重大修改意见（Major Revision），并被美国艺术与科学院院士哥伦比亚大学 Donald Green 教授列为其研究生课程的参考文献。

完成这一研究后，卢鑫将目光转向学界新兴的“溢出效应”的领域。溢出效应是一个生活中常见的现象，例如，一个人接种疫苗不仅可以降低自己患病的概率，还可能降低他人患病的风险。针对这一方向，他选择将网络随机化试验中估计溢出效应和全局效应的问题作为研究课题。为了深入研究这一问题，他查阅了大量相关文献，并研究了社区发现和网络形成机制等背景知识。最终，他完成了这一研究，目前该工作正在统计学顶级期刊 Journal of the Royal Statistical Society Series B 审稿中。

4 志存高远，展望未来

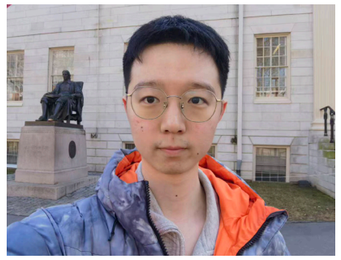
因果推断是统计学的重要组成部分，广泛应用于农业、医学、政治、经济、金融等多个领域。尽管国内在因果推断研究上起步较晚，但作为清华学子，他自觉承担起推动这一领域发展的责任。回顾自己的博士生涯，卢鑫几乎每天都沉浸在学术研究中。舍弃了一些体验生活和享受生活的时间，对于这个选择，他从未感到后悔。

卢鑫下一步的计划是在获得博士学位后，继续开展博士后研究工作，他的目标是争取国外因果推断领域顶尖高校的博士后职位。面对前方那段孤独的旅程，即使充满语言和文化障碍，他依然无怨无悔。

他希望在未来的日子里，能够在统计学领域留下清华学子的名字，并为推动该领域的进步与发展做出自己的贡献。

张皓博：哈佛交流见闻

张皓博，男，汉族，共青团员，统计与数据科学系 2021 级直博生，师从林乾副教授，研究方向为深度学习的数学基础、神经网络的泛化理论。以第一作者身份在统计学领域顶刊 Biometrika、机器学习领域顶刊 Journal of Machine Learning Research、机器学习国际会议 ICML 发表文章。曾获得 2025 年国家奖学金，2025 年国际泛化统计协会（ICSA）中国会议—青年研究员奖，2024、2023 年综合优秀奖学金，2022 年工业工程系系设奖学金。



一、学术合作与研究讨论

在导师林乾老师的指导下，我博士阶段的研究方向主要聚焦神经网络的数学基础，或者说是深度学习的可解释性。近年来，神经正切核（Neural Tangent Kernel, NTK）理论为我们提供了一个理解深度神经网络训练动态的重要窗口——它揭示了充分宽的神经网络的训练动态可以被对应的神经正切核的核回归近似。于是我首先在这一背景下对核回归展开了一系列研究，包括错定情形下核回归的最优性、核插值的泛化能力、高维核回归等理论问题。至此，我们已经对核回归理论有了比较完善的认知，然而，这种近似尽管在理论上简洁优雅，却也忽略了神经网络最核心的特征之一——表征学习（feature learning）的能力。围绕这一问题，我的导师林乾老师在 2023 年开始希望从“自适应核”（adaptive kernel）的角度刻画神经网络表征学习能力以及其优势，将大致的思路具体化一些之后，与当时还在哈佛大学统计系的刘军老师展开了几次讨论。在当时的构想当中，自适应核理论的一个重要组成部分是论证自适应核模型与一种“过参数化的高斯序列模型”的等价关系、并研究后者模型中“过参数化”带来的优势，这方面与许多统计学经典理论联系紧密，因此我们希望与刘老师展开合作。基于双方研究兴趣的契合，在林乾老师的推荐下，我有幸于 2024 年 3 月至 6 月赴哈佛大学统计系，加入刘军老师课题组进行三个月的学术交流，并最终合作完成了一篇综述论文。

在交流期间，我首先就关于此篇综述的背景和具体提纲与刘军老师进行了讨论，刘老师建议我不要仅仅注重从技术角度描述理论结果，而应尝试以“叙事”的方式讲述研究的逻辑主线：为何这一理论重要？它与既有理论的关系是如何的？在他看来，好的学术论文不仅要论证结论的正确性，更要让读者能听懂这个故事并有兴趣阅读进去。这番话让我在论文结构设计和论述逻辑上受益匪浅。随后，我也在刘军老师课题组的组会上就论文内容进行了两次汇报，与课题组的同学们进行了思想的碰撞，产生了很多有意义的新想法，这些都对最终论文的完成起到了至关重要的作用。

二、在哈佛的学习与学术氛围

哈佛统计系的学术氛围给我留下了深刻印象。这里的讨论极为开放且富有创造力。课堂上，学生们经常就老师的讲解提出问题或者追问，老师也会鼓励大家从不同角度表达观点。这种氛围让我感受到课堂不再是单向的信息传递，而是共同探索未知的过程。

哈佛统计系每周都会举办系列学术报告，邀请来自世界各地的学者分享最新研究成果，报告厅常常座无虚席，覆盖的科学问题也非常多元。我印象最深刻的一次是来自斯坦福的 Andrea Montanari 教授的报告，之前我就阅读过许多他的论文，他在深度学习理论领域非常有影响力，他在报告开始时的一句“Forget understanding deep learning – I don't think we have understood logistic regression.”（原话可能有出入）直接逗笑了现场的听众，也风趣幽默地表达了深度学习理论的困难性。虽然全英文的报告对我来说在初期有些挑战，但逐渐习惯之后，我发现这种环境极大地锻炼了自己的学术表达与跨文化沟通能力。

在交流期间，我也结识了很多优秀的同学，有哈佛统计系的中国留学生，也有和我一样来进行学术交流的同学，能够感受到他们对于自己的科研方向都有着很深刻的见解，都很有自己的想法。我们一起听了很多统计系的课程，也参加了统计系每周举办的博士生分享活动，每次由 1-2 名博士生分享他们近期的研究内容，感受哈佛博士生们的科研节奏也令我受益匪浅。

三、生活体验与文化交流

哈佛大学所在的波士顿是美国最古老的城市之一，也是全球著名的学术重镇。查尔斯河贯穿城市中心，两岸风景优美。春暖花开的季节，我常常在河边骑车，看到桥上穿行而过的地铁以及河上划帆船的人们，感受到城市与自然的和谐共生。我还曾与同学一起前往麻省理工学院、波士顿大学、波士顿学院感受不同的校园氛围，在波士顿公园散步感受闲适的氛围，或到昆西市场品尝地道的当地美食。波士顿的华人社区也非常活跃，置身于波士顿市中心的中国城或者 Allston 的中国超市，总能感受到熟悉的亲切感。我也体验了当地的体育与文化活动。波士顿是 NBA 凯尔特人队的主场城市，我有幸在 TD 花园球馆观看了一场比赛。现场气氛热烈、观众热情高涨，那种全场随球起伏的欢呼让我深切体会到 NBA 体育文化的感染力。不过遗憾的是没有赶上这一年凯尔特人夺冠后的冠军游行，那场全城的庆祝盛典一定格外热烈而令人难忘。这些生活经历让我的留学生活不再只是科研的延伸，也成为我理解异国文化的重要窗口。

四、结语

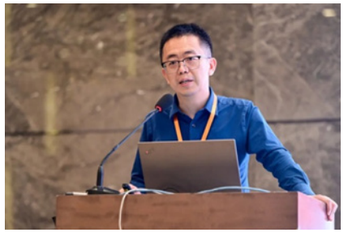
总的来说，这三个月的学术交流经历极大地丰富了我的博士生涯。不论是学术合作的深度，还是文化交流的广度，都让我受益匪浅，这段时间也将成为我珍贵的经历与记忆。非常感谢清华大学统计系为我提供了这一宝贵机会，也感谢林乾老师与刘军老师的悉心指导与支持。未来，我希望能将这段经历中获得的启发延续到自己的研究与工作中，为推动统计学与人工智能领域的交叉发展贡献力量。

北清携手，一路“统”行

2024 年 6 月 6 日，第八届北大 - 清华统计论坛成功举办，论坛由北京大学统计科学中心和清华大学统计学研究中心（现统计与数据科学系）联合发起，本届论坛由清华大学统计学研究中心承办。除北清两校师生外，还受到了很多兄弟高校和业界的学者关注及参与。本届论坛由清华大学统计学研究中心林乾副教授主持。

陈松蹊院士致辞

清华大学杨宇红教授作大会报告



北京大学王汉生教授作大会报告

海报展示环节

优秀成果汇报

经过两校老师从海报质量、工作难度、创新性和科学性四个维度对学生的海报进行严格评审，清华大学白露佳、韩庭萱，北京大学苏武、何沛予荣获“第八届北大 - 清华统计论坛优秀海报奖”，获奖同学分别做成果汇报展示。此外，清华大学宋爽、北京大学顾嘉荣获“第八届北大 - 清华统计论坛优秀毕业生奖”。

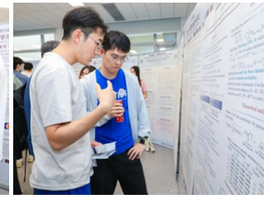
北京大学姚方教授、清华大学侯琳副教授为“优秀海报奖”获得者颁奖

陈松蹊院士、清华大学杨宇红教授为“优秀毕业生”获得者颁奖

两校教师与本年度毕业生合影留念

2025 | 第九届北大 - 清华统计论坛成功举行

2025 年 6 月 5 日，第九届北大 - 清华统计论坛在北京大学智华楼成功举办。论坛由北京大学统计科学中心和清华大学统计与数据科学系联合发起，至今已举办九届，本届论坛由北京大学统计科学中心主办。论坛聚焦跨学科对话，除北京大学和清华大学两校师生外，还吸引了众多兄弟高校和业界学者参与，线上线下参会者共计约 150 人。



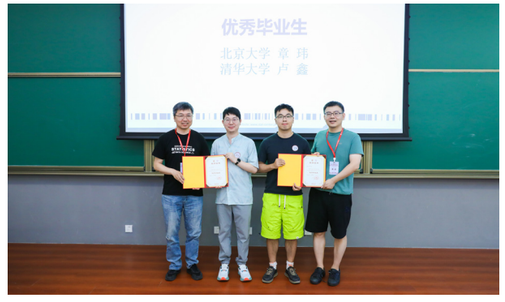
研究成果展示

两校老师从海报质量、工作难度、创新性和科学性四个维度对学生海报进行严格评审，最终评选出 4 份优秀海报奖。北京大学张国宇、谢添雨，清华大学李易诚、余成荣获“第九届北大 - 清华统计论坛优秀海报奖”，获奖同学随后分别进行了成果汇报展示。此外，北京大学章玮、清华大学卢鑫荣获“第九届北大 - 清华统计论坛优秀毕业生”称号。

优秀成果汇报

至此，第九届北大 - 清华统计论坛圆满落幕，期待来年再次相会！

“优秀海报奖”获奖者与颁奖嘉宾合影



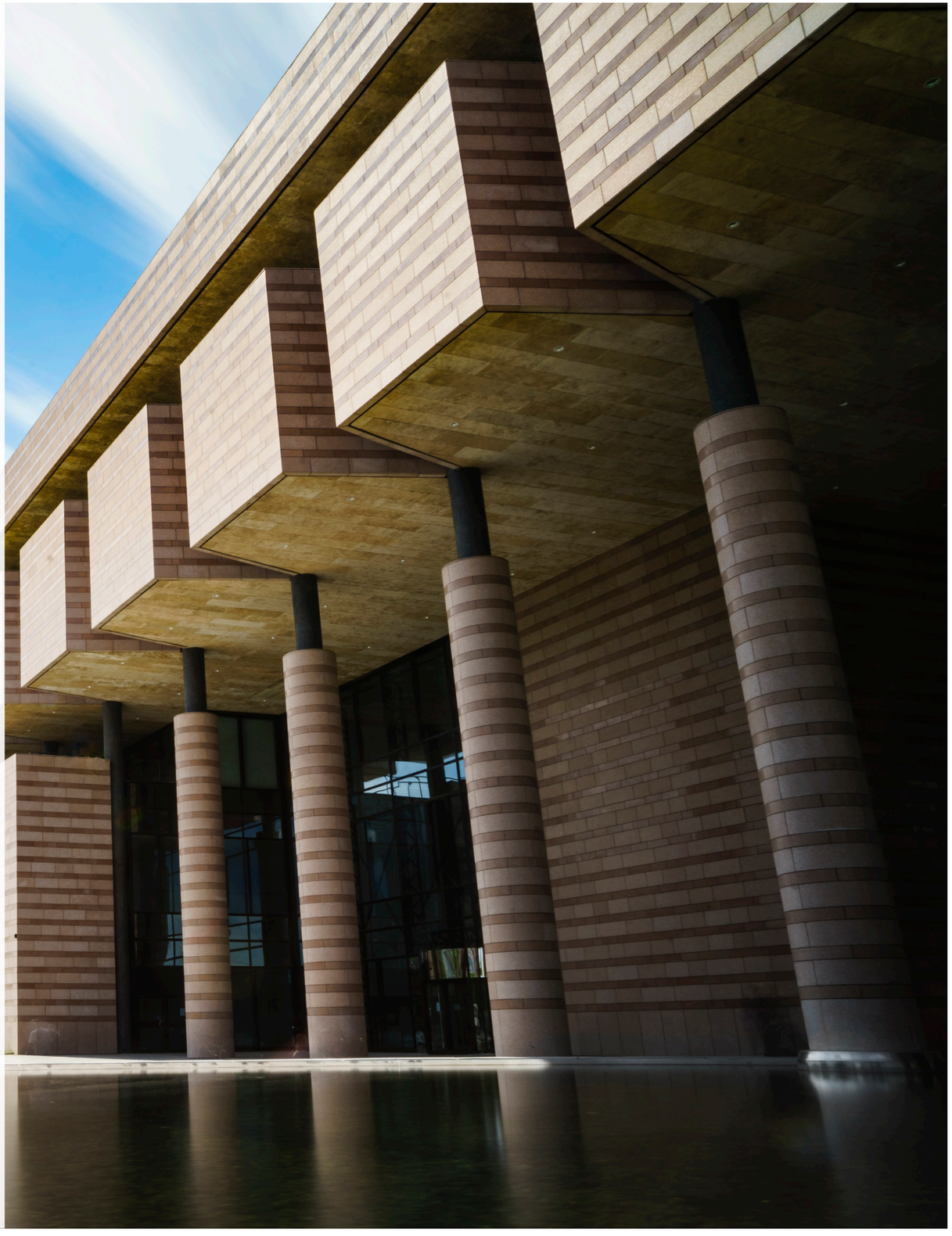
“优秀毕业生”与颁奖嘉宾合影

TSINGHUA UNIVERSITY
DEPARTMENT OF
STATISTICS AND DATA SCIENCE

第四部分

科研成果

创新突破与学术贡献



学术论文

122 篇 发表论文

16 篇 统计四大

5 篇 Nature 子刊

20 篇 AI 领域重要期刊和会议

13 篇 交叉领域重要期刊

28 篇 统计领域重要期刊

前沿论文与理论突破—高质量论文精选

- # incidates co-first author, * indicates corresponding author
- Hanyue Chen, **Songxi Chen*** and Jingkun Qiu (2025). Comments on 'Consistent Climate Fingerprinting' by McKitrick. *Climate Dynamics*, 63: 261.
- Jingkun Qiu, **Songxi Chen** and Qiman Shao (2025). Self-Normalized Cramér Type Moderate Deviation Theorem for Gaussian Approximation. *The Annals of Statistics*, 53(3): 1319-1346.
- **Xiangyu Zheng** and **Songxi Chen*** (2025). Segmented Linear Regression Trees. *Acta Mathematica Sinica, English Series*, 41(2): 498-521.
- Peifeng Tong, **Songxi Chen*** and Chengyong Tang (2025). Multivariate Calibrations with Auxiliary Information. *Statistica Sinica*, 35(3): 1519-1536.
- Peifeng Tong, Haoran Yang, Xinru Ding, Yuchuan Ding, Xiaokun Geng, Shan An, Guoxin Wang, and **Songxi Chen*** (2025). Debiased Estimation and Inference for Spatial-Temporal EEG/MEG Source Imaging. *IEEE Transaction on Medical Imaging*, 44(3): 1480-1493.
- Wu Su, Binghao Wang, Hanyue Chen, Lin Zhu, Xiaogu Zheng and **Songxi Chen*** (2025). A New Global Carbon Flux Estimation Methodology by Assimilation of Both in Situ and Satellite CO2 Observations. *npj Climate and Atmospheric Science*, 7(1): 287.
- Peifeng Tong, Bosi Dong, Xiangdong Zeng, Lei Chen and **Songxi Chen*** (2024). Detection of Interictal Epileptiform Discharges Using Transformer-Based Deep Neural Network for Patients with Self-Limited Epilepsy with Centrotemporal Spikes, *Biomedical Signal Processing and Control*, 101: 107238.
- Han Yan and **Songxi Chen** (2024). Statistical Inference for Four-Regime Segmented Regression Models, *The Annals of Statistics*, 52(6): 2668-2691.
- Jia Gu and **Songxi Chen** (2024). Statistical Inference for Decentralized Federated Learning. *The Annals of Statistics*, 52(6): 2931-2955.
- Haoxuan Sun, Shouxia Wang, Xiaogu Zheng and **Songxi Chen*** (2024). High-Dimensional Ensemble Kalman Filter with Localization, Inflation, and Iterative Updates. *Quarterly Journal of the Royal Meteorological Society*, 150(765): 4870-4884.

- Shuang Sun and **Lijian Yang** (2025+). Statistical Inference for Multivariate Functional Panel Data. *Statistica Sinica*.
- **Qirui Hu** and **Lijian Yang** (2025+). Statistical Inference for Functional Data over Multi-Dimensional Domain. *Statistica Sinica*.
- **Qirui Hu** (2025+). Testing Relevant Hypotheses in Functional Variance Function via Self-Normalization. *Scandinavian Journal of Statistics*.
- **Leheng Cai** and **Qirui Hu** (2025+). Global Inference and Test of Eigen-Systems of Image Data over Complicated Domain. *Journal of Computational and Graphical Statistics*.
- **Leheng Cai** and **Qirui Hu** (2025+). Simultaneous Inference for the Distribution of FPC Scores of Functional Data. *Statistica Sinica*.
- **Leheng Cai**, Xu Guo, Heng Lian and Liping Zhu (2025+). Statistical Inference for High-Dimensional Convoluted Rank Regression. *Journal of the American Statistical Association*.
- **Leheng Cai**, Xu Guo and Wei Zhong (2025+). Test and Measure for Partial Mean Dependence based on Machine Learning Methods. *Journal of the American Statistical Association*.
- Hongyu An, Boping Tian and **Lijian Yang** (2025+). From Scarcity to Insight: Extreme Events Analysis with a Partially Linear Single-Index Varying-Coefficient Model in High-Dimensional Settings. *Journal of Nonparametric Statistics*.
- Xiaowen Liang, Boping Tian and **Lijian Yang** (2025+). Quantile Regression of Partially Linear Varying Coefficient Models with Nonignorable Nonresponse Data. *Journal of the Korean Statistical Society*.
- Xiaowen Liang, Boping Tian and **Lijian Yang** (2025+). Smoothed Partially Linear Varying Coefficient Quantile Regression with Nonignorable Missing Response. *Metrika*.
- Chen Zhong, Yuanyuan Zhang and **Lijian Yang** (2025). Inference and Prediction for ARCH Time Series via Innovation Distribution Function. *TEST*, 34(1): 48-68.
- **Lijian Yang** (2025). Strict Monotonicity of Stochastic Process Extreme Distributions. *Statistics & Probability Letters*, 217: 110292.
- **Lijian Yang** (2025). Continuity of Gaussian extreme distributions. *Statistics & Probability Letters*, 216: 110274.
- **Zening Song** and **Lijian Yang*** and Yuanyuan Zhang (2025). Hypotheses Testing of Functional Principal Components. *Statistica Sinica*, 35: 49-65.
- **Yinghuai Yi**, **Zening Song** and **Lijian Yang** (2025). Inference for ARMA Time Series with Mildly Varying Trend. *Journal of Nonparametric Statistics*.

- **Lijian Yang** (2024). Exact Quantiles of Gaussian Process Extremes. *Statistics & Probability Letters*, 213: 110173.

- **Yongzhen Feng**, Jie Li and Xiaojun Song (2024). Testing Conditional Quantile Independence with Functional Covariates. *Biometrics*, 80(2): ujae036.

- **Qirui Hu** (2024). Change Point Test and Detection of Functional Variance Function with Stationary Error. *Journal of Multivariate Analysis*, 202: 105311.

- **Qirui Hu** and Jie Li (2024). Statistical Inference for Mean Function of Longitudinal Imaging Data over Complicated Domains. *Statistica Sinica*, 34: 955-982.

- **Leheng Cai** and **Qirui Hu** (2024). Simultaneous Inference and Uniform Test for Eigensystems of Functional Data. *Computational Statistics and Data Analysis*, 192: 107900.

- Yi Liu, **Qirui Hu** and Linglong Kong (2024). Tuning-Free Estimation and Inference of Cumulative Distribution Function under Local Differential Privacy. *Proceedings of the 41th International Conference on Machine Learning*.

- Lei Ding, Yang Hu, Nicole Denier, Enze Shi, Junxi Zhang, **Qirui Hu**, Karen Hughes, Linglong Kong, and Bei Jiang (2024). Probing Social Bias in Labor Market Text Generation by ChatGPT: A Masked Language Model Approach. *Advances in Neural Information Processing Systems*.

- Chen Zhong and **Lijian Yang** (2023). Statistical Inference for Functional Time Series: Autocovariance Function. *Statistica Sinica*, 33(4): 2519-2543.

- Yi Liu, **Qirui Hu**, Lei Ding and Linglong Kong (2023). Online Local Differential Private Quantile Inference via Self-Normalization. *Proceedings of the 40th International Conference on Machine Learning*.

- Hongyi Zhou, Wenqing Su, Qixian Zhong and **Ying Yang*** (2027) . Semiparametric Inference for Functional Survival Models. *Statistica Sinica*.

- Fei Ye, Jingsong Xiao, Yulai Miao, Weidong Ma and **Ying Yang** (2026), Consistent community detection approach in the nonparametric weighted stochastic blockmodel with unspecified number of communities, *Journal of Statistical Planning and Inference*, 242, 106339.

- Junhui He, Guoxuan Ma, Jian Kang and **Ying Yang** (2025). Scalable Bayesian Inference for Heat Kernel Gaussian Processes on Manifolds. *Journal of the Royal Statistical Society Series B: Statistical Methodology*.

- Wenqing Su, Xiao Guo,Xiangyu Chang and **Ying Yang*** (2025). Randomized Spectral Clustering for Large-Scale Multi-Layer Networks. *Statistics and Computing*, 35:190.

- Hongyi Zhou, Jin Zhu, Pingfan Su, Kai Ye, **Ying Yang**, Shakeel Gavioli Akilagun and Chengchun Shi (2025). AdaDetectGPT: Adaptive Detection of LLM Generated Text with Statistical Guarantees. *Conference on Neural Information Processing Systems (NeurIPS)*.

- Linru Fu#, Che Wang#, Ye Tao, Zhaoyang Liu, Zhijing Sun*, Lan Zhu*, **Ke Deng*** (2025). High-Quality Triage and Accurate Diagnosis of Gynecological Diseases via Artificial Intelligence. *Artificial Intelligence in Medicine*, 170: 103267.

- Yang Yang, **Ke Deng*** and Yu M. Zhu* (2025) Bayesian Optimization with Pareto-Principled Training for Economical Hyperparameter Optimization. Online published in *Statistica Sinica*.

- Zhe Du#, **Zhaoyang Liu#**, Linru Fu, **Che Wang**, Zhijing Sun*, Lan Zhu* and **Ke Deng*** (2025). Personalized Surgical Treatment Recommendation with Joint Consideration of Multiple Decision-Making Dimensions. *npj Digital Medicine*, 8: 168.

- **Jiaze Xu**, **Changzai Pan** and **Ke Deng*** (2025). Unsupervised Learning of Domain-Specific Texts via a Dual-Dictionary Model. *The Annals of Applied Statistics*, 19(2): 1147-1166.

- Yuxuan Xie, **Shirui Chen**, Fan Zhou, Jun Wang, Yinan Liu, Yucheng Gao, Xiangliang Yan, **Ke Deng** and Chao Chen (2025). Development of a Hybrid Algal Population Prediction (HAPP) Model by Algae Growth Potential Estimation and Time Series Regression and Its Application in One Reservoir in China. *Water Research*, 287: 124419.

- **Zhaoyang Liu**, **Tingxuan Han**, Donald B. Rubin and **Ke Deng*** (2025). A Bayesian Criterion for Rerandomization. Online published in *Journal of the American Statistical Association*.

- Zhichao Tian, **Yang Yang**, Sui Zhou, Tian Zhou, **Ke Deng**, Chunlin Ji, Yejun He and **Jun S. Liu** (2024). High-Dimensional Bayesian Optimization for Metamaterial Design. *Materials Genome Engineering Advances*, 2(4): e79.

- **Yichao Li#**, Wenshuo Wang#, **Ke Deng*** and **Jun S. Liu*** (2024). Differentiable Particle Filters with Smoothly Jittered Resampling. *Statistica Sinica*, 34: 1241-1262.

- Shaojun Guo, **Dong Li**, Xinghao Qiao and **Yizhu Wang** (2025). From Sparse to Dense Functional Data in High Dimensions: Revisiting Phase Transitions from a Non-Asymptotic Perspective. *Journal of Machine Learning Research*, 26(15): 1-40.

- **Dong Li**, Xinghao Qiao and **Zihan Wang** (2025). Factor-Guided Estimation of Large Covariance Matrix Function with Conditional Functional Sparsity. *Journal of Econometrics*, 251: 106070.

- Shaojun Guo, Xinghao Qiao, Qingsong Wang and **Zihan Wang*** (2025). Factor Modelling for High-Dimensional Functional Time Series. *Journal of Business & Economic Statistics*.

- **Cheng Yu**, **Dong Li**, Feiyu Jiang and Ke Zhu (2025). Matrix GARCH Model: Inference and Application. *Journal of the American Statistical Association*, 120(551): 1747-1762.

- Yipeng Zhuang, **Dong Li**, Philip L. H. Yu and Wai Keung Li (2024). On Buffered Moving Average Models. *Journal of Time Series Analysis*.

- **Yuxin Tao**, **Dong Li** and Xiaoyue Niu (2024). Grouped Network Poisson Autoregressive Model. *Statistica*

Sinica, 34(3): 1603-1624.

- **Xinyu Zhang** and **Dong Li*** (2024). Smooth Transition Moving Average Models: Estimation, Testing and Computation. *Journal of Time Series Analysis*, 45(3): 463-478.

- **Xuanling Yang**, Zhoufan Zhu, **Dong Li** and Ke Zhu (2024). Asset Pricing via the Conditional Quantile Variational Autoencoder. *Journal of Business & Economic Statistics*, 42(2): 681-694.

- **Xinyu Zhang**, **Dong Li*** and **Howell Tong** (2024). On the Least Squares Estimation of Multiple-Threshold-Variable Autoregressive Models. *Journal of Business & Economic Statistics*, 42(1): 215-228.

- Wenbo Guo, Zikang Yin, Qinglin Mei, Lianshuo Li, Yonghui Gong, Xinqi Li, Wei Zhang, Wenjie Lei, Bingqiang Liu, **Lin Hou**, Mei Yang, Jin Gu (2025). BCMA: An Integrative and Versatile Database for Multi-Scale and Multi-Omics Molecular Atlas of Breast Cancer. *Computational and Structural Biotechnology Journal*.

- **Shuang Song**, Lijun Wang, **Lin Hou*** and **Jun S. Liu*** (2024). Partitioning and Aggregating Cross-Tissue and Tissue-Specific Genetic Effects to Identify Gene-Trait Associations. *Nature Communications*, 15(1): 5769.

- Keyi Li, Xiaoyang Chen, **Shuang Song**, **Lin Hou**, Shengquan Chen and Rui Jiang (2024). Cofea: Correlation-Based Feature Selection for Single-Cell Chromatin Accessibility Data. *Briefings in Bioinformatics*, 25(1): bbad458.

- **付子初，杨茗，侯琳 *** (2025). 单细胞扰动数据分析 . ***中国科学：数学*** , 55(7): 1-16.

- **Songchi Zhou** and **Sheng Yu***(2025). High-throughput Biomedical Relation Extraction for Semi-structured Web Articles Empowered by Large Language Models. *BMC Medical Informatics and Decision Making* , 25: 351.

- Shan Gao, Kaixian Yu, Yue Yang, **Sheng Yu**, Chenglong Shi, Xueqin Wang, Niansheng Tang*, Hongtu Zhu*(2025). Large Language Model Powered Knowledge Graph Construction for Mental Health Exploration. *Nature Communications*, 16(1): 7526.

- Yue Yang, Kaixian Yu, Shan Gao, **Sheng Yu**, Di Xiong, Chuanyang Qin, Huiyuan Chen, Jiarui Tang, Niansheng Tang and Hongtu Zhu (2025). Alzheimer’ s Disease Knowledge Graph Enhances Knowledge Discovery and Disease Prediction. *Computers in Biology and Medicine*, 192: 110285.

- **Hongyi Yuan#**, **Songchi Zhou#**, and **Sheng Yu***(2024). EHRDiff: Exploring Realistic EHR Synthesis with Diffusion Models. *Transactions on Machine Learning Research*.

- **Huaiyuan Ying**, **Zhengyun Zhao**, Yang Zhao, Sihang Zeng and **Sheng Yu*** (2024). CoRTEx: Contrastive Learning for Representing Terms via Explanations with Applications on Constructing Biomedical Knowledge Graphs. *Journal of the American Medical Informatics Association*, 31(9): 1912-1920.

- **Hongyi Yuan** and **Sheng Yu*** (2024). Efficient Symptom Inquiring and Diagnosis via Adaptive Alignment of Reinforcement Learning and Classification. *Artificial Intelligence in Medicine*, 148: 102748.

- **Zhengyun Zhao**, Qiao Jin, Fangyuan Chen, Tuorui Peng and **Sheng Yu*** (2023). A Large-Scale Dataset of Patient Summaries for Retrieval-Based Clinical Decision Support Systems. *Scientific Data*, 10(1): 909.

- Keming Lu, Yuanren Tong, Si Yu, Yucong Lin, Yingyun Yang, Hui Xu, Yue Li*, and **Sheng Yu*** (2023). Building a trustworthy AI differential diagnosis application for Crohn’ s disease and intestinal tuberculosis. *BMC Medical Informatics and Decision Making*, 23(1): 160.

- **Haoyang Yu**, **Ke Zhu*** and **Hanzhong Liu** (2025). Sharp Variance Estimator and Causal Bootstrap in Stratified Randomized Experiments. *Statistics in Medicine*, 44(13-14): e70139.

- **Wanjia Fu**, Yingying Ma and **Hanzhong Liu*** (2025). Regression Adjustment in Covariate-Adaptive Randomized Experiments with Missing Covariates. *Statistics in Medicine*, 44(25-27): e70304.

- **Xin Lu**, Fan Yang and Yuhao Wang (2025). Debiased Regression Adjustment in Completely Randomized Experiments with Moderately High-Dimensional Covariates. *The Annals of Statistics*, 53(4): 1535-1558.

- **Xin Lu** and **Hanzhong Liu*** (2024). Tyranny-of-the-Minority Regression Adjustment in Randomized Experiments. *Journal of the American Statistical Association*, 120(550): 846-858.

- Wenqi Shi, Anqi Zhao and **Hanzhong Liu*** (2024). Rerandomization and Covariate Adjustment in Split-Plot Designs. *Journal of Business & Economic Statistics*.

- **Ke Zhu**, **Hanzhong Liu*** and Yuehan Yang* (2024). Design-Based Theory for Lasso Adjustment in Randomized Block Experiments with a General Blocking Scheme. *Journal of Business & Economic Statistics*.

- **Hanzhong Liu**, **Jiyang Ren** and Yuehan Yang* (2024). Randomization-Based Joint Central Limit Theorem and Efficient Covariate Adjustment in Randomized Block 2K Factorial Experiments. *Journal of the American Statistical Association*, 119(545): 136-150.

- **Ke Zhu** and **Hanzhong Liu*** (2024). Rejoinder to Reader Reaction “On Exact Randomization-Based Covariate-Adjusted Confidence Intervals” by Jacob Fiksel. *Biometrics*, 80(2): ujae052.

- Fuyi Tu, Wei Ma and **Hanzhong Liu*** (2024). A Unified Framework for Covariate Adjustment under Stratified Randomization. *Stat*, 13(4): e70016.

- Yujia Gu, **Hanzhong Liu** and Wei Ma* (2023). Regression-Based Multiple Treatment Effect Estimation under Covariate-Adaptive Randomization. *Biometrics*, 79(4): 2869-2880.

- **Ke Zhu** and **Hanzhong Liu*** (2023). Pair-Switching Rerandomization. *Biometrics*, 79(3): 2127-2142.

- **Qingyi Pan** and **Yicheng Li** (2025). Functional Virtual Adversarial Training for Semi-Supervised Time Series Classification. *Advances in Neural Information Processing Systems*.

- **Haobo Zhang**, **Weihao Lu** and **Qian Lin*** (2025). The Phase Diagram of Kernel Interpolation in Large Dimensions. *Biometrika*, 112(1): asae057.

- **Guhan Chen**, **Yicheng Li** and **Qian Lin*** (2024). On the Impacts of the Random Initialization in the Neural Tangent Kernel Theory. *Advances in Neural Information Processing Systems*, 37: 35909-35944.
- **Yicheng Li** and **Qian Lin*** (2024). Improving Adaptivity via Over-Parameterization in Sequence Models. *Advances in Neural Information Processing Systems*, 37: 52438-52480.
- **Yicheng Li**, **Haobo Zhang** and **Qian Lin*** (2024). Kernel Interpolation Generalizes Poorly. *Biometrika*, 111(2): 715-722.
- **Yicheng Li**, Zixiong Yu, **Guhan Chen** and **Qian Lin** * (2024). On the Eigenvalue Decay Rates of a Class of Neural-Network Related Kernel Functions Defined on General Domains. *Journal of Machine Learning Research*, 25(82): 1-47.
- **Haobo Zhang**, **Yicheng Li** and **Qian Lin*** (2024). On the Optimality of Misspecified Spectral Algorithms. *Journal of Machine Learning Research*, 25(188): 1-50.
- **Yicheng Li** and **Qian Lin*** (2023). On the Asymptotic Learning Curves of Kernel Ridge Regression under Power-Law Decay. *Advances in Neural Information Processing Systems*, 36: 49341-49364.
- **Weichi Wu**#, Zhou Zhou# and Yongmiao Hong (2025+) Inference for time-varying factor models under local stationarity. *Journal of Econometrics*.
- **Weichi Wu***, Sofia Olhede and Patrick Wolfe (2025). Tractably modelling dependence in networks beyond exchangeability. *Bernoulli*, 31(1): 584-608.
- **Lujia Bai** and **Weichi Wu*** (2025). Uniform Variance Reduced Simultaneous Inference of Time-Varying Correlation Networks. *IEEE Transactions on Information Theory*.
- **Tianpai Luo**#, **Xinyuan Fan**# and **Weichi Wu*** (2025).Simultaneous Statistical Inference for Off-Policy Evaluation in Reinforcement Learning. *Advances in Neural Information Processing Systems*.
- **Xinyuan Fan**#, Feiyan Ma#, Chenlei Leng* and **Weichi Wu*** (2025). Low-Rank Graphon Learning for Networks. *Advances in Neural Information Processing Systems*.
- Yang Han, **Weichi Wu*** and Wenyang Zhang (2025). A New Approach for Homogeneity Pursuit in Short Panel Data Analysis. *Journal of the American Statistical Association.*
- Tianpai Luo and **Weichi Wu*** (2025). Simultaneous Inference for Monotone and Smoothly Time-Varying Functions under Complex Temporal Dynamics. *Journal of the American Statistical Association.*
- Mohsen Asaadi, Fan Yang* and **Weichi Wu*** (2025). Operational Zone-Specific Univariate Alarm Design for Incipient Faults. *Journal of Process Control*, 155: 103536.
- **Xinyuan Fan**, **Bufan Li**, Chenlei Leng and **Weichi Wu*** (2025). Learning Changes in Graphon Attachment Network Models. *The 42nd International Conference on Machine Learning.*

- Yufei Huang, Changhu Wang, Junjie Tang, **Weichi Wu*** and Ruibin Xi* (2025). A Generic Family of Graphical Models: Diversity, Efficiency, and Heterogeneity. *The 42nd International Conference on Machine Learning.*
- Xian Chen, Kun Huang, **Weichi Wu*** and Hai Jiang* (2025). Detecting Multiple Changepoints by Exploiting Their Spatiotemporal Correlations: A Bayesian Hierarchical Approach. *INFORMS Journal on Data Science.*
- **Zirui Wang**, Wodan Ling and Tianying Wang (2025). A Semiparametric Quantile Regression Rank Score Test for Zero-Inflated Data. *Biometrics*, 81(2): ujaf050.
- **Zirui Wang** and Tianying Wang (2025). A Semiparametric Quantile Single-Index Model for Zero-Inflated Outcomes. *Statistica Sinica*.
- **Lujia Bai** and **Weichi Wu*** (2024). Detecting Long-Range Dependence for Time-Varying Linear Models. *Bernoulli*, 30(3): 2450-2474.
- **Weichi Wu** and Zhou Zhou* (2024). Multiscale Jump Testing and Estimation under Complex Temporal Dynamics. *Bernoulli*, 30(3): 2372-2398.
- **Lujia Bai** and **Weichi Wu*** (2024). Difference-Based Covariance Matrix Estimation in Time Series Nonparametric Regression with Application to Specification Tests. *Biometrika*, 111(4): 1277-1292.
- **Dong Huang**, Xianwen Song and **Pengkun Yang** (2025). Information-Theoretic Thresholds for the Alignments of Partially Correlated Graphs. *IEEE Transactions on Information Theory*.
- Muxing Wang, **Pengkun Yang**, Lili Su (2025). On the Convergence Rates of Federated Q-Learning across Heterogeneous Environments. *Transactions on Machine Learning Research*.
- **Dong Huang** and **Pengkun Yang** (2025). Sample Complexity of Correlation Detection in the Gaussian Wigner Model. *The 42nd International Conference on Machine Learning*.
- **Pengkun Yang** and Jingzhao Zhang (2025). Fast and Multiphase Rates for Nearest Neighbor Classifiers. *The 38th Annual Conference on Learning Theory*.
- **Yichen Lyu** and **Pengkun Yang** (2025). Identifiability and Estimation in High-Dimensional Nonparametric Latent Structure Models. *The 38th Annual Conference on Learning Theory*.
- **Yun Ma**, Yihong Wu and **Pengkun Yang** (2025). On the Best Approximation by Finite Gaussian Mixtures. *IEEE Transactions on Information Theory*.
- Lili Su, Jiaming Xu and **Pengkun Yang** (2024). Global Convergence of Federated Learning for Mixed Regression. *IEEE Transactions on Information Theory*, 70(9): 6391-6411.
- Xingjian Li, **Pengkun Yang**, Yangcheng Gu, Xueying Zhan, Tianyang Wang, Min Xu and Chengzhong Xu (2024). Deep Active Learning with Noise Stability. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(12): 13655-13663.

- **Dong Huang**, Xianwen Song and **Pengkun Yang** (2024). Information-Theoretic Thresholds for the Alignments of Partially Correlated Graphs. *The 37th Annual Conference on Learning Theory*.

- Xingjian Li, Abulikemu Abuduweili, Humphrey Shi, **Pengkun Yang**, Dejing Dou, Haoyi Xiong and Chengzhong Xu (2023). Semi-Supervised Transfer Learning with Hierarchical Self-Regularization. *Pattern Recognition*, 144: 109831.

- Yuchao Chen, Haoyue Tang, Jintao Wang, **Pengkun Yang** and Leandros Tassiulas (2023). Sampling for Remote Estimation of an Ornstein–Uhlenbeck Process through Channel with Unknown Delay Statistics. *Journal of Communications and Networks*, 25(5): 670-687.

- Zihan Tong, Shun Kong Cheung, Zheyu Ren, **Pengkun Yang** and Qiming Shao (2023). Modeling of Multi-Level Spin-Orbit Torque-MRAM: Scalability, Stochasticity, and variations. *2023 IEEE International Magnetic Conference - Short Papers (INTERMAG Short Papers)*.

- Zhen Miao, **Jianqiao Wang**, Kernyu Park, Da Kuang and Junhyong Kim (2025). Depth-Corrected Multi-Factor Dissection of Chromatin Accessibility for scATAC-seq Data with PACS. *Nature Communications*, 16(1): 1-15.

- Bohan Lyu, **Xin Cong***, Heyang Yu, Pan Yang, Cheng Qian, Zihe Wang, Yujia Qin, Yining Ye, Yaxi Lu, Chen Qian, Zhong Zhang, Yukun Yan, Yankai Lin, Zhiyuan Liu and Maosong Sun (2025). Enhancing Open-Domain Task-Solving Capability of LLMs via Autonomous Tool Integration from GitHub. *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*.

- Runchu Tian, Yanghao Li, Yuepeng Fu, Siyang Deng, Qinyu Luo, Cheng Qian, Shuo Wang, **Xin Cong***, Zhong Zhang, Yesai Wu, Yankai Lin, Huadong Wang and Xiaojiang Liu (2025). Distance between Relevant Information Pieces Causes Bias in Long-Context LLMs. *Findings of the Association for Computational Linguistics: ACL 2025*.

- Shengda Fan, **Xin Cong***, Yuepeng Fu, Zhong Zhang, Shuyan Zhang, Yuanwei Liu, Yesai Wu, Yankai Lin, Zhiyuan Liu and Maosong Sun (2025). WorkflowLLM: Enhancing Workflow Orchestration Capability of Large Language Models. *The Thirteenth International Conference on Learning Representations*.

- Yining Ye, **Xin Cong***, Shizuo Tian, Yujia Qin, Chong Liu, Yankai Lin, Zhiyuan Liu and Maosong Sun (2025). Rational Decision-Making Agent with Learning Internal Utility Judgment. *The Thirteenth International Conference on Learning Representations*.

- Guoxin Chen, Zhong Zhang, **Xin Cong***, Fangda Guo, Yesai Wu, Yankai Lin, Wenzheng Feng and Yasheng Wang (2025). Learning Evolving Tools for Large Language Models. *The Thirteenth International Conference on Learning Representations*.

- Cheng Qian, Bingxiang He, Zhong Zhuang, Jia Deng, Yujia Qin, **Xin Cong***, Zhong Zhang, Jie Zhou, Yankai Lin, Zhiyuan Liu and Maosong Sun (2025). Tell Me More! Towards Implicit User Intention Understanding of Language Model Driven Agents. *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*.

科研项目

30+ 项 在研项目　**19** 项 纵向项目　**7** 项 横向项目

4 项 人才项目　**2702.8** 万 总经费

科研成果
创新突破与学术贡献

科研项目一览

项目来源	项目类型	期限	项目金额（万元）	负责人
纵向项目				
国家重点实验室	委托项目	2025 年–2026 年	47.8	陈松蹊
国家自然科学基金	面上项目	2022 年–2025 年	51	YANG LIJIAN
国家自然科学基金	面上项目	2024 年–2027 年	43.5	邓　柯
重点研发计划（国内）	课题牵头（项目校外牵头）	2023 年–2026 年	240	邓　柯
国家自然科学基金	重点项目	2020 年–2024 年	50	邓　柯
科技部	重点研发计划（项目骨干）	2020 年–2023 年	54	邓　柯
北京市自然科学基金	重点研究专题（课题负责人）	2019 年–2023 年	120	邓　柯
全国哲学社会科学基金	重大项目（课题负责人）	2018 年–2023 年	10	邓　柯
北京市自然科学基金	非共识创新项目（储备）	2025 年–2028 年	50	李　东
国家自然科学基金	面上项目	2025 年–2028 年	42	李　东
重点研发计划（国内）	子课题	2021 年–2024 年	60	侯　琳
国家自然科学基金	面上项目	2021 年–2024 年	51	侯　琳
国家自然科学基金	面上项目	2022 年–2025 年	50	俞　声
北京市自然科学基金	非共识创新项目（储备）	2025 年–2028 年	50	刘汉中
国家自然科学基金	面上项目	2021 年–2024 年	52	刘汉中
国家自然科学基金	重大研究计划（培育）	2024 年–2026 年	70	林　乾
国家自然科学基金	面上项目	2023 年–2026 年	46	吴未迟
重点研发计划（国内）	项目牵头	2024 年–2029 年	300	杨朋昆
国家自然科学基金	青年科学基金	2022 年–2024 年	30	杨朋昆
横向项目				
企事业单位	委托项目	2025 年–2026 年	88	邓　柯
联合机构	委托项目	2024 年–2029 年	300	邓　柯
中央部委	委托项目	2023 年–2024 年	81	邓　柯
企事业单位	委托项目	2023 年–2023 年	79.9	邓　柯
企事业单位	委托项目	2025 年–2026 年	50	从　鑫
企事业单位	委托项目	2025 年–2026 年	40	俞　声
企事业单位	委托项目	2023 年–2024 年	26.6	林　乾
人才类项目				
国家自然科学基金	优秀青年科学基金项目	2024 年–2026 年	200	侯　琳
中组部	青年人才项目	2023 年–2027 年	200	俞　声
中组部	青年人才项目	2022 年–2024 年	120	林　乾
中组部	青年人才项目	2023 年–2025 年	100	杨朋昆

国家发明专利、软件著作权

序号	名称	获批时间	研发团队或发明人
1	尿失禁自动诊断方法、装置及存储介质	2023/7/25	清华大学 中国医学科学院北京协和医院 邓柯、孙智晶、朱兰、王掣
2	POP 治疗方案形成方法及系统	2023/7/25	中国医学科学院北京协和医院 清华大学 孙智晶、邓柯、朱兰、刘朝阳
3	盆腔器官脱垂自动诊断方法、装置及存储介质	2023/10/24	中国医学科学院北京协和医院 清华大学 孙智晶、朱兰、邓柯、王掣
4	基于相似图片组代表特征向量的图片检索方法及相关设备	2023/11/14	清华大学 邓柯、王海沣
5	文本切分方法及装置	2024/5/28	清华大学 邓柯、潘长在
6	单细胞基因扰动数据因果推断分析软件	2024/8/23	清华大学 付子初、侯琳
7	空间区域的同时置信曲面获取及系统	2024/09/17	清华大学 黄昆、杨立坚
8	关联性规则挖掘方法、装置及存储介质	2024/9/24	清华大学 邓柯
9	中文命名实体识别及分类方法和装置	2024/12/31	清华大学 上海起承文化发展有限公司 邓柯、潘长在、米成、陈静、李梦琦、李宜斐
10	事件相关电位信号的比较方法及系统	2025/1/24	清华大学 黄昆、杨立坚

学生获奖一览

- 马昕桐 -
第二十五届京津冀青年概率统计学术研讨会 钟家庆优秀论文奖

- 罗天派 -
北京应用统计学会学术研讨会 优秀论文奖

- 于丁一 -
2023 年联合国大数据黑客松大赛 中国赛区三等奖

- 韩庭萱 -
2023 年联合国大数据黑客松大赛 中国赛区三等奖

- 韩庭萱 -
北京生物医学统计与数据管理研究会 2024 年年会 论文特等奖

- 付子初 -
北京生物医学统计与数据管理研究会 2024 年年会 论文一等奖

- 于浩洋 -
北京生物医学统计与数据管理研究会 2024 年年会 论文二等奖

- 李弘梓 -
北京生物医学统计与数据管理研究会 2024 年年会 论文优秀奖

- 孙弘毅 -
北京生物医学统计与数据管理研究会 2024 年年会 论文优秀奖

- 余　成 -
第七届全国统计学博士研究生学术论坛 优秀论文二等奖

- 王梓涵 -
第八届全国统计学博士研究生学术论坛 优秀论文一等奖

- 张皓博 -
2025 年国际泛华统计协会中国会议 青年研究员奖

- 潘庆一 -
青年统计学家协会 2025 年年会博士生论坛 优秀论文

- 王孜睿 -
第十届全国高校研究生统计论坛 十佳论文

杨立坚教授课题组在应用概率论、时间序列、函数型数据三个方向取得了一系列重要进展

成果 01

杨立坚研究团队解决了高斯过程极值分布精确分位数的存在性、唯一性（部分解决）这两个基础问题，为基于高斯过程的各类统计推断提供了不可缺少的概率论支持。（Yang 2024 Statistics & Probability Letters; Yang 2024+ Statistics & Probability Letters）

成果 02

博士生易盈淮与合作者提出了适变趋势（mildly varying trend）概念，区别于常用的缓变趋势（slowly varying trend）；运用核和样条估计，在趋于无穷的区间上构造了自回归移动平均（ARMA）时间序列适变趋势函数的渐近同时置信带，证明了 ARMA 系数极大似然估计的默示有效性。（Yi, Song and Yang 2024+ Journal of Nonparametric Statistics)

成果 03

博士生孙爽建立了多元函数型面板数据均值函数的默示有效样条估计，构造了均值函数的各类同时置信区域，用于脑电数据的推断。（Sun and Yang 2024+ Statistica Sinica)

成果 04

课题组毕业生钟晨博士运用概率分布函数核估计，构造了自回归条件异方差（ARCH）时间序列的相合多步预测区间，建立了新息分布函数的渐近同时置信带、构造了对称性检验。（Zhong, Zhang and Yang 2024+ TEST)

成果 05

博士生胡祺睿运用样条回归解决了函数型数据方差函数的变点检测问题并构造变点的置信区间；在杨立坚指导下完成的工作论文获 2024 年国际数理统计学会（Institute of Mathematical Statistics）颁发的汉南研究生奖（IMS Hannan Graduate Student Travel Award）；与课题组毕业生李杰博士对不规则区域上的纵向图像数据均值函数建立了三角剖分默示有效双样条估计，构造了均值函数的同时置信区域，相关专利“空间区域的同时置信曲面获取及系统”已获中国知识产权局授予发明专利权。（Hu 2024 Journal of Multivariate Analysis；Hu and Li 2024 Statistica Sinica)

成果 06

博士生蔡乐衡、胡祺睿解决了函数型数据特征根置信区间和特征函数同时置信带以及个数趋于无穷的特征系统的同时推断问题；解决了不规则区域上图像数据特征根置信区间和特征函数同时置信带以及个数趋于无穷的特征系统的同时推断问题；对函数型数据中不可观测的函数型主成分，构造了概率分布函数的同时置信带。（Cai and Hu 2024 Computational Statistics & Data Analysis；Cai and Hu 2024+ Journal of Computational & Graphical Statistics；Cai and Hu 2024+ Statistica Sinica）

成果 07

博士生冯永真与合作者对于标量型响应变量和函数型协变量，提出了一种新的在连续分位数水平下进行非参数条件独立性检验的方法，并建立了基于函数型协变量随机投影的检验统计量。研究了该检验统计量在原假设和备择假设下的渐近性质；运用了多重自助法快速计算临界值，并将其用于对脑电数据的分析。（Feng, Li and Song 2024 Biometrics）

当格律诗与统计学相遇：清华数字人文团队联合研究成果被人工智能顶级会议接受

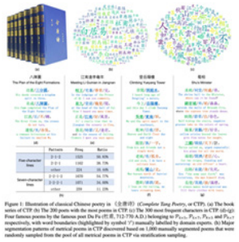


Figure 1: Illustration of classical Chinese poetry in 《全唐诗》 (*Complete Tang Poetry*, or CTP). (a) The book series of CTP. (b) The 300 poets with the most poems in CTP. (c) The 300 most frequent characters in CTP. (d)–(g) Four famous poems by the famous poet Du Fu (杜甫, 712-770 A.D.) belonging to $P_{5:0:0}$, $P_{5:0:5}$, $P_{5:5:0}$, and $P_{5:5:5}$, respectively, with word boundaries (highlighted by symbol '|') manually labelled by domain experts. (h) Major segmentation patterns of metrical poems in CTP discovered based on 1,000 manually segmented poems that were randomly sampled from the pool of all metrical poems in CTP via stratification sampling.

2023 年 10 月，2023 年自然语言处理实证方法会议（The 2023 Conference on Empirical Methods in Natural Language Processing，简称 EMNLP 2023）公布了论文接受结果，我中心邓柯老师课题组博士研究生潘长在同学的研究论文 “TopWORDS-Poetry: Simultaneous Text Segmentation and Word Discovery for Classical Chinese Poetry via Bayesian Inference” 入选 “主会长文” 单元。自然语言处理实证方法会议（EMNLP）是自然语言处理和人工智能领域的顶级会议，由国际计算语言学协会语言数据特别兴趣小组 (SIGDAT) 于 1996 年发起，是在早期超大型语料库研讨会 (Workshop on Very Large Corpora, WVLC) 基础上发展起来的。

潘长在同学的入选论文针对中国古典诗歌的文本分析展开研究，力图运用贝叶斯统计方法解决这类特殊文本的命名实体识别和分词问题。中国古代诗歌具有独特的写作风格和格律要求，富含其他通用文本中罕见的特殊诗歌词汇，其分析至今仍充满挑战。

潘长在的论文从中国古典格律诗歌主要具有两种模式的事实出发，建立了一种可以对诗歌文本分词进行有效指导的贝叶斯先验分布；并将这个先验分布与无监督中文分析方法 TopWORDS 的分析框架相结合，提出了一种同时进行文本分词和词语发现的无监督中国古代诗歌分析方法 TopWORDS-Poetry，可以借助诗歌本身隐含的格律信息对诗歌文本进行精准分词，并有效发现诗歌中的特殊词汇（如人名、地名、典故等）。

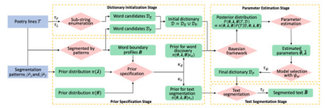


Figure 2: Pipeline of the TopWORDS-Poetry algorithm

Figure 3: Results of the Top 100 most significant discovered names (a), addresses (b), and literary allusions (c) with more than 2 Chinese characters discovered by TopWORDS-Poetry.

将该方法应用到《全唐诗》诗歌文本的分析，我们通过小样本人工验证证实了 TopWORDS-Poetry 方法能够挖掘出大量有意义的人名、地名、文学典故等词汇，并在诗歌分词上有优良的表现。相关试验还表明，TopWORDS-Poetry 方法相比于通用大语言模型 ChatGPT 和文献中其他已有方法，在综合表现上有明显的优势。

该论文是清华大学数字人文团队共同合作的成果，潘长在是第一作者，邓柯副教授作为通讯作者与清华大学中文系的李飞跃副教授共同指导了相关研究工作。清华大学数字人文团队由来自清华大学人文学院、计算机系和统计学研究中心（现统计与数据科学系）3 家单位的学者组成，凝聚了横跨文科、工科、理科的跨学科交叉研究力量，致力于运用前沿人工智能技术和统计学方法，解决新兴交叉领域 “数字人文” 中的重要问题。

该研究工作得到了国家科技部创新研究计划、清华大学国强研究院、北京市自然科学基金的支持。

- 作者团队 -

潘长在　博士
第一作者

李飞跃　副教授
共同作者

邓柯　副教授
通讯作者

邓柯副教授荣获 2022 年度教育部 “高等学校科研优秀成果奖”（科学技术类）自然科学奖一等奖

2023 年 6 月，教育部公布了 2022 年度 “高等学校科学研究优秀成果奖” 评选结果，统计学研究中心（现统计与数据科学系）邓柯副教授与清华大学自动化系生物信息学张学工教授团队联合报送的 “基因差异表达及调控的模式识别 “项目荣获自然科学奖一等奖。

项目名称：基因差异表达及调控的模式识别
项目完成人：张学工、汪小我、谢震、李衍达、古槿、邓柯


高等学校科学研究优秀成果奖
（科学技术）
证 书
项目名称：基因差异表达及调控的模式识别
奖励类别：自然科学奖
奖励等级：一等奖
获 奖 者：清华大学
2023 年 6 月 6 日
证书编号：2022-013-D01


高等学校科学研究优秀成果奖
（科学技术）
证 书
项目名称：基因差异表达及调控的模式识别
奖励类别：自然科学奖
奖励等级：一等奖
获 奖 者：邓柯
2023 年 6 月 6 日
证书编号：2022-013-R06

杨朋昆副教授获批国家重点研发计划青年科学家项目

2025 年 4 月，清华大学统计与数据科学系举办国家重点研发计划“数学和应用研究”重点专项青年科学家项目—8.1 统计学与人工智能数学基础项目启动暨实施方案论证会。会议围绕项目研究的实施方案、任务分工及具体研究内容展开研讨。项目负责人为杨朋昆副教授。

会议由南开大学王兆军教授、统计与数据科学系主管科研工作副系主任邓柯副教授联合主持，复旦大学徐文伟教授，清华大学杨立坚教授、杨宇红教授、北京大学张志华教授，清华大学李建教授，中国人民大学朱利平教授等十余位专家参与论证，共同为项目实施提供指导。清华大学科研院项目办曹立鹏老师、统计系陈松蹊教授、刘军教授出席启动会。

启动会开始，邓柯副教授首先介绍了该项目的启动情况，并代表统计系感谢南开大学、复旦大学、北京大学、人民大学等高校专家的支持。

科研院项目办的曹立鹏老师代表学校发言。曹立鹏表示，青年科学家项目的竞争异常激烈，清华大学作为项目依托单位，将全力保障项目支撑，确保经费管理、成果标注等环节符合国家要求，并预祝项目如期顺利完成。

陈松蹊教授和刘军教授分别代表院系发言。两位老师分别感谢了各位专家对项目的帮助与支持，并肯定了杨朋昆团队的努力和能力，期待未来五年团队能够取得国际领先的成果，同时，也对清华大学统计和数据科学在人工智能和大数据框架上的学术地位表示憧憬与期待。

项目组汇报环节，杨朋昆副教授作为项目负责人介绍了该项目的研究背景与目标设置、任务分解与研究内容、关键节点与实施计划、项目组织管理机制以及成果形式及测试方法等，来自西南财经大学的周岭教授、中国人民大学的张琼助理教授以及清华大学交叉信息研究院的张景昭助理教授，作为项目骨干分别介绍了三个任务的具体内容。

专家论证环节，与会专家围绕项目实施方案提出多项建议，包括加强任务协同、凝练技术成果、注重产学研结合，聚焦关键科学问题，探索新方法以适应数据科学需求等，为项目实施指明了方向。此外，与会专家还特别指出，希望团队能够依托清华平台，推动统计与计算机学科的年轻学者联合攻关，为交叉领域注入活力。

南开大学王兆军教授

复旦大学徐文伟教授

清华大学杨立坚教授

清华大学杨宇红教授

北京大学张志华教授

清华大学李建教授

与会专家合影

邓柯课题组提出双词典模型新方法，突破专业领域中文文本处理瓶颈

2025 年 6 月，清华大学统计与数据科学系邓柯副教授团队在中文文本处理研究中取得重要突破，其成果“A Dual-Dictionary Model for Mining Domain-Specific Chinese Texts”发表于《应用统计学年鉴》(Annals of Applied Statistics, AOAS)。该研究提出了一种名为 TopWORDS-MEPA（简称 TWM）的中文自然语言处理（NLP）方法，能够在仅有少量训练信息的条件下，对多种类型的专业领域中文文本进行有效处理，高质量同步完成文本分词、命名实体识别、元模式发现、关系抽取等重要任务，为解决中文文本处理难题提供了一种有效的新思路。


The Annals of Applied Statistics
2025, Vol. 19, No. 2, 1147–1166
https://doi.org/10.1214/25-AOAS2035
© Institute of Mathematical Statistics, 2025

A DUAL-DICTIONARY MODEL FOR MINING DOMAIN-SPECIFIC CHINESE TEXTS

BY JIAZE XU[1,a], CHANGZAI PAN[1,b] AND KE DENG[2,c]

一、研究背景：领域中文文本处理的挑战

中文文本处理因独特的语言结构面临无显性词汇边界、未知领域术语丰富等诸多挑战，导致传统包括大语言模型在内的众多 NLP 方法在处理专业领域文本时常常表现不稳定，时常不能准确识别不常见的专业术语，造成不准确的语义理解。

邓柯团队聚焦这一痛点，针对历史文献、医疗记录、文学作品等典型领域文本，开发了基于双字典模型（Dual-Dictionary Model, DDM）的无监督学习方法，实现了元模式发现、命名实体识别、文本分词、关系抽取等多项任务的同步高效完成。与现有方法相比，TWM 具有三大优势：

弱监督学习：无须大量标注数据，仅需少量先验知识即可启动学习；

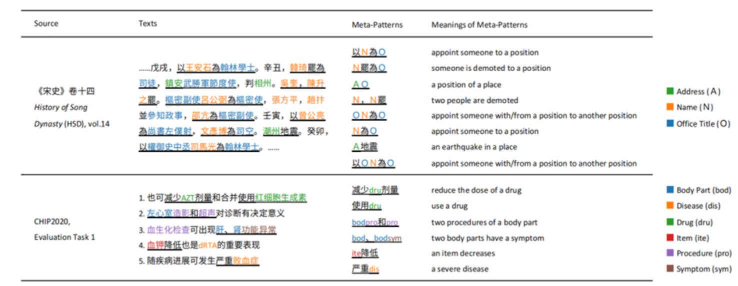
透明可解释：元模式与实体分类结果可直接解读，避免深度学习中常见的“黑箱”问题；

跨领域泛化：在历史文献、医疗文本、武侠小说等领域表现优异，展现强大适应性。

二、方法创新：从 TopWORDS 到 TopWORDS-MEPA 的“双词典升级”

TopWORDS 是邓柯团队 2016 年提出的一种无监督中文文本分析方法，通过对目标文本进行系统的扫描建立潜在词汇词典，并运用统计推断和模型选择技术对其进行深度筛选，能够在标注语料缺乏和词汇词典未知的情况下，有效实现未登录新词识别。尽管 TopWORDS 发现新词能力很强，但由于其仅实现了词汇层建模，缺乏对更高级句法结构的捕捉，语义理解能力有限，在实际应用中面临一系列局限性。

本研究提出的 TopWORDS-MEPA 方法引入的元模式字典是一大关键创新，将词语词典模型升级为“词语 - 元模式”的双词典模型。元模式字典由锚定字符（如“以”“为”等固定汉字）和词类别指标（如 N 代表人名，O 代表官职等）组成，形成类似“以 N 为 O”的混合序列，用于捕捉文本中反复出现的短语级句法和语义结构。图 1 展示了历史文献和医疗病例文本中的元模式示例。



Source	Texts	Meta-Patterns	Meanings of Meta-Patterns
《宋史》卷十四 *History of Song Dynasty* (HSD), vol.14	……戊戌，以王安石為翰林學士。辛丑，韓琦罷為司徒，鎮安武勝軍節度使，判相州。吳奎、陳升之罷。樞密副使呂公弼為樞密使，張方平，趙抃並參知政事，邵亢為樞密副使。壬寅，以曾公亮為尚書左僕射，文彥博為司空。潮州地震。癸卯，以權御史中丞司馬光為翰林學士。……	以N為O	appoint someone to a position
		N罷為O	someone is demoted to a position
		A O	a position of a place
		N、N罷	two people are demoted
		O N為O	appoint someone with/from a position to another position
		N為O	appoint someone to a position
		A地震	an earthquake in a place
		以O N為O	appoint someone with/from a position to another position
CHIP2020, Evaluation Task 1	1. 也可减少AZT剂量和合并使用红细胞生成素 2. 左心室造影和超声对诊断有决定意义 3. 血生化检查可出现肝、肾功能异常 4. 血钾降低也是dRTA的重要表现 5. 随疾病进展可发生严重败血症	减少dru剂量	reduce the dose of a drug
		使用dru	use a drug
		bodpro和pro	two procedures of a body part
		bod、bodsym	two body parts have a symptom
		ite降低	an item decreases
		严重dis	a severe disease

Address (A); Name (N); Office Title (O)
Body Part (bod); Disease (dis); Drug (dru); Item (ite); Procedure (pro); Symptom (sym)

图 1：特殊领域中文文本中的元模式示例

TWM 方法基于贝叶斯推理框架，并通过 EM 算法与扩展 Baum-Welch 算法，高效学习模型参数，同时利用弱分类器引入领域先验知识，流程图见图 2。

三、应用场景：从古籍文本到医学文本的智慧挖掘

在实际数据分析中，TWM 展现出显著优势。图 3 展示了 TWM 从不同文本数据集中自主发现的部分元模式，这些元模式为进一步的关系抽取和文本理解提供了新视角。在命名实体识别和分类方面，TWM 展现出与 LLMs（如 ChatGPT、ChatGLM）可比较的优异性能和良好的互补性。将 TWM 与 LLMs 适当结合，可以获得比二者更加优异命名实体识别和分类效果。

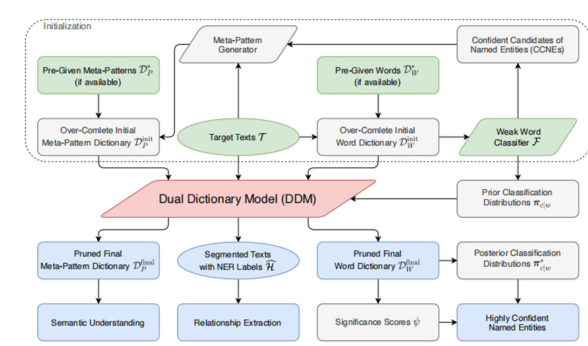


图 2：基于双词典模型的 TopWORDS-MEPA 方法流程图

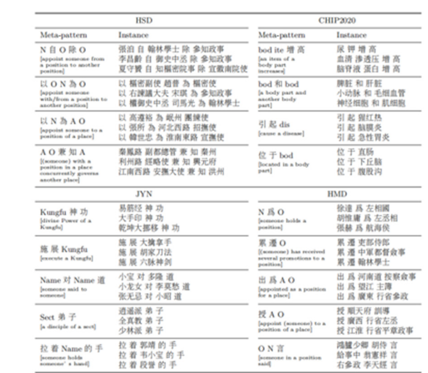


图 3：TopWORDS-MEPA 方法从《宋史》（HSD）、《明史》（HMD）、电子病历（CHIP2020）、《金庸小说全集》（JYN）中自主发现的部分元模式

- 作者团队 -

（注：徐嘉泽、潘长在为邓柯课题组已毕业博士）

徐嘉泽
第一作者

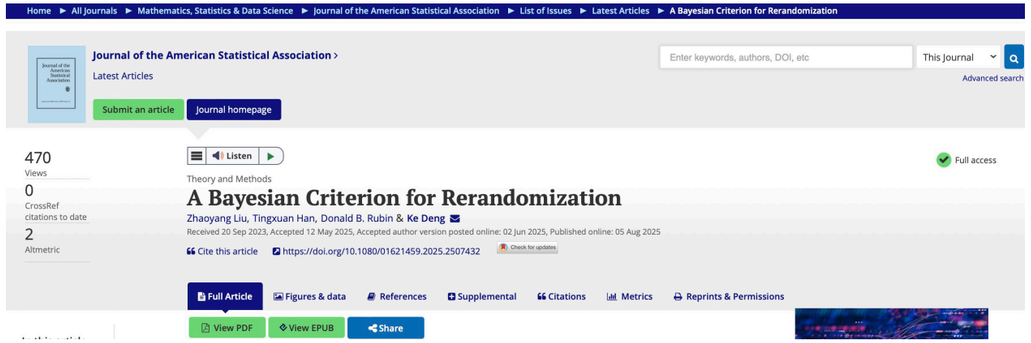
潘长在
第二作者

邓 柯
通讯作者

原文链接：https://projecteuclid.org/journals/annals-of-applied-statistics/volume-19/issue-2/A-dual-dictionary-model-for-mining-domain-specific-Chinese-texts/10.1214/25-AOAS2035.short

邓柯课题组 JASA 论文提出 rerandomization 的贝叶斯准则

2025 年 7 月，清华大学统计与数据科学系邓柯副教授课题组在统计学国际顶尖期刊 Journal of the American Statistical Association (JASA) 在线发表题为 “A Bayesian Criterion for Rerandomization” 的研究论文，提出一种基于贝叶斯的重随机化方法，可以有效提升处理效应估计的准确性。邓柯课题组 2017 级博士生刘朝阳博士和 2021 级博士生韩庭萱是论文的共同第一作者，邓柯副教授为通讯作者，与美国科学院院士 Donald B. Rubin 教授共同指导了该研究。



- 作者团队 -

刘朝阳
第一作者

韩庭萱
第一作者

Donald B. Rubin
作者

邓　柯
通讯作者

随机化试验是因果推断的金标准，因为它可以在平均意义下消除不同处理组下协变量的不平衡性。然而在一次具体的试验中，不同处理组中的协变量不平衡是一种常见现象，当协变量维数较高时这个问题尤为严重。解决这一问题的一个自然的方法是不断地尝试不同的随机化分配方案，直到得到一个协变量平衡性可接受的分配方案，即 “重随机化”（rerandomization）。重随机化方法能够使协变量取得更佳的平衡性，并且得到更高效的因果效应估计量，因而在近年获得越来越多的关注。经典的基于 Mahalanobis 距离的重随机化准则（ReM）及其系列变种通常对所有协变量施加同等约束，忽略或未能充分利用协变量重要性的信息。

本文从贝叶斯的角度出发深入研究了实施 rerandomization 的基本准则。通过将有关协变量相对重要性的认知形式化为一个先验分布，并利用它来指导重随机化过程，建立了基于贝叶斯准则的重随机化框架（ReB）。可以证明，许多基于 ReM 的重随机化过程是 ReB 取特定先验时的特例。并且，当先验信息较为准确时，通过 ReB 获得的因果效应的均值差估计量相比于 ReM 得到的估计量更准确，即有更小的渐进方差。当协变量的维数更高时，ReB 相对于 ReM 的这种优势会更加显著。

当有关协变量重要性的信息事先不可获取时，本文建议采用一种两阶段的试验设计方法来实现 ReB。在第一阶段，通过对小部分样本做完全随机化或 ReM，获取关于协变量重要性的先验信息。在第二阶段，使用第一阶段获取的先验来实施 ReB。将两阶段分别获取的因果效应估计量进行结合以获得最终的估计量。可以证明，采用这种两阶段策略来实现的 ReB 同样可以获得有关因果效应在渐近意义上更有效的估计。

本文不仅从贝叶斯角度建立了新的理论框架来理解和解释重随机化，而且提出了更有效的实施重随机化的方法。本研究中的所有理论分析都是基于设计的框架，即随机性完全来源于分配的随机性，而没有施加任何模型假设。

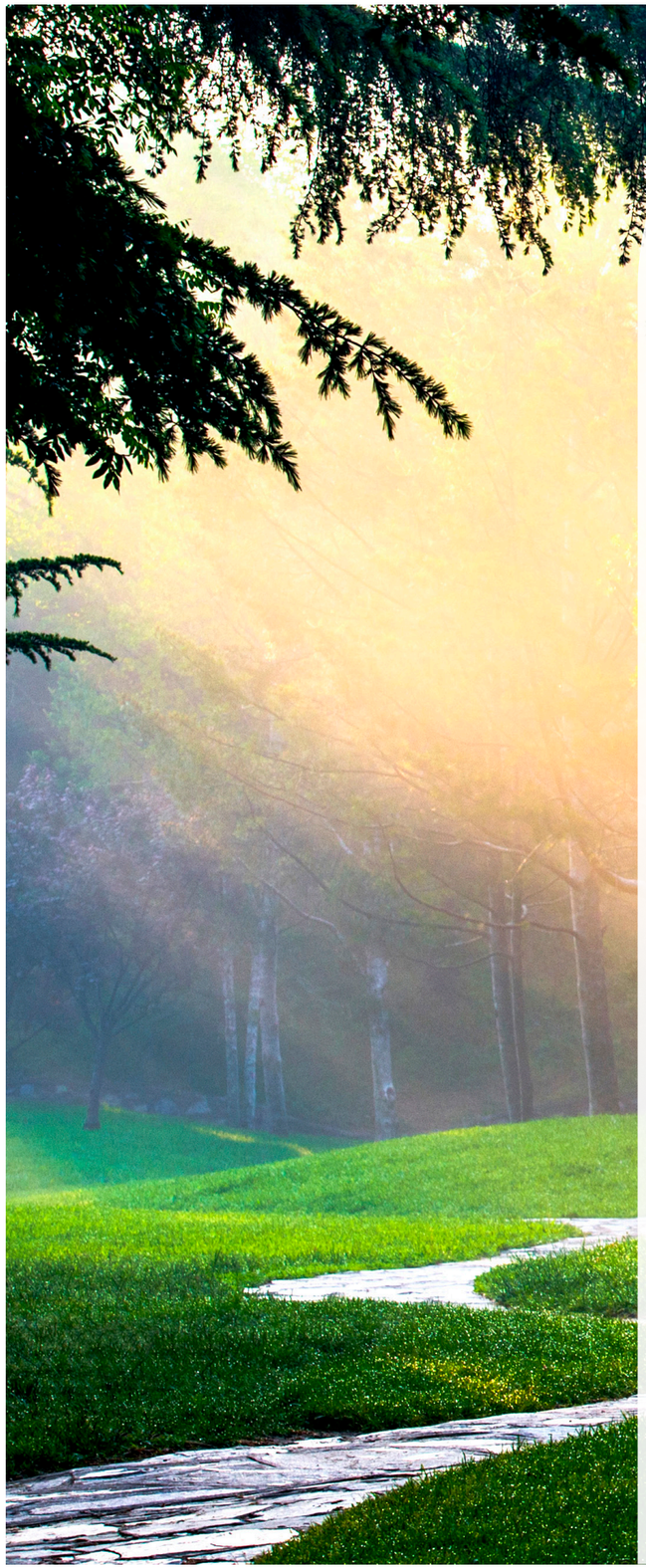
原文链接：https://www.tandfonline.com/doi/10.1080/01621459.2025.2507432

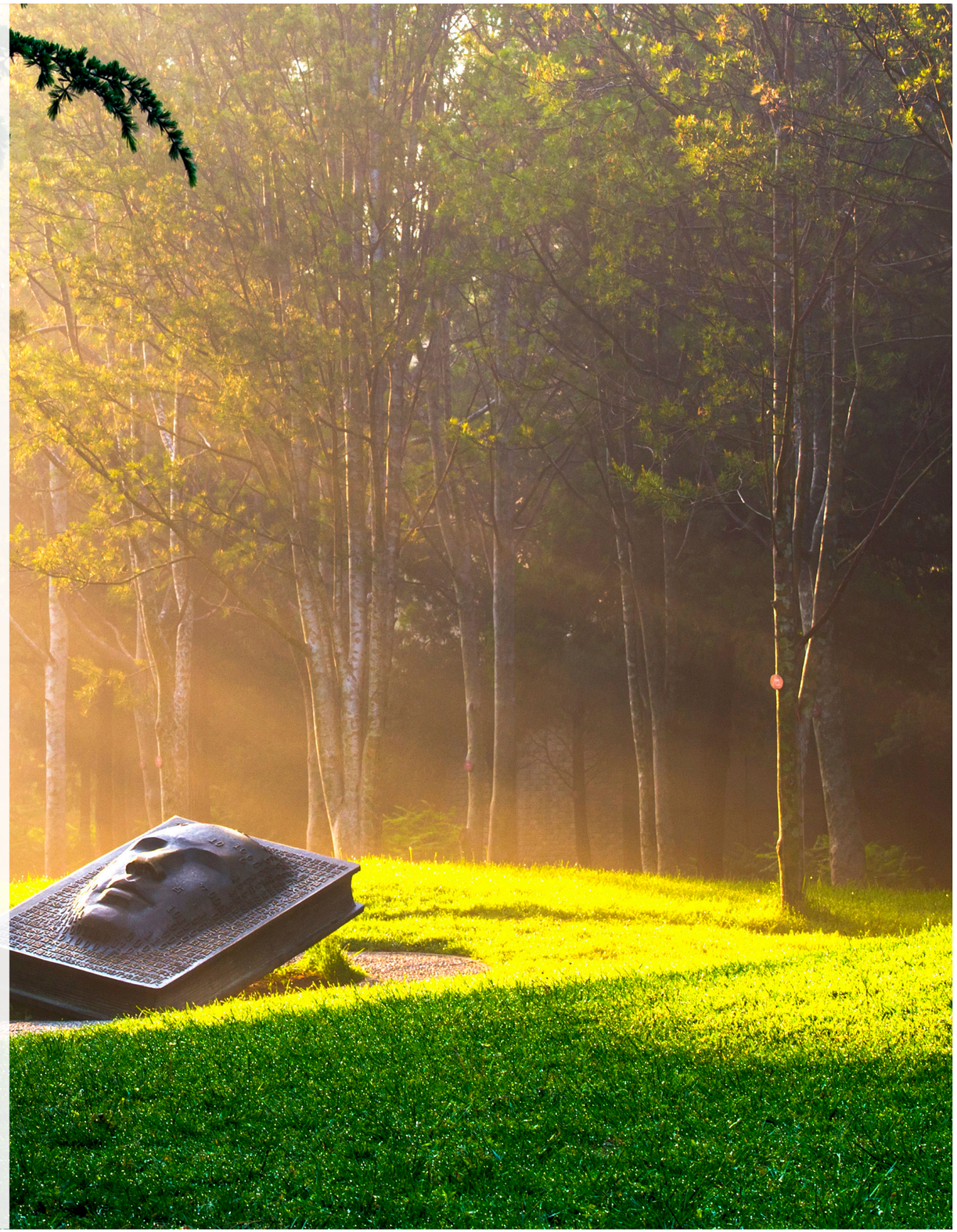
TSINGHUA UNIVERSITY
DEPARTMENT OF
STATISTICS AND DATA SCIENCE

第五部分

学术交流

全球视野与开放合作





统计系主办 2025 清华大学统计 +AI 前沿峰会

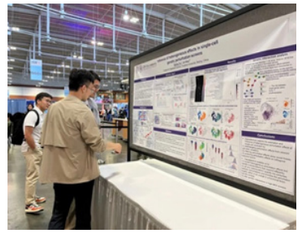
2025 年 7 月 2 日至 3 日，清华大学统计与数据科学系功举办“统计与数据科学高峰论坛”，论坛以“统计筑基探 AI 前沿，数据赋能领交叉创新”为主题，吸引了斯坦福大学、哥伦比亚大学、宾夕法尼亚大学等海内外数十位顶尖学者的参与，共同探讨统计学与人工智能的深度融合与前沿发展。会上，陈松蹊院士致辞时回顾，清华大学统计学科自 2015 年成立研究中心后，历经十年耕耘于 2024 年正式成立统计与数据科学系，这是清华面向数字时代的关键布局；筹委会主任刘军院士则在系成立一周年之际，向海内外学者发出加入邀约，共推国内统计与数据科学发展。合作单位黄大年茶思屋科技网站总编辑张群英女士代表致辞时表示，作为合作媒体，黄大年茶思屋始终致力于构建思想碰撞 - 技术融合 - 产业落地的数字桥梁，期待论坛成为学术洞察与产业需求对接的转化枢纽，共育数字时代新质生产力。

圆桌讨论环节，北京大学朱松纯教授、清华大学孙茂松教授、迈阿密大学王岚教授与华为拉格朗日数学与计算中心主任芮祥麟博士聚焦统计理论革新、算法设计突破及统计 - 人工智能交融路径三大前沿议题，展开深度思辨。斯坦福大学 Wing Hung Wong 教授、北京大学朱松纯教授、清华大学孙茂松教授、哥伦比亚大学郑甜教授、加州大学圣巴巴拉分校 Annie Qu 教授、 宾夕法尼亚大学苏炜杰副教授、迈阿密大学王岚教授和天普大学 Edoardo Airoldi 教授分别作主旨报告。此外，论坛还举办 16 场主题报告，融合专家访谈、学术沙龙等形式，深度解构统计理论与人工智能融合的前沿进展，为数据科学赋能交叉学科提供新的方法和思路。

清华统计与数据科学系亮相 JSM 2025

2025 年 8 月 2 日至 7 日，JSM 2025（The Joint Statistical Meetings）在美国田纳西州纳什维尔音乐城会议中心盛大举办。其间，清华大学统计与数据科学系学科展位及发展交流会成功举行，成为本次国际学术盛会中展现清华统计学科风采的重要窗口，为推动海外招聘及国际学术交流与合作注入了新活力。

当地时间 8 月 4 日晚，清华大学统计与数据科学系学科发展交流会如期召开。来自全球各地高校、科研机构及企业的近 200 名统计学者、校友代表齐聚一堂，共话统计与数据科学领域的前沿动态与发展机遇。交流会开场，统计与数据科学系筹委会主任、美国科学院院士刘军教授致辞，对远道而来的各位嘉宾和校友表示热烈欢迎，向长期以来关注和支持清华统计学科发展的海内外同仁表达衷心感谢。同时刘军教授着重介绍了系里为引进人才提供的各项保障，并真诚邀请海外优秀学者和青年才俊加入清华，共同探索学术前沿。交流环节中，现场气氛热烈，嘉宾们围绕学科发展方向、科研合作模式、人才培养机制等话题展开深入探讨，众多校友也分享了自身学术经历与感悟，为学科发展建言献策。



与此同时，自当地时间 8 月 3 日起至 8 月 6 日，清华大学统计与数据科学系的展位在会议中心持续开放。展位上，统计与数据科学系副系主任邓柯副教授通过展板、宣传册等形式，全面展示了系里的学科建设历程、师资力量、科研平台及学生培养成果。不少华人学者和国际同行主动前来咨询，详细了解清华统计学科建设的最新进展，以及人才引进政策等具体信息。多位海外青年学者表达合作意愿，希望能与清华开展联合研究或加入清华学术团队，为统计与数据科学事业发展贡献力量。

此外，本次会议期间，我系邓柯副教授课题组赵花丽同学荣获国际数理统计协会（Institute of Mathematical Statistics, 简称 IMS）颁发的“2025 年度汉南研究生旅行奖”（2025 Hannan Graduate Student Travel Award）。侯琳副教授课题组付子初等同学参加了海报展示等环节。

此次清华大学统计与数据科学系在 JSM 2025 期间举办的系列活动取得圆满成功。这不仅是一次充分展示清华统计学科实力与风采的机会，有效提升了学科的国际知名度和影响力，更通过搭建高水平的交流平台，让全球学界深入了解了中国统计与数据科学领域的发展成就，为引进海外优质人才、拓展国际合作渠道奠定了坚实基础。清华大学统计与数据科学系将以此次活动为契机，持续深化国际学术交流与合作，携手全球同仁共同推动统计与数据科学领域的创新发展。

我系师生参加第三届全国统计与数据科学联合会议（JCSDS 2025）

2025 年 7 月 11–13 日，盛夏杭州，太虚湖畔群贤毕至。第三届全国统计与数据科学联合会议（JCSDS 2025）在此盛大启幕，汇聚全球 21 个国家与地区的近两千名学者，共探统计与数据科学的新范式、新使命。

浙江大学副校长周江洪，清华大学讲席教授陈松蹊院士，国际数理统计学会（IMS）主席、宾夕法尼亚大学沃顿商学院讲席教授蔡天文，国际统计学会（ISI）主席、圣路易斯华盛顿大学教授何旭铭，全国工业统计学教学研究会会长、中国科学院数学与系统科学研究院研究员陈敏等嘉宾出席并致辞。

陈松蹊发言

作为全国统计与数据科学联合会议指导委员会成员，我系陈松蹊院士回顾了联合会议的发展历程，并强调了学科建设近期的突破性成果。他呼吁学界加强统计学科科普宣传，在数据时代积极回应国家需求，为数字中国建设培养更全面的数据赋能人才。

大会报告环节汇聚全球顶尖统计学者，共 6 场主旨报告由全球统计与数据科学领域的顶尖学者担纲，分别是美国国家工程院院士、香港中文大学（深圳）数据科学学院校长学勤讲座教授吴建福，中国科学院院士、北京大学讲席教授鄂维南，哈佛大学流行病学 Mitchell L. 和 Robin LaFoley Dong 讲席教授 James M. Robins，哈佛大学人口与转化数据科学 John Rock 讲席教授蔡天西，香港理工大学讲席教授黄坚，康奈尔大学管理学 Rudd Family 讲席教授 Will Cong，议题跨越人工智能、因果推断、大模型时代统计方法等前沿方向。清华大学统计师生积极参与学术活动筹备与交流，并在会议期间成功举办人才招聘宣讲会及互动展位，向全球同行集中展示清华在统计与数据科学领域的开放创新生态，并发布面向海内外高层次人才招募岗位，欢迎统计学、数据科学、AI 与机器学习及相关交叉学科优秀人才加盟。

专题研讨环节以“邀请报告 + 贡献报告 + 圆桌论坛”等方式，汇聚全球顶尖学者与青年才俊，围绕经济统计与计量经济、医学与生物统计、机器学习与人工智能及跨学科融合四大主轴展开深度对话。157 场主题研讨、631 项前沿报告，从高维因果推断、空间多重组学数据建模，到金融计量中的深度学习、大规模缺失数据的

不确定性量化，系统呈现了当代统计与数据科学在理论创新、方法与突破。我系多名师生作学术报告交流：

邓柯副教授：在“因果推断”主题会场作《Imputation-Based Randomization Tests for Randomized Experiments with Interference》特邀报告；

李东副教授：在“金融统计及其应用”主题会场作《Large covariance matrix estimation with factor-assisted variable clustering》特邀报告；

侯琳副教授：在“单细胞数据分析的统计模型与机器学习方法“主题会场作《细胞通讯网络的统计推断》特邀报告；

林乾副教授：在“统计与机器学习”主题会场作《Supporting Evidences for the Adaptive Feature Approach》特邀报告；

杨朋昆副教授：在“统计与机器学习”主题会场作《Learning with Shared Representations: Statistical Rates and Optimal Algorithms》特邀报告；

王健桥助理教授：在“大规模遗传与健康数据分析的新方法”主题会场作《Heritability estimation with similarity representation》特邀报告；

余成博士：在“因果推断在公平与稳健决策中的最新进展”主题会场作《Individualized Inference for Causal Fairness through Conformal Mediation Analysis》贡献报告；

卢鑫博士：在“因果推断在公平与稳健决策中的最新进展”主题会场作《Rerandomization for covariate balance mitigates p-hacking in regression adjustment》贡献报告；

赵花丽同学：在“针对生物医学因果推断与高维分析的前沿方法”主题会场作《A Debiased High-dimensional Regression Calibration Method for Errors-in-variables Logcontrast Models》贡献报告；

孙弘毅同学：在“临床决策与脑疾病研究的进阶统计学习方法”主题会场作《Revisiting the Phenolog Prediction Problem》贡献报告；

于浩洋同学：在“数字经济与网络结构弹性建模的创新方法”主题会场作《Minimax Optimal Design with Spillover and Carryover Effects》贡献报告。

作为国内统计与数据科学领域规模最大、影响最广的年度会议，本届盛会标志着中国统计与数据科学领域国际化与交叉融合迈上新台阶。

部分参会师生风采集锦

统计系师生参加北京生物医学统计与数据管理研究会 2024 年年会暨第九届学术交流研讨会

2025 年 1 月，北京生物医学统计与数据管理研究会（简称“BBA”）2024 年年会暨第九届学术交流研讨会在中国医学科学院阜外医院成功举办。会上，多位专家学者发表了精彩报告，其中，我系讲席教授陈松蹊院士作了题为《脑电数据的自动标记与统计分析》的专题报告，我系邓柯副教授作了题为《统计学助力生物医学研究与实践》的学术分享。

陈松蹊院士报告题目为《脑电数据的自动标记与统计分析》，重点介绍了癫痫发作检测中小样本学习集合方法的应用，通过数据与临床知识的融合显著减少医生手动标注的时间；脑电 /MEG 源成像中 SCICA 伪差去除技术和基于变点检验的分段平稳模型在脑内活动量化中的作用，以及引入纠偏估计提升源成像精度的创新；此外，还对 EEG-X 平台的功能亮点进行了剖析，包括跨平台协作支持、多模态脑电分析、自动 AI 标注（如发作间期放电识别、SEEG 电极标注）及 QEEG 指标计算，充分展现了智能化脑电分析技术的前沿进展与实际应用价值。

邓柯副教授就《统计学助力生物医学研究与实践》介绍了近期在生物医学研究领域的工作基础、研究成果和相关进展。

此外，会议同时为 BBA 第五届青年优秀论文获奖者颁奖。我系 2021 级博士研究生韩庭萱荣获特等奖，我系 2021 级博士研究生付子初荣获一等奖，我系 2022 级博士研究生于浩洋、数学系博士研究生赵花丽荣获二等奖，我系 2021 级博士研究生李弘梓、2022 级博士研究生孙弘毅荣获优秀奖。

陈松蹊 院士

邓柯 副教授

北京生物医学统计与数据管理研究会（BBA）是北京市一级学会，成立于 2015 年 1 月 30 日。BBA 致力于开展生物医学统计与数据管理等学术研讨活动，促进生物医学统计与数据管理学科的繁荣和发展，促进生物医学统计与数据管理人才的成长和提高，促进生物医学统计与数据管理理论与方法在医药卫生领域中的应用。

近年来，我系师生深度参与研究会主办的系列学术活动。多名教师应邀作专题报告，学生群体则在研究会年度论文评选中多次获奖。

医学院与统计与数据科学系座谈会举行

2025 年 9 月 8 日，清华大学医学院与统计与数据科学系座谈交流会举办。国际著名统计学家、美国国家科学院院士刘军，副教务长、医学院院长黄天荫参加座谈会。会议由医学院常务副院长祁海主持。

座谈会现场

黄天荫表示，在医学人才培养方面，医学院鼓励学生积极开展理工科与社科的跨学科研究，推动卓越医师科学家项目发展，为学生提供跨学科学习机会与系统全面的成长路径。在科研平台建设方面，医学院将在医学园区规划中统筹教学、科研和临床试验等各类平台建设。通过建设覆盖全社区的平台中心，实现附属医院间数据互联互通，打造学术–临床一体化体系，促进教学、科研与临床的深度融合，为培养兼具医学与数据科学素养的复合型人才提供有力支撑。

刘军介绍了数据科学系现有师资力量、学科发展历程、学科架构等基本情况，以及学院重点依托其学科交叉属性在人才培养方面的优势与特色。学院十分注重复合型人才的培养，希望培养具备坚实的数理基础和统计思维并掌握数据科学方法的统计学人才。学院积极推动交叉研究，与生物医学领域合作十分密切，希望双方通过共建“数据科学交叉研究院”等研究平台，实现校内科研、教学与临床实践的深度融合。

祁海介绍了医学院近期发展情况。医学院依托清华大学多学科交叉优势，以“培养卓越医师科学家”为使命，积极探索医学教育、科研与临床的深度融合。统计与数据科学系副主任邓柯、侯琳介绍了统计系在医学影像、多组学等领域的最新科研进展。

在讨论环节，双方认为应充分发挥清华大学的顶尖学科优势，推动统计方法与医学应用深度融合。未来将集中资源推动数据共享与跨学科科研，聚焦医学影像、文本分析、遗传学等领域，进一步整合临床数据推动清华附属医院在肝胆外科、心脏外科、妇产科、骨科以及罕见病等领域研究。

双方将发挥各自优势，加强清华人工智能医院对电子病历系统（EMR）成果的应用，推动病历数据系统化，为临床决策和医学研究提供支持。通过正在筹建的“数据科学交叉研究院”，双方将致力于建设高水平医疗大数据与人工智能创新平台，促进跨学科联合攻关，进一步对接临床数据，提升临床医生数据能力与科研水平。

此次座谈进一步深化了医学院与统计与数据科学系的交流合作，加强学校各院系与医学院、附属医院之间的沟通与联系，充分发挥学校交叉学科优势，推动跨学科合作与医疗创新持续发展，为医学院在医疗数据和人工智能领域的进一步发展奠定良好基础。

学术活动

时间	主讲人	工作单位及职称	报告题目
2023.07.07	胡天阳	华为技术有限公司博士	Towards a Statistical Understanding of Neural Network Classifiers
2023.07.31	蒋继明	加州大学戴维斯分校教授	Pseudo-Bayesian Classified Mixed Model Prediction
2023.08.07	冷琛雷	华威大学教授	A Two-way Heterogeneity Model for Dynamic Networks
2023.09.06	王文佳	香港科技大学（广州）助理教授	Random Smoothing Regularization in Kernel Gradient Descent Learning
2023.09.18	张世华	中科院数学与系统科学研究院研究员	Intelligent Spatial Transcriptomics: Methods and Applications
2023.09.20	丁　亮	复旦大学研究员	Kernel Packet: An Exact and Scalable Algorithm for Gaussian Process Regression with Matérn Correlations
2023.09.21	张正武	北卡罗莱纳大学教堂山分校助理教授	Modeling Human Brain Connectivity: From Discrete Networks to Continuous Functions
2023.09.25	Fabrizio Ruggeri	意大利国家研究理事会（National Research Council）教授	Unsupervised Statistical Tools for Anomaly Detection: The Case of Healthcare Frauds
2023.10.09	许晶晶	字节跳动研究员	其于代码大模型的代码智能体
2023.10.16	杨宇红	清华大学教授	Profile Electoral College Cross-Validation
2023.10.18	吕绍高	南京审计大学教授	Robust Structure Learning and L_p-Regularization for Graph Neural Networks
2023.10.28	林丹瑜	北卡罗来纳大学教堂山分校教授	Research and Training in Biostatistics
2023.11.06	曾　靖	中国科学技术大学副教授	Robust Sliced Inverse Regression: Optimal Estimation for Heavy-Tailed Data in High Dimensions
2023.11.09	Ruodu Wang	滑铁卢大学教授	E-backtesting
2023.11.09	吴　磊	北京大学助理教授	Understanding the Implicit Regularization of Stochastic Gradient Descent: A Dynamical Stability Perspective
2023.11.23	官永涛	香港中文大学（深圳）教授	Group Network Hawkes Process
2023.11.22	何　勇	山东大学教授	Matrix Kendall's tau in High-dimensions: with Applications to Matrix Factor Model and 2-Dimensional (sparse) Principal Component Analysis
2023.11.23	郑智超	新加坡管理大学副教授	Estimating Patient Health Transition from Data Censored by Treatment-Effect-Based Policies
2023.12.13	王振富	北京大学助理教授	Mean Field Limit for Large Systems of Interacting Particle Systems and its Applications
2023.12.18	杨　剑	西湖大学教授	Mapping Genes for Complex Human Traits and Diseases
2024.01.10	Jun Yang	哥本哈根大学助理教授	Stereographic Markov Chain Monte Carlo
2024.03.18	张　敏	清华大学教授	Semiparametric Causal Inference Methods for Binary Outcomes Subject to Censoring
2024.04.02	姚志刚	新加坡国立大学副教授	Principal Flow, Sub-Manifold and Boundary
2024.04.26	Ying Chen	新加坡国立大学副教授	Neural Tangent Kernel in Implied Volatility Forecasting: A Nonlinear Functional Autoregression Approach
2024.05.27	Danyu Lin	北卡罗来纳大学教堂山分校教授	Evaluating the Effectiveness of COVID-19 Vaccines
2024.05.20	张正军	中国科学院大学教授	基于汉密顿聚类刻画当代非对称因果假说
2024.05.30	徐加明	杜克大学副教授	Recent Advances on Random Graph Matching

时间	主讲人	工作单位及职称	报告题目
2024.06.03	张心雨	爱荷华大学博士	Spectral Change Point Estimation for High Dimensional Time Series by Sparse Tensor Decomposition
2024.06.17	陶　然	范德堡大学副教授	Efficient Designs and Analysis of Two-phase Studies with Longitudinal Binary Data
2024.06.20	Jun Yu	澳门大学教授	On the Spectral Density of Fractional Ornstein-Uhlenbeck Processes
2024.06.24	Judy Wang	乔治华盛顿大学教授	Conformal Prediction in Non-Exchangeable Data Contexts
2024.07.01	Edoardo M. Airoldi	天普大学教授	Designing Experiments on Social, Healthcare and Information Networks
2024.07.16	谢　瑶	佐治亚理工学院教授	Generative Models for Statistical Inference
2024.07.18	Hongzhe Lee	宾夕法尼亚大学教授	Regressing Multivariate Gaussian Distribution on Vector Covariates for Co-expression Network Analysis
2024.07.22	陈　嵘	罗格斯大学教授	Kronecker Product Approximation for Matrix Approximation, Denoising and completion
2024.07.22	Dennis Lin	普渡大学教授	AI, BI & SI-Artificial, Biological and Statistical Intelligences
2024.09.27	J. S. Marron	北卡罗来纳大学教堂山分校教授	Data Integration Via Analysis of Subspaces (DIVAS)
2024.10.11	汤家豪	清华大学杰出访问教授	Threshold Models in Time Series Analysis: 40 Odd Years On
2024.10.14	李　凡	杜克大学教授	Covariate Adjustment in Randomized Experiments with Missing Outcomes and Covariates
2024.10.21	Tailen Hsing	密歇根大学教授	A Functional-Data Perspective in Spatial Data Analysis
2024.10.28	张　鹏	清华大学自动化系博士	Cross-Cutting Research of Traditional Medicine from the Perspective of Big Data and AI

时间	主讲人	工作单位及职称	报告题目
2024.11.04	秦昭晖	埃默里大学教授	CryptoGWAS-Running GWAS Without a Deterministic Phenotype
2024.11.11	胡懿娟	北京大学教授	Statistical Inference of Microbial Networks
2024.11.18	于天维	香港中文大学（深圳）教授	Adaptive Graph Diffusion for Meta-Dimension Reduction
2024.11.26	Dianbo Liu	新加坡国立大学助理教授	Optimization of Discrete Optimization in Machine Learning
2024.12.09	刘振亚	中国人民大学教授	Estimation and Inference on the State-Varying FAVAR Model
2024.12.16	王玮宁	荷兰格罗宁根大学教授	A Quasi-Bayesian Approach
2024.12.24	杨　帆	清华大学教授	Identifiability of Causal Effects with Data Missing Not at Random
2024.12.30	姚　远	香港科技大学教授	Controlling the False Discovery Rate in Transformational Sparsity: Split Knockoffs
2025.01.06	郭绍俊	中国人民大学副教授	Empirical Likelihood-based Inference in Nonparametric Regression and Regression Discontinuity Designs: Bias Correction and Wilks Phenomenon
2025.01.14	张云舒	宾夕法尼亚大学博士	Double Sensitivity Analysis in Causal Inference with Unmeasured Confounding and Informative Sampling
2025.02.19	许进超	阿卜杜拉国王科技大学教授	Achieving Optimal Approximation Rate of Nonlinear Neural Networks through Linearization with Fixed Neurons
2025.03.03	潘建新	北京师范大学 - 香港浸会大学联合国际学院教授	Regression Models for Spatially Correlated Binary Data
2025.03.17	姚琦伟	伦敦政治经济学院教授	Autoregressive Networks and Stylized Features
2025.03.21	黄　坚	香港理工大学教授	Continuous Normalizing Flow for Learning Probability Distributions

时间	主讲人	工作单位及职称	报告题目
2025.03.24	黄东明	新加坡国立大学助理教授	Sliced Inverse Regression with Large Structural Dimensions
2025.04.21	李洪哲	宾夕法尼亚大学教授	AI-Ready Data and Advanced Data Science for Precision Health
2025.04.21	BUTUCEA	法国国立统计与经济管理学院教授	Nonparametric Inference under Local Differential Privacy
2025.04.28	周　舟	多伦多大学教授	Wasserstein and Convex Gaussian Approximations for Non-stationary Time Series of Diverging Dimensionality with Applications
2025.05.9	邱怡轩	上海财经大学副教授	Deep Sparse Masks via Optimal Transport
2025.05.12	林毓聪	北京理工大学副教授	Multimodal Data Fusion for the Auxiliary Diagnosis and Treatment of Complex Liver Diseases
2025.06.23	万　林	中国科学院大学教授	Learning Collective Multicellular Dynamics from Time-Series scRNA-seq Data
2025.07.04	陈　豪	加州大学戴维斯分校教授	Two-Sample Hypothesis Testing for High-Dimensional and Non-Euclidean Data
2025.07.04	廖振宇	华中科技大学副研究员	A Random Matrix Approach to Neural Networks: From Linear to Nonlinear, and From Shallow to Deep
2025.07.07	高子珺	南加州大学马歇尔商学院助理教授	Powerful Randomization Tests for Subgroup Analysis
2025.07.14	任之湄	宾夕法尼亚大学助理教授	ACS: An Interactive Framework for Conformal Selection
2025.08.15	郭心舟	香港科技大学助理教授	In-Sample Evaluation of Subgroups Identified by Generic Machine Learning
2025.08.21	申舒婷	新加坡国立大学助理教授	Optimal Assortment Inference within an Online Learning Framework
2025.09.01	Mu Niu	格拉斯哥大学副教授	Intrinsic Nonparametric Regression on Complex Domains and High-Dimensional Point Cloud

学术社区服务

陈松蹊院士出任《中国科学：数学》编委并荣任新刊《统计学习与数据科学》主编

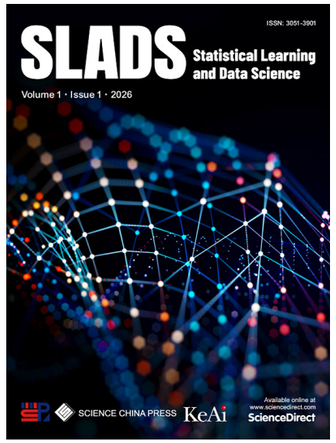
中国科学院院士、清华大学统计与数据科学系讲席教授陈松蹊学术任职再添重要篇章。陈松蹊教授自2023年起担任《中国科学：数学》编委后，2025年10月10日正式出任新创刊的国际学术期刊《统计学习与数据科学》（Statistical Learning and Data Science，简称SLADS）主编。

《中国科学：数学》由中国科学院主管，中国科学院和国家自然科学基金委员会共同主办，是涵盖基础数学、应用数学、计算数学与统计学等领域的综合性权威刊物。陈松蹊教授凭借其在《The Annals of Statistics》《Journal of the American Statistical Association》等国际顶级期刊的丰富任职经验，为该刊在统计理论、数据科学交叉方向的稿件质量把控与学术视野拓展提供重要支撑，助力中国数学期刊建设迈向新高度。

2025年10月全新英文学术期刊《统计学习与数据科学》正式创刊。该刊采用完全开放获取模式，并全额减免文章处理费，实现“作者无忧发表，读者免费阅读”。期刊引入OpenReview平台，推行透明同行评审，所有评审意见、作者回应及编辑决策将在文章接受后全部公开。

该刊设有三大核心板块：统计学、机器学习与人工智能、数据科学，覆盖数据驱动的前沿研究领域。期刊承诺投稿后7日内完成初审，3.5个月内完成终审决定，录用后未经排版版本立即在线发布，最终接受后7天内完成在线正式出版。

期刊顾问委员会云集国际顶尖学者，包括David Donoho、范剑青、Michael I. Jordan、刘军、王永雄等著名科学家，为新刊的学术质量与国际化发展提供强力保障。



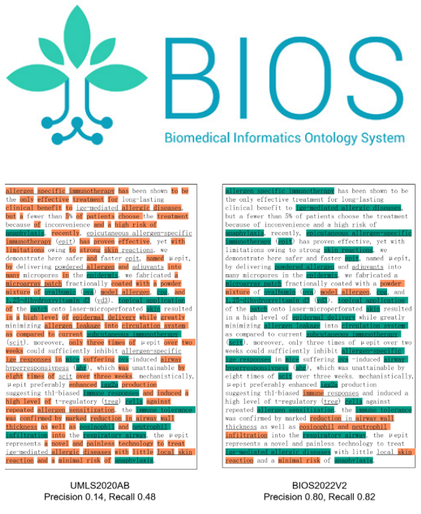
俞声副教授受任 JASA 副主编

清华大学统计与数据科学系俞声副教授受任《美国统计学会会刊》（Journal of the American Statistical Association，简称 JASA）副主编（Associate Editor），任期从 2024 年 9 月起。

《美国统计学会会刊》创立于 1888 年，按季度出版。该刊为统计科学领域的顶级期刊。所刊文章主要关注统计在经济学、社会科学、生物学、物理学、工程学、健康科学以及人文学科等各个学科中的应用、理论和方法。

俞声，清华大学统计与数据科学系长聘副教授。研究方向为医学文本类智能，包括自然语言处理、大型语言模型、知识图谱、搜索引擎、电子病历分析等。俞声开发的电子病历自然语言处理系统 NILE 被 10 个国家和地区的医学研究机构使用。俞声与哈佛大学合作发明的高通量表型提取技术，使疾病表型识别算法开发速度从每年 1-2 种疾病提高到每年超过 1000 种疾病，并应用于“Million Veteran Program”等美国国家级精准医学研究项目以及 Mass General Brigham 等医院的生物样本库、科研患者注册库建设。该系列论文获评医学信息学顶刊 JAMIA 的编辑选择奖、国际医学信息学学会 2019 年年鉴最佳论文奖，并按标准化生物医学实验方法发表于 Nature Protocols。2015 年，俞声入职清华大学统计学研究中心（现统计与数据科学系前身），获多项国家基金和国家高层次人才特殊支持计划青年拔尖人才项目支持，带领团队围绕中英文电子病历和智能诊疗开发了自动化知识图谱构建、电子病历分析、生物医学机器翻译、临床诊断决策支持等一系列技术。

2021 年起，俞声课题组与粤港澳大湾区数字经济研究院（IDEA 研究院）合作，主持开发了拥有 2210 万概念、4602 万中英文术语、9985 万关系三元组的生物医学信息学本体系统 BIOS。该系统 2022 年成为世界最大的原生单体生物医学知识图谱，体量达到美国国立卫生院国家医学图书馆自 1986 年起开始建造的一体化医学语言系统（UMLS）的数倍，为我国医疗行业大数据处理与人工智能开发建立了公共基础。



BIOS 与 UMLS 术语覆盖度和质量对比

在 BIOS 知识图谱的基础上，俞声课题组进一步训练开发了电子病历结构化专用大模型，可一次性输出电子病历中的全部医学实体及相应的语义类型、叙述状态、身体部位、数值、单位、修饰语、状态、目的等属性，使医学自然语言处理更加智能化、一体化，服务医疗信息化建设，加速医学研究样本筛选与数据获取，是大模型技术在医疗垂直领域的独特应用。

电子病历结构化大模型功能示意

"query": "Unit No: [Unit No]\n\nAdmission Date: [Admission Date]\n\nDischarge Date: [Discharge Date]\n\nDate of Birth: [Date of Birth]\n\nSex: M\n\nService: MED\n\nAllergies: Food Extracts\n\nAttending: [Attending]\n\nChief Complaint: as above\n\nMajor Surgical or Invasive Procedure:\nCardioversion times two ([Cardioversion 1] and [Cardioversion 2])\nAV node ablation ([AV node ablation])\nPacemaker placement ([Pacemaker placement])\n\nBrief Hospital Course: influenza vaccine was administered to the patient prior to discharge.\n\nDischarge Disposition: Extended Care\n\nFacility: [Facility for the Aged - [Facility]]\n\nDischarge Diagnosis:\nAtrial Fibrillation\nAcute Renal Failure\nCongestive Heart Failure\nPneumonia\nChronic subdural hematoma\n\nDischarge Condition: Stable and improved.\n\nDischarge Instructions:\nPlease take all medications as prescribed.\nPlease check your weight daily. If your weight increases by more than 3 pounds, contact your primary care physician."

"output_0821_model": "[{"phrase": "allergies", "semantic_type": "Disease, Syndrome or Pathologic Function", "assertion_status": "title", "body_location": "null", "modifier": "null", "value": "not applicable", "unit": "not applicable", "purpose": "not applicable"}, {"phrase": "food extracts", "semantic_type": "Chemical or Drug", "assertion_status": "present", "body_location": "not applicable", "modifier": "not applicable", "value": "null", "unit": "not applicable", "purpose": "null"}, {"phrase": "surgical or invasive procedure", "semantic_type": "Therapeutic or Preventive Procedure", "assertion_status": "title", "body_location": "null", "modifier": "not applicable", "value": "not applicable", "unit": "not applicable", "purpose": "null"}, {"phrase": "cardioversion times", "semantic_type": "Therapeutic or Preventive Procedure", "assertion_status": "present", "body_location": "null", "modifier": "not applicable", "value": "not applicable", "unit": "not applicable", "purpose": "null"}, {"phrase": "cardioversion", "semantic_type": "Therapeutic or Preventive Procedure", "assertion_status": "present", "body_location": "null", "modifier": "not applicable", "value": "not applicable", "unit": "not applicable", "purpose": "treat cardiac arrhythmia"}, {"phrase": "cardioversion", "semantic_type": "Therapeutic or Preventive Procedure", "assertion_status": "present", "body_location": "null", "modifier": "not applicable", "value": "not applicable", "unit": "not applicable", "purpose": "treat cardiac arrhythmia"}, {"phrase": "av node ablation", "semantic_type": "Therapeutic or Preventive Procedure", "assertion_status": "present", "body_location": "null", "modifier": "not applicable", "value": "not applicable", "unit": "not applicable", "purpose": "treat AV node dysfunction"},…

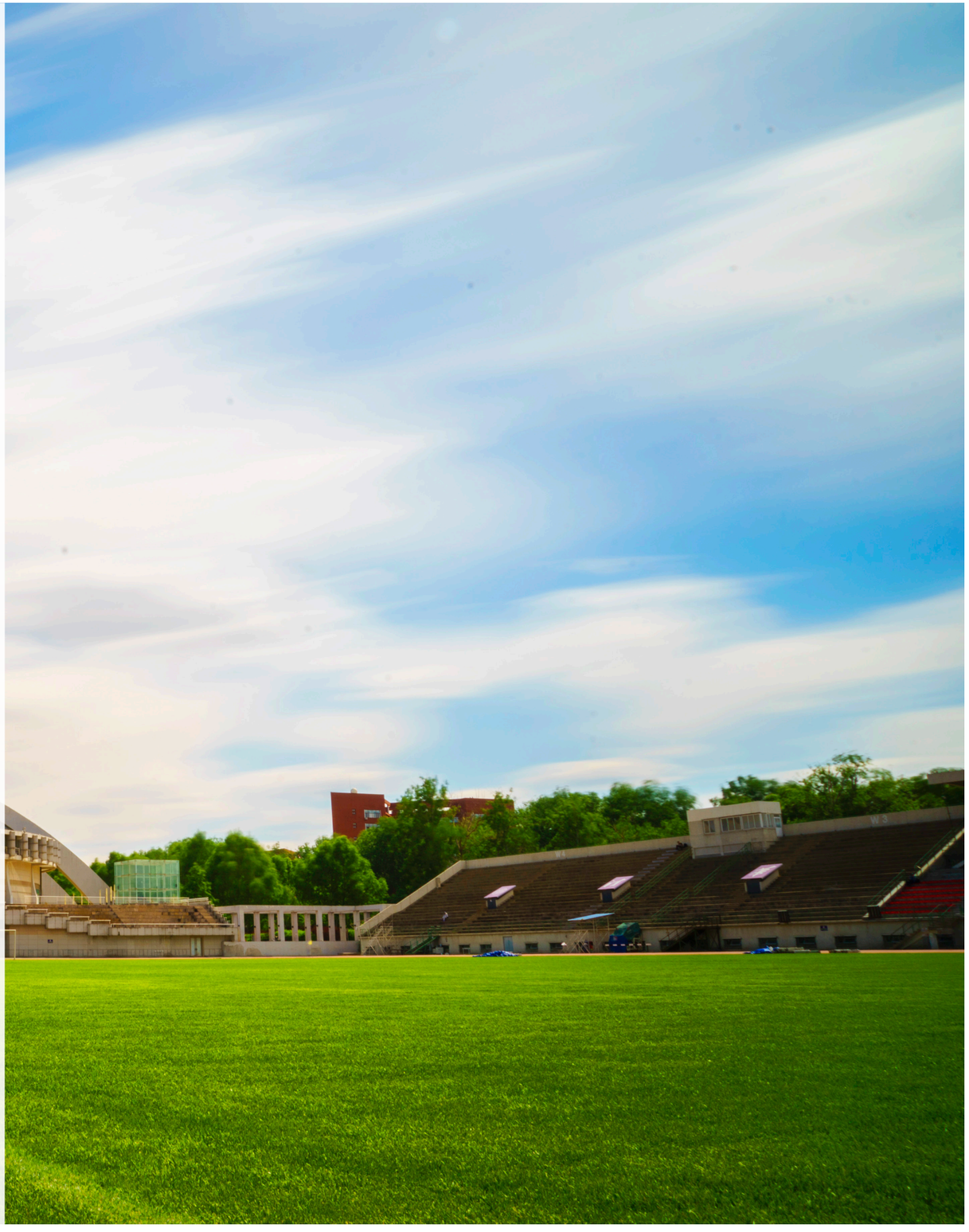
大模型实际输出示意

TSINGHUA UNIVERSITY
DEPARTMENT OF
STATISTICS AND DATA SCIENCE

第六部分

社会影响

服务国家·贡献世界



解读大气污染：我国迎来后疫情时期空气质量大范围改善

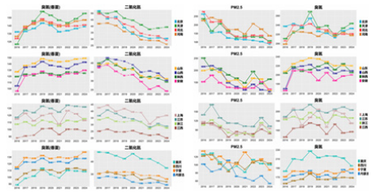
2025 年 3 月，清华大学统计与数据科学系、北京大学统计科学中心、西南财经大学统计交叉创新研究院、江西财经大学财经数据科学重点实验室联合发布《空气质量评估报告（十二）》。研究团队从 2014 年开始以年报的频率追踪中国大气污染高暴露区空气质量变化，2025 年评估范围已覆盖全国十六省 155 城市，反映全国 57.2% 的人口暴露情况。本年度首次采用结构化统计决策与大语言模型润色相结合的方法，实现报告主体的自动生成。

一、报告内容摘要

（一）后疫情期首现多数污染物大范围改善

在去年发布《空气质量评估报告（十一）》，研究团队警示了后疫情时期春夏臭氧和颗粒物污染抬头的趋势。时隔一年，我们迎来了一个阶段性的好消息：空气污染在疫情结束后首次呈现大范围改善特征。除臭氧外，其他污染物年平均浓度均实现了 5% 以上的同比降幅。秋冬 PM2.5 污染治理也取得明显进展，155 个城市中有 128 个数值较 2023 年同期下降。

在浓度整体下降的背景下，污染的“热点区”格局稳定：河南 PM2.5 居高不下，宁夏保持 PM10 和二氧化硫浓度双高态势；天津、上海、重庆三个直辖市的二氧化氮浓度居十六省市前列，遥遥领先各自周边省市；一氧化碳浓度极值依然出现在山西与重庆。



臭氧及其前体物二氧化氮的浓度趋势变化

近年各省 PM2.5 和臭氧年 90% 分位数浓度最大值变化趋势

（二）PM2.5 尚未退场 臭氧挤占 C 位

在多数污染物协同下降背景下，臭氧却持续上升。2024 年，155 个城市中超过 60% 春夏臭氧浓度比上年反弹，平均浓度呈现了连续三年升高的态势，且同比升幅较近 6 年均值进一步扩大，仅 40 个城市春夏臭氧浓度低于 2018 年水平。与 2023 年相比，10 个省市春夏臭氧浓度反弹，宁夏、浙江上升 5% 以上；安徽、河南、北京、山东连续三年反弹。目前，春夏臭氧浓度高值区集中在天津、山东、河南，三省（市）春夏臭氧平均浓度超过世界卫生组织准则值 2 倍有余；而邻近的河北只有衡水和邢台两市达到同等污染强度。相比之下，上海市臭氧治理成绩较为亮眼，近两年春夏臭氧浓度持续下降，累计降幅 10.4%。从峰值浓度来看，近年来，十六省市中只有重庆、山东、河北臭氧极端污染呈持续改善趋势，高浓度臭氧日的健康风险不容忽视。

（三）“十四五”收官 多省空气质量宣告阶段性“毕业”

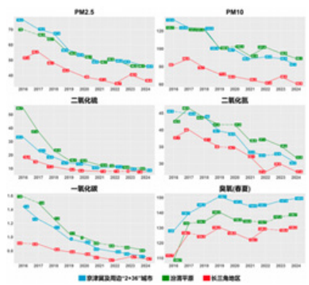
2025 年是“十四五”的收官之年，也是国务院《空气质量持续改善行动计划》和生态环境部《臭氧污染防治攻坚行动方案》制定空气质量改善目标的关键节点。报告基于 2024 年数据对十六省市“十四五”空气质量改善目标的完成情况进行了评估，内蒙古八市、江苏和四川省已提前完成目标；基于 2020 年至 2024 年平均降幅推算，上海、北京和山东预计能够在 2025 年完成目标，山西、浙江、江西、河北、天津、安徽、陕西、宁夏、重庆和河南今年须加速度，以超过 2020-2024 年平均降速的水平削减方有希望达标，其中陕西、宁夏、河南、重庆达标形势严峻，需要尽快调整策略，寻求有效的治理方案。

省市	2025 年 $PM_{2.5}$ 浓度目标
北京	控制在 32 微克/立方米以内
天津	控制在 37 微克/立方米以内
河北	比 2020 年下降 20%
河南	低于 42.5 微克/立方米
山西	降至 39 微克/立方米以下
山东	达到 38 微克/立方米
陕西	不超过 35 微克/立方米
宁夏	控制在 30.5 微克/立方米以内
内蒙古	比 2020 年下降 7.1%
上海	控制在 30 微克/立方米以下
江苏	总体达标，设区市比 2020 年下降 10%
浙江	达到 24.3 微克/立方米
安徽	控制在 35 微克/立方米以下
江西	控制在 24.8 微克/立方米以内
四川	除宜宾市、自贡市外其余城市实现达标
重庆	下降到 31 微克/立方米

（四）治“气”12 年 污染红黑榜洗牌

经过十二年大气污染防治，我国空气污染版图已悄然重构。2024 年，河南和陕西包揽了 PM2.5 浓度最高的 10 个城市，其中安阳、焦作、渭南已连续 5 年在列，咸阳已连续 3 年在列。区域层面，汾渭平原除臭氧外其余污染物平均浓度均高于其他 2 个重点区域。从 2019 至 2024 年累计降幅来看，长三角地区的南京、淮南、无锡、淮北、滁州、镇江降幅排名靠前，综合改善情况好于区域内其他城市；而京津冀“2+36”城市的三门峡、德州、日照，汾渭平原的咸阳、铜川、宝鸡降幅排名相对拉了后腿。河北在二氧化氮和一氧化碳污染治理方面表现突出，年均浓度较 2021 年实现三连降，累计降幅均超过 15%。山西、河南和陕西均有 4 种污染物浓度较 2022 年实现两连降。

绿色代表无污染物超标，橙、紫、红、黑色分别代表 1–4 个污染物超标



三个重点区域六种污染物年平均浓度时间序列图

2024 年，四个直辖市空气质量呈现显著差异。天津作为北方重要工业基地，六种污染物浓度均高于河北省平均水平。其中臭氧与二氧化氮污染尤为严重，并且已连续 2 年呈现同步上升。北京市除二氧化氮外其他 5 种污染物浓度显著低于周边“山河四省”，但二氧化硫、一氧化碳出现小幅回升趋势。与二者相比，上海与重庆的燃煤指示性污染物二氧化硫、一氧化碳浓度偏高。

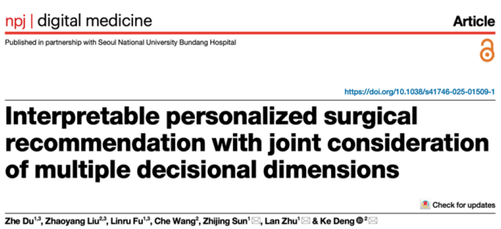
二、展望和建议

自 2018 年报告（五）起，研究团队持续呼吁空气质量级别标准修订。本次报告进一步指出，对于大气污染防治重点区域及周边城市，2024 年 PM2.5 优良天数比例已达 80% 以上，若将“良”的等级上界提高到 WHO“过渡时期”第 2 级指标，即 PM2.5 浓度 50 微克 / 立方米，“4+151”中 124 个城市 PM2.5 优良率仍可保持 70% 以上。这说明在当前阶段，我国已具备执行 WHO 建议的 PM2.5“过渡时期”第 2 级指标的基础。

报告认为，适时提高空气质量“优良”的门槛将形成强化公众健康防护、降低相应的医疗支出负担，激励空气治理成果巩固提升、促进空气质量持续改善，以及强化城市污染治理梯度格局，进一步调动污染区域治理积极性的三重效益。这有利于充分激活各级城市治理动能，协力实现“美丽中国”2035 年 PM2.5 的 25 微克 / 立方米目标，引领中国百姓走向更高质量的生活。

产学研融合创新

邓柯副教授团队携手协和医院团队联合推出智能手术推荐系统成果，破解盆底疾病治疗难题


npj | digital medicine
Article
Published in partnership with Seoul National University Bundang Hospital
https://doi.org/10.1038/s41746-025-01509-1
Interpretable personalized surgical recommendation with joint consideration of multiple decisional dimensions
Check for updates
Zhe Du[1,3], Zhaoyang Liu[2,3], Linru Fu[1,3], Che Wang[2], Zhijing Sun[1] ✉, Lan Zhu[1] ✉ & Ke Deng[2] ✉

清华大学统计与数据科学系邓柯副教授团队与北京协和医院妇产科朱兰、孙智晶教授团队合作发文，针对盆腔器官脱垂（POP）这一困扰全球女性的常见疾病，研发出多维度智能推荐系统（Multi-Dimensional Recommendation, MUDI），通过数据驱动科学平衡手术疗效、风险、成本、复杂度与患者需求，为医生和患者提供透明、个性化的手术方案推荐。相关成果发表于数字医学顶级期刊 npj Digital Medicine，同时获得国家发明专利授权。

- 作者团队 -

（注：刘朝阳、王掣为邓柯课题组已毕业博士）


杜　喆
第一作者、北京协和医院


刘朝阳
第一作者、清华大学


付琳茹
第一作者、北京协和医院


王　掣
作者、清华大学


孙智晶
通讯作者、北京协和医院


朱　兰
通讯作者、北京协和医院


邓　柯
通讯作者、清华大学

一、从“经验主导”到“数据量化”

POP 手术方案选择过程复杂，需综合评估患者解剖特征、并发症风险、医疗成本等十余项决策维度。传统方法高度依赖专家经验，而基层医疗资源短缺导致患者难以获得最优治疗。为解决经验主导带来的治疗资源不平衡的问题，基于协和医院近 20 年积累的手术及随访数据库的丰富信息及多中心的外部验证，MUDI 系统提供了专业、透明、便捷的最优治疗方案推荐。其创新性突破体现为：一是多维度量化引擎，构建手术特征图谱，涵盖疗效、风险、成本、复杂度及患者偏好五大核心维度；二是动态个性化推荐，基于临床数据学习医生决策偏好，实时调整各维度权重，支持“一患一策”；三是透明可解释逻辑，推荐结果附带量化依据，直观展示“为何选 A 而非 B”，助力医患高效沟通。

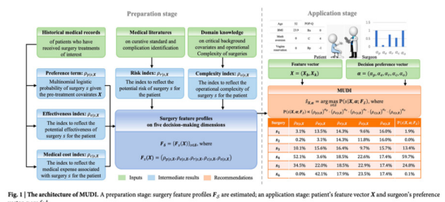


Fig. 1 | The architecture of MUDI. A preparation stage: surgery feature profiles F_{ij} are estimated; an application stage: patient's feature vector X and surgeon's preference vector α are fed.

图 1 MUDI 系统框架

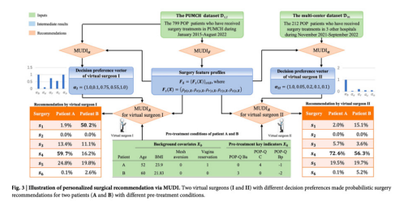


Fig. 3 | Illustration of personalized surgical recommendation via MUDI. Two virtual surgeons (I and II) with different decision preferences made probabilistic surgery recommendations for two patients (A and B) with different pre-treatment conditions.

图 2 通过 MUDI 系统实现个性化治疗方案推荐

二、准确率比肩顶尖专家，基层医生决策水平提升 27%

基于中国多家三甲医院超 1200 例 POP 患者数据开展验证，在内部验证方面，MUDI 的 Top-1 推荐准确率达 60%；在跨医院验证方面，MUDI 的 Top-1 推荐准确率仍能保持在 62%，与国内顶尖盆底专家水平持平，Top-3 推荐准确率高达 92%，显著优于普通专科医生（81%）和未受专科培训的基层医生（61%）。

三、从“黑箱预测”到“科学逻辑驱动”

与传统机器学习模型相比，MUDI 展现显著优势：一是抗过拟合提升模型稳定性，跨医院验证中性能波动 <2%，而神经网络等模型准确率下降超 10%；二是可解释性提升医生信任度，模型应用试验中 49%-97% 的医生愿参考 MUDI 调整决策，基层医生推荐准确率提升 21%-27%；三是可灵活扩展提升模型普适性，框架支持新增决策维度（如术后性功能保护），可快速迁移至其他疾病或更复杂的治疗决策场景。

四、在线平台开放，基层医疗有了“智能助手”

为便于临床应用，团队同步推出 MUDI 在线平台，基层医生面对盆腔器官脱垂手术决策时，直接访问网页即可使用一键式推荐、动态交互、知识库联动等功能，在系统辅助下作出更优、更个性化的决策，使更多患者享受到更高水平的医疗服务。

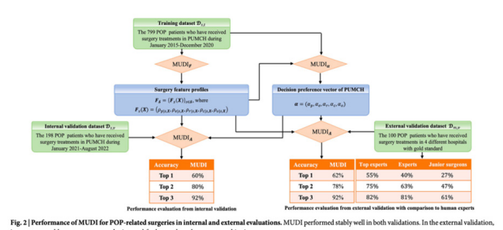


Fig. 2 | Performance of MUDI for POP-related surgeries in internal and external evaluations. MUDI performed stably well in both validations. In the external validation, it was comparable to top urogynecologists and far better than the experts and junior surgeons.

图 3 MUDI 数据验证结果

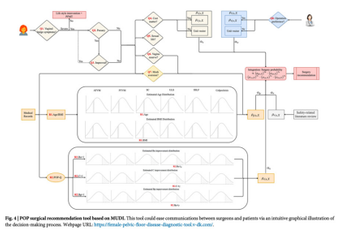


Fig. 4 | POP surgical recommendation tool based on MUDI. This tool could ease communications between surgeons and patients via an intuitive graphical illustration of the decision-making process. Webpage URL: https://female-pelvic-floor-disease-diagnostic-tool-t-dk.com/.

图 4 在线开放平台

邓柯副教授团队携手协和医院团队研发 SmartGyne 系统，实现妇科疾病高效分诊与诊断

近日，清华大学统计与数据科学系邓柯副教授团队与北京协和医院妇产科朱兰教授、孙智晶教授团队合作，在人工智能与医学交叉领域取得重要进展。团队成功研发出面向妇科疾病的全流程智能分诊和诊断系统 SmartGyne，实现了从初诊分诊到专科疾病诊断的一体化、智能化辅助决策。相关研究成果已正式发表于国际权威期刊《Artificial Intelligence in Medicine》，同时获得国家发明专利。

- 作者团队 -

（注：刘朝阳、王掣为邓柯课题组已毕业博士）

付琳茹
第一作者、北京协和医院

王　掣
第一作者、清华大学

刘朝阳
作者、清华大学

潘长在
作者、清华大学

杜　喆
作者、北京协和医院

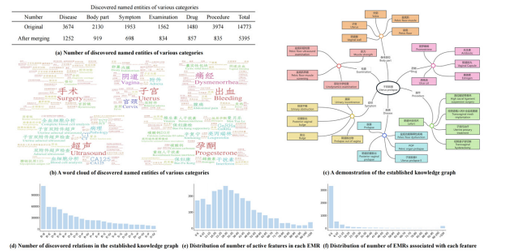
孙智晶
通讯作者、北京协和医院

朱　兰
通讯作者、北京协和医院

邓　柯
通讯作者、清华大学

一、从“海量病历”到“智能图谱”，破解妇科疾病分诊复杂性

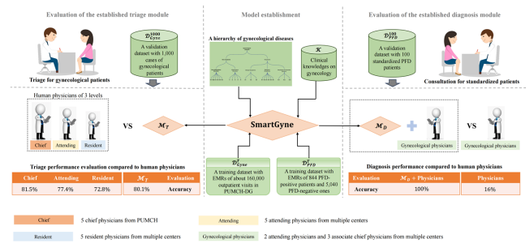
妇科疾病种类繁多、症状重叠度高，传统分诊高度依赖医生经验，基层医疗机构分诊能力不足易导致误诊、漏诊或转诊延误。SmartGyne 系统基于北京协和医院妇科门诊 23 万余份高质量电子病历，利用自然语言处理技术与知识图谱构建算法，自动提取并标准化关键特征，构建了覆盖 5395 个妇科专有概念、50 万余条关联关系的知识图谱，为 AI 分诊及诊断提供强大知识底座。



图一 知识图谱示例和结构化特征表的主要统计数据

二、分诊诊断一体化，准确率媲美顶尖专家

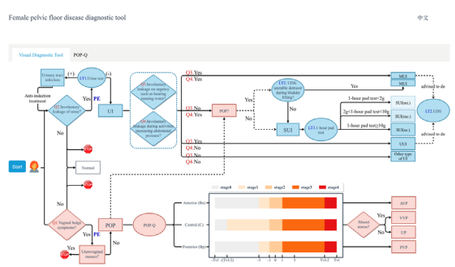
该系统创新性地提出“分诊＋诊断”双层决策架构：分诊模块对常见妇科疾病的整体分诊准确率达 80.1%；以盆底功能障碍性疾病（PFD）为示范，诊断模块准确率高达 99.4%。在与人类医生的对比测试中，SmartGyne 的分诊能力与主任医师水平相当，显著高于住院医师和主治医师。在诊断环节中，基于标准化患者（SP）的验证表明，使用 SmartGyne 辅助的医生诊断准确率从 16% 大幅提升至 100%，诊断效率明显提高。



图二 SmartGyne 的分诊及诊断表现

三、系统优势显著，助力医疗资源下沉

SmartGyne 既适用于三甲医院预检分诊、减轻人工初筛压力，也适用于基层医院辅助诊断、提升诊断准确率，同时提供清晰决策路径与可视化支持，医生无须深厚专业背景也能快速掌握使用。团队已公开系统源代码及相关知识图谱，并推出 PFD 诊断在线演示平台，助力各级医疗机构尤其是资源薄弱地区医院提升妇科疾病诊疗能力，推动高质量医疗公平可及。SmartGyne 的成功研发，不仅是人工智能技术在妇产科领域应用的重要里程碑，也为其他专科 AI 系统开发提供了可借鉴的新范式。



图三 PFD 诊断在线开放平台

俞声副教授团队携手哈佛大学团队推出电子病历结构化大模型

清华大学统计与数据科学系俞声课题组、哈佛大学生物统计系 / 生物医学信息系 /CELEHS 研究中心 Tianxi Cai 课题组与生物医学信息系 Isaac Kohane 课题组联合推出中英文电子病历结构化大型语言模型 GENIE（Generative Note Information Extraction，中文名：病历精灵）。

GENIE/ 病历精灵
https://huggingface.co/THUMedInfo/GENIE_zh_7b
https://huggingface.co/THUMedInfo/GENIE_en_8b

电子病历自由文本的分析处理是医学信息技术中重要的底层技术。通过将病历文本结构化为计算机系统容易处理的格式，病历数据可以产生非常广泛而有价值的应用，包括促进医学研究（患者筛选、临床数据提取）、院内与跨机构数据交换与治理、卫生统计与政策研究、医保数据分析，和各种临床决策支持系统的开发等。

由于电子病历涉及庞大的医学背景知识和特殊的行业书写习惯，其分析一直以来极具挑战性。从系统建设的角度，电子病历分析系统相比基础的自然语言处理系统还需要增加诸多专业分析模块，给部署与维护造成巨大挑战。GENIE 的先进性在于将所有自然语言处理环节和模块简化为一个单一模型，将所有分析结果一次性输出为 JSON 格式，并利用大模型结合背景知识分析复杂语言的优势，提供比传统系统更智能和更具可迁移性的病历分析能力。

相比 GPT-4o 等通用大模型，GENIE 具有免提示词、免样例、可在消费级硬件上本地部署、一次性输出多种分析、效果稳定等优势，特别是具有巨大的成本优势，适合机构进行大规模部署运行。

GENIE 目前提供以下分析内容：

• 术语识别：识别病历中的术语，并结合上下文自动进行缩写辨义和还原。可推断出常见隐形术语。对药品的商品名会自动额外输出其通用名称。

• 语义类型：判断术语对应医学概念在 BIOS 生物医学信息学本体系统中的分类。

• 叙述状态：对术语确定性、主语等属性的分类，包括存在、不存在、疑似、条件、假设、标题、非患者本人等。

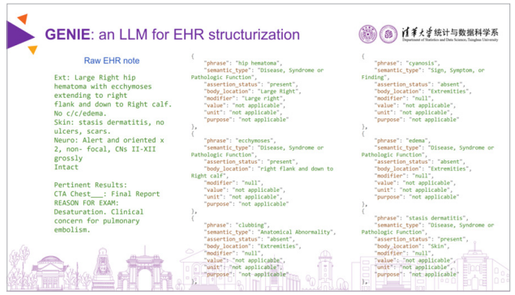
• 身体部位：疾病、症状、检查、手术等涉及的身体部位。

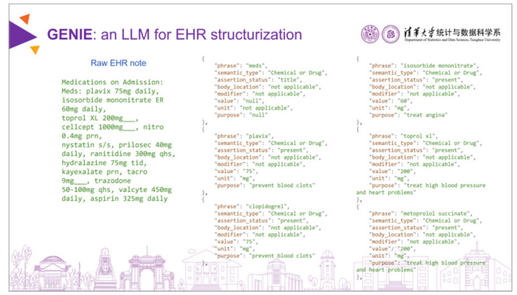
• 修饰词：身体部位以外的修饰，如急慢性、严重程度、数量、形态等。

• 数值：术语关联的数值，如化验结果等。可对连续数值（常见如生命体征、化验面板）进行主语推断。

• 单位：数值对应的单位，对省略的单位可进行推断。

• 目的：检查、手术、化验、用药等的目的。





刘军院士做客北京大学许宝騄讲座，分享粒子滤波的前世今生

2025 年 3 月 7 日，清华大学统计与数据科学系刘军教授受邀出席 2025 年度北京大学许宝騄讲座，并作学术报告“Particle Filtering：Past Present and Future”。讲座现场一座难求，来自清华、北大和多个兄弟院校的统计学者齐聚一堂，共同交流和探讨统计前沿思想。讲座由北京大学数学科学学院院长陈大岳教授主持。

许宝騄先生是我国概率统计学科的奠基人，在国内外享有崇高声誉。他既是民国时期中央研究院院士，也是中国科学院学部委员。许先生生前一直任北京大学教授。为缅怀先哲、激励后学，2009 年北京大学数学科学学院和北京国际数学研究中心决定联合主办一年一度的北京大学许宝騄讲座。每年邀请一名著名学者到北京大学访问，做一次公众演讲。历年演讲人分别是山东大学彭实戈教授、斯坦福大学黎子良教授、伦敦经济学院汤家豪教授、乔治亚理工学院吴建福教授、斯坦福大学王永雄教授、台湾“中央研究院”统计科学研究所郑清水教授、美国加州大学 David Aldous 教授、浙江大学林正炎教授、中国科学院马志明研究员、美国纽约大学牛查克教授、北京师范大学陈木法教授、北京大学耿直教授、中国科学院施展研究员、东京大学舟木直久教授、南方科技大学邵启满教授和美国芝加哥大学 Greg Lawler 教授。

讲座现场

刘军教授于许宝騄先生雕像前

刘军 教授

从历史到前沿：粒子滤波的演进

粒子滤波起源于序列重要性采样（SIS），最初用于解决与自回避随机游走（Self-Avoiding Random Walk）相关的组合优化难题。通过条件概率链式法则，SIS 构建了动态权重调整机制，避免了高维空间中的暴力搜索。随后，粒子滤波引入“预测 - 更新 - 重采样”循环，将蒙特卡洛模拟与贝叶斯推断无缝融合，成为动态系统状态估计的黄金标准。

近年来，粒子流（Particle Flow）与 KL 散度最小化的结合为 SMC 注入新活力。传统重采样依赖离散粒子跳跃，而粒子流通过构造连续概率流形（如 Stein 变分梯度下降），驱动粒子沿梯度方向平滑移动至目标分布的高概率区域，显著提升了高维空间中的样本效率。

重采样的核心作用与新进展

重采样作为粒子滤波的核心机制，通过淘汰低权重粒子缓解退化问题，其价值在于隐式构造更匹配目标分布的提议分布。然而，其理论本质仍存争议，尤其在高维场景下。本报告深入探讨了近期重采样技术的突破，包括：

分层重采样（Stratified Resampling）：通过控制方差优化粒子分布；

最优传输重采样（Optimal-Transport Resampling）：引入最优传输理论，提升粒子迁移效率；

基于梯度的粒子迁移（Gradient-Based Moves）：利用梯度信息实现更平滑的粒子调整。

面向未来的挑战与机遇

粒子滤波在处理复杂问题时展现出巨大潜力，但如何在高维场景下进一步提升效率、量化重采样的理论效应，仍是亟待解决的难题。

与会教师合影

刘军院士走进清华大学附属中学开展专题讲座

一排从左至右：邵光砚（数学老师）、刘军、张颖（语文老师）

高 81 级毕业合影

5 月 6 日刘军院士在清华附中举办“现代统计机器学习浅谈”讲座

2025 年 4 月 30 日（当地时间 4 月 29 日），美国国家科学院公布了新一届院士与外籍院士名单，120 名院士及 30 名外籍院士荣耀当选。其中，清华附中高 79 级校友、清华大学统计与数据科学系刘军教授荣膺美国国家科学院院士。这一殊荣，是对刘军教授多年来在学术领域卓越贡献的高度赞誉，也为正值建校 110 周年的清华附中增添了熠熠光辉。

1981 年，刘军从清华附中毕业。回首五年的校园时光，他感慨万千：“清华附中是我梦想启航的地方，这里有我人生的第一位‘数学引路人’邵光砚老师。”初二时，他加入数学小组，第一次真切地感受到了数学世界的浩瀚与神秘。尽管当时数学竞赛尚未广泛流行，但清华附中开放包容的学术氛围，如同一颗种子，在他心中种下了探索真理的渴望。

带着对数学的热爱，刘军考入北京大学数学系，随后赴美深造。在 Rutgers 大学学数学时逐渐察觉到数学理论与现实需求之间存在一定距离，然后转入芝加哥大学统计系学统计并攻读了博士学位。“统计学成为了我连接理论与实践的桥梁。”他坦言，统计学不仅为他提供了解决问题的有力工具，更塑造了他独立判断的能力。“在学术生涯中，做出几次正确的判断，远比经历无数次错误更有意义。”他毅然走出舒适区，不断探索未知、勇于创新，逐步形成了独具特色的研究方法和学术见解。

作为哈佛大学终身教授，刘军在统计学领域成果丰硕。他参与推动的生物信息学研究，为 AI 制药领域奠定了坚实基础；其早期提出的算法优化理论，至今已被引用近千次。面对 AI 技术的迅猛发展，他着重强调：“统计学是 AI 的基石，而新时代的挑战在于如何让模型更具可解释性、如何对预测结果赋予可信度指标、如何在各个细分领域让 AI 真正落地以提高生产效率”。他鼓励年轻学子以批判性思维直面复杂问题，在简化与复杂之间找到恰当的平衡点。

从清华附中的数学小组到哈佛大学的讲台，刘军以实际行动诠释了“自强不息，厚德载物”的深刻内涵。让我们以刘军等各届优秀校友为熠熠灯塔，汲取他们奋进的力量，在传承清华附中深厚文化底蕴与教育精髓中不断开拓创新，于探索未知的漫漫征途上奋勇前行！

媒体访谈 | 刘军院士接受《每日经济新闻》访谈分享：AI 更高层次发展可能需要突破当前模式下潜在瓶颈

在 2025 金融街论坛年会上，美国国家科学院院士、清华大学兴华卓越讲席教授、清华大学统计与数据科学系主任刘军教授接受了《每日经济新闻》记者（以下简称“NBD”）的现场采访。

在采访中，刘军教授表示，AI（人工智能）要实现更高层次的发展，可能需要突破当前大模型靠统计概率“预测下一个词元（Token）”模式内在形成的潜在瓶颈。虽然现行方法对这一模式有各种细节上的改进，但还没有找到另一个更高级的主导模式。

刘军教授一直从事贝叶斯统计理论、蒙特卡洛方法、统计机器学习、状态空间模型和时间序列、生物信息学等方向的研究，并做出杰出贡献，对大数据处理和机器学习领域有深远影响。刘军在采访中也谈到统计学自身发展。他指出，数十年来，生物医学和其他大规模数据生成技术的发展驱动了统计学基础学科持续前行。

刘军教授提出的“Gibbs 保守串抽样和指针”曾是生物学者寻找 DNA 和蛋白序列中精巧模式的最流行的两种算法之一，在了解基因调控和蛋白同源性方面有非常成功的应用。

NBD：大型语言模型依托大数据与统计概率，通过持续预测下一个字生成语言回复，这与外界以为的 AI 按照语义来推理判断有很大不同。你如何看待这一问题？

刘军：如果认为大型语言模型理解语义，那就是浪漫叙事。大语言模型的基石就是“Next Token Prediction”，即一个字一个字地预测，并未真正“理解”语言本身，尽管 DeepSeek、ChatGPT 等工具经常会给出惊艳的结果。“Next Token Prediction”在统计专业上又叫“Auto Regressive Model”，即自回归模型，通过词语（时间）序列间的关联，一步步向前预测。从这个角度看，它有可能成为 AI 模型向更高层次发展的一个潜在瓶颈，因此，下一步语言模型或许需要考虑如何突破这种思路。

事实上，目前已有人在尝试新的思路，不再是一个字一个字地预测，而是可以一段一段地生成，类似于先搭建一个句子框架，再填充具体字词。

在这一思路下，训练时每一个字是隐码，即字码所在位置为空，相当于通过去噪的方式生成结果。据反映，这一方法的结果还不错，但目前看很难说比“Next Token Prediction”效果更好。

这种整体规划式生成语言的模式，更像人类思考和表达过程，这一模式的继续发展可能会带来新的惊喜，但其前景尚存在不确定性。

NBD：统计学基础领域发展至今，已相当成熟。当前，该学科基础领域还面临哪些待解问题？

刘军：统计学是一门开放学科，换言之，它并非有一套固定的问题等着去解决，也不会因某类问题的解决而宣告“完成发展”。

统计学科的很多问题源自实践。比如，由于大家关注大模型的相关问题，统计学中高维数据方面的问题也获得更多讨论，这就是应用驱导下的问题和方法探索。

回顾统计学最初发展，该学科主要受天文星象学研究和社会人口研究驱动。进入 20 世纪，由于遗传学发展、农业育种、工业实验设计发展等，统计学进一步发展。

以英国统计学家费希尔为例，他同时也是知名遗传学家。针对群体遗传研究的需求，他提出了著名的概率论进化模型；针对农业实验的需求，他又提出了随机拉丁方设计方法，以及方差分析等统计推断理论与方法。

数十年来，医学、生物学的快速发展驱动统计学不断前行。我自身也在从事生物信息学方面的研究。以分子生物学为例，基因芯片信息中隐含着细胞内基因的表达与否。通过分析这些基因的遗传与变异规律，可判断特定变异与疾病的关联，进而为针对性药物开发提供支持。这些过程均需要统计学不断更新自身方法以适配需求。

NBD：外界也比较关注统计学的另一个应用场景，即股票投资。这也是一个概率决策的过程。从这一角度，统计学专业背景的投资者能否在股票投资上表现更优？

刘军：据我所知，投资机构确实愿意雇用具有统计专业背景的人。但对于个人投资，统计学学得扎实并不意味着个人投资业绩一定就好，因为投资还需要研究宏观经济等多个方面，并且需要大量训练、大量资金和精力。因此，对于个人而言，精力上可能不足以应对，资金量也无法支持频繁买卖。整体看，还是大型头部投资机构和对冲基金在投资上表现更佳。

刘军院士做客武大珞珈讲坛，谈蒙特卡洛方法视角下的人工智能发展

2025 年 5 月 12 日，清华大学统计与数据科学系刘军院士做客武汉大学珞珈讲坛。武汉大学校长张平文院士出席并为其颁赠珞珈讲坛纪念牌。

刘军从 Geoffrey Hinton 的研究轨迹开始，阐述了人工智能再次兴起的关键，为大家介绍了蒙特卡洛方法的研究历史、发展历程和关键人物，并详细解释了近年来蒙特卡洛方法的研究方向和具体实例。蒙特卡洛方法作为一种基于随机采样的数值计算技术，包含着多种方向——重参数化、扩散采样、重采样、最优传输以及变分近似技术，蕴含着序贯蒙特卡洛算法（SMC）、马尔科夫链蒙特卡洛算法（MCMC）等丰富的算法。蒙特卡洛方法能够成为处理优化积分问题的重要工具，在人工智能领域具有广泛应用和重要意义。

武汉数学与智能研究院副院长杨志坚教授主持讲坛。互动交流环节，在场师生踊跃提问，刘军围绕蒙特卡洛方法的细节、与 AI 发展的融合、统计学的重要性和学习方法等方面展开交流探讨。

刘军：统计和数据科学是 AI 落地的一把钥匙

2025 年 10 月 27 日至 30 日，2025 金融街论坛年会在北京金融街举行。金融科技大会作为论坛年会特定版块，与金融街论坛同期举行。10 月 30 日上午，由中国民主建国会北京市委员会金融委员会、首都经济贸易大学北京数字经济发展研究院、中关村金融科技产业发展联盟合作承办的“2025 金融科技大会平行论坛——AI+ 金融专题论坛”以“智领未来金融科技创新”为主题，在中关村金融科技特色产业园举办。清华大学兴华卓越讲席教授、统计与数据科学系主任刘军出席论坛并发布题为《AI 落地数字经济和金融的思考》的主题演讲。刘军教授认为“人工智能的落地必须与统计学和数据科学深度融合”，他指出：“单纯依赖大数据而缺乏严谨的统计思考可能导致模型偏差与决策失误”“统计与数据科学是 AI 落地的钥匙”；并通过城市治理，金融风控等案例，展示了“AI+ 统计”如何为数字经济与金融提供可靠、可解释的解决方案。

一、AI 的四大基础思想：从底层逻辑理解大模型

要理解 AI 如何落地，需先厘清其核心技术的演变才能理解技术怎么做迭代。我经常给大家打个比喻，你可以把 AI 大模型想象成是一个波音飞机，统计学和机器学习相当于空气动力学和热能学这类基础学科直接支持飞机，那么下一代飞机的发展不可能只关注如何把飞机的窗户做好？如何让椅子更舒服？飞机制造的工艺学涉及很多领域，还有许多经济方面的内容。我们正在尝试使用各种具体的 Agent，当然也需要了解其工作原理，因此我总结了 AI 发展的四大底层逻辑：1. 深度学习模型：从线性回归开始最原始的 AI 思想是线性回归——用 x 推测 y，本质是统计中的“预测”。几百年前高斯提出的线性模型是统计学的基础，后来发展出非线性模型，但核心仍是“用简单函数拟合复杂世界”。尽管神经网络发展得很早，但真正用到非常大型的网络是深度神经网络（Deep Neural Network）的出现，也就是最近二十年的事情。大模型的发展对很多原始的科技直觉具有一定的反思作用。以前科学界都认为模型越简单越好，理解一个事情要简单、直接，所以大家也希望基于一个非常简单的模型做推导。但大模型的出现对这种思路是一个冲击，使用一个参数非常多的模型也能对整个事物描绘得很好，而且会更好。大模型的出现同时催生出很多统计学习方法，如对比学习、混合专家模型、自监督学习等。但大模型到底是什么，怎么才能做好，怎么能有泛化性？这一系列的问题都有待于大家去思考。大家可能都知道大模型有幻觉的问题，同时也有智能涌现的特点，但这些特点到底是怎么产生的？为什么会出现？有些人把它看成玄学，认为机器慢慢具有智能了，我不信这一套，我认为这是瞎扯。实际上在物理、数学领域都有类似的现象，有一个叫 Phase Transition 的理论就试图对这类现象进行解释，比如说水变成冰，怎么在一个温度下水忽然变成冰，或者冰在一个温度下忽然变成水。大模型的推导能力可能也存在类似的现象，有些数学家正在研究这个事情。2. 大语言模型的基础：文字和文本在欧式空间的嵌入表达另外一个革命性的思想，就是把文本离散的数据变成向量，通过空间向量去表示两个词的距离。比如说“父亲”跟“母亲”这两个词是有空间距离的，我们认为“父亲”和“母亲”是紧密连接的，但有时候“父亲”跟“皇帝”也有一定的联系，但是不太知道怎么把它们搁在一起？这个就叫深层知识表征（Embedding），是非常革命性的想法。这种思路是所有 DeepSeek、Chat-GPT 的基础，就是所谓的 Next Token Prediction，因为这种方法可以把这些词扩展到句子或者短文。把词汇变成空间向量就可以用机器学习的方法实现 Next Token Prediction，这是一个基本的框架，同时也提示我们，大模型的价值远不止于对话交互——那些看似“老人都能理解”的生成结果，仅是其能力的冰山一角。真正值得挖掘的，是深入其底层训练过程中动态生成的深层知识表征（Embedding），并将其有机整合到模型内部或外部再次调用，可能会产生更好的结果。这种“向底层要价值”的思路，往往能解锁更精准、高效的应用潜力。比如用 AI 做对冲基金的公司们，就会用这个思路，他们从来不会用现有模型做这件事情，而是把模型拆开了，应用到具体的问题里面去。3.Stable Diffusion：生成式 AI 的基础稳定扩散（Stable Diffusion）的核心思想可以用一个直观的类比理解：它像一场“逆向的加噪实验”——我们先将一张清晰图像逐步加入随机噪声，直到它完全变成一片混沌的噪声；而模型的训练目标，正是反向学习“如何从这片噪声中一步步还原出原始图像”。具体来说，这个过程通过迭代的加噪 - 去噪训练 实现：模型首先学习“从清晰图像到噪声”的正向变换（类似给图像“加噪”），再通过深度神经网络反向推导“从噪声到图像”的逆向生成路径。每一次加噪都是对图像信息的“破坏”，而模型的任务是捕捉这种破坏的规律，最终掌握“从混沌中重建秩序”的能力——这正是生成式模型的底层逻辑。这一“加噪 - 去噪”的迭代学习框架，已成为几乎所有生成式模型（如图像生成、文本生成）的基石。它的巧妙之处在于，通过模拟“破坏 - 重建”的过程，让模型在不直接学习“好图像是什么样”的前提下，自下向上掌握了生成高质量内容的底层规律。这是一个非常奇妙的想法，现在基本所有生成式模型都用这一套框架。4. 强化学习关于 AlphaGO，大家在金融里面也会用这种策略，相当于对话式的模型调优，你可以先有一个策略出来，然后通过模型推导看结果，再根据这个结果去优化策略。强化学习就是这样一个思路。但是最重要的一点，在往前搜索的时候是要一些随机的蒙特卡洛搜索，并不需要把所有情景都概括。

二、关于大模型的应用：关注数据缪误

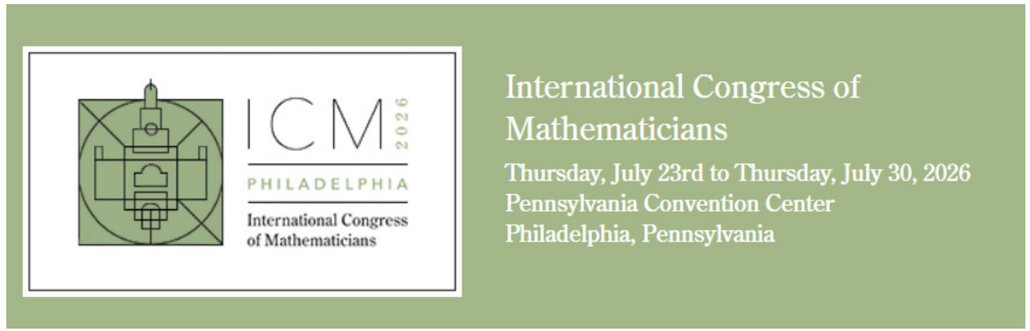
以上四点是比较核心的思路，概括了现在 AI 技术的底层逻辑。我再提一点需要注意的事项，就是关于数据的使用，即使是针对小数据也要进行遴选，因为数据往往是有偏差的。在这里，我们并不是说大模型有什么问题，但是咱们在使用的时候一定要重视数据的质量，要不然推理的结论就会出现偏差。这里面有一个例子，大家知道，谷歌多年来一直在持续研发大模型，曾经有一个非常轰动的工作发表在《自然》杂志上，核心是运用用户搜索信息预测流感爆发。作为最有钱，体量最大的互联网公司，这个对他们来讲本不应该是一件很难的事情，但是他们也犯了一些低级错误。比如，让跟季节有关的信号主导了对流感的预测，造成了对流感的预测本质上变成了对季节的预测，从而出现了偏差。这种被“大数据”和“人工智能”掩盖的错误往往更隐蔽、更有迷惑性，但危害也更大、更严重。总的来讲，对数据的理解是大模型应用非常重要的一件事情。

三、AI+ 统计赋能金融科技

我们系和我自己在应用方面做了一些尝试。一是数字治理。我们帮海关总署做监管方案，因为他们数据非常小，直接用大模型很难做，这就需要用非常精巧的 Agent 完成这个事，其中包括很多基于统计原理进行实验设计和分析等，还有市场监管的方面也可以用这些基本的想法。二是风险控制。在我国，风险控制是非常重要的目标和手段。我们用大模型把大量文本数据整合起来，然后用文本的信息挖掘企业之间的关联关系，这里面使用了一些技术手段，包括随机分析，蒙特卡洛模拟和图神经网络等方法做风险预估，实现智能风控。这个事我们正在尝试，似乎是一个很有前景的方向。现在很多的银行也在使用这种方法去预测美元的风险，现有的技术条件下如能整合更多的模态数据，预测的结果可以做的更好。反洗钱是金融监管很重要的一个目标，用统计和 AI 技术整合在一起，可帮助金融机构识别风险资金。最后再提一点公共管理方面的，包括用生成式 AI 挖掘弱势人群的需求，用一些动态的模型可以去做交通拥堵方面的控制。由于时间关系，最后总结几点，一个是再次强调对数据的理解，对我们发展有用的 Agent 具有非常重要的意义；另外，咱们国家对数据安全非常重视，但对数据进行分析和挖掘也非常重要，应该更加鼓励。

统计和数据科学是 AI 落地的一把钥匙。

陈松蹊教授受邀在 2026 国际数学家大会作 45 分钟报告



国际数学联盟近日公布了 2026 年国际数学家大会（ICM 2026）报告人名单，清华大学陈松蹊教授获邀作 45 分钟学术报告。国际数学家大会是数学领域最具影响力的学术盛会，每四年举办一次，45 分钟报告代表学者在相关研究方向的重要贡献得到国际数学同行的高度认可。

陈松蹊教授长期从事统计学与数据科学领域的研究，致力于发展和推广统计学理论和方法在超高维数据统计推断、高分辨率数据同化、数学地球物理学、气候变化、中国空气质量评估应用研究、计量经济学以及人口统计等方面的应用，并取得了一系列突破性成果。此次受邀报告，彰显了其在国际数学界的学术影响力。

ICM 2026 将于 2026 年 7 月在美国费城举行，预计汇聚全球顶尖数学家分享前沿成果。

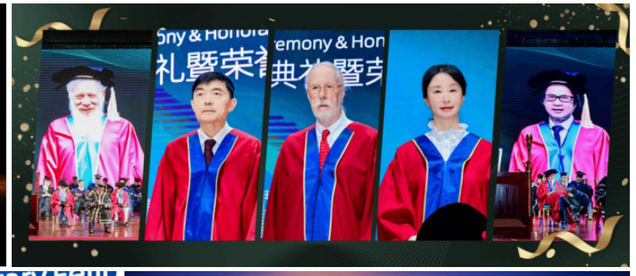
Song-Xi Chen
17 - Statistics, ML
University Chair Professor
Tsinghua University

陈松蹊院士获颁北师香港浸会大学（BNBU）荣誉院士荣衔

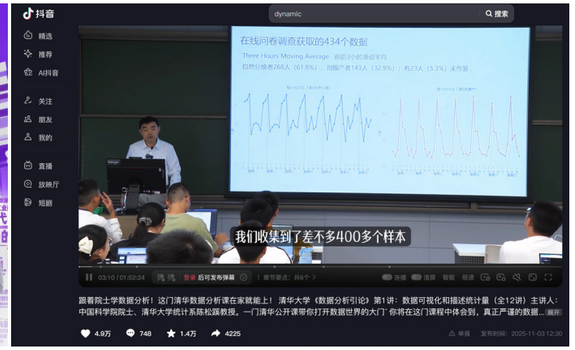
2025 年 6 月 8 日，北师香港浸会大学（BNBU）举办第十七届毕业典礼暨荣誉院士颁授典礼，五位杰出人士获颁荣誉院士荣衔。中国科学院院士、清华大学统计与数据科学系讲席教授、中国数学会概率统计分会理事长陈松蹊教授，诺贝尔经济学奖得主、美国国家科学院院士 Robert John Aumann 教授，美国著名汉学家、威斯康星大学麦迪逊分校荣休讲座教授倪豪士（William H. Nienhauser, Jr.）教授，全国脱贫攻坚先进个人、上海真爱梦想公益基金会创始人潘江雪女士，香港 SML 集团创始人、饶宗颐学术馆之友创会会长孙少文博士同获此殊荣。

典礼后，荣誉院士出席媒体见面会。陈松蹊院士指出，当前各行各业对数据分析的需求巨大，但高校培养的人才数量仍显不足。就统计学人才培养的问题，他建议高校可以借鉴工商管理硕士（MBA）的模式，鼓励相关行业的在职人士到高校进修数据分析技能，再将所学知识应用到实际工作中。“通过 AI+ 统计学，我们将能够减少计算负担，提高 AI 算法的可解释性和可重复性，并使 AI 应用更加绿色环保。”

恭贺荣誉院士晚宴上，香港浸会大学校长资深顾问黄伟国与北师港浸大校长陈致、常务副校长周荫强依次颁发荣誉院士证书。



跟着院士学数据分析！清华大学《数据分析引论》12 讲公开课抖音上线



2025 年 11 月，中国科学院院士、清华大学统计与数据科学系讲席教授陈松蹊主讲的《数据分析引论》公开课在清华大学抖音平台上线后迅速引发广泛关注，第一期课程“数据可视化和描述统计量”收获 4.9 万个点赞与 4225 次转发，成为近期热门的线上教学课程。

这门面向全民开放的清华精品课程共 12 讲，将系统讲授数据分析的完整方法论体系。陈松蹊院士在首讲中强调：“真正严谨的数据分析远不止于简单的数据整理和可视化——它是一整套系统的方法论，涵盖了数据的收集、对随机现象的理解、规律的发现、因果关系的辨析，直至智能预测的全过程。”

作为国际知名的统计学家，陈院士将带领观众深入数据科学核心领域，从基础理论到前沿应用，帮助学习者构建扎实的分析思维框架。课程以完全免费的形式呈现，公众足不出户即可畅享顶尖学府的优质教学资源，这既是陈院士践行社会责任、以学术回馈社会的具体行动，也彰显了顶尖学者推动知识普惠、服务全民教育的时代担当。

目前该系列课程已在清华大学官方抖音账号持续更新，为广大数据分析爱好者打开通往数据世界的大门。

课程网址：https://www.douyin.com/video/7568351745437519146

陈松蹊院士引领书院学子走进统计与数据科学

2024 年 10 月 17 日下午，书院“从游讲堂”第 23 讲在新清华学堂三层实验剧场举行。中科院院士、清华大学统计与数据科学系陈松蹊教授受邀，以《走进统计与数据科学》为题，为清华学子带来了一场精彩纷呈的学术讲座。数学系杨瑛教授、统计与数据科学系侯琳副教授、致理书院副院长闫永彬副教授等出席讲座。讲座由书院管理中心主任苏芃主持。

讲座伊始，陈松蹊院士由科学研究领域指引探索者们穿越未知迷雾的永恒灯塔——“规律”，引出统计学的世界观、方法及名称由来，逐层揭开统计学的神秘面纱，使同学们更加直观地理解这门学科的内涵和外延。

随后，陈松蹊院士带领大家系统回顾了统计学和国内外统计学教育及学科的发展史。他详细梳理了从古典统计到现代统计的演变历程，并着重讲解了其中具有里程碑意义的科学家及其研究理论和成果，如弗朗西斯·高尔顿（Francis Galton）、卡尔·皮尔逊（Karl Pearson）和罗纳德·艾尔默·费希（Sir Ronald Aylmer Fisher）等，他们的杰出贡献不仅奠定了统计学的基础，更为后续的研究提供了重要的思想支撑。关于统计学教育和学科的发展，他指出，国际上始于 1907 年卡尔·皮尔逊（Karl Pearson）在伦敦大学学院（UCL）成立第一个统计系，二十世纪三四十年代在美国全面盛行，现在英语国家几乎都有统计系；而我国，虽于西南联大时期已开设了非常丰富的统计学课程，但目前为止，设置独立统计学院系的综合类大学并不多，师资也较为紧缺，前行仍旧任重道远。值得欣喜的是，清华大学今年 7 月份成立统计与数据科学系，聚焦多个前沿领域，预计提供本、硕、博等多层次教育项目，致力于培养能够应对现代数据科学挑战的复合型人才，这标志着清华大学在统计学与数据科学领域的进一步拓展和深化。

此外，陈松蹊院士结合自身科研经历、研究进展等展示了统计学如何在农业、医学、社会科学、环境科学、经济学、民生等各领域发挥关键作用、为科学研究提供精准的数据分析和科学推断。陈院士风趣幽默、深入浅出的讲解方式和生动形象的案例分析，使大家深刻体会到了统计学作为一门基础性学科的重要性和广泛应用前景。

问答交流环节中，陈松蹊院士详尽耐心地回应并解答了同学们关于统计学与数据科学在未来各行业的具体应用、如何赋能工科学科、当下要解决的问题与前景，以及 AI 在统计学中起何作用、学科交叉背景下如何面临新领域中专业知识不熟悉的挑战等问题，每一位听众都受益匪浅。

讲座最后，苏芃主任向陈松蹊院士表示感谢，未央书院封羽真同学、致理书院闫永彬副院长为陈院士献花、赠送书院讲座纪念奖牌。

书院将梅贻琦先生的“从游”理念融入育人文化，于 2020 年开始，设立“从游讲堂”，邀请在人才培养和科学研究领域有着深厚积淀的学者，于春风化雨间赋予同学们更多砥砺前行的力量。本次讲座共吸引校内外 100 余名师生热情参与。

陈松蹊院士“走进统计与数据科学”本科招生宣讲纪实
——构筑高中学子通向学术前沿之梯

2025 年春，陈松蹊院士先后走进包头九中“崇德博学”讲堂、长沙一中逸夫楼、福州三中力行楼学术厅，为三地近千名拔尖学子同步开启《走进统计与数据科学》的主题宣讲。这场跨越理论与实践的学术盛宴，循百年统计长河，以跨学科视角融通古今：既用广博案例溯源学科历史、解码统计基因，更以生动叙事揭示数据科学在当代社会的核心价值，点燃少年投身数据科学的热情。

作为国际知名统计学家，陈松蹊院士一登台便以谦和的风度点燃全场热情。他系统梳理统计学科百年脉络，介绍清华大学统计与数据科学系的创建及统计学在工业、科技、医药、人工智能等领域的广泛应用；从数据科学赋能国家治理、农业升级、工业优化等角度，深入阐释回归分析、极大似然估计等核心方法；并以个人研究为“样本”展示跨学科生命力：团队借助时空统计模型揭示京津冀雾霾扩散机制，为环境治理提供科学支撑；构建海洋物理高分辨率数据集，分享“海洋治理”等前沿领域的创新成果。

他强调培养数据思维与科学精神并重，激发学生们对统计学的思考和对学习数据科学的向往。台下学子屏息凝神，眼中闪动着求知的光芒。

每场讲座均设互动问答环节，陈松蹊院士回答数学与统计区别、统计应用等问题，肯定学生思辨能力。讲座结束，同学们纷纷表示收获“统计学钥匙”，立志开启通往数据时代的大门。

陈松蹊院士在“进博会”发表主题演讲，探讨统计学在智慧海关建设中发挥的重要作用

央视新闻频道“朝闻天下”报道本次论坛

芬兰农林部部长萨丽·艾萨叶女士发表主题演讲

陈松蹊院士发表主题演讲

陈松蹊（后排左一）院士见证“技术性贸易措施评议基地”签约

陈松蹊院士在论坛间隙与芬兰农林部长萨丽·艾萨叶亲切交谈

2023 年 11 月 7 日，由海关总署主办的“2023 非关税贸易措施高质量发展论坛”在上海国家会展中心隆重召开，该论坛作为“第六届中国国际进口博览会”配套高峰论坛之一引起广泛关注。论坛以“智慧海关助力贸易安全与通关便利化”为主题，围绕“智慧监管筑牢国门生物安全边境”、“规制一致性推动智慧通关便利化”等议题展开探讨。海关总署总工程师韩森代表海关总署致欢迎词，芬兰农林部部长萨丽艾萨叶女士、中国科学院院士陈松蹊、中国工程院士沈建忠、单杨、庄松林等专家学者分别发言，来自海关总署各司局及地方海关的官员及专家、相关专业机构代表、知名企业家及多国驻华使节百余人受邀出席了论坛。

在本次论坛中，陈松蹊院士作了题为“运用统计学智慧推进通关便利化”的主题发言。陈院士在发言中指出，随着“一带一路”倡议取得巨大成功，中国以自身的快速发展带动全球贸易格局发生了深刻的变化。面对新形势、新机遇，运用前沿技术手段推动通关便利化，在严守国门安全的前提下，提高通关的规范化、智能化水平，更好地服务于国家“高质量发展”和“一带一路”战略，具有重大的经济效益和政治意义。然而，从科学逻辑上来讲，“推动通关便利化”与“守护国门安全”却是一对矛盾：推动通关便利化，客观上以简化口岸查验措施为前提；但这势必会削弱海关在口岸发现潜在问题的能力，为守护国门安全带来隐患。如何处理好这对矛盾，在有效推动通关便利化的同时，强化而不是

陈松蹊院士与邓柯副教授在论坛现场

削弱国门安全，是对中国海关管理能力的一大考验。要做好这件事情，从科学逻辑上讲，只有一种途径，那就是：持续提升中国海关的风险感知和预警能力，在口岸查验强度降低的情况下，大幅度提高查验的针对性和精准度，从而保证国门安全。要实现这个目标，运用科学、系统的方法，对错综复杂的政务大数据进行有效治理，并实施有目标、有深度、高质量的政务大数据分析，是重中之重。作为一门系统研究数据采集、处理、分析、解读的学科，统计学是实现上述目标的核心基础，由统计学所产生的数据智能是建设智慧海关的关键力量。

陈松蹊院士还在发言中特别介绍了清华大学统计学中心邓柯副教授团队与海关总署在“进出口食品安全风险评估监管”和“技术性贸易措施综合指数研究”等方面的合作成果，为统计学科与海关总署更紧密的合作打下了良好的基础。陈院士还指出运用统计学理论与方法更好更快地推动“智慧海关”建设，仍有巨大的空间；充分运用海关口岸贸易数据，深入研究由于国际贸易所引起的“碳排放国际转移”，将对于我国更好应对碳排放挑战具有重大的战略意义。相关发言在与会听众中引起了强烈共鸣。

央视新闻特别提及邓柯副教授在海关大数据分析方面的工作为提升我国口岸检测能力做出了重要贡献

央视新闻对论坛活动进行了报道。在新闻视频中，高度评价了统计学在海关大数据分析和智慧海关建设中所发挥的作用，并特别播报了陈松蹊院士的发言片段。

教育部统计学“101 计划”持续推进：教材建设会与骨干教师研修班相继举办

教育部基础学科系列统计学“101 计划”自 2025 年 4 月启动教材建设以来取得系列重要进展。4 月 12-13 日，由统计学“101 计划”专家组主办、江西财经大学承办、高等教育出版社协办的教材建设推进会在江西景德镇举行，来自清华大学（牵头高校）、北京大学、中国人民大学等 16 所参与高校的 80 余位海内外知名统计学者参会。

统计学“101 计划”牵头专家、清华大学讲席教授、中国科学院院士陈松蹊教授在开幕式上致辞。他指出，在 AI 与数据驱动的新时代，统计学作为支撑理工农医领域方法论的核心学科，正面临学科边界模糊化的挑战。编写高质量教材是吸引优秀人才、夯实学科根基的关键举措，不仅关乎统计学科的发展，更是为学科未来数十年发展奠定基础。陈院士呼吁编写团队以“打造传世经典”为目标，力争 11 月出版系列教材，集中精力提升教材质量，共同完成这项具有历史意义的教材编纂工程。

2025 年 7 月 28-30 日，该计划核心课程骨干教师研修班在昆明举办，由云南大学承办，高等教育出版社、全国高校教师网络培训中心协办。本次研修班正式发布了统计学“101 计划”核心课程体系白皮书，标志着教材建设成果向教学实践转化迈出关键一步。系列活动体现了“101 计划”在教材编纂与师资培养两大基础环节的协同推进，为统计学科未来数十年发展筑牢基础。

总 编 辑： 刘　军、邓　柯、侯　琳
执行编辑： 侯禹珊、于浩洋

见统计 STATISTICS　见万物 UNIVERSE

数学为基 | 统计为核 | 数据为用

致力于培养具备理论深度与跨界创新能力的拔尖人才
助力中国统计与数据科学人才培养再上新台阶

清华大学统计与数据科学系

地址：北京市海淀区清华大学自强科技楼 4 号楼
（吕大龙楼）715
电话：+86-10-62786091
邮箱：stats@tsinghua.edu.cn
网址：www.stat.tsinghua.edu.cn

Department of Statistics and Data Science, Tsinghua University
Address: Room 715, Lyu Dalong Building, Tsinghua University, 100084
Tel: +86-10-62786091
E-mail: stats@tsinghua.edu.cn
Website: www.stat.tsinghua.edu.cn

扫描二维码
关注官方微信公众号
清华大学统计与数据科学系

设计支持：彩虹七炫品牌设计